Web Usage Mining

Team Members:
Manasa Pola
Sravya Balagala
Himaja Tirumalasetti
Vaishali Kankanala
Neeharika Dasari

Phase 1: Acquire Data

What Data?

NASA Kennedy Space Center Logs

Server access logs

Each file contains a single line per HTTP request

Host making the request
Timestamp with time zone offset
HTTP request
Requests' reply code
Bytes in the reply

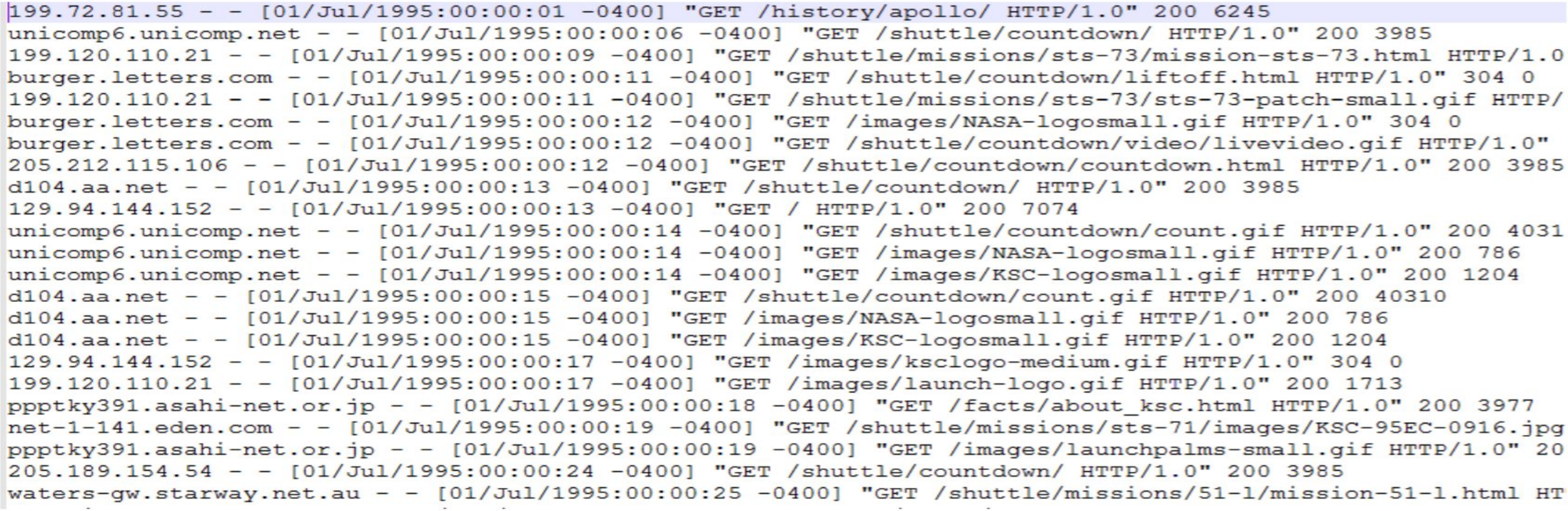
Size of the server log - July 1995

160 MB
1.5 Million records

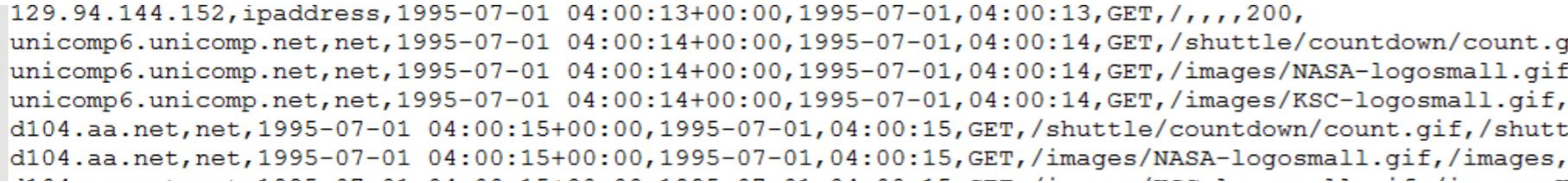
Configuration Setup

1. Python
2. MongoDB
3. Java Driver
4. Hadoop (pseudo-distributed mode)
5. MongoDB Connector for Hadoop
6. Pig Scripts
7. Pig JAR
8. R
9. Tableau

Phase 2: Preprocessing Server Logs

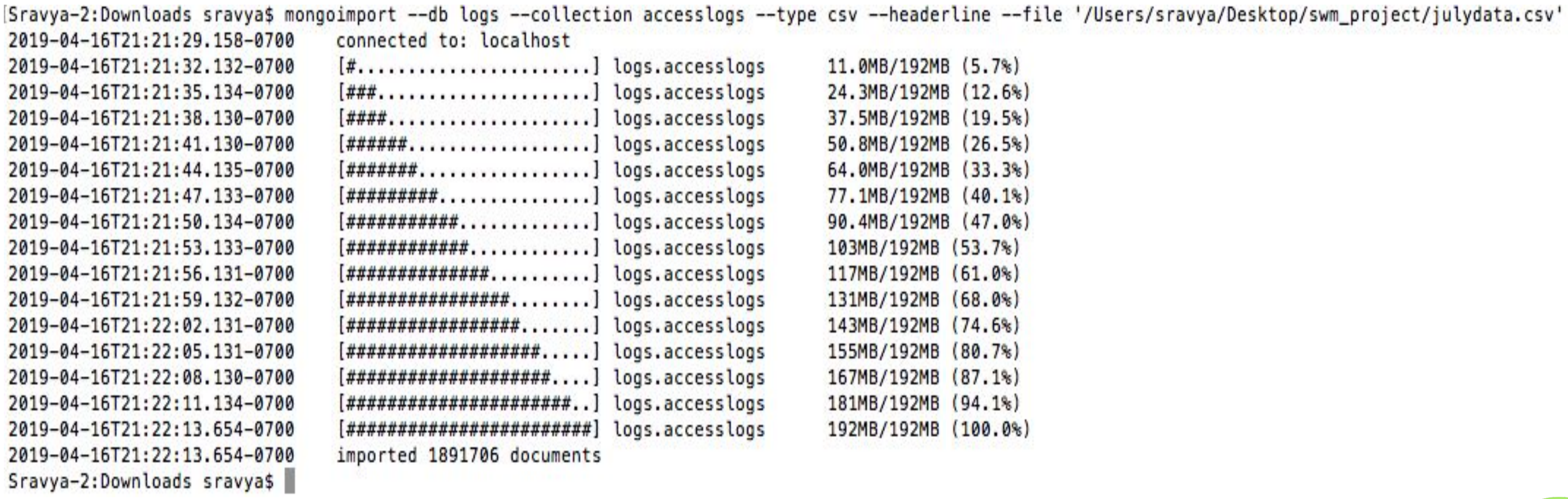


```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0
burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0"
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985
d104.aa.net - - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 4031
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:18 -0400] "GET /facts/about_ksc.html HTTP/1.0" 200 3977
net-1-141.eden.com - - [01/Jul/1995:00:00:19 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0916.jpg
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:19 -0400] "GET /images/launchpalms-small.gif HTTP/1.0" 20
205.189.154.54 - - [01/Jul/1995:00:00:24 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
waters-gw.starway.net.au - - [01/Jul/1995:00:00:25 -0400] "GET /shuttle/missions/51-l/mission-51-l.html HT
```



