DATABASE ECOSYSTEM IS THE WAY TO DATA LAKES

A.V. Bogdanov [1,2], N.L. Shchegoleva [1,2], I.V. Ulitina [1]

[1] *Saint Petersburg State University, 7/9 Universitetskayanab.,St. Petersburg 199034, Russia*

[2] *Plekhanov Russian University of Economics, Stremyannylane, 36, Moscow 117997, Russia*

E-mail: n.shchegoleva@spbu.ru

The paper examines the existing solutions design of various data warehouses. The main trends in the development of technologies are identified. An analysis of existing big data classifications allowed us to offer our own measure to determine the category of data. On its basis, a new classification of big data has been proposed (taking into account the CAP theorem). A description of the characteristics of the data for each class is given. The developed big data classification is aimed at solving the problems of selecting tools for the development of an ecosystem. The practical significance of the results obtained is shown by the example of determining the type of big data of actual information systems.

Keywords: Big Data, Data API, Big Data Ecosystem, Data Lake Concept,

Alexander Bogdanov, Nadezhda Shchegoleva, Irina Ulitina

1. Introduction

There are two globally intensively developed approaches (Data Lake and Big Data). They have been discussed for so long that they have become phenomena, which proves their importance and significance for the further development of modern information technologies. This is due to the fact that nowadays is characterized by an "overabundance" of information. The main task facing humanity now is to learn how to process data quickly and safely. The choice of tools for working with an extremely large amount of data (Big Data) of different categories and from different sources is a very difficult task that arises every time you create a new system. Often this is a research task and we can evaluate the effectiveness of its solution only after a long time, when new requirements and new tasks arise that the stack of selected technologies may not cope with.

Big Data Landscape help to understand how many tools and frameworks are available now. If we compare 2012 and 2017, then a rapid increase in the number of technologies is noticeable. They developed on the principle of integration with artificial intelligence and many tools appeared that were created specifically to solve new business problems. 2018 is characterized by the further development and implementation of data processing technologies. Since companies interested in managed data.

Of the current trends, it should be noted: companies that are faced with new volumes of data begin to implement their solutions at the local level, which means that the same problem is solved repeatedly; large companies began to support the development of data ecosystems (data platforms) and tools for building such systems; many existing ecosystems have limited functionality.

Nowadays, a wide range of different Big Data technologies has been developed for Databases, Data Management, Data Integration (ETL Solutions), Business Intelligence, Data Mining, etc. Many of the listed technologies are well developed and have great functionality. Some technologies are available under open source licenses. However, there are very few works devoted to API This is a serious problem, because for an effective solution to the problem using Big Data, it is necessary not only to choose a certain set of tools from a huge amount, but also to ensure their interaction among themselves.

The importance of Data API is determined by the fact that it should provide a business capability delivered over the Internet to internal or external consumers: network accessible function; available using standard web protocols; well-defined interfaces; access for third-parties. The key features of Data API are management tools that make it actively advertised and subscribe-able; available with SLAs; secured, authenticated, authorized and protected; monitored and monetized with analytics. So API is unified approach to data integration of conventional APIs (Web, Web Services, REST API – not built for analytics), databases (SQL, NoSQL, ODBS and JDBC connectors), database Metaphor + API = Data API.

2. New specification of Big Data types

In work [1] it was shown that the solution to this problem is should be based on new specification of Big Data types. This article proposes a method for determining the types of Big Data, the formation of ecosystems (software stacks) for different types of data, and substantiates the Data Lake concept.

Let's consider in more detail the data itself. According to the CAP theorem, they can be divided into 6 classes (fig. 1), however only 5 classes out of 6 potential are possible because PA-class cannot exist by itself and modern corporate architectures are distributed - that is, divided - by default. Then we have the following data classes.

– C - class (consistency) It is characterized by data that: agreed - this is a guarantee that simultaneous reading from different places will return the same value; that is, the system does not return outdated or conflicting data; stored in one place (usually); may not have backups (there is too much data to do backup for them);often analytical data with a short life span.

– A - class (availability) It is characterized by data that: should always be available; can be stored in different places; have at least one backup or at least one other storage location; are important data, but do not require significant scaling.

