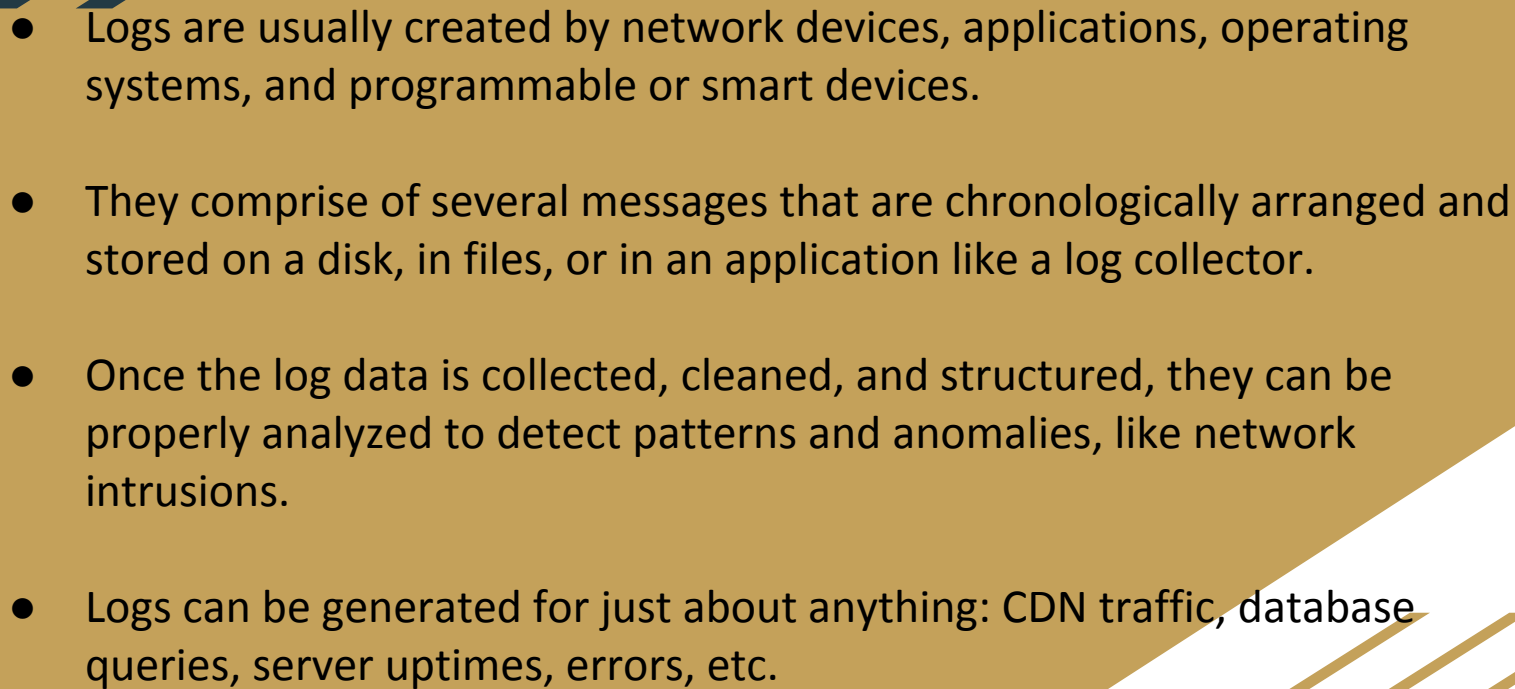


MapReduce based Log analysis using Hadoop

Akansh Kumar, Uddhav Raj, Rohit Ajit Mesharam, T.S.Ashutosh

What are Log Files?


- 
- Logs are usually created by network devices, applications, operating systems, and programmable or smart devices.
 - They comprise of several messages that are chronologically arranged and stored on a disk, in files, or in an application like a log collector.
 - Once the log data is collected, cleaned, and structured, they can be properly analyzed to detect patterns and anomalies, like network intrusions.
 - Logs can be generated for just about anything: CDN traffic, database queries, server uptimes, errors, etc.



