- Hortonworks DataFlow (HDF) – a virtual machine with a full set of tools for stream data processing;
- Cloudera Distribution including Apache Hadoop (CDH) [4] – an open source software distribution containing Apache Hadoop and key components such as Apache Flume, Apache Hive, Apache Kafka, etc.;
- MapR Converged Data Platform [5] is a single platform, implemented on a single code base, combining key technologies: distributed file system, multi-model NoSQL database, publish/subscribe streaming event engine, ANSI SQL and a wide range of open source data analysis technologies;
- Microsoft HDInsight [6] is a service running in the Windows Azure cloud that allows you to quickly launch such popular open source platforms as Apache Hadoop, Spark and Kafka;
- Arenadata Hadoop (ADH) [7] is also a Russian distribution, which includes current stable versions of all the most popular tools, such as Apache Hive, Apache Spark and Apache Atlas.

Conclusion

Deploying big data storage and analytics systems presents a complex technical and engineering challenge. To facilitate the process of deploying storage systems and big data analysis, special distributions can used in the form of virtual machines or cloud services.

References

1. Cielen D., Meysman A., Ali M. Introducing data science: big data, machine learning, and more, using Python tools. – Manning Publications Co., 2016.
2. Sawant N., Shah H. Big Data Application Architecture //Big data Application Architecture Q & A. – Apress, Berkeley, CA, 2013. – p. 9-28.
3. Get Started with Hortonworks Sandbox
4. Cloudera CDH.
5. MapR Converged Data Platform.
6. HDInsight. A simple, cost-effective, open-source, enterprise-grade analytics service.
7. Бородаенко В., Ермаков А. Универсальная платформа обработки больших данных / Виктор Бородаенко, Александр Ермаков // «Открытые системы. СУБД» 2017, № 03

BIG DATA ANALYTIC AND FORECASTING.

Mustafayeva Seving
Associate professor of Computer Engineering department

ABSTRACT

The main goal of solutions in this area is the distribution of data storage and processing. Today, there are a huge number of architectural solutions and tools used for big data. The technology for storing and analyzing big data is promising based on forecast analysis. Big data characterized by various characteristics, referred to as “Vs”. Analysis of the literature shows that today the most important characteristic of big data is data heterogeneity. It is the analysis of heterogeneous data that can give tangible results when modeling data. Big data is a powerful tool that helps firms advance, improve bottom lines, and refine decision-making processes. This impact clearly demonstrated by the big data statistics and trends discussed in this article.

Key words: big data, modeling, heterogeneous data

Introduction

Development of modern society and technology at the present stage is inextricably linked not only with the informatization of new regions any activity related to widespread implementation of technology study and analysis of data to develop competent management personnel solutions.

Among the tasks for which big data is used are the following:

- creation of a global competitive infrastructure transmission, processing and storage of data;
- ensuring information security during transmission, processing and storage data that guarantees the protection of the interests of individuals, businesses and states;
- creation of end-to-end digital technologies;
- implementation of digital technologies and platform solutions in spheres of public administration and provision of public services, including in the interests of the population and small and medium-sized businesses, including individual entrepreneurs;
- transformation of priority sectors of the economy and social areas including healthcare, education, industry, agriculture, construction, urban management, transport and energy infrastructure, financial services through the introduction of digital technologies and platform solutions.

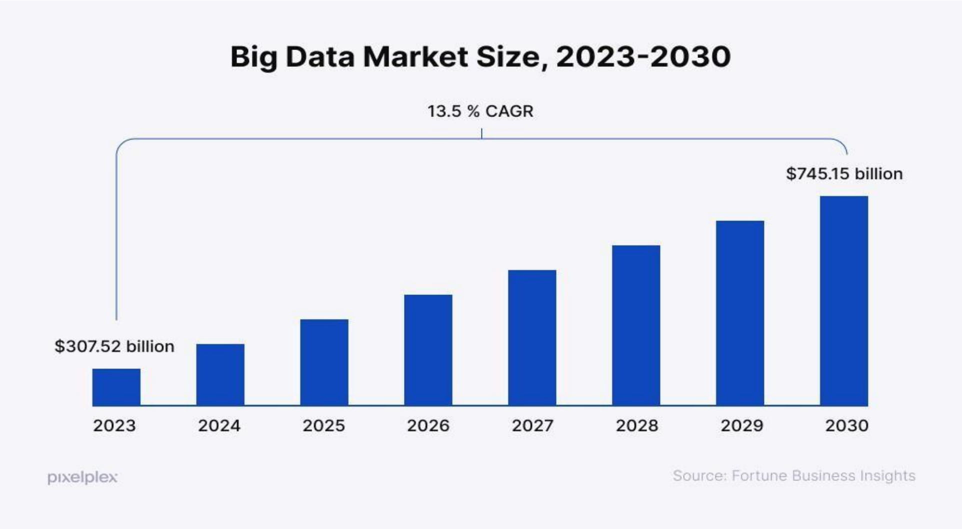

Figure 1. Big Data Analytics Market 2023-2030

The platform paradigm proposed by IDC involves the development several platforms, the first of which is computer systems on mainframe based. The first platform originated from the late 1950s and some of its elements still exist today. The second platform based on a client-server architecture and began in mid-1980s, when PCs connected to databases and applications mainframes. It still exists today.

