ЗАТВЕРДЖЕНО

Науково-методичною радою
Державного університету
«Житомирська політехніка»

протокол від 31 серпня 2023 р.
№ 10

КОНСПЕКТ ЛЕКЦІЙ
з навчальної дисципліни
«Аналіз великих даних у фінансах / Big Data Analytics in Finance»

для здобувачів вищої освіти освітнього ступеня «магістр»
072 «Фінанси, банківська справа та страхування»
освітньо-професійна програма «Фінанси, банківська справа та страхування»
факультет бізнесу та сфери обслуговування
кафедра фінансів та цифрової економіки

Рекомендовано на засіданні кафедри фінансів та цифрової економіки
28 серпня 2023 р.,
протокол № 09

Розробник: к.е.н., доцент кафедри фінансів та цифрової економіки,
ОВАНДЕР Наталія

Житомир
2023

ЗМІСТ

ВСТУП

Конкурентні переваги та інновації в цифровому світі базуються на точному аналізі та розумінні великого масиву даних, які постійно створюються. Для того, щоб використовувати знання, надані великими даними, і в режимі реального часу діяти на їх основі, важливою є бізнес аналітика, яка в епоху великих даних стає найбільш затребуваною професією у наступному десятилітті XXI століття від стартапів до великих корпорацій. Високопрофесійні менеджери по роботі з великими даними поєднують в собі бізнес-знання, знання технології Big Data та аналітичні навички для більш швидкого прийняття рішень, що забезпечують новаторське зростання та інновації в будь-якій організації.

Навчальна дисципліна призначена для підготовки нового покоління професіоналів в області менеджменту/бізнес-аналітики, які володіють методами обробки великих даних, сучасними аналітичними інструментами, методами та моделями прийняття управлінських рішень, які інформаційно та інноваційно зорієнтовані на створення нових цінностей для клієнтів.

За допомогою найновіших цифрових технологій (хмарні обчислення, мобільні технології, соціальні технології, машинне навчання, інтернет речей) менеджери/бізнес-аналітики здатні ідентифікувати, збирати, аналізувати великі масиви даних, інтерпретувати та трансформувати їх для глибшого розуміння бізнесу в реальному часі з метою більш швидкого прийняття кращих рішень, підвищення ефективності роботи компанії, оптимізації її конкурентоспроможності, пом’якшення ризиків.

Метою викладання навчальної дисципліни є – підготувати фахівців зі знаннями у галузі великих даних; надання фахівцям навичок у галузі діяльності з удосконалення організації праці, виробництва та управління даними; вивчити принципи, методи та форми організації управління великими даними.

Завданнями вивчення навчальної дисципліни є:

– оволодіти теоретичними основами і набути практичних навиків щодо аналізу економічної інформації;

– вміти видобувати знання шляхом інтеграції та аналізу великих даних, отриманих з різноманітних та різнорідних джерел інформації;

– оволодіти теоретичними основами щодо методів оцінювання достовірності моделі та її параметрів, прогнозних характеристик моделі, побудованих на основі великих даних;

– демонструвати знання сучасного рівня технологій інформаційних систем, практичні навички використання прикладних і спеціалізованих комп’ютерних систем та середовищ з метою їх запровадження у професійній діяльності.

Тема 1. Поняття великих даних та їх класифікація. Концепції великих даних

1. Суть Big Data
2. Історія виникнення терміну Big Data
3. Характеристики Big Data
4. Основні типи Big Data

1. Суть Big Data

За прогнозом компанії IDC, обсяг даних у світовому масштабі щороку зростатиме приблизно на 61%. До 2025-го цей показник досягне 175 Збайт. За часів, коли інформація генерується дуже швидкими темпами, важлива роль відводиться її обробці та аналізу. Для вирішення цих завдань використовують технології Big Data. Розглянемо докладніше, що таке великі дані, у чому їхня особливості й де їх застосовують.

Big Data (великі дані) – це поєднання структурованих, напівструктурованих та неструктурованих даних, які можуть бути видобуті для отримання інформації та використані в проектах машинного навчання, прогнозного моделювання та інших передових програм аналітики.

Системи, які обробляють і зберігають Big Data, стали загальним компонентом архітектур управління даними в великих організаціях.

Компанії використовують накопичені в їх системах Big Data для поліпшення операцій, забезпечення кращого обслуговування споживачів, створення персоналізованих маркетингових кампаній на основі конкретних уподобань клієнтів і, зрештою, підвищення прибутковості.

Підприємства, які використовують великі дані, мають потенційну конкурентну перевагу перед тими, хто цього не робить. Вони можуть приймати швидші та більш обґрунтовані ділові рішення, за умови, що вони ефективно використовують дані.

Наприклад, Big Data можуть надати компаніям цінну інформацію про своїх клієнтів. Вона може бути використана для вдосконалення маркетингових кампаній з метою збільшення залучення клієнтів та коефіцієнтів конверсії.

Крім того, використання великих даних дозволяє компаніям дедалі краще орієнтуватися на споживача.

Історичні дані та дані в реальному часі можуть бути використані для оцінки мінливих уподобань споживачів. Це дозволить підприємствам оновлювати та вдосконалювати свої маркетингові стратегії та ставати більш чутливими до бажань та потреб клієнтів.

Великі дані також використовуються медичними дослідниками для виявлення факторів ризику захворювання та лікарями для діагностики захворювань та станів у окремих пацієнтів.

Крім того, дані, отримані з електронних медичних записів, соціальних мереж, Інтернету та інших джерел, надають організаціям охорони здоров’я та

державним установам найсвіжішу інформацію про загрози інфекційних захворювань чи спалахи захворювання.

В енергетичній галузі Big Data допомагають нафтогазовим компаніям визначати потенційні місця буріння та контролювати експлуатацію трубопроводі. Так само комунальні служби використовують їх для спостереженням за електричними мережами.

Фірми фінансових послуг використовують системи Big Data для управління ризиками та аналізу ринкових даних у реальному часі.

Виробники та транспортні компанії покладаються на великі дані для управління своїми ланцюгами поставок та оптимізації шляхів доставки.

Інші сфери використання включають – реагування на надзвичайні ситуації, запобігання злочинності та побудова розумних міст.

Великі дані надходять з безлічі різних джерел, таких як системи ділових операцій, бази даних клієнтів, медичні записи, журнали кліків в Інтернеті, мобільні додатки, соціальні мережі, сховища наукових досліджень, машинно генеровані дані та датчики даних в реальному часі, що використовуються в Інтернеті речей.

Дані можуть залишатися в необробленому вигляді в системах великих даних або попередньо оброблятися за допомогою інструментів інтелектуального аналізу даних або програмного забезпечення для того, щоб вони стали готові до конкретного використання в аналітиці.

Можна навести наступні приклади Big Data:

Порівняльний аналіз. Включає вивчення показників поведінки користувачів та спостереження за діями клієнтів у реальному часі з метою порівняння продуктів, послуг та авторитету однієї компанії з продуктами її конкурентів.

Відстеження соціальних мереж. Це інформація про те, що люди говорять у соціальних мережах про конкретний бізнес чи товар. Ці дані можуть бути використані, щоб допомогти визначити цільову аудиторію для маркетингових кампаній.

Маркетинговий аналіз. Сюди входить інформація, яка може бути використана для просування нових продуктів, послуг та ініціатив.

Аналіз задоволеності споживачів та їх настроїв. Вся зібрана інформація може показати, як клієнти ставляться до компанії чи бренду, як можна зберегти їх лояльність до бренду та як покращити зусилля щодо обслуговування клієнтів.

Big Data та блокчейн – прорив в області аналізу даних

Постійне прискорення зростання обсягу даних є невід'ємним елементом сучасних реалій. Соціальні мережі, мобільні пристрої, дані з вимірювальних пристроїв, бізнес-інформація – це лише кілька видів джерел, здатних генерувати гігантські масиви даних.

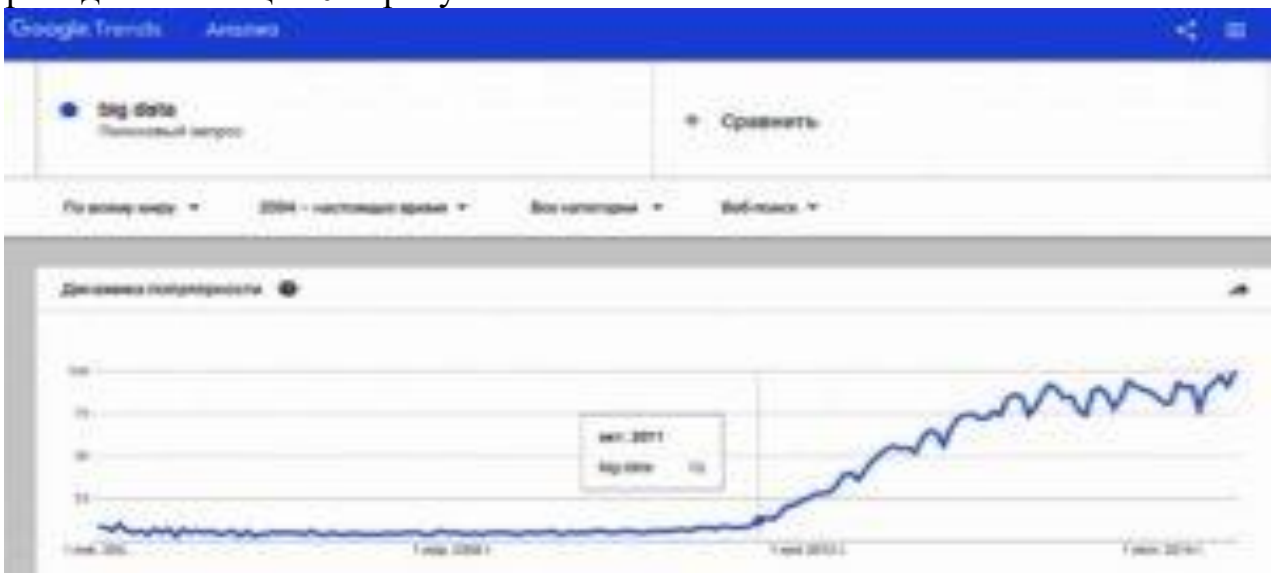
В даний час термін Big Data (Великі дані) став досить поширеним. Далеко

не всі ще усвідомлюють те, наскільки швидко й глибоко технології обробки великих масивів даних змінюють найрізноманітніші аспекти життя суспільства. Зміни відбуваються в різних сферах, породжуючи нові проблеми і виклики, в тому числі і в сфері інформаційної безпеки, де на першому плані повинні знаходитися такі найважливіші її аспекти, як конфіденційність, цілісність, доступність і т. д.

На жаль, багато сучасних компанії вдаються до технології Big Data, не створюючи для цього належної інфраструктури, яка змогла б забезпечити надійне зберігання величезних масивів даних, які вони збирають і зберігають. З іншого боку, в даний час стрімко розвивається технологія блокчейн, яка покликана вирішити цю та багато інших проблем.

Що таке Big Data? По суті, визначення терміна лежить на поверхні: «великі дані» означають управління дуже великими обсягами даних, а також їх аналіз. Якщо дивитися ширше, то це інформація, яка не піддається обробці традиційними методами через її великих обсягів.

Сам термін Big Data (великі дані) з'явився відносно недавно. Згідно з даними сервісу Google Trends, активне зростання популярності терміна припадає на кінець 2011 року:



У 2010 році вже стали з'являтися перші продукти і рішення, безпосередньо пов'язані з обробкою великих даних. До 2011 року більшість найбільших IT-компаній, включаючи IBM, Oracle, Microsoft і Hewlett-Packard, активно використовують термін Big Data в своїх ділових стратегіях. Поступово аналітики ринку інформаційних технологій починають активні дослідження даної концепції.

В даний час цей термін набув значної популярності і активно використовується в самих різних сферах. Однак не можна з упевненістю сказати, що Big Data – це якесь принципово нове явище – навпаки, великі джерела даних існують вже багато років. У маркетингу ними можна назвати бази даних по покупкам клієнтів, кредитних історій, способу життя і т. д. На

протязі багатьох років аналітики використовували ці дані, щоб допомагати компаніям прогнозувати майбутні потреби клієнтів, оцінювати ризики, формувати споживчі переваги і т. д.

В даний час ситуація змінилася в двох аспектах:

– з'явилися більш складні інструменти і методи для аналізу і зіставлення різних наборів даних;

– інструменти аналізу доповнилися безліччю нових джерел даних, що обумовлено повсюдним переходом на цифрові технології, а також новими методами збору і вимірювання даних.

Дослідники прогнозують, що технології Big Data найактивніше будуть використовуватися у виробництві, охороні здоров'я, торгівлі, держуправлінні і в інших найрізноманітніших сферах і галузях.

Big Data – це не якийсь певний масив даних, а сукупність методів їх обробки. Визначальною характеристикою для великих даних є не тільки їх обсяг, але також і інші категорії, що характеризують трудомісткі процеси обробки і аналізу даних.

В якості вихідних даних для обробки можуть виступати, наприклад:

– логи поведінки інтернет-користувачів;

– Інтернет речей;

– соціальні медіа;

– метеорологічні дані;

– оцифровані книги найбільших бібліотек;

– GPS-сигнали з транспортних засобів;

– інформація про транзакції клієнтів банків;

– дані про місцезнаходження абонентів мобільних мереж;

– інформація про покупках в великих ритейл-мережах і т.д.

Згодом обсяги даних і кількість їх джерел безперервно зростає, а на цьому тлі з'являються нові і удосконалюються вже наявні методи обробки інформації.

Основні принципи Big Data:

– Горизонтальна масштабованість – масиви даних можуть бути величезними і це означає, що система обробки великих даних повинна динамічно розширюватися

при збільшенні їх обсягів.

– Отказоустойчивость – навіть при збої деяких елементів обладнання, вся система повинна залишатися працездатною.

– Локальність даних. У великих розподілених системах дані зазвичай розподіляються по значній кількості машин. Однак у міру можливості і в цілях економії ресурсів дані часто обробляються на тому ж сервері, що і зберігаються.

Для стабільної роботи всіх трьох принципів і, відповідно, високу ефективність зберігання і обробки великих даних необхідні нові проривні технології, такі як, наприклад, блокчейн.

Для чого потрібні великі дані?

Сфера застосування Big Data постійно розширюється:

– Великі дані можна використовувати в медицині. Так, встановлювати діагноз пацієнту можна не тільки спираючись на дані аналізу історії хвороби, але також беручи до уваги досвід інших лікарів, відомості про екологічну ситуацію району проживання хворого і багато інших чинників.

– Технології Big Data можуть використовуватися для організації руху безпілотного транспорту.

– Обробляючи великі масиви даних можна розпізнавати обличчя на фото та відеоматеріалах.

– Технології Big Data можуть бути використані ритейлерами – торговельні компанії можуть активно використовувати масиви даних із соціальних мереж для ефективного налаштування своїх рекламних кампаній, які можуть бути максимально орієнтовані під той чи інший споживчий сегмент.

– Дана технологія активно використовується при організації передвиборних кампаній, в тому числі для аналізу політичних уподобань в суспільстві.

– Використання технологій Big Data актуально для рішень класу гарантування доходів (RA), які включають в себе інструменти виявлення невідповідностей і поглибленого аналізу даних, що дозволяють своєчасно виявити ймовірні втрати, або спотворення інформації, здатні привести до зниження фінансових результатів.

– Телекомунікаційні провайдери можуть агрегувати великі дані, в тому числі про геолокації; в свою чергу ця інформація може становити комерційний інтерес для рекламних агентств, які можуть використовувати її для показу таргетированной і локальної реклами, а також для рітейлерів і банків.

– Великі дані можуть зіграти важливу роль при вирішенні відкриття торгової точки в певній локації на основі даних про наявність потужного цільового потоку людей.

Таким чином найбільш очевидне практичне застосування технології Big Data лежить в сфері маркетингу. Завдяки розвитку інтернету і поширенню всіляких комунікаційних пристроїв поведінкові дані (такі як число дзвінків, купівельні звички і покупки) стають доступними в режимі реального часу.

Технології великих даних можуть також ефективно використовуватися в фінансах, для соціологічних досліджень і в багатьох інших сферах. Експерти стверджують, що всі ці можливості використання великих даних є лише видимою частиною айсберга, оскільки в набагато більших обсягах ці технології використовуються в розвідці і контррозвідці, в військовій справі, а також у всьому тому, що прийнято називати інформаційними війнами.

У загальних рисах послідовність роботи з Big Data складається з збору даних, структурування отриманої інформації за допомогою звітів і дашборда, а також подальшого формулювання рекомендацій до дії.

Розглянемо коротко можливості використання технологій Big Data в маркетингу. Як відомо, для маркетолога інформація – головний інструмент для прогнозування і складання стратегії. Аналіз великих даних давно і успішно застосовується для визначення цільової аудиторії, інтересів, попиту і активності споживачів. Аналіз великих даних, зокрема, дозволяє виводити рекламу (на основі моделі RTB-аукціону – Real Time Bidding) тільки тим споживачам, які зацікавлені в товарі чи послузі.

Застосування Big Data в маркетингу дозволяє бізнесменам:

– краще дізнаватися своїх споживачів, залучати аналогічну аудиторію в Інтернеті;

– оцінювати ступінь задоволеності клієнтів;

– розуміти, чи відповідає пропонований сервіс очікуванням і потребам;

– знаходити і впроваджувати нові способи, що збільшують довіру клієнтів;

– створювати проекти, які користуються попитом і т. д.

Наприклад, сервіс Google.trends може вказати маркетологу прогноз сезонної активності попиту на конкретний продукт, коливання і географію кліків. Якщо зіставити ці відомості до статистичних даних, що збираються відповідним плагіном на власному сайті, то можна скласти план з розподілу рекламного бюджету із зазначенням місяця, регіону та інших параметрів.

На думку багатьох дослідників, саме в сегментації і використанні Big Data полягає успіх передвиборної кампанії Трампа. Команда майбутнього президента США змогла правильно розділити аудиторію, зрозуміти її бажання і показувати саме той меседж, який виборці хочуть бачити і чути. Так, на думку Ірини Белишевим з компанії Data-Centric Alliance, перемога Трампа багато в чому стала можливою завдяки нестандартному підходу до інтернет-маркетингу, в основу якого лягли Big Data, психолого-поведінковий аналіз і персоналізована реклама.

Політтехнологи та маркетологи Трампа використовували спеціально розроблену математичну модель, яка дозволила глибоко проаналізувати дані всіх виборців США систематизувати їх, зробивши надточний таргетинг не тільки за географічними ознаками, але також і по намірам, інтересам виборців, їх психотипу, поведінковими характеристиками і т. д. Після цього маркетологи організували персоналізовану комунікацію з кожною з груп громадян на основі їх потреб, настроїв, політичних поглядів, і навіть кольору шкіри, використовуючи практично для кожного окремого виборця свій меседж.

Що стосується Хілларі Клінтон, то вона в своїй кампанії використовувала «перевірені часом» методи, засновані на соціологічних даних і стандартному маркетингу, розділивши електорат лише на формально гомогенні групи (чоловіки, жінки, афроамериканці, латиноамериканці, бідні, багаті і т. д.) .

В результаті виграв той, хто гідно оцінив потенціал нових технологій і методів аналізу. Примітно, що витрати на передвиборну кампанію Хілларі

Клінтон були в два рази більше, ніж у її опонента:

Основні проблеми використання Big Data

Крім високої вартості, одним з головних чинників, що гальмують впровадження Big Data в різні сфери, є проблема вибору оброблюваних даних: тобто визначення того, які дані необхідно отримувати, зберігати і аналізувати, а які – не брати до уваги.

Ще одна проблема Big Data носить етичний характер. Іншими словами виникає закономірне питання: чи можна подібний збір даних (особливо без відома користувача) вважати порушенням меж приватного життя?

Не секрет, що інформація, що зберігається в пошукових системах Google і Яндекс, дозволяє IT-гігантам постійно допрацьовувати свої сервіси, робити їх зручними для користувачів і створювати нові інтерактивні додатки. Для цього пошуковики збирають призначені для користувача дані про активність користувачів в інтернеті, IP-адреси, дані про геолокації, інтересах і онлайн-покупках, особисті дані, поштові повідомлення і т. д. Все це дозволяє демонструвати контекстну рекламу відповідно до поведінкою користувача в інтернеті. При цьому зазвичай згоди користувачів на це не питається, а можливості вибору, які відомості про себе надавати, не дається. Тобто за замовчуванням в Big Data збирається все, що потім буде зберігатися на серверах даних сайтів.

З цього випливає наступна важлива проблема, що стосується забезпечення безпеки зберігання та використання даних. Наприклад, безпечна та чи інша аналітична платформа, якій споживачі в автоматичному режимі передають свої дані? Крім того, багато представників бізнесу відзначають дефіцит висококваліфікованих аналітиків і маркетологів, здатних ефективно оперувати великими обсягами даних і вирішувати з їх допомогою конкретні бізнес-завдання.

Незважаючи на всі складнощі з впровадженням Big Data, бізнес має намір збільшувати вкладення в цей напрямок. За даними дослідження Gartner, лідерами інвестують в Big Data галузей є медіа, ритейл, телеком, банківський сектор і сервісні компанії.

Перспективи взаємодії технологій блокчейн і Big Data. Інтеграція

технології розподіленого реєстру з Big Data несе в собі синергетичний ефект і відкриває бізнесу широкий спектр нових можливостей, в тому числі дозволяючи:

– отримувати доступ до детальної інформації про споживчі переваги, на основі яких можна вибудовувати докладні аналітичні профілі для конкретних постачальників, товарів і компонентів продукту;

– інтегрувати докладні дані про транзакції і статистикою споживання певних груп товарів різними категоріями користувачів;

– отримувати докладні аналітичні дані про ланцюги поставок і споживання, контролювати втрати продукції при транспортуванні (наприклад, втрати ваги внаслідок всихання і випаровування деяких видів товарів);

– протидіяти фальсифікаціям продукції, підвищити ефективність боротьби з відмиванням грошей і шахрайством і т. д.

Доступ до докладним даними про використання та споживанні товарів значною мірою розкриє потенціал технології Big Data для оптимізації ключових бізнес-процесів, знизить регуляторні ризики, розкриє нові можливості монетизації і створення продукції, яка буде максимально відповідати актуальним споживчими перевагами.

Як відомо, до технології блокчейн вже проявляють значний інтерес представники найбільших фінансових інститутів, включаючи Citibank, Nasdaq, Visa і т. д. На думку Олівера Буссманна, IT-менеджера швейцарського фінансового холдингу UBS, технологія блокчейн здатна «скоротити час обробки транзакцій від декількох днів до декількох хвилин ».

Потенціал аналізу фінансової інформації з блокчейна за допомогою технології Big Data величезний. Технологія розподіленого реєстру забезпечує цілісність інформації, а також надійне і прозоре зберігання всієї історії транзакцій. Big Data, в свою чергу, надає нові інструменти для ефективного аналізу, прогнозування, економічного моделювання і, відповідно, відкриває нові можливості для прийняття більш виважених управлінських рішень.

Тандем блокчейна і Big Data можна успішно використовувати в охороні здоров'я. Як відомо, недосконалі і неповні дані про здоров'я пацієнта в рази збільшують ризик постановки невірного діагнозу і неправильно призначеного лікування. Критично важливі дані про здоров'я клієнтів медустанов повинні бути максимально захищеними, мати властивості незмінності, бути перевіряються і не повинні бути піддані будь-яким маніпуляціям.

Інформація в блокчейне відповідає всім перерахованим вимогам і може служити в ролі якісних і надійних вихідних даних для глибокого аналізу за допомогою нових технологій Big Data. Крім цього, за допомогою блокчейна медичні установи змогли б обмінюватися достовірними даними зі страховими компаніями, органами правосуддя, роботодавцями, науковими установами та іншими організаціями, такими, що потребують медичної інформації.

Big Data та інформаційна безпека. У широкому розумінні, інформаційна

безпека є захищеність інформації і підтримуючої інфраструктури від випадкових або навмисних негативних впливів природного або штучного характеру.

В області інформаційної безпеки Big Data має справу з такими викликами:

– проблеми захисту даних і забезпечення їх цілісності;

– ризик стороннього втручання і витоку конфіденційної інформації;

– неналежне зберігання конфіденційної інформації;

– ризик втрати інформації, наприклад, внаслідок чиїхось зловмисних дій;

– ризик нецільового використання персональних даних третіми особами і т. і.

Одна з головних проблем великих даних, яку покликаний вирішити блокчейн, лежить у сфері інформаційної безпеки. Забезпечуючи дотримання всіх основних її принципів, технологія розподіленого реєстру може гарантувати цілісність і достовірність даних, а завдяки відсутності єдиної точки відмови, блокчейн робить стабільною роботу інформаційних систем. Технологія розподіленого реєстру може допомогти вирішити проблему довіри до даних, а також надати можливість універсального обміну ними.

Інформація – цінний актив, а це значить, що на першому плані має стояти питання забезпечення основних аспектів інформаційної безпеки. Для того, щоб вистояти в конкурентній боротьбі, компанії повинні йти в ногу з часом, а це значить, що їм не можна ігнорувати ті потенційні можливості і переваги, які містять в собі технологія блокчейн і інструменти Big Data.

Великі дані – джерело інновацій. Це аж ніяк не новина, але від цього великі дані не стають менш значущими. Саме вони допомагають компаніям рухатись в бік діджитал-трансформації. Бізнес та технічні лідери використовують великі дані, щоб скористатися низкою переваг: від вдосконалення користувацького досвіду до нових потоків прибутку через оцінку ефективності цілих організації.

Розглянемо декілька реальних кейсів використання Big Data у телекомунікаційній, фінансовій, медичній та інших індустріях.

Рис.1.1. Сфери застосування Big Data

Телеком. Зі всіх індустрій, що одержать користь від аналізу великих даних, телеком має найкращу позицію завдяки величезним обсягам даних, які телекомунікаційні компанії через мережі операторів. Самі лише мобільні оператори володіють інформацією про профілі користувачів, їхні девайси, геолокацію, зразки поведінки тощо. Великі телеком-компанії, як от AT&T, CenturyLink, Swisscom, T-Mobile, та Vodafone, вже впровадили аналітику великих даних у розробку свого програмного забезпечення. Завдяки цьому вони зможуть краще передбачати попит, планувати навантаження на свої мережі та глибше розуміти свій ринок. Найголовніше — вони зможуть покращити користувацький досвід, що є ключовим аспектом боротьби за лідерство у будь-якій бізнес сфері.

Приклади використання великих даних у сфері телеком. Наразі одним з пріоритетних напрямків використання великих даних поміж телеком-компаніями є моніторинг стану мережі та обладнання. Наприклад, AT&T за годину збирає понад 30 мільярдів точкових даних, щоб оцінити якість роботи мережі та передбачити можливі збої у роботі обладнання. Таким чином, компанія заощаджує сотні тисяч доларів та домагається безперебійного сервісу для своїх клієнтів.

Сільське господарство. Зі світовим населенням понад 7 мільярдів людей, змінами клімату та виснаженням фермерських земель, сучасне сільське господарство стикається із серйозними проблемами. Задля подолання цих викликів, індустрія залучає інноваційні технічні розробки, як от Інтернет речей, хмарні технології, великі дані та аналітику. Використовуючи розумні сенсори та зв’язані пристрої, ми створюємо нове покоління "розумних" ферм, заснованих на використанні великих даних.

Сучасні компанії спираються на великі дані, щоб:

- Аналізувати типи ґрунту та його родючості
- Оптимізувати використання ресурсів
- Збільшувати врожайність сільськогосподарських культур
- Прогнозувати погодні умови
- Керувати каналами збуту
- Приклади використання великих даних у сільському господарстві

Щоб максимізувати врожайність, фермери повинні враховувати безліч факторів, включно з погодою, якістю ґрунту, рівнем вологості та поживних речовин, частотою та дозування добрив та пестицидів тощо.

John Deere, один зі світових лідерів у сфері сільського господарства, створив цілу екосистему, яка поєднує обладнання, що оснащене датчиками та хмарним порталом. Ця система відстежує активність у режимі реального часу, аналізує продуктивність та приймає рішень щодо того, що, де та коли саджати.

Рис. 1.2. Використання Big Data в сільському господарстві

Фінанси. Сфера застосування аналітики даних у фінансових та банківських справах величезна. Починаючи від внутрішніх структурованих даних (торговельні системи, дані з ринків та фондових бірж) та закінчуючи неструктурованими даними (соціальні медіа, відгуки користувачів), фінансові інституції знають, як використовувати інформацію задля свого успіху.

- Поглиблена сегментація користувачів. За допомогою таких даних, як демографічні відомості, моделі поведінки, дані про девайси тощо, фінансові компанії створюють точніші портрети своїх споживачів.
- Дані про фондові ринки у режимі реального часу. Алгоритми машинного навчання аналізують ціни на акції, а також соціальні та політичні тренди, які можуть потенційно вплинути на фондовий ринок.

- Безпека та запобігання шахрайству. Аналіз великих даних у режимі реального часу дозволяє фінансовим компаніям відстежувати будь-яку підозрілу активність та попереджати ненадійні транзакції.
- Точний аналіз ризиків. Беручи до уваги традиційні та нетрадиційні джерела даних, алгоритми машинного навчання краще визначають потенційні ризики з кредитуванням.

Приклади використання великих даних у фінансовій сфері

Оцінювання ризиків кредитування є одним з найголовніших напрямків діяльності фінансових установ у контексті великих даних. Наприклад, саме на цьому фокусується компанія Kreditech, що надає онлайн кредити. На додачу до стандартних відомостей про клієнтів, компанія використовує дані з їхніх постів у соціальних мережах, геолокаційну інформацію, дані про покупки в інтернеті тощо. Потім програма на основі штучного інтелекту обробляє ці дані та визначає, чи існують потенційні ризики надання тому чи іншому клієнтові кредиту – і все це за лічені хвилини.

Роздрібна торгівля. Згідно з прогнозом Global Big Data Analytics, у 2026 році ринок роздрібної торгівлі сягне 14 мільярдів доларів, зростаючи на 23,4%. Це означає, що втримувати увагу покупців серед різноманіття товарів стане дедалі складніше. Щоб успішно працювати у надзвичайно конкурентній індустрії, продавці товарів використовують великі дані, які допомагають їм краще зрозуміти поведінку споживачів та стати справді клієнтоорієнтованими.

Приклади використання великих даних у роздрібній торгівлі. Ритейл-гігант Amazon точно знає, що таке великі дані та як ними користуватися. Компанія зберігає понад 1,000,000,000 GB даних на своїх серверах. Ця інформація використовується у багатьох бізнес-процесах, наприклад, для надання покупцям релевантних рекомендацій. Amazon відстежує, на які товари покупці дивляться та які врешті купують, і надсилає їм персоналізовані рекомендації щодо майбутніх покупок. Таким чином, близько 35% прибутку компанії складається саме з таких замовлень на основі рекомендацій.

Однак, великі дані стають у пригоді не тільки світовим гігантам, але й невеликим бізнесам теж. Так, наприклад, м’ясна крамниця Pendleton & Son у Лондоні почала суттєво програвати новому супермаркету, що відкрився на тій самій вулиці. Тоді власники крамниці вирішили встановити сенсори руху, щоб визначити, наскільки їхні вітрини приваблюють покупців. Аналіз даних допоміг їм не тільки обрати оптимальний зовнішній вигляд вітрин, а ще й надав цінний інсайт. Так, власники зрозуміли, що кількість потенційних покупців біля їхньої крамниці збільшується ввечері, отже почали працювати довше та пропонувати вуличну їжу перехожим, що поверталися додому з пабів. Таким чином, Pendleton & Son змогли значно збільшити свій прибуток та витримати конкуренцію.

Медицина. Клінічні дослідження, цифрові медичні карти, телемедицина та інші MedTech рішення – це маркери справжньої технічної революції у сфері

охорони здоров'я. Завдяки великим даним, медичні аналітики досягають не просто видатних, а рятівних результатів. Опираючись на медичні дані, лікарі можуть точніше діагностувати та прогнозувати перебіг хвороби, що покращує якість життя пацієнтів та заощаджує їхні витрати.

Приклади використання великих даних у медицині. Компанія Аріхіо, один з провідних постачальників послуг у медичній аналітиці, використовує машинне навчання, щоб перевести прийняття медичних рішень на наступний рівень разом з операційною ефективністю. Аналізуючи медичні записи пацієнтів, компанія допомагає лікарям отримати деталі історії хвороби та стану здоров'я пацієнта загалом. У 2018 році Аріхіо проаналізували понад 4.5 мільйона медичних записів, зменшуючи затрати часу та зусиль медичних працівників на 80%.

Рис. 1.3. Використання Big Data в медицині

Медіа та розваги. На сьогодні більше половини населення планети користується соціальними мережами. Для того, щоб витримувати конкуренцію, медіакомпанії мають надавати своїм користувачам першокласний контент та безперебійний досвід користування за допомогою різних каналів комунікації. Саме тут у пригоді стають великі дані. Завдяки зібраній інформації, компанії отримують дані про популярність контенту, взаємодію користувачів, активність у соціальних мережах, підписки, реакції на маркетингові кампанії тощо. Проаналізувавши дані, компанії можуть:

- Передбачати поведінку користувачів
- Створювати персоналізований контент
- Вдосконалювати досвід користування платформами
- Запроваджувати ефективніші рекламні кампанії
- Приклади використання великих даних у сфері медіа та розваг

Навряд чи є кращий приклад аналітики великих даних у медіа, аніж історія компанії Netflix. Стрімінговий гігант використовує великі дані, щоб задовольняти потреби понад 195 мільйонів підписників. За допомогою

машинного навчання, компанія аналізує вподобання своїх глядачів та пропонує їм відповідний контент – 75% відсотків переглядів на Netflix забезпечують персоналізовані рекомендації від платформи.

На завершення. Великі дані надають компаніям з будь-якої індустрії можливість отримати практичну інформацію та дізнатись приховані закономірності. Аналізуючи отриману інформацію, компанії мають змогу не тільки зміцнити свої позиції на ринку, але й запропонувати своїм користувачам клієнтоорієнтований та приємний сервіс.

2. Історія виникнення терміну Big Data

Уперше термін Big Data з'явився у 2008 році. Його ввів редактор журналу Nature Кліффорд Лінч. Це поняття використовували в спецвипуску видання, яке присвятили активному росту обсягів інформації у всьому світі.

Незважаючи на настільки недавню появу терміна, великі дані існували й раніше, однак вони не мали великої цінності, оскільки для вивчення інформації потрібні істотні обчислювальні потужності, велика кількість часу і високі фінансові витрати. З появою технологій для обробки багатогігабітних даних (платформи Hadoop) ситуація змінилася, і Big Data знайшли застосування в різних сферах.

В 2011 рокці поняття Big Data почало набирати популярність, в основному, у великих корпораціях таких як Microsoft, IBM, Oracle, EMC, HP та інших.

В 2011 році компанія Gartner відмітила великі дані як тренд номер два в інформаційно-технологічній інфраструктурі після віртуалізації. За прогнозами мається на увазі, що впровадження технологій Big Data суттєво вплине на інформаційні технології в сферах виробництва, охороні здоров'я, торгівлі, державного управлінні, а також в галузях, в яких реєструються індивідуальні переміщення ресурсів. З 2013 року Big Data починають викладати в університетах в рамках вузівських програм з науки про дані і інженерії.

Інноваційні розробки в області Big Data починалися не в маленьких стартапах, як це часто буває в IT-індустрії, а в великих компаніях. Так, наприклад, технологія розподіленої обробки даних MapReduce була розроблена компанією Google, а Hadoop, що є вільним програмним забезпеченням для виконання розподілених обчислень на кластерах з сотень і тисяч вузлів, відразу після створення активно підтримала компанія Yahoo. Більшість програмних продуктів в області Big Data є вільними, а їх адаптацією і просуванням займаються ті самі стартапи. Традиційні постачальники рішень в області зберігання і обробки даних, такі як IBM уважно ставляться до нових розробок в області Великих Даних і намагаються використовувати їх в своїх продуктах разом зі своїми технологіями.

3. Характеристики Big Data

Згідно з компанією Meta Group, у Big Data є три ключові характеристики – так звані 3V: Volume, Velocity і Variety (рис. 1.4).

- Volume – великий обсяг даних;
- Velocity – регулярне оновлення даних і постійна їхня обробка;
- Variety – можливість одночасної обробки різних типів інформації: тексту, зображень, відео тощо.

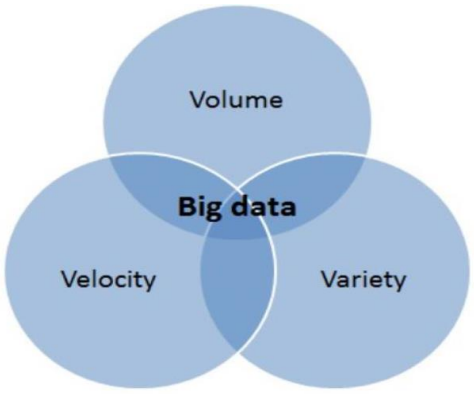


Рис.1.4. Характеристики Big Data [5]

Оскільки технологія Big Data постійно розвивалася, з часом у компанії IBM запропонували доповнити цей список четвертим V – veracity (правдивістю). У IDC до переліку додали viability (життєздатність) і value (цінність).

Існує безліч характеристик для великих даних, але спробуємо розглянути основні.

Сфера великих характеризується такими ознаками:

Volume (об'єм): накопичена база даних є гігантський обсяг інформації, для якого обробка і зберігання традиційними способами є трудомісткими процесами. Такий обсяг потребує нових підходів і в більш вдосконалених інструментах.

Velocity (швидкість): цей показник вказує як на зростаючу швидкість накопичення, так і на швидкість обробки даних.

У багатьох випадках набори великих даних оновлюються в режимі майже реального часу, замість щоденних, щотижневих або щомісячних оновлень, характерних багатьом традиційним сховищам даних.

Програми аналітики великих даних співвідносять та аналізують вхідні дані, а потім надають відповідь або результат на основі запиту. Це означає, що аналітики даних повинні детально розуміти наявні дані та мати певне розуміння того, які відповіді вони шукають, щоб переконатися, що отримана інформація є дійсною та актуальною.

Управління швидкістю передачі даних також має важливе значення, оскільки аналіз великих даних поширюється на такі сфери, як машинне навчання та штучний інтелект, де аналітичні процеси автоматично знаходять

закономірності у зібраних даних та використовують їх для отримання знань.

Останнім часом стали більш затребувані технології обробки даних в реальному часі.

Variety (різноманіття): дана характеристика означає можливість одночасної обробки структурованої і неструктурованої інформації різних форматів. Головною відмінністю структурованої інформації є можливість класифікації. Прикладом такої інформації може служити інформація про клієнтських транзакцій.

Veracity (достовірність даних): в даний час достовірність наявних даних є найважливішим критерієм для користувачів. Недостовірна інформація призводить до утруднення аналізу даних.

Достовірність даних стосується ступеня визначеності в наборах даних.

Невизначені необроблені дані, зібрані з різних джерел, таких як платформи соціальних медіа та веб-сторінки, можуть спричинити серйозні проблеми з якістю даних.

Наприклад, компанія, яка збирає масиви великих даних із сотень джерел, може виявити неточні дані, але аналітикам потрібна інформація про шляхи надходження даних, щоб простежити, де дані зберігаються, щоб вони могли виправити проблеми.

Неякісні дані призводять до неточного аналізу та можуть підірвати цінність бізнес-аналітики, оскільки це може призвести до недовіри керівників до даних у цілому.

Кількість невизначених даних в організації повинна бути врахована перед тим, як їх використовувати для аналізу великих даних. Командам IT та аналітики також потрібно забезпечити наявність достатньо точних даних для отримання достовірних результатів.

Value (цінність накопиченої інформації): великі дані повинні бути корисні в удосконаленні бізнес-процесів, складанні звітності або оптимізації витрат компаній.

Дуже важливо, щоб організації застосовували такі практики, як очищення даних, і існував механізм підтвердження, що дані стосуються відповідних питань бізнесу, перш ніж використовувати їх у проекті аналізу великих даних.

Перші три характеристики визначають так званий принцип «Трьох V».

Вирішальну роль у великих даних відіграють обсяг інформації, швидкість обробки, а також різноманітність з'являються даних.

Обсяг відноситься до наборів даних, розмір яких виходить за межі можливостей програмних засобів типової бази даних збору, зберігання, обробки і аналізу даних.

Різноманітність визначає здатність обробки безлічі типів, джерел і форматів даних від сенсорів, розумних пристроїв, соціальних мереж. Також різноманітність характеризується здатністю інтегрувати все більше число джерел, що містять різні структуровані, напівструктуровані дані, вилучаються з

web-сторінок, web log файлів, e-mail, документів та ін.

Швидкість визначає реакцію на поточну інформацію за час, обмежене додатком. Прикладом є потокова обробка (наприклад, GPS даних в реальному часі).

4. Основні типи Big Data

Існує два типи даних – традиційні дані та великі дані.

Традиційні дані зберігаються в базах даних, які містять структуровані таблиці з текстовою, цифровою та іншою інформацією. Один комп'ютер може з легкістю управляти таким видом даних.

Традиційні дані можуть надходити з різних джерел. Як правило, це бувають дані про користувачів і клієнтів, наприклад, інформація про слухачів курсів з Data Science: повне ім'я, адреса, контактна інформація, кількість відвідувань або звернень до сервісного центру та ін.

У свою чергу, *великі дані* набагато перевершують в кількості традиційні дані. Такий тип даних розподіляється між комп'ютерами, але big data дуже важко використовувати ефективно. Ми отримуємо великі дані з абсолютно різних джерел – соціальних мереж (Facebook, Twitter, LinkedIn, Quora і т.д), фінансів, мобільних телефонів, курсів та інших ресурсів.

Big Data також охоплюють широкий спектр типів даних, включаючи наступні:

- структуровані дані в базах даних та сховищах даних на основі мови структурованих запитів (SQL);
- неструктуровані дані, такі як текстові та файли документів, що зберігаються в кластерах Hadoop або системах баз даних NoSQL
- напівструктуровані дані, такі як журнали веб-сервера або потокові дані з датчиків.

Всі різні типи даних можна зберігати разом за допомогою технологій які, як правило, базуються на Hadoop або службі зберігання хмарних об'єктів.

Крім того, програми для Big Data часто містять кілька джерел даних, які в іншому випадку не можуть бути інтегровані.

До ключових джерел великих даних належать:

- інформація з Інтернету: соціальних мереж, блогів, ЗМІ, форумів, сайтів;
- показання різних пристроїв: IoT-датчиків, аудіо- та відеореєстраторів, розумних гаджетів, смартфонів, стільникового зв'язку тощо;
- корпоративні відомості: архіви, внутрішні відомості підприємств і організацій та ін.

Завдяки аналітиці великих даних (Big Data Analytics) можна швидко і якісно інтерпретувати різну інформацію, знаходити закономірності і складати прогнози. Наприклад, за допомогою Big Data визначають, у якій частині міста існує потреба в певних товарах чи послугах, яка продукція зацікавить

потенційних покупців, передбачають сплески захворювань і навіть місця, де найімовірніше відбудуться злочини. Чим більше відомостей вдасться вивчити, тим точнішим буде кінцевий результат.

Наприклад, метеорологи беруть дані про погоду за останні 100 років і аналізують їх. У результаті вони виявляють закономірності, у який період року/місяця настає потепління, похолодання чи починається сезон опадів. На основі цих відомостей вони можуть спрогнозувати погоду на найближчий період.

Тема 2. Візуалізація великих даних

1. Поняття візуалізації. Мова візуалізації
2. Сфера застосування візуалізації та завдання які вона виконує
3. Види візуалізації
4. Коротка характеристика інструментів для візуалізації даних

1. Поняття візуалізації. Мова візуалізації

Візуалізація даних допомагає сприймати та запам'ятовувати інформацію.

