International Journal of Research
eISSN: 2348-6848 & pISSN: 2348-795X Vol-5 Special Issue-13
International Conference on Innovation and Research in Engineering, Science & Technology
Held on 23rd & 24th February 2018, Organized by Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur, 441108, Maharastra, India.

Comparative Analysis on Big Data vs RDBMS

manisha patle[1], shubhangi bulle[2] , ruchita bhagat[3] , monali wade[4] & ashwin pawar[5]
[1]manishapatle09@gmail.com,[2]shubhangibulle96@gmail.com,[3]ruchitabhagat82@gmail.com,
[4]monaliwade66@gmail.com,[5]ashwinpawar435@gmail.com
[12345]dept. of computer science and engineering
Tulsiramji Gaikawad Patil college of Engineering and Technology

Abstract:

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; heterogeneous mixture; autonomous sources; complex and evolving associations; hadoop; Hadoop Distributed File System(HDFS).

1. Introduction

An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, big data is:

- Large datasets
- The category of computing strategies and technologies that are used to handle large datasets

In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

So how is data actually processed when dealing with a big data system? While approaches to implementation differ, there are some commonalities in the strategies and software that we can talk about generally.

The general categories of activities involved with big data processing are:

- Ingesting data into the system
- Persisting the data in storage
- Computing and analyzing data
- Visualizing the results

1.1 Characteristics of big data

Veracity: The verity of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data.

Variety: Big data problems are often unique because of the wide range of both the sources being processed and their relative quality. Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers.

Velocity: Another way in which big data differs significantly from other data systems is the speed that information moves through the system.

Volume: The sheer scale of the information processed helps define big data systems. These at each stage of the processing and storage lifecycle.

1.2 Introduction to RDBMS:

Most of the problems faced at the time of implementation of any system are outcome of a poor database design. In many cases it happens that system has to be continuously modified in multiple respects due to changing requirements of users. It is very important that a proper planning has to be done.

RDBMS stands for Relational Database Management System. RDBMS data is structured in database tables, fields and records. Each RDBMS table consists of database table rows. Each database table row consists of one or more database table fields. RDBMS store the data into collection of tables, which might be related by common fields. RDBMS also provide relational operators to manipulate the data stored into the database tables. Most RDBMS use SQL as database query language.

1.3 Big data characteristics:

Huge Data begins with huge volume, heterogeneous, self-sufficient sources with circulated and decentralized control, and tries to investigate complex and developing connections among information. These qualities make it a compelling test for finding helpful information from the Big Data. Creator can envision that various visually impaired men are attempting to scrutinize a mammoth elephant which will be the Big Data in this connection.

The objective of every visually impaired man is to draw a photo (or conclusion) of the elephant as indicated by the piece of data he gathers amid the procedure. Since every individual's perspective is constrained to his neighborhood area, it is not astounding that the visually impaired men will each finish up autonomously that the elephant "feels" like a rope, a hose, or a divider, contingent upon the locale each of them is restricted to.

1.2 Advantages:

1. The capability of Hadoop for volumes to manage vast amounts of data, in or out of the cloud, with validation and verification.
2. Identifying significant information that can improve decision quality.
3. Addresses speed and scalability, mobility and security, flexibility and stability.
4. Integration of both structured and unstructured data.
5. Replicability, extension to new domain.
6. 'Total' datasets, 'whole universe'.
7. No sampling needed, data for all behaviour and over whole existence.
8. Ready made manipulability.
9. Powerful relation of data to object.

2. Conclusion & Further Work

Big Data is the term for a gathering of complex information sets, Data mining is a systematic procedure intended to investigate data (usually extensive measure of information ordinarily business or business sector related-otherwise called "Big data") in inquiry of steady examples and afterward to accept the discoveries by applying the distinguished examples to new subsets of information.

To bolster Big Data mining, elite figuring stages are required, which force precise plans to unleash the full force of the Big Data. We see Big Data as a developing pattern and the requirement for Big information mining is ascending in all science and building areas.

With Big Data innovations, we will ideally have the capacity to give most pertinent and most precise social detecting criticism to better comprehend our general public at constant. In proposed work we compared data mining operations in Big Data and RDBMS system with clustering and other query operations.

Through Experimentation and result analysis of time comparison we can see that the data mining in Big Data (HDFS file system) is efficient as compared to RDBMS in terms of time required for accessing information.

In future we tend to compare much more different operations and find a way to provide much efficient system using different algorithms. Big Data systems can be made more efficient using different clustering and classification techniques.

3. Acknowledgements

This work was supported in part by a grant from the National Science Foundation.
Contribution of others who might have given suggestions or review comments.

4. References

1. U. A. Acar. 2009. Self-adjusting computation: An Overview, in Proc. of PEPM'09, New York, NY, USA.
1. T. Karagiannis, C. Gkantsidis, D. Narayanan and A. Rowstron. 2010. Hermes: Clustering users in largescale e-mail services, in Proc. of SoCC '10, New York, NY, USA.
2. Memcached-A distributed memory object caching system, http://memcached.org/, 2013.
3. P. Scheuermann, G. Weikum, and P. Zabback. 1998. Data partitioning and load balancing in parallel disk systems, The VLDB Journal. 7(1): 48-66.
4. M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica. 2008. Improving MapReduce performance in heterogeneous environments, in Proc. of OSDI'2008, Berkeley, CA, USA.
5. H. Herodotou, F. Dong and S. Babu. 2011. No one (cluster) size fits all: Automatic cluster sizing for data-

intensive analytics, in Proc. of SOCC'2011, New York, NY, USA.

6. D. Logothetis, C. Olston, B. Reed, K. C. Webb and K. Yocum. 2010. Stateful bulk processing for incremental analytics, in Proc. of SOCC'2011, New York, NY, USA.
7. C. Olston, G. Chiou, L. Chitnis, F. Liu, Y. Han, M. Larsson, A. Neumann, V. B. N. Rao, V. Sankarasubramanian, S. Seth, C. Tian, T. ZiCornell and X. Wang. 2011. Nova: Continuous pig/Hadoop workflows, in Proc. of SIGMOD'2011, New York, NY, USA.
8. C. Olston, B. Reed, U. Srivastava, R. Kumar and A. Tomkins. 2008. Pig latin: A not-so-foreign language for data processing, in Proc. of SIGMOD'2008, New York, NY, USA, 2008
9. Luiz A. Barroso, Jeffrey Dean, and Urs Holzle. ¨ Web search for a planet: The Google cluster architecture. IEEE Micro, 23(2):22–28, April 2003.