Big Data Virtualization: Why and How?

Alexander Bogdanov[1], Alexander Degtyarev[1,2], Nadezhda Shchegoleva[1], Vladimir Korkhov[1,2], Valery Khvatov[3]

[1] Saint Petersburg State University, St. Petersburg, Russia,
a.v.bogdanov@spbu.ru, deg@csa.ru, n.shchegoleva@spbu.ru, v.korkhov@spbu.ru
[2] Plekhanov Russian University of Economics, Moscow, Russia
[3] DGT Technologies AG., http://dgt.world/
valery.khvatov@gmail.com

Abstract. The increasing variability, dynamic and heterogenous nature of big data, as well as the flexibility required by data producers and consumers lead to the necessity of organizing access to data without requiring information about its structure or belonging to any particular information system, i.e. data virtualization. Data virtualization complements the concept of a virtual supercomputer, allowing us to consider the computing environment as a single information continuum, where computing tools are part of the general data model and data is considered in the context of three basic components: integration, analysis, and data presentation. In this paper, we present the concept of unified, generalized and encapsulated representation of data coming from a heterogenous set of data sources, based on the extension of the Logical Data Warehouse (LDW) and the Distributed Data Network in the form of a distributed ledger – the virtual DLT (vDLT). The main difference between the vDLT and LDW approaches is the decentralized management of data using a consensus mechanism. We discuss data virtualization practices, the methodology of constructing a virtualized data environment, and compare core steps and approaches for each of the two directions.

Keywords: Big Data, Data virtualization, Virtual Personal Supercomputer, data network, data marketplaces, distributed ledger technologies.

1 Introduction

Background and environment

According to the analyst predictions [1], the global data volume will reach 163 zettabytes by 2025. Half of that data will be produced by enterprises. Another property of the data growth will be the fact that no less than three quarters of corporate data will be sourced and processed outside of data processing centers and clouds. That means that this data will come from peripheral, or edge solutions outside of the confines of any particular organization.

This property means that the approach to the data virtualization problem must change. In the era of post digital modernism, the scope of virtualization needs to be

redefined from hardware realization to increasing abstraction of the logical structure. Data consumers are not interested in the data storage format or methods for its extraction. Previous schemes of working with data were focused on creating a single repository, however, as data volumes grow, such integration is frequently physically impossible, while the information stored in a repository quickly loses its relevance.

Peter Druker, on August 24, 1998 [2] said: "The next information revolution is well underway, but it is not happening where information scientists, information executives, and the information industry in general are looking for it. It is not a revolution in technology, machinery, techniques, software or speed. It is a revolution in CONCEPTS. … So far, for 50 years … the information revolution has centered on the 'T' ("Technology") in IT ("Information Technology"). The next information revolution asks, what is the MEANING of information and what is its PURPOSE?". The new generation of IT systems in 2020 places the "I" into the "IT" and, undoubtedly, such transformation requires full virtualization; where data virtualization will help achieve the necessary level of encapsulation of information from technical implementation.

Below are the core reasons that necessitate the implementation of data virtualization mechanisms:

- The complex nature of digital business, which substitutes the traditional vertical and horizontal integration with platform and ecosystem approaches that connect many distributed sources of data in real time;
- The change of paradigms in analytic applications: from regulated reports towards self-analytics, formation of dynamic data models by qualified users – data scientists;
- Greedy artificial intelligence algorithms that depend on large and high-quality data sets. As a result, aggregating legacy data into a corporate data warehouse is no longer sufficient for advanced analytics;
- Increasing data variability and the increasingly dynamic nature of confidentiality models, which take into account the need to protect personal data and trade secrets. Data virtualization works in conjunction with edge computing technologies to generate aggregated results without the need for transferring the detailed ones.

The present study is aimed at structuring approaches to data virtualization and forming a basic methodology for constructing an abstract data infrastructure.

Personal virtual supercomputer and virtual data

Data virtualization complements the concept of a personal supercomputer [3], allowing us to consider the computing environment as a single information continuum, where computing tools are part of the general data model and data is considered in the context of three basic components: integration, analysis, and data presentation.

Today's computing environments are increasingly heterogenous. The main problem of using heterogenous computing systems is the impossibility of balancing load. Dynamic balancing is not possible due to the delay time in communicators, while static balancing is unattainable due to different processor performance.

This situation is improved by virtualizing processors, since the performance of a virtual processor can be adjusted to an algorithm. After that, the key problem moves into the field of communications. If two processors are connected through more than

two communicators, then the total delay time makes it impossible to exchange data between the corresponding computing processes. Therefore, it is necessary to virtualize communications as well by connecting two data-exchanging processors directly through a virtual channel.

