

Bridging the Gap in Modern Computing Infrastructures: Issues and Challenges of Data Warehousing and Cloud Computing

Iyabo Awoyelu¹, Theresa Omodunbi¹ & James Udo²

¹ Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

² Department of Computer Science, University of Uyo, Uyo, Nigeria

Correspondence: Iyabo Awoyelu, Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria. Tel: 234-803-373-6280. E-mail: iawoyelu@oauife.edu.ng

Received: August 30, 2013 Accepted: October 17, 2013 Online Published: October 28, 2013

doi:10.5539/cis.v7n1p33

URL: <http://dx.doi.org/10.5539/cis.v7n1p33>

Abstract

Data warehousing and cloud computing are modern trends in computing businesses. Data warehouse (DWH) is subject-specific, time-changing, non-volatile, integrated collection of data and it is a process that helps decision maker in the process of informed decision making. It is also an integrated software component of the cloud which provides support for timely and accurate response to complex queries. The complex analytics involving huge amounts of data with the help of tools such as: Online Analytical Processing (OLAP) and data mining are built on DWH model. Similarly, cloud computing provides a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider's interaction. These computing infrastructures when allowed to work together can provide decision makers with immense benefits at minimal costs. Therefore, it is pertinent to consider the contending issues and challenges faced by bridging the gap between DWH and cloud computing as well as proffer possible solutions. The future directions and conclusions are also pointed to and drawn from the paper.

Keywords: data warehousing, cloud computing, modern computing, internet, infrastructures

1. Introduction

Data warehousing and cloud computing are recent trends in modern computing. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider's interaction. On the other hand, data warehouse (DWH) is a subject-oriented, time-variant, non-volatile, integrated collection of data and the processes that help decision maker in the process of informed decision making (Inmon, 2002). DWH can be seen as an integrated software component of the cloud which provides support to timely and accurate response to complex queries as well as complex analytics involving huge amounts of data with the help of tools such as Online Analytical Processing (OLAP) and data mining.

According to Huth and Cebula (2011), there is either of the following provision in cloud computing: software can be provided as a service to customers (Software-as-a-Service or SaaS) or computational resources can be provided for a customer to use the cloud as a platform (Platform-as-a-Service or PaaS) or Infrastructure as a Service (IaaS). Cloud computing can be of immense benefits by providing infinite scalability. Infinite scalability is the illusion of availability of infinite computing resources. Cloud computing also provides reasonable speed of deployment by offering full-fledged services in a reduced time when it is compared to in-house deployment. There is also elasticity, reliability and reduced cost due to economies of scale, a pay-per-use payment model and backups are made available.

However, there are several challenges when deploying data warehouses in the cloud. These challenges, although it can be thoroughly checked, can pose security, computation and network problems. These problems are often encountered mainly due to incompatible nature of functional requirements needed to deploy DWH in the cloud environment and vice versa. Some of these challenges are importation of the data needed for the DWH into the cloud for storage and getting large amounts of data from cloud storage to virtual nodes provided by the cloud

providers for computing. Similarly, cloud providers tend to offer low-end nodes (e.g. virtual machines) for computations, whereas local data warehousing systems tend to be well-provisioned in terms of CPU, memory and disk bandwidth. Often times, applications running in the cloud could experience Wide Area Network (WAN) latency. Loss of control can lead to issues involving loss of data security and trust. These challenges can be addressed if a comprehensive cloud-based system administration and data lifecycle management among other measures are taken into consideration.

In summary, the growing need of cloud computing will allow its evolving more in future to accommodate mission critical DWH, thus resulting in revolutionizing the different areas of data warehousing. The evolving nature of cloud computing and DWH will help the small and medium sized businesses to use more analytical data because of lower operational cost.

The remaining part of this paper is organized as follows: Sections 2 and 3 discuss data warehouse and cloud computing respectively in detail and related works in the two areas. Section 4 discusses the major issues and challenges in cloud computing. It also addresses remedies to the challenges while Section 5 concludes the paper with future directions of these modern computing infrastructures.

