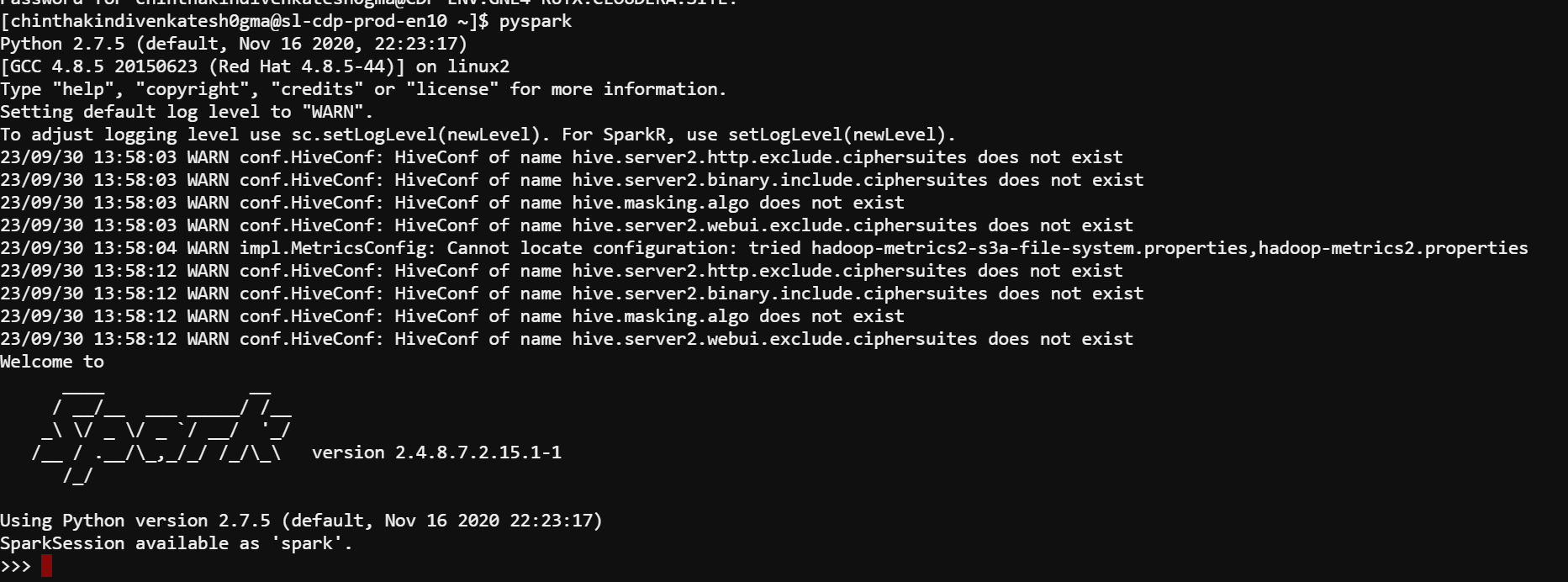
Retail DB Analysis:

Log in to PySpark shell

**Task 2.1:**



**Task 2.2:**

Explore the customer records saved in the "customers-tab-delimited" directory on HDFS

**REQUIREMENT:**

2.2.1 Show the client information for those who live in California

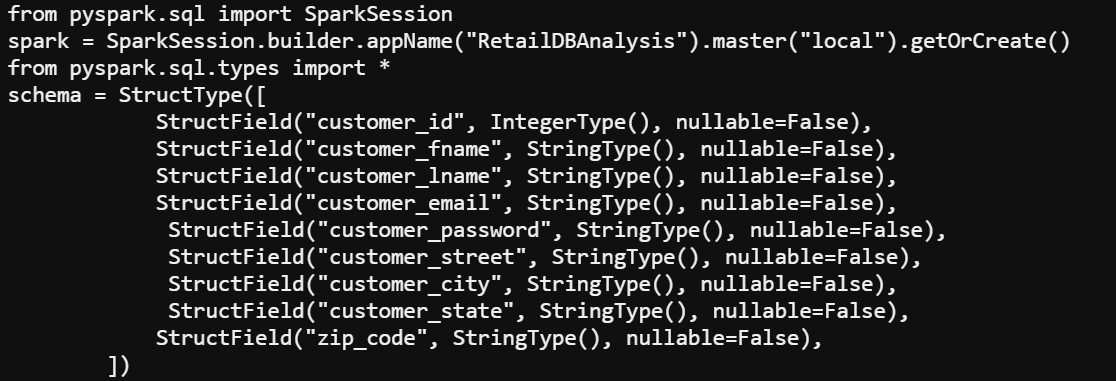
2.2.2 The final output must be in text format

2.2.3 Save the results in the result/scenario1/solution folder

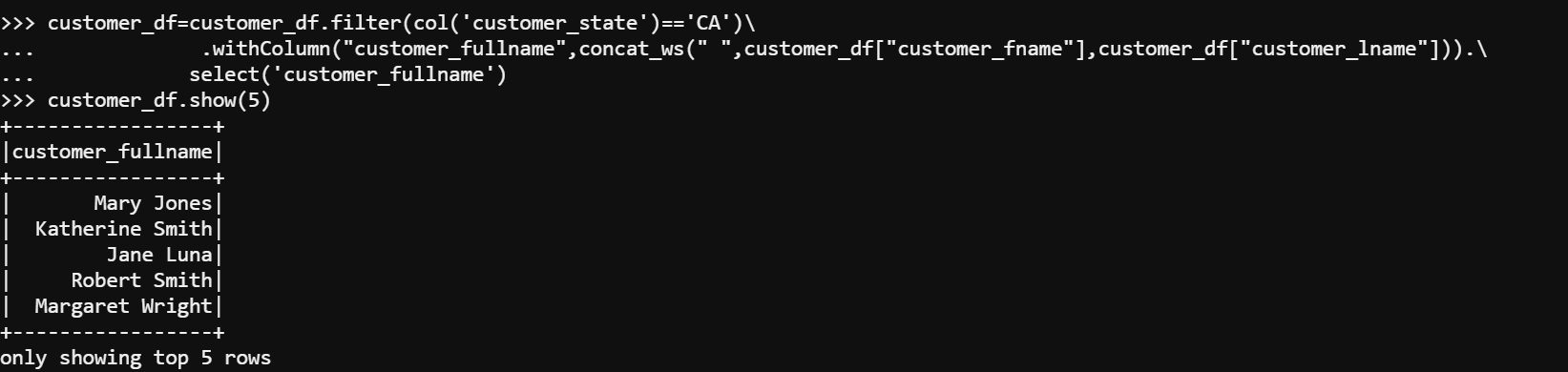
2.2.4 Only records with the state value "CA" should be included in the result

2.2.5 Only the customer's entire name should be included in the output

Example: “Robert Hudson”

****







**Task 2.3:**

**Task 2.3:**

Explore the order records saved in the “orders parquet” directory on HDFS

**REQUIREMENT:**

2.3.1:  Show all orders with the order status value "COMPLETE"

2.3.2: The output should be in JSON format

2.3.3: Save the data in the "result/scenario2/solution" directory on HDFS

2.3.4: The "order date" column should be in the "YYYY-MM-DD" format

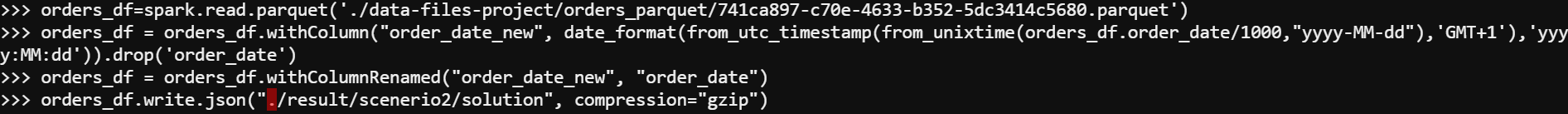
2.3.5: Use GZIP compression to compress the output

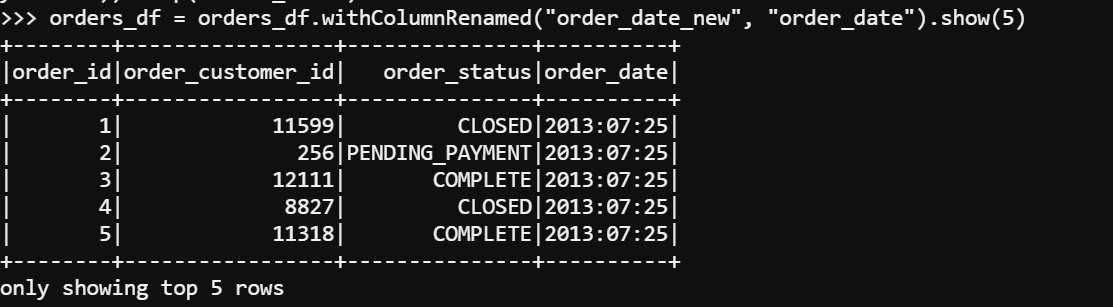
2.3.6: Only the column names listed below should be included in the output:

2.3.6.1 Order number

2.3.6.2 Order date

2.3.6.3 Current situation



****

**Task 2.4**

Explore the customer records saved in the "customers-tab-delimited" directory on HDFS

**REQUIREMENT:**

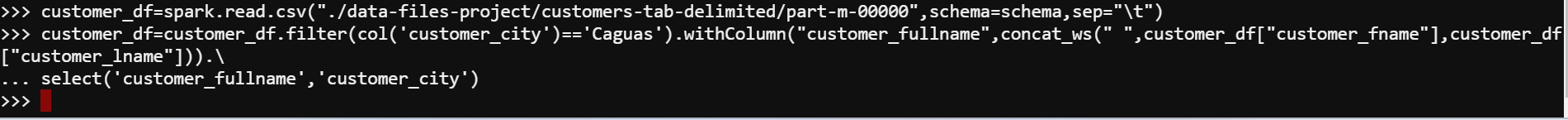
2.4.1   Produce a list of all consumers who live in the city of "Caguas"

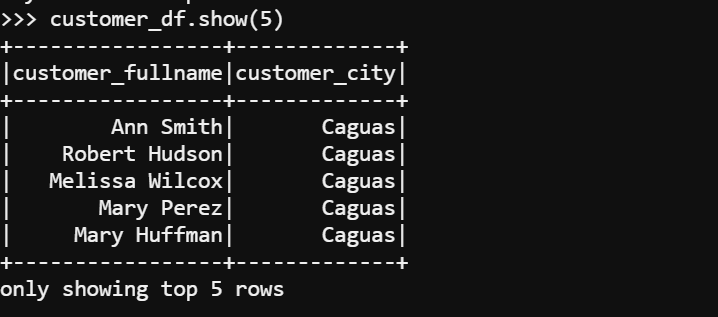
2.4.2   Save the data in the result/scenario3/solution directory on HDFS

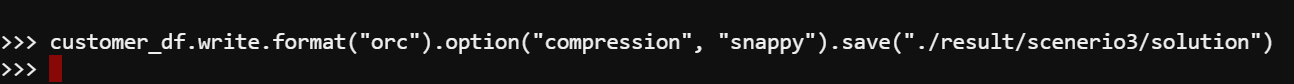
2.4.3   The result should only contain records with the value "Caguas" for the customer city

2.4.4    Use snappy compression to compress the output

2.4.5    Save the file in the orc format







**Task 2.5**

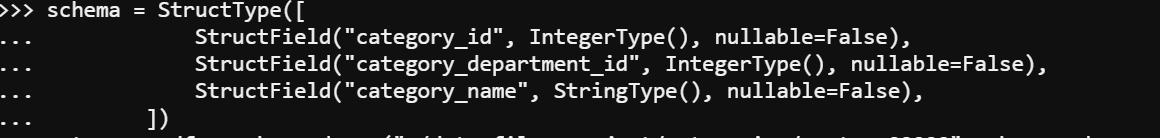
Explore all the category records stored in the “categories” directory on HDFS

**REQUIREMENT:**

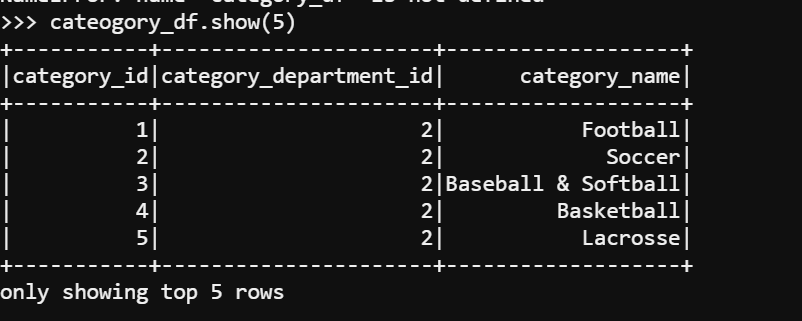
2.5.1    Save the result files in CSV format

2.5.2    Save the data in the result/scenario4/solution directory on HDFS

2.5.3    Use lz4 compression to compress the output









**Task 2.6**

Explore all product records that are saved in the “products\_avro” database

**REQUIREMENT:**

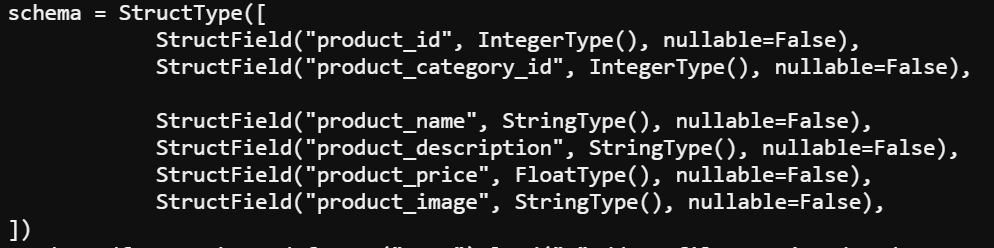
2.6.1 Only products with a price of more than 1000.0 should be included in the output

2.6.2   Save the output files in parquet format

2.6.3   Remove data from the table if the product price is lesser than 1000.0

2.6.4   Save the data in the result/scenario5/solution directory on HDFS

2.6.5   Use snappy compression to compress the output









**Task 2.7**

Explore the “products\_avro” stored in product records

**REQUIREMENT:**

2.7.1 Only products with a price of more than 1000.0 should be in the output

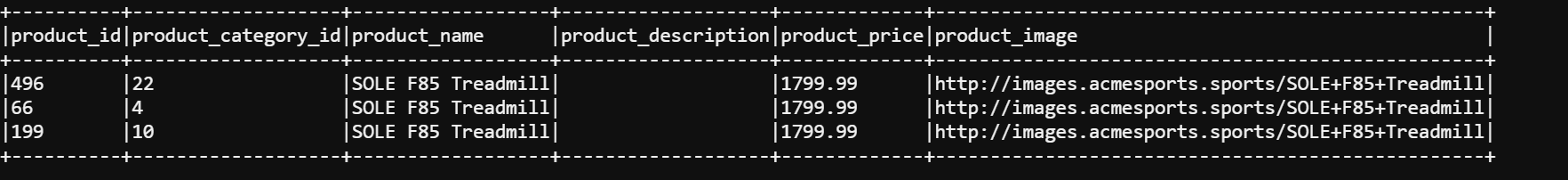
2.7.2 The pattern "Treadmill" appears in the product name

2.7.3 Save the output files in parquet format

2.7.4Save the data in the result/scenario6/solution directory on HDFS

2.7.5 Use GZIP compression to compress the output







**Task 2.8**

Explore the order records that are saved in the “orders parquet” table on HDFS

**REQUIREMENT:**

2.8.1 Output all PENDING orders in July 2013

2.8.2 Output files should be in JSON format

2.8.3 Save the data in the result/scenario7/solution directory on HDFS.

2.8.4 Only entries with the order status value of "PENDING" should be included in the result

2.8.5 Order date should be in the YYY-MM-DD format

2.8.6 Use snappy compression to compress the output, which should just contain the order date and order status

