**Step 1 – Instructions 1 & 2:**

Script:

*--Load data*

*ad\_data\_network\_1 = LOAD '/user/alexhowell3369627/Homework\_1/ad\_data1.txt' USING PigStorage('\t')*

*AS (keyword : chararray*

*, campaign\_id : chararray*

*, date : chararray*

*, time: chararray*

*, display\_site : chararray*

*, was\_clicked : int*

*, cpc : int*

*, country : chararray*

*, placement : chararray*

*);*

*-- Limit Data*

*limit\_data = LIMIT ad\_data\_network\_1 10;*

*--Display Limited Data*

*DUMP limit\_data;*

Results:



**Step 1 – Instructions 3 & 4:**

Script:

*--Load data*

*ad\_data\_network\_1 = LOAD '/user/alexhowell3369627/Homework\_1/ad\_data1.txt' USING PigStorage('\t')*

*AS (keyword : chararray*

*, campaign\_id : chararray*

*, date : chararray*

*, time: chararray*

*, display\_site : chararray*

*, was\_clicked : int*

*, cpc : int*

*, country : chararray*

*, placement : chararray*

*);*

*--Filter*

*USA\_Filter = FILTER ad\_data\_network\_1 BY (country == 'USA');*

*--Reorder Columns*

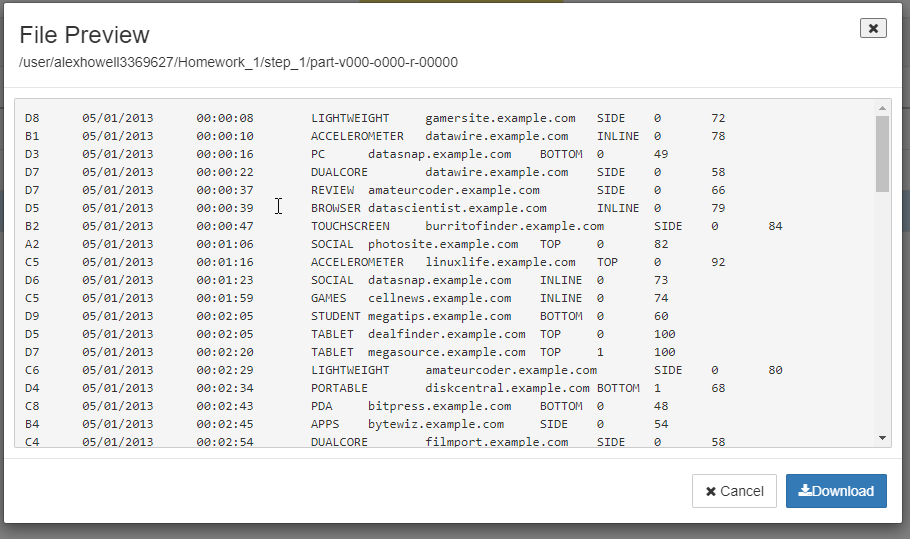
*Reordered = FOREACH USA\_Filter GENERATE campaign\_id, date, time, UPPER(TRIM(keyword)), display\_site, placement, was\_clicked, cpc;*

*--Store data*

*STORE Reordered INTO '/user/alexhowell3369627/Homework\_1/ad\_network\_data\_etl/step\_1';*

**Step 1 – Instruction 5:**

Results:



**Step 2 – Instruction 1:**

Script:

*--Load data*

*ad\_data\_network\_2 = LOAD '/user/alexhowell3369627/Homework\_1/ad\_data2.txt' USING PigStorage(',')*

