*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VII (Nov – Dec. 2014), PP 37-40*
*www.iosrjournals.org*

# Big Data Mining using Map Reduce: A Survey Paper

Shital Suryawanshi[1], Prof. V.S.Wadne[2]
*[1]PG Student, Computer Engineering Department, Savitribai Phule Pune University JSPM's Imperial College of Engg. & Research, Wagholi, Pune, India.*
*[2]Assistant Professor, Department of Computer Engineering Department, Savitribai Phule Pune University JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India.*

***Abstract:*** *Big data is large volume, heterogeneous, distributed data. Big data applications where data collection has grown continuously, it is expensive to manage, capture or extract and process data using existing software tools. For example Weather Forecasting, Electricity Demand Supply, social media and so on. With increasing size of data in data warehouse it is expensive to perform data analysis. Data cube commonly abstracting and summarizing databases. It is way of structuring data in different n dimensions for analysis over some measure of interest. For data processing Big data processing framework relay on cluster computers and parallel execution framework provided by Map-Reduce. Extending cube computation techniques to this paradigm. MR-Cube is framework (based on mapreduce)used for cube materialization and mining over massive datasets using holistic measure. MR-Cube efficiently computes cube with holistic measures over billion-tuple datasets.*

***Keywords:*** *big data, data cube, cube materialization, Map Reduce, MR-Cube.*

## I. Introduction

In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. upto 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1].for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012[10].this shows that it is very difficult for big data applications to manage, process and retrieve data from large volume of data using existing software tools. It's become challenge to extract knowledgeable information for future use [15]. There are different challenges of Data mining with Big Data. We overlook it in next section. Currently Big Data processing depends upon parallel programming models like MapReduce, as well as providing computing platform of Big Data services. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter. Due to the large size of data it is becoming expensive to analysis data cube. The Map-Reduce based approach is used for data cube materialization and mining over massive datasets using holistic (non algebraic) measures like TOP-k for the top-k most frequent queries. MR-Cube approach is used for efficient cube computation.

Our paper is organized as follows: first we will see key challenges of Big Data Mining then we overlook some methods like cube materialization, MapReduce and MR-cube approach.

## II. Challenges In Big Data Mining

Big Data has different characteristics such as it is large volume, heterogeneous, autonomous source with distributed and centralized control, seek to explore complex and evolving relationship among data [1].These different characteristics of Big Data make it challenge for discovering useful information or knowledge from it. After analyzing and research challenge form a three tier structure framework to mention different challenges at different tier, as shown in fig.1.

The challenges at tier I focus on low-level data accessing and arithmetic computing procedures, Challenges on information sharing and privacy. Big Data often stored on different location and it is continuously growing that's why an effective computing platform to take distributed large scale data storage into consideration for computing. Tier II concentrate on high-level semantics, application domain knowledge for different applications of big data and the user privacy issues. This information provides benefits to Big data access but also add a technical barriers to Big Data access (Tier I) and mining algorithms (Tier II). The Outmost tier is tier III which challenges the actual mining algorithms. At this tier III the mining challenges concentrate on algorithm designs in tacking the difficulties which is raised by the big data volumes, distributed data distribution, complex and dynamic characteristics

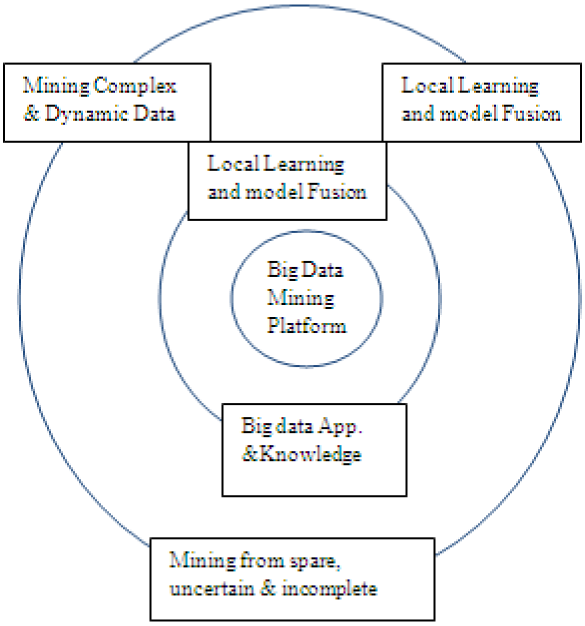
Mining Complex & Dynamic Data
Local Learning and model Fusion
Local Learning and model Fusion
Big Data Mining Platform
Big data App. &Knowledge
Mining from spare, uncertain & incomplete

**Fig.1. A Big Data processing Framework**

Tier III contains three stages. In first stage sparse, heterogeneous, uncertain, incomplete and multisource data is preprocessed by data fusion technique. In second stage after preprocessing stage complex and dynamic data are mined. Third stage is for local learning and model fusion, where the global knowledge is obtained by local learning and model fusion is tested and the relevant information is feedback to preprocessing stage.

