Web Log Analysis using Hive

A **Server log** is a **log file** automatically created and maintained by a server consists of a list of activities it performed. A typical example is a **web server log** which maintains a history of page requests.

The W3C maintains a standard format (the Common Log Format) for web server log files. Information about the request, including client IP address, request date/time, the page requested, HTTP code, byte served, user agent and referrer are typically added. This data can be combined into a single file, or separated into distinct logs, such as an **access log**, **error log**, or **referrer log**.

Hive is a data warehouse infrastructure that provides data summarization and ad-hoc querying. Hive provides an SQL dialect, called Hive Query Language (HQL) for querying data stored in a Hadoop cluster.

Hive's data model provides a high-level, table-like structure on top of HDFS.

Web Log Format:-

```
64.242.88.10 - - [07/Mar/2014:16:20:55 -0800] "GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1" 200 5253
```

Ipaddress %h : (64.242.88.10) ip address of the client (hostname).

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

Userid %u : (-) This is the userid of the person request the document as determined by the HTTP authentication. "–" present then the requested information is not available NA

Timestamp %t : [07/Mar/2014:16:20:55 -0800] time at which server finished processing 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

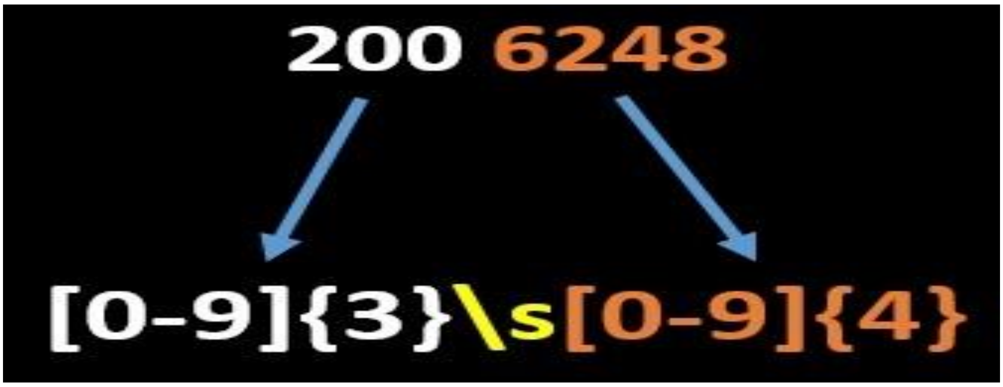
Request %r : "GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1" request made by client. Denoted by "GET"

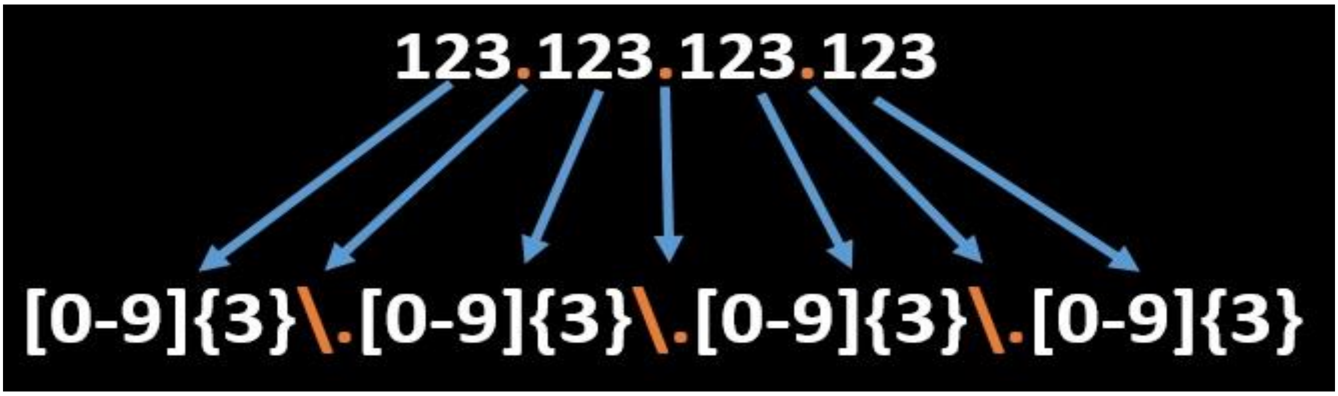
Status code %s : 200 is the HTTP status code returned to the client.
2xx is a successful response,
3xx a redirection,
4xx a client error,
5xx a server error.

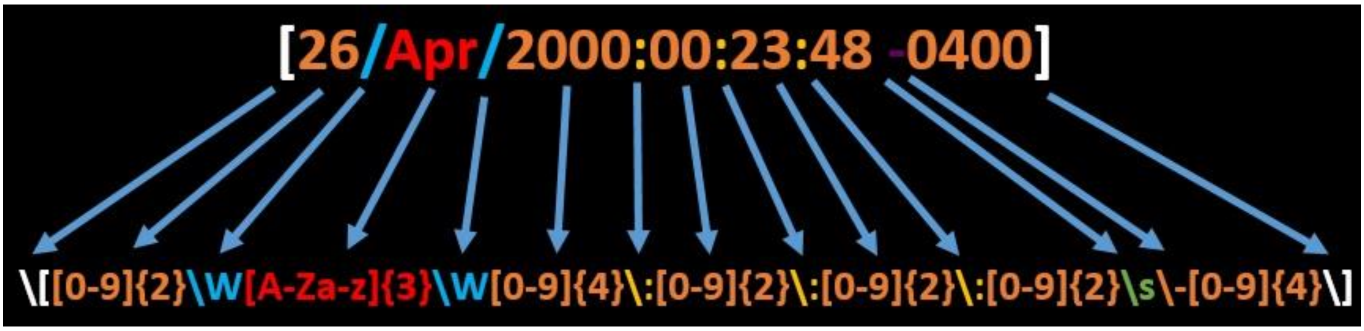
Size of Object %b : 5253 is the size of the object returned to the client, measured in bytes.

Regular Expression Regex:-

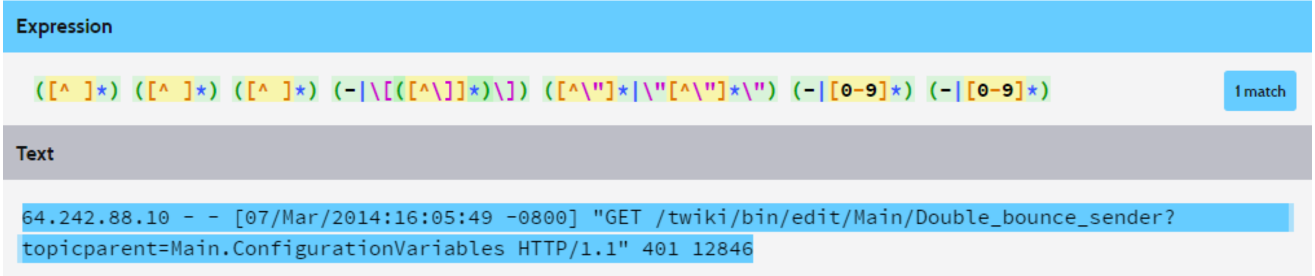
Refer the link to learn Regular Expression Regular Expression tutorial






[26/Apr/2000:00:23:48 -0400]
\[[0-9]{2}\W[A-Za-z]{3}\W[0-9]{4}\:[0-9]{2}\:[0-9]{2}\:[0-9]{2}\s\-[0-9]{4}\]

Below regular expression, we will use for log pattern matching.



```
([^ ]*) ([^ ]*) ([^ ]*) (-|\[([^\]]*)\]) ([^\"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)
```

```
64.242.88.10 - - [07/Mar/2014:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?
topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
```

Download Input **log_access** file

In the below HiveQL script, we are using RegexSerDe class to process the log file with the help of above regular expression.

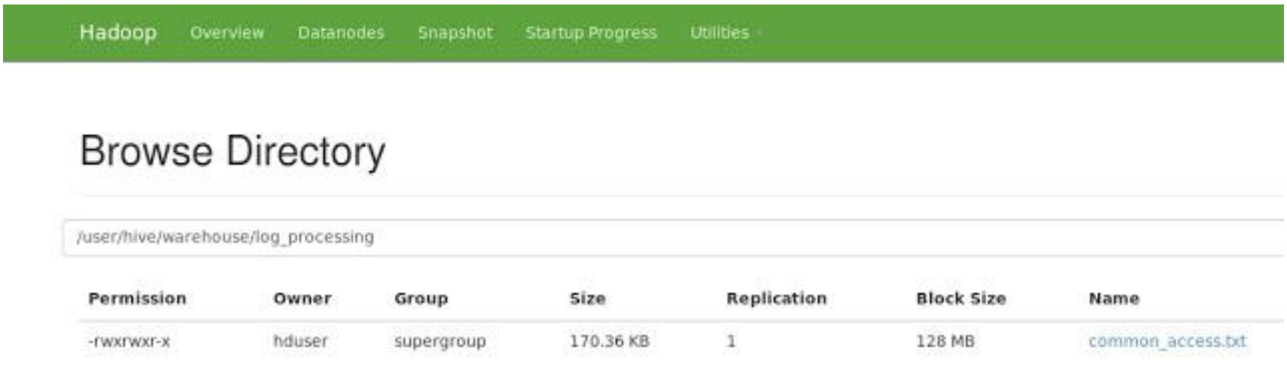
Hive script **logprochiveregex.hql**

```
--Hive is a data warehouse infrastructure that provides data summarization and ad-hoc querying.
--Hive provides an SQL dialect, called Hive Query Language (abbreviated HQL) for querying data stored
in a Hadoop cluster.
--Hive's data model provides a high-level, table-like structure on top of HDFS.
--Deleting existing table
DROP TABLE IF EXISTS log_processing;

--Here we are creating Managed table for log data
CREATE TABLE log_processing (
ipaddress STRING,
logname STRING,
userid STRING,
time STRING,
request STRING,
status STRING,
size 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"
)
STORED AS TEXTFILE;

--Loading the log data into the table log_proc
LOAD DATA LOCAL INPATH "/home/hduser/HIVE/common_access.txt" INTO TABLE
log_processing;

--describe the schema of the table
describe formatted log_processing;

--select the records from the table
SELECT * FROM log_processing ORDER BY time LIMIT 10;
```

