Analysis of MapReduce Methods

[1]Ms. Rohini Kharbade, [2]Prof. Manish B. Gudadhe

[1]Student, [2]Assistant Professor
St. Vincent Pallotti College of Engineering and Technology

Abstract: **As the uses of social media networking platforms increasing hence, huge volume of data is generated with that it required large amount of memory for storing the data. Since, there are some issues like storing the large data set, parallel and distributed large data set also including some challenges like critical path problem, reliability problem, equal split issues, single split issues and aggregation of issues. Hence, to overcome this problem we have MapReduce method which allows us to use parallel computations, distributed processing without considering issues like fault tolerance and reliability. For such reasons, this survey paper mentioned the various implementation methods of MapReduce and also, comparison between various implementation methods with respect to the volume and variety.**

***Keywords*: MapReduce. Big Data. Hadoop**

Introduction

The increasing use of internet technologies has resulted into large number of data from many social networking domains like Twitter, Facebook, and LinkedIn. Since all data is not always required, but only the relevant one suited for particular information need, some methods are required to supply user with application specific data. The technique to this huge amount of data and to extract value out of this volume and variety rich data are collectively called Big Data. Considering that there is no definitive solution for their storage, querying, analysis. Hence, in this project brief the storage architecture, data analytics, data protection, data management & security. Basically, social networking domain are unstructured data depends upon the parameter of services. Map Reduce model processes the unstructured dataset available in a clustering format. As the name indicates it mainly has two jobs map and reduce job. It splits the input data into smaller chunks and processes it in parallel and gives expected outcomes [1].

MapReduce is processing technique and program model for distributed computing based on java. The main advantages of MapReduce is to execute processing the data over various computing nodes. In the MapReduce model, the processing of data primitives are called mappers and reducers. Division of data processing application into reducers and mappers is sometimes non-trivial. When we write an application in the MapReduce form, the application go over to run over hundreds, thousands, or even tens of thousands of machines in a cluster is just a change in configuration. That is the reason many programmers has attracted to use the MapReduce model[1].

Following are some examples of Map-Reduce function usage in the industries:

1. At Google: Index building for Google Search .Article clustering for Google News. Statistical machine translation
2. At Yahoo!: Index building for Yahoo! Search. Spam detection for Yahoo! Mail
3. At Facebook: Data mining. Ad optimization. Spam detection
4. At Amazon: Product clustering. Statistical machine translation[3]

Execution of Moving data using MapReduce

In order to reduce the execution time of distributed computation, MapReduce tries to decrease moving the data from one node to another node by distributing the computation so that it is processed on the same node where the data is actually stored. In this way, the data store on the same node, but the code is shuffle through the network. This is natural because the code is much smaller than the data.
For execute a MapReduce task, the user has to execute two functions, map and reduce, and those executed functions are distributed to nodes that contain the data by the MapReduce framework. Every node executes the given functions on the data it has in order to reduce network shuffling data[2].

Implementation Model of MapReduce

There are total seven implementation methods involve in MapReduce.

1. Google MapReduce
2. Hadoop
3. GridGain
4. Mars
5. Titled-MapReduce
6. Phoenix
7. Twister

1. Google MapReduce

Google works on original MapReduce implementation which is main target to execute large clusters on networked machine. Its library automatically executes parallel processing and data distribution. (GFS) Google file system known as a distributed file system which makes a duplication of data blocks. GFS is help for increased reliability, fault tolerance and intended to view machine failures as a default rather than irregularity. MapReduce is highly extensible, so, it is run on cluster in place of thousands of low-cost machine. There are some advantages of Google MapReduce like Scalability: MapReduce is very high in scalability which can be scale across thousands of nodes. Parallel Nature: Structured and Unstructured data can work on the same time. Memory Requirements: Google MapReduce does not require large memory like other Hadoop ecosystems. It can work on minimum memory and can be produce quick result. Cost Reduction: As MapReduce is highly scalable, it's reduced the cost of storage and processing. Also, there are some disadvantages like slow processing speed of tasks and only support for batch processing [2].

2. Hadoop

Hadoop has the most popular implementation method because it provide open source feature. Hadoop provides Hadoop different frameworks which use for Hadoop application to run on large clusters. It is use for performing the cloud-based large-scale data parallel application for providing reliability and data transfer capabilities. It enables distributed, data-intensive and parallel applications by analysing a great job into smaller tasks and a large chunk of data set into smaller partitions in such a way that each task processes a different partition in parallel. So that Hadoop provides (HDFS) Hadoop distributed file system. HDFS contains a single NameNode and a master server to adjust the file system and client provide to give access to files. File divides into one or more blocks to store in set of Data Nodes. Data Nodes are used for execute 'Read' and 'Write' requests from the file's system client[10]. It is also execute creation of blocks, data replication and data deletion. Some advantages of Hadoop are Hadoop is use to provide distributed storage and computational capabilities. Hadoop is use for distributed storage and computational capabilities. Hadoop is highly extensible. Hadoop distributed file system provides large block size of processing data. Data Parallelization provided in Hadoop. Also disadvantages like as Hadoop uses Hadoop Distributed file system and MapReduce, both are master processes in single point of failure. Hadoop neither provide storage nor network level encryption. Hadoop Distributed file system does not handle small files efficiently and has not provided transparent compression[3].

