

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

I am deeply indebted to numerous individuals and institutions whose guidance, support, and resources played a pivotal role in the successful completion of this endeavor. First and foremost, I extend my sincere gratitude to the esteemed Birla Institute of Technology and Science, Pilani, for providing me with the invaluable opportunity to pursue the M.Tech. Software Engineering program. My heartfelt appreciation goes to Prof. Pramod Bide, whose insightful feedback and mentorship steered me in the right direction throughout this project.

I am profoundly grateful to Mr. Shubham Tulsyan, Senior Project Manager, and Mr. Prashant Gorade, Senior Software Engineer, whose unwavering encouragement, valuable suggestions, and consistent support from the inception of this work were instrumental in its successful completion. Their expertise and guidance were invaluable assets throughout this journey.

I would also like to express my sincere gratitude to Schlumberger India Technology Center Ltd., Pune, for their generosity in providing the necessary resources and facilities crucial for the successful execution of this dissertation work. Additionally, I extend my appreciation to BITS, Pilani, and the Training department for their commendable efforts in organizing and facilitating this enriching course.

Lastly, words cannot adequately express my gratitude to my family and parents, whose constant love, support, and motivation were the driving force behind my perseverance and determination to fulfill the requirements of this course. Their unwavering belief in me has been a source of strength throughout this journey.

**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.

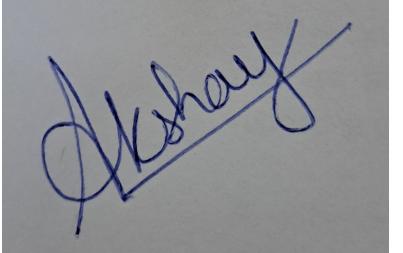
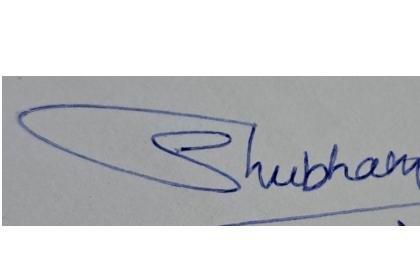
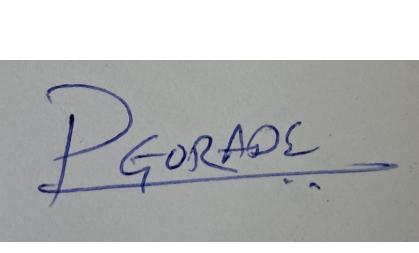
Abstract

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.

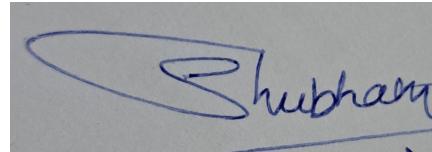
Azure Key Vault and Azure Active Directory ensure secure key management and identity & access management, respectively, contributing to a secure and compliant data environment.

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 Tulysyan	Name: Mr. Prashant Gorade
Date : 28 th April 2024	Date : 28 th April 2024	Date : 28 th 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 a wide range of services, for data storage, data processing, and data analytics. Azure provides a robust set of tools and 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 is 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 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:

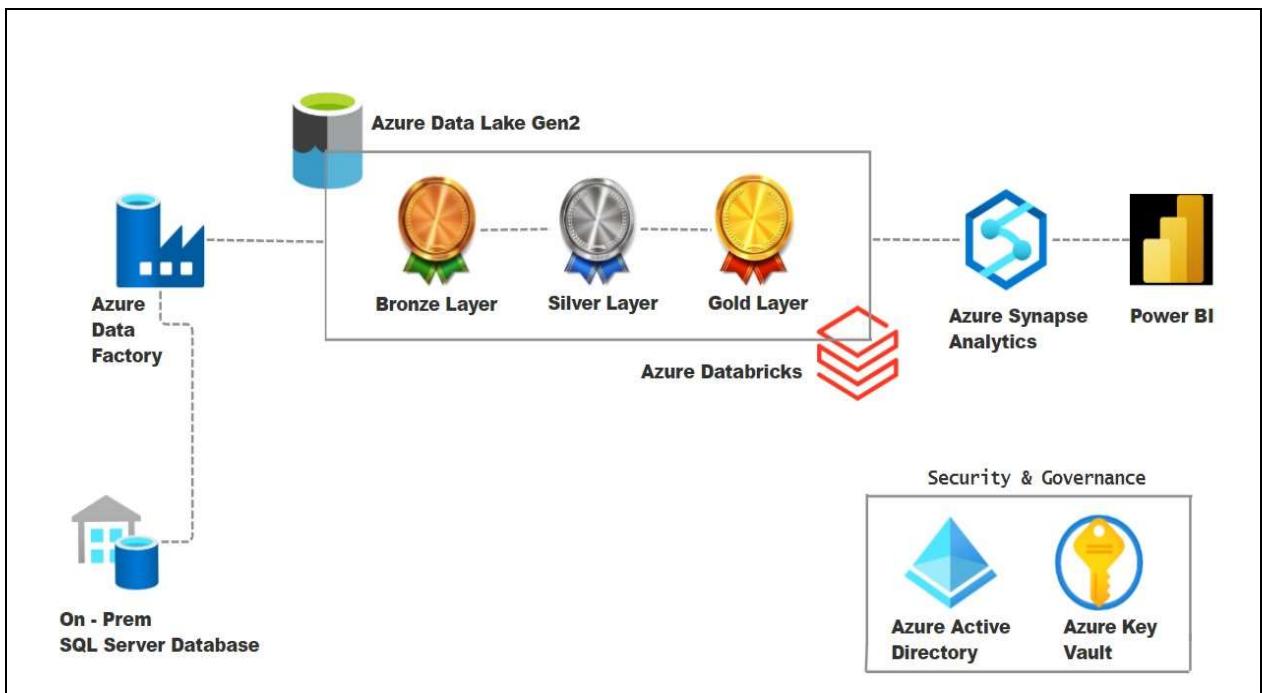
- **Data Connectivity:** Ability to connect to various data sources, including Azure Data Lake Gen2, Azure Synapse Analytics, other cloud & on-premises data sources.
- **Data Modeling:** Advanced data modeling capabilities, including support for relationships, hierarchies, and calculated columns.
- **Interactive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

