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

Конспект лекції укладено на основі джерел:

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

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

Фрэнкс Билл Укрощение больших данных: как извлекать знания из массивов информации с помощьюглубокой аналитики / Билл Фрэнкс ; пер. с англ. Андрея Баранова. М. : Манн, Иванов иФербер, 2014

Шандрівська О. Є., Кириленко А. А.Особливості ідентифікації ризиків ринку big data. URL: http://science.lpnu.ua/sites/default/files/journal-paper/2021/jun/23774/menedzhment121-84-97.pdf

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 BigData 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. 1.

Таблиця 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. 2.

Таблиця 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.

Таблиця 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 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. 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. 4).

Таблиця 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 постачальників	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 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. 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 ІТ-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 ІТ-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. 6).

Таблиця 6

Матриця SWOT-аналізу ринку Big Data

	Можливості (О)	***Загрози (Т)***
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 ІТ– 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. 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. 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. 8.

Таблиця 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 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.

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, Fasebook.

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 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.