RECTIFY AND ENVISION THE SERVER LOG DATA USING APACHE FLUME

Nilofar Begum[1], Prof. A. Ananda. Shankara[2]
School of computing and Information Technology, Reva University, Bangluru , Karnataka, India

Abstract: Apache Flume is a Distributed data collection service that gets flows of log data from their source and aggregates them to where they have to be processed .It can move large amount of Streaming log data from one place to another e.g. : from web server to Hadoop cluster. The main goals are reliability, scalability, and extensibility. Most Enterprises use Apache Flume's powerful streaming capabilities to land data from high-throughput streams in the hadoop distributed file system. Main sources of these streams are application log data, machine data, social media, geological data and sensor data. In this paper , we propose a novel for visualizing the server log data to enable how an enterprise security breach analysis and response might be performed. Apache Hadoop take the server log data to the next level for speeding and improve the security forensics and provide a low cost platform which shows the compliance. In this project we are focusing on a network security use case. Specifically we are looking how Hadoop will help the administrator of a large enterprise network system and respond to a distributed denial of service attack.
Keywords: visualize server log data, Apache Flume, Streaming log data.

I. INTRODUCTION

Big data Environment in the 21st century has bloomed up with an enormous data sets compressing of audio, video, textural images and applicatory files hence handling these files has thrown a challenge for the current network system. Gartner defines Big data is a high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Hadoop is open source platform for structuring Big Data, and solves the problem of formatting it for subsequent analytics purposes. Hadoop uses a distributed computing concept consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale and support extremely large data stores Hadoop is new and immature technology compared to relational databases. Hadoop was originally developed especially for the large-scale batch processing of log data and was not designed to meet enterprise expectations for data security and integrity. To overcome these issues are being addressed by Apache Flume for both enterprise data security and Hadoop technology which provided massive storage for any kind of data, enormous processing power and to handle the virtual limitless concurrent tasks.

Server log data are the log files generated by computer which captures the network and server operations data. These are useful for managing network operations like security and regulatory compliance. And Log analysis is the common use case for inaugural Hadoop project. The earliest uses of Hadoop were for the large scale analysis of stream logs that records data about the web server pages that people visit and in which order they visit them.

Apache Flume is an open -source frame work and designed for Big data eco-system. It is nicely integrated with the hadoop ecosystem which contains various file formats. i.e. Apache Flume is a continuous data ingestion system which is distributed, Durable storage, Failover and / or replication. and easy to install , configure reconfigure and run. Basically Apache Flume is designed for log aggregation system developed by Cloud era Engineers. The main job is to handle any type of streaming log data and it is Low-cost of installation, operation and maintenance and specially it is Highly customizable and extendable.

Visualizing weblogs is very important to many organizations, but it is a challenging task that requires special tools due to large volumes of data and the complexity of analysis. To benefit most from the solution, we need to be able to process data in real-time, batch and and more. The typical weblogs analysis solution on Hadoop may be quite complex and require making different technologies to work together, while an integrated platform such as Apache Flume can help with most difficult parts.

In summery, the challenges for Refining and Visualizing server log monitoring lie in the complexity of the data and in the high demands for monitoring the different tasks. In dealing with such tasks, the following questions will asked by data analysts.

- How to design or develop the applicable visualization system which will fully adaptable to the data sources, i.e., the Hadoop Flume.
- How to maintain immediate data access to achieve visualization response ?
- How to keep these data in safety and maintain security ?
- How to detect anomalies with server monitoring logs ?

To address these problem, we proposed and implemented visualization of server log system using Apache Flume, which is responsible for both data refining and data visualization. Data Refining means refines disparate data within a common context which increase the awareness and understand the data, remove data variability and redundancy, and develop an integrated data resource. The data visualization offers the extensible multi-view layouts for

hierarchical data and sequential data. We have implemented our approach and applied it in a distributed computing environment. The experiments demonstrate that our Hadoop framework is amenable for real-time monitoring tasks and clearly facilitates anomaly detection. The important components of our system are summarized as follows:

- The website or web server is instrumented to capture different user interactions on the page and then logged to web server logs.
- Apache Flume™ is configured to extract logs from log files in real-time and transport them to a Hadoop HDFS.
- To support data processing to generate insights, the data is batched and written to HDFS
- For real-time processing, the data is directly fetch into a processing system like Apache Flume..
- Workflows using Map Reduce/Apache Pig™ are created to cleanse log data and generate insights periodically. The output data produced is then written back to HDFS. These scheduled scripts actually analyze the logs on various dimensions and extract the results. The results are by default stored into HDFS. Real-time aggregates are stored in HBase.
- Hive is set up to expose the raw web log and output of data analysis to be accessed using SQL. Schemas for web log and insights have to be modeled and maintained.