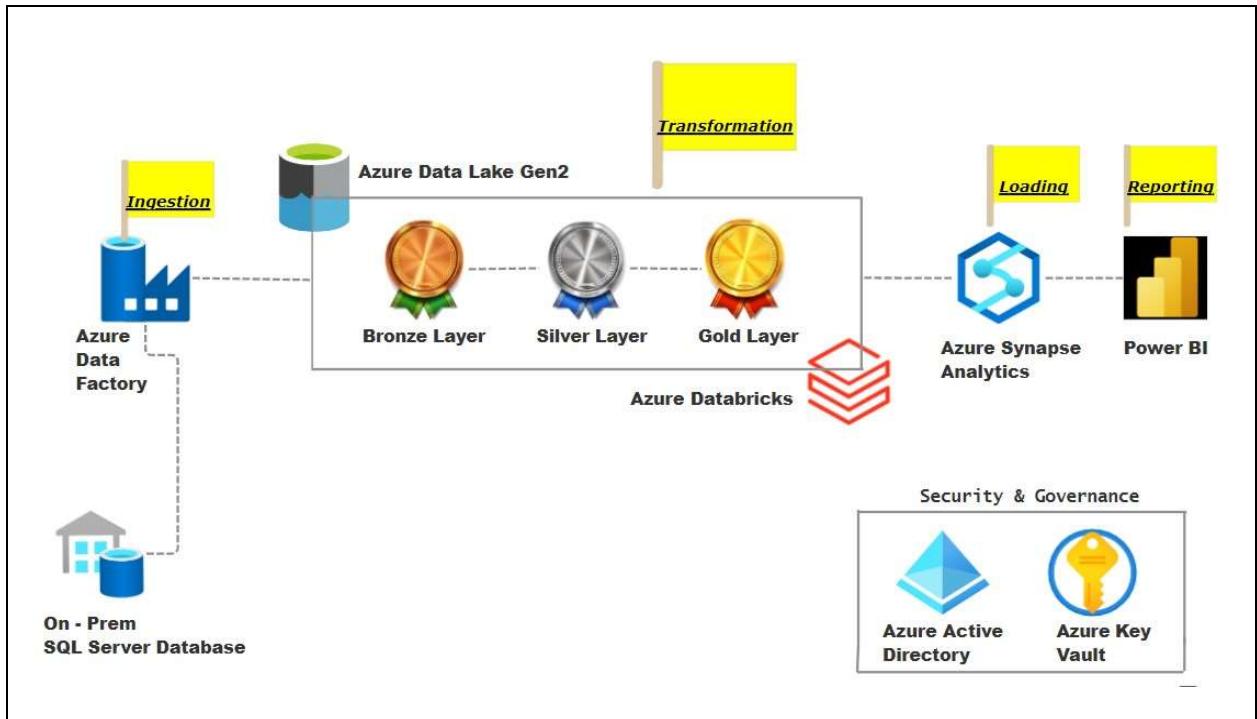
The transformation and processing tasks performed by Databricks can include:

- **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.)

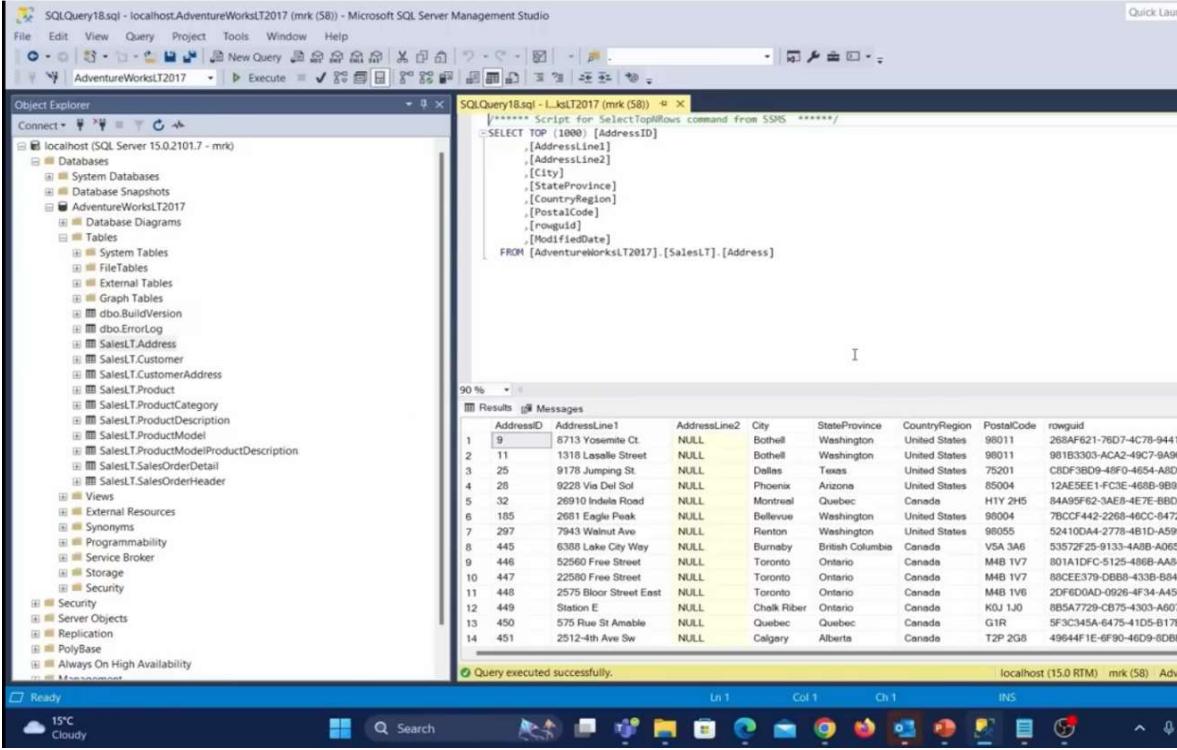
The screenshot shows the Microsoft Azure portal interface for the resource group 'rg-mtech-dissertation-2022mt93331'. The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Settings (Deployments, Security, Deployment stacks, Policies, Properties, Locks), and Cost Management (Cost analysis, Cost alerts (preview)). The main content area is titled 'Essentials' and shows a list of resources under 'Resources'. The table lists five resources:

Name	Type	Location	Actions
adf-sheilke-akshay-2022mt93331	Data factory (V2)	East US	...
az-databricks-2022mt93331	Azure Databricks Service	East US	...
kv-shelke-akshay-001	Key vault	East US	...
shelke-akshay-synw-001	Synapse workspace	East US	...
storageaccmtechdemo	Storage account	East US	...

At the bottom of the portal interface, there is a search bar, a ribbon of icons, and a status bar showing weather (30°C Clear), date (28-04-2024), and time (23:48).