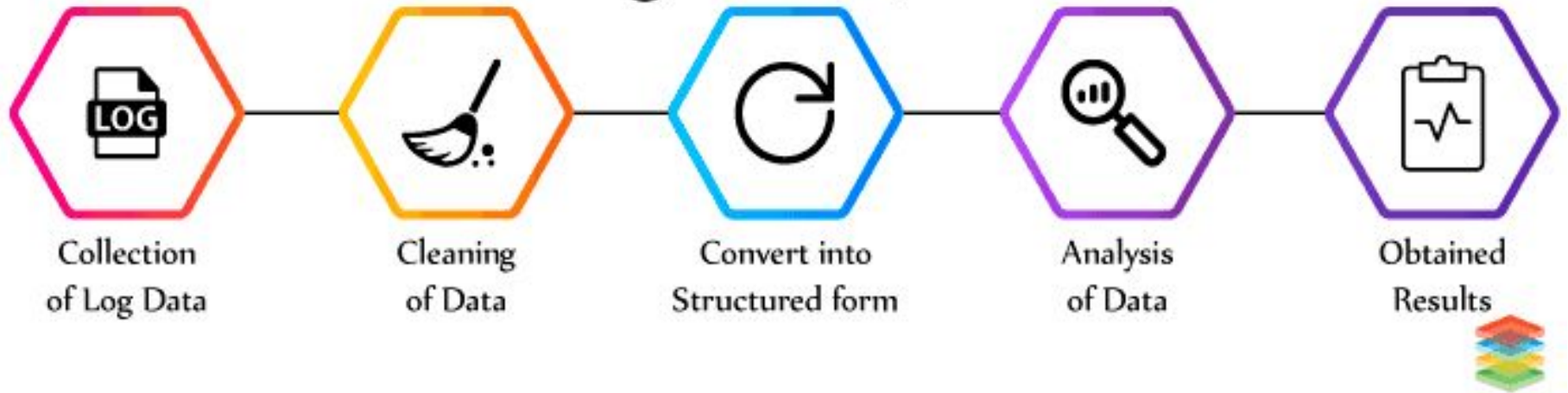
What is Log Analysis ?



What is log analysis ?

- Log is the main source of system operation status, user behavior, system's actions etc.[1]
 - Computers, networks, and other IT systems generate records called audit trail records or logs that document system activities.
 - Log analysis is the evaluation of these records and is used by organizations to help mitigate a variety of risks and meet compliance regulations.
- 

Log Analysis



Applications of Log Analysis

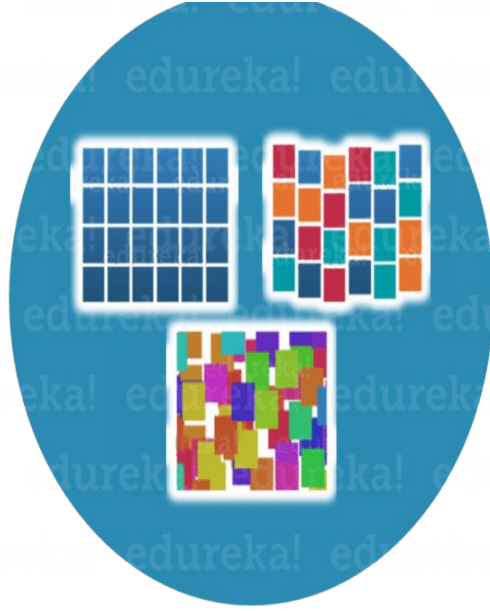
- To comply with internal security policies and outside regulations and audits.
- To understand and respond to data breaches and other security incidents.
- To troubleshoot systems, computers, or networks or understand the behaviour of users.
- Log analysis also helps companies save time when trying to diagnose problems, resolve issues, or manage their infrastructure or applications.
- To monitor the system health ensuring application availability and performance levels are met.
- Log data is a useful source of info that provides insight into how customers interact with products.



