Australian Journal of Basic and Applied Sciences, 9(11) May 2015, Pages: 617-623

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Cloud based Analytical Study on Big Data using Hadoop Tools

Saranya R and Saravanan. M.S

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Chennai- 602105, India

ARTICLE INFO

Article history:
Received 12 March 2015
Accepted 28 April 2015
Available online 1 June 2015

Keywords:
Data Storage, Hadoop tool, Web Services, Data Center and Cloud Infrastructure

ABSTRACT

Cloud computing has a remarkable impact on demand, resilient computing and data storage resources, without much cost to invest for organizing habitual data centers. Cloud computing has become a viable, mainstream solution for data processing, storage and distribution. Adoption accelerates because of using various web services from 262 billion objects stored in its S3 cloud storage in 2010, to over 1 trillion in 2015. However, companies that work with big data have been unable to realize the full potential of the cloud, due to the inherent bottlenecks of moving big data in, out and across cloud infrastructures. The big data analytics in cloud infrastructure can be deployed using Hadoop Clusters, HIVE, Sci DB and Earth DB1 tool. This article will analyze the Hadoop tool with the existing tool. Hence the state of the art of the Hadoop toolfor big data analytics in cloud storage was studied and finally observations are given to understand to implement in the field of big data analytics in future.

To Cite This Article: Saranya R and Saravanan. M.S, Cloud based Analytical Study on Big Data using Hadoop Tools. ***Aust. J. Basic & Appl. Sci.,*** *9(11): 617-623, 2015*

INTRODUCTION

Big data is a central section of the cloud and provides unique opportunities for using conventional as well as ordered database information and analytics for business applications with social networking, sensor network data, earth science and chemical science and far less structured multimedia. Big data is generally defined as the seizure, managing, and exploration of data that goes further than classic structured data, which may be queried by relational database management systems, often to unstructured files. The big data can be analyzed using the volume, variety and velocity. The volumes of data are superior to those conventional database set-ups can get, by meting out choices of break down generally into a choice between vastly parallel processing architectures and databases. Hadoop is a podium for distributing the data through a number of servers with Map Reduce approach. Hadoop Distributed File System (HDFS), used by Hadoop to store data and makes it open to number of computing nodes. The prominence of velocity is the increasing rate at which fast moving data flows in and out of an organization. Hence the big data can be handled and analyzed using cloud storage. The Hadoop tool is used to handle the big data efficiently in the cloud storage. This article gives a complete study on the state of the art of big data on cloud storage using Hadoop tool.

Related work:

The storage of huge amount of data in cloud must have a high transfer speed to and fro, and also between cloud infrastructures. The big data analytics has lot of success stories in the field of semantic computing, earth science, chemical science and high performance computing, etc. The traditional WAN based transport methods cannot move terabytes of data at the speed dictated by businesses. Hence to achieve better transfer speed that are unsuitable for such volumes, introducing unacceptable delays in moving data into and out of and within the cloud. Therefore state of the art of big data in cloud storage has given in this article. The following sub sections deals about the challenges in big data, big data in cloud computing and performance of big data in cloud storage.

Challenges in Big Data

Parallel processing is the technique used in big data practice. Large scale dispensation provides a lot of difficulties in implementing. Instead of putting effort in improving parallel processing, more concentration has to be provided to address the existing problems and enhance parallelism. After going through many optimization techniques, we found that there is more need for optimized code for multiprocessors. Enhanced optimization techniques for big data processing can be enhanced by making the most of multi core dynamic optimization into a wide

Corresponding Author: M.C. Agarana, Department of Mathematics, College of Science & Technology, Covenant University, P.M.B. 1023, Ota, Nigeria.
Tel: +2348023214236; E-mail: michael.agarana@covenantuniversity.edu.ng

Small Molecule Interaction Database (SMID)(Arun Kejariwal, 2012).

