Architecting an Enterprise Data Lake, A Covid19 Case Study

Bushra*[1], Mohsin Ali Memon[2], Salahuddin Saddar[3]

[1] P.A.Q Construction Company, Hyderabad, Sindh, Pakistan.
[2, 3] Mehran University of Engineering and Technology, Jamshoro, Sindh.

* Corresponding author. Tel.: +92 3359200889; email: engnrrbushraqureshi@gmail.com
Manuscript submitted August 6, 2020; accepted December 28, 2020.
doi: 10.17706/jsw.16.4.174-181

Abstract: Data is increasing at an enormous rate every day. Traditionally data has resided in silos across any organization, so it's difficult to have a complete picture for data driven business decision making. Data lake addresses the problem of rate of increase of data by providing "schema on read", better integration and cheaper storage. It also solves the data silos problem by providing a central platform for a variety of data housing needs. However, implementing a data lake becomes challenging as the implementation needs to address the additional needs like metadata management, data discovery, data governance, data lifecycle management, security and centralized access controls mechanisms. This paper intends to provide a comprehensive architecture of data lake to address these challenges. We have also conducted and documented our experiments with publicly available datasets about COVID19 to validate the design and applicability of the proposed architecture for business analytics purposes.

Key words: Big data, data lake, data governance, data lake management, serverless architecture.

1. Introduction

Data is increasing at an overwhelming rate and it will unlock unique user experiences and a new world of business opportunities. [1] Data lake is a centralized place to store all your data (structured, semi-structured, unstructured) in its pristine form at any scale in a cost effective way. Various types of analytics can then be applied to gain insights and make data driven decisions. According to a survey in February 2020, where 47% of the respondents worldwide confirmed the benefits of data lake. [2] An Aberdeen survey found that organizations with data lake implementation outperformed similar companies by 9% in organic revenue growth. [3] However due to the sheer size of data in the data lakes and the absence or incompleteness of a comprehensive schema or data catalogue, data discovery has become an important problem in data lakes. [4] Additionally Gartner stated that "Through 2022, over 80% of Data Lake projects will fail to deliver value as finding, inventorying and curating data will prove to be the biggest inhibitor to analytics and data science success". [5] Data cataloguing mechanism incorporated in data lake solution will not only help in data discovery but it can also serve in breaking the barriers to the adoption of Data Lake.

Implementing data management and security is considered difficult due to open/flexible nature of data lake architecture. [6]-[8] Effective data governance and management strategy help in implementing necessary user access controls. Additionally in the presence of data privacy and security compliance such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA),

organizations are required to document the complete life cycle of data asset. [4] It has also been found that data lake lacks the ability to determine data quality or the lineage of findings. [5] Therefore necessary data lifecycle management is needed. Data lifecycle or lineage provides information about the complete journey of data like its sources, dependencies, transformations and usage enabling users to make decisions more effectively.

In this paper we discuss a comprehensive data lake solution in a serverless fashion with cataloguing mechanism, lineage information, and security and access control methods. The other side of the fact is that data governance is also important to prevent data lake from turning into data swamp so it has been incorporated in the solution. This solution has been implemented and experimented which have been discussed to assess the usability of this solution.

2. State-of-the-Art

Data lake due to its flexible architecture is a popular option among data driven organizations for big data storage. Basic concepts and architecture of data lake have been explored and explained widely in research studies. [6], [9]-[11] It is often compared with traditional concepts like data warehouse as both concepts share many similarities [3], [9]. Recently, architectural designs [10] of a data lake solution and different deployment models [9] are a focus of research.

Data lake solutions have been used in several industries. In commercial banking, due to regulatory requirements data governance is often an important topic and involves creating product roadmap for introducing data governance in data lake solutions. [12] Mehmood *et al.* [13] presents the first implemented solution for cutler project based on data lake architecture which allow data collection, storage, processing of diverse data on which analysis is done. A smart grid big data ecosystem based on lambda architecture is presented to address these challenge in traditional grid. [14] In Kondylakis *et al*. [15] data management infrastructure has been implemented for the iManageCancer EU project.

