iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** I **Month of publication:** January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1457

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

A Big Data Approach for Traffic Classification using Data Mining

Srashti Chouhan[1], Savita Sisodia[2]
[1]*Research Scholar –CSE Truba College of Engineering and Technology*
[2]*Head-IT Truba College of Engineering and Technology*

Abstract: Retrieval of knowledge is the biggest task which needs full attention. For achieving the aim and objective of the desired work for traffic classification, knowledge discovery for network data is performed in college campus, which involves some steps which should be taken for carefully which leads to generate large network traffic data.
Network traffic classification generates large data volume, whose processing is not possible with traditional data mining concept. For fast processing of data execution is required in distributed or parallel manner which divides the entire work using fast processing into multiple parts. Using the MapReduce component, a Hadoop based application is developed with variable sizes and data sample collected from college campus. The collected data samples are the real world data sample which evaluates the performance of the work. For distributed and parallel processing and fetching results MapReduce component is used.
Keywords: HDFS, IET Campus Traffic Data, MapReduce, Hadoop

I. INTRODUCTION

Emerging technologies are the backbone of society which are playing a role of tremendous change in digital work. Network traffic are increasing and becoming complex due the emerging technologies, applications and services. This is the reason why managing and controlling of network is required, traffic analysis is done to manage request. Many methods are used to operate network traffic over a single server. Subsequently, if the traffic data increases, then memory, processing speed and storage capacity affects and limits the existing method. This increasing data which is of large amount is called as Big Data. Big data requires some specialized technique for the efficient classification.

A. Big Data

Big data is classified in three different ways, which is listed below:

1) Numerous databases
2) Data can not be divided on the basis of relational database.
3) Processing, capturing and generating of data is very fast. Big data is categorized into 3 V's, which are Volume, Velocity and Variety. These V's can be classified as follows :

B. Volume.

Parameters like streaming of data from social networking, data storage throughout the year which is transaction based, increasing sensors, unstructured data etc

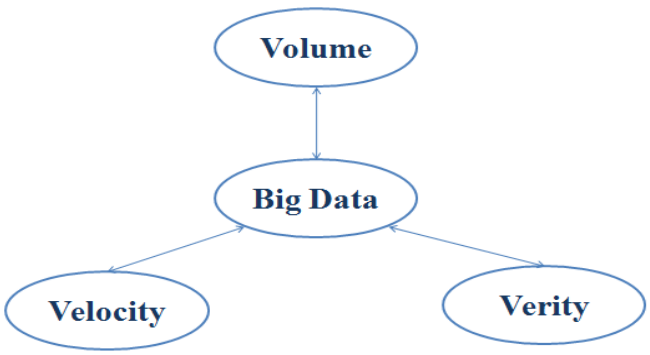


Figure 1.1 3V's of Big Data

C. Velocity.

With an unparalleled speed data is streaming and deals in a timely manner. Many of the techniques like smart metering, sensors and RFID are implemented to deal with near-real time data of large amount.

D. Variety

Data can be of any type in any format. It can be in structured or unstructured form. Traditional database classifies structured data as a numeric data. And unstructured data is classified as audio, video, text, mail, documents, stock, transaction history etc.

E. Network Classification

Into several ways Network classification can be achieved :

1) Protocols
2) IP address
3) Computer Nodes
4) Port Number

Network usage and consumption of bandwidth can be observed using Network Traffic. For enhancement of network and relevant applications pre-plan design is extended. Implementation of several topologies of network based on the amount of traffic in network in the system.

F. Network Traffic

Classification of Network traffic can be categorized as :

1) High Bandwidth Consumption Traffic and heavy traffic.
2) Working Hours consumption and Non-real Traffic.
3) Interactive Traffic can be defined as high response is expected. It can be poor due to absence of traffic classification.
4) Latency Sensitive Traffic : To competition for bandwidth. Following benefits are provided by organizations for properly analyzing network traffic :
5) Identification of bottlenecks in network : High amount of bandwidth is consumed by the applications or user, which is the important part of network traffic. To handle this type of issues several solutions are implemented.
6) Network security : Unwanted traffic and Unwanted amount of traffic is a way of for attack in network. Prevention from such attack can be valuable insight.
7) Network engineering : Future requirement can be analyzed based on the usage of network and knowing the amount of network required.

