IJESMR

International Journal of Engineering Sciences & Management Research

A SURVEY ON BIG DATA CONCEPTS AND HADOOP MAPREDUCE ALGORITHM

Thirunavukarasu B[*1], Sowbaranika S[2], Keerthiga M[3], Dr Kalaikumaran T[4] , Dr Karthik S[5]
[*1,2,3] Student, Dept of Computer Science and Engineering, SNS College of Technology, INDIA
[4] Professor and Head, Dept.of Computer Science and Engineering, SNS College of Technology, INDIA
[5] Professor and Dean, Dept.of Computer Science and Engineering, SNS College of Technology, INDIA
*Correspondence Author: bs.thirunavukarasu@gmail.com

Keywords Bigdata, mapreduce, Blocks, racks, hdfs, datanode, namenode.

ABSTRACT

In any engineering field the data associated with knowledge is important one for taking decisions for solving problems in the current system development. In current scenario Organization are supposed to work with huge amount of data. Based on those data analysis, predictions, manipulations are made. Replicas of blocks are created to improve the Redundancy in distributed system. A deep understanding is needed inorder to remove some drawbacks in these fields.

I. INTRODUCTION

Big data is an unstructured large set of data even more than peta byte data. Unstructured data is a data which is in the form of logs. There won't be any items like row, column, etc., Big Data can be of only digital one. Data Analysis become more complicated because of their increased amount of data set. Certain tools are used in order to gain business analysis on their data. Predictions, analysis, requirements etc., are the main things that should be done using the unstructured big data. Big data is a combination of three v's those are namely Volume, Velocity and Variety. Big data will basically processed by the powerful computer. But due to some scalable properties of the computer, the processing gets limited.

Organizations of all kind can gather, store, and efficiently process large quantities of data. The ultimate goal for gathering such data is to gain necessary information from the data to improve business processes of the respective organization using statistical and data mining tools. The tremendous needs of data mining have been sequentially updated in the market-basket analysis, fraud detection, consumer profiling, medicine, agriculture, and many other domains. Well before data mining became popular, commercial organizations and government agencies were using statistical methods to analyse data to benefit consumers and society.

Today, data mining draws from and adds to the many statistical analysis techniques. One of the key objectives of data mining is the discovery of new and useful relationships and patterns in the data. Some of these discoveries occur when data is mined specifically for the purposes of discovering such relationships. These unplanned discoveries are facilitated when users are provided access to the stored data. Unfortunately, privacy and confidentiality issues are increasingly creating strong barriers that prevent us from realizing the full benefits of data. In many instances the data that was collected explicitly for analytical purposes sits in a secure facility where only a few authorized individuals are provided access to the data.

Obviously this limits the usefulness of the data and defeats the very purpose for which they were gathered. Numerical data are of particular importance in this regard. They pose the greatest threat yet offer the greater benefits. They pose the greatest threat since they tend to be almost unique and an intruder with numerical data can easily compromise the privacy and confidentiality of sensitive records. They offer the greatest benefit since much of the business intelligence comes from numerical data. Hence, it is important

II. BIG DATA

Big data is an unstructured large set of data even more than peta byte data. Unstructured data is a data which is in the form of logs. There won't be any items like row, column, etc., Big Data can be of only digital one. Data Analysis become more complicated because of their increased amount of data set. Certain tools are used in order to gain business analysis on their data. Predictions, analysis, requirements etc., are the main things that should be done using the unstructured big data. Big data is a combination of three v's those are namely Volume, Velocity and Variety. Big data will basically processed by the powerful computer. But due to some scalable properties of the computer, the processing gets limited.

2.1. ANALYSIS ON BIGDATA

Big data analytics is the application of advanced analytic techniques to very large, diverse data sets that often include varied data types and streaming data.

Big data analytics explores the granular details of business operations and customer interactions that seldom find their way into a data warehouse or standard report, including unstructured data coming from sensors, devices, third parties, Web applications, and social media - much of it sourced in real time on a large scale. Using advanced analytics techniques such as predictive analytics, data mining, statistics, and natural language processing, businesses can study big data to understand the current state of the business and track evolving aspects such as customer behaviour. New methods of working with big data, such as Hadoop and MapReduce, also offer alternatives to traditional data warehousing.

