Analysis of Network Traffic Data Using Hadoop and Apache Hive

Swathi Prabhu[1], Pooja Parashuram Bajantri[2], Prajna B Nayak[3], Priya Shetty[4], Rachana N[5]

Abstract— Computer network traffic can be classified into various classifications based on parameters such as port number, IP address or protocol. The sequence and pattern of the traffic in the network is illuminated using flow analysis. This helps network administrator to understand the network usage and to examine the behavior of the user using the network and to monitor the operations going on in the network. Flow analysis helps in fault tolerance, resource allocation and network capacity planning. Since this is data age, due to the fast growing network day by day the volume of the traffic is getting huge. So it is very difficult to collect, store and analyze this large data on a single machine. It requires scalable tools to measure, classify and analyze this traffic data. Due to the limited computational capability and storage capacity traditional tools are failed. Hadoop is a leading distributed framework which is designed to execute tremendous data that can be of hundreds of terabytes and even petabytes of data. In this paper a Hadoop framework based network traffic data analysis is done. Here the input is appended to the Hadoop Distributed File System (HDFS) and process the data using Map-Reduce, Hive. The output is displayed using Tableau and RStudio.

Index Terms— Hadoop, HDFS, Hive, Internet Traffic, Map-Reduce, Network flow analysis.

1 INTRODUCTION

A network consists of two or more computers that are linked in order to share resources, exchange files or allow electronic communications. The computers on a network may be linked through cables, telephone lines, radio waves, satellites or infrared light beams.

Data analytics has become a key element of the business decision process over the last decade. Classic systems based on relational databases and expensive hardware, while still useful for some applications, are increasingly unattractive compared to the scalability, economics, processing power, availability offered by today's network driven distributed solutions.

Network traffic analysis [1] is nothing but analyzing the amount of data moving across a network at a given point of time. Network packets which are mostly the network data are the loads in the network. The main objective of monitoring network traffic is to check the availability and smooth operations of the network. Network traffic requires reviewing of each incoming and outgoing packets. It can be classified using HTTP, IP, ICMP, TCP and UDP. To address the challenges in the traditional systems we use Apache Hadoop. It is an open source software framework for storage and large scale processing of datasets.

[1] Assistant Professor, Dept. of CSE, SMVITM, Bantakal. prabhuswathi2@gmail.com
[2][3][4][5] Student, Dept. of CSE, SMVITM, Bantakal. poojaupadhya96@gmail.com [2], nayak.prajna97@gmail.com [3], shettypriya495@gmail.com [4], rachanasmvitm@gmail.com [5]

Advantages of Hadoop System are as follows:

1. Scalability: Ability to store data in a distributed fashion in a large set of servers
2. Cost-effective Solution: Hadoop's scale out architecture with Map-Reduce programming model, allows the storage and processing of BigData in affordable manner.
3. Flexibility: The Map-Reduce model of Hadoop Framework can deal with both structured and unstructured datasets.
4. Fast: Parallel Processing allows multiple processes to divide the task and hence take fewer-time.

The data analysis is carried out through Hadoop Map-Reduce and Apache Hive.

Map-Reduce[7] is a programming model used in the Hadoop ecosystem to process large amount of data in a distributed manner. Map-Reduce programs are inherently parallel, thus able to perform very large-scale data analysis. Although it is a powerful tool for processing, it is harder to understand and to develop. Hence we can use Apache Hive, a project developed on top of Hadoop framework which is easy to understand and develop.

Apache Hive is the SQL-On-Hadoop technology to query heterogeneous data stored in different databases and filesystems that are fused with Hadoop. This approach is significantly faster and has new features that will support performing inserts and updates to tables. To visualize the data, we made use of Tableau Public and RStudio.

Tableau public is an open source tool for pattern discovery using data visualization. Learning and using Tableau is a very low time consuming activity. But there are limitations, such as we can only read from txt, Excel or Odata sources and there is no security for the data. Hence we used RStudio for analysis.

RStudio provides enhanced security and authentication, advanced resource management and priority email support.

2 ARCHITECTURE

2.1 Apache Hadoop

Apache Hadoop 1.0 and apache Hadoop 2.0 are the two versions of Hadoop. The 1.0 version was the initial release, where there were only two modules: Map-Reduce and HDFS (Hadoop Distributed File System). The Map-Reduce was the programming model used for large scale processing and was also taking care of cluster resource management. The HDFS is used to manage the storage across the network of machines.

Apache Hadoop 2.0 represents a generational shift of architecture in the apache Hadoop. With the introduction of YARN (Yet another Resource Negotiator), Apache Hadoop became a significantly strong platform. YARN module takes care of cluster resource management in Hadoop 2.0, which allows the multiple applications to run in Hadoop. Only the data processing is done by Map-Reduce module. The Fig 1. depicts the Hadoop Architecture.

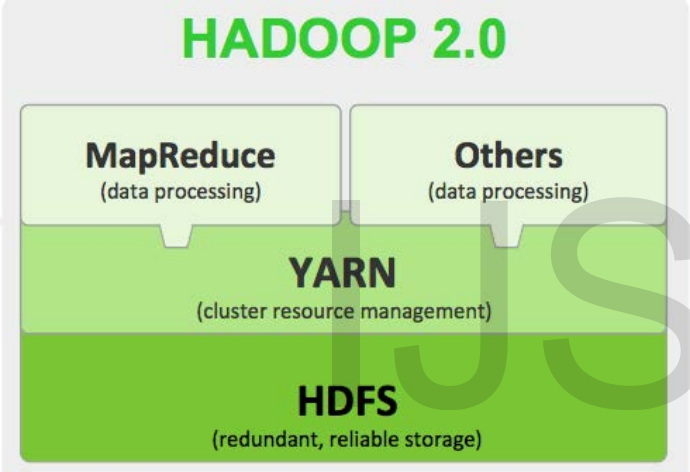


Fig 1. Hadoop Architecture

2.2 Map-Reduce

Google[3] was the first to propose the Map-Reduce programming model for page ranking or web log analysis. Map-Reduce is a programming model and an associated implementation for processing and generating big datasets with the parallel, distributed algorithm on a cluster. A Map-Reduce program is composed of two phases: a Map phase and a Reduce phase. Each phase has key-value[7] pairs as input and output.

Map phase: Master node takes large problem input and slices it into smaller sub problems and attributes these to worker nodes.

Reduce Phase: Master node collects the answers to the sub problems from the worker nodes and combines them in a predefined way to get the output.

2.2 Apache Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query and analysis. Initially Hive was developed by Facebook, later the Apache software foundation took it up and further developed as an opensource under the name Apache Hive. It is a platform used to develop SQL like scripts to do Map-Reduce operations.

Some of the features are as follows:

- It stores schema in the database and process the data into HDFS.
- It provides SQL like query language called HiveQL or HQL.
- It is designed for OLAP.

2.3 Comparison between Map-Reduce and Hive

The data processing is done using both Hadoop's Map-Reduce programming model and Apache Hive. The comparison between the two programming model is given in Table 1.

TABLE 1
COMPARISON BETWEEN MAP-REDUCE AND HIVE

Hadoop Map-Reduce	Apache Hive
Compiled language	SQL-Like query language
Low level of abstraction	Higher level of abstraction
More lines of code	Lesser lines of code
More development effort is required	Development effort is less
Code efficiency is high	Code efficiency is less

2.4 Tableau Public

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depicts the Trends, Variations and Density of the data in the form of Graphs and Charts. Tableau can connect to files, relational and Big Data sources to acquire and process the data.The software allows data blending and real time collaborations which makes it unique. It is used by businesses, academic researchers and many government organizations for visualize data analysis.Tableau provides solutions for all kinds of industries, departments and data environments. Speed of analysis, Self-reliant, Visual discovery, Blend diverse data sets, Real-time collaborations and Centralized data are some of the features of the Tableau Public.

2.5 RStudio

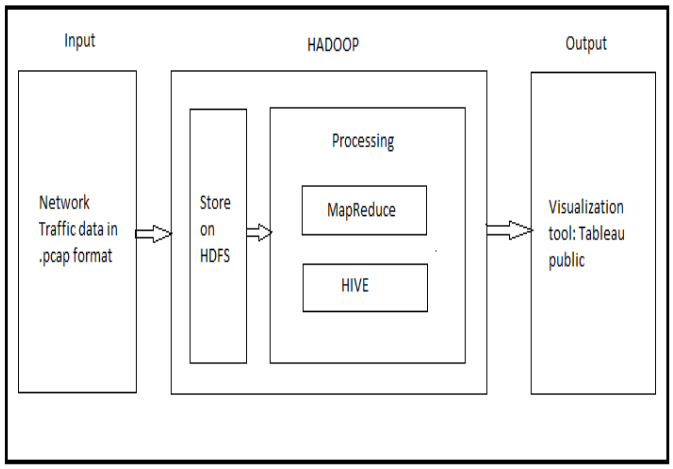
RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire. RStudio integrates with R as an IDE (Integrated Development Environment) to facilitate the user with further functionality. RStudio is a development environment facilitates with souce code editor, build automation tools and the debugger. Interactive, Easy debugging, Friendly environment for package development are some of the advantages of using RStudio.

