THE WIDESPREAD INTRODUCTION TO MAPREDUCE ANDHADOOP FOR THE BEGINNERS OF BIG DATA

B.Kalpana

Assistant Professor, Department of Computer Science and Engineering, Panimalar Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT

Modern world takes data for granted when available at loose. The data when misused can lead to biggest loss to the nation's prosperity. The paper concentrates on the basics of big data which is available in our modern world for providing many services. The paper talks about the source of data conversion into information and the wide rated usage. Some of the examples of current data and its sources consist of many traditional software system applications in combination with many recent techniques and technologies, Images, Handheld devices, servers and so on. The consumers or the predators of the generated information are the business intelligent tools commonly available in the market, databases and some specific warehouses. Without depending on the specific area there are huge volumes of data generated in the day to day life. The important role as an enterprise is how to manage such volumes of data for a better performance and how to use the data in a safe manner to the extend.

Keyword: *Big data, MapReduce, Hadoop*

---

1 Introduction to Big data:

Big data cannot be defined directly since it concentrates on some of the correlated terms. In terms of types of data used 85 to 90 percent of the available data is mostly unstructured and many of the organization are not capable of handling the unstructured data. For the past 10 years many technologies were invented to struggle hard on relational data. But big data is just unbeatable and have the capability to adapt to many of the analytical and probabilistic environment and applications. The big data solutions provided are the master of the masters to shape any kind of business and to give a new solution based system.

1.1 Overview of Big data:

Big data is not enterprise data but it is compatibility a large data in terms of latitude or longitude. They are generated around the enterprise and its related applications. It majorly comprises of semi-structured data and semi/unstructured information in very large volumes.

In order to get a very big result oriented output from the big data it has to be coupled with the basic level enterprise dates using queries and reports.

Eg 1: An e-cart wants to link the web page visitor's selection cart details with the back end log as the purchase details using the SQL or by using Access.

Eg 2: A mobile phone dealer want to display his website with recently released mobiles in all the 4 views (top view, front view, bottom view, side view) in terms of unstructured images with zooming options.

1.2 Big data and its challenges:

Some of the day to day increasing challenges faced by big data can be listed out as:

- Storage Capacity: Image capturing, storing and managing is a major limitation.
- Cost of Implementation: TEEL (Transforming Exact Extraction and Loading) operations encountered in big data implementations.
- Constant Information Accumulation: Information gets accumulated when there is a drastic growth in semi structures data. The process is rapid, fast and very quick regardless of the current challenges.
- Power Factors for Performance: The Power factors in the paper concentrates on the processing of data in terms of information and integrity. The performance factor is always a threat to the larger volume of data in terms of an enterprise.

The above explained complication shows the pull backs for implementing big data. When more insights are not given in detecting the risks leads to loss of revenue and higher expenses for solving the risks.

The solution for the above specified limitations can be providing by means of MapReduce. There are many recently developed techniques for handling unstructured data but a better solution can be provided using Hadoop.

2 Introduction to MapReduce:

The basic building block of MapReduce is the well known concept of divide and conquers technique. The process of breaking a very large problem into smallest units and computing the smaller problems in parallel, obtaining the results and combining all the results together is called as divide and conquer technique. Even though it may appear to be old-fashioned the method is a best solution for solving many unique problems and is the best example for understanding the strength of parallel processing and dynamic computing. It has the capability to fully digest a very large semi structured data which are been produces during the day to day operations. MapReduce can be studied extensively in the MapReduce design documents.

The innovative old fashioned technology is most powerful for handling the mountain of unstructured data.MapReduce is considered to be the base for the other technologies like Hadoop, and Hadoop Distributed File System (HDFS).

3. Introduction to Hadoop:

MapReduce is considered to be the best for beginners but it needs a lot of technological resources, knowledge in development of the resources to implement it in a specific organization. The feasibility of the same is not possible by all organizations. The limitations led to the development of Hadoop.

