Concept of MapReduce Scheduling and Processing Techniques

Mr. A. Antony Prakash[1*], Dr. A. Aloysius [2]
[1] Department of Information Technology, St.Joseph's college, Trichy,India
[2] Department of Computer Science, St.Joseph's college,Trichy, India

Abstract: In the Big Data society, MapReduce has been viewed as one of the key empowering approaches for taking care of persistently expanding requests on processing assets forced by enormous informational indexes. The purpose behind this is the high adaptability of the MapReduce worldview which considers greatly parallel and circulated execution over a substantial number of processing hubs. This paper recognizes MapReduce issues and difficulties in taking care of Big Data with the target of giving a diagram of the field, encouraging better arranging and administration of Big Data extends, and distinguishing open doors for future research in this field. The recognized difficulties are assembled into four primary classes comparing to Big Data undertakings composes: information stockpiling (social databases and NoSQL stores), Big Data examination (machine learning and intelligent investigation), and web based preparing, and security and protection. In addition, current endeavors went for enhancing and stretching out MapReduce to address recognized difficulties are exhibited. Therefore, by recognizing issues and difficulties MapReduce faces when taking care of Big Data, this examination energizes future Big Data research.

***IndexTerms* – Big Data, Mapping, Reducing, HDFS, Hadoop**

1. INTRODUCTION

For a many years, User who needs to store and examine information would store the information in a database what's more, process it by means of SQL inquiries. The Web has changed the vast majority of the presumptions of this period. On the Web, the information is unstructured and huge, and the databases can neither catch the information into a construction nor scale it to store and process it.They manufactured a structure for vast scale information handling getting from the "outline" "diminish" elements of the utilitarian programming worldview. They called the worldview MapReduce.

Existing and customary devices and application ends up lacking to process vast measure of information. Hadoop clarified and addressed the issue of taking care of and handling such terabytes and petabytes of information.Undertakings like Google, Facebook other Internet colossal associations process client requests, inquiries and remaining tasks at workload.

The flexibility of Linux converged with smooth and predictable versatility of cloud condition makes it fit for providing the perfect structure for investigating and preparing Big Data, while disposing of the requirement for expensive equipment and programming. Hadoop is viewed as a favored decision in open source cloud computing network for giving a efficient stage to Big Data processing.

Big data is an developing term that describes any huge amount of data like structured, semi-structured and unstructured data that has the potential to be mined for information.
Big data analysis is the process of applying superior analytics and idea techniques to large data sets to discover hidden patterns and unknown correlations for effective decision making.

A distributed file system is a client/server-based application that allows customers to access and process data kept on the server as if it were on their personal computer.When a consumer entrees a file on the server, the server sends the user a copy of the file, which is cached on the user's computer while the data is being treated and is then returned to the server.

Volume:
Volume is refers to the amount of data. For example how amount of data generated in social media, email, face book we have post images or video and like and comment in that image or video and creating the amount of data this also one example of Volume.

Velocity:
Velocity is refers to the speed at which large amounts of data are being generated, collects the data and analyzed. Example how fast data (images, videos, text, etc) transfer in social media.

Variety:

Variety is refers to the different type of data that means both structured and unstructured data. First we are used to transfer any data in table or excel sheet format use. But now a days we used to transfer data in images, videos, email, etc.....

Hadoop is an open source software framework for storing data and running application on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtual limitless concurrent task or jobs.

Map:

Takes a set of data and converts it into another set of data , where individual elements are broken into tuples (keys/values).

Reduce:

Task, which takes output from the map as an input and combine those data tuples into a similar set of tuples. The reduce job is constantly performed subsequent to the map job.

II. Literature Review

MapReduce and the Google File System. It additionally manufactured a DBMS framework known as BigTable. It winds up conceivable to look a huge number of pages and restore the outcomes in milliseconds or less with the assistance of the calculations that drive both of these significant class seek administrations started with Google's MapReduce structure [1]. Technologies for analysing huge information are developing quickly and there is noteworthy enthusiasm for new diagnostic methodologies, for example, MapReduce, Hadoop and Hive, and MapReduce expansions to existing social DBMSs [2].The utilization of MapReduce system has been generally came into center to deal with such gigantic information effectively.For the most recent couple of years, MapReduce has showed up as the most prevalent figuring worldview for parallel, group style and examination of huge measure of information [3].MapReduce picked up its ubiquity when utilized effectively by Google. In genuine, it is a scalableand blame tolerant information handling device which gives the capacity to process tremendous voluminous information in parallel with some low-end figuring hubs [4].

