1. READ THE DATASET IN DATABRICKS COMMUNITY

Databricks provides a collaborative environment for big data analytics based on Apache Spark.**Upload Dataset:**

- Go to the Databricks workspace.
- Create a new notebook or open an existing one.
- Upload your dataset to Databricks. You can do this by clicking on the "Data" tab on the left sidebar, then selecting "Add Data" and uploading your file.

Read Dataset:

- Use Spark APIs to read the dataset. The most common method is to use spark.read with the appropriate file format.

2. HOW MANY TYPES OF MODES WE HAVE IN SPARK?

There are three types of Mode we have

Failfast:this option is used when reading data from a structured source. In the context of CSV, it means that if Spark encounters a malformed or corrupted record, it fails immediately and does not proceed with reading the rest of the data.

Dropmalformed: When reading data, you can use the "dropmalformed" option to drop any records that are malformed or corrupted instead of failing the entire job. This allows the job to continue processing the remaining valid records.

Permissive: The "permissive" option is used to treat malformed records as null values. It allows Spark to read as much data as possible, even if some records do not conform to the expected structure. Malformed records are treated as null, and the job continues processing.

3. WHAT IS CLUSTER IN SPARK?

A Databricks cluster is a collection of computation resources and configurations that can run data engineering, data science, and data analytics workloads. A cluster is a group of machines called "nodes" that work together to process data and queries efficiently.

4. WHAT IS TABLE IN SPARK ?

A table refers to a structured set of data organized in a tabular form, similar to a table in a relational database. Databricks provides a feature called Databricks Tables, which are managed tables that make it easy to store, query, and analyze structured data in a distributed environment using Apache Spark.

Data Storage:

- Tables in Databricks can be used to store data in various formats such as Parquet, Delta Lake, CSV, JSON, and more. These tables are often used to persist structured data for analysis and reporting.

Managed and External Tables:

- Databricks supports both managed tables and external tables. Managed tables have their data managed by Databricks, and the underlying files are stored in DBFS (Databricks File System). External tables, on the other hand, reference data that is stored outside of Databricks (e.g., in an existing storage system), and Databricks manages metadata only.

- Schema Management:
  - Tables in Databricks have a schema that defines the structure of the data, including column names, data types, and other metadata. The schema ensures that data is organized in a consistent manner.
- SQL Queries:
  - Users can perform SQL queries on Databricks Tables using Spark SQL. This allows for efficient querying and analysis of structured data, making it easy to apply SQL operations on the tables.
- Integration with Databricks Notebooks:
  - Tables seamlessly integrate with Databricks Notebooks, allowing users to reference and analyze data stored in tables directly within their notebook environment. This integration facilitates collaborative data analysis and exploration
- Performance Optimization:
  - Databricks Tables are optimized for performance through features such as indexing and caching. This helps improve query performance, especially for large datasets.

5. WHAT WOULD YOU DO IF YOU WANT TO SHOW THE HEADER WHILE SHOWING UP 5 RECORDS OF TABLE? WRITE THE CODE

WE DO coding attach screen shot

6. WHAT IS COUNT ? PERFORM IN SPARK

The `count()` function in PySpark returns the total number of rows in a DataFrame. The `count` method can be used to count the number of rows in a DataFrame.

7. WHAT IS GROUP BY ? PERFORM IN SPARK

the `groupBy()` function groups data in a Spark DataFrame or RDD based on one or more columns. It returns a GroupedData object that can be used to perform aggregation operations.