[1]Wasudeo Purushottam Rahane

[2]Pramod D. Patil

[3]Rajesh D. Bharti

Examine Heuristic Data Lake Management Using AWS: A Big Data Handling Approach

Abstract: - In this era of technology, the most valued asset can be 'Data'. With the increasing number of data, the value of it keeps increasing. Data storage and data manipulate for to achieve some particular goals or business requirements increasing in number and storing it has become a complex and tedious task. With the use of some advanced technologies like hadoop, it simplified the data storing process, but due to rapid development and excessive use of AI and ML, tons of data is collected. The quintessence is to ascertain an extra cost effective storage alternative. This paper provides with an effective solution to store data over the cloud with numerous benefits over traditional data storage methods by developing a data lake using AWS a Cost Effective Data Lake Management algorithm (CEDLMA). Furthermore, the functionalities of data lake include managing and storing sorted as well as unsorted data, gathering various analytics from the data lake as per business requirements. Proposed work is evaluated with AWS's IAM and S3 services.

Keywords: Data Lake, Data Storage Techniques, Big data, Data Lake, AWS, IAM, S3.

I. INTRODUCTION

To introduce the concept of Data-lake [1] let us understand a following scenario. Envision a data lake as a huge virtual cloud storage where 'n' number of users can upload / download data and later use it for processing it to extract useful results or conclusions. Talking more about the traditional methods for storing data i.e. data warehouses they work fine until the data is predefined and desperate. However, with the modernization in technology and excessive use of Artificial Intelligence and Machine Learning, the number of raw and unfiltered data is increasing [1][2]. To cope up with this situation data lakes prove to be an easy and efficient means of data storage. Major issues such as frequent data loss, low data quality [3], high data storage cost are addressed in data lakes. While the data keeps generating, organizations need a way to use their data part from just storing into an effective tool for enhancing their businesses [4]. Data in S3 data lake has a data longevity rate of 99.9999999 %, which puts it ahead of other competitors [1]. A data lake is essentially a accessible, adaptable storage running that stores kind of raw data in its native format after being ingested form heterogeneous sources. To begin with, the examination of the pre-requisite conditions for achieving data high availability is typically absent from standard data availability computing models, which only address cloud storage systems from the standpoint of series or parallel connection design. Furthermore, industry's conventional approaches to enhancing data availability involve designing more intricate organizational frameworks or utilizing physical infrastructure with higher performance. [5][11]. The cloud storage system as a whole is not highly available, through regardless of how the service side is built, if the client side cannot move service nodes.[6][13]

This paper helps us to answer questions like is storing data a costly and complex thing? How can one establish a secure data lake environment ready for business use?[14] How can one integrate existing data warehouse with upcoming data lake technologies.

[1] Research Scholar, Dr. D. Y. Patil Institute of Technology, Savitribai Phule Pune University, Pune, India

wasudev.rahane@gmail.com

[2]Professor, Dr. D. Y. Patil Institute of Technology, Savitribai Phule Pune University, Pune, India

pdpatiljune@gmail.com

[3]Professor, Dr. D. Y. Patil Institute of Technology, Savitribai Phule Pune University, Pune, India

rdbharti@gmail.com

II. LITERATURE REVIEW

The authors [1] give us a descriptive idea about what is a data lake, its major characteristics, with some information of existing data storage systems like Hadoop, AWS, and Azure etc [1][7]. Hadoop is one of the traditional data storing methods uses Map Reduce for data analysis[8]. On the other hand AWS provides exciting features like enhanced data privacy, smooth data integration and user friendliness. The authors [12] compares the functionalities of data lake and data warehouses. We get a brief idea about the existing data storing methods with all the positives and negatives. [13] The research problems of data lake have been addressed by a number of systems and solutions suggested in the last decade. Even while Data Lake is a popular term right now and has a lot of excitement surrounding it, its precise definition and uses are still somewhat unclear.

Previous efforts to organize, the data lake only offer a slight perspective on a portion of the data lake research questions.[14] Moreover, none of these efforts discourse the essentials of upcoming problems with data lakes, like how to support data lakes.[16]

The quintessence is to ascertain an extra cost effective storage alternative. Data that can originate in factual time, scale data of several extent although saving period in describing data schema, transformations, structures, effective and affordable approach to store, manage and analyze data to improve performance of the applications and meet all possible requirements of the user. When the amount of data generated was low, traditionally data was stored on physical drives and magnetic disks.[17] Users interacted with the data storage mechanics physically when they require any data in real time. Users need to predefine some data structures for storing the data, think about efficient memory management techniques, efficient algorithms, and process the data before entering the storage [5]. Later querying the data is a slow process with increasing number of data the time required for fetching the data from the storage increases. Some of the major drawbacks of traditional system are:

- Physical devices are required for storage.
- User must run maintenance tools manually.
- High initial investment and effort.
- More prone to cyber threats like virus attacks.