2. Data Warehouse

DWH has been defined by many authors to suit their purposes. In all definitions, there has been a common issue that is agreed on. This is the issue of DWH as storage for historical data which is used for analytical purposes by decision makers. According to Inmon (2002), DWH is a subject-oriented, time-variant, non-volatile, integrated collection of data that helps decision maker in the process of informed decision making. DWHs are computer based information systems housing “second hand” data that are from either another application or from an external system or source. A DWH is constructed over several heterogeneous data sources such as operational databases and plain files. It contains historic data and is subject-oriented and static; that is, users do not update but it is created on a regular time-frame on the basis of the operational data of an organization (Adriaans & Zantinge, 2001). DWH was developed due to limitations posed by databases in handling timely and accurate responses to complex queries which are major driver to business competitiveness. Therefore, DWH supports timely and accurate response to complex queries and DWH systems are used for complex analytics involving huge amounts of data e.g. Online Analytical Processing (OLAP). The operations that aid analysis in the DWH are drill-down and roll-up. The drill-down analysis allows data to be viewed in its lowest level of details or granularity, while the roll-up operations allow data to be viewed in its summarized form. DWH is made up of data from different data sources and elicited requirements from users of the system (Udo et al., 2012). The DWH system can be accessed via OLAP and data mining tools.

2.1 DWH Architecture

The general architecture of DWH consists of component that allows DWH to function in the desired manner. These components are as illustrated in Figure 1.

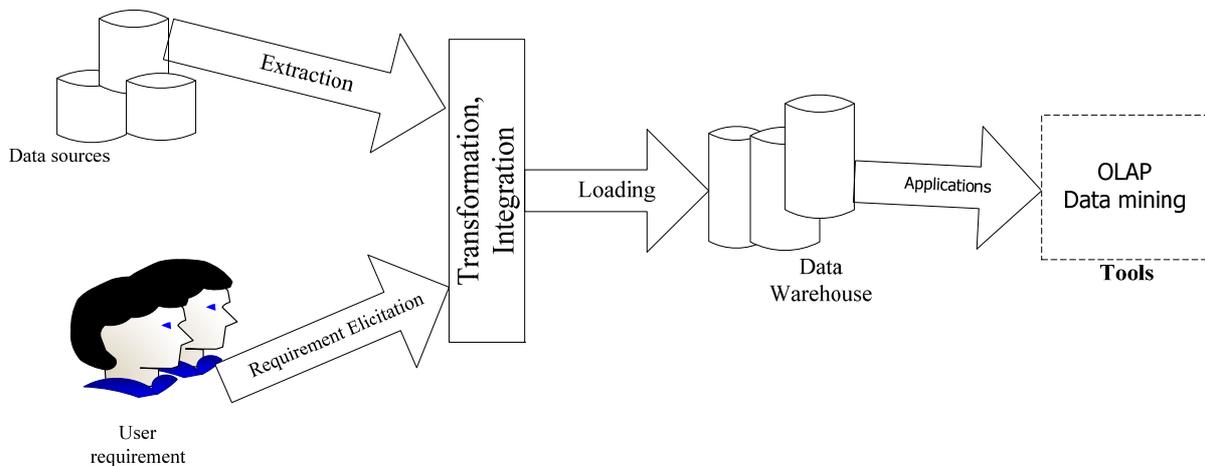


Figure 1. Architecture of a Data Warehouse

2.1.1 Data Sources

In DWH, the available data are sourced from the transaction processing systems and data marts for extraction, transformation, integration and loading into the DWH. These data from different source systems may pose a difficulty in the course of integration to a single unit because of format incompatibility. These problems are often overcome by extracting, transforming and integrating data from source systems (e.g. databases, data marts etc.) before it is loaded into the main storage (i.e. DWH).

