**2. Обгрунтування вибраного напряму роботи**

Перш ніж почати проектування та розробку системи необхідно чітко визначити які вимоги така система повинна виконувати. Тільки після повного розуміння та означення вимог програмне рішення може бути правильно спроектованим, реалізованим та відлагодженим. Типові сховища даних можуть забезпечувати не тільки зберігання великих обємів даних в придатному для аналітичних запитів вигляді, але й дозволяти обробку нових даних, що надходять, в реальному часі. Такі сховища даних можуть використовуватися для надання статистичної інформації про певну предметну область з врахуванням нових даних, що постійно надходять. В залежності від того, які вимоги ставляться до сховища даних, його структура та перелік використаних технологій значною мірою відрізнятиметься.

**2.1 Дослідження вимог до сховищ даних**

В загальному випадку розрізняють два основних види сховищ даних:

1. Сховища даних що забезпечують виконання аналітичних запитів до великих обємів даних
2. Сховища даних які забезпечують окрім виконання аналітичних запитів ще й обробку в реальному часі нових вхідних даних.

Перший вид сховищ використовується в більшій мірі в наукових та дослідницьких організаціях. Також значного поширення вони набули в організаціях, які оперують великими обємами даних для побудови прогнозів та прийняття рішень на основі цих прогнозів. Характерною рисою таких систем є те, що вимоги до часу відповіді (латентності) не є критичними. В залежності від обсягів даних та використаної інфраструктури час відповіді таких систем може вимірюватися десятками хвилин або годинами.

Характерною особливістю другого виду систем є забезпечення можливості обробки великих обсягів нових вхідних даних з (або без) використанням даних, що вже збережені в сховищі. До таких програмних рішень ставляться досить жорсткі вимоги по часу відклику. Вони найчастіше використовуються в організаціях, які мають справу з даними, що швидко змінюються в часі і потребують максимально швидкого реагування.

Схему руху даних даних в системі першого виду можна зобразити так, як показано на рис 2.1

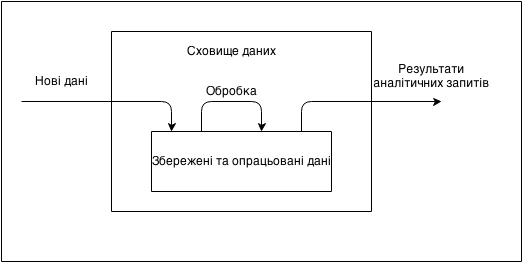


Рис. 2.1 Схема руху даних в сховищах для аналітичних запитів

Як видно зі схеми, в таких системах процес отримання нових даних є незалежним по відношенню до споживання даних зі сховища. Аналітичні запити до сховища даних використовують збережені та опрацьовані в системі дані. Дані можуть надходити в такий вид сховищ з найрізноманітніших джерел: вивантажуватись з систем керування базами даних, отримуватись у вигляді текстових файлів, у спеціальних текстових форматах і т.п. Основною вимогою до таких систем є забезпечити можливість зберігання та подальшого використання завантажених в систему даних. Час на опрацювання вхідних даних та приведення їх до вигляду, придатного для аналітичних запитів може займати суттєвий час. Основною причиною цього можуть бути:

- великі обсяги даних;

- складність трансформацій, що мають бути виконані над вхідними даними;

- необхідність використання всіх, раніше збережених даних для трансформації нових (в такому випадку час опрацювання зростатиме з накопиченням нових записів).

Споживачами даних, збережених у таких сховищах, зазвичай є аналітичні системи, які виконують велику кількість складних запитів до значних частин збереженої інформації. Обсяги інформації, в таких системах можуть сягати кількох десятків або й навіть сотень терабайт. Великою перевагою для проектувальників та розробників таких систем є те, що не ставляться жорсткі вимоги до часу опрацювання та підготовки даних. В першу чергу це дає можливість знехтувати певними оптимізаційними питаннями, які могли б відобразитися у вигляді значних часових затрат на розробку системи. Також це надає можливість зменшити вимоги до апаратної частини системи.