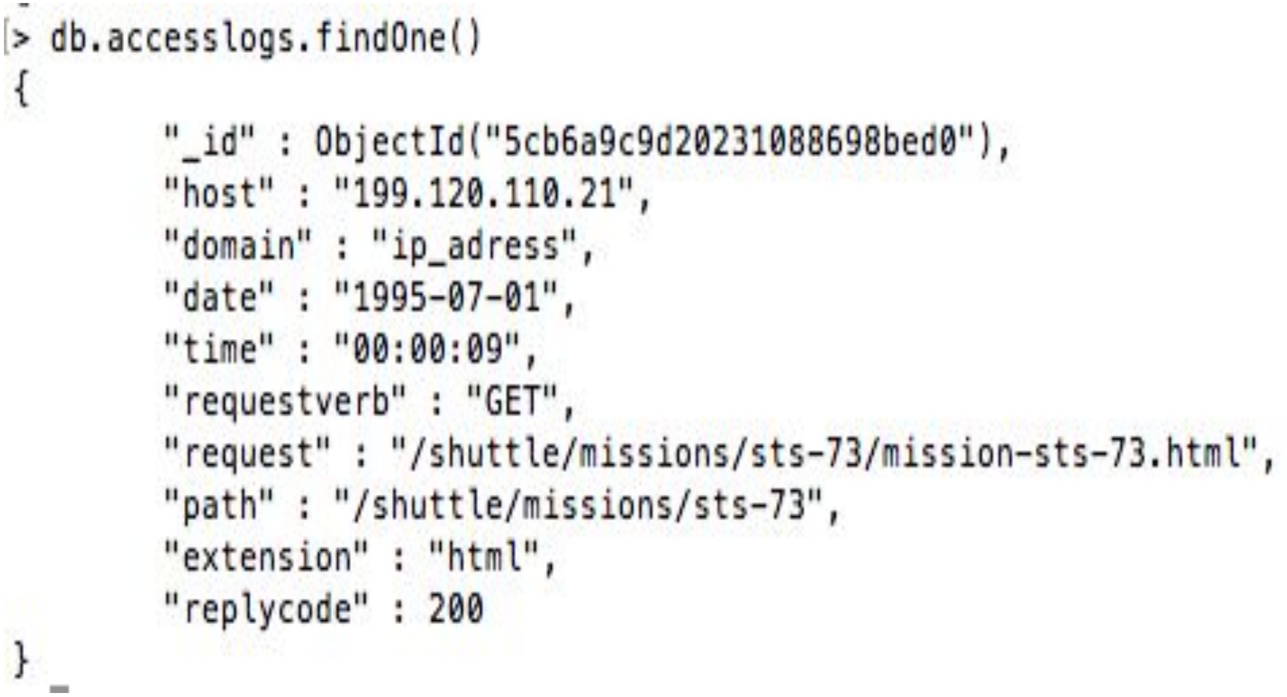
```
129.94.144.152,ipaddress,1995-07-01 04:00:13+00:00,1995-07-01,04:00:13,GET,/,,,,200,
unicomp6.unicomp.net,net,1995-07-01 04:00:14+00:00,1995-07-01,04:00:14,GET,/shuttle/countdown/count.g
unicomp6.unicomp.net,net,1995-07-01 04:00:14+00:00,1995-07-01,04:00:14,GET,/images/NASA-logosmall.gif
unicomp6.unicomp.net,net,1995-07-01 04:00:14+00:00,1995-07-01,04:00:14,GET,/images/KSC-logosmall.gif,
d104.aa.net,net,1995-07-01 04:00:15+00:00,1995-07-01,04:00:15,GET,/shuttle/countdown/count.gif,/shutt
d104.aa.net,net,1995-07-01 04:00:15+00:00,1995-07-01,04:00:15,GET,/images/NASA-logosmall.gif,/images,
```

Phase 3 : Processing Server Logs



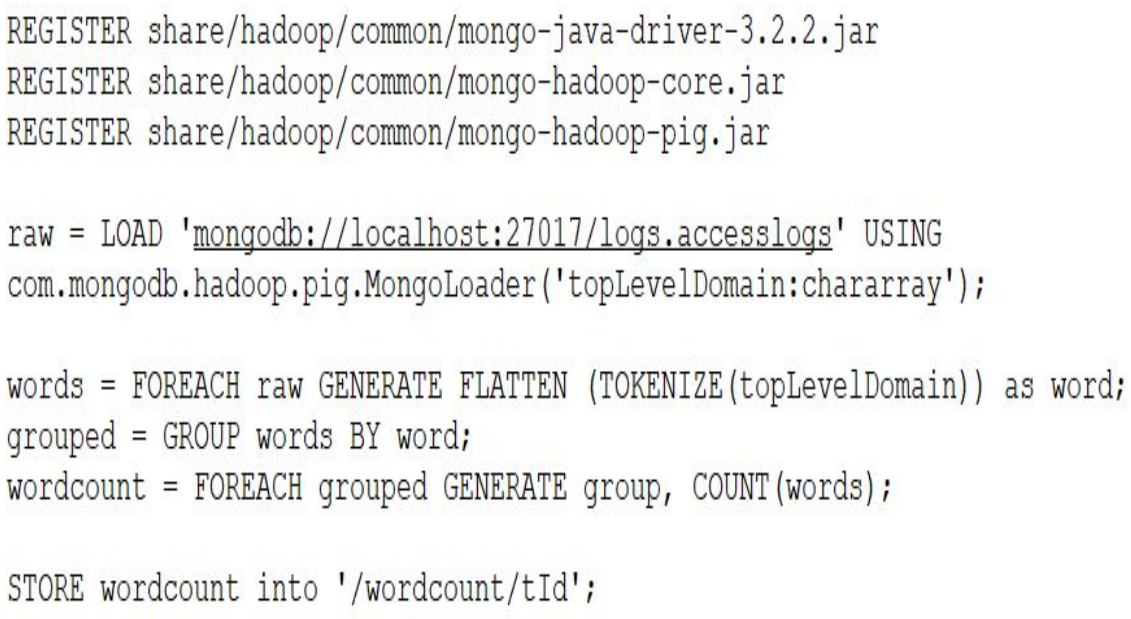
```
Sravya-2:Downloads sravya$ mongoimport --db logs --collection accesslogs --type csv --headerline --file '/Users/sravya/Desktop/swm_project/julydata.csv'
2019-04-16T21:21:29.158-0700    connected to: localhost
2019-04-16T21:21:32.132-0700    [#.......................] logs.accesslogs      11.0MB/192MB (5.7%)
2019-04-16T21:21:35.134-0700    [###.....................] logs.accesslogs      24.3MB/192MB (12.6%)
2019-04-16T21:21:38.130-0700    [####....................] logs.accesslogs      37.5MB/192MB (19.5%)
2019-04-16T21:21:41.130-0700    [######..................] logs.accesslogs      50.8MB/192MB (26.5%)
2019-04-16T21:21:44.135-0700    [#######.................] logs.accesslogs      64.0MB/192MB (33.3%)
2019-04-16T21:21:47.133-0700    [#########...............] logs.accesslogs      77.1MB/192MB (40.1%)
2019-04-16T21:21:50.134-0700    [###########.............] logs.accesslogs      90.4MB/192MB (47.0%)
2019-04-16T21:21:53.133-0700    [############............] logs.accesslogs      103MB/192MB (53.7%)
2019-04-16T21:21:56.131-0700    [##############..........] logs.accesslogs      117MB/192MB (61.0%)
2019-04-16T21:21:59.132-0700    [################........] logs.accesslogs      131MB/192MB (68.0%)
2019-04-16T21:22:02.131-0700    [#################.......] logs.accesslogs      143MB/192MB (74.6%)
2019-04-16T21:22:05.131-0700    [###################.....] logs.accesslogs      155MB/192MB (80.7%)
2019-04-16T21:22:08.130-0700    [####################....] logs.accesslogs      167MB/192MB (87.1%)
2019-04-16T21:22:11.134-0700    [######################..] logs.accesslogs      181MB/192MB (94.1%)
2019-04-16T21:22:13.654-0700    [########################] logs.accesslogs      192MB/192MB (100.0%)
2019-04-16T21:22:13.654-0700    imported 1891706 documents
Sravya-2:Downloads sravya$
```

Snapshot of a Document from MongoDB



```
> db.accesslogs.findOne()
{
        "_id" : ObjectId("5cb6a9c9d20231088698bed0"),
        "host" : "199.120.110.21",
        "domain" : "ip_adress",
        "date" : "1995-07-01",
        "time" : "00:00:09",
        "requestverb" : "GET",
        "request" : "/shuttle/missions/sts-73/mission-sts-73.html",
        "path" : "/shuttle/missions/sts-73",
        "extension" : "html",
        "replycode" : 200
}
```

Processing Server Logs

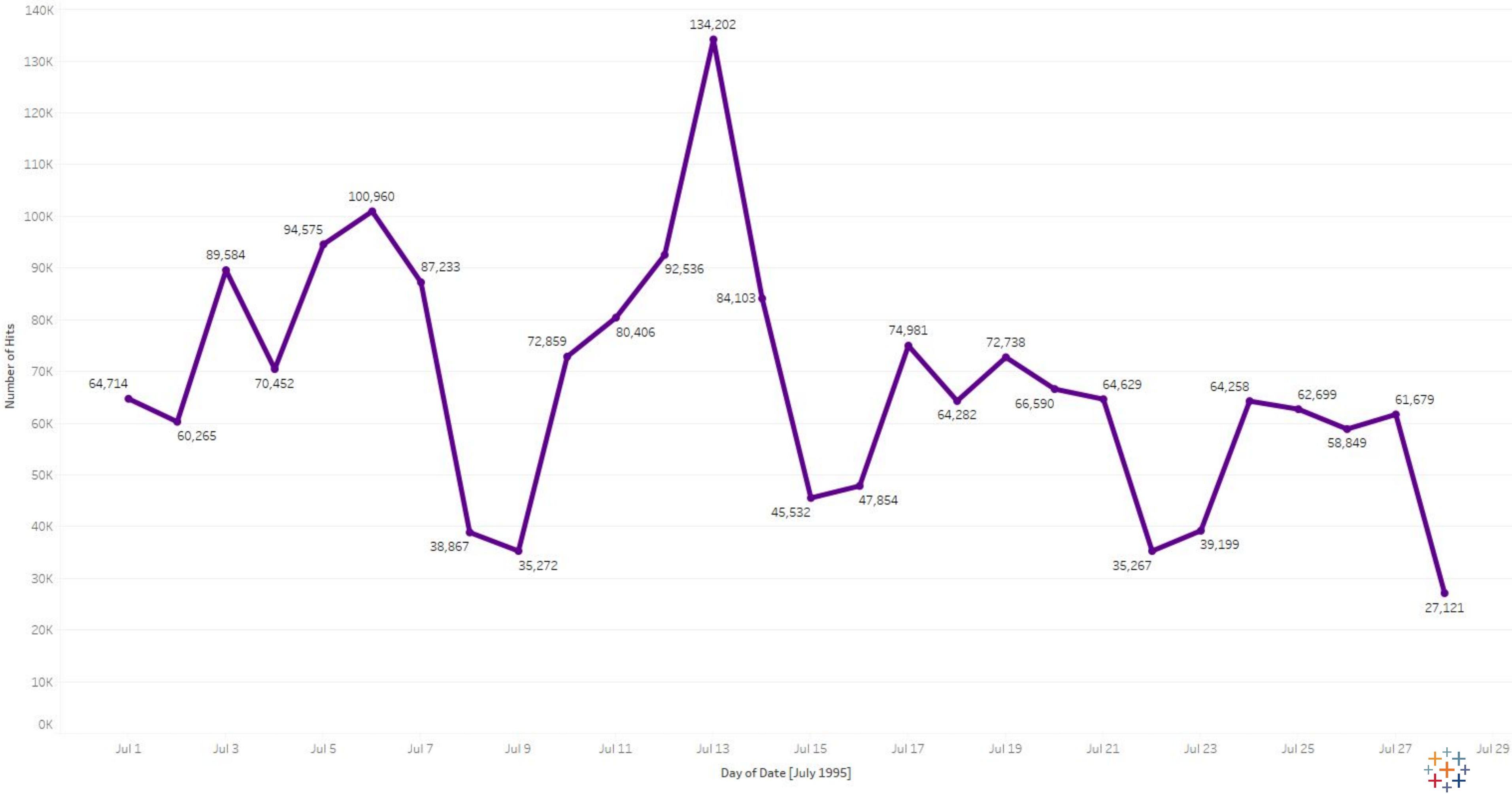


```
REGISTER share/hadoop/common/mongo-java-driver-3.2.2.jar
REGISTER share/hadoop/common/mongo-hadoop-core.jar
REGISTER share/hadoop/common/mongo-hadoop-pig.jar

raw = LOAD 'mongodb://localhost:27017/logs.accesslogs' USING
com.mongodb.hadoop.pig.MongoLoader('topLevelDomain:chararray');

words = FOREACH raw GENERATE FLATTEN (TOKENIZE(topLevelDomain)) as word;
grouped = GROUP words BY word;
wordcount = FOREACH grouped GENERATE group, COUNT(words);

STORE wordcount into '/wordcount/tId';
```

Pig Script:

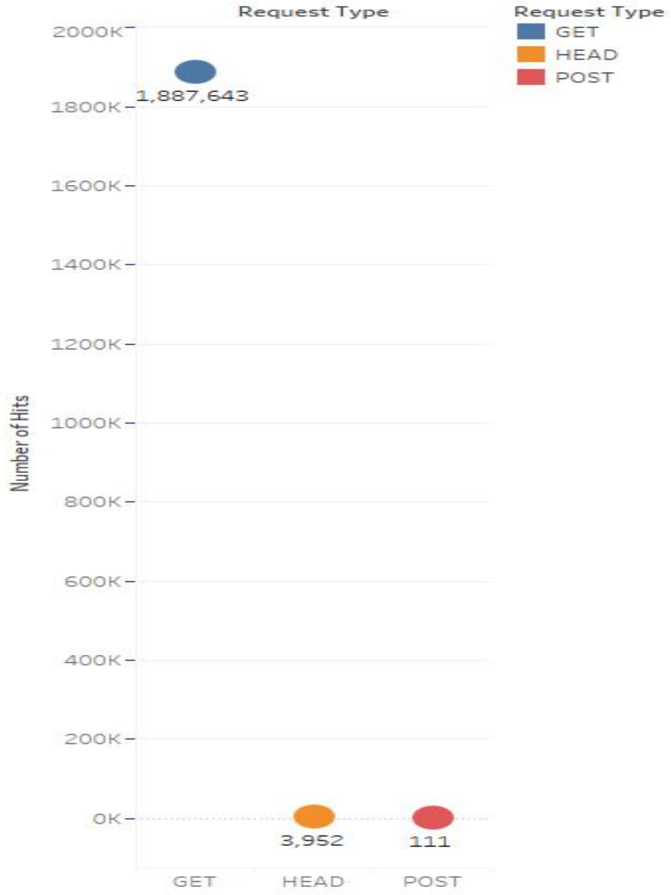
1. **Load data from MongoDB using MongoLoader**
2. **Tokenize collection imported**
3. **Group by and Count**
4. **Store output on HDFS**

