Data Engineering 101

Amazon Redshift

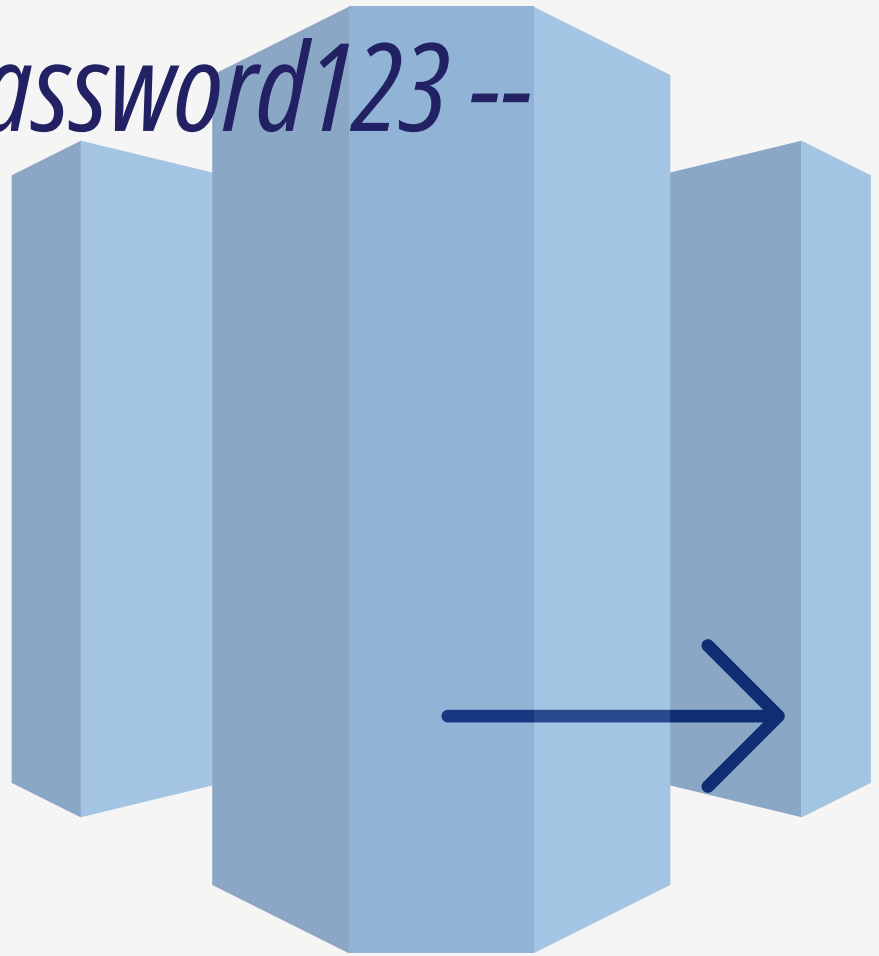
Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Redshift

Cluster

An Amazon Redshift cluster is a set of nodes that work together to store and process data. Each cluster contains one or more databases.

Creating a cluster:
aws redshift create-cluster --cluster-identifier my-cluster --node-type dc2.large --master-username admin --master-user-password Password123 --number-of-nodes 2



Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Redshift

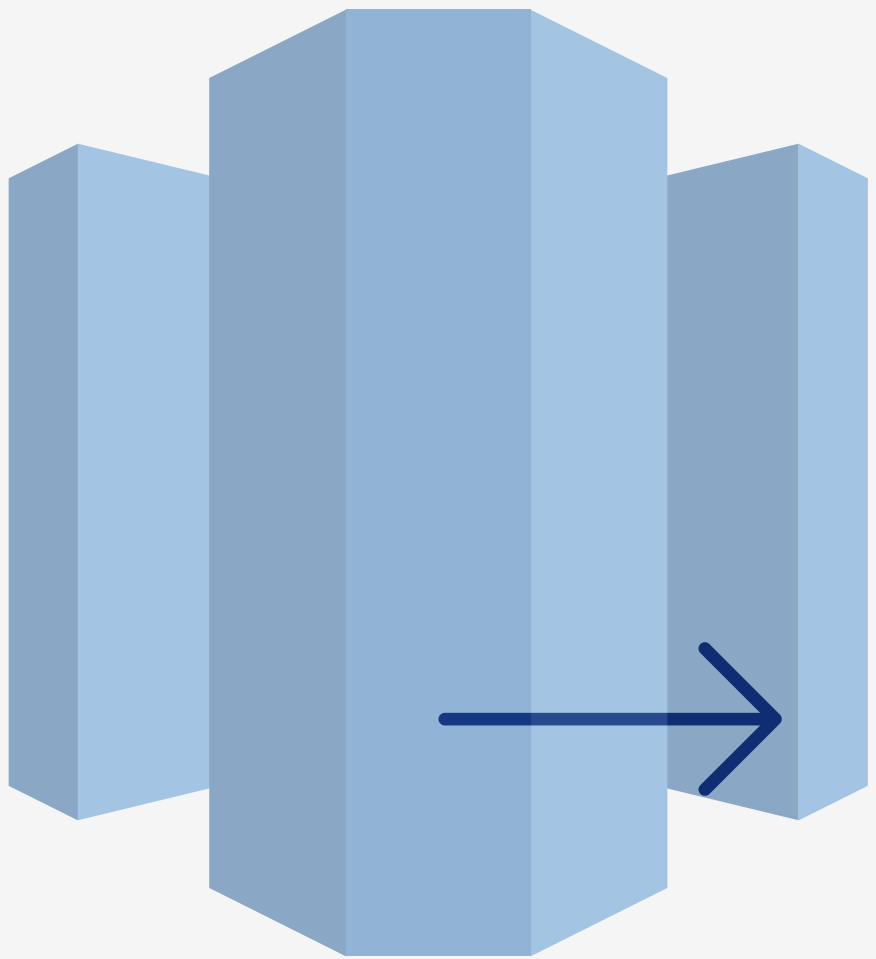
Node Types

2

Amazon Redshift offers different node types optimized for different workloads, including Dense Compute (DC) and Dense Storage (DS).

DC2 instances are ideal for performance-intensive workloads, while DS2 instances are optimized for large storage needs.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

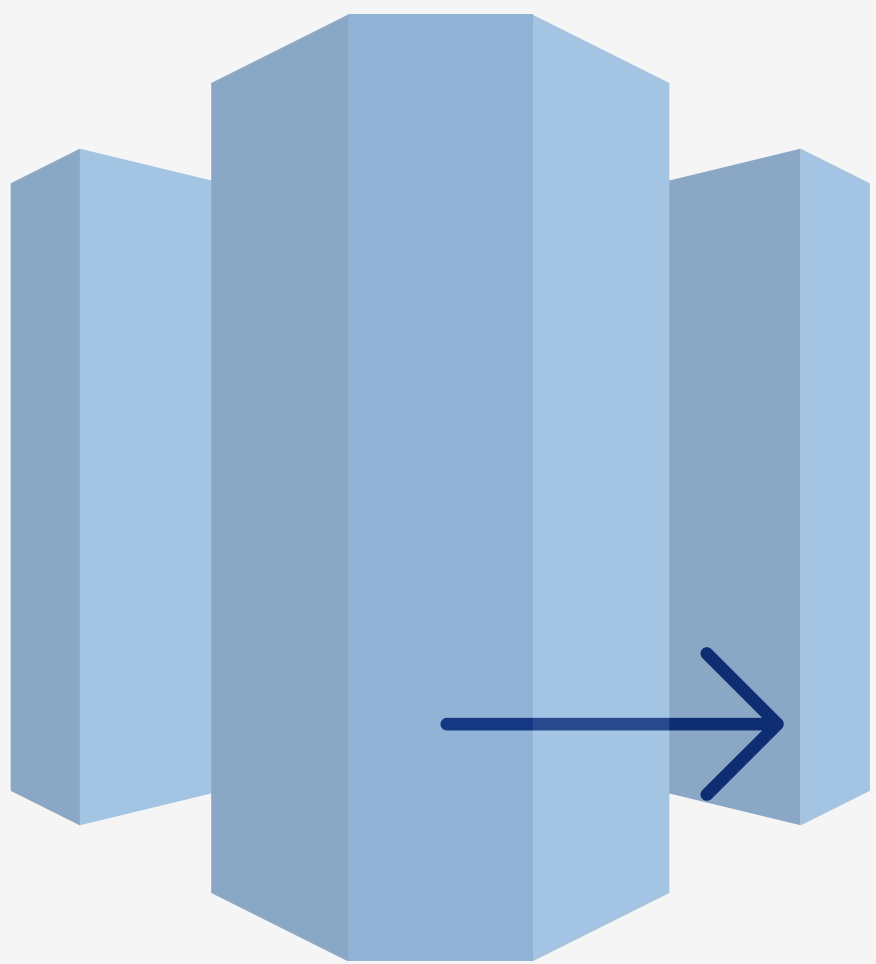
Leader Node

3

The leader node manages communications with client applications and all nodes in the cluster, receiving queries and distributing them to the compute nodes.

The leader node coordinates query processing and aggregation of results before sending them to the client.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

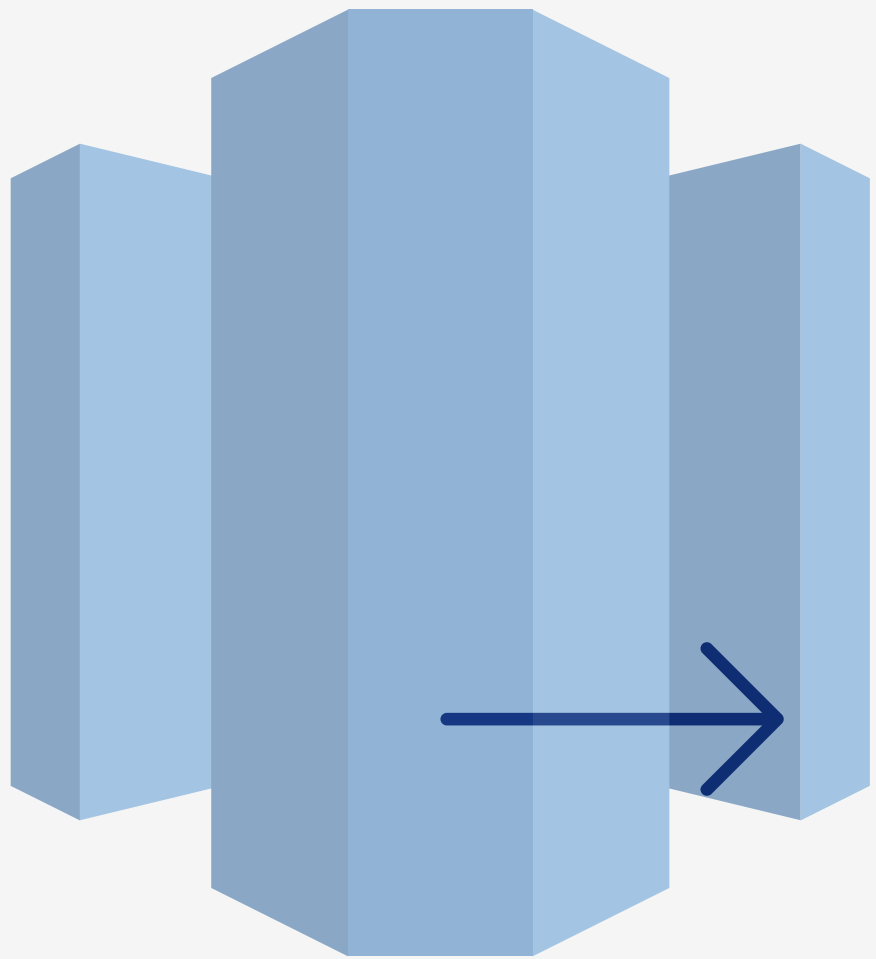
Compute Node

4

Compute nodes execute the queries and store data. They send intermediate results back to the leader node for aggregation.

Compute nodes store table data and perform query processing.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

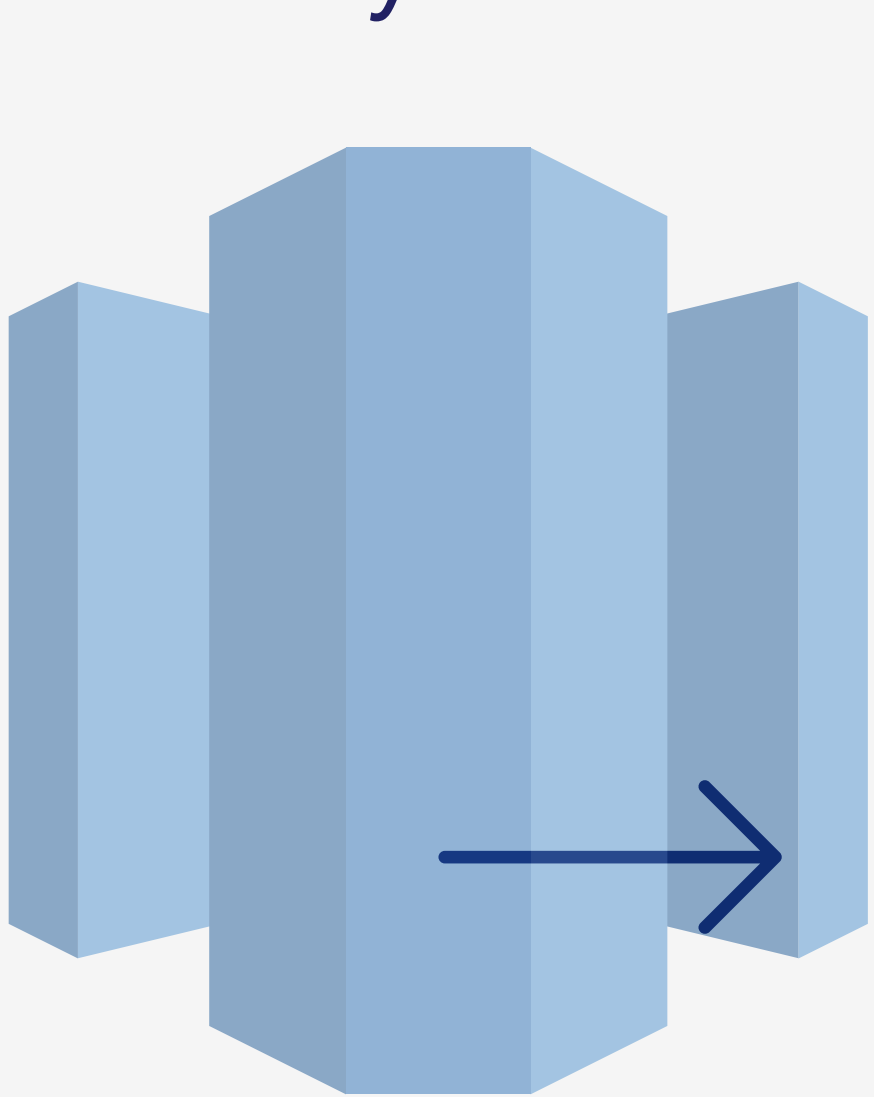
Columnar Storage

5

Amazon Redshift stores data in a columnar format, which allows for more efficient data compression and query performance, especially for read-intensive operations.

Columnar storage reduces I/O and speeds up query performance, as only the columns needed by a query are scanned.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

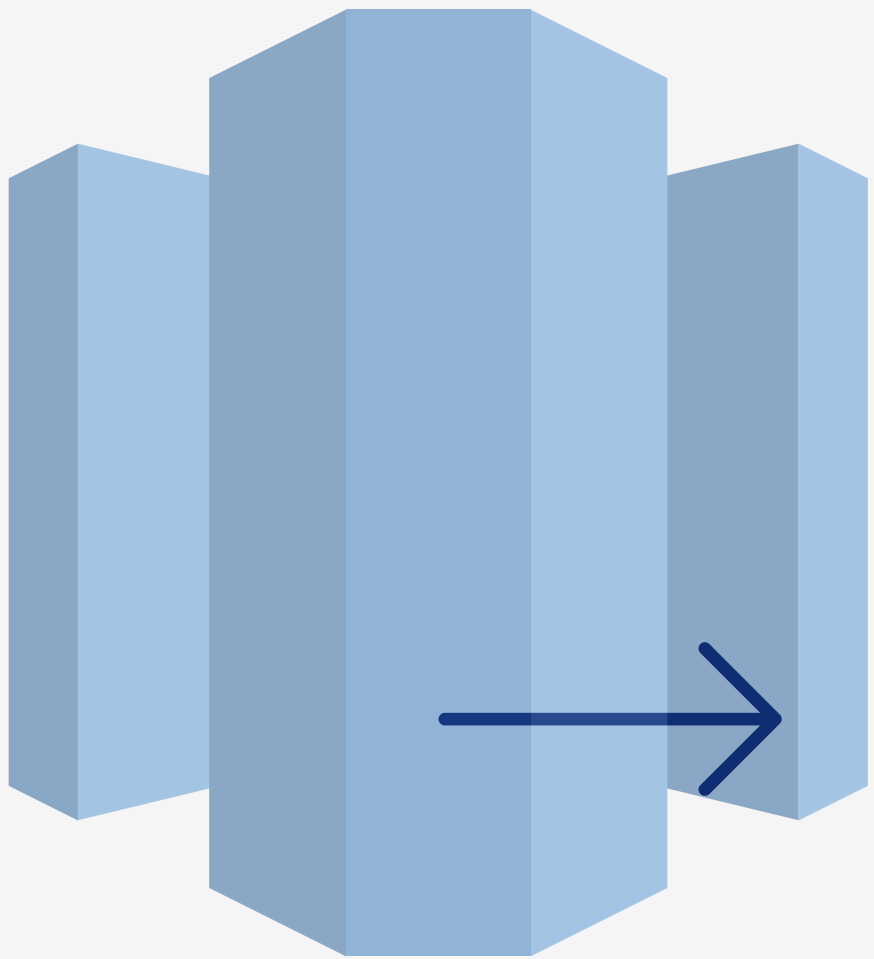
Sort Keys

6

Sort keys determine the order in which data is physically stored in Amazon Redshift tables, optimizing query performance by reducing the amount of data scanned.

Define a sort key: CREATE TABLE sales (id INT, date DATE, amount FLOAT) SORTKEY (date);

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

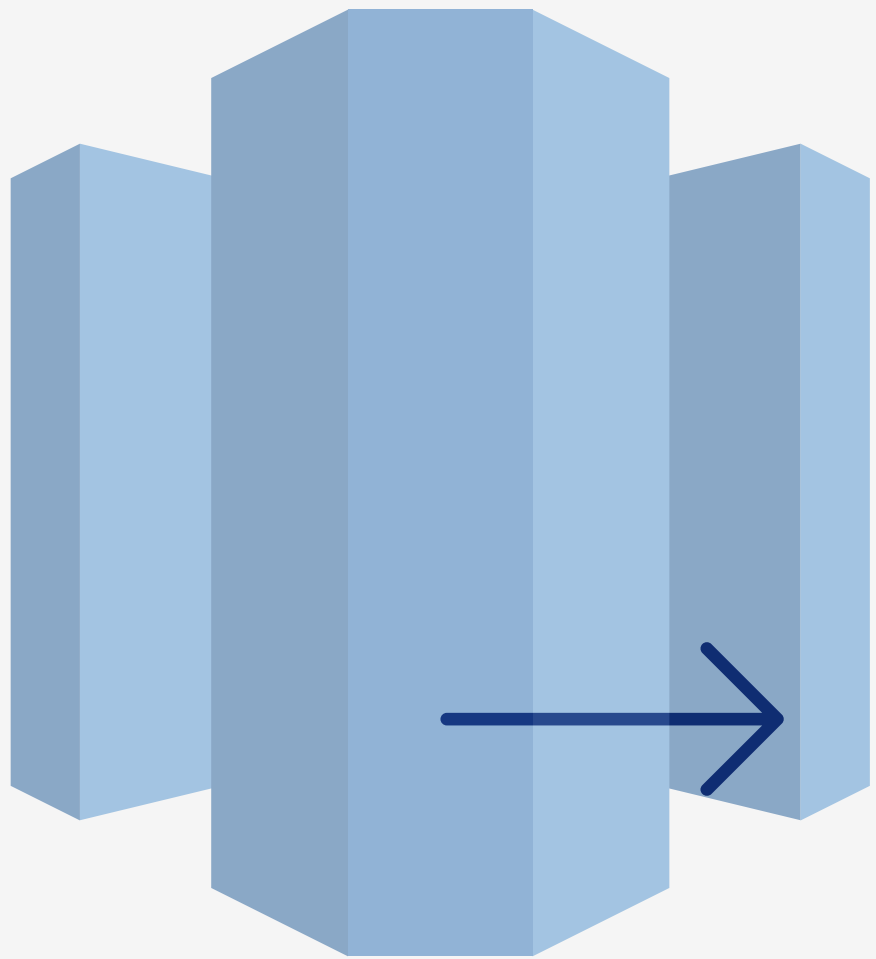
Distribution Keys

1

Distribution keys determine how data is distributed across the compute nodes. Proper selection of distribution keys can minimize data movement and optimize performance.

Define a distribution key: CREATE TABLE sales (id INT, date DATE, amount FLOAT) DISTKEY (id);

Shwetank Singh
GritSetGrow - GSGLearn.com



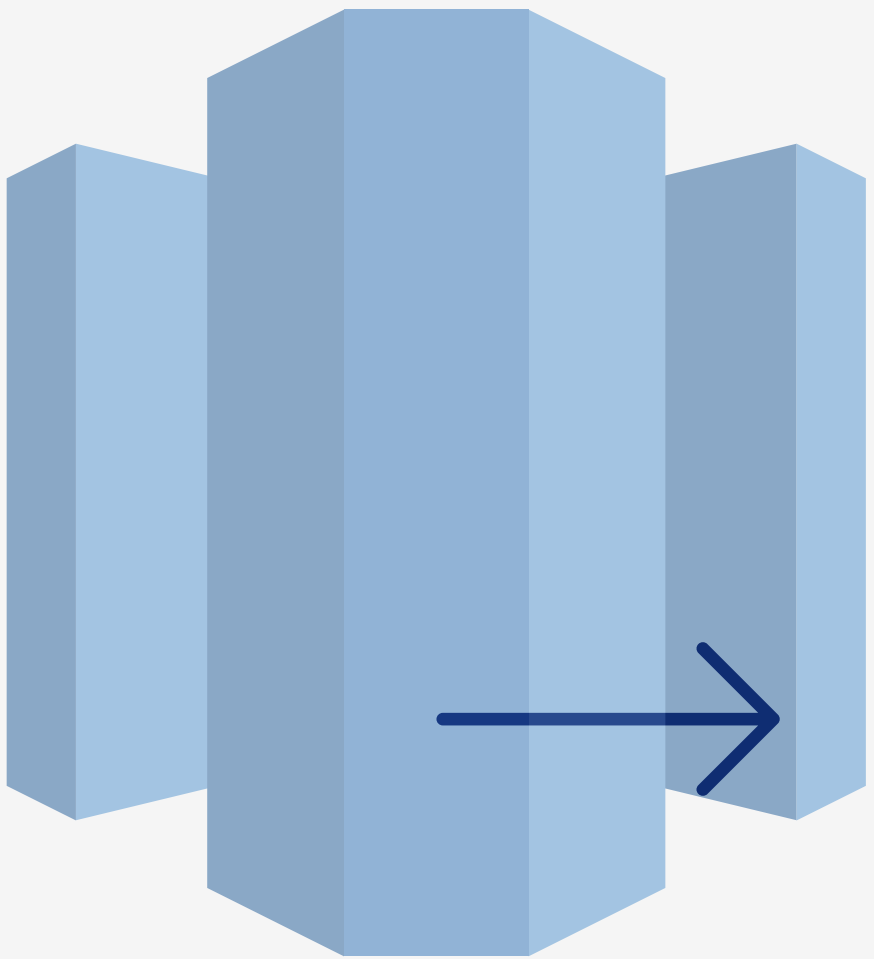
Data Engineering 101: Redshift

Compression

Amazon Redshift automatically compresses data to save storage and improve query performance. Compression types include LZO, Zstandard, and Delta.

COPY sales FROM 's3://bucket-name/sales_data.csv' COMPUPDATE ON;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

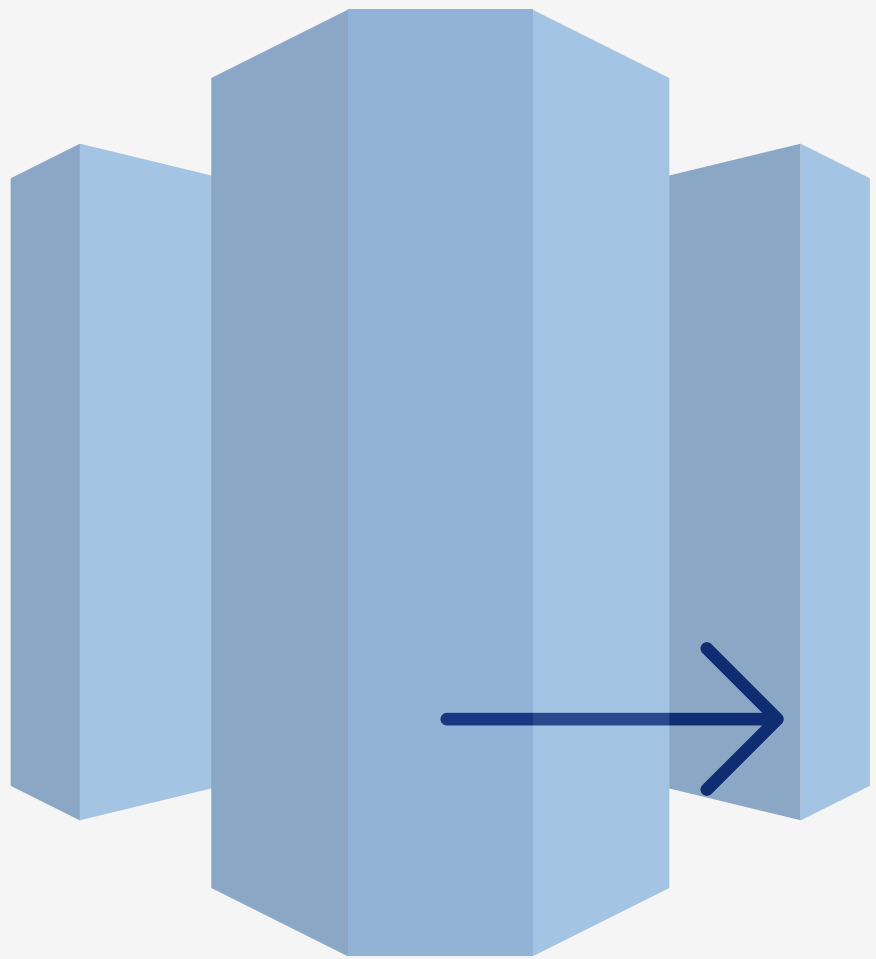
Vacuum

9

The VACUUM command reclaims space and sorts tables to optimize performance after large DELETE or UPDATE operations.

VACUUM FULL sales;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

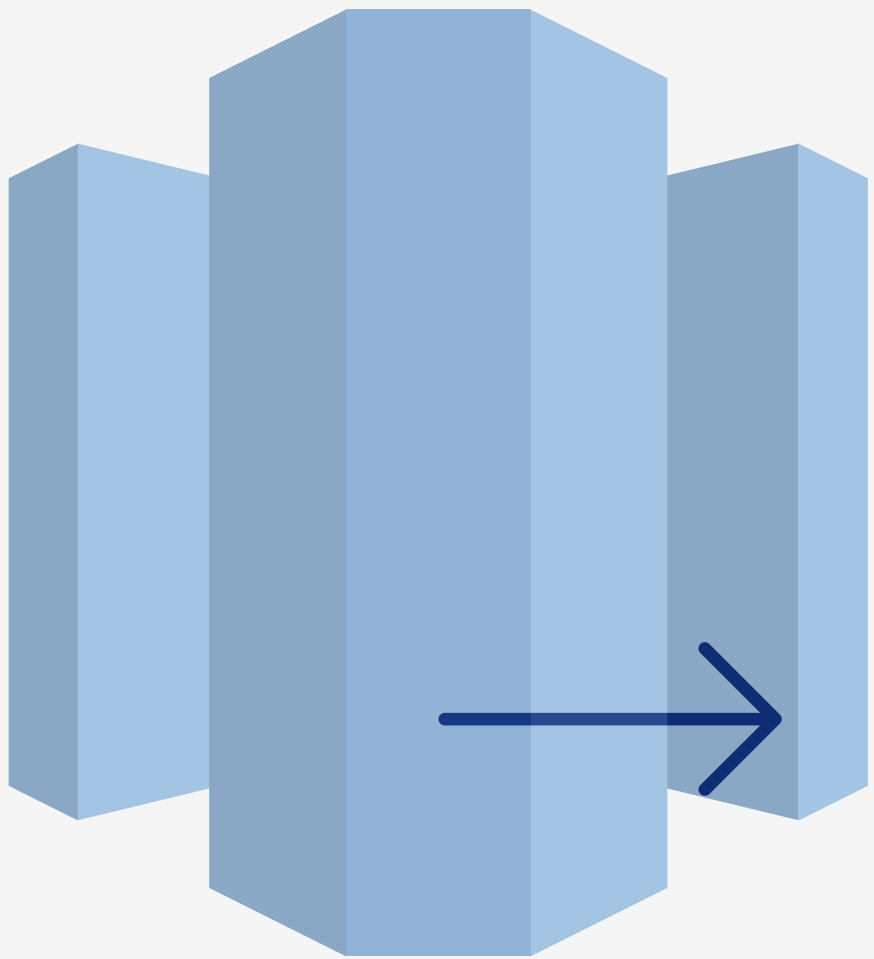
10

Analyze

The ANALYZE command updates table statistics to help the query planner create optimal execution plans.

ANALYZE sales;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

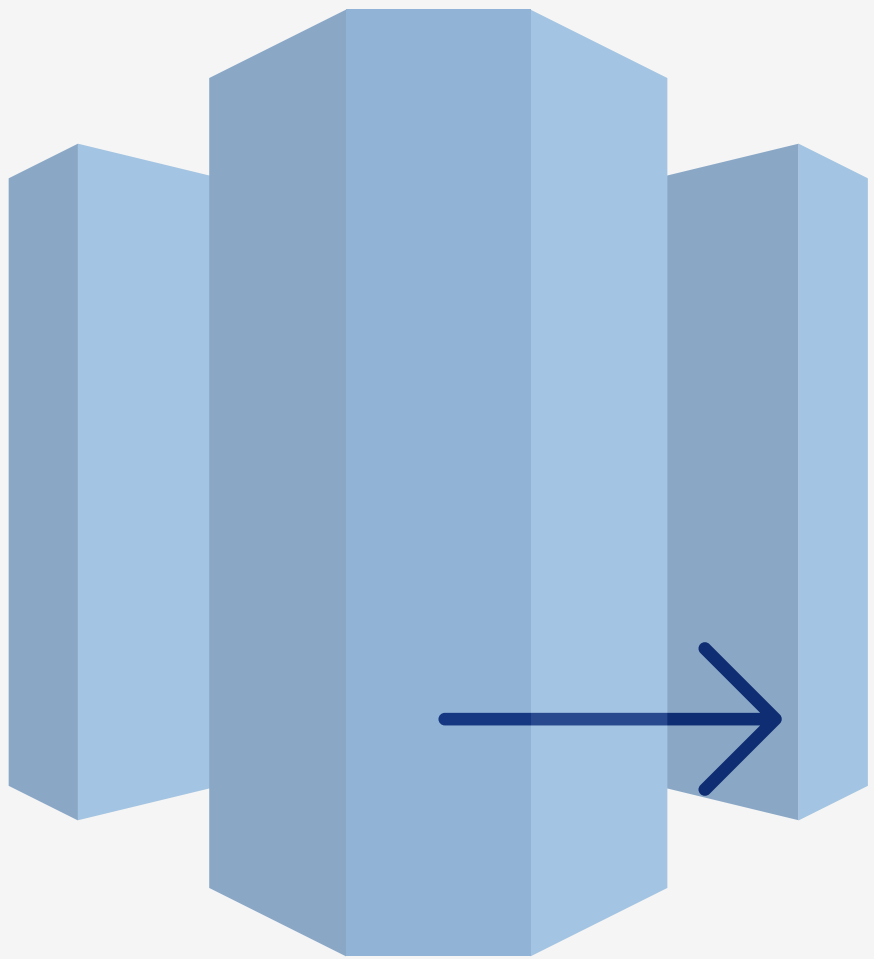
11

Materialized Views

Materialized views store the results of a query physically, allowing for faster retrieval in subsequent queries.

CREATE MATERIALIZED VIEW mv_sales AS SELECT date, SUM(amount) FROM sales GROUP BY date;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

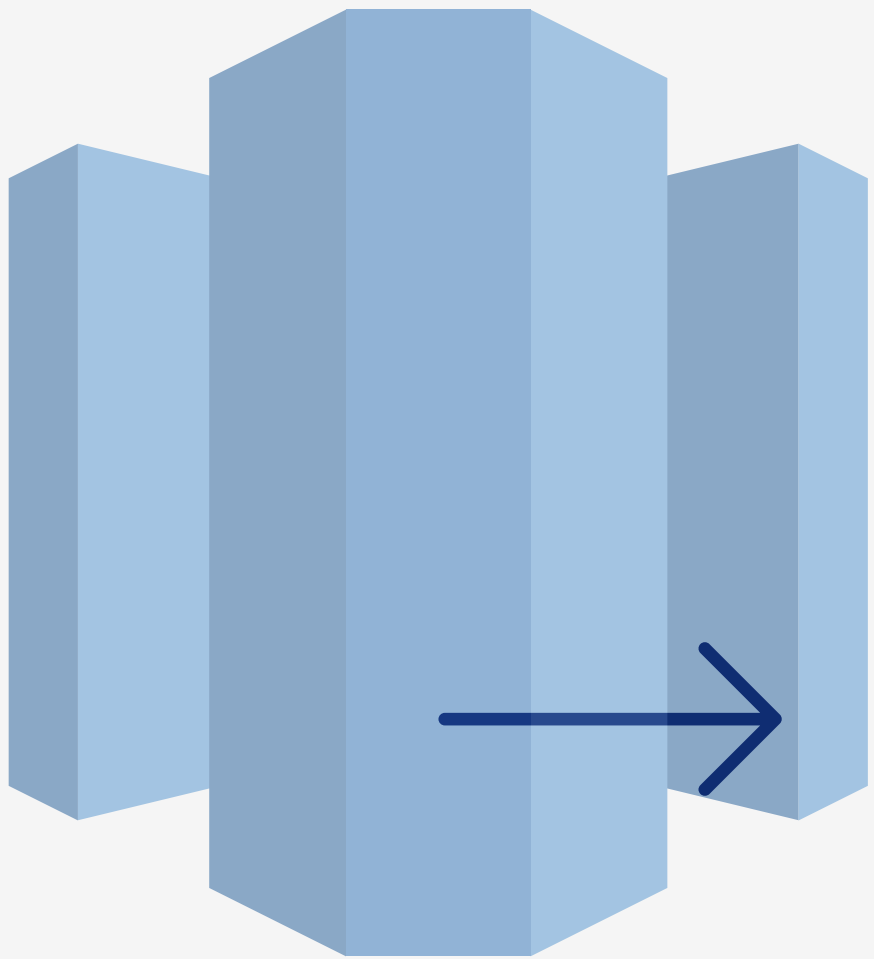
17

Snapshots

Redshift snapshots capture the current state of your data, which can be used for backup or recovery. Snapshots can be manual or automatic.

CREATE SNAPSHOT my_snapshot FROM my-cluster;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

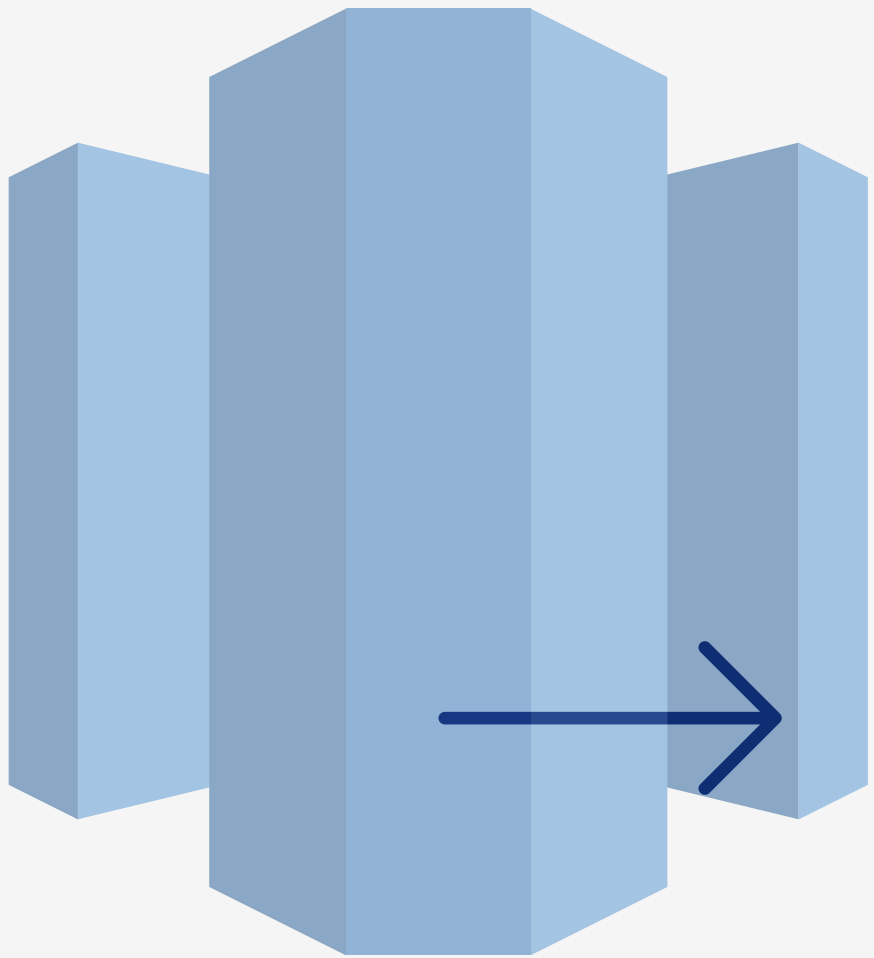
13

Backup and Restore

Amazon Redshift automatically takes incremental snapshots and allows users to manually create and restore from these snapshots.

RESTORE FROM SNAPSHOT my_snapshot;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

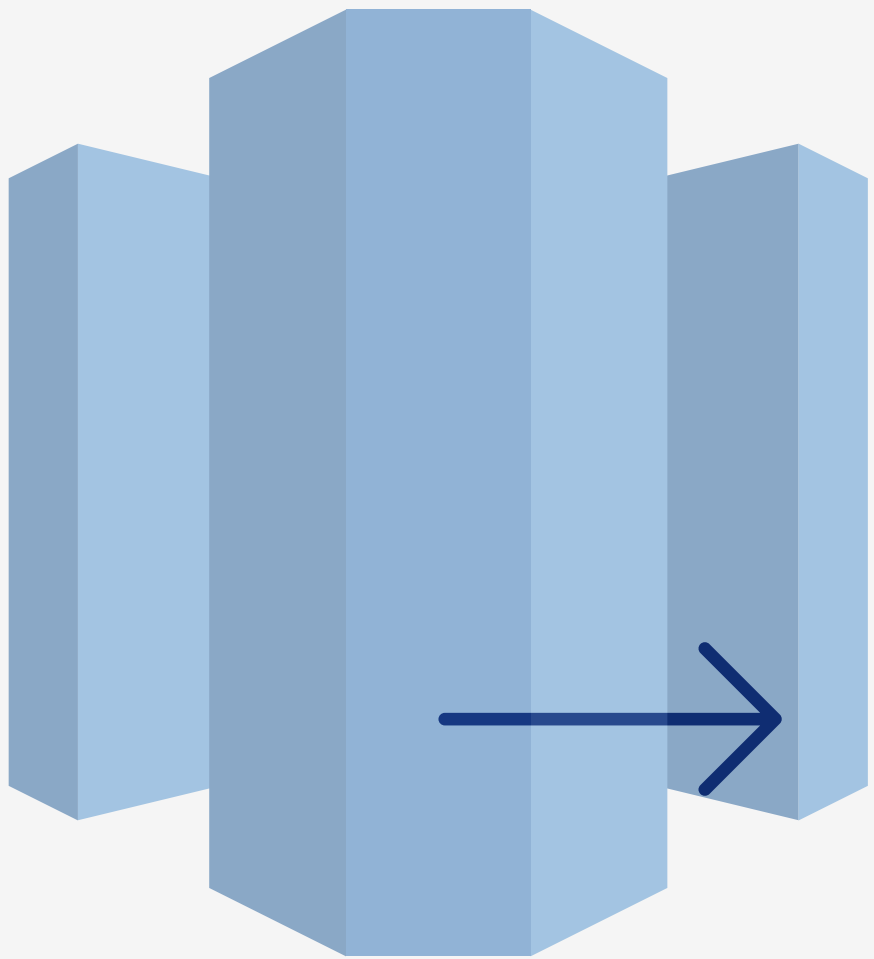
Concurrency Scaling

14

Concurrency scaling allows Redshift to automatically add additional capacity to handle large numbers of queries concurrently.

ENABLE CONCURRENCY SCALING in the cluster configuration to manage high query loads.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

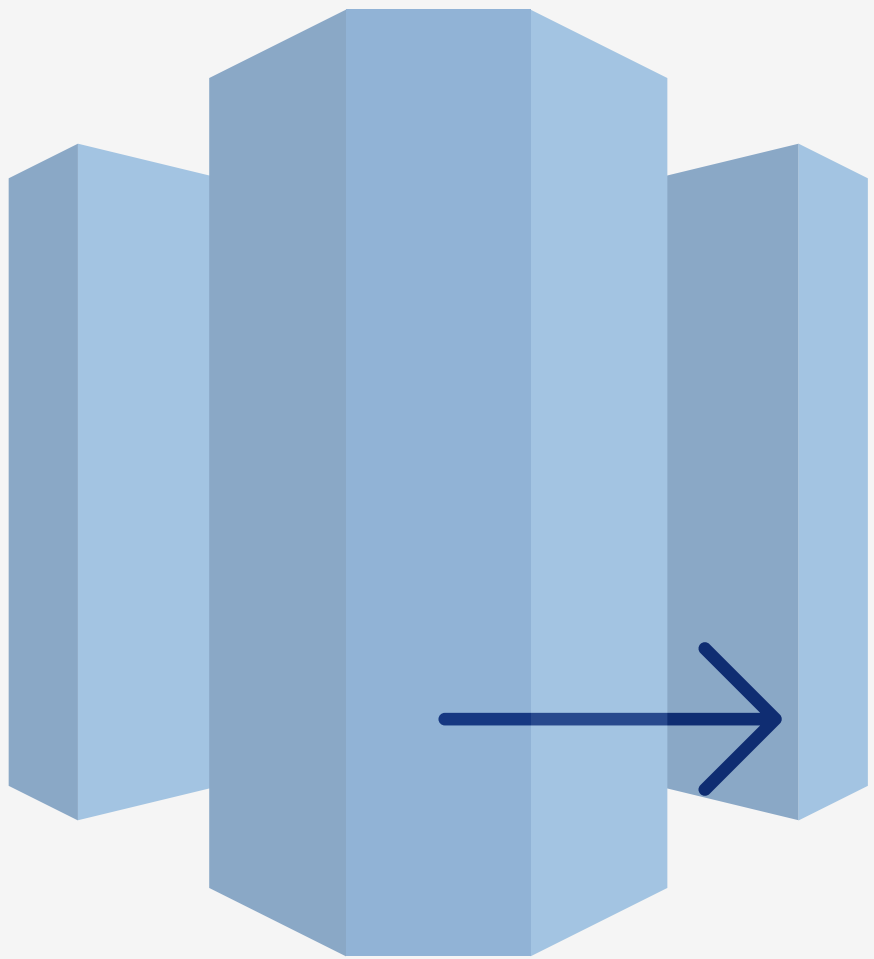
15

Elastic Resize

Allows for quickly resizing the cluster, adding or removing nodes to adjust to workload demands.

aws redshift modify-cluster --cluster-identifier my-cluster --number-of-nodes 4

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

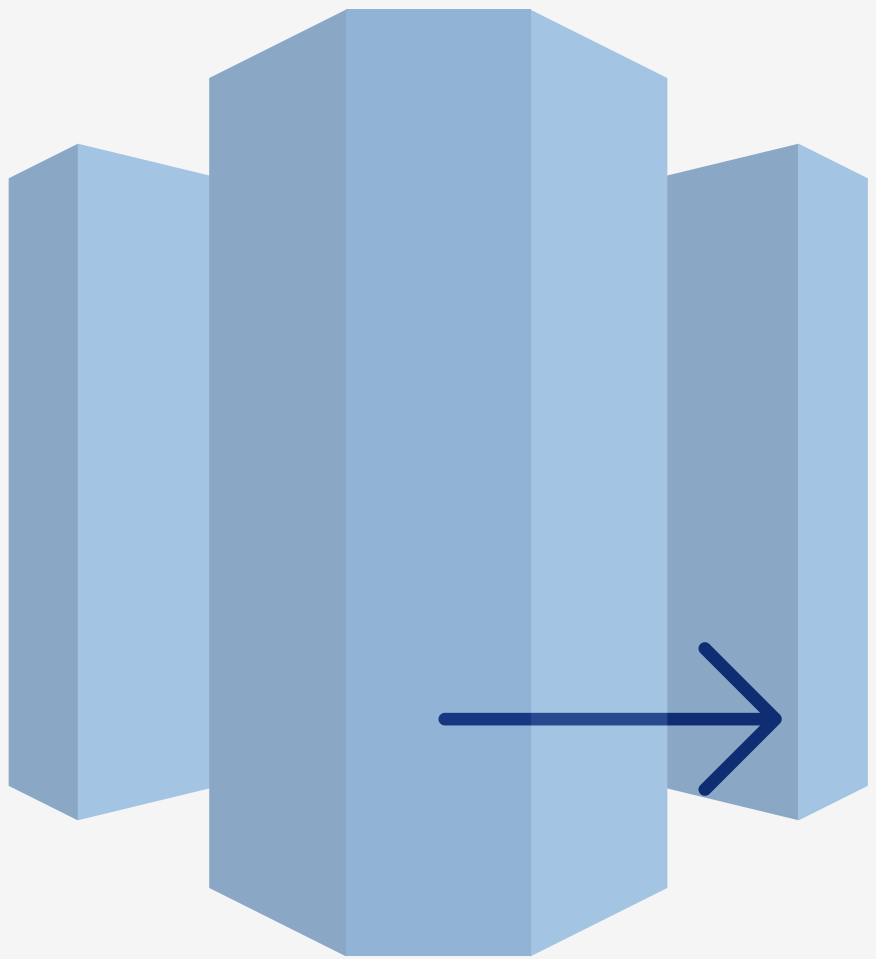
16

Redshift Spectrum

Redshift Spectrum enables querying data directly from S3 without loading it into Redshift tables.

*SELECT * FROM spectrum_table; with spectrum_table defined as an external table pointing to S3 data.*

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

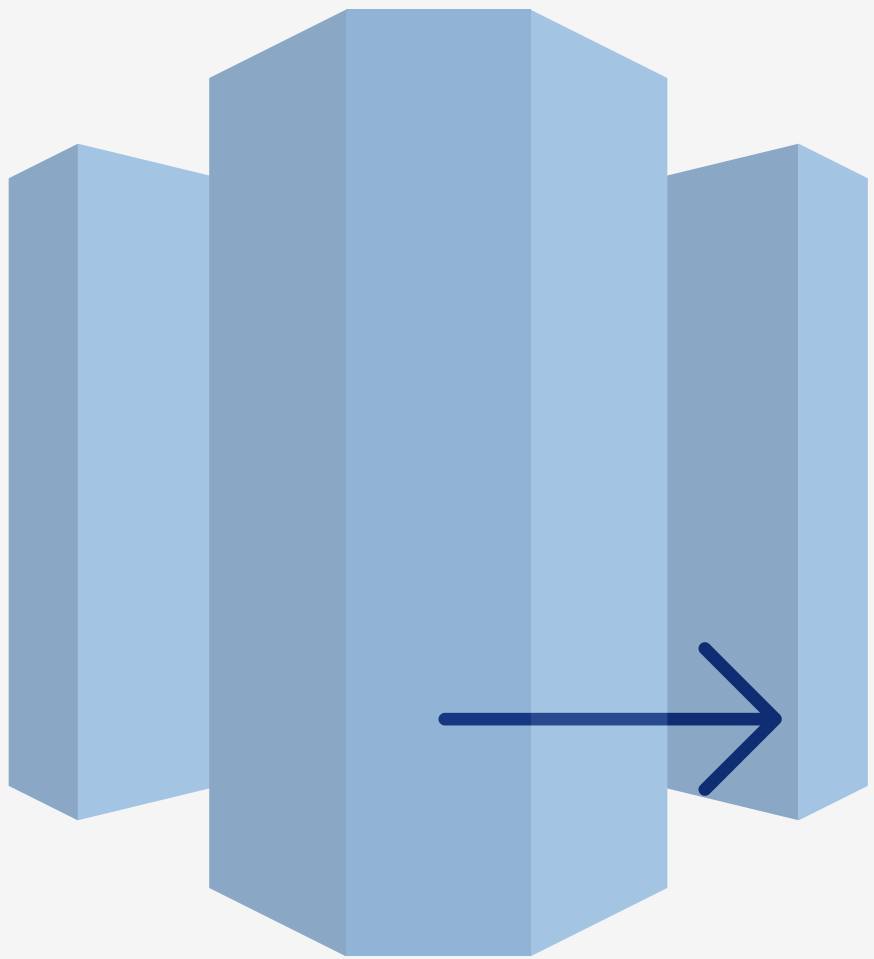
17

External Tables

External tables allow Amazon Redshift to query data stored outside of Redshift, typically in Amazon S3, using Redshift Spectrum.

CREATE EXTERNAL TABLE spectrum.sales (...) STORED AS PARQUET LOCATION 's3://bucket-name/sales_data/';

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

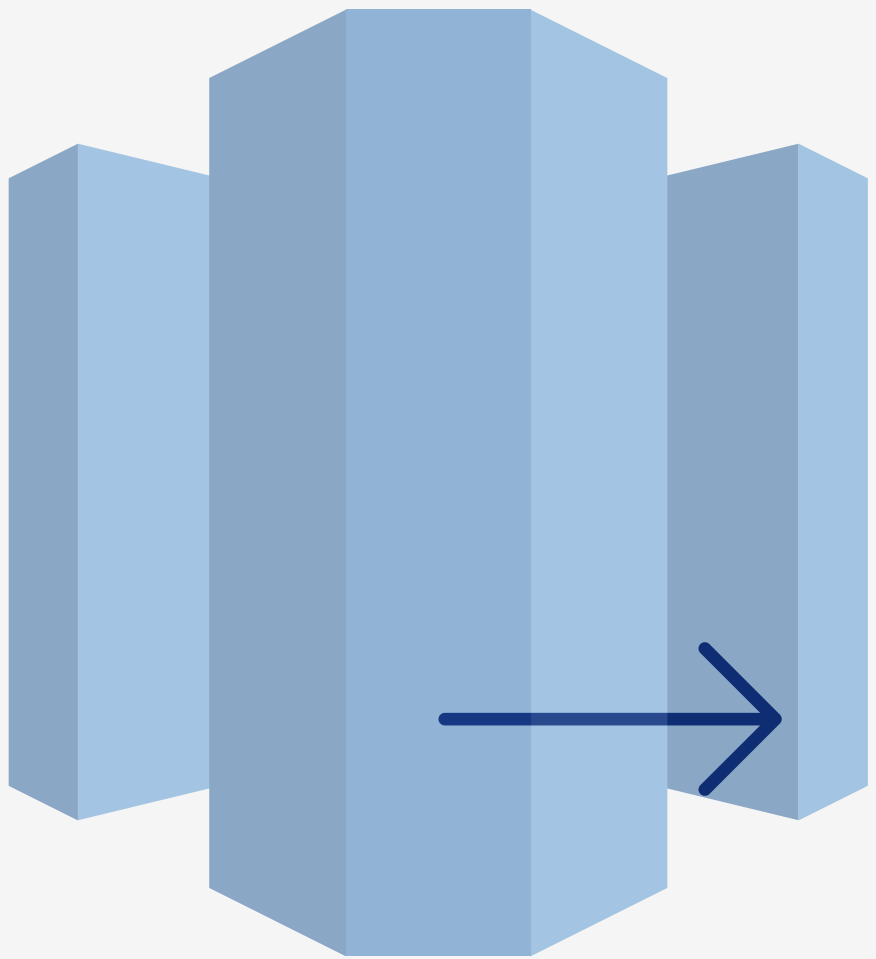
18

WLM (Workload Management)

WLM allows you to define queues that allocate resources based on query priority, enabling better management of multiple workloads.

ALTER WLM CONFIGURATION ADD QUEUE myqueue WITH MEMORY_PERCENTAGE=25;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

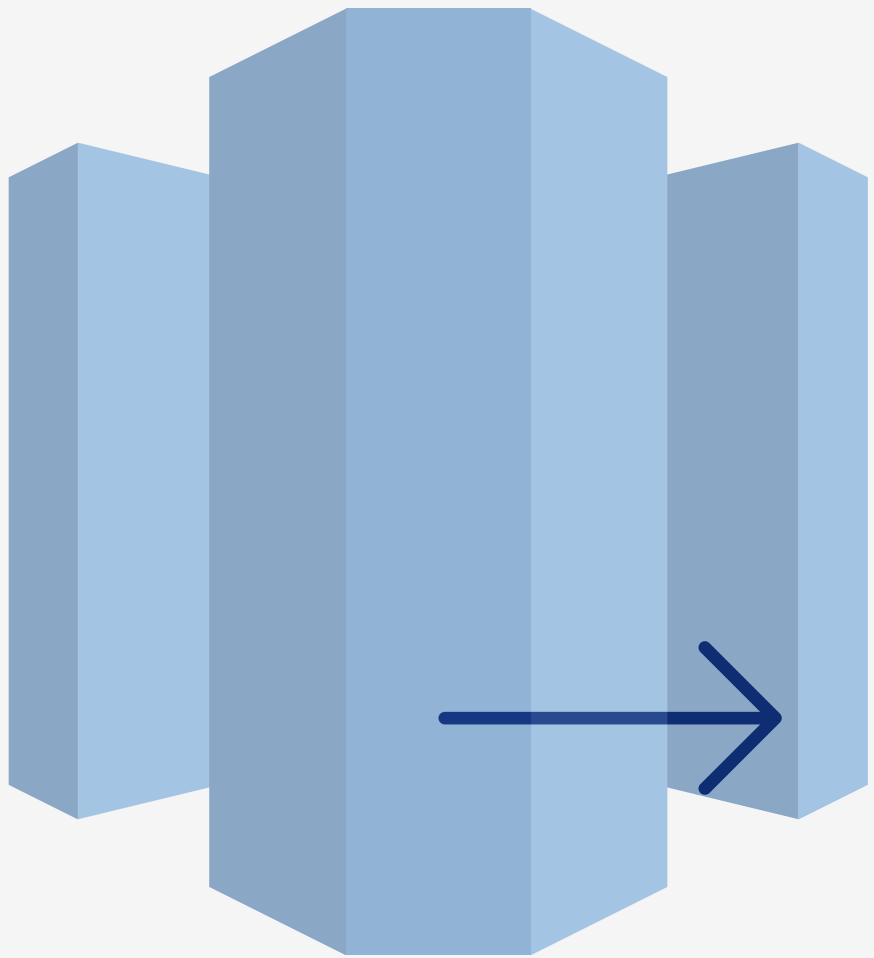
19

RA3 Instances

RA3 instances decouple compute and storage, allowing users to scale compute and storage independently.

CREATE CLUSTER with ra3.16xlarge instance types for compute/storage decoupling.

Shwetank Singh
GritSetGrow - GSGLearn.com



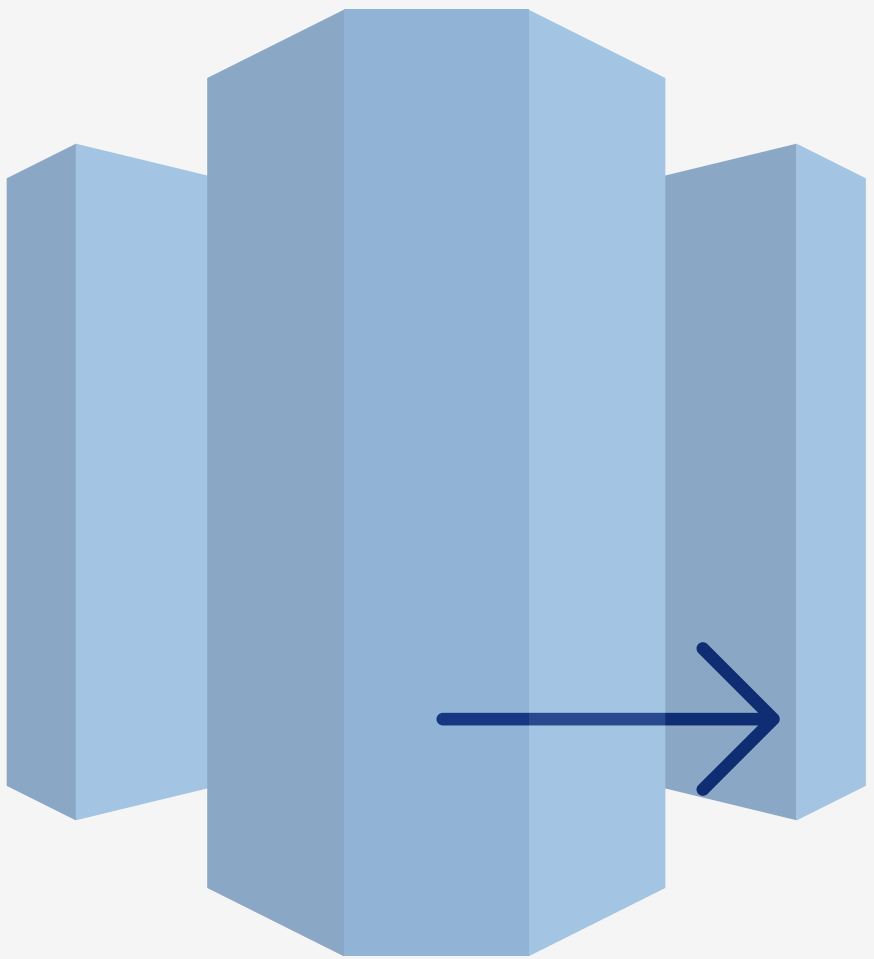
Data Engineering 101: Redshift

20

Query Monitoring Rules (QMR)

QMR helps in monitoring and managing runaway queries by setting rules that define when a query should be canceled or alerted.

CREATE QUERY MONITORING RULE
abort_long_running_query AS rule_action = log;



Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Redshift

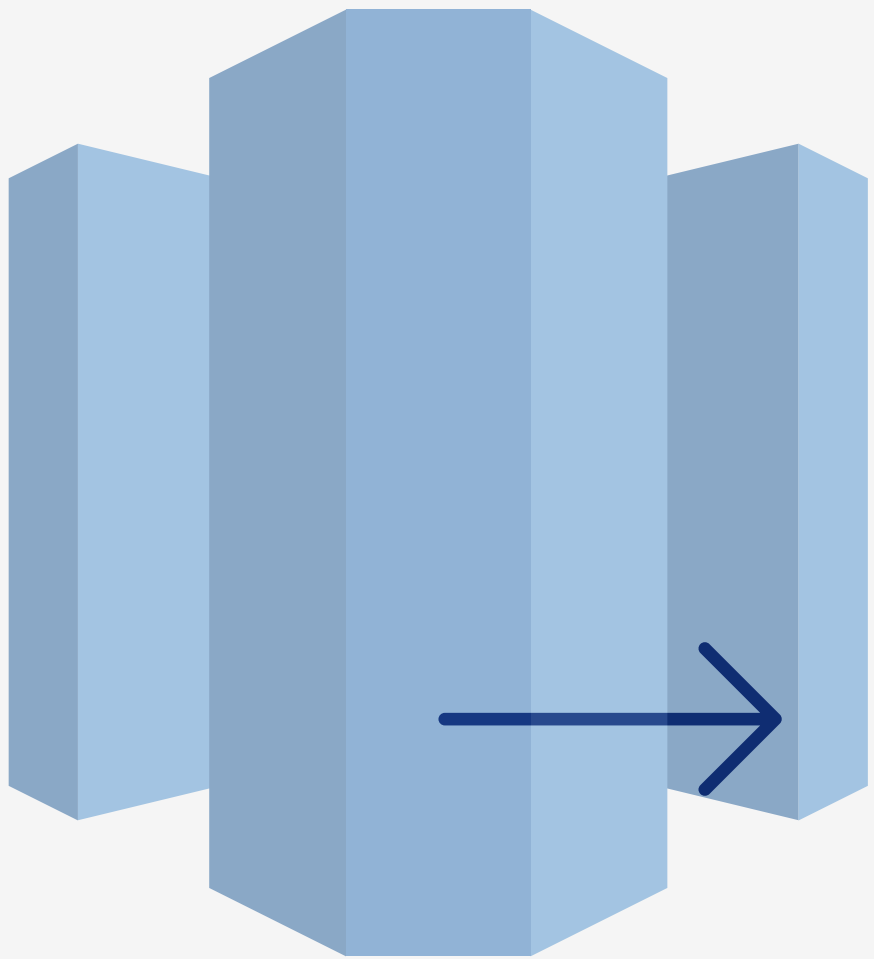
21

Automatic Table Optimization

Redshift automatically chooses the best sort and distribution keys for tables based on usage patterns, optimizing query performance.

Automatic optimization suggestions can be viewed and applied by reviewing the recommendations in the AWS Management Console.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

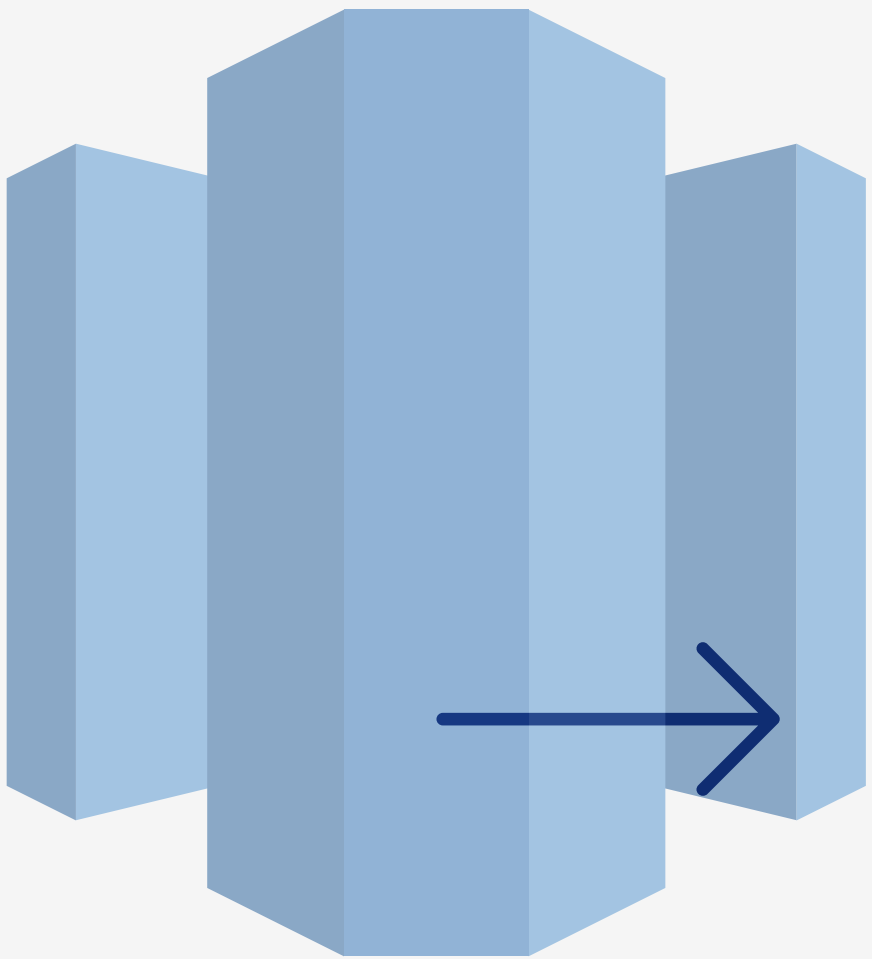
22

Stored Procedures

Stored procedures allow you to write procedural code that runs on the Redshift server, helping automate tasks such as data transformation.

CREATE PROCEDURE sp_myproc() BEGIN ... END;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

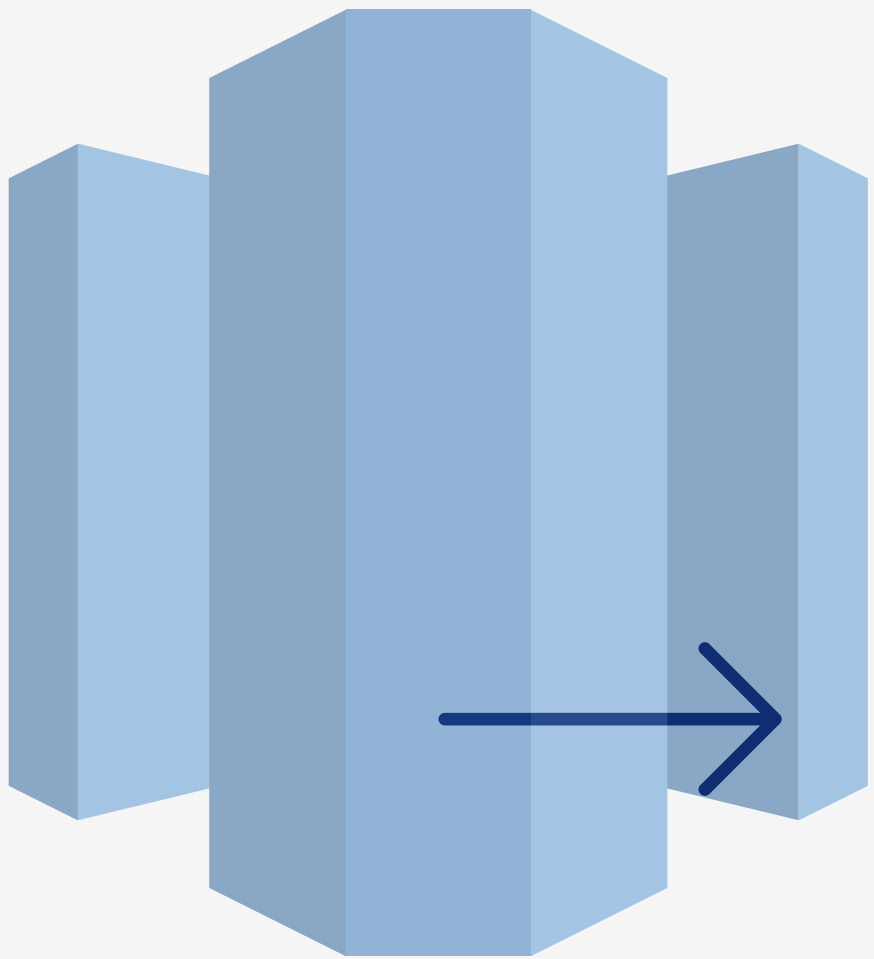
23

UDF (User Defined Functions)

UDFs let you write custom functions in SQL or Python to perform complex calculations or data manipulations within queries.

*CREATE FUNCTION myfunction(val INT) RETURNS INT IMMUTABLE AS $$ BEGIN RETURN val * 2; END; $$ LANGUAGE plpgsql;*

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

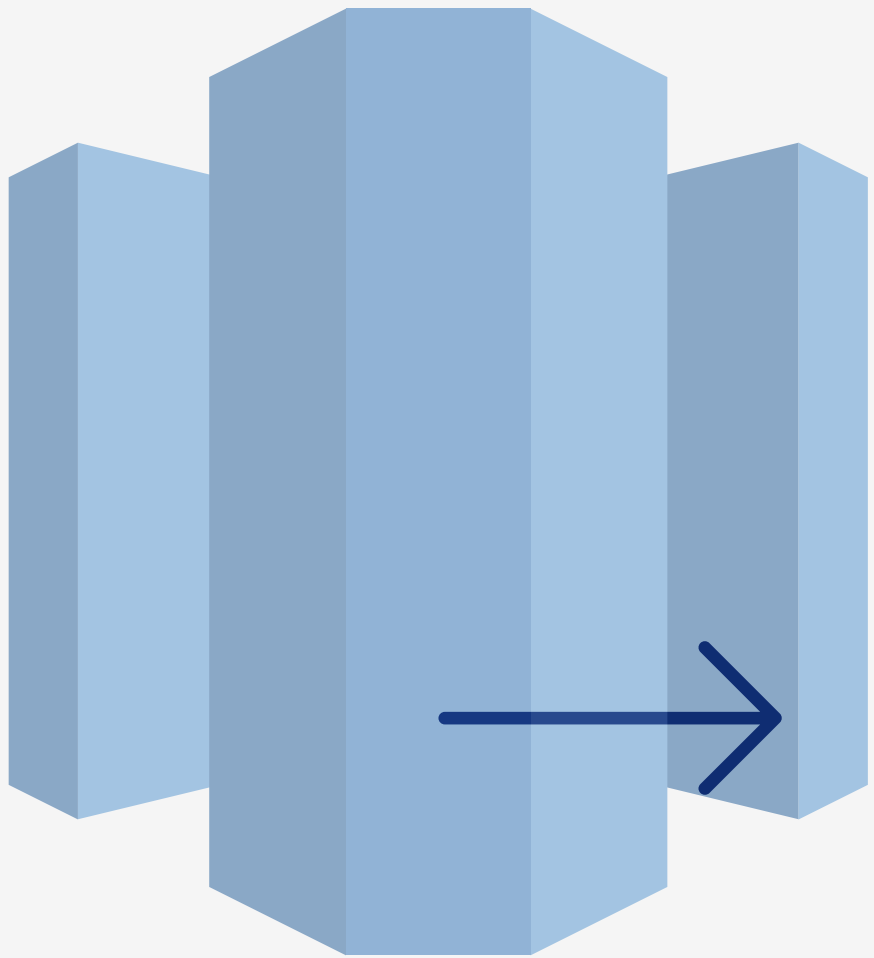
Data Sharing

24

Amazon Redshift data sharing allows secure and efficient sharing of live data across different Redshift clusters without needing to copy data.

ALTER DATASHARE myshare ADD SCHEMA public; to share schema across clusters.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

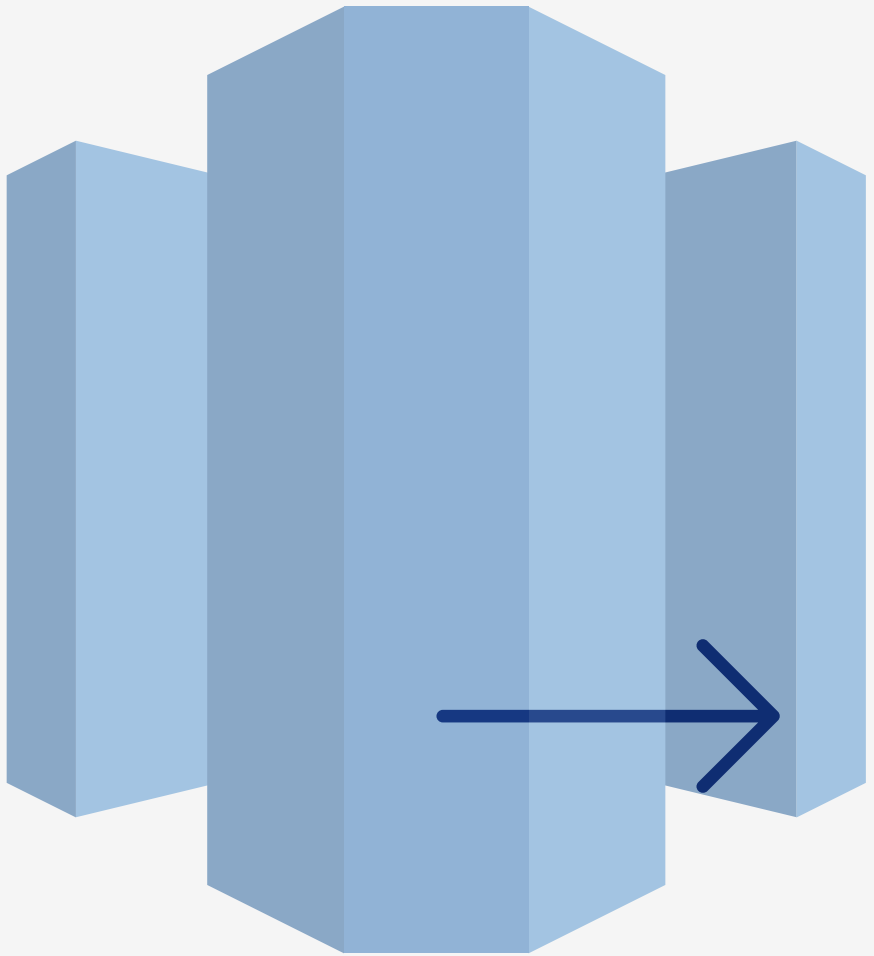
Enhanced VPC Routing

25

Enhanced VPC Routing forces all COPY and UNLOAD traffic between your cluster and data repositories in S3 to go through your Amazon VPC.

ENABLE ENHANCED VPC ROUTING in the cluster configuration to route traffic securely through VPC.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

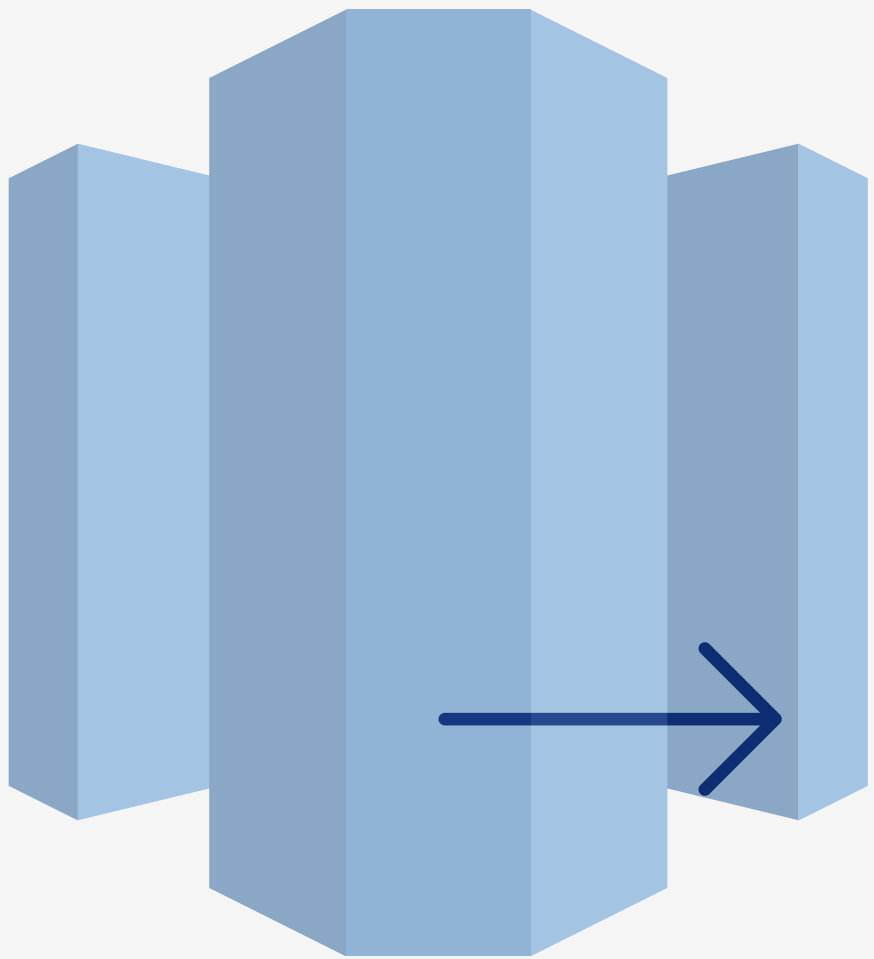
Column-Level Encryption

76

Redshift supports column-level encryption, allowing you to encrypt specific columns of your data at rest using AWS KMS keys.

CREATE TABLE sensitive_data (ssn CHAR(11) ENCODE BYTEDICT ENCRYPTED);

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

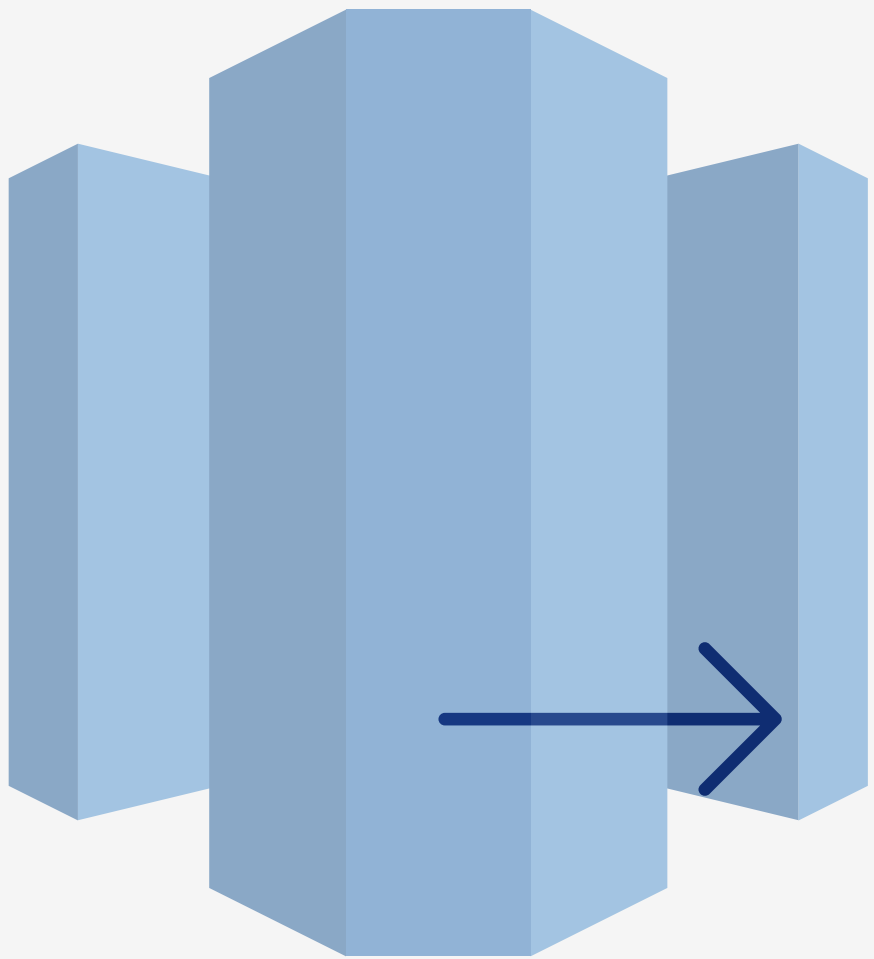
27

Data API

Redshift Data API provides a way to run SQL commands against Redshift clusters without needing to manage connections, useful for serverless applications.

*aws redshift-data execute-statement --cluster-identifier my-cluster --database mydb --sql "SELECT * FROM sales;"*

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

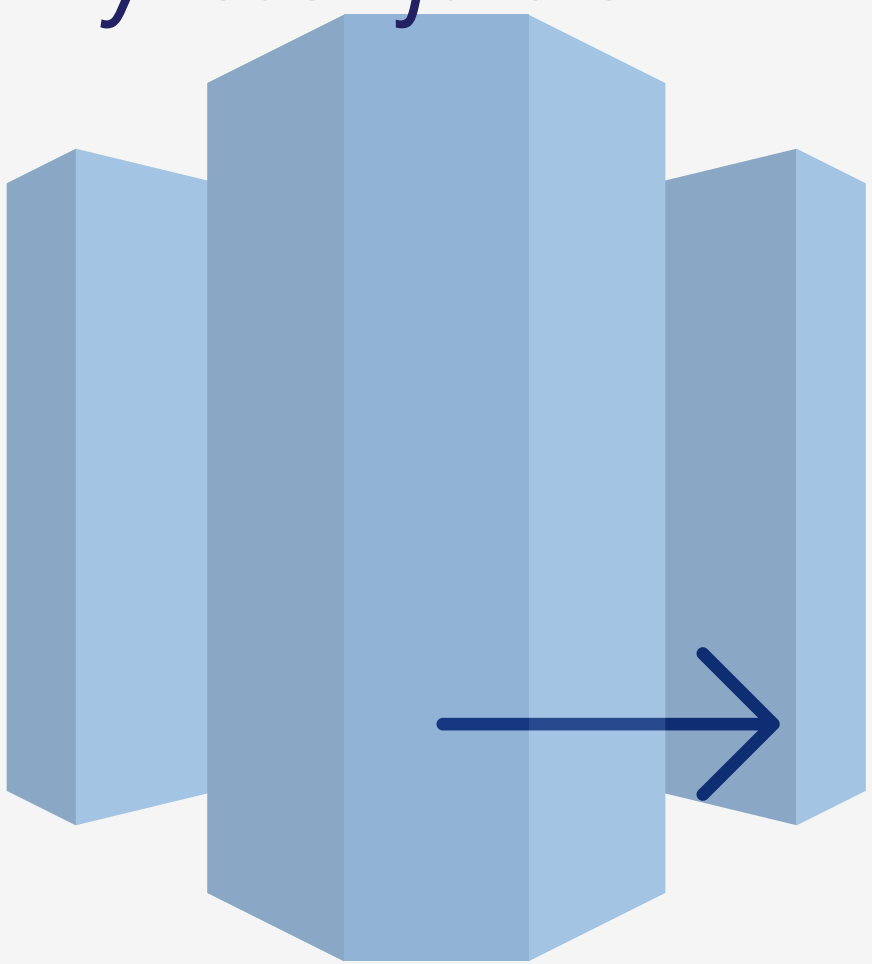
78

UNLOAD Command

The UNLOAD command exports result sets from Redshift tables to Amazon S3 in various formats, such as text or Parquet.

```
UNLOAD ('SELECT * FROM sales') TO 's3://bucket-name/unload/' IAM_ROLE
'arn:aws:iam::123456789012:role/MyRedshiftRole'
FORMAT AS PARQUET;
```

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

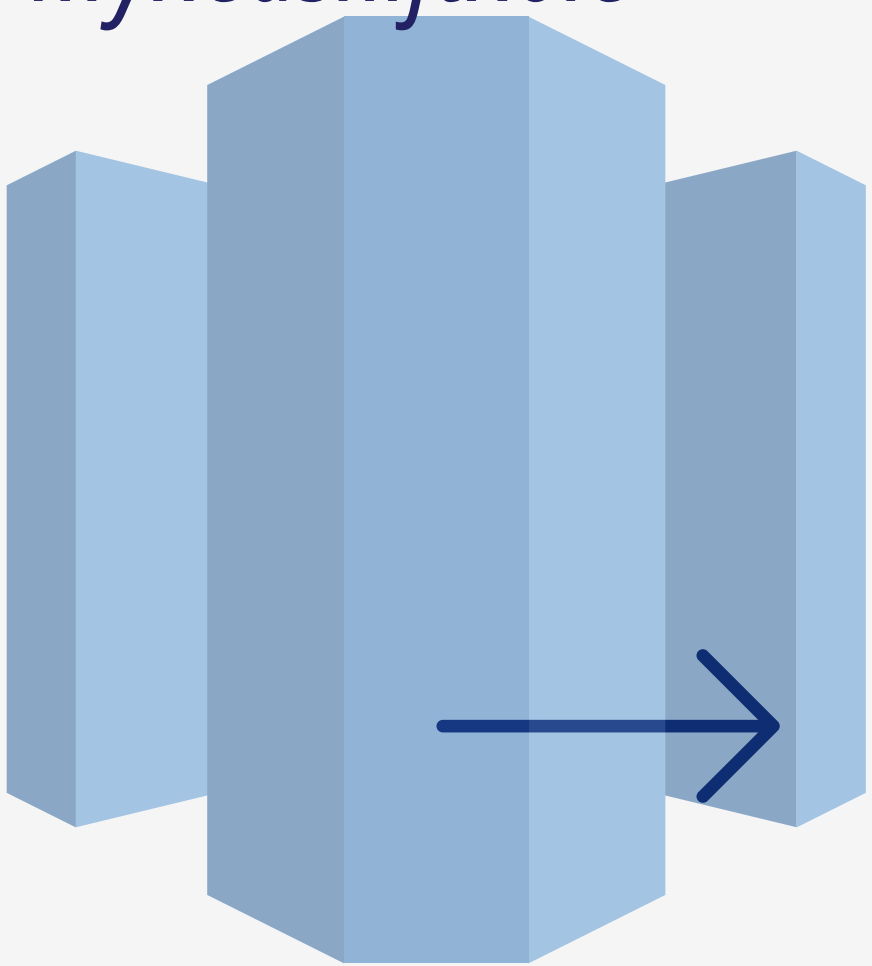
79

COPY Command

The COPY command loads data from Amazon S3, DynamoDB, or other sources into Redshift tables. It supports various data formats and parallelism.

```
COPY sales FROM 's3://bucket-name/sales_data.csv'
IAM_ROLE
'arn:aws:iam::123456789012:role/MyRedshiftRole'
CSV;
```

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

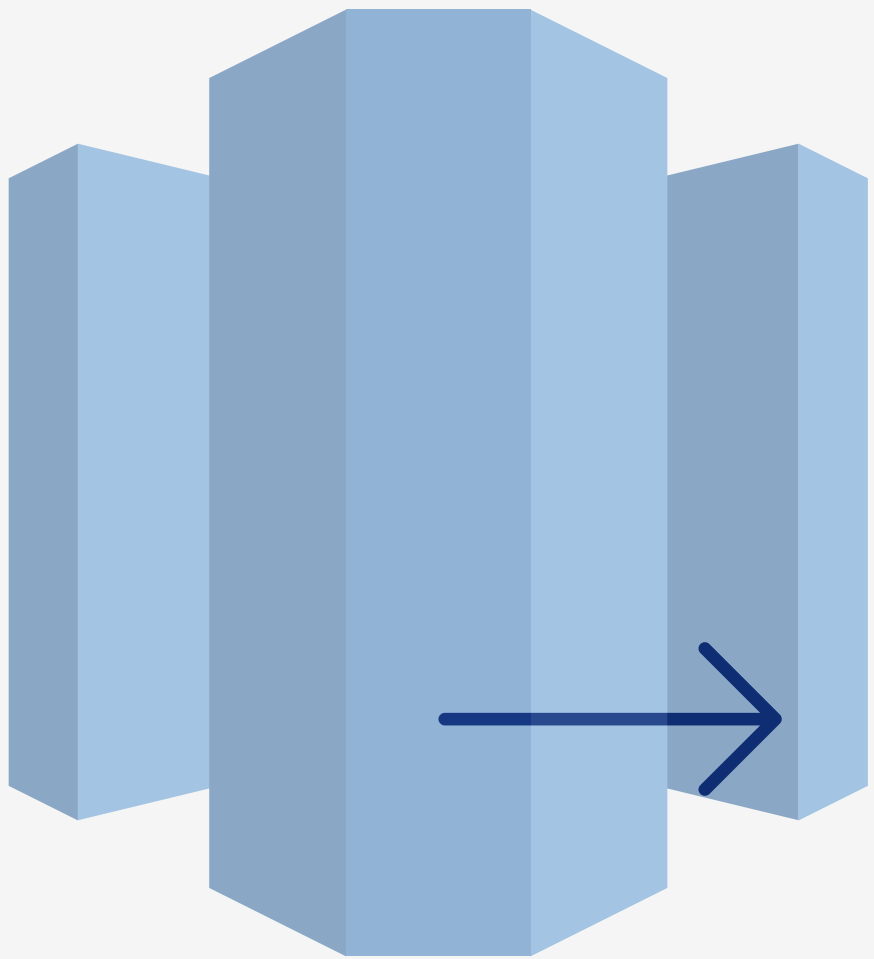
30

Concurrency Scaling

Enables Amazon Redshift to automatically add additional capacity to handle large numbers of queries concurrently.

ALTER SYSTEM SET concurrency_scaling=ON;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

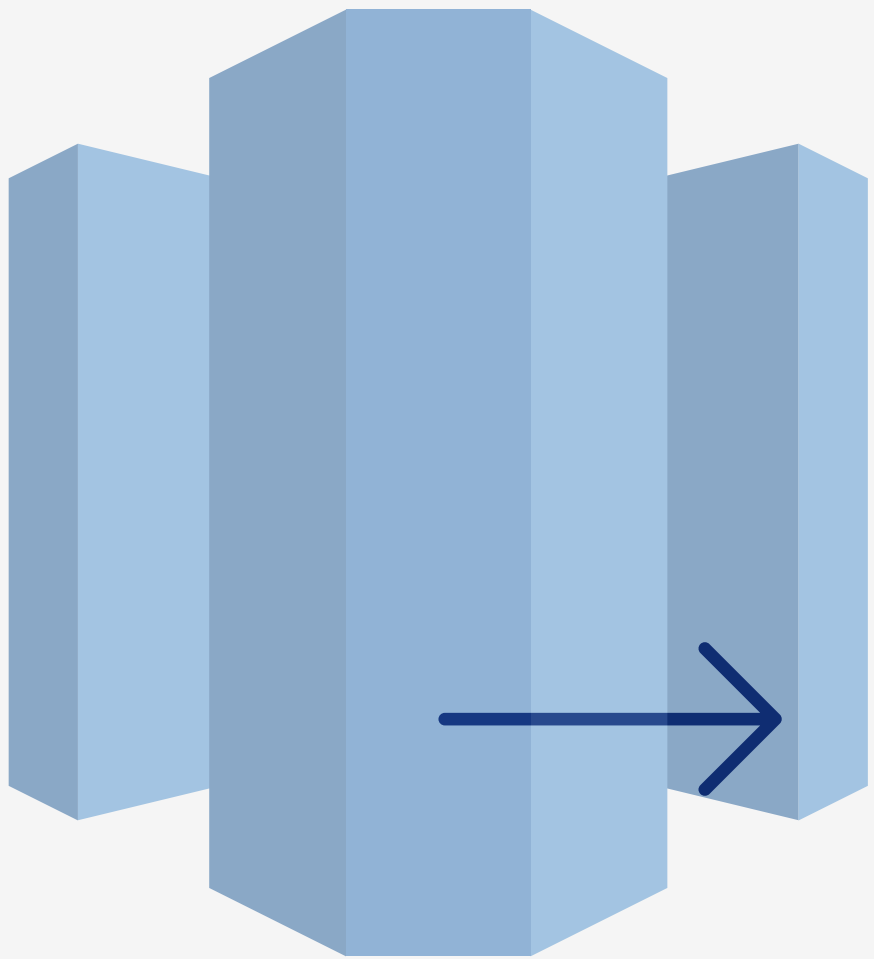
31

Redshift ML

Allows users to create machine learning models directly within Redshift using SQL queries, powered by Amazon SageMaker.

*CREATE MODEL my_model FROM (SELECT * FROM sales) TARGET amount;*

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

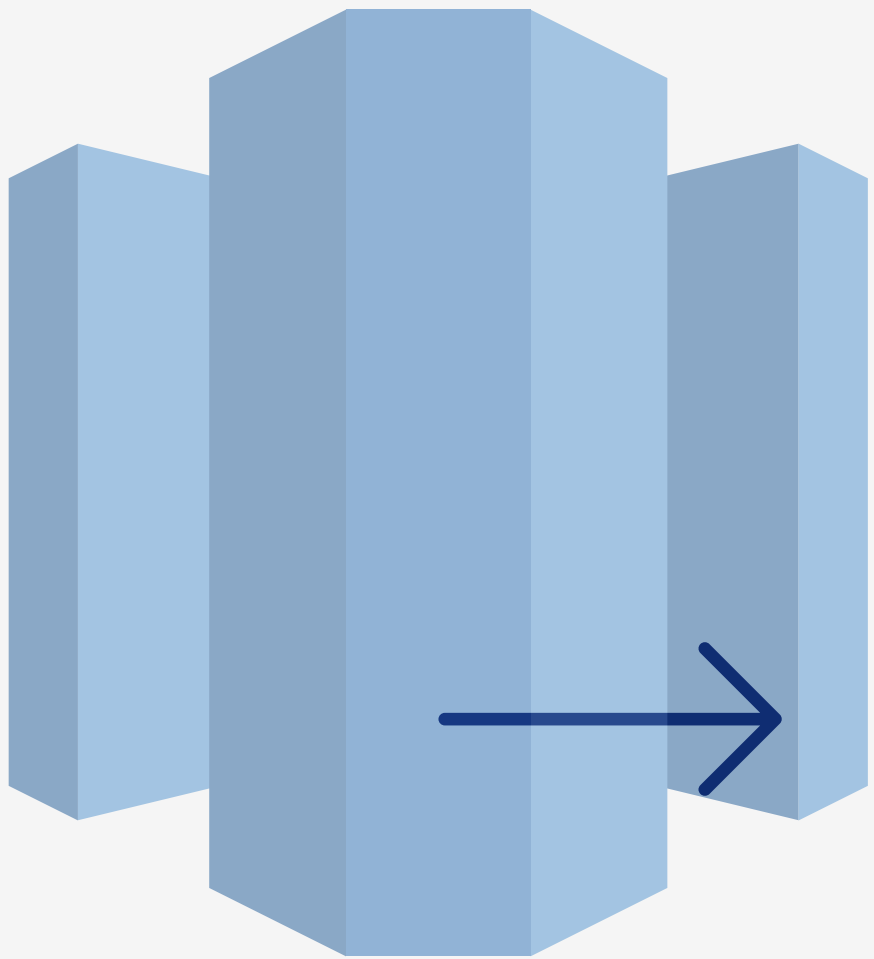
32

Partitioning in Spectrum

Partitioning in Spectrum helps optimize queries on external tables by reducing the amount of data scanned by splitting it into partitions.

ALTER TABLE spectrum.sales ADD PARTITION (year=2023, month=1) LOCATION 's3://bucket-name/sales/2023/01/';

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

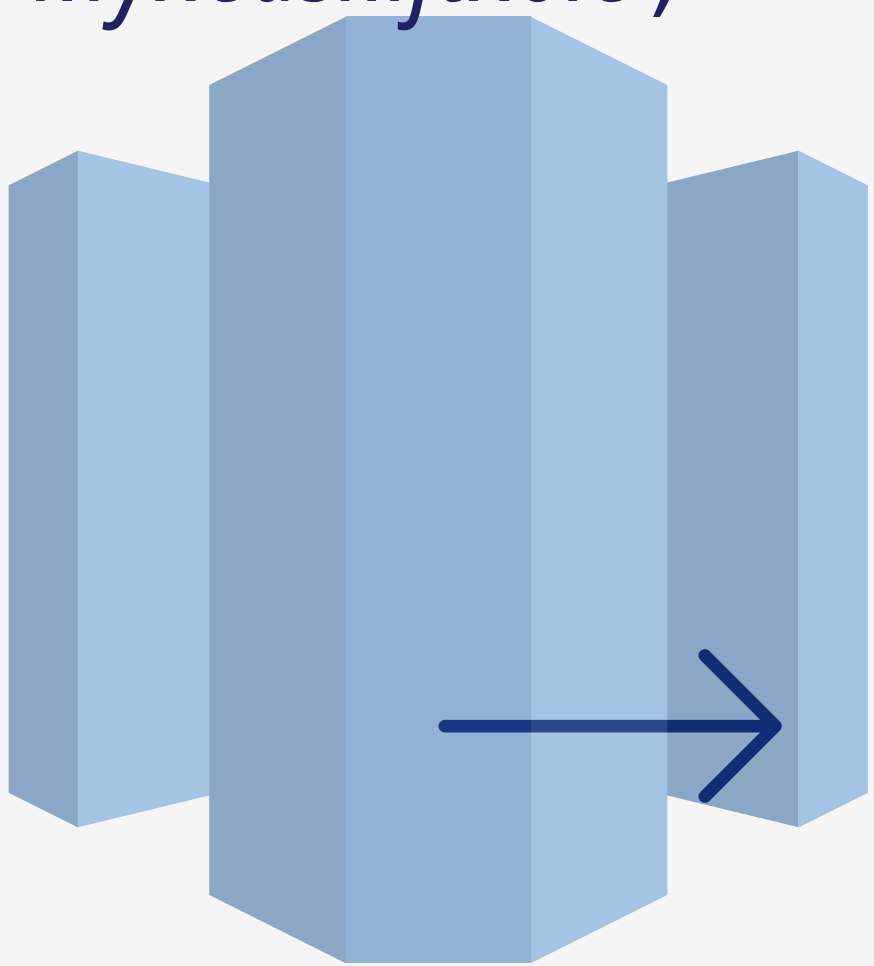
33

AWS Glue Data Catalog

AWS Glue Data Catalog is a fully managed service that lets you store and retrieve metadata about your data, which can be queried by Redshift Spectrum.

CREATE EXTERNAL SCHEMA spectrum FROM DATA CATALOG DATABASE 'mycatalogdb' IAM_ROLE 'arn:aws:iam::123456789012:role/MyRedshiftRole';

Shwetank Singh
GritSetGrow - GSGLearn.com



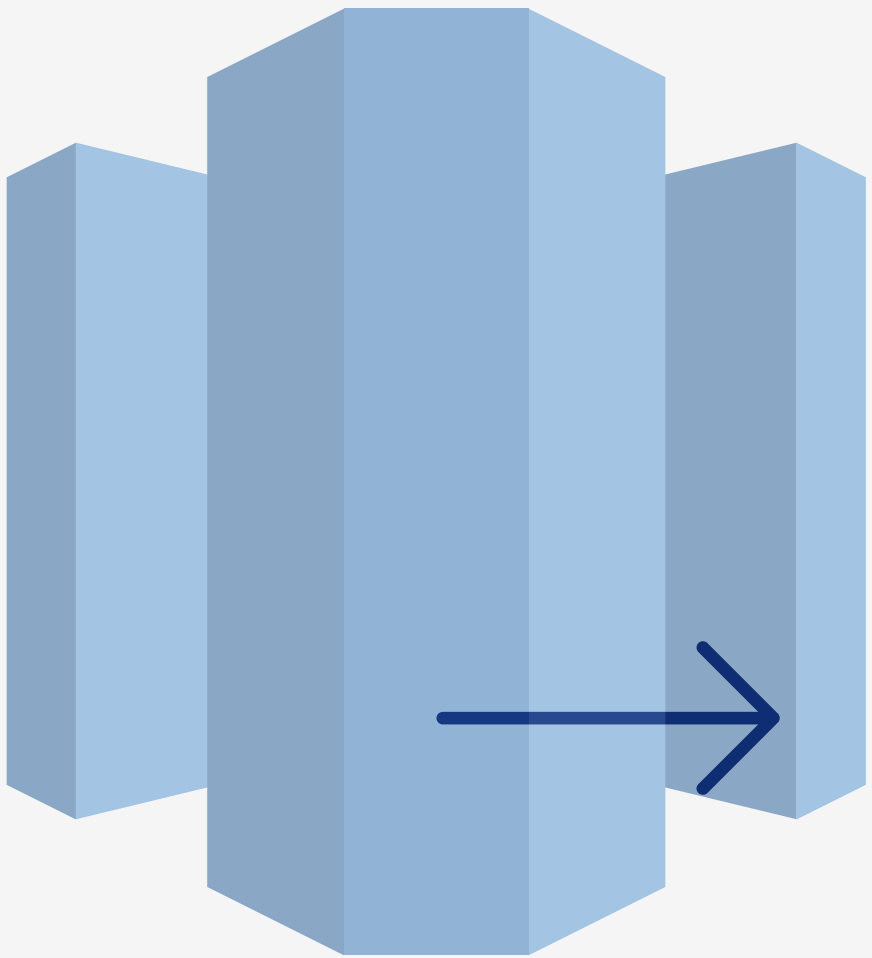
Data Engineering 101: Redshift

34

Concurrency Limits

Redshift manages concurrency by allocating resources to different queries based on the defined WLM settings and query priority.

*Monitoring concurrency: SELECT * FROM stv_wlm_query_state;*



Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Redshift

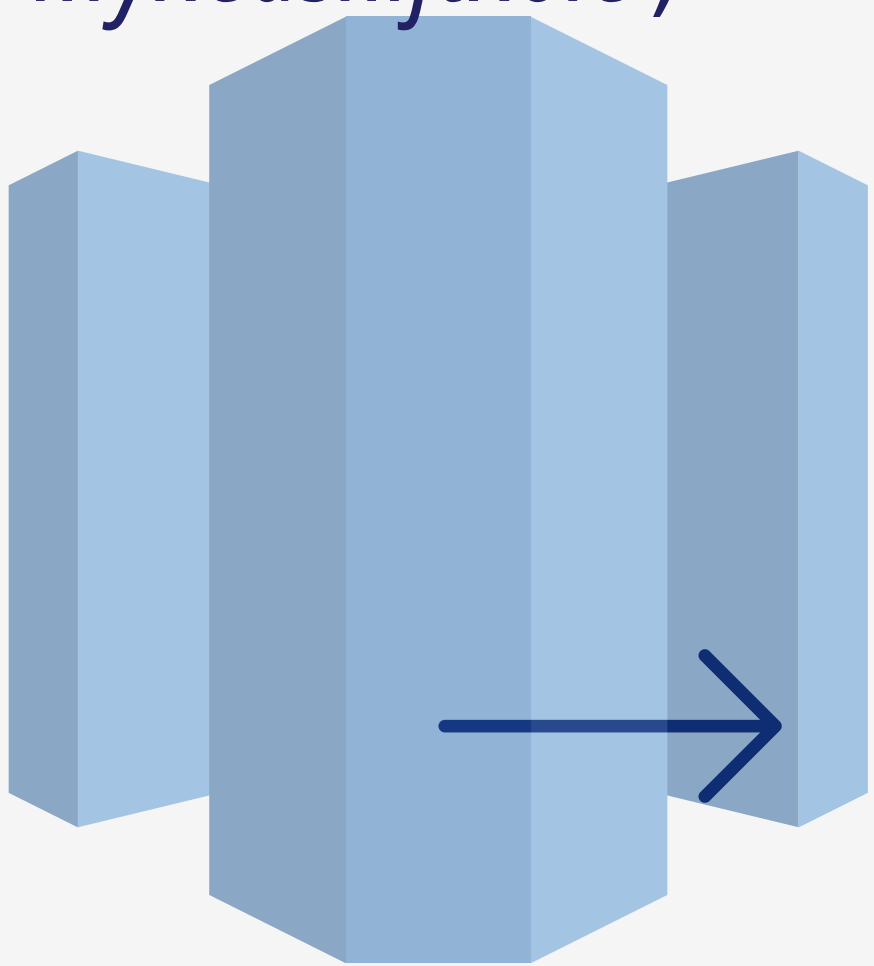
35

Amazon S3 Integration

Redshift integrates with Amazon S3 for data ingestion and backup, enabling seamless data transfer between Redshift and S3 for large-scale data processing.

Data ingestion: COPY mytable FROM 's3://bucket-name/data.csv' IAM_ROLE 'arn:aws:iam::123456789012:role/MyRedshiftRole';

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

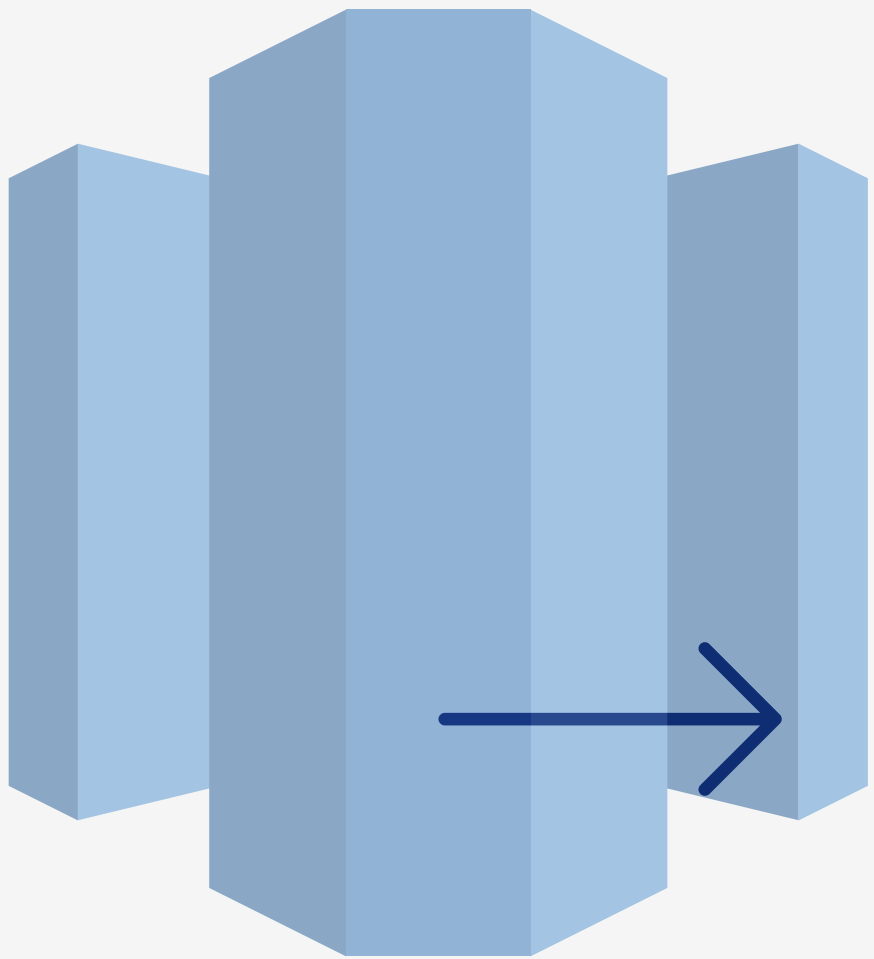
36

Security Groups

Redshift uses Amazon VPC security groups to control inbound and outbound traffic to your Redshift clusters, providing network-level security.

aws redshift modify-cluster --cluster-identifier my-cluster --vpc-security-group-ids sg-12345678

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

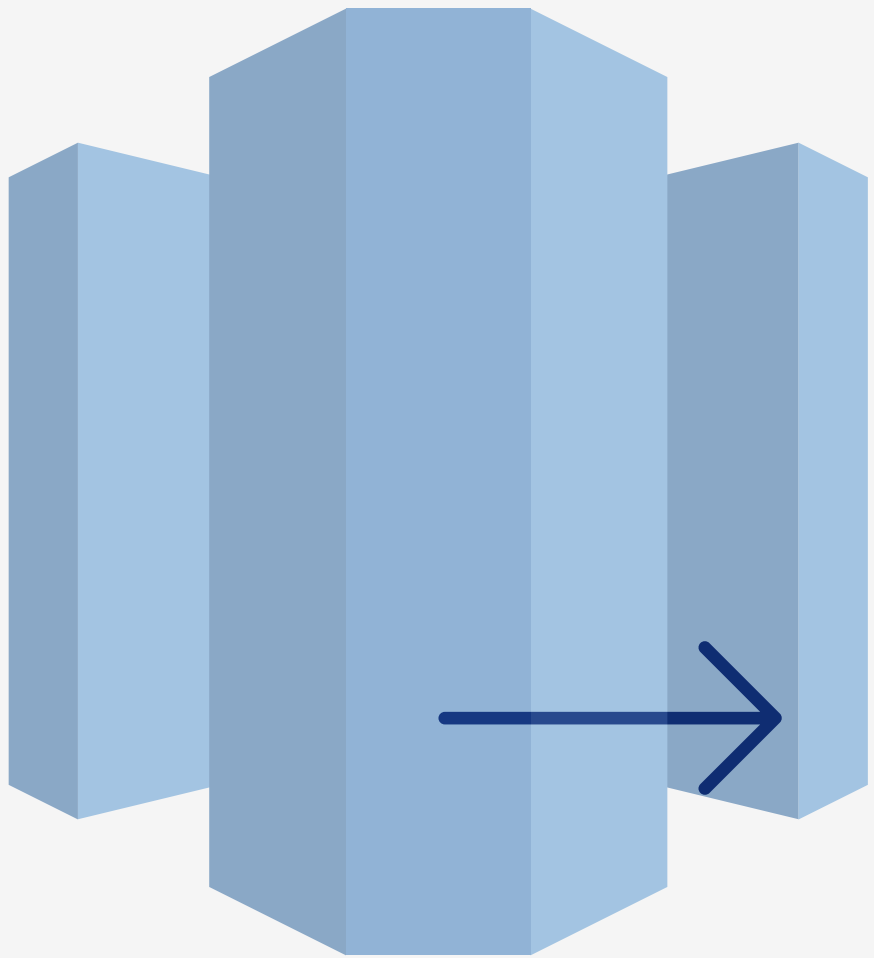
37

Audit Logging

Audit logging in Redshift allows you to track database events and query activity for security and compliance purposes by saving logs to Amazon S3.

ENABLE AUDIT LOGGING in cluster configuration to store logs in S3.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

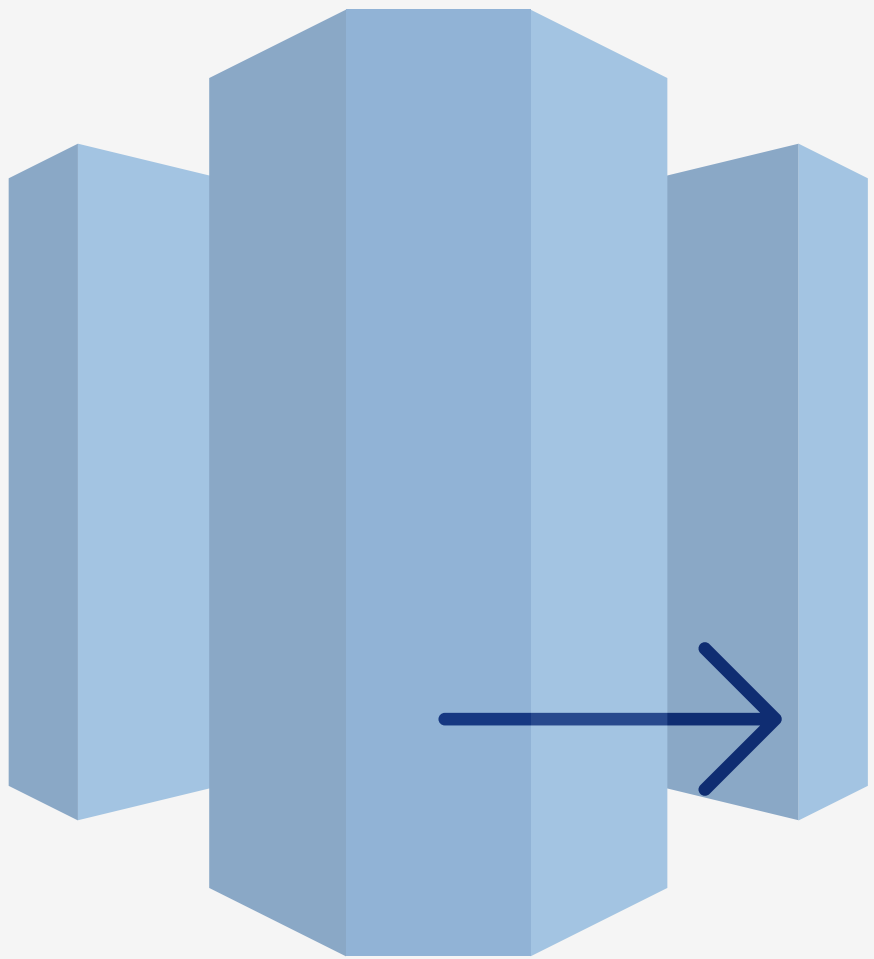
38

Automated Snapshots

Redshift automatically creates snapshots of your data to protect against data loss, which can be configured for specific intervals and retention periods.

Configure snapshots: aws redshift modify-cluster-snapshot-schedule --cluster-identifier my-cluster --snapshot-schedule my-schedule

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

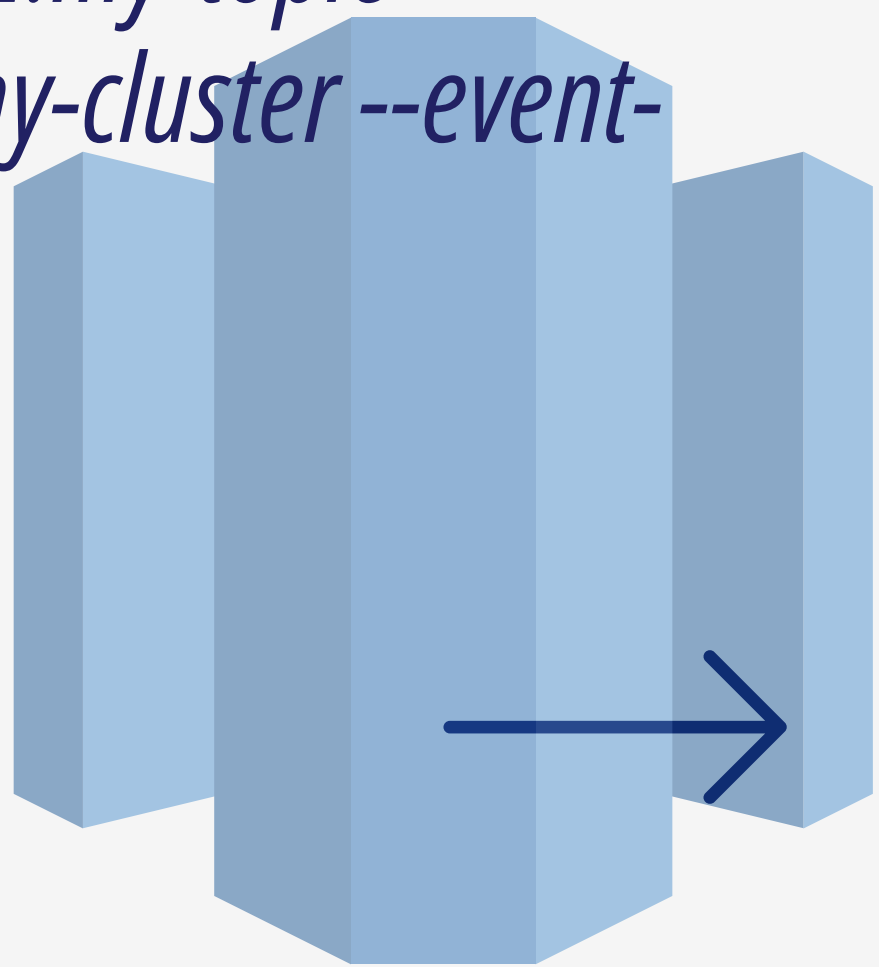
39

Event Notifications

Amazon Redshift can send notifications for specific events such as cluster creation, deletion, or failure, using SNS (Simple Notification Service).

aws redshift create-event-subscription --subscription-name my-subscription --sns-topic-arn arn:aws:sns:region:123456789012:my-topic --source-type cluster --source-ids my-cluster --event-categories availability, security

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

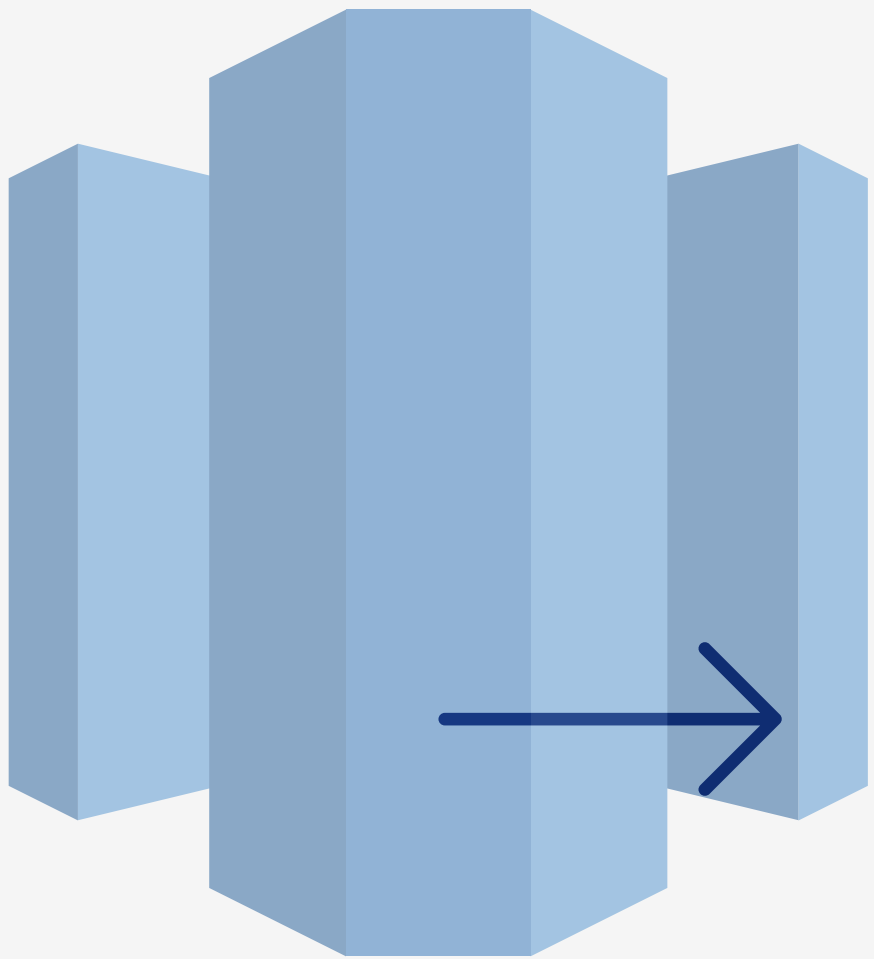
Cluster Parameter Groups

40

Cluster parameter groups allow you to configure database engine settings for your Amazon Redshift cluster, which can be applied at runtime or during a reboot.

Modify parameter group: aws redshift modify-cluster-parameter-group --parameter-group-name my-param-group --parameters "parameterName=value"

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

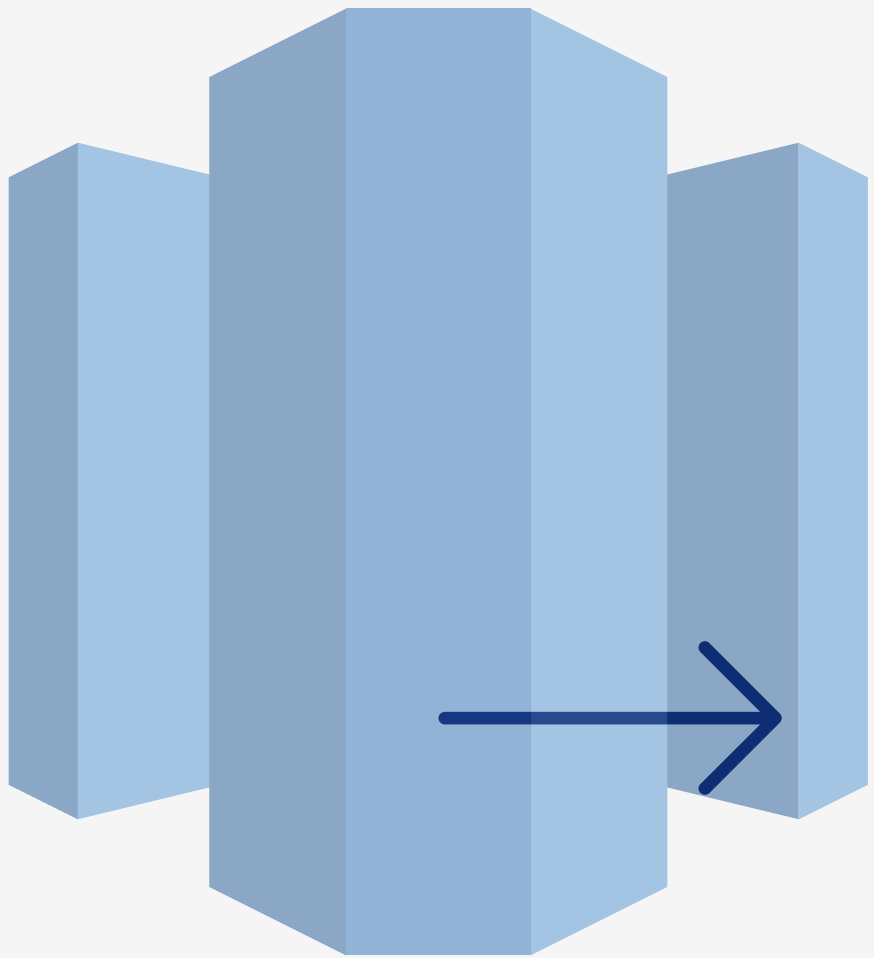
41

Reserved Instances

Reserved instances allow you to save on long-term costs by committing to a one- or three-year term for Redshift clusters, offering significant discounts over on-demand pricing.

Purchase Reserved Instance: aws redshift purchase-reserved-node-offering --reserved-node-offering-id offering-id --node-count 1`

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

42

Elastic IP Address

An Elastic IP address (EIP) is a static IPv4 address that you can associate with your Redshift cluster, allowing for consistent access even after a cluster restart.

Allocate and associate EIP: aws ec2 associate-address --instance-id my-instance-id --allocation-id eipalloc-12345678`

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101: Redshift

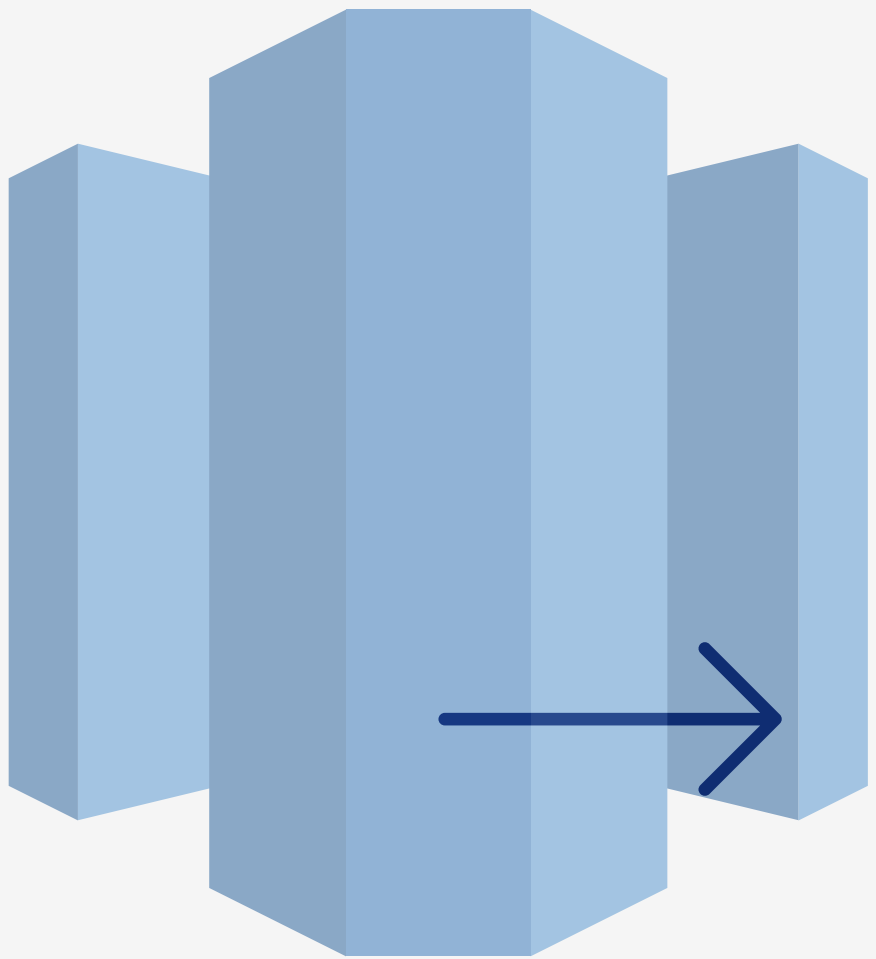
43

Cluster Resizing

Cluster resizing allows you to add or remove nodes in your Redshift cluster to adjust for changes in workload, supporting both classic and elastic resize options.

aws redshift modify-cluster --cluster-identifier my-cluster --number-of-nodes 4 for elastic resize.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

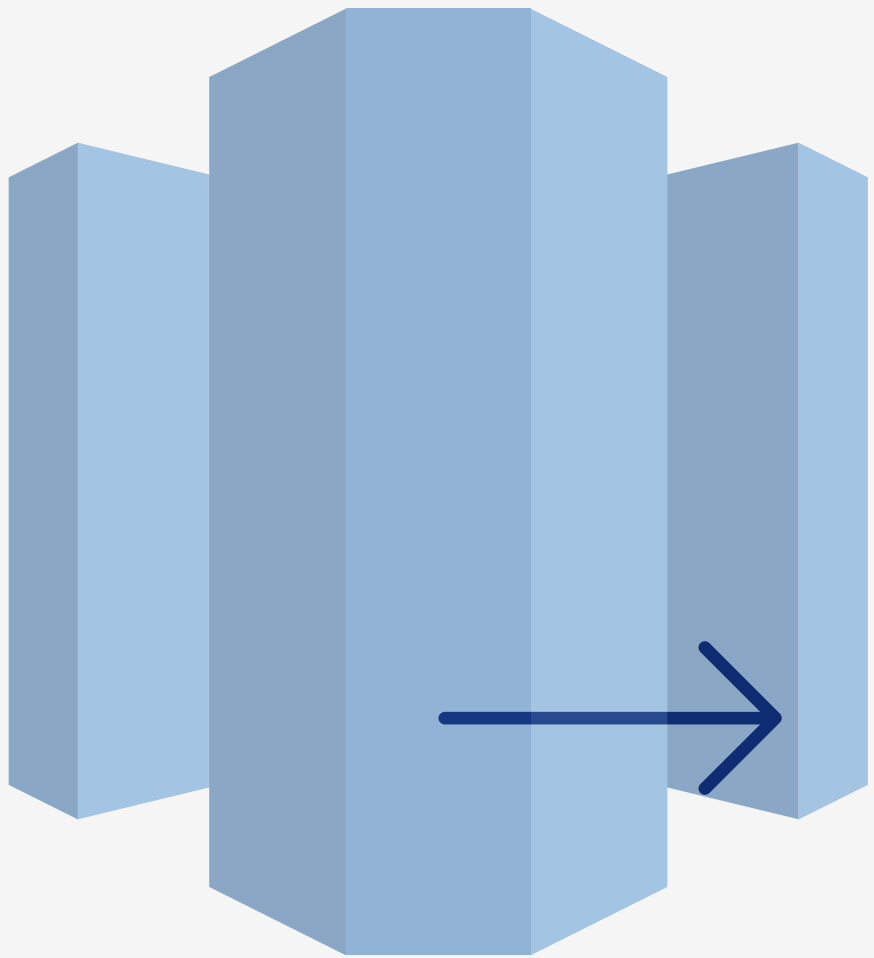
Data Transfer Costs

44

Data transfer costs in Redshift refer to the fees incurred when moving data between Redshift and other AWS services, such as S3, over the internet or across regions.

Monitoring data transfer: Check the AWS Cost Explorer for data transfer costs associated with your Redshift usage.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

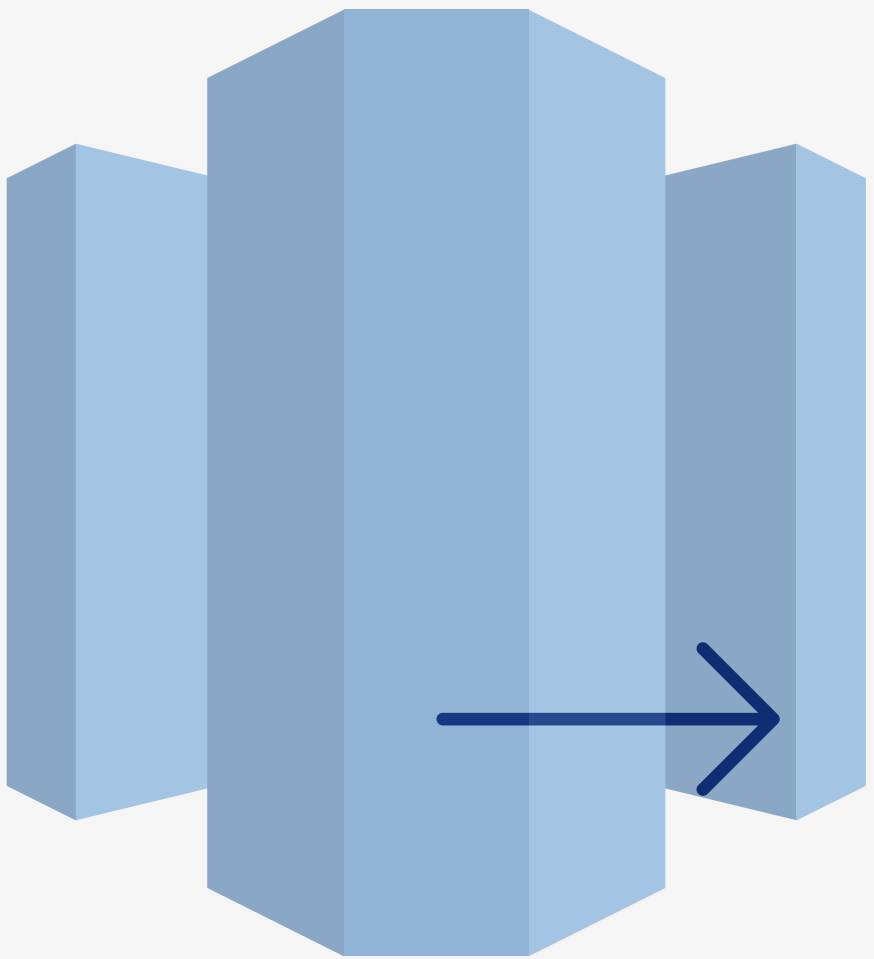
45

Enhanced Logging

Enhanced logging in Redshift captures detailed information about each query, including execution time, plan, and resource consumption, which can be analyzed for performance tuning.

Enable enhanced logging by setting up logging parameters in your Redshift cluster settings.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

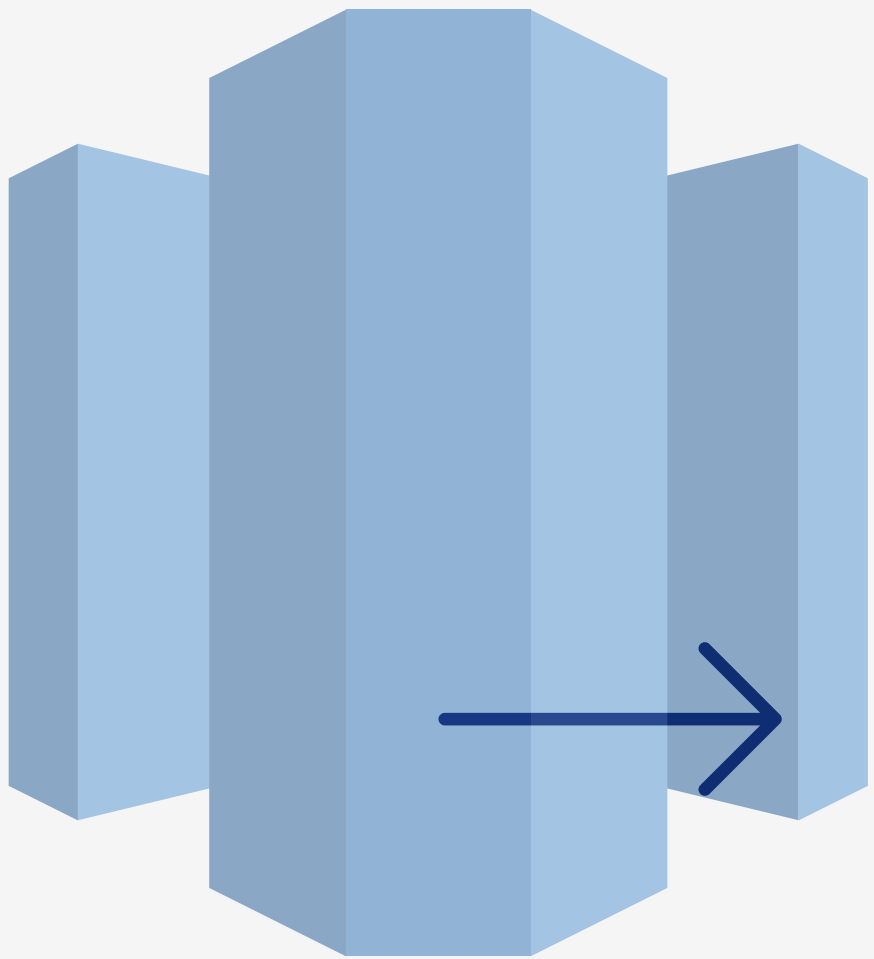
46

Encryption at Rest

Redshift supports encryption of data at rest using AWS Key Management Service (KMS) or customer-managed keys, ensuring that data is protected even when stored.

Enable encryption: aws redshift create-cluster --cluster-identifier my-cluster --encrypted --kms-key-id my-kms-key-id

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

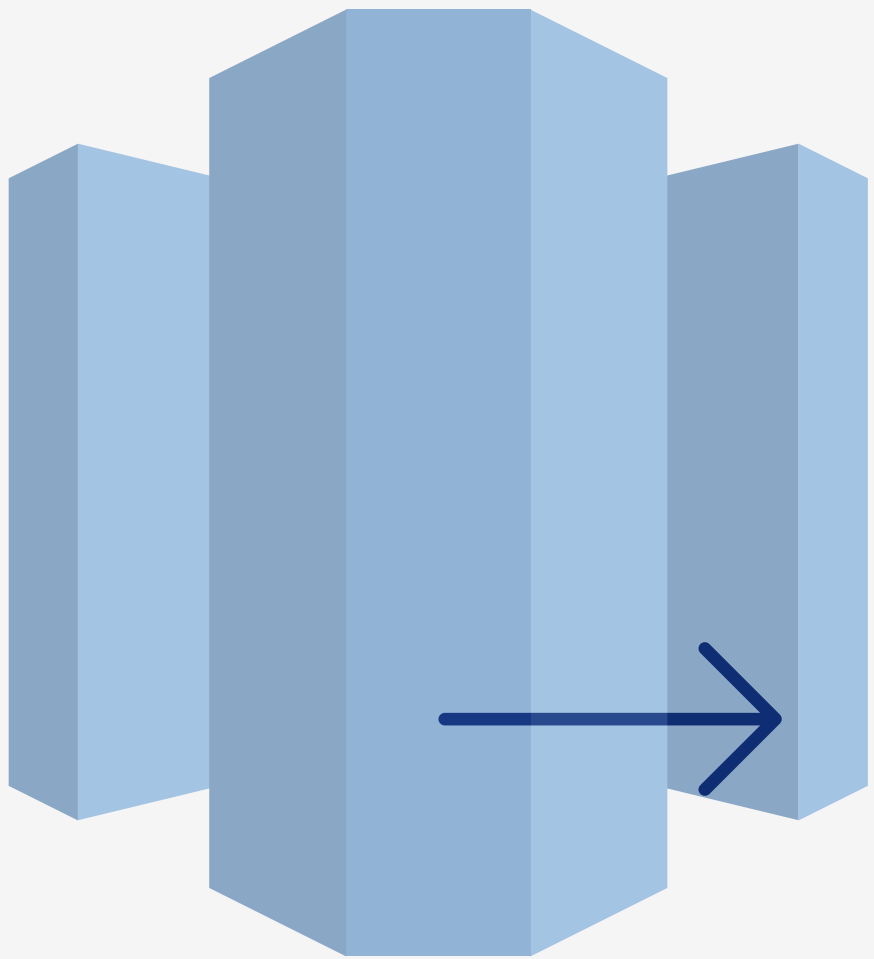
47

Query Caching

Redshift caches the results of queries to improve performance for repeated queries by storing the results and serving them directly when the same query is executed again.

Query results are cached by default. Use the EXPLAIN command to see if a cached result is being used.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

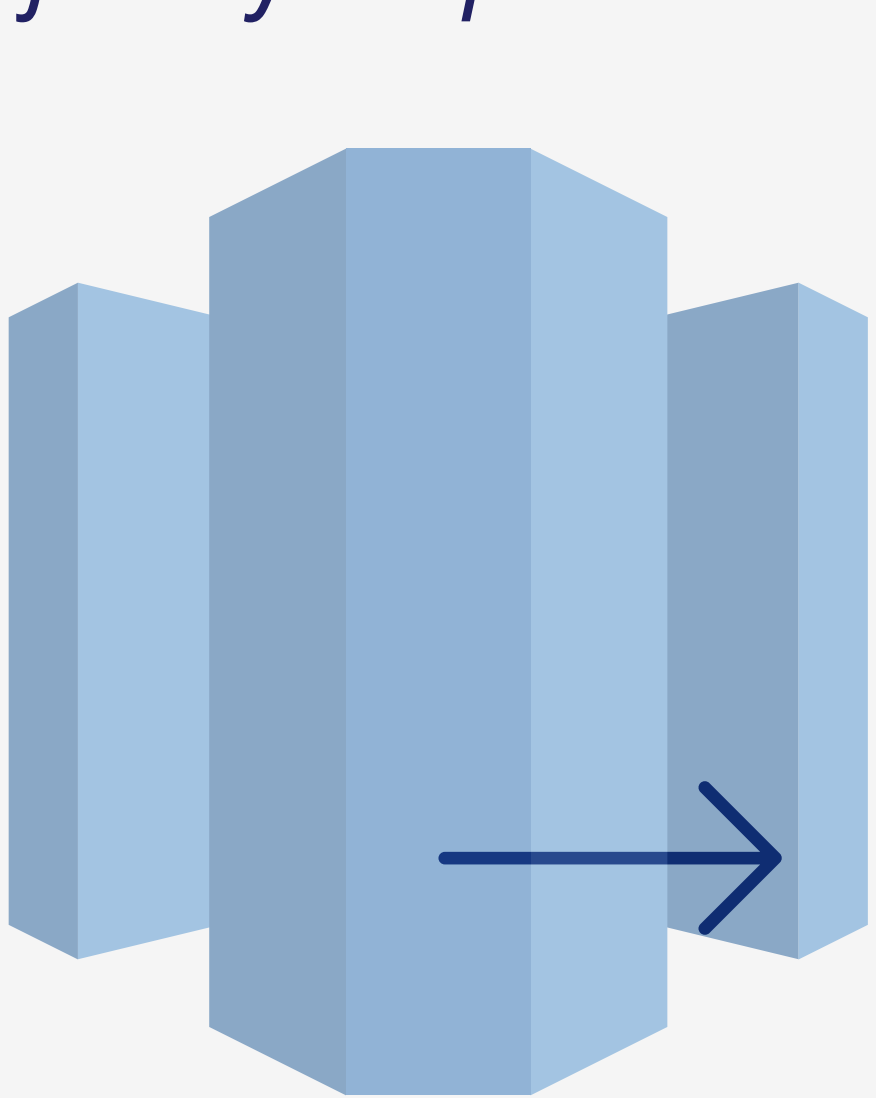
48

Manual Snapshots

Manual snapshots are user-initiated backups of your Redshift cluster that can be retained for an indefinite period, allowing you to restore the cluster to a specific point in time.

Create a manual snapshot: aws redshift create-cluster-snapshot --snapshot-identifier my-snapshot --cluster-identifier my-cluster

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

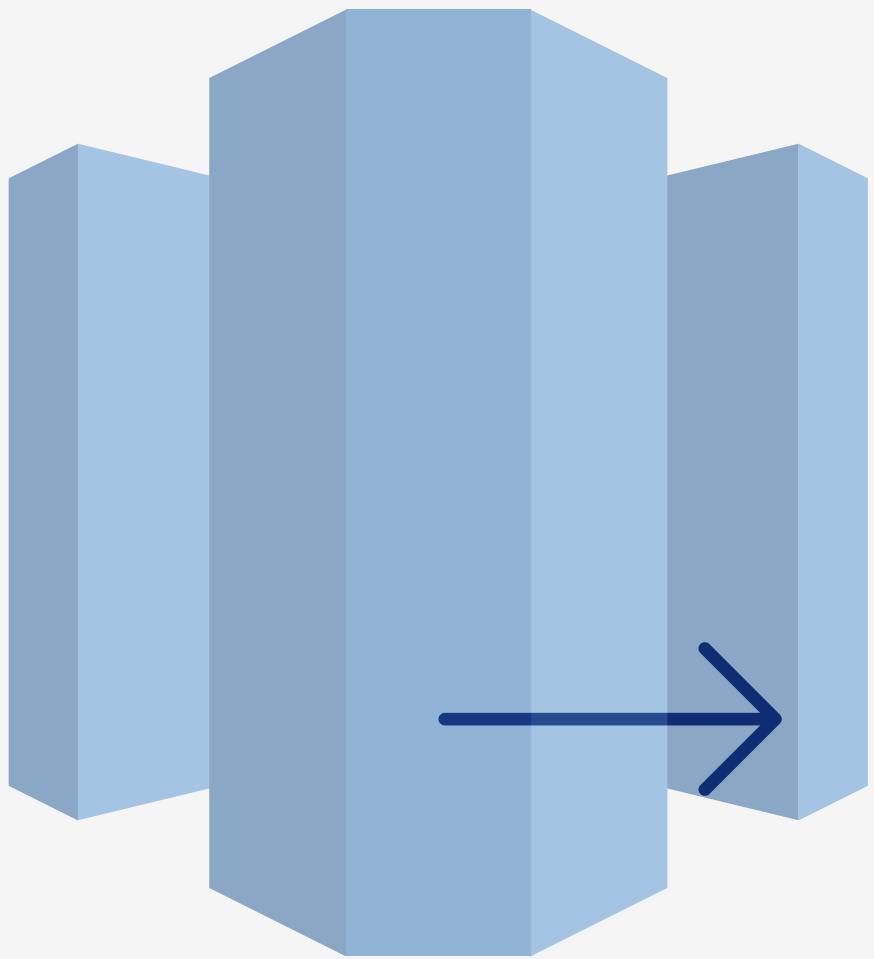
49

Database User Management

Redshift allows you to create, manage, and delete database users and groups, controlling access to data and operations within the cluster.

CREATE USER myuser WITH PASSWORD 'mypassword'; GRANT SELECT ON mytable TO myuser;

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

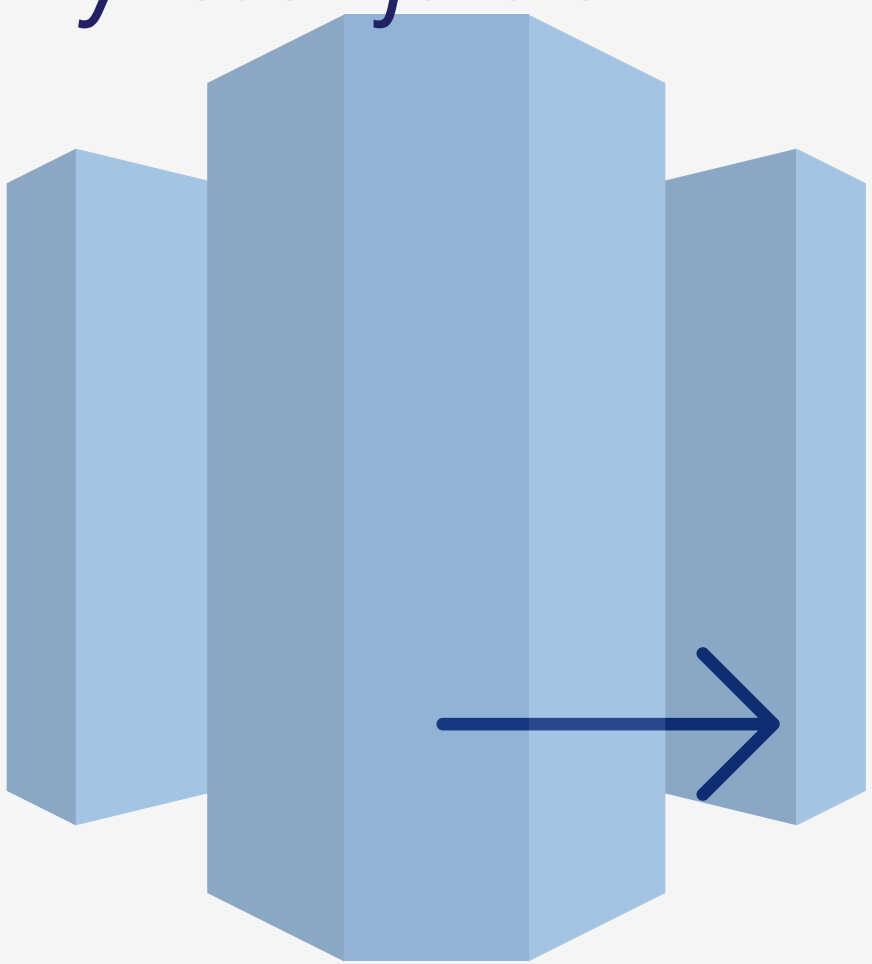
50

IAM Role Integration

Redshift integrates with AWS IAM roles to allow fine-grained access control to AWS services, enabling secure data access and operations within Redshift.

aws redshift create-cluster --cluster-identifier my-cluster --iam-roles arn:aws:iam::123456789012:role/MyRedshiftRole

Shwetank Singh
GritSetGrow - GSGLearn.com



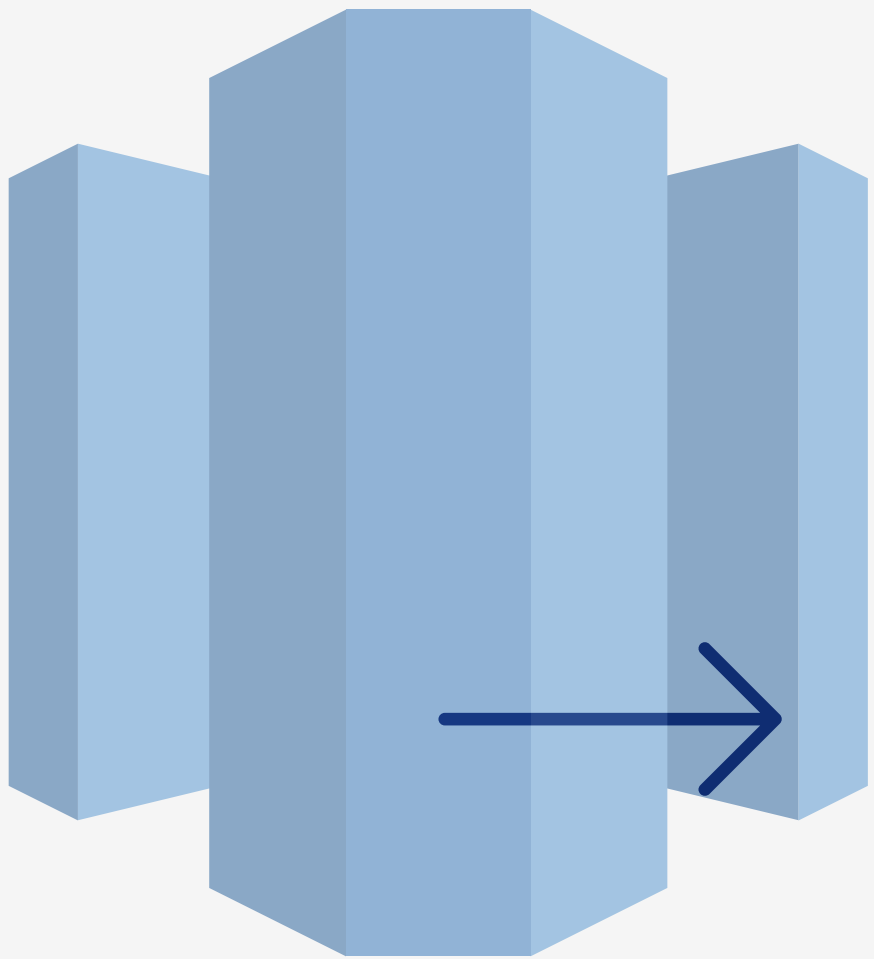
Data Engineering 101: Redshift

Federated Authentication

Federated authentication allows you to authenticate Redshift users with credentials from other identity providers, such as Microsoft AD or AWS Cognito.

Set up federated authentication with SAML 2.0 integration for your Redshift cluster.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

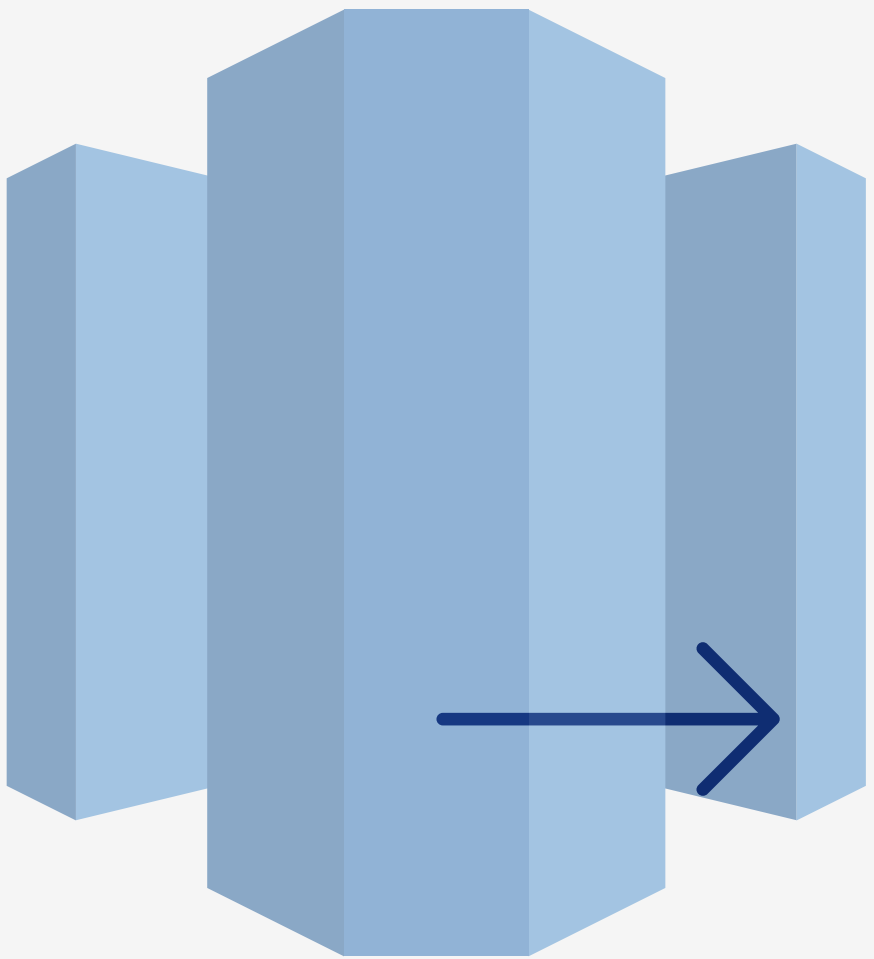
57

Automatic WLM Tuning

Redshift can automatically tune your Workload Management (WLM) settings to optimize query performance based on historical query patterns and workload characteristics.

Enable automatic WLM tuning in the Redshift console or using the AWS CLI.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

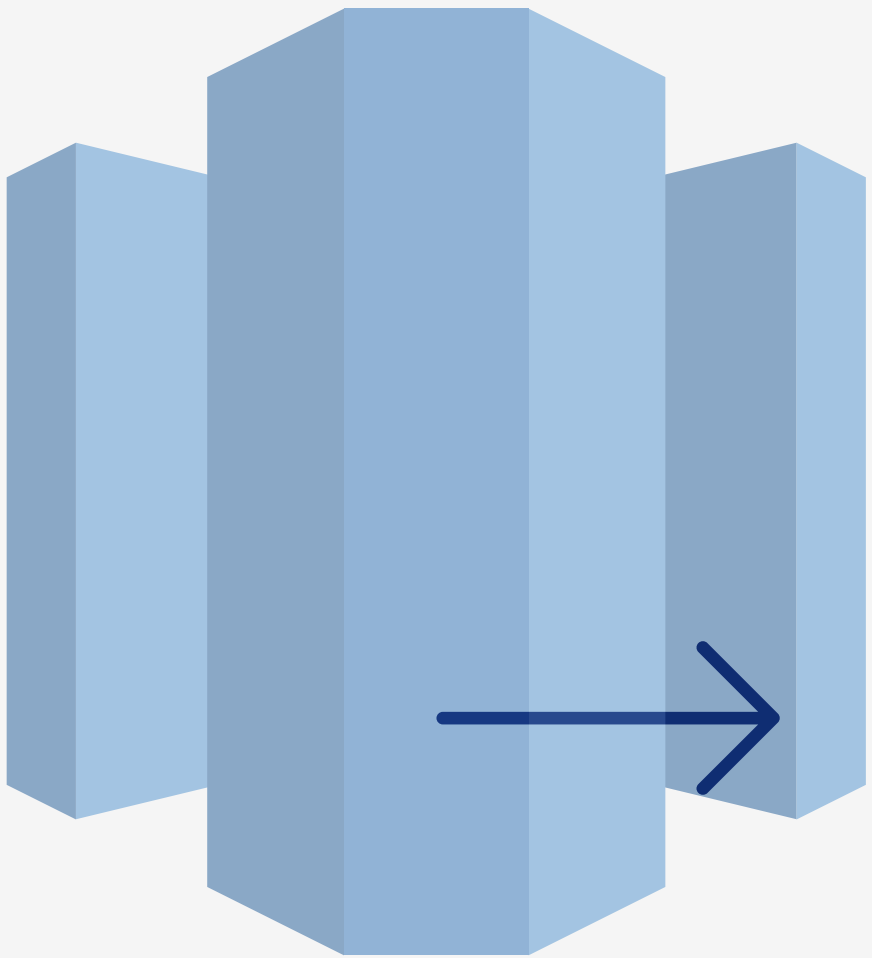
53

Cluster Maintenance

Redshift performs regular maintenance on your clusters during predefined maintenance windows to apply updates, patches, and fixes.

Configure maintenance window: aws redshift modify-cluster --cluster-identifier my-cluster --preferred-maintenance-window sun:05:00-sun:05:30

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

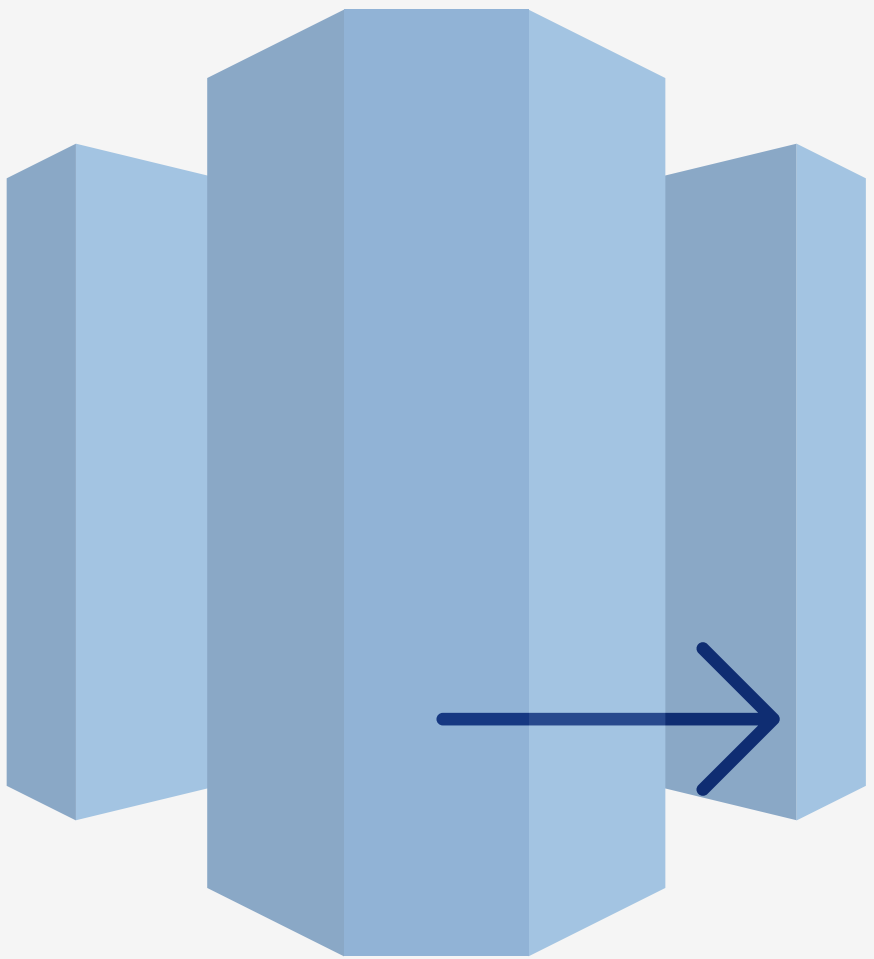
Database Auditing

54

Redshift supports auditing database activities, allowing you to track changes to database configurations, access controls, and query execution for compliance and security.

Enable and configure database auditing in your Redshift cluster settings.

Shwetank Singh
GritSetGrow - GSGLearn.com



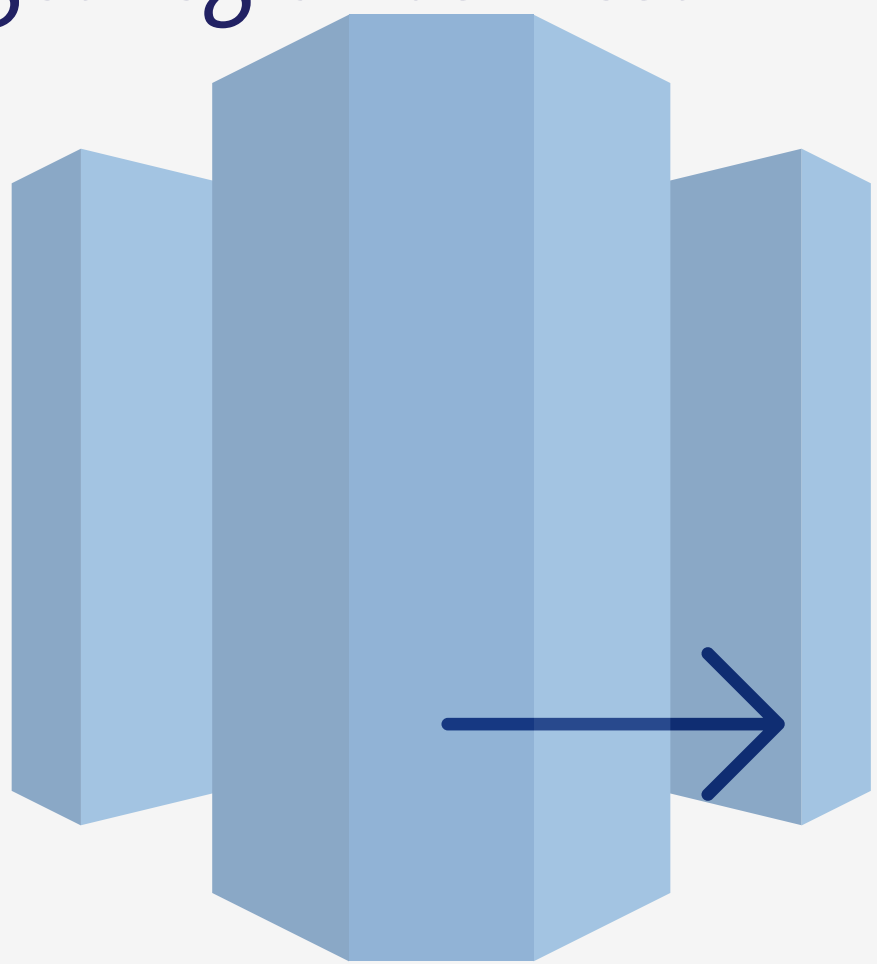
Data Engineering 101: Redshift

Cross-Region Snapshots

Cross-region snapshots enable you to copy your Redshift snapshots to another AWS region, providing disaster recovery and backup capabilities across regions.

aws redshift copy-cluster-snapshot --source-snapshot-identifier my-snapshot --target-snapshot-identifier my-snapshot-copy --target-region us-west-2

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

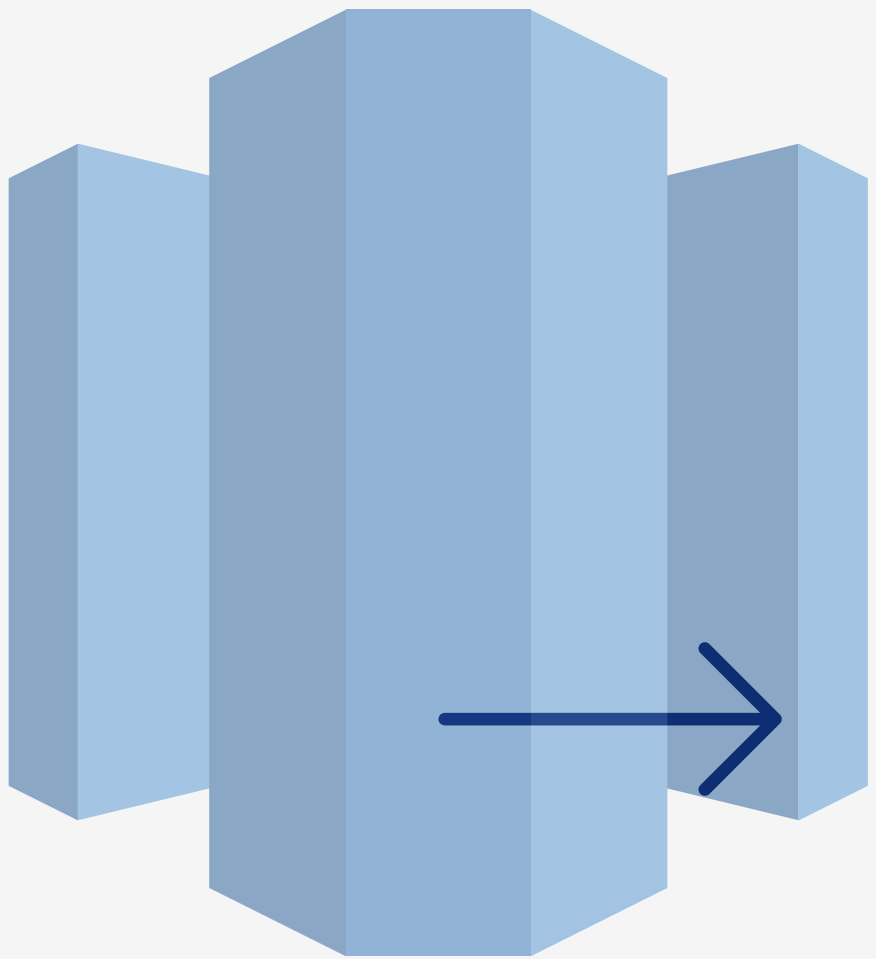
56

Performance Insights

Performance Insights provide a dashboard to visualize and monitor the performance of your Redshift cluster, helping identify and resolve performance bottlenecks.

Enable Performance Insights in the Redshift console to start monitoring your cluster.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

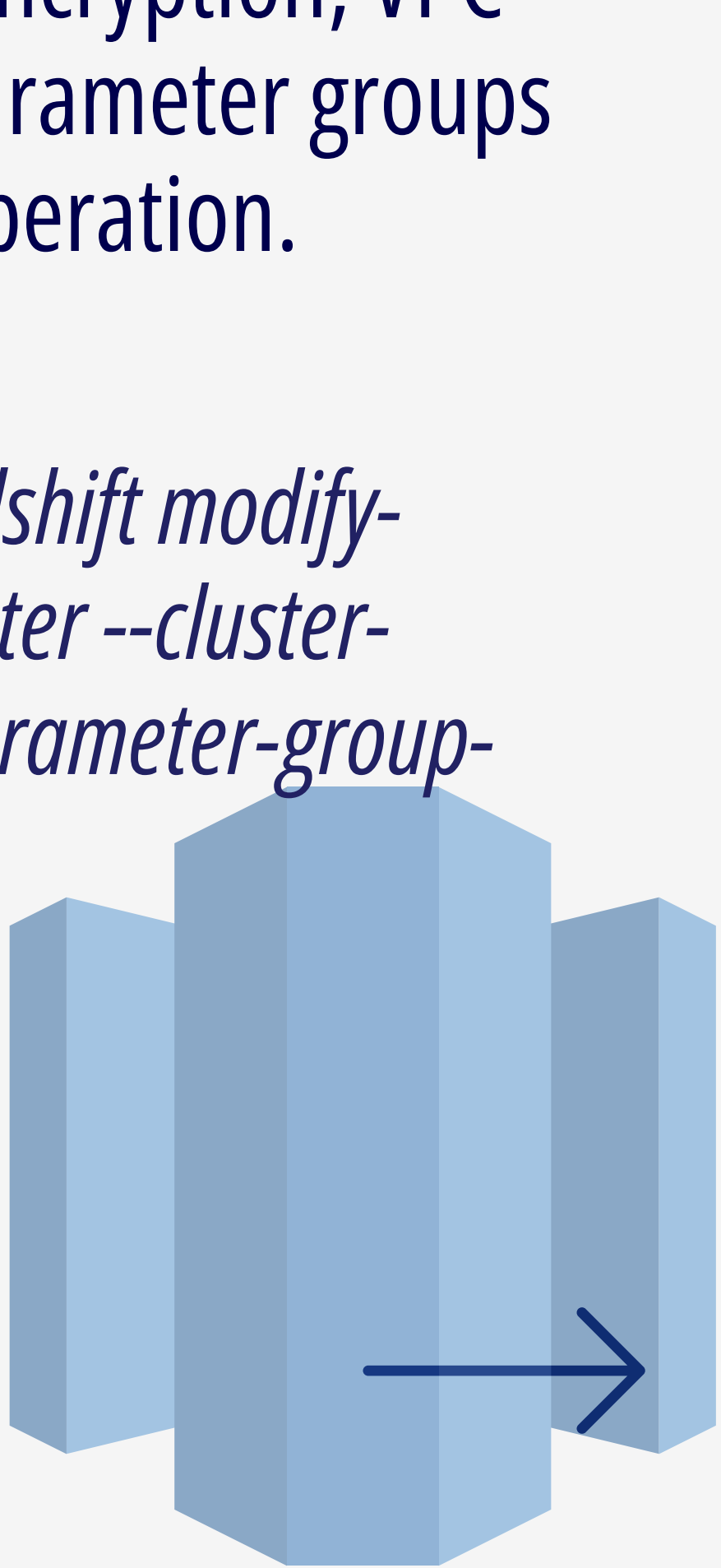
57

Cluster Security Configuration

Redshift clusters can be configured with security features such as SSL encryption, VPC security groups, and cluster parameter groups to ensure secure access and operation.

Configure SSL encryption: aws redshift modify-cluster --cluster-identifier my-cluster --cluster-security-groups sg-12345678 --parameter-group-name my-parameter-group

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

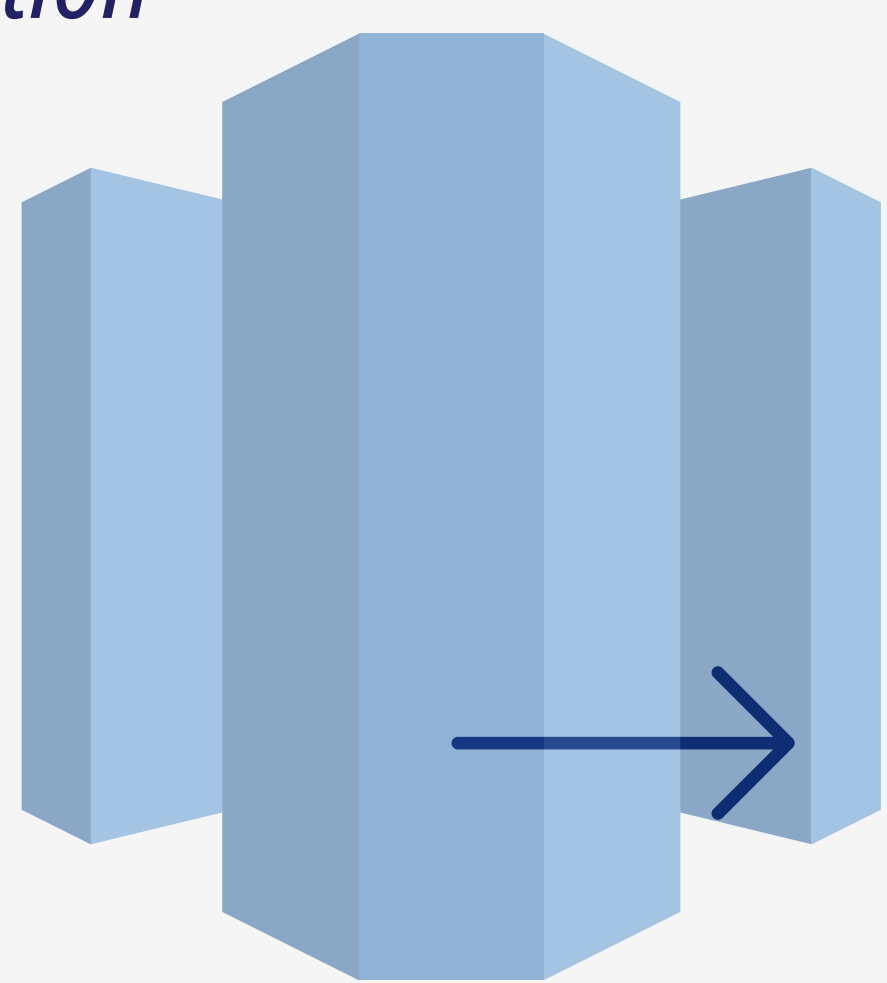
58

IAM Database Authentication

IAM Database Authentication allows you to use IAM credentials to authenticate to your Redshift database, simplifying the management of database access.

Enable IAM Database Authentication: aws redshift modify-cluster --cluster-identifier my-cluster --enable-iam-database-authentication

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

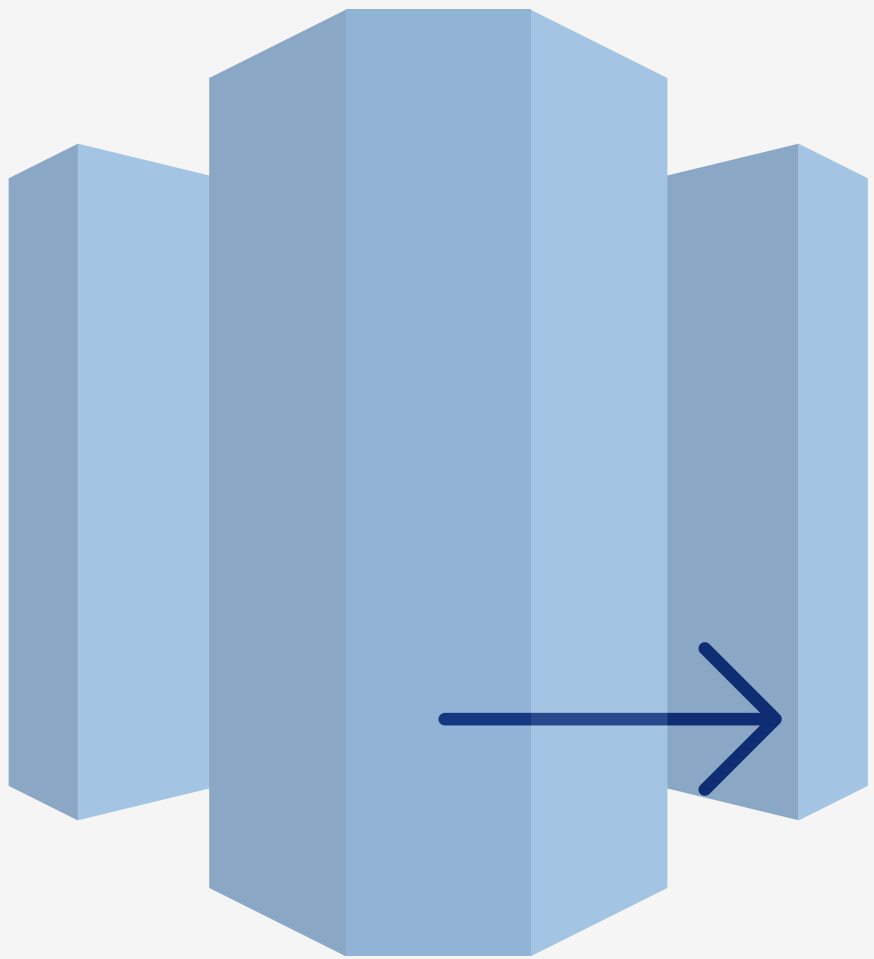
Data Transfer Accelerator

59

Redshift's data transfer accelerator speeds up data transfers between your S3 buckets and Redshift, reducing the time required for large-scale data imports and exports.

Enable data transfer accelerator in your Redshift configuration settings.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

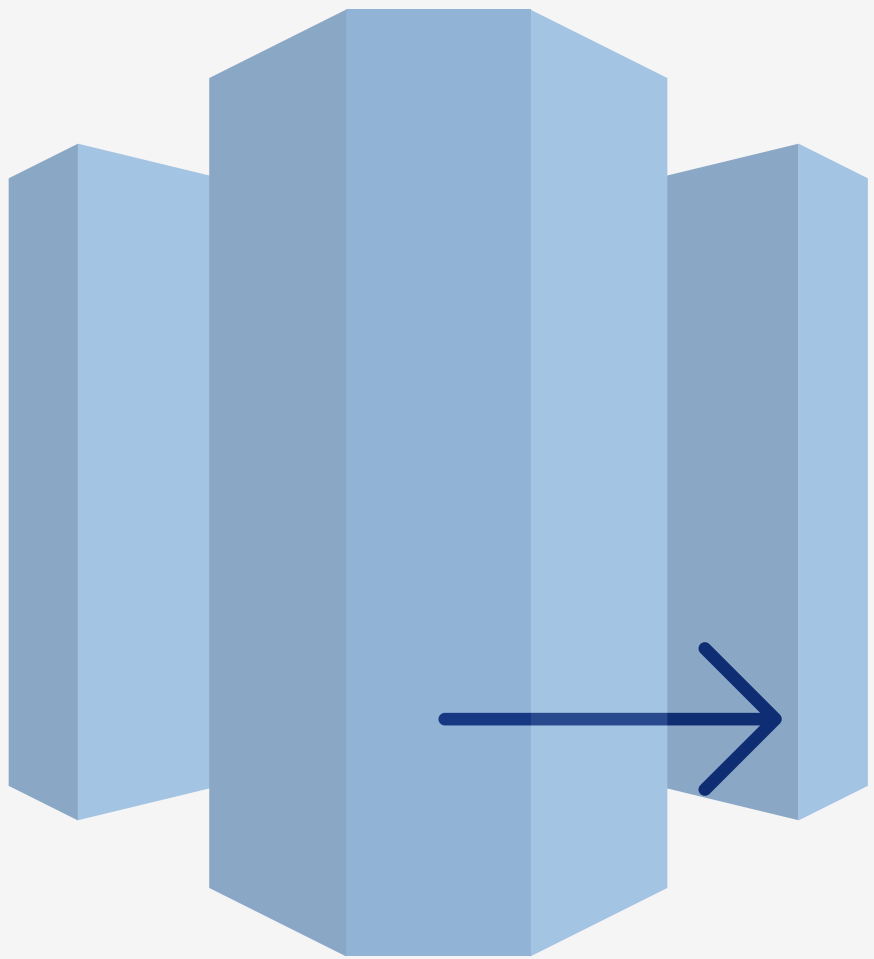
60

Query Optimizer

The Redshift query optimizer analyzes and optimizes SQL queries for performance, ensuring efficient use of resources and quick query execution times.

Use the EXPLAIN command to see the query execution plan and optimization strategies applied by the optimizer.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

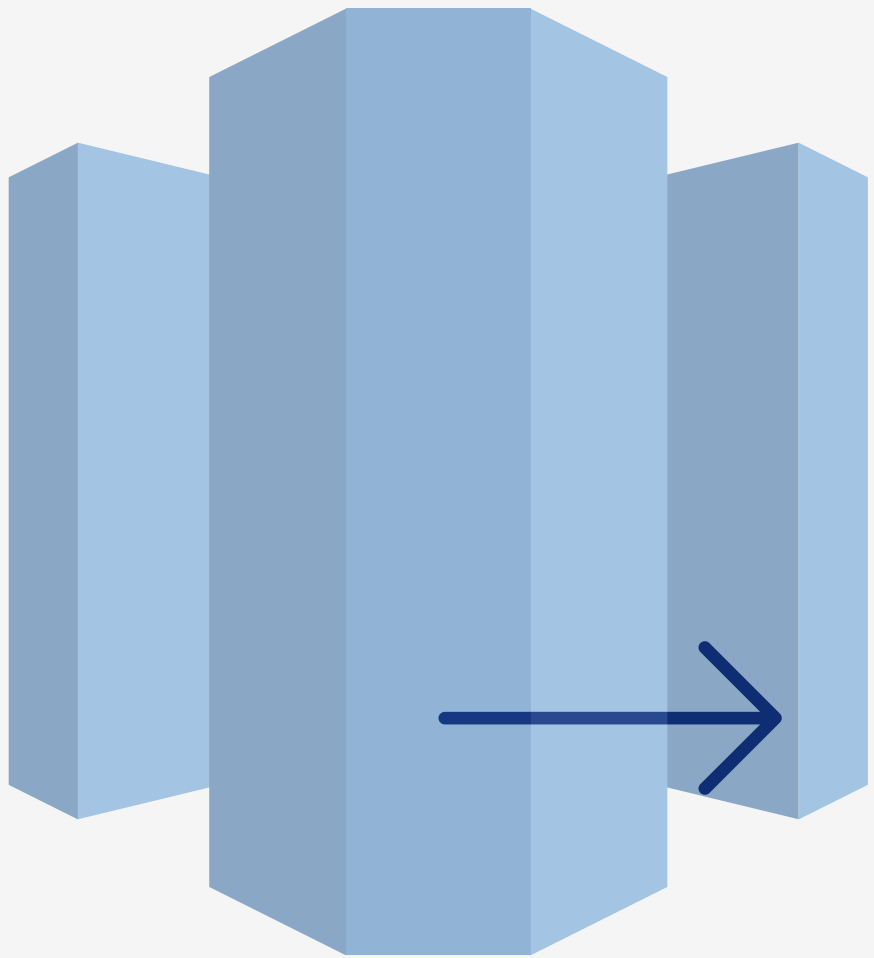
61

Instance Hour Billing

Redshift billing is based on instance hours, which are the number of hours your cluster's nodes are running. Costs depend on the instance type and region.

Monitor instance hour usage: Use the AWS Cost Explorer to view instance hour billing details for your Redshift cluster.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

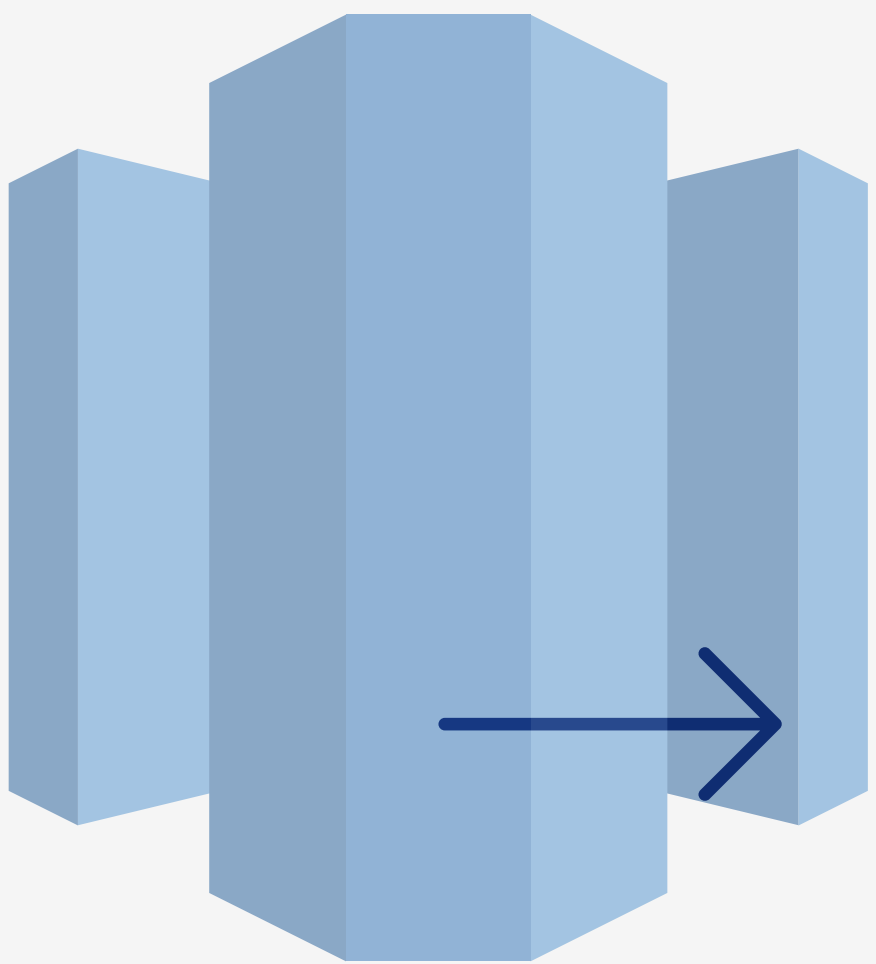
62

Data Lake Integration

Redshift integrates with AWS Data Lake services, allowing you to query and analyze data stored in the data lake without moving it into Redshift.

Set up a Data Lake integration with Redshift Spectrum to query S3 data without loading it into the cluster.

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

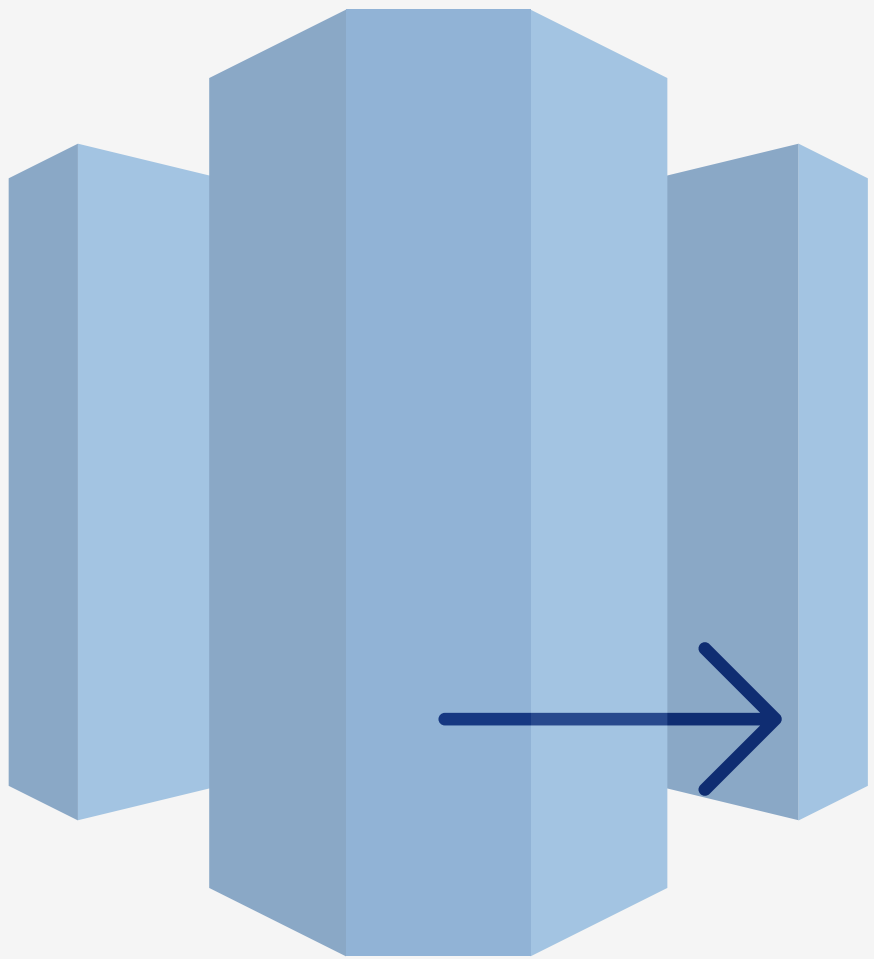
63

Database Audit Logging

Audit logging in Redshift captures logs of database activities, including connections, disconnections, and SQL queries, for security and compliance monitoring.

Configure audit logging: aws redshift enable-audit-logging --cluster-identifier my-cluster --bucket-name my-log-bucket

Shwetank Singh
GritSetGrow - GSGLearn.com



Data Engineering 101: Redshift

64

Lambda Integration

Amazon Redshift can invoke AWS Lambda functions from within SQL queries, allowing you to perform complex processing or integrate with other AWS services.

Use Lambda UDFs: ČREATE FUNCTION mylambda_udf() RETURNS float AS 'arn:aws:lambda:

Shwetank Singh
GritSetGrow - GSGLearn.com