Challenges related to data ingestion, data extraction, data cleaning, dataset discovery, metadata management, data integration, and data versioning have been discussed in [4]. Some other challenges and concerns that have been focused includes difficulty in determining data quality or the lineage of findings, lack of governance and mechanism to maintain metadata, absence of security and access controls have also been discussed in literature. [6] SWOT analysis in [7] revealed that strength, weakness and opportunities of data lake, which are important to know, includes cost effectiveness, data management and security, and data discovery and exploration respectively. Data swamp has been considered as the biggest pitfall of data lake. [16] Suriarachchi et al [8] identifies the data management and traceability problems in data lake. Challenges identified in Balachandran *et al*. [11] includes: data governance, metadata management and enhancement. Shepherd et al [17] discusses challenges, opportunities and proposed a theoretical framework for data lake adoption. In Giebler *et al.* [18] research gaps and challenges concerning data lake architecture, data governance and comprehensive strategy to realize data lakes have been identified. These challenges have also been addressed.

As metadata on the origin of data is just as important as other components [17], [19]. However it is insufficiently considered in all investigated metadata models. [18] In Suriarachchi et al [8] a reference architecture to overcome provenance challenges, implementation and evaluation of proposed solution has been provided using Hadoop technologies.

Accessibility is another important factor when building solutions. It has been assumed that by 2021, organizations that offer a curated catalogue of data to users will realize twice the business value from their data and analytics investments than those that do not. [5] In Brackenbury *et al.* [19] a methodology is proposed for discovery and management that collects user feedback along the dimensions of data, its origin

and the characteristics in order to identify data that could be integrated or managed similarly.

As Data Lakes do not have predefined schema metadata helps in understanding the data. In Sawadogo *et al*. [20] metadata management approach has been discussed for textual documents. In Nogueira et al [21] a metadata model more precisely a data vault has been presented which allows easy schema evolution. In Yebenes *et al*. [22] a data governance framework for Third Generation Platforms has been presented. In this architecture the central element is data life cycle management which is supported by metadata management, data quality and risk management components.

As data lake stores huge amount of data which often needs to be shared within organization or to the public so security and access control is necessary. In Chen *et al.* [23] a framework for preserving data privacy has been presented which discusses the data sharing protocol along with the pay per use billing policy. This paper proposes a serverless robust data lake architecture addressing the challenges discussed so far.

3. Data Lake Architecture

In this section proposed architecture is discussed using various components. The focus will be on the components that needs to be there regardless of where or which tool and technology is used for the implementation of the concept. Various components of this architecture are:

3.1. Data Ingestion

In this section proposed architecture is discussed using the data ingestion layer is the backbone of any analytics architecture. Analytics system depends on consistent and accessible data. Data ingestion is the process of collecting raw data from various siloed databases or files and integrating it in one place which serves as a single source of truth. In this architecture REpresentational State Transfer (REST) ingestion Application Programing Interface (API) has been created to facilitate users in ingesting data to the data lake. From a security point of view this API is only accessible to the authenticated users. Lineage information is also stored whenever the user requests the API to store the data.

3.2. Data Storage

The data from external sources are stored in the storage i.e. data lake in its raw format. Storage is flexible enough to support various data formats like CSV, JSON, Apache Parquet, ORC, AVRO, XML and others. It is designed for high data durability, scalability on demand, security, performance, and offers cost effective storage.

3.3. Data Processing

Data originating from different sources in heterogeneous forms makes data curation necessary to standardize this data. Extract Transform and Load (ETL) operations are performed on the data, if needed. This curated data is then stored into the data lake. During data processing phase that can include: combining, filtering and other operations, lineage information is stored to keep track of how the data is transforming from time to time.

3.4. Data Discovery

Data discovery requires extraction of metadata attributes and inferred schema. This extracted information is then used to populate the data catalogue. The data catalogue is used for data discovery purposes and further understanding of data in the consumption phase. This also requires an update mechanism to keep the data catalogue synchronized with changes happening to data.

3.5. Data Analytics

When metadata is available in the catalogue, business users can utilize this data by querying the available datasets. We can analyze the available data to find interesting patterns and various forms of data visualization will assist in representing the data by making explicit the trends and patterns in data.

3.6. Data Governance

Data governance provides structure and management to the data and makes it more accessible and meaningful. Various governance frameworks have been suggested in the literature. From the literature available and discussed in this paper it is known that data governance mainly focuses on the following three areas: 1) Data Lifecycle Management 2) Data Security 3) Metadata Management.

Data lifecycle management is essential to identify the data and trace its source, transformation it suffers, its location, dependencies and joins with the other data. The lifecycle data is stored when the data is ingested and each time it undergoes some transformation to keep traces of the data. Visualization over lineage information will aid in understanding the complete journey of the data from its origin to its usage.

