Amazon Redshift

PDF | RSS

Amazon Redshift is a fast, fully-managed, petabyte-scale data warehouse service that makes it simple and cost-effective to analyze all your data efficiently using your existing business intelligence tools. It is optimized for data sets ranging from a few hundred gigabytes to a petabyte or more, and is designed to cost less than a tenth of the cost of most traditional data warehousing solutions.

Amazon Redshift delivers fast query and I/O performance for virtually any size dataset by using columnar storage technology while parallelizing and distributing queries across multiple nodes. It automates most of the common administrative tasks associated with provisioning, configuring, monitoring, backing up, and securing a data warehouse, making it easy and inexpensive to manage and maintain. This automation enables you to build petabyte-scale data warehouses in minutes instead of weeks or months.

Amazon Redshift Spectrum enables you to run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in S3, generates a plan to minimize the amount of S3 data that needs to be read, and then requests Redshift Spectrum workers out of a shared resource pool to read and process the data from S3.

With the federated query feature in Amazon Redshift, you can query and analyze data across operational databases, data warehouses, and data lakes. It also enables you to integrate queries from Amazon Redshift on live data in external databases with queries across your Amazon Redshift and S3 environments. With cross-database queries, you can query data from any database in the Amazon Redshift cluster, regardless of which database you are connected to. Cross-database queries eliminate data copies and simplify your data organization to support multiple business groups from the same data warehouse.

AQUA (Advanced Query Accelerator) is a new distributed and hardware-accelerated cache that enables Amazon Redshift to run up to ten times faster than other enterprise cloud data warehouses by automatically boosting certain types of queries. AQUA is included with certain node types in the Amazon Redshift RA3 cluster.

There is also a feature called **Data sharing** which enables you to share data across Amazon Redshift clusters without needing to manually copy or move it for read purposes. You can have live access to data, and users can see the most up-to-date and consistent information as it's updated in the Amazon Redshift cluster.

Ideal usage patterns

Amazon Redshift is ideal for online analytical processing (OLAP) using your existing business intelligence tools. Organizations are using Amazon Redshift to:

- Analyze global sales data for multiple products
- Store historical stock trade data
- Analyze ad impressions and clicks
- Aggregate gaming data
- Analyze social trends
- Measure clinical quality, operation efficiency, and financial performance in health care
- Analyze data across the data lake (S3) and Amazon Redshift.

Cost model

An Amazon Redshift data warehouse cluster requires no long-term commitments or upfront costs. This frees you from the capital expense and complexity of planning and purchasing data warehouse capacity ahead of your needs. Charges are based on the size and number of nodes of your cluster.

There is no additional charge for backup storage up to 100% of your provisioned storage. For example, if you have an active cluster with 2 XL nodes for a total of 4 TB of storage, AWS provides up to 4 TB of backup storage on Amazon S3 at no additional charge. Backup storage beyond the provisioned storage size, and backups stored after your cluster is terminated, are billed at standard Amazon S3 rates. There is no data transfer charge for communication between S3 and Amazon Redshift.

For more information, see Amazon Redshift pricing.

Performance

Amazon Redshift uses a variety of innovations to obtain very high performance on data sets ranging in size from hundreds of gigabytes to a petabyte or more. It uses columnar storage, data compression, and zone maps to reduce the amount of I/O needed to perform queries.

Amazon Redshift has a massively parallel processing (MPP) architecture, parallelizing and distributing SQL operations to take advantage of all available resources. The underlying hardware is designed for high performance data processing, using local attached storage to maximize throughput between the CPUs and drives, and a 10 GigE mesh network to maximize throughput between nodes. Performance can be tuned based on your data warehousing needs: AWS offers Dense Compute (DC) with SSD drives as well as Dense Storage (DS) options.

Durability and availability

Amazon Redshift automatically detects and replaces a failed node in your data warehouse cluster. The data warehouse cluster is read-only until a replacement node is provisioned and added to the DB, which typically only takes a few minutes. Amazon Redshift makes your replacement node available immediately and streams your most frequently accessed data from S3 first, to allow you to resume querying your data as quickly as possible.

Additionally, your data warehouse cluster remains available in the event of a drive failure; because Amazon Redshift mirrors your data across the cluster, it uses the data from another node to rebuild failed drives. Amazon Redshift clusters reside within one Availability Zone, but if you wish to have a multi-AZ set up for Amazon Redshift, you can set up a mirror and then self-manage replication and failover.

Scalability and elasticity

With a few clicks in the console or an API call, you can change the number, or type, of nodes in your data warehouse as your performance or capacity needs change. Amazon Redshift enables you to start with a single 160 GB node and scale up to a petabyte or more of compressed user data using many nodes. For more information, see Clusters and Nodes in Amazon Redshift in the *Amazon Redshift Management Guide*.

While resizing, Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel. During this process, you pay only for the active Amazon Redshift cluster. You can continue running queries against your old cluster while the new one is being provisioned. After your data has been copied to your new cluster, Amazon Redshift automatically redirects queries to your new cluster and removes the old cluster.

Interfaces

Amazon Redshift has custom JDBC and ODBC drivers that you can download from the **Connect Client** tab of the console, which enables you to use a wide range of familiar SQL clients. You can also use standard PostgreSQL JDBC and ODBC drivers. For more information about Amazon Redshift drivers, see Amazon Redshift and PostgreSQL.

There are numerous examples of validated integrations with many popular BI and ETL vendors. Loads and unloads are attempted in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster as well as to and from S3 and DynamoDB. You can easily load streaming data into Amazon Redshift using Amazon Kinesis Data Firehose, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. Metrics for compute utilization, memory utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the console or CloudWatch API operations.

Anti-patterns

Amazon Redshift has the following anti-patterns:

- **Small datasets** – Amazon Redshift is built for parallel processing across a cluster. If your data set is less than a hundred gigabytes, you won't get all the benefits that Amazon Redshift has to offer, and Amazon RDS may be a better solution.
- **Online transaction processing (OLTP)** – Amazon Redshift is designed for data warehouse workloads producing extremely fast and inexpensive analytic capabilities. If you require a fast transactional system, you may want to choose a traditional relational database system built on Amazon RDS or a NoSQL database offering, such as DynamoDB.
- **Unstructured data** – Data in Amazon Redshift must be structured by a defined schema, rather than supporting arbitrary schema structure for each row. If your data is unstructured, you can perform extract, transform, and load (ETL) on Amazon EMR to get the data ready for loading into Amazon Redshift.
- **BLOB data** – If you plan on storing large binary files (such as digital video, images, or music), you may want to consider storing the data in S3 and referencing its location in Amazon Redshift. In this scenario, Amazon Redshift keeps track of metadata (such as item name, size, date created, owner, location, and so on) about your binary objects, but the large objects themselves are stored in S3.

Next topic: Amazon OpenSearch Service
Previous topic: Amazon DynamoDB