УДК 004.27

N. A. Zhilyak, Mohamed Ahmad El Seblani
Belarusian State Technological University

CLASS TECHNOLOGY ANALYSIS OF BIG DATA

The article discusses an overview of some technologies of BIG DATA class. The article includes the classification and analysis of methods of processing large amounts of data. The theoretical aspects associated with the emergence of the phenomenon of big data, explores the epistemology and heuristic possibilities of big data. The practical significance of the chosen theme is to develop new methods and algorithms for analysing large amounts of data (BIG DATA), allowing early detection of possible loss or distortion of information, which in turn may lead to reduction in financial losses. This article will be useful for specialists dealing with the problems of organization and processing of databases, in particular BIG DATA.

Key words: Big Data, business, factor, scoring technology.

Introduction. The category of large Big Data includes information which is no longer possible to process by conventional methods, including structured data, media and random objects. Some experts believe that in order to work with them to replace the traditional monolithic systems have new massively parallel solutions. From the name we can assume that the term “great data” simply refers to the large amounts of data management and analysis. According to the report McKinsey Institute “big data: the new frontier for innovation and competition” (Big data: The next frontier for innovation, competition and productivity), the term “great data” refers to data sets whose size is beyond the capabilities of typical databases for named, storage, management and analysis of information. And global repository of data, of course, continue to grow [1].

Big Data suggest something more than just an analysis of huge amounts of information. The problem is not that organizations create huge amounts of data, but the fact that most of them are presented in a format that bad associated traditional structured format database – a web-based magazines, videos, text documents, computer code, or, for example, geospatial data. Everything is stored in a variety of different storage facilities, sometimes even outside the organization. As a result, corporations can have access to a huge amount of their data and do not have the necessary tools to establish the relationship between these data and make on the basis of their significant conclusions. Add to this the fact that the data is now updated more and more, and you get a situation where the traditional data analysis methods can not keep up with the vast amounts of constantly updated data, which ultimately paves the way for big data technologies. The aim of the further work with big data is the development of methods and algorithms for processing large data scoring model [2].

Main part. The Big Data movement has only magnified the complexities that have existed in data architectures for decades. Any architecture based primarily on large databases that are updated incrementally will suffer from these complexities, causing bugs, burdensome operations, and hampered productivity. Although SQL and NoSQL databases are often painted as opposites or as duals of each other, at a fundamental level they are really the same. They encourage this same architecture with its inevitable complexities [3].

In fact the concept of big data involves work with the vast volume of information and varied composition, very frequently updated and located in different sources in order to increase efficiency, create new products and improve competitiveness. Forrester Consulting Company gives a brief formulation: “Large data combined techniques and technologies that extract meaning from data practicality” extreme limit.

Craig Baty, Executive Director of Marketing and Director of Fujitsu Australia Technology, pointed out that business analysis is descriptive results of the analysis process, achieved business in a certain period of time, whereas the speed of large data allows us to analyze predictive capable of offering business recommendations future. Big Data technologies can also analyze more data types in comparison with business intelligence tools, which makes it possible to focus not only on the structured storage.

Matt Slocum from O'Reilly Radar says that although big data and business analytics have the same target (search for answers to the question), they differ from each other in three dimensions.

Big data is designed to handle larger amounts of information than a business analyst, and this, of course, corresponds to the traditional definition of big data.

Big data is designed to handle more quickly received and changing information, which means that in-depth study and interactivity. In some cases, the results are generated faster than loading a web page [4].

Big data is intended for processing unstructured data, use of which we are only beginning to study after they were able to organize the collection and storage, and we need algorithms and the ability to dialogue in order to facilitate the search trends contained within these arrays [3].

According to the Oracle white paper published by "Oracle Information Architecture: Architect Guide great data" (Oracle Information Architecture: An Architect's Guide to Big Data), when working with large data, we come to the information other than during business analysis.

Analysis of Big Data, which raises the question of how to work with unstructured information, generate analytical reports, as well as the implementation of predictive models [4].

Market Big Data projects intersect with the market of business intelligence (BA), the volume of which in the world, according to experts, it amounted to about 100 billion dollars in 2012. It includes a networking component, servers, software and technical services.

