International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-6, August 2013

NOHT: Situational Awareness by Hadoop Framework VAST 2012 Mini Challenge 1

G.Phani Sindhuri, P.Kiran Kumar, T.Bhavani

Abstract—Big Data is the collection of large and complex data sets which becomes difficult to manage and process using traditional tools. Big data Analytics is the process of examining large amounts of data to make better business decisions. One of the Major challenge pose by VAST 2012 is to symbolize the "Bank of Money" network issues identified by the sprouting technologies to provide situation awareness by observing the visualization of the network. This Paper introduces usage and importance of NOSQL database and distribution of data and its processing in parallel using Apache Hadoop Framework and for fast ad-hoc visualization Tableau software is used to address this challenge.

Index terms-Visual Analytics Science & Technology (VAST)[1] , Not only SQL (NOSQL)[5][6], cloudera distribution for Hadoop (CDH)[9].

I. INTRODUCTION

In the current digital world the data such as structured or semi-structured or unstructured are exploding day-by-day to peta bytes or to exa bytes. so the complexity of maintaining it also increases in its volume, velocity and variety. The traditional enterprise have some limitations to capture, to store ,to transfer ,to share and to process in a tolerable amount of time and even to visualize that much data. so the concept of Big data came to existence. Big data is the platform for transforming all data into actionable items for the business intelligence. Big data analytics is the process of examining useful patterns for competitive advantage for business benefits.

Big Data Analytics and Apache Hadoop open source framework are the emerging technologies to address business trends that are agitating data management and its processing traditionally[8].

The VAST 2012 challenge1 was to identify the network situational awareness of an international bank called "Bank of Money". To achieve horizontal scalability and to aid performance while handling large volume of data with ease , NOSQL stands for "Not only SQL" database such as MongoDB, an open source document oriented database which is designed and supported by 10gen is used. 10gen provides repositories for .deb and .rpm packages for consistent setup, upgrade, system integration, and configuration[5][6].

Manuscript published on 30 August 2013.
***** Correspondence Author (s)
G.Phani Sindhuri*, Computer Science &Engineering, JNTUK / Sasi institute of technology &Engineering ,Tadepalligudem, India.
P.Kiran Kumar, Computer Science &Engineering, JNTUK / Sasi institute of technology &Engineering ,Tadepalligudem, India.
T.Bhavani,Computer Science &Engineering, JNTUK / Sasi institute of technology &Engineering ,Tadepalligudem, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

To achieve better management of enormous data sets across distributed clusters of servers and for an affordable solution for the processing of amorphous information, an open source software platform, Hadoop Framework which is managed by Apache Software Foundation is utilized.

It is a java based framework that enables processing of huge amount of data sets in a distributed environment. Even when an individual server or cluster fails, applications will continue to run using this Framework[3][8].

To bring huge data from Hadoop to life, a visualization tool, Tableau software is availed. It is a business-intelligence software to rapidly transform data into smart business analytics. It creates more interactive, sharable visualizations with ease even for complex analytical problems[7].

II. LITERATURE SURVEY

NOHT refers to the following:

- NO-NOSQL database such as MongoDB , it is a open document oriented database which is written in c++.It have full index support , auto sharding features and also atomic modifiers. There are different ways to implement NOSQL databases such as Document databases, Graph stores, and wide-column stores. That means NOSQL databases can be used to store and work with structured as well as unstructured data[6].
- H-Hadoop Framework, a solution to the big data problem. The sizes of the database is increasing in the current enterprises exponentially. There is a need to process that large volume of data on daily basis. Hadoop framework allows distributed processing of data sets across clusters of computers using a simple processing model.CDH is the open source distribution of Cloudera for Apache Hadoop and its related projects[3][8].
- T-Tableau software for visualization, It is a technology from Stanford university that helps us for simple drag &drop to analyze data. It allows to connect the data in few clicks and then visualize easily. The main advantage of Tableau is its speed. Tableau's data engine is blazing fast for massive data. We can publish interactive dashboards to the web or a server in seconds. It conveys the best way to represent data with different color schemes to focus on important data[7].

III. PROPOSED SYSTEM

To visualize the situational awareness for the "Bank of money " network ,the data is initially stored in NOSQL database such as MongoDB . NoSQl data base is used with out using normal SQL data base because very large amount of bank data is present in the bank world.

© Copyright: All rights reserved.

