ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки

КВАЛІФІКАЦІЙНА РОБОТА

Методи обробки даних в розподілених системах

Виконав: студент групи КСМзм-22-1 Курченко С.І.

Керівник: доц. каф. ЕОМ Козлов Ю.В.

Аналіз предметної області

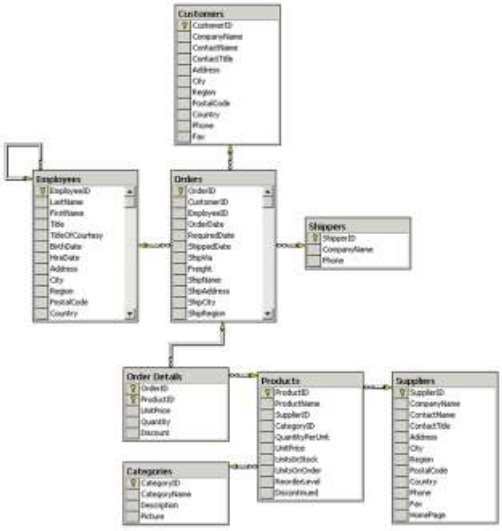
Метою кваліфікаційної роботи є аналіз методів обробки та зберігання даних в розподілених системах.

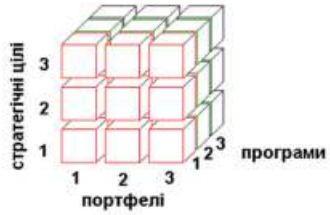
Об'єкт досліження: методи доступу до даних в паралельних розподілених системах.

Завдання:

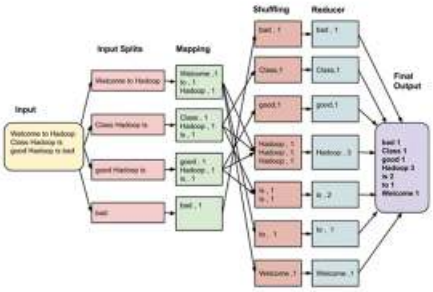
- аналіз і порівняння методів доступу до даних в паралельних розподілених сховищах даних на платформі MapReduce/Spark;
- розробка модифікованого методу виконання запитів до паралельного розподіленого сховища даних на базі технології MapReduce/Spark.

