A REPORT

ON

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

BY

Student Name: SHELKE AKSHAY NANDKUMAR

BITS ID: 2022MT93331

AT

Schlumberger India Technology Centre Ltd., Pune

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
VIDYA VIHAR, PILANI, RAJASTHAN - 333031.

April 2024

A REPORT

ON

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

BY

Student Name: SHELKE AKSHAY NANDKUMAR
BITS ID: 2022MT93331
Course No. / Course Title: SEZG628T / Dissertation
Discipline: M. Tech. Software Engineering
Research Area: Cloud Computing

Prepared in partial fulfilment of the WILP Dissertation/Project/Project Work Course

AT

Schlumberger India Technology Centre Ltd., Pune

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
VIDYA VIHAR, PILANI, RAJASTHAN - 333031.
April 2024

Acknowledgement

While working on this project, I came across many people, without their guidance and support, it was very difficult to get the correct direction in a very short span of time. It is a pleasure to convey my gratitude to all of them.

First, I would like to thank BITS, PILANI for providing me the opportunity of taking part in M. Tech. Software Engineering program. I am highly grateful to Prof. Pramod Bide from BITS faculty for his proper feedback and guidance which greatly helped me to do this project.

Further, I would like to do Big thanks to my mentor Mr. Shubham Tulsyan (Senior Project Manager) & Mr. Prashant Gorade (Senior Software Engineer) to their invaluable encouragement, suggestions, and support from an early stage of this work.

Also, I am also thankful to Schlumberger India Technology Center Ltd., Pune for giving me this opportunity and making available all the resources required for this dissertation. I am also thankful to BITS, Pilani and the Training department for all their efforts in organizing this course.

Lastly, without my family and my Parent's constant support, I could not have made this journey, they motivated me to keep focused and encouraged me to fulfill the requirement of the course.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
(RAJASTHAN)
WILP Division

Organization: Schlumberger India Technology Centre Ltd., Pune
Location: Pune, Maharashtra, India
Duration: 6 months (Dissertation)
Date of Start: 13th January 2024
Date of Submission: 29th April 2024
Title of the Project: Cloud-First Approach: Engineering a Solution for Efficient On-Premises Data Migration to Cloud Platforms

ID No./Name of the student: 2022MT93331 / Shelke Akshay Nandkumar

Name (s) and Designation (s) of your Supervisor and Additional Examiner:
1) Mr. Shubham Tulsyan (Senior Project Manager)
2) Mr. Prashant Gorade (Senior Software Engineer)

Name of the Faculty mentor: Mr. Pramod Bide

Key Words: Cloud Computing, Azure, Data Engineering, Data Transformation, Data Reporting

Project Areas: The project area is focused on building a cloud based End-to-End Data Engineering system on Microsoft Azure Cloud Platform, covering phases of the data lifecycle such as ingestion, transformation, loading, and reporting using various Azure services and tools.

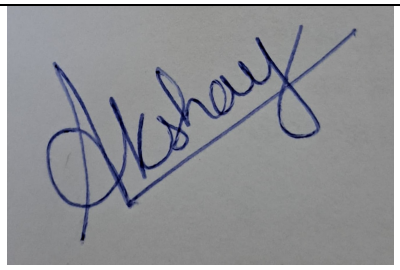
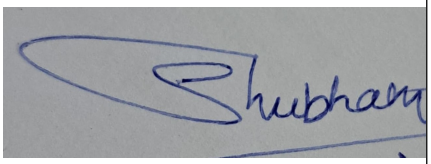
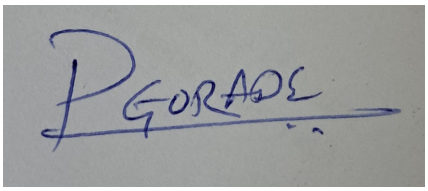
Abstract

In response to the escalating demands for comprehensive data management solutions based on cloud computing, this project aims to construct an integrated End-to-End Azure Data Engineering system. The project covers Data Ingestion, Transformation, Loading, and Reporting, utilizing various Azure services and tools.

Azure Data Factory orchestrates workflows for data movement, while Azure Data Lake Storage (Gen2) provides a secure and scalable data repository. Azure Databricks enables robust data transformation, and the transformed data is loaded into Azure Synapse Analytics for reporting and analysis.

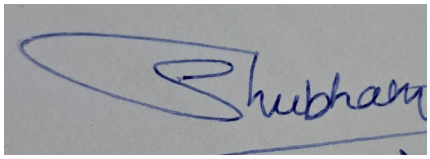
Azure Key Vault and Azure Active Directory ensure secure key management and identity & access management, respectively, contributing to a secure and compliant data environment.

The use case involves ingesting tables from an on-premise SQL Server database using Azure Data Factory, storing data in Azure Data Lake, transforming raw data using Azure Databricks, and loading cleaned data into Azure Synapse Analytics. Microsoft Power BI integrates with Azure Synapse Analytics for interactive dashboard creation.

		
Signature of Student	Signature of Supervisor	Signature of Additional Examiner
Name: SHELKE AKSHAY NANDKUMAR	Name: Mr. Shubham Tulsyan	Name: Mr. Prashant Gorade
Date : 28th April 2024	Date : 28th April 2024	Date : 28th April 2024
Place : Pune, Maharashtra, India	Place : Pune, Maharashtra, India	Place : Pune, Maharashtra, India

CERTIFICATE

This is to certify that the dissertation entitled Cloud-First Approach: Engineering a Solution for Efficient On-Premises Data Migration to Cloud Platforms by SHELKE AKSHAY NANDKUMAR, with ID No: 2022MT93331, for the partial fulfillment of the requirements of the MTech Software Engineering degree of BITS, embodies the bona fide work done by him under my supervision.



_____

Signature of the Supervisor

Mr. Shubham Tulsyan
Senior Project Manager,
Schlumberger India Technology Centre, Pune

Place: Pune

Date: 28th April. 2024

Contents

1. Broad Area of Work 5
2. Background Research and Literature Review 5
3. Problem Definition 7
4. Project Objectives 8
5. Scope of Work 10
6. Tools & Technologies used for Data Migration 11
7. Overview of Architectural Design 14
8. Use Cases for Data Migration 21
9. Detailed Steps for Data Migration (Cloud Infra, Coding etc.) 23
10. Conclusions – Results on PowerBI dashboards 33
11. Future Scope and Limitations 37
12. Plan of Work 38
13. Literature References: 39
14. Abbreviations: 40
15. Glossary 41

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:

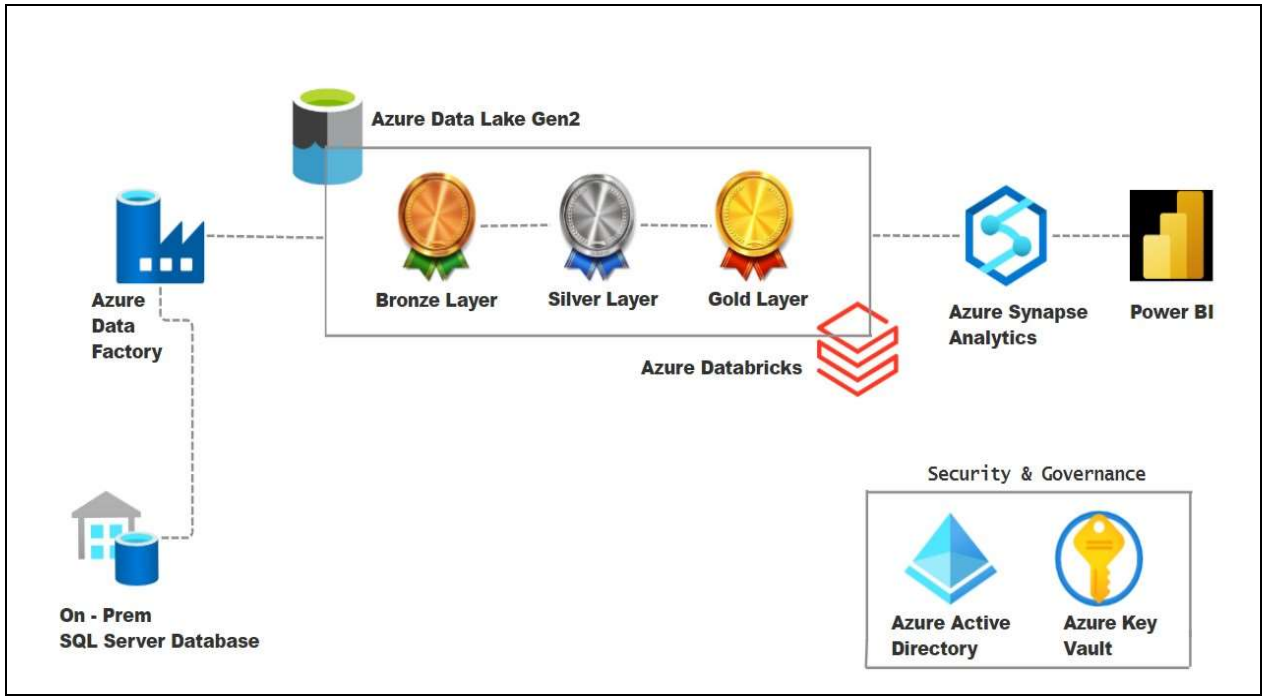
- **Data Connectivity:** Ability to connect to a various data sources, including Azure Data Lake Gen2,Azure Synapse Analytics,other cloud & on-premises data sources.
- **Data Modeling:** Advanced data modeling capabilities, including support for relationships, hierarchies, and calculated columns.
- **Interactive 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.

