INTERNATIONAL JOURNAL ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 27th Feb 2021. Link : https://ijiemr.org/downloads/Volume-10/Special

DOI: 10.48047/IJIEMR/V10/I03/59

Title: BIG DATA ANALYSIS PROBLEMS

Volume 10, Issue 03, Pages: 258-260.

Paper Authors

Elmurad Satimbaevich Babadjanov[1], Khamdam Bazarbaevich Kenjaev[2],

Aydos Turdimuratovich Atamuratov[3], Shaxsanem Turdimuratovna Atamuratova[4]

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.o

BIG DATA ANALYSIS PROBLEMS

Elmurad Satimbaevich Babadjanov[1], Khamdam Bazarbaevich Kenjaev[2],

Aydos Turdimuratovich Atamuratov[3], Shaxsanem Turdimuratovna Atamuratova[4]

[1]Nukus branch of Tashkent University of Information Technologies, PhD of the Department of Information Security, Nukus

[2]Nukus branch of Tashkent University of Information Technologies, senior teacher of the Department of Computer systems, Nukus

[3]Nukus branch of Tashkent University of Information Technologies, of the Department of Information Learning Technologies, Nukus

[4]Nukus branch of Tashkent University of Information Technologies, Student of Faculty of Telecommunication Technologies and Vocational Education, Nukus

elmurbes@gmail.com, k_xamdam@mail.ru, tatunfatm@umail.uz,

Abstract. There has been studied the process of big data processing in the article, which examines the collection of information for big data, big data systems, methods of their analysis, and the main problems in the analysis of big data and the causes of its occurrence are discussed. Then the steps of data collection are considered. The last section deals with the problems of data analysis in big data.

Keywords. Database, information, data collection, processing, big data, operational and analytical processing of data, SAP Khan method, Greenplum Chorus method, Aster Data nCluster method, MapReduce, classic B-tree method.

I. Introduction.

Today, the approaches and methods developed in the creation of information storage technologies for Big Data analysis are used, and the principles of data collection, processing and analysis are taken into account when making adjustments to them on the basis of quantitative indicators. However, some features of traditional operations may conflict with Big Data processing features.

Significant differences in operational and analytical data processing tasks began to manifest themselves at the beginning of the development of database technology. The term database was proposed by Bill Inmon as early as the 1970s, but the increase in interest in these technologies only occurred 20 years later, firstly, there was a real need for such systems, and secondly, they were necessary to increase.

The database processing cycle consists of collecting, cleaning, uploading, analyzing data, and finally presenting the results of the analysis. It would be illogical to go into detail about these steps, but the basic thesis needs to be clearly defined - if you try to use information storage technology to analyze Big Data, then you are not only looking at analysis algorithms, but all stages of data processing. we also need to pay attention.

DATA COLLECTION.

The information corresponding to the information system is taken from the operational database, converted into the required form, verified and then uploaded to the system. The recorded operations are performed with a certain periodicity, and here the question arises: when working with Big Data, is such a “periodicity” always possible and should it be as fast as possible for analysis? The time interval between the appearance of information and its availability for analysis may be less than the time required to perform information storage operations. An example of such a task is to monitor social networks to detect negative statements or to detect the loss of confidential information. These events should be identified and neutralized as soon as possible. However, we are dealing with an informal presentation of the information here, which first requires the development of algorithms for the production of undefined text at high speed in the first stage.

In the initial operation of a database (e.g., when searching for data), previously collected content can be used, which means that it is difficult to work with Big Data. The problem is primarily with their constant distribution, which is not easy to analyze, but easy to collect - for example, if we are talking about telecommunication systems, ma 'data "added" on regional servers (Figure 1). In terms of analysis, it is easier to distribute data over time rather than regionally (each server is responsible for a specific time), and so on. But again, the data will only be available after it has been transferred to the servers needed for analysis.

