

# 1. READ THE DATASET IN DATABRICKS COMMUNITY

The screenshot shows the Databricks Spark test interface. The code in the editor reads a CSV file from the Filestore and displays the first 5 rows of the resulting DataFrame. The output shows the first 5 rows of the Titanic dataset.

```
1 df1 = spark.read.format("csv")\
2   .option("header", "false")\
3   .option("inferSchema", "false")\
4   .option("mode", "FAILFAST")\
5   .load("dbfs:/FileStore/shared_uploads/vijaiey88@gmail.com/titanic-3.csv")
6
7 df1.show(5)
8
```

Command took 2.23 seconds -- by vijaiey88@gmail.com at 3/9/2024, 11:02:37 PM on My Cluster

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

Failfast: Throws an error when it meets corrupted records.

The screenshot shows the Databricks Spark test interface. The code in the editor reads a CSV file from the Filestore and displays the first 10 rows of the resulting DataFrame. The output shows an error due to corrupted records.

```
1 df1 = spark.read.format("csv")\
2   .option("header", "true")\
3   .option("inferSchema", "false")\
4   .option("mode", "DROPMALFORMED")\
5   .load("dbfs:/FileStore/shared_uploads/vijaiey88@gmail.com/titanic-3.csv")
6 df1.show(10)
```

Command took 1.65 seconds -- by vijaiey88@gmail.com at 3/9/2024, 11:48:18 PM on My Cluster

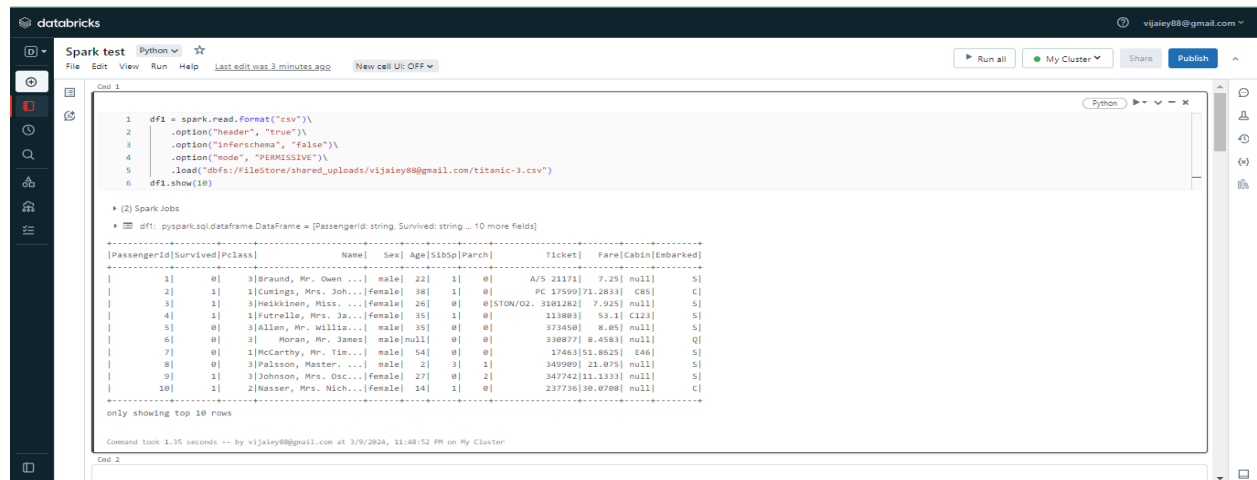
Dropmal formed: Ignores the whole corrupted records

The screenshot shows the Databricks Spark test interface. The code in the editor reads a CSV file from the Filestore and displays the first 5 rows of the resulting DataFrame. The output shows the first 5 rows of the Titanic dataset after dropping malformed records.

```
1 df1 = spark.read.format("csv")\
2   .option("header", "false")\
3   .option("inferSchema", "false")\
4   .option("mode", "FAILFAST")\
5   .load("dbfs:/FileStore/shared_uploads/vijaiey88@gmail.com/titanic-3.csv")
6
7 df1.show(5)
8
```

Command took 2.23 seconds -- by vijaiey88@gmail.com at 3/9/2024, 11:02:37 PM on My Cluster

Permissive: Set other fields to null when it meets corrupted records.



