CS 5620 – Take-Home Final Exam

First Name: ....Mareedu.....

Last Name: .....Sai Sritha.....

Student 700 Number: .....700 74 29050.....

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.

Question-1:

Three modes for hive metastore are:

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

Embedded Metastore	Local Metastore	Remote Metastore
In Embedded metastore both metastore service and hive services runs in the same JVM by using Embedded Derby database. This model 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. This mode is not suggest for production.	To overcome the limitation of embedded metastore, this mode allows to have many hive sessions this can be achieved by using any JDBC databases like MySQL. Which runs in a different machine than that of hive service and metastore service running in the same JVM.	In Remote Mode, Metastore runs on its own seperate JVM, not in the hive Service JVM. Thrift network API's can be used to communicate with the metastore Server and other processes.

Question-2:

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.

Question-3 :-

```
Create table test (line String);
load data inpath 'test'
overwrite into table test;
Select Sum (length (line)) from test;
```

Question - 4 :

a) Show databases ;

b) Show tables ;

c) Show tables in Companydb ;

d) Select * from Companydb . employees ;

e) Select * from Companydb . products limit 5 ;

f) Set hive . ~~extens~~ execution . engine ;

g) Set hive . metastore . ware house . dir ;

h) To display the content under users home directory. dfs -cat /* ;

to list the files

dfs -ls / ;

i) ! pwd ;

j) load data local inpath 'E:/foo.txt' into table mytable ;

k) describe extended mytable ; (or) describe formatted mytable ;

Question - 5 :

Explode Function is Used to Split o/p in individual token rather than a list.

Ex:- Welcome
to
programming Hive !

Question-6 :

→ **Schema On Write** : In traditional database, db has control over storage. Here data is being checked against the schema when written into the db during load.

→ **Schema On Read** : Hive has no control over storage and the data schema is not verified during load time, rather it is verified while processing the query.

Hive uses this process called Schema on Read.

Question-7 :

To share the data between tools we create external tables where external tables doesn't take owner ship 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 (
                    Student_id int,
                    Student_name string,
        Row format Delimited fields Terminated by ','
        Location '/data/dataset-2020';
```

Question - 8:

8.a)
```
Create Table Customers (
    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 (
        cust_id int,
        cust_name string, street string,
        city string,
        zip string,
        region string)
    partitioned by (country string);

Local data inpath '/data/customers/usa'
  into table staging_table
  partition (country = 'usa');

Load data inpath '/data/customers/canada'
  into table staging_table
  partition (country = 'canada');

Load data inpath '/data/customers/mexico'
    into table staging_table
  partition (country = 'mexico')
```

Now, In order to allow dynamic partition first set the below properties in hive.

```
Set hive.exec.dynamic.partition = true;
Set hive.exec.dynamic.partition.mode = nonstrict;
```

Now insert the data to customers from staging_table

```
Insert into table customers partition (Country)
Select cust_id, cust_name, street, city, zip, region, Country from
                                                    stagging_table.
```

8.c) Yes, We can use dynamic partition to avoid creating lot of partitions as we have many countries. In the above question 8.b we have loaded data using single SQL query using dynamic partition.

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 the contents of one directory by this way for larger datasets, partition can dramatically improve query performance.

Question-9:

```
Lines = sc.textFile('./data/log files/*')
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 - not required as we are not using 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. lazy evaluation occurs when Spark transformation comes in picture. Spark maintains 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 Ex:- filter.

No, it does not apply to every operation in Spark, lazy evaluation comes for transformations only.