# Data Warehousing on AWS

## Introducing Amazon Redshift

Amazon Redshift is a big scale data warehousing solution made to lower the cost and effort associated with deploying data warehouse systems. It is made to be able to handle petabytes of information with the help of BI (business intelligence) tools. Because of the advantages that this technology provides, it's currently one of the fastest growing AWS services on the market.

## Modern Analytics and Data Warehousing Architecture

Data usually moves to data warehouses from a variety of systems and databases in the form of structured, semi-structured, and unstructured data. Those who access this data generally do so through BI tools, SQL clients, and other tools.

Data warehouses specialize on high data amount and use denormalized schemas like the Snowflake schema and the Star schema, while OLTP databases focus on high transaction requirements and therefore using normalized schemas. AWS analytics services convert their data to answers. With these you can use intelligent tiering and Amazon Elastic Compute Cloud.

Analytics pipelines are made to handle large volumes of incoming data from databases, applications and devices. It gets divided into four stages: collect, store, process, and analyze and visualize the data.

## Data Warehouse Technology Options

Row-oriented databases store whole rows in a physical block. Examples of row-oriented databases are Oracle, MSSQL, MySQL, PostgreSQL. These systems are better for OLTP than for analytics. Column-oriented databases organize each column in it's own set of physical blocks. They tend to be more efficient for read-only queries. They are also better for data warehousing.

MPP (Massively Parallel Processing) architectures makes it so that you can increase performance of large scale data warehouses by using all the resources in the cluster. Basically, MPP data warehouses improve performance by adding more nodes.

## Amazon Redshift Deep Dive

Amazon Redshift is a data warehousing solution based on ANSI SQL meant to be cost-effective and efficient. This system lets the user make use of Redshift Spectrum to query and write data back to the data lake in an easier way by using open file formats. Some of the open file formats that you can use are Parquet, ORC, JSON, Avro, CSV and S3. It also uses high performing hardware, Advanced Query Accelerator, efficient storage, high-performance query processing, materialized views, auto workload management to maximize throughput and performance and result caching. Amazon Redshift detects and replaces any failed node in your data warehouse cluster and it also attempts to maintain at least three copies of data. At last, but not least, you can get elastic resize for quickly resizing your Amazon cluster and concurrency scaling to support unlimited concurrent users and concurrent queries.

## Operations

Amazon Redshift automates cluster performance and cost optimization. It also provides a feature called Amazon Redshift Advisor for recommendations about changes to make. Redshift supports custom JDBC and ODBC drivers to use a wide range of familiar SQL clients. Also, Amazon Redshift can be run inside the Amazon Virtual Private Cloud, which is a virtual private cloud; this data can only be accessed from the cluster's leader node, with it having multiple means of authentication. You can also restrict traffic based on the rules you configure, it supports ssl connections between client applications and your Redshift cluster, the data then is encrypted in transit.