Also, the use of Big Data technologies relevant for the class revenue assurance solutions (RA), designed to automate the activities of companies. Modern revenue assurance systems include inconsistencies detection tools and in-depth analysis of data, allowing early detection of loss or distortion of information that could lead to a decrease in financial results. Against this background, Russian companies, confirming the presence of Big Data technologies in demand in the domestic market, noted that factors that stimulate the development of Big Data in Russia are data growth, accelerate management decision-making and improve their quality.

Unfortunately, today, only 0.5% of analyzed digital data accumulated, despite the fact that there are objectively industry-wide problem which could be solved by making analytical grade Big Data. Development of IT-markets already have results, which can assess the expectations associated with the accumulation and processing of large data. One of the main factors which hinders the implementation of Big Data – projects, in addition to the high cost, it is considered the problem of selecting data to be processed: that is, to determine which data need to extract, store and analyze, and what – is not taken into account.

There are many hardware and software combinations that allow you to create effective solutions for Big Data of various business disciplines, from social media and mobile applications to intelligent analysis and visualization of business data. An important advantage of Big Data – it is compatible with the new tools are widely used in business database, which is especially important when dealing with cross-disciplinary projects, for example, such as the organization of multi-channel sales and customer support.

The sequence of work with Big Data includes data collection, structuring the information obtained via reports and dashboards (dashboard), creating insights and contexts, as well as the formulation of recommendations for action. Since working with Big Data implies high costs of data collection, which is the result of processing is not known beforehand, the main task is a clear understanding of what data are needed, and not how much they have in stock. In this case, the collection of data is converted into the process of obtaining the necessary solely for specific tasks of information [4].

Based on the definition of Big Data, we can formulate the main principles of work with the following data:

– horizontal scalability. Since data can be arbitrarily long – any system that involves processing of big data must be scalable. 2 times increased the volume of data in 2 times increased the amount of iron in the cluster, and all continued to work;

– fault tolerance. The principle of horizontal scalability implies that the machines in the cluster can be many. For example, Hadoop cluster Yahoo has more than 42,000 machines. This means that some of these cars is guaranteed to fail. Methods of working with big data should consider the possibility of such failures and survive them without any significant consequences;

– the data locality. In large distributed systems data spread over a large number of machines. If the data is physically located on the same server, and processed on the other – the data transfer costs can exceed the cost of the treatment itself. Therefore, one of the most important design principles big data solutions is the principle of data locality – if possible, process data on the same machine on which they are stored.

All modern means of big data one way or another followed these three principles. In order for you to follow – you must invent some methods, techniques and paradigms of development, development tools data. One of the classical methods I will explore in today's article.

MapReduce is a distributed processing model proposed by Google for processing large amounts of data on computer clusters. MapReduce is illustrated by the following (Fig. 1).

MapReduce assumes that the data is organized in records. Processing of data occurs in three stages:

1. The Stage Map. At this stage the data predostavlyayutsya function map () that the user defines. The work of this stage is pre-processing and filtering. The work is very similar to the map operation in functional programming languages – user-defined function is applied to each input record.

The map() function applied to one input record and outputs a set of pairs key-value. Many ie only issues a single entry may not give anything, and can give out a few pairs key-value. What is the key and the value to solve, but the key is a very important thing, since the data with one key in the future will fall into one instance of the reduce function.

