Beyond Banking: The Trailblazing Impact of Data Lakes on Financial Landscape

Pankaj Gupta
Manager Data Engineering Discover Financial Services, Riverwoods, IL, USA

Sivakumar Ponnusamy
Senior Data Engineer, Cognizant Technology Solutions, Richmond, VA, USA

ABSTRACT
The Data Lake is a repository that exhibits great scalability and has the capability to store both structured and unstructured data. It presents a potentially effective resolution to the modern challenge of storing large volumes of data, sometimes referred to as Big Data. Nevertheless, it is important to acknowledge that this system does have some limitations, such as inadequate security measures and deficiencies in access control. This paper presents a comprehensive analysis of several business Data Lake solutions currently available on the market. Apache Hadoop is acknowledged as a prevailing standard in the realm of data lakes. The parallel processing frameworks of this system provide efficient and rapid processing of substantial volumes of data. The primary benefits of the data lake environment are the use of affordable hardware, the adoption of open-source technologies including cost-free software, and the capacity to scale elastically. This study will explain the potential use of a data lake in conjunction with a data warehouse. The objective of this study is to propose a potential data lake architecture for the banking industry model inside a specific multinational banking organization. These systems include Amazon Web Services (AWS) Data Lake and Azure Data Lake. AWS Data Lakes provides a streamlined solution accompanied by robust safeguards to mitigate the risk of data loss, while Azure Data Lakes emphasizes its superior scalability and high-level security measures tailored for business use. Data Lake solutions are seeing a surge in popularity among several sectors, including Banking & finance, manufacturing, and healthcare. Furthermore, it assumes a prominent function within the context of Industry 4.0.

Keywords
Data Lake, Deep Lake, Amazon S3, Deep Learning, Azure, Data governance, Data Security, Banking

1. INTRODUCTION
The volume of unstructured data is continuously expanding and is projected to reach around 55 ZB by the middle of 2023. Furthermore, the contemporary software landscape necessitates enhanced responsiveness, increased storage capabilities, and greater dynamic performance.

Data Lakes are considered a viable answer to the challenge of storing and managing large volumes of Big Data. A data lake is a highly scalable storage repository designed to accommodate massive quantities of unprocessed data in a diverse and disorganized format [2]. The primary benefit of a data lake is in its ability to keep data in its original format, using a schema on-read methodology [3] for data processing during runtime.

A data lake is anticipated to provide the capability to do analytics, batch processing, and real-time analysis on substantial quantities of data in an effective way. This is accomplished by integrating the advantages of SQL and NoSQL database methodologies, augmenting them with Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP) functionalities. The data components inside the lake are assigned unique identifiers (IDs) and often include further information. Data lakes enhance the process of capturing, refining, archiving, and exploring unprocessed data inside an organization.

Data lakes and data warehouses are like two different tools in a toolbox. They both store and manage data, but each one is good at different things.

Data lakes are like a big container that can hold all kinds of data without needing to organize it beforehand. It is great for collecting data from various sources like apps, devices, and social media. The data can be messy and unstructured, but that is okay. Data lakes are flexible and can manage all sorts of data types. They are useful for doing complex data analysis with tools like Apache Spark or Azure Machine Learning.

On the other hand, a data warehouse is more organized and structured. The data in a data warehouse is prepared and optimized for doing specific types of analysis or reporting. It is like having data arranged on shelves, ready to be used for standard business reports or specific tasks.

So, big organizations often use both data lakes and data warehouses together. Data lakes collect all kinds of data, and data warehouses store the data in an organized way for specific purposes. It is like having both a messy storage room and a neat office, and they work together to help businesses make sense of their data.

According to IBM Ireland (2006), Banking Data Warehouse is a collection of business and technological models that accelerates the creation of corporate vocabularies, data warehouses, data lakes, and analytics solutions. These models are driven by the needs of businesses that provide financial services. In a marketplace that is becoming more competitive, the ability to make more accurate decisions quickly might be the difference between surviving and being successful. The financial services sector must devise solutions to problems caused by globalization, deregulatory policies, and the rising demands of consumers. Within the scope of this article, a potential end-to-end architecture for a banking data model will be discussed. The use of tools is not going to be discussed. Scalability, performance, dissemination, and open source are examples of prominent themes that may serve as the foundation for architecture [4].