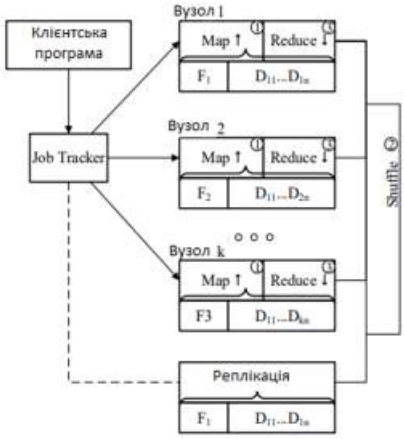
Технологія OLAP. База даних Northwind 3



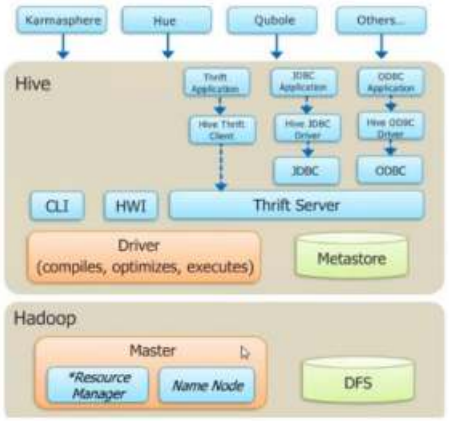


Технологія MapReduce. Процес виконання пар завдань 4

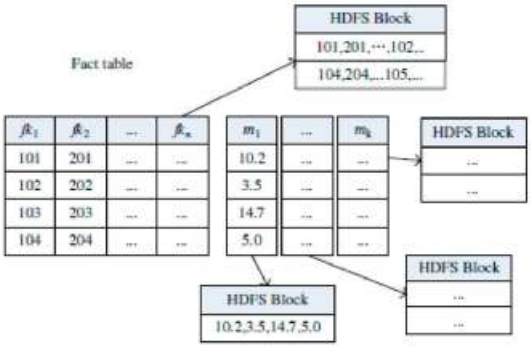




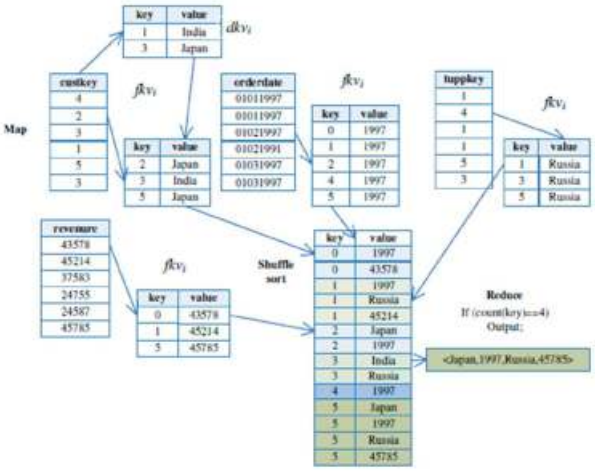
Взаємозв'язок компонентів Hive 5



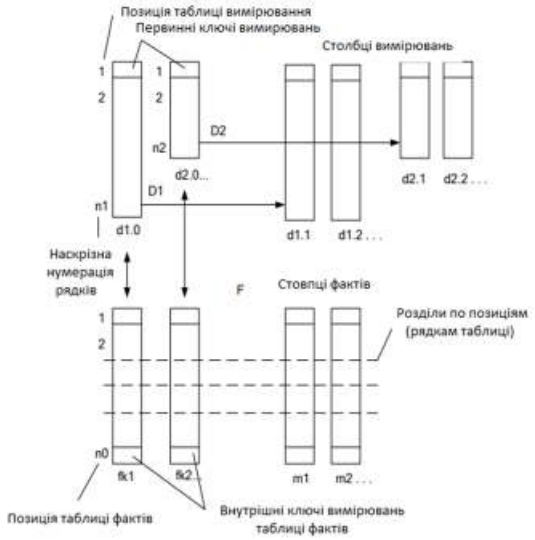
Принцип обробки інформації по методу MFRJ 6



Принцип обробки інформації по методу MFRJ 7



Модифікований метод обробки та зберігання даних без кешування таблиць вимірювань в оперативній пам'яті 8



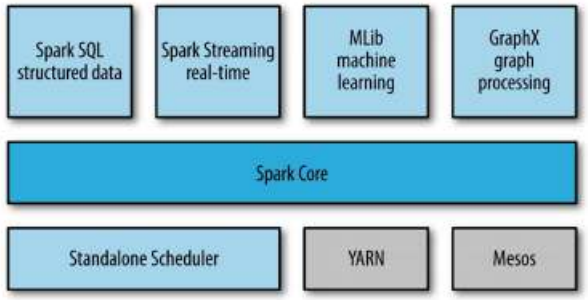
9

Переваги модифікованого методу

- ✓ відсутне дублювання таблиць вимірів в вузлах, і тому немає необхідності передачі їх фрагментів по мережі;
- ✓ немає необхідності хешувати таблиці вимірювань в оперативній пам'яті;
- ✓ таблиці вимірювань і фактів можуть бути рівномірно розподілені по вузлах, не треба міняти політику розподілу блоків, прийняту в MapReduce за замовчуванням;
- ✓ використовується уніфікований метод зберігання таблиць (RCFile), що дозволяє виконувати декомпресію тільки тих стовпців, які використовуються в запиті.

10

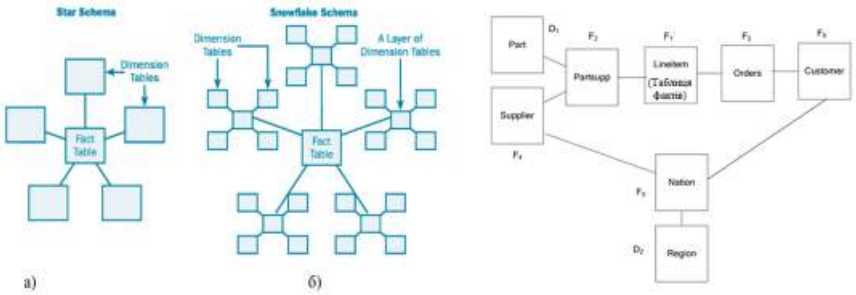
Основні компоненти платформи Spark. Порівняння Hadoop та Spark



1) Запити в Spark виконуються швидше. У Hadoop проміжні дані зберігаються і на локальному диску (перед shuffle), і в файловій системі HDFS.
2) Стійкість до збоїв в Hadoop вище
3) В Spark простіше програмувати складні процеси обробки (відмінні від простих запитів Select). В Spark досить послідовно закодувати необхідні операції map, join, save і реалізувати функції map на мові високого рівня