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

On Premises SQL Database



The screenshot shows the Microsoft SQL Server Management Studio interface. In the Object Explorer, the connection is to 'localhost (SQL Server 15.0.2101.7 - mrk)'. The 'AdventureWorksLT2017' database is selected. In the 'Tables' node, the 'Address' table is expanded. A query window titled 'SQLQuery18.sql - L:\SQL2017 (mrk (58))' contains the following script:

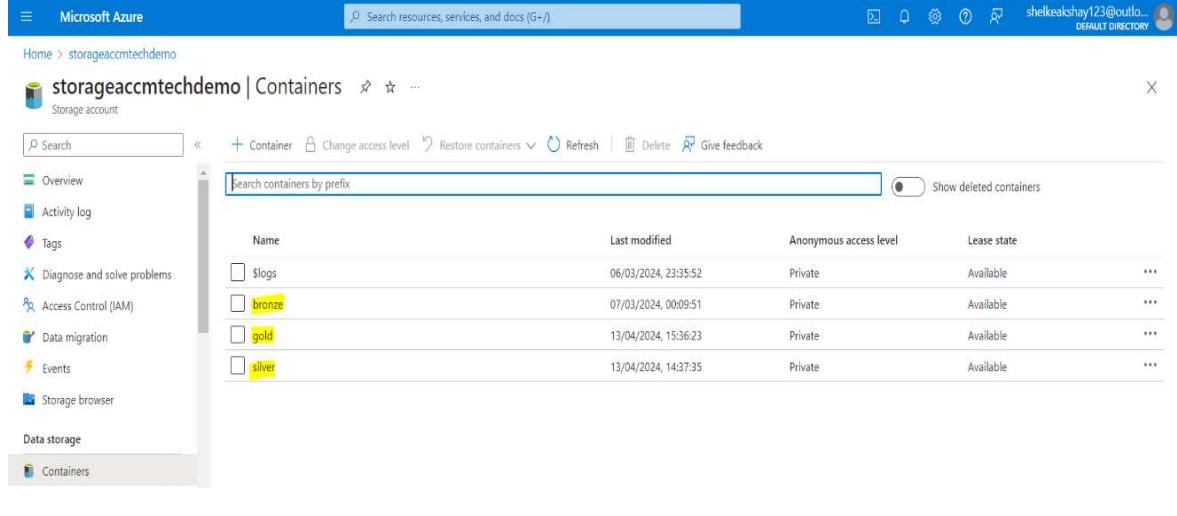
```
***** Script for SelectTopNRows command from SSMS *****
--SELECT TOP (1000) [AddressID]
--      ,[AddressLine1]
--      ,[AddressLine2]
--      ,[City]
--      ,[StateProvince]
--      ,[CountryRegion]
--      ,[PostalCode]
--      ,[rowguid]
--      ,[ModifiedDate]
FROM [AdventureWorksLT2017].[SalesLT].[Address]
```

The results grid shows 14 rows of address data. The columns are: AddressID, AddressLine1, AddressLine2, City, StateProvince, CountryRegion, PostalCode, and rowguid. The data includes various addresses across the United States and Canada.

AddressID	AddressLine1	AddressLine2	City	StateProvince	CountryRegion	PostalCode	rowguid
9	8713 Yosemite Ct.	NULL	Bothell	Washington	United States	98011	268AF621-76D7-4C7B-9441-981B3303-ACA2-49C7-9496-
11	1318 Lasalle Street	NULL	Bothell	Washington	United States	98011	C8DF3B09-48F0-4654-A10D
25	9178 Jumping St.	NULL	Dallas	Texas	United States	75201	12AE5EE1-FC3E-468D-BB02
28	9228 Via Del Sol	NULL	Phoenix	Arizona	United States	85004	84A95F62-3AE8-4E7E-BB05
32	26910 Indira Road	NULL	Montreal	Quebec	Canada	H1Y 2H5	7BC0CF442-226B-46CC-8472-
185	2681 Eagle Peak	NULL	Bellevue	Washington	United States	98004	524100A4-2778-4B1D-A599-
297	7943 Walnut Ave	NULL	Renton	Washington	United States	98005	53572F25-9133-4A8B-A02-
445	6380 Lake City Way	NULL	Burnaby	British Columbia	Canada	V5A 3A6	801A1DFC-5125-466B-AA84-
446	52560 Free Street	NULL	Toronto	Ontario	Canada	M4V 1V7	88CEE379-DBB8-413B-B84E
447	22580 Free Street	NULL	Toronto	Ontario	Canada	M4V 1V7	20FD60AD-0920-4F34-A450
448	2575 Bloor Street East	NULL	Toronto	Ontario	Canada	K0J 1J0	885A7729-CB75-4303-A007-
450	575 Rue St Amble	NULL	Chalk River	Ontario	Canada	G1R	5F3C145A-6475-41D5-B17B-
451	2512-4th Ave Sw	NULL	Calgary	Alberta	Canada	T2P 2G8	49644F1E-6F90-4ED9-BD8B-

At the bottom of the results grid, it says 'Query executed successfully.'

Azure Data Lake Storage (Gen2)



The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar shows the navigation path: Home > storageaccmtechdemo. The main area displays the 'Containers' section for the 'storageaccmtechdemo' storage account. The container names listed are \$logs, bronze, gold, and silver. The 'gold' and 'silver' containers are highlighted with yellow boxes.

Name	Last modified	Anonymous access level	Lease state
\$logs	06/03/2024, 23:35:52	Private	Available
bronze	07/03/2024, 00:09:51	Private	Available
gold	13/04/2024, 15:36:23	Private	Available
silver	13/04/2024, 14:37:35	Private	Available

Storage Mount in Azure Databricks

This screenshot shows a Databricks notebook titled "1_storage_mount" running in Python 3. The notebook contains two cells. Cell 1 contains the following code:

```

config = {
    "fs.azure.account.auth.type": "CustomAccessToken",
    "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

try:
    dbutils.fs.mount(
        source = "abfss://bronze@storageaccmtechdemo.dfs.core.windows.net/",
        mount_point = "/mnt/bronze",
        extra_configs = config)
    print("Mount Point created successfully")
except:
    print("Mount Point already exists")

```

Cell 2 shows the output of the command `dbutils.fs.ls("/mnt/bronze/SalesLT")`:

```

[FileInfo(path="dbfs:/mnt/bronze/SalesLT/Address/", name="Address/", size=0, modificationTime=1711712578800),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/Customer/", name="Customer/", size=0, modificationTime=1711712585000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/Product/", name="Product/", size=0, modificationTime=1711712577000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/ProductCategory/", name="ProductCategory/", size=0, modificationTime=1711712576000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/ProductDescription/", name="ProductDescription/", size=0, modificationTime=1711712589000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/ProductModel/", name="ProductModel/", size=0, modificationTime=1711712579000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/ProductModelDescription/", name="ProductModelDescription/", size=0, modificationTime=1711712577000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/SalesOrderDetail/", name="SalesOrderDetail/", size=0, modificationTime=1711712579000),
 FileInfo(path="dbfs:/mnt/bronze/SalesLT/SalesOrderHeader/", name="SalesOrderHeader/", size=0, modificationTime=1711712581000)]

```

This screenshot shows a Databricks notebook titled "1_storage_mount" running in Python 3. It contains two cells. Cell 1 is identical to the one in the previous screenshot. Cell 2 contains the following code:

```

config = {
    "fs.azure.account.auth.type": "CustomAccessToken",
    "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

try:
    dbutils.fs.mount(
        source = "abfss://silver@storageaccmtechdemo.dfs.core.windows.net/",
        mount_point = "/mnt/silver",
        extra_configs = config)
    print("Mount Point created successfully")
except:
    print("Mount Point already exists")

```

Cell 3 shows the output of the command `dbutils.fs.ls("/mnt/silver")`:

```

[FileInfo(path="dbfs:/mnt/silver/")]