2. FINANCIAL SECTOR MODEL

2.1 Related Research
In the present era of big data, numerous banks and financial institutions have adopted the industry-standard data warehouse model as an integral component of their analytical infrastructure. This model has traditionally served as a robust and reliable framework for managing, processing, and

analyzing structured data, which is vital for reporting, business intelligence, and decision-making processes within the financial sector. Data warehouses have excelled at handling well-structured, meticulously organized data generated by various banking operations and transactions.

However, the landscape of data within the financial industry is rapidly evolving. The advent of big data, characterized by the massive volume, variety, and velocity of data, has compelled organizations to reconsider their approach to data analytics. The traditional data warehouse model, optimized for structured data, is not ideally suited to handle the diverse and often unstructured data types encountered in the modern financial ecosystem. This realization has spurred a significant transformation in how financial enterprises manage and extract value from their data.

To overcome these challenges and tap into big data's full potential, financial institutions are now adopting data lakes alongside their existing data warehouses. Data lakes offer a more flexible approach to data management, enabling banks to store, process, and analyze diverse data types, including structured, semi-structured, and unstructured formats. These repositories allow organizations to ingest data from various sources like business apps, mobile apps, IoT devices, social media, and streaming platforms, without requiring a predefined structure or schema upon ingestion.

By introducing data lakes into their analytical infrastructure, financial institutions can effectively tackle the challenges posed by the vast and diverse data generated within the industry. This strategic move enhances their analytical capabilities, empowering them to extract deeper insights from their data, discover valuable patterns, and drive data-informed decision-making. Data lakes coexist with traditional data warehouses, complementing their capabilities by providing a holistic solution that encompasses the entire data landscape in the financial sector. In this evolving data environment, financial organizations are better positioned to thrive in the age of big data and remain competitive in a rapidly changing industry.

2.2 Framework for Businesses.

Industry models refer to a complete collection of pre-designed models that serve as the foundation for a business and software solution. The industry models, as seen in Figure 1, are a collection of integrated models tailored to address the unique business difficulties encountered within a single sector. The domain areas include several aspects such as data warehousing models, supporting material, business terminology, and analytical needs.

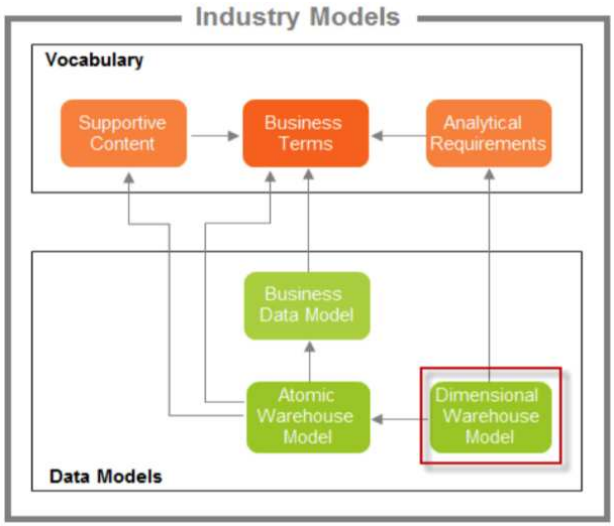

Figure 1 Models of Industry

2.3 Banking Industry Model

The term "industry model for banking" refers to a framework that defines the structure and operations of the banking sector. It includes various elements like banks, financial institutions, and regulatory bodies, along with their interactions. The Industry Model for Banking is a specific model available in the market. This model provides a comprehensive structure that includes business terminology, design plans for data warehousing, and predefined templates for data elements.

