EMORY UNIVERSITY

Managing Big Data

Homework #5

Due: turned in by Mon 2/12/2020 before class

Carl Xi

(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**
- 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)**
- **Sqoop queries a source database for schema information**
- Sqoop requires ODBC connectivity
- **Sqoop can write data to and from Hive tables**
- 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**
- 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**
- An RDBMS for big data
- An application for ingesting data to HDFS

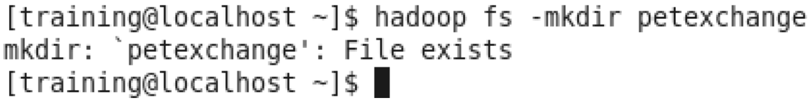
Part II. Hands on (80 points)

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.

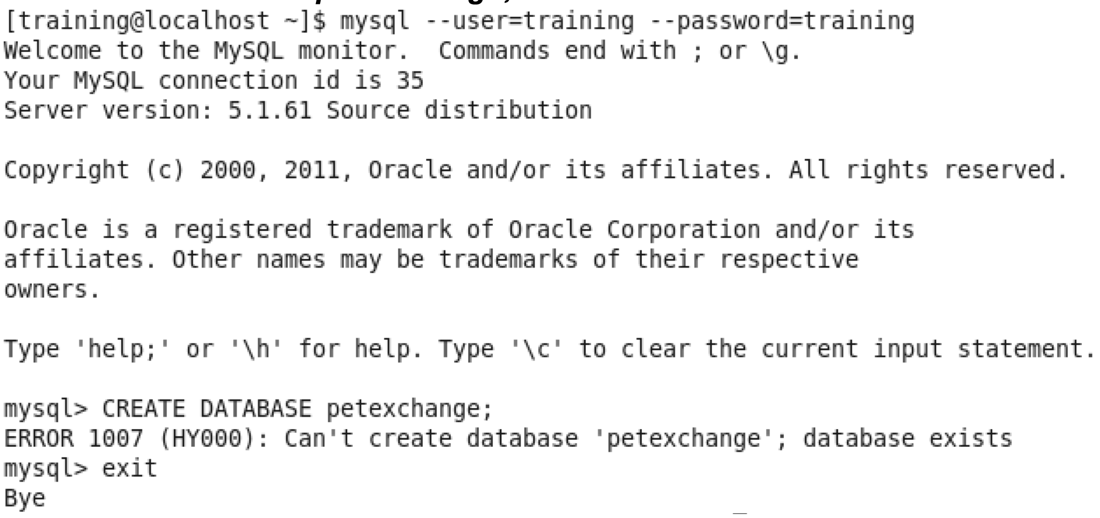
hadoop fs -mkdir petexchange



```
[training@localhost ~]$ hadoop fs -mkdir petexchange
mkdir: `petexchange': File exists
[training@localhost ~]$ █
```

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.

CREATE DATABASE petexchange;



