Homework #6 **Due: turned in by Wed 11/08/2017 before class**

___Wenbo Liu___

(put your name above)

Hands on (50 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: (45 points)

1. In Hadoop, create a new directory ('*petexchange'*) in your home directory.
   hadoop fs -mkdir 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.
   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.
   b. Report the number of rows imported.
      Procedure:
      i. Login to VMware, download petsexchange.out
      ii. mv petsexchange.out petsexchane.sql
      iii. mysql --user=training --password=training
      iv. CREATE DATABASE petexchange_data;
      v. SOURCE /home/training/Downloads/petsexchange.sql
      vi. SHOW TABLE;
      vii. DESCRIBE posts;
      viii. ALTER TABLE posts
      ix. DROP COLUMN Body;
      x. Quit
      xi. Then we use sqoop to do the transformation
      xii. sqoop import \--connect jdbc:mysql://localhost/petexchange_data \--username training --password training --fields-terminated-by '\t' --table posts --target-dir petexchange/posts

```
17/11/03 10:42:01 INFO mapred.JobClient:     HDFS: Number of read operations=4
17/11/03 10:42:01 INFO mapred.JobClient:     HDFS: Number of large read operatio
ns=0
17/11/03 10:42:01 INFO mapred.JobClient:     HDFS: Number of write operations=4
17/11/03 10:42:01 INFO mapred.JobClient:   Job Counters
17/11/03 10:42:01 INFO mapred.JobClient:     Launched map tasks=4
17/11/03 10:42:01 INFO mapred.JobClient:     Total time spent by all maps in occ
upied slots (ms)=43696
17/11/03 10:42:01 INFO mapred.JobClient:     Total time spent by all reduces in
occupied slots (ms)=0
17/11/03 10:42:01 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
17/11/03 10:42:01 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
17/11/03 10:42:01 INFO mapred.JobClient:   Map-Reduce Framework
17/11/03 10:42:01 INFO mapred.JobClient:     Map input records=11130
17/11/03 10:42:01 INFO mapred.JobClient:     Map output records=11130
17/11/03 10:42:01 INFO mapred.JobClient:     Input split bytes=417
17/11/03 10:42:01 INFO mapred.JobClient:     Spilled Records=0
17/11/03 10:42:01 INFO mapred.JobClient:     CPU time spent (ms)=6070
17/11/03 10:42:01 INFO mapred.JobClient:     Physical memory (bytes) snapshot=39
2916992
17/11/03 10:42:01 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=290
0422656
17/11/03 10:42:01 INFO mapred.JobClient:     Total committed heap usage (bytes)=
63438848
17/11/03 10:42:01 INFO mapreduce.ImportJobBase: Transferred 1.7051 MB in 29.6471
 seconds (58.8938 KB/sec)
17/11/03 10:42:01 INFO mapreduce.ImportJobBase: Retrieved 11130 records.
```

xiii. Therefore there are 11130 rows of data

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

```
[training@localhost Downloads]$ hadoop fs -ls  petexchange/posts
Found 6 items
-rw-r--r--   1 training supergroup          0 2017-11-03 10:55 petexchange/posts
UCCESS
drwxr-xr-x   - training supergroup          0 2017-11-03 10:54 petexchange/posts
ogs
-rw-r--r--   1 training supergroup     507896 2017-11-03 10:54 petexchange/posts
rt-m-00000
-rw-r--r--   1 training supergroup     374768 2017-11-03 10:54 petexchange/posts
rt-m-00001
-rw-r--r--   1 training supergroup     435043 2017-11-03 10:54 petexchange/posts
rt-m-00002
-rw-r--r--   1 training supergroup     470226 2017-11-03 10:54 petexchange/posts
rt-m-00003
[training@localhost Downloads]$
training@localhost:~/...
```

4. Create a local folder named '*petexchange'* in your home directory for holding a sample of the posts data.

First step: making sure we know where our home directory is.

```
[training@localhost ~]$ echo $HOME
/home/training
```

Then make the directory and check the result
mkdir petexchange
ll

```
[training@localhost ~]$ mkdir petexchange
[training@localhost ~]$ ll
total 190292
drwxr-xr-x   2 training training      4096 Jun  6  2014 Desktop
drwxr-xr-x   2 training training      4096 Jun  7  2014 Documents
drwxr-xr-x   2 training training      4096 Nov  3 10:41 Downloads
drwxr-xr-x   9 training training      4096 Feb  4  2013 eclipse
-rw-r--r--.  1 training training 194791349 Dec 10  2013 kiji-bento-albacore-1.0.
elease.tar.gz
drwxr-xr-x.  2 training training      4096 Dec 10  2013 lib
drwxr-xr-x   2 training training      4096 Jun  7  2014 Music
drwxrwxr-x   2 training training      4096 Nov  3 11:06 petexchange
drwxr-xr-x   2 training training      4096 Jun  7  2014 Pictures
drwxr-xr-x   2 training training      4096 Jun  7  2014 Public
drwxr-xr-x.  5 training training      4096 Dec 10  2013 scripts
drwxr-xr-x. 14 training training      4096 May  7  2013 src
drwxr-xr-x   2 training training      4096 Jun  7  2014 Templates
drwxr-xr-x.  6 training training      4096 Dec 10  2013 training_materials
drwxr-xr-x   2 training training      4096 Jun  7  2014 Videos
drwxrwxr-x  23 training training      4096 Oct 25 11:09 workspace
drwxrwxr-x.  4 training training      4096 Dec 11  2012 workspace.save.dev
```

a. This folder should be created in the local filesystem. Not in Hadoop.

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.

   hadoop fs -cat petexchange/posts/part-m-00000 | head -25 >petexchange/posts.txt

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]$ ll
total 8
-rw-rw-r-- 1 training training 4511 Nov  5 17:02 posts.txt
[training@localhost petexchange]$ cat posts.txt
1       1       58      null    2013-10-08 21:29:52.0   null    37      5971
null    null    user9   2013-10-30 19:36:21.0   2013-10-30 19:36:21.0   What
ses a dog to lunge at an unknown child and how should the owner respond?
dogs><behavior><aggression>     2       2       4       null    null
2       1       25      null    2013-10-08 21:40:34.0   null    19      1677
0       null    129     null    2013-10-09 18:18:40.0   2013-10-29 14:27:20.0
ow do I walk a small dog afraid of loud noises in an urban area?        <dogs>
raining><fear><sound>   5       1       null    null    null
3       1       46      null    2013-10-08 21:44:31.0   null    21      5516
3       null    null    user87  2013-11-08 05:17:26.0   2015-04-12 12:53:58.0
hat is required to house break a rabbit?        <rabbits><toilet-training>
0       3       null    null
4       1       null    null    2013-10-08 22:00:01.0   null    6       173
2       null    null    user87  2013-11-08 05:16:34.0   2013-11-08 05:16:34.0
hat is the best way to toilet train a puppy?    <dogs><toilet-training> 2
null    2013-10-13 14:57:17.0   null
5       2       null    3       2013-10-08 22:00:44.0   null    10      null
8       null    null    null    null    2013-10-08 22:00:44.0   null    null
```

In this part, you should start Amazon EMR and then connect to the master node using putty. Please provide a screenshot showing that you successfully managed to connect using putty. (5 points)

a) Hint: Use the instructions posted on Canvas.

b) Do NOT forget to terminate the cluster.

Amazon EMR

- Clusters
- Security configurations
- VPC subnets
- Events
- Help

Clone | Terminate | AWS CLI export

Cluster: ISOM671-Big_Data

Waiting Cluster ready after last step completed.

Summary | Application history | Monitoring | Hardware | Events | Steps | Configurations | Bootstrap actions

Connections: Enable Web Connection – Hue, Ganglia, Resource Manager … (View All)
Master public DNS: ec2-18-216-106-222.us-east-2.compute.amazonaws.com SSH
Tags: -- View All / Edit

Summary

ID: j-W2K6XWKRQA1C
Creation date: 2017-11-05 16:26 (UTC-5)
Elapsed time: 23 minutes
Auto-terminate: No
Termination protection: Off Change

Configuration details

Release label: emr-5.9.0
Hadoop distribution: Amazon 2.7.3
Applications: Ganglia 3.7.2, Hive 2.3.0, Hue 4.0.1, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://aws-logs-238816346879-us-east-2/
EMRFS consistent view: Disabl
Custom AMI ID: --

Network and hardware

Availability zone: us-east-2c
Subnet ID: subnet-0f13b142
Master: Running 1 m4.large
Core: Running 2 m4.large
Task: --

Security and access

Key name: ISOM671-Big_Data
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-62bed50a (ElasticMapReduce-master)
Security groups for Core & Task: sg-f5bcd79d (ElasticMapReduce-slave)