Схеми сховищ даних

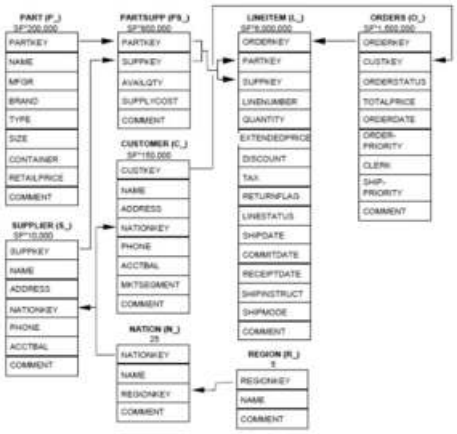
11



а) б)

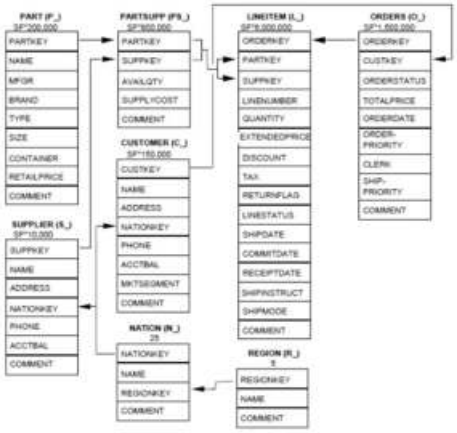
Інфологічна схема сховища даних тесту TPC-H

12



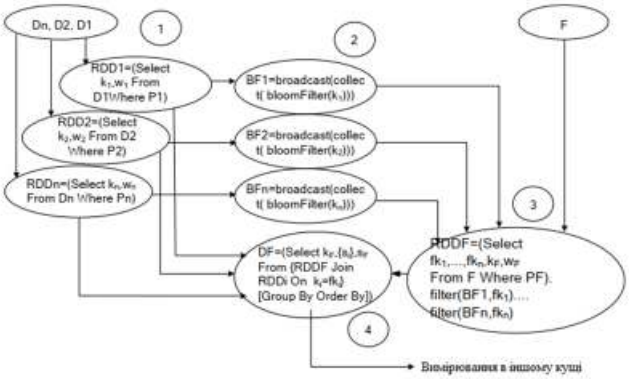
Інфологічна схема сховища даних тесту TPC-H

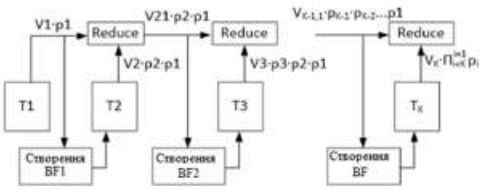
12



Загальна схема реалізації куща вихідного запиту

14



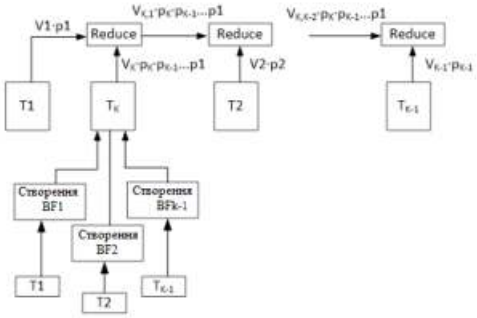


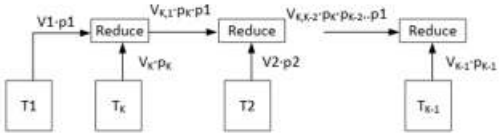
$$\Delta = \sum_{i=1}^{k-1} \left(V_{I+1} P_{I+1} \left(1 - \prod_{J=I}^{1} P_J \right) - V_{BLI} \right)$$

$$\Delta = \sum_{i=1}^{k-1} \left(V_{I+1} - V_{BLI} \right.$$

Схема з'єднання K-1 таблиць вимірів з таблицею фактів з використанням фільтрів Блума