Наш мозок влаштований таким чином, що візуальні образи він сприймає набагато краще, ніж текстовий, цифровий або табличний контент. Тому, часто ми можемо не помічати важливу інформацію у масивних об'ємах тексту. Візуалізація покликана донести до користувача те, що він зазвичай не бачить. Веб-дизайнери та контент-мейкери можуть влучно використовувати цю природну особливість людини, щоб передавати їй велику кількість даних. А добре продумані візуалізації, особливо персоналізовані, можуть не тільки донести інформацію, а ще й закарбуватися в пам'яті. Це спричинено тим, що користувач реагує на дизайн візуалізації так само, як і на самий контент. Якщо загальне оформлення або певні елементи звертаються до його досвіду, особливих якостей, переживань тощо, то реакція користувача на них і їхній візуальний вплив будуть сильнішими. В його пам'яті залишиться певний досвід.

Рис.2.1. - це демонстрація особливостей благодійної діяльності зірок залежно від статі. Візуалізована аналітика подається в дуже простій формі (3 секції: які сфери благодійності більше підтримують жінки, чоловіки та які сфери — однаково). Така подача даних допоможе користувачеві запам'ятати деяку інформацію. Якби цю інформацію не було візуалізовано, її б важче було аналізувати, порівнювати дані та фіксувати для себе цікаві моменти.

Візуалізація даних не тільки допомагає опрацьовувати масиви інформації, а ще й спроможна переконати користувача. Коли дизайнер оформлює інформацію у візуальному форматі, він працює з абстрактними даними і робить їх реальними, надає їм форми та об'ємності загальній картині.

Такою “живою” інформацією легше впливати на користувача, бо вона створює образи в його свідомості, апелює до почуттів й починає говорити з ним на емоційному рівні. Тому, довіра до неї значно більша, ніж до тексту. Цікавий факт: якщо в матеріал додати графік, діаграму тощо, то контент миттєво стане більш переконуючим. Цим успішно можна користуватися у вебі, спонукаючи користувачів на певні думки, переконання та змінюючи їхні настрої.

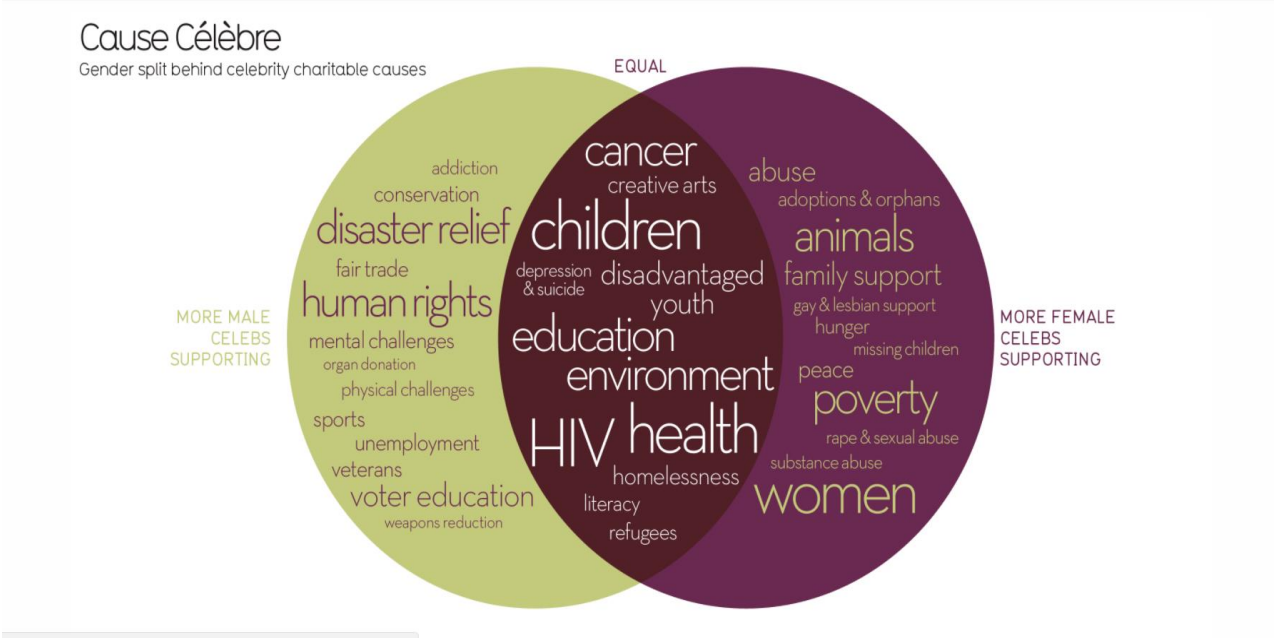


Рис. 2.1. Благодійна діяльність зірок залежно від статі https://informationisbeautiful.net/visualizations/cause-celebre-gender-split-behind-celebrity-charitable-giving-causes/

Рис. 2.2 - візуальне пояснення вартості розриву відносин між Великобританією та Євросоюзом у порівнянні з реаліями життя Королівства. Дані на контрасті показують чого вартий Brexit і формують відповідну думку в користувача.

Візуалізація даних допомагає зацікавити. В інтернеті в користувача мало часу та сил не цілеспрямовано читати аналітичні матеріали. Тому сучасні ЗМІ почали активно залучати візуалізацію даних. Її використовують, щоб цікаво подавати (доповнювати) великі матеріали, оскільки візуалізація здатна перетворити складне, в щось просте для розуміння. Найкраще для цього підходять інфографіки, бо вони живо та лаконічно можуть описати цілу проблему або продовжену в часі подію. Інфографіки здатні цікаво розповісти користувачу історію, щоб той особисто не вивчав проблему, читаючи аналітичні статті. Тому, коли людина бачить інфографіку, в неї з’являється зацікавленість в матеріалі, бо роздивитися картинку і “послухати” історію значно зручніше та швидше.

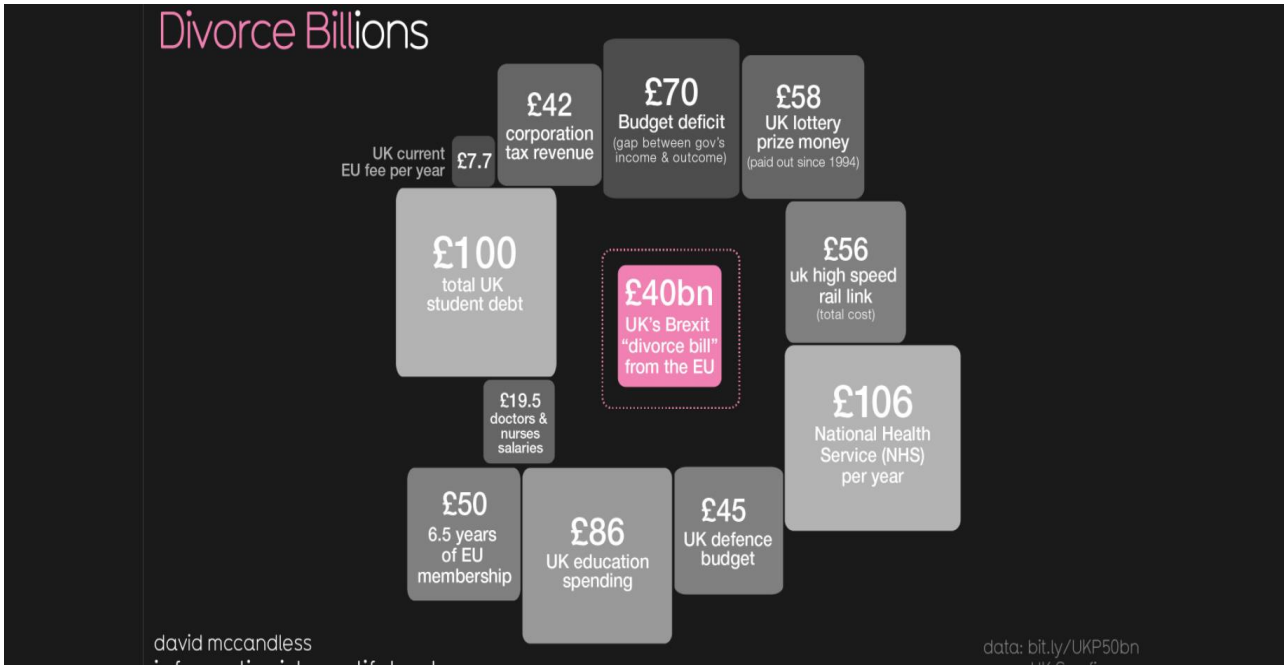


Рис. 2.2. Візуальне пояснення вартості розриву відносин між Великобританією та Євросоюзом https://informationisbeautiful.net/visualizations/uks-brexit-divorce-bill-ions/

Через це інтернет-медіа використовують одну техніку, якою приваблюють користувачів натиснути на матеріал та переглянути його. Вони попередньо вказують у заголовках, що в матеріалі присутня візуалізація даних, пишучи поруч з назвою слово “Інфографіка”.

*Візуалізація корисно розважає користувача***.** Інтерактивність — це ще один плюс візуалізованих даних. Завдяки їй можна привабити аудиторію на свій ресурс і залишити там надовго, пропонуючи цікаві взаємодії з контентом. Інтерактивні інфографіки, діаграми тощо не тільки повідомляють користувачу масив інформації, а й пропонують цікаво її опрацювати, погратися з нею. В режимі гри юзер легше сприймає та краще усвідомлює дані, бо він має безпосередню взаємодію з ними. Користувач відчуває себе важливим, бо від нього залежить, що буде далі на екрані, він стає частиною графіку і може керувати ним. Це його безумовно приваблює.

Рис. 2.3 - Великий масив даних, який обертається навколо юзера. Йому пропонують ввести дату свого народження, вираховують його вік і пропонують подивитися, що відомі люди зробили, коли їм було стільки ж років. Це приваблює аудиторію, бо людям завжди цікаво, коли справа стосується їх особистості. Цікаво порівняти себе з іншими і паралельно дізнатися багато історичних фактів: наприклад, що в його віці Чингісхан вже побудував свою першу армію, а Білл Гейтс заснував Microsoft. В такому розважальному форматі інформацію дуже просто та ефективно подавати користувачу, якщо, звичайно, вона естетично оформлена.

http://www.informationisbeautiful.net/visualizations/who-old-are-you/

Рис. 2.3.

2. Сфера застосування візуалізації та завдання які вона виконує

Розглянемо галузі використання візуалізації.

Статистика та звіти. Дані за якийсь період часу показуються разом. Наприклад, статичної картинкою в додатку до звіту або налаштованим графіком в сервісі статистики, з можливістю зміни параметрів його відображення.

Довідкова інформація. Доповнення до основного тексту, наочно ілюструє його згаданими даними. Наприклад, дати загальне уявлення про динаміку одного з показників, або відобразити якийсь процес і його етапи; може бути - показати структуру якогось явища.

Інтерактивні сервіси. Продукти харчування й проекти, в яких інфографіка є частиною функціональності. Так, в якості засобу навігації по сервісах може бути діаграма процесу. Майже все, що пов'язано з роботою з картами в спеціалізованих системах на кшталт диспетчерських і більшої частини комп'ютерних ігор.

Ілюстрації. Красиве відображення даних для створення самостійних ілюстрацій.

Креслення і схеми. Спеціалізовані документи, що показують структуру і процес роботи складних інженерних та природних систем.

Експерименти і мистецтво. Візуалізація даних у вигляді складних і громіздких зображень, які складно «прочитати» побіжно - обсяг даних і взаємозв'язків між ними такий, що потрібно розбиратися з картинкою по

частинах; або просто абстрактні зображення, автоматично згенерували. Останнім часом напрямок все більш популярно і періодично виходить за рамки комп'ютерної графіки - наприклад, у вигляді графіків-скульптур.

Виділимо умовно три типи візуалізації.

– Наукова візуалізація. При моделюванні різних об'єктів або процесів з'являються великі обсяги даних.

– Інформаційна візуалізація. Опис / уявлення якоїсь абстрактної інформації, отриманої при зборі та обробці багаторівневих даних, для аналізу яких необхідно застосовувати різні кількісні та якісні заходи оцінки.

– Візуалізація роботи програмного забезпечення.

Візуалізація BigData має певні завдання:

– візуалізація потоків даних;

– візуальний інтелектуальний аналіз даних;

– візуальний пошук і рекомендації;

– опис ситуацій на основі великих даних з використанням візуалізації;

– масштабовані методи паралельної візуалізації;

– сучасні апаратні засоби і архітектури для аналізу і візуалізації даних;

– людино-комп'ютерний інтерфейс і візуалізація великих даних;

– додатки візуалізації великих даних.

Можна сформулювати вимоги до такого роду візуалізації:

– оцінка придатності (адекватності в візуалізації) видів відображення,

– природність (звичність для користувачів),

– стійкість до масштабування,

– можливість виведення надвеликих обсягів даних,

– можливості для представлення складних структур, а також об'єктів особливого інтересу, особливих точок, аттракторов, сингулярностей.

3. Види візуалізації

Розглянемо традиційні види візуалізації.

– Графікі і діаграми

– Інфографіка і схеми

– Презентація і аналіз даних

– Інтерактивний сторітеллінг

– Бізнес аналітика і дашборда

– Наукова і медична візуалізація

– Карти і картограми

Графіки і діаграми. Напевно, самий звичний вид візуалізації даних. Використовується як для презентації даних, так і для аналізу. Зустріти їх можна і на роботі, і в журналі, і в науковому звіті. Зазвичай знання про існуючі типи діаграм і графіків ми отримуємо зі школи або з стандартного набору в Excel. Однак світ графіків і діаграм не обмежується точковим графіком, стовпчиковою

і круговою діаграмою. Існують близько 15 загальновідомих типів діаграм, а всього їх понад 60, при цьому їх кількість збільшується з кожним днем люди придумують нові типи для візуалізації складних і незвичайних даних.

Інфографіка стала дуже популярна в останні роки, хоча існують вже давно. Інфографіка відноситься до журналістики даних, де графіки і схеми пояснюють будь-які факти з обраної теми. Зазвичай інфографіка статична і являє собою довге «простирадло» з картинками і текстом. Відмінною особливістю інфографіки є те, що в ній наводяться вже готові висновки, тобто читача проводять за руку з обраної теми і при цьому приправляють це все цифрами і картинками. Часто використовується мальований або анімаційний стиль. Часто використовується не до місця або «для краси», хоча звичайно ж є чудові і цікаві приклади.

Презентація та аналіз даних. Один самих звичних способів використання візуалізації даних - презентація інформації у вигляді діаграм або інфографіки. І якщо з цим все зрозуміло, то використання візуалізації для аналізу інформації в основному використовується тільки бізнес-аналітиками та вченими. У чому полягає відмінність?

При аналізі даних за допомогою візуалізації використовують створення великої кількості різних візуальних уявлень одних і тих же даних. Робиться це для можливості знаходження прихованих, на перший погляд, взаємозв'язків і залежностей, а також первинної оцінки набору даних для можливості застосування в подальшому більш складних інструментів аналізу. Цей підхід називається Exploratory data analysis (EDA), що можна перекласти як розвідувальний аналіз даних. Основна відмінність від презентації даних - візуалізація тут може бути «чорновий», але виконується швидко і однією людиною або невеликою робочою групою.

Інтерактивний сторітеллінг. Сторітеллінг - це підношення будь-якої корисної інформації в формі цікавої розповіді. Інтерактивний сторітеллінг - розповідь, з яким слухач може взаємодіяти. Користувач може управляти відображенням інформації і знаходити ті залежності, які не знайшов автор. У цьому сенсі він близький до розвідувального аналізу даних, але відрізняється тим, що дані заздалегідь оброблені і представлені в зручному для аналізу вигляді, а також є підказки або заздалегідь прописані сценарії використання.

Тому, найчастіше інтерактивний сторітеллінг називають інтерактивної інфографікою, але для того щоб їй стати недостатньо просто до статичної інфографіку додати спливаючі віконця.

Дашборди і бізнес аналітика. Візуалізація активно використовується в бізнесі. Принцип «говорите з даними» допомагає компаніям заробляти більше, а клієнтам отримувати кращий сервіс. Для разового аналізу зазвичай використовується Excel або R. Однак це незручно якщо необхідно стежить за якимись показниками на постійній основі. Для відстеження використовують дашборди - дисплеї, на яких виведені всі необхідні показники в одному місці в

вигляді графіків, діаграм і таблиць. Проектування ефективних дашбордів - складна і неординарна завдання. Найчастіше їх перевантажують непотрібною інформацією або намагаються використовувати всі можливі типи шаблонних графіків. Часто для того, щоб спроектувати хороший дашборд, необхідне створення нових типів візуалізації інформації. Тематика активно розвивається за рахунок все більшого застосування аналітики в бізнесі. Також дашборди застосовуються і для особистого використання (фітнес трекери, аналіз особистих витрат і т. п.)

Візуалізація в медицині та науці. Специфічний вид візуалізації Його метою зазвичай є виділення закономірностей або аномалій. Від звичайної візуалізації даних відрізняється тим, що часто буває тривимірної і вимагає спеціальної підготовки для інтерпретації.

Карти і картограми. Карти - одні з найдавніших способів візуалізації, що відображають навколишню реальність. Картограма - карта з нанесеною на неї інформацією у вигляді кольору або інших способів. Картограми можуть бути використані для відображення будь-якої інформації - від щільності населення, до частоти використання мобільних телефонів в кожному районі країни.

Хмара тегів. Кожному елементу в хмарі тегів присвоюється певний ваговий коефіцієнт, який корелює з розміром шрифту. У разі аналізу тексту величина вагового коефіцієнта безпосередньо залежить від частоти вживання (цитування) певного слова або словосполучення.

Дозволяє читачеві в стислі терміни отримати уявлення про ключові моменти скільки завгодно великого тексту або набору текстів.

Кластерграмма. Метод візуалізації, що використовується при кластерному аналізі. Показує, як окремі елементи безлічі даних співвідносяться з кластерами в міру зміни їх кількості. Вибір оптимальної кількості кластерів - важлива складова кластерного аналізу.

Історичний потік. Допомагає стежити за еволюцією документа, над створенням якого працює одночасно велику кількість авторів. Зокрема, це типова ситуація для сервісів wiki в тому числі. По горизонтальній осі відкладається час. За вертикальної - внесок кожного з співавторів, тобто обсяг введеного тексту. Кожному унікальному автору присвоюється певний колір на діаграмі

Просторовий потік. Використовується для відстеження просторового зміни інформації.

4. Коротка характеристика інструментів для візуалізації даних

Кожен день ми тонемо у величезній кількості найрізноманітнішої інформації: від етикеток на продуктах до звітів Всесвітньої організації охорони здоров'я. І подавати інформацію так, щоб вона виділялася серед іншої, стає все складніше і складніше.

Якщо ви шукаєте спосіб просто і зрозуміло розповісти про складні дані,

географію, пояснити неочевидні взаємозв'язки, складні або прості ідеї, то вам потрібна візуалізація. Вона зручна тим, що відразу привертає увагу до ключового послання, демонструє закономірності, які важко вловити в тексті або в таблиці з цифрами.

Існує багато спеціальних інструментів для візуалізації: деякі з них зовсім прості: потрібно тільки завантажити дані та вибрати, як вони будуть відображатися. Інші програми більш складні і комплексні — вимагають настройки і, наприклад, знань JavaScript.

Ми підібрали найрізноманітніші варіанти: і для тих, кому потрібен швидкий зрозумілий результат, і для просунутих користувачів. Є з чого вибрати.

Plotly. Будує дуже докладні графіки. Ця програма створює діаграми, презентації та дашборди. Ти можеш виконати аналіз за допомогою JavaScript, Python, R, Matlab, Jupyter або Excel. Також є кілька варіантів імпорту даних. Бібліотека візуалізації та інструмент для створення діаграм в режимі онлайн дозволяють створювати по-справжньому красиві графіки.

DataHero. Добре підходить, щоб зібрати інформацію з безлічі сервісів в єдину систему. У DataHero можна інтегрувати дані з хмарних сервісів і створювати діаграми та дашборди. Не потребує ніяких спеціальних технічних знань, тому це відмінний інструмент, яким може користуватися вся команда.

Chart.js. Чудово підходить для невеликих проєктів. Незважаючи на те, що програма пропонує всього 6 видів діаграм, безкоштовна бібліотека Chart.js підійде для невеликих проєктів. Для побудови діаграм програма використовує HTML5 Canvasі створює швидко реагуючий на зміни простий дизайн.

Tableau. Створює набори даних, якими можна ділитися в режимі реального часу.

Tableau Public — це практично безкоштовний інструмент візуалізації з графіками, діаграмами, картами та іншим. Ви легко зможете завантажити інформацію в систему, а потім спостерігати за тим, як все оновлюється. Для прискорення процесу можна працювати одночасно з іншими учасниками проєкту.

Raw. Безкоштовний веб-додаток з простим інтерфейсом. Це додаток з відкритим кодом, який можна безкоштовно скачати, змінити і налаштувати під себе. У ньому можна робити векторні візуалізації у форматах SVG або PNG.

Dygraphs. Підходить для візуалізації великої кількості даних. Це безкоштовний додаток, що дозволяє досліджувати та пояснювати великі обсяги даних. Ви можете налаштувати програму так, як потрібно саме вам, вона працює в усіх основних браузерах. Є функція стиснення графіків для смартфонів і планшетів.

ZingChart. Створює діаграми за допомогою HTML5 Canvas. ZingChart — це бібліотека діаграм на JavaScript. Завдяки багатофункціональному API можна створювати інтерактивні Flash або HTML5-діаграми. У програмі понад 100

варіантів діаграм, щоб ви могли вибрати відповідний для ваших цілей і формату даних.

InstantAtlas. Створює гарні візуалізації у вигляді інформаційних карт. Якщо вам потрібен інструмент для візуалізації даних з карт, зверніть увагу на InstantAtlas. У ньому можна створювати інтерактивні динамічні та вузькопрофільні звіти, які об'єднують статистику та картографічну інформацію.

Timeline. Створює інтерактивний таймлайн.Timeline — це зручний віджет, який реагує на рухи мишки користувача. Він спрощує створення графіків з великою кількістю інформації, і видає їх в компактному вигляді. До кожного елементу можна додати більш розгорнуту інформацію, яка буде відображатися при натисканні — жодна деталь не буде упущена.

Exhibit. Перетворює візуалізацію даних на гру. Цей інструмент дозволяє легко створювати інтерактивні карти та інші візуалізації, які можна використовувати в навчальних цілях. Добре підходить для статистичних та історичних наборів даних, таких як прапори різних країн або місця народження відомих людей.

Modest Maps. У цій програмі можна робити інтерактивні карти та вбудовувати їх на сайт. Цей плагін підходить для дизайнерів, що вважають за краще допрацьовувати функціонал під особисті потреби з урахуванням власного користувальницького досвіду. API підключається досить просто, є можливості для додавання власного коду. Основну бібліотеку можна розширити за допомогою додаткових плагінів із корисними опціями.

Leaflet. Дозволяє використовувати дані з OpenStreetMap і візуалізувати їх за допомогою HTML5 та CSS3. Ще один інструмент для створення карт, в якому можна створити повністю інтерактивну візуалізацію.

Основна бібліотека сама по собі дуже маленька, але існує величезна кількість плагінів, які розширюють функціонал до рівня профі. Наприклад, можна додати анімовані позначки, маски та зони активності. Ідеально підходить для проєктів, де потрібно показати дані, накладені на географічну розмітку (включаючи нестандартне проєктування).

WolframAlpha. Дуже добре справляється зі створенням діаграм. Інструмент добре створює діаграми за запитами даних, не потребує додаткового налаштування. Якщо ви хочете візуалізувати загальнодоступні дані, то підійде простий конструктор віджетів.

Visual.ly. Спрощує візуалізацію даних настільки, наскільки це можливо. Visual.ly — це одночасно і галерея, і інструмент для створення інфографіки. Використовуючи простий набір опцій, можна створювати красиві візуалізації даних. Це не просто візуалізація даних, а щось фантастичне, мрія інфоманіяка!

Visualize Free. Visualize Free — це безкоштовний інструмент, в якому можна використовувати загальнодоступні дані або завантажувати власні і створювати інтерактивні візуалізації. Візуалізації виходять далеко за рамки простих графіків. Для роботи потрібен Flash, але результат може виводитись і в

HTML5.

FusionCharts. Комплексне рішення для побудови діаграм на JavaScript та HTML5. FusionCharts Suite XT пропонує більше 90 графіків і макетів, 965 карт з даними, готові бізнес-панелі та демки. JavaScript API дозволяє легко інтегрувати плагін в будь-який AJAX-додаток або JavaScript-фреймворк. Діаграми, карти та інформаційні панелі неймовірно інтерактивні, їх легко налаштовувати і вони працюють на всіх пристроях і платформах. У додатку також є порівняльний аналіз топових бібліотек діаграм JavaScript.

jqPlot. Чудове рішення для лінійних і точкових діаграм. До плагіну додається кілька приємних додаткових функцій, таких як автоматичне створення трендових ліній та інтерактивних точок, які можуть коригувати відвідувачі сайту, відповідно оновлюючи набір даних.

D3.js. Створює незвичайні діаграми. D3.js — це бібліотека JavaScript, що створює діаграми у форматах HTML, SVG та CSS. Можна використовувати різні джерела даних. Ця бібліотека може сильно підвищити рівень візуалізації складних наборів даних. Програма безкоштовна і використовує веб-стандарти, тому дуже зручна і доступна для користувачів. Також є цікаві варіанти інтерактивної підтримки.

JavaScript InfoVis Toolkit. Фантастична бібліотека, написана Ніколасом Бельмонте. Модульна структура дозволяє завантажувати тільки те, що абсолютно необхідно для створення візуалізацій. Є ряд унікальних стилів та анімаційних ефектів. Бібліотеку можна використовувати безкоштовно (хоча заохочуються донати).

Highcharts. Плагін пропонує великий вибір опцій. Highcharts — це графічна бібліотека JavaScript з величезним діапазоном доступних варіантів діаграм. Результат візуалізується з використанням SVG в сучасних браузерах і VML в Internet Explorer. Графіки автоматично підтримують гарну анімацію, а фреймворк — потоки даних в реальному часі. Highcharts можна завантажити безкоштовно і використовувати в некомерційних цілях (або купити ліцензію для комерційного використання). Також можна відтворювати демки, використовуючи JSFiddle.

Excel. Графічно зовсім не гнучкий, але це хороший спосіб вивчити дані. Наприклад, створивши "теплові карти", подібні до цієї. Деякі досить складні речі можна робити за допомогою Excel: починаючи з "теплових карт" по клітинам до приблизних діаграм. Як інструмент для початкового рівня він дозволяє швидко вивчити дані або створити візуалізацію для внутрішнього використання. Але є обмеження: стандартний набір кольорів, ліній та стилів ускладнює створення графіки. Проте, він підходить в якості засобу швидкої передачі ідей.

Для цих цілей можна використовувати і електронні таблиці Google. В них можна створювати ті самі діаграми, що і в API Google Chart.

Crossfilter. Кросфільтр в дії: обмежуючи діапазон введення на якомусь

одному графіку, ми зачіпаємо всі дані. Це чудовий інструмент для панелей моніторингу або інших інтерактивних інструментів з великими обсягами даних.

У міру того, як з'являються все більш складні інструменти, що дозволяють людям продиратися крізь дані, графіки та діаграми перетворюються на інтерактивні віджети графічного інтерфейсу. Бібліотека JavaScript Crossfilter може бути і першим, і другим. Вона не тільки відображає дані, але і дозволяє побачити реакцію інших пов'язаних діаграм при обмеженні діапазону даних.

R. Потужна безкоштовна програма для статистичних обчислень і створення графіки. R — це найбільш складний з перерахованих тут інструментів.

Як статистичний збірник, застосовуваний для аналізу великих наборів даних, R — дуже складний інструмент, який вимагає часу на навчання, але пропонує потужну підтримку від інших фахівців та пакетну бібліотеку, яка постійно розширюється. А ще в ньому є власна пошукова система.

Навчитися працювати з цією програмою буде складніше, ніж із будь-якою іншою з перерелічених тут, але це того варте.

Weka. Weka — це набір алгоритмів машинного навчання для задач інтелектуального аналізу даних. Потужний засіб для вивчення та опрацювання інформації. Weka -— хороший інструмент для класифікації та кластеризації даних, але в ньому можна створювати і прості графіки.

Тема 3. Поняття ринку великих даних. Життєвий цикл аналітики даних. Збір та підготовка даних

1. Ринок великих даних: переваги, недоліки та ризики
2. Поняття життєвого циклу великих даних
3. Джерела даних
4. Збір та підготовка даних

1. Ринок великих даних: переваги, недоліки та ризики

Детермінантою сучасного етапу розвитку економіки є перехід до нового технологічного укладу, який обумовлює зміну продуктивних сил та виробничих відносин. Виклик суспільству, сформований цифровою трансформацію, сприяв зародженню нових технологічних продуктів та послуг, формуванню нових форм соціально-економічних відносин та способів цифрової взаємодії між суб’єктами товарних ринків, інтеграції окремих галузевих ринків та секторів економіки. Високо динамічна цифровізація економіки, заснована на перевагах від використання Big Data, пришвидшує використання в управлінських та виробничих процесах штучного інтелекту, робототехніки, хмарних технологій тощо. Однак, динамічне формування глобального цифрового ринку у міжнародній економічній системі супроводжується значними соціально-економічними протиріччями між країнами із розвинутою ринковою економікою

та інституціонально недостатньо розвинутими країнами, до яких належить Україна.

Існуюча ринкова ситуація вимагає посилення орієнтації соціально-економічного розвитку окремих держав в частині забезпечення збалансування процесів трансформації національних ринків окремих країн з позиції поліпшення їх конкурентних позицій завдяки становленню у них цифрової економіки. Прикладні аспекти цифрової трансформації суспільства базуються на використанні технологій Big Data. Останні стають інструментом стратегічного планування, підвищення операційної ефективності, рівнів маркетингово-логістичного сервісу клієнтів в таких компаніях, як Nasdaq, Facebook, Google, IBM, VISA, Master Card, Bank of America, HSBC, AT&T, Coca Cola, Starbucks та Netflix тощо. Підвищення точності прогнозування попиту споживачів, моделювання та візуалізація у процесі створення моделей нових продуктів і послуг, підтримка прийняття рішень, управління маркетинговими та логістичними ризиками, підвищення маржі на етапах створення доданої вартості тощо – лише деякі можливості системи інформаційно-аналітичного забезпечення підприємств на засадах використання масивів Big Data та цифрової обробки інформації. Очікується, що підвищення адаптаційної здатності завдяки роботі з Big Data даними, розвиток технологій захисту інформації та діджиталізація процесів виробництва та збуту продукції сприятиме підвищенню інформаційної безпеки та запобіганню кіберзагрозам підприємств, які працюють в умовах ринкової глобалізації та підвищених ризиків, сформує засади для забезпечення економічної безпеки підприємств. Це дозволяє стверджувати, що тенденції розвитку глобального ринку Big data суттєво позначаються на розвиткові інших галузей, в т. ч. суміжних, а відтак свідчать про актуальність та перспективність даного дослідження.

Світовий технологічний прогрес нерозривно пов’язаний із зростанням обсягу інформації, зокрема у цифровому вимірі та Інтернет-мережі. За прогнозами, до 2021 року глобальний IP-трафік досягне значення 3,3 ЗБайт, і 1,7 Мбайт нової інформації створюватиметься щосекунди.

Глобальні дані (англ. Big Data) – позначення структурованих і неструктурованих масивів даних значних обсягів, що не піддаються обробці за допомогою традиційних способів та підходів. У більш широкому сенсі Big Data – це набір інструментів та методів, які надають можливість аналізувати великі масиви інформації. Застосування технологій Big Data за ефективністю займає третє місце після контент-маркетингу (content marketing) та штучного інтелекту (artificial intelligence). Показники ідентифікації динаміки розвитку глобального ринку Big Data за період 2011 – 2020 рр. наведено у табл. 3.1.

Таблиця 3.1

Аналіз показників ідентифікації динаміки розвитку глобального ринку Big Data

Показник	Рік										2020/2011
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	
Сукупний дохід, млрд. дол. США	7,6	12,25	19,6	18,3	22,6	27,36	35,29	42,16	48,79	61,01	8,03
Чисельність користувачів мережі Інтернет, млн. осіб	2242	2478	2669	2853	3060	3345	3701	3924	4131	4540	2,02
Втрати суб'єктів ринку Big Data від витоку даних, млн. дол. США	5,5	5,4	3,14	3,5	3,79	4,0	3,62	3,86	3,92	4,14	0,75

Аналіз динаміки сукупного доходу ринку Big Data свідчить про його постійне зростання, що також пов'язано зі зростанням втрат суб'єктів ринку від ризиків витоку даних. За прогнозами аналітичної спільноти Wikibon, доходи від глобального ринку Big Data збільшаться з 42 млрд. дол. США у 2018 році до 103 млрд. дол. США у 2027 році, досягнувши загального річного темпу зростання у розмірі 10,48%. Поділ сукупного доходу глобального ринку Big Data за основними сегментами джерел цього доходу поданий у табл. 3.2.

Таблиця 3.2

Аналіз динаміки розвитку основних сегментів глобального ринку Big Data, за показником сукупного доходу, млрд.дол. США

Сегмент	Рік					2020/2016
	2016	2017	2018	2019	2020	
Послуги	8	14	16	19	21	2,63
Апаратне забезпечення	9	10	12	14	15	1,67
Програмне забезпечення	11	11	14	17	20	1,82
Разом	28	35	42	50	56	2,00

Відповідно до поділу ринку Big Data можна зробити висновок, що найбільшу частку у ньому займає ринок послуг (37,5 % у 2020 р.), проте, за прогнозом Statista, вже із 2021 року роль програмного забезпечення значно зросте і переважатиме інші категорії у структурі ринку Big Data. Основними факторами стрімкого збільшення розмірів ринку Big Data є зростання обізнаності організацій щодо пристроїв інтернету речей (Internet of Things).

Ключовими гравцями на ринку Big Data у 2019 році є Китай, США, Канада, Франція та Великобританія. Динаміка показників розміру ринків Big Data цих країн за 2017 – 2019 рр. подана у табл. 3.3.

Таблиця 3.3

Аналіз динаміки доходів країн-лідерів ринку Big Data, млн. дол. США

Країна	2017	2018	2019	2019/2018, %
США	9782,3	12341,0	15209,0	123,2
Велика Британія	1452,4	1882,1	2354,9	125,1
Китай	747,2	1460,6	2 392,6	163,8
Канада	453,7	558,7	768,8	130,6
Франція	232,0	340,7	469,5	137,8

Серед великих учасників ринку Big Data є також суб'єкти регіону APAC: Індія, Південна Корея та Японія, які роблять акцент на вдосконаленому управлінні даними для забезпечення розвитку бізнес-рішень та бізнес-процесів. Очікується, що зростаюча діджиталізація та поширення впровадження технологій Big Data, таких як Hadoop та Apache суб'єктами регіону, а також сприятливі урядові постанови зумовлюватимуть ріст ринку Big Data APAC.

Значну частку глобального ринку Big Data становлять персональні дані фізичних осіб. Для прикладу, показник монетизації користувачів соціальної мережі Facebook у 4 кварталі 2019 року склав 8,52 доларів США за особу. За результатами опитування користувачів Інтернету в Канаді у 2015 році, більшість осіб згодні надавати підприємствам свої дані в обмін на персоналізовані послуги чи винагороди від підприємств, які отримуватимуть ці дані. В Україні деякі підприємства практикують таку систему обміну зі своїми клієнтами; також є створені онлайн- платформи для опитувань (наприклад, Opinion.com.ua), за участь в яких користувачі отримують винагороду у вигляді віртуальних коштів, які, при накопиченні до певного обсягу, можливо конвертувати у реальні.

Попри те, що користувачі у більшості випадків погоджуються надавати свої персональні дані певним організаціям, вони є стурбованими щодо захисту персональної інформації цими організаціями. Споживачів хвилює ймовірність поширення організаціями їх персональних даних третім сторонам (37%), а також ризик витоку даних через недостатньо потужну систему інформаційної безпеки підприємств (29%). Для захисту своїх персональних даних користувачі вживають такі заходи, як регулярна перевірка кредитної історії на наявність незнайомих трансакцій (80%), перевірка програмного захисту ПК (77%), знищення (подрібнення) документів, котрі містять персональні дані (70%), використання різних паролів для різних користувацьких акаунтів тощо. Окрім того, більше 50% користувачів знають про своє право на огляд, коректування, оскарження та зупинку подання своєї персональної інформації будь-яким підприємствам, які нею володіють.

Для дослідження глобального ринку Big data доречним є використання методики аналізу п'яти сил конкуренції Портера. Результати проведення даного аналізу подані у таблиці 3.4.

Результати аналізу ринку Big Data за п'ятьма силами конкуренції М. Портера свідчать про те, що найбільш вагомою із конкурентних сил на ринку є високий рівень конкурентної боротьби, за якого суб'єктам ринку рекомендовано зосереджувати увагу на потенційних потребах та сподіваннях своїх клієнтів для посилення бази диференціації та чіткого позиціонування своїх послуг.

Для проведення ефективного дослідження тенденцій ринку Big Data необхідним є здійснення оцінки чинників його внутрішнього та зовнішнього середовищ. Одним із найбільш поширених інструментів для реалізації цього завдання є проведення SWOT-аналізу ринку. Перелік ідентифікованих сильних,

слабких сторін, можливостей та загроз глобального ринку Big Data подано у табл. 3.5.

Відповідно до проведеної оцінки сильних, слабких сторін, можливостей та загроз ринку сформовано матрицю SWOT-аналізу, квадранти якої містять перелік можливих стратегій для більш ефективного використання сильних сторін, мінімізації слабких сторін, зниження загроз від зовнішніх факторів та ефективного використання можливостей (табл. 3.4).

Таблиця 3.4

Аналіз п’яти сил конкуренції Портера на ринку *Big Data*

Параметр	Значення	Опис	Напрям роботи
Загроза появи нових гравців на ринку	Середнє	Оскільки галузь є прибутковою і має відносно невисокий поріг для входу, то до неї залучатимуться більше нових учасників, що, відповідно, становить загрозу для вже існуючих на цьому ринку компаній. Наприклад, у 2020 році спостерігалось збільшення кількості спеціалізованих стартапів на ринку Big Data: Apheris, Cinnamon AI, Dataiku, DataKitchen та ін.	Нарощення потенціалу збільшення витрат на оперування масивами Big Data, в т. ч. для зменшення ймовірності входу нових учасників; розвиток лояльності споживачів до бренду, щоб запобігти переходу клієнтів до нових конкурентів (бренд Netflix використовує Big Data для покращення таргетованої реклами, утримуючи клієнтів на своїй платформі).
Ринкова влада поста-чальників	Середнє	Постачальники здійснюють тиск на бізнес– організації, застосовуючи зменшення доступності товару, зниження якості або підвищення цін тощо. Багато постачальників ПЗ для баз даних, такі як Ahana, Cockroach Labs, Databricks, починають використовувати власні інструменти управління Big Data, створюючи конкуренцію вже існуючим на ринку підприємствам.	Формування ефективних взаємовідносин із кількома постачальниками; розвиток спеціалізованих постачальників, бізнес яких залежить від фірми (на прикладі WallMart і Nike, UserTesting та Facebook, Tamr і Toyota)); редизайн та диверсифікація товарних ліній підприємств.
Ринкова влада споживачів	Середнє	Покупці здійснюють тиск на бізнес-організації з метою отримання високоякісної продукції за доступними цінами з високим рівнем сервісу. Ця сила безпосередньо впливає на здатність учасників ринку досягти бізнес-цілей.	Збільшення диверсифікованості клієнтської бази; введення нових товарів, орієнтуючись на нові сегменти ринку. Н-д, великі компанії IBM, Docker, Atlassian та Instacart використовують платформу Segment для реалізації зазначеного напряму.
Загроза появи товарів–замінників	Низьке	Висока ймовірність використання субститутів з інших галузей для задоволення потреб споживачів (н-д, сервіси Dropbox та Google Drive є замінниками апаратних накопичувачів) може бути обумовлена нижчою ціною, вищим рівнем якості та ефективністю від використання тощо.	Чітке позиціонування переваг пропонованого товару над товарами-замінниками (даний елемент у сфері Big Data своїх підприємств використовували The Marriott hotels, Amazon, Netflix та Uber Eats); спрямування зусиль на підвищення лояльності та довіри споживачів; покращення якості продукції, що пропонується;

Рівень конкурентної боротьби	Високе	Гостра конкуренція сприяє зниженню середньоринкових цін та зростанню загальної прибутковості галузі. Найбільшими гравцями ринку Big Data є компанії IBM, Google, Oracle, Microsoft та Amazon Web service, які чинять конкурентних тиск на менші підприємства, що зумовлює обмеження потенціалу зростання усіх підприємств.	Зосередження уваги на наявних потребах та сподіваннях своїх клієнтів для посилення бази диференціації; інвестування в науково-дослідну діяльність для визначення нових сегментів клієнтів. У 2020 році 55% інвестицій у Big Data великих компаній були спрямовані на пошук IT рішень, 26% – у зовнішній технічний консалтинг.

Таблиця 3.5

SWOT-аналіз глобального ринку Big Data

Сильні сторони (Strengths):	**Слабкі сторони (Weaknesses):**
1) розширення бізнесу внаслідок збільшення обсягу інформації, якою він володіє; 2) зростання кількості відгуків клієнтів через соціальні мережі; 3) встановлення стратегічних партнерських стосунків з постачальниками, дилерами та іншими зацікавленими особами завдяки застосуванню Big Data; 4) перманентне підвищення кваліфікації працівників для утримання конкурентоспроможності організацій; 5) налагоджена IT-система підприємства сприяє більш швидкому прийняттю ефективних управлінських рішень; 6) високі доходи, зумовлені прийняттям ефективних управлінських рішень, володіння результатами дослідження ринку завдяки технологіям Big Data; 7) здійснення прогнозування високої точності та визначення потенційних ризиків на основі використання великих масивів даних;	1) традиційні підходи управління інформацією є неефективними; 2) робота з великими обсягами інформації потребує інноваційного програмного та апаратного забезпечення; 3) ринок Big Data характеризується високою плинністю кадрів, що зумовлює зростання витрат на підбір та навчання нових працівників; 4) ефективне застосування Big Data потребує залучення висококваліфікованих кадрів, які потребують заробітної плати відповідного рівня; 5) більшість продуктів володіють низькою часткою ринку, що зумовлює залежність Big Data від тієї меншості продуктів, які володіють більшою часткою ринку; це спричиняє вразливість Big Data до зовнішніх загроз; високі витрати на дослідження та розробку; 7) зберігання даних у хмарних середовищах Big Data вважається відносно ненадійним
1) зростання чисельності населення, що означає збільшення кількості потенційних споживачів та обсягу даних, що збираються ; 2) зростання кількості підприємств, які впроваджують e-commerce у свою діяльність; 3) приріст активних споживачів за рахунок інтеграції Big Data у соціальні мережі; 4) збільшення частки автоматизованих процесів, що сприяє зниженню витрат; 5) зростання популярності IT– спеціалізації у ВНЗ; 6) глобалізація економіки, що дозволяє підприємствам поширювати свою діяльність на інші країни.	1) побоювання носіїв даних щодо конфіденційності можуть спричинити публічний/ приватний опір Big Data; збільшення кількості кібератак; 3) витік чи втрата даних внаслідок оброблення їх третьою стороною; 4) обробка некоректних даних спричиняє помилкові управлінські рішення; 5) посилення обмежень щодо збору даних споживачів урядом; 6) велика популярність Big Data спричиняє збільшення припливу нових гравців; 7) прискорення процесу насичення ринку внаслідок його високої актуальності, що у майбутньому зумовить перенасичення цього ринку; 8) перехід висококваліфікованих працівників підприємства до конкурента.

