Amazon Redshift Reimagined: RA3 and AQUA

Debu Panda

Senior Product Manager, Amazon Redshift

Amazon Web Services

AWS re:Invent

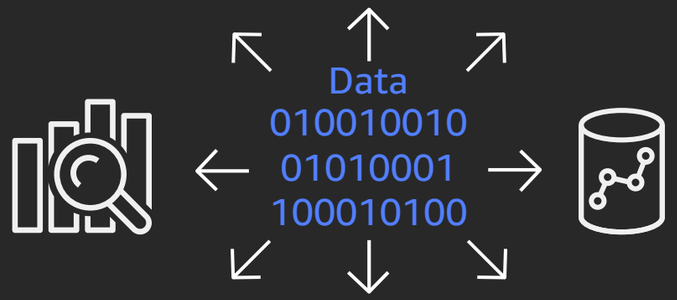
aws

Agenda

- Data Warehouse trends
- Innovations in Amazon Redshift
- Amazon Redshift new instance RA3 and AQUA
- Ease of use
- Data lake integration
- Performance Improvements
- Next steps
- Q & A

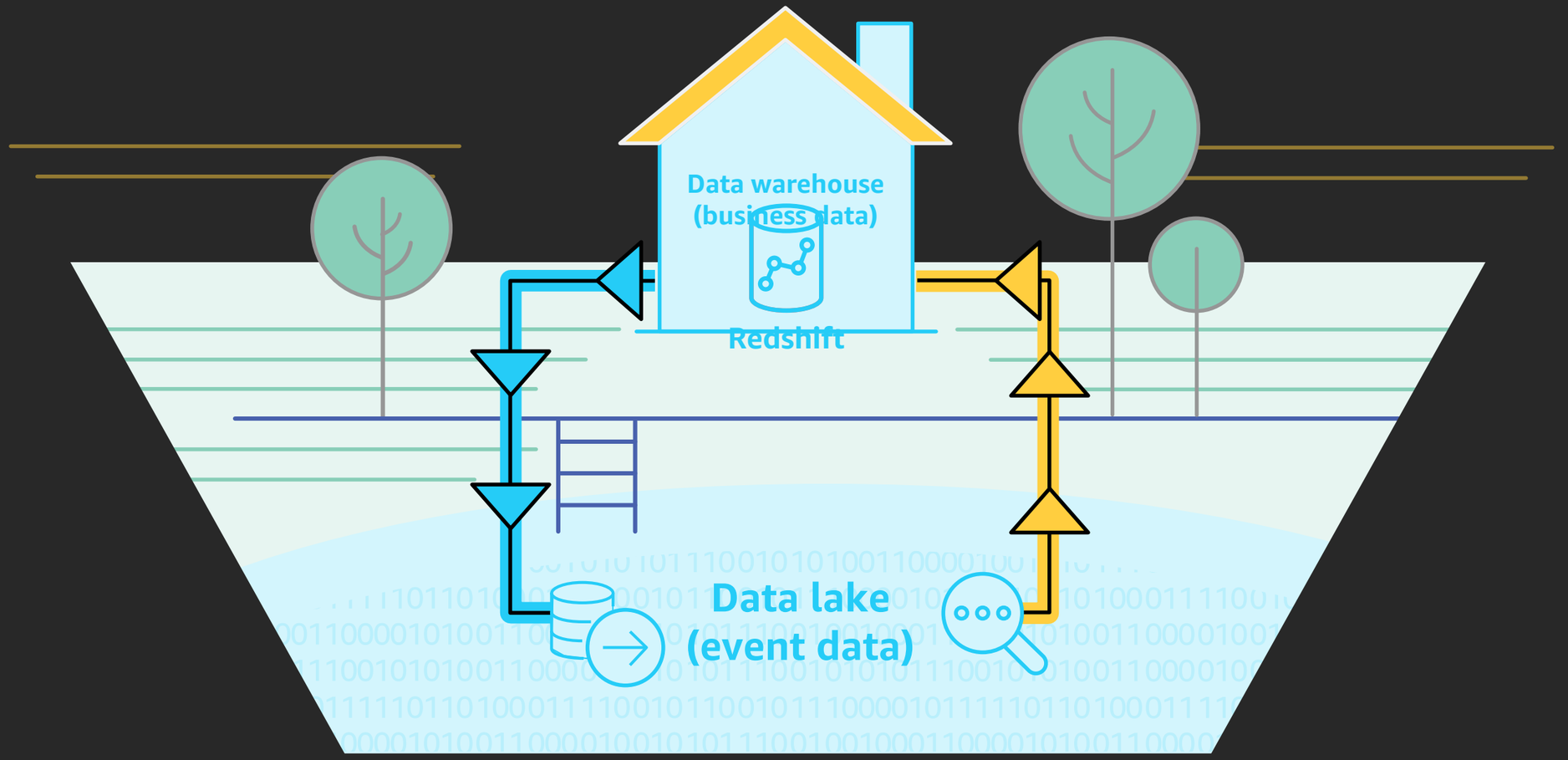
Our customers are asking for…

Migrations to the cloud



Exponential growth of event data

End-to-end insights from analyzing all your data



Customers moving to data lake architectures

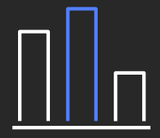
Redshift enables you to have a lake house approach

Amazon Redshift benefits

10’s of thousands of customers use Amazon Redshift to process exabytes each day

Data lake & AWS integrated

Lake Formation catalog & security, Exabyte scale query (spectrum & federated), AWS integrated (DMS, CloudWatch)



Best performance

Up to 3x faster than other cloud data warehouses

Lowest cost

Up to 75% less than other cloud data warehouses & predictable costs

Most scalable

Virtually unlimited elastic linear scaling

Most secure & compliant

AWS-grade security, (e.g. VPC, encryption with KMS, Cloud Trail), Certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPAA

Fully managed

Easy to provision & manage, automated backups, AWS support, 99.9% SLAs

Amazon Redshift

is the most popular cloud data warehouse

21st Century Fox
23andMe
amazon
AMGEN
ancestry
AstraZeneca
CIVIS
coinbase
coursera
DOW JONES
edmunds
EA
EQUINOX
FT Financial Times
FOURSQUARE
HEARST
iflix
intuit
Jack in the box
Kimberly-Clark
Klarna
Liberty Mutual
London Stock Exchange
MANTHAN
MERCK
Nasdaq
The Ohio State University
OPTUM
PADDYPOWER
betfair
Pfizer
Prezi
QANTAS
REDFIN
robinhood
SCHOLASTIC
Schumachergroup
SCOPELY
SEIZE THE PLAY
SOUNDCLOUD
Sysco
tinder
yelp

Amazon Redshift has been innovating quickly

Robust result set caching

Large # of tables support ~20000

Copy command support for ORC, Parquet

IAM role chaining

Elastic resize

Groups

Amazon Redshift Spectrum: date formats, scalar json and ION file formats support, region expansion, predicate filtering

Auto analyze

Health and performance monitoring w/Amazon CloudWatch

Automatic table distribution style

CloudWatch support for WLM queues

Performance enhancements—hash join, vacuum, window functions, resize ops, aggregations, console, union all, efficient compile code cache

Unload to CSV

Auto WLM

~25 Query Monitoring Rules (QMR) support

AQUA

Concurrency Scaling

200+

new features in the past 18 months

Manage multi-part query in AWS console

Auto analyze for incremental changes on table

DC1 migration to DC2

Resiliency of ROLLBACK processing

Spectrum Request Accelerator

Apply new distribution key

Amazon Redshift Spectrum: Row group filtering in Parquet and ORC, Nested data support, Enhanced VPC Routing, Multiple partitions

Faster Classic resize with optimized data transfer protocol

Performance: Bloom filters in joins, complex queries that create internal table, communication layer

