ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 8, Issue 3 , March 2021

The Impact of Cloud Computing in Deployment of Effective Data Warehouse Technology

DR. EZE CHIMDIYA CHIEMEKA, DR. AGOHA UCHECHI KENNEDY, OJI NKECHI BLESSING, ONYEMECHARA IDONG-ESIT COMFORT

Lecturer, Department of Computer Science, Federal College of Education (Technical) Omoku Rivers State, Nigeria
Lecturer, Department of Computer Science, Federal Polytechnic Nekede, Owerri Imo State, Nigeria
System Analyst, Federal University of Technology Owerri, Imo State, Nigeria
Lecturer, Department of Computer Science, Federal Polytechnic Nekede, Owerri Imo State, Nigeria

ABSTRACT: The recent years have seen a tremendous surge in data generation levels. The amount of data being generated globally is increasing at rapid rates with digitization taking over industries; more and more organizations are generating digital data like never before. Cloud Computing has emerged as a new paradigm for hosting and delivering services over the internet and it is the major source for data sharing, storage and processing. A data warehouse is a specialized database focused on data quality and presentation, providing tangible data assets that are actionable and consumable by user and also support decision-making process. In this paper we will focus on the impact of cloud computing in the deployment of Data warehouse technology. For this purpose, we first provide a functional overview of cloud computing, its characteristics and requirements for cloud-based Data Warehousing systems and at the end we provide the challenges to deploy the data warehousing on the cloud.

KEYWORDS: Cloud Computing, Data Warehouse, Deployment, Cloud Services, Cloud

I. INTRODUCTION

The amount of data being generated globally is increasing at rapid rates. In fact, studies by the Gigabit Magazine depict that the amount of data generated in 2020 will be over 25 times greater than it was 10 years ago. Furthermore, it has been estimated that by 2025, the cumulative data generated will triple to reach nearly 175 zettabytes. Demands from business decision makers for real-time data access is also seeing an unprecedented rise at present, in order to facilitate well-informed, educated business decisions. In order to make data useful, actionable and scalable for their business, enterprises need an efficient and cost-effective way to store, label, and interpret this data. One of the most lucrative ways to do this is through data warehousing. Cloud computing is basically for storing and accessing of applications from the computer (Remote). Whereas Datawarehousing 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. Cloud computing is about moving services, computation and data for cost and business advantage off-site to an internal or external, location-transparent, centralized facility. By making data available in the cloud, it can be more easily and ubiquitously accessed, often at much lower cost, increasing its value by enabling opportunities for enhanced collaboration, integration, and analysis on a shared common platform. In this paper we tried to attempt how we can apply the cloud computing strategies to the Data Warehousing systems.

II.BACKGROUND OF STUDY

The traditional data warehouses solved the problem of processing and synthesizing large data volumes, but they presented new challenges for the analytics process. Cloud data warehouses took the benefits of the cloud and applied them to data warehouses bringing massive parallel processing to data teams of all sizes. Software updates, hardware, and availability are all managed by a third-party cloud provider. Scaling the warehouse as business analytics needs grow is as simple as clicking a few buttons (and in some cases, it is even automatic). The reduced overhead and cost of ownership with cloud data warehouses often makes them much cheaper than traditional warehouses. With all of your data in one place, the warehouse acts as an efficient query engine for cleaning the data, aggregating it, and reporting it often quickly querying your entire dataset with ease for ad hoc analytics needs.

Figure 1 Cloud based warehouse

III. RELATED WORKS

Dating back to the 1970s, the data warehousing market emerged when computer scientist Bill Inmon first coined the term ‘data warehouse’. Created as on-premise servers, the early data warehouses were built to perform on just a gigabyte scale. They have undergone significant transformation since then, with modern warehouses housing largescale terabyte capacities.

Data warehouse, also known as a decision support database, refers to a central repository, which holds information derived from one or more data sources, such as transactional systems and relational databases. The data collected in the system may in the form of unstructured, semi-structured, or structured data. This data is then processed, transformed, and consumed to make it easier for users to access it through SQL clients, spreadsheets and Business Intelligence tools.

