EXPLORING THE MAP REDUCE AND R TECHNIQUES USED IN LARGE SCALE DATA ANALYSIS

PRADEEP KUMAR SHRIWAS

Research Scholar, Dept. of Computer Science & Engineering,
Sri Satya Sai University of Technology & Medical Sciences,
Sehore, Bhopal-Indore Road, Madhya Pradesh, India,

Dr. Satendra Kurariya

Research Guide, Dept. of Computer Science & Engineering,
Sri Satya Sai University of Technology & Medical Sciences,
Sehore, Bhopal Indore Road, Madhya Pradesh, India.

ABSTRACT

Agriculture, banking, data mining, education, chemistry, finance, cloud computing, marketing, health care stocks, and the stock market are just a few of the many industries that rely heavily on Big Data. Analyzing large amounts of data with the goal of discovering previously unseen patterns, previously unfathomable relationships, and other useful information that can be used to make better choices is known as "big data analytics." Big data's rapidly increasing popularity may be attributed to its wide range of possible uses and broad applicability. The advent of the big data era has presented both advantages and difficulties for the field of intelligence studies. Data mining, data visualisation, semantic processing, etc. are only a few of the new methods of intelligence study that are being examined in this article as they emerge in the context of the big data environment. In the meanwhile, it offers a rundown of contemporary resources like Weka, Sitespace, etc. Both academia and industry are realising the value of large-scale data analysis. In order to analyse massive amounts of data, statistical languages provide a wealth of features while being intuitive to use. To put it simply, Hadoop has altered the dynamics and economics of supercomputers. It allows for scalability at a reasonable price. The article provides an introduction to Hadoop and R for large-scale data processing.

Keywords: - Big Data, Mining,Advanced,Techniques, Large Data.

I. INTRODUCTION

Technically speaking, in the era of big data, computer technologies like as visualisation and data mining provide a potent technological viewpoint for intelligence study, while intelligence understanding gives direction for the development of other approaches. In a similar vein, many big data analysis techniques are now extensively employed in the military intelligence, science and technology intelligence, and the cognitive process of intelligence analysis to direct the development of intelligence analysis tools. Automated systems are crucial for dealing with massive data sets. To prevent misunderstanding, it is important to make the information's substance and relationships clear using a number of technological ways and different technologies.

While large and smaller databases have many similarities, there are notable distinctions between the two. To begin with, they can often deal with a very large amount of records (e.g., millions and greater.) Due to its larger size, such a system has a higher overhead than a medium-sized database, but the faster query times make up for it. Second, transaction logging is present in most big databases. An administrator may "rollback" or "undo" erroneous changes to the database with the use of this function. Finally, huge databases are built to accommodate several users at once. Many hundreds of people may simultaneously query a huge database. It is common practise for large-scale databases to need a substantial financial investment in computer hardware, as well as the availability of knowledgeable personnel and sufficient funds for management, upgrades, and maintenance of the system. Microsoft SQL Server, Oracle, Informix, Interbase, and Sybase are all examples of enterprise-level databases. Only unique research problems

warrant the use of large-scale solutions. For instance, large-scale database management solutions may be used by traffic management systems for cities, counties, and states to organise and maintain their massive data sets.

II. APPLICATION OF BIG DATA TECHNIQUES IN DIFFERENT FIELDS

The development of intelligence research, determined it could not remain on the qualitative analysis or simple statistics. Therefore, the research on intelligence technology puts forward new requirements. The McKinney global institute has Released its research report: big data: the next Frontier for innovation, competition, and productivity. This Report is divided into six parts, including the second part, which discussed the techniques of big data in three aspects: big data Analysis techniques, big data techniques and visualization. In Big data analysis techniques, analysis techniques that Suitable for numerous industries are enumerated, including Clustering analysis, crowdsourcing, data mining, natural Language processing, network analysis, predictive modelling, Regression, visualization, etc. Most of those techniques are the Existing techniques; also some of them are developed based on the original techniques with the development of internet and the demand for large-scale data mining. These techniques can be roughly divided into big data storage and processing Techniques, big data query and analysis techniques, big data Analysis and visualization techniques three categories. Among them, the first two techniques are the foundation of big data, and the last one is the most frequently used in intelligence Analysis domain currently and should be paid more attention And in-depth study. Big data advanced analysis and Visualization techniques mainly include the analysis of data mining and advanced analysis, visual analytics and knowledge Representation, and semantic analysis.

A. Data mining and advanced analysis

The term "data mining" is often used to describe the process of extracting useful information from massive data sets using sophisticated computational methods. Data mining is a subfield of computer science that uses a variety of tools and techniques to analyse databases for hidden insights and insights from raw data. Data mining's purpose is to discover models from massive data stores. Numerous subfields of data mining have been identified by the Task, with correlation analysis and classification analysis being the most common using methods such as cluster analysis, sequence analysis, and decision trees/neural networks, etc.

The data mining algorithm is the heart of large data analysis. Each data mining algorithm provides an objective, scientific look at the features of data based on the specific kind and structure of the data being mined. Also, many commonly used data mining techniques, generally approved by many statisticians, may go deep into data and mine its actual Worth due to the management of huge data.

From the point of view of the data mining notion, there is an obvious tie to the field of information theory. Data Mining has the unique characteristics and implementation procedure that may be used to information research problems. Current informational studies show that many data mining techniques are primarily employed for basic tasks like counting statistics or the frequency of popular terms. These relatively straightforward applications are only useful for preliminary data processing in preparation for deep mining, which is required for knowledge discovery. As a result, data mining may be used to the field of information research; this is not just a byproduct of the growing field of data mining, but also a reflection of the maturation of the information research discipline as a whole.

B. Visual Analytics and Knowledge Representation

If you have a huge data collection with dispersed information and a complex data structure, you may utilise a method called "visual analytics," which involves doing relation analysis via interactive visualisation to help users make decisions and create flawless analytical figures and tables. All occurrences, the data analysis procedure, and the general trend of the data stream are shown graphically and tabularly. Information visualisation, on the other hand, is concerned with the design, development, and use of graphical representations of data that has been automatically created. In order to reach the end objective of decision making, visual analytics evolves out of information visualisation and places special emphasis on the selection of analysis methods and the coupling of analysis methods with visualisation techniques. One of the hottest areas of study in the field of information science, visual analytics has the potential to greatly enhance the efficacy of information analysis in this context.

Using an information visualisation tool, you may study data from a fresh perspective, avoiding the drawbacks of conventional ways of information research. The previously secret knowledge is exposed, explained, and analysed.

Because of its ability to provide useful conclusions for decision making, the efficiency and effectiveness of data analysis are greatly enhanced. Normal end-users and big data exporters are both part of the audience for big data analysis. They can't function without visual analytics as a prerequisite. Because reviewing statistics is as natural as reading a book, and visual analytics may immediately convey the characteristics of huge data.

C. Semantic Analysis

What we call "semantics" is the study of how meaning works. To gather data for the code generation phase, semantic analysis determines if there are any semantic errors in the source code. Semantic analysis involves looking at the bigger picture to ensure the code is organised properly. The goal of semantic analysis is to help computers comprehend, integrate, and reuse organised and unstructured information by combining natural language processing, information indexing, database approaches, and other methods. Semantic analysis is a broad field, but some of the fundamental methods include semantic labelling, knowledge sampling, indexing, modelling, inference, and so on. Deep mining, which uses the semantic process for various forms of information and data mining methods for the structured data with the extracted semantics, may benefit from a solid grounding in semantic approaches.

III. ANALYSING LARGE DATA BY USING MAP REDUCE AND R

Researchers' perspectives have shifted as a result of the explosive development of web-based applications and services during the last decade. The standard method for storing and processing massive amounts of data has been upgraded. The businesses are prepared to invest in trustworthy solutions. Google's MapReduce is a programming methodology and its corresponding implementation for distributed cluster processing of massive amounts of data. In order to use the web as papers, an enormous library of already existing documents must be continually transformed when new documents are added to the index. Databases aren't well suited for storing data processing jobs. The indexing engine at Google uses thousands of computers to store and analyse billions of changes every day. The MapReduce framework is very useful in these circumstances. Hadoop includes the HDFS distributed file system and a version of Google's MapReduce programming model. Hadoop processes data in its native format without any intermediate steps; HDFS handles file distribution and replication among the Hadoop cluster's nodes. The MapReduce methodology is used to process the data.

It takes more time to analyse data because there is always more data to evaluate and that data is expanding, changing, and being altered at a rapid pace in most businesses. Hadoop and MapReduce are capable of handling graph data structures, making it possible to handle the massive and varied data sets. Large, batch-oriented processing, which is often related to task completion in a linear fashion, finds its ideal use in the Map Reduce architecture. Both the MapReduce framework developed by Google and the Hadoop system released as open source place an emphasis on implementing the framework in batches, with the results of each map and reduce stage being written to persistent storage before being used in the subsequent step. Due to the increased likelihood of slowdowns and failures at worker nodes, this materialisation enables a straightforward that is crucial in big deployment. Google's MapReduce is a programming methodology and its related implementation for distributed cluster processing of massive amounts of data.

