Data warehousing in the cloud

Business

In the recent times, the need to store and access information has become a vital aspect for success of individuals and businesses across the globe (Patil and Rao, 2011). It is therefore crucial for the organizations to employ business tools, which will enable them to succeed both in the short and long run.

“ Vertica Systems Inc.” (2008) argues that since the beginning of the 21st century, the amount of information recorded as well as stored in various parts of the globe has increased tremendously. This can be attributed to the issue of globalization, which has enabled individuals and companies to share information regardless of their geographical location (Kimball and Margy, 2002). Some of the mostly used data formats include relation databases engines, usually interconnected through the Intranet, and recently the WWW (World Wide Web) sites connected via the Internet (Gelder, 2010). “ Vertica Systems Inc.

” (2008) indicates that the high level of interconnectivity of the data sources offers opportunity for firms to access large amount of information spread through numerous sources of data. It is notable that there are several modern applications that could enormously benefit from the available information. They range from a variety of domains like BI (Business Intelligence), which is a trade analysis, leisure, libraries, education and science among others. The most striking feature of all these notable applications is that they depend on data having multiple origins and types in order for them to function properly. Chaudhuri (1997) indicates that this requirement demands for an integrating tool allowing this kind of application, thus making effective use of the diverse data.

This can be done by supporting the browsing as well as querying of all tailored subsets of information. Gelder (2010) argues that in contrast to some types of information integration such as the demand approach, data warehousing has offered one of the most viable alternative solutions to the traditional approaches. This paper will critically look at data warehousing and how it is being carried out in the cloud. Data Warehousing As indicated above, the ability to store and access information has remained one of the key ingredients for the success of most individuals as well as companies globally (Chaudhuri, 1997). It can be noted that in 1990s many organizations of scale started to require data timely to enhance sustainability.

Consequently, they established that most of the traditional systems of information technology were highly ineffective to offer them relevant data quickly and effectively. It was from this challenge that data warehousing came up. This is due to the fact that crucial data could be used in completing strategic reports for management among other usages of information. Patil and Rao (2011) define data warehousing as a single, consistent, and complete storage of data, which can be obtained from various sources. This data is made available by numerous users in ways they can clearly understand thus being able to use the information in several business contexts.

“ Vertica Systems Inc.” (2008) argues that data warehousing is an integrated, subject-oriented, non-volatile as well as timely-varying collection of data, which is used for decision making in organizations. Furthermore, it can be noted that data warehousing can be used to save crucial information from loss in unforeseen accidents like mass data corruption, bomb blasts,

and natural calamities. As indicated by Patil and Rao (2011), designing a data warehouse demands one to address some notable technical as well as non-technical aspects. Some of them include: Determining the objectives and goals of an organization Identification of different requirements of the audiences Identification of ETL tool Indentifying methods to be employed by the end users to access information, which include both analysis and reporting The chart and figure below indicates an overview of a data warehouse (see Chart 1 and Figure 1).

Figure 1. The General Structure of Data Warehouse (Inmon, 1996) Chart 1. Data Warehouse Overview (Patil and Rao, 2011) Gelder (2010) candidly indicates that most multinational organizations use data warehousing to protect crucial data from any losses. For instance, in the US, such asAppleInc.

, Coca ColaCompany among other international corporations, data warehouses are established as well as maintained in India or China. Inmon (1996) notes that data warehousing has ensured availability as well as better querying performances in comparison to the other approaches, such as a demand approach. This is due to the fact that it is easy to retrieve data directly from one dedicated site, especially when data analysis and improved querying process is required. The approach is also important if the sources of data are expensive to access, or they are unavailable or unreliable.

Kulkarni and Shinde (2011) argue that most organizations have been able to implement data warehouses in analyzing all the available data in their multiple operational systems thus enabling them to compare their current as

well as historical values. By doing this, they have been able to analyze past efforts, manage business and plan their future (Gelder, 2010). However, “ Vertica Systems Inc.” (2008) notes that the quality of decisions, which are facilitated by data house are only as effective as the quality of data contained in data houses. Patil and Rao (2011) note that data warehousing architecture for an organization consists of various components, which co-exist as well as complement one another as indicated on the table below (see Table 1).

Table 1. Data Warehouse Variants (Kulkarni and Shinde, 2011) According to Chaudhuri (1997), data warehouse systems require different kinds of data bases than the convectional data base systems. While to original database systems performed small transactions, such as OLTP (On-line Transaction Processing), data warehousing systems can be used for more complex analytics, mostly involving large quantities of data like OLAP (On-line Analytic Processing), Patil and Rao (2011) note that both OLAP and OLTP mostly co-exist, thus enabling businesses to deal with emerging challenges. It is notable that in OLAP tools, data is mostly represented as a data hypercube as indicated in the figure below (see Figure 2). Figure 2.

Three dimensional OLAP cube having customers, products and time dimensions (Inmon, 1996) Systems Benefiting from Data Warehousing As noted by Gelder (2010), there are various systems, which benefit from data warehouse in an organization. The first type of a system includes the monolithic systems. This is a kind of a system, where a firm controls both single data, offering the data feed and the back-end warehouse stores. For some online-purchasing stores, like theAmazon. com among others, the

underlying data warehouse serves as the container of all those logged overtime, thus enabling the firm to conduct offline analysis (Lomet and Zwilling, 2009). The second category benefiting from data warehouse in an organization is the closed environments. This environment, which is usually well-distributed, is composed of small number of the independent sources of data, mostly controlled by owners having joint cooperative goals (Davenport and Harris, 2007). For instance, in a hospital environment, information system attempts to integrate data sources, which are maintained by various units such as pharmacy, personnel department and registration among others (Lomet and Zwilling, 2009). Comparison between Operating Systems and Data Warehouses There exists a notable difference between operating systems and data warehouse. Patil and Rao (2011) indicate that operating systems are mostly optimized to preserve data integrity as well as the speed of recording transactions made by a business.

This is done by using entity-relationship and database normalization models. The designers of the operating systems mostly follow the Codd rules to ensure that data integrity is well-maintained. On the other hand, data warehouses are usually optimized for data analysis. Therefore, they are denormalized through dimension-based models (Patil and Rao, 2011). Future Trends of Data Warehousing As indicated above, I.

T. sector remains one of the key pillars of growth in most parts of the world (“ Vertica Systems Inc.”, 2008). In order to achieve this, data warehousing should continue to increase to enable organizations store and retrieve information easily. Consequently, data warehousing will have to use

performance and optimization as one of the main differentiators (Ganczarski, 2009).

This will be in addition to focusing on the aspects of optimization storage for the data warehouses though usage-based and compression placement strategies. Patil and Rao (2011) stipulate that the usage of data mart will continue to increase in the future. This is due to the effectiveness of data marts in optimizing warehouse environments through offloading process. Data Warehousing Application in Cloud Computing The term ‘ cloud computing’ can be used to describe an arrangement, in which computing resources, be they software of hardware (computational power) platforms, are made available via a computer network such as the Internet, or other WAN (Wide Area Network) configurations. In a cloud where software is availed as the binding factor, the configuration is referred to as SaaS (Software as a Service).

Where computational resources, such as hard drive space or processing output are the resource-availed, the configuration is called PaaS (Platform as a Service). The cloud is practically seen as a virtual environment because the users, or customers, do not need to know the interior of the platform or the computational resources available for them to use (R. Buyya, 2009). A provider who offers PaaS services to a client may offer networking, computing or storage resources to the client through a computer network such as the Internet. The most general tariff arrangement is the Pay Per Use model, in which a customer pays only foe what they have used.

Virtual Machines (VMs) are a form of computer nodes applied across the networks and allocated to customers. These nodes allow the service provider to share physical computing power resources among different clients. The virtual machines may be physically separated or they may run on the same machine. These applications are such that the end user does not need to know the underlying mechanism of the network (Gelder, 2010). Cloud computing has become a favorable option for many firms, companies and individuals due to its many advantages over conventional networked resource sharing configurations.

Firstly, the arrangement allows customers to infinite scalability in terms of resources. Customers may access additional resources in terms of storage, bandwidth, processing power and similar needs almost instantly upon request. This is lacking in a conventional network with fixed capabilities and rigid, expensive, reconfiguration requirements. Secondly, the Pay Per Use (PPU) model is convenient for most clients, as opposed to the prefixed tariffs generally applicable to conventional networks with resource sharing. This is mainly made possible due to the fact that a very large number of customers can usually subscribe to a cloud environment, as opposed to the fixed capabilities of a conventional client server network.

A cloud environment usually offers better reliability than a conventional network, owing to its very effective but high cost initial investments. Investing in the best facilities is a key element for a cloud platform provider, where cost is not a key consideration. The reliability factor, however, has recently come under scrutiny as more outages are experienced on cloud environment. The number of outages are, however, far lower than that

experienced with ordinary networks (Armburst et al, 2009). The major problems with cloud computing have to do with the rising number of downtime incidences as well as the seldom, but still significant occurrences of bottlenecks, especially in a high I/O (Input – Output) environment, in which huge numbers of clients are writing and/or reading data from the cloud therefore causing delays in service (Armburst et al, 2009).

In addition, clients will lose control of resources that they previously had when they switched to cloud computing because they do not own the platform. In other cases, large corporations transferring data may actually find that the cloud is more expensive than conventional data transfer methods such as physical hard drive transportation. In the specific field of data warehousing, the more important aspect of cloud computing is the PaaS, because database administration is more to do with data storage, organization and retrieval than data analysis and advanced statistical treatment. Using Data Warehousing in the Cloud In order to connect a data warehouse to a cloud, data housed traditionally iin a remote database is distributed into a cloud environment either at a specific physical machine or in different machines sharing resources. The warehouse owner is unlikely to be the cloud’s owner, therefore data control is compromised.

Queries directed to the warehouse are routed to the cloud through the cloud’s node or virtual machine and returns routed via the same node. Due to the fact that data warehousing is mostly concerned with input and output functions such as data reading and data writing, the downtime aspects of a cloud are very significant, as are speeds of the network and cost implications. Connecting a data warehouse to a cloud has as its most

significant aspect the issue of connecting a usually high capability warehouse network through a limiting node (VM) to a cloud (Gelder, 2010). The table below shows specifications for some common data warehousing service providers. The data will be compared with typical specifications for nodes to the leading cloud network providers.

(Patil and Rao, 2011) The cloud end providers have the following configurations. (Patil and Rao, 2011) It can be seen from the tables that the hardware capabilities are very different with the cloud nodes being of far lower capabilities than the high end data warehouses. In order to connect the data warehouses to clouds and still maintain usability, the following adjustments are made. Distributing Workload across Multiple Nodes By using many nodes and parallelization, and partitioning techniques, it is possible to build warehousing applications that are highly distributed, and parallel in a cloud. Parallel applications use different bus paths, effectively reducing bottlenecks caused due to packet queuing. Node partitioning involves splitting the resources in a node to seem like different machines in one.

It may be challenging when the node is small for meaningful partitions. New developments may provide nodes with capacities as high as 23 GB of RAM, 2 TB in node storage and up to 33. 5 GHz speed (“ Vertica Systems Inc.”, 2008). In such cases, a node may be applicable for large organizations databases. Warehouse Bus Parallelization and Node Partition Parallelization is a technique of relaying queries through different data paths in order to avoid queuing in cases, where huge data volumes are involved.

To do this, machines-operating parallel ports, as opposed to the ones running serial ports may be used. Node partitioning involves splitting the physically available capacity of a cloud's machine to appear as though there are multiple machines. This helps the node handle more functions in cycle (Gelder, 2010). Data Compression Alternatives Data compression enables data passing through the bandwidth restricted network links between a cloud node and data warehouse to move fast. This, in turn, requires higher CPU usage in compression and decompression as well as additional software for this work. In order to reduce unpredictability of a data system as large as a cloud with multiple concurrent users, nodes are placed as near as possible to each cluster of warehouses to ease packet path and enhance precision.

Multi-Tenancy Capabilities Different users may be drawing data from the same physical computer. The cloud system must be able to differentiate schema of different users and ensure confidentiality of data transferred between the cloud and the users. The systems are also constructed in such a way that analytical support for system usage is given such as time of access and duration among other uses. Security Measures Clouds may be placed in different geographical and political areas from the data warehouse owners; this lack of control over data may become significant, especially if sensitive data is at risk of exposure or loss. The issue must be handled before joining a cloud system. Usually, this can be achieved through encryption, but at the expense of CPU usage (" Vertica Systems Inc.

", 2008). From the study above, it is clear that the benefits brought about by data warehousing are crucial for the growth and development of many organizations across the globe. Furthermore, we have seen the fact that data

warehousing is possible through cloud computing (Gelder, 2010). However, this is still in its development phase, especially due to the limiting aspects of bandwidth capabilities for link networks, hardware capability differences between high end warehouses and lower capacity cloud nodes, multi-tenancy issues as well as security associated with the cloud environments. It is, however, a very promising technology that will potentially reduce investment costs while enhancing resource access time and convenience. In order to successfully use this technology, it is important to develop nodes mechanisms in terms of specifications of processor speed, physical memory and hard disk capabilities.

This may mean investing in ultra-powerful machines, but as the cost of memory decreases, it is a viable option