Amazon Redshift Spectrum: Concurrency scaling

Amazon Lake Formation integration

Auto-Vacuum sort, Auto-Analyze and Auto Table Sort

Auto WLM with query priorities

Snapshot scheduler

Performance: join pushdowns to subquery, mixed workloads temporary tables, rank functions, null handling in join, single row insert

Advisor recommendations for distribution keys

AZ64 compression encoding

Console redesign

Stored procedures

Spatial Processing

Column level access control with AWS lake formation

RA3

Performance of Inter-Region Snapshot Transfers

Federated Query

Materialized Views

Manual Pause and Resume

Amazon Redshift: Key 2019 innovations

Performance & scalability

- NEW! RA3 with Amazon Redshift managed storage
- PREVIEW! AQUA for Amazon Redshift
- NEW! AZ64
- PREVIEW! Materialized views
- NEW! Spatial data support
- PREVIEW! Query federation across Amazon Redshift & Aurora
- NEW! Concurrency Scaling

Ease of use

- NEW! New management console
- NEW! Auto WLM: Query priorities
- NEW! Elastic resize scheduler
- NEW! Auto-Vacuum, Auto-Analyze & Auto Table Sort
- NEW! Stored procedures
- NEW! Distribution and sort key advisor
- NEW! Cross-instance restore
- NEW! Faster cross regional copy and change
- NEW! Auto Data Distribution
- NEW! Deferred Maintenance
- NEW! Elastic resize

Data lake integration

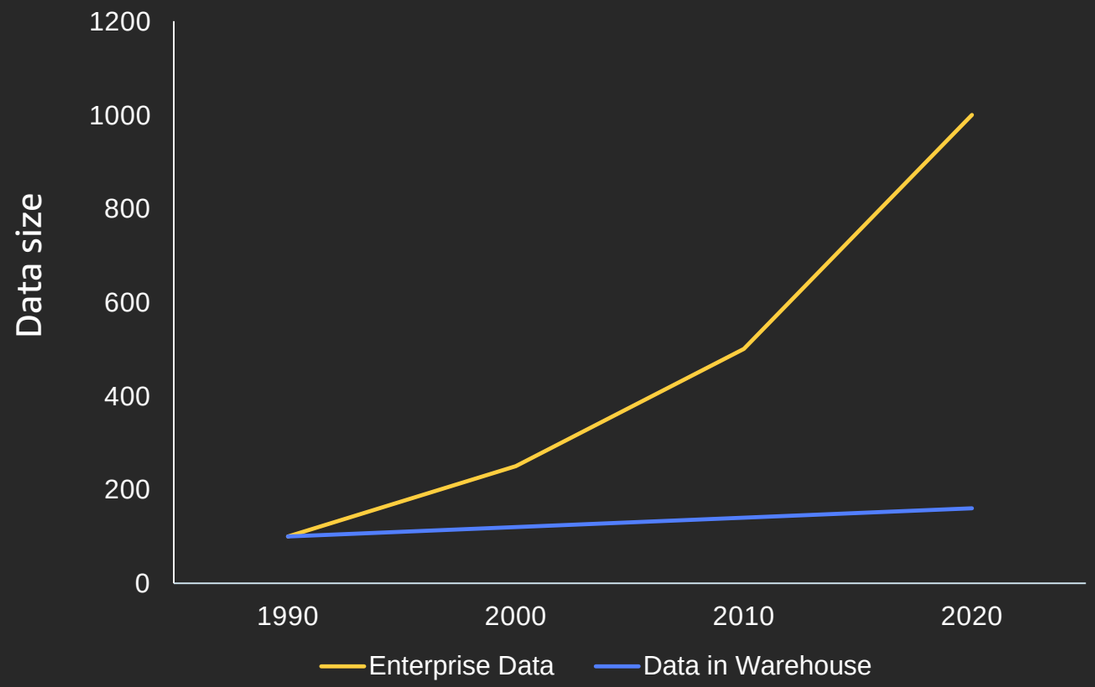
- NEW! Data lake export in Parquet
- NEW! Spectrum Request Accelerator
- NEW! Amazon Lake Formation integration

RA3 for Amazon Redshift Unpacked

aws

Exponential data growth causing management, performance and scaling challenges

What about data growth?



Compute and storage needs can vary independently

Compute scaling:

- To handle increased query load during peak times
- To ingest data fast and make available for querying in near real time

Storage scaling:

- To store data for querying or auditing
- To pay separately for data storage at cheaper rates, independent of compute usage

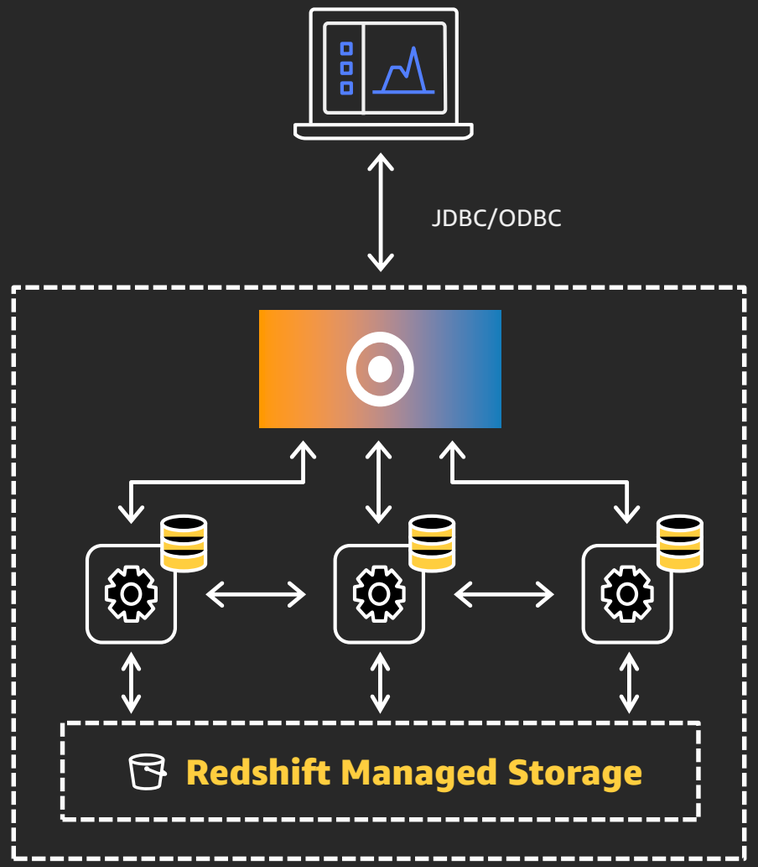
New 3rd generation Redshift compute instance: RA3

Managed storage | High-speed cache | High-bandwidth networking

Scale data warehouse only based on steady state compute needs

Pay separately for storage and compute

Automatic, no changes to any workflows, no need to manage storage



RA3: Unmatched performance at unbeatable price

RA3.16xl

On demand price—$13.04/node/hr

For storage pay as you use at $0.024/GB-Month

Can scale to tens of PB of data (8PB compressed)

Coming soon
RA3.4xl

2x Performance and 2x storage capacity compared to DS2.8XL for the same price

Up to 3x price-performance compared to any other Cloud DW

The minimum size RA3.16XL cluster scales up-to 128TB (compressed), size only for your compute need

DEMONSTRATION – BUILDING A NEW AMAZON REDSHIFT DATA WAREHOUSE WITH RA3

AWS re:Invent

aws

RA3: Migration considerations

Migrate using restore from snapshot

- Get a new RA3 cluster in minutes
- Validate the new RA3 cluster and delete the old cluster
- Use modify cluster to rename the RA3 cluster to old cluster's name
- Reduces the flexibility of Elastic Resize

Another option is classic resize

- Classic resize copies data from old to new cluster and renames the cluster upon completion (Classic resize is slower than restore)
- Retains full flexibility of Elastic Resize

DEMONSTRATION – MIGRATING TO RA3

Customer experience with RA3

Western Digital

"Our data is nearly ***doubling*** every year and we run ***6 Redshift clusters with total 78 nodes and 631+ TB compressed data*** stored to get insights that our business analysts and leadership depend on. The new Redshift ***RA3 instances*** offer us the ability to process our growing data more cost-effectively while we ***doubling our storage capacity compared to our previous Redshift cluster***"

Fayaz Syed, Sr. Manager, Big Data Platform

"We load ***billions of events per day*** into Amazon Redshift and have ***hundreds of terabytes of data that is expected to double every year***. While we store and process all our data, most of the analysis only ***uses a subset of the data***. The new Redshift RA3 instance with managed storage ***delivered 2x performance*** for most of our queries compared to our previous DS2 instance based Redshift cluster."

