IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. III (May-Jun. 2016), PP 12-17
www.iosrjournals.org

Distributed Heterogeneous Big Data Mining Adaptation in the Cloud: a Survey

Jayaraj T[1], Dr. S. Muruganantham[2]
[1] *(Research and Development Centre, Bharathiar University, Coimbatore, India)*
[2] *(ST Hindu College, Nagercoil, India.)*

Abstract: *Big Data Mining in the Cloud is the capability of integrating and extracting useful information from large datasets or streams of data in the on demand computing environment which were not possible before due to its volume, [variety], velocity, veracity and value. It includes three important fields such as Cloud, Big Data and Mining and makes Big Data's Knowledge Integration and Extraction from distributed heterogeneous Environment in the Cloud. The Cloud based Hadoop with MapReduce analyses Big Data in the Cloud. This paper presents study of related work and concepts relating to distributed heterogeneous Big Data Mining to the Cloud technology.*

Keywords: *Big Data, Big Data Mining, Cloud technology, Hadoop with Mapreduce, Distributed Heterogeneous Big Data.*

I. Introduction

In the Big Data era, with the proliferation of social media, sensors, websites, and mobile devices, the huge amount of data is generated with a wide variety of data types [1]. As the size of Big Data is growing, the demand to powerfully store, process and analyze this large amount of data to make it usable is also growing rapidly [2]. With the help of Big Data analysis, the researchers, business users and market analysts can easily gain the insights from the available data which resulting in numerous business advantages [3] [4]. Hadoop is the most accepted technology for Big Data analysis since it provides a reliable, flexible, economical, and scalable solution. It exploits the Hadoop Distributed File System (HDFS) [5] to store the large volume of data and perform the analysis using MapReduce with parallel processing [6]. MapReduce is rising as a key programming model for large-scale data-parallel applications. [7].

While there is no doubt that the Big Data analysis using Hadoop has created substantial benefits to businesses and consumers alike, there is a commensurate risk that goes along with using Big Data. The significant risk is that it requires a large computational infrastructure to ensure successful data management and data processing. Cloud computing is the type of computing follows the concept of Service-Oriented Architecture and which is designed to overcome many distributed organization computing challenges [8] [9]. Cloud computing is make use of of NoSQL databases and adopt a non-relational model for data storage [10]. The NoSQL database is used for providing scalable data analytics [11]. The Cloud-based Hadoop overcomes the infrastructure issues, and MapReduce analyzes the Big Data without maintaining the expensive computing hardware, software and dedicated space [12]. Hence, the Big Data analysis using Cloud-based Hadoop drives the more cost-effective, consumer-driven and agnostic technology solutions.

Despite this benefit, adoption of Hadoop in Cloud reduces the performance of MapReduce due to its homogeneous assumption in task scheduling [13]. The performance of MapReduce depends on the task scheduling and resource allocation but, the current scheduling approach performs poorly in a heterogeneous environment [14]. This is because the MapReduce with Hadoop assumes that all nodes in the cluster have the same processing power and capability and hence, it equally allocates the task to every node. This assumption cannot be held in a Cloud environment, and hence it leads to decrease the data locality and fault tolerance capability of MapReduce as it waits for the low-processing node. Hence, with the help of suitable resource provisioning techniques and considering heterogeneous job scheduling algorithms, the efficient Big Data analysis can be enabled in the Cloud-hosted MapReduce.

II. Distributed & Heterogeneous Big Data

Big data refers to huge, heterogeneous, distributed and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term encompasses the complexity and range of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics. Big data also has new sources, like machine generation (e.g., log files or sensor networks), mobile devices (video, photographs, and text messaging), and machine-to-machine. Characteristics of big data are volume, variety, velocity, veracity and value. *Volume, It* estimates that 2,500,000,000,000,000,000 bytes of data are created now each day. V*elocity*, increasing data rates because of network bandwidth. V*ariety*,

additional unstructured data types. V*eracity* of data, large set of decisions and analysis. Finally, the Value can be extracted and analyzed for useful data findings.

