

Тема. Big-data аналіз.

План.

1. Історія виникнення та напрями big-data аналізу.
2. Процедури та прийоми big-data аналізу.
3. Кейси з використання big-data аналізу.

1. Історія виникнення та напрями big-data аналізу.

Термін «великі дані» вперше був використаний для позначення збільшення обсягів даних у середині 1990-х років. У 2001 році Даг Лейні, тодішній аналітик консалтингової компанії Meta Group Inc., розширив визначення великих даних.

Ці три фактори стали відомі як 3V великих даних. Gartner популяризував цю концепцію після придбання Meta Group і найму Лейні в 2005 році.

До 2011 року аналітика великих даних почала міцно закріплюватися в організаціях і громадськості.

Зовсім недавно більша кількість користувачів прийняла аналітику великих даних як ключову технологію цифрової трансформації. Серед користувачів – роздрібні торговці, фірми, що надають фінансові послуги, страховики, організації охорони здоров'я, виробники, енергетичні компанії та інші підприємства.

Великі дані – це термін, що описує великі набори різноманітних даних – структурованих, неструктурованих і напівструктурованих – які постійно генеруються з високою швидкістю та у великих обсягах. Зараз зростає кількість компаній, які використовують ці дані, щоб отримати значущу інформацію та покращити процес прийняття рішень, але вони не можуть зберігати та обробляти їх за допомогою традиційних пристроїв зберігання та обробки даних.

Великі дані. Аналітика зазвичай відбувається в режимі реального часу – під час генерації даних – і відкриття представлені майже миттєво.

Їх **атрибути** (відомі як чотири V).

Обсяг – це те, що є «великим» у Big Data. Це стосується від терабайтів до петабайтів інформації, що надходить із різних джерел, таких як пристрої Інтернету речей, соціальні мережі, текстові файли, бізнес-транзакції тощо. Щоб ви могли зрозуміти масштаб, 1 петабайт дорівнює 1 000 000 гігабайт. Під час перегляду один HD-фільм на Netflix займає понад 4 гігабайти. Тепер уявіть, що 1 петабайт містить 250 000 фільмів. Великі дані – це не 1 петабайт, а тисячі й мільйони їх.

Швидкість – це швидкість, з якою дані генеруються та обробляються. Це представлено у вигляді пакетного звітування, обробки майже в режимі реального часу та потокової передачі даних. Найкращий сценарій – коли швидкість створення даних відповідає швидкості їх обробки. Візьмемо, наприклад, транспортну галузь. Один автомобіль, підключений до Інтернету з підключеним телематичним пристроєм, генерує та передає 25 гігабайт даних щогодини з майже постійною швидкістю. І більшість цих даних потрібно обробляти в режимі реального часу або майже в реальному часі.

Різнороманітність – це вектор, що демонструє різноманітність великих даних. Це не просто структуровані дані, які містяться в реляційних базах даних

у вигляді рядків і стовпців. Він доступний у різноманітних формах, які відрізняються від однієї програми до іншої, і більшість великих даних є неструктурованими. Скажімо, проста публікація в соціальних мережах може містити текстову інформацію, відео чи зображення, позначку часу. тощо

Правдивість – це міра того, наскільки правдивими, точними та надійними є дані та яку цінність вони приносять. Дані можуть бути неповними, непослідовними або шумними, що знижує точність процесу аналітики. Через це правдивість даних зазвичай класифікують як хорошу, погану та невизначену. Це дуже корисно під час роботи з різноманітними наборами даних, як-от медичні записи, у яких будь-які невідповідності чи двозначності можуть мати шкідливі наслідки.

Сфера застосування аналізу

Аналітика ланцюга поставок і каналів. Прогнозні аналітичні моделі можуть допомогти з попереджувальним поповненням, мережами постачальників B2B, управлінням запасами, оптимізацією маршрутів і повідомленням про можливі затримки доставки.

Залучення та утримання клієнтів. Дані споживачів можуть допомогти маркетинговим зусиллям компаній, які можуть діяти відповідно до тенденцій, щоб підвищити задоволеність клієнтів. Наприклад, механізми персоналізації для Amazon, Netflix і Spotify можуть покращити взаємодію з клієнтами та створити лояльність клієнтів.

Запобігання шахрайству. Фінансові установи використовують інтелектуальний аналіз даних і машинне навчання, щоб зменшити ризик шляхом виявлення та прогнозування моделей шахрайської діяльності.

