Amazon Redshift

Analyze Data Across Your Lake House with Amazon Redshift

Rajesh Francis

Sr. Analytics Specialist Solutions Architect

Discussion Topics

- Data Trends and Lake House architecture
- Amazon Redshift Architecture
- Analyze all your data
- Reference Architectures
- Performance and Scale
- Price for performance

> **Bradley Todd**
> **Liberty Mutual, Technology Architect**
>
> Redshift allows us to quickly spin up clusters and provide our data scientists with a fast and easy method to access data and generate insights

Data Trends

Amazon Redshift

Large-Scale Data Warehousing Service

Used by more customers for their data warehouse workloads than anyone else

docomo
WARNER BROS. GAMES
yelp
Jack in the box
Pfizer
AstraZeneca
playrix
ancestry
coursera
Nasdaq
duolingo
EA
EQUINOX
FINANCIAL TIMES
intuit
Liberty Mutual INSURANCE
London Stock Exchange
FOX
QANTAS
SCHOLASTIC
Sysco
tinder

NTT DOCOMO

Moved >10 PB of data from on-premises to cloud

WARNER BROS.

Performance, scale, cost-effeciency

Yelp

Enabling a data-driven organization with concurrency scaling

Jack in the Box

Improved ops by moving off of on-premises DW

Pfizer

Provide scientists with near real-time analysis

Challenges of data analytics at scale

VARIETY

Variety of sources and data types

Multiple analytics needs

Data volume and velocity

PERFORMANCE

Slow performance

Difficult to manage systems

Complex to scale

COST

Increasing and unpredictable cost

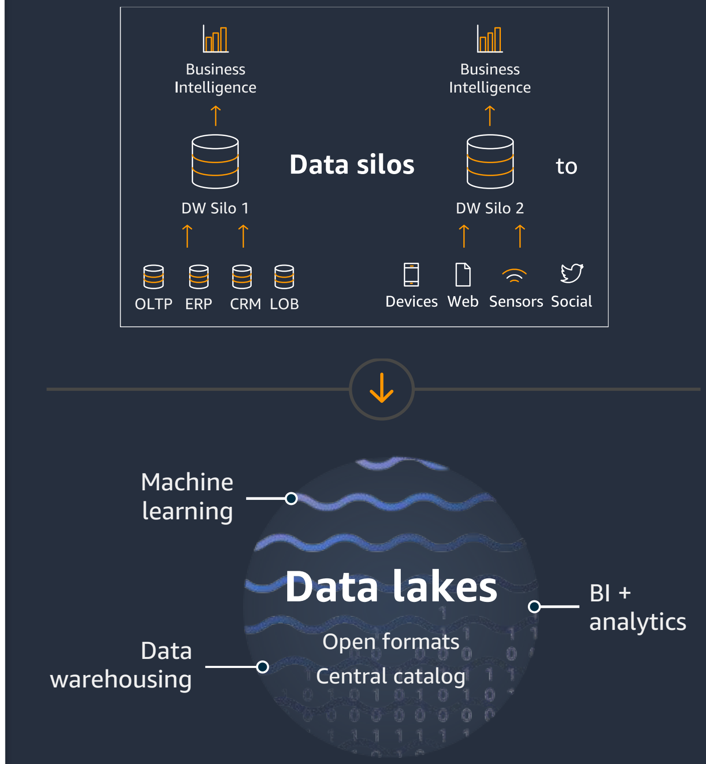
Inflexible tools

Security, compliance

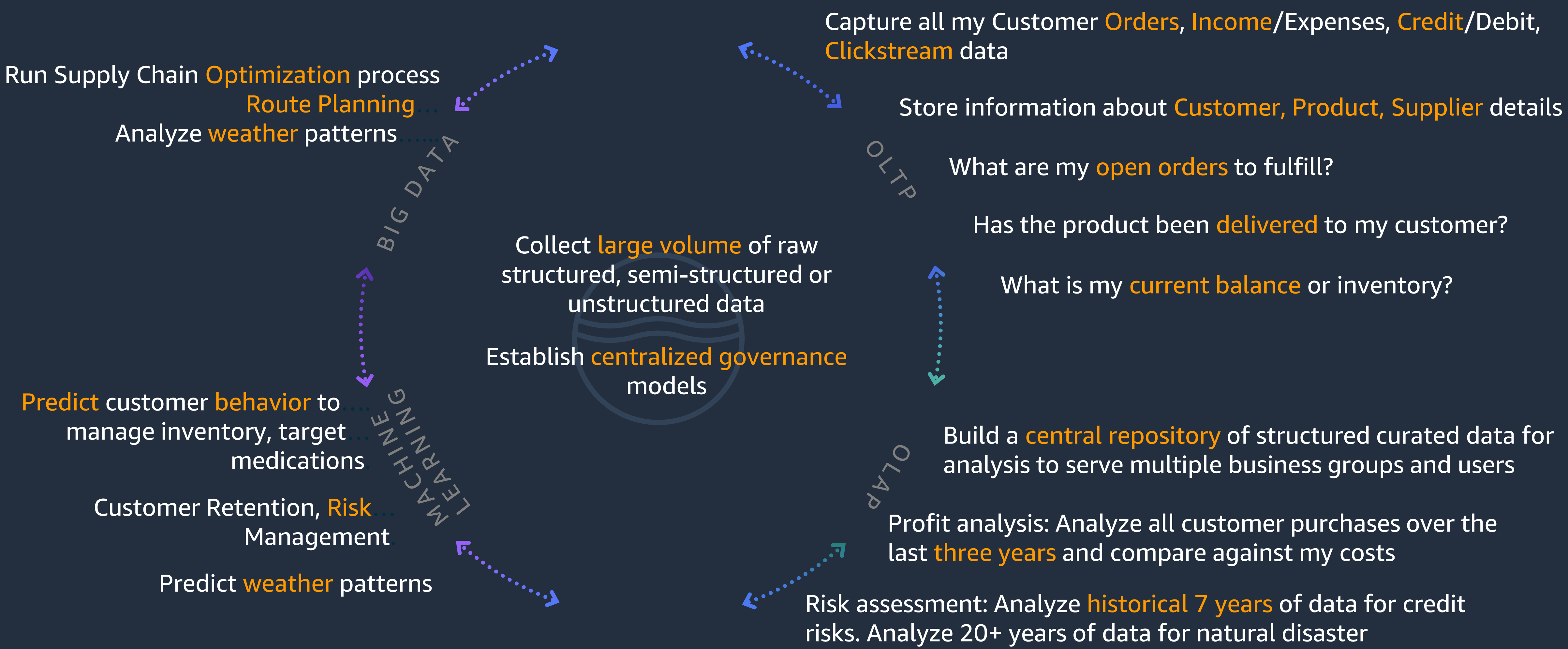
Traditional architectures lead to dark data

Traditional architectures & on-prem data warehousing lead to *dark data –* data that is collected but challenging to extract insights from that data.

Scale	• Can’t scale easily or on-demand • Long lead times for hardware procurement & upgrades
Cost	• High operational costs • Compute tied to storage
Anti-democratization	• Proprietary formats • Data silos • Need to ingest, transform data before analysis • Limits on users and data
Architecture	• One size fits all approach



Business use cases for Lake House architecture



Collect large volume of raw structured, semi-structured or unstructured data

Establish centralized governance models

Run Supply Chain Optimization process

Route Planning

Analyze weather patterns

Capture all my Customer Orders, Income/Expenses, Credit/Debit, Clickstream data

Store information about Customer, Product, Supplier details

What are my open orders to fulfill?

Has the product been delivered to my customer?

What is my current balance or inventory?

Build a central repository of structured curated data for analysis to serve multiple business groups and users

Profit analysis: Analyze all customer purchases over the last three years and compare against my costs

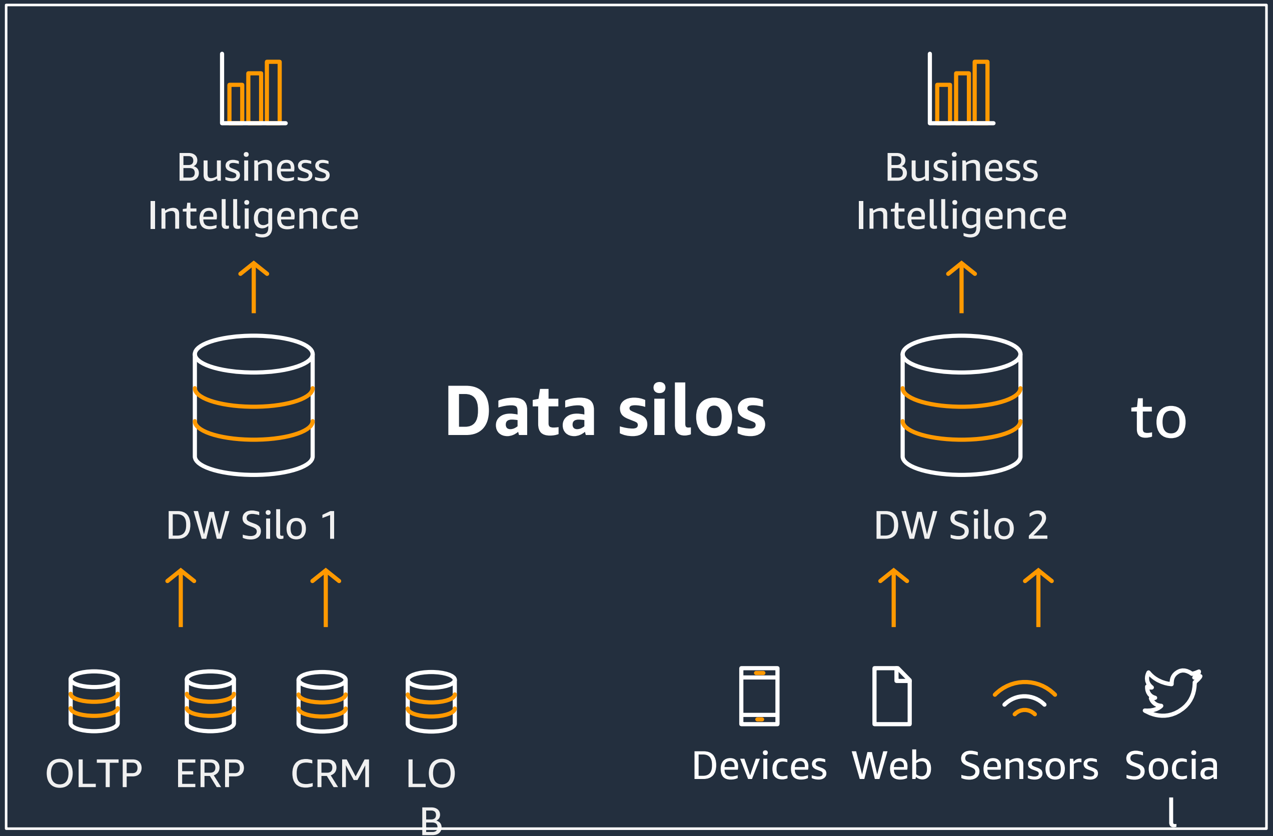
Risk assessment: Analyze historical 7 years of data for credit risks. Analyze 20+ years of data for natural disaster

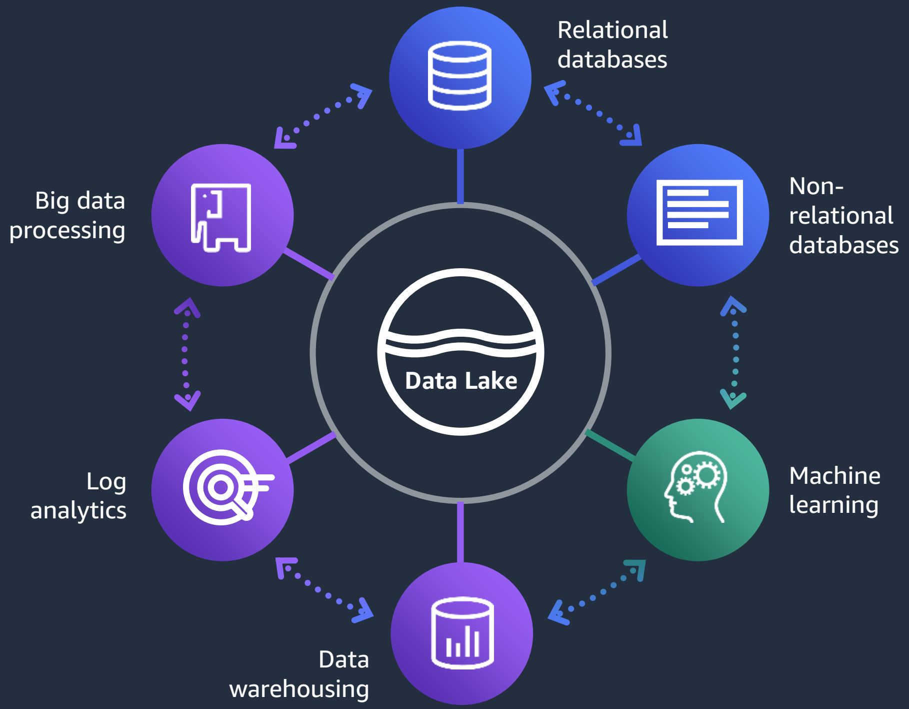
Predict customer behavior to manage inventory, target medications

