TEM Journal. Volume 7, Issue 3, Pages 645-650, ISSN 2217-8309, DOI: 10.18421/TEM73-22, August 2018.

Analysis of Apache Logs Using Hadoop and Hive

Aleksandar Velinov [1], Zoran Zdravev [1]

[1]Goce Delcev University, Faculty of Computer Science, Krste Misirkov 10-A, Stip, Republic of Macedonia

***Abstract* – In this paper we consider an analysis of Apache web logs using Cloudera Hadoop distribution and Hive for querying the data in the web logs. We used public available web logs from NASA Kennedy Space Center server. HDFS (Hadoop distributed file system) was used as a logs container. The apache web logs were copied to the HDFS from the local file system. We made an analysis for the total number of hits, unique IPs, the most common hosts that made request to the NASA server in Florida, the most common types of errors. We also examined the ratio between the number of rows in the logs and the time of execution.**

***Keywords* – Logs, Hadoop, Hive, analysis.**

1. Introduction

The word „data“ represents a collection of information in either an organized or unorganized format. Organized data refers to data that are sorted into a row/column structure. These data are also called structured data. Every column represents a characteristic of the data and it must be strictly defined in terms of row name and row type (alpha, numeric, date). Every row represents a single observation [1]. Relational databases (RDBMS) and spreadsheets are often used to store these types of data. Structured data can easily be ordered, processed and analyzed by data mining tools. It can be searched using standard search algorithms and manipulated in well-defined ways [2]. There are a lot of such useful tools. Unorganized or unstructured data are data that is in a free form such as text or raw audio/signals [1]. These data do not have a pre-defined data model and must be parsed further to become organized or to be in structured format. It is difficult to analyze this type of data. Besides the types of data explained by Ozdemir [1], there are also semi-structured data. Semi-structured data (such as what you might see in log files) are a bit more difficult to understand than the structured data. Normally, this kind of data is stored in the form of text files, where there is some degree of order - for example, tab delimited files, where columns are separated by a tab character. So, instead of being able to issue a database query for a certain column and knowing exactly what you are getting back, users typically need to explicitly assign data types to any data elements extracted from semi-structured data sets [2]. For our experimental analysis, we used semi-structured data or logs.

The science that deals with the study and analysis of data is called “Data science”. Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems [3]. The data are used as a basis for analysis and obtaining conclusions. Algorithms are used as a way to quickly and more easily obtain the necessary analyzes. The technology is used to implement the corresponding data analysis algorithms.

One of the basic goals of the Data Science is turning the data into business value. As can be seen from Figure 1., this process for getting business value from data starts from the place where the data are stored (Data warehouses).

Data from warehouses are used for quantitative data analysis (Discovery of Data Insight). This aspect of data science is used for uncovering findings from data [3]. That can help companies to make better decisions in the future. For example, film companies using the data can understand what drives user interest and to make decision which film to be broadcast. Website administrators can use the log data to understand the most common errors on a

DOI: 10.18421/TEM73-22
https://dx.doi.org/10.18421/TEM73-22

Corresponding author: Aleksandar Velinov,
Goce Delcev University, Faculty of Computer Science, Krste Misirkov 10-A, Stip, Republic of Macedonia
Email: aleksandar.velinov@ugd.edu.mk

Received: 09 March 2018.
Accepted: 29 June 2018.
Published: 27 August 2018.

page. By revealing the cause of the errors, in the future, the availability of the website can be improved. Production companies can analyze client orders data in order to optimally plan the production in the future. The conclusion from this would be that the analysis of the data brings us a better future. Data scientists use data explorations to mine out insights. When given a challenging question, data scientists become detectives. They investigate leads and try to understand pattern or characteristics within the data. This requires a big dose of analytical creativity [3]. The data scientists can also apply quantitative techniques and to get a better insights such as segmentation analysis, synthetic control experiments, time series forecasting, inferential models, etc. The goal is to make a forensic view of the data and to get better understandings.