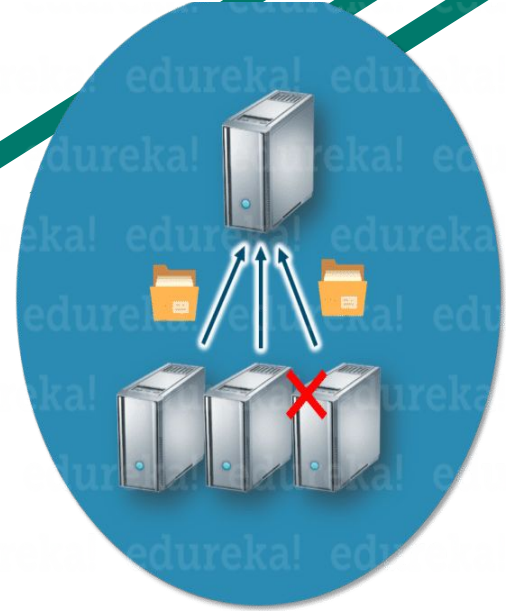
Problems in Traditional Log Analysis



Storing huge and exponentially growing
datasets



Processing data having complex structure
(structured, un-structured, semi-
structured)



Bringing huge amount of data to
computation unit becomes a bottleneck

Storing the colossal amount of data.

- Due to the increasing scale and complexity of distributed systems, the size of logs must be very large.
- Storing this huge data in a traditional system is not possible.
- The storage will be limited only to one system and the data is increasing at a tremendous rate.
- Thus, it's inefficient for common methods to analyze system logs on single node.



Storing Heterogeneous Data.

- In traditional approach, the main issue was handling the heterogeneity of data i.e. structured, semi-structured and unstructured.
- A single system to store all these varieties of data, generated from various sources is not possible or inefficient.


Accessing and Processing speed.

- Even if we increase the storage capacity of a system, but the disk transfer speed or the access speed is not increasing at similar rate.
- It is very time-consuming to diagnose a great amount of log messages produced by a large scale distributed system on just one node.[usb 3.0]





Solution to the Problem

- The problem of log files analysis is complicated because of not only its volume but also its disparate structure.
 - Therefore, there is a great demand to adopt a distributed method for anomaly detection techniques based on log analysis.
 - Distributed paradigms are designed to compute large volumes of data in a parallel fashion[3].In which the workload is divided across a large number of machines or nodes.[2]
 - The Solution is Hadoop.
- 

WHY HADOOP ?

- In 2009, Erik Paulson and Andrew Pavlo [4] compared the Hadoop MapReduce and SQL DBMS suggested that Hadoop MapReduce tunes up the task faster and load data faster than DBMS.[1]
- RDBMS technology is a proven, highly consistent, matured systems supported by many companies. Whereas, Hadoop is in demand due to Big Data, which mostly consists of unstructured data in different formats. Hadoop specializes in semi-structured, unstructured data like text, videos, audios, Facebook posts, logs, etc.
- As log files is one of the type of Big Data which grows rapidly so Hadoop is the best suited platform for storing log files and parallel implementation of MapReduce program for analyzing them.

Literature Review


- There are many already implemented tools for system monitoring like vmstat ,netstat ,etc.[1]
- Vmstat (virtual memory statistics) is a computer system monitoring tool that collect and display summary information about operating system,memory ,processes,interrupts.
- Netstat is command-line tool network utility tool that display network connections for the Transmission Control Protocol routing tables.

- Netstat is used for finding problems in the network and to determine the amount of traffic on the network as a performance measurement.
- These system are not distributed system.so these tools are very inefficient when used for analysis of log files in distributed systems as log file are very big in size.
- So we have proposed an idea to deal with distributed system network.


- The framework used is Hadoop which provides a software Framework for distributed storage and parallel processing of big data using MapReduce Programming model.[2]
- MapReduce process log in a high efficient and stable way.



What is Hadoop?




Hadoop is a framework that allows us to store Big Data in a distributed environment, so that, we can process it parallely.





