AWS re:Invent

ANT350-R

What's New with Amazon Redshift ft. Dow Jones

Colleen Camuccio
VP, Program Management
Dow Jones

Vidhya Srinivasan
General Manager, Amazon Redshift
AWS

AWS re:Invent

aws

DOW JONES

AWS re:Invent

aws

Data challenges we face with customer data

Multiple versions of the truth

“ I see so many different versions of the same metric

Limited visibility into performance

“ I have no idea how my sales team performs vs. peer groups

Wasted time spent hunting for data

“ It takes four days to analyze my customers’ usage patterns

Missing insights impairs decision-making

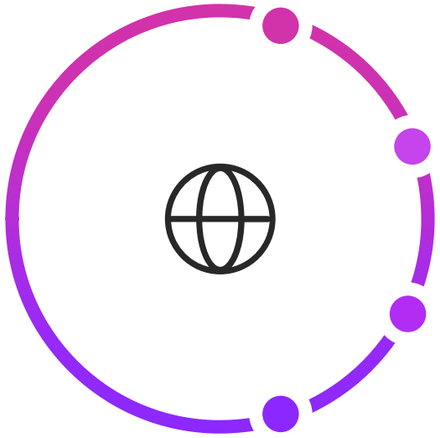
“ I have difficulty determining the value of the customers we are acquiring

Inability to segment customers

“ I still can’t tell how many business travelers subscribe

We set out to build a world-class data platform

Key objectives



Enable
new and existing revenue streams

Improve
customer experience through targeted communications

Reduce
operational cost

Promote
more informed decision-making

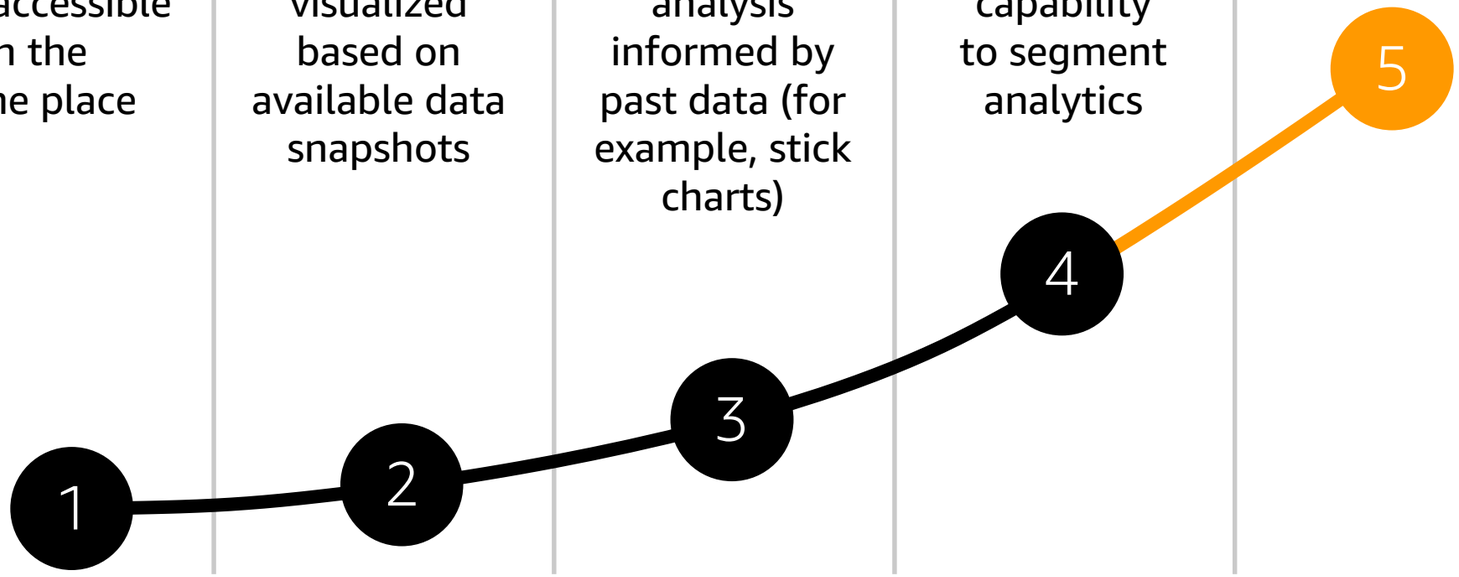
We ramped up an AWS Landing Zone

Created
an environmental council to make key architectural decisions

Partnered
closely with AWS through the process

Platform evolution

Stage	Visualize	Trend	Cluster	Predict
Single location	Interactive & visual	Historical analysis	Integrated insights	Advanced modeling & prediction
All customer data accessible in the same place	Insights visualized based on available data snapshots	Complex analysis informed by past data (for example, stick charts)	Advanced capability to segment analytics	

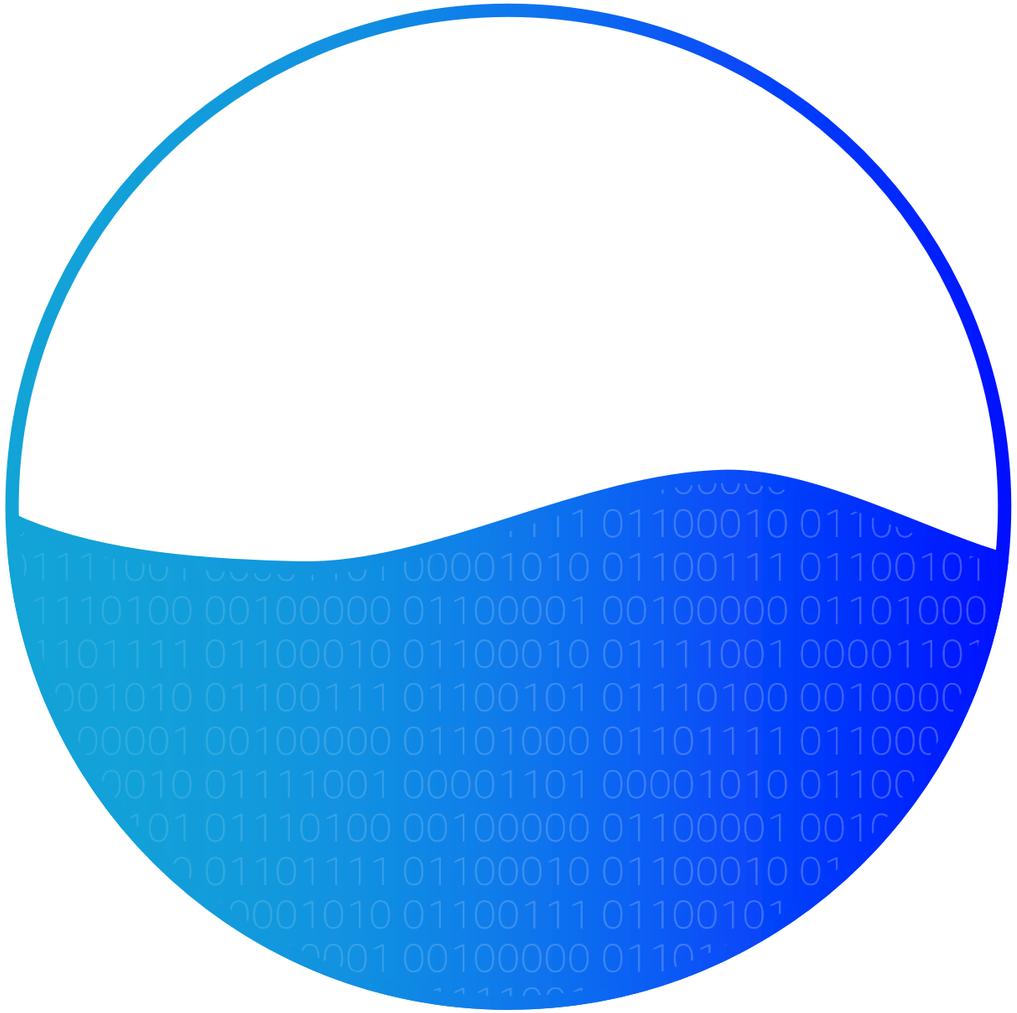


End goal

Predictive analytics & machine learning

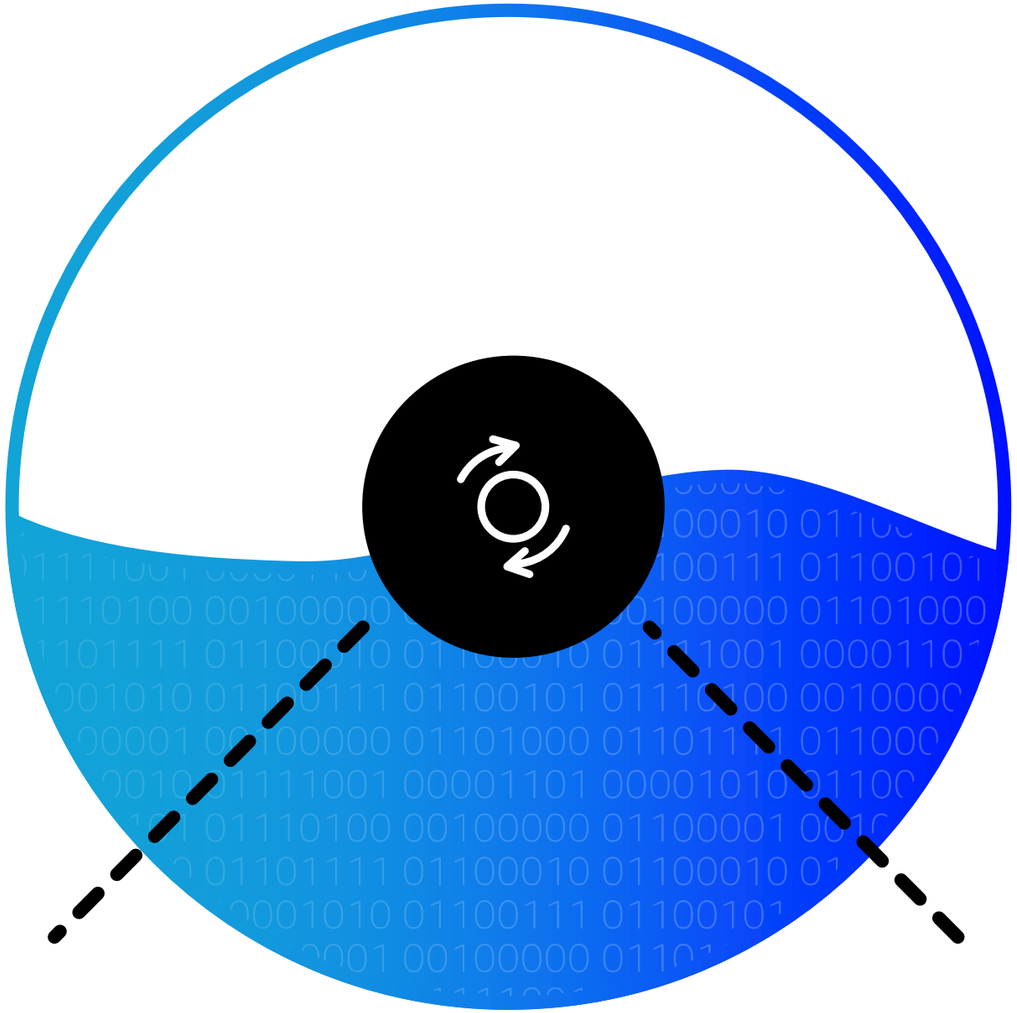
The pace of evolution for each customer lifecycle step is based on

- Knowledge & Insight
- Risk & Compliance
- News & Information



A **data lake** is a storage repository that holds a vast amount of raw data in its native format