In recent trends, the usage of data has become large in all the fields like business, social networks and mobile applications. For handling this huge data, the analysis also needed to be done in large scale. The most important technique that was introduced is parallel database and parallel computing framework with their implementation as Hadoop. But this framework is incompetent in the query processing for the real world environment. The proposed system frontwards a real time multi-dimensional query mode with insertion in dynamic manner. It makes use of Z curve to map multi-dimensional data to one dimension and bloom filter for reducing storage space and lookup time needed. The indexing of data in this environment is carried out by the DC – Tree (dynamic indexed) for creating a real time model of query. In this model, the node which is responsible for the data query & query for update location called as master node and the data node to store the data with dynamic indexing(Dan Wei Chen, 2013).

Traffic information system may create substantial crisis in addition to difficult targeted visitors data with all the means of obtaining real time initial GPS UNIT (Global Placement System) data, complementing positions with a guide in addition to making targeted visitors circulation information, which produces great market segments intended for even-worse targeted visitors issue with The far east. However, various problems would come when we reuse these types of substantial targeted visitors data intended for historical past data mining making use of on-hand repository supervision instruments or perhaps standard data finalizing approaches, including substantial storage devices, large functionality finalizing, open program. “Big Data” system usually incorporates data packages with sizes past the capacity associated with commonly-used software program instruments to help catch, handle, in addition to process the results inside a bearable passed moment. With this particular difficulty in addition to the benefit of “Big Data”, the RTIC-C system to help take care of sense making around substantial levels associated with targeted visitors data primarily based in cloud processing technique. RTIC-C models a new spread data supervision program to back up substantial degree associated with data storage devices; a new parallel spread processing composition intended for different types associated with mining programs depending on Map-Reduce process; a new good World Wide Web solutions program to back up third-party mining programs. Findings with a substantial targeted visitor’s data packages confirmed that RTIC-C accomplishes extensive functionality comparing with standard targeted visitors data mining programs(Jianjun, 2013).

Now a day’s the level of data processed is in a higher extent. Analyzing large data sets is more difficult to perform. The simplest solution to this process is divide them into simple tasks and assigned to different components. The best part of this solution is achieved by re usage of components in various pipelines. The performance of this method is affected because of poor interactions between components of different pipelines. The proposed framework named as probabilistic pipeline model which consists of graphical model for pipeline inference & inference algorithm for resolving uncertainty in the pipeline. The two algorithms used here are canonical inference and full Bayesian inference. But a trade-off of efficiency vs. accuracy occurs between these two models, Top K inference, Fixed Beam inference and Adaptive inference provides solution to these problems(Karthik Raman, 2013).

Big Data in Cloud Computing

Cloud Computing have greater influences on today’s IT methods. The basic revelation provided by cloud is avoiding the necessity of building huge infrastructure for maintaining the big sized data and the person is needed to maintain. Just one of the principal hurdles involving cloud processing will be of which not only the software, but in addition your data should be relocated towards the cloud (Changqing, 2012). Network pace drastically boundaries how much info that could be travelling between the cloud and also the person, concerning different sites on the same cloud service, or even without a doubt concerning different cloud providers. Therefore, it's imperative that you hold applications near at the information themselves. That document investigates where approach heap controlling involving the computational sources and also the info surrounding area can certainly end up being managed at the same time. The β balancedness balls-into-bins idea answers the balancing load problem. Balls define jobs and the bins define cloud and the idea was defining various definitions for neighborhood to make the ball to choose the bin which is near to it and used in various type of network model for checking their correctness. But this model provide inefficient way for internet working, which was surmount by creating overlay of networks which avoided the pressure in most of the overloaded nodes (Petra Berenbrink, 2013).

In the Cloud age, the IT professional are in urge to enrich their skills by acquiring knowledge about cloud skills. Informative businesses including universities have to offer informative Cloud curriculums with regards to learners. Conventional methods are used as starters for providing educational cloud but its high cost leads to super saturation. In super saturation the usage of logical resources was high so its show 10 percent of more running instances than conventional. Additionally the machine language skills are also expected to sound good. Mahout and Jubatus lead into a race, even Jubtaus shows well remark in real time processing than mahout, the nature of mingle with Hadoop make Mahout as a good choice. Virtual Machine Monitor(VMM) environment is formed by

making Mahout as machine learning tool based on Linux environment which is suited for host OS window environment also. The 90 minutes of processing contains the process of installation and environment variable setup for JDK, Maven, Hadoop and Mahout in the above order will create the VMM environment. With the help of SSH, an access to local host can be provided. The stack based (install mahout for start to run an instance) and image based (mahout has already installed in the images) are the two choices of environments. The few challenges of this VMM are difficult to make out within 90 minutes in classroom environments and if something moves wrong then need to restore the entire one from starting (Yuichiro Takabe, 2013).