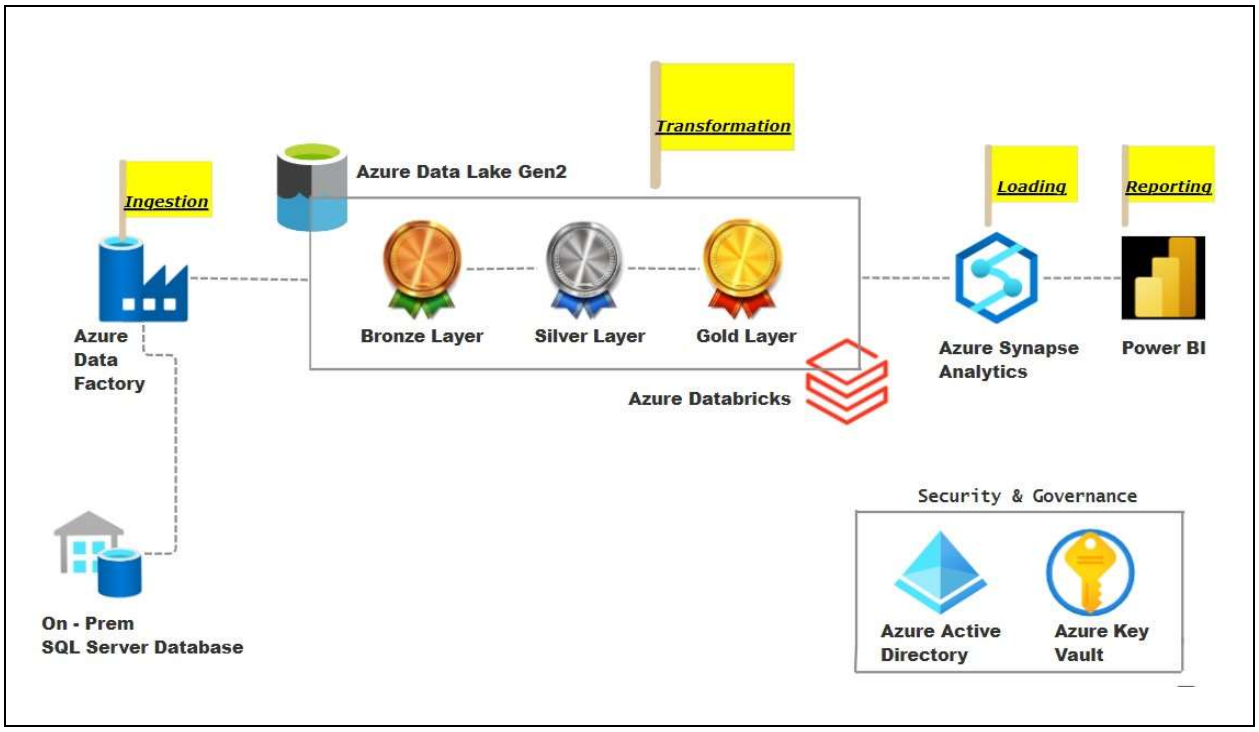
The transformation and processing tasks performed by Databricks can include:

- **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:

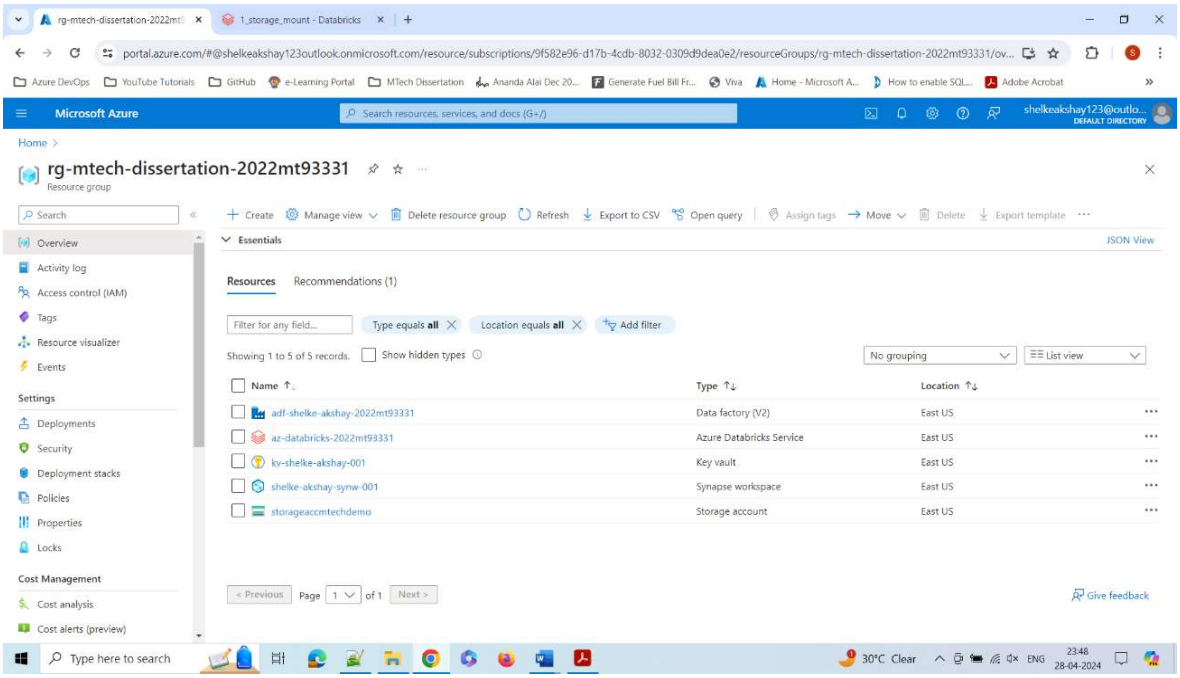
- **IoT and Sensor Data Migration:** Manufacturing companies and supply chain operations generate massive volumes of IoT and sensor data from various sources, such as machinery, equipment, and logistics systems. This solution can facilitate the migration of these data to cloud platforms, enabling real-time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

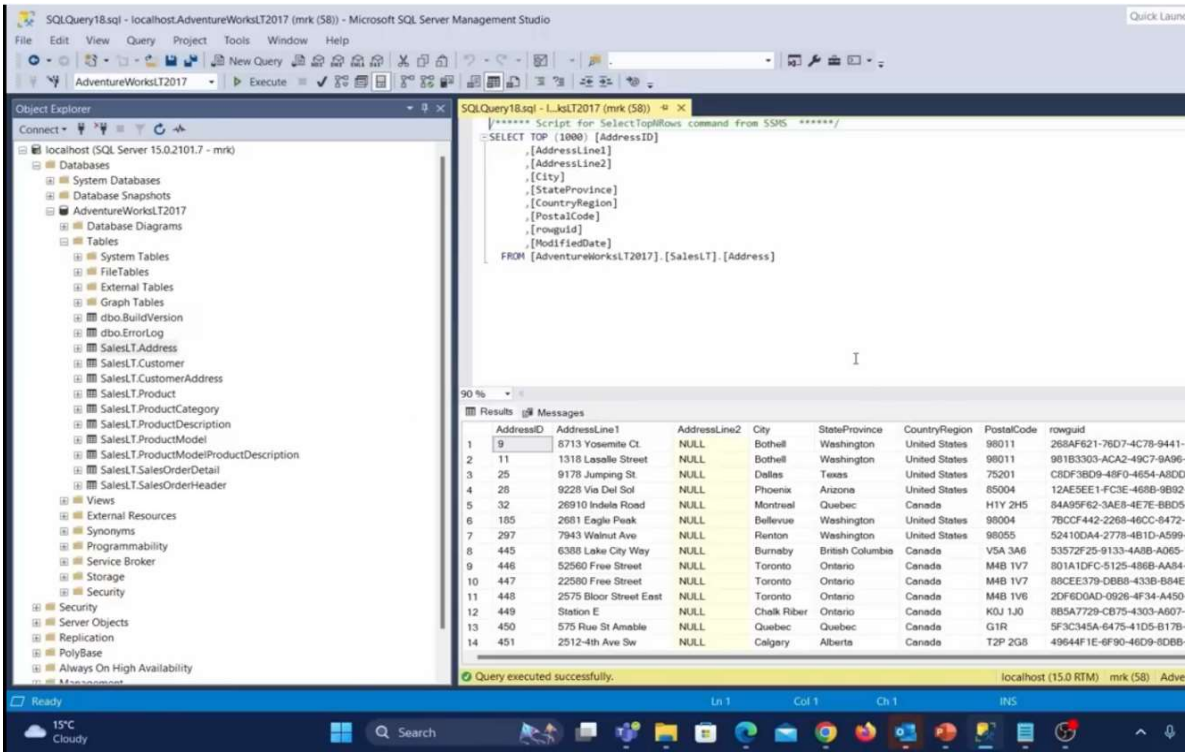
9. Detailed Steps for Data Migration (Cloud Infra, Coding etc.)

Cloud Infra (Azure Resource Group)

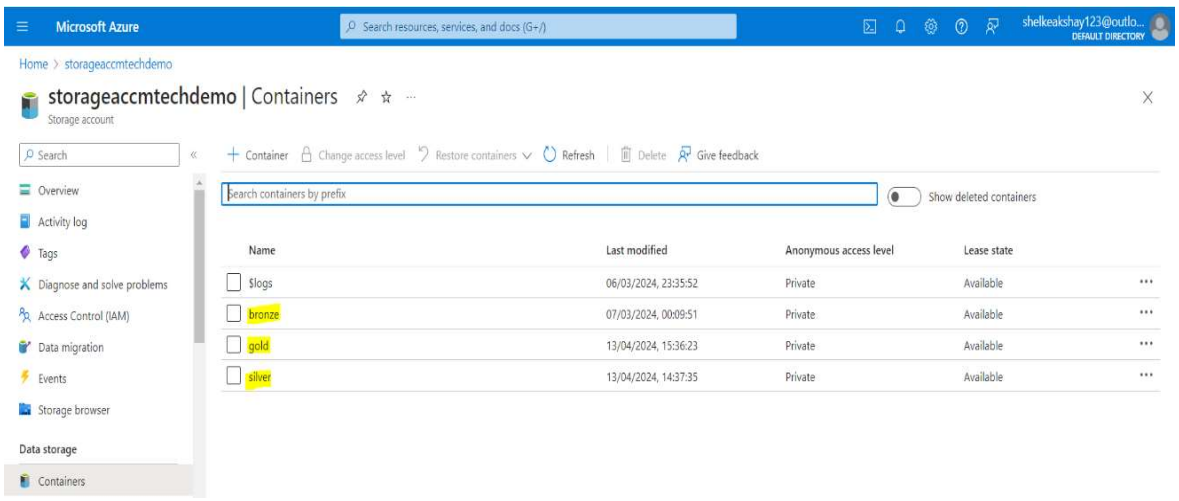


- Azure Data Lake Storage (Gen2) – Storage Account
    - To store data parquet files and delta tables
- Azure Key Vault
    - To store secrets like database credentials
- Azure Databricks
- Azure Data Factory
- Azure Synapse Analytics

