DOI: http://dx.doi.org/10.26483/ijarcs.v15i3.7099

ISSN No. 0976-5697

Volume 15, No. 3, May-June 2024

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An overview of cloud data warehouses: Amazon Redshift (AWS), Azure Synapse (Azure), and Google BigQuery (GCP)

Praveen Borra
Computer Science
Florida Atlantic University
Boca Raton USA

***Abstract*:** Massive volumes of structured and semi-structured data originating from various sources can be efficiently stored, managed, and analysed using a data warehouse. It is specifically designed for query and analysis purposes, rather than for transaction processing, offering a strong foundation for making data-driven decisions. Organizations and businesses are drawn to cloud platforms for numerous reasons, including faster development and deployment, superior performance, low latency, and cost efficiency. The study delves into the cloud data warehouse services provided by three prominent cloud service providers: GCP, Microsoft Azure, and Amazon Web Services (AWS). Three of the most prominent cloud providers—AWS, Azure, and GCP—offer cloud data warehouse services: Amazon Redshift, Azure Synapse, and Google BigQuery. Selecting the right cloud data warehouse is a complex task due to the myriads of options available from these major providers. By carefully evaluating these factors, organizations can decide which cloud data warehouse service best meets their needs, balancing performance, latency, integration, security, and cost considerations to optimize their cloud strategy.

***Keywords*:** Data warehouse, Cloud service provider, AWS, Azure, and GCP Amazon Redshift, Azure synapse, and Google BigQuery

I. INTRODUCTION

Massive volumes of organised and semi-structured data from many sources can be stored, managed, and analysed in a data warehouse. It is specifically designed for query and analysis purposes, rather than for transaction processing, offering a solid foundation for making data-driven decisions. Organizations and businesses are drawn to cloud platforms for numerous reasons, including faster development and deployment, superior performance, low latency, and cost efficiency. These advantages apply to new analytic applications, applications currently running on-premises, and those already on other cloud platforms. Key considerations for organizations include performance, latency, cost estimation, scalability, integration, security, and compliance [1][8][9][10].

Database systems that are housed on cloud computing platforms, such as Microsoft Azure, Amazon Web Services (AWS), or Google Cloud Platform (GCP), are known as cloud data warehouses. Its purpose is to facilitate complicated searches and data analysis by managing massive volumes of structured and semi-structured data. Here are some notable features and benefits of a cloud data warehouse: Scalability, managed services, cost efficiency, high performance, integration, accessibility, and security [17].

Cloud data warehouses are highly adaptable to different workloads and storage requirements. These systems provide extensive managed services, handling tasks like hardware provisioning, software updates, backups, and disaster recovery. Cloud data warehouses charge customers just for the resources they really need, allowing them to pay as they go. They provide fast results for analytics and queries on enormous datasets by making use of technologies like data compression, columnar storage, and massively parallel processing (MPP). In addition to supporting a single data environment, they interact effortlessly with a variety of data sources, BI tools, and cloud services. Because they are hosted in the cloud, these data warehouses can be accessed from any location with an internet connection, making remote work and collaboration much easier. In order to keep data kept in the cloud safe, cloud providers employ robust security measures. These procedures include encryption, access limits, and adhering to industry standards.

Amazon Redshift, Google BigQuery, and Microsoft Azure Synapse Analytics are some of the most well-known cloud data warehouse offerings. To help with improved corporate decision-making and insights, these services are commonly used by organisations to store, manage, and analyse huge datasets [17].

II. AMAZON REDSHIFT

Amazon Redshift is used by many clients every day to improve data analytics and obtain useful business insights. Fast and economical company decisions are made possible by Amazon Redshift's fully managed, AI-driven, massively parallel processing (MPP) architecture. For use cases involving artificial intelligence (AI), machine learning (ML), and near real-time applications, AWS's zero-ETL strategy enables smooth data integration.Thanks to advanced security features and precise governance, you can securely share and collaborate on data within and across organizations, AWS regions, and even with third-party providers [2].

One such cloud data warehouse solution is Amazon Redshift, which provides full management and scalability. Without the hassle of setting up a conventional data warehouse, you can access and analyse data using Amazon Redshift Serverless. Even with highly unpredictable and demanding workloads, optimal performance is guaranteed by automatically

provisioning resources and dynamically scaling capacity. There are no fees when the data warehouse is not in use; you just pay for the resources that you really use. With the Amazon Redshift query editor v2 or your chosen BI tool, you can load data and begin querying right away. You'll experience great pricing performance and familiar SQL functionality in an easy-to-use, zero-administration environment [2].
For enterprise-level query and administrative needs, consider an Amazon Redshift data warehouse, a relational database with great performance. Many client applications, including BI, reporting, data, and analytics tools, are compatible with Amazon Redshift.Amazon Redshift analytical queries use a multi-stage process to retrieve, compare, and evaluate massive datasets in order to get the desired outcomes. Columnar data storage, sophisticated data compression encoding algorithms, and highly parallel processing are the ways in which Amazon Redshift improves storage and query efficiency. An outline of the Amazon Redshift system architecture is given in this section [3][11].

Amazon Redshift Architecture

Amazon Redshift data warehouse architecture components are introduced in this part. Check out the image below for a visual illustration of these components and how they work together. Strong data management and analytics skills are made possible by each component's vital function in keeping the data warehouse efficient and scalable.

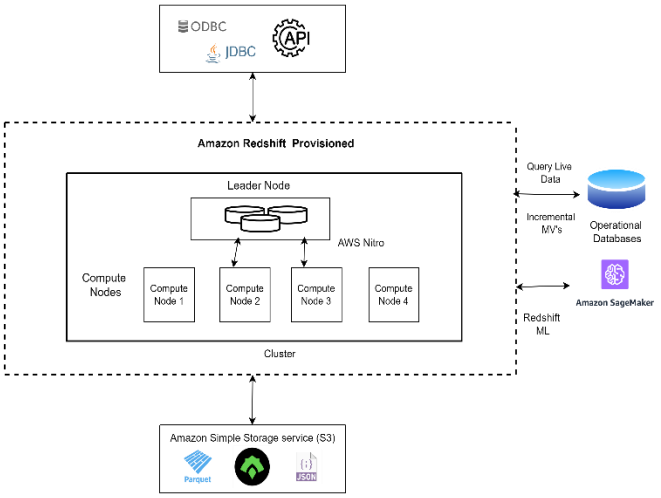

Figure 1: Amazon Redshift Architecture

The Amazon redshift architecture includes several key components: The following components make up the cluster: client applications, compute nodes, redshift managed storage, database slices, and a leader node.

Client Applications

Due to its PostgreSQL base, Amazon Redshift is compatible with the majority of SQL client applications and works effortlessly with a wide range of data loading, ETL, and business intelligence (BI) analytics tools. Refer to the documentation for Amazon Redshift and PostgreSQL to learn about the main distinctions between the two databases.

Clusters

An Amazon Redshift data warehouse's core component is the cluster, wherein a leader node coordinates the activities of one or more compute nodes; this node is responsible for communicating with clients and handling external communication, rendering the compute nodes invisible to programmes running outside the cluster.

Leader Node

By processing and developing execution plans for complicated queries, sending compiled code and specified data sections to each node, and coordinating interactions with compute nodes, the leader node handles communication with client applications. For more information, refer to the documentation for the SQL functions supported on the leader node [3]. Compute nodes can only receive SQL statements when queries reference tables on those nodes. Additionally, certain SQL functions are only available on the leader node and can cause errors when used with compute node tables.

Compute Nodes

Compute nodes in Amazon Redshift execute compiled code from the leader node, providing intermediate results for aggregation, with each node equipped with dedicated CPU and memory that can scale by adding nodes or upgrading types; various node types cater to diverse compute needs detailed in the Amazon Redshift Management Guide [3].