Components of Hadoop are:

1. HDFS- It creates an abstraction for distributed storage. We can see it as a single logical unit but actually we are storing our data across multiple nodes.
 2. Yarn- It performs the processing activities by allocating resources and scheduling tasks.
- 



HDFS
(Storage)

Allows to dump any kind of data
across the cluster

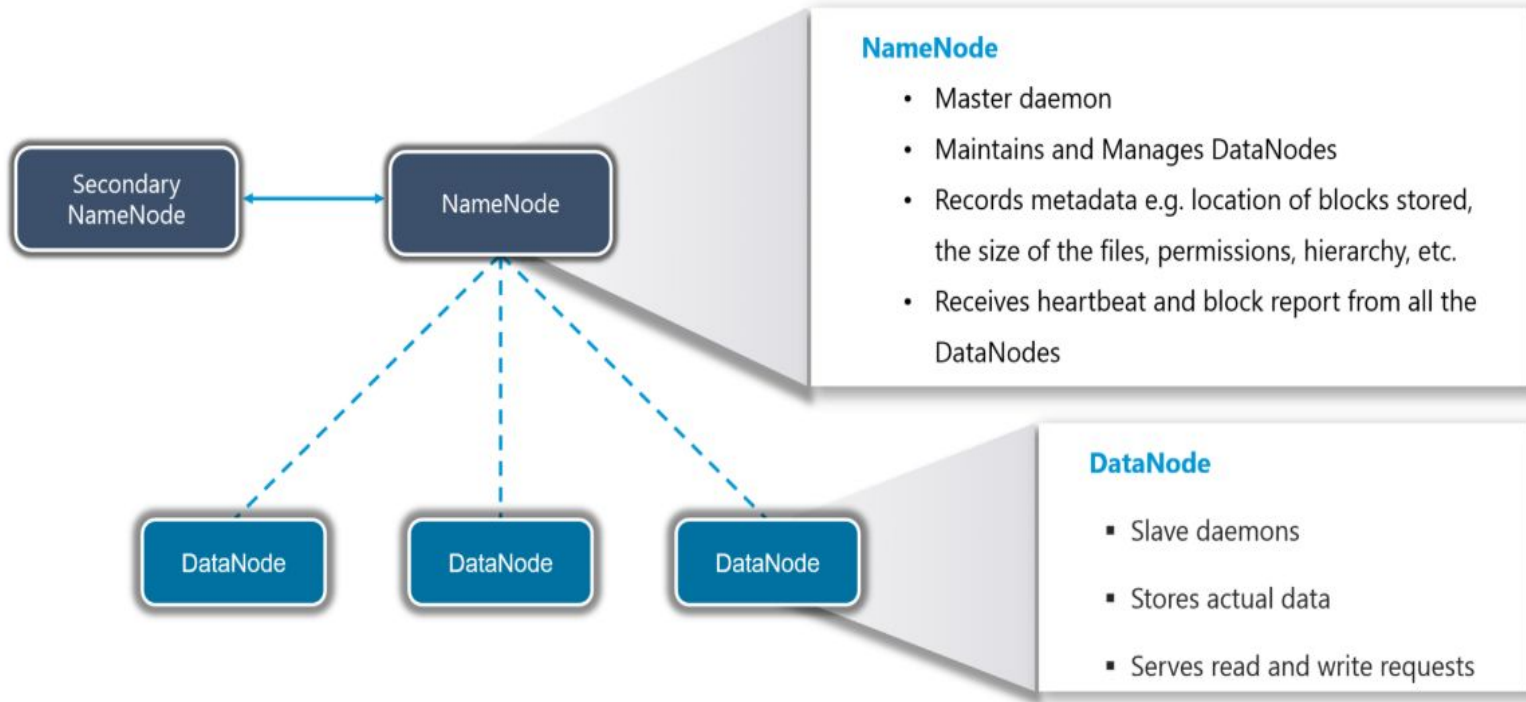


YARN
(Processing)

