Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776769

MAP REDUCE-MECHANISM TO MANAGE BIG VOLUME DATA SETS

Ms. Harpreet Kaur
Ph.D Research Scholar in Computer Science & Applications
Desh Bhagat University, harpreet.hcl@gmail.com

Prof.(Dr.) R.K.Bathla
Senior Professor,Department of Computer Science & Applications
Desh Bhagat University, Prof.bathla@gmail.com

ABSTRACT
Today the term big data produces a lot of recognition. Agencies store huge Amounts of data and extract useful information from those databases to discover desired Patterns and interrelationships among them that human brains can't understand by human brains.The data volume range crosses our ability to process and generate from divergent Structured, Unstructured and semi structure resources. The main focus is on the framework of big data Hadoop , Map Reduce Environment and various tools related to big data which plays a very vital role to handle huge Data volume.Social networking sites like Facebook, Twitter produce large volumes of data which will be unmanageable within a few years. In order to manage these data sets, the proposed method uses various algorithms for processing this huge amount of data. Main Purpose of the Map Reduce programming model is processing and producing large datasets clusters that control a variety of real-world tasks. It has a main two function map and reduce and runtime system performs parallel processing across various machines and also handles all other networking jobs. Basically it focuses on the analysis of data, especially based on the user's needs. For this purpose Map Reduce performs the task of mapping, combining, partitioning, joining and reducing. It runs on large Data Sets on different machines which are highly quantifiable Programs. Many Map Reduce Programs are executed on Google's Data sets every day from many years.

Keywords: Big Data, Machine learning, Map Reduce, Virtualization,Algorithms of Map Reduce , Map Reduce Framework.

• 1.INTRODUCTION

Due to the explosion of machine-generated data like data records, web-log files, and sensor data and from growing social networks Data volume is also growing exponentially [1]. According to the 2011 Digital Universe Study, 130 exabytes of data were created and stored in 2005[3]. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 Exabytes in 2015[3].The growth of data will never stop and advancements in technology has given rise to an ecosystem of software and hardware products. Big Data technologies as a new generation of technologies and architectures designed to extract value from very large volumes of data. Big data increases the processing capacity of traditional database systems because data is too big, moves too fast, or is not handled by existing database architectures.Today datasets are within the range of terabytes but soon they could reach

petabytes or even Exabytes. The data has to be collected, stored and distributed at levels that would quickly overwhelm traditional management techniques. The velocity of data in terms of the frequency is also an attribute of big data. The data which is coming from different organizations are at millisecond rates and very high in speed .Big data contains a variety of data such as text, sensor data, audio, and video, click streams, log files etc which is composed from social media, document management and various government resources. It is in both structured and unstructured forms. Data Veracity is defined by accuracy and reliability of the data. A data set may have very reliable data with low precision based on the various methods and tools.

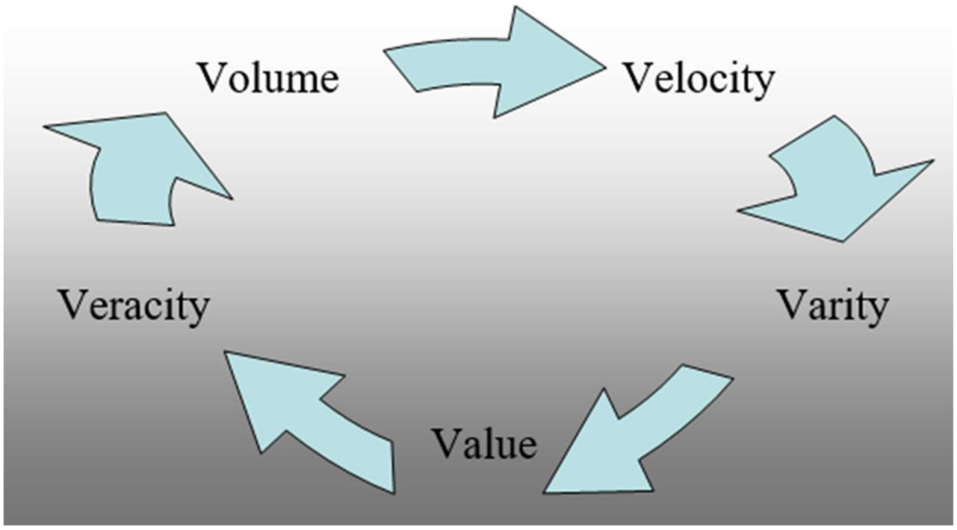

Fig 1 Bi\g Data Characteristics

The data generated is usually categorized as structured, semi-structured and unstructured.

