UDK 004.62

ANALYSIS MODEL MAPREDUCE TECHNOLOGY «BIG DATA»

ANTON STANOVOY, SERGEI SURTO
Polotsk State University, Belarus

Although they have existed for several years big data have not previously been of great value, because their processing and analysis were difficult. This required substantial computing power, long time and financial costs. Everything changed when the technology of processing multi-gigabyte arrays of information in fast RAM appeared. A breakthrough in this area is associated with the launch of the free Hadoop platform, including libraries, utilities and frameworks for working with Big Data. Hadoop components are used today in most commercial platforms and systems of companies such as SAP, Oracle, IBM, and so on. Today, the term Big Data, as a rule, is used to refer not only to the data arrays themselves, but also to tools for their processing and the potential benefits that can be obtained as a result of painstaking analysis.

Introduction. The widespread introduction of the term «big data» is associated with Clifford Lynch, the editor of Nature magazine, who prepared a special issue for September 3, 2008 with the topic «How can technology influence the future of science, opening up opportunities for working with large amounts of data?» about the phenomenon of explosive growth in the volume and diversity of the processed data and technological perspectives in the paradigm of the likely jump «from quantity to quality»; The term was proposed by analogy with the «big oil», «big ore» metaphors in the business English-speaking environment. Despite the fact that the term was introduced in an academic environment and, above all, the problem of growth and diversity of scientific data was understood, since 2009 the term has been widely spread in the business press, and by 2010 the appearance of first products and solutions related exclusively and directly to the problem of processing big data. By 2011, most of the largest suppliers of information technology for organizations in their business strategies use the concept of big data, including IBM, Oracle, Microsoft, Hewlett-Packard, EMC, and the main analysts of the information technology market devote dedicated research concepts. In 2011, Gartner noted big data as trend number two in the information technology infrastructure (after virtualization and as more significant than energy saving and monitoring). At the same time, it was predicted that the introduction of big data technologies would have the greatest impact on information technologies in manufacturing, health care, trade, government, as well as in areas and industries where individual movements of resources are recorded [1].

Since 2013, big data as an academic subject has been studied in new university programs on data science and computational sciences and engineering.

In 2015, Gartner eliminated big data from the maturity cycle of new technologies and stopped releasing a separate maturity cycle of big data technologies in 2011-2014, motivating this by moving from the stage of hype to practical application. The technologies that figured in the dedicated maturity cycle, for the most part, moved into special cycles on advanced analytics and data science, on BI and data analysis, corporate information management, resident computing, and information infrastructure. One of the main big data models is Mapreduce [2].

Main part. MapReduce is a distributed data processing model proposed by Google for processing large amounts of data on computer clusters. MapReduce is well illustrated by the picture (Figure).

MapReduce assumes that data is organized in the form of some records. Data processing occurs in 3 stages:

1. Stage Map. At this stage, the data is processed using the map () function that the user defines. The job of this stage is to pre-process and filter the data. The operation is very similar to the map operation in functional programming languages — a user-defined function is applied to each input record. The map () function applied to a single input record and produces a set of key-value pairs. The set - i.e. can give only one record, can give nothing, and can give several key-value pairs. What will be in the key and in the meaning is up to the user, but the key is a very important thing, since the data with one key will fall into one instance of the reduce function in the future.

2. Stage Shuffle. It passes unnoticed by the user. At this stage, the output of the map function is "sorted into baskets" – each basket corresponds to one output key of the map stage. In the future, these baskets will serve as an input for reduce.

3. Stage Reduce. Each "basket" with values, formed at the shuffle stage, goes to the input of the reduce () function. The reduce function is set by the user and calculates the final result for a single «basket». The set of all values returned by the reduce () function is the final result of the MapReduce task [3].