Data warehousing also facilitates easier data mining, which is the identification of patterns within the data which can then be used to drive higher profits and sales. Data warehousing industry application scope spans across several domains related to analytics and even cloud in some cases, including BFSI, healthcare, manufacturing, telecom & IT, retail and government, among others. There are several companies in the technological sphere making significant strides in advancing data warehousing technologies. One of the most prominent is Teradata, which is a leading data warehouse company, with over 30 years of experience in the domain. The Teradata software is used extensively for various data warehousing activities across many industries, most notably in banking. The company works consistently to enhance its business intelligence solutions through innovative new technologies including Hadoop-based services. Three models of Cloud computing are accessible today. There is Software as a Service (SaaS), Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). SaaS model is unified with which most people are well known regardless of the fact that they don't comprehend its hidden innovations. Google's Gmail for instance, is a standout amongst the most broadly known and ordinarily utilized SaaS stages. SaaS Big Data Applications (BDAs) exist at the most elevated amount of the cloud stack. Customers are prepared to utilize them out-of-the-crate. There is no perplexing costly framework to set up or programming to introduce and oversee. In this same model, Salesforce has extended its offerings through a progression of acquisitions including Radian6 and Buddy Media. Salesforce now offers cloud based social, information and examination applications. More current participants like AppDynamics, BloomReach, Content Analytics, New Relic and Rocket Fuel all arrangement with expansive amounts of cloud-based information.

IV. THE CLOUD COMPUTING PARADIGM

A. What is cloud computing?

Cloud computing is a method of providing a set of shared computing resources that includes applications, computing, storage, networking, development, and deployment platforms as well as business processes. Cloud computing turns traditional siloed computing assets into shared pools of resources that are based on an underlying Internet foundation. Clouds come in different versions, depending on your needs.

Figure 2. Cloud Computing

B.TYPES OF CLOUD ENVIRONMENTS

Open Community Clouds

The most open type of cloud environment is an open community cloud — a cloud environment that doesn't require any criteria for joining other than signing up and creating a password. These environments may be privately or publicly owned and include social networking environments, such as Facebook, LinkedIn, and Twitter. There are also open community sites that enable individuals with a common interest to participate in online discussions. For example, there may be a community of professionals in a certain industry that want to share ideas.

Controlled Open Mode

Some public clouds offer a higher level of service because they are true commercial environments. Commercial public clouds are those environments that are open for use by any one at any time, but these clouds are based on a pay-per-use model. For example, a SaaS vendor that charges per-user, per-month or per-year is one example of this kind of environment. In addition, vendors can offer analytics as a service to customers on a per-use or per-task basis.

Public/Private Hybrid Clouds

Companies often want the flexibility of the cloud but with the security and predictability of the data center. In these cases, a private cloud provides an environment that sits behind a firewall. Unlike a data center, a private cloud is a pool of common resources optimized for the use of the IT organization. Unlike a public cloud, a private cloud adheres to the company's security, governance and compliance requirements. Whatever service level is required for the company applies to the private cloud.

C. CLOUD COMPUTING CHARACTERISTICS

Cloud computing has five essential characteristics. They are on-demand capabilities, broad network access, resource pooling, rapid elasticity and measured service. These are the characteristics that distinguish it from other computing paradigms.

- On-demand Capabilities: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops and workstations).
- Resource Pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned per consumer demand.
- Rapid elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.
- Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth and active user accounts).

D. CLOUD DEPLOYMENT MODELS

Cloud deployment models are grouped broadly into four models: private cloud, public cloud, community cloud and hybrid cloud. Private cloud is the most secure way to utilize cloud computing. The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. Community cloud is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns. It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises. Public cloud is provisioned for open use by the public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider. Hybrid cloud is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability.

E. CLOUD SERVICE DELIVERY MODELS

The three categories of services offered by cloud providers are discussed below.

- **Infrastructure as a Service** is the delivery of computer hardware (servers, networking technology, storage, and data center space) as a service. It may also include the delivery of operating systems and virtualization technology to manage the resources.
- **Platform as a Service** includes the delivery of more than just infrastructure. It delivers what you might call a solution stack — an integrated set of software that provides everything a developer needs to build an application — for both software development and runtime.
- **Software as a Service** is the delivery of business applications designed for a specific purpose. Software as a Service comes in two distinct modes:
    - **Simple multi-tenancy:** Each customer has its own resources that are segregated from those of other customers. It amounts to a relatively inefficient form of multi-tenancy.

- **Fine-grain multi-tenancy:** This offers the same level of segregation but is far more efficient. All resources are shared, but customer data and access capabilities are segregated within the application.

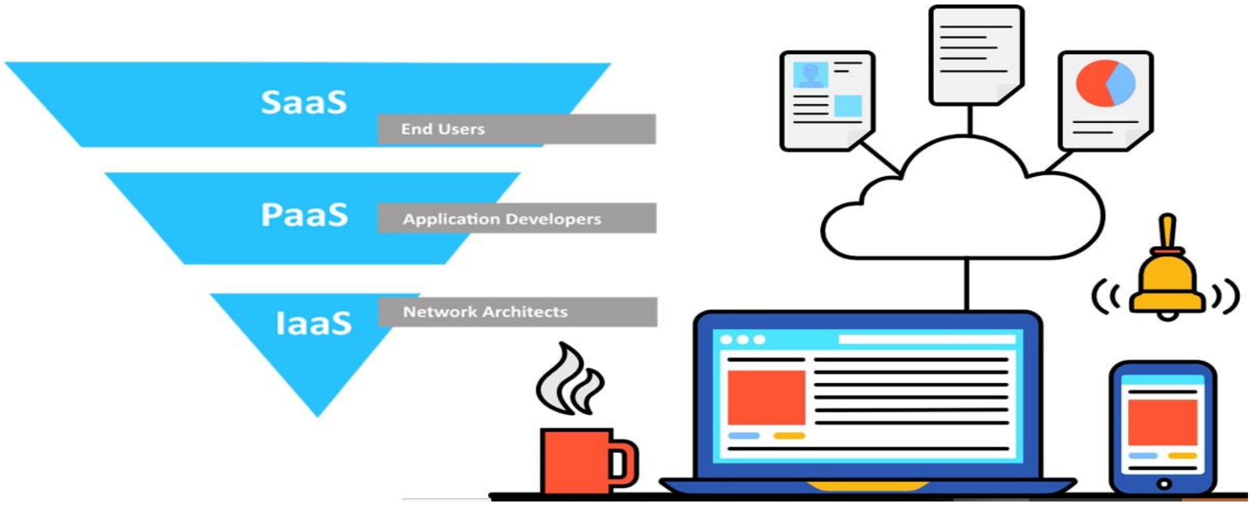

Figure 3. Cloud Service Delivery Models

VI. REQUIREMENTS FOR CLOUD-BASED DATA WAREHOUSING SYSTEMS

In order to adapt and become more suited for the cloud environment data warehousing systems would have to comply with the requirements listed below. While data warehousing systems in general have many more functional requirements, we only state the most important ones specifically relevant to moving towards the cloud.