1) Structured data: This type of data information includes organization databases, data warehouses and enterprise solutions. Data is stored into a relational scheme, it responds to simple queries based on an organization's parameters and requirements.

2) Unstructured data: These data are raw data that has been extracted from applications on the Internet. It has not been further processed to organize in meaningful formats. These formats cannot easily be indexed into relational tables for querying for example images, audio and video files.

3) Semi-structured data: This type of data is a combination of structured and unstructured data such as social media data, location type data, and user-generated data. It has no fixed schema and contains self-describing tags or other markers. For example weblogs and social media feeds.

To structure the data Hadoop Map Reduce and collaborative filtering approach are used .MapReduce model are used to manage the user's heavy data and for solving problems such as

who are the common friends/followers between you and another user on Facebook, Twitter or common connections in LinkedIn between two users. Some other factors like who reads your post on Facebook or Twitter can be quantified using the MapReduce programming model, it also compute who reads your profile etc.

2. METHODOLOGY USED IN BIG DATA

In order to extract value from extremely large volumes of multi-structured data from multiple different sources, companies need advanced big data applications that will enable them to quickly access and analyze that data. Some of the framework which is used to solve the problem in computation of big data is:

a)Hadoop: Hadoop is a free, Java-based programming framework that supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop provides scalable, cost effective, and flexible and fault tolerant solutions and lowers the risk of an entire system failure. It is used by popular companies like Google, Yahoo, Amazon and IBM etc.It has two main sub projects Map Reduce and Hadoop Distributed File System (HDFS).

i) Map Reduce divides the data into individual chunks which are processed by Map jobs in parallel [1].

The outputs of the maps sorted by the framework are then input to the reduced tasks. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework and input and the output of the job are both stored in a file-system.

ii) Hadoop Distributed File System (HDFS) spans all the nodes in a Hadoop cluster for data storage [1].

It links together file systems on local nodes to make it into one large file system.Hadoop also refers to a collection of other software projects that uses the Map Reduce and HDFS framework.

(b) HPPC: It is also known as the Data Analytics Supercomputer (DAS) developed by LexisNexis Risk Solutions.HPPC supports both batch and real-time data processing. (High Performance Computing Cluster) is a massive parallel-processing computing platform that Solves Big Data problems.

(c) Storm: It uses the open source Eclipse Public License [10].It does real-time processing what Hadoop does for batch processing.

(d) GridGain: It offers an alternative to Hadoop's Map-Reduce that is compatible with the Hadoop Distributed File System. It offers in-memory processing for fast analysis of real-time Data.

(e) Spark: It is open source cluster computing system that focus to make data analytics Fast for processing.

(f) GraphLab: A machine learning toolkits fully redesigned for providing distributed API,HDFS Integration and a wide range of new tools

(g) Dryad: It is investigating programming models for writing parallel and distributed programs To scale from a small cluster to a large data-center.

(h) Apache Flink – This is also an open source distributed data processing platform. Join, map,Group are used for Distributed programs.

(i) Storm: It is a free and open source distributed real-time computation system and Easy to process unbounded streams of data, doing real-time processing for batch Processing.
(j) Disco: It is a lightweight, open-source framework for distributed computing based on the Map Reduce methodology.
(k) Phoenix -It is a shared-memory implementation of Google's Map Reduce model for data-Intensive processing tasks.
(l) Plasma: It is a distributed file system for large files, implemented in user space and runs the Famous algorithm scheme for mapping and rearranging large files.

3.MACHINE LEARNING TOOLS FOR BIG DATA

Machine learning is a method of data analysis using algorithms iteratively. The iterative aspect of machine learning is important because they learn from previous computations to produce reliable, repeatable decisions and results .Machine learning algorithms including Fraud detection, Web search results, Email spam filtering, Pattern and image recognition, Network intrusion detection, Prediction of equipment failures etc.

Virtualization technology has become fundamental in modern computing environments such as cloud computing [14][13][10].It allows us to achieve a high utilization of the available hardware resources, security, reliability, scalability(e.g., [15][16][19]).The virtualization infrastructure provided by Virtual Box[20].Virtual Box is an open-source multi-platform. Virtual Box is designed in levels and provides SDK.Lower level is the hypervisor known as heart of the virtualization engine. Above the hypervisor there are modules that provide additional functionality for example Remote Desktop Protocol. The API level is implemented above these functional blocks.VirtualBox comes with a web service that, once running, acts as HTTP server, accepts SOAP connections [18][2] and processes them. It is possible to write client programs in any programming language such as Java, C++, NET PHP, Python, and Perl. Most widely adopted machine learning methods are supervised learning which is mostly used and unsupervised learning only 10 to 20 percent. Semi-supervised and reinforcement learning are two other technologies that are sometimes used [21].
Supervised learning receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors and then modifies the model accordingly using methods like regression, prediction and gradient boosting.
Unsupervised learning is used against data that has no historical labels and works well on transactional data. It explores the data and finds some structure within.Methodes includes self-organizing maps, nearest-neighbor mapping, and k-means clustering.
Semi-supervised learning uses both labeled and unlabeled data. Methods used are classification, regression and prediction same as supervised learning. Early examples of this include identifying a person's face on a webcam [2].
Reinforcement learning is often used for robotics, gaming because it is discovered through trial and error. This type of learning has three primary components: the learner or decision maker, the environment and actions [2] [21].

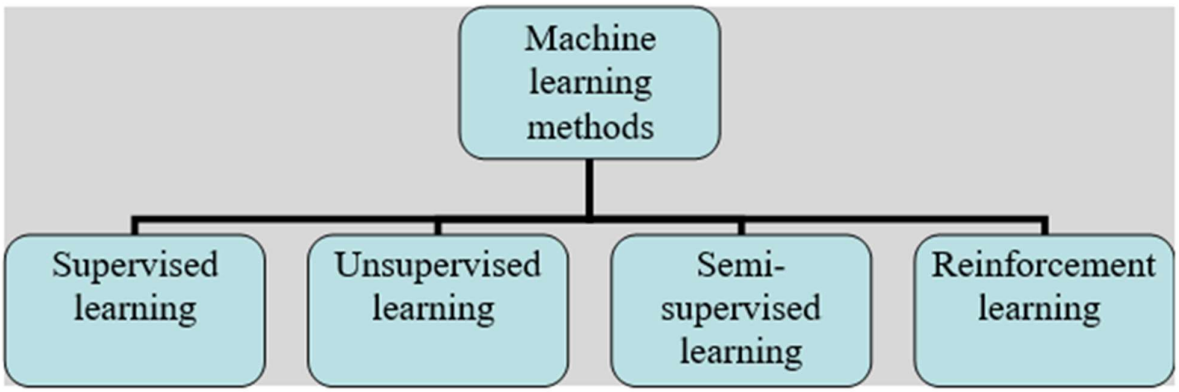

Fig 2 Methods of Machine Learning

Machine learning is used to reproduce known patterns and knowledge, automatically apply that to other data, and then automatically apply those results to decision making and actions. [2]
Some Algorithms of machine learning are [21]:

- Neural networks.
- Decision trees.
- Associations and sequence discovery.
- Random forests.
- Support vector machines.
- Nearest-neighbor mapping.
- K-means clustering.
- Self-organizing maps.
- Bayesian networks.
- Kernel density estimation
- Singular value decomposition.
- Gaussian mixture models.
- Sequential covering rule building.

4. MAP REDUCE PROPOSED SYSTEM

MapReduce Framework is a Technique to process data parallel by the distribution of data. The huge volume data divided into chunks has to be checked for interdependencies to avoid critical problems while aggregation of these resulting sets to get the required structured data. The data have to be clustered based on their deadline scheduled for processing, priorities and data dependencies. If processing of one data requires the output of other data as its input, then it can be combined together to form a cluster. The clusters can also be formed on the basis of priority and processing of the data clusters. MapReduce technique is mainly used for parallel processing of data sets across various clusters known as filtering, performed by the map function and generating computation results by aggregation and known as reduce function. To Process the heterogeneous data map Join Reduce technique is used. In case of a single node failure incomplete reduce tasks will be re-executed instead of the entire map and reduce tasks. Google released a paper on MapReduce technology in December, 2004. This became the root of the Hadoop Processing Model. So, MapReduce is a programming model that allows us to perform parallel and distributed processing on huge data sets.

- MapReduce works by breaking the process into two phases:

○ 1. Map phase: The first is the map job that takes the set of data and converts it into a further set of data,

○ where particular elements are broken down into tuples (key/value pairs).

○ 2. Reduce phase: Second the reduce job takes the output from a map as input and combines those data

○ tuples into a smaller set of tuples. As the sequence of MapReduce, the reduce job is always performed

○ after the map job. Further we can say it contains three tasks such as Mapper, Combiner and Partitioner.

○ Mapper involves the mapping of data, combiner combines the mapped data and partitions splits the

○ data into small clusters. On large data sets Map Reduce allows us to perform distributed and parallel

○ processing in a distributed environment.

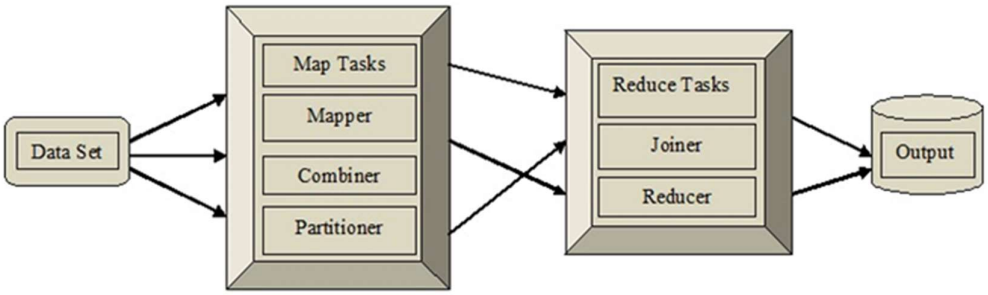

Fig. 3 Flow of Map Reduce Process

Let us consider an example to understand how a MapReduce works are as follows:

Disk, Buffer, Rom, Computer, Computer, Rom, Disk, Computer and Buffer This is a word count example on the new.txt using MapReduce. So, the work will be focused on finding the different words and counting the occurrences of those different words.

- Take input in three splits which will distribute the work among all the map nodes.
- Then assign the value (1) for each mapper word. The logic behind giving a fixed value equal to 1 is that every word will occur once.
- In the next step key-value pair will be generated where the key is nothing but the individual words and value is one. For example in the first line (Disk Buffer River) three key-value pairs – Disk, 1; Buffer, 1; Rom, 1.
- In the partition process after sorting, shuffling performed and all the rows with the same key are sent to the corresponding reducer which will have a unique key and a list of values similar to that

  every key like Buffer, [1,1]; Computer, [1,1,1].., etc.
- Reducer counts the values present in that list of values and counts the number of ones in the list, in the last all the output key/value pairs collected and written in the output file.

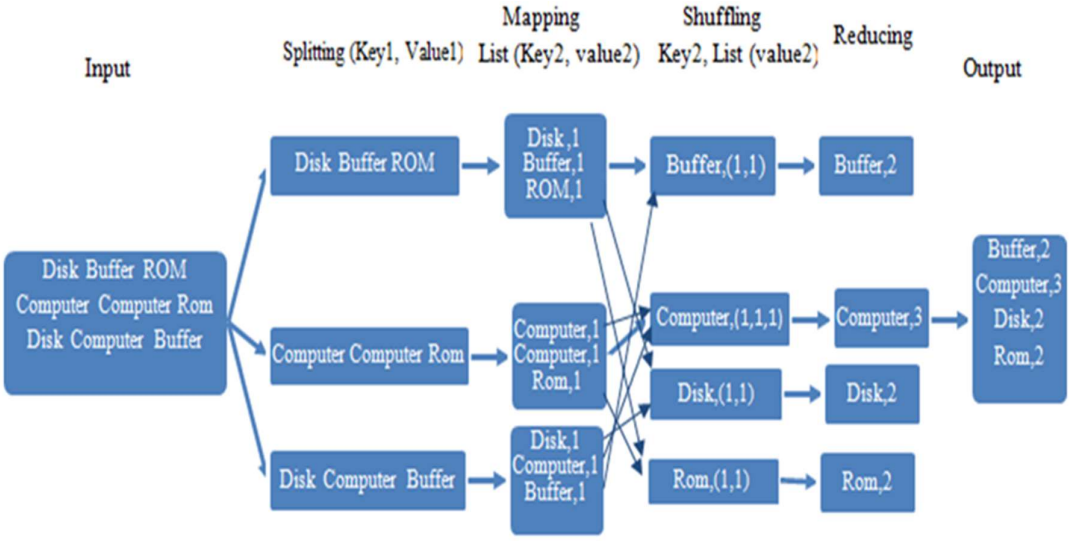

Fig. 4 Execution Overview of MapReduce

5 ALGORITHMS OF MAP REDUCE

○ Basic methodology used by MapReduce is Mapping and Reducing which is derived from Mapper Class and Reducer Class has their own work like mapper gives input, adds hard count, map and sorting and gives input to Reducer class and it performs matching pairs and reduces them. Some of the terms used in this process are Job Tracker which schedules all the jobs and follows the jobs given using Task Tracker which actually tracks the jobs and gives the report to the Job Tracker. The Main Job of Master Node is Job Tracking. It accepts job requests from different clients. On the other hand mapping and reducing programs are run on Slave Node.

Filtering Algorithms Mapping Phase uses the filtering phenomenon in which finding files or data items with particular characteristics. It also searches patterns in web logs or files.

Sorting Algorithms Keys automatically sorted Mapper output and forward to Reducer which plays a very important role in sorting and analyzing data. Sorting techniques are applied in the mapper class itself.

Searching Algorithms Searching data items is a special terminology which acts an important role in MapReduce.Most of the time combiner phase and Reducer phase performs searching.

Aggregation Algorithms Reducer plays a very important role in this phase,Mapper has its own individuality. Reducer performs Computational tasks like minimum, maximum, sum, average, counts the number of Tweets per day etc.

Indexing Algorithms Indexing plays a very crucial role, for example Search engines like Google and Bing use inverted indexing techniques as in the similar way map reduce works according to the same concept. Indexing is used to point to a particular set of data items and its locations. Batch indexing is also used by Mapper on the input files.

Joining Algorithms In Map Reduce Joining of two large dataset needs lots of code implementation by comparing the size of each dataset. The main focus is on smaller dataset and distributed it to every data node in the cluster. Further according to requirement either Mapper or Reducer side uses this smaller dataset to produce output records. Join key plays a very important role in this process.

TF-IDF Algorithms

Full Form of TF is Term Frequency and IDM is Inverse Document Frequency which is a text processing algorithm. Frequency describes the number of times a term appears in a document and calculated using total number of words counted in a file dividing by total number of words in that file.TF-IDF is also known as web analysis algorithm.IDM is calculated by the number of files in the database divided by the number of files where a specific term appears.

Link analysis algorithm This algorithm allocates weights to each vertex in a graph by calculating the weight of each vertex depending on the weight of its nearest neighbors.

Page Rank Algorithm Page Rank can be described as a join followed by an update with two aggregations which are repeated until stopping conditions occur. It is a part of relational algebra.

K-Means Clustering Algorithm This methodology is used to categorize semi structured or unstructured data sets. This is the most efficient technique that deals with huge amounts of data with simplicity.Basically It works with the number of clusters and the distance between each data item is calculated with each of the centroids of the respective cluster. Least distance data item is given to the cluster and distance recalculated which is known as Euclidean Distance used for comparison of points.

○ 6. Conclusion

During the last decade Big-data computing is the biggest innovation in the technical world. In every field we have to deal with a huge volume of data and need to process and organize it. Data. Also discussed machine learning concept and its algorithms for computation of big data. Virtualized environments which decline the application workloads and provide great functionality are also discussed in this paper. But big data is bound to lead to some challenges, there needs to be further improvement in the applications and technologies used to handle big data.So the performance and workload can be enhanced and issues related to big data improvised and provide opportunity for greater success.This paper explains, the unstructured data is structured and processed by using Map Reduce framework automatically arranged according user's requirements which is done through splitting, mapping, shuffling and reducing.MapReduce is the most efficient technique for processing and analysis of large volume of data sets. MapReduce programming Technique has been successfully used at Google for many different purposes due to its easy use. Even having less experience programmers works with parallel and distributed systems, because it hides the details of fault tolerance, load balancing and locality optimization.MapReduce computations are used to processed large variety of problems and performs filtering,sorting,data mining, and machine learning for the generation of data for Google's production web search service.

7. REFERENCES

[1]Jain, V.K. & Kumar, S. (2015). Big Data Analytic Using Cloud Computing. Advances in Computing

and Communication Engineering, Second International Conference, 667-672.

[2] Cuzzocrea, A., Mumolo, E. & Corona, P. (2015). Cloud-based Machine Learning Tools for Enhanced

Big Data Applications. Cluster, Cloud and Grid Computing, 15th IEEE/ACM International

Symposium, 908-914.
[3]Ovum. What is Big Data: The End Game. [Online] Available from: http://ovum.com/research/what-is-big-data-theend-game.
[4] IBM. Data growth and standards. [Online] Available from: http://www.ibm.com/developerworks/xml/library/xdatagrowth/index.html?ca=drs.
[5]IDC. The 2011 Digital Universe Study: Extracting Value from Chaos. [Online] Available from:http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf .
[6]James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity.
[Online]Availablefrom:http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/
big_data_the_next_frontier_for_innovation .
[7]Edd Dumbill. What is big data? [Online] Available from: http://radar.oreilly.com/2012/01/what-is-big-data.html.
[8]Carl W. Olofson, Dan Vesset. Worldwide Hadoop – Map Reduce Ecosystem Software 2012-2016
Forecast [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=234294 .
[9]Inukollu, V .N., Arsi S. & Ravuri S. R.(May 2014). SECURITY ISSUES ASSOCIATED WITH BIG
DATA IN CLOUD COMPUTING. International Journal of Network Security & Its Applications
(IJNSA), Vol6(No3).
[10] Big Data: 3 Open Source Tools to Know Firmex http://www. firmex.com/blog/big-data-3-open-source-tools-to-know.
[11]Cisco Cloud Computing Data Center Strategy, Architecture,and Solutions[Online]http://www.cisco.com/web/strategy/educa-tion/index.html
[12]Abouzeid et al.. (2009). HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. PVLDB, 2(1), 922-933.
[13] Bleikers, S.,Vogel, C. &Gro T.(2014).Cloud radar: near real-time detection of security failures in
dynamic virtualized infrastructures, ACSAC, 26–35.
[14] Bruder, G., Steinicke, F. & Nchter A.(2014). Poster: Immersive point cloud virtual environments,
3DUI, 161–162
● [15] Carlson, M.(2014).Systems and Virtualization Management: Standards and the Cloud. A report on
● SVM 2013. Journal of Network Systems Management,22(4), 709–715
● [16] Cho Y. & Choi J.(2013)An integrated management system of virtual resources based on
● virtualization API and data distribution service,CAC , 26
● [17] Hsu et al..(2014),A GABased Approach for Resource Consolidation of Virtual Machines in Clouds,
● ACIIDS ,342–351

- [18] SOAP, HTTP://WWW.W3.ORG/TR/SOAP
- [19] Cheriet et al..(2014),Life cycle assessment of videoconferencing with call management servers
- relying on virtualization, ICT4S
- [20] Virtualbox, HTTP://WWW.VIRTUALBOX.ORG/MANUAL
- [21] http://stackoverflow.com/questions/19310293/is-hadoop-the-only-framework-in-big-data-space
- [22]http://www.ittoday.info/ITPerformanceImprovement/Articles/2014-07Raghupathi.html
- [23] http://www.sas.com/en_us/insights/analytics/machine-learning.html
- [24] Jaseena, K.U & David, J.M. (2014). ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA
- MINING. Computer Science & Information Technology, 131–140.
- [25] Gupta, S & Chaudhari, M.S. (2015). Big Data Issues and Challenges Data analysis, storing,
- processing, issues, challenges and future scope. International Journal on Recent and Innovation
- Trends in Computing and Communication, 3(2), 062– 067.

[26] Subramaniyaswamy,V., et al.(2015) 2nd International Symposium on Big Data and Cloud Computing, "Unstructured Data Analysis on Big Data using Map Reduce" , Procedia Computer Science 50,456 – 465

[27]Mohebi, A., et al.(2015) Published online in Wiley Online Library, "Iterative big data clustering Algorithms: a review", DOI: 10.1002/spe.2341

[28] Mishra, S. and Badhe, V. 2016 International Journal of Engineering and Computer Science, "Improved Map Reduce K Mean Clustering Algorithm for Hadoop Architecture", Volume 5, Issues 7, 17144-17147

[29] Dean, J. and Ghemawat, S. 2008 COMMUNICATIONS OF THE ACM, "MapReduce: Simplified Data Processing on Large Clusters", Vol. 51, No. 1

[30] P. and Chwala, R. 2017 International Journal of Scientific Research and Management, "Optimized Map Reduce Based Shuffling Mechanism for Density Clustering", Volume 5, Issues 8, 6771-6776

[31] https://www.tutorialspoint.com/map_reduce/map_reduce_api.htm

[32] https://intellipaat.com/blog/what-is-mapreduce/

[33]Vishnupriya, N. and Francis,F.S. 2015 International Advanced Research Journal in Science, Engineering and Technology, "Data Clustering using MapReduce for Multidimensional Datasets", Vol. 2, Issues 8

[34] Akhtar, C. G. and Pasha, S. B.2016, International Journal of Advanced Research in Computer
Science and Software Engineering," Hadoop Map Reduce Auditing Mechanisms" Website:
www.ijarcsse.com, Volume 6, Issue 4
[35] Vu, T-T. And Huet, F.2013 IEEE/ACM International Symposium on Cluster, Cloud, and Grid
Computing, CCGrid, "A Lightweight Continuous Jobs Mechanism for MapReduce Frameworks",
Pp.269-276, https://hal.archives-ouvertes.fr/hal-00916103
[36] https://www.edureka.co/blog/mapreduce-tutorial/
[37] Stonebreaker, M., et al. (2010) communications of the acm "MapReduce and Parallel DBMS's:
Friends Or Foes?" Vol. 53,No 1
[38] Dean, J. and Ghemawat, S. 2004 Google, Inc.," MapReduce: Simplified Data Processing on Large
Clusters"
[39] Chang, F., et al. (2006) Google, Inc "Big table: A Distributed Storage System for Structured Data"
[40] Lee, R., et al.,"YSmart: Yet Another SQL-to-MapReduce Translator", Facebook Data
Infrastructure Team
[41] Chambers,C.,et al.(2010) Google, Inc "FlumeJava: Easy, Efficient Data-Parallel Pipelines"
[42] Archenaa,J. and Anita, E.A.Mary 2015, 2nd International Symposium on Big Data and Cloud
Computing " A Survey of Big Data Analytics in Healthcare and Government", Procedia Computer
Science 50,408 – 413
[43] Ghazi,M.R. and Gangodkar, D. 2015 International Conference on Intelligent Compute
Communication& Convergence, "Hadoop, Map Reduce and HDFS: A Developers Perspective",
Procedia Computer Science 48, 45-50
[44] Maitrey, S. and Jha, C.K. 2015 3rd International Conference on Recent Trends in Computing,
"MapReduce: Simplified Data Analysis of Big Data", Procedia Computer Science 57, 563-571