Amazon Redshift Warehouse on Cloud
Part 2

RAJAS WALAVALKAR

Associate Solution Architect - Quantiphi
&
Community Builder

11th JUNE
3:00 pm

Live stream on AWS UG INDIA

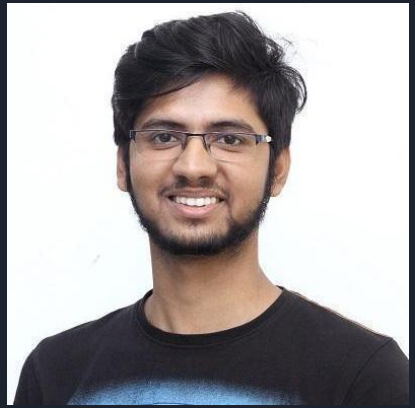
AWS
User Groups
Mumbai

@awsugmum

Introduction

Rajas Walavalkar - Associate Solution Architect at Quantiphi Analytics & AWS Community Builder

Expertised in AWS Data Analytics Services, Big Data Tools, Data Lake, Warehouse and Business Intelligence & Dashboarding tools



Linkedin - https://in.linkedin.com/in/rajas-walavalkar-4ab21b15b

Medium - rajaswalavalkar.medium.com

Agenda

1. Redshift Distribution Keys and Sort Key
    a. Significance of Dist keys & Sort Keys
    b. Types of Dist Keys and their use cases

2. Redshift Serverless Introduction

3. Redshift Spectrum
    a. Redshift Spectrum Overview
    b. Glue Catalog and Spectrum Integration
    c. Use Cases for Redshift Spectrum

4. Hands-on
    a. Create an External Table (Spectrum table)
    b. Create Materialized view to join local & spectrum tables

5. Redshift Features
    a. Concurrency Scaling
    b. Workload Management Overview & Example

1. Redshift Distribution Keys & Sort Keys

What are DIST Keys and SORT Keys?

Amazon Redshift DISTKEY and SORTKEY are a powerful set of tools for optimizing query performance.

Because Redshift is a columnar database with compressed storage, it doesn't use indexes that way a transactional database such as MySQL or PostgreSQL would. Instead, it uses DISTKEYs and SORTKEYs to optimize the query performance

Significance of DIST Keys and SORT Keys?

DIST Keys determine where data is stored in Redshift.

Sort Key determines the order in which rows in a table are stored

Query performance suffers when a large amount of data is stored on a single node

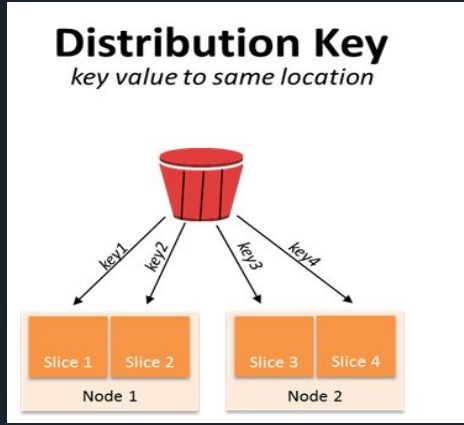
The query optimizer uses this sort Keys while determining optimal query plans

Distribution Styles

1. KEY distribution Style:

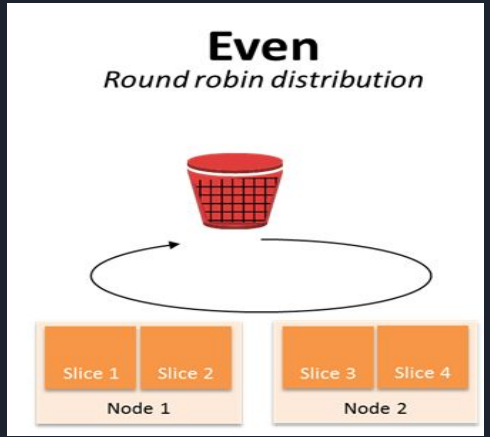
   The data is distributed across slices by the leader node matching the values of a designated column **(DIST KEY)**. So all the entries with the same value in the column end up in the same slice.

   Note : It is beneficial to select a KEY distribution if a table is used in JOINS.



