PySpark project (using databricks)

# Business Requirements

Clients have data in csv format not properly organized, Business wants to grow the business and for that we must clean the data and process and build certain KPI for the business.

We have two files:

1.customer related data

2.product related data

# KPI

* Total amount spent by each customer
* Total amount spent by each category
* Total amount of sales each month
* Yearly sales
* Quarterly sales
* Total number of orders by each category
* Top 5 ordered items
* Top ordered items
* Frequency of customer visited
* Total sales by each country
* Total sales by order source

# Dataset

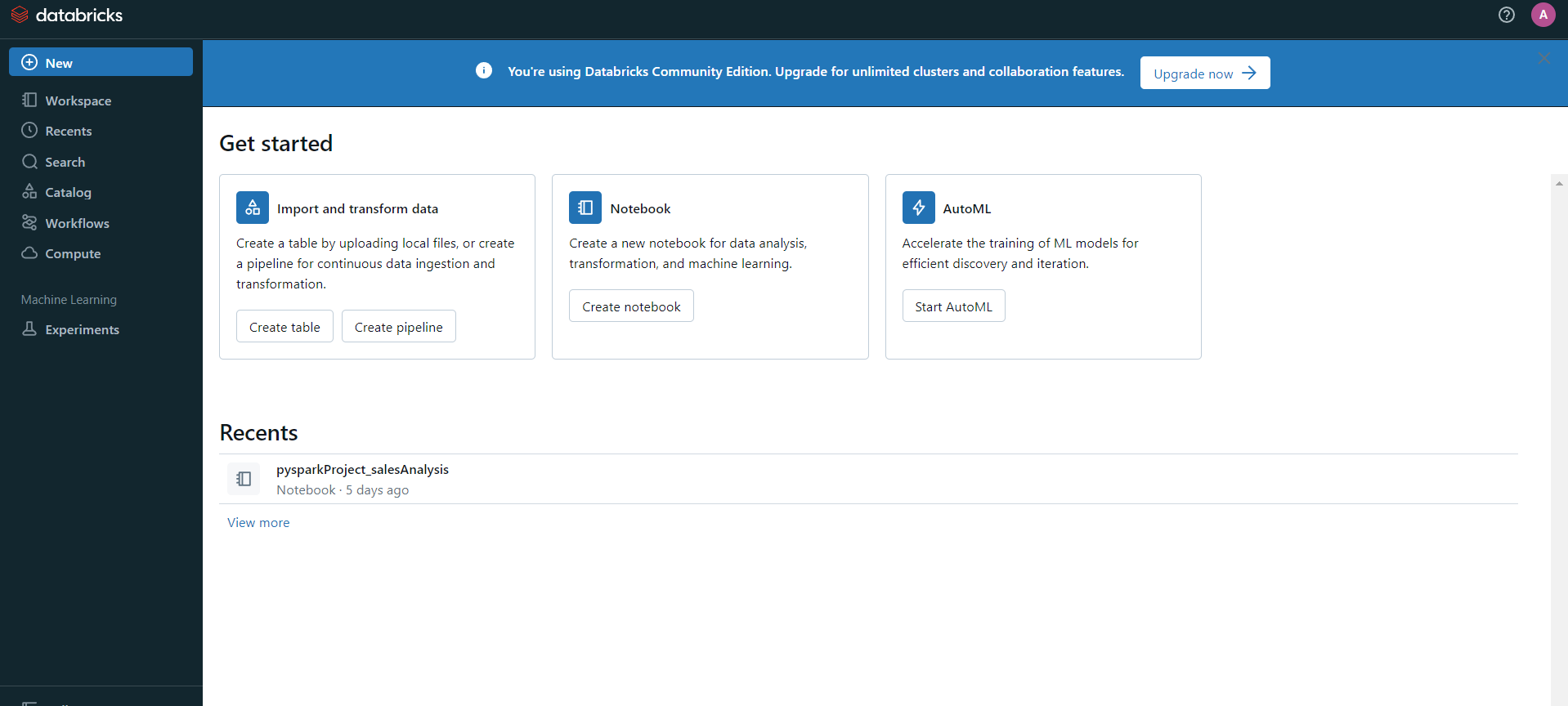
Customer data

* Product\_id
* Customer\_id
* Order\_date
* Location
* Source\_order

Product data

* Product\_id
* Product\_name
* Price

# Home page of databricks



# Compute

A screenshot of a computer

Description automatically generated

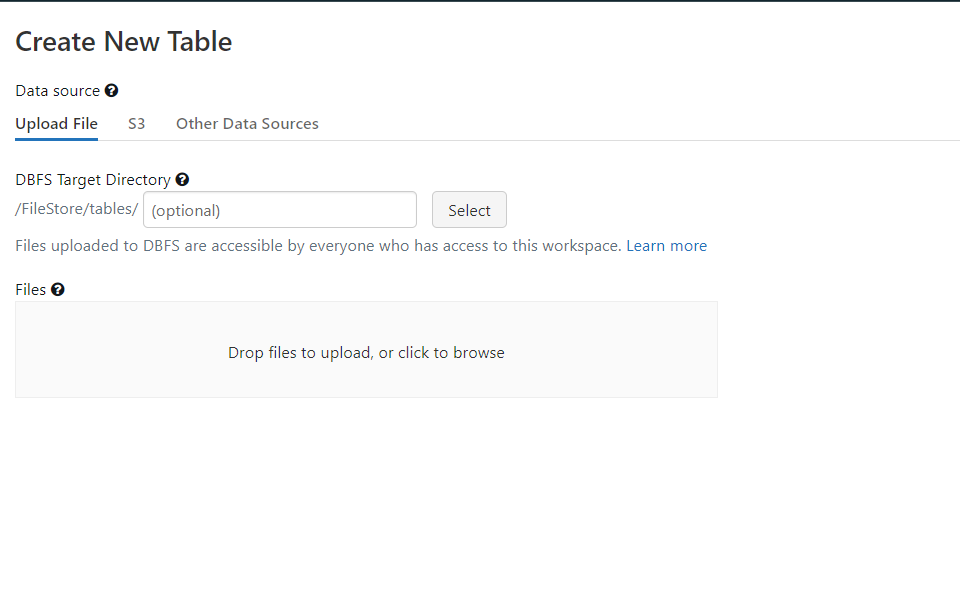
We will create cluster using compute for utilising the notebook for operating PySpark.

# Notebook

Now we will create jupyter notebook in the environment of databricks.A screenshot of a computer

Description automatically generated

# Tables



We will upload tables from our device to databricks and use them in our notebook.

# Dashboard

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

This is the resultant dashboard which is implemented using jupyter notebook in databricks.