So the database must have the capability of enlarging itself horizontally and the data should be secured . NOSQl database is capable for those constraints. A MongoDB deployment typically involves multiple servers distributed across multiple data centers. As the data involves about bank transactions, even any system or node fails during any transaction, the data could not be lost. The Hadoop framework can preserve the data even for any disaster .The data is distributed among the clusters and maintained by the Hadoop Framework. So the data is safe at Hadoop framework .If we want the available datasets are useful for predicting any future situation or useful for other applications we need to get the data to a visualized form. Then any one can understand the diagrammatical description and that will be used for analytics .So we are using a simple business intelligence visualization tool called as Tableau Software. The Situation of Bank of money network is visualized by the tableau software . The following data flow diagram shows the process of working with NOHT tool.

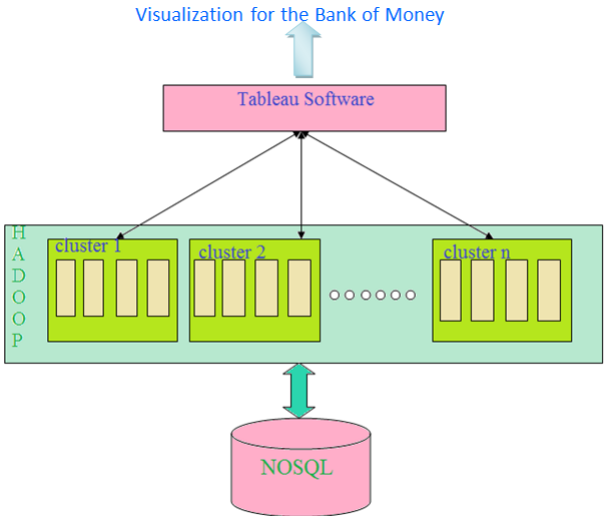

Fig 1 : The figure shows the step wise procedure to work with NOHT tool for the situational awareness for the bank of money network.

IV. DEVELOPMENT PROCESS FOR NOHT

The VAST 2012 challenge was to investigate the cyber situational awareness of a fictitious international bank called the 'Bank of Money'. Two datasets were provided for the first challenge, the first contained meta-data that describes the computer network of the bank. This consisted of nearly 1 million individual machines divided over 4,05geographically dispersed facilities which included datacenters, regional offices and branches. The second dataset contained the output of status updates for all computers on the banks network for two days. The status of each machine is stored every 15 minutes and provides details about deviations from corporate policies and the current activity for each individual machine[1][2].

The development of NOHT is divided into 3 phases. Database development, development of Hadoop Framework, and the development of visualization framework

A. Database Development:

As the challenge involves large amounts of data sets with different data types, instead of storing data in traditional relational database systems, MongoDB stores data and it performs integration of different types of data easier and faster. If database needs to enlarge its features horizontally(add new attributes) ,NOSQL databases allow insertion of data without predefined schema and it does not slow down the process. Instead of giving burden on single server, NOSQL Databases support automatic spread across the number of servers.

MongoDB database have strong consistency. It also have the property called as "Automatic Replication" to provide high availability of data even when a disaster occurs at any resource at different branches. The performance of NOSQL database is more compared to Relational Databases. MongoDB provides more flexible index support and also Advanced Security using advanced firewall configurations[5][6].

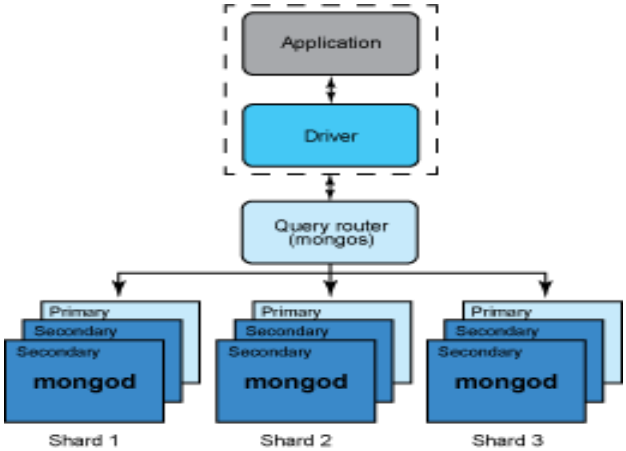

Fig 2:The diagram shows the overview of MongoDB.

Shading is one of the main properties of MongoDB. It is the process of storing data across multiple machines .As the size of data increases single machine may not be sufficient to store data and also to perform read and write operations efficiently. With sharding, we can support data growth and the demands of Read and Write operations. To route a query to a cluster the mongos first determines the list of shards that must receive the query. Then it establishes the cursor on all targeted shards[6].

B. Development of Hadoop Framework:

Hadoop is a generic & flexible infrastructure for distributed computation which is completely written in java. It is a Reliable ,scalable and more powerful framework. Hadoop is not a single product rather an ecosystem of software products that together implements fully featured and flexible big data analytics[8].

Hadoop focuses on Batch processing. The goal of Hadoop and NOSQL is to give massive scalability and to support Big Data. It provides high throughput access to application data and it is suitable for the applications that have large datasets. Map Reduce framework produced by Hadoop will process huge amount of data across multiple machines in a cluster in parallel. It then merges all the sub-problem solutions together and writes out the solution into files[8].

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

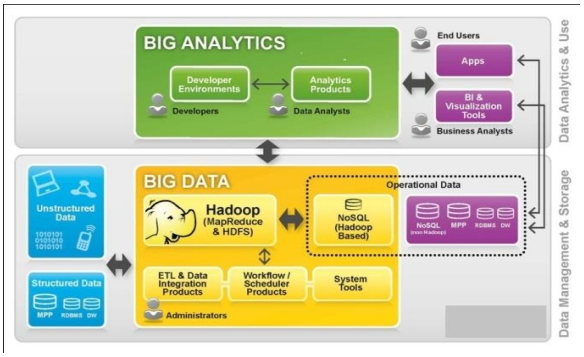

Fig3: The figure shows the Architecture for Big Data Analytics. structured or unstructured data(Big Data) is stored and managed by Hadoop and NOSQL databases and then performing analytics on the data.

MongoDB is the database for big data systems across variety of measures. MongoDB connector for Hadoop makes easy to Hadoop users to integrate the data in MongoDB. Real time data from MongoDB database can be read and processed by Hadoop and Map Reduce jobs and the results can be written back to MongoDB to support real time data processing and for querying.

HDFS(Hadoop distributed File System)is the default memory area for all the machines in Hadoop cluster. Hadoop is a scalable and fault-tolerant(user specific replication)-Even if a node fails, the system redirects work to another location of the data and continues processing without missing any fragment/block.

C. Development of visualization Frame work:

The process of converting data into graphical/visual form is called data visualization. The data from Hadoop is brought and visualized by Tableau software, which is a simple drag & drop technology to analyze data and it is a strong analytics tool in a data-driven world. The data can be analyzed and visualized quickly and easily and it is also easy to share information. To make the data in hadoop more meaningful, the tableau connector with Hadoop through the partnership with cloudera distribution (CDH) is used[7][9].

Cloudera connector for Tableau is a free ODBC driver that enables connection to Apache Hive, is a technology which enables easy data summarization, and analyze large data sets stored in Hadoop using SQL-like query language called HiveQL. Hive translates HiveQL statements into set of Map Reduce jobs then executes on Hadoop cluster[8][9].

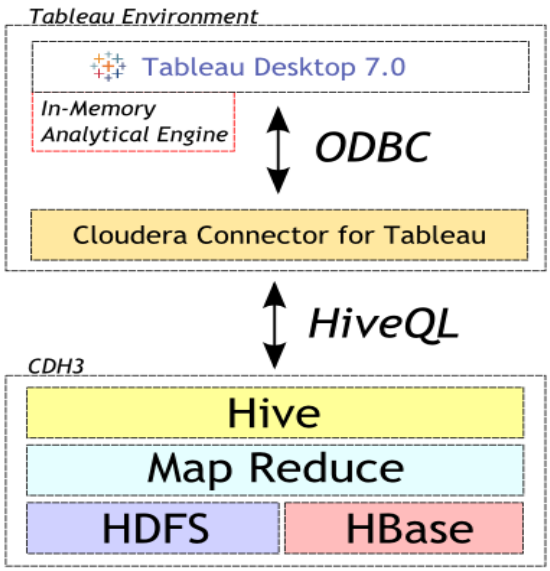

Fig4 :The figure shows the connection to Tableau desktop with Hive using cloudera connector .

