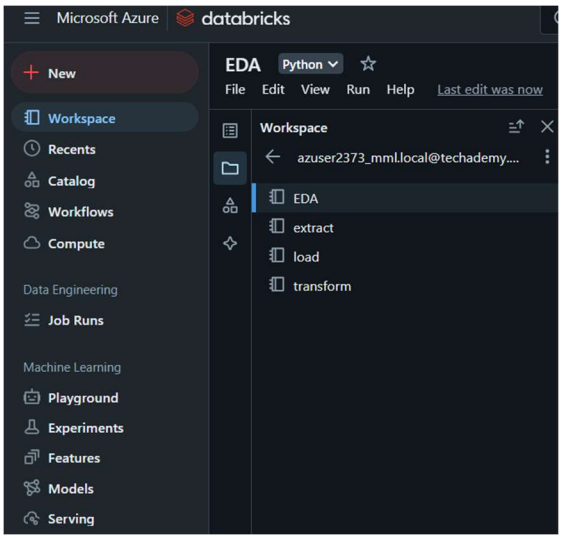
ETL Pipeline

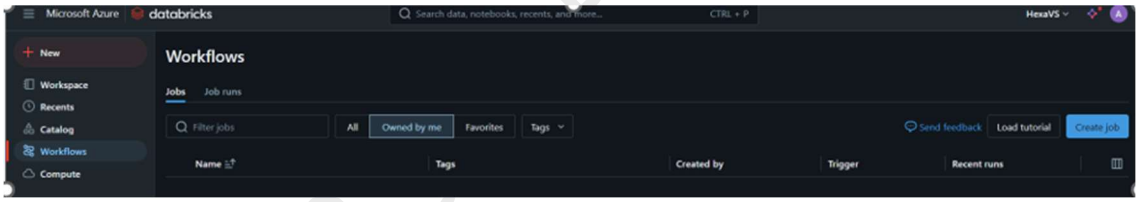
Prerequisite: Azure Databricks service.

First start with creating job, later will explain the code.

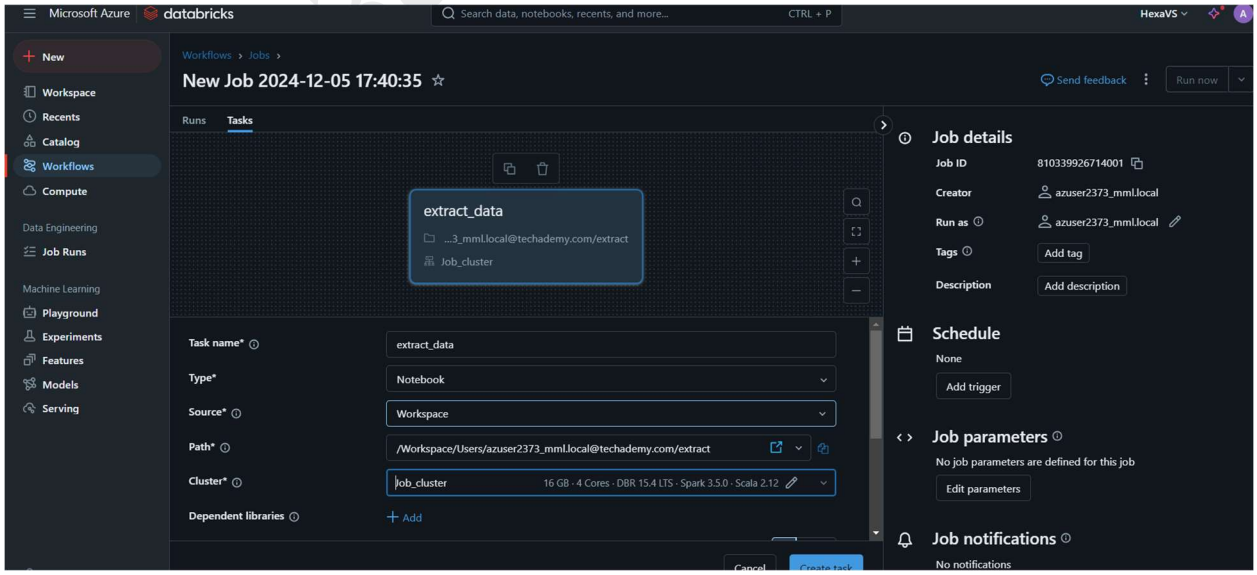
0. Ensure all notebooks are present



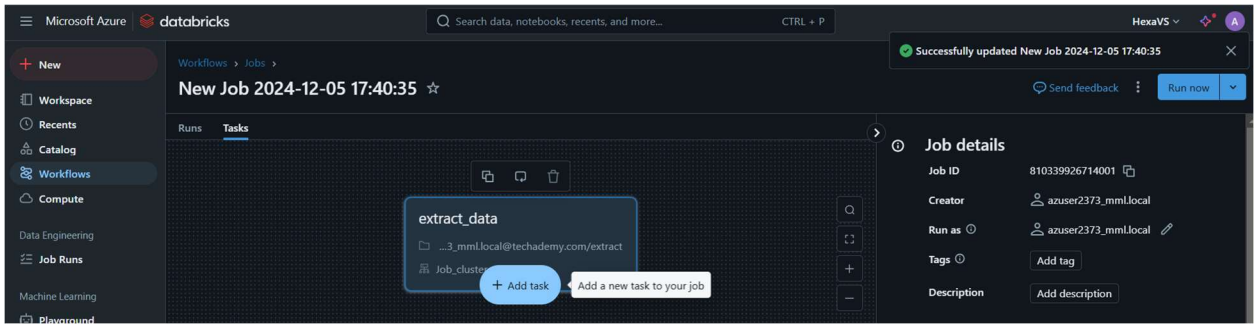
1. Go to Workflow. Click on create job



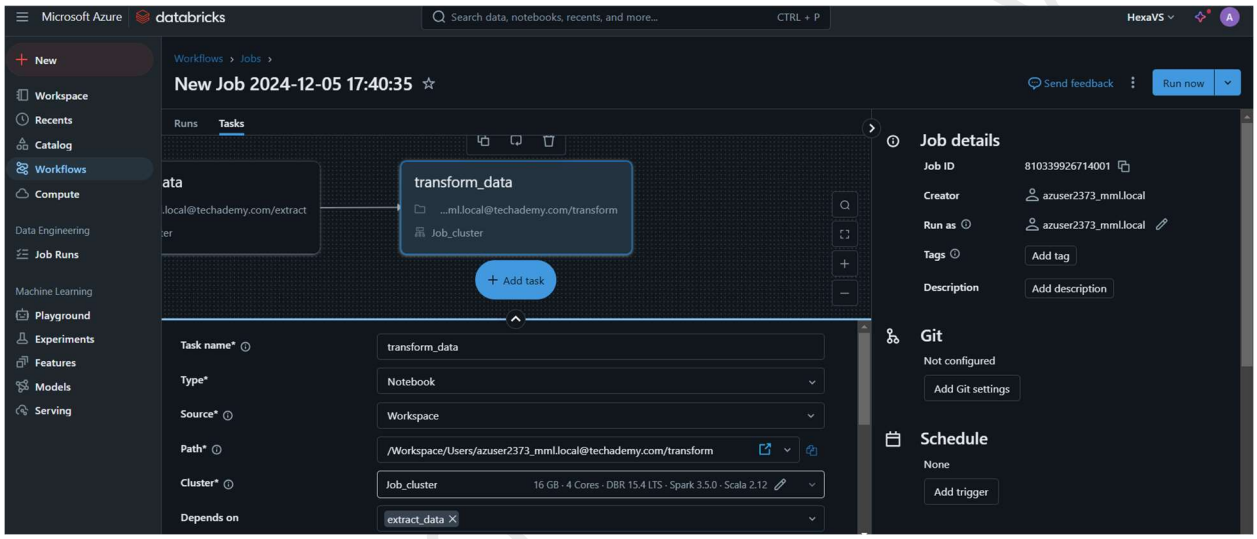
2. Give Name to task, choose the notebook which has code, choose job cluster type, other details.



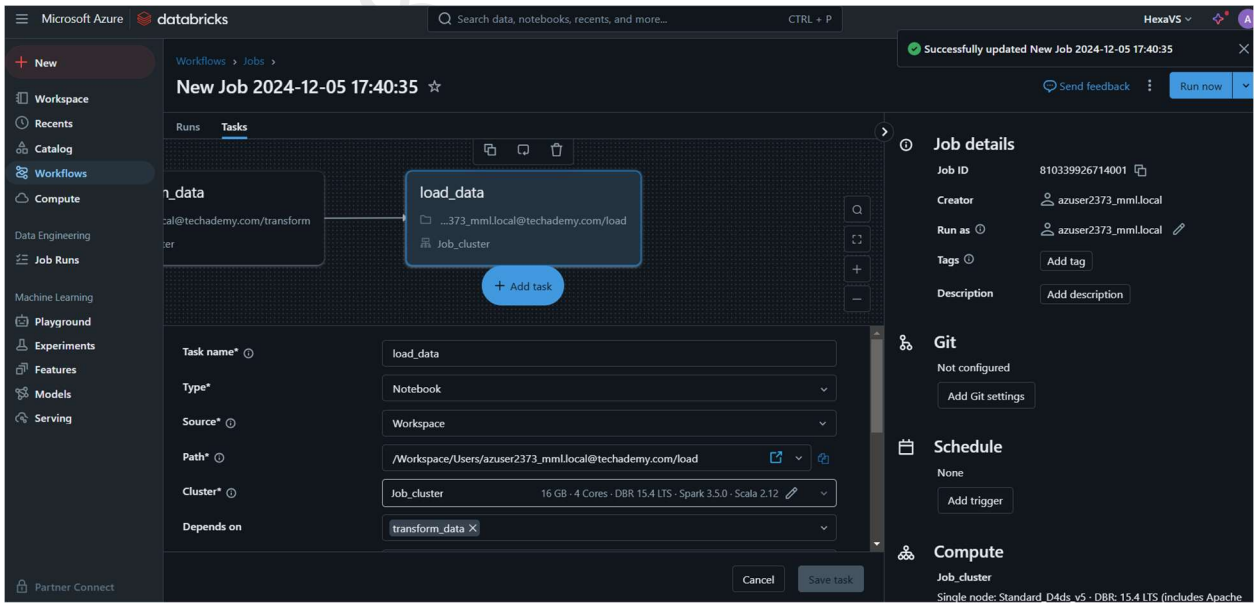
3. click on add task



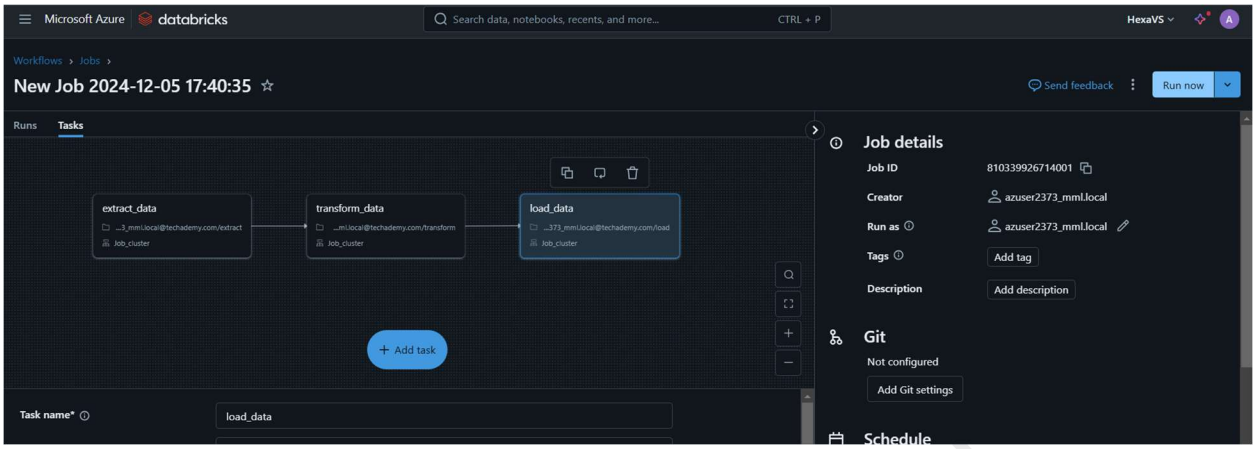
4. Similar to extract_data task, give all necessary details for transform_data.



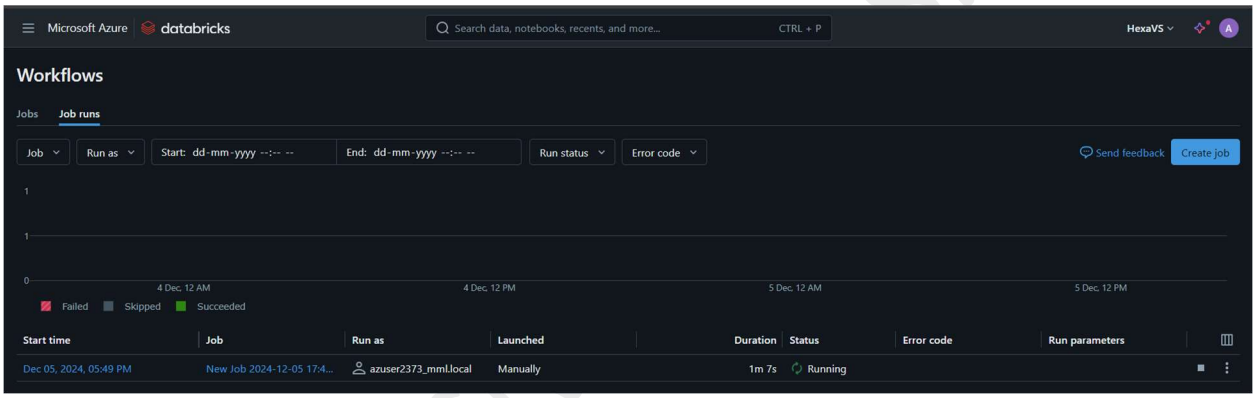
5. Similar to above one do for load_data task



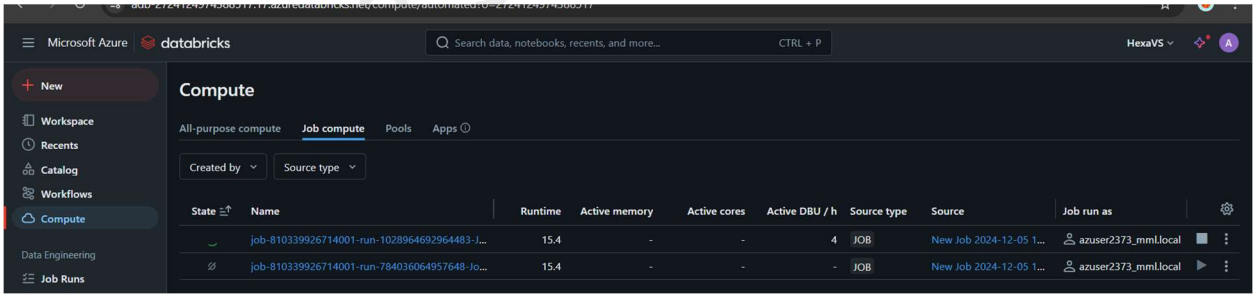
6. Once all setup is done click on Run now



7. Can be monitored in workflow – job run tab

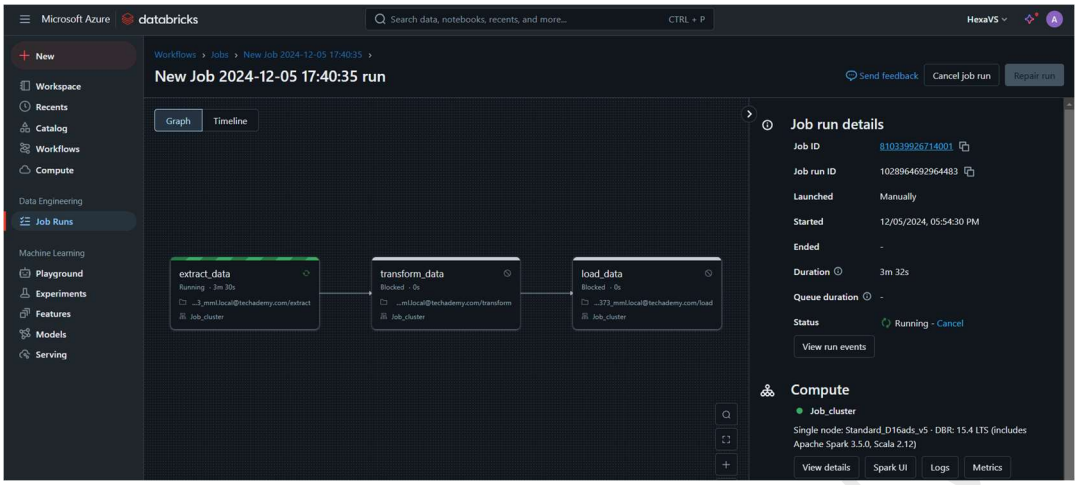


8. As soon as we clicked on Run Job, job cluster is invoked to perform job

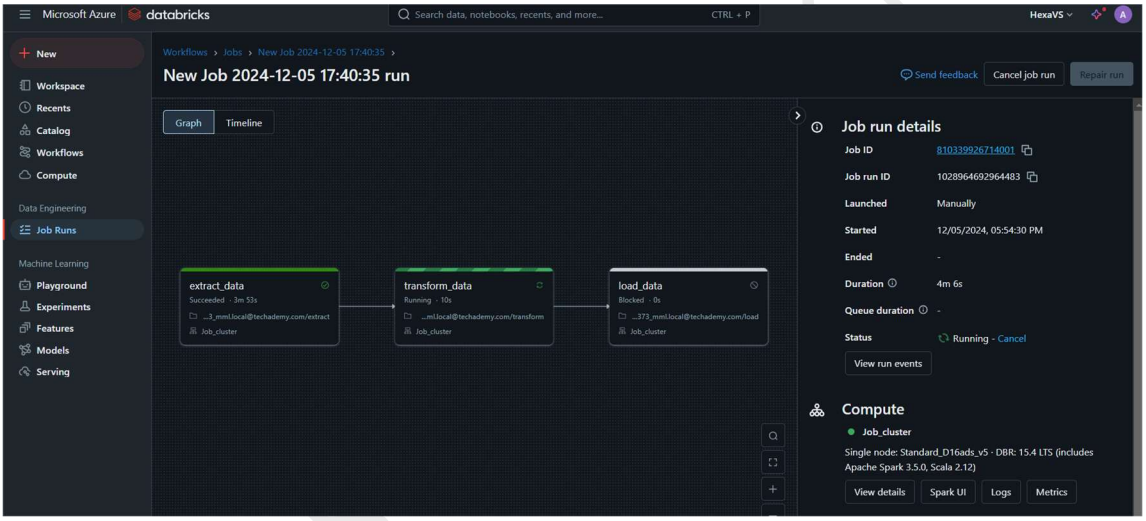


9. After Successful creation of job cluster, it starts to perform the ETL pipeline tasks. Here it starts by extracting data task.

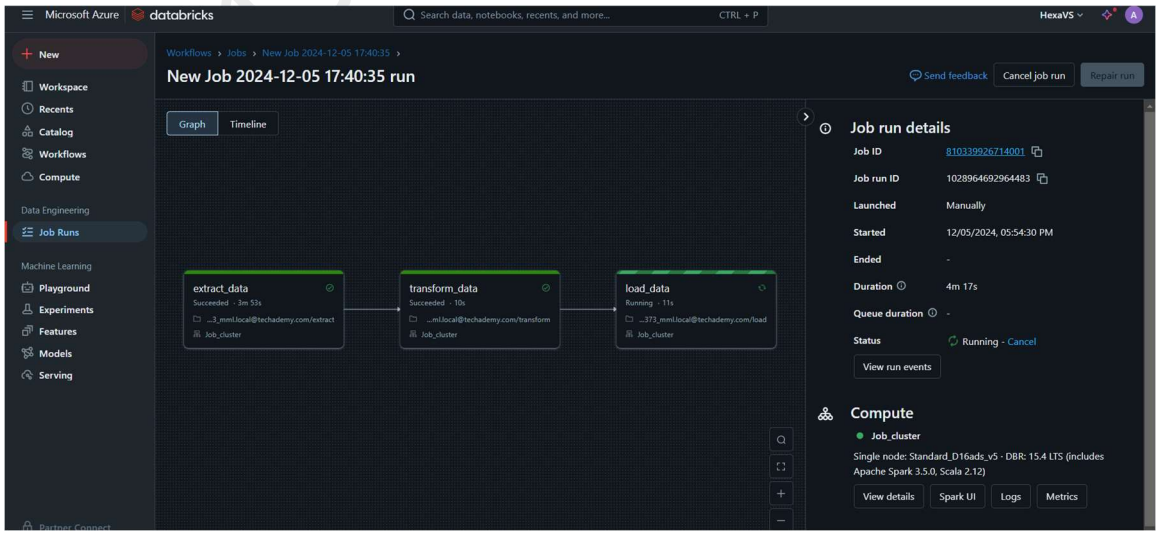
We can check the status or running of data via Directed Acyclic Graph representation.



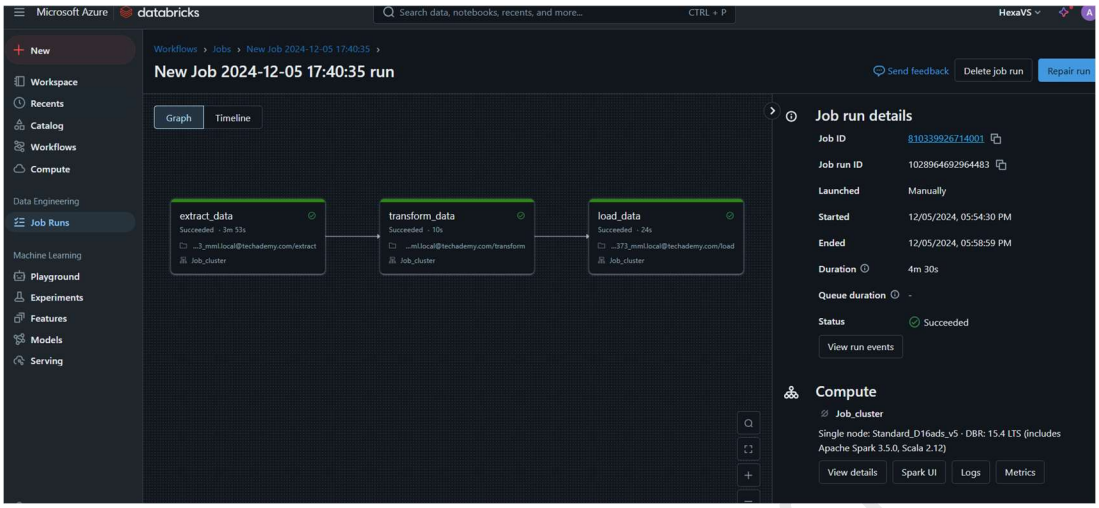
10. After the data extraction, data transformation takes place.



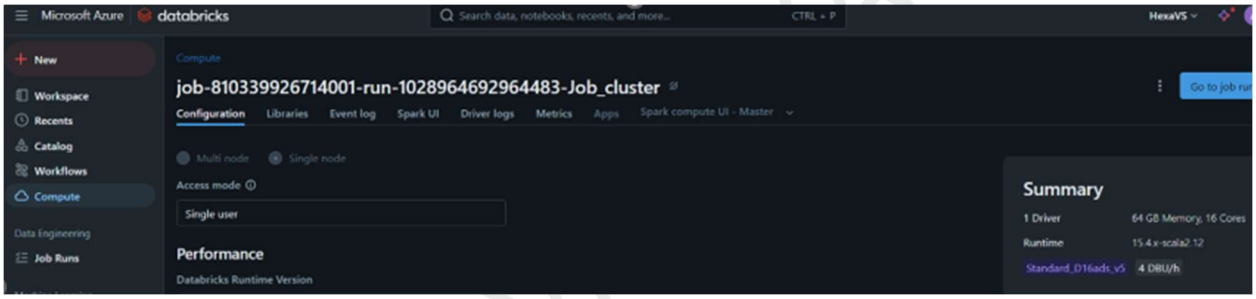
11. After data transformation, data loading takes place.



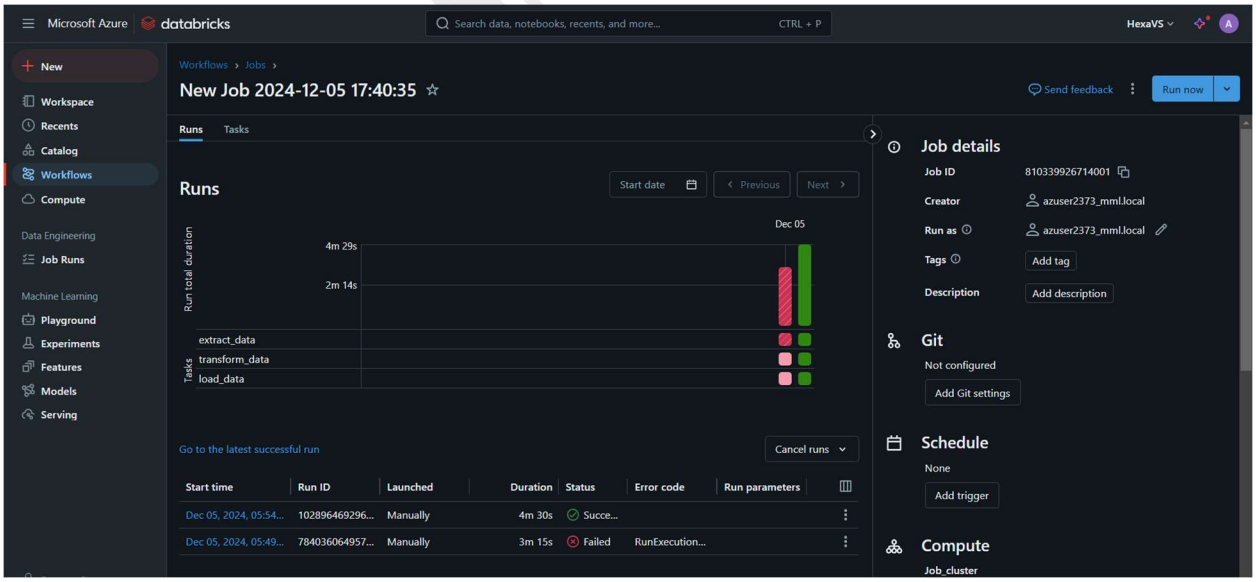
12. Once all the task in pipeline completed successfully, the DAG turns to complete green.



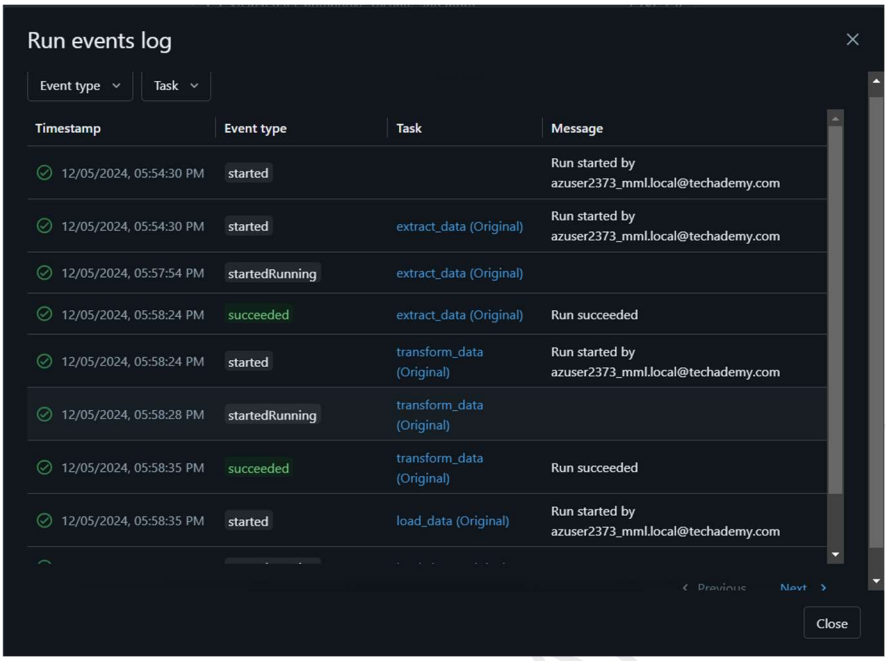
13. As soon as the running job completed, either success or failure, the job cluster terminates.



14. We can check the complete job run in Workflow tab.

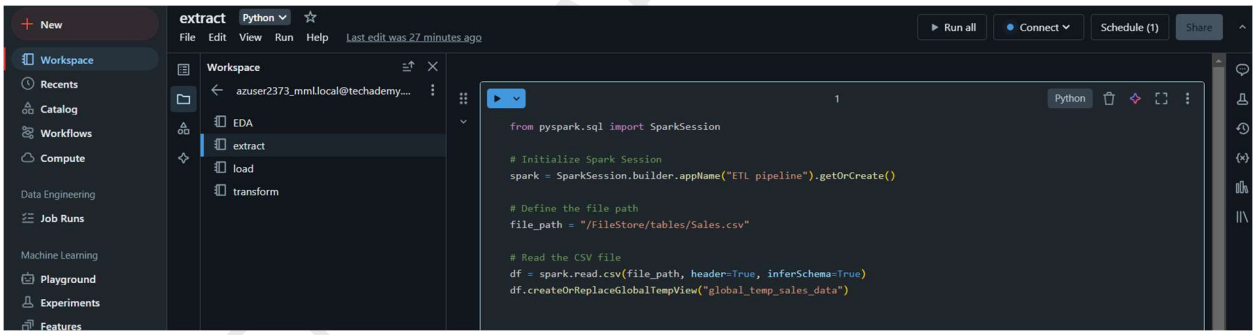


15. Event log is stored for every actions.



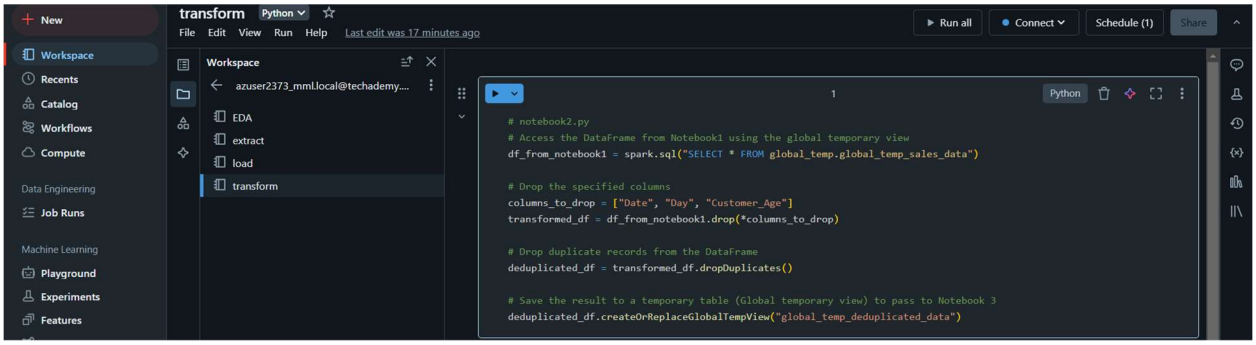
----------CODE DETAILS----------

extract data from Filestore and creates Global temp view.

