**Лекція №12. Сучасні тенденції розвитку технологій в СУБД. Бази знань. Великі Дані**

**Тенденції розвитку систем баз даних**

***Тенденції зростання обсягів інформації систем*.** Має місце тенденція зростання обсягів інформації, що зможе зберігатися в СУБД. Ця тенденція базується, з одного боку, на досягненнях виробників комп’ютерів в галузі створення нових носіїв для зберігання інформації, а, з іншого боку, на зростаючих обсягах інформації, яку потрібно зберігати і обробляти. Тенденція зростання систем пояснюється також і тим, що збільшуються обсяги зберігання графічної інформації, відео- та аудіо-записів.

Завдання обслуговування великих баз даних потребує неординарних технологічних рішень. Бази даних навіть невеликих розмірів сьогодні розміщують на дискових носіях, що звуться вторинними устроями зберігання інформації. Нижче перелічені технологічні рішення, що дозволяють СУБД справлятися із невпинно зростаючими потоками даних.

***Паралельні обчислення*.** Здатність системи зберігати великі обсяги даних важлива, але вона не вирішує повністю проблему обслуговування великих баз даних, які вимагають великої швидкодії. Існують два шляхи підвищення швидкодії, а саме: реалізація прямого доступу до даних на основі використання структур індексів і застосування засобів паралельних обчислень. Паралелізм обчислень реалізується у декількох напрямках.

Ідея паралелізму використовується у алгоритмах обробки запитів до бази даних, які дозволяють “розділяти ” множину запитів на частини таким чином, щоб дозволити паралельним комп’ютерам, або розподіленим мережам комп’ютерів ефективно використовувати усі наявні обчислювальні ресурси.

***Системи клієнт/сервер і багаторівневі архітектури***

Значний пласт сучасного програмного забезпечення підтримує архітектуру *клієнт/сервер* , у відповідності із якою запити, сформовані одним процесом (*клієнтом*), відсилаються для обробки другому процесу (*серверу*). СУБД також використовує таку архітектуру, розділяючи свої функції між процесом сервера і клієнтськими процесами.

Простіший варіант архітектури “*клієнт/сервер*” передбачає, що СУБД цілком представляє собою сервер, за виключенням інтерфейсів запитів, які взаємодіють із користувачем і відсилають запити і інші команди на мові SQL на сервер з метою їх виконання. Результати виконання запитів сервером повертаються клієнтом у формі таблиць, або відношень. В останній час широке розповсюдження отримали багаторівневі архітектури, у яких СУБД відводиться роль постачальника Web-сайтов. СУБД діє як сервер бази даних, але його клієнтом є *сервер додатків* (application server), який керує підключеннями до бази даних, транзакціями, авторизацією та іншими процесами. Клієнтами серверів додатків у свою чергу можуть бути Web-сервери, що обслуговують кінцевих користувачів та інші програмні додатки.

***Дані мультимедіа***

Інша важлива тенденція розвитку сучасних СУБД пов’язана із підтримкою даних мультимедіа. Застосовуючи термін мультимедіа (multimedia), треба під ним розуміти інформацію у вигляді сигналу певного виду - аудіо і відеозаписи, сигнали, що отримуються від радарів, супутникові зображення, також документи і графіка різних форматів. Загальною властивістю даних мультимедіа являється великий обсяг, здатний змінюватися у широких межах, що відрізняє їх від традиційних одиниць представлення інформації – цілих чисел, рядків фіксованої довжини і т.п.

Щоб забезпечити можливість виконання маніпуляцій складною інформацією (наприклад, операцій по обробці графічних зображень), СУБД повинна дозволяти користувачам визначати нові функції, які необхідні в той чи в іншій ситуації. Розширення функціонального потенціалу систем часто здійснюється на основі об’ектно-орієнтованого підходу.