II. RELATED WORK

In this section we presents the related work in this domain, now a days internet usage is one of the emerging area where data analysis is important to track the user behavior in order to better serve users. sever log data is a collection of log file which consists of several file which is automatically generated and maintained by server consisting of list of different activities it performed e.g. web server log which consists of history of page requests. And Information about web request, include client IP address, request date and time, HTTP code, user agent and referrer are typically added. these all information is combined into a single or separated into several logs, such as access log, referrer log or error log. These log files are specifically not accessible to common internet users, only to the web masters or authenticated person. The Hadoop cluster contains and access the large amount of semi structured data in a parallel computational model. And these log files are generated from the website or web server comprise of large amount of data which cannot be handled by a Relational database system of other programming languages for computation. The below fig shows how log files can be process and keeps track on sessions accessed by the user, using Hadoop system architecture.

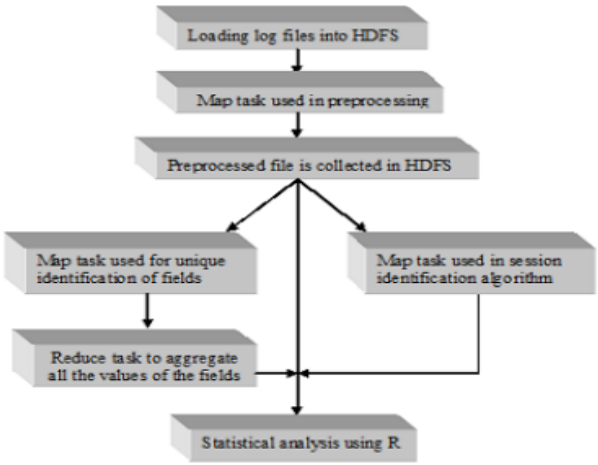


Fig : processing of log files using Hadoop framework.

In this architecture the work is divided into two phases i.e. storage and processing phase is made in HDFS. First load the log files which contains a number fo recors that corresponds to automatic requests originated by web server which includes huge amount of erroneous, misleading and incomplete information. After loading the data the applicable resources will stored in the HDFS as text file. The stored web logs are used for further analysis of session identification which is utilized by the user and also identify unique user, and unique URLs accessed. The processed log files are used to find the user identification, In Map reduce identification of the unique values are based on key-value pair. These log files are consists of different fields and the user identified through IP address, which belongs to key and their corresponding count as values. After all keys are found, combiner is used to aggregate the all values of specific key, then combined result will passed to the reduce task and counts the total IP accessed will listedas text data in HDFS.

A service for streaming logs into Hadoop

Flume Works. The architecture of Flume is based on transferring the streamlined codebase that is easy-to-use and easy-to-extend. Our proposed system architecture is designed Flume with the following components: Event, Source, Sink, Channel Agent and Client. An Event is a single unit of data which is transported by Flume. The Source is an entity which data enters into Flume. A different type of sources allow data to be collected, e.g. syslogs. Sink is the entity that sends the data to the destination. An example is the HDFS sink that writes events to HDFS. Channel –is the bridge between the Source and the Sink. Sources ingest the events into channel and sinks drain the channel. Agent is the any physical JVM running Flume. It consists of sources, sinks and channels. The Client produces and transmits the Event to the Source operating within the Agent

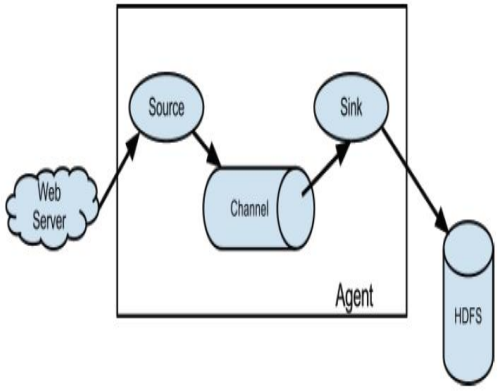


Fig: Data flow model

The flow of Flume is starts from the client which is web server. The web server transmits the event to the source operating under the agent. The source receives this event and then transfers it to one or more channels. These channels are drained by sinks operating within the same agent. which allows the decoupling of ingestion rate from drain rate . The sink removes the event from the channel and puts it into an external repository like HDFS (via Flume HDFS sink) or forwards it to the Flume source of the next Flume agent (next hop) in the flow. Flume agents can be linked together by connecting with sink of one agent to the another agent.