Data security is needed at every layer. Data is secured by encryption in both at rest and in transit. Access controls take place right from data ingestion to consumption. Ingestion API is only accessible to the authenticated and authorized users using specific application maintained roles. The proposed architecture also incorporates audit controls and all user actions are logged in an audit trail for regulatory and compliance purposes. The proposed architecture provides a holistic approach toward data lake security by providing authentication, authorization, protection, encryption and audit in a single platform.

4. Architectural Implementation

This section discusses implementation of the architecture on cloud computing platform. For the implementation of proposed architecture Amazon Web Services (AWS) serverless and managed services are used as shown in Fig. 1.

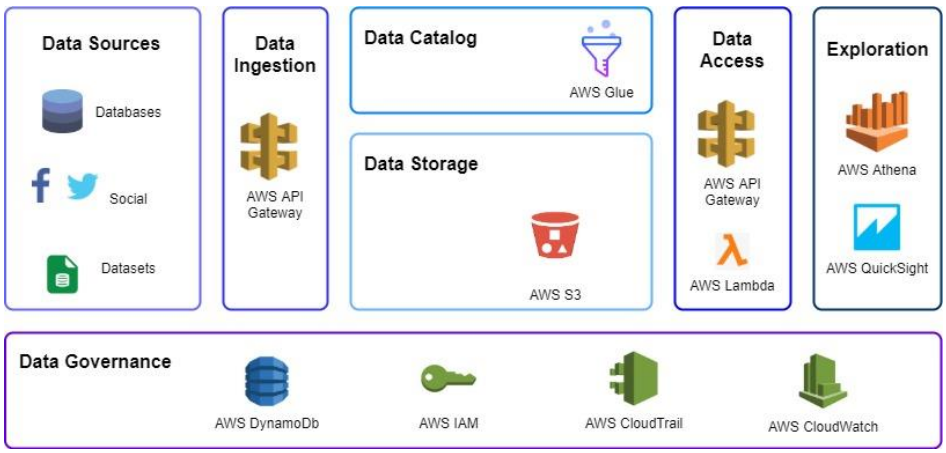

Fig. 1. Physical model.

Data Ingestion REST API is created using Amazon API Gateway. AWS IAM authorization is used with the ingestion API, so that it is accessible to users having specific permissions or role. This API is then integrated with the lambda function, which imports the data to the storage i.e. AWS S3. This data in its raw format is standardized by executing spark scripts using AWS Glue job. Once ETL is performed, AWS Glue crawlers will help in building data catalogue. It will crawl the data sources, extract their metadata and discover schemas then populate catalogue with new and modified table and partition definitions, and maintain schema versioning. This catalogue serves as a central metadata repository that will help users in data discovery. When metadata is available in the catalogue, business users can utilize this data by querying the available

datasets using AWS Athena. Athena can be used for exploratory analytics and ad hoc queries. Athena along with QuickSight helps in analytics and visualization to identify interesting patterns.

Data Governance activities are carried out throughout the architecture. Lineage information is gathered whenever a data source is added or data is transformed and it is stored in AWS Dynamodb in json format. This information can now be visualized using any json visualization tool. Data auditing information is also stored in AWS DynamoDB to track changes performed in the database and by whom, when and how these were performed. AWS CloudWatch and AWS CloudTrail are used to assist in monitoring resources, and keep a log of all actions that have taken place inside the AWS environment. The access to the data lake is controlled using IAM roles.

5. Practical Application of the Ecosystem

The main objective of developing comprehensive data lake solution is to evaluate the usability of the proposed architecture. In our experiments, we used publicly accessible datasets provided by Kaggle. Dataset, named as COVID-19 dataset, contains data from various sources like WHO, worldometer official site and others. Lineage information or data life cycle management of this practical implementation is shown in Fig. 2 Dataset1 contains information regarding measures taken for COVID19 which was used for analytics to generate ReportA. Dataset2 contains country-wise data containing information regarding the corona cases situation worldwide, and worldometer data containing information regarding the current(2020-07-02) situation of countries, which was used upon which analytics was applied to obtain ReportB and ReportC as shown in Fig. 2. To compare population with the number of confirmed, recovered, total tests, and deaths on 2020-07-02 we combined the information of population from worldometer with the country wise information. This analysis report is shown by Report D in Fig. 2 is an example of filtering and combining two datasets. The results of analysis is shown is Fig. 3.

