View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by MAT Journals

MAT JOURNALS
e-ISSN: 2456-9437
Volume 4 Issue 1

A Review Paper on Management of Big Data

Kaustubh[1*], Vikash Singh[1], Madhur Ahuja[1], Gaganjot Kaur[2]
[1]B. Tech Students, [2]Professor
Department of CSE, Manav Rachna University
Faridabad, Haryana, India
Email: *kaustubhd287@gmail.com*
DOI: *http://doi.org/10.5281/zenodo.2580487*

Abstract

Big Data is everywhere around us.In previous era,there is an increase in demand of big data,business analytics and work environment.The big data is dominantly practice-driven,the corporations are exploring how big volume of data can be used to produce and store values for the company and governments.Though the machine learning and web analytics is to guess the action of the individuals,choice of the consumer or search action.Big Data is rapid tool that not only study patterns, but also give the diving possibility of an event.Corporations have stepped on the trend of using the big capacity of data that is forever increasing,often in terra or petabytes,to better guess results with large accuracy.The United Nation has passed another initiative named Global Pulse that manipulates new digital data sources such as mobile calls or mobile payments.Now big data has become a mainstream as a corporate terms,there is a small published management scholarship that handles the provocation of using such tools.In this research paper, we explore about the concepts of Big Data and how to manage it.

Keywords:*Big Data, Managing Big Data*

INTRODUCTION

Big data produced from an increasing range of sources such as internet clicks,mobile transactions, user-generated content and social media and resolved generated content through sensors networks,or business transactions such as sales queries and purchase transactions. The healthcare, operations and finance all add to big data pervasiveness. Big data nessecisate the use of powerful computational techniques to unveil trends and pattern.The misnomer invariably attracts researchers' attention to the size of the dataset.The nature of data of defining parameter of big data is fine-grained itself. For example, a participant in a cycle race generates 20 gigabytes of data from its 150 sensors on the cycle that can help analyze technical performance of the components, but also cyclist reactions, pit stop delays, and communication between crew and cyclist that contribute to overall performance.

Sources of Big Data

Big data is also the binding for different types of granular data. The five main sources of high volume data is Public data, Private data, Data Exhaust, Community Data, and Self-Quantification Data.

Sources

UGC

This variety of data is sourced from applications with user population in millions. For example, Twitter, RSS feeds, Blog posts, multimedia sharing services.
This data is structure and pattern less as it is directly uploaded (shared) by its users.

TD

This variety is generated by the huge number of transactions taking place all over the world. For example Web logs (which is in a huge amount at this time), transactions made of all kinds. This data is structured with the help of a set of rules or a schema.

Page 11-15 © MAT Journals 2019. All Rights Reserved

SD
This variety of data is generated from data driven and dependent science experiments and processes. The incoming data is usually large sized and frequent in transfer.For example, Data from the Hubble, Genetics and the research data of unanswered questions.This data may be highly structured or structure less as well, depending on its source.

Whatever may be the source, there is a dire need to manage and provide safekeeping for the big data. Most of the big data is unstructured, which means that they contain a vast amount of knowledge which need to be segregated first. This structuring helps to fit the big data in a well defined data model.

If the data is in a relational structure (schema), it requires the current RDBMSs available.

The current RDBMS softwares are single threaded and their performance falls almost exponentially with the data limit crosses a hundred GBs. These softwares traditionally run on ACID which hinders their performance in case of big data.

Some modern DBMS handling softwares like Oceanbase works on parallel and distributed computing, which helps the software to scale accordingly with the data.

Figure 1: *Big Data and Analytics*

Public Data: These data are grasp by governments, governmental organizations and local communities that can potentially be saddled for wide rangeing business and management application. Examples of such data include transportation, energy use, and healthcare that can be accessed by under certain restrictions to guard individual privacy.

