Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

An Efficient Method for Enhancing Visualization and Performance of Unstructured Big Data

Shivam Singh
IT Professional-II
ICAR-Indian Agricultural Statistics Research Institute (IASRI)
New Delhi, India

***Abstract*- The newest wave of Big Data is generating new opportunities and new challenges for businesses across each trade. The challenge of information integration, incorporating knowledge from social media and alternative unstructured knowledge into standard BI surroundings is one in all the foremost imperative problems. As knowledge is increasing day by day, in several sites likes IRCTC, Facebook etc., presence of databases and database management system isn't enough. We'd like to prepare and obtain the important info from the pool of information i.e. big data. Massive knowledge could be a term for knowledge sets that area unit therefore giant or advanced that ancient processing application software package area unit inadequate to alter them. The term "big data" usually refers merely to the utilization of prophetical analytics, user behavior analytics, or sure alternative advanced knowledge analytics strategies that extract price from knowledge, and rarely to a selected size of information set. In this paper we are going to analyze the big data and create that data into a user friendly image. Because the data is therefore massive, it's terribly tough to search out easy visualizations.**

Keyword-Big Data, HDFS, Hive, Flume.

1. INTRODUCTION

Big data implies truly a major information; it is a gathering of vast datasets that can't be handled utilizing conventional processing systems. Huge information is not only information; rather it has turned into an entire subject, which includes different tools, strategies and structures. It is for the most part characterized as high volume, speed and assortment data resources that request financially savvy, inventive types of data preparing for improved understanding and decision making. Benefit of Big Data is Real-time enormous information isn't only a procedure for putting away petabytes or exabyte of information in an information stockroom, It's about the capacity to settle on better choices and take important activities at the ideal time.

1.1 HDFS

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing. Features of HDFS

1. It is suitable for the distributed storage and processing.
2. Hadoopgives a charge interface to communicate with HDFS.
3. The implicit servers of name node and data node help clients to effortlessly check the status of bunch.
4. Streaming access to file system data.
5. HDFS gives data authorizations and confirmation.

1.2 HIVE

Apache Hive is an data distribution center framework based on top of Hadoop for giving information synopsis, question, and analysis. Hive gives a SQL-like interface to question information put away in different databases and document frameworks that coordinate with Hadoop. Customary SQL questions must be actualized in the Map ReduceJava API to execute SQL applications and inquiries over distributed data. Hive gives the vital SQL deliberation to incorporate SQL-like Queries (HiveQL) into the basic Java API without the need to actualize questions in the low-level Java API. Since most data warehousing applications work with SQL-based questioning dialects, Hive underpins simple convenience of SQL-based application to Hadoop.

1.3 FLUME

Apache Flume is a tool/service/data ingestion system for gathering accumulating and transporting a lot of spilling data, for example, log files, occasions (and so on.) from different sources to a concentrated information store. It is an exceedingly dependable, dispersed, and configurable tool. It is primarily intended to duplicate spilling data (log data) from different web servers to HDFS. A portion of the striking components of Flume are as per the following:

1. Flume ingests log data from various web servers into an incorporated store (HDFS, HBase) proficiently.
2. Using Flume, we can get the information from numerous servers promptly into Hadoop.
3. Flume backings an extensive arrangement of sources and destination type.
4. Flume backings multi-bounce streams, fan-in fan-out streams, relevant directing, and so on.
5. Flume can be scaled horizontally.

2. RELATED WORK

A different amount of research work has been done to increase the scalability and visualization of unstructured big data. *Vibha Bhardwaj et.al and Rahul Johari et.al.*They demonstrates that how ordering procedures like Map Reduce are connected to Big Data and its relevance in numerous applications like, re-arranged word, Mutual Friend Problem, Word Count and so forth and the other overall utilizations of big data are additionally examined

