INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

BIG DATA ANALYTICS WITH ADVANCEMENT OF CHALLENGES IN UNSTRUCTURED DATA

N. Jayalakshmi[1] , Nimmatisatheesh[2]

Department of Computer Applications, PSNA College of Engineering & Technology, Dindigul- 624622, India
[1]Jayapsna@gmail.com [2]nimmatisatheesh@gmail.com

Abstract: - Big Data is Latest technology which is very complex to identify the large datasets due to its big size and complexity. It is very difficult to manage in current technology and era. For better exploration of structure and unstructured data requires interactive exploration techniques for the visualization. The visualization mainly used the three important terms: DATA, INFORMATION and KNOWLEDGE. In existing tool there are challenges regarding Human perception means not proper understanding of data and limited screen space not proper visibility of objects properly the large number and complexity of unstructured data opens up many new possibilities for the analyst. Text mining and natural language processing are two techniques with their methods for knowledge discovery from textual context in documents. This is an approach to organize a complex unstructured data and to retrieve necessary information. The paper is to find an efficient way of storing unstructured data and appropriate approach of fetching data. Unstructured data targeted in this work to organize, is the public tweets of Twitter. Building a Big Data application that gets stream of public tweets from twitter which is latter stored in the HBase using Hadoop cluster and followed by data analysis for data retrieved from HBase by REST calls is the pragmatic approach of this project.

Keyword: Unstructured Data, Hadoop, HBase, Data Mining.