Large and complicated datasets, whether they be organised, semi-structured, or unstructured, are a challenge for Big Data since they cannot be stored in memory and hence must be handled in a different way. They need local processing, meaning calculations must take performed at the physical location of the underlying data. The 3 Vs model of Big Data (velocity, volume, and variety) would normally be mentioned.

Velocitymeans that the analytics have to be deployed quickly, in real time, with minimum latency. A constant flow of data from a social networking site, for instance, or the combination of data streams from several sources, would be an example of this.

Volumemeans the number of records in the database. Depending on the data's source and destination, its size might be measured in bytes, megabytes, gigabytes, terabytes, or petabytes.

Varietymeans the many forms that information may take, including text, sound, images, and video. As part of Big Data, there are also large datasets.

The term "Big Data Analytics" refers to the practise of applying a very straightforward model to data sets that would overwhelm a more conventional data analysis framework. Hadoop is quickly becoming one of the key solutions for storing and executing operations on data as the quantity of data acquired by companies and corporations expands, particularly the amount of unstructured data.

A. Big Data Analytics With Hadoop

R and Hadoop integration seems to be a logical fit. Both are free and publicly available data-driven endeavours. To work together, though, several basic obstacles must be overcome.

Iterative vs. batch processing - If we take a look at how most individuals do analytics, we can see that it is an iterative process. The user should form a hypothesis before beginning work in R for big data analysis, and then proceed to explore and attempt to comprehend the data, test out various statistical approaches, investigate data at various levels of granularity, etc. This is why R is such a robust platform, and why it provides the best possible conditions for doing this kind of study.

Hadoop, on the other hand, has a batch approach, where tasks are queued and then performed, which may take anything from minutes to hours.

In-memory vs. in parallel - An further complication arises from the fact that R is built to operate with all data in memory, but Hadoop (map/reduce) algorithms are meant to operate in parallel on separate slices of data.

B. Large Scale Data Management Systems With Mapreduce

In 2004, Google first implemented the MapReduce programming paradigm. Using this technique, developers with little to no background in parallel programming may create applications that can scale to handle massive datasets. The Hadoop ecosystem's processing backbone is the MapReduce framework. This framework makes it possible to specify an operation, apply it to a massive data collection, partition the issue and the data, and execute the partitioned and parallelized tasks in parallel.

C. Reducing The Large Data By Map reduce

MapReduce Framework is responsible for processing data in the Hadoop environment. It permits the definition of an operation to be applied to a massive data collection, partitioning the issue into manageable chunks that can then be run in parallel. This kind of work is often implemented as a MapReduce job in either Java or Python. These tasks return their results to HDFS or HBASE. You may do the data analysis in R.

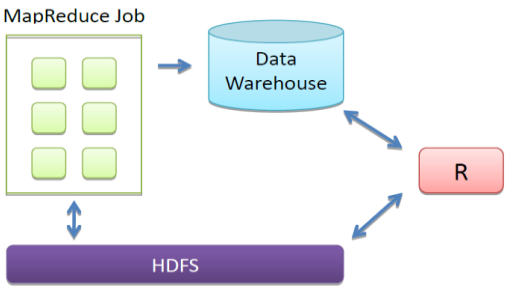

Figure 1: Data Analysis by R

IV. COMBINING MAPREDUCE AND R TOGETHER FOR BIG DATA ANALYTICS

Understanding data storage in Hadoop, how it can be used from R, and why it is vital is crucial if we are to rise to the difficulties of huge data analysis. HDFS, Hadoop's File System, provides the framework for storing data (Hadoop Distributed File System). R developers have the option of using a standalone client to read and write HDFS files. To begin interacting with the Hadoop environment, you must first start a R Session. This feature enables the analyst to simply work with a data subset and start an ad hoc research without consulting other parties, but is still limited by the memory limits of R. It also allows models and other R objects to be saved and retrieved for use in subsequent

MapReduce tasks. The output of running MapReduce tasks is stored in HDFS. The ability to run R code inside the framework of a MapReduce task expands the scope and depth of the analyses that may be performed on massive datasets. Work-compatible problems that can be solved in a simultaneous fashion.