*AS (campaign\_id : chararray*

*, date : chararray*

*, time: chararray*

*, display\_site : chararray*

*, placement : chararray*

*, was\_clicked : int*

*, cpc : int*

*, keyword : chararray*

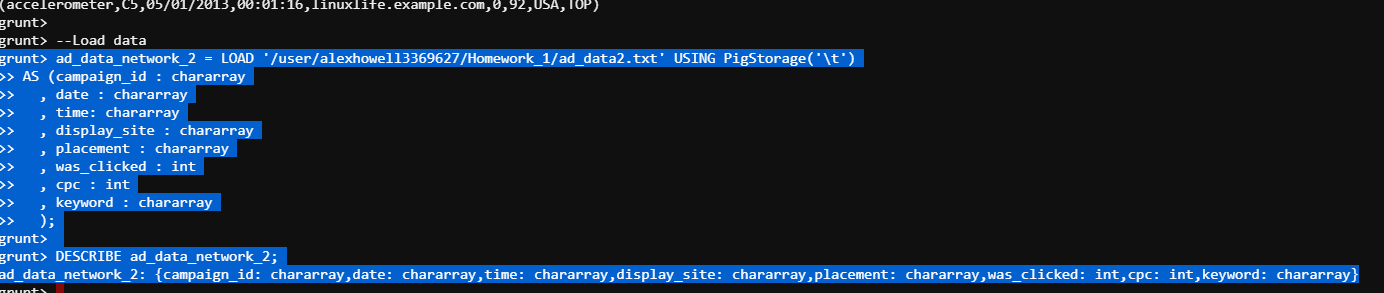
*);*

*--Describe loaded data*

*DESCRIBE ad\_data\_network\_2;*

**Step 2 – Instruction 2:**

Results:



**Step 2 – Instructions 3 & 4:**

Script:

*--Load data*

*ad\_data\_network\_2 = LOAD '/user/alexhowell3369627/Homework\_1/ad\_data2.txt' USING PigStorage(',')*

*AS (campaign\_id : chararray*

*, date : chararray*

*, time: chararray*

*, display\_site : chararray*

*, placement : chararray*

*, was\_clicked : int*

*, cpc : int*

*, keyword : chararray*

*);*

*--Get Rid of Duplicate Records*

*distinct\_ad\_data = DISTINCT ad\_data\_network\_2;*

*--Reorder Columns*

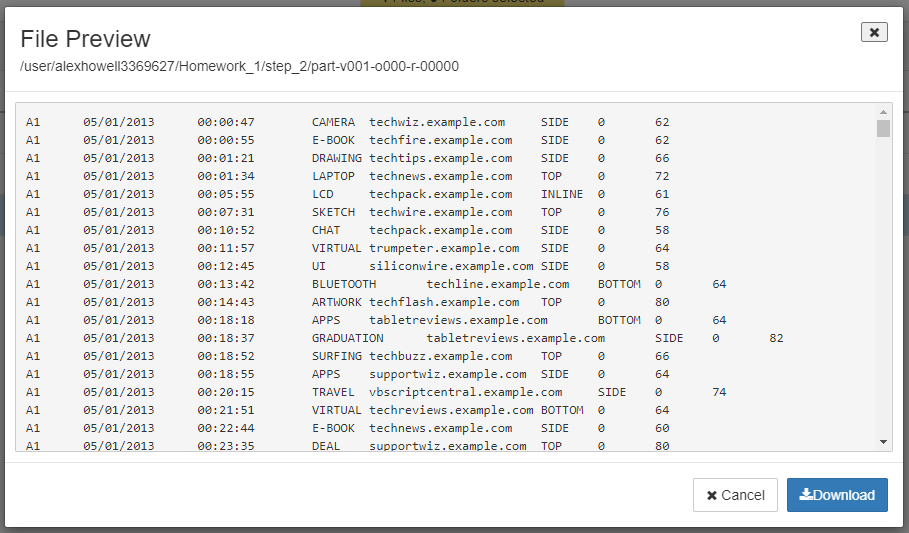
*Reordered = FOREACH distinct\_ad\_data GENERATE campaign\_id, REPLACE(date, '-', '/'), time, UPPER(TRIM(keyword)), display\_site, placement, was\_clicked, cpc;*

