Hadoop and Spark Developer - CCA 175

Problem Scenario 1

PLEASE READ THE INTRODUCTION TO THIS SERIES. CLICK ON HOME LINK AND READ THE INTRO BEFORE ATTEMPTING TO SOLVE THE PROBLEMS

Video walk through of the solution to this problem can be found here ***[Click here]***

Click here for the video version of this series. This takes you to the youtube playlist of videos.

Problem 1:

1. Using sqoop, import orders table into hdfs to folders **/user/cloudera/problem1/orders**. File should be loaded as Avro File and use snappy compression
2. Using sqoop, import order_items table into hdfs to folders **/user/cloudera/problem1/order-items**. Files should be loaded as avro file and use snappy compression
3. Using Spark Scala load data at **/user/cloudera/problem1/orders** and **/user/cloudera/problem1/orders-items** items as *dataframes*.
4. **Expected Intermediate Result:** Order_Date , Order_status, total_orders, total_amount. In plain english, please find total orders and total amount per status per day. The result should be sorted by order date in descending, order status in ascending and total amount in descending and total orders in ascending. Aggregation should be done using below methods. However, sorting can be done using a dataframe or RDD. Perform aggregation in each of the following ways
   - a). Just by using Data Frames API - here order_date should be YYYY-MM-DD format
   - b). Using Spark SQL - here order_date should be YYYY-MM-DD format
   - c). By using combineByKey function on RDDS -- No need of formatting order_date or total_amount
5. Store the result as parquet file into hdfs using gzip compression under folder
   - /user/cloudera/problem1/result4a-gzip
   - /user/cloudera/problem1/result4b-gzip
   - /user/cloudera/problem1/result4c-gzip
6. Store the result as parquet file into hdfs using snappy compression under folder
   - /user/cloudera/problem1/result4a-snappy
   - /user/cloudera/problem1/result4b-snappy
   - /user/cloudera/problem1/result4c-snappy
7. Store the result as CSV file into hdfs using No compression under folder
   - /user/cloudera/problem1/result4a-csv
   - /user/cloudera/problem1/result4b-csv
   - /user/cloudera/problem1/result4c-csv
8. create a mysql table named result and load data from **/user/cloudera/problem1/result4a-csv** to mysql table named result

Solution:

Try your best to solve the above scenario without going through the solution below. If you could then use the solution to compare your result. If you could not then I strongly recommend that you go through the concepts again (this time in more depth). Each step below provides a solution to the points mentioned in the Problem Scenario.

Step 1:

```
sqoop import \
--connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
--username retail_dba \
--password cloudera \
--table orders \
--compress \
--compression-codec org.apache.hadoop.io.compress.SnappyCodec \
--target-dir /user/cloudera/problem1/orders \
--as-avrodatafile;
```

Step 2:

```
sqoop import \
--connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
--username retail_dba \
--password cloudera \
--table order_items \
--compress \
--compression-codec org.apache.hadoop.io.compress.SnappyCodec \
--target-dir /user/cloudera/problem1/order-items \
--as-avrodatafile;
```

Step 3:

```
import com.databricks.spark.avro._;
  var ordersDF = sqlContext.read.avro("/user/cloudera/problem1/orders");
  var orderItemDF = sqlContext.read.avro("/user/cloudera/problem1/order-items");
```

Step 4:

```
var joinedOrderDataDF = ordersDF
.join(orderItemDF,ordersDF("order_id")===orderItemDF("order_item_order_id"))
```

Step 4a:

```
import org.apache.spark.sql.functions._;

var dataFrameResult =
dataFrameResult.show();


joinedOrderDataDF.
```

```
groupBy(to_date(from_unixtime(col("order_date")/1000)).alias("order_formatted_date"),col("order_status")).
agg(round(sum("order_item_subtotal"),2).alias("total_amount"),countDistinct("order_id").alias("total_orders")).
orderBy(col("order_formatted_date").desc,col("order_status"),col("total_amount").desc,col("total_orders"));
```

