МЕТОДИКА ВИЗНАЧЕННЯ НАЙКРАЩОГО АЛГОРИТМУ ОБРОБКИ ТА АНАЛІЗУ ВЕЛИКИХ ОБСЯГІВ ДАНИХ

Студент О.С. Маришева, д.т.н., доцент, професор Любченко В.В.

Одеський національний політехнічний університет
Україна, Одеса
opu@opu.ua/marysheva.alex@gmail.com

Робота базується на визначенні найкращого алгоритму обробки та аналізу великих обсягів даних. Шляхом імітаційного моделювання двох популярних підходів було виявлено найбільш оптимальний та ефективний щодо часу виконання для досліджуваної предметної області – метод з використанням користувальницьких функцій сховища даних Hive.

Ключові слова: великий обсяг даних, алгоритми обробки даних, MapReduce, імітаційне моделювання, UDF.

Цифрові технології присутні у всіх сферах життя людини, через що обсяг записуваних в світові сховища даних щомиті зростає та створює проблему – умови зберігання інформації повинні змінюватися впроваджуючи нові можливості, що дозволить нарощувати її обсяг, зберігати, структурувати та обробляти існуючу інформацію максимально швидко.

За вихідну точку, з якої почався розвиток інструментів для обробки великих даних, можна прийняти створення Hadoop – набор утиліт, бібліотек та фреймворків для розробки і виконання розподілених програм, на початку 2000-х років [1]. Саме в Hadoop розподілена файлова система HDFS (Hadoop Distributed File System) була об'єднана з фреймворком MapReduce. В результаті з'явився інструмент, який стало можливо використовувати для вирішення найрізноманітніших завдань зі збору та обробки великих даних. Але створення цього інструменту не стало вирішенням глобальної проблеми, опанування цієї технології є дуже складним кроком, який потребує чималих знань спеціаліста у багатьох суміжних областях та багато коштів на створення та підтримку серверів. У зв'язку з цим виникає необхідність ретельного вивчення усіх подробиць використання цього стеку технологій, їх ризиків, переваг та недоліків перед впровадженням, що, нажаль, недостатньо вивчено на сьогоднішній день.

Метою даної роботи є аналіз двох основних алгоритмів для обробки та аналізу великих обсягів даних та порівняння їх ефективності. Допоміжним засобом для моделювання проблеми і її вирішення є імітаційне моделювання, яке без витрат на розробку та тестування може зображати рішення у вигляді прототипу.

Предметна область, яка досліджувалась – система для аналізу закономірності працевлаштування випускників ВНЗ, яка ґрунтується на даних з соціальної мережі «Вконтакті». Інформація, отримана з соціальної мережі була визначена за типом, тому обмежень на використання технологій не було. Аналіз існуючих алгоритмів виявив два основних варіанти реалізації: використання моделі MapReduce та використання користувальницьких функцій сховища даних Hive.

При вирішенні задачі алгоритмом MapReduce процес виконання будується з двох фаз: фази відображення Map і фази згортки Reduce. Функція Map виконує первинну обробку даних, що лежать в HDFS. Вона формує ключі до кожного значення відповідно до задачі. Результати роботи функції Map передаються функції Reduce, яка об'єднує результати та формує остаточний результат. Імітаційна модель досліджуваного підходу зображена на рис. 1.

Другий підхід – UDF дозволяє реалізувати функції або логіку роботи програми, яку складно змоделювати в мові запитів HiveQL, мовою програмування Java. У функції можна задати логіку роботи вибірки з урахуванням необхідних розрахунків, тим самим перенести реалізацію окремих в деяких випадках вкрай витратних операцій вибірок в саме сховище даних, що спри-яє зменшенню часу виконання програми, позбавляє від необхідності багаторазово витягувати дані зі сховища і записувати назад, максимально розмежує функціональні обов'язки сховища даних і призначеної для користувача програми. Імітаційна модель другого алгоритму зображена на рис. 2.

