IJFEAT
INTERNATIONAL JOURNAL FOR ENGINEERING APPLICATIONS AND TECHNOLOGY

PLAGIARISM DETECTION ON BIGDATA USING APACHE STORM WORKING ON THE CIA PRINCIPLE

Priyanka V. Narkhede[1], Radhika V. Mahalle[2] , Priya A. Lokhand[3]
Student, Department of CSE, DES'sCOET Dhamangaon Rly, Maharashtra, India, priyankan364@gmail.com
Student, Department of CSE, DES'sCOET Dhamangaon Rly, Maharashtra, India, radhikavmahalle@gmail.com
Student, Department of CSE, DES'sCOET Dhamangaon Rly, Maharashtra, India, priyalokhande71@gmail.com

Abstract
Plagiarism is the act of copying another person's data such as text documents, pictures, ideas or any creative contents and represents it as one's own without giving citation to the main author. Plagiarism occurs in every field such as in schools, colleges as well as novels, scientific papers, art designs and source code. Plagiarism has a wider meaning, paraphrasing someone else's texts by replacing a few words by synonyms or interchanging some sentences in own way is also plagiarism. As the lots of data available on the internet and the use of smart phones increases the availability of information which causes the occurrence of plagiarism. In this paper, we are using SCAM algorithm for comparing similarity within data. We have implementing the SCAM algorithm for plagiarism detection in the big data which works with the Apache storm for better performance and speed. The Apache Storm is an open source computation system, simple and can be used with any programming language. The data security is most important everywhere. CIA principle is
Used for the data security.

Index Terms: ***Apache Storm, Bigdata, CIA, Plagiarism, SCAM***

---

1. INTRODUCTION

Plagiarism is an ancient art of fraud. Plagiarism word is derived from the two words plagiarius and plagiare. The word plagiarism means abductor one, such as kidnapper, who abducts and the plagiare means to steal. Plagiarism is nothing but presenting someone's ideas, computer programmers, thought, text documents, pictures, videos, etc. For detecting the plagiarism to check the similarity score between the documents the SCAM (Standard Copy Analysis Mechanism) algorithm is used. Here the SCAM algorithm is modified which works on Hadoop Framework. Big Data is very vast and large amount of data which is very complex for processing with traditional computing techniques. As the amount of data produced by the generation increases, the size of big data growing rapidly every year. The Bigdata has the 3V characteristics i.e. Velocity, Volume and Variety. The speed of change and updating is fast in bigdata. The volume of bigdata rises rapidly. The bigdata is in 3 varieties i.e. structured, unstructured and semi structured. To handle this big data the Hadoop Map Reduced based algorithm is used which can be used for processing of large data sets. It provides the
distributed storage and computation can be performed on the clusters of computer. Hadoop uses two main modules for processing the data HDFS and MapReduce. The big data is data produced by the various applications and servers. For processing the big data hadoop is used. But the server's bandwidth of hadoop is less. For the plagiarism Detection System we have to locate all web servers through DNS. Thus for speed and scalability we use the apache storm. Apache storm is a lightning fast cluster computing designed for fast computation. Apache storm is designed for processing the large amount of data which is very complex and big for processing. The use of Apache Storm is beneficial because it is fault tolerant and fast, flexible and it supports any programming language. The biggest advantage of using Apache storm is that if the supervisor or nimbus dies it doesn't affect the whole system; it will continue the execution of system from where it previously stopped. Some components of storm are tuple, stream, bolts and spouts. These are also called as the processing primitives of storm. The tuple is main data structure in the storm which supports all data types. It is a list of elements. Stream consists of an unordered sequence of tuples. Spouts is the source of stream for reading data from or taking input from the data sources like Apache Kafka queue, Twitter Streaming API, Kestrel queue, etc. Bolts are the processing units. It process the data send by spouts and produce an output stream. Bolts also perform the various operations. For implementing tspouts and bolts the 'ISpout' and 'IBolt' interfaces are used respectively. The zookeeper framework used in Apache architecture provides simple interface and services. Security isn't a piece of software; it's a property of entire system, including its users. To get closer to the goals of security a simple but

widely-applicable security model is the CIA triad; standing for Confidentiality, Integrity and Availability; three key principles which should be guaranteed in any kind of secure system. Confidentiality is the ability to hide information from those people unauthorised to view it. Integrity is the ability to ensure that data is an accurate and unchanged representation of the original secure information. One type of security attack is to intercept some important data and make changes to it before sending it on to the intended receiver. The Integrity comes with some properties such as Accuracy, Audibility, Verifiability, Traceability, etc. It is important to ensure that the information concerned is readily accessible to the authorised viewer at all times. For getting the security of the data the CIA principle is used in the system.

2. SYSTEM DESIGN

The plagiarism detection system is designed by using the SCAM which the measure to calculate the similarity percentage between the documents. The SCAM algorithm compares the documents as test document and original document and gives the percentage of similarity between both the documents. The SCAM algorithm is modified so that it works on topology based and for making it capable to process the big data. The Apache Storm works by making the topologies. After making the topology when the topology gets submitted it processes the topology and collect all the tasks which are carried out and the order in which all the tasks get executed. To make the system compatible with every server's bandwidth the Apache Storm is used. The use of Apache Storm gives the faster processing and scalability. Zookeeper framework in Apache storm allows distributed processes to coordinate with each other through a shared hierarchical name space of data registers, known as znodes. The core abstraction in storm is the "stream". The process of detecting plagiarism includes the some stages such as collection, analysis, investigation and confirmation, etc. In final step we evaluating similarity with scam in this step scam formula are used to detect over lap and irrespective of the differences in document sizes. Topology in Apache storm is similar to MapReduce jobs. In a topology each node contains processing logic. Links between nodes represent the data flows between those processing links. To create and destroy the topology in the Apache form the two commands are used which are described below.