Since standard communication protocols are relatively slow, there exist instruments specifically meant for accelerating data exchange between virtual processors. Since each of the virtual processors is connected with its physical memory on one end and with the virtual net channels on the other, it is possible to build a collective virtual memory, i.e. create a method for effective reconfiguration of communication channels between processors. In this scheme, the main limiting factor for computing and data processing is the size of the file system of each node. Therefore, a virtual file system may be built for the entire calculating complex. This way, the entire system turns into a virtual SMP-system in which you can use the programming paradigm of shared memory. Tests have shown that using collective virtual memory even with accelerating protocols does not allow scaling of calculations by more than 40-80 threads. Therefore, a toolkit was developed that made it possible to avoid pulling data to a virtual node, but rather pull the computational process to the data. This is all the more effective, the greater the amount of processed data.

Data virtualization introduces an additional level of abstraction – not only in working with memory and computational processes, but also in processing the data structure, data streams, and network processing topology. At the same time, individual data processing subsystems, such components as cleaning, validation, and storage may be built as virtualized entities based on containerization technology. This would create an internal virtual data exchange network based on internal API.

Core principles of data virtualization

At present, infrastructure virtualization is well developed and has a set of principles and practices (partitioning, encapsulation, hardware independence) that allow for the construction of virtual machines, containers, virtual data centers, and clouds. Data virtualization also requires an understanding of general principles that can be generalized on the basis of an analytical note [4]:

- Data virtualization is built taking into account the business requirements, data management processes, and the life cycle of information objects;
- The availability of data and information in the framework of data virtualization is ensured by the stability of sources, the separation of complex data sets by workloads and granulation, as well as the continuous monitoring of data quality;
- The main component of data virtualization is a unified data layer based on adaptive mechanisms for working with data. As an example, it can be implemented as a series of interconnected and bidirectional data pipelines that meet the needs of business. Such pipelines can be built using basic data objects – data snapshots, increments, representations, etc.;
- The aforementioned data objects must allow for continuous repeated usage in compliance with the idempotency principle (repeatability of results) and automa-

tion of processing based on maintaining data catalogs (metadata, data classes, data clustering), thereby creating a single consistent logical data model;

- Loosely connected service architecture must provide a universal data publishing mechanism (DaaS – Data as a Service). The corresponding implementation would provide the end users with the results of processed data streams in a format that is accessible in a short time, convenient, and protected from external interference.

Data virtualization as part of the data management process

The construction of a data virtualization system cannot be successful without taking into account the general system of managing data. According to [5], data management is understood as a complete set of practices, concepts, procedures, and processes that allow organizations to gain control over data. The practical use of the Data Management Body of Knowledge (DMBoK) allows for the organization of practical solutions to deploy a virtual data environment that would be more sensitive to the integrity of the organizational approach to data management, rather than classical solutions like corporate data warehousing.

A complete set of DMBoK practices includes maintaining the relevant policies (Data Governance and Data Stewardship), shaping of data architecture, managing integration, data security, metadata, and etc. Like a conventional data warehouse, a virtualized structure requires a number of solutions for defining the data profile, management components, and methods for working with master data. Using a single framework lowers implementation risks raises the effectiveness of data use and lowers the time required to implement new solutions built on modern data.

An alternative approach to data management is given by the DCAM framework [6]. The main differences of these approaches fall on three fundamental provisions:

- While DMBoK considers data management as a process within the general IT structure, DCAM considers data management as a distinct group of processes;
- DMBoK builds its data management model on knowledge areas and business functions, which in the context of data virtualization means services. DCAM considers data models as attributes of an organization's business model.
- DMBoK separates data modelling into a separate process, so that the main results of the data architecture are data streams. DCAM derives data models from the declarative architecture.

While both approaches allow for the creation of virtualized data management environments, the DMBoK approach seems to be the most relevant for dynamic environments with a large number of sources, i.e. where data virtualization has the largest effect.

2 Architectural Aspects of Data Virtualization

In this work, we consider two different approaches to data virtualization – based on the extension of the basic storage construction methodology, the so-called logical data warehouse (LDW), and in the form of a distributed ledger – the virtual DLT (vDLT).

Logical Data Warehouse

The Logical Data Warehouse (LDW) is a data management architecture with clearly defined abstract layers for integrating external sources and building data marts (analytic showcases). The most typical solution in this approach is the Denodo data virtualization platform [7].

