IST 615

Cloud Management

Final Project Report

Team 4

Submitted by:
Diya Shah
Priyanka Sheth
Sampada Regmi
Vrushali Lad

Table of Contents

PROJECT DESCRIPTION

Analyzing Employee Attrition in Healthcare using Azure Services

Employee attrition in the healthcare sector poses a critical challenge, leading to substantial financial costs, compromising the quality of patient care, and disrupting organizational stability. High turnover rates among healthcare professionals, particularly specialized staff like registered nurses, result in increased recruitment expenses and indirect costs, impacting team morale and patient satisfaction. The lack of continuity in care due to attrition also raises concerns about the consistency and quality of healthcare services delivered. Thus, the problem statement focuses on understanding the underlying causes and implications of employee attrition in healthcare and devising strategic solutions to mitigate its adverse effects on healthcare organizations, staff, and patient care.

Impact Analysis of Employee Attrition in Healthcare

1. **Financial Implications:**
    - ***High Cost of Replacement:*** A significant financial burden arises from the need to replace employees, especially specialized staff such as registered nurses. The American Society for Healthcare Human Resources Administration highlights that the average cost of replacing a registered nurse is approximately $48,000. This cost encompasses various factors, including recruitment, training, and the temporary loss of productivity as new hires reach full proficiency.
    - ***Indirect Costs*****:** Beyond direct replacement costs, there are also indirect costs such as the impact on team morale, the additional workload on existing staff, and potential errors or decreased care quality during transition periods.

2. **Impact on Quality of Care:**
    - ***Quality and Continuity of Care*****:** High attrition rates can lead to a decrease in the quality and continuity of patient care. Experienced staff plays a crucial role in maintaining high care standards, and their departure can create knowledge gaps and inconsistency in patient care.
    - ***Staffing Shortages*****:** Persistent staffing shortages can lead to increased workloads for remaining staff, potentially resulting in burnout, further attrition, and a cycle that's hard to break.

3. **Patient Satisfaction**
    - ***Impact on Patient Experience*****:** The turnover of healthcare providers can negatively affect the patient experience. Patients value continuity in their healthcare providers,

and frequent changes can lead to dissatisfaction and a lack of trust in the healthcare system.

- ***Perception of Care*:** High employee turnover can be perceived as a reflection of the organization's internal problems, which can negatively impact the public's perception of the quality of care provided.

Broader Organizational Effects

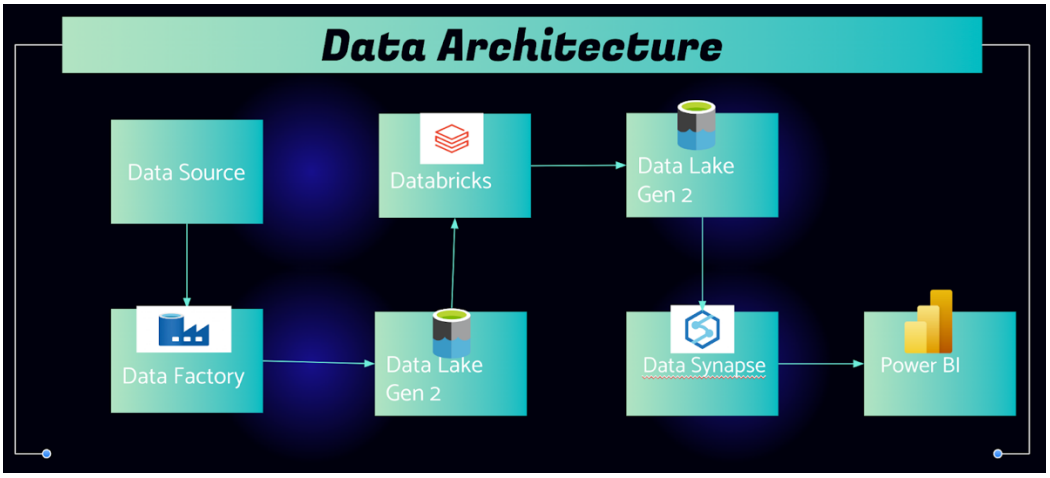
- ***Organizational Culture*:** High turnover can disrupt the culture of a healthcare organization, affecting morale and employee engagement.
- ***Training and Development*:** Continuous attrition requires constant training and development of new staff, diverting resources from other critical areas.

The Solution

To address these challenges, we've implemented a comprehensive solution focused on managing and reducing employee attrition within healthcare organizations. Our approach involves delving into the root causes of turnover, fostering a supportive work culture, and leveraging predictive analytics to proactively identify employees at risk of leaving.

Our solution comprises a data pipeline utilizing Azure cloud services. It begins by extracting data from CSV files stored in Azure Blob Storage. Subsequently, we employ Azure Databricks to clean and preprocess this data before storing it in Azure Data Lake Storage (ADLS). From there, we utilize Azure Synapse Analytics for further analysis and to derive meaningful insights. Finally, we harness Microsoft Power BI to visualize these insights and create intuitive dashboards that provide actionable information for decision-makers.

Architecture Diagram



Details on Azure Services

Azure Data Factory:

We have used Azure Data Factory to ingest the CSV file from Kaggle and load it into Azure Blob Storage. Azure Data Factory is a managed data integration service that makes it easy to move data between different sources and destinations. In our project, Azure Data Factory serves as the backbone for data ingestion and initial processing. By automating the data pipeline, it ensures that our data is consistently up-to-date, reliable, and ready for analysis pipeline, particularly for feeding into our machine learning model for employee attrition prediction.

Azure Databricks:

Azure Databricks will be used to transform the raw data into a clean and usable format for analysis. In our project, Azure Databricks plays a crucial role in transforming raw data into a format that is clean, structured, and suitable for analysis. This transformation process is essential for ensuring the accuracy and reliability of our predictive analytics.

Data Cleaning Processes on Databricks

1. ***Identify and Remove Duplicate Records:*** Databricks allows us to scan the dataset for duplicate entries, which can skew analysis results. These duplicates are efficiently identified and removed to maintain data integrity. So, in this we
2. ***Handling Missing Values:*** The platform provides tools to detect missing values within the dataset. Depending on the nature and context of the data, these missing values can either be filled with appropriate substitutes or the records can be excluded from the analysis.
3. ***Standardizing Data Formats:*** Data from various sources often comes in different formats. Azure Databricks enables the standardization of these diverse data formats into a uniform structure, facilitating easier analysis and integration with other data.
4. ***Validating Data Quality:*** The platform is also instrumental in validating the quality of data. This involves checking for inaccuracies or inconsistencies in the dataset and correcting them as necessary.

DATA LAKE GEN2 (AZURE LAKE STORAGE GEN2)

Azure Data Lake Storage is used to store the raw and transformed data that we cleaned using Data bricks. Hence it offers massively scalable and secure data storage for big data analytics.
Data Lake Gen2 enhances capabilities and is optimized for analytics workloads. It allows you to perform analytics on the data in place without the need to move it, which can save time and

resources. It supports access control at the file and folder level and integrates seamlessly with various analytics frameworks.

AZURE SYNAPSE ANALYTICS:

Azure Synapse Analytics will be used to load the clean data and analyze it using SQL. Azure Synapse Analytics is a limitless analytics service provided by Microsoft Azure that brings together enterprise data warehousing and Big Data analytics. It offers a unified experience to ingest, prepare, manage, and serve data for immediate business intelligence and machine learning needs.

Loading Clean Data

- ***Integration with Azure Ecosystem:*** After the data is cleaned and transformed using Azure Databricks, it is loaded into Azure Synapse Analytics. This service seamlessly integrates with Azure Databricks and other Azure storage solutions, ensuring a smooth data transfer process.
- ***Data Warehousing:*** Azure Synapse provides a highly scalable and secure data warehousing solution, allowing for the storage of large volumes of structured and semi-structured data.

SQL-Based Analysis

- ***SQL Pool:*** Azure Synapse features a SQL pool (formerly SQL Data Warehouse), which is a distributed query system optimized for large-scale data warehousing. The SQL pool allows for running complex queries on large datasets efficiently.
- ***Analytical Capabilities:*** Users can perform a range of analytical tasks including aggregations, joins, and window functions using familiar SQL language, making it accessible to those with SQL expertise.
- ***Real-Time Insights:*** The service enables real-time analytics, allowing for the rapid processing and analysis of data, crucial for timely decision-making.

MICROSOFT POWER BI

Microsoft Power BI is a powerful business analytics service provided by Microsoft. It enables users to visualize data and share insights across an organization or embed them in an app or website. Power BI is known for its user-friendly interface, robust data connectivity, and advanced data visualization capabilities. In our project, Microsoft Power BI is employed to transform the analyzed data into visual representations, making it easier to understand and communicate the insights derived, especially in the context of employee attrition.

Visualization and Reporting

- ***Data Connectivity:*** Power BI seamlessly connects with Azure Synapse Analytics, enabling the direct use of processed and analyzed data.

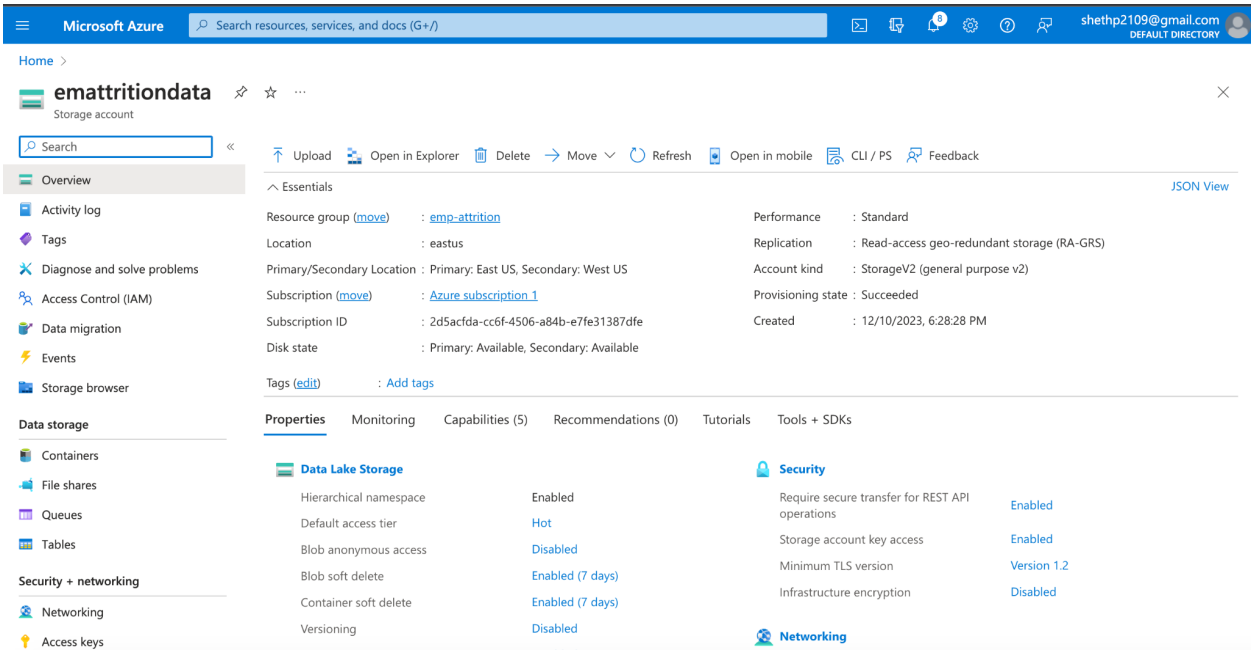
- ***Interactive Dashboards:*** It allows for the creation of interactive dashboards and reports. These visualizations help in identifying trends, patterns, and anomalies in the data related to employee attrition.
- ***Custom Visualizations:*** The service offers a wide range of visualization tools - from basic charts and graphs to complex data plots. These can be customized to suit specific needs, providing a clearer understanding of the underlying data.

Generating Insights

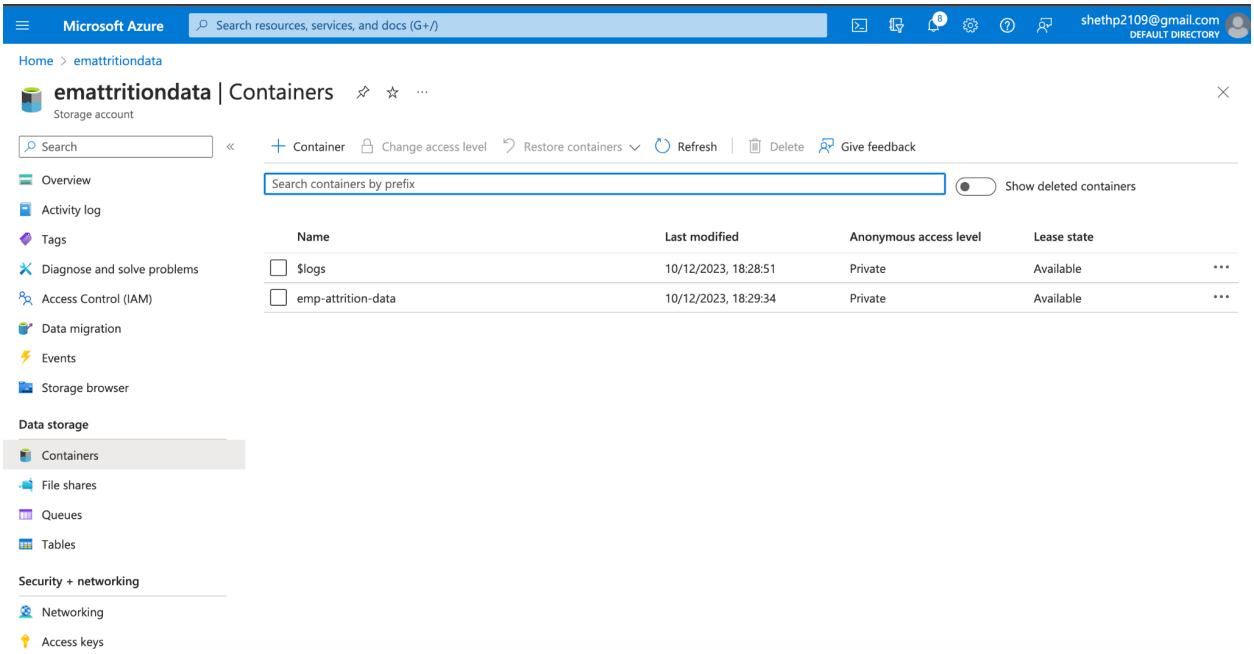
- ***Data Exploration:*** Users can explore the data in various ways, drilling down into specifics or viewing the broader trends, which is crucial for understanding the factors contributing to employee attrition.
- ***Shareable Reports:*** Reports and dashboards created in Power BI can be easily shared with stakeholders, providing them with valuable insights into employee behavior and attrition risks.
- ***Real-Time Analytics:*** Power BI supports real-time analytics, enabling HR and management teams to make timely, data-driven decisions.

STEPS FOR BUILDING THE ADF PIPELINE

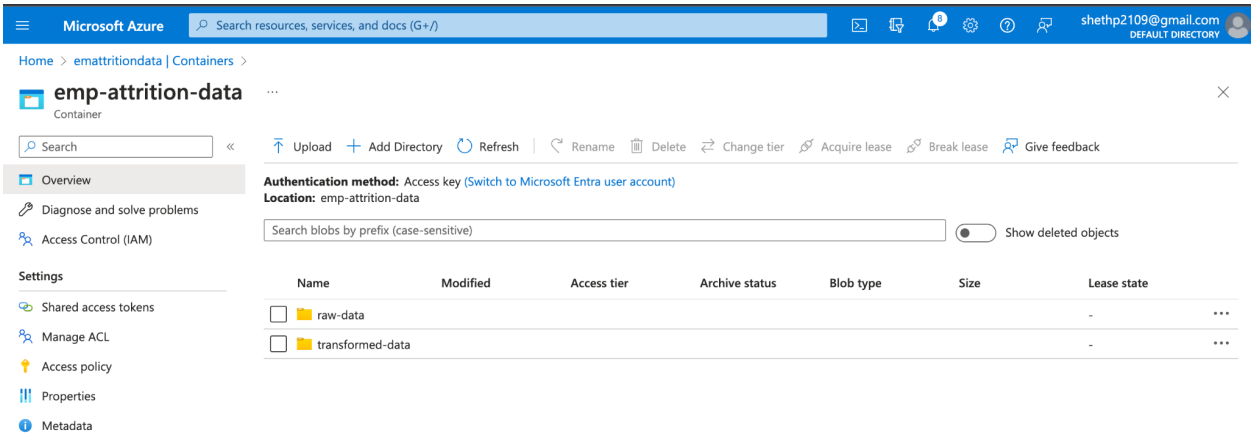
1. Created storage account emattritiondata



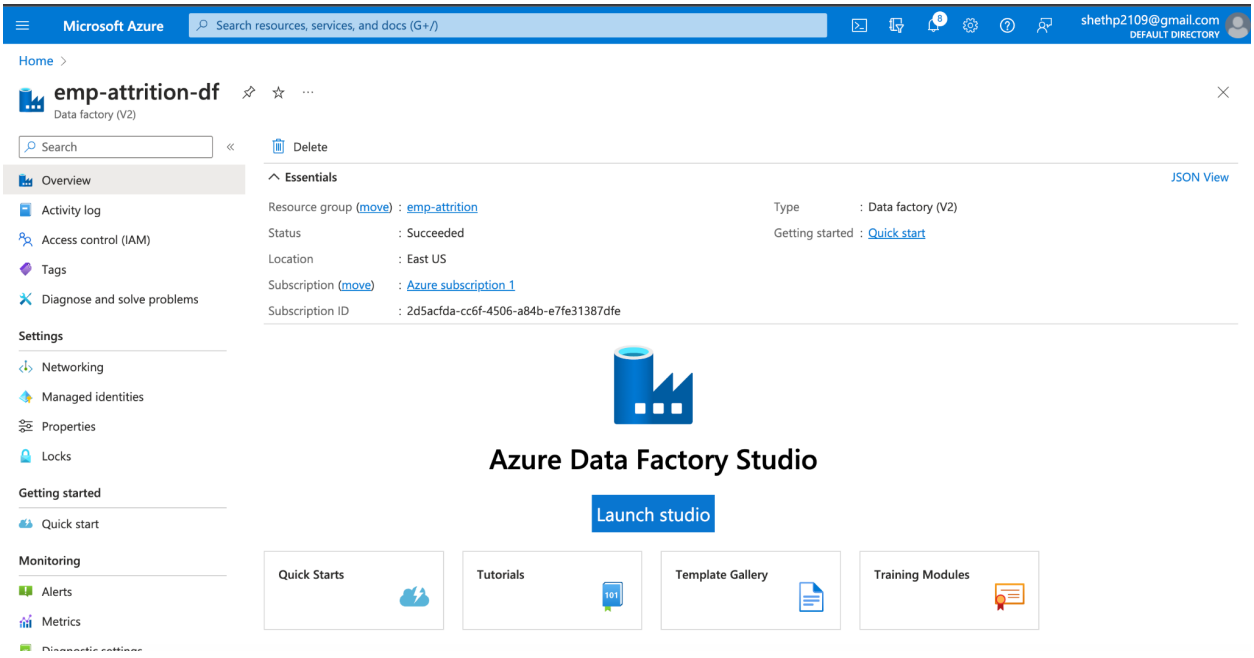
2. Created container named emp-attrition-data



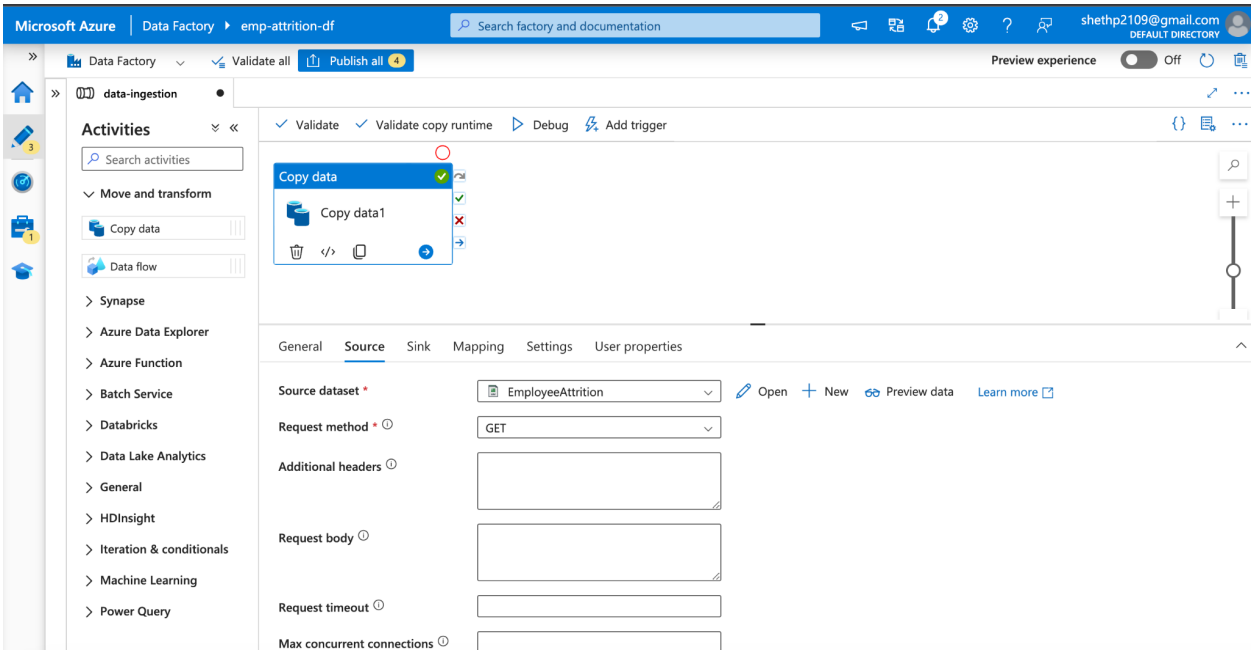
3. Created 2 folders raw-data and transformed-data



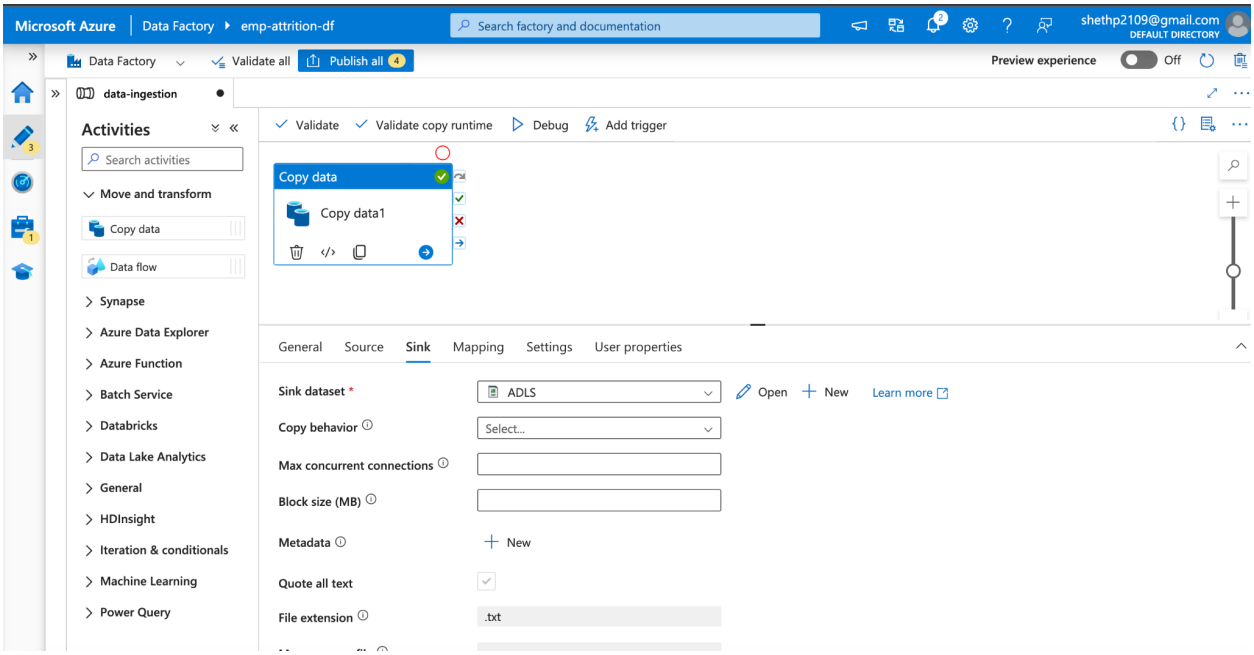
4. Created Azure Data Factory named emp-attrition-df



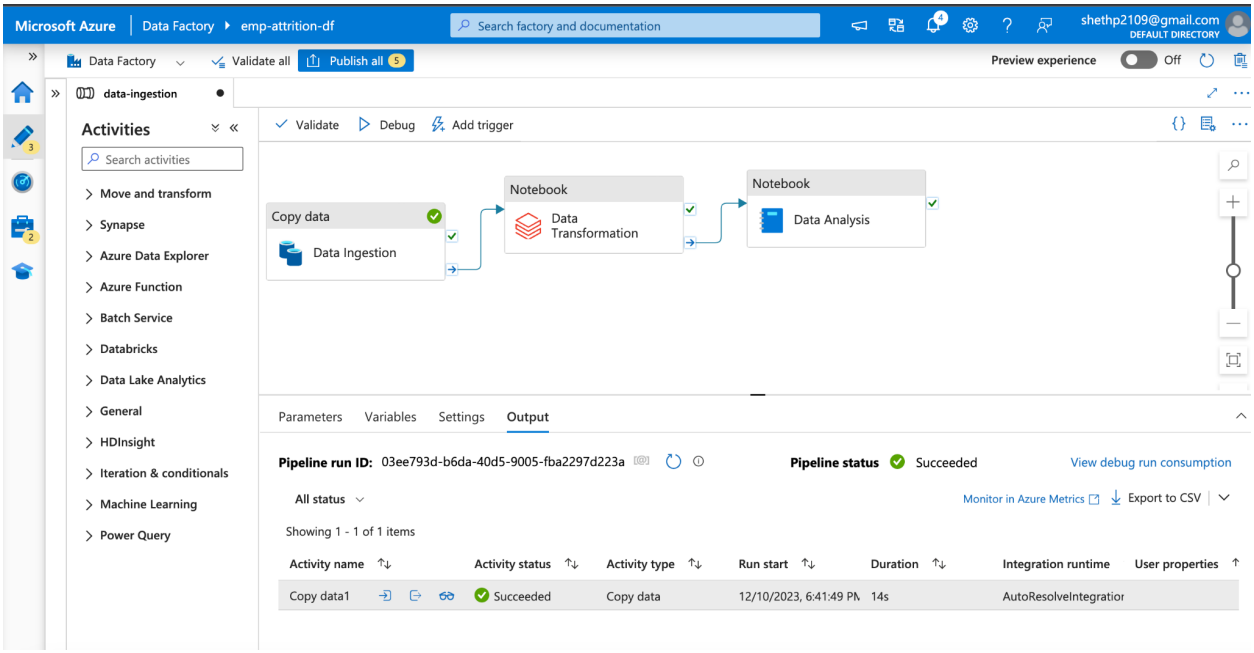
5. Added Copy data activity with source as blob storage.



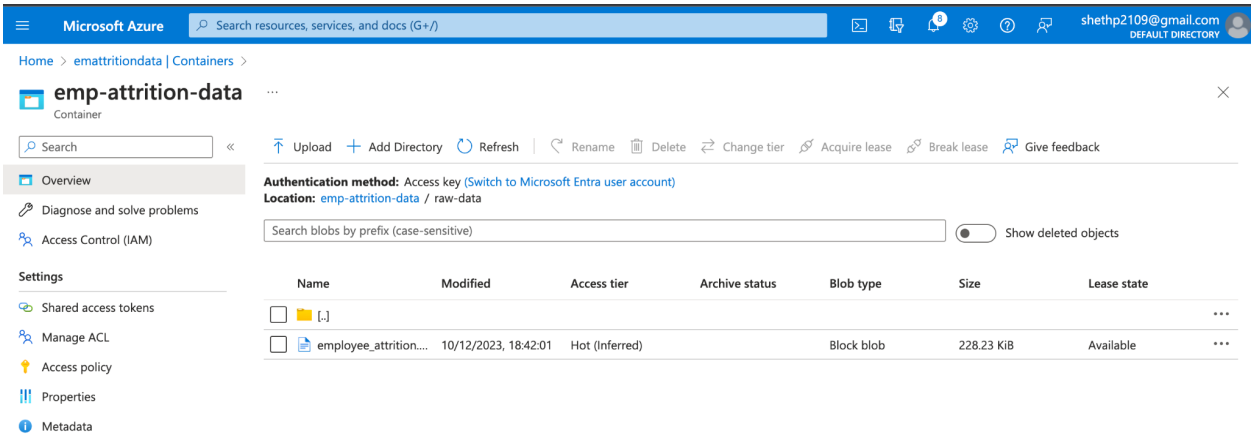
6. Added Copy data activity with sink as ADLS



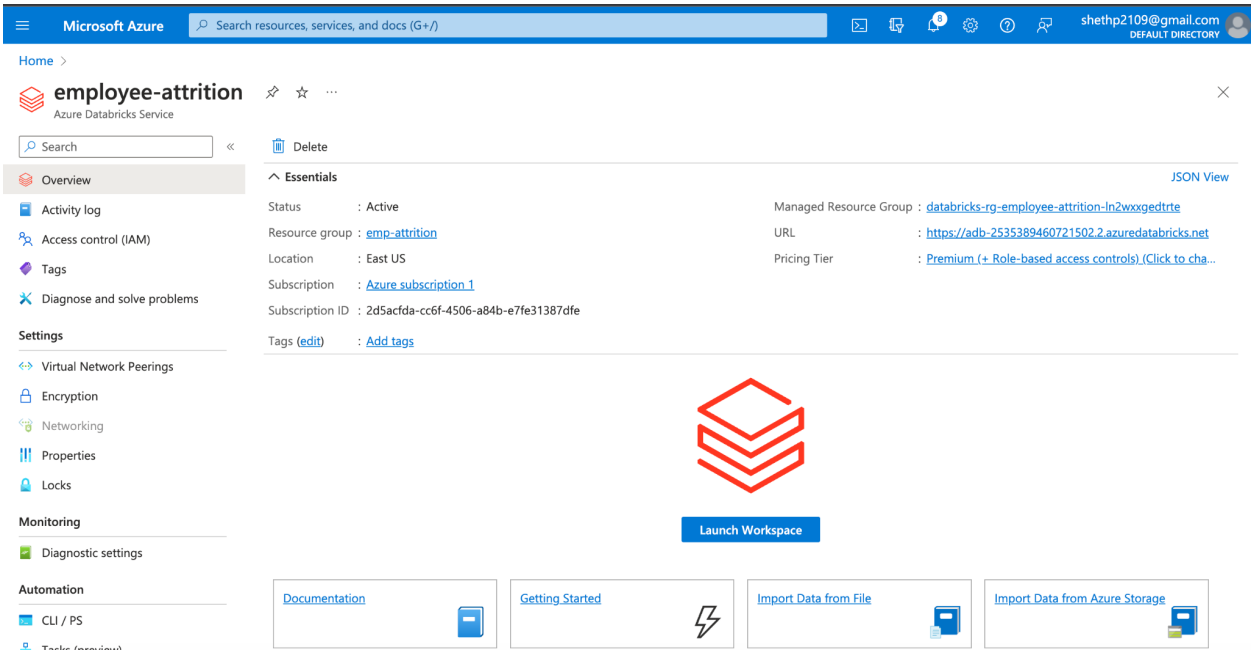
7. Added data bricks notebook in the pipeline to transform and clean data



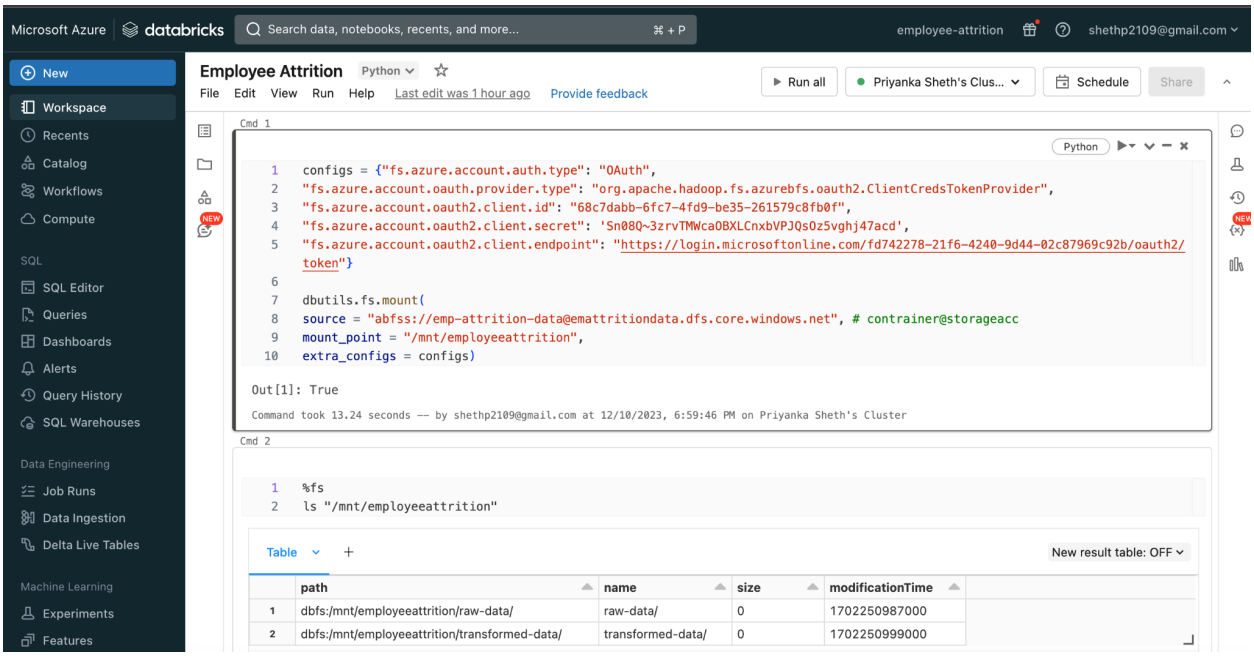
8. Csv file stored in blob storage



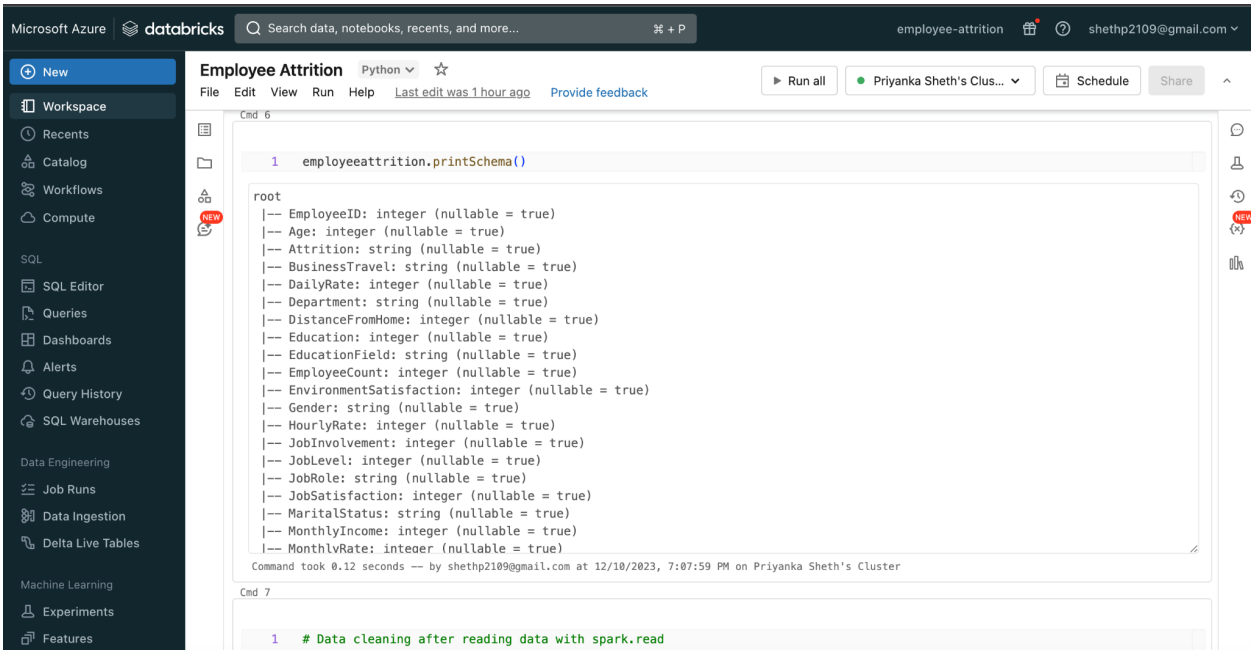
9. Created data bricks service as employee-attrition

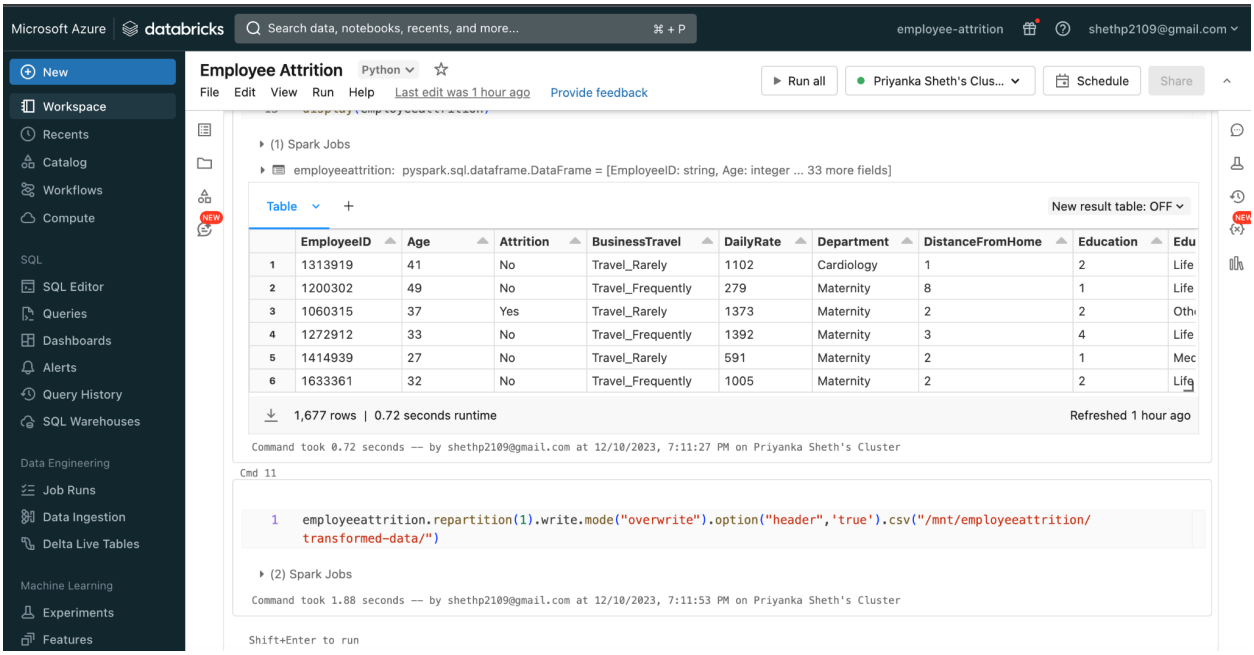


10. Code to connect databricks to blob storage

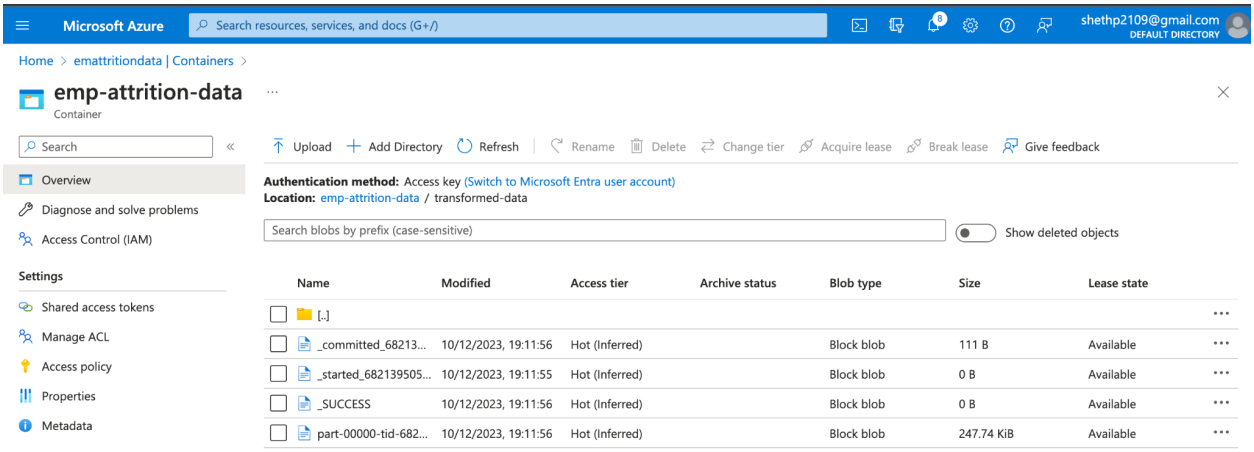


11. Reading and cleaning data.

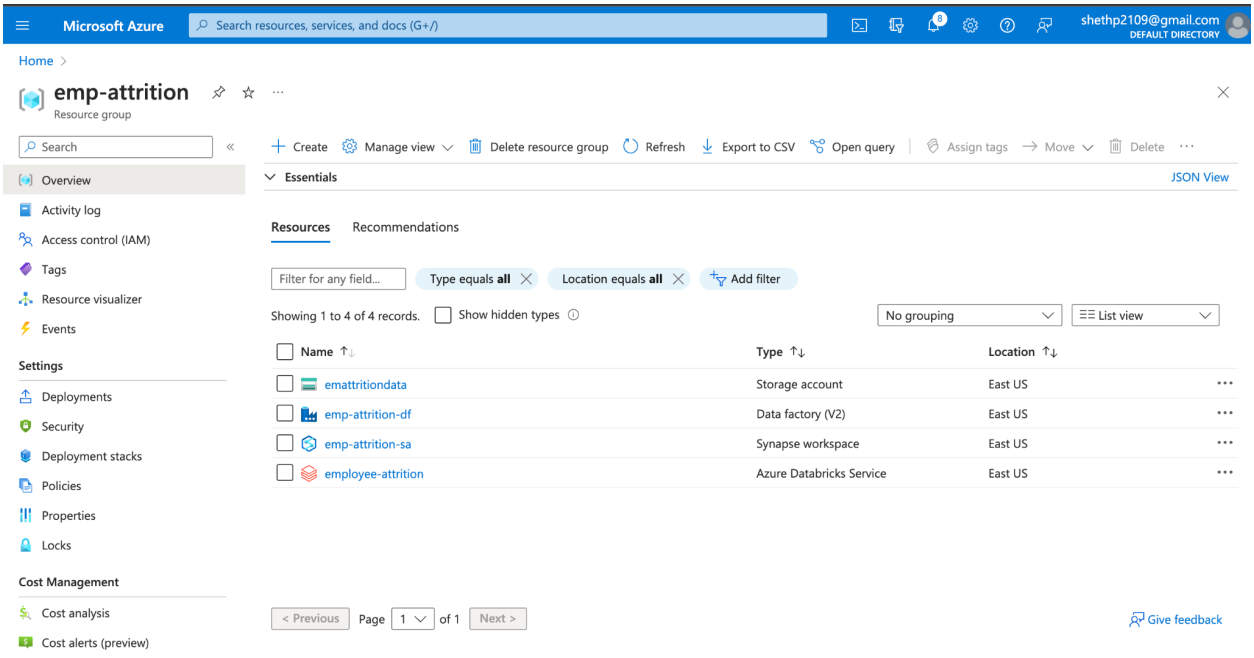




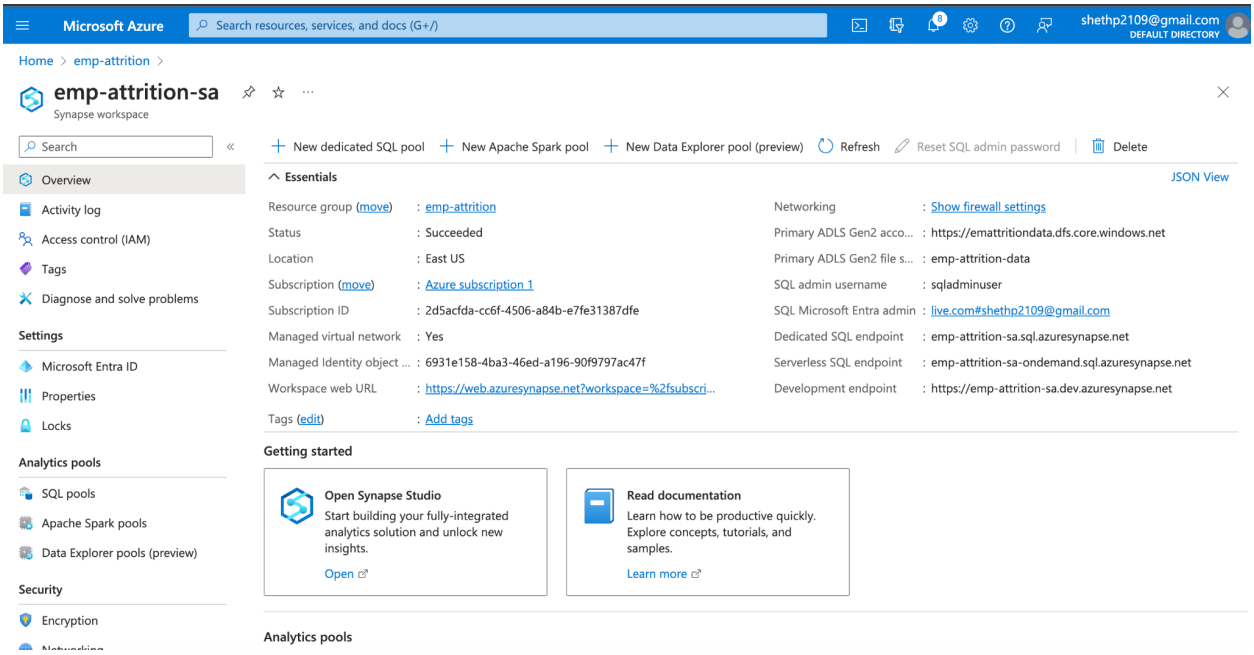
12. Writing cleaned data to transformed folder in storage account.



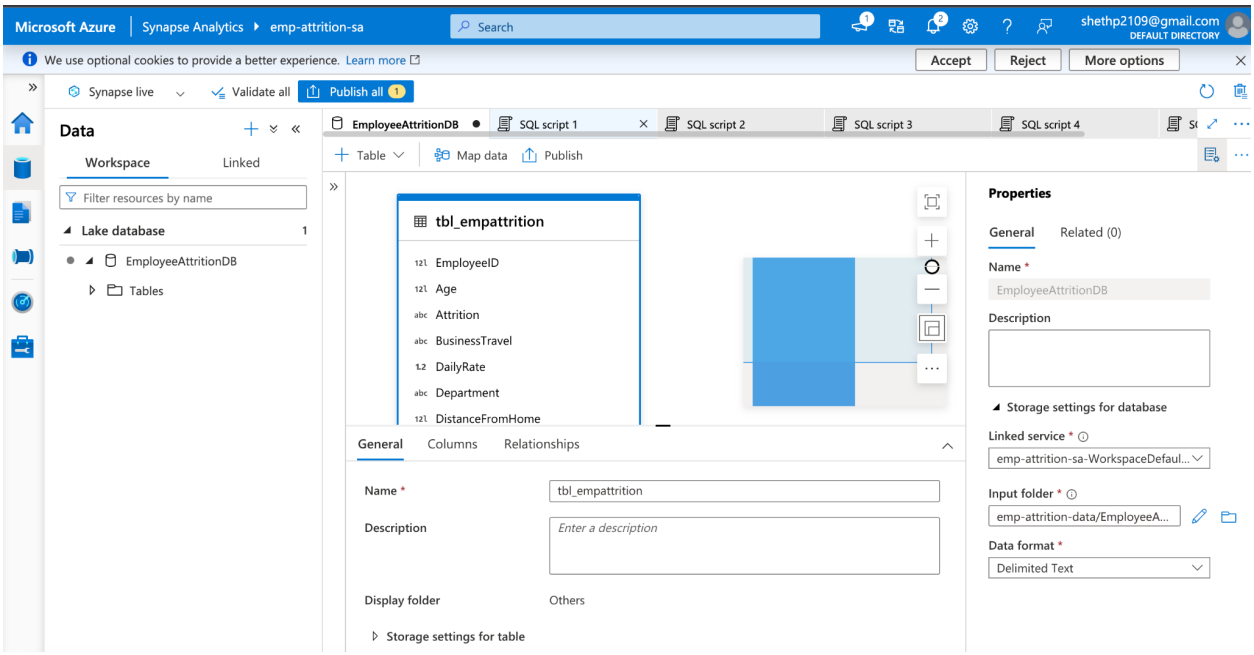
13. List of resource group.



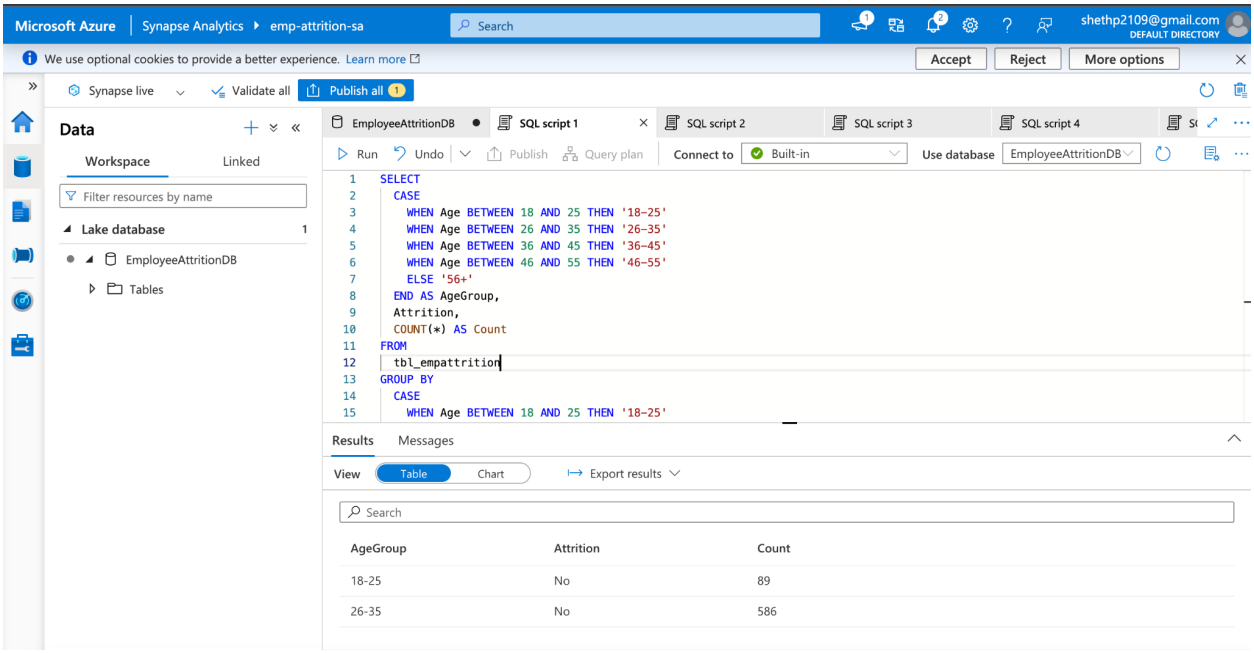
14. Created Azure Synapse workspace.

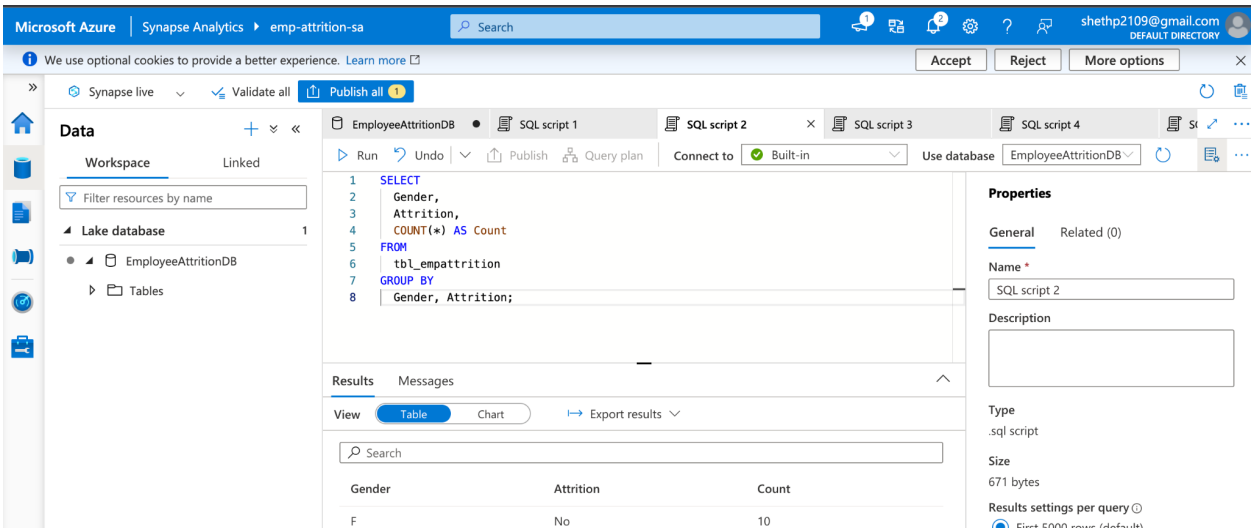


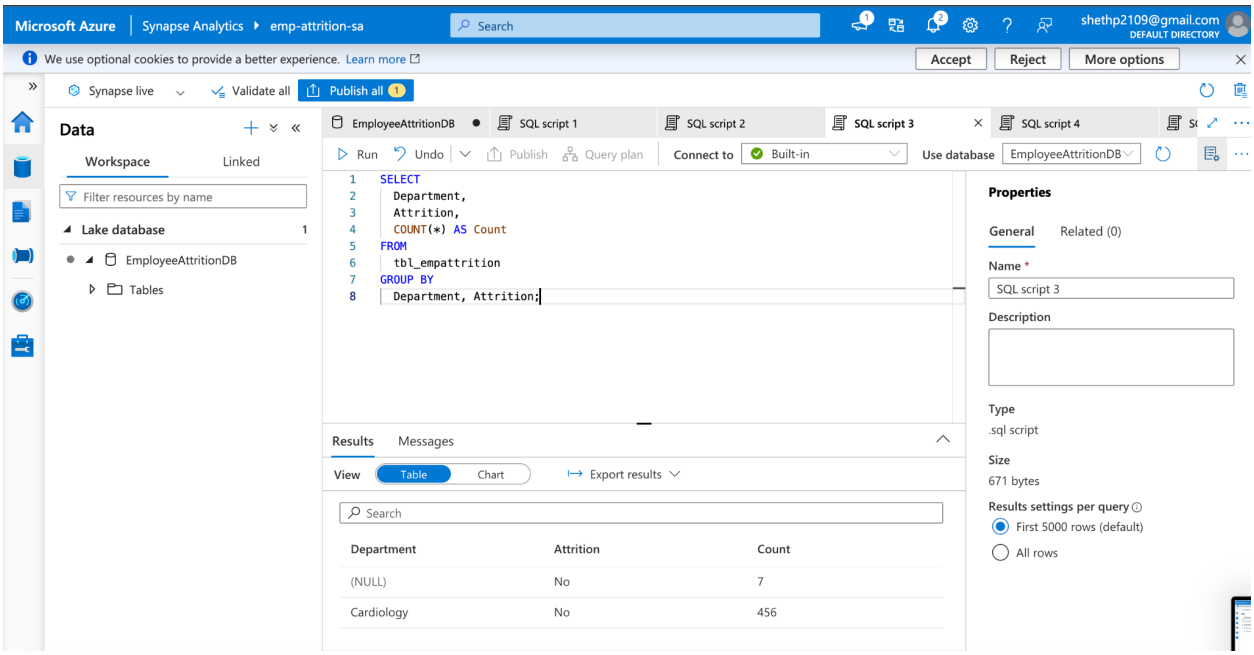
15. Created EmployeeAttritionDB in synapse.

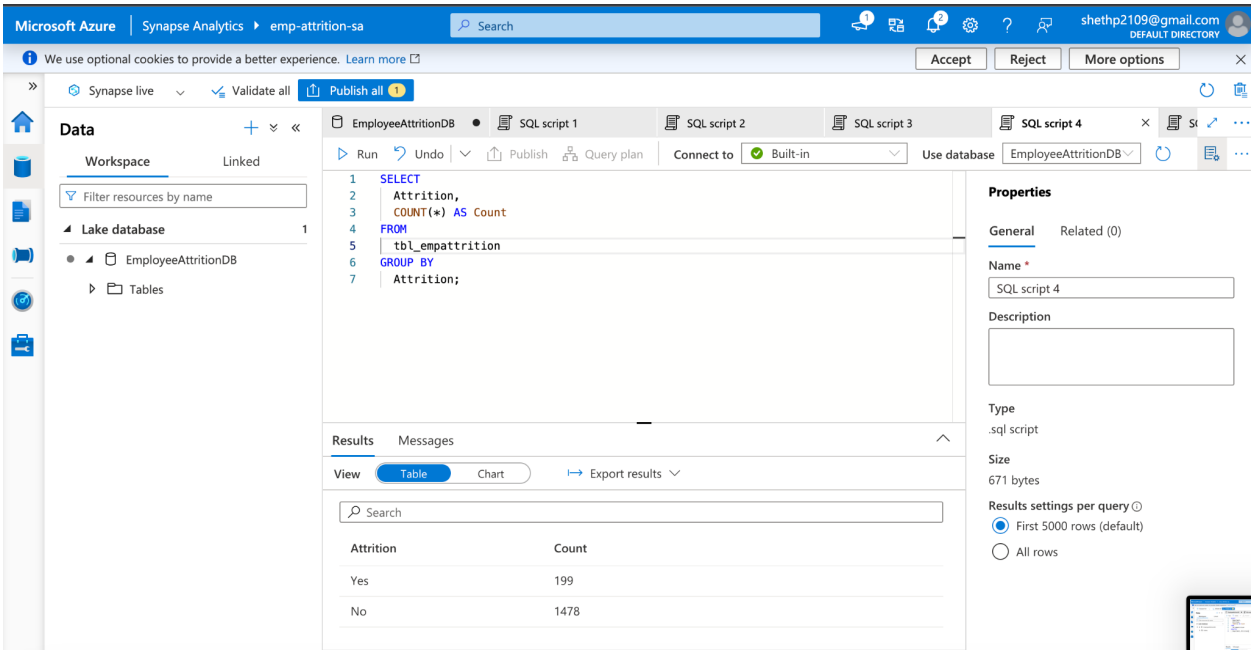


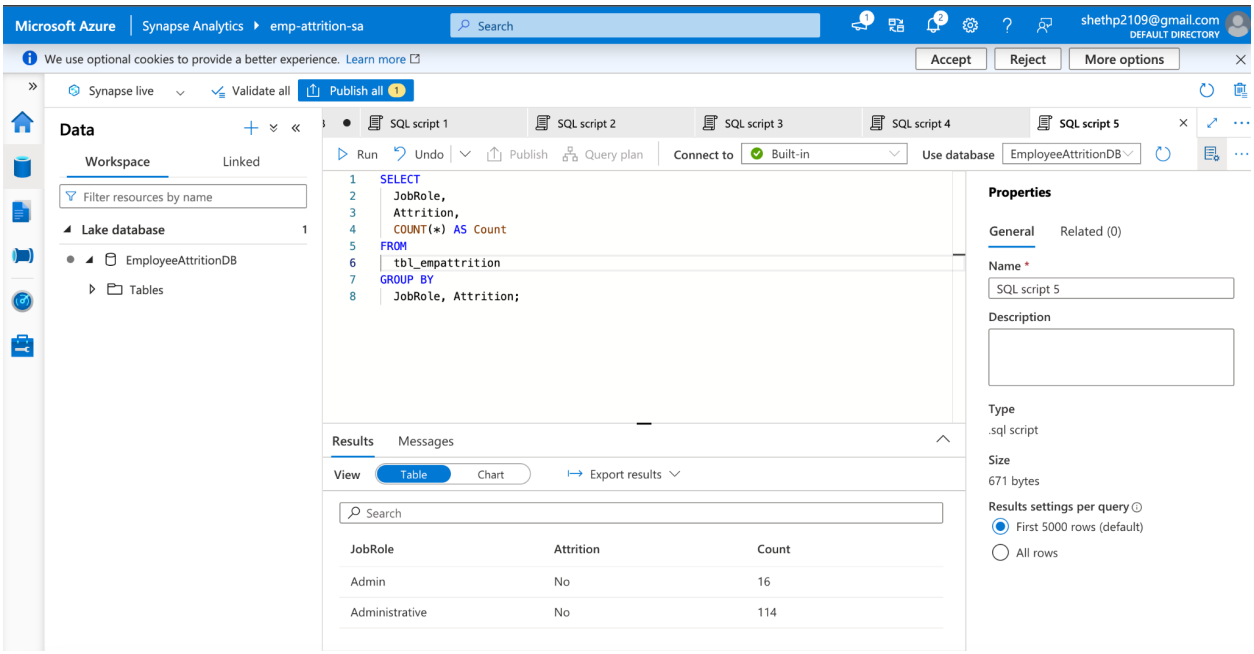
16. Writing SQL scripts to fetch data.



