
AWS
re:Invent

ANT202

How Amazon Redshift powers large-scale analytics for Amazon.com

Maor Kleider

Senior Manager, Amazon Redshift
Product and Database Engineering

Ryan Wilke

Solutions Architect, Amazon.com,
Business Data Technologies

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Data analytics at scale
- The evolution of Amazon Redshift
- Amazon.com journey with Amazon Redshift
- Amazon.com RA3 architecture and usage

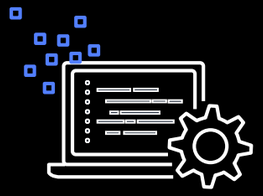


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data analytics at scale



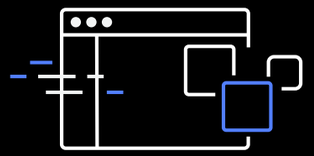
Multiple analytics needs



Multiple systems



Data volume, variety, velocity



Performance, concurrency



Security, compliance

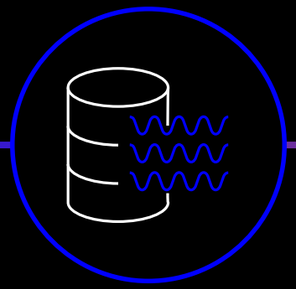


Cost, flexibility

Amazon Redshift

THE MOST WIDELY USED CLOUD DATA WAREHOUSE WITH TENS OF THOUSANDS OF CUSTOMERS

ANALYZE ALL YOUR DATA



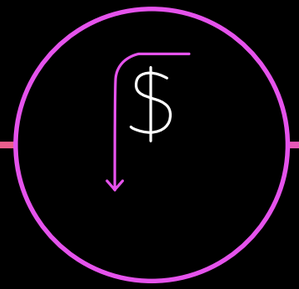
Take a **lake house approach** by analyzing all your data across your data warehouse, your Amazon S3 data lake, and operational databases with consistent security and governance policies

PERFORMANCE AT ANY SCALE



Get up to **3x better price performance** than other cloud data warehouses with a **self-tuning** system, and boost queries up to **10x with AQUA**

LOWER YOUR COSTS

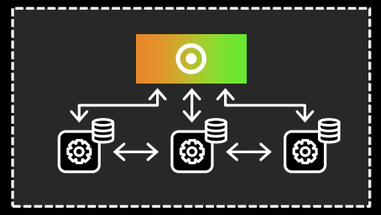


Start small and pay only for what you use with **predictable** monthly costs; Amazon Redshift is at least **half the cost of** other cloud data warehouses

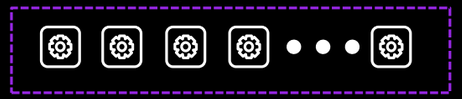
The evolution of Amazon Redshift (2017–2020)

THE MOST POPULAR AND FASTEST CLOUD DATA WAREHOUSE

Redshift Spectrum for data lake analytics



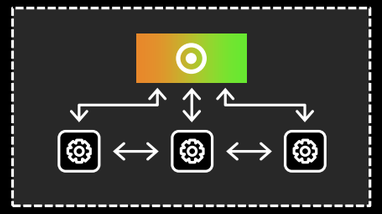
Spectrum



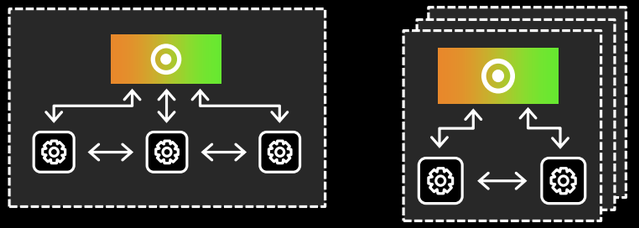
Concurrency scaling for bursty workloads