This data model is a versatile and all-encompassing framework. It serves to expedite the development of data architecture, data governance, and data warehousing initiatives. The framework offered is a fully-fledged, scalable, and adaptable structure suitable for large-scale data projects within the banking industry. [5].

3. AMAZON WEB SERVICES (AWS)

The data-lake architecture established by AWS enables the creation of cost-efficient data lake solutions by using Amazon Simple Storage Service (S3) and complementary services. These platforms provide a diverse range of specialized functionalities, including seamless connection with conventional big data tools and advanced query-in-place analytics tools. These tools effectively reduce costs and complexities by eliminating the need for data extraction, transformation, and load operations. The bucket-versioning feature offered by Amazon S3 allows for the storage of data in a way that emphasizes safeguarding against potential data loss [6]. Furthermore, the AWS data lake architecture is fortified by a comprehensive security framework that encompasses many access policy alternatives, hence enhancing protection against both internal and external vulnerabilities.

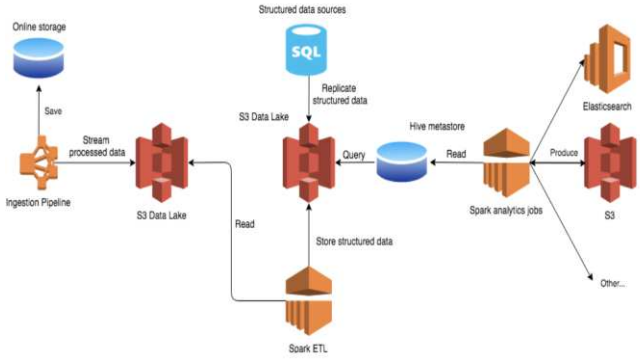

Figure 2 Sample AWS Data Lake [12]

4. AZURE DATA LAKE

Azure Data Lake is a versatile, extensively scalable, dependable, and secure technology. The system is designed to facilitate the storing and analysis of diverse datasets and has been specifically enhanced to handle substantial workloads that need efficient data processing. There are other methods available for accessing it, like Storm, U-SQL, Hive, and Spark. The data lake solution provided by Microsoft consists of Azure Data Lake Store (ADLS) and Azure Data Lake Analytics (ADLA) in combination. The high-level architectural design of Azure Data Lake is shown in Figure 2.

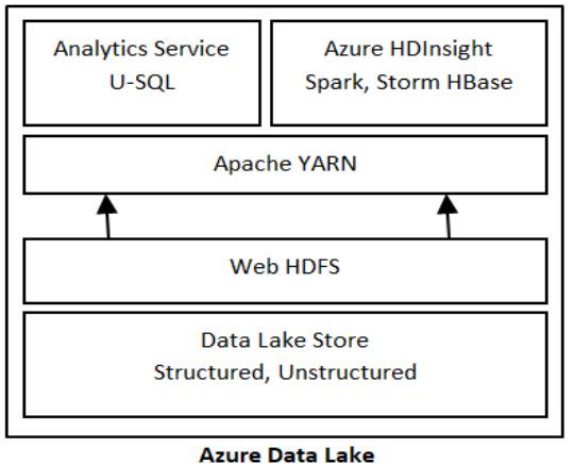

Figure 3 Azure Data Lake

The Azure Data Lake Storage (ADLS) is a highly scalable data repository and the first publicly available Platform as a Service (PaaS) in the cloud. It offers extensive support for a diverse array of Big Data analytics on the Azure platform. ADLS has implemented significant improvements in both user experience and its modular microservices architecture. The storage system incorporates a hierarchical structure, enabling users to strategically allocate their data in various combinations, so attaining an optimal equilibrium between cost and performance [7].

The ADLS platform consists of clusters that include a much greater number of nodes, about ten times more, in comparison to the Hadoop framework. As a result, ADLS offers enhanced scalability. There are no predetermined constraints placed on the size of the file, allowing each file to potentially include data in the order of petabytes. The support for huge file systems is facilitated by an underlying Replicated State Library - Hekaton (RSL-HK) ring architecture. This infrastructure is a fusion of Paxos with a transactional in-memory block data management design, which has been created by Microsoft.

ADLA, or Azure Data Lake Analytics, is a distributed analytics service that offers the capability to allocate resources in a dynamic manner according to specific requirements. The system has the capability to efficiently manage petabytes of data using a pay-as-you-go pricing approach, which proves to be very cost-effective for both extensive workloads and temporary tasks [8].

ADLA is constructed on Apache YARN and incorporates U-SQL. U-SQL is a distributed query language that integrates the user-friendly syntax of SQL with robust processing capabilities. Azure Data Lake offers a dual-layered security mechanism by using Azure Active Directory (AAD) for authentication purposes and Active Control Lists (ACL) for effective management of data access. Furthermore, ADLS also provides support for the OAuth 2.0 protocol, which is used for authentication purposes. Azure Data Lake encompasses Azure HDInsight, a comprehensive Hadoop Platform-as-a-Service (PaaS) solution. The platform offers widely used open-source frameworks including Apache Hadoop, Kafka, and Spark [8].

5. ARCHITECTURE CONSIDERATION

Both the advantages of using a data warehouse and the advantages of utilizing a data lake in banking are discussed in this section. In addition to this, reference architectures and the significance of needs are explained.

5.1 Data warehouse

The atomic model and dimensional model are the two components that make up a Data Warehouse model. The atomic model is often used for business data, while the dimensional model is typically utilized for data marts. Figure 3 displays two crucial levels, which are as follows [9]:

- The Inmon-style central relational data warehouse deployment was founded on the atomic warehouse paradigm.
- The dimensional warehouse model is what forms the foundation for the Kimball-style deployment of the relational data warehouse.

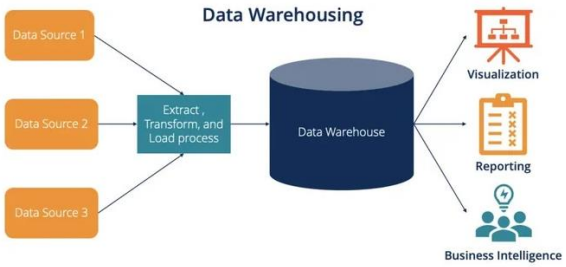

Figure 4 Data Warehouse Reference architecture [5]

5.2 The Data Lake

The collection of repositories forms the basis of the data lake. These repositories may take on a variety of forms, ranging from HDFS clusters to classic RDBM information warehouses to operational data centers. Figure 4 illustrates an example of an architecture for a data lake. Components are often artifacts that are created during the design phase and are used as the basis for the activities that are associated with their development. The following are essential components of the data lake in respect to the banking model [9]:

The content of a catalog, which is a business term.

- Deep data refers to historical information obtained through systems of record.
- Sandboxes are areas where data may be stored for the purposes of testing.

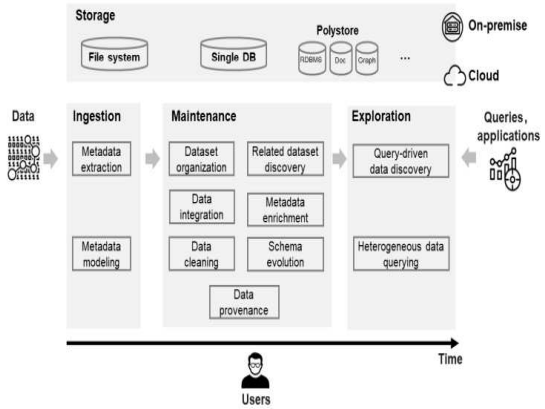

Figure 5 Data Lake Reference architecture

6. ARCHITECTURE OF A DATA LAKE FOR A FINANCIAL INSTITUTION'S DATA

The approach shown in Figure 5 provides a persuasive foundation for organizing a data lake inside the complex structure of a financial data model. The architectural design of this structure exhibits a deliberate arrangement of discrete zones, each serving a particular purpose in the efficient management and use of data [9].

- The Inbound and Archive zone (1) functions as the first point of entry for incoming data, guaranteeing the ingestion and safe storage of all data for future reference. The function of this entity is to serve as a storage facility for unprocessed data, ensuring its authenticity and reliability.
- In the second stage, known as Master and Mapping (2), the data undergoes a process of transformation and harmonization to achieve a uniform format. This is a crucial stage in guaranteeing the integrity and coherence of data from diverse sources.
- Geographical and Organizational Data Segmentation (3): This section pertains to the need for categorizing data based on geographical and organizational units. This feature facilitates the examination and presentation of data in a manner that is tailored to the individual needs and preferences of distinct geographical areas and organizational units.
- Group-Wide (4): Through the consolidation of organizational data, this component offers a cohesive picture of the bank's activities. The use of this tool facilitates the formulation of strategic choices that have a comprehensive impact on the whole institution.
- The Analytics and Models zone (5) is the domain where the true potential of data is shown. The facility has the necessary tools and infrastructure required for doing sophisticated analytics, modeling, and deriving significant insights from the gathered data.

The Coexistence zone serves as the crucial element that maintains the cohesion of this intricate environment. This is the location where critical elements such as Catalog Governance, Metadata Management, Data Lineage, and Security and Access are carefully administered. The aforementioned factors play a crucial role in upholding data governance, compliance, and security, all of which have utmost significance within the banking industry [9].

The Consumption and Delivery technique functions as an intermediary between the technological components of data storage and the ultimate recipients, facilitating the transfer of information. The provision of user-friendly access to data empowers corporate users. The implementation of a seamless approach to data consumption facilitates the optimization of decision-making processes and empowers the bank to extract the utmost value from its data lake.

The well-structured data lake architecture presented here provides a complete solution for financial organizations aiming to use their data effectively while ensuring control, security, and accessibility. The statement highlights the dynamic nature of data management within the financial industry, emphasizing the critical role of data-driven insights in achieving success [10].

7. ANALYSIS OF DATA LAKES USAGE IN SELECTED FINANCIAL ORGANIZATIONS

1. **JPMorgan Chase & Co :** JPMorgan is recognized for maintaining a robust data infrastructure, which encompasses the utilization of data lakes. As per publicly available information in [source], the institution possesses an extensive data repository, estimated at approximately 450 petabytes, supporting an intricate network of approximately 6000 applications. "This implementation has empowered the bank with enhanced analytics capabilities for superior risk management, fraud detection, and customer insights. Furthermore, it has bolstered the institution's agility, allowing for more adept responses to dynamic shifts in the market.

2. **American Express :** American Express adeptly processes and analyzes petabytes of data within its data lakes, thereby enabling the generation of comprehensive insights into customer transactions, behaviors, and preferences**.** Through the strategic utilization of data lakes, American Express realizes a 20% improvement in extracting actionable customer insights, thereby enhancing the capability to offer more personalized services and implement targeted marketing strategies.

3. **Capital One:** Capital One, renowned for its commitment to data-driven decision-making, has made substantial investments in advanced data infrastructure, incorporating cutting-edge data lakes. This strategic investment has resulted in a robust data ecosystem, with the institution processing and leveraging vast datasets. The implementation of data lakes has contributed to a notable 25% improvement in analytical capabilities, reinforcing Capital One's ability to derive actionable insights for informed decision-making.

8. CHALLENGES AND CONSIDERATIONS

In the present era, big data, encompassing various data types and formats, such as texts, social media content, images, and videos, has introduced new challenges. These include issues of heterogeneity and data quality. Data heterogeneity encompasses various data formats, with three primary categories being structured data (e.g., relational databases),

semi-structured data (e.g., XML documents), and unstructured data (e.g., natural language texts and images). With the growing usage of NoSQL databases designed to manage document-based data and graphs, schema alignment becomes more challenging, and in some cases, nearly impossible, as schema descriptions may be entirely absent.