In the traditional Enterprise Data Warehouse (EDW) scenario, data comes in from transactional databases, business applications, ERP systems, or any other source of data. The data is later standardized, cleaned, and transformed in the ETL process (extraction, transformation, loading), to ensure reliability, consistency and correctness for use by a wide range or organizational applications, analytics, and reports.

LDW extends this concept by integrating logical sources of information and building data presentation services in various interpretations. LDW stores data at its source and dispenses with heavy ETL routines. The LDW architecture contains the following core components:

- Connect – a layer of connectors to data sources (real-time);
- Combine (the Unified Data Layer) – services for integrating, improving quality, transforming data;
- Publish – normalized views (data marts) for supporting applied services (Enterprise Applications, BI, reporting, Mobile, Web)

The work of the core processes is enabled by orchestration modules, metadata management, cache, monitoring and other modules. The figure below provides a reference LDW model built into the general data management architecture, including on top of traditional sources, one of which can be a regular data warehouse.

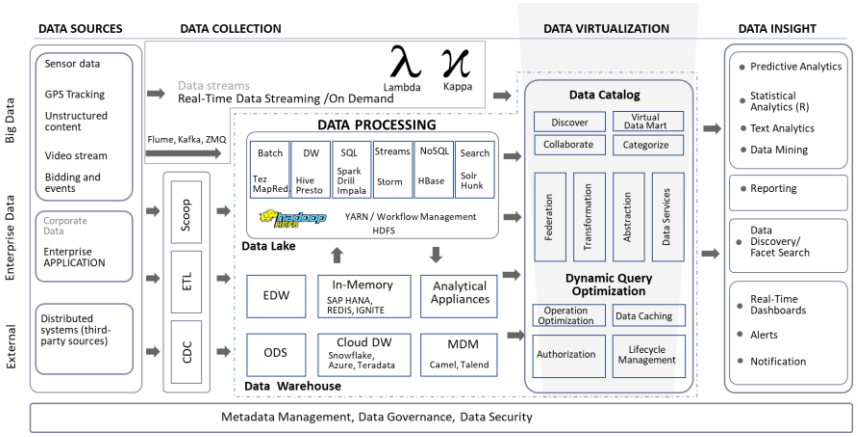

Fig. 1. LDW Hybrid Reference Architecture

As was mentioned above, the Unified Data Layer plays a significant role in LDW. It must possess the following important characteristics:

- An extensible data mode, including variable data structures;
- The existence of mechanisms for structuring data and extracting information from unstructured sources;
- Quality assessment system in the form of DQI (Data Quality Index);
- Access control based on multi-mode access;
- The system of data federation, as well as the optimization of the system of queries to heterogenous sources of information;
- The subsystem for publishing data access services taking into account domain structure (the so-called DDD approach, Domain Driven Design, which groups services into thematic groups).

Distributed Data Network

The Distributed Data Network or Virtual Distributed Ledger Technology (vDLT) is an alternative approach to data virtualization, solving it by collecting and managing data through the blockchain class of mechanisms. Blockchain, or, in a broader sense, DLT systems form data registries that are shared among the entire network of data source nodes or its separate parts (sharding).

The main difference between the vDLT and LDW approaches is the decentralized management of data using asymmetric cryptography and a mechanism for real-time data matching – a consensus mechanism. This approach is explored further in [8] and [9]. As an example, data virtualization may be implemented using the DGT platform [10]. Below is its reference architecture (DGT):

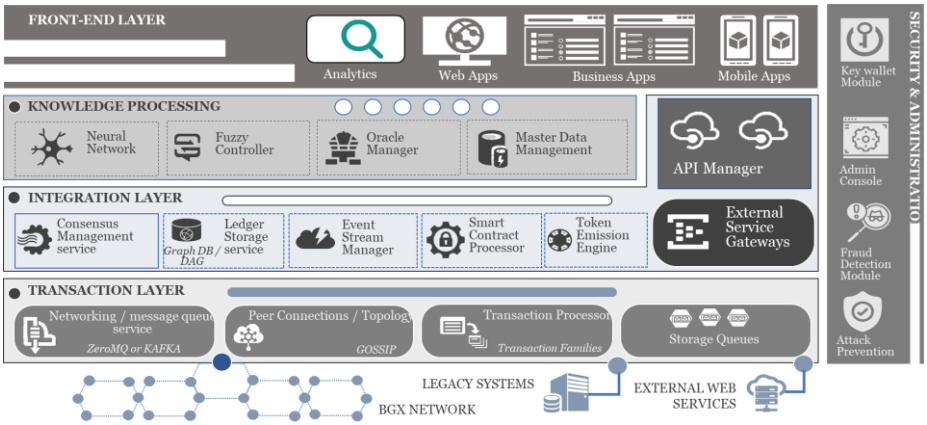

Fig. 2. DGT Reference Architecture

It should be noted that not every DLT architecture is capable of supporting data virtualization. Data virtualization can be accomplished by a decentralized network if it has the following attributes:

- Support for a virtual node model (allocation of data storage, consensus mechanisms and other components as independent containerized components);
- Support for multiple transaction families with the possibility of their full or partial overload (resetting);
- Support for registry partitioning (sharding) mechanisms. For example, in DGT, such partitioning is achieved through private branches within a DAG-based registry (Directed Acyclic Graph);
- Support for a federative network structure, which allows for flexible data routing (see also the NFV approach below);
- Support for end-to-end search within a distributed registry;
- Transaction support for unstructured, binary data.

These features allow us to receive data from distributed sources and then distribute them through the network. Within the Distributed Data Network concept, each network can be both a source of data and a receiver, which establishes this model as oriented towards dynamic data virtualization schemes. The advantage of such solutions is the support for edge calculations, which enables the drawing of privacy zones almost fully arbitrarily. A limitation of these solutions is the insufficient support for massive batch downloads (MPP).

Network Virtualization

The aforementioned vDLT approach is focused on networks of transferring value (Internet-of-Value), where information accompanies some value, while consensus mechanisms establish an environment of trust. At the same time, the network environment also requires virtualization, not only within closed node clusters, but also in the broader network space.

Traditional network services run on fixed, dedicated hardware. Virtualization of network services, such as routing, load balancing, and firewalls brings an additional impact on data virtualization by simplifying data routing and changing the data access model, as well as providing additional data protection and verification mechanisms. The most common approach is the NFV concept – Network Functions Virtualization [11]. At present, this concept is also being developed by the industry specifications group of the European Telecommunication Standards Institute (ETSI, [12]). Virtualization of network services allows for the creation of dynamic networks, fixing one of the greatest vulnerabilities of virtualized data environments – instability with respect to flickering sources.

Data Protection & Confidentially

An important driver of data virtualization is the support of advanced data protection mechanisms, such as working across overlapping security borders, anonymizing data, and data resistance to changes (refusal of signature).

Below are the main mechanisms that ensure the necessary level of protection for virtualized data:

- The use of an addressing system based on asymmetric cryptography (a pair of private and public keys), as well as an asymmetric digital signature mechanism;
- End-to-end encryption of transmitted transactions (the "envelope" concept, which leaves a set of public data available, while hiding the message itself – payload);
- Forming a Merkle-based data registry with properties of immutability;
- Stability as per the BFT (Byzantine Fault Tolerance) and CFT (Crash Fault Tolerance) types. A reliable mechanism for ensuring the consensus of data allows for reliable integrability of information and its quality at the time of entry into the registry.

It is important to protect the virtualized environment in accordance with the legal requirements, such as the European General Data Protection Regulation (GDRP), or the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA), or the Privacy Act highlighting Australia's privacy principles, the California Consumer Privacy Act (CCPA), or the Health Insurance Portability and Accountability Act (HIPAA). Data anonymization can be implemented through the general Differential Privacy approach [13].

3 Data Virtualization Practice

Constructing a virtualized data environment has its own methodology, presented in such works as Gartner's [15]. A specific plan differs in relation to using an LDW or a Distributed Data Network based strategy. The core steps and approaches for each of the two directions are presented below:

Constructing LDW:

- Setting architectural principles and describing a functional model, which would have a fundamental meaning for data virtualization;
- Forming a data management policy based on the Information Governance model, data consumption, as well as planning the system for access and workloads (Capacity Model);
- Developing the Data Model considering Master Data Management, Metadata;
- Designing Data Quality Measures;
- Developing an allocation model (Cloud-First Approach vs. On-Premises);
- Developing data adapters and data integration policy;
- Agile Data Marts and Sandboxes;
- Data Ingest Design (migration vs. greenfield);
- Growing Agile and Self-Service Platforms;
- Data Acquisition;
- Data Lake Analytics Consumption

Constructing a Distributed Data Network

- Formation of architecture principles and describing a functional model;
- Planning the network processing topology (Agent-based, Federated, Mesh);
- Developing a transaction family and validation rules;
- Setting up consensus mechanisms and storage systems (DAG);
- Designing API and Data Bridges;
- Workload planning and network launch.

