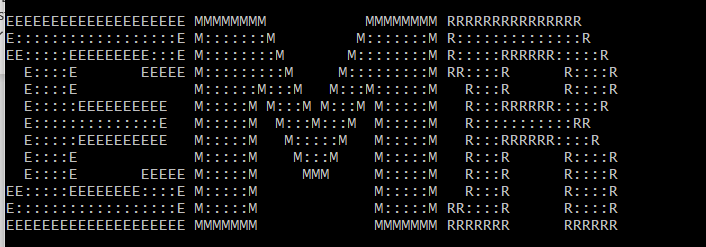
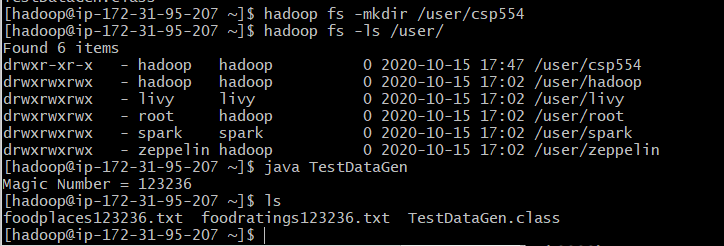
***Creating an EMR Cluster***

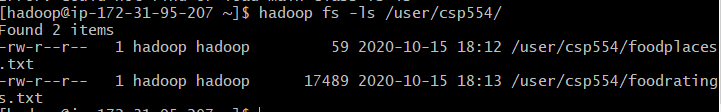


***Running TestDataGen.Class***

***Magic Number: 123236***



***Moving the files created using testDatagen.class to newly created Directory csp554 using copyFropmLocal command***



***Exercise 1:***

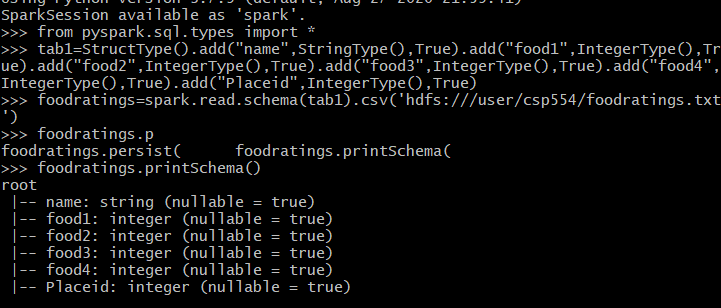
***Loading the foodratings.txt file to foodratings DataFrame***

***Command Used:***

***from pyspark.sql.types import \****

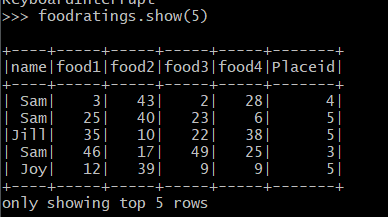
***tab1=StructType().add("name",StringType(),True).add("food1",IntegerType(),True).add("food2",IntegerType(),True).add("food3",IntegerType(),True).add("food4",IntegerType(),True).add("Placeid",IntegerType(),True)***

***foodratings=spark.read.schema(tab1).csv('hdfs:///user/csp554/foodratings.txt')***



***Showing top 5 Rows***

***Command Used: foodratings.show(5)***



***Exercise 2:***

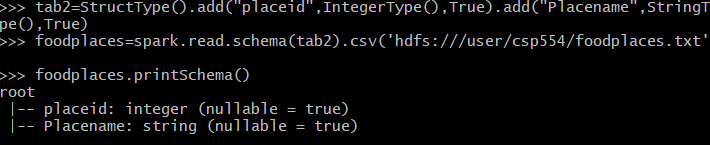
***Loading the foodplaces.txt file to foodplaces DataFrame***

***Command Used:***

***tab2=StructType().add("placeid",IntegerType(),True).add("Placename",StringType(),True)***

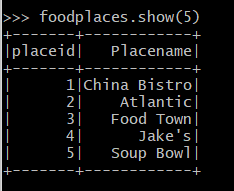
***foodplaces=spark.read.schema(tab2).csv('hdfs:///user/csp554/foodplaces.txt')***

