Data Warehouse as a Service (DwaaS)

Sreekanth Tangudu[1], Velagapudi Sreenivas[2], K.V.D Kiran[3]

IV/IV B.Tech, Department of CSE, K L University, Andhra Pradesh, India
[1] tsreekanth1699@gmail.com
Asst Professor, Department of CSE, K L University, Andhra Pradesh, India
[2] velagapudi@kluniversity.in
Asst Professor, Department of CSE, K L University, Andhra Pradesh, India
[3] kiran cse@kluniversity.in

***Abstract[1][2]*- Our everyday data processing activities create massive amounts of data. Today, with data available in many different formats, from varying sources, and via diverse access methods, companies need a solution for managing business information. Data warehousing is must for running an enterprise of any size to make intelligent decisions. It enables the competitive advantage. Data warehousing is essentially tells about data and its relationships and it is foundation for Business Intelligence(BI).Cloud Computing has emerged as a new paradigm for hosting and delivering these services on demand over the internet. This paper discusses the concept of hosting data warehouse on cloud and providing it as a service.**

***Key words* – data warehouse, Cloud Computing, Data Warehouse as a Service, Business Intelligence, ETL Process.**

I. INTRODUCTION

In recent times there is rapid and agile development in the field of computing and network and the processing power has increased tremendously and the success of the internet boosted up the development of computing platforms which paved new ways of computing in the field of computers. [3]

There cloud computing deployment models can be categorized into three types, in fact they can also be called as various types of cloud computing. They are

- Public Cloud
- Private Cloud
- Hybrid Cloud

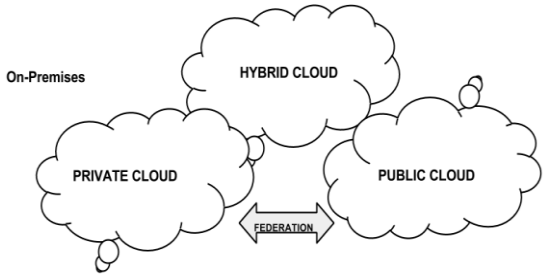

Fig 1. *Types of Cloud Computing*

In cloud technology the information is shared from clients to the organization through the virtual data centers. This virtual data centers has all the required information. The cloud technology model includes:[4]

- SaaS (Software as a service)
- PaaS(Platform as a service)
- IaaS(Infrastructure as a service)

This is stated as in Venn diagram as,

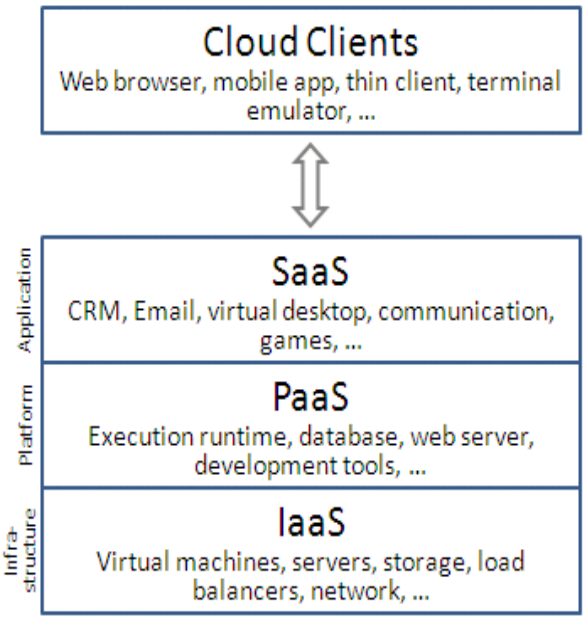

Fig 2. *Overview of Cloud Computing Services*

Whereas Data warehousing [5] refers to the combination of many different databases across an entire enterprise used to store the information and generate the query regarding the required data. The sources will help to access the information, save downloads & update the information viz. suppose your one file has some data and you want to add

same data or updating the data in a file, so these sources will help you. The characteristics of a data warehouse as set forth by William Inmon:[6]

- Subject Oriented
- Integrated
- Nonvolatile
- Time Variant

II. PROVIDING THE FEATURES OF DATA WAREHOUSE THROUGH CLOUD

In this article we propose an architecture explaining how to extract the data from various sources of the enterprise at different sites and transform the data and to load into the target server and also how to create, publish and manage the data in data warehouse. Also how a data warehouse can be delivered as a service to the client using a cloud computing model.