Page 11-15 © MAT Journals 2019. All Rights Reserved

Private Data: These data are also held by private firms, non-profit organsations and individual that display private informations that cannot be assigned from public sources.Examples of Private Data is that it include consumer transactions, organizational supply chains using RFID tags, movement of company goods and resources, website browsing, and mobile phone usage among several others

The Four Vs of Big Data

Volume: It means that all types of the data is created or generate by the different types of the resources and expand continue. The good reason of the grathering of huge amount of data includes the pattern and information is hide to data analysis. The aforesaid intiative is called mobile data challenge motivated by nokia. [1]. These challenges tells an interesting result which is similar to the examine of the human behavior.

Variety: The data is collected from the different sources such as mobile phone, social sites, sensors,images, and the data can be unstructured or structured. Example of unstructured is mobile phone. The Internet guests also generated an huge set of both unstructured and structured data.[2].

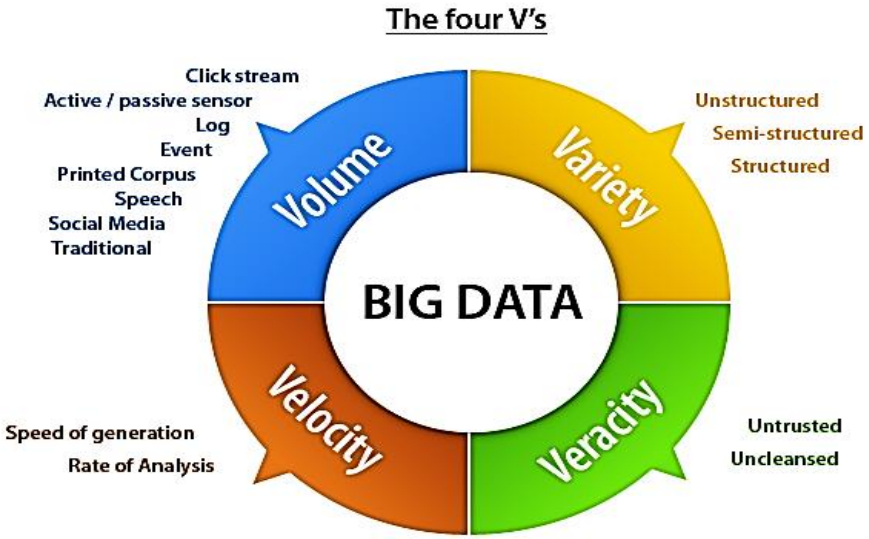

Figure 2: *4Vs of Big Data*

Velocity It indicates the rate of the data is transferred. Due to the absorption of complementary data collection the content of data is changes regularly. ThePreviously data or collection data arrive from more than one sources.[3]

Value The big data most important aspect value. It works is to from the huge datasets it discover the hidden values and the datasets are containing different types and rapid generation

Big Data in Management Research

The regularly change in the field of digital economy has challenge the concepts of business and traditional ecomonic .User genrates large volume of data which is transfered and analysed in different sector. A mere tweet is came from the trusted source then a chain reaction in the blogs and social network which causes loss or profits of billions of dollars to that source. The information for this stiutation makes good but even more difficult to value because it has impact on real time decision making. In private or public area the mobile technology and banking services are the sourcses of Big data low tech services like water and electricity can changes societies and communities. There is additional use of big data which have

Page 11-15 © MAT Journals 2019. All Rights Reserved

implications for communities and societies. Most of the previous database systems was design for enterprise infrastructures and but not design for meet the future goals. This called up for the data management systems for cloud infrastructures. Due to regularly usage of internet application genrates the huge amount of data.Nowdays the industries adopted the hadoop framework which is running succesfully in managing the data.but hadoop is also need some improvement [4]. Many applications are starts with small database. But growing with the popularity therefore their data footprints also grow and a point came where the grow limit is stop due to database efficiency

The another important area for the data management the huge number of application that need to be support in the cloud in which each has small data footprint.this is called as large multitenant system.the example of the multitenant system is salesforce.com which has database multitenancy considered in case of the SaaS.In this same database tables, is shared by different tenants.In context of different cloud paradigms different models of multitenancy are relevant. The techniques which are well define in the the past big data create new challenges and chances to these technology nowdays.big data took new approach to data reduction. The two challenges for big data reduction are:

1-Machine learning
2-Parallel Processing

Machine learning to improve the traditional data reduction techniques machine learning is a way.big data needs technologies which take elapsed time to process huge quantity of data this is because traditional techniqies not briefly give all records.As a result from that techniques machine learning is able to understand the data trend.so machine learning gives solution to this big data.

Parallel Processing To reduce the data reduction it is another way. massively parallel processing is also include in this and also distributed databases,cloud computing,scalable stroage is part of this.Among all above cloud computing is the best or successfulbecause cloud computing is a model which provide on -demand network access to shared pool.with some improvement in cloud computing is a way tio reduce or minimze the rate of big data. And also cloud computing uses the parallel process to reduce the data reduction

Querying and Indexing

In this era of big data which technically should now be called huge data, there is a dire need to manage and store this data efficiently. The motive is to store the data not as archives but with a goal to extract it whenever needed and in lesser time. Till recently we were using traditional structured RDBMSs to store our big data. It is time we realise that those cannot help us anymore.

We come across applications with growing popularity and the need to scale. The application itself may do that but the concern is of data management. We need distributed and parallel computation to overcome that.

Techniques

Segregation

This is the first step of managing big data. This is where the concept of tags comes in. Most of the social media giants use tagging to segregate and lessen the data heap.

Distributed

A single machine (no matter how powerful) is not a good idea to store our data on. The size or the complexity of data may prove to be too much for a machine.

Key-Value

The limitation of traditional rdbmss to scale has given the popularity to Key-

Page 11-15 © MAT Journals 2019. All Rights Reserved

Value stores. Examples of this are PNUTS, Cassandra etc. These systems have been extensively used in personal and professional infrastructures. These are also know as NOSQL systems. Its popularity is the result of its tremendous success in dealing with data with PBs in size and multidimensional complexity.

Tree

Big data doesn't only refer to the data that is big in size, but also that is high in complexity. Data which has considerable size but is complex and heterogamous might prove to be a challenge. Although K-V systems are capable of handling it, a tree syntax representation is becoming popular with complex data. Querying big data might be difficult in tree usage, but a better fault tolerance goes a long way.

An example of tree usage is the BATON system (Balanced tree overlay network). It is a balanced tree structure, which is capable of providing a peer to peer networking with exact and range query capabilities. It guarantees the exact query in a O (log N) complexity in an N nodes network.

Hadoop Framework

Hadoop is an open-source framework from Apache which helps the user to enable distributed processing of large datasets. One correlating feature is the Map Reduce framework by Google. It is a programming model processing superpowers. It is based on GFS and implemented through Hadoop. It mainly processes big data on the cloud and scaling applications deployed on the cloud.

CONCLUSION

This paper develops an overview of the various methodologies used in big data management. However, there are other methods that are used or were used in production. The pin point details of such methods are beyond the current scope of this paper.

REFERENCES

1. J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: Big data for mobile computing research, Workshop on the Nokia Mobile Data Challenge, in: Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, pp. 1–8.
2. D.E. O'Leary, Artificial intelligence and big data, IEEE Intell. Syst. 28 (2013) 96–99.
3. M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 19
4. Apache Hadoop, http://hadoop.apache.org.

Cite this article as: *Kaustubh, Vikash Singh, Madhur Ahuja, & Gaganjot Kaur. (2019). A Review Paper on Management of Big Data. Journal of Data Mining and Management, 4(1), 11–15.*
http://doi.org/10.5281/zenodo.2580487

Page 11-15 © MAT Journals 2019. All Rights Reserved