Azure Databricks Case Study

Submitted By-
Subrat Shukla, DE-1

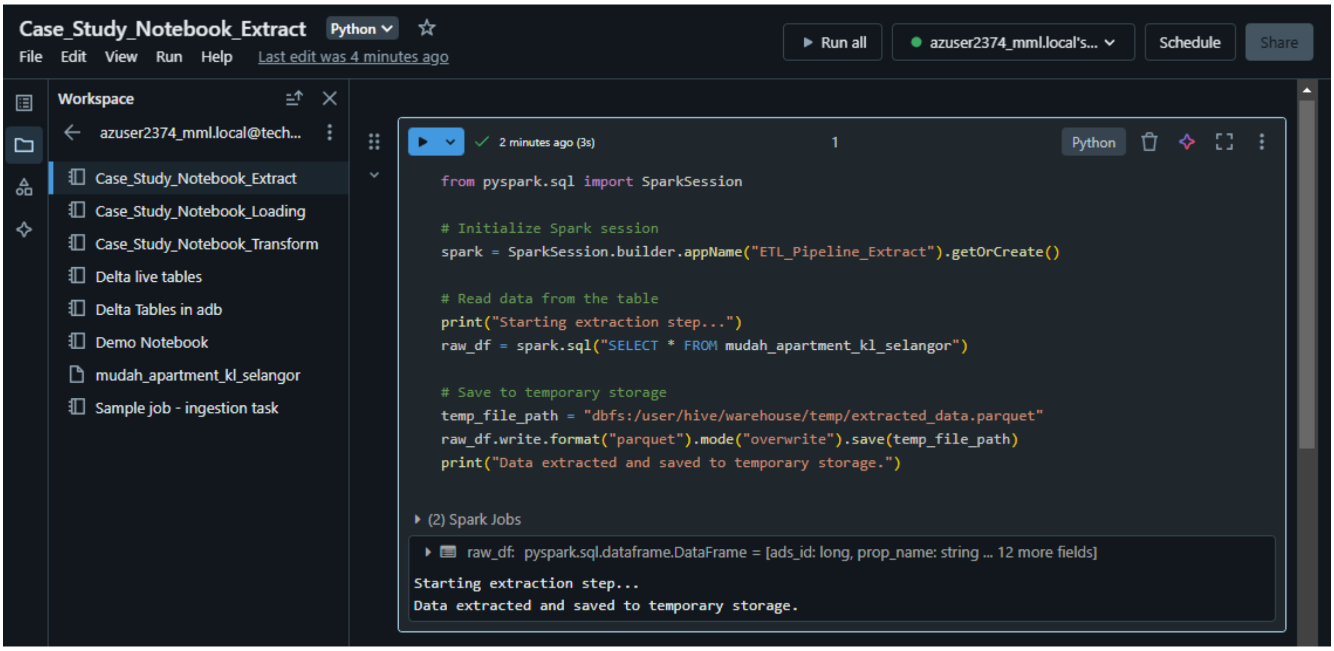
Task given: Create an ETL pipeline of ingestion & transform and load queries on any data set and initiate the pipeline from workflow using the notebook.

Steps:

- Create a notebook with ETL queries.
- Run the notebook from workflow pipeline in azure databricks workspace.

1. Notebook for Ingestion (Extract):

This notebook will load the raw data into the Databricks environment, either from a file or a table.