Операції. Аналіз фінансових даних допомагає організаціям виявляти та зменшувати приховані операційні витрати, у свою чергу економлячи гроші та підвищуючи продуктивність.

Охорона здоров'я. Аналітику великих даних можна використовувати, щоб отримати ключову інформацію з даних пацієнтів, що допомагає постачальникам послуг відкривати нові діагнози та варіанти лікування.

Персоналізація. Поточкові платформи та онлайн-магазини аналізують залучення користувачів, щоб створити більш персоналізований досвід у формі рекомендацій, цільової реклами, додаткових продажів і програм лояльності.

Покращене прийняття рішень. Статистика, яку бізнес-користувачі витягують із релевантних даних, може допомогти організаціям швидше та краще приймати рішення.

Розробка продукту. Аналітика великих даних може надавати інформацію про життєздатність продукту, рішення щодо розробки, вимірювання прогресу та спрямовувати вдосконалення в напрямку того, що підходить клієнтам бізнесу.

Управління ризиками. Аналітика великих даних може виявляти нові ризики на основі шаблонів даних для ефективних стратегій управління ризиками.

Ціноутворення. Дані про продажі та транзакції можна аналізувати для створення оптимізованих моделей ціноутворення, що допомагає компаніям приймати рішення щодо ціноутворення, які максимізують дохід.

Геоаналітика. Аналіз трафіка та локації під відкриття нових об'єктів, розширення мережі магазинів, проведення коректної комунікації з потенційними клієнтами.

2. Процедури та прийоми big-data аналізу.

Переваги використання аналітики великих даних включають:

1. Швидкий аналіз великої кількості даних із різних джерел у різних форматах і типах.

2. Швидке прийняття більш обґрунтованих рішень для ефективної розробки стратегії, що може принести користь і покращити ланцюжок постачання, операції та інші сфери прийняття стратегічних рішень.

3. Економія коштів, яка може бути результатом підвищення ефективності та оптимізації бізнес-процесів.

4. Краще розуміння потреб, поведінки та настроїв клієнтів, що може привести до кращого маркетингового розуміння, а також надати інформацію для розробки продукту.

Удосконалені стратегії управління ризиками з кращою інформацією, які базуються на великих обсягах вибірки даних.

Проблеми аналітики великих даних

Доступність даних. З більшими обсягами даних зберігання та обробка ускладнюються. Великі дані слід зберігати та обслуговувати належним чином, щоб вони могли використовуватися менш досвідченими спеціалістами з обробки даних та аналітиками.

Підтримка якості даних. Оскільки великі обсяги даних надходять із різноманітних джерел і в різних форматах, керування якістю великих даних потребує значного часу, зусиль і ресурсів для їх належного обслуговування.

Безпека даних. Складність систем великих даних створює унікальні виклики безпеці. Належне вирішення питань безпеки в такій складній екосистемі великих даних може бути складним завданням.

Вибір відповідних інструментів. Вибір із величезної кількості доступних на ринку інструментів і платформ для аналізу великих даних може бути заплутаним, тому організації повинні знати, як вибрати найкращий інструмент, який відповідає потребам користувачів та інфраструктурі.

Через потенційну нестачу навичок внутрішньої аналітики та високу вартість найму досвідчених спеціалістів із обробки даних та інженерів деяким організаціям важко заповнити прогалини.

Джерела отримання інформації

- 1) дані потоку кліків в Інтернеті;
- 2) журнали веб-сервера;
- 3) хмарні програми;
- 4) мобільні додатки;
- 5) контент соціальних мереж;
- 6) текст із електронних листів клієнтів і відповідей на опитування;

- 7) записи мобільних телефонів;
- 8) машинні дані, отримані датчиками, підключеними до Інтернету речей.

Зібрані, оброблені та очищені дані аналізуються за допомогою аналітичного програмного забезпечення. Це включає інструменти для:

- 1) інтелектуальний аналіз даних, який просуває набори даних у пошуках закономірностей і зв'язків;
- 2) прогнозна аналітика, яка створює моделі для прогнозування поведінки клієнтів та інших майбутніх дій, сценаріїв і тенденцій;
- 3) машинне навчання, яке використовує різні алгоритми для аналізу великих наборів даних;
- 4) глибоке навчання, яке є більш просунутим відгалуженням машинного навчання;
- 5) програмне забезпечення для інтелектуального аналізу тексту та статистичного аналізу;
- 6) штучний інтелект (AI);
- 7) основне програмне забезпечення бізнес-аналітики;
- 8) засоби візуалізації даних.

