International Journal of Electrical Sciences & Engineering (IJESE)
Online ISSN: 2455 6068; Volume 1, Issue 1; January 2016 pp. 25-28

Hadoop Framework for Flow Analysis and Congestion Control of the Big Data

Rakshitha Kiran P.[1], Jeffrey Shafer[2]

[1]Assistant Professor, MCA (VTU) Department
Dayananda Sagar College of Engineering, Bangalore
rakshitha610@gmail.com
[2]University of California
shaferjef@gmail.com

Abstract: **Flow categorization of the Computer Network traffic displays the sequence and pattern of the traffic in the network. This helps the network administrator to monitor the operations going on in the network, to understand the network usage and to examine the behavior of the user using the network. Tapping of the internet traffic can avoid a huge amount of problems. Flow analysis helps in fault tolerance, traffic engineering, resource allocation and network capacity planning. Due to fast growing network, the volume of the traffic is getting very big day by day. So it is very difficult to collect, store and analyze this huge data on a single machine. Hadoop is a leading framework which is designed to execute tremendous datasets that can be hundreds of terabytes and even petabytes of data. Hadoop performs brute force scan for multiple traces of input data and produces the output for traffic flow identification, flow clustering. In this paper a Hadoop based traffic analysis of internet traffic is done. Once the analysis is done flow control mechanisms are carried out to avoid any kind of congestion in flow.**

Keywords: **Hadoop, bigdata;**

1. INTRODUCTION

Data turns into big data when its volume, variety or velocity exceeds the potential of the storage systems to fetch, gather, store analyze and finally process the data. In many organizations have necessary equipments to handle this large amount of data. The data that is produced by various streams can be structured or unstructured so these systems lack the capacity to mine the data. It's not just about the volume of the data that is produced every day, but its also about the speed of the data in which the data arrives. For example, online social networking site facebook produces every day 2. 5 billion pieces of content and 500+terabytes of data, similarly data from twitter, weather forecast, call detail reports etc produces very huge amount of data every day.

Hadoop is a emerging framework performs distributed processing of the large data, big data, across big clusters of computer. . Hadoop is designed to process tremendous datasets and performs brute force scan on the multiple traces of the data and produces a efficient output.

Flow analysis of the traffic helps us to elucidate the pattern and the sequence of the traffic in the network. This information is needed for the network controller to monitor the various operations going in the network. Network controller understands the type of network, its usage, the behavior of the user using the network, monitors the network traffic and bottleneck.

In this paper we talk about the flow analysis and congestion control mechanism for the big data by making use of Hadoop framework. Hadoop performs Map and reduce techniques over the large datasets and gives a summarized output. The output is then analyzed for congestion.

2. LITERATURE SURVEY

The literature survey makes a very significant part in the research process. It is a place from where research ideas are drawn and developed into concepts and finally theories. In the paper The State of Enterprise Network [1] the author tells about the analysis done on the passive and active techniques that were used to measure the state of the communication using Internet. TCP is a major protocol used on the internet traffic but nowadays UDP has risen in the recent years. The author in the paper Traffic Classification on the fly [2] proposes a technique for analysing the traffic which are associated with the TCP connection. The detection of application using TCP flows is an important step in network security. the author proposes a method where the observation lies on the output of first five packets which are associated with the TCP connection to identify application. Based on the observation the author proposes a traffic classification mechanism which has 2 phases: First, Learning Phase and second is traffic classification phase. In the paper Network Traffic Characteristics of Data Centres in the Wild[3] the author analyzes data sets from 10 different data centers which

includes 5 commercial cloud based data centres, 2 private enterprise and 3 university campus data centers. The author collects and analyzes SNMP statistics, packet-level traces and topology. Examination is done on the basis the range of applications that deployed in those data centers, their placements, their flow level and also packet level transmissions of those applications. In the paper The Hadoop Distributed Filesystem: Balancing Portability and Performance [4] the author tells about the Hadoop framework. Hadoop is a popular framework which is open source which implements MapReduce and HDFS for the analysis of large data sets. HDFS is distributed file system which is used for storing purpose and MapReduce is processing technique used to process the large data sets. In the paper Toward Scalable Internet Traffic Measurement and Analysis with Hadoop [5] author tells traffic measurement and analysis of scalable Internet is difficult job because of large amount of data set is required for the computation and the storage resources. Traffic measurement and analysis is used to find out network usage and user behaviour. But because of very high growth of traffic and very high speed it's very difficult to analyse them. Hadoop is popular framework which is also open source is used for this traffic management and analysis. It makes use of HDFS and MapReduce techniques for massive data analysis because it supports scalable processing of data and storage of data.