With the fast improvement in sensor engineering along with Wi-Fi circle, exploration along with improvement regarding targeted traffic connected apps, such as real occasion targeted traffic map along with on-demand take a trip option endorsement get attracted considerably more attentions than ever prior to. Both are equally archived along with real-time information associated with these kinds of apps might end up being very large, based about how many started sensors. Growing Cloud national infrastructure may elastically handle this kind of large information along with handily supplying almost unrestricted processing along with safe-keeping sources in order to organized apps, to use investigation not merely for long-term arranging along with choice producing, but additionally analytics for close to real-time choice support. On this paper, we propose Clever Site visitors Cloud, a new application national infrastructure make it possible for targeted traffic information order, along with control, examines along with existing the results inside a versatile, scalable along with secure way using a Cloud podium. The actual offered national infrastructure grips sent out along with parallel information supervision along with investigation using ontology databases as well as the well-known Map-Reduce composition. We've got prototyped the national infrastructure inside a professional Cloud podium along with we designed a new real-time targeted traffic situation map using information compiled via commuters' mobiles (WenQiang Wang, 2012).

The cloud computing resources and big data are used for scientific computing and discovery. Using Hadoop cluster we evaluate VM types and workload types. Large scale data analysis requires large scale computing power by HPC clusters. In this we can use cutting- edge software for many research groups for the task of HPC clusters. Word count, sorting, K-means are the three important classes of MapReduce. In this cloud computing provides highly efficient software and high utilization of computing resources like HPC cloud. And we have to increase the performance of virtual machines for a Hadoop clusters. It helps to reduce the computational cost in the virtualized environment(Moussa Taifi, 2014).

The framework provides parallel processing and distributed data storage processing in set of raw data. In the Hadoop distributed file system needs cross domain services. Big data analytics having some security domains in MLS environment. Decisions are made based on level of the data. When the client request for the file, it can be retrieved from the associate blocks. The design and implementation of the frame work has a set of prototype designs. The CD Hadoop clusters having the physical data nodes. The primary data node is the instance in the main handler of HDFS blocks. WE can use Hadoop in MLS environment. The resources are constrained by the underlying trusted operating system(Thuy Nguyen, 2013).

Large amount of host and accelerator in heterogeneous clusters are used to transfer data in HPC. The system architecture allocate for data movement. GPU can provide additional computation power, GAS Net supports the partitioned global address space for private memory region. Oncilla-supported cluster combines the host memory and accelerator memory (GDDR) into a large. It implements to combine several hardware and software components. Local and remote accesses cannot be distinguished based on location, the replacement of GPU allocations with Oncilla-managed GPU allocations, the addition of two remote allocations using the Oncilla API, and the replacement of cudaMemcpy calls with calls to ocm_copy(Jeff Young, 2013).

Cloud computing shows its significance in all major fields of this era. Internet Service Provider (ISP) and Internet Content Provider (ICP) are the two major areas which offer various services for the clouds. In recent trends cloud shows its significance in mobile computing environment also. The important pro of cloud is the user can get the needed storage and power at all the time when connected with the cloud. The problem is not enough simulator environments for cloud processing. CloudSim is an open source simulator for cloud and also used as datacenters. The implementation is done in two steps: first step is doing modification in native cloudsim code and second step is to add the code of MapReduce model in cloudsim (Jongtack Jung and Hwangnam Kim, 2012).

