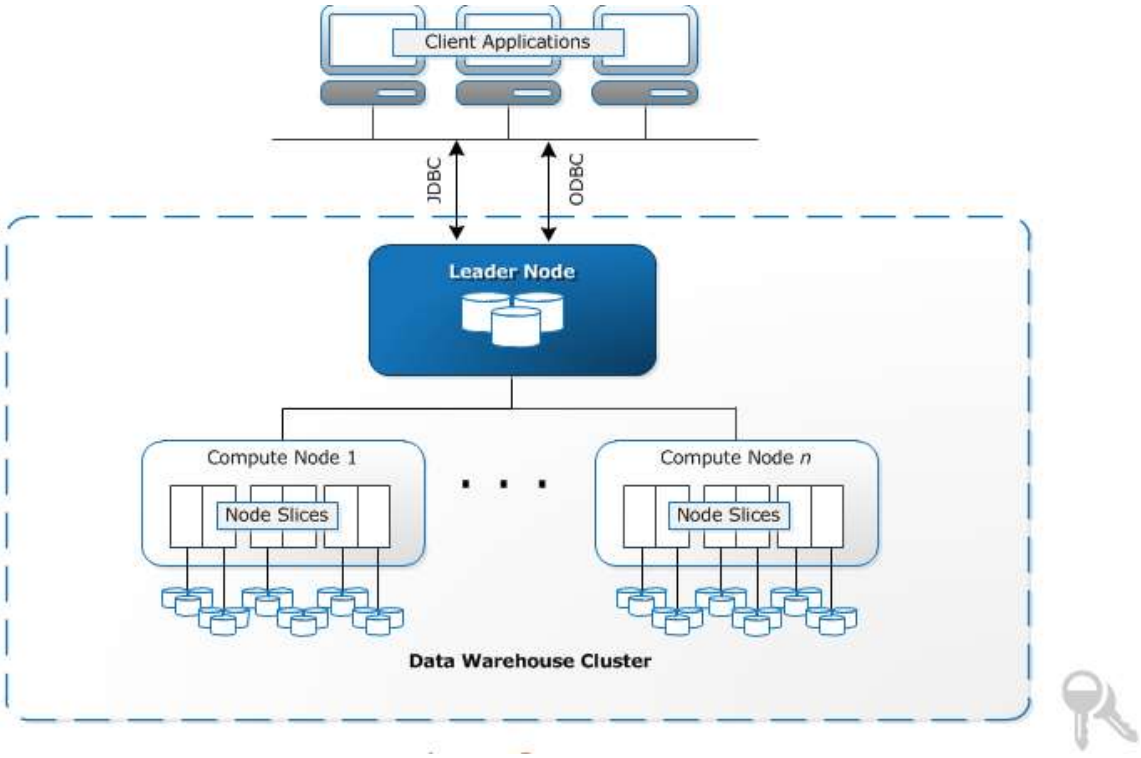
AWS RedShift

Data Warehousing is the process of storing and analyzing data from multiple sources to provide meaningful business insights. It involves transforming the data from multiple sources into a common format for both storage and analysis. In general, this process is known as **ETL** or **Extract, Load,** and **Transform**. Data Warehouses like Amazon Redshift leverage this process to cater to their customers.

Amazon Redshift is a fully managed data warehouse solution from AWS that allows you to store and query large volumes of analytical data.

It is fast and scalable providing a 10x performance lift over other data warehouses by using machine learning and parallel processing of queries against columnar storage held on very high performance disk.

Amazon Redshift Architecture



1. **Redshift Cluster:** Redshift uses a cluster of nodes as its core infrastructure component. A cluster usually has one leader node and a number of compute nodes. In cases where there is only one compute node, there is no additional leader node.

2. **Compute Nodes:** Each compute node has its own CPU, memory, and storage disk. Client applications are oblivious to the existence of compute nodes and never have to deal directly with compute nodes.
3. **Leader Node:** The leader node is responsible for all communications with client applications. The leader node also manages the coordination of compute nodes. Query parsing and execution plan development is also the responsibility of the leader node.

   On receiving a query the leader node creates the execution plan and assigns the compiled code to compute nodes. A portion of the data is assigned to each compute node. The final aggregation of the results is performed by the leader node.

Redshift allows the users to select from two types of nodes – **Dense Storage nodes** and **Dense Compute nodes.**

Customers can select them based on the nature of their requirements – whether it is storage heavy or compute-heavy. Redshift's cluster can be upgraded by increasing the number of nodes or upgrading individual node capacity or both.