Проведення SWOT-аналізу ринку Big Data (табл. 3.5) дозволило виявити

особливості його функціонування. Сильними сторонами ринку Big Data є: розширення бізнесу внаслідок збільшення обсягу інформації, якою він володіє; зростання кількості відгуків клієнтів через соціальні мережі; встановлення стратегічних партнерських стосунків з постачальниками, дилерами та іншими зацікавленими особами завдяки застосуванню Big Data; перманентне підвищення кваліфікації працівників для утримання конкурентоспроможності організацій; налагоджена IT-система підприємства, яка сприяє більш швидкому прийняттю ефективних управлінських рішень; високі доходи, зумовлені прийняттям ефективних управлінських рішень, володіння результатами дослідження ринку завдяки технологіям Big Data. Виявленими можливостями ринку є: зростання чисельності населення, що означає збільшення кількості потенційних споживачів та обсягу даних, що збираються; зростання кількості підприємств, які впроваджують e-commerce у свою діяльність; приріст активних споживачів за рахунок інтеграції Big Data у соціальні мережі; збільшення частки автоматизованих процесів, що сприяє зниженню витрат; зростання популярності IT-спеціалізації у ВНЗ; глобалізація економіки, що дозволяє підприємствам поширювати свою діяльність на інші країни.

Відповідно до проведеної оцінки сильних, слабких сторін, можливостей та загроз ринку сформовано матрицю SWOT-аналізу, квадранти якої містять перелік можливих стратегій для більш ефективного використання сильних сторін, мінімізації слабких сторін, зниження загроз від зовнішніх факторів та ефективного використання можливостей (табл. 3.6).

Таблиця 3.6

Матриця SWOT-аналізу ринку Big Data

	Можливості (O)	***Загрози (T)***
Сильні сторони (S)	**Стратегії SO:** - вихід підприємств на нові ринки завдяки ринковій глобалізації та ефективному впровадженню Big Data (S1, O6); - використання підприємствами соціальних мереж для збору даних про споживачів та їх залучення у процеси Big Data підприємства (S1, S2, O3); - запровадження та вдосконалення системи e- commerce підприємств завдяки можливостям налагоджених IT–систем підприємства із Big Data (S5, O2); - зниження цін на продукцію завдяки зниженим витратам та ефективній взаємодії з контрагентами (S5, O4).	**Стратегії ST:** - формування потужної дистрибуційної мережі (наприклад, Facebook Inc.) для більшого охоплення ринку та боротьби з новими учасниками ринку (S1, S3, T6); - використання сильного фінансового становища для інвестицій у права інтелектуальної власності; це дасть додаткові переваги над конкурентами (S6, T6); - проведення постійного підвищення кваліфікації працівників сприятиме збору та обробці більш релевантних даних (S4, T4).

Слабкі сторони (W)	**Стратегії WO:** - збільшення заробітної плати працівників, надання заохочуваних пакетів та вигод працівникам для зменшення плинності кадрів та покращення морального стану працівників. Це можливо завдяки зниженню витрат через автоматизацію процесів (W3,W4,O4); - залучення молодих кваліфікованих працівників, які отримали спеціалізовану вищу освіту у вітчизняних ВНЗ (W3, O5).	**Стратегії WT:** - збільшення витрат на дослідження ринку та розробки за для підвищення конкурентних переваг підприємства та мінімізації збору та обробки некоректних даних (W6, T4, T6); - забезпечення стимулів та організація кращих робочих умов для збереження висококваліфікованих кадрів на підприємстві (W3, W4, T8).

На основі виявлених сильних сторін ринку та його можливостей запропоновано такі стратегії: вихід підприємств на нові ринки завдяки ринковій глобалізації та ефективному впровадженню Big Data, використання підприємствами соціальних мереж для збору даних про споживачів та їх залучення у процеси Big Data підприємства, запровадження та вдосконалення системи e-commerce підприємств завдяки можливостям налагоджених IT–систем підприємства із Big Data, зниження цін на продукцію завдяки зниженим витратам та ефективній взаємодії з контрагентами.

Проведений SWOT-аналіз глобального ринку Big Data та аналіз чинників ринкового середовища дозволили ідентифікувати такі ризики суб'єктів ринку Big Data.

1. Ризик втрати даних внаслідок хакерських атак. Ймовірність даного ризику зростає при збільшенні обсягу даних підприємства. Наприклад, у грудні 2013 року база даних роздрібної мережі Target зазнала хакерської атаки, яка призвела до витоку даних кредитних карт більш ніж 40 мільйонів клієнтів.
2. Ризик знищення конфіденційності даних. Наприклад, у березні 2020 р. готельна мережа Marriott Internaional оголосила про несподіване отримання доступу до даних 5,2 мільйонів клієнтів через використання облікових записів співробітників.
3. Ризик зростання витрат на збір, обробку та зберігання даних. Помилка у плануванні бюджету може призвести до спіральних витрат, що у майбутньому спричинить анулювання доданої вартості, створеної завдяки використанню Big Data.
4. Ризик проведення неефективної аналітики зібраних даних.
5. Ризик збору неправдивих, некоректних, неякісних даних. Велика частка проектів є невдалими через використання неактуальних, застарілих або помилкових даних. За результатами дослідження MarketingWeek, 60% інтернет-користувачів Великої Британії навмисно подають недостовірну інформацію при наданні своїх особистих даних в Інтернеті, намагаючись зберегти свої дані приватними .
6. Ризик невідповідності дій над даними чинному законодавству.
7. Ризик формування висновків із низькою точністю.

8. Ризик порушення інтелектуальної власності третьої сторони.

9. Ризик виникнення етичних дилем. У 2014 році система охорони здоров'я Carolinas HealthCare здійснювала придбання даних про своїх пацієнтів. Попри те, що деякі пацієнти можуть схвалювати такий підхід, такі дії є вторгненням у приватне життя клієнтів. Це засвідчує про виникнення етичних дилем на підприємствах, які використовують Big Data.

10. Ризик хибної організації (структуризації) зібраних даних. Матриця ризиків суб'єктів ринку Big Data подана на рис. 3.

Значний	-	-	Знищення конфіденційності даних (Р2)	Хакерська атака (Р1)
Великий	Порушення законодавства (Р6)	Неефективна аналітика (Р4)	Зростання витрат (Р3)	Неправдиві дані (Р5)
Помірний	Хибна структуризація даних (Р10)	Неточні висновки (Р7) Етичні дилеми (Р9)	Порушення авторських прав (Р8)	-
Незначний	-	-	-	-
Збиток / Ймовірність	Дуже низька (<9%)	Низька (від 10 до 24%)	Середня (від 25 до 49%)	Висока (від 50%)

Рис. 3.1. Матриця ризиків суб'єктів ринку Big Data

За результатами побудови матриці ризиків суб'єктів ринку Big Data визначено, що найбільш вагомими ризиками даного ринку є зниження інформаційної безпеки суб'єкта внаслідок хакерських атак та знищення конфіденційності даних. Суб'єктам ринку Big Data необхідно здійснювати систематичний контроль захисту внутрішнього інформаційного середовища для своєчасної ідентифікації зазначених ризиків та їх якнайшвидшого усунення. Якісна інтерпретація ризиків глобального ринку Big Data подано у табл. 3.5.

Найбільш поширений ризик, на думку авторів, полягає у втраті даних внаслідок хакерських атак. Це активізує застосування методики оцінювання даного ризику за послідовністю: ризик → загроза, яку він несе → вразливість → обґрунтування рівня ймовірності настання ризику (високий, 3 б.; середній – 2 б., низький – 1 б.) → обґрунтування рівнів наслідків прояву ризику → визначення загального рівня ризику → розробка положень щодо ефективності пом'якшувальних заходів → оцінювання чистого ризику. Деталізований опис та оцінка ризику втрати Big Data внаслідок хакерських атак подано у табл. 3.8.

Таблиця 3.7

Якісна інтерпретація ризиків на глобальному ринку Big Data

Ідентифіковані ризики	Коментар
Чинники зовнішнього середовища	
Високий рівень хакерських атак	Згідно із «Звітом про захист від кіберзагроз за 2015 рік», 71% організацій у 19 галузях, які функціонують у Пн. Америці та Європі, стали жертвами кібератак у 2014 році. 46% усіх підприємств Великої Британії протягом 2018 року виявили щонайменше 1 порушення захисту даних або кібератаку.
Високі законодавчі обмеження щодо дій над даними	Уряд кожної країни (або груп країн) самостійно здійснює законодавче регулювання Big Data на території свого впливу. У травні 2018 року набув чинності Загальний регламент ЄС про захист персональних даних (GDPR), який обмежує права організацій щодо дій над персональними даними своїх споживачів.
Значні законодавчі обмеження щодо дій над даними	Регулювання Big Data у США регулюється за допомогою різних статутів: у сфері медичного обслуговування – Закон «Про переносність та підзвітність» від 1996 р. (HIPPA), у сфері шкільної освіти – Закон «Про сім'ю та конфіденційність», 1974 р. (FERPA).
Ненадійність хмарних сховищ для зберігання даних	Ненадійність хмарних сховищ може бути спричинена підвищенням частоти збоїв у хмарних сховищах, які включають переповнення, відсутність ресурсів даних, збої у базах даних, програмному забезпеченні, апаратному забезпеченні та у з'єднанні із мережею Internet.
Висока динамічність глобального ринку Big Data	Висока динамічність ринку відображається темпами його зростання. За даними Technavio, протягом наступних 3 років очікується зростання глобального ринку Big Data на 17% .
Високий рівень недовіри індивідуальних носіїв даних до компаній із Big Data	Високий рівень недовіри відображається у стурбованості споживачів щодо надійності зберігання їх персональних даних організаціями. За результатами дослідження MarketingWeek, 60% інтернет-користувачів Великої Британії навмисно подають недостовірну інформацію при наданні своїх особистих даних в Інтернеті, намагаючись зберегти свої дані приватними.
Чинники внутрішнього середовища	
Висока плинність кадрів на ринку Big Data	Однією із найбільших проблем організацій є утримання висококваліфікованих працівників. Ринок Big data характеризується високою плинністю кадрів, що пояснюється великим інтелектуальним навантаженням на працівників, великим вибором потенційних місць працевлаштування із різними методами заохочення та винагороди персоналу.
Низька якість досліджень та розробок на основі Big Data	Якість досліджень залежить від фінансових та інтелектуальних засобів, залучених у проведення досліджень. Недостатні витрати на дослідження спричиняють отримання менш точних та ефективних результатів дослідження.
Висока вартість утримання та обслуговування Big Data	Технологія Big Data у межах підприємства потребує залучення потужного апаратного та програмного забезпечення та висококваліфікованих працівників. В усіх випадках придбання (найм) та утримання цих ресурсів становить велику частку витрат у бюджеті підприємства.

Таблиця 3.8

Оцінювання ризику втрати даних внаслідок хакерських атак

Послідовність визначення ризику	Обґрунтування етапу оцінювання
Ризик	Витік або втрата даних внаслідок хакерських атак.
Загроза	Втручання третіх сторін у інформаційні потоки підприємств з метою викрадення, зміни чи знищення масивів даних. Зловмисні акти є найпоширенішою причиною порушення безпеки даних (48%), решта – викликані збоєм IT-системи підприємства або помилками людини .
Вразливість	Недостатньо потужна система захисту інформації підприємства.
Рівень ймовірності настання ризику (3б.)	У 2017 р. спостерігалось збільшення кількості кібератак на 600% (від 6 тис. атак у 2016 р. до 50 тис. у 2017 р.). Окрім того, активність цільових нападів зросла на 10% у 2017 році порівняно із 2016 р.. За підрахунками Varonis, станом на березень 2020 року кожні 39 секунд у світі відбувається 1 кібератака. Відповідно до звіту про кібербезпеку Ponemon Institute, 83% фінансових компаній і 44% підприємств роздрібної торгівлі зазнають близько 50 атак на місяць. Щодня з'являється близько 230 тис. нових зразків зловмисного програмного забезпечення, а варіанти Ransomware у 2017 році зросли на 46%.
Рівень наслідків прояву ризику (3б.)	Згідно з доповіддю про глобальні ризики Всесвітнього економічного форуму за 2018, кібератаки знаходяться у трійці найбільших ризиків для глобальної світової стабільності протягом наступних п'яти років. За оцінками Varonis, середня вартість зламаних даних перевищить 150 млн. дол. США до кінця 2020 року. Внаслідок витоку даних роздрібної мережі Target у 2013 році акції мережі знизились на 2,2%, а ринкова оцінка вартості втрат склала 890 млн. дол. США. Цільовий прибуток зменшився на 1,59 млрд. дол. США.
Загальний рівень ризику	9 балів
Рівень ефективності пом'якшувальних заходів (2 б.)	Незважаючи на великий ризик атак, більше половини малого бізнесу (51%) не виділяють бюджет на зменшення кіберризиків. Проте уряд держав передбачає виділення коштів на захист від кіберзлочинів та ліквідацію їх наслідків. Некласифіковані федеральні видатки США на кіберзабезпечення зросли з 7,5 мільярдів доларів у 2007 році до 28 мільярдів доларів у 2016 році.
Рівень чистого ризику	6 балів

Рівень чистого ризику становить 6 балів. Відповідно до шкали рівнів ризику, дана оцінка відповідає рівню «високий». На основі отриманого результату можливо стверджувати, що діяльність суб'єктів ринку Big Data є вразливою до можливих хакерських вторгнень та потребує впровадження більш ефективних заходів для зниження рівня ризику та забезпечення надійного захисту даних підприємств та їх клієнтів.

За допомогою SWOT-аналізу ідентифіковано такі основні ризики галузі: ризик знищення конфіденційності даних, ризик збору неправдивих даних, ризик порушення інтелектуальної власності третьої сторони та ін. Сформована матриця ризиків свідчить про те, що, найбільш вагомими ризиками даного ринку є зниження інформаційної безпеки суб'єкта внаслідок хакерських атак (ймовірність справдження від 50%, значна величина збитків) та знищення

конфіденційності даних (ймовірність справдження від 25%, значна величина збитків). Проведена якісна інтерпретація ризиків на ринку Big Data дозволила охарактеризувати вплив несприятливих факторів внутрішнього та зовнішнього середовищ, а саме: недостатній рівень ненадійності хмарних сховищ для зберігання даних, високий рівень недовіри носіїв даних до компаній, що використовують Big Data, висока плинність кадрів на ринку та ін. Оцінювання ризику втрати даних внаслідок хакерських атак дозволило ідентифікувати його як ризик із високим рівнем важливості (6 б.). На основі отриманого результату зроблено висновок, що діяльність суб'єктів ринку Big Data є вразливою до можливих хакерських вторгнень та потребує впровадження більш ефективних заходів для забезпечення надійного захисту даних підприємств та їх клієнтів.

2. Поняття життєвого циклу великих даних

Аналіз великих даних відрізняється від традиційного аналізу даних в першу чергу з огляду на характеристику оброблюваних даних, таких як об'єм, швидкість і різноманітність. Для задоволення різних вимог до проведення аналізу великих даних необхідна поетапна методологія для організації дій і завдань, пов'язаних з придбанням, обробкою, аналізом і повторного використанням даних.

З точки зору впровадження великих даних і перспективного планування важливо, щоб крім життєвого циклу великих даних були враховані питання навчання, обладнання і кадрового забезпечення необхідного для аналітики даних.

Життєвий цикл аналітики великих даних можна розділити на дев'ять етапів:

1. Оцінювання бізнес-ситуації
2. Ідентифікація даних
3. Збір і фільтрація даних
4. Виокремлення даних
5. Перевірка і очищення даних
6. Агрегування і подання даних
7. Аналіз даних
8. Візуалізація даних
9. Використання результатів аналізу

Кожен життєвий цикл аналітики великих даних повинен виходити з чітко визначеної бізнес-ситуації, яка дає чітке уявлення про обґрунтування, мотивацію і цілі проведення аналізу. На етапах оцінювання бізнес-ситуації необхідно скласти економічне обґрунтування, оцінити і затвердити його до початку виконання реальних практичних завдань аналізу.

Оцінюючи бізнес-ситуацію потрібно чітко сформулювати мету для проведення аналізу великих даних, або іншими словами поставити проектне завдання.

Формування проектного завдання повинно включати такі моменти:

– Чітко сформульована мета досліджень.
– Призначення і контекст проекту.
– Попередній опис методики аналізу.
– Ресурси, які ви маєте намір використовувати.
– Доказ практичної можливості бути реалізованим проекту (або можливості перевірки концепції.)
– Пред'являються результати і критерій успіху.
– Календарний план.

На підставі отриманої інформації оцінюються витрати на проект, а також людські та інформаційні ресурси, необхідні для його успішного завершення.

Існує багато методологій проведення аналізу даних, включаючи популярний міжгалузевий стандартний процес обробки даних (CRISP-DM), який використовується більш ніж 40% аналітиків даних. Близько 27% аналітиків даних використовують власну методологію. Решта використовують інші методики. Максимально схожий на науковий метод, життєвий цикл аналізу даних розроблений для використання в бізнес-середовищі. Стрілки спрямовані в обидва боки між деякими кроками. Це підкреслює той факт, що життєвий цикл може зажадати багатьох ітерацій, перш ніж ті, хто приймає рішення, будуть досить впевнені, щоб рухатися вперед.

Рис. 3.2. Життєвий цикл аналізу даних

Як і в науковому методі, життєвий цикл аналізу даних починається з питання. Наприклад, ми можемо задати питання: «Який злочин був найбільш поширеним у Києві, 20 серпня 2015 року?» Кожен крок життєвого циклу аналізу даних включає багато завдань, які необхідно виконати, перш ніж перейти до наступного кроку. Нижче наведено короткий опис кожного кроку.

 Збір даних – процес пошуку даних, визначення, чи є достатньо даних для завершення аналізу. Наприклад, ми шукаємо відкритий набір даних про злочинність для Києва протягом серпня 2015 року.

 Підготовка даних – цей крок може включати багато задач з перетворення даних у формат, відповідний інструменту, який буде використовуватися. Набір даних про злочини вже може бути підготовлений до аналізу. Однак зазвичай можна внести деякі коригування, які допоможуть відповісти на питання.

 Вибір моделі – цей крок включає вибір методики аналізу, яка найкраще відповість на питання із наявними даними. Після вибору моделі вибирається інструмент (або інструменти) для аналізу даних.

 Аналіз даних – процес тестування моделі на даних та визначення надійності моделі та аналізованих даних.

 Представлення результатів – це, як правило, останній крок для аналітиків даних. Це процес донесення результатів до осіб, які приймають рішення. Іноді аналітику даних просять рекомендувати дії. За даними про злочини 20 серпня, гістограма, кругова діаграма або якесь інше представлення можуть бути використані для повідомлення про те, який злочин найбільш поширений. Аналітик може запропонувати посилити присутність поліції у певних районах, щоб стримувати злочинність у конкретний день, наприклад, 20 серпня.

 Прийняття рішень – заключний крок у життєвому циклі аналізу даних. Організаційні лідери включають нові знання як частину загальної стратегії. Процес починається заново зі збору даних.

3. Джерела даних

При зборі даних для подальшого аналізу важливим є визначитися з їх джерелами. Іноді потрібно збирати дані, як то кажуть, з нуля, але в багатьох випадках компанії вже мають певну базу даних, а іншу їх частину часто можна придбати у третіх сторін. Крім того слід мати на увазі що все більше організацій відкривають безкоштовний доступ до високоякісних даних для громадського і комерційного використання.

Перш за все оцінюють актуальність і якість даних, вже зібрані компанією. У багатьох компаніях існують спеціальні програми супроводу ключових даних, так що велика частина роботи з очищення даних може бути вже виконана. Ці дані можуть зберігатися в офіційних сховищах, якими керують IT-професіонали, але може бути і ситуація коли вони зберігаються в файлах Excel на комп'ютерах працівників компанії.

Знайти дані навіть в межах компанії може бути досить складно. В процесі зростання компанії її дані виявляються розсіяними по багатьом місцям. Дані можуть бути розкидані через те, що працівники переходять на інші посади або

йдуть з компанії. Документація і метадані не завжди входять в число пріоритетів керівництва.

Отримання доступу до даних навіть в рамках окремо взятої компанії також може бути складаним завданням. Розуміють цінність і конфіденційність даних в компаніях досить часто встановлюються правила, за яких будь-якому працівнику доступні дані лише необхідні для його роботи. Ці правила перетворюються в фізичні і електронні бар'єри, які носять назву «китайські стіни». У більшості країн, в тому числі і в Україні, такі «стіни» щодо клієнтських даних є обов'язковими і суворо регламентованими.

Для проведення глибокого аналізу даних якими володіє компанія, як правило, недостатньо. Необхідні дані недоступні всередині організації, необхідно віднайти в зовнішньому світі. Багато компаній спеціалізуються на зборі цінної інформації. Наприклад, Nielsen та GFK добре відомі в цьому відношенні в сфері роздрібної торгівлі. Інші компанії надають дані для того, щоб ви, в свою чергу, удосконалювали надані ними послуги і екосистеми. Зокрема, до цієї категорії відносяться Twitter, Facebook.

Хоча деякі компанії вважають дані дорогим ресурсом, в наші дні все більше урядових установ і організацій безкоштовно діляться своїми даними. Це можуть бути досить якісні і правдиві дані в залежності від установи, яка створює їх і керує ними. Надана інформація відноситься до самих різних галузей. Інформація може принести користь як доповнення власних даних компаній, але вона також стане в нагоді тим, хто займається самонавчанням в галузі data science. Нижче в таблиці наведена невелика добірка постачальників відкритих даних, яких з кожним днем стає все більше і більше.

Таблиця 3.9

Постачальники відкритих даних

Сайт з відкритими даними	Опис
Data.gov	Центр відкритих даних уряду США
https://open-data.europa.eu/	Центр відкритих даних Європейської комісії
Data.worldbank.org	Проект відкритих даних всесвітнього банку
Data.gov.ua	Портал відкритих даних Міністерства цифрової трансформації України

4. Збір та підготовка даних

Етап ідентифікації даних, присвячений визначенню наборів даних, необхідних для аналітичних проектів і їх джерел.

Залучення більш широкого спектра джерел даних може збільшити ймовірність виявлення прихованих закономірностей і кореляцій. Наприклад, щоб дати аналітичне висновок, може бити корисно визначити якомога більше типів пов'язаних джерел даних, особливо коли неясно, що саме потрібно шукати.

На етапах збору і фільтрації даних, вони збираються з усіх джерел, які були попередньо ідентифіковані. Потім отримані дані піддаються

автоматизованій фільтрації для видалення пошкоджених або таких, що не мають особою значення для цілей аналізу.

Залежно від типу джерела дані можуть надходити як набір файлів, наприклад, дані, отримані у стороннього постачальника, або можуть вимагати інтеграції API, наприклад, з Twitter. У багатьох випадках деякі або й більшість отриманих даних можуть бути нерелевантними і можуть бути відкинуті в процесі фільтрації.

Дані, що класифіковані як «спотворені», можуть включати записи з відсутніми або безглуздими значеннями або неприпустимо типами даних. дані, відфільтровані для одного аналізу, можуть бити значимість для іншого типу аналізу. Тому рекомендується зберегти точну копію вихідного набору даних перед початком фільтрації.

Необхідно зберегти як внутрішні, так і зовнішні дані після генерування або використання всередині компанії. Для пакетної аналітики ці дані зберігаються на диску перед початком аналізу. У разі аналітики в реальному часі дані спочатку аналізуються, а потім зберігаються на диску.

Деякі дані, ідентифіковані як вхідні дані для аналізу, можуть надходити в форматі, несумісному для роботи з великими даними. Необхідність звертатися до несумісних типів даних більш імовірна при роботі з даними із зовнішніх джерел.

Необхідна ступінь виокремлення і перетворення залежать від типів аналітики і можливостей вирішення для великих даних.

Неправильні дані можуть спотворювати і фальсифікувати результати аналізу. На відміну від традиційних корпоративних даних, де структура даних заздалегідь визначена і дані попередньо перевірений, дані вводяться в аналіз великих даних можуть бити неструктурованих, без будь-яких вказівок на достовірність. Ця складність також може ускладнити отримання набору відповідних обмежень перевірки.

Етап перевірки і очищення даних призначений для створення складних правил перевірки і видалення любих відомих неприпустимо даних.

Рішення для великих даних часто отримують надлишкові дані в різних наборах даних. Ця надмірність може використовуватися для дослідження взаємопов'язаних наборів даних, щоб збирати параметри перевірки і заповнювати відсутні достовірні дані.

Для пакетної аналітики перевірка даних і їх очищення можуть бути виконані за допомогою автономної операції ETL.

Для аналітики в реальному часі потрібно більш складна система внутрішньої пам'яті для перевірки і очищення даних по мірі їх надходження з джерела. Походження може відігравати важливу роль у визначенні точності і якості сумнівних даних. Дані, які здаються неприпустимими, можуть як і раніше мати значимість, оскільки вони можуть приховувати закономірності і тенденції.

Дані можуть бути розподілені за кількома наборами даних, вимагаючи об'єднання наборів даних через загальні поля, наприклад дату або ідентифікатор (ID). В інших випадках одні й ті ж поля даних можуть відображатися в декількох наборах даних, таких як дата народження.

Етап агрегування і представлення даних призначений для інтеграції декількох наборів даних разом для досягнення уніфікованого представлення.

Великі обсяги, оброблювані рішеннями для великих даних, можуть зробити агрегування даних довготривалим і трудомістким. Узгодження цих відмінностей може зажадати складної логіки, яка виконується автоматично без втручання людини.

Тема 4. Методи та інструменти аналізу великих даних

1. Методи аналізу великих даних

2. Програмне забезпечення для аналізу великих даних

3. Платформи великих даних

1. Методи аналізу великих даних

Стандартна бізнес-практика великомасштабного аналізу даних ґрунтується на понятті “корпоративного сховища даних” (*Enterprise Data Warehouse, EDW*), запити до якого надходять від програмного забезпечення “бізнес-аналітики” (*Business Intelligence, BI*). Інструменти BI дають змогу створювати звіти та інтерактивні інтерфейси, узагальнення даних за допомогою агрегатних функцій (наприклад, обчислити кількість або середнє) до різноманітних розподілів ієрархічних даних на групи.

Традиційно вважається, що ретельно спроектоване сховище даних відіграє центральну роль у разі правильного застосування інформаційних технологій. Сховище даних традиційно контролюють спеціально призначені працівники IT, які не тільки супроводжують систему, а й ретельно контролюють доступ до неї, щоб керівні особи могли гарантовано розраховувати на високий рівень обслуговування .

Кількість внутрішньокорпоративних великомасштабних джерел даних істотно зростає: великі бази даних сьогодні виникають навіть на основі єдиного джерела потоків даних про відвідування Web-сайтів (*click-stream*), журналів програмних систем, архівів електронної пошти і форумів тощо. Загальновизнаною стала значущість аналізу даних. Численні компанії демонструють, що складний аналіз даних сприяє зменшенню витрат та навіть прямому зростанню доходів. Результатом цих можливостей є масовий перехід до збирання та використання даних у декількох організаційних одиницях корпорацій.

У цьому змінному кліматі збирання розрізнених великомасштабних даних доцільним є підхід, який називають *могутнім аналізом даних* (МАД;

Magnetic, Agile, Deep (MAD) data analysis). Акронім MAD походить від трьох аспектів цього середовища, що відрізняють його від орто-доксальних сховищ даних, а саме: Магнетична (*magnetic*); Гнучкість (*agile*); Ґрунтовність (*deep*).

Великі дані (*англ. Big data*) – серія підходів, інструментів і методів опрацювання структурованих та неструктурованих даних величезних обсягів і значного різноманіття для отримання зрозумілих для людини результатів, ефективних в умовах безперервного приросту, розподілу по численних вузлах обчислювальної мережі, що сформувалися в кінці 2000-х років, альтернативних традиційним системам управління базами даних і рішень класу Business Intelligence. Є три типи завдань, пов'язаних з Великими даними (Big Data).

1) *зберігання і управління.* Обсяг даних в сотні терабайт або петабайт не дає змоги легко зберігати їх та керувати ними за допомогою традиційних реляційних баз даних;

2) *неструктурована інформація.* Більшість Великих даних неструктуровані;

3) *аналіз Великих даних*. Як аналізувати неструктуровану інформацію? Як на основі Великих даних складати прості звіти, будувати та впроваджувати поглиблені прогностичні моделі?

Робота з Великими даними не схожа на звичайний процес бізнес-аналітики, коли просте додавання відомих значень приносить результат. Працюючи з великими даними, результат одержують, очищаючи їх за допомогою послідовного моделювання: спочатку висувається гіпотеза, будується статистична, візуальна або семантична модель, на її підставі перевіряється достовірність висунутої гіпотези і потім пропонується наступна. Цей процес вимагає від дослідника або інтерпретації візуальних значень, або складання інтерактивних запитів на основі знань, або розроблення адаптивних алгоритмів "машинного навчання", здатних отримати потрібний результат. Причому час життя такого алгоритму може бути доволі коротким.

За інтенсивного розвитку бізнесу для збереження конкурентоспроможності підприємства та опрацювання значних обсягів накопичених структурованих та неструктурованих даних допомогти може інформаційна технологія Великі дані. Актуальним є застосування методів і технологій аналізу Великих даних та інтегрованої платформи для бізнес-аналітики. Метою роботи є дослідження особливостей класифікації методів і технологій аналітики Великих даних з урахуванням означення та особливостей застосування технології Великих даних.

Описання методів і технологій аналітики Великих даних (Big Data Analytics). Формальна модель великих даних як інформаційної технології така:

$$\mathbf{BD} = \langle \mathbf{Vol}\text{BD}, \mathbf{Ip}, \mathbf{A}\text{BD}, \mathbf{T}\text{BD} \rangle,$$

де $\mathbf{Vol}\text{BD}$ – множина типів обсягів; **Ip** – множина типів джерел даних (інформаційних продуктів); $\mathbf{A}\text{BD}$ – множина методик аналізу Великих даних; $\mathbf{T}\text{BD}$ – множина технологій обробки Великих даних. На основі означення Великих даних можна сформулювати основні принципи роботи з такими даними: горизонтальна масштабованість; стійкість до відмов; локальність даних. Усі сучасні засоби роботи з Великими даними так чи інакше відповідають цим трьом принципам. Для того, щоб їх дотримуватися, необхідно придумувати якісь методи, способи і парадигми розроблення засобів опрацювання даних.

Сьогодні наявна множина $\mathbf{A}\text{BD} = \{\mathbf{A}i\}$ різноманітних методик аналізу масивів даних, в основу яких покладено інструментарій, запозичений з статистики та інформатики.

Необхідність у нових засобах для аналізу обґрунтована тим, що даних стає більше, більше їх зовнішніх і внутрішніх джерел, тепер вони складніші та різноманітніші (структуровані, неструктуровані та слабкоструктуровані), використовуються різні схеми індексації (реляційні, багатовимірні, noSQL). Колишні способи опрацювання даних вже неефективні – *Big Data Analytics* поширюється на великі й складні масиви, тому ще використовують терміни *Discovery Analytics* (аналітика, що відкриває) і *Exploratory Analytics* (аналітика, що пояснює).

Сьогодні не розмежовують вживання термінів Big Data і Big Data Analytics. Ці терміни описують як самі дані, так і технології управління та методи аналізу.

Big Data Analytics є розвитком концепції Data Mining. Ті самі завдання, сфери застосування, джерела даних, методи і технології. За роки, що минули з моменту появи концепції Data Mining до настання ери Великих даних, революційно змінилися обсяги даних, що аналізуються, з'явилися системи високопродуктивних обчислень, нові технології, зокрема MapReduce і її численні програмні реалізації. З появою соціальних мереж постали і нові завдання.

Data Mining – це процес підтримки ухвалення рішень, що ґрунтується на пошуку в сирих даних прихованих закономірностей, раніше невідомих, нетривіальних, практично корисних та доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності. Data Mining – це особливий підхід до аналізу даних. Акцент робиться не тільки на добуванні фактів, а й на генерації гіпотез.

Якщо підхід DataMining доповнити технологією MapReduce і вимогою 4V (Volume (обсяг), Velocity (швидкість), Variety (різноманітність), Veracity (достовірність), то це відобразить функціональні зв'язки Big Data

Analytics (рис. 4.1).

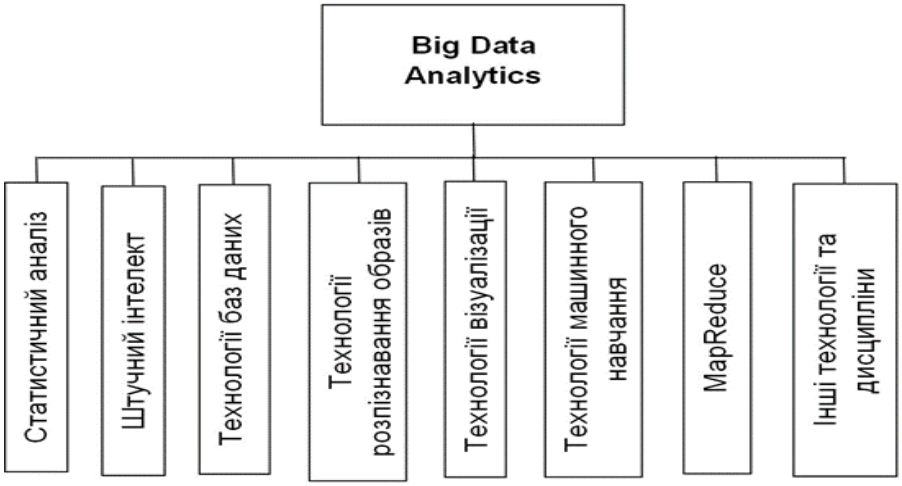


Рис. 4.1. Функціональні зв'язки аналітики Великих даних

Аналіз великих обсягів даних і необхідності зрозуміти значення з індивідуальної поведінки потребує методів оброблення, які виходять за межі традиційних статистичних методів.

Методики і методи аналізу, які застосовують до великих даних, також описано в звіті McKinsey: методи DataMining; краудсорсинг; консолідація та інтеграція даних; машинне навчання; нейронні мережі, мережевий аналіз, оптимізація, зокрема, генетичні алгоритми; розпізнавання образів; аналітика, прогнозування; імітаційне моделювання; просторовий аналіз; статистичний аналіз; візуалізація аналітичних даних.

BothManyika (2011) і Chen (2012) запропонували такий список методів аналітики Великих даних (в алфавітній послідовності): A/B тестування (A/B testing), правило навчання асоціації (Association rule learning), класифікація (Classification), кластерний аналіз (Cluster analysis), злиття і інтеграція даних (Data fusion and data integration), Ансамблі навчання (Ensemble learning), генетичні алгоритми (Genetic algorithms), машинного навчання (Machine learning), обробки природної мови (Natural Language Processing), Нейронні мережі (Neural networks), мережевий аналіз (Network analysis), розпізнавання образів (Pattern recognition), Прогнозне моделювання (Predictive modelling), регресія (Regression), Настроїв аналіз (Sentiment Analysis), Обробка сигналів (Signal Processing), Просторовий аналіз (Spatial analysis), статистика (Statistics), кероване і некероване навчання (Supervised and Unsupervised learning), моделювання (Simulation), аналіз часових рядів та візуалізації (Timeseries analysis and Visualization).

Опишемо групи методів і технологій аналітики Великих даних,

які класифікуються з урахуванням функціональних зв'язків та формальної моделі цієї інформаційної технології, а саме: методи Data Mining, технології Text Mining, технологія MapReduce, візуалізація даних, інші технології та методики аналізу (рис. 4.2).

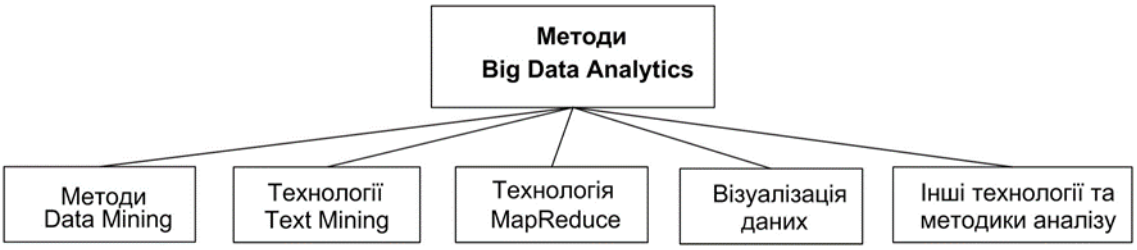


Рис. 4.2. Групи методів аналітики Великих даних

Методи інтелектуального аналізу даних (Data Mining). Застосування методів і технологій Data Mining дає змогу розв'язати такі задачі: класифікація (*Classification*); кластеризація (*Clustering*); асоціація (*Associations*); послідовність (*Sequence*), або послідовна асоціація (*sequential association*); прогнозування (*Forecasting*); визначення відхилень (*Deviation Detection*), аналіз відхилень або викидів; оцінювання (*Estimation*); аналіз зв'язків (*Link Analysis*); візуалізація (*Visualization, Graph Mining*); підбивання підсумків (*Summarization*) – опис конкретних груп об'єктів за допомогою аналізованого набору даних.

Методи Data Mining поділяють на дві групи: навчання з учителем (*Supervised Learning*); навчання без учителя (*Unsupervised Learning*). Інша класифікація поділяє все різноманіття методів Data Mining на дві групи: статистичні й кібернетичні методи. Ця схема поділу ґрунтується на різних підходах щодо навчання математичним моделям.

Опишемо найпридатніші з них для аналізу Великих даних.

Асоціативні правила (*Association Rule Learning*). Набір методик для виявлення взаємо-зв'язків, тобто асоціативних правил, між змінними величинами у великих масивах даних. Для ана-лізу ринкового кошика застосовують **аналіз прихованих закономірностей** (A*ssociation Analysis*).

Класифікація (*Classification*). Набір методик, які дають змогу передбачити поведінку споживачів у певному сегменті ринку (прийняття рішень про покупку, відтік, обсяг споживання тощо).

Метод дерев рішень (*Decision Trees*) є одним з найпопулярніших методів розв'язання завдань класифікації та прогнозування. У найпростішому вигляді дерево рішень – це спосіб подання правил в ієрархічній, послідовній структурі. Метод дерев рішень зазвичай називають "наївним" підходом.

Кластерний аналіз (*Cluster Analysis*). Статистичний метод класифікації об'єктів за групами у результаті виявлення наперед не відомих загальних ознак. Приклад – сегментування ринку.

Для вирішення завдання кластеризації на графах застосовують алгоритм Girvanand Newman методу MLP (Markov Cluster Algorithm).

Для аналізу Великих багатовимірних даних розроблено методологію “Dynamic Quantum Clustering” (DQC), що реалізує парадигму пошуку як “нехай дані говорять про себе самі”. Метод DQC (як і багато інших методів аналітики Великих даних) “працює” без попереднього знання про ті “структури”, їх тип і топології, які можуть бути “приховані” в даних і виявлені в результаті його застосування. Метод добре працює з багатовимірними даними і час аналізу лінійно залежить від розмірності.

Регресія (*Regression*). Набір статистичних методів для виявлення закономірності між зміною залежної змінної та однієї або декількох незалежних.

Аналіз часових рядів (*Time Series Aanalysis*). Набір запозичених зі статистики та цифрової обробки сигналів методів аналізу повторюваних з плином часу послідовностей даних. **Аналіз викидів** (*Outlieran Aalysis*) застосовують для виявлення шахрайства, особистого маркетингу, медичного аналізу.

Машинне навчання (*Machine Learning*). Напрям в інформатиці (історично за ним закріпилася назва “штучний інтеллект”), який має на меті створення алгоритмів самонавчання на основі аналізу емпіричних даних. Машинне навчання сьогодні використовується: для розпізнавання спаму або не спаму повідомлень електронної пошти; для отримання знань про переваги користувача та надання рекомендацій, що ґрунтуються на цій інформації; для визначення кращого контенту для залучення потенційних клієнтів; для встановлення ймовірності виграшу справи та відповідності юридичним нормам пред’явлених рахунків.

Кероване і некероване навчання (*Supervised and Unsupervised Learning*). Набір методик, що ґрунтуються на технологіях машинного навчання, які дають змогу виявити функціональні взаємозв’язки в аналізованих масивах даних. Некероване навчання має спільні риси з кластерним аналізом.

Ансамблі навчання (*Ensemble Learning*). У цьому методі задіється множина предикативних моделей, за рахунок чого поліпшується якість прогнозів.

Еволюційні алгоритми, генетичні алгоритми (*Evolution Analysis, Genetic Algorithms*). Генетичні алгоритми навіяні природою еволюційних процесів – тобто таких механізмів, як успадкування, мутації та природний добір. Ці механізми використовуються для “еволюціонування” корисного вирішення проблем, які потребують оптимізації. У цій методиці можливі рішення подають у вигляді “хромосом”, які можуть комбінуватися і мутувати. Як і в процесі природної еволюції, виживає найпристосованіша особина.

Нейронні мережі (*Neural Networks*) – це клас моделей, що ґрунтуються на аналогії з роботою мозку людини та призначені для розв'язання різноманітних задач аналізу даних після проходження етапу навчання на даних. За допомогою нейронних мереж можна, наприклад, передбачати обсяги продажів, показники фінансового ринку, розпізнавати сигнали, розробляти самонавчальні системи.

Візуалізація даних

Візуалізація (*Visualization*). Методи графічного подання результатів аналізу великих даних у вигляді діаграм або анімації для спрощення інтерпретації, полегшення розуміння отриманих результатів. Візуалізація аналітичних даних – зображення інформації у вигляді рисунків, графіків, схем і діаграм з використанням інтерактивних можливостей та анімації для результатів, а також вихідних даних для подальшого аналізу.

Наочне представлення результатів аналізу Великих даних має принципове значення для їхньої інтерпретації. Сприйняття людини обмежене, і вчені продовжують вести дослідження у галузі вдосконалення сучасних методів подання даних у вигляді зображень, діаграм або анімацій. Новими прогресивними методами візуалізації є: хмара тегів; кластерограма; історичний потік; просторовий потік.

Технології Text Mining. Підґрунтям технології **Text Mining** – статистичний та лінгвістичний аналіз, методи штучного інтелекту. Ця технологія застосовується для проведення аналізу, забезпечення навігації та пошуку в неструктурованих текстах. Застосування інформаційних систем класу Text Mining дає змогу користувачам набувати нових знань.

Технології Text Mining – набір методів, які призначені для видобування відомостей з текстів на основі сучасних ІКТ, що дає змогу виявити закономірності, які забезпечують користувачам отримання корисних даних та нових знань. Основна мета Text Mining – надати аналітику можливість працювати з великими обсягами початкових даних за рахунок автоматизації процесу здобуття потрібних даних.

Основними методами технології Text Mining є: класифікація (*classification*); кластеризація (*clustering*); побудова семантичних мереж або аналіз зв'язків (*Relationship, Event andFact Extraction*); здобуття феноменів, фактів, понять (*feature extraction*); автоматичне реферування, створення анотацій (*summarization*); відповідь на запити (*question answering*); тематичне індексування (*thematic indexing*); пошук за ключовими словами (*keyword searching*); засоби підтримки та створення таксономії (*oftaxonomies*) і тезаурусів (*thesauri*).

Прикладом ефективного застосовування технологій Text Mining є проведення контент-аналізу. **Контент-аналіз** (*Content Analysis*) – це якісно-кількісне, систематичне опрацювання, оцінювання та інтерпретація форми і змісту тексту.

Інші технології та методики досліджень

Опишемо декілька технологій і дисциплін дослідження даних з погляду технології Великих даних.

A/B тестування (*A/B testing, Splittesting*). Методика маркетингового дослідження, в якій контрольна вибірка по черзі порівнюється з іншими. Метод використовується для оптимізації Web-сторінок відповідно до заданої мети.

Обробка природної мови (*Natural Language Processing (NLP*). Набір запозичених з інформатики та лінгвістики методик розпізнавання природної мови людини.

Аналіз настроїв (*Sentiment Analysis*). В основу методик оцінки настроїв споживачів покладено технології розпізнавання природної мови людини. Аналіз настроїв допомагає дослідникам визначити настрої спікерів або авторів щодо теми.

Мережевий аналіз (*Network Analysis*). Набір методик аналізу зв'язків між вузлами в мережах. Стосовно соціальних мереж дає змогу аналізувати взаємозв'язок між окремими користувачами, компаніями, спільнотами тощо.

