BIG Data Analytics: A Framework for Unstructured Data Analysis

T.K.Das[1] , P.Mohan Kumar[2]
School of Information Technology and Engineering, VIT University,
Vellore- 632014, India
[1]tapan.das@vit.ac.in
[2]pmohankumar@vit.ac.in

Abstract - Nowadays, most of information saved in companies are unstructured models. Retrieval and extraction of the information is essential works and importance in semantic web areas. Many of these requirements will be depend on the unstructured data analysis. More than 80% of all potentially useful business information is unstructured data, in kind of sensor readings, console logs and so on. 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 an 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: Data is exploding at an astounding rate. While it took from the dawn of civilization to 2003 to create 5 Exa bytes of information, we now create that same volume in just two days! By 2012, the digital universe of data will grow to 2.72 zetta bytes (ZB) and will double every two years to reach 8ZZB by 2015. For perspective: That's the equivalent of 18 million Libraries of Congress. Billions of connected devices-ranging from PCs and smart phones to smart phones to sensor devices such as RFID readers and traffic cams-generate this flood of complex structured and unstructured data.

Unstructured data is heterogeneous and variable in nature and comes in many formats, including text, document, image, video and more. Unstructured data is growing faster than structured data. According to a 2011 IDC study, it will account for 90 percent of all data created in the next decade. As a new, relatively untapped source of insight, unstructured data analytics can reveal important interrelationships that were previously difficult or impossible to determine.

Big data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners and the business and ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities.

Impact of big Data on IT:

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

Volumes. The massive scale and growth of unstructured data outpace traditional storage and analytical solutions.

Variety. Big Data is collected from new sources that haven't been mined of insight in the past. Traditional data management processes can't cope with the heterogeneity and variable nature of big data, which comes in formats as different as e-mail, social media, videos , images, blogs and sensor data as well as "shadow data" such as access journals and Web search histories.

Velocity. Data is generated in real time with demands for usable information to be served up as needed.

Hadoop: Apache Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware. The entire Apache Hadoop "platform" is now commonly considered to consist of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects – including Apache Hive, Apache HBase, and others.

HBase: HBase is an open source; non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File system), providing BigTable-like capabilities for Hadoop. It provides a fault-tolerant way of storing large quantities of sparse data.

II.RELATED STUDY

Emerging Technologies for Big Data Analytics:

New technologies are emerging to make unstructured data analytics possible and cost-efficient. The new approach redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources. It utilizes easily scalable “shared nothing” architecture, distributed processing frameworks, and non relational and parallel relational databases.

• Analytics applications architecture.- New data processing systems make the computing grid work by managing and pushing the data out to individual nodes, sending instructions to the networked servers to work in parallel, collecting individual results and then reassembling then to produce meaningful results. Processing the data where it resides is faster and more efficient then first transporting it to a centralized system.

• 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.

Distributed Frameworks: The Emergence of Apache Hadoop

Apache Hadoop is evolving as the best new approach to unstructured data analytics. Hadoop is an open-source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms and an application layer that manages distributed processing, parallel computation, workflow and configuration management. In addition to offering high availability, Hadoop is more cost-efficient for handling large unstructured data sets than conventional approaches, and it offers massive scalability and speed.

The entire Apache Hadoop platform is now commonly considered to consist of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects including Apache Hive, Apache HBase, and others.

HBase:

HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File system), providing BigTable-like capabilities for Hadoop. That is, it provides a fault-tolerant way of storing large quantities of sparse data. HBase is not a direct replacement for a classic SQL database, although recently its performance has improved, and it is now serving several data-driven websites, including Facebook's Messaging Platform. Use Apache HBase when you need random, real time read/write access to your Big Data. This project's goal is the hosting of very large tables -- billions of rows X millions of columns -- atop clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, column-oriented store modeled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

Features of HBase**:**

- Linear and modular scalability.
- Strictly consistent reads and writes.
- Automatic and configurable shading of tables
- Automatic failover support between Region Servers.
- Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.
- Easy to use Java API for client access.
- Block cache and Bloom Filters for real-time queries.
- Query predicate push down via server side Filters
- Thrift gateway and a REST-ful Web service that supports XML, Protobuf and binary data encoding options
- Extensible jruby-based (JIRB) shell
- Support for exporting metrics via the Hadoop metrics subsystem to files or Ganglia; or via JMX

III.OUR APPROACH

Our proposed approach comprises of following three phases.