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

.For the purpose of fulfillment an expansive correlation of various data mining methods too has been displayed.
*Nishant Agnihotri et.al.*and*Aman Kumar Sharma et.al.*This paper proposes an approach for breaking down continuous surges of information from web-based social networking for particular parameters of customer's proposals for development of administrations and fulfillment from the administrations gave.
*Omesh Kumar et.al.*and*Abhishek Goyal et.al.*In this work, they are extending the data visualization capacity of enormous data examination results. We are attempting to demonstrate the outcomes in a vastly improved manner to enhance the basic leadership emotionally supportive network. The better the visual result of results, the better and compelling is the utilization of result outcomes. We are attempting to improve an incorporation with enormous dataset comes about for better utilization of huge data analytics.
Qunchao Fu et.al. and, Wanheng Liu et.al. In this paper, there is a data handling strategy for visual showing on huge data. It givesa data diminishment technique to support visual outlines for enormous information, and check the plan space is expansive plot of multidimensional data. For communication, to precompute multivariate data tiles and parallel handling.
Harneet Kaur et.al. and *Jing (Selena) He et.al.* In this their mobile application improvement separates the tweets and dissects the tweets. This gets more thoughts and furthermore advances those thoughts and sentiments through informal organizations where it achieves most extreme number of individuals in least time. Their application can be a basic framework for the researches to work on the real data analysis projects.

3. PROPOSED APPROACH

First of all we are having three types of data as follows:

1. Semi-structured data is a type of organized data that does not accommodate with the formal structure of data models related with social databases or different types of data tables, however in any case contains labels or different markers to isolate semantic components and uphold chains of importance of records and fields inside the data.
2. Structured data refers to kinds of data with an abnormal state of association, for example, data in a social database. At the point when data is exceedingly organized and unsurprising, web crawlers can all the more effectively compose and show it in inventive ways.
3. Unstructured information (or unstructured data) refers to data that either does not have a pre-characterized information show or is not sorted out in a pre-characterized way. Unstructured data is normally message substantial, yet may contain information, for example, dates, numbers, and certainties too.

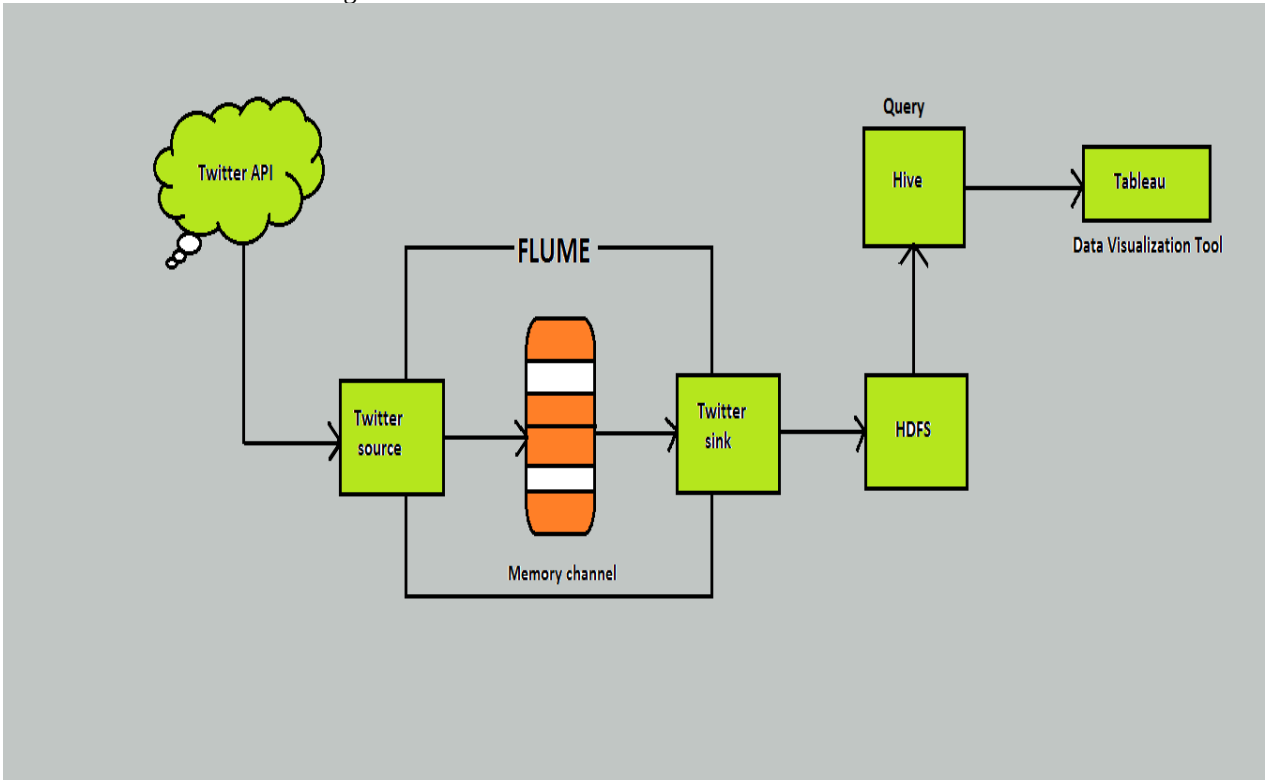

Fig-1: Overall block diagram

In this overall block diagram flume is a tool which is used to fetch the data from the web source like twitter with the help of agent which is situated at the source as agent fetch the data and transfers it to the memory channel and then that data is stored in the hdfs .The data is always stored in the form of log files. Hive helps in solving the query.
Tableau is a Business Intelligence tool for visually analyzing the data. Clients can make and disseminate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

intelligent and shareable dashboards which portray the patterns, varieties and thickness of the information in type of diagrams and outlines. Tableau can associate with records, social and Big data sources to secure and handle information. The product permits information mixing and continuous coordinated effort, which makes it exceptionally remarkable. It is utilized by organizations, scholarly analysts and numerous legislatures to do visual information examination.

Streaming/Log Data-Generally, most of the data that is to be analyzed will be produced by various data sources like applications servers, person to person communication destinations, cloud servers, and endeavor servers. This information will be as log files and events. Basically a log file is a file that lists events/activities that happen in a working

HDFS put command-The principle challenge in dealing with the log data is in moving these logs delivered by various servers to the Hadoop environment. Hadoop File System Shell gives charges to embed information into Hadoop and read from it. We can embed information into Hadoop utilizing the put command as demonstrated as follows ($ Hadoop fs –put /path of the required file /path in HDFS where to save the file).

Failure Handling -In Flume, for every event, two exchanges happen: one at the sender and one at the beneficiary. The sender sends events to the collector. Not long after subsequent to getting the data, the beneficiary confers its own exchange and sends a "got" signal to the sender. After getting the signal, the sender confers its exchange. (Sender won't submit its exchange till it gets a signal from the collector.)

3.1 Proposed Algorithm

1. To configure Flume, we have modify three files namely, flume-env.sh, flumeconf.properties, and .bash.rc.
2. To verify the installation of flume we will open the bin folder and then type the command (. /flume-ng.).
3. After installing flume we have configure the configuration file which is a java property file having key- value pairs .We have pass the values to the keys in the file before this we have to give the name to the source, sink and channel.
4. Then after this we have develop a twitter application name as Twitter Data according to which we have generated the access token, consumer key, consumer secret, access token secret all these values will help in configure the file.
5. After configuration we have started the flume agent by the command ($ bin/flume-ng agent –conf. / conf / -f conf / twitter. Conf –Dflume . root. Logger =DEBUG, console –n TwitterAgent).
6. Now after this we have started Hadoop by browsing sbin directory of Hadoop and started yarn and Hadoop distributed file system by the command (cd /$Hadoop_Home/sbin/ $ start-dfs.sh /$ start-yarn.sh).
7. After this we have created the directory in HDFS by using the command mkdir.
8. After creating the directory we have given the fetching command so as the data can be fetched from twitter for that we uses the command ($ cd $FLUME_HOME $bin/flume-ng agent –conf . / conf/ -f conf/twitter . conf – Dflume logger=DEBUG, console –n TwitterAgent).
9. At last we will verify the HDFS by accessing the Hadoop Administration Web UI using the URL (http://localhost:50070/).
10. After fetching the data we will visualize that data with the help of tableau.

4. RESULT

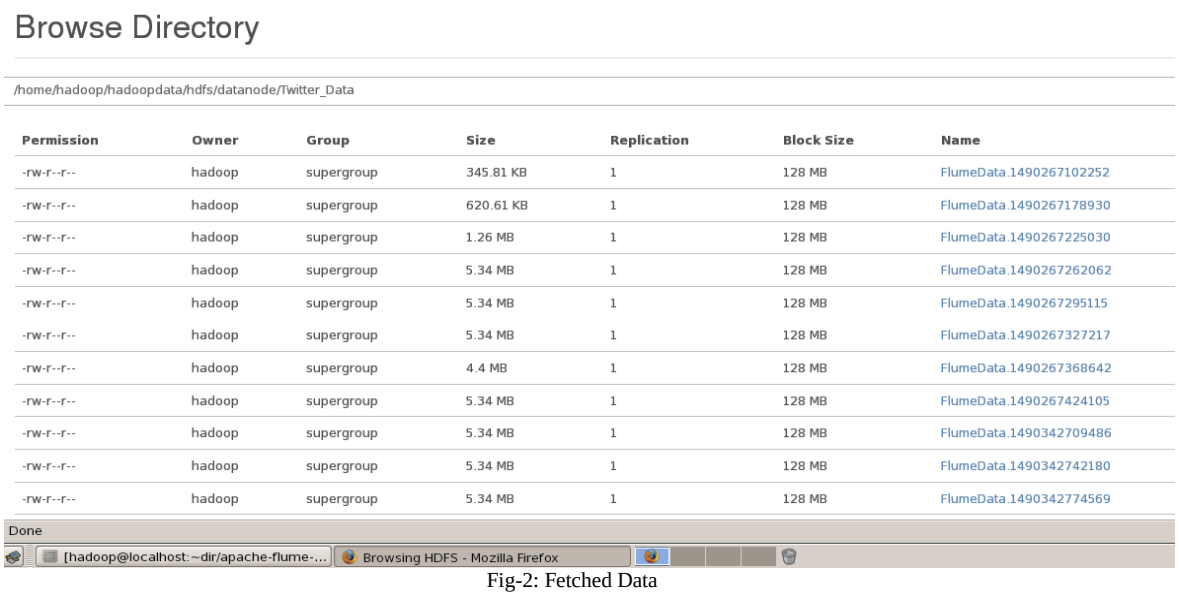

Fig-2: Fetched Data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

In this figure the Twitter data has been fetched with the help of flume. According to that we will make this data into a user friendly image.

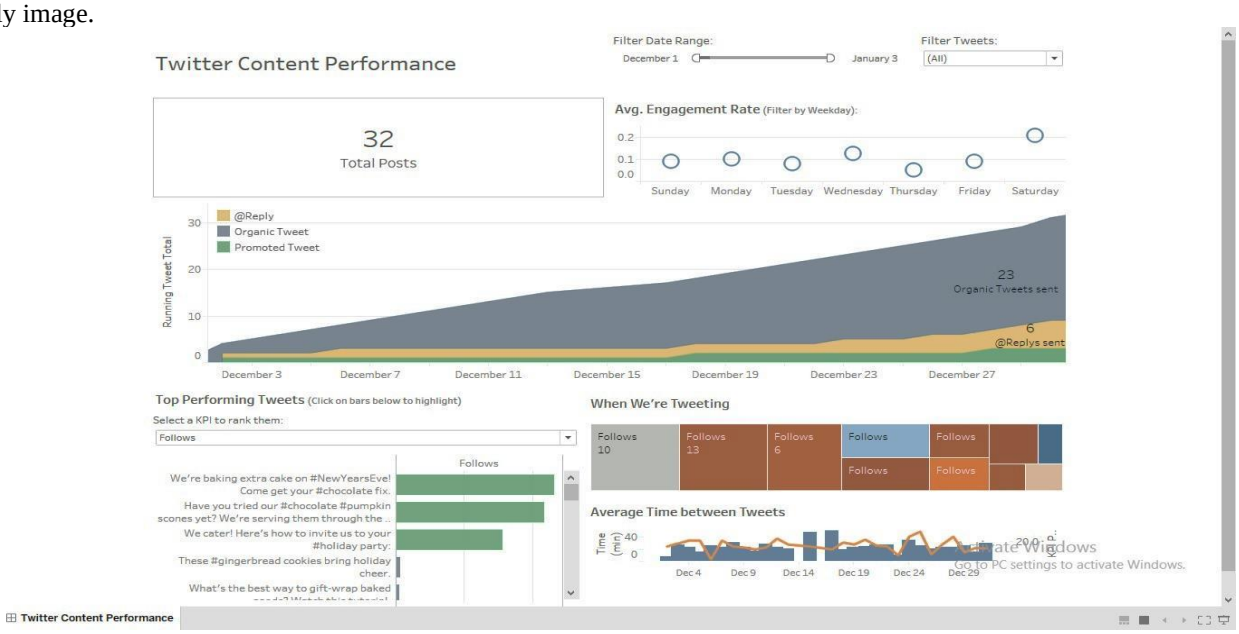

Fig-3: Visual output of Twitter data

Graphical results will be filtered on the basis of different parameters. This all gives a like analysis capability to our tool. This all gives effective results to decision making team.

In our visualization tool, we have taken the data of twitter. Through the graphical representation in tableau we analyzed the total number of posts, reply, organic and promoted tweet. We have also analyzed when we are tweeting and average time between the tweets and also filter the data range by week days, by hours.

5. CONCLUSION AND FUTURE WORK

As twitter post are very important source of opinion on different issues and topics. It can give a keen insight about a topic and can be a good source of analysis. Analysis can help in decision making in various areas. It renders unstructured tweets in a tabular format for easier management. Apache Hadoop is one of the best options for twitter post analysis. Once the system is set up using FLUME and HIVE, it helps in analysis of diversity of topics by just changing the keywords in query. Also it does the analysis on real time data, so is more useful. The analysis will helpful in finding people mood for election voting and can be helpful in strategy planning. Opinion mining can also be done on that data for finding polarity (Positive, Negative, Neutral) of tweets collected.

Future scope for Big Data and analytics are:

1. Visual data discovery tools will be growing 2.5 times faster than rest of the Business Intelligence (BI) market. By 2018, investing in this enabler of end-user self-service will become a requirement for all enterprises.
2. Over the next five years spending on cloud-based Big Data and Analytics (BDA) solutions will grow three times faster than spending for on premise solutions. Hybrid on/off premise deployments will become a requirement.

REFERENCES

[1] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, "The rise of big data on cloud computing: Review and open research issues", *ELSEVIER*, 2015.

[2] J. Singh, V Singla, "Big Data: Tools and Technologies in Big Data", *International Journal of Computer Applications*, 2015.

[3] M. S. Aravinth, M. S. Shanmugapriyaa, M. S. Sowmya, Arun, "An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing", *International Journal for Innovative Research in Science and Technology*, pp. 252-255, 2015.

[4] M. Chen, S. Mao, Y Liu, "Big data: A survey", *Mobile Networks and Applications Springer*, vol. 19, no. 2, pp. 171-209, April 2014.

[5] S. R. Qureshi, A Gupta, "Towards efficient Big Data and data analytics: A review", *IEEE International Conference on IT in Business Industry and Government (CSIBIG)*, pp. 1-6, March 2014.

[6] P. Bedi, V. Jindal, A Gautam, "Beginning with Big Data Simplified", *IEEE International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pp. 1-7, 2014.

[7] A. Pal, S Agrawal, "An experimental approach towards big data for analyzing memory utilization on a Hadoop cluster using HDFS and MapReduce", *IEEE First International Conference on Networks & Soft Computing (ICNSC)*, pp. 442-447, August 2014.

[8] L.M Patnaik, "Big Data Analytics: An Approach using Hadoop Distributed File System", *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, pp. 239-243, May 2014.

[9] J. Zhang, M. L. Huang, "5Ws model for bigdata analysis and visualization", *IEEE 16th International Conference on Computational Science and Engineering*, pp. 1021-1028, 2013.

[10] DunrenChe, MejdlSafran, and ZhiyongPeng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013.

[11] Daniel Keim et al., "Big-Data Visualization", *IEEE Computer Society*, 2013.

[12] Petra Isenberg et al., "Data Visualization on interactive surfaces: A Research Agenda", *Technical report*.

[13] Alpa, "Data Visualization for University Research papers", *International Journal of Soft Computing and Engineering*, vol. 2, no. 6, January 2013.