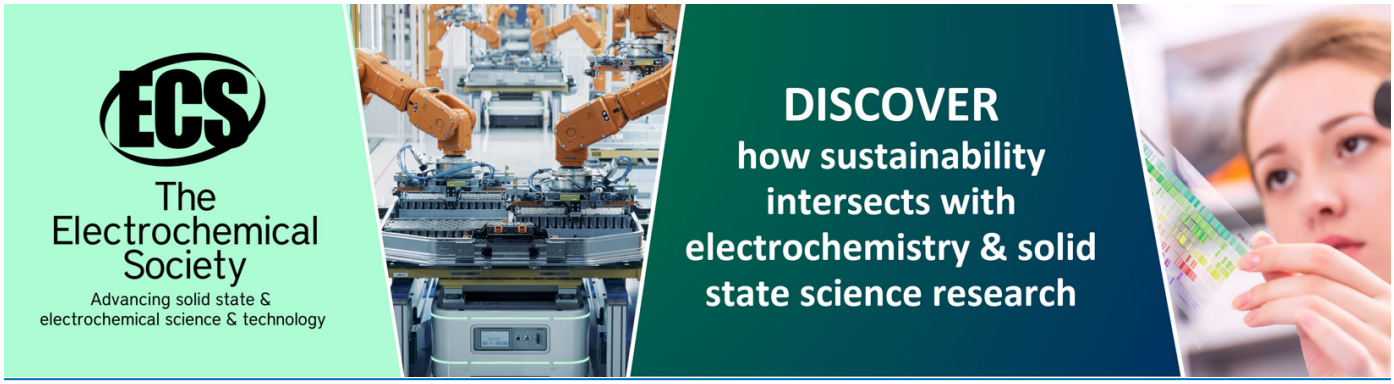
MapReduce and Apache spark: technology analysis, advantages and disadvantages



MapReduce and Apache spark: technology analysis, advantages and disadvantages

T Q Urazmatov and X Sh Kuzibayev

Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Urgench, Uzbekistan

E-mail: tohir20314@gmail.com

Abstract. Nowadays, it is absolutely illogical and impossible to process big data using traditional software methods and hardware. because too much data available does not allow this. However, there are some effective ways to perform such operations. This article discusses the main problems and solutions for processing big data. Today, there are a number of technologies and algorithms that process and analyze big data. This article mainly discusses, analyzes, and summarizes the advantages and disadvantages of the MapReduce architecture and Apache spark technology, and the results are presented in tabular form.

1. Introduction

Big data size means data that can be more than a hundred terabytes and petabytes. In addition, this information is regularly updated. For example, data from contact centers, social networks, stock trading data, and so on. Also, the concept of big data sometimes includes ways to process them. If we talk about terminology, then Big Data means not only data, but also the principles of processing large volumes of data, their subsequent use, the order of definition of a particular block of data in large arrays. Questions related to such processes do not lose their relevance.[13] Their solution is important for systems that produce and collect a variety of data over many years. Big data have the following main features.

- Volume - about 1 petabyte and higher.
- Velocity - high speed data creation, reception and processing
- Variety - heterogeneity of data, lack of different formats and possible structure.

Often two more factors are added to these parameters:

- Variability - different intensities of income influencing the choice of processing methods
- Value - the difference in the level of complexity of the data obtained.

2. Materials and methods

Differences between MapReduce and Apache spark

MapReduce is a programming mechanism for processing and creating large data sets with a parallel, distributed algorithm in a computer cluster. MapReduce consists of several components, including:

- JobTracker - the main node that manages all work and resources in the cluster.
- TaskTrackers - agents placed on each machine in the cluster to run the map and reduce tasks.
- JobHistoryServer - a component that tracks the work done and is usually deployed as a separate function or with JobTracker.

