Cloud Data Warehousing for SMEs

Sérgio Fernandes[1] and Jorge Bernardino[1,2]
[1]ISEC – Superior Institute of Engineering of Coimbra, Polytechnic Institute of Coimbra, 3030-190 Coimbra, Portugal
[2]CISUC – Centre of Informatics and Systems of the University of Coimbra, University of Coimbra,
3030-290 Coimbra, Portugal

Keywords: Database as a Service, Cloud Computing, Cloud Databases, Database Management System, DBaaS, DWaaS.

Abstract: The emergence of cloud computing caused a revolution in the universe of Information Technology. With cloud computing solutions it is possible to access powerful features, hardware and software in less time and with considerably lower costs by using the model "pay-as-you-go". At the same time, this turnover increased information, and data warehouses must respond to this new reality. Small and medium enterprises (SMEs) were deprived of owning a traditional data warehouse due to the costs involved, but the cloud has made it possible to overcome this barrier. This paper provides an overview of Data Warehouse (DW) in the cloud and presents the main characteristics of the following solutions: Amazon Redshift, IBM dashDB, Snowflake, Teradata Active Data Warehouse Private Cloud, Treasure Data, and Microsoft Azure.

1 INTRODUCTION

In recent years, the significant growth in data volume and its complexity makes organizations look for ways to keep the focus on their business instead of in their IT infrastructure. Advances in infrastructure and technology have led companies to trust the cloud model, including Data Warehouses (DW). According to recent studies, in 2014 the use of large scale DW increased by 25% to 31% in implementing cloud-based DW services (Henschen, 2015).

A Data Warehouse (DW) is a central repository of integrated data, time-variant and non-volatile to support management decision process. This handy vision can quickly identify potential customers and to identify trends and patterns, which can also serve to support a marketing campaign or product promotions.

However, these systems require high computational resources in order to manage the large volumes of data involved and providing constantly updated information to their users.

To put that in practice the enterprise can acquire the computational power and space, however the problems are the high costs involved into the initial acquisition and maintenance. Therefore, companies began to use the services in the cloud.

Cloud computing is presented as a medium that allows easy acquisition of computational resources, scalable infrastructure with minimal setup and it is a service at a lower cost, using the model "pay-as-you-go" (Malliga, 2012). Many organizations have chosen to implement or migrate their systems, processes and data in order to reap the benefits of elasticity that this environment provides, so that IT administrators need not worry about work peaks.

The cloud models can be classified in three types: public, private and hybrid (Kaur et al., 2012). Cloud providers usually offer three different basic services: Infrastructure as a Service (IaaS); Platform as a Service (PaaS); and Software as a Service (SaaS).

While there are companies that support the cost of their data storage services, small and medium enterprises (SMEs) could not afford to have a Data Warehouse because of the associated costs. The Cloud Data Warehousing reduces costs and eliminates the heavy lifting in creating the infrastructure, allowing entrepreneurs focus more on their businesses.

Another service model, data warehouse as a service (DWaaS), is a source of new solutions that have the same or better level of performance. This paper aims to analyse the following solutions of Cloud DW for companies and assess their characteristics: Amazon Redshift, IBM dashDB,

Fernandes, S. and Bernardino, J.
Cloud Data Warehousing for SMEs.
DOI: 10.5220/0005996502760282
In *Proceedings of the 11th International Joint Conference on Software Technologies (ICSOFT 2016) - Volume 1: ICSOFT-EA*, pages 276-282
ISBN: 978-989-758-194-6

Snowflake, Teradata ADW Private Cloud, Treasure Data, and Microsoft Azure.

The remainder of this paper is organized as follows. Section 2 provides a brief evolution of database technology. Section 3 presents the data warehouse solutions in the cloud and their characteristics. In Section 4 existing solutions and its main features are compared. Finally, section 5 presents the main conclusions and suggests future work.