By goodness of its effortlessness, adaptability, and adaptation to non-critical failure, MapReduce is getting to be universal, increasing huge energy from both industry and scholarly world. We can accomplish superior by breaking the preparing into little units of work that can be kept running in parallel over a few hubs in the cluster.

In the MapReduce structure, a disseminated document framework (DFS) at first parcels information in different machines and information is spoken to as (key, esteem) sets. The MapReduce system executes the principle work on a solitary ace machine where we may preprocess the information before guide capacities are called or postprocess the yield of diminish capacities. A couple of guide and decrease capacities might be executed once or various occasions as it relies upon the attributes of an application [6]. Hadoop is a mainstream open-source execution of MapReduce for the examination of huge datasets. It utilizes a conveyed client level filesystem to oversee capacity assets over the bunch [7].

Hadoop clarified and addressed the issue of taking care of and preparing such terabytes and petabytes of data [8]. Enterprise like Google, Facebook other Internet giant associations process client requests, questions and remaining burdens. Billions of inquiries and solicitations are served each hour on Internet. Along these lines, relocating towards distributed computing to exploit diminished expenses and enhanced execution is getting to be need of a hour [9]. master hub can likewise assume a job of slave. Therefore, notwithstanding expert daemons, ace master can run the slave daemons too. In a general sense, the daemons running on ace master assume liability for planning and dealing with the slave daemons on all hubs which carryout work for information stockpiling and processing [10][11]. A MapReduce job for the most part breaks and partitions the input information into lumps which are first prepared by "map stage" in parallel and after that by "map stage"[12][13][14]. Hadoop structure deals with the yield of the Map stage which are then given as a contribution to Reduce stage to start parallel lessen errands These information and yield records are put away in document framework. As a matter of course, the MapReduce structure gets input datasets from HDFS[15][16]. Hadoop is a renowned open-source use of MapReduce for the examination of broad datasets. It uses a passed on customer level filesystem to manage limit resources over the gathering [17]. The as of late

presented MapReduce strategy has picked up a great deal of consideration from mainstream researchers for its relevance in substantial parallel information investigations [18]. Hadoop, for handling substantial information volume employments utilizes MapReduce programming model. Hadoop makes utilization of various schedulers for executing the occupations in parallel. The default scheduler is FIFO (First In First Out) Scheduler. Other schedulers with need, pre-emption and non-pre-emption choices have likewise been produced. As the time has passed the MapReduce has achieved few of its restrictions. So with the end goal to defeat the confinements of MapReduce, the people to come of MapReduce has been produced called as YARN (Yet Another Resource Negotiator). Thus, this paper gives a review on Hadoop, few planning strategies it utilizes and a brief prologue to YARN[19]. MapReduce planning calculations with and without our procedure yet additionally with a current information territory improvement method (i.e., the defer calculation created by Facebook). Trial results demonstrate that our system regularly prompts the most astounding information region rate and the least reaction time for guide undertakings. Besides, in contrast to the postpone calculation, it doesn't require a multifaceted parameter tuning process [20].

III. Workflow of Map Reduce Algorithm

While handling vast arrangement of information, we should address versatility and effectiveness in the application code that is preparing the extensive measure of information. Map reduce algorithm is extremely effective in management big data.Let us solve the simple example use mapping and reducing to solve a problem.

Let us assume to processing a vast amount of data and trying to find out what level of your user base where discussing about games.

To start with, we will distinguish which we are going to map from the data to presume that its something identified to games.

Next, we will compose a mapping utility to recognize such patterns in our data. For instant, the keywords can be Gold medals, Bronze medals, Silver medals, Olympic cricket, basketball, cricket, and so on.