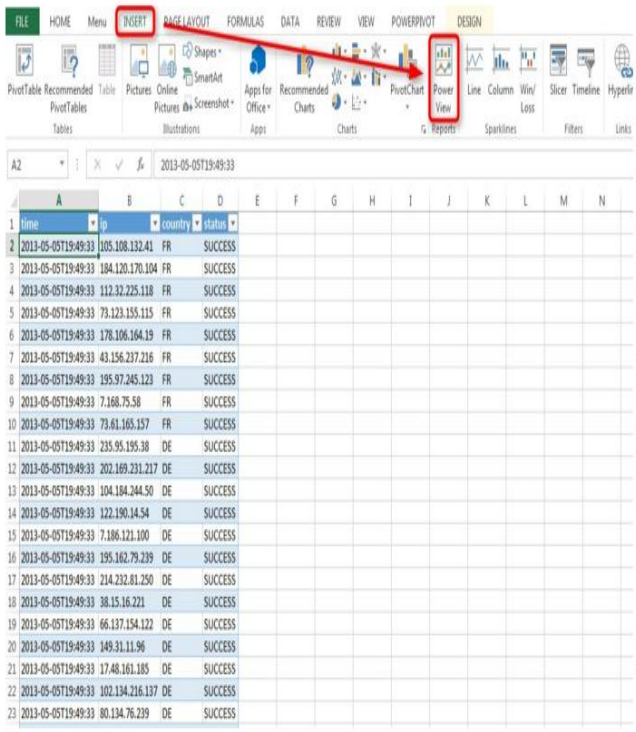
Refine and Visualize server logs using Apache Flume

The Five main stages for refine and visualize server logs are First download and extract the server log files, configure the Flume, Generate the server log files, Import these server log data into Excel, and finally Visualize the server log data by using the Excel power view.

In first step the downloaded server log files are in compressed folder so we have to extract the server log data. For configuring the Flume first We have to create a few directories for Flume and we need to update the flume agent config. and need to set up a flume agent and edit the flume configuration file. After the flume will start which will begin collecting data for us. For generating the server log data we are using a Python script and create an Hcatalog table from the data. When the log file has been generated, the time stamp will appear and we have to create an Hive table from the log files. For importing the server log data we are using Excel Professional Plus 2013 which access the generated log data and it filters these data in table and it displays the data in the form of table. After successfully importing the log data into Microsoft Excel, we can use the Excel power view to analyze and visualizing the data. In our project we are analyzing the data for a denial-of-service attack : Review the network traffic, Zoom in on one particular region and generate the list of attacking IP address.

III. EXPERIMENTAL RESULTS

In this paper we are demonstrating how to analyze and visualize an enterprise security and response might be performed. How to stream the server log files into Hadoop with Apache Flume. and how to visualize the data with Microsoft Excell . After downloading and generating the server log data into Flume we are importing the these log data into Excel 2013 which access the server log data. The Figure 1: show the successfully imported data into Microsoft Excel.



time	ip	country	status
2013-05-05T19:49:33	105.108.132.41	FR	SUCCESS
2013-05-05T19:49:33	184.120.170.104	FR	SUCCESS
2013-05-05T19:49:33	112.32.225.118	FR	SUCCESS
2013-05-05T19:49:33	73.123.155.115	FR	SUCCESS
2013-05-05T19:49:33	178.106.164.19	FR	SUCCESS
2013-05-05T19:49:33	43.156.237.216	FR	SUCCESS
2013-05-05T19:49:33	195.97.245.123	FR	SUCCESS
2013-05-05T19:49:33	7.168.75.58	FR	SUCCESS
2013-05-05T19:49:33	73.61.165.157	FR	SUCCESS
2013-05-05T19:49:33	235.95.195.38	DE	SUCCESS
2013-05-05T19:49:33	202.169.231.217	DE	SUCCESS
2013-05-05T19:49:33	104.184.244.50	DE	SUCCESS
2013-05-05T19:49:33	122.190.14.54	DE	SUCCESS
2013-05-05T19:49:33	7.186.121.100	DE	SUCCESS
2013-05-05T19:49:33	195.162.79.239	DE	SUCCESS
2013-05-05T19:49:33	214.232.81.250	DE	SUCCESS
2013-05-05T19:49:33	38.15.16.221	DE	SUCCESS
2013-05-05T19:49:33	66.137.154.122	DE	SUCCESS
2013-05-05T19:49:33	149.31.11.96	DE	SUCCESS
2013-05-05T19:49:33	17.48.161.185	DE	SUCCESS
2013-05-05T19:49:33	102.134.216.137	DE	SUCCESS
2013-05-05T19:49:33	80.134.76.239	DE	SUCCESS

Fig: successfully imported Horton works Sandbox data into Microsoft Excel.