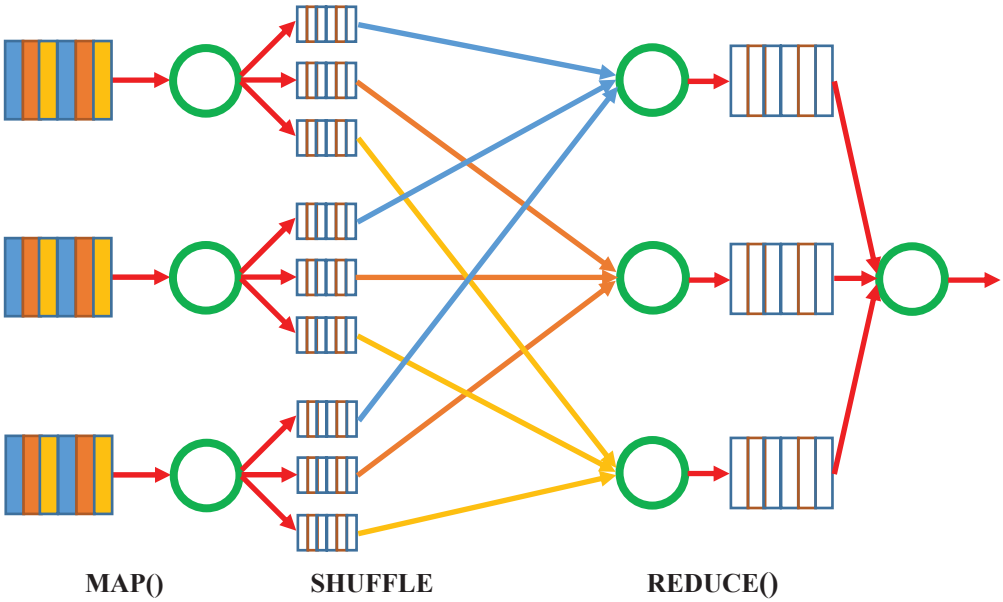

Fig. 1. The distributed processing model

2. Stage Shuffle. Runs invisibly to the user. In this stage, the output of the map function and "versed in the baskets" – each basket corresponds to one key of the output stage of the map. In the future, these will serve as input to reduce.

3. Stage Reduce. Each "basket" with the values generated at the stage of shuffle, gets the input of reduce().

The reduce function is specified by the user and calculates the final result for the individual "basket". The set of all values returned by the function reduce () is the final result of a MapReduce task.

A few additional facts about MapReduce:

1) for All runs map function work independently and can work in parallel on different machines of the cluster;

2) all runs of the reduce function are independent and can run in parallel, including on separate machines in the cluste;

3) shuffle within itself is parallel sorting, so it can also run on different machines in the cluster. Paragraphs 1 to 3 allow you to perform the principle of horizontal scalability;

4) map Function, usually used on the same machine on which data is stored – it allows to reduce data transfer over the network (the principle of data locality);

5) mapReduce is always a full scan of the data, no indexes, no. This means that MapReduce is hardly suitable when the response is required very quickly.

Those who are accustomed to working with relational databases, often use a very convenient Join operation that allows to simultaneously process the content of some tables, combining them according to some key. When working with big data this problem is also sometimes. Consider the following example.

There are logs of two web servers, each log is as follows:

\t\t. Example piece of log.

```
1446792ll139
178.7ll8.82.1/sphingosine/unllhurrying.css
1446792ll139 126.3ll1.163.222 /accentually.jsll
1446792139  154.1ll64.149.83
    /pyroacid/unkemllptly.jpg
1446792ll139 202.2ll7.13.181/Chawia.jsll
1446792ll139 67.12ll3.248.174
    /morphograllphical/dismain.css
1446792ll139 226.7ll4.123.135 /phaneritell.php
1446792ll139 157.1ll09.106.104
    /bisonant.css
```

You need to count for each IP address on which of the 2 servers he often came by. The result should be presented in the form:

\t. An example of the result:

```
178. ll78.82.1 first
126. ll31.163.222      second
154.164.149.83         second
226. llll74.123.135    first
```

Unfortunately, unlike relational databases, in General the Union of two logs according to the key (in this case IP address) is a rather heavy operation and can be solved using MapReduce pattern and Reduce the Join (Fig. 2).

It is important that at this moment on reduser get entries from both logs and the type field can be used to identify from which of the two logs got to a specific value. So the data is enough to solve the original problem. In our case, reducere just have to count for each key of the records with which the type was found more and bring this type.

