Cloud Based Data Warehousing: Challenges and Criteria

Farah Yasser
Department of Business Informatics
German University in Cairo
Cairo, Egypt
fyseoudy@gmail.com

Ayman Alserafi
Department of Business Informatics
German University in Cairo
Cairo, Egypt
ayman.alserafi@guc.edu.eg

***Abstract*— In this digital age, organizations are doing their best to leverage their data for making valuable insights to gain competitive advantage. Cloud Computing has transformed the way organizations store, analyze, and manage their data as it removes the burden of building and maintaining the hardware and software infrastructure needed to manage the data. Moreover, the pay-as-you-go nature of the cloud makes it cost effective as organization pay only for the resources utilized; also, organizations benefit from the vast scalability of the cloud that enables the utilization of more resources at peak workloads. As a result, this research focuses on investigating the Cloud Based Data Warehouse solutions capabilities and challenges. We experiment using a prototypical implementation in the Google BigQuery platform. We propose criteria to evaluate cloud-based Data Warehouses consisting of the ETL process, Data Querying, Data Visualization, Security and Encryption, Access control, Audit logging, Backups, and Disaster Recovery. This research enables the exploration and testing of Cloud Data Warehouse solutions to explain how it can be adopted by organizations.**

Keywords—Business Intelligence (BI), Cloud Computing, Cloud Based Data Warehousing, Data Warehouse, Data Warehouse as a Service (DWaaS).

I. Introduction

Business Intelligence (BI) can be identified as the process of allowing users to enhance their decision making by transforming raw data into meaningful information using technology [1]. Data warehousing is an integral part of BI as it enables organizations to have a single point of access for all their data. The implementation of Data Warehouse (DW) can be either implemented on premises or as a service over the Internet using cloud computing. The purpose of this paper is to examine how Cloud Computing can be used to host a DW. The deployment of a Cloud Data Warehouse (CDW) will enable businesses to take advantage of the cloud features and benefits; however, it is still a very vague field. As a result, this paper continues from where the research is currently standing by proposing criteria to assess and select CDWs.

II. Previous Research

A. Data Warehousing

A DW is a central repository that stores data from different sources [2]. The stored data in the warehouse is integrated, non-volatile, subject oriented, and time variant [3]. It allows organizations to integrate different types of historical data gathered from various systems across the organization for the purpose of having a single point of access to the information needed for decision making. As a result, quick and informative decisions can be made [4]. The DW is used for the purpose of giving decision makers a multidimensional view of data [5].

Due to the heterogeneity of data sources, data integration is a challenge in data warehousing; however, this issue is addressed by the Extract-Transform-Load (ETL) process. ETL is a data integration process [4] that is responsible for the extraction, cleaning, conforming, and loading of data into the DW [6]. First, Extraction is responsible for gathering data from various external and internal sources to the organization. Second, Transformation is responsible for conversion and adaptation of data to the target format [2]; it also applies transformation rules to data such as filter, select, derive, and convert [3]. Third, Loading is responsible for updating the data that was collected and transformed into the DW [2]. For BI to support decision making, it retrieves data from the DW and this data can be used by data mining tools for making predictions and insights or it can be presented using dashboards and other reporting tools [6]. Dimensional modelling is used in Data Warehousing which offers a highly effective and intuitive retrieval, aggregation, and analyses of historical data [6]. Furthermore, it provides high-level query access [4].

B. Cloud Computing

Cloud Computing refers to the use of software and hardware components for delivering a service over the Internet; to further explain the concept, it allows for the storage and access of data as well as programs over the Internet rather than the computers' hard drive [7]. It has four deployment models which are Private Cloud that is created for the exclusive use of an organization [8], Public cloud that is available for open use by the public [8], Hybrid Cloud that is a combination of both the public cloud and the private cloud [7], and Community cloud that is created to be used exclusively by organizations with similar concerns who together form a community of consumers [8].