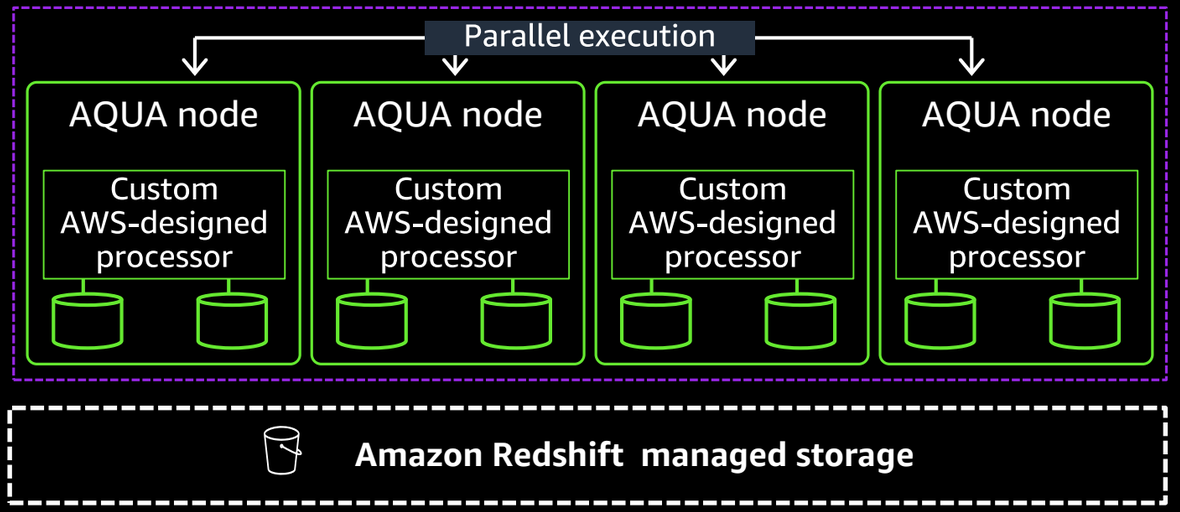
RA3 with independent compute and storage scaling



Data sharing across clusters (preview)