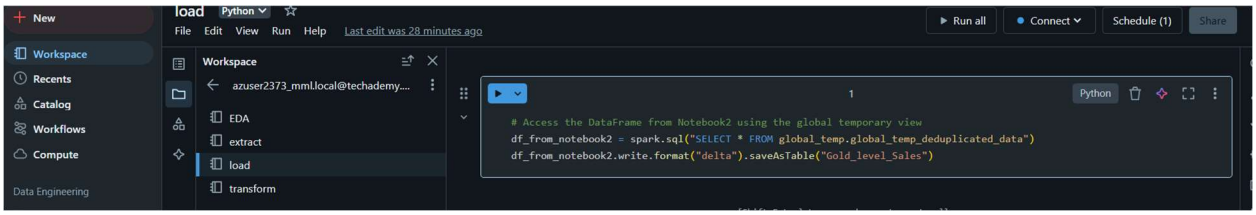


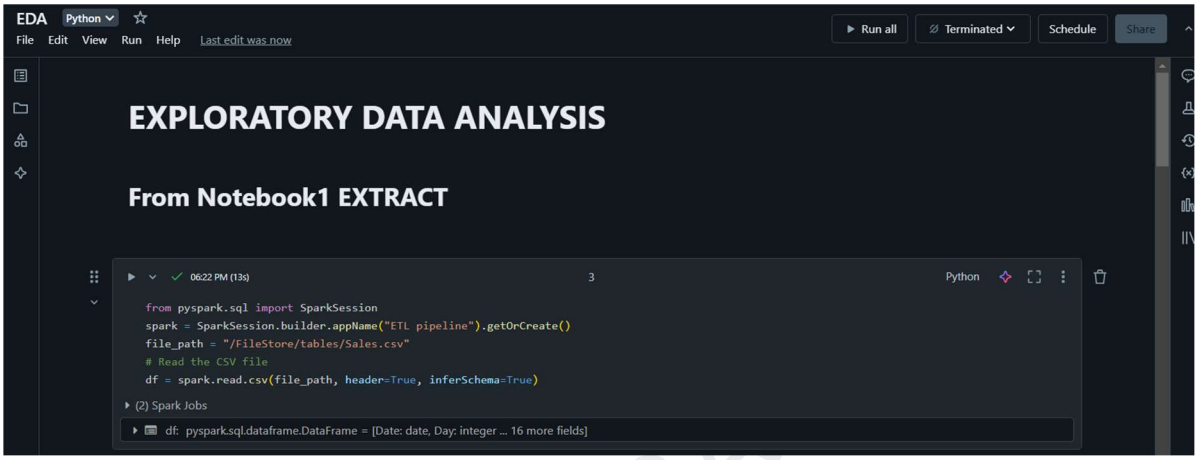
Data from notebook1 – Extract undergoes transformation.

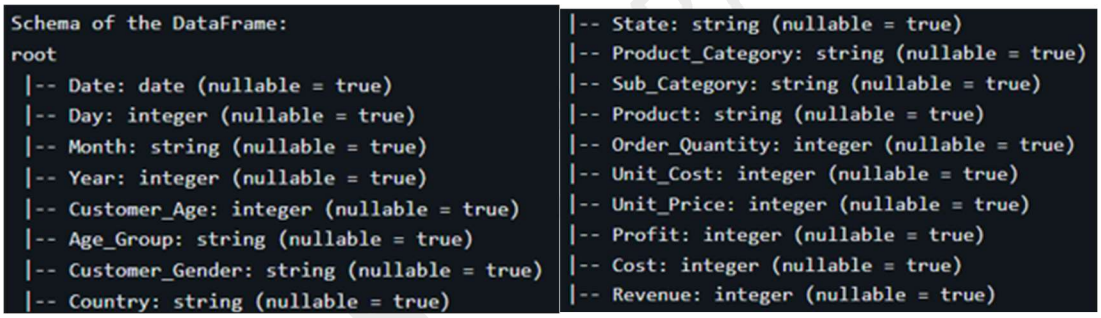
1. Drop Date, Day, Customer_age columns.
2. Drop duplicate values.