Analytics, providing deep insights on Big Data to optimize every customer touch point. Using personalized workspaces and self-service templates, analytics are rapidly assembled, customized and shared across business teams.

2.2. HADOOP

Hadoop is a framework of tools. The main objective of Hadoop is to support running of applications on big data. Hadoop is an open source tool distributed under Apache licence. It should have the efficiency to work with big volume of data (Volume), data coming in high speed (Velocity) and data of all sort of variety. Hadoop breaks the large set of data into pieces. It breaks the computation as well into smaller pieces. Once all these computations are finished their results are combined together. Hadoop have two main components, they are

- MapReduce
- File system

Where file system is called HDFS. Many tools or projects are grouped and gathered under Hadoop and the objective of this projects is to perform the task. Hadoop works on the distributed model. And basically Hadoop is an LINUX based set of tools and hence we have LINUX on these low cost numerous computers. And all these computers will be called as slaves.

III. HDFS

HDFS is a file system designed for storing very large files with streaming data access patterns, running clusters on commodity hardware. Where HDFS is not a goal fit may be,

- Low latency data access
- Lots of small files
- Multiple writers arbitrary file modifications

We cannot store the data fetched using Hadoop, Hadoop is used only for reading out huge data set in single shot.

3.1. HDFS COMPONENTS:

- Namenode
- Datanode

Namenode

Namenode associated with the job tracker, Job tracking operations are done on the masternode or Namenode. Namenode is a reliable machine which is used to maintain and manage the blocks which present on the Datanodes. It does not store any data. The Namenode is an expensive one and have double and triple redundant machines.

Datanodes

Datanodes are the slaves of the Hadoop distributed file system. They are deployed and each machine provide octal storage. Responsible for serving read and write requests from the client. Datanode associated with the task tracker, Task tracking operations are done on the data node or commodity hardware. Task Tracker sends a heartbeat to task tracker to indicate that they are still alive.

In the existing HDFS Architecture, we mainly focusing on the blocks and Racks. The racks are the place where a group of DataNodes are arranged. These racks are of multiple in order to make use of replication. Every time the operation is performed, DataNode gets the redundant data due to replication. So certain storage is maintained to keep track of already fetched data.

3.2. RACK AWARENESS:

Namenode determines the DataNode to write the 1st block to. If the client is running on a dataNode, it will try to place it there. Otherwise it chooses random order to place it within the same rack. By default, data is replicated to two other places in the cluster. A pipeline is built between the three Datanodes that makeup the pipeline. The second DataNode is randomly chosen node on the rack other than that of the first replica of the block. This is to increase redundancy. The third replica is placed in the random node within the same rack as the second replica placed. The data is piped from the second DataNode to the third one.

To ensure that the write was successful before continuing, acknowledgement packets are sent back from the third DataNode to the second one, from the second dataNode to the first. And from the first dataNode to the client. Thus acknowledgement reaches the client indicating that the blocks are received successfully. This process occurs for the each blocks that make up the file, in this case, the second and third block. It should be noted that, for every block, there is a replica on at least two racks. Now when the client is done writing the dataNode pipeline and has received acknowledgements, it tells the NameNode that it is complete. The NameNode will check that the blocks are at least minimally replicated before responding.