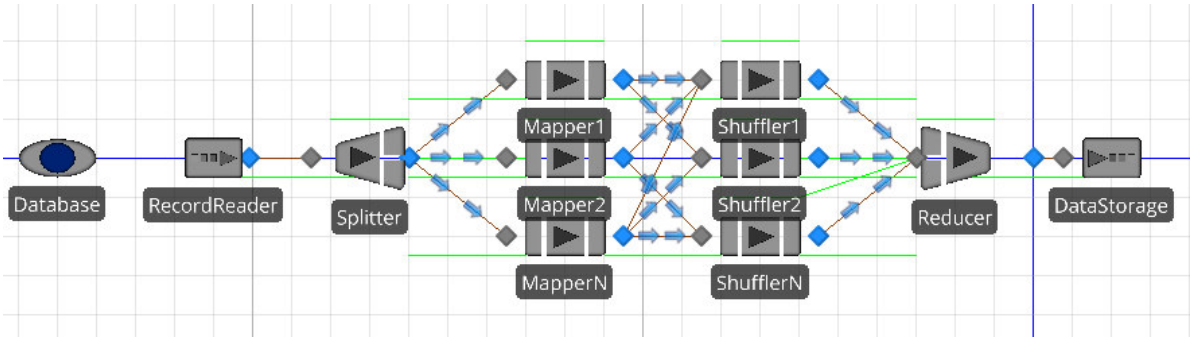

Рис. 1 – Імітаційна модель моделі розподілених обчислень MapReduce

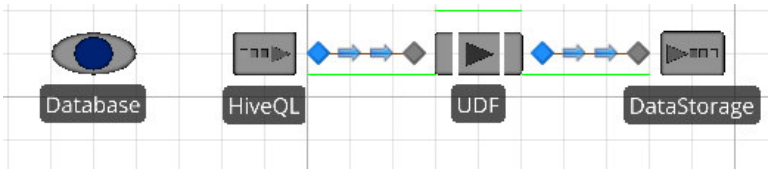

Рис. 2 – Імітаційна модель користувальницьких функції сховища даних

Після проведення експериментів по обробці даних двома зазначеними алгоритмами були зібрані результати порівняння, які зображені у таблиці 1.

Таблиця 1 – Порівняльний аналіз алгоритмів

	Час виконання, с				
	50 об'єктів	500 об'єктів	5000 об'єктів	50000 об'єктів	5000000 об'єктів
MapReduce	8.146	10,165	16,247	21,299	28,315
UDF	0,504	0,889	1,792	2,412	4,159

Виконавши порівняння двох підходів обробки «великих» даних, можна зробити висновок, що використання користувальницьких функцій є більш швидким по часу виконання і хоча їх створення є більш складною процедурою, це можна виправдати їх ефективністю, коли обсяги даних порівняно не-великі, відносно можливостей технологій. Важливим зауваженням є те, що з приростом кількості даних, модель MapReduce починає вигравати в ефективності тому що виконується незалежно від сховища та не завантажує його, сповільнюючи час відклику, що є критичним, коли запитів дуже багато. Але для ситуацій, коли є бажання впровадити системи BigData, кількість клієнтів невелика і обсяги даних вимірюються в Гігабайтах, використання користувальницьких функції є найбільш прийнятним підходом.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Сухобоков А. А., Лахвич Д. С. Влияние инструментария Big Data на развитие научных дисциплин, связанных с моделированием. – МГТУ ім. Н.е. Баумана, Москва, Росія: Електронний журнал, 2015. – С. 211 – 212.

Marysheva O., Liubchenko V.

Method of determining the best algorithm for processing and analyzing large volumes of data

The work based on determining the best algorithm for processing and analyzing large volumes of data. Proceeding from the simulation of two popular approaches was determined the most optimal and time-efficient algorithm for the study area - a method using the user-defined functions of the Hive data warehouse.

Keywords: large amount of data, data processing algorithms, MapReduce, simulation modeling, UDF.