Execution of the Hive script **logprochiveregex.hql**

```
[hduser@localhost bin]$ hive -f /home/hduser/HIVE/logprochiveregex.hql

Logging initialized using configuration in jar:file:/usr/local/hadoop-2.6.0/hive/lib/hive-common-2.1.0.jar!/hive-
log4j2.properties Async: true
OK
Time taken: 8.815 seconds
OK
Time taken: 1.496 seconds
Loading data to table default.log_processing
OK
Time taken: 2.1 seconds
OK
# col_name            data_type            comment

ipaddress            string
logname              string
userid               string
time                 string
request              string
status               string
size                 string

# Detailed Table Information
Database:            default
Owner:               hduser
CreateTime:          Fri Mar 24 09:32:53 PDT 2017
LastAccessTime:      UNKNOWN
Retention:           0
Location:            hdfs://localhost:9000/user/hive/warehouse/log_processing
Table Type:          MANAGED_TABLE
Table Parameters:
 numFiles            1
 numRows             0
 rawDataSize         0
 totalSize           174447
 transient_lastDdlTime 1490373175

# Storage Information
SerDe Library:       org.apache.hadoop.hive.serde2.RegexSerDe
InputFormat:         org.apache.hadoop.mapred.TextInputFormat
OutputFormat:        org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:          No
Num Buckets:         -1
Bucket Columns:      []
Sort Columns:        []
```

```
Storage Desc Params:
 input.regex         ([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)
 output.format.string %1$s %2$s %3$s %4$s %5$s %6$s %7$s
 serialization.format 1
Time taken: 0.861 seconds, Fetched: 37 row(s)
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
Starting Job = job_1490354948174_0003, Tracking URL = http://localhost:8088/proxy/application_1490354948174_0003/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1490354948174_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-24 09:36:07,239 Stage-1 map = 0%,  reduce = 0%
2017-03-24 09:39:03,514 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 41.82 sec
2017-03-24 09:39:05,651 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 42.37 sec
2017-03-24 09:40:06,470 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 45.91 sec
MapReduce Total cumulative CPU time: 45 seconds 910 msec
Ended Job = job_1490354948174_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 45.91 sec   HDFS Read: 184082 HDFS Write: 1423 SUCCESS
Total MapReduce CPU Time Spent: 45 seconds 910 msec
OK
64.242.88.10 - - [07/Mar/2014:16:05:49 -0800] "GET
/twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2014:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2
HTTP/1.1" 200 4523
64.242.88.10 - - [07/Mar/2014:16:10:02 -0800] "GET /mailman/listinfo/hsdivision HTTP/1.1" 200 6291
64.242.88.10 - - [07/Mar/2014:16:11:58 -0800] "GET /twiki/bin/view/TWiki/WikiSyntax HTTP/1.1" 200 7352
64.242.88.10 - - [07/Mar/2014:16:20:55 -0800] "GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1" 200 5253
64.242.88.10 - - [07/Mar/2014:16:23:12 -0800] "GET
/twiki/bin/oops/TWiki/AppendixFileSystem?template=oopsmore&param1=1.12&param2=1.12 HTTP/1.1" 200 11382
64.242.88.10 - - [07/Mar/2014:16:24:16 -0800] "GET /twiki/bin/view/Main/PeterThoeny HTTP/1.1" 200 4924
64.242.88.10 - - [07/Mar/2014:16:29:16 -0800] "GET
/twiki/bin/edit/Main/Header_checks?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2014:16:30:29 -0800] "GET /twiki/bin/attach/Main/OfficeLocations HTTP/1.1" 401 12851
64.242.88.10 - - [07/Mar/2014:16:31:48 -0800] "GET /twiki/bin/view/TWiki/WebTopicEditTemplate HTTP/1.1" 200 3732
Time taken: 432.103 seconds, Fetched: 10 row(s)
```



Browse Directory

/user/hive/warehouse/log_processing

Permission	Owner	Group	Size	Replication	Block Size	Name
-rwxrwxr-x	hduser	supergroup	170.36 KB	1	128 MB	common_access.txt