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

# A Study of Enterprise Data Lake Solutions

**Tanmay Sanjay Hukkeri[1], Vanshika Kanoria[2], Jyoti Shetty[3]**

*[1]Dept. of Computer Science and Engineering, RV College of Engineering, Bengaluru, India*
*[2]Dept. of Computer Science and Engineering, RV College of Engineering, Bengaluru, India*
*[3]Prof, Dept. of Computer Science and Engineering, RV College of Engineering, Bengaluru, India*

---

**Abstract -** *Data Lake is a highly scalable repository capable of storing structured and unstructured data and uses schema on-read approach. It serves as a promising solution to today's Big Data storage conundrum. However, it also possesses a few shortcomings including proper security and access management. This paper provides a study of some of the existing enterprise Data Lake solutions. Apache Hadoop is widely considered to be a standard for data lakes. It ensures high speed processing of large amounts of data through its parallel processing frameworks. Many enterprises have sought to develop wrappers over Hadoop to address concerns over its raw state and lack of data security that addresses these concerns. This includes platforms like Amazon Web Services (AWS) Data Lake and Azure Data Lake. AWS Data Lakes provide a much simpler solution along with fail safes to prevent data loss while Azure Data Lakes boast about much larger scalability and enterprise level security. Such Data Lake solutions are growing increasingly popular across several fields like banking industries, business intelligence, manufacturing industries and healthcare. It also plays a significant role in Industry 4.0.*

***Key Words***: **Big Data, Data Lake, Hadoop, Amazon S3, Azure**

## 1. INTRODUCTION

The vast multitude of unstructured data is constantly increasing, and is expected to be around 44ZB by mid-2020. In addition, today's rapidly evolving software world also demands faster response time, larger and more efficient storage capacities and improved dynamic response.

The fundamental concerns of a Big Data platform today are [1]:

1. Volume: The size of the data, ranging from a few megabytes of data such as document stores, to several petabytes of data for storing media.

2. Variety: Diversity of data, in terms of format, schema etc.

3. Velocity: Speed of data creation and processing

4. Veracity: Correctness and Accuracy of the data source.

5. Value: Extracting information from data

Data Lakes serve as a promising solution to this Big Data storage conundrum. A data lake is a massively scalable storage repository that is capable of holding vast amounts of raw data in an unstructured, heterogeneous manner [2]. The key advantage of a data lake lies in its capacity to store data in its native form, using a schema on-read approach [3] to process data at runtime. A data lake is also expected to provide the ability to perform analytics, batch processing and real-time analysis on large volumes of data in an efficient manner. This is achieved by combining the benefits of SQL and NoSQL database approaches, supplementing them with Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP) capabilities. The data elements are tagged in the lake with an id, and often carry additional metadata. Data lakes improve the capture, refinement, archival and exploration of raw data within an enterprise.

In spite of all these advantages, there are a few risks associated with data lakes as well. The first is that of proper security and access management [4]. With the growing importance of data security and privacy in today's world, a well-designed data lake architecture needs to pay considerable attention towards ensuring a highly secure storage platform. One other concern is that of performance. With the additional application specific processing required using the "schema on-read" approach, care should be taken to ensure that this does not affect performance severely. In general, however, the benefits of a well-designed, secure data lake far outweigh any potential disadvantages.

The paper is divided as follows. Section 2 provides an in-depth analysis on the differences between Data Warehouses and Data Lakes. Section 3 then provides a deep-dive on the various enterprise data lake solutions. Section 4 discusses the merits of these solutions and compares and contrasts them. Section 5 details some of the existing applications of data lakes. Section 6 provides a conclusion to the study conducted.

## 2. DATA WAREHOUSES VS DATA LAKES

Data Lakes are generally regarded as the successor to Data Warehouses, providing enhanced capability to store unstructured data. Table 1 depicts the fundamental differences between data lakes and data warehouses [5].