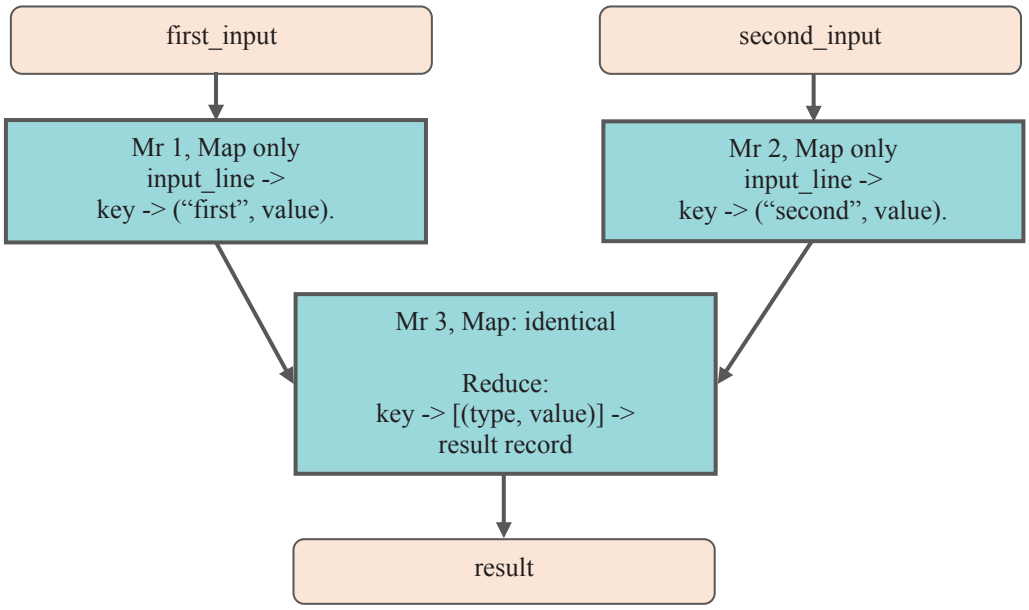

Fig. 2. The MapReduce pattern

Another example – how banks can use Big Data to prevent fraud. If the customer says about loss of the card, and when making a purchase with the help of the bank sees in real time the location of the customer's phone in the shopping area, where the transaction, the bank may verify the information on the client's request, I did not try whether he deceive him. Or the opposite situation, when a customer makes a purchase at the store, the bank sees that the card, on which the transaction, and the client's mobile phone are in one place, the bank may conclude that the card uses the owner. Due to such advantages Big Data, expand the boundaries of which are endowed with traditional data warehouse.

Conclusion. After analyzing the different approaches to big data revealed that an important factor in achieving business goals when working with large data is to choose the right strategy, which includes analytics, that highlights the demands of consumers, as well as the use of innovative technologies in the area of Big Data.

In practice, a chain of MapReduce tasks can be quite complex sequences that MapReduce tasks can be connected as series and parallel to each other.

Thus, it must be noted that while the “Big Data” – this is a great potential that still need to be able to take advantage of.

Thus one of the goals is the development and analysis of new methods and algorithms for the organization of unstructured information for big data Analytics on the basis of scoring model, and classifying the customer base into different groups with unknown characteristics that separates these groups.

References

1. Konstantin B. Optimizations in computing the Duquenne-Guigues basis of implications – *Annalise of Mathematics and Artificial Intelligence*, 2014, vol. 70, no. 1–2, pp. 5–24.

2. Obiedkov S. Modeling ceteris paribus preferences in formal concept analysis, in: Formal Concept Analysis. Vol. 7880. Berlin, Heidelberg, Springer, 2013. P. 188–202.

3. Raman V., Swart G. How to wring a table dry: Entropy compression of relations and querying of compressed relations. *Proceedings of International Conference on Very Large Data Bases (VLDB)*, 2006, pp. 34–46.

4. Stonebraker M., Cetintemel U. One size fits all: An idea whose time has come and gone. *Proceedings of the International Conference on Data Engineering (ICDE)*, 2008, pp. 18–64.

Information about the authors

Mohamed Ahmad El Seblani – PhD student. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: msiblani@hotmail.com

Zhilyak Nadezhda Aleksandrovna – PhD (Engineering), Assistant Professor, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: gh_nadya@mail.ru

Received 14.09.2017