Step 4b:

```
joinedOrderDataDF.registerTempTable("order_joined");

var sqlResult = sqlContext.sql("select to_date(from_unixtime(cast(order_date/1000 as bigint))) as order_formatted_date, order_status, cast(sum(order_item_subtotal) as DECIMAL (10,2)) as total_amount,
count(distinct(order_id)) as total_orders from order_joined group by to_date(from_unixtime(cast(order_date/1000 as bigint))), order_status order by order_formatted_date desc,order_status,total_amount desc,
total_orders");

sqlResult.show();
```

Step 4c:

```
var comByKeyResult =
joinedOrderDataDF.
map(x=> ((x(1).toString,x(3).toString),(x(8).toString.toFloat,x(0).toString))).
combineByKey((x:(Float, String))=>(x._1,Set(x._2)),
(x:(Float,Set[String]),y:(Float,String))=>(x._1 + y._1,x._2+y._2),
(x:(Float,Set[String]),y:(Float,Set[String]))=>(x._1+y._1,x._2++y._2)).
map(x=> (x._1._1,x._1._2,x._2._1,x._2._2.size)).
toDF().
orderBy(col("_1").desc,col("_2"),col("_3").desc,col("_4"));

comByKeyResult.show();
```

Step 5:

- sqlContext.setConf("spark.sql.parquet.compression.codec","gzip");
- dataFrameResult.write.parquet("/user/cloudera/problem1/result4a-gzip");
- sqlResult.write.parquet("/user/cloudera/problem1/result4b-gzip");
- comByKeyResult.write.parquet("/user/cloudera/problem1/result4c-gzip");

Step 6:

- sqlContext.setConf("spark.sql.parquet.compression.codec","snappy");
- dataFrameResult.write.parquet("/user/cloudera/problem1/result4a-snappy");
- sqlResult.write.parquet("/user/cloudera/problem1/result4b-snappy");
- comByKeyResult.write.parquet("/user/cloudera/problem1/result4c-snappy");

Step 7:

- dataFrameResult.map(x=> x(0) + "," + x(1) + "," + x(2) + "," + x(3)).saveAsTextFile("/user/cloudera/problem1/result4a-csv")
- sqlResult.map(x=> x(0) + "," + x(1) + "," + x(2) + "," + x(3)).saveAsTextFile("/user/cloudera/problem1/result4b-csv")
- comByKeyResult.map(x=> x(0) + "," + x(1) + "," + x(2) + "," + x(3)).saveAsTextFile("/user/cloudera/problem1/result4c-csv")

Step 8:

a) login to MYSQL using below : mysql -h localhost -u retail_dba -p
(when prompted password use cloudera or any password that you have currently set)

b) create table retail_db.result(order_date varchar(255) not null,order_status varchar(255) not null, total_orders int, total_amount numeric,
constraint pk_order_result primary key (order_date,order_status));

c)

```
sqoop export \
--table result \
--connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
--username retail_dba \
--password cloudera \
--export-dir "/user/cloudera/problem1/result4a-csv" \
--columns "order_date,order_status,total_amount,total_orders"
```

86 comments:

Venkat Williams May 6, 2017 at 12:48 PM

Are you covering all possible problems in each scenario?

Reply

Replies

Arun Kumar Pasuparthi May 6, 2017 at 1:11 PM

While that is my goal, it may not be possible to cover all scenarios in a single problem questions. But the goal is to cover all exam topics in the series of problems. I am not targeting to post more than 10. So far i have posted 3 as of the time this comment is posted. I have 2 getting ready to be published by End of day today. Will do 5 more by end of this week with other possible combinations.

Unknown December 16, 2017 at 8:20 PM

Great Help buddy..

Reply

Venkat Williams May 6, 2017 at 12:54 PM

It would be if you could prepare each problem scenarios possibly inclusive of all exam objectives in future.. may be like 1- sqoop import, 1 -sqoop export/1- flume, 8- sprak questions.. (3- spark - etl, 4-spark sql, 1-spark configuraton). This structure would help us to consider these scenarios as real mock exams with solutions..

