Big Data Storage and Data Virtualization

Author, Ajit Singh

A Data Science Foundation White Paper

April 2019

---

www.datascience.foundation

ABSTRACT

The major objective of this paper is to present Big Data Storage techniques and Data Virtualization. The Data virtualization servers have focused on making big data processing easy. They can hide the complex and technical interfaces of big data storage technologies, such as Hadoop and NoSQL, and they can present big data as if it is stored in traditional SQL systems. This allows us as developers to use our own existing skills and to deploy our traditional ETL, reporting, and analytical tools that all support SQL. Additionally, the products and our existing skills can extend the data security mechanisms for accessing and processing big data across multiple big data systems. But with scale and performance rising, making big data processing is not enough and easy anymore. As such, the next challenge for data virtualization is parallel to big data processing. In this paper, All of the above regular issues are covered in my paper along with their proper prospects.

Keywords: Big Data, Storage, Data Virtualization, Prospect of data virtualization.

1. INTRODUCTION

The big data concept has been adopted by all kinds of organizations. Data can be "big" because of its enormous volume, because it's not structured in the traditional way, or because it's streaming in with enormous quantities. The business advantages of big data systems are clear, they can improve, deepen, and strengthen an organization's analytical capabilities with relatively attractive infrastructure economics.

Until now the data virtualization servers have focused on making big data processing easy. They can hide the complex and technical interfaces of big data storage technologies, such as Hadoop and NoSQL, and they can present big data as if it's stored in traditional SQL systems. This allows developers to use their existing skills and to deploy their traditional ETL, reporting, and analytical tools that all support SQL. Additionally, the products and existing skills can extend the data security mechanisms for accessing and processing big data across multiple big data systems. Butcwith scale and performance rising, making big data processing easy is not enough anymore. As such, the next challenge for data virtualization is parallel big data processing.

2. Big Data Storage Technologies

The technology is available to store, process, and analyze big data. Systems such as the Hadoop platform and numerous NoSQL products have been designed and optimized to work with big data. With specialized analytical tools all that data can be analyzed fast and efficiently, self-service data preparation tools help business users and data scientists understand big data by applying artificial intelligence techniques. In addition, architectures exist, such as data lakes, to help organize and process big data.

Technique 1: Parallel Processing of Big Data Requests - Crucial for big data technologies is that they can process requests on large data sets fast. Therefore, they apply several techniques, one is parallel processing. When all the data of a file is stored on one disk and only one process can access the file, access is serial; see Figure 1. Two records of that same file are never processed simultaneously, only serially. Processing real big data sets serially takes too long. The first technique to improve the performance of requests on big data was introduced years ago in SQL systems and is called file or data partitioning and is the basis for parallel processing. This technique has also been implemented in big data technologies.

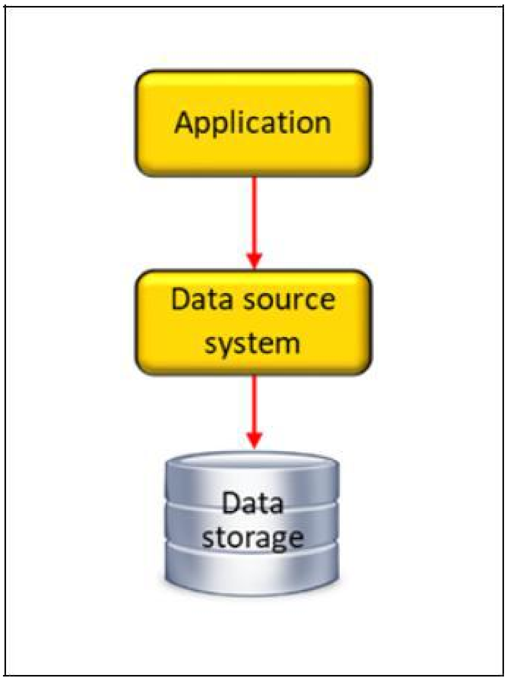

Figure 1 When all the data of a file is stored on one disk and only one process can access the file, data access is by definition serial and not parallel.

Technique 2: Massive Parallel Processing - For a long time, data partitioning and parallel processing features offered sufficient performance improvements, until big data came along. Purely the sheer volume of big data demands a level of parallelization higher than was common. Unfortunately, most SQL systems can only parallelize processing efficiently across a limited number of nodes. The Hadoop platform and the NoSQL products, on the other hand, are designed specifically to support massive parallel processing. They can distribute data and processing across hundreds of disks and nodes and have a similar number of workers running in parallel. Their internal architectures make these systems big data-ready.Figure 2 Big data technology is able to process requests in parallel by piding them in subrequests and sending them to the workers for parallel processing.

