International Journal of Engineering & Technology, 7 (4.6) (2018) 13-16

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Review: Map Reduce Framework for Cloud Computing

Mekala Sandhya[1], Ashish Ladda [2], Dr.Uma N Dulhare[3]
[1,2]balaji Institute Of Technology & Science, Warangal, Telangana, India
[3]muffakham Jah College Of Engg. &Tech, Hyderabad, Telangana, India

**Corresponding author E-mail: sandhyamekala1@gmail.com*

Abstract:

In this generation of Internet, information and data are growing continuously. Even though various Internet services and applications. The amount of information is increasing rapidly. Hundred billions even trillions of web indexes exist. Such large data brings people a mass of information and more difficulty discovering useful knowledge in these huge amounts of data at the same time. Cloud computing can provide infrastructure for large data. Cloud computing has two significant characteristics of distributed computing i.e. scalability, high availability. The scalability can seamlessly extend to large-scale clusters. Availability says that cloud computing can bear node errors. Node failures will not affect the program to run correctly. Cloud computing with data mining does significant data processing through high-performance machine. Mass data storage and distributed computing provide a new method for mass data mining and become an effective solution to the distributed storage and efficient computing in data mining.

Keywords: *Data Mining, Cloud, Map Reduce Framework, HDFS (Hadoop Distributed File System), Parallel Programming, Distributed Databases*

1 Introduction:

Data Mining is the approach of accessing the exact data i.e. required data from large amount of database. Where the user can get

this information with in very short time. So many Data mining software's came into the market which can be performed on complex calculations and can be analyzed on set of data in very short time. Data mining aims at knowledge analysis, discovering frequent patterns, and sequential patterns, unknown & hidden patterns from multiple data streams. Data mining utilizes tools, procedures, algorithms and methodologies to taking out from large data. Data mining tools are used for predictive modeling, presenting information in required format such as graph or table and efficient handling of complex and relational data. Data mining allows finding information hidden in the data that is not always apparent, given that, given the gigantic volume of existing data; a large part of that volume will never be analyzed.

Cloud computing is an area or place where you can store large amount of data. In today's generation cloud computing is most merging technology where user can access the data from anywhere, any place, at any time. It also provides a most important feature to the user i.e. "As You Pay as You Get" i.e. how much the user is using the storage that much only they need to pay for it. Cloud computing deals with the resources of infrastructure for massive and complex data, software distribution for users to subscribe the software and platform for users are able to use prebuilt environment to run a new application. The main objective of cloud computing is to access resources & services needed to perform tasks efficiently. Essentially, cloud computing is a multi-user, multi-tasking, concurrently supported system. Efficient, simple and fast is its core philosophy.

Map Reduce model delegates the data-intensive computations to a cluster of remote servers, through a distributed file system, will distribute the workload, optimizing time and resources. It facilitates a parallel development pattern to simplify the implementation of applications in distributed environments. The original intention of the distributed parallel programming model was to make more efficient use of hardware and software resources to enable users to use applications or services faster and easier. In distributed parallel programming mode, complex background tasks and resource scheduling are transparent to the user.

Hadoop Distributed File System (HDFS) In the current field of cloud computing, the open source system HDFS developed by Google's GFS and Hadoop are the two popular cloud computing distributed storage systems. Most ICT vendors, including Yahoo, Intel's "cloud" plan are used HDFS data storage technology. Future developments will focus on very large data storage, data encryption and security guarantees, and continued improvements in I / O rates.

GFS (Google File System) Technology: GFS meet the needs of a large number of users, in parallel to provide services for a large number of users. Making cloud computing data storage technology with high throughput and high transmission rate characteristics.

Parallel Computing is a mechanism where two or more process can be executed concurrently on different processors at the same time. In order to handle this overall control/coordination mechanism is employed. The parallel computing will increase the performance of

the system i.e. within less time so many processes can be executed simultaneously.

Virtualization is one of the enabling technology in Cloud computing. Virtualization is basic building block is of cloud. Virtualization provides hardware i.e. processor, memory, secure remote access, multiple storage locations and energy saving technique. Virtualization is the creation of a virtual (rather than real) version of something, such as an operating system, a server, a storage device, or network resources. Perhaps something is known about virtualization if the hard disk has ever been divided into different partitions. A partition is the logical division of a hard drive to create, in effect, two separate hard disks. The operation of system virtualization is to use software to allow a piece of hardware to run multiple images of the operating system at the same time. This technology originated decades ago in mainframes, and allowed administrators to avoid wasting processing power, which was expensive. In 2005, virtualization software was adopted faster than anyone imagined, including experts.

Distributed Databases: The distributed storage is not exactly the same as the traditional network storage which uses a centralized storage server to store all the data. The storage server becomes a bottleneck problem. It adopts a scalable system architecture, which utilizes multiple storage servers and location server which improves system reliability, availability and efficiency, but also expands easily.

2. Literature Survey:

There are so many data mining algorithms are available in the existing system. In this paper we will have glance on the existing system and how it is differ from our approach. They have proposed a recommendation algorithm in reference [7] .The data items of features relevant to the recommendation task. In order to support personalized recommendation, this random walk for each user is initialized respectively. In addition, for the demand of calculation, we must give the size of the graph. In this way, we use multiple-way clusters to gather the nodes with high correlation. In reference [9] proposes a distributed algorithm constructing a decision tree on a heterogeneous distributed database.

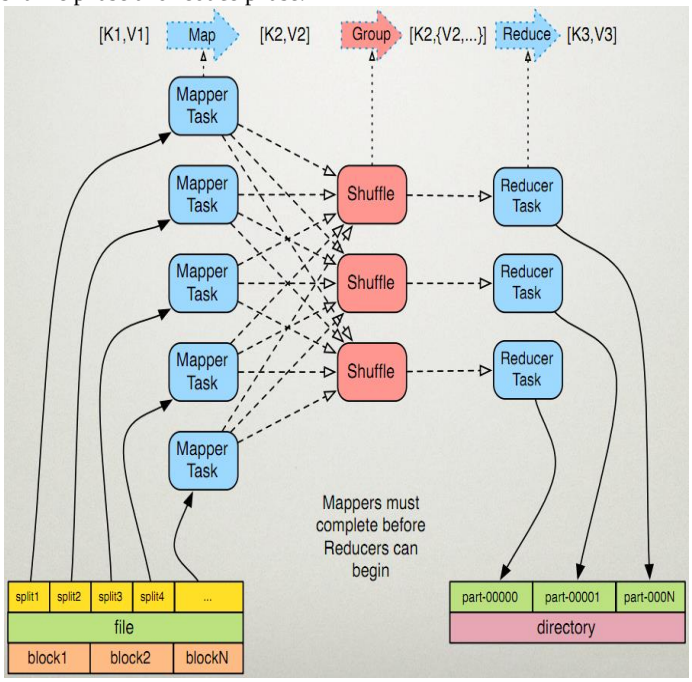
In reference [10] represents a distributed algorithm to learn parameters of a Bayesian network from distributed heterogeneous data sets. In reference [11] proposes two parallel algorithms which both use the concept of exception points called outliers. An outlier is an observation that is so different from other observations that it can be doubted reasonably that may be it is generated by a different mechanism. If there is an outlier, it can be found that some individuals or groups have behaved differently from most of the individuals or groups of the database. The two algorithms are almost linear speedup.

In reference [12] mainly discusses the effects of clusters composed of "weak" clustering member's to clustering ensemble. "Weak" clustering refers to clustering only better than the random division. The paper selects three consensus functions to do cluster fusion of cluster members. In reference [13] proposes a sequential pattern mining method MILE which deals with multiple data streams. Using MILE (Mining in the multiple streams) algorithm makes the mining process much easier.

In reference [14] a parallel association rule mining algorithm based on Apriori algorithm on Hadoop platform. Reference [15] proposes a parallel k-means clustering algorithm on Hadoop platform.

3. Map Reduce Functionalities:

The task process is divided into two phases: the map phase and the reduce phase, in which it uses key / value as input and output. The Programmers need to do is to define the function of these two stages: map function and reduce function. In the below Figure it is clearly explained how it works. The operations of map reduce is performed in a chronological order: input split, map phase, combiner phase, shuffle phase and reduce phase.



3.1 Fig: Basic diagram for Map Reduce Approach

Input Split: Before performing the map calculations, map reduce calculates input splits based on the input file, each input split is for a map task, input split is stored Is not the data itself, but an array of fragment lengths and a location where data is recorded. The input splits are often closely related to the hdfs block. If we set the size of the hdfs block to be 64 MB , If we enter three files of size 3mb, 65mb and 127mb respectively, then map reduce divides the 3mb file into an input split, 65mb for two input split and 127mb for two Input split, in other words if we do an input slicing adjustment before the map calculation, for example, merging small files, then there will be 5 map tasks to be executed and the data size of each map will be uneven , This is also a key point of map reduce optimization calculation.

Map phase: In this the programmer need to write a best approach for map function, so map function efficiency is relatively good control, and the general map operations are localized operation is carried out in the data storage node;

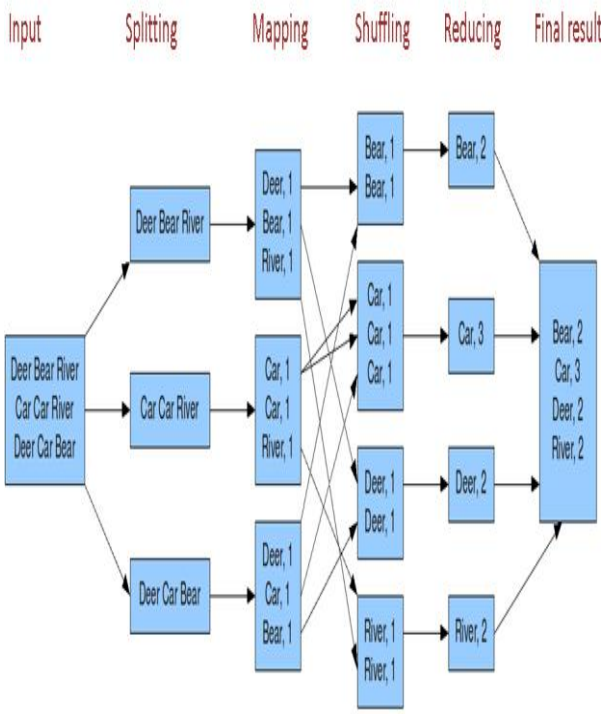
Combiner Stage: The combiner stage is a programmer's choice, and the combiner is actually a reduce operation, so we see that the word Count Class is loaded with reduce. Combiner is a localized reduces operation, which is the follow-up operation of the map operation, mainly to do a simple operation of combining and repeating key values before the intermediate file is calculated by the map. For example, we calculate the word frequency in the file and the map calculation If you hit a Hadoop word will be recorded as 1, but this article Hadoop may appear n many times, then the map output file redundancy will be a lot, so reduce the calculation before the same key to do a merge operation , Then the file will be smaller, thus improving the transmission efficiency of broadband, after all,

Hadoop computing power broadband resources are often the bottleneck of computing is the most valuable resource, but the combined operation is at risk, the principle of using it is combiner input For example: if the calculation is only for the sum, the combiner can be used for the maximum and the minimum, but the average is calculated. With the combiner, the final reduce result will be incorrect.

Shuffle Phase: In this phase the output of the map as input to reduce, which is where map reduce optimization focuses? Here we will be mostly concentrating on the principle of shuffle stage, because most of the books did not make it clear shuffle stage. Shuffle is the beginning of the map phase of the output operation, the general map reduce calculations are massive data, the map output cannot put all the files into memory operations, so map is written to disk process is very complicated, not to mention the map output time Sorting the results, the memory overhead is great, map in the output will be opened in memory, a ring memory buffer, the buffer is dedicated to the output, the default size is 100MB, and in the configuration file for this buffer Zone set a threshold, the default is 0.80 (the size and threshold can be configured in the configuration file), while map will start a daemon thread for the output operation, if the buffer memory reaches the threshold 80% of the time, the daemon thread will write the contents of the disk, the process is called spill, the other 20% of the memory can continue to write to write data to the disk, write to disk and write to memory operation isNot interfere with each other, if the buffer zone is full, then the map will block the operation of writing to memory, write disk operation is completed and then continue to write to memory operation, I mentioned before writing to disk there will be a Sort operation, this is written to the disk operation, not when writing to memory, if we define the combiner function, then sort before the implementation of the combined operation. Each time spill operation is written to the disk operation will write an overflow file, which means that in the map output spill several times will have a number of overflow files, so the map output all done, the map will merge these outputs file. There is a Partitioner operation in this process. Many people are confused for this operation. In fact, the Partitioner operation is very similar to the input split in the map phase. A Partitioner corresponds to a reduce job. If our map reduce operation has only one reduce Operation, then there is only one Partitioner, if we have multiple reduce operations, then the corresponding Partitioner there will be multiple, Partitioner and therefore reduce the input fragment, the programmer can programmatically control, mainly based on the actual key and value .Depending on the type of real business or for better load reduction requirements, this is a key part of improving reduce efficiency. Reduce the stage is the merger map output file, Partitioner will find the corresponding map output file, and then copy operation, the copy operation will reduce when you open a few replicate threads, the default number of these threads is 5, the programmer can also be configured File changes the number of replication threads, the replication process and map write to disk process is similar, there are thresholds and memory size, the threshold can be configured in the configuration file, and the memory size is the task tracker directly reduce the size of the memory, copy Reduce when sorting operations will be carried out and the operation of the merger of documents, these operations will be done to reduce the calculation.