2.1.2 User Requirement

In a cloud-based environment, there is a cost in terms of time and resources attached to migrating data to data warehouse in the cloud. However, to minimize the cost of operating data warehouse in a cloud, user requirements which often change in time (Udo et al., 2012) should be considered during requirement elicitation and preprocessed to form an integral part of data in the data warehouse to meet the changing needs of the users.

2.1.3 Elicitation, Extraction, Transformation, Integration and Loading

The process of wrapping data into a DWH from different source systems entails extraction and transformation. Similarly, the changing user requirements are elicited from the user. These processes help to remove inconsistencies and duplicate data from the DWH. The transformation specifically changes data format of different source systems and user requirements to a compatible format in the DWH. The transformed data and user requirement are both integrated and loaded into the DWH. However, the loaded DWH is suitable for use by various applications such as Online Analytical and Processing (OLAP) and data mining tools.

2.2 DWH Tools

There are various tools that allow data in the DWH to be viewed and manipulated. These tools such as OLAP and data mining tools support operations such as drill-down and roll-up operations. Once the data is loaded into the data warehouse, it is ready for use by OLAP and data mining tools. In OLAP tools, data is usually presented in the form of a data hypercube which is made up of dimensions and measures.

2.3 Types of Data Warehouses

The various types of DWHs that are considered in this work are as follows: point-to-point, centralized and distributed data warehouses.

(1) Point-to-point DWH

This is otherwise called a virtual data warehouse. In this strategy, the end-users are allowed to get at operational databases directly with appropriate tools that are enabled to perform the function of data access.

(2) Centralized DWH

Centralized DWH is preferred whenever the need for data sharing among specific functional areas, departments and other units of an organization is necessary (Awoyelu, 2009). It is a single physical database that contains all of the data for a specific functional area, department, division, or enterprise. Central DWHs are often selected where there is a common need for informational data and there are large numbers of end-users already connected to a central computer or network. A central DWH may contain data for any specific period of time. These data are got from different data sources with DWH connecting them together.

(3) Distributed DWH

This is a DWH in which the certain components of it are located across a number of different physical databases. It integrates data from a central registry and individual local sites and gives on-demand access to data (Awoyelu, 2009).

2.4 DWH in the Market Place

The issues of the shape of DWH market with its major product suppliers such as IBM, Oracle and Teradata are reviewed in the work of Gelder (2011). These issues are based on functionalities offered by the different companies. For instance, according to Gelder (2011), IBM DWH systems have various specifications for various functionalities. The range of values for core is 8 - 320 core processors, memory range of 64 GB – 2560 GB and 14 TB – 560 TB of storage spread across multiple modules. Oracle with the advent of Exadata Database Machine X2-8, provides a capacity of 128 processor cores, 2 TB memory for database processing and 168 cores for storage processing and 24 TB capacity of storage. Moreover, Teradata storage ranges from 5.8 – 343 TB in fully populated six cabinets.

Therefore, the market trend in DWH is promising as the storage capacities and processor cores of each of the product suppliers are increasing.

At the present, the storage capacities in DWH is hovering in terabytes with optimization to reduce the data storage cost. Advanced analytics is also introduced to help clients to discover new opportunities that can be used to optimize business outcomes. In future, it is expected that the processor cores will increase geometrically with adequate partitioning. The storage capacities for both database and storage processing will be counting in Petabytes. The degree of parallelizability and OLAP tools will be capable of sustaining large volumes of data in just a smallest possible time. These efforts will be geared towards an era of high scalability in DWH, thereby rendering it more complex.

3. Cloud Computing

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (NIST, 2010). It allows organizations to be less concerned about computational management and concentrate on more specific tasks that increase their productivity. Cloud services can be accessed through a web browser, mobile applications or lightweight application installed on desktop. These services can be installed or accessed for a particular purpose in an organization with less stress and cost. For example, an organization with a huge number of computer systems using particular software, for instance, Microsoft Office will utilize their expenditure by subscribing for similar product in the cloud thereby reducing the cost of installing same product on each of their computers. Just light-weight software will be downloaded and used by the organization. All that is paramount is a reliable internet connection to access information needed from their organizations. Cloud computing allows organizations to get up and running irrespective of where they are, effectively carry out their assignments and thereby increase profit.

