Analyzing Apache Access Logs in Apache Hive

The following statistics are analyzed:

1. A count of response code's returned from the server.
2. The content size of responses returned from the server to host.
3. The top ten most popular URL’s in the Apache log
4. The average, min, and max content size of responses returned from the server.

The steps to process data with Apache Hive

Before proceed the below steps, we have to install the Cloudera Quickstart vm 5.5 and VMwareplayer. The Hadoop 2.6, Java 1.7, Eclipse Luna, Hive, Hbase, Spark, and all required libraries have been included in cloudera.

1. Download the apache log file from http://www.monitorware.com/en/logsamples/apache.php and unzip it.

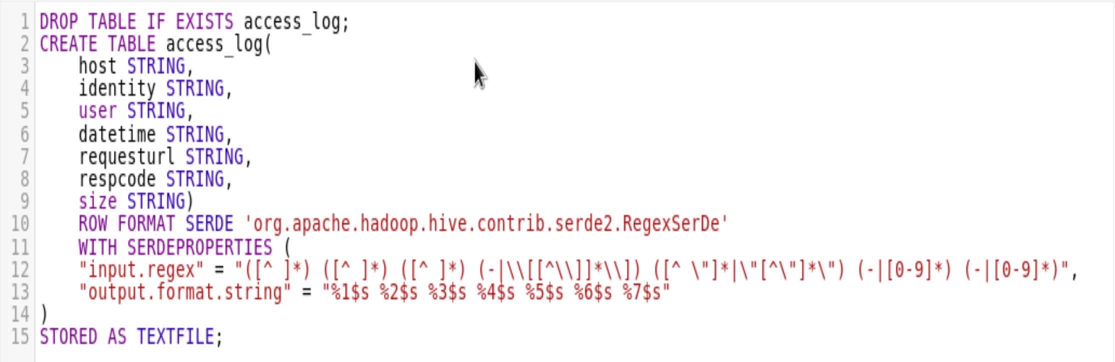
2. Create a loganalyzer/input directory named path in HDFS.

   ```
   hadoop fs -mkdir -p /user/cloudera/hive/input
   ```

3. Copy the log file from the local file system to directory within the HDFS.

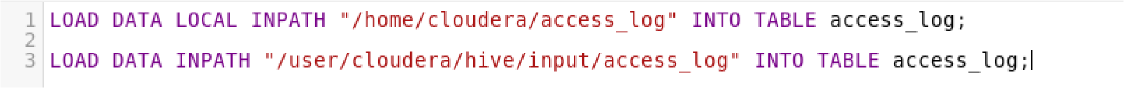
   ```
   hadoop fs -put access_log /user/cloudera/hive/input/
   ```

4. Create appropriate table for string Apache logs.



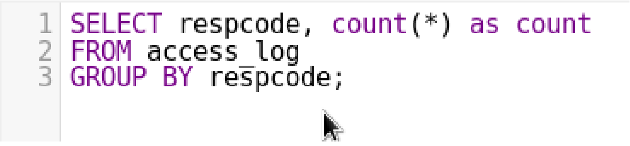
```
DROP TABLE IF EXISTS access_log;
CREATE TABLE access_log(
    host STRING,
    identity STRING,
    user STRING,
    datetime STRING,
    requesturl STRING,
    respcode STRING,
    size STRING)
    ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
    WITH SERDEPROPERTIES (
    "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;
```

5. Load access_log file, depending location of file (local file system or HDFS) do on of followings.



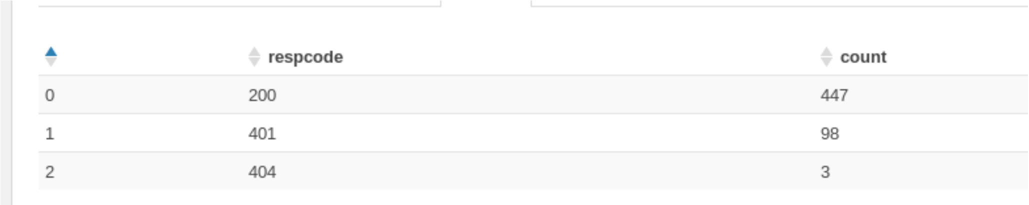
```
LOAD DATA LOCAL INPATH "/home/cloudera/access_log" INTO TABLE access_log;

LOAD DATA INPATH "/user/cloudera/hive/input/access_log" INTO TABLE access_log;
```

6. List the count of response code's returned from the server



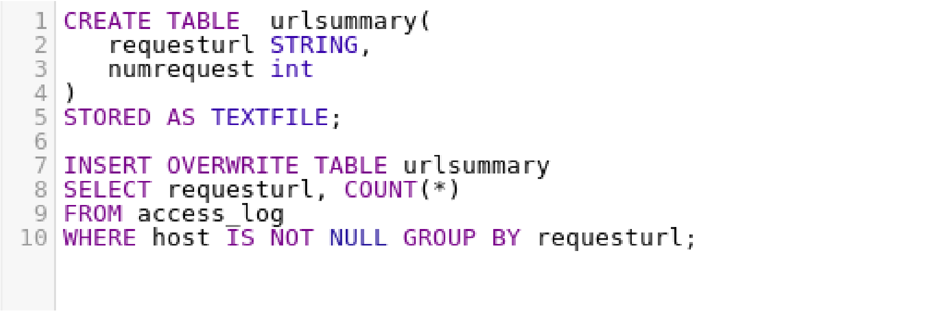
```
SELECT respcode, count(*) as count
FROM access_log
GROUP BY respcode;
```

Result

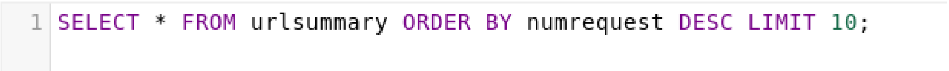


	respcode	count
0	200	447
1	401	98
2	404	3

7. List the top 10 most popular URL’s in the Apache log

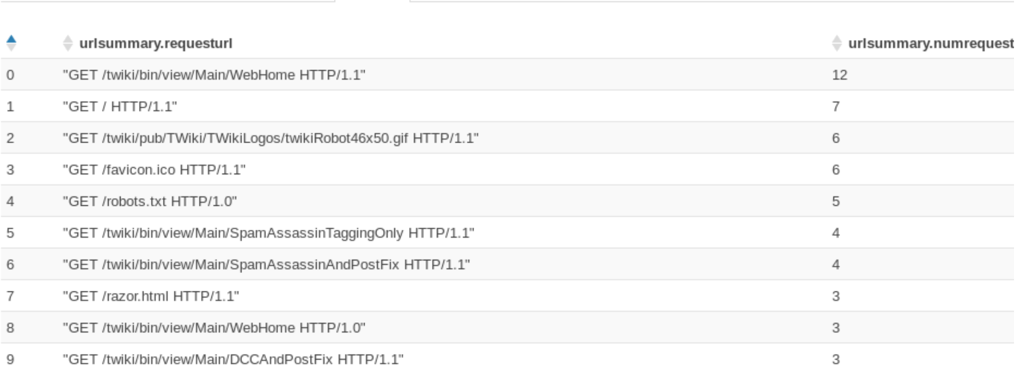


```
CREATE TABLE  urlsummary(
   requesturl STRING,
   numrequest int
)
STORED AS TEXTFILE;

INSERT OVERWRITE TABLE urlsummary
SELECT requesturl, COUNT(*)
FROM access_log
WHERE host IS NOT NULL GROUP BY requesturl;
```



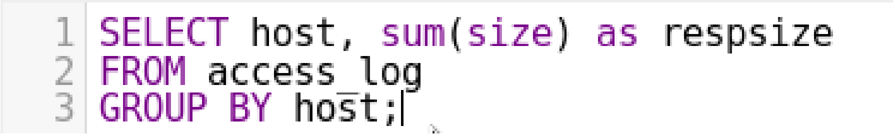
```
SELECT * FROM urlsummary ORDER BY numrequest DESC LIMIT 10;
```

