Тема 16. Технології обробки великих даних (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.

Аналітики компанії IBS «весь світовий обсяг даних» оцінили такими величинами:

- 2003 г. – 5 ексабайт даних (1 ЕБ = 1 млрд гігабайтів)
- 2008 г. – 0,18 зетабайта (1 ЗБ = 1024 ексабайти)
- 2015 г. – більше 6,5 зетабайт
- 2021 г. – 44–48 зетабайта (прогноз)
- 2025 г. – цей об'єм збільшується ще у 10 разів.

Великі дані – це сукупність технологій, покликаних здійснювати три операції:

- Обробляти більші, у порівнянні зі «стандартними» сценаріями, об'єми даних.
- 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.
- 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 дикредитованим і пропонують від нього відмовитись.

Більш того, у жовтні 2015 року компанія Gartner виключила Big Data з числа популярных 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.

Три принципи роботи з великими даними

Визначальними характеристиками для великих даних є, окрім їх фізичного об'єму, й інші, які підкреслюють складність задачі обробки і аналізу цих даних. 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.

У подальшому з'явилась інтерпретація з чотирьома V (додалась veracity – достовірність), п'ятю V (viability – життєздатність і value – цінність), семи V (variability – змінність та visualization – візуалізація). Але компанія IDC, наприклад, інтерпретує саме четверте V як value (цінність), пікреслюючи економічну доцільність обробки великих об'ємів даних у відповідних умовах [7].

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

Горизонтальна масштабованість. Це — базовий принцип обробки великих данных. 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. Власне, для роботи з великими даними розробляються підходи і технології.

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.

SQL - мова структурованих запитів, що дозволяє працювати з базами даних. За допомогою SQL можна створювати і модифікувати дані, а управлінням масиву даних займається відповідна система управління базами даних.

NoSQL (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.

Основні риси:

– базова доступність - запити гарантовано завершується (успішно чи безуспішно);

– гнучкий стан - стан системи може змінюватися з часом, навіть без введення нових даних, для досягнення узгодження даних;

– узгодженість в кінцевому рахунку - дані можуть бути деякий час неузгодженими, але приходять до узгодження через деякий час.

Hadoop — проект фонду Apache Software Foundation, набір утилітів, 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.

Використовується для реалізації пошукових і контекстних механізмів багатьох високонавантажених веб-сайтів, у тому числі, для Yahoo! та Facebook. Розроблено на Java в рамках обчислювальної парадигми MapReduce, згідно з якою додаток розділяється на велику кількість однакових елементарних завдань, здійсненних на вузлах кластера і природним чином приводяться в кінцевий результат.

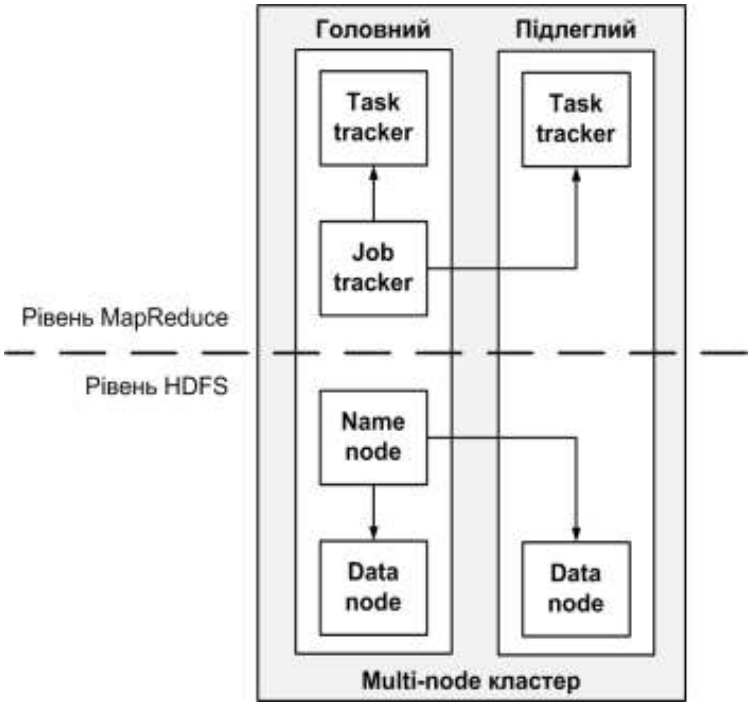

Рисунок 16.1 – Обчислювальна парадигма MapReduce

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.