***foodplaces.printSchema()***



***Showing top 5 Rows***

***Command Used: foodplaces.show(5)***



***Exercise 3:***

1. ***Creating a table using the below command***

***Command used:***

***foodratings.createOrReplaceTempView("foodratingsT")***

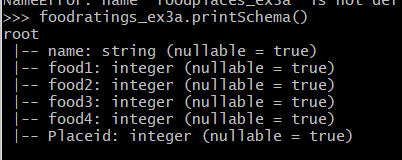
***foodplaces.createOrReplaceTempView("foodplacesT")***

1. ***Creating a new table from the fooodratingsT created at above step***

***Command Used:***

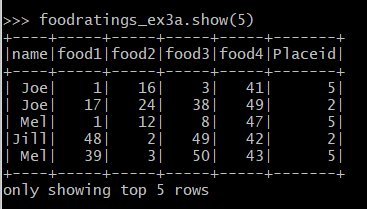
***foodratings\_ex3a=spark.sql("select \* from foodratingsT where food2<25 and food4>40")***





***Showing top 5 Rows***

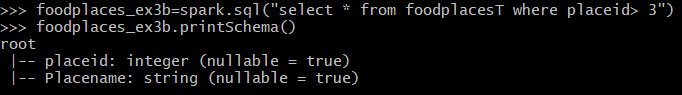
***Command Used: foodpratings\_ex3a.show(5)***



1. ***Creating a new table from the fooodplacesT created at above step***

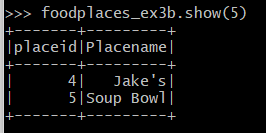
***Command Used:***

***foodplaces\_ex3b=spark.sql("select \* from foodplacesT where placeid> 3") foodplaces\_ex3b.printSchema()***



***Showing top 5 Rows***

***Command Used: foodplaces.show(5)***



***Exercise 4:***

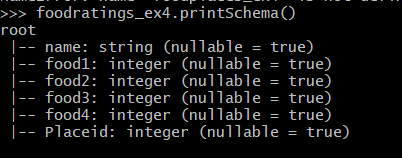
***Creating a new DataFrame using the below command***

***Command Used:***

***foodratings\_ex4=foodratings.filter((foodratings.name=='Mel') & (foodratings.food3<25))***

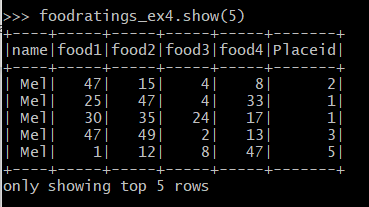
***foodratings\_ex4.printSchema()***





***Showing top 5 Rows***

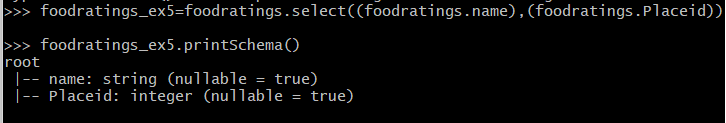
***Command Used: foodratings\_ex4.show(5)***



***Exercise 5: Creating a new DataFrame using columns name and PlaceId***

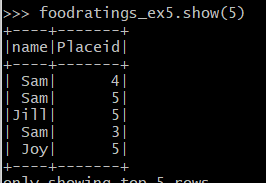
***Command Used:***

***foodratings\_ex5=foodratings.select((foodratings.name),(foodratings.Placeid))***



***Showing top 5 Rows***

***Command Used: foodratings\_ex5.show(5)***

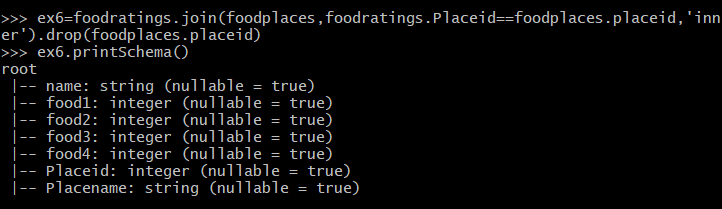


***Exercise 6: Creating a new Dataframe using below command***

***Command Used:***

***ex6=foodratings.join(foodplaces,foodratings.Placeid==foodplaces.placeid,'inner').drop(foodplaces.placeid)***

***ex6.printSchema()***



***Showing top 5 Rows***

***Command Used: ex6.show(5)***