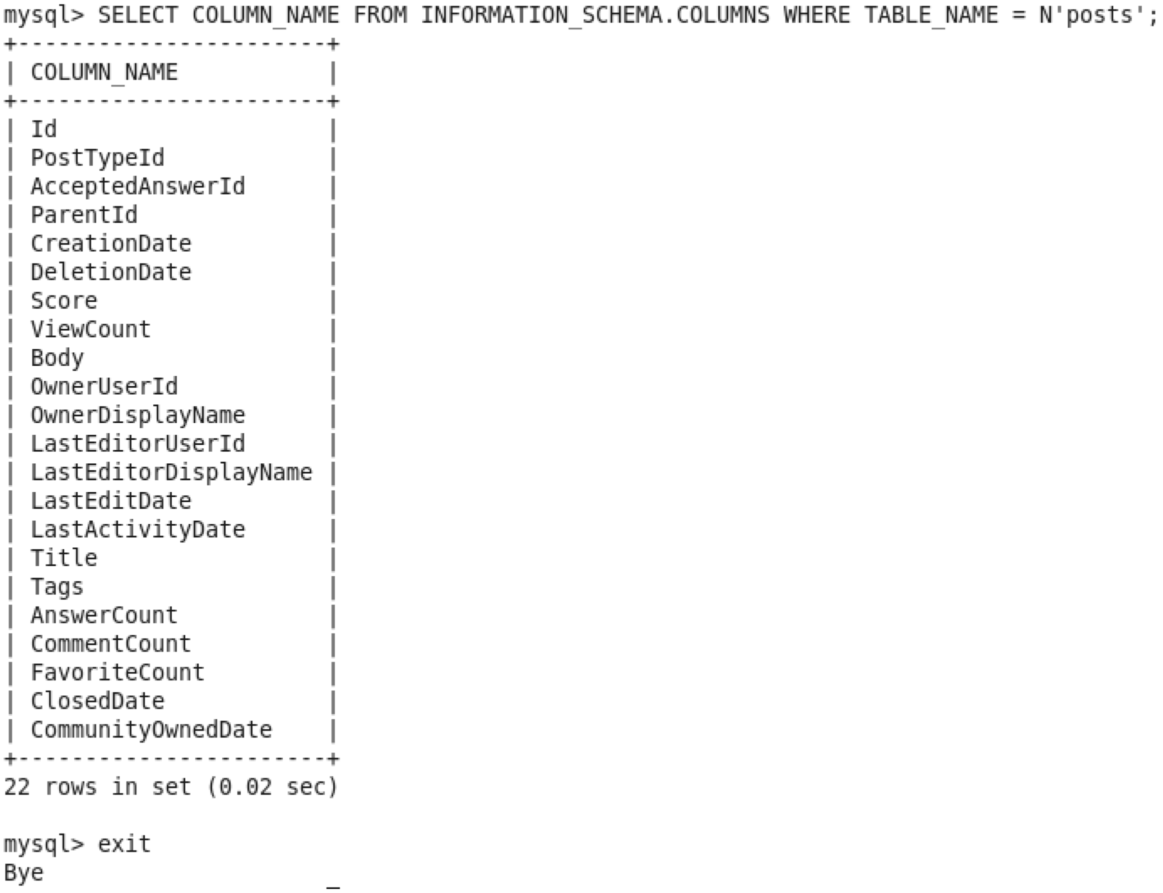
```
[training@localhost ~]$ mysql --user=training --password=training
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 35
Server version: 5.1.61 Source distribution

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;
ERROR 1007 (HY000): Can't create database 'petexchange'; database exists
mysql> exit
Bye
```



```
mysql> SELECT COLUMN_NAME FROM INFORMATION_SCHEMA.COLUMNS WHERE TABLE_NAME = N'posts';
+-----------------------+
| COLUMN_NAME           |
+-----------------------+
| Id                    |
| PostTypeId            |
| AcceptedAnswerId      |
| ParentId              |
| CreationDate          |
| DeletionDate          |
| Score                 |
| ViewCount             |
| Body                  |
| OwnerUserId           |
| OwnerDisplayName      |
| LastEditorUserId      |
| LastEditorDisplayName |
| LastEditDate          |
| LastActivityDate      |
| Title                 |
| Tags                  |
| AnswerCount           |
| CommentCount          |
| FavoriteCount         |
| ClosedDate            |
| CommunityOwnedDate    |
+-----------------------+
22 rows in set (0.02 sec)

mysql> exit
Bye
```

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.

```
mysql --user=training --password=training petexchange < petsexchange.out
sqoop import \
--connect jdbc:mysql://localhost/petexchange \
--username training --password training \
--fields-terminated-by '\t' \
--warehouse-dir /petexchange \
--table posts \
--columns "Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score,
ViewCount, OwnerUserId, OwnerDisplayName, LastEditorUserId, LastEditorDisplayName,
LastEditDate, LastActivityDate, Title, Tags, AnswerCount, CommentCount, FavoriteCount,
ClosedDate, CommunityOwnedDate"
```

