International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Architecting data lake-houses in the cloud: Best practices and future directions

Aravind Nuthalapati *

Microsoft, USA.

International Journal of Science and Research Archive, 2024, 12(02), 1902–1909

Publication history: Received on 30 June 2024; revised on 08 August 2024; accepted on 10 August 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.12.2.1466

Abstract

As the volume of data has grown exponentially, what this means for organisations are significant opportunities and challenges. Three significant challenges face traditional data warehouses when it comes to big-data volume, velocity and variety. Data lakes was a step in the evolution to resolve these challenges, but many times data quality and governance aspects has unmet expectations with traditional Data Lake solutions. Data lakehouses step in as a hybrid approach that combines the freedom of cloud data lakes with rigorously managed warehouses, consolidating operational and analytics workloads on one platform. The study discusses the design and implementation of modern cloud-based data lakehouses, elaborating vital aspects such as storage layer, metadata management system along with its access control policy enforcement. Data lakehouses take advantage of cloud technologies to provide scalable, cost-effective ways to improve the decision-making process based on sound data. We talk about best practices and look to the future, struggling with issues in data governance and integration to optimise organisational data strategies.

Keywords: Data Lakehouse; Cloud Computing; Data Management; Big Data Analytics; Data Governance

1. Introduction

The exponential rise in data, is a unique opportunity but also bigtime hurdles for the organisations of all industries given where we stand today with digital interfaces [1]. As data becomes an increasingly important resource to business and institutions in seeking competitive advantage, the requirement for well-conceived systems that allow effective management of such data has become crucial [2] - [4]. Data warehouses, for instance, have long been the workhorse behind business intelligence and decision-making [5] - [7]. Nevertheless, the ubiquitous three Vs of big data volume, velocity and variety represent an order-of-magnitude increase in scale from what these systems were designed for [8]. To solve these challenges, data lakes were innovated as they provide a flexible and scalable storage solution for massive amounts of raw data in different formats [9] - [11]. However, data lakes have failed to deliver this potential primarily because of the problems and issues with Data Quality, Governance & Security leading many organisations struggling in realising its value [12] & [13]. A complementary idea called data lakehouses has arisen as a potential fix, striving to deliver the unbounded capability of Data Lakes together with structured management components resembling those one would find in Data Warehousing environments [14] - [16]. Data lakehouses combine the best of both warehouse-like structure and lack of proprietary systems, providing a single platform that can handle diverse data types and analytics workloads [17] & [18].

However, with the emergence of cloud computing, a new chapter has unfolded in data management providing scalable and cost-effective infrastructure which complements what is required by snowflake [19]. Cloud-based data lakehouse uses the best of both cloud services to create a single, unified solution for Data management which brings together elements of recording and storage and displaying and optimising at scale warehouse list ready-to-query-optimised-and secured for high-performance [20] - [22]. The journal details architectural principles, best practices and design patterns

* Corresponding author: Aravind Nuthalapati

for building data lakehouses at cloud scale describing how they can revolutionise data-powered decision-making in industry.