2. EVEN distribution Style:

   In Even Distribution the Leader node of the cluster distributes the data of a table evenly across all slices, using a round-robin approach.

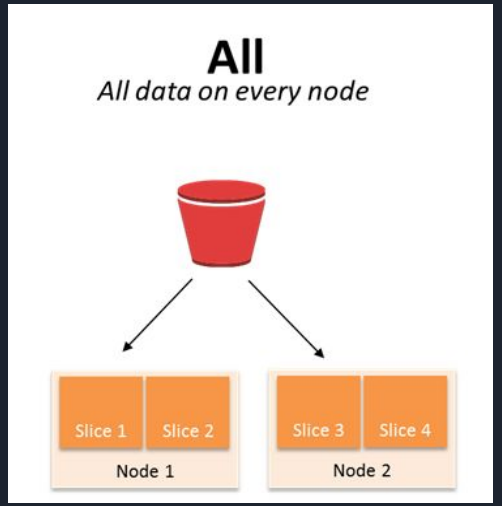


Distribution Styles

3. ALL distribution Style:

- The leader node maintains a copy of the table on all the computing nodes resulting in more space utilization.
- Since all the nodes have a local copy of the data, the query does not require copying data across the network. This results in faster query operations.
- The negative side of using ALL is that a copy of the table is on every node in the cluster. This takes up too much space and increases the time taken by Copy command to upload data into Redshift

Note : Choose ALL styles for small tables that do not often change. For example, a table containing telephone ISD codes against the country name.



4. AUTO distribution Style:

In this, Redshift defines the distribution style for the table depending on the past history of usage & size of tables. It will also update the distribution style in future depending changes in size of data

2. Redshift Serverless & Redshift Spectrum

Redshift Serverless

What is the need of Serverless?

- We’re seeing the use of data analytics expanding among new audiences within organizations, like developers and line of business analysts who don’t have the expertise or the time to manage a traditional data warehouse

FEATURES

There is no need to set up and manage clusters.

It makes super easy to run analytics in the cloud with high performance at any scale.

You pay for the duration in seconds when your data warehouse is in use

Example - Pay on when you are querying or loading data. There is no charge when your data warehouse is in idle state

Just load your data and start querying.

Redshift Spectrum Overview

Amazon Redshift Spectrum is a feature that lets a data analyst conduct fast, complex analysis on objects stored on the AWS S3 Buckets without loading the data on the Redshift Cluster

This can save time and money because it eliminates the need to move data from a storage service to a database, and instead directly queries data inside an S3 bucket in its raw format

FEATURES

Scalable without Manual Intervention

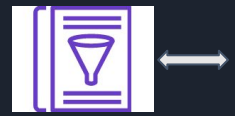
No ETL required - Query Data in its raw format

On Demand & Pay Per Pricing

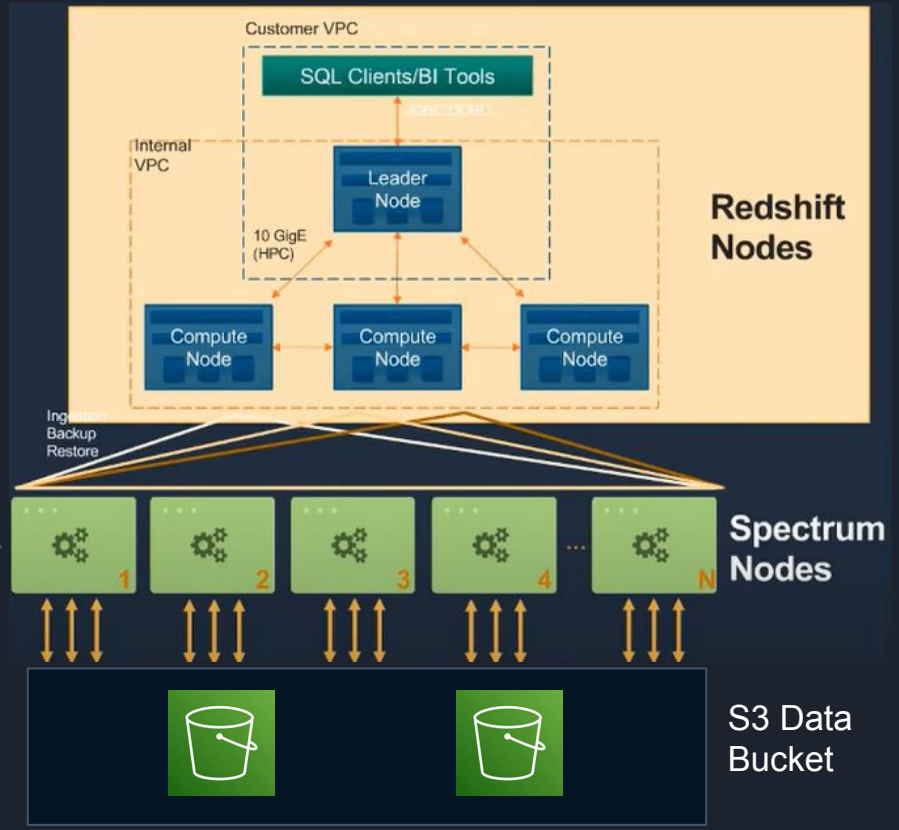
High Concurrency - Multiple cluster can access the data from same source concurrently

Redshift Spectrum Architecture

- Spectrum Nodes are a part of internal architecture, as a user we are not responsible to deploy these spectrum nodes
- Query Projections, Filters, Aggregations & Joins are pushed down to Spectrum nodes for S3 Data
- Spectrum can leverage Glue Catalog for getting the schema of the raw data on S3
- Cost of Spectrum Queries are based on the data pulled from S3 data lake and not on the computation done (1 TB of data = $5)



Glue Data Catalog



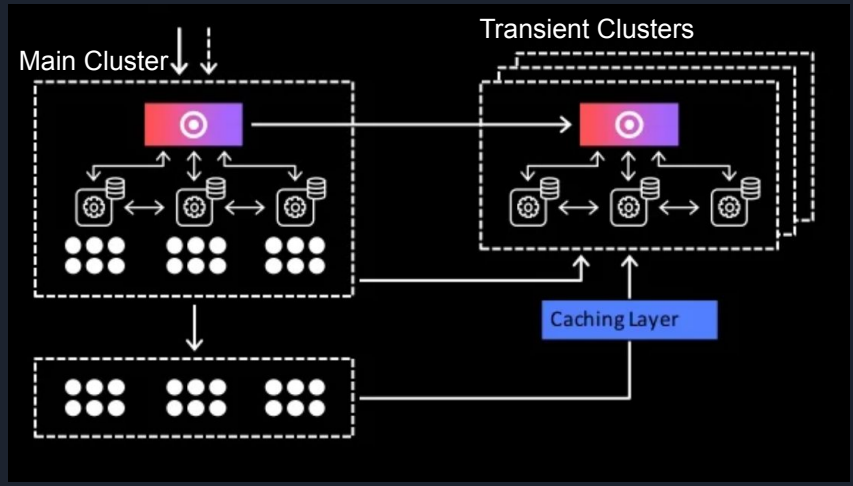
4. Let’s do some Hands-on

5. Redshift IMPORTANT Features

Concurrency Scaling

Amazon Redshift automatically deploys and removes compute nodes as needed to serve the changing requirements and quey loads

It basically automatically adds a transient clusters in seconds to serve a sudden spike in concurrent requests



How does it Work?

- When Queries in the WLM queue begins to wait, then Redshift routes them to new concurrent transient clusters
- Once the Queries are completed the transient cluster is deleted automatically

Note: *Amazon provides* ***one free hour of*** *concurrency scaling option if your cluster is running for* ***24 hours****. This means that in month you can get* ***30 Hours of free*** *concurrency scaling if the cluster is running all the time*

Workload Management (WLM)

Amazon Redshift provides WLM feature to manage the different workloads of queries running on Redshift

There are majorly two types of WLM which Redshift provides

1. Manual WLM

   In manual WLM the end user has the ability to manage and configure the Memory allocations and the query concurrency limit for the queues

2. Auto WLM

   While in Auto WLM Redshift cluster automatically detects the amount of memory required for a query to execute and accordingly adjusts the concurrency query limit as per the configuration.

How to decide which WLM queue to configure….?

Walkthrough of Workload Management

WLM

AWS
User Groups

Thank you!

AWS
User Groups