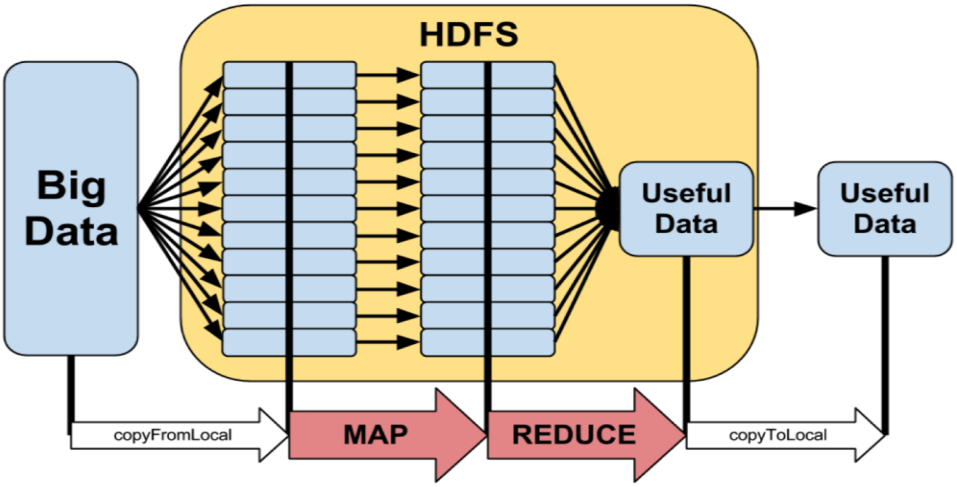


Figure 1. MapReduce architecture.

Apache Hadoop is one of the open-source software systems. It is developed for measure applications from a single server to hundreds of machines and run applications in accumulations with traditional equipment. The Apache Hadoop platform falls into two categories.

- Hadoop shared file system
- Recycled layer

The Hadoop storage layer, i.e. HDFS, is responsible for storing data, and the MapReduce is responsible for processing data in the Hadoop cluster. MapReduce is a programming model that provides extensions for tens of thousands of servers in a Hadoop accumulation. MapReduce is a machining technique and programming language for shared computing based on the Java programming model. MapReduce is a strong framing for processing large shared sets of configured or non-structured information in a Hadoop cluster stored in a Hadoop shared file system.[4] MapReduce's strong feature is its scalability.

Apache Spark is a cluster model shared for fast computing in the processing of big data. Apache Spark is a distributed recycling engine, but it does not come with a integral cluster resource manager and distributed warehouse system. You need to connect the selected cluster manager and the warehouse system. The Apache Spark consists of a set of libraries similar to the existing ones for the Spark kernel and Hadoop. [8] The core is a set of shared execution mechanisms and models. Apache Spark helps languages such as Java, Scala, Python, and R for developing shared applications. More libraries have been throwing out top of the Spark core to help workloads that use streaming, SQL, graphics, and machine learning. Apache Spark is a data processing mechanism for bulk and flow, which includes SQL queries, graphics, and machine learning. Apache spark can work separately, as well as in the Hadoop YARN steam supervisor, so that it can read available Hadoop information.

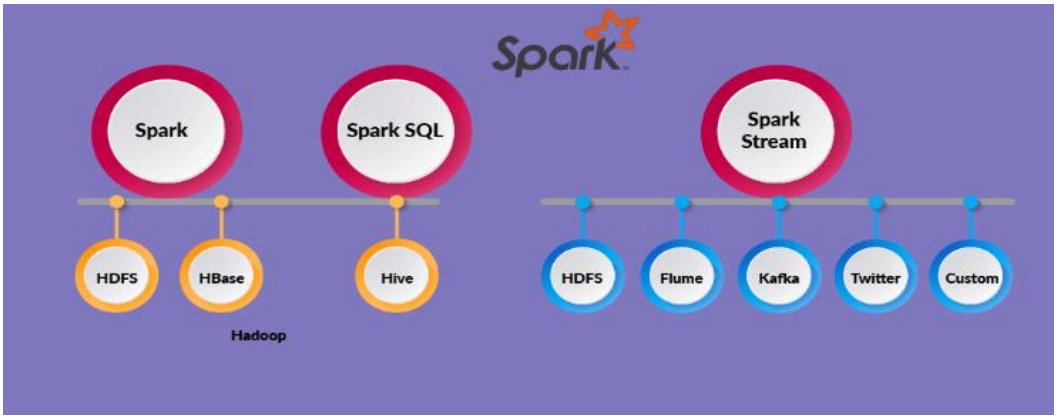


Figure 2. Apache spark technology.

The basic distinction among MapReduce and Apache spark

- MapReduce and Apache Spark are similar in terms of data types and information sources.
- MapReduce is based on hard drive, while Apache spark benefits fast remembrance and can benefit disk for cultivationing.
- Hadoop MapReduce is designed for non-volatile data, while Apache spark has superior production for non-volatile data, especially in allocated clusters.
- The basic distinction among MapReduce and Spark is that MapReduce benefits non-volatile memory, while Spark uses a permanently distributed information set.
- Apache spark and Hadoop MapReduce are both fault forbearing, but relatively more fault tolerant than Hadoop MapReduce Spark.
- Hadoop MapReduce jerrycan be an inexpensive choice because Hadoop is a service and Apache spark is economical due to high memory availability.
- Spark can do mass processing 20-200 tact's quicker than MapReduce, but both instruments are applied to process large amounts of data.[7]
- Hadoop MapReduce needs basic Java planning expertise, while Apache spark planning is not difficulted because it has an interactive way.