V. VISUALIZATIONS

Bank of Money is an international organization which is having number of offices/branches in the bank world. Number of people involves in the bank transactions like balance enquiry, withdrawal, deposit etc from different regions. The data is stored in MongoDB database and the data is distributed and managed by Hadoop framework to help from disasters and to execute parallelism. The MongoDB connector for Hadoop connects the data from the database to Hadoop framework. The data from all the Hadoop clusters will bring to tableau software, a visualization tool. It is very easily workable tool with a simple drag & drop Technique. With in very few seconds we can visualize large amount of data in understandable way with interactive nature.

NOHT tool gives the following situational visualization at different regions.

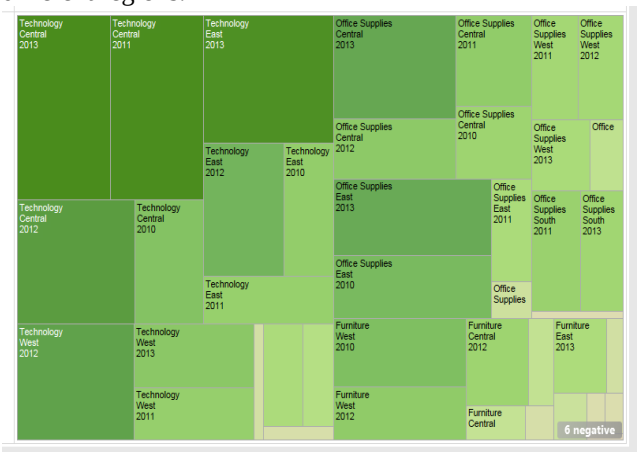

Fig5:Figure shows the snapshot of NOHT describing the situation of Bank of Money enterprise at different regions.

The NOHT tool shows the density of people who are working from different regions in the bank world.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

VI. CONCLUSION

We provided a solution to the VAST 2012 mini challenge1. NOHT- provides the situational awareness for the mini challenge1 given by VAST 2012. It was developed for improving the efficiency and reliability using Hadoop, an open source Apache framework and MongoDB ,a leading NOSQL database is used for the faster access of the data and for easy scaling and to improve the performance and Tableau software, an open source visualization tool is used for fast-analytics and more interactive to the user.

ACKNOWLEDGMENT

We want to thank our co-faculty members for their suggestions and encouragement.

REFERENCES

[1] Kristin Cook, Georges Grinstein, Mark Whiting, Michael Cooper, Paul Havig, Kristen Liggett, Bohdan Nebesh, "VAST Challenge 2012: Visual Analytics for Big Data".
[2] Williams, F.C.B., Faithful, W.J., Roberts, J.C., "SitaVis – Interactive Situation Awareness Visualization of Large Datasets."
[3] Patel A.B,Birla M,Nair U,Addressing big data problem using Hadoop and Map Reduce
[4] Abousalh-Neto, N. A., Kazgan, S., "Big Data Exploration through Visual Analytics."
[5] MongoDB vs. Oracle Database Comparision, Romanian , Boicea.A;Fac. Of Autom. Control & Comput.sci .,Politeh.Univ. of Bucharest, Bucharest, Romania; Radulescu.F;Agapin.
[6] MongoDB:http://www.mongodb.org
[7] Tableau Software: http://www.tableausoftware.com
[8] Apache Hadoop: http://www.hadoop.apache.org
[9] cloud era Hadoop: http://www.cloudera.com/content/cloudera/en/products/cdh.html

First Author : G Phani Sindhuri

Photo :

Education :B. Tech(CSE) ,pursuing M. Tech(CSE)
Occupation :Asst.Prof at Sasi Institute of Technology & Engineering Tadepalligudem, Andhra Pradesh, India

Second Author : P Kiran Kumar

Photo :

Education :M. Tech(CSE)
Occupation :Asst.Prof at Sasi Institute of Technology & Engineering, Tadepalligudem, Andhra Pradesh, India

Publications :A new approach for distribution, execution and management of application soft wares and Enterprise VMs based on Virtualization
Membership :A Member of CSI
Started the Research work in Big Data Analytics for Ph.D.

Third Author :T.Bhavani

Photo :

Education :M. Tech(CSE)
Occupation :Asst.Prof at Sasi Institute of Technology & Engineering Tadepalligudem, Andhra Pradesh, India
Publications :Share point 2007 for business intelligence
Started the Research work in Data Mining for Ph.D.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.