*--Store data*

*STORE Reordered INTO '/user/alexhowell3369627/Homework\_1/ad\_network\_data\_etl/step\_2';*

**Step 2 – Instruction 5:**

Results:



**Step 3 – Instructions 1 & 2:**

Script:

*-- A: This glob loads only the ad\_data1 and ad\_data2 directories*

*data = LOAD '/user/alexhowell3369627/Homework\_1/ad\_network\_data\_etl' AS (campaign\_id:chararray,*

*date:chararray, time:chararray,*

*keyword:chararray, display\_site:chararray,*

*placement:chararray, was\_clicked:int, cpc:int);*

*-- B: Includes only records where the ad was clicked*

*clicked = FILTER data BY (was\_clicked == 1);*

*-- C: Groups the data by the 'display\_site' field*

*persite = GROUP clicked BY display\_site;*

*-- D: Creates relation with 'display\_site' and sum of click cost*

*costofclicks = FOREACH persite GENERATE group, SUM (clicked.cpc) AS Total;*

*-- E: Sorts the new relation in ascending order of cost*

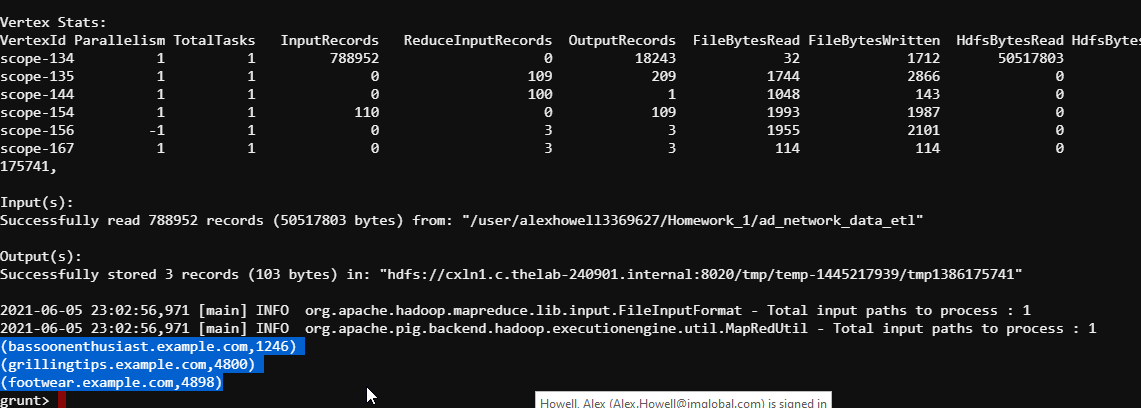
*costofclicksasc = ORDER costofclicks BY Total ASC;*

*-- F: Displays just the first three records to the screen*

*first3 = LIMIT costofclicksasc 3;*

*DUMP first3;*

Results:



**Step 4 – Instructions 1 - 3:**

Script:

*-- Load only the ad\_data1 and ad\_data2 directories*

*data = LOAD '/user/alexhowell3369627/Homework\_1/ad\_network\_data\_etl' AS (campaign\_id:chararray,*

*date:chararray, time:chararray,*

*keyword:chararray, display\_site:chararray,*

*placement:chararray, was\_clicked:int, cpc:int);*

*-- Include only records where the ad was clicked*

*clicked = FILTER data BY (was\_clicked == 1);*

*-- A: Group the data by the 'keyword' field*

*perkeyword = GROUP clicked BY keyword;*

*-- Create relation with 'keyword' and sum of click cost*

*costofclicks = FOREACH perkeyword GENERATE group, SUM (clicked.cpc) AS Total;*

*-- B: Sort the new relation in descending order of cost*

*costofclicksasc = ORDER costofclicks BY Total DESC;*

*-- C: Display just the first five records to the screen*

*first5 = LIMIT costofclicksasc 5;*

*DUMP first5;*

Results:

