DATA LAKE AS A FUTURE OF BIG DATA

[1]THUREIN KYAW LWIN, [2]A.V.BOGDANOV

Saint Petersburg State University, Russia
E-mail: [1]trkl.mm@mail.ru, [2]Bogdanov@csa.ru

Abstract - We introduce the data lake view to Big Data processing. We present a new approach to distributed Database System and we propose to use some effective Data tools for databases management. Thus we can provide a fresh approach for design, architecture, performance and concurrency control of Distributed Database Systems (DDBS). Data Lake is a new storage concept that is gaining ground in the Cloud. Often the Data Lake term is being used as part of the Big Data solution. In theory, it is where you can store raw data in its native format, usually in Hadoop and Hadoop Distributed File System (HDFS). Data lakes' architecture acts as a storage space for data, so engineers and IT teams can easily access it for future.

Keywords - Big Data, IT, CAP, information processing, analysis data, big data development, Data Lakes, Large Data

I. INTRODUCTION

What are the data lakes and how they can be used for big data? Big Data, today this term is in headlines and of very much interest the field of information technology. Big Data is quickly becoming a business best resource. Refined and analyzed data can help companies to improve operations and uncover insights. But as data begins to flow in from multiple sources, it's important to prevent it from being siloes by pooling it in a central repository architecture that acts as a storage space for data, so engineers and IT teams can easily access it for future use. Moreover, a data lake can form the firm foundation of a modern data platform. To get a deeper understanding of data lakes and why they're the optimal candidate for housing big data, it's important to dive into what makes them so different from data warehouses. So, in this article we will show, how data lakes are capable of storing such vast and diverse amounts of data, and what is the underlying architecture of these massive repositories?

II. PRINCIPLES OF BIG DATA

Most of the large information technology suppliers in their business strategies used the concept of "big data" for marketing purposes and do not explore its influence on the choice of methods and tools for data processing.

We want to stress the crucial value of CAP theorem for Big Data project. The CAP theorem is a heuristic statement that in any realization of distributed computations, it is possible to provide the following three properties:

1) Consistency of data - in all compute nodes at one time the data does not contradict each other;
2) Availability - any request to a distributed system is terminated with a correct response;
3) Partition tolerance - the splitting of a distributed system into several isolated sections does not lead to an incorrect response from each section.

We also want to illuminate that CAP theorem describes a system of a very specific model. It is a single node \ register (read-write) and that's it. For example, it does not take into account transactions that affect several objects. They are simply outside the limits and conditions of the theorem, until you realize in some way those objects as a single object. The only type of the failure of CAP theorem is network sharing, i.e. the nodes remain active, but the network between some of them does not work. This kind of error somehow happens, but it's not the only thing that can go wrong (nodes can be broken or rebooted), the disk space may end, we can catch a bug in the software, and so on. In the process of creating a distributed system, we need to keep in mind the much larger range of possible errors and trade-offs which they need. Too much focusing on the CAP theorem leads to ignoring other important problems. Moreover, the CAP theorem says nothing about latency that can develop much more problems than accessibility. In fact, CAP-compatible systems can be arbitrarily slow and still be called "available". If we can bring everything to the extreme, we are sure that our users will not be inclined to call the system "affordable".
In modern conditions, organizations create a large number of unstructured data, such as text documents, images, video recordings, machine codes, tables, etc. All this information is stored in a variety of repositories, sometimes even outside of the organization. Companies can have access to a huge array of own data and do not have the necessary tools, which could establish the relationship between these data and make significant conclusions based on them. Traditional methods of analyze information cannot keep up with the huge volumes of growing and updated data, which eventually opens a concept of Big Data ecosystem [1].

III. DATA LAKE ARCHITECTURE

A data lake is a repository for structured, unstructured, and semi-structured data. Data lakes are much different from data warehouses since they

allow data to be in its rawest form without needing to be converted and analyzed first. In simpler terms, all types of data that are generated by both humans and machines can be loaded into a data lake for classification and analysis later on. Data warehouses, on the other hand, require data to be properly structured before any work can get done [2].

Data lakes are built upon a schema-on-read data model. A schema is essentially the skeleton of a database outlining its model and how data will be structured within it. Think of a blueprint. The schema-on-read data model means we can load our data in the lake as-is without having to worry about its structure. This allows for much more flexibility. Data warehouses, on the other hand, are comprised of schema-on-write data models. This is a much more traditional model for databases. Every set of data, every relationship, and every index in the schema-on-write data model must be clearly defined ahead of time. This limits flexibility, especially when adding in new sets of data or features that could potentially create gaps within the database [3].

The schema-on-read data model acts as the backbone of a data lake, but the processing framework (or engine) says how data actually gets loaded into one. There are two processing frameworks of data for data lakes:

- Batch processing – Millions of blocks of data processed over long periods of time (hours-to-days). The least time-sensitive method for processing big data.
- Stream processing – Small batches of data processed in real-time. Stream processing is becoming increasingly valuable for businesses that harness real-time analytics.

Hadoop, Apache Spark, and Apache Storm are among the more commonly used big data processing tools which are capable of either batch or stream processing. Some tools are particularly useful for processing unstructured data such as sensor activity, images, social media posts, and internet clickstream activity [4]. Other tools prioritize processing speed and usefulness with machine learning programs. Data warehouses rely on structure and clean data, whereas data lakes allow data to be in its most natural form. This is because advanced analytic tools and mining software intake raw data and transform it into useful insight [3].

IV. INCREASED DATA LAKE FLEXIBILITY

After Data Lake emphasized flexibility and data availability one can provide users and subsequent applications data without a schema. In other words, data is in a “natural” format, regardless of origin. Most of the Data lake applications do not support partial or incremental downloads. An organization cannot load or reload portions of its data into a data lake. As a rule, it's all or nothing. As noted, a data lake holds a vast amount of raw, unstructured data in its native format, whereas the data warehouse is much more structured into folders, rows, and columns. As a result, data lake is much more flexible with its data than a data warehouse [5].

Each data element in a lake is assigned a unique identifier and tagged with a set of extended metadata tags. When someone performs a business query, based on a certain metadata, all the data tagged is then analyzed for the query or question. Unlike a data warehouse, data lakes don’t have an underlying database. Instead, data lakes use a flat file system. With a database, you have to choose data and columns before you write to it. The trade-off is that it might take a while to insert the data into a database, but when you do a query it is a lot faster than in a data lake, which has to process the data as it is read.

A. Data Lakes are Parallel with Hadoop and Relational Databases

Hadoop is open source distributed file system that more and more organizations are using today. Hadoop has grown for many reasons, first of them is due to the fact that it works with real data, that could not be operated by relational database management systems (RDBMS). Honestly, its open source, fault tolerance and parallel processing also occupy an important place in the list. RDBMs simply were not designed to handle gigabytes or petabytes of unstructured data. Try uploading thousands of photos, videos, tweets, articles and emails to your traditional SQL server or Oracle database, and also run reports or write to SQL only[6]!

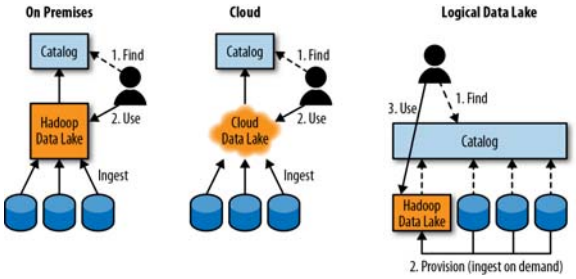


Figure 1. Different Data Lake Architecture

V. DATA LAKE WITH BIG DATA

The consept of Big Data is expanding with data being generated every second from multiple communication channels. Consequently, Data Lake solves the need to store this unstructured and unorganized data. You can consider Data Lake as a parallel system able to store big data, and a system to perform computations on the data without moving the data [7].

Big Data is used to describe both the technology ecosystem around, and to some extent the industry that deals with, data that is in some way too big or too

complex to be conveniently stored and/or processed by traditional means. Sometimes this can be a matter of sheer data volume: Once you get into the 100s of terabytes or petabytes, your good old fashioned RDBMS databases tend to throw in the towel, and we are forced to spread our data across many disks, not just one large one. And at those volumes we'll want to parallelize our workloads, leading to things like MPP databases, the Hadoop ecosystem, and DAG-based processing [8]. Despite some of the challenges of data lakes, the fact remains that more than 80 percent of all data is unstructured. As more businesses turn to big data for future opportunities, the application of data lakes will rise. Unstructured data like social media posts, phone call recordings, and clickstream activity contain valuable information that cannot be withheld in data warehouses. While data warehouses are strong in structure and security, big data simply needs to unconfined so it can flow freely into data lakes. Many of the differences between data lakes and data warehouses depend on the various structures of data itself [9]. The data lake is a daring new approach that harnesses the power of big data technology and marries it with agility of self-service. Most large enterprises today either have deployed or are in the process of deploying data lakes.

