Databricks Case Study

-Sarthak Niranjan Kulkarni (Maverick)

- sarthakkul2311@gmail.com - (+91) 93256 02791

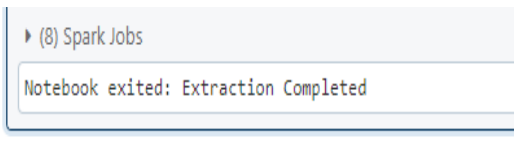
06/12/2024 (Friday)

Download Data Set from Kaggle:

students-copy.csv: https://www.kaggle.com/datasets/zeeshier/student-information-dataset

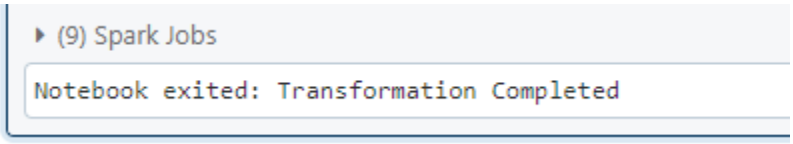
Extraction: -

```
from pyspark.sql import SparkSession


# Initialize Spark Session

spark = SparkSession.builder.appName("ETL_Students_Extract").getOrCreate()


# data path

data_path = "dbfs:/user/hive/warehouse/students_copy"


# Extract data

students_df = spark.read.format("delta").load(data_path)

extracted_data_path = "dbfs:/user/hive/warehouse/students_extracted"

students_df.write.format("delta").mode("overwrite").save(extracted_data_path)

students_df.show()

dbutils.notebook.exit("Extraction Completed")
```



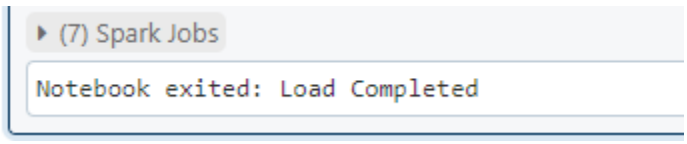
Transformation: -

```
from pyspark.sql import SparkSession

from pyspark.sql.functions import col


# Initialize Spark Session

spark = SparkSession.builder.appName("ETL_Students_Transform").getOrCreate()


# Read data

extracted_data_path = "dbfs:/user/hive/warehouse/students_extracted"

students_df = spark.read.format("delta").load(extracted_data_path)


# Transformation: filtering students above age 21

transformed_df = students_df.filter(col("Age") > 21)

transformed_data_path = "dbfs:/user/hive/warehouse/students_transformed"

transformed_df.write.format("delta") \

    .option("mergeSchema", "true") \

    .mode("overwrite") \

    .save(transformed_data_path)

transformed_df.show()

dbutils.notebook.exit("Transformation Completed")
```



Load: -

```
from pyspark.sql import SparkSession


# Initialize Spark Session

spark = SparkSession.builder.appName("ETL_Students_Load").getOrCreate()

# Read the transformed data


transformed_data_path = "dbfs:/user/hive/warehouse/students_transformed"

transformed_df = spark.read.format("delta").load(transformed_data_path)


# Load: Save the transformed data (original Delta table location)

data_path = "dbfs:/user/hive/warehouse/students_copy"


transformed_df.write.format("delta") \
    .option("overwriteSchema", "true") \
    .mode("overwrite") \
    .save(data_path)


transformed_df.show()

dbutils.notebook.exit("Load Completed")
```



Workflow: ETL-Case Study