Databricks Case Study

-Sarthak Niranjan Kulkarni (Maverick)

- [sarthakkul2311@gmail.com](mailto: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

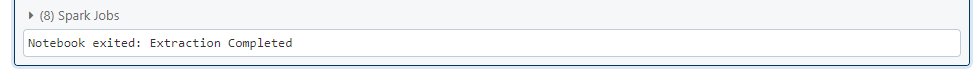
students\_df = spark.read.format("delta").load(data\_path)

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")

A screen shot of a computer

Description automatically generated

**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")

A screen shot of a computer

Description automatically generated

**Workflow: ETL-Case Study**

A screenshot of a chat

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated