Experiment 10: Week 11:

10. Install and Run Pig then write Pig Latin scripts to sort, group, join, project, and filter your data.

ALGORITHM

STEPS FOR INSTALLING APACHE PIG

1) Extract the pig-0.15.0.tar.gz and move to home directory

2) Set the environment of PIG in bashrc file.

3) Pig can run in two modes

Local Mode and Hadoop Mode
Pig –x local and pig

4) Grunt Shell

Grunt >

5) LOADING Data into Grunt Shell

```
   DATA = LOAD <CLASSPATH> USING PigStorage (DELIMITER) as
(ATTRIBUTE : DataType1, ATTRIBUTE : DataType2…..)
```

6) Describe Data

```
   Describe DATA;
```

7) DUMP Data

```
   Dump DATA;
```

8) FILTER Data

```
   FDATA = FILTER DATA by ATTRIBUTE = VALUE;
```

9) GROUP Data

```
   GDATA = GROUP DATA by ATTRIBUTE;
```

10) Iterating Data

```
   FOR_DATA = FOREACH DATA GENERATE GROUP AS GROUP_FUN,
ATTRIBUTE = <VALUE>
```

11) Sorting Data

SORT_DATA = ORDER DATA BY ATTRIBUTE WITH CONDITION;

12) LIMIT Data

LIMIT_DATA = LIMIT DATA COUNT;

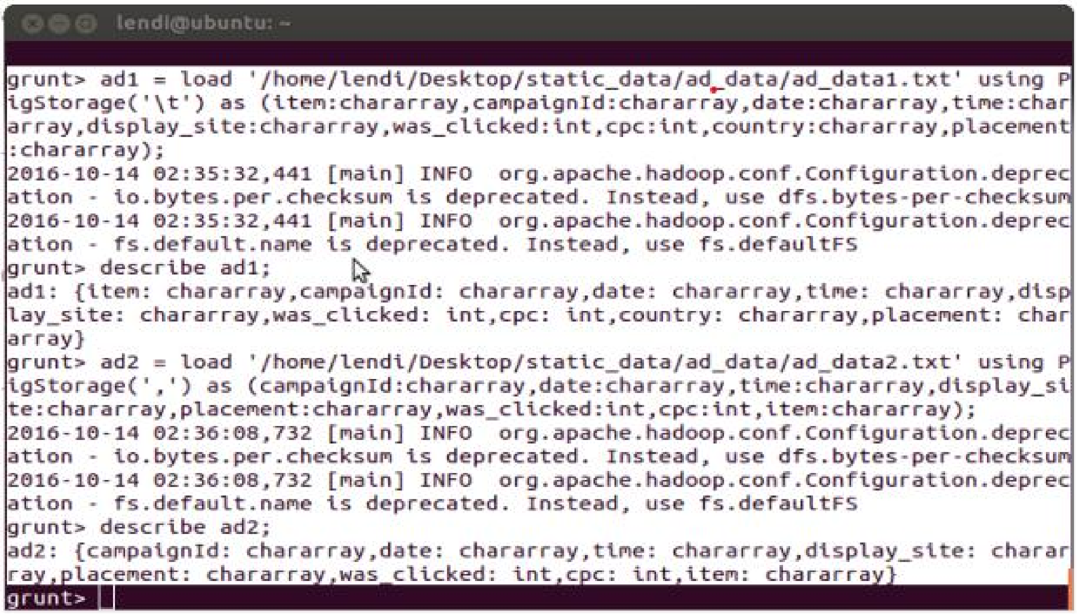
13) JOIN Data

JOIN DATA1 BY (ATTRIBUTE1,ATTRIBUTE2….) , DATA2 BY (ATTRIBUTE3,ATTRIBUTE….N)

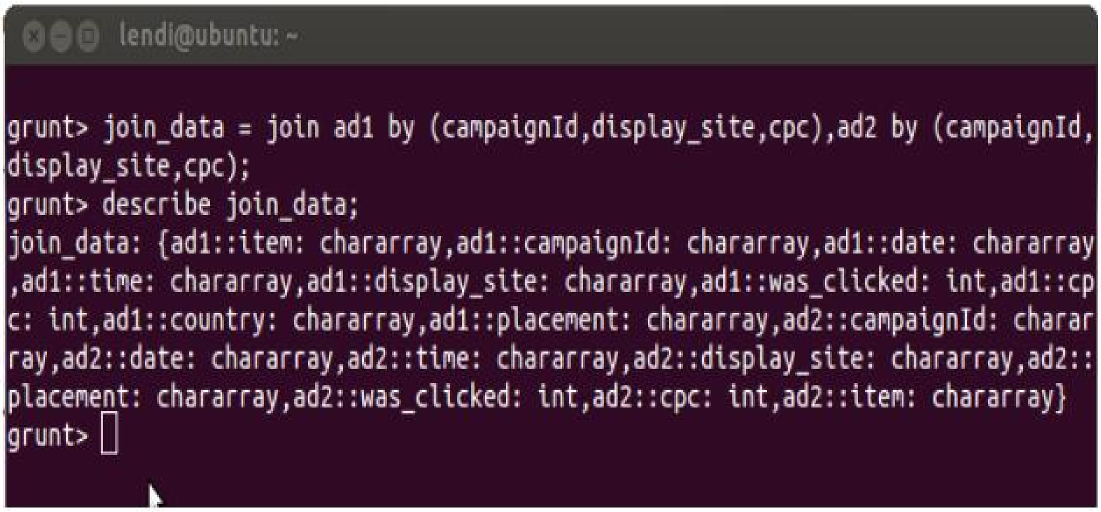
INPUT:

Input as Website Click Count Data

sOUTPUT:



```
grunt> ad1 = load '/home/lendi/Desktop/static_data/ad_data/ad_data1.txt' using P
igStorage('\t') as (item:chararray,campaignId:chararray,date:chararray,time:char
array,display_site:chararray,was_clicked:int,cpc:int,country:chararray,placement
:chararray);
2016-10-14 02:35:32,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-10-14 02:35:32,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe ad1;
ad1: {item: chararray,campaignId: chararray,date: chararray,time: chararray,disp
lay_site: chararray,was_clicked: int,cpc: int,country: chararray,placement: char
array}
grunt> ad2 = load '/home/lendi/Desktop/static_data/ad_data/ad_data2.txt' using P
igStorage(',') as (campaignId:chararray,date:chararray,time:chararray,display_si
te:chararray,placement:chararray,was_clicked:int,cpc:int,item:chararray);
2016-10-14 02:36:08,732 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-10-14 02:36:08,732 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> describe ad2;
ad2: {campaignId: chararray,date: chararray,time: chararray,display_site: charar
ray,placement: chararray,was_clicked: int,cpc: int,item: chararray}
grunt> 
```



```
grunt> join_data = join ad1 by (campaignId,display_site,cpc),ad2 by (campaignId,
display_site,cpc);
grunt> describe join_data;
join_data: {ad1::item: chararray,ad1::campaignId: chararray,ad1::date: chararray
,ad1::time: chararray,ad1::display_site: chararray,ad1::was_clicked: int,ad1::cp
c: int,ad1::country: chararray,ad1::placement: chararray,ad2::campaignId: charar
ray,ad2::date: chararray,ad2::time: chararray,ad2::display_site: chararray,ad2::
placement: chararray,ad2::was_clicked: int,ad2::cpc: int,ad2::item: chararray}
grunt> 
```