Table 1 Lakehouse Architectures model comparison

Author	**Lakehouse Architectures**	**Advantages**	**Disadvantages**	**Use Cases**
Michael Armbrust et al. [23]	Lakehouse: Unifying Data Warehousing and Analytics	Combines data lakes and warehouses, supports both BI and ML	Complexity in unifying storage and processing, metadata management challenges	Business intelligence, advanced analytics
Sergey Melnik et al. [24]	Dremel: Interactive Analysis of Web-scale Datasets	Fast query execution, efficient storage	Limited to structured data, requires specialised skills	Web analytics, large-scale data analysis
Matei Zaharia et al. [25]	Resilient Distributed Datasets on Spark	Fault tolerance, in-memory processing	High memory usage, cost of resources	Large-scale data processing, machine learning
Daniel J. Abadi et al. [26]	HadoopDB: Hybrid of MapReduce and DBMS	Combines flexibility of MapReduce with DBMS features	Complexity in configuration, integration challenges	Analytical workloads, data warehousing
Ronnie Chaiken et al. [27]	SCOPE: Parallel Processing Framework	Scalability, ease of data parallelism	Complexity in task management, resource requirements	Big data processing, batch analytics
Michael Stonebraker et al. [28]	C-Store: Column-oriented DBMS	Improved query performance, efficient storage	Limited to structured data, integration complexity	Data warehousing, analytical processing
B. Saha et al. [29]	Apache Tez: Unifying Data Processing Applications	Efficient data modeling, flexible application building	Integration challenges, learning curve	ETL workflows, stream processing
Reynold S. Xin et al. [30]	GraphX: Distributed Graph System on Spark	High performance, scalable graph processing	Resource-intensive, complex setup	Graph analytics, social network analysis
Fabian Hueske and Stephan Ewen [31]	Apache Flink: Stream and Batch Processing	Real-time data processing, fault tolerance	Complexity in stream management, potential latency issues	Real-time analytics, data streaming
Manos Karpathiotakis et al. [32]	Adaptive Query Processing on Raw Data	Dynamic query optimization, supports unstructured data	Complexity in processing, requires tuning	Ad-hoc queries, exploratory analysis

Lakehouse Architectures model comparison is presented in Table 1. This re-conflation of analytical and operational models heralds a new era in data management, the lakehouse architecture, which can be considered as an evolution post-data lake to bridge storage improved query performance. Michael Armbrust, et al. By unifying these paradigms, the lakehouse architecture is flexible enough to support structured (e.g., SQL) and semi-structured / unsemi-sturctured data as well as workloads such business intelligence and machine learning on a single datalake. By doing so, organisations can take advantage of the scalability and elasticity afforded by data lakes while still reaping the benefits that come from using a relational database like transactional guarantees and management features. Nonetheless, as soon as you work on a system of that scale there are issues around managing metadata and orchestrating storage and processing frameworks which become hard to solve. New platforms optimized for big data storage and processing specifically on Apache-Licensed ecosystem tooling: other examples include Apache Beam (Abadi et al., 2009), and Apache Flink (Hueske and Ewen, 2016).

2. The Evolution of Data Management Architectures

2.1. Traditional Data Warehouses

Data warehouses have been the foundation of enterprise data management for years, serving as a focal point to aggregate structured information from across different operational systems. They are specifically designed for Online Analytical Processing (OLAP), which enables organisations to execute complex queries and make analytical decisions. Data warehouses work with the "schema-on-write" method where data must be shaped and formatted before it is inserted into storage, which ensures that all entries are uniform in quality, consistency and structural integrity.

On the other hand, conventional data warehouses have their own set of limitations in coping up with modern diverse and dynamic data environments. Data warehouses are ideal places for structured data but can often choke on the heads of unstructured or semi-structured events emanating from sources such as social media feeds, IoT devices & web logs.

Finally, data warehouses often hold very strict schema requirements which are troublesome handling changing formats and types or the need for enterprise-grade repositories that can handle such rapid deployments.

2.2. Emergence of Data Lakes

As a response to the inefficiencies of data warehouses, schema-on-read strategy was used as foundation for new kind of storage and process technology called Data Lakes. Data lakes are the raw data in its native format, that is stored and has been extended to include high-volume files without having had time or resource for schema inscription. That flexibility enables organizations to help more analytic use cases at scale such as Advanced Analytics and Machine Learning.

Data lakes offer several advantages, including:

- Scalability: Data lakes are a resilient option for storing an immense amount of data, as they can use distributed storage systems such as the Hadoop Distributed File System (HDFS) to distribute and replicate metadata over multiple servers.
- Flexibility: By storing data in its raw format, data lakes allow organizations to explore and analyze data in different ways, supporting a wide range of analytics applications.
- Cost Efficiency: This drives down the costs of data preparation and management by helping companies avoid having to go through large amounts of transformation processes, that would need: controlling structures in place for a data warehouse.