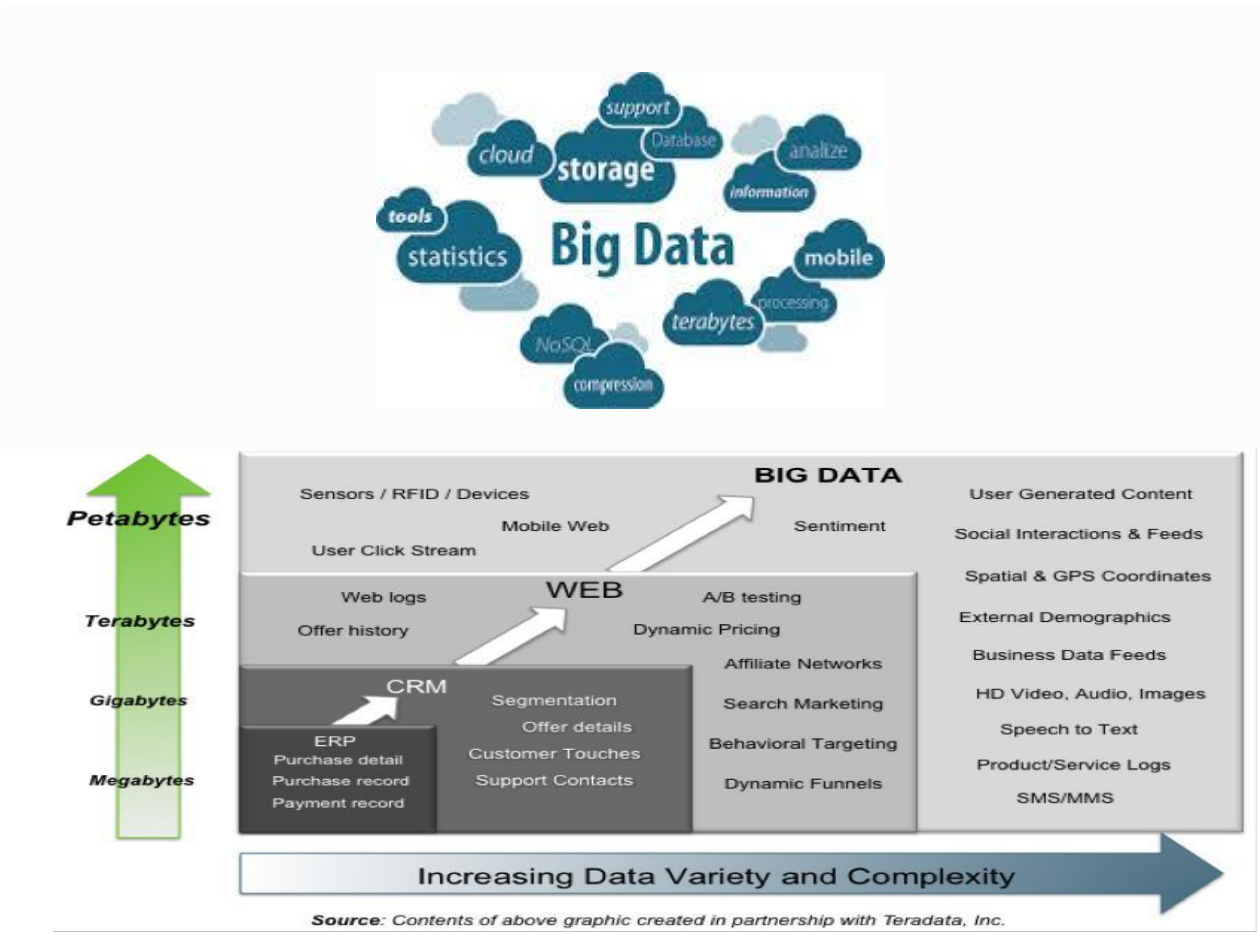
I. INTRODUCTION

Twitter: Twitter is an online social networking service and micro blogging service that enables its users to send and read text-based messages of up to 140 characters, known as "tweets". Public tweets of the Twitter are taken as the Big Data source.

Big Data: Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.

Gartner defines Big Data as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

According to IBM, 80% of data captured today is unstructured, from sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. All of this unstructured data is Big Data.



Unstructured data — everything from social media posts and sensor data to email, images and web logs — is growing at an unprecedented pace. Here are just a few mind-blowing statistics: Twitter sees about 175 million tweets each day and has more than 465 million accounts. 571 new websites are created every minute of every day. And the world creates 2.5 quintillion bytes of data per day from unstructured data sources like sensors, social media posts and digital photos. Clearly, unstructured data is growing exponentially, and government is no exception. What does that mean? "Unstructured" means just that — the elements within the data have no structure. For example, even a simple blog post has many elements embedded in it — the date and time it was posted, the content, embedded links, author, etc. That makes searching and analysis much more difficult than for structured data, like transactions.

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch and different sizes from terabytes to zettabytes. Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety. Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions.

The five Vs characterize what bug data is all about, but also define the major issues IT needs to address

Volume

We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, musics and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents ***Big Data***.

Variety

Data can be stored in multiple formats. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world has data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represent represents ***Big Data***.

Velocity

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent ***Big Data***.

Veracity

When we are dealing with a high volume, velocity and variety of data, it is inevitable that not all of the data is going to be 100% correct – there will be dirty data. The question is, how clean is good enough for the analysis to be performed? Often the data does not need to be perfect, but does need to be close enough to gain relevant insight. Dependent on the application, the veracity, or verification of the data may be essential, or simply "nice to have"

Value

This is the most important aspect of big data. It costs a lot of money to implement IT infrastructure systems to store big data, and businesses are going to require a return on investment. At the end of the day, if you can't extract value from your data, there is no point in building the capability to store and manage it. This is one of the areas where Deloitte Analytics can help, analyzing the data to provide an ROI and build competitive advantage.

.

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage..

II. RELATED

STUDY Emerging Technologies for Big Data Analytics:

Hadoop Virtualization: Virtualization can significantly degrade Hadoop performance and is often avoided, despite bringing obvious advantages in management and utilization. New and evolving tools from Serengeti, Cloudera, Pivotal, MapR and Hortonworks are likely to improve performance and offer the possibility of building high-performance, cost-effective data centers on the Hadoop Distributed File System (HDFS).

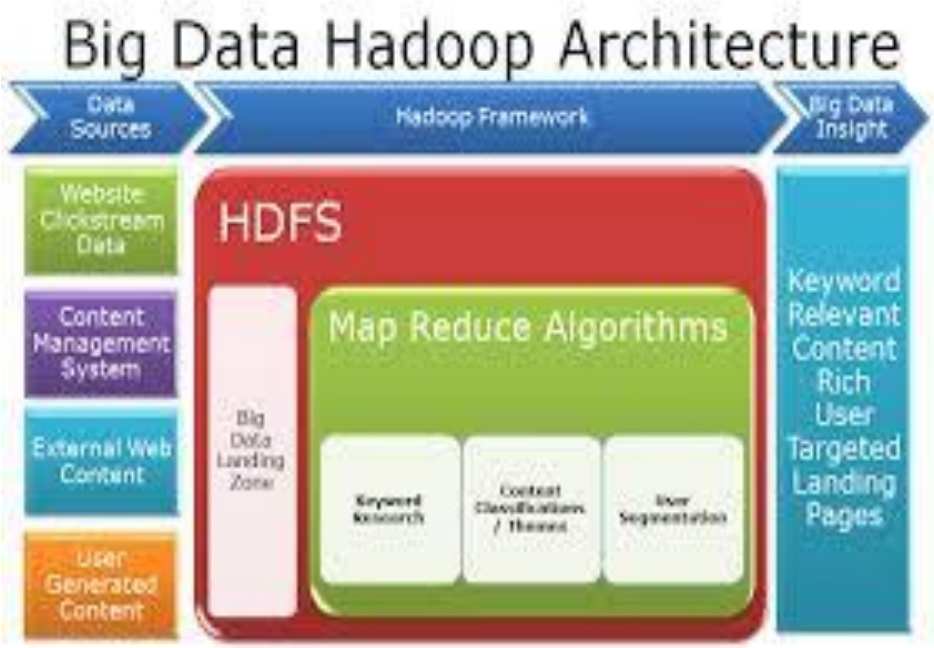
Secure Big Data File Systems: HDFS does not offer the same level of access control and protection as a traditional file system. Options for securing big data are quickly emerging. One such option is a fast, one- and two-way format, preserving encryption of sensitive fields, especially for cloud environments and analyses where disguised identifiers can be used. Nodelevel access control (e. g., from Kerberos), is adequate for some

applications; HDFS encryption is in the works for future Hadoop releases and available from third parties. Finally, row-level access control is available in secure versions of Apache HBase, notably Accumulate

Attached Storage: Attached storage has several advantages over internal JBOD (Just a Bunch of Disks). First, it allows storage to scale independently of compute. Second, the higher reliability allows for a lower replication factor. Third, storage vendors provide SLAs that are unavailable in many JBOD solutions. F

In-Memory: Big data solutions are often a trade-off between ease of access and speed of analysis. In-memory databases offer the best of both worlds: random access and high-speed reads and writes.

Analytics applications architecture.-



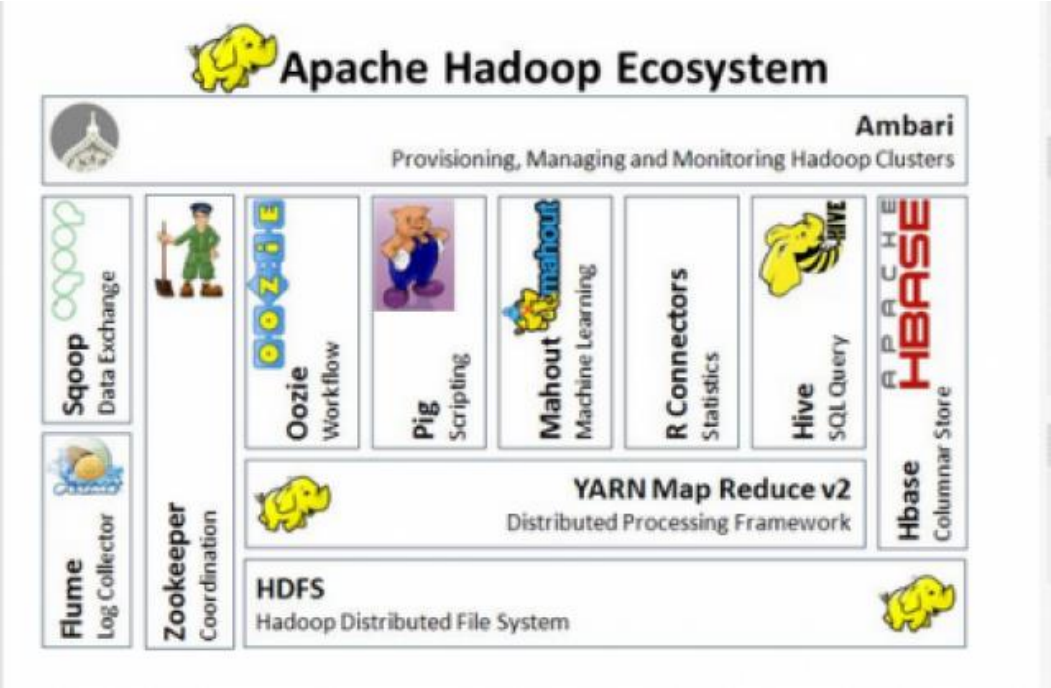
Data architecture. - To handle the variety and complexity of unstructured data, databases are shifting from relational databases to non relational. Unlike the orderly world of relational databases, which are structured normalized, and densely populated, non relational databases are scalable, network oriented, semi structured, and sparsely populated. NOSQL database solutions do not require fixed table schemas, avoid join operations, and scale horizontally.

