CCA175 SAMPLE EXAM QUESTIONS

**Question 1 (1.1):**

You have been given MySQL DB with following details.

user=retail\_user  
password=itversity  
database=retail\_db  
table=retail\_db/orders  
table=retail\_db/order\_items  
jdbc URL = jdbc:mysql://ms.itversity.com /retail\_db  
Please accomplish following activities.  
1. Copy "retail\_db/order\_items" table to hdfs in respective directory p90\_order\_items .  
2. Do the summation of entire revenue in this table using pyspark.  
3. Find the maximum and minimum revenue as well.  
4. Calculate average revenue  
Columns of ordeMtems table : (order\_item\_id , order\_item\_order\_id ,  
order\_item\_product\_id, order\_item\_quantity,order\_item\_subtotal,order\_  
item\_subtotal,order\_item\_product\_price)

*Solution:*

*1)*  Copy table from MySQL into HDFS => use Sqoop Import.p90\_order\_items

Write the following Sqoop query:

sqoop import \

--connect jdbc:mysql://ms.itversity.com/retail\_db \

--username retail\_user \

--password itversity \

--table order\_items \

--target-dir /user/por160893/sqoop\_import/p90\_order\_items \

--num-mappers 1

Check the imported table to see if it is import is correct:

hadoop fs -cat /user/por160893/sqoop\_import/p90\_order\_items/part-m-00000 | head

*2)* Check the order of the columns in the imported table, “p90\_order\_items”, on where it is on MySql, origiinally in table “order\_items”:

mysql -u retail\_user -h ms.itversity.com -p

show databases;

use retail\_db;

describe order\_items;

The order of the columns are: order\_item\_id, order\_item\_order\_id, order\_item\_product\_id, order\_item\_quantity, order\_item\_subtotal, order\_item\_product\_price.

Next, launch PySpark on the shell by typing: pyspark

Want to add up all the subtotals over all orders in HDFS table “p90\_order\_items”.

First, convert the HDFS table into a RDD to be used in PySpark:

ordersRdd = sc.textFile("/user/por160893/sqoop\_import/p90\_order\_items/part-m-00000“)

Next, since we have used sc.textFile, all the columns of the imported table is of type String in RDD, “ordersRDD”.

We are only interested in the “order\_item\_subtotal” of this RDD and we need to convert it to its original Float type to sum up the values.

extracted = ordersRdd.map(lambda k:k.float(split(",")[4]))

Check the contents of the RDD by lookin at the first 10 rows of the RDD:

for i in extracted.take(10):print i

Next, sum up all the values in the column “extracted”

sumOrderSubtotals = extracted.reduce(lambda a, b: a+b)

3) Need to find the minimum and maximum subtotal amount in the RDD “sumOrderSubtotals”.

min = extracted.reduce(lambda a,b: a if a<b else b)

and

max = extracted.reduce(lambda a,b: a if a>b else b)

4) Finally, find the average order amount.

Already have the total amount of the orders. Next, need to find the number of orders.

numberOrders = extracted.count()

Now, calculate the average with the total order amount and the number of orders:

averageOrders = sumOrderSubtotals/numberOrders

**Question 2 (2.2):**

You have been given MySQL DB with following details.

user=retail\_user  
password=itversity  
database=retail\_db  
jdbc URL = jdbc:mysql://ms.itversity.com /retail\_db  
Now accomplish following activities.  
1. Import departments table from mysql to hdfs as textfile in departments\_text directory.  
2. Import departments table from mysql to hdfs as sequncefile in departments\_sequence directory.  
3. Import departments table from mysql to hdfs as avro file in departments\_avro directory.  
4. Import departments table from mysql to hdfs as parquet file in departments\_parquet directory.

*Solution:*

*1)*  Want to import the departments table into a new directory for each file type => must use --target-dir for each case.

Imported as a text file into “departments\_text” directory:

**sqoop import**

**--connect jdbc:mysql://ms.itversity.com/retail\_db**

**--username retail\_user**

**--password itversity**

**--table departments**

**--target-dir /user/por160893/sqoop\_import/departments\_text**

**--num-mappers 1**

**--as-textfile**

2) Imported as a sequence file into “department\_sequence” directory:

**sqoop import**

**--connect jdbc:mysql://ms.itversity.com/retail\_db**

**--username retail\_user**

**--password itversity**

**--table departments**

**--target-dir /user/por160893/sqoop\_import/departments\_sequence**

**--num-mappers 1**

**--as-sequencefile**

3) Imported as a sequence file into “department\_avro” directory:

**sqoop import**

**--connect jdbc:mysql://ms.itversity.com/retail\_db**

**--username retail\_user**

**--password itversity**

**--table departments**

**--target-dir /user/por160893/sqoop\_import/departments\_avro**

**--num-mappers 1**

**--as-avrodatafile**

4) Imported as a parquet file into “department\_parquet” directory:

**sqoop import**

**--connect jdbc:mysql://ms.itversity.com/retail\_db**

**--username retail\_user**

**--password itversity**

**--table departments**

**--target-dir /user/por160893/sqoop\_import/departments\_parquet**

**--num-mappers 1**

**--as-parquetfile**

**Question 3 (2.11):**

You have been given following mysql database details as well as other info.

user=retail\_user  
password=itversity  
database=retail\_db  
jdbc URL = jdbc:mysql://ms.itversity.com /retail\_db  
Please accomplish following.  
1. Import departments table in a directory.  
2. Again import departments table same directory (However, directory already exist hence it should not overrride and append the results)  
3. Also make sure your results fields are terminated by '|' and lines terminated by '\n\

*Solution:*

*1)*  Want to import the departments table into a new directory, say directory called departmentsNew1

**sqoop import**

**--connect jdbc:mysql://ms.itversity.com/retail\_db**

**--username retail\_user**

**--password itversity**

**--table departments**

**--target-dir /user/por160893/sqoop\_import/departmentsNew1**

**--num-mappers 1**

**--fields-terminated-by "|"**

**--lines-terminated-by "\n"**

2) Next, import the same table into the same directory => must use the “—append” command to add a table to an existing directory.

**sqoop import   
--connect jdbc:mysql://ms.itversity.com/retail\_db   
--username retail\_user   
--password itversity   
--target-dir /user/por160893/sqoop\_import/departmentsNew1   
--append   
--table departments   
--num-mappers 1   
--fields-terminated-by "|"   
--lines-terminated-by "\n"**

3) Finally, check your imported tables, which should generate the same results:

**hadoop fs -cat /user/sqoop\_import/departmentsNew1/part-m-00000**

and

**hadoop fs - cat /user/sqoop\_import/departmentsNew1/part-m-00001**