The main purpose of cloud computing is to share resources. Cloud computing has made storage and computing resources easy to share and use in terms of accessibility and flexibility. Cloud providers now host a number of applications (Gelder, 2011) and allow clients to bring their own applications while they host for them. The term cloud means hosting and managing these computing resources from a remote area like “up there” with assurance of better and reliable services. It gives capabilities for all sorts of software – back end, front end, applications etc. - in a way that just a web browser and internet connection is needed to get the application running. Figure 2 depicts the overview of cloud computing. Services such as applications (both bespoke and standard), data, information, communications through emails, telephony networks etc. can be accessed by different electronic devices (e.g. PDA, laptops, mobile phones) through the use of internet. Depending on the service an organization subscribed for on the internet, the facilities are delivered.

3.1 Types of Cloud Computing

Cloud computing can be classified according to their availability to users. Cloud computing notable types are Public, Private and Hybrid cloud.

Public Cloud allows availability of cloud resources to general public and users are billed pay as use or free use. The cloud is available for multiple customers. Organizations can subscribe to public cloud if their information is meant for general public. Private cloud is part of cloud infrastructure operated solely for a single customer. It is dedicated to manage a customer or organization at a time to make their applications or data private and secure. An example is a financial institution of a particular country that coordinates multiple bank transactions. Hybrid cloud is a type of cloud that merges both private and public clouds, thereby linking the infrastructure of both types, offering the benefits of multiple deployment models. In this case, sensitive data are deployed in private while data that are less sensitive are deployed in public data.

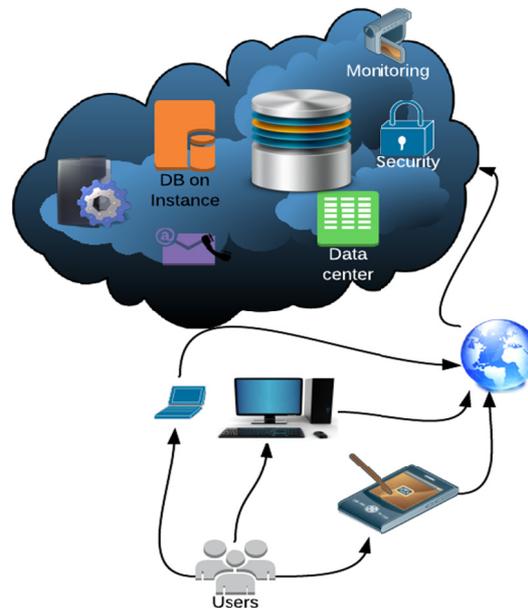


Figure 2. Schematic representation of cloud computing

3.2 Components of Cloud Computing

In cloud computing environment, there are various resources that are necessary for efficient functioning of the cloud. These resources include network, servers, storage, application and services. These resources, according to Duncan et al. (2009) and Monaco (2012), are components of Cloud computing which can be categorized into Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure as a service (IaaS).

SaaS provides software for users and organizations as the need demands and allow them to access those software from their remote systems there by eradicating the stress of installing software one every new system bought or formatted and cost of maintenance and support by IT staff. It also reduces the cost of purchasing software for every system used and allows more space on the system compared to installing the whole package (e.g. antivirus) on each system used in the organization.

PaaS is a service in which cloud provider provides computing platforms such as OS, servers, programming language environments, databases etc. to host organizations' products in the cloud. They provide a scalable platform for merging both existing system (local server) and new system (cloud server) thereby optimizing the efficiency of the organization. This allows organizations to bring their software while cloud providers host such for them. The good thing about this is that the organizations can customize their data as they want and save their heads from maintenance and support issue.