Jonathan Burket, Senior Software Engineer

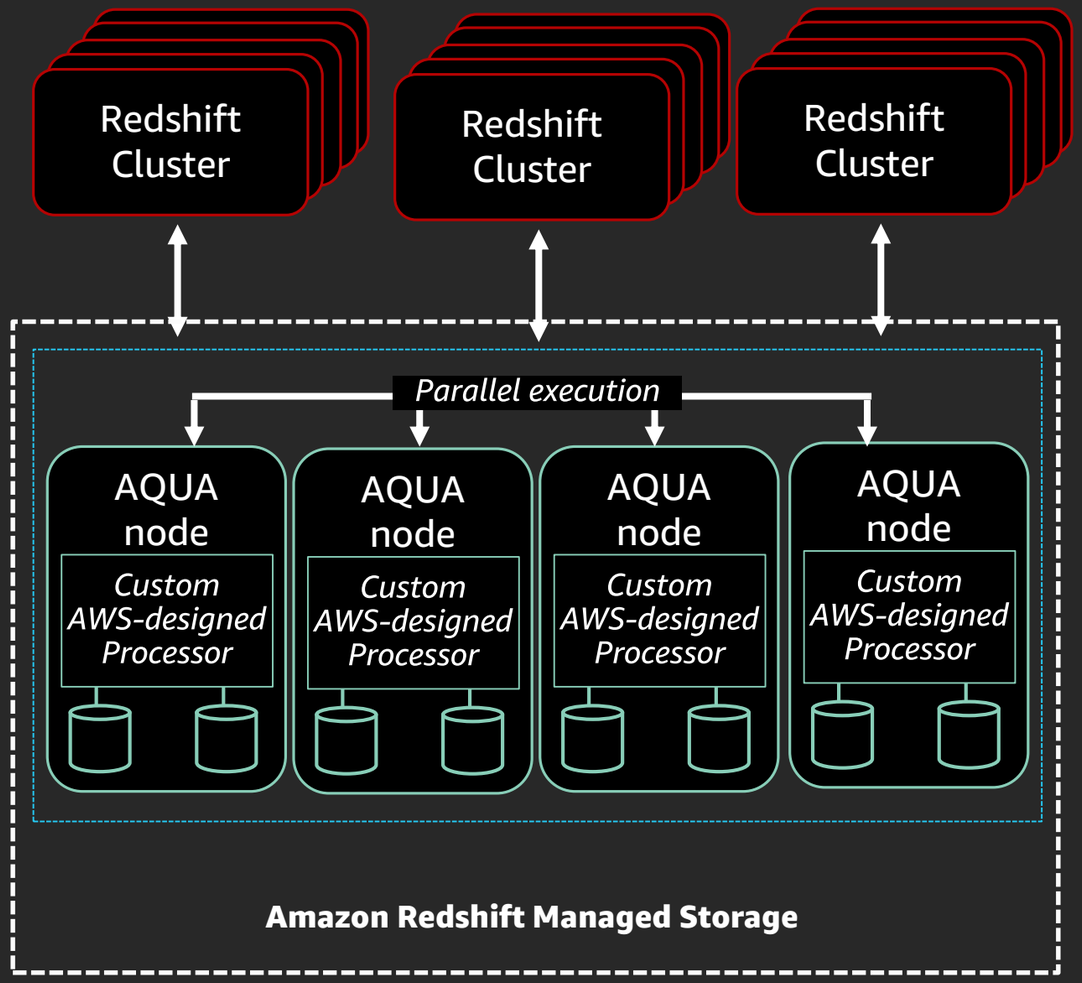
"We are thrilled with the new RA3 instance type. We have observed a ***1.9x performance improvement over DS2*** and ***1.5x performance improvement over DC2*** in our workload while ***keeping the same costs and providing scalable managed storage***. This allows us to keep pace with explosive data growth and have the necessary fuel to train our machine learning systems."

Stephen Moy, Software Engineer

AQUA for Amazon Redshift

aws

AQUA (Advanced Query Accelerator) for Amazon Redshift



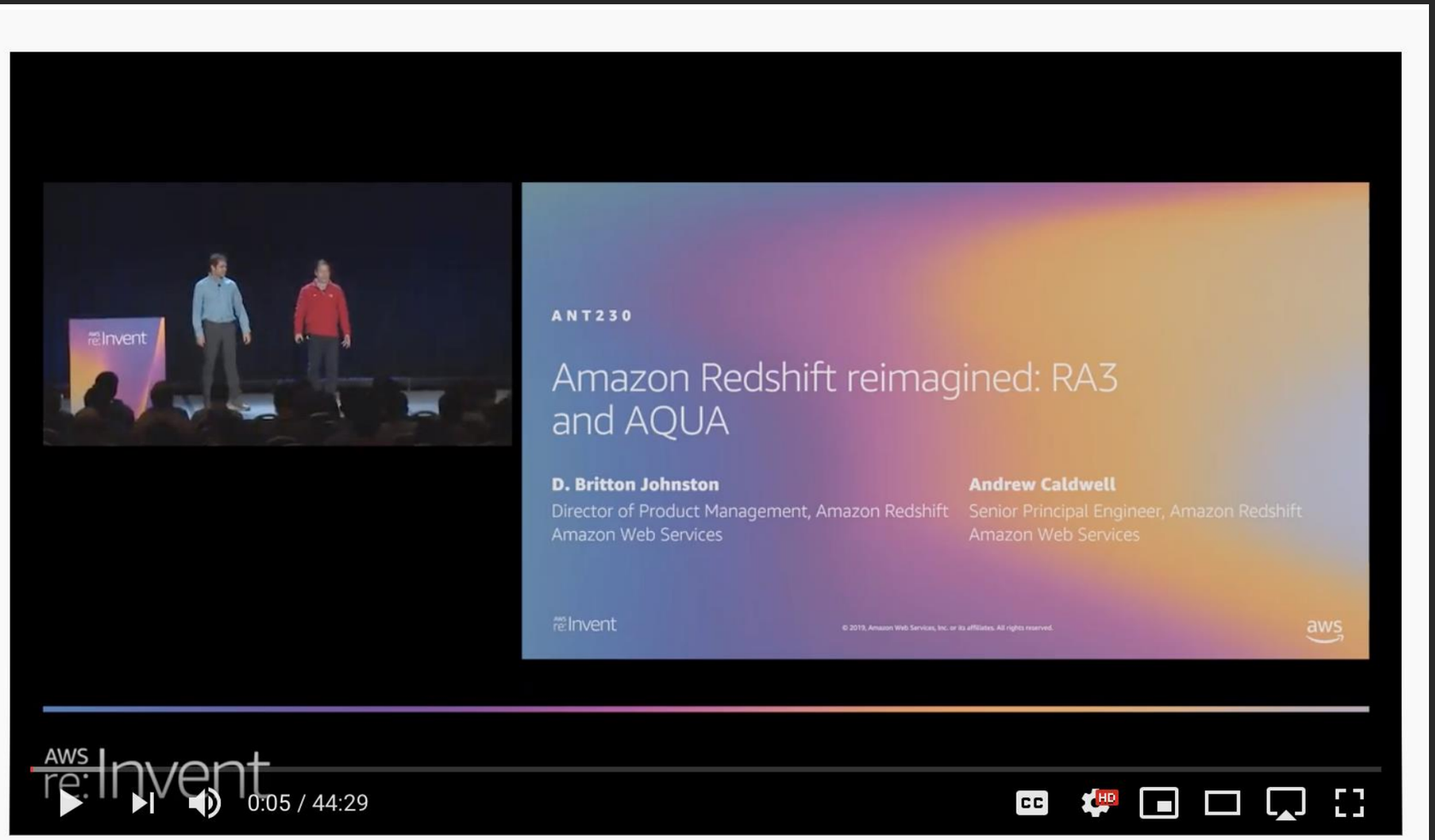
New distributed & hardware-accelerated processing layer