3. HADOOP FRAMEWORK

In the traditional framework an organization had a system which would do both storing and processing of the big data. In this framework the data would be stored in Relation Database Management System(RDBMS) like MS SQL Server or Orcale Database system or any sophisticated softwares. These softwares were required to interact with the database. This approach works good when data is of less volume. But when it comes to dealing with large data it will be very difficult. So to overcome this drawback we go for Hadoop framework.

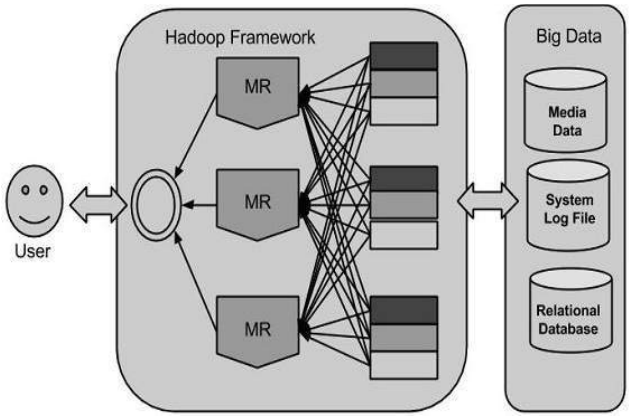

Fig. 1. Hadoop framework

Hadoop makes use of MapReduce algorithm where the data will be processed in parallel on various CPU. In this way hadoop framework is capable of runnig over multiple clusters and thus gives statistical analysis for huge amount of data.

4. PROPOSED METHODOLOGY

The proposed methodology consist of network where the large amount of packets flow, a system to capture the flow of packets, hadoop cluster to perform analysis. .

Packet flow:

Hadoop framework is used for flow analysis and congestion control for big data. Initially we will start capturing a large data from large network. The packet flow from this network will be captured in the form of text file. This packet flow is given as input to the Hadoop Cluster. Hadoop Cluster will accept the internet traffic flow and performs analyzes using by making use of 2 main functions called Map and Reduce. Hadoop cluster then generates an output file which is in the form of text file. The output obtained is then monitored for any congestion, if found flow control mechanism is done.

Flow Analysis:

Define Based on the input given to the hadoop cluster an output file is generated. The input file is a text file that consist of the various information like source address, destination address, protocol, length of the packet, information about the packet. This input file will be given to the hadoop cluster for analysis. The hadoop cluster will then perform Map and Reduce on the input file. Here the mapping is done based on the source ip address and destination ip address. The packets from same source ip address to same destination ip address are mapped together and finally reduce is applied. The summarized detailed of the packet are then stored in a separate output text file. The output file generated will have a detailed analysis of the packets that are exchanged between the hosts.

Congestion Control:

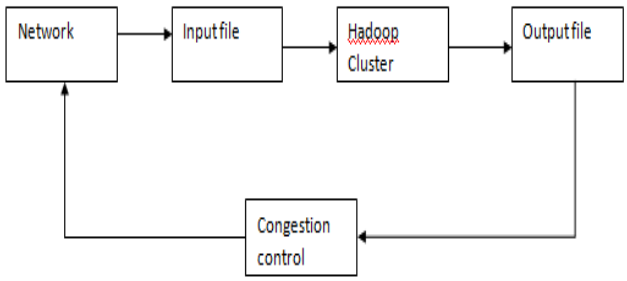

Fig. 2. flow diagram of system

Every path from the source to destination host is provided with some priority. The output text file generated will have details about amount of data passed from one particular source to destination. If the bytes of data have crossed the threshold then priority of that particular route will be updated and the packets will be forced to take up other congestion free path. Fig below shows flow diagram of the system.

