G+ More ▾ Next Blog» Create Blog Sign In

Hadoop and Spark Developer - CCA 175

Problem Scenario 3

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 this solution is available at ***[Click Here]***

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

Problem 3: Perform in the same sequence

1. Import all tables from mysql database into hdfs as avro data files. use compression and the compression codec should be snappy. data warehouse directory should be **retail_stage.db**
2. Create a metastore table that should point to the orders data imported by sqoop job above. Name the table **orders_sqoop**.
3. Write query in hive that shows all orders belonging to a certain day. This day is when the most orders were placed. select data from **orders_sqoop**.
4. query table in impala that shows all orders belonging to a certain day. This day is when the most orders were placed. select data from **order_sqoop.**
5. Now create a table named **retail.orders_avro** in hive stored as avro, the table should have same table definition as order_sqoop. Additionally, this new table should be partitioned by the order month i.e -> year-order_month.(example: 2014-01)
6. Load data into orders_avro table from orders_sqoop table.
7. Write query in hive that shows all orders belonging to a certain day. This day is when the most orders were placed. select data from **orders_avro**
8. evolve the avro schema related to **orders_sqoop** table by adding more fields named (order_style String, order_zone Integer)
9. insert two more records into orders_sqoop table.
10. Write query in hive that shows all orders belonging to a certain day. This day is when the most orders were placed. select data from **orders_sqoop**
11. query table in impala that shows all orders belonging to a certain day. This day is when the most orders were placed. select data from **orders_sqoop**

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-all-tables \
--connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" \
--username retail_dba \
--password cloudera \
--warehouse-dir /user/hive/warehouse/retail_stage.db \
--compress \
--compression-codec snappy \
--as-avrodatafile
-m 1;
```

Step 2:

```
hadoop fs -get /user/hive/warehouse/retail_stage.db/orders/part-m-00000.avro
avro-tools getschema part-m-00000.avro > orders.avsc
hadoop fs -mkdir /user/hive/schemas
hadoop fs -ls /user/hive/schemas/order
hadoop fs -copyFromLocal orders.avsc /user/hive/schemas/order
```

Launch HIVE using 'hive' command in a separate terminal

Below HIVE command will create a table pointing to the avro data file for orders data

```
create external table orders_sqoop
STORED AS AVRO
LOCATION '/user/hive/warehouse/retail_stage.db/orders'
TBLPROPERTIES ('avro.schema.url'='/user/hive/schemas/order/orders.avsc')
```

Step 3-Run the query in Hive:
Run this query in Hive.

```
select * from orders_sqoop as X where X.order_date in (select inner.order_date from (select Y.order_date, count(1) as total_orders from orders_sqoop as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) inner);
```

Step 4-Run the query Impala:
Lanch Impala shell by using command impala-shell

1. Run 'Invalidate metadata'
2. Run below query

```
select * from orders_sqoop as X where X.order_date in (select a.order_date from (select Y.order_date, count(1) as total_orders from orders_sqoop as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) a);
```

Step 5 and 6:

Search This Blog

Search

CCA 175 Hadoop and Spark Developer Preparation

- Home
- CCA 175 Prep Plan
- Problem Scenario 1
- Problem Scenario 2
- **Problem Scenario 3**
- Problem Scenario 4
- Problem Scenario 5 [SQOOP]
- Problem Scenario 6 [Data Analysis]
- Problem Scenario 7 [FLUME]
- File Formats
- Youtube Playlist

A leader with a unique blend of deep technology expertise and strong management skil

Arun Kumar Pasuparthi

G+ Follow 212

View my complete profile

Report Abuse

Blog Archive

April 2017 (1)

```
create database retail;

create table orders_avro
    > (order_id int,
    > order_date date,
    > order_customer_id int,
    > order_status string)
    > partitioned by (order_month string)
    > STORED AS AVRO;

 insert overwrite table orders_avro partition (order_month)
select order_id, to_date(from_unixtime(cast(order_date/1000 as int))), order_customer_id, order_status,
substr(from_unixtime(cast(order_date/1000 as int)),1,7) as order_month from default.orders_sqoop;
```

Step 7 - Query Hive