Thanks for your efforts on this so far..

Reply

Replies

Arun Kumar Pasuparthi May 6, 2017 at 1:10 PM

Sure, i will create the next set of problems in the pattern you suggested.

Venkat Williams May 6, 2017 at 1:22 PM

Thanks for accepting the suggestion. Appreciate your efforts.

Reply

SV May 8, 2017 at 1:08 PM

where can we get these sample datasets to come up with solution and test it on our vm? Thx

Reply

Replies

Arun Kumar Pasuparthi May 8, 2017 at 2:03 PM

All of datasets come prepackaged in the cloudera quickstart vm. I have not used anything that is not available in the QuickStart. Install it from link below link
Http://www.cloudera.com/downloads/quickstart_vms/5-10.html

SV May 8, 2017 at 2:15 PM

Thx. Will download and try it. I took the Cloudera course but that doesnt involve complicate aggregate functions etc. So this blog is helpful.

Arun Kumar Pasuparthi May 8, 2017 at 2:18 PM

Yes this blog will help, I have uploaded a video walk through of strategy as well as for each problem. Please go through the videos and let me know what you think

Reply

Sharad May 21, 2017 at 11:51 PM

Hi Arun, are you covering problems based on the new syllabus by cloudera?

Reply

Arun Kumar Pasuparthi May 22, 2017 at 3:34 AM

Yes. Please check the prep plan post where I explained the same in detail

Reply

Replies

Sharad May 22, 2017 at 6:15 AM

Thanks!

Reply

Bereket June 1, 2017 at 12:53 PM

good job it is helpful blog

Reply

Bereket June 1, 2017 at 12:54 PM

i tied step4c: to round the big decimal by adding this format "%.2f".format(x)

joinorders.map(x=>((x(1).toString,x(3).toString),(x(8).toString.toDouble,x(0).toString.toInt))).combineByKey((x:(Double,Int))=>(x._1,Set(x._2)),(x:(Double,Set[Int]),y:(Double,Int))=>(x._1+y._1,x._2 + y._2),(x:(Double,Set[Int]),y:(Double,Set[Int]))=>(x._1+ y._1,x._2++ y._2)).map(x=>(x._1._1,x._1._2,"%.3f".format(x._2._1),x._2._2.size))

Reply

Replies

vihary c December 16, 2017 at 8:25 PM

Need somehelp to understand combinebykey() .. please suggest good online material

Reply

Sarcos : lider de enlace y logistica June 1, 2017 at 2:01 PM

hello Arun

thanks for your blog , really useful, just one question. why do you use combineByKey instead of group by..because performance? or the exam ask for use it.. (still for me os complicated to use and I dont find eny site who can explain better)

Reply

Replies

Arun Kumar Pasuparthi June 4, 2017 at 5:11 PM

combineByKey gives more control. If you can solve the same problem using group by key then i recommend using it. however, remember that you may be asked questions to complete a programming sentence and hence you need to be prepared with using all major spark functions.

Reply

Bereket weldeslassie June 7, 2017 at 9:09 AM

thank you Arun. your blog helped me to clear exam cca175.

Reply

Replies

Arun Kumar Pasuparthi June 7, 2017 at 9:35 AM

Very very happy to know. Congratulations!

William KOUOGUE August 5, 2017 at 2:17 AM

Hello please which editor is use during CCA175 exam. Thank you

Reply

Satish Kumar Kakollu June 9, 2017 at 4:03 PM

Hi Arun,

While doing sqoop export do we have to takecare of anything say suppose my table in mysql is having 7 rows when I describe I see table with 7 columns with one field is having not null . I have to export using hdfs file . I used below query it did not work our please correct me

sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/retail_db --username --password --table departments --export-dir /hdfs/ocation --bath --columns colunames --batch --outdir output

anything wrong in the about command my job is failing with some file format issue . Please help me

Thanks,
Satish.

Reply

Replies

Arun Kumar Pasuparthi June 9, 2017 at 4:36 PM

your question is about export the command you posted is doing import. are you sure this is the command you want me to review? also i dont know what bath option is... i know you used batch which i kind of understand but not bath.

Reply

Satish Kumar Kakollu June 9, 2017 at 5:00 PM

Hi Arun ,

Miss typed it is export.

Thanks,
Satish.

Reply

Unknown June 10, 2017 at 12:23 AM

Hi Arun,

I took the exam yesterday and I met a problem when loading multiple files(parquet,avro) into the spark shell. However, today when I try directly load them into the spark shell just like

avroLoadingTest = sqlContext.read.format("com.databricks.spark.avro").load("/user/raku/test/avroLoadingTest")

everything is fine. Now I am confused by this thing, because before the test I had tried loading single file only, so I am not sure is loading multiple files in a directory in one time not supported or I just got some grammar mistake in the exam. Is it alright to use a single sentence like above to load multiple files in a directory into spark? Thank you.

P.S. The scenario you gave are really helpful

Reply

Replies

Arun Kumar Pasuparthi June 12, 2017 at 10:05 AM

yes you can load all the files when you specifiy a directory location. you can also load files from different directories at once.

adarsh singh June 27, 2017 at 12:38 PM

Hi Arun.. Is there possibility of having multiple types of files at one directory ? or we will get one file format at a time?
Thanks for your response in advance

Reply

bereket June 12, 2017 at 8:50 AM

Hi Arun if it is possible can you have spark Databricks exam certification on your blog.

Reply

Unknown June 20, 2017 at 10:41 PM

Hi Arun thanks for the post ...can i know this is how exactly question is exam?? i mean each question will have at least 10 task to complete? please let me know i'm going to take exam in next month

Reply

Arun Kumar Pasuparthi June 21, 2017 at 3:06 AM

each question will have one or two tasks to complete.

Reply

Chengalvala Abhishek June 21, 2017 at 4:28 AM

Hi Arun,

Are you planning to more problem scenarios or is the 7th one the last scenario?Your blog is really helpful.Thanks for creating the problem scenarios

Reply

Unknown June 22, 2017 at 11:05 AM

This comment has been removed by the author.

Reply

Vamsee Krishna Basineni June 22, 2017 at 11:16 AM

This comment has been removed by the author.

Reply

Replies

Arun Kumar Pasuparthi June 22, 2017 at 12:53 PM

Buddy you removed the comment by the time i enjoyed viewing it :). All the best for your CCA 175 exam buddy.

Vamsee Krishna Basineni August 14, 2017 at 7:23 PM

This comment has been removed by the author.

Arun Kumar Pasuparthi August 14, 2017 at 7:35 PM

congrats

Reply

taoufik elk July 1, 2017 at 6:40 PM

Hi Arun,

Really a huge effort you are doing here so big thank to you.

i have two questions:

1/how can i export parquet file into mysql using sqoop? you did the export for the csv file, but when i tried the same command using the parquet directory it gave me some exceptions.

2/Are you sure the queries in 4a, 4b and 4c give the same result? its not the case for me.

Thanks!

Reply

Replies

Roy Ryder July 29, 2017 at 1:30 PM

This comment has been removed by the author.

Reply

Arun Kumar Pasuparthi July 1, 2017 at 7:40 PM

Below are responses to both questions

1. Create a hive table that is backed by parquet file, then use hive table as source
2. Watch the video

Reply

Replies

Roy Ryder July 29, 2017 at 1:34 PM

Hi Arun --
Great content! Thanks for doing this.

Sorry, I agree with Taufik in that Results for 4a, 4b, and 4c do not match. SQL and DF result sets do but CombineByKey resultset does not match with the other two from second row on. For example, the second row total amount for dataframe and SQL res is 16333.16; while for combineByKey res for the same row is 10999.05 (off by about 4000). What you are doing appears to be correct -- I am not sure what need to be done differently, can't tell why the result for combineByKey is wrong. MySQL query supports the other two result sets.