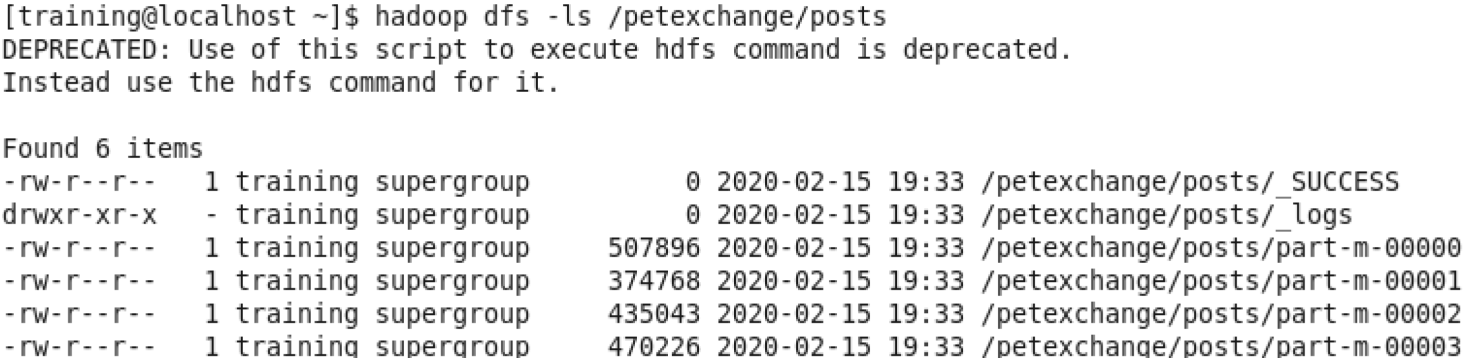
```
Note: Recompile with -Xlint:deprecation for details.
20/02/15 19:33:33 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-training/comp
ile/3340aa19cff784d684ded923d816e27b/posts.jar
20/02/15 19:33:33 WARN manager.MySQLManager: It looks like you are importing from mysql.
20/02/15 19:33:33 WARN manager.MySQLManager: This transfer can be faster! Use the --direc
t
20/02/15 19:33:33 WARN manager.MySQLManager: option to exercise a MySQL-specific fast pat
h.
20/02/15 19:33:33 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToN
ull (mysql)
20/02/15 19:33:33 INFO mapreduce.ImportJobBase: Beginning import of posts
20/02/15 19:33:34 WARN mapred.JobClient: Use GenericOptionsParser for parsing the argumen
ts. Applications should implement Tool for the same.
20/02/15 19:33:35 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`Id`), M
AX(`Id`) FROM `posts`
20/02/15 19:33:35 INFO mapred.JobClient: Running job: job_202002121253_0001
20/02/15 19:33:36 INFO mapred.JobClient:  map 0% reduce 0%
20/02/15 19:33:48 INFO mapred.JobClient:  map 50% reduce 0%
20/02/15 19:33:55 INFO mapred.JobClient:  map 100% reduce 0%
20/02/15 19:33:56 INFO mapred.JobClient: Job complete: job_202002121253_0001
20/02/15 19:33:56 INFO mapred.JobClient: Counters: 23
20/02/15 19:33:56 INFO mapred.JobClient:   File System Counters
20/02/15 19:33:56 INFO mapred.JobClient:     FILE: Number of bytes read=0
20/02/15 19:33:56 INFO mapred.JobClient:     FILE: Number of bytes written=839908
20/02/15 19:33:56 INFO mapred.JobClient:     FILE: Number of read operations=0
20/02/15 19:33:56 INFO mapred.JobClient:     FILE: Number of large read operations=0
20/02/15 19:33:56 INFO mapred.JobClient:     FILE: Number of write operations=0
20/02/15 19:33:56 INFO mapred.JobClient:     HDFS: Number of bytes read=417
20/02/15 19:33:56 INFO mapred.JobClient:     HDFS: Number of bytes written=1787933
20/02/15 19:33:56 INFO mapred.JobClient:     HDFS: Number of read operations=4
20/02/15 19:33:56 INFO mapred.JobClient:     HDFS: Number of large read operations=0
20/02/15 19:33:56 INFO mapred.JobClient:     HDFS: Number of write operations=4
20/02/15 19:33:56 INFO mapred.JobClient:   Job Counters
20/02/15 19:33:56 INFO mapred.JobClient:     Launched map tasks=4
20/02/15 19:33:56 INFO mapred.JobClient:     Total time spent by all maps in occupied slo
ts (ms)=32511
20/02/15 19:33:56 INFO mapred.JobClient:     Total time spent by all reduces in occupied
slots (ms)=0
20/02/15 19:33:56 INFO mapred.JobClient:     Total time spent by all maps waiting after r
eserving slots (ms)=0
20/02/15 19:33:56 INFO mapred.JobClient:     Total time spent by all reduces waiting afte
r reserving slots (ms)=0
20/02/15 19:33:56 INFO mapred.JobClient:   Map-Reduce Framework
20/02/15 19:33:56 INFO mapred.JobClient:     Map input records=11130
20/02/15 19:33:56 INFO mapred.JobClient:     Map output records=11130
20/02/15 19:33:56 INFO mapred.JobClient:     Input split bytes=417
20/02/15 19:33:56 INFO mapred.JobClient:     Spilled Records=0
20/02/15 19:33:56 INFO mapred.JobClient:     CPU time spent (ms)=4210
20/02/15 19:33:56 INFO mapred.JobClient:     Physical memory (bytes) snapshot=386760704
20/02/15 19:33:56 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=2908643328
20/02/15 19:33:56 INFO mapred.JobClient:     Total committed heap usage (bytes)=63438848
20/02/15 19:33:56 INFO mapreduce.ImportJobBase: Transferred 1.7051 MB in 23.3478 seconds
(74.7835 KB/sec)
20/02/15 19:33:56 INFO mapreduce.ImportJobBase: Retrieved 11130 records.
```

