International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Data Lake: A Centralized Repository

Karuna Thomas[1], Praseetha S Nair[2]

[1]*Student, Department of CSE, Musaliar College of Engineering and Technology, Pathanamthitta, Kerala*
[2]*Assistant Professor, Department of CSE, Musaliar College of Engineering and Technology, Pathanamthitta, Kerala*

Abstract - *Data never gets static. New data's are added and existing data's are modified. With the rapid growth of unstructured data, which represents a huge percentage of overall data which is termed as big data. The Big Data is not only about massive data capture and storage but also about combining the existing data with newly arrived data. To meet the latent advantages of Big Data successfully, an enterprise needs perfect infrastructure in place to obtain, store, merge and enrich huge amounts of unstructured raw data. It should also have the ability to perform analytics on these huge volumes of data, near-real-time analysis, batch processing. The Data Lake concept is proposed to address these business needs in an effective manner. It is one of the data capture and processing capability empowering concept for Big Data analytics. It is a repository for storage that can store large quantities of structured, semi-structured, and unstructured data. Data Lake is massive, easily accessible, flexible enough and scalable. It originates from the field of business. Many companies are establishing data lakes to make it easier to access their vast data stores.*

Key Words: Big Data, Big Data analytics, Data Lake

1. INTRODUCTION

The data is too voluminous, moves too fast, the structures of existing database architectures make it impossible to manage, with time Big Data is growing exponentially. Therefore, there must be an alternative way of processing those data. Big data includes the data that exceeds the processing capacity of the traditional databases within an acceptable time and value. Big data is a field that interacts with ways to analyze, systematically retrieve information from, or otherwise address data sets that are too large or complex for traditional applications software for data processing. Data with many cases provide greater statistical power while data with higher complexity i.e. with attributes or columns may result in higher false discovery rates. Big data challenges include collection of data, storage, analysis, search, share, distribute, visualize, querying, update and data privacy.

1.1 Characteristics of Big Data

- **Volume:** It refers to the size of the data which is enormous.
- **Variety:** Variety refers to heterogeneous sources and the type of data, structured, semi-structured and unstructured.
- **Velocity:** It is the speed at which the data are being created or processed for the analysis result.
- **Veracity:** It is the extended definition for big data, which refers to the noises and abnormalities in data, data quality and the data value.

Hadoop [4] is an open-source framework that enables to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to extend from single servers to thousands of machines, each offering local storage and computation which provide massive storage for any kind of data. MapReduce is a parallel programming model for writing distributed applications for efficient, reliable and fault-tolerant processing of large amounts of data on large clusters of commodity hardware. It generates big data sets with a parallel, distributed algorithm on a cluster. The program MapReduce runs on Hadoop. MapReduce involves two task: Map and Reduce. A data warehouse is a system used for storing, reporting, data analysis and is considered a core component of business intelligence. It is the core repositories of integrated data from one or more diverse sources. It stores current and historical data in single source that are used for creating analytical reports. In Big Data era a new term was introduced, called Data Lake. Data Lake's intention is to manage every data produced in an organization in order to give a more valuable insight at a finer level. The Data Lake is a data-centered architecture that features as a repository that can store vast amounts of data in different formats. The Data Lake concept intends to challenge reliable, conventional data warehouses facilities to store complex heterogeneous data. The idea was initiated initially by James Dixon, chief technology officer at Pentaho in 2010 [2]. If a data warehouse or data mart tends to be cleaned and ready as a bottle of water for consumption, then Data Lake can be considered as a large water body of data in a natural state that is cleaned for ready use. A Data Lake stores information which is different in type while ignoring almost everything. Data's generated in an organization are of the different types, structures, or formats will be stored in their original forms in Hadoop clusters or other similar framework. When the data needs to be used by parts of the organizations, that stored data will be loaded and transformed as required by parts of that organization. Due to these Data Lake seem to be challenging to traditional ways of storing data i.e. data warehouse and data marts.

