International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 28593-28602

http://www.ripublication.com

Big Data Literature Survey

Shobhanjaly P. Nair[1], N. Gomathi[2], D. Archana Thilagavathy[3]
[1]Assistant Professor, Department of CSE,Vel Tech Dr. RR & Dr. SR Technical University,
[2]Associate Professor, Department of CSE,Vel Tech Dr. RR & Dr. SR Technical University,
[3]Assistant Professor, Department of ECE,Vel Tech Dr. RR & Dr. SR Technical University,
Chennai, Tamil Nadu,India.
Email: - anjaly.cse@gmail.com

Abstract

This paper throws light on various methods and concepts used for Big Data analysis and their advantages and drawbacks, also enhancement of each method to overcome their drawback to serve better. There is a consistent development in Big Data field which lead to an enormous progress and improvement in the concepts used for storing data, analyzing them, extracting and furnishing data when required by user. The article highlights about different techniques used and their intensification from time to time.

Key words: Big Data, HACE Theorem, Hadoop method, MapReduce

Introduction

In the current engineering world, data are created and stored in a large amount. Data are collected from anywhere in any format, this overpowering quantity of data could not be stored in a small database, to overcome restrictions due to large and complex data sets since the terrific development in the field of information and communication technology, **big data** has been transpired. Big data is a cluster of huge amount of assorted, unstructured, distributed and complex datasets, and facilitate improved data processing function which is difficult to be processed in traditional application. Big data is a technology which involves new techniques to integrate and discover huge amount of concealed values from enormous dataset where varied, complex and large scale of data are stored. Big data promises to focus on important concepts for better decision making process, which in turn enhances organizational efficiency and effectiveness, also have fascinated interest from academia to carry out research in the field of big data. The big data concept, offers high volume, velocity, and variety,

which helps in mining latest knowledge to improve economic development and technical advancement. Enormous data are gathered from different kinds of users and devices, and are stored and processed data centers.

Literature Survey

The decision making ability of big data is improved by implementing analytical tools and methods that helps in making sense to the given data, and thereby improve organizational efficiency and effectiveness. The article suggested by Rhoda et al, explains three analytical tools and their barriers when used in public sectors. The analytical tools used are, *descriptive analytics, predictive analytics and prescriptive analytics. Descriptive analytics is used to* generate standard reports, ad-hoc reports, and alerts; *Predictive analytics* is related with forecasting and statistical modeling; and *prescriptive analytics* centers on optimization and randomized testing. To leverage data these tools could be implemented in private sector organization but it's a difficult task to be implemented in public sector. Government services could be enhanced with the use of these analytics but there are few drivers and barriers disturbing the use of big data in e-government. This can be defeated by developing e-government to transformational government (t-government).Flourishing t-government will assist in product advancement, better customer service delivery, improved management of the open sector infrastructure, and lower operational costs. T-government takes place through; *aggregation, syndication, consumption, co-creation.* To upgrade the efficiency, automating data analysis and redesigning processes are implemented and to upgrade effectiveness, data are segmented and are made more transparent. To enhance the transformative value of big data in government, appropriate resource allocation should be done, ensuring that the data is properly captured and controlled.

Big Data concentrates on accommodating massive, distributed, independent and composite database which can handle data of any kind. Use of Big data is progressing continuously in all area which includes science (bio-medical, physics, biology etc), production, engineering and many more, with the technical improvement in networking, data storage, and data collection methods. X Wu et, presents a new technique called HACE theorem (**H**eterogeneous, Autonomous, **C**omplex & **E**volving), which helps in developing a Big Data processing model, from the knowledge mining viewpoint . The theorem is data-driven form and involves demand-driven aggregation of data resources, mining, analyzing the data, user interest modeling, and security and privacy considerations.

As there is an increase in the amount of data stored, it has become a necessity to develop a more effective storage system; this has led its way for an emerging big data field. Big data can store volumes and volume of data and could be extracted when required. Also extracting data is more effective since all data are of distinct characteristics even though big data stores distributed and unstructured data. But all big data stored information is real time facts which make it more difficult to extract, for which system architecture is improved for data acquisition, transmission, storage, and large-scale data processing mechanisms. This article is a literature survey and

discussion by Wen et al, about big data analytics platform, to help novice to understand the concept in a better way. Big data faces lots of problems, primarily in gathering data because of dissimilarity in sources from where data is collected and secondly in storing these large amounts of data and to give an assurance in accuracy of data extracted. This problem could be reduced by using a better RDBMS system and for traffic regulations use MapReduce technique. There are mainly three types of definition such as *Attribute Definition, Comparative Definition, Architectural Definition* used to structure big data. A systematic framework is constructed to breakdown big data systems into four sequential modules, such as *data generation, data acquisition, data storage, and data analytics*. Real world scenarios are evaluated and the big data system is verified. A new approach called Hadoop framework is introduced to deal with big data issues. After several evaluation scaling and potential examinations instructions to use a big data system is sketched.