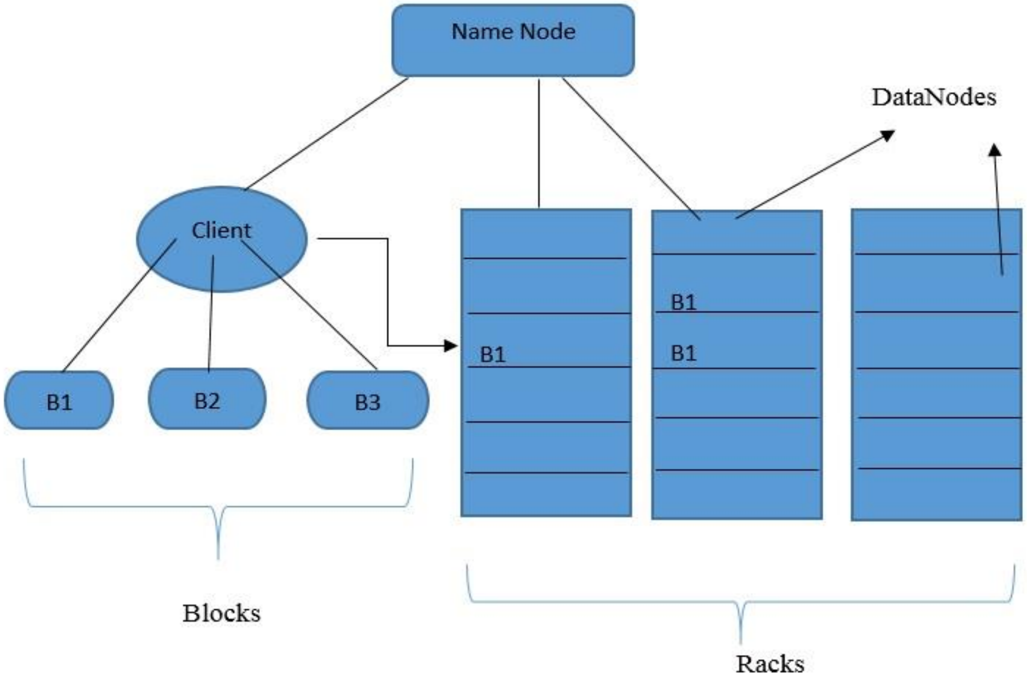

Fig. 1 Existing Replica allocation on Racks

In this methodology if the first DataNode gets dead due to some reasons then, only the replica of the second rack can be used. This increases fetching time as well. The client needs to move to other racks located elsewhere to fetch the DataNode. There in that rack one could find two replicas. The client can use any replica. Once this was done, the performed operation and data should be sent to client from the second rack. This needs higher amount of bandwidth for effective performance and the data transfer. This can be considered as one of the drawback in HDFS architecture.

IV. MAPREDUCE ALGORITHM

4.1. MAPREDUCE:

MapReduce is the heart of Hadoop. It is basically a programming model that can be written in language of choice. Mostly it is practiced to use java, python, etc. Map in the case process the data locally on the DataNode. Key value pair is generated by the Map that helps to process the data.

This key value pair reduces the work and generates the output data. Basically the file is divided into input splits. Each input split must be given to unique DataNodes. For this each input splits one needs Map. And hence number of maps is equal to number of splits. The map performs the operations required to be done with the data whereas Reduce will reduces the number of map program created. The reducing will increase the efficiency. The input splits is equal to number of Maps but the reduced size was not the same. The reduction is based on certain algorithms that make the wright reduction. There are plenty of MapReduce Algorithms, among which Google MapReduce Algorithm is most widely followed. Minimal MapReduce, sorting, searching, BFS, TF-IDF are some of the other used MapReduce Algorithms.