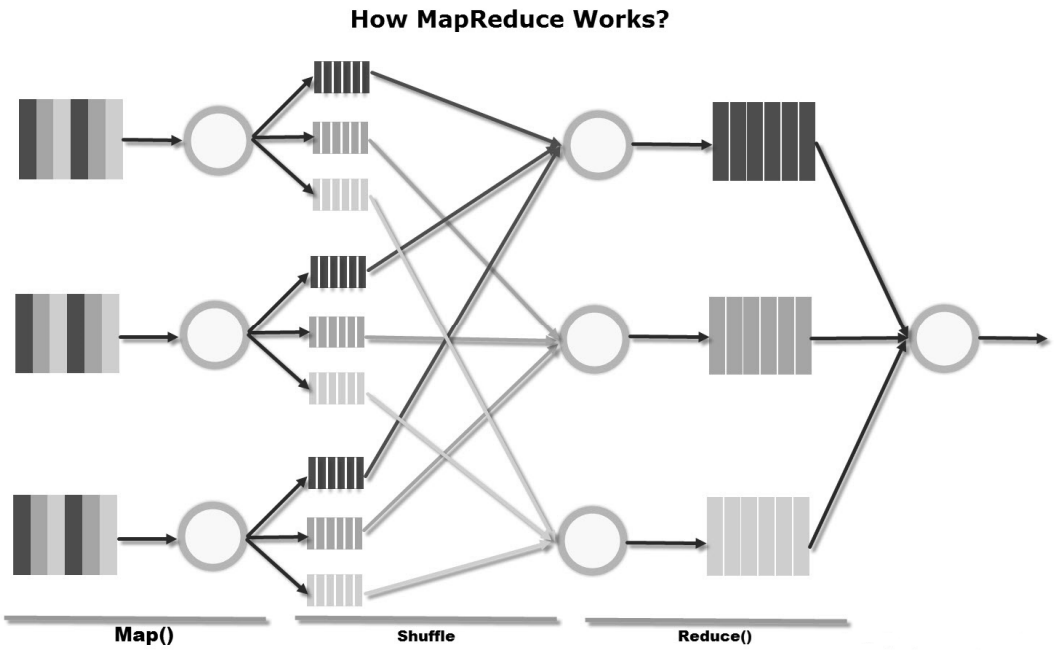

Figure 1. – Model MapReduce

Some additional facts about MapReduce:

1) All launches of the map function work independently and can work in parallel, including on different cluster machines.

2) All launches of the reduce function work independently and can work in parallel, including on different cluster machines.

3) Shuffle inside represents parallel sorting, so it can also work on different cluster machines. Items 1–3 allow you to perform the principle of horizontal scalability.

4) The map function, as a rule, is used on the same machine on which the data is stored - this reduces the transmission of data over the network (the principle of data locality).

5) MapReduce is always a full scan of the data, there are no indices. This means that MapReduce is poorly applicable when a response is required very quickly [4].

Conclusion. Currently, enterprises have to work with large amounts of information, which is often updated and comes from different sources. With the help of Big Data technology, enterprises can analyze huge amounts of data and identify useful patterns that give them competitive advantages.

For easier perception and quick decision-making, it is necessary to present the results of data analysis visually. At the moment there are several types of data arrays. But existing visualization methods are still underdeveloped and need improvement.

Companies that have already implemented Big Data technologies will have a great competitive advantage in the future.

REFERENCES

1. Новостной портал [Электронный ресурс]. – Режим доступа: https://www.kommersant.ru/doc/2614791. – Дата доступа: 19.02.2019.
2. Энциклопедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Big_data. – Дата доступа: 19.02.2019.
3. Веб-сайт в формате коллективного блога с элементами новостного сайта, созданный для публикации новостей, аналитических статей, мыслей [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/dca/blog/267361/. – Дата доступа: 19.02.2019.
4. Personal blog [Electronic resource]. – Mode of access: / Pinal Dave https://blog.sqlauthority.com/2013/10/09/big-data-buzz-words-what-is-mapreduce-day-7-of-21/. – Date of access: 19.02.2019.