**Case Study**

**EDA Analysis on Sales Data from ADLS using Azure Databricks**

## **1. Objective**

The objective of this case study was to:

* Extract a dataset stored in Azure Data Lake Storage Gen2 (ADLS).
* Perform Exploratory Data Analysis (EDA) using Azure Databricks and PySpark.
* Store the processed data as a Delta table and run SQL queries for insights.

## **2. Data Source**

* Storage account: hexdatastoragegen2
* Container: datacontainer
* File path: /sales/sales\_data.csv
* Data format: CSV file containing sales transactions with columns:  
   order\_id, order\_date, store, product, category, quantity, price, customer\_age.

## **3. Data Extraction**

The dataset was read directly from ADLS into Databricks using Storage Account Key authentication.

Code snippet:

spark.conf.set("fs.azure.account.key.hexdatastoragegen2.dfs.core.windows.net", "<storage\_account\_key>")

df = spark.read.format("csv") \

.option("header", "true") \

.option("inferSchema", "true") \

.load("abfss://[datacontainer@hexdatastoragegen2.dfs.core.windows.net](mailto:datacontainer@hexdatastoragegen2.dfs.core.windows.net)/sales/sales\_data.csv")

## **4. Exploratory Data Analysis (EDA)**

EDA was conducted in PySpark with the following steps:

1. Schema inspection and preview of records using printSchema() and show().
2. Summary statistics for numerical columns with describe().
3. Row count and missing value check.
4. Aggregations and insights:  
   * Total quantity sold per category.
   * Average price per category.
   * Total sales per store.

Example aggregation:

df.groupBy("category").sum("quantity").show()

## **5. Delta Table Creation**

The DataFrame was saved as a Delta table for running SQL queries efficiently.

Code snippet:

df.write.format("delta").mode("overwrite").saveAsTable("sales\_delta")

## **6. SQL Queries on Delta Table**

SQL queries were executed directly in Databricks SQL cells.

Example 1 — Total Sales by Store:

SELECT store, SUM(price) AS total\_sales

FROM sales\_delta

GROUP BY store

ORDER BY total\_sales DESC;

## **7. Key Insights**

* The dataset contains multiple categories of products, with Electronics contributing significantly to sales.
* Certain stores have consistently higher sales revenue compared to others.
* Delta tables in Databricks allow faster queries and better data management.

## **8. Conclusion**

This case study demonstrated how to:

* Connect Azure Databricks to ADLS Gen2 using Storage Account Key authentication.
* Perform EDA using PySpark.
* Save the results in Delta format for optimized queries.