Reply

Murali Rachakonda July 18, 2017 at 7:13 PM

Hi Arun, Thanks for a great block. I was going through the Sqoop documentation "sqoop-1.4.6-cdh5.10.1". I have two questions 1) Do we need to by-heart compression technique names because it is not mentioned in the documentation 2) Is the compression technique is mentioned in any other document? Thanks in advance.

Reply

Replies

Arun Kumar Pasuparthi July 18, 2017 at 8:08 PM

Working (hands on) knowledge of compression techniques is absolutely required not only for the exam but also on a day to day life as a big data developer. You need to memorize the compression methods and i recommend that you go through the file formats link of this blog. best of luck

Reply

Arun Kumar Pasuparthi July 18, 2017 at 8:06 PM

This comment has been removed by the author.

Reply

Karthika July 19, 2017 at 8:21 AM

Hello Arun,
Thanks for the posts . They are really helpful.
But I am confused with analyzing the right way to solve a given problem as there could be many solutions possible. For example , creating a hive table in say parquet file format can be done through Hive or sqoop import . How do I know which is the right solution during certification. Please suggest. Suggestions from people with certification experience will be helpful too.
Thanks once again for all the effort .

Reply

ssrk July 19, 2017 at 3:42 PM

Hi Arun,
Thank you for the blog. The questions are of great detail and the questions were well structured

Reply

Replies

Arun Kumar Pasuparthi July 19, 2017 at 8:05 PM

it depends. the exam may ask you to do it in a certain method or just ask you for the result. if the exams question is asking is you only for a result then you are free to choose whatever method you want. but always remember that spark allows a lot of flexibility whereas sqoop is very limited. so choose a technology that helps you solve the problem in the fastest way. Time is premium, i have heard from so many that they could not complete the exam as they took too much time debugging their procedure. so be careful choosing a solution..

Reply

Sagar Bunny July 23, 2017 at 1:17 PM

Hello Arun,
I have booked exam and waiting to take it once preperation is done. Am confusing a lot with scala i decided to learn more on dataframes as you have provided in problrm1 and 2 as it is easy compare to scala is it ok to choose dataframes in exam for all type of problems can you add some more videos regarding dataframes. that would be greatly appreciated.

Reply

Murali Rachakonda July 27, 2017 at 2:35 PM

This comment has been removed by the author.

Reply

Unknown July 28, 2017 at 12:18 PM

Hi Arun,

thanks for your blog.
from exam point of view should we be aware of all the below 3 types? If i'm good in sql, can i ignore a &c?? Please advise..

a). Just by using Data Frames API - here order_date should be YYYY-MM-DD format
b). Using Spark SQL - here order_date should be YYYY-MM-DD format
c). By using combineByKey function on RDDS -- No need of formatting order_date or total_amount

Reply

Replies

Arun Kumar Pasuparthi August 20, 2017 at 6:10 AM

be aware of all three. you may be asked to fill in the blanks and execute the program.awareness about all three will make you better prepared.

Reply

aakash68404 July 30, 2017 at 5:45 AM

Hello , I have been trying the problems scenarios given in your blog . but could you please provide the sample files for order-items so we can try in our laptop .for example lets say problem 1 sqooping order-items data . So I need sample data for this. Appreciate your response :)

Reply

Arun Kumar Pasuparthi July 30, 2017 at 5:51 AM

Buddy, you need to have CDH downloaded from cloudera. CDH has all the data in mysql database that comes with it. Watch the video for an understanding how that data is pulled from mysql to hdfs.

Reply

Unknown August 5, 2017 at 2:10 AM

This comment has been removed by the author.

Reply

William KOUOGUE August 5, 2017 at 2:23 AM

Hello please which editor is use during CCA175 exam. Thank you

Reply

Replies

Arun Kumar Pasuparthi August 20, 2017 at 6:09 AM

use gedit.