Query acceleration with computational cache





RA3 nodes with Redshift managed storage

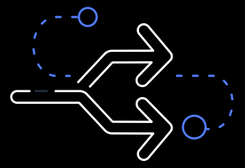
SCALE COMPUTE AND STORAGE INDEPENDENTLY



Managed storage

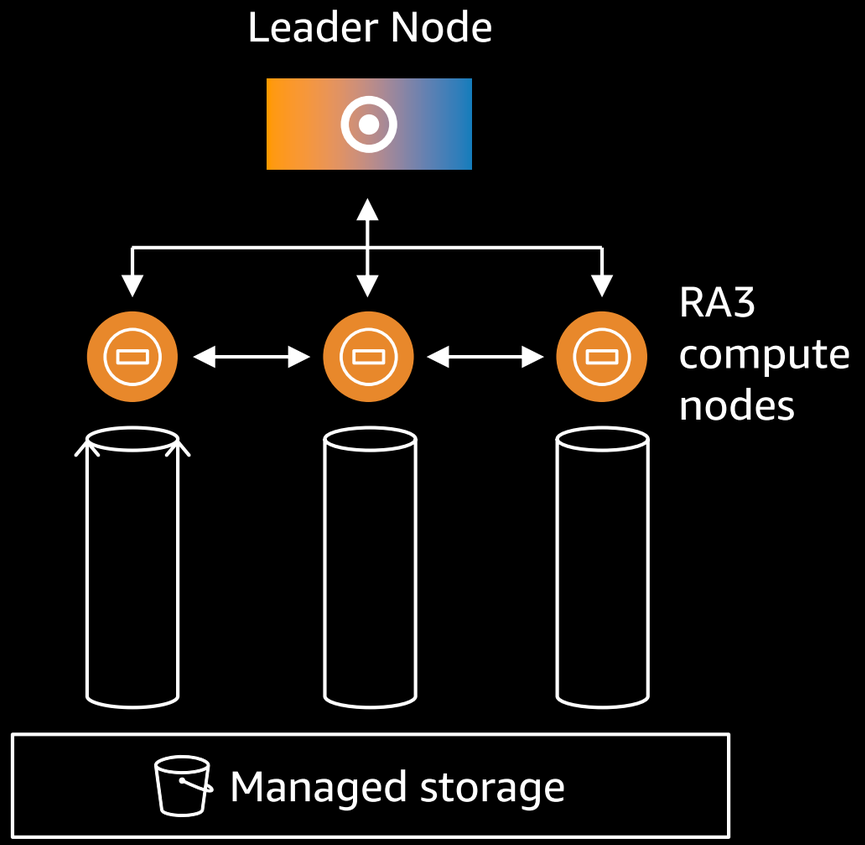


Large high-speed cache



High-bandwidth networking

- Size 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



RA3 compute nodes

MULTIPLE NODE SIZES AVAILABLE TO PROVISION THE RIGHT AMOUNT OF COMPUTE

RA3.16xl

48 vCPU, 384 GiB memory, 8 GB IO

