A Review on Enterprise Data Lake Solutions

Aakash Aundhkar[1], Shweta Guja[2]

[1,2]Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Pune, Maharashtra, India
a.aundhkar7100@gmail.com

To Cite this Article
Aakash Aundhkar, Shweta Guja, "A Review on Enterprise Data Lake Solutions", Journal of Science and Technology, Vol. 06, Special Issue 01, August 2021, pp11-14: .

Article Info
Received: 15.07.2021 Revised: 24.07.2021 Accepted:10.08.2021 Published: 16.08.2021:

Abstract: *Data Lake is a highly flexible storage solution that can store both structured and unstructured data and operates on the schema-on-read approach. It acts as a potential alternative to the current Big Data storage issue. However, it does have certain flaws, such as inadequate authentication and access control. This paper examines a few of the current business Data Lake strategies. Apache Hadoop is generally regarded as the data lake industry standard. Its parallel processing systems ensure high-speed processing of massive volumes of data. Many businesses have attempted to build Hadoop wrappers in order to resolve questions about its raw state and lack of data protection. Platforms including Amazon Web Services (AWS) Data Lake and Azure Data Lake fall under this category. AWS Data Lake provides a more straightforward approach with failsafes to avoid data failure, while Azure Data Lake offers much greater scalability and enterprise-level reliability. Data Lake systems are becoming increasingly common in a variety of sectors, including finance, business intelligence, engineering, and healthcare.*[1]

Key Word*:* *Data lake, Hadoop, Data Analysis*

I. INTRODUCTION

The accelerated growth of technological devices along with easily accessible internet have resulted in unprecedented data volumes and accumulation. It is estimated that the total amount of data collected, consumed and captured in the world will be 74 zettabytes in 2021. These data can be structured, semi structured or unstructured which makes it difficult to manage and process them using the old conventional ways.[1]. Addition to this, today's software environment demands stable, quick and efficient data storage capabilities.

The Big Data is based on the 5 V's [1,2]:

1. Volume : If we consider big data as the pyramid then the volume is the base of this pyramid. It is the amount of data which ranges from a few kilobytes of document to terabyte of raw footage of a movie.
2. Velocity : Many a times velocity is considered more important than volume as it has its own advantages. It consists of the time required for creating, processing and transferring of data.
3. Variety : The type of data received in terms of schema, format ranging from a simple XML to a video or a sms.
4. Veracity : Veracity is equivalent to quality measuring how accurate and clean the data is.
5. Value : Value sits at the top of the pyramid. It refers to the ability of providing maximum information from the available data.

Data lake seems like an assuring solution to this big data conundrum. It is a centralised enormous storage repository which is capable of storing both and structured unstructured raw data at any scale. Data lake uses the schema on read approach which process the data at runtime. Data lake also provides various analytical tools ranging from machine learning algorithm for helping in taking decision to dashboards and visualisation of big data. This is achieved by using SQL and NOSQL approaches along with online analytical processing (OLAP) and online transaction processing (OLTP) capabilities. [1].

Despite all these benefits, there are several risks associated with data lake, one of which is proper security and access management. With the collection of new data everyday, data lake should be able to provide a well secured platform which ensures adequate security.

II. DATALAK

Data lake is a massive, easily available and scalable data storage repository which can practically store any type of data without any predefined schema. Data lake is often considered as a successor of data warehouse.It allows same data to be processed and structured in various ways which is critical while working with unstructured data as there are no well defined methods for data retrieval , manipulation or analysis, and different techniques are typically used. [2] .

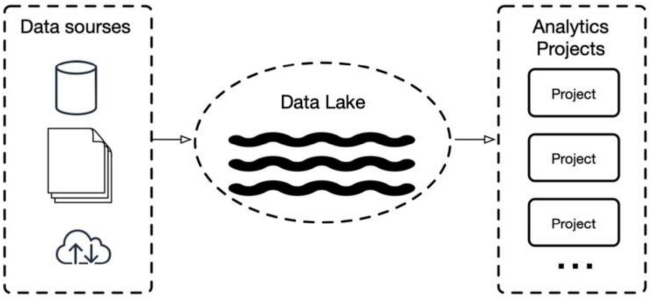

Figure 1: A common view of data lake

Enormous amount of structured and unstructured data coming from multiple sources such as web servers, FTP and IOT gets collected in Data lake. These data can be used multiple times using different techniques according to user requirements.