```
select * from orders_avro as X where X.order_date in (select inner.order_date from (select Y.order_date, count(1) as total_orders from
orders_avro as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) inner);
```

Step 8 - Evolve Avro Schema

```
1. hadoop fs -get /user/hive/schemas/order/orders.avsc
2. gedit orders.avsc

3.{
  "type" : "record",
  "name" : "orders",
  "doc" : "Sqoop import of orders",
  "fields" : [ {
    "name" : "order_id",
    "type" : [ "null", "int" ],
    "default" : null,
    "columnName" : "order_id",
    "sqlType" : "4"
  }, {
    "name" : "order_date",
    "type" : [ "null", "long" ],
    "default" : null,
    "columnName" : "order_date",
    "sqlType" : "93"
  }, {
    "name" : "order_customer_id",
    "type" : [ "null", "int" ],
    "default" : null,
    "columnName" : "order_customer_id",
    "sqlType" : "4"
  },{
    "name" : "order_style",
    "type" : [ "null", "string" ],
    "default" : null,
    "columnName" : "order_style",
    "sqlType" : "12"
  }, {
    "name" : "order_zone",
    "type" : [ "null", "int" ],
    "default" : null,
    "columnName" : "order_zone",
    "sqlType" : "4"
  }, {
    "name" : "order_status",
    "type" : [ "null", "string" ],
    "default" : null,
    "columnName" : "order_status",
    "sqlType" : "12"
  } ],
  "tableName" : "orders"
}

4. hadoop fs -copyFromLocal -f orders.avsc /user/hive/schemas/order/orders.avsc
```

Step 9 - Insert 2 records from Hive shell

```
insert into table orders_sqoop values (8888888,1374735600000,11567,"xyz",9,"CLOSED");
insert into table orders_sqoop values (8888889,1374735600000,11567,"xyz",9,"CLOSED");
```

Step 10 -Run the query in Hive:

Run this query in Hive.

```
select * from orders_sqoop as X where X.order_date in (select inner.order_date from (select Y.order_date, count(1) as total_orders from
orders_sqoop as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) inner);
```

Step 11-Run the query Impala:

Lanch Impala shell by using command impala-shell

1. Run 'Invalidate metadata'
2. Run below query

