




Selecting a Data Warehouse Provider: A Daunting Task

João Ferreira ^a, Nuno Lourenço ^b and João R. Campos ^c

CISUC/LASI, DEI, University of Coimbra, Coimbra, Portugal

Keywords: Data Warehouses, ETL Tools, Cost Comparison, Pricing Models.

Abstract: In the contemporary landscape of rapid data accumulation, organizations increasingly rely on data warehouses to process and store vast datasets efficiently. Although the most challenging task is appropriately designing a data warehouse, selecting a provider is far from the trivial task it should be. Each provider offers a distinct array of services, each with its pricing model, which requires significant effort to analyze and determine which configuration meets the specific needs of the organization. In this paper, we highlight the inherent challenges of making fair comparisons among data warehouse solutions, providing the context of a start-up in the space traffic management industry as a case study. We defined several critical attributes for corporate decision-making: cost, processing capabilities, and data storage capacity. We systematically compare four leading technologies: Google BigQuery, AWS Redshift, Azure Synapse, and Snowflake. Our methodology employs a set of metrics designed to assess warehouse solutions, encompassing storage pricing, processing capabilities, scalability, and the integration of ETL tools. The process and the results highlight the challenges of this evaluation. It underscores the need for a standard approach to characterize the provided service specifications and pricing to allow for a fair and systematic assessment and comparison of alternative solutions.

1 INTRODUCTION

The exponential escalation in the volume of data generated globally reflects an era marked by the continuous increase in data storage capacity, with the majority being hosted in cloud environments. Technological advances and the increasing digitization of information in all sectors of society drive this growth (Aftab and Siddiqui, 2018). As a result, there has been a significant increase in the use of Data Warehouse (DW) solutions, which are essential to handling and extracting knowledge and value from these data (Berisha et al., 2022).


Data warehouses extend beyond the traditional concept of centralized data repositories, functioning as dynamic and sophisticated structures. Designed to store and organize large volumes of data from various sources, such as transactional systems, relational databases, or heterogeneous sources, DWs are fundamental in various areas, including business, health, science, and technology (Serra, 2024).


The explosion in the number of companies that build data architectures has been remarkable in recent


years, and this trend is expected to continue. Daily data generation is estimated at around 44 zettabytes (Berisha et al., 2022), driven by sources like social media, IoT devices, and third-party software (Kim et al., 2017). From 2018 to 2021, the data volume nearly doubled to approximately 84 zettabytes, with projections suggesting that it could reach 149 zettabytes by 2024 (Ahlawat et al., 2023).

Companies can enhance revenue by analyzing data to detect trends and make forecasts (Serra, 2024). However, designing a DW that effectively supports business processes and analytics remains challenging (Santos and Costa, 2022). Selecting a suitable provider requires evaluating processing capacity, hardware, scalability, and cost. This process is often complex and time-consuming due to heterogeneous architectures and pricing models, which hinder objective comparisons (Soma, 2022). The lack of standardization further complicates the task, especially for startups with limited resources and time.

This paper adopts the perspective of a startup selecting its initial cloud-based DW provider, where the investment decision must balance current needs with future scalability and operational costs. Rather than offering a comprehensive assessment of platform functionalities, the focus is on evaluating costs and

^a  <https://orcid.org/0000-0003-4242-7961>

^b  <https://orcid.org/0000-0002-2154-0642>

^c  <https://orcid.org/0000-0002-4623-764X>

configuration aspects that impact the initial setup and expected evolution of the data infrastructure.

Recent studies have expanded knowledge of data warehouses and cloud services, highlighting the value of comparative analyses. Uddin (Uddin and Hos-san, 2024) examined AI integration in DWs for big data optimization, summarizing topics from 25 papers. Villamizar (Villamizar et al., 2017) analyzed the economics of microservices and AWS Lambda, noting cost-efficiency despite management complexity. AlJamal (AlJamal et al., 2019) and Biplob (Badi-uzzaman Biplob et al., 2018) explored IaaS and HPC, focusing on virtual machines and real-time ETL tools. Harby (Harby and Zulkernine, 2022) introduced the Data Lakehouse model to handle both structured and unstructured data. However, we did not find studies in the literature that address the specific challenges of selecting a suitable DW provider from a startup perspective.

In this paper, we systematically compare four DW solutions: Google Big Query (Google, 2025), AWS Redshift (Services, 2025), Azure Synapse (Microsoft, 2025), and Snowflake (Inc., 2025), highlighting the complexities of cost comparison. Our analysis revealed significant challenges stemming from non-standardized virtual machine hardware, diverse services, and varying pricing policies among providers. This lack of uniformity complicates objective evaluations of DW solutions, as performance and capacity differ significantly across similar categories. With the challenges highlighted in this work, we aim to contribute towards the standardization of DW solution providers.

2 DW PRICE MODEL

Data warehouse service providers typically adopt multifactor pricing models that are designed to reflect resource usage, scalability, and performance. Here are some of the most common pricing models adopted by these providers (sourced from the official DW websites):

- **Instance or Node-Based Pricing:** This model depends on the number and compute nodes or instances, suitable for platforms like Amazon Redshift and Azure Synapse Analytics, allowing users to choose configurations based on their workload needs.
- **Resource-Based Pricing:** Charges are based on actual resource usage such as CPU and memory. This model is often used by services that offer automatic scalability, like Google BigQuery, which

also considers the volume of data processed and storage.

- **Storage-Based Pricing:** Costs are calculated based on the amount of data stored in the data warehouse, typical for platforms that separate compute and storage resources, such as Snowflake and Google BigQuery.
- **Performance or Speed-Based Pricing:** Pricing is determined by the speed of query processing and the performance level selected, allowing customers to pay more for higher speeds and lower latency.
- **Subscription:** Involves a fixed monthly or annual fee, providing access to a predefined package of features and capabilities, with discounts often available for longer-term commitments.

Although these various pricing approaches might provide more flexible and granular billing because there are no standards or unified practices, comparing the pricing of multiple providers becomes time-consuming and difficult.

3 METHODOLOGY

This study emerged from the development of a DW solution for a company in the space traffic management sector. A systematic methodology was adopted to compare existing DW platforms, aligning the analysis with the company's requirements and the capabilities of major providers.

Requirements and Contextual Analysis: The initial evaluation focused on four leading providers with infrastructure in Europe, considering scalability, ease of deployment, and potential for future expansion. Key criteria included storage, processing, backup options, integration with ETL tools, and cost models, favoring on-demand pricing and fixed-rate plans for predictable budgeting. The European location was selected due to the company's base of operations, alongside regional advantages such as strong data protection laws, robust infrastructure, and proximity to end users.

Chosen Providers: Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, and Snowflake were selected for their market leadership, global reach, and advanced technical capabilities. All offer scalable solutions, integrated services (including ETL tools), and robust infrastructure in Europe, aligning with the company's geographic and operational needs.

Defining the Evaluation Process: We selected baseline hardware configurations aligned across

providers in terms of vCPU, memory, and storage. The analysis compared technical features and costs, including capacity-based and serverless pricing, long-term contract discounts (1–3 years), and regional data transfer costs. We also evaluated ETL-related expenses, such as pipeline orchestration and flow execution.

Data Adequacy Check: We then verified whether the collected data supported a fair and reliable comparison. If gaps were identified, a **Collect More Data** step was triggered to obtain further details from providers, ensuring the dataset was robust enough to support informed and accurate conclusions.

For **Data Analysis**, given the recognized disparity in the units of measurement, scales, and pricing models provided by the different providers, it was essential to carry out data adjustments and transformations. Among others, this process involved standardizing some units of measurement.

4 PLATFORMS

This section examines DW solutions from leading providers: Google BigQuery (Google, 2025), AWS Redshift (Services, 2025), Azure Synapse (Microsoft, 2025), and Snowflake (Inc., 2025), selected for their compatibility with the partner company’s requirements. We analyze pricing and hardware specifications to highlight each platform’s capabilities and cost structure. All prices were obtained from official provider websites during the first half of 2025.

4.1 Google BigQuery

Google BigQuery is a serverless data warehouse that uses standard SQL, easing adoption for users of traditional databases. It integrates with Google’s analytics and machine learning tools and operates on a scalable architecture. The service provides various computing nodes tailored to specific processing needs, as outlined in Table 1.

Table 1: BigQuery Data Warehouse Node Configurations.

Name	vCPU	GB per CPU
E2	up to 32	8
N2	up to 128	8
N2D	up to 224	8
C3	up to 176	2 or 4 or 8
C3D	up to 360	2 or 4 or 8

Table 2 shows the pricing structure of various BigQuery editions, quantified by the use of slots, which

represent fundamental units of computational power similar to virtual CPUs. Queries are broken down into tasks and processed in parallel across multiple slots, with the number of slots enhancing processing speed by enabling more simultaneous task execution.

Table 2: Pricing and Commitment Models.

On Demand - Access 2000 slots Simultaneous		
Standard Edition		
Model	Hourly cost	Details
Pay as you go	\$0.044 / slot hour	\$/per sec
Enterprise Edition		
Pay as you go	\$0.066 / slot hour	\$/per sec
1 yr commit	\$0.0528 / slot hour	\$/for 1 yr
3 yr commit	\$0.0396 / slot hour	\$/for 3 yrs
Enterprise Plus Edition		
Pay as you go	\$0.11 / slot hour	\$/per sec
1 yr commit	\$0.088 / slot hour	\$/for 1 yr
3 yr commit	\$0.066 / slot hour	\$/for 3 yrs

The lack of detailed machine configurations may create uncertainty for users needing to align performance with budget, which is essential for precise planning or specific technical demands. BigQuery’s one- and three-year commitments offer pricing incentives for stable workloads, enabling significant savings over pay-as-you-go rates.

4.2 Amazon AWS

Amazon Redshift (Services, 2025), part of AWS, offers RA3 instances with managed storage and DC2 Dense Compute instances, as shown in Table 3 for the Europe (Spain) region. RA3 enables independent scaling of compute and storage, while DC2 synchronizes both, favoring compute-intensive workloads with steady data volumes. Key attributes include vCPU, ECU, memory, I/O (GB/s), and hourly cost.

Table 3: Instance types and specifications for Redshift.

Type	vCPU	Mem	I/O	Price
RA3 with Redshift Managed Storage				
ra3.16xlarge	48	384 G	8.00	\$14.424
ra3.4xlarge	12	96 G	2.00	\$3.606
ra3.xlplus	4	32 G	0.65	\$1.202
ra3.large	2	16 G	0.36	\$0.601
Dense Compute DC2				
dc2.8xlarge	32	244 G	7.50	\$5.6
dc2.large	2	15 G	0.60	\$0.30

Table 4 provides an analysis of the price and commitment models for Amazon Redshift instances in a European region (Spain). It categorizes costs into various commitment terms: one-year commitments

with no initial payment, partial initial payment, and full initial payment. Each category describes not only the initial fees and monthly fees but also the effective hourly rates, the annual costs per terabyte, and the comparative savings compared to on-demand rates (due to space constraints, on-demand rates, and three-year payment values are omitted in this article but are indirectly represented in the savings column).

Table 4: RI Pricing and Commitment Models for Redshift.

Instance	RI upfront Fee	Monthly Fees	Effective Hourly Rate	Savings on demanded
No advance payment - One Year (EU - Spain)				
dc2.large	USD 0	USD 175.20	USD 0.240	20%
ra3.4xlarge	USD 0	USD 1,839.60	USD 2.520	30%
dc2.8xlarge	USD 0	USD 3,212.00	USD 4.400	21%
ra3.16xlarge	USD 0	USD 7,373.00	USD 10.100	30%
Partial advance - One Year				
dc2.large	USD 858	USD 73.00	USD 0.198	34%
ra3.4xlarge	USD 10,600	USD 883.30	USD 2.420	33%
dc2.8xlarge	USD 16,951	USD 1,416.20	USD 3.875	31%
ra3.16xlarge	USD 42,311	USD 3,525.90	USD 9.660	33%
Full upfront payment - One Year				
dc2.large	USD 1,682	USD 0.00	USD 0.192	36%
ra3.4xlarge	USD 20,849	USD 0.00	USD 2.380	34%
dc2.8xlarge	USD 33,218	USD 0.00	USD 3.792	32%
ra3.16xlarge	USD 83,395	USD 0.00	USD 9.520	34%

Table 5 provides a detailed overview of the costs associated with data transfers within Amazon Redshift, describing a pricing model that differentiates between inbound and outbound data movements.

Table 5: Data Transfer Costs for Amazon Redshift.

Category	Cost
Inbound	Free
Outbound - From Redshift to Internet	
First 10 TB per month	\$ 0.09 per GB
Next 40 TB per month	\$ 0.085 per GB
Next 100 TB per month	\$ 0.07 per GB
More than 150 TB per month	\$ 0.05 per GB
Outbound to Other Regions	\$ 0.02 per GB

Inbound transfers are free, encouraging users to upload data without incurring charges. In contrast, outbound transfers are tiered, with costs decreasing as data volumes start at USD 0.09 per GB for the first 10 TB and reducing to USD 0.05 per GB beyond 150 TB per month; additionally, a reduced rate of USD 0.02 per GB for outbound transfers to other AWS regions.

4.3 Snowflake

Snowflake is a cloud-based DW that decouples compute from storage (Inc., 2025). Table 6 presents specifications for various instance types, from basic (e.g., 'CPU - XS') to advanced (e.g., 'GPU - L'), detailing vCPUs, memory (GiB), storage, GPU type and memory, and resource usage limits where applicable.

Table 6: Snowflake Node Specifications.

Instance	vCPU	Memory	GPU Type	GPU Memory
CPU-XS	2	8 GiB	n/a	n/a
CPU-S	4	16 GiB	n/a	n/a
CPU-M	8	32 GiB	n/a	n/a
CPU-L	32	128 GiB	n/a	n/a
HighMem-S	8	64 GiB	n/a	n/a
HighMem-M	32	256 GiB	n/a	n/a
HighMem-L	128	1024 GiB	n/a	n/a
GPU-S	8	32 GiB	1 NVIDIA A10G	24 GiB
GPU-M	48	192 GiB	4 NVIDIA A10G	96 GiB
GPU-L	96	1152 GiB	8 NVIDIA A100	320 GiB

The complexity of comparisons increases due to the differing computational needs between CPU-centric and GPU-enhanced instances. Snowflake's pricing includes Credit Costs for Virtual Storage Services, where usage is quantified in "credits" as detailed in Table 7, and On-Demand Storage Costs based on data volume, as outlined in Table 8. Storage costs are segmented by tiers, with prices varying according to the amount of data stored. The tiered pricing encourages larger data storage by offering lower rates for higher volumes.

Table 7: Credits Per Hour for Snowflake.

Service	XS	S	M	L	2XL	4XL
Standard Warehouse	1	2	4	8	32	128
Snowpark Optimized	N/A	N/A	6	12	48	192
SnowPark Container Service Compute						
CPU	0.11	0.22	0.43	1.65	-	-
High-Memory CPU	N/A	0.56	2.22	8.88	-	-
GPU	N/A	1.14	5.36	28.24	-	-

Table 8: On Demand - Credit and Standard Storage Pricing - Snowflake.

On Demand - Credit Pricing						
Provider	Region	Standard	Enterprise	Business Critical	VPN	
AWS	Dublin	2.6	3.9	5.2	7.8	
AZURE	Ireland	2.6	3.9	5.2	7.8	
GCP	Netherlands	2.6	3.9	5.2	7.8	
On Demand - Standard Storage Pricing						
		Tier 1	Tier 3	Tier 5	Tier 7	
AWS	Dublin	23	19.94	16.86	13.8	
AZURE	Ireland	23	19.94	16.86	13.8	
GCP	Netherlands	20	20	20	20	

4.4 Microsoft Azure Synapse Analytics

Azure Synapse Analytics integrates big data and data warehousing, supporting sources from relational databases to data lakes. It offers over 125 VM types across six categories (e.g., General Purpose, Compute-Optimized, GPU) and 31 OS options on Windows and Linux, ensuring broad configurability.

For this analysis, we focus on the Azure D2ads

v5 instance (Table 9) because its hardware closely aligns with the discussed platforms, facilitating direct comparison and consistent data presentation. Other instances, which are not detailed here, have similar descriptions but vary in values and specifications.

Table 9: Azure Dads v5 Series Pricing.

Instance	vCPU/RAM/Storage	Pay as you go	Savings (% off)
D2ads v5	2 / 8 GiB / 75 GiB	\$75.19/mo	1-Yr: 31, 3-Yr: 54
D4ads v5	4 / 16 GiB / 150 GiB	\$150.38/mo	1-Yr: 31, 3-Yr: 54
D8ads v5	8 / 32 GiB / 300 GiB	\$300.76/mo	1-Yr: 31, 3-Yr: 54
D16ads v5	16 / 64 GiB / 600 GiB	\$601.52/mo	1-Yr: 31, 3-Yr: 54
D32ads v5	32 / 128 GiB / 1.2 TB	\$1,203.04/mo	1-Yr: 31, 3-Yr: 54
D48ads v5	48 / 192 GiB / 1.8 TB	\$1,804.56/mo	1-Yr: 31, 3-Yr: 54
D64ads v5	64 / 256 GiB / 2.4 TB	\$2,406.08/mo	1-Yr: 31, 3-Yr: 54
D96ads v5	96 / 384 GiB / 3.6 TB	\$3,609.12/mo	1-Yr: 31, 3-Yr: 54

Table 9 outlines the pricing models for the Azure Dads v5 series, featuring a tiered pricing structure based on vCPUs, RAM, and storage capacities across different payment plans. Pay-as-you-go rates increase with resource capacity, reflecting the direct costs of computing resources. In contrast, long-term commitment plans, such as the 1-year and 3-year savings plans, offer cost reductions ranging from 31% to 54%.

5 DISCUSSION

A meticulous comprehension and rigorous evaluation of the financial implications and the Return on Investment (ROI) are imperative prerequisites prior to the initiation of a Data Warehouse (DW) development within a cloud environment. The genesis of this study was a case analysis that reviewed these aspects. The evaluation process necessitates an estimation of operational expenses encompassing storage, data processing, and information transfer, alongside an examination of the cost reductions afforded by the scalability and elasticity inherent to cloud infrastructures. Of equal significance is the analysis of both tangible and intangible benefits derived from the deployment of a cloud-based DW, such as enhanced response times for analytical queries, improved operational efficiency, and bolstered support for data-driven decision-making.

However, conducting a balanced comparison across available market solutions poses a substantial challenge due to the variance in technical configurations among platforms. These differences encompass distinct processing architectures, data partitioning techniques, and query optimization strategies (as presented in detail in section 4).

To ensure a fair and accurate assessment, we strove to standardize the configurations between the different services, keeping the vCPU, memory, and

storage capacity specifications as closely aligned as possible. This methodological approach allows us to provide a clear view of the variations in cost and performance. Table 10 summarizes the configurations and hourly prices of four popular cloud DW platforms. The price column for BigQuery is empty, indicating missing data.

Table 10: Warehouse characteristics and its configurations.

Warehouse	Node Name	vCPU	Memory	Storage	Price/h
BigQuery	E2	up to 32	8 GB	128 GB	-
Redshift	dc2.large	2	15 GB	-	\$0.3
Snowflake	CPU_X64_S	4	16 GB	-	\$0.22
Azure	D4ads v5	4	16 GB	150 GB	\$0.22

Table 11 shows a comparison of data warehouse solutions according to cost based on a consultation carried out in March 2024 for Snowflake, BigQuery, Redshift, and Azure Synapse.

Table 11: Comparative Costs of Data Warehouse Solutions.

Warehouse	Storage (1TB)/Mo	Compute Cost/h	Serverless Cost/TB	1-Yr Saving	3-Yr Saving
Snowflake	\$23	\$2.3	-	-	-
BigQuery	\$20	\$0.044/slot	\$6	-	-
Redshift	\$24	\$0.3	\$5	~34%	~63%
Azure Synapse	\$20	\$0.2216	\$4.6	~24%	~47%

5.1 Processing and Storage Costs

In this section, we explore in detail the results obtained in the comparisons, focusing on the variations in hardware configurations and the pricing models of the selected data warehouses.

5.1.1 Storage Cost per TB per Month

BigQuery and Azure Synapse have an identical monthly cost of \$20 per TB per month, which is also in line with Snowflake \$23. Redshift is slightly more expensive, at \$24.

5.1.2 Capacity-Based Cost

This model requires customers to choose and pay for a specific capacity of resources that are reserved regardless of usage. Snowflake offers hourly-based costs that are considerably higher (\$2.3/h) compared to other options. BigQuery offer (\$0.044 per slot per hour) and Redshift (\$0.3/h). Azure Synapse offers the best choice (\$0.2216/h).

5.1.3 Serverless Computing Costs

This model automates resource scaling and charges based on actual usage, eliminating server manage-

ment. BigQuery and Redshift offer serverless computing at competitive rates (\$6/TB and \$5/TB, respectively), suitable for variable workloads. Azure Synapse offers the lowest rate at \$4.6/TB. Snowflake does not provide a serverless option. This approach is advantageous for companies with fluctuating data needs, allowing them to pay only for utilized resources without managing infrastructure.

5.1.4 Discounts for Contracts

Redshift and Azure Synapse offer notable discounts for one- and three-year commitments, up to 63% and 47%, respectively, making Redshift attractive despite its higher base price.

Each DW solution presents distinct advantages. BigQuery's flexible pricing suits variable workloads. Snowflake and Synapse offer versatile configurations and competitive discounts, while Redshift's long-term savings benefit organizations able to commit in advance.

Table 12 shows the values associated with storing and backing up data in the same region (Europe/Europe) and another continent (Europe/North America).

Table 12: Comparative costs of In/Out storage solutions.

Warehouse	Storage(\$) 1TB/Mo	Inbound	Out. EU/EU (\$/TB)	Out. EU/US (\$/TB)	Est. Cost (\$)
Snowflake	23	Provider	20 / TB	50 / TB	€2,000
BigQuery	20	Free	0.2 / GB	0.5 / GB	€300
Redshift	24	Free	Free	0.2 / GB	€200
Azure	20	Free	0.019 / GB	0.047 / GB	€320

5.1.5 Variability in In/Out Costs

Inbound: All providers except Snowflake offer free inbound data. Snowflake indicates a 'Cloud Provider Cost,' implying potential costs from the cloud storage provider, which varies based on the customer's choice, making direct comparisons challenging.

Outbound: Outbound rates differ significantly in values and units (per GB vs. per TB) and depend on the storage region and destination. For instance, a lower per GB rate may become substantial when scaled to TB.

5.2 Regional Storage Implications

In this subsection, we will explore the financial, legal, and logistical complexities associated with storing data in Europe with different backup strategies, whether within the same continent or transcontinentally in the United States.

5.2.1 Europe/Europe

This cost refers to data storage within Europe, including backups in other European regions. Price variations reflect differing policies and infrastructures across providers. Redshift, for example, offers the first 10 TB per month free. For this analysis, geographically close servers were selected: Snowflake (Dublin, Ireland, Netherlands), AWS (Frankfurt, Ireland, London, Milan, Paris, Spain, Stockholm, Zurich), Azure (Paris, Frankfurt, Milan, Ireland, Madrid, London, Netherlands), and BigQuery (Belgium, Berlin, Frankfurt, London, Madrid, Milan, Paris).

5.2.2 Europe/America

Table 12 presents the cost of storing data in Europe with backups in the U.S. Transfer costs vary significantly due to inconsistent pricing units and models. Snowflake charges €50/TB, reflecting high international bandwidth and security costs. BigQuery applies €0.5/GB, suiting variable transfer needs. Redshift is more competitive at €0.2/GB, indicating infrastructure efficiency. Azure offers the lowest rate at €0.047/GB, making it an appealing option for users with frequent or large-scale transatlantic transfers.

These costs are shaped by multiple complex factors that directly affect data strategy and financial planning.

Physical Distance and Infrastructure: Variations in distance between European and U.S. data centers impact latency and transfer costs, which tend to rise with reliance on long-distance networks and international gateways.

International Transfer Fees: Cross-border transfers incur higher fees due to ISP charges and regulatory overheads tied to international data movement.

Data Protection Regulations: Divergent laws, like the GDPR in Europe and U.S. state-specific rules, require additional compliance and security measures, increasing operational complexity and associated costs.

5.3 ETL Tools

Table 13 shows information on the costs associated with the ETL tools Azure Data Factory, AWS Glue, and Cloud Data Flow. Each tool offers a unique billing model that can significantly impact the total cost of operation, depending on the volume of data, frequency of task execution, and complexity of the data operations involved.

Table 13: Comparative Costs of ETL Tools.

ETL Tools	Orchestration (\$/run)	Execution (\$/unit)
Azure Data Factory	\$0.001 per run	\$0.000045 per min
AWS Glue	\$0.025 per second	\$0.00025 per GB
Cloud Data Flow	\$0.05 per Vcpu-hour	\$0.023 per GB

5.3.1 Azure Data Factory

Features an advantageous pricing model for large-scale operations, with a charge of 1 dollar per 1,000 executions. This cost makes it ideal for scenarios involving numerous pipelines that are executed frequently but are relatively light in terms of data processing. In addition, the cost of 0.168 per hour for executing and debugging data flows effectively supports processes that require less processing time, offering a cost-effective option for maintaining continuous operations without incurring high costs. However, it is essential to note that there is a charge for inactive pipelines, stipulated at 0.80 per month, which can introduce additional costs if not well managed.

5.3.2 AWS Glue

Adopts a billing methodology that emphasizes processing time, with a rate of 0.29 dollars per second. This model can result in high costs for processing that takes long periods or is highly complex. Such a pricing structure can prove disadvantageous for ETL projects that require substantial processing power or extend over long time intervals, making cost management a critical aspect. In addition, charging 0.00025 dollars per GB of data processed presents a competitive cost for operations that handle large volumes of data. This pricing per volume of data processed offers cost predictability that benefits large-scale operations, facilitating financial planning for organizations that handle large amounts of data.

5.3.3 Cloud Data Flow

The cost per vCPU-hour in Cloud Data Flow, set at US\$0.05, can make this tool a relatively more expensive option for CPU-intensive processing, especially when compared to alternative solutions that do not implement vCPU-based charging. This charging is based on the allocated processing capacity, applied independently of the actual utilization of resources.

Cloud Data Flow offers a cost of 0.023 dollars per GB of stored data, which remains constant regardless of data usage or activity. This pricing model is similar to the one adopted by AWS Glue, providing a reasonable and predictable cost for data storage.

5.4 DW Selection

To validate the proposed comparison methodology, we applied it to a real-world scenario involving the selection of a cloud DW platform for a startup in the early stages of building its analytics infrastructure. The main requirements of this scenario included: (i) limited initial investment capacity, (ii) expectation of workload growth over time, (iii) preference for serverless capabilities to simplify management, and (iv) ease of integration with existing tools such as Power BI.

Among the evaluated options, Azure Synapse was identified as the most suitable solution given this context. Azure presented a well-balanced hardware configuration and efficient costs for 1 TB per month storage at \$4.60/TB in its serverless model, in addition to progressive offers for long-term contracts (detailed in Table 11). It stood out for its competitive cost of serverless computing, as presented in section 5.1.3, and moderate prices for capacity-based operations, detailed in section 5.1.2. The native integration with Power BI adds even more value to our decision.

The selection process followed the evaluation criteria and cost models described in the Methodology section. This included standardizing hardware configurations (Table 10), comparing operational cost structures, and considering regional storage implications (Table 12). The structured comparison supported the decision, and the methodology proved helpful in narrowing the options in a practical and reproducible way.

However, it is essential to note that this choice is mainly linked to the specific characteristics of our task and could vary for other problems.

Professionals in similar startup contexts, where budget constraints, platform simplicity, and native integration with existing tools are essential, may benefit from applying the same methodology to guide their selection.

Another decisive factor in selecting the DW to use is prior knowledge of the providers' cloud platforms. Although this is ultimately a qualitative factor, if the company already uses other services from the same providers, it might reduce learning and adaptation costs.

