BIG DATA AS THE FUTURE OF INFORMATION TECHNOLOGY

A. Bogdanov [a], A. Degtyarev, V. Korkhov, T. Kyaw, N. Shchegoleva

St. Petersburg State University 13B Universitetskaya Emb., St. Petersburg 199034, Russia

E-mail: [a] a.v.bogdanov@spbu.ru

Currently, the problem of "Big Data" is one of the most, if not the most urgent in computer science. Its solution implies the possibility of processing uncorrelated and heterogeneous data of large volume, the implementation of their integration from distributed sources by consolidation or federalization methods and ensuring the security of access and storage of these data. Only the creation of technology that provides processing and storage of dissimilar, uncorrelated data of large volume can be considered a breakthrough result corresponding to the world level. To effectively address these issues, a new definition of this concept is proposed, namely, "Big Data" is characterized by the situation when the conditions for implementing the CAP theorem are relevant. The CAP theorem is a heuristic statement that in any realization of distributed computations it is impossible to provide the following three properties: Consistency, Availability and Partition Tolerance. Thus, depending on which of the properties cannot be implemented, we are dealing with different types of "Big Data". And this, in turn, means that a standard approach based on the MapReduce concept has a limited scope of applicability. Various possibilities for implementing data processing in different cases are discussed, and a conclusion is made about the need to create an ecosystem of "Big Data".

Keywords: Big Data, CAP theorem, Big Data types, Data API, Hadoop, data validation

1. Introduction

There is no metric for Big Data at the moment. We need to introduce it to understand what kinds of Big Data exist and how to treat these different kinds of Big Data. Only the creation of technology that provides processing and storage of dissimilar, uncorrelated data of large volume can be considered a breakthrough result corresponding to the world level. To effectively address these issues, a new definition of this concept is proposed, namely, “Big Data” is characterized by the situation when the conditions for implementing the CAP theorem are relevant. We will also briefly discuss the possible ways to deal with the CAP related issues in distributed Database system.

The Big Data concept itself means not just large data layers, among them the huge stored and processed arrays with hundreds of gigabytes, and even petabytes of data. Data can be processed and some useful information can be extracted from it. So, we can define BigData as a collection of information processing technologies to obtain information. Important to note, that the volumes of data processed through BigData are constantly growing, as well as the speed of its processing. At the moment, BigData is not only the giants of the IT world, but also systems faced by anyone working in the IT field.

2. Metric based on the CAP theorem

The CAP theorem claims [1] that in distributed system we cannot achieve Consistency, Availability and Partition tolerance at the same time, only two out of three. This can be used as a definition of BIG DATA, as based on applicability of CAP conditions to the data; different kinds of Big Data are defined by appropriate combinations of C, A and P.

We introduce the CAP theorem for a distributed database system project as a process in which with growing of data volume, complexity and system dimensions the condition of CAP theorem is realized.

The CAP theorem states that at any given point in time, a distributed database (system) does not guarantee the existence of all such components as Consistency, Availability and Partition Tolerance.

The diagram in Figure 1(a) clarifies that only two of the three attributes of the distributed system can be supported simultaneously. We found that CAP theorem diagrams of this type almost always missed illustrating two of the most important factors. One is that the intersection of all three attributes should never be available (Cross area), and second is the CAP theorem always refers to distributed networks.

The consequence of this definition is the fact, that pre- and post processing time grows substantially and thus the measure of BIG DATA can be estimated by the formula

$$T = \frac{t_{preproc} + t_{proc} + t_{postproc}}{t_{proc}},$$

$t_{preproc}$ – preprocessing time, t_{proc} – processing time, $t_{postproc}$ – postprocessing time.

3. Practical examples of Big Data types

Based on the analysis, the following types of Big Data with different functionalities were identified, correlated with various application areas and priority development areas [2]:

1. Big Data of the same nature, interconnected components, undergoing changes in the process of use, with the need to track and restore their previous state. A typical area of application is banking and finance, where the main task is to securely and quickly conduct transactions and track the flow of funds. Similar challenges arise in the insurance industry.

2. Big Data, characterized not so much by volume, as by heterogeneity and complex hierarchy. A typical representative of this class are industrial applications. As the most common task here, we consider the electronic history of a technical object: a ship, an aircraft, a helicopter; by which we mean

all the data about the object and the attached procedures from the idea / project of the product to its disposal.

3. Big Data, which feature is inaccuracy, vagueness and incompleteness. The main task in this class of data is to restore the gaps and obtain complete, interconnected domain objects. Such tasks are characteristic of poorly formalized areas, the humanitarian sphere and art. A typical representative of this class of problems is, for example, archeology.

4. Big Data that combines the diversity, incompleteness and lack of clear models in the subject area. This class, carrying on itself some features of the previous two classes, has very important specific features in terms of data mining and obtaining new knowledge about the subject area. Typical relatives of this class are medicine and biology, research in these areas, as well as the associated health care of the population.