But data lakes can also suffer from the same governance, quality, and security problems as traditional architectures. When properly managed and governed, data lakes can avoid becoming "data swamps" where relevant data is impossible to find or useless. Data lakes risk undermining their potential to provide constructive and actionable insights when they are left without data quality controls, or if metadata management is subpar.

2.3. Rise of Data Lakehouses

The main reason data lakehouses came into being is the problems that people are experiencing with both Data Warehosues and Data Lakes. Data lakestack is a hybrid architecture that marries the scale and adaptability of data lakes with the discipline to handle structured data in single method, possibly laying rest for arguments about Data Lake vs. As such, this is seen as an end-to-end way to have a common platform that would support both batch and real-time processing in addition to advanced analytics as well machine learning applications.

Key features of data lakehouses include:

- Unified Storage: Data lakehouses leverage the same storage layer that can handle both structured and unstructured data, so companies have a single platform on which they manage all types of their data.
- Open Data Formats: Since data lakehouses are built on open-data formats like Parquet, Avro or ORC they support quick and seamless access to the data that can be easily processed by a range of tools across different platforms.
- Transactional Capabilities: Data lakehouses have ACID (Atomicity, Consistency, Isolation, Durability) transactions for data consistency and reliability of analytical workloads with built-in data lakehouses

- Metadata Management: Robust metadata management systems enable data lakehouses to organise and catalogue data, ensuring data discoverability, governance, and compliance.
- Advanced Analytics Support: Data lakehouses delivering native support for advanced analytics and machine learning capabilities, allowing businesses to unearth more profound insights than ever before.

3. The Role of Cloud Computing in Data Lakehouses

We have also seen that the first-rate fit for link introduces is a combination of hyperscale and value high-quality infrastructure, which cloud computing services offer in modern-day information architectures. Cloud suppliers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) provide huge quantities of offerings had to install and perform records Lakehouses Specifically, with cloud storage solutions like AWS S3 or Azure Data Lake Storage; data processing frameworks such as Apache Spark and Google BigQuery, machine learning platforms what provide a framework to develop an AML- platform for the power sector.

These mechanisms are then leveraged by cloud-based data lakehouses to construct a single and integrated solution, which provides the following benefits.

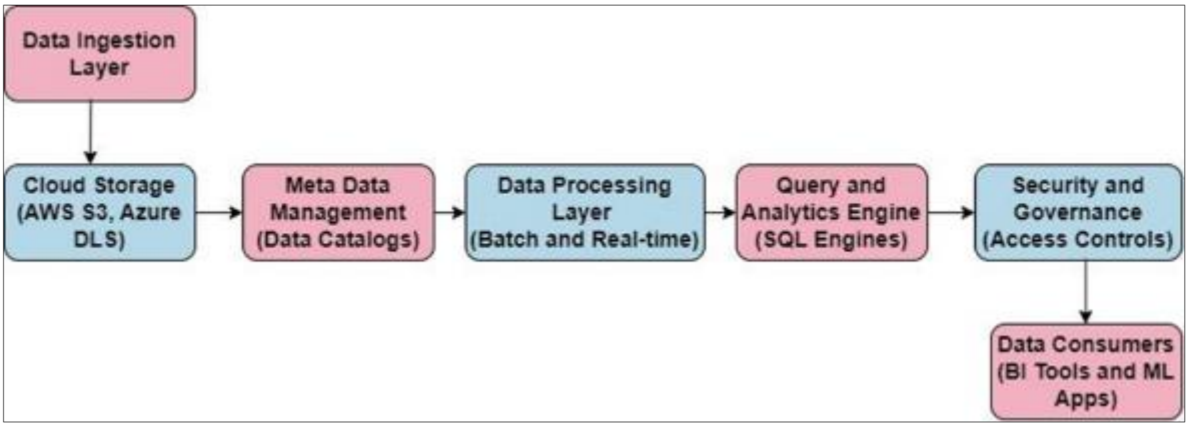
Data Ingestion Layer
Cloud Storage (AWS S3, Azure DLS)
Meta Data Management (Data Catalogs)
Data Processing Layer (Batch and Real-time)
Query and Analytics Engine (SQL Engines)
Security and Governance (Access Controls)
Data Consumers (BI Tools and ML Apps)

Figure 1 Overall Architecture of a Cloud-Based Data Lakehouse

Figure 1 shows the architecture of a cloud-native data lakehouse, with its components and how different aspects of data management fit together. Central to this architecture is the Data Ingestion Layer: a user-defined component that receives and processes data from any number of sources such as IoT devices, traditional databases and social media platforms being collected for real-time or batch-driven analytics. When stream and batch need to be handled, a layer is described which has the freedom of big data in respect of volume and varieties. After being ingested, it is stored in a Cloud Storage Layer that uses large-scale cloud storage solutions (i.e., AWS S3 or Azure Data Lake Storage). Those platforms can handle a complex range of data types, which provide the lakeshouse with comprehensive support for both structured and unstructured data.

At the heart of lakehouse architecture is a Metadata Management Layer that organizes and indexes data making it easy to find, regardless if its in a Data Lake (a la HDFS) or stored as Parquet on top of an existing cloud object store. Making Available Comprehensive Metadata Management ServicesDefaulting it with metadata management services that makes data governance and compliance easy, thus improving the quality of the data on which our business users depend upon. Underwater Data Processing Layer consists of distributed computing frameworks such as Apache Spark to perform comprehensive data preprocessing and enrichment, for both real-time processing & batch processing. It is also extended with the Query and Analytics Engine which enables users to run advanced SQL based queries, making it possible for analytics people access all kinds of data. Access management, including encryption strategies is a core part of the Security and Governance Layer to mitigate security issues. Lastly, Data Consumers (also known as BI tools and ML apps) use the data.

Cloud platforms let you scale storage and compute resources systematically, thus making it easier for firms to support the growing data sizes as well as fluctuating workloads whilst benefitting from cost efficiency. This kind of scalability is required for the variety of data types and analytics workloads that lakehouses are intended to drive. Organizations can significantly reduce costs via the pay-as-you-go pricing model of cloud services since businesses only need to bear expenses for resources that they use. This is immensely useful for data lakehouses functioning in environments like the cloud, where user workloads may vary and require resource capacity to be manually enlarged or reduced. The cloud-based lakehouse is open to many data formats, can integrate with numerous Data Sources and any analytics tool.

The cloud comes with native built-in advanced analytics capabilities, as well as machine learning services which organisations can use to uncover hidden patterns in the data. Commvault can help unlock the full potential of your data, providing organisations with capabilities that enable innovation and competitive opportunity. Deploy the data lakehouse across different regions with cloud platforms to leverage a global infrastructure which can significantly increase availability and performance of their solutions. For large organisations with wide spread operations or customers, this global reach becomes of great significance to be able to provide services consistently around the world reliably.

4. Key Architectural Components of Data Lakehouses

The architecture of a cloud-based data lakehouse consists of several key components that work together to provide a unified data management platform:

- Storage Layer: Architected to take advantages of cloud storage protocols: Keeps raw and processed data in open-format (Parquet, Avro, ORC). This layer provides support for structured and unstructured data (easier access to the stories).
- Metadata Management: It is the main tool that you will use to organize data and have proper cataloging of Data for discovery, accessibility for governance & compliance. When used collectively with data catalogs and lineage tools metadata layers allow for a more 360-degree view of all your organizations data. Both Apache Atlas and AWS Glue Data Catalog are tools that assist organizations in managing metadata well enough to do data governance, compliance etc.
- Data Processing Layer: Supports distributed computing frameworks, e.g., Apache Spark (in batch mode) or Apache Flink. For large, complex data workloads there is a computational distribution and heuristic load on this layer that allows processing transformation into the required form for analysis. Thanks to managed services such as AWS Glue, Azure Data Factory and Google Cloud Dataprep/Dataflow the data processing/orchestration complexity in cloud platforms is (at least partially) wiped away.

