Exercise 1 :

**Magic Number : 62216**

Code :

from pyspark.sql import \*

food\_schema = 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)

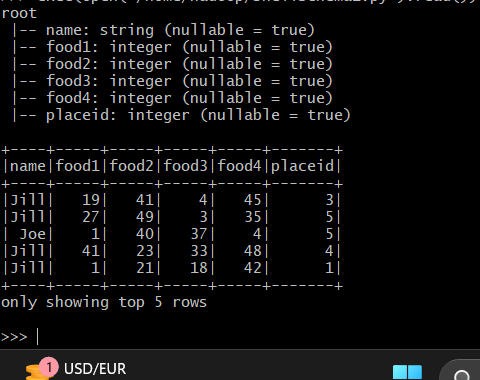
])

foodratings = spark.read.csv("/user/hadoop/foodratings62216.txt", schema=food\_schema, header=True)

foodratings.printSchema()

foodratings.show(5)

output :



**Exercise 2:**

**Code :**

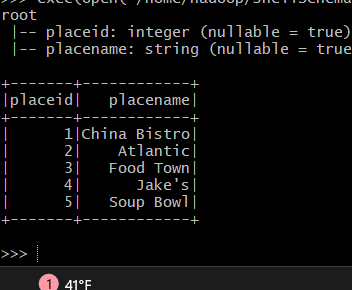
from pyspark.sql import \*

struct1 = StructType().add("placeid", IntegerType(), True).add("placename", StringType(), True)

foodplaces = spark.read.schema(struct1).csv('/user/hadoop/foodplaces62216.txt')

foodplaces.printSchema()

foodplaces.show(5)



Exercise 3:

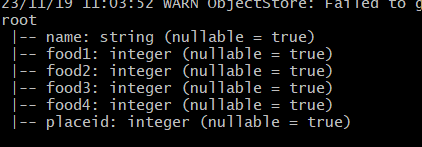
3.1:

Code :

foodratings.createOrReplaceTempView("foodratingsT")

foodplaces.createOrReplaceTempView("foodplacesT")

output :



3.2

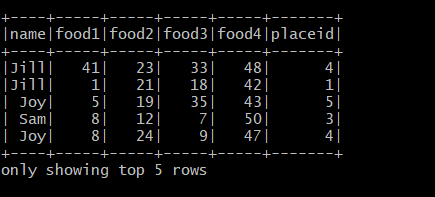
Code :

foodratings\_ex3a = spark.sql("SELECT \* FROM foodratingsT WHERE food2 < 25 AND food4 > 40")

foodratings\_ex3a.printSchema()

foodratings\_ex3a.show(5)

output :



3.3

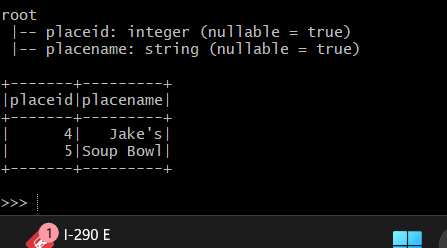
Code :

foodplaces\_ex3b = spark.sql("SELECT \* FROM foodplacesT WHERE placeid > 3")

foodplaces\_ex3b.printSchema()

foodplaces\_ex3b.show(5)

output :



Exercise 4:

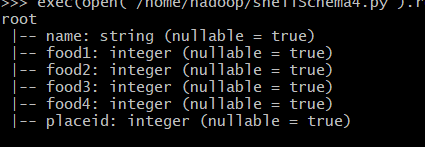
Code :

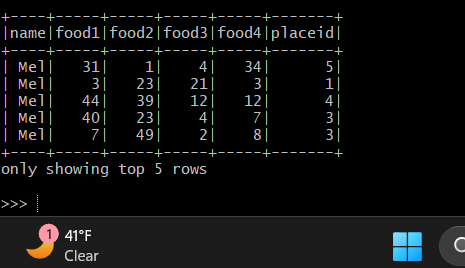
foodratings\_ex4 = foodratings.filter((foodratings['name'] == "Mel") & (foodratings['food3'] < 25))

foodratings\_ex4.printSchema()

foodratings\_ex4.show(5)

output :





Exercise 5:

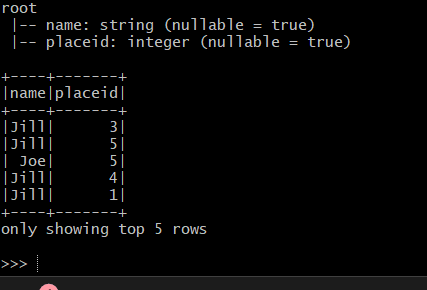
Code :

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

foodratings\_ex5.printSchema()

foodratings\_ex5.show(5)

output:



Exercise 6:

Code :

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

ex6.printSchema()

ex6.show(5)

output :