b. Report the number of rows imported.

11130 records were imported.

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

hadoop dfs -ls /petexchange/posts

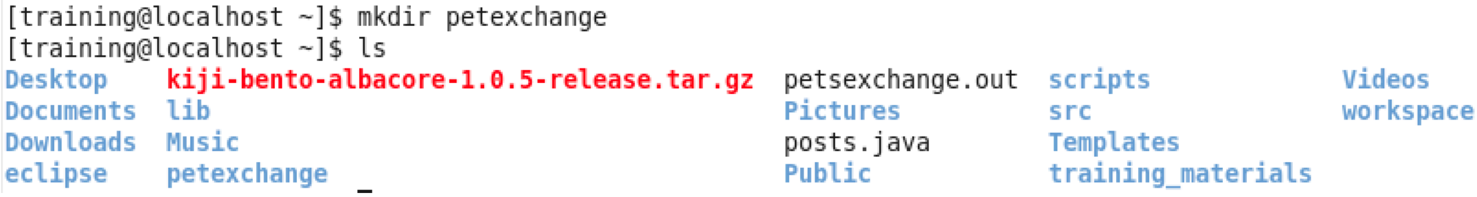


```
[training@localhost ~]$ hadoop dfs -ls /petexchange/posts
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 6 items
-rw-r--r--   1 training supergroup          0 2020-02-15 19:33 /petexchange/posts/_SUCCESS
drwxr-xr-x   - training supergroup          0 2020-02-15 19:33 /petexchange/posts/_logs
-rw-r--r--   1 training supergroup     507896 2020-02-15 19:33 /petexchange/posts/part-m-00000
-rw-r--r--   1 training supergroup     374768 2020-02-15 19:33 /petexchange/posts/part-m-00001
-rw-r--r--   1 training supergroup     435043 2020-02-15 19:33 /petexchange/posts/part-m-00002
-rw-r--r--   1 training supergroup     470226 2020-02-15 19:33 /petexchange/posts/part-m-00003
```

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

mkdir petexchange

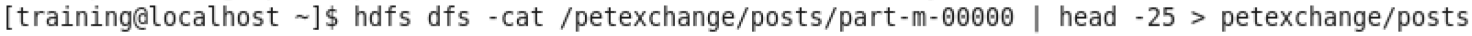


```
[training@localhost ~]$ mkdir petexchange
[training@localhost ~]$ ls
Desktop    kiji-bento-albacore-1.0.5-release.tar.gz  petsexchange.out  scripts             Videos
Documents  lib                                       Pictures          src                 workspace
Downloads  Music                                     posts.java        Templates
eclipse    petexchange                               Public            training_materials
```

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.

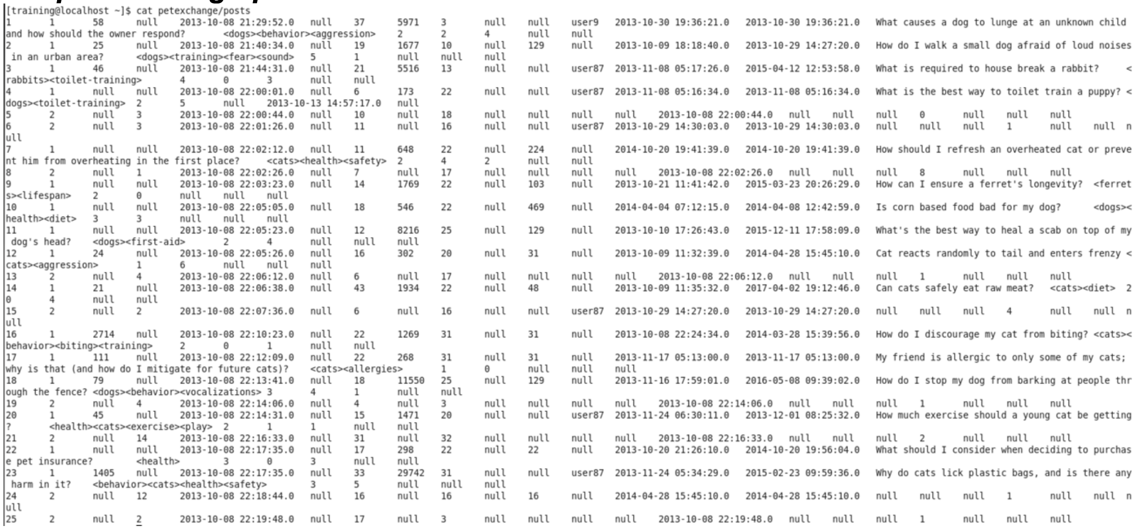
hdfs dfs -cat /petexchange/posts/part-m-00000 | head -25 > petexchange/posts



```
[training@localhost ~]$ hdfs dfs -cat /petexchange/posts/part-m-00000 | head -25 > petexchange/posts
```

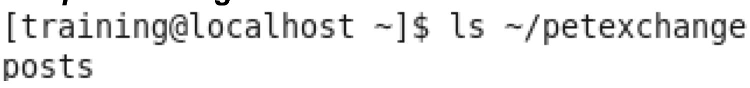
6. After you take the sample, check if a file posts has been created under the local folder *petexchange*. If yes, view the content of the file to make sure that it is valid.

cat petexchange/posts



```
[training@localhost ~]$ cat petexchange/posts
1       1       58      null    2013-10-08 21:29:52.0   null    37      5971    3       null    null    user9   2013-10-30 19:36:21.0   2013-10-30 19:36:21.0   What causes 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    10      null    129     null    2013-10-09 18:18:40.0   2013-10-29 14:27:20.0   How do I walk a small dog afraid of loud noises
 in an urban area?      <dogs><training><fear><sound>   5       1       null    null    null
3       1       46      null    2013-10-08 21:44:31.0   null    21      5516    13      null    null    user87  2013-11-08 05:17:26.0   2015-04-12 12:53:58.0   What is required to house break a rabbit?       <
rabbits><toilet-training>       4       0       3       null    null
4       1       null    null    2013-10-08 22:00:01.0   null    6       173     22      null    null    user87  2013-11-08 05:16:34.0   2013-11-08 05:16:34.0   What is the best way to toilet train a puppy? <
dogs><toilet-training>  2       5       null    2013-10-13 14:57:17.0   null
5       2       null    3       2013-10-08 22:00:44.0   null    10      null    18      null    null    null    null    2013-10-08 22:00:44.0   null    null    null    0       null    null    null
6       2       null    3       2013-10-08 22:01:26.0   null    11      null    16      null    null    user87  2013-10-29 14:30:03.0   2013-10-29 14:30:03.0   null    null    null    1       null    null  n
ull
7       1       null    null    2013-10-08 22:02:12.0   null    11      648     22      null    224     null    2014-10-20 19:41:39.0   2014-10-20 19:41:39.0   How should I refresh an overheated cat or preve
nt him from overheating in the first place?     <cats><health><safety>  2       4       2       null    null
8       2       null    1       2013-10-08 22:02:26.0   null    7       null    17      null    null    null    null    2013-10-08 22:02:26.0   null    null    null    8       null    null    null
9       1       null    null    2013-10-08 22:03:23.0   null    14      1769    22      null    103     null    2013-10-21 11:41:42.0   2015-03-23 20:26:29.0   How can I ensure a ferret's longevity?  <ferret
s><lifespan>    2       0       null    null    null
10      1       null    null    2013-10-08 22:05:05.0   null    18      546     22      null    469     null    2014-04-04 07:12:15.0   2014-04-08 12:42:59.0   Is corn based food bad for my dog?      <dogs><
health><diet>   3       3       null    null    null
11      1       null    null    2013-10-08 22:05:23.0   null    12      8216    25      null    129     null    2013-10-10 17:26:43.0   2015-12-11 17:58:09.0   What's the best way to heal a scab on top of my
 dog's head?    <dogs><first-aid>       2       4       null    null    null
12      1       24      null    2013-10-08 22:05:26.0   null    16      302     20      null    31      null    2013-10-09 11:32:39.0   2014-04-28 15:45:10.0   Cat reacts randomly to tail and enters frenzy <
cats><aggression>       1       6       null    null    null
13      2       null    4       2013-10-08 22:06:12.0   null    6       null    17      null    null    null    null    2013-10-08 22:06:12.0   null    null    null    1       null    null    null
14      1       21      null    2013-10-08 22:06:38.0   null    43      1934    22      null    48      null    2013-10-09 11:35:32.0   2017-04-02 19:12:46.0   Can cats safely eat raw meat?   <cats><diet>  2
0       4       null    null
15      2       null    2       2013-10-08 22:07:36.0   null    6       null    16      null    null    user87  2013-10-29 14:27:20.0   2013-10-29 14:27:20.0   null    null    null    4       null    null  n
ull
16      1       2714    null    2013-10-08 22:10:23.0   null    22      1269    31      null    31      null    2013-10-08 22:24:34.0   2014-03-28 15:39:56.0   How do I discourage my cat from biting? <cats><
behavior><biting><training>     2       0       1       null    null
17      1       111     null    2013-10-08 22:12:09.0   null    22      268     31      null    31      null    2013-11-17 05:13:00.0   2013-11-17 05:13:00.0   My friend is allergic to only some of my cats;
why is that (and how do I mitigate for future cats)?    <cats><allergies>       1       0       null    null    null
18      1       79      null    2013-10-08 22:13:41.0   null    18      11550   25      null    129     null    2013-11-16 17:59:01.0   2016-05-08 09:39:02.0   How do I stop my dog from barking at people thr
ough the fence? <dogs><behavior><vocalizations> 3       4       1       null    null
19      2       null    4       2013-10-08 22:14:06.0   null    4       null    3       null    null    null    null    2013-10-08 22:14:06.0   null    null    null    1       null    null    null
20      1       45      null    2013-10-08 22:14:31.0   null    15      1471    20      null    null    user87  2013-11-24 06:30:11.0   2013-12-01 08:25:32.0   How much exercise should a young cat be getting
?       <health><cats><exercise><play>  2       1       1       null    null
21      2       null    14      2013-10-08 22:16:33.0   null    31      null    32      null    null    null    null    2013-10-08 22:16:33.0   null    null    null    2       null    null    null
22      1       null    null    2013-10-08 22:17:35.0   null    17      298     22      null    22      null    2013-10-20 21:26:10.0   2014-10-20 19:56:04.0   What should I consider when deciding to purchas
e pet insurance?        <health>        3       0       3       null    null
23      1       1405    null    2013-10-08 22:17:35.0   null    33      29742   31      null    null    user87  2013-11-24 05:34:29.0   2015-02-23 09:59:36.0   Why do cats lick plastic bags, and is there any
 harm in it?    <behavior><cats><health><safety>        3       5       null    null    null
24      2       null    12      2013-10-08 22:18:44.0   null    16      null    16      null    16      null    2014-04-28 15:45:10.0   2014-04-28 15:45:10.0   null    null    null    1       null    null  n
ull
25      2       null    2       2013-10-08 22:19:48.0   null    17      null    3       null    null    null    null    2013-10-08 22:19:48.0   null    null    null    1       null    null    null
```

ls ~/petexchange



```
[training@localhost ~]$ ls ~/petexchange
posts
```