УДК 004.6

ВИКОНАННЯ РОЗПОДІЛЕНИХ ЗАПИТІВ У МУЛЬТИБАЗОВИХ СХОВИЩАХ ДАНИХ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЇ MAPREDUCE

Яцишин Андрій

Національний технічний університет України «Київський Політехнічний Інститут»

Анотація

У даній доповіді розглядається питання застосування технології MapReduce до проектування мультибазових сховищ даних. Дане рішення позиціонується як гібрид використанн архітектури Virtual Database та технології MapReduce та може бути використано для змішаного (Інтернет та Інтранет) збереження даних.

Abstract

Application of MapReduce technology is considered in this article. This solution is positioned as hybrid use of Virtual Database and MapReduce technologyand can be used for mixed (Internet and Intranet) data storage.

Вступ

На сьогоднішній час розподілені обчислення, зокрема через мережу Інтернет, є поширеним способом обробки та збереження даних. Зокрема, часто використовується платформа даних HortonWorks [1], яка використовує ПЗ Apache Hadoop та технологію MapReduce [2]. Однак при використанні цієї технології контроль цілісності даних лежить на застосуваннях, що обробляють ці дані. Альтернативою є використання архітектур паралельних баз даних, а також так званої "віртуальної бази даних яка надає єдиний доступ до даних, розміщениху різних базах. Вона зберігає цілісність даних за рахунок використання баз даних. Було запропоновано декілька рішень з використання методології MapReduce [3],[4]. Однак поєднання цих технологій не враховує переваг запропонованої автором концепції мультибазових сховищ даних [5-8], у якій сховище напівавтоматизовано проектується як з урахуванням джерел даних, та і з урахуванням запитів. Дана робота висвітлює зусилля автора по поєднанню всіх трьох технологій.

Для кращого розуміння поставленої задачі порівняємо архітектуру Virtual Database і платформу даних HortonWorks.

Зокрема, можна виділити наступне:

- Перевагами HortonWorks є висока швидкодія виконання запитів за рахунок використання розподілених обчислень та надійність за рахунок розподіленої надлишковості даних.
- Недоліками HortonWorks є покладання забезпечення цілісності лежить на застосуванні, тому в окремих випадках цілісність може не підтримуватися.
- Ключовою відмінністю Virtual Database використовує бази даних як носіїв даних, що забезпечує цілісність даних, крім того, за рахунок розподілу даних між базами можлива паралельна обробка запитів, що підвищує їх швидкодію

Оскільки концепція мультибазових сховищ даних [1-3] відповідає архітектурі віртуальних баз даних, до неї також може бути також застосований механізм MapReduce. Застосування цього механізму здійснюється на двох рівнях :

1. Рівень баз даних (за допомогою принципу, використаного в [N2]) для всіх носіїв даних, що полягає у наступній процедурі виконання запитів:

1.1. Модуль Mapper будує пари "ключ-значення" з результатів запитів, що їх надають носії даних;

1.2. Модуль Reduce розподілено пари "ключ-значення" і формує результат підзапиту;

1.3. Модуль Collect формує результат запиту.

2. Рівень елементів даних – використовувується БД HBase на базі Hadoop замість MongoDB, БД XML та файлового сховища даних (в останньому випадку для файлу має бути описано його формат і механізм перетворення в пари "ключ-значення). При цьому взаємодія з цим носієм відповідає роботі з MongoDB.

Варто зауважити, що час виконання запитів при застосуванні технології MapReduce у випадках баз даних NoSQL і при застосуванні її для сховища різний, це пов'язано з тим, що дані у першому випадку отримуються з різних джерел і у випадку NoSQL час отримання даних є одинаковим і залежить від продуктивності обладнання, а в випадку сховища він залежить від різних баз даних і способів їх обробки. Однак це і дозволяє в ряді випадків зменшити час виконання запитів.

У загальному випадку неможливо точно сказати, яка база даних буде краще обробляти запити до певних даних. Тому проектування мультибазових сховищ даних відбувається за двофазним алгоритмом:

1. Фаза проектування сховища, на якій дані розміщуються у носіях сховища на базі їх структурованості. На цій фазі дані розміщуються за детермінованим алгоритмом.
2. Фаза оптимізації сховища, на якій носії оптимізуються для обробки визначених наборів запитів, зібраних за деякий час. На цій фазі виконується генетичний алгоритм, який визначає оптимальні засоби оптимізації.

Крім переваги у швидкості виконанні запитів, може використовуватися поєднання двох архітектур - "внутрішньої" коли бази даних централізовано розміщуються за периметром мережі, і "зовнішньої" коли збереження даних відбувається у вузлах, підключених до мережі Інтернет. Це дозволяє мати гнучкість як у формі обробки запитів, так і у профілюванню безпеки даних, коли дані з підвищеними вимогами до захисту інформації зберігаються у внутрішній мережі, а їх відображення (чи вітрини) через розподілене зберігання даних доступні ззовні периметру.

Отже, отримано рішення, що використовує всі три технології, і забезпечує наступне:

1. Автоматизований розподіл даних у сховищі за допомогою системи керування мультибазовим сховищ даних
2. Оптимізація зберігання у носіях сховища на базі статистики запитів
3. Обробка операцій сховища за рахунок використання механізму MapReduce
4. Використання реляційних та багатовимірних баз даних, що дозволяє забезпечувати цілісність даних
5. Використання нереляційних баз даних, зокрема Apache Hive на базі Apache Hadoop, та баз даних XML (або обробки XML у Hadoop), що дозволяють підвищити швидкодію за рахунок розподіленості обчислень.

Список використаних джерел:

1. Платформа даних Hortonworks [Електронний ресурс]. Режим доступу : http://hortonworks.com.
2. Технологія MapReduce. Режим доступу : http://wiki.apache.org/hadoop/MapReduce
3. Gupta A. Virtual data technology [Текст]. ACM SIGMOD Record 26, 4 / Gupta A., Harinarayan V., Rajaraman A. – 1997
4. Y. Yuan. MapReduce-based distributed data integration using virtual database [Текст]. Future Generation Computer
5. S. Sathya. Application of Hadoop MapReduce technique to Virtual Database system design. Proceedings of IEEE Conference on Emerging Trends in Electrical and Computer Technology/ S. Sathya, M. Victor Jose – 2011

6. Томашевський В.М., Яцишин А.Ю. Проектування мультибазових сховищ даних на основі двохфазного алгоритму [Текст]. Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. /Томашевський В.М., Яцишин А.Ю. – К.: Век+, – 2011. – No 55. – 211