|  |  |
| --- | --- |
|  | **Managing Big Data** |

|  |  |
| --- | --- |
| **Homework #7** | **Due: turned in by Mon 11/13/2017 before class** |

\_Xi (Athena) Li\_

(put your name above)

Total grade: \_\_\_\_\_\_\_ out of \_\_\_100\_\_\_ points

# **Part I: Short Answers (30 points)**

The answers will be graded along the lines of validity, informativeness, and presentation style. Be sure to include sources if you use any.

### Pig and SQL (10 points)

Name at least two important differences between Pig and SQL.

* *Pig Latin is a data flow programming language, whereas SQL is a declarative programming language.*
* *Pig supports complex, nested data structures. SQL operates on flatter data structures.*

### Pig's Strengths as Ad-hoc Query Tool (10 points)

One of Pig's most compelling attributes is its ability to conduct ad hoc queries. Name at least two reasons that makes pig a flexible tool for conducting ad hoc queries.

* *PIG is used to build complex jobs behind the scenes to spread the load across many servers and process massive quantities of data in an endlessly scalable parallel environment.*
* *PIG is a high level data flow language which creates step-by-step procedures on raw data to derive valuable insights. It offers major advantages in efficiency and flexibility to access different kinds of data.*

### Pig's Local Mode (10 points)

What does it mean to run Pig in “Local Mode”? What is the purpose of the “Local Mode”?

*After we created a pig script, since pig is slow as an interpreter, we may want to test it through a subset of the original dataset. If our pig script can run successfully using the small local file, we can then run the script in the HDFS file.*

B. (2012, September 21). PIG and Big Data – Processing Massive Data Volumes at High Speed. Retrieved November 12, 2017, from http://www.bodhtree.com/blog/2012/09/18/pig-and-big-data-processing-massive-data-volumes-at-high-speed/

# **Part II. Hands on (50 points)**

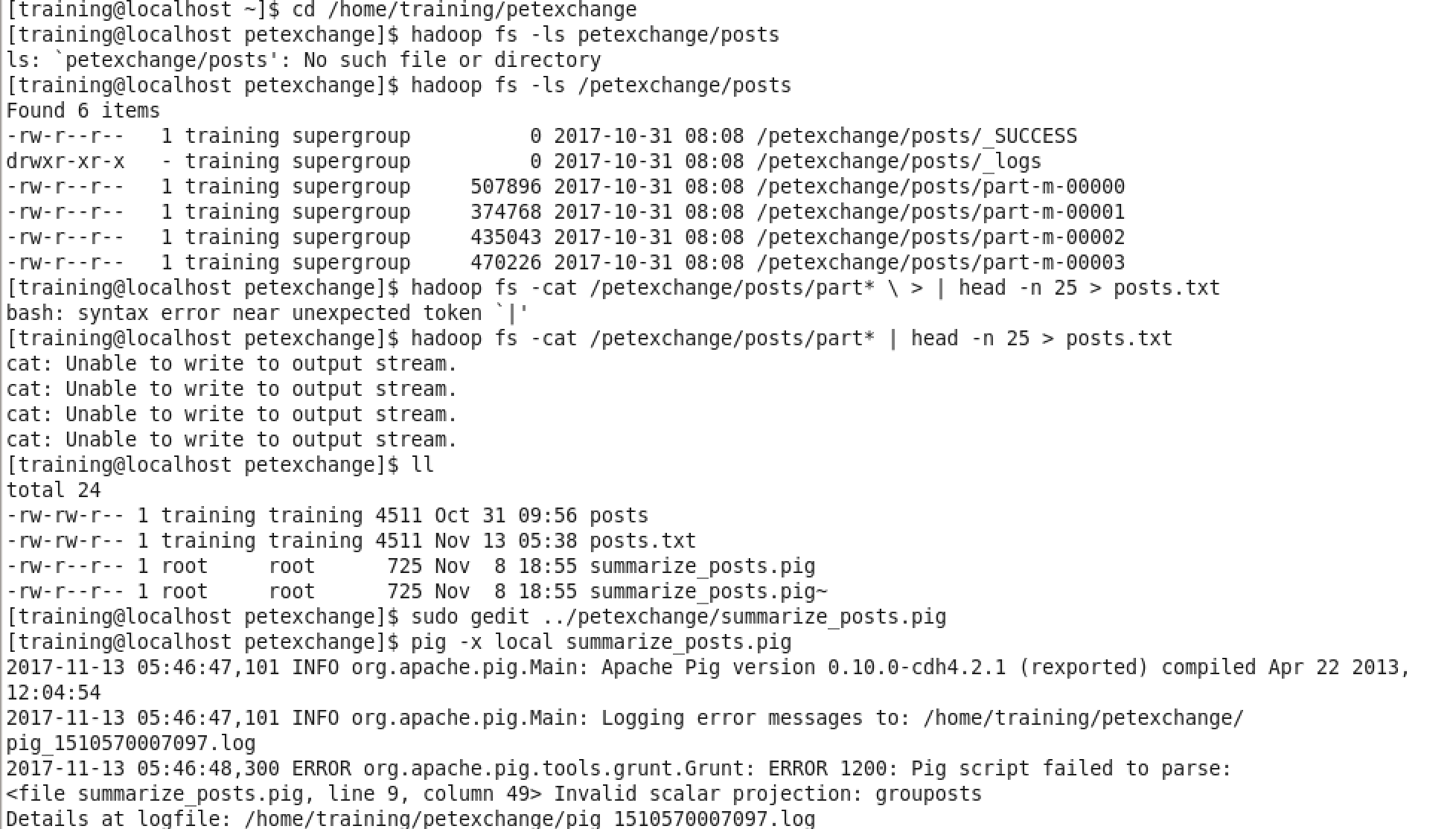
For this part of the assignment you can use the VM that you have used for first few Hadoop labs in this class. Please include a copy of all commands and their step numbers in the PDF file you submit. Please also submit a separate pure-text file that contains all the commands. The latter is for occasional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

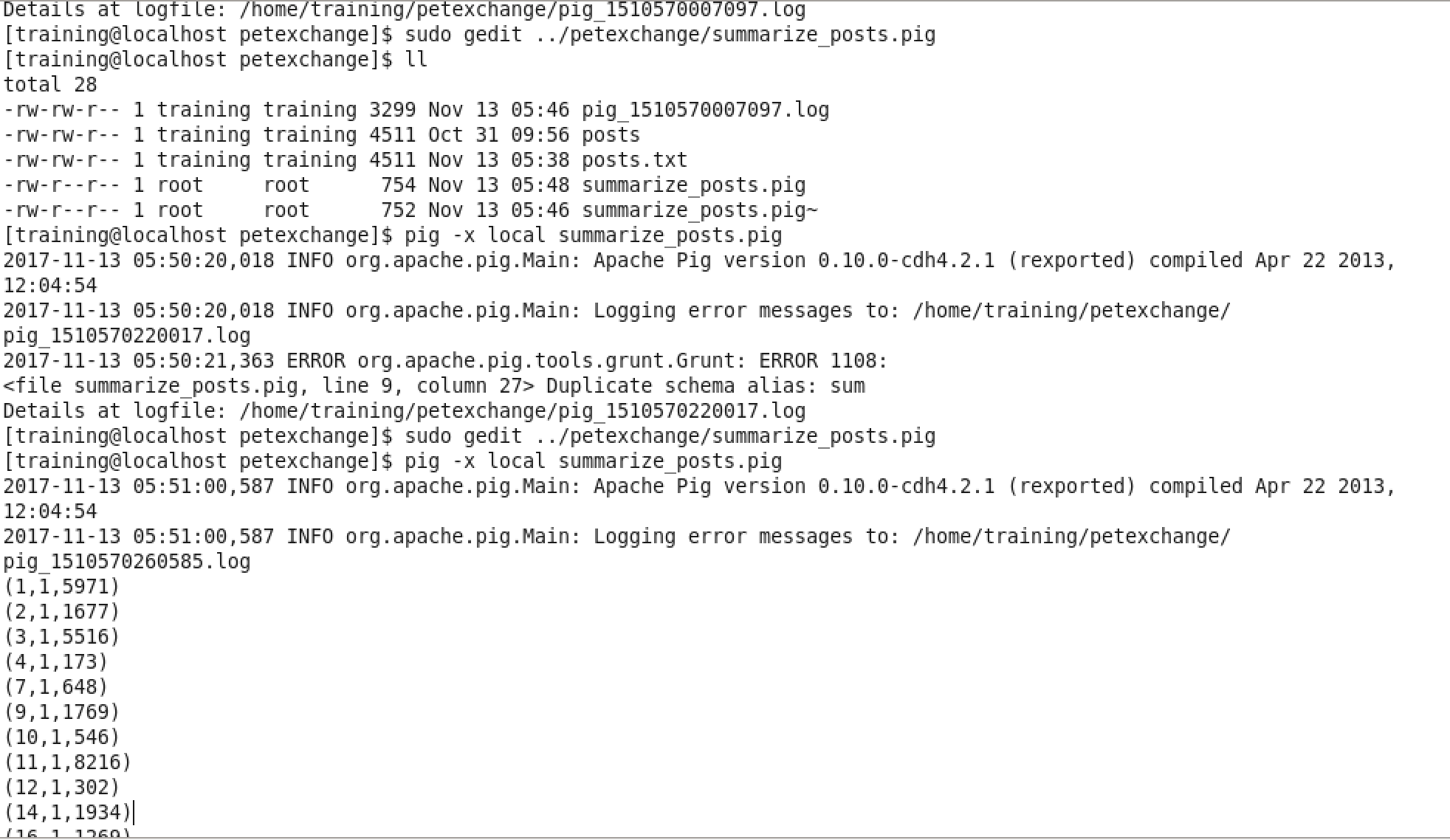
In this part of this exercise, you will use the data you imported into HDFS from the pets\_stackexchange database in part II of assignment 6. In particular, in the previous assignment you were asked to complete the following steps:

1. In Hadoop, create a new directory “**/petexchange**”.
2. Import the database table “posts” into Hadoop, and put it under “**/petexchange**”.
3. Create a local directory **petexchange** under the home (local) directory.
4. Take the first 25 records from posts data file and put it under the local petexchange folder you have just created.

**As part of assignment 7, you have to complete the following steps: (40 points).**

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





**In this part of the assignment, you should complete the following labs that are posted on Canvas: (10 points)**

1. “Lab: Analyzing Disparate Data Sets with Pig”
2. “Lab: Using Pig for ETL Processing”
3. “Lab: Analyzing Ad Campaign Data with Pig”

You should provide a screenshot for each lab illustrating the successful completion of the last step of each lab. Please feel free to skip the “bonus lab” parts.

*Hint*: You can use any of the material that’s already uploaded on Canvas in order to complete these labs.