II. RELATED WORK

Hadoop Distributed Server and Big data parameters and dimensions are spotlighted in this chapter. Large data processing are analyzed on the basis of architecture and research paper. Y. Chen et al. In[1] Described that Big data is a good strategy to work and enhance for any business and organization.

Big data should be properly analyzed and systematically synthesized for the powerful intelligence business. 40% from the total population are connected, from them many are connected with the vast organizations, therefore increasing and generating data in every single second based on the survey of “Compression Algorithm in Hadoop”. P. Prakash et at. In[2] proposed about HDFS is a Hadoop Distributed File System which is described and with a scalable file system MapReduce works for system infrastructure, providing quick query access ans data management.

Configuring Hadoop creates unique nodes for managing and tracking data throughout the job tracker, name node, data node, client node. NS. Ghemawat et al. InData nodes are the central location where data is managed and stored and it consist of multiple database infrastructure which are scaled horizontally and computed across resources infrastructure. These repositories have enormous amount of data nodes. The architecture shown below is different from traditional infrastructure.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

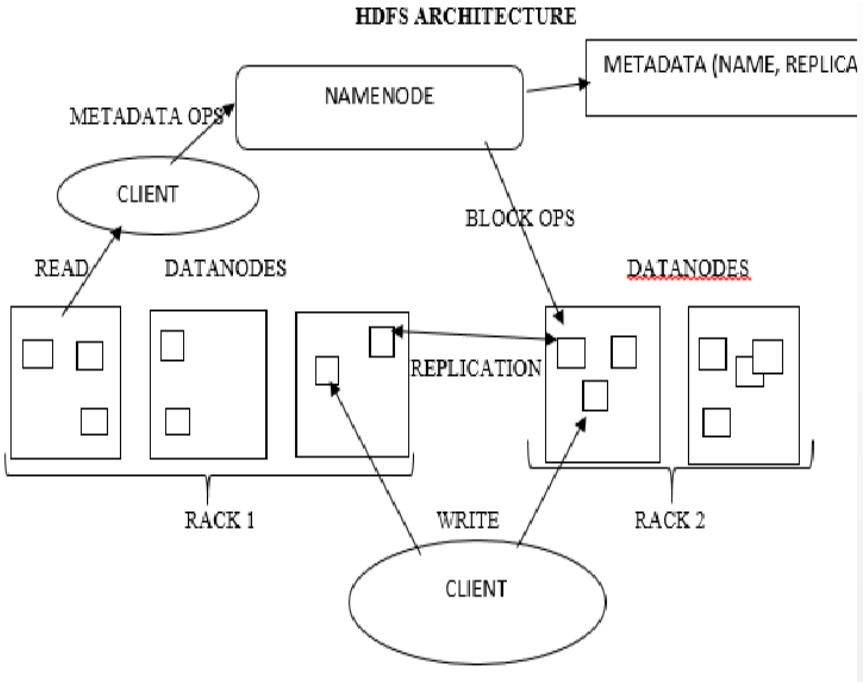


Figure 2.1 Hadoop Architecture

Table 1: Comparative Table

AUTHOR	TITLE	PROPOSED WORK
Y. Chen	Understanding TCP incast throughput collapse in datacenter networks,	Described that Big data is a good strategy to work and enhance for any business and organization.
P. Prakash	The TCP outcast problem: Exposing unfairness in data center networks	Proposed about HDFS is a Hadoop Distributed File System which is described and with a scalable file system MapReduce works for system infrastructure
S. Ghemawat	MapReduce: Simplified data processing on large clusters	Data nodes are the central location where data is managed and stored and it consist of multiple database infrastructure which are scaled horizontally and computed across resources infrastructure.

III. PROBLEM DOMAIN

A. Overview

The enhancement in technology increases computer dependency. Network establishes communication and is a collection of interconnected nodes. All the Business depends on Internet and network infrastructure to process and increase the growth of there business. Use of Network traffic and communication is increasing the heavy use of Internet based applications and services. Expectations of user increase with the rapid growth of solution and generate large data for processing and storing.

B. Detailed Problem Statement

In this work problem states that a solution should be developed to increase the performance of huge data processing and also should maintain security while storage, computation and transaction process. The analytical study of Big data concludes that much attention is provided to big data from many researchers and people belonging to IT sector because of its wide scope in digital environment applications. A research observed that 2 billion of the population uses Internet out of the 7.2 billion population. Moreover, many study observed that mobile technology is very expensive and is used by more then 5 billion people individually. So as a result, these users are generating large amount of data with the more and more use of mobile devices. This large amount of data is known as Big data. Researchers also express that by the year 2020 use of Internet will be increased upto 50 billion. The complete work indicates that development is required for processing large data because production will be higher in few coming years by 50%.

IV. METHODOLOGY USED

The proposed solution methodology is described in this chapter using some basic steps and implementations. By analyzing theoretically and in a systematic way methodology is described so that solution of methodology can be explored. With the supportive action systematic view is presented with the help of analyzed methodology. The required threshold are maintain by the theoretical analysis so that suitable actions can be explored. Entire scenario is considered in the proposed solution for the model proposed and there similar factors which is affective for the system performance. A research work is done in order to reduce the performance issue of large datasets for getting systematic solution for the classification of network traffic. In order to achieve network traffic classification, some steps has to be performed, and this knowledge discovery is performed using network data of college campus. Large network traffic is generated as a result, in which knowledge is explored and analyzed using special methods. The entire observation concludes the processing of large datasets in network traffic classification, so the traditional methodology for data mining can not be performed. Parallel and distributed execution is done to divide the entire task into multiple parts for the fast processing of data. Using MapReduce a Hadoop based application is developed, which classifies the the data sample size and time in college campus. To evaluate proposed solution performance live data sample is used. For fetching results, parallel processing and distribute, Map Reduce component is used. The entire work is developed in scenarios, and these scenario are listed below in two forms :

A. *Scenario 1:* To deploy the proposed implementation, configuration of single node machine using Hadoop server is done. All the execution is done on single machine; no slave partition process is involved in this scenario. Main focus in the implementation is the performance; performance of large and small data samples on single machine is measured.

B. *Scenario 2:* With the objective of high performance Hadoop computation environment is implemented for distributed and parallel processing, it is similar to scenario 1. For very fast result one master node and two slave nodes are implemented in it for distributive processing. Basic focus is the performance, performance of the proposed solution for multiple machine processing and large data processing. It also detect and measures the overhead on small data size of multiple machine.

V. PROPOSED WORK

A. Proposed Solution

He entire proposed solution is explained below as :

1) *Step1:* Database is considered as the initial step for any application to implement. So in our implementation of proposed solution database is considered as a initial step. All the network traffic is collected using the tool Wire-shark and for the variable time period data sets are prepared. Different data size are collected for variable time period and then proposed system is observed for different data size.
2) *Step2:* College campus is selected for the collection of data and during the selected hours sampling is done. It has been observed that in one hour approximately 300 MB of data is generated, this includes the entire observation during the implementation.
3) *Step3:* It supports .csv file format for different data sample files. The important information like time period, data size, destination address, source address, protocol type etc are constituted in it. Different transactions are done from different system address to demonstrate data sample. Load on different system and servers can be measured with the help of it and also traffic can be checked for particular time. The applications and protocols which are mostly used can be measured with the help of it.
4) *Step4:* For forming the group of large datasets into multiple parts clustering approach is used, it is used to form a group of relevant objects.

5) *Step5:* DBSCAN algorithm has been implemented using an API package of an apache commons Math 3.6.1. The respective clustering algorithm can be implemented using classes which it directly provides.
6) *Step 7***:** Clustering algorithm is used to demonstrate traffic approach using the single node and multi-node configuration.
7) *Step 8 :* Hadoop constitute of two major components which is MapReduce and HDFS

VI. RESULT ANALYSIS

Through the computation time performance of system analysis is done. It is classified into four sections and these sections are compared with data sample. Below given scenarios analyze the entire result in two scenario:
Scenario 1. Based on Hadoop environment, proposed solution is configured using single node.
Scenario 2. Based on Hadoop environment, proposed solution is configured using multi node.
Observation of the proposed solution using size of data sample after execution describing results for scenario 1 and 2.

