

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")
```

▶ (8) Spark Jobs

Notebook exited: 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")
```

► (9) Spark Jobs

Notebook exited: 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

Microsoft Azure

Search data, notebooks, recents, and more...

CTRL + P

Sarthak_workspace

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Partner Connect

Workflows > Jobs >

ETL-CaseStudy

Send feedback

Run now

Runs

Tasks

Extraction
...mml.local@techademy.com/Extract-2
Sarthak Cluster

Transformation
...local@techademy.com/Transform-2
Sarthak Cluster

Load
...lvmml.local@techademy.com/Load-2
Sarthak Cluster

+ Add task

No task selected
Choose a task from the graph to edit its properties

Job details

Job ID
751031295642777

Creator
azuser2371_mml.local

Run as
azuser2371_mml.local

Tags
Add tag

Description
Add description

Git
Not configured
Add Git settings

Schedule
Activate Windows
Go to Settings to activate Windows.

Microsoft Azure

Search data, notebooks, recents, and more...

CTRL + P

Sarthak_workspace

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Partner Connect

Workflows > Jobs >

ETL-CaseStudy

Send feedback

Run now

Runs

Tasks

Run total duration
35s
17s
Tasks
Extraction
Transformation
Load
Go to the latest successful run
Cancel runs

Start time
Run ID
Launched
Duration
Status
Error code
Run param...

Dec 06, 2024, 1...
482634321...
Manually
36s
Succeeded

Job details

Job ID
751031295642777

Creator
azuser2371_mml.local

Run as
azuser2371_mml.local

Tags
Add tag

Description
Add description

Git
Not configured
Add Git settings

Schedule
Activate Windows
Go to Settings to activate Windows.

Microsoft Azure

Search data, notebooks, recents, and more...

CTRL + P

Sarthak_workspace

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Partner Connect

Workflows > Jobs > ETL-CaseStudy >

ETL-CaseStudy run

Send feedback

Delete job run

Repair run

Graph

Timeline

Extraction
Succeeded - 13s
...mml.local@techademy.com/Extract-2
Sarthak Cluster

Transformation
Succeeded - 11s
...local@techademy.com/Transform-2
Sarthak Cluster

Load
Succeeded - 10s
...lvmml.local@techademy.com/Load-2
Sarthak Cluster

Job run details

Job ID
751031295642777

Job run ID
482634321168044

Launched
Manually

Started
12/06/2024, 11:54:32 PM

Ended
12/06/2024, 11:55:07 PM

Duration
36s

Queue duration
-

Status
Succeeded

View run events

Compute
Activate Windows
Go to Settings to activate Windows.

Sarthak Cluster