Reply

Naseer Ahmedf August 20, 2017 at 5:47 AM

Thanks for your blog.
I have one question. Is it mandatory to practice the questions on Cloudera Platform only. I have experience working on Hortonworks platform. Can I directly attempt the exam by practicing on Hortonworks Platform

Reply

Replies

Arun Kumar Pasuparthi August 20, 2017 at 6:08 AM

practice on cloudera platform. this gives you acquaintance. Even seasoned big data developers failed the CCA 175 exam recently. not due to complexity of the exam but due to unfamiliarity of the environment. they ran out of time and hence could not complete the required number of problem scenarios.

Reply

Unknown August 23, 2017 at 1:55 PM

var dataFrameResult =
dataFrameResult.show();

ERROR IN MY MACHINE PLEASE REPLY ME

Reply

aswin R August 27, 2017 at 6:58 PM

Hi Arun,

I have cleared CCA spark and Hadoop Developer because of your blogspot.

Thank you very much!! Hope to see more problem scenarios more like this.

Regards,
Aswin Ramakrishnan

Reply

sathya September 1, 2017 at 5:52 AM

Superb explanation & it's too clear to understand the concept as well, keep sharing admin with some updated information with right examples.Keep update more posts.

Informatica Training in Chennai

Dataware Housing Training in Chennai

Reply

Manish Tewari September 15, 2017 at 2:55 AM

Hi Arun sir Can I get the data of problems which you have explained.

Reply

Unknown October 4, 2017 at 9:31 AM

Beautiful Blog.....Great content presented in the best possible way that the entire big data knowledge can be streamlined

Reply

Annoyed October 27, 2017 at 7:54 PM

This comment has been removed by the author.

Reply

Annoyed October 27, 2017 at 7:56 PM

Hello Arun - I think you have done the coding in Scala in all your videos so if someone wants to take the exam using Python do you have the code snippet for the same ? If yes can you please share the same -Auro

Reply

Trep Helix November 7, 2017 at 2:00 AM

Hi Arun,

I have booked exam and holding up to take it once planning is finished. Am mistaking a considerable measure for Scala I chose to take in more on information outlines as you have given in issue and 2 as it is simple contrast with Scala is it alright to pick information outlines in exam for all kind of issues would you be able to include some more recordings in regards to information outlines. that would be significantly refreshing.

Regards,
Trep Helix

Reply

crescendo November 7, 2017 at 11:55 PM

Hi Arun,

In certificate exam, this avro databrocks will work in the spark-shell. Because as per cloudera they do not provide "Databricks API" in the CDH install

import com.databricks.spark.avro._;

Reply

Doubts Several November 8, 2017 at 3:31 AM

Hi Arun,
I've WATCHED YOUR YOUTUBE VIDEO but I still DON'T UNDERSTAND THIS part of 4c exercise. I'm totally lost with the combineByKeyResult:

```
combineByKey((x:(Float, String))=>(x._1,Set(x._2)),
(x:(Float,Set[String]),y:(Float,String))=>(x._1 + y._1,x._2+y._2),
(x:(Float,Set[String]),y:(Float,Set[String]))=>(x._1+y._1,x._2++y._2)).
map(x=> (x._1._1,x._1._2,x._2._1,x._2._2.size))
```

Completeley lost with the combineByKey. Could you please explain me what you are doing here?
I've got my exam in two days.

Thank you

Reply

Cynix Technologies November 8, 2017 at 5:02 AM

nice blog keep updating your blog and i am waiting for your next update alsoBig Data Hadoop Online course Hyderabad

Reply

sohil shivani November 10, 2017 at 2:18 AM

HI Arun,

This is really informative blog..

But CCA-175 exam content is changed. Do you help with new playlist for the same?

Reply

Replies

Venkat Williams November 10, 2017 at 2:25 AM

Hi Sohil,

These would suffice the CCA175 exam content as well.

Reply

Loving November 15, 2017 at 2:28 AM

This comment has been removed by the author.

Reply

Jitesh November 15, 2017 at 2:31 AM

Hi Arun/Venkat,

Can we submit more than one file as output because "spark.sql.shuffle.partitions" value is more than 1, by default. OR we have submit one output file only.