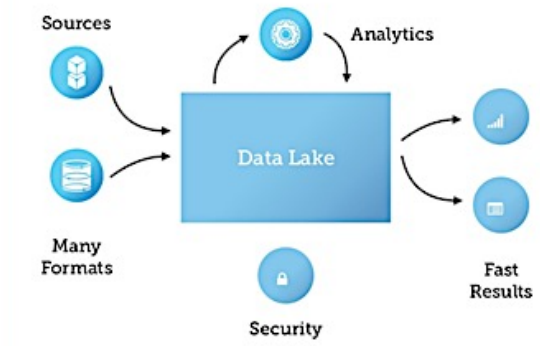


Figure 2. Data lake in Big Data Solutions.

VI. CONCLUSION

Data lakes may be flexible, scalable, and quick to load, but it does come at a price. All data is loaded from source systems and No data is turned away. Data is stored at the leaf level in an untransformed or nearly untransformed state. Data is transformed and schema is applied to fulfill the needs of analysis. Data lakes are a fairly new concept in IT, which means some of the tools are still working out the security kinks. One of these kinks is ensuring only the right people have access to sensitive data loaded into the lake[10]. To conclude data lake is a large data storage repository that holds data in its native format until it is desired. And in simple data lake is the evolution of an Enterprise Data Warehouse (EDW) into an active repo for structured, semi-structured, and unstructured data that retains all features against which we can run all our data analyzing and process. The other way to define data lake is formed by the joining NoSQL & Hadoop. It's the primary landing zone for disparate sources like clickstreams, weblogs, sensor data, etc. Data lake helps business to take more holistic business decisions [11]. Data Lake will continue to be popular with Big Data and Internet of Things streaming data from various connected devices. The still emerging concept will need further maturity, security and organization, before it will be embraced completely by businesses.

ACKNOWLEDGMENT

This Research was supported by, Faculty of Applied Mathematics and Control Processes, Department of Computer Modeling and Multi processor system of Saint-Petersburg State University. We thank our colleagues from our Department who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] Thurein Kyaw Lwin, Alexander Bogdanov, Alexander Degtyarev , Vladimir Korkhov // BIG DATA AS THE FUTURE OF INFORMATION TECHNOLOGY// The 8th International Conference "Distributed Computing and Grid-technologies in Science and Education" Laboratory of Information Technologies (LIT) of the Joint Institute for Nuclear Research (JINR) ., 10 - 14 September 2018. Vol-1787, ISSN: 1613-0073, P 25.

[2] Andy Patrizio // What is a data lake and Flexible big data management explained// InfoWorld Journal SEP 24 2018.

[3] Devin Pickell, // what is a data lakes and why it is essential for Big data// https://learn.g2.com/what-is-a-data-lake

[4] Tyndal Susan. Great data: all you need to know. PC Week / RE, 2012, No. 25 (810). URL: http://www.pcweek.ru/idea/article/ detail.php? ID = 141962 (reference date on July 10, 2014)

[5] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[6] Chris Campbell, // Top Five Differences Between Data lakes and Data warehouses// Ebook: Data Lakes in a Modern Data Architecute,.// https://www.blue-granite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses

[7] Data lakes and big data analytics: the what, why and how of data lakes //https://www.i-scoop.eu/big-data-action-value-context/data-lakes// unpublished.

[8] Shantha Kumari.,// Data Lake is it the Future for Big Data// Sysfore IT Blog January 20 2016.

[9] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[10] A. Klimentov et al. A Roadmap for HEP Software and Computing R&D for the 2020s. arXiv:1712.06982 [physics.comp-ph], Dec 2017.

[11] MLassnig, A. Klimentov, K. De et al. ATLAS & Google — The Data Ocean Project, 23rd International Conference on Computing in High Energy and Nuclear Physics, CHEP 2018, Sofia, Bulgaria, 9–13 Jul 2018.

★★★