Result



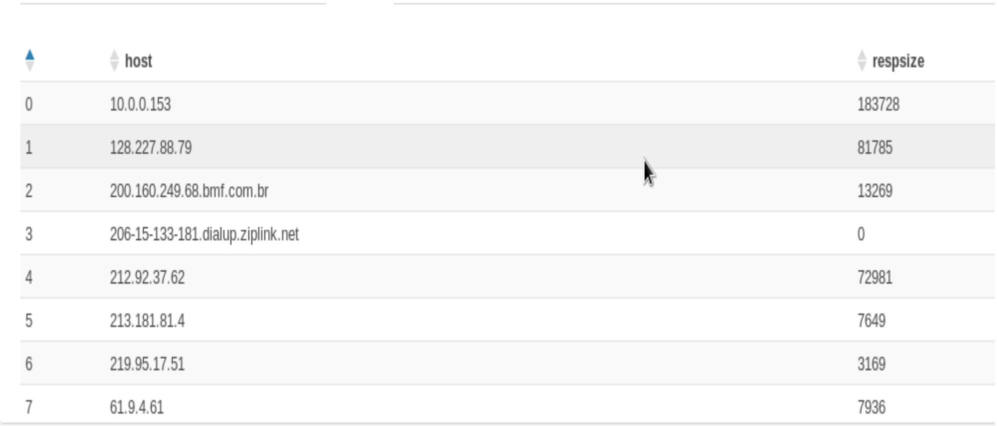
	urlsummary.requesturl	urlsummary.numrequest
0	"GET /twiki/bin/view/Main/WebHome HTTP/1.1"	12
1	"GET / HTTP/1.1"	7
2	"GET /twiki/pub/TWiki/TWikiLogos/twikiRobot46x50.gif HTTP/1.1"	6
3	"GET /favicon.ico HTTP/1.1"	6
4	"GET /robots.txt HTTP/1.0"	5
5	"GET /twiki/bin/view/Main/SpamAssassinTaggingOnly HTTP/1.1"	4
6	"GET /twiki/bin/view/Main/SpamAssassinAndPostFix HTTP/1.1"	4
7	"GET /razor.html HTTP/1.1"	3
8	"GET /twiki/bin/view/Main/WebHome HTTP/1.0"	3
9	"GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1"	3

8. List the content size of responses returned from the server.



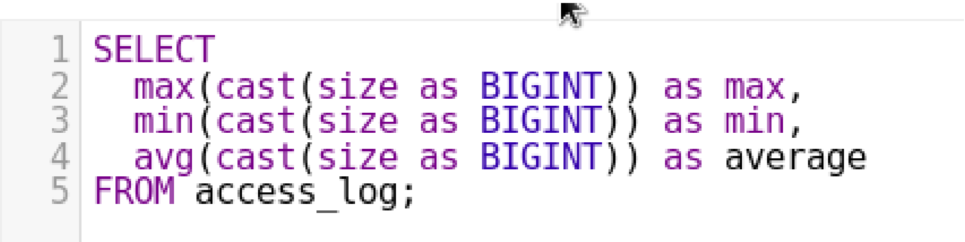
```
SELECT host, sum(size) as respsize
FROM access_log
GROUP BY host;
```

Result



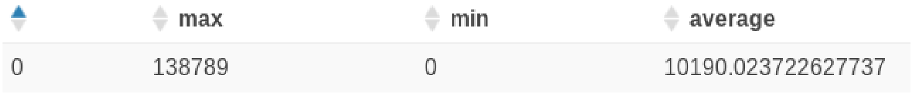
	host	respsize
0	10.0.0.153	183728
1	128.227.88.79	81785
2	200.160.249.68.bmf.com.br	13269
3	206-15-133-181.dialup.ziplink.net	0
4	212.92.37.62	72981
5	213.181.81.4	7649
6	219.95.17.51	3169
7	61.9.4.61	7936

9. List the average, min, and max content size of responses returned from the server



```
SELECT
  max(cast(size as BIGINT)) as max,
  min(cast(size as BIGINT)) as min,
  avg(cast(size as BIGINT)) as average
FROM access_log;
```

Result



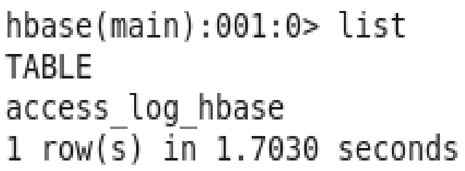
	max	min	average
0	138789	0	10190.023722627737

Hive and HBase Integration

Steps for hive and hbase integration

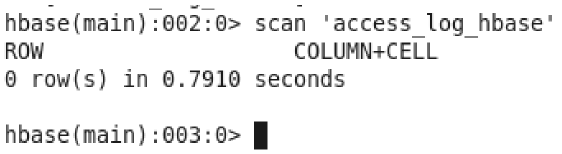
1. Create a table access_log_hbase and columnfamily as m in hbase
   Create 'access_log_hbase', 'm'