3. GridGain

GridGain is basically based on Java which is open-source implementation for memory processing of big data in distributed environment. It is high performance in nature and it is use for a distributed, real-time in memory and scalable data grid to have a connection between application and data sources. GridGain provide fast implementation In-Memory MapReduce with In-Memory data Grid technology which is easy to use and easy to scale software. GridGain presents between business, analytics, transactional or BI application and its use for long term data storage such as RDBMS(Relational Database Management System), ERP or Hadoop HDFS. GridGain gives in-memory data platforms for high performance, low latency data storage and processing. Its also breaks a tasks into one or more subtasks and send them into a nodes, so that it's helps to improve load balancing capabilities. GridGain having some advanatages like GridGain is a distributed computation middleware that allows us to easily farm out arbitrary tasks to nodes. GridGain use to modify task execution to non-deterministic nature. Also, disadvantages like GridGain use for distributed computation but does not use for Hadoop Distributed File System. GridGain framework does not use the sort of data between Mapper and Reducer[3].

4. Mars

Mars is a used for Graphics Processing Units (GPUs).Mars has basically three components such as Map, Reduce and Groups. Map takes the input data from the disk for pre-processes, modifying the input data to key in the main memory. After that, it push input records from the main memory to the graphics processing main device memory. In the MapStage, Map divides input records to GPU threads that the workload for all threads is even. Each thread works the user-defined MapCount function to evaluate a local histogram of the number. Then, the runtime performs a graphics processing unit -based prefix sum on the local histograms to obtain the output size and the write position for each thread. .At last, after the output buffer is assigned to the device memory, each graphics processing unit thread implements the user-defined Map function and outputs the results. Advantage of Mars are Mars is the first implementation of MapReduce on GPUs and atomic-free. It permits users and programmers to take benefits of different processor on a single machine. Disadvantages are it is expensive in pre-processing phase. Map and Reduce function in pre-processing design need to be execute twice[3].

5. Titled-MapReduce

Titled- MapReduce is used in chip multiprocessor. The chip multiprocessor is very popular and it's runs on data-parallel applications in clusters on the single machine. MapReduce is used to program large scale clusters. It has controlled multicore platform, by solving large scale data-parallel issues. Titled-MapReduce, uses as "tiling strategy" to breaks a large MapReduce job into a number of one or more smaller sub-jobs and its handles the sub-jobs iteratively. Titled-MapReduce helps to increase the general Reduce phase to execute the partial result till all the iteration execute, instead of the intermediate data. So, the output of the Reduce phase is consistent with the output of the general Reduce phase. Advantages like it provides better data storage and task parallelism. Tiled-

MapReduce explores several optimizations, its helps to improve the memory and saves the memory up to 85%. Disadvantages like Tiled-MapReduce search the possible use of the tiling strategy in the single-node version of MapReduce[4].

6. Phoenix

Phoenix is use for shared-memory systems in MapReduce and its target to support for efficient implementation on multiple cores without give any difficulties to the programmer with parellel management. It's has own separate Application Programming Interface (API) for providing runtime to apply data parallelization, resource management and fault recovery. It is used by application programmer to achieve a target for multi-core and multiprocessor system. Shared-memory threads executes parallel task in Phoenix. At the first, a user give the runtime with the Map or Reduce functions for applying on the data. The runtime uses multiple worker threads to execute the computation tasks. In the Map phase the input data are breaks into smaller chunks, and the Map function is invoked on each chunk. This result gives intermediate key/value pairs. In Reduce phase, for each unique key, the Reduce function is called with the values for the same key as an argument and reduces them to a single key/value pair. The results of the Reduce tasks are combined and sorted by keys to get the last output.
Advantages of Phoenix are MapReduce runtime targeted for shared-memory multi-cores and multiprocessor. It provides similar performance for most applications. Phoenix automatically provides key scheduling decisions during parallel execution[4]. Phoenix provides significant speedups with both systems for all processor counts and benchmarks. Also, Disadvantages like Phoenix perform well for small-scale systems but does not large-scale shared memory system. Store and retrieve intermediate data in data structure are crucial in overall system performance[5].

