УДК 004.912 : 004.048

Bisikalo Oleg V.
Maksymova Anastasiia T.

ANALYSIS OF THE MODERN METHODS OF STORAGE DATA USING BIG DATA TECHNOLOGIES

Vinnytsia national technical university

Abstract

Now storage and processing of large volumes of data has become a topical issue. More and more operations require fast execution. Hadoop is a distributed file system that also includes a framework for analyzing and transforming very large amounts of data using the MapReduce paradigm.

Keywords: Hadoop, Big Data, Storage Data.

Анотація

У сучасному світі зберігання та обробка великих об’ємів даних стала актуальною темою. Все більше операцій вимагають швидкого виконання. Hadoop - це розподілена файлова система, яка також включає в себе рамки для аналізу і перетворення дуже великих обсягів даних за допомогою парадигми MapReduce.

Ключові слова: Hadoop, великі дані, зберігання даних.

Introduction

To solve the problem of analytics of collected archival data, different approaches can be used. Recently, two paradigms are often compared: traditional (relational databases) and distributed systems for processing large data. Despite the fact that modern SQL-based databases work well enough with the given volume to the appropriate task, this task will be implemented using tools and tools included in the Apache Hadoop ecosystem.

Structured data. One of the differences is that relational databases are structured in their architecture, and many Hadoop applications deal with unstructured data, for example, with text data.

Scaling. Scaling commercial relational databases is expensive. By their nature, they are aimed at vertical scaling: in order to deploy a larger database, it is necessary to purchase equipment that is more powerful.

Storage schema. The underlying principle of relational DBMSs is the placement of data in tables that have a relational structure defined by the schema.

Processing. In essence, SQL is a high-level declarative language. Asking for the data, you say what result you would like to receive, and you provide the DBMS with the decision how to achieve the desired. In the MapReduce paradigm, it is assumed that you describe the specific steps of data processing yourself, which in some way resembles the DBMS-generated SQL query execution plan.

Results of research

Specifics. Hadoop was designed for offline processing and analysis of large amounts of data. It is not intended for random reading and updating of several records, that is, it cannot serve as a substitute for online transaction processing systems. In fact, now and in the near future, Hadoop is best used to work with data warehouses, in which the record is made once, and reading is repeated.

Hadoop is a framework designed to build distributed applications for working with very large data. Hadoop implements the computational paradigm MapReduce, in which the application is divided into many independent parts, each of which can be executed on a separate node.

Hadoop is a distributed file system that also includes a framework for analyzing and transforming very large amounts of data using the MapReduce paradigm [DG04]. Although the HDFS

interface was designed in the same way as the Unix file system interfaces, the developers sacrificed the accuracy of following the standards in order to improve the performance of the applications used.

Hadoop clusters in Yahoo! in aggregate consist of 40,000 servers and store 40 petabytes of application data, with the largest cluster of 4000 servers.

In a fully configured cluster, a set of daemons, or resident programs on different network servers, works underneath. Each demon plays its role; some run only on one server, others on a few. Demons in its function are divided into nodes and calculations.

User applications access the file system using the HDFS client, the library that exports the HDFS file system interface.

Like most traditional file systems, HDFS supports read, write and delete operations, as well as create and delete directories. The user describes the files and directories using paths from the namespace. The user application does not need to worry about the fact that the metadata and file data from the file system are stored on different servers or that the blocks have multiple copies.

When an application reads a file, the HDFS client first requests from the metadata server a list of application data servers storing copies of the data blocks of the required file. The list is sorted based on the distance to the client within the network topology. The client connects directly to the application data server and requests the transfer of the required block. When a client writes, it primarily requires the server of metadata to select the application data servers to store copies of the first block of the file. The client organizes the channel between several servers and sends data. When the first block is transferred, the client requests the selection of the following application data servers to store copies of the next block. A new channel is organized and the client sends the data of the next block. The choice of application data servers for each block is likely to be different.

Unlike traditional file systems, HDFS provides an API that allows you to locate the locations of the file data blocks. This circumstance allows such applications as the MapReduce framework to schedule tasks on the servers where the necessary data is located, thereby increasing the speed of data reading. Usually, the files are subjected to three-time replication. In the case of working with important files or files, which are accessed very often, an increased degree of replication increases the resistance to failures and the speed of reading data.

The interactions between the client, the metadata server, and the application data servers are shown in figure 1.