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

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

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

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

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

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.

Великі дані у промисловості

Згідно звіту компанії McKinsey «Global Institute, Big data: The next frontier for innovation, competition, and productivity», дані стали таким само важливим

фактором виробництва, як трудові ресурси чи виробничі активи. За рахунок використання великих даних, компанії можуть отримувати відчутні конкурентні переваги. 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 і т.д..

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.

Проте зміна класу досліджень – від оперативного до аналітичного, поява нових типів даних, необхідність швидкого доступу до них, зумовила збільшення інтересу до проблеми інтеграції та опрацювання даних з метою підвищення якості управлінських рішень. Найвищий пік активності дослідженьу сфері інтеграції припадає на 90-ті рр. XX ст. та на наш час у зв'язку збурхливим розвитком методів Business Intelligence та збільшенням можливостей сховищ даних (збільшення обсягів збережених даних, наявністьпроцедур аналітичного опрацювання даних – OLAP).

Особливістю сучасних досліджень є аналіз не лише типів даних (описів), але й семантики. Особливо активний розвиток засобів для оперативного збору різнотипних даних, завантаження їх у сховище даних, аналізу та прогнозування спостерігається в сферах енергетики та адміністративного керування, нафтогазовому секторі.

Проблема швидкого отримання різнотипових даних (сенсорних числових, текстових документів, графіків тощо) з метою формування на їх основі оперативних рішень постала ще у роки 2-ї світової війни і активно розвивалась для застосування в атомних проектах, управлінні ракетами, навігації, управлінні бойовими діями.

Опрацювання та аналіз таких різнотипових даних використовується для моделювання розвитку подій та ситуацій, а також в системах підтримки прийняття рішень. Започаткували вивчення цієї проблеми фон Нейман, розробки компанії IBM, науковці школи Лєбєдєва С.О. (спеціалізована ЕОМ), Глушкова В.М. (системний аналіз, теорія конфліктних ігор, проблемно орієнтовані системи моделювання та опрацювання даних) що призвело до розвитку мов блокового програмування, систем підтримки прийняття рішень.

Схема отримання інформації органами керування регіоном передбачає створення статистичних звітів іншими об’єктами галузі за наперед визначеноюформою з покроковим агрегуванням інформації від одного об’єкту до іншого

Це приводить до того, що особа, яка отримує інформацію, бачить її лише в агреговану вигляді за жорстко визначеними критеріями групування, а деталізована інформація потрапляє зі значним запізненням. Тому рішення, які можуть бути прийняті у такому випадку, недостатньо враховують усі особливості перебігу процесів розвитку регіону. Процес консолідації даних (на рисунку) для аналізу та прогнозування розвитку регіону генерує наступні задачі:

– підвищення оперативності отримання, аналізу та використання інформації, необхідної для підтримання прийняття рішень щодо керування регіоном.
– підвищення якості та дієвості керуючих рішень завдяки оперуванню достовірною інформацією, отриманою безпосередньо з відповідного джерела;
– визначення нових аспектів діяльності регіону завдяки аналізу даних, які не потрапляли у традиційні звіти, і тому не враховувалися при прийнятті рішень;
– своєчасного виявлення негативних тенденцій розвитку з метою їх подальшого усунення.

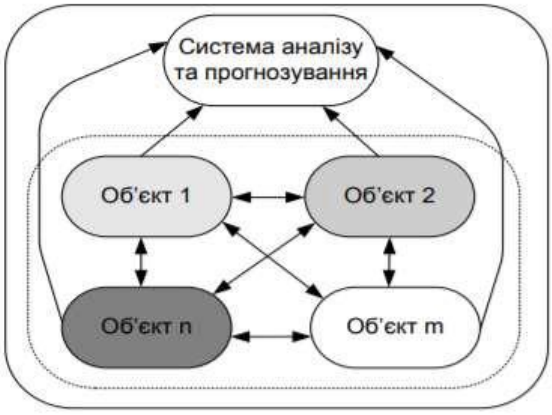

Рисунок 16.2 – Схема рівноправного обміну даними

Інформаційний бум призвів до збільшення кількості даних, накопичених у багатьох предметних галузях у сотні та тисячі разів. До таких областей відноситься і сфера державного управління. Кількість зібраної інформації зростає експоненційно. Так, за дослідженням IDC Digital Universe Study, проведеним на замовлення компанії EMC, сумарний обсяг світових даних у2005 році складав 130 ексабайт, до 2011 року він зріс до 1227 EB, а за останній рік знову подвоївся, досягши 3 ZB (зетабайт). Прогноз, здійснений тим же дослідженням, показує, що до 2021 року обсяг цифрових даних зросте до 7.9ZB. Розмір окремих баз даних зростає так само швидко і подолав петабайтний бар'єр. Більшість зібраних даних на даний час не аналізується, або ж проходить лише поверхневий аналіз.