Customer Retention, Risk Management

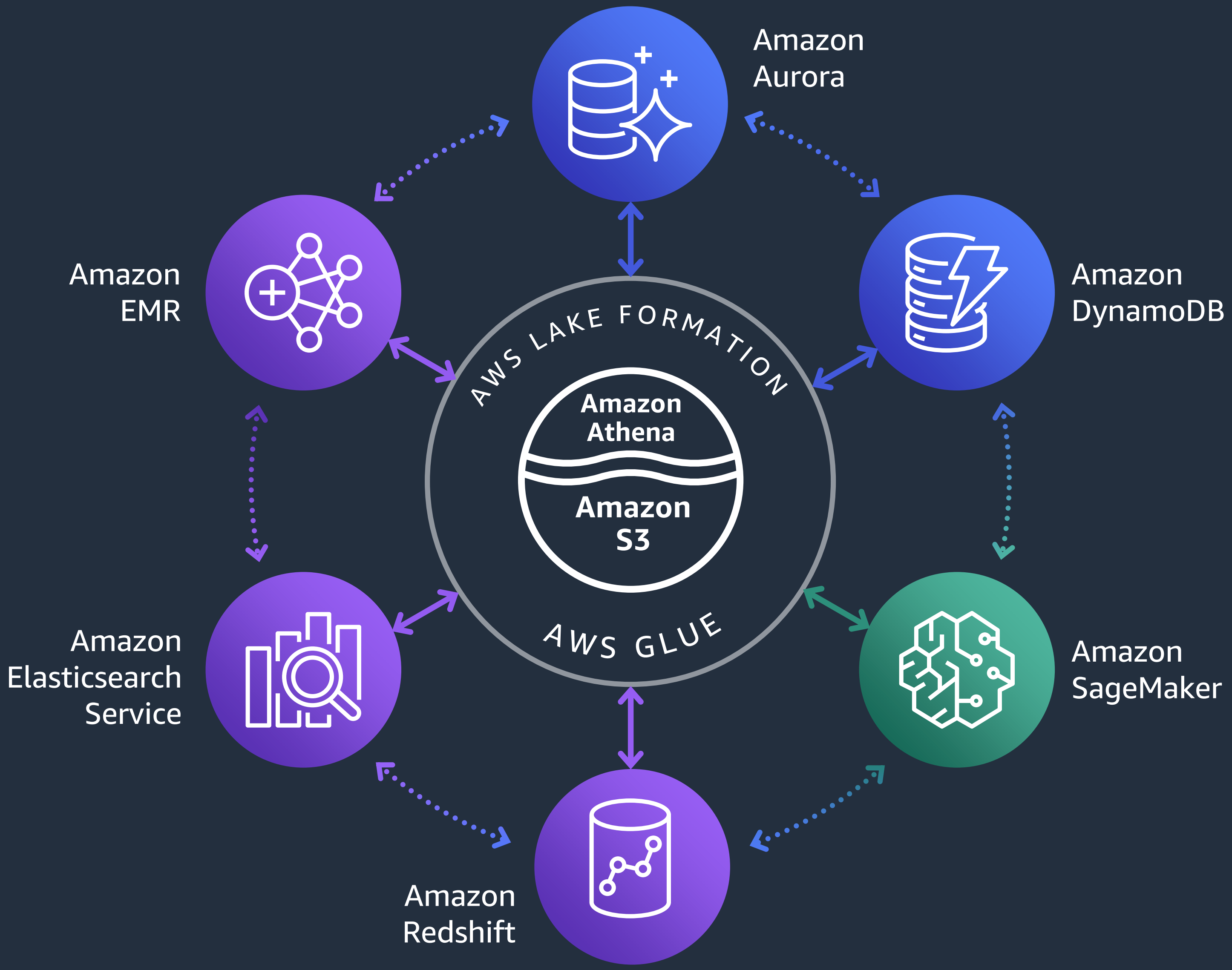
Predict weather patterns

From traditional Data Warehouse to Lake House approach





Lake House architecture on AWS



Scalable data lakes

Purpose-built data services

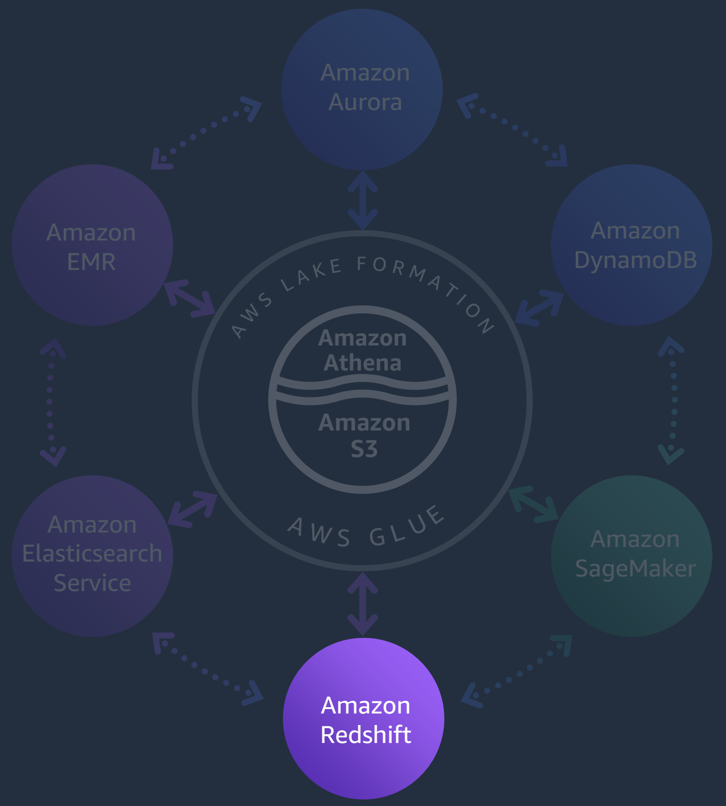
Seamless data movement

Unified governance

Performant and cost-effective

Amazon Redshift

Analyze all your data with the fastest and most widely used cloud data warehouse



Analyze all your data

Deepest integration with your data lake

Performance at any scale

Up to 3x better price performance than other cloud DW

Lower your costs

At least 50% less expensive than other cloud DW

Amazon Redshift innovates to meet your needs

Analyze all your data

Lake house with AWS integration

- NEW! Amazon Redshift ML
- NEW! Data sharing
- NEW! Super data type with JSON support
- UPDATED! Federated query
- NEW! AWS Lambda UDF
- NEW! Partner console integration
- Amazon Redshift Spectrum + AWS Lake Formation
- Data lake export

Performance & scale

Fast and self-tuning

- UPDATED! RA3 nodes & managed storage
- NEW! AQUA
- NEW! Performance tuning: automated
- UPDATED! Materialized views
- NEW! 100K tables
- NEW! HyperLogLog
- Concurrency scaling

Low cost & best value

Predictable costs

- UPDATED! Automatic workload manager
- NEW! Cross-AZ cluster recovery
- NEW! Data API
- On-demand and RIs
- Pause and resume
- Cost controls
- Built-in security features

Customers – sample list

Tens of thousands of customers process exabytes of data with Amazon Redshift daily

NTT docomo

NTT DOCOMO

Moved >10 PB of data from on-premises to cloud

WARNER BROS. GAMES

WARNER BROS.

Performance, scale, cost-efficiency

yelp

Yelp

Enabling a data-driven organization with concurrency scaling

Jack in the box

Jack in the Box

Improved ops by moving off of on-premises DW

Pfizer

Pfizer

Provide scientists with near real-time analysis

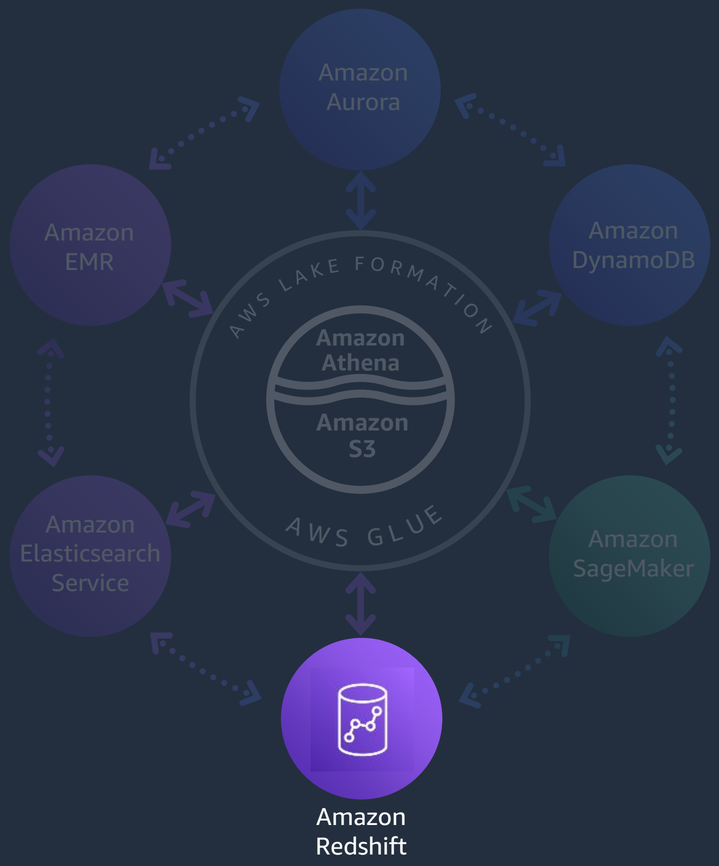
AstraZeneca playrix ancestry coursera Nasdaq duolingo EA EQUINOX FINANCIAL TIMES

intuit Liberty Mutual INSURANCE London Stock Exchange Magellan HEALTH FOX QANTAS SCHOLASTIC Sysco tinder

aws

Amazon Redshift

Analyze all your data



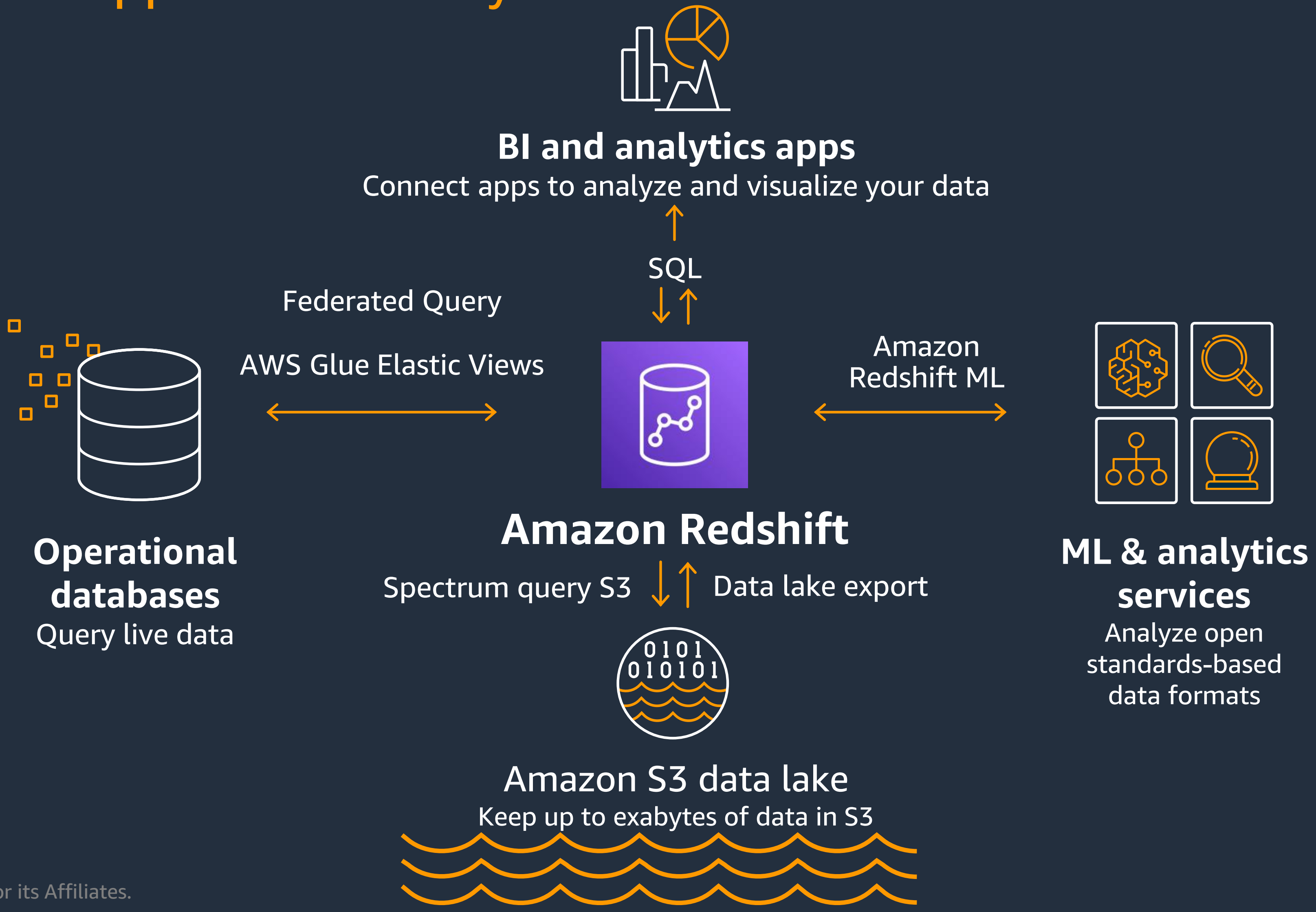
Analyze all your data
Deepest integration with your data lake