With AQUA, Amazon Redshift is up to **10x 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 in Preview with RA3. No code changes required

More on RA3 and AQUA



AWS re:Invent 2019: [NEW LAUNCH!] Amazon Redshift reimagined: RA3 and AQUA (ANT230)

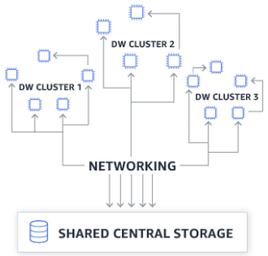
Sign up for Preview

aws

Advanced Query Accelerator (AQUA) for Amazon Redshift is now Available for Preview

AQUA is a new distributed and hardware-accelerated cache that enables Redshift to run up to 10x faster than any other cloud data warehouse.

Existing data warehousing architectures with centralized storage require data be moved to compute clusters for processing. As data warehouses continue to grow over the next few years, the network bandwidth needed to move all this data becomes a bottleneck on query performance.



If you look at hardware trends since 2012, SSD storage throughput has increased 12x while the ability for CPUs to process data in memory has only scaled 2x. This means that even if you removed the network from being the bottleneck, the ability for CPUs to process all this data is limited.

AQUA takes a new approach to cloud data warehousing. AQUA brings the compute to storage by doing a substantial share of data processing in-place on the innovative cache. In addition, it uses AWS-designed processors and a scale-out architecture to accelerate data processing beyond anything traditional CPUs can do today.

Please submit the information below to request an invitation to the preview. We will contact you with instructions if you are approved.

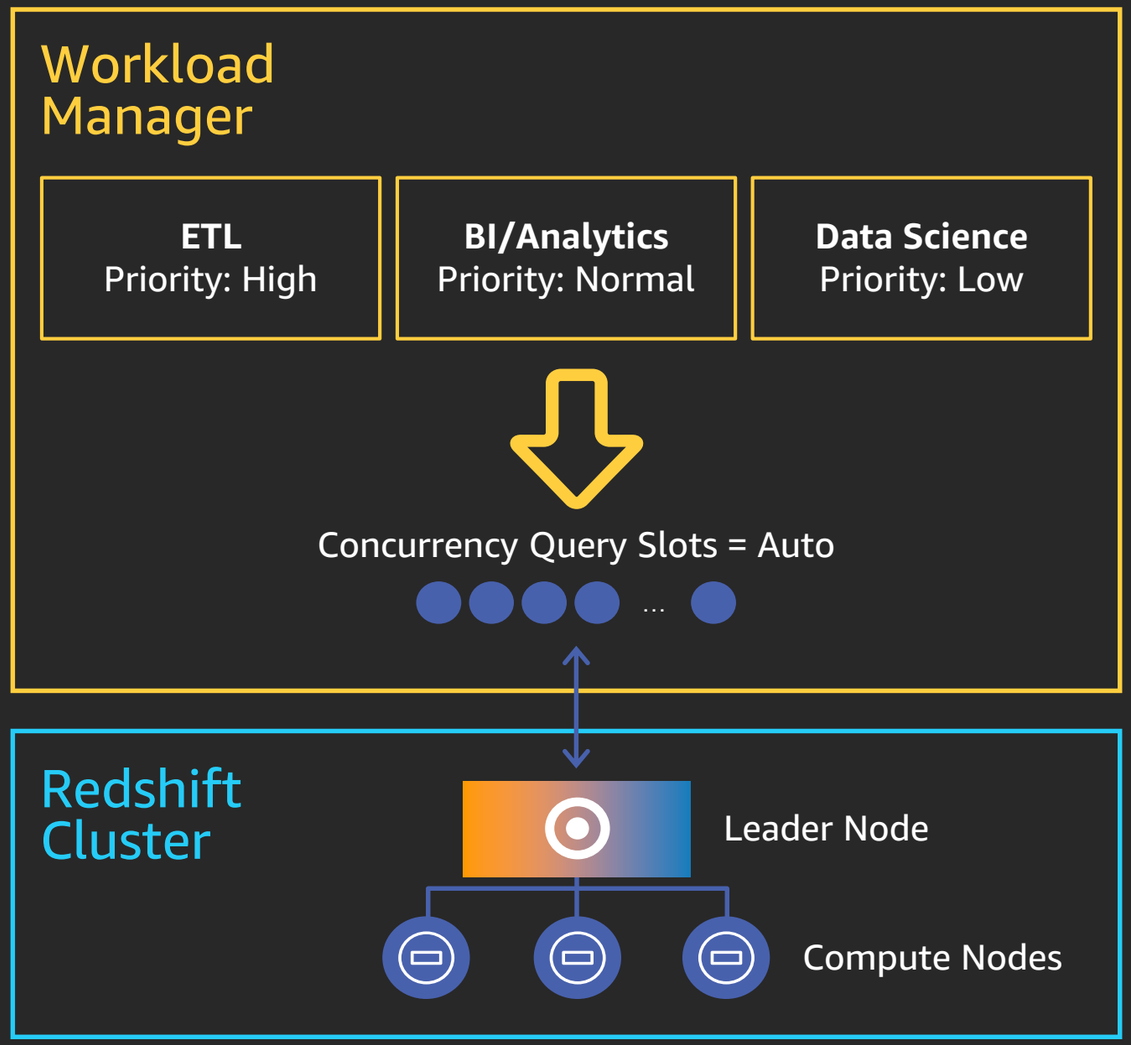
* Business Email Address:
* First Name:
* Last Name:
* Phone Number:
* Company Name:
* Country / Region: Select...
* Postal Code:
* Industry: Select...
* Job Role: Select...
* Job Title:
* Level of AWS Usage: Select...
* Use Case: Select...
* What is your AWS Region? Select...
* AWS Account ID:
* I am completing this form in connection with my:
  - Business interests
  - Personal interests

By completing this form, I agree that I'd like to receive information from Amazon

Ease of Use

aws

Efficiency with automatic workload management and query priorities



- Manages concurrency dynamically
- Supports efficient sharing of cluster resources
- Maximizes throughput and performance
- Leverages machine learning to classify queries based on resources needs
- Ability to influence workload performance based on business priorities
- Ensures low priority queries make progress

Amazon Redshift turbo charges query performance with machine learning based automatic optimizations