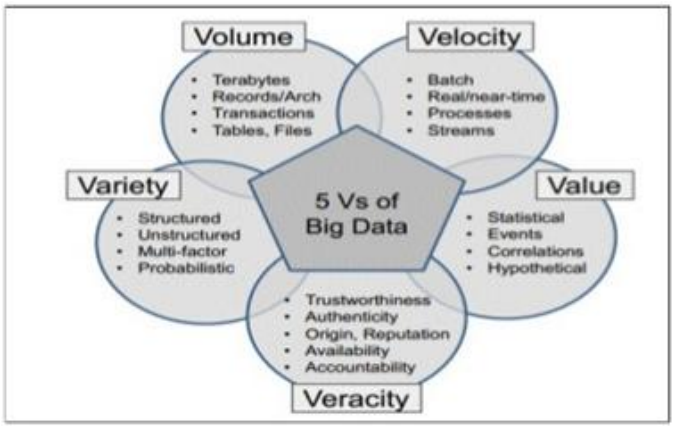

Figure 1: Characteristics of Big Data

When compared to relational database records big data is less structured. Highly developed data mining techniques and associated tools can help take out information from huge, difficult datasets and create new insights in big data mining techniques in a limited time [15]. When dealing out a query in big data, speed is an important claim [16]. However, the process may take time because mostly it cannot traverse all the related data in the whole database in a short time. In this case, index will be an optimal choice. At present, indices in big data are only aiming at simple type of data, while big data is becoming more complicated. The combination of appropriate index for big data and up-to-date preprocessing technology will be a desirable solution when we encountered this kind of problems.

There is several application paradigms used for big data problems. The traditional serial algorithm is inefficient for the big data. Cloud's reduced cost model to use hundreds of computers for a short time costs of the big data application. By using online big data application, a lot of companies can greatly reduce their IT cost. However, security and privacy affect the entire big data storage and processing, since there is a massive use of third-party services and infrastructures that are used to host important data or to carry out critical operations. The scale of data and applications grow exponentially, and bring huge challenges of dynamic data monitoring and security protection. Unlike traditional security method, security in big data is mainly in the form of how to process data mining without exposing sensitive information of users. Besides, current technologies of privacy protection are mainly based on static data set, while data is always dynamically changed, including data pattern, variation of attribute and addition of new data. Thus, it is a challenge to implement effective privacy protection in this complex circumstance.

III. Big Data Mining Adaptation to the Cloud

Hadoop is the preferred choice to handle and analyze these large volumes of data, since it manages the structured and unstructured data using HDFS and process the data in a parallel manner using MapReduce. However, the real world Big Data applications require a large computational infrastructure for successful data analysis. The rise of Cloud based Hadoop assists to achieve this requirement by providing the resources dynamically according to the user demands. However, the adaptation of Hadoop in Cloud decreases the performance of MapReduce, as it provides the homogeneous task assignment. The numerous existing approaches have been proposed and applied to a Cloud environment to solve these problems. However, the various issues, including the current capability of Cloud technologies to process the massive amount of data hinder the successful implementation of effective solutions. The effective resource provisioning and job scheduling algorithms with considering the homogeneous environment can increase MapReduce performance in the Cloud environment.

Cloud-based Hadoop processing in Big Data Analysis: The first part formulates the process of Big Data integration and knowledge extraction of heterogeneous data on homogeneous environment. The second part formulates the Big Data process on the heterogeneous Cloud environment.

Cloud Assisted Big Data Integration and Knowledge Extraction

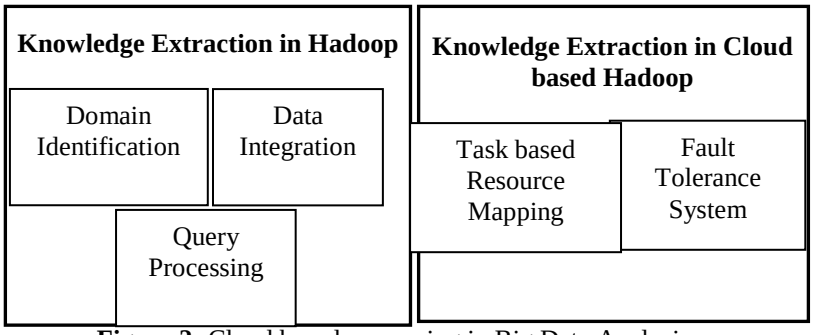

Figure 2: Cloud based processing in Big Data Analysis

The trend in Cloud computing is the rising use of NoSQL databases as the chosen technique for store and retrieve information. NoSQL implement a non-relational model for data storage. Non-relational models used more than 60 years in the form of object-oriented, hierarchical, and relational databases, but this new technology uses column centric document oriented methods [17]. The causes for such raise in interest are better performance, capacity of managing unstructured data, and suitability for distributed environments. NoSQL databases with importance on their advantages and limitations for Cloud computing. The big differences among the features of different technologies, and there is no single system that would be the most appropriate for every need. Therefore, it is significant for adopters to understand the requirements of their applications and the capabilities of different systems so that the system whose features better match their requests is selected. Cloud technology uses a service oriented work flow based data mining concepts in distributed execution that reduces data analytics end time. Application developers can design data analysis tasks, scientific computation methods, and complex simulation techniques as workflows that incorporate single Web services and execute them concurrently on virtual machines in the cloud.

IV. Related Works

a. Distributed heterogeneous Environment with Big Data Mining

The conventional technique exploits the data warehouse to integrate the various source data. The huge amount of data tables is extracted using the technique Extract, Transform Load (ETL) and their relationship are identified by the foreign key. The data extraction is done using the extended SQL. However, it is not an adaptable solution for Big Data as it is not capable of supporting the Big Data characteristics [18]. Due to the data warehouse limitation there is a need to find the new approaches to effectively integrate and extract the information from huge amount of structured and unstructured data. The ONDINE approach initializes the ontological resources to integrate the heterogeneous data. The approach semantically annotates the data tables and the outputs are represented in an RDF format. Finally, the SPARQL queries are exploited to extract the needed information from the annotated RDF. This approach focuses only on the data extracted from the web and does not provide any information about the Big Data integration from heterogeneous sources [19]. Similarly, the approach analyses the Big Data by integrating the Linked Data which is an interesting potential external source. However, it also concentrates son logistics domain for supply chain management and does not deal with the multi-domain Big Data [20].

Google uses distributed files in the form of GFS [21], it forms huge files that supports fault-tolerance by data partitioning and duplication. Google's cloud computing platform also used to read input and store output of MapReduce [22]. Hadoop uses a distributed file system as its data storage layer called HDFS [23]. Now a days more and more IT companies have rising needs to store and analyze the ever increasing data, such as logs, crawled web content, and click streams, usually in the variety of pita bytes, composed from a range of web application and services. Moreover, simple distributed file systems mentioned above cannot assure service providers like Google, Yahoo!, Microsoft and Amazon. Service providers have their purpose to provide possible users to big data management in the cloud systems. Big table refers very large size of distributed structured data not having full relational data model with thousand of commodity servers [24]. PNUTS [25] is a database designed to sustain Yahoo!'s web applications of complex queries. The Dynamo [26] is a highly available and scalable distributed key/value based data store built for supporting internal Amazon's applications. It provides a simple primary-key only interface to meet the requirements of these applications. However, it differs from key-value storage system. Face book proposed the design of a new cluster-based data warehouse system, Llama[27], a hybrid data management system which combines the features of row-wise and column-wise database systems. Hadoop new column-wise format CFile provides better performance in data analysis.

Intelligent Data Understanding System (INDUS) is the data integration component to flexibly integrate the information from heterogeneous, distributed, autonomous information sources. It integrates the data from multiple sources with multiple domains using the ontology and stores the information in relation database based on the ontology. Moreover, the integrated information can be accessed using application programs or relational database operations. Although, it integrates the various sources multi domain data, it does not deal with the unstructured data [28]. In Big Data 80% of information are presented in unstructured formats, hence it is not applicable for Big Data. Similarly, the Data Representation and Integration Framework (DRIF) approach semantically enrich the data and provides a common framework to encapsulate entire structure data from heterogeneous sources into an RDF representation. However, it also concentrates mainly on structured data [29]. The Multi-Intelligence Data Integration Services (MIDIS) is a domain ontology based data integration approach which semantically annotates both structured and unstructured Big Data. However, the domain ontology comprises set of concepts related to a particular domain and hence, it is suitable for data integration and extraction with heterogeneous multi domain data [30]. Cura is the first effort that is aimed at developing a novel usage model and resource management techniques for achieving global resource optimization in the cloud for MapReduce services [31]. From the existing approaches it is clear that there is a need for effective approach for integrating the heterogeneous data from various domains to extract the useful knowledge.

b. Adaptation of Cloud based Hadoop for Big Data analysis

Big Data and cloud integration will invariably intersect as the market evolves [32]. For cloud-based big data analytics, several frameworks like Google MapReduce, Spark, Hadoop, Twister, Hadoop Reduce and Hadoop++ are available [33]. MapReduce is a scalable and fault-tolerant data processing tool helps to perform the Big Data analysis in a parallel and timely manner over a massive amount of data. MapReduce is referred to as a latest advance of processing big data in cloud computing environments, it is also criticized as a "major step backwards" compared with DBMS [34]. However, adopting MapReduce in the Cloud has inherent limitations on its performance and efficiency. The following are the existing approaches tried to increase the MapReduce performance on Cloud infrastructure. With the intention of eliminating the data locality problem in Hadoop MapReduce while adapting in the Cloud the data placement mechanism is developed in [35]. According to this mechanism, the data set is allocated to the heterogeneous node according to its capacity. However, the approach does not consider the data replication. Moreover, hence, if a node crash, then data will be lost. The data locality is modeled as a linear sum assignment problem, and it is evaluated in different running scenario to increase the performance of MapReduce [36]. Initially, it analyzes the importance of data locality in MapReduce cluster performance and introduces an algorithm called lsapsched by mathematically reformulating the scheduling problem. To reduce the data locality issue the reduce task is scheduled on the node with the maximum portion of input for that reduce task. However, the reducer has waited until all the maps have to complete their job to the output of map [37].

In the context of Big Data analytics, Hadoop, an open source MapReduce implementation, allows for the creation of clusters that use the Hadoop Distributed File System (HDFS). In addition to exploiting concurrency of huge number of nodes, HDFS reduce the failures of number of nodes. It has been used by Thusoo [38] to develop an analytics platform to process Face book's large data sets. Among the drawbacks of Cloud storage techniques and MapReduce implementations, there is the fact that they need the customer to learn a new set of APIs to build analytics solutions for the Cloud. There are plenty of solutions for Big Data related to Cloud computing. Analytics can be descriptive, predictive, and prescriptive; Big Data can have various levels of variety, velocity, volume, veracity and value. Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools. Although Cloud infrastructure offers such elastic facility to supply computational resources on demand, the area of Cloud supported analytics is still in its early days.

Longest Approximate Time to End(LATE) scheduling is a scheduling algorithm tried to increase the response time of short jobs by executing the duplicates of tasks. It analyzes the current speculative execution scheduler in the Hadoop MapReduce and how its homogeneous assumption affects the MapReduce performance on Cloud environment. However, due to the static manner computation of progress of nodes, it obtains only poor performance [39]. Self-adaptive MapReduce Scheduling Algorithm (SAMR) is an extended version of LATE and able to compute the node progress in dynamically. SAMR identifies the slow tasks dynamically with the help of historical information, and it dynamically updates the execution value. However, it does not take into account the fact of different types of jobs are differ in their map and reduce stage weights [40].

Energy-aware MapReduce scheduling algorithms (EMRSA-I and EMRSA-II) identify the works of map and reduce tasks to the machine slots for minimizing the energy consumption of executing each MapReduce job. Both the algorithms are suitable in real world Big Data application as they provide fast execution of data. However, this approach is effective only for single MapReduce job [41]. The cost-aware and SLA based algorithms are developed for providing the Cloud resources and schedule MapReduce tasks by

taking the budget and deadline as constraints [42]. Moreover, the admission control and cost-effective resource scheduling algorithm are designed to provide the analytic solution with the defined SLA guarantees [43]. With the intention of providing cost-effective solution, the Cura multiples the existing Cloud resources according to the requirements of the job. Moreover, it implements an efficient resource allocation scheme. And leverages the MapReduce to build the suitable cluster configuration and unique optimization to deal with workloads for achieving the goal. However, it lacks in scalable scheduling algorithms [44]. The existing approaches indicate that the effective job scheduling and resource provisioning techniques can improve the MapReduce performance in the Cloud environment.

V. Significance and Applications of Big Data Adaptation in Cloud

With the help of Cloud computing the Big Data can be effectively collected, stored, shared, and transferred at very high speed. The Cloud computing infrastructure has served as an efficient to address the data storage necessary to perform data analysis. Cloud computing allocates the services dynamically to the Big Data analysis, according to the user needs via virtual resources and the Internet. Hence, it reduces the costs of hardware and the application software also no longer installed on a local machine. The Big Data analysis makes use of parallel data processing in a distributed environment with the help of Cloud computing thus enables the massive data analysis with affordable cost and reasonable time.

With the assistance of cloud computing infrastructure, the businesses and users can access the application services from anywhere in the world. Big Data and Cloud Computing together, allow the data analysts and decision makers to discover new insights for Intelligence analysis and allow the business users to get the answers to their ad-hoc analysis questions faster with higher precision. The Big Data analysis, such as Fraud Detection, Ad serving, Log analysis, Web Crawling and Recommendation Engine achieves better performance results with the help of Cloud infrastructure. Big Data technologies along with the Cloud help the organizations to differentiate through productivity and efficiency improvement.

VI. Conclusion

Big Data Mining is still a challenging and time demanding job that requires expensive software, large computational infrastructure, and effort. Cloud assisted big data Analysis helps in alleviating these problems by providing resources on-demand with costs proportional to the actual usage in the digital world. It analyzed the possibilities of handling cloud based Hadoop with MapReduce in the Cloud assisted Big Data Integration and Knowledge Extraction. This paper analyzed the related work, identifies the challenges and benefits in adopting Hadoop to the Cloud for the heterogeneous data integration and knowledge extraction with Big Data.

References

[1]. "Analytics: Real-world use of big data in telecommunications", IBM Report, 2013.
[2]. Labrinidis, Alexandros, and H.V. Jagadish, "Challenges and opportunities with big data", Proceedings of the VLDB Endowment, Vol.5, No.12, pp.2032-2033, 2012.
[3]. Big Data Analytics, Advanced Analytics in Oracle Database, An Oracle White Paper, 2013.
[4]. Big Data Meets Big Data Analytics, www.sas.com/offices, 2012.
[5]. Shvachko, K., Kuang, H., et al, "The Hadoop distributed file system. In Mass Storage Systems and Technologies (MSST)", 2010, IEEE, pp.1-10.
[6]. Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, Vol.51, No.1, pp.107-113, 2008.
[7]. Zaharia, Matei, et al, "Improving MapReduce performance in heterogeneous environments", Proceedings of the 8th USENIX conference on Operating systems design and implementation, USENIX Association, pp.29-42, 2008.
[8]. M. Armbrust, A. Fox, R. Griffith, et al., "A view of Cloud computing", Commun. ACM, vol. 53, pp. 50-58, 2010.
[9]. Hauck, Michael, et al. "Challenges and opportunities of cloud computing", Karlsruhe Reports in Informatics, Vol.19, 2010.
[10]. Marcos D. Assunção , Rodrigo N. Calheiros , Silvia Bianchi , Marco A.S. Netto, Rajkumar
[11]. Buyya "Big Data computing and clouds: Trends and future directions" Journal of Parallel
[12]. and Distributed Computing 79-80 (2015) 3-15.
[13]. Mahdi Bohlouli, Fabian Merges, MadjidFathi, "Knowledge Integration of Distributed
[14]. Exterprises using Cloud based Big Data Analytics", In Proc. International Conference on
[15]. Electro/Information Technology 2014 conference.
[16]. Liu, Huan, "Big data drives Cloud adoption in enterprise", IEEE Internet computing, Vol.4, pp.68-71, 2013.
[17]. Fernández, Alberto, et al, "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol., No.5, pp.380-409, 2014.
[18]. "Fair Scheduler," hadoop.apache.org/docs/r1.0.4/fair scheduler.html
[19]. Big data in the cloud - Data velocity, volume,variety and veracity ARTICLE · JULY 2013
[20]. READS 254 1 AUTHOR: Sam Siewert Embry-Riddle Aeronautical University
[21]. X. Zhou, J. Lu, C. Li, and X. Du, "Big data challenge in the management perspective,"
[22]. *Communications of the CCF*, vol. 8, pp. 16–20, 2012.
[23]. N. Leavitt, Will NoSQL Databases Live Up to Their Promise? Computer 43 (2) (2010) 12–14.
[24]. Chaudhuri, Surajit, and Umeshwar Dayal. "An overview of data warehousing and OLAP technology." ACM Sigmod record, Vol.26, No.1, pp.65-74, 1997.

