УДК 004.01

Чіома Е.В., *студентка 3 курсу спеціальності 122 «Комп'ютерні науки»*
Катаєва А.І., *старший викладач кафедри комп'ютерних наук та інформаційних технологій*

СХОВИЩА ДАНИХ: АРХІТЕКТУРА ТА ПРОЕКТУВАННЯ

Донецький національний університет імені Василя Стуса, м. Вінниця

Міжнародний ринок дата-центрів щорічно зростає приблизно на 11% [1]. Основною провідною групою є підприємства, підключені пристрої та користувачі, які забезпечують постійну появу нових даних. З об'ємом ринку ростуть вимоги до надійності зберігання, масштабування та рівнем доступності сховищ даних.

Сховище даних – це предметно-орієнтована система, в якій зібрані дані з різних джерел всередині підприємства та використовуються для підтримки прийняття управлінських рішень.

Існують концептуальні підходи, що висвітлюють класичні принципи проектування традиційних сховищ даних.

Трирівнева архітектура

Нижній рівень – містить сервер бази даних, що використовується для отримання даних з безлічі різних джерел, наприклад, з транзакційних баз даних.

Середній рівень – містить сервер OLAP, який перетворює дані в структуру. Така технологія найкраще підходить для виконання аналізу і застосування складних запитів. Сервер OLAP може працювати в якості розширеної системи керування базами даних або з використанням багатовимірної моделі, яка реалізує багатовимірні дані і операції.

Верхній рівень – це рівень клієнта, який містить інструменти для високорівневого аналізу даних і створення звітів.

Моделі сховищ даних

Існує три загальних моделі сховищ даних: віртуальне сховище, вітрина даних і корпоративне сховище даних.

Віртуальне сховище даних – це набір окремих баз даних, які можна використовувати спільно, щоб користувач міг ефективно отримувати доступ до всіх даних.

Модель вітрини даних використовується для звітності та аналізу конкретних бізнес-ліній. У цій моделі сховища – агреговані дані з ряду вихідних систем, що відносяться до конкретної бізнес-сфери, наприклад: продажі або фінанси.

Модель корпоративного сховища даних передбачає зберігання агрегованих даних, що охоплюють всю організацію. Ця модель розглядає

сховище даних як ядро інформаційної системи підприємства з інтегрованими даними всіх бізнес-одиниць [2].

Проектування сховищ даних

Розрізняють два підходи до проектування традиційних сховищ даних: за Біллом Інмоном та Ральфом Кімболом.

Підхід Ральфа Кімбола ґрунтується на важливості вітрин даних, які є сховищами даних та належать конкретним напрямкам бізнесу. Такий підхід працює за принципом «знизу-вгору», значно полегшує звітність та аналіз.

Підхід Білла Інмона передбачає, що сховище є централізованим для всіх корпоративних даних. За таким принципом організація спочатку створює нормалізовану модель сховища даних, а потім формує вітрини розмірних даних на основі моделі сховища. Такий підхід є спадним [2].

Структурування сховищ даних

Схеми «зірка» та «сніжинка» – це два способи структурувати сховище даних.

Схема типу «зірка» має централізоване сховище даних, яке зберігається в таблиці фактів. Схема розбиває таблицю фактів на ряд денормалізованних таблиць вимірів. Таблиця фактів містить агреговані дані, які будуть використовуватися для складання звітів, а таблиця вимірювань описує збережені дані. Проста конструкція зіркоподібної схеми значно спрощує написання складних запитів.

Схема «сніжинка» відрізняється тим, що використовує нормалізовані дані. Нормалізація означає ефективну організацію даних так, щоб всі залежності даних були визначені і кожна таблиця містила мінімум надмірності. Тому найчастіше такі таблиці вимірювань розгалужуються на окремі таблиці вимірювань. Дана схема використовує менше дискового простору і краще зберігає цілісність даних. Основним недоліком є складність запитів, необхідних для доступу до даних [3].

