WEB SERVER LOG ANALYSIS

Assume you host an e-commerce website. In order to understand your customers better, you want to analyse your Apache web logs to discover how people are finding your site. The web server logs, however, are too large to import into a MySQL database, and they are not in a relational format. You need another way to analyse them.

Data Description:

The compressed file contains the client requests captured by a Web server.

```
21.125.155.111 - - [01/Jan/2012:12:07:48 +0530] "GET /digital-cameras/digital-camera/sony-qx-dsc-qx100-point-shoot-digital-camera-black.html HTTP/1.1" 200 1470 "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17" "-"
```

You will find the below fields listed in that file.

Host: 21.125.155.111 (IP address of the client (remote host) which made the request)

Identity: - (Identity of the client)

User: - (userid of the person requesting the document)

Date, Time and Timezone: [01/Jan/2012:12:07:48 +0530],

Request Line: "GET /digital-cameras/digital-camera/sony-qx-dsc-qx100-point-shoot-digital-camera-black.html HTTP/1.1",

Status code: 200 *(Note: 2xx is a successful response, 3xx a redirection, 4xx a client error, and 5xx a server error.)*

Object size: 1470 is the size of the object returned to the client, measured in bytes.

Agent: "Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17"

Referrer URL-

More details about apache web server log structure –

http://articles.slicehost.com/2010/8/27/reading-apache-web-logs

Problem Statement:

1. Load data into HDFS using **HDFS client**

2. Develop **MR program** to parse logs and convert request string into structured format (/a/b/c/d => a b c d)
50.57.190.149 - - 22/Apr/2012:07:12:41 +0530 GET /computers/laptops.html?brand=819 HTTP/1.0
computers - -laptops.html brand=819 200 12530 - -

Develop **MR/Pig/Hive** program to extract data for the following KPIs.
3. Count of page views by individual user

4. Top / Bottom 5: catagery-1/ catagery-2 / page /users / entry pages (Exclude status code other than 200, also exclude record related to css/js/image)

5. Total page views / Category wise pageviews / Unique pageviews

6. Count of status code = 200 / 404 / 400 / 500

7. Load results into tables in MySql Database using **Sqoop**.