```
select * from orders_sqoop as X where X.order_date in (select a.order_date from (select Y.order_date, count(1) as total_orders from
orders_sqoop as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) a);
```

30 comments:

Bapu May 20, 2017 at 3:03 AM

Hi, Arun. Let me say Thank you for exercises. please provide more(no problem without videos).

For Task-1, is it really necessary to use -m 1? please explain.

Reply

Replies

Arun Kumar Pasupathi May 21, 2017 at 6:34 PM

-m 1 (i.e number of mappers) can be anything based on the number of nodes you want to leverage in a cluster to number of files you want sqoop to create. I used 1 in order to help the flow of the video series. one file is easier to compare, query and manage in terms of size and data while walking through the solution. This is not mandatory.

Bapu May 22, 2017 at 12:24 AM

Thank You.
&
Expecting more exercises from you.

Reply

Bala Hassan June 20, 2017 at 12:35 AM

Hi Arun,
That's a nice blog out there, am learning a lot of things and thanks for the same!

I have a question on documentation available during exam. Will we have access to the complete documentation(as listed in cloudera website) or will it be a stripped down version of the online documentation?

Thanks,
Hassan.

Reply

Arun Kumar Pasupathi June 20, 2017 at 3:27 AM

Go to the link https://www.cloudera.com/more/training/certification/cca-spark.html and scroll to the bottom. You will have access to the same documentation that is available on this page during the exam as well. But i recommend dont rely on documentation as your primary or secondary means. it should be a final resort in case you cannot find the answer in your brain or via help commands.

Reply

Varun Mishra June 21, 2017 at 2:43 AM

Hi Arun,

This query does not support to HiveContext in spark.

Reply

Replies

Arun Kumar Pasupathi June 21, 2017 at 3:05 AM

can you post the query you are referring to?

Varun Mishra June 27, 2017 at 12:05 AM

step-3,
The same query working in hive but not in spark in hive context mode.
I don't it is question for me also?

select * from orders_sqoop as X where X.order_date in (select inner.order_date from (select Y.order_date, count(1) as total_orders from orders_sqoop as Y group by Y.order_date order by total_orders desc, Y.order_date desc limit 1) inner);

Reply

SriniNN June 25, 2017 at 7:23 PM

Hi Arun, Appreciate your efforts for putting all the material together. Have a few Qs, please respond.
1. Does each Q have sub-questions like the one you have above? I see scenario 3 and there are 11 sub-Q underneath it. Have you put these together just to signify all the valid Qs on the topic?
2. For some of the Qs, there is no mention of where I should store the result. Do I have to create files in my HOME and name them 1a,1b to signify the Q and the sub-Q? I have read that they dont look at the type of solution, but only interested in the result.
3. I also thought that since they are only interested in the stored result, it doesn't matter which lang we use. As long as i am comfortable in 1 lang (eg: scala), you don't need to know Python or Java
4. Coming from the HortonWorks world, I don't know Impala, do I need to learn Impala?
5. A Spark problem can be solved using spark core, sql or DF. Do I need to know all the functions in a regular Spark core or can I solve this using Spark SQL as I have more familiarity with SQL

Reply

Arun Kumar Pasupathi June 25, 2017 at 8:19 PM

1. Does each Q have sub-questions like the one you have above? I see scenario 3 and there are 11 sub-Q underneath it. Have you put these together just to signify all the valid Qs on the topic?
Ans: in the real exam, dont expect so many subquestions. As i mentioned in the introduction page, this is just to test your knowledge.
2. For some of the Qs, there is no mention of where I should store the result. Do I have to create files in my HOME and name them 1a,1b to signify the Q and the sub-Q? I have read that they dont look at the type of solution, but only interested in the result.
A. in the exam, you will be given clear instructions on where to store the result. For the practice mode i created, you can store wherever you want if the destination is not predetermined in the question. once again, the intent is to help you practice the concepts and not get lost in the nuances which can be tackled with common sense during the exam.
3. I also thought that since they are only interested in the stored result, it doesn't matter which lang we use. As long as i am comfortable in 1 lang (eg: scala), you don't need to know Python or Java
A. if the question is asking you for a result then it does not matter which language you use. however, if the question is asking you to complete a partially created solution to a problem then you will have to understand phython as well just in case that partially created solution is in python.
4. Coming from the HortonWorks world, I don't know Impala, do I need to learn Impala?
A. Impala is probably the easiest to learn. you just have to type impala-shell. Majority of the syntax is same as hive. Most importantly, impala works so fast that you will love it better than hive.
5. A Spark problem can be solved using spark core, sql or DF. Do I need to know all the functions in a regular Spark core or can I solve this using Spark SQL as I have more familiarity with SQL