Видобування даних є процесом виявлення нових нетривіальних закономірностей з великих масивів інформації. Необхідно зазначити, що прикметник «великий» у застосуванні до даних має тенденцію до постійного зростання значення. Наприклад, за даними Тіма Суонсона кількість операцій, що здійснюється щодня в криптовалюті Bitcoin, перевищила 100 000 операцій(оригінал у «Щоденний обсяг транзакцій в Bitcoin подолав 100-тисячнийбар'єр».

Основними проблемами, які виникають при обробці даних, є відсутність методів аналізу, придатних до застосування через їх різнотипність (для регіону – це і числові дані, і геодані, слабоструктуровані звіти тощо), потреба у значних людських ресурсах для підтримки процесу аналізу даних, висока обчислювальна складність наявних алгоритмів аналізу та стрімке зростання обсягу зібраних даних. Це в свою чергу призводить до постійного зростання часу, що витрачається на аналіз даних навіть при регулярному оновленні комп'ютерних засобів, а також – необхідність роботи із розподіленими базами даних, можливості яких більшість існуючих методів аналізу даних не використовують ефективно.

Визначення Великих даних

Концепція Великих даних не нова, вона виникла за часів мейнфреймів і пов'язаних з ними наукових обчислень. Як добре відомо, наукомісткість обчислень завжди було складним завданням. Як правило, вона нерозривно пов'язана з обробкою великих обсягів інформації.

Проте, безпосередньо термін «Великі дані» (Big Data) з'явився порівняно недавно. Він є одним з небагатьох, що має відомий день народження – 3вересня 2008 р. Тоді було випущено спеціальний випуск найстарішого британського наукового журналу Nature. Журнал присвячений пошукам відповіді на питання: «Як технології можуть вплинути на наукове майбутнє, що відкриває можливості для роботи з Великими даними». Згідно зі звітом McKinsey інституту під назвою «Великі дані: Наступний рубіж для інновацій, конкуренції і продуктивності», термін «Великі дані» відноситься до наборів даних, розмір яких перевищує ємність звичайної бази даних (БД) для видобування, зберігання, управління і аналізу інформації.

EWeek подає визначення, запропоноване дослідницькою компанією Gartner: «Великі дані характеризуються обсягом, різноманітністю і швидкою плинністю структурованих і неструктурованих даних в процесорах і пристроях зберігання даних, а також перетворення даних для задач бізнес-консалтингу для підприємств».

Великі дані (Big Data) в інформаційних технологіях (за визначенням К. Лінча) – набір методів та засобів опрацювання структурованих і неструктурованих різнотипних динамічних даних великих обсягів з метою їх аналізу та використання для підтримки прийняття рішень.

Є альтернативою традиційним системам управління базами даних і рішенням класу Business Intelligence. До цього класу відносять засоби паралельного опрацювання даних (NoSQL, алгоритми MapReduce, Hadoop).

На думку компанії DCA (Data-Centric Alliance) під Big Data розуміють не якийсь конкретний об'єм даних і навіть не дані, а методи їх обробки, які дозволяють розподілено обробляти інформацію. Ці методи можна застосовувати

як до великих масивів даних (таких як дані всіх сторінок в мережі Інтернет), так і до малих масивів (інформація про денні поступлення товару в магазин).

Визначальними характеристиками для Великих даних є обсяг (volume, в сенсі величини фізичного обсягу), швидкість (velocity в сеансах як швидкості приросту, так і необхідності високошвидкісної обробки та отримання результатів), різноманіття (variety, в сенсі можливості одночасної обробки різних типів структурованих і слабоструктурованих даних).

Хмарні технології підтримують інфраструктуру віртуалізації та її профілювання для конкретних структур даних або для підтримки конкретних наукових робочих процесів.

Розмаїття (Variety) визначається за допомогою:

- реляційних даних (таблиці / транзакції);
- текстових даних (Web), напівструктурованих даних (XML);
- даних на основі графових моделей (соціальна мережа, Semantic Web, RDF);
- потокових даних;
- великих публічних даних (онлайн, погода, фінанси і т.д.).