Сховища такого виду є поширеними уже на протязі тривалого часу і зазвичай будуються з використанням БД та великих серверних станцій. Використання таких технологій цілком достатньо і раціонально у випадку, якщо обєми вхідної і оброблюваної інформації не перевищують кількасот гігабайт. З ростом даних такі системи стають надто повільними, а програмних оптимізацій стає недостатньо щоб задовольнити вимоги. Єдиним шляхом є покращення апаратного забезпечення, а це надто дорогий крок.

Системи другого виду забезпечують не тільки можливість виконання аналітичних запитів, але й обробку та видачу вхідної інформації у реальному часі. Такі системи здатні реагувати на події, що виникають, та реагувати на них з надзвичайно низькою латентністю. Сховища такого виду зазвичай мають кілька інтеграційних точок: інтерфейс для виконання аналітичних запитів, та інструмент для передачі результатів обробки вхідної інформації, що безпереревно надходить в систему.

Схема руху даних, які потребують обробки в реальному часі зображена на рис. 2.2.



Рис. 2.2. Схема руху даних в сховищах з підтримкою відповіді в реальному часі

Цей вид сховищ даних є розширеним варіантом першого виду. Такі системи забезпечують можливість не тільки виконання аналітичних запитів, але й надають інтерфейс для виконання обробки вхідних даних в реальному часі.

Інформація в сховище надходить кількома різними шляхами. Це можуть бути великі завантаження файлів з файлових систем або таблиць з баз даних. Ці дані зберігаються та опрацьовуються для подальшого використання інструментами побудови аналітичних звітів. Окрім цього існує також канал, по якому в систему надходить інформація, яка потребує дуже швидкої обробки і передачі вже опрацьованих даних користувачам системи. Такі дані можуть оброблятися за наборами складних правил а також потребувати доступ до уже наявних даних. Вимоги до латентності такої системи є критичними. На відміну від першого типу сховищ даних, в яких дозволений час для опрацювання і перетворення даних може становити десятки хвилин або й години, в системах другого типу дані повинні опрацьовуватися за лічені секунди. В певних випадках час відклику має становити кількасот мілісекунд. Складністю побудови таких систем є в першу чергу те, що при значному збільшені вимог до часу відклику, вимоги до обсягів даних, які мають опрацьовуватися не зменшується. Це означає, що має бути досягнута надзвичайно велика паралелізація обчислень. Потік вхідних даних, які потребують обробки в реальному часі є безперервним, тому не допускаються жодні затримки або збої в системі. Швидкість такого потоку може досягати кількох Мб/с.

Сховища такого виду можуть використовуватися в організаціях, що надають оперативну інформцію в сферах, де потрібно опрацьовувати значні обєми даних в реальному часі. Наприклад метеостанції, медичні заклади, дослідницькі центри, які працюють з вимірюваннями швидкоплинних процесів.

Перелік вимог до систем значною мірою вплине на структуру програмного рішення. В магістерській кваліфікаційній роботі буде досліджено та розроблено шляхи побудови як першого так і другого виду сховищ даних. Для першого типу також буде запропонована реалізація, на основі платформи розподілених обчислень Hadoop.

2.2. Огляд шляхів побудови сховищ даних на основі платформи Hadoop

Вимоги до системи є найбільш визначальним фактором, що впливає на весь процес проектування та розробки програмного рішення. Так і у випадку з сховищами даних, від способу обробки даних залежитиме набір технологій і спосіб їх взаємодії в межах системи. Платформа Hadoop надає можливість зберігання та паралельної обробки великих масивів даних на кластерах зі звичайного недорого обладнання. Такі кластери є набагато дешевшими аніж великі серверні станції, що використовуються для встановлення на них баз даних. Системи на основі Hadoop володіють майже лінійною масштабованістю, а це в свою чергу дозволяє легко і недорого розширювати апаратне забезпечення.

Для правильної побудови сховища даних потрібно враховувати те, як дані мають рухатись в системі. Під параметрами руху маються на увазі:

* Кількість джерел даних
* Інтерфейси, які ці джерела надають;
* Обсяги;
* Швидкість надходження;
* Правила перетворень
* Вимоги до часу зберігання
* Інтерфейси систем-споживачів
* Вимоги до часу запиту

Всі ці параметри визначають спосіб взаємодії між усіма елементами системи та алгоритми їх роботи. Також від значень цих параметрів залежить спектр конкретних інструментів та технологій, які будуть використовуватися для побудови сховища даних.