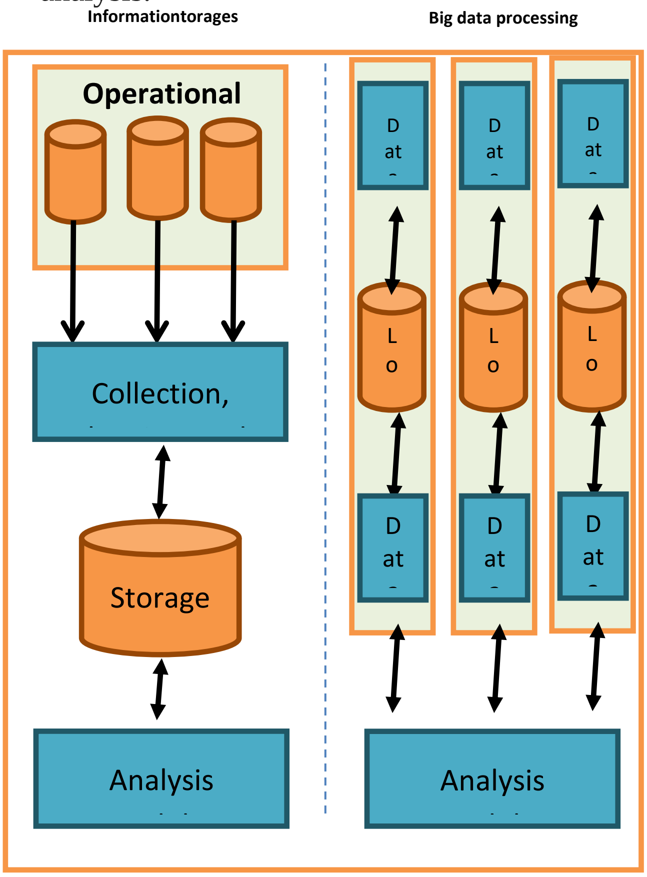

Figure. 1. Data collection

Thus, in traditional data storage systems, all data always passes through a single logical block that is responsible for converting, controlling, cleaning, loading, and performing these operations. However, large-scale data processing cannot form such a unit. It should be noted that to date, there is not much work with a fast input data flow, but as a logically integrated system, it is possible to implement a block for collection, cleaning, conversion, and loading.

ANALYSIS

Traditionally, databases provide approximately the same set of data analysis tools: multidimensional analysis (OLAP), regression, classification, clustering, and search. Today, there are also software products in this area - for example, SAP Khan, Greenplum Chorus, Aster Data nCluster, which allow you to work in Big Data. To understand the potential of such solutions, a review of the underlying algorithms, as well as an analysis of their large parallelization paths, is the key to processing large amounts of data. However, it is important to consider not only the basic parameters of the Big Data (intensity and volume of network exposure) characteristic of distributed distributed data processing technologies (e.g., MapReduce) but also all available methods for working with Big Data. -Used to manage multidimensional data from trees to complex structures.

CONCLUSION

Big Data creates characteristics that are not shared by traditional datasets. These features pose significant challenges for data analysis and motivate the development of new statistical methods. Unlike traditional datasets, where the sample size is usually larger than the dimension, Big Data is characterized by a huge sample size and high dimensionality. First, we discuss the impact of large sample sizes on understanding heterogeneity: on the one hand, large sample sizes allow us to uncover hidden patterns associated with small subpopulations and weak generality across the population. On the other hand, modeling the internal heterogeneity of Big Data requires more sophisticated statistical methods. Second, there are several unique high-dimensional phenomena, including noise accumulation, spurious correlation, and random endogeneity. These unique features make traditional statistical procedures invalid.

REFERENCES

1. Virdjiniya Andersen, Bazi dannix Microsoft Access. Problemы i resheniya: Prakt. posob. /Per. s angl.—M.: Izdatelstvo EKOM, 2001.—384 s.: ill.
2. Tomas Konnolli Karolin Begg, Bazi dannix, proektirovanie, realizatsiya i soprovojdeniya, teoriya i praktika, Universitet Peysli, Shotlandiya, izd. M.-SPB.- Kiev, 2003.
3. DeytK.Dj., Vvedenie v sistemi baz dannix.
4. Ulman Djeffri D., Djennifer Uidom, Vvedenie sistemi baz dannix. Per.sangl. M.: «Lori», 2000.
5. Meyer D., Teoriyarelyatsionnых baz dannix. Per.sangl. M.: Mir, 1987.
6. Kirillov V.V. Strukturizovanniy yazik zaprosov (SQL). – SPb.: ITMO, 1994. – 80 s.