5. EXPERIMENTAL RESULTS

Experiments were conducted by making use of 2 host and 4 nodes. The packets would travel from one host to another host and vice versa. Packet flow was captured from both the hosts. The packet information is in the form of text file.

Figure below shows the screenshot of the network and packet information that are exchanged between the nodes.

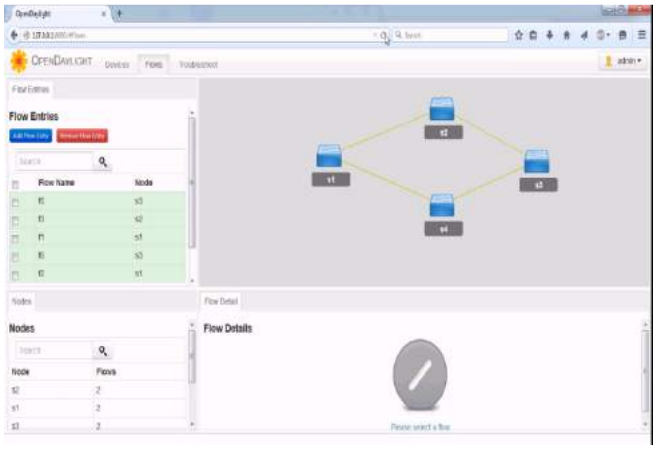

Fig. 3. screenshot of network topology created

Fig. 4. screenshot of packet information

The packet information collected will be given for the Hadoop cluster for the analysis. Hadoop cluster performs analysis based on Map Reduce technique. The input file i. e. the text file is stored in the separate folder called input. We need to specify the path of input file for the hadoop cluster. Fig below shows screenshot of the hadoop performing Map Reduce technique and then generating the input file.

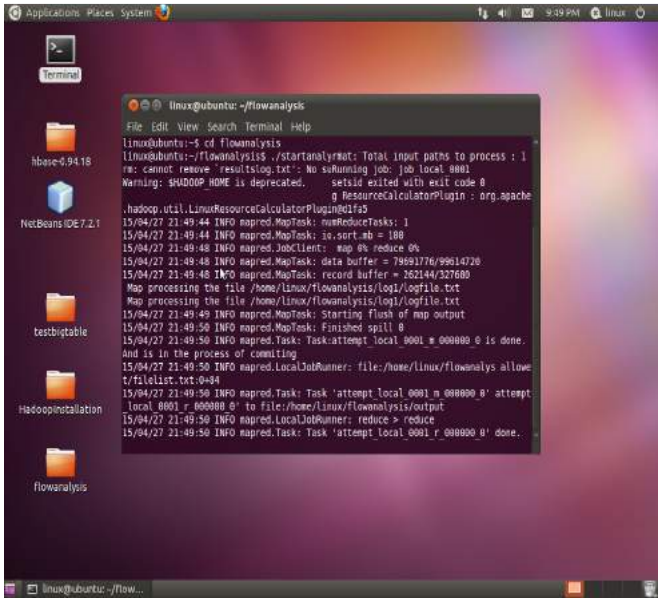

Fig. 5. Screenshot of Hadoop performing MapReduce

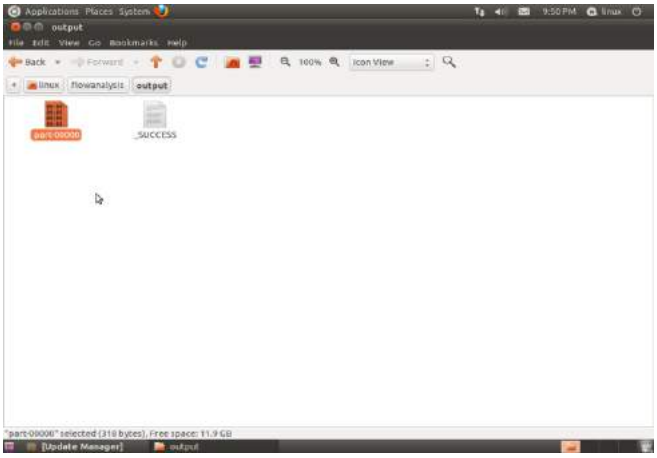

Fig. 6. Screenshot of output file generated

The output text file generated by the hadoop cluster is shown above which is named as part-00000. This output file consists of the summary of the flow of the packet information collected above. Here the packets from same source to same destination are grouped up and the also the byte information of the packets are added. The screenshot below shows the analyzing the output file and updating the flow.

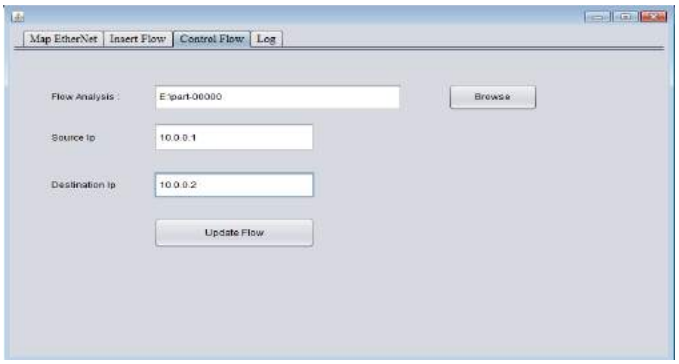

Fig. 7. Screenshot of analyzing output file and updating flow

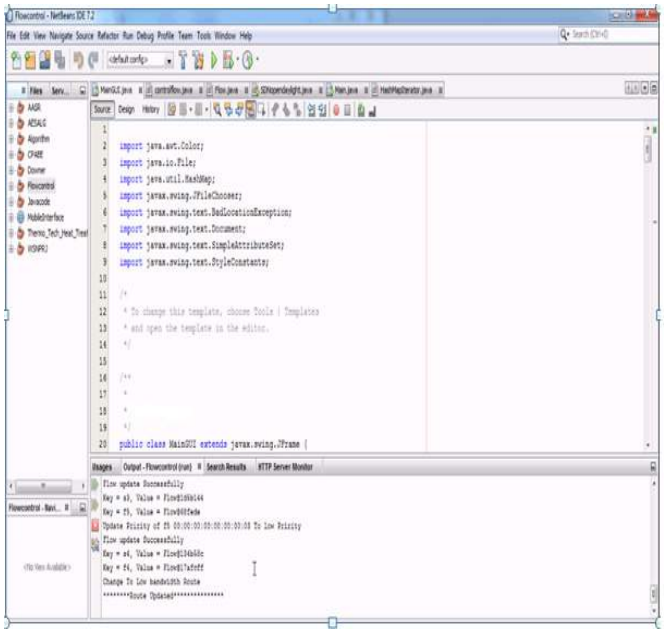

Fig. 8. Screenshot of updated flow

6. CONCLUSION

This paper gives the work on flow analysis and congestion control on the Hadoop platform. Hadoop is a popular framework which can process a huge amount of data just in few seconds. Here we have provided a detailed analysis on the classification of the packets based on the address, protocol and its bytes of the packets. In this system we show a flow control mechanism applied on the packets flowing in the network where the large amount of packets flow, a system to capture the flow of packets, hadoop cluster to perform analysis and also to avoid congestion.

REFERENCES

[1] M. Yu, L. Jose, and R. Miao," Software defined traffic measurement with opensketch," in Proceedings 10th USENIX Symposium on Networked Systems Design and Implementation NSDI, vol, 13, 2013.

[2] Scscc J. Shafer, S. Rixner, and Alan L. Cox, "The Hadoop Distribution Filesystem: Balancing Portability and Performance", in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement ACM 2010.

[3] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in Proceedings of the 10th ACM. SIGCOMM conference on Internet measurement. ACM, 2010, pp. 267–280.

[4] A. W. Moore and K. Papagiannaki," Toward the accurate identification of network applications," in Passive and Active network Measurement. Springer, 2005, pp. 41-54.

[5] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly", ACM SIGCOMM Computer Communication Review, vol 36, no. 2, pp. 23-26, 2006.

[6] Yuanjun Cai, Min Luo, "Flow Identification and Characteristics Mining from Internet Traffic using Hadoop" in 978-1-4799-4383-8/14/ at IEEE 2014.

[7] Apache Hadoop Website, http://hadoop. apache. org/