To illustrate, consider the following scenario: An R model is used to assign a score to a dataset. This is accomplished by distributing the model across the Hadoop cluster's Task nodes, executing a MapReduce job to load the model into R on a task node, scoring the data either row by row or in aggregates, and saving the results back to HDFS. Assuming the simplest of circumstances, a Map task will suffice. This acts as a stand-in for the "apply" group of operators in R. The Revolution Analytics framework may also be used for a variety of other applications, including the generation of quintiles, crosstabs, summaries, data conversions, and stochastic computations. These implementations do not presume anything about the data's structure.

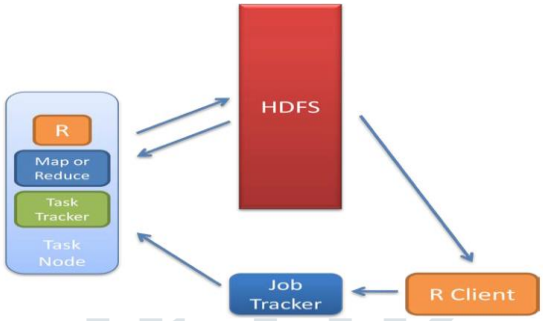

Figure 2: Data Analysis by MapReuce and R on HDFS

R excels in data analysis thanks to its extensive collection of tools but struggles when confronted with very big data sets.

The limitations of Rare as below –

- Requires installation of R on all Task Tracker nodes
- Does not automatically parallelize algorithms
- Different slot/memory configurations is recommended to leave memory and CPU resources for R

Hadoop, on the other hand, excels in storing and processing massive volumes of data, on the order of several terabytes or even many petabytes. These data collections are too large to be handled using memory alone. It is possible to match the analytical might of R with the storage and processing might of Hadoop, or to execute the analysis of a data set on restricted chunks (also called sampling). When R code is run inside the framework of a MapReduce job, the scope and depth of the analyses that can be performed on massive datasets is greatly expanded.

V. CONCLUSION

In the age of big data, cutting-edge methods and resources are required to enhance intelligence studies. As a result, in the subject of intelligence studies, the possibilities and difficulties presented by big data approaches and technologies are mixed. However, the current methods and tools for handling intelligence data, including archiving, processing, and analysis, are not sufficient for the volume of data being collected. In order to better integrate, process, organise, and use big data in order to enhance the quality of data service and boost the intelligence impact in the area of knowledge management and application, we need actively research the application of new techniques and ways to do so. The R community is always working to make the language more scalable. More possibilities exist for parallelizing R over many processors or nodes if Hadoop is employed. Large data utilisation and rising needs in scientific research and high-performance computing are the primary impetuses for this effort.

REFERENCES: -

1. Tyson Condie, Neil Conway, Peter Alvaro, Joseph M.Hellerstein, Khaled Elemeleegy, Russel Sears, "Map Reduce Online", Proceedings in NSDI'10 Proceedings of the 7th USENIX conference on Networked systems design and implementation, pages 1-14, Oct 9 2009

2. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, "Bigtable: A Distributed Storage System for Structured Data", ACM Transactions on Computer Systems (TOCS), Volume 26 Issue 2, June 2008, Article No.4 , pages1-14, 2006

3. Tang Zhixiong, Xian Donglai. The Implement Method of AnalyzingCustomer Retention by Data Mining Technology. Information &Communications. 2011 (2): 99-100

4. Tang Tianbo, Gao Feng. Case Study of Visual Analytics in IntelligenceResearch. Information Studies: Theory & Application. 2009,8 (32): 63-67

5. Liu Zhilong. Data Analysis and Data Mining Application in StatisticsIndustry. Statistics and consultation. 2014 (1): 36-38

6. Hajirahimova, Makrufa & Aliyeva, Aybeniz. (2015). Review of statistical analysis methods of large-scale data. 10.1109/ICAICT.2015.7338519.

7. Pavlo, Andrew & Paulson, Erik & Rasin, Alexander & Abadi, Daniel & DeWitt, David & Madden, Samuel & Stonebraker, Michael. (2009). A Comparison of approaches to large-scale data analysis. 165-178. 10.1145/1559845.1559865.

8. Memon, Mashooque & Soomro, Safeeullah & Jumani, Awais & Kartio, Muneer. (2017). Big Data Analytics and Its Applications. Annals of Emerging Technologies in Computing. 1. 10.33166/AETiC.2017.01.006.