On-demand price – $13.04 / node / hour

RA3.4xl

12 vCPU, 96 GiB memory, 2 GB IO

On-demand price – $3.26 / node / hour

NEW

RA3.xlplus

4 vCPU, 32 GiB memory, 650 MB IO

On-demand price – $1.09 / node / hour

Scale to tens of PB of data, pay as you go at $0.024 / GB-month for storage

Save more with reserved instances – 33% discount over OD with 1-year and 60% with 3-year

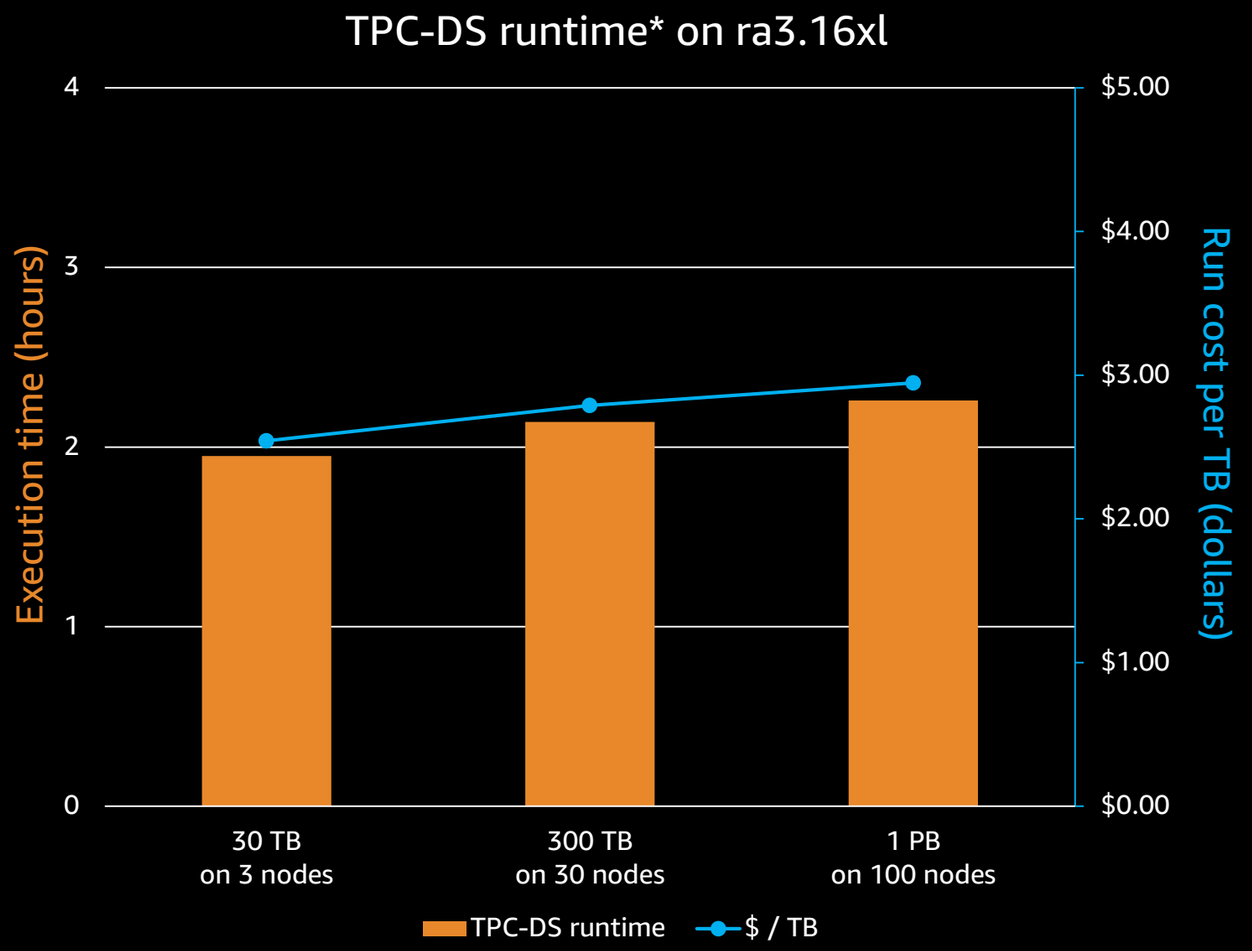
Amazon Redshift: Fast performance at any scale

AMAZON REDSHIFT PROVIDES UP TO 3X BETTER PRICE PERFORMANCE THAN OTHER CLOUD DATA WAREHOUSES

Performance scales with data volume, from 1TB to 1PB+

Predictable costs as data volume grows in size

Simple scaling of compute power independently of storage



* Execution of TPC-DS 2.10 Power Run

Concurrency scaling

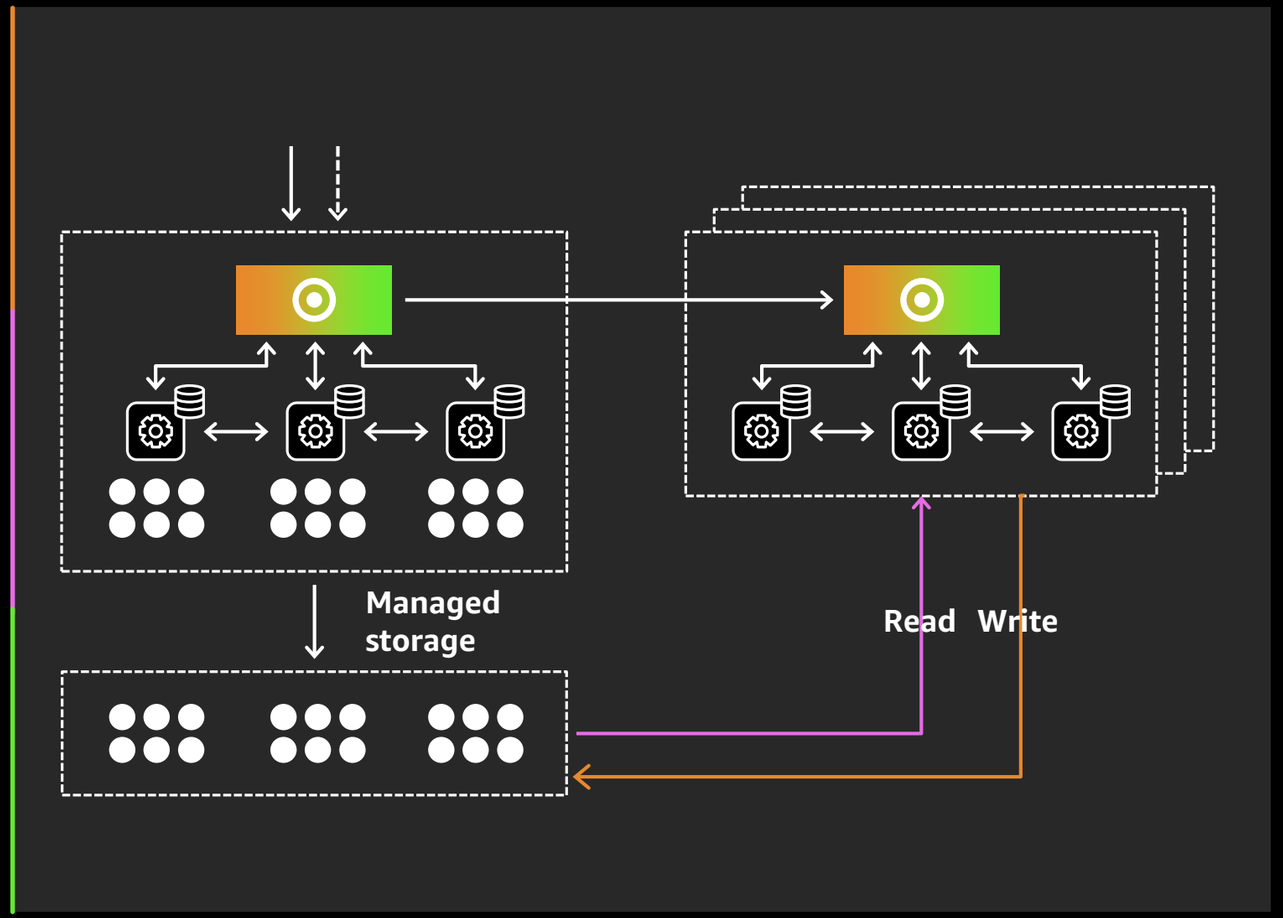
COMPUTE ELASTICITY AND SCALABILITY TO HANDLE UNPREDICTABLE USER DEMAND

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

Support virtually unlimited concurrent users and queries while maintaining SLAs

Per-second billing for additional clusters used

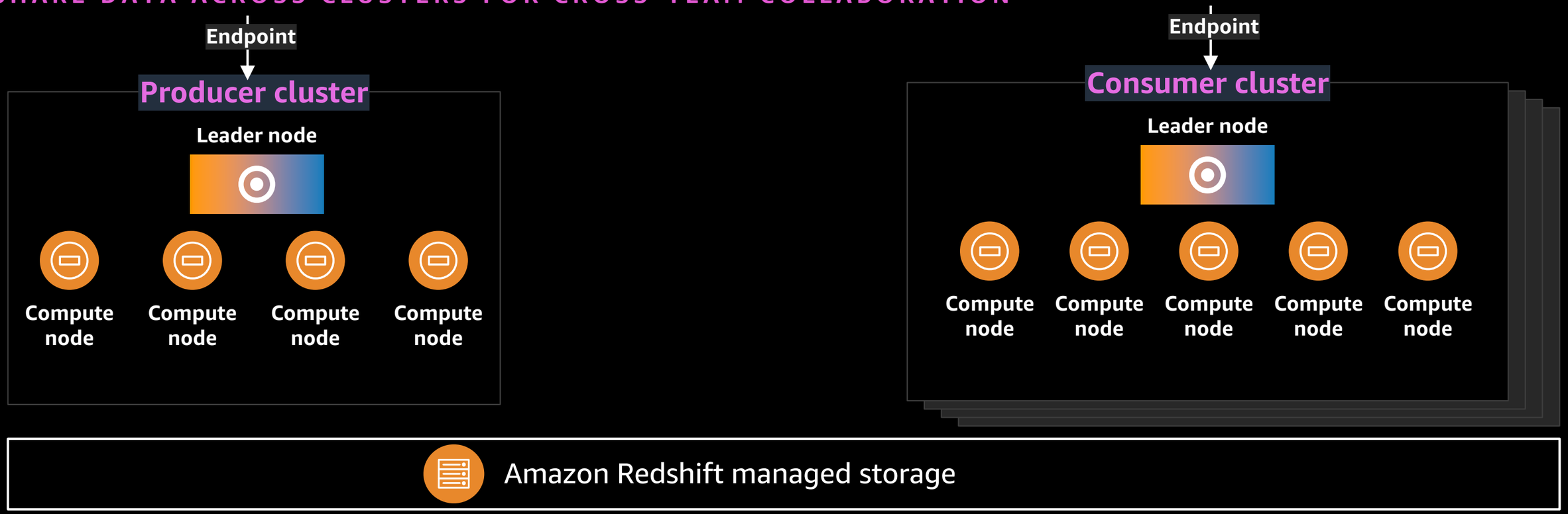
Cost controls and free one hour usage per day