5. Big Data requiring so much computing power for a one-time processing that can not be provided locally. This class of data arises when e.g. solving oil and gas production problems, when in order to receive a response and make technical decisions, it is necessary to simultaneously use very large data arrays that do not fit even in the theoretically possible supercomputer RAM. Such tasks can arise both against the background of complex data processing procedures and data omissions.

The central applied task of the research is to develop common approaches to the study and processing of financial, socio-economic, biomedical, archaeological, technical and geophysical systems using the methodology for processing extremely large amounts of data and their intellectual analysis to solve a number of applied problems.

Currently, social, economic, biomedical, geophysical and technical systems are becoming so complex that an avalanche-like growth of the diversity of presentation forms, their structural complexity and the amount of information associated with the processes occurring in them occurs. This circumstance qualitatively changes the requirements both to the approaches of analyzing such large-scale systems, and to creating decision-making tools on the effective influence on the processes occurring in them. Examples of such systems include partially state-regulated social institutions (labor market, system of higher and secondary vocational education), the banking sector, as well as the transport, municipal and social environment of megacities. Also, with the emergence of a global information and communication environment, qualitatively new properties are acquired by extremism, sociocultural and cyber threats. Big Data is a constant element of any medical activity. And in connection with the same problems, the relevance of their systematization, storage and processing becomes an extremely essential element of the development of the entire healthcare system. To solve the problems arising from the analysis of this class of systems, as well as for effective and timely decision-making, new scientific and technological approaches and technologies developed on their basis are needed.

Today, the direction associated with the intellectualization of data processing and analysis methods is developing intensively. Intelligent Data Analysis Systems (IDAS) are designed to minimize the efforts of the decision maker in the data analysis process, as well as in setting up analysis algorithms. Many IDASes allow not only solving classical decision-making problems, but also are able to identify causal relationships, hidden patterns in the system being analyzed.

4. Data API

To be able to work with BIG DATA it is necessary to have a set of tools, that we call ecosystem, out of which the most important is API (that in our case we prefer to call DataAPI). The formal definition is [3]:

API is a business capability delivered over the Internet to internal or external consumers

- Network accessible function
- Available using standard web protocols
- With well-defined interfaces
- Designed for access by third-parties

The key features of 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

And thus we can in unified manner build in our system:

- Conventional APIs: Web, Web Services, REST API – not built for analytics;
- Database paradigm: SQL, NoSQL, ODBS and JDBC connectors – familiar to analysts;
- Database Metaphor + API = Data API;
- Specific API for every type of Big Data (every "V" and their combinations) – under a generic paradigm.

That means that DataAPI can from one hand help with different data sets integration and from the other provide means for access and process the data from any location (Figure 1(b)).

5. Tools for different Big Data types

Today in the field of the information technology, the most popular model of work with Big Data is based on the MapReduce paradigm and Apache Hadoop project [4]. Most products for working with Big Data have highly effective system for processing huge amounts of information and its analytics in real time. The expected effect of the introduction of Big Data may vary depending on the type of activity and the policy enterprises [5] while working with large-scale data, knowledge manipulation methods, various methods of recognition theory and classification, methods of intelligence analysis and generalization of data, intellectual approaches in the form of genetic algorithms, neural networks and other branches of artificial intelligence.

The main tasks of the Hadoop platform are storage, processing and data management. The Hadoop platform allows reducing processing time and subbing data preparation, expands opportunities for analysis, allows new information and unstructured data.

There are several Hadoop distributions: Hortoworks, Cloudera, MapR, IBM BigInsights, etc. Hadoop is very popular and useful, including such IT giants as Facebook, Alibaba, Amazon, Linkedin, eBay. The reason lies primarily in the ability of Hadoop without preparation to accept and analyze huge data sets of different structures from a variety of sources, as well as in its performance and availability. In addition, Hadoop includes the HDFS file system, which can significantly reduce the cost of handling terabytes of storage.

In the case of the CAP theorem, Hadoop-based databases are referred as CP systems (consistent with partition tolerance). With data volumes stored redundantly across several slave nodes, outages of high portions (partitions) of a Hadoop cluster can be tolerated. Hadoop is refereed to be consistent since it has a centralized meta-data store which manages a single, consistent view of data stored in the cluster. We cannot say that Hadoop always guarantees availability, because if the NameNode itself fails, applications cannot access data in the cluster.

In other cases different instruments are needed. For example, for the AC case different variants of RDBMS systems are used, and standard grid tools are still very popular and combination of RDBMS and grid tools can be most effective in geographically distributed networks.

For the AP case mostly used tools are NoSQL DBs. Most popular ones out of them are Cassandra and CouchDB. For more effective data processing these DBs are often used together with some BigData engines, based on Apache Spark architecture, but not to loose the data one should carefully organize the work with hierarchial memory (Figure 1(c)). We feel that most effective way to do it is to use virtualization on all stages of building of distributed system.

6. Tests, to clear it up

It is still unclear, what is the critical value of parameter T, depending on architecture, types of data and access organization. So we feel that a lot of experiments should be made to make our approach work, in particular within the Virtual supercomputer paradigm developed by the authors [6,7]. The most important experiments are the following:

Step 1: Data Staging Validation

1. Data from various sources like RDBMS, weblogs etc. should be validated to make sure that correct data is pulled into system.

