Analysis of big data processing technologies

T Q Urazmatov, B B Nurmetova and X Sh Kuzibayev

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

E-mail: tohir20314@gmail.com

Abstract. This article describes large amounts of information. Examples of large volumes are provided and some of their processing technologies have been analyzed. The advantages and disadvantages of the methods of processing parallel processing technologies of large volumes of data processing, the methods of selection of newly selected methods of selection and models are considered. These methods are compared with the distributed technology of data processing. This article discusses a few of the large volumes of data processing principles. They are: the principle of horizontal measurements, the principle of inverse stability, and the principles of data positioning. On the basis of horizontal measurements, any data processing system is expanding, with data doubling. In contrast to the principle of horizontal stability, the number of machines should not be increased. In the location of data, the data is distributed to a large number of machines. If the data is located on a server and processed in another, the cost of data transmission may exceed the cost of processing. Therefore, the most effective principle of data processing is the principle of information location processing. It is desirable to recycle data in a machine that will save them if possible. All models of large volumes of data processing are available in these three methods. This article analyzes a larger model of data that works with these principles. The MapReduce model also operates on the basis of the above three principles. This is also a model for distributed data processing to process large amounts of data in the computer cluster. In this article, several methods and models are analyzed and their advantages and disadvantages are discussed.

1. Introduction

Although Big Data is already available in scientifically-related areas, this phenomenon has been highlighted in recent years. The reason is that new technologies, devices and communication tools are rapidly developing. And the dramatic increase in the amount of data produced by human beings through the widespread proliferation of social networks. Today, human information is being generated every 10 minutes. New types of algorithms and technologies are developed and widely used in practice. By collecting, processing, and using billions of aggregated data around us, we can have important historical, current, and future information about a particular situation or object. First of all, it is important to identify this source. Big Data concepts are used for operators that can perform large-scale tasks. For example, Facebook has more than 10 million new photos per hour. Big Data analysis can help you identify the rules beyond human capability. This will allow us to further improve our everyday life - government management, medicine, telecommunications, finance, transportation, manufacturing and other industries,
and to find alternative solutions to the problems. Today, Big Data is widely used by financial institutions, and the key issues such as increasing the competitiveness and efficiency of data analysis are being solved. The use of critical information, in addition to creating additional tasks and requirements for the information expert, will lead to the emergence of new jobs such as Data mining and Data Scienceist. "Data mining" is an analysis of the data to find out what's going on.

When Big Data is over 100 GB:
Big Data is data that can not be processed in Excel;
Big Data is data that can not be processed on a single computer;
Large amounts of information are the continuous growth of methods, methods and solutions for the various structured and not structured data processing. Larger volumes of data are not the amount of data, but their processing methods. Some of the following data sources can be an example of a bulk data source:
- user history logs;
- stream of GPS signals coming from vehicles for traffic companies;
- all bank customers' transactions;
- information on all purchases in the retail chain and so on.

The number of large-scale sources of information is growing rapidly, that is, their processing technology is becoming increasingly popular.

2. Principles of processing Big Data
Based on the Big Data definition, it is possible to formulate the basic principles of processing such data. These are:

1. Horizontal measurements. Any system that involves the processing of large data as much as possible for large data processing can be expanded. The size of the data doubles, the number of components in the cluster doubles, and the process continues.

2. Reversal stability. There are many components in the cluster based on horizontal measurements. For example, the Hadoop cluster of Yahoo has more than 42,000 machines( can be used to look at the size of the cluster of different organizations). This means that some of these machines fail. Information processing methods that take into account the existence of such errors and should work without any serious consequences.

3. The location of the data. In large distributed systems, data is distributed to a large number of machines. If the data is located on the same server and processed on the other - the data transmission value may exceed the cost of processing it. Therefore, BigData solutions principle to settle one of the most important principle for the principles of information, if possible, to protect them in accordance with the purpose of the data processing machine [3].

The whole modern way of working with big information is based on these three principles. Let's look at the models that work with these principles.