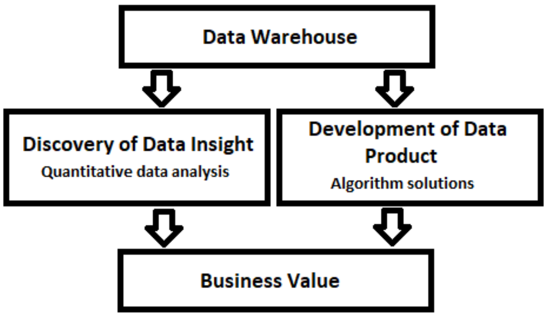
Data Warehouse
Discovery of Data Insight
Quantitative data analysis
Development of Data Product
Algorithm solutions
Business Value

Figure 1. Turn data into Business Value

Other way in turning data into business value is Development of data products (algorithm solutions in productions). A "data product" is a technical asset that utilizes data as input, and processes that data to return algorithmically-generated results [3]. As we have said before, the data are input also in this step. Algorithms are used to generate results that can help companies in their work, and work processes. The example for data product is recommendation engine that makes recommendations using the data as input. Some engines suggest items to buy (Recommendation engine of Amazon). Spam filter is a data product that uses an algorithm behind the scene that processes the incoming mail and decides whether the message is spam or not (Spam filter of Gmail). Self-driving cars use a computer vision. This is also a data product. Machine learning algorithms are used for recognition of traffic lights, pedestrians, other cars on the roads, etc. This differs from the first approach (Data insights) where the results from the analysis are used to make better business decisions. The data products include an algorithm that is integrated into core applications and automatically contributes to the improvement of the decisions. Data products are developed, tested and incorporated into production systems by data scientists. Using these two processes (Discovery of Data Insight and Development of Data Products), the data are turned into business value.

According to the special report, every year attempts are periodically made to estimate how much data are generated worldwide, and in what form [4]. According the IDC and EMC, in 2013 there were 4.4 zettabytes (ZB) data - that is 4.4 trillion gigabytes, and predicted this would grow to 44ZB (44 trillion gigabytes) in 2020, more than doubling every two year [5]. The latest estimate is made by IDC and Seagate (Data Age 2015 report) [6]. The estimate is that the number of data in 2025 will be 163ZB - a tenfold rise from the 16.1ZB created in 2016 [5]. This introduces the concept of big data.

There are many definitions for the concept of Big Data. Gollapudi says that Big Data can be defined as an environment comprising of tools, processes, and procedures that fosters discovery with data at its center. This discovery process refers to our ability to derive business value from data and includes collecting, manipulating, analyzing, and managing data [7]. Other analysts use other definitions. Most of them use the 3V model to define Big Data. The three Vs stand for volume, velocity and variety. Volume refers to the fact that Big Data involves huge amounts of data. Velocity reflects the speed at which this data is generated, processed and analyzed. One of the key factors in a Big Data World is speed. Analyzing historical data is the focus of traditional analytics. Real-time analytics extends this concept and includes in-flight transitory data. These data can come from commercial systems in companies, from websites, online generated forms etc. Variety describes the fact that Big Data can come from many different sources, in various formats and structures [8]. Social media sites and sensor networks generate a stream of data and text data (geographical information, images, videos and audio).

Our goal in this research is to make a statistical analysis of web logs and to inspect the ratio between the number of rows in the logs and execution time. To do this, we used special tools for big data processing and analyzing.

In Section 2 of this paper, presented is the related work for statistical analysis of web server logs. In Section 3, Hadoop is presented as a platform for processing and analyzing big data. This section also explains Hive as a tool for querying, analyzing and summarizing the data in the HDFS (Hadoop Distributed File System). The structure of Apache web server logs is presented in Section 4. Section 5 represents the methodology and technology used to analyze web logs. The experimental results are shown in Section 6. Section 7 represents the ratio between the number of log rows and the execution time.

2. Related work

Harish and Kavitha in their paper represent a statistical analysis of web server logs using Apache Hive in Hadoop Framework [9]. They propose a methodology which uses the HDFS (Hadoop Distributed File System) as a logs storage and Apache Hive (HiveQL) for processing and analyzing the logs data. They applied the Hadoop MapReduce programming model for analyzing web server log files. Hadoop MapReduce framework provides parallel distributed processing and reliable data storage for large volumes of log files [9]. Their methodology includes multiple steps: Loading log files from Local Storage to HDFS, Creating a Hive table to store raw log files and moving raw log files to hive table, Preprocessing the raw log files using HiveQL, Preprocessed data is stored in HDFS, Statistical Analysis using HiveQL and Visualizing the results using Jaspersoft iReport.

Lavanya and Srinivasa explored the customer behavior using analysis of web server logs using Hive in Hadoop Framework [10]. They proposed the same methodology as the Harish and Kavitha [9]. Web access logs taken from web page were used for analysis. Using the report from the analysis business group can show what parts of the site need to be enhanced (for example the errors of the web site), who are the potential clients, which product is the most purchased, periods of the year when it has the most sales etc.

Gavandi, Gori, Ingawale and Yadav also used Hadoop for web server log processing [11]. They used a different methodology compared to the previous two approaches. They used FLUME framework to transfer the data from the local file system to the HDFS. FLUME has agents running on Web servers. Flume Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tunable reliability mechanisms for failover and recovery [12]. This log data is processed by MapReduce to produce Comma Separated Values i.e. CSV [11]. Finally, they used Microsoft Excel to produce statistical information and generate reports.

3. Hadoop Distribution and Hive

Apache Hadoop was developed to process and analyze big data. It is an open source platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware (everyday computer hardware that is affordable and easily available, and running applications against that data. A group of computers (nodes) that can work together on the same problem, form cluster. Hadoop cluster typically consists of a few master nodes (Namenode), which control the storage and processing systems in Hadoop, and many slave nodes (Datanode), which store all the cluster's data and is also where the data gets processed. Hadoop uses network of affordable computer resources to turn the data into business value (get business insight). Hadoop is simply the name that the son of Doug Cutting gave to his stuffed elephant. Doug Cutting is co-creator of Hadoop. He named this framework as Hadoop because the name is unique and easy to remember [2]. This framework was developed after the discovery of Google MapReduce model and Google File System. It is well-adopted, standards-based and as we already say it is open source what makes it accessible to everyone.

Hadoop consists of two main components: a distributed processing framework named MapReduce and HDFS (Hadoop distributed file system). An application that is running on Hadoop gets its work divided among the nodes (machines) in the cluster, and HDFS stores the data that will be processed [2]. MapReduce is supported by a component called YARN. MapReduce enables the development of a wide variety of applications which processes vast amounts of data in-parallel on large clusters (thousands of nodes). A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks [13]. MapReduce uses a single master named as JobTracker and one slave TaskTracker per cluster-node. Jobs component tasks are scheduled on the slaves by the master. The master is also responsible for monitoring and re-executing the failed tasks. The tasks are executed by the slaves which are managed by the master. MapReduce Framework uses <key, value> pairs as input and produces <key, value> pairs as output of the job. The key and value classes have to be serializable by the framework and hence need to implement the Writable interface. The key classes have to implement the WritableComparable interface to facilitate sorting by the framework [13].

The types of input and output of a MapReduce job are as follows:

(input) <k1, v1> -> map -> <k2, v2> -> combine -> <k2, v2> -> reduce -> <k3, v3> (output)

Apache Hive is a data warehouse system for Hadoop. Using Hive we can query, analyze and summarize the data in the HDFS. Hive uses HiveQL as a query language which is similar to SQL. Hive is used to convert unstructured data into structured format. After that we can use HiveQL to query the data. Hive also supports serializer/deserializers (SerDe) for complex or irregular structured data. With Hive we can create internal and external tables. Internal tables are used to store data in the Hive data warehouse. External tables are used to store data outside the data warehouse. Here is an example which we used in our research for creating external table from logs stored in HDFS:

```
CREATE EXTERNAL TABLE apachelog (
HOST STRING,
IDENTITY STRING,
USER STRING,
TIME STRING,
request STRING,
status 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
LOCATION '/user/cloudera/NASA_logs'
```