Table 6.1 Sample Data with Number of Records

Sample Data Size	Number of Records(Rows)
29 MB	250039
112 MB	1008609
284 MB	2557544
642 MB	6121452

Table 6.2 comparative study for the configuration of single node(sec)

Process Name	29MB	112MB	284MB	642MB
Send	13	21	90	337
Receive	11	19	54	254
Merge	16	19	14	16
Dbscan	10	16	11	11
Total Time(Sec)	50	65	169	618

Table 6.3 Comparative study for the configuration of multi node(ms)

Process Name	29MB	112MB	284MB	642MB
Send	112	17	93	185
Receive	106	16	37	182
Merge	106	18	15	13
DBSCAN	106	10	12	10
Total Time(Sec)				

VII. CONCLUSION

The approach used is effective in future because of preservation, protection, privacy and mechanism. Implementation and configuration justifies the performance using desirable security. The entire work concludes about network, protocol and trends with determined load on every single IP address. The concluded work observes realistic information and also increases the performance for large datasets. For data size of 642 MB in single node, multi node is observed with 40% overhead reduction.

VIII. FUTURE WORK

Proposed solution analyzed better solution comparing with traditional system. Future implementation can overcome the observed limitations, which is the future scope of the proposed solution. Authentication needs username and password in concluded work, which results in vulnerability, leads to security attack. Future implementation can be done as more enhancement in authentication case. RSA algorithm is used in our work for encryption, this leads to the enhancement in overheads and heavy size of cipher text. So another future scope is reduction in overhead.

REFERENCES

[1] Y. Chen, R. Griffith, J. Liu, R. H. Katz, and A. D. Joseph, “Understanding TCP incast throughput collapse in datacenter networks,” in Proceedings of the 1st Workshop on Research on Enterprise Networking, ser. WREN ’09. New York, NY, USA: ACM, 2009, pp. 73–82. [Online]. Available: http://doi.acm.org/10.1145/1592681.1592693

[2] [2] P. Prakash, A. Dixit, Y. C. Hu, and R. Kompella, “The TCP outcast problem: Exposing unfairness in data center networks,” in Proceedings of the 9th Conference on Networked Systems Design and Implementation, ser. NSDI’12. Berkeley, CA, USA: USENIX Association, 2012, pp. 30–30. [Online]. Available: http://dl.acm.org/citation.cfm?id=2228298.2228339

[3] [3] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” in Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation, ser. OSDI’04. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10. [Online]. Available:http://dl.acm.org/citation.cfm?id=1251254.1251264.

[4] [4] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, “Data center TCP (DCTCP),” in Proceedings of the SIGCOMM 2010 Conference, ser. SIGCOMM ’10. New York, NY, USA: ACM, 2010, pp. 63–74. [Online]. Available: http://doi.acm.org/10.1145/1851182.1851192

[5] [5] Y. Chen, R. Griffith, J. Liu, R. H. Katz, and A. D. Joseph, “Understanding TCP incast throughput collapse in datacenter networks,” in Proceedings of the 1st Workshop on Research on Enterprise Networking, ser. WREN ’09. New York, NY, USA: ACM, 2009, pp. 73–82. [Online]. Available: http://doi.acm.org/10.1145/1592681.1592693