Automates table maintenance

Optimizes for peak performance as data and workloads scale

Leverages machine learning

Prescriptive recommendations with ability to apply changes dynamically

Automatic Analyze

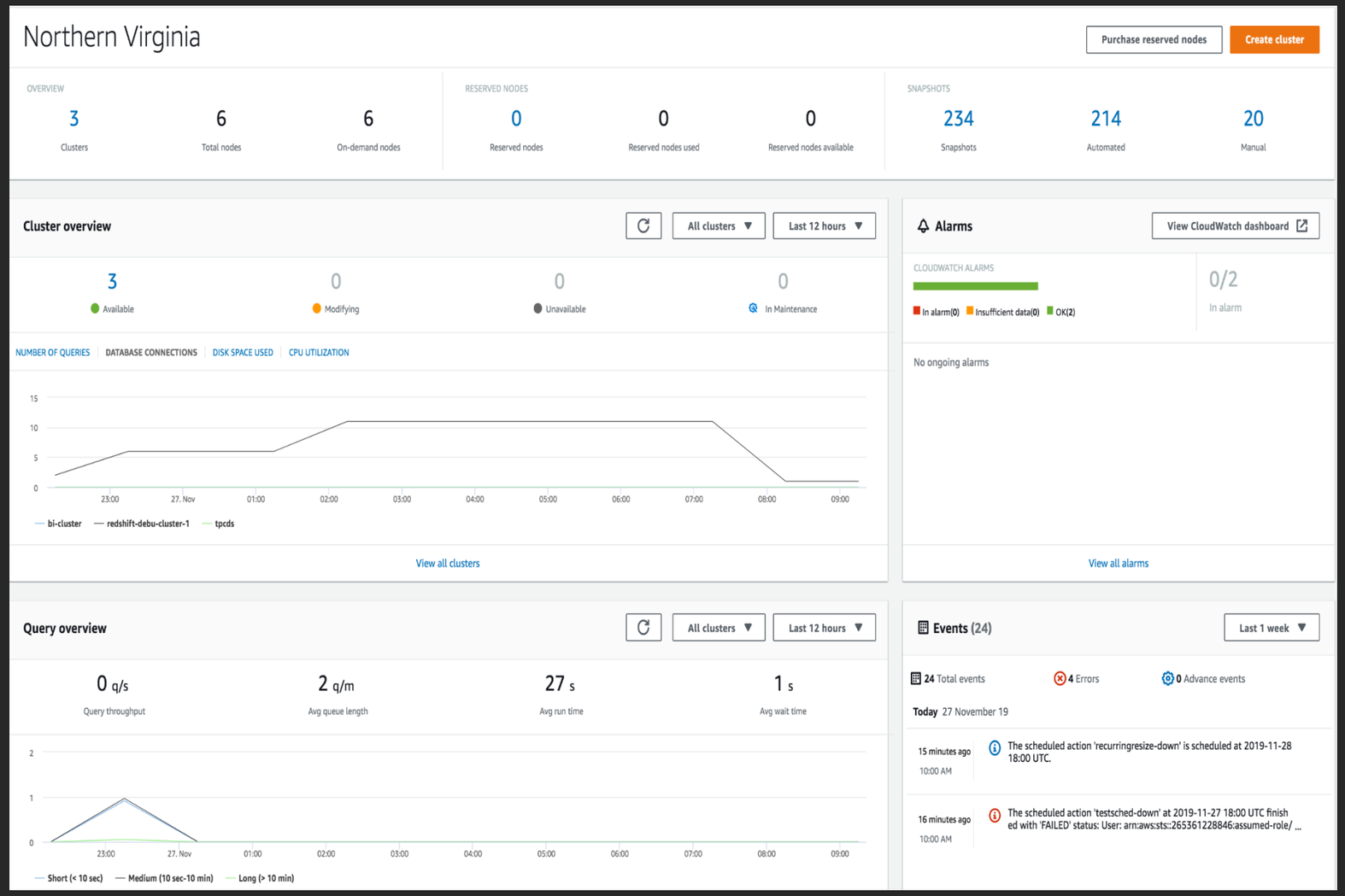
Automatic Table Distribution Style

Distribution/Sort key advisors

Automatic Vacuum Delete

Automatic Table Sort

New management console: intuitive, fast, and feature-rich



- Manage multiple data warehouses with command center-like dashboard
- Reduced time to diagnose when unexpected happens
- Share Query Editor with non-admin users
- Visually analyze your query results right in the query editor

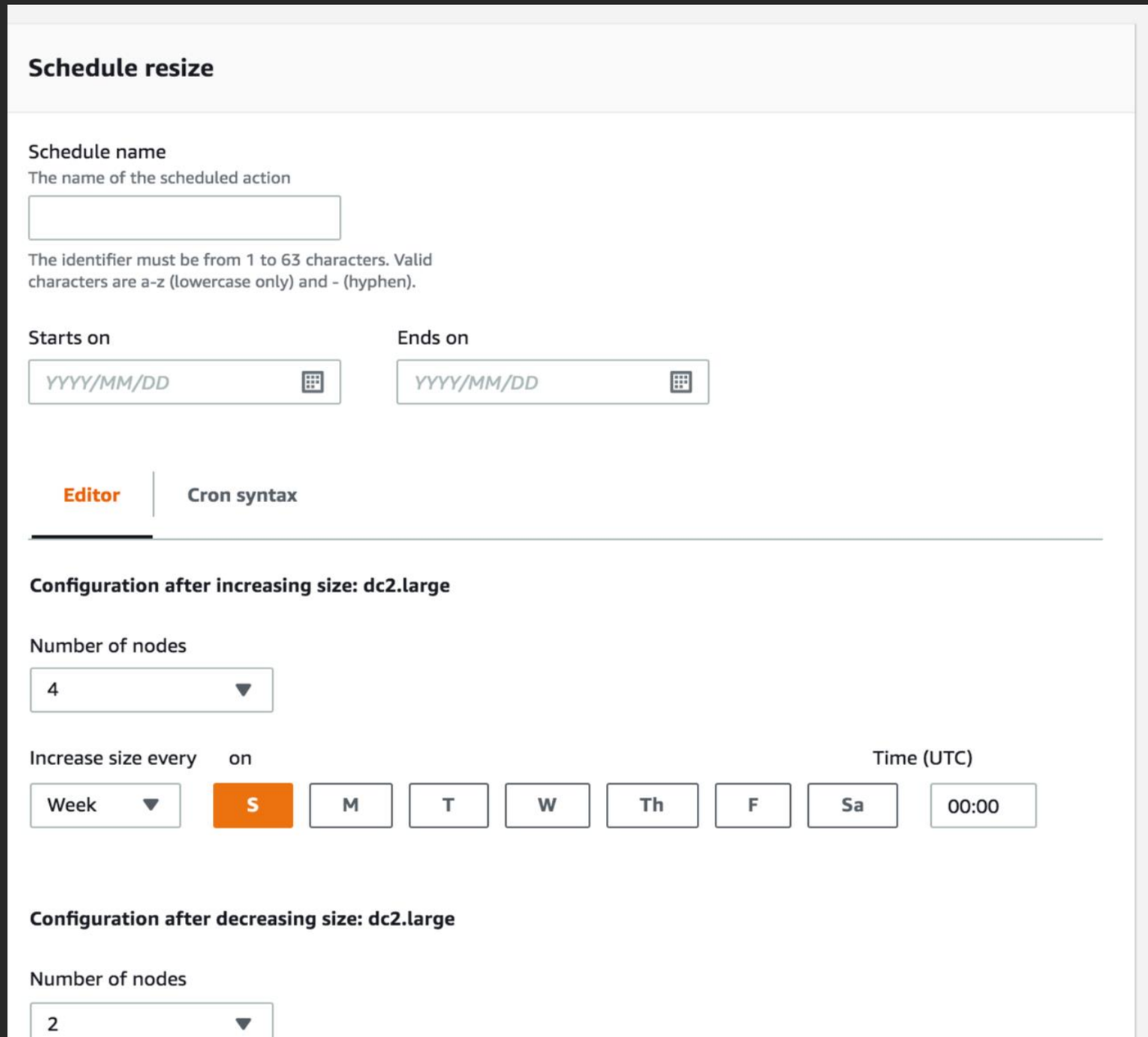
Scheduler for elastic resize: optimize cost easily