Before introducing the concept of Big Data, it was difficult to save several customers queries and information. The technical people found it hard and they would delete the information and query as soon as the solution is provided to user. But, the world of technology is developing and therefore it has become the need of the hour to store all data and other information for future retrieval. The facts are retained for further analysis to improve the knowledge of information stored, easy to recognize customer pattern, first call resolution and to capitalize on every day conclusion for a new technical employee which would minimize the agitation and help in providing customer contentment. A D Barrachina and A O'Driscoll, has proposed a Proof of Concept (PoC) , which serves as an end to end solution that make use of the Hadoop programming model, extended ecosystem and the Mahout Big Data Analytics. These concepts used as records for classifying similar support calls for large technical support data sets. The proposed solution is evaluated on a VMware technical support dataset. The continuous solution proposed by the authors achieves the concept by Data pre processing, the Automated periodic refresh of new technical support data, Parallelized clustering, Real-time platform access, User query of results. These concepts were used in real world applications which provided proof of concept. The tools used could be enhanced further to achieve a better result.

According to R.Lu., et al, Big data, is used to extract new data for economic development and technical improvement. Big data has gained significant attention, and many research works are directed to big data processing. This is mainly due to its high volume, velocity, and variety (referred to as "3V") challenges. Big data also helps to face security and privacy challenges. The collected data must be legitimate; else data mined will not be credible. Unless privacy problem is rectified, people won't be interested to share their information. The existing privacy problems are Privacy requirements in big data collection, Privacy requirements in big data storage and Privacy requirements in big data processing and methods to eradicate them are Privacy-preserving aggregation, Operations over encrypted data and De-identification. The security problem are deemed as a new dimension, "veracity", in this research, author aims to exploit existing problems of big data in terms of privacy, and work towards efficient and privacy-preserving computing. An efficient and privacy-preserving cosine similarity computing protocol is developed after analyzing

the architecture of big data and privacy requirements. The cosine similarity concept measures two similar objectives capture by vectors $\vec{a}$ and $\vec{b}$. Though the privacy problems are analyzed and new concepts to fix them are introduced, it's necessary to address unique privacy issues as it helps to improve research efficiency to a large extend.

Big data, promises to determine important approaches to betterment the decision making, because it's been used extensively in academia circles and industries. Large data are generated from a different kind of users and devices, which in turn is stored and processed in efficient data centers for knowledge discovery. To collect and gather geologically distributed and rapidly generated data we require a very eminent network infrastructure. The network architecture must be expanded to interconnect multiple data centers and the server nodes within data center. X. Yi et al, have briefed about the challenges to build a network infrastructure for big data. They have designed a three segmented network highway- the access networks that connect data sources, the Internet backbone that bridges them to remote data centers, as well as the dedicated network among data centers and within a data center. This paper explains about all the three segments with two real-world case study, since in real world applications data are collected from user of different categories and through different ways and data center stores distributed data, which is processed and accurate information must be given as output, all which requires an efficient internet network infrastructure. With the swift progress in the field of big data applications, creating an enhanced network infrastructure will remain to be task to be accomplished. In future the network system must also provide standardization of data and privacy and security.

The rapid development in the big data field has made the processing of data complicated for computing, large storage and passing information in data centers. To overcome these issues we are forced to use an expensive big data system. Lin Gu et al, have suggested a cost minimization method to support a economic big data processing. Since data and computation are interconnected in big data, the computation of a data could be performed only based on the availability subsequent data. The problems which lead to find a effective big data system with cost-efficiency is misuse of resources space, variation in transmission rates and Quality –of- service. As per this research paper, the three features which assist in cost minimization in geo-distributed data center are task assignment, data placement, and data movement. Also a 2-D Markov chain is suggested to get the mean completion time by illustrating the task completion time with respect to data communication and computation. The task is molded as Mixed- integer nonlinear programming and it is linearized by the Gu et al., concept. The research article proves the benefits of optimization of three key factors to get a cost effective big data processing by various testing.

The development of information and communication technology has led to the necessity of expanding the data storage space. Since all the information received are saved in the database. This has paved the way for introducing big data concepts, which has become a global topic of significance. To enhance the efficiency of big data, distributed wireless sensor networks (WSNs) are used. Data generated by a single sensor though is of valid use are not are not given priority, but those data collected from different distributed sensors are given more priority. Research in this

area is more challenging due to distributed sensor's high energy usage. A way to deal with this issue is to make use of sink node's mobility to simplify the resource collecting. But there are other obstacles with sink node such as determining the sink node's trajectory and cluster formation ahead of data gathering. D Takaishi et al, in this paper has suggested a new mobile sink routing and data gathering technique by network clustering based on modified expectation-maximization technique, also several favorable clusters are drawn to reduce the energy usage. In this research article the author has proposed an energy minimized clustering algorithm by using the Expectation-Maximization (EM) algorithm for 2-dimensional Gaussian mixture distribution, it focus on reducing the sum of square of wireless communication distance as the energy usage is relative to square of the wireless communication distance also on data request flooding issue to find the finest clusters. The clustering is done using Low-Energy Adaptive Clustering Hierarchy (LEACH) algorithm, again a drawback in this algorithm is; only nodes with long lasting energy are given more significance. Therefore a better algorithm called *K*-hop Overlapping Clustering Algorithm (KOCA) is used, which depends on probabilistic cluster head selection and nodes' location. WSN is divided into sub-networks. Power-Efficient Gathering in Sensor Information Systems (PEGASIS) and KAT mobility (*K*-means And TSP mobility) are centralized clustering algorithm used. The research work offers efficient data collection and to identify best clusters.

The distinctiveness of big data must be identified, but it is complicated using traditional models and algorithms. Data collected through remote sensing contain extensive images that are complicated based on their structural, spectral and textual aspects. L Wang et al, have proposed wavelet transform to exhibit remote sensing big data that occupies huge amount of space, interconnected in spectral domain and constant in time domain. The big data set is disintegrated into estimated multi-scale coefficients using proposed method. In addition to this a two-component Gaussian mixture model (GMM) is used to find the density function of wavelet coefficients, and to check if it is at zero and shows tailed shape curve. The change in GMM depending on variations in bands, time, and scale are widely examined. It is evident in a remote sensing big data, that wavelet coefficient of each cluster is distinctive based on different bands and scales. In addition to different bands and scale the article also introduces different textures of big data which affects probability density function and GMM parameters of the wavelet coefficients.

The research article helps to sample and transmute the remote sensing big data into a wavelet form. By conducting different experiments such as, one is by decomposing all bands individually by wavelet transform and compare. Another method is by focusing on the statistical characteristics of remote sensing data set images that are of long-term sequence. In the third type of experiment different texture data sets are crumbled into varied scales using wavelet transform the Statistical Characteristics of Remote Sensing Big Data in the Wavelet Transform area is estimated.

J C Bertot et al, have purposed an article which helps to observe the ways in which governments build social media and information and communication technologies (ICTs) into e-government transparency scheme, to promote alliance with members of the public and the ways in which the members of the public are able to utilize the

same social media to monitor government activities. The method used is an iterative strategy that involved conducting a literature review, content analysis, and web site analysis, offering multiple perspectives on government transparency efforts and the ability of e-government scheme to foster collaborative transparency through embedded ICTs and social media. This helps in identifying key initiatives, potential impacts, and future challenges for collaborative e-government as a means of intelligibility. The paper is one of the first to examine the interrelationships between ICTs, social media, and collaborative e-government to facilitate transparency. By improving lines of communication, ICTs – like the telegraph and then telephones – were able to provide a tool of increased effectiveness in colonial administration and control. Therefore telegraph, telephone and social media applications of the internet all have been improved and facilitate to enhance existing approaches to transparency and foster new cultures of openness both by giving governments new tools promote transparency and reduce corruption and by empowering members of the public to collectively take part in monitoring the activities of their governments.

A combined approach to learn about Bayesian network working, from a distributed heterogeneous data has been proposed by R. Chen et al. In this approach, first a local Bayesian network at each site is learned using the local data. Then each site identifies the observations that are most likely to be evidence of coupling between local and non-local variables and transmits a subset of these observations to a central site. Second, Bayesian network is learnt at the central site using the data transmitted from the local site. The local and central Bayesian networks are pooled to obtain a collective Bayesian network, which models the entire data. Experimental results and theoretical justification demonstrate the feasibility of the proposed method.

Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. S. Melnik et al, describes the architecture and implementation of Dremel, and explain how it complements MapReduce-based computing in their proposed article. A novel columnar storage representation for nested records is introduced and experiments on few-thousand node instances of the system are discussed. Dremel, is a distributed system used for interactive analysis of large datasets. Also it is a custom, scalable data management solution built from simpler components. Its performance efficiency on trillion-record, multi-terabyte datasets of real data is one of the best feature of using Dermel. The key aspects of Dremel, includes its storage format, query language, and execution. Future enhancements can be, planed, to cover in-depth areas such as formal algebraic specification, joins, extensibility mechanisms.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key pair to generate a set of intermediate key pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. J. Dean and S. Ghemawat implements MapReduce which runs on a large cluster of service machines and is highly scalable: a typical MapReduce computation, processes many terabytes of data on thousands of machines. Programmers find the system easy to use as hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day since the run-time system takes care of the details of partitioning the input data, scheduling the

program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication.

J. Lu and D. Li discuss about the bias problem when estimating the population size of big data such as online social networks (OSN) using uniform random sampling and simple random walk in their article. The paper shows analytically, that the relative bias can be approximated by the reciprocal of the number of collisions; thereby, a bias correction estimator is introduced. Estimators are usually evaluated by both bias and variance. The purpose of this paper is not to evaluate the overall performance of the estimator, instead, it shows that there is a bias, and the bias of N can be too large to neglect when sample size is small relative to the big data being studied. The bias correction formula works for both uniform random sampling and random walk sampling.

High throughput concern is one of the major hindrances for opening up the new era of big data stream computing on cloud. Dawei Sun et al has briefed, the definition of data stream graph and the principles for high throughput objectives are systematically analyzed through this paper. An elasticity adaptive data stream graph model EADSG is put forward, which includes: (1) modeling data stream graph in big data stream computing environments by referring to the distributed stream computing theories, identifying the bottlenecks at special data throughput, and getting hot or cold regions in a data stream graph; (2) optimizing the data stream graph and allocating resources by the elasticity adaptive strategy EADSG, to maximize the throughput and minimize response time, to meet high throughput objectives; and, (3) evaluating the high throughput objectives in big data stream computing environments, and to consider both high throughput and high response time objectives. Theoretical as well as experimental results conclusively demonstrate that the EADSG strategy has high potential as the strategy provides efficient throughput enhancements and significant response time reducing.

Methods and Concepts Used

The different methods and concepts used till date, their functioning with their merits and demerits are shown in Table I.

Table 1: Methods Merits and Demerits

S. No	Concept Name	Working	Merits	Demerits
1	HACE theorem: **H**eterogeneous, **A**utonomous, **C**omplex & **E**volving	Displays abstract view of the Big Data processing framework. It has three tier namely, data accessing and computing, data privacy and domain knowledge, and Big Data mining algorithms.	Holds heterogeneous, independent, unstructured, and diverse data sources which are decentralized control. Can extract any data without much fuss from the huge quantity of varied data. Good scalability	There are possibilities of missing certain data due to complex self-sufficient data set. Other issues may arise due to interferences like noise, privacy alert.
2	MapReduce	Does processing group-aggregation tasks which includes website ranking	Facilitates automatic data parallelization and the distribution of large-scale computation applications to large clusters of commodity servers. Thus has high scalability and can process unlimited	streams the map output**,** Name node keeps track of the metadata.
3	Hadoop framework	An open-source software framework that supports massive data storage and processing.	Scalability and Cost Efficiency. Flexibility and Fault tolerance	Hadoop has no schema and no index. Therefore the input read is transformed into data objects, all which lead to low efficiency.
4	Privacy-preserving cosine similarity computing protocol	Efficiently calculates the cosine similarity of two vectors without disclosing the vectors to each other.	By applying homomorphic encryption (HE), such as Pallier encryption (PE), PCSC will not disclose details of comparing vectors to each other. Thus provides privacy and security.	This technique may not be effective in a few explicit big data analytics.
5	Expectation-Maximization (EM) algorithm	EM algorithm focus on reducing the sum of square of wireless communication distance as the power usage is proportional to the square of the wireless communication distance.	Overcomes ``data request flooding problem" (energy inadequacy that arise when all the nodes transmit data request messages to their respective neighboring nodes) and decides the optimal number of clusters. EM algorithm can reduce required energy significantly	Algorithm do not work well when node density is low, becomes a worst clustering algorithm. EM algorithm can connect only to a less number of nodes when the number of nodes is low and node density is low.
6	*K*-hop Overlapping	Based on probabilistic cluster head selection		But in WSNs, minimizing data

	Clustering Algorithm (KOCA) and k-hop Connectivity ID (k-CONID)	and nodes' location, a multiple overlapping clusters, called KOCA algorithm was designed. Similarly in k-CONID the nodes exchange their random IDs with each other, and the node that has the minimum ID within k-hop is selected as a cluster head.		transmission is difficult since this algorithm is probabilistic cluster based whereas we need a centralized clustering algo-Rithm to get the information from sub-divided WNS nodes. Thus, the algorithm won't attain optimization.

Conclusion

Big data is used by wide-range of organizations varying from health centers, government organizations to industries and institutions where large amount of information must be stored. In the past, numerous methods were proposed and every concept was enhanced constantly to surmount the drawbacks and to help in big data filed in a better way. From the drawbacks of above mentioned methods and concepts, there is lot of scope to develop better solution for gathering and storing data which could be extracted back when required in the Big Data field.

References

[1] Rhoda C. Joseph, *Pennsylvania State University Harrisburg*, Norman A. Johnson . 2013. Leveraging Big Data: Big Data and Transformation Government. *University of Houston.* IEEE Computer Society 1520-9202**.**

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

[3] Yonggang Wen, Han Hu, Tat-Seng Chua, and Xuelong . 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*. VOLUME 2, 2169-3536.

[4] Arantxa Duque Barrachina and Aisling O'Driscoll. 2014. Big data methodology for Categorising technical support requests using Hadoop and Mahout. *Journal Of Big Data* . 1:1 doi:10.1186/2196-1115-1-1.

[5] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao. 2014. Toward Efficient and Privacy-Preserving Computing in Big Data Era. *IEEE Network*. 0890-8044/14.

[6] Xiaomeng Yi, Fangming Liu, Jiangchuan Liu, and Hai Jin. 2014. Building a Network Highway for Big Data: Architecture and Challenges. *IEEE Network* , 0890-8044/14, IEEE 5.

[7] Lin Gu, Deze Zeng, Peng Li, and Song Guo. 2014. Cost Minimization for Big Data Processing in Geo-Distributed Data Centers. *IEEE, TRANSACTIONS ON EMERGING TOPICS IN COMPUTING,* VOLUME 2, NO. 3, 2168-6750.

[8] Daisuke Takaishi, HirokI Nishiyama, Nei Kato and Ryu Miura. 2014. Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks. *IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING*, VOLUME 2, NO. 3, 2168-6750.

[9] Lizhen Wang, Hui Zhong, Rajiv Ranjan, Albert Zomaya, and peng liu. 2014. Estimating the Statistical Characteristics of Remote Sensing Big Data in the Wavelet Transform Domain. *IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING*, VOLUME 2, NO. 3, 2168-6750.

[10] J.C. Bertot, P.T. Jaeger, and J.M. Grimes,2012, "Promoting Transparency and Accountability through ICTs, Social Media, and Collaborative E-Government," *Transforming Government: People, Process and Policy*, vol. 6, no. 1, pp. 78–91.

[11] R. Chen, K. Sivakumar, and H. Kargupta, 2004. "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187.

[12] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis., 2010. Dremel: Interactive analysis of web-scale datasets," Proc. VLDB Endowment, vol. 3, nos. 1_2, pp. 330_339.

[13] J. Dean and S. Ghemawat, 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM*, vol. 51, no. 1, pp. 107_113.

[14] J. Lu and D. Li, Bias correction in a small sample from big data, *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2658_2663, Nov. 2013.

[15] Dawei Sun, Ge Fu, Xinran Liu, Hong Zhang. 2014. Optimizing Data Stream Graph for Big Data Stream Computing in Cloud Datacenter Environments International Journal of Advancements in Computing Technology(IJACT) Volume 6, Number 5.