- High performance: Data needs to move from persistent cloud storage towards compute nodes or between compute nodes in a fast way. Moving significant amounts of data (up to terabytes) can be needed when new nodes become active (in order to deal with changing workloads for example) as well as during query processing. This is a challenge because storage and network bandwidth in the cloud are generally not fast compared to traditional data warehousing systems. A possible (partial) solution is to use compression to reduce bandwidth usage while paying the price of higher CPU usage.
- Flexibility: In order to achieve elasticity data warehousing systems in the cloud will have to be able to scale up and down automatically once workloads, amounts of users or data volumes increase or decrease. Partitioning can be used in order to distribute data across different nodes in the cloud, comparable to the situation in traditional (cluster) data warehousing systems.
- Multi-tenancy: Data warehousing systems in the cloud will have to be able to deal with multiple different users potentially using the same database server. Data warehousing systems in the cloud will have to make sure that database schema's are strictly separated. Different tenants must not be able to see each other's data even though data may be placed at the same physical machine or even in the same table.
- Infrastructure: In terms of infrastructure data warehousing systems in the cloud will need API's that allow local system entrance without going via the main server continuously. This can help in reducing latency for example, because if a data warehouse server is on the other side of the world, latency can potentially become an issue (although, like we discussed, it will not likely be a bottleneck). Having to communicate over large geographical distances is a new issue in data warehousing applied in the cloud, because traditional data warehousing systems are generally situated at or close to the owning organizations.
- Privacy: Data warehousing systems in the cloud must be able to encrypt data locally in order to ensure privacy. The possibility to operate on encrypted data would be a valuable addition. This does lead to previously discussed technical issues [Ahituv, Lapid & Neumann, 1987) because encrypted data tends to require different forms of analysis.
- Security: Data warehousing systems in the cloud must be secure so that the data warehouse including all forms

of communication to the data warehouse are accessible only to the original customer. Security can lead to many issues discussed before.

- Monitoring: Monitoring capabilities have to be available for customers and administrators in order to analyze the systems operations. Potential bottlenecks (e.g. 'killer-queries') must also be detected and cancelled when required. Monitoring capabilities are important in achieving for example elasticity and load balancing; changing usage patterns need to be detected in order to be able to decide whether or not scaling up/down is required, or whether or not loads need to be rebalanced.
- Availability and Reliability: Data Warehousing systems in the cloud must be available at all times to the customer and data warehousing systems in the cloud must be highly reliable. Having a separate replication server could help, because it possibly allows dealing with system failures, like for example power-outages, automatically.

VI. . REASONS TO MOVE TO A CLOUD DATA WAREHOUSE

- Cost Efficiency - This is the biggest advantage of cloud computing, achieved by the elimination of the investment in stand-alone software or servers. By leveraging cloud's capabilities, companies can save on licensing fees and at the same time eliminate overhead charges such as the cost of data storage, software updates, management etc. Renting your infrastructure can make good financial sense.
- Continuous availability - Public clouds offer services that are available wherever the end user might be located. This approach enables easy access to information and accommodates the needs of users in different time zones and geographic locations. As a side benefit, collaboration booms since it is now easier than ever to access, view and modify shared documents and files. Moreover, service uptime is in most cases guaranteed, providing in that way continuous availability of resources. The various cloud vendors typically use multiple servers for maximum redundancy. In case of system failure, alternative instances are automatically spawned on other machines.
- Scalability and Elasticity - Scalability is a built-in feature for cloud deployments. Cloud instances are deployed automatically only when needed and thus, you pay only for the applications and data storage you need. Hand in hand, also comes elasticity, since clouds can be scaled to meet your changing IT system demands.
- Fast deployment and ease of integration - A cloud-based application can be up and running with just a few hours rather than weeks or months and without spending a large sum of money in advance. This is one of the key benefits of cloud. On the same aspect, the introduction of a new user in the system happens instantaneously, eliminating waiting periods. Resiliency and Redundancy- A cloud deployment is usually built on a robust architecture thus providing resiliency and redundancy to its users. The cloud offers automatic failover between hardware platforms out of the box, while disaster recovery services are also often included.
- Increased Storage Capacity- The cloud can accommodate and store much more data compared to a personal computer and in a way, offers almost unlimited storage capacity. It eliminates worries about running out of storage space and at the same time it spares businesses the need to upgrade their computer hardware, further reducing the overall IT cost.

VII. CHALLENGES TO DEPLOY DATA WAREHOUSING SYSTEMS ON CLOUD

Besides advantages, there are also some potential disadvantages and bottlenecks to keep in mind when discussing cloud computing. The potential disadvantages listed below underline why it is challenging to deploy data warehousing systems in the cloud.

- Communication with the cloud happens by means of a WAN link. When dealing with huge amounts of data (i.e. terabytes) this might become a bottleneckArmbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009). Data transfer over a WAN link is generally not very fast compared to data transfer in local systems. A WAN link can also lead to latency issues.

- Performance in the cloud can be unpredictable (Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009). When a large number of customers are active in the cloud at the same time, this may decrease performance. Especially when sharing I/O to write to traditional disk Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009). This potential disadvantage does not have to be present; it depends on the architecture of the cloud by the cloud provider and careful design of the cloud might go a long way in eliminating these issues.

- Loss of control (Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009). When an organization starts to use the cloud as a platform, it loses some of the control it previously had. One can for example no longer increase performance by buying better hardware, nor can one fix downtime themselves anymore. Loss of control also leads to significant security and thrust issues.

- High costs. Costs can both be an argument for and against using the cloud. When many terabytes of data are involved, data transfer to the cloud can become expensive. For example: Amazon charges roughly $100 to transfer 1TB of data towards their cloud (Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009). It is even claimed that physically shipping data on disk to a different location is the cheapest way to send large quantities of data(Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, Lee, Pattersom, Rabkin, Stoica & Zaharia, 2009) (some cloud providers allow customers to do this). Exploiting economies of scale can be difficult for large organizations once dependent on cloud provider prices.

VIII. CONCLUSION

We have discussed the possibilities of data warehousing via the cloud. It is our belief that data warehousing systems in the cloud have great potential, due to its elasticity, scalability, deployment time, reliability and reduced costs (due to e.g. elasticity). In order for a data warehousing system to be able to utilize the capabilities of the cloud it will have to be both highly parallel and distributed, while complying with many requirements discussed in this paper. Furthermore, we believe that in the short term data marts have high scope in the cloud compared to data warehousing systems because they tend to be smaller due to their specialized nature allowing them fit in less or just a single node. This makes distributing and parallelizing less of an issue. Security issues will likely always be involved in the decision of moving a data warehousing systems or data marts into the cloud.

REFERENCES

[1]Vaibhav C. Gandhi, Jignesh A. Prajapati and Pinesh A. Darji, “Cloud Computing with Data Warehousing”, International Journal of Emerging Trends & Technology in Computer Science, Volume 1: Issue 3, Page 72,73, ISSN 2278-6856, September – October 2012.William H. Inmon, Book, “Building the Data Warehouse”, Page. 1-30.

[2]K. Kala Bharathi and K. Sandhya Sree, “Recent Developments on Data Warehouse and Data Mining In Cloud Computing”, International Journal of Computer Science Engineering and Technology (IJCSET), Vol 5, Issue 2, pp.31-34, Feb 2015

[3]A. Aboulnaga, K. Salem, A.A. Soror, U.F. Minhas, P. Kokosielis, S. Kamath (2009). Deploying Database Appliances in the Cloud. *IEEE Data Eng. Bull*.

[4]A. Ganapathi, H.Kuno, U.Dayal, J.L. Wiener, A. Fox, M. Jordan, D. Patterson (2009). Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning. *Proceedings of the 2009 IEEE International Conference on Data Engineering*.

[5]A. Gupta, I.S. Mumick (1995). Maintenance of Materialized Views: Problems, Techniques, and Applications. *Data Engineering Bulletin*.

[6]Agarwal, D., Das, S. and Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM 978-1-4503-0528-0/11/0003. Retrieved from: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pd [2] Aydin, N. (2015). Cloud Computing for E-Commerce, Journal of Mobile Computing and Application. Vloume 2, Issue, 1, pp 27-31.

[7] Barthelus, L. (2010). adopting cloud computing within the healthcare industry: opportunity or risk? Online Journal of Applied Knowledge Management, Volume 4, Issue 1.

[8] Dan, S and Roger, C. (2010). Privacy and consumer risks in cloud computing, Computer Law and Security Review, Vol 26, pp: 391-397.

[9] Fan, J., Han, F. & Liu, H., 2013. Challenges of Big Data Analysis. ResearchGate, 1(1), pp.1-38.