Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2023

Benchmark of Market Cloud Data Warehouse Technologies

Jorge Oliveira e Sá[a]*, Renata Gonçalves[b], Claus Kaldeich[c]

†

[a]*Algoritmi Center, University of Minho, Guimarães, Portugal*
[b]*University of Minho, Guimarães, Portugal*
[c]*Universidad Paraguayo-Alemana, Asunción, Paraguay*

Abstract

Over the past two decades, the way computing resources are been developed, deployed, upgraded, and applied changed dramatically, with more and more software and hardware solutions being transferred to cloud technologies. Data Warehouses (DW), defined as a way of organizing corporate data in an integrated manner over (sequential) time periods, "structured & disposed" in order to generate a "single data source", were also affected by the evolution, thus giving rise to the concept of Cloud Data Warehouse (CDW). This technology allows users to be more technologically free, as they do not need to spend time investing in software and hardware, they only pay for the resources they used and the infrastructure itself has greater flexibility and scalability. However, selecting the most suitable platform or technology for a CDW can be a complex task due to the large number of factors that can influence the decision and due to the existing offer in the market.
The objective of this paper is to describe the process of benchmarking a set of CDW platforms, with the goal of analyzing and exposing each one's performance results. These platforms are Snowflake, Google BigQuery, Amazon Redshift, and Azure Synapse. The metrics to be measured are data loading and query running time, and alias running times. For this benchmark, the dataset used was Star Schema Benchmark (SSB), a dataset based on the well-known TPC Benchmark™ H (TPC-H).

* Corresponding author. Tel.: +351253510315 fax: +351253510300.
E-mail address: jos@dsi.uminho.pt

Keywords: Data Warehouse; Cloud Computing; Cloud Data Warehouse; Cloud Data Warehouse Technologies

1. Introduction

In today's society, the effective organization of data to maximize its utility is paramount. Data management has thus become a critical function within organizations. Proper data handling not only streamlines operations but also furnishes valuable insights for strategic management, aiming for optimization and cost reduction.

For organizations grappling with vast datasets scattered across multiple systems, Data Warehouses (DW) offer a solution. A DW consolidates enterprise data into a unified, time-sensitive repository, facilitating streamlined data analysis.

The landscape of computing resources has undergone a seismic shift in the last two decades, with a significant migration of software and hardware solutions to cloud technology. This evolution has given rise to Cloud Data Warehouses (CDW). A CDW represents a physical infrastructure managed by a cloud service provider, obviating the need for customers to make initial investments in hardware or software. This hands-off approach allows users to focus on data analysis.

The surge in CDW service providers makes choosing the most suitable technology a complex endeavour. Numerous factors influence this decision, and the market offers a multitude of options. While CDW technologies share commonalities, key differences necessitate thorough investigation. For organizations, a benchmark review of CDW technologies can be invaluable, leveraging a pre-established knowledge base and a defined set of metrics to facilitate the decision-making process.

This article aims to conduct a benchmark analysis of a selection of CDW technologies currently available in the market. It involves setting up CDW environments for each technology, employing a predefined dataset, and evaluating them using a set of metrics. The goal is to derive insights into the performance of each CDW technology, with the following results:

- Provide a literature review on cloud data storage, elucidating its core concepts.
- Introduce prominent CDW technologies and offer insights.
- Establish CDW environments using four different CDW technologies with a predefined dataset.
- Share the benchmark results, incorporating a set of metrics, to assess the performance of each CDW technology.

2. Literature Review

The concepts associated with Cloud Computing (CC), an emerging field of computer science that takes the Information Technology (IT) sector to a new level, are investigated. Next, several topics about DW and CDW are studied, thus presenting their main characteristics and advantages, in order to obtain a basic knowledge about the topic.

2.1. Cloud Computing

Cloud Computing (CC) has revolutionized the IT industry, offering a dominant model for IT resource provisioning. It provides access to a vast pool of resources, including servers and storage [1]. Over the past two decades, CC has become essential across various sectors, from education to industries [2]. This shift from physical products to service-oriented delivery has prompted organizations to migrate their IT resources to the cloud, driven by advantages like cost reduction and enhanced agility [3,4].

2.2. Data Warehouse

DW is the systematic management of corporate data. It organizes data with a focus on its evolving history and non-volatile integration [5]. DW takes a subject-oriented approach, simplifying data analysis and decision-making within specific business areas [5]. A central feature is that DW becomes the sole source of data for decision-making,

consolidating information from various sources [5]. DW empowers companies to identify trends, enhance market competitiveness, and boost profits [6]. Key DW components include databases, ETL tools, metadata, data marts, and access/reporting tools [5]. These elements form a robust infrastructure for harnessing data's strategic power.

2.3. Cloud Data Warehouse

CDW have revolutionized data management, forcing companies to reconsider their on-premises DW [7]. CDW, as defined by the International Business Machine Corporation [8], is a cloud-managed physical infrastructure. Customers no longer need to invest in hardware or software, and they can focus on data analysis without technical concerns.

CDW offers flexibility, enabling easy adjustments to resource needs, user numbers, and geographic locations [9]. It boasts attributes like elasticity, scalability, reliability, and availability. CDW providers often offer a comprehensive ecosystem, enhancing user operations. Multi-tenancy allows multiple users to access and utilize the CDW efficiently [7].

2.4. Advantages of the Migration to a Cloud Data Warehouse

Processing vast amounts of data in Data Warehouses (DW) demands substantial processing power and storage, which can be challenging for IT departments to provide consistently. Cloud Computing (CC) offers a solution by enabling scalability during peak periods, with organizations only paying for what they use. Competition among cloud providers has further improved CC's performance, making hosting DW in the cloud a viable option.

Migrating DW to the cloud is popular due to its flexible architecture, enabling data access from various sources, whether cloud-based or remote servers accessed via the internet [10]. This shift reduces capital expenditures, favoring lower operational expenses, driving organizations to embrace the cloud [7].

However, security concerns arise due to the sensitive and confidential nature of DW data. Adoption of CC poses security challenges, including issues related to computing, network, and functional requirements that may not align seamlessly with DW needs in the cloud environment [11,12].

2.5. Selection of the Cloud Data Warehouse Provider

Selecting a Cloud Data Warehouse (CDW) provider is a critical decision. Companies must assess their specific needs to find the best-suited CDW. Testing with a small dataset across multiple CDWs can help evaluate performance and costs. When migrating workloads to the cloud, the choice of environments and services affects configurations and workload preparation. CDW providers should ideally be chosen in parallel with migration planning. CDW vendors have focused on efficient clustering, Machine Learning (ML), serverless resources, multi-cloud support, and simplified deployments to differentiate themselves [13].

3. Research Methodology

The principles, practices, and procedures established by the Design Science Research Methodology (DSRM) for Information Systems [14] were followed for the development of this article, in order to manage the research process represented in Figure 1.

The DSRM methodology comprises six key steps [14]:

1. Problem Identification and Motivation: Identifying the factors that lead to defining the problem.
2. Definition of Objectives: Setting objectives to solve the identified problem and gather essential knowledge for practical implementation.
3. Design and Development: Conducting a state-of-the-art review to better understand relevant concepts.
4. Demonstration: Validating the previously developed work.
5. Evaluation: Drawing conclusions based on the work conducted.
6. Communication: Presenting and disseminating the results.

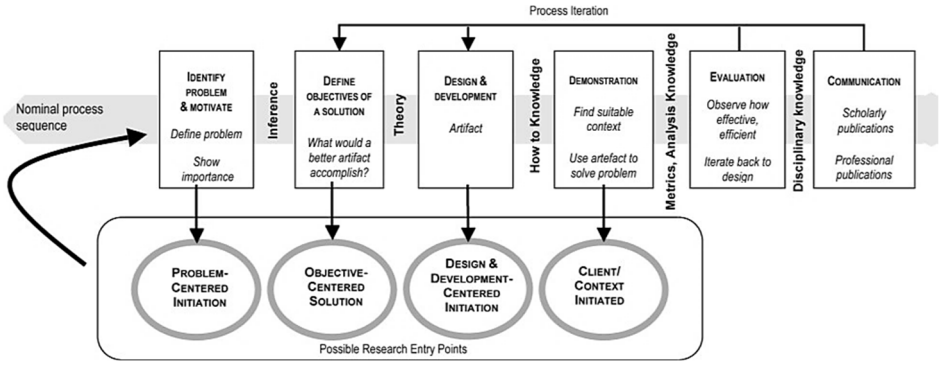

Fig. 1. Design Science Research Methodology for Information Systems, adapted by [14]

4. Cloud Data Warehouse Technologies

4.1. Snowflake

Snowflake is a fully managed Software as a Service (SaaS) platform introduced in 2012. It combines data warehousing, data lakes, data engineering, data science, and more in one platform. Users benefit from its cloud-agnostic nature, working seamlessly across AWS, GCP, and Azure, and it supports ANSI SQL. Snowflake offers features like storage and compute separation, real-time computing, data sharing, cloning, and third-party tool compatibility. It simplifies data transformation and modelling for data engineers, empowering stakeholders in critical decision-making processes through reports and dashboards [15].

4.2. Google Big Query

BigQuery is a Platform as a Service (PaaS) introduced in 2010 within the Google Cloud Platform. It serves as a fully managed enterprise CDW with an integrated query engine, enabling users to effortlessly manage and analyze data, incorporating features like machine learning, geospatial analysis, and business intelligence. BigQuery's serverless architecture eliminates the need for users to manage infrastructure and allows data analysis via SQL queries. It efficiently handles data volumes in the order of billions of rows and supports both Google Standard SQL and Legacy SQL. BigQuery's scalable and distributed analytics engine enables rapid querying of terabytes of data in seconds and petabytes in minutes. It maximizes flexibility by separating the compute engine that analyses data from the storage options. Additionally, it facilitates data ingestion from external sources and supports continuous data updates through streaming [16].

4.3. Amazon Redshift

In 2013, AWS disrupted the DW industry with the introduction of Amazon Redshift, a pioneering petabyte-scale CDW under the PaaS model. Amazon Redshift revolutionized data analysis by enabling cost-effective examination of large datasets using conventional Business Intelligence (BI) tools. This innovation marked a departure from expensive, inflexible, and expertise-intensive on-premises data storage solutions. Powered by a database management and query processing system based on PostgreSQL, Amazon Redshift seamlessly integrates with most existing SQL applications with minimal adjustments. As a native AWS service, it seamlessly collaborates with other AWS technologies, positioning itself as a central hub for connecting various services [15].

4.4. Azure Synapse

Azure Synapse Analytics, previously Azure SQL Data Warehouse, is a Microsoft PaaS solution tailored for data integration, DW, and big data analytics. Launched in late 2020, it serves as a unified platform for organizations to gather and consolidate public, operational, and historical data. Utilizing SQL, specifically Transact-SQL (T-SQL), it empowers customers to adjust their analytics infrastructure to match their processing requirements. Azure Synapse offers pricing options based on dedicated or serverless usage, with the ability to pause and resume compute charging, ensuring cost-effectiveness. Its versatility and ease of administration make it adaptable for a wide range of usage patterns. Beyond analytics, Azure Synapse serves as a central hub for connecting various Azure services, including Apache Spark for streaming, Artificial Intelligence (AI), Machine Learning (ML), SQL, and BI workloads [17].

5. Cloud Data Warehouse Environments

5.1. Dataset: Star Schema Benchmark

Initially, the project employed a dataset derived from the well-known industry standard TPC Benchmark™ H (TPC-H). TPC-H is a decision support reference that encompasses business-oriented ad-hoc queries and concurrent data modifications. Criticisms emerged regarding TPC-H's adherence to Ralph Kimball's model, which advocates for the use of a "star schema" in decision support systems. In response, a modified version called the "Star Schema Benchmark" (SSB) was introduced to assess star schema optimization and address TPC-H's shortcomings. SSB is a simplified benchmark featuring four query sets, four dimensions, and a fact table. Figure 2(a) illustrates the relational model of the TPC-H benchmark, while Figure 2(b) depicts the SSB's relational model with necessary modifications, including a LINEORDER fact table and four dimension tables: CUSTOMER, PART, SUPPLIER, and DATE, each with a specified number of tuples below the table name [17].

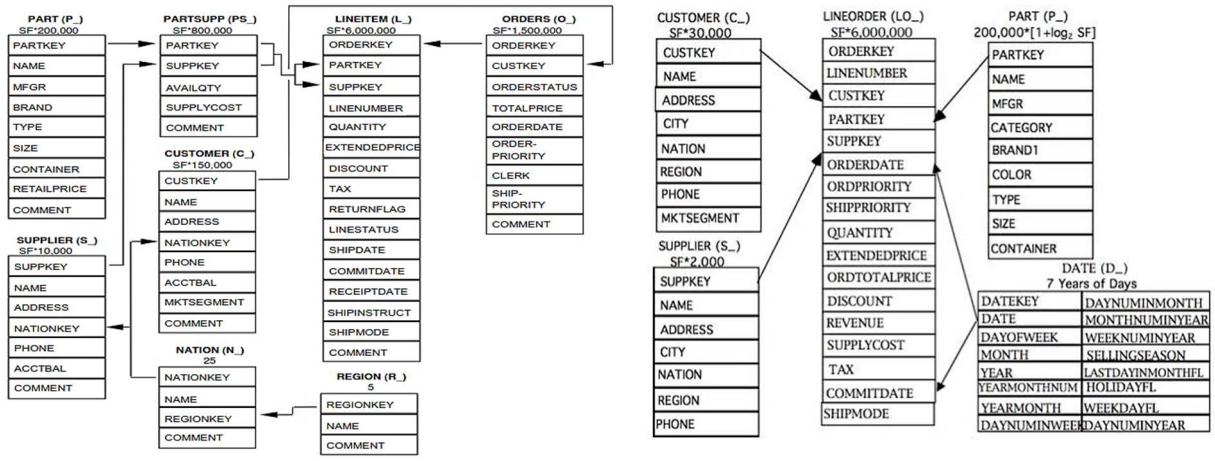

Fig. 2. (a) TPC-H model (b) SSB model

5.2. Cloud Data Warehouse Configuration

In order to keep the configurations of the applied technologies as close as possible to run accurate benchmarks, the server technologies configuration were defined with two cores each. The only exception is Google BigQuery, which uses a serverless technology and therefore there is no such configuration possible. Table 1 describes the settings selected for each of the technologies used in the present study.

Table 1. Cloud Data Warehouse Configuration by Technology

Technology	Configuration
Snowflake	XS (2 cores)
Google BigQuery	Using the serverless DWaaS (Data Warehouse as a Service)
Amazon Redshift	dc2.large (2 cores)
Azure Synapse	DW1000c (2 computer nodes)

5.3. Construction of the Cloud Data Warehouse Environments

The project began by creating a database and a Storage Area (SA) schema, acting as a temporary data extraction zone between data sources and the Data Warehouse (DW). Data from the TPC-H dataset was used to populate the SA schema, and all required SSB tables were created.

Next, the Support Application Data Lake (SADL) schema was established across all technologies. Dimension and fact entities, including DIM_CUSTOMER, DIM_SUPPLIER, DIM_PART, DIM_DATE, and FACT_LINEORDER, were created. The main objective was to load data into these tables, following the modifications outlined in [18].

To achieve this, distinct procedures were developed for each table, while the core code within these procedures was similar, differences in execution arose due to varying platform capabilities, with some functions available on one platform but not on others, necessitating diverse implementations.

With the SADL schema tables filled, the environment is ready to perform the tests.

The main objective is to assess CDW technology performance using two key metrics:

1. Evaluating data loading times for SADL schema tables by measuring stored procedure execution times.
2. Measuring the running times of predefined queries.

To ensure accuracy and maintain consistent benchmark conditions, caching will be disabled across all technologies, and the same WiFi connection will be used for all measurements. Each test, whether for tables or queries, will be repeated five times, and the resulting averages will be calculated. The set of queries used is defined in the SSB document [18].

6. Results and Discussion

In the benchmark, each query ran five times, and the average of these runs was used for accuracy. Caching was disabled across all technologies to prevent prior runs from affecting results.

Figure 3 displays the average data loading times (in seconds) for SADL tables. Google BigQuery generally exhibited slower performance, likely due to its serverless nature, while other technologies used 2-core systems. However, for the smallest table, DIM_SUPPLIER, BigQuery loaded data slightly faster than Azure Synapse, which consistently ranked third among the 2-core systems.

The benchmark's conclusion is that Snowflake and Amazon Redshift are the most competitive technologies in this environment. Snowflake performs better when loading data into tables with smaller volumes, such as DIM_SUPPLIER and DIM_DATE, while Amazon Redshift excels with larger data volumes.

Analyzing average query running times in Figure 4, Azure Synapse generally performs slower, except for Queries 4.1 and 4.2, which are complex and affect performance. Google BigQuery shows slower performance in most queries of set 4, but it outperforms Synapse, except for Queries 4.1 and 4.2. In queries 1.1 and 2.2, BigQuery consistently ranks third but performs better than Synapse.

In summary, Snowflake and Redshift consistently perform well in this benchmark and are recommended choices for users considering the metrics used in selecting a CDW technology.

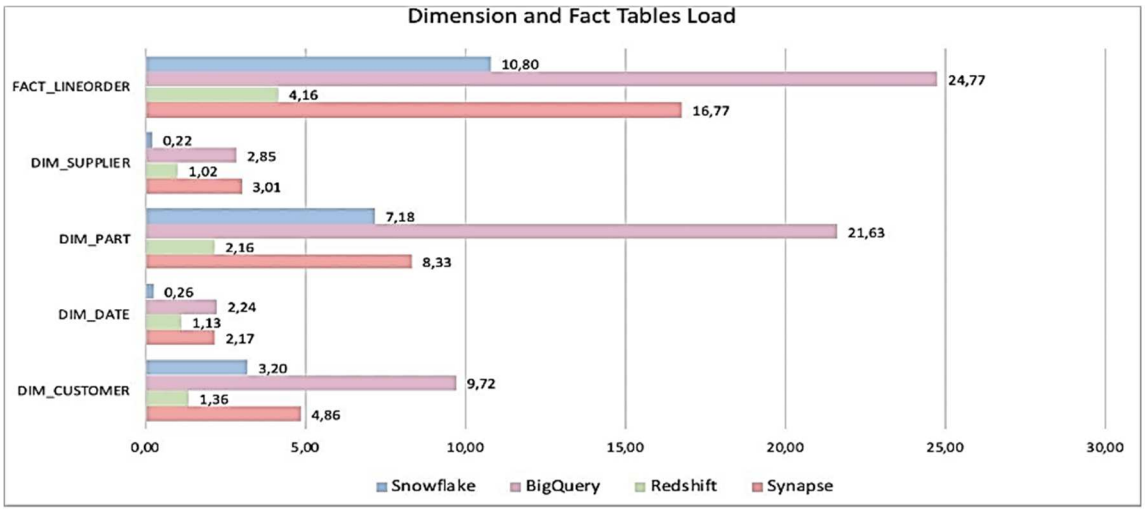

Fig. 3. Average of the data load times to the SADL tables

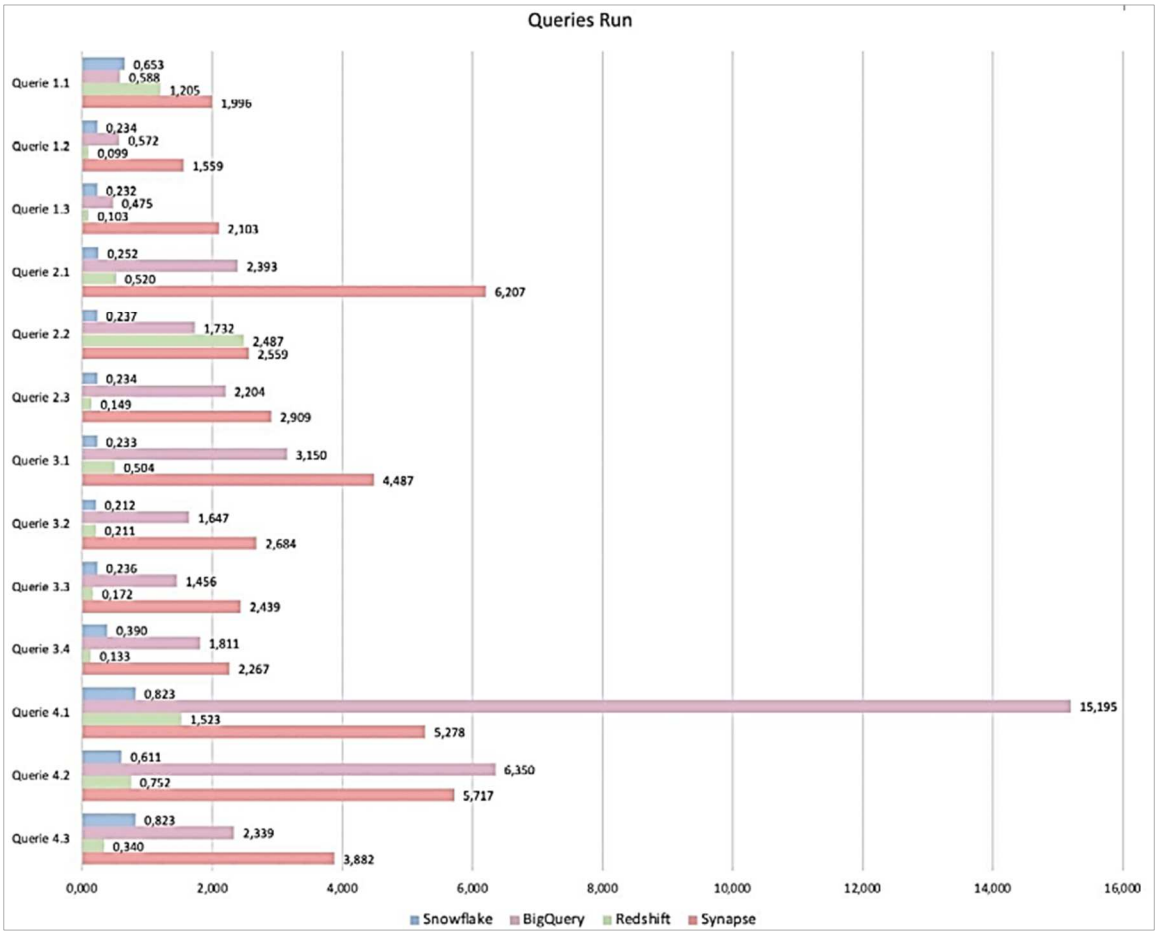

Fig. 4. Average of the queries running times

7. Conclusions, Limitation, and Future Works

In conclusion, this article has provided valuable insights into the use of Cloud Data Warehousing (CDW) technologies and has addressed the initial problem by exploring various performance metrics. The following achievements were made:

- A literature review on CDW and its core concepts was presented.
- Several CDW technologies with significant market presence were introduced.
- CDW environments were constructed using four different technologies, employing a predefined dataset.
- Benchmark results for these four technologies were shared, utilizing a set of metrics to assess their performance.

However, it's important to acknowledge certain limitations of this research, particularly related to the uniform configuration used for implementing the CDW across the four selected technologies. Exploring different configurations, such as varying memory and CPU cores, could provide additional insights.

This article may serve as a foundation for future research in the CDW domain, including comparisons with on-premises platforms or more complex database models. Conducting benchmarks on larger CDWs could yield further interesting findings and results.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020

References

[1] Sunyaev, A., Cloud computing. Internet Computing, 195-236, 2020

[2] Gill, S. Tuli, S. Xu, M. Singh, I. Singh, V. Lindsay, D. and Garraghan, P. Transformative effects of IOT, blockchain and Artificial Intelligence on Cloud Computing: Evolution, vision, trends and open challenges. Internet of Things, 8, 100118, 2019

[3] Nieuwenhuis, J. Ehrenhard, L. and Prause, L. The shift to cloud computing: The impact of disruptive technology on the enterprise software business ecosystem, Technological Forecasting and Social Change, 129, 308-313, 2018

[4] Alkhalil, A. Sahandi, R. and John, D, An exploration of the determinants for decision to migrate existing resources to cloud computing using an integrated toe-DOI model. Journal of Cloud Computing, 6(1), 2017

[5] Singhal, U., Competitions, quizzes, hackathons, scholarships and internships for students and corporates.Dare2Compete. Available: https://dare2compete.com/blog/characteristics-of-data-warehouse, 2021

[6] Raslan, A. and Calazans, T., Data warehouse: conceitos E Aplicações. *Universitas: Gestão E TI*, 4(1)

[7] Virant, R. Kamišalić, A. and Šestak, M. A comparison of traditional and modern data warehouse architectures.

Available: https://dk.um.si/IzpisGradiva.php?id=80484, 2021

[8] IBM Cloud Education. IaaS vs. PaaS v. SaaS. Available: https://www.ibm.com/cloud/learn/iaas-paas-saas, 2021

[9] Rehman, U. Ahmad, U. and Mahmood, S., A comparative analysis of traditional and Cloud Data Warehouse. Available: https://www.researchgate.net/profile/Sajid-Mahmood-11/publication/329704023_A_Comparative_Analysis_of_Traditional_and_Cloud_Data_Warehouse/links/5c3c6 38c92851c22a3736ebb/A-Comparative-Analysis-of-Traditional-and-Cloud-Data-Warehouse.pdf, 2018

[10] Thompson, J. and Van der Walt, S. Business intelligence in the cloud, SA Journal of Information Management, 12(1), 2010

[11] Guermazi, E. Ayed, B. and Ben-Abdallah, H. Adaptive Security for Cloud Data Warehouse as a Service. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7166672, 2017.

[12] Awoyelu, I. Omodunbi, T. and Udo, J. Bridging the gap in modern computing infrastructures: Issues and challenges of data warehousing and cloud computing. Computer and Information Science, 7(1), 2013.

[13] Wurm, H., Data Warehouse Technology Value Matrix 2022-oracle.com. Available: https://www.oracle.com/a/ocom/docs/database/nucleus-research-data-warehouse-technology-matrix-report.pdf, 2022

[14] Peffers, K. Tuunanen, T. Rothenberger, A. and Chatterjee, S. A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45-77. 2007

[15] Kline, L. Redshift vs Snowflake: The definitive guide. Available: https://hightouch.io/blog/redshift-vs-snowflake-the- definitive-guide/, 2022

[16] Kline, L. Bigquery vs snowflake: The definitive guide. Available: https://hightouch.io/blog/big-query-vs-snowflake- the-definitive-guide/, 2021

[17] Kline, L. Azure Synapse vs snowflake: The definitive guide. Available: https://hightouch.io/blog/azure-synapse-vs-snowflake-the-definitive-guide/, 2021

[18] O'Neil, P., O'Neil, B. and Chen, X., Star schema benchmark. Available: https://www.cs.umb.edu/~poneil/StarSchema PDF, 2009