2. Comparing source data with the data pushed into the Hadoop system to make sure they match.
3. Verify the right data is extracted and loaded into the correct HDFS location

Step 2: “Map Reduce” Validation

1. Map Reduce process works correctly
2. Data aggregation or segregation rules are implemented on the data
3. Key value pairs are generated
4. Validating the data after Map Reduce process

Step 3: Output Validation Phase

5. To check the transformation rules are correctly applied
6. To check the data integrity and successful data load into the target system
7. To check that there is no data corruption by comparing the target data with the HDFS file system data.

Step 4: Architecture and Performance Testing

1. Data ingestion and throughput;
2. Data processing sub-component performance.

Another consequence of our approach is that a user before processing his data should, with a help of DataAPI, determine what type of Big Data he is dealing with and form proper software stack. To do this some of experiments can be of a great help:

1. Estimate the total system parameters (maximum number of users for simultaneous operation, the ability to scale services, the availability of personalized access).
2. Evaluate the project (having its own server capacity, cost comparison with the cost of building rental of services).
3. Evaluate time data access, query performance evaluation for cloud infrastructures.
4. Construct the automatic allocation system and send requests in a distributed database.

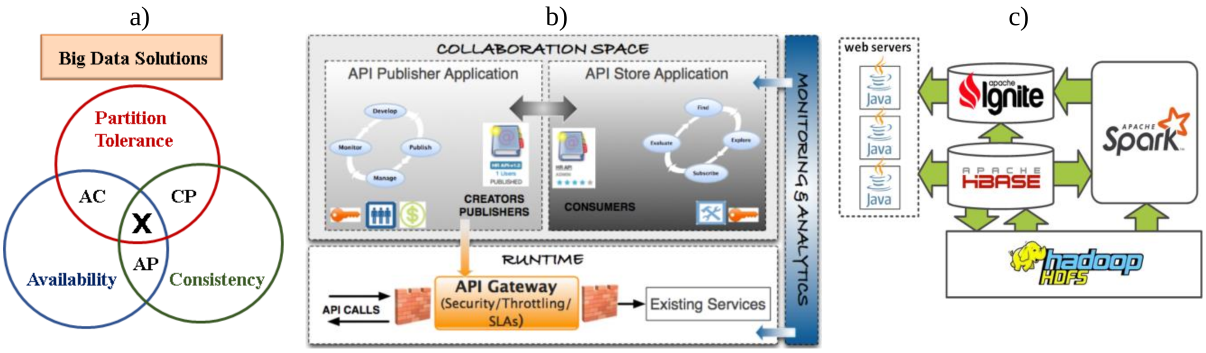

Figure1. CAP theorem diagrams

Acknowledgment

This work was partially supported by the grant of Saint Petersburg State University no. 26520170, RFBR 16-07-00886 and BGX company.

7. Conclusion

From a short review above we can see, that proposed definition works in a sense, that

1. We can make new specification of BigData types, that can help with a choice of tools to process them;
2. DataAPI is a substantial part of toolkit since it is an important engine both for integration with other systems and for access to data and to the results of their processing;
3. Future research in this field is definitely connected with building of BIG DATA Ecosystem, that will help to determine what type of data we are dealing with and what tools are proper for its effective processing;
4. Large amount of tests is still needed both for determination the limits of different data types and for choosing the proper software stacks for definite data types;
5. The proposed measure helps on the first stage of research but we feel it should be more detailed to cover the problems of matching of different data types to processing architecture.

References

[1] Eric A. Brewer. Towards robust distributed systems. In *Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing* (PODC '00). ACM, New York, NY, USA. 2000. DOI: 10.1145/343477.343502

[2] Yuri Demchenko, Cees de Laat, Peter Membrey. Defining architecture components of the Big Data Ecosystem. International Conference on Collaboration Technologies and Systems (CTS), 2014, pp 104-112. DOI: 10.1109/CTS.2014.6867550

[3] Chris Haddad. Six Tactics For Building Successful APIs, API Strategy & Practice Conference - WSO2, September 24-26, 2014, Chicago

[4] Apache Hadoop web site. URL: https://hadoop.apache.org/

[5] Thi Mai Le, Shu-Yi Liaw. Effects of Pros and Cons of Applying Big Data Analytics to Consumers' Responses in an E-Commerce Context. Sustainability. 9. 798. 10.3390/su9050798.

[6] Alexander Bogdanov, Alexander Degtyarev, Vladimir Korkhov. Desktop supercomputer: what can it do? Physics of Particles and Nuclei Letters, 2017, Volume 14, Issue 7, pp 985–992 DOI: 10.1134/S1547477117070032

[7] Alexander Bogdanov, Alexander Degtyarev, Vladimir Korkhov, Vladimir Gaiduchok, Ivan Gankevich. Virtual Supercomputer as basis of Scientific Computing, in series: Horizons in Computer Science Research, vol. 11, eds.: Thomas S. Clary, pp. 159-198, Nova Science Publishers, 2015, ISBN: 978-1-63482-499-6