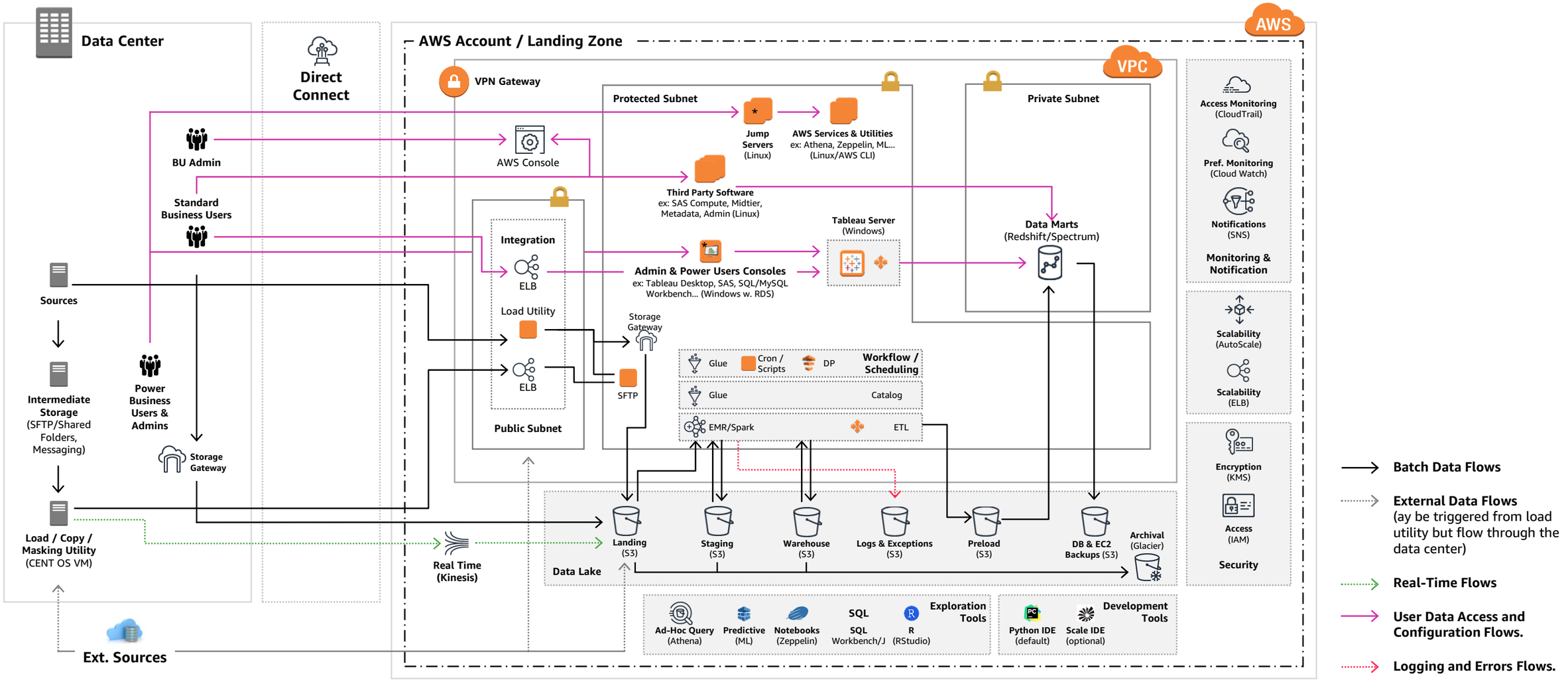
We ensure the data is loaded correctly each day. We aim to collect every relevant customer event that occurs for any of our members.



To make sense of this customer data, we will construct easy-to-understand **data marts**

that will normalize this data to a state where it can be consumed by self-serve tools and ad hoc analysis.

Reference architecture



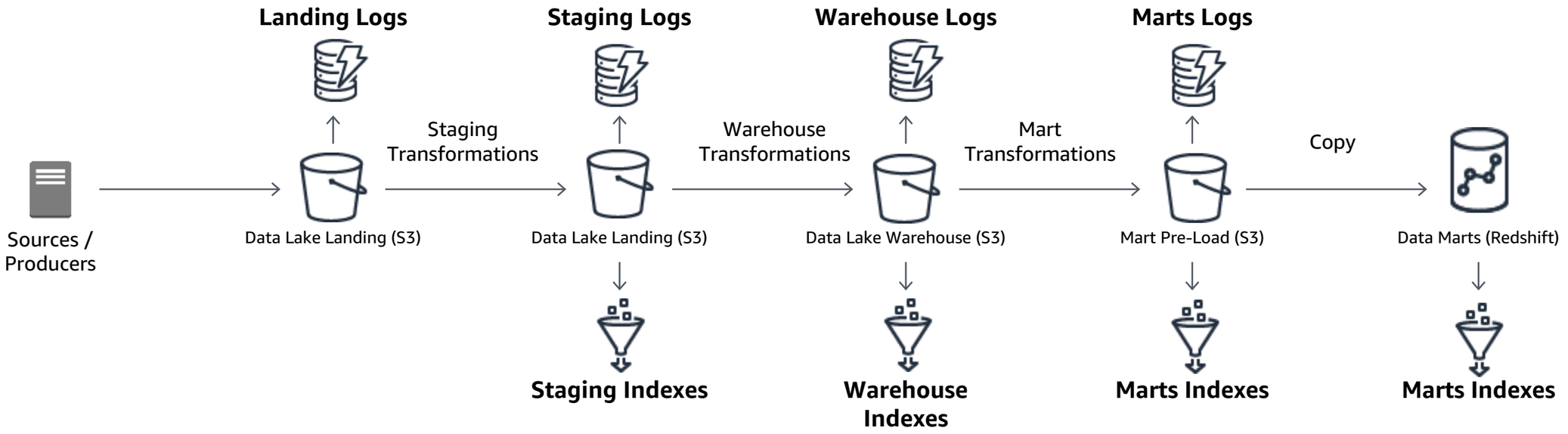
Creating our data platform

Use traditional data warehouse practices to process and store our data

Leverage the staging area where we prepare the source data, standardize it, catalog it and make it available as the data lake

Prepare intermediate tables in the warehouse layer and store them in aggregated data marts

Use the Copy command to push the data to Amazon Redshift



Amazon Athena Vs. Amazon Redshift

Access data in **Amazon Redshift**

Best performance

All data is stored in **Amazon S3** but created as external tables in Amazon Redshift

Redshift Spectrum to query

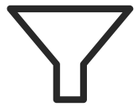
More cost-effective measure

Use the **Copy command** to copy data from **Amazon S3** to **Amazon Redshift** as a physical table

Case-by-case basis

When performance takes priority over cost

Analytics tools connected directly to Redshift create custom dashboards for our users



Harness the power of both the Data Lake and the Data Marts

Create a BI layer where we join multiple data sets to create custom insights for our business partners

Users can also connect directly to Redshift to query the tables and generate their own insights

Analytics tools connected directly to Redshift create custom dashboards for our users

Users can download the tool of their choice, as long as it contains a JDBC/ODBC connection to Redshift

Users are armed with a data dictionary and a set of guidelines that allow them to harness the power of Redshift

Best Practices: Querying the data

Select a specific set of columns

instead of Select *

Conduct a "Group By"

instead of a "Select Distinct"

Use "With Clauses"

instead of "Sub Queries" as it will create a temp table which can be re-used or referenced multiple times

Amazon Redshift is best at aggregated queries

use aggregate functions whenever possible

When investigating the data, use filters

to make the data set size smaller or limit the number of records to the first 1,000 records or so

Always use the data partitions

when querying the data

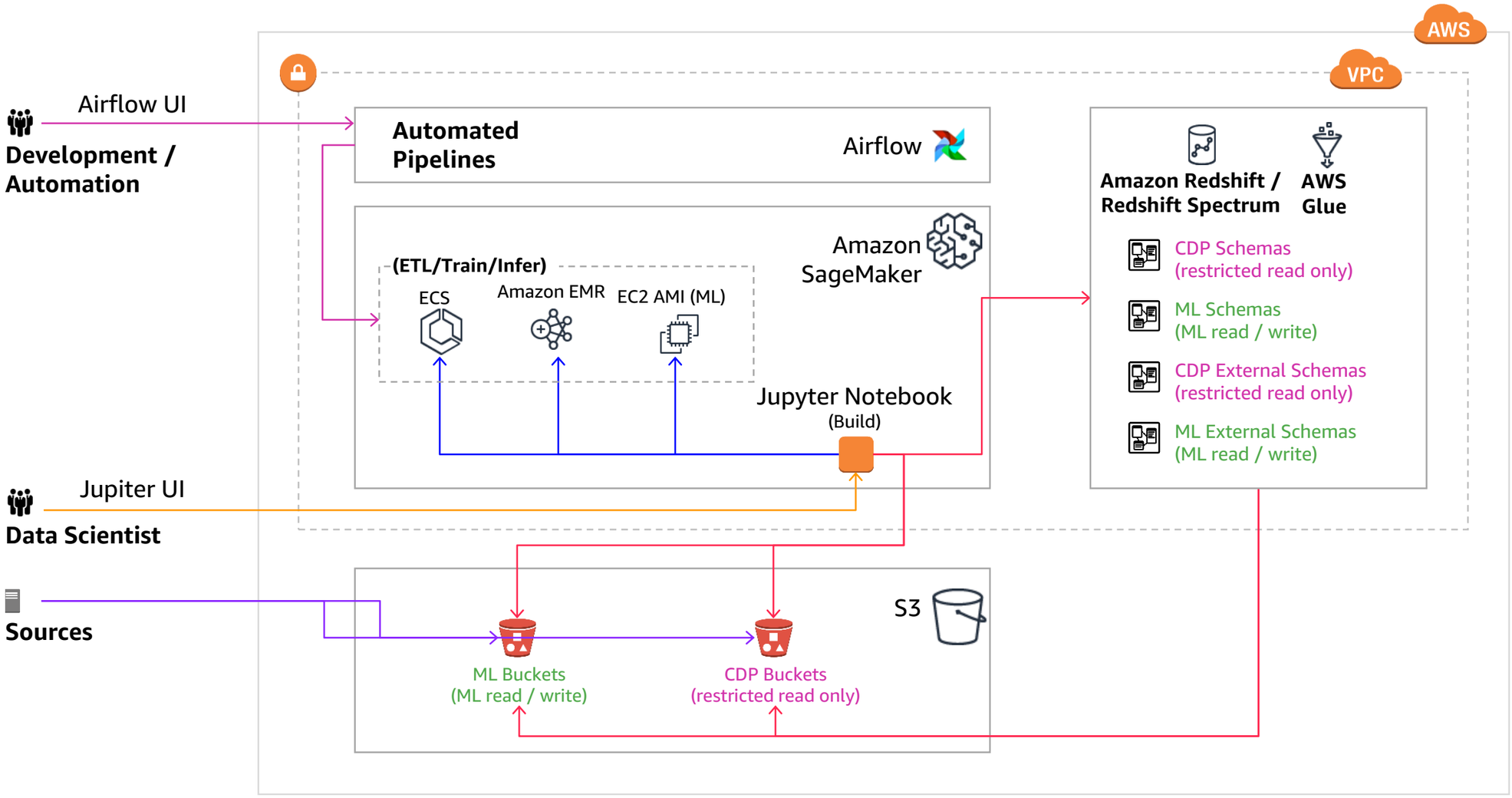
What is next: Our platform is entering into the final stages of evolution

Prepare and organize
the volume of data we collect on our customers and complete the customer data lake

Catalog and cleanse
the data so that is ready to be used in predictive models and is primed for machine learning

Use machine learning
to train a model to triage rejected records and re-process them to improve our match rates during the warehouse and data mart transformation steps

Reference machine learning architecture



Amazon SageMaker: “Analyst Sandbox” within Amazon S3

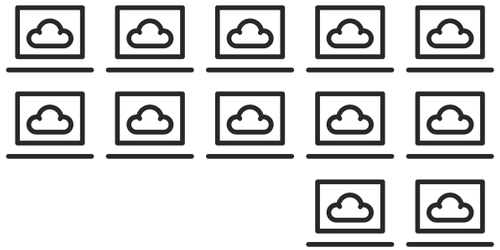
Query the
Data Lakes and
Data Marts
in **Amazon Redshift**

Create curated
data sets and
store them in **Amazon S3**
to be cleansed,
labeled and provided
as input into the data mart

Access pre-setup
Jupyter notebook
instances to
both query and
prepare data

Amazon SageMaker will be responsible for managing the AWS resources
behind the scenes while the analysts focus on the key modeling activities