Hadoop is defined as the standard open source framework which is very well adopted and built on the famous MapReduce for Google and GFS papers (Google File System). Hadoop has the capability to unify the process of parallel processing which majorly takes the advantages of Big Data System by using consistent amount of servers. Hadoop also removes the limitations faced by distributed processing system. The developer majorly focuses on the Result Oriented System rather than the process oriented environment. The Maintenance of Hadoop is taken over by the Apache software Foundation in combination with Yahoo! for providing a no-profit contribution. Some of the critical functionalities of Hadoop when implemented are

1. Requirement of Latency
2. Providing Excellent Support for the Service Level Agreements (SLA)
3. Compound Architecture
4. Various platform compatibility for any Specific Application and
5. Cost Compatibility

The detailed explanation for the functionalities is as follows,

3.1 Requirement of Latency

Hadoop has the capability to perform and execute various kinds of tasks with less overhead. The processes are prioritized based on the group and real time implementation.

3.2 Providing Excellent Support for the Service Level Agreements (SLA)

Hadoop also has the capability to deliver,

- Several levels of predictability
- National/International support
- Availability of resources
- Quality performance
- Wide scalability and
- Higher level of utilization.

3.3 Compound Architecture

Hadoop takes its data and information from various sources including traditional and recent trends. Hadoop has the capability of excellent storage of data which includes independence of data from MapReduce programming and higher flexibility from Distributed File System (DFS).

3.4 Various Platform Compatibility for any Specific Application

The platform compatibility provided by Hadoop is non-match able which includes

- Open standards compatible nature
- Execution of pipelining with many types of programming languages, and
- Global architecture with no pattern rights with no specific vendor lock out.

3.5 Cost Compatibility

Hadoop has many inbuilt benefits to the organization which can be implementing with less cost. It has a very attractive FLOC (FLexibility Owner Cost) and HRIP (High Return of Investment Policies).

4 Correlationand Association between Big data, Hadoop and MapReduce:

Modern organization comes across tons and tons of newly emerging data which take up many different forms. Big data has the unique capability to provide a detailed insight which can transform the data according to the business handled. It also has the capability for the upcoming modern enterprises with MapReduce to support many kinds of open architectures.

MapReduce is the recently developed framework for solving/breaking a very large complicated problem into smaller units and then solving them and finally combining the results which uses specifically the famous " divide and conquer" technique in combination with commodity servers created by Google and employed successfully. It also has the capability to solve the problem very quickly. But when it considered to be very complex so which has consequently led to the development of Hadoop.

Hadoop is also called as the master stack to implement the complex concepts of the MapReduce and to explore a variable Big data. Hadoop has now occupied a huge market in 2015 and also works like a king of kings in front of the commercial sources.

5 Important role of MapReduce in 2015

Traditional Technique handled the cases of excellent powerful data by using highly compatible servers. The servers had a high memory and very much compatible processor speed. They also ended up with Relational Data

Base Management System (RDBMS), with powerful software etc. But Google had faced large volume data problems and hence resulted in MapReduce technique to handle the problems of multiple servers.

Working of MapReduce:

Let us study about the working of MapReduce in three broad sections such as

- MapReduce needs how much data.
- MapReduce Architecture
- Execution of MapReduce.

5.1MapReduce needs how much data:

Many hundreds and thousands megabytes and gigabytes of data are handled by MapReduce. It's very complex to use and handle these tons of data with MapReduce. Even though MapReduce has the capability to work with raw/pure data but the unstructured nature will lead to many passes. But the typical patterns in Relational Data Base Management System (RDBMS) and Distributed File System (DFS) use the Hadoop Distributed File System (HDFS) where the processing speed increases.

5.2MapReduce Architecture:

MapReduce Architecture comprises of 2 steps Map and Reduce to step into Action.

5.2.1 Map

MapReduce techniques uses KV pairs i.e. key/value pairs. The key is used to describe the type of data that the user is looking for which is normally equated to a single column. The key can take up a Name, Account number or Aadhar number of a person. The value is the instance of a particular data which is directly associated to the key. The value can be integer, floating point or a string value.

The Mapping phase of the MapReduce architecture is used to take the data from the Map function in terms of KV pairs. It then produces the intermediate results along with the Input.

