2022mt93331

by Akshay SHELKE

Submission date: 26-Mar-2024 05:40PM (UTC+0530)
Submission ID: 2330961453
File name: 2022mt93331.pdf (1.81M)
Word count: 5937
Character count: 36598

Dissertation Title:

Cloud-First Approach: Engineering a Solution for Efficient On-Premises Data Migration to Cloud Platforms

Course No. : SEZG628T

Course Title: Dissertation

Dissertation Work Done by:

Student Name: SHELKE AKSHAY NANDKUMAR

BITS ID: 2022MT93331

Degree Program: M.Tech. Software Engineering

Research Area: Cloud Computing

Dissertation work carried out at:

Schlumberger India Technology Centre Ltd, Pune

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

VIDYA VIHAR, PILANI, RAJASTHAN - 333031.

March 2024

Dissertation Title:

Cloud-First Approach: Engineering a Solution for Efficient On-Premises Data Migration to Cloud Platforms

Course No. : SEZG628T

Course Title: Dissertation

Dissertation Work Done by:

Student Name: SHELKE AKSHAY NANDKUMAR

BITS ID: 2022MT93331

Degree Program: M.Tech. Software Engineering

Research Area: Cloud Computing

Dissertation work carried out at:

Schlumberger India Technology Centre Ltd, Pune

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

VIDYA VIHAR, PILANI, RAJASTHAN - 333031.

March 2024

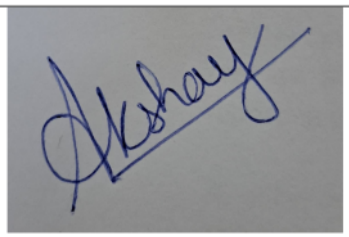
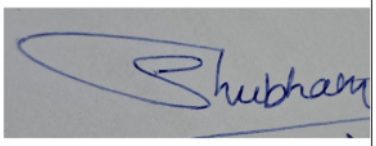
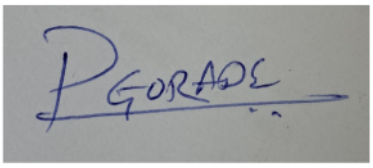
Abstract

In response to the escalating demands for comprehensive data management solutions, this project endeavors to construct an intricate yet seamlessly integrated End-to-End Azure Data Engineering system. The project spans crucial phases of the data lifecycle, addressing Data Ingestion, Transformation, Loading, and Reporting, with a strategic amalgamation of diverse Azure services and tools.

At the heart of the project, Azure Data Factory (ADF) serves as the orchestrator, efficiently managing workflows critical for the fluid movement of data. Leveraging Azure Data Lake Storage (Gen2) ensures a secure and the scalable repository for accommodating a variety of data types, fostering flexibility and reliability in storage solutions. The integration of Azure Databricks into the project introduces a robust data transformation layer, allowing for the refinement of raw data into structured and actionable format. This transformed data, is then seamlessly loaded into the Azure Synapse Analytics, setting the stage for insightful reporting and analysis.

Recognizing the pivotal role of security and governance in modern data ecosystems, Azure Key Vault and Azure Active Directory(AAD) play crucial roles in the project. Azure Key Vault(AKV) ensures secure key management, safeguarding sensitive information, while Azure Active Directory(AAD) enhances identity & access management, contributing to a secure and compliant data environment.

The use case revolves around ingesting tables from an on-premise SQL Server database using Azure Data Factory, followed by the storage of this data in Azure Data Lake. Azure Databricks comes into play to meticulously transform the raw data into a refined state, optimizing it for subsequent analysis. The cleaned data is then efficiently loaded into Azure Synapse Analytics, forming the foundation for creating actionable insights. The reporting phase is facilitated by Microsoft Power BI, which seamlessly integrates with Azure Synapse Analytics to craft interactive dashboards.

		
Signature of Student	**Signature of Supervisor**	**Signature of Additional Examiner**
Name: SHELKE AKSHAY NANDKUMAR	**Name: Mr. Shubham Tulsyan**	**Name: Mr. Prashant Gorade**
Date : 23rd March 2024	**Date : 23rd March 2024**	**Date : 23rd March 2024**
Place : Pune, Maharashtra, India	**Place : Pune, Maharashtra, India**	**Place : Pune, Maharashtra, India**

Contents

1. Broad Area of Work

In the realm of contemporary data management, enterprises face challenges in establishing efficient and cohesive data workflows. Organizations often grapple with issues related to data integration, transformation, and reporting, especially when dealing with diverse data sources and formats. The need for a streamlined End-to-End data engineering solution becomes critical as businesses strive to get actionable insights from their datasets while maintaining security, compliance, and efficiency. The absence of a comprehensive data engineering platform can lead to disjointed processes, increased complexity, and hindered decision-making capabilities. Addressing these challenges is the primary focus of this project.

The broad area of work for this project falls within the domain of Azure Data Engineering and Integration. It encompasses the end-to-end processes involved in managing and deriving value from data within the Azure ecosystem. Specifically, the project revolves around orchestrating seamless workflows for data ingestion, transformation, loading, and reporting utilizing a suite of Azure services and tools. The broader context includes addressing challenges related to data management, analytics, and the need for scalable, secure, and integrated solutions for enterprises dealing with diverse data sources.

2. Background Research and Literature Review

The project is grounded in the context of the evolving landscape of data management and analytics. Data migration to cloud platforms has become a critical endeavor for organizations seeking to leverage the benefits of the cloud computing, such as scalability, cost-effectiveness, and advanced analytics capabilities. However, the process of migrating data from on-premises infrastructure to cloud platforms is often fraught with challenges, necessitating a comprehensive understanding of the underlying technologies, best practices, and potential pitfalls.

One of the key considerations, in data migration is the selection of appropriate cloud platforms and services. Azure, Amazon Web Services (AWS),Google Cloud Platform (GCP) are among the leading cloud providers,offering the wide range of services, for data storage, data processing, and data analytics.Azure provides a robust set of tools and the services tailored for data migration, including Azure Databricks, Azure Data Factory, Azure Data Lake (Gen2), and Synapse Analytics.

Azure Databricks is the cloud-based Apache Spark platform, which used for scalable and efficient data processing and analytics. It provides the collaborative environment for

the data engineers,data scientists,and the data analysts to work together on the data-intensive workloads. Azure Data Factory, on the other hand, is a fully managed data integration service that simplifies the process of ingesting, preparing, and transforming data from various sources.

Azure Data Lake (Gen2) is a highly secure and scalable data lake solution that integrates seamlessly with other Azure services, enabling organizations to store and then analyze vast amounts of structured and the unstructured data.Azure Synapse Analytics, is a cloud-based analytics service which combines traditional data warehousing capabilities with big data analytics.

Research has shown that organizations often face challenges in migrating data due to the complexity of their on-premises infrastructure, data silos, and legacy systems. Data migration requires careful planning, understanding of data lineage, and adherence to governance and security protocols. Failure to address these challenges can lead to data loss, security breaches, or compliance issues.