[25]. Patrice Buche, Juliette Dibie-Barthelemy, Liliana Ibanescu, and Lydie Soler, "Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource", IEEE Transactions on Knowledge and Data Engineering, Vol.25, No.4, 2013.
[26]. Silva Robak, Bogdan Franczyk, Marcin Robak , "Applying Big Data and Linked Data Concepts in Supply Chains Management" Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, pp.1203–1209, 2013.
[27]. S. Ghemawat, H. Gobioff, and S. Leung, "The google file system," in *ACM SIGOPS Operating*
[28]. *Systems Review*, vol. 37, no. 5. ACM, 2003, pp. 29–43.
[29]. J. Dean and S. Ghemawat, "Mapreduce: simplified dat aprocessing on large clusters,"
[30]. *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[31]. D. Borthakur, "The hadoop distributed file system: Architecture and design," *Hadoop*
[32]. *Project Website*, vol. 11, 2007.
[33]. Douglas and Laney, "The importance of 'big data': A definition," 2008.
[34]. B. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H. Jacobsen, N. Puz,
[35]. D. Weaver, and R. Yerneni, "Pnuts: Yahoo!'s hosted data serving platform," *Proceedings of*
[36]. *the VLDB Endowment*, vol. 1, no. 2, pp. 1277–1288, 2008.
[37]. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin,
[38]. S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available
[39]. keyvalue store," in *ACM SIGOPS Operating Systems Review*, nvol. 41, no. 6. ACM, 2007,
[40]. pp. 205–220.
[41]. Y. Lin, D. Agrawal, C. Chen, B. Ooi, and S. Wu, "Llama: leveraging columnar storage for
[42]. scalable join processing in the mapreduce framework," in *Proceedings of the 2011*
[43]. *international conference on Management of data*. ACM, 2011, pp. 961–972.
[44]. Castillo, J. A. R., Silvescu, et al., "Information extraction and integration from heterogeneous, distributed, autonomous information sources-a federated ontology-driven query-centric approach", Proceedings IEEE International Conference on Information Reuse and Integration, pp.183-191, IEEE, 2003.
[45]. Salmen David, Malyuta Tatiana, Hansen Alan, Cronen Shaun, Smith Barry, "Integration of Intelligence Data through Semantic Enhancement", 2011.
[46]. Boury-Brisset, Anne-Claire, "Managing Semantic Big Data for Intelligence", In STIDS, pp.41-47, 2013.
[47]. Palanisamy, B., Singh, A., & Liu, L., "Cost-effective resource provisioning for mapreduce in a Cloud", IEEE Transactions on Parallel and Distributed Systems, Vol. 26, No.5, pp.1265-1279, 2015.
[48]. WHITE PAPER on "Data Integration Platforms for Big Data and the Enterprise" Customer
[49]. Perspectives on IBM, Informatica, and Oracle April 2015
[50]. Samiya Khan1, Kashish Ara Shakil and Mansaf Alam "CLOUD-BASED BIG DATA ANALYTICS –
[51]. A SURVEY OF CURRENT RESEARCH AND FUTURE DIRECTIONS."
[52]. Changqing, Uchechukwu Awada, Keqiu Li, "Big Data Processing in Cloud computing
[53]. environments" , Proceedings of the international symposium on parallel architectures,
[54]. algorithms and networks,I-span Dec2012.
[55]. Xie, Jiong, et al., "Improving mapreduce performance through data placement in heterogeneous Hadoop clusters", Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on. IEEE, pp.1-9, 2010.
[56]. Guo, Z., Fox, G., & Zhou, M, "Investigation of data locality and fairness in MapReduce", In Proceedings of third international workshop on MapReduce and its Applications, pp. 25-32, ACM, 2012.
[57]. M. Hammoud and M. F. Sakr, "Locality-aware reduce task scheduling for Mapreduce," in Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science, ser. CloudCOM '11. Washington, DC, USA: IEEE Computer Society, pp.570–576, 2011.
[58]. A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J.S. Sarma, R. Murthy, H. Liu, Data
[59]. warehousing and analytics infrastructure at Facebook, in: Proceedings of the 2010
[60]. International Conference on Management of Data, ACM, New York, NY, USA, 2010, pp.
[61]. 1013–1020.
[62]. Zaharia, Matei, et al, "Improving MapReduce performance in heterogeneous environments", Proceedings of the 8th USENIX conference on Operating systems design and implementation, USENIX Association, pp.29-42, 2008.
[63]. Chen, Q., Zhang, D., Guo, M., et al., "Samr: A self-adaptive mapreduce scheduling algorithm in heterogeneous environment", In Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, pp. 2736-2743, IEEE, 2010.
[64]. Mashayekhy, Lena, et al. "Energy-aware scheduling of mapreduce jobs for big data applications.", Vol.25, No.10, 2014.
[65]. Alrokayan, M., Vahid Dastjerdi, A., Buyya, R., "SLA-Aware Provisioning and Scheduling of Cloud Resources for Big Data Analytics", Proceedings IEEE International Conference of Cloud Computing in Emerging Markets (CCEM), pp.1-8, 2014 .
[66]. Yali Zhao, Rodrigo N. Calheiros, Graeme Gange, and et al., "SLA-Based Resource Scheduling for Big Data Analytics as a Service in Cloud Computing Environments", 2015.
[67]. Palanisamy, B., Singh, A., & Liu, L., "Cost-effective resource provisioning for mapreduce in a Cloud", IEEE Transactions on Parallel and Distributed Systems, Vol. 26, No.5, pp.1265-1279, 2015