2. CONCEPT

Data Lake is an efficient, data-driven model for large-scale capturing of a wide range of data types. It is optimized for rapid ingestion of raw, detailed source data and for the exploration, analysis and processing of such data by operations. Data Lake is an approach enabled by a massive low-cost technology using data repository which enhances the capture, refinement, archiving and exploration of raw data within a company. A data lake may contain data in different format i.e. raw, unstructured, or semi-structured data where most of the data may be of unrecognized organizational value. Data Lake performs various modifications to the data. When it is to be used the data structure will be defined. The basic concept of Data Lake is all data that the organization emits will be stored in a single data source. The data is stored at the lake in their original format. Complex pre-processing and data-loading transformation to data warehouses will be eliminated in Data Lake. It can also reduce the costs of data ingestion. Once the data is placed in the lake, it is available in the organization for analysis by all employees of that organization. Data Lake is a changeover from schema-on-write to schema-on-read. It draws more attention from business field. The data lake concept is closely tied to Apache Hadoop and its open source project ecosystem. It is becoming popular because it provides a cost-effective and technologically feasible way to tackle the big data challenges. The Data Lake is being realized by organizations as an evolution from their existing data architecture.

2.1 Characteristics of Data Lake

- **Consolidation:** One of the principles of a data lake architecture is the centralisation of data silos. This centralisation brings number of benefits, including being easier to manage and govern.
- **A data lake is a collection of data, not a platform for data:** A data lake is a collection of data or multiple collections in different data format that is usually managed on Hadoop.
- **Collect and Store All Data at Any Scale:** The data lakes allow data to be collected and stored on any scale. Cloud object storage services offers virtually unlimited space at very low cost in terms of the Big Data. Business data can be collected in real time, using high velocity streaming technology.
- **Locate, Curate, and Secure Data:** Having a centralized data lake makes it easier to keep track of data stored, who has access to the data, what kind of data are stored and what it is being used for.
- **A data lake is a data-driven design pattern:** The data lake is a data-driven design pattern, business data hubs, and logical data warehouses which captures wide range of data.
- **Increased Agility:** Having a centralized data lake enables companies to innovate in new ways of data processing. Can introduce new use cases without the need to reengineering the architecture.

2.2 Capabilities of Data Lake

In the Big Data era there are always new types of data that enterprise needs to capture and analyse [5]. The first example of data lakes was created at Internet Company to handle web data and then people find out many other data suit sorts. Therefore data lakes became popular in the ecosystem of enterprise data management.

The data lake supports the following capabilities:

- To capture and store raw data at scale for a low cost: Because the volume of data continually growing so that the cost of data store became more important than before.
- To store many types of data in the same repository: There are structured data from traditional DBMS, multi-structured data include multiple attributes that are undefined and multi-media data.
- To perform transformations on the data: The key use case for data lake is to perform pre-processing and ETL (Extract, Transform, Load) transformation of data for further exploration by other system.
- To define the structure of the data at the time it is used: The data lake avoid complex, costly data modeling and data integration effort.
- To perform new types of data processing: The data lake support all the data and all the ways for data processing.

2.3 Data Lake Compared To Data Warehouse

The differences between data warehouses and data lakes are significant [3]. Data Lake and Data Warehouse concepts, structures, and implementation differ in many respects. Data Warehouse with files or folders data storage uses a hierarchical format, the data lake uses a flat architecture. It is a database optimized for analysing the relational data from transactional systems and business line applications. In order to optimize the data structure and schema in advance for fast SQL queries, the results are typically used for operational reporting and analysis. Data is cleaned, enriched, and transformed so that it can act as the single source of truth that users can trust, related from transactional systems, operational databases, and business line applications. Data warehouses have well-defined regulatory and storage capabilities. Higher cost storage used to quickest query results. The quality of data is highly curated data which serves as the central version of the truth. The users are Business analysts. Designed to collect only quality-controlled data and conform to an enterprise data model, the data warehouse can only answer a limited number of questions. A data lake is different, because it stores business line related data, and non-relational data from mobile apps, IoT devices,