Let us take the following piece in a big data set and see how to progression it.

“Hi, how are you”
“We love cricket”
“He is an awesome cricket player”
“Happy New Year”
“Olympics will be held in USA”
“Records broken today in Olympics”
“Yes, we won 2 Gold medals”
“He qualified for Olympics”

Mapping Phase

So the mapping phase of our algorithm will be as follows:

1. Declare a “Map” function
2. Loop: For each words equal to “cricket”
3. Increment counter
4. Return key value “cricket”=>counter

Similarly, we can describe n number of mapping functions for mapping different words: “Olympics”, “Gold Medals”, “cricket”, etc.

Reducing Phase

The reducing function will realize the key from all these mappers in form of key value pair and then prepare it. So, input to the reduce utility will look like the following:

- reduce("cricket"=>2)
- reduce("Olympics"=>3)

Our algorithm will continue with the following steps:

5. Declare a function reduce to acknowledge the values from map function.
6. Where for each key-value pair, insert value to counter.
7. Return "games"=> counter.

At the end, we will get the output like "games"=>5.

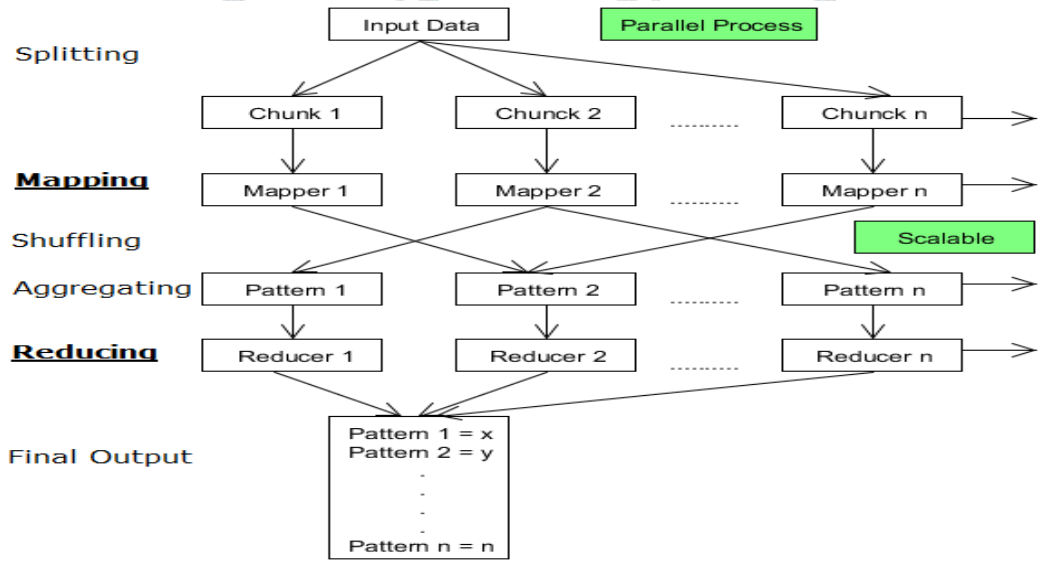
Presently, getting into a major picture we can compose n number of mapper works here. Let us chance to state that you needto know who all where wishing one another. For this situation you will write a mapping function to map the words like "Wishing", "Wish", "Happy", "Merry" and then will write a corresponding reducer function.

Here you will require one function for rearranging which will recognize between the "games" and "wishing" keys returned by mappers and will send it to the rticularp reducer function.

Thus you may require a function for splitting initially to give inputs to the mapper functions in form of chunks.

Flow of Map Reduce Algorithm

To following diagram summarizes the the flow of Map reduce algorithm:



In the above map reduce flow:

1. The input information can be partitioned into n number of pieces depending upon the amount of data and handling limit of individual unit.
2. Next, it is passed to the mapper functions. Please note that all the chunks are processed simultaneously at the same time, which embraces the parallel processing of data.
3. After that, shuffle happens which prompts to aggregation of similar patterns.
4. Finally, reducers combine them all to get a consolidated output as per the logic.
5. This algorithm embraces scalability as depending on the size of the input data, we can keep increasing the number of the parallel processing units.