Several studies have explored different approaches and methodologies for data migration to cloud platforms. One such approach is the Extract, Transform, Load (ETL) process, which does involve extracting data from on-premises sources,transforming it into the suitable format,and loading it into the cloud storage or processing systems. Another approach is the Extract, Load, Transform (ELT) process, which does involve loading the data into the cloud first and then transforming it using cloud-based processing services.

Research has also highlighted the importance of data quality and governance during the migration process. Data quality issues, such as inconsistencies, duplicates, and missing values, can significantly impact the accuracy and reliability of analytical insights derived from the migrated data. Implementing robust data governance frameworks and data lineage tracking mechanisms is crucial to ensure data integrity and compliance with regulatory requirements.

Furthermore, researchers have explored the potential of the cloud-based data lakes and data warehousing solutions for enabling advanced analytics and ML (machine learning) capabilities. By leveraging the scalability and processing power of cloud platforms, organizations can unlock insights from vast amounts of data, enabling data-driven decision-making and innovation.

Overall, the literature review highlights the growing importance of data migration to cloud platforms and the need for comprehensive solutions that address the challenges associated with this process. By combining theoretical research and practical implementation, this Dissertation project aims to contribute to the body of knowledge in this field and provide a robust solution for efficient on-premises data migration to cloud platforms.

3. Problem Definition

Data migration from on-premises infrastructure to cloud platforms is a critical endeavor for organizations seeking to leverage the benefits of cloud computing, like scalability, cost-effectiveness, and advanced analytics capabilities. However, this process is often fraught with challenges that must be addressed to ensure a successful and efficient migration.

One of the **primary challenges in data migration** is the complexity of on-premises infrastructure and data silos. Organizations usually have data stored in various formats, locations, and legacy systems, making it not easy to consolidate and migrate data seamlessly. Furthermore, data governance and security requirements must be adhered to during the migration process to maintain data integrity, privacy, and compliance with regulatory standards.

Another challenge lies in the sheer volume of data that must be migrated. As organizations generate and store massive amounts of data, the migration process can become time-consuming and resource-intensive, leading to potential performance bottlenecks and increased costs. Ensuring data quality and consistency during the migration process is also crucial to enable accurate and reliable analytical insights from the migrated data.

Moreover, the migration process requires careful planning and execution to minimize downtime and business disruptions. Failure to properly plan and execute the migration can result in the loss of data, security breaches, or compliance issues, potentially leading to significant financial,reputational consequences for the organization.

To address these challenges, a comprehensive solution is needed that encompasses various stages of the data migration process, including data extraction, transformation, loading, and integration. ***This solution should leverage advanced technologies and best practices to streamline the migration process, while ensuring data security, governance, and compliance.***

Additionally, ***the solution*** should provide organizations with the ability to optimize data storage, processing, and querying in the cloud environment, enabling them to harness the full potential of cloud computing for advanced analytics and data-driven decision-making.

By addressing these challenges and providing a robust and efficient data migration solution, organizations can seamlessly transition their data assets to the cloud, unlocking new opportunities for scalability, cost-efficiency, and innovation.

4. Project Objectives

The primary objective of this Dissertation project is to engineer a comprehensive and efficient solution for on-premises data migration to cloud platforms, following a cloud-first approach. By leveraging cutting-edge technologies and best practices, the project aims to address the challenges and complexities associated with data migration, enabling organizations to unlock the benefits of cloud computing seamlessly. Key objectives include:

- **Efficient Data Workflow:** Develop a streamlined and efficient data workflow covering ingestion, transformation, loading, and reporting.

- **Seamless Integration:** Showcase the seamless integration of Azure services, such as Azure Data Factory, Data Lake Storage Gen2, Databricks, Synapse Analytics, Key Vault, AAD, and Power BI, to construct a cohesive data platform.

- **Data Governance and Security:** Implement robust data governance and security measures using Azure Key Vault and Azure Active Directory to ensure compliance and safeguard sensitive information.

- **Insightful Reporting:** Utilize Microsoft Power BI to create interactive dashboards, providing actionable insights for informed decision-making.

- **Use Case Demonstration:** Apply the solution to a practical use case involving the ingestion of on-premise SQL Server data, transformation through Databricks, loading into Synapse Analytics, and reporting through Power BI.

By achieving these objectives, the project aims to demonstrate the versatility and efficacy of Azure services in addressing contemporary data management challenges and empowering enterprises with a comprehensive data engineering platform.

One of the key objectives is to develop a streamlined and scalable data migration framework that can handle large volumes of data from various on-premises sources. This framework should facilitate the extraction, transformation, and loading of data into cloud storage and processing systems, ensuring data integrity and consistency throughout the migration process.

Another critical objective is to implement robust data governance and security measures to ensure compliance with industry standards and regulatory requirements. This includes establishing data lineage tracking mechanisms, implementing data encryption and access controls, and adhering to data privacy and protection regulations, such as the General Data Protection Regulation (GDPR).

The project also aims to optimize data storage and processing in the cloud environment, leveraging the scalability and cost-effectiveness of cloud platforms. This objective involves exploring techniques for efficient data partitioning, indexing, and compression, as well as leveraging cloud-based services for distributed processing and analytics.

Furthermore, the solution should enable organizations to harness the power of advanced analytics and machine learning capabilities in the cloud. By migrating data to cloud platforms, organizations can leverage scalable computing resources and cutting-edge analytics tools to derive insights from large and complex datasets, enabling data-driven decision-making and innovation.

Additionally, the project aims to develop a user-friendly and intuitive interface or dashboard for monitoring and managing the data migration process. This interface should provide real-time visibility into the migration status, data quality metrics, and any potential issues or bottlenecks, allowing organizations to take proactive measures and ensure a smooth migration experience.

To achieve these objectives, the project will leverage a combination of theoretical research and practical implementation. Extensive literature review and analysis of existing data migration solutions, best practices, and industry standards will be conducted to establish a solid foundation for the project.

Moreover, the project will involve the selection and integration of appropriate cloud platforms and services, such as Azure Databricks, Azure Data Factory, Azure Data Lake Gen2, and Azure Synapse Analytics. These tools will be utilized to build a robust and scalable data migration pipeline, enabling efficient data extraction, transformation, loading, and integration.

Throughout the project, emphasis will be placed on thorough testing, validation, and performance optimization to ensure the solution meets the highest standards of reliability, efficiency, and scalability.

By achieving these objectives, the Dissertation project will contribute to the broader field of data engineering and cloud computing, providing organizations with a comprehensive and efficient solution for on-premises data migration to cloud platforms. This solution will enable organizations to unlock the full potential of cloud computing, driving innovation, cost-efficiency, and data-driven decision-making.

5. Scope of Work

The scope of work for this project is comprehensive, covering various facets of data engineering and integration within the Azure environment. Key aspects of the scope include:

- **Data Ingestion:** Involves the extraction and ingestion of data from on-premise SQL Server databases into the Azure environment using Azure Data Factory.