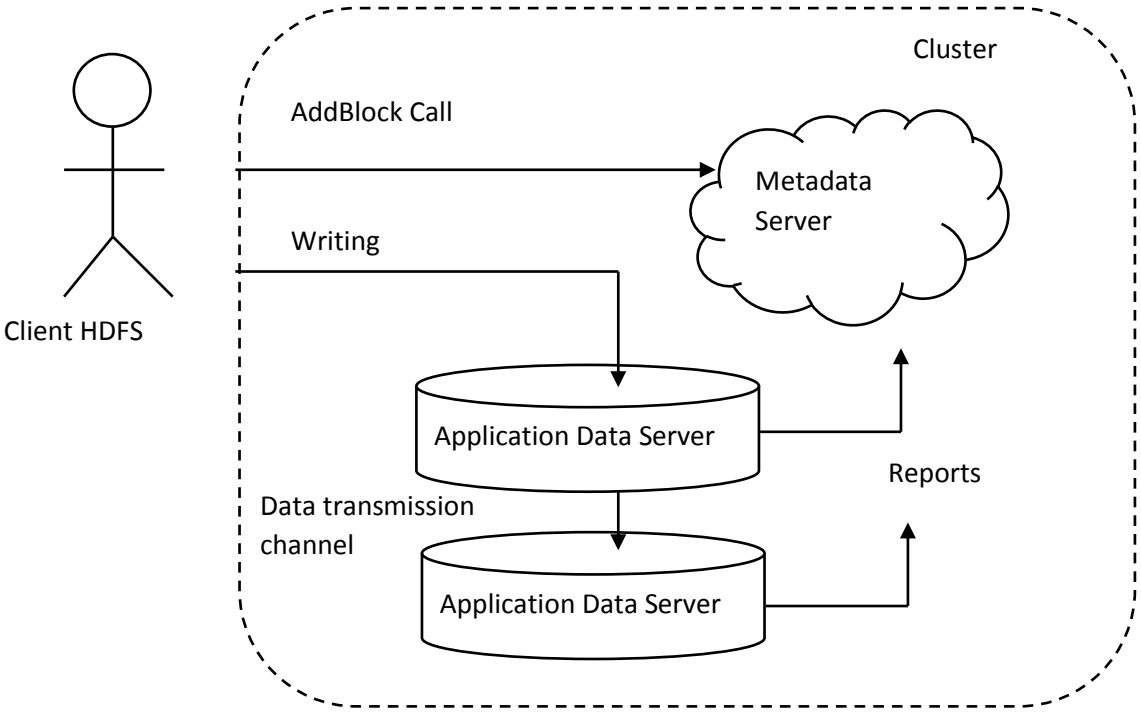

Figure 1 – Interactions between the client, the metadata server, and the application data servers

Conclusions

The proposed method allow the distribution of data and computing resources between many (thousands) of nodes, as well as the execution of calculations provided by applications in parallel with the delivery of the necessary data. The Hadoop cluster allows you to scale the computing resources, storage capacity and bandwidth of channels for I / O operations by simply adding purchased servers.

References

1. Schmarzo B. Big Data: Understanding How Data Powers Big Business / B. Schmarzo. – M., Wiley, 2013. – 240 p.
2. Arvind S. Big Data Analytics: Disruptive Technologies for Changing the Game / S. Arvind. – M., 2012. – 323 p.
3. Кучерук Ю.В. Як Big Data (великі дані) впливають на буття людини / Ю.В. Кучерук, І. О Головашенко, М.В. Глушко ; Нац. Ун-т «Вінницький національний технічний університет». – Вінниця : Вид-во Нац. Ун-ту «Вінницький національний технічний університет», 2017. [Електронний ресурс]: Інформаційний портал. Режим доступу: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/17811/2555.pdf?sequence=3&isAllowed=y

Maksymova Anastasiia T. – student of the group 3ACIT-17m, control computer system department, Vinnytsia national technical university.

Supervisor – Bisikalo Oleg V. – Dr. of Sci., professor, Dean of the Faculty of Computer Systems and Automatics, Vinnytsia national technical university.

Максимова Анастасія Тарасівна – студентка групи 3АКІТ-17м, кафедра комп'ютерних систем управління, Вінницький національний технічний університет.

Науковий керівник– Бісікало Олег Володимирович – доктор технічних наук, професор, декан факультету комп'ютерних системта автоматики, Вінницький національний технічний університет.