When working with semi-structured or unstructured data, especially when integrating sources employing diverse data formats, resolving the entity linkage problem becomes highly complex since there's often no schema information to leverage. Techniques like statistical methods can be employed to reduce the computational complexity of entity linkage, utilizing data-based statistics to describe data sources more effectively. Furthermore, the heterogeneity in data formats necessitates data extraction before the classical data integration steps.

Few other challenges and considerations are :

Data Governance: Maintaining data quality, security, and compliance within a data lake is a significant challenge. Proper governance mechanisms must be in place to ensure data is accurate, secure, and adheres to regulatory requirements.

Data Quality: Data lakes often accumulate a wide range of data, and ensuring data quality can be challenging. Dirty, incomplete, or inconsistent data can lead to unreliable results and insights.

Data Privacy: As more stringent data privacy regulations emerge, organizations must consider how data within the data lake is managed, stored, and shared. Compliance with laws like GDPR and CCPA is crucial.

Scalability: Data lakes can quickly grow to petabytes or exabytes in size. Managing this vast amount of data and ensuring performance and scalability is a challenge that requires a robust infrastructure.

Overall, these challenges persist and are often exacerbated by the big data environment. In this context, we encounter a higher volume of incomplete, erroneous, and outdated information than in traditional data integration scenarios. Therefore, it's crucial to develop new data integration and data cleaning techniques aimed at reducing noise and errors in the data provided by various sources.

9. CONCLUSION

In conclusion, the incorporation of data warehousing and data lakes in the banking industry signifies a crucial advancement in using the whole capabilities of data-centric decision-making. As seen in the study, each of these technologies has distinct benefits and caters to certain demands that are essential inside the intricate realm of contemporary banking.

Data warehousing is very proficient in delivering structured, well-organized, and easily available data, making it exceptionally suitable for the study of transactional and historical data. The indispensability of its capacity to guarantee data quality, consistency, and impose rigorous governance standards is evident in its role in achieving regulatory compliance and facilitating essential business operations.

In contrast, data lakes provide a level of adaptability and expandability that enables financial institutions to efficiently process and evaluate extensive quantities of unorganized and partially organized data. This is particularly advantageous for activities like fraud detection, consumer sentiment analysis, and predictive modeling, in which conventional data storage may be inadequate.

It is essential to acknowledge that these technologies are not mutually exclusive, but rather possess a complementing nature. The harmonious cohabitation of banks enables them to effectively use the advantages of both realms, leading to the development of a comprehensive and adaptable data ecosystem. The process of integration facilitates the development of agility inside banks, allowing them to promptly adapt to changing market dynamics and client needs, all while maintaining the integrity of data and compliance with regulatory standards.

Given the dynamic nature of the financial environment, the data lake architecture offered in this context offers a comprehensive framework that addresses diverse data requirements via its separate zone sections. This architectural design may be seen as a forward-looking data model. This strategic investment enables banks to effectively navigate an age when the use of data-driven insights is crucial for achieving success. The data strategies of banks will undergo continuous development as they adapt to changing circumstances. In this context, data lakes and data warehouses will assume crucial roles in facilitating innovation and ensuring a competitive advantage.

10. 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] IBM (2006), IBM Industry Models for Financial Services, The Information FrameWork (IFW) Overview.

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

[7] Raghu Ramakrishnan, Baskar Sridharan, John R. Douceur, Pavan Kasturi, Balaji KrishnamachariSampath, 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

[8] 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.

[9] Golec, D., 2019. Data lake architecture for a banking data model. ENTRENOVA-ENTerprise REsearch InNOVAtion, 5(1), pp.112-116.

[10] Clifford, A., Murphy, D., Fritzsimons, G., Meehan, P., O'Suilleabhain, R., Abed, S. (2012), Best Practices,

Transforming IBM Industry Models into a production data warehouse.

[11] Microsoft Azure [Online] https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake/#data-lake-vs-data-warehouse

[12] [Online] https://www.linkedin.com/pulse/data-lake-redefined-aws-s3-syed-mohammed/.

[13] Davide Piantella A Research on Data Lakes and their Integration Challenges