Великі розміри об’єктів даних мультимедіа заставляють розробників СУБД модифікувати функції менеджерів зберігання даних, щоб забезпечити можливість розміщення у базі даних об’єктів або кортежів обсягом у гігабайт і більше. Однією із серйозних проблем, яка виникла внаслідок непомірного зростання обсягів елементів даних, зв’язана із доставкою клієнту результатів відправленого ним запиту. У контексті реляційної бази даних результатом виконання запиту являється набір кортежів, які повертаються клієнту як єдине ціле, результатом запиту може бути відео кліп обсягом 1 гігабайт. Доставка його клієнту як неподільне ціле не може бути не може бути у всіх випадках коректною. По-перше, його доставка клієнту може заблокувати на значний час обробку сервером усіх інших запитів, що чекають своєї черги. По-друге, клієнта може задовольнити невелика частка кліпу, якщо він зможе переглянути якийсь початковий фрагмент кліпа і ,тільки потім, мати можливість сформулювати у запиті, який вигляд повинний мати результат запиту. В-третіх, клієнт повинен мати можливість вказувати у запиті спосіб передачі результуючого кліпу (наприклад, передавати інформацію частками на протязі певного часу). Таким чином СУБД, що підтримує формати даних мультимедіа, повинна пропонувати інтерактивний режим обслуговування запитів.

***Інтеграція інформації***

Проблема інтеграції інформації виникає при створенні великих компаній за рахунок поглинання ними середніх компаній. Централізація управління у рамках всієї компанії вимагає створення єдиної центральної бази даних. У той же час слід враховувати, що підрозділи компанії, можливо, у свій час інвестували значні кошти у створення власних баз даних. Заміняти *успадковані бази даних* (legacy databases) їх новою базою даних не є раціональним рішенням. Більш розумно побудувати “зверху ” існуючих успадкованих баз даних нову інформаційну структуру, здатну представити всю продукцію компанії у погодженому і послідовному вигляді.

Один із самих популярних підходів до рішення цієї задачі пов’язаний із використанням технологій *сховищ даних* (data warehouses), які передбачають копіювання інформації із успадкованих баз даних із відповідною трансляцією і наступним зберіганням у центральній базі даних. При зміні стану успадкованої бази даних необхідні зміни вносяться і у вміст сховища, хоча необов’язково автоматично і миттєво. Дуже часто реплікацію даних здійснюють вночі, коли ймовірність завантаження успадкованих баз даних найбільш низька.

Успадкована база даних, таким чином, продовжує виконувати усі звичайні функції, передбачені у період її проектування, а нові, такі як підтримка електронних каталогів у Web, покладається на сховище даних. Вміст сховищ даних використовується також для цілей аналізу і планування: аналітики компанії отримують можливість звертатися до сховища даних із запитами, що дозволяють визначати тенденції продаж продукції, оптимізувати асортимент, спланувати подальший розвиток підприємства. Сховища даних відкривають перспективи застосування новітніх технологій “*розробки даних* ” (data mining) – пошуку цікавих і незвичайних зразків інформації і використання їх для оптимізації бізнес-процесів

Перспективи розвитку архітектур СУБД пов'язані з розвитком концепції обробки нетрадиційних даних та їх інтеграцією, обміном даними з різних СУБД, багатокористувацької технології в локальних мережах. Одна з найважливіших тенденцій розвитку СУБД - розробка "універсальних" СУБД, які здатні інтегрувати в базі традиційні і нетрадиційні дані - тексти, малюнки, звук і відео, сторінки HTML (HyperText Markup Language, мова розмітки гіпертексту) тощо. Це особливо актуально для Web.

Є два підходи до побудови таких СУБД: об'єктно-реляційний - удосконалення реляційних СУБД та об'єктний.

Шляхом створення об'єктно-реляційних СУБД пішли такі фірми, як IBM, Informix та Oracle. IBM розробила об'єктно-реляційну СУБД DB2 для ОС OS/2.

Інформаційні сховища на базі СУБД з паралельною обробкою розраховані на багатопроцесорні системи. Такі СУБД поділяються за типом архітектури - без поділу ресурсів та зі спільним використанням дискового простору. У першому випадку за кожним із процесорів закріплено виділені ділянки пам'яті і диски, що дає велику швидкість обробки. У другому - всі процесори розподіляють між собою як оперативну пам'ять, так і місце на диску.

Прикладами СУБД без розподілу ресурсів є: DB2 (IBM), Informix Online Dynamic (Informix). СУБД зі спільним використанням пам'яті - це: AdabasD версія 6.1 (Software AG).