– CA – class: data must be consistent and accessible; potentially a monolithic system, without the possibility of scaling or scaling under the condition of instant exchange of information about the changed data between the master-slave nodes; there is no resistance to distribution, if scaling is provided for (branches), then each branch works with a relatively independent database.

In this case, the CA class is divided into 3 subclasses:

1). Big data of large sizes that cannot be represented in a structured way or they are too large (stored in Data Lake or Data Warehouse): data has any format and extension (text, video, audio, images, archives, documents, maps, etc.); whole data collected, the so-called "raw data"; large data that is unreasonable to place in the database (unstructured data in the case of data warehouses); multidimensional data.

Medical data that cannot be stored in tabular form (x-ray, MRI, DNA, etc.) are the example of this type.

2). Data of a specific format that can be represented in a structured form (biological data, DNA and protein sequences, data on a three-dimensional structure, complete genomes, etc.) characterized by multidimensional data; data must be analyzed and their sizes reach gigantic values. Medical and bioinformatics data that need to be searched and stored in a relational table with extensions of xml, json, etc. are the example of this type.

3). Other data well presented in relational databases witch: have a clear structure or can be represented in the concept of a relational database; the size of the stored data does not matter (provided that lightweight objects or links to large objects are stored in the storage); transactional required;

"Raw" data may be, but is not recommended (an exception - if the logs are stored), customer data, logs, clicks, weather statistics or business analytics, personal data, rarely updated, customer base, etc. are the example of this type.

– CP – class It is characterized by data that : must be consistent and at the same time there is support for the distributed state of the system, which has the potential for scaling; structured, but can easily change their structure; must be presented in a slightly different format (graph, document), that is, data for social networks, geographic data and any other data that can be presented in the form of a graph; have a complex structure, because of which there is a potential need for storing files in a document-oriented format; they accumulate very quickly, so a distribution mechanism is needed; no permanent availability requirements.

Frequently recorded, rarely read statistics, as well as temporary data (web sessions, locks, or short-term statistics) stored in a temporary data store or cache are the example of this type.

– PA – class It is characterized by data that: should be available and at the same time there is high support for the distributed state of the system, which has the potential for scaling; have a complex structure, the potential need to store files in a different format with the ability to change the scheme without the need to transfer all the data to a new scheme (Cassandra); accumulate quickly.

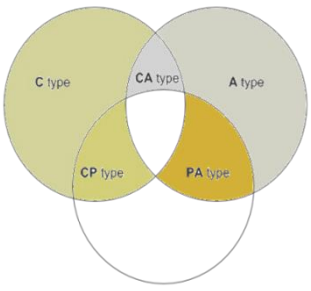

Figure 1. Graphical representation of Big Data classes

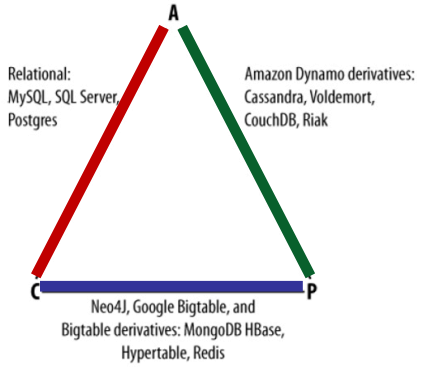

Figure 2. Database Type Classification [2]

This class is suitable for data that is historical in nature. The main task here is to store large amounts of data with the potential growth of this information every day, statistical and other processing of information online and offline in order to obtain certain information (for example, about the interests of users, mood in conversations, to identify trends and etc.)

Before determining the type of system, we must estimate the total system parameters (maximum number of users for simultaneous operation, the ability to scale services, the availability of personalized access), evaluate the project (having its own server capacity, cost comparison with the cost of building rental of services), evaluate time data access, query performance evaluation for cloud infrastructures, construct the automatic allocation system and send requests in a distributed database.