In January 2016, The Economist suggested the following: “The third platform is based on online cloud computing and interaction with all kinds of devices, including wireless ones such as smartphones, hardware and sensors (collectively known as the Internet of Things”) [1].

It is also worth noting that a fourth platform is sometimes mentioned [2]. This platform is characterized by the active implementation of smart technologies, IoT, massively distributed grid computing and, possibly, quantum computing. Thus, we can conclude that technologies for obtaining, storing and analyzing big data are an integral part of modern technology platforms.

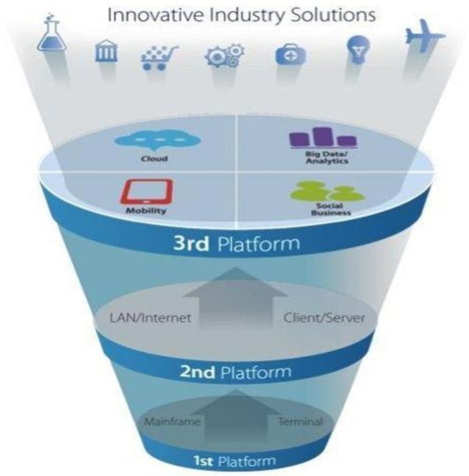

Figure 2. IDC's third technology platform

The third platform includes social Internet technologies, mobile, cloud solutions, big data technologies and partly the Internet of Things (IoT) was allocated at the beginning of 2010 and is developing to this day.

In January 2016, The Economist suggested the following: “The third platform is based on online cloud computing and interaction with all kinds of devices, including wireless ones such as smartphones, hardware and sensors (collectively known as the Internet of Things”) [1].

It is also worth noting that a fourth platform is sometimes mentioned [49]. This platform characterized by the active implementation of smart technologies, IoT, massively distributed grid computing and, possibly, quantum computing. Thus, we can conclude that technologies for obtaining, storing and analyzing big data are an integral part of modern technology platforms. According to an Accenture study [3], 79% of business executives agree that companies that do not use big data will lose their competitive position and may face extinction. Moreover, 83% of executives have implemented Big Data projects to increase the company's competitive advantage in the market.

Big data technologies actively take advantage of intelligent technologies. Thus, 59% of executives say that Big Data in their company will be improved through the use of AI. Sales and Marketing, Research and Development (R&D), Supply Chain Management (SCM) including Distribution, Workforce Management locations and operations is where advanced analytics, including big data, make the biggest contribution to revenue growth today. McKinsey's research [4] provides a comprehensive overview of how analytics and big data are enabling the creation of entirely new ecosystems that serve as the foundational technology for artificial intelligence. (AI). McKinsey believes that analytics and big data make the most valuable contributions to the basic materials and high technology industries. Nearly 50% of respondents to a McKinsey Analytics survey believe that analytics and big data have revolutionized the way sales and marketing do business.

According to NewVantage Venture Partners [5], big data brings the greatest benefits to enterprises by reducing costs (49.2%) and creating new opportunities for innovation (44.3%). 69.4% of enterprises have started using big data to create data-driven business processes. The Hadoop and Big Data market projected to grow from $17.1 billion in 2017 to $99.31 billion in 2022, achieving a 28.5% CAGR . The largest period of projected growth is in 2021 and 2022, when the market projected to grow by $30 billion within one year [6].

According to the forecast [7], big data applications and analytics will grow from $5.3 billion in 2018 to $19.4 billion in 2026, achieving a CAGR of 15.49%. The global big data market, which includes professional services, will grow from $16.5 billion in 2018 to $21.3 billion in 2026. Comparing the global demand for modern data analytics technologies and big data-related hardware, services and software, the dominance of the latter category becomes clear. The software segment is projected to grow faster total relative to other categories from $14 billion in 2018 to $46 billion in 2027 , reaching a CAGR of 12.6% [7].