***База знань***

База знань - це база даних, яка використовується для обміну та управління знаннями, сприяє збору, організації та пошуку знань. Багато баз знань побудовані навколо штучного інтелекту і не тільки зберігають дані, але і знаходять рішення для подальших проблем, використовуючи дані попереднього досвіду, що зберігаються як частина бази знань.

Системи управління знаннями залежать від технологій управління даними, починаючи від реляційних баз даних до сховищ даних. Деякі бази знань - це трохи більше, ніж індексована енциклопедична інформація; інші - інтерактивні та поводяться / відповідають відповідно до інформації, запропонованої користувачем.

Різні методи представлення знань, включаючи кадри та сценарії, являють собою знання. Пропоновані послуги - це пояснення, міркування та розумна підтримка рішення.

Два основні типи баз знань є читабельними та машиночитаними.

* Читані людські бази знань дозволяють людям отримувати доступ та використовувати ці знання. Вони зберігають довідкові документи, посібники, інформацію про усунення несправностей і часто відповідають на запитання. Вони можуть бути інтерактивними та вести користувачів до вирішення проблем, але вони розраховують на надання користувачем інформації для керівництва процесом.
* Машиночитані бази знань зберігають знання, але лише у системах, доступних для читання. Рішення пропонуються на основі автоматизованих дедуктивних міркувань і не є настільки інтерактивними, оскільки це спирається на системи запитів, які мають програмне забезпечення, яке може відповісти на базу знань, щоб звузити рішення. Це означає, що інформація про машиночитану базу знань, що ділиться з іншими машинами, зазвичай лінійна і обмежена в інтерактивності, на відміну від взаємодії людини, яка базується на запитах.

Управління знаннями (КМ) містить цілий спектр стратегій, що застосовуються в організації для створення, представлення, аналізу, поширення та забезпечення можливості переживання досвіду. Він орієнтований на конкурентні переваги та підвищення ефективності діяльності організацій.

Знання можуть бути класифіковані по наступних категоріях:

* поверхневі - знання про видимі взаємозв'язки між окремими подіями і фактами в наочної області;
* глибинні - абстракції, аналогії, схеми, що відображують структуру і процеси в наочної області.

Знання можна розділити на процедурні і декларативні. Історично первинними були процедурні знання, тобто знання, "розчинені" в алгоритмах. Вони управляли даними. Для їх зміни необхідно було змінювати програми. Проте з розвитком штучного інтелекту пріоритет даних поступово змінювався, і все більша частина знань зосереджувалася в структурах даних (таблиці, списки, абстрактні типи даних), тобто збільшувалася роль декларативних знань.

Сьогодні знання набули чисто декларативної форми, тобто знаннями вважаються речення, записані на мовах подання знань, наближених до природніх і зрозумілих неспеціалістам.

Стратегії здобуття знань

Існує декілька стратегій здобуття знань. Найбільш поширені:

* придбання;
* витягання;
* формування.

Під придбанням знань розуміється спосіб автоматизованої побудови бази знань за допомогою діалогу експерта і спеціальної програми (при цьому структура знань заздалегідь закладається в програму). Ця стратегія вимагає істотного попереднього опрацювання наочної області. Систем придбання знань дійсно набувають готові фрагменти знань відповідно до структур, закладених розробників систем. Більшість цих інструментальних засобів орієнтована на конкретні експертні системи з жорстко визначеною предметною областю і моделлю подання знань, тобто не є універсальними.

Термін витягання знань стосується безпосереднього живого контакту інженера по знаннях і джерела знань.

Термін формування знань традиційно закріпився за розробкою моделей, методів і алгоритмів аналізу даних для здобуття знань і вчення, що активно розвивається. Ця область включає індуктивні моделі формування гіпотез на основі повчальних вибірок, вчення аналогічно і інші методи.

База знань – сукупність знань, що відносяться до деякої предметної області, формально представлених так, щоб на їх основі можна було здійснювати міркування. Бази знань найчастіше використовуються в контексті експертних систем, де з їх допомогою подаються навики і досвід експертів, зайнятих практичною діяльністю у відповідної області (наприклад, в медицині або в математиці). Зазвичай база знань є сукупністю правил виводу.

