ASSIGNMENT-5

BIG DATA FRAMEWORK

FACULTY :RAMESH BABU K SIR

NAME : Lavanya Bandla

REG NO : 19MCB0008

LAB : L15+L1

Web log Analysis

Project code :

There are two parts in this projects: 1)Find out
the most viewed page

Steps to find the most viewed page:

Step 1: First and foremost, we have to register the Piggybank jar to use its classes.

Step 2: Next, load the data using CombinedLogLoader() and specify the schema.

Step 3: Group the data by page link to count the page hits of each unique link.

Step 4: For every grouped data (grouped by link) we have to generate the link and its total count. Here, we have used flatten() to explode the tuples and then count the hits.

Step 5: Once COUNT is received, we need to order it in descending order and generate the only first result.

Step 6: use dump to get the desired result.

Commands:

```
 grunt> most_viewed = LOAD '/homr/Desltop/Sample_log.docx' USING PigStorage(‘,') AS
 (ip:chararray, datechararray,file_typechararray,numberchararray,linkchararray,spacechara
 rray,pagechararray); grunt>
 dump most_viewed;
grunt> grouped_link = GROUP most_viewed BY link; grunt>
dump grouped_link;
grunt> final_table = FOREACH grouped_link GENERATE COUNT(most_viewed.link) as
total_count,group;
 grunt> dump final_table;
 grunt> final_answer = LIMIT final_table 1;
 grunt> dump final_answer;
```

2)Total hits per day

Find total hits per unique day

Based on each unique day we need to find the total hits. For example, on 24th of a

particular month, there were X hits, on 27th of the month, there can be Y hits.

The assumption has been made that logs are of a single month.

To solve this problem, we have to use DateExtractor() available in Piggybank jar.

This will take the timestamp as input and will give corresponding “day” against

each timestamp.

Step 1) Define the DateExtractor() in the Pig Grunt shell as shown

Step 2) Use the above class defined to extract the day and group by it.

Step 3) To find the unique hits per day, run the below command.

Step 4) Dump the result and see the output.
The first column of the output is the date, and the second is the total number of hits on that day.

```
grunt> day_hits = LOAD '/homr/Desltop/Sample_log.docx' USING PigStorage ('') AS (ip:chararray, datechararray,hhint,mmint,sschararray,linkchararray);grunt> dumpday_hits;
```

```
grunt> day_hits = LOAD '/homr/Desltop/Sample_log.docx' USING PigStorage('') AS (ip:chararray,datechararray,hhint,mmint,sschararray,linkchararray); grunt> dumpday_hits;
```

```
grunt> hits_final = FOREACH hits_grouped GENERATE
COUNT(day_hits.ip_date) as number_of_hits,group;
grunt> dump hits_final;
```