Є такі види вартості (Value) у Великих даних як статистичні дані, події, метадані тощо. Швидкість (Velocity) (Speed) Великих даних подана як:

- дані генеруються швидко і повинні бути опрацьовані швидко,
- он-лайн аналіз даних,
- підтримка прийняття рішень з неповними даними.

Достовірність (Veracity) – поняття, зворотне до невизначеності, яка виникає через невідповідність даних, їх неповноту, латентність [18]. Аналіз даних в системах територіального управління зводиться до вирішення трьох конкретних завдань:

- соціально-економічна оцінка стану природного середовища в регіоні вданий час і перспективі, розроблення на її основі системи заходів по повному запобіганню чи максимальному пом’якшенню негативного впливу господарської діяльності на навколишнє середовище;

– визначення й врахування можливих наслідків змін у природному середовищі в результаті господарської діяльності і техногенних процесів, їх вплив на спеціалізацію і комплексний розвиток господарства регіону;
– врахування прогнозів еколого-економічних процесів у контексті загального комплексного прогнозу соціально-економічного розвитку регіону шляхом формування ряду критеріїв і обмежень як по ресурсах, так і за допомогою показників якісного стану навколишнього середовища.

Незважаючи на те, що термін був введений в академічному середовищі, первинною була проблема зростання кількості і різноманітності наукових даних. Станом на 2009 рік термін став широко поширений у діловій пресі, а до 2010 року з'явилася перша низка інформаційно-технологічних продуктів і рішень, що стосуються виключно проблем обробки великих обсягів даних. З 2011 року більшість найбільших постачальників інформаційних технологій для організацій в їх бізнес-стратегії використовують концепцію Великих даних, у тому числі IBM, Oracle, Microsoft, Hewlett-Packard, EMC.

Завдання, що виникають під час опрацювання, обробки, інтерпретації, збору та організації Великих даних, з'явилися в численних секторах, включаючи бізнес, промисловість, некомерційні організації. Обсяги даних, такі як операції замовника у роздрібній торгівлі, моніторинг погоди, бізнес-аналіз, можуть швидко випереджувати потужність традиційних методів та інструментів аналізу даних. З'явилися нові методи та інструменти, включаючи бази даних NoSQL, MapReduce, обробка природної мови, машинне навчання, візуалізація, придбання, і серіалізація. Усе це стає необхідним, щоб повною мірою усвідомити, що відбувається, коли зростають Великі дані, як вони застосовуються і де починають відігравати вирішальну роль. Також необхідно знати вимоги до існуючих методів розроблення систем і аналізу даних.

Великі дані є терміном, який використовується для ідентифікації наборів даних, з якими ми не можемо впоратися з використанням існуючих методологій

та програмних засобів через їх великий розмір і складність. Багато дослідників намагаються розробити методики і програмні засоби для передачі даних або видобування інформаційних гранул з Великих даних.

Особливості Великих даних, а саме:

- робота з неструктурованою та структурованою інформацією,
- орієнтація на швидке опрацювання даних, призводять до того, що традиційні мови запитів виявляються малоефективними для роботи з даними.

Концепція «Великі дані» досі не дуже добре окреслена, хоча активно використовується для бізнесу та технологій. Аналіз згаданих вище джерел, науково-популярних журналів, і блогів дають змогу виділити наступні фокуси обговорення:

- джерела Великих даних,
- апаратне забезпечення та інфраструктура,
- програмне забезпечення і зберігання,
- інформаційні технології (методи і засоби обробки даних).
- використання Великих даних, бізнес-аналіз.

В якості джерел Великих даних можна виділити пристрої та людей. Приклади перших джерел: національні та міжнародні проекти, такі як Великий адронний колайдер (LHC) в ЦЕРН, Лабораторія фізики елементарних частинок в Європі, Великий синоптичний оглядовий телескоп на півночі Чилі; промисловість (SCADA, фінанси і т.д.).

Приклади другого типу джерел на рисунку: соціальні мережі, охорона здоров’я, роздрібна торгівля, особисті дані про місцезнаходження, управління громадським сектором і т.д. Для збору і обробки Великих даних доцільно використовувати технології хмарних обчислень. Хмарні обчислення – це нова парадигма для розміщення кластерів даних і надання різних послуг локальною мережею або через Інтернет.