Day Vs Number of Hits



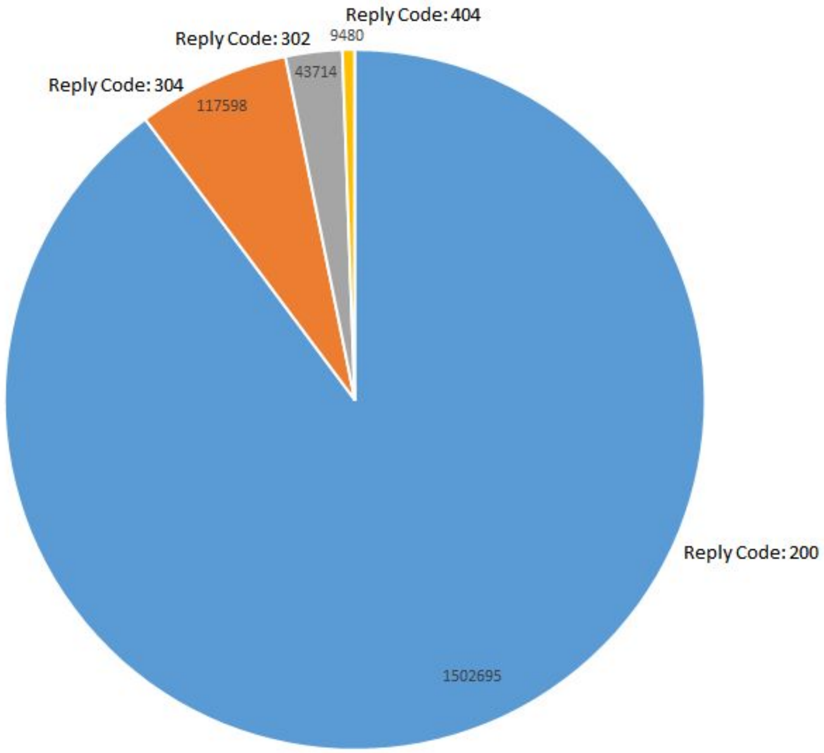
The trend of sum of Number of Hits for Date Day.

Request Type VS Number of Hits

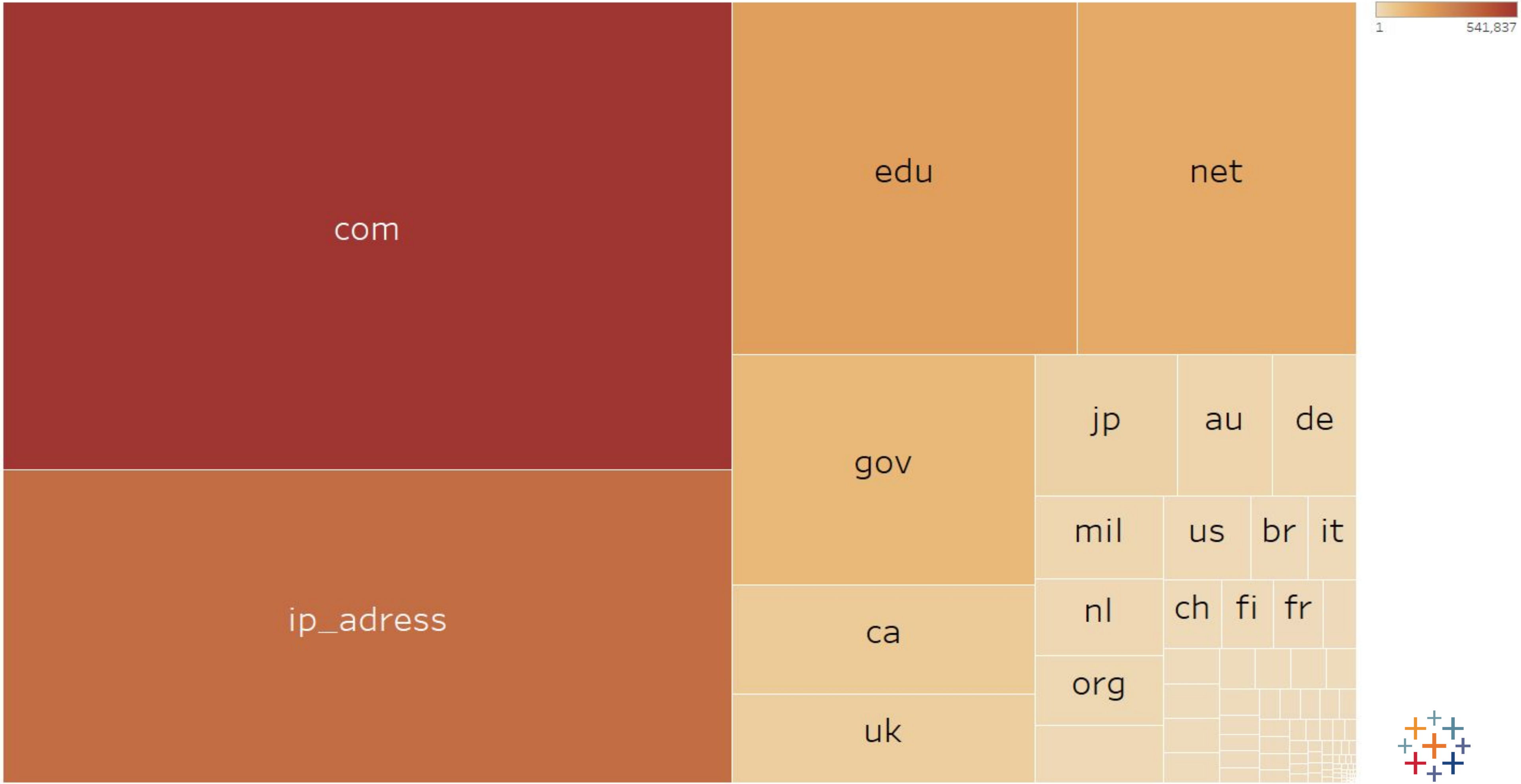


Sum of Number of Hits for each Request Type. Color shows details about Request Type.