Data sharing

PREVIEW

SHARE DATA ACROSS CLUSTERS FOR CROSS-TEAM COLLABORATION



Simple and direct way to share data across Amazon Redshift clusters

Instant, granular, and high performance data access without data copies and data movement

Live and consistently updating views of data across all consumers

Secure and governed collaboration within and across organizations and with external parties

Native semi-structured data support

PREVIEW

New data type: **SUPER**

Easy, efficient, and powerful JSON processing

Fast, row-oriented data ingestion

Fast, column-oriented analytics with materialized views over SUPER

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

```
{
"id":1,
"name":{"given":"Jane", "family":"Doe"},
"phone":[{"type":"work", "num": "9255550101"},
         {"type":"cell", "num": 6505550190}]
}
{
"id":2,
"name":{"given":"Jorge", "family":"Souza"},
"phone":[{"type":"work", "num": 5105550188}]
}
```

```
SELECT name.given AS firstname, ph.num
FROM customers c, c.phone ph
WHERE ph.type = 'cell';

firstname | num
----------+---------------
"Jane"    | 6505550190
```

Amazon.com large-scale analytics



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Amazon.com analytics mission

“Accelerate business insights and data-driven innovation by providing secure, compliant, reliable, intuitive, and cost-efficient technologies that our customers trust.”