7. Twister
Twister uses to identify the improvement in the programming model as well as its architecture to provide MapReduce to develop MapReduce runtime. The Map or Reduce tasks use with these two types of data products are studied which can be used to read any static data at the Map and Reduce tasks. Example, key and value pairs are stated as variable data in the Map phase of the computation and the static data which is already read produces a group of output (key, value) pairs. It also found an optional reduction phase named "combine" to associate the results of the Reduce phase into a single value. The user program and the merged operation are done in a single process space, which allows its output directly accessible to the user program. Advantages of twister are Twister helps to provide an orderly support for Iterative MapReduce computations. It provides some quality to support MapReduce computations like distinction on static and variable data. It merge different phases to gather all Reduce Final Output. And Disadvantage are in Scheduling Task, Google's MapReduce and Hadoop use a dynamic scheduling mechanism which is more effective than Twister. Hence, Scheduling task is not work properly. It is require the user to break large datasets into multiple files[6].

Comparison of implementation of methods with respect to volume and variety:

Methods	Feature	Impact with respect to volume	Impact with respect to variety	Uses in Existing system
Google MapReduce	Google performs aim to have large cluster and its library automatically handles parallelization and data distribution.	1. Sequential Execution mode is used for computing large volume of data. 2. Execution for large amount of data might take long time to process.	1. It supports fast and efficient processing on unstructured data on which any data can be arranged in structured data. 2. Output file generated by MapReduce might be removed before running for avoiding file exist exception.	Google is used in social media platforms like Facebook & Twitter.
Hadoop	It is an Open source platform. It is applicable for large-scale cloud-based system. It has HDFS for data distribution.	1. Data- Parallel application is used for processing the data. 2. HDFS used for holding terabytes or petabytes of data for providing high throughput access to the information.	1. Execution of large dataset is faster by providing the reliability and data transfer capabilities. 2. Hadoop has a low-level APIs.	Hadoop is used in social media platforms like Amazon, Facebook, Twitter
	It is used for open-source grid computing which is made for	1. GridGain offers a computing made for Java using distributed, real-time and in-memory.	1. GridGain provides 3 × performance solutions for Hadoop acceleration. 2. It is used in transactional and Non-	GridGain typically used in business, transactional or BI applications and long term

GridGain	java. Processing of big data is fast.	2. GridGain has low popularity and it is used in small dataset. 3. Performance is high.	Transactional live data with low latency.	data storage such as ERP and RDBMS.
Mars	It is Developed for Graphics Processing Unit. It provides tasks portioning, data distribution and parallelization.	1. Mars is 16 times faster than its CPU-based quad-core machine. 2. As Mars is deleveloped for GPU and it hides the programming complexity of GPU for simple and familiar MapReduce Interface.	Mars is used in Graphics Processing Unit.	Mars mainly use in Facebook by analyzes enterprise social networks like yammer.
Titled-MapReduce	Titled-MapReduce is used for Multicore Platforms. It is used to partition large MapReduce Job into a smaller sub-jobs and it handles the partitions job iteratively using multicore platforms.	1. It provides separate resources for executing sub jobs. 2.The Tiled-MapReduce programming model extension is given that allows exploiting multicore environments for data-parallel applications	1. An analysis is given wherein iteratively processing small chunks of data is more efficient than processing a large chunk of data for MapReduce on multicore platforms	It is used in online MapReduce Model
Phoenix	It is used for efficient implementation on multiple cores using API(Application Programming Interface).	1. Phoenix provides fast access to large amount data. 2. Performing single millisecond reads, writes, and updates, as well as fast table scans and we can scan 100 million rows in 20 seconds.	1. Phoenix is delivered as a client-embedded JDBC driver. 2. Accessing, Storing and Retrieving large amount of data is very easy in phoenix.	Phoenix is used in Digital Marketing, Google+, YouTube
Twister	It provides high quality services. It is used to identify the performance of programming model and architecture.	1. Twister takes long time for processing data	1. Twister is efficiently used for small-scale data.	Phoenix is used in latest social media "Phoenix Business Journal" where you can find business news, stories, all updates of day,etc.

Future Scope: As surveyed, MapReduce technique is the most efficient for large amount of data for parallel processing an distributed processing without considering issues like fault tolerance and reliability. In this paper determines and differentiate various implementation methods of MapReduce with respect to the volume and variety.

References:

[1] https://acadpubl.eu/jsi/2017-114-7-ICPCIT-2017/articles/12/34.pdf
[2] https://pergamos.lib.uoa.gr/uoa/dl/frontend/file/lib/default/data/2820037/theFile.
[3]https://www.researchgate.net/publication/286902556_Analytical_review_on_Hadoop_Distributed_file_system.
[4] https://ipads.se.sjtu.edu.cn/_media/publications/ostrich-taco13.pdf
[5] https://core.ac.uk/download/pdf/38466216.pdf
[6]https://www.researchgate.net/publication/220717676_Twister_A_Runtime_for_Iterative_Mapreduce
[7] https://www.researchgate.net/publication/269504730_Tiled-MapReduce
[8]https://www.academia.edu/26926208/Analysis_of_social_networking_data_using_Map_Reduce_and_Hadoop4.
[9]https://www.researchgate.net/publication/257929971_Survey_of_Parallel_Data_Processing_in_Context_with_MapReduce