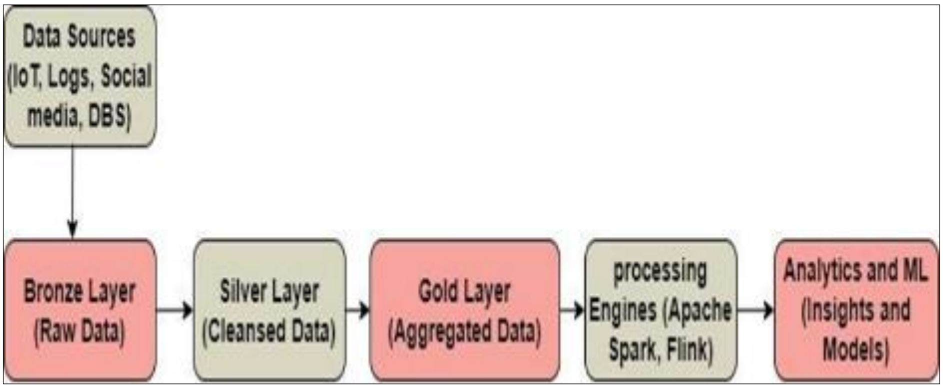
Data Sources (IoT, Logs, Social media, DBS)
Bronze Layer (Raw Data)
Silver Layer (Cleansed Data)
Gold Layer (Aggregated Data)
processing Engines (Apache Spark, Flink)
Analytics and ML (Insights and Models)

Figure 2 Data Flow and Processing in a Lakehouse

Figure 2 shows how data flow and processing pipeline work for a cloud-based Data Lakehouse Architecture. The diagram below illustrates how data is moving through different stages starting with the ingestion of a variety of sources like IoT devices, logs, social media platforms or traditional databases. The Bronze Layer is the place where all different types of data are first stored, and in its raw format. This step is important as it keeps the data still available and unmodified, so its original presence is maintained, can be examine or further analysis in future. Lakehouse architecture shines because you can store data in a raw form with dynamic, unforced schemas that make it possible to support many analytics strategies and offer strong capabilities for static or interactive use cases.

Data as is goes low the pipeline to be cleansed and prepped for usage moves onto silver layer. In this stage, duplicates are removed, errors fixed and data is transformed into structured formats to better analysis. The clean data is then maintained in the Gold Layer, and it can be aggregated and enriched to enable different analytics workloads. This tiered data arrangement - in bronze, silver and gold layers (in order) is good for differentiation of the level quality of the same set where a layer protects to be used those lower levels are less reliable. Right now, Processing Engines are the backbone of this transformation and Apache Spark or Apache Flink helps them to re-engineer in a way so as to handle stream based real-time data flow along with batch processing techniques. Finally, the enriched and processed data is consumed by **Analytics** & **Machine Learning** applications, enabling organisations to generate insights, build predictive models, and support data-driven decision-making processes. This comprehensive pipeline ensures that the lakehouse architecture delivers high-quality data ready for advanced analytics and business intelligence.

- Query Engine: Provides SQL-based querying capabilities, enabling users to perform complex analytics on data residing in the lakehouse. Query engines like Presto, Hive, or Google BigQuery offer high-performance querying and optimization features, allowing organizations to derive insights quickly and efficiently. These query engines integrate with cloud storage and processing services, ensuring seamless data access and analysis.
- Security and Governance: Implements access controls, encryption, and compliance audits to protect sensitive data and ensure regulatory adherence. Cloud platforms often provide integrated security services to manage identities and permissions effectively, safeguarding data across the data lakehouse. Services like AWS Identity and Access Management (IAM), Azure Active Directory, and Google Cloud Identity provide robust security and governance capabilities.