1. Establishing connection, followed by Streaming the public tweets from the Twitter using Java.(data is retrieved from parsing XML reply)
2. Building a Hbase and then Storing the data in it after sentimental analysis.(All communication is done through REST CALLS)
3. 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:

i) Register the application with the Twitter development for getting the Access Tokens and other credentials necessary for the authentication process.
ii) 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)
iii) Sending the Requests- Java coding for interacting to the server and pass the requests Http Client.
iv) 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

i) Initially Hadoop need to be installed, which is available for free as it is a open source.
ii) Hbase need to be configured to the system, mostly the project works on the single node cluster.
iii) Storing data – data obtained from the twitter were to be stored in the database by REST Calls.
iv) Organizing the Big tables.

Phase 3: User interacting front end**:**

i) 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.
ii) Graphical representation for the user.

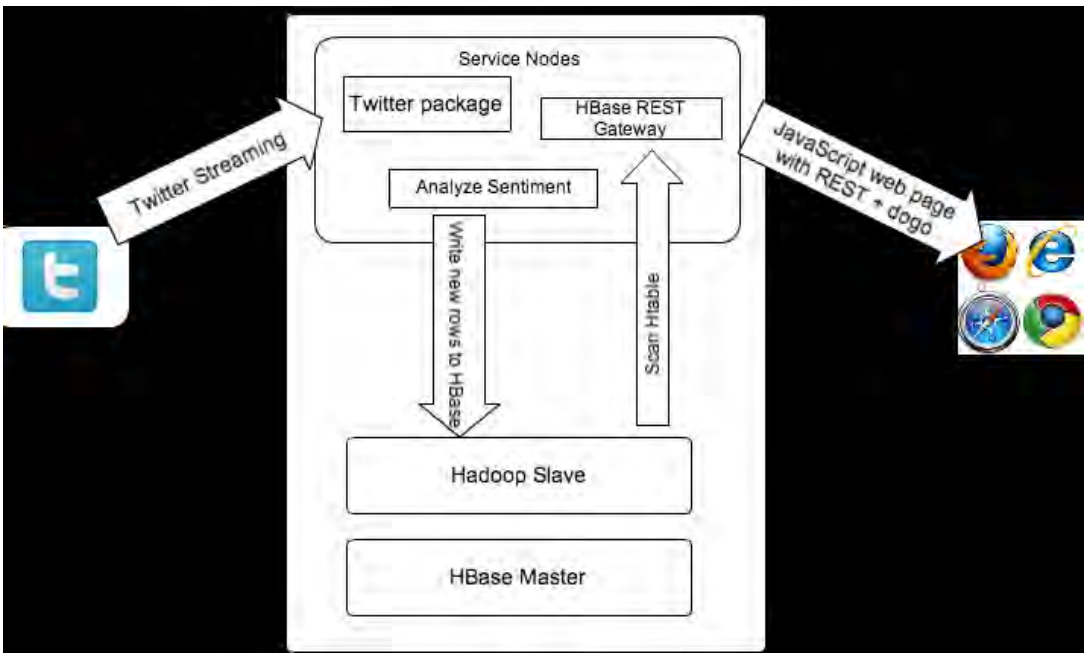

Fig 1. Flowchart of the proposed model

IV.CONCLUSION AND FUTURE WORK

In this paper we presented a framework for analyzing unstructured data. . This is a ongoing project. We have completed the first phase where unstructured data is pulled from public tweets of Twitter and the XML data is parsed to store in a NOSQL database like HBASE, In the future we would build the HBase and by using Text mining algorithms, we would try to get an insight from those data. Result of mining the data would be published in next paper**.**

REFERENCES

[1] Chaiken R. et. al.: SCOPE: easy and efficient parallel processing of massive data sets. PVLDB 1(2), 2008.
[2] Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Communications of the ACM 53(1): 72-77 (2010).
[3] The Apache Hadoop Project.http://hadoop.apache.org/core/, 2009.
[4] S. Das, Y. Sismanis, K. Beyer, R. Gemulla, P. Haas, and J. McPherson. Ricardo: Integrating R and Hadoop. In *SIGMOD*, 2010.
[5] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton.Mad skills: New analysis practices for big data. *PVLDB*,2(2):1481–1492, 2009.
[6] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach,M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. In *OSDI*, pages 205–218, 2006