Big Data is carry out computing on the PB (Petabyte) or even on EB (Exabyte) data with complex computing process, so parallel computing infrastructure, programming language support and software model utilizing to efficiently analyze and mine distributed data. MapReduce mechanism is suitable for large scale data mining task on clusters.

## III. Method Overview

### 3.1 Data Cube

Data cube provide multi-dimensional views in data warehousing. If n dimensions given in relation then there are 2^n cuboids and this cuboids need to computed in the cube materialization using algorithm[2]which is able to facilitate feature in MapReduce for efficient cube computation. In data cube Dimension and attributes are the set of attributes that user want to analyze. Cube lattice is formed representing all possible groupings of this attributes, based on those attributes. After that by grouping attribute into hierarchies and eliminating invalid cube regions from lattice we get more compact hierarchical cube lattice. Finally cube computation task is to compute given measure for all valid cube groups. There are different techniques of cube computations [3] like multi- dimensional aggregate computation, BUC(Bottom-Up Computation), star cubing for efficient cube computation. There are limitations in these techniques: 1) They are designed for a single node or for a cluster with less nodes [19], so it is difficult to process data with a single or few machines. 2) Many analyses over logs, involve computing holistic measure where as many techniques uses the algebraic measures. 3) Existing techniques failed to detect and avoid data skew. There is need of technique to compute cube in parallel on holistic measure over massive dataset. Hadoop based MapReduce can handle large amount of data in cluster with thousand of machines. So this technique is good option for analysis of data.

### 3.2 Map Reduce

MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. The nature of this programming model and how it can be used to write programs which run in the Hadoop environment is explain by this model. Hadoop [11] is an open source implementation for this environment. Map and Reduce are two functions.

The main job of these two functions are sorting and filtering input data. During Map phase data is distributed to mapper machines and by parallel processing the subset it produces <key ,value>pairs for each record. Next shuffle phase is used for repartitioning and sorting that pair within each partition. So the value corresponding same key grouped into {v1, v2,….}values. Last during Reduce phase reducer machine process subset <key, {v1, v2,}>pairs parallel in the final result is written to distributed file system.

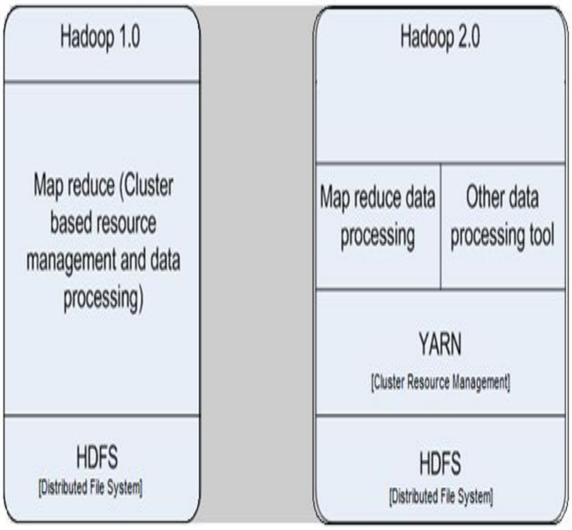
Hadoop 1.0
Map reduce (Cluster based resource management and data processing)
HDFS
[Distributed File System]
Hadoop 2.0
Map reduce data processing
Other data processing tool
YARN
[Cluster Resource Management]
HDFS
[Distributed File System]

**Fig.1:- Architecture of Hadoop1.0 and 2.0**

MR1 is used in Hadoop1.0 but due to some resource management issues like inflexible slot configuration, scalability. After Hadoop version 0.23, MapReduce changed significantly. Now it known as MapReduce 2.0 or YARN (Yet Another Resource Negotiator). MapReduce 2.0 has two major functionalities of job tracker which are spit into resource management and job scheduling into separate daemons [4]. Fig 1 shows the architecture of both Hadoop versions. In Hadoop1.0 Job Tracker has a responsibility for managing the resources and scheduling jobs across the cluster. But in Hadoop2.0 the architecture of YARN allows the new Resource Manager to manage the usage of resources across all applications. And Application Masters takes the responsibility of managing the job execution.

This new approach improves the ability to scale up the Hadoop clusters to a much larger configuration than it was previously possible. In addition to this, YARN permits parallel execution of a range of programming models. This includes graph processing, iterative processing, machine learning, and general cluster computing.

### 3.3 MR-cube Approach

MR-Cube MR-Cube is a MapReduce based algorithm introduces for efficient cube computation [5] and for identifying cube sets/groups on holistic measures. MR-Cube algorithm is used for cube materialization and identifying interesting cube groups. Complexity of the cubing task is depending upon two aspects: size of data and size of cube lattice. Size of data impacts size of large group and intermediate size of data, where as the cube lattice size impacts on intermediate data size and it is controlled by the number/depth of dimension. First we identify the subset of holistic measures that can easily compute in parallel than an arbitrary holistic measure. We can call it Partially Algebraic Measures. The technique of partitioning large groups based on algebraic attribute called Value partitioning. Value partitioning is used to effectively distribute the data; we can easily compute it with Naïve algorithm [9]. Value partitioning performs on only on group that are likely reducer friendly and dynamically adjust the partition factor. Partition factor is ratio by which a group is partitioned.

There are different approaches for detecting reducer unfriendly groups. One of the approach is sampling approach where we estimate the reducer unfriendliness of cube region based on the number of groups it is estimated and perform partitioning for all small groups within the list of cube region that are estimated to be reducer unfriendly.

### 3.4 Cube Materialization

Cube materialization task comes under the MR-Cube approach. Materializing the cube means computing measures for all cube groups satisfying the pruning condition. After materializing cube we can identify the interesting cube groups for cube mining algorithm. The main MR-CUBE-MAP-REDUCE task is perform using annotated lattice. The combine process of identifying and value partitioning unfriendly regions followed by partitioning of regions is referred as annotate.

Based on the sampling results cube regions have deemed as reducer unfriendly and require partitioning. Each tuple in dataset the MR-Cube-Map emits key:value pairs for each batch area. In required keys are appended with hash based on value partitioning. The shuffle phase then sorts them by key yielding reducer tasks. The BUC algorithm is then run on each reducer and cube aggregates are generated. The value partitioned group are merged during post processing to produce the final result.

## IV. Conclusion

In real-world applications managing and mining Big Data is Challenging task, As the data concern large in a volume, distributed and decentralized control and complex. There are several challenges at data, model and system level. We need computing platform to handle this Big Data. The MapReduce framework is one of the most important parts of big data processing, and batch oriented parallel computing model. In earlier versions of MapReduce the components were designed to address basic needs of processing and resource management. Recently, it has evolved into a improved version known as MapReduce 2/YARN that provides improved features and functionality. With Big Data technologies we able to provide most relevant and accurate social sensing feedback to better understand to society at real-time. MR-Cube efficiently distributes the computation workload across machines and completes the cubing task.

## Acknowledgement

I would like to thank all of them who help me in my work.

## References

[1]. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

[2]. Zhengkui Wang, Yan Chu, Kian-Lee Tan, Divyakant Agrawal, Amr EI Abbadi, Xiaolong Xu, "Scalable Data Cube Analysis over Big Data" appliarXiv:1311.5663v1 [cs.DB] 22 Nov 2013

[3]. Dhanshri S. Lad #, Rasika P. Saste, "Different Cube Computation Approaches: Survey Paper" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4057-4061

[4]. The Apache Software Foundation"http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html"

[5]. Arnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan, Fellow, IEEE, "Data Cube Materialization and Mining over MapReduce" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 6, NO. 1, JANUARY 2012

[6]. A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[7]. D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.

[8]. A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033,2012.

[9]. A. Nandi, C. Yu, P. Bohannon. And R. Ramakrishnan, "Distributed Cube Materialization on Holistic Measures, " Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.

[10]. F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[11]. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo" Yahoo! Developer Network Blog, 2008.

[12]. "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[13]. A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011.

[14]. K. Yury, "Applying Map-Reduce paradigm for parallel closed cube computation," Proc. First Int'l

[15]. Hadoop. http://hadoop.apache.org/.

[16]. P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquini. Incoop: Mapreduce for incremental computations. In SOCC, 2011.

[17]. Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. Haloop: Efficient iterative data processing on large clusters. PVLDB, 3(1):285–296, 2010.

[18]. Iman Elghandour and Ashraf Aboulnaga. Restore: Reusing results of mapreduce jobs. PVLDB, 5(6):586–597, 2012.