Оптимізація (*Optimization*). Набір числових методів для редизайну складних систем і процесів для поліпшення одного або декількох показників. Допомагає у прийнятті стратегічних рішень, наприклад, складу виведеної на ринок продуктової лінійки, у проведенні інвестиційного аналізу тощо.

Розпізнавання образів (*Pattern Recognition*). Набір методик з елементами самонавчання для передбачення поведінкової моделі споживачів.

Прогнозне моделювання (*Predictive Modeling*). Набір методик, які дають змогу створити математичну модель наперед заданого ймовірного сценарію розвитку подій.

Обробка сигналів (*Signal Processing*). Запозичений з радіотехніки набір методик, який має на меті розпізнавання сигналу на тлі шуму і його подальшого аналізу.

Просторовий аналіз (*Spatial Analysis*). **Просторовий аналіз** – використання топологічної, геометричної та географічної інформації в даних. Набір частково запозичених зі статистики методик аналізу даних. Джерелом великих даних у цьому випадку є геоінформаційні системи (ГІС).

Статистика (*Statistics*). Наука про збирання, організацію та інтерпретацію даних, зокрема розроблення опитувальників і проведення експериментів. Статистичні методи часто застосовують для оцінкових суджень про взаємозв'язки між тими чи іншими подіями.

Моделювання (*Simulation*). Моделювання поведінки складних систем часто використо-вується для прогнозування, передбачення і опрацювання різних сценаріїв під час планування.

Краудсорсинг (*Crowdsourcing*). Методика збирання даних з великої

кількості джерел. Краудсорсинг – категоризація та збагачення даних силами широкого, невизначеного кола осіб, з метою використання їхніх творчих здібностей, знань і досвіду із застосуванням інформаційно-комунікаційних технологій.

Злиття та інтеграція даних (*Data Fusion and Data Integration*). Набір технік, що дають змогу інтегрувати різнорідні дані з різноманітних джерел інформації для проведення глибинного аналізу. Цей набір методик дає змогу аналізувати коментарі користувачів соціальних мереж і зіставляти з результатами продажів у режимі реального часу.

Технологія MapReduce. Створення і підтримка сховищ даних обсягом в терабайт, петабайт і більше уможливилась завдяки технологіям розподілених файлових систем. Розподілені системи опрацювання даних, замість зберігання даних в одній файловій системі, зберігають та індексують дані на декількох (навіть тисячах) жорстких дисках і серверах. Створюється також “карта” (*map*), на якій міститься інформація про місцезнаходження тих чи інших даних. Однією з найвідоміших систем, що використовують цей підхід, є **Hadoop**. Щоб опрацювати дані в розподіленій файловій системі, необхідно виконувати низькорівневі обчислення, такі як підсумовування, агрегування тощо, в місці їхнього фізичного розміщення в розподіленій файловій системі. Створити карту (*map*) виконаних обчислювальних алгоритмів і відстежувати локальні результати, а потім акумулювати результати (reduced). Цей підхід і шаблон проведення обчислювальних алгоритмів отримав назву **MapReduce** . MapReduce – це фреймворк для обчислення деяких наборів розподілених завдань з використанням великої кількості комп’ютерів (“нод”), що утворюють кластер. Опрацьовуватися можуть дані, які зберігаються або в файловій системі (неструктуровано), або в базі даних (структуровано).

Багато практичних завдань можна реалізувати у цій моделі програмування. Є безліч інструментів для проведення такого агрегування даних у розподіленій файловій системі, що дає змогу легко здійснювати цей аналітичний процес.

Наведений опис методів і технологій аналізу Великих даних дає змогу побудувати онтологію відповідно до підходу METHONTOLOGY, який відображає процес ітеративного проектування. За методологією METHONTOLOGY глосарій термінів містить всі терміни (концепти та їхні екземпляри, атрибути, дії), важливі для аналізу Великих даних, і їхні природно-мовні описи.

Глосарій термінів онтології аналізу Великих даних містить означені вище терміни, які можна семантично розділити на три групи: структура завдання (групи технологій аналітики, зв’язки), дані, що наповнюють задачу (методи, що застосовують для кожної групи), і результати обчислень (рекомендації щодо використання Великих даних для підвищення ефективності

ухвалення рішень). Онтологія аналізу Великих даних розроблена засобами Protégé-OWL.

Великі дані мають вагоме практичне значення як технологія, призначена для вирішення актуальних повсякденних проблем, але породжує ще більше нових. Великі дані здатні змінити наш спосіб життя, праці й мислення.

Однією з умов успішного розвитку світової економіки на сучасному етапі стає можливість фіксувати й аналізувати величезні масиви і потоки інформації. Є думка, що країни, які оволодіють найефективнішими методами роботи з Великими даними, чекає нова індустріальна революція. Напрям “Big Data” концентрує зусилля в організації зберігання, оброблення, аналізу величезних масивів даних.

Міжнародна консалтингова компанія McKinsey, що спеціалізується на розв’язанні задач, пов’язаних зі стратегічним управлінням, виділяє 11 методів і технік аналізу, що застосовуються до великих даних.

Методи классу Data Mining (видобуток даних, інтелектуальний аналіз даних, глибинний аналіз данних) — сукупність методів виявлення у даних раніше невідомих, нетривіальних, практично корисних знань, необхідних для прийняття рішень. До таких методів, зокрема, належать: навчання асоціативним правилам (association rule learning), класифікація (разгалуження на категорії), кластерний аналіз, регресійний аналіз, виявлення і аналіз відхилень тощо.

Краудсорсінг — класифікація і збагачення даних силами широкого, неозначеного кола особистостей, що виконують цю роботу без вступу у трудові стосунки.

Змішання та інтеграція даних (data fusion and integration) — набір технік, що дозволяють інтегрувати різнорідні дані з розмаїття джерел з метою проведення глибинного аналізу (наприклад, цифрова обробка сигналів, обробка природньої мови, включно з тональним аналізом).

Машинне навчання, включаючи навчання з учителем і без учителя – використання моделей, побудованих на базі статистичного аналізучи машинного навчання для отримання комплексних прогнозів на основі базових моделей.

Штучні нейронні мережі, мережевий аналіз, оптимізація, у тому числі генетичні алгоритми (genetic algorithm — евристичні алгоритми пошуку, що використовуються для розв’язання задач оптимізації і моделювання шляхом випадкового підбору, комбінування і варіації потрібних параметрів з використанням механізмів, аналогічних натуральному відбору у природі).

З точки зору обробки в основу технологій Big Data покладені два основних принципи:

– розподіленого зберігання даних;
– розподіленої обробки, з урахуванням локальності даних.

Розподілене зберігання вирішує проблему великого обсягу даних, дозволяючи організовувати сховище з довільного числа окремих простих носіїв. Зберігання може бути організовано з різним ступенем надмірності, забезпечуючи стійкість до збоїв окремих носіїв Розподілена обробка з урахуванням локальності даних означає, що програма обробки доставляється на обчислювач, що знаходиться якомога ближче до оброблюваних даних. Це принципово відрізняється від традиційного підходу, коли обчислювальні потужності і підсистема зберігання розділені і дані повинні бути доставлені на обчислювач. Таким чином, технології Big Data спираються на обчислювальні кластери з безлічі обчислювачів, забезпечених локальної підсистемою зберігання.

Доступ до даних і їх обробка здійснюються спеціальним програмним забезпеченням. Найбільшвідомим і інтенсивно розвиваються проектом в області Big Data є Apache Hadoop. В даний час на ринку інформаційних систем і програмного забезпечення синонімом Big Data є технологія Hadoop, яка представляє собою програмний фреймворк, що дозволяє зберігати і обробляти дані за допомогою комп'ютерних кластерів, використовуючи парадигму MapReduce. Основними складовими платформи Hadoop є:

– відмовостійка розподілена файлова система Hadoop Distributed File System (HDFS), за допомогою якої здійснюється зберігання;

– програмний інтерфейс Map Reduce, який є основою для створення програмного забезпечення, що обробляють великі обсяги структурованих і неструктурованих даних паралельно на кластері, що складається з тисяч машин;

– Apache Hadoop YARN, що виконує функцію управління даними.

Відповідно до підходу MapReduce обробка даних складається з двох кроків: Map і Reduce. На кроці Map виконується попередня обробка даних, яка здійснюється паралельно на різних вузлах кластера.

На кроці Reduce відбувається зведення попередньо оброблених даних в єдиний результат.

В основі моделі роботи Apache Hadoop лежать три основних принципи. По-перше, дані рівномірно розподіляються на внутрішніх дисках безлічі серверів, об'єднаних HDFS.

По-друге, не дані передаються програмі обробки, а програма - до даних. Третій принцип - дані обробляються паралельно, причому ця можливість закладена архітектурно в програмному інтерфейсі Map Reduce. Таким чином, замість звичної концепції «база даних + сервер» у нас є кластер з безлічі недорогих вузлів, кожен з яких є і сховищем, і обробником даних, а саме поняття «база даних» відсутня.

Платформа Hadoop дозволяє скоротити час на обробку і підготовку даних, розширює можливості по аналізу, дозволяє оперувати новою інформацією та неструктурованими даними.

Компанія Oracle розбиває життєвий цикл обробки інформації на три етапи і використовує для кожного з них власне рішення:

1) Збір, обробка та структурування даних.

В якості вирішення застосовується Oracle Big Data Appliance - це встановлений Hadoop-кластер, Oracle NoSQL Database і засоби інтеграції з іншими сховищами даних. Завдання Oracle Big Data Appliance полягає в зберіганні та первинній обробці неструктурованою або частково структурованої інформації, тобто як раз в тому, що у систем на базі Hadoop виходить найкраще.

2) Агрегація і аналіз даних.

Для роботи зі структурованими даними використовується комплекс Oracle Exadata. Модулі інтеграції Oracle Big Data Appliance дозволяють оперативно завантажувати дані в Oracle Exadata, а також отримувати доступ до даних «на льоту» з Oracle Exadata.

3) Аналітика даних в реальному часі.

2. Програмне забезпечення для аналізу великих даних

Процес інформатизації суспільства та економіки загалом, з кожним роком набирає обертів та проникає у всі галузі та сфери життя. Така тенденція є позитивною для розвитку та функціонування постіндустріального суспільства, однак вона наділена специфічними особливостями. «Big Data» – сукупність великої кількості неструктуризованої інформації яка зростає у геометричній прогресії щороку, та значно ускладнює процес пошуку та аналізу необхідної інформації в мережі. Разом з формуванням великих об’ємів інформації з’являється програмне забезпечення, що дозволяє обробляти та класифікувати необхідну інформацію для спеціалістів з маркетингу. Підбір такого програмного забезпечення є незамінним інструментом майбутнього для великого комплексу аналітичних дій на підприємстві та якості отриманих результатів, що уже є 50% успіху при плануванні будь яких стратегій підприємства.

Сучасне бізнес середовище, як ніколи, є досить турбулентним та інформаційно перевантаженим, кількість різноманітних даних з внутрішнього та зовнішнього середовища постійно зростає, стає складнішою та менш структурованою. Існуючі підходи та методи аналізу інформації уже не виконують повноцінно свої функції та стають менш актуальними, виникає потреба пошуку нових можливостей. Найкраще з такими викликами справляються методики Big Data Analytics.

Big Data Analytics – це комплекс методик та підходів, що направлені на акумулювання, систематизацію та обробку великої кількості різної за своїми характеристиками інформації та формуванні на їх основі відповідних висновків та гіпотез. Відповідно функціональні зв’язки Big Data Analytics є досить розгалуженим та включають у себе такі елементи: технології візуалізації, статистичний аналіз, штучний інтелект, технології баз даних, технології розпізнавання образів, об’єднавши які та класифікувавши

можна отримати перелік методів аналітики Великих даних.

До таких методів можна віднести: методи Data Mining, технології Text Mining, технологія MapReduce та візуалізація даних.

Для отримання обґрунтованого управлінського рішення дані проходять крізь послідовність процесів накопичення у сховищі даних, аналітичну обробку та видобуток знань. Для реалізації цього процесу використовують методи та алгоритми, які входять до складу технології Data Mining. Близько 80% роботи над Data Mining полягає в зборі та підготовці даних, що проводиться ще до запуску інструментів видобутку знань. Найбільш поширинеми методами Data Mining є: Базові методи, нечітка логіка, генетичні алгоритми, нейронні мережі.

Технологія MapReduce – модель розподілених обчислень у комп`ютерних кластерах, представлена компанією Googleю. Згідно з цією моделлю, додаток розділяється на значну кількість однакових елементарних завдань, що виконуються на вузлах кластера і потім, природнім шляхом зводяться у кінцевий результат.

Технології Text Mining. Підґрунтям технології Text Mining – статистичний та лінгвістичний аналіз, методи штучного інтелекту. Ця технологія застосовується для проведення аналізу, забезпечення навігації та пошуку в неструктурованих текстах. Застосування інформаційних систем класу Text Mining дає змогу користувачам набувати нових знань.

Візуалізація (*Visualization*). Методи графічного подання результатів аналізу великих даних у вигляді діаграм або анімації для спрощення інтерпретації, полегшення розуміння отриманих результатів. Візуалізація аналітичних даних – зображення інформації у вигляді рисунків, графіків, схем і діаграм з використанням інтерактивних можливостей та анімації для результатів, а також вихідних даних для подальшого аналізу [3, с. 173–210].

Аналізуючи особливості Big Data Analytics, а саме, розгалуженість, багатофункціональність та складність існуючих методик, можна чітко стверджувати, що даний інструментарій є перевантаженим для практичного використання на підприємствах. Бізнес середовище не може витрачати багато часу на усі етапи аналізу, а тому, потребує автоматизації даних процесів. Відповідно на рику почало з’являтись програмне забезпечення, функції якого дають змогу виконувати ряд специфічних функцій, що до обробки великих даних.

Програмні засоби великих даних можна класифікувати за задачами, які вони вирішують. У процес створення рішення повинні бути інтегровані різні засоби для зберігання, управління та аналізу великих даних. Інструменти великих даних відповідно до їх задач включають наступні групи:

ПЗ зберігання великих даних; ·

ПЗ управління великими даними; ·

ПЗ обробки великих даних;

методи та засоби візуалізації великих даних; ·

методи та засоби аналітики великих даних.

Повноцінне програмне забезпечення по обробці великих данних повинен відображати інструменти для зберігання, управління та обробки такої інформації, інструменти та методи аналітики, візуалізації та оцінювання у різні етапи процесу побудови рішення.

Досліджуючи такий ринок програмного забезпечення, необхідно відмітити, що за останні п’ять років він значно збільшився і почав кластеризуватись. В процесі таких згрупувань створилась платформа Business Intelligence. Business Intelligence (BI) – це термін-метафора, який не має дослівного тлумачення та «ієрархічно-синергетичний комплекс автоматизованих засобів нетривіального аналізу первинних даних і візуалізації його результатів для підтримки рішень (Decision Support)».

У бізнес середовищі за допомогою BI можна виконувати такі завдання:

–класифікація споживачів;

–виявлення асоціативних правил у споживчому попиті та їх використання для збільшення продажів; – багатомірний аналіз обсягів продажів, маркетингових затрат засобами OLAP;

–оптимізація асортименту;

–прогнозування обсягів продажів та інших показників за допомогою методу регресійного аналізу; –сегментування ринку за допомогою кластерного аналізу

–оцінка ефективності та оптимізація маркетингових кампаній; – оптимізаційне управління ціновою політикою.

Основними операціями, які проводяться з інформацією в бізнес середовищі це – накопичення, аналіз, та побудова на її основі прогнозів та виявлення тенденцій (Табл. 4.1).

Таблиця 4.1.

Програмне забезпечення, що використовується при роботі з Big Data у бізнес середовищі

Призначення	Продукт
Для збору інформації про внутрішнє та зовнішнє середовище	Marketing Geo, Mapinfo, ArcGI “Infostreamcorporate”, “Stikler” KonSi-Competitive Intelligence&Benchmarking CRM-системи ERP-системи:1С, SAP ERP, Галактика ERP.
Інтеграція данних	ERP Integration, ETL Integration, Portal Inte-gration, CRM Integration, MS Office Applica-tions and Big-Data Connectors
Аналіз данних та їх уніфікація	SAS Business Intelligence, Microsoft BI, IBM Cognos BI, SAP Business Intelligence, Oracle Business Intelligence. Tableau 9.0, Qlik Sense 2.0 і Microsoft Power BI
Візуалізація данних	Visual Querying, Storyboarding, Geospatial Integration, Autocharting and Animations

Аналізуючи вище наведені програми та методи, необхідно відмітити, що усі вони є дієвими та ефективними, та мають місце на практичне застосування,

їхня актуальність для бізнесу є досить індивідуальною та залежить від особливостей підприємства та очікуваних результатів від їхнього використання.

В умовах розбудови постіндустріального суспільства, ключовим благом якого є інформація та інтелект, неможливо вести бізнес без роботи з цифровими даними та різноманітними інформаційними технологіями. Актуальність та вагомість, на сьогодні програмного забезпечення, яке використовується для аналізу «Big Data» є очевидним, оскільки дає змогу ефективно та в короткі строки опрацювати та систематизувати великі дані та значно полегшити роботу при планування, аналізі та прогнозуванні.

3. Платформи великих даних

Платформа великих даних – це інструмент, розроблений постачальниками програмного забезпечення (ПЗ) для управління даними з метою покращення масштабованості, доступності, продуктивності та безпеки організацій, які працюють з великими даними.

Платформа призначена для обробки в режимі реального часу об’ємних багато-структурних даних. Різні користувачі можуть її використовувати для виконання різних задач. Так, наприклад, інженери да-них – для очищення, агрегування та підготовки даних для аналізу, бізнес-користувачі – для запуску запитів, а вчені вважають її корисною при аналізі шаблонів з наборів великих даних за допомогою алгоритмів машинного навчання.

Це платформа інформаційних технологій (ІТ) класу підприємства, яка забезпечує властивості та функціональність прикладної системи в одному рішенні для розробки, розгортання, обробки та управління великими даними. Програмне забезпечення (ПЗ) аналітики великих даних допомагає розкрити приховані шаблони, невідомі кореляції, ринкові тенденції, вподобання клієнтів та іншу корисну інформацію з широкого різноманіття наборів даних.

Головне питання організації роботи з великими даними на корпоративному рівні: обрати реляційну (SQL) чи нереляційну (NoSQL) базу даних? Головною причиною відмови від SQL баз даних (БД) є не правильна робота з самою базою. Більшість компаній не можуть собі дозволити тримати спеціалістів для постійного налагодження баз даних, а для того, щоб розпочати використовувати NoSQL БД не потрібно додаткових розробок. При розробці NoSQL БД особлива увага приділяється забезпеченню високої масштабованості та гнучкості рішень. NoSQL БД – це, перш за все, швидкий доступ до даних, що зберігаються в оперативній пам’яті, гнучкість використання та можливість швидкого розподілення даних між вузлами. Однак можливі такі сценарії, коли дані згодом виходять з-під контроля або вже просто не вміщуються в оперативній пам’яті.

Основні властивості та переваги платформ великих даних

До основних властивостей платформ великих даних можна віднести:

 забезпечення ефективного зберігання та обробки даних, а також їх інтеграції, управління, витягнення, трансформації, та завантаження (*ETL*);

 використання системи Hadoop: забезпечують функції для масового зберігання даних будь-якого типу, величезну потужність обробки та можливість обробляти практично необмежену кількість паралельних задач;

 потокові обчислення: забезпечують функції для *затягування* даних у по-тік, обробки даних та передачі їх назад єдиним потоком;

 функції розвиненої аналітики та машинного навчання;

 функції управління життєвим циклом контенту та документів;

 функції інтеграції великих даних з будь-якого джерела;

 управління даними: містять комплексну систему безпеки, рішення для управління даними та забезпечують дотримання вимог щодо захисту даних.

До головних переваг платформи великих даних та ПЗ аналітики великих даних можна віднести:

 точні дані. Платформа великих даних пропонує точні дані, що сприяє прийняттю правильних рішень. Її аналітичні засоби зменшують ризик отримання недостовірних даних, які виникають внаслідок використання сирих, не проаналізованих даних;

 підвищення ефективності праці. Платформа спрощує отримання джерела необхідної інформації. Пропонує також інформацію, що може стати у нагоді в майбутньому, таким чином, зберігаючи час та підвищуючи ефективність роботи користувачів;

 швидкі відповіді на складні питання. Ефективне управління бізнесом вимагає швидких адекватних відповідей на критичні питання, які впливають на успішність бізнес-операції. Платформа великих даних дозволяє робити це більш надійно. Деякі критичні питання, відповіді на які вимагають тижнів або місяців, за наявності правильного інструменту можуть вирішуватись лише за кілька годин або хвилин;

 безпека даних. Забезпечує безпечну інфраструктуру, яка гарантує безпеку даних.

Задачі та ПЗ великих даних

Програмні засоби великих даних можна класифікувати за задачами, які вони вирішують. У процес створення рішення повинні бути інтегровані різні засоби для зберігання, управління та аналізу великих даних. Інструменти великих да-них відповідно до їх задач включають наступні групи:

 ПЗ зберігання великих даних;

 ПЗ управління великими даними;

 ПЗ обробки великих даних;

 методи та засоби візуалізації великих даних;

 методи та засоби аналітики великих даних.

Таким чином, відповідний фреймворк повинен відображати інструменти для зберігання, управління та обробки великих даних, інструменти та методи

аналітики, візуалізації та оцінювання у різні етапи процесу побудови рішення. Аналітика великих даних може застосовуватись до виявлення знань та обґрунтованого прийняття рішень.

Зберігання та управління великими даними

Традиційні методи структурованого зберігання та витягування даних, такі як реляційні БД, вітрини або SQL сховища даних, мають певні обмеження, які роблять їх не придатними для роботи з великими даними, а саме вони:

- не дозволяють включати нові джерела даних без їх попереднього очищення та інтеграції,
- не дозволяють швидко виробляти та адаптувати дані,
- не забезпечують можливості синхронізації логічного та фізичного вмісту БД з швидкою еволюцією даних,
- не забезпечують поточні потреби аналізу даних.

Необхідність порівняно недорогого зберігання та обробки гігантських обсягів неструктурованої інформації призвела до створення спеціалізованого ПЗ, яке дозволило розподіляти дані за кластерами з сотень та тисяч вузлів, а також обробляти їх у паралельному режимі. Засоби нового покоління для зберігання та управління не структурованими (не реляційними) даними, а саме NoSQL БД, дозволили використовувати репозиторій даних без додаткових розробок, підготовки або на-лагодження, забезпечили високу масштабованість, розподілення даних між вузла-ми та швидкий доступ до даних, що зберігаються в оперативній пам’яті. NoSQL БД дозволяють записувати задачі управління даними на прикладному рівні. Кожна база, в даному випадку, є колекцією незалежних документів, де кожний документ підтримує власні дані та схеми та може мати метадані – оглядову інформацію про дані документа. Прикладна про-грама може мати доступ до багатьох БД, розташованих у різних місцях.

Нові вимоги до зберігання, управління та обробки даних обумовили виникнення Hadoop – фреймворка з відкритим кодом під крилом Apache Sotware Foundati-системи на базі відносно недорогого об-ладнання масового попиту. З часом Hadoop був розширений набором бібліотек та утиліт, та сформував навколо себе екосистему проектів з розподіленої обробки даних. Розглянемо його більш детально.

Apache Hadoop фреймворк забезпечує розподілене зберігання та обробку дуже великих наборів даних на комп’ютерних кластерах з промислового комп’ютерного обладнання. Тобто, замість того, щоб використовувати один великий комп’ютер, Hadoop дозволяє кластеризувати апаратне забезпечення для паралельного виконання аналізу масивних наборів даних. Сервіси Hadoop забезпечують виконання наступних функцій:

- зберігання даних;
- обробка даних;
- доступ до даних;
- управління даними;

 безпека та операції з даними.

Екосистема Hadoop складається з багатьох модулів (процедур, бібліотек та властивостей), які розглядаються як частини фреймворка. Кожний модуль виконує певну задачу, необхідну для виконання аналітики великих даних. Ядро Hadoop складається з двох основних модулів: розподіленої файлової системи Hadoop Distributed File System (HDFS) та базового інструменту для обробки даних MapReduce та підтримується майже всіма відомими постачальниками систем великих даних. Також до найбільш використовуваних від-носять планувальник завдань та управлін-ня кластерами YARN і множину загальних утиліт Hadoop Common.

Розподілена файлова система. HDFS дозволяє зберігати дані у простому доступному форматі. Це досягається завдяки використанню великої кількості пов’язаних пристроїв зберігання даних та механізму MapReduce для їх обробки.

"Файлова система" є методом, що застосовується комп’ютером для зберігання даних таким чином, щоб їх можна було знаходити та використовувати. Зазвичай, він визначається операційною системою (ОС) комп’ютера, але система Hadoop використовує власну файлову систему, яка надбудовується "над" файловою системою хост-комп’ютера. Це означає, що доступ до даних можна отримати з будь-якого комп’ютера, на якому встановлена будь-яка підтримувана ОС.

Hadoop розділяє файли на великі блоки та розподіляє їх між вузлами у кластері. Потім він передає пакетований код у вузли для обробки даних у паралельно-му режимі. В даному підході вузли маніпулюють даними, до яких вони мають доступ. Це дозволяє обробляти набір да-них швидше та ефективніше, ніж в більш традиційній архітектурі суперкомп’ютера, яка спирається на паралельну файлову систему, де обчислення та дані розподіляються у високошвидкісній мережі.

HDFS є розподіленою, масштабованою та портативною файловою системою, що написана на Java для Hadoop фреймворк. Вона забезпечує виконання команд та Java інтерфейсів (API), подібних до інших файлових систем, для зв'язку використовує протокол TCP/IP. Клієнти для спілкування один з одним використовують виклики віддаленої процедури (RPC). Надійність зберігання даних досягається шляхом реплікації між декількома хостами. Щоб зменшити трафік у мережі, Hadoop необхідно знати, які сервери є най-ближчими до даних або інформації, яка може забезпечити встановлення мостів з HDFS.

Hadoop може працювати безпосередньо з будь-якою розподіленою файловою системою, яка може бути встановлена основною операційною системою.

Прикладами файлових систем, що підтримуються Hadoop (окрім HDFS), є:

 FTP (зберігає всі свої дані на віддалених FTP-серверах);

 сховище об’єктів Amazon S3 (Simple Storage Service), орієнтоване на кластери, які розміщені на інфраструктурі Amazon Elastic Compute Cloud типу сервер-на-запит;

 Windows Azure Storage Blobs (WASB), розширення HDFS, яке дозволяє розподілення Hadoop для доступу до даних в Azure блог-сховищах без постійного пе-реміщення даних у кластер.

Існують також файлові системи, які не розповсюджуються разом з Hadoop, але постачаються як альтернативні, що використовуються за замовченням, з де-якими його комерційними рішеннями. Наприклад: IBM General Parallel File System, Parascale, Appistry (драйвер файлової системи Hadoop для використання з CloudIQ Storage), драйвер файлової системи IBRIX Fusion, альтернативна файлова система MapR FS, що заміщує HDFS системою повністю випадкового доступу для читан-ня/запису файлів.

Модуль MapReduce названий за двома головними операціями, які він виконує, а саме: читання даних з БД і переведення їх у формат, що підходить для аналізу (map), та виконання математичних операцій (reduce).

Функціонування MapReduce забезпечується двома компонентами: JobTracker та TaskTracker. Клієнтські прикладні програми направляють завдання MapReduce до JobTracker, JobTracker працює з доступними у кластері вузлами TaskTracker, щоб наблизитись до потрібних для виконання цих завдань даних. JobTracker відомо, які вузли містять дані, та які інші комп’ютери є поруч (рис. 4.3).

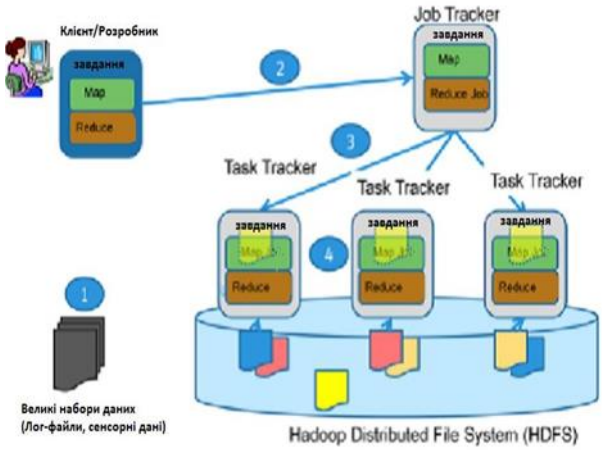


Рис. 4.3. Робота Map Reduce модуля

❶ Дуже великий набір даних. HDFS зберігає репліки даних у вузлах даних.

❷ Клієнт виконує Map та Reduce завдання на конкретному наборі даних та відсилає їх JobTracker.

❸ JobTracker розподіляє завдання серед TaskTracker. TaskTracker запускає механізм відображення (map), результат роботи якого зберігається у HDFS.

❹ Запускається завдання Reduce на даних, що вже оброблені завданням Map

Якщо робота не може бути виконана на тому вузлі, де розміщені дані, пере-вага надається вузлам, що розміщені на тій самій стойці. Таким чином, скоро-чується трафік на головній магістральній мережі. Якщо TaskTracker виходить з ладу або зазнає затримку, здійснюється пере-планування частини завдань. TaskTracker в кожному вузлі породжує окремий процес віртуальної машини Java (JVM). Це дозволяє запобігти виходу з ладу TaskTracker, якщо запущене на виконання завдання зруйнує свою JVM. Кожні кілька хвилин з TaskTracker до JobTracker над-силається імпульс, щоб перевірити його стан. Стани JobTracker і TaskTracker та інформація про їх роботу відображаються у контейнері сервлетів Jetty та її можна також переглядати у веб-браузері.

Слід зазначити, що даний підхід має певні обмеження:

 Алгоритм розподілення роботи TaskTracker є дуже простим. Кожний TaskTracker має множину наявних слотів. Кожна активна задача займає один слот. JobTracker розподіляє роботу на TaskTracker з наявним слотом, найближчий до даних. При цьому не приймається до уваги поточна завантаженість системи призначеної машини – її реальна доступність.

 Якщо один TaskTracker дуже повільний, це може затримати роботу MapReduce в цілому. Однак, коли дозволе-не паралельне виконання, окрема задача може виконуватися на декількох підпорядкованих вузлах.

В первинному варіанті, для впоряд-кування завдань з робочої черги, Hadoop підтримує First-In-First-Out (FIFO) планування та опціональне планування пріоритетів, які використовуються за замовченням. Згодом до планувальника завдань бу ла додана можливість використовувати альтернативні планувальники, такі як Fair (Facebook AI Research) або Capacity. Планувальник Fair є розробкою Facebook. Розробники мали за мету забезпечити швидку відповідь для невеликих завдань та якість сервісу для виробничих завдань. Завдання групуються у пули і ресурси роз-поділяються між цими пулами. За замовченням для кожного користувача є окремий пул, так що кожний користувач от-римує рівну частку кластера. На відміну від планувальника Hadoop, що використовується за замовченням та формує чергу завдань, Fair дозволяє коротким завданням завершуватись у розумний час, не очікуючи довго своєї черги. Це також є простим способом спільного використання кластера між кількома користувачами, яке також може працювати з пріоритетами завдань. Ці пріоритети використовуються як ваги для визначення частки загального часу обчислення, яке отримує кожне завдання.

Планувальник Capacity, розроблений Yahoo, підтримує декілька властивостей, подібних властивостям Fair, а саме: кожній черзі виділяється частка загального ресурсу, вільні ресурси виділяються чергам за їх потужністю.

У черзі завдання з більшим пріоритетом мають пріорітетніший доступ до її ресурсів.

HDFS не обмежується MapReduce завданнями. Вона може працювати з інши-ми прикладними програмами, багато з яких є розробками Apache, наприклад, база даних HBase, система машинного навчання Apache Mahout, система Apache Hive Data Warehouse. Теоретично, Hadoop може використовуватися для будь якого типу завдань, які є швидше пакетно-орієнтованими, ніж тими, що виконуються у реальному часі, а також для завдань з дуже інтенсивними даними і завдань, для яких корисна паралельна обробка даних. Hadoop також може використовуватися для доповнення системи реального часу, такої як, наприклад, лямбда архітектура, Apache Storm, Flink або Spark.

Hadoop Common та планувальник Yarn. Hadoop Common забезпечує інструменти, які дозволяють комп'ютерним системам користувача читати дані, що зберіга-ються у файловій системі Hadoop.

YARN керує ресурсами систем, які зберігають дані та виконують їх аналіз.

Використання Hadoop. Hadoop також містить безліч інших інструментів з відкритим кодом, призначених для створення додаткових функцій на компонентах ядра Hadoop.

Так, Apache Tez є фреймворком наступного покоління, який може використовуватися замість Hadoop MapReduce, в якості двигуна. Amazon EMR включає конектор EMRFS, який дозволяє Hadoop використовувати для зберігання даних сховище Amazon S3. Amazon EMR також може використовуватися для легкого встановлення та налаштування у кластері таких інструментів, як Hive, Pig, Hue, Ganglia, Oozie та HBase. Окрім Hadoop на Amazon EMR можна запускати інші фреймворки, такі як Apache Spark для обробки даних у пам'яті або Presto для виконання інтерактивних запитів.

Гнучка природа системи Hadoop дозволяє компаніям, коли вони потребують змін, додавати або змінювати власну систему даних, використовуючи дешеві та легко-доступні частини від будь-яких постачальників інформаційних систем. Сьогодні Hadoop є найбільш використовуваною системою для зберігання та обробки даних на виробничому апаратному забезпеченні. Hadoop використовують майже всі великі постачальники он-лайн продуктів, та кожний має можливість його вільно модифікувати відповідно до своїх цілей. Ці зміни, які вносять до ПЗ експерти, наприклад, Amazon чи Google, відсилаються до спільноти розробників, де вони часто використовуються в подальшому для вдосконалення "офіційного" продукту. Така форма колаборативної розробки є ключовою властивістю ПЗ з відкритим кодом.

Слід зазначити, що використання базових модулів Hadoop Apache є складним навіть для фахівця галузі інформацій-них технологій, тому були

розроблені комерційні версії продукту такі, як наприклад, Cloudera, що спрощують задачу інсталяції та запуску Hadoop, а також пропонують послуги навчання та підтримки. Завдяки гнучкій природі Hadoop, компанії при розширенні бізнесу мають можливість корегувати та розширювати операції аналізу даних. Підтримка спільноти відкритого коду робить аналіз великих даних доступним для кожного.

Apache Spark – фреймворк (містить більш ніж 80 операторів для роботи з даними) з відкритим кодом, який був створений для розподіленої обробки великих даних. На відміну від класичного обробника з ядра Hadoop, що реалізує дворівневу концепцію MapReduce з дисковим сховищем, він використовує спеціалізовані примитиви для рекурентної обробки в оперативній пам’яті, завдяки чому дозволяє отримувати значне прискорення роботи для деяких класів задач. Зокрема, можливість багатократного доступу до даних користувача, що завантажені в оперативну пам’ять, робить бібліотеку дуже привабливою для алгоритмів машинного навчання. Фактично, Spark є переосмисленим MapReduce, але працює у 10-100 разів швидше, залежно від того, працює він в пам’яті або на диску. Spark підтримує мови програмування Scala, Python, Java, R.

Головним поняттям у Spark є Resili-ent Distributed Dataset (RDD) – це розподілена структура даних, яка розміщується в оперативній пам’яті (рис. 4.4). Кожний RDD є фрагментом даних, що розподілені по вузлах кластера. RDD є незмінними структурами, тому після виконання перетворень створюються нові RDD. RDD обробляються паралельно за допомогою трансформацій/дій, які виконуються одночасно во всіх розділах (partition). RDD є відмовостійкими: якщо розділ втрачається в результаті відмови вузла, він може бути відновлений з вихідних джерел.

Рис. 4.4. Розподілення RDD

Фактично RDD являє собою набір даних, над яким можна виконувати пере-творення двох типів: трансформації та дії. Відповідно, вся робота з цими структурами полягає у послідовності цих перетворень.

Трансформація. Як правило, перетворює якимось чином елементи даного на-бору даних. Результатом застосування її до RDD є новий RDD. Далі наведений неповний перелік найрозповсюдженіших транс-формацій, кожна з яких повертає новий RDD (рис. 4.4):

 map(f) – застосовує функцію f до кожного елемента набора даних;

 filter(f) – повертає всі елементи набору даних, на яких функція f повернула істинне значення;

 distinct([numTasks]) – повертає набір даних, який містить унікальні елементи вихідного набору даних;

Також підтримуються наступні операції над множинами:

 union(Dataset) – об'єднання з на-бором даних Dataset,

 intersection(Dataset) – перетин з набором даних Dataset,

 cartesian(Dataset) – результатом операції є новий набір даних, який містить пари (A,B), де A належить вихідному набору даних, а B – набору даних Dataset.

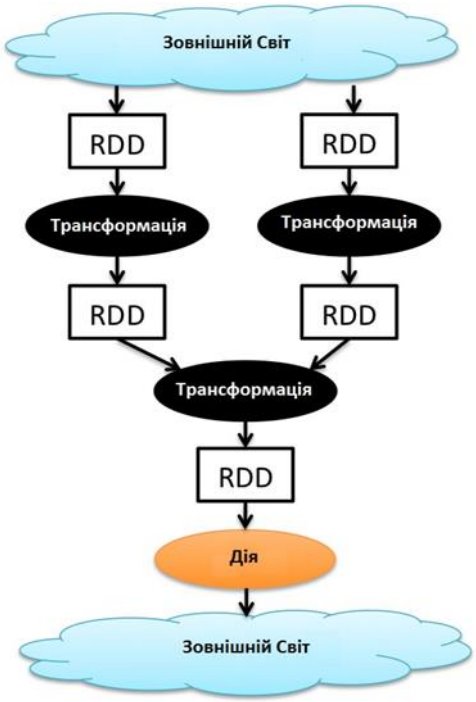


Рис. 4.5.Робота з RDD

Дії. Застосовуються, коли необхідно матеріалізувати результат – як правило, зберегти дані на диску, або вивести частину даних у консоль. Найбільш розповсюдженими діями, які можна застосувати до RDD, є:

 saveAsTextFile(path) – зберігає дані у текстовому файлі (hdfs) на локальну машину або у будь-яку іншу файлову систему, яка підтримується, path визначає шлях для збереження файлу;

 collect() – повертає елементи набора даних у вигляді масива. Як правило, використовується після застосування до набору даних фільтрів та перетворень для візуалізації або додаткового аналізу результату;

 take(n) – повертає у вигляді масива перші n елементів набору даних;

 count() – повертає кількість елементів у наборі даних;

 reduce(f). Функція f (приймає на вхід 2 аргументи, повертає одне значення) повинна обов'язково бути комутативною та асоціативною.

Spark не змушує думати в парадигмі MapReduce, а дозволяє створювати зрозумілий код, який спрямований саме на виконання поставленої бізнес-задачі.

Фреймворк бере на себе розподілення та фрагментацію кода та даних, які автоматично передаються на кластер.

Spark має також колекцію бібліотек (набір готових алгоритмів, підходів та практик), що дозволяють комбінувати існуючі рішення в межах одного програмного кода для досягнення поставленої мети. На цей час Spark містить наступні бібліотеки:

- Spark SQL;
- Spark Streaming (аналіз у реальному часі);
- MLib (машинне навчання);
- GraphX (робота з графами).

Spark SQL. Це модуль Apache Spark, який є частиною ядра Spark та інтегрує реляційну обробку даних та процедурний API Spark. Він може працювати ра-зом з Hive (HiveQL/ SQL) або його заміщу- вати. Окрім цього, модуль здатний взаємодіяти з інструментами бізнес-аналітики.

Spark SQL підтримує реляційну обробку даних як в межах програм Spark (через RDD), так і з зовнішніх джерел даних. Він може взаємодіяти з новими джерелами даних, включаючи слабоструктуровані дані та зовнішні бази даних, що підтримують федеративні запити.

Spark SQL реалізує та оптимізує реляційну обробку, підтримуючи наступні підходи:

- перетворення даних у більш ефективні формати (з точки зору сховища, мережі та операцій введення/ виведення), зокрема, в різні формати, що орієнтовані на стовбці (columnar format);
- розбиття даних на секції;
- зменшення кількості операцій читання на основі статистики;
- оптимізація операцій над даними;
- виконання оптимізації наскільки можливо пізніше, коли доступна вся інформація по конвейєрах даних.

Spark SQL використовує оптимізатор запитів Catalyst для інтелектуального планування запитів.

Spark SQL може підтримувати пакетний та потоковий SQL. Ядро Spark забезпечує обробку пакетних навантажень через RDD. RDD можуть посилатися на статичні набори даних, а за допомогою розвиненого API Spark можна маніпулювати RDD в оперативній пам’яті із застосуванням «ледачих» обчислень.

Spark Streaming. Реалізує абстракцію DStream (discretized stream, дискретизований потік), що являє собою безперервний потік даних. DStream може бути створений з потоку вихідних даних; на основі таких джерел, як Kafka або Flume, або за допомогою виконання операцій з іншими DStream. По суті, DStream є послідовністю RDD (рис. 4.6).

Рис. 4.6. Структура DStream

RDD, що створений за допомогою DStream, можна перетворювати у Data-Frame та виконувати SQL запити до нього. Доступ до потоку може надаватися для будь-якої зовнішньої прикладної програми, що підтримує SQL, за допомогою JDBC-драйвера. Пакети потокових даних зберігаються у пам'яті вузла. До цих даних можна будувати інтерактивні запити, використовуючи SQL або API Spark. Для вико-нання SQL-запитів до Dstream використо-вується StreamSQL, що поєднує Spark Streaming з Catalyst. StreamSQL є розши-ренням SQL, яке додатково забезпечує під-тримку наступних потокових операцій:

- виборка (SELECT) з потоку для обчислення функцій або фільтрації даних (за допомогою умови WHERE);
- з'єднання (JOIN) потоку з одним або декількома наборами даних для створення нового потоку;
- застосування віконних функцій та виконання агрегацій. Потік можна налаштувати таким чином, щоб він створював набори даних обмеженого розміру. За допомогою віконних функцій можна виконувати складний відбір повідомлень на основі значень полів. Після створення обмеженого пакета можна виконувати аналітику на ньому.

В основу підходу для реалізації аналітики реального часу покладено лямбда-архітектуру, що застосовується для створення аналітичних систем реального часу в контексті великих потокових даних (рис. 4.7).

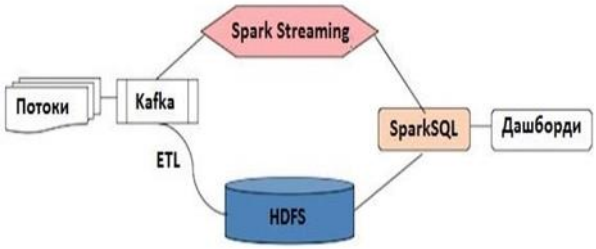


Рис. 4.7. Логічна схема реалізації аналітики великих даних в реальному часі за допомогою Spark SQL

MLib. Бібліотека для машинного навчання. Її метою є зробити машинне навчання масштабованим та простим. Вона містить розповсюджені алгоритми і утіліти машинного навчання, та дозволяє розпаралелювати на кластері алгоритми машинно-го навчання (класифікація, регресія, кластеризація і т. і.) лише за пару строк коду. Окрім цього, SparkMLib якісно працює з локальними даними, використовуючи пакет лінійної алгебри Breeze. MLib має добре

продуманий API, працює з даними у будь-якому форматі на базі Hadoop та не потребує попереднього встановлення.

GraphX. Розподілений фреймворк обробки графів на основі Apache Spark. Графи є наявною та простою для розумін-ня моделлю даних. Розподілені обчислення кардинально спростили зберігання та обробку графів.

Головним механізмом ітерації графа в GraphX є розроблений Google алгоритм Pregel. Головною ідеєю цього алгоритму є передача повідомлень між вузлами у графі, які називають супер кроками завдяки послідовності ітерацій. Ітерація часто формується як ”think like a vertex”, тобто стан поточного вузла залежить лише від стану його сусідів. Використання Pregel є особливо доцільним, коли задачу складно вирішити за допомогою звичайного MapReduce.

Головним примитивом для обходу графа у GraphX є триплет: поточний вузол, вузол, до якого здійснюється перехід, та ребро між ними. Pregel вимагає визначення відстані між вузлами за замовченням, як правило, це PositiveInfinity – UDF (user defined function) функція для кожного вуз-ла, що дозволяє обробити вхідне повідомлення та порахувати наступний вузол, а також UDF функції для злиття двох вхідних повідомлень. Ці функції повинні бути комутативними та асоціативними.

GraphX містить статичну та динамічну версії реалізації алгоритму PageRank, який для кожного вузла графа призначає вагу серед решти вузлів. Наприклад, якщо користувач Твітера має велику кількість підписок від інших користувачів, то він буде мати високий рейтинг, тобто, його можна буде легко знайти у пошуковій системі. Статична версія має фіксовану кількість ітерацій, тоді як динамічна версія буде працювати доки рейтинг не почне зходитися до заданого значення.

Через те що GraphX побудований на основі незмінних RDD, графи теж незмінні, тому GraphX непридатний для роботи з графами, які оновлюються, тим більше транзакціями, як у графових БД.

GraphX надає два окремі API для реалізації масово паралельних алгоритмів (таких як PageRank): Pregel-подібний та більш загальний — MapReduce API.

Основні типи NoSQL сховищ

На сьогодні виділяють чотири основних типи NoSQL сховищ:

- *сховище «ключ-значення».* В ньому є велика хеш-таблиця, що містить ключі та значення. (Приклади: Riak, Amazon DynamoDB);
- *документоорієнтоване сховище.* Зберігає документи, які складаються з тегованих елементів. (Приклад: CouchDB);
- *стовпчикове сховище.* У кожному блоці зберігаються дані лише з однієї колонки. (Приклади: HBase, Cassandra);
- *сховище на основі графів.* Мережеве сховище, яке використовує вузли та ребра для відображення та зберігання да-них. (Приклад: Neo4J).

Сховище типу «ключ-значення». Відсутність схеми у сховищах типу «ключ-значення» є важливою перевагою для зберігання великих даних. Ключ може бути синтетичним або автосгенерованим, а значення може бути представлено строкою, JSON, блобом (BLOB, Binary Large Object) тощо. Такі сховища, як правило, використовують хеш-таблицю, яка містить унікальний ключ та посилання на певний об'єкт даних. Існує поняття блока – логіч-ної групи ключів, що фізично не поєднують дані у групи. У різних блоках можуть бути ідентичні ключи.

Продуктивність обробки даних значно збільшується за рахунок механизмів хешування, що працюють на основі маппінгів. Щоб прочитати значення, необхідно знати ключ та блок, оскільки насправді ключ є хешем (блок + ключ).

Модель «ключ-значення» проста в реалізації. Такі сховища є доступними та толерантними до розділення, але явно програють у питаннях погодженості даних. В якості недоліків сховищ типу «ключ-значення» можна зазначити:

 модель не надає стандартних можливостей баз даних таких, як атомарність транзакцій або погодженість даних при одночасному виконанні декількох транзакцій. Такі можливості повинні надаватися самою прикладною програмою.

 при збільшенні об'ємів даних, підтримка унікальних ключів може стати проблемою. Для її вирішення необхідно якось ускладнити процес генерації строк, щоб вони залишалися унікальними серед дуже великої множини ключей.

Документоорієнтоване сховище. Дані, які представлені парами ключ-значення, стискаються як сховище документів, що є схожим зі сховищем «ключ-значення». Але на відміну від сховища «ключ-значення», документи, які зберігаються, мають визначену структуру та кодування даних. Деякі зі стандартних розповсюджених кодировок, що використовуються – це XML, JSON та BSON.

Однією з ключових відмінностей між сховищами «ключ-значення» та документоорієнтованим є те, що останнє включає метадані, які пов'язані зі вмістом, що зберігається. Це надає можливість робити запити на основі цього вмісту. Найпопулярнішими прикладами документоорієнтованих сховищ є CouchDB та MongoDB. CouchDB використовує JSON для зберігання даних, JavaScript в якості мови запитів з використанням MapReduce та HTTP для API. Дані та відношення не зберігаються в таблицях так, як це відбувається у традиційних реляційних БД, а за сутністю є набором незалежних документів. Той факт, що такі сховища працюють без схеми, спрощує задачу додавання полів до JSON-документа без необхідності попереднього заявлення про зміни.

Стовпчикове сховище. У стовпчикових NoSQL сховищах дані зберігаються у комірках, що згруповані у стовпчики, а не строки даних. Стовпчики логічно групуються у стовпчикові сімейства. Стовпчикові сімейства

можуть складатися з практично необмеженої кількості стовпчиків, які можуть створюватися під час роботи програми або під час визначення схеми. Читання та запис відбувається із використанням стовпчиків, а не строк.

Порівняно зі зберіганням даних у строках, як у більшості реляційних БД, переваги зберігання у стовпчиках полягає у швидкому пошуці /доступі та агрегації даних. Реляційні БД зберігають кожну строку як безперервний запис на диску. Різні строки зберігаються у різних місцях на диску, в той час як стовпчикові сховища зберігають всі комірки, що відносяться до стовпчика, як безперервний запис, що прискорює пошук/доступ.

Стовпчикові сховища використовують наступну модель даних:

 стовпчикове сімейство – структура, яка може легко групувати колонки та суперколонки;

 ключ – постійне ім’я запису. У ключів може бути різна кількість стовпчиків, тому сховище може розширюватися нерівномірно;

 простір ключів – визначає зов-нішній рівень організації, як правило, ім’я прикладної програми/БД.

 стовпчик – має впорядкований список елементів – кортежів з іменами та значеннями.

Найвідомішими представниками стовпчикових сховищ є Google BigTable та HBase з Cassandra.

BigTable є високопродуктивним, стислим та пропрієтарним сховищем даних від Google. Воно має наступні атрибути:

 розрідженість – деякі комірки можуть бути порожніми;

 розподіленість – дані розділені між багатьма вузлами;

 постійність – дані зберігаються на диску;

 багатомірність – більш ніж одне вимірювання;

 співставлення – ключ та значення;

 відсортованість.

На стовпчики можна посилатися за допомогою стовпчикового сімейства.

Графове сховище. У графовому сховищі немає строгого формату SQL або представлення таблиць та стовпчиків, замість цього використовується гнучке графічне представлення, яке ідеально підходить для вирішення проблем масштабованості. Графові структури використовуються разом із ребрами, вузлами та властивостями, що забезпечує безіндексну суміжність. При використанні графового сховища дані можуть бути легко перетворені з однієї моделі в іншу (рис. 4.8).

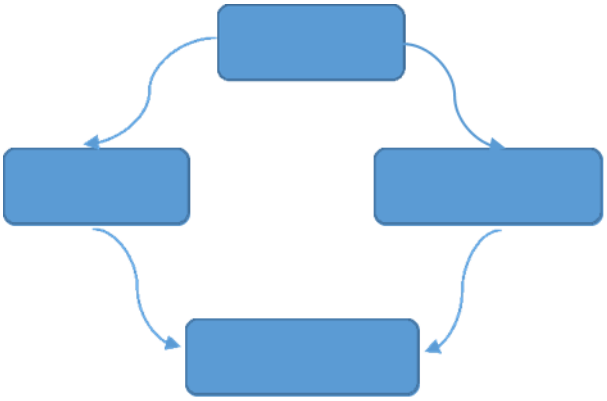


Рис. 4.8. Принципи використання графової моделі

 Такі сховища використовують ребра та вузли для представлення даних.

 Вузли пов’язані між собою певними відношеннями, які представлені реб-рами між ними.

 Вузли та відношення мають деякі властивості.

Розмічений, спрямований, атрибутований мультиграф (рис. 4.9) – це граф, який містить вузли, які помічені певними властивостями та які мають зв’язки один з одним, що представлені спрямованими ребрами. Наприклад, зв’язок «Аліса знає Боба» виражена ребром з відповідними властивостями. Будь-який рейтинг «вам рекомендовано», представлений на різних сайтах, часто вираховується виходячи з того, як інші користувачі оцінили продукт. Графові БД відмінно підходять для вирішення такого типу задач.

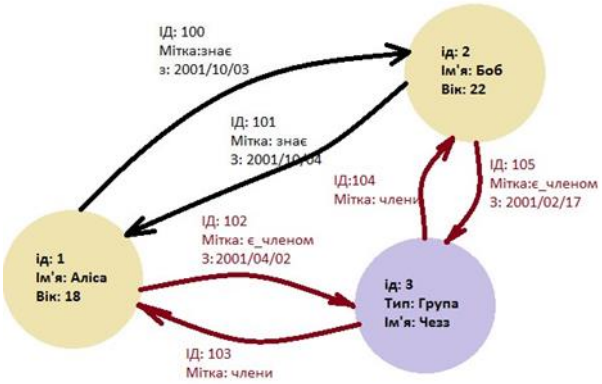


Рис. 4.9. Приклад атрибутованого мультиграфу

Прикладами найпопулярніших графових сховищ є InfoGrid та Infinite Graph. InfoGrid дозволяє з’єднувати множини ребер та вузлів, що спрощує представлення набора інформації зі складними взаємними посиланнями. InfoGrid пропонує два типи сховищ:

 MeshBase — підходить для автономного розгортання;

 NetMeshBase — підійде для великих розподілених графів та має додаткові можливості для взаємодії з іншими подібними сховищами.

Постачальники та ПЗ великих даних

В області розробки підходів до ро-боти з пам’яттю найбільшим чемпіоном є SAP зі своєю Hana платформою, але, слід зазначити, що зараз Microsoft та Oracle також вводять спеціальні опції для роботи з пам’яттю для своїх провідних баз даних. Постачальники ПЗ, що фокусуються на використанні аналітичних БД, включаючи Actian, HP Vertica та Teradata ввели спеціальні опції для співвідношення високих-ОЗУ-дисків, а також пропонують інструменти для розміщення конкретних даних у пам’яті для виконання ультра-швидкого аналізу. Прогрес, що має місце у зростанні пропускної здатності та обчислювальної потужності, вдосконалив й можливості потокової обробки та проведення аналізу в реальному часі.

До великих постачальників сервісів обробки даних можна віднести IBM, Microsoft, Oracle, SAP, які пропонують все від ПЗ інтеграції даних та систем керування базами даних (DBMS) до ПЗ для бізнес-аналізу та аналітичної обробки, а також Hadoop опцій для роботи з пам’яттю та потокової обробки. Teradata більш вузько зосереджений на керуванні даними, та подібно Pivotal, він має тісні зв’язки з лідером аналітичного ринку SAS.

Багато постачальників пропонують реалізовані окремі опції хмарних технологій, але такі розробники, як 1010data та Amazon Web Services (AWS) використовують хмарну модель у повному обсязі в своєму ПЗ. З них двох Amazon має найширшу вибірку продуктів і є очевидним вибором для тих, хто працює з великими навантаженнями та зберігає велику кількість даних на AWS платформі. 1010data має високо масштабований сервіс БД та підтримує можливості управління інформацією, бізнес-аналіз та аналітику, що обслуговуються в стилі приватної хмари.

Hadoop довів свою користь та переваги у вартості там, де є екстремальними об’єм та різноманітність даних. На сього-дні це найбільш відомий та поширений програмно-апаратний комплекс для роботи з великими даними. Він виявився на-стільки гарним, що став фундаментом декількох комерційних реалізацій на його основі, а саме: Cloudera, MapR та Horton-works, кожна з яких пропонує власний дистрибутив. На сьогодні всі постачальники традиційних BI-систем, як Micro-Strategy або SAS, забезпечують інтерфейс з Hadoop. Виробники MPP-систем (масово-паралельних архітектур) у свою чергу забезпечують суттєво більш міцну інтеграцію з Hadoop, коли дані, що зберігаються і в Hadoop, і в реляційній СКБД, можуть оброблятися в одному SQL-запиті. Oracle, IBM, Teradata. Cloudera, Hortonworks та MapR, що також включили Hadoop до своїх продуктових лінійок, роблять все можливе, щоб перемістити Hadoop з високо-масштабованого зберігання даних та Map Reduce обробки у світ аналітики.

Менші постачальники такі як Ac-tian, InfiniDB/Calpont, HP Vertica, Info-bright та Kognitio, навпаки фокусуються загалом скоріше на аналітиці, ніж на обробці транзакцій.

Такі постачальники аналітики, як Alpine Data Labs, Revolution Analytics та SAS, працюють з платформами, які забезпечуються сторонніми постачальниками СКБД та дистриб'ютерами Hadoop, хоча, зокрема, SAS розмиває це розмежування зі зростаючою підтримкою для середовищ SAS-керованих рядів даних «у-пам'яті» та Hadoop. NoSQL та NewSQL СКБД фокусуються на високо-масштабованій обробці транзакцій, а не на аналітиці.

Взагалі програмні засоби для роботи з великими даними не заміщують решту інструментів обробки, бізнес-аналітики, візуалізації та прогнозування, а лише допомагають підтримати терабайти нових даних та спрямовують їх у потрібне русло.

Так, відповідно до аналітичних платформ для великих даних, деякі експерти вважають найбільш універсальною платформу Pentaho, а для вирішення задач машинного самонавчання, таких як, на-приклад, кластеризація, класифікація, регресія та інші, краще підходять Mahout та Spark. Серед найбільш технологічних MPP – платформ спеціалісти виділяють Vertica та Teradata Aster. Останнім часом з'явилася множина платформ, які підтримують швидку аналітику для великих да-них, наприклад, MemSQL або Splice Machine.

Окремої уваги заслуговує Intel платформа з відкритим кодом для Hadoop. Привабливість рішення Intel для Hadoop обумовлює також й фактор "апаратного забезпечення", а саме – оптимізація, що виконана Intel з урахуванням архітектури процесорів Xeon та специфіки роботи твердотільних накопичувачів з контролерами Intel, дозволяє досягти значного приросту продуктивності. Процесори Xeon прискорюють операції шифрування або дешифрування за алгоритмом AES (Advanced Encryption Standard), що реалізується за допомогою додаткового набору команд AES-NI (New Instruction). Окрім цього, платформа Intel для Hadoop також пропонує розширені можливості у галузі обробки потокових даних.

Різноманіття платформ для роботи з великими даних доповнюється величез-ною кількістю прикладних програмних продуктів, комерційних чи безкоштовних, для аналітичної обробки таких даних. Нижче наведений невеликий перелік най-поширеніших прикладів такого ПЗ:

 Cluvio – сучасна платформа аналітики даних, що дозволяє виконувати SQL запити, обробляти дані, візуалізувати результати та створювати гарні, інтерактивні дашборди за лічені хвилини. Підтримує потужне вбудовування, що дозволяє дода-вати аналітичні властивості до будь-якого веб-сайту або веб-застосунку.

 IBM SPSS Statistics дозволяє виявляти нові зв'язки між даними та будувати прогнози. Він дозволяє отримувати легкий доступ до даних,

управляти та аналізувати набори даних, не маючи попереднього статистичного досвіду. Це дозволяє практично виключити довготривалу підготовку даних та швидко створювати, маніпулювати та розповсюджувати інформацію для прийняття рішень.

 Qlik Sense Desktop – безкоштовний продукт, що надає можливість інтерактивного створення звітів та дашбордів з діаграмами та графіками. Програма візуалізації спрощує аналіз даних та допомагає створювати інформовані бізнес-рішення швидше, ніж будь-коли раніше. Перетворення електронних таблиць у більш чіткі візуалізації робить процес аналізу простішим та швидшим для перегляду всіх користувачів.

 Elasticsearch – розповсюджений пошуковий та аналітичний движок на базі Apache Logstash, Kibana та Beats складають "Elastic Stack", розроблений фірмою Elastic. Також Elasticsearch забезпечується хостінг Elastic Cloud.

 Cyfe – бізнес-панель для управління даними компанії за допомогою звітів, попередньо побудованих віджитів тощо.

 Forestpin Analytics – платформа аналізу даних для знаходження нерівностей, кореляцій та дублювань за допомогою простого дашборда.

Кількість підприємств, що використовують великі дані, безперервно зростає. Практика останніх років продемонструвала, що застосування результатів аналізу великих масивів даних може принести реальний ефект. Але, окрім переваг існує велика кількість проблем, вирішення яких вимагає застосування досить значних ресурсів.

Для систем, що отримують аналітичні дані в масштабі, близькому до реального часу, ключовими є вимоги не лише до продуктивності, але й до часу відгуку (наприклад, IBM каже про час відгуку, менший за мілісекунди). Це дуже обмежує вибір аналітичних платформ. Неможливо використовувати колосальні обчислювальні можливості Hadoop, якщо накладні витрати на ініціювання та завершення тривіальної MapReduce-програми складають десятки секунд. Забезпечити прийнятний час відгуку можуть або досить недешеві MPP-платформи (такі як Netezza, Teradata, Greenplum), або розподілені системи з розвиненою індексацією або високим рівнем резидентності даних в оперативній пам’яті.

Багато аналітичних систем все ще використовують реляційну модель даних, внаслідок чого вибір платформ обмежується такими рішеннями, як GridGain або Gigaspaces XAP. Для роботи з потоковими даними в режимі он-лайн були створені технології Storm, Spark Streaming та Akka. Але аналіз даних за допомогою SQL на Hadoop не дозволяє досягти того максиму-му, який пропонує платформа.

Компанії обирають Hadoop, щоб збирати складні та різноманітні дані: історія відвідувань веб-сайтів, логи, дані про використання мобільних пристроїв й інформації з соцмереж та багато іншого. Цими даними складно оперувати у СКБД. Можна витягувати структуровані дані з Hadoop для SQL-

аналізу, але більш перспективними є такі підходи як машинне самонавчання та інші, що дозволяють спів-віднести нові дані зі вже накопиченою, проаналізованою та структурованою інформацією. BI та SQL системи досить добре себе проявили, але постійно виникають нові потреби та нові питання, що виходять за межі поточних можливостей. Уже не достатньо просто управляти даними. Окрім цього, компанії не можуть покладатися лише на аналітику, вони потребують також рішень зі сфери BI, системи збору та передачі оперативної інформації та інше. Межа між цими поняттями почала стиратися, а SAS, Alpine Data Labs та інші стали підтримувати кластеризовані серверні середовища, вимогливі до пам’яті та Hadoop.

Тема 5. Методи моделювання та прогнозування економічного розвитку

1. Основні поняття, цілі та завдання економічного прогнозування

2. Часовий ряд та його компоненти

3. Особливості простих методів прогнозування

4. Кореляційно-регресійні методи та моделі

1. Основні поняття, цілі та завдання економічного прогнозування

Методи та моделі соціально-економічного прогнозування засновані на теорії прогнозування (прогностиці), яка сформувалась порівняно недавно, хоча питаннями пророкування, прогнозування людство цікавилось протягом усієї історії розвитку. Найбільш яскравими прикладами вирішення завдань пророкування є прогнози піфій – віщунів дельфійского оракула, які інтерпретувались жерцями. Віщун Тирезій, описаний Гомером, передбачив Одиссею його майбутнє. Прометей ("провидець", "прозорливець") був єдиною із земних істот, яка могла передбачити майбутнє. Саме тому, що Прометей пророкував майбутнє, Зевс прикув його до скелі й віддав на муки.

Становлення наукового передбачення пов'язане з науковими дослідженнями XX століття. Так, наприклад, відомі дослідження Дж. Форрестера, в яких розроблені моделі світової, промислової, системної динаміки. Такі моделі дозволяли оцінити рівень багатства, бідності, якості життя залежно від рівня забруднення, виробництва продуктів харчування тощо. У розвинених країнах широко впроваджується практика контрактних замовлень на прогнозні розробки, виконувані для урядових закладів і великих компаній. У США центрами подібних досліджень стають «РЕНД-Корпорейшн», Decision and Design, Гудзонський інститут, корпорація Цортона, що спеціалізуються на економічному прогнозуванні. Створюється найвідоміша міжнародна прогнозна організація – «Римський клуб», головною лінією діяльності якого є стимулювання та координація досліджень глобальних проблем.

У своєму розвитку прогнозування пройшло через різноманітні форми,

відповідні до типів державного регулювання змішаної економіки. Історично першою формою економічного прогнозування стала кон'юнктурна, пов'язана з посиленням впливу бюджету на темпи та пропорції економічного зростання в міру збільшення державних видатків у ВВП.

В умовах структурної перебудови економіки і її прискореного розвитку виникла необхідність узгодження бюджетів з показниками економічних прогнозів, на яких ґрунтувалися оцінки податкових надходжень і розмірів дохідної частини бюджету. Це призвело до розробки середньо-строкових і довгострокових прогнозів, прикладами яких є «Вибір шляхів економічного росту» (1976 – 1985 рр.) у Канаді, Прогноз Міністерства праці на 1986 – 1995 рр. у США, «Десятирічний план подвоєння національного доходу» (1961 – 1970 рр.) у Японії.

По мірі вдосконалювання й ускладнення прогнозна діяльність стала відділятися від бюджетної методично й організаційно. Якщо на першому етапі національні економічні прогнози складалися в міністерствах фінансів, то з початку 60-х років XX століття в економічно-розвинених країнах почали створюватися спеціальні прогнозно-планові органи (Генеральний комісаріат із планування у Франції, Економічна консультативна рада в Японії, Центральне планове бюро в Нідерландах та ін.). Були створені спеціальні прогнозні системи. На сьогодні відомі прогнозні системи ІНПРОГС, ПАТТЕРН, ФЕІМ, ПРОФАІЛ тощо.

З часом виникли необхідні умови для створення *теорії прогнозування (прогностики)*, під якою розуміється наукова дисципліна, що вивчає загальні принципи та методи прогнозування розвитку об'єктів будь-якої природи, закономірності процесів розробки прогнозів. Як наука прогностика сформувалась в 70-80 роки XX сторіччя. Дана наука має набір термінів, які вживаються для позначення певних понять.

Слово «прогноз» походить від грецьких слів «про» та «гнозис» і перекладається як «передбачення», «завбачення». Однак таке пояснення не дає цілісного уявлення про сутність та основні завдання прогнозу. Загальним поняттям, яке поєднує всі різновиди отримання інформації про майбутнє, є передбачення, яке поділяється на наукове і ненаукове (інтуїтивне, повсякденне, релігійне й ін.). Наукове передбачення базується на знанні закономірностей розвитку природи, суспільства, мислення; інтуїтивне – на передчуттях людини, повсякденне – на життєвому досвіді; релігійне – на вірі в надприродні сили, які визначають майбутнє. Виділяють дві взаємозалежні форми конкретизації передбачення: форму передрікання, яку прийнято вважати дискриптивною або описовою та передвказівну форму, яку ще називають переддискриптивною або розпорядчою.

За допомогою передрікання проводиться опис можливих або бажаних перспектив, станів, рішень, які стосуються проблем майбутнього. Передвказання пов'язане з власне вирішенням цих проблем, використанням

інформації про майбутнє для цілеспрямованої діяльності особистості і суспільства.

Передрікання найчастіше проявляється у формах передчуття, передвгадування, прогнозування. Передчуття (просте передбачення) містить інформацію про майбутнє на рівні інтуїції – підсвідомості. Передвгадування (складне передбачення) несе інформацію про майбутнє на основі життєвого досвіду, це більш-менш вірні здогади про майбутнє, проте незасновані на спеціальних наукових дослідженнях. Нарешті, прогнозування (яке часто вживають у попередніх значеннях) повинно означати при такому підході спеціальне наукове дослідження, предметом якого виступають перспективи розвитку явища.

Передвказання виступає у формах цілевбачання, планування, програмування та проектування, які є складовими управлінських рішень. Цілевбачання – це встановлення ідеально припустимого результату діяльності. Планування – проекція в майбутнє людської діяльності для досягнення наперед заданої мети при визначених засобах, перетворенні інформації про майбутнє в директиви для цілеспрямованої діяльності. Програмування означає встановлення основних положень, які потім розгортаються в плануванні, або послідовності конкретних заходів щодо реалізації планів.

Проектування – створення конкретних образів майбутнього, деталей програм.

Залежно від ступеня конкретизації та характеру впливу на хід досліджуваних процесів прийнято виокремлювати три форми передбачення: гіпотезу, прогноз та план, програму).

Гіпотеза характеризує наукове передбачення на рівні загальної теорії. Вихідною базою для побудови гіпотези є теорія а також відкриті на її базі закономірності розвитку і причинно-наслідкові взаємозв'язки функціонування досліджуваних об'єктів. На рівні гіпотези подається якісна характеристика розвитку досліджуваної системи або процесу, яка описує загальні закономірності поведінки. Гіпотеза має найменший рівень визначеності.

Прогноз – це ймовірнісне науковообгрунтоване судження про перспективи, можливі стани того чи іншого явища в майбутньому та (або) про альтернативні шляхи і терміни їх здійснення. Порівняно з гіпотезою, прогноз має значно більшу визначеність, так як ґрунтується не тільки на якісних, а й на кількісних параметрах.

Прогноз описує передбачення на рівні прикладної теорії. Таким чином, прогноз відрізняється від гіпотези меншим ступенем невизначеності та більшою достовірністю. В той же час, зв'язки між прогнозною інформацією та станом досліджуваного (прогнозованого) об'єкта не є строго детермінованими та однозначними - прогноз носить імовірнісний характер. В цьому контексті доцільно навести коротке визначення прогнозу, дане Е.Янчем: прогноз (forecast) – це імовірнісне судження про майбутнє з відносно високим рівнем

ймовірності.

Прогноз, порівняно з гіпотезою, має значно більший ступінь визначеності. Ще більший ступень визначеності має план.

План являє собою постановку точно визначеної цілі та передбачення визначених детально описаних подій в розвитку досліджуваного об'єкта. В ньому фіксуються шляхи та засоби розвитку у відповідності з поставленими задачами, обгрунтовуються прийняті управлінські рішення. В плані передбачення отримує найбільшу чіткість та визначеність. Як прогноз, так і план грунтуються на результатах та досягненнях прикладних економіко-математичних методів.

План – це образ досліджуваного об'єкта, система заходів, спрямованих на досягнення поставленої однієї або декількох цілей. В аспекті соціально-економічних процесів його визначають як систему цільових показників розвитку соціально-економічної системи, а також як вказівку на етапи та способи їх досягнення, розподіл ресурсів, визначення очікуваних результатів і способів їх використання. У плані, на відміну від прогнозу, конкретизуються ресурси. Відмінність між планом і прогнозом полягає в такому:

1) прогнозування має характер дослідження, наукового опису майбутнього (пророкування), а план – характер цілевстановлення (предуказання);

2) прогноз має ймовірнісний характер, а план – нормативний. Прогноз може бути дійсним і неправильним, що не можна сказати про план, тому що план є системою заходів;

3) прогноз має варіантний зміст, а план – однозначне рішення;

4) вимога до планів – їх ресурсна забезпеченість, тоді як прогнози можуть пророкувати ймовірність досягнення мети при неповному забезпеченні ресурсами;

5) у процесі планування проявляється вплив суб'єктивних чинників – волі та бажання людини, що ухвалює рішення. При прогнозуванні враховуються об'єктивні дані, що визначають вид і рішення розроблених моделей прогнозу.

Взаємозв'язок плану та прогнозу може бути різноманітним. Прогноз може передувати плану, й, навпаки, можуть формуватися прогнози виконання плану. Із цим етапом зв'язане поняття так званого випереджувального прогнозу, мета якого полягає в забезпеченні керівників об'єктивною інформацією.

Загальні риси прогнозів і планів – випереджальний характер інформації, яка закладена в них, що відрізняє передбачення від соціально- економічного аналізу й статистики.

Програма – сукупність заходів, необхідних для реалізації декількох проблем. Програма може випереджати деякий план або конкретизувати окремий його етап. Як правило, поняття "програма" є більш широким, ніж поняття "план".

Прогнозування соціально-економічних процесів (СЕП) — це наукова дисципліна, яка вивчає основні підходи до розробки прогнозів розвитку національної економіки та соціальної сфери, ґрунтується на науковому пізнанні соціально-економічних явищ та використанні всієї сукупності методів, засобів і можливостей прогностики.

Процес розроблення прогнозів називають прогнозуванням. Подібно до будь-якої діяльності (зокрема й творчої) характер прогнозування визначають його суб'єкт і об'єкт, застосовувані засоби й методи, а також навколишнє середовище.

Прогнози виражаються у вербальній, математичній, графічній або іншій формах.

Суб'єктами прогнозування соціально-економічного розвитку є органи державної влади й місцевого самоврядування, корпорації й підприємства, також науково-дослідні й консалтингові організації, окремі експерти, яких залучають для розроблення й упровадження прогнозів.

Об'єктом соціально-економічного прогнозування є соціально-економічні процеси (СЕП) — тобто сукупність економічних і соціальних процесів формування та функціонування соціально-економічної системи, які характеризують динаміку зміни її параметрів на певному рівні господарювання.

Іншими словами, прогноз є пошуком реалістичного, економічно правильного рішення для управління об'єктом чи процесом, а прогнозування є необхідним і важливим науково-аналітичним етапом загального процесу планування.

Вільна енциклопедія "Вікіпедія" слово "прогноз" (πρόγνωσις – передбачення, пророкування) визначає як пророкування майбутнього за допомогою наукових методів або сам результат пророкування. Прогноз – це наукова модель майбутньої події, явища тощо.

Попри певні відмінності в наведених визначеннях, можна сформулювати ***основні властивості даного поняття.***

1. Прогноз пов'язаний з певним майбутнім станом і (або) шляхами та термінами його досягнення.

2. Прогноз ґрунтується на проведенні певного дослідження, деякого обґрунтування.

3. Прогноз має ймовірнісний характер, тому він не може мати директивний характер.

Процес розробки прогнозів називається *прогнозуванням.* Поняття "пророкування" та "прогнозування" відрізняються за ступенем вірогідності оцінок майбутнього. Логічні формули різноманітних видів процесів вироблення інформації про майбутнє (передбачення) можна записати, як: прогнозування – "імовірно, буде", пророкування – "буде".

Із дослідженням майбутніх процесів і явищ, з передбаченням, крім поняття "прогнозування", пов'язані також поняття "планування" й

"програмування". Відмінності цих понять полягають в аналізованих часових горизонтах. Приблизні часові горизонти для зазначених процесів співвідносяться як 4 : 2 : 1, відповідно, для прогнозування, програмування та планування.

Прогнозування та планування можуть використовувати однакові методи та показники, будуватися на основі загальної інформаційної бази. Прогнозування є дослідницькою попередньою базою для планування. Прогноз передує розробці плану, але може також слідувати, визначаючи можливості досягнення запланованих рівнів показників.

Прогностика в сучасному її стані включає:

1. основи прогнозування: основні поняття; принципи прогнозування; види та призначення прогнозів; параметри прогнозів; етапи прогнозування;
2. об'єкт прогнозування: характеристики об'єкта прогнозування; вихідну інформацію про об'єкт; аналіз об'єкта прогнозування;
3. апарат прогнозування: фактографічні методи; експертні методи; методи верифікації.

Одним із найбільш важливих напрямів суспільного розвитку є соціально-економічне прогнозування. Останнє розглядається як процес розробки прогнозів, заснований на наукових методах пізнання соціально-економічних явищ і використання сукупності методів і засобів соціально-економічної прогностики.

До основних понять прогнозування, крім раніше зазначених, належать такі.

Етап прогнозування – частина процесу розробки прогнозів, що характеризується завданнями, методами та результатами. Розподіл на етапи позв’язаний зі специфікою побудови систематизованого опису об'єкта прогнозування, збору даних, з побудовою моделі, верифікацією прогнозу.

Модель прогнозування – модель об'єкта прогнозування, дослідження якої дозволяє отримати інформацію про можливі стани об'єкта прогнозування в майбутньому й (або) шляхи та терміни їх здійснення.

Метод прогнозування – спосіб дослідження об'єкта прогнозування, спрямований на розробку прогнозу. Методи прогнозування є підставою для методик прогнозування.

Методика прогнозування – сукупність спеціальних правил і прийомів (одного або декількох методів) розробки прогнозів.

Верифікація прогнозу – оцінювання вірогідності, точності або обґрунтованості прогнозу.

Прогнозним фоном називають сукупність зовнішніх чинників, що впливають на прогноз.

Прийом прогнозування – одна або кілька математичних або логічних операцій, спрямованих на отримання конкретного результату в процесі розробки прогнозу.

Об'єкт прогнозування – процеси, явища та події, на які спрямована діяльність суб'єкта прогнозування.

Система прогнозування (прогнозувальна система) – сукупність методик, технічних засобів, призначених для прогнозування складних явищ або процесів.

2. Часовий ряд та його компоненти

Ряди динаміки (часові ряди) характеризують процеси розвитку соціально-економічних явищ. Цим процесам властиві дві взаємопов’язані риси: динамічність та інерційність. Динамічність проявляється зміною рівнів і варіації показників, що характеризують процес, інерційність — сталістю механізму формування процесу, напрямку та інтенсивності динаміки протягом певного часу. Поєднуючи ці риси, динамічний ряд у будь-який момент *t* містить залишки минулого, основи сучасного і зародки майбутнього.

Діалектична єдність мінливості й сталості, динамічності й інерційності формує закономірність розвитку. Під впливом безлічі факторів довгострокової і короткострокової дії в одних рядах рівні протягом тривалого часу зростають або зменшуються з різною інтенсивністю, в інших зростання і зменшення рівнів чергуються з певною періодичністю (наприклад, одинадцятирічні цикли градових опадів, зумовлені циклами сонячної активності). З року в рік більш-менш регулярно повторюються сезонні піднесення і спади (використання виробничих потужностей і робочої сили, попит на ринку споживчих товарів тощо). Окрім закономірних коливань рівнів, динамічним рядам притаманні також випадкові коливання, пов’язані з масовим процесом.

Ряди, в яких рівні коливаються навколо постійної середньої, називаються стаціонарними. Економічні ряди, як правило, нестаціонарні. Для більшості з них характерна систематична зміна рівнів з нерегулярними коливаннями, коли піки і западини чергуються з різною інтенсивністю. Скажімо, економічні цикли (промислові, будівельні, фондового ринку тощо) повторюються з різною тривалістю і різною амплітудою коливань. Рисунок 1 ілюструє характер динаміки виплат страхового відшкодування VAR2, коливання якого залежать від кількості постраждалих об’єктів. Поквартальні ($n = 18$) обсяги виплат коливаються від 7,9 до 19,2 млн. грн., на графіку вони представлені відхиленнями від мінімального рівня.

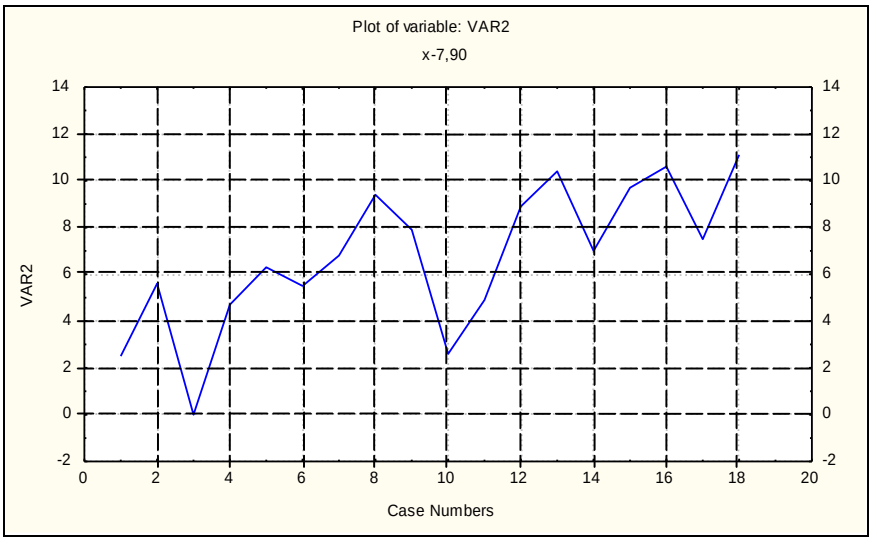


Рис. 5.1. Динаміка виплат страхового відшкодування

При моделюванні динамічних процесів причинний механізм формування властивих їм особливостей у явному вигляді не враховується. Будь-який процес розглядається як функція часу. Певна річ, час не є фактором конкретного соціально-економічного процесу, змінна часу t просто акумулює комплекс постійно діючих умов і причин, які визначають цей процес.

У моделях динаміки процес умовно поділяється на чотири складові:

- довгострокову, детерміновану часом еволюцію — тренд $f(t)$;
- періодичні (циклічні) коливання різних частот C_t;
- сезонні коливання S_t;
- випадкові коливання e_t.

Зв’язок між цими складовими може представлятися адитивно (сумою) або мультиплікативно (добутком):

$$y_t = f(t) + C_t + S_t + e_t, \quad (1)$$

$$y_t = f(t)\, C_t\, S_t\, e_t. \quad (2).$$

Така умовна конструкція дає змогу, залежно від мети дослідження, вивчати тренд, елімінуючи коливання, або вивчати коливання, елімінуючи тренд. При прогнозуванні здійснюється зведення прогнозів різних елементів в один кінцевий прогноз.

Характерною властивістю будь-якого динамічного ряду є залежність рівнів: значення y_t певною мірою залежить від попередніх значень: y_{t-1}, y_{t-2} і т. д. Для оцінювання ступеня залежності рівнів ряду використовують *коефіцієнти автокореляції* r_p з часовим лагом $p = 1, 2, \ldots, m$.

Переважна більшість методів прогнозування з використанням часових рядів ґрунтується на ідеї екстраполяції, тобто уявному продовженні на майбутнє тенденції зміни значень досліджуваного показника, яка спостерігалася в минулому до моменту розрахунку прогнозу. При цьому робиться припущення, що фактори, які впливали на результуючий показник в минулому, суттєво не змінять характер свого впливу на період прогнозування.

Під тенденцією розуміють деякий загальний характер змін досліджуваного показника, обумовлений внутрішніми взаєзв'язками факторів, які впливають на розвиток процесу.

Кожна з вищезазначених складових моделі має свій характер впливу на результуючий показник:

- *тренд* віддзеркалює вплив причинно-наслідкових закономірностей, властивих досліджуваному процесу та обумовлених довгодіючими факторами його природи;
- *сезонна складова*, з допомогою якої враховують можливі повторення впливу деяких тимчасових факторів на результуючу змінну протягом відносно короткого проміжку часу – пори року, місяця, тижня;
- *циклічна складова*, з допомогою якої враховують можливі періодично повторювані умови змін в розвитку досліджуваного процесу;
- *випадкова складова*, з допомогою якої враховують вплив на результуючу змінну випадкових і неспостережуваних факторів.

Через труднощі в одночасному врахуванні характеру впливу на результат усіх чотирьох компонент, виділяють тільки дві складові: закономірну та випадкову. Закономірна складова об'єднує тренд, сезонну та циклічну компоненти і називається *трендом*. В цьому контексті тренд характеризує природну закономірність зміни значень досліджуваного показника в часі, звільнену від впливу випадкових факторів.

Отже, рівні $y(t)$ часового ряду доцільно представити у вигляді залежності

$$y(t)=f(t)+\varepsilon(t), \qquad (3)$$

де $f(t)$ – аналітичне представлення тренду з урахуванням можливих циклічних та сезонних складових;

$\varepsilon(t)$ – міра відхилення наявних експериментальних даних від відповідних аналітичних величин, обумовлена дією випадкових факторів.

Причиною широкого використання методів екстраполяції тенденції зміни часових рівнів є відсутність іншої інформації, крім дискретних значень самого показника. Іншими словами – інколи стає проблематично, нерентабельно або ж взагалі неможливо зібрати інформацію про значення факторів, які впливають на значення досліджуваного показника протягом тривалого періоду часу.

Використання методів екстраполяції ґрунтується на наступних припущеннях:

1) тенденція зміни в часі кількісної міри досліджуваного показника може бути представлена певною аналітичною залежністю;

2) умови, які визначали тенденцію зміни в минулому, несуттєво зміняться в недалекому майбутньому.

Основні види тенденції в часових рядах:

– *тенденція середнього рівня* – відображається, як правило, рівнянням лінії, навколо якої змінюються фактичні рівні досліджуваного явища, тобто $y(t)=f(t)+\varepsilon(t)$. Зміст даної функції полягає в тому, що значення тренду в окремі

моменти часу є математичними сподіваннями ряду динаміки;

– *тенденція дисперсії* – характеризує тенденцію зміни відхилень між емпіричними рівнями та детермінованою компонентою ряду;

– *тенденція автокореляції* – характеризує зв'язок окремими рівнями ряду динаміки.

Перед побудовою тренду доцільно перевірити виконання гіпотез, чи дійсно рівні ряду змінюються в часі, чи наявні значення різниць їх величин обумовлені лише дією випадкових факторів. Для цього проводять перевірку гіпотез про зміну середнього, дисперсії, автокореляції.

Основні етапи прогнозування з використанням трендових моделей:

1. Попередній аналіз даних (перевірка ряду на стаціонарність).
2. Формування набору моделей (відбір декількох, як правило, нелінійних моделей, котрі візуально найкраще описують закономірність розвитку процесу).
3. Кількісна оцінка параметрів моделей.
4. Перевірка оцінених моделей на адекватність, а параметрів – на статистичну значимість.
5. Вибір найкращої моделі (на основі логічного, економічного та математико-статистичного аналізу).
6. Розрахунок точкового та інтервальних прогнозів.
7. Верифікація прогнозу.

3. Особливості простих методів прогнозування

Метод екстраполяції – один з основних методів у прогнозуванні економічних явищ та процесів. Сутність методу – на основі статистичних даних досліджують закономірності й тенденції розвитку економічних явищ та процесів. Даний метод ґрунтується на припущенні, що ті фактори, які впливали на розвиток певного явища в минулому, будуть діяти і в майбутньому. При формуванні прогнозу за допомогою екстраполяції виходять з тенденцій зміни кількісних характеристик об'єкта дослідження.

Методи екстраполяції можуть бути простими і складними. *Прості методи* прогнозування на основі екстраполяції використовують в управлінні виробництвом, оскільки вони мають ряд переваг:

- достатньо простий апарат дослідження;
- можливість використання для розрахунків портативних і нескладних обчислювальних засобів;
- швидкість виконання розрахунків в оперативному режимі;
- наявність відносно невеликого масиву інформації.

Складні методи екстраполяції передбачають виявлення основної тенденції, тобто застосування статистичних формул, що описують тренд. Методи цієї групи можна розділити на два основні типи: аналітичні й адаптивні.

В основу аналітичних методів прогнозування покладений принцип отримання за допомогою методу найменших квадратів оцінки детермінованої компоненти, що характеризує основну тенденцію.

Метод аналітичного вирівнювання тренду (метод найменших квадратів) може бути застосований тільки в тому випадку, коли розвиток явища досить добре описується побудованою моделлю й умови, що визначають тенденцію розвитку в минулому, істотно не зміняться у майбутньому.

Основу екстраполяційних методів прогнозування складають *динамічні ряди.* Ряд динаміки – це числова послідовність, що характеризує зміну економічного явища у часі.

При побудові динамічних рядів слід в першу чергу приділити увагу на порівнянність рівнів ряду. Це значить, що усі рівні повинні виражатися в однакових одиницях виміру, розраховуватися по єдиній методології, включати єдине коло об'єктів.

Завдання прогнозування економічних показників по рядах динаміки зводиться до наступного:

1. Заданий один часовий ряд показників Y_t (t=1,2,3.,n). Потрібно спрогнозувати значення показника Y для $t > t_n$.

2. Дана система рядів динаміки, в яких показник одного ряду залежить від інших. Необхідно знайти цю залежність, спрогнозувати в кожному ряду показники і лише потім спрогнозувати по знайденій залежності основний показник.

