EMORY UNIVERSITY

Managing Big Data

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

__Wenbo Liu____

(put your name above)

Total grade: ________ out of ___100___ points

General Submission Guidelines

In this and future assignments, there are typically two types of problems: short answers and hands-on exercises. For short answers, if you use others' work as part of your answer, please properly cite your source. If the source involves a URL, the URL should be provided. Please refer to the following example for the bibliography style:

> This phenomenon has been mentioned in several sources include a web page (Kehoe 1992) and a journal paper (Yeh 1996). A recent newspaper article (Greiner 2011) provides further details about this phenomenon.
>
> - Kehoe, Brendan P. "Zen and the Art of the Internet." January 1992, http://freenet.buffalo.edu/~popmusic/zen10.txt
> - Yeh, Michelle. "The 'Cult of Poetry' in Contemporary China." Journal of Asian Studies 55 (1996): 51-80.
> - Greiner, Lynn. "Wrists on fire? Tech gear for what ails you." Globe and Mail (Toronto) January 27, 2011. http://www.theglobeandmail.com/

The answers for homework assignments should be submitted in a PDF file. When the homework involves script files (e.g. pig, scala, 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: 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.

1. Pig and SQL (10 points)

Name at least two important differences between Pig and SQL.

- Different languages: Pig uses Pig Latin, which is different from standard SQL.
- Different use case: SQL typically runs on top of RDBMS, while Pig runs on top of HDFS

2. 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 procedural, therefore it’s straightforward and you have the control over every step.
- Pig is rather flexible when dealing with unstructured data compared to other tools
- Lazy-evaluation: before you produce any results, Pig remembers your procedures but don’t execute them until you want output. This can be more efficient in many cases.
- Enjoys all the benefit of Hadoop – parallelization, fault tolerance, and scalability.

3. Pig's Local Mode (10 points)

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

Running Pig in Local mode means Pig looks for location/path on your local directories instead of on HDFS and run simulated MapReduce job. Sometimes you want to test your code in a smaller subset locally before you run it in your HDFS. Therefore, you can copy a smaller subset of your file from HDFS to your local machine and test your code using PIG local mode.

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.

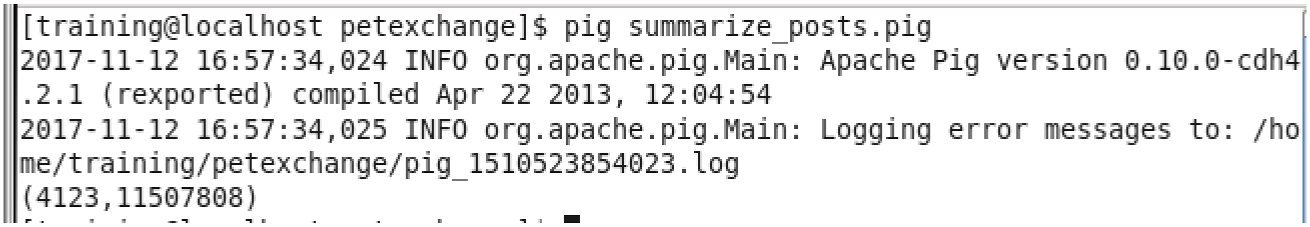
As part of assignment 7, you have to complete the following steps: (40 points).

1. Create a pig script called "summarize_posts.pig" to do the following:
   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.

```
data = LOAD '/user/training/posts' AS (Id:int,
PostTypeId:int,
AcceptedAnswerId:int,
ParentId:int,
CreationDate:chararray,
DeletionDate:chararray,
Score:int,
ViewCount:int,
OwnerUserId:int,
LastEditorUserId:int,
LastEditorDisplayName:chararray,
LastEditDate:chararray,
LastActivityDate:chararray,
Title:chararray,
Tags:chararray
);

fp = FILTER data by PostTypeId == 1;
data_sub = FOREACH fp GENERATE Id,CreationDate,Title,Tags,Score,ViewCount;
data_grp = GROUP data_sub ALL;
result = FOREACH data_grp GENERATE COUNT(data_sub) AS count,SUM(data_sub.ViewCount) AS sum;
```

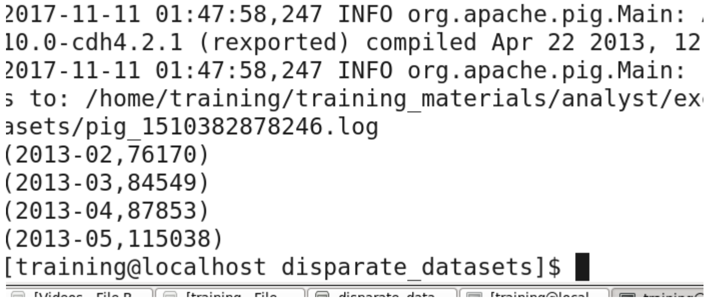
DUMP result;



```
[training@localhost petexchange]$ pig summarize_posts.pig
2017-11-12 16:57:34,024 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2017-11-12 16:57:34,025 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/petexchange/pig_1510523854023.log
(4123,11507808)
```

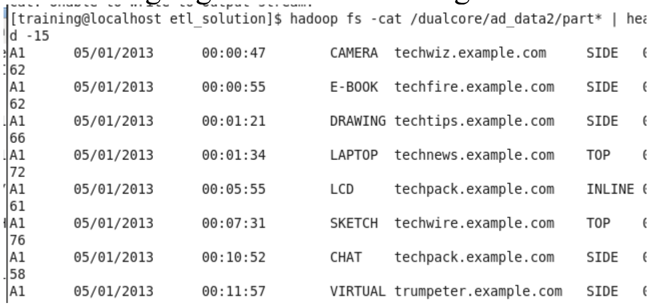
In this part of the assignment, you should complete the following labs that are posted on Canvas: (10 points)

1. “Lab: Analyzing Disparate Data Sets with Pig”



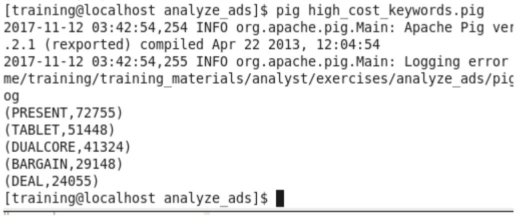
```
2017-11-11 01:47:58,247 INFO org.apache.pig.Main:
10.0-cdh4.2.1 (rexported) compiled Apr 22 2013, 12
2017-11-11 01:47:58,247 INFO org.apache.pig.Main:
s to: /home/training/training_materials/analyst/ex
asets/pig_1510382878246.log
(2013-02,76170)
(2013-03,84549)
(2013-04,87853)
(2013-05,115038)
[training@localhost disparate_datasets]$ 
```

2. “Lab: Using Pig for ETL Processing”



```
[training@localhost etl_solution]$ hadoop fs -cat /dualcore/ad_data2/part* | hea
d -15
A1      05/01/2013      00:00:47        CAMERA  techwiz.example.com     SIDE
62
A1      05/01/2013      00:00:55        E-BOOK  techfire.example.com    SIDE
62
A1      05/01/2013      00:01:21        DRAWING techtips.example.com    SIDE
66
A1      05/01/2013      00:01:34        LAPTOP  technews.example.com    TOP
72
A1      05/01/2013      00:05:55        LCD     techpack.example.com    INLINE
61
A1      05/01/2013      00:07:31        SKETCH  techwire.example.com    TOP
76
A1      05/01/2013      00:10:52        CHAT    techpack.example.com    SIDE
58
A1      05/01/2013      00:11:57        VIRTUAL trumpeter.example.com   SIDE
```

3. “Lab: Analyzing Ad Campaign Data with Pig”



```
[training@localhost analyze_ads]$ pig high_cost_keywords.pig
2017-11-12 03:42:54,254 INFO org.apache.pig.Main: Apache Pig ver
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2017-11-12 03:42:54,255 INFO org.apache.pig.Main: Logging error
me/training/training_materials/analyst/exercises/analyze_ads/pig
og
(PRESENT,72755)
(TABLET,51448)
(DUALCORE,41324)
(BARGAIN,29148)
(DEAL,24055)
[training@localhost analyze_ads]$ 
```

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.