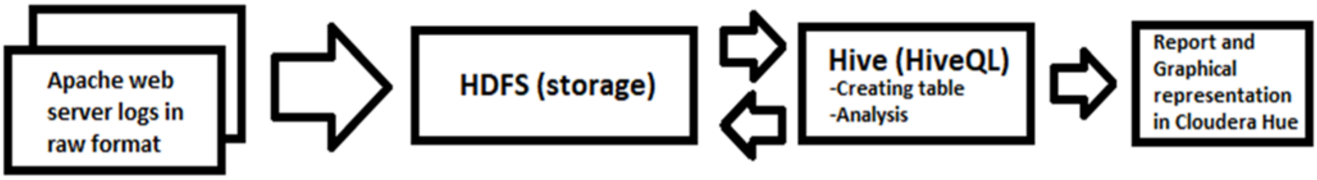
Apache web server logs in raw format
HDFS (storage)
Hive (HiveQL)
-Creating table
-Analysis
Report and Graphical representation in Cloudera Hue

Figure 2. Log data processing and graphical representation

4. Apache web logs

In our research for analysis we used Apache web server logs. The Apache access logs stores information about events that occurred on your Apache web server. Logs record the information about visitor IP address, what pages they are viewing, status codes etc. The common Apache access log format is as follows:

LogFormat "%h %l %u %t \"%r\" %>s %b" common

The sections in the log are the following [14]:

- %h – The IP address of the client.
- %l – The identity of the client determined by identd on the client's machine. Will return a hyphen (-) if this information is not available.
- %u – The userid of the client if the request was authenticated.
- %t – The time that the request was received.
- \"%r\" – The request line that includes the HTTP method used, the requested resource path, and the HTTP protocol that the client used.
- %>s – The status code that the server sends back to the client.
- %b – The size of the object requested.

One row of log is similar to the following:

127.0.0.1 - aleksandar [13/Feb/2017:10:21:00 -0700] "GET /sample-image.png HTTP" 200 2489

We used public available logs from NASA [15]. These traces contain two months data of all HTTP requests to the NASA Kennedy Space Center WWW server in Florida.

5. Used technology and methodology

In our research we used Cent OS 6 Linux Distribution and Cloudera Hue framework which contains all the tools for working with big data. Cloudera provides a scalable, flexible, integrated platform that makes it easy to manage rapidly increasing volumes and varieties of data. Cloudera products enable you to manage Apache Hadoop framework and Hadoop projects. It also helps on data scientists to store, analyze and process data. Cloudera has multiple tools for data analysis such as: CDH (Cloudera distribution of Apache Hadoop and other related open-source projects including Apache Impala, Cloudera Search, Hive, Pig etc.), Cloudera Manager and Cloudera Navigator. In our research we used CDH (HDFS), Hue and Hive as a query framework for creating table with data from the semi-structured logs.

As we can see from Figure 2., firstly we copy the data from the local file system to the HDFS. After storing the data in the HDFS we used Hive to create structured data from the raw log data that we stored in HDFS. Using special HiveQL query we created external table which contains all the data from the log files in structured format.

After creating a table, we executed some queries to turn data into values. At the end we represented the results from analysis with graphics using Cloudera Hue.

There are a lot of log analysers that we can find on Internet such as: Deep Log Analyzer, GoAccess, Logmatic. Io etc. For big data it is recommended to use Hadoop HDFS and query tools (like Hive, Pig and Impala) to analyze the data.

6. Experimental results

Firstly, we found the total number of successful hits. According to the related work in Part 2 of this paper, preprocessing is preliminary phase of the log analysis. In our research, we did not make preprocessing and used a Hive query with WHERE clause to select only the successful hits. The total number of rows in the Hive table is 3461612. The number of hits is 771203. Table 1. and Figure 3. show the most common status codes. Code 200 shows that users have successful access to the webpage. Other codes are types of errors, such as client errors (4XX) and server errors (5XX) and redirection (3XX). We can see that the status code 200 is the most common, than the status code 302 and 302 (redirection). The most common types of errors are client errors (404).

The total number of unique IPs that accessed the page is 40297. From Table 2. we can see 10 unique IP addresses sorted by the number of accesses. The total number of hits in July is 374906. The total number of hits in August is 396297.

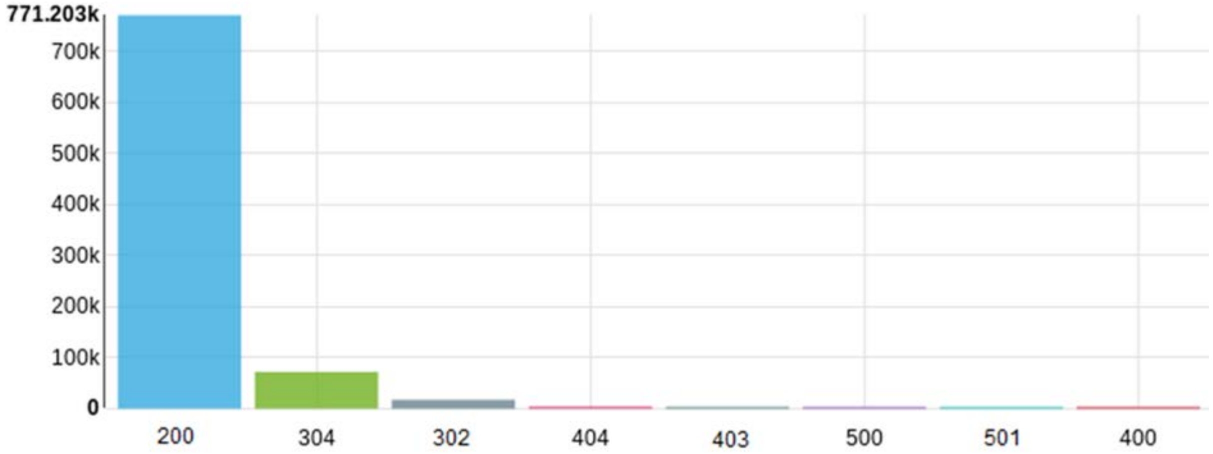

Figure 3. The most common status codes

Table 1. Status and number of hosts with status

	status	num_hosts_status
1	200	771203
2	304	72283
3	302	17714
4	404	4290
5	403	54
6	500	53
7	501	19
8	400	10

Table 2. Ten unique IPs sorted by the number of accesses

	host	cnt
1	163.206.89.4	9697
2	198.133.29.18	4204
3	163.205.156.16	2963
4	163.206.137.21	2787
5	163.205.1.45	2017
6	128.159.122.180	1680
7	128.217.62.1	1598
8	163.205.11.31	1549
9	163.205.1.19	1419
10	128.159.122.110	1348

7. Execution time vs Number of Log Rows

In Figure 4., we can see the execution time when creating a table from logs, depending on the number of rows in the logs. We did this research using logs with different number of rows (1000-1000000 rows). As we can see from Figure 4., the execution time (computer time) increases with increasing the number of rows as exponential function.

In Figure 5., we can see the dependence of the execution time on the number of rows in the logs when performing a query to find the number of hits on website. Execution time increases by increasing the number of rows. The execution time if we have 1000 rows in the logs is 96.945 seconds, while the execution time if we have 1000000 rows is 145.605 seconds.

