Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

STREAMING AND ANALYSIS OF TWITTER DATA USING APACHE HADOOP ECOSYSTEM AND BIGINSIGHTS

N.V.MUTHU LAKSHMI[1], J. S. MANI JANGAM[2], K. SANDHYA RANI[3]
[1]Assistant Professor, Department of Computer Science, SPMVV, Tirupati, A.P.,India
[2]Research Scholar, Department of Computer Science, SPMVV, Tirupati, A.P. ,India
[3]Professor, Department of Computer Science, SPMVV, Tirupati, A.P. ,India
[1]nvmuthulakshmi@yahoo.co.in, [2]jsmanij@gmail.com,[3]sandhyaranikasireddy@yahoo.co.in

ABSTRACT

At present the whole mankind lives in a world that is "drowning in data" and this data is continuously generated on massive scale mostly by online interactions among people, by the transactions taking place between people and systems, also by sensor enabled instrumentation called Big data. Big data is high in volume, velocity and also high in variety that is, it can be a structured, semi-structured, or un-structured data. Big data can reveal the issues hidden by data that is too costly to process and perform the analytics such as user's transactions, social and geographical data issues faced by the industry. It is difficult to process and stream the big data within the specified resources. Using online streaming tool of Big data eco-system, big data(eg: twitter data) can be collected and stored on clusters of commodity servers later it could be analyzed, visualized by using analytics and visualization tools such as hunk, BigInsights and so on. Twitter is one of the largest social media website where people around the world would share and respond their opinions on various topics in the form of tweets. Millions of tweets are exchanged every day, zettabytes per year. This massive data is considered as a Big data and can be used for industrial or business purpose after organizing as per the need and processing. This paper addresses how the tweets are streamed, processed, analyzed and visualized using Apache Hadoop and other tools in a distributed fashion.

Keywords: Big Data, Hadoop, MapReduce, BigInsights, Twitter Data.

1. INTRODUCTION

Today's organizations are facing growing challenges from their business perspective especially their urge for value need to be obtained from huge amount of data generated and the complexity of data which is both structured, semi-structured and unstructured. Big data is the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers [1] in a reasonable amount of time.

The first organizations to embrace Big data were on-line and startup firms. Firms like Google, Twitter, Wallmart, eBay, LinkedIn, and Facebook were built around Big data from the beginning. Big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task [2].

As per the statistics of industries in real time, there is 2.5 million items added per minute by every individual. In the same way 300,000 tweets, 200 million emails, 220,000 photos are generating per minute, and other enterprises like RFID's 5 TB and is greater than 1PB data for gas turbines producing per day. In the year 2012 this whole data is 2.8 zettabytes only, now i.e. in 2015 it is increased to 20 zettabytes and by 2020 this number may reaches to 40 zettabytes. In this world, 90% of data is unstructured and which becomes difficult to stream to process for enterprises. So organizations and individual firms need deeper insight to overcome this problem.

As per several definitions, Big data is the data which requires more processing capability than the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit into the structures of database architectures. To gain value from big data, an alternative approach is required to process [3]. Among many Big data technologies, Hadoop is popular and widely used to address the Big data challenges. There are various platforms which provide Hadoop for enterprises to stream their data like Apache Hadoop, IBM's BigInsights, Microsoft's Azure HD Insights, cloudera tools and hortonworks etc. All these tools perform various functionalities for analyzing the data based on the different problem domains.

Characteristics of Big Data

The three characteristics of Big data are 3V's: Volume, Velocity, and Variety. The following figure shows the structure of 3V's.

Volume

It is the size of the dataset, out of this, required knowledge is extracted. It may be in KB, MB, GB, TB, or PB, etc., based on the type of the application that generates or receives the data [3]. An example of it would be facebook which generates 25TB of data daily where as twitter generates 400 million tweets daily. When the enterprise data increases to high, the percentage of data to store, process, retrieve, analyze, and visualize also increases at high rates[4].

Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

Velocity

Velocity refers to real-time speed at which the analytics need to be applied. A typical example of this would be to perform analytics on a continuous stream of data originating from a social networking site or aggregation of different sources of data [5].

Variety of data

Variety represents all types of data. In Big data systems, the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, and a raw feed directly from a sensor source. None of these things come ready for integration into an application. Big data processing is done by converting unstructured data into structured and which gives as input to the application [6]. For example, tweets data is converted into structured JSON format because the actual data is unstructured which consists of all types of data such as text, image, and videos etc.

Fig: 1.1: Characteristics of Big Data

2. ARCHITECTURES

A. Apache Hadoop

As per the latest trend, business success heavily depends on the ability to store and analyze large datasets to extract other organizations data, etc. Organizations are discovering important predictions by sorting and analyzing Big data. Since 80% of this data is "unstructured", it must be formatted (or structured) in a way that makes it suitable for data mining and subsequent analysis [7]. Apache Hadoop is the open source software which provides large scale data processing capability. It is the core platform for structuring Big data, and solves the problem of formatting it for subsequent analytics purposes [7]. It uses a distributed computing architecture consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale from single server to thousands of machines and support extremely large data stores.

Hadoop has two main components: Hadoop Distributed File System (HDFS) – is a file system, and MapReduce - a programming paradigm that perform processing of massive datasets.

- **Hadoop Distributed File System (HDFS) Architecture:**

HDFS is highly fault tolerant which was designed using low-cost hardware holds very large amount of data and are stored across multiple machines. Some of the important features of HDFS are storage and processing in distributed environment, streaming access to file system data and provides security through file permissions and authentication. The following figure shows the architectural overview of HDFS.

In the architecture each cluster includes one name node (single point of failure) and many number of data nodes in the racks. Where the name node consists of the metadata i.e. the location where the data is present and data nodes are present in the racks which are the actual data locations. Client or end users can read the data from data nodes by contacting the name node for locations and read operation is performed, but in the writing operation client needs to update the metadata information to the name node after writing the data into data nodes and updates are not possible [8].

As it is in distributed file system, it follows some replication policy while writing the data into clusters. First replica placed in a random node or local node, second replica in different rack and third replica in the same rack as second replica. Here replication policy provides protection against a rack failure during the server crashes or system failures.

Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

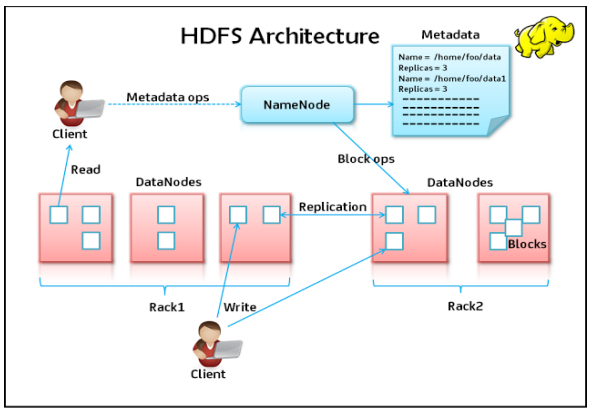

Fig: 1.2: Architecture of HDFS

- **Map Reduce Paradigm**

Robust processing capabilities of Apache Hadoop are based on MapReduce. It is a framework for performing highly parallelized processing of huge datasets, using large clusters of nodes [8].

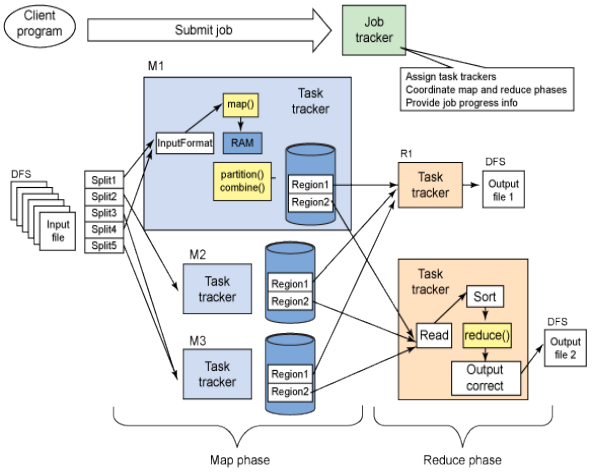

Fig: 1.3: MapReduce Job Processing

MapReduce architecture consists of two phases- Map phase and Reduce phase. Initially the input data set in the job tracker splits into multiple copies as <key, value> pairs where key is the word and value indicates the number of times the word occurred and assigns each sub-job to task trackers. Then it shuffles the values to arrange in an order. Finally in the reduce phase, the results from each task tracker are combined to produce final results.

B. IBM's Platform- Infosphere BigInsights Architecture

IBM Infosphere BigInsights is a software platform which is a distribution of Apache Hadoop with added capabilities that are specific to IBM. It is designed to help firms to discover and analyze business insights in large volumes of different range of data. Examples of such data includes log records, news feed social media, sensors information and transactional data etc [9, 10, 11].

The Fig 1.4 illustrates the IBM's Big data platform, which integrated software libraries for processing, streaming, and persistent data. In the top layer it has different components for application development, management, and visualizations tools. Accelerators layer2 perform main business logic, data processing capabilities, and also running various analytics to manage 3V's of the system. Finally, it integrates Hadoop ecosystem and data warehousing tools in the next layer. In addition to open source software, BigInsights integrates a number of IBM developed technologies to help the organizations to optimize the productivity quickly [9]. Examples include a text analysis engine, supporting analytics tools, data exploration tool and platform enhancements to improve the runtime performance for industry analysts.

Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

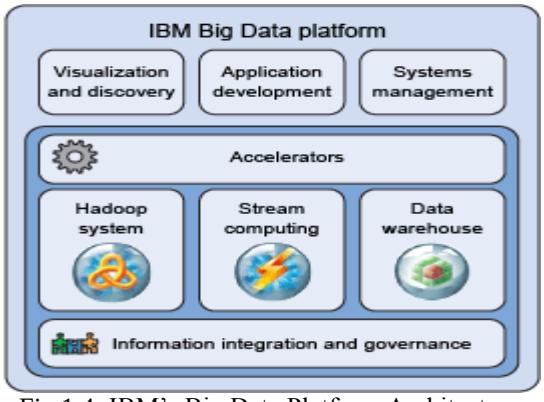

Fig 1.4: IBM's Big Data Platform Architecture

2. PROBLEM STATEMENT

Traditional Enterprise Data Warehouses and database systems do not have the ability to keep up with rapidly increasing social media data. This issue could be solved by building a dashboard to monitor the sentiment of Twitter traffic related to any given topic in near real-time (i.e., with a delay of 1-2 minutes), allowing users to take advantage of near real-time Twitter sentiment for business insights or any other purpose using Apache Hadoop ecosystem. There are several ways to define and analyse the social media data such as Facebook, Twitter etc, where anyone can perform different operations, queries on the collected data of any form like text, audio, video and so on. The traditional methods generally use some coding techniques written using JAVA, .NET, Python, etc for downloading the required data from the twitter. Twitter data can be extracted from Twitter by downloading the libraries that are provided by the twitter, later the raw data can be filtered to find the positive, negative and moderate words. All these words are considered for doing opinion mining or sentiment analysis [12, 13]. After performing the analysis, these words are stored into a database such as RDBMS, but they are having limitations in creating and accessing the data effectively.

This paper presents how to overcome the limitations of traditional techniques using Hadoop ecosystem for simplifying the data processing from large clusters. To stream and analyze the twitter data easily, Apache Hadoop ecosystem and IBM BigInsights tools can be used. Twitter data is crawled from twitter APIs, processed using Jaql script and stored into HDFS, then analyzed as positive, negative words using the MapReduce methods and finally it returns tweet ids as a result then visualizes the results as charts by using bigsheets tool of BigInsights.

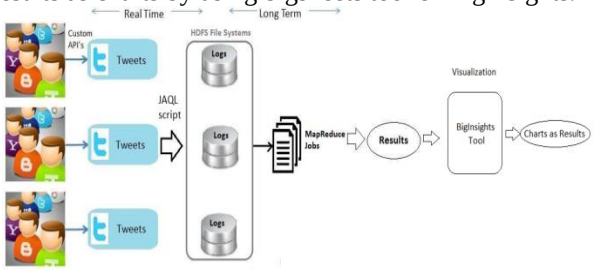

Fig: 2.1 Proposed Architecture for Twitter Data

3. METHODOLOGY

For twitter analysis, the proposed system has the following steps to overcome the problem faced in the existing system.

- Creating Twitter Application.
- Collecting sample twitter data archives.
- Processing data with jaql script
- Running MapReduce for processing the tweets
- Creating Bigsheets Master Workbook and charts.
- **Creating Twitter Application**

First step to stream and analyze the twitter data is to create a twitter application using twitter API by logging into twitter account.

Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

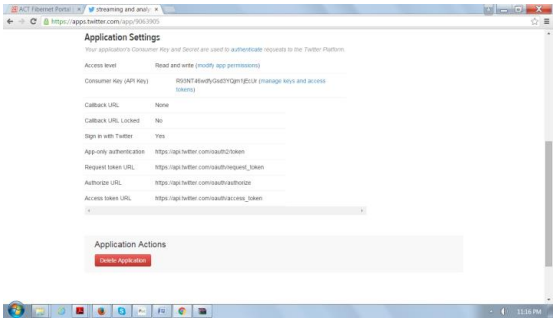

Fig: 3.1 Creation of Twitter Application using Twitter API

➢ **Collecting Sample Twitter Data Archives**

Second step to process the twitter data is to collect sample data archives from the twitter application by requesting authentication tokens from dev application page of twitter that gives the twitter data archives in JSON format.

➢ **Processing Data with jaql Script**

The collected data is in unstructured format. Jaql script is used to extract the important data, transform them into a simpler structure to convert into comma-delimited file and then store the data into HDFS to perform processing using MapReduce.

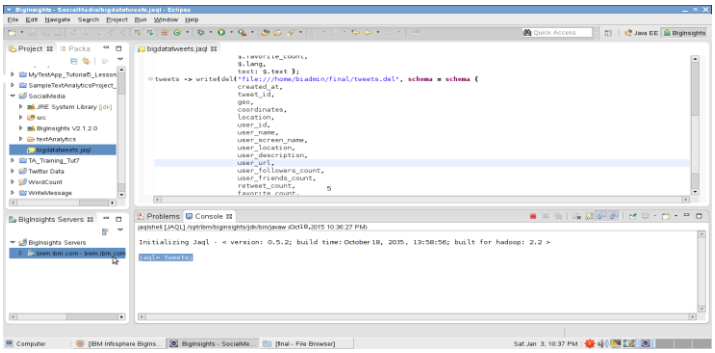
Fig: 3.2 jaql Script for Stream and Storing Twitter Data in HDFS

➢ **Running MapReduce Job for Processing the Tweets**

This step takes the input of produced data archives present in the HDFS and executes MapReduce job to count the occurrences of words in all tweets. It executes mapper, reducer classes and outputs a file which contains tab-separated data in the format WORD <tab> number of occurrences and stores into HDFS.

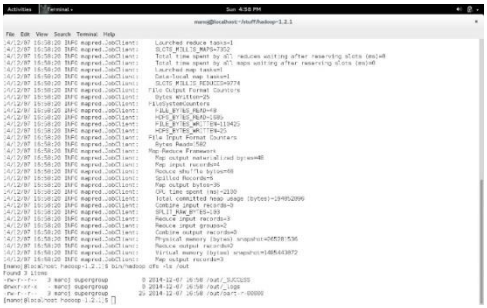
Fig: 3.3 Running Mapreduce Job for Processing Twitter Data

➢ **Creating Master Workbook and Charts**

In final step, import output values to IBM Infosphere BigInsights tool and then it runs the cluster. After uploading the data, tweets data can be in the files tab. Charts can be created as follows:

- In the File menu, choose navigate the comma-delimited file with tweets,
- Click “sheet” radio button, edit the reader from Line to CSV (Comma Separated Value) data, click on confirm changes.

Vol 05, Article 12435; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

- Now table structure for the input data can be seen, then click on save as master book and save it with an appropriate name.
- Open previously saved workbook and build it from the BigSheets web console. To add it to the sheets, click on Add sheets and choose the sheet type, press ok.
- Go to calculate tab and create a new column to show the results then click on apply settings and save the workbook.
- Finally, click on Add chart and choose the type of chart to show the final results.

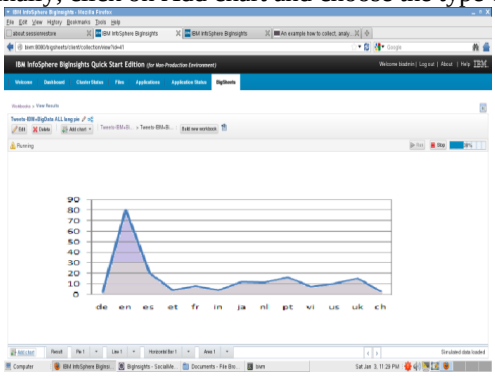

Fig: 3.4 Visualization Chart-1

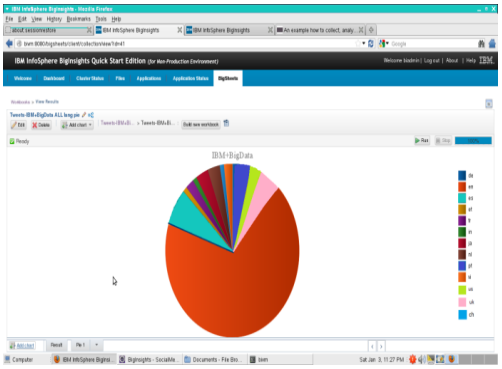

Fig: 3.5 Visualization Chart-2

4. CONCLUSION

Apache Hadoop and its ecosystem addresses the problems which are encountered when traditional methods were applied to deal Big data. But there are other platforms such as cloudera, Horton Works, Microsoft Azure HD Insights, IBM BigInsights that could also be used to overcome the said problems, because they all could integrate with Apache Hadoop ecosystem. This paper presented, a new way of streaming, analyzing and visualizing the twitter data (big data) using BigInsights tool and also the integration of it with Apache Hadoop. The above said tools not only applicable for streaming, processing, analyzing, and visualizing the twitter data but also could be enhanced to apply other types of Big data from various sources like facebook, RFID's, cellular devices and different organizations which requires fast processing, analysis and visualization. In this paper, it is concluded that processing time for analysis of massive twitter data is minimised and retrieving capabilities are also made efficient by using the proposed method when compared to other traditional processing methods for Big data.

REFERENCES

[1] Peter Lake, Paul Crowther, "Concise Guide to Databases: A Practical Introduction", Springer-Verlag London 2013.

[2] Thomas H. Davenport, "Big Data at Work: Dispelling the Myths, Uncovering the Opportunities", Harvard Business Review Press, Boston, Massachusets.

[3] O'Reilly Radar Team, Planning for Big data, A CIO's Handbook to changing the Data Landscape.

[4] Paul C. Zikopoulos, Chris Eaton, Dirk deRoos "Understanding Big Data", ISBN 978-07179053-6.

[5] Vignesh Prajapati, Big Data Analytics with R and Hadoop, Packt Publishing, 2013.

[6] Mr. Swapnil A. Kale, Prof. Sangram S.Dandge, Understanding the Big Data problems and their solutions using Hadoop MapReduce, ISSN 2319 – 4847, Volume 3.

[7] http://www.hadoopuniversity.co.in/

[8] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System" , In the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[9] Steven Hurley, James C. Wang, IBM System x Reference Architecture for Hadoop: IBM BigInsights Reference Architecture.

[10] http://www.ibm.com/developerworks/data/library/techarticle/dm-1110biginsightsintro/

[11] https://www.ibm.com/developerworks/library/bd-socialmediabiginsights/

[12] Penchalaiah.C, Murali.G Suresh Babu.A, Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive, Computer Science and EngineeringDept, JNTUACEP, Pulivendula, Vol. 1 Issue 8, October 2014.

[13] Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde, Real Time Sentiment Analysis of Twitter Data Using Hadoop.