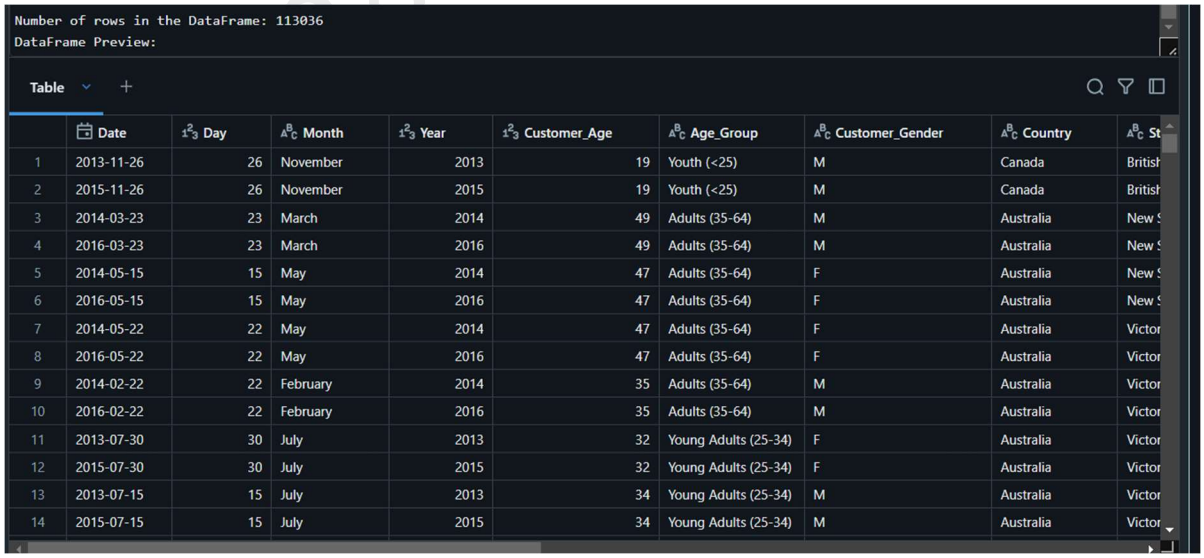


Finally the data is loaded as delta table.









	Date	Day	Month	Year	Customer_Age	Age_Group	Customer_Gender	Country	St
1	2013-11-26	26	November	2013	19	Youth (<25)	M	Canada	British
2	2015-11-26	26	November	2015	19	Youth (<25)	M	Canada	British
3	2014-03-23	23	March	2014	49	Adults (35-64)	M	Australia	New S
4	2016-03-23	23	March	2016	49	Adults (35-64)	M	Australia	New S
5	2014-05-15	15	May	2014	47	Adults (35-64)	F	Australia	New S
6	2016-05-15	15	May	2016	47	Adults (35-64)	F	Australia	New S
7	2014-05-22	22	May	2014	47	Adults (35-64)	F	Australia	Victor
8	2016-05-22	22	May	2016	47	Adults (35-64)	F	Australia	Victor
9	2014-02-22	22	February	2014	35	Adults (35-64)	M	Australia	Victor
10	2016-02-22	22	February	2016	35	Adults (35-64)	M	Australia	Victor
11	2013-07-30	30	July	2013	32	Young Adults (25-34)	F	Australia	Victor
12	2015-07-30	30	July	2015	32	Young Adults (25-34)	F	Australia	Victor
13	2013-07-15	15	July	2013	34	Young Adults (25-34)	M	Australia	Victor
14	2015-07-15	15	July	2015	34	Young Adults (25-34)	M	Australia	Victor

The transformed data doesn't have Date, Day, Customer_age features.

Number of records before transformation is 113036

Number of records after transformation is 100409

```
gold_level_sales_df: pyspark.sql.dataframe.DataFrame = [Month: st
root
 |-- Month: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Age_Group: string (nullable = true)
 |-- Customer_Gender: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- State: string (nullable = true)
 |-- Product_Category: string (nullable = true)
 |-- Sub_Category: string (nullable = true)
 |-- Product: string (nullable = true)
 |-- Order_Quantity: integer (nullable = true)
 |-- Unit_Cost: integer (nullable = true)
 |-- Unit_Price: integer (nullable = true)
 |-- Profit: integer (nullable = true)
 |-- Cost: integer (nullable = true)
 |-- Revenue: integer (nullable = true)

Number of records in Gold_level_Sales: 100409
```

DataFrame Preview:

Table

	Month	Year	Age_Group	Customer_Gender	Country	State	Product_Category	Sub_Cate
1	May	2014	Young Adults (25-34)	M	United States	California	Accessories	Bike Racks
2	August	2013	Adults (35-64)	M	United States	California	Accessories	Bike Racks
3	January	2016	Adults (35-64)	M	Canada	British Columbia	Accessories	Bike Racks
4	February	2014	Adults (35-64)	F	Australia	New South Wales	Accessories	Bike Racks
5	January	2016	Adults (35-64)	M	Canada	British Columbia	Accessories	Bike Racks
6	May	2014	Adults (35-64)	F	United Kingdom	England	Accessories	Bike Racks
7	June	2016	Adults (35-64)	M	United States	California	Accessories	Bike Racks
8	November	2015	Adults (35-64)	F	Canada	British Columbia	Accessories	Bike Stands
9	May	2016	Adults (35-64)	M	United States	California	Accessories	Bike Stands
10	April	2016	Youth (<25)	M	France	Nord	Accessories	Bike Stands
11	May	2014	Adults (35-64)	F	United States	Washington	Accessories	Bike Stands
12	November	2013	Adults (35-64)	M	Australia	Victoria	Accessories	Bottles and C
13	July	2013	Adults (35-64)	M	France	Seine (Paris)	Accessories	Bottles and C
14	November	2015	Young Adults (25-34)	F	Canada	British Columbia	Accessories	Bottles and C