Hadoop and Pig exercise

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'.

Part 1: Hadoop hands on

Questions:

1. In Hadoop, create a new directory ('*petexchange'*) in your home directory.

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.

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

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.

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.

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.

Answers:

1. ~/scripts/developer/training_setup_dev.sh
   hadoop fs -mkdir petexchange

```
[training@localhost home]$ hadoop fs -ls /user/training
Found 6 items
drwxr-xr-x   - training supergroup          0 2017-11-01 09:14 /user/training/movie
drwxr-xr-x   - training supergroup          0 2017-11-01 09:25 /user/training/movierating
drwxr-xr-x   - training supergroup          0 2017-11-03 10:45 /user/training/petsexchange
drwxr-xr-x   - training supergroup          0 2017-10-25 11:14 /user/training/shakespeare
drwxr-xr-x   - training supergroup          0 2017-10-25 11:20 /user/training/weblog
drwxr-xr-x   - training supergroup          0 2017-11-01 08:26 /user/training/wordcounts
```

2. mysql --user=training --password=training
   mysql > CREATE DATABASE petsexchange

mysql --user=training --password=training petsexchange <
"/home/training/Downloads/petsexchange.out"

ALTER TABLE posts DROP COLUMN body;

sqoop import \
--connect jdbc:mysql://localhost/petsexchange \
--username training --password training \
--fields-terminated-by '\t' \
--warehouse-dir petexchange \
--table posts

```
mysql> ALTER TABLE posts DROP COLUMN body;
Query OK, 11130 rows affected (0.69 sec)
Records: 11130  Duplicates: 0  Warnings: 0
INFO mapreduce.ImportJobBase: Transferred 1.7051 MB in 23.2881 seconds (74.9753 KB/sec)
INFO mapreduce.ImportJobBase: Retrieved 11130 records.
```

11130 rows imported.

3. hadoop fs -ls /petexchange/posts

```
[training@localhost ~]$ hadoop fs -ls /petsexchange/posts
Found 6 items
-rw-r--r--   1 training supergroup          0 2017-11-03 10:54 /petsexchange/posts/_SUCCESS
drwxr-xr-x   - training supergroup          0 2017-11-03 10:54 /petsexchange/posts/_logs
-rw-r--r--   1 training supergroup     507896 2017-11-03 10:54 /petsexchange/posts/part-m-00000
-rw-r--r--   1 training supergroup     374768 2017-11-03 10:54 /petsexchange/posts/part-m-00001
-rw-r--r--   1 training supergroup     435043 2017-11-03 10:54 /petsexchange/posts/part-m-00002
-rw-r--r--   1 training supergroup     470226 2017-11-03 10:54 /petsexchange/posts/part-m-00003
```

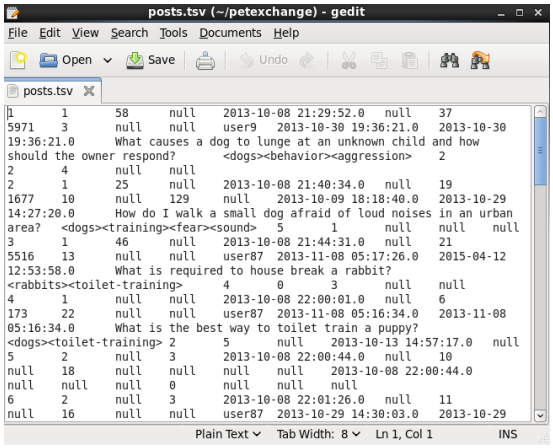
4. cd /home
   mkdir petexchange

```
[training@localhost ~]$ ls -l
total 190316
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 09:40 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-1.0.5-release.tar.gz
drwxr-xr-x.  2 training training      4096 Dec 10  2013 lib
-rw-rw-r--   1 training training      8627 Nov  1 09:13 movie.java
-rw-rw-r--   1 training training      9011 Nov  1 09:25 movierating.java
drwxr-xr-x   2 training training      4096 Jun  7  2014 Music
drwxrwxr-x   2 training training      4096 Nov  3 10:15 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
```

5. hadoop fs -cat $home/petexchange/posts/part-m-00000 | head -25 > petexchange/posts.tsv

6. cd ~/petexchange
   ls -l
   gedit posts.tsv &

```
[training@localhost ~]$ cd ~/petexchange
[training@localhost petexchange]$ ls -l
total 8
-rw-rw-r-- 1 training training 4511 Nov  5 15:57 posts.tsv
[training@localhost petexchange]$ gedit posts.tsv &
[2] 15539
```



Part 2: Pig hands on

Questions:

1. Create a pig script called “summarize_posts.pig” to do the following:
   a. Load the posts data, choosing appropriate data types wherever necessary for the next steps.
   b. Filter data so only posts with postypeid=1 remain (these are the original posts)
   c. Re-order the fields keeping only the following fields: id, creationdate, title, tags, score, and viewcount.
   d. Calculate the total number of posts and total (i.e., sum) viewcount.
   e. Print on screen (or write on a file) the information you calculated in the previous step.

Answers:

```
posts = LOAD'/petsexchange/posts' AS
(Id:int,
PostTypeId:int,
AcceptedAnswerId:int,
ParentID:int,
CreationDate:datetime,
DeletionDate:datetime,
Score:int,
Viewcount:int,
OwnerUserId:int,
OwnerDisplayName:chararray,
LastEditorUserId:int,
LastEditorDisplayName:chararray,
LastEditDate:datetime,
LastActivityDate:datetime,
Title:chararray,
Tags:chararray,
AnswerCount:int,
CommentCount:int,
FavoriteCount:int,
ClosedDate:datetime,
CommunityOwnedDate:datetime);

origposts = FILTER posts BY PostTypeId == 1;
reorderfields = FOREACH origposts GENERATE Id, CreationDate, Title, Tags, Score, Viewcount;
grouped = GROUP reorderfields ALL;
summarize = FOREACH grouped GENERATE COUNT(reorderfields.Id),SUM(reorderfields.Viewcount);
DUMP summarize;
```

Local:

```
[training@localhost petexchange]$ pig -x local summarize_posts.pig
2017-11-11 17:16:28,895 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2017-11-11 17:16:28,895 INFO org.apache.pig.Main: Logging error messages to: /home/training/petexchange/pig_1510438588894.log
(16,71350)
```

HDFS:

```
[training@localhost petexchange]$ pig summarize_posts.pig
2017-11-11 18:01:33,793 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2017-11-11 18:01:33,793 INFO org.apache.pig.Main: Logging error messages to: /home/training/petexchange/pig_1510441293789.log
(4123,11507808)
```