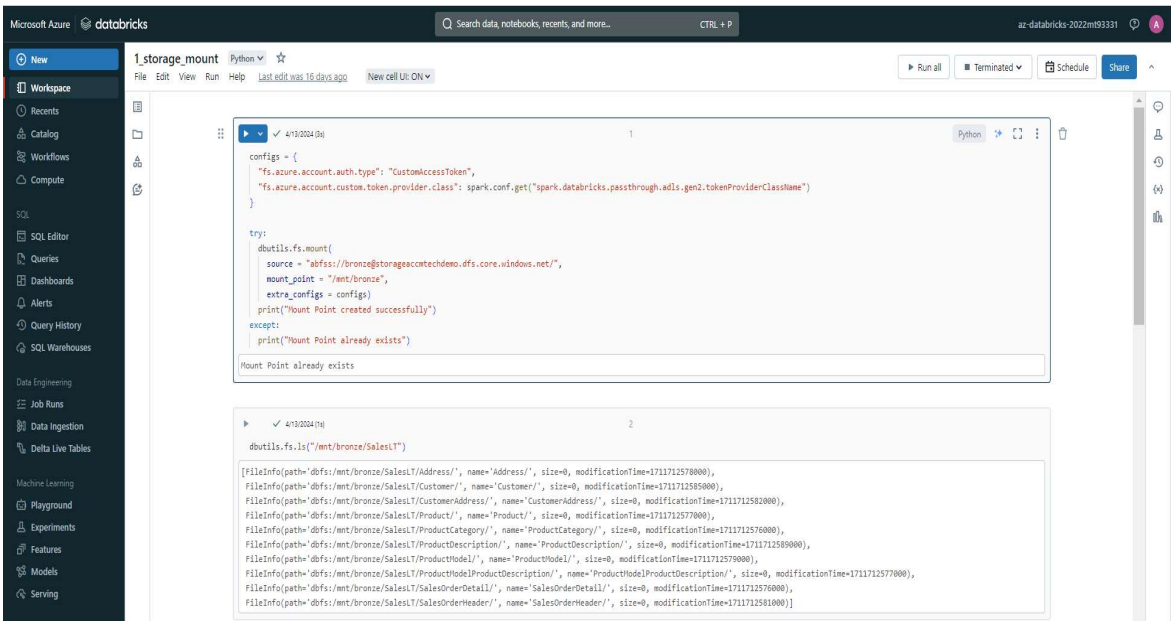
On Premises SQL Database

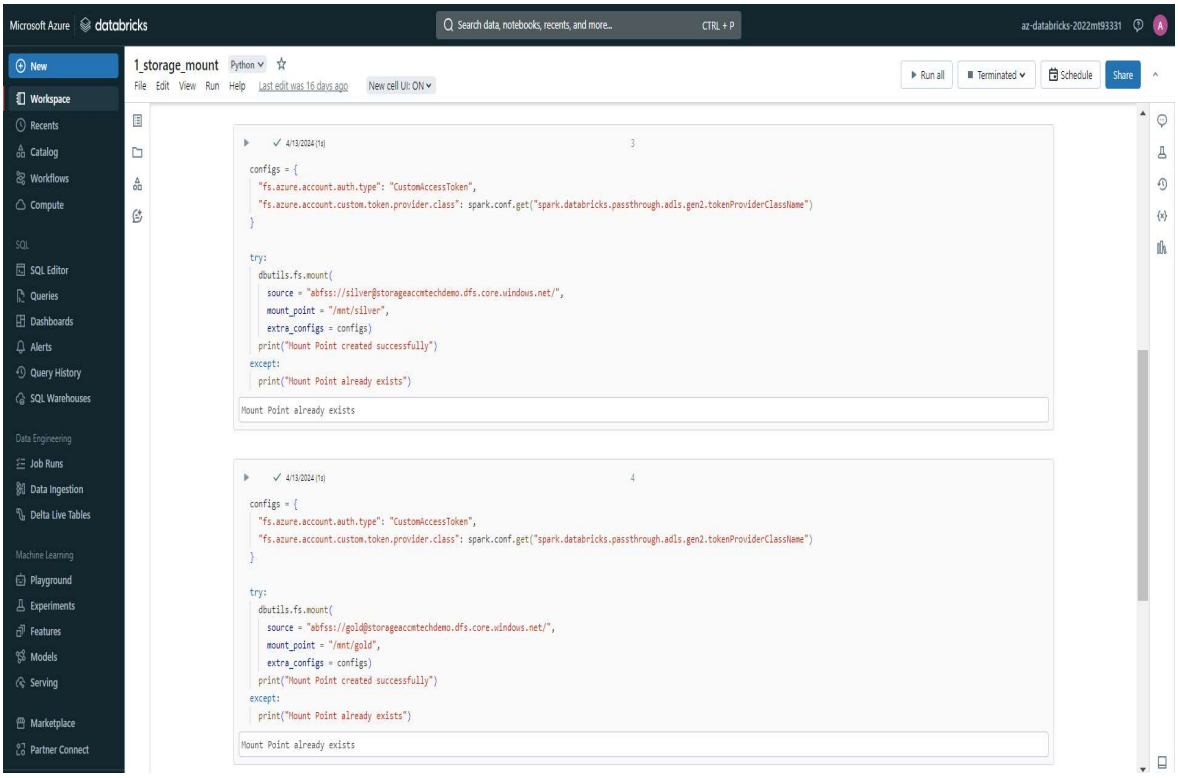


Azure Data Lake Storage (Gen2)