2. List the table in hbase



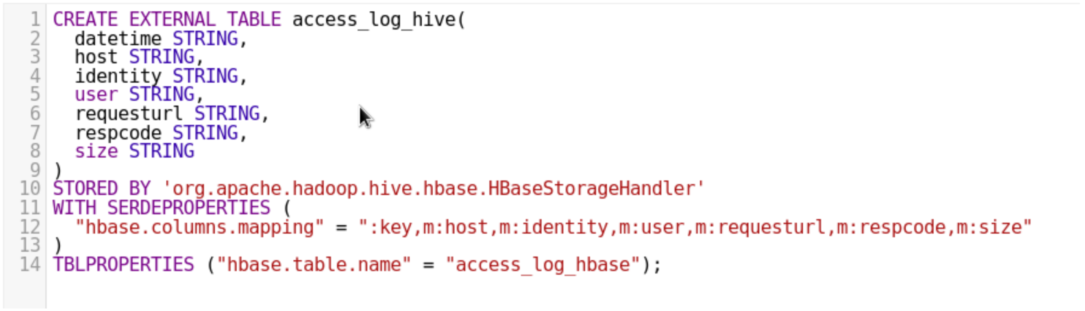
```
hbase(main):001:0> list
TABLE
access_log_hbase
1 row(s) in 1.7030 seconds
```

3. View the data in **access_log_hbase**



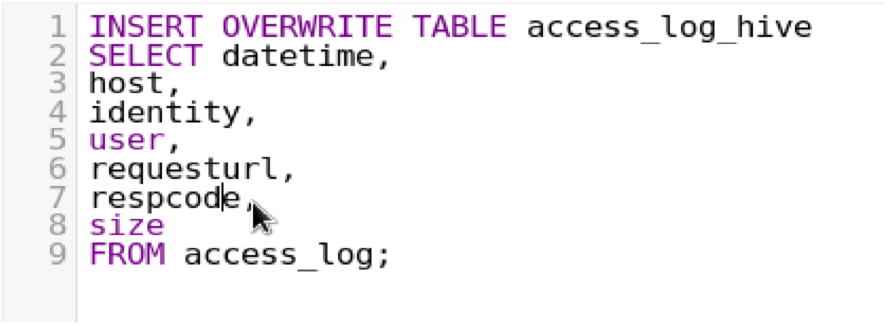
```
hbase(main):002:0> scan 'access_log_hbase'
ROW                        COLUMN+CELL
0 row(s) in 0.7910 seconds

hbase(main):003:0> 
```

4. Create an external table in hive as **access_log_hive**



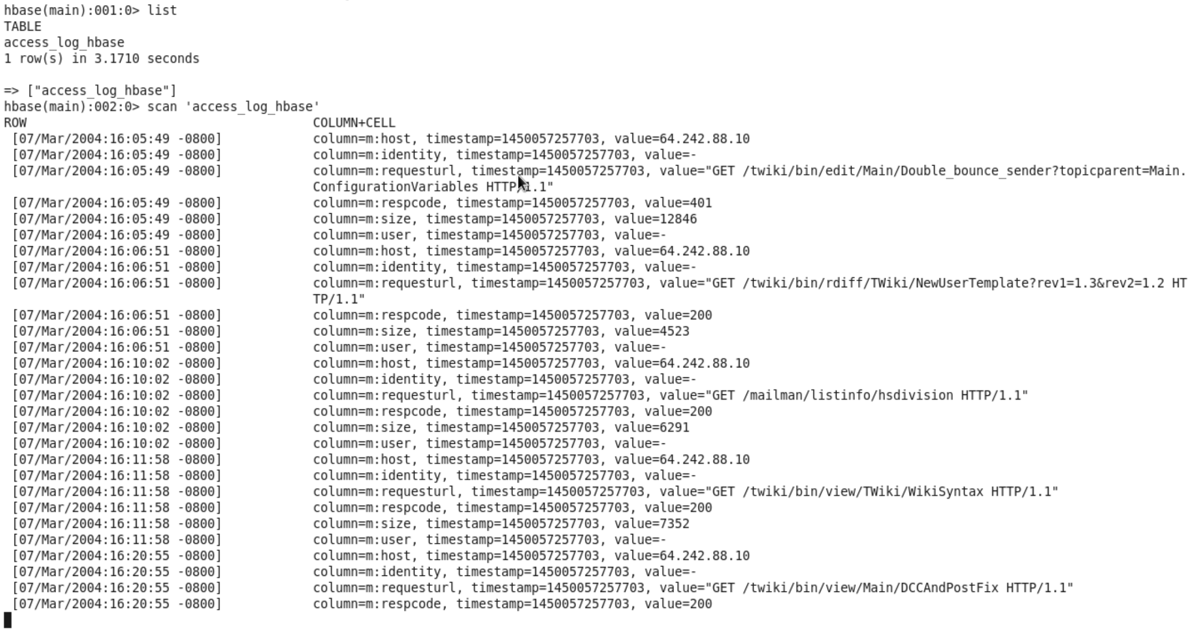
```
CREATE EXTERNAL TABLE access_log_hive(
  datetime STRING,
  host STRING,
  identity STRING,
  user STRING,
  requesturl STRING,
  respcode STRING,
  size STRING
)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES (
  "hbase.columns.mapping" = ":key,m:host,m:identity,m:user,m:requesturl,m:respcode,m:size"
)
TBLPROPERTIES ("hbase.table.name" = "access_log_hbase");
```