3. Classification of systems for working with Big Data

For Classification of systems for working with Big Data we can use CAP theorem. In accordance with it, 3 classes of systems can be distinguished:

- CA - System provides high available consistency.
- CP - The system provides strong alignment with the separation tolerance.
- PA - The system assumes full availability, weakened consistency.

The results of the study of various databases can be summarized in the form of the following schemes, which show which databases can be attributed to which types are shown on fig. 2 [2].

To summarize the opinions of experts, the ideal Big Data solution should combine the technical specifications: scalability, fault tolerance, high availability, the ability to work with data in a wide access, but protected, support for analytics, data science and content applications, support for automation of data processing workflows, integration with popular solutions, self-healing ability.

In addition, when developing an ecosystem, the composition should include tools that provide: Data collection; Data storage: the ability to organize heterogeneous data storage, providing real-time access to data, data storage using distributed storage (including the organization of data lakes); Data research: data extraction, formation of a complete picture of data, data collaboration; Data management: creating directories, providing access control, data quality control and building information processing chains; Data production: building a data infrastructure, data delivery, integration issues. The most complete technological equipment should contain the components: Data Source Layer: Streaming – relational databases (RDBMS), social networks, web services, sensors, etc.; Batch – data received from the database and file system; Ingestion layer – priority is given to data, they acquire category and sent to the Data Storage Layer, where they are stored in a very different form (depending on the type of storage selected); Analytics Layer – basic data processing. Management Components – data cataloging, data processing chain building and auditing; Exploratory Environment – the implemented functionality is tested, experiments are being conducted (storage replacement, data processing, data enrichment using additional data from the main storage); Data from this level may be published on Publishing Layer – level of data presentation. Using a special API, access to the representation of this data is formed – Data Access Layer.

An algorithm for testing work of Big Data applications consist of the following steps:

Step 1. Data Staging Validation consists in: data from various sources like RDBMS, weblogs etc. should be validated to make sure that correct data is pulled into system; comparing source data with the data pushed into the Hadoop system to make sure they match; verify the right data is extracted and loaded into the correct HDFS location.

Step 2. "Map Reduce» Validation provides: Map Reduce process works correctly; data aggregation or segregation rules are implemented on the data; key value pairs are generated; validating the data after Map Reduce process.

Step 3. Output Validation Phase intended for: checking the transformation rules are correctly applied; checking the data integrity and successful data load into the target system; checking that there is no data corruption by comparing the target data with the HDFS file system data.

Step 4. Architecture and Performance Testing: data ingestion and throughput and data processing sub-component performance.

4. New technologies for working with Big Data

Data Warehouse and Data Lake can be used to store data. Data Warehouse - is a database optimized for analyzing relational data; its structure is predefined in order to optimize it for fast SQL queries. Data is cleared, enriched and transformed. However, this is no longer enough even for a traditional business that is interested in using methods of working with information based on examples

of working with big data approaches and using appropriate tools to obtain results that can be used. Data Lakes is a technology that has been widely used in a number of large and medium-sized projects and is currently one of the most popular. Data Lake - this is the concept of a centralized storage that allows you to store structured data from relational databases (rows and columns), semi-structured data (CSV, magazines, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video) in its original state and with unlimited scalability.

Data Lake is a more flexible technology. However, with the incorrect use of the capabilities of the Data Lake, it will be difficult to find the necessary data, as a result of which control over the data will be lost and the Data Lake will turn into a "swamp".

Despite all the advantages, a number of problems interfere with the development of the Data Lake. The article [3] lists 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. The authors of the article believe that to create ecosystems of software tools for processing any big data, it is necessary to use the approach described in [4, 5].

The proposed universal solution concept assumes the following modules: 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.

Conclusion

Thus by using the Data API you can select technologies and form stack. Then make connection between technologies, develop an interface for convenient user work and form an effective ecosystem for each data class. Together, these ecosystems form a universal data platform, which, in essence, is a Data Lake.

The classification of Big Data proposed in the article, the tool stack formed for data processing, taking into account all their characteristics on its basis combined in Data API, as well as the use of the concept of a universal Data Lake will create a ecosystems for convenient user work with for all data types in accordance with the proposed classification. Together, these ecosystems form a universal data platform, which, in essence, is a Data Lake. If tasks of convenient and efficient Data API and the Big Data Ecosystem are solved, then two globally intensively developed approaches – Data Lakes and Big Data (both so far, to a large extent, remain concepts) will naturally merge.