Не дивлячись на всі недоліки, найбільшого поширення набула продукційна модель подання знань. При використанні продукційної моделі база знань складається з набору правил. Програма, що управляє перебором правил, називається машиною виведення. Машина виведення (інтерпретатор правил) виконує дві функції: по-перше, перегляд існуючих фактів з робочої пам'яті (бази даних) і правил з бази знань і додавання (в міру можливості) в робочу пам'ять нових фактів, а по-друге, визначення порядку перегляду і вживання правил. Цей механізм управляє процесом консультації, зберігаючи для користувача інформацію про отримані висновки, і запрошує у нього інформацію, коли для спрацьовування чергового правила в робочій пам'яті виявляється недостатньо даних.

***Елементи експертних систем***

Експертна система – це комплекс комп'ютерного програмного забезпечення, що допомагає людині приймати обґрунтовані рішення. Експертні системи використовують інформацію, отриману заздалегідь від експертів – людей, які в якій-небудь області є кращими фахівцями. Всі експертні системи включають, принаймні, три основні елементи: базу знань, машину виводу і інтерфейс користувача.

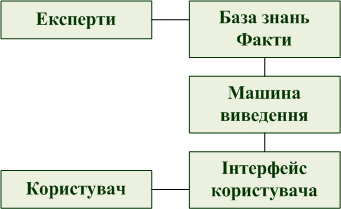


Рисунок 1 – Елементи експертної системи

***База знань*.**База знань містить відомі факти, виражені у вигляді об'єктів, атрибутів і умов. Окрім описових уявлень про дійсність, вона включає вирази невизначеності – обмеження на достовірність факту. В цьому відношенні вона відрізняється від традиційної бази даних внаслідок свого символьного, а не числового або буквеного вмісту. При обробці інформації бази даних користуються заздалегідь визначеними логічними правилами Відповідно, база знань, що представляє вищий рівень абстракції, має справу з класами об'єктів, а не з самими об'єктами. База знань створюється людьми – консультантами.

***Машина виведення*.**Головним в експертній системі є механізм, що здійснює пошук в базі знань за правилами раціональної логіки для здобуття рішень. Ця машина виведення приводиться в дію при здобутті запиту користувача і виконує такі завдання:

* порівнює інформацію, що міститься в запиті користувача, з інформацією бази знань;
* шукає певну мету або причинні зв'язки;
* оцінює відносну визначеність фактів, ґрунтуючись на відповідних коефіцієнтах довіри, пов’язаних з фактом.

Машина виведення призначена для побудови висновків. Її дія аналогічно міркуванням експерта-людини, яка оцінює проблему і пропонує гіпотетичні рішення. У пошуку цілей на основі запропонованих правил, машина виведення звертається до бази знань до тих пір, поки не знайде вірогідну дорогу до здобуття прийнятного результату.

***Інтерфейс користувача*.**Завдання інтерфейсу користувача полягає в організації обміну інформацією між оператором і машиною виводу. Інтерфейс прямого введення повинен уміти розпізнавати мову, або, принаймні, достатню кількість ключових слів і фраз, щоб уловлювати їх зв'язок з даною проблемою і її пропонованими рішеннями.

***Аспект людський*.**У роботі з експертними системами беруть участь як мінімум три групи людей. По-перше, адміністрація встановлює призначення експертної системи, обмежує предметну область, яку повинна охоплювати система, і точно визначає, які вигоди організація зможе отримувати з її використання. По-друге, фахівець із збору знань (інженер - когнітолог) збирає інформацію, необхідну для бази знань, порівнює відповідні дані і евристично організовує інформацію. По-третє, потенційний користувач вказує, як використовуватиметься система, якого роду проблеми належить вирішувати, і яким чином здійснюватиметься взаємодія програми з оператором. І, нарешті, системі потрібний експерт (частіше група експертів) у встановленої наочної області, для здобуття від нього знань, як у формі фактичної інформації, так і відносно аналітичних методів, які застосовуються для вирішення проблем в цій області.

***Машинний аспект*.**Комп'ютерну частину системи представляють компоненти програмного забезпечення, які обробляють отриману інформацію про дійсність, закладену в символьному вигляді в базу знань. У базу знань поступають факти. Зв'язок між фактами представлений евристичними правилами – виразами декларативного знання про стосунки між об'єктами. Кожне таке правило має складову "якщо і компонент "те" (висновок-дія), які визначають прямий або зворотний причинно-наслідковий зв'язок. Дійсні твердження лише вірогідні, іншими словами, міра їх визначеності не завжди абсолютна. Загальноприйнята схема полягає в тому, щоб варіювати рівень довіри від 0, що представляє мінімальну міру визначеності, до 100 – вища міра.

