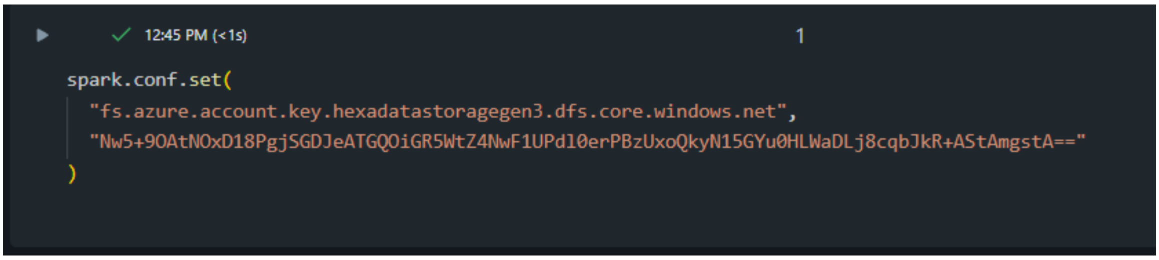
CASE STUDY – 1 (Azure Databricks)

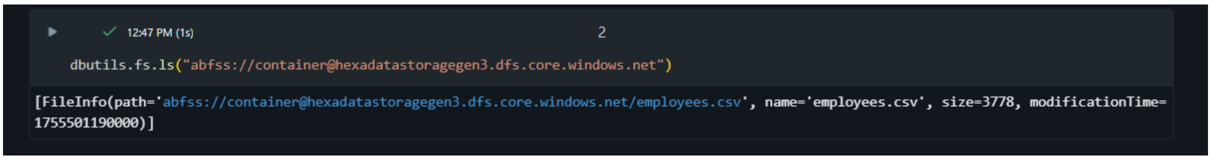
TRAINEE NAME: Swathi Baskaran

Connecting container with Azure Databricks



```
spark.conf.set(
  "fs.azure.account.key.hexadatastoragegen3.dfs.core.windows.net",
  "Nw5+9OAtNOxD18PgjSGDJeATGQOiGR5WtZ4NwF1UPdl0erPBzUxoQkyN15GYu0HLWaDLj8cqbJkR+AStAmgstA=="
)
```

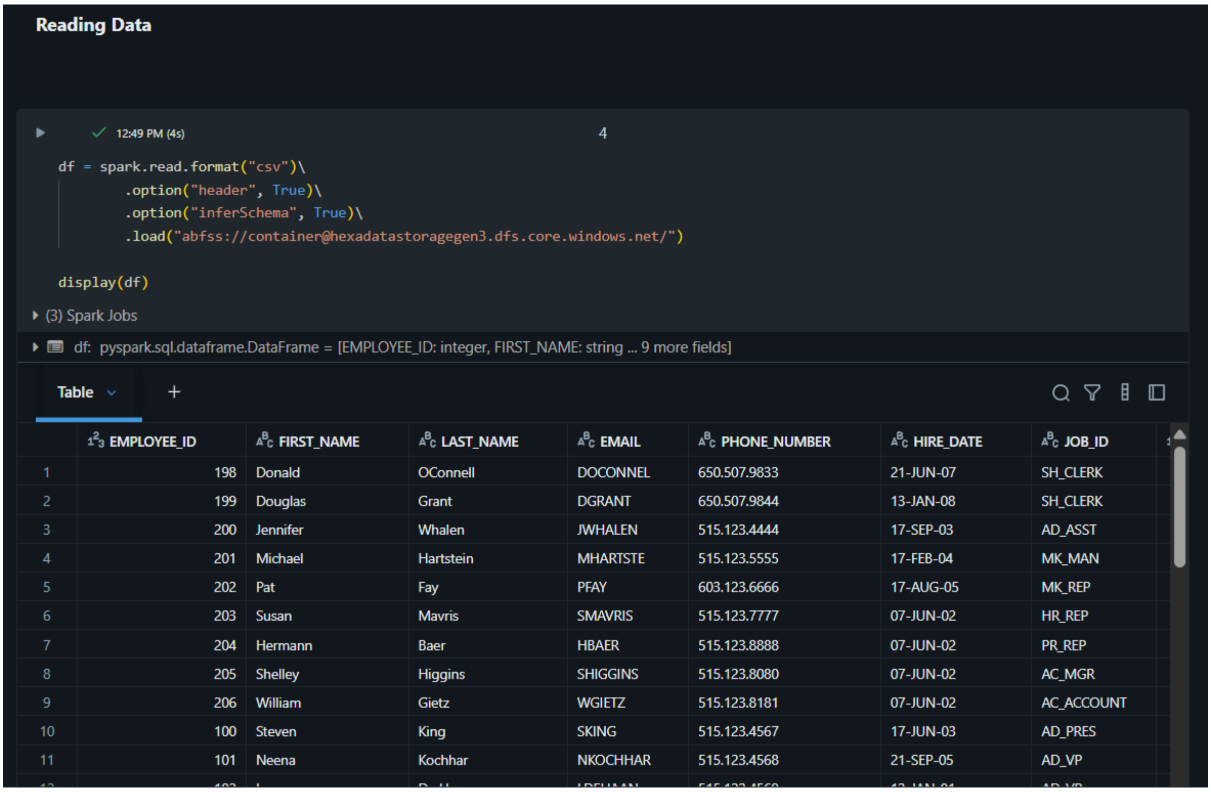
Checking the files that are present in the given location



```
dbutils.fs.ls("abfss://container@hexadatastoragegen3.dfs.core.windows.net")
```

```
[FileInfo(path='abfss://container@hexadatastoragegen3.dfs.core.windows.net/employees.csv', name='employees.csv', size=3778, modificationTime=
1755501190000)]
```

Reading Data from ADLS



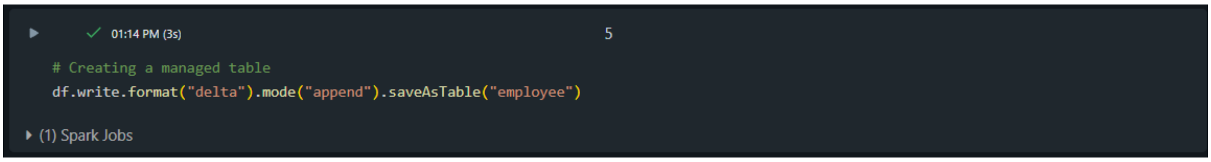
```
df = spark.read.format("csv")\
        .option("header", True)\
        .option("inferSchema", True)\
        .load("abfss://container@hexadatastoragegen3.dfs.core.windows.net/")

display(df)
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [EMPLOYEE_ID: integer, FIRST_NAME: string ... 9 more fields]

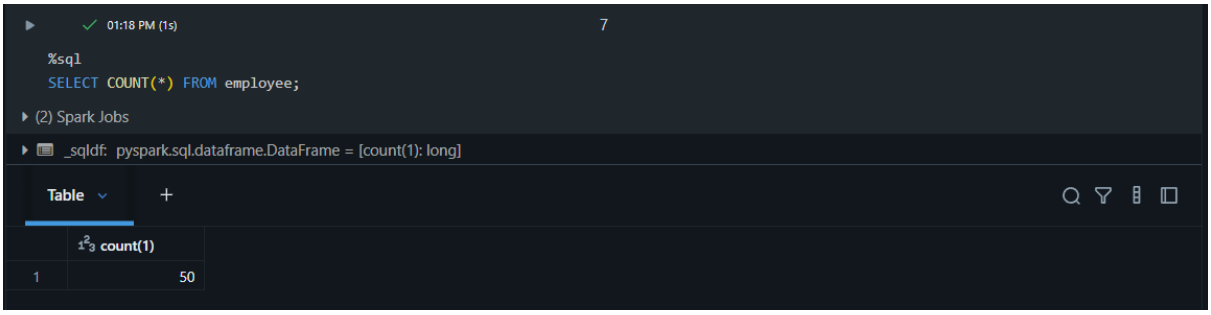
	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID
1	198	Donald	OConnell	DOCONNEL	650.507.9833	21-JUN-07	SH_CLERK
2	199	Douglas	Grant	DGRANT	650.507.9844	13-JAN-08	SH_CLERK
3	200	Jennifer	Whalen	JWHALEN	515.123.4444	17-SEP-03	AD_ASST
4	201	Michael	Hartstein	MHARTSTE	515.123.5555	17-FEB-04	MK_MAN
5	202	Pat	Fay	PFAY	603.123.6666	17-AUG-05	MK_REP
6	203	Susan	Mavris	SMAVRIS	515.123.7777	07-JUN-02	HR_REP
7	204	Hermann	Baer	HBAER	515.123.8888	07-JUN-02	PR_REP
8	205	Shelley	Higgins	SHIGGINS	515.123.8080	07-JUN-02	AC_MGR
9	206	William	Gietz	WGIETZ	515.123.8181	07-JUN-02	AC_ACCOUNT
10	100	Steven	King	SKING	515.123.4567	17-JUN-03	AD_PRES
11	101	Neena	Kochhar	NKOCHHAR	515.123.4568	21-SEP-05	AD_VP

Storing data in Delta Format



```
# Creating a managed table
df.write.format("delta").mode("append").saveAsTable("employee")
```

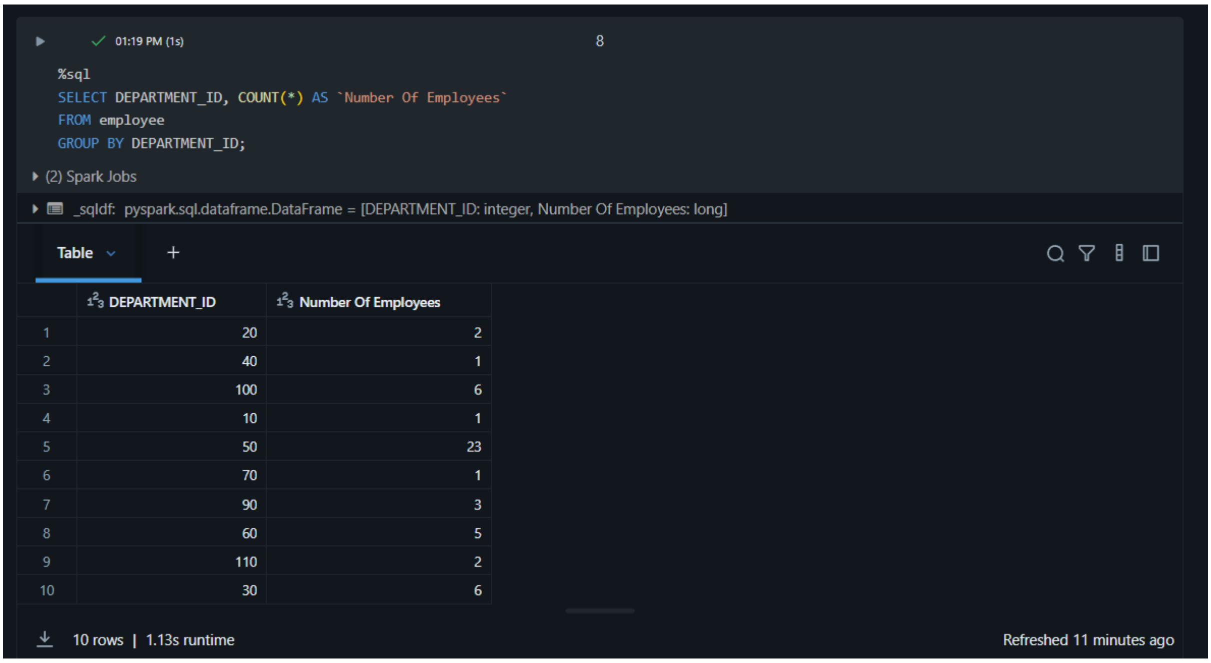
Querying Delta Tables



```
%sql
SELECT COUNT(*) FROM employee;
```

```
_sqldf:  pyspark.sql.dataframe.DataFrame = [count(1): long]
```

	count(1)
1	50



```
%sql
SELECT DEPARTMENT_ID, COUNT(*) AS `Number Of Employees`
FROM employee
GROUP BY DEPARTMENT_ID;
```

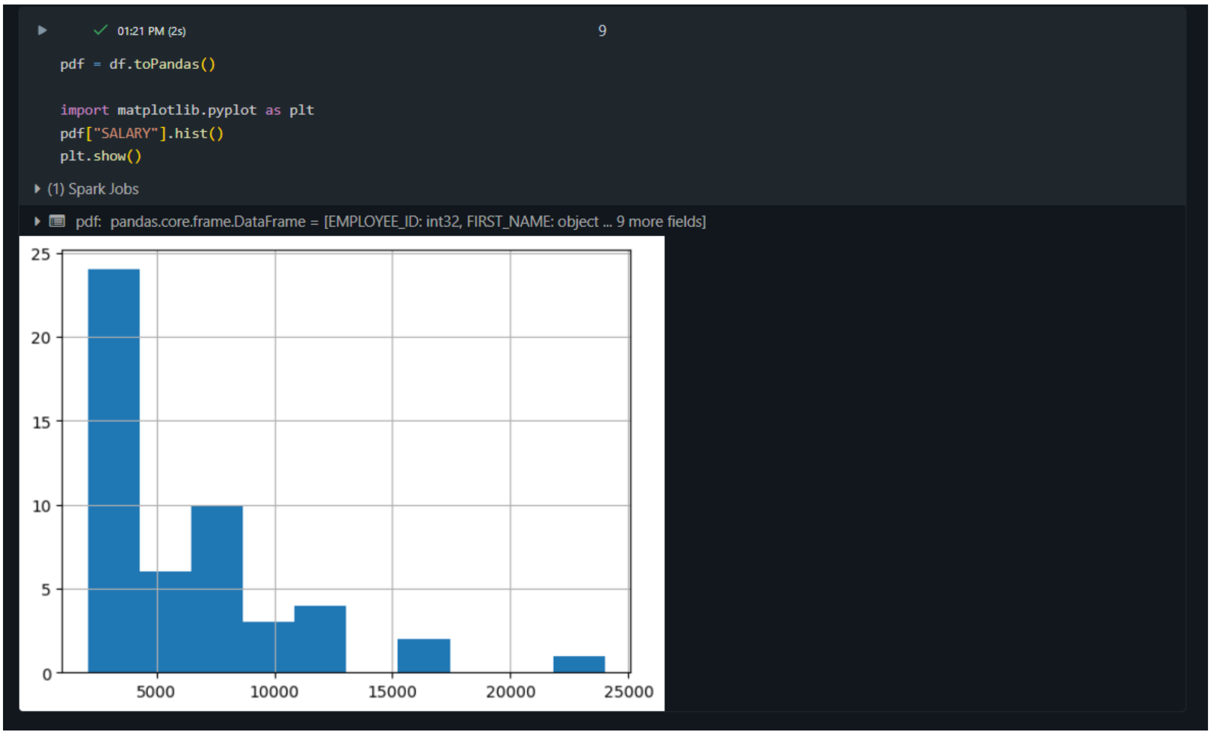
```
_sqldf:  pyspark.sql.dataframe.DataFrame = [DEPARTMENT_ID: integer, Number Of Employees: long]
```

	DEPARTMENT_ID	Number Of Employees
1	20	2
2	40	1
3	100	6
4	10	1
5	50	23
6	70	1
7	90	3
8	60	5
9	110	2
10	30	6

10 rows | 1.13s runtime

Refreshed 11 minutes ago

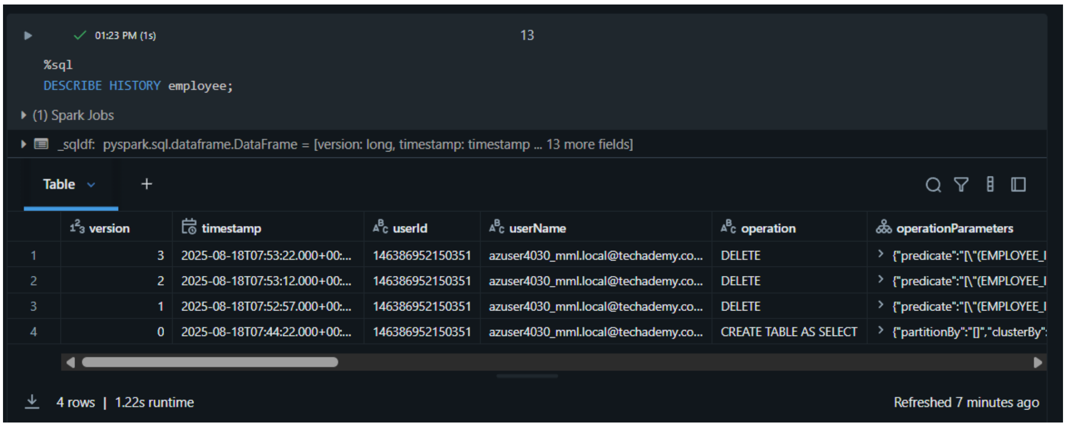
Visualizing Data



```
pdf = df.toPandas()

import matplotlib.pyplot as plt
pdf["SALARY"].hist()
plt.show()
```

```
(1) Spark Jobs
pdf:  pandas.core.frame.DataFrame = [EMPLOYEE_ID: int32, FIRST_NAME: object ... 9 more fields]
```

Time Travel



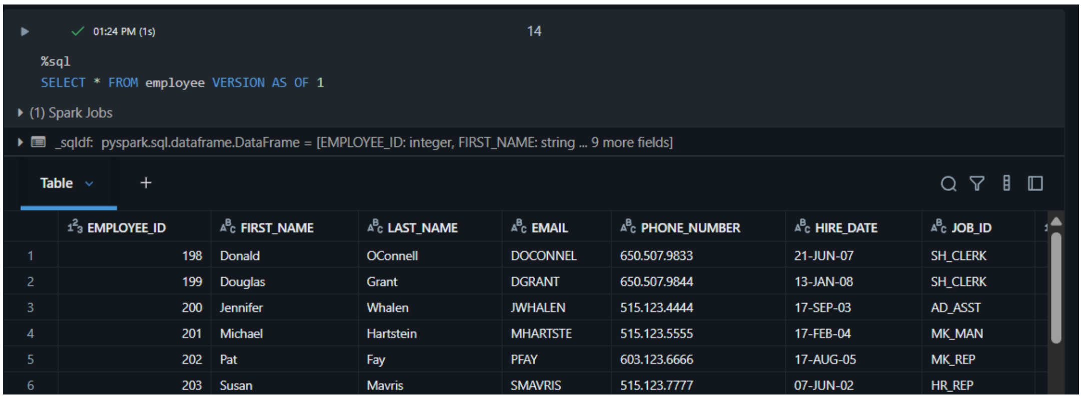
```
%sql
DESCRIBE HISTORY employee;
```

```
(1) Spark Jobs
_sqldf:  pyspark.sql.dataframe.DataFrame = [version: long, timestamp: timestamp ... 13 more fields]
```

	version	timestamp	userId	userName	operation	operationParameters
1	3	2025-08-18T07:53:22.000+00:...	146386952150351	azuser4030_mml.local@techademy.co...	DELETE	{"predicate":"[\"(EMPLOYEE_I
2	2	2025-08-18T07:53:12.000+00:...	146386952150351	azuser4030_mml.local@techademy.co...	DELETE	{"predicate":"[\"(EMPLOYEE_I
3	1	2025-08-18T07:52:57.000+00:...	146386952150351	azuser4030_mml.local@techademy.co...	DELETE	{"predicate":"[\"(EMPLOYEE_I
4	0	2025-08-18T07:44:22.000+00:...	146386952150351	azuser4030_mml.local@techademy.co...	CREATE TABLE AS SELECT	{"partitionBy":"[]","clusterBy":

4 rows | 1.22s runtime

Refreshed 7 minutes ago

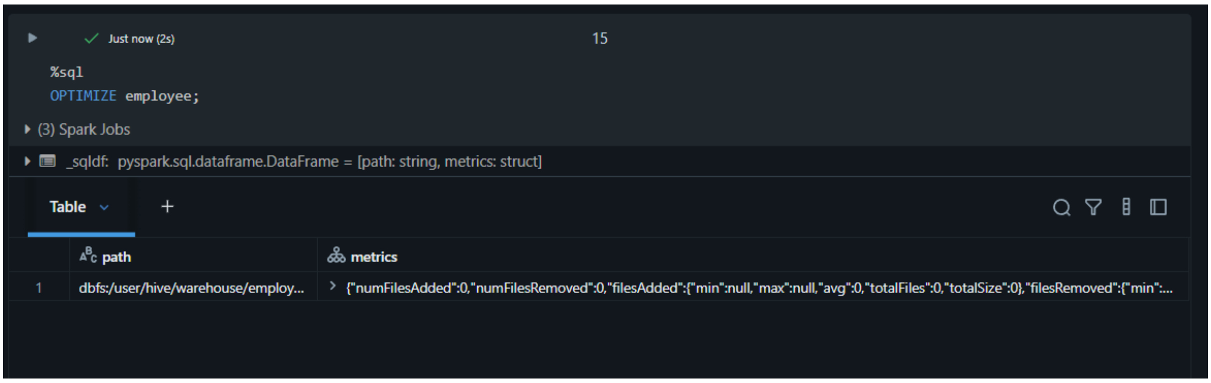


```
%sql
SELECT * FROM employee VERSION AS OF 1
```

```
(1) Spark Jobs
_sqldf:  pyspark.sql.dataframe.DataFrame = [EMPLOYEE_ID: integer, FIRST_NAME: string ... 9 more fields]
```

	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID
1	198	Donald	OConnell	DOCONNEL	650.507.9833	21-JUN-07	SH_CLERK
2	199	Douglas	Grant	DGRANT	650.507.9844	13-JAN-08	SH_CLERK
3	200	Jennifer	Whalen	JWHALEN	515.123.4444	17-SEP-03	AD_ASST
4	201	Michael	Hartstein	MHARTSTE	515.123.5555	17-FEB-04	MK_MAN
5	202	Pat	Fay	PFAY	603.123.6666	17-AUG-05	MK_REP
6	203	Susan	Mavris	SMAVRIS	515.123.7777	07-JUN-02	HR_REP

Optimization



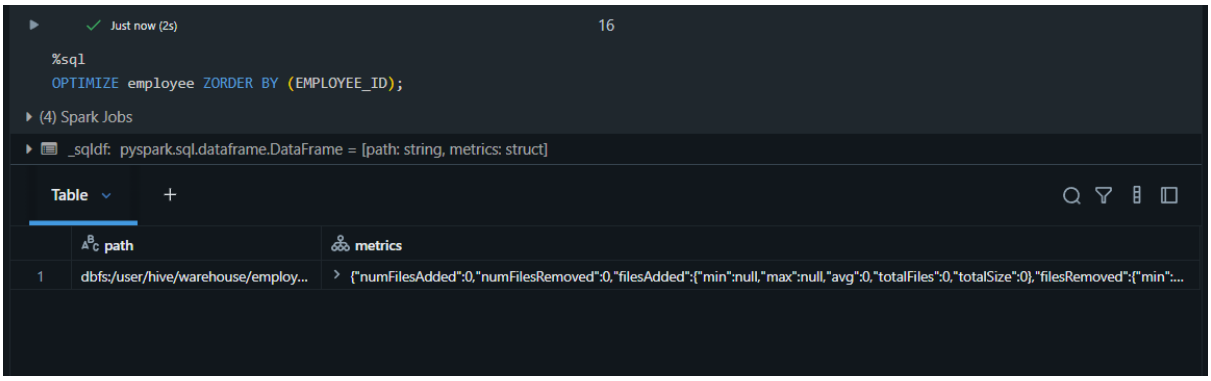
```
%sql
OPTIMIZE employee;
```

(3) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [path: string, metrics: struct]

	path	metrics
1	dbfs:/user/hive/warehouse/employ...	{"numFilesAdded":0,"numFilesRemoved":0,"filesAdded":{"min":null,"max":null,"avg":0,"totalFiles":0,"totalSize":0},"filesRemoved":{"min":...

ZOrder Optimization



```
%sql
OPTIMIZE employee ZORDER BY (EMPLOYEE_ID);
```

(4) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [path: string, metrics: struct]

	path	metrics
1	dbfs:/user/hive/warehouse/employ...	{"numFilesAdded":0,"numFilesRemoved":0,"filesAdded":{"min":null,"max":null,"avg":0,"totalFiles":0,"totalSize":0},"filesRemoved":{"min":...