

CS 5620 – Take-Home Final Exam

First Name: SAI KUMAR

Last Name: CHILUVERI

Student 700 Number: 700739829

Instructions:

1. PLEASE ANSWER ALL QUESTIONS.
2. PLEASE WRITE THE QUESTION NUMBER AND ITEM NEXT TO EACH OF YOUR ANSWERS (Examples: Question 2., Question 3.c., Question 4.g. etc.)
3. WRITE YOUR ANSWERS ON CLEAN SHEETS.
4. WRITE BRIEFLY AND NEATLY.
5. PLEASE SCAN YOUR ANSWER SHEET THEN UPLOAD IT INTO BLACKBOARD. YOU ARE ONLY ALLOWED ONE ATTEMPT. NO EMAIL SUBMISSION WILL BE ACCEPTED
6. PLEASE WRITE YOUR FIRST NAME, LAST NAME, AND NUMBER.
7. PLEASE TURN IN YOUR GENUINE ANSWERS. DON'T SHARE YOUR ANSWERS WITH ANYONE. ANY VIOLATION FOUND WILL RESULT IN FAILING THE COURSE AND BEING REPORTED TO THE DEPARTMENT.
8. IF YOU HAVE NO ACCESS TO A SCANNER, YOU CAN USE A PHONE APP TO SCAN YOUR ANSWER SHEETS. ALL OF YOUR ANSWERS MUST BE SUBMITTED IN ONE PDF FILE ON BLACKBOARD.
9. GOOD LUCK!

Q.1. What are the three configuration modes for running Hive cli service with respect to the metastore service and the metastore database? Briefly, state the difference between them.

Q.2. **True or False:** In order to run Hive queries in a Hadoop cluster, Hive must be installed on every node in the cluster. Justify your answer.

Q.3. Write the necessary commands in Hive to find the total length of all lines of a file that exists in HDFS. Assume that the filename is transactions.txt and it is stored under the user's home directory. The function length(<input-str-argument>) can be used to find the length of strings in Hive. You need to create and populate one or more tables to solve the problem, the fewer the number of tables the better is your answer. You should be able to figure out the commands without actually using Hive console.

Q1: Three modes for hive metastore are .

1. Embedded Metastore
2. Local Metastore
3. Remote Metastore

Embedded Metastore

In embedded Metastore both metastore service and hive service runs in the same JVM by using embedded Derby database. This mode has a limitation that it allows only a single user session to connect to metastore. If we try to open second session it results in error. It is not suggested for production.

Local Metastore

To overcome the limitation of Embedded metastore, this mode allows to have many hive sessions. This can be achieved by using any JDBC database like MySQL which runs in different machine than of hive service and metastore service running in the same JVM.

Remote Metastore

In remote mode, metastore runs on its own separate JVM, not in the hive service JVM. Thrift Network API's can be used to communicate with the metastore server & other process.

Q2:

False;

Because Hive is a Hadoop client and it runs on top of Hadoop it's not required to install Hive on every node in the cluster to run the Hive queries.

Q3:

```
Create Table test (line string);
load data inpath 'text'.
overwrite into table test;
select sum (length (line)) from test;
```


Question 4:

- Show database;
- Show tables;
- Show tables in companydb;
- Select * from companydb.employees;
- Select * from companydb.products limit 5;
- Set hive . execution . engine;
- Set hive . metastore . warehouse . dir;
- To display the content under users home directory
dfs - cat /*;
To list the files
!dfs - ls/;
- in prod
- Load data local inpath '/*oo.txt' into table mytable;
- Describe extended mytable; (or)
Describe for mapped mytable;

Question 5:

Explode function is used to split output in individual tokens rather than a list.

Example: welcome
to
programming
hive b.

Question 6:

Schema on Write: In traditional database, database has control over storage. Here data is being checked against (Control over storage). Here data is being checked against the schema when written into db during check load.

Schema on Read: Hive has no control over storage and the data scheme is not verified while processing the query. Hive uses this process called schema on Read.

Question 7:

To share the data between tools we create external table where external tables doesn't take ownership of data. When we delete the (drop) external table it does not delete the data but metadata for that table will be deleted.

Create external Table If not exists mytable c

student-id int,
student name string)

Row format Delimited fields terminated by ','
location 'data / data set - 2020';

Question 8:

8.a: Create Table customers C

cust - id int,

cust - name string,

street string,

city string,

zip int,

region string)

partitioned by (country string);

8.b: To load the data into customers table without moving files in HDFS, we will create one staging table.

create table staging.table C

cust - id int,

cust - name string, street string,

city string,

zip string,

region string)

Partitioned by (country string);

load data inpath 'data/customers/canada'

into table staging.table.

8.d: Based on filter condition, where clause on partition table, we no need to scan thousands of records it is only necessary to scan thousands of records.

8.e: Yes, we can use dynamic partition to avoid creating Id of partitions as we have many countries. In the 8.b, we have loaded data using single Sql

Question 9: lines = sc.textFile('data/logfiles/*')
lines.count()

Question 10: error-rdd = lines.filter(lambda, line: "error" in line.lower())
error-rdd.count()

Question 11: string-count = lines.map(lambda line: len(line))
string-count.sum()

Question 12: Import numpy as np (Requires only if we use it)
numbers = sc.parallelize(range(0, 100))
numbers.sum()

Question 13:
Lazy Evaluation in spark means execution of RDD will not start until an action is triggered. This occurs when spark transformation comes in picture. spark maintain the record of operations called through DAG (Directed Acyclic Graph) since transformations are lazy in nature we can execute operation by calling an action only.
Example: filter.
→ ~~Not~~ It does not apply to every operation in spark, lazy evaluation comes for transformations only.