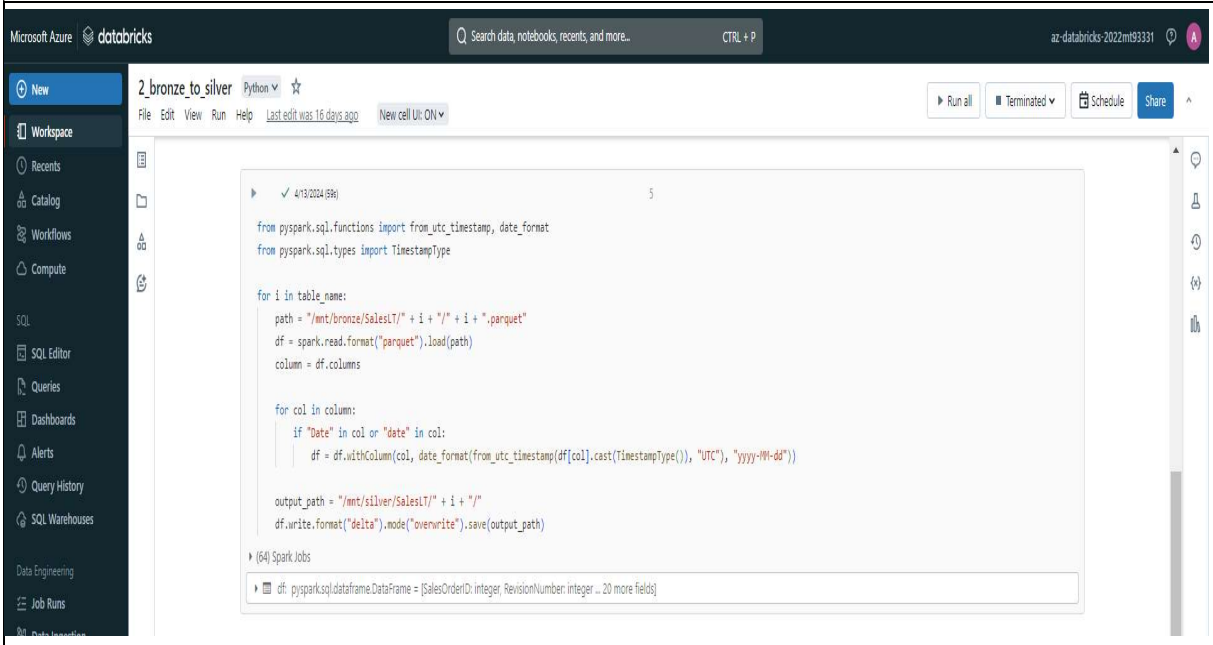


Storage Mount in Azure Databricks

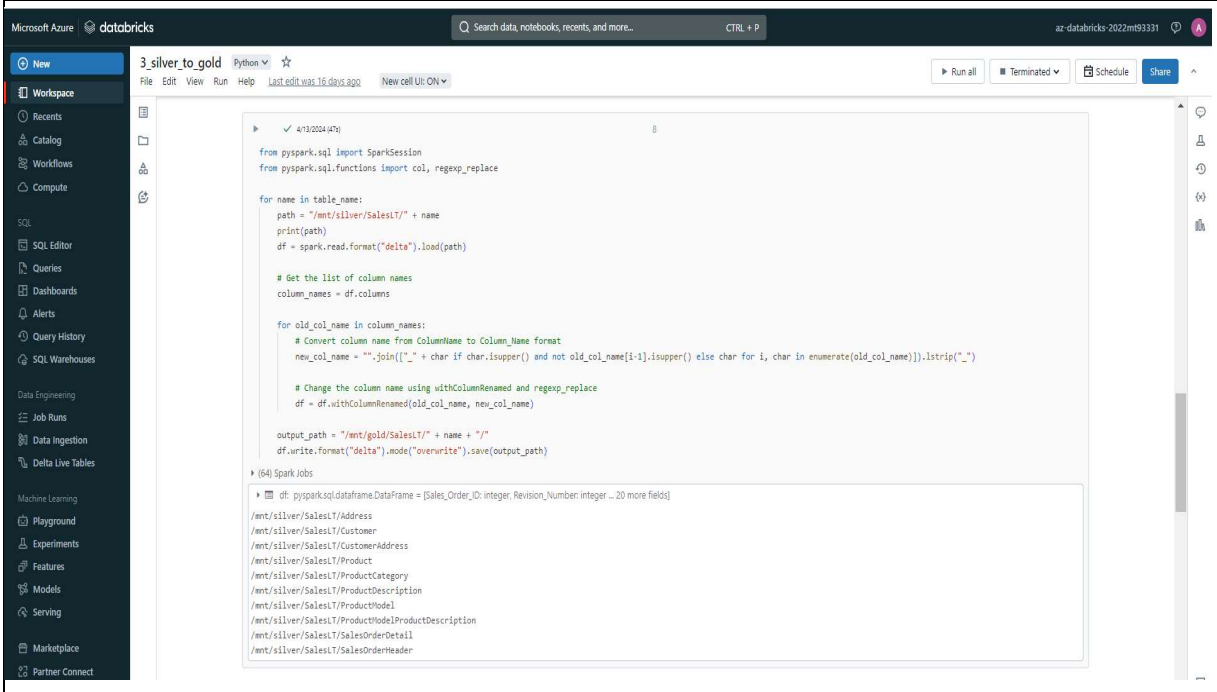




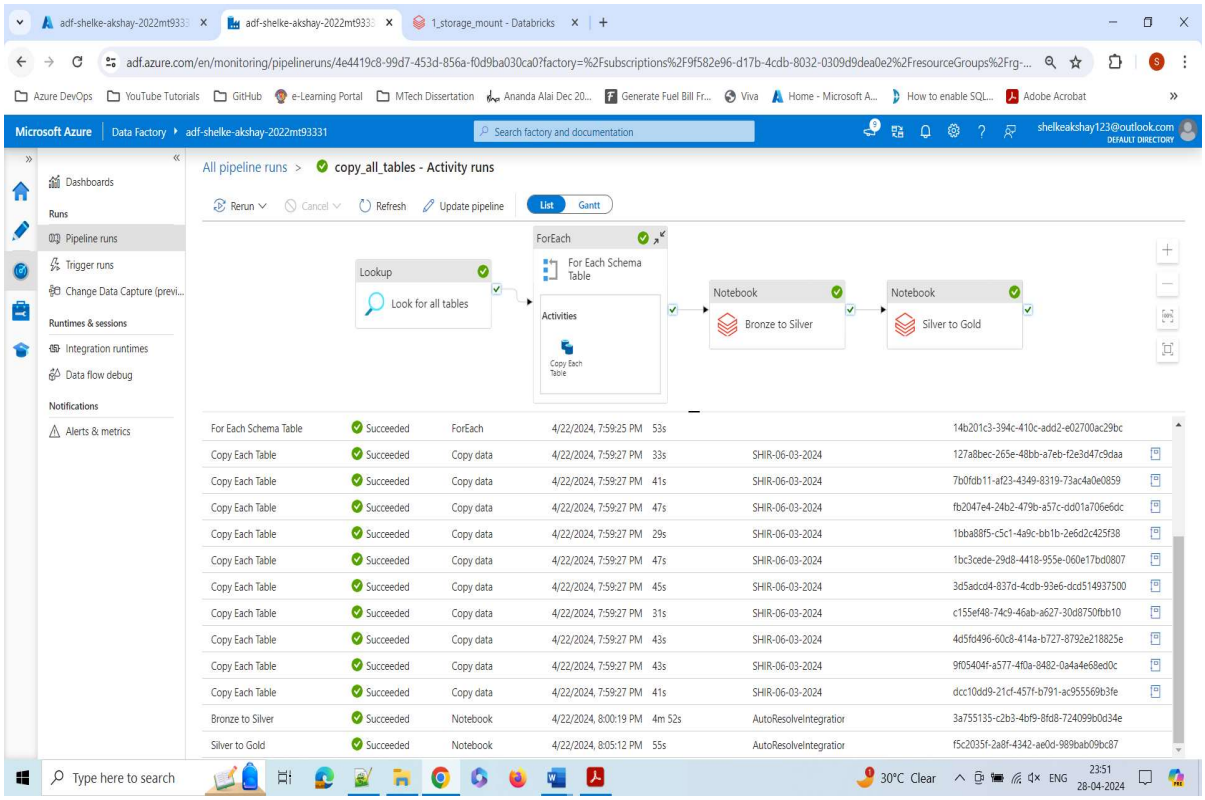
Code – Bronze to Silver Data Transformation



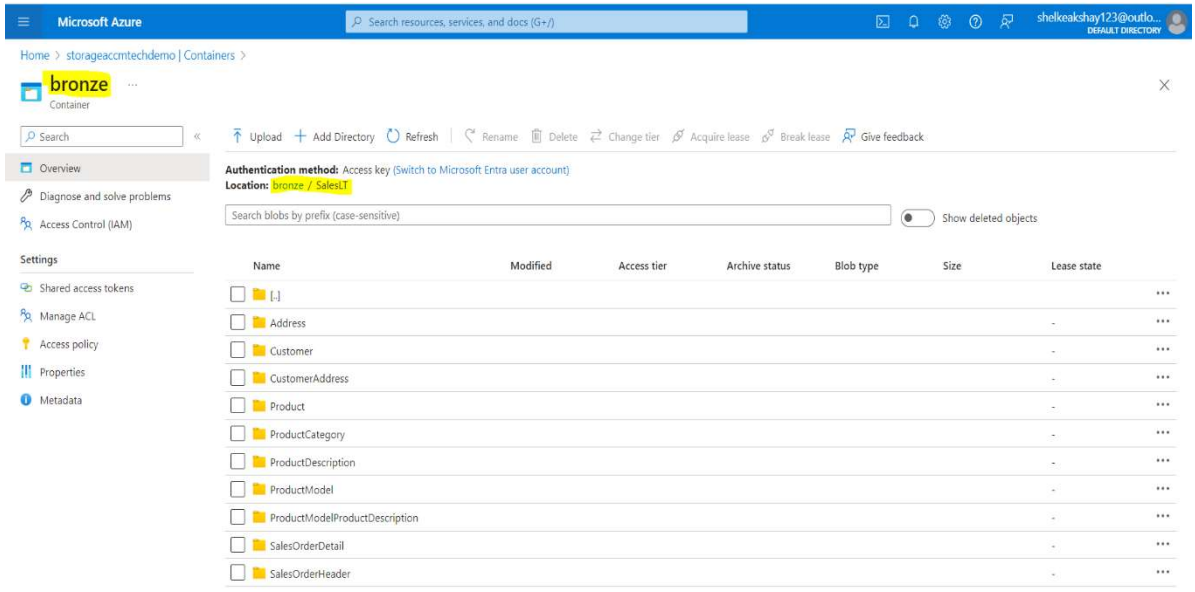
Code – Silver to Gold Data Transformation

