Data Warehousing in the Cloud: Amazon Redshift vs Microsoft Azure SQL

Pedro Joel Ferreira[1], Ana Almeida[1] and Jorge Bernardino[2]
[1]*Instituto Superior de Engenharia do Porto, Rua Bernardino de Almeida, Porto, Portugal*
[2]*Instituto Superior de Engenharia de Coimbra, Rua Pedro Nunes, Coimbra, Portugal*

Keywords: Cloud Computing, Data Warehousing, Cloud Data Warehousing.

Abstract: A data warehouse enables the analysis of large amounts of information that typically comes from the organization's transactional systems (OLTP). However, today's data warehouse systems do not have the capacity to handle the massive amount of data that is currently produced, then comes the concept of cloud computing. Cloud computing is a model that enables ubiquitous and on-demand access to a set of shared or non-shared computing resources (such as networks, servers, or storage) that can be quickly provisioned or released only with a simple request and without human intervention. In this model, the features are almost unlimited and in working together they bring a very high computing power that can and should be used for the most varied purposes. From the combination of both these concepts, emerges the cloud data warehouse. It advances the way traditional data warehouse systems are defined by allowing their sources to be located anywhere as long as it is accessible through the Internet, also taking advantage of the great computational power of an infrastructure in the cloud. In this paper, we study two of the most popular cloud data warehousing market solutions: Amazon Redshift and Microsoft Azure SQL Data Warehouse.

1 INTRODUCTION

Data warehouses (DW) are defined as customized data storage that aggregate data from multiple sources and store it in a common location to be able to run reports and queries over it. This concept arose from the requisite to integrate corporate data across multiple application servers that an organization might have, so that it would be possible to make data accessible to all users who need to consume information and make decisions based on it. Many companies use data warehouses to compile regular financial reports or business metric analyses.

Integration of a Cloud Data Warehouse solution demands a very well defined strategy that would involve Cloud Computing capabilities. The success of the implementation depends on the existence of a service-oriented strategy at the organization level, which would provide the necessary infrastructure for the Cloud implementation. Without SOA and BPM, integration of a Data Warehousing solution based on Cloud Computing serves no financial purpose, involving high costs for the present systems reengineering. Also, in order to be successful, Cloud strategy has to be led according to the business strategy of the organization.

One of the biggest challenges with data warehousing in the Cloud is how the data is transferred up into it. Pumping gigabytes, terabytes, or even petabytes of data up into Cloud over the public Internet can not only come with security concerns, but also performance challenges.

When selecting a Data Warehousing solution, we have to take into account the newest trends on the DW and Cloud Computing market, the present and future needs and the opportunity of integration. Therefore, in order to be successful, the selection of a Cloud DW solution has to be achieved objectively based on good criteria that have been analyzed and weighted according to the present and future needs of the organization. Cloud Data Warehousing is a potential cost saver for big companies, and removing a cost barrier that have held data warehousing back from small and mid-sized businesses.

Cloud Data Warehouse must be designed to take away the undifferentiated heavy lifting of running infrastructure at heavy scale, allowing the customers to focus on their core competencies – its business.

The growing interest in cloud-based data warehousing is driven by the high return on investment. Nonetheless, the adoption of cloud computing for data warehousing faces security

Ferreira P., Almeida A. and Bernardino J.
Data Warehousing in the Cloud: Amazon Redshift vs Microsoft Azure SQL.
DOI: 10.5220/0006587103180325
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 318-325
ISBN: 978-989-758-271-4

challenges given the proprietary nature of the enclosed data. Moving to the cloud is a hard decision not only because the data owners like to have their data near to them (on-premises) but also due to security and data confidentiality issues, because of this last two the decision makers tend to delay the decision of moving to the cloud.

There are many benefits to implement a DW in the cloud, such as cost, efficiency, elasticity, flexibility in the choice of provider, more competitiveness and less time involved in installation and maintenance. It is assumed that cloud system efficiency is even more efficient with the use of parallel computing and the cloud is motivation for small and medium enterprises by the possibility of expansion at an affordable cost.