15

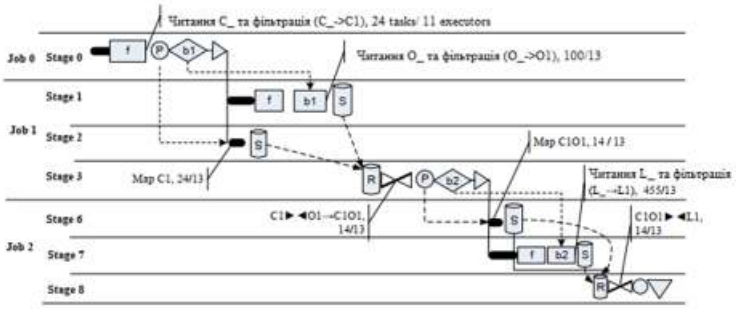




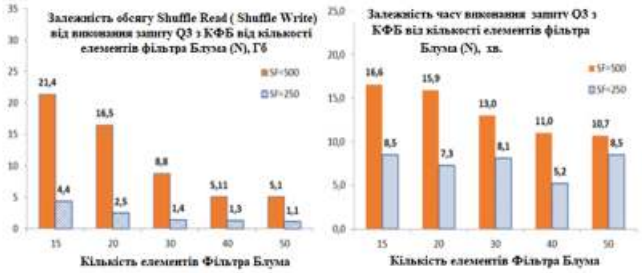
$$\sum_{I=1}^{K-1} V_I P_I + \sum_{I=1}^{K-2} V_{K,I} * \prod_{J=K}^{1} P_J + V_K \prod_{J=K}^{1} P_J + \sum_{I=1}^{K-1} V_{BLI}$$

Процес виконання запиту в середовиищі Spark

16



Результати експериментів 17



Висновки 18

У ході виконання кваліфікаційної роботи проведено порівняльний аналіз методів обробки та зберігання даних в розподілених системах. Також були розглянуті методи доступу до даних в паралельних розподілених сховищах даних на платформі MapReduce/Spark. Розроблен модифікований метод обробки даних та виконання запитів до паралельного розподіленого сховища даних на базі технології MapReduce/Spark з використанням фільтру Блума.

ДОДАТОК Б
ТЕЗИ ДОПОВІДІ

Problems of Informatization: the eleventh international scientific and technical conference

МЕТОДИ ОБРОБКИ ДАНИХ В РОЗПОДІЛЕНИХ СИСТЕМАХ

Курченко С.І., Климова І.М.
Харківський національний університет радіоелектроніки, Харків, Україна

Ні для кого не є секретом, що в сучасному світі обсяги даних стрімко зростають. Розподілені системи стають важливим інструментом для обробки великих обсягів даних. Розподілені системи – це мережі комп'ютерів, які працюють разом, об'єднані для досягнення спільної мети. Ці системи забезпечують величезну масштабованість, надійність та доступність, але також ставлять перед фахівцями нові виклики щодо обробки даних. Розподілені системи стали ключовим елементом сучасного інформаційного ландшафту. Вони забезпечують можливість обробляти великі об'єми даних з високою продуктивністю та надійністю. Вибір методу обробки залежить від конкретних потреб та вимог до системи. Паралельні розподілені системи використовуються для обробки великих об'ємів даних, забезпечуючи швидкість, надійність та масштабованість. Однак одним з ключових викликів є забезпечення ефективного доступу до даних в таких системах. В даній доповіді розглянемо методи доступу до даних в розподілених системах.

Метою доповіді є аналіз існуючих методів обробки даних [1, 2] в розподілених системах та методів доступу до даних в паралельних розподілених сховищах даних на платформі MapReduce. В доповіді наводяться результати досліджень. MapReduce - це модель програмування і асоційоване виконання для обробки та генерації великих наборів даних. Вперше представлений Google у 2004 році, цей підхід з того часу став фундаментом для великих систем обробки даних, таких як Hadoop. MapReduce базується на двох основних процедурах: Map та Reduce. Ці процедури приймають набір вхідних даних і конвертують його у проміжний набір даних у форматі пар "ключ-значення", а також приймають проміжний набір даних і комбінують дані за однаковими ключами. MapReduce пропонує високу масштабованість і надійність для обробки великих наборів даних. Хоча цей підхід має свої обмеження, він продовжує бути ключовою технологією в світі великих даних.

Доступ до даних в паралельних розподілених системах є складною задачею, яка вимагає виваженого підходу до архітектури та вибору технологій. Правильно підібрані методи і рішення дозволяють ефективно працювати з даними, забезпечуючи швидкість, надійність та масштабованість.

Список літератури

1. Варшавский П.Р., Еремеев А.П. Методы правдоподобных рассуждений на основе аналогий и прецедентов для интеллектуальных систем поддержки принятия решений// Новости искусственного интеллекта. - 2006. - №3. С. 39-62.

2. А.А. Барсегян, И.И. Холод, М.Д. Тесс, М.С. Куприянов, С.И. Елизаров. Анализ данных и процессов. 3-е изд. – СПб.: БХВ-Петербург, 2009.

Рисунок Б.1 – Наукова публікація