and social media. The data or schema structure is not defined when the data is being captured. One can store all your data in the without careful design or the need to know what questions you might need answered in the future. To uncover different perspectives, different types of analysis on your data be performed such as SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used. Contains non-relational and relational data from IoT devices, websites, mobile apps, social media, and enterprise apps. Using low cost storage the results of the query are getting faster. Business analysts, Data developers and Data Scientist are the users. Processing is not performed or a little processing is done for adapting the structure to an enterprise. Data lakes have the greatest advantage of flexibility.

3. DATA LAKE ARCHITECTURE

Data Lake is a repository which is designed to store huge amounts of data in its native format. Much of Data Lake implementation is frequently based on Apache Hadoop, which enables load heterogeneous and voluminous data with low cost. But it does not allow users to process data and does not records any user operations. Variety of data to be stored in the Hadoop Cluster will be extracted from heterogeneous data stores. HADOOP is a widely popular big data tool particularly suitable for large data workload batch processing. There are two key components in Hadoop – HDFS (Hadoop Distributed File System) and MapReduce. HDFS File System handles the single point of failure and scalability by replicating multiple copies of blocks of data into different cluster nodes. MapReduce approach will process all data stored in this data block. Data will be retrieved as a list of pairs with key value i.e. Map Phase. The same data keys will be shuffled, sorted and listed into groups to accomplish the necessary operations i.e. reduce phase. All enterprise-generated data will be dumped into the Data Lake Hadoop Cluster. Later data lakes use stream processing framework such as Apache Spark, Apache Flink, for the real-time load. The required data will be transformed in the query time according to the need of the analytics systems. Saving data from heterogeneous sources with different formats and structures and handling different data velocities (i.e. different processing velocities of big data) requires careful consideration when building data pipes for data transfer to the lake. Data Lake can often include a semantic DB, a conceptual model, and add a context layer to define the significance of data in its interrelation with other data. It can be said that the Data Lake strategy includes the storage of all data types (data variety) from SQL and NoSQL databases and the combination of OLTP and OLAP concepts. SQL databases are used to store the structured data which resides in a fixed field. NoSQL databases are used to store semi-structured and unstructured data (Key-value, Columnar, Document, and Graph Stores). They can however also be used to store structured data. All the transactional data from these databases (Dilute-E) will be stored (Load-L) into the data lake without changing their format. The data in the Data Lake is transformed according to the parts of enterprise system when it is required at the query time. Necessary work for query operations has to be done at the level of the application. Many other architecture were proposed to overcome the drawbacks of Hadoop Data Lake architecture. Such as Data Lake architecture with different zones etc. were proposed.

4. DEPTH OF DATA LAKE

Traditionally data had been stored in data warehouses, pulled from multiple sources, transformed and structured and defined by very specific parameters. Data warehouses were useful because it regulated and trusted the data within them. However, this model does not fit 80 % of the data, because it is considered semi-structured or completely unstructured. This means it does not have a pre-defined data model, or is not organized in predefined way. So, to store this data, instead of having to properly integrate it all into one model, data lakes allowed data to remain the way it is, to be handled at another time.

4.1 Benefits of Data Lake

- Collect and store raw data at low cost like the Hadoop-based data lake.
- Perform transformations on the data.
- Store many data types in the same repository like structured, semi-structured and unstructured.
- Availability of data, Data Lake ensures that in a business sector all the employees has access to the data. This is called data democratization.
- Data Lake takes benefit from large amounts of consistent data and deep learning algorithms to arrive at real-time decision analysis.
- Data Lake offers diverse analytical options and language support. It has Hive / Impala / Hawq supporting the SQL. Additionally, it provides features for addressing advance requirements. For example, PIG to analyze the data flow, can use Spark MLlib for machine learning.