- **Data Transformation:** Encompasses the process of refining raw data into a structured and actionable format using Azure Databricks, ensuring data quality and usability.

- **Data Loading:** Focuses on the efficient loading of cleaned and transformed data into Azure Synapse Analytics, facilitating advanced analytics and reporting.

- **Security and Governance:** Incorporates the implementation of robust security measures using Azure Key Vault and Azure Active Directory for secure key management, identity, and access management, ensuring compliance and governance.

- **Reporting and Analytics:** Utilizes Microsoft Power BI for creating interactive dashboards, providing stakeholders with insightful visualizations and analytics capabilities.

- **Monitoring and Governance:** Establishes a framework for monitoring and governance throughout the data lifecycle, ensuring data integrity, security, and compliance.

- **Use Case Implementation:** Demonstrates the entire data engineering workflow in a practical scenario by ingesting on-premise SQL Server data, transforming it using Databricks, loading it into Synapse Analytics, and creating interactive reports with Power BI.

The scope emphasizes a holistic approach to Azure Data Engineering, integrating various Azure services to provide end-to-end solutions for enterprises aiming to harness the power of their data effectively and securely.

6. Tools & Technologies used for Data Migration

This Dissertation project will leverage a range of powerful tools and technologies to engineer an efficient solution for on-premises data migration to cloud platforms. The chosen tools and technologies are designed to address the various challenges and requirements associated with the data migration process, ensuring a seamless and secure transition to the cloud environment.

Python:

Python is a versatile and widely-used programming language that will play a crucial role in this project. Its extensive ecosystem of libraries and frameworks will be utilized for various tasks, including data extraction, transformation, integration. Some of the key Python libraries that will be leveraged include:

- **Pandas**: A powerful data manipulations,analysis library that provides high-performance data structures and the data analysis tools.

Azure Databricks (ADB):

This is a cloud-based Apache Spark platform that will serve as the foundation for distributed data processing and analytics in this project. It provides an environment for the data engineers,data scientists, and data analysts to work together on data-intensive workloads. Key features of Azure Databricks that will be leveraged include:

- **Spark Core:** The core engine for the large-scale data processing,supporting batch, streaming, and interactive workloads.
- **Spark SQL:** A module for the structured data processing,enabling distributed SQL queries and data analysis.

Azure Data Factory (ADF):

ADF is a fully managed data integration service that will streamline the process of ingesting, preparing, and transforming data from various on-prem and cloud sources. Its visual authoring interface and automated orchestration capabilities will be utilized to build the scalable and reliable data pipelines. Key features of ADF include:

- **Data Movement:** Ability to move data between on-premises and cloud sources, supporting a wide range of connectors and formats.
- **Data Transformation:** Built-in and custom data transformation activities for data cleansing, enrichment, and transformation.
- **Control Flow:** Orchestration and scheduling of data pipelines, including dependency management and error handling.
- **Logging and Monitoring:** Comprehensive logging and monitoring capabilities for tracking pipeline execution and pipeline troubleshooting issues.

Azure Data Lake (Gen2):

It is a highly scalable and highly secure data lake solution that will serve as the central storage repository for the structured and the unstructured data in this project. Its tight integration with other Azure services and support for various data formats make it as an ideal choice for storing and then analyzing large volumes of data. Key features of Azure Data Lake Gen2 include:

- **Unlimited Storage:** Virtually unlimited storing capacity and scalability, enabling organizations to do store and process vast amounts of data.
- **Secure Access Control:** Granular access control and auditing capabilities for data security and governance.
- **Performance Optimization:** Support for partitioning and indexing, enabling efficient data querying and analysis.
- **Integration with Analytics Services:** Seamless integration with Azure Databricks, Azure Synapse Analytics, and other Azure data services.

Azure Synapse Analytics:

Azure Synapse Analytics, formerly known as Azure SQL Data Warehouse, is a cloud-based analytics service that combines traditional data warehousing capabilities with big data analytics. It will be utilized in this project for large-scale data warehousing, advanced analytics, and business intelligence (BI) workloads. Key features of Azure Synapse Analytics include:

- **Massively Parallel Processing (MPP):** Highly scalable and parallel processing architecture for querying and analyzing large datasets.
- **Data Virtualization:** Ability to query data from various sources, including Azure Data Lake Gen2, without data movement.
- **Advanced Analytics:** Support for machine learning (ML) and advanced analytics through integration with Azure Databricks and other Azure services.
- **BI Integration:** Native integration with Power BI for data visualization and reporting.

Power BI:

It is a business analytics service from Microsoft that will be used in this project for data visualization and reporting. Its set of features and intuitive user interface make it an ideal choice for creating interactive dashboards and reports, enabling data-driven decision-making. Key features of Power BI include:

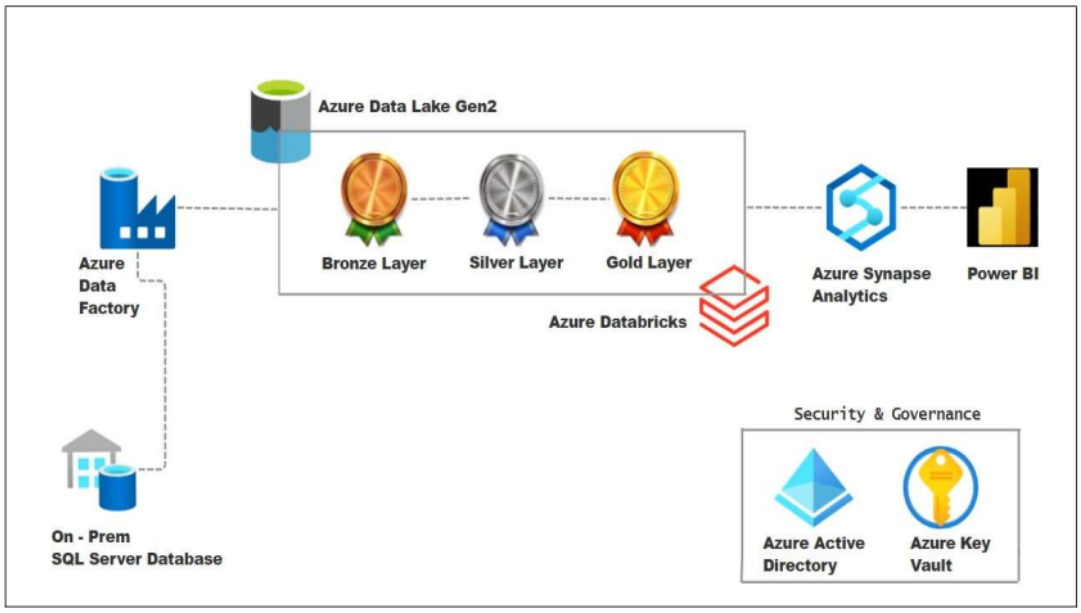
- **Data Connectivity:** Ability to connect to a various data sources, including Azure Data Lake Gen2,Azure Synapse Analytics,other cloud & on-premises data sources.
- **Data Modeling:** Advanced data modeling capabilities, including support for relationships, hierarchies, and calculated columns.
- **Interactive Visualizations:** A vast array of visualization types and customization options for creating compelling and insightful reports.
- **Sharing and Collaboration:** Seamless sharing and collaboration features, enabling secure distribution of reports and dashboards.

In addition to these tools and technologies, various other libraries, frameworks, and best practices will be leveraged throughout the project. This includes implementing robust data governance and security measures, adhering to industry standards and regulatory requirements, and ensuring efficient data partitioning, indexing, and compression.

By combining the power of Python, Azure Databricks, Azure Data Factory, Azure Data Lake Gen2, Azure Synapse Analytics, and Power BI, this Dissertation project will deliver a comprehensive and efficient solution for on-premises data migration to cloud platforms. The chosen tools and technologies will enable organizations to do unlock the full potential of cloud computing, driving innovation, cost-efficiency, and data-driven decision-making.

7. Overview of Architectural Design

In today's data-driven world, the organizations are constantly striving to unlock the true potential of their data assets. The ability to ingest, store, process, and analyze data from various sources has become crucial for driving business intelligence, making informed decisions, and gaining a competitive edge. Microsoft Azure,offers a comprehensive suite of services and tools to build robust and scalable data platforms.



The provided architectural diagrams showcase a modern data platform built on Azure services, enabling organizations to streamline their data lifecycle from ingestion to reporting. This documentation will delve into the intricacies of these diagrams, explaining each component's role and how they collectively contribute to a cohesive and efficient data ecosystem.

Architectural Diagram Overview:

The architectural diagrams depict a data platform designed to handle the entire data lifecycle, from ingestion to reporting, while ensuring security and governance. The key components of this architecture are:

1.Data Ingestion

2.Data Storage

3.Data Transformation and Processing

4.Data Analytics and Reporting

5.Security and Governance

Let's dive deeper into each component and its significance within the overall architecture.

1. Data Ingestion:

It is the initial stage of the data lifecycle, where raw data is gathered from various sources and brought into the data platform. In the provided diagrams, Azure Data Factory is utilized as the primary tool for orchestrating data ingestion.

Azure Data Factory(ADF) is a cloud-based data integration service that allows organizations to create,schedule,and manage data pipelines. These pipelines can ingest data from various sources,including on-premises databases, cloud storage services, SaaS applications, and more. The diagrams illustrate an on-premises SQL Server database as one potential data source, showcasing the ability to integrate both on-premises and cloud-based data sources seamlessly.

The ingestion process typically involves extracting data from the source systems, performing any necessary transformations or validations, and loading the data into a centralized data storage solution. In this architecture, the ingested data is landing in the Azure Data Lake Gen2, a highly scalable and secure data lake storage service offered by Azure.

2. Data Storage:

Azure Data Lake Gen2 serves as the central repository for storing and managing data within this architecture. It is designed to handle massive volumes of structured, semi-structured, and unstructured data, making it an ideal choice for modern data platforms that deal with diverse data types.

The Data Lake Gen2 is organized into three distinct layers: Bronze, Silver, and Gold. This layered approach, also known as the Data Lake House architecture, facilitates effective data management and governance.

a. Bronze Layer:

The Bronze layer is the landing zone for raw, unprocessed data ingested from various sources. This layer serves as a secure and reliable storage location for the initial data ingestion, ensuring data integrity and auditability. The Bronze layer acts as a single source of truth for the organization's raw data, enabling data lineage and traceability.

b. Silver Layer:

The Silver layer contains transformed and cleaned versions of the data from the Bronze layer. In this layer, data undergoes processing and transformations to enhance its quality, consistency, and usability. Common operations performed in the Silver layer include data deduplication, schema normalization, data type conversions, and data enrichment from external sources.

c. Gold Layer:

The Gold layer represents the highest level of data refinement and curation. It stores the aggregated, curated, and enriched data that is ready for consumption by analytical tools and reporting applications. The data in the Gold layer is typically organized in a format optimized for efficient querying and analysis, such as columnar storage or denormalized schemas.

By separating data into these layers, the architecture promotes data governance, auditability, and versioning. It also enables efficient data processing by allowing transformations and computations to be performed incrementally, reducing redundancy, and minimizing the overall processing overhead.

3. Data Transformation and Processing:

Data transformation and processing are critical steps in the data lifecycle, ensuring that raw data is converted into a usable and valuable format for analysis and decision-making. In the provided architecture, Azure Databricks is leveraged as the primary platform for data transformation and processing tasks.

Azure Databricks is a fully managed Apache Spark-based analytics service that enables organizations to process and transform large volumes of data efficiently. It provides a collaborative environment for data engineers, data scientists, and analysts to develop, deploy, and manage Spark-based applications and workflows.

Within the architectural diagrams, Azure Databricks is represented as the central component responsible for transforming and processing data as it moves from the Bronze layer to the Silver and Gold layers in the Data Lake. Databricks allows users to write and execute Spark jobs using various programming languages, including Python, Scala, R, and SQL.

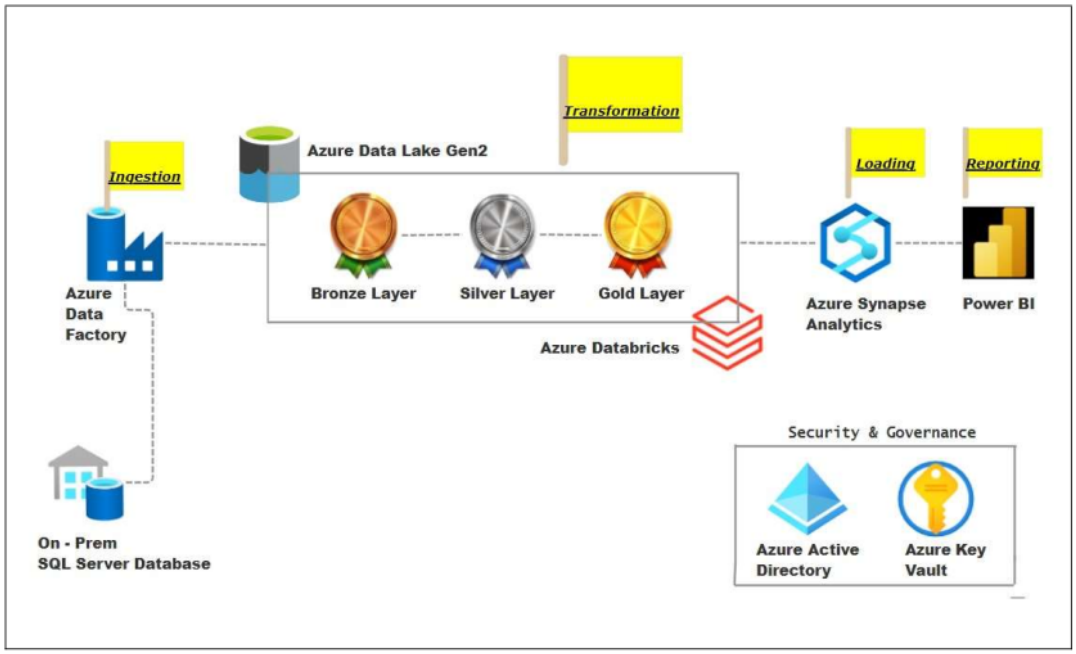
The transformation and processing tasks performed by Databricks can include:

- **Data cleaning and validation**: Removing duplicates, handling missing values, and enforcing data integrity rules.
- **Data enrichment**: Combining data from multiple sources, joining datasets, and adding contextual information.
- **Feature engineering**: Deriving new features or transforming existing ones to enhance the data's predictive power for machine learning models.
- **ETL (Extract, Transform, Load) pipelines**: Extracting data from the Bronze layer, applying transformations, and loading the processed data into the Silver and Gold layers.
- **Data aggregation and summarization**: Performing rollups, aggregations, and summarizations to create curated datasets for reporting and analysis.

By leveraging Databricks' scalable and distributed computing capabilities, organizations can efficiently process and transform large volumes of data, enabling faster time-to-insight and more accurate decision-making.

4. Data Analytics and Reporting:

Once the data has been transformed and curated in the Gold layer, it is ready for consumption by analytical tools and reporting applications. The architectural diagrams showcase two key components for data analytics and reporting: Azure Synapse Analytics and Power BI.



a. Azure Synapse Analytics:

Azure Synapse Analytics is a limitless analytics service that brings together data integration, enterprise data warehousing, and big data analytics. It provides a unified experience for querying and analyzing data across structured and unstructured sources, enabling organizations to gain valuable insights from their data assets.

Within the architecture, Azure Synapse Analytics is positioned to interact with the Gold layer of the Data Lake, where the curated and aggregated data resides. By leveraging Synapse Analytics' powerful querying and analytical capabilities, organizations can perform complex analyses, generate reports, and uncover actionable insights from their data.

Synapse Analytics supports various query languages, including SQL and Spark, allowing analysts and data scientists to leverage their preferred tools and programming languages. It also integrates with other Azure services, such as Azure Machine Learning and Azure Databricks, enabling advanced analytics and machine learning workloads.

b. Power BI:

Power BI is a business intelligence and data visualization tool offered by Microsoft. It empowers organizations to create visually appealing and interactive reports, dashboards, and data visualizations, enabling better data storytelling and decision-making.

In the architectural diagrams, Power BI is shown as a separate component connected to Azure Synapse Analytics. This integration allows Power BI to seamlessly access and visualize the data stored in the Gold layer of the Data Lake, leveraging the analytical power of Synapse Analytics.

Power BI provides a user-friendly interface for creating custom reports, dashboards, and visualizations, making it accessible to a wide range of users, from business analysts to executives. It supports a variety of data sources, including Azure services, on-premises databases, and cloud-based applications, enabling a unified view of data across the organization.

5. Security and Governance:

Ensuring data security and governance is of paramount importance in any modern data platform. The architectural diagrams highlight two key Azure services responsible for managing security and governance aspects: Azure Active Directory and Azure Key Vault.

a. Azure Active Directory (AAD):

AAD is a cloud-based identity and access management service. It plays a crucial role in managing the user identities,enabling secure access to Azure resources, and enforcing access control policies within the data platform.

In the context of the architectural diagrams, Azure AD is responsible for authenticating and authorizing users and applications accessing the various components of the data platform, such as Azure Data Factory(ADF),Azure Databricks(ADB),and Azure Synapse Analytics. It ensures that only authorized personnel and applications can interact with the data, preventing unauthorized access and potential data breaches.

Azure AD supports features like multi-factor authentication,conditional access policies, and role-based access control(RBAC), providing granular control over who can perform specific actions within the data platform. It does integrate seamlessly with the other Azure services, enabling a consistent and secure authentication experience across the entire data ecosystem.

b. Azure Key Vault(Azure KV):

Azure Key Vault(Azure KV) is a secure and centralized storage solution for managing cryptographic keys, secrets, and certificates. It plays a critical role in protecting sensitive information, such as database connection strings, API keys, and encryption keys, within the data platform.

In the architectural diagrams, Azure KV is depicted as a separate component responsible for securely storing and managing encryption keys and the other secrets required by various components of the data platform. For example, the sample Python code snippet demonstrates how Azure KV can be used to securely retrieve the account key for accessing the Azure Data Lake (Gen2) storage account.

By centralizing the management of secrets and encryption keys, Azure KV simplifies key lifecycle management, enables secure key distribution, and provides robust access control mechanisms. It also supports key rotation and auditing, ensuring compliance with industry standards and regulatory requirements.

With Azure AD and Azure Key Vault working in tandem, the data platform achieves a robust security and governance framework, protecting sensitive data and ensuring adherence to organizational policies and industry best practices.

8. Use Cases for Data Migration

The solution engineered through this Dissertation project, aimed at enabling efficient on-premises data migration to cloud platforms, has numerous potential use cases across various industries and domains. Here are some notable examples:

1. Healthcare Industry:

- **Electronic Health Records (EHR) Migration:** Healthcare organizations often struggle with managing and analyzing large volumes of patient data stored in on-premises systems. This solution can facilitate the migration of EHR data to cloud platforms, enabling scalable storage, advanced analytics, and improved patient care through data-driven insights.
- **Medical Imaging Data Migration:** Medical imaging data, like X-rays,CT scans, MRI images, can be migrated to cloud storage using this solution, allowing healthcare professionals to access and share these data seamlessly, improving collaboration and diagnosis accuracy.

2. Financial Services:

- **Historical Transactions Data Migration:** Banks and financial institutions generate massive amounts of transactional data that need to be securely stored and analyzed for regulatory compliance, fraud detection, and customer insights. This solution can enable the migration of historical transactions data to cloud platforms, ensuring scalability, security, and advanced analytics capabilities.
- **Risk Data Migration:** Financial institutions must manage and analyze vast amounts of risk data to comply with regulations and mitigate potential risks. This solution can facilitate the migration of risk data to cloud platforms, enabling real-time risk monitoring, modeling, and reporting.

3. Retail and E-commerce:

- **Customer Data Migration:** Retailers and e-commerce companies often have customer data scattered across various on-premises systems. This solution can help migrate customer data to cloud platforms, enabling centralized data management, advanced customer segmentation, and personalized marketing strategies.
- **Sales and Inventory Data Migration:** By migrating sales and inventory data to cloud platforms using this solution, retailers can gain real-time visibility into sales trends, inventory levels, and supply chain operations, enabling data-driven decision-making and optimized inventory management.

4. Manufacturing and Supply Chain:

- **IoT and Sensor Data Migration:** Manufacturing companies and supply chain operations generate massive volumes of IoT and sensor data from various sources, such as machinery, equipment, and logistics systems. This solution can facilitate the migration of these data to cloud platforms, enabling real-time monitoring, predictive maintenance, and supply chain optimization.

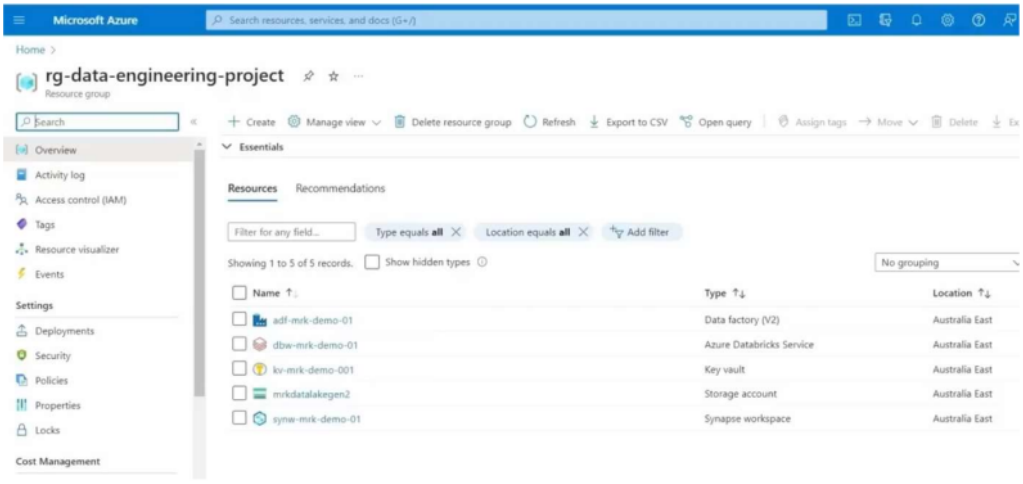
- **Product Lifecycle Data Migration:** Product lifecycle data, including design specifications, manufacturing processes, and quality control data, can be migrated to cloud platforms using this solution, improving collaboration, data accessibility, and product quality assurance.

5. Media and Entertainment:

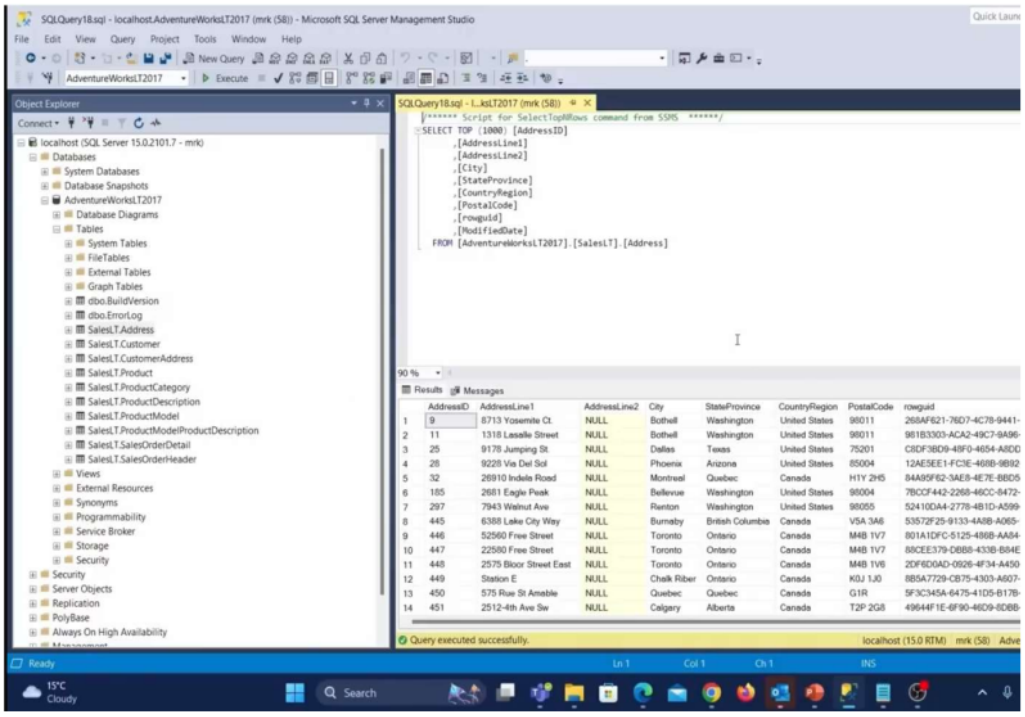
- **Video and Audio Content Migration:** Media and entertainment companies often have large repositories of video and audio content stored on-premises. This solution can enable the migration of these data to cloud platforms, allowing for scalable storage, content distribution, and advanced analytics for audience engagement and content recommendation.

9. Detailed Steps for Data Migration (Cloud Infra, Coding etc.)

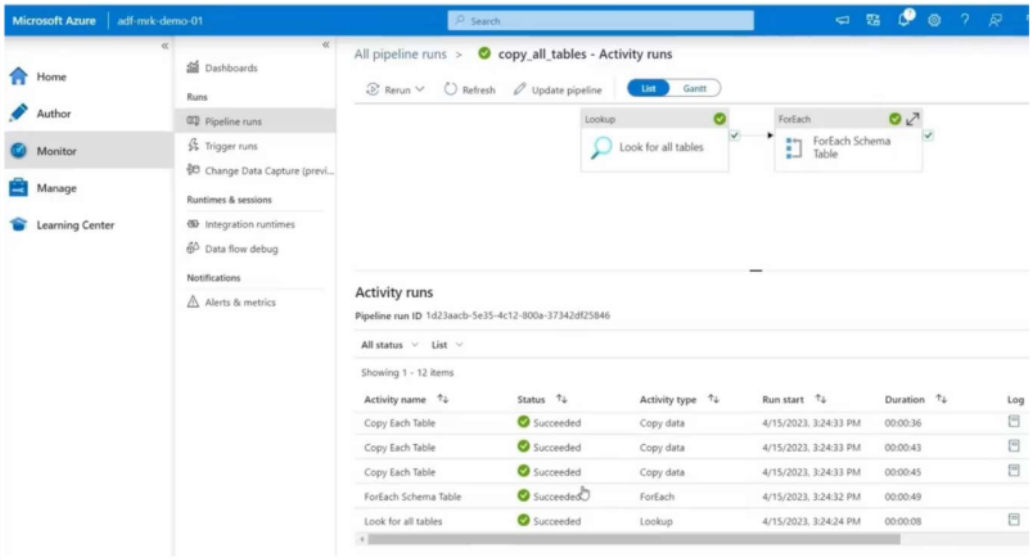
Cloud Infra (Azure Resource Group)