**Table -1:** Differences between Data Warehouses and Data Lakes

| Dimension | Data Warehouses | Data Lakes |
|---|---|---|
| **Data** | Has to be stored in a structured, processed form | Can store even unstructured, raw data |
| **Schema** | Schema on-write | Schema on-read |
| **Scalability** | Scalable to large volumes with moderate cost | Scalable to extremely large volumes with very low cost |
| **Architecture Design** | Hierarchical (with folders and files) | Flat (each data element has its own ID tag) |
| **Complexity** | Complex Joins | Complex Processing |
| **Cost/Efficiency** | CPU/IO is user efficiently | Storage and Processing cost is very low |
| **Agility** | Less agile/flexible, restricted to rigid configuration | Configuration is highly agile and flexible |

## 3. STUDY OF EXISTING SYSTEMS

### 3.1 Hadoop

Apache Hadoop is widely considered as the standard for data lakes and other big data applications. It provides a highly scalable and parallel processing framework that ensures efficient and high-speed processing of a large amount of data. Additionally, its data replication process ensures that no data is lost as it is replicated across the clusters. One key advantage of Hadoop lies in it being open source, which allows immense flexibility in choosing application specific and highly customizable datasets.

The two main components of the Hadoop architecture are the Hadoop Distributed File System (HDFS) and the Map-Reduce Processing algorithm. These are elucidated below.

1. Hadoop Distributed File System (HDFS): HDFS serves as the Hadoop native file system that is capable of storing an enormous amount of data. It stores the data over a cluster of commodity hardware (even up to over 1000 servers). The file system is based on the master-slave architecture design. Additionally, it follows the policy of "Steaming access patterns", which is a write-once, read any number of times approach [6].
   By using its own file structure architecture, Hadoop allows access to data from any computer bearing a supported OS [7]. Additionally, the HDFS architecture supports high performance across large datasets by scaling horizontally, i.e.; processing on a large number of servers via the Map Reduce approach. It also tackles the problem of single point of failure by replicating data blocks across several nodes in the cluster. The general block size in HDFS is 64 MB in size, but this can be extended up to 128 MB based on application requirements. Figure 1 depicts the types of services in HDFS architecture.

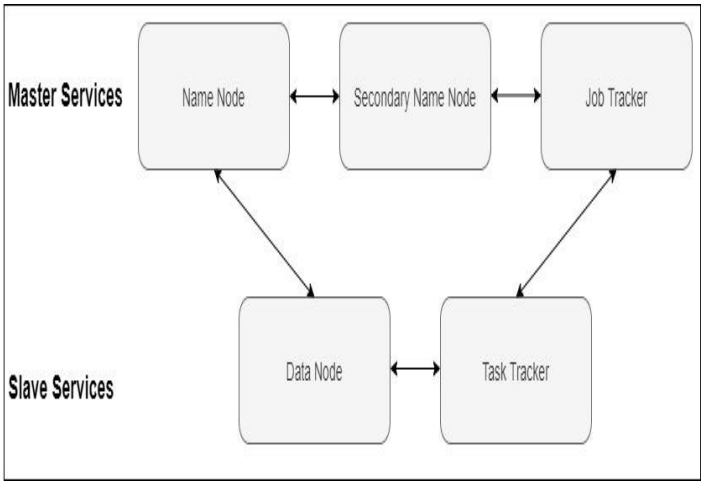


**Fig – 1:** HDFS services

2. Map-Reduce: The Map-Reduce approach towards processing is based on a software project developed by Google in 2004. It dictates an efficient approach towards processing enormously large volumes of data by distributing it across several hardware units and processing in parallel [8]. The two major steps are elucidated below.
    a. Map: This function is used to filter or parse the data and apply the necessary transformations. The data is shuffled and sorted into various groups.
    b. Reduce: The pre-processed data from the Map stage is fed to this stage, which then performs the necessary operations on each sub group of data. This set of processed data is then returned to the application.

    Figure 2 demonstrates the Map-Reduce algorithm for a simple word count operation.