III. DATA WAREHOUSE V/S DATA LAKES

Data warehouses are traditional data storage methods whereas data lakes are a modern updated version of existing data warehouses[15].

Table 1.0, below depicts the key alterations among data warehouses and Data Lake.

Table 1 Data lake and Data ware house dimensions

Sr. No	**Dimensions**	**Data Warehouses**	**Data Lakes**
1]	Structured Format	✓	✗
	Unstructured format	✗	✓
	raw format	✗	✓
	processed format	✓	✗
2]	Schema	✓	✗
	on-write	✓	✗
	on-read	✗	✓
3]	Scalability	✓	✗
	Volume	Large	Extremely large
	Cost	Moderate	Low
4]	Architectural Design	✓	✗
	Hierarchical	✓	✗

	Flat	✗	✓
5]	Design Complexity	✓	✗
	Joins	✓	✗
	Processing	✗	✓
6]	Efficiency	✓	✓
7]	CPU/IO	User efficient	Moderately user efficient

Major benefits of data-lake over data-warehouses are:

- Democratize data
- Improve data quality
- Higher Scalability and Versatility
- Schema Flexibility
- Advanced Analytics
- Data storage in native format

IV. SYSTEM ARCHITECTURE

The AWS cloud consists of core AWS services, which include AWS Lambda (function) micro-services, OpenSearch service Amazon (inheritor to amazon elastic search) intended for providing reliable exploration proficiencies, amazon's Cognito to authenticate user, for data transformation AWS glue data transformation and for data analysis Athena [7]. The elucidation controls the scalability, safety and stability of amazon's S3 to accomplish a prescient dataset of organizational catalog. Amazon dynamo DB that deals with the corresponding metadata. When a data set is categorized, the descriptive and attributes tags are obtainable for exploration [18-20]. A user be able to search and surf obtainable datasets trendy the created data lake console and build the data list required by the users. The system retains track of data set, a user chooses and creates a manifest file with protected access links to the results when the user checks out [21-23].

Proposed Methodology

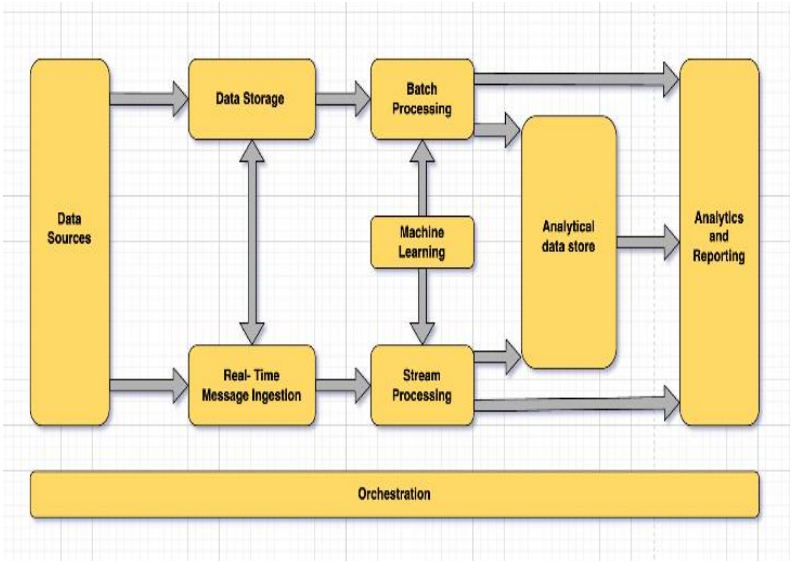

Figure. 1: Data Lake process

Proposed System Architecture for data lake

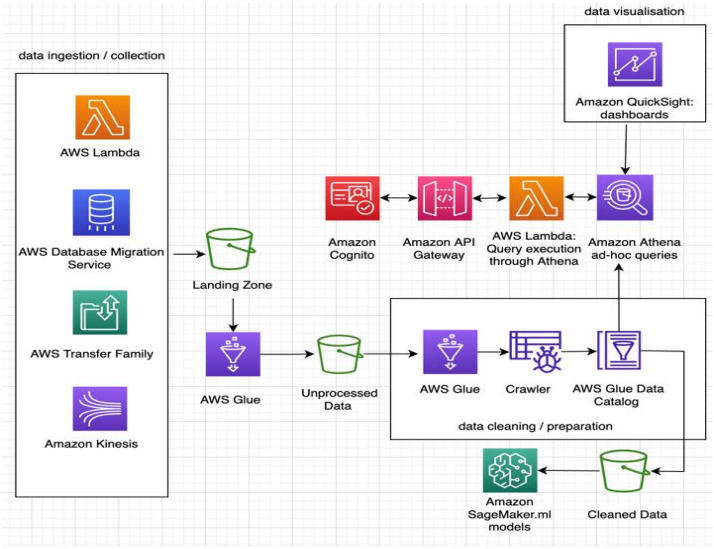

Figure 2: System Architecture

Developing a data lake by selecting an appropriate application for building a data lake, we need to design the data lake, which will follow the given steps:

CEDLM: Data Lake Algorithm:

Step_1: map present data with incoming data and present user with incoming user

Step_2: If user is not present in existing user list - create new user

- Check the required role is present to attach to user or not
- If IAM role is present

i. Check existing IAM role and policies

ii. If condition satisfied attach present policy to the Role

iii. Attach Role to the newly user

- If IAM role is not present:

i. Create security policy

ii. Create new IAM Role

iii. Attach the required permissions and role policy to the Role

iv. Attach Role to the newly created user

Step_3: Create S3 bucket

i. Create security policy for S3 bucket

ii. Create Role

iii. Assign Role to the respective User as per requirement.

Step_4: Data Lake Configuration

i. Airflow to programmatically author, schedule and monitor workflow.

ii. Create Database using RDS (Relational database service)

iii. Administration and Development of Database.

iv. Register S3 bucket as Data Storage.

Step_5: Glue and Crawler Configuration

i. Create AWS Glue catalog

ii. Classify data

iii. Determine the format, schema, and associated properties data.

iv. Group data in tables

v. Write metadata for GLUE, ATHENA to vision the S3 facts as a database schemas using Crawler.

vi. Assign permissions to the crawler, Athena, and Glue

vii. Assign role the respective services for proper communication with each other.

Step_6: Query Data

Run queries using Athena

Mathematical modeling:

C-Total Storage Capacity of the data lake (in bytes)

P-Probability of data longevity (as a decimal)

I-Data loss rate (as a percentage)

Tr- Average data retrieval time (in seconds).

Cost- Total cost of data storage.

Now, let's represent some relationships:

1. Data Longevity(D):
   $D = P * 100\%$
2. Data Loss Rate (L):
   $L \leq 100\ \%$
3. Data Retrieval Efficiency (E):
   $E = \frac{1}{T1}$
4. Cost of the Data Storage:
   Cost = Infrastructure Cost + Maintenance Expenses + Data Retrieval Costs

Result Analysis

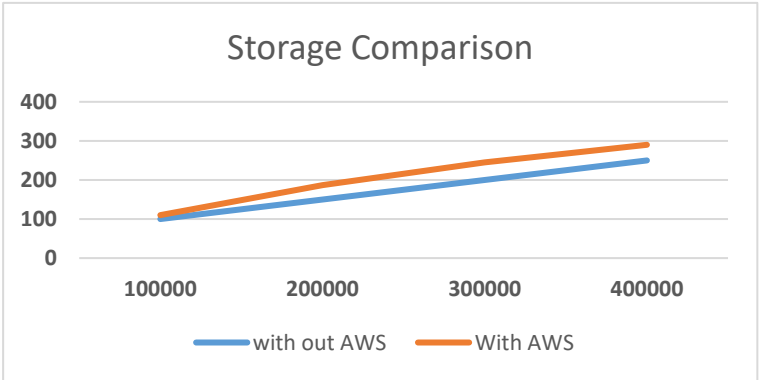

Figure 3: Data Lake storage

V. CONCLUSION

With increasing number of data from numerous sources, developing a Data Lake is a need, to store large amount of data efficiently that comes in real life. Security and user accessibility plays a major role nowadays. AWS offers excellent and efficient end-to-end framework with solutions like security, managing, monitoring data at low cost effectively. Apart from being a widely use technology, AWS also has various integrity features to upscale existing data warehouses into data lakes at low cost without any major complexities.

REFERENCES

[1] R. Hai, C. Koutras, C. Quix and M. Jarke, "Data Lakes: A Survey of Functions and Systems", in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 12, pp. 12571-12590, 1 Dec. 2023, doi: 10.1109/TKDE.2023.3270101

[2] D. Oreščanin, T. Hlupić and B. Vrdoljak, "Managing Personal Identifiable Information in Data Lakes", in IEEE Access, vol. 12, pp. 32164-32180, 2024, doi: 10.1109/ACCESS.2024.3365042.

[3] Xu, J. "An accurate management method of public services based on big data and cloud computing", J Cloud Comp 12, 80 (2023). https://doi.org/10.1186/s13677-023-00456-0

[4] Aakash Aundhkar, Shweta Guja, "A review on Enterprise Data Lake Solutions", Journal of Science and Technology, Volume 06, Issue :01|August 2021

[5] E. Zagan and M. Danubianu, "Data Lake Architecture for Storing and Transforming Web Server Access Log Files", in IEEE Access, vol. 11, pp. 40916-40929, 2023, doi: 10.1109/ACCESS.2023.3270368.

[6] F. Nargesian, K. Pu, B. Ghadiri-Bashardoost, E. Zhu and R. J. Miller, "Data Lake Organization", in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 237-250, 1 Jan. 2023, doi: 10.1109/TKDE.2021.3091101.

[7] R. S, A. S. Karthik, M. H. S. M. K. Karthik, M. Jayasurya and S. Yashwanth, "Examining Amazon Customer Reviews using PySpark and AWS: A Data Lake Approach", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10307845.

[8] Z. Dong, "Research of Big Data Information Mining and Analysis : Technology Based on Hadoop Technology", 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, 2022, pp. 173-176, doi: 10.1109/BDICN55575.2022.00041.

[9] Tanmay Sanjay Hukkeri, Vanshika Kanoria, Jyoti Shetty, "A study of Enterprise Data Lake Solutions", International Research Journal of Engineering and Technology (IRJET) Volume : 07 Issue : 05|May 2020

[10] Amra Munshi, Yasser Abdel-Rady I Mohamed, "Data Lake Lambda Architecture for Smart grids big data analytics", IEEE Issue: 23 July

[11] Bozena M-M，Marek S，Dariusz M. "Soft and decarative fishing of information in Big Data Lake", IEEE Transactions on Fuzzy Systems，2018，1(99):1-6.

[12] Cravero, O. Saldana, R. Espinosa, and C. Antileo, "Big data architecture for water resources management: A systematic mapping study," IEEE Lat. Am. Trans., vol. 16, no. 3, pp. 902-- 908, 2018.

[13] Sophia Boing Righetto, Eduardo Luiz Martins, Andre Luiz Pereria, "Data Lake Architecture for Distribution System Operator", 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT) | 978-1-7281-8897-3/21/$31.00©2021 IEEE DOI: 10.1109/ISGT49243.2021.9372181

[14] ByungRai Cha, Jong won Kim, Design and Implementation of connected data lake for a reliable data transmission.

[15] Tanmay Sanjay Hukkeri, Vanshika Kanoria, Jyoti Shetty, "A study of Enterprise Data Lake Solutions", International Research Journal of Engineering and Technology (IRJET) Volume : 07 Issue : 05|May 2020

[16] Yi-Hua Chen, Hsin-Hsin Chen, and Po-Chun Huang, "Enhancing the Data Privacy for Public Data Lakes", Proceedings of IEEE International Conference on Applied System Innovation 2018

[17] J. Sawadogo, Pegdwende and Darmont, "On data lake architectures and metadata management," J. Intell. Inf. Syst. Springer, pp. 1--24, 2020.

[18] Mukund Rajeshwar,Rajesh Bharati, "Function as a Service in Cloud Computing: A survey", International Journal of Future Generation Communication and NetworkingVol. 13, No. 3, (2020), pp. 3291–3297

[19] Filiana, A. G. Prabawati, M. N. A. Rini, G. Virginia, and B. Susanto, "Perancangan Data Warehouse Perguruan Tinggi untuk Kinerja Penelitian dan Pengabdian kepada Masyarakat," J. Tek. Inform. dan Sist. Inf., vol. 6, no. 2, pp. 174–183, 2020, doi: 10.28932/jutisi.v6i2.2557

[20] G. W. Darma, K. S. Utami, and N. W. S. Aryani, "Data Warehouse Analysis to Support UMKM Decisions using the Nine-step Kimball Method", Int. J. Eng. Emerg. Technol., vol. 1, no. 1, pp. 65–68, 2019.

[21] Shashikant Athawale, Virat Giri, S.L. Bangare, "Collateral extension in provocation of security in IoT", Int. J. Future Gener. Commun. Netw. (Web of Science), 2233-7857, 14 (1) (2021), pp. 3703-3716.

[22] S.L. Bangare, P.S. Bangare, K.P. Patil, "Internet of Things with green computing", Turkish J. Physiother. Rehabil., 2651-4451, 32 (3) (2021), pp. 12494-12497

[23] S.L. Bangare, S. Gupta, M. Dalal, A. Inamdar, "Using node.Js to build high speed and scalable backend database server", Proc. NCPCI. Conf. International Journal of Research in Advent Technology, 4 (2016): 19.