

SPARK TEST

>Dipenkumar Padhiyar

1.READ THE DATASET IN DATABRICKS COMMUNITY

> In the Databricks Community, we can read a dataset using Spark APIs. For example, if you have a CSV file stored in DBFS (Databricks File System), we can use the following code to read

```
file_path = "/FileStore/tables/iris.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)
display(df)
```

```
1 file_path = "/FileStore/tables/iris.csv"
2 df = spark.read.csv(file_path, header=True, inferSchema=True)
3 display(df)
```

▶ (3) Spark Jobs


▶  df: pyspark.sql.dataframe.DataFrame = [sepal.length: double, sepal.width: double ... 3 more fields]

Table ▾ +

| | sepal.length ▲ | sepal.width ▲ | petal.length ▲ | petal.width ▲ | variety ▲ | |
|---|----------------|---------------|----------------|---------------|-----------|--|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Setosa | |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Setosa | |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Setosa | |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Setosa | |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Setosa | |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Setosa | |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Setosa | |

⬇ 150 rows | 2.23 seconds runtime

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

> There are three types of modes in Spark:

- Local Mode
- Cluster Mode
- Client Mode

3. WHAT IS CLUSTER IN SPARK?

> In Spark, a cluster is a group of machines (nodes) that work together to process data. It typically consists of a master node and one or more worker nodes.

4.WHAT IS TABLE IN SPARK?

> A table is a structured collection of data organized into rows and columns. Tables can be created from various data sources like Parquet files, CSV files, JSON files, or existing RDDs (Resilient Distributed Datasets). Spark provides APIs to manipulate and query tables using SQL or DataFrame operations.

Create New Table

Data source ?

Upload File S3 Other Data Sources

DBFS Target Directory ?

/FileStore/tables/ (optional) Select

Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

Files ?

Drop files to upload, or click to browse

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

> `df.show(5, truncate=False)`

```
1 df.show(5, truncate=False)
```

► (1) Spark Jobs

```
+-----+-----+-----+-----+-----+
|sepal.length|sepal.width|petal.length|petal.width|variety|
+-----+-----+-----+-----+-----+
|5.1         |3.5         |1.4         |0.2         |Setosa |
|4.9         |3.0         |1.4         |0.2         |Setosa |
|4.7         |3.2         |1.3         |0.2         |Setosa |
|4.6         |3.1         |1.5         |0.2         |Setosa |
|5.0         |3.6         |1.4         |0.2         |Setosa |
+-----+-----+-----+-----+-----+
only showing top 5 rows
```

6. WHAT IS COUNT? PERFORM IN SPARK

> In Spark, **count()** is an action that returns the number of elements in a DataFrame or RDD.
To perform a count operation in Spark, you can use:

`count = df.count()`

```
1 count = df.count()
```

```
2 count
```

► (2) Spark Jobs

150

7. WHAT IS GROUP BY? PERFORM IN SPARK

> `groupBy()` is a transformation that groups the data in a `DataFrame` based on the specified column(s).

```
grouped_df = df.groupBy("column_name").agg({"aggregating_column": "sum"})
```