Тенденція ряду динаміки – це загальний напрям розвитку процесу, явища, показника, довгострокова закономірність. Тенденція виражається за допомогою тренду – рівняння, в якому основним чинником виступає час.

При прогнозуванні методами екстраполяції виходять з інерційності явищ (процесів), що досліджуються і прогнозуються.

Ступінь інерційності залежить від розміру і масштабу процесу, що вивчається. На мікрорівні вплив окремого фактору може миттєво змінити ситуацію, в той час, коли на макрорівні, через дії багатьох факторів, які здійснюють часом протилежний один одному вплив, інерційність зберігається у більшій мірі.

При значній інерційності економічних процесів (явищ), що досліджуються, можна з достатнім ступенем імовірності сподіватися, що закономірності, які виникли в «передісторії», будуть з незначними змінами діяти і в прогнозованому періоді.

При побудові ряду оперативних соціально-економічних прогнозів дослідник стикається з проблемою здобуття достатнього обсягу необхідної інформації. Як правило, в цій ситуації доводиться мати справу з короткими часовими рядами, довжина яких може не перевищувати десяти точок. У зв'язку з неповнотою кількісної інформації використовувати досить складні методи формального прогнозування не доцільно. На практиці при побудові кількісних

прогнозів використовують відносно прості методи екстраполяції. Ці методи дозволяють отримати хоча і «грубі», однак кількісні оцінки, на основі яких приймаються управлінські рішення.

Методи простої екстраполяції засновані на припущенні про практично незмінний характер поточного процесу, про відсутність істотних змін у стані зовнішнього та внутрішнього середовища об'єкту прогнозування. Це, у свою чергу, накладає певні обмеження на можливості використання даних методів. Як правило, вони використовуються для отримання оперативних і короткострокових прогнозів за умов неповної інформації.

Операцію екстраполяції в загальному вигляді можна подати у вигляді визначення значення функції:

$$y_{j+h}=f(y_j^*, h) \qquad (4)$$

де y_{j+h} – значення рівня, що екстраполюється;

h – період упередження;

y_j^* – рівень, прийнятий за базу екстраполяції.

Існують різноманітні прийоми екстраполяції, серед яких будуть розглянуті: метод побудови «наївної моделі»; метод екстраполяції на основі середньої; метод екстраполяції на основі середнього темпу зростання; екстраполяція на основі лінійного тренду, побудованого по двох крайніх точках або двох середніх групових точках.

4. Кореляційно-регресійні методи та моделі

Економетричні моделі в залежності від обсягу вибірки статистичних даних поділяються на узагальнені та вибіркові.

Узагальненою вважається регресійна модель, побудована по статистичних даних генеральної сукупності і має вигляд: $y = \beta_0 + \beta_1 x + u$, де β_0, β_1 – параметри моделі, u – випадкова величина (відхилення).

Вибіркова модель будується по статистичних даних вибіркової сукупності. У загальному вигляді вибіркова регресійна модель між факторною ознакою $X = \{x_1, x_2, \ldots, x_n\}$ та результативною ознакою $Y = \{y_1, y_2, \ldots, y_n\}$ з урахуванням фактору випадкових величин (помилок) $U = \{u_1, u_2, \ldots, u_n\}$ записується у вигляді:

$$y = a_0 + a_1 x + u \qquad (5)$$

де a_0 , a_1 – невідомі параметри економетричної моделі;

u – випадкова величина (відхилення).

Тут і надалі з метою уникнення неоднозначності великими літерами X ,Y ,U ми позначаємо дискретні (векторні) величини, а малими x, y, u - неперервні.

Причини обов'язкової присутності в регресійних моделях випадкової змінної (відхилення) *u* такі:

1. Невключення до моделі всіх пояснюючих змінних. Будь-яка регресійна

модель є спрощенням реальної ситуації. Остання завжди являє собою взаємодію різних чинників, багато з яких в моделі не враховуються, що обумовлює відхилення реальних значень залежної змінної від її модельних значень. Проблема полягає ще й в тому, що наперед не відомо, які фактори при певних умовах дійсно є визначальними, а якими можна нехтувати.

2. Неправильний вибір функціональної форми моделі. Через недостатню вивченість процесу чи явища, що моделюється, може бути невірно підібрана аналітична функція, якою проводиться моделювання.

3. Агрегація змінних. У багатьох моделях розглядаються залежності між чинниками, які представляють складну комбінацію інших, простіших змінних. Наприклад, чинник сукупний попит є складною композицією індивідуальних попитів, які впливають на результативний показник. Це може виявитись причиною відхилення реальних значень від модельних.

4. Помилка вимірювань. Якою б якісною не була модель, помилки вимірювань змінних вплинуть на невідповідність модельних значень емпіричним даним, що також відобразиться на величині випадкового члена (відхилення).

5. Обмеженість статистичних даних. Часто будуються моделі, що виражаються безперервними функціями. Але для цього використовується набір даних, що мають дискретну структуру. Ця невідповідність знаходить свій вираз у випадковому відхилені.

6. Непередбачуваність людського чинника. Ця причина може «зіпсувати» найякіснішу модель. Дійсно, при правильному виборі форми моделі, скрупульозному підборі пояснюючих змінних все одно неможливо спрогнозувати поведінку кожного індивідуума.

Таким чином, відхилення (випадкова величина) є віддзеркаленням впливу всіх описаних вище причин. До того ж, цей перелік може бути доповненим.

Метод математичної статистики, який вивчає кореляційні зв’язки між явищами, називається ***кореляційним аналізом***. Кореляційний аналіз представляє собою інструмент, який дозволяє кількісно оцінити зв’язки між великою кількістю взаємодіючих економічних явищ, при цьому деякі з них невідомі. Застосування кореляційного аналізу дає можливість перевірити різні економічні гіпотези про наявність і силу зв’язку між двома явищами або явищем та групою явищ, а також гіпотезу про форму зв’язку.

Задачею регресійного аналізу є обчислення невідомих параметрів a_0, a_1 рівняння регресії $\hat{y} = a_0 + a_1 x$. При цьому необхідно досягти «найкращої» апроксимації. Найчастіше при цьому використовують метод найменших квадратів, що передбачає мінімізацію виразу:

$$Q(a_0, a_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} u_i^{\,2} \to \min \qquad (6)$$

де $y_i, \hat{y}_i$ фактичні (емпіричні) та розрахункові (теоретичні) значення результативної ознаки.

На рисунку 5.3 пряма є теоретичною лінією регресії.

Розглянемо геометричну інтерпретацію комбінації цих двох складових (рис.15.1.1). Показники $x_1, x_2, .., x_n$ – це гіпотетичні значення пояснювальної змінної. Якщо би співвідношення між y та x були однаковими, то відповідні значення y були би представлені точками $B_1, B_2, .., B_n$ на одній прямій. Наявність випадкового члена збурення приводить до того, що насправді значення y отримують іншим. Відзначимо на графіку реальні значення y при відповідних значення x з допомогою точок $A_1, A_2, \ldots, A_n$.

Із множини прямих необхідно вибрати «найкращу» з точки зору мінімізації суми квадратів відхилень u_i : $u_i = y_i - \hat{y}_i = y_i - a_0 - a_1 \cdot x_i; i = \overline{1,n}$. Відхилення або помилки u_i ще іноді називають залишками. Теоретичну лінію регресії необхідно проводити таким чином, щоб сума квадратів відхилень була мінімальною.

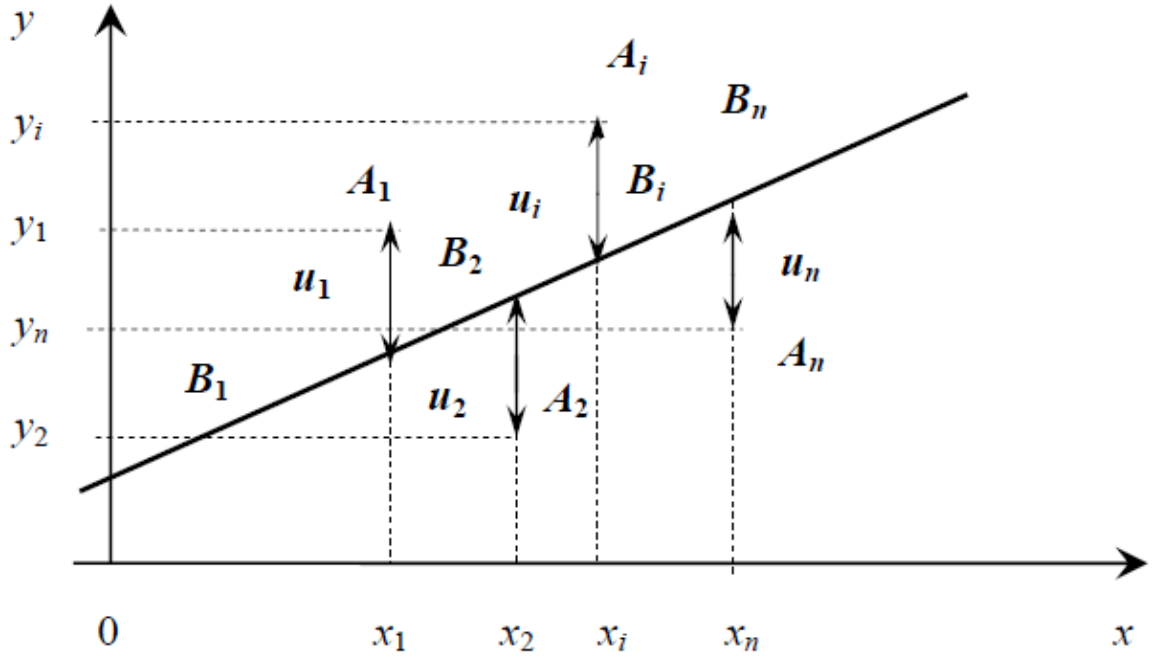


Рис. 5.2. Фактична залежність між y та x

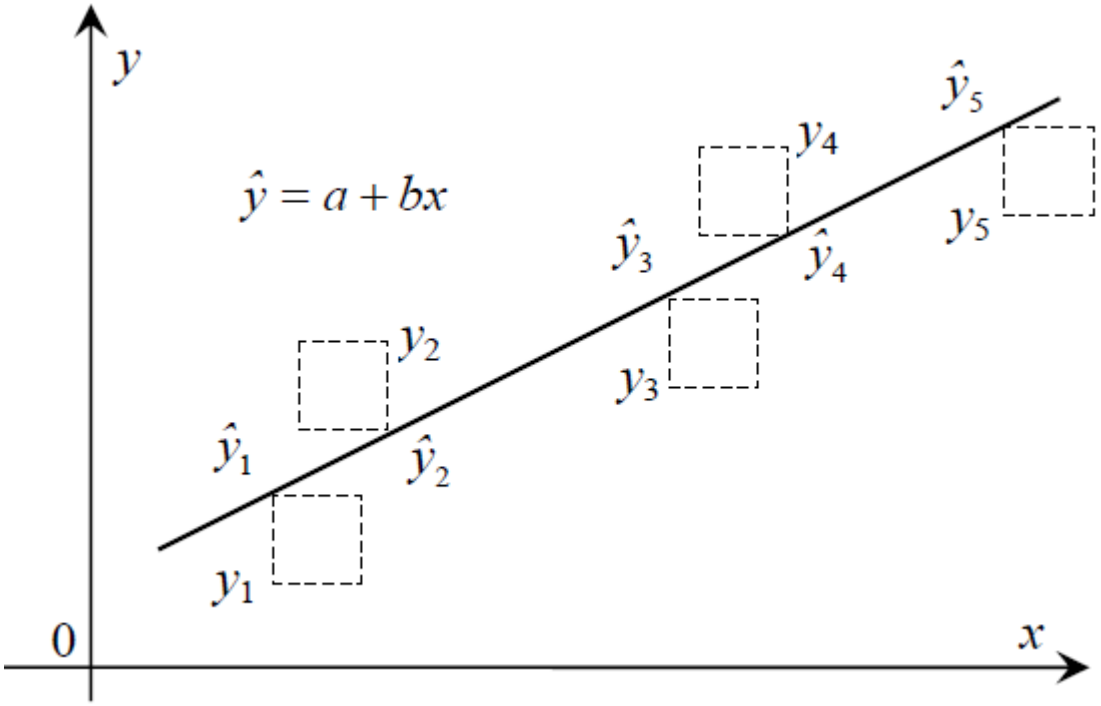


Рис. 5.3. Графічна інтерпретація методу найменших квадратів

У цьому і полягає метод найменших квадратів: невідомі параметри a_0 та a_1 визначаються таким чином, щоб мінімізувати $\sum_{i=1}^{n} u_i^2$. Мінімум функції (6) досягається за умови, коли перші похідні дорівнюють нулю. Тому підставивши в вираз (9.2), взявши частинні похідні $\frac{\partial Q}{\partial a_0}$ і $\frac{\partial Q}{\partial a_1}$, після елементарних перетворень одержимо систему нормальних рівнянь:

$$\begin{cases} na_0 + a_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ a_0 \sum\limits_{i=1}^{n} x_i + a_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i x_i \end{cases} \qquad (7)$$

де n – кількість спостережень або довжина вибірки.

Шляхом розв'язання системи нормальних рівнянь на основі метода найменших квадратів оцінюються параметри лінійної економетричної моделі a_0 та a_1:

$$a_0 = \frac{\sum\limits_{i=1}^{n} y_i \cdot \sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n} y_i x_i \cdot \sum\limits_{i=1}^{n} x_i}{n \cdot \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2} \qquad (8)$$

$$a_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (9)$$

Параметри a_0 , a_1 мають таку економічну інтерпретацію або зміст: параметр a_0 характеризує деяке середнє значення результативного показника y , а параметр a_1 показує, як в середньому зміниться y при зміні x на одну одиницю.

Постійна величина a_0 визначає точку перетину прямої регресії з віссю ординат і є середнім значенням у точці x_0=0. Зрозуміло, що економічна інтерпретація a не тільки утруднена, а й взагалі неможлива. Величина a_0 у рівнянні регресії лише виконує функцію вирівнювання і має розмірність y. При цьому слід відзначити, що завдяки постійній a_0 функція регресії непомилкова. Рівняння регресії інтерпретується тільки в області скупчення точок і, як наслідок, тільки між найменшим і найбільшим значенням змінної x, яка спостерігається. Більш практичний інтерес представляє економічний зміст величин a_1 та $\hat{y}$.

Відповідно до рівняння a_1 визначає середню величину зміни результативного показника при зміні пояснювальної змінної x на одну одиницю. Знак a_1 визначає напрямок цієї зміни, а розмірність цього коефіцієнта є відношенням розмірності залежної змінної до розмірності пояснювальної змінної.

Приклад. Нехай залежність денного виробітку робітника від рівня механізації праці описується рівнянням регресії: y = 2,142 + 0,051x. У цьому рівнянні параметр a_0 є середнім денним виробітком при виконанні операції вручну, а a_1 – перевищення середнього виробітку при механізованому виконанні операції. А тому параметр a_1 (коефіцієнт нахилу) показує, що при підвищенні рівня механізації на 1% денний виробіток зростає в середньому на 0,051 одиниць.

Отже, при моделюванні та аналізі багатьох соціально-економічних явищ та процесів виникає задача виявлення та оцінки зв’язку між ними, одне з яких є незалежною змінною (x), чи фактором, а інше (y) – залежною або результативною ознакою. Форма зв’язку між змінними x та y встановлюється шляхом логічного аналізу їх природи та зовнішнього вигляду кореляційного поля та емпіричної лінії регресії, а тіснота зв’язку – величиною коефіцієнта кореляції.

Тіснота (щільність) зв’язку між змінними x та y оцінюється **коефіцієнтом парної кореляції або коефіцієнтом кореляції Пірсона** r_{xy} (якщо зв’язок лінійний) і **кореляційним відношенням** η_{xy} (якщо зв’язок нелінійний).

Коефіцієнт кореляції являє собою ступінь асоціативності між двома змінними.

Для обчислення коефіцієнта кореляції пропонуються різні формули.

Розглянемо деякі з них:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \tag{10}$$

де $\overline{xy}$ – середнє значення добутку змінної x та змінної y ; $\bar{x}$, $\bar{y}$ – середнє значення змінних x та y:

$$\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i , \qquad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i , \qquad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \tag{11}$$

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \overline{y})}{n \cdot \sigma_x \cdot \sigma_y} \tag{12}$$

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left[n \cdot \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i \right)^2 \right] \cdot \left[n \cdot \sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i \right)^2 \right]}} \tag{13}$$

де n – довжина вибірки або кількість спостережень;

$Cov(x, y)$ – коефіцієнт коваріації між змінними x та y ;

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1} \tag{14}$$

$Var(x)$ – дисперсія змінної x:

$$Var(x) = \sigma_x^{\,2} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{15}$$

$Var(y)$ – дисперсія змінної y:

$$Var(y) = \sigma_y^{\,2} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{16}$$

Визначена таким чином величина має назву коефіцієнта кореляції за вибіркою.

Властивості коефіцієнта кореляції:

1. Він може бути позитивним або негативним, знак r залежить від знаку чисельника (9.10), що є мірою коваріації за вибіркою двох змінних.

2. Коефіцієнт кореляції змінюється в інтервалі:

$$-1 \le r_{xy} \le 1$$

3. За своєю природою він симетричний, тобто коефіцієнт кореляції між x_i і y_i (r_{xy}) той же, що й між y_i і x_i (r_{yx}). Тому, іноді скорочено будемо його позначати просто r.

4. Він незалежний по відношенню до вибору початку системи координат і масштабу вздовж осей координат, тобто, якщо ми визначимо $X_i^* = aX_i + C$ і $Y_i^* = bY_i + d$, де $a > 0$, $b > 0$, a, b і d – константи, то r_{xy} між X^* і Y^* той ж, що й між початковими змінними X і Y.

5. Якщо X і Y статистично незалежні, коефіцієнт кореляції між ними дорівнює нулю, але якщо $r = 0$, це не означає, що дві змінні незалежні. Іншими словами, нульовий коефіцієнт кореляції не обов'язково означає незалежність (рис. 9.5 *з*).

6. Коефіцієнт кореляції є мірою тільки лінійної асоціативності або лінійної залежності; він незастосовний для опису нелінійної залежності. Так, на рис. 2.13, *з* $y=x^2$ є точна залежність, хоча $r = 0$.

7. Хоча r є мірою лінійної асоціативності між двома змінними, це необов'язково означає будь-який причинно-наслідковий зв'язок, як було відзначено раніше.

При коефіцієнті кореляції рівному 0, між y та x не існує кореляційного зв'язку. Якщо коефіцієнт кореляції знаходиться в інтервалі $-1 \le r_{xy} \le 1$ або $0 \le r_{xy} \le 1$ 0, між y та x існує обернена або пряма кореляційна залежність.

За щільністю зв'язку можна виділити:

а) слабкий зв'язок, якщо $r_{xy} \le 0{,}3$;

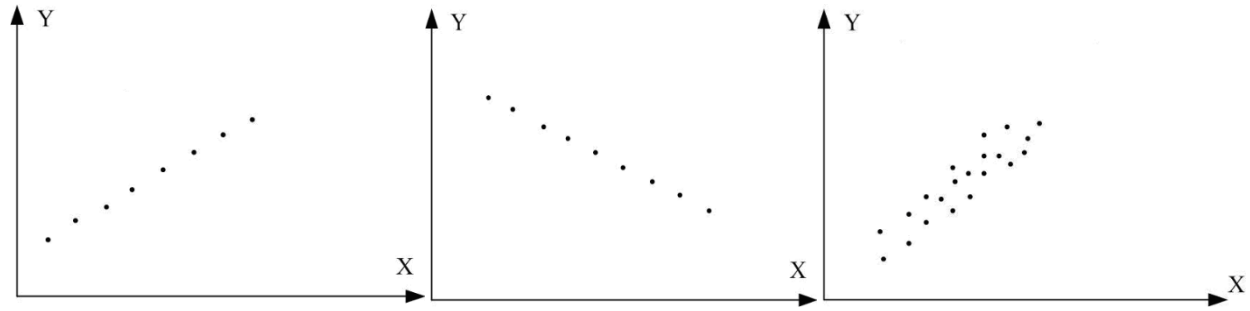
б) середній зв'язок, якщо r_{xy} = 0,31-0,65;

в) сильний зв'язок, якщо r_{xy} = 0,66-0,95.

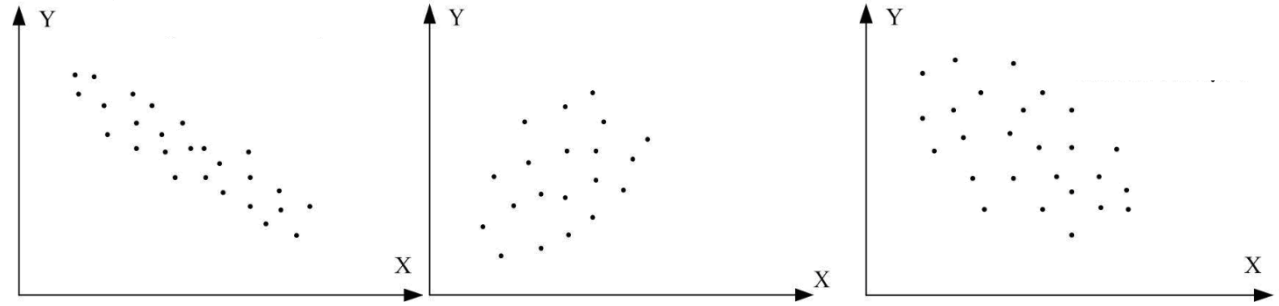
За значенням коефіцієнта кореляції можна зробити такі висновки:

- якщо r_{xy} набуває значення, яке близьке до -1, то між факторами існує щільний зворотний (обернений) зв'язок;
- якщо $r_{xy} = 0$, то зв'язок відсутній;
- якщо r_{xy} близьке до +1, то між факторами існує щільний прямий зв'язок;
- якщо $|r_{xy}| = 1$, то між досліджуваними показниками існує функціональний зв'язок.

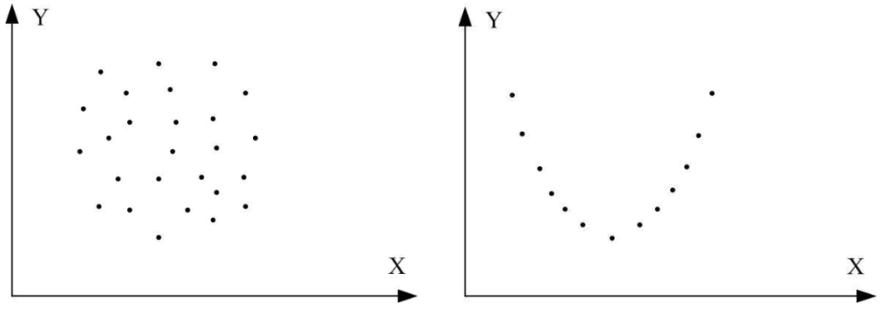
Відзначимо, що знак коефіцієнта кореляції r вказує на напрям зв'язку між ознаками x y, в той час як $|r_{xy}|$ характеризує щільність зв'язку.



а) r=1; *б*) r=-1 *в)* r близький до 1



г) r близький до –1; *д)*r додатній, близький до 0; *е)* r від'ємний, близький до 0



ж) r=0; *з)* r=0

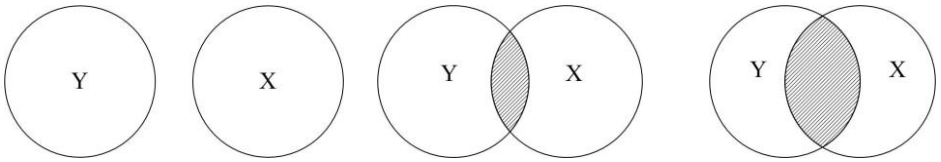
Рис. 5.4. Кореляційний коефіцієнт для різних випадків вибірок

Коефіцієнт детермінації r^2 : міра «якості підгонки».

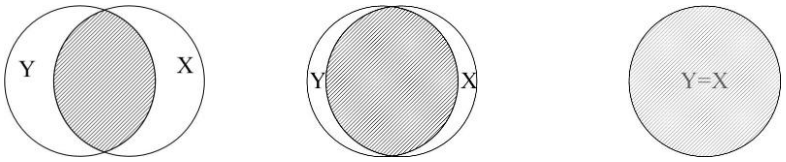
Звернемося зараз до розгляду питання якості підгонки лінії регресії до множини даних, тобто дослідимо, наскільки «добре» лінія вибіркової регресії підходить до цих даних. Із рис.9.3 видно, що якби всі спостереження знаходилися на лінії регресії, ми отримали б «точну підгонку», але на практиці це окремий випадок. У загальному ж випадку будуть як позитивні відхилення u_i, так і негативні. Ми прагнемо, щоб ці залишки були наскільки можливо малі. Коефіцієнт детермінації $\boldsymbol{r^2}$ (випадок двох змінних) або $\boldsymbol{R^2}$ (множинна регресія) являє собою сумарну міру якості підгонки лінії регресії до даних спостереження.

Перш ніж з'ясовувати, як підраховується r^2, розглянемо евристичне пояснення r^2 за допомогою графічних діаграм, відомих як діаграма Венна (Venn) (рис. 9.6).

На рис. 9.6 коло *Y* зображає дисперсію залежної змінної *Y*, а коло *X* – дисперсію пояснювальної змінної *X*. Перетин двох кіл (заштрихована область) являє собою область, у якій дисперсія *Y* пояснюється дисперсією в *X* (скажімо, за регресією МНК). Чим більша область перетину, тим більше дисперсія *Y* пояснюється за допомогою *X*. Коефіцієнт детермінації r^2 зображає числову міру області перетину. На рис. 5.5 бачимо, що при русі зліва направо область перекриття збільшується, тобто послідовно зростає частина варіації *Y*, з'ясована за допомогою *X*, - r^2 зростає. Коли перекриття немає, r^2, очевидно, дорівнює нулю, а коли відбувається повне перекриття, то $r^2=1$, оскільки 100% дисперсії *Y* пояснюється за допомогою *X*. Отже, r^2 лежить між 0 і 1.



а) $r^2=0$ *б)* r^2 близький до 0 *в)* r^2 близький до 0,5



г) r^2 більш ніж 0,5 *д)* r^2 близький до 1 *е)* r^2 дорівнює 1

Рис. 5.5. Діаграма для пояснення r^2

Для визначення варіації результативного показника під впливом факторів обчислюють **коефіцієнт детермінації** r^2_{yx}. Припустимо, що r^2_{yx}=0,8; тоді можна сказати, що 80% варіації результативного показника відбувається під впливом фактору *x*, а решта 20% приходиться на інші фактори та випадкові величини.

При виявленні зв'язку між варіацією факторної ознаки *(x)* і варіацією результативної ознаки *(y)* використовують такі *дисперсії:*

1) дисперсія, яка вимірює загальну варіацію за рахунок дії всіх факторів, або *загальна дисперсія*:

$$\sigma^2_{загальна} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n} \qquad (17)$$

2) *пояснювальна дисперсія*, яка вимірює варіацію результативної ознаки *y*

за рахунок дії факторної ознаки *x* або дисперсія, що пояснює регресію:

$$\sigma^2_{пояснювальна} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{n} \quad (18)$$

3) *залишкова (непояснювальна) дисперсія*, яка характеризує варіацію ознаки *y* за рахунок всіх факторів, крім *x* (тобто при виключенні *x*) або дисперсія помилок:

$$\sigma^2_{непояснювальна} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n} \quad (19)$$

тоді по правилу додавання дисперсій:

$$\sigma^2_{загальна} = \sigma^2_{пояснювальна} + \sigma^2_{непояснювальна} \quad (20)$$

або

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (21)$$

де $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ – загальна сума квадратів, яка позначається через *TSS (total sum squares)*; вона відображає дисперсію величини y_i (емпіричне або фактичне значення) відносно її середнього значення;

$\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ – сума квадратів, що пояснює регресію та позначається через *ESS (explained sum squares);* відображає дисперсію оціненої (теоретичної) величини $\hat{y}_i$ відносно середнього значення y_i;

$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ – сума квадратів помилок, яка позначається через *RSS (residual sum squares)*; відображає залишкову або нез'ясовану дисперсію величини y_i щодо лінії регресії $\hat{y}_i$ або просто залишкова сума квадратів.

Вирази (9.17), (9.18) запишемо у скороченому вигляді:

$$TSS=ESS+RSS \quad (22)$$

Формула (9.19) показує, що загальна варіація спостережуваних величин *Y* щодо їх середнього значення може бути розбита на дві частини, одна відповідає лінії регресії, а інша – випадковим відхиленням, оскільки не всі спостережувані *Y* лежать на лінії регресії. На рис. 9.7 це розбиття пояснене геометрично.

Таким чином, ми розклали загальну дисперсію на дві частини: дисперсію, яка пояснює регресію, та дисперсію помилок (або дисперсію випадкової

величини).

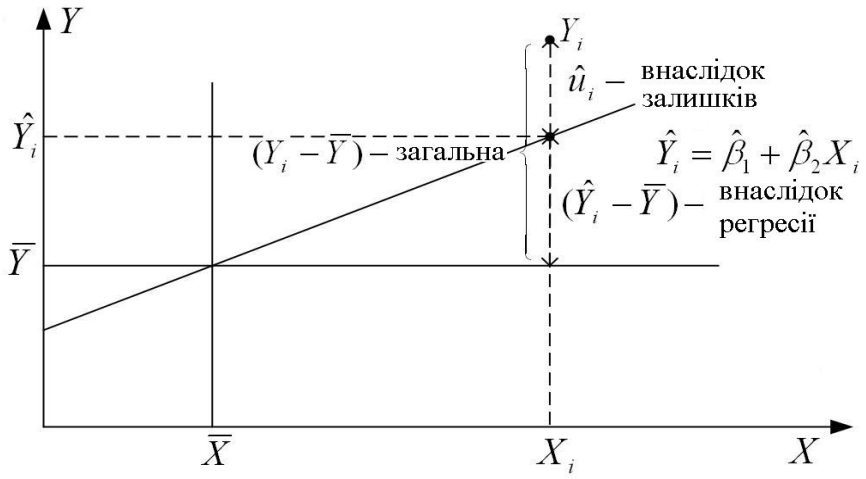


Рис. 5.6. Розбиття варіації Y_i на дві компоненти

Поділивши обидві частини виразу (22) на *TSS*=$\sigma^2_{загальна}$, отримаємо:

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

або

$$1 = \frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}} + \frac{\sigma^2_{непояснювальна}}{\sigma^2_{загальна}} \quad (23)$$

Із виразу (23) випливає, що перша частина $\frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}}$ є складовою дисперсії, яку можна пояснити через лінію регресії. Друга частина $\frac{\sigma^2_{непояснювальна}}{\sigma^2_{загальна}}$ є питомою вагою помилок у загальній дисперсії, тобто часткою дисперсії, яку не можна пояснити через регресійний зв'язок.

Частина дисперсії, що пояснюється регресію, називається **коефіцієнтом детермінації** і позначається r^2. Коефіцієнт детермінації використовується як критерій адекватності моделі, бо є мірою пояснювальної сили незалежності змінної *x*.

Таким чином, коефіцієнт детермінації:

$$R^2 = \frac{\sigma^2_{пояснювальна}}{\sigma^2_{загальна}} \quad (24)$$

або

$$R^2 = \frac{ESS}{TSS} \quad (25)$$

Або в альтернативному вигляді:

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum u_i^2}{\sum (Y_i - \bar{Y})^2}. \quad (26)$$

Визначена таким чином величина r^2, відома як коефіцієнт детермінації, і є мірою якості підгонки лінії регресії, що широко застосовується.

Властивості коефіцієнту детермінації r^2***:***

1. Коефіцієнт r^2 завжди додатній (випливає з виразів (25)-(26)).

2. r^2 має межі $0 \le r^2 \le 1$. При значенні $r^2 = 1$ ми маємо випадок точної підгонки, тобто $\hat{y}_i = y_i$ для кожного *i*. Водночас випадок $r^2=0$ означає відсутність зв'язку між регресантом і регресором (тобто параметр перед *x* - a_1 для всіх *i*). В цьому випадку кращим прогнозом для будь-якої величини *Y* є її середнє значення. При цьому лінія регресії - паралель осі *X*.

Коефіцієнт кореляції *r* кількісно близько пов'язаний з коефіцієнтом детермінації r^2, але концептуально вони дуже різні. Коефіцієнт кореляції можна визначити за формулою

$$r = \pm\sqrt{r^2} \quad (27)$$

або

$$r = \frac{\sum x_i y_y}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}. \quad (28)$$

Між коефіцієнтом кореляції і нахилом a_1 та середнім квадратичним відхиленням $б_x$, $б_y$ існує певний зв'язок. Це дає можливість розрахувати параметри вибіркового рівняння регресії $y = a_0 + a_1 x + u$ через ці величини.

Оскільки

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} \quad (29)$$

$$a_1 = \frac{Cov(x,y)}{\sqrt{Var(x)}} = \frac{Cov(x,y)}{\sigma_x^2} \quad (30)$$

можна записати вираз для коефіцієнта кореляції через параметр a_1 :

$$r_{xy} = \left(\frac{Cov(x,y)}{\sigma_x^2}\right) \cdot \left(\frac{\sigma_x}{\sigma_y}\right) = a_1 \frac{\sigma_x}{\sigma_y} \quad (31)$$

Запишемо формули для розрахунку параметрів економетричної моделі:

$$a_1 = \frac{Cov(x,y)}{Var(x)} = \frac{Cov(x,y)}{\sigma_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sigma_x^2} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x^2} = r_{xy}\cdot\frac{\sigma_y}{\sigma_x} \qquad (32)$$

$$a_0 = \bar{y} - r_{xy}\frac{\sigma_y}{\sigma_x}\cdot\bar{x} = \bar{y} - a_1\cdot\bar{x} \qquad (33)$$

Необхідно відмітити, що при лінійній формі зв'язку коефіцієнт кореляції r_{xy} є оцінкою точності апроксимації, тобто адекватності моделі і дорівнює кореляційному відношенню η_{xy}.

Регресійний аналіз і аналіз дисперсії

Звернемося до регресійного аналізу з погляду аналізу дисперсії.

Раніше нами була доведена така рівність:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (34)$$

тобто TSS=ESS+RSS, яке розкладає загальну суму квадратів (TSS) на два доданки: пояснена сума квадратів (ESS) і сума квадратів залишків (RSS). Вивчення цих доданків у TSS відоме під терміном (*ANOVA*, analysis variance) аналізу дисперсії з погляду регресії (табл. 5.1).

Таблиця 5.1

ANOVA-таблиця для регресійної моделі

Джерело варіації	*Ступені свободи, df*	*Сума квадратів відхилень,* SS	*Середні суми квадратів відхилень, MS*
Регресії	k_1=*m-1*	$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MSE = \frac{ESS}{k_1}$
Залишків	k_2=*n-m*	$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSR = \frac{RSS}{k_2}$
Загальної змінної	*n-1*	$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$MST = \frac{TSS}{n-1}$
SS – сума квадратів (sum squares) MS – середня сума квадратів (mean sum squares)			

З кожною сумою квадратів пов'язані кількість її степенів вільності (свободи) *df* - це кількість незалежних спостережень, на яких вона заснована. Іншими словами це числа, що показують скільки незалежних елементів

інформації зі змінних y_i потрібно для розрахунку відповідної суми квадратів.

Після побудови моделі обчислюється також середня відносна похибка апроксимації, %:

$$\varepsilon = \frac{100}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \qquad (35)$$

Середня похибка апроксимації показує в процентах середнє для всіх значення відхилення результативного показника від розрахункових значень. Модель можна вважати адекватною, якщо середня похибка апроксимації буде знаходитись у межах 12-15%.

Тема 6. Аналіз великих даних в банківській сфері

1. Місце великих даних у фінансовій сфері

2. Приклади використання великих даних в провідних країнах

3. Використання великих даних в банках України

1. Місце великих даних у фінансовій сфері

Фінансові послуги взагалі і банківський сектор зокрема належать до вертикальних ринків, на яких стрімке зростання обсягів даних буде і надалі стимулювати впровадження технологій великих даних. За даними досліджень IDC, в Західній Європі рівень поширення технологій великих даних у фінансовій галузі помітно вище середнього і саме в цій галузі плани подальшого впровадження рішень найбільш численні.

Технології великих даних допомагають фінансовим організаціям більш ефективно вирішувати цілий ряд завдань, таких як:

Поліпшення взаємодії з клієнтами. Обмежуючись лише масовим просуванням продуктів, багато банків досі не враховують потреби окремих груп клієнтів. Як наслідок, вони витрачають зайві ресурси, намагаючись продавати не той продукт не тому клієнту. Для вирішення цієї проблеми рекомендується застосування технологій великих даних, які дозволяють ефективно обробляти великі обсяги даних і отримувати корисну інформацію.

Забезпечення відповідності законодавству та галузевим стандартам. Ця проблема не втрачає своєї актуальності для банківського сектора. Вимоги PSD2, Basel III, FACTA - лише деякі приклади нормативи, які продовжують істотно впливати на витрати банків на оптимізацію операційної діяльності. Крім цього, існують і національні нормативи - наприклад, циркуляр Банку Італії №263 зі змінами до вимог інформаційної безпеки і безперервності бізнес-процесів.

Модернізація базових банківських систем. Багато європейських банків

потребують ґрунтовному оновленні своїх базових систем, але повністю замінити їх новими системами неможливо через ряд причин. Цей процес буде проходити поетапно, за модульним принципом, починаючи з найбільш застарілих систем.

Впровадження мобільних рішень. Сюди входять мобільний банкінг, мобільні платежі, платежі з використанням технології NFC. Можливість виконання банківських операцій на мобільних пристроях стає однією з найбільш затребуваних послуг, і більшість банків вже займаються розробкою нових інструментів. Очікується, що інвестиції в мобільні послуги триватимуть, але проблеми з доступом до даних, безпекою і цілісністю даних залишаються значними.

Технології великих даних впроваджуються в багатьох галузях - від цільового мобільного маркетингу до виявлення ознак шахрайства та від управління коштами до залучення клієнтів

Фахівці з банківського сектора говорять про те, що віддача від інвестицій в технології великих даних для поліпшення управління персоналом, залучення клієнтів, підвищення операційної ефективності, оптимізації процесів і виявлення ризиків перевершує очікування. До найбільш значущих напрямків використання технологій великих даних відносяться наступні:

- підвищення операційної ефективності;
- поліпшення якості обслуговування клієнтів;
- управління ризиками і дотримання вимог законодавства.

Ось тільки кілька прикладів використання технологій великих даних європейськими фінансовими організаціями: алгоритмічна торгівля (тобто система біржової торгівлі, забезпечує підтримку прийняття рішень про проведення транзакцій на фінансових ринках із застосуванням різноманітних математичних інструментів), аналіз уподобань (аналіз великих обсягів неструктурованих даних, таких як коментарі та дописи в соціальних мережах, з метою оцінки ставлення до тих чи інших брендів, організацій і т.д.) і аналіз чинників впливу (застосування аналітичного інструментарію для передбачення того, який фактор з найбільшою ймовірністю вплине на рішення клієнта).

2. Приклади використання великих даних в провідних країнах

Фінансові організації отримують величезні обсяги даних з тисяч різних джерел, і для управління ними зазвичай потрібне спеціальне програмне забезпечення. Звичайний інструментарій, орієнтований на інтелектуальну підтримку бізнесу, призначений для роботи зі структурованими даними, тоді як нове покоління засобів аналітики призначається для роботи з неструктурованими даними з різних джерел. Аналітика, заснована на великих даних, забезпечує оперативний доступ до інформації про бізнес-процесах, подіях і операціях, банк отримує можливість негайно розсилати відповідні повідомлення, оновлювати інформаційні зведення для керівництва,

пропонувати стимули готовому покинути його клієнту, перенастроювати банківське обладнання та запобігати шахрайству.

Розглянемо приклад з UniCredit Business Integrated Solutions (UBIS). Через відсутність єдиної точки доступу UBIS доводилося витрачати чимало часу на аналіз даних з багатьох джерел.

Для вирішення цієї проблеми UBIS звернулась до компанії Splunk, постачальника хмарного програмного забезпечення і послуг для пошуку, моніторингу, аналізу та візуалізації великих даних, що надходять від веб-сайтів, додатків, серверів, мереж, датчиків і мобільних пристроїв. Обробка даних, що надходять в режимі реального часу, а також аналіз раніше накопичених обсягів за допомогою рішень Splunk дозволили UBIS моментально виявляти проблеми і проводити профілактику форсмажорних подій. Зокрема, UBIS використовує Splunk для моніторингу транзакцій із заданою періодичністю і видачі сигналів тривоги при досягненні встановлених порогів або настанні певних умов. Завдяки профілактичному моніторингу групі обслуговування клієнтів вдалося значно підвищити якість обслуговування.

Сьогодні бізнес-аналітики UBIS використовують програмне забезпечення Splunk для складання щотижневих звітів для керівництва, що відображають поточну ситуацію за такими виробничими показниками, як:

– кількість клієнтів, що обслуговуються в відділеннях банків з використанням мобільного і інтернет-банкінгу;
– кількість нових відкритих банківських рахунків;
– кількість обслуговуваних позик / кредитних карт;
– кількість проведених платежів / банківських транзакцій.

Для відображення результату операцій мобільного банкінгу на спеціалізованих інформаційних панелях в рішеннях Splunk використовуються карти Google Maps.

Аналіз великих даних відкриває нові можливості підвищення операційної ефективності завдяки доступу в режимі реального часу до інформації. Це дозволяє приймати обґрунтовані рішення і витрачати менше часу на непродуктивні ручні операції. Крім того, застосування засобів великих даних дозволяє перерозподіляти ресурси на користь більш важливих завдань.

Aareal Bank (Німеччина), в 2012 році зайнявся реорганізацією своїх корпоративних даних з метою поліпшення можливості їх аналізу.

Для вирішення цих завдань Aareal Bank звернувся до корпорації SAP. Запропоноване SAP рішення полягало в модернізації вже використовуваного банком додатку SAP NetWeaver Business Warehouse (BW) і перехід на платформу зберігання і обробки даних SAP HANA.

Після реалізації проекту Aareal Bank отримав такі конкретні вигоди:

– отримання більш повного і глибокого уявлення про стан справ в таких галузях, як аналіз продуктів, управління ризиками та фінансами,

планування кредитних портфелів і звітність, без звернення за допомогою до відділу інформаційних технологій;

– прискорення підготовки звітів для керівництва на 67%;

– скорочення термінів обробки запитів користувачів в чотири рази;

– підвищення продуктивності праці на 50-70% завдяки скороченню часу очікування відповіді системи, що дозволяє більш оперативно приймати обґрунтовані рішення;

– зниження обсягів сховища даних більш ніж в десять разів

Технології великих даних часто використовуються для ефективного залучення клієнтів. В цьому випадку використовуються дані про результати проведених маркетингових кампаній, дані з систем взаємини з клієнтами (CRM). Аналіз цих даних, а також даних з соціальних мереж, дозволяє виробляти персональний підхід до кожного клієнта або групі клієнтів.

Банк Raiffeisen Bank Austria (RBA) (Хорватія) для підвищення ефективності цільових маркетингових кампаній почав аналізувати клієнтську базу з використанням технологій великих даних. Для цього було вибрано нове CRM-рішення американського постачальника програмних засобів бізнес-аналітики SAS, що включає модель даних банку, а також засоби аналітики для сегментації клієнтської бази, продажу супутніх продуктів, пропозиції альтернативних продуктів і утримання клієнтів. Після впровадження нового CRM-рішення банк зміг побудувати передбачувальні моделі для обслуговування кредитних карт і позик, а потім приступити до запуску кампаній для тестування нових підходів до сегментації клієнтської бази. Коли якість моделей підтвердилося, банк зміг змінити свою маркетингову політику від орієнтації на окремі продукти до орієнтації на клієнта і почати пропонувати клієнтам продукти відповідно до їх потреб і поточними стосунками з банком.