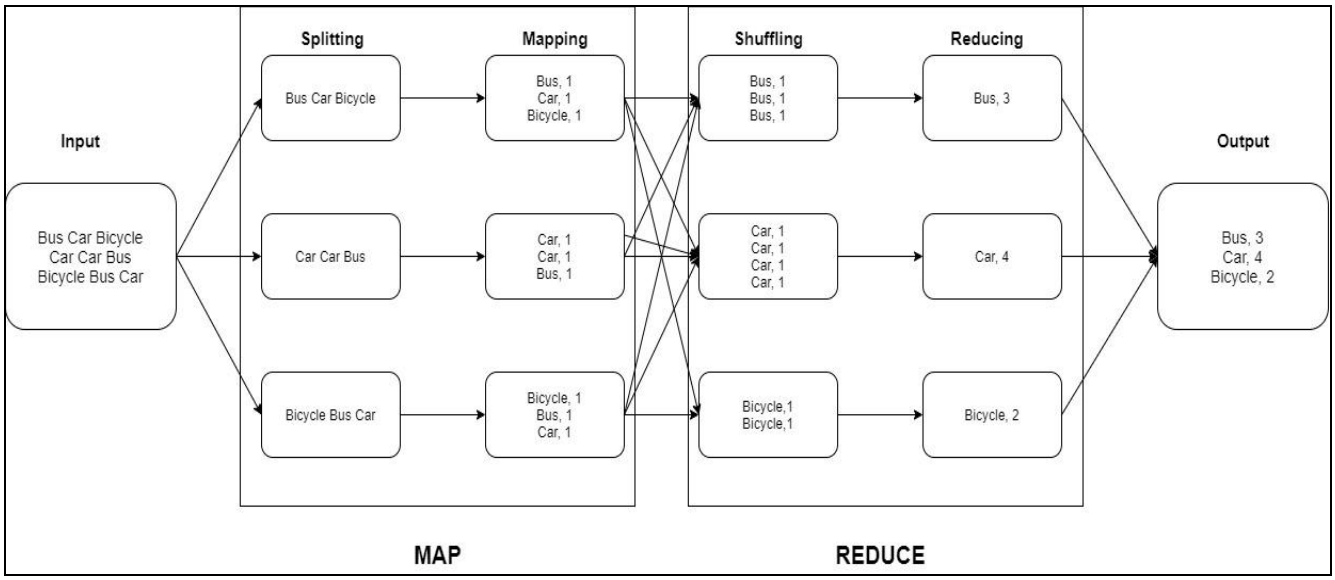


**Fig- 2:** Word Count Example for Map-Reduce Algorithm

While using the Hadoop Architecture and its distributed file system, one key concern is that of data security and privacy. This has three major facets, authentication (access control to the clusters), authorization (restricting access to explicit data) and audit (logs of various users and their actions).

### 3.2 AWS

AWS has developed a data-lake architecture that allows the development of cost-effective data lake solutions via the use of Amazon Simple Storage Service (S3) and other supporting services. These provide a wide array of specialized features such as seamless integration with traditional big data tools, as well as innovative query-in-place analytics tools that significantly eliminate cost and complexity by removing the need for processes such as data extraction, transformation and load. Amazon S3 also provides a novel bucket-versioning facility to store data in a manner that prioritizes protection against data loss [9]. Additionally, the AWS data lake architecture is also supported by a robust security design involving access policy options, which help protect against both internal and external threats.

The Amazon S3 data lake architecture boasts of a 99.999999999% data durability percentage, putting it well ahead of its competitors. Colloquially speaking, the Amazon S3 architecture has the capacity to reliably store over 10,000,000 data assets for over 10,000 years [10]. Additionally, the S3 architecture also provides virtually unlimited scalability, facilitating shifts in storage capacity from something as low as a few gigabytes, to even a large number of petabytes.

Some of the key features of the Amazon S3 data lake include:

- Separation of storage from computation and processing. As opposed to traditional data lake architectures that enforce tight coupling, the loose coupling provided by Amazon S3 facilitates cost and workflow optimization.
- Robust frameworks to secure, protect and manage data.
- Centralized data architecture that ingests data from a variety of sources into a centralized platform.
- Easy integration with current third-party data processing tools, with flexible support for future tools as well.
- Provision for a complete and cohesive set of standardized Application Program Interfaces (APIs) for optimal data querying and processing.