Performance of Big Data on Cloud Storage:

The big data concept addresses the issue of maintaining a large collection of data which is not possible to manage using on hand database management tools. A concrete approach for processing these collections of data provides by cloud computing. Big data applications utilize clouds as private and public based on their notion of virtualization. Few applications need to inherit feature from both, cross cloud provides the requirement with limitations of privacy and time cost tradeoff. Existing HIRESOME – I (History record-based Service optimization method) provide cross cloud service but

failed to achieve promise QoS. The proposed method HIRESOME – II enhances existing by reduce time complexity using a service history record specifying transactions by its QoS values and privacy concern by not reviling information of all records (Wanchun Dou, 2013).

In recent years cloud storage offers service to various business organizations with reasonable cost. Even though lot of provisions such as parallel computing, control of access provide by cloud, efficacy and competent service of data remains as tailback for the performance. ADSC (Adaptive Data Service Coordinator) is the proposed method which follows the content sensitive transaction analysis and adoption. Data needed based queries to virtual machines are monitored and collected by ADSC. Fuzzy ART analyze this data to find the similarity, redundancy and then rearrange the sequence which will improve the performance (Chih-Wei, 2013).

Program recommender systems are actually shown since valuable resources for offering appropriate advice to consumers. In a final decade, the number of customers, services along with online information continues to grow rapidly, yielding the huge data research problem for service recommender techniques. Consequently, traditional program recommender techniques often suffer from scalability along with inefficiency complications when running or analyzing such large-scale facts. Moreover, nearly all of existing program recommender techniques present identical ratings along with rankings regarding services to different consumers without thinking about diverse users' choices, and therefore ceases to meet users' tailored requirements. In this paper, the proposed method uses Keyword-Aware Program Recommendation technique, named KASR, to handle the preceding challenges. It is aimed at preventing any personalized program recommendation checklist and recommending the most appropriate services on the users successfully. Specifically, keyword – candidate list stores keywords that are utilized to reveal users' choices and domain thesaurus mean to group and check similarity of keyword in possible manner. User-based Collaborative Blocking algorithm is actually adopted to get appropriate advice. To enhance its scalability along with efficiency within big facts environment, Hadoop a widely adopted techniques afford computing platform which using Map Reduce for its parallel running paradigm. Ultimately, extensive trials are done on real-world facts sets, along with results illustrate that KASR significantly improves the accuracy along with scalability regarding service recommender techniques over existing approaches(Shunmei Meng, 2013).

We used the Hybrid Infrastructure as a Service (HIaaS) technique in a cloud computing service model for crowd sourcing marketplace to Big data processing. Traditional process techniques cannot perform efficient data set. In crowd sourcing it can offer distributed problem solving environment are assigned. SaaS, PaaS and IaaS layers are found in the HIaaS service. Basically all the software components of the platform can be deployed as web services so that service invocations to the components. In HIaaS service the crowd sourcing marketplace for processing big data jobs. HRPO model based on stochastic programming primarily derived from the HIaaS, so minimum cost for HIaaS providers in Big data job (Sivadon Chaisiri, 2013).

Recently big data plays a vital role in scientific, engineering and business fields. Time is the important constraint in this analysis. In medical field the data needed to maintain was named as EHR (Electronics Health Records). The main constraints in maintaining EHR are unstructured temporal data, integration of scattered data for profound analysis and unclear temporal constraints. TIMER (Temporal Information Modeling, Extraction and Reasoning) framework answers theses constraints. TEO (Time Event Ontology) form basis of TIMER for model temporal information and to find whether structured data or unstructured one. For answering time related queries TIMER combine with DL-based and SWRL-based reasoning with SWRL Built-Ins library. Drawback of this framework is difficult to make it as a automatic information extraction tool. The solution to this drawback is TCTM (Temporal & Reference Model) annotation model i.e. any automatic annotation tool. Also addressed data sparseness, high dimensionality and easy to add newly attributes. Expectation Maximization based (EM) variational methods is adopted to answer joint inference (Dingcheng, 2012).