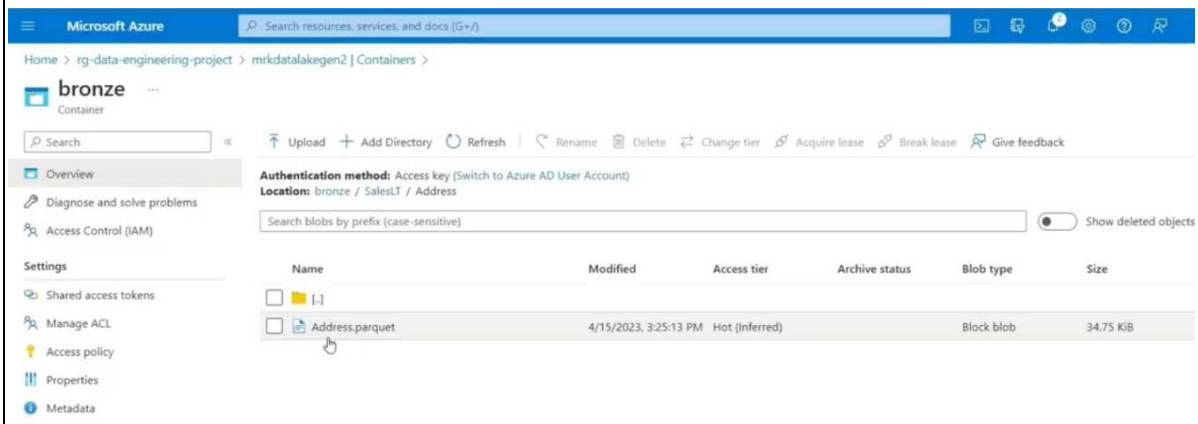


Azure Data Factory(ADF) Pipeline – Data Ingestion + Data Transformation

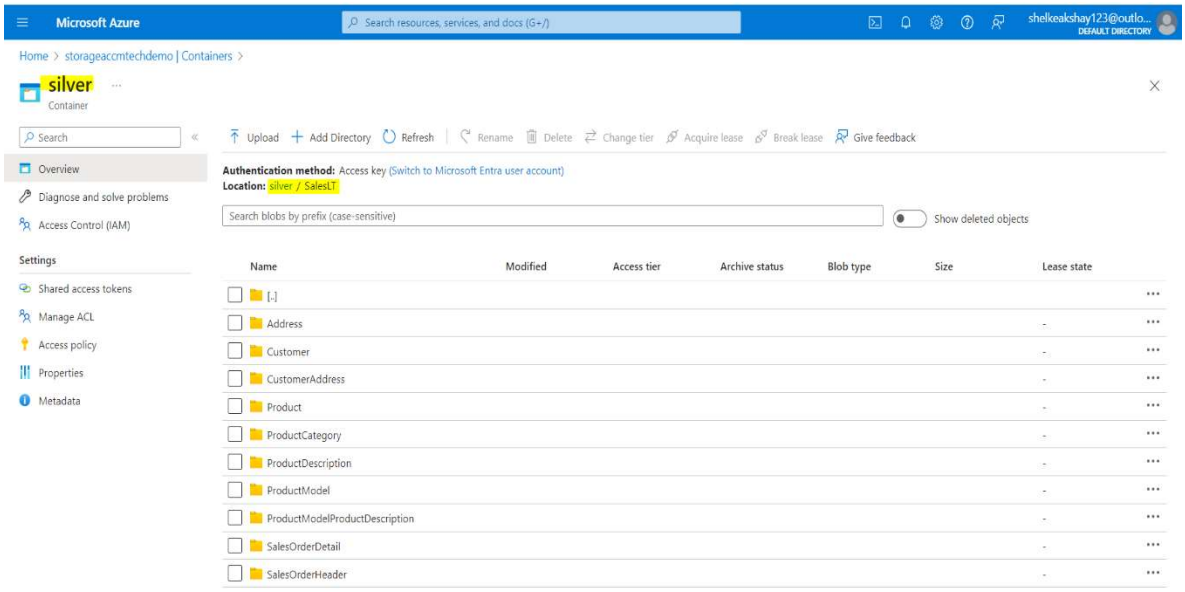


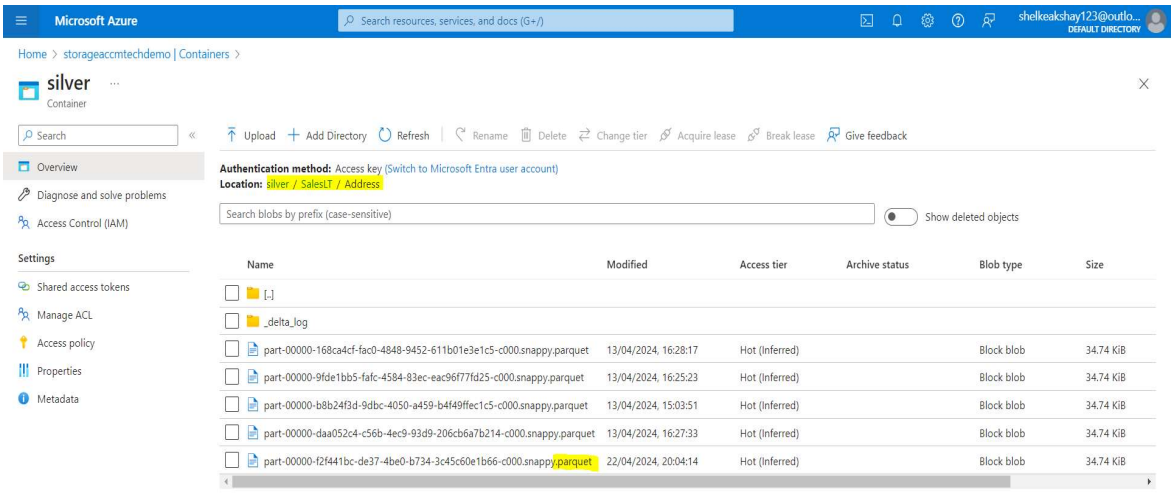
Bronze Layer Parquet Files (Data Ingestion Step)



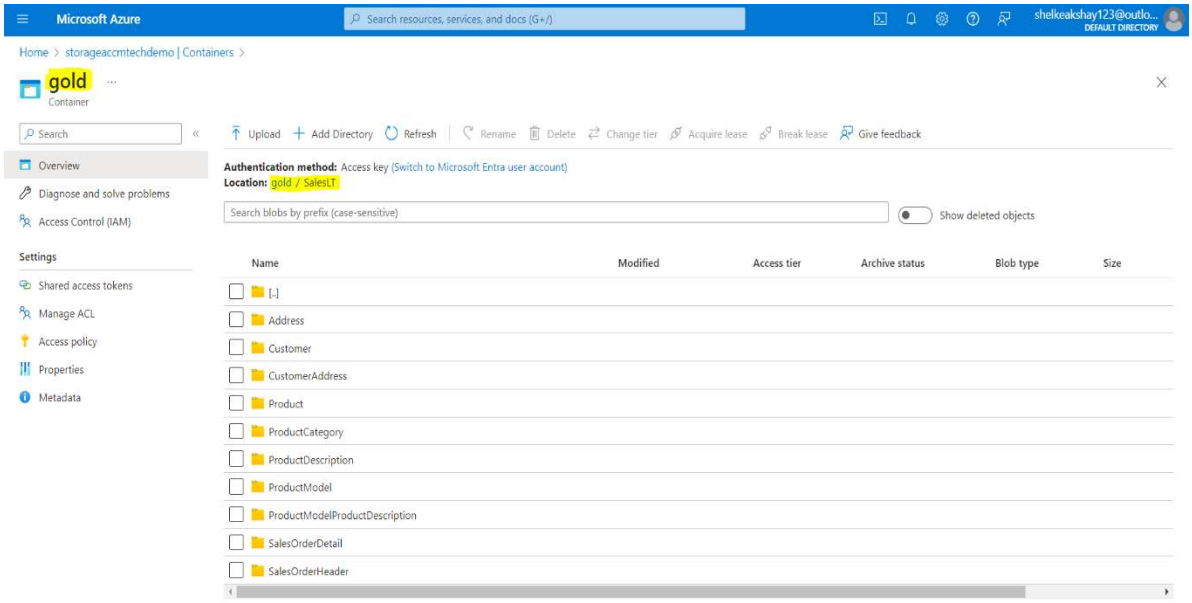


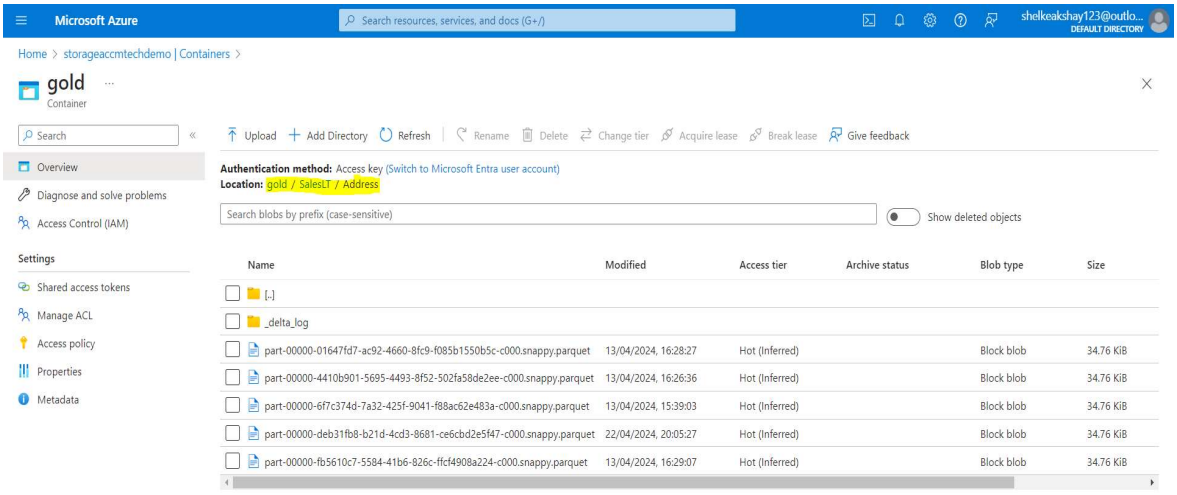
Silver Layer Parquet Files (Post Data Transformation Step)



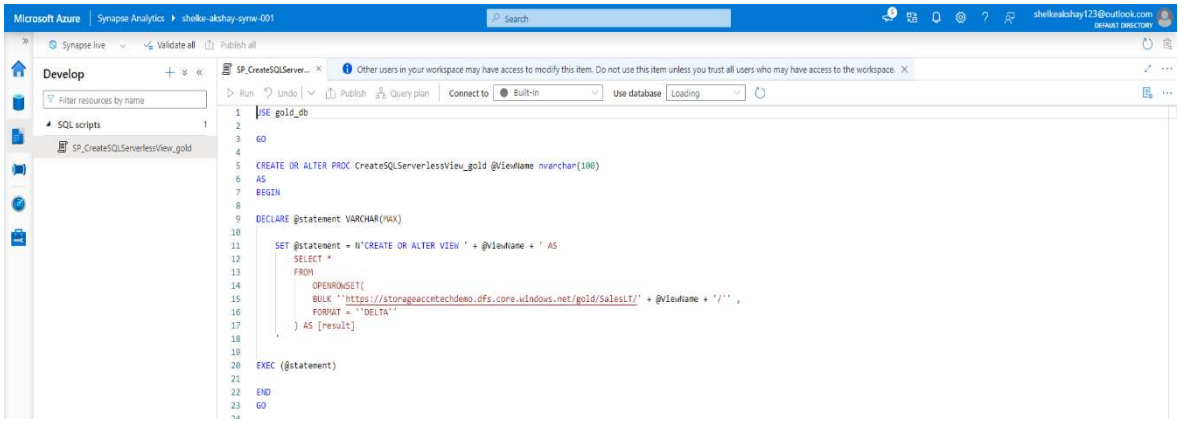


Gold Layer Parquet Files (Post Data Transformation Step)

