LogData Analysis

We will see web log analysis using Apache Hive
Server logs contain lots of information from web servers, application logs, user-generated. This case study will show you how to derive insights from the web server logs. The insights can be used for monitoring servers, user behaviour, fraud detection, improving business intelligence etc.

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.

1) Create a table in database and load the data
2) Load data from storage
3) Find the top endpoints that received server side error
4) Which resource is requested most frequently by the hosts
5) Display the top 10 host who made maximum requests to the server
6) Find the total count of different response codes returned by the server
7) How many hosts have accessed the server more than 1500 times
8) Find the average, maximum and minimum size of resource returned by the server
9) Find the total number of unique host sending request to server

```
hive> create table if not exists myLog(host string,time string,request string,pagelink string)
    > row format serde 'org.apache.hadoop.hive.serde2.RegexSerDe' with
    > serdeproperties("input.regex"="([^ ]*) ([^ ]*) ([^ ]*) (-
    > |\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)" ,
    > "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s
    > %7$s %8$s") STORED AS TEXTFILE;

hive> load data local inpath 'file:///home/cloudera/khasimbabu/HIVE_Excercise/sample_log'
overwrite into table mylog;

hive> select * from mylog limit 5;

OK

mylog.host      mylog.time      mylog.request   mylog.pagelink

Failed with exception java.io.IOException:org.apache.hadoop.hive.serde2.SerDeException: Number of matching groups doesn't match the number of columns

Time taken: 0.043 seconds
```