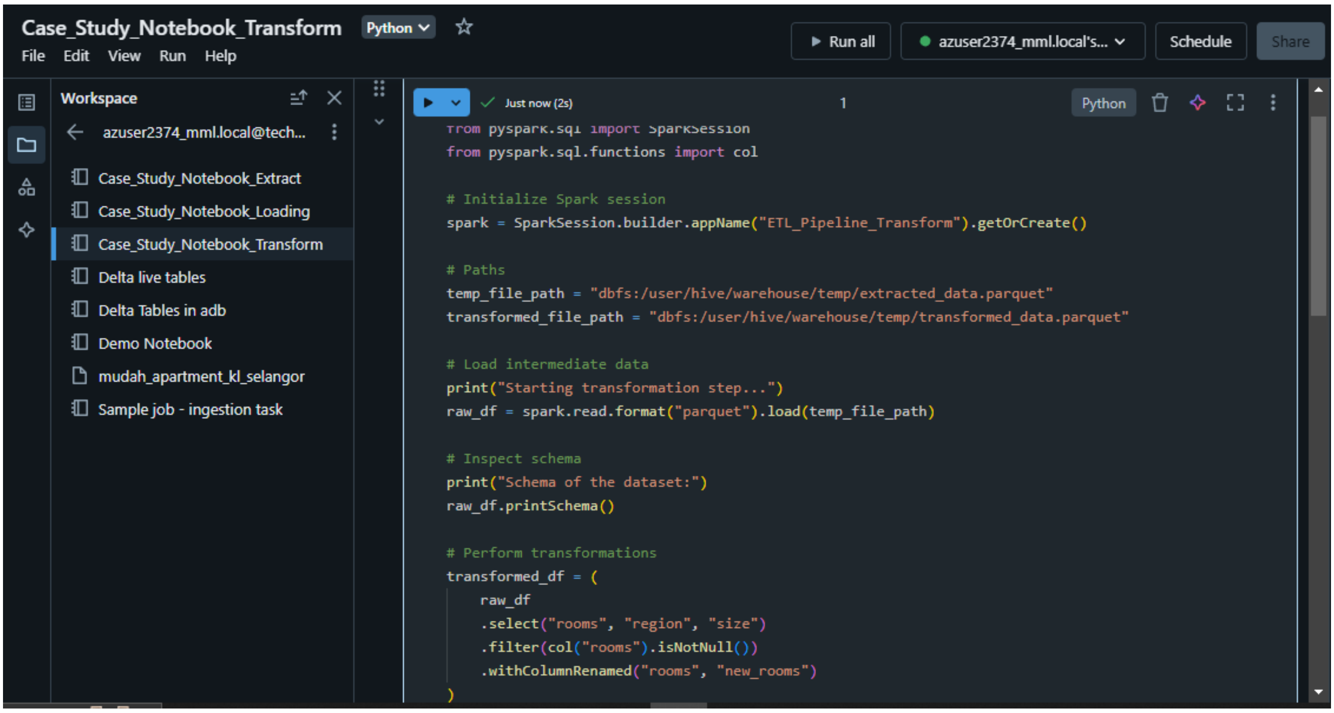

Case_Study_Notebook_Extract
Python
File
Edit
View
Run
Help
Last edit was 4 minutes ago
Run all
azuser2374_mml.local's...
Schedule
Share
Workspace
azuser2374_mml.local@tech...
Case_Study_Notebook_Extract
Case_Study_Notebook_Loading
Case_Study_Notebook_Transform
Delta live tables
Delta Tables in adb
Demo Notebook
mudah_apartment_kl_selangor
Sample job - ingestion task
2 minutes ago (3s)
1
Python
(2) Spark Jobs

```
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("ETL_Pipeline_Extract").getOrCreate()

# Read data from the table
print("Starting extraction step...")
raw_df = spark.sql("SELECT * FROM mudah_apartment_kl_selangor")

# Save to temporary storage
temp_file_path = "dbfs:/user/hive/warehouse/temp/extracted_data.parquet"
raw_df.write.format("parquet").mode("overwrite").save(temp_file_path)
print("Data extracted and saved to temporary storage.")
```

