Problem Statement: End-to-End ETL Pipeline Using Azure Blob Storage, Azure Data Factory, Databricks, and Power BI

Objective:

To build an automated ETL pipeline that extracts raw data from a Git repository, stores it in Azure Blob Storage using ADF, processes it in Databricks, saves the transformed data back to Blob Storage, and visualizes it in Power BI Cloud.

Notes:

Dataset Information:

⇒ The raw datasets are available in the following GitHub locations:
Students Data: stu.csv
Courses Data: course.csv

⇒ stu.csv contains student details (ID, name, age, course).

⇒ course.csv contains course information (course name, fee).

Data Ingestion in ADF:

⇒ In Azure Data Factory (ADF), use "HTTPS" as the linked service to fetch data from GitHub.

⇒ Set authentication type to "Anonymous" to access public GitHub files.

⇒ Store the fetched data in Azure Blob Storage (container).

Accessing Blob Storage in Databricks:

⇒ In Azure Databricks, use the storage account access key (available in Azure Security Settings) to connect to Blob Storage.

⇒ Read the raw CSV files, perform the inner join on "course", and save the results in a separate output folder (container).

```
# Define storage account and container details
storage_account_name = "s777"
container_name = "c101"
access_key =
"oo41afBqbKVlFejIMgkSGhSkXQs1O3yF7MYyG9GQjvLbCCXrRClCO0D2lB+HIp8h66smbSt0Q
Uaw+AStFkeY7g=="

# Set access key for Azure Blob Storage
spark.conf.set(f"fs.azure.account.key.{storage_account_name}.blob.core.win
dows.net", access_key)

# Load student.csv file
student_df =
spark.read.csv(f"wasbs://{container_name}@{storage_account_name}.blob.core
.windows.net/student.csv",
                            header=True, inferSchema=True)

# Display the dataframe
student_df.show()

# Load course1.csv file
course1_df =
spark.read.csv(f"wasbs://{container_name}@{storage_account_name}.blob.core
.windows.net/course1.csv",header=True, inferSchema=True)

# Display the dataframe
course1_df.show()
```

```
from pyspark.sql import SparkSession

# Join student_df and course1_df on 'course' field
joined_df = student_df.join(course1_df, on="course", how="inner")

# Show the joined DataFrame
joined_df.show()

# Define output path in Azure Blob Storage
```

```python
output_path = f"wasbs://{container_name}@{storage_account_name}.blob.core.windows.net/output/joined_data.csv"

# Write the joined DataFrame to a CSV file
joined_df.write.mode("overwrite").option("header", True).csv(output_path)
```

```python
joined_df.write.mode("overwrite").format("parquet").save(output_path)
```