INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

The Graph based Query monitoring System in Data Warehousing Environment for resource allocation and redirection of workloads

Uma Pavan Kumar Kethavarapu[#], Sridevi.S.Erady [*], Govinda Raj Pandit [$]

[#]*Associate Professor, AIMS Institutions, Bangalore*

[1]*umapavanmtech@gmail.com*

[*] *Assistant Professor, IT Dept., AIMS Institutions, Bangalore*

[2]*sreedevi_erady@yahoo.com*

[$] *Associate Professor, IT Dept., AIMS Institutions, Bangalore*

[3]*grpadith@gmail.com*

***Abstract*— This paper gives the presentation about various levels of organizations ranging from small, medium, large and enterprise. The activities are common in all the levels of data warehousing environment, the only thing to consider is the amount of data and the number of systems and number of users participating in the processing of the data. The paper presents the dominating set model and balanced behaviour of the data warehousing recourses to handle the activities in better way, we are proposing a model of Query Monitor (QM) as a basic component so as to redirect and balance the workloads based on the availability of the systems in case of enterprise data the usage of big data along with Hadoop technology is presented. The importance of this work is usage of graph based model so as to get the work load estimation, creation of the resource table and redirecting the loads to the identified systems which will help the organizations and enterprises for better management of the resources.**

***Keywords*— big data, dominating set, query monitor, enterprise, hadoop, resource table.**

I. INTRODUCTION

The organizations ranging from small number of people to large are necessarily requiring the security implications so as to handle the data and activities in effective and efficient manner [1]. The ultimate goal of any organization is to serve the needs of their customers. Some organizations are planning
the transactions such as day-to-day activities in report format, some companies are maintaining the data in servers and the strategic decisions are taken from the reports generated by the server data. In some cases the data may not be available in a single location or in a single server but that may be distributed in various servers. In this case the data should be gathered from different locations and that is consolidated to form a report and further to maintain the strategic decisions. In all the above cases the security is a mandatory aspect so as to produce the correct decision by the people or company. With the usage of data warehousing tools it is possible to achieve Extraction, Transformation and Loading (ETL Process).Some of the tools to handle ETL process are Informatica, Abinitio, Data stage with parallel jobs. In data warehousing environment the reporting side we depend on Online Analytical Processing (OLAP) some of the tools are Business Objects, Cognos.These tools are used to generate the reports which are helpful for various levels of users such as end users, Business analysts, Data base designers and administrators. In data mining the pattern recognition and hidden data patterns are identified, by using data mining techniques the efficient searching is done on the data sets. Example for data mining tool is Weka.To handle all these activities the data must be secured in all aspects such as in server side, client side, networkside.A strong mechanism is required to handle the efficient security mechanism with out loss of the data from intruders and unauthorized access so as to maintain the data in secured manner. Research is done with respect to security aspects in data warehousing but lot to be done yet. The aim of the data warehousing environment is to handle strategic decisions; to achieve this data should be in secured manner new mechanisms and new methods are required for perfect management of the data, organization and users [2]. The current research focuses on invention of new metrics and mechanisms for the sake of strong secured systems. In this view we prepared some papers and the work was published in various national and international journals and we share our ideas in various national and international conferences. Further with the help of soft set computing and fuzzy systems and the latest trend in current data warehousing industry is hadoop technology and big data by using this combination of technologies we are moving to build a high-end security model for handling the security in data warehousing environment.

II. CATEGORIES OF DATA WAREHOUSING ENVIRONMENT BASED ON SCALE

The small scale industries with specified number of users such as 15 to 200 and the distribution capacity of the systems is with in the local area network then it is less complex to attempt a security measure for those kind of organizations. It is enough to maintain a single server which will handles the details of all users and projects, so the main concern of security is based on that server only. The authentication

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

mechanism is helpful in the handling of the security in data warehousing.

In case of medium organizations with 200 to 500 users and the distribution capacity of the systems is between Local Area Networks with some considerable complexity in the processing of the data between LANs.The best policy of maintain security for this kind of scenario is maintain security for the servers as that of small organizations along with that we need to track the data transmission with some tracing tools in such a way that if any misuse or illegal usage of the network and fraud data interpretation all these kinds of false usage need to govern so as to process the data with in medium level organizations to handle the data warehousing projects.

In case of Large organizations with 500 to 1000 capacity users and distribution capacity is wide area networks then the required security mechanisms involve the server security as that of small and medium organizations ,network tracer usage as in case of medium organization along with that to handle bulk data transfer in case of wide area networks. The usage of encryption and each time password authentication for the users who are trying to access the sensitive data in the data transfer. The categorization of users is important in large organizations to handle the data with out any false usage.

For all the above mentioned organizations the best suited method is dominating set usage[10], in this method first we need to estimate the workload of ach system participating in an activity, suppose if all the systems are common load then the work is shared equally, suppose if some of the systems are having less work to do then the work should be distributed equally to all the systems to handle this in data structures the concept AVL tree is used to balance the systems load, and it is possible to identify the system that will serve the most of the systems need such we can name as a dominating system, if such systems are available as a group it is a dominating set to process the data and to handle common activities required by the user the dominating sets are used.

In case of enterprise with number of user capacity such as greater than 1000 and distribution capacity is between various companies in different countries where they can share the same project ,and integration of the work done by different users at different locations is required. In this scenario the server security, network trace tools usage, using the encryption in the data transfer along with that one time password usage by the user while connecting with the server. The procedure is at the time of connection with the system the user need to enter the authentication to prove the identity, after that to handle the sensitive data the server will generate one password for that specific user and the same will be send to the registered users personal id, through that only the user is able to complete his task.

III. Graph Based Resource Management in Data Warehousing

In graph theory the dominating set and reachability are the aspects where we can make use with data warehousing. We are taking one scenario of existing some N systems in a network, each system is having some capacity so as to serve the user requests. Suppose if any of the system reached to its maximum capacity of serving the user requests then how to handle the situation. Another problem we are concentrating is the dominating set where is a cluster of machines/server where we can get the entire data so as to serve the user requests. The following integrations so as to achieve the above mentioned requirements.

Case 1: Establishing a query monitor (QM), the functionality of QM is it will periodically estimate the status of workloads and how many requests are waiting for a system. Suppose any system is exceeding the number of requests allocated to its capacity then the QM will look up in the resource table which consists of the following information.

System name/No	IP address	Current Load(number of hits)	Requests in queue
A	207.46.130.1	56	5
B	207.46.130.12	67	10
C	207.46.130.20	93	3
D	207.46.130.15	100	8
E	207.46.130.24	43	78

Table1: Resource Table by QM

From the above table it is clear that system D is reached to its maximum capacity. (We are assuming that number of requests processed by each system is 100) so the queue with 8 requests cannot be assigned to the system Then the QM will look up the resource table and decides that system E is having less load but it has to process the 78 requests in the queue and there is one more possibility that based on priority of the requests in the queues the QM will redirect the new requests to the system Otherwise it will go for either A or B provided the same things should be considered by the QM,the above process will be performed by the QM recursively so as to come up with one solution of subset with the possible number of systems which are chosen so as to serve the requests in the queue.

System name/No	Action Performed By QM
A	**ADJUST CURRENT QUEUE ALONG WITH SYSTEM E QUEUE**
B	**ADJUST CURRENT QUEUE ALONG WITH SYSTEM E QUEUE**
C	**ADJUST THE CURRENT QUEUE**
D	**NO ALLOCATION FOR SYSTEM D REDIRECTED TO A OR B**
E	**ALLOCATE CURRENT QUEUE AND REDIRECTS TO A,B**

Table2: Conclusion Table by QM

Procedure (Resource Monitoring)

Inputs: SYSTEMS as NODES of Graph
Current Processes as Edges of Graph

Output: Subset of systems along with allocated requests

Mechanism:

Step1: QM will traverse the Graph in BFS notation so as estimate the current workloads of the all the systems.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

Step2: The result of step1 is the above table with the details of current status and queue information.

Step3: Based on the resource table the QM will decide the outcome with possible number of systems along with the assignments of new requests based on their capacity of work load.

Step4: Assigning and monitoring the status of workloads again will be done by the QM.

Case 2: The second case we are considering is the dominating set, which we are concentrating on most of the application relevant data which is residing either in a single server/cluster through which all the systems are going to acquire their needs based on the requirement. Here we are considering the case where there is a chance of reaching the dominating set itself beyond the capacity. In that case we are making use of QM conclusion to elect a system from the list as proxy to redirect a specific module data/resources to the available system with a 2-way authentication mechanism which is in the following manner.

Procedure (Redirect Workload)

Inputs: Resource graph by QM, Dominating set

Outputs: Redirect System info to User along with authentication

Mechanism:

Step1: Estimation of server capacity so as to serve the requests made by all the systems.

a) If server is able to carry on with the work load no involvement of QM.

b) else QM will redirect the requests based on the modules and number of requests to a specific module and a mapping will be established with conclusion table with allocation details.

Step3: QM will send the system information and authentication to the authorized users so as to get the final outcome to the module that user has requested.

Step4: User will have the details of system along with the authentication information.

The following figure explain the abstract view of the resource allocation in the graphs. A hypervisor is like one monitoring tool like our QM where it can have the provision of monitoring and estimating the work load of the available systems. As per the user given constraints the QM will observe the available resources along with the CPU resources and it will construct the Resource table for the current scenario as we explained in the case1.

After that the QM will take care about the dominating set and if required the redirection of requests from the dominating set is done with the QM conclusion table as described in case 2.So finally the QM will depends on resource allocation and redirection of requests.

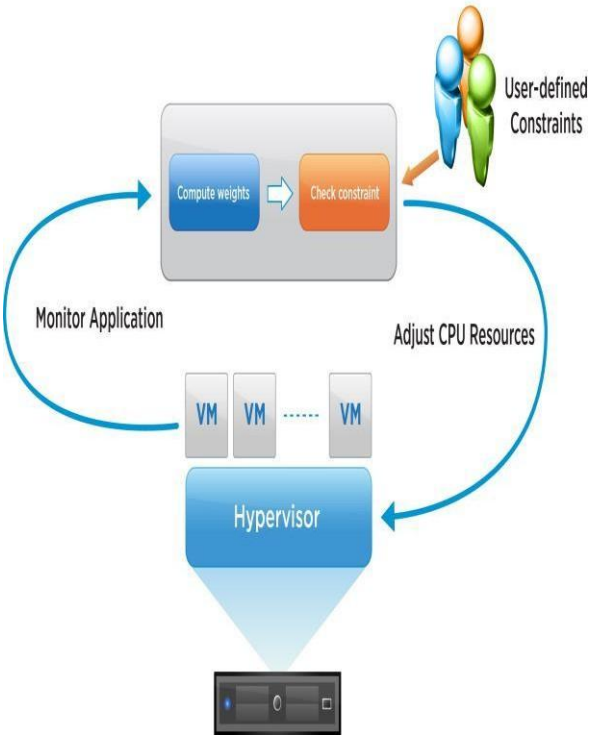

Fig 1:Resorce Allocation in Graphs

IV. FOR ENTERPRISES BIG DATA USAGE AND HADOOP TECHNOLOGY

Big data technology enables massive data aggregation beyond what has been previously possible [11]. Given the state of today's security systems, most organizations are a long way from using these types of advanced technologies for security management. Security professionals need to get more value from the data already collected and analysed. They also need a better understanding of both current issues and impending challenges related to data. Starting with a foundational set of data management and analytic capabilities enables organizations to effectively build and scale security management as the enterprise evolves to meet Big Data challenges. When dealing with –Big Data,| the volume and types of data about IT and the business are too great to process in an ad hoc manner. Moreover, it has become increasingly difficult to secure meaningful information from the data being collected. According to the Verizon Data Breach Investigations report (2012), 91 percent of breaches led to data compromise within –days| or less, whereas 79 percent of breaches took –weeks| or more to discover. The following are major needs of big data usage

- –Scaling out| rather than –scaling up|, since centralizing all this data will be practically impossible.
- Analytics and visualization tools that support security analyst specialties. Security professionals require specialized analytic tools to support their work.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

- Network forensics analysts need full reconstruction of all log and network information about a session to determine precisely what happened.

Threat intelligence to apply data analytic techniques to the information collected. Organizations require a view of the current external threat environment in order to correlate with information gathered from within the organization itself. This correlation is key for analysts to gain a clear understanding of current threat indicators and what to look for.

Security organizations today need to take a ―Big Data‖ approach. Eliminate tedious manual tasks in routine response or assessment activities.

- Use business context to point analysts toward highest impact issues. Security teams need to be able to map the systems they monitor and manage back to the critical applications and business processes they support.
- Present only the most relevant data to analysts. Security professionals often refer to ―reducing false positives.‖
- See ―over the horizon.‘ Defence against modern threats is a race against time. The system needs to provide early warning –and eventually predictive model
- Start by implementing a security data infrastructure that can grow with you. This involves implementing an architecture that can not only collect detailed information about logs, network sessions, vulnerabilities, configurations, and identities, but also human intelligence about what systems do and how they work.
- Deploy basic analytic tools to automate repetitive human interactions.
- Create visualizations and outputs that support major security functions. Some analysts will only need to see the most suspicious events with some supporting detail. Malware analysts will need a prioritized list of suspect files and the reasons why they are suspect.

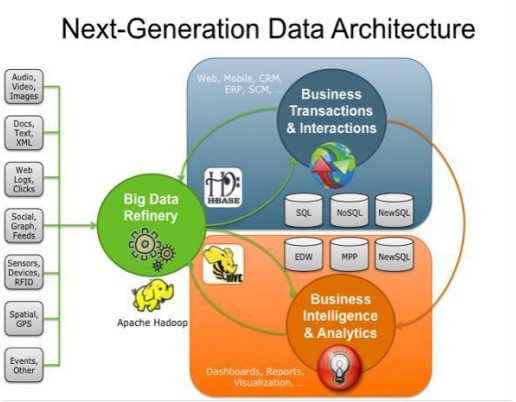

Fig. 2 Big Data Architecture

A. *Hadoop Basics*

A software framework that supports distributed computing using Map Reduce [12]

- Distributed, redundant file system (HDFS)

Job distribution, balancing, recovery, scheduler, etc.

- **Map Reduce:** A programming paradigm that is composed of two functions (~ relations)

Map Reduce

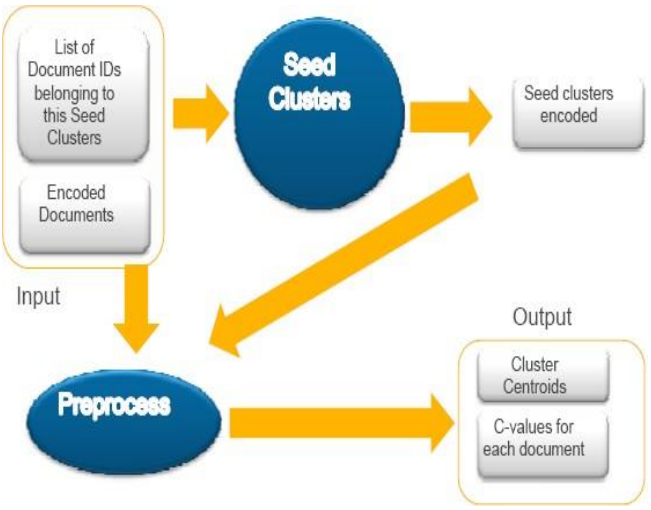

Fig. 3 Processing of Data in Hadoop

The Security of Big Data Infrastructure considers the following principles

- Evaluate the security capabilities of big data infrastructure
- Do the available tools provide needed security features?
- What security models can be used when implementing big data infrastructure?
- . Identify techniques to enhance security in big data frameworks (e.g., data tagging approaches, sHadoop)
- Conduct experiments on enhanced security framework implementations

V. CONCLUSION AND FUTURE WORK

The overall concept we discussed in this paper is starting from the small organization to enterprise the activities of data warehousing are common that is OLTP, ETL and OLAP.But the differentiation is regarding with data, whether the warehousing uses parallel and distributed environments. Depending on the level of organization the data handling mechanism will be changed.Upto the large organizations according to our view the usage of dominating sets with balancing nature of the load of

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

the system is better, where as in case of enterprises we need to move for big data analytics another important criterion is the security implementation depending on number of users and kind of activity taken in the data warehousing environment. Some security measurements and requirements are specified in the paper. The future scope and enhancement for this work is constructing a common security mechanism starting from requirements gathering up to maintenance phase, and more over irrespective of the kind of data warehousing the common security mechanism is required to implement.

REFERENCES

[1] Dr..S.L Gupta,Sonali Mathur,Palal Modi, Data Warehouse Vulnerability and Security, International Journal Of Scientific & Engineering Research Volume 3,Issue5,May-2012.

[2] J. Mohamed Salah Gouider,Amine Farahat,Building Data Warehouse National Social Security Fund Of The Republic Of Tunisia, International Journal Of Database Management Systems(IJDMS),Vol 2,No2,May 2010.

[3] S. V.M.NavaneethaKumar, Dr.C.Chandrasekhar, Security Of data Warehousing Server, International Journal Of Computer Applications, Vol-III, No.4, Dec 2010.

[4] M. Robert Winter,Olivera Marjanovic,Barbara H.Wixom,Introduction to the Business Intelligence and Data Warehousing Minitrack,45 th Hawaii International Conference On System Sciences,IEEE-2012.

[5] R. Shashank Saroop,Manoj Kumar, Comparison Of Data Warehouse Design Approaches From User Requirement to Conceptual Model:A Survey, International Conference On Communication System and Network Technologies,IEEE-2011.

[6] Stefano Rizzi,Albrt Abello,Jens Lechtenborger,Juan Trujillo, Research in Data Warehouse Modeling and Design: Dead or Alive?ACM Transactions NOV-2012.

[7] M. Veronique Limere,Aditya Pradhan,Melih Ceilk,Mallory Soldner,Warehousing Efficiency In a Small Warehouse,IEEE 2011.

[8] *Marcel Danilescu,Data Security management applying trust policies for small organizations,ad hoc organizations and Virtual OrganizationsJournal of Accounting and management,Vol 2,no.3-2012.*

[9] –Xuejian Yan,Xueqing Li,A Multidimensional Data Analysis Based on MDA for Educational Data Warehousing, The 6th International Conference on Computer Science and Education,IEEE,Aug-2011.

[10] Uma Pavan Kumar Kethavarapu,Dr.S.Saraswathi,,The Requirements of Parallel Data Warehousing Environment to Improve the Performance with Dominating sets for Next Generation Users.Intenational Journal Of Computer Science and Information Security,Vol.10,No.5,May 2012.

[11] Sam Curry, Engin Kirda, Eddie Schwartz, William H.Stewart,Amit Yoran,Big data Fuels Intelligence-Driven Security, January 2013.

[12] HDFS Architecture Guide, White Paper Apache Software Foundation 2012.

The Author Named **Uma Pavan Kumar Kethavarapu** he received his **M.Tech** from NIET, Sattenapalli affiliated to **JNTUK**, Kakinada, He is having total of 9 years of teaching experience in various levels such as lecturer, Assistant professor and Associate Professor. His research interest are **Data warehousing, Data bases, distributed and parallel systems. He is having papers published in national, international journals and attended various national and international conferences.**

The Author Named **Sreedevi.S.erady** she received her MCA from DOEACC centre, affiliated to Calicut University, Kerala, she is having total of 7 years of teaching experience in various levels such as lecturer, Assistant professor. The author is SET qualified from Karnataka. Her research interest are Data **warehousing, Data mining, and Security aspects in data bases and data warehousing, Distributed environments.**

The author named **Govinda raj pandit** MSC CS from University of Mysore, he is having total of 13 years of teaching experience. His research interests are **Network security, data warehousing, algorithms, data structures and operating systems.**