```

Code – Bronze to Silver Data Transformation

```

2_bronze_to_silver Python ✘
File Edit View Run Help Last edit was 16 days ago New cell UI: ON
Run all Terminated Schedule Share
4/13/2024 (5h)
from pyspark.sql.functions import from_utc_timestamp, date_format
from pyspark.sql.types import TimestampType

for i in table_name:
    path = "/mnt/Bronze/SalesLT/" + i + "/" + ".parquet"
    df = spark.read.format("parquet").load(path)
    column = df.columns

    for col in column:
        if "Date" in col or "date" in col:
            df = df.withColumn(col, date_format(from_utc_timestamp(df[col].cast(TimestampType()), "UTC"), "yyyy-MM-dd"))

    output_path = "/mnt/silver/SalesLT/" + i + "/"
    df.write.format("delta").mode("overwrite").save(output_path)

(64) Spark jobs
df: pyspark.sql.dataframe.DataFrame = [SalesOrderID: integer, RevisionNumber: integer ... 20 more fields]

```

This screenshot shows a Databricks workspace with a notebook titled '2_bronze_to_silver' in Python. The code reads parquet files from the 'Bronze' folder and writes them to the 'Silver' folder as Delta tables. It uses the `from_utc_timestamp` function to convert UTC timestamps to a specific date format.

Code – Silver to Gold Data Transformation

```

3_silver_to_gold Python ✘
File Edit View Run Help Last edit was 16 days ago New cell UI: ON
Run all Terminated Schedule Share
4/13/2024 (4h)
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, regexp_replace

for name in table_name:
    path = "/mnt/silver/SalesLT/" + name
    print(path)
    df = spark.read.format("delta").load(path)

    # Get the list of column names
    column_names = df.columns

    for old_col_name in column_names:
        # Convert column name from ColumnName to Column_Name format
        new_col_name = "".join(["_" + char if char.isupper() and not old_col_name[i-1].isupper() else char for i, char in enumerate(old_col_name)]).lstrip("_")

        # Change the column name using withColumnRenamed and regexp_replace
        df = df.withColumnRenamed(old_col_name, new_col_name)

    output_path = "/mnt/gold/SalesLT/" + name + "/"
    df.write.format("delta").mode("overwrite").save(output_path)

(64) Spark jobs
df: pyspark.sql.dataframe.DataFrame = [Sales_Order_ID: integer, Revision_Number: integer ... 20 more fields]
/mnt/silver/SalesLT/Address
/mnt/silver/SalesLT/Customer
/mnt/silver/SalesLT/CustomerAddress
/mnt/silver/SalesLT/Product
/mnt/silver/SalesLT/ProductCategory
/mnt/silver/SalesLT/ProductDescription
/mnt/silver/SalesLT/ProductModel
/mnt/silver/SalesLT/ProductModelProductDescription
/mnt/silver/SalesLT/SalesOrderDetail
/mnt/silver/SalesLT/SalesOrderHeader

```

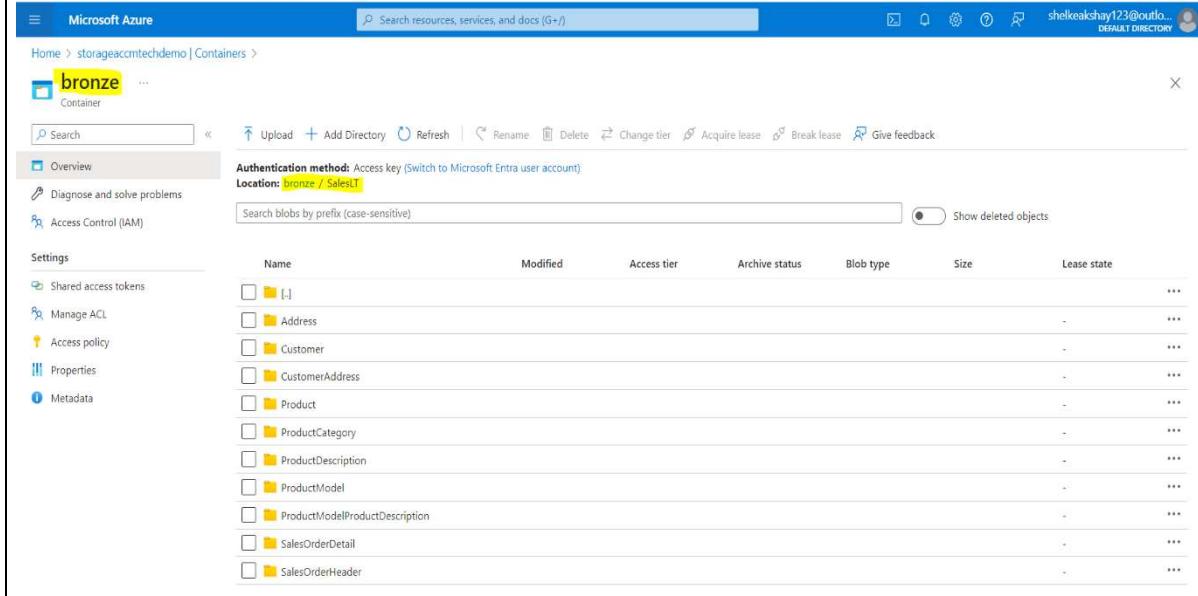
This screenshot shows a Databricks workspace with a notebook titled '3_silver_to_gold' in Python. The code reads Delta tables from the 'Silver' folder, renames columns to follow a specific naming convention (e.g., ColumnName to Column_Name), and then writes the data to the 'Gold' folder as Delta tables.

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

The screenshot shows the Azure Data Factory pipeline interface. The pipeline named 'copy_all_tables' is displayed, consisting of a 'Lookup' activity followed by a 'ForEach' loop. The 'ForEach' loop iterates over 'For Each Schema Table' activities, which then lead to 'Activities' (specifically 'Copy Each Table') and finally 'Notebook' activities ('Bronze to Silver' and 'Silver to Gold'). All activities in the pipeline have succeeded.