- Integration with services such as Amazon Kinesis Firehose that provides key transformation functions such as data batching, compression, lambda functions and encryption.

### 3.3 Azure Data Lake

Azure Data Lake is a flexible, highly scalable, reliable and secure system. It accommodates the storage and analysis of a wide range of data and has been optimized for large workloads requiring high throughput. It can be accessed through a variety of ways including Storm, U-SQL, Hive and Spark. Azure Data Lake Store (ADLS) and Azure Data Lake Analytics (ADLA) together form the data lake solution offered by Microsoft. Figure 3 depicts a high-level architecture diagram of Azure Data Lake.

The ADLS is a hyper-scaled repository and the first public Platform as a Service (PaaS) on cloud which supports a wide range of Big Data analytics on Azure. ADLS has made significant enhancements in terms of the user experience and its modular microservices architecture.[11] These microservices provide features like low-latency small appends and a service to mitigate the noisy neighbor problem. It has a tiered storage system that allows users to store their data in any combination to achieve the optimum balance between cost and performance. ADLS has clusters which have a much larger number of nodes (10 times) as compared to Hadoop. Therefore, ADLS provides larger scalability. There are no fixed limits imposed on the file size with each file containing potentially up to petabytes of data. The large file systems are supported through an underlying Replicated State Library - Hekaton (RSL-HK) ring infrastructure, a combination of Paxos and a transactional in-memory block data management design developed by Microsoft.

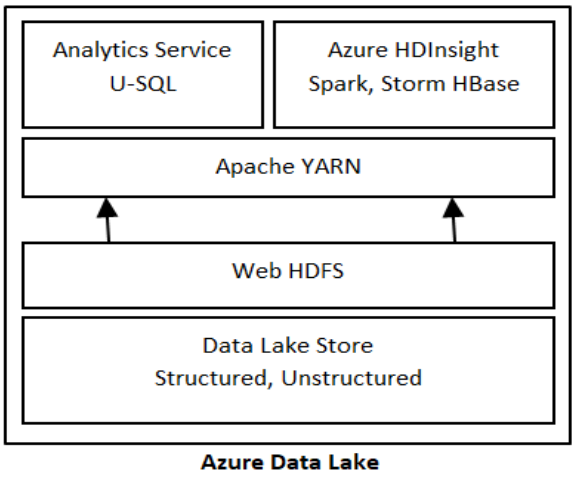


**Fig-3:** Azure Data Lake

The ADLA is a distributed analytics service that can provide resources dynamically based on requirement. It can process petabytes of data with a pay-as-you-go model that is very cost effective for large workloads as well as short term jobs. ADLA is built using Apache YARN and includes U-SQL. U-SQL is a distributed query language that combines the simplicity of SQL with a high processing power.

Azure Data Lake provides a two-fold security using Azure Active Directory (AAD) for authentication and Active Control Lists (ACL) to manage access to the data. Additionally, ADLS also supports OAuth 2.0 protocol for authentication.

Azure Data Lake also includes Azure HDInsight which is a full-stack Hadoop PaaS. It provides popular open-source frameworks like Apache Hadoop, Kafka and Spark.

## 4. DISCUSSION

The above study went into an analysis on the various modern-day enterprise solutions for data storage. The current section provides an analysis on their suitability in modern day big data solutions.

One of the key challenges associated with using a vanilla Hadoop platform is the level of complexity associated with it. As a consequence of its raw state, as well as its data analysis tools which can be extensively complicated, Hadoop can be challenging to work with. Thus, modern day firms often move towards developing custom wrappers over the generic platform. By developing these software packages, the process of using Hadoop in an end-to-end manner is simplified.

One such solution which utilizes the above approach is the AWS Data Lake solution, which uses Amazon S3 as a storage medium instead of the HDFS. This, along with reduced complexity, provides several other advantages as well. Key among these is the bucket versioning capability, which puts in place fail safe measures to prevent data and project loss, as well as isolates different steps of the analytics project from each other, to facilitate optimal separation of concerns. Additionally, Amazon S3 also provides supporting personal data loss protection schemes as well. In contrast however, one of the key disadvantages of this platform is it performs slower when compared to the HDFS performance. In general, however, the advantages outweigh the shortcomings, and the speed can be optimized by using well-tuned data analytics algorithms.