Redshift Managed Storage

Data is stored in Redshift Managed Storage (RMS) by Amazon Redshift. Storage can be independently scaled up to petabytes with Amazon S3. To optimise performance, SSD-based local storage is used as tier-1 cache. Automatic adjustments are made based on workload patterns.

Node Slices

To facilitate the parallel distribution of data and workload, Amazon Redshift partitions each compute node into slices, with memory and disc space given to each slice. The number of slices is defined by the size of the cluster's nodes and is maintained by the leader node.To improve the efficiency of data operations that include parallel processing, you can optimise data distribution among node slices by specifying a distribution key when you create a table in Amazon Redshift. For advice, see the article on Choose the optimum distribution style.

Internal Network

To provide secret, high-speed network communication between the leader node and computing nodes, Amazon Redshift uses proximity, unique communication protocols, and high-bandwidth connections. Client applications cannot access the isolated network where the compute nodes are operating.

Databases

Amazon Redshift clusters house databases where user data resides on compute nodes, with SQL client interaction managed by the leader node overseeing query execution across nodes,

supporting OLTP functions while prioritizing high-performance analysis of extensive datasets [4].

III. AZURE SYNAPSE

Enterprise data warehousing and powerful Big Data analytics are combined in Microsoft Azure Synapse, a comprehensive analytics solution. It offers flexibility in querying data, supporting both serverless and dedicated resource models at scale, and empowering organizations to manage and analyze vast datasets efficiently [4]. Quickly gain insights from data warehouses and big data systems with Azure Synapse, an advanced analytics service. Data Explorer is used for log and time series analytics, Pipelines are used for seamless data integration and ETL/ELT workflows, and it has tight integration with Azure services such as Azure Cosmos DB, Azure ML, and SQL technologies for enterprise data warehousing [4].

Microsoft Azure Synapse Architecture

With Azure Synapse Analytics, you can combine Big Data analytics with business data warehousing. The node-based design of Azure Synapse's Dedicated SQL pool (formerly SQL DW) allows applications to issue T-SQL commands by connecting to a Control node. By distributing tasks among Compute nodes, the Control node's distributed query engine prepares queries for parallel processing. To ensure correct query results, these Compute nodes store user data in Azure Storage and conduct parallel queries with the help of the Data Movement Service (DMS) [5]. The data warehouse architecture of Microsoft Azure Synapse is described in this section. Check out the image below for a visual illustration of these components and how they work together. Strong data management and analytics skills are made possible by each component's vital function in keeping the data warehouse efficient and scalable.

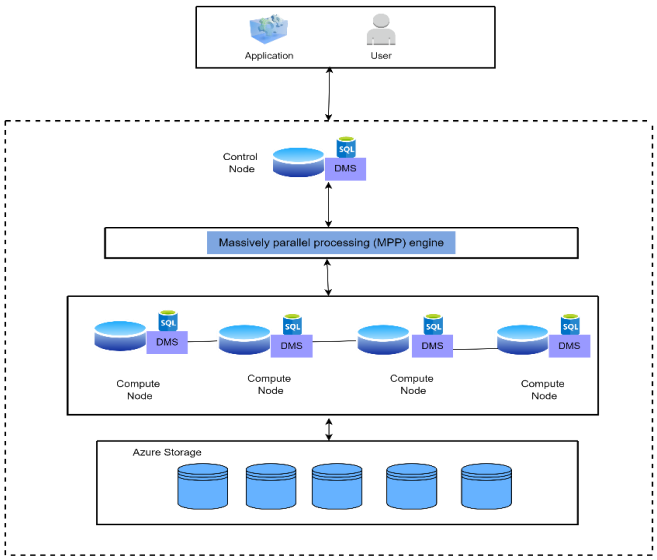

Figure 2: Microsoft Azure Synapse Architecture

The Microsoft Azure Synapse architecture includes several key components: azure storage, control node, compute node, and distributions.

Azure Storage

To safely handle user data, Dedicated SQL pool SQL (formerly SQL DW) makes use of Azure Storage; storage charges are charged independently according to utilisation. Data is organized into distributions optimized for system performance, allowing users to define distribution patterns such as Hash, Round Robin, or Replicate when configuring tables [5][12].

Control node and Compute node

The Control node serves as the primary interface for applications, managing interactions and optimizing parallel queries through its distributed query engine. Depending on the selected service level for Synapse SQL, compute nodes dynamically remap distributions across a maximum of 60 nodes, providing processing power. In the system views, you can see that each compute node has its own unique identifier.

Distributions

Azure Synapse SQL distributions are the building blocks for storing and processing distributed data in parallel. Hash-distributed tables maximize performance for joins and aggregations by using a hash function to assign rows to specific distributions based on a designated distribution column. Round-robin distributed tables evenly distribute data across distributions without optimization, ideal for rapid data loading but may require reshuffling during joins. Replicated tables cache complete copies on each Compute node, offering fast query performance for small tables at the cost of increased storage and write overhead, making them less suitable for large datasets.

IV. GOOGLE BIGQUERY

Google BigQuery is a managed enterprise data warehouse that offers a wide range of features for managing and analysing data. These features include business intelligence tools, geographic analysis, and machine learning. Its serverless design enables seamless SQL query execution without the need for infrastructure management. Key features include federated queries for integrating external data sources and streaming support for real-time data updates. Google BigQuery's scalable, distributed analysis engine ensures rapid query performance, handling massive datasets efficiently from terabytes to petabytes. This is facilitated by a dual-layer architecture: a dedicated storage layer for data ingestion and optimization, and a compute layer for executing analytics tasks, leveraging Google's robust network infrastructure for high-performance operations [6].

BigQuery Architecture

This section introduces the components of the Google BigQuery's architecture. These components are illustrated in the figure below, providing a visual representation of how they interconnect and function together. Strong data management and analytics skills are made possible by each component's vital function in keeping the data warehouse efficient and scalable.

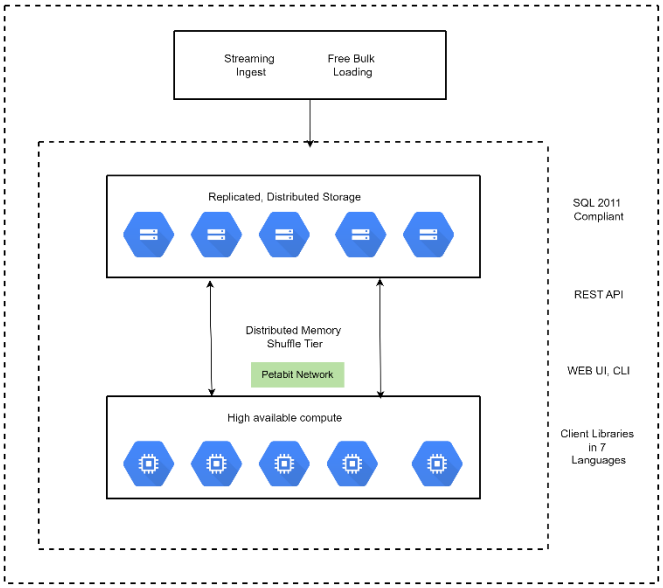

Figure 3: Google BigQuery Architecture

The Goggle BigQuery's architecture includes several key components: compute and storage.

Google Big Query's serverless architecture separates storage and computing, enabling independent and scalable adjustments based on demand. In contrast to conventional node-based cloud data warehousing and on-premise MPP systems, this adaptability offers cost-effective advantages by doing away with the necessity for constant operation of costly compute resources. Data may be quickly ingested and analysed using Standard SQL without the burden of managing database operations and system engineering considerations, making it suitable for organisations of all sizes [7][13].

Compute

When running SQL queries in a massive multi-tenant cluster, BigQuery makes use of Dremel. Dremel transforms queries into execution trees, with 'slots' at the leaves performing data retrieval and computation, while 'mixers' at the branches handle aggregation. Slots are dynamically allocated to maintain fairness among concurrent user queries, with individual users potentially accessing thousands of slots.

Storage

Data is stored in a columnar format that is optimised for efficient reading of structured data by Colossus, Google's global storage system. When compared to conventional data warehouses, Colossus's ability to manage replication, recovery, and remote management makes it easy to scale to massive data volumes without breaking the bank on additional computing power.

With the help of Google's petabit Jupiter network and the'shuffle' function, compute and storage are able to communicate with one another quickly. Borg, Google's forerunner to Kubernetes, orchestrates BigQuery by managing resources for mixers and slots.

Without the downtime and upgrade problems of conventional systems, Google is always working to improve these technologies so that BigQuery users have better performance, durability, efficiency, and scalability.

V. Comparision of Amazon Redshift, Microsoft Azure and Google BigQuery

Google BigQuery, Amazon Redshift, and Microsoft Azure Synapse are three data warehousing and analytics solutions that offer different perspectives and capabilities.

AWS Redshift focuses on scalable enterprise data warehousing optimized for OLAP workloads, utilizing a massively parallel processing (MPP) architecture [15][17]. Azure Synapse combines data warehousing with big data analytics, integrating SQL and Spark pools closely with Azure services [16][17]. Google BigQuery offers a serverless, scalable platform for rapid SQL querying and analysis of large datasets, deeply integrated with the broader Google Cloud environment [17][18]. These platforms vary in architecture, scalability options, querying capabilities, integration possibilities, and performance characteristics, catering to diverse requirements in enterprise data management and analytics solutions.

Table 1: Comparison of Amazon RedShift, Microsoft Azure, and Google BigQuery

Feature / Service	AWS Redshift	Azure Synapse	Google BigQuery
Primary Focus	AWS Redshift focuses on enterprise data warehousing, optimized for OLAP workloads.	Azure Synapse integrates data warehousing with big data analytics capabilities.	Google BigQuery is designed as a scalable data warehouse and analytics platform.
Architecture	AWS Redshift employs a Massively Parallel Processing (MPP) architecture with separate compute and storage nodes.	Azure Synapse combines SQL pools and Spark pools, tightly integrating with Azure services.	Google BigQuery operates on a serverless architecture, decoupling computing, and storage.
Querying Support	AWS Redshift supports standard SQL with optimizations for its architecture.	Azure Synapse offers SQL querying with dedicated SQL pools and Spark for large-scale data processing.	Google BigQuery supports ANSI SQL queries and federated queries across diverse data sources.
Integration	AWS Redshift integrates deeply with AWS services like S3, Glue, and QuickSight.	Azure Synapse seamlessly integrates with Azure services including Blob Storage, Data Lake, and Power BI.	Google BigQuery tightly integrates with Google Cloud services such as Storage, Dataflow, and AI Platform.
Scalability	AWS Redshift scales compute and storage independently, accommodating large-scale data warehouses.	Azure Synapse scales compute and storage resources, supporting both on-demand and provisioned options.	Google BigQuery automatically scales to handle large datasets and varying workloads.
Performance	AWS Redshift is optimized for high query concurrency and efficient data warehousing operations.	Azure Synapse provides scalable processing capabilities for both structured and unstructured data.	Google BigQuery delivers fast query performance across terabytes to petabytes of data.
Cost Management	AWS Redshift charges based on compute node hours and storage usage, with options for cost-effective pricing models.	Azure Synapse operates on a pay-as-you-go model, with flexibility in pricing for SQL pools and serverless querying.	Google BigQuery charges based on data processed and storage used, offering flat-rate options for predictable workloads.

This table summarizes the key differences in focus, architecture, querying support, integration with cloud ecosystems, scalability, performance, cost management, language support, and overall ecosystem integration for AWS Redshift, Azure Synapse, and Google BigQuery [1][16][17][18].