Виділяють чотири ключові типи аналітики великих даних.

Описова аналітика – це поширений вид аналітики, який дозволяє дізнатися, що сталося і коли.

Діагностична аналітика пояснює, чому та як щось сталося, визначаючи шаблони та зв'язки в доступних даних.

Прогностична аналітика використовує історичні дані, щоб виявити закономірності та спрогнозувати, що, ймовірно, станеться в майбутньому.

Наказова аналітика дає конкретні рекомендації щодо того, що слід зробити краще.

Аналіз даних

Обробка природної мови – це технологія, яка використовується для того, щоб комп'ютери розуміли людську мову та реагували на неї, будь то текст чи вимовлені слова.

Інтелектуальний аналіз тексту – це розширений аналітичний підхід, який використовується для розуміння великих даних у текстових формах, таких як електронні листи, твіти, дослідження та публікації в блогах.

Аналіз даних датчиків – це дослідження даних, які постійно генеруються різними датчиками, встановленими на фізичних об'єктах. Якщо це зробити вчасно та належним чином, це може допомогти не лише отримати повну картину стану обладнання, але й виявити несправності та передбачити несправності.

Аналіз викидів або виявлення аномалій – це техніка, яка використовується для визначення точок даних і подій, які відрізняються від решти даних. Він широко використовується в діяльності з виявлення шахрайства.

3. Кейси з використання big-data аналізу.

Електронна комерція: **Amazon** аналізує великі дані, щоб покращити свій механізм рекомендацій.

Не дивно, що гігант електронної комерції та технологій Amazon збирає та аналізує величезну кількість даних про кожного клієнта компанії. Це може бути така інформація, як історія веб-перегляду чи покупок користувача, демографічні дані тощо. Аналіз цих даних допомагає Amazon скласти повну картину кожного користувача та відповідно створити більш цілеспрямовані рекламні кампанії.

Marriott приймає рішення на основі аналізу великих даних.

Marriott – це американська багатонаціональна компанія, яка володіє різними закладами гостинності по всьому світу. Компанія є чудовим прикладом того, як аналітику великих даних можна використовувати для прийняття бізнес-рішень і отримання конкурентних переваг у галузі.

Marriott застосовує підхід автоматизації динамічного ціноутворення до управління доходами, що дозволяє компанії робити точні прогнози щодо попиту та моделей поведінки клієнтів. Щоб ця модель працювала, потрібні дані в реальному часі різних видів. Це можуть бути такі показники, як дохід на доступний номер, заповнюваність і скасування, поведінка бронювання тощо, або дані про погоду, події, глобальну та місцеву економічну ситуацію. Аналізуючи величезні обсяги цих даних, мережа готелів може зрозуміти, як її заклади справляються з конкурентами, і завчасно коригувати свою стратегію ціноутворення для кращих результатів.

Обладнання: GE використовує вдосконалену аналітику великих даних для оптимізації вітрової електростанції.

General Electric – глобальна цифрова промислова компанія, що надає послуги, обладнання та програмні рішення в різних галузях – від охорони здоров'я до авіації та екологічної енергетики. Компанія встановила датчики в обладнання в усіх галузях промисловості, щоб контролювати кожен окремий аспект, який може вплинути на роботу обладнання.

Наприклад, кожна з 200 їхніх вітряних турбін містить майже 50 датчиків, які безперервно передають масу оперативних даних у хмару. Дані датчика використовуються для конфігурації напрямку та кроку лопатей турбіни, щоб забезпечити фіксацію максимальної енергії обертання. Крім того, ці дані дають команді з експлуатації об'єкта можливість оцінити стан і продуктивність кожної турбіни. Використання Big Data допомагає компанії точно налаштувати процеси та зменшити час простою та втрати.

Компанія **American Express** ставить аналітику великих даних в основу прийняття рішень. Станом на 2019 рік налічувалося понад 110 мільйонів карток із понад 8 мільярдами транзакцій. Це багато даних, з яких можна вчитися. Як і інші фінтех-компанії, American Express вважає кібербезпеку своїм головним пріоритетом. Тому фірма інвестувала кошти у створення моделі ML, яка використовує різні типи даних, як-от деталі витрат, інформацію про членство картки, інформацію про продавця тощо, щоб виявити будь-яку підозрілу діяльність і миттєво прийняти рішення щодо запобігання шахрайству з кредитними картками.