Моніторинг і аналіз даних про банківські транзакції може стати ключовим моментом до підвищення якості обслуговування клієнтів. Одним з ефективних способів отримання інформації про стан програмного забезпечення, що працює на великій кількості різнорідних апаратних платформ, що використовуються клієнтами банку, може виявитися впровадження рішення по збору та аналізу великих даних. В результаті такого аналізу з'являється можливість оперативного виявлення і усунення недоліків, підвищення якості обслуговування клієнтів.

Так італійський банк Sinergia вирішуючи питання найбільш ефективного способу доступу до банкоматів співпрацює з корпорацією NCR і постачальником рішень для великих даних і аналітики Inetco.

Ці компанії запропонували об'єднати рішення для управління банкоматами APTRA Vision виробництва NCR і платформу моніторингу транзакцій і аналітики Insight, розроблену Inetco.

APTRA Vision надає функції моніторингу обладнання, необхідні Sinergia для підтримки в робочому стані зростаючого парку банкоматів різних

виробників. INETCO Insight забезпечує відсилання повідомлень в режимі реального часу і повну інформацію по клієнтським транзакціях, що ініціюється через банкомат або будь-яку іншу точку контакту з клієнтом, таку як мобільний додаток, інтернет-кіоск або система електронного банкінгу.

Дане рішення дозволяє зменшити число проблем з виконанням транзакцій, знизити час непрацездатності банкоматів або неможливості їх використання, а також виявляти недоліки в роботі обладнання або програмного забезпечення ще до отримання скарг від клієнтів і визначення підозрілих операцій. Це призводить до підвищення як операційної ефективності, так і рівня задоволеності клієнтів.

Використання технологій великих як інструменту управління ризиками дозволяє відстежувати поведінка клієнтів з метою виявлення підозрілої активності, запобігати шахрайство, а також більш ефективно готувати звіти про діяльність організації для контролюючих органів. Отримання цілісної картини стану справ в організації і забезпечення доступу до архівувати даними дозволяють скоротити витрати на аналіз і прогнозування.

В банку Rabobank Nederland (Нідерланди) постало питання, що до отримання повної картини руху даних і визначенням того, як модифікація однієї з систем позначиться на роботі інших. Ключовим компонентом рішення Rabobank стало створення центрального пункту управління всіма даними Data Portal Financing (DPF). Основний для нього стала платформа інтеграції даних PowerCenter, розроблена американською компанією Informatica. після впровадження цієї платформи все системи стали обмінюватися даними через портал DPF. новий портал забезпечив краще представлення даних і дозволив виконувати набагато більш детальний аналіз, використовуючи більш надійну інформацію, зібрану з різних систем. Аналіз тенденцій на ринку нерухомості і динаміки попиту на нові продукти дозволив Rabobank швидко і точно коригувати свої пропозиції ринку. Дотримання нормативних вимог значно спростилося за рахунок більш наочного відображення потенційних ризиків.

3. Використання великих даних в банках України

Банківський сектор в Україні займає провідні позиції у використанні великих даних: банки володіють великим масивом інформації. Ці знання відкривають великі можливості для розвитку, але щоб ними скористатися, потрібен спосіб швидкої обробки, оскільки вручну проаналізувати весь масив практично неможливо.

Важливо зазначити, що використовується статистична інформація й не порушуються закони України «Про Персональні дані» і «Про інформацію». Для того, щоб структурувати, проаналізувати й отримати максимум користі від цих даних, і використовується великі дані.

Найчастіше в банківській сфері система великих даних застосовується для вивчення наявних і потенційних клієнтів банку. Наприклад, банки

проводять обробку інформації щодо покупок, соціальних мереж людей, аналізують транзакції. Так з'являється можливість пропонувати актуальний банківський продукт і виключити ненадійних клієнтів. Завдяки всьому цьому банку вдається підвищити свій прибуток.

Сфери застосування великих даних в банківському секторі України.

Основні функції великих даних у банках — це скоринг, запобігання шахрайства та аналіз аудиторії.

1. Аналіз аудиторії. Вивчення цільової аудиторії допомагає створювати популярні пропозиції й оптимізувати їхнє просування. За запитом бізнесу експерти Київстар можуть розробити портрет клієнта, щоб краще пізнати цільовий сегмент, його ключові характеристики та переваги. A Look-alike аудиторія («пошук схожих») допоможе в залученні нових клієнтів і підвищенні ефективності реклами.

2. Скоринг на основі великих даних в банках. Використовуючи скоринг на базі великих даних від Київстар, можна точніше й об'єктивно оцінити потенційних клієнтів на предмет надійності та платоспроможності, навіть якщо в них немає кредитної історії, і виявити ризик шахрайства.

3. Теплові карти й геоаналітика. Ці інструменти допоможуть знайти оптимальні місця для нового відділення, банкомата або термінала, на основі характеристик центрального апарата й місць її скупчення.

4. Таргетована розсилка. Суть таргетованої розсилки в тому, щоб донести інформацію щодо послуг, акцій та товарів тим, кому це цікаво. Для налаштування таргетингу можна вибирати різні сегменти центрального апарату, спираючись на захоплення, географічне положення, вік тощо.

Аналіз аудиторії банку за допомогою великих даних. Що більше бізнес знає про «болі» та бажання своїх наявних і потенційних клієнтів, то з більшою ймовірністю зможе утримати перших і залучити других. Адже серед безлічі пропозицій, представлених на ринку, споживач буде шукати варіант, який максимально відповідає його вимогам. І завдання банку — розробити такий продукт і вчасно запропонувати його зацікавленим людям.

На підставі великих даних від Київстар експерти складають портрет клієнта, щоб краще пізнати потреби й побажання центрального апарату. Для створення портрета Data-фахівці вивчають і аналізують великі дані, визначаючи важливі закономірності. Отримані результати допомагають у розробленні релевантних пропозицій і в запуску ефективнішої реклами для залучення нових клієнтів.

Також великі дані від Київстар дає можливість скористатися інструментом Look-alike — «пошук схожих». Дуже добре підходить для банківської сфери, оскільки ґрунтується на даних з наявної клієнтської бази. Суть процесу в тому, щоб знайти цільову аудиторію, схожу на вже наявних клієнтів. Для цього дата-фахівці, використовуючи методи аналізу Великих даних і машинне навчання, знаходять ключові особливості наявних клієнтів і на

їх основі визначають аудиторію, у якій є збіги з клієнтами банку за тими чи іншими критеріями.

Крім того, під час роботи з великими даними можна розрахувати величину життєвого циклу потенційного клієнта (CLV), отже, дізнатися, наскільки вигідним буде співпраця з ним. Так можна прогнозувати свої доходи й постійно їх збільшувати, залучаючи й утримуючи перспективнішу аудиторію.

Скоринг із застосуванням великих даних від Київстар. Щоб зменшити кількість проблемних позичальників банків, дата-фахівці Київстар розробляють скорингові моделі на основі великих даних. Водночас використовують дані від телеком-оператора, які актуальніші, оскільки на відміну від інформації з Бюро кредитних історій, постійно оновлюються.

Отже, переваги скорингу із застосуванням великих даних від Київстар:

– Створення скорингової моделі на базі актуальної інформації.

– Висока швидкість обробки даних.

– Поліпшення ситуації з видаванням кредитів клієнтам без кредитної історії.

– Надання кредитного та фінансового скорингу.

– Антифрод-скоринг для відсіювання шахраїв.

– Запобігання шахрайства в банках за допомогою великих даних.

Одне з основних напрямів у роботі банків — захист даних і запобігання шахрайства. Для цього використовують антифрод-скоринг, спрямований на виявлення неблагонадійних клієнтів.

Антифрод-скоринг на основі великих даних від Київстар дає можливість охопити більше інформації та зробити прогнози точнішими. Він ефективно працює навіть у таких ризикових фінансових послугах, як мікрокредитування і кредитування онлайн.

Отже, великі дані в банках допомагає за короткі терміни аналізувати й систематизувати великі масиви інформації, познайомитися ближче з цільовою аудиторією, дізнатися, де шукати нових клієнтів і як налагодити з ними комунікацію. Також великі дані від одного з найбільших телеком-операторів України — Київстар — це надійне джерело актуальної інформації для створення ефективних скорингових моделей, розроблення нових продуктів з урахуванням особливостей цільового сегмента й маркетингових кампаній для просування бізнесу.

Тема 7. Аналіз великих даних в державному управлінні та соціальній сфері

1. Цифрова соціологія

2. Технології великих даних в управлінні просторово-економічним розвитком міста і регіону

3. Джерела відкритих даних в Україні

1. Цифрова соціологія

Цифрова соціологія виникла у відповідь на певний ажіотаж навколо того, як «нові дані» перетворюють способи пізнання суспільства. Вона розуміється як «обчислювальна соціальна наука». Деякі критично налаштовані автори ототожнюють її з певною формою аналізу даних. Але цифрова соціологія пропонує альтернативу вузьким визначенням цифрових соціальних досліджень. Навпаки, цифрові соціологи прагнуть досліджувати набагато ширший набір взаємодій між даними, людьми та технологіями, які переповнюють, перевищують і не «вписуються» у просту історію про нові форми аналізу даних, що займають місце таких старих методів соціальних досліджень, як опитування чи польові роботи.

Цифрова соціологія, як вважають її апологети, призначена зрозуміти сучасний даніфікований світ, який потребує нових способів мислення про соціальне. Її послідовники розробляють поняття, інструменти та практики для аналізу перетину соціального та цифрового. Мета цифрової соціології – досліджувати закономірності соціального життя сучасної людини, інтегрованої у цифровий інтернет-простір. Об'єкт – цифрове суспільство як нова соціокультурна реальність. Предмет – соціальні відносини, що виникають у цифровому середовищі, цифрове соціальне життя, що включає в себе різноманітні соціальні феномени, що виникають у цифровому середовищі, а також їхній взаємозв'язок з матеріальною соціальною реальністю. Проте з самого початку свого формування цифрова соціологія зазнавала найбільшої критики у зв’язку з відсутністю теоретичного підґрунтя. Сьогодні можна константувати, що вона так й не знайшла шляхів перетворення у форму спеціальної соціологічної теорії. Натомість, соціологія поступово акумулює всі методи, розроблені цифровими соціологами, що сприяє широкій зацікавленості до напрямку, що отримав назву «цифрові методи» та по суті є розгорнутою методологією проведення соціологічних досліджень на основі великих даних, яка наразі вже розглядається «не як інновація, а як мейнстрім».

Лідером розвитку цифрових методів є провідна європейська дослідницька група «The Digital Methods Initiative» (DMI), яка складається з соціологів та дослідників нових медіа, що розробляють методи та інструменти перепрофілювання інтернет-пристроїв і платформ (таких як Twitter, Facebook, Google) для дослідження соціальних та політичних питань.

Соціальні наслідки – третій аспект великих даних. Вплив великих даних на суспільство зазвичай описують через історії успіху стартапів, які впроваджують технології Big Data. Проте цим вплив великих даних не обмежується. Піонер у галузі великих даних Алекс Петланд писав: «З Big Data ми можемо почати реально розглядати деталі соціальної взаємодії та те, як вони розігруються і більше не обмежуються такими середніми показниками, як ринкові індекси або результати виборів. Це приголомшлива зміна. Можливість

бачити деталі ринку, політичні революції та вміти передбачати та контролювати їх – це, безумовно, можна порівняти з даром Прометею, який може бути використаний як для добра, так й для зла».

Перші ж спроби дослідження великих даних з метою соціологічного аналізу призвели до необхідності вирішення низки методологічних питань, а також питань «перепрофілювання» методів обробки онлайнових даних, які використовуються інтернет-платформами, з метою вирішення соціологічних завдань. У результаті сформувався підрозділ соціологічної науки – цифрова соціологія, яка фокусується на розумінні використання великих даних як частини повсякденного життя.

У літературі багато прикладів позитивних наслідків застосування великих даних, але «темний бік великих даних» до останнього часу дуже мало дискутувався. Чому? Мабуть, тому, що сутність феномену великих даних й досі не повністю осягнута, досі немає розуміння, що великі дані стосуються не тільки комп'ютерних учених, а всіх, навіть тих, хто про них нічого не знає. Ми всі вже перетворилися на «суб'єктів даних». «Дані не лише формують наші соціальні відносини, вподобання та життєві шанси, але й наші демократії».

Останнім часом голосно звучать заклики до встановлення меж втручання великих даних у людське життя. Цьому значно сприяв нещодавній скандал, пов'язаний з Facebook, коли особиста інформація 87 мільйонів користувачів потрапила до компанії Cambridge Analytica (CA), яка використала могутність великих даних з метою впливу на думку виборців шляхом точно розрахованої персоналізованої реклами. Цей скандал має позитивний бік: «Хороша річ, що стосується скандалу з CA, – це те, що він розпочав тривалу дискусію щодо використання даних соціальних медіа».

Існує безліч ентузіазму та оптимізму щодо того, як уряди всіх рівнів можуть використовувати великі дані, алгоритми та штучний інтелект. Проте одночасно зростає стурбованість ризиками, які виникають із цими новими системами. Оскільки уряди широко використовують великі дані, існує низка реальних та потенційних наслідків, про які треба знати та бути напоготові: 1) заохочується перманентний збір даних про громадян, у зв'язку з чим громадяни перетворюються на «суб'єктів даних»; 2) все більше поширюються автоматизовані сервіси прийняття управлінських рішень, алгоритми яких невідомі широкому загалу; 3) громадяни стають «безперервно пізнаними», але мають мало можливостей дізнатися, як їхні дані збираються та використовуються; 4) акценти зміщуються від причинного зв'язку до кореляції, від запобігання до випередження на основі прогнозів, зроблених штучним інтелектом; 5) заохочується розширення партнерства між державним та приватним секторами, а також участь IT-корпорацій в управлінні та наданні державних послуг.

Зрозуміло, що без належної прозорості, підзвітності та контролю системи великих даних можуть бути використані таким чином, що порушує

громадянські права. Ключовим тут є визнання, що при великих даних несприятливі наслідки можуть бути навіть не навмисними зловживаннями, а просто «глюком системи». На даний момент громадяни не мають інформації або ресурсів, необхідних для змістовної взаємодії з цими змінами, які вже відбуваються.

Приклад Китаю, який з 2014 року запроваджує систему соціального рейтингу, яку ще називають системою соціального кредиту, просто жахає. Згідно з «Програмою створення системи соціального кредиту», вже до 2020 року в країні повинна бути запущена система оцінки кожного громадянина, що буде працювати в режимі реального часу, а результати її роботи мають публікуватися в інтернеті. Артикульована мета програми – побудова суспільства, в якому «бути чесним і вартим довіри стане престижним й бажаним». Критерії оцінювання обирає державна влада, яка має право визначати, чи варто вам довіряти. На додачу, рейтинг буде публічною інформацією, від якої залежатиме, чи отримаєте ви кредит, нову посаду, можливість подорожувати і навіть підуть з вами на побачення чи ні.

Алгоритм обчислення рейтингу повністю не розкривається, але оприлюднено 5 головних факторів, що впливають на рейтинг: 1) кредитна історія; 2) здатність виконувати взяті на себе зобов'язання; 3) верифіковані особисті дані, наприклад, адреса чи номер мобільного телефону; 4) особисті переваги та поведінка; 5) стосунки між громадянами.

Як же нас бачить штучний інтелект очима своїх алгоритмів? Якщо ви вчасно не оплатили комунальні рахунки або кредит, то ваш соціальний рейтинг знижується. Система не бачить і не враховує контексти, наприклад, що рахунки не оплачені, оскільки ви лежали в лікарні. Крім того, штучний інтелект оцінює людей за типами продуктів, які вони споживають. Так, особа, що проводить багато часу за відеоіграми, вважатиметься ненадійною. Ті, хто часто купує підгузки, імовірно мають маленьку дитину, отже вважаються відповідальними. П'ятий фактор означає врахування те тільки поведінки певної особи, а й всіх її онлайн-друзів, тобто буде враховуватися, якими словами і про що ви переписуєтеся у чатах та соціальних мережах, які повідомлення та коментарі залишаєте (на підтримку уряду, чи ні, схвалюєте економічний стан країни, чи ні). Навіть ставити «лайк» громадянам Китаю доведеться дуже обережно, оскільки він може змінити їхній особистий соціальний рейтинг довіри. Таким чином, система спрямовано схилятиме громадян уникати сценаріїв поведінки, які не подобаються уряду.

У західних ІТ-виданнях було багато критики стосовно китайської системи соціального рейтингу. Проте ми не маємо ніяких гарантій, що в інших країнах не створюються передумови для впровадження подібних систем. Китайський уряд хоча би повідомив громадян про свої «соціальні інновації». А тим часом алгоритми Facebook знають наших друзів та можуть розпізнати нас на фотографії. Google взагалі « знає про нас більше, ніж ми знаємо про себе». Ще

у 2015 році OECD (Організація економічного співробітництва та розвитку) підрахувала, що на 100 жителів США припадає не менше 24,9 підключених пристроїв. Всі компанії, які виробляють і обслуговують такі гаджети, збирають безліч інформації про користувачів та їхні звички, щоб розуміти і прогнозувати онлайн-поведінку. «Уряди більшості країн займаються моніторингом. Американське АНБ навіть і не заперечує масштабного стеження за всіма громадянами. Жителі США вже живуть у світі, де алгоритми вирішують, чи є особа небезпечною або походить із зони ризику, чи є вона порядним громадянином та чи можна їй довіряти. Захід набагато ближче до Сходу, який трансформує кредитний рейтинг у соціальну систему оцінювання SCS. От тільки, на відміну від китайців, американців про це офіційно не повідомили».

Якщо не бути пильними та не втручатися у формування даніфікованого суспільства, заснованого на великих даних, ми можемо опинитися в суспільстві безпрецедентної диктатури, суспільстві, де стеження є нормою, а життя пересічних громадян перетвориться на нескінченний конкурс популярності, в якому всі, окрім обраних, змагаються за високий рейтинг, поступаючись власними принципами і вподобаннями.

Як вплинути на ситуацію? Перш за все, потрібна повна відкритість, надійні та прозорі механізми, що управлятимуть процесом і забезпечать належне використання інформації. Щоб довіряти системі, вона не повинна бути «чорним ящиком», в ній не має залишитися жодного нерозкритого елементу, а також не має бути можливостей корегування рейтингів для обраних. Крім того, потрібно встановити рейтинг довіри для самих оцінювачів.

Підводячи підсумок, зазначимо, що великі дані – це не тільки джерело інформації про сучасне суспільство, а й інструмент його дослідження. Цей інструмент поки що далеко не повною мірою опанований соціологами. Проте робота в цьому напрямку йде, лише хотілося би, щоб наша країна активніше залучилася до неї. Безумовно, треба бути відкритими для інновацій, критично осмислювати нові епістемологічні підходи, активніше впроваджувати практику соціологічного аналізу на основі великих даних. Все так, проте, на нашу думу, більш нагальною проблемою на сьогодні є необхідність визнати, що даніфікація немунича, вона вже відбулась. І якщо дані не використовуються кимось (владою чи корпораціями) з корисною метою, то це не означає, що такої можливості нема. У зв’язку з цим соціологічна спільнота повинна залучитися до розробки справедливої політики даних.

2. Технології великих даних в управлінні просторово-економічним розвитком міста і регіону

Технології Big Data мають великі перспективи для застосування не тільки в економіці та бізнесі, а й у державному управлінні. При цьому, органами державної влади, регіонального управління та місцевого самоврядування має бути усвідомлення того, що ефективність їхнього управління залежить не так

від наявності даних, як від їх повноти, достовірності, візуалізації, своєчасного надання. Потрібно перейти від механізму традиційної аналітики до прогнозної, з використанням моделей сценарного розвитку. Саме аналітика такого типу дозволить побачити і проаналізувати можливі результати та зміни у випадку прийняття тих чи інших управлінськихрішень.

Потенціал реалізації Big Data у сфері державного, регіонального та місцевого управління досить значний. Серед можливих напрямів застосування технології великих даних в органах державної влади варто виділити:

1) Оцінка громадської думки про ефективність заходів щодо реалізації соціально-економічної політики:

- здійснення якісного та кількісного порівняння публікацій в ЗМІ про результати впровадження управлінських рішень;
- відстеження динаміки зміни громадської думки щодо діяльності органів державної влади;
- визначення характеру публікацій в ЗМІ про діяльність органів державної влади по конкретних напрямах діяльності, оцінювання адекватності її відображення в публічному просторі;
- аналіз реакції аудиторії в соціальних мережах, відвідувачів сайтів на оприлюднену інформацію про впроваджені управлінські рішення.

2) Відстеження громадської думки про діяльність органів державної влади:

- впізнаваність і розуміння основних завдань та напрямів діяльності;
- еволюція образу органу влади або окремого публічної особи.

Вважаємо, що технології аналізу великих даних потрібно використовувати не тільки для аналізу результатів діяльності органів влади, а й для забезпечення ефективності управління розвитком регіону чи міста. Тому технології Big Data доцільно використовувати для:

1) Розвитку інфраструктури з переходом від кількісної забудови територій до якісної.

Будівництво нового житла в містах повинне здійснюватися там, де вже є сформований інфраструктурний комплекс, або буде можливим його формування у перспективі. Покращення якості життя і підвищення її комфорту (розвинена інфраструктура, зручність громадського транспорту тощо) має бути прерогативою сфери житлового будівництва.

2) Визначення основних потреб і настроїв населення.

Big data дозволяє аналізувати інформаційний масив з різних джерел (комп'ютери, мобільні пристрої, Інтернет, соціальні мережі), на основі чого можна отримати якісні знання про потреби і можливості людей, які проживають на певних територіях регіону. Тобто створити інформаційний портрет середньостатистичного жителя міста чи регіону.

3) Моніторингу використання бюджетних коштів на підтримку певних галузей.

4) Оцінювання міграційних процесів.

Межі бізнесу, міст, регіонів та держави все більше і більше будуть розмиватися. Яскравими прикладами чого є відвідування сільськими дітьми садочків та шкіл у містах, відсутність обов'язковості прописки за місцем проживання породжує процеси переміщення населення між містами, регіонами та унеможливлює адекватність його оцінки тощо.

Наприклад, на основі аналізу даних про переміщення абонентів мережі Київстар чи будь-якого іншого оператора мобільного зв'язку можна проаналізувати туристичні та міграційні потоки населення між сільської та міської місцевостями в межах одного регіону, між регіонами. В подальшому порівняти із офіційними статистичними даними. Результати моніторингу дадуть можливість місцевій владі удосконалити маркетингову політику туристичних напрямків, регіональну політику розвитку депресивних територій тощо.

5) Оптимізації різних сфер економіки регіону, наприклад, житлово-комунальної сфери.

Із розвитком технологій Big Data стає можливим проведення безперервного моніторингу і управління цілими галузями.

Використання Big Data для контролю та управління житлово-комунальним господарством, що дозво-лить зменшити рівень заборгованості за комунальні послуги, ефективно використовувати газ в опалювальний період, оновлювати старий житловий фонд, визначати проблемні зони тощо.

6) Покращення життєдіяльності міст шляхом оптимізації потоків транспорту.

Завдяки використанню аналізу даних можливість поєднання транспортної інфраструктури з іншими видами комунальних послуг в єдине ціле стає більш реальним. Такий підхід зменшить затори на дорогах, завантаженість доріг, узгодить між собою курсування різних видів транспорту.

7) Виявлення загроз економічній безпеці регіону.

Модель великих даних дозволяє реалізувати багато ефективних методів моніторингу для виявлення загроз економічній безпеці регіону.

8) Визначення та врахування можливих наслідків змін у природному, соціально-економічному середовищі регіону, їх вплив на регіональний просторовий розвиток.

9) Виявлення територій перспективного розвитку.

У Концепції «Закону України «Про території перспективного розвитку», схваленій розпорядженням Кабінету Міністрів України від 13 січня 2010 р. №110-р зазначається, що однією із основних проблем, регіонального та місцевого розвитку є поглиблення диспропорцій соціально-економічного розвитку територій. Одні території розвиваються прискореними темпами і там концентруються фінансові, інноваційні та трудові ресурси, інші − внаслідок ліквідації підприємств, зменшення обсягів виробництва, відсутності

інфраструктури належного рівня мають значні соціально-економічні проблеми та потребують державного стимулювання для активізації економічної діяльності. Такі території є непривабливими для інвесторів, через що відчувають дефіцит ресурсів для розвитку. В результаті чого на таких територіях спостерігаються такі негативні явища, як економічний занепад, зростання безробіття, руйнування виробничої та іншої інфраструктури.

Критеріями, за якими можна визначити території перспективного розвитку є:

- наявність економічно активного населення;
- високий рівень безробіття;
- високий рівень дотацій з державного бюджету місцевим бюджетам;

наявність земель несільськогосподарського призначення для створення спеціальних митних зон та відповідної інфраструктури з метою залучення інвестицій для забезпечення розвитку виробництва, оптової торгівлі, транспорту, туризму тощо.

Для виявлення територій перспективного розвитку потрібно опрацювати та проаналізувати великий масив статистичних даних, щоб якомога краще на їх основі оцінити всі сфери життєдіяльності міста чи регіону та визначити пріоритетні сфери підприємницької діяльності. Разом з тим застосування сучасних технологій Big Data дозволить виділити зони активного споживання товарів і послуг, які можна буде активно розвивати на конкретній території.

Зазначим, управління розвитком підприємницької діяльності в регіоні на основі Big Data повинне складатися із декількох етапів, серед яких визначення ролі і місця конкретної сфери підприємницької діяльності у розвитку території, визначення основних видів послуг, що будуть надаватись у регіоні, створення портрету споживача тих послуг з використанням математичного моделювання (кореляційно-регресійний аналіз, імітаційне моделювання тощо), створення інформаційної моделі споживача послуг, формування зон сфери підприємницької діяльності в регіоні, розробка рекомендацій щодо прийняття управлінських рішень (рис. 7.1).

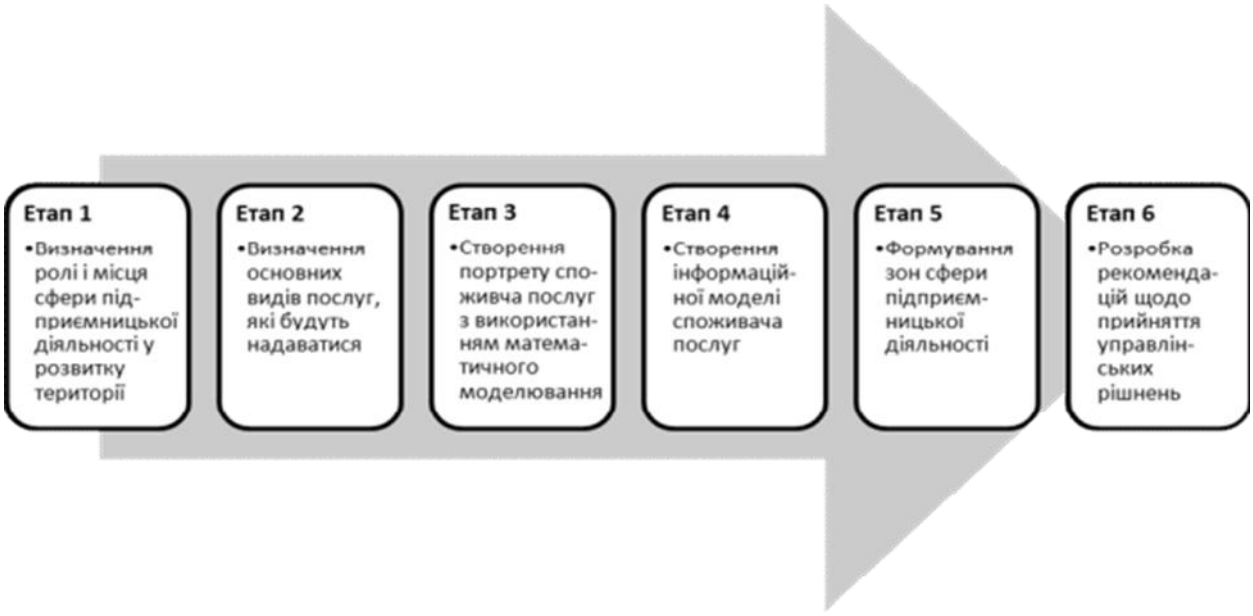


Рис. 7.1. Етапи процесу застосування Big Data для виявлення пріоритетних сфер підприємницької діяльності на територіях перспективного розвитку

Як вже зазначалося, сучасне середовища міста чи регіону характеризується накопиченням великих даних, що потребують обробки з метою прийняття обґрунтованих управлінських рішень з управління містом чи регіоном. Застосування технологій Big Data концептуально змінюють підхід до управління, переорієнтовуючи з конку-рентних переваг регіону до виявлення раніше не визначених точок розвитку. На цій думці сходяться більшість науковців, які займаються дослідженням впливу інформаційних технологій на розвиток економік регіонів та держави. Впровадження інформаційних технологій швидше за все потрібно розглядати не як конкурентну перевагу, а як джерело загроз з боку конкурентів. Адже, фактори і технології успішного розвитку одних територій приваблюють інвесторів для суб'єктів-конкурентів, які у своєму розвитку рухаються в тому ж напрямку. Тому визначення нових точок розвитку є більш суттєвою основою для збільшення швидкості розвитку і відриву від суперників. Основою для цього мають бути не результати аналізу даних за поточний і минулий періоди, які вже давноне є відображення ростучи стагнації, а саме прогнозні значення.

Крім того, практичне застосування інформаційних технологій в управлінні повинне змінювати фокус управлінського впливу на вирішення найбільш актуальних задач в максимально короткі терміни. Інакше механізм регіонального управління не буде відповідати новому формату суспільних очікувань, що поглиб-люватиме проблеми в низці галузей економіки та в соціальній сфері.

В Україні з прийняттям Концепції реформування місцевого самоврядування та територіальної організації влади в Україні, схваленої розпорядженням Кабінету Міністрів України від 1 квітня 2014 р. № 333-р [7],

ухваленням Закону України «Про співробітництво територіальних громад» від 17 червня 2014 року№ 1508-VII [6] та Закону України «Про добровільне об'єднання територіальних громад» від 5 лютого 2015 року № 157-VIII [5] розпочато процес децентралізації, в результаті чого громади здобувають ресурс, фінанси та повноваження для забезпечення повноцінного розвитку територій, створення комфортного та безпечного середовища для життя. Паралельно із отриманням ресурсів для вирішення економічних, соціальних та екологічних проблем у влади на місцях зростає відповідальність перед суспільством, через що нею має бути обрано ефективний та прозорий механізм для управління, в якому інноваційним способам збору та аналізу даних за технологією Big Data має бути відведено чільне місце. На відміну від статистичних вибірок для дослідження соціально-економічного становища, які встигають застаріти до моменту їх аналізу, «великі дані» можуть оброблятися в режимі реального часу, що підвищує якість і швидкість прийняття управлінських рішень. «Великі дані» в області міського та регіонального управління доповнюють традиційні типи інформації про місто, регіон та розширюють сферу їх застосування.

Так, завдяки «великим даним» стає можливим моніторинг поведінкових моделей і аналіз міського способу життя, способу життя на різних територіях регіону на перетині таких категорій як населення, економічний роз-виток, забудова населених пунктів, інфраструктура і т. д. Їх результати можуть бути покладені в основу розробки стратегії розвитку міста, стратегії регіонального розвитку, комплексних програм соціального економічного розвитку території, малого і середнього підприємництва, освіти, туризму та рекреації, інфраструктури тощо.

Очевидно, що для аналізу даних по технології Big Data потрібна відповідна технологічна інфраструктура, у якій виділяють три рівні. Перший рівень інфраструктури забезпечує збір даних з різних джерел (сенсори температури і чистоти повітря, відеокамери, зчитувальні пристрої в громадському транспорті, мобільні телефони, реєстри звернень громадян до органів влади тощо) в рамках міста чи регіону. Залежно від типу джерела використовується відповідний інструмент для збору даних.

Другий рівень інфраструктури включає інструментарій для зберігання і обробки даних з метою побудови прогнозів, визначення взаємозв'язку між різними потоками інформації тощо. Для зберігання і обробки даних, в залежності від завдань використовують як платформи класу Hadoop, так і традиційні (реляційні) сховища даних.

На третьому рівні знаходяться платформи з відкритими даними, інструменти візуалізації даних, вітрини даних, апаратні системи для прогнозування, моделювання та моніторингу, автоматизовані системи реагування на ті чи інші події, які використовуються органами влади для обміну даними і для прийняття рішень на їх основі.

Процес формування і ухвалення управлінських рішень щодо просторового розвитку регіону (міста) із застосуванням технологій Big Data доволі складний і багатосторонній та складається з декількох стадій. Загальну схему процесу прийняття управлінських рішень на регіональному (місцевому) рівні з використанням технології Big Data наведено на рис. 7.2.

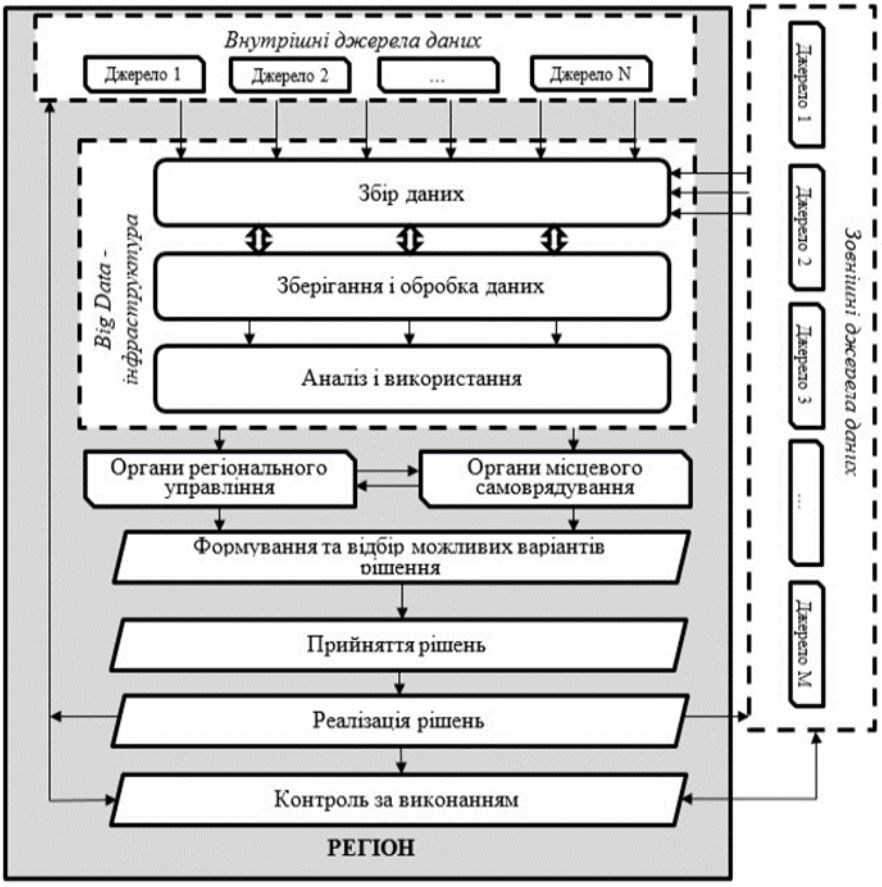


Рис.7.2. Процес прийняття управлінських рішень на регіональному (місцевому) рівні з використанням технології Big Data

Одним з ключових питань при впровадженні управлінських рішень, що базуються на аналізі даних із використанням технології Big Data, є оцінка ефекту від цих рішень у розвитку регіональної економіки та у формуванні збалансованого простору регіону. Тут можливими є наступні ефекти:

прямий ефект − операційний або економічний ефект, який отримують органи влади від реалізації управлінського рішення;

- синергетичний ефект, отриманий в результаті реалізації декількох управлінських рішень спрямованих на розвиток окремої сфери життя (транспорт, охорона здоров'я і т.п.) регіону (міста);
- загальний ефект, який отримують органи влади, населення та економіка.

Впровадження і використання методів Big Data для автоматизації обробки інформації в сфері регіонального управління та в управлінні містом є доцільним, так як технологія дозволяє вирішити, ряд питань на етапі

формування управлінських рішень щодо комплексного перспективного просторово-економічного розвитку. Адже, збір та обробка великих масивів даних дозволяє проводити ситуаційний аналіз і прогнозувати розвиток територій з урахуванням кон'юнктури ринку, демографічної ситуації, економічних та інших чинників, передбачати розвиток соціально-економічної ситуації в масштабах міста, регіону, всієї країни з метою запобігання соціальної напруженості, негативним змінам на ринку праці, в бізнес-середовищі в режимі реального часу.

3. Джерела відкритих даних в Україні

Держава збирає та накопичує інформацію про різні аспекти нашого життя. Ці дані дуже цінні, і велика частина з них має бути доступною та відкритою для бізнесу, стартапів, урядовців, журналістів, громадськості. Відкриті дані допомагають контролювати роботу державних органів, покращувати державні сервіси та створювати нові. Більше відкритих даних – більше можливостей для всіх.

Ми прагнемо аби Україна увійшла в трійку світових лідерів за рівнем розвитку сфери відкритих даних, адже наша ціль - побудувати прозору та підзвітну державу для громадян, які користуються сервісами на основі відкритих даних. Саме тому в Україні запущено Дія.Відкриті дані https://diia.data.gov.ua/ - центр компетенцій в сфері відкритих даних, що має на меті підвищення рівня знань про відкриті дані, їхній вплив та користь для кожного, а також допомогти Україні стати однією з найпрозоріших країн світу.

Відкриті дані — це:

Публічна інформація. Тобто інформація, яка була отримана/створена розпорядником на законних підставах.

Машиночитаний формат. Тобто формат, що дозволяє автоматизоване оброблення даних електронними засобами. Це означає, що не тільки людина, але й комп'ютер має розпізнавати дані.

Вільний і безоплатний доступ. Як до самих даних, так і до подальшого їх використання. Тож спокійно створюйте власні рішення на основі даних – як для себе, так і для бізнесу та громадськості.

Суспільна користь. Влада, бізнес та громадяни можуть використовувати відкриті дані для одержання економічної, екологічної та іншої соціальної користі.

Прозорість та боротьба з корупцією. Відкриті дані допомагають запобігати порушенням з боку влади. Приміром, усі охочі можуть контролювати планування та витрати бюджету, державні закупівлі тощо.

Драйвер інновацій та економічного розвитку. Дані по-справжньому цінні коли їх використовують. Великий і малий бізнес створює послуги та сервіси, віднаходить можливості для економії та вдосконалення своєї діяльності.

Єдиний державний веб-портал відкритих даних https://data.gov.ua/

YouControl — це українська команда, яка створює сервіси для аналізу підприємств https://youcontrol.com.ua/. Вона допомагає бізнесу уникати фінансових ризиків, а журналістам та громадським активістам — розслідувати суспільно важливі справи.

YouControl — аналітична система для комплаєнсу, аналізу ринків, ділової розвідки та розслідувань. Система формує повне досьє на кожну компанію України на основі відкритих даних, відстежує зміни в держреєстрах та візуалізує зв'язки між афілійованими особами. Унікальна технологія дозволяє за хвилину отримати актуальну (на час запиту) інформацію про компанію або ФОП із понад 100 офіційних джерел даних. Функція моніторингу щоденно повідомляє про зміни, спираючись на дані з офіційних джерел.

Opendatabot - платформа для роботи з відкритими даними, якою користується 250 000 користувачів в чат-ботах та 1000 компаній https://opendatabot.ua/.

Пошуково-аналітична система .007- Web-ресурс на основі відкритих даних про використання публічних коштів. Проект надає можливість проведення пошуку та візуалізації даних з відкритих джерел про використання державою бюджетних коштів. Основний акцент зроблено на простоті використання та представленні специфічної інформації з масивів великих даних. Допоможе зокрема знайти платежі, договори, акти виконаних робіт, додаткові угоди, специфікації розпорядників та одержувачів бюджетних коштів, з'ясувати, хто є засновниками організацій, які витрачають бюджетні кошти, провести аналітику трансакцій, договорів, контрагентів і тендерів. «BIBot» - потужний інструмент для виявлення фінансових зв'язків на основі платежів або тендерних угод https://www.007.org.ua/.

Тема 8. Аналіз великих даних у маркетингових дослідженнях

1. Роль великих даних в реалізації стратегій цифрового маркетингу

2. Проблеми та наслідки використання великих даних в маркетингу

3. Аналіз ефективності поштових розсилок

4. Веб-аналітика як важливий інструмент цифрового маркетингу

1. Роль великих даних в реалізації стратегій цифрового маркетингу

Невпинний технологічний прогрес змушує організації змінювати свою традиційну операційну діяльність, корегувати процеси, імплементувати нові інформаційні системи та підтримувати існуючі в актуальному стані. Розвиток інформаційних технологій приводить до пришвидшення темпів залучення людей до мережі Інтернет. Щодня з'являються терабайти нової інформації. У цих умовах технології обробки та аналізу даних стають життєво необхідними для існування сучасного бізнесу. Big Data – одна зі сфер інформаційних

технологій, яка найбільш швидко розвивається: згідно зі статистикою загальний обсяг даних подвоюється кожні кілька років. Постійно зростає кількість даних, що передаються мобільними мережами. За оцінками Cisco, у 2014 р. обсяг мобільного трафіку становив 2,5 ексабайта на місяць, а у 2019 р. він дорівнював 24,3 ексабайтам. Аналітики компанії IBS «весь світовий обсяг даних» оцінили такими величинами: 2003 р. – 5 ексабайтів даних (1 ЕБ = 1 млрд. гігабайтів), 2008 р. – 0,18 зетабайти (1 ЗБ = 1024 ексабайти), 2015 р. – понад 6,5 зетабайтів, 2020 р. – 40–44 зетабайти, 2025 р. – цей обсяг збільшиться ще у 10 разів (прогноз). З першого кварталу 2020 р. до першого кварталу 2021 р. глобальний середньомісячний трафік мобільних даних досяг 66 ексабайтів (або 66 млн. терабайтів), що на 68 % більше, ніж минулого року.

Аналітика великих даних вже набула поширення в багатьох сферах економіки. Термін «Big Data» означає методи обробки даних величезних обсягів, які дають змогу розподілено аналізувати цю інформацію. Невід'ємною частиною будь-якого бізнесу нині є наявність product placement у мережі Інтернет. Це може бути сайт, сторінка в соціальних мережах, профіль на відеохостингу Youtube. Усі ці складові є частиною цифрового маркетингу. Сучасні виклики економіки спонукають компанії переглядати свої рекламні кампанії та способи просування у мережі, враховуючи цифрові технології маркетингу.

Нові технології можуть впливати на поведінку споживачів, процеси управління та організаційну стратегію, в результаті чого генерується майже нескінченна кількість даних. Ефективний маркетинг має першорядне значення, щоб запропонувати правильний продукт потрібним клієнтам у потрібний час. Для цього необхідно керувати інформацією систематично, а зі згенерованих даних потрібно вилучати необхідну інформацію. Виникає запит на формалізовані способи отримання своєчасної та точної інформації, що стосується ринку, продуктів і клієнтів, а також загального ділового середовища. Величезні проблеми та можливості, пов'язані з новими технологіями та збільшенням доступності даних, змушують маркетингові організації (агентства, ЗМІ, рекламодавців) змінювати свої бізнес-моделі. Нові концепції та технології, зокрема Інтернет речей (IoT), Big Data, машинне навчання (ML), штучний інтелект (AI) та інші, вимагають масштабної адаптації від сучасного бізнесу.

Big Data на сьогоднішній день є одним із ключових драйверів розвитку інформаційних технологій. Це зумовлено, насамперед, тим, що в епоху інформаційних технологій, особливо після буму соціальних мереж, по кожному користувачеві Інтернету стала накопичуватися значна кількість інформації, що в кінцевому підсумку дало розвиток напряму Big Data. Варто зазначити, що до цієї сфери відноситься обробка саме великого обсягу інформації, який важко обробляти традиційними способами.