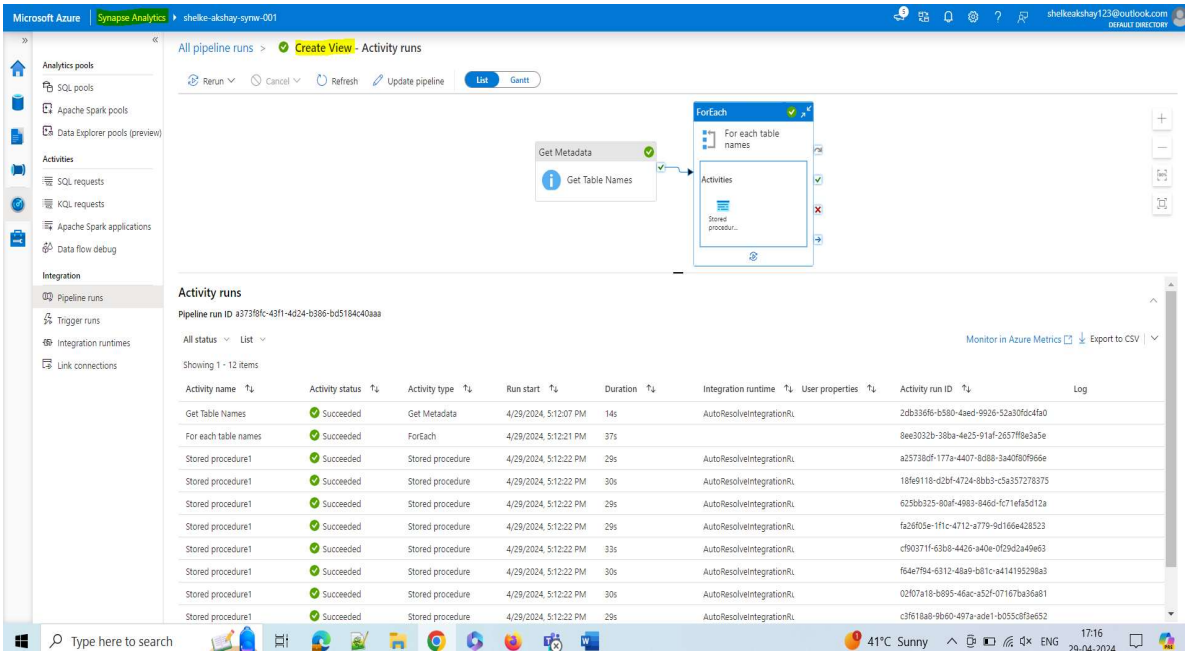


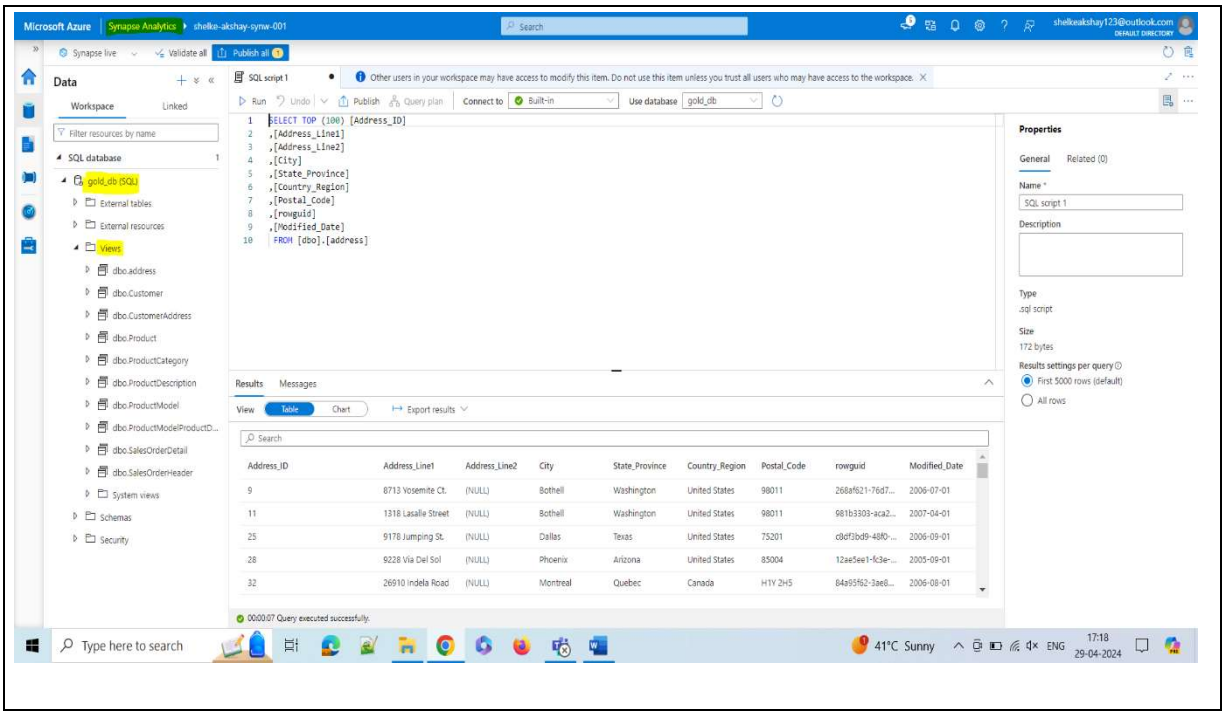


Code – Stored Procedure to create views on gold dataset



Azure Synapse Analytics – Data Loading




Microsoft Azure
Synapse Analytics
shelke-akshay-synw-001
Search
shelkeakshay123@outlook.com
DEFAULT DIRECTORY
Synapse live
Validate all
Publish all
Data
Workspace
Linked
Filter resources by name
SQL database
gold_db (SQL)
External tables
External resources
Views
dbo.address
dbo.Customer
dbo.CustomerAddress
dbo.Product
dbo.ProductCategory
dbo.ProductDescription
dbo.ProductModel
dbo.ProductModelProductD...
dbo.SalesOrderDetail
dbo.SalesOrderHeader
System views
Schemas
Security
SQL script 1
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
gold_db
SELECT TOP (100) [Address_ID]
,[Address_Line1]
,[Address_Line2]
,[City]
,[State_Province]
,[Country_Region]
,[Postal_Code]
,[rowguid]
,[Modified_Date]
FROM [dbo].[address]
Properties
General
Related (0)
Name *
SQL script 1
Description
Type
.sql script
Size
172 bytes
Results settings per query
First 5000 rows (default)
All rows
Results
Messages
View
Table
Chart
Export results
Search
Address_ID
Address_Line1
Address_Line2
City
State_Province
Country_Region
Postal_Code
rowguid
Modified_Date
9
8713 Yosemite Ct.
(NULL)
Bothell
Washington
United States
98011
268af621-76d7...
2006-07-01
11
1318 Lasalle Street
(NULL)
Bothell
Washington
United States
98011
981b3303-aca2...
2007-04-01
25
9178 Jumping St.
(NULL)
Dallas
Texas
United States
75201
c0df3bd9-48f0-...
2006-09-01
28
9228 Via Del Sol
(NULL)
Phoenix
Arizona
United States
85004
12ae5ee1-fc3e-...
2005-09-01
32
26910 Indela Road
(NULL)
Montreal
Quebec
Canada
H1Y 2H5
84a95f62-3ae8...
2006-08-01
00:00:07 Query executed successfully.
Type here to search
41°C Sunny
ENG
17:18
29-04-2024

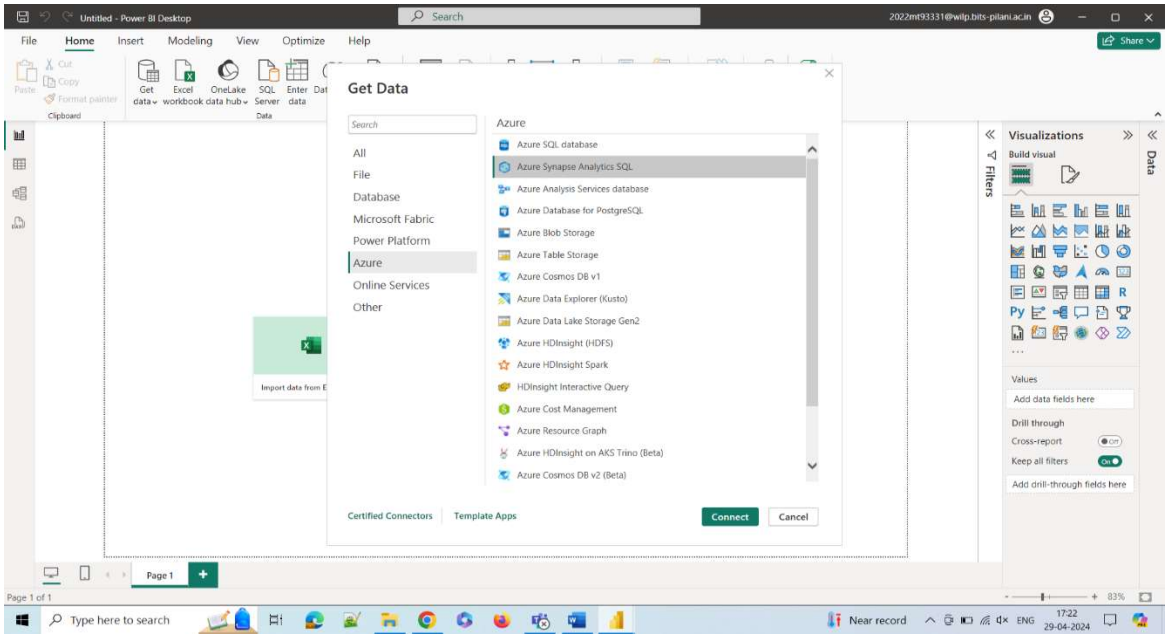
10. Conclusions – Results on PowerBI dashboards

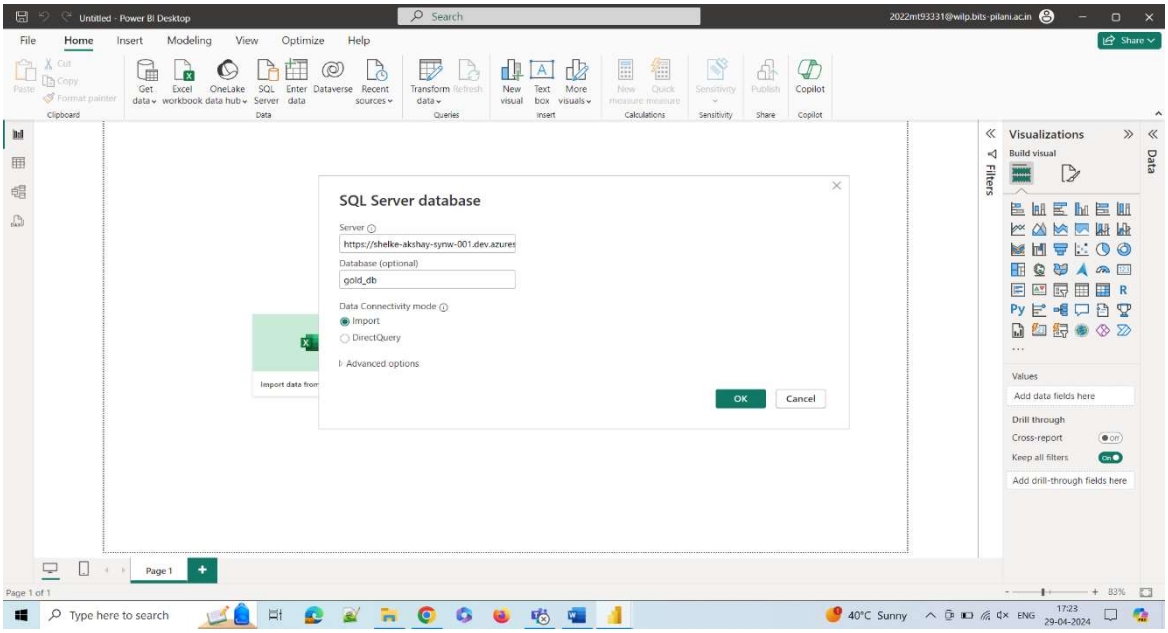
The project aimed to develop a comprehensive and efficient solution for on-premises data migration to cloud platforms using Microsoft Azure services. The key objectives were achieved, including:

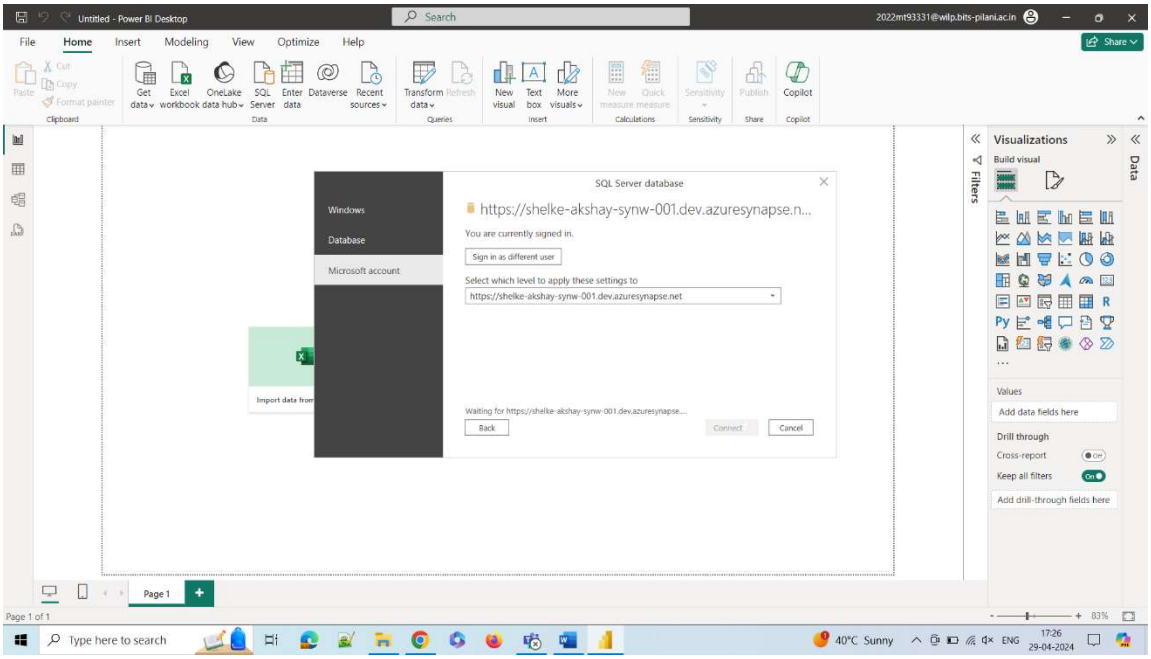
- Implementing a streamlined and efficient data workflow for ingestion, transformation, loading, and reporting using Azure Data Factory, Databricks, Data Lake Gen2, Synapse Analytics, and Power BI.
- Demonstrating seamless integration of various Azure services to construct a cohesive data platform.
- Implementing robust data governance and security measures using Azure Key Vault and Azure Active Directory.
- Creating interactive dashboards and reports in Power BI for actionable insights and informed decision-making.
- Applying the solution to a practical use case involving on-premise SQL Server data migration to the Azure cloud.

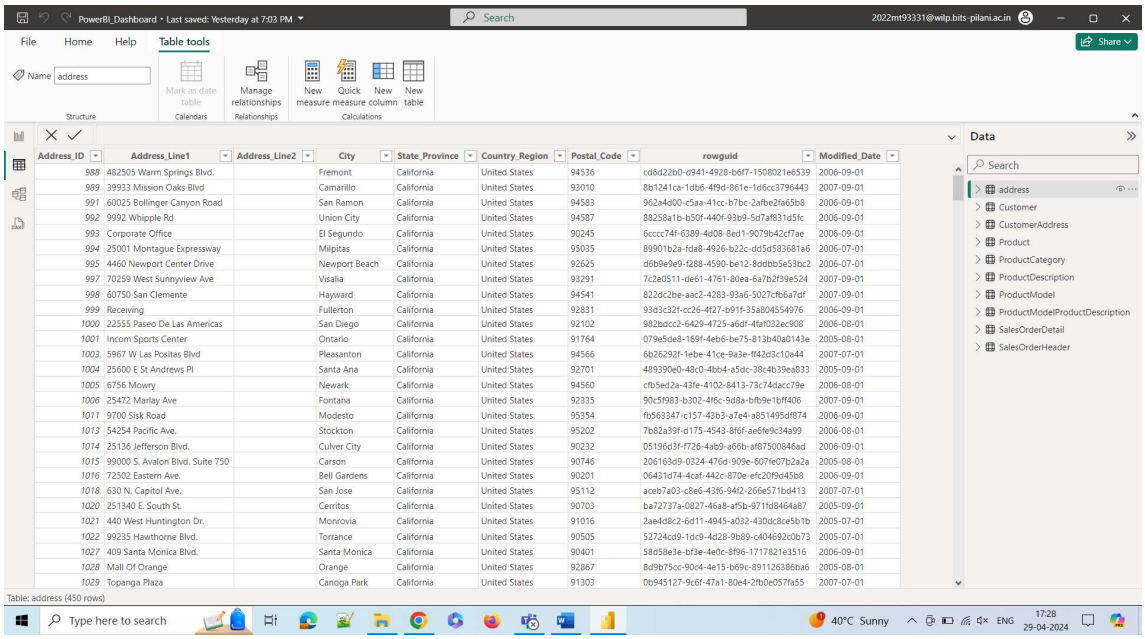
The project successfully delivered a secure, and efficient solution for enterprises aiming to harness the power of their data assets effectively in the cloud environment.

