ON THE WAY FROM VIRTUAL COMPUTING TO VIRTUAL DATA PROCESSING

A. Bogdanov[1,a], A. Degtyarev[1,2,b], N. Shchegoleva[1,c], V. Khvatov[3,d]

[1] *St. Petersburg University, 7–9, Universitetskaya emb., St. Petersburg, Russia, 199034*

[2] *RUE, SL CTABD, 36, Stremyanny lane, Moscow, Russia, 117997*

[3] *DGT Technologies AG., http://dgt.world/*

E-mail: [a] a.v.bogdanov@spbu.ru, [b] a.degtyarev@spbu.ru, [c] n.shchegoleva@spbu.ru, [d] valery.khvatov@gmail.com

Concept of a virtual personal supercomputer is proposed to solve the problems of distributed data processing. It boils down to data virtualization and a two-level processing system that allows data to be processed at their location. On this basis, a new computing platform DGT has been created, which can serve as the basis for the Big Data ecosystem.

Keywords: Big Data, virtual private supercomputer, data virtualization, kappa paradigm for data processing, DGT platform.

Alexander Bogdanov, Alexander Degtyarev, Nadezhda Shchegoleva, Valery Khvatov

1. Introduction

Efficient processing of large amounts of data has come to the fore in recent years. The main problem of distributed data processing is described by the CAP theorem (Brewer's theorem) [1]. It says that simultaneously when working with distributed data, only two of the three requirements can be satisfied: Consistency (data do not contradict each other in all computational nodes at the same time), Availability (any request to a distributed system finalizes with a correct response), Partition tolerance (the system continues to operate despite an arbitrary number of messages being dropped by the network between nodes). It is also possible to reformulate and extend the CAP theorem. It is called the PACELC theorem, according to which, in the case of network partition (P) in a distributed computer system, it is necessary to choose between availability (A) and consistency (C), but in any case, even if the system works normally in no separation, you have to choose between latency (L) and consistency (C).

The development of information technologies in the direction of finding a solution to this problem led to the emergence of the Data Lakes concept. This concept is that storage can be of different types, including portals, archives, storefronts, databases of different kinds, data clouds, and networks. These stores can have synchronous or asynchronous computer connections. Since the data type is often not known in advance, there is a need for a highly flexible storage system that allows you to easily switch between different sources and systems. However, moving away from some problems, the user ultimately encounters other, equally complex problems. First of all, with a low speed of work with data. However, no other way of working in a distributed environment has yet been proposed.

The intensive development of information technology determines the annual increase in data processing capabilities. At the same time, users of large computing and data centers are faced with the fact that the architecture with which they have to deal remains old. Permanent replacement of equipment requires both high financial resources and, equally important, very high personnel qualifications. A disaster occurs when new data types arise and the existing architecture is completely unattended to these challenges. However, you have to somehow use this user-pressured architecture to solve your problems.

2. Virtual private supercomputer

How to solve this problem in the current conditions? A similar question was raised by the authors in the field of computing more than ten years ago [2,3]. At that time, it seemed that there was an unsolvable problem that more and more computer power was required to solve complex problems, but they could be achieved in those conditions only by building a large cluster, distributed or hybrid system. All these solutions led to high network losses, which made it impossible to use a large number of nodes to solve almost any real problem in which there is at least a small interaction between processors. The way out was found in the concept of a personal virtual supercomputer [2], when all possible computer elements are virtualized: processors, memory, network, address space, etc. This approach did not allow, of course, to create a large universal supercomputer similar to the NEX SX, but made it possible to organize a high-performance virtual SMP system. Such system can solve any single complex computing problem basing on the standard computing equipment available to the researcher. Such approach takes possibility also to develop a methodology for restructuring the system to a specific task [4].

The idea of virtual private supercomputer allows you to circumvent the problem of dynamic load balancing, since you can work statically in a virtual environment. In addition, with this approach, it is possible to migrate the task to the data, and not vice versa, as is only possible with the MPP computing paradigm. Therefore, one possible solution for today's distributed data problem is to add a data virtualization element. Thus, combining the concept of a virtual private supercomputer with a classification of Big Data, taking into account various storage schemes, would solve this problem.

3. New paradigm of data processing

Let us consider how this approach can be applied to working with data. To do this, first of all, you need to understand that now researchers are forced to abandon the previously proposed paradigm of

working with data. The old λ-paradigm described monolithic systems based on the old architecture, when different data types could only be processed separately using their own instrumentation for each type, and then summarize all the results into a single source. The recent emergence of a large number of new data types has led to the transition to a new κ-architecture, when data processing is sent to an external level, and work with tools, and not with data. In practice, this allows you to send business analytics to the periphery of the system and receive only the result of data processing.

The proposed new paradigm is that we do not work with domains where data is located, but directly with data. What should actually be done for this?

1. You need to locate the data, you want to process
2. Process data where it is
3. Insert them into the required infrastructure

At the same time, it is not necessary to collect everything in a single center, as it is implemented in the classic example – a distributed database. It is not possible to perform a transaction without changing the system as a whole.

The idea is not to work with the system "as a whole," but to be able to change something somewhere locally and get only the result. That is, abandoning the Big Data concept, we allocate only small parts of the data that are required to solve the task for processing. They can be isolated and combined only on the basis of virtualization technology.

This paradigm moves from a "data lakes" to a mesh data network. In fact, the requirements for this network are dictated by the need to solve the following problems: when you virtualize a system, each operation collects only those parts of the data that are needed to perform the operation and eliminate the overwhelming amount of data that is not involved in this particular operation. This is similar to a virtual private supercomputer in computing [4].

This naturally leads us to the following data virtualization paradigm [5] (see Figure 1):

- the layer where the data are produced;
- the layer where the data are processed;
- the layer where the data are utilized.

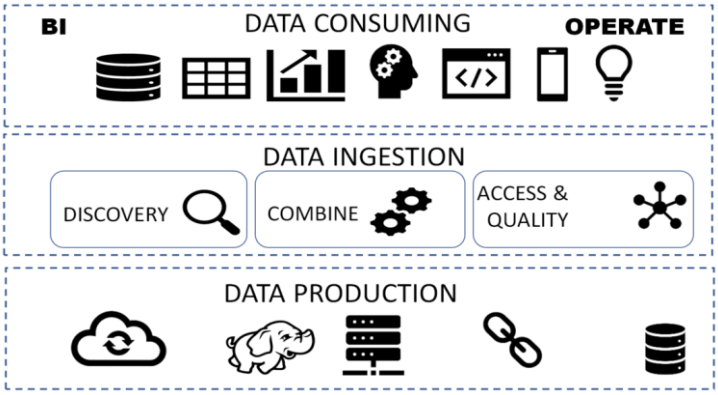

Fig.1. Data virtualization

In the framework of this approach, the user on the one hand can work only with a part of the data belonging to one huge array, and on the other hand he can "bypass" the requirements of the CAP (C - consistency, A - availability, P - partition tolerance) theorem, providing two (CA, AP or CP) of the three properties of the theorem for different data sets. This allows you to work with the selected part of the data as local data, and to apply different combinations of tools for its processing, which best correspond to their type (CA, AP or CP) [6]. Since data processing is done in according with the Data Lakes concept on the same server where they are located, it is not necessary to develop a data isolation technique, since they are not stored in the DBMS.

This approach is in line with the modern concept of Data marketplaces developed by leading corporations such as Amazon, Intel, etc. The structure of the Data marketplace [5] is shown in Figure 2.

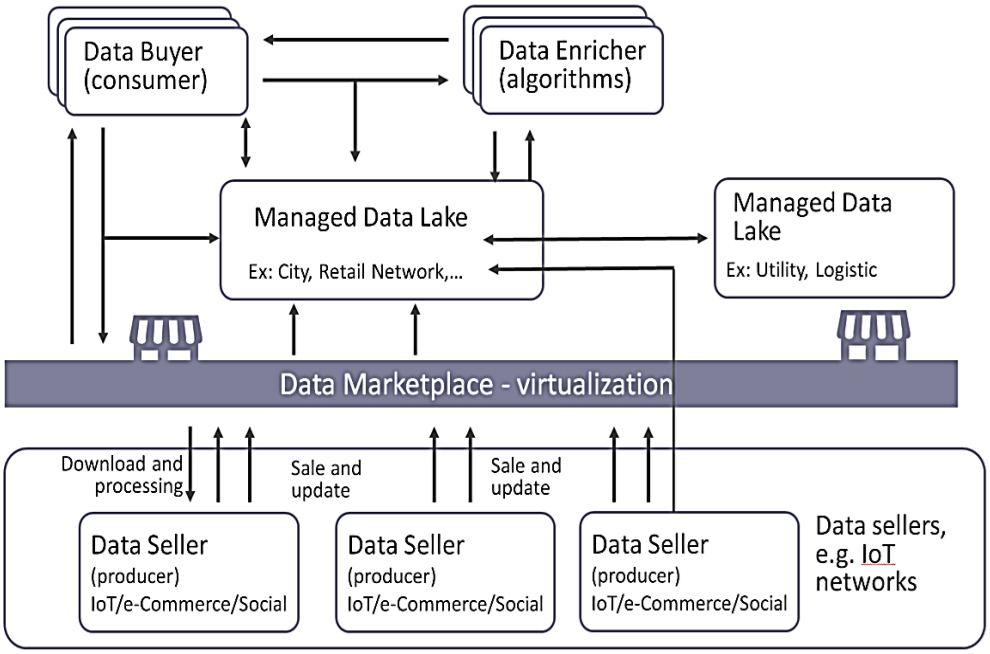

Fig.2. Data marketplaces structure

4. Decentralized storage platforms

A recent example of such a decentralized storage platform is the DGT Network [7], which forms a virtual data network, connecting different data sources across the boundaries of corporate information into a single analytic system accessed by authorized users in a way that ensures differentiated confidentiality. This platform is developed in accordance with *Byzantine tolerance protocol* [8] and this is the first in the world, and so far the only protocol supported by proven mathematical theorems.

DGT Network provides horizontal integration by creating separate clusters of enterprise-managed nodes that communicate through the secure F-BFT protocol and write them to a single ledger (Direct Acyclic Graph). Although this ledger serves as a "single source of truth" for its participants, differential anonymity protects the corporate confidentiality of the source data, while allowing analysts to provide relevant information to participants in real time.

In some ways, the competitor to the DGT Network is Data Marketplace, launched by the IOTA Foundation in 2017. IOTA has launched an open source distributed ledger that connects IoT devices to microtransactions processing in exchange for cryptocurrency. Unlike the blockchain, the distributed IOTA - IOTA Tangle ledger does not group transactions into blocks, as a typical blockchain. Instead of this it considers them as a stream of separate transactions connected together using a relatively simple network algorithm. For participation, a node must perform a small amount of computational work to verify the two previous transactions.

Since the Internet of Things does not require any consensus, the DGT Network has a significant advantage, allowing you to implement projects of different levels of complexity from cryptocurrencies to Internet banks.

Why is this so important right now? When working with Big Data, the central issue is data quality. If the user works in the field where the CAP theorem is true, then he must sacrifice something. When working with distributed data, data quality is a key issue.

Two critical trends guide the direction in which Big Data quality structures are growing: data decentralization and virtualization. The first trend illustrates the need to adapt distributed ledger technologies for data quality control, and the second shows the need to abandon verification according to a given data structure (since it can vary).

A two-tiered approach to data processing is therefore proposed:

- Pre-processing of incoming data with identification of main information objects and verification of their attributes;

- Handling quality attributes for all available data based on differences in transactional information versions.

This is a solution to the problem, since the data analysis task is much simpler. It means that it does not require large loads and more local. Division into two stages allows you to remove a large amount of unnecessary data from the process, which makes it possible to significantly reduce time costs and ensure a high level of parallelization. The principle in this approach is that the quality of the data does not suffer.

5. DGT Quality Framework

The basic strategy of DGT Quality Framework is based on allocating Master Data processing to a separate data processing type for a distributed environment [9]. In fact, master data refer to all static information that is used to identify critical elements of an organization and its business processes. Assigning inbound operational information to objects requires identification, as does generating consistent data sets for analysis. Therefore, master data, transactional data, and analytical data are interdependent and part of the same context. Errors and discrepancies in master data can cause the same or even greater damage than differences in transaction data.

Master data support the consistency of a common information array between different information systems, departments, and organizations. The most important characteristic of master data is the slow rate of change in information exchange between several participants. When working with master data, you can select the following management styles:

– on the basis of transactions;
– centralized master data;
– common master data.

The following information exchange characteristics shall be taken into account

– limitations of centralized solutions
– data access in real time
– smart data processing
– storage of logs

As part of the DGT Quality Framework approach, these challenges are solved through innovative technologies that provide fast decision-making and reduce data mismatch loss:

- The integration layer of the system is built on a high-performance DGT core, which ensures the formation of a single Master Data leger and its distribution among participants in information exchange.
- “Smart” modules that track data in real time and participate in the creation of consistent datasets while measuring quality measures.
- A developed API that can connect not only to various enterprise systems and analytical tools, but also to various data management and profiling tools.

Base architecture of the framework [9] is shown on Figure 3.

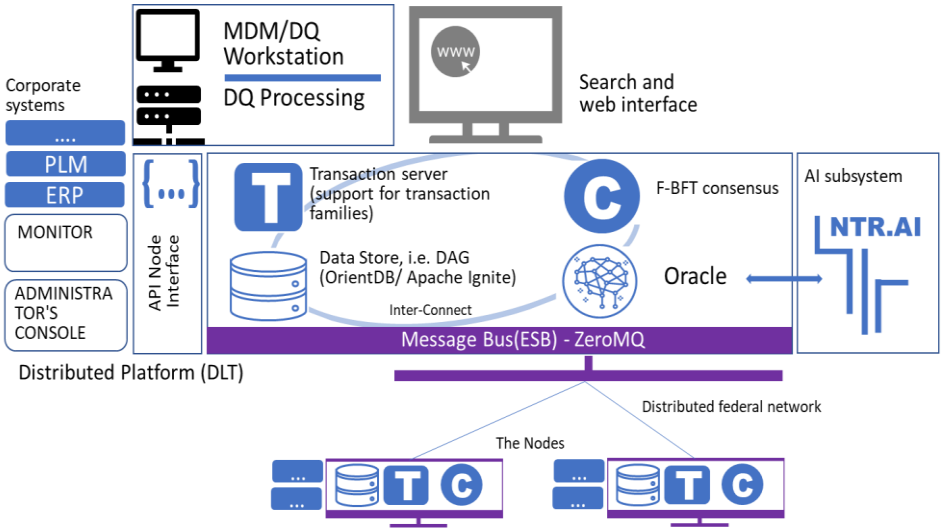

Fig.3. Base architecture of the framework

6. Conclusion

The use of distributed ledger technologies to support basic data between organizations will provide a single information space for groups of companies that are integrated horizontally or vertically. This technology enables real-time quality indicators to be calculated and information exchanged effectively in operational data, to improve the quality of analytical data and, ultimately, to make the decision-making process qualitative.

Data virtualization is a method of organizing of access to data without requiring knowledge of its structure or location in a particular information system. This makes it possible to achieve the main goal - to simplify access to and use of data by turning it into a service, thus significantly shifting the paradigm from storage to use. This is provided by the proposed virtualization concept, which supports the scalability and operational efficiency required for Big Data environments through the implementation of:

- portioning, that is sharing resources and transitioning to streaming data;
- insulation, which is an object-oriented approach to data taking into account the application of the subject area;
- encapsulation that allows you to save logical storage as a separate object.

All this ensures differentiated data security and confidentiality. Therefore, the proposed data virtualization is more than just a modern approach, it is a completely new view of the data and how to work with it.

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] Gankevich, I., et al, Constructing Virtual Private Supercomputer Using Virtualization and Cloud Technologies //Lecture Notes in Computer Science, 8584, 341-354 (2014) DOI: 10.1007/978-3-319-09153-2_26

[3] Virtual Supercomputer as basis of Scientific Computing / A. Bogdanov, A. Degtyarev, V.Korkhov, V. Gaiduchok, I. Gankevich // Horizons in Computer Science Research — New York: Nova Science Publishers, Inc., 2015. — Vol. 11, — 203p., 159-198 p.

[4] Bogdanov, A., Degtyarev, A., Korkhov, V. Desktop supercomputer: what can it do? // Phys. Part. Nuclei Lett. 14(7), 985-992 (2017) DOI: 10.1134/S1547477117070032

[5] Bogdanov A., et al, Evolving Principles of Big Data Virtualization. //Lecture Notes in Computer Science, 12254, 67-81 (2020) DOI: 10.1007/978-3-030-58817-5_6

[6] Bogdanov, A. V., Shchegoleva, N. L. & Ulitina, I. V., 2019, Database ecosystem is the way to data lakes *Proceedings of the 27th Symposium on Nuclear Electronics and Computing (NEC 2019).* Korenkov, V., Strizh, T., Nechaevskiy, A. & Zaikina, T. (ed.). RWTH Aahen University, pp. 147-152 (CEUR Workshop Proceedings ; vol. 2507).

[7] DGT, the Decentralized Enterprise Platform. http://dgt.world/

[8] Bogdanov A., et al A DLT Based Innovative Investment Platform. //Lecture Notes in Computer Science, 12251, 72-86 (2020) DOI: 10.1007/978-3-030-58808-3_7

[9] Bogdanov A., et al Data Quality in a Decentralized Environment. //Lecture Notes in Computer Science, 12251, 58-71 (2020) DOI: 978-3-030-58808-3_6