IaaS provides users with physical machines such as networks and computers (such as memory, CPU) as requested to help keep organizations' sensitive information in separate virtual machines. IaaS manages basic computing resources allocated to each task by the organization. These machines are also kept with cloud providers in one of their data centers and managed by them.

With cloud services, organizations can host their applications in cloud (Platform-as-a-service), subscribe for software (software-as-a-service) and even request for physical computers that may serve as virtual machines (infrastructure-as-a-service) which are kept in the data center of cloud providers.

From cloud computing components, notable advantages of cloud computing are reliability, scalability, elasticity, high performance, device and location independence, security, reduced cost, ease of maintenance, multi-tenancy. Although some researchers (Amburst et al., 2009) considered these merits as debatable, cloud providers, of recent, have improved their services and they are of no doubt observed these merits as the importance of using cloud computing.

4. Data Warehouse in the Cloud

Cloud computing is increasingly improving their sets of services in a way to accommodate large capacity, instant deployment of resources with secure networks and low cost and at the same time satisfy users in terms of

performance and reliability (Malathy et al., 2013). Competitions among cloud providers have increased cloud computing performance and thereby give possibility of housing DWH in the cloud. DWH serves as a PaaS in the cloud in which different data sources can be linked to. The data sources can be in cloud or in a remote server that is accessed through the internet. The following subsections explained deploying DWH in cloud, challenges and suggested solutions to make DWH a home in cloud.

4.1 Deploying Data Warehouses in the Cloud

DWH deals with huge amount of data with referencing to other data sources which might not be in the cloud. Hosting DWH in the cloud will encompass PaaS and hybrid cloud. Its deployment is explained in Figure 3. From the diagram, the data needed for data warehouse move between the organization and the data warehouse. Frequent changes from the operational data sources such as OLTP (Online Transaction Processing), documents and web services are passed through the ETL engine to clean up the data for irregularity or replicas to ensure the data warehouse performed as desired and then transferred to the cloud as storage. The processed data in the data warehouse can be queried by the OLAP (Online Analytics Processing) tool and data mining for decision making.

4.2 Benefits of Data Warehouses in the Cloud

There are numerous benefits of deploying DWHs in the cloud environment. These among others include: cost efficiency, time efficiency, flexibility and more competitiveness.

In terms of cost efficiency, hosting infrastructure or hardware is expensive given that DWH requires a lot of different software. Therefore, customers can save huge cost by using the cloud and paying for the resource as per their use. For small and medium sized companies, it is also needless to worry about the ownership and maintenance of the software and hardware which pays a great benefit to such businesses. Cloud computing provides organizations with spending less or decreased capital expenses. It only requires an operational expense which is very low compared to capital expenses.

