International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Big Data, RDBMS and HADOOP - A Comparative Study

Roopa Raphael[1], Raj Kumar T[2]

[1, 2]Department of Computer Science and Engineering, College of Engineering Kallooppara, Kerala

Abstract: ***Big Data and Hadoop are the recently trending words in the world of internet. What is meant by Big Data, Hadoop, from where did they evolve and what is being replaced by them is a major question that exists today? RDBMS and its benefits had created a revolution in the field of data handling. What caused a reframing or casting away of RDBMS aside from data analysis and paving way for the new data handling system known as HADOOP is a great question. To how much extent have these new technologies been influencing the internet era and the world of data handling and analysis, its users and why are answered here.***

Keywords: RDBMS, Big Data, Map Reduce, Hadoop Distributed File System, Google file System.

1. Introduction

Large Volume of Data is growing because the organizations are continuously capturing the collective amount of data for better decision making process**.** Volume of data increases by online contents like blogs, posts, social networking site interactions, photos that are created by the users and servers and records continuously the messages about what the online users are doing. The business today is affected by this unexpected data growth. Daily 2.5 quintillion bytes of data are created according to the estimation done by IBM and it is said to be in such huge amount that 90% of data in the world created in last 2 years [2]. It is a mindboggling figure and but the sad part is instead of having more information, people feels less conversant.

Industrial organizations are essentially intended to make sense from the massive growth of Big Data in order to develop the analytic platforms for producing the traditional structure data which includes semi-structured and unstructured sources of information. Likewise, industrial organizations can reap benefits of Big Data processing for better decision making process. The Big Data‘s success depends on its analysis. Big Data analysis is an unending process and for this there are a required set of activities instead of an isolated activity.

Big data is a buzzword, or catch-phrase, utilizes to describe a massive volume of both structured and unstructured data that is so huge that it's complicated to process using traditional database and software techniques. In most enterprise situations the data is either very large or it moves too fast or it exceeds current processing capacity. Big data has the prospective to aid organizations to rally operations and make sooner, more intellectual decisions [3].

Though big data doesn't mention to any specific amount, this term is often used when speaking about petabytes and Exabyte‘s of data [4]. An e xample of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data entailing of billions to trillions of records of millions of people—all from diverse sources. The data is structured in an unpredictable way such that the data is often incomplete and inaccessible [3].

I. 4V’s of Big Data

1. Volume

The main characteristic that makes data ―big‖ is the sheer volume. It makes no logic to emphasis on minimum storage units because the total amount of information is growing exponentially every year. In 2010, Thomson Reuters evaluated in his annual report that it assumed the world was ―awash with over 800 exabytes of data and growing‖. In that same year, EMC, a hardware company that makes data storage devices, believed it was nearer to 900 exabytes and would grow by 50 percent every year. No one essentially knows how much new data is being produced, but the amount of information being collected is huge.

2. Variety

Variety is one the most interesting developments in technology as more and more information is digitized. Out-of-date data types (structured data) include things on a bank statement like date, amount, and time. These are things that are adequate in a relational database. Structured data is augmented by unstructured data, which is where things like Twitter feeds, audio files, MRI images, web pages, web logs are put — anything that can be captured and stored but doesn‘t have a *meta model* (a set of rules to frame a notion or idea — it defines a class of evidence and how to express it) that neatly defines it.

By *unstructured data*, on the other hand, we mean there are no rules. Pictures, voice recordings, tweets — they all can be poles apart but express ideas and thoughts based on human understanding. One of the objectives of big data is to use tools to take this unstructured data and make sense of it.

3. Veracity

Veracity refers to the trustworthiness of the data. Can the manager depend on the fact that the data is descriptive? Each worthy manager knows that there are intrinsic discrepancies in all the data gathered.

4. Velocity

Velocity is the frequency of incoming data that needs to be processed. Think about how many SMS messages, Facebook status updates, or credit card swipes are being led on a specific telecom carrier each minute of every day, and we'll have a good appreciation of velocity. A streaming application like Amazon Web Services is an example of an application that levers the velocity of data.

