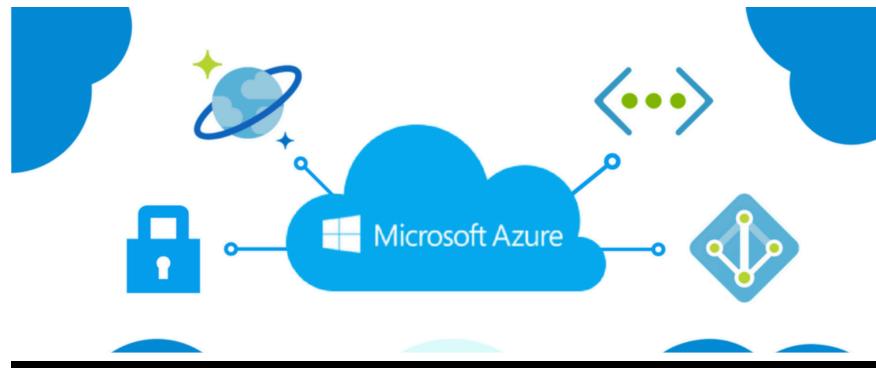


**IST 615**

## **Cloud Management**



## **Final Project Report**

**Team 4**

**Submitted by:**

Diya Shah  
Priyanka Sheth  
Sampada Regmi  
Vrushali Lad

## Table of Contents

<b><i>PROJECT DESCRIPTION .....</i></b>	<b>3</b>
<b><i>ARCHITECTURE DIAGRAM.....</i></b>	<b>4</b>
<b><i>DETAILS ON AZURE SERVICES.....</i></b>	<b>5</b>
<b>AZURE DATA FACTORY: .....</b>	<b>5</b>
<b>AZURE DATABRICKS:.....</b>	<b>5</b>
<b>AZURE SYNAPSE ANALYTICS: .....</b>	<b>6</b>
<b>Loading Clean Data .....</b>	<b>6</b>
<b>SQL-Based Analysis .....</b>	<b>6</b>
<b>MICROSOFT POWER BI .....</b>	<b>6</b>
<b>Visualization and Reporting.....</b>	<b>6</b>
<b>Generating Insights .....</b>	<b>7</b>
<b><i>STEPS FOR BUILDING THE ADF PIPELINE .....</i></b>	<b>8</b>
<b><i>APPROACH TWO: DATA CLEANING PROCESS USING ADF DATA FLOW .....</i></b>	<b>19</b>
<b><i>CHALLENGES FACED .....</i></b>	<b>20</b>
<b><i>CONCLUSION.....</i></b>	<b>21</b>

## 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

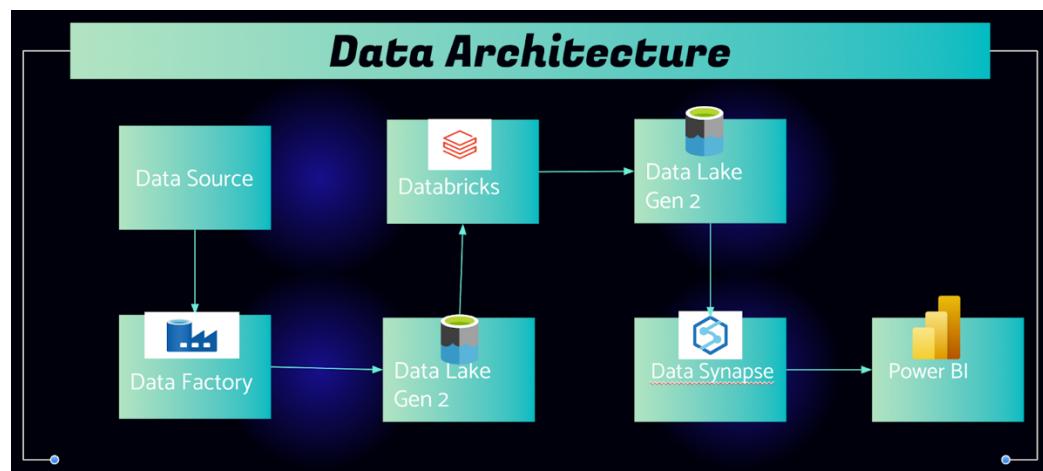
- **Organizational Culture:** High turnover can disrupt the culture of a healthcare organization, affecting morale and employee engagement.
- **Training and Development:** Continuous attrition requires constant training and development of new staff, diverting resources from other critical areas.