The below figure 2: displays a global view of the network traffic by country. The color orange represents successful, authorized network connections. Blue represents connections from unauthorized sources.

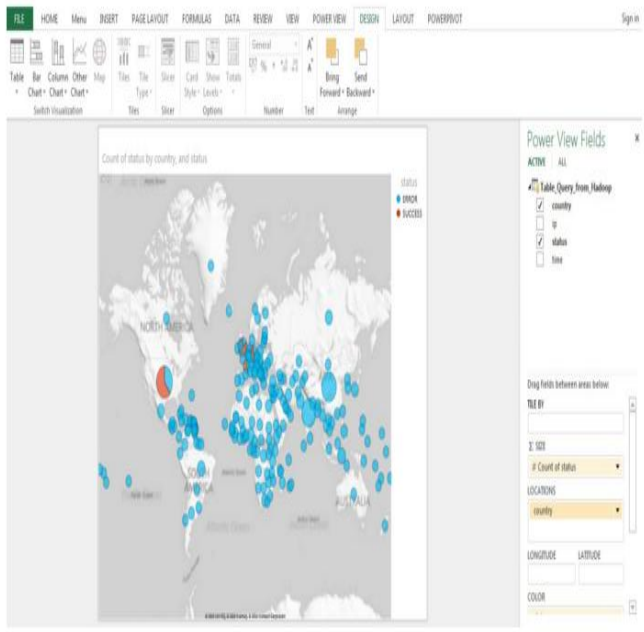
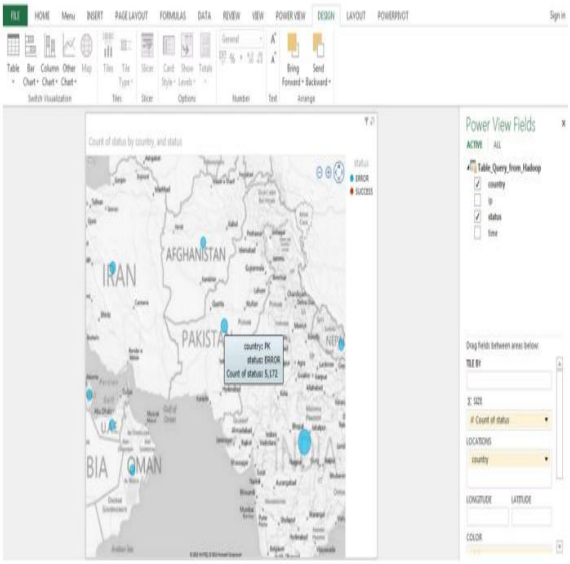


Figure 2: Global view of network traffic by country.

Assume that recent denial-of-service attacks have originated in Pakistan. The following figure 3 represents a closer look of traffic from that country



IV. CONCLUSION

This present situation the Big data technologies are successfully incorporated for all real domain problems like Web log analysis, fraud detection and many more. This paper reveals the importance of the Big data technology i.e. Apache Flume (Hadoop) where Flume Handles the large amount of data for analyzing and visualizing the server log files. In our proposed work we were able to block unauthorized access, and restore VPN access to authorized users. and also we can protect the company network from similar attacks in the future.

REFERENCES

[1] Joseph McKendrick, "Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data strategies survey," September 2012.
[2] Jeffrey Dean and Sanjay Ghemawat, "Map Reduce: Simplified Data Processing on Large Clusters," Google, Inc
[3] "Hadoop", http://hadoop.apache.org.
[4] http://hortonworks.com/blog/hadoop_tutorial_visualizing_server_logs/
[5] "Hortonwork",http://hortonworks.com/hadooptutorial/ howtorefineandvisualizeserverlogdata/
[6] "Flume",https://flume.apache.org/FlumeUserGuide.html
[7] http://www.rittmanmead.com/2014/05/tricklefeedingwebserverlogfilestohdfsusingapacheflume/