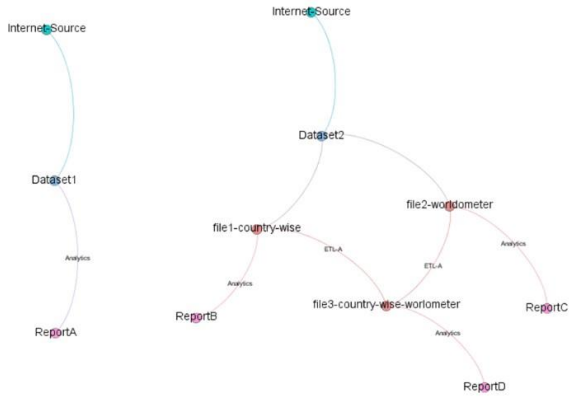

Fig. 2. Lineage information.

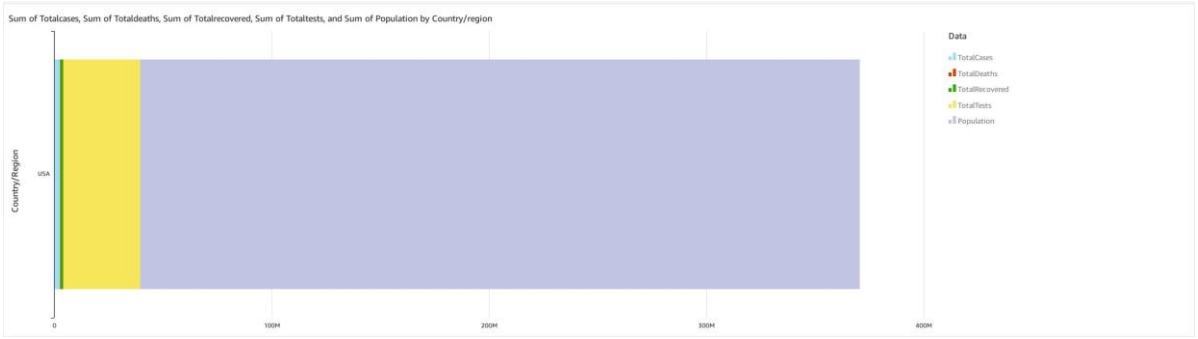

Fig. 3. Comparison of total cases, deaths, recovered, tests with the population of USA.

6. Conclusion and Future Work

Data lake is a concept used for storing big data and later using it for analytics, visualization or making data driven decisions. This paper presents a comprehensive data lake solution addressing some of the main challenges like data discovery, data governance, data lifecycle management, and security. This ecosystem is designed to handle enormous amount of data due to its flexible, scalable on demand, secure and durable architecture. In the proposed solution data is stored in the data lake using ingestion API. This API is only accessible to users who have specific role and permissions. Metadata is extracted from the curated data and managed using data catalogue. Data catalogue assists in discovering, understanding, contributing and using the data. Analytics and visualization can be done on these available datasets. Security and access to the data source is controlled using user roles. Data lifecycle management, security and metadata management together helps in building data governance to the solution. The presented eco-system was implemented and setup on a cloud computing platform i.e. AWS. Experiments have been performed to assess the architecture and its business analytics capabilities. However the solution components can also be used on different cloud computing platforms or infrastructures. As an avenue for future work, we plan to integrate ML based real time lineage information extraction to this solution as lineage information is necessary to get a holistic view of the data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Bushra conducted the research; Mohsin Ali Memon and Salahuddin Saddar supervised the research work and provided active feedback throughout the work; all authors had approved the final version.

References

[1] Reinsel, D. (2017). Data age 2025: The evolution of data to life-critical. Retrieved July 26, 2020, from: https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf

[2] Neeb, A. Hadoop and data lakes: Relevance. Retrieved July 26, 2020, from: https://bi-survey.com/data-lakes-relevance

[3] Michael, L. (2015). Angling for insight in today's data lake. Retrieved July 26, 2020, from: https://s3-ap-southeast-1.amazonaws.com/mktgapac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today's+Data+Lake.pdf

[4] Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: challenges and opportunities. *VLDB Endowment*, 12(12). Retrieved July 26, 2020, from: https://dl.acm.org/doi/10.14778/3352063.3352116.

