# Program 4

**AIM: Web Log Analysis using Mapper-Reducer on single node cluster.**

Web logs are data that is generated by web servers for requests they receive. There are various web servers such as Apache, Nginx, Tomcat, and so on. Each web server logs data in a specific format.

The server *access log* records all requests processed by the server.

Of course, storing the information in the access log is only the start of log management. The next step is to analyze this information to produce useful statistics. We can store log data in many different format but here we are going to use data from the Apache Web Server, which is in *combined access logs.*

***Combined Log Format***

A typical configuration for the access log might look as follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined
CustomLog log/acces_log combined
```

**Example:** `127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"`

**127.0.0.1 (%h)**
This is the IP address of the client (remote host) which made the request to the server.

**- (%l)**
The "hyphen" in the output indicates that the requested piece of information is not available

**frank (%u)**
This is the userid of the person requesting the document as determined by HTTP authentication.

**[10/Oct/2000:13:55:36 -0700] (%t)**
The time that the server finished processing the request.

The format is:
[day/month/year:hour:minute:second zone]
day = 2*digit
month = 3*letter
year = 4*digit
hour = 2*digit
minute = 2*digit
second = 2*digit
zone = (`+' | `-') 4*digit

**"GET /apache_pb.gif HTTP/1.0" (\"%r\")**

The request line from the client is given in double quotes. The request line contains a great deal of useful information. First, the method used by the client is GET. Second, the client requested the resource /apache_pb.gif, and third, the client used the protocol HTTP/1.0.

**200 (%>s)**

This is the status code that the server sends back to the client. This information is very valuable, because it reveals whether the request resulted in a successful response (codes beginning in 2), a redirection (codes beginning in 3), an error caused by the client (codes beginning in 4), or an error in the server (codes beginning in 5).

**2326 (%b)**

The last entry indicates the size of the object returned to the client (in bytes), not including the response headers.

**"http://www.example.com/start.html" (\"%{Referer}i\")**

The "Referer" (sic) HTTP request header. This gives the site that the client reports having been referred from. (This should be the page that links to or includes /apache_pb.gif). This is the page that is linked to this URL.

"**Mozilla/4.08 [en] (Win98; I ;Nav)" (\"%{User-agent}i\")**

This is the browser identification string.

We can write map reduce programs to analyze various aspects of web log data. In this program, we are going to write a map reduce program that reads a web log file, give URL address and their counts. (URL, Total Count)

**** Steps: Write down same as you have written in program 2 & 3 (for web log data).**
**** Results**

This way, you can write similar programs for the following:

- Most number of referral sites (hint: use a referral group from the matcher)
- Number of client errors (with the Http status of 4XX)
- Number of of server errors (with the Http status of 5XX)