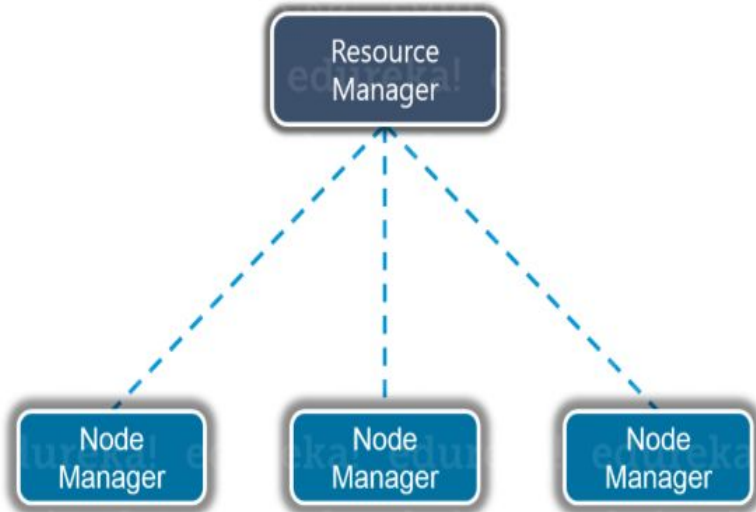
Allows parallel processing of the
data stored in HDFS



HDFS



YARN



ResourceManager


- Receives the processing requests
- Passes the parts of requests to corresponding NodeManagers

NodeManagers


- Installed on every DataNode
- Responsible for execution of task on every single DataNode



What is MapReduce?



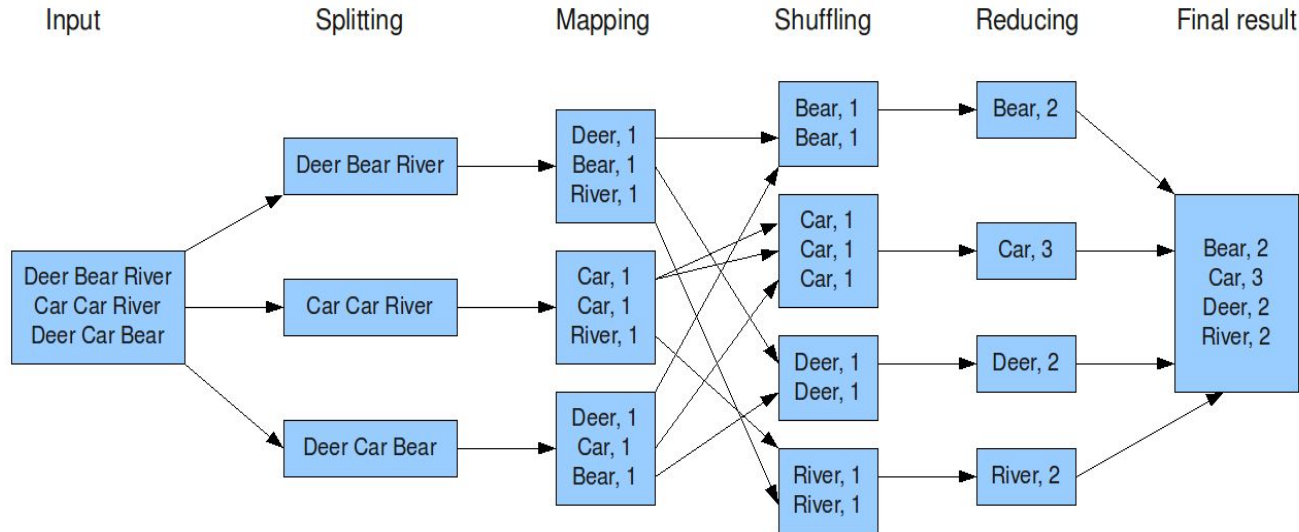
MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.



1. Map stage – The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
2. Reduce stage – This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer job is to process the data that comes from the mapper. After processing, it produces a new set of output, which is stored in the HDFS.