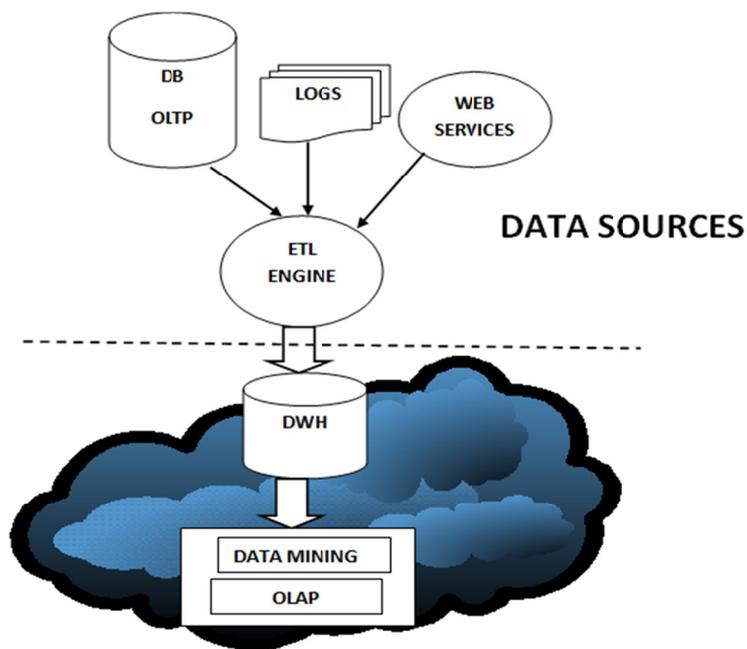


Figure 3. Deploying DWH in cloud

Similarly, regarding time efficiency of DWH in the cloud, customers do save a lot of time, as they need not spend weeks of effort for buying, installing and configuring hardware and software. Cloud can easily be made available for use thereby reducing the amount of time wastage.

Also, flexibility is provided in cloud computing because customers are provided with more options to choose. With low operating costs and easy availability of software and hardware infrastructure, organizations can migrate from one technology to another comparatively easily.

More competitiveness is also provided in the cloud between providers of similar service. Users can evaluate similar software and hardware from different vendors easily and quickly. This healthy competitiveness can give customers more options of choice.

4.3 Issues of DWH in the Cloud

Moving the data needed for the data warehouse into the cloud for storage may be a challenge. This is because organizations depend on the internet and the infrastructure of the cloud provider. DWH deals with large volume of data from different sources and thereby need a “big data” capability in cloud. Although many challenges have been presented in the work of Gelder (2011), this paper focuses attention on the listed points as the major issues of DWH in cloud. These are as follows:

- **Performance:** With performance measure, the rate of data transfer and the scalable nature of the DWH are brought to the fore. The cloud computing technology often uses WAN link for communication (Gelder, 2011). The slow nature of the WAN link makes it a bottleneck for organizations transferring large quantity of data. Ability to process bulk data posed by the large volume of data in the cloud makes DWH unsuitable for cloud. Importing data into DWH and at the same time querying the DWH makes it inefficient. Data are transferred offline in case of local DWH and thereby does not concern with importing and querying but cloud DWH is slow in this area. Uploading and querying the DWH at the same time may slow down or even hang the system. There is need to improve bulk data handling and the performance of processing bulk data in the cloud.
- **Cost:** The cost in terms of data transfer and infrastructural procurement in case of a small organization (for instance small and medium-sized businesses) also poses a challenge. High cost of transferring data to cloud environment is an issue that portends danger to the growth of cloud. If organizations pay exorbitantly to transfer large terabytes of data into the cloud, what will be the level of patronage of cloud in future when DWH capacity will be counting in Petabytes and there is no corresponding reduction in the cost of data transfer? Therefore, it is pertinent to address high cost of data transfer and infrastructure procurement to encourage users. Some cloud providers, such as Amazon, charge as high as 100 dollars to move 1TB of data into the cloud. However some cloud service providers, such as Amazon, allow users to ship physical disks of their data to their sites which is cheaper and also increase performance (Gelder, 2011).
- **Security Issues:** The local data warehousing system gives the organizations full control. With cloud computing however, there is a loss of control which can give rise to issues such as security and trust. Sharing of data over the internet compared to it in a private network increases data exposure. Cloud computing also shares infrastructure between clients and controls information technology workloads between lots of dissimilar physical machines and maybe data centers that are geographically apart. This means that organizations are totally unaware of where their data is located or the way it is protected. Identity management becomes a concern in cloud computing compared to data stored in house. The diverse privacy regulation from country to country is equally a core issue (Buyya et al., 2009).
- **Availability and interoperability issues:** Availability is also a major concern. Services from the cloud service providers may not always be available when needed. Though data centers of the cloud service providers are in diverse places geographically, usually workloads are not spread and balanced between these data centers. Another issue is the poor regional coverage of the cloud service providers, this usually results to organizations being served from far data centers. At present, there is no standardization for cloud service providers; this makes it difficult for organizations wishing to switch to another cloud service provider difficult. Deploying data needed for a data warehouse into the cloud also faces some interoperability issues such as those with the organization present infrastructure, the application and the cloud service provider. Moving the IT operations of an organization into the cloud usually takes a while.