Mining a large volume of data in real time analysis, a tedious work. Clustering algorithm provides solution but with a problem of scalability if a big data needed to be mining. Latter parallel algorithm addressed this problem but need high computation time for few applications such as graph based applications. Thus PIC (Parallel Iterative Clustering) algorithm is proposed as new one but data have to fit in memory, not possible in all time. The new framework introduced parallelization in PIC named as Parallel- PIC which improves scalability and response time. p-PIC using MapReduce to answer the node failure but fault tolerant remains as questionable(Jayalatchumy, 2014).

Map Reduce paradigm is the mostly used techniques to deal with large amount of data such as big data. The Problem in mining process was the entire huge dataset needed to analyze each and every time of map reduce process. Using the intermediate result of the previous outcome of Map reduce offers partial satisfactory because some situations difficult to perform. The proposed solution named as Itchy offer efficient way to deal with the above problem. Optimizer used by Itchy automatically performs decision about right way i.e., not analyzing the entire huge sets. For recomputed the intermediate results Itchy automatically find provenance of intermediate results sometimes if needed (high cost, small results)

Itchy stores the intermediate works to avoid performing provenance. In Map reduce additionally Itchy make available of techniques needed to merging results from various jobs. Comparison of Hadoop and Incoop in various benchmark bases, Itchy shows some superiority in the results (Jorg Schad, 2013).

To support the demands of fast growing data, the scaling up in methodologies also needed. An application domain called Bio informatics which deals with the big challenge in now a day's environment. Genome sequencing is the core part which produced some megabytes for a individual genome. Genome comparison is the basic idea of genome sequencing. For performing this comparison using Hadoop, an application built on top of Hadoop for storing the results of comparison in HBase tables whether intermediate result or end result and BigTable methods of Google. GFS (Google File System) used as storage area for BigTable. For mapping, BigTable Map is used by combining row key, column key and timestamp. MapReduce for parallel execution. HDFS provides the open source implementation. A genome is sequence of four types of bases combined as a pairs (Adenin(A) pairing with thymin (T) and guanine (G) pairing with cytosine (C)). The genome comparison is a workflow in which the preprocessing the gene, comparing it and storing it in the HBase table and the similarity is denoted by dotplot generation [14].

Data processing on a cloud based cluster would provide added benefits such as fault tolerant, heterogeneous, ease of use, free and open, efficient, provide performance and "tool plug-ability" which most DBMS do not provide. Combining different types of software such as Mosaic of Antarctica(MOA) and Hadoop is a possible solution for online analytics of scientific data.

Discussion on various tools in big data and hadoop:

This section inculcates a deep discussion on various open source as well as proprietary versions of several kinds of framework and tools that are used currently to develop the applications depending on the concepts of big data.

The table holds the discussion purpose of usage of the particular tool and its potency.

S.NO	NAME OF THE TOOL	YEAR	PURPOSE OF THE TOOL	STRENGTH OF THE TOOL
1	SPARK	FEB 2014	In Memory Cluster Computing	100 times quicker than Hadoop Map Reduce
2	DREMEL	JAN 2004	Classes of software that involve extended hold up as the standards contain anti-virus suites and multiplayer online games	So, Dremel is a higher level of abstraction than MapReduce and it fits as part of an entire data segment.
3	APACHE DRILL		interactive study of huge datasets	Processes the petabytes of data and trillions of records in seconds
4	IMPALA	JUL 2014	distributed query execution engine	Performance, Cost savings, security
5	STORM	APR 2013	process abundant streams of data, especially for real-time processing	Can be used with any programming language
6	GRID GRAIN	AUG 2014	Offers with different interfaces, and classes such as listeners and adapters, to enable developers to gain access to the grid	written in Java, run anywhere benefits that come with Java and cost-effective
7	EC2	AUG 2006	Used for computation of complex applications	Offers flawless encryption and decryption techniques
8	MESOS	AUG 2014	Petabytes of data can be held, revise millions of rows of data per second and handle trillions of queries a day.	Keeps operational even if one of the data centers stop working.
9	GRASS	SEP 2006	To perk up GIS Supervised Classification Workflow	Offers Graphical User Interface and manages multispectral image data and stores spatial data.
10	AMAZON – S3	MAR 2006	For Storage	S3 affords 99.999999999% durability and 99.99% availability of things