**Big Data - Великі дані**

Це набір даних, який є величезним і складним, так що традиційні програми обробки даних є недостатніми для роботи з ними. Існують проблеми з управлінням таким величезним обсягом даних, як збір, зберігання, аналіз даних, передача даних, обмін даними тощо.

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.

Набір даних VVV (volume, velocity, variety — фізичний об’єм, швидкість приросту даних і необхідність їх швидкої обробки, здатність обробляти дані різних типів) був розроблений компанією Meta Group у 2001 році з метою вказати на рівну значимість управління даними по всім трьом аспектам. Ці характеристики (англійською – 3 V):

* великий обсяг даних (**Volume**)

***Обсяг*** є найбільш часто цитованою характеристикою великих даних. Це сукупність даних з різних джерел, починаючи від чітко визначених і закінчуючи слабко визначеними. Вони походять від людських або машинних джерел.

* різноманітність даних, широкий вибір типів даних, що зберігаються в системах великих даних (**Variety**)

Big Data 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 в озері даних, яке, як правило, базується 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.

* швидкість, з якою дані генеруються, збираються та обробляються (**Velocity**)

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

Нещодавно до опису того, що таке Big Data, було додано кілька нових V:

* достовірність (**Veracity**)

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.

Погані 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 ІТ та аналітики також потрібно забезпечити наявність достатньо точних даних для отримання достовірних результатів.

* цінність (**Value**)

Не всі зібрані дані мають реальну ділову цінність, і використання неточних даних може послабити результат, що надаються додатками аналітики. 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.

* мінливість (**Variability**)

Варіабельність також часто згадується, коли мова йде про набори великих даних, які є менш послідовними, ніж звичайні, і можуть мати кілька значень або можуть бути відформатовані по-різному.

Хоча великі дані кількісно не прирівнюються до будь-якого конкретного обсягу даних, розгортання великих даних часто включає в себе терабайти (TB), петабайти (PB) і навіть екзабайти (EB) даних.

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.

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.

Необхідність швидкої передачі і обробки великих даних спричиняє вимоги до базової обчислювальної інфраструктури.

 Обчислювальна потужність, необхідна для швидкої обробки величезних обсягів та різновидів даних, може перегрузити сервер або кластер серверів.

Організації повинні застосовувати адекватну обробну потужність для завдань пов’язаних із Big Data, щоб досягти необхідної швидкості операцій.

Це потенційно може вимагати сотень або тисяч серверів, які можуть розподіляти обробні роботи та працювати спільно в кластерній архітектурі, часто заснованій на таких технологіях, як Hadoop та Apache Spark.

Як результат, загальнодоступні хмарні обчислення зараз є основним засобом розміщення систем великих даних. Публічний хмарний провайдер може зберігати петабайти даних і збільшувати необхідну кількість серверів. Бізнес платить лише за фактично використаний час зберігання та обчислень, а хмарні екземпляри можна вимкнути, поки вони знову не знадобляться.

Щоб ще більше покращити рівень обслуговування, провайдери загальнодоступних хмарних служб пропонують можливості передачі великих даних за допомогою керованих служб:

* Amazon EMR (раніше Elastic MapReduce)
* Microsoft Azure HDInsight
* Google Cloud Dataproc

У хмарних середовищах великі дані можна зберігати в наступних системах:

* Розподілена файлова система Hadoop (HDFS)
* Дешевші сховища хмарних об’єктів, такі як Amazon Simple Storage Service (S3)
* Бази даних NoSQ
* Реляційні бази даних

Для організацій, які хочуть розгорнути локальні системи великих даних, зазвичай використовуються технології з відкритим кодом Apache на додаток до Hadoop та Spark і включають наступне:

* YARN, вбудований менеджер ресурсів і планувальник робіт Hadoop
* програма програмування MapReduce, яка також є основним компонентом Hadoop;
* Kafka, платформа обміну повідомленнями та передачі даних від програми до програми
* HBase бази даних
* Системи запитів SQL-on-Hadoop, такі як Drill, Hive, Impala та Presto.