As for the other solutions, we highlight that BigQuery is suitable for large-scale analysis within the Google Cloud ecosystem, with a query-based pricing model, which is ideal for companies that need real-time streaming analysis and integrations with machine learning. AWS Redshift, on the other hand, is best for companies that can plan data usage and commit to the long term. It offers significantly re-

duced prices through one- and three-year commitments, ensuring long-term cost-effectiveness. Finally, Snowflake stands out for its performance in multi-cloud environments and its architecture, which separates storage from computing. It enables nearly unlimited scalability and a strictly pay-as-you-go model, ideal for companies with variable data processing needs.

6 CONCLUSIONS

This position paper highlights the challenges faced by startups when selecting a cloud-based data warehouse (DW) provider. While modeling and implementing a DW are complex tasks, the provider selection process, though expected to be straightforward, is in practice intricate due to the diversity in technical configurations, pricing models, and service structures.

To support this decision, we conducted a systematic comparison of four leading DW technologies: Google BigQuery, AWS Redshift, Azure Synapse, and Snowflake. The analysis focused on critical attributes such as cost, processing capacity, storage, and integration with ETL tools. We applied the methodology in the context of a real-world startup project in the space traffic management domain, demonstrating how such a structured evaluation can guide informed decision-making.

Our findings underscore the significant heterogeneity across platforms, which complicates fair comparisons and increases the cognitive load on decision-makers. Based on this evidence, we advocate for greater standardization in the description of DW service offerings, particularly regarding resource specifications and pricing transparency, to facilitate more accessible and equitable evaluations across different organizational contexts.

ACKNOWLEDGEMENTS

This work was supported by Project No. 7059 - Neuraspace - AI fights Space Debris, reference C644877546-00000020, supported by the RRP - Recovery and Resilience Plan and the European Next Generation EU Funds, following Notice No. 02/C05-i01/2022, Component 5 - Capitalization and Business Innovation - Mobilizing Agendas for Business Innovation.

This work was also partially financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025.

REFERENCES

- Aftab, U. and Siddiqui, G. F. (2018). Big data augmentation with data warehouse: A survey. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2785–2794.
- Ahlawat, P., Borgman, J., Eden, S., Huels, S., Iandiorio, J., Kumar, A., and Zakahi, P. (2023). A new architecture to manage data costs and complexity. *Boston Consulting Group (BCG)*, pages 1–12.
- Aljamal, R., El-Mousa, A., and Jubair, F. (2019). A user perspective overview of the top infrastructure as a service and high performance computing cloud service providers. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 244–249.
- Badiuzzaman Biplob, M., Sheraji, G. A., and Khan, S. I. (2018). Comparison of different extraction transformation and loading tools for data warehousing. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*, pages 262–267.
- Berisha, B., Mëziu, E., and Shabani, I. (2022). Big data analytics in cloud computing: an overview. *Journal of Cloud Computing*, 11(1):24.
- Google (2025). Bigquery: Cloud data warehouse. Accessed: 2025-10-02.
- Harby, A. A. and Zulkernine, F. (2022). From data warehouse to lakehouse: A comparative review. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 389–395.
- Inc., S. (2025). Snowflake: The data cloud. Accessed: 2025-02-02.
- Kim, T.-h., Ramos, C., and Mohammed, S. (2017). Smart city and iot.
- Microsoft (2025). Azure synapse analytics. Accessed: 2025-01-03.
- Santos, M. Y. and Costa, C. (2022). *Big data: concepts, warehousing, and analytics*. CRC Press.
- Serra, J. (2024). *Deciphering Data Architectures*. "O'Reilly Media, Inc."
- Services, A. W. (2025). Aws official site. Accessed: 2025-10-01.
- Soma, V. (2022). Comparative study of big query, redshift, and snowflake. *North American Journal of Engineering Research*, 3(2).
- Uddin, M. K. S. and Hossan, K. M. R. (2024). A review of implementing ai-powered data warehouse solutions to optimize big data management and utilization. *Academic Journal on Business Administration, Innovation & Sustainability*, 4(3):10–69593.
- Villamizar, M., Garcés, O., Ochoa, L., Castro, H., Salamanca, L., Verano, M., Casallas, R., Gil, S., Valencia, C., Zambrano, A., et al. (2017). Cost comparison of running web applications in the cloud using monolithic, microservice, and aws lambda architectures. *Service Oriented Computing and Applications*, 11:233–247.