Observation:

In day to day data can involve with huge number of applications to maintain and to make decision for the real time competitive world in business and research. The data being used for the massive manipulation for mining and learning to predict the perfect analysis for any of the real time world applications such as modern society, including mobile services, retail, manufacturing, financial services, life sciences, earth sciences and chemical sciences and physical sciences. Big data also has its revolutionary step in many fields such as scientific, research, education, healthcare etc. The observations from the above literature survey are:

1. The performance slit between the software in usage and the optimized code needed for performing parallel processing remains, which was answered by wide SMID and dynamic optimization.
2. Using parallel database and parallel computing environment the difficulty was querying real time environment that was answered by DC-Tree with bloom filter and Z-curve.
3. Mining of real time data such as large traffic datasets in the parallel computing environment a

difficult one answered by cloud based traffic mining platform.

4. Pipelining is also a method to performing parallel computing process which has the uncertainty problem of pipelining and also answered by probabilistic pipeline model.

5. Cloud answered the parallel computing processing and it has the physical gap between the data and the application. It was a difficult one. The network pace offered for transmission was also remarkable. Balls to bins provide solution for load balancing and keep the data closer.

6. The cloud computing creates its own era in technical fields influence education side to admire this skills to learners. VMM environment provides answer for this. Mahout is the machine language used in VMM.

7. To answer the real time environment such as traffic management, various cloud infrastructure services were under process, hence the better infrastructure using ontology based system. The ontology database is used to store the traffic data using and map reduce processing framework for parallel and distributed computing environment.

8. The selection of suitable HPC configuration for effective processing, that is low cost and high performance was a difficult one. Analyzing the map reduce algorithm, it gives the benchmark to provide the solutions.

9. The parallel processing and distributed data storage for analytics was analyzed and handled effective by Hadoop tool using Map reduce algorithm to run a set of jobs. Extension of this work can be used in the cross domain SELinux environment.

10. Maintaining the host and accelerator across various cloud infrastructure was important one because of rapid development in Big Data. The new system architecture for cluster resource allocation and data movement achieved by Oncilla and global address space were used for this.

11. The major expectation in cloud usage was its continuation in service like power and storage environment. Simulator which satisfies this requirement was CloudSim, open source simulator.

12. To reduce the time complexity and privacy issues in cross domain service, the HIRESOME– II was used. It records the history maintained by its Qos value for answering the queries.

13. To increase the efficiency of cloud computing, ADSC method is used to improve the query processing.

14. The query processing gives the better output expected by the used. Instead of providing suggestion for a query based on ranking, KASR provides suggestion based on the preference in user demands.

15. HIasS a cloud computing framework which used the human intelligence in for processing big data in a crowd source market place.

16. Providing time reduction in answering a query for health records, TCTM provides automatic annotation tool for processing.

17. To address scalability, response time reduction and node failure, Parallel-PIC model was used along with MapReduce.

18. In MapReduce analyzing all the data every time was questionable. Itchy provides the better solution than by the MapReduce.

19. In the field of Bioinformatics, to provide efficiency the Genome Sequencing is the techniques used along with Hadoop.

20. The MOA tool is used to analyze the Big data earlier to Hadoop, it has a lot of research values to know and introduce the Hadoop.

Conclusion:

The big data in cloud storage can be achieved through a lot of processing tools. The parallel processing is the one which can handle the large number of processes coming from various sources of network through high end technology solutions. The parallel processing can also be achieved through a cloud network such as cloud storage and infrastructure as a service such as IBM, Google, Saleforce.com and EMC service providers. The high speed connectivity of these services leads to better solutions for the big data analytics to analyze and predict to provide better solutions for various applications. The query processing also required for analyzing better results using the various query processing tools on big data using HPCS. The MOA and Hadoop tools are used for analyzing the query processing and prediction of data processing. The Hadoop tool can predict spaghetti structures in multimedia and earth science and chemical science applications. Therefore this article gives a complete survey and state of the art of the big data in cloud storage using Hadoop tool. But the recent big data tool called “Mesa” also can create more impact on future research.