Користувачі можуть самостійно встановити версії технологій з відкритим кодом або звернутися до комерційних платформ великих даних.

Окрім проблем обсягу обробки та вартості, проектування архітектури великих даних є ще однією загальною проблемою для користувачів. Системи великих даних повинні бути адаптовані до конкретних потреб організації, це робота, що вимагає від ІТ-команд та розробників програм складання набору інструментів з усіх доступних технологій.

Розгортання та управління системами великих даних також вимагають нових навичок порівняно з тими, якими володіють адміністратори баз даних (DBA) та розробники.

Обидві ці проблеми можна полегшити за допомогою керованої хмарної служби, але ІТ-менеджери повинні пильно стежити за використанням хмар, щоб переконатися, що витрати не виходять з-під контролю. Крім того, перенесення локальних наборів даних та обробка робочих навантажень у хмару часто є складним процесом для організацій.

Зробити дані в системах великих даних доступними для аналітиків також є проблемою, особливо в розподілених середовищах, що включають поєднання різних платформ та сховищ даних. Щоб допомогти аналітикам знаходити відповідні дані, команди ІТ та аналітики все частіше працюють над створенням каталогів даних, що включають функції управління метаданими та функціями лінійки даних.

Якість даних та управління даними також повинні бути пріоритетами для забезпечення чистоти, послідовності та правильного використання наборів Big Data.

Протягом багатьох років компанії не мали обмежень щодо даних, які вони збирали від своїх клієнтів. Однак, оскільки збір та використання великих даних збільшився, зростає і зловживання ними. Стурбовані громадяни вимагають прийняття законів щодо прозорості збору даних та конфіденційності даних споживачів.

**GDPR в ЄС**. Резонанс щодо порушень конфіденційності особистої інформації змусив Європейський Союз прийняти Загальний регламент про захист даних (GDPR), який набрав чинності в травні 2018 року. Він обмежує типи даних, які організації можуть збирати, і вимагає згоди окремих осіб або дотримання інших визначених законних підстав для збору персональних даних. GDPR також включає положення, яке дозволяє жителям ЄС просити компанії видаляти їх дані.

**CCPA в США.** Хоча в США немає подібних федеральних законів, Каліфорнійський закон про конфіденційність споживачів (CCPA) має на меті надати жителям Каліфорнії більше контролю над збором та використанням їх особистої інформації компаніями. Закон CCPA був підписаний у 2018 році і набув чинності 1 січня 2020 року. Крім того, урядовці США розслідують практику обробки даних, зокрема серед компаній, які збирають споживчі дані та продають їх іншим компаніям для невідомого використання.

Людська сторона аналітики великих даних

Зрештою, цінність та ефективність Big Data залежать від працівників, яким доручено аналізувати дані та формулювати відповідні запити для спрямування проектів аналізу великих даних. Деякі інструменти для обробки Big Data відповідають спеціалізованим нішам і дозволяють менш технічним користувачам використовувати повсякденні ділові дані в програмах прогнозної аналітики.

Інші технології, такі як пристрої Big Data на базі Hadoop, допомагають компаніям впровадити відповідну обчислювальну інфраструктуру для реалізації проектів Big Data, мінімізуючи при цьому потребу в апаратному та програмному забезпеченні.

Big Data можна порівняти з малими даними, іншим терміном, який часто використовується для опису даних, обсяг і формат яких дозволяють легко проаналізувати ці дані самостійно.

Зазвичай цитується аксіома: **“великі дані – для машин; малі – для людей”**.

Джерело: [searchdatamanagement.techtarget.com/](https://searchdatamanagement.techtarget.com/definition/big-data)

«Великі дані (Big Data) – позначення структурованих и неструктурованих даних величезних обсягів і значного розмаїття, що піддаються ефективній обробці програмних інструментів, які горизонтально масштабуються та з’явились у кінці 2000-х років, і альтернативних традиційних систем управління базами даних і рішенням класу рішень Business Intelligence».

Можна користуватись і більш простим визначенням, яке цілком відповідає усталеним і більш простим визначенням, що цілком відповідають думці журналістів і маркетологів. «Великі дані – це сукупність технологій, покликаних здійснювати три операції:

1. Обробляти більші, у порівнянні зі «стандартними» сценаріями, об’єми даних.