Cloud Computing is known for delivering highly scalable computing resources as a service to several customers over the Internet [9]. The services provided are called XaaS, where X refers to the service provided by the cloud environment, such as software, hardware, infrastructure, data, or any other service [9]. Infrastructure as a Service (IaaS) is the cloud service where computing infrastructure is provisioned to consumers as a service based on their demand and duration of their need; resources such as servers, storage, network equipment, CPU, and data center facilities are outsourced to the service users [10]. Platform as a Service (PaaS) is the cloud service that offers to developers application deployment and development platform through the Internet in order for them to build, deploy, and improve their applications smoothly [11]. Software as a Service (SaaS) is the cloud service where users are enabled to use applications without the need to locally install and run them on their devices [10]. Database as a Service (DBaaS) is the cloud computing service where users are given access to a database over the Internet without deploying the physical hardware infrastructure as well as installing software configurations [12].

C. Cloud Based Data Warehousing

CDW is referred to as Data Warehouse as a Service (DWaaS); it enables the users to generate, maintain, and query their data over the Internet [13]. CDW enables organizations to access latest technologies in a cost-effective manner because it removes the burden of building and setting up the hardware and software components needed for the DW which requires a costly upfront investment [14]. Furthermore, the benefits of CDWs include cost reduction and scalability which is a result of the pay-as-you-go nature of cloud computing; it allows users to utilize more resources without having to expand their own hardware and software components. The service is also available to different sized enterprises with different levels, it enables better collaboration between users as well as ubiquitous accessibility [15].

After analysing the available literature for CDWs, it became known that research in this field is lacking particularly for the deployment of the DW in the cloud; furthermore, there are no remarkable development in the field [16]. So, it raised the question: **what are the criteria for selecting a CDW solution and how to implement it in practice?**

III. Research Methodology

The purpose of this paper is to investigate how can a DW be implemented in the cloud. The methodology used in this research is design science which is used to test certain criteria for the implementation of the CDW solution to guarantee comparable performance to that of the on-premises solution. Furthermore, a prototypical CDW solution was chosen and tested; then the results were shown to an expert and an interview was conducted to compare the results to the on-premises solution capabilities.

A. Criteria for Testing a Cloud Data Warehouse

The initial criteria that are tested to evaluate the CDW platform is shown in Table I and it includes the ETL process, Data Querying, Data Visualization, Security and Encryption, Access control, Audit logging, Backups, and Disaster recovery.

TABLE I. Criteria for testing Cloud DataWarehouse solutions

Criteria	Description
ETL	The process of data integration that is responsible for the extraction, cleaning, conforming, and loading of data into the DW [6].
Data Querying	The process of requesting data from the DW [17].
Data Visualization	The process of providing an interactive visual representation of data [18].
Security and Encryption	The process of protecting the data against unauthorized access [19].
Access Control	The process of determining who can access the data and the resources of the DW [20].
Audit Logging	The process of documenting the activity of DW [21].
Backups and Disaster Recovery	The process of creating copies of the DW data and using these copies in case of data loss or corruption to ensure continuity of business operations [22].

B. Prototypical Design and Devlopment

- A CDW solution is chosen which is Google BigQuery [1] - a cost-effective and serverless (not requiring server provisioning for querying data) DW that has machine learning and BI tools that enable it to scale according to the user data and work across different clouds [23].
- A dataset was used to test the solution which is AdventureWorksDW that is created by Microsoft to test the performance of DW solutions [24]. This data set is uploaded to the CDW platform as explained in the ETL process in the results section below. Furthermore, SQL queries were used to analyze the performance metrics, such as the execution time, of the CDW platform.
- The criteria in Table I are tested for the CDW solution.
- To evaluate the solution an interview was conducted with Shaimaa El Gammal who is a Data Engineering Manager at a large Egyptian bank called CIB, which uses an on-premises DW solution that we compare its activities to that of the cloud-solution used in our prototype. The interviewee was shown the solution and how each of the criteria was tested to provide expert feedback.

IV. Results

This Section presents the results of testing the criteria using Google BigQuery.

A. ETL

- AdventureWorksDW dataset was uploaded in Microsoft SQL Server to extract the needed data. Then these tables were uploaded into the BigQuery project by creating a dataset then creating a table in the dataset and choosing to upload the table as a CSV file as shown in Figure 1.
- BigQuery platform provides the creation of a cloud storage bucket in Google Cloud Platform- which are containers where data is stored in the cloud [25]- and uploading AdventureWorksDW directly to it to be extracted in the project; however, this was not tested due to license limitations; furthermore, it enables the creation of tables from Google Big Table, Amazon S3, and Azure Blob Storage.
- Transformation can be done using SQL in the BigQuery project. Moreover, it can be done using the BigQuery API with Python on Google Colab as shown in Figure 2 or any other platform then loading the data back into BigQuery as shown in Figure 3.

[1] https://cloud.google.com/bigquery

Figure 1: Creating Table in BigQuery

```
!pip install google-cloud-bigquery

from google.colab import auth
auth.authenticate_user()

from google.cloud import bigquery
project_id = 'farah-thesis'
client = bigquery.Client(project=project_id)
```

Figure 2: Connecting BigQuery API

```
[ ] from pandas.io import gbq

[ ] gbq.to_gbq(dim_date, 'farah-thesis.AdventureWorks.DimDate', project_id=project_id, if_exists='replace')

100%|██████████| 1/1 [00:00<00:00, 8630.26it/s]
```

Figure 3: Loading data into BigQuery

B. *Data Querying*

- SQL queries can be run in the BigQuery SQL Workspace as shown in Figure 4. The query retrieves the total sales revenue of each product sold in 2013 by calculating the revenue of each product, grouping the data by the product name, and ordering the data by the total revenue in descending order.
- Execution details are provided which include elapsed time, slot time consumed, bytes shuffled, and the execution graph.

C. *Data Visualization*

- BigQuery provides the exploration of queries' data with Looker Studio to create visualized interactive reports.
- Data could be visualized and explored using Python on Google Colab notebooks.

D. *Security and Encryption*

- Google BigQuery automatically encrypts the data using Google managed encryption keys.
- Customer managed encryption keys can be used where organizations can use their own encryption keys to protect their data, but this was not tested due to license limitations.
- Google BigQuery has several data protection and governance features which include: securing resources with Identity and Access Management (IAM), securing data with classification as column level security and row level security, data discovery which consists of cloud data loss prevention and data catalog, encryption, monitoring, auditing, and logging.

E. *Access Control*

- IAM features offered by BigQuery manage the permissions given to the users for the BigQuery resources and data sets as well as authorizing the users' identities.
- The permissions include different roles and under each role there are different permissions that can be provided.

F. *Audit Logging*

- BigQuery has a logging tab as shown in Figure 5 which contains in Part A the logging operations provided by the platform with Logs Explorer, Logs Dashboard, Log-based Metrics, Log Router, Logs Storage, and Log analytics. It contains in part B details about the query workload results, which are severity levels, resource types, and execution time ranges.
- The audit logs show activities that occur in the DW, the resource accesses, and the data modifications.
- Another feature that is offered is the project history which contains data about each job created in the project along with ID, creation time, owner, action, and summary.

G. *Backups and Disaster Recovery*

- BigQuery offers Disaster planning features like recovering data from backups; however, they could not be tested due to license limitations.

After testing the criteria in the Google BigQuery environment, an expert interview was conducted to explain how each criterion was addressed in the on-premises solution and how to improve the proposed criteria. The BigQuery results are compared against the on-premises DW used in the bank.

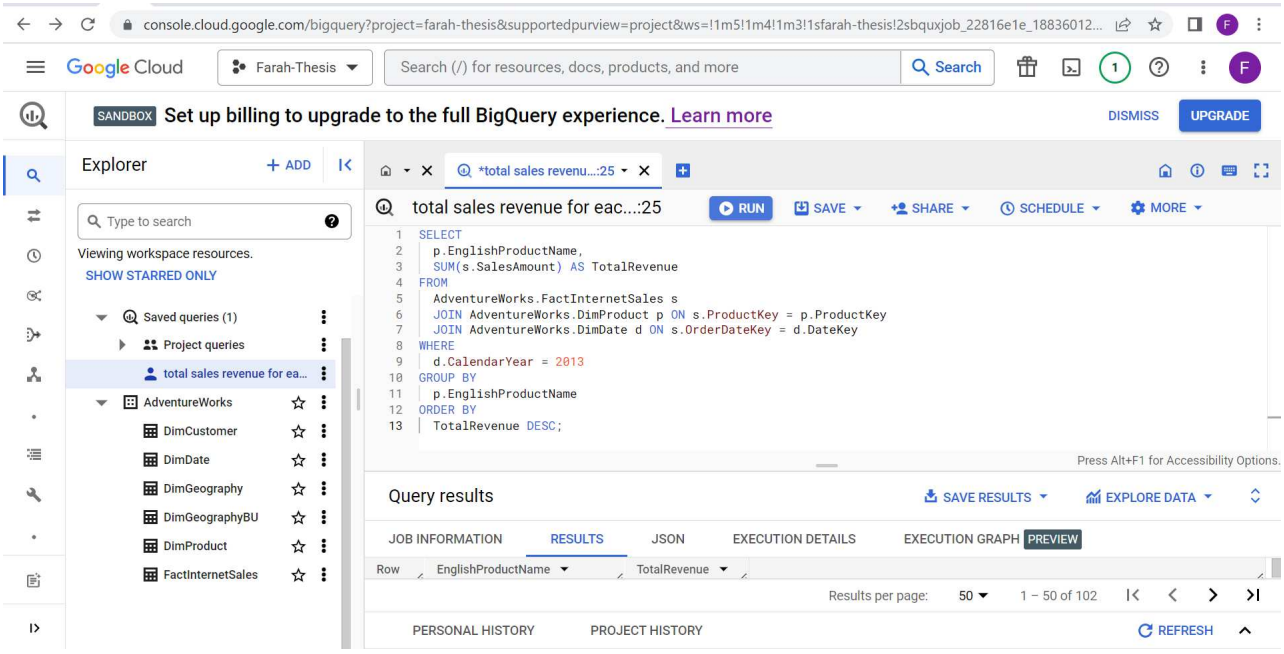

Figure 4: SQL engine

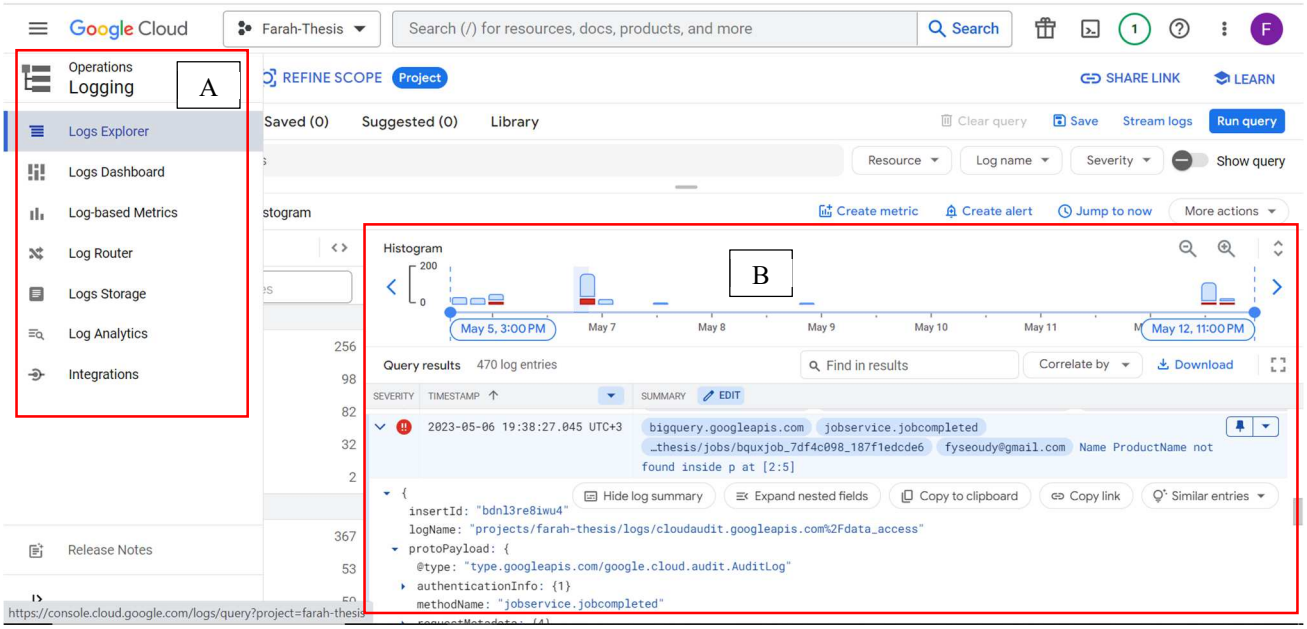

Figure 5: Logging in BigQuery

Interview Findings:

- Extract, Load, Transform (ELT) is mainly used in the on-premises DW solution instead of ETL where data is extracted first then loaded into staging table then transformation is applied on the data. The interviewee explained that what was performed in testing Google BigQuery is also ELT as data was loaded first then transformed.
- The creation of indices in the on-premises DW solution enables better query performance. This was not tested in the CDW prototype.
- It was stated that there is a system in the on-premises solution for workload management and capacity planning that was not provided in the CDW.
- The interviewee agreed on the proposed criteria; however, she suggested to add the ELT process instead of ETL which is covered in BigQuery and is tested under ETL. She also suggested including criteria about the creation of indices, performing in depth data analyses, and checking if a comprehensive management system can be built in the CDW, which will be investigated in future research.

V. CONCLUSION

A DW is a repository that stores integrated, non-volatile, subject oriented, and time variant data and it is a single point of access for all data inside the organization. Furthermore, it can be deployed on premises or in the cloud. Cloud Computing refers to the provisioning of services over the Internet using hardware and software components; it can provide anything as a service using the XaaS model, for instance, Software as a Service, Infrastructure as a Service, Platform as a Service, and DW or Data Base as a Service.

The CDW gives the users the ability to generate, maintain, and query their data over the Internet. However, it is an area with limited research, so it raised the question of how can a DW be deployed as a Service in the Cloud. As a result, this paper proposed criteria to test CDW solutions which are the ETL process, Data Querying, Data Visualization, Security and Encryption, Access control, Audit logging, Backups, and Disaster Recovery.

As was demonstrated in our prototype, our proposed criteria can be systematically used to assess DWaaS cloud-computing services. We implemented an example using Google BigQuery. Furthermore, an expert interview was conducted to validate the criteria as well as compare them with the equivalent activities conducted in the on-premises solution. The limitations of this paper include basic license restrictions where features could be only tested in premium subscriptions, including: the creation of a cloud storage bucket to upload the data, the creation of customer managed encryption keys, disaster recovery, and indexing. For future research, it is recommended to acquire a premium license to test the other features of the system. Moreover, this paper can be expanded by exploring cloud based data analytics or machine learning.

REFERENCES

[1] M. R. Llave, "Business Intelligence and Analytics in Small and Medium-sized Enterprises: A Systematic Literature Review," *Procedia Computer Science*, vol. 121, pp. 194–205, 2017, doi: 10.1016/j.procs.2017.11.027.

[2] A. Martins, P. Martins, F. Caldeira, and F. Sá, "An Evaluation of How Big-Data and Data Warehouses Improve Business Intelligence Decision Making," *Trends and Innovations in Information Systems and Technologies*, pp. 609–619, 2020, doi: 10.1007/978-3-030-45688-7_61.

[3] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management*, vol. 36, no. 5, pp. 700–710, Oct. 2016, doi: 10.1016/j.ijinfomgt.2016.04.013.

[4] L. W. Santoso and Yulia, "Data Warehouse with Big Data Technology for Higher Education," *Procedia Computer Science*, vol. 124, pp. 93–99, 2017, doi: 10.1016/j.procs.2017.12.134.

[5] A. Dhaouadi, K. Bousselmi, M. M. Gammoudi, S. Monnet, and S. Hammoudi, "Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons," *Data*, vol. 7, no. 8, p. 113, Aug. 2022, doi: 10.3390/data7080113.