Новітня архітектура

Сучасні сховища даних переходять в хмару не дотримуючись традиційної архітектури, а кожне з них пропонує свою власну унікальну архітектуру. Найпопулярнішими хмарними сховищами є Amazon Redshift і Google BigQuery.

Amazon Redshift потребує, щоб обчислювальні ресурси були підготовлені та налаштовані у вигляді кластерів, які містять набір з одного або декількох вузлів. Кожен вузол має свій власний процесор та оперативну пам'ять. На кожному вузлі дані зберігаються в блоках, які називаються зрізами (показано на рис. 1). Redshift використовує архітектуру MPP (Massively Parallel Processing), розбиваючи великі набори даних на шматки (слайси) в кожному вузлі. Запити виконуються швидше, тому що обчислювальні вузли обробляють запити в кожному слайсі одночасно.

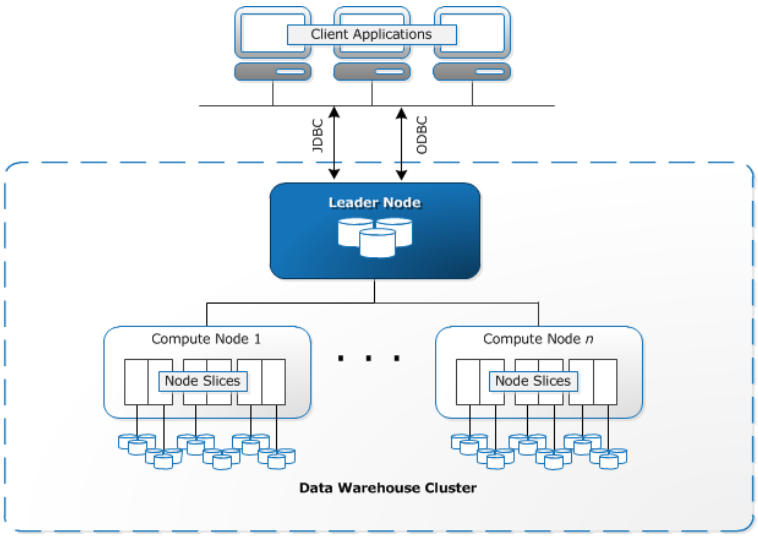

Рисунок 1 - Архітектури сховища даних Amazon Redshift

Архітектура **Google BigQuery** не вимагає сервера, а це означає, що Google динамічно управляє розподілом ресурсів комп'ютера. BigQuery дозволяє клієнтам завантажувати дані з Google Cloud Storage та інших популярних джерел даних [3]. Альтернативним варіантом є потокова передача даних, що дозволяє розробникам додавати дані в сховище даних в режимі реального часу, рядок за рядком, коли вони стають доступними. Для виконання запитів до даних використовуються прості команди SQL, а деревоподібна архітектура відправляє запити тисячам машин за лічені секунди.

Висновок. Хмарні сховища даних – це великий крок вперед в порівнянні з традиційними підходами до архітектури. Однак користувачі та підприємства, як і раніше стикаються з рядом проблем при їх налаштуванні: завантаження даних в хмарні сховища даних – нетривіальне завдання, яке потребує використання сторонніх сервісів; оновлення, вставки і видалення можуть бути складними і повинні виконуватися обережно, щоб не допустити зниження продуктивності запитів тощо. Тому вибір та налаштування сховища даних слід виконувати зважаючи на фактори необхідності придбання фізичного обладнання, подальшого масштабування та швидкість виконання запитів.

Список використаної літератури:

1. *Впровадження хмарних сховищ: огляд та статистика – [Електронний ресурс]. Режим доступу: https://www.businesswire.com/news/home/20170904005104/en/High-Adoption-Cloud-based-Storage-Boost-Growth-Hyperscale*
2. *Пасічник В.В., Шаховська Н.Б. – «Сховища даних: Навчальний посібник» - Львів: "Магнолія 2006", 2008.-492 с.*
3. *Сховища даних – [Електронний ресурс]. Режим доступу: https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/*