Analysis of Industrial Software Solutions for Data Processing and Storage

Mikhail Konovalov

System Analyst of Raiffeisenbank JSC, Moscow, Russia
Advisor of Russian Academy of Natural History
konov328@gmail.com

Abstract. *This article covers three technological trends of data processing and storage, which definitely became the most popular in the modern market of development of industrial information technologies – databases, data warehouses and big data technology. Key properties, architectural features and main differences, thanks to which a decision is made in favour of one or another option at the initial stage of the project are considered. The main features, purposes and problems usually encountered when developing solutions using these technologies are analysed. Issues of effective organization of the process of collecting and processing large volumes of the different data types are disclosed as well as criteria by which information relevant to different types of processing can be attributed, since not all data can be used for analytics.*

Also reviewed the newest markets trends and most popular customer needs as well the most advanced data processing facilities and the main reasons why these technological solutions most likely to be in increasing demand in the future.

Keywords: *information system, software, data processing, data storage, database, data warehouse, big bata, IS, DWH, analysis, normalization.*

Introduction

In the modern peace of information technologies development and the growing diversity of systems offered in today's market, it is extremely important to carefully choose technologies, as well as specific solutions and manufacturers, analyzing the needs of the enterprise, the architectural landscape, integration features and the budget. To a large extent project success and sometimes development of the entire company depend on this stage.

This article is concerned with three technological trends: databases, data warehouses (DWH) and big data.

The main functions of the database are the ability to quickly save new data or find an existing record, change and save or delete it. Databases are designed for operational work with a relatively small amount of data.

DWH are designed to store and process large arrays of data. Systems of this class are used, for example, in accounting, processing of information archives, telephone networks, banking operations, etc.

Big data is one of the most advanced technologies for working with large amounts of information that rapidly gains popularity. This term began to be used and quickly became popular only 9-10 years ago.

The variety of high-tech solutions for processing and storing various kinds of data is growing steadily, and it is difficult to imagine an industrial information system without them. Meanwhile, each technology has its advantages and peculiar features, and various difficulties and niceties, which are also covered in this article, are found in the course of development and operation of such systems.

Software solutions for data processing

One of the main task in developing a system for storing and processing large amounts of data is efficient organization of these processes taking into account the needs of the enterprise. Data shall be easily retrieved, modified, deleted. At the same time it is necessary to remember about scalability so that in the future, when the needs of the company grow, there will be no need to incur additional costs or completely abandon the current solution in favour of a new one.

DWH are used to solve problems related to the need to select significant amounts of data and convert it for further research and analysis. The author of the DWH concept - Ralph Kimball - described a DWH as “a place where people can access their data”. Another, more detailed definition of a DWH can be as follows: a DWH is a subject-oriented, time-related, uncorrected, integrated set of data collected from other systems to present information to users, conduct a statistical or business analysis, provide reports and make strategic and tactical decisions in a company. The term "subject-oriented" means that data is combined into categories and stored in accordance with the areas it describes, but not with the applications it uses. The term “integrated” means that data is combined so that it meets all the requirements of the enterprise as a whole, but not the only one function of the business.

It may also be necessary to process absolutely different information from a variety of sources: documents, blogs, social networks, any customer data, or even information about actions performed by customers, information from measuring devices, etc. But this is mainly text information. In addition, processing of images, audio and video data, etc. may be required.

Such data can be processed using big data technology. Here is the definition of this technology. Big data is a series of approaches and methods for the processing of a large amount and a large variety of data that is difficult to process using conventional methods. The objective of big data processing is to get new information. At the same time the data can be both processed (structured) and isolated (unstructured).

But not all data is subject to processing using this technology. There are criteria by which information relevant to such processing can be attributed, since not all data can be used for analytics. The key concept of big data is based on these defining characteristics, the so-called 3Vs:

Volume. Data is measured by the physical volume of the “document”, which is subject to analysis.

Velocity. Data is not static in its development, but is constantly growing. Therefore, the meaning of this characteristic is not only in rapidly growing data volumes, but also in the need for their quick processing to obtain the required results.

Variety. Formats of data can differ. It means that the data can be fragmented, structured, unstructured, or partially structured. And the point is to process different types of data simultaneously.

In addition to the considered 3Vs, the fourth V is added in some sources. Veracity. And even the fifth V - viability or value. Some sources mention 7Vs, but 3Vs is enough for the basic understanding.

Features and problems of industrial data processing

Classic databases are designed for operational work with a relatively small amount of data, but often there is a need to select significant amounts of data and convert them for further research and analysis. To perform of such task requires a different architecture and data organization. For this purposes there are data warehouses (DWH).

Since DWHs are designed to support decision-making, specific requirements discussed below are imposed on them.

Integration. Due to a large number of different sources, the same data, parameters can be stored in different ways and have different formats and values. Such inconsistencies should be eliminated by software automatically. The data shall be processed and unified, so it could meet the requirements of the entire enterprise. This can be one of the most time-consuming tasks when designing a DWH.

Subject orientation. I.e. it is required to upload the range of information reduced as much as possible to a DWH and use only the data that is necessary to solve the problem.

Time reference. Data in a DWH is never deleted, but is stored for 5-10 years or more. It is necessary to identify regularities and develop forecasts. There is the latest version of the object/parameter record for each time point. The data is not modified, since it can result in violation of its integrity.

Support of data internal consistency. This requirement results from the preceding ones, since multiple data sources and a denormalized structure can threaten consistency within a DWH, but it should not be allowed. There are special mechanisms ensuring data consistency.

It is required to ensure fast extraction of large amounts of data. It should represent an environment that is optimized in such a way as to quickly obtain ready-made cuts or data arrays from very large volumes, while performing complex, arbitrary, non-standardized queries, individual for the organization, department or even analyst. So, it is necessary to abandon the main principle - normalization, i.e. splitting tables so that each value is found in a DWH only once. So, a DWH is denormalized, and the same value can be met both in a detailed form and in an aggregated form.

Based on the definition of big data, we can formulate the following three basic principles of working with such data.

Horizontal scalability. Since data volumes are constantly growing and there can be as much information as you like, a system that presupposes processing of such data must be scalable. For example, if the amount of data increases twice, then it should be possible to increase the capacity of hardware twice per cluster, and the system will continue to work without loss of performance.

Data locality. Data is stored on a large number of machines in large distributed systems. But if data is physically located on one server and processed on another one, then the resources required for data transfer may exceed expenses for its processing. Therefore, when designing solutions for big data, one of the most important principles is the principle of data locality, the essence of which is that the data is processed and stored on the same machine.

Fault tolerance. The principle of horizontal scalability discussed above implies that there can be many machines in a cluster. For example, Yahoo cluster consists of more than 40,000 machines. In this case it is assumed that some of these machines will fail on a regular basis. Methods of working with big data should take into account the likelihood of such failures and maintain system performance without any significant consequences.

Advanced data processing software solutions

A logical question arises in the process of researching modern data processing and storage technologies. What is the purpose of methods and approaches called big data, what is unique in them and how is it possible to use the information processed using the technologies under consideration?

Firstly, ordinary databases and DWHs cannot store and process such huge amounts of data (hundreds and thousands of terabytes). And it is not about analytics, it is about data storage.

For example, a database is designed for fast processing of relatively small amounts of data or a stream of small records. This main task is solved with the help of big data - successful storage and processing of large amounts of data.

Secondly, diverse information is structured in big data. It comes from various sources (images, photos, videos, audio, and text documents) to a single, understandable and acceptable for further work type.

Thirdly, analytics is developed and accurate forecasts are made based on the received and processed information in big data.

As a result, having full understanding of your company and business, including statistical information and numbers, detailed data on competitors, new and detailed information about your customers will allow to succeed in attracting new customers, significantly increase the level of services provided to current customers, understand the market and competitors better and, therefore, get ahead by dominating them.

One of the most obvious and popular examples to date, which is described in many sources on the Internet, is associated with Apple that collects data about its users using manufactured devices: phones, tablets, watches, and computers. It is the availability of such a system that allows the corporation to own a huge amount of information about its users and use it to get profit. And there are a lot of such examples today.

Given the above results that big data allows achieving, it is possible to explain the desire of companies trying to conquer the market to invest in these modern methods of data processing today to get advantages over competitors and reduce costs tomorrow.

Conclusion

Complex information systems such as industrial databases or DWHs allow solving entire ranges of tasks and provide various services - from conducting client transactions to planning, forecasting and making tactical and strategic decisions at the level of the largest enterprises. Companies are willing to spend huge amounts of money on the development of the IT infrastructure realizing the importance of this component in improving the efficiency, making a profit, increasing competitiveness in the market. Popularity and scalability of such solutions are growing steadily and it is already difficult to imagine an industrial information system without them.

Meanwhile, regardless of the specifics and industry, the value and importance of information about potential and current customers of the company, direct competitors and upcoming trends in the market is becoming increasingly apparent. These conditions are becoming necessary to maintain competition in the modern and dynamic world. In connection

with it and judging by already existing examples of successfull introduction of big data, it can be assumed that the importance of these technologies will continue increasing in the future. Today, the world's largest companies, which hold leading positions in various areas of business, invest billions of US dollars in the development of this area.

Under conditions of severe competition, undoubtedly, companies that know and understand needs of their customers better will have advantages and, therefore, will be able to offer the most relevant and suitable solutions and products.

References

1. BIG DATA Academy: Introduction in analytics of large data arrays: Information // National Open University INTUIT. URL: https://www.intuit.ru/studies/courses/12385/1181/info (Reeference date: November 29, 2018).
2. Ralph Kimball, Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Third Edition. — John Wiley & Sons, 2013.
3. Data Warehousing and Big Data // ORACLE. URL: http://www.oracle.com/technetwork/database/bi-datawarehousing/overview/index.html (Reference date: November 29, 2018)
4. Viktor Mayer-Schönberger, Kenneth Cukier. Big Data. A Revolution That Will Transform How We Live, Work, and Think, 2014.
5. Analytical Review of Big Data Market // Habr. URL: https://habr.com/company/moex/blog/256747/ (Reference date: November 29, 2018).