A Review: Data Mining for Big Data

Miss. Punde Archana, Miss. Daundkar Kavita, Miss. Shelar Sarswatee

Abstract—***Instead of relying on expensive, proprietary hardware and different systems to store and process data, enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With no data is too big. And in today's hyper-connected world where more and more data is being created every day, breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.***

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This project presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

Keywords— **Big data, Data mining, Hace theorem, 3V's, Privacy.**

I. INTRODUCTION

This Better Practice Guide aims to improve government agencies' competence in big data analytics by informing government agencies about the adoption of big data including:

Identifying the business requirement for big data capability including advice to assist agencies identify where big data analytics might support improved service delivery and the development of better policy.

Developing the capability including infrastructure requirements and the role of cloud computing, skills, business processes and governance. Considerations of information management in the big data context including assisting agencies in identifying high value datasets, advising on the government use of third party datasets, and the use of government data by third parties. Promoting privacy by design.

Promoting Privacy Impact Assessments (PIA) and articulating peer review and quality assurance processes and big data project management including necessary governance arrangements for big data analytics initiatives. Government agencies have extensive experience in the application of information management principles that currently guide data management and data analytics practices, much of that experience will continue to apply in a big data context.

This better practice guide is intended initially as an introductory and educative resource for agencies looking to introduce a big data capability and the specific challenges and opportunities that accompany such an implementation. Often there will be elements of experience with implementing and using big data to a greater or lesser degree across government agencies. In this guide we aim to highlight some of the changes that are required to bring big data into the mainstream of agencies operations. More practical guidance on the management of big data initiatives will be developed subsequent to this better practice guide as part of a guide to responsible data analytics.

As outlined greater volumes and a wider variety of data enabled by new technologies presents some significant departures from conventional data management practice. To understand these further we outline the meaning of big data and big data analytics contained and explore how this is different from current practice.

II. OBJECTIVE FOR BIG DATA

1. Habituate with the concepts of "Cloud Computing", "Cloud Security", "Cloud Security Issues", "Cryptographic techniques".
2. Provides security to the data in Cloud using Cryptographic schemes.
3. Prevent unauthorized access to the data stored in cloud.
4. Provides Independent and Concurrent access to the data in Cloud.

5. Provide Role-Based Access, i.e. provides access only if the user has access permission.
6. Build own private cloud using Eucalyptus.

III PROPOSED SYSTEM

The propose system is built on Ubuntu 12.04LTS operating system and require Eucalyptus framework to create actual private cloud environment.

HACE THEROM:-

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data .These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naïve sense, we can imagine that a number of blind men are trying to size up a giant Camel, which will be the Big Data in this context. The goal of each blind man is to draw a picture (or conclusion) of the Camel according to the part of information he collects during the process. Because each person view is limited to his local region, it is not surprising that the blind men will each conclude independently that the camel "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that the camel is growing rapidly and its pose changes constantly, and each blind man may have his own (possible unreliable and inaccurate) information sources that tell him about biased knowledge about the camel (e.g., one blind man may exchange his feeling about the camel with another blind man, where the exchanged knowledge is inherently biased). Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the camel in a real-time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the camel and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process .The term Big Data literally concerns about data volumes,

HACE theorem suggests that the key characteristics of the Big Data are

1. Huge with heterogeneous and diverse data sources:

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This huge volume of data comes from various sites like Twitter, MySpace, Orkut and LinkedIn etc.

2. Decentralized control:-

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers

3. Complex data and knowledge associations:-

Multistructure, multisource data is complex data, Examples of complex data types are bills of materials, word processing documents, maps, time-series, images and video. Such combined characteristics suggest that Big Data require a "big mind" to consolidate data for maximum values.

IV. THREE V'S IN BIG DATA

Volume: there is more data than ever before, its size Continues increasing, but not the percent of data that Our tools can process.

Variety: there are many different types of data, as text, Sensor data, audio, video, graph, and more **Velocity:** data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time Nowadays, there are two more V's:

Variability: there are changes in the structure of the data and how users want to interpret that data **Value**: business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach Gartne summarizes this in their definition of Big Detain 2012 as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

There are many applications of Big Data, for example the following:

Business: costumer personalization, churn detection

Technology: reducing process time from hours to seconds

Health: mining DNA of each person, to discover, monitor and improve health aspects of every one

Smart cities: cities focused on sustainable economic development and high quality of life, with wise management of natural resources.

These applications will allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness

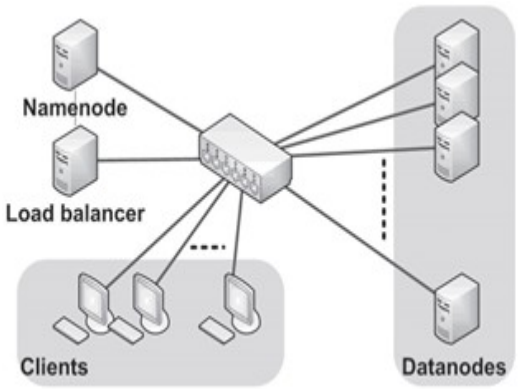

Fig 1: System Architecture

PERFORMANCE RESULTS

1. Cloud Controller (CLC):-

It is the entry point into the private cloud for end user, project manager's developers and administrator. It also helps in manage virtualized resources. Walrus: It implements bucket-based storage, which is available inside and outside the cloud system.

2. Cluster Controller (CC):-

It executes on a machine that has network connectivity to the machines that are running on Node Controller and Cloud Controller .It manages the Virtual Machine (VMs) Network. All Node Controllers associated with a single CC must be in the same subnet. Walrus is the storage system, which allow user to store data, organized as bucket and object, it is also used to create, delete, and list buckets.

3. Storage Controller (SC):-

Provides block-level network storage including support for Amazon Elastic Block Storage (EBS) semantics.

4. Node Controller (NC):-

Is installed in each compute node to control Virtual Machine activities, including the execution, inspection, and termination of VM instances.

This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

V.CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data(usually large amount of data-typically business or market related-also known as "big data")in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. We regard Big data as an emerging trend and the need for Big data mining is rising in all science and engineering domains. With Big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCES

[1] Xindong Wu, Gong-Quing Wu and Wei Ding " Data Mining with Big data ", *IEEE Transactions on Knowledge and Data Engineering* Vol 26 No1 Jan 2014

[2] Bharti Thakur, Manish Mann **"**Data Mining for Big Data: A Review" Volume 4, Issue 5, May 2014

[3] Wei Fan and Albert Bifet "Mining Big Data: Current Status and Forecast to the Future",Vol 14,Issue 2,2013

[4] Dunren Che1 , Mejdl Safran1 , and Zhiyong Peng ,"From Big Data to Big Data Mining: Challenges, Issues, and Opportunities"2013

[5] Sagiroglu, S.; Sinanc, D. ,"Big Data: A Review" May 2013

[6] Garlasu, D.; Sandulescu, V.; Halcu, I. ; Neculoiu, G.,"A Big Data implementation based on Grid Computing", Jan. 2013

[7] Garlasu, D.; Sandulescu, V.; Halcu, I. ; NeculoiuG.,"A Big Data implementation based on Grid Computing", Jan. 2013

[8] Aditya B. Patel, Manashvi Birla, Ushma Nair "Addressing Big Data Problem Using Map Reduce" reports Dec. 2012

[9] Thomas H. Davenport Real Time Literature Review About the Big data According 2012

[10] Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44, 2009.

[11]U. Fayyad. Big Data Analytics: Applications and Opportunities, 2012.

[12] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data:, 2013.

[13] J. Gama. Knowledge Discovery from Data Streams. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[14] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.

[15] Gartner, http://www.gartner.com/it-glossary/bigdata.

[16] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza.7 {11, New York, NY, USA, 2012. ACM.

[17]Intel. Big Thinkers on Big Data, big-thinkers-on-Big-data.html, 2012.

[18] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.

[19] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