REFERENCES

Arun Kejariwal, 2012. Big Data Challenges a Program Optimization Perspective. IEEE Second International Conference on Cloud and Green Computing.

Changqing, Ji., Li. Yu, Wenming Qiu, Uchechukwu Awada, Li. Keqiu, 2012. Big Data Processing in Cloud Computing Environments. IEEE International Symposium on Pervasive Systems, Algorithms and Networks.

Chih-Wei, Lu., Chih-Ming Hsieh, Chih-Hung Chang and Chao-Tung Yang, 2013. An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation. IEEE 37th Annual Computer Software and Applications Conference Workshops.

Dan Wei Chen, Jun Zhuang, 2013. A real time index model for big data based on DC-Tree IEEE

International Conference on Advanced Cloud and Big Data.

Dingcheng, Li., Cui Tao, Hongfang Liu, Christopher Chute, 2012. Ontology-based Temporal Relation Modeling with Map-Reduce Latent Dirichl*et al*locations for Big EHR data. IEEE Second International Conference on Cloud and Green Computing.

Jayalatchumy, D.P., Thambidurai, 2014. Parallel Processing of Big Data using Power Iteration Clustering over Map Reduce . IEEE World Congress on Computing and Communication Technologies

Jeff Young, Se Hoon Shon, Sudhakar Yalamanchili, Alex MerriUt, Karsten Schwan, IEEE, 2013 Oncilla: A GAS Runtime for Efficient Resource Allocation and Data Movement in Accelerated Clusters IEEE

Jianjun, Yu., Fuchun Jiang,Tongyu Zhu, 2013. RTIC-C: A Big Data System for Massive Traffic Information Mining. IEEE International Conference on Cloud Computing and Big Data.

Jongtack Jung and Hwangnam Kim, 2012. MR-CLOUDSIM: Designing and Implementing MapReduce Computing Model On CloudSim.

JorgSchad, Jorge-ArnulfoQuian´e-Ruiz, Jens Dittrich, 2013. Elephant, do not Forget Everything! Efficient Processing of Growing Datasets. IEEE Sixth International Conference on Cloud Computing

Karthik Raman, Adith Swaminathan, Johannes Gehrke, Thorsten Joachims, 2013. Beyond Myopic Inference in Big Data Pipelines. IEEE

Moussa Taifi, Y. Justin Shi, 2014. Map Reduce Performance Evaluation on a Private HPC Cloud. IEEE 41st International Conference on Parallel Processing Workshops.

Paul Heinzlreiter, T. Michael Krieger, Iris Leitner, 2012. Hadoop-based Genome Comparisons. IEEE Second International Conference on Cloud and Green Computing.

Petra Berenbrink, Andre Brinkmann, Tom Friedetzky, Dirk Meister, Lars Nagel, 2013. Distributing Storage in Cloud Environments. IEEE 27th International Symposium on Parallel & Distributed Processing Workshops and PhD Forum

Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, 2013. KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications. IEEE Transactions on Parallel and Distributed Systems.

Sivadon Chaisiri, 2013. Utilizing Human Intelligence in a Crowd Sourcing Marketplace for Big Data Processing. International Conference on Parallel and Distributed Systems

Thuy Nguyen, D., A. Mark Gondree, Jean Khosalim, E. Cynthia Irvine, 2013. Towards A Cross-Domain MapReduce Framework. IEEE Military Communications Conference.

Wanchun Dou, Xuyun Zhang, Jianxun Liu and Jinjun Chen, Senior Member, 2013. HireSome-II: Towards Privacy-Aware Cross- Cloud Service Composition for Big Data Applications. IEEE Transactions On Parallel And Distributed Systems.

WenQiang Wang, Xiaoming Zhang, Jiangwei Zhang, Hock Beng Lim, 2012. Smart Traffic Cloud: An Infrastructure for Traffic Applications. IEEE 18th International Conference on Parallel and Distributed Systems.

Yuichiro Takabe, Minoru Uehara, 2013. Rapid Deployment for Machine Learning in the Educational Cloud. IEEE International Conference on Network-Based Information Systems.