Key Results

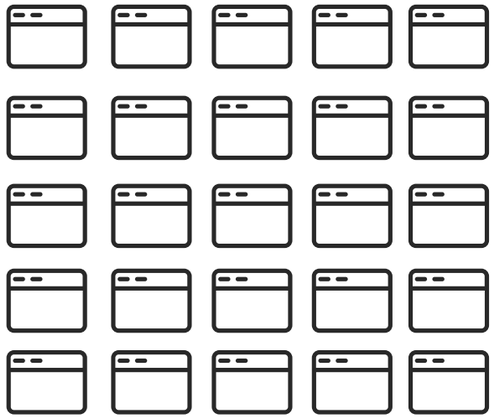


We have spun up

12

different dashboards

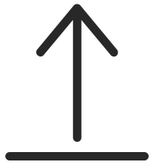
to support our stakeholders



These dashboards are powered by

30

different fact & dimension tables in Redshift



Across the whole of the warehouse we hold upwards of

118 TB of Data

The platform is currently accessed by

100+ users

DOW JONES

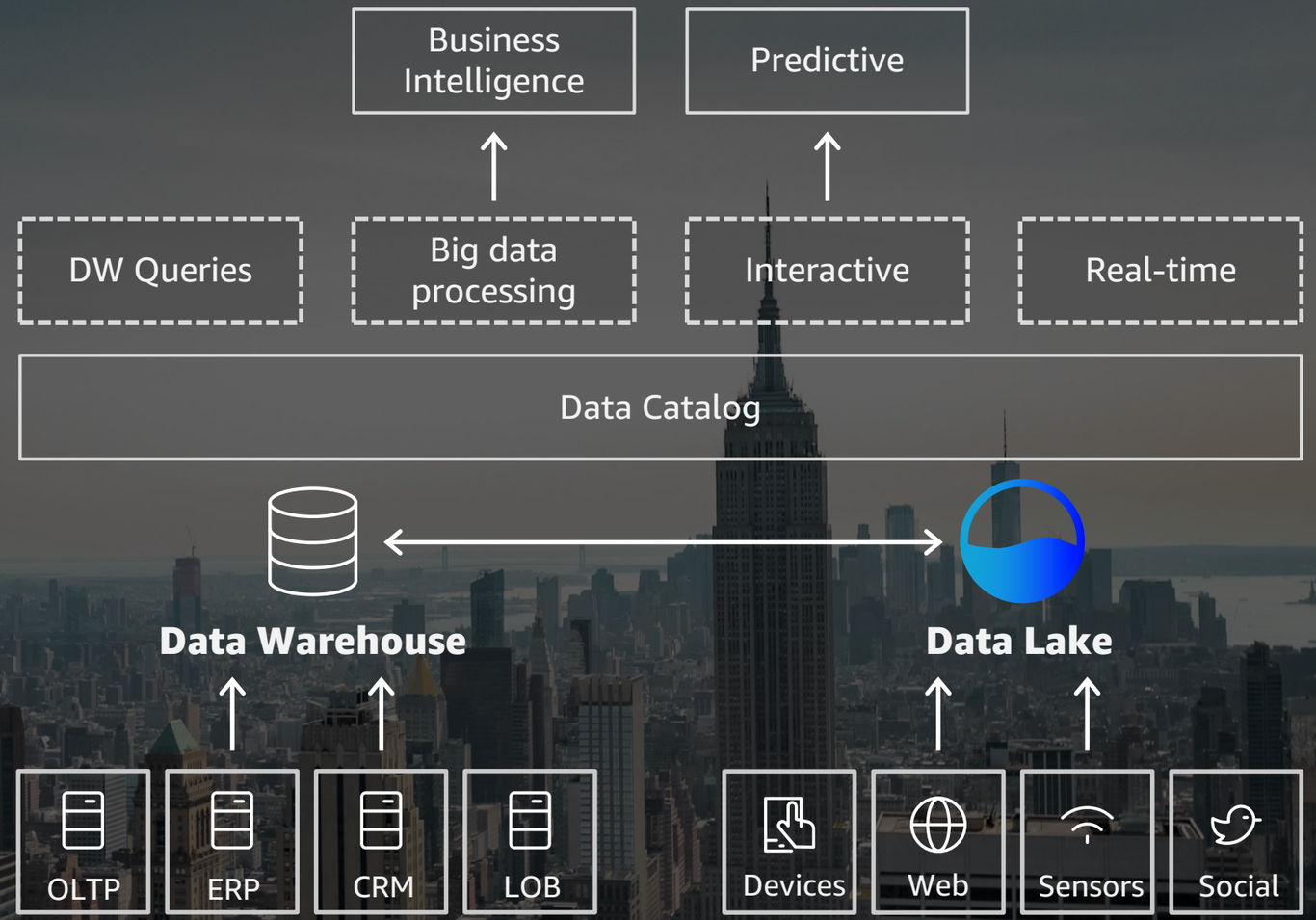
Thank you.

Bucket strategy matters

Glue and EMR/Spark are great tools for data transformation

Redshift provides a central location to view high-performance reports and run queries with data in S3

Scale on a per job or per query basis



Analytics

Storage

Data movement

Data warehousing

amazon REDSHIFT

Big data processing

amazon EMR

amazon S3

amazon DATA PIPELINE

aws

2012 Analytics Portfolio

AWS databases and analytics

Broad and deep portfolio, built for builders

Business Intelligence & Machine Learning


QuickSight


SageMaker


Comprehend


Rekognition


Lex


Transcribe


DeepLens

AWS Marketplace

250+ solutions

Databases


QLDB NEW
Ledger Database


ElastiCache
Redis, Memcached


Aurora
MySQL, PostgreSQL


RDS
MySQL, PostgreSQL, MariaDB, Oracle, SQL Server


Neptune
Graph


DynamoDB
Key value, Document


Timestream NEW
Time Series


RDS on VMWare NEW

Analytics


Redshift
Data warehousing


EMR
Hadoop + Spark


Elasticsearch service
Operational Analytics


Athena
Interactive analytics


Kinesis Analytics
Real-time

Blockchain


Managed Blockchain NEW


Blockchain Templates

730+ Database solutions

600+ Analytics solutions

25+ Blockchain solutions


S3/Glacier


Lake Formation NEW
Data Lakes

Data Lake


Glue
ETL & Data Catalog

20+ Data lake solutions

Data Movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Data Pipeline | Direct Connect

30+ solutions

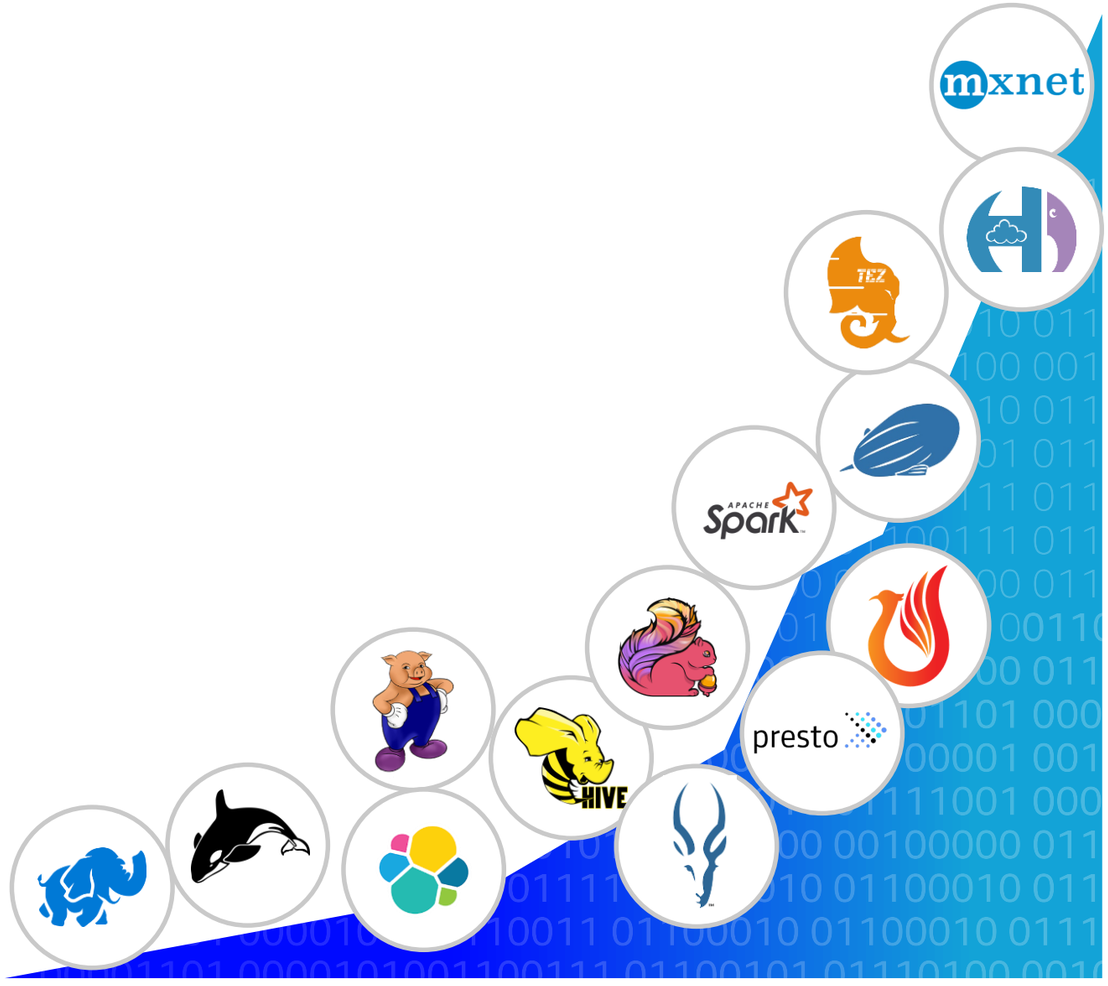
AWS re:Invent

aws

There is **more data** than people think

Data	Data platforms need to	
grows	live for	scale
>10x	**15**	**1,000x**
every 5 years	years	

* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.



There are **more ways** to analyze data than ever before

11 **8** **5** **4**

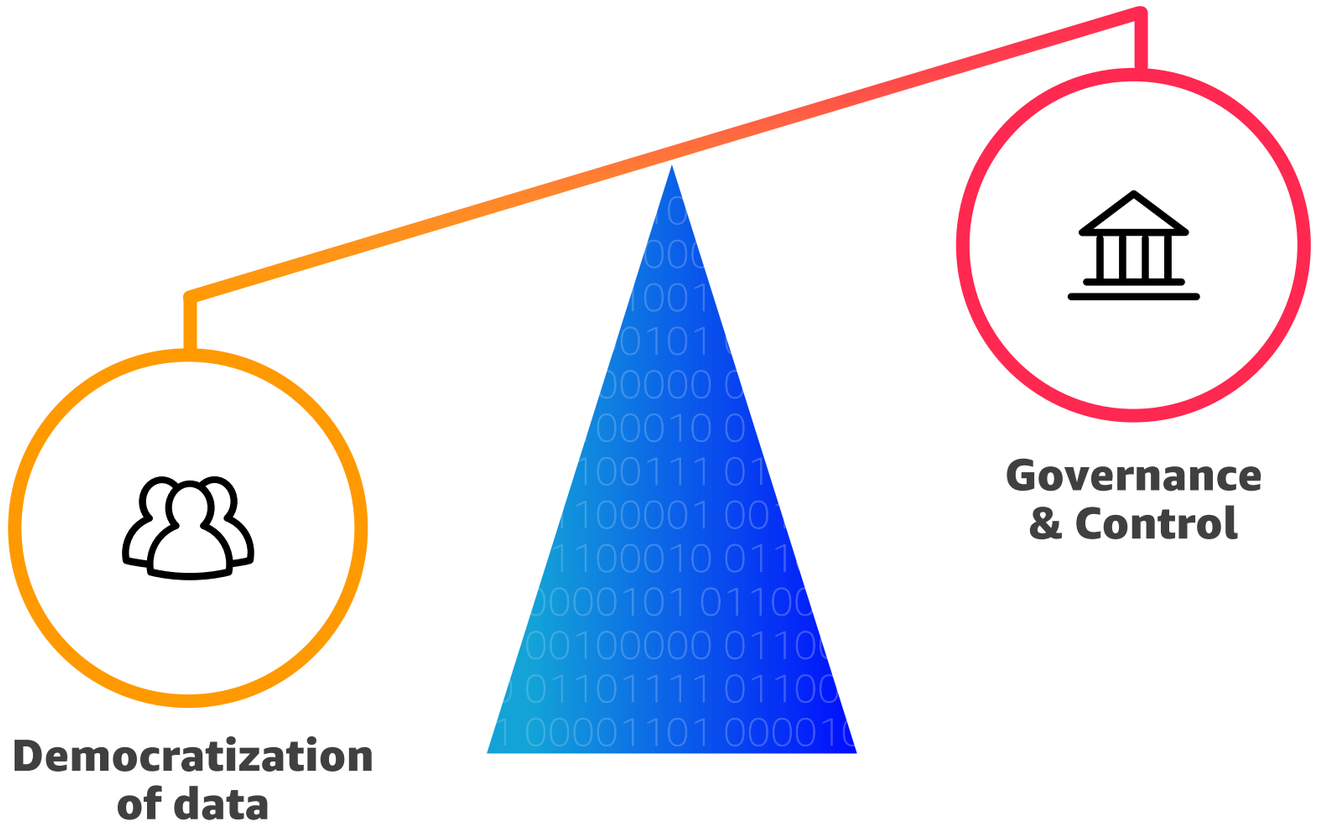
Years ago

Hadoop	Elasticsearch	Presto	Spark

Didn't exist

There are **more people** working with data than ever before

How do I provide democratized access to data to enable informed decisions while at the same time enforce data governance and prevent mismanagement of the data?



Managing the evolving data landscape

Flexible

Open APIs and open data formats

Choice

Use the best analytic tool for the job, without data movement

Scale

Platforms that scale up to 1,000x

Secure

Full auditability, access controls, and data governance

Amazon Redshift

The 4 things that matter most

Speed

Simplicity

Scale

Security

AWS re:Invent

aws

More customers use

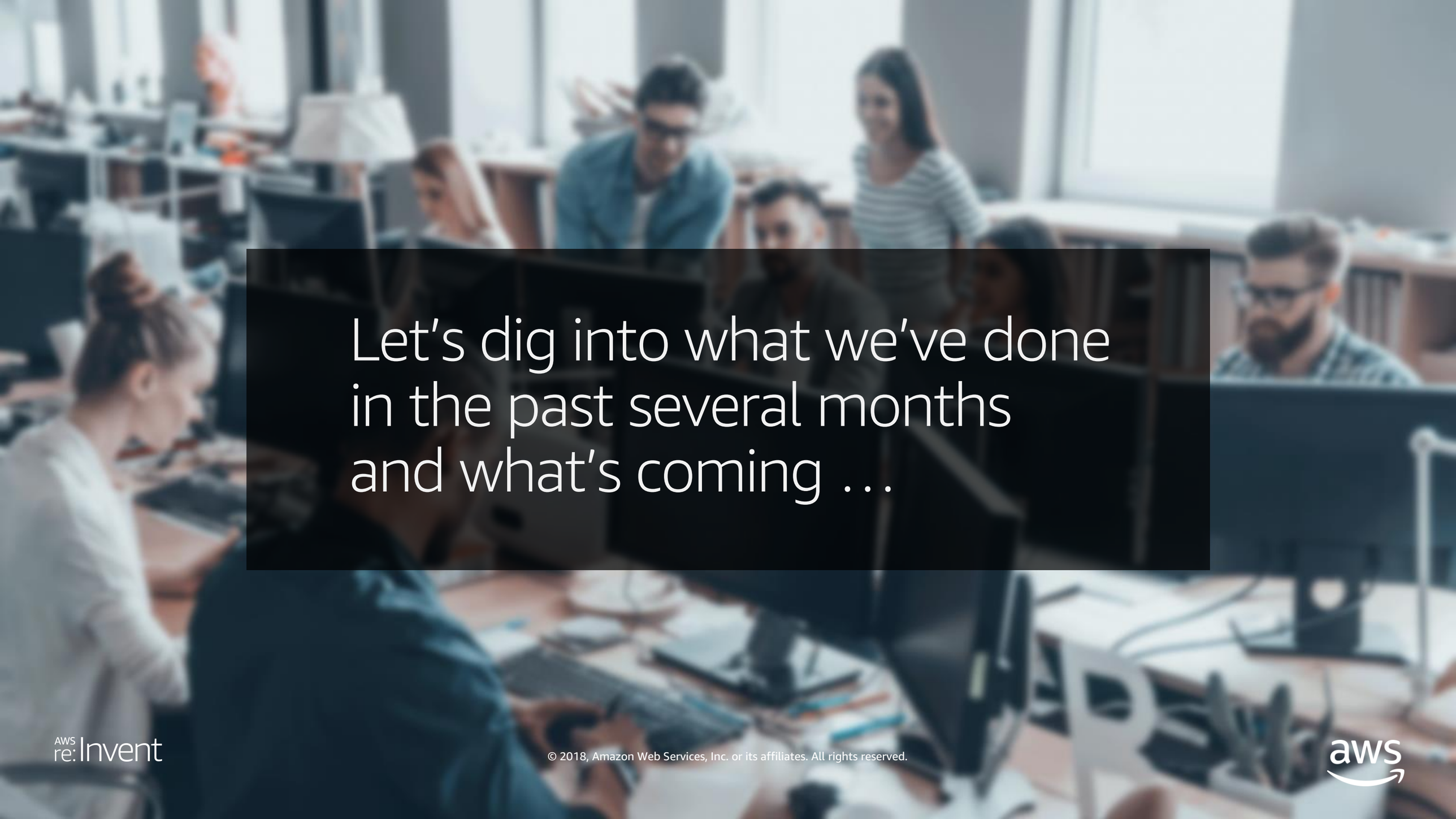
Amazon Redshift

for their data warehouse workloads than anyone else

AWS re:Invent

aws

AWS re:Invent



Redshift is **growing fast** and **innovating faster**

Since we last spoke …

Automatically enabled short query acceleration

Support for lateral column alias reference

New Quick Starts

New CloudWatch metrics

Customized Recommendations with Advisor

Current and trailing tracks for release update

Federated authentication with single sign-on

Improved performance for commits

COPY from Parquet and ORC file formats

Support for Parquet and ORC in Kinesis Data Firehose

Improved workload management console experience

Query Editor

Support for late-binding views

SQL Scalar user-defined functions

Integration with AWS Glue

Support for Nested Data with Spectrum

Spectrum support for DATE data type

Improved performance for UNION ALL queries

Free upgrade from DC1 to DC2 RIs

Query monitoring rules (QMR)

Support for Zstandard high compression encoding

Query processing improvements

220+

features and enhancements released*

Support for Python UDF logging module

Enhanced VPC routing

Additional Spectrum regions

Support for Scalar JSON and Ion data types

Late materialization for faster query processing

Support for DATE data type with Spectrum

Short Query Acceleration

Utilization reports

Machine learning integration to accelerate dashboards and interactive analysis

Improved resource management for memory-intensive queries

Faster string manipulation

Automatically hopping queries without restarts

Support for uppercase column names

Result Caching for Repeat Queries

Support for LISTAGG DISTINCT

Support for ORC and Grok file formats

Integration with QuickSight

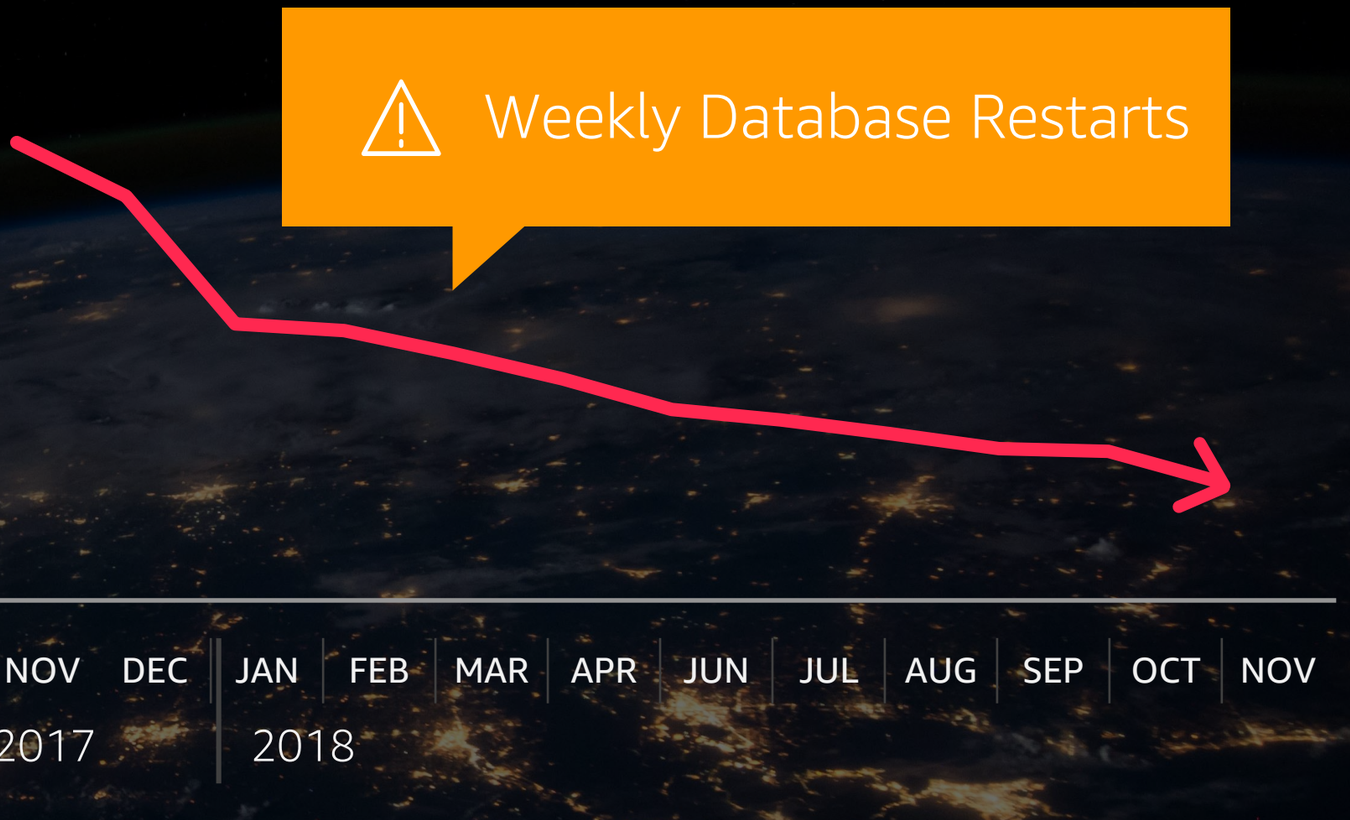
DMS support with Redshift

3.5x Improved Throughput

Improved performance for repeat queries

***Since re:Invent 2017**

Improvements in availability since we last talked



Improvements to speed

Compiled code cache

Support for lateral column alias reference

Query planning

Cluster resize operations

COPY operation when ingesting data from Parquet and ORC formats

Query processing improvements

Result caching

Queries operating over CHAR and VARCHAR columns

2x the number of tables in a cluster

Late materialization

Queries with intermediate subquery results that can be distributed

Single-row inserts

Complex EXCEPT subqueries

DC2 nodes

Short query acceleration

Improvements for the COPY operation when ingesting data from Parquet and ORC formats

Commit processing enhancements

Resource management for memory-intensive queries

Query rewrites that pushdown selective joins into a subquery

Faster string manipulation

Queries that refer to stable functions with constant expressions

Performance improvement for queries that refer to stable functions over constant expressions

Hash join memory utilization optimizations and cache line prefetching

Expressions on the partition columns of external tables

Joins involving large numbers of NULL values in a join key column

***Since re:Invent 2017**

AWS re:Invent

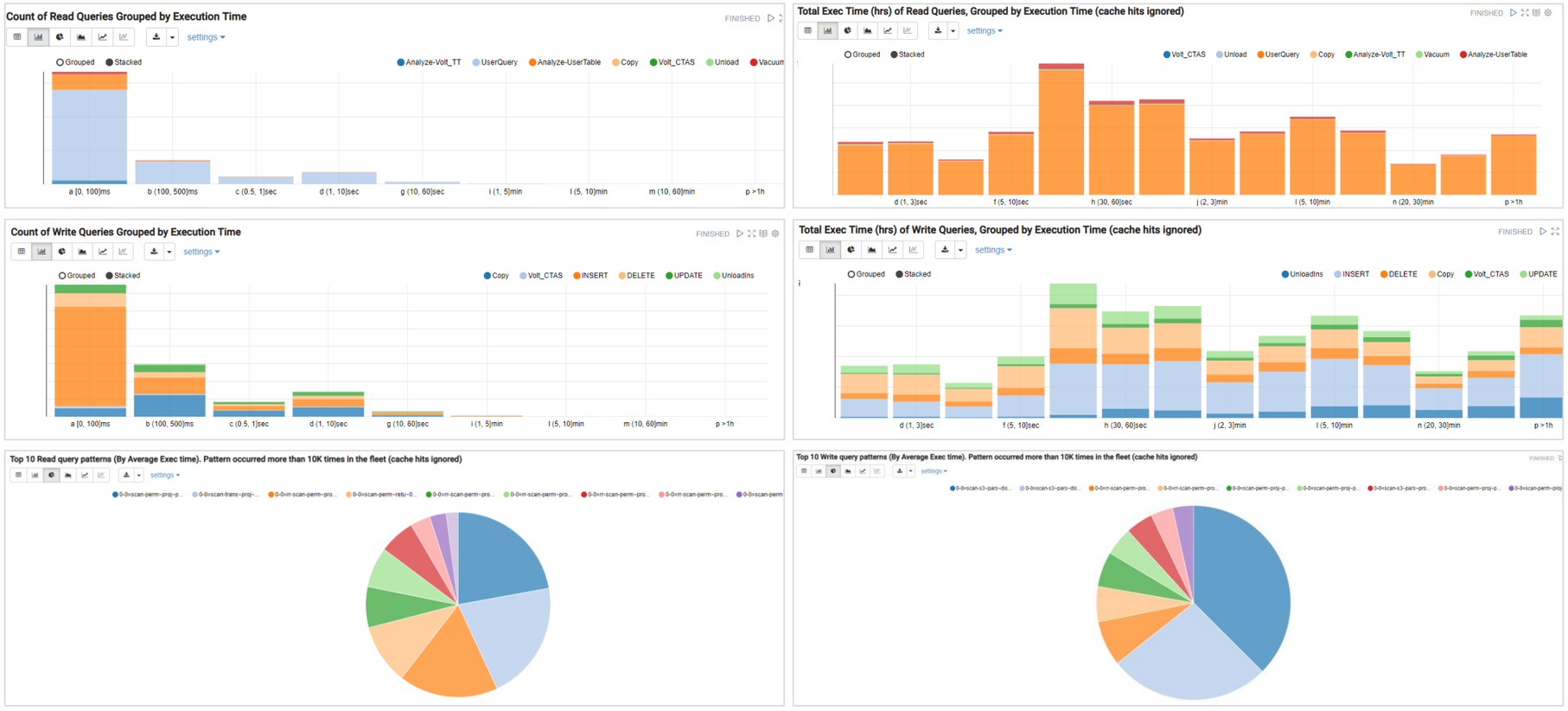
aws

Increases in performance in **real-world workloads**

How do we improve real-world performance?

17x	10x	3x	2x
for repetitive queries	for bulk-deletes	for single-row inserts	for commits

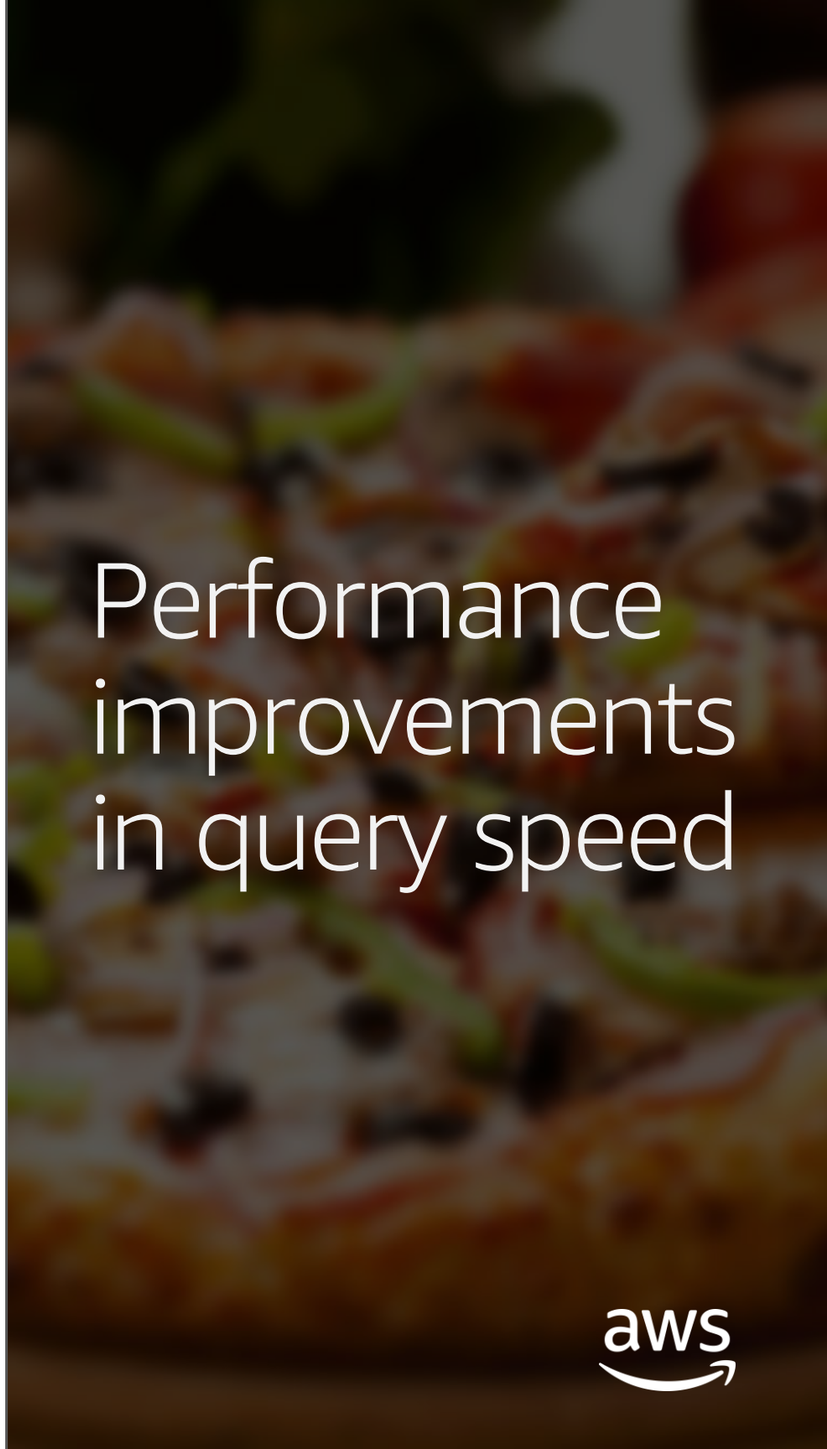
How we leverage **fleet telemetry**



“Redshift’s query performance and scalability has been increasing, even though our data has grown. In the last 10 months, we have seen commit performance **increase by 500%** without any increase in cost.”

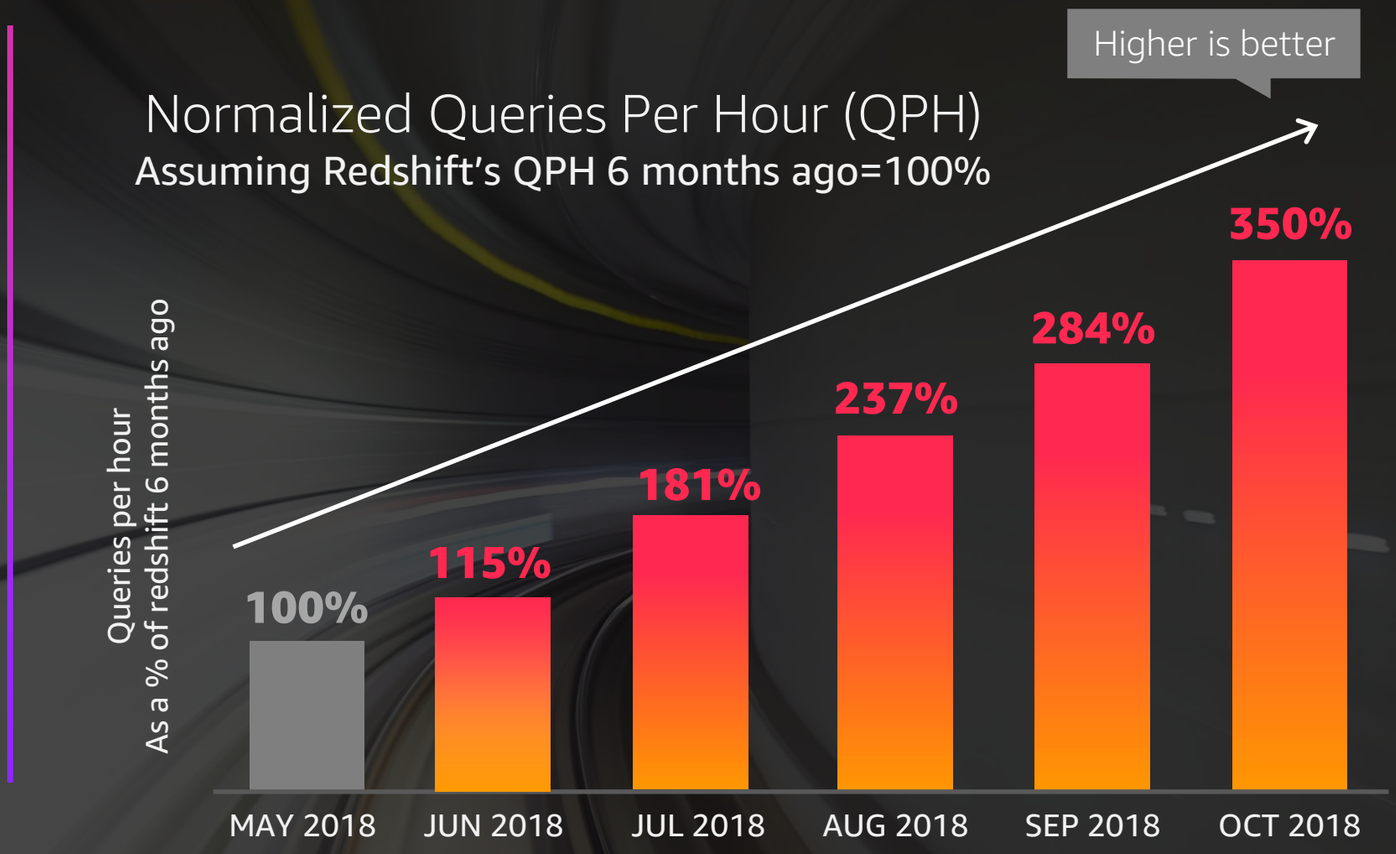
- Minero Aoki
Senior Data Engineer, Cookpad Inc.

cookpad



aws

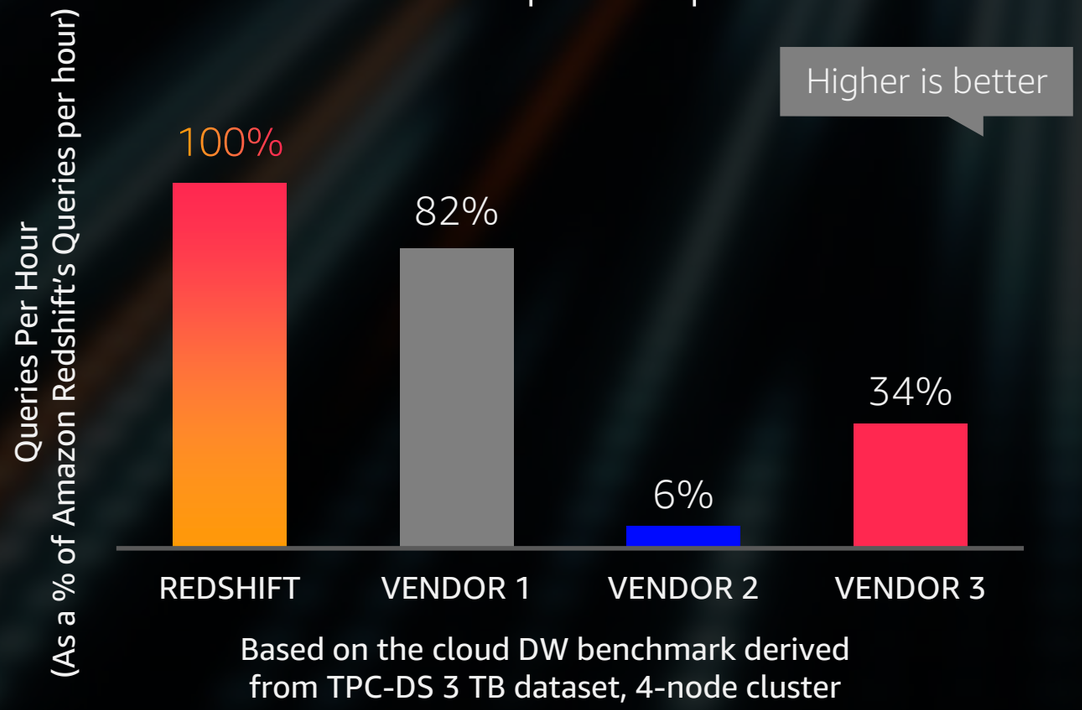
Amazon Redshift is now over 3x faster on standard benchmarks than 6 months ago