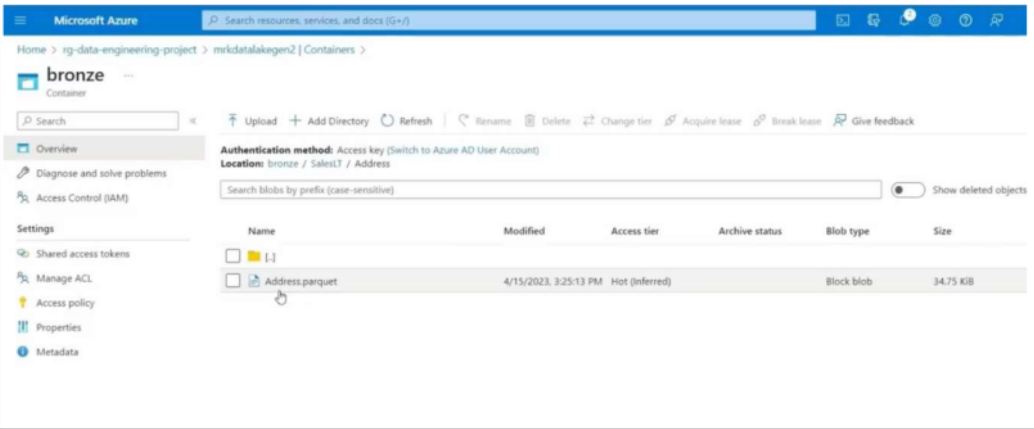
On Premises SQL Database



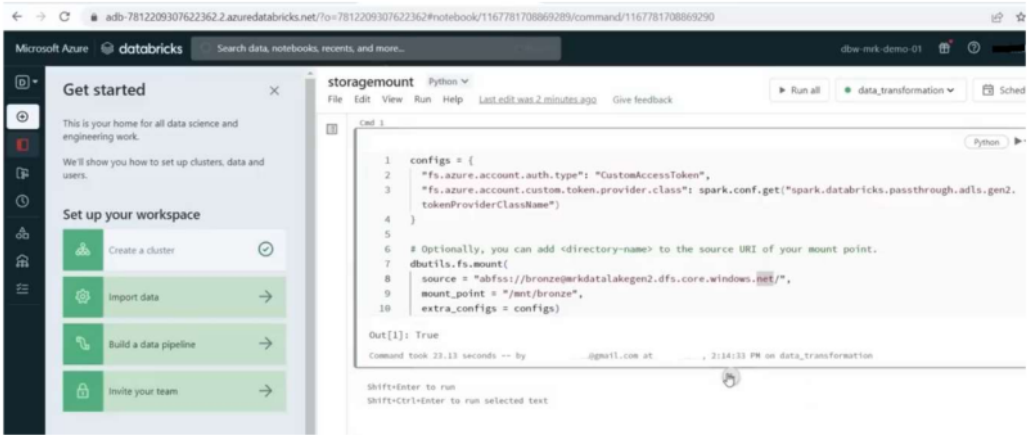
Azure Data Factory(ADF) Pipeline

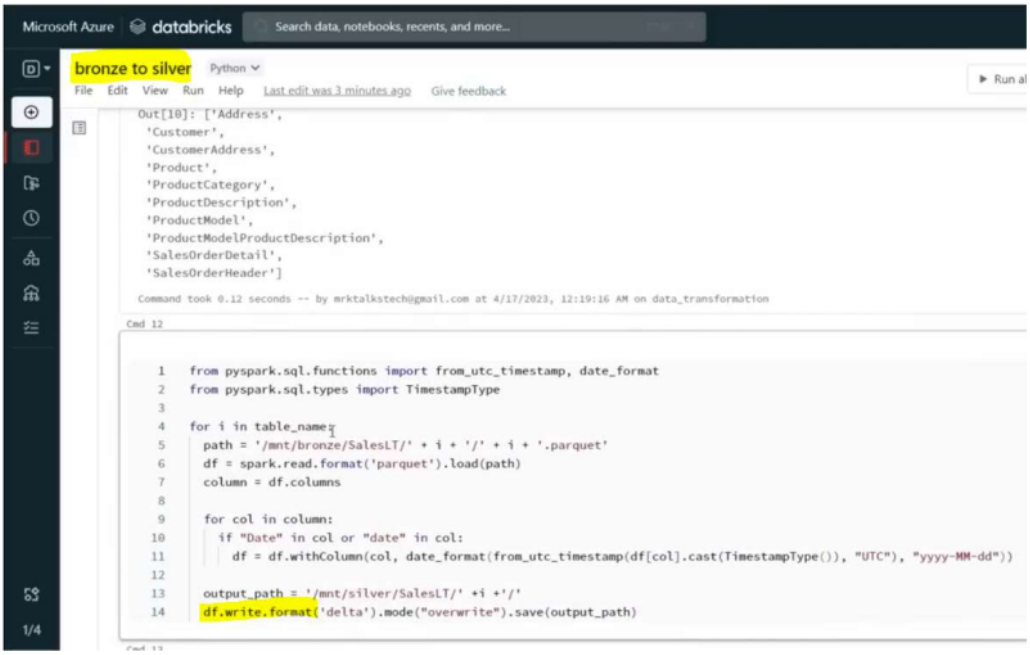


Broze Layer Parquet Files (Data Ingestion Step)

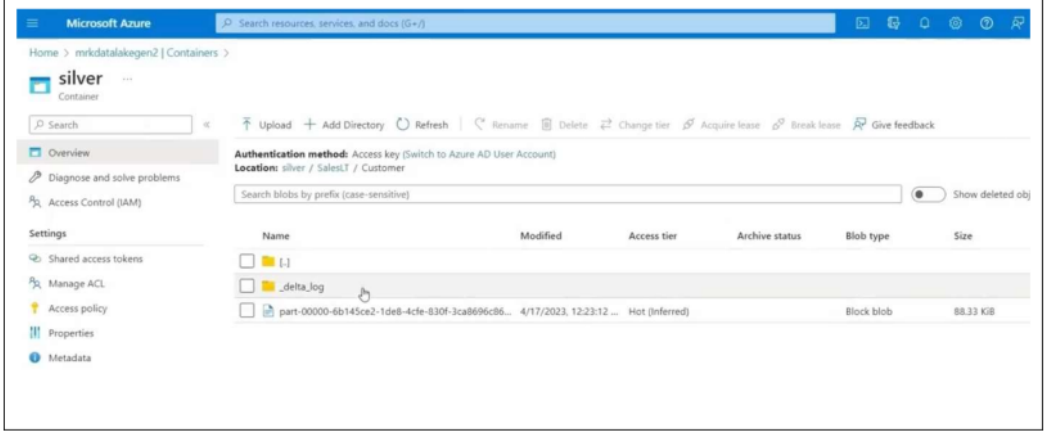


Broze to Silver (Data Transformation Step)





Silver Layer Parquet Files (Post Data Transformation Step)