4 Conclusion

Data virtualization is a methodology of organizing data access without the need to obtain information about data structure or placement in a specific information system. The main goal of data virtualization is to simplify the access and use of data by converting it into a service (Data as a Service), which essentially shifts the paradigm from storage to usage.

The three major principles of data virtualization supporting the scalability and operational efficiency for big data environments are as follows:

- Partitioning: sharing resources and moving to streaming data.
- Isolation: transition to the object representation of data with reference to the domain model.
- Encapsulation: logical storage as a single entity

In the present study, we structured the approaches to data virtualization and proposed a basic methodology for constructing an abstract data infrastructure. We outlined an approach based on the concept of a virtual personal supercomputer, discussed core principles of data virtualization and architectural aspects of data virtualization. Two different approaches were analysed: extension of the Logical Data Warehouse (LDW) model and the Distributed Data Network based on the virtual distributed ledger technology (vDLT). Finally, we analyzed data virtualization practices and compared of the core steps and approaches for each of the two directions.

Data services and APIs are changing the way distributed information is accessed. Data virtualization is more than just a modern approach; it is a completely new way to see data. So far, the information revolution focused on the Technology; the next information revolution seeks answers to the Meaning and Purpose of information. Data virtualization is a way to address these global issues.

References

1. IDC. “Data Age 2025, The Digitization of the World From Edge to Core”, IDC White Paper, November 2018, https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, Last Access 12 Jul 2020;
2. Drucker, Peter F., "The Next Information Revolution", Forbes, 24 August 1998
3. Bogdanov A., Private cloud vs Personal supercomputer. - Distributed computing and GRID technologies in science and education, JINR, Dubna,2012, pp. 57 – 59.
4. Capgemimi, Data Virtualization. How to get your Business Intelligence answers today, 2017. https://www.capgemini.com/wp-content/uploads/2017/07/ data_virtualization._how_to_get_your_business_intelligence_answers_today.pdf/ Last Access 12 Jul 2020;
5. DAMA International. DAMA-DMBOK: Data Management Body of Knowledge, 2nd edition. Technics Publications, 2017;
6. EDM Council. “Data Management Capability Assessment Model, DCAM 1.2.2. (Assessor’s Guide)” EDM Council, 2 Dec. 2018, http://www.edmcouncil.org/dcam Last Access 12 Jul, 2020;
7. DENODO, Data Virtualization Mainstream, Whitepaper, Data Virtualization Goes Mainstream, https://whitepapers.em360tech.com/wp-content/files_mf/ 1415789864DataVirtualizationGoesMainstream.pdf. Last Access 12 Jul, 2020;
8. Fei Richard Yu, Jianmin Liu, Ying He, Pengbo Si, Yanhua Zhang, Virtualization for Distributed Ledger Technology (vDLT), *in* IEEE Access, vol. 6, pp. 25019-25028, 2018, doi: 10.1109/ACCESS.2018.2829141.
9. Bogdanov, A.V., Degtyarev, A.B., Korkhov, V.V., Kamande, M., Iakushkin, O.O., Khvatov, V. About some of the blockchain problems (2018) Selected Papers of the 8th International Conference "Distributed Computing and Grid-Technologies in Science and Education", GRID 2018, CEUR Workshop Proceedings, 2267, pp. 228-232.
10. DGT Bluepaper v 0.2. http://dgt.world/ (Last Access: 12 Jul 2020);
11. C. Liang and F. R. Yu, ‘‘Wireless network virtualization: A survey, some research issues and challenges,’’ IEEE Commun. Surveys Tuts., vol. 17, no. 1, pp. 358–380, 1st Quart., 2015

12. ETSI, NFV Release 3 Description, https://docbox.etsi.org/ISG/NFV/Open/Other/ReleaseDocumentation/NFV(20)0000016_NFV_Release_3_Description_v0_5_0.pdf Last Access 12 Jul, 2020;
13. Cynthia Dwork, Aaron Roth, The Algorithmic Foundations of Differential Privacy, University of Pennsylvania & Microsoft Research, 2015. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf Last Access 12 Jul 2020
14. Bogdanov, A. Degtyarev, and V. Korkhov. Desktop Supercomputer: What Can It Do? Physics of Particles and Nuclei Letters, Vol. 14, No. 7, pp. 985–992, 2017.
15. Gartner, The Practical Logical Data Warehouse: A Strategic Plan for a Modern Data Management Solution for Analytics, 2019, https://www.gartner.com/doc/3867565 Last Access 12 Jul 2020.