Amazon Redshift

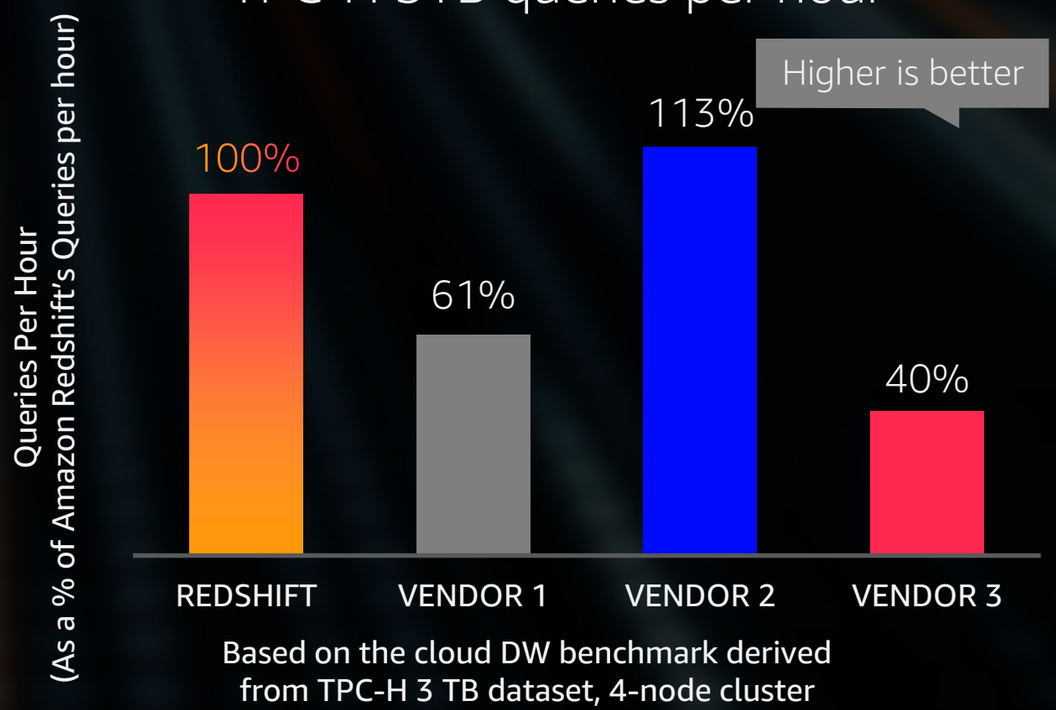
up to 16x faster

TPC-DS 3TB queries per hour



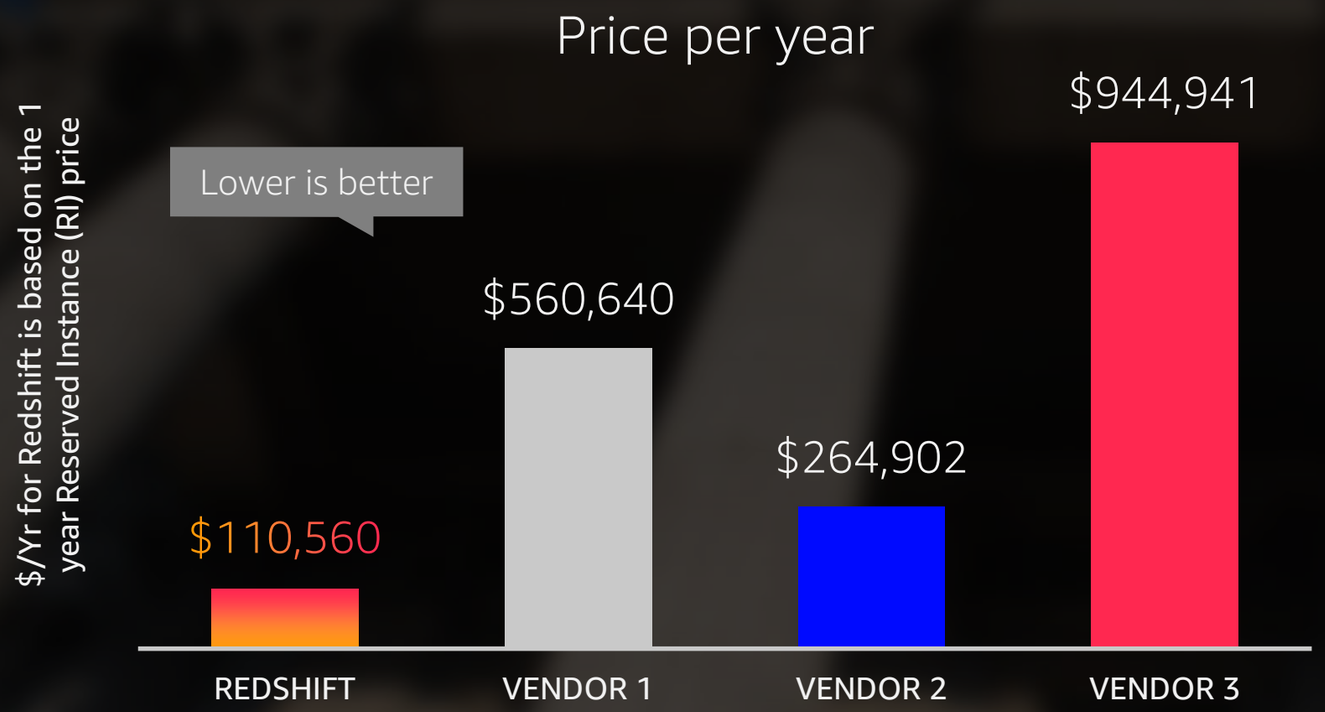
Based on the cloud DW benchmark derived from TPC-DS 3 TB dataset, 4-node cluster

TPC-H 3TB queries per hour



Based on the cloud DW benchmark derived from TPC-H 3 TB dataset, 4-node cluster

Amazon Redshift is the most cost-effective cloud data warehouse



The only data warehouse with reserved instances saving

up to 75%

The best price-to-performance

Fleet telemetry on query wait times

87% of Redshift customers don't have significant wait times

Remaining 13% have bursts of activity averaging 10 minutes at a time

Concurrency Scaling for bursts of user activity (Preview)

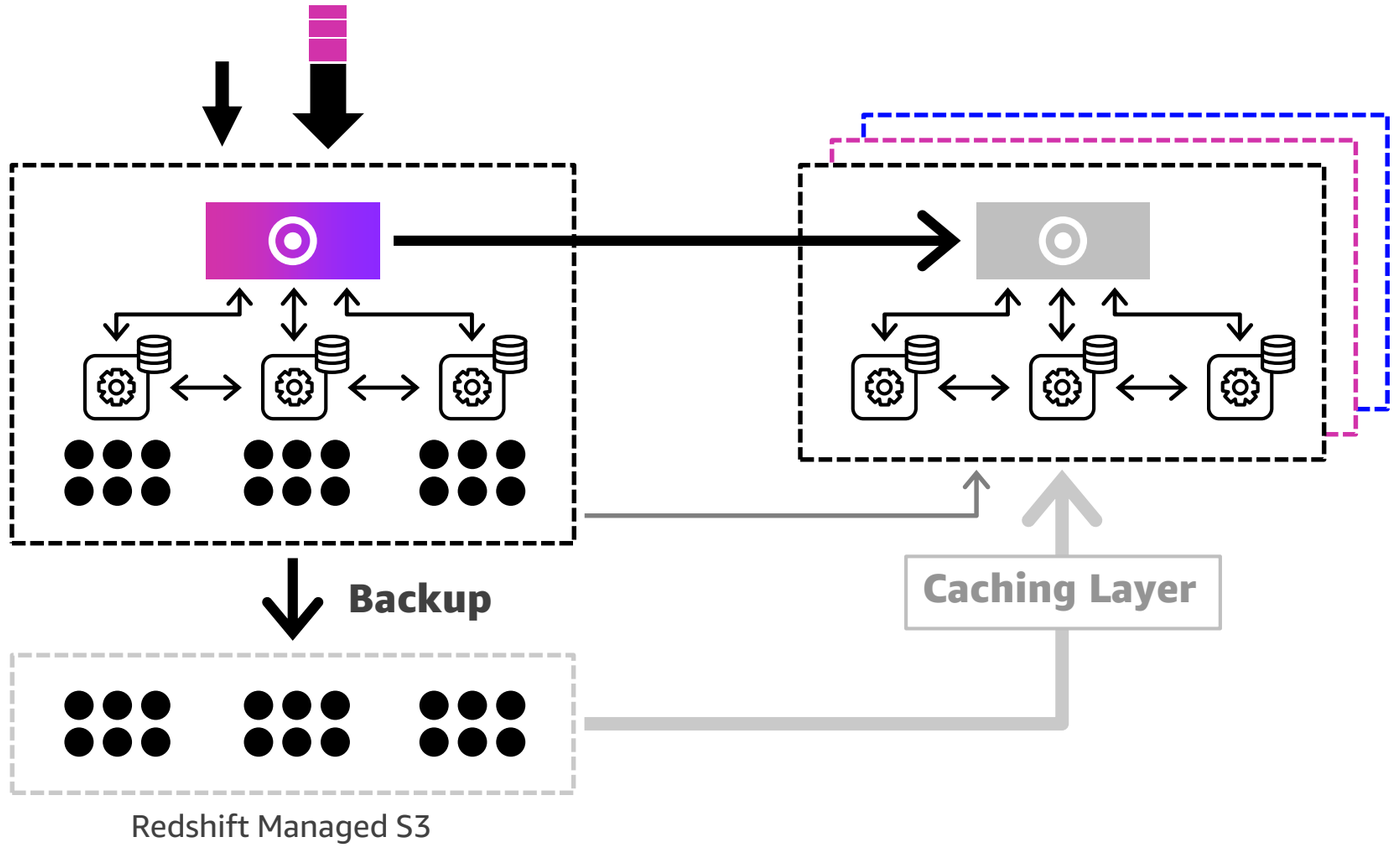
New!

Automatically creates more clusters on-demand

Consistently fast performance even with thousands of concurrent queries

No advance hydration required

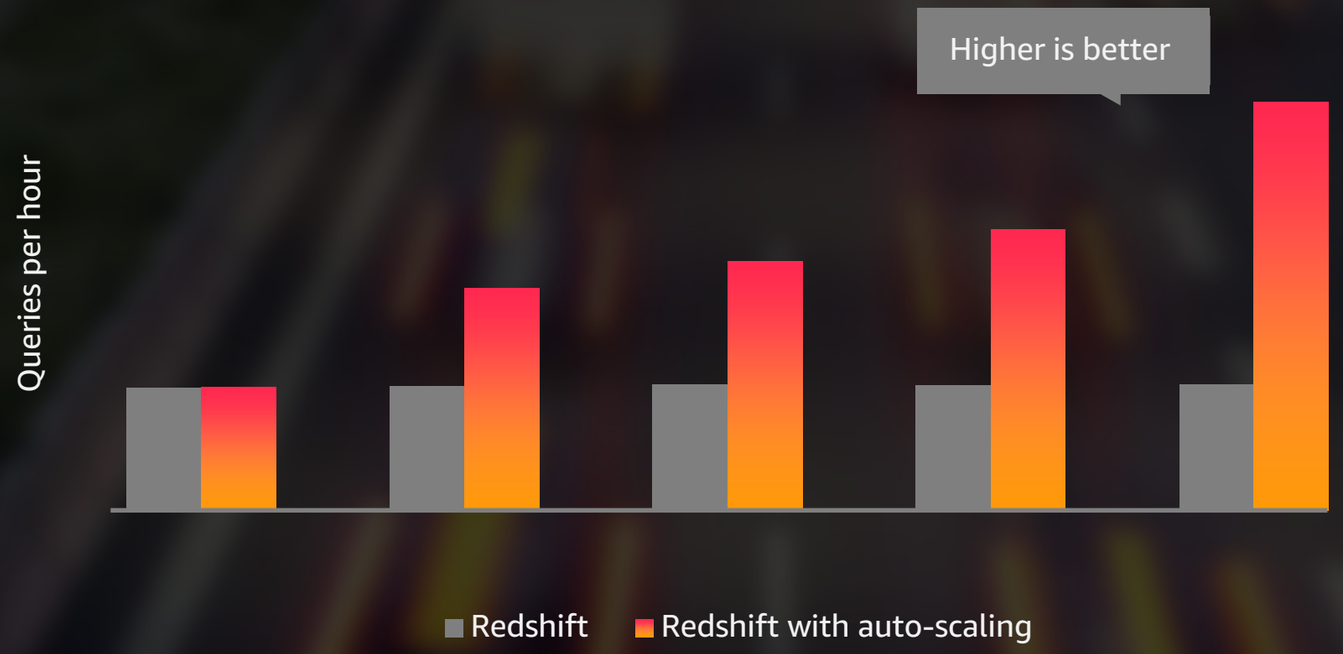
Quickly scale to serve changing query workload



Results with Concurrency Scaling

Concurrency Scaling is free for more than 97% of Redshift customers.