2 EVOLUTION OF DATABASE TECHNOLOGY

Relational database management system (RDBMS) is a DBMS based on the relational model, where information is stored in tables with rows and columns. Each row contains a unique identifying, the primary key, which can be used to create the relationship with other columns. This model is based in relationships between keys, which has as disadvantage the cost to process joins between tables. This model allows companies to store information with moderate volume of data.

With the explosion of Web 2.0 and mobile applications, Big Data comes and the industry felt the need for development of large RDBMS. One of the proposed models was the Object Oriented Database Management Systems (OODBMS) representing the objects used in object-oriented programming (Malliga, 2012), which uses all mechanisms of identity, encapsulation, inheritance, and the grouping of classes. It allows change and improve performance. This is followed some time later by the object-relational model database management system (ORDBMS) that is similar to a relational database but with a database based on object-oriented model. Currently, there are several models on the market, as described above. The most used ones are still based on traditional RDBMS (DB-Engines Ranking). The NoSQL databases increased by about 7% in 2014, due to the fast and flexible developments that made it possible to reduce deployment costs (Henschen, 2015). NoSQL engines are thought to store unstructured data and thus to avoid the limitations of structured data in the transactions of large volumes of information (Abramova and Bernardino, 2013). The evolution of the database involves migrating to the cloud.

2.1 Database for Cloud

If a small and medium enterprise (SME) needs to create a new competitive platform, it needs a database-compatible cloud. The use of shared disk enables scalability, i.e. elastic support and offering high availability for SLAs. Relational databases were not designed from the ground to the cloud or, for that matter, to deal with Big Data.

The databases available to the cloud can be based on SQL or NoSQL. The solutions market Amazon Relational Database Service (MySQL), Microsoft SQL Azure (MS SQL), Heroku, Postgres SQL, Xeround Cloud Database (MySQL) and Enterprise DB (Postgres SQL) are based on the SQL model. Already the Amazon Dynamo DB solutions, Amazon Simple DB, Database.com (SalesForce), Cassandra, MongoDB, CouchDB, Big Table, Hbase, Redis and Google App Engine Datastore are all based on NoSQL model (Malliga, 2012; Nayak et al.; Pathak, 2013).

This methodology is designated by the DataBase as a service (DBaaS). A recent Aberdeen Group study reveals that companies using cloud-based analysis systems increased user adoption by 35% to 52% as well as increased self-service Business Intelligence systems in 42% to 65%. Other advantage is the implementation of cloud solutions turns out to have a smaller operation time (Lock, 2014).

2.2 Data Warehouse for Cloud

To achieve some productive results of the analysis of a big-data it is necessary to have a data warehouse (DW). The DW is a system that permits the analysis of large volumes of data thus requiring large computational processing power. It is also necessary to have some investment or to adopt some Cloud solutions (see Figure 1). This is what led to the creation of the service model data warehouse as a service (DWaaS).

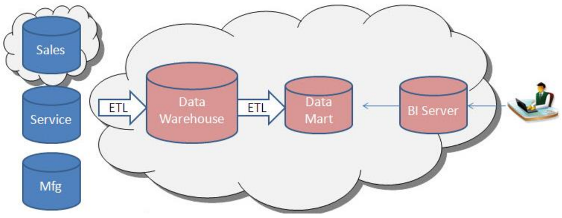

Figure 1: Data Warehouse in the Cloud (Deployment Options for Cloud Computing).

Currently, there are solutions to ensure speed, simplicity, scalability and analytical power modernizing the data warehouse and respond to today's big data requirements. In the next section, we will address some of the currently existing solutions.

3 CLOUD DATA WAREHOUSING

A data warehouse stored on the Cloud is queried via analytical queries which are usually incorporated into web services. Figure 2 exemplifies the evolution of database systems over the years and the last comes the latest model called data warehouse as a service (DWaaS).

In this section we will present the following solutions that support DWaaS and its main characteristics: Amazon Redshift, IBM dashDB, Snowflake, Teradata Active Data Warehouse Private Cloud, Treasure Data, and Microsoft Azure.

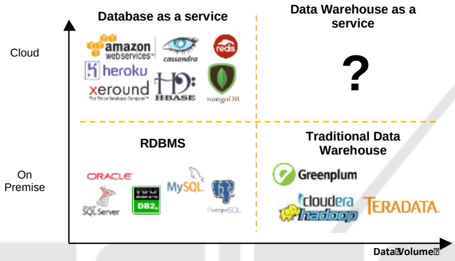

Figure 2: Evolution of database systems.

3.1 Amazon Redshift

The solution Redshift in the Amazon AWS cloud has shown to be very reliable. This platform was created in 2012 and it provides petabytes of storage in the cloud across multiple nodes. Figure 3 illustrates the Redshift architecture built around two types of nodes: a single Leader Node and a cluster of Compute Nodes. The customer may use current intelligence tools to perform quick and accurate analysis of data, usually in catalogued reports. Most administration tasks are executed in an automated platform. This solution runs on a SQL platform, based on PostgreSQL and is built to integrate with Amazon Web Services as well as Amazon S3 (Redshift). It is available in the PaaS model and is based on Massively Parallel Processing (Showflake Computing: A New Take on a Data Warehousing in the Cloud).

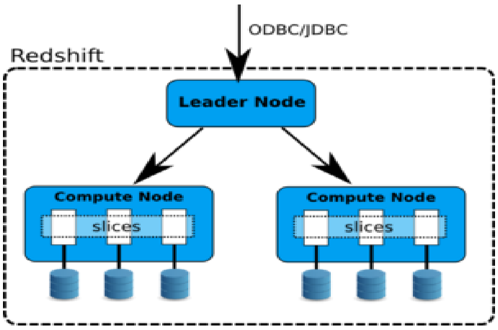

Figure 3: Amazon Redshift architecture (Redshift).

3.2 IBM dashDB

IBM presents the dashDB solution available through IBM Bluemix (Cloud Platform) or Cloudant (DBaaS). This platform ensures that robust stability is fully managed in the cloud and simple with fast provisioning at SoftLayer. Safety is also ensured by SoftLayer Secure Cloud Infrastructure.

The solution ensures flexibility supporting data volumes and processing speed. OLTP convergence includes supporting the enterprise data warehousing and analysis for the same database or in the cloud parallel processing database (see Figure 5). The performance is supported by the low latency with IBM in-memory, compression, hardware acceleration and parallel processing of analysis algorithms libraries. The dashDB supports integration DB2 database or PureData Analytics System (PDA), and IBM Watson Analytics can extract Business Intelligence (Dashdb).

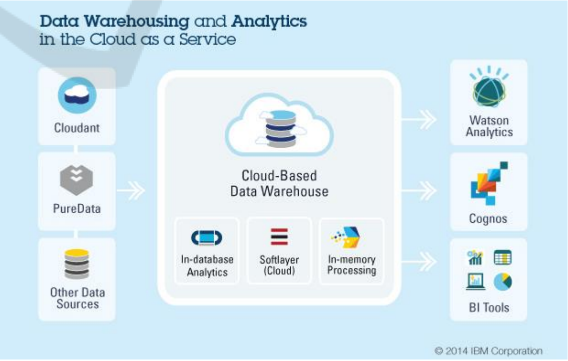

Figure 4: IBM dashDB (Dashdb).

3.3 Snowflake

The Snowflake Elastic Warehouse solution appears to be the first to create a data warehouse from scratch for the cloud. It guarantees a 90% lower cost compared to traditional storage solutions and the only solution able to take advantage of flexibility elasticity, scalability and cloud (Snowflake). The company has specialists with many years of experience that have identified the need to reinvent the databases to achieve the power of cloud computing.

Snowflake's Elastic Data Warehouse is the first in SQL data solution built for the cloud and its patented architecture separates physically and logically integrates the storage and computing. As illustrated in Figure 6, it takes full advantage of the cloud flexibility that would be impossible in a traditional implementation. The solution gathers all business data, structured or semi-structured in one place for analysis and thus allow the use of SQL tools with main focus on data analysis. The storage system is based on JavaScript Object Notation (JSON) that is currently being used in various applications.

There is a guarantee that the client does not have to spend time in the management of resources, whether hardware or software. The multi-dimensional elasticity enables optimum performance at every scale and allows to upload information and make an appointment at the same time because there are no conflicting resources (Snowflake).

Snowflake was designed from the ground in a non-MPP engine, and creates a PaaS solution that automates the design, provisioning, and data warehouse design allowing multidimensional elasticity that is one of the main advantages of cloud. The solution is available on the Amazon platform or may be accessed through a tool-dedicated connection (Showflake Computing: A New Take on a Data Warehousing in the Cloud).

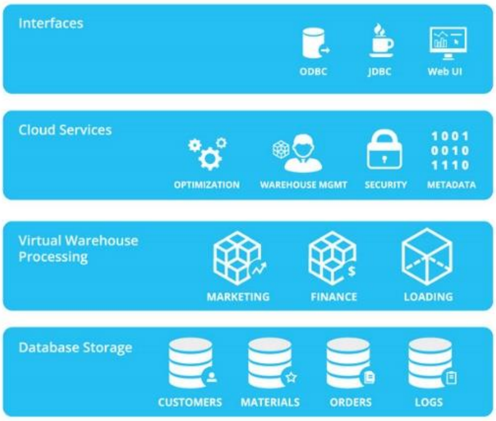

Figure 5: Snowflake logically integrates the storage and computing but there is physical separation (Snowflake).

3.4 Teradata

The Teradata Active Data Warehouse Private Cloud solution provides migration project to the cloud ensuring best practice in the management and security systems. It also provides training and customer support (Teradata).

Teradata offers a simple solution producing a single view of business. In data integration, it includes processes to ensure the integrity, cleaning or processing.

Teradata next-generation solution leverages and extends the cloud resources through virtualized resources, scalability, elasticity, self-service and consistent performance (see Figure 7).

The solution integrates ETL and BI ecosystem in a secure environment, reliable and hardware management / software for Teradata. The use of JSON allows flexibility in managing information. The company ensures that the solution is ideal for organizations that require a real-time storage or look for real-time analysis via dashboards (Teradata Turns to the Cloud, Offers Data Warehouse as a Service).

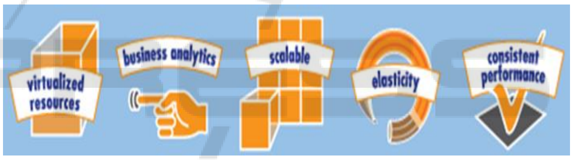

Figure 6: Teradata ADW Cloud Characteristics (Teradata Turns to the Cloud, Offers Data Warehouse as a Service).

3.5 Treasure Data

The Treasure Data Cloud service is available at a reduced price, fast to implement, easy to use without the need for specialized IT resources eliminating complexity (Treasure). The solution was developed based on Hadoop and other open source technologies in order to keep costs down. Ensures loading faster and easier data as well as processing large volumes of data quickly in real time. With Treasure it is not necessary to learn a new language, due to the use of SQL-like and the JDBC driver that allows the use of BI tools.

As shown in Figure 8, this service is available through Amazon S3 that is scalable, reliable and secure, allowing the elasticity and scalability. In addition to this, it does not require an investment in infrastructure (Treasure).

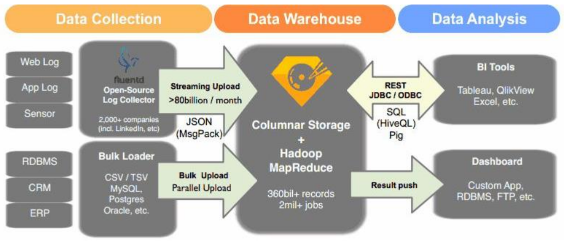

Figure 7: Treasure Data (Treasure).

3.6 Microsoft Azure

Microsoft argues that the Azure Virtual Machines platform is ideal for running SQL Server applications because of new virtual images optimized for performance and a simple configuration (Microsoft Azure), as shown in Figure 9.

The SQL Server engine was redesigned for the multiple parallel processing (MPP) with Parallel Data Warehouse (PDW). The MPP architecture enables powerful distributed computing and climbing accompanying the client's need and reduces latency improving performance (The Microsoft Modern Data Warehouse, 2013).

The company states that the warehouse relational data were not designed to handle a large volume of data. The proposed solution integrates the traditional data warehouse with non-relational data and thus supports any volume of information and real-time processing (The Microsoft Modern Data Warehouse, 2013).

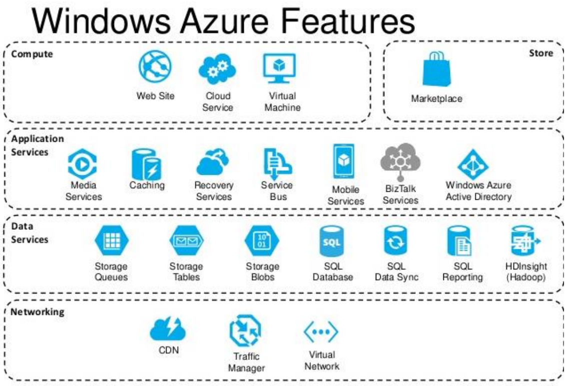

Figure 8: Microsoft Azure (The Microsoft Modern Data Warehouse, 2013).

4 EVALUATION

In this section we evaluate the platforms described in section 3. Table 1 illustrates the comparison of Cloud Data Warehouse platforms concerning the following aspects: manufacturer, security, delivery method, BI (Business Intelligence), ETL (Extract-Transform-Load), Cloud, payment model, and elasticity. These solutions in the cloud prevent SMEs to make a substantial investment in

Table 1: Cloud Data Warehouse platforms comparison.

	Manufacturer	**Security**	**Delivery Method**	**BI**	**ETL**	**Cloud**	**Payment Model**	**Elasticity**
Redshift	Amazon	Yes	?	Yes	Yes	Amazon S3	Free 750hours 1000$/Year /TB	Yes
dashDB	IBM	Yes – AES 256	?	Yes	Yes	IBM Bluemix or Cloudant	Free to 1 GB. 50,00 US$/Monthly	Yes
Snowflake	Snowflake	Yes	PaaS	Yes	Yes	Amazon e Azure	?	Yes
Teradata	Teradata		?	Yes	Yes		From 1000$ to 5000$/ Monthly	Yes
Treasure Data	Treasure Data	Yes	?	Yes	Yes	Amazon S3	?	Yes
Azure	Microsoft		PaaS	Yes	Yes	Azure	Try for free From 0.14€ to 1.9€/hour	Yes

infrastructure or hardware or software and being also independent of a supplier who cannot even meet their needs. Therefore, SMEs can experience these solutions and decide which is the most advantageous solution. In our opinion, the Snowflake solution is a good solution that was developed and designed for the cloud and by professionals with many years of experience in the field (Showflake Computing: A New Take on a Data Warehousing in the Cloud).

All other solutions identified belong to major suppliers in the area and meet the needs of building a data warehouse in the cloud. However, it is necessary to consider all the features and consulting enterprise solution for some more information and support.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented the evolution of databases, their need to shift to cloud and some existing solutions on the market for data warehousing, emerging a new paradigm, Data Wwarehouse as a Service (DWaaS).

For SMEs (Small and Medium Enterprises), which generally do not have the opportunity to acquire a lot of technology, cloud computing is attractive because it offers flexibility in purchasing additional IT resources at anytime, which is especially useful in the rapid changes in today's business model.

In this work we analyse and compare the most popular Cloud Data Warehouse platforms: Amazon Redshift, IBM dashDB, Snowflake, Teradata Active Data Warehouse Private Cloud, Treasure Data, and Microsoft Azure. All the platforms described in this paper are accessible without the concern existing in the acquisition and maintenance of IT resources.

DWaaS is seen as a good option for businesses looking for an alternative to ownership and management of the engine information. Thus, all management activities, backup, recovery, security, and performance and capacity management is replaced by a monthly cost. Some market solutions include the ETL process and BI.

The market is gaining more confidence in cloud services and SMEs need a partner to assist their membership of these services so they can move their business to the cloud.

Many service providers only provide BI solutions and the market is still immature, alternative services varied widely in both cost and performance of the type of services offered by leading vendors discussed in the previous section.

As future work we pretend to evaluate these solutions in a SME with real information doing a complete assessment. This evaluation will be based on a standard performance benchmark, but also analysing the interface offered by platforms and the initial steps necessary to build a complete data warehouse.

REFERENCES

Malliga, P., "Database Services for Cloud Computing – An Overview" *International Journal of Computers & Technology, Volume 2 No. 3*, June, 2012.

Redshift [online] http://aws.amazon.com/pt/redshift/ (Accessed June 2015).

Dashdb [online] http://www-01.ibm.com/software/data/dashdb/what-is.html (Accessed February 2016).

Snowflake [online] http://www.snowflake.net/ (Accessed February 2016).

Teradata [online] http://www.teradata.com/Teradata-ADW-Private-Cloud/ (Accessed February 2016).

Treasure [online] https://www.treasuredata.com/ (Accessed February 2016).

Microsoft Azure [online] https://azure.microsoft.com/ (Accessed February 2016).

Kaur, Harjeet, Agrawal, Prateek and Dhiman, Amita, "Visualizing Clouds on Different Stages of DWH – An Introduction to Data Warehouse as a Service", *2012 Int. Conference on Computing Sciences*.

DB-Engines Ranking [online] http://db-engines.com/en/ranking (Accessed February 2016).

Nayak, A., Poriya, A. and Poojary, D., "Type of NoSQL Databases and its Comparison with Relational Databases", *International Journal of Applied Information Systems (IJAIS).*

Pathak, A. R., "Survey of Confidentiality and Integrity in Outsourced Databases" *Int. Journal of Scientific Engineering and Technology*, 1 April 2013.

Showflake Computing: A New Take on a Data Warehousing in the Cloud [online] http://tdwi.org/articles/2015/02/17/snowflake-computing.aspx (Accessed February 2016).

Henschen, D., "5 Analytics, BI, Data Management Trends For 2015" [online] https://www.ironsidegroup.com/2015/02/02/top-5-trends-in-cloud-data-warehousing-and-analytics-for-2015/ (Accessed March 2016).

Lock, M., "Rapid Insight with Results: Harnessing Analytics in the Cloud", June 2014.

Deployment Options for Cloud Computing [online] http://www.b-eye-network.com/blogs/eckerson/archives/2011/07/deployment_opti.php (Accessed March 2016).

Teradata Turns to the Cloud, Offers Data Warehouse as a Service [online] http://www.cio.com/article/2381531/business-intelligence/teradata-turns-to-the-cloud--offers

-data-warehouse-as-a-service.html (Accessed March 2015).

Microsoft Corporation, “The Microsoft Modern Data Warehouse”, 2013. (Accessed March 2016) Available http://download.microsoft.com/download/c/2/d/c2d2d5fa-768a-49ad-8957-1a434c6c8126/the_microsoft_modern_data_warehouse_white_paper.pdf.

Abramova, V. and Bernardino, J., 2013. NoSQL databases: MongoDB vs Cassandra. In *International C* Conference on Computer Science and Software Engineering (C3S2E '13). ACM*, New York, USA.

Neves, P. C. and Bernardino, J.. Big Data in the Cloud: A Survey. *Open Journal of Big Data (OJBD) Vol 1 (2)*, pp. 1-18, 2015.