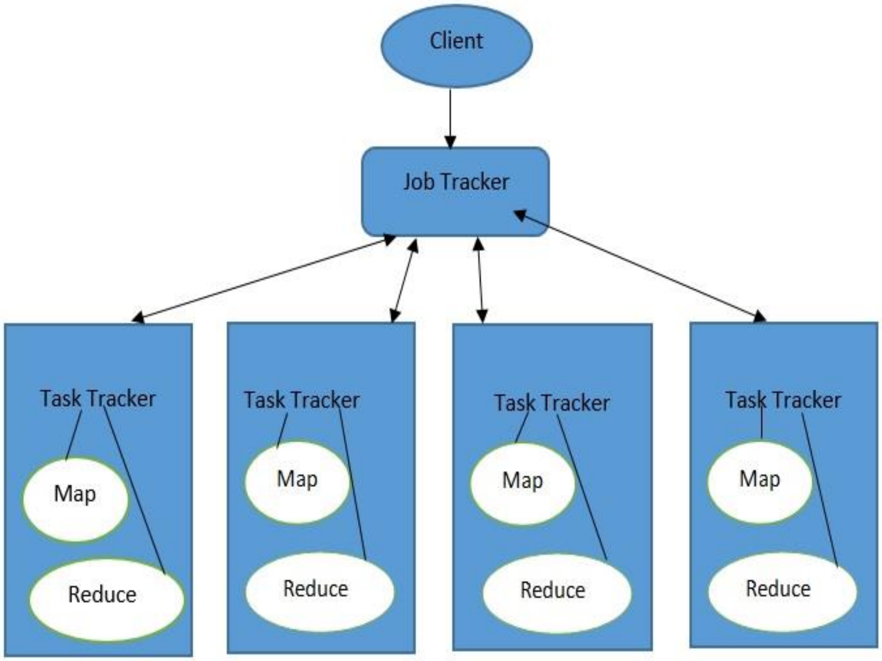

Fig. 2 Map Reduce inside the Task Tracker

4.2. ALGORITHMS FOR MAPREDUCE

- Sorting
- Searching
- TF-IDF
- BFS
- PageRank
- More advanced algorithms

4.3. MAPREDUCE ALGORITHM BY GOOGLE:

MapReduce Jobs Tend to be very short, code-wise Identity Reducer is very common "Utility" jobs can be composed Represent a data flow, more so than a procedure

Sort Algorithm Takes advantage of reducer properties: (key, value) pairs are processed in order by key; reducers are themselves ordered. Mapper: Identity function for value

(k, v) _ (v, _)

_ Reducer:

Identity function (k', _) -> (k', "")

Sort: The Trick

_ (key, value) pairs from mappers are sent to a particular reducer based on hash(key)

_ Must pick the hash function for your data such that k1 < K2 => hash (k1) < hash (k2)

V. CONCULSION AND FUTUREWORK

In this survey the basic concepts in big data was well studied and the basic operations on how a system will interact with other aspects was been surveyed. The Operation of map reduce algorithm within Hadoop was well aware by the effective analysis of the content. In Future any algorithms could be proposed by that a Well-equipped Block racking could be done and some technics can be implemented to overcome the numerical confidential on Big data

VI. REFERENCES

[1] Sarathy R., K. Muralidhar, R. Parsa. 2002. Perturbing non-normal confidential variables: The copula approach. Management Science 48 1613-1627.

[2] K. Muralidhar and R. Sarathy, "A Theoretical Basis for Perturbation Methods," Statistics and

Computing, vol. 13, pp. 329-335, 2003

[3] K. Muralidhar and R. Sarathy, "Data shuffling - A new masking approach for numerical data,"

Management Science, vol. 52, pp. 658-670, 2006.

[4] L. T. Willenborg and T. D. Waal, Elements of statistical disclosure control. New York: Springer, 2001.

[5] R. Nelsen, "An introduction to Copulas," New York: Springer, 2007

[6] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," Management Science, vol. 45, pp. 1399-1415, 1999.

[7] K. Muralidhar, R. Sarathy, and R. Parsa., "An improved security requirement for data perturbation with implications for e-commerce," Decision Sciences, vol. 32, pp. 683-698, 2001.

[8]. Hadoop, Applications powered by Hadoop: http://wiki.apache.org/hadoop/ PoweredB

[9]. Presentation by Randal E. Bryant, Presented in conjunction with the 2007 Federated Computing Research Conference, http://www.cs.cmu.edu/~bryant/ presentations/DISC-FCRC07.ppt.

[10]. L. Barroso, J. Dean, and U. Holzle, Web search for a planet: The Google cluster architecture, IEEE Micro, 23(2), 2003, pp. 22_28.

[11].MapReduce in Wikipedia, http://en.wikipedia.org/wiki/MapReduce (accessed September2009).

[12].Hadoop in Wikipedia, http://en.wikipedia.org/wiki/Hadoop (accessed September 2009).

BIOGRAPHY

Mr. B.Thirunavukarasu is presently pursuing B.E Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Chennai, Tamilnadu, India. His research interests includes BigData , Data Mining and Business Analytics. He has published 2 papers in National conferences and 4 papers in International journals. He is an active entrepreneur involving web services and mobile application development.

Ms. S.Sowbaranika is presently pursuing B.E Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Chennai, Tamilnadu, India. Her research interests includes Data Mining and Network Security. She has published 2 papers in National Conferences.

Ms. M Keerthiga is presently pursuing B.E Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Chennai, Tamilnadu, India. Her research interests includes Big Data, Data Mining and Software Engineering. She has published 3 papers in National Conferences and 1 in journal.

Dr.T.Kalaikumaran is presently Professor & HoD in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University, Chennai Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University, Chennai.. He is interested in the research areas of data mining, spatial data mining, machine learning, uncertain data classification and clustering, pattern recognition, database management system and informational retrieval system. He is a member of CSI and IEEE.

Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Anna University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.