10. Plan of Work

Phases	Start Date-End Date	Work to be done	Status
Dissertation Outline	13 January 2024 – 20 January 2024	Literature Review and prepare the Dissertation Outline	COMPLETED
Design & Development	21 January 2024 – 15 February 2024	Design and the Development Activity	COMPLETED
Testing	16 February 2024 – 21 March 2024	Software Testing,User Evaluation & Conclusion	COMPLETED
Mid Semester Report	22 March 2024 – 29 March 2024	Prepare and submit mid semester report	COMPLETED
Final phase of Development And Testing	30 March 2024 – 11 April 2024	Development and Testing Activity	PENDING
Dissertation Review	11 April 2024 – 22 April 2024	Fine tune the software system design. Submit the Dissertation to Supervisor & Additional Examiner for review and feedback	PENDING
Submission	23 April 2024 – 30 April 2024	Final Review and the submission of Dissertation	PENDING

11. Literature References:

To explore the latest research and new development going on this field is necessary to work on research and implementation project. In this project literature review is more inclined towards design of data migration software application which includes cloud-based technologies and frameworks. The following references considered for literature review.

1. Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K. & Zaharia, M. (2015, May). Spark sql: Relational data processing in spark. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 1383-1394). ACM.

2. Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). IEEE.

3. Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A. & Stoica, I. (2016, June). Apache spark: a unified engine for big data processing. In Communications of the ACM (Vol. 59, No. 11, pp. 56-65). ACM.

4. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

5. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A. & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

6. Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. Knowledge and Information Systems, 46(2), 241-284.

7. Anjani Kumar, Abhishek Mishra, Sanjeev Kumar. "Architecting a Modern Data Warehouse for Large Enterprises", Springer Science and Business Media LLC, 2024

8. Gani, A., Hameed, A., & Siddiqa, A. (2015, December). A survey on indexing techniques for big data: NoSQL databases. In 2015 Fifth International Conference on Advanced Computing & Communication Technologies (pp. 293-297). IEEE.

9. Agrawal, D., Das, S., & El Abbadi, A. (2011, March). Big data and cloud computing: current state and future opportunities. In Proceedings of the 14th International Conference on Extending Database Technology (pp. 530-533). ACM.

10. Mavridis, I., & Karatza, H. (2017). Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. Journal of Systems and Software, 125, 133-151.

12. Abbreviations:

ADF	Azure Data Factory
ADB	Azure Databricks
EHR	Electronic Health Records
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
IoT	Internet of Things
SQL	Structured Query Language
ETL	Extract, Transform, Load
ELT	Extract, Load, Transform
ML	Machine Learning
AWS	Amazon Web Services
GCP	Google Cloud Platform
MPP	Massively Parallel Processing
BI	Business Intelligence
AAD	Azure Active Directory
AKV	Azure Key Vault
RBAC	Role-Based Access Control
GDPR	General Data Protection Regulation
NIST	National Institute of Standards and Technology
NoSQL	Not only SQL

2022mt93331

ORIGINALITY REPORT

20%	11%	14%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Anjani Kumar, Abhishek Mishra, Sanjeev Kumar. "Architecting a Modern Data Warehouse for Large Enterprises", Springer Science and Business Media LLC, 2024 Publication	3%
2	Submitted to Birla Institute of Technology and Science Pilani Student Paper	2%
3	ijcttjournal.org Internet Source	1%
4	Puthiyavan Udayakumar. "Design and Deploy a Secure Azure Environment", Springer Science and Business Media LLC, 2023 Publication	1%
5	www.coursehero.com Internet Source	1%
6	fastercapital.com Internet Source	1%
7	Submitted to Johns Hopkins Unversity Student Paper	<1%

8	Submitted to Sheffield Hallam University Student Paper	<1 %
9	www.c-sharpcorner.com Internet Source	<1 %
10	datascijedi.blogspot.com Internet Source	<1 %
11	Submitted to HCUC Student Paper	<1 %
12	www.gobiit.com Internet Source	<1 %
13	www.microsoftpressstore.com Internet Source	<1 %
14	labs.sogeti.com Internet Source	<1 %
15	Submitted to Institute Of Business Management & Research, IPS Student Paper	<1 %
16	"High Performance Computing in Biomimetics", Springer Science and Business Media LLC, 2024 Publication	<1 %
17	docs.microsoft.com Internet Source	<1 %
18	Submitted to The Robert Gordon University Student Paper	<1 %

19	Submitted to Brigham Young University Student Paper	<1%
20	Submitted to Washington University of Science and Technology Student Paper	<1%
21	Submitted to BPP College of Professional Studies Limited Student Paper	<1%
22	Submitted to Regent Independent School and Sixth Form College Student Paper	<1%
23	apessay.elementfx.com Internet Source	<1%
24	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1%
25	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1%
26	Submitted to Northcentral Student Paper	<1%
27	hevodata.com Internet Source	<1%
28	soniacomp.medium.com Internet Source	<1%

29	Submitted to Algonquin College Student Paper	<1 %
30	Submitted to American InterContinental University Student Paper	<1 %
31	Submitted to Glasgow Caledonian University Student Paper	<1 %
32	Submitted to Heriot-Watt University Student Paper	<1 %
33	Submitted to Kaplan Professional Student Paper	<1 %
34	medium.com Internet Source	<1 %
35	Robert Stackowiak. "Azure Internet of Things Revealed", Springer Science and Business Media LLC, 2019 Publication	<1 %
36	Roberto Barriga Rodríguez. "Optimization of Fluid Bed Dryer Energy Consumption for Pharmaceutical Drug Processes through Machine Learning and Cloud Computing Technologies", Universitat Politecnica de Valencia, 2023 Publication	<1 %
37	www.mdpi.com Internet Source	<1 %

38	ien.ro Internet Source	<1%
39	pdfcoffee.com Internet Source	<1%
40	Babeș-Bolyai University Publication	<1%
41	Mavridis, Ilias, and Helen Karatza. "Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark", Journal of Systems and Software, 2017. Publication	<1%
42	Submitted to Northampton College Student Paper	<1%
43	businessyield.com Internet Source	<1%
44	itvmo.gsa.gov Internet Source	<1%
45	pdfs.semanticscholar.org Internet Source	<1%
46	www-development.newrelic.com Internet Source	<1%
47	www.tandfonline.com Internet Source	<1%

48	Adam Aspin. "Pro Power BI Desktop", Springer Science and Business Media LLC, 2020 Publication	<1%
49	Matt How. "The Modern Data Warehouse in Azure", Springer Science and Business Media LLC, 2020 Publication	<1%
50	Sudhir Rawat, Abhishek Narain. "Understanding Azure Data Factory", Springer Science and Business Media LLC, 2019 Publication	<1%
51	Bhadresh Shiyal. "Beginning Azure Synapse Analytics", Springer Science and Business Media LLC, 2021 Publication	<1%

Exclude quotes	On	Exclude assignment template	On
Exclude bibliography	On	Exclude matches	< 5 words