Performance at any scale
Up to 3x better price performance than other cloud DW

Lower your costs
At least 50% less expensive than other cloud DW

Analyze all your data

With a lake house approach to analytics



BI and analytics apps

Connect apps to analyze and visualize your data

SQL

Federated Query

AWS Glue Elastic Views

Operational databases

Query live data

Amazon Redshift

Amazon Redshift ML

ML & analytics services

Analyze open standards-based data formats

Spectrum query S3

Data lake export

Amazon S3 data lake

Keep up to exabytes of data in S3

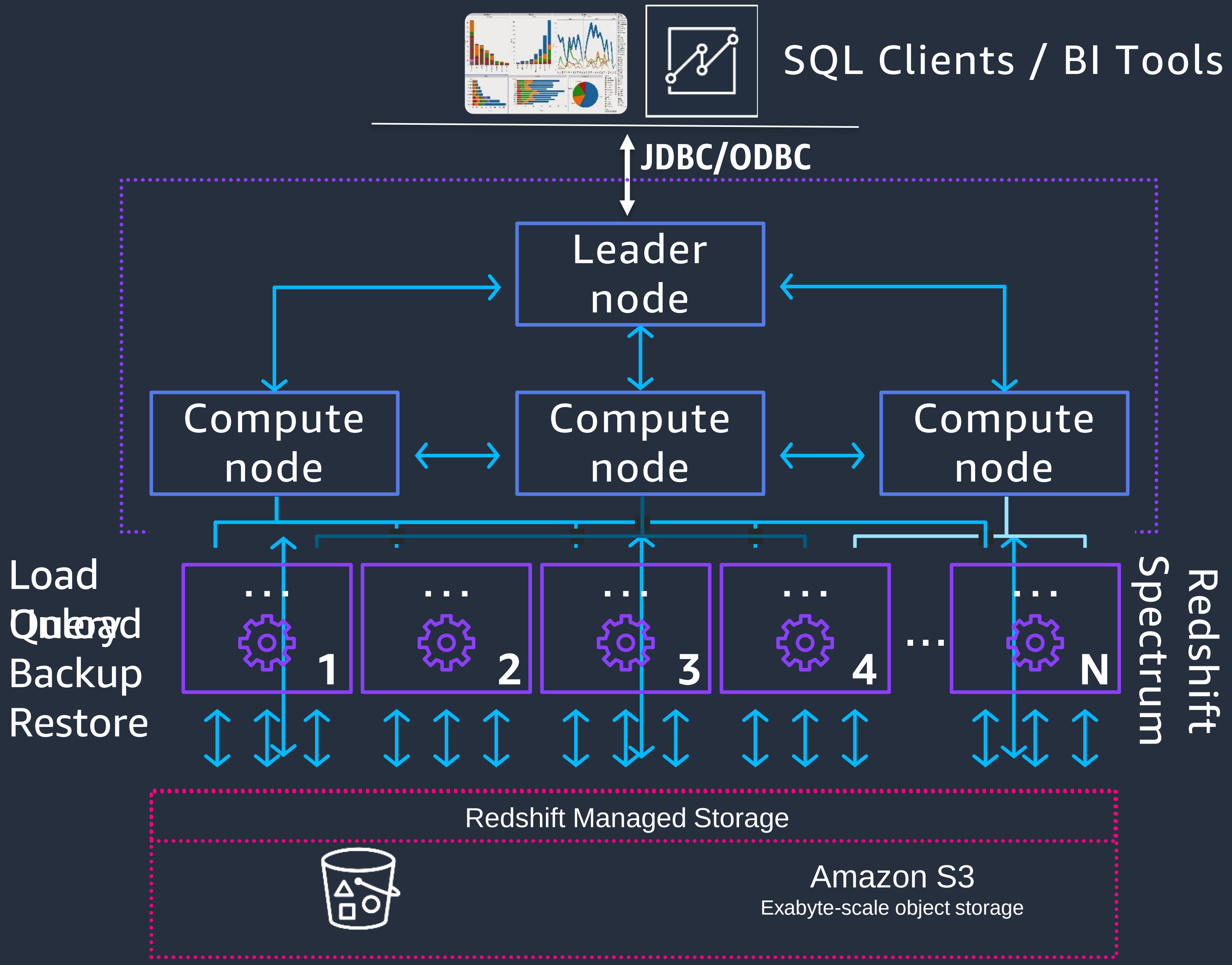
Redshift Architecture Overview

> **Anil Chalasani**
>
> **Gainsight, VP Product Operations**
>
> Using Redshift's DC2 node, we generate reports 35 percent faster. This enables our customers to spend more time curating and visualizing their data in Gainsight to take advantage of opportunities to drive customer success

Redshift cluster architecture

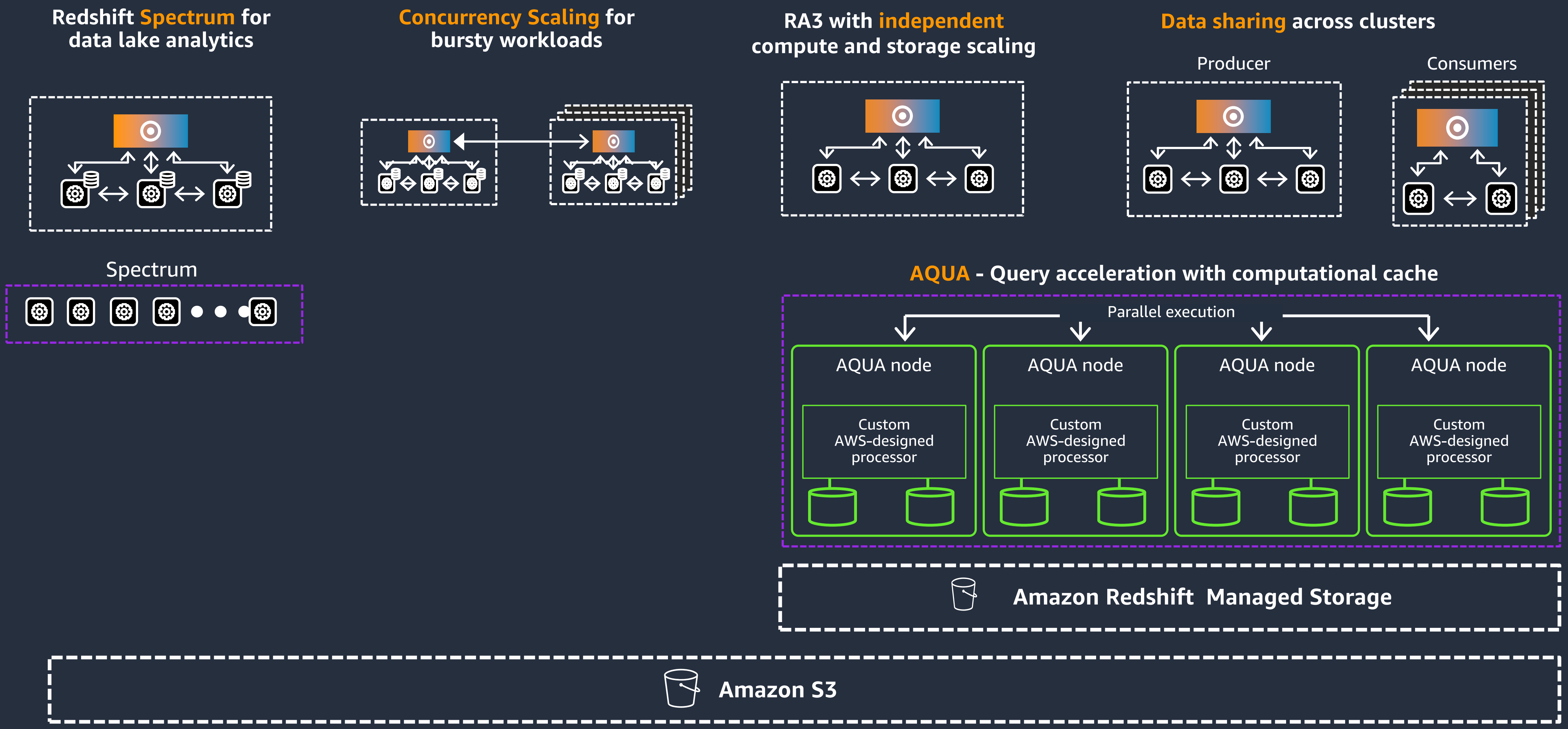
- Leader node
  - SQL endpoint
  - Stores metadata
  - Coordinates parallel SQL processing & ML optimizations
  - Leader node is no-charge for clusters with 2+ nodes
- Compute nodes
  - Local, columnar storage
  - Executes queries in parallel
  - Load, unload, backup, restore from S3
- Amazon Redshift Spectrum nodes

  Execute queries directly against data lake
- Massively parallel, shared nothing architecture



Evolving architecture (2017–2020)

Incremental features released in the last few years



Redshift Spectrum for data lake analytics

Spectrum

Concurrency Scaling for bursty workloads

RA3 with independent compute and storage scaling

Data sharing across clusters

Producer

Consumers

AQUA - Query acceleration with computational cache

Parallel execution

AQUA node

Custom AWS-designed processor

AQUA node

Custom AWS-designed processor

AQUA node

Custom AWS-designed processor

AQUA node

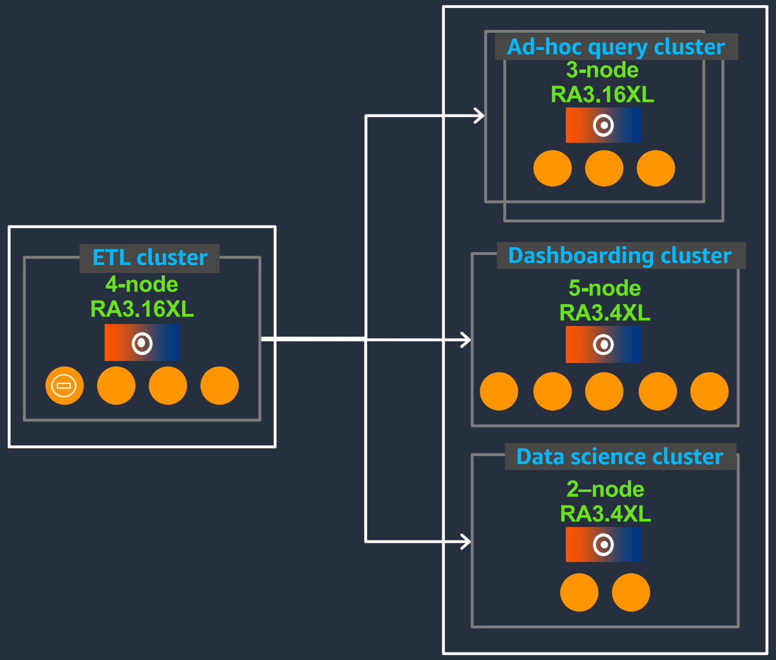
Custom AWS-designed processor

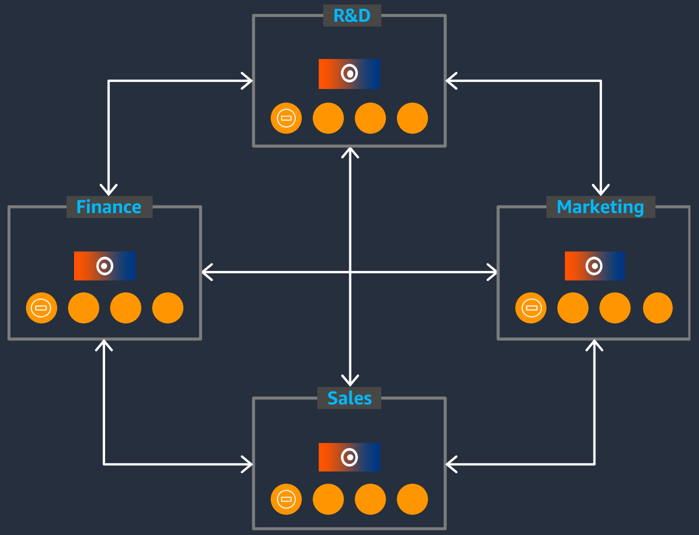
Amazon Redshift Managed Storage

Amazon S3

Data Sharing

A SECURE AND EASY WAY TO SHARE DATA ACROSS AMAZON REDSHIFT CLUSTERS





"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."

