Create a Dataframe

data = [('Shahrukh','Khan','3/11/1965','M',3000),

('Salman','Khan','','11/12/1967','M',4000),

('Hritik','Roshan','12/3/1969','M',4000),

('Alia’,'Bhatt','3/5/1993','F',4000),

('Anushka', Sharma','5/10/1989','F',-1)

]

columns = ["firstname","lastname","dob","gender","salary"]

df = spark.createDataFrame(data=data, schema = columns)

To get the schema

df.printSchema()

To show the first 20 elements.

df.show()

To Read a CSV File

df = spark.read.option("header",True) .csv("/user/training/customer.csv")

df.printSchema()

df.show()

OR

df = spark.read.format("csv").load("/user/training/customer.csv")

// or

df = spark.read.format("org.apache.spark.sql.csv").load("/user/training/customer.csv")

df.printSchema()

To select columns

df.select("country","city","zipcode","state").show(5)

To read a CSV file into a table

spark.read.option("header",True) \

.csv("/user/training/customer.csv") \

.createOrReplaceTempView("customer")

To select using SQL

spark.sql("SELECT country, city, zipcode, state FROM CUSTOMER").show(5)

Filter Rows

df.select("country","city","zipcode","state") \

.where("state == 'AZ'") \

.show(5)

Filter Rows using SQL

spark.sql(""" SELECT country, city, zipcode, state FROM customer

WHERE state = 'AZ' """) \

.show(5)

Order By

df.select("country","city","zipcode","state") \

.where("state in ('PR','AZ','FL')") \

.orderBy("state") \

.show(10)

Order by in SQL

spark.sql(""" SELECT country, city, zipcode, state FROM customer

WHERE state in ('PR','AZ','FL') order by state """) \

.show(10)

Group By

df.groupBy("state").count().show()

To write a DF to a csv File

df.write.option("header",True).csv("/user/training/cust\_result")

Write DF with a header

df2.write.options(header='True', delimiter=',') \

.csv("/tmp/spark\_output/zipcodes")