III. ETL PROCEDURE

In computing, Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that involves:

- Extracting data from outside sources[7]
- Transforming it to fit operational needs, which can include quality levels
- Loading it into the end target (database, more specifically, operational data store, data mart or data warehouse).

3.1 Extract:

The first part of an ETL process involves extracting the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as through web spidering or screen-scraping.

3.2 Transform:

The transform [8] stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database.

3.3 Load:

The load phase loads the data into the end target, usually the data warehouse (DW). Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information, frequently updating extract data is done on daily, weekly or monthly basis. Other DW (or even other parts of the same DW) may add new data in a historical form, for example, hourly. As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

3.4 Virtual ETL Procedure:

As of 2010 data virtualization [9] had begun to advance ETL processing. The application of data virtualization to ETL allowed solving the most common ETL tasks of data migration and application integration for multiple dispersed data sources. So-called Virtual ETL operates with the abstracted representation of the objects or entities gathered from the variety of relational, semi-structured and unstructured data sources. ETL tools can leverage object-oriented modeling and work with entities' representations persistently stored in a centrally located hub-and-spoke architecture. Such a collection that contains representations of the entities or objects gathered from the data sources for ETL processing is called a metadata repository and it can reside in memory or be made persistent. By using a persistent metadata repository, ETL tools can transition from one-time projects to persistent middleware, performing data harmonization and data profiling consistently and in near-real time.

.

IV. MANAGING STORED DATA

How do we deal with stored data? We need effective strategies to handle the stored Data warehouse data. To provide guideline for stored data management we turned to the techniques used in real life. In this paper we will discuss about how to manage Data warehouse with cloud computing. Data warehousing over Cloud computing has recently

emerged as a compelling paradigm for managing and delivering services over the Internet. In a data warehouse all data from an organization can be brought together in one place. Data warehouse systems require a different kind of database. Data warehousing systems are also used for complex analytics involving huge amounts of data (OLAP or online analytical processing). Data warehouse may support OLAP (Online Analytical processing) tools, allowing the decision maker to navigate the data in the data warehouse.

V. ARCHITECTURE

5.1 Steps to create data warehouse over cloud computing

- Develop the API which is used for data extraction from the various sources.
- After extracting the data then transform the data in various dimensions on the server.
- The data is loaded into the data warehouse.
- Ready for usage by OLAP and data mining tools.
- Analyze the data In addition to one data warehouse where all data come together, an organization may also choose to use data marts which carry only past data of the data warehouse. Data marts are more specialized and therefore easier to deploy.
- A data mart is set up by a single department or division within an organization for a single purpose.
- Then quickly implement a needed system, without affecting or changing the existing data warehouse.
- We move towards the cloud is just as relevant for data marts as for data warehouses.

  The fundamental requirements of a data mart are simpler and human resources to be managed on the department level.

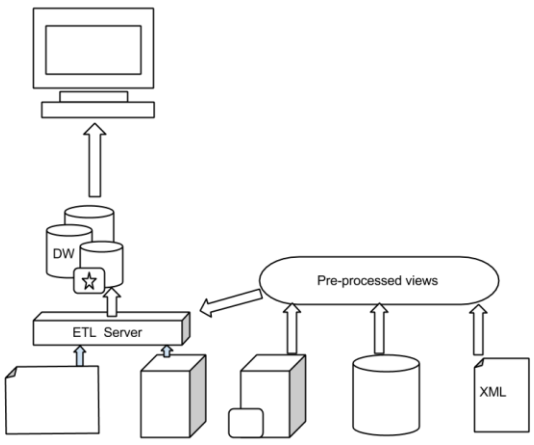

Fig 3. *Data warehousing over Cloud Computing*

5.2 Managing Data warehouse and applications over Cloud

The vast majority of SaaS solutions are based on a multi-tenant architecture. With this model, a single version of the application, with a single configuration (hardware, network, operating system), is used for all customers ("tenants"). To support scalability, the application is installed on multiple machines (called horizontal scaling). The data warehouse resides in the datacenter and the queries and the managing the data is done by the service provider. Analytic applications such as EII products enable loose coupling between homogeneous-data consuming client applications and services and heterogeneous-data stores. Such client applications and services include Desktop Productivity Tools (spreadsheets, word processors, presentation software, etc.), development environments and frameworks (Java EE, .NET, Mono, SOAP or Restful Web services, etc.), business intelligence (BI), business activity monitoring (BAM) software, enterprise resource planning (ERP), Customer relationship management (CRM), business process management (BPM and/or BPEL) Software, and web content management (CMS) can be applied over these data warehouse.

VI. PRATICAL IMPLEMENTATION OF OUR PROPOSAL

In this chapter we will present the implementation details for our proposal. The following chapter contains various features that are provided by our prototype.

6.1 Data Manipulation

6.1.1 Enterprise Information Integration

Data sharing is done at company level through the first stage. This represents the different data sources that feed data into the data warehouse. The data source can be of any format -- plain text file, relational database, other types of database, Excel file, etc., can all act as a data source. All these raw data are integrated at the application sever by providing various API's to upload these data.

6.1.2 Restoration of Data

Depending on the different parts constituting the information, the second part is to collect the data from the uploading sites and reproducing the original data. This step is done at company level.

6.2 Performing ETL Operations

There are several layers of operations involved in the ETL process. The following chapter contains the procedural way to carry out these ETL Operations.

.

.

Data Extraction Layer

Data gets pulled from the data source into the data warehouse system. There is likely some minimal data cleansing, but there is unlikely any major data transformation. There are the following methods of physical extraction:

- Online Extraction
- Offline Extraction

Online Extraction

The data is extracted directly from the source system itself. The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner (for example, snapshot logs or change tables. With online extractions, you need to consider whether the distributed transactions are using original source objects or prepared source objects.

Offline Extraction

The data is not extracted directly from the source system but is staged explicitly outside the original source system. The data already has an existing structure (for example, redo logs, archive logs or transportable table spaces) or was created by an extraction routine.

We will consider the offline extraction policy with full or incremental extraction depends on the user.

Staging Area

This is where data sits prior to being scrubbed and transformed into a data warehouse / data mart. Having one common area makes it easier for subsequent data processing / integration.

ETL Layer

This is where data gains its "intelligence", as logic is applied to transform the data from a transactional nature to an analytical nature. This layer is also where data cleansing happens. The ETL design phase is often the most time-consuming phase in a data warehousing project, and an ETL tool is often used in this layer.

Data Storage Layer

This is where the transformed and cleansed data sit. Based on scope and functionality, 3 types of entities can be found here: data warehouse, data mart, and operational data store (ODS). In any given system, you may have just one of the three, two of the three, or all three types.

Data Logic Layer

This is where business rules are stored. Business rules stored here do not affect the underlying data transformation rules, but do affect what the report looks like.

Data Presentation Layer

This refers to the information that reaches the users. This can be in a form of a tabular / graphical report in a browser, an emailed report that gets automatically generated and sent every day, or an alert that warns users of exceptions, among others. Usually an OLAP tool and/or a reporting tool is used in this layer.

Metadata Layer

This is where information about the data stored in the data warehouse system is stored. A logical data model would be an example of something that's in the metadata layer. A metadata tool is often too used to manage metadata.

These are the different layers of ETL operations done at the server and then the data is uploaded to the data center which hosts the data ware house.

VII. ADVANTAGES OF DATA WAREHOUSE ON CLOUD COMPUTIING

Here are the key benefits of a data warehouse once it's launched in cloud.

- For businesses struggling to manage their data, the cloud can provide a low-cost alternative to investing in infrastructure to manage it all on their own sites.
- A combination of potential benefits related to cost and manageability, combined with concerns about security and data transfer.

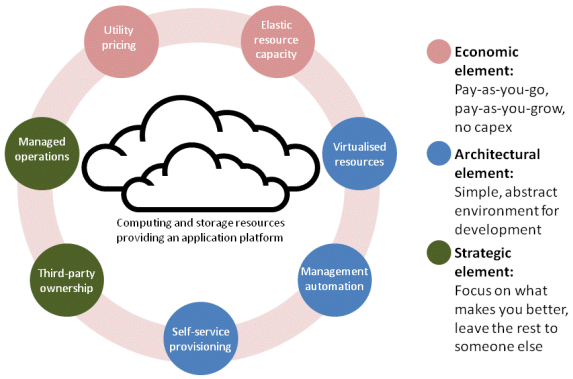

Fig 4. *Benefits of data warehouse in cloud.*

- Data warehouses have traditionally been defined as customized data storage services that aggregate data from multiple different sources and collect it in a central location to be able to run reports and queries of it.

- One issue with data warehousing is many times this is highly critical, proprietary information that some may be reluctant to ship off to a cloud provider.

VIII. CONCLUSION

Cloud computing is one of the newest and most innovative platform choices to come along in years. This is where numerous hardware servers and other resources are pooled and virtualized, so that they can be freely allocated to applications and software platforms as resources are needed. This enables software applications to dynamically scale as workloads increase. As the workload of an application decreases, resources are freed up for use by other systems. Furthermore, dynamic allocation and reallocation give the provider of the cloud fuller utilization of server resources, with less administrative work, as compared to traditional data center approaches. Although cloud computing and similar virtualization techniques are firmly established with operational applications today, they're just now starting to be used as DW platforms of choice. The dynamic allocation of a cloud is useful when the data volume of the warehouse varies unpredictably, making capacity planning difficult. According to users TDWI[10] interviewed for this report, a data warehouse (or analytic database) in a cloud is often set up for analytics, to accommodate the large and unpredictable volumes of data that business analysts and other power users collect, analyze, and then archive. But TDWI has also encountered users who have moved their entire enterprise data warehouse to a cloud.

TDWI presented survey respondents with a long list of options for next generation data warehouse platforms. These options include a mix of vendor-oriented product features and product types, as well as user-oriented techniques and best practices. The list includes options that have arrived fairly recently (clouds, SaaS, open source, appliances), have been around for a few years but are just now experiencing broad adoption (real-time data warehousing, advanced analytics, MDM, MPP), or have been around for years and are firmly established (DBMSs designed for transaction processing, SMP, EDWs). After all, generational change (when managed well) addresses features and techniques based on business requirements, not the vintage or novelty of available options.

It is a familiar and proven revenue model to give away the razor and charge a little bit extra for the razor blade. Technology lock-in! It is an easy prediction to make that something like that will occur once the computing model has been demonstrated to be scalable, reliable and popular.

IX. REFERENCES

[1] Journal of Internet Services and Applications, "Cloud computing: state-of-the-art and research challenges", Page. No 1

[2] Vishal Jain1 and Mahesh Kumar Madan 2, "Information retrieval through Multi – Agent System with Data Mining in Cloud Computing", "International Journal of Computer Technology and Applications, Volume 2 :Issue 4, Page 1,2012

[3] K.A. Delic & J.A. Riley, "Enterprise Knowledge Clouds," Next Generation Km Syst. Int. Conf. Inform., Process, Knowledge Management, Cancun, Mexico, pp. 49–53, 2009.

[4] Sreenivas V &Dr C Narasimham "Enhancing the Security for information with Virtual Data Centers in Cloud" Springer lecturer notes

[5] "The Bottom-Up Misnomer". 2003-09-17. Retrieved 2012-02-14.

[6] Abdullah, Ahsan (2009). "Analysis of mealy bug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool, Volume 69, Issue 1". Computers and Electronics in Agriculture 69: 59–72. doi:10.1016/j.compag.2009.07.003.

[7] Inmon, Bill (1992). *Building the Data Warehouse*. Wiley. *ISBN* 0-471-56960-7.

[8] Kimball, Ralph (1996). *The Data Warehouse Toolkit*. Wiley. ISBN 0-471-15337-0.

[9] Hillard, Robert (2010). *Information-Driven Business*. Wiley. ISBN 978-0-470-62577-4.

[10] Next generationData Warehouse Platforms By Philip Russom. TWDI Best Practises

[11] "An overview on data ware house." - Oracle Solutions www.oracle.com/solutions/091023.pdf