Survey on MapReduce and Hadoop

[1]Prof. A. S. Kamble, [2]Kapratwar Abhishek, [3]Patel Ahad, [4]Khandagale Nikita, [5]Helwade Rushikesh

[1]Assistant Professor, [2,3,4,5]B.E Student
Department of Computer Engineering,
Modern Education Society's College of Engineering, Pune(MS), India.

***Abstract*: The word Big Data indicates to catch, regulate, process and analyze large amount of data that is structured, unstructured or semi-structured which has varying speed and growth. Analyzing such large and varying size of data is challenging using traditional structural databases. One of the important patterns in Big Data processing is HDFS which is better than the traditional structural databases. Hadoop is freely available software program. Hadoop being classified as a pivot platform for structuring Big Data and provides facility for analyzing the stored data. Hadoop can expand from a single machine to n number of machines depending on the demand of the users. This paper helps users to understand how Hadoop is better than the traditional structural databases for processing of Big Data and also gives a summary about how Map Reduce is used alongside with HDFS to do analysis on large datasets.**

***Index Terms*: Hadoop, Map Reduce, Big Data, HDFS.**

I. INTRODUCTION

Hadoop is an open source utilization of map reduce concept which is used in parallel processing in distributed environment to achieve high performance computing on different nodes. Working of Map/Reduce takes place in two phases which includes Map & Reduce. Map-Reduce and HDFS defines the core system of Hadoop. Map-Reduce model does the computation while distributed storage is done by HDFS.

The end user of the Map-Reduce repository specifies the data processing in two steps: Map & Reduce. Google's own web search service like data-mining, machine-learning, ordering and various supplementary systems uses Map-Reduce.

The Apriori algorithm is generally used to find the frequent item sets in dataset or Boolean Association rules. The prior knowledge of the frequent itemset properties is needed by the Apriori algorithm. To increase the efficiency of level-wise formation of frequent item sets and the efficiency of level wise frequent itemset generation can be improved by using Apriori property: First is, all non-empty subset of frequent itemset should be frequent. And second is, if any itemset is infrequent, its superset will be infrequent.

GFS i.e Google-File-System and HDFS i.e Hadoop-distributed-file-system work on end user level processes working on top of typical operating-system. Google-File-System is a distributed file system for data exhausting operations in application. It provides facilitation of storing data on hardware in a fault tolerant manner that throughputs high performance to a huge number of end users.

II. MAP-REDUCE

As shown in Figure 1 Map-Reduce is a concurrent programming schema that was made by Google. In this framework a user uses two functions for computation, Map & Reduce. Map function takes input as a key-value pair and produces an output of a list of values associated with the key. The mapping function is developed in such a way that multiple map functions can be carried out simultaneously. The output of map function is given as an input to the reduce function which performs processing on them to generate the designated outcomes.

It will automatically parallelize the computation and it also handles the complex issues such as load balancing, data distribution, etc. The first implementation of Map-Reduce as well as Hadoop is aimed for large cluster parallelization[1]. Because of its fault tolerance capacity, Map-Reduce has gained large popularity.

Map reduce is highly scalable because we can partition a job into number of smaller tasks that can run on several machines. Google did execute over one hundred percent Map-Reduce jobs per day and processed 20PB of data simultaneously in the year 2008. By 2010 Google had created numerous Map Reduce Programs which performs hundreds of parallel computation task. Google File system and HDFS both are targeted at massive data computing applications. Both favors high bandwidth. Both run on clusters. In both the system the file system works on end user level processes that is running on top of typical operating-system.