3 SYSTEM OVERVIEW

The proposed system involves four main phases: Collection

and Conversion of input data set, storing in HDFS, processing of data using Map-Reduce and Hive, visualization using Tableau Public and RStudio.The working procedure of the system is represented by Fig 2.

Fig. 2 Proposed System Architecture



Four phases are as follows:

A. First Phase:

a) Collection of dataset: We are using pcap file as a dataset, collected from [6]. And also we used Wireshark tool to capture the packets.

b) Conversion of data set from .pcap format to .csv file format: Wireshark tool is used for conversion of file format.

B. Second Phase:

a) Setting up of Hadoop platform: Installed Hadoop on Ubuntu 16.04 LTS version.

b) Load dataset into HDFS

C. Third Phase:

a) Processing by Map-Reduce: We used Map-Reduce programming model in which we defined separate mapper and reducer class to process the pcap data.

b) Processing by Hive: Initially we loaded the data into External hive table and then the data queried using HiveQL, a SQL like querying language.

D. Fourth Phase:

Analysis using visualization tool: Tableau Public, an open source visualization tool is used to display the analyzed result. We have also made use of RStudio for analysis.

4 EXPERIMENTS

4.1 Experimental Setup

Pre-requisites:

- RAM: Minimum 4GB
- Operating System: Ubuntu 16.04 LTS
- Java Version: JDK 1.8

Hadoop: we used Hadoop 2.7.4 version with Multi-node Cluster containing one master and a slave.

Apache Hive: we used Hive 2.2.0 version which runs on top of Hadoop Distributed File System (HDFS).

Tableau Public: A visualization tool of version 10.5 to display results graphically.

RStudio: A statistics Computational tool of version 1.1 to display the results.

4.2 Overall Results

1. PCAP dataset

   The dataset employed here is basically a pcap file which contains fields such as time of the packet capture, source IP address, destination IP address, protocol, length and other information about the packet. A screenshot of dataset is shown in Fig 3.

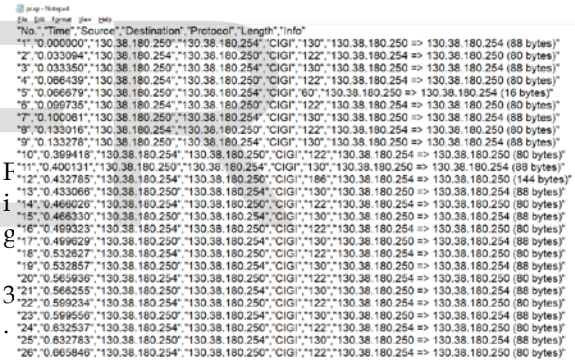
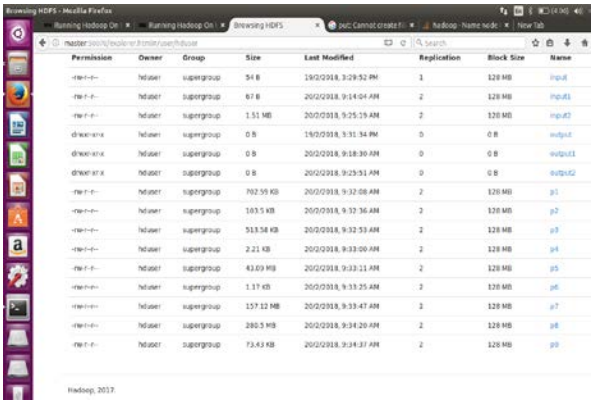


Fig 3. Example Dataset



2. Copying the Network Traffic data on HDFS is shown in the Fig 4.

Fig 4. File Storage in HDFS

3. Processing the dataset using Map-Reduce

The Fig 5. shows the result of Map-Reduce on multi-node cluster. The time taken is 19.23 seconds.

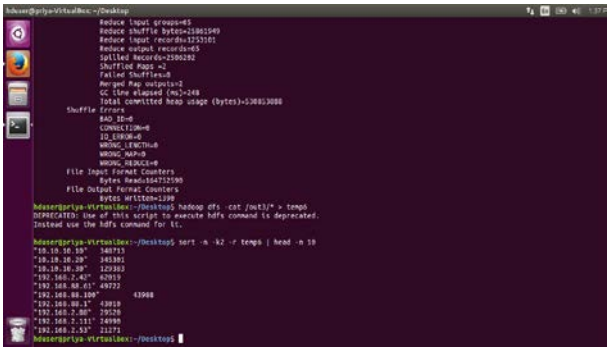


Fig 5. Result of Map-Reduce

4. Processing the data using Hive tool
The time taken to retrieve the Top-20 IP address which generates more traffic from the input dataset using Hive on multi-node cluster is 17.442 seconds which is given in Fig 6.

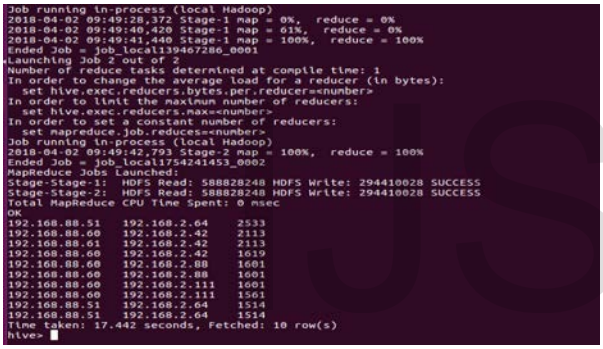


Fig 6. Hive Query processing

5. Visualization for Top-20 protocols used in the data transmission using Tableau Public is given in Fig 7.

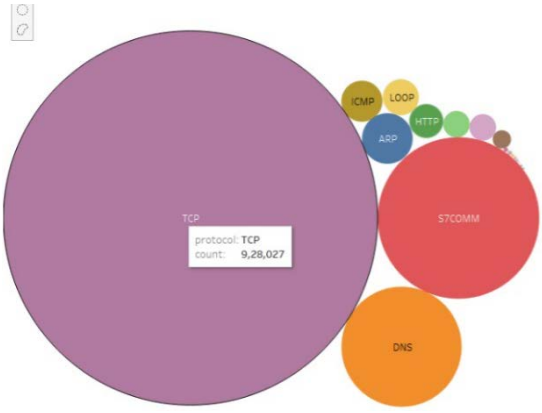


Fig 7. Procol based analysis

The Top-20 IP address that generates more traffic in the network is shown in Fig 8.

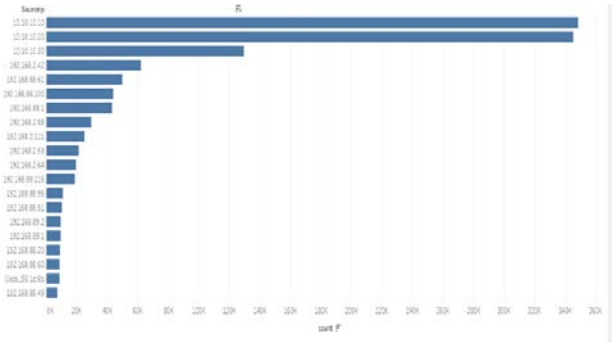


Fig 8. IP address based analysis

Visualization for Top-10 IP addresses that generates more traffic using RStudio is shown in Fig 9.

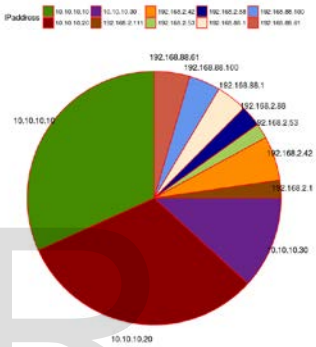


Fig 9. IP based analysis using RStudio

5 Conclusion and Future Work

The network traffic analysis is very useful for the network admins at ISP level to monitor the packet flow and plan for the future. In this paper, we have compared between Hive and Hadoop's Map-Reduce programming model.

As a future work, if the input dataset is unstructured we can use Apache Pig for processing and for the dynamic (Real-time) data processing we can use Apache Spark. Also, the size of the cluster can be increased for the faster and massive data processing.

References

[1] Preethi Joshi, Arati Bhandari, Kalyani jamunkar, Kanchan Warghade,

Priyanka Lokhande, "Network Traffic Analysis Measurement and Classification using Hadoop", IJARCCE, vol. 5, Issue 3, March 2016.

[2] Rakshitha Kiran P, "Hadoop Technology for Flow Analysis of the Internet Traffic", International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, Issue 2, February 2015.

[3] Youngseok Lee, Wonchul Kang, Hyeongu Son, "An Internet Traffic Analysis method with MapReduce", IEEE, 2010.

[4] Hadoop, http://hadoop.apache.org.

[5] Wireshark, http://wireshark.org.

[6] The training dataset collection – http://www.netresec.com/.

[7] Tom White, Hadoop-The Definitive Guide, 4th Edition, 2015.