Azure Data Lake is another solution designed to enhance the end-to-end user experience. The ADLS provides various specialized microservices to achieve this. The files stored across the tiered storage are managed by the system which automatically takes care of the compliance and security also. This resolves the weaknesses of the current cloud approaches like the overhead of moving the data from

**Table -2:** Comparison of AWS, Hadoop, Azure

| Sl. No | Feature | Hadoop | AWS | Azure |
|---|---|---|---|---|
| 1. | Unstructured Data Storage | Supported | Supported | Supported |
| 2. | Scalability | Horizontal Scaling | Horizontal Scaling enhanced by AutoScaling<br>Manual Vertical Scaling | Horizontal Scaling enhanced by AutoScaling<br>Vertical Scaling via Webhooks |
| 3. | Security | Supports basic access control | Data Protection using bucket versioning, security using access control, ACL | Enterprise level two-fold security using ADLA and ACL |
| 4. | Data Visualization Capability | Not Supported | Supported through QuickSight tool | Supported through PowerBI tool |
| 5. | Data Durability | Lower durability (<99.9% a year) | 99.999999999 % objects retained per year (very high retention) | 99.999999999 % durability through local redundant storage |
| 6.. | File Size Capacity | No maximum file size | Object size restricted to 5 GB | Unlimited |
| 7.. | Usage | Open source Development | Multi-purpose (Database backup, storage logs) | Designed for Analytics |
| 8.. | Complexity of Development | Challenging | Simplified with the use of custom enterprise wrappers | Simplified with the use of custom enterprise wrappers |
| 9.. | Fault Tolerance | Erasure Coding used (high tolerance) | Regions and Availability Zones, Amazon Machine Image, AutoScaling, Elastic Load Balancing | Erasure Coding |
| 10. | Performance | High throughput<br>Low Latency<br>Read - 350 mbps/node<br>Write- 200 mbps/node | Low throughput<br>High Latency<br>Read - 120 mbps/node<br>Write- 100 mbps/node | High throughput<br>High Latency<br>Improves with more data |
| 11, | Price | Free to Use (Open Source) | Per hour charge with rate dependent on hardware capacity ($0.8560 for a On-Demand / Windows / General Purpose / 4 CPUs / 16 GB Memory) | Per hour charge with rate dependent on hardware capacity ($0.5970 for a On-Demand / Windows / General Purpose / 4 CPUs / 16 GB Memory) |

inexpensive storage to compute layers and running a file manager to locate a file. The underlying RSL-HK Ring Infrastructure significantly simplifies persistent metadata state management and fault tolerance, thus increasing the scalability. The ADLS clusters support 10 times the number of nodes in Hadoop. A storage provider abstraction exposes a small set of operations that allows users to add new tiers through different providers (HDFS, Cosmos) to optimize their cost-performance tradeoff. One disadvantage is that the storage costs of Azure are more than AWS, however, it provides more functionalities like the Data Factory which can combine data from diverse sources. Table 2 provides a comparative analysis of Hadoop, AWS and Azure.

## 5. APPLICATIONS OF DATA LAKES

### 5.1 Banking data model

Banking Data Warehouse is a family of business and technical models that accelerate the design of enterprise vocabularies, data warehouses, data lakes, and analytics solutions, driven by financial-services business requirements (IBM Ireland, 2006).[12]

Data Lake architecture is tailor made for businesses which expect substantial data growth with high speed data generation. The businesses with uncertainty of data as well as different forms of data would profit from leaning towards data lakes.

### 5.2 Healthcare

Healthcare is a field in which data is increasing every day. The quality of healthcare can improve by analyzing data generated. The processing of high volume and diverse data

requires a big data platform.[13] Some findings have shown that Apache Hadoop environment provides error detection and easy scalability. However, it does not support stream processing. Apache Spark provides a very quick response time while Apache Flink grants multiple loads of data. It also has a complex fault tolerance mechanism.