Reply Codes Vs Count

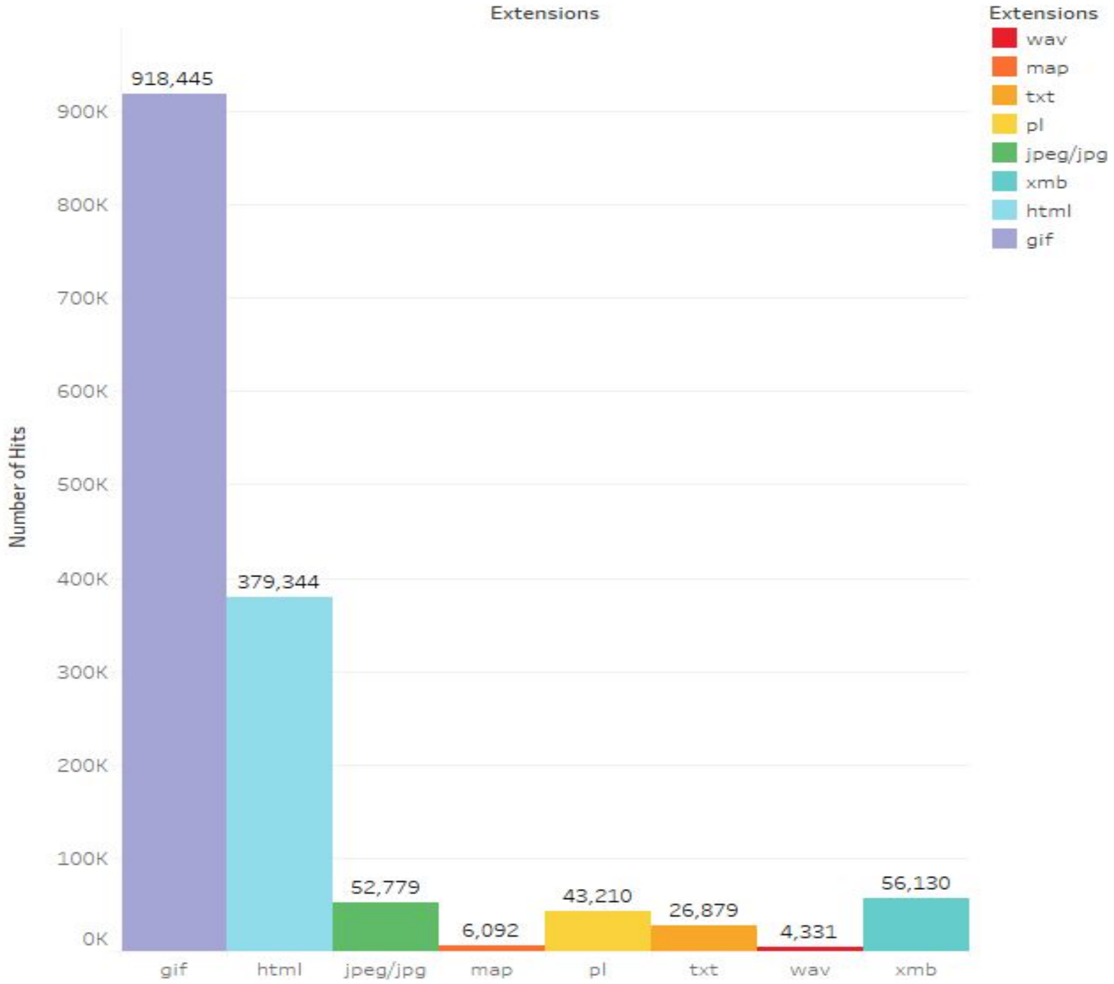


Domain Name Vs Number of Hits



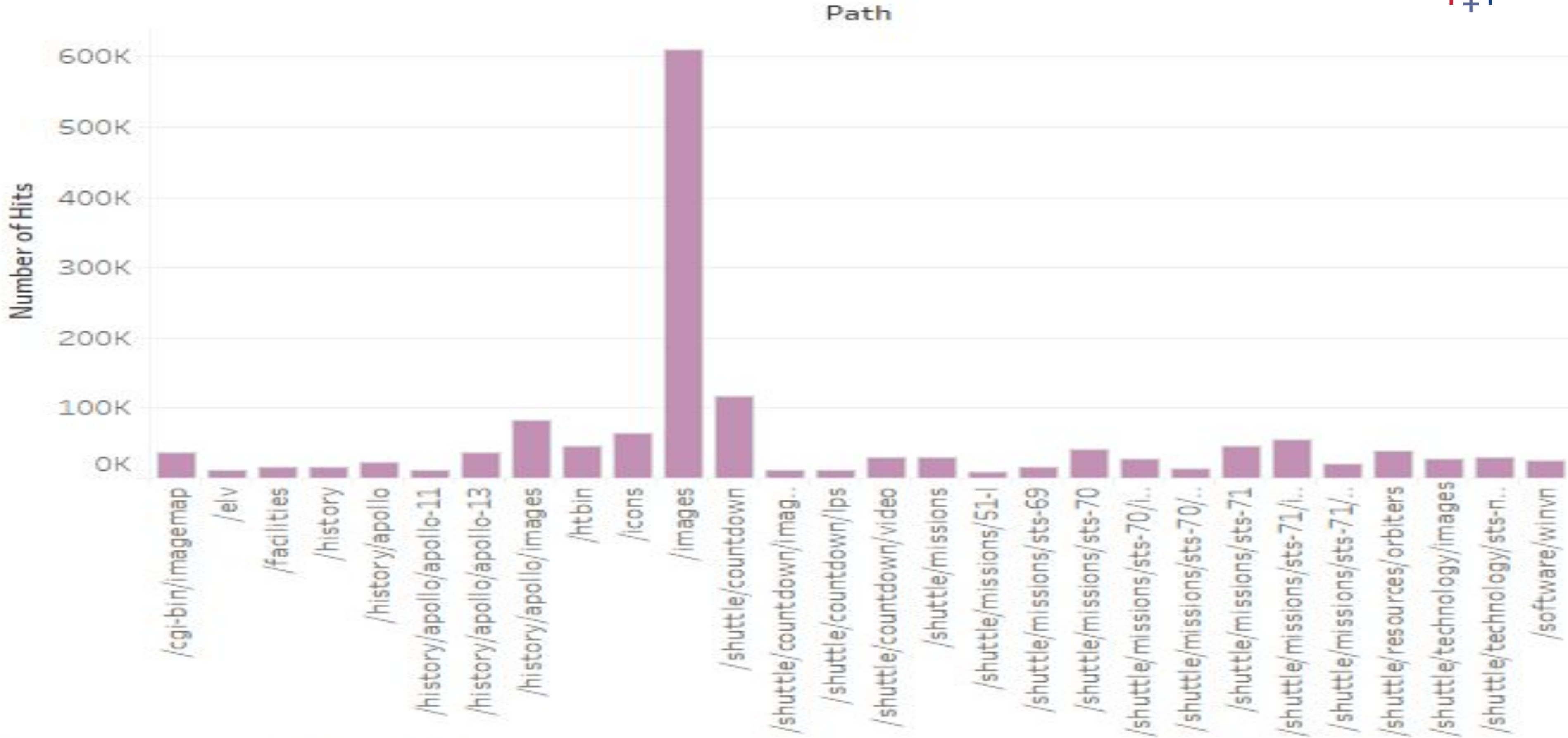
Domain Name. Color shows sum of Number of Hits. Size shows sum of Number of Hits. The marks are labeled by Domain Name. The view is filtered on Domain Name, which keeps 103 of 103 members.

Extensions vs Number of Hits



Sum of Number of Hits for each Extensions. Color shows details about Extensions as an attribute.