MapReduce
MapReduce is a distributed data processing model offered by Google to process large amounts of data in computer clusters.

MapReduce is clearly defined in the following figure:
Figure 1. MapReduce model.

MapReduce model, the data is carried out in three stages.
1. Map step. At this stage, the map () function is used by the user to define the data. The task of this phase is to process and filter data. In this case, each entry is assigned a special function. The map () function applies to a single entry and creates a set of key-value pairs.
2. Shuffle step. This step does not appear to the user. The map () function at this stage is divided into "baskets". These "baskets" are then the input for the Reduce () step.
3. Reduce () step. Each "basket" with the value generated during the Shuffle transition goes to Reduce() step. The Reduce () step is marked by the user and calculates the final result for a single basket. A set of all the values returned by the Reduce () function is the final result of the MapReduce function.

3. Some additional info about MapReduce
The functionality of the function begins independently and parallelly operates with different terminal machines.
Shuffle represents parallel sorting, so it can work on multiple terminal devices.
Allows 1-3 positions to perform the horizontal-sized principle.[6]
The concept of "data science" includes all the methods of data warehouse design and processing of numbered data. "Data Science" is a modern contributor to a considerable amount of business information. Today, the e-government system is a widespread system all around the world, including Uzbekistan. This system envisages the use of modern information and communication technologies in the provision of public services to citizens, business entities and public authorities. We will consider some approaches to processing large volumes of data.[5]

4. Combining models (Combination)
The possibility of increasing the speed but not limited to, optimization of the performance of the base can be made by combining a variety of models. It is known that the processing speed depends on the complexity of the mathematical apparatus used. Analyzing how a simple mechanism is used, so the data can be analyzed quickly.[4]
Initially, the most common algorithms are used. Some of these algorithms use these algorithms as inefficient than using complex algorithms, and the data is sent to the next sophisticated algorithm. The
remaining data will also be sent to the next processing stage so that the chain of processing algorithms will be formed. Eventually, the most complex algorithms and data will be processed, but the analysis of the data will be slightly smaller than the size of the first prototype. As a result, the overall amount of data is considerably reduced.

Let's look at a practical example of this approach. In analyzing the problem, the RIS analysis allows for predicting the stability of the demand for various commodities. The products of the X group are sold consistently, so using the algorithms of prognosis can provide a good quality forecast. Y group products are sold less stable, and for each of them there is no need to build a model, which provides us with time-savvy algorithm and job forecasting. Z group products are randomly sold, so they do not normally need to build a prognostic model, they have simple computational formulas, such as average monthly sales. According to statistics, 70% of the products are in the Z group. About 25% of the product is in Group Y, and 5% of the product is X group. Thus, the construction and implementation of complex models is about 30% of the product. Therefore, the use of the above approach will reduce the time of analysis and forecasting time by 5 to 10 times.

5. Parallel processing
Another effective strategy for Big Data refinement is the separation of data into segments and the creation of models for each segment, with a more integrated look at the results. Most of the large amounts of data different from each other. Therefore, it is desirable to group the data separately into separate groups. In this case, instead of creating a complex model for everyone, we can build a simpler model for each segment. This approach enables us to increase the speed of analysis and reduce memory demand in the analysis process. Moreover, in this case, analytical processing will have a positive effect on the time it takes to parallelize the process. In addition, analysts can create models for each segment [3].

This approach, in addition to speed improvements, has another advantage. At the same time, the results can be summarized in a short time by grouping large volumes of data and gradually analyzing these groups on small models.[2]
6. Refined selection
The modeling for Big Data is sufficient for a small set of reusable, but not all of its data. Properly selected sample will contain the information needed to create a quality model. The analytical process is divided into two parts: build models and work with new built-in data. Creating a complex model is a process that requires resources. Working with new data from a built-in model requires substantial resources.

7. Conclusions
The following conclusions were made in this article. These are:

1. Horizontal measurements. Big data processing for large data processing container of any possible expansion of the system. The size of the data doubles, the number of components in the cluster doubles, and the process continues.

2. Reversible stability. There are many components in the cluster based on horizontal measurements. For example, in the Yahooning Hadoop cluster, there are more than 42,000 vehicles available (available for cluster sizes of different organizations). This means that some of these cars fail. Larger data handling techniques should take into account such errors and should work without any serious consequences.

3. The location of the data. In large distributed systems, data is distributed to a large number of machines. If the data is located on the same server and processed on the other - the data transmission value may exceed the cost of processing it. Therefore, one of the most important principles for Big Data solutions is to re-process data in a machine that stores information, if possible.
These three principles apply to all modern models. Based on these principles, the MapReduce model works in a distributed way. In this model, the location of the information is primarily based on the location. If the information location principle does not work, the transmission of data will be more expensive than processing data.

In summary, we can say that parallel algorithms are more efficient in processing data. This technology greatly reduces processing time and seamless use of memory by processing segments at great lengths of processing information. Models combine the technology is much more complex process, but reduces the processing time of 5-10 times. In the recycled selection technology, the selected information is provided and the part that is being processed is processed. This will increase processing capacity.

References
[1] Franks B 2012 Taming the Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics p 45
[2] Gantz J 2013 The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East (United States)
[3] Afzali G and Mohammadi S 2016 Privacy Preserving Big Data Mining (Association Rule Hiding)
[4] Kachalov D L, Mishustin A V and Farkhadov M P 2018 Modern Big Data Processing Techniques in Large-Scale Systems (Institute of Management Problems, RAS named after V.A. Trapeznikova)
[5] Cuzzocrea A, Song I and Davis K C 2011 Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! Proceedings of the ACM International Workshop on Data Warehousing and OLAP pp 101-4
[6] 2012 Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making (Capgemini Reports) pp 1–24
[7] Elgendy N 2013 Big Data Analytics in Support of the Decision Making Process (German University in Cairo) p 164
[8] 2012 EMC: Data Science and Big Data Analytics EMC Education Services pp 1–508
[9] He Y, Lee R, Huai Y, Shao Z, Jain N, Zhang X and Xu Z 2011 RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems IEEE International Conference on Data Engineering (ICDE) pp 1199–208
[10] Herodotou H, Lim H, Luo G, Borisov N, Dong L, Cetin F B and Babu S 2011 Starfish: A Self-tuning System for Big Data Analytics Proceedings of the Conference on Innovative Data Systems Research pp 261–72
[11] Kubick W R 2012 Big Data, Information and Meaning Clinical Trial Insights pp 26–8
[12] Lee R, Luo T, Huai Y, Wang F, He Y and Zhang X 2011 Ysmart: Yet Another SQL-to-MapReduce Translator IEEE International Conference on Distributed Computing Systems (ICDCS) pp 25–36
[13] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C and Byers A H 2011 Big Data: The Next Frontier for Innovation, Competition, and Productivity McKinsey Global Institute Reports pp 1–156
[14] Mouthami K, Devi K N and Bhaskaran V M 2013 Sentiment Analysis and Classification Based on Textual Reviews International Conference on Information Communication and Embedded Systems (ICICES) pp 271–6
[15] Plattner H and Zeier A 2011 In-Memory Data Management: An Inflection Point for Enterprise Applications (Springer, Heidelberg)
[16] Russom P 2011 Big Data Analytics TDWI Best Practices Report pp 1–40
[17] Sanchez D, Martin-Bautista M J, Blanco I and Torre C 2008 Text Knowledge Mining: An
Alternative to Text Data Mining *IEEE International Conference on Data Mining Workshops* pp 664–72

[18] Serrat O 2009 Social Network Analysis *Knowledge Network Solutions* **28** 1–4

[19] Shen Z, Wei J, Sundaresan N and Ma K L 2012 Visual Analysis of Massive Web Session Data *Large Data Analysis and Visualization* (LDAV) pp 65–72