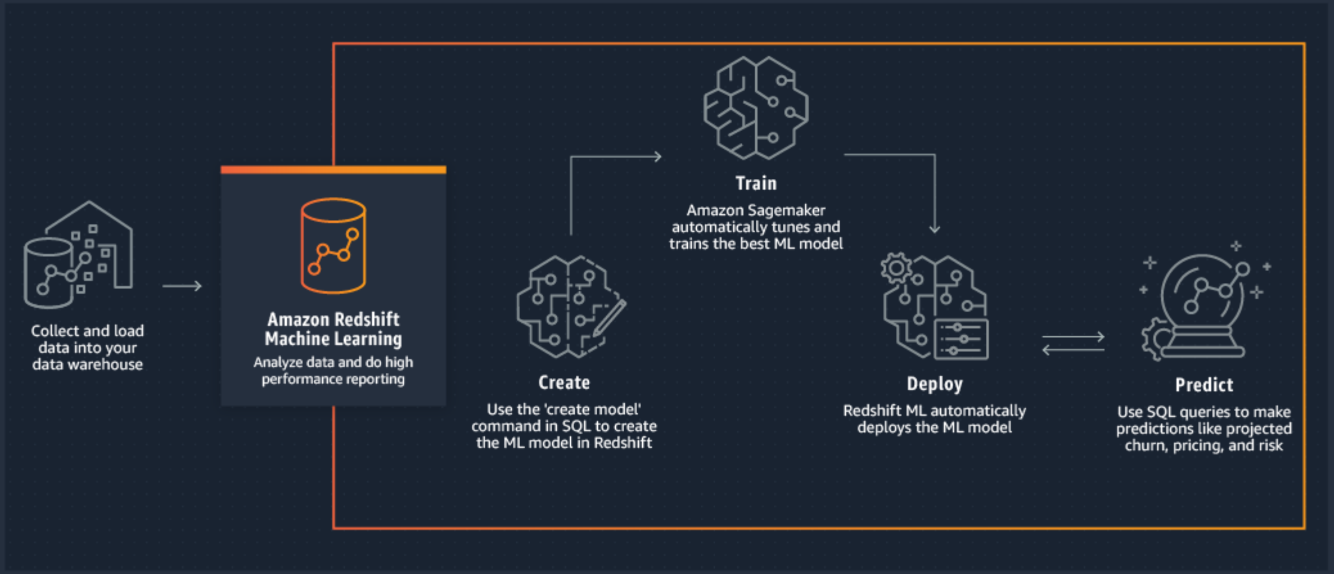
Steven Moy, Yelp

- 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

Amazon Redshift ML

EASILY CREATE AND TRAIN ML MODELS USING SQL QUERIES WITH AMAZON SAGEMAKER

- ✓ Use case: Product recommendations, fraud prevention, reduce customer churn
- ✓ Create, train, and apply ML models using SQL
- ✓ Automatic selection of ML algorithms or select your algorithm with XGBoost
- ✓ Automatic pre-processing, creation, training, deployment of your model
- ✓ Deploy inference models locally in Amazon Redshift; run an inference as invoking a user-defined function as part of SQL statements

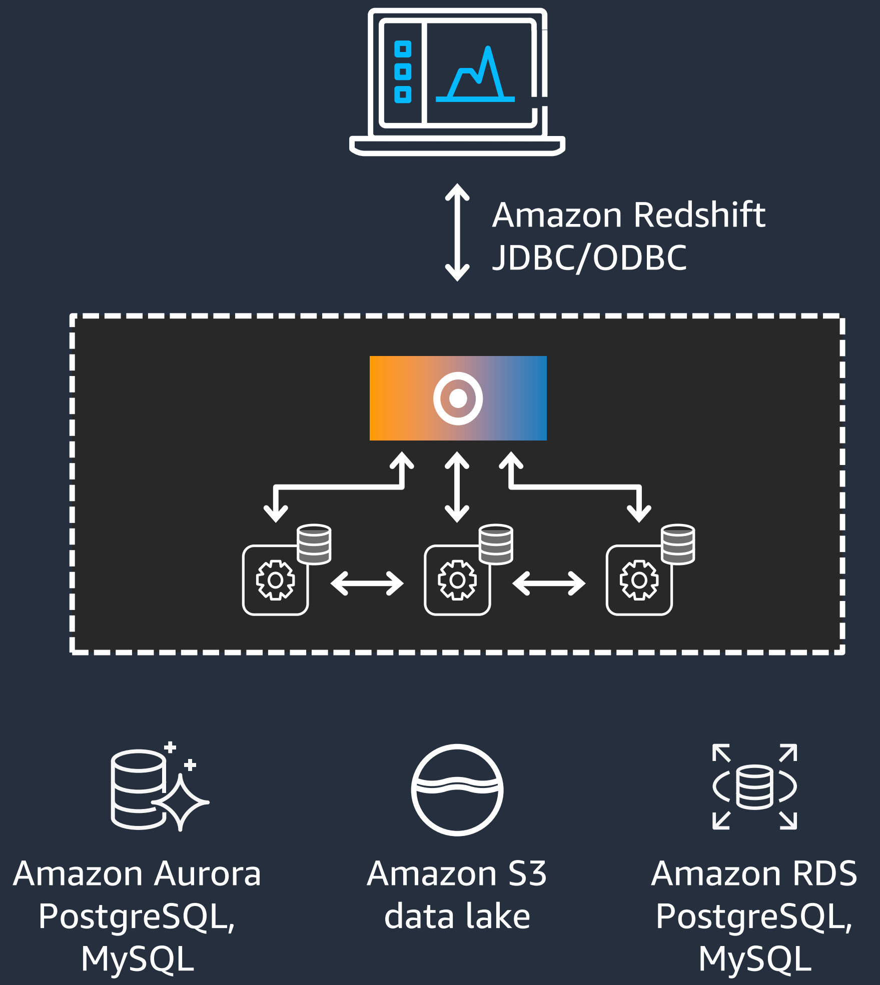

Collect and load data into your data warehouse
Amazon Redshift Machine Learning
Analyze data and do high performance reporting
Create
Use the 'create model' command in SQL to create the ML model in Redshift
Train
Amazon Sagemaker automatically tunes and trains the best ML model
Deploy
Redshift ML automatically deploys the ML model
Predict
Use SQL queries to make predictions like projected churn, pricing, and risk

```
CREATE MODEL demo_ml.customer_churn
FROM (SELECT c.age, c.zip, c.monthly_spend,
c.monthly_cases, c.active FROM
customer_info_table c)
TARGET c.active;
```

Amazon Redshift Federated Query

UNIFIED ANALYTICS ACROSS DATABASES, DATA WAREHOUSE, AND DATA LAKE

- ✓ Use case: Integrate operational data with DW and data lake for real-time analytics
- ✓ Analytics on operational data without data movement and ETL delays
- ✓ Query and join data from one or more Amazon RDS and Aurora PostgreSQL databases
- ✓ Flexible and easy way to ingest data avoiding complex ETL pipelines
- ✓ Intelligent distribution of computation to remote sources to optimize performance
- ✓ Amazon RDS and Aurora MySQL support



Native semi-structured data support

New data type: **SUPER**

Easy, efficient, and powerful JSON processing

Fast row-oriented data ingestion

Fast column-oriented analytics with materialized views over SUPER/JSON

Access to schema-less nested data with easy-to-use SQL extensions powered by the PartiQL query language

id INTEGER	name SUPER	phones SUPER
1	{"given":"Jane", "family":"Doe"}	[{"type":"work", "num":"9255550100"}, {"type":"cell", "num": 6505550101}]
2	{"given":„Richard", "family":„Roe"},	[{"type":"work", "num": 5105550102}]

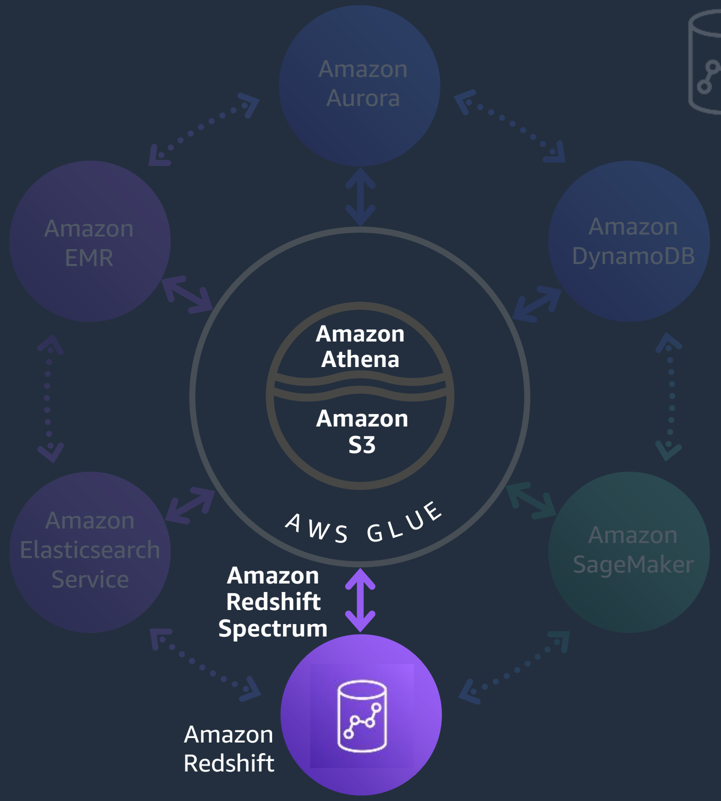
```
SELECT name.given AS firstname, ph.num
FROM customers c, c.phones ph
WHERE ph.type = ‘cell’;

firstname | num
----------+---------------
"Jane"    | 6505550101
```

Redshift Spectrum Overview

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

Benefits

- Enables the Lake House pattern to query exabytes of data in an S3 data lake
- Data is queried in-place, no loading of data
- Keep warm data local and cold/historical data in data lake
- Write query results from Redshift direct to S3 external tables
- Powered by a separate fleet of powerful Spectrum nodes
- Create materialized views on S3 data



Fast @ Exabyte scale

Elastic & highly available

On-demand, pay-per-query

High concurrency: Multiple clusters access same data

No ETL: Query data in-place using open file formats

Full Amazon Redshift SQL support

Steps to define and create External Schema & Tables

1. Define an external schema in Amazon Redshift using the AWS Glue data catalog or your own Apache Hive Metastore

   ```
   CREATE EXTERNAL SCHEMA <schema_name>
   ```

2. Register external tables using Athena, your Hive Metastore client, or from Amazon Redshift `CREATE EXTERNAL TABLE` syntax

   ```
   CREATE EXTERNAL TABLE <table_name>
   [PARTITIONED BY <column_name, data_type, …>]
   STORED AS file_format
   LOCATION s3_location
   [TABLE PROPERTIES property_name=property_value, …];
   ```

3. Query external tables using

   ```
   <schema_name>.<table_name>
   ```

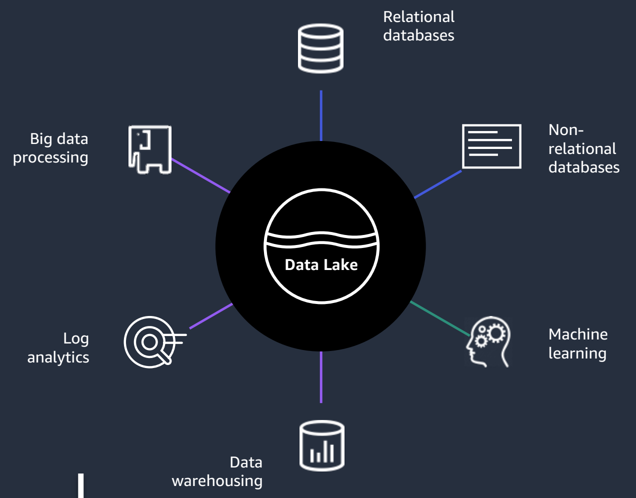
Examples from the Redshift documentation

```
create external schema hive_schema
from hive metastore
database 'hive_db‘
uri '172.10.10.10' port 99
iam_role 'arn:aws:iam::123456789012:role/MySpectrumRole';
```

```
create external table lakehouse.sales(
salesid integer,
listed integer,
saledate date,
qtysold smallint,
pricepaid decimal(8,2),
saletime timestamp)
row format delimited
fields terminated by ‘\t‘
stored as textfile
location ‘s3://sampledbusw2/tickit/lakehouse/sales/’
table properties ('numRows'='170000');
```

```
select lakehouse.sales_event.salesmonth, event.eventname,
sum(lakehouse.sales_event.pricepaid) from
lakehouse.sales_event, event where
lakehouse.sales_event.eventid = event.eventid and
salesmonth = '2008-02' and (event = '101' or event = '102')
group by event.eventname, lakehouse.sales_event.salesmonth
order by 3 desc;
```

Data Lake query services: How to choose?



- Data warehouse, highly-relational, complex joins
- Lake house architecture
- Sub-second latency
- Joins between data warehouse data & an S3 data lake

Amazon Redshift

- Interactive ad-hoc queries
- Serverless
- No data warehouse, not 24x7
- Log analysis
- Offload S3 workload from Datawarehouse

Amazon Athena

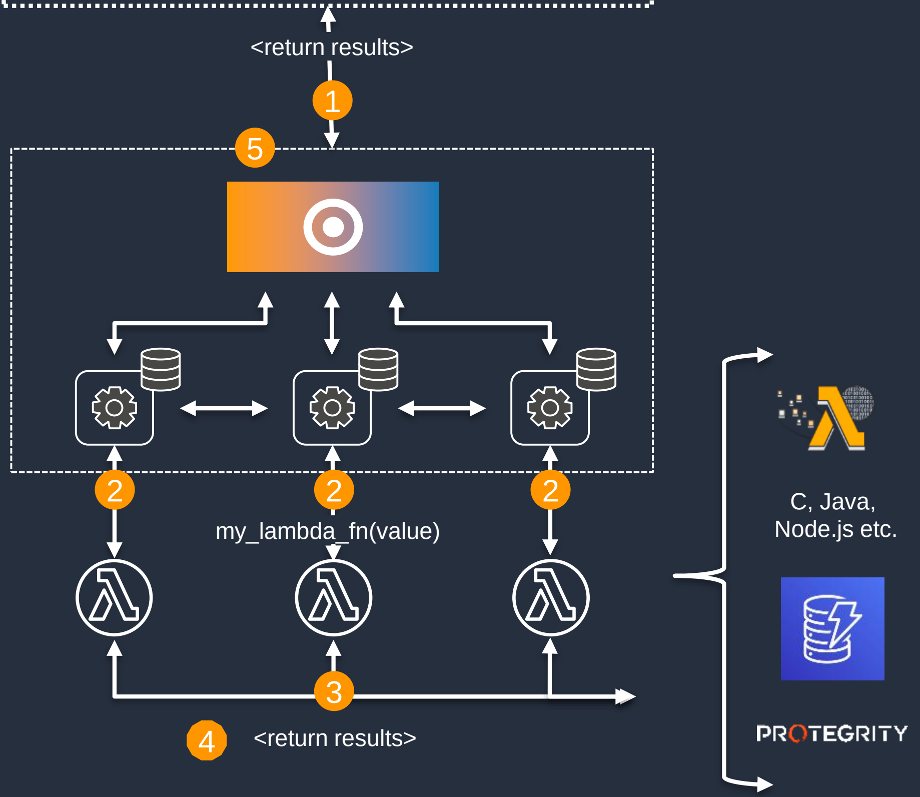
- Process large volume of data
- Use big data tools like Apache Hadoop, Spark, Presto, Hive
- Run Jupyter-based EMR notebooks

Amazon EMR

Lambda UDFs

INTEGRATE WITH EXTERNAL SERVICES USING AWS LAMBDA

```
SELECT my_lambda_fn(t.value)
FROM my_table  t
WHERE customer = “c1”
```



Invoke AWS Lambda programs as UDFs in Amazon Redshift SQL queries

Simple integration with external services

- Tokenization with third-party vendors like Protegrity
- More languages runtimes (C++, Java, etc.)
- Access Amazon DynamoDB, Amazon SageMaker, etc.

Concurrent and batch processing

Cost controls and error controls

Native partner integration

INTEGRATE WITH SELECT PARTNERS FROM THE AMAZON REDSHIFT CONSOLE

etleap

Fivetran

MATILLION

SISENSE

segment

Ingest business data and gain insights in minutes

Deep links into partner platforms streamline account setup

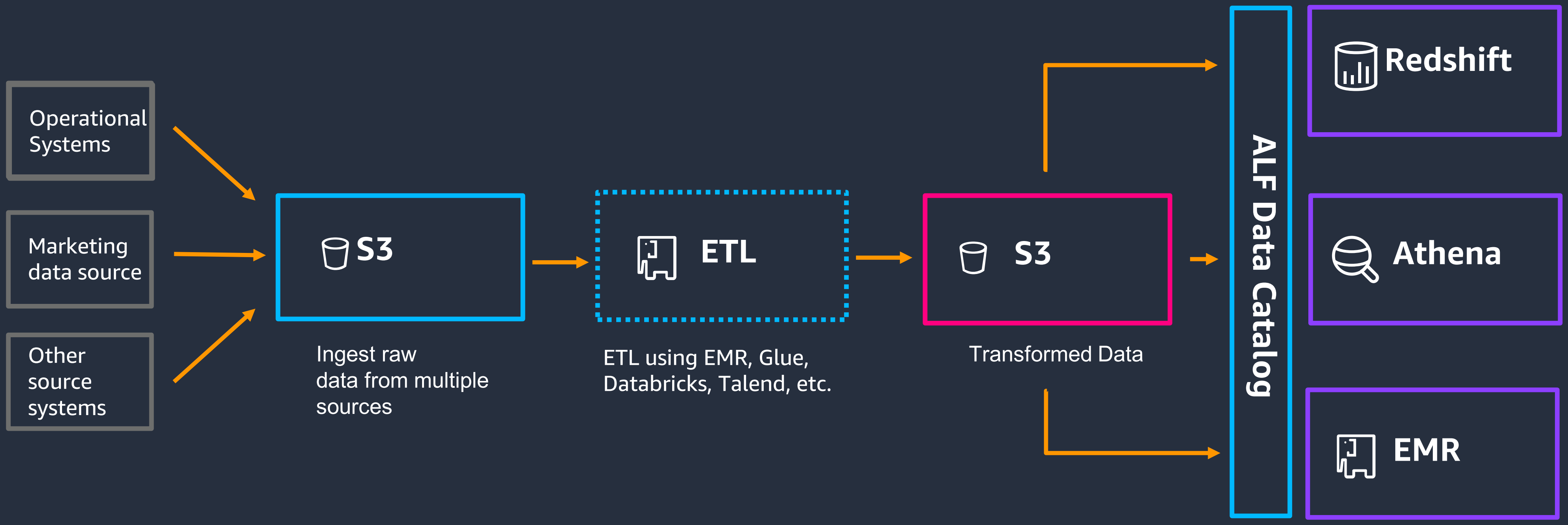
Centralized dashboard for all partner integration statuses

Reach out to redshift-partners@amazon.com to integrate your product into the Amazon Redshift console today!

Reference Architectures

Reference Architecture

Data Lake first approach - Collect raw data in S3, curate and load to target

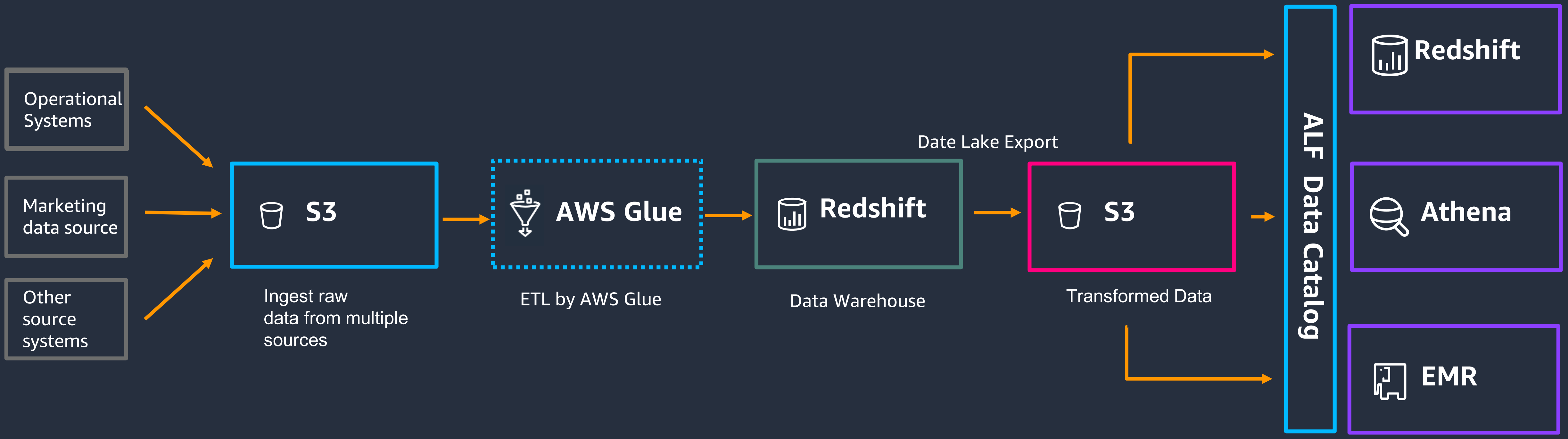


ALF = AWS Lake Formation Data Catalog

Note: Lake Formation leverages the Glue Data Catalog

Reference Architecture

Data warehouse first approach - Load raw data to Redshift and publish refined data on S3 Data Lake