How to use MapReduce:

- In outline processing of big data sets
- When no halfway suspension is needed.

How to use Apache spark:

- When processing data quickly and interactively;
- When joining the data set;
- When processing schedules;
- When performing repetitive work;
- Real-time processing;
- In machine learning

3. Results

As a result of our research, we have the following facts. We compared MapReducate and Apache spark based on 20 types of parameters. We placed the results of the studies in the table below.

Table 1. MapReduce and Apache spark differentiation list.

Parameters	(a) Apache spark	(b) MapReduce
SQL	Spark supports via SQL	Hive supports query language
Security	The security features of the Apache spark are evolving and becoming more sophisticated.	The MapReduce framework is safer compared to the Apache spark
Scheduler	Apache spark has its own planner	Depending on the external planner
Palatability	Both scales are restricted to 2000 branches per cluster	Both scales are restricted to 2000 branches per cluster
Resistance to errors	Apache spark interests RDD and other preservation imitations for mistake opposition	Use replication for error tolerance
Processing speed	100 times faster in memory and 10 times faster on disk	Slower than Apache spark because if there is an input / output delay on the disk
Machine learning	Apache spark has built-in machine learning APIs	MapReduce is more suitable with Apache Mahout when integrated with machine learning.
Language supporting	Apache spark supports Java, Scala, Python and R.	The main language is Java, but languages such as C, C ++, Ruby, Python, Perl, Groovy are also supported.
Interactive mode	Interactive	Not interactive
Infrastructure	Medium and high-level hardware	Brand equipment
Ease of use	Apache spark is uncomplicated to use because of the APIs	Thanks to the MapReduce JAVA API, it is a bit more complicated compared to the Apache spark
Duplication elimination	Apache spark processes each entry exactly once, thus eliminating duplication.	MapReduce does not help this feature
Delay	Faster compared to MapReduce Framework	Very high delay
Information cultivationing	Mass cultivationing, as well as real-time information cultivationing	For mass cultivationing only
Costs	More expensive due to the huge quantity of RAM	Cheaper compared to Apache spark
Complexity	not difficult to compose and troubleshooting	Codes are hard to compose and troubleshooting
Suitabilitiy	Apache spark can merge with each of information resource and file forms helped by the Hadoop cluster.	Suitable with each of information resource and file forms
Coding	Fewer code rows	More code lines
Category	Data Analytics Engine	Data processing mechanism
Apache	An open-source substructure for high-speed information cultivationing	An open-source substructure for information cultivationing

4. Conclusion

Both of the above algorithms are important tools for processing large amounts of information. The main advantage of MapReduce is that the processing of cluster nodes is easy to solve. Apache Spark and MapReduce perform high-level calculations when both are used together. Apache Spark is mainly used for real-time data processing. In MapReduce and Apache Spark, data processing is limited to a thousand nodes per cluster. That's enough for a lot of data right now. MapReduce is a relatively secure system in terms of data and algorithm security. Apache Spark security is now evolving. Comparing the two financially, Spark requires a lot of money because it requires a high amount of RAM. Another key feature of both algorithms is their error tolerance level. MapReduce uses replication for error tolerance, while Apache Spark uses RDD models. Another key feature is the supported programming languages. MapReduce and Apache Spark also support several programming languages, including Java, C, C ++, Python, and other programming languages. In short, MapReduce and Apache Spark are the main tools for processing large amounts of data.

References

[1] Franks B 2012 *Taming the Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics* (Wiley and SAS Business)
[2] Gantz J and Rainsel D 2013 *The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (United States)
[3] Hadoop and Big Data: http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html.
[4] Afzali G A and Mohammadi Sh 2016 Privacy Preserving Big Data Mining: Association Rule Hiding. 10.7508/jist.2016.02.001. http://www.jist.ir/Article/139504261512112857
[5] Kachalov D L, Mishustin A V and Farkhadov M P Institute of Control Problems of the Russian Academy of Sciences named after V.A. Trapeznikova. Modern methods of processing big data in large-scale systems http://www.hozir.org/mapreduce-and-apache-spark-technology-analysis-advantages-and.html?page=4
[6] Urazmatov T Q, Nurmetova B B and Kuzibayev X Sh 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **862** 042006