Reduce phase: the same as the map function is also prepared by the programmer, the final result is stored in hdfs.

Example for Map Reduce Framework:



3.2 Figure: Map Reduce Frame work with example

4. Challenges and Open Research Issues in Data mining

1) To know various authors work whether the algorithms can be applied in Map Reduce, that is, whether the algorithms can parallel is quite obvious.
2) To find and extract includes association rule classification, clustering algorithms & stream data mining algorithms.
3) To implement the novel Map Reduce framework efficient, scalable and simplified programming model for large scale distributed data processing on a large cluster of commodity computers and also used in cloud computing.
4) To analyze and review the overview of parallel Apriori algorithm implemented on Map Reduce framework
5) To identify the Map and Reduce functions used to implement them like 1-phase vs. k-phase, I/O of Mapper, Combiner and Reducer, using functionality of Combiner inside Mapper

5 Future Scope:

1) Implement the Map Reduce framework for Hadoop Distributed File System (HDFS) considerably reduce the time complexity of the database scan.

2) Applying partitioned Pincer-Search Algorithm that can segregates the data into so-called cluster and addresses each of these as separate problems

3) By Map Reduce the complexity for the database scan can be significantly improved.

6. Conclusion:

Map Reduce is a software framework devised by Google to support distributed parallel processing on huge datasets on parallel computers such as clusters. As this implementation, various proposals including Hadoop have been proposed and widely used. Finally the evaluation of proposed work incorporating the parallel distributed data mining platform PD Miner which is based on Hadoop tool at TB level's mining.

References

[1] K. Chen and WM. Zheng, "Cloud computing: System instances andcurrent research," Journal of Software, vol. 20, no. 5, pp. 1337-1348,2009 (In Chinese).

[2] K. Sharma, G. Shrivastava, and 0V. Kumar, "Web Mining: Today andTomorrow," In Proceedings of the IEEE 3rd International Conference onElectronics Computer Technology, Athens, vol. 1, pp. 399–403, April2011.

[3] highlyscalable.wordpress.com/2012/02/01/MapReduce-patterns

[4] "Pincer-Search Algorithm for Discovering Maximum FrequentSet" – AkashSaxena, NITJ

[5] "Pincer-Search: An Efficient Algorithm for Discovering theMaximum Frequent Set" – Dao-I Lin, Zvi M. Kedem, 1999

[6] "Study of Data Mining algorithm in cloud computing usingMapReduce Framework" – Viki Patel, Prof. V. B. Nikam,V.J.T.I, Mumbai, 2013

[7] H. Cheng, P. Tan, S. Jon , and W. F. Punch, "Recommendation viaQuery Centered Random Walk on K-partite Graph," In Proceedings ofthe IEEE International Conference on Data Mining, Omaha, pp. 457–462, October 2007.

[8] A. Javed and A. Khokhar, "Frequent pattern mining on message passingmultiprocessor systems," Distributed and Parallel Databases, vol. 16, pp.321-334, 2004.

[9] C. Giannella, K. Liu, T. Olsen, and H. Kargupta, "Communication efficient construction of decision trees over heterogeneously distributeddata," In Proceedings of the Fourth IEEE International Conference onData Mining, pp. 67-74, 2004.

[10] R. Chen, S. Krishnamoorthy, "A New Algorithm for LearningParameters of a Bayesian Network from Distributed Data," InProceedings of the 2002 IEEE International Conference on Data Mining,Maebashi City, pp. 585–588, 2002.

[11] E. Lozano, E. Acuna, "Parallel Algorithms for Distance- based andDensity-based Outliers," In Proceedings of The Fifth IEEE InternationalConference on Data Mining, Houston, pp. 27-30, November 2005.

[12] A Topchy, A K Jain, W F Punch, "Combining Multiple WeakClusterings," In Proceedings of the 3rd IEEE International Conferenceon Data Mining, pp. 331-338, 2003.

[13] G. Chen, X. Wu, X. Zhu, "Sequential pattern mining in multiplestreams," In Proceedings of the 30th International Conference on Datamining.Houston, pp. 585-588, 2005.

[14] M. Cheng, "Web data mining Based on cloud computing," ComputerScience, vol. 38, no. 10A, pp. 146-149, 2011 (In Chinese).

[15] WZ. Zhao, HF. Ma, YL, "Fu. Research on Parallel k-means AlgorithmDesign Based on Hadoop Platform," Computer Science, vol. 38, no.10pp. 166-168, 2011 (In Chinese).