2. Уміти працювати з даними, що швидко надходять у дуже великих об’ємах. Тобто даних не просто багато, а їх постійно стає все більше й більше.   
3. Вміти працювати зі структурованими і мало стуктурованими даними паралельно і у різних аспектах.»

Вважається, що ці «вміння» дозволяють виявляти приховані закономірності, що вислизають від обмеженого людського сприйняття. Це дає безпрецедентні можливості оптимізації багатьох сфер нашого життя: державного управління, медицини, телекомунікацій, фінансів, транспорту, виробництва і так далі. Не дивно, що журналісти і маркетологи так часто використовували словосполучення Big Data, що багато експертів вважають цей термін дискредитованим і пропонують від нього відмовитись.

Більш того, у жовтні 2015 року компанія Gartner виключила Big Data з числа популярних трендів. Своє рішення аналітики компанії пояснили тим, що до складу поняття «великі дані» входить значна кількість технологій, які вже активно застосовуються на підприємствах, вони частково стосуються інших популярних сфер і тенденцій і стали повсякденним робочим інструментом.

Виходячи з вищеназваних визначень, основні принципи роботи з великими даними такі:

* **Горизонтальна масштабованість.** Це — базовий принцип обробки великих даних. Як вже було зазначено, великих даних з кожним днем стає все більше. Відповідно, необхідно збільшувати кількість обчислювальних вузлів, за якими розподіляються ці дані, при чому обробка має відбуватись без погіршення продуктивності
* **Відмовостійкість.** Цей принцип витікає з попереднього. Оскільки обчислювальних вузлів у кластері може бути багато (іноді десятки тисяч) та їх кількість, не виключено, буде збільшуватись, зростає ймовірність виходу машин з ладу. Методи роботи з великими даними мають враховувати ймовірність таких ситуацій і передбачати превентивні заходи
* **Локальність даних.** Оскільки дані розподілені по великій кількості обчислювальних вузлів, то, якщо вони фізично знаходяться на одному сервері, а обробляються на іншому, витрати на передачу даних можуть бути невиправдано великими. Тому обробку даних бажано проводити на тій же машині, на якій вони зберігаються

Ці принципи відрізняються від тих, які характерні для традиційних, централізованих, вертикальних моделей зберігання добре структурованих даних.

***Технології і тенденції роботи з Big Data***

Початково у сукупність підходів і технологій включались засоби масово-паралельної обробки невизначено-структурованих даних, такі як СУБД NoSQL, алгоритми MapReduce і засоби проекту Hadoop. У подальшому до технологій великих даних почали відносити й інші рішення, що забезпечують схожі за характеристиками можливості обробки надвеликих масивів даних, а також деякі апаратні засоби.

* **MapReduce** — модель розподілених обчислювань у комп’ютерних кластерах, представлена компанією Google. 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.
* **NoSQL** (від англ. Not Only SQL, не лише SQL) — загальний термін для різних нереляційних баз даних і сховищ, не означає якусь конкретну технологію чи продукт. Звичайні реляційні бази даних добре підходять для досить швидких і однотипних запитів, а на складних і гнучко побудованих запитах, характерних для великих даних, навантаження перевищує розумні межі і використання СУБД стає неефективним.
* **Hadoop** — набор утиліт, бібліотек і фреймворків, що вільно розповсюджується, для розробки і виконання розподілених програм, які працюють на кластерах із сотень і тисяч вузлів. Вважається однією з основоположних технологій більшості даних.
* **R** — мова програмування для статистичної обробки даних і роботи з графікою. Широко використовується для аналізу даних і фактично став стандартом для статистичних програм.
* **Апаратні рішення.** Корпорації Teradata, EMC та ін. др. пропонують апаратно-програмні комплекси, призначені для обробки великих даних. Ці комплекси поставляються як готові до установки телекомунікаційні шафи, що містять кластер серверів і керівне програмне забезпечення для масово-паралельної обробки. Сюди іноді відносять апаратні рішення для аналітичної обробки в оперативній пам’яті, зокрема, апаратно-програмні комплекси Hana компанії SAP і комплекс Exalytics компанії Oracle, незважаючи на те, що така обробка початково не є масово-паралельною, а об’єми оперативної пам’яті одного вузла обмежуються кількома терабайтами.

Консалтингова компанія McKinsey, окрім технологій NoSQL, MapReduce, Hadoop, R, які розглядає більшість аналітиків, включає у контекст придатності для обробки великих даних також технології Business Intelligence і реляційні системи управління базами даних з підтримкою мови SQL.

***Методи і техніка аналізу великих 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 даних) — 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 (розгалуження на категорії), кластерний аналіз, регресійний аналіз, виявлення і аналіз відхилень тощо.
* **Краудсорсинг** — 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 природної мови, включно з тональним аналізом).
* **Машинне навчання**, включаючи навчання з учителем і без учителя – використання моделей, побудованих на базі статистичного аналізу 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)
* **Розпізнавання образів**
* **Прогнозна аналітика**
* **Імітаційне моделювання** (simulation) — метод, що дозволяє будувати моделі, що описують процеси так, як вони би проходили у дійсності. Імітаційне моделювання модна розглядати як різновид експериментальних випробувань.
* **Просторовий аналіз** (spatial analysis) — клас методів, що використовують топологічну, геометричну і географічну інформацію, що вилучається із даних.
* **Статистичний аналіз** — аналіз часових рядів, A/B-тестування (A/B testing, split testing — метод маркетингового дослідження; при його використанні контрольна група елементів порівнюється із набором тестових груп, у яких один чи кілька показників були змінені, щоб з’ясувати, які зі змін покращують цільовий показник.
* **Візуалізація аналітичних даних** — подання інформації у вигляді малюнків, діаграм, з використанням інтерактивних можливостей і анімації, як для отримання результатів, так і для використання у якості вихідних даних для подальшого аналізу. Дуже важливий етап аналізу великих даних, що дозволяє показати найважливіші результати аналізу у найбільш зручному для сприйняття вигляді.

Технології Big Data можуть бути корисними при вирішенні наступних задач:

* прогнозування ринкової ситуації;
* маркетинг і оптимізація продажів;
* вдосконалення продукції;
* ухвалення управлінських рішень;
* підвищення продуктивності праці;
* ефективна логістика;
* моніторинг стану основних фондів.

На виробничих підприємствах великі дані генеруються також внаслідок впровадження підприємства, великі дані генеруються також внаслідок впровадження технологій Промислового інтернету речей. У ході цього процесу основні вузли і деталі станків і машин оснащуються датчиками, виконавчими пристроями, контролерами та, іноді, недорогими процесорами, здатними виробляти граничні (туманні) обчислення. В процесі виробничого процесу здійснюється постійний збір даних і, можливо, їх попередня обробка (наприклад, фільтрація). Аналітичні платформи обробляють результати у найбільш зручному для сприйняття вигляді і зберігають для подальшого використання. На основі аналізу отриманих даних робляться висновки про стан обладнання, ефективність внесених змін у технологічні процеси.

Завдяки моніторингу інформації у режимі реального часу персонал підприємства має змогу:

* скорочувати кількість простоїв;
* підвищувати продуктивність обладнання;
* зменшувати витрати на експлуатацію обладнання;
* запобігати нещасним випадкам.

Останній пункт особливо важливий. Наприклад, оператори, що працюють на підприємствах нафтопереробної промисловості, отримують у середньому біля 1500 аварійних повідомлень на день, тобто більше одного повідомлення у хвилину. Це призводить до підвищеної втоми операторів, яким доводиться постійно приймати миттєві рішення про те, як реагує платформа на той чи інший сигнал. Але аналітична платформа може відфільтровувати другорядну інформацію, і тоді оператори отримують можливість зосередитись у першу чергу а критичних ситуаціях. Це дозволяє їм більш ефективно виявити чи попередити аварії і, можливо, нещасні випадки. В результаті підвищуються рівні надійності виробництва, промислової безпеки, готовності технологічного обладнання, відповідності нормативним вимогам.

Крім того, за результатами аналізу великих даних можна розрахувати строки окупності обладнання, перспективи змін технологічних режимів, скорочення обслуговуючого персоналу – тобто приймати стратегічні рішення стосовно подальшого розвитку підприємства.