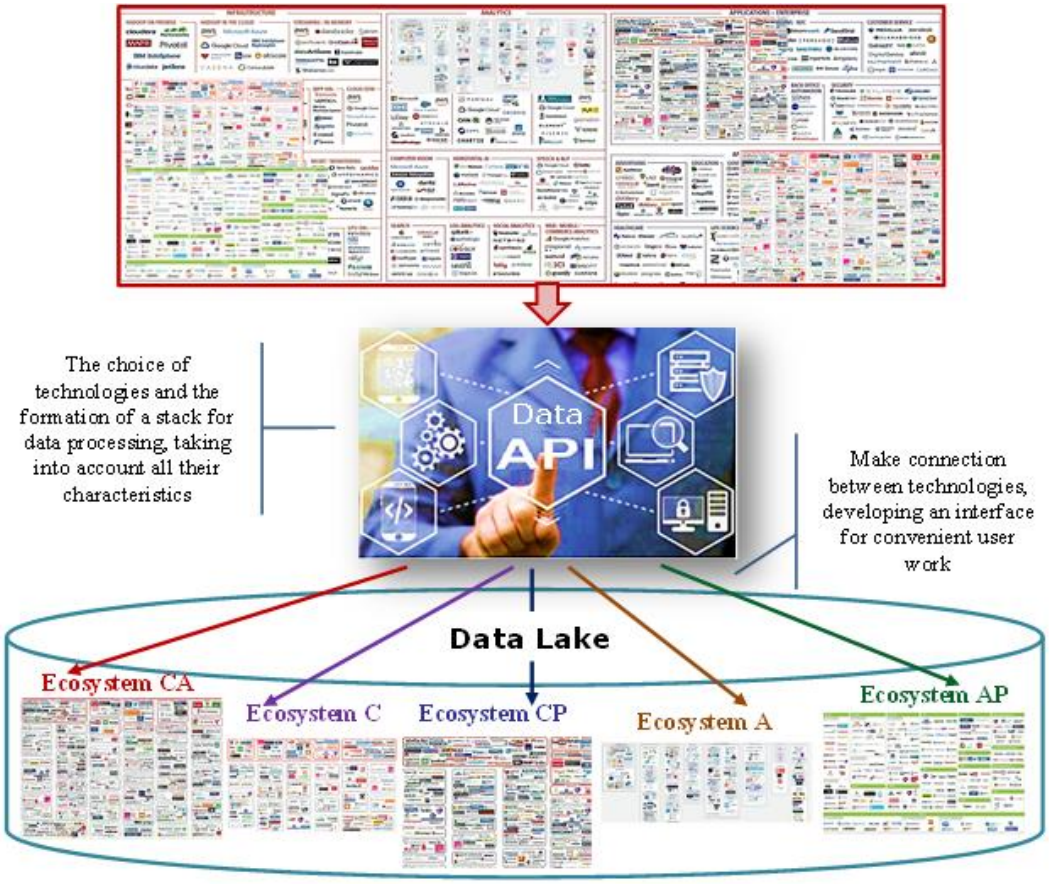

Figure 4. Data Lake concept

References

[1] Bogdanov, A. Degtyarev, V. Korkhov, T. Kyaw, N. Shchegoleva (2019). Is the Big Data the future of information technologies? In J. Busa, M. Hnatic, & P. Kopcansky (Eds.), *The 20th Small Triangle Meeting on theoretical physics* (pp. 15-28). Kosice, Slovakia: Institute of Experimental Physics, Slovak Academy of Sciences.

[2] Всё, что вы не знали о CAP теореме // habr.com URL: https://habr.com/en/post/328792/

[3] Alexander Bogdanov, Irina Ulitina. The Impact of big data on the choice of used storage. -CSIT, proceedings of the conference, September 23 – 27, 2019, Erevan, pp. 57 – 64.

[4] 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.

[5] 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.