Redshift

* + Overview
    - Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and existing Business Intelligence(BI) tools
    - Clustered peta-byte scale data warehouse
    - RS is a SQL based data warehouse used for **analytics** applications
    - Redshift is an Online Analytics Processing(OLAP) type of DB
    - RS is used for running complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution
    - RedShift is ideal for **processing** large amounts of data for BI
    - Extremely cost-effective as compared to some other on-premises data warehouse platforms
    - PostgreSQL compatible with JDBC and ODBC drivers available; compatible with most BI tools out of the box
    - Features parallel processing and columnar data stores which are optimized for complex queries
    - **Option to query directly from data files on S3 via RedShift Spectrum**
    - Redshift is 10x faster than a traditional SQL DB
    - RS can store huge amounts of data but cannot ingest huge amounts of data in real time
    - RedShift uses columnar data storage:
      * Data is stored sequentially in columns instead of rows
      * Columnar based DB is ideal for data warehousing and analytics
      * Requires fewer I/Os which greatly enhances performance
    - RedShift provides advanced compression:
      * Data is stored sequentially in columns which allows for much better performance and less storage space
      * RS automatically selects the compression scheme
    - RS provides good query performance and compression
    - RS provides Massively Parallel Processing(MPP) by distributing data and queries across all nodes
    - RS uses EC2 instances so you need to choose your instance type/size for scaling compute vertically, but you can also scale horizontally by adding more nodes to the cluster
    - You cannot have direct access to your RS cluster nodes as a user, but you can through applications
    - HDD and SDD storage options
    - The size of a single node is 160GB and clusters can be created up to a petabyte or more
    - Multi-node consists of:
      * Leader node:
        + Manages client connections and receives queries
        + Simple SQL end-point
        + Stores metadata
        + Optimizes query plan
        + Coordinates query execution
      * Compute nodes:
        + Stores data and performs queries and computations
        + Local columnar storage
        + Parallel/distributed execution of all queries, loads, backups, restores and resizes
        + Up to 128 compute nodes
      * **Amazon RS Spectrum is a fewature of RS that enables you to run queries against exabyts of unstructured data in S3, with no loading or ETL required**
    - Availability and Durability
      * RS uses replication and continuous backups to enhance availability and improve durability and can automatically recover from component and node failures
      * **Only available in one AZ but you can restore snapshots into another AZ**
      * Alternatively, you can run data warehouse clusters in multiple AZ's by loading data into two RS dat warehouse clusters in separate AZs from the same set of S3 input files
      * RS replicates your data within your data warehouse cluster and continuously backs up your data to S3
      * RS always keeps three copies of your data:
        + The original
        + A replica on compute nodes(within a cluster)
        + A backup copy on S3
      * RS provides continuous/incremental backups
        + Multiple copies within a cluster
        + Continuous and incremental backups to S3
        + Continuous and incremental backups across regions
        + Streaming restore
      * RS provides fault tolerance for the following failures:
        + Disk failures
        + Node failures
        + Network failures
        + AZ/region level disasters
      * For node failures the data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB
      * HA for RS
        + Currently, RS does not support Multi-AZ deployments
        + The best HA option is to use multi-node cluster which supports data replication and node recovery
        + A single node RS cluster does not support data replication and you'll have to restore from a snapshot on S3 if a drive fails
      * RedShift can asynchronously replicate your snapshots to S3 in another region for DR
      * Single-node clusters do not support data replication(in a failure scenario you would need to restore from a snapshot)
      * Scaling requires a period of unavailability of a few minutes(typically during the maintenance window)
      * During scaling operations RS moves data in parallel from the compute nodes in your existing data warehouse cluster to the compute nodes in your new cluster
      * By default, Redshift retains backups for 1 day. You can configure this to be as long as 35 days
      * If you delete the cluster you can choose to have a final snapshot taken and retained
      * Manual backups are not automatically deleted when you delte a cluster
    - Security
      * You can load encrypted data from S3
      * Supports SSL Encryption in-transit between client applications and Reshift data warehouse cluster
      * VPC for network isolation  
        Encryption for dat at rest(AES 256)
      * Audit logging and AWS CT integration
      * RS takes care of key management or you can manage your own through HSM or KMS
    - Charges
      * Charged for ocmpute nodes hours, 1 unit per hour(only compute node, not leader node)
      * Backup storage - storage on S3
      * Data transfer - no charge for data transfer between RS and S3 within a region but for other scenarios you may pay charges