Connect to gold_db



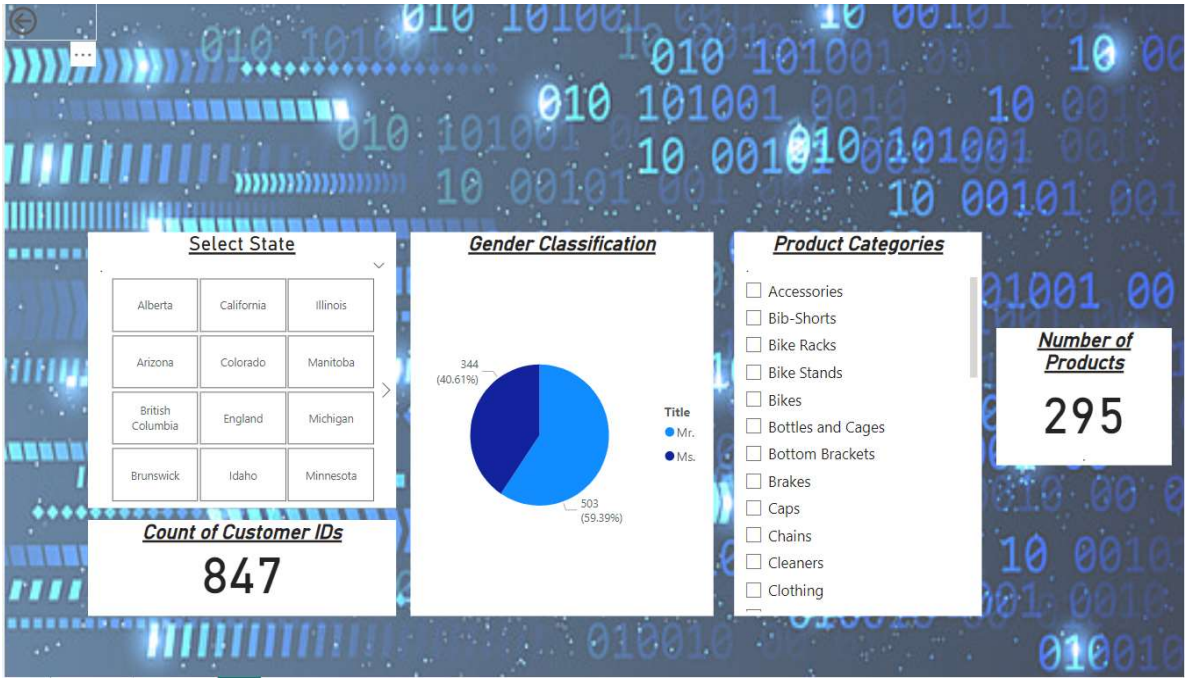




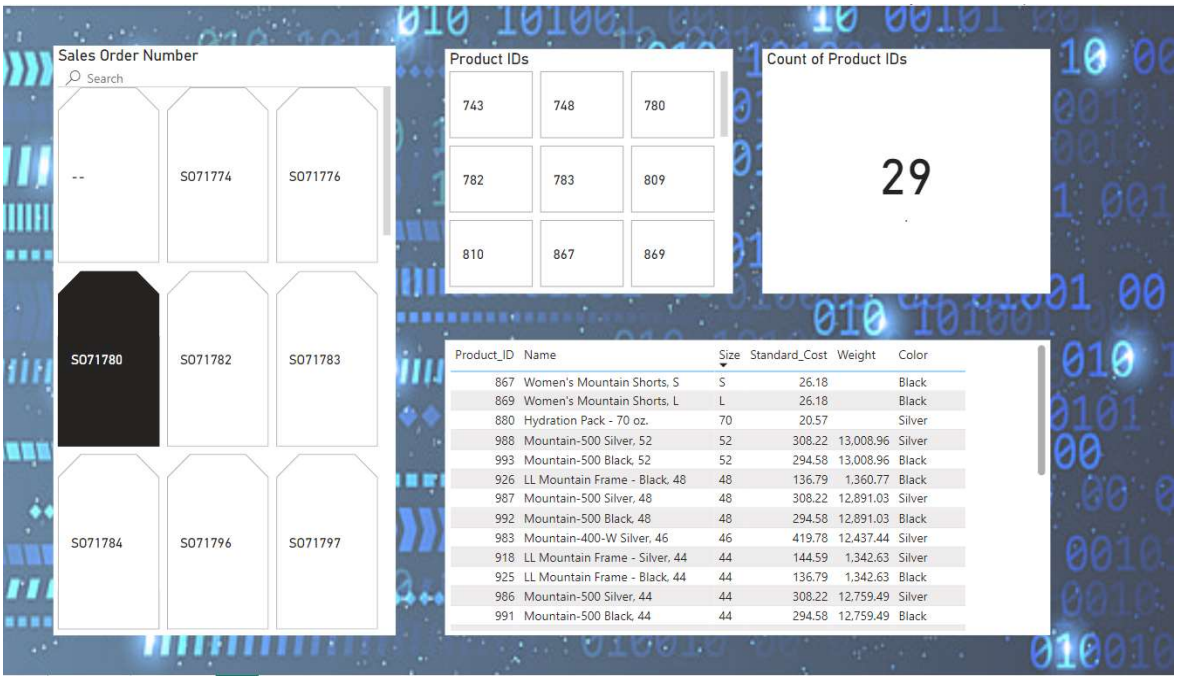


Address_ID	Address_Line1	Address_Line2	City	State_Province	Country_Region	Postal_Code	rowguid	Modified_Date
988	482505 Warm Springs Blvd.		Fremont	California	United States	94536	cd6d22b0-d941-4928-b6f7-1508021e6539	2006-09-01
989	39933 Mission Oaks Blvd		Camarillo	California	United States	93010	8b1241ca-1db6-4f9d-861e-1d6cc3796443	2007-09-01
991	60025 Bollinger Canyon Road		San Ramon	California	United States	94583	962a4d00-c5aa-41cc-b7bc-2afbe2fa65b8	2006-09-01
992	9992 Whipple Rd		Union City	California	United States	94587	88258a1b-b50f-440f-93b9-5d7af831d5fc	2006-09-01
993	Corporate Office		El Segundo	California	United States	90245	6cccc74f-6389-4d08-8ed1-9079b42cf7ae	2006-09-01
994	25001 Montague Expressway		Milpitas	California	United States	95035	89901b2a-fda8-4926-b22c-dd5d583681a6	2006-07-01
995	4460 Newport Center Drive		Newport Beach	California	United States	92625	d6b9e9e9-f288-4590-be12-8ddbb5e53bc2	2006-07-01
997	70259 West Sunnyview Ave		Visalia	California	United States	93291	7c2e0511-de61-4761-80ea-6a7b2f39e524	2007-09-01
998	60750 San Clemente		Hayward	California	United States	94541	822dc2be-aac2-4283-93a6-5027cfb6a7df	2007-09-01
999	Receiving		Fullerton	California	United States	92831	93d3c32f-cc26-4f27-b91f-35a804554976	2006-09-01
1000	22555 Paseo De Las Americas		San Diego	California	United States	92102	982bdcc2-6429-4725-a6df-4faf032ec908	2006-08-01
1001	Incom Sports Center		Ontario	California	United States	91764	079e5de8-169f-4eb6-be75-813b40a0143e	2005-08-01
1003	5967 W Las Positas Blvd		Pleasanton	California	United States	94566	6b26292f-1ebe-41ce-9a3e-ff42d3c10a44	2007-07-01
1004	25600 E St Andrews Pl		Santa Ana	California	United States	92701	489390e0-48c0-4bb4-a5dc-38c4b39ea833	2005-09-01
1005	6756 Mowry		Newark	California	United States	94560	cfb5ed2a-43fe-4102-8413-73c74dacc79e	2006-08-01
1006	25472 Marlay Ave		Fontana	California	United States	92335	90c5f983-b302-4f6c-9d8a-bfb9e1bff406	2007-09-01
1011	9700 Sisk Road		Modesto	California	United States	95354	fb563347-c157-43b3-a7e4-a851495df874	2006-09-01
1013	54254 Pacific Ave.		Stockton	California	United States	95202	7b82a39f-d175-4543-8f6f-ae6fe9c34a99	2006-08-01
1014	25136 Jefferson Blvd.		Culver City	California	United States	90232	05196d3f-f726-4ab9-a66b-af87500846ad	2006-09-01
1015	99000 S. Avalon Blvd. Suite 750		Carson	California	United States	90746	206163d9-0324-476d-909e-607fe07b2a2a	2005-08-01
1016	72502 Eastern Ave.		Bell Gardens	California	United States	90201	06431d74-4caf-442c-870e-efc20f9d45b8	2006-09-01
1018	630 N. Capitol Ave.		San Jose	California	United States	95112	aceb7a03-c8e6-43f6-94f2-266e571bd413	2007-07-01
1020	251340 E. South St.		Cerritos	California	United States	90703	ba72737a-0827-46a8-af5b-971fd8464a87	2005-09-01
1021	440 West Huntington Dr.		Monrovia	California	United States	91016	2ae4d8c2-6d11-4945-a032-430dc8ce5b1b	2005-07-01
1022	99235 Hawthorne Blvd.		Torrance	California	United States	90505	52724cd9-1dc9-4d28-9b89-c404692c0b73	2005-07-01
1027	409 Santa Monica Blvd.		Santa Monica	California	United States	90401	58d58e3e-bf3e-4e0c-8f96-1717821e3516	2006-09-01
1028	Mall Of Orange		Orange	California	United States	92867	8d9b75cc-90c4-4e15-b69c-891126386ba6	2005-08-01
1029	Topanga Plaza		Canoga Park	California	United States	91303	0b945127-9c6f-47a1-80e4-2fb0e057fa55	2007-07-01

PowerBI dashboard page 1



PowerBI dashboard page 2



11. Future Scope and Limitations

Future Scope:

- Extending the solution to support real-time data streaming and processing for low-latency analytics.
- Incorporating advanced machine learning and AI capabilities for predictive analytics and automated decision-making.
- Enhancing the solution to handle multi-cloud environments and hybrid architectures.
- Implementing advanced data governance and lineage tracking mechanisms for improved compliance and auditing.
- Exploring serverless computing options for increased scalability and cost optimization.

Limitations:

- The solution is primarily focused on data migration and may require additional components for advanced use cases like real-time analytics or IoT data processing.
- While the solution demonstrates data migration from on-premises SQL Server, additional connectors and adaptations may be required for other data sources or formats.
- The project scope is limited to the Microsoft Azure ecosystem, and additional work may be required for integrating with other cloud platforms or on-premises systems.
- Advanced data governance and lineage tracking features may require further development and integration with external tools or services.
- The solution does not cover aspects of performance optimization, cost optimization, or auto-scaling, which may be relevant for large-scale deployments.

12. 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	COMPLETED
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	COMPLETED
Submission	23 April 2024 – 30 April 2024	Final Review and the submission of Dissertation	COMPLETED

13. 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.

14. 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

15.Glossary

Cloud-First Approach: Strategic prioritization of cloud-based solutions over traditional on-premises infrastructure for IT deployments.

Engineering: Application of scientific principles to design and build systems to solve specific problems.

Efficient: Achieving maximum productivity with minimal resources.

On-Premises: Traditional hosting of computing infrastructure within an organization's physical premises.

Data Migration: Transfer of data from one system or storage location to another.

Cloud Platforms: Infrastructure and services provided by cloud service providers.

Azure Data Factory: Cloud-based data integration service for orchestrating data movement and transformation workflows.

Azure Databricks: Apache Spark-based analytics platform for big data processing and analytics.

Azure Synapse Analytics: Cloud-based analytics service integrating data warehousing and big data analytics.

Power BI: Business analytics service for creating interactive visualizations and reports.

Data: Information stored and processed by computer systems.

Infrastructure: Physical or virtual components supporting computing environments.

Checklist of Items for the Final Dissertation / Project / Project Work Report

This checklist is to be attached as the last page of the final report.

This checklist is to be duly completed, verified and signed by the student.

1.	**Is the final report neatly formatted with all the elements required for a technical Report?**	**Yes**
2.	Is the Cover page in proper format as given in Annexure A?	**Yes**
3.	Is the Title page (Inner cover page) in proper format?	**Yes**
4.	(a) Is the Certificate from the Supervisor in proper format? (b) Has it been signed by the Supervisor?	**Yes** **Yes**
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	**Yes** **Yes**
6.	Is the title of your report appropriate? **The title should be adequately descriptive, precise and must reflect scope of the actual work done.** Uncommon abbreviations / Acronyms should not be used in the title	**Yes**
7.	Have you included the List of abbreviations / Acronyms?	**Yes**
8.	Does the Report contain a summary of the literature survey?	Yes
9.	Does the Table of Contents include page numbers? (i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1) (ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures) (iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) (iv). Are the Captions for the Figures and Tables proper? (v). Are the Appendices numbered properly? Are their titles appropriate	**Yes** **Yes** **Yes** **Yes** **Yes** **Yes**
10.	Is the conclusion of the Report based on discussion of the work?	**Yes**
11.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report	**Yes** **Yes** **Yes**
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a PowerPoint Presentation, or a user manual. Source code of software need not be included in the report.	**Yes**

Declaration by Student:
I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

Place: Pune, Maharashtra, India **Signature of the Student**

Date: 29th April 2024 **Name: SHELKE AKSHAY NANDKUMAR**

ID No.: 2022MT93331