In this paper we analyze the characteristics of two of most popular cloud data warehousing platforms: Amazon Redshift and Microsoft Azure SQL Data Warehouse. The remainder of this paper is structured as follows. Section 2 describes the related work. Section 3 presents the cloud computing and Section 4 presents a summary of cloud data warehousing area. Section 5 presents two cloud data warehousing market solutions: Amazon Redshift and Microsoft Azure SQL Data Warehouse. Section 6 presents a brief comparison of the cloud data warehouse solutions. Finally, conclusions and future work are summarized in Section 7.

2 RELATED WORK

Multiple research works have been done to compare and evaluate existing Big Data platforms. However, most of the research focus only on a specific capability, technology or purpose.

Almeida and Bernardino (2015a; 2015b) focus on the capability of mining data, and in a mix of technical parameters and features that are suitable for Small and Medium Enterprise environments. On the other hand, Morshed et al. (2016) focused their work on platforms addressing distributed real-time data analytics, and concluded that the platforms present on their research do not cover all the features that are required for distributed computation in real-time.

Kaur et al. (2012) describe some solutions from multiple vendors in the cloud to support a Data Warehouse. They state that each service provider implements different levels of use, some provide ETL mechanisms or Business Intelligence (BI) solutions while others provide the storage infrastructure and the client then chooses the supplier for their needs and the services required. In the paper it is considered that the market is still immature and services vary in price and performance.

Hemlata Verna (2013) is concerned about how to manage the data through recycle, reuse, reduce and recover information in cloud environment. Popeangã (2014) concentrate her work on studying what architecture is suited for a data warehouse in the cloud.

Mathur et al., (2011), focused their work on distributed databases in the cloud. They propose IaaS cloud servers that can be used to store these databases at low initial cost.

Therefore, there are few related works which evaluate solutions based on specific capabilities, technology or purpose. Our work will be of a broader scope in functionalities and applications in order to be used by SMEs.

3 CLOUD COMPUTING

There is no definitive definition of cloud computing, but Kimball and Ross (2013) propose that "Cloud computing is a self-provisioning and on-demand delivery of computing resources and applications over the Internet with a pay-as-you-go pricing model".

Cloud, by definition, is a self-service system that allows end users to setup applications and services in a cloud computing environment without the intervention of an IT service provider (Armbrust et al., 2010). Cloud computing addresses large amounts of computing power by aggregating computing resources and offering a single view of the system.

Cloud brought new features and possibilities to improve its user's life but also originated a new market segment for IT with new business opportunities. Many organizations build their business around the cloud not only to use its services but also to offer business solutions in this environment. Like any other service, we can identify at least two actors that are directly connected with it:

- User or consumer: is the one who uses the functionalities or resources provided by the cloud. It is the entity or organization that uses the cloud computing services, whether they are related with software, platform or infrastructure;
- Provider: is the entity or organization responsible for making cloud services available to consumers. The provider is also responsible for managing the necessary infrastructure which supports its services.

There are different cloud implementation models and they are the different means by which consumers access cloud services. These models are different and are characterized by the target audience they serve. If this assumption is taken into account, there are only two models: private clouds and public clouds.

Private clouds are typically used in the distribution of services in an internal organization environment. The infrastructure that supports the private cloud can be delegated to a service provided and stays private at all, it has not to be acquired and managed by the private cloud users themselves.

When managed internally by users themselves, they have full control over the processes, data or applications used in the cloud. However, they lose some of the general principals or benefits of cloud computing, such as access to infrastructure at reduced prices, elasticity, resources availability or rapid deployment times.

Unlike private clouds, public clouds are used for the distribution of services to the general public, typically via the Internet. The service provider is responsible for the entire support infrastructure. This infrastructure is fully shared by the various users of this cloud. Thanks to this sharing and optimization of resource management processes, providers are able to maximize their use and enable them to be supplied at reduced prices to consumers.

4 CLOUD DATA WAREHOUSING

Nowadays, companies have a greater collection of data than ever before. This includes a huge variety of sources, including cloud-based applications or even company datamarts. In order to make good decisions, get insights and achieve a competitive advantage, companies need to have their data properly analyzed, on time.

The conventional data warehouse architecture is widespread in a large number of companies working with massive and diverse data sets but is a very closed and complex model to respond with the agility that companies currently need (Tereso and Bernardino, 2011). People who make analysis need to wait a number of hours or even days for data to flow into the data warehouse before it becomes available for analysis. In most cases, the storage and compute resources required to process that data are insufficient (or the same) and this leads to hanging or crashing systems (Goutas et al., 2016).

One of the major concerns on moving to the cloud is the time to "live" with both on-premises and cloud data warehouse system because it is not a good idea to move at once the whole data warehouse. To ease this concern, a data virtualization solution can be used to help out the migration and coexistence of the both data warehouse systems while migration to cloud is ongoing.

Cloud data warehousing was born from the convergence of three trends – huge changes in data sources, volume and complexity; the need for data access and analytics; and better technology that increased the efficiency of data access, analytics and storage. Traditional data warehouse systems were not designed to handle the volume, variety and complexity of today's data (Almeida et al., 2008).
A data warehouse in the cloud is a database which information is consumed over the Internet, a typical database as a service (DBaaS). Cloud data warehousing is a cost-effective way for organizations to use and take advantage of high technology without high upfront costs to purchase, install and configure the required hardware, software and infrastructure (Talia, 2013).

In the next section we will analyse two of the most popular cloud data warehousing market solutions.

5 CLOUD DATA WAREHOUSING MARKET SOLUTIONS

In this section, the architecture of two cloud data warehousing solutions is described.

The following sections describe the characteristics of most popular platforms: Amazon Redshift and Microsoft Azure SQL Data Warehouse.

5.1 Amazon Redshift

Gartner reports Amazon Web Services (AWS) is often considered the leading cloud data warehouse platform-as-a-service provider (Gartner, 2016a).

Recognized by Gartner as a leader, Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse that makes simple and cost-effective to analyze all our data using existing business intelligence tools.

Amazon Redshift engine is a SQL-compliant, MPP, query processing and database management system designed to support analytics workload. The storage and compute is distributed across on or more compute nodes (Gupta et al., 2015).

The core infrastructure of Amazon Redshift data warehouse is a cluster and it is composed by:

- A leader node;
- One or more compute nodes.

The leader node accepts connections from the client applications and dispatch the work to the compute node: it parses and develops execution plans to carry out database operations and based on the execution plan it compiles code, distributes the compiled code to the compute nodes and assigns a portion of the data to each node (see Figure 1).

The leader node distributes SQL statements to the compute nodes only when a query references tables that are stored on the compute nodes, otherwise they run exclusively on the leader node ("Data Warehouse System Architecture - Amazon Redshift," n.d.).

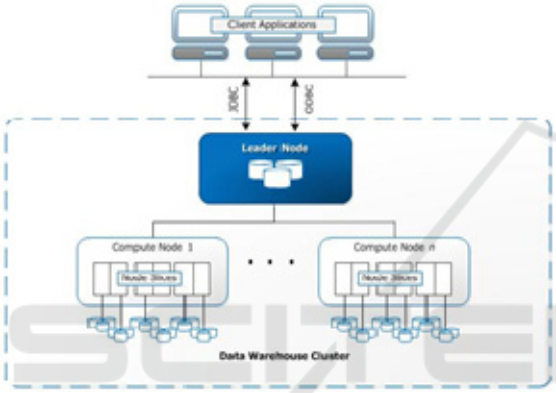

Figure 1: Amazon Redshift system architecture (source: aws.amazon.com/redshift).

The compute nodes execute the compiled code sent by the leader and send the results back for final aggregation. Each compute node has its own dedicated CPU, memory and storage – it is easy to scale the cluster by upgrading the compute nodes or adding new ones. The minimum storage for each compute is 160GB and scale up to 16TB to support a petabyte of data or more. The compute node is partitioned into slices and each of it is allocated a portion of the node's memory and disk space, where it processes a portion of the workload assigned to the node – the leader node manages the distribution of data of the workload to the slices and then they work in parallel to complete the operation.

The cluster contains one or more databases. Amazon Redshift is a relational database management system and provides the same functionality as a typical RDBMS including all related with OLTP but it is optimized for high-speed performance analysis and reporting of very large datasets ("Data Warehouse System Architecture - Amazon Redshift," n.d.).

The database engine is based on PostgreSQL. Another interesting characteristic of Amazon Redshift is that it is a columnar database, which means that each record is not saved as a unique block of data but it is stored in independent columns. The query performance can greatly be improved by selecting a limited subset of columns rather than the full record.

The data warehouse functionality is comparable to the high level databases. The ease of use and scalability of Redshift is definitely a huge advantage of this solution.

5.2 Microsoft Azure SQL Data Warehouse

Microsoft Azure SQL data warehouse is a cloud-based, scale-out database capable of processing massive volumes of data, both relational and non-relational. It is a massively processing (MPP) distributed database system. "It provides SaaS, PaaS and IaaS services and supports many different programming languages, tools and frameworks, including non-Microsoft software ("SQL Data Warehouse | Microsoft Azure," n.d.).

SQL Data Warehouse is based on the SQL Server relational database engine and integrates with the tool that its users may be familiar with. This includes ("SQL Data Warehouse | Microsoft Azure," n.d.):

- Analysis Services;
- Integration Services;
- Reporting Services;
- Cloud-based tools.

The Microsoft Azure SQL Data Warehouse is composed by a Control Node, Compute nodes and Storage. It also has a service called Data Movement Service that is responsible for the data movement between the nodes ("SQL Data Warehouse | Microsoft Azure," n.d.).

Like the Leader Node of Amazon Redshift, the Azure Control Node manages and optimizes queries and is responsible for the coordination of all the data movement and computation required to run parallel queries (see Figure 2). When a request is made to SQL Data Warehouse, the control node transforms it into separate queries that run on each compute node in parallel.

The compute nodes are SQL databases that store the data and process queries. When data is added, it is distributed to the compute nodes and when the data is requested these nodes are the workers that run queries in parallel. After processing, they pass

the results back to the control node so it can aggregate the results and return the final result to the user.

All the data stored in Azure SQL Data Warehouse is stored in Azure Blob Storage – it is a service that stores unstructured data in the cloud as objects/blobs. Blob Storage can store any type of text or binary data, such as a document, media file or application installer. When compute nodes interact with data, they write and read directly to and from blob storage. Compute and Storage are independent.

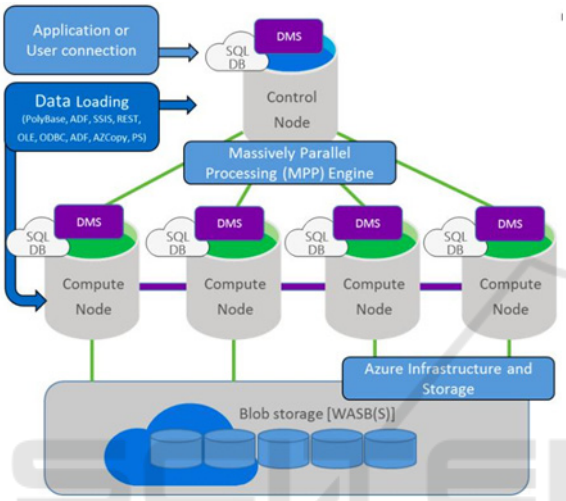

Figure 2: Microsoft Azure SQL Data Warehouse system architecture (source: "SQL Data Warehouse | Microsoft Azure," n.d.).

As previously referred, Data Movement Service (DMS) is responsible for all the data movements between the nodes. It gives the compute nodes access to data they need for joins and aggregations. It is not an Azure service but a Windows service that runs alongside SQL Database on all the nodes and it is only visible on query plans because they include some DMS operations since data movement is necessary to run a query in parallel.

Azure is an enterprise-level SQL data warehouse that extends the SQL Server family of products and services into the cloud. Azure can also scale storage and computing so the customers only have to pay for what they require.

6 COMPARING CLOUD DATA WAREHOUSES

Data warehouses in the cloud are gaining popularity because cloud vendors offer DW services at lower costs. While Amazon Redshift is the number one in the market, according to Gartner (2016), Azure offers a competitive platform to be considered.

Redshift and Azure SQL Data Warehouse both support petabyte scale systems. Both of them have leader or control nodes and compute nodes. The biggest difference between Azure SQL Data Warehouse and Redshift is the decoupling of storage and compute resources.

About scalability, in Redshift, when the cluster is modified, the changes are immediately applied. While the new clusters are being provisioned, the current cluster is available in read only mode, so during this process the data is available for read operations. After the new clusters are provisioned, the data is copied.

In Azure SQL Data Warehouse, the scaling of the clusters can happen in few minutes. The scale out can be done for compute and storage units independently. Azure SQL Data Warehouse also supports pausing a compute operation. There is no cost applied when the compute nodes are in pause state; only a storage cost is charged.

From Data sources, data can be integrated with Redshift from Amazon S3 storage. If there is an on-premises database to be integrated with Redshift, the data needs to be extracted from the database to a file and then imported up into S3.

Azure SQL Data Warehouse is integrated with Azure Blob storage. It uses a similar approach as Redshift to import the data from SQL Server. The SQL Server data is exported to a text file and then copied across to Azure Blob storage.

Comparing the adoption of public clouds, in particular AWS and Azure, we can see that AWS is the number one adopted cloud solution for the most respondent users of the 2017 Rightscale survey, as shown in Figure 3.

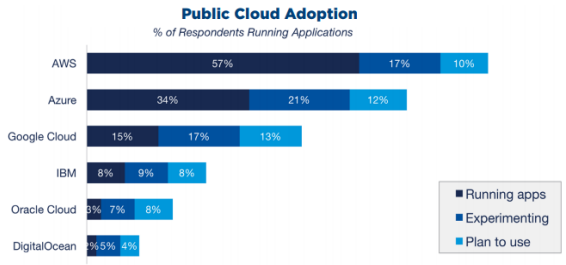

Figure 3: Adoption of public cloud in 2017 (source: "RightScale 2017 - State of the cloud report," 2017).

Although AWS continues to lead in public cloud adoption (57 percent of respondents currently run applications in AWS), this number has stayed the same as in 2016.

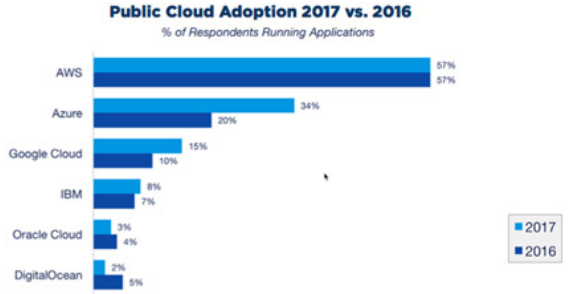

Figure 4: Adoption of public cloud in 2017 vs 2016 (source: "RightScale 2017 - State of the cloud report," 2017).

In contrast, over the last year, happened a significant growth in the percentage of respondents running applications in Azure and Google, the second and third public cloud providers (see Figure 4). Overall Azure adoption grew from 20 to 34 percent of respondents, reducing the AWS lead. Google also increased from 10 to 15 percent.

A comparison of system properties is shown in Table 1.

RedShift and Azure SQL have a database model based on relational database management system (RDBMS) which supports the relational data model.

Amazon Redshift is built around industry-standard SQL, with added functionality to manage very large datasets and support high-performance analysis that data. Although it is based on PostgreSQL, there are some unsupported features, data types and functions. Some SQL features are also implemented differently, for example:

- CREATE TABLE
- ALTER TABLE
- INSERT, UPDATE and DELETE

Amazon Redshift does not support tablespaces, table partitioning, inheritance, and certain constraints. The Amazon Redshift implementation of CREATE TABLE enables users to define the sort and distribution algorithms for tables to optimize parallel processing. ALTER COLUMN actions are not supported. ADD COLUMN supports adding only one column in each ALTER TABLE statement. Using INSERT, UPDATE, and DELETE the WITH is not supported. For the complete list of unsupported features, data types and functions, our suggestion is to check on AWS documentation, in particular the one that concerns about Amazon Redshift and PostgreSQL ("Amazon Redshift and PostgreSQL - Amazon Redshift," n.d.).

Table 1: Comparison Redshift Vs Azure SQL Data Warehouse.

System Properties	**Amazon Redshift**	**Microsoft Azure SQL Data Warehouse**
Database model	Relational DBMS	Relational DBMS
Developer	Amazon (based on PostgreSQL)	Microsoft
Licence	Commercial	Commercial
Cloud based	Yes	Yes
Implementation language	C	C++
XML support	No	Yes
SQL standard support	Does not fully support	Yes
Supported programming languages	All languages supporting JDBC/ODBC	.Net, Java, JavaScript PHP, Python Ruby
Server-side scripts	User defined Python functions	Transact SQL
Support for concurrent data manipulation	Yes	Yes
MapReduce API support	No	No
In-memory support	Yes	No
Control over node configuration	Yes	No

In-Memory OLTP is a technology for optimizing performance of transaction processing, data ingestion, data load, and transient data scenarios. In-Memory support is available on RedShift but not in Azure SQL Data Warehouse because it is only available for OLTP workloads in SQL Server since version 2014 and Azure SQL Database.

In MapReduce support property, both databases do not offer an API for user-defined Map/Reduce methods. In case of RedShift, it is possible to combine MapReduce with RedShift by processing input data with MapReduce and import results to RedShift. In case of Azure SQL Data Warehouse, Polybase unifies data in relational data stores with non-relational ones, combining data from both

RDBMS and Hadoop so that users don't need to understand HDFS or MapReduce.

Redshift and Azure SQL Data Warehouse offer many similar capabilities, so it is not necessarily a matter of one provider being better or worse than the other. It all depends on what our business needs, but each solution has pros and cons. See Table 2 to find some pros and cons of Amazon RedShift and Microsoft Azure SQL Data Warehouse cloud solutions.

Table 2: Pros and Cons of Redshift and Azure SQL Data Warehouse.

	Pros	Cons
Amazon Redshift	Performance through use of local storage	Compute cannot be scaled independent of storage (and vice versa)
	Loading data from S3 is very fast	Can't pause resources
	Beyond Petabyte	Queries that require joins against multiple columns can suffer in performance
	Columnar data store allows high performance queries on large volumes of data	
	Familiarity with PostgreSQL makes adopting Redshift easier	
Microsoft Azure SQL Data Warehouse	Resources can be paused during idle time in workload	Can only run 32 concurrent queries (maximum)
	Scale separate compute and storage resources and pay only for what is used	Not fully supports T-SQL
	Excellent integration with Azure Services	

7 CONCLUSIONS AND FUTURE WORK

Data warehouses are defined as customized data storage that aggregate data from multiple sources and store it in a common location to be able to run reports and queries over it. Many companies use data warehouses to compile regular financial reports or business metric analyses.

In this paper we analyse cloud data warehousing area that is the convergence of three trends – huge changes in data sources, volume and complexity; the need for data access and analytics; and better technology that increased the efficiency of data access, analytics and storage. Traditional data warehouse systems were not designed to handle the volume, variety and complexity of today's data.

The integration of a Cloud DW solution needs a very well defined strategy that would involve Cloud Computing capabilities. The success of the implementation depends on the existence of a service-oriented strategy at the organization level, which would provide the necessary infrastructure for the Cloud implementation.

In this study we conclude that there are several challenges when deploying data warehouses into the cloud:

- Importing data for the data warehouse into the cloud for storage can be a challenge, because when using the cloud, a customer is dependent on the Internet connection and the infrastructure of the cloud provider. It can be necessary to use a dedicated communication line to mitigate the connection problems but it as a cost;
- Getting large amounts of data from cloud storage to compute nodes provided by the cloud solution for computing can lead to a performance issue;
- Loss of control can lead to issues involving security and trust.

In this work we analyse and evaluate two of the most popular cloud data warehousing solutions: Amazon Redshift and Microsoft Azure SQL Data Warehouse.

Some conclusions concerning the two cloud data warehouse solutions have been taken:

- Using Redshift to scale our data warehouse we must increase both the compute and storage units. With Azure SQL DW, compute and storage is decoupled so we can scale them individually. This is a very

different economic model that can save customers a lot of money as they don't have to purchase additional storage when they just need more compute power, or vice-versa.

- Azure SQL DW has the ability to pause compute when not in use so we only pay for storage, as opposed to Redshift in which we are billed 24/7 for all the virtual machines that make up the nodes in our cluster.
- RedShift is easier to configure than Azure SQL Data Warehouse and takes less time to be online and available after its setup.

As future work we intend to analyze these two platforms with data from a company, and a recommendation will be given of what is the best cloud data warehouse solution in the market based on a set of criteria.

REFERENCES

Almeida, R., Vieira, J. Vieira, M. Madeira, H. and Bernardino, J. "Efficient Data Distribution for DWS". In *International Conference on Data Warehousing and Knowledge Discovery - DaWaK, pages 75–86, 2008.*

Almeida, P., and Bernardino, J. "Big Data Open Source Platforms". BigData Congress 2015: 268-275

Almeida, P., and Bernardino, J. "A comprehensive overview of open source big data platforms and frameworks", *International Journal of Big Data (IJBD), 2(3), 2015, pp. 15-33.*

Amazon Redshift and PostgreSQL - Amazon Redshift [WWW Document], n.d. URL http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html (accessed 9.2.17).

Amazon Redshift vs. Microsoft Azure SQL Data Warehouse vs. Microsoft Azure SQL Database Comparison [WWW Document], n.d. URL https://db-engines.com/en/system/Amazon+Redshift%3BMicrosoft+Azure+SQL+Data+Warehouse%3BMicrosoft+Azure+SQL+Database (accessed 9.2.17).

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M., 2010. A View of Cloud Computing. Communications of ACM 53, 50–58.

Combining Hadoop/Elastic Mapreduce with AWS Redshift Data Warehouse [WWW Document], n.d. URL http://atbrox.com/2013/02/25/combining-hadoop elastic-mapreduce-with-aws-redshift-data-warehouse/ (accessed 9.3.17).

Data Warehouse System Architecture - Amazon Redshift [WWW Document], n.d. URL https://docs.aws.amazon.com/redshift/latest/dg/c_high _level_system_architecture.html (accessed 1.1.17).

Database Manag. Solut. Anal. URL https://www.gartner.com/doc/reprints?id=1-2ZFVZ5B&ct=160225&st=sb (accessed 1.2.17).

Gartner, 2016. Magic Quadrant for Data Warehouse and Database Management Solutions for Analytics [WWW Document]. *Magic Quadr. Data Wareh.*

Goutas, L., Sutanto, J., Aldarbesti, H., 2016. The Building Blocks of a Cloud Strategy: Evidence from Three SaaS Providers. *Communications of ACM 59, 90–97.*

Gupta, A., Agarwal, D., Tan, D., Kulesza, J., Pathak, R., Stefani, S., Srinivasan, V., 2015. Amazon Redshift and the Case for Simpler Data Warehouses, in: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15. ACM, New York, NY, USA, pp. 1917–1923.*

Hemlata Verna, 2013. Data-warehousing on Cloud Computing, in: *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013.*

Kaur, H., Agrawal P., and Dhiman, A., "Visualizing Clouds on Different Stages of DWH - An Introduction to Data Warehouse as a Service," *2012 Int. Conf. on Computing Sciences, Phagwara, 2012, pp. 356-359.*

Key Concepts & Architecture Snowflake Documentation [WWW Document], n.d. URL https://docs.snowflake.net/manuals/user-guide/intro-key-concepts.html (accessed 2.16.17).

Mathur, A., Mathur, M. & Upadhyay, P., 2011. Cloud Based Distributed Databases: The Future Ahead. In: *International Journal on Computer Science and Engineering (IJCSE), 3(6), pp.2477-81.*

Miller, J. A., Bowman, C., Harish, V.G., Quinn, S., 2016. Open Source Big Data Analytics Frameworks Written in Scala, in: *2016 IEEE International Congress on Big Data (BigData Congress), pp. 389–393.*

Morshed, S. J., Rana, J., Milrad, M., 2016. Open Source Initiatives and Frameworks Addressing Distributed Real-Time Data Analytics, in: *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Presented at the 2016 IEEE Int. Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 1481–1484.*

Popeangã, J., 2014. Shared-Nothing Cloud Data Warehouse Architecture, in: *Database Systems Journal vol. V, no. 4/2014.*

RightScale 2017 - State of the cloud report [WWW Document], 2017. URL: https://assets.rightscale.com/uploads/pdfs/RightScale-2017-State-of-the-Cloud-Report.pdf (accessed 8.11.17).

SQL Data Warehouse, MicrosoftAzure [WWW Document], n.d. URL https://azure.microsoft.com/en-us/services/sql-data-warehouse/ (accessed 11.13.16).

Talia, D., 2013. Clouds for Scalable Big Data Analytics. Computer 46, 98–101. doi:10.1109/MC.2013.162

Tereso, M., and Bernardino, J. "Open source business intelligence tools for SMEs". *Information Systems and Technologies (CISTI), 6th Iberian Conference on, IEEE (2011) 1–4.*