MapReduce Example

The overall MapReduce word count process







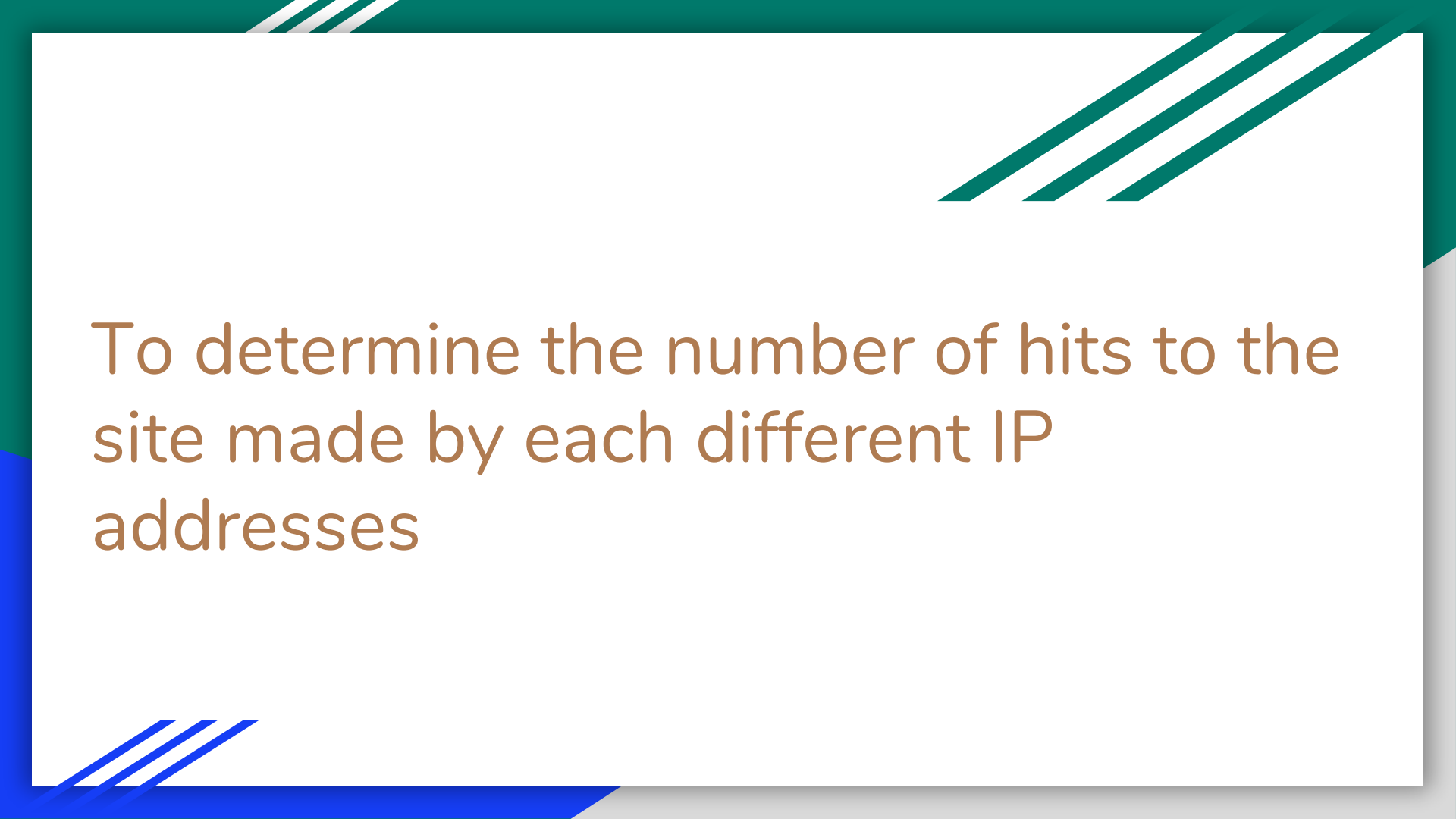
The Log File Analysis

Log File Example


```
1 10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET / HTTP/1.1" 403 202
2 10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET /favicon.ico HTTP/1.1" 404 209
3 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET / HTTP/1.1" 200 9157
4 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 200 10469
5 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/reset.css HTTP/1.1" 200 1014
6 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/960.css HTTP/1.1" 200 6206
7 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/the-associates.css HTTP/1.1" 200 15779
8 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/the-associates.js HTTP/1.1" 200 4492
9 10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lightbox.js HTTP/1.1" 200 25960
10 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/search-button.gif HTTP/1.1" 200 168
11 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/dummy/secondary-news-3.jpg HTTP/1.1" 200 5604
12 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/dummy/primary-news-1.jpg HTTP/1.1" 200 10556
13 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/dummy/primary-news-2.jpg HTTP/1.1" 200 9925
14 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/closelabel.gif HTTP/1.1" 200 979
15 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/home-logo.png HTTP/1.1" 200 3892
16 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/dummy/secondary-news-2.jpg HTTP/1.1" 200 5397
17 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/loading.gif HTTP/1.1" 200 2767
18 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/dummy/secondary-news-4.jpg HTTP/1.1" 200 5766
19 10.223.157.186 - - [15/Jul/2009:15:50:36 -0700] "GET /assets/img/home-media-block-placeholder.jpg HTTP/1.1" 200 68831
20 10.223.157.186 - - [15/Jul/2009:15:50:37 -0700] "GET /assets/img/dummy/secondary-news-1.jpg HTTP/1.1" 200 5766
21 10.223.157.186 - - [15/Jul/2009:15:50:37 -0700] "GET /assets/swf/home-media-block.swf HTTP/1.1" 200 123884
22 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET / HTTP/1.1" 200 9157
23 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/css/960.css HTTP/1.1" 304 -