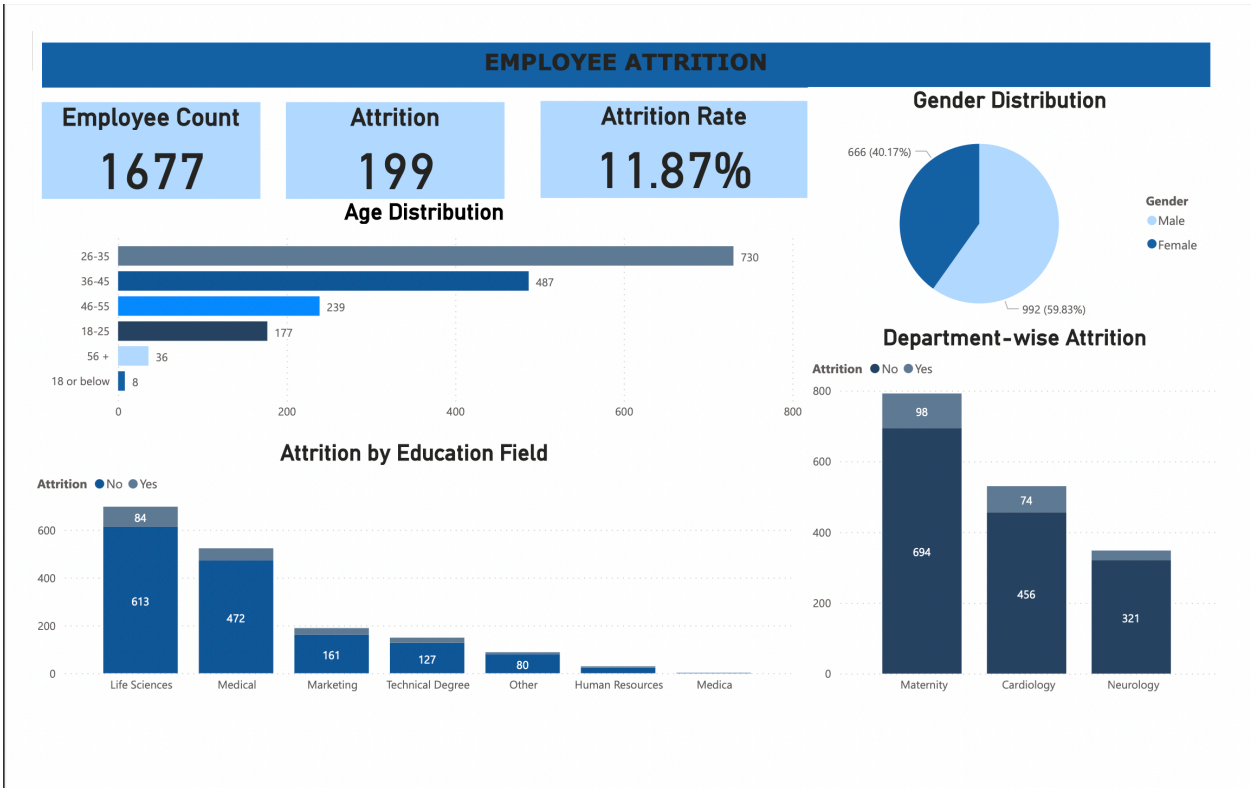



Microsoft Azure
Synapse Analytics
emp-attrition-sa
Search
shethp2109@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 1
Data
Workspace
Linked
Filter resources by name
Lake database 1
EmployeeAttritionDB
Tables
EmployeeAttritionDB
SQL script 1
SQL script 2
SQL script 3
SQL script 4
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
EmployeeAttritionDB
SELECT
Department,
Attrition,
COUNT(*) AS Count
FROM
tbl_empattrition
GROUP BY
Department, Attrition;
Properties
General
Related (0)
Name *
SQL script 3
Description
Type
.sql script
Size
671 bytes
Results settings per query
First 5000 rows (default)
All rows
Results
Messages
View
Table
Chart
Export results
Search
Department
Attrition
Count
(NULL)
No
7
Cardiology
No
456


Microsoft Azure
Synapse Analytics
emp-attrition-sa
Search
shethp2109@gmail.com
DEFAULT DIRECTORY
We use optional cookies to provide a better experience. Learn more
Accept
Reject
More options
Synapse live
Validate all
Publish all 1
Data
Workspace
Linked
Filter resources by name
Lake database 1
EmployeeAttritionDB
Tables
EmployeeAttritionDB
SQL script 1
SQL script 2
SQL script 3
SQL script 4
Run
Undo
Publish
Query plan
Connect to
Built-in
Use database
EmployeeAttritionDB
SELECT
Attrition,
COUNT(*) AS Count
FROM
tbl_empattrition
GROUP BY
Attrition;
Properties
General
Related (0)
Name *
SQL script 4
Description
Type
.sql script
Size
671 bytes
Results settings per query
First 5000 rows (default)
All rows
Results
Messages
View
Table
Chart
Export results
Search
Attrition
Count
Yes
199
No
1478



17. Final PowerBI dashboard development.

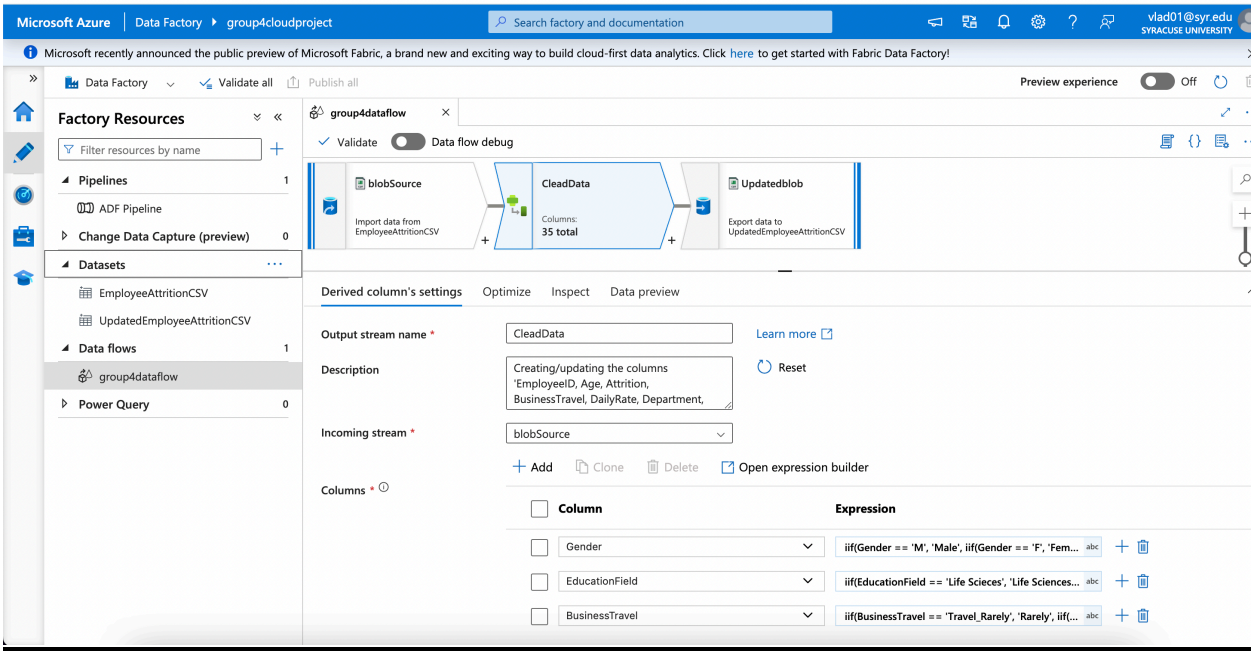


Approach Two: Data Cleaning Process Using ADF Data Flow

Our data refinement process revolves around **Azure Data Factory's Data Flow,** ensuring the information used for predicting employee attrition is robust and reliable. Initially, we source raw data from Azure Blob Storage, encompassing diverse facets such as employee demographics, job satisfaction, and performance metrics.

Within the Data Flow structure, a systematic sequence of steps is employed to cleanse and enhance the data quality. This involves eliminating duplicates, rectifying missing values, and standardizing formats to ensure consistency. Additionally, we implement feature engineering to derive new insights from the data. This structured approach guarantees that our data undergoes a comprehensive cleaning regimen before proceeding.

The refined data then seamlessly progresses through the Data Flow stages, culminating in its transition to Azure Synapse Analytics. This platform serves as our analytical hub, where the thoroughly cleansed data becomes the bedrock for in-depth exploration and analysis. This meticulous cleaning process lays the groundwork for subsequent analyses, empowering us to extract actionable insights crucial for understanding and mitigating employee attrition in healthcare settings.



CHALLENGES FACED

Error while creating Databricks cluster.

We encountered an error during the creation of a Databricks cluster. The process failed and did not complete as expected. The inability to create a cluster prevents the execution of data processing jobs, which impacts our data analytics workflows. Diagnostic Information: Attempted to create a cluster. The error message received Preliminary checks were performed on network settings and subscription limits. These might be the insufficient permissions or incorrect role assignments. Network configuration issues preventing communication with other Azure resources. Exceeding quota limits for the subscription or resource group.

Error while deploying the triggers

There was a problem with setting up some automated tasks in Azure Functions, just like the trouble we had before with setting up a data processing group in Databricks. This problem was causing some delays because it stops these tasks from working right. Last time, we made sure the internet connection was okay and that we hadn't used too much of our allowed resources. We think the issue might be because we don't have the right permissions, there's a problem with how the network is set up, or we've used more resources than we're allowed to.

Access Errors while creating Linked service for data bricks

During the setup of a Linked Service in Azure Data Factory for Databricks, access errors were encountered with Azure Key Vault. There were issues with permissions or access policies that are preventing the successful linkage of the Key Vault secrets to the Databricks workspace. The technical steps to resolve this would involve verifying the access policies within Azure Key Vault to ensure that the Data Factory service principal has the necessary permissions to read secrets.

CONCLUSION

The integration of Azure services—Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and Microsoft Power BI—has been integral in our data processing journey, imparting significant insights and learnings. Azure Data Factory streamlined the movement of data from diverse sources, notably from Kaggle, enabling efficient ingestion into Azure Blob Storage. This storage architecture, structured with dedicated folders like 'Employee Attrition', facilitated secure and organized storage for our critical CSV file.

Azure Databricks emerged as the linchpin for data transformation, unraveling the complexities within our datasets. Its functionalities empowered us to address duplicate records, handle missing values, and standardize data formats, ensuring accuracy for subsequent analysis. Transitioning to Azure Synapse Analytics, our exploration into SQL-based analysis was eye-opening. The scalable SQL pool and analytical prowess enabled us to derive profound insights, unveiling trends and patterns related to employee attrition, ultimately aiding informed decision-making.

The visualization prowess of Microsoft Power BI breathed life into our analyzed data, converting it into interactive dashboards and reports. This transformation allowed for a nuanced exploration of employee attrition trends, enhancing our understanding and empowering actionable decision-making.

Beyond the technical aspects, this journey has been a rich learning experience. It illuminated the significance of structured data handling, the power of scalable analytics tools, and the transformation of raw data into actionable insights. This holistic understanding of leveraging Azure services for data-driven decision-making, specifically in the context of managing employee attrition, has been invaluable for our team’s growth and future endeavors.