Amazon Redshift

What is Redshift?

- Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. This enables you to use your data to acquire new insights for your business and customers. The first step to create a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster.
- Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call.
- Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period. Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures.

Benefits of Redshift

- **Query and export data to and from your data lake**: No other cloud data warehouse makes it as easy to both query data and write data back to your data lake in open formats. You can query open file formats such as Parquet, ORC, JSON, CSV, and more directly in S3 using familiar ANSI SQL. To export data to your data lake you simply use the Redshift UNLOAD command in your SQL code and specify Parquet as the file format and Redshift automatically takes care of data formatting and data movement into S3

Benefits of Redshift

- **RA3 instances**: RA3 instances deliver 3x the performance of any cloud data warehouse service. These Amazon Redshift instances maximize speed for performance-intensive workloads that require large amounts of compute capacity, with the flexibility to pay separately for compute independently of storage by specifying the number of instances you need.
- **Efficient storage and high performance query processing**: Amazon Redshift delivers fast query performance on datasets ranging in size from gigabytes to petabytes. Columnar storage, data compression, and zone maps reduce the amount of I/O needed to perform queries.

Benefits of Redshift

- **Machine learning to maximize throughput and performance**: Advanced machine learning capabilities in Amazon Redshift deliver high throughput and performance, even with varying workloads or concurrent user activity. Amazon Redshift utilizes sophisticated algorithms to predict and classify incoming queries based on their run times and resource requirements to dynamically manage performance and concurrency while also helping you to prioritize your business critical workloads.
- **Result caching**: Amazon Redshift uses result caching to deliver sub-second response times for repeat queries. Dashboard, visualization, and business intelligence tools that execute repeat queries experience a significant performance boost. When a query executes, Amazon Redshift searches the cache to see if there is a cached result from a prior run. If a cached result is found and the data has not changed, the cached result is returned immediately instead of re-running the query.

Benefits of Redshift

- **Petabyte-scale data warehousing**: Amazon Redshift is simple and quickly scales as your needs change. With a few clicks in the console or a simple API call, you can easily change the number or type of nodes in your data warehouse, and scale up or down as your needs change. With managed storage, capacity is added automatically to support workloads up to 8PB of compressed data.
- **Petabyte-scale data lake analytics**: You can run queries against petabytes of data in Amazon S3 without having to load or transform any data with the Redshift Spectrum feature. You can use S3 as a highly available, secure, and cost-effective data lake to store unlimited data in open data formats. Amazon Redshift Spectrum executes queries across thousands of parallelized nodes to deliver fast results, regardless of the complexity of the query or the amount of data.

Workload Management and Spectrum

- **Amazon Redshift workload management (WLM)** enables users to flexibly manage priorities within workloads so that short, fast-running queries won't get stuck in queues behind long-running queries.
- **Amazon Redshift Spectrum** is a feature within Amazon Web Services' Redshift data warehousing service that lets a data analyst conduct fast, complex analysis on objects stored on the AWS cloud. With Redshift Spectrum, an analyst can perform SQL queries on data stored in Amazon S3 buckets

Reserved Nodes

- Reserved node pricing is less expensive than on-demand pricing because compute nodes are billed at discounted hourly rates. If you intend to keep your Amazon Redshift cluster running continuously for a prolonged period, you should consider purchasing reserved node offerings.
- No Upfront – You pay nothing upfront, and commit to pay monthly over the course of one year.
- Partial Upfront – You pay a portion of the Reserved Instance upfront, and the remainder over a one- or three-year term.
- All Upfront – You pay for the entire Reserved Instance term (one or three years) with one upfront payment.

Automated Backups

- Amazon Redshift automatically takes incremental snapshots that track changes to the cluster since the previous automated snapshot. Automated snapshots retain all of the data required to restore a cluster from a snapshot. You can create a snapshot schedule to control when automated snapshots are taken, or you can take a manual snapshot any time.
- When you restore from a snapshot, Amazon Redshift creates a new cluster and makes the new cluster available before all of the data is loaded, so you can begin querying the new cluster immediately.
- To ensure that your backups are always available to your cluster, Amazon Redshift stores snapshots in an internally managed Amazon S3 bucket that is managed by Amazon Redshift.

Encryption

- Amazon Redshift uses a hierarchy of encryption keys to encrypt the database. You can use either AWS Key Management Service (AWS KMS) or a hardware security module (HSM) to manage the top-level encryption keys in this hierarchy.
- The process that Amazon Redshift uses for encryption differs depending on how you manage keys. Amazon Redshift automatically integrates with AWS KMS but not with an HSM.
- When you use an HSM, you must use client and server certificates to configure a trusted connection between Amazon Redshift and your HSM.

CloudWatch Alarms

- Using Amazon CloudWatch alarms, you watch a single metric over a time period that you specify. If the metric exceeds a given threshold, a notification is sent to an Amazon SNS topic or AWS Auto Scaling policy.
- CloudWatch alarms do not invoke actions because they are in a particular state. Rather the state must have changed and been maintained for a specified number of periods.
- Using CloudWatch metrics for Amazon Redshift, you can get information about your cluster's health and performance and see information at the node level. When working with these metrics, keep in mind that each metric has one or more dimensions associated with it.

Redshift Pricing Models

- **On-demand pricing**: Amazon Redshift on-demand pricing allows you to pay for capacity by the hour with no commitments and no upfront costs, you simply pay an hourly rate based on the type and number of nodes in your cluster. Partial hours are billed in one-second increments following a billable status change such as creating, deleting, pausing or resuming the cluster. The pause and resume feature allows you to suspend on-demand billing during the time the cluster is paused. During the time that a cluster is paused you only pay for backup storage. This frees you from planning and purchasing data warehouse capacity ahead of your needs, and enables you to cost-effectively manage environments for development or test purposes.

Redshift Pricing Models

- **Amazon Redshift Spectrum Pricing**: It enables you to run SQL queries directly against the data in your S3 data lake, out to exabytes - you simply pay for the number of bytes scanned.
- **Concurrency Scaling Pricing**: Each cluster earns up to one hour of free concurrency scaling credits per day, which is sufficient for 97% of customers. This enables you to provide consistently fast performance, even with thousands of concurrent queries and users. You simply pay a per-second on-demand rate for usage that exceeds the free credits.
- **Redshift managed storage pricing**: It gives you the flexibility to pay only for the data that you store in RA3 clusters independent of number of compute nodes provisioned. You simply pay hourly for the total amount of data in managed storage.