# Project report on using Pig for ETL Processing:

**a) Introduction:**

In this project we used Pig to explore, correct and reorder data in files from two different ad networks. The first part was to experiment with sample data using in local mode, and once the ETL script works as expected, we used them to process the complete data sets in HDFS by using Pig in MapReduce mode. Now, the ETL to be written, is for *Dualcore* that supposedly has started using online advertisements to attract new customers to its e-commerce site. The two ad networks provide data about the ads they've placed like the site where the ad was placed and the when it was placed the keyword that triggered the ad to display, where the user clicked the ad and the per-click cost.

The data though, is in a different format for each network, also containing some invalid records. It is required that the file be processed before it gets analyzed.

**b) The format of the data in the original data input file:**

The original input data files are having format such that it contains nine fields namely, keyword(chararray), campaign_id(chararray), date(chararray), time(chararray), display_site(chararray), was_clicked(int), cpc(int), country(chararray), placement(chararray).

**c) Data load method in pig:**

Once the LOAD statement is defined in script is edited for the two data sets from the two add networks, we may run both the scripts in the local mode first for a sample of data to check if the fields are in the expected order and the values appear similar to that defined in the table of the input file.

**d) Data processing procedure:**

Further, for the first data set by the first ad network, the data is processed by filtering out all the where the country field does not contain USA and then create a new relation containing the fields namely, campaign_id, data, time, keyword, display_site, placement, was_clicked and cpc. Further keyword field is converted to uppercase and any leading or trailing whitespace is removed.

And for the second data set by the second as network, the data is processed by removing duplicate records and creating a new relation containing fields in order, campaign_id, date, time, keyword, display_site, placement, was_clicked and cpc. Also the UPPER and TRIM functions are used to correct the keyword field. In the date field, the separator is replaced by '/'.

**e) Data output & results:**

Running the pig in MapReduce mode to analyze the whole input file in HDFS, on both the data sets by the two ad networks processed their data as defined in the procedure and results in modified data sets that is filtered out and with some of the fields modified.

# Report on Analyzing Ad Campaign Data with Pig

**a) Introduction:**

This project involves writing the Pig scripts that analyzes the data to optimize advertising so that it can be economical for Dualcore while attracting new customers. It involves two factors, sites having lowest total cost and keywords that are most expensive.

**b) The format of data in original input data file:**

The input for this project is both of the earlier processed data sets by the two ad networks. Once the data is processed in the earlier project, the output of both of them is used as inout for this project.

**c) Data load method in Pig:**

Both the data files are loaded using the LOAD statement in the Pig script and defining the location of the data directory in the HDFS.

**d) Data Processing procedure:**

The data processing procedure for the for getting the low cost sites includes creating a ne relation to include only records where field *was_clicked* has a value 1 and then groupingit by the diplay_site field. Further a new relation is created that includes two fields display_site and total cost of all the clicks on that site and it is sorted in ascending order by cost.

For getting the high cost keywords, the data is processed by grouping the field keyword and sorting in descending order of cost.

**e) Data output and results:**

Running the first Pig script results in all the sites with their cost in ascending order so that lowest cost sites can be obtained. And further running the next Pig script results all the keywords along with their cost sorted in descending order so that highest costing keyword can be obtained.

# Report on Bonus Lab #1, #2 & #3

**a) Introduction:**

This project required calculating the count of all of Ad clicks, estimating maximum cost for the next Ad campaign and calculating the click through rate (percentage of ads shown that users actually clicked).

**b) The format of data in original input data file:**

The input for this project is both earlier processed data sets by the two ad networks. Once the data is processed in the earlier project, the output of both is used as input for this project while calculating ad click count, next ad campaign cost estimation and CTR.

**c) Data load method in Pig:**

Both the data files are loaded using the LOAD statement in the Pig script and defining the location of the data directory in the HDFS.

**d) Data Processing procedure:**

For calculating count of the total clicks the input data file is processed by filtering the fields where the field *was_clicked* has the value 1 and grouping the records so that the aggregate functions can be called. Further COUNT function is used to calculate the total number of clicked ads.

For estimating maximum cost for next Ad campaign, the data is processed by considering the possibility that any ad might be clicked and changing the aggregate function to the one that returns the maximum value in the cpc field. Further , the value returned by aggregate function is multiplied by the total number of clicks expected to have in next campaign.

For calculating click-through-rate, the data is processed by filtering the records to include only records where the ad was clicked and creating a new relation on the line that follows the FILTER statement which counts the number of records within the current group and calculating CTR by dividing number of clicks by the total number of ads shown and putting it in field ctr.

**e) Data output & result:**

The output of the pig script results in the total number of ad-clicks. Running the second Pig script results in estimated maximum cost of the next Ad campaign. Running the third Pig script results in sites with their click-through-rate.