Hands on demo in Jypyter notebook

First step

pip install pyspark

Second step

from pyspark.sql import SparkSession

spark=SparkSession.builder.appName("Examples").getOrCreate()

df=spark.read.csv('diabetes.csv',header=True,escape="\"")

df.show(5,0)

Third step

df.printSchema()

Fourthstep

print((df.count(),len(df.columns)))

Fifthstep

df.groupby('Outcome')

Sixthstep

# cleaning data

for col in df.columns:

print(col+":",df[df[col].isNull()].count())

Seventhstep

from pyspark.sql import \*

for i in df.columns[1:6]:

data=df.agg({i:'mean'}).first()[0]

print("Mean value for {} is {}".format(i,int(data)))