## 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

The screenshot shows the Microsoft Azure Storage account 'emattritiondata' overview page. The account was created on 12/10/2023 at 6:28:28 PM. It is located in the eastus region and is part of the 'emp-attrition' resource group. The primary/secondary location is Primary: East US, Secondary: West US. The subscription is 'Azure subscription 1'. The storage kind is StorageV2 (general purpose v2) with standard performance and read-access geo-redundant storage (RA-GRS) replication. The account is provisioned successfully. The Data Lake Storage section shows that hierarchical namespace is enabled, default access tier is Hot, blob anonymous access is disabled, blob soft delete is enabled (7 days), container soft delete is enabled (7 days), and versioning is disabled. The Security section shows that require secure transfer for REST API operations is enabled, storage account key access is enabled, minimum TLS version is Version 1.2, and infrastructure encryption is disabled. The Networking section is shown below.

### 2. Created container named emp-attrition-data

The screenshot shows the Microsoft Azure Storage account 'emattritiondata' containers page. It lists two containers: '\$logs' and 'emp-attrition-data'. Both containers were created on 10/12/2023 at 18:28:51 and 18:29:34 respectively. They both have private anonymous access levels and are available. The '\$logs' container has three dots next to it, while the 'emp-attrition-data' container has four dots. The left sidebar shows the 'Containers' option under 'Data storage' is selected. Other options like 'File shares', 'Queues', and 'Tables' are also listed.

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

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar has a tree view with 'Home > emattritiondata | Containers > emp-attrition-data'. The main area shows a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
raw-data					-	---
transformed-data					-	---

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

The screenshot shows the Microsoft Azure Data Factory Studio interface. The left sidebar has a tree view with 'Home > emp-attrition-df'. The main area shows the 'Essentials' section with the following details:

Resource group (move)	: emp-attrition	Type	: Data factory (V2)
Status	: Succeeded	Getting started	: <a href="#">Quick start</a>
Location	: East US		
Subscription (move)	: <a href="#">Azure subscription 1</a>		
Subscription ID	: 2d5acfda-cc6f-4506-a84b-e7fe31387dfe		

Below the essentials, there is a large blue icon of a factory building, followed by the text 'Azure Data Factory Studio' and a 'Launch studio' button. At the bottom, there are four cards: 'Quick Starts' (cloud icon), 'Tutorials' (book icon), 'Template Gallery' (document icon), and 'Training Modules' (certificate icon).

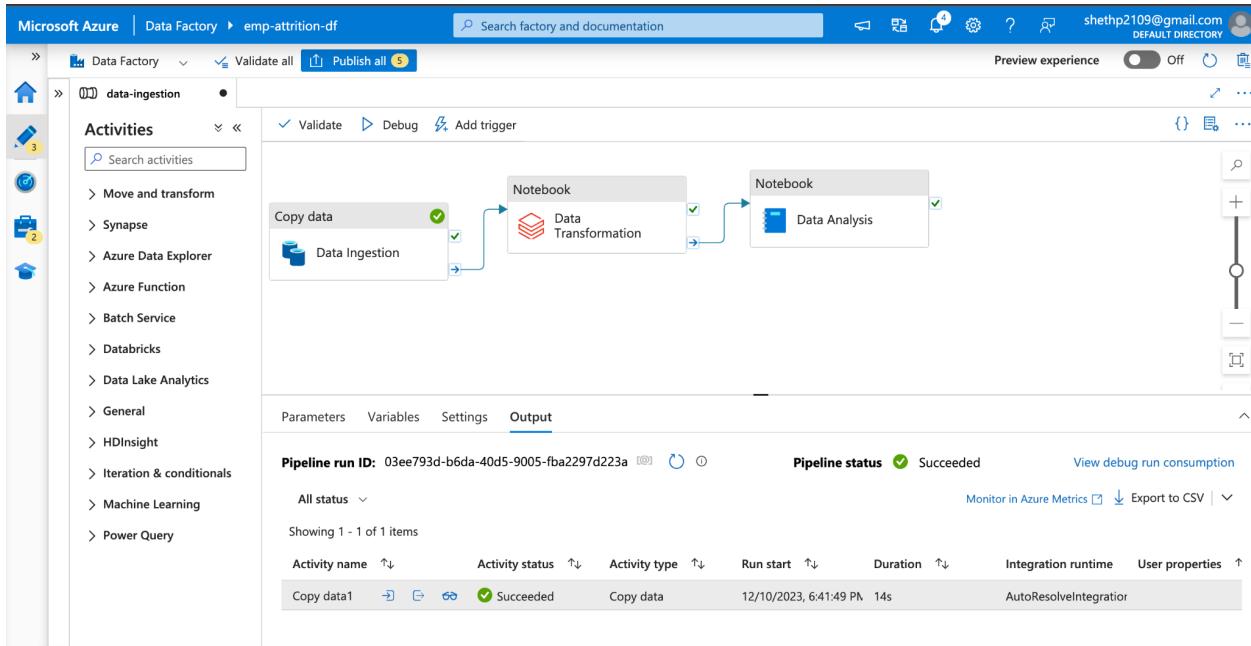
## 5. Added Copy data activity with source as blob storage.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, there's a navigation sidebar with icons for Home, Data Flow, Copy Data, Data Explorer, Synapse, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The main area displays a pipeline named 'data-ingestion'. A 'Copy data' activity is selected, indicated by a green checkmark icon. The 'Source' tab is active, showing the configuration for the source dataset 'EmployeeAttrition' (selected via a dropdown menu), request method 'GET', and other parameters like Additional headers, Request body, Request timeout, and Max concurrent connections. The 'Sink' tab is also visible but not selected.

## 6. Added Copy data activity with sink as ADLS

This screenshot is similar to the previous one but focuses on the 'Sink' tab of the 'Copy data' activity configuration. The 'Sink' tab is now active, showing the configuration for the sink dataset 'ADLS' (selected via a dropdown menu). Other sink-related settings include 'Copy behavior' (set to 'Select...'), 'Max concurrent connections', 'Block size (MB)', 'Metadata' (with a '+ New' button), 'Quote all text' (checkbox checked), and 'File extension' (set to '.txt'). The 'Source' tab is visible but not selected.

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



## 8. Csv file stored in blob storage

The screenshot shows the Microsoft Azure Blob Storage interface for a container named 'emp-attrition-data'. The container has one blob named 'employee\_attrition...'. The blob details show it was modified on 10/12/2023, 18:42:01, has a size of 228.23 KiB, and is a Block blob.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
employee_attrition...	10/12/2023, 18:42:01	Hot (Inferred)		Block blob	228.23 KiB	Available

## 9. Created data bricks service as employee-attrition

The screenshot shows the Microsoft Azure portal interface for the 'employee-attrition' Databricks Service. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (shethp2109@gmail.com). The main content area displays the service's details under the 'Essentials' tab. Key information includes:

- Status: Active
- Resource group: [emp-attrition](#)
- Location: East US
- Subscription: [Azure subscription 1](#)
- Subscription ID: 2d5acfda-cc6f-4506-a84b-e7fe31387dfe
- Managed Resource Group: [databricks-rg-employee-attrition-In2wxgedtre](#)
- URL: <https://adb-2535389460721502.2.azure.databricks.net>
- Pricing Tier: Premium (+ Role-based access controls) [\(Click to change\)](#)

On the left sidebar, there are sections for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Virtual Network Peering, Encryption, Networking, Properties, Locks), Monitoring (Diagnostic settings), and Automation (CLI / PS). Below the essentials section, there are links for Documentation, Getting Started, Import Data from File, and Import Data from Azure Storage. A large red 'Databricks' logo is centered below the essentials section, and a 'Launch Workspace' button is available.

## 10. Code to connect databricks to blob storage

The screenshot shows the Databricks workspace for the 'Employee Attrition' cluster. The left sidebar lists various notebooks, workspace items, and data engineering tasks. The main area displays a Python notebook with two command cells. The first cell contains the following code:

```
1 configs = {"fs.azure.account.auth.type": "OAuth",
2 "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3 "fs.azure.account.oauth2.client.id": "68c7dabb-6fc7-4fd9-be35-261579c8fb0f",
4 "fs.azure.account.oauth2.client.secret": "Sn080-3zrvTMca08XLNxbvPQ0s025vhj47acd",
5 "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/fd742278-21f6-4240-9d44-02c87969c92b/oauth2/token"}
6
7 dbutils.fs.mount(
8 source = "abfss://emp-attrition-data@emattritiondata.dfs.core.windows.net", # contrainer@storageacc
9 mount_point = "/mnt/employeeattrition",
10 extra_configs = configs)
```

The output of the first cell is 'Out[1]: True'. The second cell contains the command '%fs' followed by 'ls "/mnt/employeeattrition"'. The output of the second cell is a table showing the contents of the mounted blob storage:

path	name	size	modificationTime
1 dbfs:/mnt/employeeattrition/raw-data/	raw-data/	0	1702250987000
2 dbfs:/mnt/employeeattrition/transformed-data/	transformed-data/	0	1702250999000

## 11. Reading and cleaning data.

This screenshot shows a Databricks notebook titled "Employee Attrition" running in Python. The sidebar on the left contains various navigation links such as New, Workspace, Recents, Catalog, Workflows, Compute, SQL, and Machine Learning. The main area displays two command cells:

```
1 employeeattrition.printSchema()  
root  
|-- EmployeeID: integer (nullable = true)  
|-- Age: integer (nullable = true)  
|-- Attrition: string (nullable = true)  
|-- BusinessTravel: string (nullable = true)  
|-- DailyRate: integer (nullable = true)  
|-- Department: string (nullable = true)  
|-- DistanceFromHome: integer (nullable = true)  
|-- Education: integer (nullable = true)  
|-- EducationField: string (nullable = true)  
|-- EmployeeCount: integer (nullable = true)  
|-- EnvironmentSatisfaction: integer (nullable = true)  
|-- Gender: string (nullable = true)  
|-- HourlyRate: integer (nullable = true)  
|-- JobInvolvement: integer (nullable = true)  
|-- JobLevel: integer (nullable = true)  
|-- JobRole: string (nullable = true)  
|-- JobSatisfaction: integer (nullable = true)  
|-- MaritalStatus: string (nullable = true)  
|-- MonthlyIncome: integer (nullable = true)  
|-- MonthlyRate: integer (nullable = true)  
|-- MonthlyWage: integer (nullable = true)  
Command took 0.12 seconds -- by shethp2109@gmail.com at 12/10/2023, 7:07:59 PM on Priyanka Sheth's Cluster
```

```
1 # Data cleaning after reading data with spark.read
```

This screenshot shows the same Databricks notebook after data cleaning. The sidebar and notebook title remain the same. The main area now displays a table view of the cleaned data:

EmployeeID	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Edu
1	41	No	Travel_Rarely	1102	Cardiology	1	2	Life
2	49	No	Travel_Frequently	279	Maternity	8	1	Life
3	37	Yes	Travel_Rarely	1373	Maternity	2	2	Other
4	33	No	Travel_Frequently	1392	Maternity	3	4	Life
5	27	No	Travel_Rarely	591	Maternity	2	1	Mec
6	32	No	Travel_Frequently	1005	Maternity	2	2	Life

Below the table, it says "1,677 rows | 0.72 seconds runtime" and "Refreshed 1 hour ago". The command history shows the repartition and write operations:

```
1 employeeattrition.repartition(1).write.mode("overwrite").option("header", 'true').csv("/mnt/employeeattrition/transformed-data/")  
1 (2) Spark Jobs  
Command took 1.88 seconds -- by shethp2109@gmail.com at 12/10/2023, 7:11:53 PM on Priyanka Sheth's Cluster
```

At the bottom, there is a note: "Shift+Enter to run".

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

The screenshot shows the Microsoft Azure Storage Explorer interface. The left sidebar has a tree view with 'Home' selected, followed by 'emattritiondata | Containers > emp-attrition-data'. The main area shows a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]	10/12/2023, 19:11:56	Hot (Inferred)		Block blob	111 B	Available
_committed_68213...	10/12/2023, 19:11:56	Hot (Inferred)		Block blob	0 B	Available
_started_682139505...	10/12/2023, 19:11:55	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	10/12/2023, 19:11:56	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-682...	10/12/2023, 19:11:56	Hot (Inferred)		Block blob	247.74 KiB	Available

## 13. List of resource group.

The screenshot shows the Microsoft Azure Resource Groups page. The left sidebar has a tree view with 'Home' selected, followed by 'emp-attrition'. The main area shows a table of resources:

Name	Type	Location	...
emattritiondata	Storage account	East US	...
emp-attrition-df	Data factory (V2)	East US	...
emp-attrition-sa	Synapse workspace	East US	...
employee-attrition	Azure Databricks Service	East US	...

## 14. Created Azure Synapse workspace.

The screenshot shows the Microsoft Azure portal interface for the 'emp-attrition-sa' Synapse workspace. The top navigation bar includes the Microsoft Azure logo, a search bar, and user information (shethp2109@gmail.com). The main content area displays the workspace details under the 'Essentials' tab. Key information includes:

- Resource group: emp-attrition
- Status: Succeeded
- Location: East US
- Subscription: Azure subscription 1
- Subscription ID: 2d5acfd-a-cf6-4506-a84b-e7fe31387dfe
- Managed virtual network: Yes
- Managed Identity object ID: 6931e158-4ba3-46ed-a196-90f9797ac47f
- Workspace web URL: <https://web.azuresynthesize.net?workspace=%2fsubscri...>
- Tags: (edit) : Add tags
- Networking: Primary ADLS Gen2 account: https://emattritiondata.dfs.core.windows.net
- SQL admin username: sqladminuser
- SQL Microsoft Entra admin: live.com#shethp2109@gmail.com
- Dedicated SQL endpoint: emp-attrition-sa.sql.azuresynapse.net
- Serverless SQL endpoint: emp-attrition-sa-on-demand.sql.azuresynapse.net
- Development endpoint: https://emp-attrition-sa.dev.azuresynthesize.net

The left sidebar contains navigation links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings (Microsoft Entra ID, Properties, Locks), Analytics pools (SQL pools, Apache Spark pools, Data Explorer pools (preview)), and Security (Encryption, Networking).

## 15. Created EmployeeAttritionDB in synapse.

The screenshot shows the Microsoft Azure portal interface for the 'EmployeeAttritionDB' table within the 'EmployeeAttritionDB' database of the 'emp-attrition-sa' workspace. The table structure is defined as follows:

Name	Type	Length	Nullable
EmployeeID	int	10	NO
Age	float	10	NO
Attrition	varchar	50	NO
BusinessTravel	varchar	50	NO
DailyRate	float	10	NO
Department	varchar	50	NO
DistanceFromHome	float	10	NO

The 'Properties' pane on the right shows the following settings:

- General: Name: EmployeeAttritionDB
- Description: (empty)
- Storage settings for database: Linked service: emp-attrition-sa-WorkspaceDefault...
- Input folder: emp-attrition-data/EmployeeA...
- Data format: Delimited Text

## 16. Writing SQL scripts to fetch data.

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. On the left, there's a sidebar with icons for Home, Databases, Workspaces, and Linked services. The main area shows a workspace named "EmployeeAttritionDB". A search bar at the top has the placeholder "Search". Below it, there are tabs for "SQL script 1", "SQL script 2", "SQL script 3", and "SQL script 4". The "SQL script 1" tab is active, displaying the following SQL code:

```
1 SELECT
2 CASE
3     WHEN Age BETWEEN 18 AND 25 THEN '18-25'
4     WHEN Age BETWEEN 26 AND 35 THEN '26-35'
5     WHEN Age BETWEEN 36 AND 45 THEN '36-45'
6     WHEN Age BETWEEN 46 AND 55 THEN '46-55'
7     ELSE '56+'
8 END AS AgeGroup,
9 Attrition,
10 COUNT(*) AS Count
11 FROM
12     |tbl_empattrition|
13 GROUP BY
14 CASE
15     WHEN Age BETWEEN 18 AND 25 THEN '18-25'
```

The "Results" tab is selected, showing a table with three columns: "AgeGroup", "Attrition", and "Count". The data is as follows:

AgeGroup	Attrition	Count
18-25	No	89
26-35	No	586

This screenshot shows the same workspace environment as the first one. The "EmployeeAttritionDB" workspace is selected. The "SQL script 2" tab is active, displaying the following SQL code:

```
1 SELECT
2     Gender,
3     Attrition,
4     COUNT(*) AS Count
5 FROM
6     |tbl_empattrition|
7     GROUP BY
8     |Gender, Attrition|;
```

The "Results" tab is selected, showing a table with three columns: "Gender", "Attrition", and "Count". The data is as follows:

Gender	Attrition	Count
F	No	10

A "Properties" panel is open on the right side of the screen, showing the following details for "SQL script 2":

- Name:** SQL script 2
- Description:** (empty)
- Type:** .sql script
- Size:** 671 bytes
- Results settings per query:** (with a link)

Microsoft Azure | Synapse Analytics > emp-attrition-sa

We use optional cookies to provide a better experience. Learn more

Accept Reject More options

Synapse live Validate all Publish all 1

Data Workspace Filter resources by name

Lake database EmployeeAttritionDB Tables

EmployeeAttritionDB SQL script 1 SQL script 2 SQL script 3 SQL script 4 Connect to Built-in Use database EmployeeAttritionDB

Run Undo Publish Query plan

```
1 SELECT
2     Department,
3     Attrition,
4     COUNT(*) AS Count
5 FROM
6     tbl_emppattrition
7 GROUP BY
8     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

We use optional cookies to provide a better experience. Learn more

Accept Reject More options

Synapse live Validate all Publish all 1

Data Workspace Filter resources by name

Lake database EmployeeAttritionDB Tables

EmployeeAttritionDB SQL script 1 SQL script 2 SQL script 3 SQL script 4 Connect to Built-in Use database EmployeeAttritionDB

Run Undo Publish Query plan

```
1 SELECT
2     Attrition,
3     COUNT(*) AS Count
4 FROM
5     tbl_emppattrition
6 GROUP BY
7     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

Microsoft Azure | Synapse Analytics > emp-attrition-sa

We use optional cookies to provide a better experience. [Learn more](#)

Synapse live Validate all Publish all 1

Data Workspace Filter resources by name

Lake database EmployeeAttritionDB Tables

SQL script 1 SQL script 2 SQL script 3 SQL script 4 SQL script 5 Connect to Built-in Use database EmployeeAttritionDB

```

1 SELECT
2   JobRole,
3   Attrition,
4   COUNT(*) AS Count
5   FROM
6   |tbl_empattrition
7   GROUP BY
8   |JobRole, Attrition;

```

Properties General Related (0)

Name \* SQL script 5

Description

Type .sql script

Size 671 bytes

Results settings per query

First 5000 rows (default)

All rows

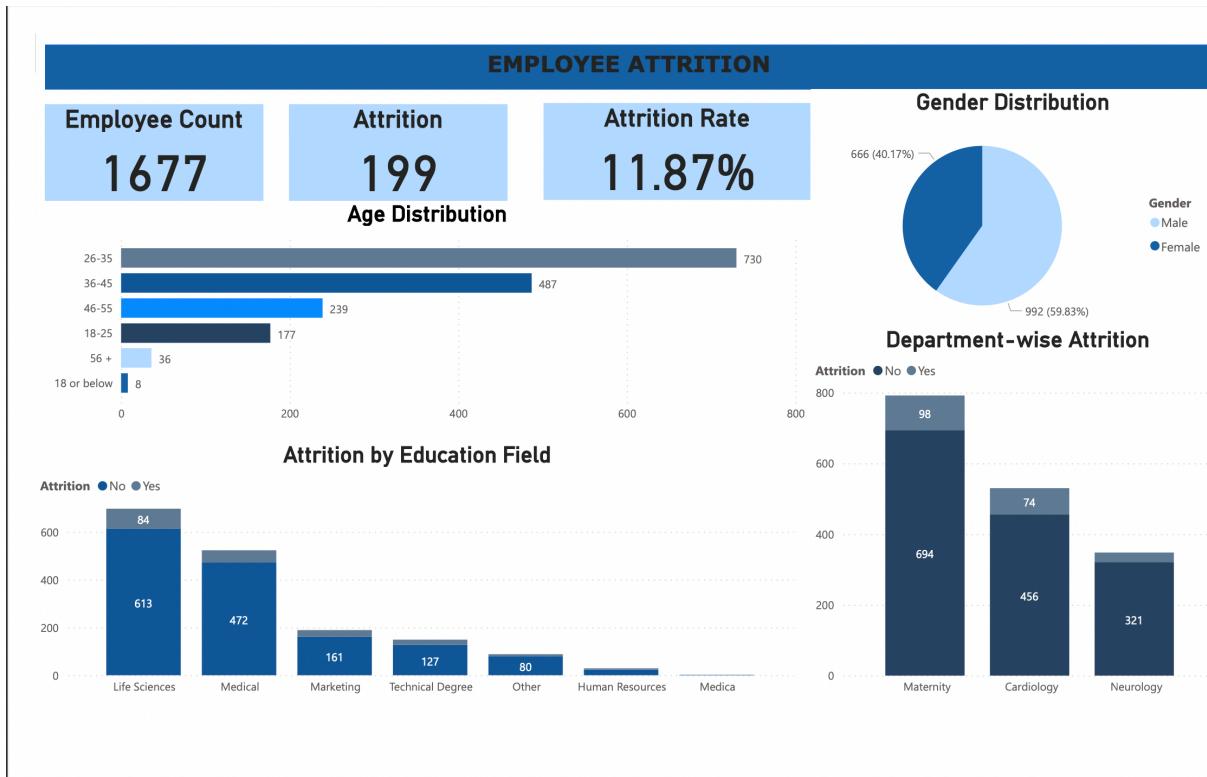
Results Messages

View Table Chart Export results

Search

JobRole	Attrition	Count
Admin	No	16
Administrative	No	114

## 17. Final PowerBI dashboard development.

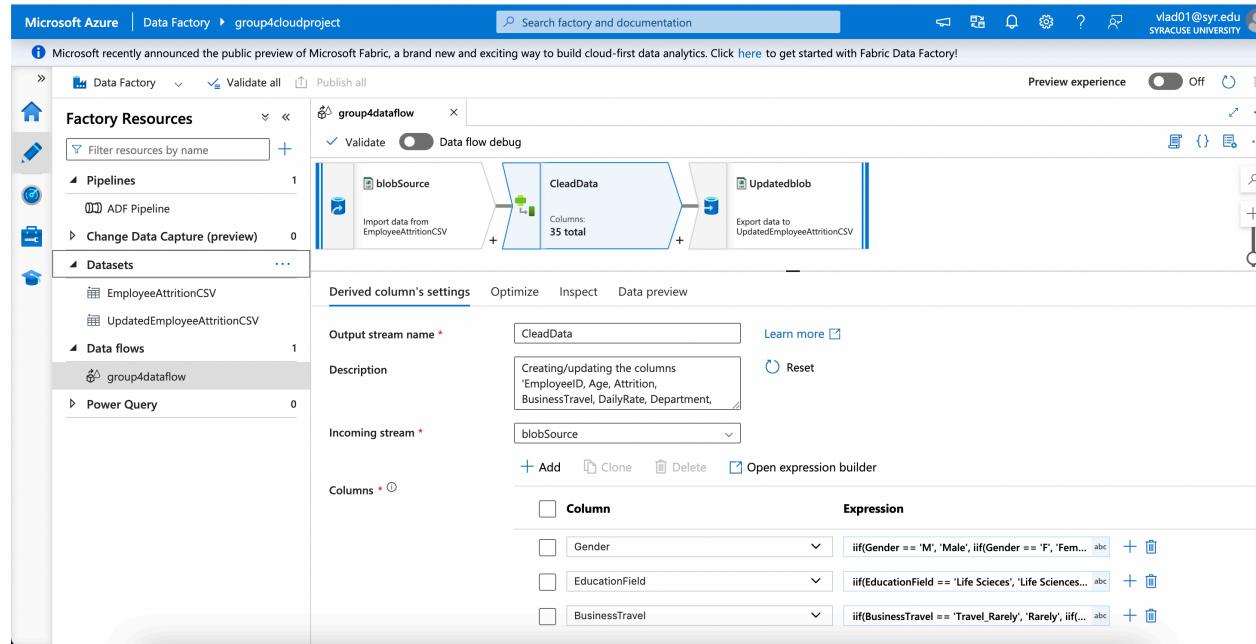


## APPROACH TWO: DATA CLEANING PROCESS USING ADF DATA FLOW

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.