IV. CONCEPTS OF MAPPING AND REDUCING

MapReduce Algorithm is mostly motivated by Functional Programming model.

MapReduce algorithm is mostly useful to process vast amount of data in parallel, reliable and efficient way in cluster environment.

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', "")

map (k, v) → <k', v'>*

reduce (k', v') → <k', v'>*

All values with the same key are reduced together

partition (k', number of partitions) → partition for k'

- Often a simple hash of the key, e.g., hash(k') mod n
- Divides up key space for parallel reduce operations

combine (k', v') → <k', v'>*

- Mini-reducers that run in memory after the map phase
- Used as an optimization to reduce network traffic

It uses Divide and Conquer method to process vast amount of data.

It separates input task into smaller and reasonable sub-tasks (They must be executable independently) to execute them in-parallel.

MapReduce Algorithm Steps

MapReduce Algorithm uses the following three main steps:

1. Map Function
2. Shuffle Function
3. Reduce Function

Here we will examine to discuss each function role and responsibility in MapReduce algorithm.

Map Function

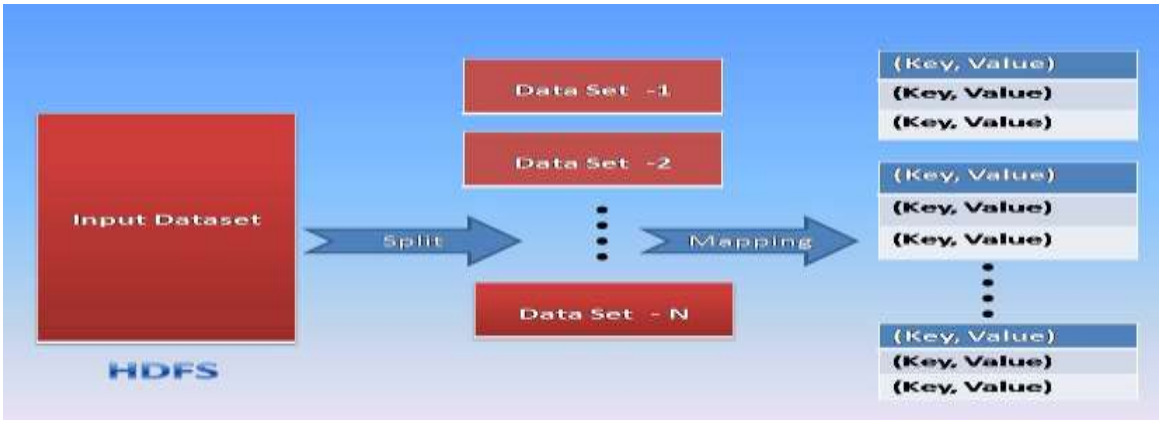
Map Function is the initial phase in MapReduce Algorithm. It takes input assignment (say DataSets. I have given only single DataSet in below diagram.) and partitions them into smaller sub-tasks. At the point perform required calcualtion on each sub-task in parallel.

This step performs the following two sub-steps:

1. Splitting
2. Mapping

- Splitting step takes input DataSet from Source and divide into smaller Sub-DataSets.
- Mapping step takes those smaller Sub-DataSets and perform required action or computation on each Sub-DataSet.

The output of this Map Function is a set of key and value pairs as <Key, Value> as shown in the below diagram.



MapReduce First Step Output:

MAP Function Output = List of <Key, Value> Pairs

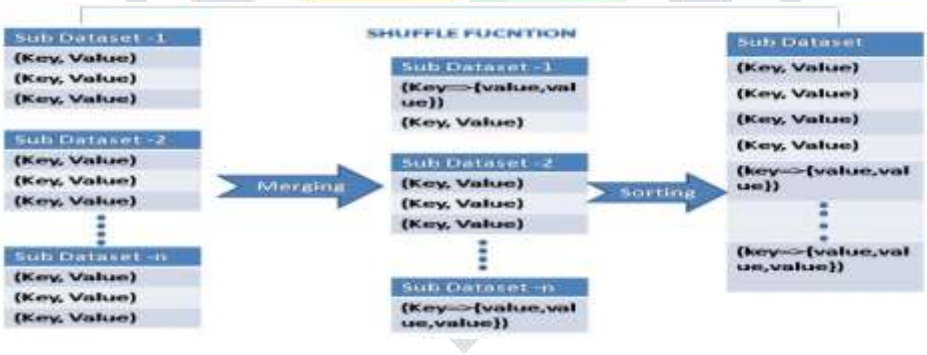
Shuffle Function

It is the second step in MapReduce Algorithm. Shuffle Function is likewise know as "Combine Function". It performs the following two sub-steps: Merging and Sorting

It takes a list of outputs coming from "Map Function" and perform these two sub-steps on each and every key-value pair.

- Merging step consolidates all key-value sets which have same keys (that is confederacy key-value pairs by looking at "Key"). This performance returns <Key, List<Value>>.
- Sorting step makes contribution from Merging step and sort all key-value pairs by utilizing Keys. This progression also returns <Key, List<Value>> yield yet with sorted key-value sets.

Finally, Shuffle Function proceeds a list of <Key, List<Value>> sorted sets to subsequent stage.

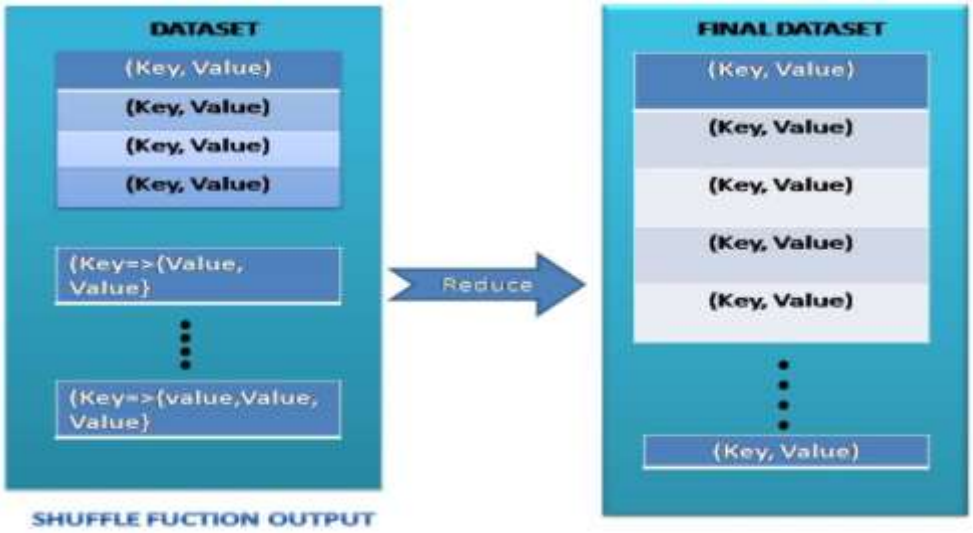


MapReduce Second Step Output:

Shuffle Function Output = List of <Key, List <Value>> Pairs

Reduce Function

It is the final step in MapReduce Algorithm. It perform just a single steps : Reduce step. It takes directory of <Key, List<Value>> sorted sets from Shuffle Function and perform reduce operation as shown below.



MapReduce Final Step Output:

Reduce Function Output = List of <Key, Value> Pairs

Final step output looks like initial step output. However ultimate step <Key, Value> sets are diverse than initial step <Key, Value> pairs. Final step <Key, Value> pairs are figured and sorted sets.

V. Conclusion

MapReduce is a preparing strategy and a program model for circulated computing dependent on java. The MapReduce algorithm contains two vital tasks, specifically Map and Reduce. map takes an arrangement of data and changes over it into another set of information, where singular elements are separated into tuples (key/esteem sets). Hadoop and MapReduce programming paradigm and HDFS are increasingly being used for processing large and unstructured data sets. Hdfs is used for distributing the file content and mapReduce is used to sort and shuffle the data. The new arrival of Hadoop known up 'yet Another Resource Negotiator (YARN)— the beyond– Map-Reduce (MR) thinking has been solidified.

REFERENCES

[1] J R Swedlow, G Zanetti, C Best. Channeling the data deluge. Nature Methods, 2011, 8: 463-465.
[2] G C Fox, S H Bae, et al. “Parallel Data Mining from Multicore to Cloudy Grids. High Performance Computing and Grids workshop”, 2008.
[3] Maitrey S, Jha. “An Integrated Approach for CURE Clustering using Map-Reduce Technique”. In Proceedings of Elsevier, ISBN 978-81-910691-6-3,2nd August 2013.
[4] D. DeWitt and M. Stonebraker.” MapReduce: A major step backwards”. The Database Column, 1, 2008.
[5] Apache. Apache Hadoop. http://hadoop.apache.org, 2010.
[6] Y. Kim and K. Shim. Parallel top-k similarity join algorithms using MapReduce. In ICDE, 2012.
[7] Jeffrey Shafer, Scott Rixner, and Alan L. Cox.” The Hadoop Distributed Filesystem: Balancing Portability and Performance”. DOP is March 30,2010
[8] Kenneth Wottrich, and T. Bressoud, “The performance characteristics of mapreduce applications on scalable clusters”, in *Proceedings of theMidstates Conference on Undergraduate Research in Computer Science and Mathematics (MCURCSM),* Denison University, Granville,USA, Nov. 2011.
[9] S. Loughran, J.M.A. Calero, A. Farrell, J. Kirschnick, and J.Guijarro, "Dynamic deployment of mapreduce architecture in the cloud."*IEEEInternet Computing*, vol. 16, no. 6, pp. 40-50, Dec. 2012.
[10] T. White, “MapReduce and the hadoop distributed file system”, *in Hadoop: The definitive guide*, 1st edition, O'Reilly Media, Inc., Yahoo press, 2012.
[11] A.Elsayed, O. Ismail, and M.E. El-Sharkawi, "MapReduce: state-of-the-art and research directions", *International Journal of Computer andElectrical Engineering*, vol. 6, no. 1, pp. 34-39, Feb. 2014
[12] J. Ekanayake, S. Pallickara, and G. Fox, “Mapreduce for data intensive scientific analyses”, in *IEEE 4thInternational*

Conference on eScience, Indianapolis, Indiana, USA, Dec. 7-12, 2008, pp. 277-284.

[13] Gray, and T.C. Bressoud, "Towards a mapreduce application performance model", in *Proceedings of the Midstates Conference onUndergraduate Research in Computer Science and Mathematics (*MCURCSM), Ohio Wesleyan University, USA, Nov.17, 2012.

[14] A. Verma, N. Zea, B. Cho, I. Gupta and R. H. Campbell, "Breaking the mapreduce stage barrier", in *Proceeding of IEEE InternationalConference on Cluster Computing (CLUSTER),*, Chicago, USA, Sep. 23-27, 2013, vol. 16, no. 1, pp. 191-206.

[15] C. Tian, H. Zhou, Y. He, and L. Zha, "A dynamic mapreduce scheduler for heterogeneous workloads." in *8th International Conference onGrid and Cooperative Computing (GCC)*, Lanzhou, China, Aug. 27-29, 2009, pp. 218-224.

[16] C. Lam, "Writing basic mapreduce programs", in *Hadoop in Action*, 1st ed., MANNING, 2011.

[17] Jeffrey Shafer, Scott Rixner, and Alan L. Cox. The Hadoop Distributed Filesystem: Balancing Portability and Performance. DOP is March 30,2010.

[18] JaliyaEkanayake, ShrideepPallickara, and Geoffrey Fox, MapReduce for Data Intensive Scientific Analyses. In Fourth IEEE International Conference on eScience (978-0-7695-3535-7/08) eScience, 2008.

[19] Amogh Pramod Kulkarni1, Mahesh Khandewal, "Survey on Hadoop and Introduction to YARN", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014.

[20] Chen He Ying Lu David Swanson," Matchmaking: A New MapReduce Scheduling Technique", Third IEEE International Conference on Coud Computing Technology and Science,2011.