5.2.2 Reduce

When the Mapping phase is completed all the intermediate results are joined together to form a final list. It also combines the intermediate values with the half generated output key in combination with the Input Key. It includes the technique of parallel processing and smashes the tons of data/information very quickly before the configuration of the technique.

Some of the very common important assumptions of MapReduce are:

- All the expensive hardware has the capability to cause failure at high rate.
- Each of the data will be in terms of some small number of MB's to large number of GB's.
- All the files in the execution are write once only (woo) but are given flexibility for add ons or appending it.
- We can have many kinds of streaming videos in large quantities at a specific time period.

5.3 Execution of MapReduce

Let us consider an example to visualize the execution of MapReduce. In this case, let us say person A is an operation manager who is running a website for online imitation jewellery. Person A has over 1,50,000 products online and have around 8000 customers /visitors of the website/per day. In an average around 4500 orders are places per day through the online website. These constructed data measures around some hundreds of Giga bytes of data when these pictures are zoomed in and zoomed out. When the dispatching department wants to know about the other related products searched with the currently placed orders again several tons of data are downloaded and investigated. The final data comes up with a sorted array of elements.

MapReduce can be implemented as follows

- Break the initial data into small number of GB's specifically 5 to 7 GBs.
- Then equally distribute these data to several other nodes.
- On each of the node the Mapping phase will produce a unique list consisting of the frequently searched word in the webpage and produces this list as the intermediate result.
- In the Reducing phase will collect and consolidate the intermediate responses of the mapping phase with the total search list across the files.

6 Application of MapReduce:

When MapReduce is used then we can handles terabytes of data at ease, many servers can be used easily even in combination with cloud and Hadoop can be used when needed at our side and on-demand.

Some of the wide applications of MapReduce are:-

- Social networking sites webpages
- Government organization
- Life and Health care services
- Retail and Financial services
- Home security products and surveillance cameras.
- Defense departments
- Security & threat analysis
- Advertising agencies
- Traffic and congestion control monitoring
- Basic Risk Modeling frame works.
- Emission control and Reporting services
- Auto/Vehicle option Modeling and
- Genomic research and analysis.

7 Conclusion

The paper gives a very detailed view of Big data along with the overview, the challenges faced in 2015. The paper also speaks in depth about MapReduce and Hadoop technologies with functional details of Hadoop. The relationship between Big data, MapReduce and Hadoop is clearly specified for understanding even in terms of a common man. The Importance and role of MapReduce is explained with architecture and application in detail. Hence MapReduce is successful in bringing the computational need for data, apart from moving the data across the wide network. When MapReduce used with combination of Hadoop the limitations of unstructured data is broken and provides a wise energy efficient network traffic control which leads to relatively very high performance.

References

1. Ddanahboyd (2010-04-29). "Privacy and Publicity in the Context of Big Data". WWW 2010 conference. Retrieved 2011-04-18.
2. Jones, MB; Schildhauer, MP; Reichman, OJ; Bowers, S (2006). "The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere" (PDF). Annual Review of Ecology, Evolution, and Systematics 37 (1): 519–544. doi:10.1146/annurev.ecolsys.37.091305.110031.
3. Boyd, D.; Crawford, K. (2012). "Critical Questions for Big Data". Information, Communication & Society 15 (5): 662. doi:10.1080/1369118X.2012.678878.

About the Author

B Kalpana is an Assistant Professor at Department of Computer Science and Engineering at Famous Anna University Affiliated Institution, Tamil Nadu, India. She received her Bachelor of Technology Degree in Information Technology from Sri Venkateswara College of Engineering Chennai with First Class and Distinction affiliated to Madras University in 2003 and Master Degree at Anna University, Chennai. She received her Diploma Degree in Computer Science from Panimalar Polytechnic Chennai with First Class and Distinction with a Gold Medal affiliated to Directorate of Technical Education in 2000.She has several high level involvements in the area of Artificial Intelligence and Big data. She has nearly 11 years of academic experience in the field of Engineering and guided many projects. She has published many papers on Dependable and secure computing and in the area of Big data. She is also an Associate Editor of Information Science & Engineering to the Editorial Review Board of esteemed International Journal of Entrepreneurship and Small & Medium Enterprises (IJESMES), Kathmandu, Nepal.