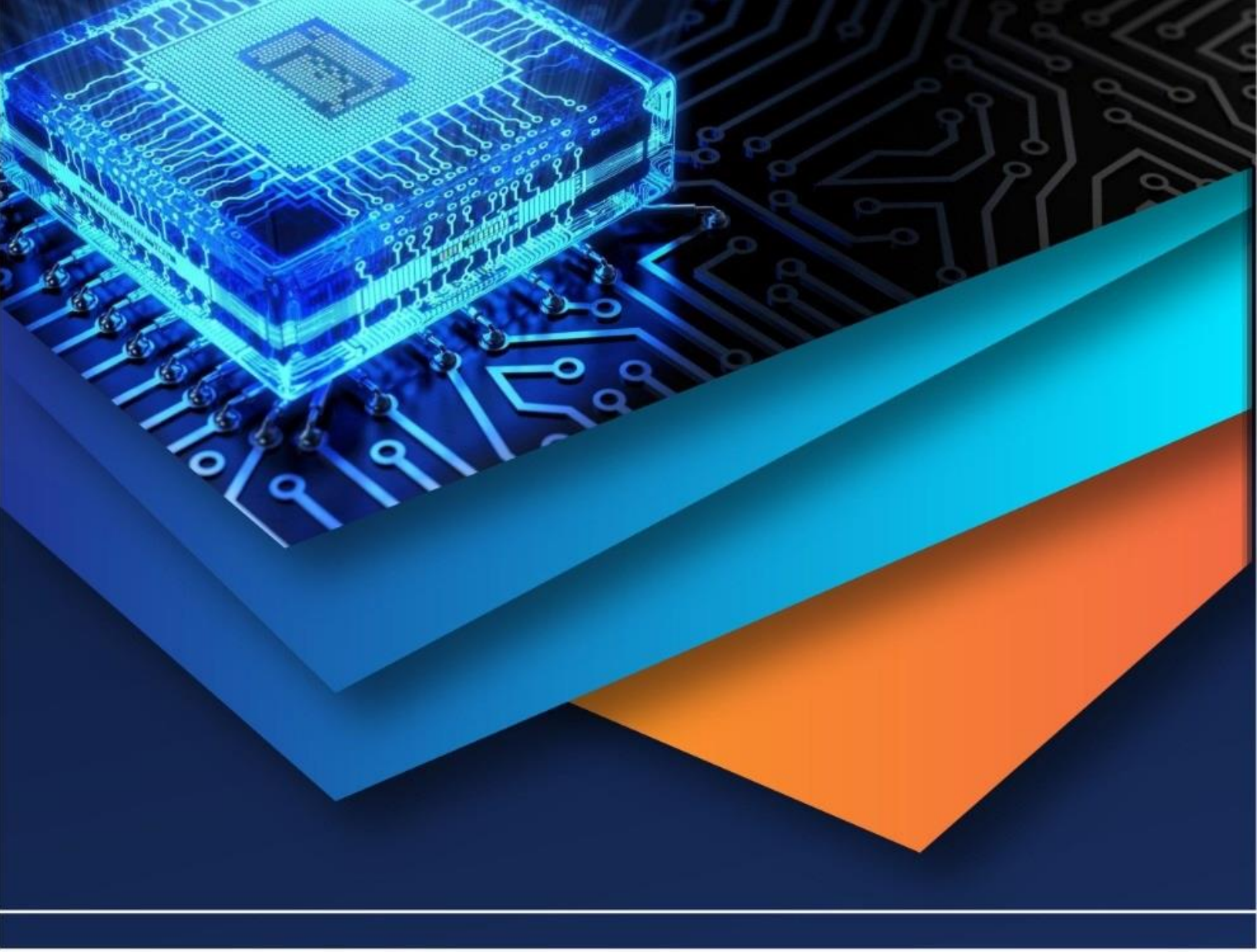
[6] [6]. S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoidance,” IEEE/ACM Transactions on Networking, vol. 1, no. 4, pp. 397–413, Aug 1993.

[7] [7]. K. Nichols and V. Jacobson, “Controlling queue delay,” Queue, vol. 10, no. 5, pp. 20:20–20:34, May 2012. [Online]. Available:http://doi.acm.org/10.1145/2208917.2209336

[8] [8]. C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, “Improving datacenter performance and robustness with multipath tcp,” in Proceedings of the ACM SIGCOMM 2011 Conference, ser. SIGCOMM ’11. New York, NY, USA: ACM, 2011, pp. 266–277. [Online]. Available: http://doi.acm.org/10.1145/2018436.2018467

[9] [9]. S. N. Ismail, H. A. Pirzada, and I. A. Qazi, “On the effectiveness of codel in data centers,” Tech. Rep.

[10] [10] P. Rygielski, S. Kounev, and S. Zschaler, “Model-based throughput prediction in data center networks,” in 2013 International Workshop on Measurements and Networking Proceedings (M&N). IEEE, Oct 2013, pp. 167–172.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY