International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights

Manoj Kumar Danthala
Department of Computer Science
Keshav Memorial Institute of Technology
Hyderabad, India

Dr. Siddhartha Ghosh
Department of Computer Science
Keshav Memorial Institute of Technology
Hyderabad, India

***Abstract*— Nowadays the term big data becomes the buzzword in every organization due to ever-growing generation of data every day in life. The amount of data in industries has been increasing and exploding to high rates-so-called big data. The use of big data will become a key basis of competition and growth for individual firms. It is difficult to process and stream the big data within the specified resources. So the simple and easy way to solve the problem of big data is with Hadoop which processes the big data in parallel of data intensive jobs on clusters of commodity servers.**

Here in this paper twitter data, which is the largest social networking area where data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphere BigInsights tool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets.

Keywords—Bigdata; Hadoop; Mapreduce; BigInsights; zettabytes; petabytes

I. INTRODUCTION

Today's organizations face growing challenges from their business values for huge amount data generation and the complexity of data which is both structured and unstructured. Big Data is a term applied to data sets of very large size such that the available tools are unable to undertake their stream, access, analyzing an application in a reasonable amount of time.

If we look at the statistics of industries in real time, there is 2.5 million items added per minute by every individuals. 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 >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's increased to 20 zettabytes and by 2020 this number reaches to 40 zettabytes. In this 80% of world's data is unstructured only which becomes difficult to stream and process for enterprises. So organizations and individual firms need deeper insight to overcome this problem.

Existing database environments, designed years ago, lacks the ability to process big data within the specified amount of time. Also these types of databases have limitations when dealing with different types of data in real time enterprises. So traditional solutions cannot help organizations to manage complex and unstructured data generated in several ways.

Using big data technologies like Hadoop is the best way to solve the big data challenges. These help industries to handle large of complex and unstructured data from various sources. 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. These are some of the tools provides Hadoop ecosystem build by industries to process their own data. All these tools perform various functionalities of analyzing the data based on the different situations.

Characteristics of BigData

The three characteristics of big data are 3V's: Volume, Variety, and Velocity. They're a helpful words through which to view and understand the nature of big data and the software platforms available to exploit them. The following figure shows the structure of 3V's.

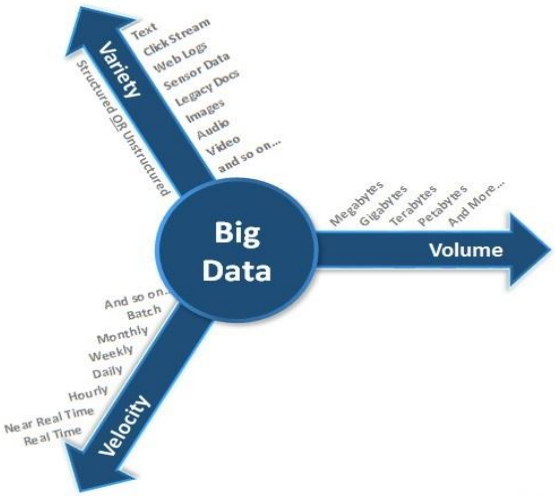


Fig 1.1: Big data characteristics

A. The Volume of data

The volume of data being stored is exploding and increasing every day. In the year 2000, 800,000 petabytes of data were stored in the world. Now it's increased to high rate due to ever-growing usage. We expect this number to reach 40 zettabytes by 2020. Twitter alone generates 12 terabytes of tweets every day. And Facebook also generates 25+ terabytes of unstructured data logs, 10 terabytes of data through google products like gmail, gtalk,youtube etc. and 10 terabytes of data through different tools like RFID, mobile phones and smart meters etc. As the amount of data available to the enterprise is high, the percent of data it can process, retrieve, analyze increases at high rates.

B. *Velocity of data*

The importance of data's velocity- the increasing rate at which data flows into an organization and the time taken to process and analyze the big data in real time. It's not only incoming data, also possible to stream fast-moving data into bulk storage for later batch processing. There are two main reasons for streaming: first is when the input data is too fast to store in to the server clusters and second is to where the application requires an immediate response to the data being requested. So the velocity of data defines both storing data into the servers and also processing, retrieving the same in a specified amount of time.

C. *Variety of data*

A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption as structured input into application. Variety represents all types of data structured data to include raw, semi structured, and unstructured data as part of decision making and insight process. If we look at twitter feed, the tweets data is in structured JSON format but the actual data is not structured, it includes all types- text, image, and videos etc.

II. ARCHITECTURES

A. *Apache Hadoop*

Each of the big data characteristics have its own set of complexities to data processing. When one of these characteristic is present, then it can be processed through traditional data processing tools and methods. But when more than one is present at a time, a new way of techniques required.

Depending on the enterprise and real time scenarios, the raw data may generate as log files, transactional data, uploaded data sets and data related to social media etc. One key difference between the different types of data is structured such as database log records, unstructured such as different types of documents or semi-structured such as log data stored within text files.

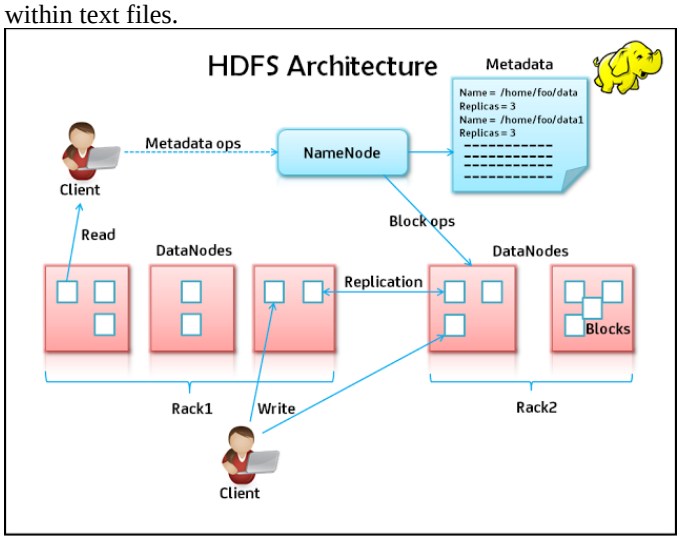


Fig 2.1: Hadoop Distributed File System

Apache Hadoop is the open source software platform which provides large scale data processing capability. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop comprises two main components: a file system, known as the Hadoop Distributed File System (HDFS), and a programming paradigm, known as Hadoop MapReduce.

The above figure shows the architectural overview of the Hadoop eco system in real time. In the architecture each cluster includes one name node and many number of data nodes in the racks. Here 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.

As it is distributed file system it follows some replication policy while writing the data into clusters. Place first replica 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.

B. *MapReduce Paradigm*

The Apache Hadoop's highest processing capabilities are based on MapReduce, a framework for performing highly parallelized processing of huge datasets, using a large clusters of nodes.

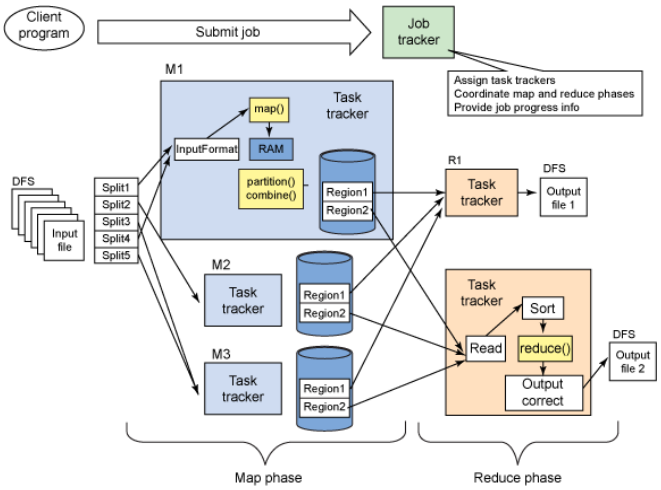


Fig: 2.2: MapReduce Architecture

MapReduce architecture consists of two phases- map phase and reduce phase. Initially the input data set in the job tracker is splits to multiple copies as key value pairs. Here key represents the word and the value is the number of times that 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 it combines the results from each task tracker and produce final outputs.

C. *IBM's Platform- Infosphere BigInsights Architecture*

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 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.

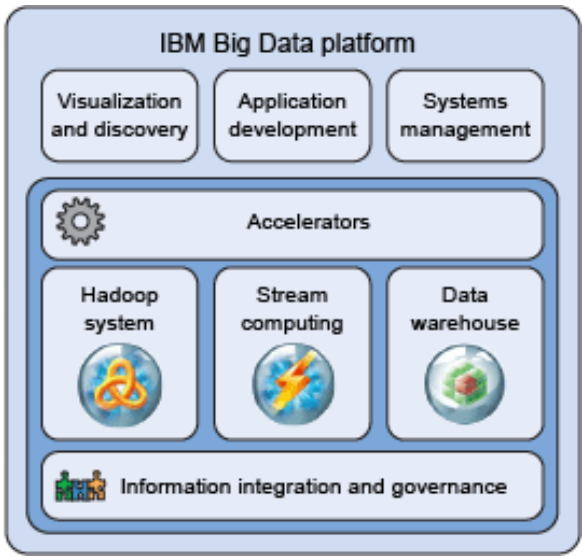


Fig: 2.3 IBM's Big Data Platform Architecture

The above figure illustrates the IBM's big data platform, which integrated software libraries for processing streaming data and persistent data. In the top layer it has different components for application development, management and visualizations tools. Accelerators present in the next layer to perform main business logic, data processing capabilities and also running various analytics to manage 3V's of the system. Then it integrates Hadoop ecosystem, and data warehousing tools in the next layers. Here Hadoop ecosystem includes many open source software's for large computing.

In addition to open source software, BigInsights integrates a number of IBM developed technologies to help the organizations become productive quickly. Examples include a text analysis engine and supporting analyzing tools, a data exploration tool for industry analysts and platform enhancements to improve runtime performance.

III. PROBLEM STATEMENT

A. *Existing System*

As discussed above, small set of social media data can be downloaded and processed easily through traditional databases. This process uses old techniques to stream and analyze the raw data. Also these uses old coding logics to filter and to find out positive, negative and moderate words from the collection of text files and stores in databases. But sometimes the data may be huge in amount and unstructured raw data which traditional databases cannot handle, process and analyzed. This is the major issue comes while streaming and processing big data in real time enterprises.

B. *Proposed System*

In this paper, we are going to overcome the problem of big data by using Hadoop and ecosystem platform, for simplifying the data processing from large clusters. Here we used Apache Hadoop and IBM BigInsights platform to stream and analyze the bigdata problems easily. Here a sample of social media bigdata such as twitter achieves processed using Jaql script and analyzed the positive and negative words using the MapReduce methods and finally shown the results as different charts by using bigsheets tool of BigInsights.

IV. METHODOLOGY

As discussed in the proposed system, to achieve the issue of big data and doing analysis we need to go through following steps.

- Collecting sample twitter data archives.
- Processing data with jaql script
- Running MapReduce technique
- Creating Bigsheets master workbook and charts.

A. *Collecting sample twitter data archives*

The first step to process the twitter data is to collect sample data archives from the twitter application. This can be done by two ways, first tweets can be directly downloaded by requesting your archive from settings in twitter application and second is by requesting authentication tokens from dev application page of twitter. Here I collected through the tokens method from the twitter dev application which gives the twitter data archives in json format.

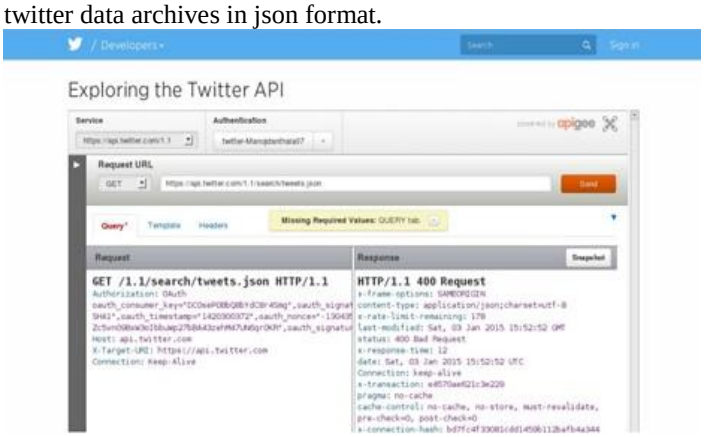


Fig: 3.1 Collecting Twitter data samples

B. *Processing data with jaql script*

Now the collected sample data is not in organized structure, so we need to transform them in a simpler structure to convert into comma-delimited file. Then store the data in hdfs to perform MapReduce operation. Here jaql script loads the tweets, extracts the important data and saves the results into HDFS.

C. *Running MapReduce technique*

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

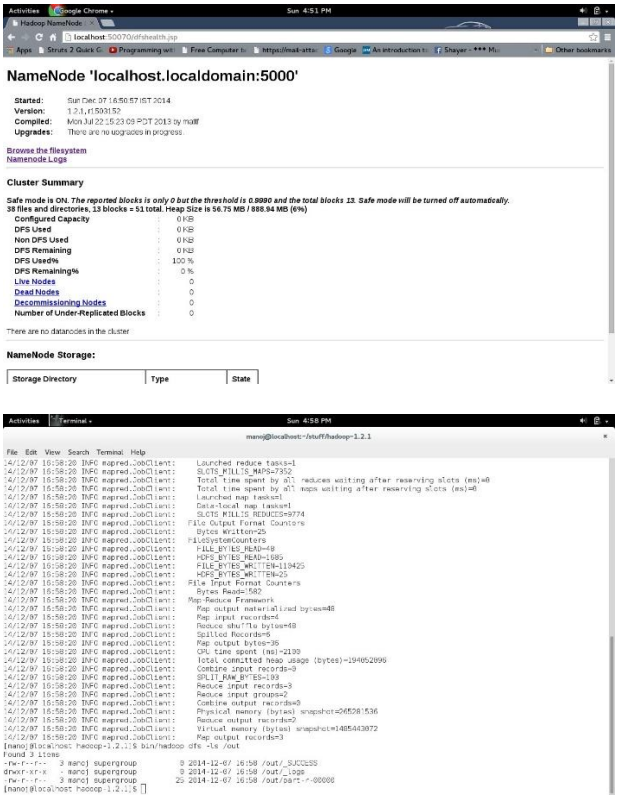
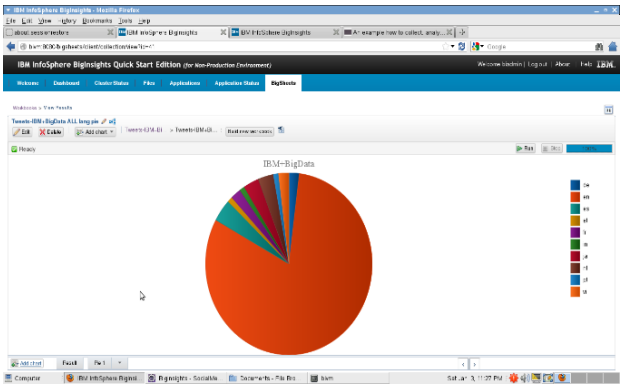


Fig: 3.2 Running MapReduce Jobs

D. Creating Master workbook and charts

Now we need to import output values to BigInsights tool and run the cluster. After uploading the data, we can see the tweets data in the files tab. Now create charts by following steps:

- On the main menu go to Files and navigate the comma-delimited file with tweets
- Click the Radio button "sheet", edit the reader from Line to comma separated value (CSV) data and confirm changes.
- Now we can see the table structure of data, then click on save as master book and set the name and save it.



- From the Bigsheets page of the web console, open previously saved workbook and build it. Then click on Add sheets and as a type of sheet choose the sheet type.
- Now 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.

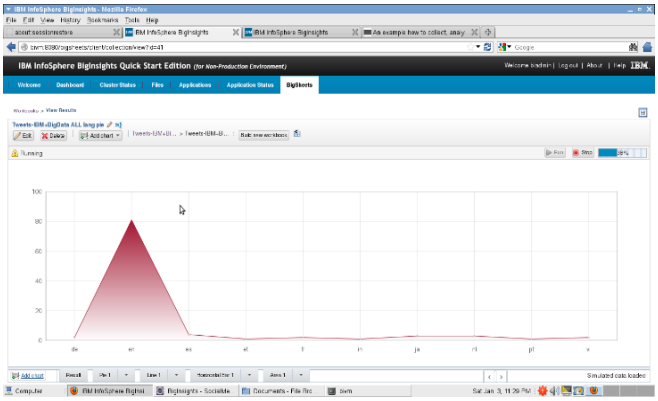


Fig: 3.3 Performance results of tweets

V. FUTURE SCOPE

Previous studies shown that the problems of big data is solved by Hadoop and eco system. But there are other platforms also to solve it which integrates the Hadoop ecosystem. In this paper, a new way of analyzing and visualizing the twitter data shown through BigInsights tool. It can be also enhanced and applied to other types of big data such as facebook, RFID's, cellular devices and different organizations where huge amounts of data is present and needs fast processing and analyzing the data. And there is other platforms to deal with bigdata such as Microsoft's Azure HD Insights, Cloudera, Hortonworks etc. So the same analysis can be done in these platforms also to speed up the performances in real time.

VI. CONCLUSION

Big data is the major problem nowadays in every organization due to ever growing data day-by-day. This issue can be solved by using Hadoop paradigm in real time. In this paper a sample twitter archives is considered as big data for processing and analyzing. Here mapreduce is performed by apache Hadoop and IBM's BigInsights tool is used to stream, transform the tweets data to find the required words. And finally the data is visualized through bigsheets to tool by creating different charts for analyzing tweets.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] 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.

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

[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] Penchalaiah.C, Murali.GSuresh 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.

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

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

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

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