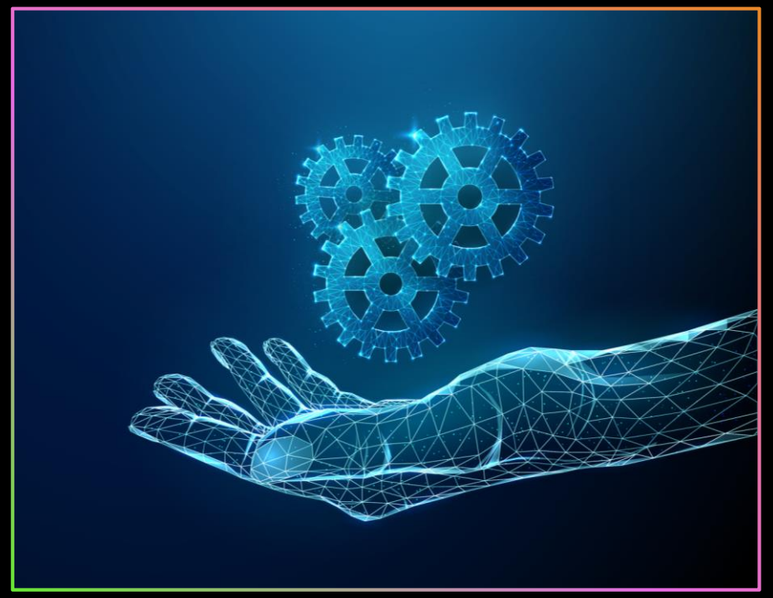
Amazon.com large-scale analytics



Diverse user base
and use cases

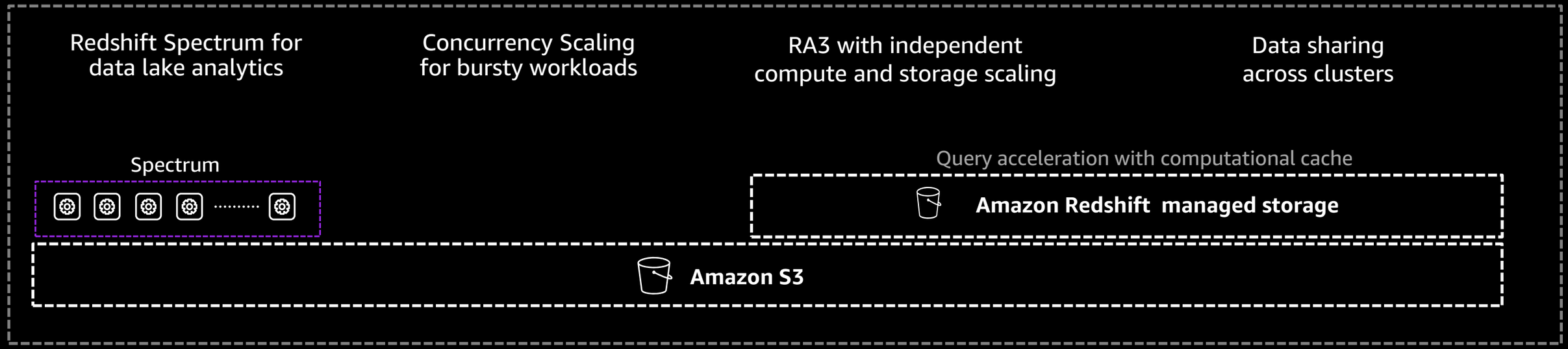


~38K active data sets;
Curated business data;
Over ~500PB in Andes data lake



~900K daily jobs
~80K active users

Our analytics journey with AWS Redshift (2016–2020)



2016–2017

Established Redshift DB as future of Amazon.com, replacing old DW DB tech

Redshift Spectrum revolutionized how customers could access large datasets with Redshift

2017–2019

Fully invested in and migrated to Amazon Redshift with data lake / marketplace (500PB data)

2020

RA3 node adoption to provide performance and cost-effective scalability boost for Amazon.com