24 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/css/the-associates.css HTTP/1.1" 304 -
25 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 304 -
26 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/js/lightbox.js HTTP/1.1" 304 -
27 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/css/reset.css HTTP/1.1" 304 -
28 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/js/the-associates.js HTTP/1.1" 304 -
29 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/img/dummy/secondary-news-4.jpg HTTP/1.1" 304 -
30 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/img/search-button.gif HTTP/1.1" 304 -
31 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/img/dummy/primary-news-1.jpg HTTP/1.1" 304 -
32 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/img/dummy/secondary-news-3.jpg HTTP/1.1" 304 -
33 10.223.157.186 - - [15/Jul/2009:15:50:51 -0700] "GET /assets/img/home-media-block-placeholder.jpg HTTP/1.1" 304 -
```



To display the number of hits for each
different file on the website



To determine the number of hits to the site made by each different IP addresses





To find the most popular and least popular file on the website



Further Improvements

Yan Liu suggested a MapReduce Framework using k-means clustering algorithm where the log files are clustered to reduce the amount of processing time. [1]

[1] System Anomaly Detection in Distributed Systems through MapReduce-Based Log Analysis, Yan Liu

Technology Stack

1. Python
2. Hadoop
3. Hadoop Streaming
4. Linux(CentOS 6.3)

References

- [1] Hemant Hingave, Prof. Rasika Ingle, "An approach for MapReduce based Log analysis using Hadoop", IEEE SPONSORED 2'ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS(ICECS '2015)
- [2] Ning Cao, Guangwei Qiao, Yan Liu, Wei Pan, "System Anomaly Detection in Distributed Systems through MapReduce-Based Log Analysis", 3rd International Conference on Advanced Computer Theory and Engineering(ICAETE), 2010
- Data Clustering, http://biocomp.bioen.uiuc.edu/oscar/tools/Hierarchical_Clustering.html
- [3] Chao Tian, Haojie Zhou, Yongqiang He, Li Zha, "A Dynamic MapReduce Scheduler for Heterogeneous Workload", Eighth International Conference on Grid and Cooperative Computing, 2009
- [4] Tom White: "Hadoop: The Definitive Guide (1st Ed.)", O'Reilly Media, Inc., United States of America, 2009.

THANK YOU