# Case Study - Azure

## EDA analysis extract adls with Azure databricks and run queries of delta tables

Since I have only Databricks Community Edition (which does not support Azure ADLS mounting), here's exactly what I can do to do My case study — simulating the ADLS part by uploading data directly — and then run EDA and Delta queries.

### Step 1: Prepare your dataset (CSV file)

id,name,age,salary

1,John,28,50000

2,Mary,35,60000

3,Sam,22,45000

4,Lisa,40,70000

### Step 2: Upload this CSV file to Databricks Community Edition

- Login to community.cloud.databricks.com
- Click Data tab on left sidebar
- Click Add Data → Upload File
- Select your data.csv and upload

### Step 3: Create a new notebook and start a cluster

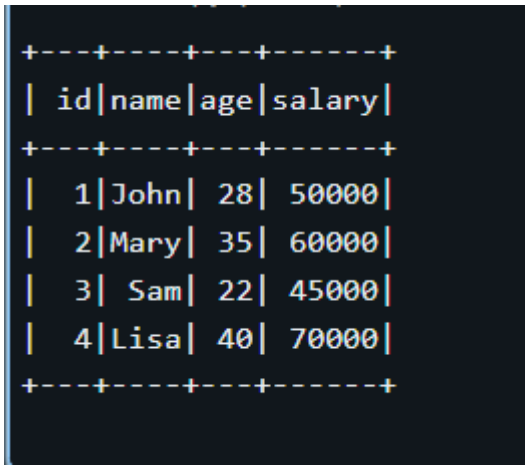
- Go to Workspace > your user folder
- Click Create > Notebook
- Name it EDA_Case_Study

- Choose language Python
- Attach to a running cluster or start one

### Step 4: Read the CSV file in your notebook

df = spark.read.format("csv").option("header", "true").load("/Volumes/workspace/default/emplo/data.csv")

df.show()

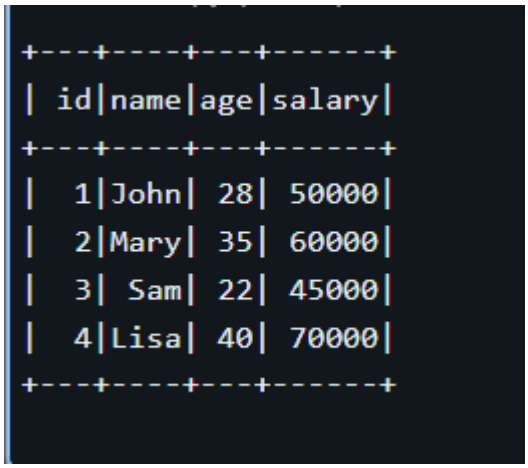


```
+---+----+---+------+
| id|name|age|salary|
+---+----+---+------+
|  1|John| 28| 50000|
|  2|Mary| 35| 60000|
|  3| Sam| 22| 45000|
|  4|Lisa| 40| 70000|
+---+----+---+------+
```

### Step 5: Convert this Data Frame to a Delta Table

df.write.format("delta").mode("overwrite").saveAsTable("data_delta")

df.show()



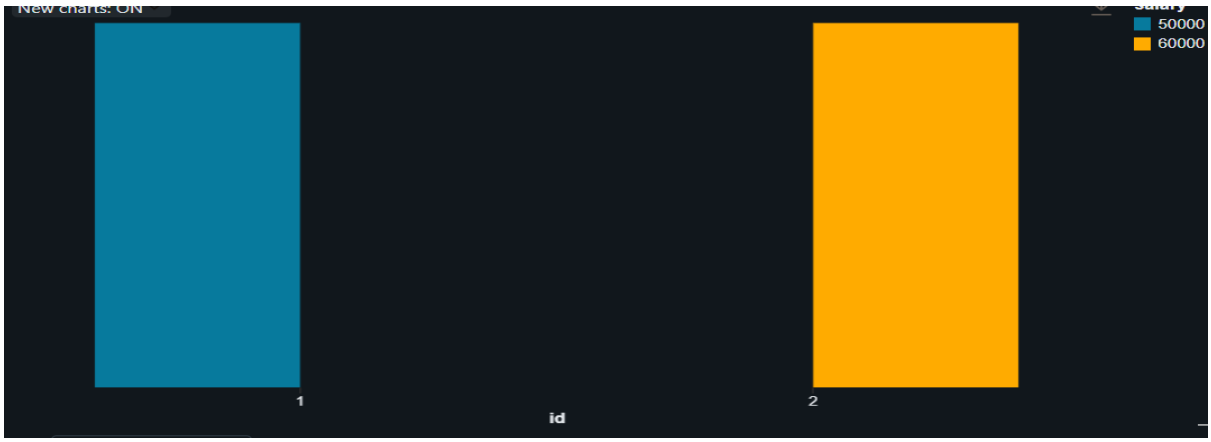
```
+---+----+---+------+
| id|name|age|salary|
+---+----+---+------+
|  1|John| 28| 50000|
|  2|Mary| 35| 60000|
|  3| Sam| 22| 45000|
|  4|Lisa| 40| 70000|
+---+----+---+------+
```

## Step 6: Run SQL queries on Delta table for EDA

## -- Show first 2 records

SELECT * FROM data_delta LIMIT 2;

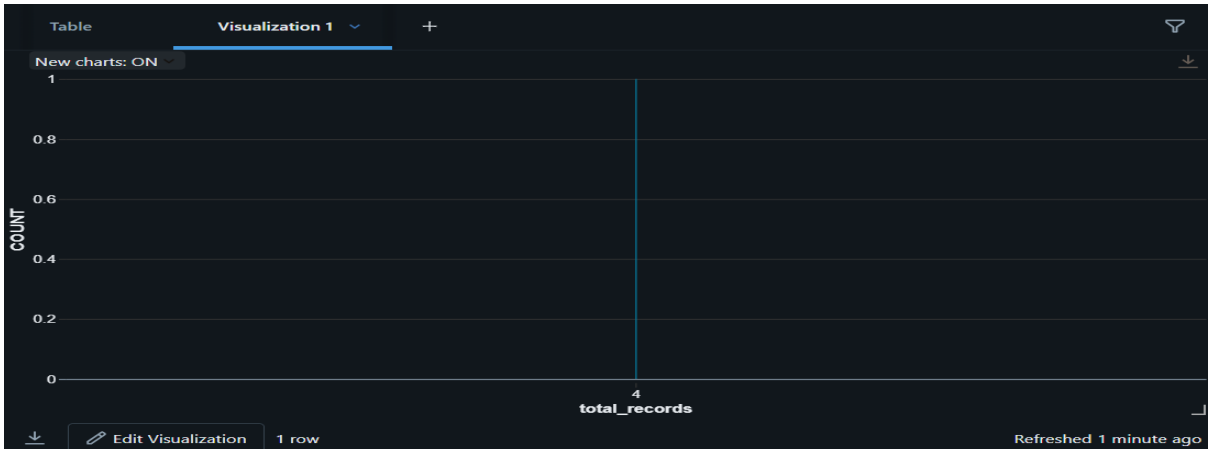
| | id | name | age | salary |
|---|---|---|---|---|
| 1 | 1 | John | 28 | 50000 |
| 2 | 2 | Mary | 35 | 60000 |




## -- Count total records

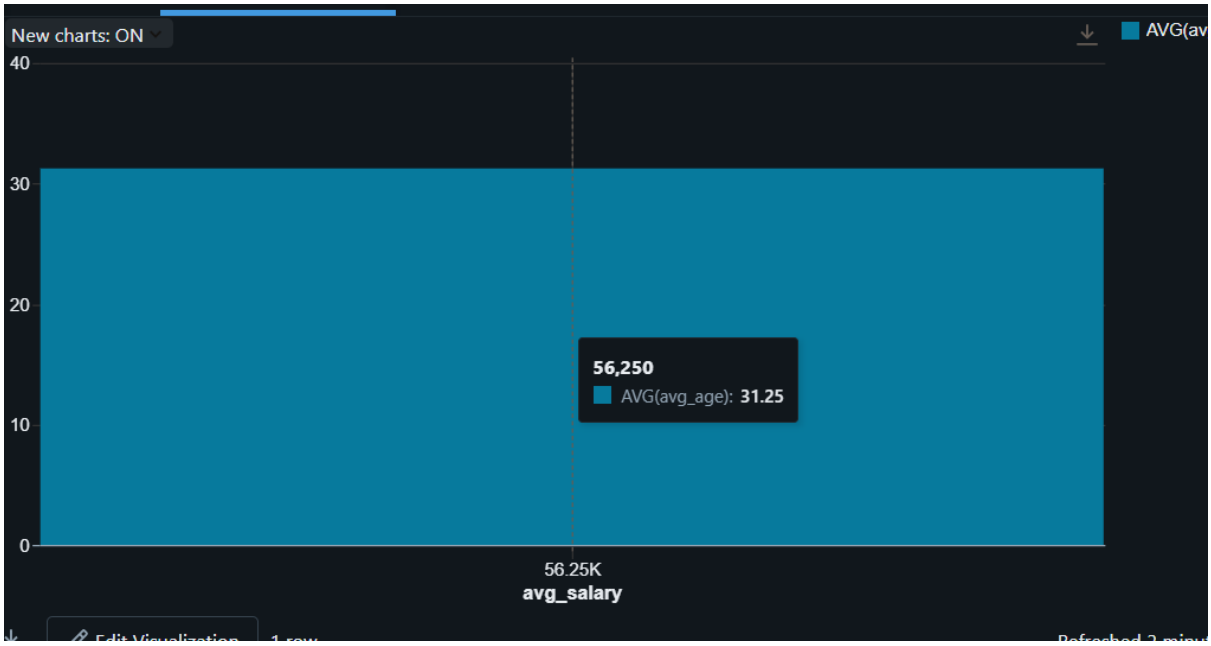
SELECT COUNT(*) AS total_records FROM data_delta;

| | total_records |
|---|---|
| 1 | 4 |



```
-- Calculate average of numeric columns (example: age, salary)
SELECT
  AVG(CAST(age AS INT)) AS avg_age,
  AVG(CAST(salary AS INT)) AS avg_salary
FROM data_delta;
```

| | 1.2 avg_age | 1.2 avg_salary |
|---|---|---|
| 1 | 31.25 | 56250 |




## Conclusion

This case study demonstrated how to perform EDA using Databricks Community Edition by uploading data manually to simulate extraction from Azure Data Lake Storage. The data was loaded into Spark, saved as Delta tables, and analyzed with SQL queries. Despite the lack of direct ADLS access, all key steps of the analysis were completed successfully, showcasing effective use of Delta Lake and Databricks for data exploration.