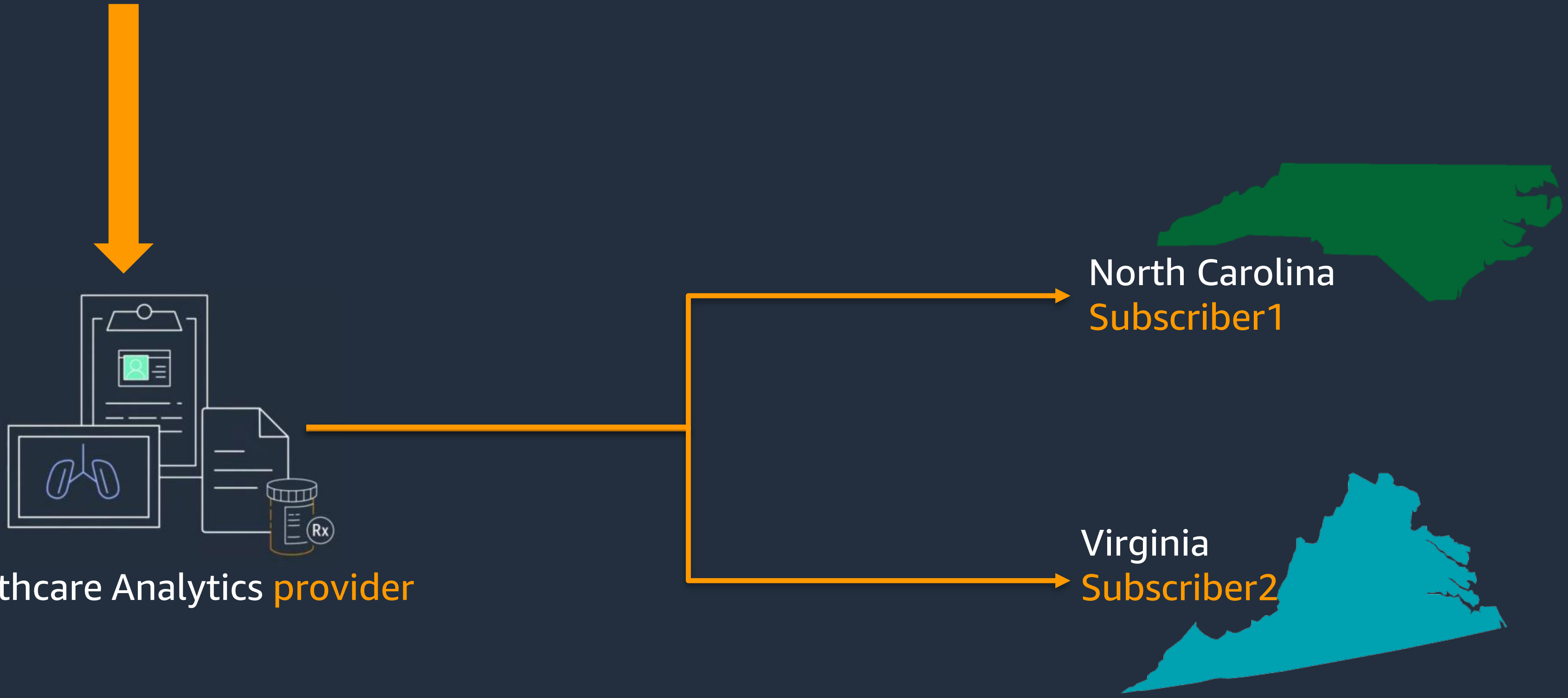
ALF = AWS Lake Formation Data Catalog

Note: Lake Formation leverages the Glue Data Catalog

Healthcare Analytics use case

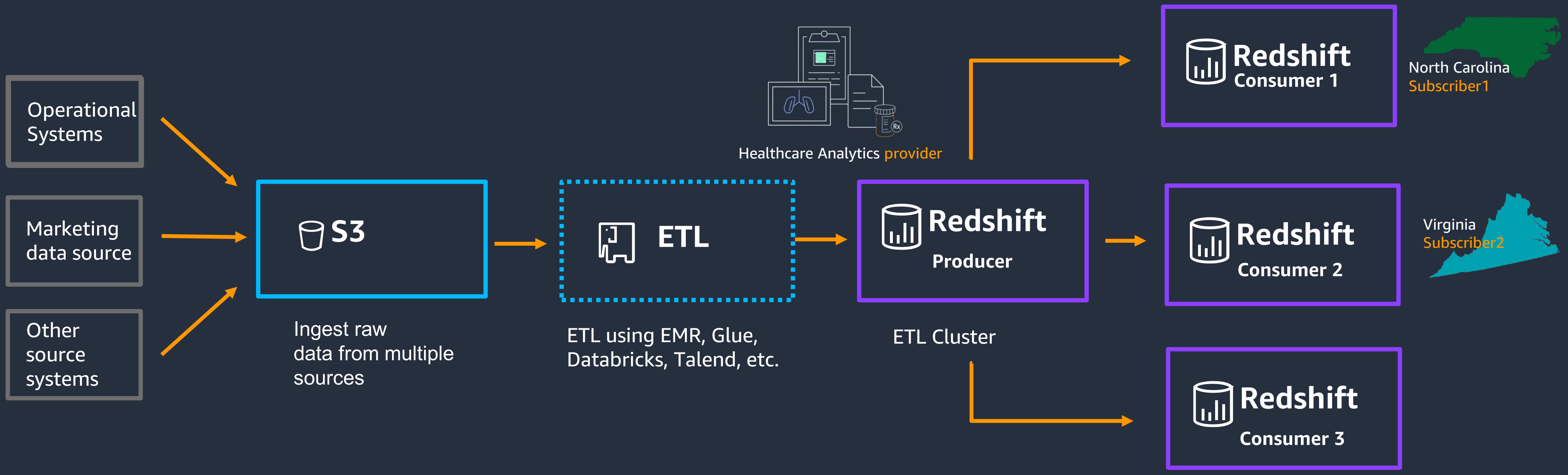
Analytics as a Service

Store & Share inpatient bed occupancy by state



Reference Architecture

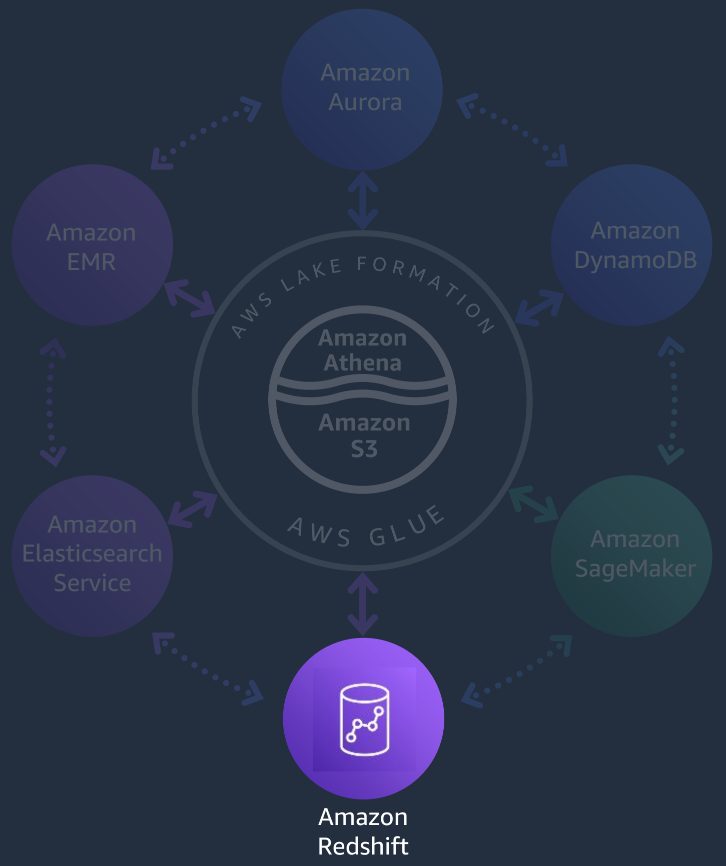
Subscriber Redshift clusters access provider Redshift cluster



Redshift Data Sharing between Producer & Consumers clusters

Amazon Redshift

Performance at any scale



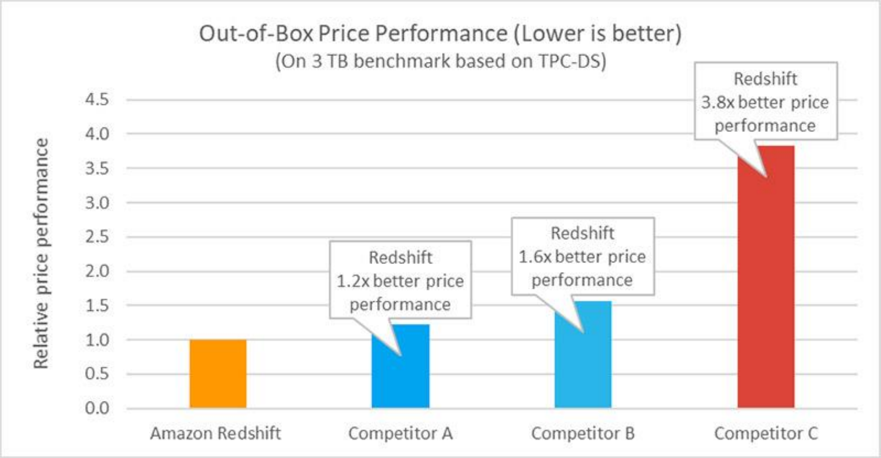
Analyze all your data
Deepest integration with your data lake

Performance at any scale
Up to 3x better price performance than other cloud DW

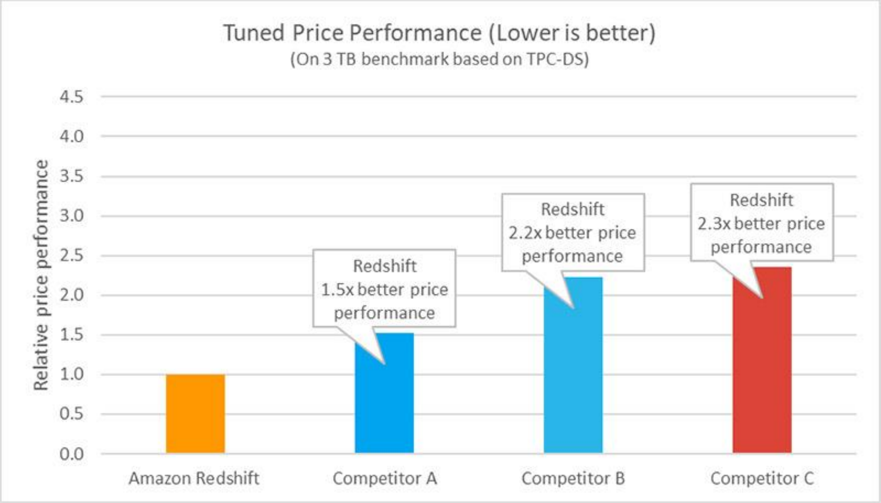
Lower your costs
At least 50% less expensive than other cloud DW

Get the best price performance and scale

Amazon Redshift has up to 3X better price performance than other cloud data warehouses


Out-of-Box Price Performance (Lower is better)

(On 3 TB benchmark based on TPC-DS)


Tuned Price Performance (Lower is better)

(On 3 TB benchmark based on TPC-DS)

Customers like Duolingo, Yelp, and Codecademy are doubling performance and saving costs with RA3

Amazon Redshift has up to 3x better price performance than other cloud data warehouses (see blog)

Scales data volume from 1TB to 1PB+ with predictable cost and performance

Boost queries up to 10x with AQUA for Amazon Redshift

Redshift Scalability

Data

Users

Workloads

Separation of Storage & Compute

> **Martin Brambley**
>
> **Sirocco Systems, Director**
>
> We saw an immediate 30 percent improvement in end-to-end ETL loading using the new DC2 node from Redshift. This is fantastic news for our clients as data volumes and demand for analytics continue to grow rapidly

RA3 nodes with managed storage

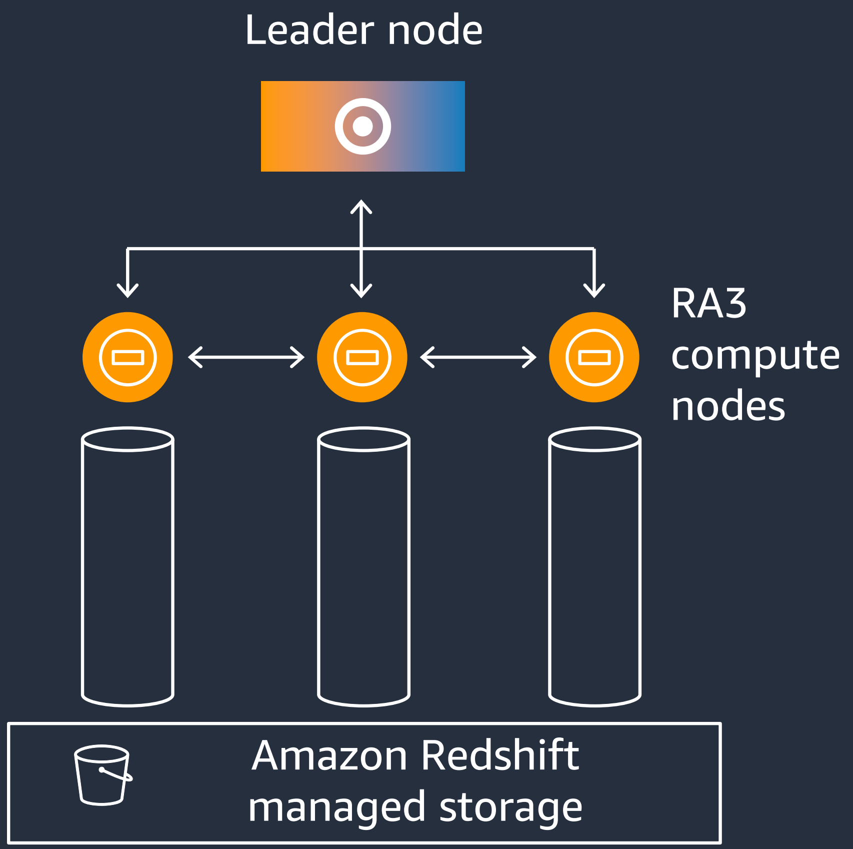
SCALE COMPUTE AND STORAGE INDEPENDENTLY

Managed storage

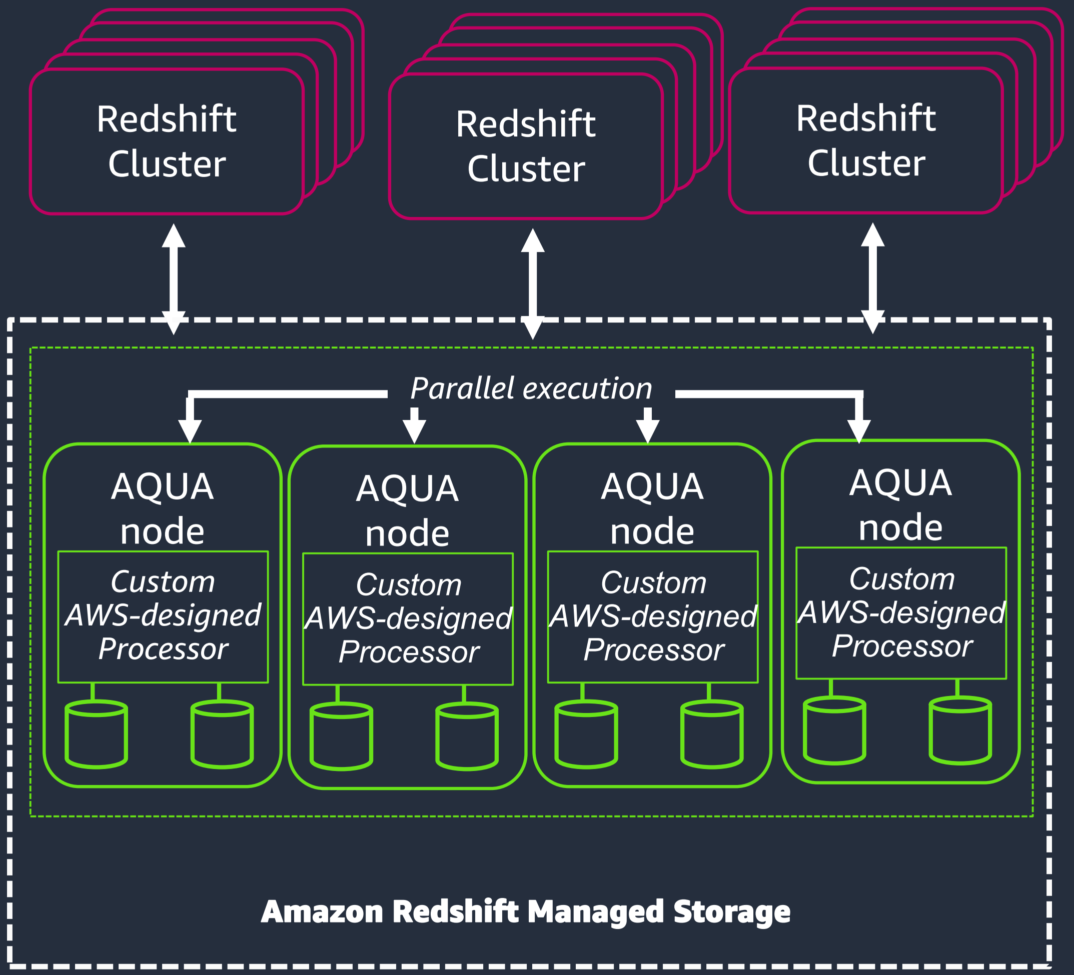
Large high-speed cache

High-bandwidth networking

- Size of data warehouse only based on steady state compute needs
- Scale and pay independently for compute and storage
- Automatic, no changes to any workflows, no need to manage storage



AQUA (Advanced Query Accelerator)


Redshift Cluster
Redshift Cluster
Redshift Cluster
Parallel execution
AQUA node
Custom AWS-designed Processor
AQUA node
Custom AWS-designed Processor
AQUA node
Custom AWS-designed Processor
AQUA node
Custom AWS-designed Processor
Amazon Redshift Managed Storage

New distributed & hardware-accelerated processing layer

With AQUA, Amazon Redshift is multiple times **faster** than any other cloud data warehouse, no extra cost

AQUA Nodes with custom AWS-designed analytics processors to make operations (compression, encryption, filtering, and aggregations) faster than traditional CPUs

Available with RA3 and no code changes required

Amazon Redshift automates performance tuning

ML-BASED OPTIMIZATIONS TO GET STARTED EASILY AND GET THE FASTEST PERFORMANCE QUICKLY

Auto property for tables – Dist, Sort

Continuously scans workload patterns and automates physical design optimization

Optimizes for peak performance as data and workloads scale

Leverages machine learning to adapt to shifting workloads

Can be enabled or disabled per table

svv_alter_table_recommendations logs the recommended changes

svl_auto_worker_action logs audit trail of changes

Automatic vacuum delete

NEW

Automatic distribution keys

NEW

Automatic sort keys

V2

Auto workload manager

Automatic table sort

NEW

MV auto-refresh and rewrite

Concurrency scaling

Compute elasticity and scalability to handle unpredictable user demand

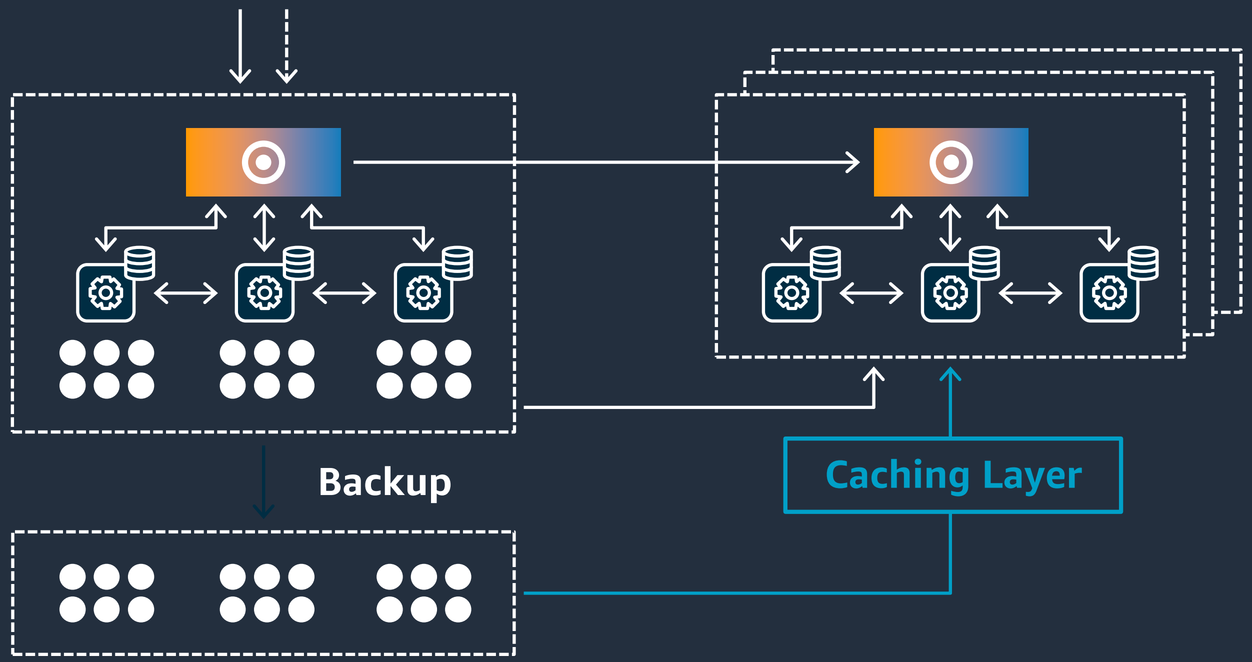
Size cluster on Steady State compute needs

Scale-out to multiple Redshift clusters from a single endpoint in seconds

Support virtually unlimited concurrent users while maintaining SLAs

Per-second billing for additional clusters used

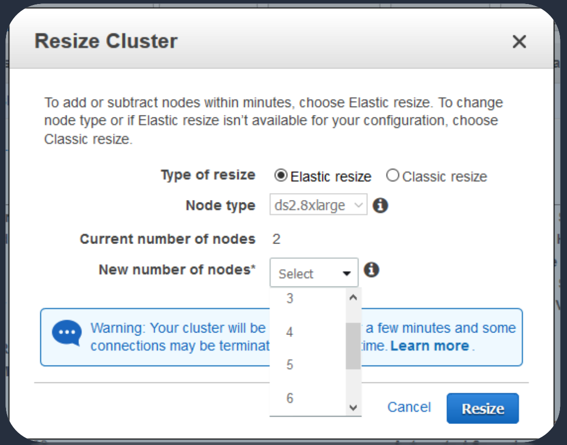
Free 1hr per day (free for 97% of clusters)



Redshift cluster resizing

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

- **Elastic Resize:**
  - In-place: Add or remove nodes to/from existing cluster
  - Scale-Out: Performance scales proportionally
  - Time: Completes within few minutes. Limited disruption to sessions and queries
  - Slice count: remains the same as original cluster
- **Classic Resize:**
  - New cluster: new cluster is provisioned and data copied
  - Time: Proportional to data volume in original cluster
  - Slice count: changes based on the new cluster



Resize Cluster ✕

To add or subtract nodes within minutes, choose Elastic resize. To change node type or if Elastic resize isn't available for your configuration, choose Classic resize.

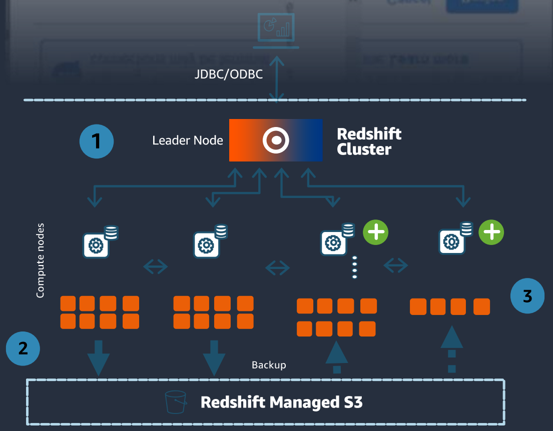
Type of resize: ● Elastic resize ○ Classic resize

Node type: ds2.8xlarge

Current number of nodes: 2

New number of nodes*: Select

Warning: Your cluster will be … a few minutes and some connections may be terminat… time. **Learn more**.



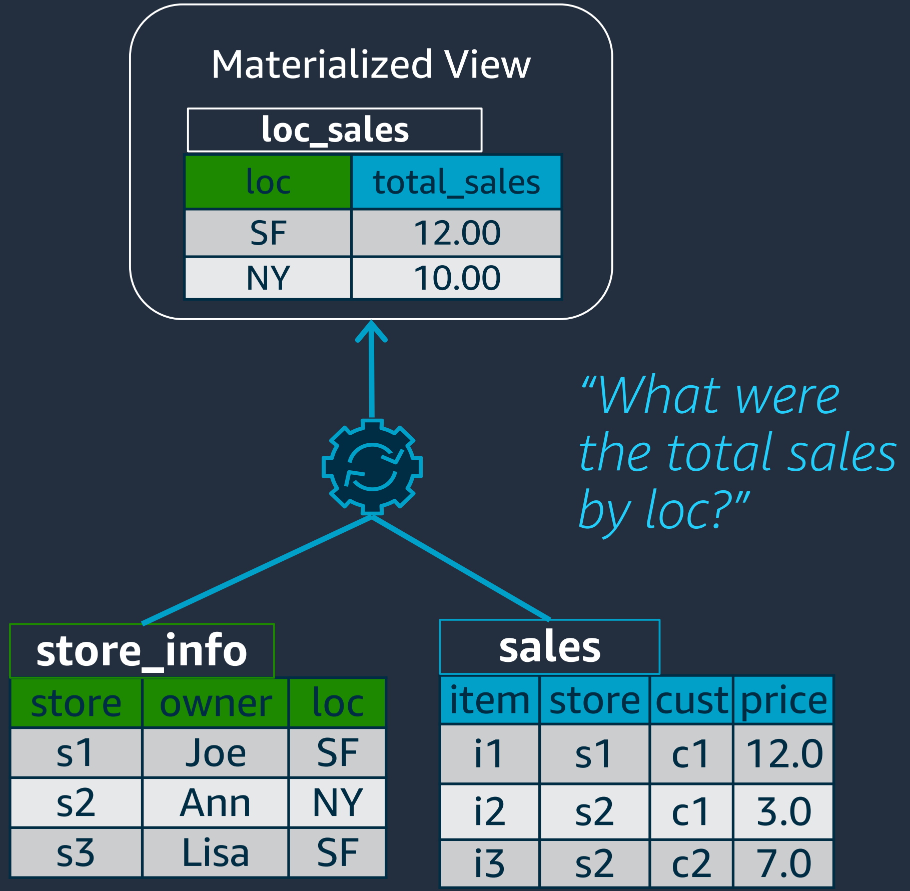
Materialized views

Compute once, query many times

- Speed-up queries by orders of magnitude
  - Joins, filters, aggregations, and projections

- Simplify and accelerate ETL/BI pipelines
  - Incremental refresh
  - Auto refresh
- Easier and faster migration to Redshift

- Auto-query rewrite
- Supports MV on Redshift local, Spectrum and Federated queries



Materialized View

loc_sales

loc	total_sales
SF	12.00
NY	10.00

"What were the total sales by loc?"

store_info

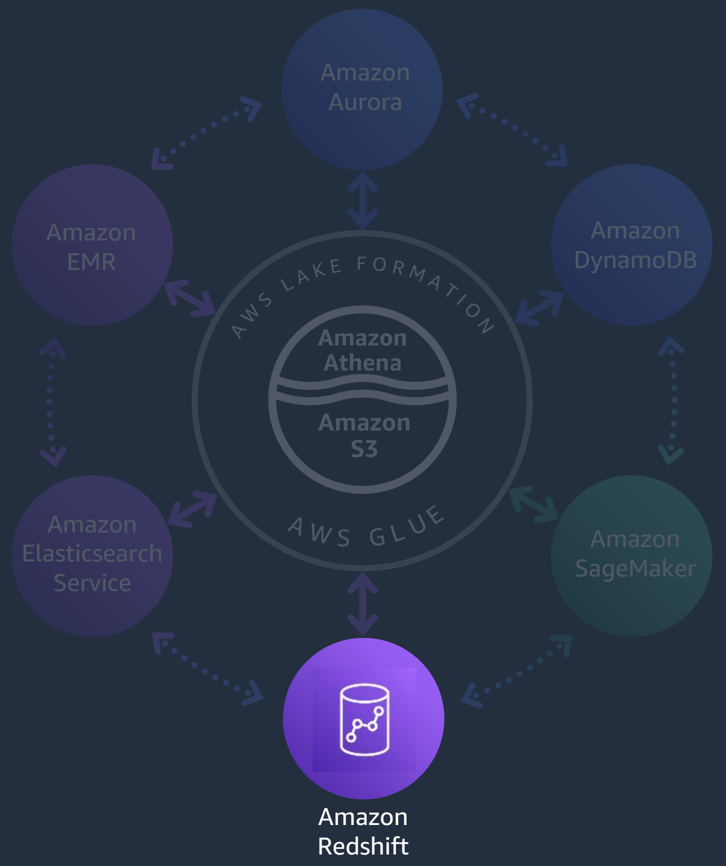
store	owner	loc
s1	Joe	SF
s2	Ann	NY
s3	Lisa	SF

sales

item	store	cust	price
i1	s1	c1	12.0
i2	s2	c1	3.0
i3	s2	c2	7.0

Amazon Redshift

Lower your costs



Analyze all your data

Deepest integration with your data lake

Performance at any scale

Up to 3x better price performance than other cloud DW

Lower your costs

At least 50% less expensive than other cloud DW

Built-in security and compliance

SECURITY AND COMPLIANCE FEATURES WITH NO EXTRA COSTS WITH AMAZON REDSHIFT

Authentication

- IAM integration
- ID federation
- Azure AD, AD, Okta, Ping Federate
- Multifactor authentication

Access control

- Column-level privileges for Amazon Redshift and data lake

Audit

- AWS CloudTrail integration
- Enable Audit logging to S3

  Connection, user, user activity

Encryption

- AWS KMS integration
- Encrypted data in motion, data at rest
- Tokenization with Lambda UDFs and third-party tools

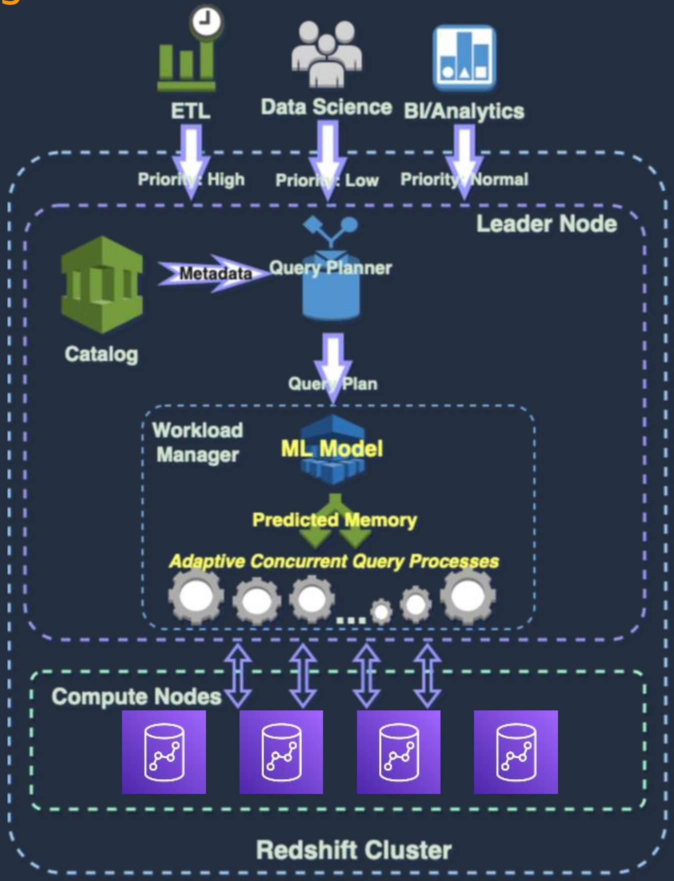
Helps achieve compliance

- SOC
- PCI
- FedRAMP
- HIPAA and others

Additional Documentation: Security Overview

Enhanced Auto WLM – Adaptive Concurrency

ML based memory prediction algorithm maximizes system resource usage and optimizes throughput with adaptive concurrency and feedback mechanisms