5. Overwrite the hive as **access_log_hive** with existing hive table **access_log**



```
INSERT OVERWRITE TABLE access_log_hive
SELECT datetime,
host,
identity,
user,
requesturl,
respcode,
size
FROM access_log;
```

6. View the data in access_log_hbase



```
hbase(main):001:0> list
TABLE
access_log_hbase
1 row(s) in 3.1710 seconds

=> ["access_log_hbase"]
hbase(main):002:0> scan 'access_log_hbase'
ROW                                      COLUMN+CELL
 [07/Mar/2004:16:05:49 -0800]            column=m:host, timestamp=1450057257703, value=64.242.88.10
 [07/Mar/2004:16:05:49 -0800]            column=m:identity, timestamp=1450057257703, value=-
 [07/Mar/2004:16:05:49 -0800]            column=m:requesturl, timestamp=1450057257703, value="GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.
                                         ConfigurationVariables HTTP/1.1"
 [07/Mar/2004:16:05:49 -0800]            column=m:respcode, timestamp=1450057257703, value=401
 [07/Mar/2004:16:05:49 -0800]            column=m:size, timestamp=1450057257703, value=12846
 [07/Mar/2004:16:05:49 -0800]            column=m:user, timestamp=1450057257703, value=-
 [07/Mar/2004:16:06:51 -0800]            column=m:host, timestamp=1450057257703, value=64.242.88.10
 [07/Mar/2004:16:06:51 -0800]            column=m:identity, timestamp=1450057257703, value=-
 [07/Mar/2004:16:06:51 -0800]            column=m:requesturl, timestamp=1450057257703, value="GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HT
                                         TP/1.1"
 [07/Mar/2004:16:06:51 -0800]            column=m:respcode, timestamp=1450057257703, value=200
 [07/Mar/2004:16:06:51 -0800]            column=m:size, timestamp=1450057257703, value=4523
 [07/Mar/2004:16:06:51 -0800]            column=m:user, timestamp=1450057257703, value=-
 [07/Mar/2004:16:10:02 -0800]            column=m:host, timestamp=1450057257703, value=64.242.88.10
 [07/Mar/2004:16:10:02 -0800]            column=m:identity, timestamp=1450057257703, value=-
 [07/Mar/2004:16:10:02 -0800]            column=m:requesturl, timestamp=1450057257703, value="GET /mailman/listinfo/hsdivision HTTP/1.1"
 [07/Mar/2004:16:10:02 -0800]            column=m:respcode, timestamp=1450057257703, value=200
 [07/Mar/2004:16:10:02 -0800]            column=m:size, timestamp=1450057257703, value=6291
 [07/Mar/2004:16:10:02 -0800]            column=m:user, timestamp=1450057257703, value=-
 [07/Mar/2004:16:11:58 -0800]            column=m:host, timestamp=1450057257703, value=64.242.88.10
 [07/Mar/2004:16:11:58 -0800]            column=m:identity, timestamp=1450057257703, value=-
 [07/Mar/2004:16:11:58 -0800]            column=m:requesturl, timestamp=1450057257703, value="GET /twiki/bin/view/TWiki/WikiSyntax HTTP/1.1"
 [07/Mar/2004:16:11:58 -0800]            column=m:respcode, timestamp=1450057257703, value=200
 [07/Mar/2004:16:11:58 -0800]            column=m:size, timestamp=1450057257703, value=7352
 [07/Mar/2004:16:11:58 -0800]            column=m:user, timestamp=1450057257703, value=-
 [07/Mar/2004:16:20:55 -0800]            column=m:host, timestamp=1450057257703, value=64.242.88.10
 [07/Mar/2004:16:20:55 -0800]            column=m:identity, timestamp=1450057257703, value=-
 [07/Mar/2004:16:20:55 -0800]            column=m:requesturl, timestamp=1450057257703, value="GET /twiki/bin/view/Main/DCCAndPostFix HTTP/1.1"
 [07/Mar/2004:16:20:55 -0800]            column=m:respcode, timestamp=1450057257703, value=200
```