### 5.3 Business Intelligence Data Analysis

Business intelligence is a combination of services and software that support the decision-making process of an organization through integration and analysis of business information. They are evolving to large scale systems. Trends in the development of such systems suggest the presence of increasing amounts of unstructured data.[14] Data Lakes can offer a possible solution to the increasing demands of the industry and could be used along with data warehouses in a Business Intelligence architecture. However, there are certain challenges in using Data Lakes for such systems.[15] These challenges include data retrieval, analytical skills and data management among others.

## 6. CONCLUSION

The need for efficient and optimal large-scale data storage is ever increasing in today's data driven world. To meet this demand for large scale storage, data lake solutions such as AWS and Azure are showing excellent promise in providing effective end to end firmware. Additionally, several industries such as Cloudera continue to work towards developing newer and more robust frameworks for data lakes. In doing so, one of the key concerns to be tackled include security and data privacy. In this increasingly hands-free and data centric world, ensuring the protection of private information, as well as information proprietary to enterprises, is fundamental. Ensuring a data-secure storage system with robust access control measures will be a significant factor in developing effective data lake solutions.

## REFERENCES

[1] Surabhi D Hegde, Ravinarayana B, Survey Paper on Data Lake, International Journal of Science and Research (IJSR), 2016.

[2] Pwint Phyu Khine, Zhao Shun Wang, Data lake: a new ideology in big data era, ITM Web of Conferences 17, 03025, 2018.

[3] Natalia Miloslavskaya and Alexander Tolstoy, Big Data, Fast Data and Data Lake Concepts, 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016.

[4] Ms. S. Divya Meena, Ms. S. Vidhya Meena, Data Lakes - A New Data Repository for Big Data Analytics Workloads, International Journal of Advanced Research in Computer Science, 2016.

[5] CITO Research, Putting the Data Lake to Work, A Guide to Best Practices, 2014.

[6] Aditya Gupta, Gaurav Aggarwal, Akash Kumar, Hadoop, HDFS and Map Reduce - A Review, International Journal of Scientific & Engineering Research, 2018.

[7] Ceyhun Ozgur, Jeffrey Coto,David Booth, Usage of Hadoop and Microsoft Cloud in Big Data Analytics, AIMS International Journal of Management, 2019.

[8] Vinayak Pujari, Dr. Yogesh K. Sharma, Rohan Rane, A Review Paper on Big Data and Hadoop, International Journal of Advance and Innovative Research, 2020.

[9] AWS, Building Big Data Storage Solutions (Data Lakes) for Maximum Flexibility, 2017.

[10] Valerio Persico, Antonio Montieri, Antonio Pescape, On the Network Performance of Amazon S3 Cloud-storage Service, 2016 5th IEEE International Conference on Cloud Networking (Cloudnet), 2016.

[11] Raghu Ramakrishnan, Baskar Sridharan, John R. Douceur, Pavan Kasturi, Balaji Krishnamachari-Sampath, Karthick Krishnamoorthy, Peng Li, Mitica Manu, Spiro Michaylov, Rogério Ramos, Neil Sharman, Zee Xu, Youssef Barakat, Chris Douglas, Richard Draves, Shrikant S Naidu, Shankar Shastry, Atul Sikaria, Simon Sun, Ramarathnam Venkatesan, Azure Data Lake Store:A Hyperscale Distributed File Service for Big Data Analytics, SIGMOD '17: Proceedings of the 2017 ACM International Conference on Management of Data , 2017.

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

[13] Nazari E, Shahriari MH, Tabesh H, Big Data Analysis in Healthcare: Apache Hadoop, Apache spark and Apache Flink, Front Health Inform, 2019.

[14] Snezhana Sulova, the Usage of Data Lake for Business Intelligence Data Analysis, International Conference of Information And Communication Technologies In Business And Education, 2019.

[15] Marilex Rea Llave, Data lakes in business intelligence: reporting from the trenches, International Conference on Enterprise Information Systems, 2018.