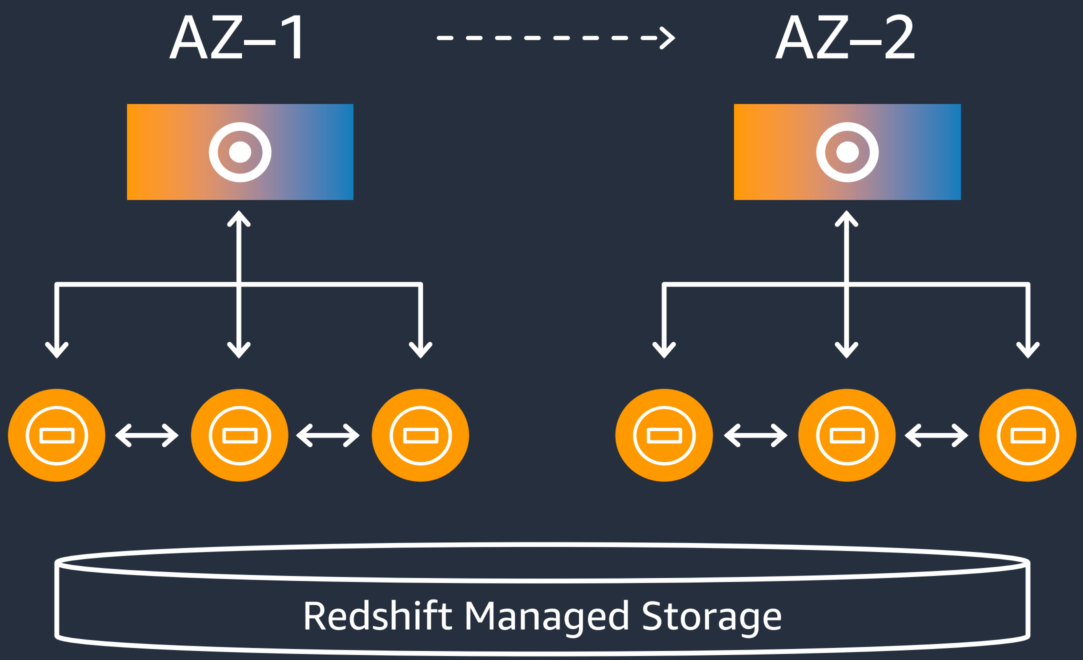
Based on the query traffic and resource utilization, Amazon Redshift automatically determines the number of concurrent queries to **optimize query throughput**

Amazon Redshift's Auto WLM allows you to create user or query group queues based on business function, priority, and SLA

Resilience: Cross-AZ cluster recovery

RELOCATE YOUR CLUSTER IN ANOTHER AZ IN RESPONSE TO FAILURE

- ✓ Amazon Redshift has a service SLA of 99.9%
- ✓ Recovery with zero data loss (RP = Zero)
- ✓ No need to restore from a snapshot
- ✓ On-demand failover
- ✓ Cluster is created in another AZ on-demand, so cost of a standby replica cluster is avoided
- ✓ Supported on the RA3 instance family



Amazon Redshift Data API

Simplifies data access from web services based applications

Simplifies data access from languages such as Python, Go, Java, Node.js and other languages

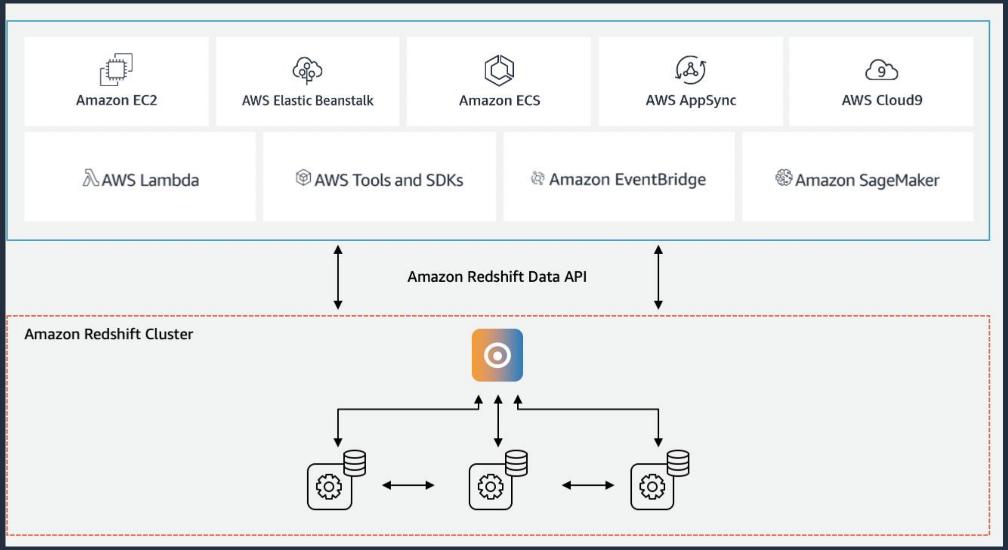
Query, Load and Unload data from CLI/SDK

Do not have to worry about configuring drivers, connection pools

Leverages IAM credentials or Secrets manager

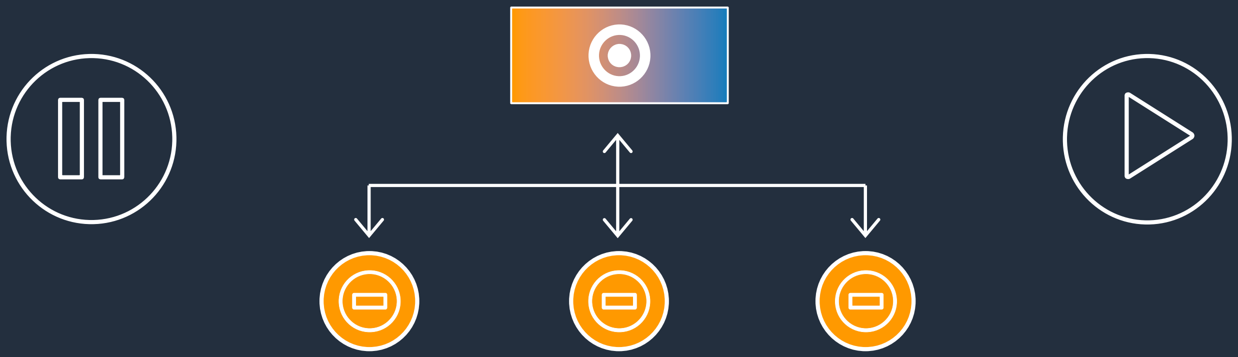
Improved notebook experience

```
aws redshift-data execute-statement
--database [DATABASE]
--query [QUERY]
--secret-arn [CREDENTIALS_ARN]
```



Cost optimization: pause/resume cluster

Easily start and stop a cluster to save cost for intermittent workloads



Pause a development or QA cluster during non-business hours

Reserved Instances can be shared between clusters if not running at the same time

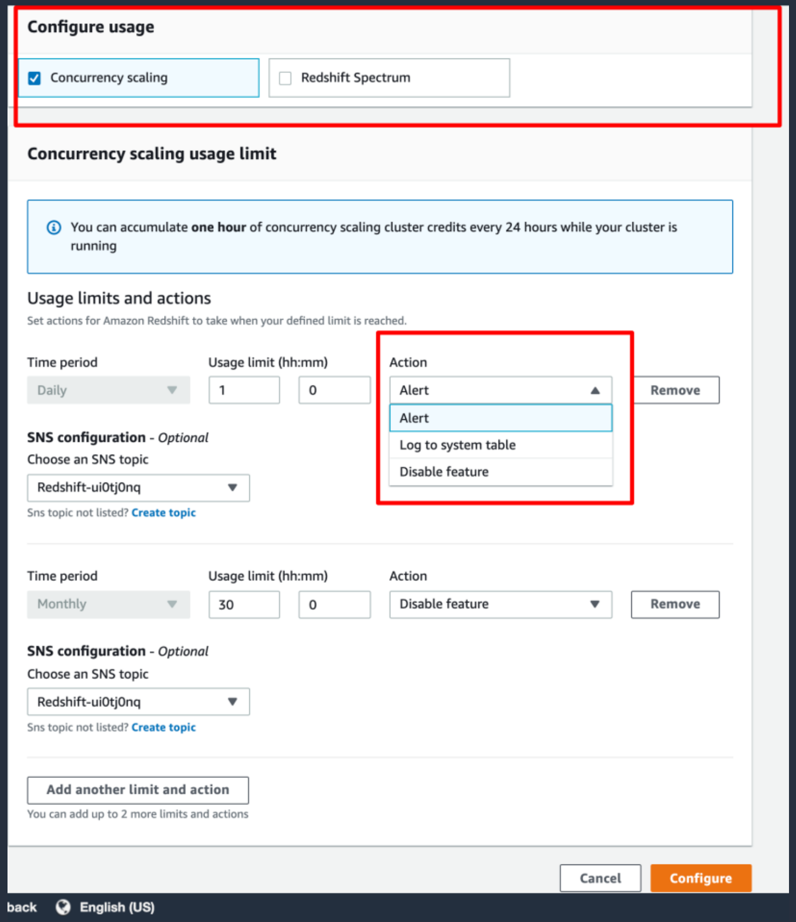
Cluster used for ingest/transform/unload to the Data Lake can be Paused after the load process.

Pause and Resume cluster operations with 1-click in the console or via API

Pause and Resume using a user-defined schedule

Cost controls

Control your spend by creating usage limits for concurrency scaling and Amazon redshift spectrum



Configure usage

☑ Concurrency scaling　□ Redshift Spectrum

Concurrency scaling usage limit

ⓘ You can accumulate **one hour** of concurrency scaling cluster credits every 24 hours while your cluster is running

Usage limits and actions

Set actions for Amazon Redshift to take when your defined limit is reached.

Time period	Usage limit (hh:mm)		Action	
Daily	1	0	Alert	Remove

SNS configuration - *Optional*

Choose an SNS topic: Redshift-ui0tj0nq

Sns topic not listed? **Create topic**

Time period	Usage limit (hh:mm)		Action	
Monthly	30	0	Disable feature	Remove

SNS configuration - *Optional*

Choose an SNS topic: Redshift-ui0tj0nq

Sns topic not listed? **Create topic**

You can add up to 2 more limits and actions

Create daily, weekly, and monthly limits

Up to 4 limits per feature

Log to system table, Alert, Disable feature

Automatically generate Amazon CloudWatch alarm

Next steps

Already using Redshift?

- ✓ Ask your account team for a free Redshift optimization session
- ✓ Learn more about what’s new

https://aws.amazon.com/redshift/whats-new/

Getting started with Redshift?

- ✓ Ask your account team for a 10-minute demo
- ✓ Learn more about starting a proof of concept

https://docs.aws.amazon.com/redshift/proof-of-concept-playbook.html

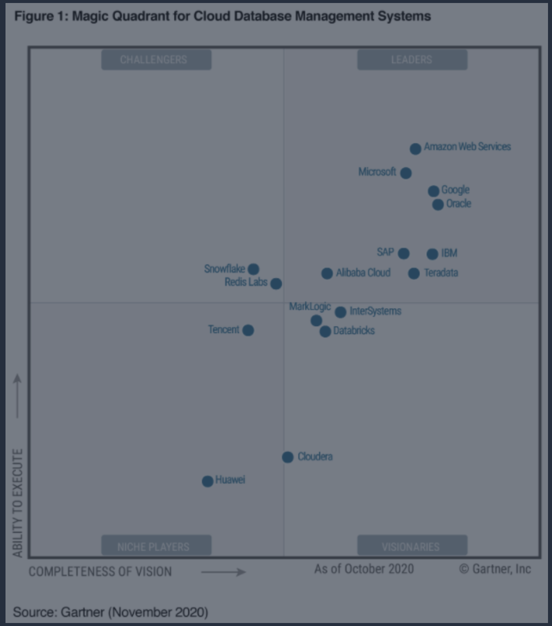
- ✓ Or request help from AWS

https://pages.awscloud.com/redshift-proof-of-concept-request

Thank You!

Gartner Magic Quadrant 2021

Figure 1: Magic Quadrant for Cloud Database Management Systems



Source: Gartner (November 2020)

AWS named a Leader in the Cloud Database Management Systems Magic Quadrant with the highest score in Ability to Execute among the 16 vendors evaluated