1)

from pyspark.sql.types import \*

foodratingstruct = StructType(

[

StructField("name", StringType(), True),

StructField("food1",IntegerType(), True),

StructField("food2",IntegerType(), True),

StructField("food3",IntegerType(), True),

StructField("food4",IntegerType(), True),

StructField("placeid",IntegerType(), True)

]

)

foodratings=spark.read.schema(foodratingstruct).csv('/user/maria\_dev/foodratings178695.txt')

foodratings.printSchema()

foodratings.head(5)

2)

from pyspark.sql.types import \*

foodplacestruct = StructType(

[

StructField("placeid", IntegerType(), True),

StructField("placename", StringType(), True)

]

)

foodplaces = spark.read.schema(foodplacestruct).csv('/user/maria\_dev/foodplaces178695.txt')

foodplaces.printSchema()

foodplaces.head(5)

3)

from pyspark.sql.types import \*

foodratings.createOrReplaceTempView("foodratingsT")

foodplaces.createOrReplaceTempView("foodplacesT")

foodratings\_ex3 = spark.sql("SELECT \* FROM foodratingsT WHERE food2 < 25 AND food4 > 40")

foodratings\_ex3.printSchema()

foodratings\_ex3.head(5)

Step c)

foodplaces\_ex3 = spark.sql("SELECT \* FROM foodplacesT WHERE placeid > 3")

foodplaces\_ex3.printSchema()

foodplaces\_ex3.head(5)

4)

foodratings\_ex4 = foodratings.filter(foodratings.name == "Mel").filter(foodratings.food3 <

25)

foodratings\_ex4.printSchema()

foodratings\_ex4.head(5)

5)

foodratings\_ex5 = foodratings.select('name','placeid')

foodratings\_ex5.printSchema()

foodratings\_ex5.head(5)

6)

ex6 = foodratings.join(foodplaces, foodratings.placeid == foodplaces.placeid,"inner").drop(foodratings.placeid)

ex6.printSchema()

ex6.head(5)