The screenshot shows a Databricks workspace with a Python script in a cell. The script reads a CSV file from a file store and displays the first 10 rows of the resulting DataFrame. The output shows a table with columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The first 10 rows of data are displayed.

```
1 df1 = spark.read.format("csv")\
2     .option("header", "true")\
3     .option("inferSchema", "false")\
4     .option("mode", "PERMISSIVE")\
5     .load("dbfs:/FileStore/shared_uploads/vijaiey88@gmail.com/titanic-3.csv")
6 df1.show(10)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. Joh...	female	38	1	0	PC 17599/71.2833	C85	C	
3	1	3	Heikkinen, Miss. ...	female	26	0	0	STON/O2. 3101282	7.925	null	S
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35	0	0	373450	8.05	null	S
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	null	Q
7	0	1	McCarthy, Mr. Tim...	male	54	0	0	17463/51.8625	54.0		S
8	0	3	Palsson, Master. ...	male	2	3	1	349909	21.075	null	S
9	1	1	Johnson, Mrs. Osc...	female	27	0	2	347742/11.1333	null		S
10	1	2	Nasser, Mrs. Nich...	female	14	1	0	237736/30.0708	null	C	

### 3. WHAT IS CLUSTER IN SPARK?

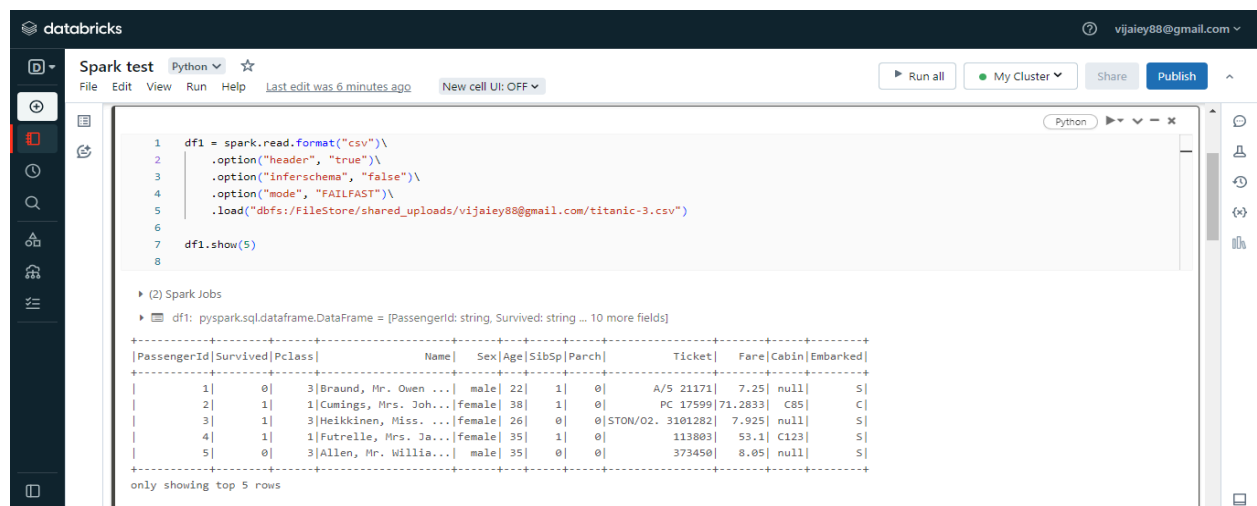
A Spark cluster is a group of machines that run Apache Spark, a big data processing engine. It consists of a driver, executors, and a cluster manager that work together to complete tasks.

### 4. WHAT IS TABLE IN SPARK ?

Table in spark is a collection of rows and columns that are stored as data files. In spark dataframe resembles this type of format which contains data in rows and columns(structured).

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

We can use header - 'true' in order to show the header in table



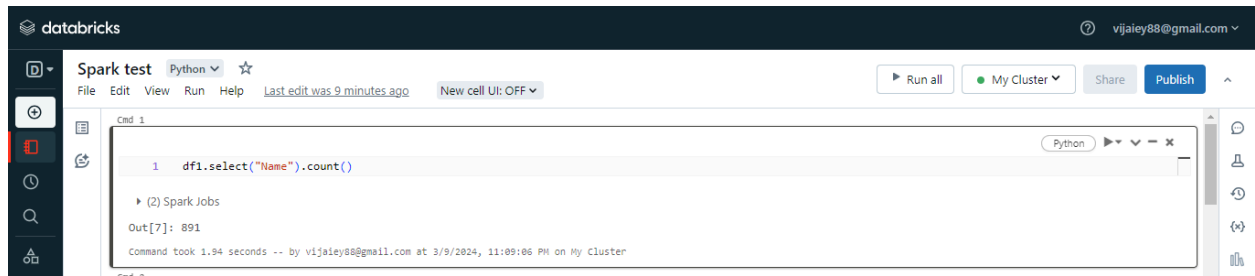
The screenshot shows a Databricks workspace with a Python script in a cell. The script reads a CSV file from a file store and displays the first 5 rows of the resulting DataFrame, including the header. The output shows a table with columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The first 5 rows of data are displayed.

