Data Warehouse

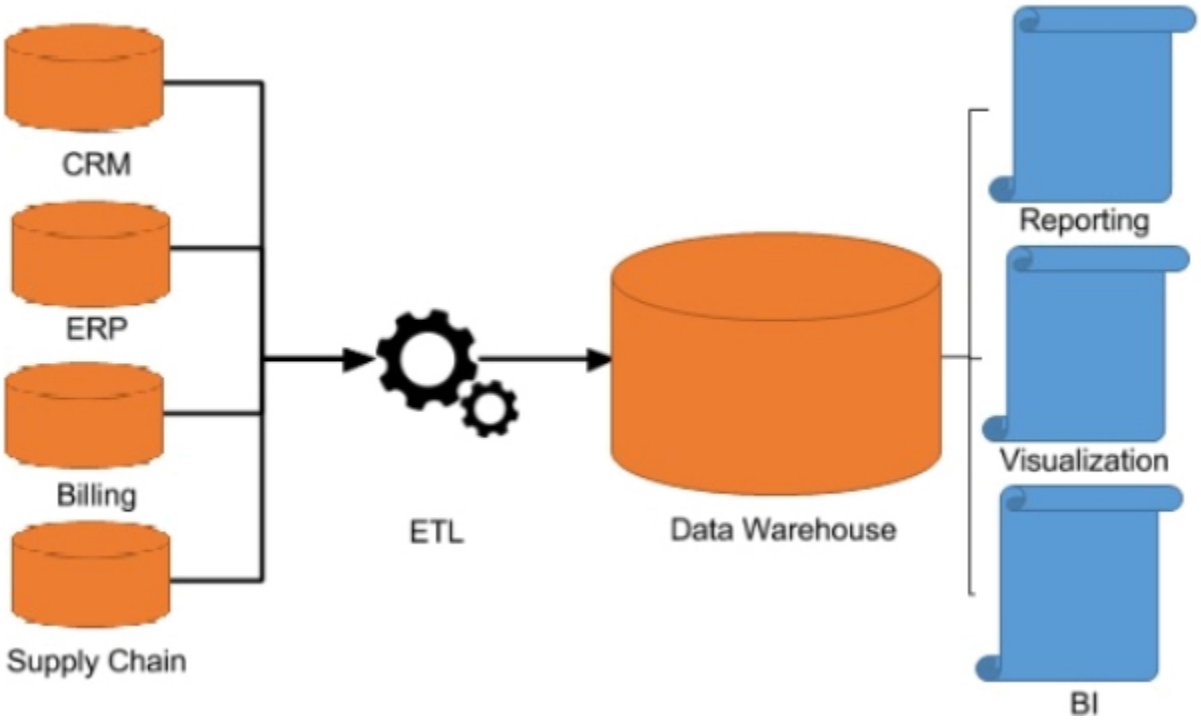
What is a data warehouse?

A data warehouse is a central repository of information that can be analyzed to make more informed **decisions**.

The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to **decision** support environments.

Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. An Extract, Transform, and Load [ETL] software tool is used to obtain data from each appropriate source.

Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytics applications.



How is a data warehouse architected?

A data warehouse architecture is made up of tiers.

The top tier is the front-end **client** that presents results through reporting, analysis, and data mining tools.

The middle tier consists of the **analytics engine** that is used to access and analyze the data.

The bottom tier of the architecture is the **database server**, where data is loaded and stored. Data is stored in two different types of ways: 1) data that is accessed frequently is stored in very **fast** storage (like SSD drives) and 2) data that is infrequently accessed is stored in a **cheap** object store, like Amazon S3. The data warehouse will automatically make sure that frequently accessed data is moved into the “fast” storage so query speed is optimized.

How do data warehouses, databases, and data lakes work together?

Typically, businesses use a combination of a database, a data lake, and a data warehouse to store and analyze data.

A data warehouse is specially designed for data analytics, which involves reading large amounts of data to understand relationships and trends across the data.

A database is used to capture and store data, such as recording details of a transaction.

Unlike a data warehouse, a data lake is a centralized repository for all data, including structured, semi-structured, and unstructured.

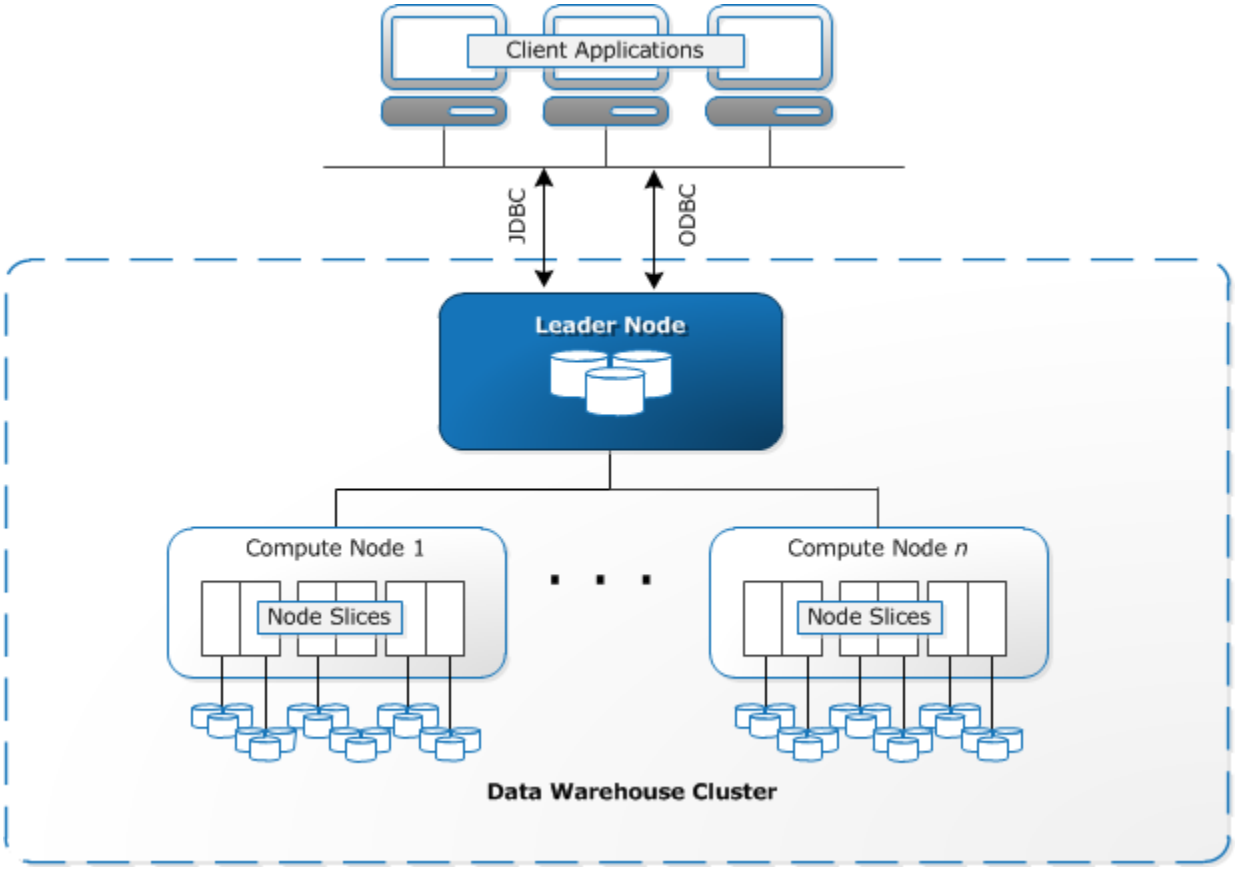
A data warehouse requires that the data be organized in a tabular format, which is where the schema comes into play. The tabular format is needed so that SQL can be used to query the data.

How does a data mart compare to a data warehouse?

A data mart is a data warehouse that serves the needs of a specific team or business unit, like finance, marketing, or sales. It is smaller, more focused, and may contain summaries of data that best serve its community of users. A data mart might be a portion of a data warehouse, too.

Case Studies

Amazon Redshift



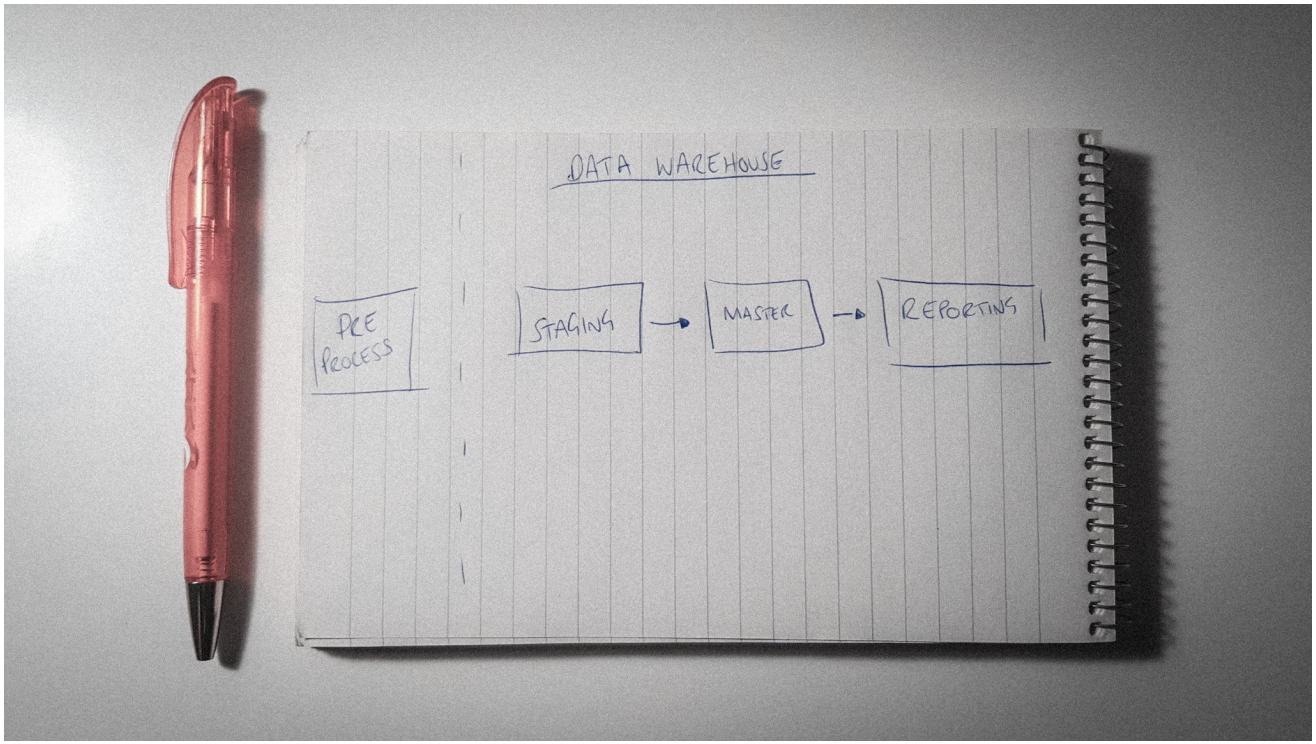
Client Applications: Amazon Redshift integrates with various data loading and ETL (extract, transform, and load) tools and business intelligence (BI) reporting, data mining, and analytics tools.

The core infrastructure component of an Amazon Redshift data warehouse is a **cluster**.

A cluster is composed of one or more **compute nodes**. If a cluster is provisioned with two or more compute nodes, an additional **leader node** coordinates the compute nodes and handles external communication. Your **client** application interacts directly only with the leader node. The compute nodes are transparent to external applications.

A compute node is partitioned into **slices**. Each slice is allocated a portion of the node's **memory** and **disk** space, where it processes a portion of the **workload** assigned to the node. The leader node manages distributing data to the slices and apportions the workload for any queries or other database operations to the slices. The slices then **work in parallel** to complete the operation.

How to architect the perfect Data Warehouse



https://lewisdgavin.medium.com/how-to-architect-the-perfect-data-warehouse-b3af2e01342e

References*

https://aws.amazon.com/data-warehouse/
https://en.wikipedia.org/wiki/Data_warehouse
https://databricks.com/glossary/data-warehouse
https://docs.aws.amazon.com/redshift/latest/dg/c_redshift_system_overview.html