A. Characteristics of Data Lake :

1. Data can be in relational as well as non relational form obtained from IoT devices, emails, etc
2. Data lake provides faster result using cheaper storage solutions.
3. Data lake can scale as much as it wants by using techniques like HDFS. [3] .

III. STUDY OF EXISTING SYSTEM

A. Hadoop

Hadoop is considered as an industry standard for data lakes and other big data solutions. It is an open source platform which is widely used in data processing and storing huge amount of data.It is capable of processing high amount of data using parallel processing framework. Hadoop is made up of two parts : a storage subsystem called Hadoop's Distributed File System (HDFS) and a processing component known as MapReduce.[4] .

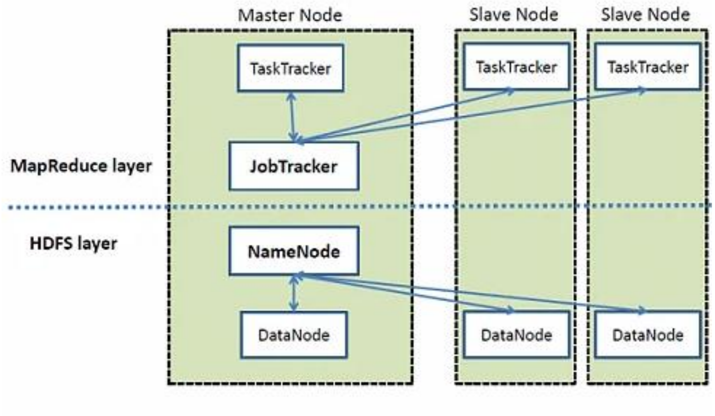

Figure 2: Architecture of Hadoop

HDFS is a distributed storage file system designed to operate on low cost hardware. It stores the data over the cluster of commodity hardware. This whole file system is based on master slave architecture. The master node for data storage is NameNode whereas the master node for parallel processing of data using MapReduce is Job Tracker. The

Hadoop architecture's slave nodes are the Hadoop cluster's other computers that store data and execute complex computations.The Map Reduce approach to data analysis is based on a Google software project from 2004. It distributes the work of processing large amount of data across various hardware units thus resulting in higher efficiency . There are two major steps in Map Reduce. First is Map which is used to parse and filter the data and perform required transformation. The other one is reduce which performs all the necessary operations on each sub group of data. [4] .

B. AWS

Amazon web services have created their own data lake architecture which enables creation of low cost data lake solutions. It uses Amazon Simple storage Service (S3) along with some other core AWS services. [1] .

These offer a variety of advanced features, including smooth integration with conventional big data solutions and creative query-in-place analytics tools that reduce expense and sophistication while eliminating the need for data analysis, transformation, and load. AWS also provides a vigorous security design which involves access policy option which will protect from both internal and external threats.

The Amazon S3 data lake architecture has a data longevity rate of 99.999999999 percent, which puts it way ahead of the competition. The Amazon S3 architecture, in layman's terms, has the potential to efficiently store over 10,000,000 data properties for over 10,000 years.

C. Azure Data Lake

Azure Data Lake is a framework which is adaptable, versatile, efficient, and stable. It can store and analyse a wide variety of data and has been designed for massive workloads with high throughput requirements. It can be accessed via Storm, U-SQL, Hive, and Spark, among other methods. The Azure Data Lake Store (ADLS) and Azure Data Lake Analytics (ADLA) together make up Microsoft's data lake solution.

The Azure Data Lake Service (ADLS) is a rambunctious repository and the first public Platform as a Service (PaaS) on Azure that supports a wide variety of Big Data analytics. The user interface and the modular micro services architecture of ADLS have been significantly improved. [5] .

The ADLA is a decentralised analytics service that can automatically provide services depending on demand. U-SQL is used in ADLA, which is developed with Apache YARN. U-SQL is a distributed query language that incorporates the ease of use of SQL with the computing ability of distributed databases. Azure Data Lake offers two levels of security: authentication via Azure Active Directory (AAD) and data access management via Active Control Lists (ACL). Azure HDInsight, a full-stack Hadoop PaaS, is also part of Azure Data Lake. Famous open-source frameworks such as Apache Hadoop, Kafka, and Spark are available.