Path vs Number of Hits

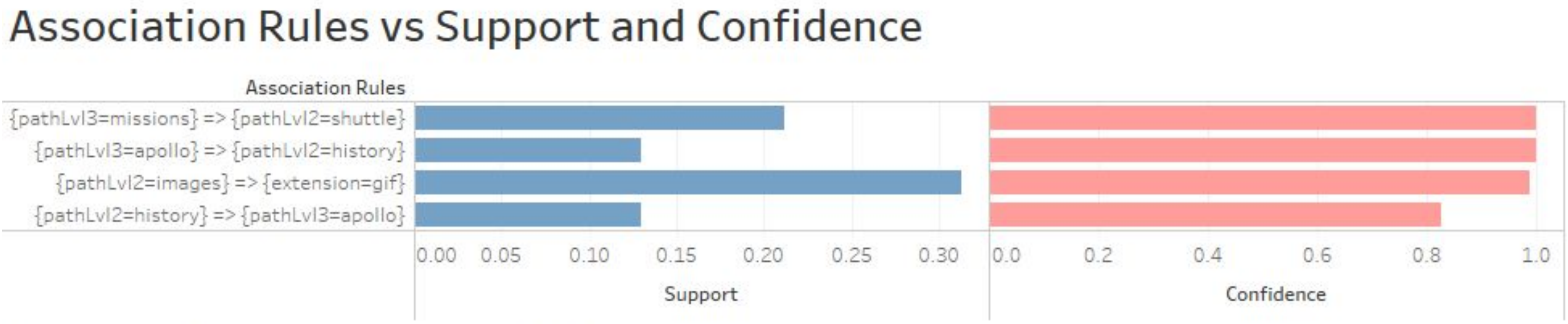


Sum of Number of Hits for each Path. Details are shown for Path. The view is filtered on Path, which excludes Null.

Phase 4: Data Mining of Server Logs (Apriori-1)

Apriori

Extract frequently co-accessed pages within a single request

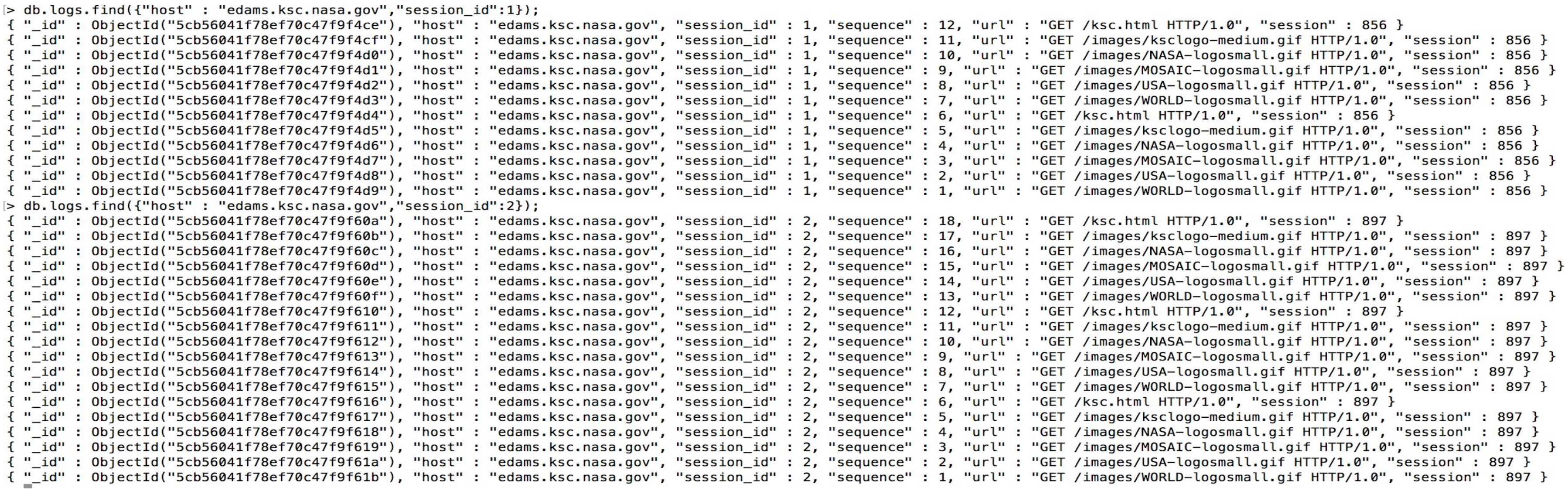


Sum of Support and sum of Confidence for each Association Rules.

Phase 4: Data Mining of Server Logs (Apriori-2)

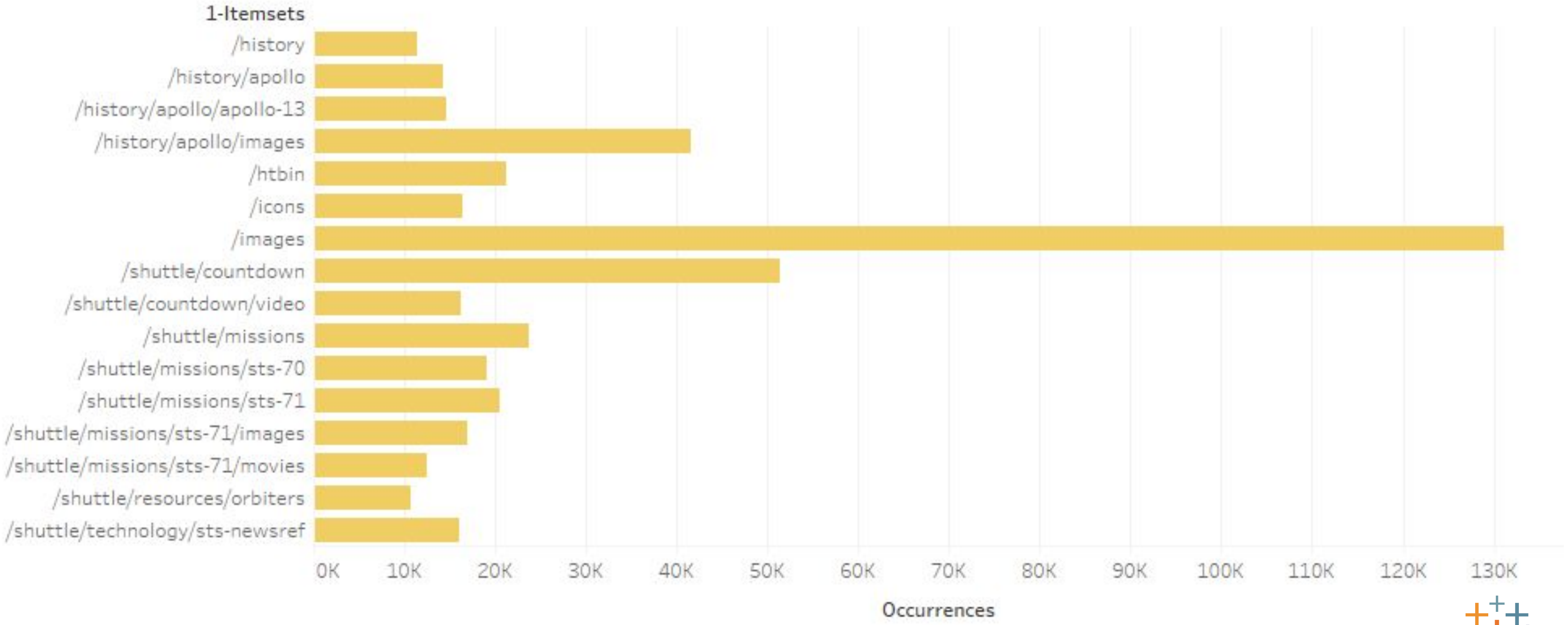
Session Identification:

- Default timeout 30 minutes
- If extends timeout user is assigned next session



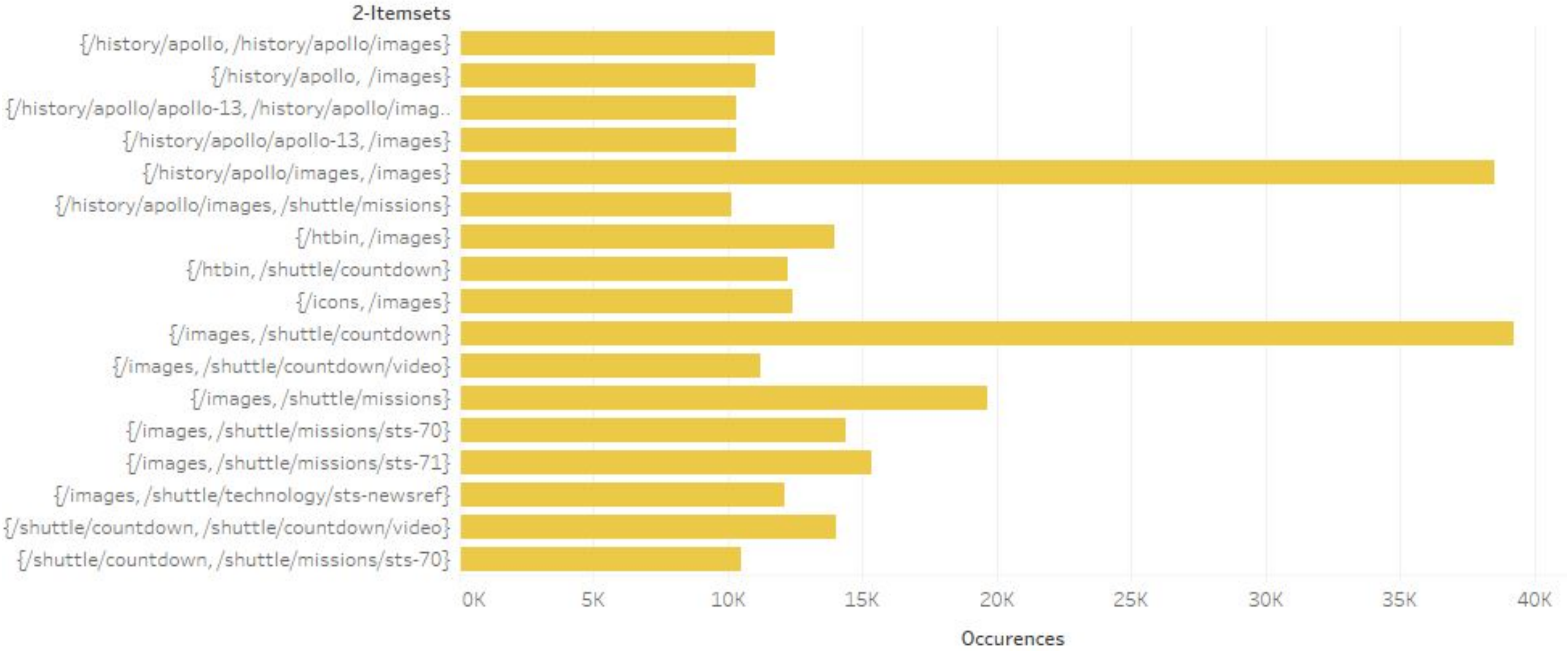
```
> db.logs.find({"host" : "edams.ksc.nasa.gov","session_id":1});
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4ce"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 12, "url" : "GET /ksc.html HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4cf"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 11, "url" : "GET /images/ksclogo-medium.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d0"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 10, "url" : "GET /images/NASA-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d1"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 9, "url" : "GET /images/MOSAIC-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d2"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 8, "url" : "GET /images/USA-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d3"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 7, "url" : "GET /images/WORLD-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d4"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 6, "url" : "GET /ksc.html HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d5"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 5, "url" : "GET /images/ksclogo-medium.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d6"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 4, "url" : "GET /images/NASA-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d7"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 3, "url" : "GET /images/MOSAIC-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d8"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 2, "url" : "GET /images/USA-logosmall.gif HTTP/1.0", "session" : 856 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f4d9"), "host" : "edams.ksc.nasa.gov", "session_id" : 1, "sequence" : 1, "url" : "GET /images/WORLD-logosmall.gif HTTP/1.0", "session" : 856 }
> db.logs.find({"host" : "edams.ksc.nasa.gov","session_id":2});
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60a"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 18, "url" : "GET /ksc.html HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60b"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 17, "url" : "GET /images/ksclogo-medium.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60c"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 16, "url" : "GET /images/NASA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60d"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 15, "url" : "GET /images/MOSAIC-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60e"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 14, "url" : "GET /images/USA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f60f"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 13, "url" : "GET /images/WORLD-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f610"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 12, "url" : "GET /ksc.html HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f611"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 11, "url" : "GET /images/ksclogo-medium.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f612"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 10, "url" : "GET /images/NASA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f613"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 9, "url" : "GET /images/MOSAIC-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f614"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 8, "url" : "GET /images/USA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f615"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 7, "url" : "GET /images/WORLD-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f616"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 6, "url" : "GET /ksc.html HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f617"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 5, "url" : "GET /images/ksclogo-medium.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f618"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 4, "url" : "GET /images/NASA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f619"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 3, "url" : "GET /images/MOSAIC-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f61a"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 2, "url" : "GET /images/USA-logosmall.gif HTTP/1.0", "session" : 897 }
{ "_id" : ObjectId("5cb56041f78ef70c47f9f61b"), "host" : "edams.ksc.nasa.gov", "session_id" : 2, "sequence" : 1, "url" : "GET /images/WORLD-logosmall.gif HTTP/1.0", "session" : 897 }
```

1-Itemsets



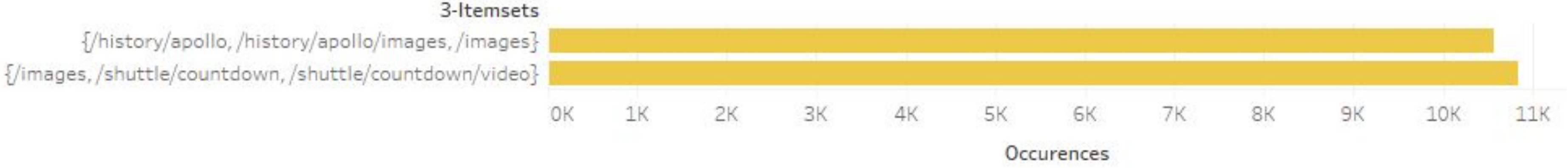
Sum of Occurrences for each 1-Itemsets.

2-Itemsets



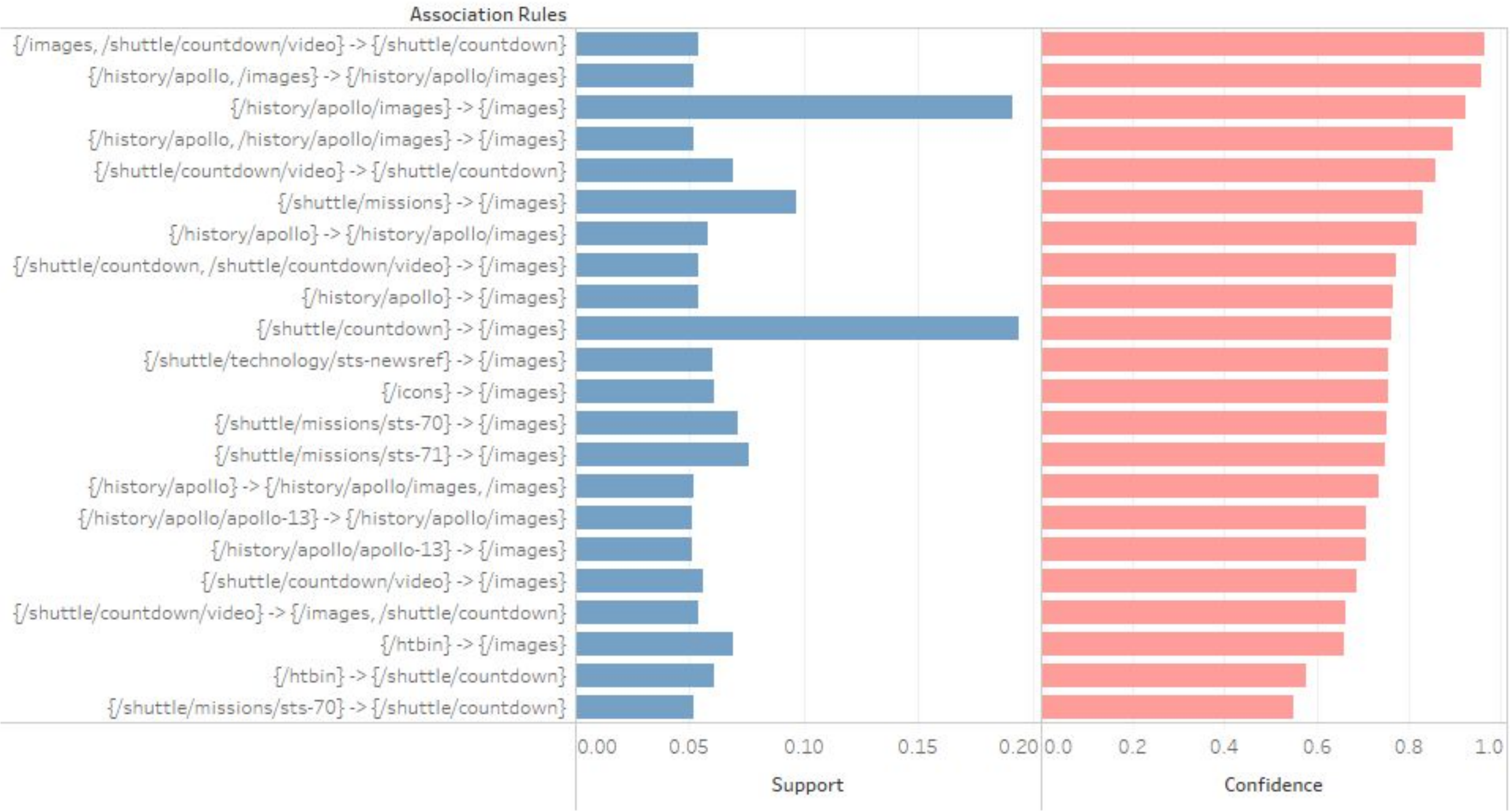
Sum of Occurences for each 2-Itemsets.

3-Itemsets



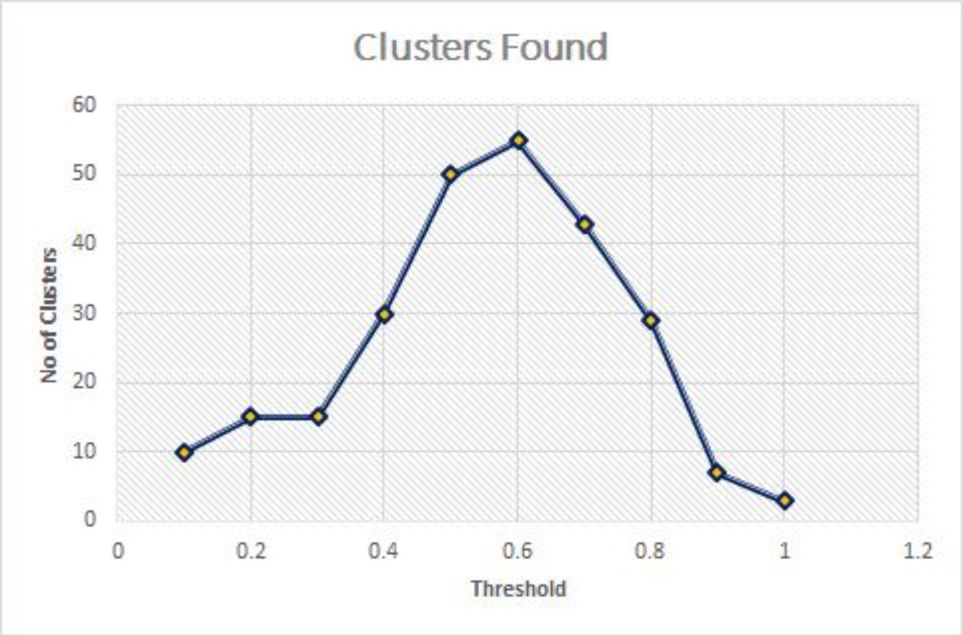
Sum of Occurences for each 3-Itemsets.

Association Rules vs Support and Confidence



Sum of Support and sum of Confidence for each Association Rules.

Phase 5: Clustering



$$C_{ij} = \alpha \times M_{ij} + (1 - \alpha) \times P_{ij}$$

M - Page Co-Occurrence matrix
P - Path Similarity Matrix

Challenges

1. Web datasets can be very large
2. It can be difficult to maintain in database.
3. Proper configuration setup to mine multi terabyte data sets.
4. Cached page views are not recorded
5. Proxy Servers
6. Use of cookies or Dynamic URL's

References

[1] Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data." *ACM SIGKDD Explorations Newsletter* 1.2 (2000): 12-23.

[2] Asadi, Tawfiq & Obaid, Ahmed. (2016). "An efficient web usage mining algorithm based on log file data." 92. 215-224.

[3] Savio, Marc Nipuna Dominic, "Predicting User's Future Requests Using Frequent Patterns" (2016). Master's Projects. 501. http://scholarworks.sjsu.edu/etd_projects/501.

[4] Mobasher, Bamshad, Robert Cooley, and Jaideep Srivastava. "Automatic personalization based on web usage mining." *Communications of the ACM* 43.8 (2000): 142-151.

[5] http://www.sthda.com/english/wiki/descriptive-statistics-and-graphics#check-your-data

[6] https://onlinehelp.tableau.com/current/desktopdeploy/en-us/desktop_deploy_download_and_install.htm

[7] https://www.r-bloggers.com/log-file-analysis-with-r/

[8] http://eecs.csuohio.edu/~sschung/cis612/CIS612_PDF_Presentation_NASA_Halley_Orogvany.pdf

GitHub Repository Link

https://github.com/ManasaPola/WebMining_WebUsageMining.git

Thank you