The Apache Hadoop framework is composed of the following modules

1. Hadoop Common: contains libraries and utilities needed by other Hadoop modules
2. Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the Commodity machines, providing very high aggregate bandwidth across the cluster
3. Hadoop YARN: a resource-management platform responsible for managing compute Resources in clusters and using them for scheduling of users' applications
4. Hadoop MapReduce: a programming model for large scale data processing

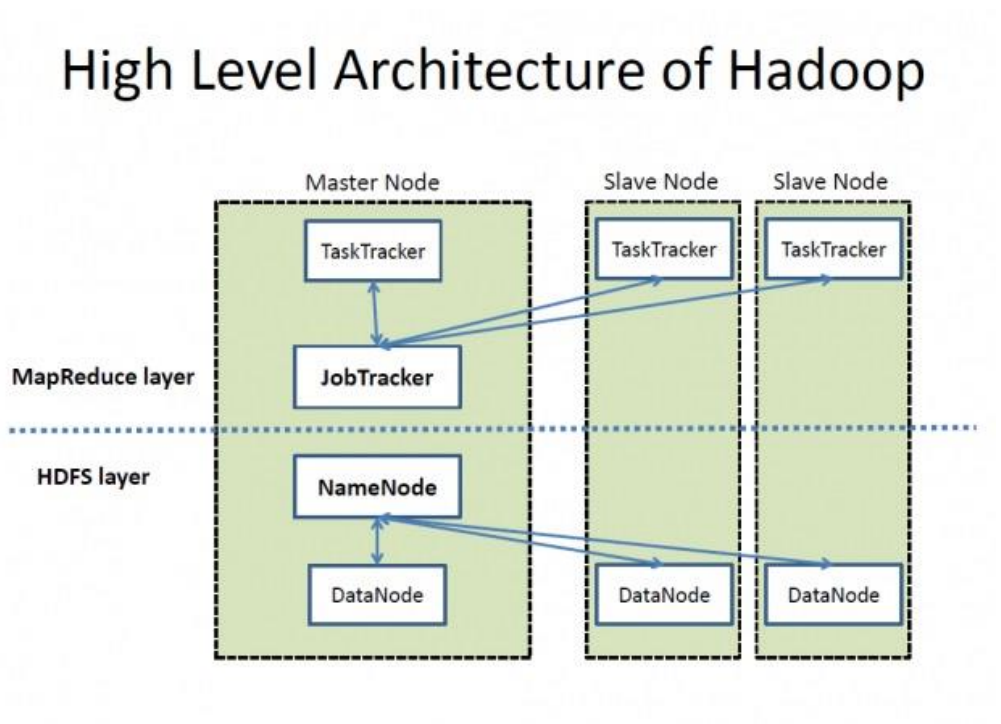
All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers.

Beyond HDFS, YARN and MapReduce, the entire Apache Hadoop "platform" is now commonly considered to consist of a number of related projects as well: Apache Pig, Apache Hive, Apache HBase, and others.



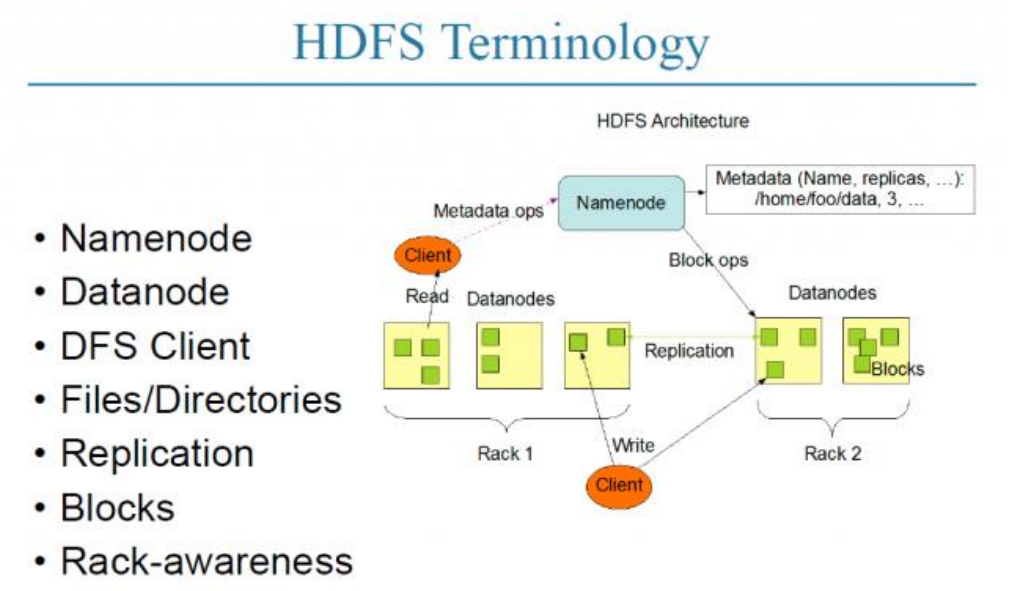
HDFS and MapReduce

There are two primary components at the core of Apache Hadoop 1.x: the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework. These are both open source projects, inspired by technologies created inside Google.



Hadoop distributed file system

The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode, and a cluster of datanodes form the HDFS cluster. The situation is typical because each node does not require a datanode to be present. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other.



HBase is a column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typical MapReduce application. HBase does support writing applications in Avro, REST, and Thrift.

An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key. An HBase column represents an attribute of an object; for example, if the table is storing diagnostic logs from servers in your environment, where each row might be a log record, a typical column in such a table would be the timestamp of when the log record was written, or perhaps the server name where the record originated. In fact, HBase allows for many attributes to be grouped together into what are known as column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families. However, it's very flexible in that new columns can be added to families at any time, making the schema flexible and therefore able to adapt to changing application requirements.

III. OUR APPROACH

Our proposed approach comprises of following three phases.

Establishing connection, followed by Streaming the public tweets from the Twitter using Java. (Data is retrieved from parsing XML reply)
Building a Hbase and then Storing the data in it after sentimental analysis.(All communication is done through REST CALLS)
Building of front end for the client to interact to the Hbase through Java framework for getting the appropriate data required.
Figure 1 depicts an overview of the proposed model, while the following subsection illustrate each phase in detail.
Phase 1: Streaming public tweets from the Twitter:
The process involved is:

Register the application with the Twitter development for getting the Access Tokens and other credentials necessary for the authentication process.
Authenticating the application- Now the application need to be authenticated for the access of the twitter database, which is done by using OAuth (OAuth is an open standard for authorization. OAuth provides a method for clients to access server resources on behalf of a resource owner)
Sending the Requests- Java coding for interacting to the server and pass the requests Http Client.
Parsing the result – the XML result obtained need to be parsed for the filtering the result obtained.
Phase 2: Building Hbase and establishing the connection to java framework
Initially Hadoop need to be installed, which is available for free as it is a open source.
Hbase need to be configured to the system, mostly the project works on the single node cluster.
Storing data – data obtained from the twitter were to be stored in the database by REST Calls.
Organizing the Big tables.
Phase 3: User interacting front end**:**
Building a front end on java script, which in turn connected to the Java framework which is connected to Hbase for fetching and analyzing the data.
Graphical representation for the user.

