CS 5620 – **Take-Home Final Exam**

First Name: …….

Last Name: …….

Student 700 Number: …….

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.

Q.4. Assume that we have started Hive shell and we are connected to the **default** database. Give a command in Hive to achieve each of the following tasks:

a) To list existing databases.
`(default)>`
b) To list existing tables.
`(default)>`
c) To list all existing tables in a database called **companydb**.
`(default)>`
d) To list all records from **employees** table that exists in **companydb**.
`(default)>`
e) To list only five records from **products** table that exists in **companydb**.
`(default)>`
f) To show the current execution engine.
`(default)>`
g) To show the default file system in Hadoop.
`(default)>`
h) To display the content of the user's home directory in HDFS.
`(default)>`
i) To print the current working directory.
`(default)>`
j) To append data from a file stored locally in the current working directory into an existing table called **mytable**. Let the filename be `foo.txt`
`(default)>`
k) To find whether the table **mytable** is an external table or not.
`(default)>`

Q.5. What is the effect of calling **explode** on the output of **split** in Hive? Consider for example
`explode(split('Welcome to Programming Hive!', ' '))`

Q.6. Briefly state the difference between *schema on read* and *schema on write*. Which one is used by Hive?

Q.7. When should we consider using external tables in Hive? Give the command in Hive shell to create a two-column external table, called **mytable**. Your command should also populate the table using **comma delimited** files found under the path '/data/dataset-2020'. Feel free to pick the column names and datatypes.

Q.8. Assume that we have customer data in HDFS organized into a directory structure by country as follows:

'/data/customers/usa'
'/data/customers/canada'
'/data/customers/mexico'
etc.

Assume also that Hive default delimiters hold for the data files. Each line in the files gives the customer id (**cust_id**), name (**cust_name**), street address (**street**), **city**, zip code (**zip**), and state/region (**region**).

a) Write a Hive command to create a partitioned table for the customer data, partitioned by country. Please choose appropriate column data types.
b) Write Hive commands to populate at least three partitions without moving any files in HDFS.
c) Can we use Hive dynamic partitioning for this scenario? Briefly justify your answer.
d) Briefly state how partitioning in Hive can speedup query evaluation.

Q.9. Give the command(s) in Spark python shell to find the total number of lines in all the files stored under the HDFS directory: '/data/logfiles'

Q.10. Repeat (Q.9) but now we are just interested in those lines that contain the word 'error', case-insensitive.

Q.11. For the same files in (Q.9.), give the command(s) in Spark python shell to find the total number of characters. Please note that in python we can find the length of a string by using the `len()` function.

Q.12. Give the command(s) in Spark python shell to find the sum of all the values in the range 0-99.

Q.13. Briefly, justify lazy evaluation in Spark? Does it apply to every Spark operation?