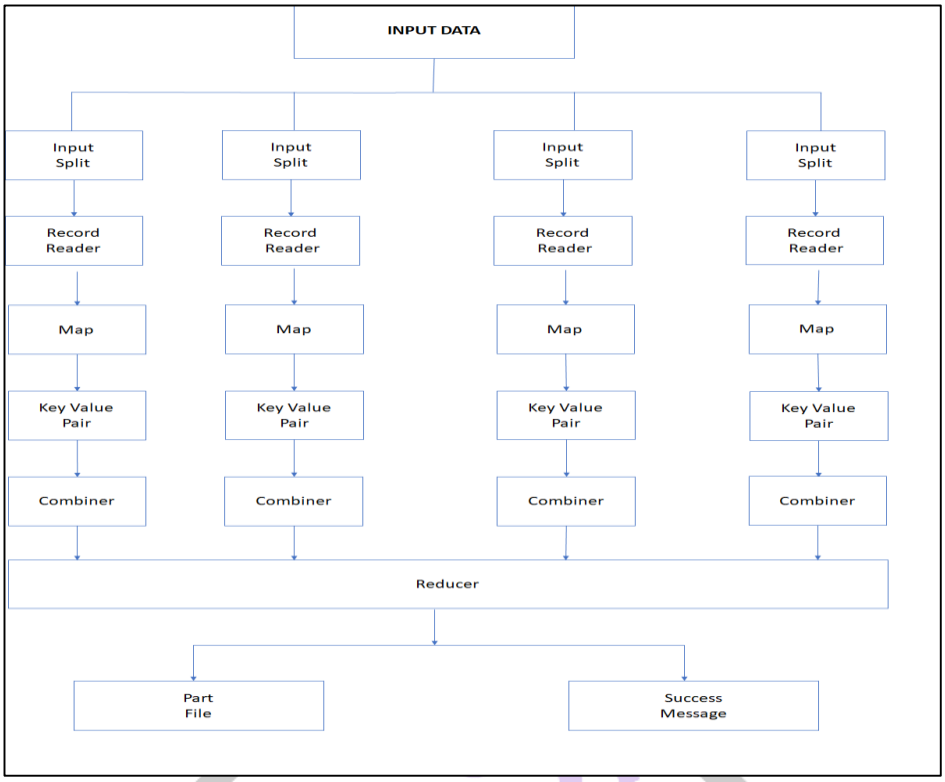

Figure 1. Map-Reduce Flowchart

III. APRIORI ALGORITHM

Apriori is one of the most classical algorithms for association rule which also executes to operate on database consisting of transactions. The main idea of Apriori is to produce candidate items in the database, and afterwards scanning the database to decide in order whether they meet the count. It uses a "bottom up"[2] fashion, the most frequent subsets are extended one by one and this group of candidates are examined against the data. The algorithm will terminate if no additional successful extensions are found. It proceeds as follows: Input taken: D, (Transaction in Database) and min_supp (threshold of minimum support), Output: L, (persistent item sets)

A. Apriori/Map-Reduce

Apriori/Map-Reduce is an algorithm that runs on concurrent map-reduce schema(ApacheHadoop) removes non-frequent item sets having occurrence of items less than minimum support.

The algorithm commences with the computation of persistent item sets at each map node. Then collect persistent itemset and remove the item sets which are having support less than minimum support at reduce node. Compute persistent item set having supplementary items by combining, ordering, and rejecting the replicated items that are present at each map node.

The time complexity of sequential Apriori algorithm is p times more than the Apriori/Map-Reduce algorithm in which p is the count of map&reduce node[3]. This algorithm can be extended to the Hadoop schema that produces test data by processing on transaction data.

IV. HADOOP

Hadoop is freely available schema that is used for dealing with large data sets in distributed background. Developed using Google's Map-Reduce schema where application or input is split in smaller parts. Hadoop ecosystem is composed of Hadoop-Kernel, Map-Reduce, HDFS and diversified elements such as Apache HBase and Zookeeper.

A. HDFS

HDFS is a distributed file system which primary aim of storing and processing of large datasets. It has various elements such as Name-Node, Data-Node, Block. It has low cost commodity hardware schema. The failure rate is high because low cost commodity hardware.

The grouping of data is done and these large group's of data are stored across thousands of nodes for further processing [4]. Hadoop is DFS which is built o commodity hardware. One of the important feature is high aggregate bandwidth.

HDFS file system yields high throughput access. Files are stored redundantly on machines to ensure that they are durable and highly available. It is a master-worker schema which comprises of a master(primary server) that controls the file-system namespace and handles the rights to the files. Along with this server there are number of datanodes generally which are solitary per node in a cluster that handles storage that is connected with the namenode. The architecture of HDFS shown in Figure 2.

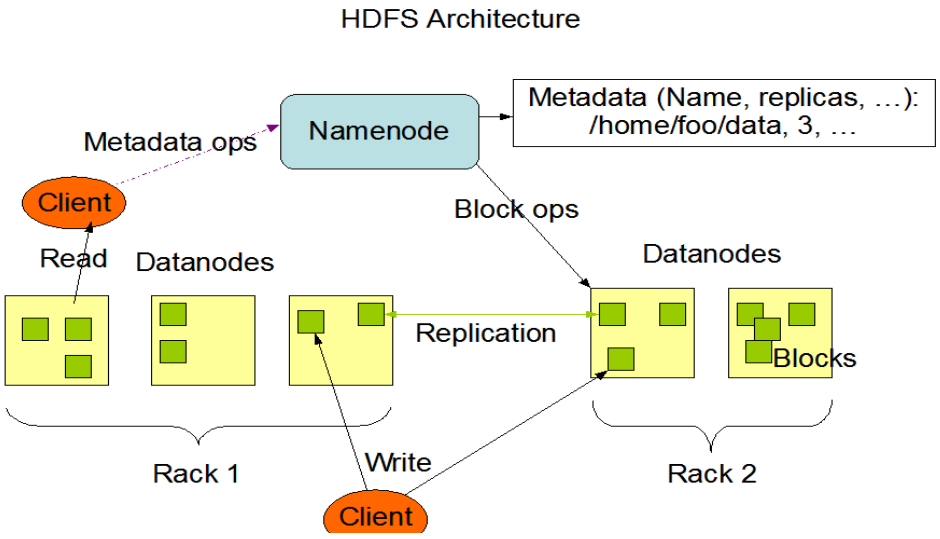

Figure 2. HDFS Architecture

V. DATA MINING & WAREHOUSING

Data mining can be very useful whenever a system has to deal with large data sets and Data warehousing is a good choice for maintaining such large dataset[5]. Data-Warehousing and Data-Mining consists the following, collecting the data from diverse origin then scrubbing and conversion of that data and updating the data tuples periodically.

A data-warehouse for a particular application can contain various fields and a cube can be made from these field's where the attributes are data tuples and its records are stored in the cells of the cube. Any particular view of this data-cube is deliberated as data-mart.

A. *Mining Techniques*

There are various data mining techniques that are available for extracting meaningful information from raw data. Following are few of them[6]:

1) Clustering: In this we form groups that are called clusters and the objects that are in a particular cluster are similar to each other than the other cluster. Clustering is used in many field's like ML, image analytics, pattern recognition. Every cluster that is formed is seen as a class of object, from which rules can be formed.

2) Decision Tree: It's a decision support tool which utilizes tree like graph-model for concluding and predicting their follow through. It's easy to comprehend and interpret and manages to produce good result with even small amount of information.

3) Factor Analysis: It's a analytical approach that is used to determine inconsistency between the examined, corresponding attribute regarding the probability of lower number of neglected, not associated attributes that are called factors

4) Regression Analysis: It's a technique that's used for modeling and analyzing many variables, when the emphasis is on relationship among one or more independent variables and one dependent variable.

REFERENCES

[1] Madhavi Vaidya, "Parallel Processing of Cluster by Map Reduce", International Journal of Distributed and Parallel System (IJDPS), January 2012.

[2] Xin Yue Yang, Zhen Liu, Yan Fu, "Map Reduce as a Programming Model for Association Rules Algorithm on Hadoop", International Conference on Information Sciences and Interaction Sciences, August 2010.

[3] Jongwook Woo, "Apriori-Map/Reduce Algorithm", International Journal of Distributed and Parallel System, 2012.

[4] Harshwardhan S. Bhosale, Prof. Devendra P. Gadekar, "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publication, October 2014.

[5] Sonali Agarwal, G. N. Pandey, M. D. Tiwari, "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, April 2012.

[6] Monika Goyal, Rajan Vohra, "Application of Data Mining in Higher Education", International Journal of Computer Science, March 2012.

[7] Jeffrey Dean, Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Magazine Communications of the ACM, January 2008.

[8] Katrina Sin, Loganathan Muthu, "Application of Big Data in Education Data Mining and Learning Analytics - A Literature Review", ICTACT Journal on Soft Computing, July 2015.

[9] B. Manjulatha, Ambica Venna, K. Soumya, "Implementation of Hadoop Operations For Big Data Processing in Educational Institutions", International Journal of Innovative Research In Computer and Communication Engineering, April 2016.