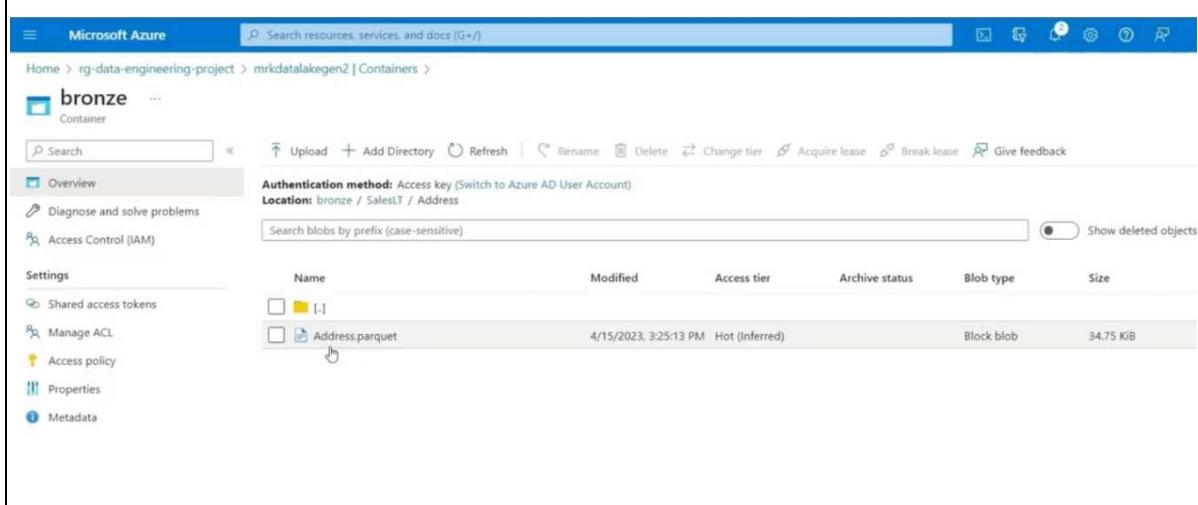
Activity	Status	Start Time	Duration	Last Modified
Lookup	Succeeded	4/22/2024, 7:59:25 PM	53s	14b201c3-394c-410c-add2-e0270ac29bc
For Each Schema Table	Succeeded	4/22/2024, 7:59:27 PM	33s	127a0bec-265e-48fb-a7eb-f2e3d47c9daa
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	41s	7bf0fdb11-af23-4349-8319-73a4a0e0859
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	47s	fb2047e4-24b2-4799-a57c-dd01a706e6dc
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	29s	1bba8f5c-5c1-4a9c-bb1b-2ef2d2c42538
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	47s	1bc3cede-29d8-4418-955e-060e17bd0807
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	45s	3d5adcc4-837d-4ccb-93e6-dd514937500
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	31s	c155ef48-74c9-46ab-a627-30c8750fb10
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	43s	4d56f496-60c8-414b-b727-8792e218825e
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	43s	9f05404f-a577-4f0a-8482-0a44e58ed0c
Copy Each Table	Succeeded	4/22/2024, 7:59:27 PM	41s	dcc10dd9-21cf-457f-b791-ac955569b3e
Bronze to Silver	Succeeded	4/22/2024, 8:00:19 PM	4m 52s	3a755135-c2b3-4fbf-86d8-724099b0d34e
Silver to Gold	Succeeded	4/22/2024, 8:05:12 PM	55s	f5c2035f-2a8f-4342-ae0d-989bab09bc87

Bronze Layer Parquet Files (Data Ingestion Step)



The screenshot shows the Microsoft Azure Storage Explorer interface for the 'bronze' container. The left sidebar shows navigation options like Overview, Diagnose and solve problems, Access Control (IAM), and Settings. The main area displays a list of blobs with the following details:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]						---
Address						---
Customer						---
CustomerAddress						---
Product						---
ProductCategory						---
ProductDescription						---
ProductModel						---
ProductModelProductDescription						---
SalesOrderDetail						---
SalesOrderHeader						---



The screenshot shows the Microsoft Azure Storage Explorer interface for the 'bronze' container. The left sidebar shows navigation options like Overview, Diagnose and solve problems, Access Control (IAM), and Settings. The main area displays a list of blobs with the following details:

Name	Modified	Access tier	Archive status	Blob type	Size
[...]					
Address.parquet	4/15/2023, 3:25:13 PM	Hot (Inferred)		Block blob	34.75 KB

Silver Layer Parquet Files (Post Data Transformation Step)

This screenshot shows the Microsoft Azure Storage Explorer interface for the 'silver' container. The left sidebar contains navigation links: Overview, Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and a blob list. The main area displays a table of blobs with columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The table lists several directory entries (e.g., 'Address', 'Customer') and numerous parquet files. A search bar at the top right allows filtering by prefix.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						...
Address						...
Customer						...
CustomerAddress						...
Product						...
ProductCategory						...
ProductDescription						...
ProductModel						...
ProductModelProductDescription						...
SalesOrderDetail						...
SalesOrderHeader						...

This screenshot shows the Microsoft Azure Storage Explorer interface for the 'silver' container, focusing on the 'Address' directory. The table now lists several parquet files, with one specific file highlighted in yellow: 'part-00000-f2f441bc-de37-4be0-b734-3c45c60e1b66-c000.snappy.parquet'. This file was generated during the data transformation step. The rest of the table structure remains the same as the first screenshot.

Name	Modified	Access tier	Archive status	Blob type	Size	
[..]						
._delta_log						
part-00000-168ca4cf-fac0-4848-9452-611b01e3e1c5-c000.snappy.parquet	13/04/2024, 16:28:17	Hot (Inferred)		Block blob	34.74 KIB	
part-00000-9fdde1bb5-fafc-4584-83ec-eac96f77fd25-c000.snappy.parquet	13/04/2024, 16:25:23	Hot (Inferred)		Block blob	34.74 KIB	
part-00000-b8b24f3d-9dbc-4050-a459-b4f49ffec1c5-c000.snappy.parquet	13/04/2024, 15:03:51	Hot (Inferred)		Block blob	34.74 KIB	
part-00000-daa052c4-c56b-4ec9-93d9-206cb6a7b214-c000.snappy.parquet	13/04/2024, 16:27:33	Hot (Inferred)		Block blob	34.74 KIB	
part-00000-f2f441bc-de37-4be0-b734-3c45c60e1b66-c000.snappy.parquet	22/04/2024, 20:04:14	Hot (Inferred)		Block blob	34.74 KIB	

Gold Layer Parquet Files (Post Data Transformation Step)

The screenshot shows the Microsoft Azure Storage Explorer interface for the 'gold' container. The left sidebar lists navigation options: Home, storageaccmtechdemo | Containers, gold (selected), Overview, Diagnose and solve problems, and Access Control (IAM). The main area displays blob details for the 'SalesLT' folder under 'Address'. The table has columns: Name, Modified, Access tier, Archive status, Blob type, and Size.

Name	Modified	Access tier	Archive status	Blob type	Size
[...]					
Address					
Customer					
CustomerAddress					
Product					
ProductCategory					
ProductDescription					
ProductModel					
ProductModelProductDescription					
SalesOrderDetail					
SalesOrderHeader					

The screenshot shows the Microsoft Azure Storage Explorer interface for the 'gold' container. The left sidebar lists navigation options: Home, storageaccmtechdemo | Containers, gold (selected), Overview, Diagnose and solve problems, and Access Control (IAM). The main area displays blob details for the 'SalesLT' folder under '_delta_log'. The table has columns: Name, Modified, Access tier, Archive status, Blob type, and Size.

