Amazon Redshift

Introduction to Amazon Redshift

Speaker Name
Speaker title, Company

Discussion Topics

- Redshift Overview
- Getting Started
- Autonomics
- Scalability
- Durability
- Security
- New Features

Redshift Overview

Typical use cases

Nasdaq Pfizer

Traditional Data Warehousing

- Mid-Market, Enterprise Customers, Large established customers
- Deliver the same compatibility at a vastly lower price

intermix.io Avazu

Software as a Service / Analytics

- Deploying a new application with embedded analytics

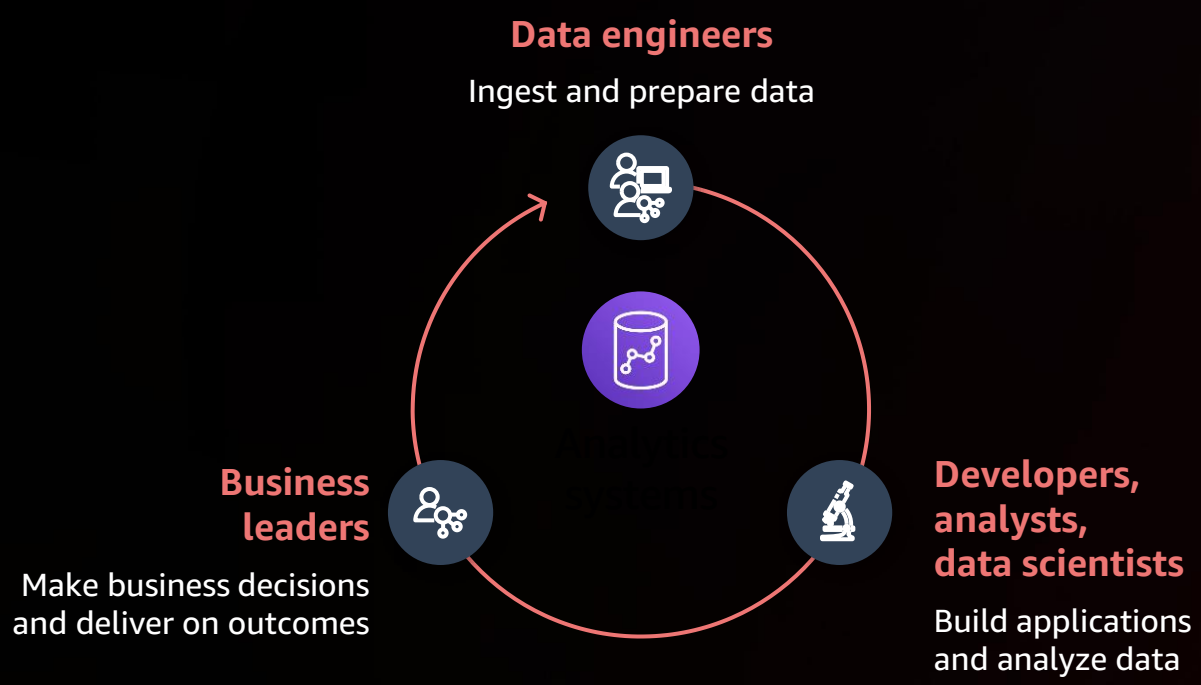
Pinterest yelp

Big Data Analytics

- BI Reporting Analytics
- Variety and volume of data coming at a high velocity – streaming data
- Requirement to store and analyze in a relational format

Easy analytics for everyone

FOCUS ON GETTING FROM DATA TO INSIGHTS IN SECONDS



Automatic provisioning and scaling

Automatically provisions and scales the underlying compute resources to deliver high performance for demanding and unpredictable workloads

Visualize your data

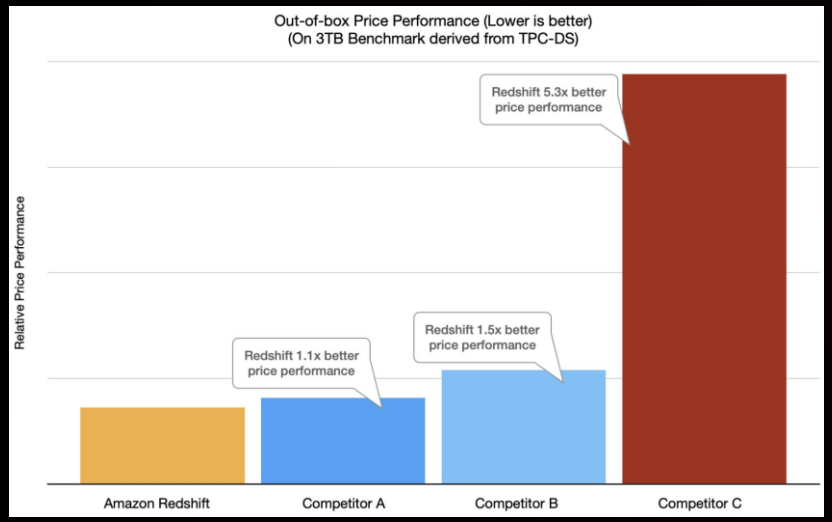
Single, visual interface for querying data to improve productivity through one-click visual analytics, collaboration, version control, and scheduling

Bypass administrative tasks

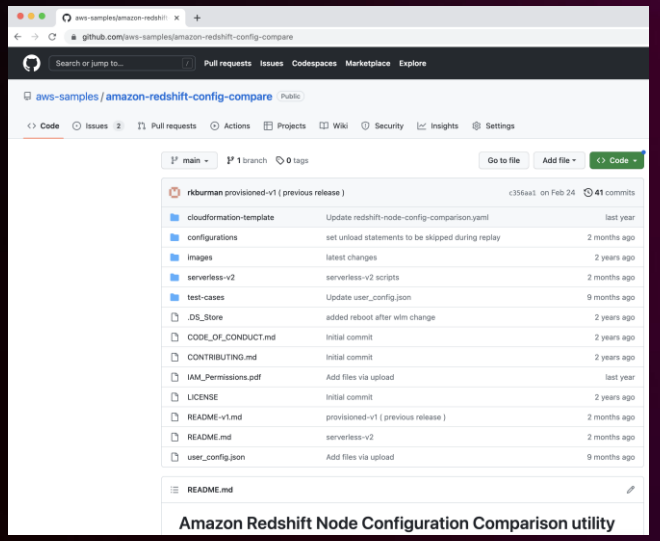
Take advantage of automated provisioning, backup, patching, tuning, and monitoring in Amazon Redshift

Amazon Redshift Price Performance

Best price performance at any scale



Amazon Redshift delivers up to 5x better price performance than other cloud data warehouses and up to 7x better price-performance on high concurrency, low latency workloads.



Amazon Redshift **Node Configuration Comparison utility** replays your workload on various configurations to optimize cost and performance.

https://aws.amazon.com/blogs/big-data/amazon-redshift-continues-its-price-performance-leadership/
https://github.com/aws-samples/amazon-redshift-config-compare

Get Started with Redshift

Redshift cluster architecture

Leader node

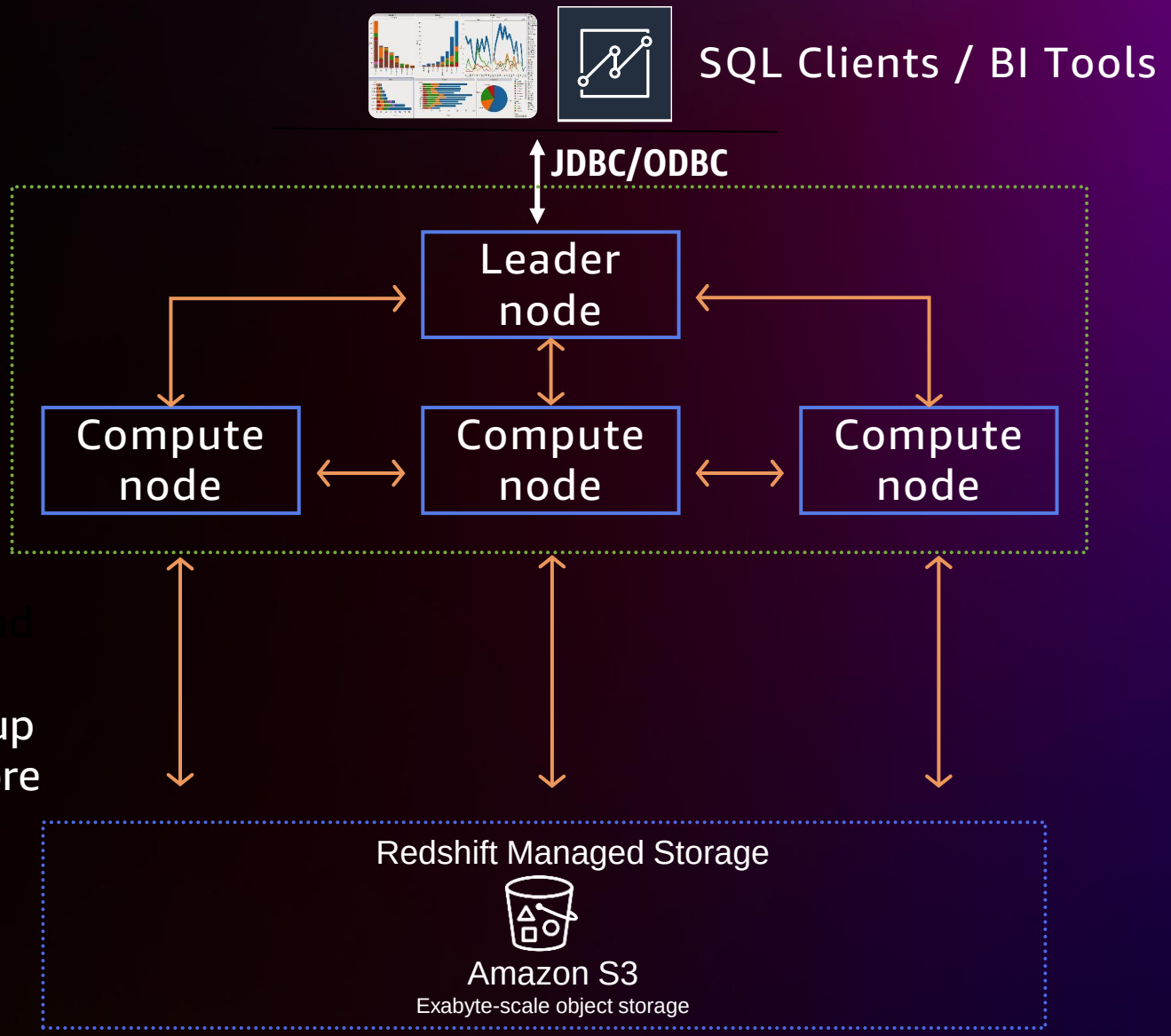
- SQL endpoint
- Stores metadata
- Coordinates parallel SQL processing &
- ML optimizations
- Leader node is no-charge for clusters with 2+nodes

Compute nodes

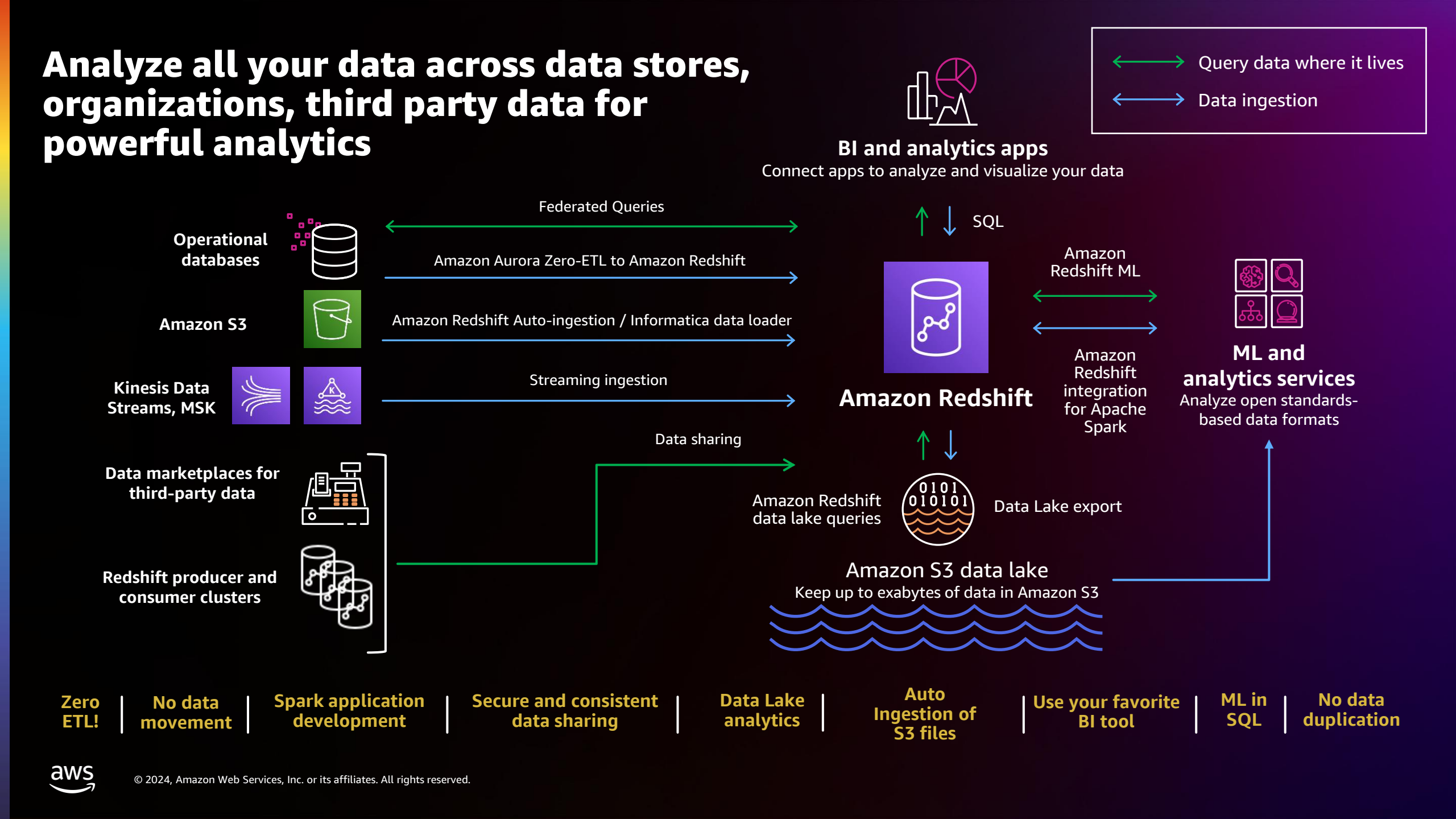
- Split into “Slices”
- Local SSDs for caching
- Executes queries in parallel
- Load, unload, backup, restore from S3

Redshift Managed Storage

- Resides in S3
- Available across entire Region
- Pay for space used (not provisioned)
- Scales independently of Compute



Analyze all your data across data stores, organizations, third party data for powerful analytics



Zero ETL! | No data movement | Spark application development | Secure and consistent data sharing | Data Lake analytics | Auto Ingestion of S3 files | Use your favorite BI tool | ML in SQL | No data duplication

Amazon Redshift innovates to meet your needs

Easy analytics for everyone

- Serverless
- Query editor v2
- Automated DW management
- Automatic materialized views
- Data API
- Amazon Redshift Advisor
- AWS CloudFormation templates
- Grafana Plugin
- Multi-AZ Deployment

Analyze all your data

- Data sharing
- AWS Data Exchange integration
- Amazon Redshift ML
- Federated query
- Geospatial enhancements
- SUPER data type with JSON
- Redshift Streaming Ingestion
- Aurora Zero ETL with Redshift
- Apache Spark Connector

Best price performance at any scale

- RA3 nodes & managed storage
- Concurrency scaling for reads and writes
- SQL enhancements & migration support
- Security, governance & compliance
- Workload management enhancements
- Auto Copy from S3
- Data Sharing Access Control via AWS Lake Formation
- Dynamic Data Masking

Amazon Redshift

Provisioned

- Fully managed, petabyte data warehouse in the cloud
- Set of nodes called a Cluster
- Budget control with Reserved Instances discounted pricing

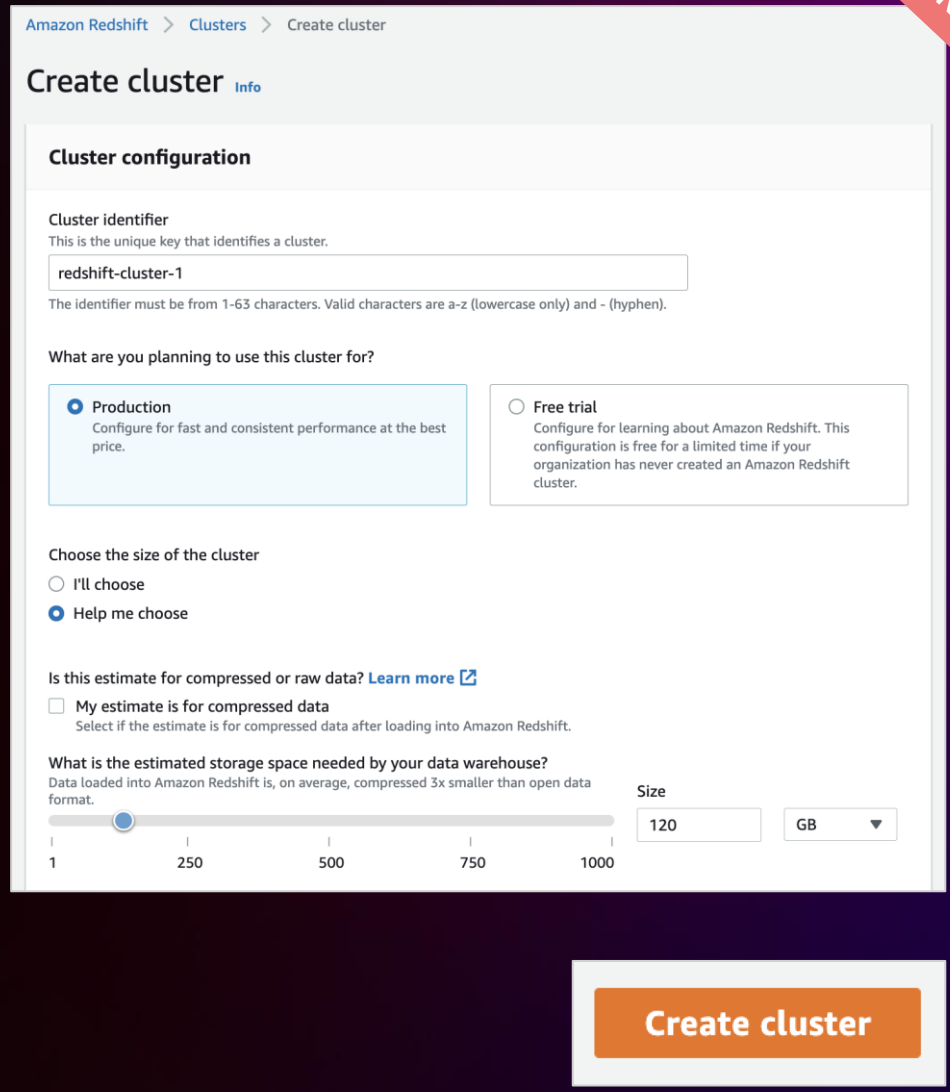
Serverless

- Fully managed, petabyte data warehouse in the cloud
- Intelligent scaling without thinking about servers
- Pay-for-use with RPU hour rates

Three steps to launch

1. Select Redshift from the AWS Service Console
2. Click the “Create cluster” button
3. Use the Quick Launch wizard to select the size of your dataset and enter basic information such as instance type, number of nodes, IAM role to use, etc.

PROVISIONED



Redshift instance types

PROVISIONED

Amazon Redshift RA3 current generation)

- Solid-state disks + Amazon S3
- Amazon Redshift Managed Storage (RMS)

Dense compute DC2

- Solid-state

A Redshift cluster can have up to 128 ra3.16xlarge nodes (16 PB of managed storage) and can support EBs of data with its Redshift Data Lake support.

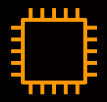
	Instance type	Disk type	Size	Memory	# CPUs	# Slices
Scale compute independent of storage	RA3	RA3 xlplus	RMS	Scales to 32 TB	32 GIB	4	2
		RA3 4xlarge	RMS	Scales to 128 TB	96 GIB	12	4
		RA3 16xlarge	RMS	Scales to 128 TB	384 GIB	48	16
Classic MPP	Compute Optimized	DC2 large	SSD	160 GB	16 GIB	2	2
		DC2 8xlarge	SSD	2.56 TB	244 GIB	32	16

Redshift Serverless

Compute separated from storage

SERVERLESS

WORKGROUP



- New normalized compute unit – Redshift Processing Unit (RPU)
- Usage is billed in RPU-hours, metered on per-second basis
- Base data warehouse, scaling capacity, data lake queries are part of same RPU-hours

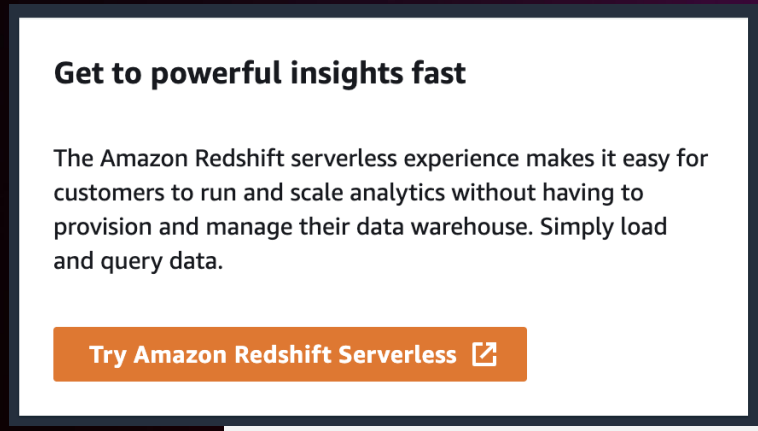
NAMESPACE

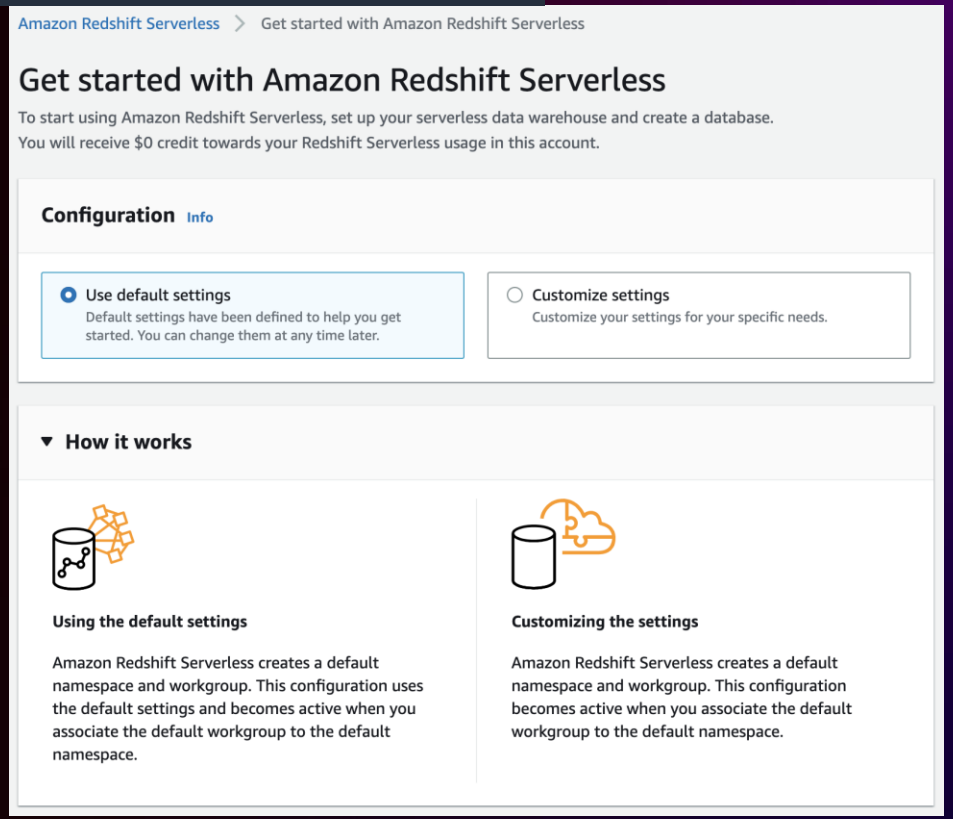
- Fixed GB-month rate pricing for the Redshift managed storage and user snapshots
- Restore their data warehouse to specific points in last 24 hours at a 30 min granularity at free of charge

Three steps to launch

1. Try Amazon Redshift serverless for your AWS account
2. Review default configuration
3. Connect from your favorite tools or Amazon Redshift Query Editor v2

SERVERLESS





Amazon Redshift Serverless

SERVERLESS

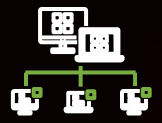
Simplified user experience

Run and scale analytics without having to manage data warehouse clusters

Intelligent and dynamic compute

Automatically provisions and scales data warehouse capacity to deliver consistently fast performance

All Redshift functionality and performance



Leverage Amazon Redshift's rich SQL capabilities, seamless data lake integration, as well as industry-leading price performance at scale

Pay for use

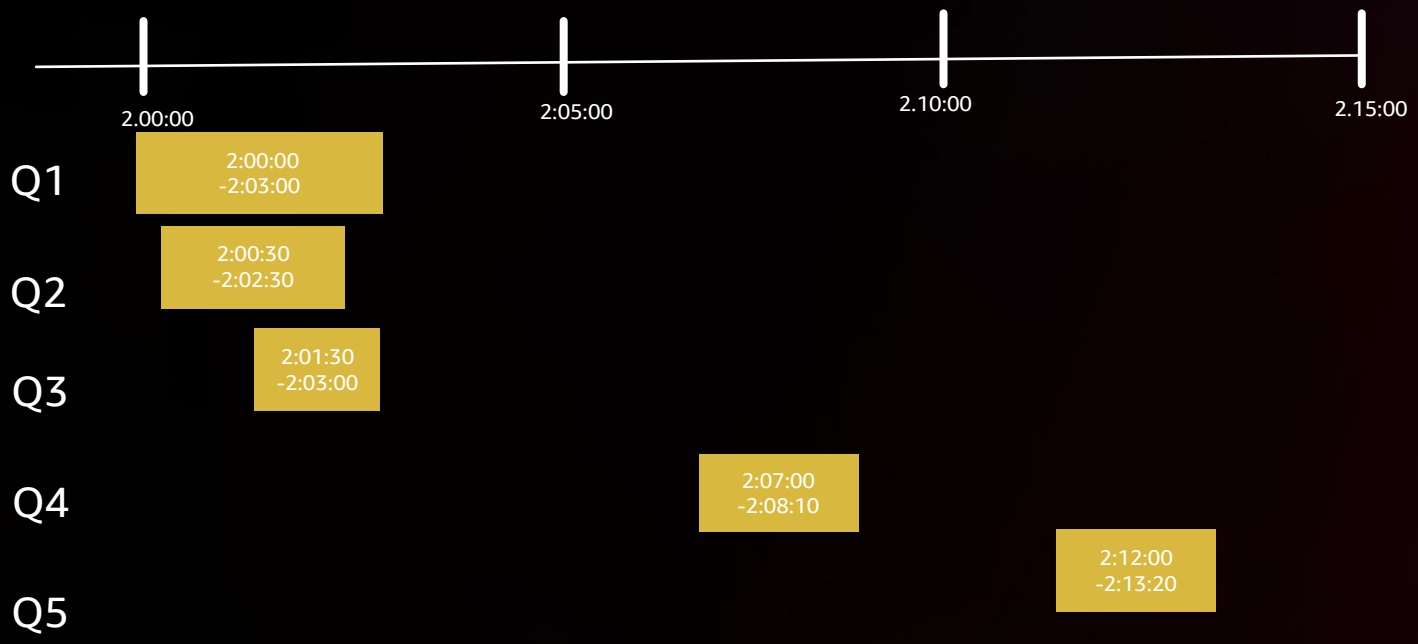
Pay for the compute capacity only for the workload duration on a per-second basis. No charges for idleness

aws

Pay for use

SERVERLESS

Pay for the compute capacity only for the workload duration (metered on a per-second basis)



Queries	Query execution time
Q1, Q2, Q3	3 minutes
Q4	1 minute 10 seconds
Q5	1 minute and 20 seconds
Total charges	5 minutes and 30 seconds

No charges for idleness!

Redshift Serverless or Provisioned Highlights

Provisioned

- Cluster of Compute Nodes
- Greater control of configuration and workload management
- Predictable cost
- Discounts with Reserved Instances

Serverless

- Workgroup is a collection of compute resources
- Workgroup resources managed by Redshift Processing Units (RPU)
- Simplified management
- Pay for use

Autonomics

Autonomics in the data warehouse

ML-BASED OPTIMIZATIONS

ATO: smart defaults

Automatic Table Optimizations (ATO)

Smart defaults for compression and distribution

Auto MVs, auto refresh, & query rewrite

Auto Materialized Views (MVs)

System generated MVs to improve performance

Amazon Redshift Advisor

Amazon Redshift Advisor

Recommendations for improved performance

Autonomics in the data warehouse

ML-BASED OPTIMIZATIONS

Automatic Table Optimizations (ATO)

Automates the physical data distribution and schema design in the storage layer.

ATO: automatic sort & distribution keys

Auto ANALYZE and VACUUM

Tables maintained automatically

Auto analyze, vacuum delete, column encoding

Workload Management (WLM)

Automatic and ML powered

Auto workload management

Redshift Scalability

Concurrency Scaling

Scale Compute elasticity and dynamically to handle unpredictable user demand

Automatically scale-out to multiple Redshift clusters from a single endpoint in seconds and scale back down when workloads decrease

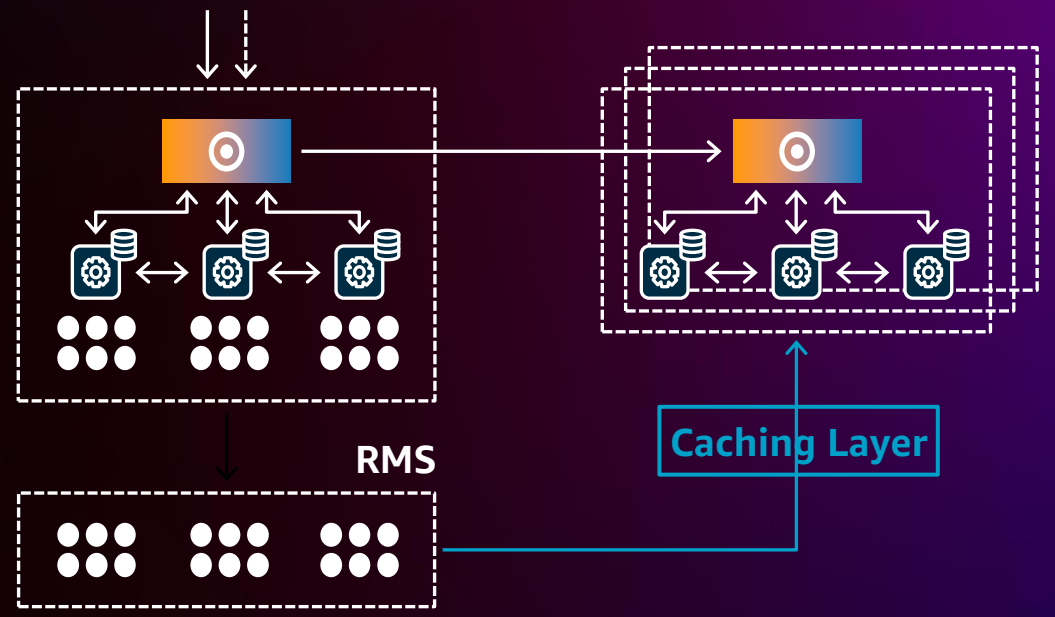
- Support virtually unlimited concurrent users while maintaining SLAs
- Choose to toggle Concurrency Scaling on/off for a given workload

Free one-hour usage per day. Per-second billing for additional clusters used

Configure usage limit to ensure Concurrency scaling does not exceed free tier

Uses machine learning to optimize query throughput

Write operations (COPY, INSERT, UPDATE, and DELETE) can run on transient Concurrency Scaling clusters when there is queueing



Workgroup Scaling

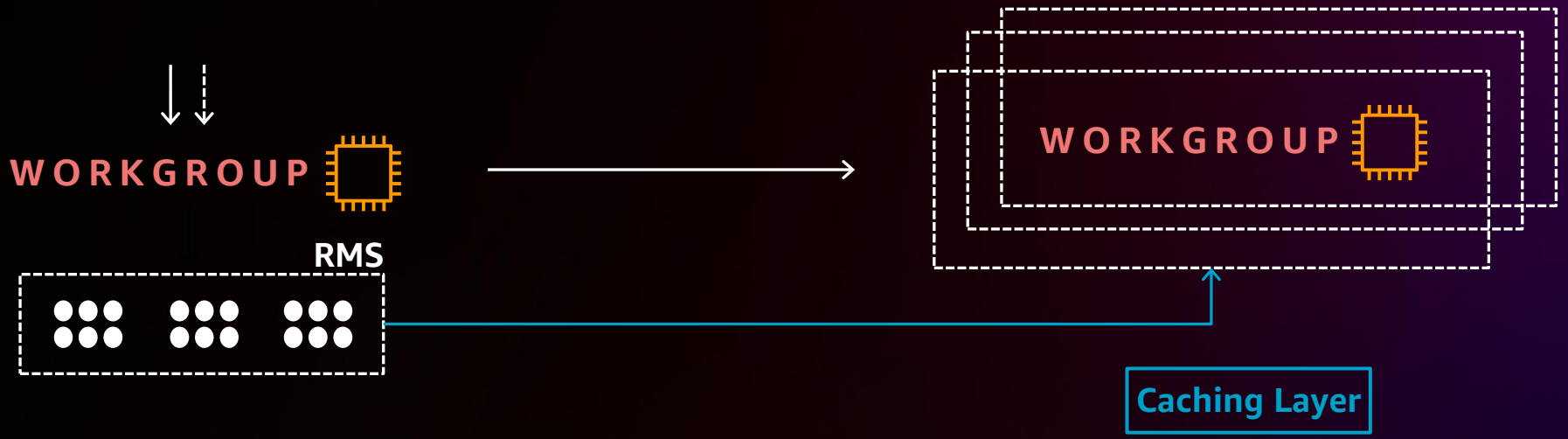
SERVERLESS

Scale Compute elasticity and dynamically to handle unpredictable user demand

Automatically scale-out to multiple Redshift workgroups from a single endpoint in seconds and scale back down when workloads decrease

- Support virtually unlimited concurrent users while maintaining SLAs
- Base RPU defines Workgroup size
- Limit RPU usage daily, weekly, or monthly

Uses machine learning to optimize query throughput

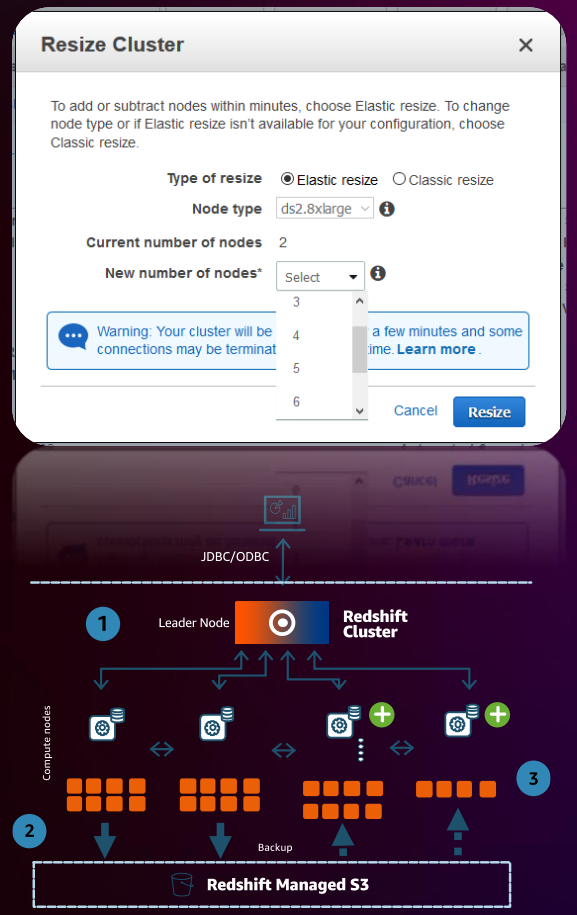


Redshift cluster resizing

PROVISIONED

- Elastic Resize
  - Add or remove nodes to/from existing cluster
  - Completes within few minutes. Limited disruption to sessions and queries
  - Slice count remains the same as original cluster
- Asynchronous Classic Resize
  - New nodes added to existing cluster
  - Available within a few minutes. Tables are rebuilt asynchronously
  - Slice count changes based on the number of nodes

Resizing a cluster is easily achieved with a few clicks on the Redshift console, and there are two resizing approaches to choose from

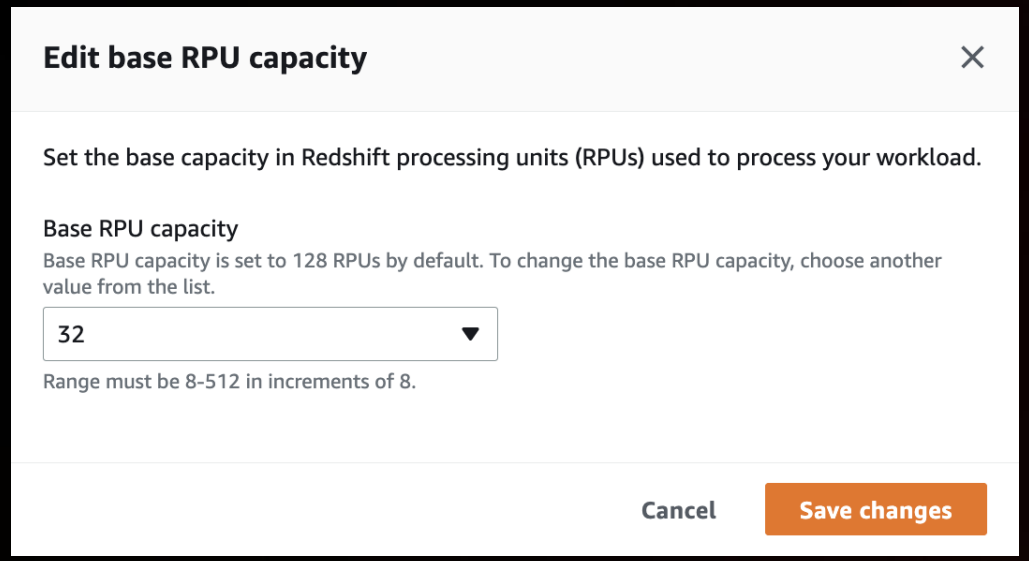


Redshift workgroup resizing

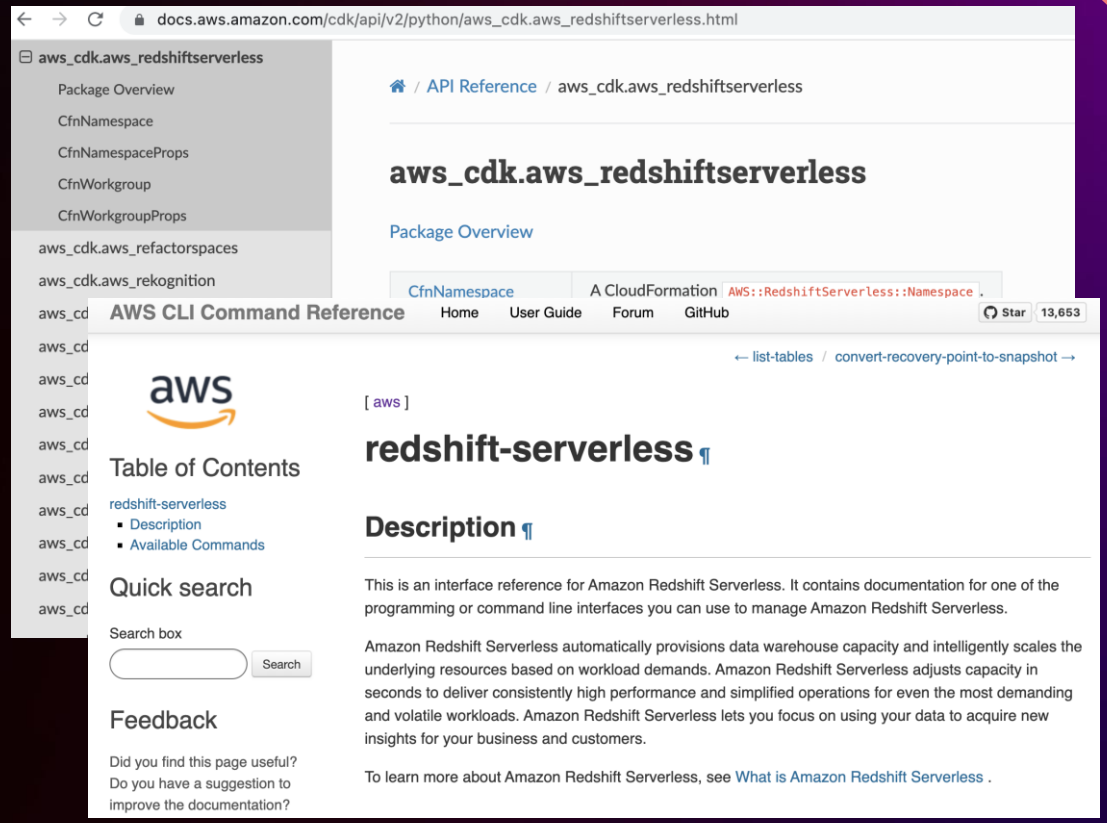
SERVERLESS

- Base RPU
  - Adjust Base RPU up or down
  - Short modifying state
  - AWS Console, AWS Command Line Interface (CLI), and AWS Cloud Development Kit (CDK) integration

Resizing a workgroup is easily achieved with a few clicks on the Redshift console



Resizing a workgroup is easily achieved with the CLI or CDK



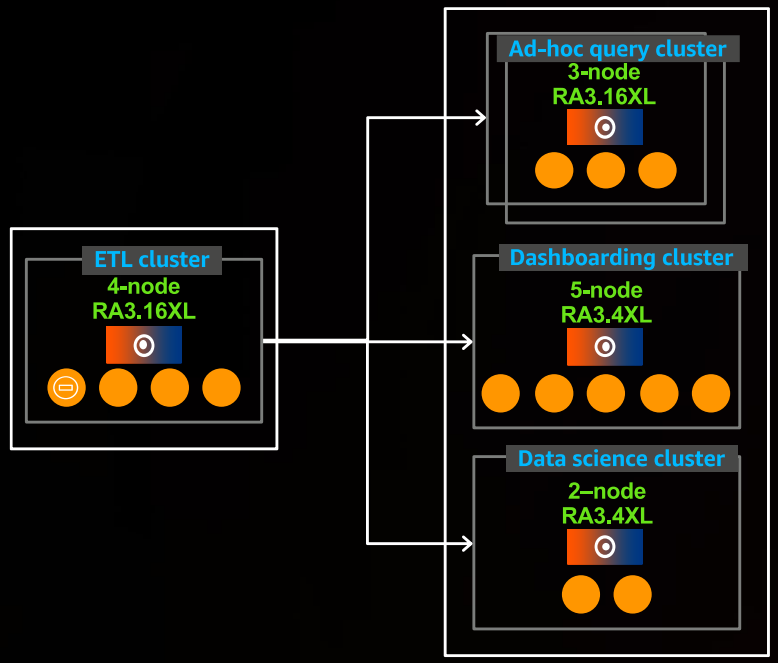
Amazon Redshift Data Sharing Overview

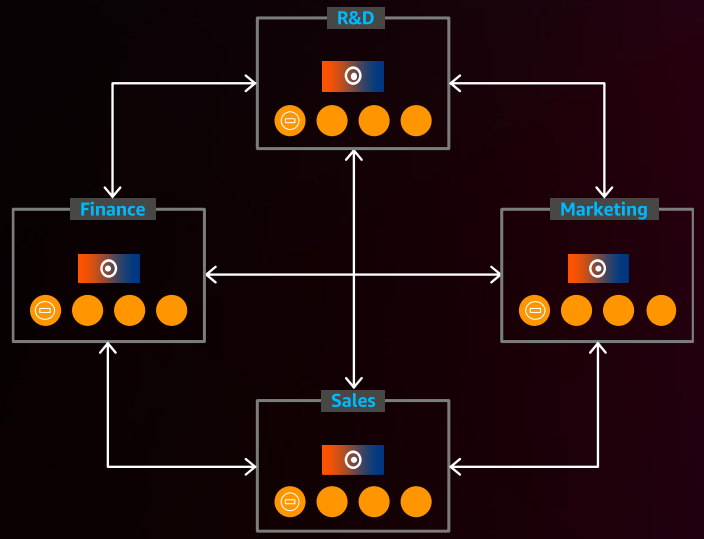
Live Data Sharing that's instant, granular and highly performant. Eliminate the complexity of data movement

Workloads accessing shared data are isolated from each other and the producer

Data Sharing

A secure and easy way to share data across Amazon Redshift clusters





- Instant, granular, high-performance data access without data copies / movement
- Live and consistently updating views of data across all consumers
- Secure and governed collaboration within and across organizations and with external parties

- Workloads accessing shared data are isolated from each other
- Use cases: Cross-group collaboration and sharing, workload isolation and chargeability, data as a service
- Sharing to other AWS analytic services – coming soon

"Data sharing feature seamlessly allows multiple Amazon Redshift clusters to query data located in our RA3 clusters and their managed storage. This eliminates our concerns with delays in making data available for our teams, reduces the amount of data duplication and associated backfill headache. We now can concentrate even more of our time making use of our data in Amazon Redshift and enable better collaboration instead of data orchestration."

Steven Moy, Yelp

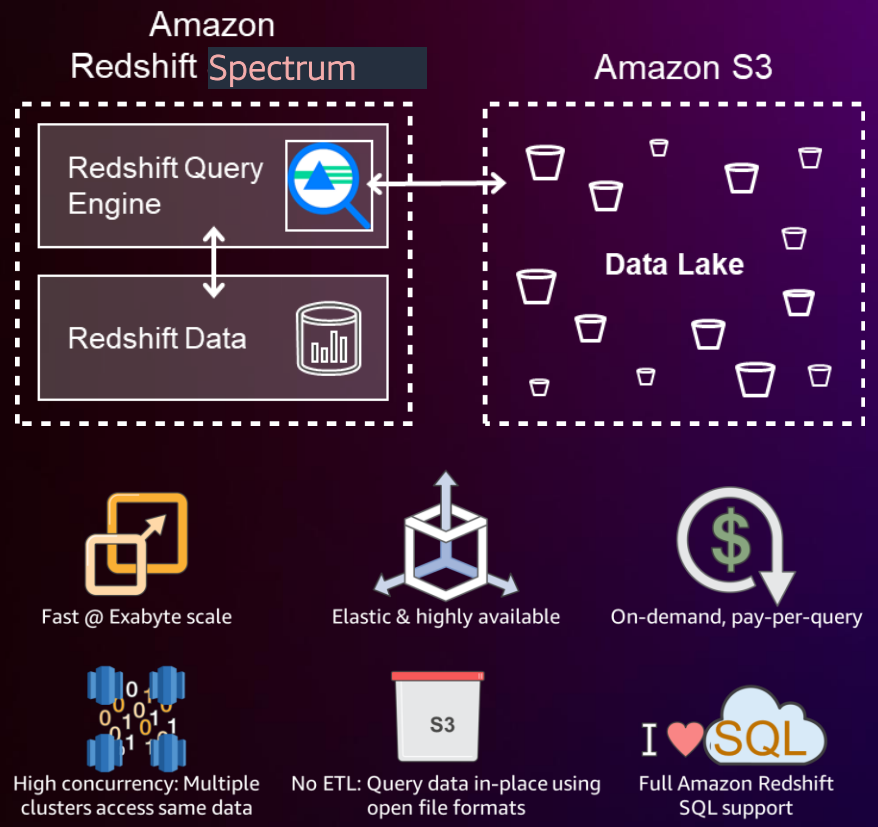
Redshift Spectrum Overview

Redshift *Spectrum* is a feature of Redshift that allows SQL queries on external data stored in Amazon S3

Benefits

- Enables the *Modern Data Architecture* pattern to query exabytes of data in an S3 data lake
- Data is queried in-place, no loading of data
- Keeps your data warehouse lean by ingesting warm data locally while keeping other data in the data lake within reach
- Write query results from Redshift direct to S3 external tables
- Create materialized views on S3 data using Redshift Spectrum queries

Run SQL queries directly against data in S3 using thousands of nodes

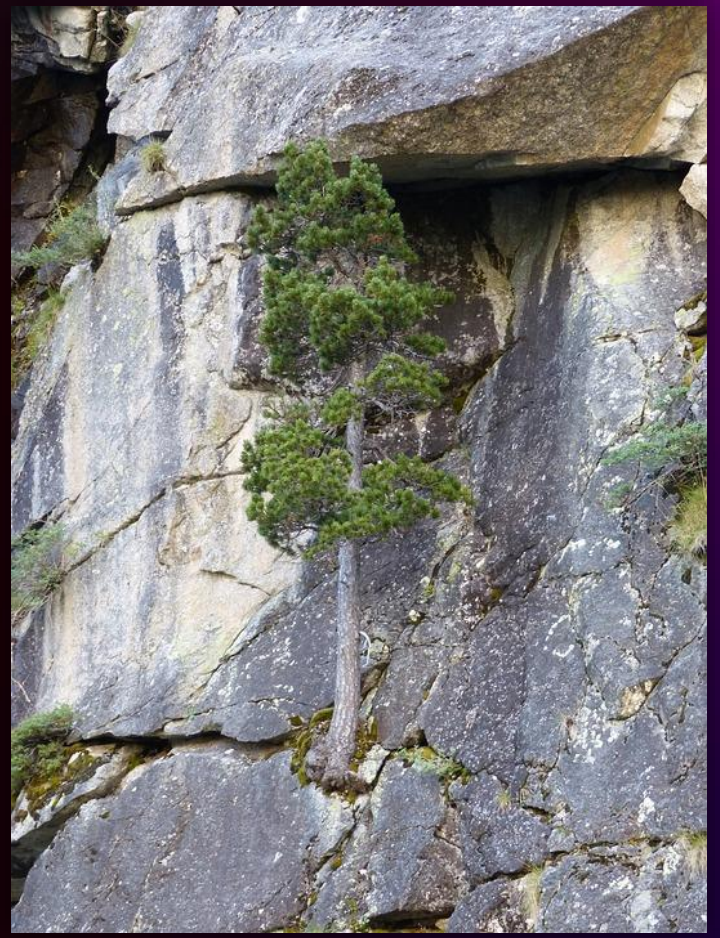


Redshift Durability

Resiliency: Overview

Key Highlights

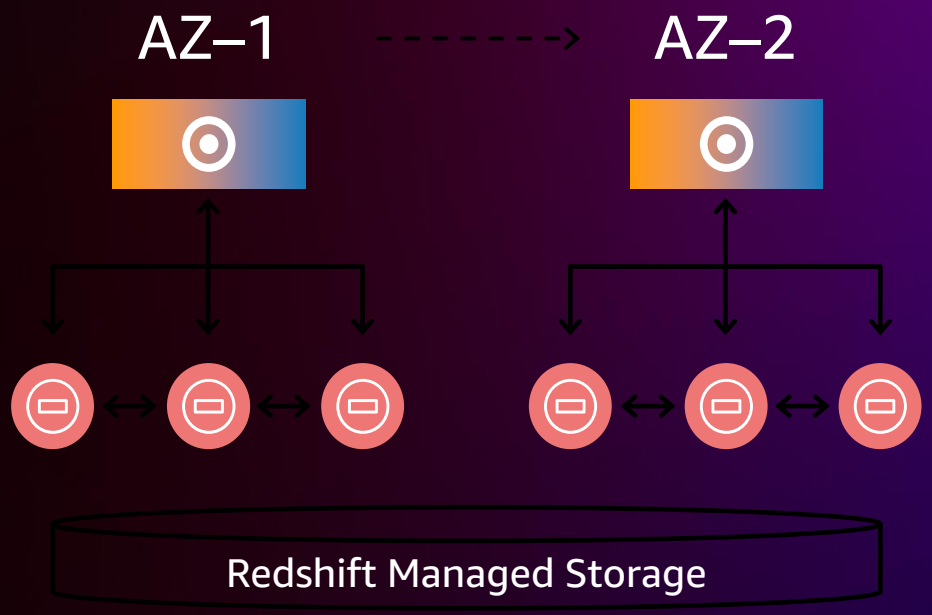
- **Amazon Redshift has a service SLA of 99.9%**
- Automatically detects and recovers from a disk or node failure
- Amazon Redshift automatically backups your data
- Amazon Redshift can automatically replicate your backups to another AWS region (e.g. DR site)
- Easily relocate provisioned cluster to another AZ
- Redshift Serverless automatically relocates workgroup to another AZ in case of failure



Cross-AZ cluster/workgroup recovery

Relocate to another Availability Zone (AZ) in response to failure

- Recovery with zero data loss (RP = Zero)
- Redshift endpoint remains unchanged after the cluster is relocated to the new Availability Zone
- No need to restore from a snapshot
- On-demand failover (provisioned only)
- Cluster/workgroup is created in another AZ, so cost of a standby replica cluster is avoided
- Redshift cluster relocation is supported for the RA3 instance types
- Included with Redshift Serverless



Security

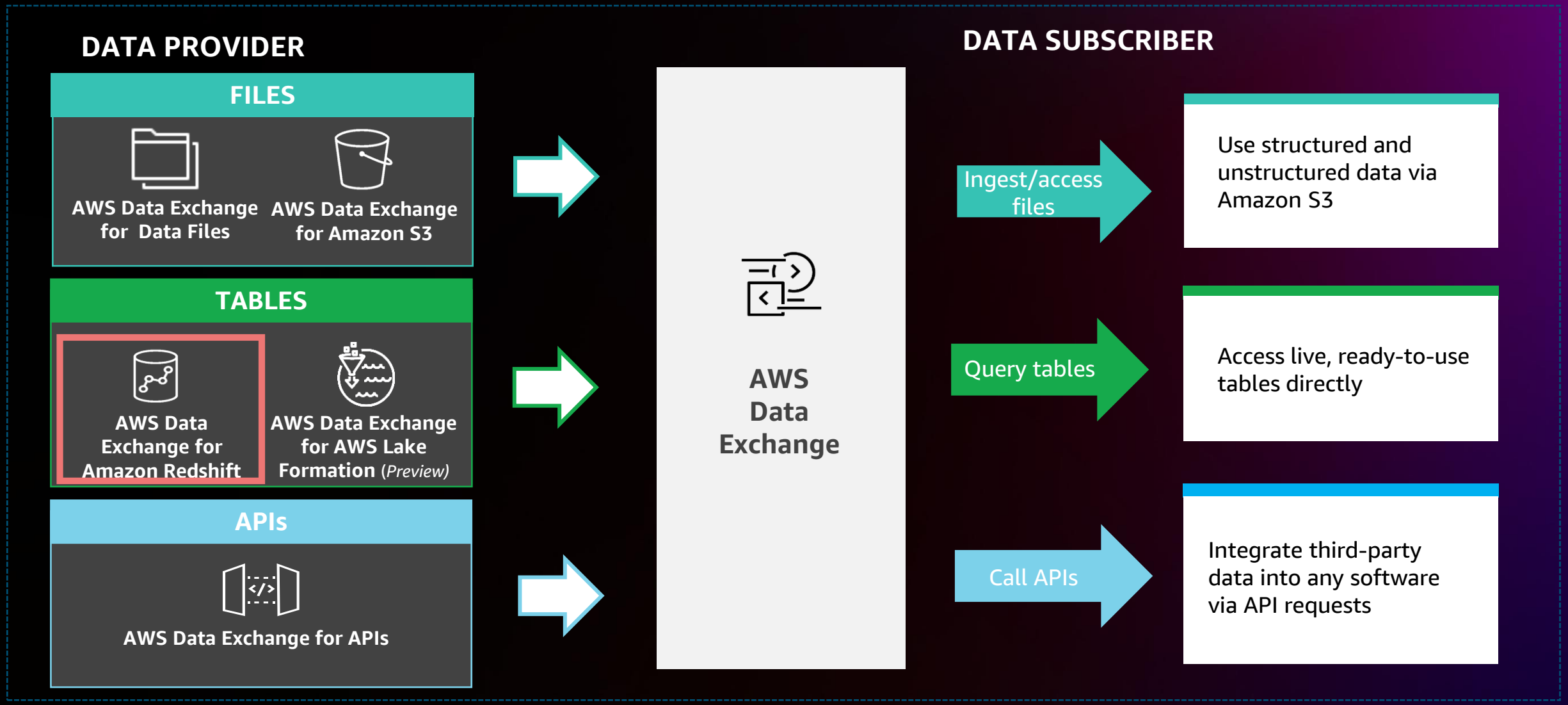
Built-in security and compliance

SECURITY AND COMPLIANCE FEATURES WITH NO EXTRA COSTS WITH AMAZON REDSHIFT

Authentication	Access control	Audit	Encryption	Helps achieve compliance
IAM integration	Role-based access control	AWS CloudTrail integration	Encrypted data in motion, data at rest	SOC
IDP integration and multifactor integration	Column-level & Row-level security	Amazon Macie integration	AWS KMS integration	PCI
	Dynamic data masking	Audit logging to Amazon CloudWatch	Faster encryption for resize/restore	FedRAMP
	AWS Lake Formation integration for data		Tokenization with Lambda UDFs and third-party tools	HIPAA and others

AWS Data Exchange for Amazon Redshift

AWS Data Exchange supports files, tables, and APIs



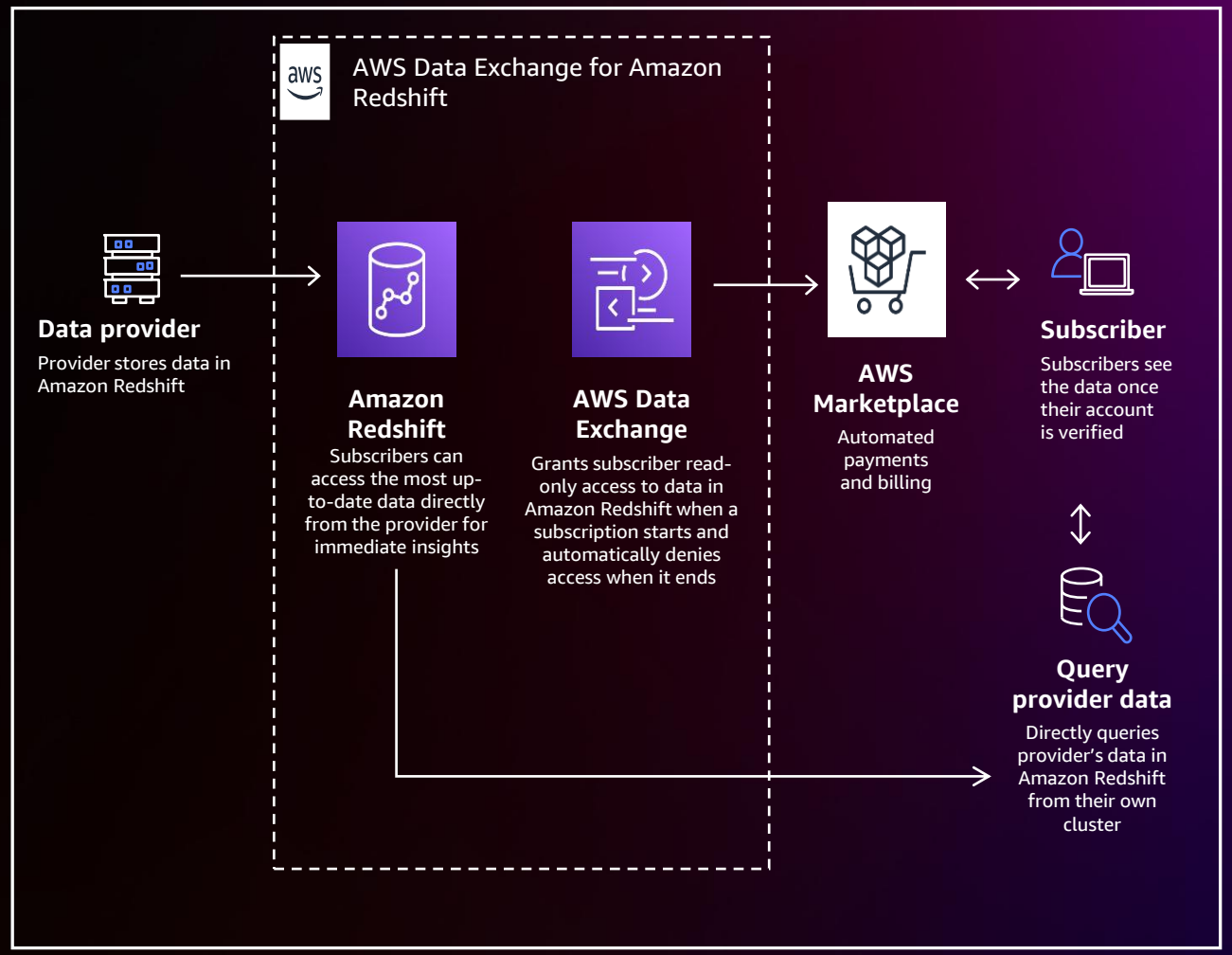
Amazon Redshift and AWS Data Exchange integration

Provider can offer custom slices of data without duplicating or modifying files

Provider can customize data for subscribers with little effort

Make data available to your subscribers across regions

Provider only pays for storage, subscribers pay for their own compute



New Features

NEW

NEW [GENERAL AVAILABILITY]

Amazon Aurora zero-ETL integration with Amazon Redshift

An easy and secure way to enable near real-time analytics on petabytes of transactional data

Supports mixed transaction and analytics workloads with Aurora and Redshift

Near real time analytics on petabytes of transaction data

Zero-ETL data pipeline

Consolidate multiple Amazon Aurora clusters into a single Amazon Redshift data warehouse for federated access to operational data stores, data warehouse and data lakes

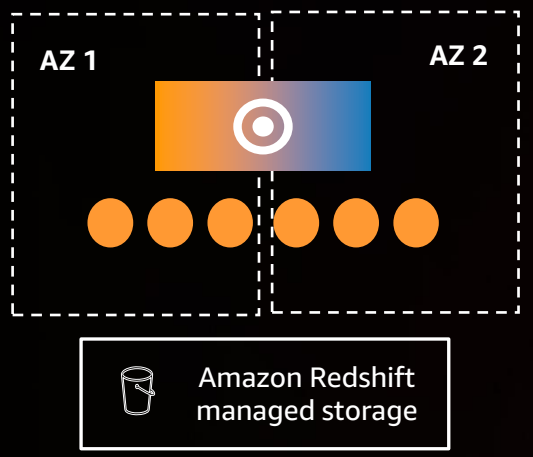
Leverage Amazon Redshift's analytics and capabilities such as built-in ML, materialized views, data sharing with transactional data

NEW

NEW [GENERAL AVAILABILITY]

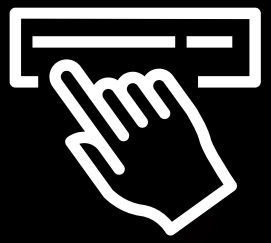
Amazon Redshift Multi-AZ

Highly resilient data warehouse



- Auto-failover with zero data loss and no manual intervention
- Easy management through a single endpoint
- Workload processing across Availability Zones
- Designed for highest levels of availability and resiliency
- Improved Concurrency
- Redshift Provisioned Clusters only

NEW



NEW [GENERAL AVAILABILITY]

Amazon Redshift supports dynamic data masking

baffle

"We are excited about utilizing the Amazon Redshift Dynamic Data Masking capability to allow our customers to achieve the goal of protecting sensitive data throughout the analytics pipeline from secure ingestion to responsible consumption."

Ameesh Divatia

CEO & Co-Founder, Baffle.io

Easily protect sensitive data by managing data masking policies through a SQL interface

User can define the way to do the data masking. Modify sensitive or PII data with fictitious content viable for software development, testing, analytics

Restrict different levels of permissions to masked data with Role Based Access Control

ID	Geo-location	Name	Phone number
123	WA	Ana	123-456-3568
124	NY	Alice	123-457-****
125	WA	Bruce	123-457-3569
126	CA	Chris	123-457-****
130	CA	Sharon	123-457-****
	Condition column		Mask column

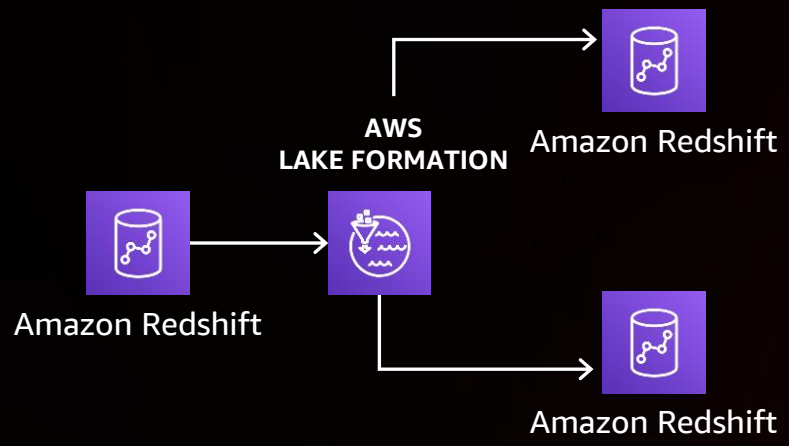
aws

NEW

NEW [GENERAL AVAILABILITY]

Data sharing access control with AWS Lake Formation

Centrally manage data sharing with AWS Lake Formation



Centrally manage granular access to data across all consuming data services

Improve security and governance with row level and column level granular permissions on data sharing

No manual scripting or complex querying

Define policies once and enforce those consistently for multiple consumers

Supports cross-region data sharing

Next Steps

- Dig deeper into Redshift with hands on labs
- Identify workload that is suited to be migrated to Amazon Redshift
- Determine scope and success criteria for an evaluation
- Understand the timeline for such an evaluation

Thank you!