Add or remove nodes in minutes using elastic resize

Scheduled cluster resize using management console or API

Removes dependency on external function and failures

Optimize cost and plan ahead for peak demand

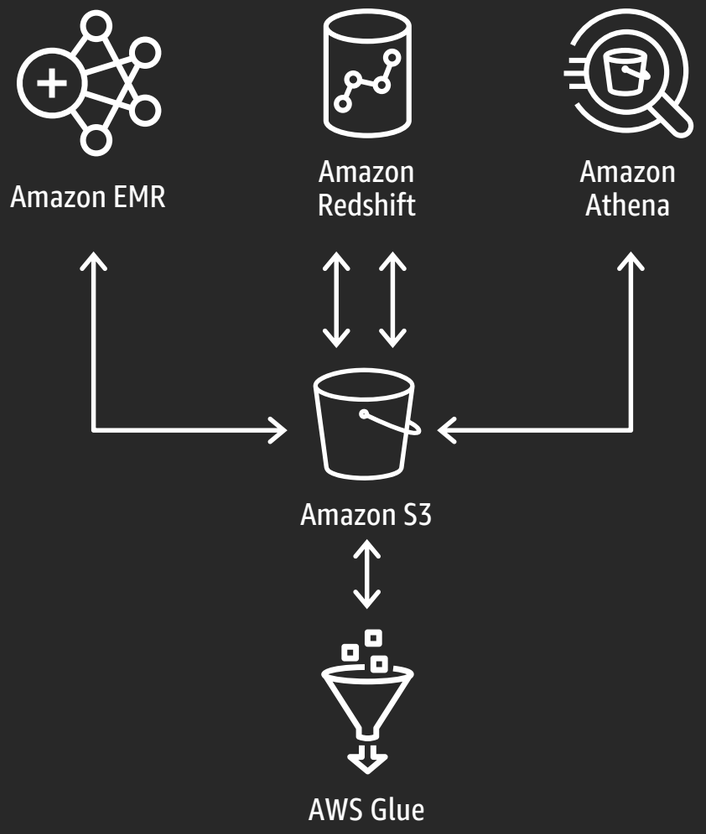


DEMONSTRATION – Scheduled Resize

Data lake Integration

aws

Data lake export: share data in Parquet format



Parquet is efficient open columnar storage format for analytics

Analyze your data with Redshift Spectrum and other AWS services such as Amazon Athena and Amazon EMR

```
UNLOAD
('select * from lineitem')
TO
's3://mybucket/unload/lineitem/'
FORMAT as PARQUET
PARTITION BY (l_shipdate);
```

DEMONSTRATION – Data Lake

Federated Query (Preview)

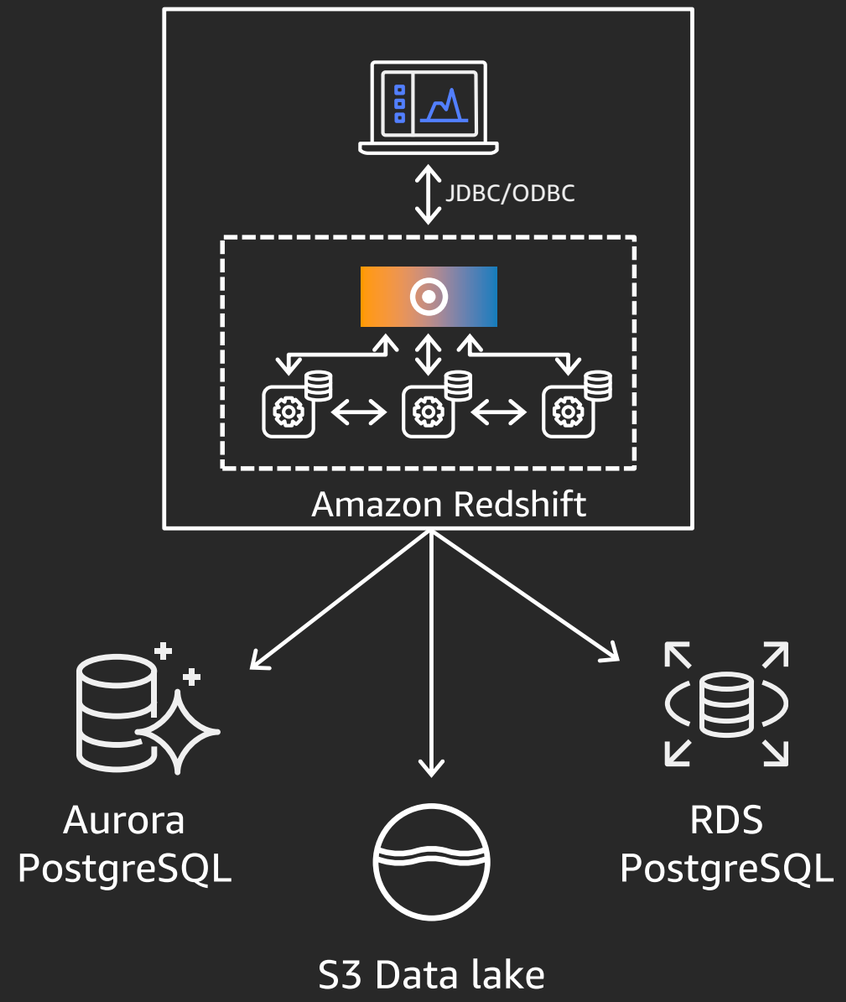
Queries on RDS and Aurora PostgreSQL databases

Analytics on live data without data movement

Unified analytics across data warehouse , data lake & operational databases

Flexible and easy way to ingest data

Performant and secure access to data



Amazon Redshift federated query: most up to date data, no ETL required

```
-- Aurora Postgres has Hot Data (2019)
-- Redshift has Recent Data (2016-2018)
-- S3 has Archival Data (1992-1998)
-- Declare a view across all backends

CREATE VIEW lineitem_all AS
SELECT * FROM s3.lineitem_1t_part
UNION ALL
SELECT * FROM public.lineitem
UNION ALL
SELECT * FROM apg.lineitem
WITH NO SCHEMA BINDING
```

```
-- Find #sales with 1 item in Jan of each year
-- Predicates are being pushed down
-- Partition pruning on the S3 data
-- Aggregates are being pushed down
-- Very intuitive syntax

SELECT EXTRACT(year FROM l_shipdate) AS year,
       EXTRACT(month FROM l_shipdate) AS month,
       COUNT(*) AS orders
FROM lineitem_all
WHERE extract(month FROM l_shipdate) = 1
AND l_quantity < 2
GROUP BY 1,2
ORDER BY 1,2;
```

Scalability and Performance

aws

Amazon Redshift: now 2x faster out of the box

Maintaining performance and scalability leadership through continuous telemetry and benchmarking

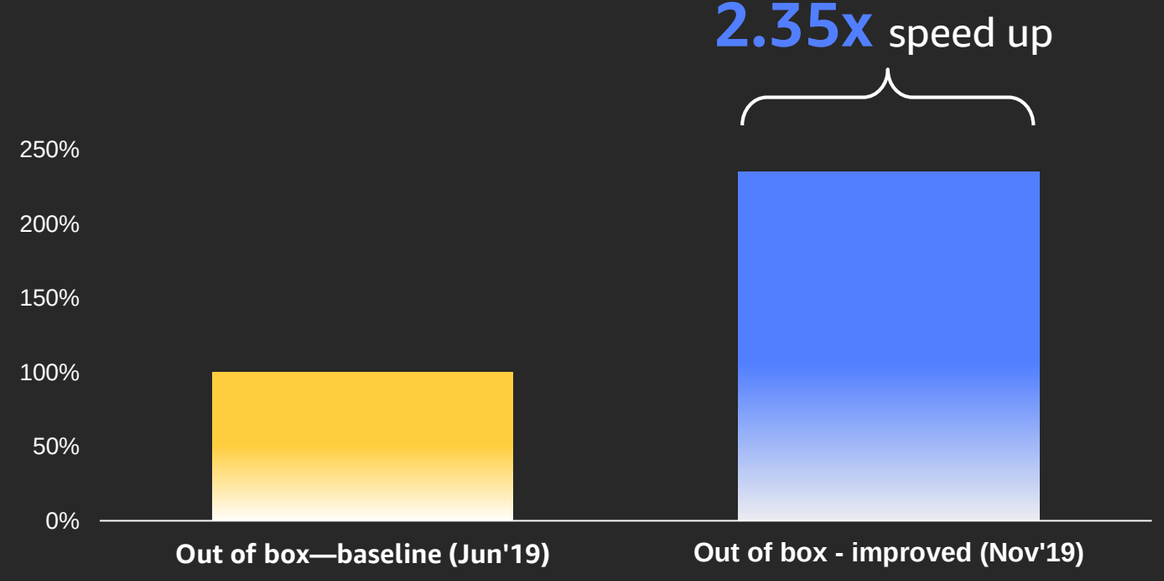
Performance innovations:

- AZ64 encoding
- Bloom filters for collocated & broadcast JOIN queries
- Enhanced planner for modern hardware CPUs and networking
- HLL (HyperLogLog) statistics
- Cache-optimized aggregation and join processing

Cloud DW 30TB benchmark*

Normalized queries/hour (QPH)

(Higher is better)



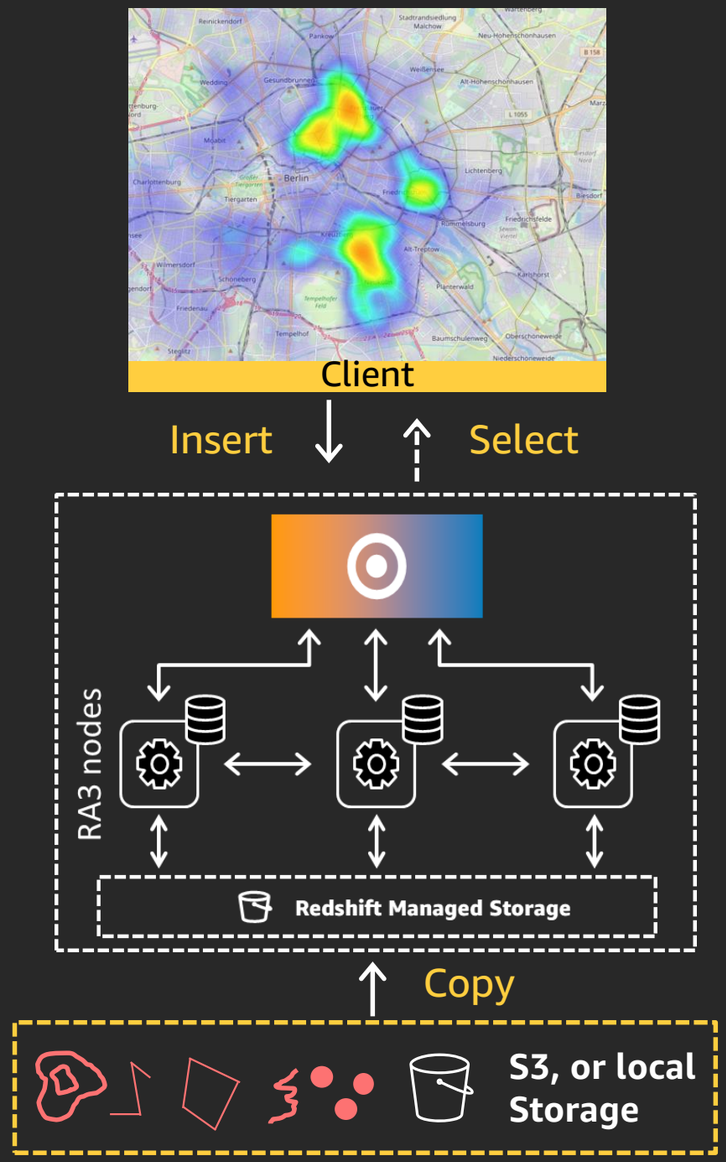
* Cloud DW benchmark is based on TPC-DS (v2.10) with no query modifications done

New data type Geometry: Ingest, store, and analyze spatial data

Seamlessly integrate spatial and business data

Geometry data types supports shapes such as MultiPoint, MultiLinestring, MultiPolygon, and GeometryCollection

40+ SQL spatial functions to construct geometric shapes, import, export, access and process the spatial data