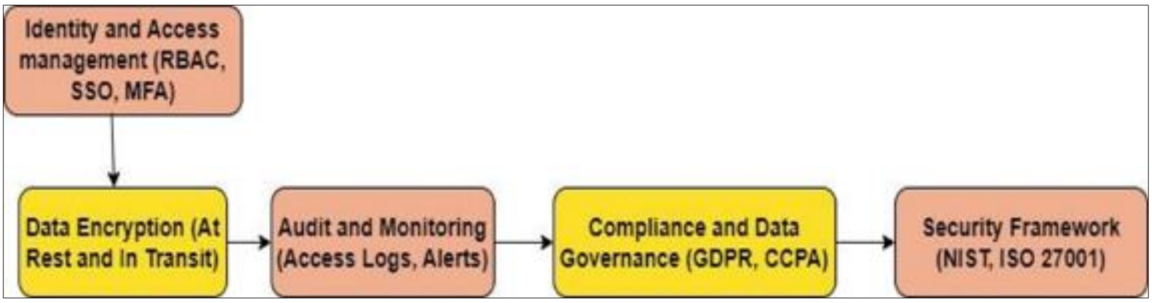
Identity and Access management (RBAC, SSO, MFA)
Data Encryption (At Rest and In Transit)
Audit and Monitoring (Access Logs, Alerts)
Compliance and Data Governance (GDPR, CCPA)
Security Framework (NIST, ISO 27001)

Figure 3 Security and Governance Framework in a Data Lakehouse

The cloud-based data lakehouse architecture opens new opportunities for integrating with a robust security and governance framework to enforce integrity, privacy (via security), and compliance as represented in Figure 3. Fundamentally with the Identity and Access Management (IAM) that enforces high-security standards such as Role-Based access control, Single Sign-On(SAML), Multi-factor Authentication(MFA). At the heart of this is a set of protocols which hook into everything else - to manage who has what permissions, and creating user profiles that grant access only to those trained on correct procedures with materials. For every organisation the IAM is fundamental and provides the desired level of security through its fine granular control over access management to monitor user activities reducing a serious business threat.

A new Data Encryption while surrounded by the IAM, will guarantee that data is in essence safe both at rest and on-the-go Encryption methods protect confidential information by encrypting data through advanced encryption algorithms that render it illegible to unauthorised individuals. This layer plays a pivotal role in keeping the data secure and accurate especially when it comes to cloud environments where data moves across storage and processing components. Also Audit and Monitoring component monitors the data access or use with deep logging and alerting systems which makes sure of all-time oversight on the subject of control over data. Compliance and Data Governance The framework finally includes a set of Compliance and Data Governance measures to meet regulatory standards like GDPR, CCPA ensuring our practices on data handling comply with legal requirements. By following these guidelines and using well-known Security Frameworks such as NIST, ISO 27001, organisations will be able to transform a comprehensive governance strategy that protects data and fosters trust among stakeholders.

5. Challenges and Future Directions

Despite the promising potential of cloud-based data lakehouses, several challenges need to be addressed to realize their full potential. These include:

- Scalability and Performance: It is important to keep a scalable and high-performance data processing/ querying capability as the amount of data that gets into these systems becoming ever bigger.
- Integration and Interoperability: Creating data lakehouses that integrate seamlessly to current systems and tools is key in increasing the insights offered by a company's data as well. Implementing standard interfaces

and APIs will support interoperability, reducing integration complexity. To succeed at it, companies need to build inter-operable data architecture that serves different types of analytical tools and raw sources.