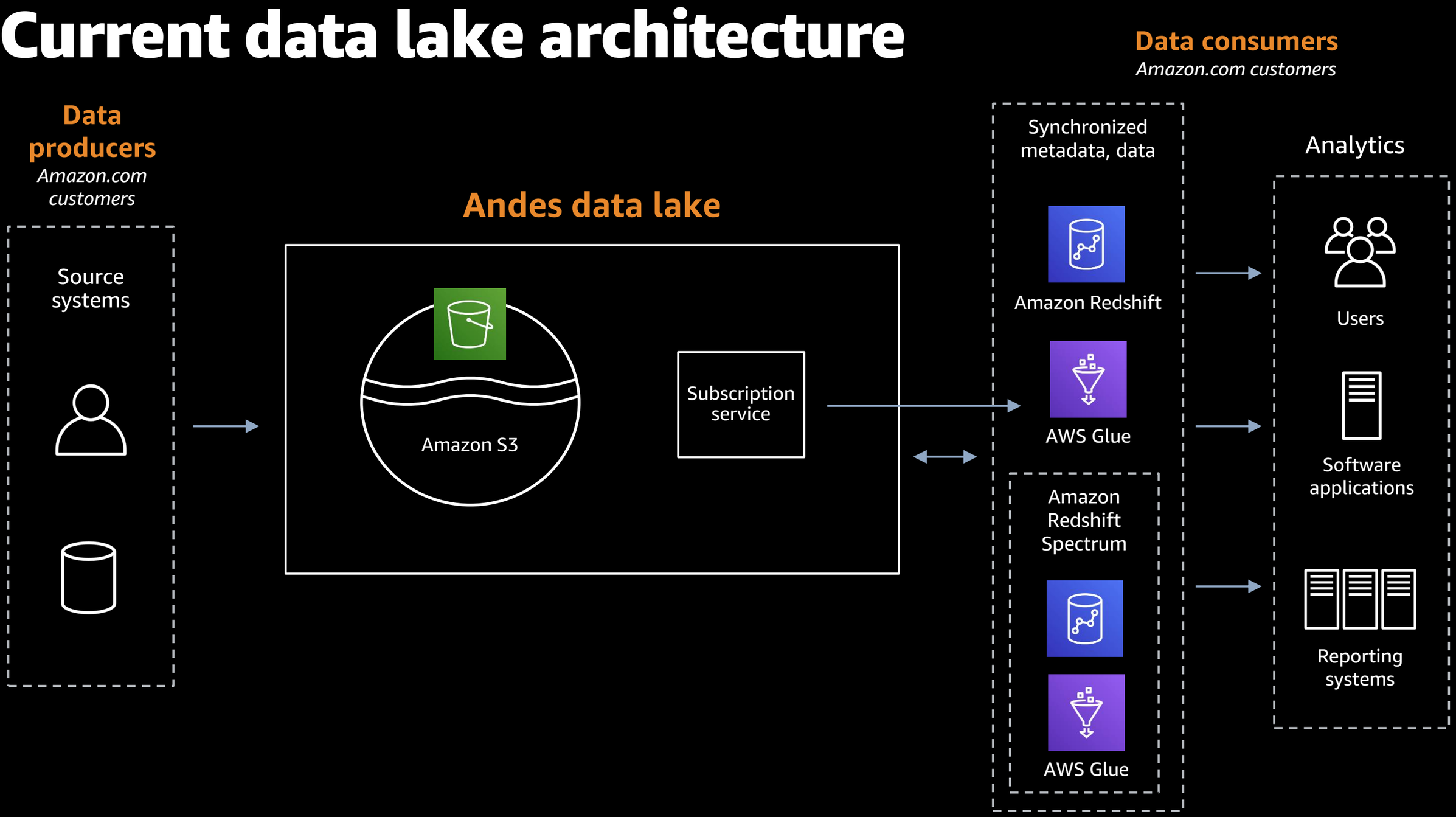
Data lake architecture



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Current data lake architecture



Amazon Prime and Prime Video large-scale architecture

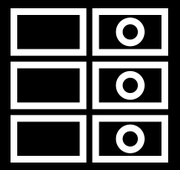
amazon Prime

prime video

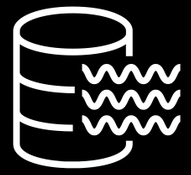
Data producers
(Amazon.com customers)

Data lake and marketplace

Data consumers
(Amazon.com customers)



Data input



Data lake



Amazon Redshift

Alexa BI and shipping large-scale architecture

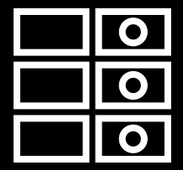
amazon shipping

amazon alexa

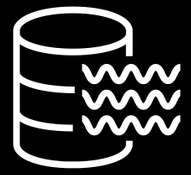
Data producers
(Amazon.com customers)

Data lake and marketplace

Data consumers
(Amazon.com customers)



Data input



Data lake



Amazon Redshift

Amazon Redshift RA3 node adoption

Amazon.com challenges

Customers want the ability to scale their data storage needs without over-provisioning compute capacity

- Scalable and cost effective
- Reduce intensive load operations on individual instances with fewer duplicate ETL pipelines

Amazon.com users ultimately desire

- Better query performance
- Reduction in cost to operate

Amazon Redshift RA3 node adoption

RA3 as solution

BDT Org performed a proof of concept for identifying the benefits of migrating from DS2 node type to the new RA3 node type with the same daily workload

Key RA3 benefits found

- Transforms and extracts 1.3x faster
- Loads 1.4x faster
- Managed storage flexibility and scale
  - High performance SSDs for your hot data and Amazon S3 for cold data
- High bandwidth S3 networking built on the AWS Nitro System
- Storage costs separate from compute costs
- Data sharing capability

Exciting future with Amazon Redshift

Materialized views (GA)

- Precomputed result set
- Results returned much faster than standard data pull
- Useful for many Amazon.com ETL pipelines (internally / externally)

AQUA (GA Jan. 2021)

- **A**ccelerates Amazon Redshift queries
- Runs data-intensive tasks closer to the storage layer

Data sharing (Preview)

- Simple and direct way to share data across Redshift clusters and with other AWS analytic services
- Reduce and / or remove need to copy data directly to team instance

SUPER data type (Preview)

- Native semi-structured data support in Amazon Redshift
- Supports nested JSON files
- Amazon.com users can leverage the powerful PartiQL language, including unnesting

Next steps

Already using Amazon 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 Amazon 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!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.