DEMONSTRATION – Spatial Data processing

Introducing Materialized Views:
Compute once, query many times

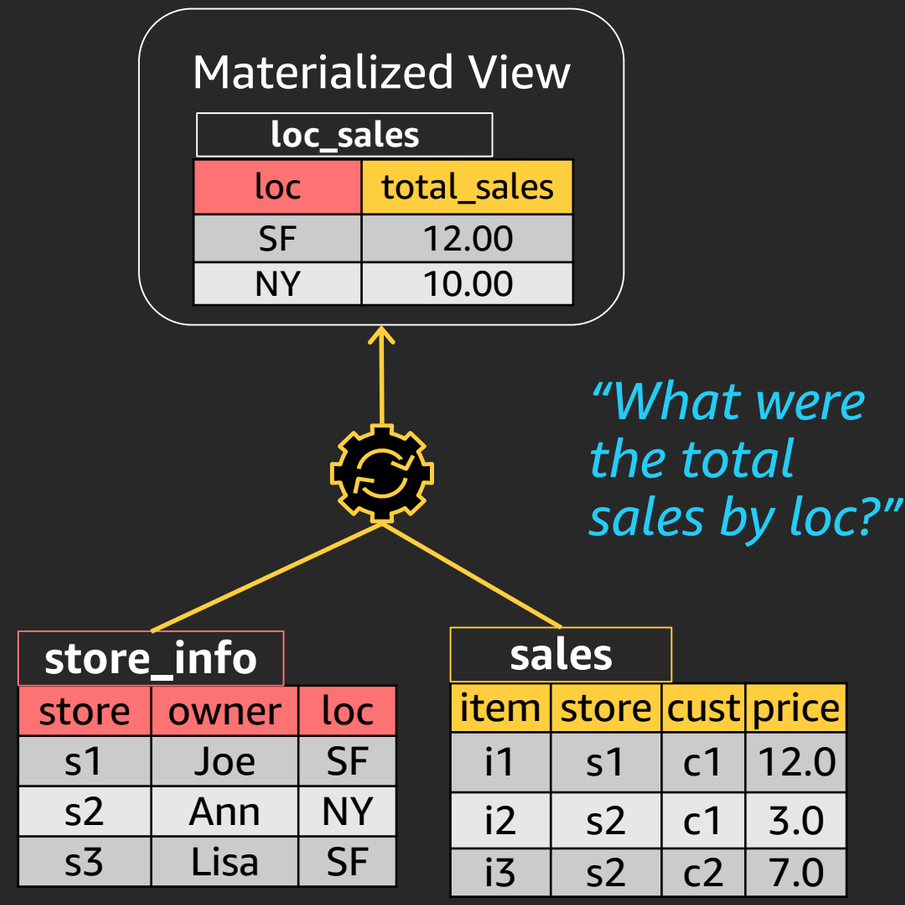
Speed-up queries by orders of magnitude

- Joins, filters, aggregations, and projections

Simplify and accelerate ETL/BI pipelines

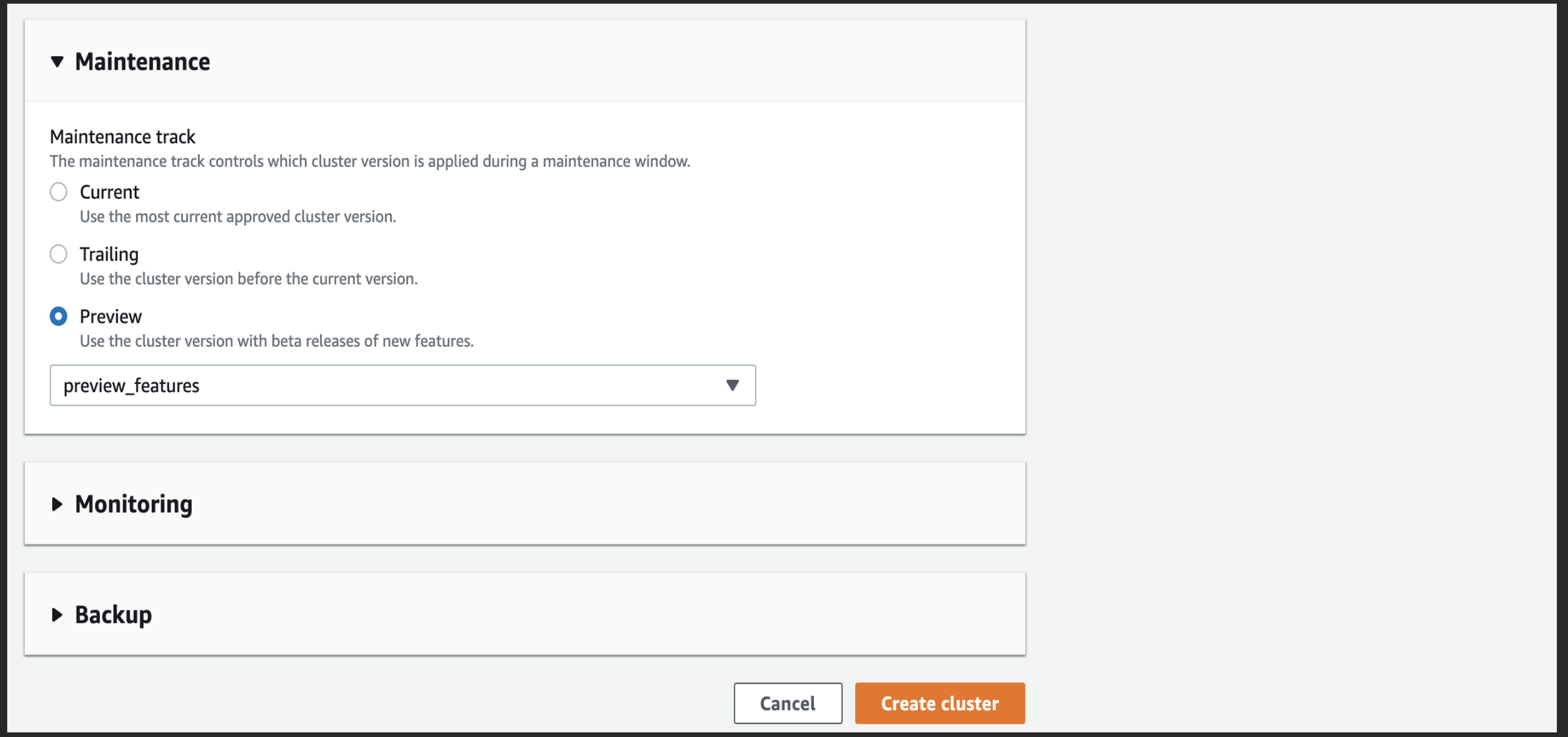
- Incremental refresh
- User triggered maintenance

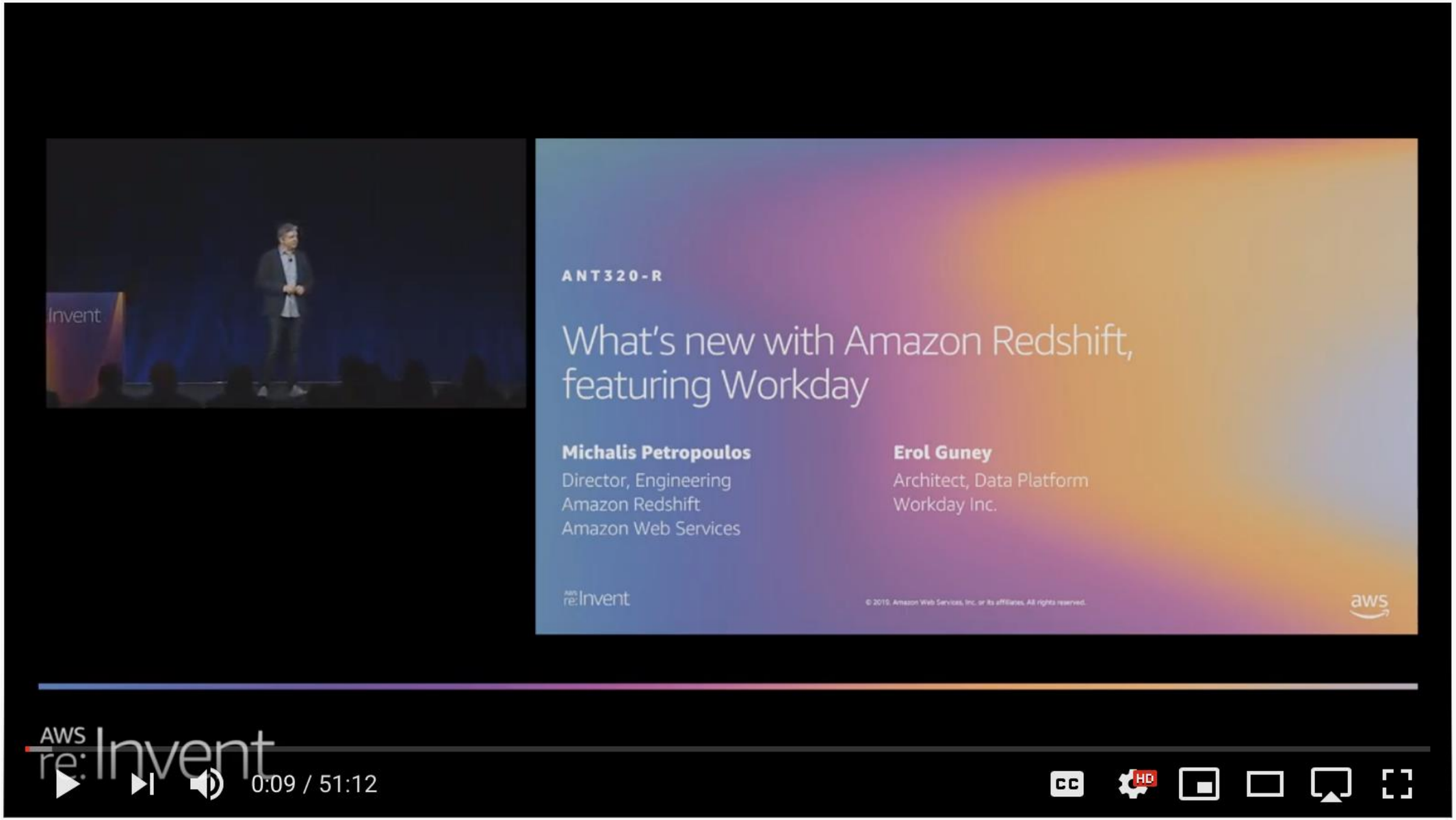
Easier and faster migration to Redshift



Sign up for preview now!

Creating Your Cluster in Preview Track





AWS re:Invent 2019: [REPEAT] What's new with Amazon Redshift, featuring Workday (ANT320-R)

Summary

- Amazon Redshift is the most popular data warehouse with Tens of thousands of customers
- Amazon Redshift has innovated 200+ features in last 18 months
- Amazon Redshift new instance RA3 provides ability to scale compute and storage independently
- AQUA enables Amazon Redshift to run up to 10x faster than any other cloud data warehouse
- Amazon Redshift has simplified several admin operations
- Amazon Redshift is very integrated with your data lake
- Out of box performance for Amazon Redshift is 2x faster