Name	Modified	Access tier	Archive status	Blob type	Size
[...]					
_delta_log					
part-00000-01647fd7-ac92-4660-8fc9-f085b1550b5c-c000.snappy.parquet	13/04/2024, 16:28:27	Hot (Inferred)		Block blob	34.76 KiB
part-00000-4410b901-5695-4493-8f52-502fa58de2ee-c000.snappy.parquet	13/04/2024, 16:26:36	Hot (Inferred)		Block blob	34.76 KiB
part-00000-67c374d-7a32-425f-9041-f88ac62e483a-c000.snappy.parquet	13/04/2024, 15:39:03	Hot (Inferred)		Block blob	34.76 KiB
part-00000-deb31fb8-b21d-4cd3-8681-ce6cbd2e5f47-c000.snappy.parquet	22/04/2024, 20:05:27	Hot (Inferred)		Block blob	34.76 KiB
part-00000-fb5610c7-5584-41b6-826c-fccf4908a224-c000.snappy.parquet	13/04/2024, 16:29:07	Hot (Inferred)		Block blob	34.76 KiB

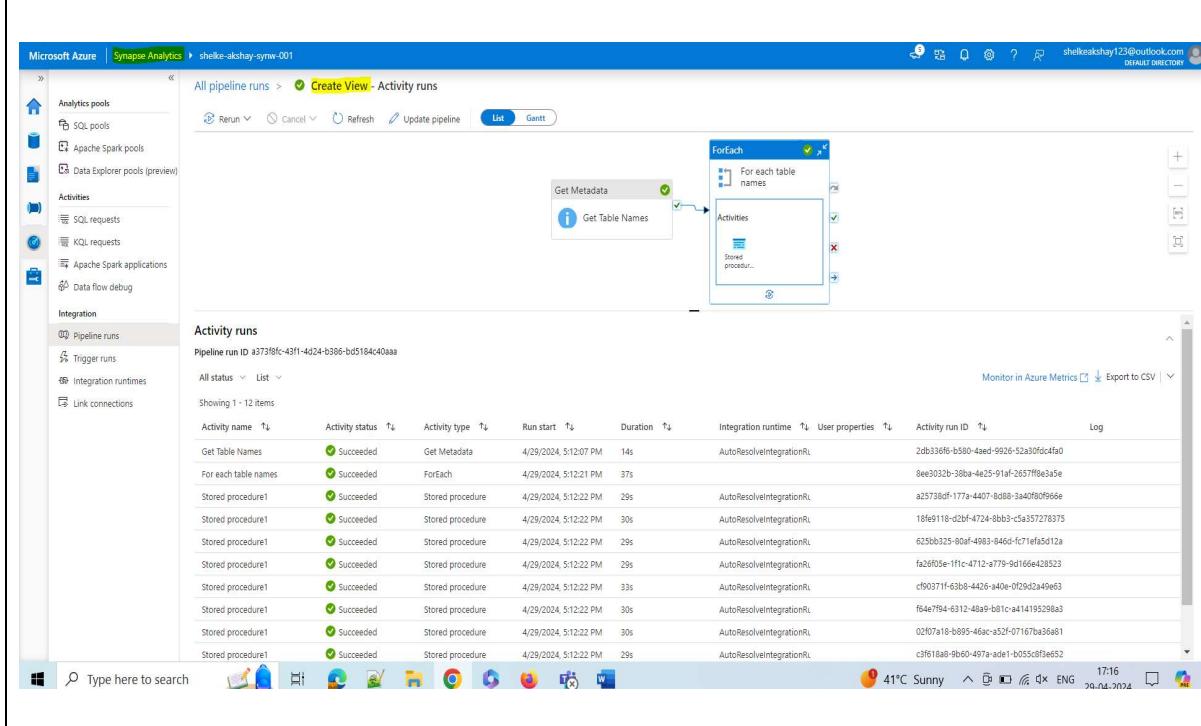
Code – Stored Procedure to create views on gold dataset

```

Microsoft Azure | Synapse Analytics > shelke-akshay-syn-001
Develop + <
Filter resources by name
SQL scripts 1
SP_CreateSQLServerlessView_gold
1 USE gold_db
2 GO
3
4
5 CREATE OR ALTER PROC CreateSQLServerlessView_gold @viewName nvarchar(100)
6 AS
7 BEGIN
8
9 DECLARE @statement VARCHAR(MAX)
10
11 SET @statement = N'CREATE OR ALTER VIEW ' + @viewName + ' AS
12     SELECT *
13     FROM
14     OPENROWSET(
15         BULK ''https://storageaccntechdemo.dfs.core.windows.net/gold/SalesLT/' + @viewName + '/',
16         FORMAT = ''DELTA''
17     ) AS [result]
18
19
20 EXEC (@statement)
21
22 END
23 GO

```

Azure Synapse Analytics – Data Loading



Microsoft Azure | Synapse Analytics > shelka-akshay-syna-001

Validate all | Publish all | SQL script 1

Data | Workspace | Linked

SQL database | gold_db (5GB) | External tables | External resources | Views | dbo.address | dbo.Customer | dbo.CustomerAddress | dbo.Product | dbo.ProductCategory | dbo.ProductDescription | dbo.ProductModel | dbo.ProductModelProductDetail | dbo.SalesOrderDetail | dbo.SalesOrderHeader | System views | Schemas | Security