VI. CONCLUSION

Amazon Redshift, Microsoft Azure, and Google BigQuery are just a few examples of cloud-based data warehouses that are analysed in this study. In summary, AWS Redshift, Azure Synapse, and Google BigQuery stand out as top-tier solutions in data warehousing and analytics, each designed with unique strengths to meet diverse organizational needs. Whether prioritizing OLAP workloads with AWS Redshift's MPP architecture, integrating SQL and Spark functionalities in Azure Synapse, or harnessing Google BigQuery's serverless scalability and seamless Google Cloud integration, businesses can effectively enhance their data management and analytics capabilities based on specific operational demands and preferences.

VII. FUTURE WORK

However, selecting the right cloud service provider requires a thorough evaluation of multiple factors. Improving the efficiency of data processing, making it more scalable, and expanding integration capabilities across new platforms like Google BigQuery, Amazon Redshift, Microsoft Azure, and others could be the focus of future study.

REFERENCES

[1] Praveen Borra, An Overview of Cloud Computing and Leading Cloud Service Providers, International Journal of Computer Engineering and Technology (IJCET), 15(3), 2024, pp. 122-133

[2] AWS (Amazon Web Services) official documentation, "Amazon Redshift Overview," Available: https://docs.aws.amazon.com/redshift/latest/dg/c_redshift_system_overview.html. Accessed: May 31, 2024.

[3] AWS (Amazon Web Services) official documentation, "Amazon Redshift Management Guide," Available: https://docs.aws.amazon.com/redshift/latest/dg/c_high_level_system_architecture.html. Accessed: May 31, 2024.

[4] Microsoft Azure official documentation, "Introduction to Azure Synapse Analytics," Available: https://learn.microsoft.com/en-us/azure/synapse-analytics/overview-what-is. Accessed: May 31, 2024.

[5] Microsoft Azure official documentation, "Azure Synapse Architecture Overview," Available: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture. Accessed: May 31, 2024.

[6] Google Cloud official documentation, "Introduction to Google BigQuery," Available: https://cloud.google.com/bigquery/docs/introduction. Accessed: May 31, 2024.

[7] Google Cloud official documentation, "BigQuery Architecture," Available: https://cloud.google.com/blog/products/data-analytics/new-blog-series-bigquery-explained-overview. Accessed: May 31, 2024.

[8] Microsoft Azure, "What is Cloud Computing?" Available: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-computing. Accessed: May 31, 2024.

[9] Google Cloud, "What is Cloud Architecture?" Available: https://cloud.google.com/learn/what-is-cloud-architecture. Accessed: May 31, 2024.

[10] Amazon Web Services, "What is Cloud Computing?" Available: https://aws.amazon.com/what-is-cloud-computing/. Accessed: May 31, 2024.

[11] Amazon Web Services, "AWS Documentation," Available: https://docs.aws.amazon.com/. Accessed: May 31, 2024.

[12] Microsoft Azure, "Azure Documentation," Available: https://learn.microsoft.com/en-us/azure/?product=analytics. Accessed: May 31, 2024.

[13] Google Cloud, "Google Cloud Documentation," Available: https://cloud.google.com/docs. Accessed: May 31, 2024.May 2024

[14] Zhang, Q., Cheng, L. & Boutaba, R. Cloud computing: state-of-the-art and research challenges. J Internet Serv Appl 1, 7–18 (2010)

[15] Y. Jadeja and K. Modi, "Cloud computing - concepts, architecture and challenges," 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), Nagercoil, India, 2012, pp. 877-880, doi: 10.1109/ICCEET.2012.6203873

[16] Praveen Borra "Exploring Microsoft Azure's Cloud Computing: A Comprehensive Assessment" ,International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) ,vol. 2, no. 8, pp. 897 - 906, 2022.

[17] Praveen Borra, Comparison and Analysis of Leading Cloud Service Providers (AWS, Azure and GCP), International Journal of Advanced Research in Engineering and Technology (IJARET), 15(3), 2024, pp. 266-278

[18] Praveen Borra "A Survey of Google Cloud Platform (GCP): Features, Services, and Applications" ,International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) ,vol. 4, no. 3, pp. 191 - 199, 2024.