IPHV8I1067X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .I
ISSN Online: 2319 – 9253
Print: 2319 – 9245

BRING BIG DATA TO THE ENTERPRISE : PROTECTING CONFIDENTIALITY IN BIG

[#1]K.VIJAYA KUMAR- M.Tech Student,
[#2] B.PRABHAKARA REDDY- Associate Professor,
Dept of CSE,
BHEEMA INSTITUTE OF TECHNOLOGY & SCIENCE, ADONI, A.P, INDIA.

ABSTRACT— Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This study paper includes the information about what is big data, Data mining, Data mining with big data, Challenging issues and its related work.

Keywords *— Big Data, Data mining, Challenging issues, Datasets, Data Mining Algorithms.*

I.INTRODUCTION

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress". Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya. However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold .The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD Big Mine‟ 12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Face book has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices and big companies as Google, Apple, Face book, Yahoo are starting to look carefully to this data to find useful patterns to improve user experience. “Big data” is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. In industries throughout the world, executives recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing. So, we sought to better understand how organizations view big data – and to what extent they are currently using it to benefit their businesses.

Big data refers to the enormous amount of structured and unstructured data that overflow the organization. If the overflowed data is used in a proper way it leads to meaningful information. When big data is compared to traditional databases it includes a large number of data which requires more processing in real time. It also provides opportunities to discover new values, to understand an in-depth knowledge from hidden values and also provides space to manage those data effectively. Big Data concern large-volume, complex, growing datasets with multiple data sources. With the fast development of networking, data storage and data collection capacity, big data are now expanding in all science and engineering domains, including physical, biological and biomedical sciences.[1]. Data Mining is a task of identifying relevant and significant information from large data set.

II. BIG DATA WITH DATA MINING

Generally big data refers to a collection of large volumes of data and these data are generated from various sources such as internet, social media, business organizations etc., With these data some useful information can be extracted with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data[2].

IPHV8I1067X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .I
ISSN Online: 2319 – 9253
Print: 2319 – 9245

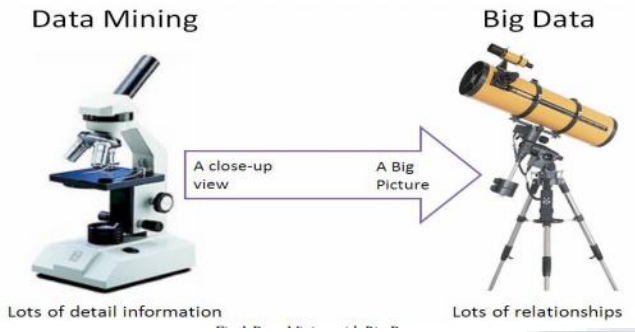

Fig.1 Data Mining with Big Data

The figure 1[3] given above portraits the relationship of big data with data mining. From the figure it is observed that big data gives lots of relationships and data mining gives lots of information.

II1. CHALLENGES IN BIG DATA

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day. The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices. Furthermore, the variety of data being generated is also expanding, and organization‟s capability to capture and process this data is limited. Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data.

A. Privacy, security and trust

The Australian Government is committed to protecting the privacy rights of its citizens and has recently strengthened the Privacy Act (through the passing of the Privacy Amendment (Enhancing Privacy Protection) Bill 2012) to enhance the protection of and set clearer boundaries for usage of personal information. Government agencies, when collecting or managing citizens data, are subject to a range of legislative controls, and must comply with the a number of acts and regulations such as the Freedom of Information Act (1982), the Archives Act (1983), the Telecommunications Act (1997) ,the Electronic Transactions Act (1999), and the Intelligence Services Act (2001). These legislative instruments are designed to maintain public confidence in the government as an effective and secure repository and steward of citizen information. The use of big data by government agencies will not change this; rather it may add an additional layer of complexity in terms of managing information security risks. Big data sources, the transport and delivery systems within and across agencies, and the end points for this data will all become targets of interest for hackers, both local and international and will need to be protected. The public release of large machine-readable data sets, as part of the open government policy, could potentially provide an opportunity for unfriendly state and non-state actors to glean sensitive information, or create a mosaic of exploitable information from apparently innocuous data. This threat will need to be understood and carefully managed. The potential value of big data is a function of the number of relevant, disparate datasets that can be linked and analysed to reveal new patterns, trends and insights. Public trust in government agencies is required before citizens will be able to understand that such linking and analysis can take place while preserving the privacy rights of individuals.

B. Data management and sharing

Accessible information is the lifeblood of a robust democracy and productive economy.2 Government agencies realize that for data to have any value it needs to be discoverable, accessible and usable, and the significance of these requirements only increases as the discussion turns towards big data. Government agencies must achieve these requirements whilst still adhering to privacy laws. The processes surrounding the way data is collected, handled, utilized and managed by agencies will need to be aligned with all relevant legislative and regulatory instruments with a focus on making the data available for analysis in a lawful, controlled and meaningful way. Data also needs to be accurate, complete and timely if it is to be used to support complex analysis and decision making. For these reasons, management and governance focus needs to be on making data open and available across government via standardized APIs, formats and metadata. Improved quality of data will produce tangible benefits in terms of business intelligence, decision making, sustainable cost-savings and productivity improvements. The current trend towards open data and open government has seen a focus on making data sets available to the public, however these „open‟ initiatives need to also put focus on making data open, available and standardised within and between agencies in such a way that allows inter-governmental agency use and collaboration to the extent made possible by the privacy laws.

C. Technology and analytical systems

The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a consequence of recent advances in the technology that allow this. If big data analytics is to be adopted by agencies, a large amount of stress may be placed upon current ICT systems and solutions which presently carry the burden of processing, analyzing and archiving data. Government agencies will need to manage these new requirements efficiently in order to deliver net benefits through the adoption of new technologies.

IPHV8I1067X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .I
ISSN Online: 2319 – 9253
Print: 2319 – 9245

IV. METHOD OVERVIEW

4.1 Data Cube

Data cube provide multi-dimensional views in data warehousing. If n dimensions given in relation then there are 2^n cuboids and this cuboids need to computed in the cube materialization using algorithm[2]which is able to facilitate feature in Map Reduce for efficient cube computation. In data cube Dimension and attributes are the set of attributes that user want to analyze. Cube lattice is formed representing all possible groupings of this attributes, based on those attributes. After that by grouping attribute into hierarchies and eliminating invalid cube regions from lattice we get more compact hierarchical cube lattice. Finally cube computation task is to compute given measure for all valid cube groups. There are different techniques of cube computations [3] like multi- dimensional aggregate computation, BUC(Bottom-Up Computation), star cubing for efficient cube computation. There are limitations in these techniques:

1) They are designed for a single node or for a cluster with less nodes [19], so it is difficult to process data with a single or few machines.

2) Many analyses over logs, involve computing holistic measure where as many techniques uses the algebraic measures.

3) Existing techniques failed to detect and avoid data skew. There is need of technique to compute cube in parallel on holistic measure over massive dataset. Hadoop based MapReduce can handle large amount of data in cluster with thousand of machines. So this technique is good option for analysis of data.

4.2 Map Reduce

MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. The nature of this programming model and how it can be used to write programs which run in the Hadoop environment is explain by this model. Hadoop [11] is an open source implementation for this environment. Map and Reduce are two functions. The main job of these two functions are sorting and filtering input data. During Map phase data is distributed to mapper machines and by parallel processing the subset it produces pairs for each record. Next shuffle phase is used for repartitioning and sorting that pair within each partition. So the value corresponding same key grouped into {v1, v2,….}values. Last during Reduce phase reducer machine process subset pairs parallel in the final result is written to distributed file system.

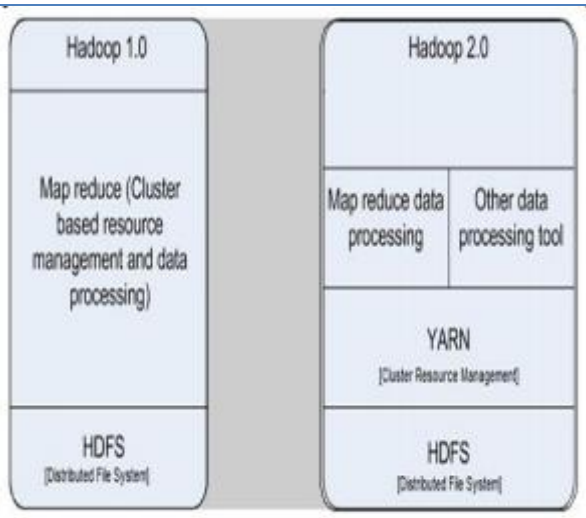

Fig.2:- Architecture of Hadoop1.0 and 2.0

MR1 is used in Hadoop1.0 but due to some resource management issues like inflexible slot configuration, scalability. After Hadoop version 0.23, Map Reduce changed significantly. Now it known as Map Reduce 2.0 or YARN (Yet another Resource Negotiator). Map Reduce 2.0 has two major functionalities of job tracker which are spit into resource management and job scheduling into separate daemons [4]. Fig 2 shows the architecture of both Hadoop versions. In Hadoop1.0 Job Tracker has a responsibility for managing the resources and scheduling jobs across the cluster. But in Hadoop2.0 the architecture of YARN allows the new Resource Manager to manage the usage of resources across all applications. And Application Masters takes the responsibility of managing the job execution. This new approach improves the ability to scale up the Hadoop clusters to a much larger configuration than it was previously possible. In addition to this, YARN permits parallel execution of a range of programming models. This includes graph processing, iterative processing, machine learning, and general cluster computing.

4.3 MR-cube

Approach MR-Cube is a Map Reduce based algorithm introduces for efficient cube computation [5] and for identifying cube sets/groups on holistic measures. MR-Cube algorithm is used for cube materialization and identifying interesting cube groups. Complexity of the cubing task is depending upon two aspects: size of data and size of cube lattice. Size of data impacts size of large group and intermediate size of data, where as the cube lattice size impacts on intermediate data size and it is controlled by the number/depth of dimension. First we identify the subset of holistic measures that can easily compute in parallel than an arbitrary holistic measure. We can call it Partially Algebraic Measures. The technique of partitioning large groups based on algebraic attribute called Value partitioning. Value

partitioning is used to effectively distribute the data; we can easily compute it with Naïve algorithm [9]. Value partitioning performs on only on group that are likely reducer friendly and dynamically adjust the partition factor. Partition factor is ratio by which a group is partitioned. There are different approaches for detecting reducer unfriendly groups. One of the approach is sampling approach where we estimate the reducer unfriendliness of cube region based on the number of groups it is estimated and perform partitioning for all small groups within the list of cube region that are estimated to be reducer unfriendly.

4.4 Cube Materialization

Cube materialization task comes under the MR-Cube approach. Materializing the cube means computing measures for all cube groups satisfying the pruning condition. After materializing cube we can identify the interesting cube groups for cube mining algorithm. The main MR-CUBE-MAP-REDUCE task is perform using annotated lattice. The combine process of identifying and value partitioning unfriendly regions followed by partitioning of regions is referred as annotate. Based on the sampling results cube regions have deemed as reducer unfriendly and require partitioning. Each tuple in dataset the MR-Cube-Map emits key:value pairs for each batch area. In required keys are appended with hash based on value partitioning. The shuffle phase then sorts them by key yielding reducer tasks. The BUC algorithm is then run on each reducer and cube aggregates are generated. The value partitioned group are merged during post processing to produce the final result.

V. RELATED WORK

On the level of mining platform sector, at present, parallel programming models like Map Reduce are being used for the purpose of analysis and mining of data. Map Reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model. For those people, who intend to hire a third party such as auditors to process their data, it is very important to have efficient and effective access to the data. In such cases, the privacy restrictions of user may be faces like no local copies or downloading allowed, etc. So there is privacy-preserving public auditing mechanism proposed for large scale data storage.[1] This public key-based mechanism is used to enable third-party auditing, so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ. For example, when we go to the doctor for the treatment, that doctor's treatment program continuously adjusts with the conditions of the patient. Similarly the knowledge. For this, Wu [2] [3] [4] proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

VI. CONCLUSION

Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data(usually large amount of data-typically business or market related-also known as "big data")in search of consistent patterns and then to validate the findings by applying the detected patterns to new subsets of data. To support Big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. In real-world applications managing and mining Big Data is Challenging task, As the data concern large in a volume, distributed and decentralized control and complex. There are several challenges at data, model and system level. We need computing platform to handle this Big Data. The Map Reduce framework is one of the most important parts of big data processing, and batch oriented parallel computing model. In earlier versions of Map Reduce the components were designed to address basic needs of processing and resource management. Recently, it has evolved into a improved version known as Map Reduce 2/YARN that provides improved features and functionality. With Big Data technologies we able to provide most relevant and accurate social sensing feedback to better understand to society at real-time. MR-Cube efficiently distributes the computation workload across machines and completes the cubing task.

REFERENCES

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

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

[3]. Dhanshri S. Lad #, Rasika P. Saste, "Different Cube Computation Approaches: Survey Paper" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4057-4061
[4]. The Apache SoftwareFoundation"http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html"
[5]. Arnab Nandi, Cong Yu, Philip Bohannon, and Raghu Ramakrishnan, Fellow, IEEE, "Data Cube Materialization and Mining over MapReduce" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 6, NO. 1, JANUARY 2012
[6]. A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20- 23, 2012.
[7]. D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 296-301, 2009.
[8]. A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032- 2033,2012.
[9]. A. Nandi, C. Yu, P. Bohannon. And R. Ramakrishnan, "Distributed Cube Materialization on Holistic Measures, " Proc. IEEE 27th Int'l Conf. Data Eng. (ICDE), 2011.
[10]. F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.
[11]. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo" Yahoo! Developer Network Blog, 2008.
[12]. "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.
[13]. A. Rajaraman and J. Ullman, Mining of Massive Data Sets. Cambridge Univ. Press, 2011.
[14]. K. Yury, "Applying Map-Reduce paradigm for parallel closed cube computation," Proc. First Int'l
[15]. P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquini. Incoop: Mapreduce for incremental computations. In SOCC, 2011.

AUTHORS PROFILE:

[1].K.VIJAYA KUMARA, M.Tech Student, Dept of CSE, from Bheema Institute Of Technology & Science, Adoni, A.P, India. He Received M.C.A Degree from Osmania University, Hyd in the year 2005. He worked as assistant professor in various professional colleges. He has Served for 5 Years as Principal of Dr. Jyothirmayi Degree College, adhoni, AP,India. His research interests are in the field of DataMining.

[2]. PRABHAKARA REDDY BAGGIDI, received B.Tech Degree in Electronics and Communication Engineering, in the year 1997from SV University, Tirupathi, India. He is awarded with M.Tech Degree in Digital Systems & Computer Electronics in the year 2002 and currently carrying out Ph.D work in association with Jawaharlal Nehru Technological University, Anantapur, India. He guided many Academic projects for the last 15years of teaching experience. His research interests are in the field of Mobile Ad Hoc Networks & Optical Networks.