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

Конспект лекції укладено на основі джерел:

Великі дані URL: https://uk.publicspeakingtip.org/big-data-7429

Кобзаренко Д.Н., Мустафаев А.Г. Учебное пособие дисциплины «Анализ больших данных» для направления подготовки 38.03.05 «Бизнесинформатика», профиль «Электронный бизнес». – Махачкала: ДГУНХ, 2019 г. – 107 с.

Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. СПб: Университет ИТМО, 2018. 52 с.

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології /Пер.з англ. Анатолія Гладуна; За наук. ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

1. Суть Big Data

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

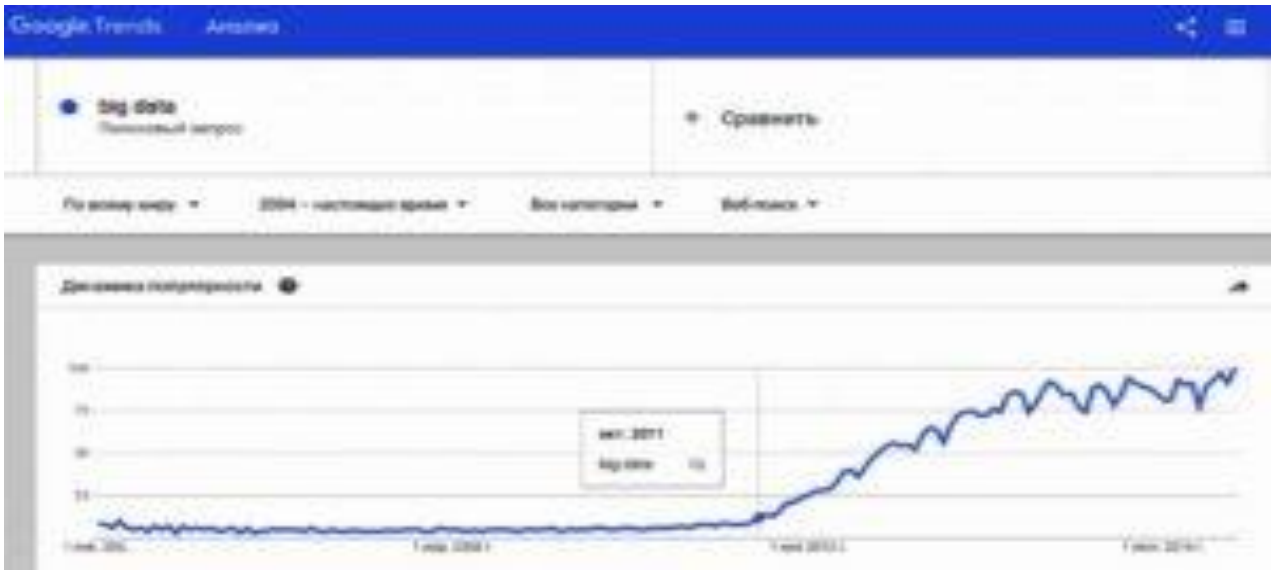
Постійне прискорення зростання обсягу даних є невід'ємним елементом сучасних реалій. Соціальні мережі, мобільні пристрої, дані з вимірювальних пристроїв, бізнес-інформація – це лише кілька видів джерел, здатних генерувати гігантські масиви даних.

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. l.

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.

ЩО ТАКЕ BIG DATA: ЯК БІЗНЕС ЇХ ВИКОРИСТОВУЄ?

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. Рітейл-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

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

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).

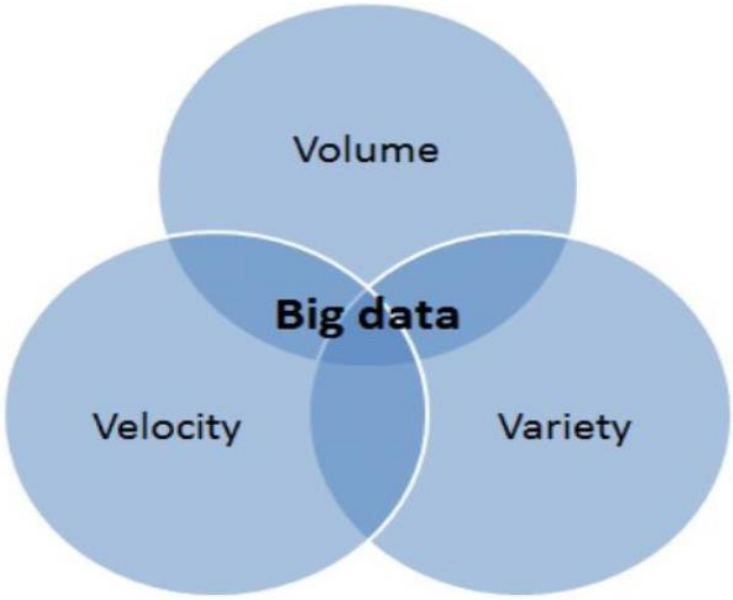


Рис.1.4. Характеристики Big Data [5]

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.