LogData Analysis

We will see web log analysis using Apache Pig

Understanding Data

ip_addr : ip address of the client (hostname).
Time: time at which server finished processing request.
Request: request made by client
Page link: web page through which client made a request.

Problem Statement 1

- **Find out the most viewed page.**

1) LOAD the Log data set.

```
grunt> load_logdata = LOAD
'file:///home/cloudera/khasimbabu/PIG_Excercise/CaseStudy4/sample_log' USING PigStorage('|') as
(ip_addr:chararray,time:chararray,request:chararray,pagelink:chararray);
```

2) Select the list of columns from dataset

```
grunt> log_details = FOREACH load_logdata GENERATE ip_addr,pagelink;
```

3) Group log data by pagelink

```
grunt> grp_logdata = GROUP log_details by pagelink;

grunt> describe grp_logdata;

grp_logdata: {group: chararray, log_details: {(ip_addr: chararray,pagelink: chararray)}}
```

4) Count and Order the Pagelink info.

```
grunt> count_grp_logdata = FOREACH grp_logdata GENERATE flatten($0),COUNT($1);

grunt> order_grp_logdata = ORDER count_grp_logdata by $1 DESC;

grunt> dump order_grp_logdata;
```

Problem Statement 2

- **Find total hits per unique day**

```
grunt> Register Piggybank.jar;

grunt> DEFINE DATE_EXTRACT
org.apache.pig.piggybank.evaluation.util.apachelogparser.DateExtract

grunt> grpd = GROUP raw_data by DATE_EXTRACT(time);

grunt> hits_per_day = foreach grpd GENERATE flatten($0), COUNT($1);

grunt> dump hits_per_day
```