[6] A. L. Antunes, E. Cardoso, and J. Barateiro, "Incorporation of Ontologies in Data Warehouse/Business Intelligence Systems - A Systematic Literature Review," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100131, Nov. 2022, doi: 10.1016/j.jjimei.2022.100131.

[7] A. Rashid and A. Chaturvedi, "Cloud Computing Characteristics and Services A Brief Review," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 2, pp. 421–426, Feb. 2019, doi: 10.26438/ijcse/v7i2.421426.

[8] P. Mell and T. Grance, "The NIST definition of cloud computing". *National Institute of Standards and Technology Special Publication*, Vol 53, pp. 1-7, 2011.

[9] M. N. Birje, P. S. Challagidad, R. H. Goudar, and M. T. Tapale, "Cloud computing review: concepts, technology, challenges and security," *International Journal of Cloud Computing*, vol. 6, no. 1, p. 32, 2017, doi: 10.1504/ijcc.2017.083905.

[10] S. Murugesan and I. Bojanova, "Cloud Computing," *Encyclopedia of Cloud Computing*, pp. 1–14, May 2016, doi: 10.1002/9781118821930.ch1.

[11] M. I. Malik, S. H. Wani, and A. Rashid, "CLOUD COMPUTING-TECHNOLOGIES," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 2, pp. 379–384, Apr. 2018, doi: 10.26483/ijarcs.v9i2.5760.

[12] C. Miyachi, "What is 'Cloud'? It is time to update the NIST definition?," *IEEE Cloud Computing*, vol. 5, no. 3, pp. 6–11, May 2018, doi: 10.1109/mcc.2018.032591611.

[13] K. Karkouda, A. Nabli, and F. Gargouri, "TrustedDW: A New Framework to Securely Hosting Data Warehouse in the Cloud," *in CATA*, pp. 397-406, March 2019.

[14] D. Baum and J. Kraynak, "Cloud Data Warehousing," John Wiley & Sons, 2020.

[15] H. Verma "Data-warehousing on cloud computing", *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 2, no. 2, pp. 411-416, 2013.

[16] M. S. Ali, S. Khan, and S. J. Miah, "Understanding towards Interactions between Business Intelligence and SMEs: Learn from Each Other," *Journal of Information Systems and Technology Management*, vol. 14, no. 2, pp. 151–168, Aug. 2017, doi: 10.4301/s1807-17752017000200002.

[17] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Communications of the ACM*, vol. 54, no. 8, pp. 88–98, Aug. 2011, doi: 10.1145/1978542.1978562.

[18] M. Sadiku, A. E. Shadare, S. M. Musa, C. M. Akujuobi, and R. Perry, "Data visualization," *International Journal of Engineering Research And Advanced Technology (IJERAT)*, vol. 2, no. 12, pp. 11-16, 2016.

[19] Y. Sun, J. Zhang, Y. Xiong, and G. Zhu, "Data Security and Privacy in Cloud Computing," *International Journal of Distributed Sensor Networks*, vol. 10, no. 7, p. 190903, Jul. 2014, doi: 10.1155/2014/190903.

[20] Microsoft, "What is access control? Microsoft Security," Retrieved June 3, 2023, from https://www.microsoft.com/en-us/security/business/security-101/whatis-access-control, 2023.

[21] Google Cloud, "Cloud Audit Logs overview," Retrieved June 3, 2023, from https://cloud.google.com/logging/docs/audit, 2023.

[22] IBM, "Backup and disaster recovery," Retrieved June 3, 2023, from https://www.ibm.com/topics/backup-disaster-recovery, 2023.

[23] Google Cloud, "Cloud data warehouse to power your data-driven innovation," Retrieved December 26, 2022, from https://cloud.google.com/bigquery#section-2, 2022.

[24] Microsoft, "AdventureWorks Database Installation and Configuration," Microsoft SQL Server Samples, Retrieved July 20, 2023 from https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms, 2023.

[25] Google Cloud, "About Cloud Storage buckets," Retrieved July 12, 2023, from https://cloud.google.com/storage/docs/buckets, 2023.