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

```
1 SELECT TOP (100) [Address_ID]
2 ,[Address_Line1]
3 ,[Address_Line2]
4 ,[City]
5 ,[State_Province]
6 ,[Country_Region]
7 ,[Postal_Code]
8 ,[rowguid]
9 ,[Modified_Date]
10 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 | Export results | Search

Address_ID	Address_Line1	Address_Line2	City	State_Province	Country_Region	Postal_Code	rowguid	Modified_Date
9	8713 Vosemitte Ct.	(NULL)	Bothell	Washington	United States	98011	268af621-76d7...	2006-07-01
11	1318 Lassalle Street	(NULL)	Bothell	Washington	United States	98011	981b3303-acd2...	2007-04-01
25	9178 Jumping St.	(NULL)	Dallas	Texas	United States	75201	c0f3bb09-48b0...	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	8495562-3ae8...	2006-08-01

0000:07 Query executed successfully.

Type here to search | Windows Start | Taskbar icons | Weather: 41°C Sunny | Date: 29-04-2024 | Time: 17:18

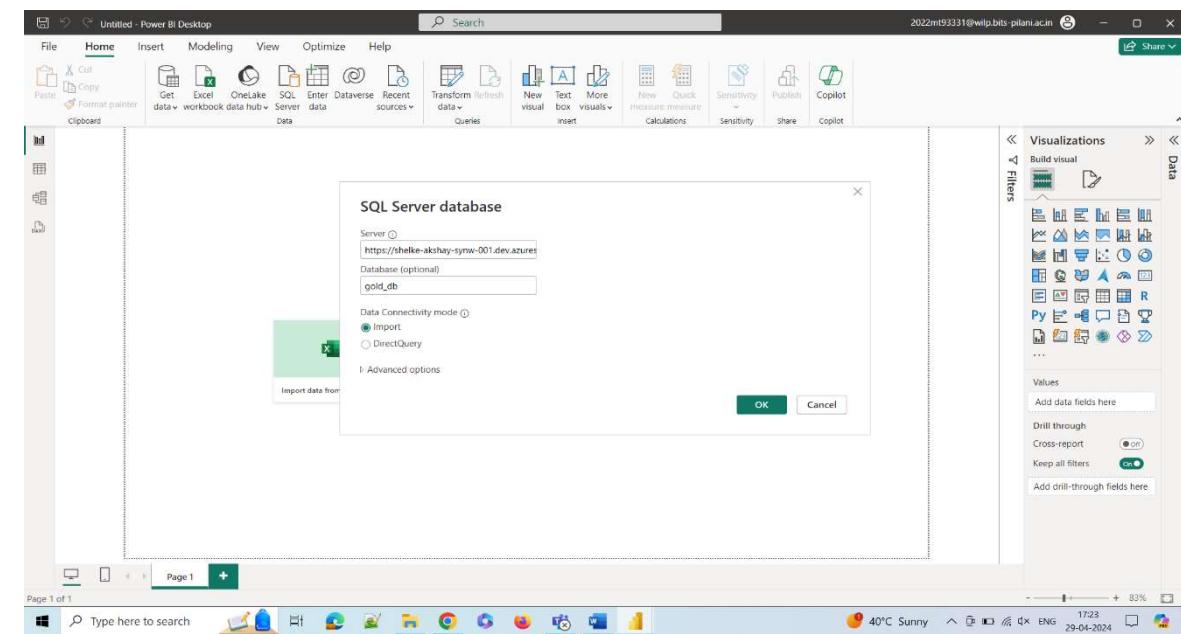
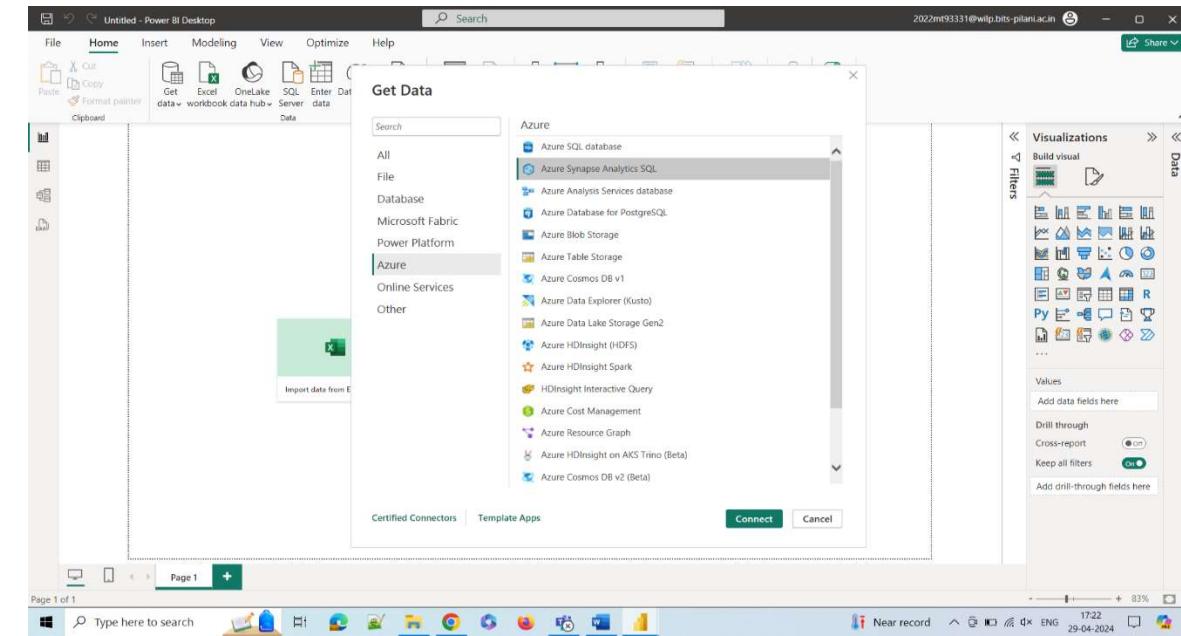
10. Conclusions – Results on PowerBI dashboards

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



Untitled - Power BI Desktop

File Home Insert Modeling View Optimize Help

Clipboard

Windows Database Microsoft account

Import data from

SQL Server database

<https://shelke-akshay-synw-001.dev.azuresql.net>

You are currently signed in.

Sign in as different user

Select which level to apply these settings to

<https://shelke-akshay-synw-001.dev.azuresql.net>

Waiting for https://shelke-akshay-synw-001.dev.azuresql.net...

Back Connect Cancel

Visualizations Build visual Data

Add data fields here

Drill through Cross-report Keep all filters Add drill-through fields here

Page 1 of 1

Type here to search

PowerBI_Dashboard • Last saved: Yesterday at 7:03 PM

File Home Help Table tools

Name address

Mark as date table Calendars Manage relationships Relationships New Quick measure column table Calculations

Structure

Data

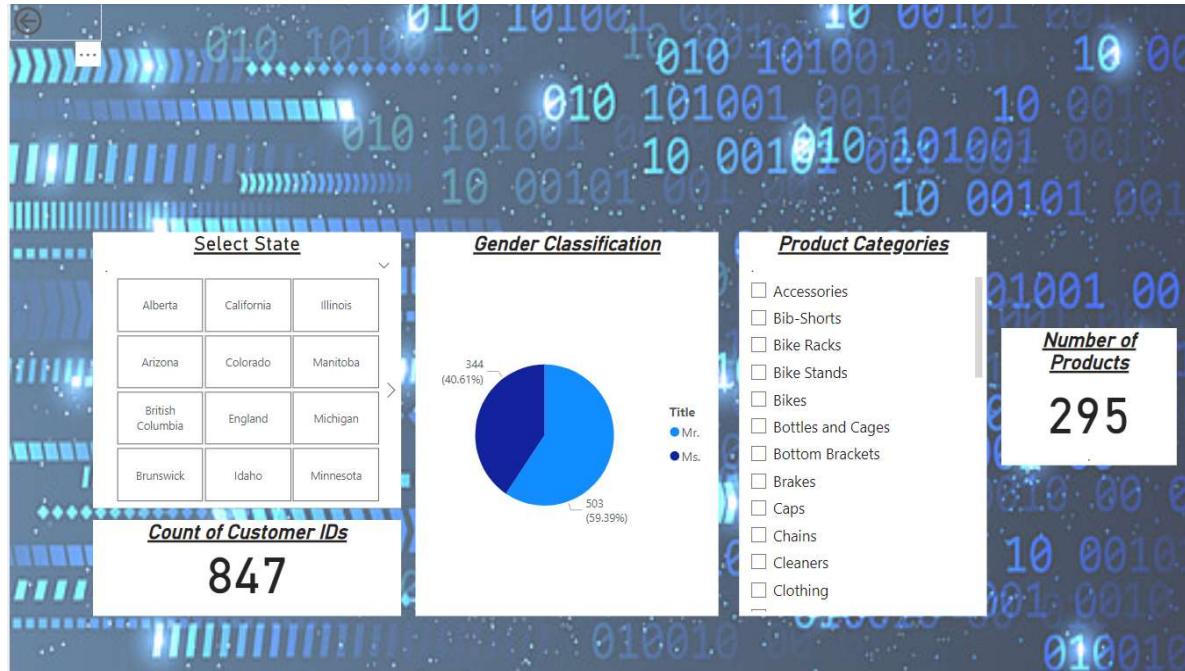
Address_ID Address_Line1 Address_Line2 City State_Province Country_Region Postal_Code rowguid Modified_Date

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	b81241ca-1db6-4f96-861e-1fdcc3796443	2007-09-01
991	60025 Bollinger Canyon Road		San Ramon	California	United States	94583	962ad00c-c5aa-41cc-b7bc-2afe2a6568b	2006-09-01
992	9902 Whipple Rd		Union City	California	United States	94587	882581b1-b50f-440f-9369-5d7aef831d5fc	2006-09-01
993	Corporate Office		El Segundo	California	United States	90245	6ccc74f-6389-4f08-8ed1-9079b42cf2fae	2006-09-01
994	25001 Montague Expressway		Milpitas	California	United States	95035	899012a2-fda8-4920-b22c-dd5d53601a	2006-07-01
995	4460 Newport Center Drive		Newport Beach	California	United States	92625	d69e9e9b-f288-4590-be12-8dabb5e53bc2	2006-07-01
997	70259 West Sunnyview Ave		Visalia	California	United States	93291	7ce0f511-de61-4761-80ea-6a7b2739e524	2007-09-01
998	60750 San Clemente		Hayward	California	United States	94541	823dc2ba-aec3-4382-93a6-5027cb6a7df	2007-09-01
999	Receiving		Fullerton	California	United States	92831	9305c32f-c2c6-4f27-891f-35ab04554976	2006-09-01
1000	23353 Paseo De Las Americas		San Diego	California	United States	92102	982b0cc2-6429-4723-8df1-4fa032ec908	2006-08-01
1001	Incom Sports Center		Ontario	California	United States	91764	079e5de8-169f-4eb8-be75-813b40a043e	2005-08-01
1003	5967 Las Postas Blvd		Pleasanton	California	United States	94566	6b26292f-1ebe-41ce-9a3e-f423c310644	2007-07-01
1004	25600 E St Andrews Pl		Santa Ana	California	United States	92701	489390e0-48c0-4bb4-a5dc-384b39e833	2005-09-01
1005	6756 Mowry		Newark	California	United States	94560	cfa5ed2a-43fe-4102-8413-734d4acc79e	2006-08-01
1006	25472 Marlay Ave		Fontana	California	United States	92335	90c5f983-b302-4ff6-9d8a-fbf8fe1bf4f0	2007-09-01
1011	9700 Sisk Road		Modesto	California	United States	95354	fb56334f-c157-43b3-a7e4-a851495df574	2006-09-01
1013	54254 Pacific Ave.		Stockton	California	United States	95202	7b82a39f-d175-4543-8f6f-a6e69c34ab	2006-08-01
1014	25136 Jefferson Blvd.		Culver City	California	United States	90232	0519663f-f726-4a9b-a66b-af87500846ad	2006-09-01
1015	99000 S. Avalon Blvd. Suite 750		Carson	California	United States	90746	200163e9-0324-4769-909e-607e67b2a2a	2005-08-01
1016	72502 Eastern Ave.		Bell Gardens	California	United States	90201	0643174a-4caf-442c-870e-ef20945458b	2006-09-01
1018	630 N. Capitol Ave.		San Jose	California	United States	95112	aceb7a03-c866-4396-94f2-366e571bd413	2007-07-01
1020	25140 E. South St.		Cerritos	California	United States	90703	ba72737a-0827-46a9-a5b-971fb464a87	2005-09-01
1021	440 West Huntington Dr.		Monrovia	California	United States	91016	2ae460c2-6d11-4945-a032-430dd0ce5a1b	2005-07-01
1022	99225 Hawthorne Blvd.		Torrance	California	United States	90505	52724cd9-1dc9-4d26-9b89-c04692d0b73	2005-07-01
1027	409 Santa Monica Blvd.		Santa Monica	California	United States	90401	58d583e-bf3e-4e0c-8f96-17717821e316	2006-09-01
1028	Mall Of Orange		Orange	California	United States	92867	8d9b75cc-90c4-4e15-b69c-891126380ba6	2005-08-01
1029	Topanga Plaza		Canoga Park	California	United States	91303	0b945127-9c6f-47a1-80e4-2fb0e057a55	2007-07-01

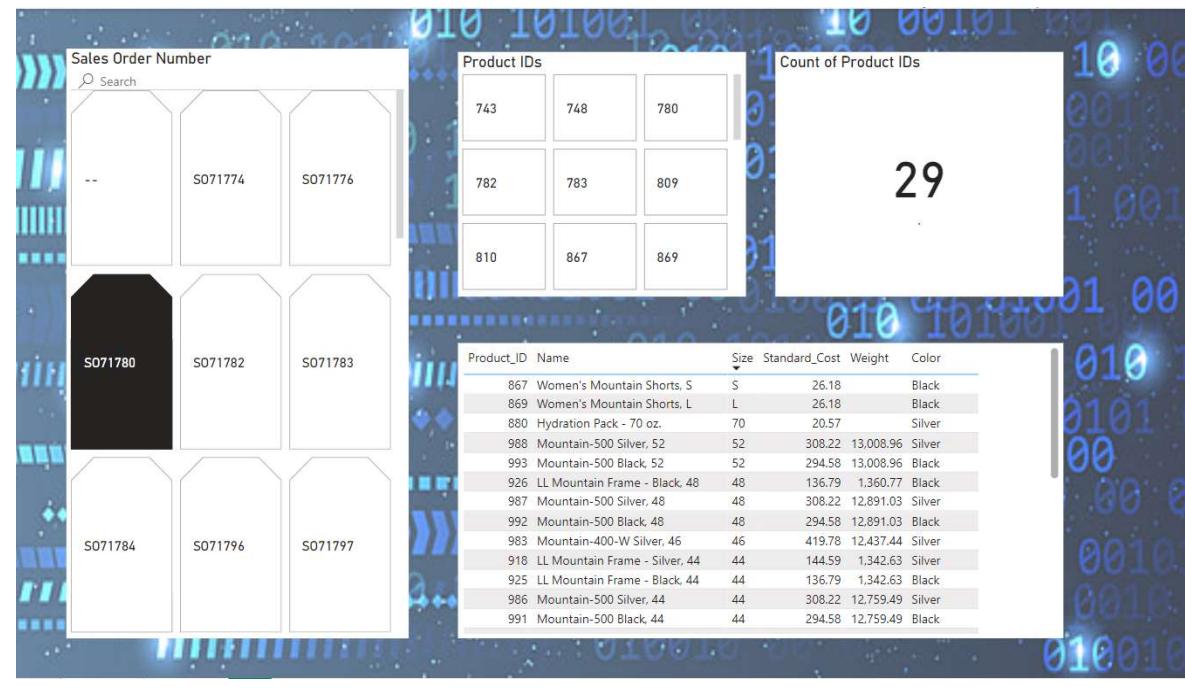
Table: address (459 rows)

Type here to search

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., Siddiq, 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., & Siddiq, 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