Cloud based Data Warehouse 2.0 Storage Architecture: An extension to traditional Data Warehousing

Kifayat Ullah Khan[1],Sheikh Muhammad Saqib[1], Bashir Ahmad[1], Shakeel Ahmad[1] and Muhammad Ahmad Jan[1]
[1]Institute of Computing and Information Technology Gomal University, D.I.Khan, Pakistan
kualizai@hotmail.com

Abstract—**Data Warehouse plays a critical role in organizational decision making. This gigantic environment provides an extremely conducive decision making grounds provided its smooth operation and maintenance is guaranteed. The recent enhancements in the Data Warehouse Architecture, DW 2.0, provide an approach which organizes the data into separate physical boundaries. The maintenance of such a massive data repository is really a hard nut to crack. If this maintenance responsibility is followed a divide and conquer approach in terms of splitting these sectors of data over the cloud then a maximum desirable output can be ensured. The authors in this research have given a model which explains how the data should be split, what data should be available on local and over the cloud storage. In this way, the organizational management can ponder maximum concentration on data analysis and stay trouble free from focusing on non-functional requirements.**

Keywords- *Data Warehouse, , Maintainance, DW 2.0 and Cloud Computing*

I. INTRODUCTION

DW (Data Warehouse) plays a significant role in strategic decision making. More the size of data, the more think tanks can understand how the things have shaped up in the past and what should be the future strategy. The need for more storage of data in the DW is proportional to the rate of incoming data from various data sources. The classical architecture of DW suggests the storage of data in de-normalized format using traditional entity-relationship modeling [2]. Similarly the storage techniques defined by Ralph Kimball differs from that of above with respect to data modeling and emphasizes on Star Schema [1].

In the industries like Telecommunication where the daily incoming data size is very large, the way data is physically organized matters a lot. It was observed that some portion of data in the data warehouse is analyzed frequently but at the same time, a portion of data is not touched at all or it comes under use very rarely. In such a scenario, the data model of DW 2.0 is very useful. DW 2.0 follows an approach which organizes the data in various physical storage sectors and the criteria to organize the data within such physical boundaries, is the age of data.

When the size of available data is too high and it is growing at a rapid pace, the cost to maintain it, is also proportional. Various cost attributes are the data storage mediums, DBAs, System Administrators, power etc. It sometimes gives a feeling that the full concentration should be emphasized on exploring the data for better decision making rather than a portion of time is spent on the issues related to its maintenance and operations.

Cloud computing is an ideal platform and solution in such circumstances where an organization wants their all energies to be devoted towards data analysis and sound decision making and has no time to worry about the operations/non-functional requirements. Cloud computing aims to take away all the pains related to storage and maintenance of data in this particular scenario. It further aims to facilitate the community by providing a variety of conveniences which are in the shape of Software-as-a-Service (SaaS), Platform-as-a-Service (Paas) and Infrastructure-as-a-Service (Iaas).

As the size of data is growing at an alarming rate in various organizations, so that's why the importance and dependence on decision support systems i.e. DW is increasing rapidly. The possibility that an organization has certain limitations/restrictions for in-housing such DSS systems, does exist. In such a case, maximum benefit can be achieved utilizing the features of Cloud Computing. In this regard, Stephen Russell et.al have studied that the cloud based decision support applications can yield correct decision outcomes if the availability of computing resources is guaranteed. Though the major emphasis of the authors is on the availability of computing resources but their study strongly supports cloud computing for decision support systems [10].

In a research study carried out by Tuncay Ercan, an effective utilization of cloud computing is explored in educational institutions. The author has first made it realize that the ratio of the use of cloud computing in educational sectors is quite low to that of financial services, business and management services and manufacturing and telecommunication services. The author has then proposed a model which aims to benefit various people and their needs in the academic institutions [5].

In a similar research work, the importance of cloud platform is illustrated by floating an idea where desktop based custom software are emphasized to be replaced by cloud based custom software in order to get the maximum possible output from the application by utilizing the key features of cloud computing. This clearly supports the cloud platform for application deployment as it varnish's the shortcomings of desktop applications [4].

Similarly the work of Daniel J. Abadi strongly recommends the cloud based architecture for decision support

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

systems. Sound logic is elaborated for using these grounds for analytical data management like the utilization for shared-nothing architecture, painless environment to easily maintain the 'A', 'C', and 'I' (atomicity, consistency, and isolation) of ACID properties for database transactions, and valid arguments not to store the particular sensitive data on cloud and thus keeping it out of analysis. The author also gives a number of sound reasons for not using the cloud environment for transactional data management [6].

The authors in this research consider Cloud Computing with respect to DW 2.0 and understand that it has a vital role to play. For example, when an organization either is not ready to take the pain of maintenance of data or due to financial limitations or for any other reason, it can utilize cloud computing in such a way that it can store the smaller portion of data locally and store the major bulk on the cloud. This strategy enables the administrators of DW to be answerable for a minimum portion of data. And the bulk of data is stored on the cloud, taking away all the worries of storage constraints, back-up and recovery of huge sized data, disaster management, and cost of retaining very seasoned professionals for its non-functional requirements.

II. Data Warehouse

DW is a central repository which stores the entire organizational data under one roof. This data is stored separately from the transactional/operational data as its main goal is to support database transactions rather querying and analysis. On the other hand, DW aims to provide an ideal platform for efficient querying and reporting [1]. The data stored in the DW has four key characteristics i.e. it is subject oriented, integrated, time-variant and non-volatile. These attributes enable the data to be in a format which is highly desirable and useful for data analysis and ultimately for decision making [2].

A. Data Warehouse 2.0

DW 2.0 is a model which emphasis a number of areas which were given not much significance in its predecessor. It aims to throw light on complete life cycle of data, focuses more on un-structured data in order to get maximum out of it, emphasizes much on the significance of metadata, tries to make use of it up to a maximum possible level, provides and introduces an option to provide online updates in the data which was never a chance in its earlier version.

DW 2.0 is defined as the "Architecture for the Next Generation of Data Warehousing". The Architecture of DW 2.0 relies on the four life cycles of data which can be termed as the data segregated into four physical sectors. The Interactive Sector, first of four, serves almost like operational system, contains real time data which can be updated and contains data for around 1 day period. The Integrated Sector resembles much with the earlier version of DW 2.0, contains detailed enterprise atomic data, ranging from 1 day time period to almost 2 years time. Unstructured data is stored in structured format. The Near Line Sector exists to reduce the volume of data and to increase the response time of the Integrated Sector. It provides an option to use cheaper medium of storage as the volume of data is too high, containing almost 3 to 4 years of data. The role of Archival Sector is also to reduce the volume of data in the Integrated Sector, cheaper medium of storage can utilized as the probability to access this data is extremely low but it does not mean that this data is no more needed. Data is a vital asset for the organization. It almost stores last 5 to 10 years of data [3].

III. Cloud Computing

Cloud computing is basically an idea which involves the maintenance of data and applications using remote server. It enables its users to use applications and manage their data on such computing resources which are remotely available and are administered by some responsible authorities. Cloud computing provides highly efficient computing. Yahoo, Gmail and so on is prime example of cloud computing [7]. Cloud Computing enables to utilize the highly reliable hosted services which are accessed using internet. It involves managing data and applications (which behave like services) hosted remotely. Cloud computing is based on three major directions which are SaaS (Software as a service), PaaS (Platform as a service) and IaaS (Infrastructure as a service). One can utilize SaaS as a front end tool. PaaS provides a platform to create various applications which are remotely accessed. While IaaS provides virtual server and memory. All those applications which are based on cloud computing are highly optimized, efficient, reliable and easy to use. It eases the life of novice users and results in increased efficiency and reliable communication medium [8] [9].

IV. CDSA (Cloud based Data Warehouse 2.0 Storage Architecture)

With the ever increased popularity of DW systems for better strategic decision making, at the same time keeping in mind a number of overheads involved in establishing such a critical and massive data stores, cloud computing has a vital role to play. The authors in this study have proposed a model, named CDSA, which makes its intended users to extract maximum out of both worlds. The architecture of DW 2.0, in authors view, is extremely a better fit for CDSA. The storage layout of DW 2.0 is already in the shape of various sectors, which makes it easy to split them in order to follow a de-centralized approach.

Establishment and maintenance of DW is not an easy task. It does require the supervision of very seasoned Database Administrators, System Administrators, extraordinary data storage, appropriate and up-to-date data security mechanisms, supreme database tuning approaches and so on. It, sometimes, gets very difficult for an organization to stay on top in all above mentioned areas otherwise they will have to compromise on certain key facts. In order to stay on high notes i.e. make better decisions and stay trouble free from the establishment and maintenance of such a gigantic environment, authors emphasize on CDSA which provides all the time in the world to the high ups for data exploration and let the Cloud take the pain of massive data. The CDSA comprises of LR (Local Repository) and CR (Cloud Repository) as shown in Figure1.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Integrated, Near Line and Archival Sectors respectively. And this is the supreme bulk which is extremely crucial for the organization, an ultimate source for decision making. Thus its

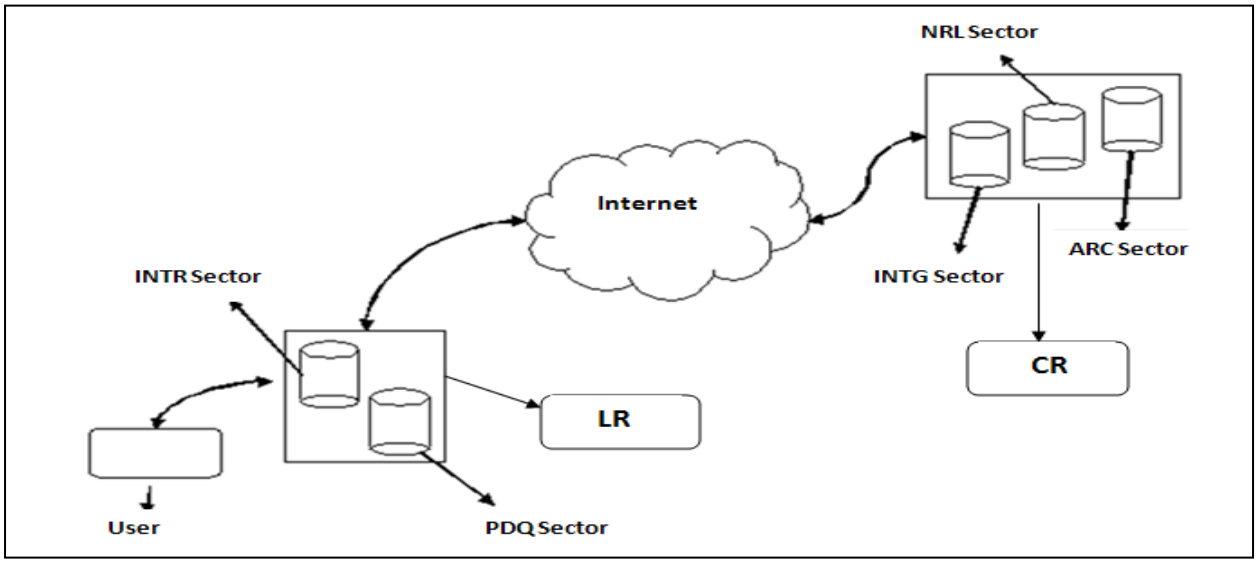

Figure1. Architecture of CDSA

A. *LR (Local Repository)*

LR in CDSA is comprised of two components i.e. the INTR Sector (Interactive Sector - one of the four sectors of DW 2.0 architecture) and PDQ (Pre-Defined Queries) Sector. The vital reason to have Interactive Sector in LR is that it consists of minimum data whose life is nearly equal to one day so that the less seasoned personal equipped with less powerful computing resources, be answerable to minimum amount of data. The Interactive Sector then supports any changes desired to be made to the data and provides reporting on current state of the data. The PDQ Sector is a new idea floated by the authors and is emphasized to be established in order to provide extremely quick response to the pre-defined queries whose sole purpose is to serve daily, weekly pre-defined reporting and routine queries. One of the main deriving forces for PDQ Sector is QUERY REWRITE which caches the data for all those similar queries who are frequently executed. So it provides a space where materialized views can be accommodated. In this way, it gives a benefit of rather to scan whole data each time the same queries are needed to be executed.

B. *CR (Cloud Repository)*

According to this research the CR is the main stakeholder. It has the charge to maintain the major bulk of data. It comprises of INTG (Integrated), NRL (Near Line) and ARC (Archival) Sectors which are administered by the cloud. CR is responsible to manage the data whose age ranges from one day to two years, two to four years and five to ten years in management and administration duties are taken up by the cloud. When no such responsibility is on the local side, the management of the organization has little to worry about the non-functional requirements and free from the establishment of an extremely massive and grand setup which consists of highly powerful computing and storage devices etc.

V. RESULTS AND CONCLUSION

The authors in this research have given practical directions as how DW 2.0 can be implemented on a cloud environment and what can be achieved using this model/technique/strategy. There are a number of advantages which can be achieved using this extended architecture. Following are some the advantages and comparisons of CDSA and other DW implementation approaches.

(i) Establishing and setting up a DW is a decision which is not the end of story. Instead the smooth operations and maintenance is really a cumbersome task. CDSA is really an opportunity which eases the life of personal associated with its smooth working.

(ii) DW 2.0 architecture can easily be implemented using CDSA. DW 2.0 is already divided into four sectors so it will not bring much trouble in separating them and hence placing these sectors on different physical locations as CDSA proposes.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(iii) Any DW needs highly professional and experienced DBAs (Database Administrators) and SAs (System Administrators) who are of course are not easy to retain and are expensive. By CDSA, an organization will not have to bother much about it. Because a minimum amount of data will be available on the local side which can easily be administered by any DBA and SA of ordinary experience.

(iv)The amount of data in DW can grow to immense level. Its can rise from Gigabytes to Terabytes and Petabytes. In traditional DW, an organization has to purchase and setup a massive storage space at the beginning but CDSA enables the top management to pay as much as data is stored. There is certainly no need to purchase massive data storage at the initial days of DW.

(v) There is certainly no need to buy servers of extremely high specifications as the bulk of data will be housed and administered on the cloud using CDSA.

(vi)When data storage is involved, a tight and high level of security of data must be ensured. CDSA provides an opportunity to house almost all of the data on the cloud, an organization does not have to worry much about it.

(vii) It is an obvious objection about the privacy of the sensitive data in DW, but as CDSA does provide local storage allowing a small portion of data will be placed in the premises of the organization which can easily manage very confidential data. This issue will have no grounds then.

(viii) CDSA introduces a PDQ Sector on the local side which will be the source for the pre-defined reporting and materialized views. It will store the amount of data which will be suitable for all the pre-defined and frequently executed queries. It will be of a size which can easily be administered by the less seasoned DBAs and SAs.

(ix) Less effort will be spent on the database backup and recovery as very small amount of data is stored on local side.

(x) CDSA will not bind any one to stick to a specific DBMS and software. Local side can go with any DBMS and software and cloud side can live with its own will.

The authors are in a strong view that following CDSA, the main stakeholders of an organization will be in a better frame of mind to devote their full energies and concentrations on data explorations rather than attending the meetings and getting reports or findings that on such and such timings, DW was not available or will not be available due to surprise problems of the infrastructure. They will not be hearing the complaints at all that they have run out of physical storage and so on.

REFERENCES

[1] R.R. Kimball, M. Ross, The Data Warehouse Lifecycle Toolkit, 2nd edition, Wiley, 2002.

[2] W.H. Inmon. Building the Data Warehouse. John Wiley, 1993

[3] W.H. Inmon, D. Strauss, G. Neuschloss. DW 2.0. The Architecture for the Next Generation of Data Warehousing. Morgan Kaufmann, 2008

[4] Sheikh Muhammad Saqib, Custom Software under the Shade of Cloud Computing, IJCSIS) International Journal of Computer Science and Information Security,Vol. 9, No. 5, May 2011

[5] Tuncay Ercan, Effective use of cloud computing in educational institutions, 2010

[6] Daniel J. Abadi, Data Management in the Cloud: Limitations and Opportunities, 2009

[7] Carnegie Mellon M.S , The Basic Concepts of Cloud Computing, http://www.articlesbase.com/databasesarticles/the-basic-concepts-of-cloud-computing-1284980.html, 2009

[8] Cloud computing,http://www.wikinvest.com/concept/Cloud_Computing

[9] Sathishkumar, Cloud Computing: Best Practices,http://www.tech2date.com/cloud-computing-bestpractices.html, 2011

[10] Stephen Russell, Victoria Yoon, Guisseppi Forgionne, Cloud-based decision support systems and availability context: the probability of successful decision outcomes, 2010

http://sites.google.com/site/ijcsis/
ISSN 1947-5500