```
1 df1 = spark.read.format("csv")\
2     .option("header", "true")\
3     .option("inferSchema", "false")\
4     .option("mode", "FAILFAST")\
5     .load("dbfs:/FileStore/shared_uploads/vijaiey88@gmail.com/titanic-3.csv")
6 df1.show(5)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. Joh...	female	38	1	0	PC 17599/71.2833	C85	C	
3	1	3	Heikkinen, Miss. ...	female	26	0	0	STON/O2. 3101282	7.925	null	S
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35	0	0	373450	8.05	null	S

## 6. WHAT IS COUNT ? PERFORM IN SPARK

Count is the function to calculate the number of records in a table



The screenshot shows the Databricks Spark test interface. The code cell contains the following Python code:

```
1 df1.select("Name").count()
```

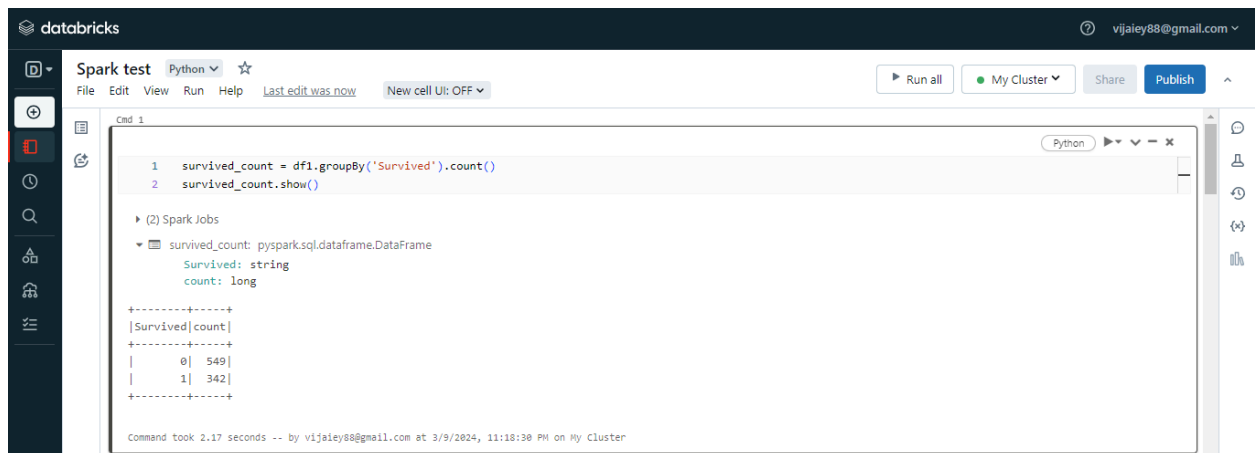
The output shows the result of the count operation:

```
Out[7]: 891
```

The command took 1.94 seconds to execute.

## 7. WHAT IS GROUP BY ? PERFORM IN SPARK

Groupby is used to separate records based on similarities



The screenshot shows the Databricks Spark test interface. The code cell contains the following Python code:

```
1 survived_count = df1.groupBy('Survived').count()
2 survived_count.show()
```

The output shows the result of the group by operation:

```
survived_count: pyspark.sql.dataframe.DataFrame
Survived: string
count: long

+-----+-----+
|Survived|count|
+-----+-----+
|0|549|
|1|342|
+-----+-----+
```

The command took 2.17 seconds to execute.