Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-6, Issue-7, July 2024
DOI: 10.53469/jrse.2024.06(07).20

Evolution of Data Warehouse Architecture from Local to Cloud

Rabia Tugba Egmir

Independent Researcher, USA
egmirrtugba@gmail.com

Abstract: ***In today's era, organizations are more committed to analyzing, studying, and prioritizing data to make data-driven business decisions. Companies' critical decisions and key performance metrics revolve around understanding data. All effective decision-making begins with reliable data, and the Data Warehouse serves as the definitive source of information. The Data Warehouses have improved from single-node static tightly coupled models to dynamic separate cloud, storage, and compute models. This research paper not only studies and evaluates data warehousing architectures from traditional on-premises to the latest cloud models but also highlights the challenges of conventional systems. Importantly, it emphasizes how modern architecture provides practical, real-world solutions, thereby reassuring the audience about the effectiveness of the research. It further discussed a couple of modern architectures of leading data warehouses and recommended their powerful features. Finally, it summarizes the key components of these data warehouses' technological advancements.***

Keywords: data warehouse, cloud, architecture, on-prem, server

1. Introduction

The Data Warehouse, a centralized repository of data and the single source of truth is a cornerstone in modern data analysis. It plays a crucial role in data-driven decision-making, consuming daily data from multiple transactional database systems and storing them in an aggregate business-ready format. This format is designed to answer challenging business questions like forecasting, predicting, and providing data insights relevant to different metrics. The insights derived from the Data Warehouse can be used as input for various data visualizations, enabling more insightful understanding and analysis. [1] The tools used in Data Warehouses have evolved significantly, from traditional systems [2] like Microsoft Power Business Intelligence to the latest cloud-based solutions like Snowflake, Redshift, and Google Big Query. Despite these advancements, all data warehouses are defined by four essential characteristics, as outlined below.

- Subject-oriented—It defines the business problem we are trying to solve; for instance, a company's profit and loss reporting would require financial data to be the subject.
- Integrated – The data from multiple source systems should follow the same naming, calculation, and object creation pattern.
- Time –variant – The historical data should be present and preserved by adding new data. So, time-sensitive data analysis could be performed.
- Nonvolatile – The data in the warehouse should not be updated or deleted; it must only be accessed for reading and referring purposes.

2. Related Work

Data Warehouse Architecture has shown several transformations over time and has always been a topic of interest in data analytical fields. The research [3] explores data warehouses into single, two, and three-layer architectures where the single layer is a source-data warehouse analysis. The two layers are a source-staging-presentation layer, and the three layers are source-staging-reconciled-presentations layer. This research is helpful when building a data warehouse from scratch and identifying which layer makes the most sense per business use case. For instance, a three-layer architecture with an extra data reconciliation step is practical where financial reconciliation is required. Another research [4] highlights various models and classification approaches, including Bill Inmon's top-down, Kimball's bottom-up, and centralized data lake. This research identifies the need to build a data warehouse depending on the organization's use case. The data mart is a subset of a data warehouse; for instance, an organization can have different data marts can be built for different internal departments such as finance and HR, or a central data lake can be built to serve all the business needs of the company if they want to focus on one standard product.

3. Traditional Data Warehouse Architecture

Companies earlier used to install traditional data warehouses on their physical servers [5] due to the design compatibility of these tools with the on-premises servers and hardware. Generally, the typical architecture of such a data warehouse is divided into three tiers, mainly the Top, Middle, and Bottom tiers. It comprises a single node, i.e., a single server architecture.

- Bottom Tier – This ingestion layer extracts data from multiple sources, such as OLTP databases, JSON, and CSV files.
- Middle Tier – This layer transforms and manipulates data per business needs.
- Top Tier – This is a consumption layer for business intelligence responsible for querying and reporting data.

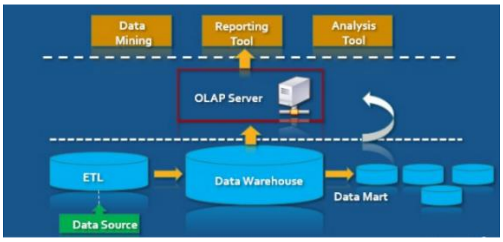

Figure 1: Traditional Data Warehouse Architecture [5]

This architecture supported two types of data loading.

- ETL – Extract, Transform, Load is used for extracting, transforming, and loading data.
- ELT – Extract, Load, Transform is used for extracting, loading, and transforming data.

3.1 Traditional Data Warehouse Limitations

Traditional data warehouses (DWH) have limitations, and it is difficult to keep up with the latest technological advancements. Below are a few significant limitations.

- Lack of data format support- Earlier, we had limited data source formats, and DWH only supported Database, CSV, and JSON files. With advancements on the Internet, a need for web-based API and hierarchical data has arisen.
- Server Disk Limitations- As the data volume increases, the server storage size must increase to accommodate the rising data needs.
- In addition to disk size, server RAM limitations and the RAM for input/output operations must be augmented to avoid data retrieval latency.
- Frequent updates- DWH is built for specific use cases and needs to be modified for new requests.
- Increased Costs- The hardware and servers will keep increasing to meet high data demands.

4. Modern Data Warehouse Architectures

Modern data warehouses are a game changer, along with cloud technologies, self-managed services, and serverless architecture [6]. Modern cloud systems have the advantages listed below.

- Cloud-based systems do not require hardware installations; hence, installing, upgrading, and maintaining massive servers is unnecessary.
- Self-managed services do not require software installation, hotfixes, patch installations, or configuration.
- These improvements reduce the workload and accountability of any company's Infrastructure team.
- Cloud-based DWH also supports various data formats and provides state-of-the-art encryption and data optimization.

In addition, the latest DWH tools, like Snowflake and Google Big Query, have unique architectures that compress data, improve performance, and encrypt data.

4.1 Snowflake Architecture

The snowflake architecture is a mix of shared-nothing and shared-disk database architecture. A shared-disk database is a centralized data repository accessible to all nodes, and a shared-nothing architecture uses Massive Parallel Processing (MPP) computing clusters, where every node's data is stored locally [6]. MPP also allows every node to operate independently, making adding or removing nodes easier. They have input and output devices and work on the same memory. Another advantage is that it provides a higher availability architecture to run computations continuously. In the shared-nothing approach, each node keeps the physical copy of data, so there is a lot of data redundancy and replication happening in the back-end. This redundancy causes extra data storage in Snowflake.

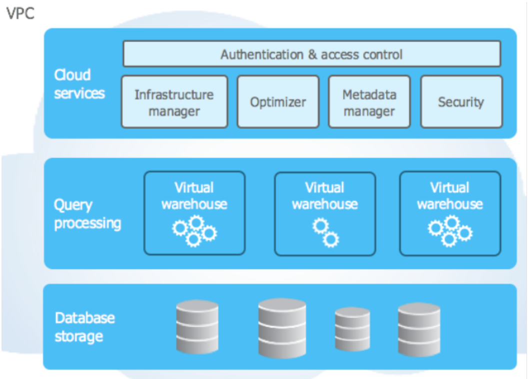

Figure 2: Snowflake Architecture [6]

The three main layers of Snowflake Architecture are as follows:

- Database Storage—When we load the data into Snowflake, its self-service management organizes it for optimization compressions with columnar format. The data objects require SQL queries to show data as they are not accessible straightforwardly.
- Query Processing – Virtual warehouses are massive parallel processing computer clusters with multiple computing nodes. They all work independently and are parallel to each other to perform query processing. As a result, each virtual warehouse has no impact on other virtual warehouses.
- Cloud Services- This controls administrative work, infrastructure, and metadata management. It also aids in query parsing, optimization, and controlling user and role access. Cloud service also uses compute services from Snowflake's cloud provider.

In addition to this robust architecture, Snowflake boasts below critical features [7]

- Improved Authentication is available using Multi-Factor Authentication, Snowflake OAuth, External OAuth, and Single Sign On.
- Cloud deployment is available on major providers such as Amazon Web Service, Google Cloud Platform, and Microsoft Azure.
- Data Encryption and Security are available for row-level column-level data. For sensitive Personal Data, object-level tagging is available.
- For Disaster Recovery, Snowflake has Snowflake-Fail-Safe, which recovers all the historical data 7 days after the disaster.
- The Time Travel feature allows users to query the changed data in tables for the point in time data.
- Multiple data formats, including JSON, Avro, ORC, Parquet, XNL, TSV, S3, Google, and Azure Cloud storage, are supported.
- Advanced data availability features such as sharing, replication, and failover support are provided.
- Extensive API support for Java, Python, and Scala, with connectors including Python, Spark, ODBC, JDBC, Go, Node.js, and PHP, is available.
- For the container application, Snowflake self-managed provides container service to deploy and scale applications fully.
- HIPPA compliance and PHI data supported.

4.2 Google Big Query Architecture

Google Big Query is also a serverless architecture with independent computing and storage resources that can scale up and down depending on the requirements [8]. As opposed to Massive Parallel Processing (MPP) architecture, the scale-on-demand architecture allows customers to scale up resources only when the data volume increases; hence, customers do not need to keep expensive resources running all the time.