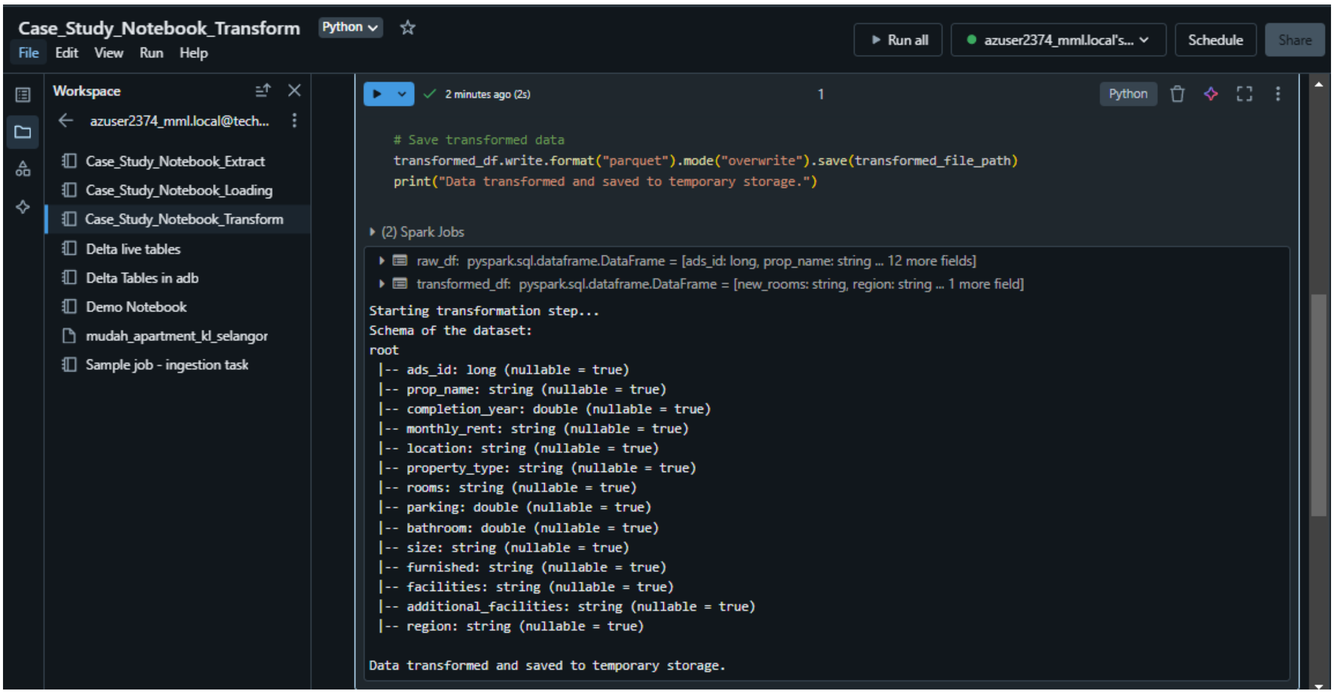
```
raw_df: pyspark.sql.dataframe.DataFrame = [ads_id: long, prop_name: string ... 12 more fields]
Starting extraction step...
Data extracted and saved to temporary storage.
```

2. Notebook for Transformation

This notebook will read the extracted data, apply transformations, and prepare it for loading.


Case_Study_Notebook_Transform
Python
File
Edit
View
Run
Help
Run all
azuser2374_mml.local's...
Schedule
Share
Workspace
azuser2374_mml.local@tech...
Case_Study_Notebook_Extract
Case_Study_Notebook_Loading
Case_Study_Notebook_Transform
Delta live tables
Delta Tables in adb
Demo Notebook
mudah_apartment_kl_selangor
Sample job - ingestion task
Just now (2s)
1
Python

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# Initialize Spark session
spark = SparkSession.builder.appName("ETL_Pipeline_Transform").getOrCreate()

# Paths
temp_file_path = "dbfs:/user/hive/warehouse/temp/extracted_data.parquet"
transformed_file_path = "dbfs:/user/hive/warehouse/temp/transformed_data.parquet"

# Load intermediate data
print("Starting transformation step...")
raw_df = spark.read.format("parquet").load(temp_file_path)

# Inspect schema
print("Schema of the dataset:")
raw_df.printSchema()

# Perform transformations
transformed_df = (
    raw_df
    .select("rooms", "region", "size")
    .filter(col("rooms").isNotNull())
    .withColumnRenamed("rooms", "new_rooms")
)
```


Case_Study_Notebook_Transform
Python
File
Edit
View
Run
Help
Run all
azuser2374_mml.local's...
Schedule
Share
Workspace
azuser2374_mml.local@tech...
Case_Study_Notebook_Extract
Case_Study_Notebook_Loading
Case_Study_Notebook_Transform
Delta live tables
Delta Tables in adb
Demo Notebook
mudah_apartment_kl_selangor
Sample job - ingestion task
2 minutes ago (2s)
1
Python
(2) Spark Jobs

```
# Save transformed data
transformed_df.write.format("parquet").mode("overwrite").save(transformed_file_path)
print("Data transformed and saved to temporary storage.")
```

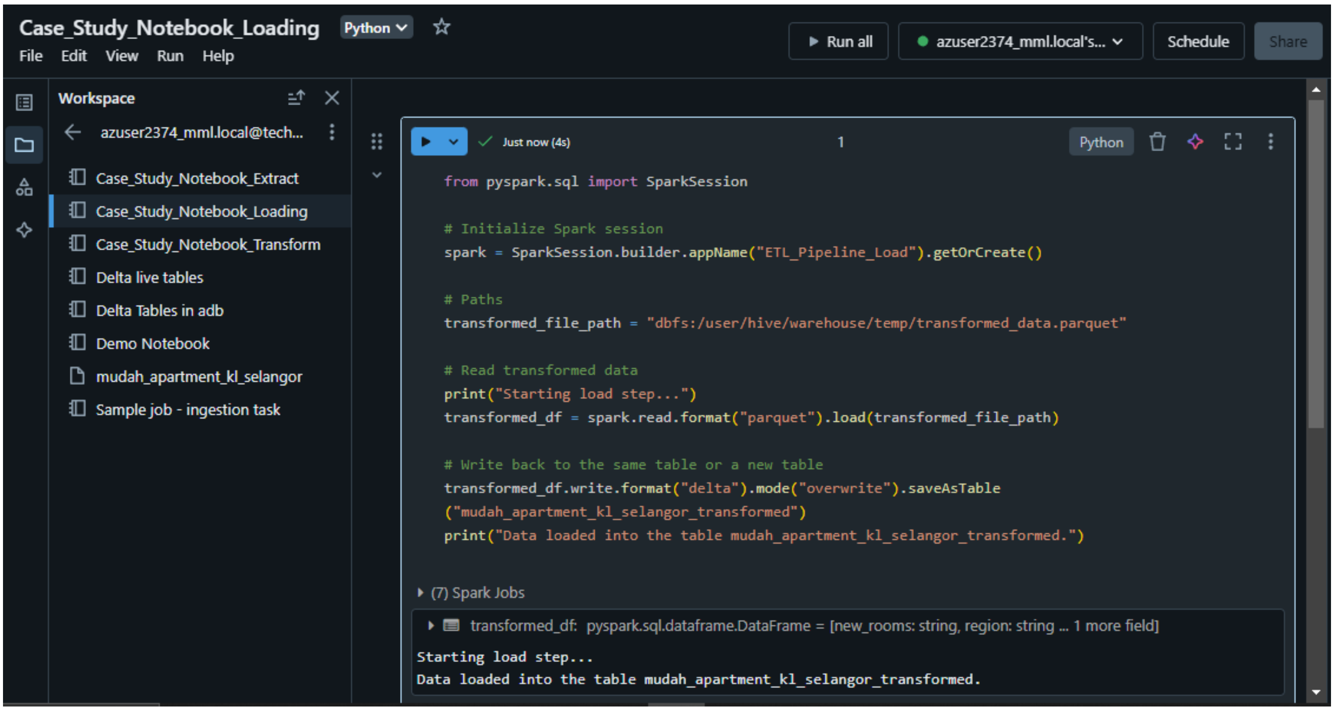
raw_df: pyspark.sql.dataframe.DataFrame = [ads_id: long, prop_name: string ... 12 more fields]

transformed_df: pyspark.sql.dataframe.DataFrame = [new_rooms: string, region: string ... 1 more field]

```
Starting transformation step...
Schema of the dataset:
root
 |-- ads_id: long (nullable = true)
 |-- prop_name: string (nullable = true)
 |-- completion_year: double (nullable = true)
 |-- monthly_rent: string (nullable = true)
 |-- location: string (nullable = true)
 |-- property_type: string (nullable = true)
 |-- rooms: string (nullable = true)
 |-- parking: double (nullable = true)
 |-- bathroom: double (nullable = true)
 |-- size: string (nullable = true)
 |-- furnished: string (nullable = true)
 |-- facilities: string (nullable = true)
 |-- additional_facilities: string (nullable = true)
 |-- region: string (nullable = true)

Data transformed and saved to temporary storage.
```

3. Notebook for Loading

This notebook will load the transformed data into the final destination, such as a Delta table or another storage format.

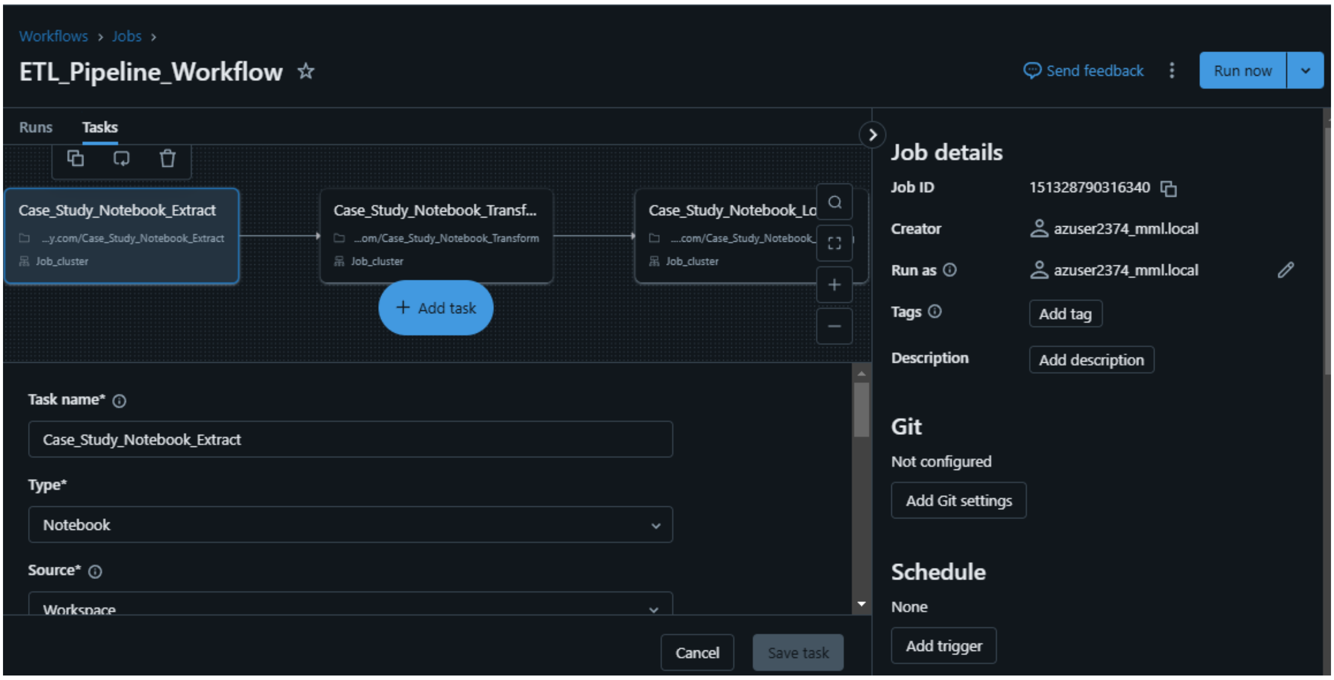

Case_Study_Notebook_Loading
Python
File
Edit
View
Run
Help
Run all
azuser2374_mml.local's...
Schedule
Share
Workspace
azuser2374_mml.local@tech...
Case_Study_Notebook_Extract
Case_Study_Notebook_Loading
Case_Study_Notebook_Transform
Delta live tables
Delta Tables in adb
Demo Notebook
mudah_apartment_kl_selangor
Sample job - ingestion task
Just now (4s)
1
Python

```
from pyspark.sql import SparkSession

# Initialize Spark session
spark = SparkSession.builder.appName("ETL_Pipeline_Load").getOrCreate()

# Paths
transformed_file_path = "dbfs:/user/hive/warehouse/temp/transformed_data.parquet"

# Read transformed data
print("Starting load step...")
transformed_df = spark.read.format("parquet").load(transformed_file_path)

# Write back to the same table or a new table
transformed_df.write.format("delta").mode("overwrite").saveAsTable
("mudah_apartment_kl_selangor_transformed")
print("Data loaded into the table mudah_apartment_kl_selangor_transformed.")
```

```
(7) Spark Jobs
transformed_df: pyspark.sql.dataframe.DataFrame = [new_rooms: string, region: string ... 1 more field]
Starting load step...
Data loaded into the table mudah_apartment_kl_selangor_transformed.
```

4. Creating Databricks Workflow:

- For Extract process-



Task name* Case_Study_Notebook_Extract

Type* Notebook

Source* Workspace

Job details

Job ID 151328790316340

Creator azuser2374_mml.local

Run as azuser2374_mml.local

Tags

Description

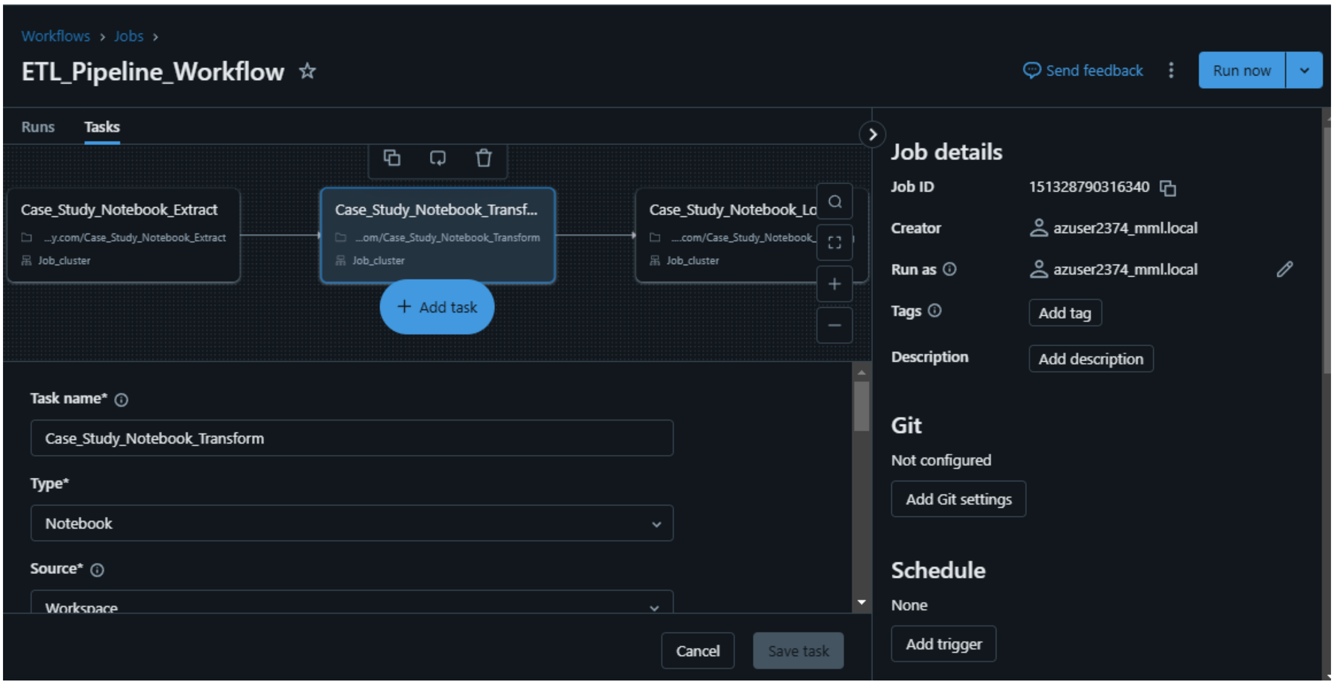
Git

Not configured

Schedule

None

- For Transformation process-



ETL_Pipeline_Workflow

Task name*: Case_Study_Notebook_Transform

Type*: Notebook

Source*: Workspace

Job details

Job ID: 151328790316340

Creator: azuser2374_mml.local

Run as: azuser2374_mml.local

Tags

Description

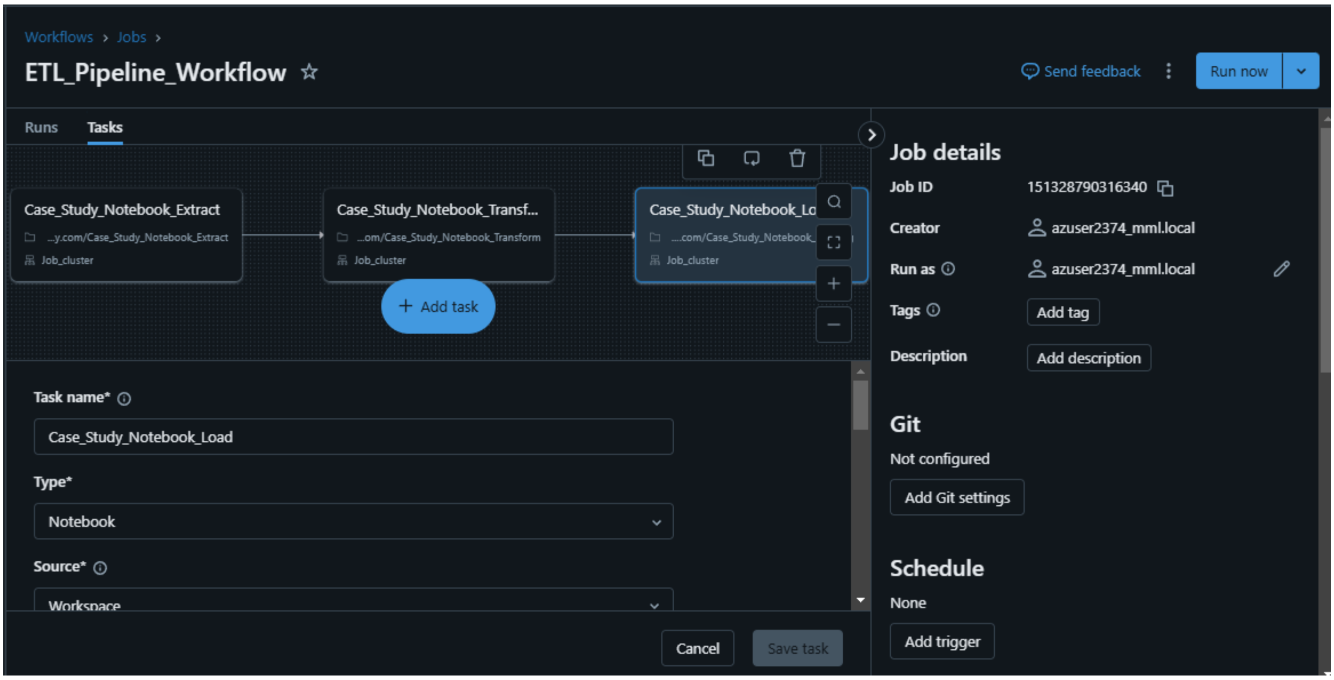
Git

Not configured

Schedule

None

- For Loading process-



ETL_Pipeline_Workflow

Task name*: Case_Study_Notebook_Load

Type*: Notebook

Source*: Workspace

Job details

Job ID: 151328790316340

Creator: azuser2374_mml.local

Run as: azuser2374_mml.local

Tags

Description

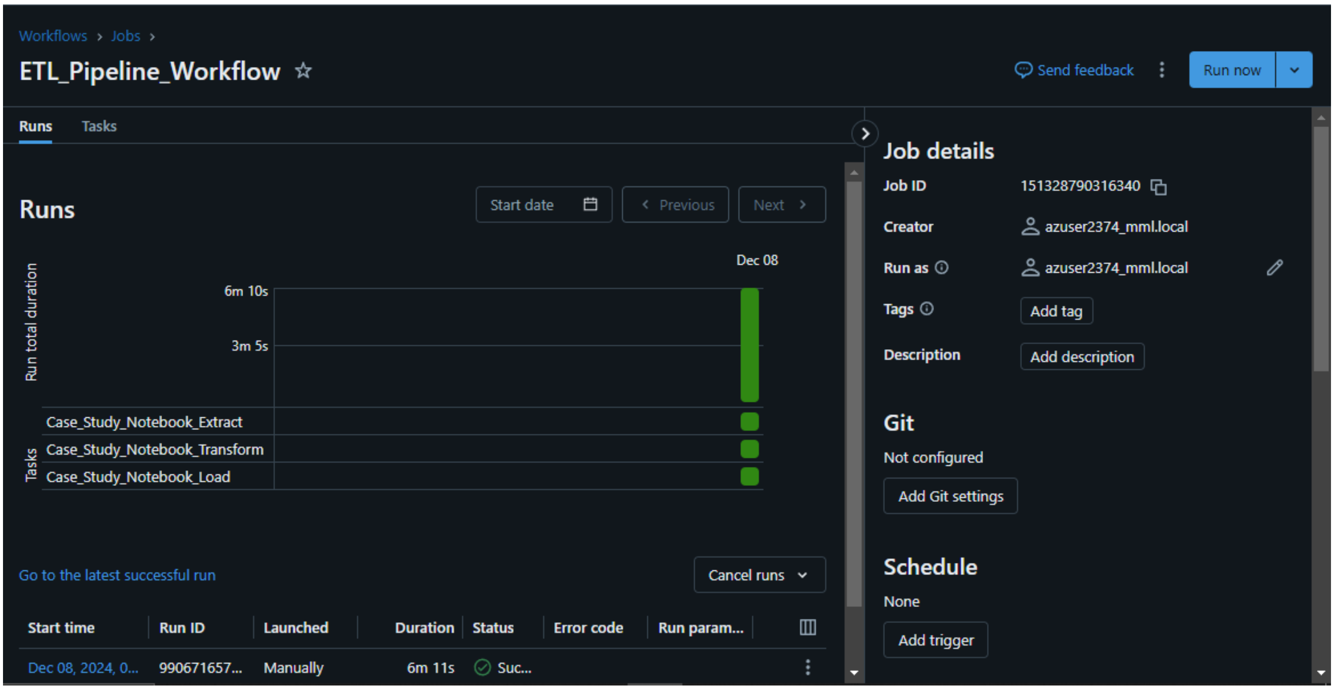
Git

Not configured

Schedule

None

5. Running the Workflows:



Job details

Job ID	151328790316340
Creator	azuser2374_mml.local
Run as	azuser2374_mml.local
Tags	Add tag
Description	Add description

Git

Not configured

Schedule

None

Start time	Run ID	Launched	Duration	Status	Error code	Run param...
Dec 08, 2024, 0...	990671657...	Manually	6m 11s	✓ Suc...		

- Hence, The ETL Pipeline consisting processes of Extraction, Transformation, Loading has successfully created and running.

--Thank You!