Big Data є відносно новою технологією, проте багато хто не усвідомлює, наскільки активно вона використовується, особливо глобальними брендами.

Важливість «великих даних» у сфері торгівлі та цифрового маркетингу складно переоцінити, адже використання цієї технології забезпечує:

1) допомогу компаніям у визначенні, які з їхніх продуктів матимуть кращий ринковий потенціал. Більше немає необхідності у великій кількості спроб і помилок. Компанія може збільшити масштаб реалізації продуктів і послуг, які цінує цільова демографічна група, і зосередити на них свої зусилля з продажу та маркетингу. Це також позбавить бізнес від заповнення складів товарами з низьким попитом;

2) надання перспективним брендам необхідної впевненості. Впевненість у своїх діях, ймовірно, не є великою проблемою для компаній, які вже досягли високої впізнаваності бренду, але це не стосується брендів, що розвиваються. Big Data дає обґрунтований прогноз щодо того, чи будуть нові продукти та послуги бренду популярними, чи ні;

3) збільшення продажів за допомогою оптимізації цін. Існує багато стратегій, які можна використовувати, щоб визначити правильну цінову політику на товари компанії. Звичайна формула – це обчислення загальної собівартості продукції плюс 10 % прибутку. Проте ці формули не завжди працюють, особливо в мережі Інтернет, де рівень конкуренції є значним. Big Data допоможе оптимізувати ціни не лише шляхом вивчення того, скільки споживачі готові витратити, аналізуючи їхні звички щодо витрат, але також враховуючи інші пов’язані фактори, такі як ціна конкурента, попит на продукт, стан галузі тощо;

4) допомогу у підвищенні ефективності маркетингових кампаній. Існує багато інструментів, які можуть допомогти «гарантувати» успіх маркетингових зусиль, але вони далеко не всі настільки ж інтуїтивно зрозумілі й надійні, як маркетинг, що базується на «великих даних». Використання Big Data у маркетингу може вказати на елементи, які дають можливість успішним маркетинговим кампаніям досягати поставлених цілей, і аспекти, які прирекли невдалі з них на провал. Маркетингова аналітика, яка ґрунтується на «великих даних», загалом допоможе приймати більш зважені рішення.

«Великі дані» сьогодні відіграють багато ролей у цифровому маркетингу. Серед найпоширеніших:

- Сегментація аудиторії: Big Data дають змогу маркетологам збирати, досліджувати й аналізувати різні аспекти поведінкових критеріїв – як люди використовують свої продукти та послуги, а також соціальні та демографічні фактори. Результати можуть допомогти ефективніше визначити вподобання споживачів, щоб маркетингові повідомлення можна було вдосконалити та оптимізувати.

- Аналіз настроїв: аналізуючи публікації в соціальних мережах, огляди та пошукові запити, маркетологи можуть краще зрозуміти, як споживачі ставляться до бренду.

- Цільовий маркетинг: рекомендації щодо продуктів, реклама в соціальних мережах та кампанії email-маркетингу використовують аналітику великих даних, щоб надати споживачам більш релевантний вміст.

- Прогнозний і директивний аналіз: маркетологи можуть працювати з ланцюгом поставок, щоб допомогти забезпечити більш адекватне за обсягами виробництво товарів через прогнозування попиту на основі Big Data.

- Вимірювання результатів: кампанії цифрового маркетингу можна вимірювати та корегувати в режимі реального часу для оптимізації бюджету.

- Аналіз профілів користувачів певних сервісів: наприклад Amazon, та розширення аудиторії за допомогою пропозиції товару користувачам зі схожим профілем.

- Моніторинг соціальних медіа: для визначення ставлення до власного продукту/бренду та продукту/бренду конкурентів, пошуку ідей для вдосконалення товару, аналізу якості обслуговування;

- Аналіз різних каналів продажів та відбір найкращих для конкретних клієнтів. - Аналіз активностей конкурентів.

Покращена лідогенерація – велика перевага, яку Big Data приносить маркетологам. Опитування McKinsey показало, що «постійні користувачі аналітики клієнтів мають у 23 рази більше шансів перевершити своїх конкурентів з точки зору залучення нових клієнтів». Використання хмарних технологій дає можливість збирати й аналізувати послідовні та персоналізовані дані з різних джерел, таких як Інтернет, мобільні додатки, електронна пошта, чат у реальному часі та навіть дії в точках продажу.

Big Data можуть допомогти маркетологам використовувати дані в реальному часі в середовищах хмарних обчислень. Здатність технології Big Data швидко й точно отримувати, обробляти й аналізувати дані в реальному часі, щоб вжити негайних та ефективних дій, не може зрівнятися з жодною іншою. Це важливо під час аналізу даних із GPS, датчиків Інтернету речей, кліків на вебсторінці або інших даних у реальному часі.

«Великі дані», як правило, бувають представлені у наступних трьох формах:

1. Структурована форма. Дані, які можуть зберігатися, бути доступними та обробленими у формі з фіксованим форматом, називаються структурованими. За тривалий час досягнуто великих успіхів у вдосконаленні техніки для роботи з цим типом даних (де формат відомий заздалегідь) і навчилися отримувати користь. Проте вже сьогодні спостерігаються проблеми, пов’язані зі зростанням обсягів до розмірів, які вимірюються в діапазоні кількох зетабайтів (1 зетабайт відповідає мільярду терабайтів). Дивлячись на ці числа, неважко переконатися в правдивості терміна Big Data та труднощах, пов’язаних з обробкою та зберіганням таких даних

2. Неструктурована форма. Дані невідомої структури класифікуються як неструктуровані. На додаток до великих розмірів, така форма

характеризується низкою складнощів для обробки та вилучення корисної інформації. Типовий приклад неструктурованих даних – гетерогенне джерело, що містить комбінацію простих текстових файлів, картинок та відео. Сьогодні організації мають доступ до великого обсягу сирих або неструктурованих даних, але не знають, як отримати з них користь. Прикладом такої категорії Big Data є результат Google пошуку.

3. Напівструктурована форма. Ця категорія поєднує в собі риси обох форм, описаних вище, напівструктуровані дані мають певну форму, але насправді не визначаються за допомогою таблиць у реляційних базах. Приклад цієї категорії – персональні дані, представлені у XML-файлі.

Як структуровані, так і неструктуровані дані можна розділити на додаткові підтипи. Спеціалісти у сфері цифрового маркетингу надають перевагу наступним типам інформації:

- Дані клієнта. Це інформація, яка допомагає маркетологам краще пізнати своїх клієнтів (імена, електронні адреси, вік, місцезнаходження, онлайн-активність та історію покупок). Також варто зібрати інформацію, яка визначить характер аудиторії компанії, наприклад активність у соціальних мережах, відповіді на опитування тощо.

- Фінансові дані. Це інформація, яка допомагає маркетологам аналізувати цифри, пов’язані з їхнім брендом (продажі та маркетингові показники (як компанії, так і конкурентів), витрати на маркетингову кампанію, отриманий дохід, пов’язаний із будь-якою запущеною кампанією тощо).

- Оперативні дані. Інформація, пов’язана з бізнес-процесами, що стосуються продажів і маркетингу, як-от логістика, онлайн-платформи, які використовуються, продуктивність співробітників, CRM-системи тощо.

Це, безумовно, не єдині типи інформації, яку можна збирати, керувати та аналізувати.

Завдяки високопродуктивним технологіям, таким як грід-обчислення або аналітика в оперативній пам’яті, компанії можуть використовувати будь-які обсяги «великих даних» для аналізу. Іноді Big Data спочатку структурують, відбираючи ті, що потрібні для аналізу. Все частіше Big Data застосовують для завдань у рамках розширеної аналітики, включаючи штучний інтелект.

Виділяють чотири основні методи аналізу Big Data:

1. Описова аналітика (descriptive analytics) – найпоширеніша. Вона відповідає на запитання «Що сталося?», аналізує дані, які надходять у реальному часі, та історичні дані. Головна мета – з’ясувати причини та закономірності успіхів чи невдач у тій чи іншій сфері, щоб використовувати ці дані для найефективніших моделей. Для описової аналітики використовують основні математичні функції. Типовим прикладом є соціологічні дослідження або дані веб-статистики, які компанія отримує через Google Analytics. У медичних установах, наприклад, цей тип аналізу часто використовується, коли велика кількість людей потрапляє до відділення невідкладної допомоги за

короткий період часу. Система описової аналітики повідомляє про дану ситуацію і надає дані в режимі реального часу з усією відповідною статистикою (дата виникнення випадку, обсяг, дані про пацієнта тощо).

2. Прогнозна або предикативна аналітика (predictive analytics) – допомагає спрогнозувати найімовірніший розвиток подій на основі наявних даних. Для цього використовують готові шаблони на основі якихось об'єктів або явищ з аналогічним набором характеристик. За допомогою предикативної (або прогнозної) аналітики, наприклад, можна прорахувати обвал або зміну цін на фондовому ринку, або оцінити можливості потенційного позичальника із виплати кредиту, чи спрогнозувати збільшення кількості клієнтів внаслідок зміни ринкової кон'юнктури.

3. Приписна аналітика (prescriptive analytics) – наступний рівень порівняно з прогнозною. За допомогою Big Data та сучасних технологій можна виявити проблемні точки у бізнесі чи будь-якій іншій діяльності та розрахувати, за якого сценарію їх можна уникнути у майбутньому, наприклад, збільшити чисельність персоналу відділу збуту через передбачення потенційної загрози невиконання контрактів внаслідок виникнення ажіотажного попиту на продукцію компанії.

4. Діагностична аналітика (diagnostic analytics) – забезпечує більш глибокий аналіз, щоб відповісти на запитання «Чому це сталося?». Це допомагає виявляти аномалії та випадкові зв'язки між подіями та діями, наприклад виявлення причин виникнення ажіотажного попиту на певну продукцію.

Дані обробляють та аналізують за допомогою різних інструментів та технологій:

- спеціальне програмне забезпечення: Apache Hadoop, Apache Cassandra, NoSQL, MapReduce, Xplenty;
- data mining – вилучення з масивів раніше не відомих даних за допомогою великого набору техніки;
- штучний інтелект та нейромережі – для побудови моделей на основі Big Data, включаючи розпізнавання тексту та зображень. За допомогою штучного інтелекту компанії аналізують клієнтський досвід та пропонують персоніфіковані продукти та послуги;
- візуалізація аналітичних даних – анімовані моделі чи графіки, створені з урахуванням великих даних.

Роль «великих даних» у цифровому маркетингу значно зросла за останні кілька років. Однак інтеграція Big Data із реальною аналітикою потребує постійного удосконалення. Оскільки методи збору даних стають все більш впорядкованими через Інтернет речей, використання мобільних пристроїв, голосовий пошук у домі та транспортних засобах за допомогою AI (штучний інтелект) помічників, дані можуть стати ще точнішими, складнішими та динамічнішими. Компанії, які використовують ці дані, потенційно можуть

проводити кращі кампанії ретаргетингу, створювати більш релевантні пропозиції щодо продуктів і надавати рекламний вміст, який найбільше відповідатиме їх цільовій аудиторії. Цифровий маркетинг, керований даними (data-driven marketing), може суттєво скоротити воронки продажів і підвищити прибутковість компаній.

2. Проблеми та наслідки використання великих даних в маркетингу

Big Data, звичайно, як і багато інших технологій, все ще не ідеальна, незважаючи на значний потенціал. Використання «великих даних» у цифровому маркетингу пов’язано з деякими проблемами та наслідками:

- Відкриваються можливості неправомірного використання особистої інформації. Не має значення, чи це звичайний споживач, маркетолог чи власник бізнесу – нікому не подобається, коли його особиста інформація продається іншим людям без згоди.

- Проблема масштабування. Великі дані – це завжди великий обсяг інформації, який потребує не тільки зберігання, а й постійного доступу. Більшість корпоративних центрів інформації не були розраховані на такі об’єми. Отже, компаніям доводиться не тільки думати про розширення власних корпоративних центрів зберігання, а й шукати способи оптимізації, наприклад використовувати загальні стандарти зберігання та обробки, переводити дані до «хмар».

- Інтеграція даних, зібраних раніше. Для маркетолога важливо мати доступ до даних щодо клієнтів, проведених кампаній, маркетингових досліджень минулих періодів. Без них часто неможливо побудувати тренд, зрозуміти специфіку поведінки споживача на ринку. Системи зберігання таких даних ніколи не призначалися для використання в режимі реального часу. Фахівці в сфері «великих даних» сходяться на думці, що використання технології Big Data не ефективне, якщо серверні системи не можуть підтримувати транзакції в реальному часі.

- Розрізненість систем збору та обробки даних. Компанії збирають дані для різних цілей, у різний спосіб, рідко інтегруючи системи збору. Ніхто не уявляв собі, що коли-небудь потрібно буде взаємодіяти з абсолютно незв’язаними системами та сховищами даних, одночасно всередині і поза компанією, і для аналізу, і для візуалізації. Навіть коли технологія може забезпечити рішення для інтеграції та взаємодії, власники бізнесу неохоче відмовляються від контролю або вимагають від IT-персоналу виставляти пріоритети щодо проектів на основі поточних інтересів бізнесу.

- ТехнологіїBig Data потребують фахівців високого рівня кваліфікації. При зборі та аналізі даних необхідно ставити правильні запитання. Відділам маркетингу, навіть зі значним бюджетом, буде важко «перекупити» талановитих аналітиків у інвестиційних та фінансових компаній. Наявні на

ринку фахівці часто спеціалізуються винятково на IT-проектах і погано знайомі з філософією і культурою маркетингу.

Проблема вироблення універсальної мови спілкування та роботи з даними всередині компанії. Для технологій Big Data життєво потрібна інтеграція. Компаніям доведеться навчати базових навичок роботи з даними велику кількість працівників. Тільки спільна робота всіх співробітників зможе дати зрушення та змусити всю компанію правильно та ефективно використовувати технології Big Data. Також необхідно постійно пам'ятати, що деякі види даних, включаючи фінансові та медичні записи, підлягають захисту та значному регулюванню, яке може змінюватися залежно від місцезнаходження.

- Потенційна загроза безпеці. Компанії які пов'язані зі зберіганням та обробкою величезних масивів даних, можуть стати справжньою мішенню для зловмисників, одна єдина помилка в системах безпеки може призвести до витоку чи крадіжки особистих даних, шантажу тощо.

- Потенційна можливість нашкодити вашій репутації. Big Data можуть давати неймовірні результати щодо впізнаваності бренду, але буває і навпаки. Є ряд споживачів, які не відчувають себе комфортно, коли компанії використовують їхню інформацію для збільшення продажів і покращення маркетингових кампаній. Відсутність розуміння клієнтами специфіки цієї технології приводить їх до відчуття, що Big Data є формою цифрової «маніпуляції». Проте коректно проведені маркетингові кампанії можуть покращити якість життя людей, інформуючи їх про продукти та послуги, якими вони будуть першочергово користуватися.

- Потенційна можливість помилок. Ця проблема може бути пов'язана з прогалинами у структурі системи або навіть через атаку зловмисників, саботаж тощо. Надто покладаючись на Big Data, які вже скомпрометовані, можна завдати серйозної шкоди продажам і цифровим маркетинговим кампаніям.

Існує ряд державних і неурядових організацій, які працюють, щоб зробити використання великих даних більш безпечним і відповідальним. Наприклад, надання користувачам Інтернету можливості вибору – приймати файли cookie чи ні – це значний крок до кращого цифрового майбутнього. Чим більше людей довіряє технології Big Data, тим краще ми зможемо використовувати її для потреб бізнесу.

Технологія Big Data сьогодні відіграє величезну роль при здійсненні практичної діяльності у сфері цифрового маркетингу. «Великі дані» дають можливість оптимізувати маркетингові кампанії та забезпечити кращий досвід роботи з клієнтами. Поява Big Data і прогнозного аналізу спричинила зміни парадигми в digital-маркетингу. Маркетинг, керований даними (data-driven marketing), тепер є стандартом для тих компаній, що хочуть утримувати лідерські позиції на ринку. З кожним роком використання Big Data допомагають пізнавати клієнтів все краще. Попри величезні можливості та

ефекти від використання «великих даних» у цифровому маркетингу, все ще існує значна кількість проблем для імплементації цієї технології в масове використання для потреб бізнесу.

3. Аналіз ефективності поштових розсилок

Email-маркетинг — це напрям performance-маркетингу, який використовує електронну пошту для комунікації бізнесу з клієнтом.

Цілі email-маркетингу:

- Підвищення впізнаваності та лояльності до бренду.
- Збільшення продажів (та ймовірності повторних).
- Побудова довірливих відносин із користувачами.
- Супровід користувачів на всіх етапах ухвалення рішення.
- Повернення неактивної аудиторії та конвертування її в покупців.

Як розробити ефективну стратегію email-маркетингу:

1. Визначити цільову аудиторію. Важливо зрозуміти, чого хочуть користувачі, і адаптувати електронну кампанію до їхніх потреб.
2. Чітко сформулювати мету розсилки. Наприклад, стимулювати продажі, підвищити впізнаваність або лояльність до бренду.
3. Розробити спосіб реєстрації для створення бази даних розсилки. Це може бути вкладка для підписки внизу вебсайту, підписка на розсилку під час реєстрації, оформлення замовлення або завантаження файлу з корисною інформацією.
4. Вибрати тип кампанії (щотижнева розсилка, особистий блог тощо).
5. Вибрати розклад і частоту розсилки. Дослідження показали, що електронні листи краще надсилати в середині робочого тижня — тобто у вівторок, середу або четвер, з дев'ятої до одинадцятої ранку. Однак, варто самостійно тестувати й перевіряти, як реагують користувачі на різний розклад розсилок.
6. Вимірювати результати.

ЯКІ МЕТРИКИ АНАЛІЗУВАТИ?

Open rate (або OR) — базова метрика в email-маркетингу. Вона показує, наскільки розсилки цікаві користувачам, а також — наскільки вдалою виявилася тема листа.

$$\text{Open rate (або OR)} = \frac{\text{Кількість відкритих листів}}{\text{Кількість надісланих}} \times 100\ \%$$

Click to open rate (або CTOR) — метрика, яка допомагає оцінити релевантність контенту в листах.

$$\text{Click to open rate (або CTOR)} = \frac{\text{Кількість кліків}}{\text{Кількість відкриттів}} \times 100\ \%$$

Click-through rate (або CTR) — це показник «клікабельності». Він вказує, наскільки ефективно працює розсилка.

$$\text{Click-through rate (або CTR)} = \frac{\text{Кількість кліків}}{\text{Кількість доставлених листів}} \times 100\ \%$$

Показник відмов (або bounce rate) — відсоток користувачів, які не отримали розсилку, бо поштові сервери її повернули.

Листи повертаються, коли їх не можна доставити за вказаною адресою. Наприклад, коли адреси користувача не існує або розмір листа завеликий. Якщо значення bounce rate підвищене, поштові провайдери вирішують, що відправник розсилає спам, та блокують його розсилки.

Метрика відписок (або unsubscribe rate) — показує, скільки людей відписалися від розсилки.

Скарги на спам (або spam complaints) — метрика, яка відстежує, коли користувачі позначають розсилку як спам.

Види електронних листів:

1. Тригерні листи. Це автоматична розсилка, створена заздалегідь на основі поведінки користувачів. На відміну від звичайних, тригерні листи надсилають не рандомно, а у запланований час або залежно від дій отримувачів.

2. Комерційні листи. Це рекламні листи зі спеціальними пропозиціями, знижками, освітньою інформацією та оголошеннями. Такі листи мають бути персоналізованими й містити вигідну пропозицію, наприклад, знижку або унікальну послугу.

3. Анонси. Це листи, в яких повідомляється про певну подію або новину. Це, наприклад, листи з анонсами акцій або інформацією про майбутні розпродажі.

4. Інформаційні листи. Це листи, у яких міститься корисна для користувачів інформація, яка може їх зацікавити.

Структура листів:

Перше, на що користувачі звертають увагу перед тим, як відкрити лист — його **тема**. Саме від неї залежить, чи прочитають повідомлення повністю, чи видалять, навіть не переглянувши. Важливо, щоб тема листа пояснювала, що саме всередині нього. Необхідно зацікавити читача, але не розкривати всю ідею повідомлення відразу.

Далі — **Preheader**. Це фрагмент тексту, який відображається в листах після імені відправника і теми. Основна мета прехедера — розкрити ідею листа і мотивувати підписників відкрити повідомлення.

Третій елемент **Header** — тобто шапка з елементами, які наштовхнуть на розуміння, який це продукт. До прикладу, у хедері може бути його назва та логотип. Шапка допомагає швидко ідентифікувати бренд: одержувачі бачать ім'я відправника в полі «Від кого». А коли відкривають лист, одразу звертають увагу на логотип.

Після цього йде **Body** (або тіло листа). Це частина, яка містить заголовок, текст, кнопки із закликом до дії — або CTA (тобто call-to-action). У CTA краще використовувати не більше ніж 2 слова, включно з дієсловом. CTA можна виділити кольором логотипа компанії, зробити контрастним, щоб він не зливався із загальним фоном, або анімованим.

Останній елемент — **Footer**. Це повідомлення, чому людина отримала лист, посилання для відписки та загальна інформація про компанію (наприклад, підпис, лінки на соцмережі та інші контакти).

4. Веб-аналітика як важливий інструмент цифрового маркетингу

Розвиток інформаційних технологій дає змогу проводити комплексне дослідження більшості процесів цифрового маркетингу, насамперед йдеться про оцінювання маркетингових кампаній у мережі Інтернет та ефективності сайтів окремих компаній. Для досягнення цих завдань використовують веб-аналітику, що дає змогу збирати комплексну інформацію про ключові процеси на Інтернет-ресурсах компанії, аналізувати отримані результати та приймати ефективні управлінські рішення. Пошукові системи заохочують компанії вкладати кошти у рекламу в Інтернеті, використовуючи інструменти веб-аналітики, які дають змогу отримати інформацію про дії клієнтів (кількість кліків на контекстну рекламу, час перегляду рекламного ролику, відвідування сторінок з рекламним контентом тощо). Необхідно зауважити, що системи веб-

аналітики значно відрізняються від систем статистики, оскільки, на відміну від останніх, дають змогу не тільки збирати комплексні дані про поведінку цільової аудиторії, але й досліджувати модель поведінки окремих користувачів та аналізувати отримані результати. Натепер є значна кількість систем веб-аналітики, серед яких доцільно виділити Google Analytics, AdWatcher, Snoobi, ClickTracks Optimizer, ClickTale, CrazyEgg та ін.

Слід зазначити, що отримані у результаті веб-аналітики дані потрібно використовувати як оціночні, що пояснюється певними неточностями збору інформації внаслідок таких причин:

1. Використання декількох пристроїв. Сучасні користувачі використовують декілька пристроїв, перебуваючи в Інтернеті, що не дає можливості об'єктивно вирахувати кількість цільової аудиторії, яка відвідує тематичні сайти або переглядає певний контент. Крім того, одним пристроєм у різні періоди часу можуть користуватися різні особи, що також приводить до викривлення даних.

2. Технічні помилки. Нестабільність роботи Інтернету чи окремих сайтів у певні періоди часу приводить до неможливості зібрати інформацію про наявні процеси у повному обсязі.

3. Побудова вибірок. На великих Інтернет-ресурсах накопичують значні обсяги інформації, обробка якої вимагає значних затрат часу, грошових ресурсів та фахівців. Для аналізу великих даних дуже часто використовується метод вибірки, що передбачає виокремлення частини сукупності даних за певними соціальними, демографічними, економічними та іншими характеристиками з подальшою екстраполяцією отриманих результатів на генеральну сукупність.

4. Відключення cookie-файлів та JavaScript. Деякі користувачі відключать певні файли та скрипти, що приводить до неможливості отри-мати об'єктивну статистику про їхню активність та дії у мережі Інтернет.

Веб-аналітика базується на формуванні певної групи науково обґрунтованих показників, зборі відповідної інформації та її всебічному аналізі. Показники веб-аналітики групуються за різноманітними ознаками:

I. За соціально-демографічними ознаками: – стать;

– вікова група; – рівень освіти;

– соціальний статус; – рівень доходів та ін.

II. За методами розрахунку:

– кількісні показники, які відображають обсяг або чисельність досліджуваного явища: кількість відвідувачів сайту, кількість переглядів певного контенту або реклами; затрати на рекламу тощо;

– якісні показники розраховуються у вигляді відносних або середніх величин та відображають якісні параметри досліджуваного явища (рівень конверсії, середній час перебування на сайті, вартість одного перегляду реклами або кліка тощо).

Під час формування системи показників важливо виокремити головні індикатори, які дадуть змогу оцінювати ефективність функціонування компанії в Інтернеті, рівень комунікації з цільовою аудиторією та ін. Використання ключових показників ефективності (KPI) передовими компаніями світу в останні роки пояснюється універсальністю зазначеного підходу та можливістю оптимізувати будь-які бізнес-процеси на рівні конкретного підприємства. Під час формування системи показників поряд із забезпеченням комплексного оцінювання ключових процесів у мережі Інтернет потрібно уникати значної деталізації аналізованих явищ та дослідження другорядних для компанії індикаторів, що при-веде лише до економічно недоцільних витрат фінансових, трудових та часових ресурсів. Веб-аналітика дає можливість використовувати значну кількість інструментів, тому маркетинговий підрозділ компанії повинен на основі наукових підходів та специфіки функціонування компанії сформувати оптимальну кількість показників, які дадуть змогу проводити всебічний кількісний та якісний аналіз. У таблиці 8.1 наведено особливості формування KPI, виходячи з напрямів оцінювання діяльності сайту компанії.

Таблиця 8.1

Формування системи показників, виходячи з напрямів оцінювання компанії

Напрями оцінювання	**Ключові показники ефективності**
Показники продажів та конверсії	– кількість продажів продукції загалом та за окремими позиціями; – кількість трафіку загалом та у розрізі тематичного контенту; – кількість та питома вага нездійснених замовлень продукції загалом та за окремими позиціями (клієнти поклали товар у корзину, проте не оплатили її).
Якісний склад трафіку упродовж майбутніх періодів часу (тиждень, місяць, квартал, рік)	– темпи приросту трафіку загалом; – темпи приросту трафіку за окремими видами контенту, що розміщується на сайті компанії (у тому числі за товарними позиціями); – частка відвідувачів сайту у загальній кількості цільової аудиторії певного ринку; – середня вартість переходу та темп приросту його вартості упродовж досліджуваного періоду часу
Налагодження комунікацій з клієнтами на постійній основі	– кількість клієнтів, що здійснюють повторні покупки впродовж певного періоду часу; – питома вага клієнтів, які повторно звернулися на сайт компанії; – частка клієнтів, що повторно придбали продукти компанії через сайт; – кількість та частка клієнтів, які придбали продукцію компанії, за різноманітними цифровими каналами;
Рівень задоволеності клієнтів компанією	– кількість брендового трафіку; – частка клієнтів, які задоволені; – питома вага клієнтів, що задоволені рівнем обслуговування; – частка клієнтів за холодними та гарячими зонами сайту; – частка клієнтів, які позитивно сприймають сайт компанії (у тому

	числі оформлення, зручність використання та пошуку, розміщення товарів та їх опис тощо)

Використання комплексної інформації, виділення основних показників для окремих напрямів маркетингової діяльності у мережі та загалом для веб-ресурсів компанії дає можливість менеджерам відповідних рівнів всебічно аналізувати наявні процеси, оцінювати їх у короткостроковій, середньостроковій та довгостроковій перспективах і на основі отриманих результатів приймати стратегічні рішення.

Комплексний аналіз діяльності підприємства за допомогою KPI передбачає відбір до зазначеної системи статистичних показників, виходячи з рівня дослідження (окремі сторінки, або рекламні заходи, або діяльність Інтернет-ресурсів компанії загалом), періоду дослідження (короткостроковий, середньостроковий або довгостроковий), діяльності (рекламна, збутова) тощо.

Правильно сформована система KPI для потреб веб-аналітики має такі переваги:

1. Отримані показники дають можливість керівництву контролювати всі етапи функціонування компанії в Інтернеті.

2. Система показників забезпечує оптимізацію прийняття управлінських рішень щодо функціонування сайту компанії та реалізації заходів у сфері цифрового маркетингу.

3. Зазначена система показників спрямована на підвищення ефективності усіх веб-процесів компанії.

4. KPI забезпечують оперативне та всебічне розуміння процесів компанії у мережі.

5. Показники ефективності процесу в майбутньому можуть слугувати вимірниками передового досвіду компанії.

6. Науково обґрунтована система KPI може бути використана для побудови візуалізованого звіту (dashboard), який дає можливість проаналізувати діяльність компанії .

Серед значної кількості показників, які можуть бути сформовані та налаштовані у системі веб-аналітики, найважливішими для керівництва компанії є індикатори ефекту від використання вкладених грошових ресурсів. Система індикаторів, що використовуються для характеристики діяльності компанії у мережі Інтернет, складається з таких показників:

1. CPA (Cost Per Action) – вартість певної дії, яку здійснив відвідувач сайту компанії. Цей показник є гнучким за своєю сутністю, оскільки різноманітні компанії використовують різні методичні підходи щодо ідентифікації дій користувачів на сайті (перегляд окремої сторінки або певного контенту, перехід за певними посиланнями, реєстрація, заповнення заявки тощо).

2. CPL (Cost Per Lead) – вартість потенцій-ного клієнта, який залишив персональні дані під час контакту з працівниками компанії або заповнення певної форми на Інтернет-ресурсі.

3. CPO (Cost Per Order) – вартість одного підтвердженого замовлення, яке було здійснено на сайті компанії.

4. ROI (Return on Investment) – коефіцієнт повернення інвестицій, який характеризує рентабельність вкладених компанією коштів загалом або в окремі процеси.

5. ROAS (Return On Ad Spend) – прибуток від розміщення реклами загалом та за окремими видами реклами.

6. ROAS (Return On Advertising Spend) – прибуток від розміщення реклами на 1 грошову одиницю загалом та за окремими видами реклами.

7. CTR (Click-Through Rate) – відношення кількості переглядів до кількості кліків на це оголошення.

8. CPC (Cost Per Click) – вартість кліку на рекламне оголошення.

9. EPC (Earnings Per Click) – прибуток у розрахунку на один клік.

10. LTV (Lifetime Value) – сукупний прибуток, який компанія отримує від клієнта за весь час співпраці з ним у мережі Інтернет.

11. CPI (Cost Per Install) – вартість встановлення мобільного додатку .

Серед зазначених показників найважливіше місце для маркетингової стратегії компанії в Інтернеті займає ROI. Якщо значення показника ROI більше за 100%, то інвестиції приносять прибуток, якщо ж ROI менше за 100% – інвестиції нерентабельні. Компанії необхідно відстежувати ROI для всіх ключових процесів у Інтернеті з певною періодичністю, що дасть змогу коригувати структуру витрат на різноманітні заходи залежно від їхньої ефективності та оптимізувати розподіл інвестицій.

На основі отриманих значень ROI може бути прийняте рішення щодо збільшення прибутку. Для досягнення поставленої мети можна реалізувати такі заходи, які сприятимуть зростанню потоків прибутку, як:

– реалізація стратегії, орієнтованої на зростання трафіку на Інтернет-ресурси компанії, що дасть змогу збільшити кількість продажів товарів чи послуг компанії;

– оптимізація затрат завдяки перерозподілу наявних грошових ресурсів: відмова від неефективної Інтернет-реклами та фейкових кліків (клікфрод).

Воронка — це декомпозиція шляху користувача.

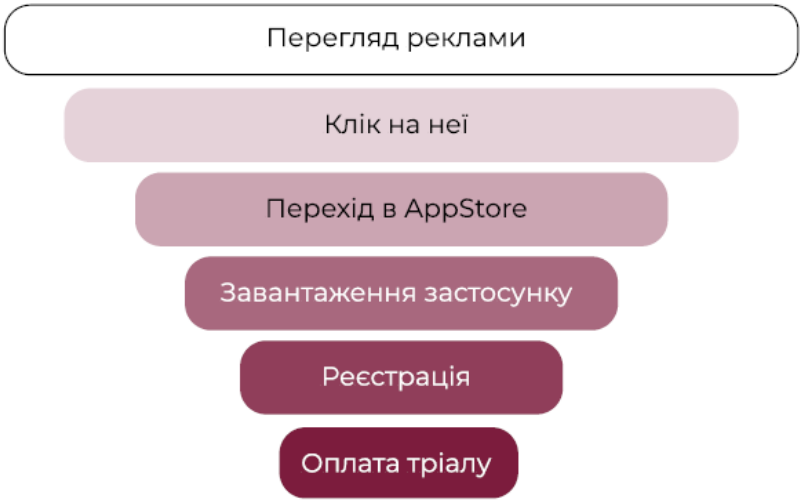
Маркетингова воронка — це шлях потенційного користувача від моменту знайомства з продуктом до здійснення ним регулярних покупок.

Маркетингова воронка — це інструмент, який дозволяє краще розуміти шлях користувачів і, усуваючи складні для користувача місця, оптимізувати його.

Воронки показують, які стадії відносин проходить споживач відносно продукту, щоб у результаті зробити покупку та сформувати довіру до бренду, тобто — лояльність.

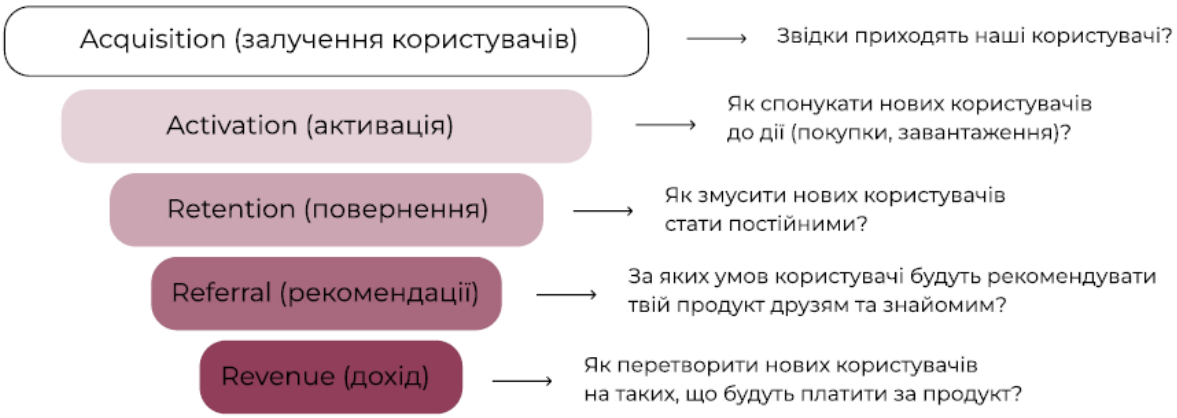
Як побудувати маркетингову воронку?

Можна розбивати шлях користувача на конкретні маленькі кроки, Наприклад:



Або можна розбити на ширші кроки — наприклад:
Awareness, Consideration, Conversion, Loyalty.

«Піратська» воронка AARRR – це воронка, яка включає такі складники: **Acquisition, Activation, Retention, Referral, Revenue.**

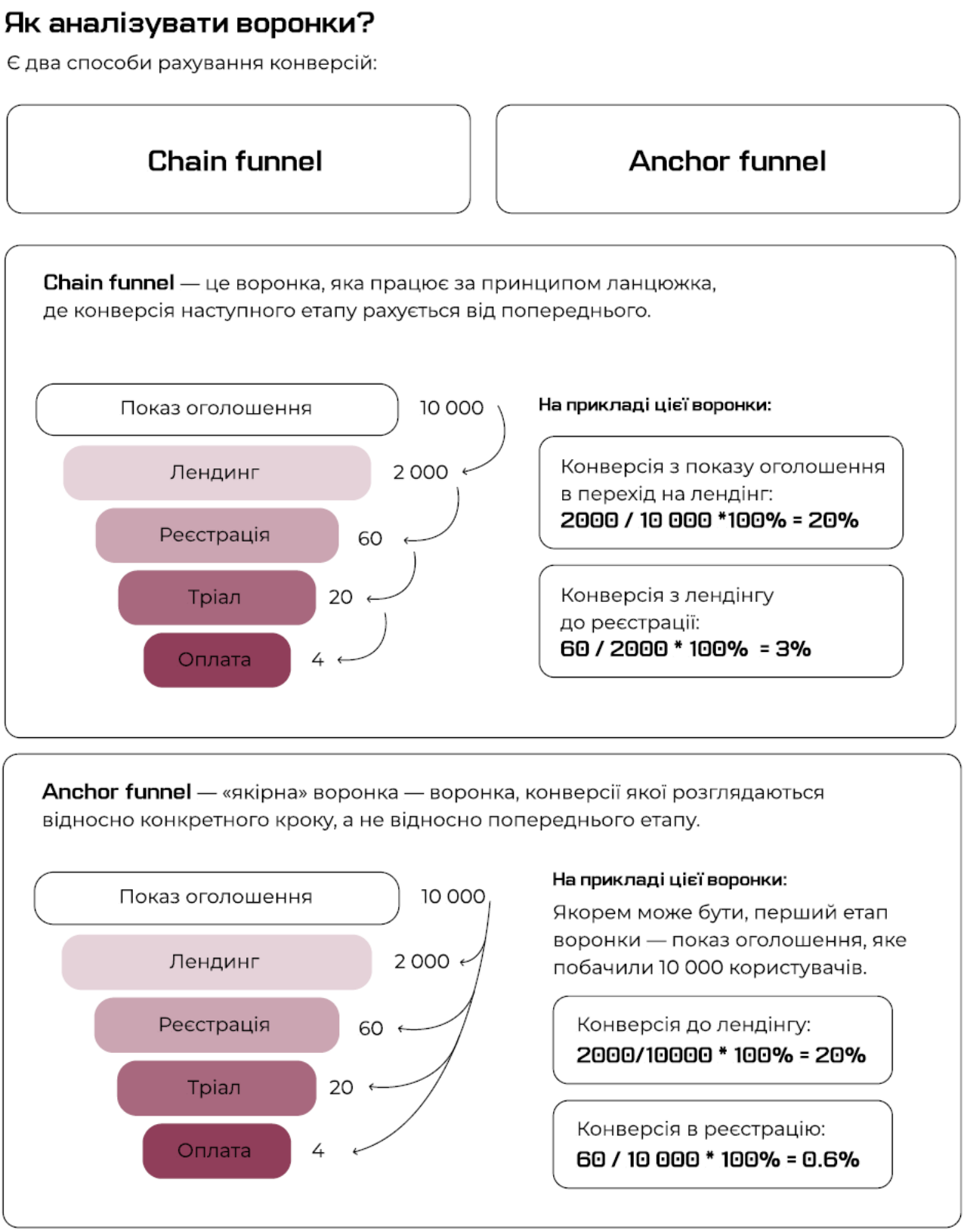


Навіщо розбивати шлях користувача?

1. **Щоб зрозуміти, хто насправді є цільовою аудиторією продукту.** Проаналізувавши низ воронки, можна зрозуміти, що об'єднує цільових користувачів. Це можуть бути — демографічні показники чи інтереси аудиторії.Така аналітика дозволяє оптимізувати креатив, таргетинг, лендінги та інші маркетингові інструменти.

2. Воронка необхідна для того, **щоб порівнювати ефективність каналів трафіку** для різних груп користувачів.Наприклад, під час закупки реклами для двох країн воронки можуть допомогти глибше проаналізувати різницю між цими ринками.

3. **Воронка дозволяє знаходити проблеми.** До прикладу, воронка може допомогти відшукати та усунути перешкоди перед покупкою продукту (такі, як-от незрозуміле розташування кнопки оплати, проблеми з платіжною системою — вони можуть завадити користувачу здійснити покупку продукту).



Специфіка побудови комунікації у цифровому маркетингу передбачає конверсію лише на певному етапі відносин між веб-ресурсами компанії та цільовою аудиторією. Система може включати побудову складної воронки продажів з інтеграцією у неї ВРП (воронки перед воронкою) та інших

додаткових маркетингових інструментів, що приводить до здійснення цільової дії клієнтами з певним лагом. Для компаній, що реалізують товари з різною специфікою, рівень конверсії загалом та за окремі проміжки часу буде відрізнятися. У цьому контексті ключова мета веб-аналітики полягає у зборі інформації про конверсію за окремими напрямами діяльності компанії та загалом на постійні основі.

Система веб-аналітики є гнучкою за своєю сутністю та дає змогу оцінювати ключові процеси на ресурсах компанії за різними напрямами і подавати результати у різноманітних формах, що дає можливість формувати різні за наповненням звіти. Окрім ROI, також є й інші методи веб-аналітики, які можуть бути корисними для керівництва компанії на етапі розроблення стратегії оптимізації маркетингової діяльності у мережі Інтернет. Нижче наведено деякі з варіантів оптимізації цифрового маркетингу компанії за допомогою інструментів веб-аналітики.

Веб-аналітика дозволяє, що дає можливість проводити комплексний аналіз різноманітних заходів на веб-ресурсах компанії з метою виявлення най-ефективніших дій, які сприятимуть максимізації трафіку та потенційно можливого прибутку. Цей принцип може бути використаний для:

1. Створення різноманітних варіантів сайтів. Компанія створює різні варіанти сайту за дизайном, наповненням та поданням цільової інформації. За допомогою веб-аналітики проводиться дослідження поведінки цільової аудиторії на різних Інтернет-ресурсах, визначаються вподобання користувачів та визначається най-кращий сайт для просування бренду. Цей підхід є ефективним рішенням під час вибору найкращої цільової сторінки (Landing Page) для просування певної продукції компанії.

2. Формування різних каналів комунікації з клієнтами. Інформація про компанію та її продукцію розміщується у різноманітній формі (текстова інформація, відео, інфографіка, пости тощо) на різноманітних ресурсах (власний сайт, соціальні мережі, блоги, форуми тощо). Інструменти веб-аналітики дають змогу оцінити конверсію від різноманітних каналів та прийняти рішення щодо доцільності використання певних каналів або зміни їхньої ролі у загальній маркетинговій стратегії у мережі Інтернет.

3. Оцінювання ефективності рекламних повідомлень. Для продуктів компанії створюється різноформатна реклама, яка просувається різноманітними каналами в Інтернеті. Використання веб-аналітики дає змогу оцінити ефективність кожного з видів реклами та оптимізувати її використання для досягнення максимального можливого ефекту для компанії.

Для оцінювання зручності використання сайту компанії, привабливості його певних зон для клієнтів веб-аналітика дає можливість використовувати інструменти, що формують карту уваги користувачів для кожної окремої сторінки. У цьому разі формуються холодні (ділянки, на які користувачі звертають найменше уваги) та гарячі (ділянки, що користуються найвищою

популярністю серед цільової аудиторії) зони сайту. Програмне забезпечення деяких компа-ній у сфері веб-аналітики дає змогу записувати відео, яке показує переміщення курсора на сторінці та перехід користувача між посиланнями усередині сайту та на зовнішні ресурси.

Система веб-аналітики дає змогу ефективно протистояти клікфроду, який є нелегальним видом діяльності та завдає значних збитків рекламним кампаніям через вимивання бюджетів на рекламу за різні товари чи послуги. Клік-шахраї навмисно та регулярно переходять за рекламними посиланнями конкурентів лише з метою завдання збитків внаслідок здійснення нецільових кліків. Клікфрод призводить до зростання вартості одного кліку для компанії, оскільки за певного рекламного бюджету за рахунок клік-шахраїв зменшується кількість клієнтів, які купують продукцію певного бренду.

Система веб-аналітики дає змогу ідентифікувати клікфрод завдяки виявленню підозрілої моделі поведінки у процесі переходу за посиланнями (швидкий перехід між посиланнями може свідчити про використання спеціалізованих ботів), встановленню надмірної активності за певними IP-адресами, виявленню фактів аномальної кіль-кості переходів за рекламним повідомленням та швидким виходом з нього, фіксації надмірної активності користувачів з країн третього світу та ін. Отримані результати можуть бути використані для боротьби з клікфродом: компенсація збитків від сервісу, який розповсюджує рекламу; встановлення та блокування реклами для зловмисників за ідентифікованими IP-адресами тощо.