- Data Governance and Quality: The primary concerns today are maintaining data quality and governance with all these disparate sources of truth. It enables data lineage tracking and metadata management which is important to ensure that the integrity of your datasets — thus maintaining trust. To this end, organizations need to operationalise comprehensive information governance frameworks that correspond with data quality and compliance alongside metaknowledge.
- AI and Machine Learning Integration: The use of AI and machine learning tools with data lakehouse solutions can enable organizations to take their advanced analytics and other predictive modeling abilities various notches up. The farther innovations dataset AI-driven data management automation The challenge for organisations will be to figure out how to seamlessly incorporate AI and machine learning in their data lakehouse architectures so that they can act on real-time insights and drive decisions.
- Security and Privacy: Organizations adopting cloud-based data lakehouses will not only need to safeguard sensitive information from prying eyes, but are also likely subject to a range of regulatory frameworks in accordance with the protection and security measures. Data security is key, including encryption as well as access controls and auditing to protect the data in compliance with regulations such as GDPR or CCPA

6. Conclusion

Cloud data lakehouses are a new development in the world of data management by blending Data Lakes (flexible) with Data Warehousing components and capabilities. Using cloud computing, they deliver a scalable and cost-effective platform to handle all your different data types. It finds a way to avoid the shortcoming of traditional architecture, which promotes better storage and analysis capabilities. However, despite all the opportunities to enjoy cloud-based data lakehouse goodness in addition to scalability of operations and integrated functionality; there are numerous accompanying challenges as well like execution environments requirement concerns at every touchpoint, platform scaling for desired performance tuning yet keeping valuation agreement leveraged from base line etc. These challenges must be addressed to deliver on the full vision of data lakehouse.

References

[1] Kim, W. C., & Mauborgne, R. A. (2023). Beyond disruption: Innovate and achieve growth without displacing industries, companies, or jobs. Harvard Business Press.

[2] Shah, T. R. (2022). Can big data analytics help organisations achieve sustainable competitive advantage? A developmental enquiry. Technology in Society, 68, 101801.

[3] babu Nuthalapati, S., & Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning.

[4] Morabito, V., & Morabito, V. (2015). Managing change for big data driven innovation. Big Data and Analytics: Strategic and Organizational Impacts, 125-153.

[5] Jishamol, T. R., & Bushara, A. R. (2016). Enhancement of Uplink Achievable Rate and Power Allocation in LTE-Advanced Network System. International Journal of Science Technology & Engineering, Volume 3, Issue 03.

[6] Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C. O., Gronauer, A., Pejakovic, V., ... & Stampfer, K. (2022). Digital transformation in smart farm and forest operations needs human-centered AI: challenges and future directions. Sensors, 22(8), 3043.

[7] Nuthalapati, A. (2023). Smart Fraud Detection Leveraging Machine Learning For Credit Card Security. Educational Administration: Theory and Practice, 29(2), 433-443.

[8] Costa, C., Chatzimilioudis, G., Zeinalipour-Yazti, D., & Mokbel, M. F. (2017, April). Efficient exploration of telco big data with compression and decaying. In 2017 IEEE 33rd international conference on data engineering (ICDE) (pp. 1332-1343). IEEE.

[9] Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data lakes: A survey of functions and systems. IEEE Transactions on Knowledge and Data Engineering, 35(12), 12571-12590.

[10] Babu Nuthalapati, S. (2023). AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking. Educational Administration: Theory and Practice, 29(1), 357-368.

[11] Errami, S. A., Hajji, H., El Kadi, K. A., & Badir, H. (2023). Spatial big data architecture: from data warehouses and data lakes to the Lakehouse. Journal of Parallel and Distributed Computing, 176, 70-79.

[12] Mahanti, R., & Mahanti, R. (2021). Data and its governance. Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools, 5-82.

[13] Muhammed Kunju, A. K., Baskar, S., Zafar, S., & AR, B. (2024). A transformer based real-time photo captioning framework for visually impaired people with visual attention. Multimedia Tools and Applications, 1-20.