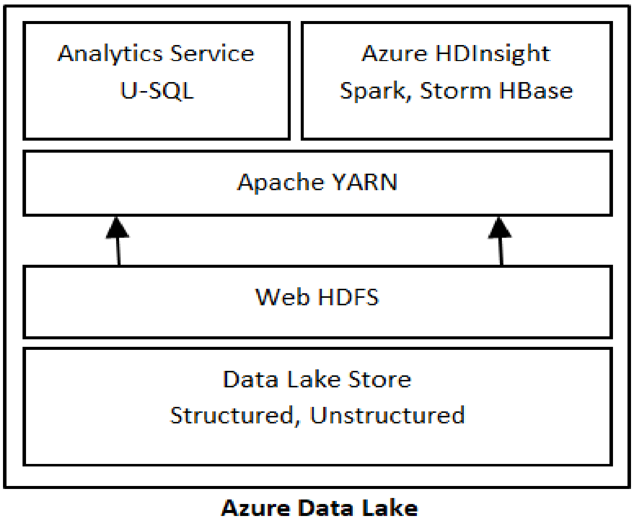

Figure 3: Architecture of Azure Data Lake

IV. APPLICATIONS OF DATA LAKE

A. Healthcare

Healthcare is a field where data is constantly growing. Through reviewing data produced, the quality of health care can be improved. A big data infrastructure is needed to handle large amounts of complex data. A data lake is a consolidated, curated, and stable archive that holds all of the data, both in its raw state and as prepared for review. It's one of the easiest ways to handle complex data integration. [6] .

B. Banking Data model

Industry models provide an outstanding ability to accelerate growth by introducing best practises and guidelines. A banking model for data warehouses is one such example. Banking Data Warehouse is a set of business and technical models that help to facilitate the development of enterprise vocabularies, data warehouses, data lakes and analytics technologies based on financial-services business requirements.[7] .

C. Business Intelligence Data Analysis

Business intelligence is a collection of resources and tools that aids an organisation's decision-making process by integrating and analysing business data. They're progressing to larger-scale applications. The emergence of increasing volumes of unstructured data is suggested by trends in the implementation of such networks. Data Lakes may be used in conjunction with data warehouses in a Business Intelligence architecture to address the industry's growing demands. [8] .

V. Literature review

The authors in [1] gives a brief idea about the various data lake solutions available in the market for the consumers. It includes Hadoop, AWS and Azure data lake architecture. Apache Hadoop is considered as an industry standard for Data lakes. Both Amazon Web Services (AWS) data lakes and Azure data lake have constructed wrappers around Hadoop in order to solve concerns about its raw data and lack of data security. In [2] , the ideas of data warehouse and data lake are explained by the author along with its positives and negatives.

In [3], authors present a quick overview about how Data lake can be used to address the challenges faced by the smart grid data management systems. The author has explained data lake lambda architecture for analysing the data. The emerging big data concepts : Data lake and fast data is explained in [4] by the authors. While both data lakes and big data are able to store and process more data, fast data gives real time insights rapidly based on limited data.

VI. Conclusion

In today's data-driven environment, the need for reliable and optimal large-scale data storage is growing. Data lake technologies like AWS and Azure are showing great potential in delivering successful end-to-end firmware to satisfy this need for large-scale storage. Several industries, including Cloudera, are also working to build newer and more stable architectures for data lakes. One of the most important issues to address is encryption and data protection. In today's increasingly hands-free and data-centric environment, protecting the security of private information, as well as information confidential to businesses, is essential.

References

[1]. Tanmay Sanjay Hukkeri, VanshikaKanoria, Jyoti Shetty, A Study of Enterprise Data Lake Solutions, International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 05 | May 2020

[2]. SnezhanaSulova ,The Usage of Data Lake For Business Intelligence Data Analysis, Conference Paper · October 2019

[3]. Surabhi DHegde, Ravinarayana B, Survey Paper on Data Lake, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[4]. Amra. Munshi, and Yasser Abdel-Rady I. Mohamed, Data Lake Lambda Architecture for Smart Grids Big Data Analytics , IEEE, date of publication July 23.

[5]. Raghu Ramakrishna, Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics.

[6]. Ekta Maini, BonduVenkateswarlu, Data Lake-An Optimum Solution for Storage and Analytics of Big Data in Cardiovascular Disease Prediction System,IJCEM International Journal of Computational Engineering & Management, Vol. 21 Issue 6, November 2018.

[7]. DarkoGolec, Data Lake Architecture for a Banking Data Model, ENTRENOVA 12-14, 2019.

[8]. Marilex Rea Llave, Data lakes in business intelligence: reporting from the trenches, Procedia Computer Science 138 (2018) 516–524.