Reply

Venkat Williams November 15, 2017 at 3:28 AM

Hi Jitesh,

By default "spark.sql.shuffle.partitions" is set to 200. This can be even configured to value 1 based performance tuning options.

Reply

Jitesh November 16, 2017 at 4:48 AM

Hi Venkat,

Thanks for your answer, But my question is still same, can we submit more than file? Because it will save time during exam.

Reply

Venkat Williams November 16, 2017 at 4:59 AM

If you are expecting yes or no kind of answer. Answer is YES.

Reply

imthiyas aalam November 27, 2017 at 2:40 PM

This comment has been removed by the author.

Reply

imthiyas aalam November 27, 2017 at 2:43 PM

I am trying to use aggregateByKey instead of combineByKey.but getting issue with count rest all looking good...could you please let me know what i am missing here,

```
val ordersDF = sqlContext.read.avro("/user/imthiyas90/problem1/orders")
val orderitemDF = sqlContext.read.avro("/user/imthiyas90/problem1/order_items")

ordersDF.registerTempTable("orders")

orderitemDF.registerTempTable("order_items")

val joinedDF = ordersDF.join(orderitemDF, ordersDF("order_id") === orderitemDF("order_item_order_id"))

val combine_agg = joinedDF.map(e => ((e(1).toString,e(3).toString),(e(8).toString.toDouble,e(0).toString.toInt))).aggregateByKey((0.0,0))((x:(Double,Int),y:(Double,Int)) =>(x._1+y._1,x._2+1),(x:(Double,Int),y:(Double,Int)) =>(x._1+y._1,x._2+y._2)).map(x => (x._1._1,x._1._2,x._2._1,x._2._2)).toDF.orderBy(col("_1").desc,col("_2"),col("_3").desc,col("_4"))
```

Reply

Aparna Sen December 1, 2017 at 3:07 AM

This comment has been removed by the author.

Reply

Aparna Sen December 1, 2017 at 3:09 AM

Hi Arun,
Thanks for providing such a wonderful blog for those who aspire to clear CCA Spark And Hadoop Certification. But what I found is all of the solutions are explained in scala, I am learning Pyspark and don't know scala. Could you please tell if you have the same solutions in Python as well?
Thanks

Reply

Sumanth Sai December 3, 2017 at 7:54 PM

Hi Arun,
I tried from_unixtime. But is returning null values.

Reply

soukya reddy December 11, 2017 at 1:45 AM

Thanks for good info.It's very good blog
Big data Hadop Online course Hyderabad

Reply

Jeyassri Balachandran December 28, 2017 at 7:03 AM

Hi, i appreciate your great efforts! Is it necessary to round the sum of order_item_subtotal to 2 decimal values? They have not mentioned in the question.

Reply

ausgeek January 1, 2018 at 4:37 PM

when it comes to saving as csv can we do following steps:-

```
val resultRDD = resultDF.rdd.
map(_.toString().replace("[","").replace("]","")).
saveAsTextFile("/user/cloudera/problem1/result4a-csv")
```

Reply

Vamsi January 2, 2018 at 7:23 PM

Hi Arun,

Thanks a ton for these exercises and the videos.

I have cleared my CCA 175 today. These exercises has helped a lot in my preparation and gave the confidence to take up the test.

Thanks
Vamsi

Reply

Replies

Funjab Kaur January 3, 2018 at 10:34 AM

This comment has been removed by the author.

Funjab Kaur January 3, 2018 at 10:42 AM

many congratulations for passing the exam. I am going to appear for exam in few days. I just wanted to know did you get templates in exam? Was there any question about avro? How many questions were there from each category? Your answers can really help since you have just given it recently. i will really appreciate if you will revert.

Reply

Home

Subscribe to: Posts (Atom)

If you have landed on this page then you are most likely aspiring to learn Hadoop ecosystem of technologies and tools. Why not make you...

(no title)
If you have landed on this page then you are most likely aspiring to learn Hadoop ecosystem of technologies and tools. Why not make you...