[14] Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2024). The Lakehouse: State of the Art on Concepts and Technologies. SN Computer Science, 5(5), 1-39.

[15] Bushara, A. R., RS, V. K., & Kumar, S. S. (2024). The Implications of Varying Batch-Size in the Classification of Patch-Based Lung Nodules Using Convolutional Neural Network Architecture on Computed Tomography Images. Journal of Biomedical Photonics & Engineering, 10(1), 39-47.

[16] Bianchini, D., De Antonellis, V., & Garda, M. (2024). A semantics-enabled approach for personalised Data Lake exploration. Knowledge and Information Systems, 66(2), 1469-1502.

[17] Azzabi, S., Alfughi, Z., & Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. Computers, 13(7), 183.

[18] Bauer, D., Froese, F., Garcés-Erice, L., Giblin, C., Labbi, A., Nagy, Z. A., ... & Wespi, A. (2021). Building and operating a large-scale enterprise data analytics platform. Big Data Research, 23, 100181.

[19] Shah, S. T. U. (2024). Optimizing Data Warehouse Implementation on Azure: A Comparative Analysis of Efficient Data Warehousing Strategies on Azure.

[20] Farhan, M. S., Youssef, A., & Abdelhamid, L. (2024). A Model for Enhancing Unstructured Big Data Warehouse Execution Time. Big Data and Cognitive Computing, 8(2), 17.

[21] AR, B. (2022). A deep learning-based lung cancer classification of CT images using augmented convolutional neural networks. ELCVIA electronic letters on computer vision and image analysis, 21(1).

[22] Leng, J., Yan, D., Liu, Q., Zhang, H., Zhao, G., Wei, L., ... & Chen, X. (2021). Digital twin-driven joint optimisation of packing and storage assignment in large-scale automated high-rise warehouse product-service system. International Journal of Computer Integrated Manufacturing, 34(7-8), 783-800.

[23] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8, p. 28).

[24] Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., & Vassilakis, T. (2011). Dremel: interactive analysis of web-scale datasets. Communications of the ACM, 54(6), 114-123.

[25] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., ... & Stoica, I. (2012). Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In 9th USENIX symposium on networked systems design and implementation (NSDI 12) (pp. 15-28).

[26] Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., & Rasin, A. (2009). HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proceedings of the VLDB Endowment, 2(1), 922-933.

[27] Chaiken, R., Jenkins, B., Larson, P. Å., Ramsey, B., Shakib, D., Weaver, S., & Zhou, J. (2008). Scope: easy and efficient parallel processing of massive data sets. Proceedings of the VLDB Endowment, 1(2), 1265-1276.

[28] Stonebraker, M., Abadi, D. J., Batkin, A., Chen, X., Cherniack, M., Ferreira, M., ... & Zdonik, S. (2018). C-store: a column-oriented DBMS. In Making Databases Work: the Pragmatic Wisdom of Michael Stonebraker (pp. 491-518).

[29] Saha, B., Shah, H., Seth, S., Vijayaraghavan, G., Murthy, A., & Curino, C. (2015, May). Apache tez: A unifying framework for modeling and building data processing applications. In Proceedings of the 2015 ACM SIGMOD international conference on Management of Data (pp. 1357-1369).

[30] Xin, R. S., Gonzalez, J. E., Franklin, M. J., & Stoica, I. (2013, June). Graphx: A resilient distributed graph system on spark. In First international workshop on graph data management experiences and systems (pp. 1-6).

[31] Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache flink: Stream and batch processing in a single engine. The Bulletin of the Technical Committee on Data Engineering, 38(4).

[32] Karpathiotakis, M., Branco, M., Alagiannis, I., & Ailamaki, A. (2014). Adaptive query processing on RAW data. Proceedings of the VLDB Endowment, 7(12), 1119-1130.