A. I recommend that you know all, this will equip you to face 'fill in the blanks' questions.

I hope this helps. Sorry for the typos and grammatical mistakes. Good luck.

Reply

SriniNN June 25, 2017 at 8:46 PM

Thank you Arun. Your response cleared my apprehensions. Much appreciated.

Reply

SriniNN June 26, 2017 at 10:57 AM

Arun, Good Morning. Can you please tell me if they would have SBT and Scala installed On the Cloudera VM on the exam?

Thank you

Reply

Replies

Arun Kumar Pasuparthi June 26, 2017 at 12:05 PM

Yes. You can use spark-shell for all the spark problems in the exam if they are related to output only. However, if they are related to completing a portion of code then you will have all the required tools.

Reply

David Boudart June 26, 2017 at 4:10 PM

Hello Arun
are you going to do a similar blog like this one (which I congrat you for this excelent work) but for the Data Engineer exam? CCP 575.

Best regards,
DB

Reply

Replies

Arun Kumar Pasuparthi July 6, 2017 at 6:56 AM

I may not David. Sorry about that :(

Reply

taoufik elk July 6, 2017 at 6:46 AM

Hi Arun,

in the question 2 you are talking about "metastore table", i think this is not a good name as it create a confusion because a "metastore" in hive is a central repository that contain all the metadata (table schemas ...) of the hive data. so i suggest to name it just "hive table" instead of "metastore table". what do you think? i'm making sense ?

Reply

Replies

Arun Kumar Pasuparthi July 6, 2017 at 6:55 AM

I understand your question. However, see what CCA 175 web page on the certification syllabus says. I tried to be consistent with CCA 175 terminology.

This text is copied from CCA 175 web page. 'Use meta store tables as an input source or an output sink for Spark applications'

Reply

vivek munjal July 13, 2017 at 5:22 AM

Hello Arun,

When i run the below query it gets failed

select * from orders_sqoop as x where x.order_date = (select ji.order_date from (select y.order_date, count(y.order_id) count11 from orders_sqoop y group by y.order_date order by count11 desc limit 1) ji);

With the error

FAILED: ParseException line 1:55 cannot recognize input near 'select' 'ji' '.' in expression specification

but when i put "in" instead of "=" it works. Can you explain this behaviour or is it the way hive works

Reply

Unknown July 14, 2017 at 1:27 PM

Hello Arun

This topic is about avro files but in new syllabus they taken out avro files.still i need to learn about avro files?

Reply

Arun Kumar Pasuparthi July 14, 2017 at 1:39 PM

Absolutely.......you need to know this

Reply

Sagar Bunny July 14, 2017 at 2:27 PM

so for cca 175 as latest update i need to avro files

Reply

Sagar Bunny July 14, 2017 at 2:30 PM

can you please tell me under which category this avro files come Data Ingest --- Transform, Stage, and Store ----Data Analysis

Reply

Murali Rachakonda July 28, 2017 at 12:24 PM

Hi Arun, Thanks for all your efforts. In step 8.evolve the avro schema related to orders_sqoop table by adding more fields named (order_style String, order_zone Integer). Is there any specific reason why you have not added new columns after order_status (which was the last field). Thanks.

Reply

Allan Mercader July 31, 2017 at 8:32 PM

This comment has been removed by the author.

Reply

▾ Replies

Allan Mercader July 31, 2017 at 9:13 PM

Never mind this... I got my answers towards the end of your video. Again, thank you from making this blog.

Reply

Lakshmi Thiagarajan September 26, 2017 at 8:11 AM

Hi Arun
Thanks for all the problem statements, its testing all our knowledge and equipping us to sit on time bound Exam.

Also , we will have to getschema using avro-tools from the .avro file when u have only the .avro file available . We do get the orders.avsc file and orders.java file when u run Sqoop import on orders and we can find it in the folder where we ran the Sqoop import from.

We can move that .avsc into HDFS to create the Hive table and evolve the schema in the same. Just a way to save time , for exam takers.

Reply

簡小宇 November 27, 2017 at 3:03 AM

Hi Arun,

I encountered the following error messages when doing step 5 and 6:

FAILED: SemanticException [Error 10096]:
Dynamic partition strict mode requires at least one static partition column.
To turn this off set hive.exec.dynamic.partition.mode=nonstrict

Please help to add the following instruction in solutions :

hive>set hive.exec.dynamic.partition.mode=nonstrict;

Please correct me if I am wrong, thanks :)

Again, thank you for providing these good practices to us. :)

Reply

▾ Replies

Arun Kumar Pasuparthi November 27, 2017 at 8:55 AM

i think my CDH VM had this setting my default and hence it was not needed. I am glad you found a fix to the problem. I cannot update the video now. but will update the posting shortly

Reply

imthiyas aalam November 27, 2017 at 3:19 PM

problem says "data warehouse directory should be retail_stage.db". it meant to say always it is /user/hive/warehouse/ even if it is not mentioned in Q..

Reply

Unknown November 28, 2017 at 10:28 PM

1.) You have used different code for snappy compression in Problem 1 and this problem. Are both the same?

Reply

Enter your comment...

Comment as: Select profile... ▾

Publish **Preview**

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...

Simple theme. Powered by Blogger.