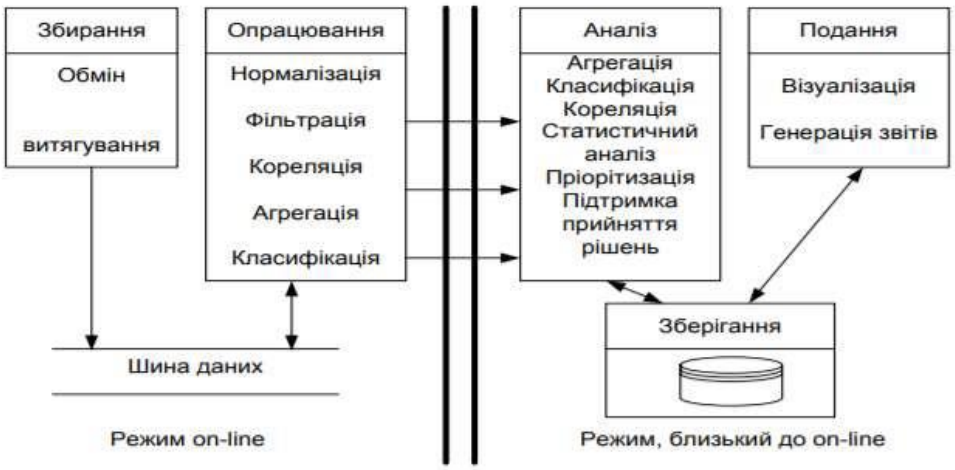

Рисунок 16.3 – Порівняльна характеристика OLAP та BigData

Хостинг кластерів даних дає змогу клієнтам зберігати та обчислювати величезну кількість даних у хмарі. Оскільки розмір окремих баз даних зростає швидко і подолав петабайтний бар'єр (наприклад, бази даних соціальних мереж),то онлайн опрацювання в режимі OLAP таких обсягів практично неможливе.

В таблиці подано відомості щодо ряду інструментів опрацювання Великих даних з відкритим вихідним кодом, які надаються через інфраструктури хмарних обчислень. Більшість інструментів забезпечується Apache і випущені під ліцензією Apache. Усі ці продукти згруповано за типами задач, що виникають в процесах опрацювання Великих даних.

Обробка і методи аналізу Big Data

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

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

Розподілене зберігання вирішує проблему великого обсягу даних, дозволяючи організовувати сховище з довільного числа окремих простих носіїв. Зберігання може бути організовано з різним ступенем надмірності, забезпечуючи стійкість до збоїв окремих носіїв.

Розподілена обробка з урахуванням локальності даних означає, що програма обробки доставляється на обчислювач, що знаходиться якомога ближче до оброблюваних даних. Це принципово відрізняється від традиційного підходу, коли обчислювальні потужності і підсистема зберігання розділені і дані повинні бути доставлені на обчислювач. Таким чином, технології Big Data спираються на обчислювальні кластери з безлічі обчислювачів, забезпечених локальної підсистемою зберігання.

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

- відмовостійка розподілена файлова система Hadoop Distributed File System (HDFS), за допомогою якої здійснюється зберігання;
- програмний інтерфейс Map Reduce, який є основою для створення програмного забезпечення, що обробляють великі обсяги структурованих і неструктурованих даних паралельно на кластері, що складається з тисяч машин;
- Apache Hadoop YARN, що виконує функцію управління даними.

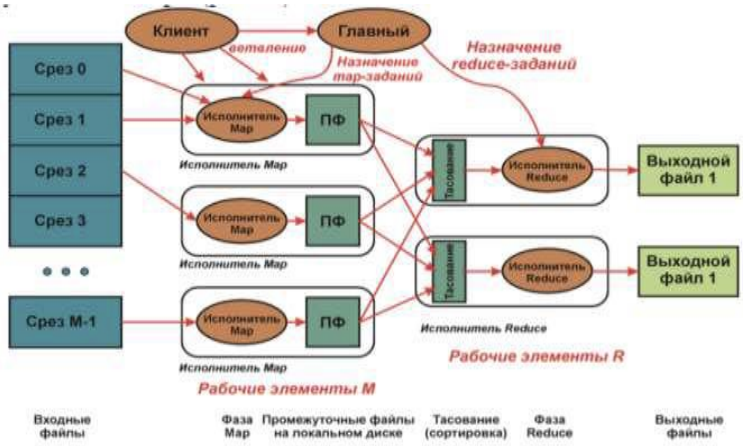

Рисунок 16.4 – Концептуальна модель MapReduce

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

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

В основі моделі роботи Apache Hadoop лежать три основних принципи.