The Big Query architecture is divided into three essential parts:

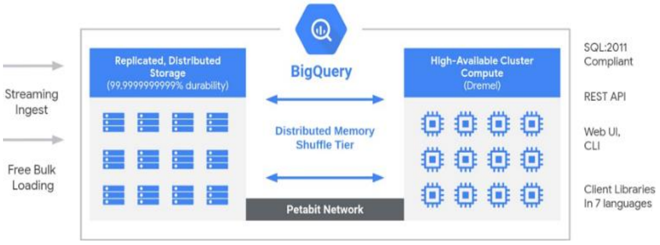

Figure 3: Google Big Query Architecture [8]

- Dremel – The compute resource aids in executing Structure Query Languages (SQL) into large multi-tenant clusters. Dremel transforms the query into an execution tree pattern with leaves and branches for optimized execution. The leaves, also known as slots, are responsible for reading data from storage, and branches, also known as mixers, perform the aggregation.
- Colossus - The storage resource that optimizes data reads using the columnar format and compression algorithm. It also covers replication, distribution management, and recovery so a single point of failure does not occur.
- Jupiter - It is a network resource allowing computing and storage to communicate. With the advancement over MapReduce and Bigtable, Jupiter Fabrics delivers more than 1 Petabit/sec. [9]
- Borg – This orchestrator ensures all queries run successfully and resources get allocated.

Big Query also provides the following powerful features:

- Gemini is an AI-powered feature that assists in writing codes, visually represents data, recommends intelligence insights, improves productivity, and reduces costs.
- Enterprise Capabilities include advanced services such as cross-region disaster recovery achieved by backup recovery, operational health monitoring, and Big Query Migration Services, which migrate data from legacy systems to the cloud.
- Streaming pipelines to provide near real-time data and in-memory service offering high concurrency and query response time in seconds.
- Easy connectivity to Google's native visualization tool Looker to offer built-in business intelligence.
- Supports all standard file formats, and in addition to this, supports open formats such as Delta, Hudi, and Iceberg.
- Built-in Data governance includes cataloging, profiling, data lineage, and data quality.

5. Conclusion

Modern data warehouses have made several advancements, the significant changes being cloud-hosted and multi-node architectures. As DWH becomes more cloud-native, there is no system maintenance, upgrades, or manual scaling of resources as needed. A cloud solution allows organizations to spend less time and budget on Infrastructure Teams and more on data research teams and tools. On the other hand, multi-node architectures where computing and storage are separated, and both can be scaled up or down based on-demand, have been pivotal. In addition to this, DWH has made technological advancements in query enhancement, database optimization, and powerful computer and storage resources.

References

[1] What is a data warehouse? - Data Warehouse explained - AWS. (n.d.). Retrieved from https://aws.amazon.com/what-is/data-warehouse/

[2] Data warehouse. (2024, April 3). In *Wikipedia*. https://en.wikipedia.org/wiki/Data_warehouse

[3] Merseedi, Karwan & Yazdeen, Abdulmajeed & Ibrahim, Abass & Abdulrazzaq, Maiwan & Mahmood, Mayyadah. (2022). Analyses the Performance of Data Warehouse Architecture Types. Applied Soft Computing. 3. 45-57. 10.30880/jscdm.

[4] Yang, Qishan & Ge, Mouzhi & Helfert, Markus. (2019). Analysis of Data Warehouse Architectures: Modeling and Classification. 604-611. 10.5220/0007728006040611.

[5] Oyero, O. (2024). A brief comparison of database, Data Warehouse, Data Mart and Data Lake and these services in Azure. Retrieved from https://techcommunity.microsoft.com/t5/nonprofit-techies/a-brief-comparison-of-database-data-warehouse-data-mart-and-data/ba-p/3944981

[6] Key Concepts & Architecture¶. (n.d.). Retrieved from https://docs.snowflake.com/en/user-guide/intro-key-concepts#snowflake-architecture

[7] Overview of crucial features¶. (n.d.). Retrieved from https://docs.snowflake.com/en/user-guide/intro-supported-features

[8] An overview of BigQuery's architecture and how to quickly get started | google cloud blog. (n.d.). Retrieved from https://cloud.google.com/blog/products/data-analytics/new-blog-series-bigquery-explained-overview

[9] A look inside Google's Data Center Networks. (2015). Retrieved from https://cloudplatform.googleblog.com/2015/06/A-Look-Inside-Googles-Data-Center-Networks.html

[10] BigQuery Enterprise Data Warehouse. (n.d.-b). Retrieved from https://cloud.google.com/bigquery?hl=en#features

Author Profile

Bhushan Fadnis received an MS in Information Science from San Diego State University, USA 2017. He has more than 12+ years of technology experience working in various MNCs and is now a Business Intelligence Engineer in a leading software company in USA.