4.2 Challenges of Data Lake

- Data lakes lacks the ability to determine the quality of the data or the result lineage. They have been discovered by other data analysts in the same data lake but cannot provide for later analysts and also a thorough check-up has not taken place.
- Data Lake accepts the data without supervision and management.
- No descriptive metadata or mechanism for the preservation of metadata leading to data swamp.
- Every time data needs to be analysed from scratch.
- Cannot guarantee performance.

- Security and access control as data in the lake can be replaced without contents being monitored.

4.3 Concerns of Data Lake

- **Data lakes are becoming data swamps:** Companies have struggled to use data lakes as more than endless data repositories. The result is that they end up hoarding data in data lakes, but without any structure or organization, so analysts who want to use the data have no idea how to do that. The data just ends up sitting in the lake and is almost never used.
- **Data never put into production:** Mode of data lakes has been that data is allowed to fester in data lakes because of how disorganized they are in most businesses. As a result, the process of extracting signals from it is complex and the data is never fresh enough to actually be put into production, or relevant in real-time. So the data remains in pilot mode in the Data Lake.
- **Asking small, not big data questions:** As well, companies have undermined the value of their data lakes by viewing them and big data in general within the data warehouse's confined perspective. Users simply ask the types of questions they've asked in the past, and fail to recognize how much more powerful data lakes can be. They fail to understand how much more signal can be extracted from big data, because in the past they could only answer questions that they know.
- **Failing to gain added value:** When companies treat their data lakes like updated versions of their data warehouses, their data lakes fail to experience value. Running processes which done on data warehouses on data lakes, such as moving data across clusters to separate data marts or BI servers, creating schemes, or extracting subsets, are all ways in which businesses limit their data lakes worth. These older processes extend the time it takes for data to be processed and analysed. New processes are required at data lakes.
- **A lack of business impact:** An imbalance between the significant investments they have made in data lakes and the relative lack of impact on business decisions that data from the data lake is having. For the insights that data lakes generate to matter, they must drive behavioural changes. For this to happen, businesses have to empower managers to act and make decisions based on data lake analytics. Additionally, businesses need to implement ways to operationalize data into real-time business processes so that data lakes can have a real impact on the end result.
- **An inability to mine data lakes for analytics:** The cause of this problem has been that too often the data is owned by someone other than the analyst, and they must check with them before using it. Companies need to streamline the processes and remove analyst barriers. Furthermore, companies need to ensure that analysts have the tools they need to work with data in the data lake in a manner that works best for them.

5. CONCLUSION

Big Data is data which exceeds the processing capacity of traditional database systems. The data is exponentially growing, and moves too furiously that it does not fit the relational database architecture. Big Data offers organizations tremendous insights. But traditional architectures are not up to the level to take up the challenges with large data such as terabytes and petabytes and even more that pours into the organizations every day. It was thus the introduction of the Data Lake concept. Data Lake is a huge repository which contain different type of data which belonging to an organization in variant schemas and formats. By saving all of this data at one location, data availability and reusability across departments and business units are increased. The Data Lake stores large volumes of data, increasingly. It is used as a multi-tenant service and stores sensitive data, and mainly works on read schema. Data lakes are becoming increasingly central to data strategies in enterprises. Data lakes best address today's data aspects with far greater data volumes and varieties, increased user expectations and rapid globalization of economies.

REFERENCES

[1] Pwint Phyu Khine, Zaho Shun Wang, 2018 Data Lake: a new ideology in big data era https://doi.org/10.1051/itmconf/20181703025.

[2] Keith D. Foote 2018 A Breif History of Data Lake

[3] Tamara Dull, 2017, Data Lake vs Data Warehouse: Key Differences.

[4] Apache Hadoop. http://hadoop.apache.org/. 2016

[5] Huang Fang 2015, Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem.