The Impact of Big Data on the Choice of Used Storages and Modern Trends in Managing Big Data

Alexander, Bogdanov

Plekhanov Russian University of Economics Stremyanny lane, 36, Moscow, 117997, Russia, St.-Petersburg State University, Universitetskaya emb. 7/9, St.-Petersburg, Russia
e-mail: bogdanov@csa.ru

Irina, Ulitina

St.-Petersburg State University, Universitetskaya emb. 7/9, St.-Petersburg, Russia
e-mail: hg.ulitina@yandex.ru

ABSTRACT

Data lake is a technology that became very popular because Big Data is a part of modern reality. However, modern databases often do not fit the data that organizations want to store because of various reasons. When should people think about using data lakes? In what cases can data lakes be used, and when will their use bring some benefits to companies? In this article, we analyze how to use this technology, discuss the prospects and possible problems in the implementation of data lakes in practice.

Keywords

Big Data, data lakes, data warehouse

1. INTRODUCTION

Big Data technologies allow you to process a huge amount of information (which can be as large as hundreds of petabytes) in order to get new, useful information. However, it is worth noting that the volume of information that is processed is constantly growing. The demand for the development of technological solutions for monitoring and analyzing data from open sources is constantly growing too. Moreover, many scientists and researchers are sure that the growth in the amount of information will be exponential at the moment and in the following years as well. Even now people are faced with the problem of losing control over their data. So, the vector of development of technologies related to Big Data is aimed at solving the emerging issues. This is undoubtedly relevant since almost all companies produce and collect data.

However, we cannot take any advantage of the usual data storage. Even a traditional business is interested in using methods of working with information based on examples of working with Big Data approaches and using appropriate tools to produce results that could be used. Data Lakes is a technology that has found wide application in a number of large and medium-sized projects and now is one of the most popular.

In our opinion, the main problem is the uncertainty of the concept of Big Data and the lack of understanding of what tools are required to be applied in certain cases. We propose using theoretical developments in the field of formulating the mathematical basis for working with Big Data [1] and the corresponding concepts of the Big Data ecosystem for formulating approaches to storage systems.

2. THEORY

Selecting tools for working with large amounts of data is a separate task for the development team. Not infrequently, the architecture had to be extremely drastically changed because of increased data loads, and control of stored data was lost, and the collection of statistics became more and more difficult. There are many examples of such situations (such as Sberbank, Uber, Amazon, Google, Instagram, etc.). Often companies initially begin to use a variety of NoSQL solutions, specialized databases such as Vertica and others, but modern databases often do not fit for various reasons for data that organizations want to store. There was a need for a solution that allows not only to store all sorts of information with the ability to download from different sources but also has a set of tools to analyze the collected information (Big Panda, Informatica, etc.).

A data lake is a concept, an architectural approach to centralized storage that allows you to store all structured and unstructured data (from mobile applications, IoT devices, and social networks) with the possibility of unlimited scaling [2].

According to the Markets and Markets forecast, by 2021, the lakes market will grow to 8.81 billion dollars with an annual growth rate of 28.3 [3].

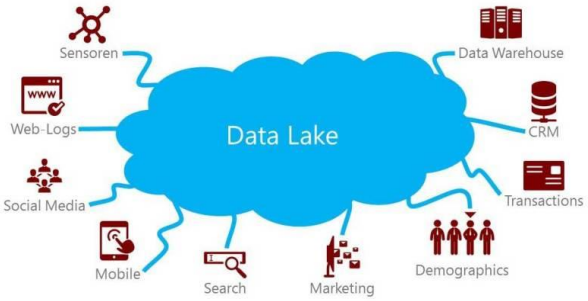

Figure 1. Different sources of Big Data [4]

Usually, among the reasons for using data lakes are the following [5]:

- The need to have all materials in case of verification (this is especially important for the financial sector, reconciliation of transactions, accounts, archiving history, etc.)
- The potential value of data in the future (data on clicks and views, which so far there is no need to analyze)
- Requirements of the law (for example, a law that was adopted in Russia for all mobile operators)
- The desire to use more advanced and complex analytical methods used in the analysis of information
- The desire to make traditional activities such as data access and search speed more efficient
- The need to respond quickly to peaks of activity

A data lake can store structured data from relational databases (rows and columns), semi-structured data (CSV, journals, XML, JSON), unstructured data (emails, documents, PDF files) and binary data (images, audio, video) [6] (Fig.1). Quite popular is the approach, in which incoming data is converted into metadata. This allows you to store data in its original state, without special architecture or the need to know which questions you may need to answer in the future, without the need to structure the data and have various types of analytics - from dashboards and visualizations to big data processing, real-time analytics and machine learning to make the right decisions.

The concept of data lakes in some cases may be replaced by analytical databases such as Vertica. Nowadays databases have more obvious business applications than data lakes, although they are far from being mutually exclusive. Unlike a database that relies on structural markers such as file types, the data lake provides data that can be moved between processes and is readable by various programs. Storage costs for this type of data management setup are usually lower than for databases.

Data lakes can be located on the servers of the company or in the cloud storage. The storage and administration of data lakes today are carried out by specialized firms: Hortonworks, Google, Oracle, Microsoft, Teradata, Cloudera, Zaloni, HVR, Podium Data, Snowflake, Amazon, and etc. Most companies offer not only storage capacity, but also tools for structuring lakes and processing data.
Cloud services can lower the barriers to implementing these big data processing platforms by eliminating the need to build, configure, and manage local infrastructure, which speeds development. The shortest possible time to market is often the driving force for the introduction of cloud services with Big Data for start-ups and developing industries, which allows them to move from concept to production without having to design, buy, configure and support all the required infrastructure [7].

However, if data lakes are used without proper skills, an unpleasant situation can occur when it becomes difficult to get the necessary data among all the collected data. As a result, control over the data in the lake data will be lost (turning the data lake into a so-called swamp [8]).

Several disadvantages of data lakes are needed to be taken into account when designing architecture. First of all, this is a security issue because the technology is relatively new, and all employees, as a rule, have access to data, and the degree of protection of lakes is low. However, if operations that are large and complex or that expect significant growth requires a flexible, scalable data storage solution, it probably means using a data lake.

In addition, it should be borne in mind that the data lake is a system that collects data from a variety of sources, including connected production equipment, delivery systems, customer reviews, sales data, forecasting algorithms, and even social networking channels. Thanks to the correct analytical software, it provides significant value in the form of real-time analysis of data. Data lakes are more flexible and accessible to a wide range of users and technology platforms but they also, by their very nature, induce to store everything and later sort their usefulness. Business analysts can use data lakes to create automated reports and produce analytical information for digital dashboards.

3. SUGGESTIONS FOR DEVELOPING NEW APPROACH

Despite all advantages, the study of the integrating data lakes revealed that in practice many companies face many difficulties when trying to create their own data warehouses based on the data lake concept. Usually, organizations start by using the existing data storage and then work with a given fixed architecture, which also delivers a number of inconveniences.

As a result of the study, it was found that the problems that arise are associated with many factors, among which the most popular are the following:

- Not enough good equipment available
- Lack of specialists with experience in deploying data lakes
- Difficulties in attempting to create a repository on equipment that is specially prepared by the company providing the solution
- Outdated approaches in existing implementations, limitations, performance gaps
- The complexity of making changes to the repository, the rigidity in the architecture, lack of flexibility
- High costs

After analyzing the existing solutions, we came to the conclusion that a universal solution that could be deployed on the most diverse equipment that would be easy to install and serving is needed. The solution should be developed in the open-source format of the project because it will allow involving in the development and refinement of the solution many specialists who are also interested in developing a product that would be profitable and easy to use. In order to be successful, such a data lake should have certain mechanisms for collecting, storing, cataloging, protecting and analyzing data as well as integration with different systems and be able to exchange results. Of course, one of the most important issues is the question of the data protection algorithm will be implemented in such a repository because the implementation of the project in open-source implies a complete putting the code on public display and refinement. However, this problem is easy to solve if you create a solution in the format of a plug-in external library in which all the necessary data encryption protocols and the required algorithms are enclosed. At the same time, it is necessary to develop a solution that would increase the reliability of data encryption but there would be no delays in working with data at the level of processing the stored data.

A data lake, as a rule, consists of a series of standard blocks [2], [9]. As a result of the analysis of existing solutions, the following functional modules were identified that are the most necessary and need to be developed in the universal solution:

- Storage for all data with the ability to create separate storage for hot/cold data, for ever-changing data or to handle fast streaming
- Security module
- Databases for structured data
- The module of tools for working with data (analysis, data engines, dashboards, etc.)
- Machine learning module
- Services for the development of add-ons, modifications and deployment of storage

We believe that the creation of such a solution, which could be deployed on the largest possible variety of equipment as well as with the possibility of using different equipment

(including different companies, with different settings, etc.), would increase the use of data lakes technology. We are sure that many companies for which the organization of their own data lakes is an inaccessible option will be able to keep up with the times and create their own repositories using the minimum amount of funds.

4. CONCLUSION

This paper concentrated on data lakes technology as a necessary part of any Big Data corporate infrastructure. We feel that current use of Big Data technology, based mostly on Hadoop, cannot cover more than 60% of applications. In many cases it is useless if not unsuitable. Approach proposed in [1], [10] makes it possible to build an ecosystem of software tools to process any kind of Big Data. And thus, we can try to build a universal storage solution working with any types of data. We are focused on the idea of developing an open-source solution that would be available for everyone who wants to organize his own data warehouse without spending a lot of effort, time and money.

We proposed the concept of a universal data lake with a description of the basic necessary functionality to solve the previously mentioned problems. This is one of the most important tasks, which should be solved in the near future because of the growth of information that needs to be processed every day. We believe that this approach will reduce the cost of creating an ecosystem for a wide variety of projects and will increase the growth of usage of data lakes in the development of various levels.

The individual components of the proposed solution were developed as part of the Big Data project at St- Petersburg State University and will be presented in a more detailed publication.

REFERENCES

[1] A. Bogdanov, A. Degtyarev, V. Korkhov, T. Kyaw, N. Shchegoleva, "Big data as the future of information technology", *Proceedings of the VIII International Conference "Distributed Computing and Gridtechnologies in Science and Education" (GRID 2018), Dubna, Moscow region, Russia, September 10 - 14, 2018,* pp 26 – 31, 2018.

[2] White Paper: The Compelling Advantages of a Cloud Data Lake, https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/ESG-White-Paper-AWS-Apr-2017+(FINAL).pdf

[3] Data Lakes Market by Software (Data Discovery, Data Integration, Data Lakes Analytics, Data Visualization), Business Functions (Marketing, Sales, Operations, Finance, HR), Service, Deployment, Organization Size, Vertical and Region - Global Forecast to 2021, https://www.marketsandmarkets.com/Market-Reports/data-lakes-market-213787749.html

[4] Introduction To Data Lake Part 1 What is Data lake, http://dwbimaster.com/introduction-to-data-lake-part-1-what-is-data-lake

[5] Angling for Insights in Today's Data Lake, https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

[6] The enterprise data lake: Better integration and deeper analytics, https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

[7] The Cloud-Based Approach to Achieving Business Value From Big Data, https://d0.awsstatic.com/analyst-reports/

[8] Blog atricle about data lakes, https://aboutdata.ru/2017/06/01/data-lakes/

[9] IBM. The Journey Continues From Data Lake to Data-Driven Organization, http://www.redbooks.ibm.com/redpapers/pdfs/redp5486.pdf

[10] I. Gankevich, Y. Tipikin, V. Korkhov, V. Gaiduchok, A. Degtyarev, and A. Bogdanov. "Factory: Master Node High-Availability for Big Data Applications and Beyond", *ICCSA 2016, Part II, LNCS 9787, Springer International Publishing Switzerland 2016,* pp. 379–389, 2016.