Running a topology: To submit a new topology the following command is used.

Storm jar all-my-code.jar backtype.store.MyTopology arg1 arg2

Killing a topology: Once the topologies are processed to destroy that topology the following command is used.

Store kill (topology name)

2.1 Purposed work

1. The Nimbus daemon distributes code to worker nodes.
2. The logic processed by worker is specified in topology.
3. A spout read the number of tuples and emits them as a stream.
4. The stream will get distributed to all the bolts for the further processing.
5. Bolts then performs the processing tasks and send the tasks to other bolts.
6. The nodes will match the input data and produce the result.

Apache storm will work in this algorithm for the faster speed of processing the data. In the topology of storm each node executes in a parallel. The developer can specify how much parallelism is wanted for each node, and then storm will spawn the number of threads across the cluster for the execution. The CIA framework used in the system ensures the security of the data.

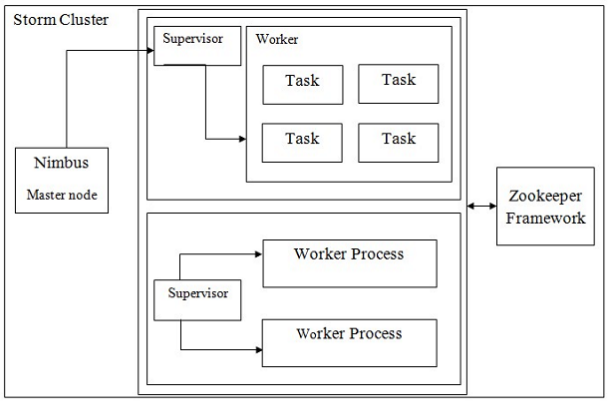

Fig-1: Architecture of Apache Storm

A working Storm cluster should have one nimbus and one or more supervisors. Another important node is Apache Zookeeper, which is used for coordination within the nimbus and the supervisors. In the workflow of Apache Storm initially, the nimbus waits for the "Storm Topology" to be submitted to it. Once a topology get submitted, it will process the topology and gather all the tasks that are carried out and the order in which task is executed. Then, the nimbus will distribute those tasks to all available supervisors. After some particular interval, all supervisors send the heartbeats to thenimbus to inform that they are alive. If the supervisor doesn't send the heartbeats to the nimbus or it get dies while execution, then the nimbus will assign that task to another supervisor. The biggest advantage of Apache Storm is that if the nimbus itself dies, it doesn't affect the whole system; Supervisors will work the assigned tasks. After completing the assigned tasks

supervisors waits for the new task for come in. In the meanwhile of execution, the dead nimbus will be restarted automatically by services monitoring tools. If the nimbus fails, it get restarted nimbus will continue from where it stopped. Similar to that the supervisors also get restarted automatically. Apache Storm guaranteed to process all the tasks at least once. As all the topologies are processed, the nimbus waits for a new topology to arrive for the execution. There are two modes in storm cluster i.e. Local mode and Production mode.

3. CONCLUSION

The developed plagiarism detection system in bigdata can be used in various applications. Here the SCAM algorithm is modified for distributed computing platform. We have selected the Apache Storm platform in order to analyze the big data. The use of Apache Storm increases the scalability and the speed of system. It gives the faster processing of data and scalability. The use of CIA principle provides the security to the data. This technique takes sometimes for finding results gives output in short time with speed and accuracy and we are easily process and handle big data sets. So Apache Storm is used for performance enhancement

REFERENCES

[1] Jayshree Dwivedi, Prof. Abhigyan Tiwary ICIMIA 2017) Plagiarism Detection on Bigdata Using Modified MapReduced Based SCAM Algorithm 978-1-5090-5960-7/17/$31.00 ©2017 IEEE.

[2] Plagiarism Detection Based on SCAM Algorithm Daniele Anzelmi, Domenico Carlone, Fabio Rizzello, Robert Thomsen, And D. M. Akbar Hussain.

[3] Plagiarism Detection-Different Methods and Their Analysis: Review S.A.Hiremath, M.S.Otari International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 7 (August 2014) http://ijirae.com.

[4] Robert Evans Apache Storm a Hands on Tutorial 978-1-479982189/15$31.00DOI10.1109/IC2E.2015.67 © 2015 IEEE.

[5] Jan Spike van der Veen, Bram van der Waaij, Elena Lazovik, Wilco Wijbrandi, Robert J. Meijer. Dynamically Scalin Apache Storm for the Analysis of Streaming Data 978-1-4799-8128-1/15 $31.00 © 2015 IEEE.

[6] Sylvia Killinen (2016)Using the Principles of the CIA Traid to Implement Software Security. © 2011-2017

[7] Saman Shojae Chaeikar, Mahammadreza Jafari, Hamed Taherdost, Nakisa Shojae Chaei kar Definitions and Criteria of CIA Security Triangle in Electronic Voting System. Vol. 1, No.1, October 2012, Page: 14-24, ISSN: 2296-1739 © Helvetic Editions LTD, Switzerland

[8] www.elvedit.com