Homework #5 **Due: turned in by Mon 2/12/2020 before class**

Adil Ashish Kumar

(put your name above)

Total grade: _____ out of ___100___ points

General Submission Guidelines

The answers for homework assignments should be submitted in a PDF file. When the homework involves script files (e.g. pig, hive, or python scripts), the script files should be submitted in addition to the PDF for purpose of easy-debugging. Such scripts should be emailed to managingbigdata.msba.emory@gmail.com

Part I: Multiple Choice Questions (20 points)

A. **HDFS enhanced authentication can be accomplished by which of the following?** (5 points)
- SSH
- Secure LINUX
- Kerberos Yes
- IP Chains

B. **What are three attributes of Apache Sqoop?** (Choose three) (5 points)
- Sqoop supports custom connectors for improved performance using certain systems (such as Netezza, Teradata, or Oracle) Yes
- Sqoop queries a source database for schema information Yes
- Sqoop requires ODBC connectivity
- Sqoop can write data to and from Hive tables Yes
- Sqoop ingests data in real-time from log files

C. **What is Hue?** (5 points)
- Hue is a machine learning dashboard for large-scale analytics
- Hue is a web application that allows you to install Hadoop clients on your cluster
- Hue is a web interface that allows you to interact and perform data analysis on your Cloudera cluster Yes
- Hue is a web application that allows you to change client configuration parameters on your cluster in real-time

D. **Which best describes HBase?** (5 points)
- A SQL-like language for processing big data
- A NoSQL database on top of HDFS Yes
- An RDBMS for big data
- An application for ingesting data to HDFS

Part II. Hands on (80 points)

For this part of the assignment you can use the same VM that you have used for first few Hadoop labs in this class. Please include a copy of commands and their step numbers in the PDF file you submit. Please also submit a separate pure-text file that contains all the commands. The latter is for occasional debugging purposes.

In this part, you will import a table from pets_stackexchange database on mysql into HDFS. The dataset is a dump from a stackoverflow site for pets related Q&As: http://pets.stackexchange.com/. You can find a copy of the dump posted on Canvas under the section 'Data'. Please complete the following steps: (80 points)

1. In Hadoop, create a new directory ('*petexchange'*) in your home directory.
```
[training@localhost ~]$ hdfs dfs -mkdir /user/petexchange
```
2. Import the database table posts into Hadoop, and put it under *petexchange*. As an intermediary step, you can first import the dump in MySQL.

First drag and drop petsexchange.out db into root directory in cloudera VM. Then I ran the following to import db into mysql

```
[training@localhost ~]$ mysql --user=training --password=training
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 18
Server version: 5.1.61 Source distribution

Copyright (c) 2000, 2011, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> CREATE DATABASE petexchange;
Query OK, 1 row affected (0.00 sec)

mysql> exit
Bye
[training@localhost ~]$ mysql --user=training --password=training petexchange <p
etsexchange.out
[training@localhost ~]$ mysql --user=training --password=training petexchange
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 20
Server version: 5.1.61 Source distribution

Copyright (c) 2000, 2011, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show tables;
+----------------------+
| Tables_in_petexchange |
+----------------------+
| badges               |
| comments             |
| posthistory          |
| postlinks            |
| posts                |
| tags                 |
| users                |
| votes                |
+----------------------+
8 rows in set (0.00 sec)
```

a. Instead of importing all columns, please skip the body field because this field sometimes contains the line break character (\n), which misleads tools such as Pig to think that it is a new record after the line break.

```
[training@localhost ~]$ sqoop import --connect jdbc:mysql://localhost/petexchange --username training --password training --w
arehouse-dir /petexchange --table posts --columns Id,PostTypeId,ParentId,CreationDate,DeletionDate,Score,ViewCount,OwnerUserI
d,OwnerDisplayName,LastEditorUserId,LastEditorDisplayName,LastEditDate,LastActivityDate,Title,Tags,AnswerCount,CommentCount,F
avoriteCount,ClosedDate,CommunityOwnedDate
```

b. Report the number of rows imported.

```
20/02/13 22:10:32 INFO mapreduce.ImportJobBase: Retrieved 11130 records.
```

3. After ingesting the data, display the content of the *petexchange/posts* folder in HDFS.

```
[training@localhost ~]$ hdfs dfs -ls /petexchange/posts
Found 6 items
-rw-r--r--   1 training supergroup          0 2020-02-13 22:10 /petexchange/posts/_SUCCESS
drwxr-xr-x   - training supergroup          0 2020-02-13 22:09 /petexchange/posts/_logs
-rw-r--r--   1 training supergroup     492410 2020-02-13 22:10 /petexchange/posts/part-m-00000
-rw-r--r--   1 training supergroup     363296 2020-02-13 22:10 /petexchange/posts/part-m-00001
-rw-r--r--   1 training supergroup     421063 2020-02-13 22:10 /petexchange/posts/part-m-00002
-rw-r--r--   1 training supergroup     455139 2020-02-13 22:10 /petexchange/posts/part-m-00003
```

4. Create a local folder named '*petexchange'* in your home directory for holding a sample of the posts data.
   a. This folder should be created in the local filesystem. Not in Hadoop.

   mkdir petexchange

5. Take the first 25 records from *petexchange/posts* and save it as a local file named '*posts'* under the *petexchange* folder you have just created.

```
[training@localhost ~]$ mkdir petexchange
[training@localhost ~]$ hdfs dfs -get /petexchange/posts/part-m-00000 ~/petexchange/posts.txt |head -n 25
```

6. After you take the sample, check if a file posts has been created under the local folder *petexchange*. If yes, view the content of the file to make sure that it is valid.

```
[training@localhost ~]$ cd petexchange
[training@localhost petexchange]$ ls -l
total 484
-rwxr-xr-x 1 training training 492410 Feb 13 22:16 posts.txt
[training@localhost petexchange]$ cat posts.txt
```