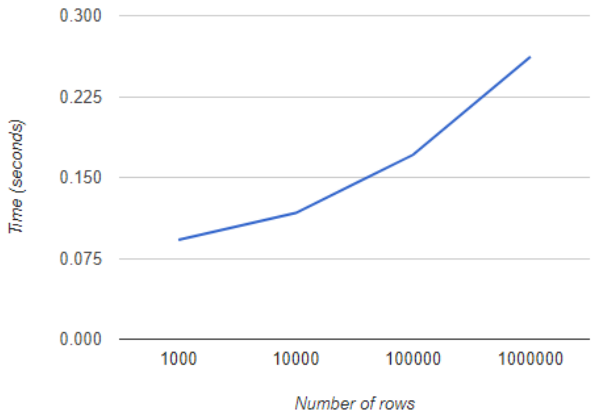

Figure 4. Execution time when creating a table from logs

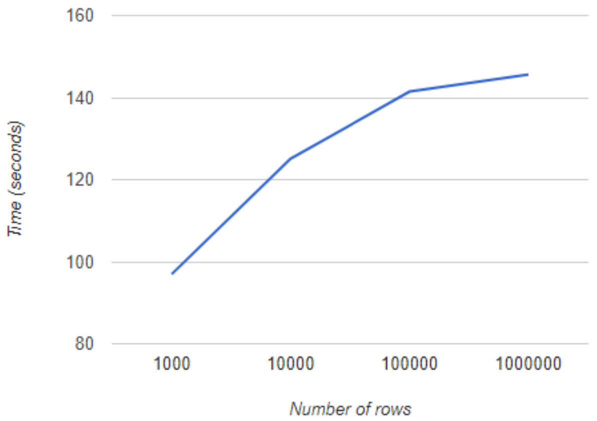

Figure 5. Time of execution (query to find the number of hits)

8. Further work

In the future, we plan to analyze IIS (Internet Information Services) logs and get a report from that analysis. IIS logs have different structure compared to Apache common log format. We saved the logs from our e-learning platform Moodle for a period of one year. These logs will be used for our further research.

9. Conclusion

The number of data is growing daily and the concept of big data is more present today. The main goal of the concept of big data is to make an analysis of the data in order to extract some business value. In this research, we made an analysis of Apache web logs. Using the logs we got some statistical analysis, such as user behaviour, number of hits, the most common types of errors, unique IPs etc. The execution time (computer time) increases with increasing the number of inspected log rows as exponential function.

References

[1]. Ozdemir, S. (2016). *Principles of Data Science*. Packt Publishing Ltd.

[2]. DeRoos, D., Zikopoulos, P., Brown, B., Coss, R., & Melnyk, R. B. (2014). *Hadoop for dummies*. John Wiley & Sons, Incorporated.

[3]. Lo, F. (n.d). *Data Jobs. What is Data Science*. Retrieved from: https://datajobs.com/what-is-data-science

[4]. Mclellan, C. (2017, September). *Turning big data into business insights*. Retrieved from: http://theadhikaris.info/wp-content/uploads/2017/09/Turning-big-data-into-business-insights.pdf

[5]. EMC Digital Universe with Research&Analysis by IDC. (n.d). *Executive Summary, Data Growth, Business Opportunities,and the IT Imperative*. Retrieved from: https://www.emc.com/leadership/digitaluniverse/2014iview/executive-summary.htm

[6]. Seagate. (n.d). *New Approaches for A New Data Age.* Retrieved from: *https://www.seagate.com/gb/en/our-story/data-age-2025/*

[7]. Gollapudi, S. (2013). *Getting started with Greenplum for big data analytics*. Packt Publishing Ltd.

[8]. Fujitsu. (2012). *Big Data, The definitive guide to the revolution in business analytics*. Retrieved from: https://www.fujitsu.com/rs/Images/WhiteBookofBigData.pdf

[9]. Harish, S., Kavitha, G. (2015). Statistical Analysis of Web Server Logs Using Apache Hive in Hadoop Framework. *IJIRCCE, International Journal of Innovative Research in Computer and Communication Engineering*, *3*(5), 4510-4515.

[10]. Lavanya, K. S., & Srinivasa, R. (2016).Customer behavior analysis of web server logs using Hive in Hadoop framework. *Int. J. Adv. Netw. Appl.(IJANA)*, 409-412.

[11]. Gavandi, P., Gori, B., Ingawle, S., & Yadav, S. (2016). Web ServerLogProcessing using Hadoop, 1st International Conference on Research, Enhancement & Advancements in Technology and Engineering.

[12]. Hortonworks. (n.d) *Apache Flume*. Retrieved from: https://hortonworks.com/apache/flume/

[13]. Hadoop. (n.d) *MapReduce Tutorial*. Retrieved from: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.pdf

[14]. KeyCDN. (2017, February 28) *Understanding the Apache Access Log*. Retrieved from: https://www.keycdn.com/support/apache-access-log/

[15]. The Internet Traffic Archive. (n.d). *NASA-HTTP*. Retrieved from: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html