An Overview on Data Virtualization

Soshya Joshi[1]

[1]*Assistant Professor, Indira GandhiColeege of Engineering and Technology for Women Chengalpattu,*

Kancheepuram, Tamil Nadu,India

Abstract*— Data Virtualization is a process wherefrom different information sources irrespective of theirphysical location and structure , data can be abstracted . These information sources can be web services , relation databasesetc. Instead of accessing manyData Warehousesa better alternative can be Data Virtualization technology. Data virtualization is a new technology but it has the same old concept of DataFederation. Traditionally data is integrated by constructing Data Warehouse Data Mart through data storage and ETL(extract, transform, load) methods. So constructing Data Warehouse ,Data Mart takeshigh cost, maintenance ,months years, resources to develop good infrastructure for storage and high performance hardware for processing of large amount of data. Data virtualization technology use only information about meta data only , no physical location movement, data providers will be externally by reliable providers. Data Virtualization a best solution where finance is a concern but need rapid solution.

Keywords— ***Data Virtualization, Cloud Computing,Data Warehouse, Data Mining, Data Mart, Meta Data, Data Visualization, Data Federation.***

I.INTRODUCTION

This paper focuses on how data warehousing and data mining helps to store huge collection of data ,the challenges involved in collecting and sorting out same truth from different heterogeneous sources. This paper also discusses about the importance of business intelligence , decision support system and how data warehouse will be helping in this process . But data warehouse is expensive to build every time and it takes time to implement even, so what can be an alternative option when one has less time and budget. Present paper emphasis on how data virtualization helps to enhance this process and the benefits of using the concept of data virtualization in terms of cost affective, greater agility, reusability, less time for implementation.

II.DATA WAREHOUSE

Data warehouse has information collection collected from heterogeneous sources, which while storing are stored under unified schema and can be accessed from a single site.

To construct a data warehouse, processes like data cleaning , integration, transformation, loading and refreshing of data need to be done. Major advantage of building a data warehouse for decision making process is – in data warehouse data are organized around major required subjects plus historical data too are stored in summarized fashion, which in turn helps to get all type of information at one place. Data warehouse is important because all data can be achieved as data gets accumulate on daily basis transaction which need to be stored for data mining purpose. Data warehouse and traditional data base both are designed in different ways as in conventional databases we use entity relationship model whereas in data warehouse designing we use multi-dimensional model. Multidimensional dimensional model enhances performance .

Data mining's great example is web mining . It's a blend of artificial intelligent with data base, to help in decision making process which generally is useful in areas like science, business, as data mining having feature of automate extraction of information. Terabytes of day get accumulate in every day transaction to so extracting information from these huge data is the only useful thing for this reason data mining is very useful.

Data - data are raw facts , it can be in the form of numbers texts anything . computer take input in the form of data which can be categorized as operational data, non operational data, meta data.

Information-when useful relation can be figured between data means any association then those data are information. In short meaningful data is information , data can be meaning only when data having any relationship with each other and makes any sense .

Knowledge- knowledge is the finding of useful pattern from the information , as information is the association between words and when a useful pattern can be figured from all those associated patterns then it is knowledge.

For data mining process both current as well as historical data are important but keeping historical data in conventional databases will affect its performance, as huge amount of data gets accumulate in the database for each day transaction. Therefore the better

way is to send historical data to another repository , thus integrating historical data from different sources and storing at one place.

The most popular definition given by Bill Inmon is-"A Data Warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process". According to this definition subject oriented means data warehouse will be used to inspect one particular subject, for example analyzing sales subject. Integrated means data are collected from heterogeneous sources and stored at one place. Time variant means data warehouse will store data of all times such as current as well as historical data. Non-volatile means data content will never be deleted, if any updating is done then data warehouse will keep track of both updated data and historical data, but will never change or erase any data from the repository.

Unlike entity relation model datawarehouse's dimensional model is very complex and consists of fact tables and dimensional because of which retrieval is very fast. There are three architectures to make a dimensional table – star schema, snow-flake schema and fact constellation.

Another important part is meta data. Meta data is the data about data. Means it is like a dictionary which stores all the meaning , information about the data stored in data warehouse plus tells how data warehouse constructed and how it can be managed , how to use data warehouse etc. all these details will be stored in meta data.

III. Data Mining

Mining means digging example gold mining means digging gold from earth crust same way data mining means extracting useful data from large amount of data. When too much data is there it becomes difficult to analyze the useful part and take decision , here data mining plays a vital part as it reduces that searching large dataset problem , data mining provides only summarized useful information which helps decision makers to take decision easily and to a great extend accurately .Best example of data mining usage is in sales sector , as with the help of data mining checking historical data can figure out which items can likely to be purchased with each other example bread with butter , bread with milk , this way sales can be increased. Data mining is not a simple method as it involves continuous feedback loop. as while selecting the data mining technique, user will decide whether the quality of data is good or not and the result generated is satisfactory or not , based on this feedback earlier step will be repeated or sometimes process will be restarted. Data mining is a type of query processing only but an advanced query processing than traditional query processing of database , as it need to extract more deep information than regular query answers. Data mining has some phase –

i)Problem definition phase – objective of this phase is to understand the business problem , it requires the team effort of data mining expert, that business expert and domain expert. Once business problem is understood data mining problem can be well defined. Next phase is data exploration phase.

ii)Data exploration- in this phase data is exchanged between data mining experts and business analyst in order to understand the meaning of meta data . Here the data is collected analyzed in order to have deep understanding of the business. For this traditional tools like statistics is being used . Next phase is data preparation phase.

iii)Data preparation- in this phase data model is created before building mining model. In this phase only data is collected, cleaned and required formatting is done. We can say data is finalized in this phase example like filling missing values or removing or formatting data, etc. formatting is needed because some mining functions accept only particular format data only.

iv)Modeling phase: Many mining functions are available as one can use different mining functions on same data model. In this phase communication between domain expert of problem definition phase may be needed sometimes. Next is evaluation phase.

iv)Evaluation phase : one of the important phase taas in this phase experts test the model and check whether model is fulfilling all the client requirements. If not then model again has to be rebuilt . Generally more than one model will be made and all are evaluated and based on the performance model is choose as model should be able to answer like whether objectives of business is achieved , whether all issues has been take into account while building model. If model is not up to the expectations then again process not be started from beginning. Before deployment phase this evaluation should be done properly.

v)Deployment phase: The result of evaluation phase is used in this phase. After selecting the best model result can be deployed in the production environment.[1]

IV.Data Virtualization

Data Virtualization is a technique which tells us how to use data for decision management without knowing any background details about the data. Background details include technical details like formatting of data , physical location details etc. Oscillates virtualization is everywhere, as it saves lot of expenses. One can use virtualization concept in operating system, database management, networking, processors. Data virtualization provides an abstract interface where user can access data. The user can use data for retrieval purpose , manipulation purpose, data entry etc. In all these purposes user need not to know where the physical location of data or where is the database servers running, or the language of database . As in many organizations multiple database management systems are being used like SQL server , Oracle so it's challenging to interpret different data from different types , integrating data and storing data. Cloud computing is one of the recent technology where data virtualization concept has been implemented. Data Virtualization is a process where from different information sources irrespective of their physical location and structure , data can be abstracted . These information sources can be web services , relation databases etc. Instead of accessing many Data Warehouses a better alternative can be Data Virtualization technology. Data virtualization is a new technology but it has the same old concept of Data Federation. Traditionally data is integrated by constructing Data Warehouse Data Mart through data storage and ETL(extract, transform, load) methods. So constructing Data Warehouse , Data Mart takes high cost, maintenance , months years, resources to develop good infrastructure for storage and high performance hardware for processing of large amount of data. Data virtualization technology use only information about meta data only , no physical location movement, data providers will be externally by reliable providers. Data Virtualization a best solution where finance is a concern but need rapid solution.

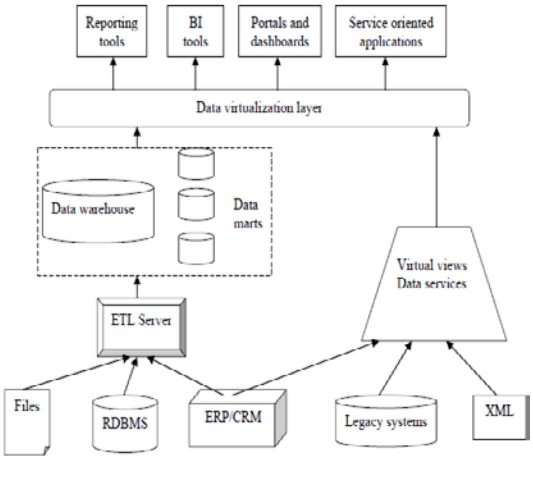

Fig.1 Data Virtualization

Advantage of using data virtualization are:-

- Data Virtualization helps in accessing data sources logically from a single place.
- It saves time and money , which will be spend on constructing data warehouse.
- One can easily and quickly access information.
- Security and flexibility as data servers will be accessed only when use requests.

V. Data Visualization different from Data Virtualization

Data visualization is visual presentation of the result which is captured after analysis. This information presentation can be in the form of pie chart , bar graph etc . , with the help of visualization result can be communicated more easily. Tableau, click view are some of the software which helps to do data visualization . Three approaches of data visualization are one way is applying visualization techniques to the information extracted from data mining. In this approach visualization can be in the form of cluster or correlation and applied directly on data mined from data bases. So first data is mined then visualization techniques are applied.

Second approach is first visualization techniques are applied over data then on that data mining techniques are applied. Here assumption is that it is easier to apply mining techniques on visual form as these forms won't be as complex and large as the conventional databases ,thus will increase efficiency but issue is loss of data. While visualizing it is believed there can be some loss of information when data is visualized from databases over than when mining techniques are applied it will be on incomplete data.

Third approach is to use visualization techniques just as an compliment to data mining means some data will be mined some data will be visualized if needed for data mining.

XII.Conclusions

It's the dream of ever manager to take quick efficient decisions in less cost and in less time. To build a decision support system for oscillates accumulating dynamic information is challenging,in that case data virtualization offers interesting features like less cost , less implementation time , reusability [4].

Acknowledgment

I would like to express my special thanks of gratitude to my mentors who gave me the golden opportunity to do this work on the topic An Overview on Data Virtualization ,which helped me to gain lot of knowledge.

References

[1] Ali Radhi Al Essa ,Bach , Christian, , University of Bridgeport, “Data Mining and Warehousing”,ASEE 2014 Zone I Conference, April 3-5,2014,University of Bridgeport, Bridgeport,CT,USA.

[2] Dr.Georges Grinstein, Dr.BhavaniThuraisingham, The MITRE Corporation, Bedford ,”Data Mining and Data

Visualization: Position Paper for the Second IEEE Workshop on Database Issues for Data Visualization”.

[3] Vijay Raj Singh, ShobhanaBansal, TCS, “ White Paper: Data Virtualization: Enabling Next Generation Business Intelligence”.

[4] Ana-Ramona BOLOGA, Razvan BOLOGA ,ACADEMY OF Economic Studies,Bucharest, Romania,“A Perspective on the Benefts of Data Virtualization Technology”, InformaticaEconomica vol.15, no.4/2011.

[5] Saas , the power to know, Big Data,[Online]. Available : http://www.sas.com

[6] Twelve Key Reasons to Use Composite Data Virtualization, Composite Software, January 2010, http://purl.manticoretechnology.com/ImgHost/582/12917/2011/ Document_ Down-loads/12Reasonsfor CompositeDV.pdf

[7] Katina Michael , University of Wollongong and Keith W. Miller , University of Missouri–St. Louis,” Big Data: New Opportunities and New Challenges”, June 2013 (Vol. 46, No. 6) © 2013 IEEE Published by the IEEE Computer Society