**Lab: Splitting a Dataset**

**About this Lab**

**Objective:**

Research the White House visitor data and look for members of Congress.

**File locations:**

n/a

**Successful**

**outcome:**

Two folders in HDFS, congress and not\_congress, containing a split of the White House visitor data.

**Before you begin:**

You should have the White House visitor data in HDFS in /user/root/whitehouse/visits.txt.

**Related lesson:**

***Advanced Pig Programming***

**Lab Steps**

1 ) Explore the Comments Field

1. If not already done, open a Terminal in your VM and type "ssh sandbox".
2. In this step, you will explore the comments field of the White House visitor data. From the Pig Grunt shell, start by loading visits.txt:



# pig

grunt> cd whitehouse

grunt> visits = LOAD 'visits.txt' USING PigStorage(',');

1. Field $25 is the comments. Filter out all records where field $25 is null: grunt> not\_null\_25 = FILTER visits BY ($25 IS NOT NULL);
2. Now define a new relation that is a projection of only column $25: grunt> comments = FOREACH not\_null\_25 GENERATE $25 AS comment;
3. View the schema of comments and make sure you understand how this relation ended up as a tuple with one field:



grunt> describe comments;

comments: {comment: bytearray}



|  |  |
| --- | --- |
|  |  |

2 ) Test the Relation

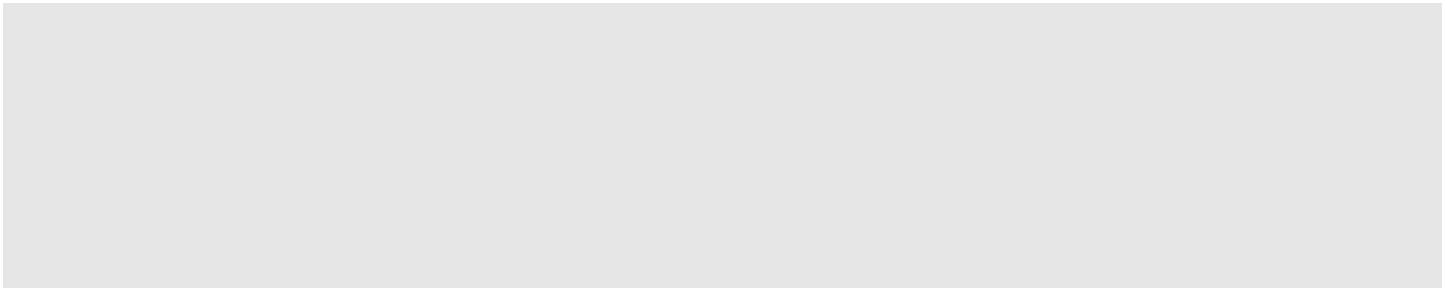
1. A common Pig task is to test a relation to make sure it is consistent with what you are intending it to be. But using DUMP on a big data relation might take too long or not be practical, so define a SAMPLE of comments:

grunt> comments\_sample = SAMPLE comments 0.001;



1. Now DUMP the comments\_sample relation. The output should be non-null comments about visitors to the White House, similar to:

grunt> DUMP comments\_sample;



(ATTENDEES VISITING FOR A MEETING)

(FORUM ON IT MANAGEMENT REFORM/)

(FORUM ON IT MANAGEMENT REFORM/)

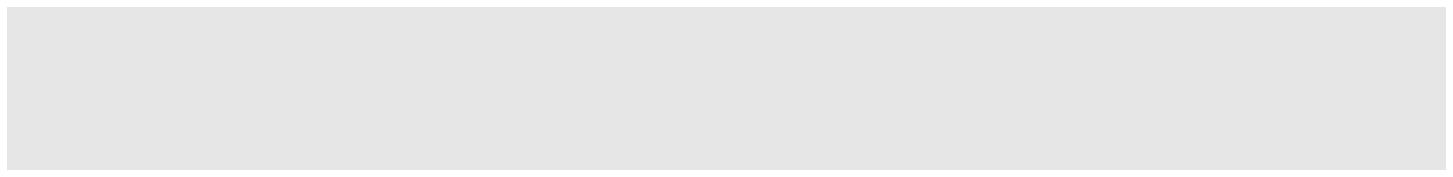
(HEALTH REFORM MEETING)

(DRIVER TO REMAIN WITH VEHICLE)

3 ) Count the Number of Comments

1. The comments relation represents all non-null comments from visits.txt. Write Pig statements that output the number of records in the comments relation. The correct result is 222,839 records.

***Solution*:**

****

comments\_all = GROUP comments ALL;

comments\_count = FOREACH comments\_all GENERATE

COUNT(comments);

DUMP comments\_count;

4 ) Split the Dataset

**Note**

Our end goal is find visitors to the White House who are also members of Congress. We could run our MapReduce job on the entire visits.txt dataset, but it is common in Hadoop to split data into smaller input files for specific tasks, which can greatly improve the performance of your MapReduce applications. In this step, you will split visits.txt into two separate datasets.

1. In this step, you will split visits.txt into two datasets: those that contain “CONGRESS” in the comments field, and those that do not.



1. Use the SPLIT command to split the visits relation into two new relations named congress and not\_congress:

grunt> SPLIT visits INTO congress IF($25 MATCHES '.\* CONGRESS .\*'), not\_congress IF (NOT($25 MATCHES '.\* CONGRESS .\*'));



1. Store the congress relation into a folder named ‘congress’. grunt> STORE congress INTO 'congress';
2. Similarly, STORE the not\_congress relation in a folder named

‘not\_congress’.



grunt> STORE not\_congress INTO 'not\_congress';



1. View the output folders using ls. The file sizes should be equivalent to the following:

grunt> ls congress



hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/congress/



\_SUCCESS<r 1>0

hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/congress/



part-m-00000<r 1>45618

hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/congress/



part-m-00001<r 1>0

grunt> ls not\_congress

hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/not\_congr



ess/\_SUCCESS<r 1>0

hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/not\_congr



ess/part-m-00000<r 1>90741587

hdfs://sandbox.hortonworks.com:8020/user/root/whitehouse/not\_congr



ess/part-m-00001<r 1>272381

1. View one of the output files in congress and make sure the string “CONGRESS” appears in the comment field:

grunt> cat congress/part-m-00000



5 ) Count the Results

***a.*** Write Pig statements that output the number of records in the congress

relation. This will tell us how many visitors to the White House have

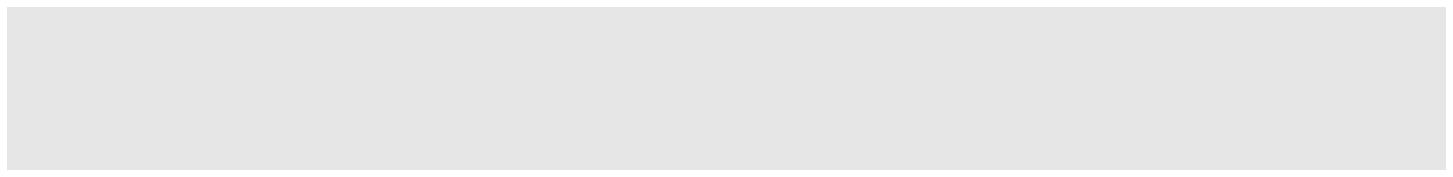
“CONGRESS” in the comments of their visit log. The correct result is 102.

**Note**

You now have two datasets: one in ‘congress,’ with 102 records, and the remaining records in the ‘not\_congress’ folder. These records are still in their original, raw format.



***Solution*:**

****

grunt> congress\_grp = GROUP congress ALL;

grunt> congress\_count = FOREACH congress\_grp GENERATE COUNT(congress);

grunt> DUMP congress\_count;

**Result**

You have just split ‘visits.txt’ into two datasets, and you have also discovered that 102 visitors to the White House had the word “CONGRESS” in their comments field. We will further explore these visitors in the next lab as we perform a join with a dataset containing the names of members of Congress.