[5] Simoni, G. D., & Zaidi, E. (2019). Augmented data catalogs: Now an enterprise must-have for data and analytics leaders. Retrieved July 26, 2020, from: https://www.gartner.com/en/documents/3957301/augmented-datacatalogs-now-an-enterprise-must-have-for-

[6] Khine, P. P., & Wang, Z. S. (2018). Data lake: A new ideology in big data era. *Procedings of the 4th Annual ITM Web Conf. on Wireless Communication and Sensor Network*.

[7] Zicari, R. (2015). The data lake: A brief SWOT analysis. Retrieved July 26, 2020, from: http://www.odbms.org/2015/05/the-data-lake-a-brief-swot-analysis/

[8] Suriarachchi, I., & Plale, B. (2016). Crossing analytics systems: A case for integrated provenance in data lakes. *Proceedings of the 12th International Conference on e-Science (e-Science),* Baltimore, MD, USA.

[9] Fang, H. (2015). Managing data lakes in big data era. *Proceeding of the 5th Annual International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, Shenyang, China.

[10] Zagan, E., & Danubianu, M. (2020). Data lake approaches: A survey. *Proceedings of* the *15th International Conference on Development and Application Systems*, Suceava, Romania.

[11] Kachaoui, J., & Belangour, A. (2019). Challenges and benefits of deploying big data storage solution. *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*, Kenitra Morocco.

[12] Paschalidi, C. (2015). Data governance: A conceptual framework in order to prevent your data lake from becoming a data swamp. M.S. thesis, Luleå University, Sweden.

[13] Mehmood, H., Cortes, E. G. M., Kostakos, P., Byrne, A., Valta, K., & Tekes, S. *et al*. (2019). Implementing big data lake for heterogeneous data sources. *Proceeding of the 35th International Conference on Data Engineering Workshops*, Macao.

[14] Munshi, A. A., & Mohamed, Y. A. I. (2018). Data lake lambda architecture for smart grids big data analytics. Retrieved July 26, 2020, from: https://ieeexplore.ieee.org/document/8417407

[15] Kondylakis, H., Koumakis, L., Tsiknakis, M., & Marias, K. (2018). Implementing a data management infrastructure for big healthcare data. *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Las Vegas, NV.

[16] Walker, C., & Alrehamy, H. (2015). Personal data lake with data gravity pull. *Proceeding 5th International Conference on Big Data and Cloud Computing*, Dalian, China.

[17] Shepherd, A., Kesa, C., Cooper, J., Onema, J., & Kovacs, P. (2018). Opportunities and challenges associated with implementing data lakes for enterprise decision-making. *Issues in Information Systems*. Retrieved July 26, 2020, from: http://www.halcyon.com/pub/journals/21ps03-vidmar

[18] Giebler, C., & Gröger, C. (2019). Leveraging the data lake: Current state and challenges. *Proceeding International Conference on Big Data Analytics and Knowledge Discovery*.

[19] Madsen, M. (2018). How to build an enterprise data lake: Important considerations before jumping in third nature inc. snapLogic. Retrieved July 26, 2020, from: https://www.snaplogic.com/resources/white-papers/build-enterprise-data-lake

[20] Sawadogo, P., Kibata, T., & Darmont, J. (2019). Metadata management for textual documents in data lakes. *Proceeding of the 21st International Conference on Enterprise Information Systems*, Heraklion, Greece.

[21] Nogueira, I. D., Romdhane, M., & Darmont, J. (2018). Modeling data lake metadata with a data vault. *Proceeding of the 22nd International Database Engineering & Applications Symposium*, New York.

[22] Yebenes, J., & Zorrilla, M. (2019). Towards a data governance framework for third Generation platforms. *Proceeding of the 2nd International Conference on Emerging Data and Industry 4.0*, Leuven, Belgium.

[23] Chen, Y., Chen, H., & Huang, P. (2018). Enhancing the data privacy for public data lakes. *Proceeding of the IEEE International Conference on Applied System Invention*, Chiba, Japan.

Bushra has completed her bachelor's degree in software engineering from MUET, Pakistan in October 2018. At present she is pursuing her master's degree in software engineering. She is enjoys working on big data, applications of cloud computing and data mining.

Mohsin Ali Memon was awarded MEXT Japanese Cultural scholarship to pursue the PhD in Japan in 2010. Currently he is working as associate professor and is involved in various research projects with masters and PhD students in Department of Software Engineering, MUET.

Salahuddin Saddar has done master's degree in software engineering from MUET, Pakistan. He is currently working as assistant professor at Department of Software Engineering, Mehran UET Pakistan.