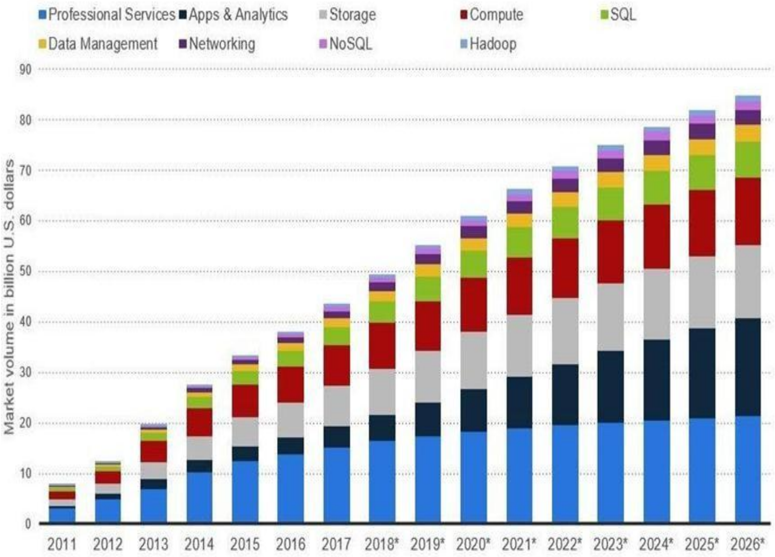

Figure 3. Big Data Application Market Forecast by Software Segment

If we consider the term “big data” directly, then the characteristic of a large volume of data is not fundamental, as it is other aspects of big data that determine the essence of the new technology. Big data is the new technological stage. in the field of Data Science, so big data should be distinguished from earlier technologies, such as business analytics. The Data Science process in relation to big data can be presented in a sequence of six stages), which are different, for example, from business analytics [8]:

1. Determining the purpose of the study.
2. Data collection stage.
3. Data preparation stage.
4. Data research stage.
5. Data modeling stage.
6. Display and automation stage.

At the first stage, a project assignment is prepared, which defines the subject of the research, the expected result from the implementation of big data technology, what data will be used, what resources will be involved. The assignment also provides a description of the output results.

At the second stage, data directly collected in accordance with the design assignment. Therefore, the assignment specifies data sources. At this stage, an analysis of data availability and quality carried out. As noted earlier, data may exist in internal and/or external sources. The data obtained in the second stage may contain various errors, so the third stage improves the quality of the data, preparing it for further use. At this stage the following phases can be applied:

- cleaning;
- integration;
- transformation.

During the cleaning phase, the algorithms eliminate incorrect values and possibly reconcile data between sources (if the same data from different sources has different values). During the integration phase, as one would expect, information obtained from several sources combined. Finally, the transformation phase converts the data into a format suitable for use in subsequent stages.

The fourth stage - data research - is aimed at clarifying the characteristics of the data, such as distribution, correlation, presence of outliers, clusters, etc. At this stage, methods of descriptive statistics and data visualization methods are actively used. In [8], this stage is called "exploratory data analysis" (EDA).

Data modeling, the fifth stage, serves to build a data model, that is, this stage directly answers the questions of the purpose of the study. This stage is often iterative, in which the researcher tries out certain sets of models and determines their characteristics. At this stage, methods of statistics, operations research, machine learning, etc. are used.

Finally, at the last, sixth stage, the data is usually visualized for presentation to the customer. If the developed technology is needed by the customer more than once, then an automatic application is developed based on the obtained data models and applied algorithms.

Conclusion

The real process of big data analysis, due to its complexity and multifactorial nature, often not performed in a sequential manner - the researcher often has to return to previous steps and make adjustments. Therefore the actual process is iterative.

References

1. Tech pundits' tenuous but intriguing prognostications about 2016 and beyond.The Economist. https://www.economist.com/business/2015/12/31/tech-pundits-tenuous-but- intriguing-prognostications-about-2016-and-beyond
2. Third platform. From Wikipedia, the free encyclopedia.
3. 3. Kukartsev, V. V. Database theory [Electronic resource]: textbook / V. V. Kukartsev, R. Yu. Tsarev, O. A. Antamoshkin. — Electron. text data. - Krasnoyarsk: Siberian Federal University, 2017. - 180 p. - 978-5-7638-3621-9.
4. Kiseleva T.V. Software engineering. Part 1 [Electronic resource]: textbook / T. V. Kiseleva. - Stavropol: North Caucasus Federal University, 2017. - 137 p.
5. Big Data Executive Survey 2017. Executive Summary of Findings.
6. Global Market Research Reports Company. Statistic MRC
7. Home – Wikibon research.
8. Cielen D., Meysman A., Ali M. Introducing data science: big data, machine learning, and more, using Python tools. – Manning Publications Co., 2016.