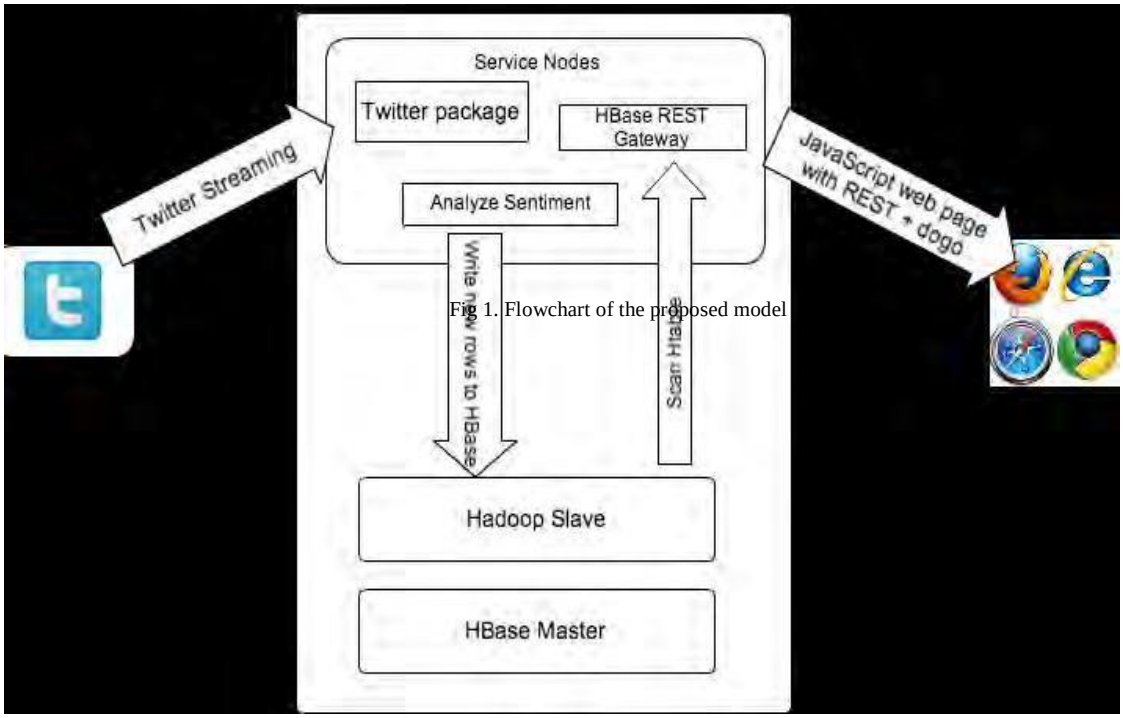


Fig 1. Flowchart of the proposed model

FUTURE SCOPE

The scope of advance tool over Existing ManyEyes tool is that only authenticated user can view the files information which are previously uploaded or used by him at particular time. No other users have access to those files. No issue regarding file saving at web. PDf Files are converted into Tabular format for defining the locations of the text as per the user requirement. The main scope of data is how to analyse the data to make the decisions and to reach at the result. The hidden objects should available properly for resolving the problem of human perception and limited screen space. The main scope is to provide the security and privacy to the datasets which is important for the user.

CONCLUSION

With the need of the decisions it is necessary to convert the data into interactive format for the better decision making. For this best use is visualization of the data in various forms to represent the data and discovering the hidden patterns. Visualization resolves the problem of Human perception and limited screen issues etc.there are various challenges related to storage, analysis and understanding of data. The security is main concerned idea for private information.

REFERENCES

[1] Hansen, C. (2013), "Big Data: A Scientific Visualization Perspective", SCI Institute Professor of Computer Science, University of Utah.
[2] Zhang, Jinson ,Huang, Mao Lin (3-5 Dec. 2013), "5Ws Model for Big Data Analysis and Visualization",
[3] Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference, Pages 1021 – 1028.
[4] Danial Keim(August 2013), "Big-Data Visualization", Computer Graphics and Applications, IEEE Volume: 33, Issue: 4.
[5] Prof. Roberto V. Zicari (2013), "The challenges and opportunities of big data".
[6] Tien, J.M. (17-19 July, 2013), "Big Data: Unleashing information".
[7] http://vis.ucdavis.edu/Workshops/BigDataVis2013/#about .