По-перше, дані рівномірно розподіляються на внутрішніх дисках безлічі серверів, об'єднаних HDFS.

По-друге, не дані передаються програмі обробки, а програма - до даних.

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

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

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

1) Збір, обробка та структурування даних.

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

2) Агрегація і аналіз даних.

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

3) Аналітика даних в реальному часі.

Для максимально оперативного аналізу отриманих даних використовується Oracle Exalytics Database Machine, яка дозволяє вирішувати аналітичні завдання фактично в режимі «online». Існує безліч різноманітних методів аналізу масивів даних, в основі яких лежить приблизно однаковий набір інструментів аналізу даних: багатовимірний аналіз (OLAP), регресія, класифікація, кластеризація і пошук закономірностей. Деякі з перерахованих методик зовсім не обов'язково можуть бути застосовані виключно до великих даними і можуть з успіхом використовуватися для менших за обсягом масивів (наприклад, A / B- тестування, регресійний аналіз).

Багатовимірний аналіз - суть методу полягає в побудові багатовимірного куба і отриманні його різних зрізів. Результатом аналізу, як правило, є таблиця, в осередках якої містяться агреговані показники (кількість, середнє, мінімальне або максимальне значення і так далі). Залежно від реалізації, існують три типи системи багатовимірного аналізу (OLAP): багатовимірна OLAP (Multidimensional OLAP - MOLAP); реляційна OLAP (Relational OLAP - ROLAP); гібридна OLAP (Hybrid OLAP - HOLAP). Серед них ROLAP-системи є найбільш прозорими і вивченими, оскільки ґрунтуються на широко поширених реляційних СУБД, в той час як внутрішній устрій MOLAP і HOLAP зазвичай більш закрита і відноситься до області «ноу-хау» конкретних комерційних продуктів.

MOLAP представляє інформацію у вигляді «чесної» багатовимірної моделі, але всередині використовуються ті ж підходи, що і в ROLAP: схеми «зірка» та «сніжинка». З точки зору СУБД база даних ROLAP - це звичайна реляційна база, і для неї необхідно підтримувати весь перелік операцій. Однак це не дозволяє, по-перше, жорстко контролювати етапи введення даних. По-друге, збирати статистику і підбирати оптимальні структури для зберігання індексів. По-третє, оптимізувати розміщення даних на диску для забезпечення високої швидкості введення / виводу. По-четверте, при виконанні аналітичних запитів через високі вимоги до швидкодії немає можливості провести глибокий статистичний аналіз і виробити оптимальний план виконання. У ROLAP

використовуються «рідні» реляційні оптимізатори запиту, які ніяк не враховують «багатовимірність» бази даних. Технології MOLAP позбавлені перелічених недоліків і завдяки цьому дозволяють домогтися більшої швидкості аналізу.

Вибір технології MOLAP / ROLAP / HOLAP при аналізі Big Data залежить від частоти оновлення бази даних. З точки зору розпаралелювання обробки, на перший погляд, все просто - будь-який багатовимірний куб може бути «розрізаний MOLAP представляє інформацію у вигляді «чесної» багатовимірної моделі, але всередині використовуються ті ж підходи, що і в ROLAP: схеми «зірка» та «сніжинка». З точки зору СУБД база даних ROLAP - це звичайна реляційна база, і для неї необхідно підтримувати весь перелік операцій. Однак це не дозволяє, по-перше, жорстко контролювати етапи введення даних. По-друге, збирати статистику і підбирати оптимальні структури для зберігання індексів. По-третє, оптимізувати розміщення даних на диску для забезпечення високої швидкості введення / виводу.

По-четверте, при виконанні аналітичних запитів через високі вимоги до швидкодії немає можливості провести глибокий статистичний аналіз і виробити оптимальний план виконання. У ROLAP використовуються «рідні» реляційні оптимізатори запиту, які ніяк не враховують «багатовимірність» бази даних. Технології MOLAP позбавлені перелічених недоліків і завдяки цьому дозволяють домогтися більшої швидкості аналізу.

Наприклад, якщо користувач запитує статистику продажів по країні за вказаний проміжок часу, а дані розподілені за кількома регіональним ОЛР-серверів, то кожен сервер повертає свою власну відповідь, які потім збираються воєдино. Якщо ж дані будуть розподілені по тимчасовому критерію, то при виконанні даного прикладу запиту все навантаження ляже на один сервер.

Проблема в тому, що, по-перше, дуже важко заздалегідь визначити оптимальний розподіл даних по серверах, а по-друге, для частини аналітичних запитів може бути заздалегідь невідомо, які дані і з яких серверів знадобляться. Стосовно до Великих Даним це означає, що існуючі підходи для багатовимірного аналізу можуть добре масштабуватися і що вони допускають розподілений збір

інформації - кожен сервер може самостійно збирати інформацію, здійснювати її очищення і завантаження в локальну базу.

Регресія - під регресією розуміють побудову параметричної функції, яка описує зміну зазначеної числової величини в зазначений проміжок часу. Ця функція будується на основі відомих даних, а потім використовується для передбачення подальших значень цієї ж величини. На вхід методу надходить послідовність пар виду «час - значення», що описує поведінку цієї величини при заданих умовах, наприклад, кількість продажів конкретного виду товару в конкретному регіоні.

На виході - параметри функції, яка описує поведінку досліджуваної величини. Незалежно від виду використовуваної параметричної функції підбір значень її параметрів здійснюється одним і тим же способом. Обчислюється сумарна різниця між що спостерігаються значеннями і значеннями, які дає функція при поточних значеннях її параметрів. Потім визначається, як слід підкоригувати значення параметрів для того, щоб зменшити поточну сумарну різницю. Ці операції повторюються до тих пір, поки сумарна різниця не досягне необхідного мінімуму або її подальше зменшення стане неможливим. З точки зору обробки даних при регресійному аналізі ключовими операціями є обчислення поточної сумарною різниці і коригування значень параметрів. Якщо перша операція розпаралелюється очевидним чином (сума обчислюється по частинах на окремих серверах, а потім підсумовується на центральному сервері), то з другої складніше.

У найбільш загальному випадку при коригуванні ваг використовують загальновідомий математичний факт: функція декількох параметрів зростає в напрямку градієнта і убуває в напрямку, протилежному градієнту. У свою чергу, обчислення градієнта полягає в обчисленні приватних похідних функції по кожному з параметрів, що зводиться до дискретного диференціювання, заснованому на обчисленні зважених сум. В результаті коригування значень параметрів також зводиться до підсумовування, яке може бути розпаралеліть.

Якщо регресійний аналіз зводиться до обчислення зважених сум, то він має приблизно тим же ступенем застосовності і при роботі з Big Data, що і багатовимірний аналіз. Таким чином, системи регресійного аналізу хвилі можуть масштабуватися і працювати в умовах розподіленого збору інформації. Класифікація - її завдання частково схожа на завдання регресії і полягає в спробі побудови і використання залежності однієї змінної від декількох інших. Наприклад, маючи базу даних про ціну об'єктів нерухомості, можна побудувати систему правил, що дозволяє на основі параметрів нового об'єкта передбачити його приблизну ціну.

Відмінність класифікації від регресії полягає в тому, що аналізується не тимчасовою ряд - подаються на вхід значення ніяк не можуть бути впорядковані. На поточний момент -розроблено безліч методів класифікації (функції Байеса, нейронні мережі, машини підтримують векторів, дерева рішень і т. Д.), Кожен з яких має під собою добре опрацьовану наукову теорію. Разом з тим всі методи класифікації будуються по одній і тій же схемі. Спочатку проводиться навчання алгоритму на порівняно невеликій вибірці, а потім - застосування отриманих правил до іншої вибірці. На першому етапі можливо копіювання масиву даних на один сервер для запуску «класичного» алгоритму навчання без розпаралелювання роботи. Однак на другому етапі дані можуть оброблятися незалежно - система правил, отримана за підсумками самонавчання, копіюється на кожен сервер, і через неї проганяється весь масив даних, що зберігається на цьому сервері. Отримані результати можуть або зберігатися там же на сервері, або відправлятися для подальшої обробки.

Таким чином, на етапі навчання класифікаторів про роботу з Big Data поки мова не йде - не існує вибірок такого обсягу, підготовлених для навчання систем, а на етапі класифікації окремі порції даних обробляються незалежно один від одного.

Кластеризація - її завдання полягає в розбитті безлічі інформаційних сутностей на групи, при цьому члени однієї групи більш схожі один на одного, ніж члени з різних (класифікація відносить кожен об'єкт до однієї з заздалегідь

визначених груп). В якості критерію схожості використовується функція-відстань, на вхід якої надходять дві сутності, а на вихід - ступінь їх схожості. Відомо безліч різних способів кластеризації (графові, ієрархічні, ітеративні, мережі Кохонена).

Проблема кластеризації Big Data полягає в тому, що наявні алгоритми припускають можливість безпосереднього звернення до будь-якої інформаційної суті у вихідних даних (заздалегідь неможливо передбачити, які саме сутності знадобляться алгоритму). У свою чергу, вихідні дані можуть бути розподілені по різних серверах, і при цьому не гарантується, що кожен кластер зберігається строго на одному сервері. Якщо розподіл даних по серверах робити прозорим для алгоритму кластеризації, то це неминуче призведе до копіювання великих обсягів з одного сервера на інший.

Рішення проблеми може бути наступним. На кожному сервері запускається свій алгоритм, який оперує тільки даними цього сервера, а на виході дає параметри знайдених кластерів і їх ваги, які оцінюються виходячи з кількості елементів всередині кластера. Потім отримана інформація збирається на центральному сервері і проводиться метакластерізація - виділення груп близько розташованих кластерів з урахуванням їх ваг.

Цей метод універсальний, добре розпаралелюється і може використовувати будь-які інші алгоритми кластеризації, однак він вимагає проведення серйозних наукових досліджень, тестування на реальних даних і порівняння отриманих результатів з іншими «локальними» методами. Таким чином, для аналізу Big Data переважна частина методів кластеризації непридатна в чистому вигляді і необхідні додаткові дослідження.

Пошук закономірностей - суть методу полягає в знаходженні правил, що описують взаємозалежності між внутрішніми елементами даних. Класичним прикладом є аналіз покупок в супермаркеті і виявлення правил виду «якщо людина купує фотоапарат, то зазвичай він купує ще до нього акумулятор і карту пам'яті». На вхід завдання пошуку закономірностей надходить неврегульована безліч сутностей, для кожної з яких відомий набір присутніх інформаційних

ознак; наприклад, такими сутностями можуть бути чеки на покупки, а ознаками - куплені товари.

Завдання пошуку закономірностей зводиться до виявлення правил виду «якщо присутні ознаки Ai, A2, ..., An, то присутні і ознаки Б1; B2, ..., Bm, при цьому кожне правило характеризується двома параметрами: ймовірністю спрацьовування і підтримкою. Перший параметр показує, як часто виконується дане правило, а другий - як часто можна застосувати дане правило, тобто як часто зустрічається поєднання ознак A1, A2, ..., An.

A / B тестування - методика, в якій контрольна вибірка по черзі порівнюється з іншими. Таким чином вдасться виявити оптимальну комбінацію показників для досягнення, наприклад, найкращою відповідної реакції споживачів на маркетингову пропозицію. Великі дані дозволяють провести величезну кількість ітерацій і таким чином отримати статистично достовірний результат.

Краудсорсінг - методика збору даних з великої кількості джерел: категоризація і збагачення даних силами широкого, невизначеного кола осіб.

Змішання і інтеграція даних - набір технік, що дозволяють інтегрувати різнорідні дані з різноманітних джерел для можливості глибинного аналізу.

Машинне навчання («штучний інтелект») - має на меті створення алгоритмів самонавчання на базі статистичного аналізу даних або машинного навчання для отримання комплексних прогнозів.

Генетичні алгоритми - в цій методиці можливі рішення представляють у вигляді «хромосом», які можуть комбінуватися і мутувати. Як і в процесі природної еволюції, виживає найбільш пристосована особина. Оптимізація - набір чисельних методів для ре- дизайну складних систем і процесів для поліпшення одного або декількох показників. Допомагає в прийнятті стратегічних рішень, наприклад, складу виведеної на ринок продуктової лінійки, проведенні інвестиційного аналізу. Візуалізація аналітичних даних - методи для подання інформації у вигляді малюнків, графіків, схем і діаграм з використанням

інтерактивних можливостей та анімації як для результатів, так і для використання в якості вихідних даних.

Хмарна платформа Oracle для Big Data

Корпорація Oracle анонсувала хмарне рішення Oracle Cloud Platform for Big Data, компонент портфоліо PaaS-сервісів.

Рисунок 16.5 – Хмарна платформа Oracle для Big Data

Ключові вимоги середовища для роботи з big data включають в себе:

- Масштабованість. Через високий і швидкого приросту даних будь-яка система завжди повинна бути готова до розширення.
- Відмовостійкість. Якась частина машин в кластері, які проводять аналіз, гарантовано буде виходити з ладу, наслідки цього не повинні позначатися на процесі обробки інформації.