assignment-1-6

July 14, 2025

[0]:
```
%pip install pandas
%restart_python
```

```
Requirement already satisfied: pandas in
/databricks/python3/lib/python3.11/site-packages (1.5.3)
Requirement already satisfied: python-dateutil>=2.8.1 in
/databricks/python3/lib/python3.11/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/databricks/python3/lib/python3.11/site-packages (from pandas) (2022.7)
Requirement already satisfied: numpy>=1.21.0 in
/databricks/python3/lib/python3.11/site-packages (from pandas) (1.23.5)
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from
python-dateutil>=2.8.1->pandas) (1.16.0)
Note: you may need to restart the kernel using %restart_python ordbutils.library.restartPython
```

[0]:
```
# 1. Data Ingestion
df = spark.read.csv("/Volumes/workspace/default/skit_assignment/Sales.csv",␣
  ↪header=True, inferSchema=True)
df.display()
```

[0]:
```
# 2. Data Exploration(First Few Rows)
df_few = df.limit(5)
df_few.display()
```

[0]:
```
# 4. Data Filtering(Transaction of Sales more than 1000)
df_filtered = df.filter(df.SALES > 1000)
df_filtered.display()
```

[0]:
```
# 5. Delta Table Management (Delta Table and mode=append)
df_filtered.write.format("delta").mode("append").saveAsTable("default.
  ↪agg_sales_data")
```

[0]:
```
%sql
TRUNCATE TABLE default.agg_sales_data
```

[0]:
```
# 6. Version Control (Check the history of the delta table)
df_version = spark.sql("DESCRIBE HISTORY default.agg_sales_data")
```

```
df_version.display()
```

[0]:
```
# 6. Version Control (Read data from a specefic version)
version = 1
df_specefic_version = spark.read.format("delta").option("versionAsOf", version).
  ↪table("default.agg_sales_data")
df_specefic_version.display()
```