5. Conclusion and Future direction

The possible solutions to the many challenges of bridging the gap between DWH and cloud computing could be in the aspects of optimizing computing infrastructures (i.e. both hardware and software) to deliver high performance systems with minimum cost. This could serve as a motivation to small and medium-sized businesses to expand its operational efficiency with modern computing tools. Moreover, storage facilities like DWH should adopt computing techniques like advanced data compression, analytics, scalability and high level security to also ensure extreme performances at a reduced cost in the face of increasing volumes of data.

However, parallelization of operations by cloud's hardware designers will further boost the computational efficiency of service platforms available for customer to run its applications.

Nevertheless, even with the aforementioned challenges, cloud providers are working on infrastructure to increase and promote DWH. A survey carried out in November 2012 on 511 business companies' shows that about 70% of these businesses are planning on using the cloud technology (Healey, 2012). This shows that majority of this companies concerns are cost, elasticity, availability and cost reduction. It is believed that cloud computing will evolve more in future to accommodate mission critical DWH. This will revolutionize the area of DWH and also help the small and medium sized businesses to use more analytical data because of its lower operational cost. The provision of data warehouses in the cloud will provide fast and cost-effective storage and ad hoc querying of terabytes of customers data. It can also be used for archiving purposes. There is also ease of being integrated with applications' cloud-based platforms. There should be more publicity awareness programs for the entire populace about the benefits DHWs in the cloud offer.

References

- Adriaans, P., & Zantige, D. (2001). *Data Mining* (3rd ed.). India: Addison Wesley Longman.
- Armbrust, M., Fox, A., Rean, G., Joseph, A. D., Katz, R. H., Konwinski, A., & Zaharia, M. (2009). Above the Clouds: A Berkeley View of Cloud Computing. *Technical Report No. UCB/EECS-2009-28*.
- Awoyelu, I. O. (2009). *Development of a Scalable Distributed Data Warehouse for Higher Educational Institutions*. Unpublished Thesis, Department of Computer Science and Engineering, ObafemiAwolowo University, Ile-Ife, Nigeria.
- Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems*, 25(6), 599-616. <http://dx.doi.org/10.1016/j.future.2008.12.001>
- Duncan, D., Chu, X., Vecchiola, C., & Buyya, R. (2009). *The Structure of the New IT Frontier: Cloud Computing – Part I*. Technical report. Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computer Science and Software Engineering, Melbourne.
- Gelder, K. V. (2011). *Elastic Data Warehousing in the Cloud: Is Sky really the Limit?* Technical report, Faculty of Exact Sciences, Vjire Universiteit Amsterdam, Netherlands.
- Healey, M. (2012). *Research: 2012 State of Cloud Computing*. Retrieved August 9, 2013, from <http://reports.informationweek.com/abstract/5/8658/cloud-computing/research-2012-state-of-cloud-computing.html>
- Huth, A., & Cebula, J. (2011). *The Basics of Cloud Computing*. United States Computer Emergency Readiness Team. Retrieved July 12, 2013, from http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf
- Inmon, W. H. (2002). *Building the Data Warehouse* (3rd ed.). John Wiley.
- Malathy, G., Rm, S., & Duraiswamy, K. (2013). Performance Improvement in Cloud Computing Using Resource Clustering. *Journal of Computer Science*, 9(6), 671-677. <http://dx.doi.org/10.3844/jcssp.2013.671.677>
- Monaco, A. (2012). A View inside the Cloud. *Institute: The IEEE news source* (p. 8). June.
- NIST. (2010). Cloud Computing Program. Retrieved July 12, 2012, from <http://www.nist.gov/itl/cloud/>
- Udo, I. J., Afolabi, B. S., & Akhigbe, B. I. (2012). Conceptual Process Model for Executive Information System Data Store: A communication-driven perspective. *Journal of Computer Science and Engineering*, 16(2), 9-14.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).