II. RDBMS

A **relational database management system** (**RDBMS**) is a database management system (DBMS) that is based on the relational model. Countless common databases presently in use are grounded on the relational database model. RDBMSs are a shared choice for storing of information in new databases used for financial records, manufacturing and logistical information, personnel data, and other applications. Relational databases have frequently switched legacy hierarchical databases and network databases because they are easier to understand and use. Nevertheless, relational databases have received unsuccessful challenge attempts by object database management systems in the 1980s and 1990s. Also by XML database management systems in the 1990s also provided the same. In spite of such efforts, RDBMSs possess most of the market share, which has also grown over the years.

RDBMS stock the data into assembly of tables, which might be linked by common fields (database table columns). RDBMS also offer relational operators to handle the data stored into the database tables. Utmost RDBMS use SQL as database query language. A significant feature of relational systems is that a particular database can be spread across several tables. This fluctuates from flat-file databases, in which all databases are self-reliant in a single table. Nearly all full-sized database systems are RDBMS's. Small database systems, yet, use other designs that offer less flexibility in posing queries.

III. HADOOP

Hadoop is an open-source software structure that ropes data-intensive distributed applications. It allows applications to work with thousands of computationally self-governing computers and with petabytes of data. Hadoop increases the storage space and the processing power by uniting many computers into one. Hadoop devises two parts: HDFS (Hadoop Distributed File System) file system and MapReduce programming paradigm. It has been formed by Dong Cutting and Mike Cafarella in 2005. It was developed to upkeep distribution for search engine project. It is certified under APACHE LICENSE 2.0. This is written in Java Runtime Environment (JRE) 1.6 or advanced version. The operating system is cross-platform. It was developed by Apache Software Foundation. Hadoop came as a derivative from Google's Map Reduce and Google File System (GFS).

The principal of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop separates files into large blocks and allocates them among the nodes in the cluster. To work on the data, Hadoop MapReduce handovers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality—nodes manipulating the data that they have on hand—to allow the data to be processed faster and more efficiently than it would be in a more predictable supercomputer architecture that depends on on a parallel file system where computation and data are connected via high-speed networking. The HDFS is a distributed file system that provides fault tolerance and is designed to run on commodity hardware. HDFS delivers high throughput access to application data and is appropriate for applications that have large data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers, and a way of running work (ie; Map/Reduce jobs) across those machines, running the work near the data. HDFS devises master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the Hadoop cluster.

IV. Map-Reduce

MapReduce is a high-level abstraction of parallel computing introduced by Google. While traditional approaches need to describe exactly the way to carry out the data to process, MapReduce programming model emphases on the processing code. The key hint of the method is to change the network transfer from the data to the code. In short, the data is circulated using a distributed file system. For achieving a task, the code is then distributed where the needed data is. Therefore, the cloud computing prototype is well matched for problems dealing with huge volumes of data. MapReduce originates from functional programming concepts. Its functions(map and reduce),takes other functions as inputs. The map purposes to divide input data into numerous inputs for applying a function on each of them (mapper). The reduce function applied by reducers combines the individual results from the mappers. These tasks are recognized by a master machine to slave ones. The master is also liable to sense node or network failures by a ping mechanism in order to reassign tasks to others nodes.

Map reduce is a software framework introduced by Google to backing up distributed computing on large data sets on clusters of computers. Map Reduce- It is a programming model. It is castoff for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediary key/value pairs and a reduce function that merges all intermediate values associated with the same intermediary key.

"Map"step: The master node takes in the input. It partitions up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in order, leading to a multi-level tree structure. The worker node then processes

the smaller problem. It then passes the answer back to its master node. Map takes each pair of data of a type in one data domain, and returns a list of pairs in a different domain: Map (ki, vi) → list (Kj, vj)

"Reduce" step: The master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve. The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain: Reduce (Kj, list (vj)) → list (vl).

V. HDFS

The HDFS file system is not restricted to MapReduce jobs. It can also be used for other applications, many of which are still under development at Apache. The list comprises the HBase database, the Apache Mahout Machine learning system, and the Apache Hive Data Warehouse system. According to theory, Hadoop can be used for any sort of work that is batch-oriented rather than real-time, is very data-intensive, and benefits from parallel processing of data. It can be used to counterpart a real-time system, such as lambda architecture.

As from the time Hadoop was developed it came into mainframe. Most of the companies like Amazon, Yahoo, and Google now use Hadoop for its data mining. Some of the commercial applications of Hadoop included:

- Log and/or clickstream analysis of various kinds
- Marketing analytics
- Machine learning and/or refined data mining
- Image-processing
- Handling of XML messages
- Web-crawling and/or text processing
- General-archiving, inclusing relational/tabular data, e.g. for compliance

2. Literature Survey

Handling of data has become far more tedious than that of some years before. As the quantity of data is growing beyond limits, and be bestowed a name for such huge amount of data as ‗BIGDATA' [1] a best possible mechanism had to be deployed for the purpose says a survey on Big Data conducted by Ms. Vibhavari Chavan and Prof. Rajesh N. Phursule. The traditional RDBMS techniques became too old to handle these data as one of its main aspects of redundancy only doubled the job. So a new technology of immense data handling known as Map Reduce was introduced. Of course, it turns out to be a huge success in the field of Big Data management. According to what the inventors of Hadoop says, it could be a better substitute for Map Reduce and it has been confirmed throughout the comparative study.

3. Results and Findings

Comparison between RDBMS and HADOOP

	RDBMS	HADOOP
Description	Traditional row-column databases used for transactional systems, reporting, and archiving.	Distributed file system that stores large amount of file data on a cloud of machines, handles data redundancy etc. On top of that distributed file system, Hadoop provides an API for processing all that stored data - Map-Reduce. On top of this basic schema a Column Database, like hBase can be built.
Type of data supported	Works with structured data only	Works with structured, semi-structured, and unstructured data
Max data size	Terabytes	Hundreds of Zettabytes
Limitations	Databases must slowly import data into a native representation before they can be queried, limiting their ability to handle streaming data.	Works well with streaming data

4. Conclusion and Future Scope

By the above comparative survey we have come to know that HADOOP is the best technique for handling Big Data compared to that of RDBMS. As world moves on, the data used increases and therefore a better way of handling such huge amount of data is becoming a tedious task. Analysis and storage of the so called Big Data is handy only by the help of the new Hadoop eco-system than the traditional RDBMS being used till now. Hadoop is a large scale, open source software framework dedicated to scalable, distributed, data-intensive computing. The framework breaks up large data into smaller parallelizable chunks and handles scheduling, maps each piece to an intermediate value, Fault tolerant, reliable, and supports thousands of nodes and petabytes of data, tried and tested in production, many implementation options.

References

[1] Ms.Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―Survey Paper On Big Data ―, Vol. 5 (6) , 2014, 7932-7939.
[2] Improving Decision Making in the World of Big Data. http://www.forbes.com/sites/christopherfrank/2012/03/25/improvingdecision-making-in-the-world-of-big-data/
[3] Apache HBase. Available at http://hbase.apache.org
[4] Apache Pig. Available at http://pig.apache.org
[5] 4 v's of Big Data. Available at http://www.dummies.com/how to/content/the-4-vs-of-big-data.html

Author Profile

Roopa Raphael graduated the Degree of Bachelor of Technology in Information Technology from Mahatma Gandhi University in 2013. She is now pursuing her master degree in Computer Science with specialization in Cyber Forensics and Information Security at College of Engineering Kallooppara under Cochin University of Science and Technology.

Raj Kumar T graduated the degree of Master of Technology from National institute of Technology Karnataka Surathkal and is presently working as Assistant Professor in Computer Science and Engineering at College of Engineering Kallooppara, Kerala .