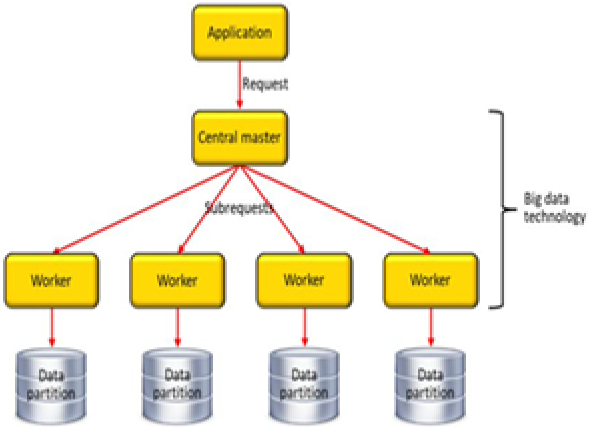

Figure 2

Technique 3: Non-Standard Programming Interfaces – Big data technologies support application programming interfaces that are suitable for developing big data systems. The consequence is that they don’t support well-known interface standards, such as SQL and SOAP. They all support proprietary and quite technical interfaces that can only be deployed from within programs developed in languages, such as Java, C#, or Python.

In addition, these systems all support different programming interfaces. Products such as Apache HBase, Cassandra, and MongoDB, all support their own native API. No standards exist for the programming interfaces of current big data technologies. There is no equivalent for SQL in the NoSQL world.

Technique 4: Specialization - Another technique used by big data technologies to improve performance is specialization. Traditional SQL database servers, such as Microsoft SQL Server and Oracle, are so-called generic database servers. They are good for almost any type of application, whether it’s a transaction system, a portal, or a reporting environment. But they don’t really excel at anything.

3. Data Virtualization - Make Big Data Processing Easy

Until now, data virtualization has brought several features to the big data table to help organizations adopt and process big data more easily.

1. Simplifying Transforming Native and Technical Programming Interfaces - Data virtualization servers have always allowed developers to use well-known and standardized interfaces, such as SQL and SOAP, to access a wide range of data sources using complex, technical, and mostly proprietary interfaces, such as Excel spreadsheets, flat files, and mainframe systems. Nowadays, data virtualization servers also allow SQL or SOAP developers to access the proprietary interfaces of big data technologies. The benefits are that they don't have to learn

new programming interfaces, most reporting and analytical tools can be used to access big data, and it unlocks big data to a wide range of developers, from BI developers to IT specialists.

2. Simplifying Polyglot Persistence - Dealing with polyglot persistence is simplified with data virtualization. All the different programming interfaces can be hidden to the developers. Developers don't have to work with a multitude of different interfaces and different big data technologies. Instead, all the big data sources are presented as one integrated database accessible through one interface. While maintaining the parallel processing power of those big data technologies, data virtualization simplifies the development of applications dealing with all these perse data storage technologies easier. Similarly, it also reduces the risks for an organization to deploy an additional specialized big data storage technology for a new big data system with a special use case.

3. Seamlessly Integrating Big Data with Traditional Data - Data virtualization simplifies integration of big data with data stored in data warehouses or transactional databases developed with traditional database technology. In fact, the entire set of data sources, including the big data stores, can be presented to the developers as one integrated database. Because of the data federation capabilities of data virtualization servers, developers won’t even know that some of the data is stored in big data systems and some isn’t. The benefit is that developers can, for example, integrate historic data (stored in a SQL-based data warehouse) with streaming data stored in Hadoop, and can combine descriptive data stored in data marts with sensor data kept in MongoDB.

4. Enhancing Data Security Capabilities for Big Data - Data virtualization servers offer extensive data security mechanisms for accessing data sources in general, including the big data systems. With a data virtualization server one integrated data security layer can be defined on a heterogeneous set of data sources. The data security features offered by data virtualization servers are much more extensive and more detailed than those of most big data technologies, allowing big data to be secured in a way that's not possible with that technology itself. In addition, security specialists only have to deal with one tool for specifying data access rules for a wide range of data sources.

4. **Prospects of Big Data Virtualization**

i. Easy Staging of Big Data
ii. Speeding Up Slow Data Sourcess
iii. Simplifying the Data Lake
iv. Accessing Remote Big Data
v. Accessing Offloaded Cold Data
vi. Processing Traditional Data Sources in Parallel

5. **Conclusion**

Analytics workloads at greater scale, with higher performance have driven demand for big data processing. Data virtualization has enriched the big data world with features that makes working with it much easier, and in addition, it fills some important functionality gaps. But until now, data virtualization has not focused on speeding up big data processing. It makes big data processing fast.

REFERENCES

- http://www.cisco.com/c/en/us/products/cloud-systems-management/data-analytics/index.html#~overview
- http://www.techtarget.com/contributor/Rick-Van-Der-Lans
- http://www.b-eye-network.com/channels/5087/articles/
- http://www.b-eye-network.com/channels/5087/view/12495
- R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007. 6 R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation