CCA175$^{Q\&As}$

CCA Spark and Hadoop Developer Exam

Pass Cloudera CCA175 Exam with 100% Guarantee

QUESTION 1

Problem Scenario 12 : You have been given following mysql database details as well as other info. user=retail_dba password=cloudera database=retail_db jdbc URL = jdbc:mysql://quickstart:3306/retail_db Please accomplish following.

1.

Create a table in retailedb with following definition.

CREATE table departments_new (department_id int(11), department_name varchar(45),

created_date T1MESTAMP DEFAULT NOW());

2.

Now isert records from departments table to departments_new

3.

Now import data from departments_new table to hdfs.

4.

Insert following 5 records in departmentsnew table. Insert into departments_new

values(110, "Civil" , null); Insert into departments_new values(111, "Mechanical" , null);

Insert into departments_new values(112, "Automobile" , null); Insert into departments_new

values(113, "Pharma" , null);

Insert into departments_new values(114, "Social Engineering" , null);

5.

Now do the incremental import based on created_date column.

Correct Answer: See the explanation for Step by Step Solution and configuration.

Solution :

Step 1 : Login to musql db

mysql --user=retail_dba -password=cloudera

show databases;

use retail db; show tables;

Step 2 : Create a table as given in problem statement.

CREATE table departments_new (department_id int(11), department_name varchar(45),

createddate T1MESTAMP DEFAULT NOW());

show tables;

Step 3 : isert records from departments table to departments_new insert into

departments_new select a.", null from departments a;

Step 4 : Import data from departments new table to hdfs.

sqoop import \

-connect jdbc:mysql://quickstart:330G/retail_db \

~username=retail_dba \

-password=cloudera \

-table departments_new\

--target-dir /user/cloudera/departments_new \

--split-by departments

Stpe 5 : Check the imported data.

hdfs dfs -cat /user/cloudera/departmentsnew/part"

Step 6 : Insert following 5 records in departmentsnew table.

Insert into departments_new values(110, "Civil" , null);

Insert into departments_new values(111, "Mechanical" , null);

Insert into departments_new values(112, "Automobile" , null);

Insert into departments_new values(113, "Pharma" , null);

Insert into departments_new values(114, "Social Engineering" , null);

commit;

Stpe 7 : Import incremetal data based on created_date column.

sqoop import \

-connect jdbc:mysql://quickstart:330G/retail_db \

-username=retail_dba \

-password=cloudera \

--table departments_new\

-target-dir /user/cloudera/departments_new \

-append \

-check-column created_date \

-incremental lastmodified \

-split-by departments \

-last-value "2016-01-30 12:07:37.0"

Step 8 : Check the imported value.

hdfs dfs -cat /user/cloudera/departmentsnew/part"

QUESTION 2

Problem Scenario 76 : You have been given MySQL DB with following details. user=retail_dba password=cloudera database=retail_db table=retail_db.orders table=retail_db.order_items jdbc URL = jdbc:mysql://quickstart:3306/retail_db Columns of order table : (orderid , order_date , ordercustomerid, order_status} ..... Please accomplish following activities.

1.

Copy "retail_db.orders" table to hdfs in a directory p91_orders.

2.

Once data is copied to hdfs, using pyspark calculate the number of order for each status.

3.

Use all the following methods to calculate the number of order for each status. (You need to know all these functions and its behavior for real exam)

-countByKey() -groupByKey()

-reduceByKey() -aggregateByKey()

-combineByKey()

Correct Answer: See the explanation for Step by Step Solution and configuration.

Solution :

Step 1 : Import Single table

sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --username=retail dba password=cloudera -table=orders --target-dir=p91_orders

Note : Please check you dont have space between before or after \\'=\\' sign. Sqoop uses the

MapReduce framework to copy data from RDBMS to hdfs

Step 2 : Read the data from one of the partition, created using above command, hadoop fs

-cat p91_orders/part-m-00000

Step 3: countByKey #Number of orders by status allOrders = sc.textFile("p91_orders")

#Generate key and value pairs (key is order status and vale as an empty string keyValue =

alIOrders.map(lambda line: (line.split(",")[3], ""))

#Using countByKey, aggregate data based on status as a key

output=keyValue.countByKey()Jtems()

for line in output: print(line)

Step 4 : groupByKey

#Generate key and value pairs (key is order status and vale as an one

keyValue = allOrders.map(lambda line: (line.split)",")[3], 1))

#Using countByKey, aggregate data based on status as a key output=

keyValue.groupByKey().map(lambda kv: (kv[0], sum(kv[1]}}}

tor line in output.collect(): print(line}

Step 5 : reduceByKey

#Generate key and value pairs (key is order status and vale as an one

keyValue = allOrders.map(lambda line: (line.split(","}[3], 1))

#Using countByKey, aggregate data based on status as a key output=

keyValue.reduceByKey(lambda a, b: a + b)

tor line in output.collect(): print(line}

Step 6: aggregateByKey

#Generate key and value pairs (key is order status and vale as an one keyValue =

allOrders.map(lambda line: (line.split(",")[3], line}}

output=keyValue.aggregateByKey(0, lambda a, b: a+1, lambda a, b: a+b}

for line in output.collect(): print(line}

Step 7 : combineByKey

#Generate key and value pairs (key is order status and vale as an one

keyValue = allOrders.map(lambda line: (line.split(",")[3], line))

output=keyValue.combineByKey(lambda value: 1, lambda ace, value: acc+1, lambda ace,

value: acc+value)

tor line in output.collect(): print(line)

#Watch Spark Professional Training provided by www.ABCTECH.com to understand more on each above functions. (These are very important functions for real exam)

QUESTION 3

Problem Scenario 51 : You have been given below code snippet.

val a = sc.parallelize(List(1, 2,1, 3), 1)

val b = a.map((_, "b"))

val c = a.map((_, "c"))

Operation_xyz

Write a correct code snippet for Operationxyz which will produce below output.

Output:

Array[(Int, (Iterable[String], Iterable[String]))] = Array(

(2,(ArrayBuffer(b),ArrayBuffer(c))),

(3,(ArrayBuffer(b),ArrayBuffer(c))),

(1,(ArrayBuffer(b, b),ArrayBuffer(c, c))))

Correct Answer: See the explanation for Step by Step Solution and configuration.

Solution :

b.cogroup(c).collect

cogroup [Pair], groupWith [Pair]

A very powerful set of functions that allow grouping up to 3 key-value RDDs together using

their keys.

Another example

val x = sc.parallelize(List((1, "apple"), (2, "banana"), (3, "orange"), (4, "kiwi")), 2)

val y = sc.parallelize(List((5, "computer"), (1, "laptop"), (1, "desktop"), (4, "iPad")), 2)

x.cogroup(y).collect

Array[(Int, (Iterable[String], Iterable[String]))] = Array(

(4,(ArrayBuffer(kiwi),ArrayBuffer(iPad))),

(2,(ArrayBuffer(banana),ArrayBuffer())),

(3,(ArrayBuffer(orange),ArrayBuffer())),

(1 ,(ArrayBuffer(apple),ArrayBuffer(laptop, desktop))),

(5,{ArrayBuffer(),ArrayBuffer(computer))))

QUESTION 4

Problem Scenario 15 : You have been given following mysql database details as well as other info. user=retail_dba password=cloudera database=retail_db jdbc URL = jdbc:mysql://quickstart:3306/retail_db Please accomplish following activities.

1.

In mysql departments table please insert following record. Insert into departments values(9999, \\'"Data Science"1);

2.

Now there is a downstream system which will process dumps of this file. However, system is designed the way that it can process only files if fields are enlcosed in(\\') single quote and separate of the field should be (-} and line needs to be terminated by : (colon).

3.

If data itself contains the " (double quote } than it should be escaped by \.

4.

Please import the departments table in a directory called departments_enclosedby and file should be able to process by downstream system.

Correct Answer: See the explanation for Step by Step Solution and configuration.

Solution :

Step 1 : Connect to mysql database.

mysql --user=retail_dba -password=cloudera

show databases; use retail_db; show tables;

Insert record

Insert into departments values(9999, \\'"Data Science"\\');

select" from departments;

Step 2 : Import data as per requirement.

sqoop import \

-connect jdbc:mysql;//quickstart:3306/retail_db \

~username=retail_dba \

--password=cloudera \

-table departments \

-target-dir /user/cloudera/departments_enclosedby \

-enclosed-by V -escaped-by \\ -fields-terminated-by--\\' -lines-terminated-by :

Step 3 : Check the result.

hdfs dfs -cat/user/cloudera/departments_enclosedby/part"

QUESTION 5

Problem Scenario 52 : You have been given below code snippet.

val b = sc.parallelize(List(1,2,3,4,5,6,7,8,2,4,2,1,1,1,1,1))

Operation_xyz

Write a correct code snippet for Operation_xyz which will produce below output.

scalaxollection.Map[lnt,Long] = Map(5 -> 1, 8 -> 1, 3 -> 1, 6 -> 1, 1 -> S, 2 -> 3, 4 -> 2, 7 ->

1)

Correct Answer: See the explanation for Step by Step Solution and configuration.

Solution : b.countByValue countByValue Returns a map that contains all unique values of the RDD and their respective occurrence counts. (Warning: This operation will finally aggregate the information in a single reducer.) Listing Variants def countByValue(): Map[T, Long]