For every 24 hours your main cluster is in use, we'll provide a one-hour credit for concurrent cluster usage.



aws

Redshift Elastic Resize (GA)

New!

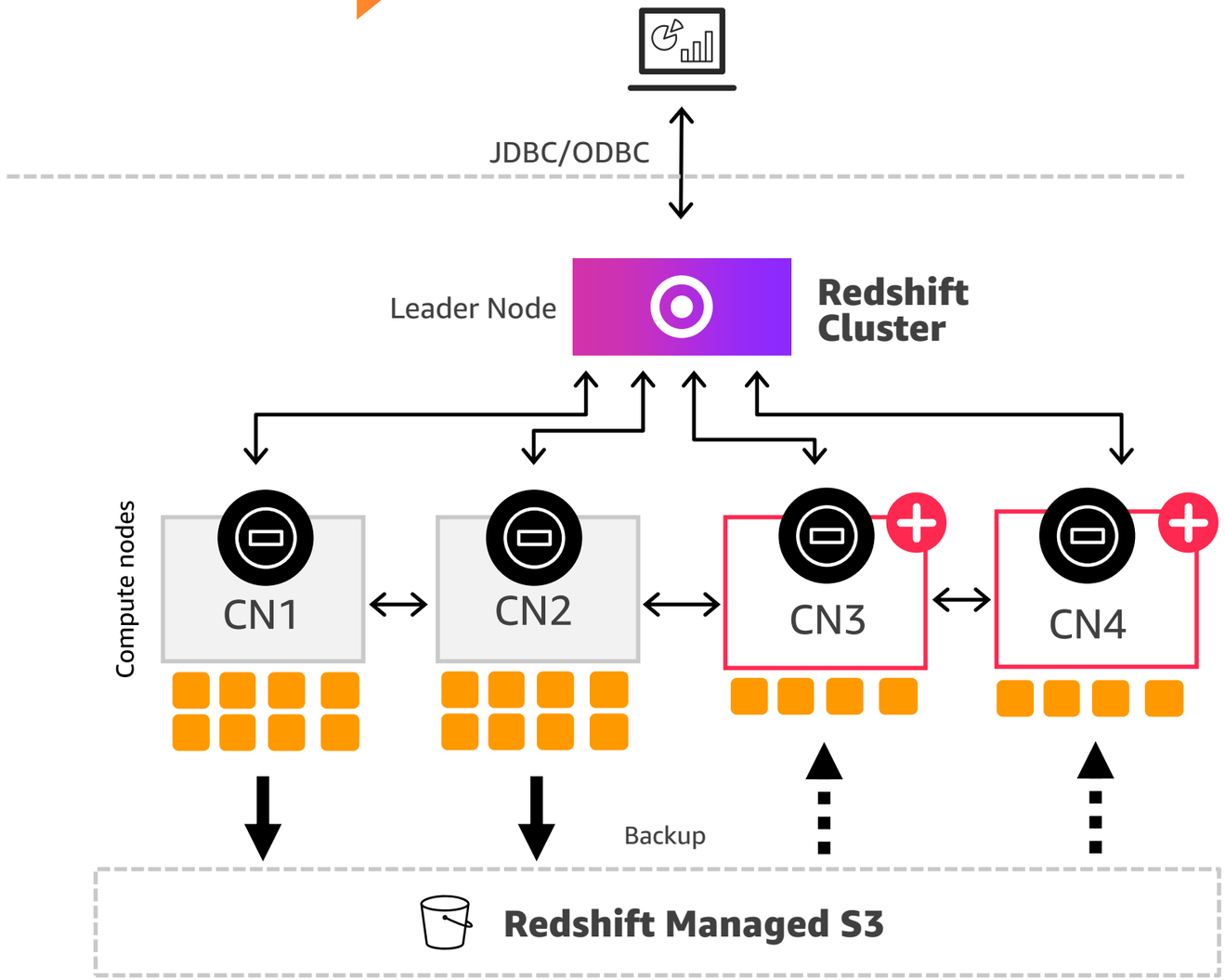
Scale up and down in minutes

Adds additional nodes to Redshift cluster

Distributes data across new configuration in minutes

Minimal transition time

Scale compute and storage on-demand



Efficiency of backup performance

Automatic vacuum delete

Enhancements to VACUUM DELETE

CloudWatch metrics for Workload Execution Breakdown

Query Editor

Cluster resize

CloudWatch metrics for Query Throughput by WLM Queues

Redshift Advisor for best practice recommendations

Manage components of a multi-part query in the AWS console

Current and trailing tracks for release updates

Improvements to simplicity

Lateral column alias reference

Stream real-time data in Parquet or ORC formats using Kinesis Data Firehose

CloudWatch metrics for Query Duration by WLM Queues

Cluster resize operations

Short query acceleration is self-optimizing

Query Monitoring Rules (QMR) now support 3x more rules

Free upgrade from for DC1 RIs to DC2

DISTSTYLE AUTO distribution style

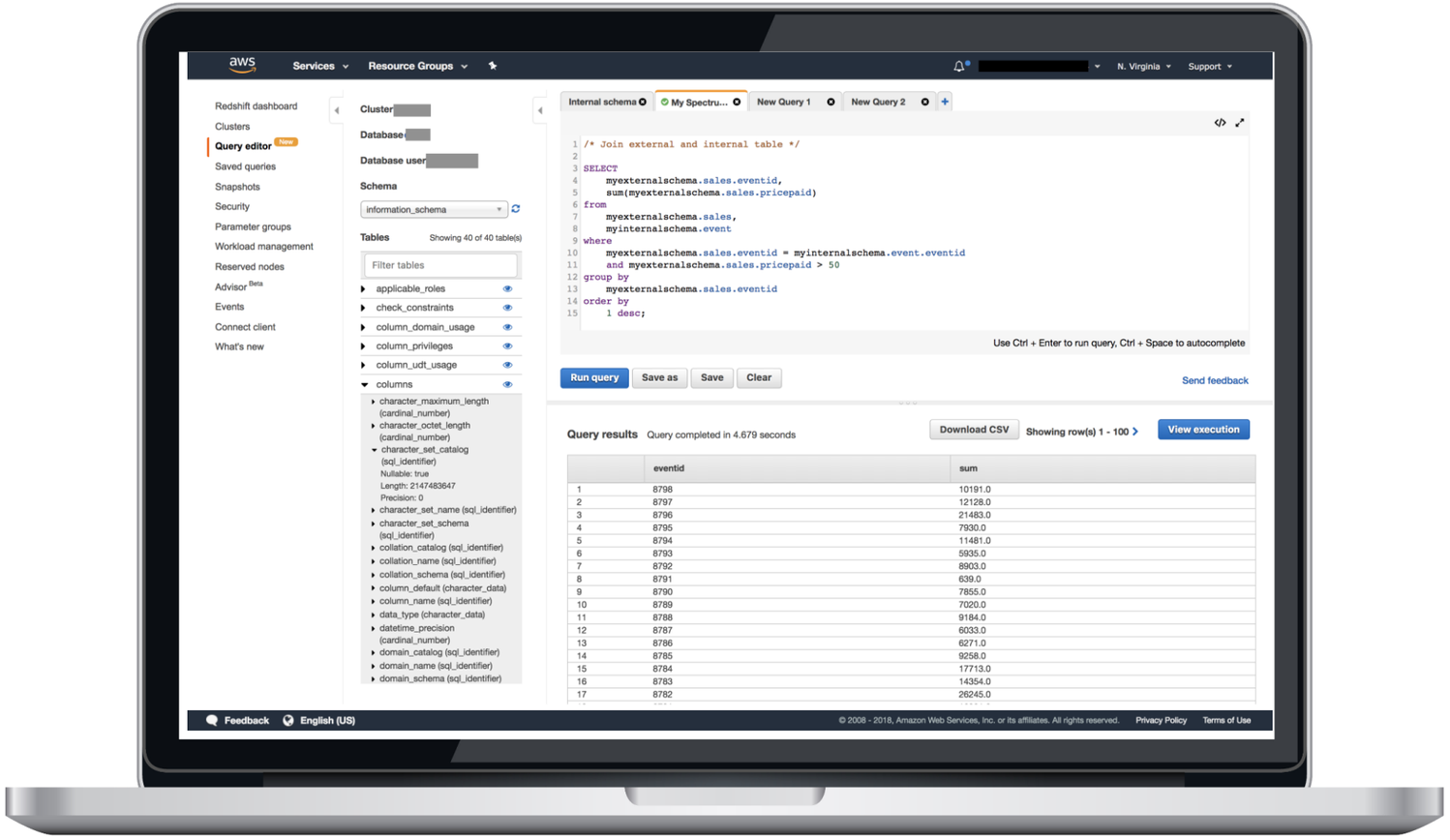
CloudWatch Query Runtime Breakdown metric

CloudWatch metrics for Query Throughput, Query Duration

***Since re:Invent 2017**

Redshift Query Editor

Launched in October!



Query data directly from the AWS console

Results are instantly visible within the console

No need to install and setup an external JDBC/ODBC client

Redshift Advisor

Launched in July!

Provides automated recommendations
to help optimize database performance and decrease operating costs

>96% of clusters
have tailored feedback

Actionable WLM
COPY, storage, and system maintenance advice

Intelligent recommendations
for tuning based on continuous workload analysis

Amazon Redshift intelligent administration

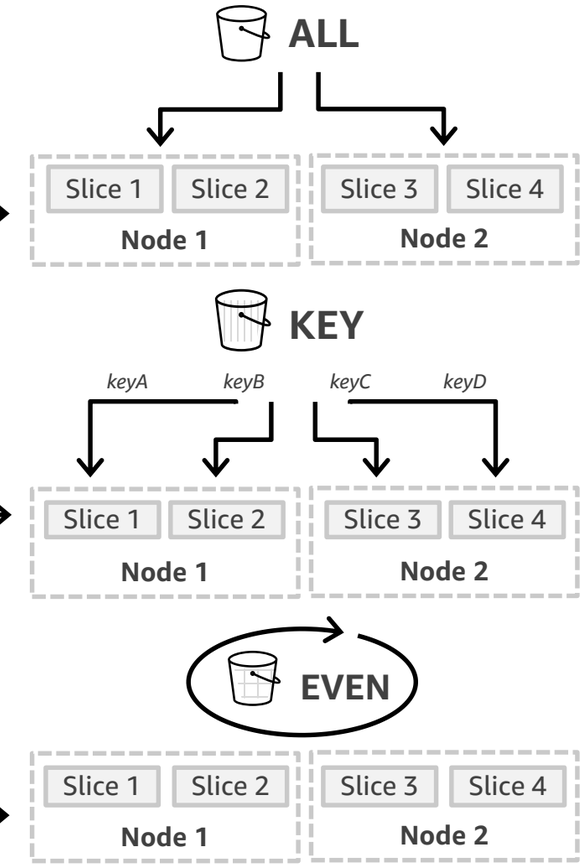
Coming Soon!

Automates data distribution in tables for improved performance and disk space utilization.

No more messing with distkeys!

Advise

Amazon Redshift recommended distribution key



Provides intelligent recommendations for tuning based on continuous workload analysis.

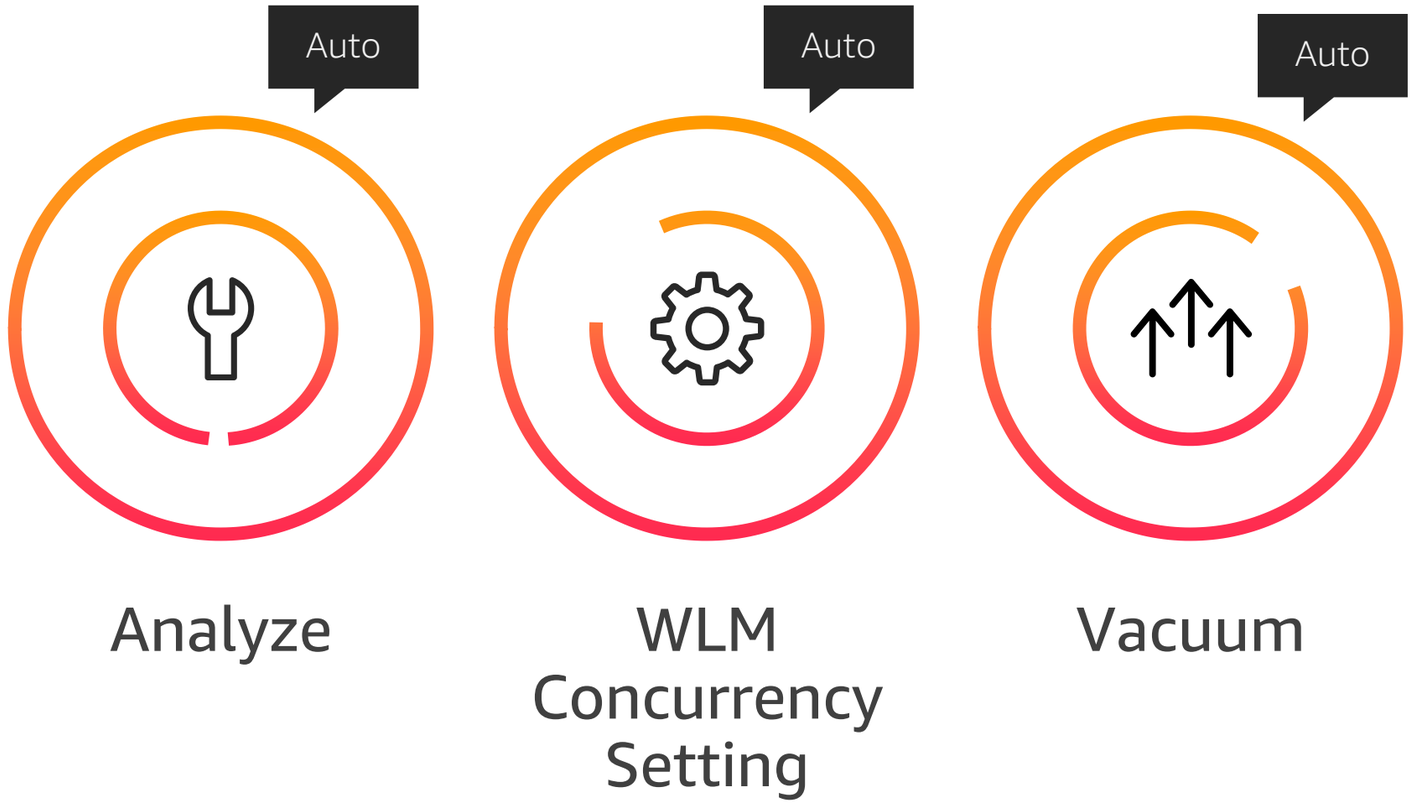
Amazon Redshift intelligent maintenance

Coming Soon!

Maintenance processes like vacuum and analyze will automatically run in the background.

Moving towards zero-maintenance.

Redshift will automatically adjust the WLM concurrency setting to deliver optimal throughput.



Run stored procedures in Amazon Redshift

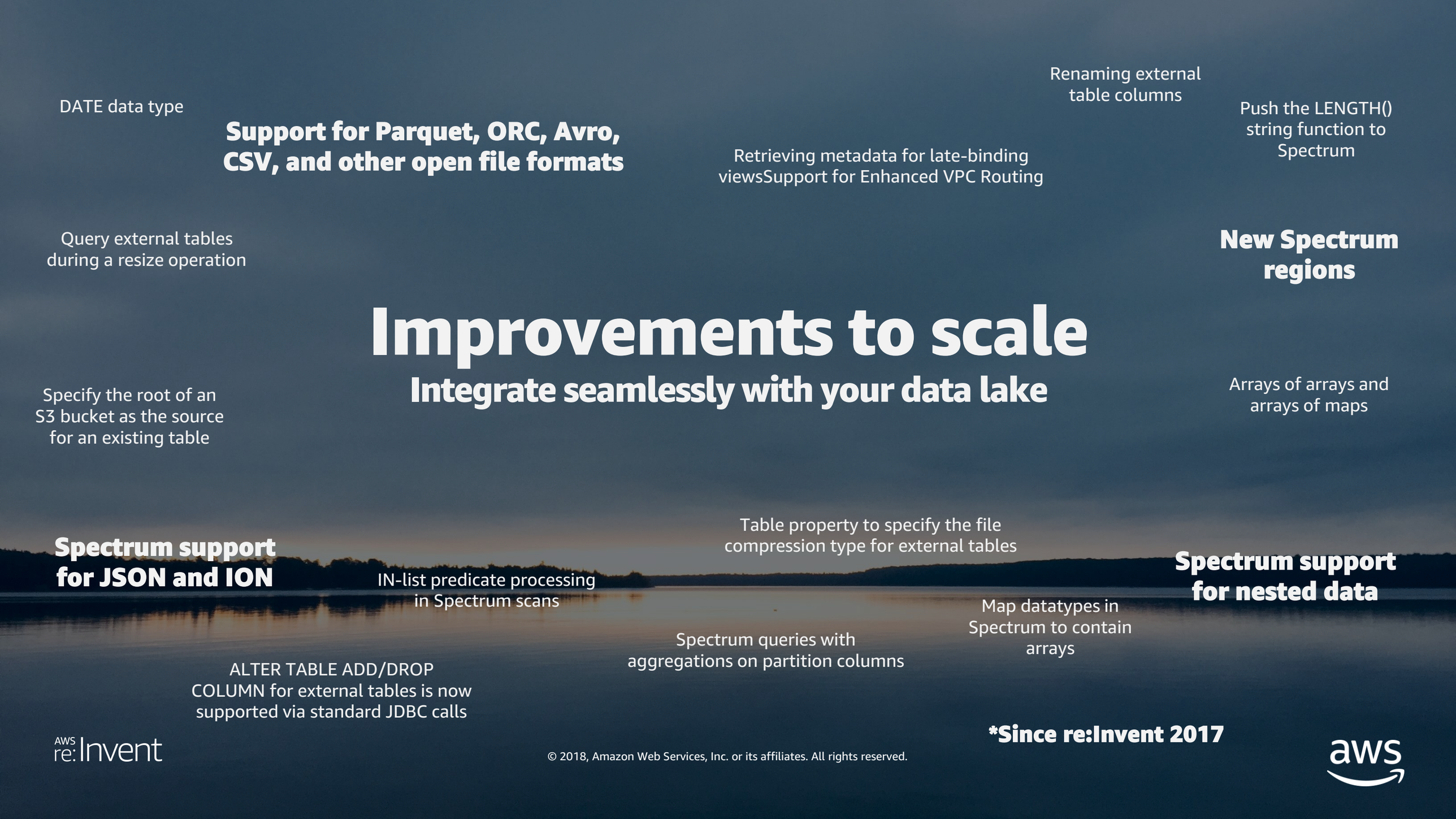
Coming Soon!

Bring your existing **Stored Procedure** and run in Redshift.

Migrating to Redshift is even easier!

Redshift will support Stored Procedure in PL/pgSQL format, enabling you to bring your existing Stored Procedure to Redshift.

Support for stored procedure provides the ability to run code where the data is to efficiently run ETL, data validation, and custom business logic.



AWS re:Invent

aws

Amazon Redshift Spectrum

Extend the data warehouse to exabytes of data in S3 Data Lake

No loading required

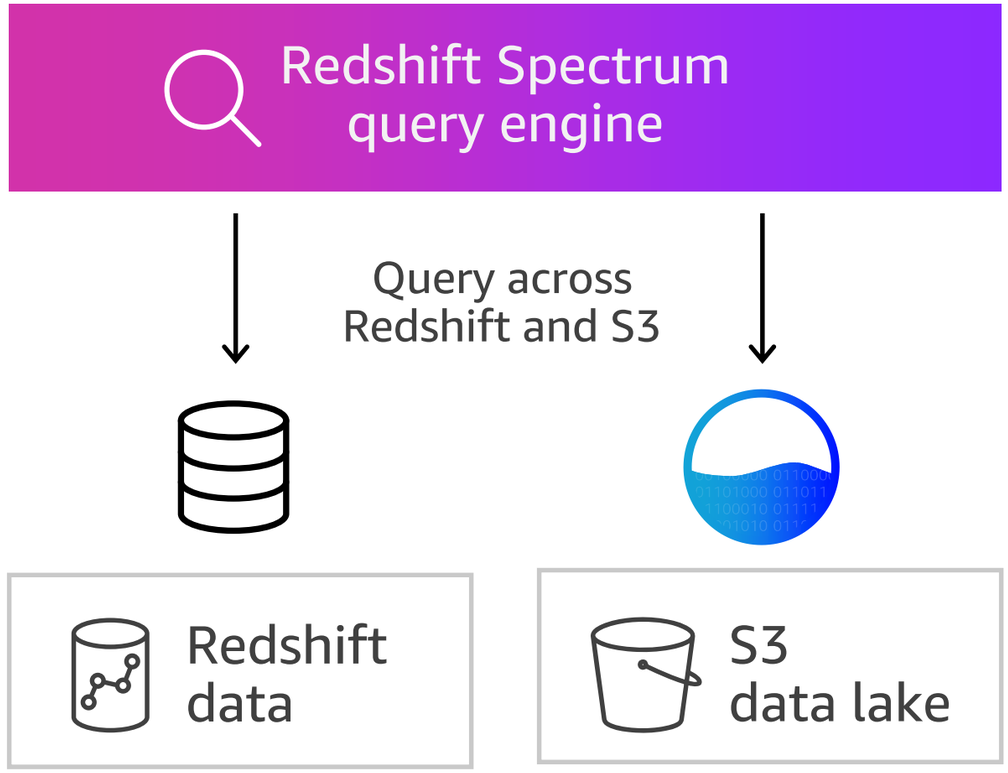
Scale compute and storage separately

Directly query data stored in S3

Parquet, ORC, Avro, Grok, and CSV data formats

→ Unload to Parquet
→ Spectrum Request Accelerator

Coming Soon!



Amazon Redshift is Scalable

Redshift Spectrum: Exabyte data lake query in under three minutes

Imagine you are the manager at a Seattle book store. An author released her 8th book in a popular series, and you need to figure out how many copies to order.

You have historical book sales data in Amazon S3

Roughly 140 terabytes of customer item order detail records for each day over the past 20 years

190 million files across 15,000 partitions in S3

One partition per day for USA and rest of world

Total data size is over an exabyte

Redshift Spectrum <3 minutes

Optimization

Compression	**5X**
Columnar file format	**10X**
Scanning with 2500 nodes	**2,500X**
Static partition elimination	**2X**
Dynamic partition elimination	**350X**
Amazon Redshift query optimizer	**40X**

* Query used a 20 node DC1.8XLarge Amazon Redshift cluster
* Not actual sales data—generated for this demo based on data format used by Amazon Retail.

Improvements to security

Encrypt your previously unencrypted cluster with 1-click

Federated authentication with single sign-on

Cross-region backups for KMS-encrypted clusters

Default access privileges

Tag-based permissions

Enhanced VPC Routing

SAS integration enhancements

Encrypt unloaded data using S3 server-side encryption with AWS KMS keys

IAM roles with COPY and UNLOAD commands

Superusers to grant users access to all rows in selected system tables

***Since re:Invent 2015**

AWS re:Invent

aws

Security is **built-in**

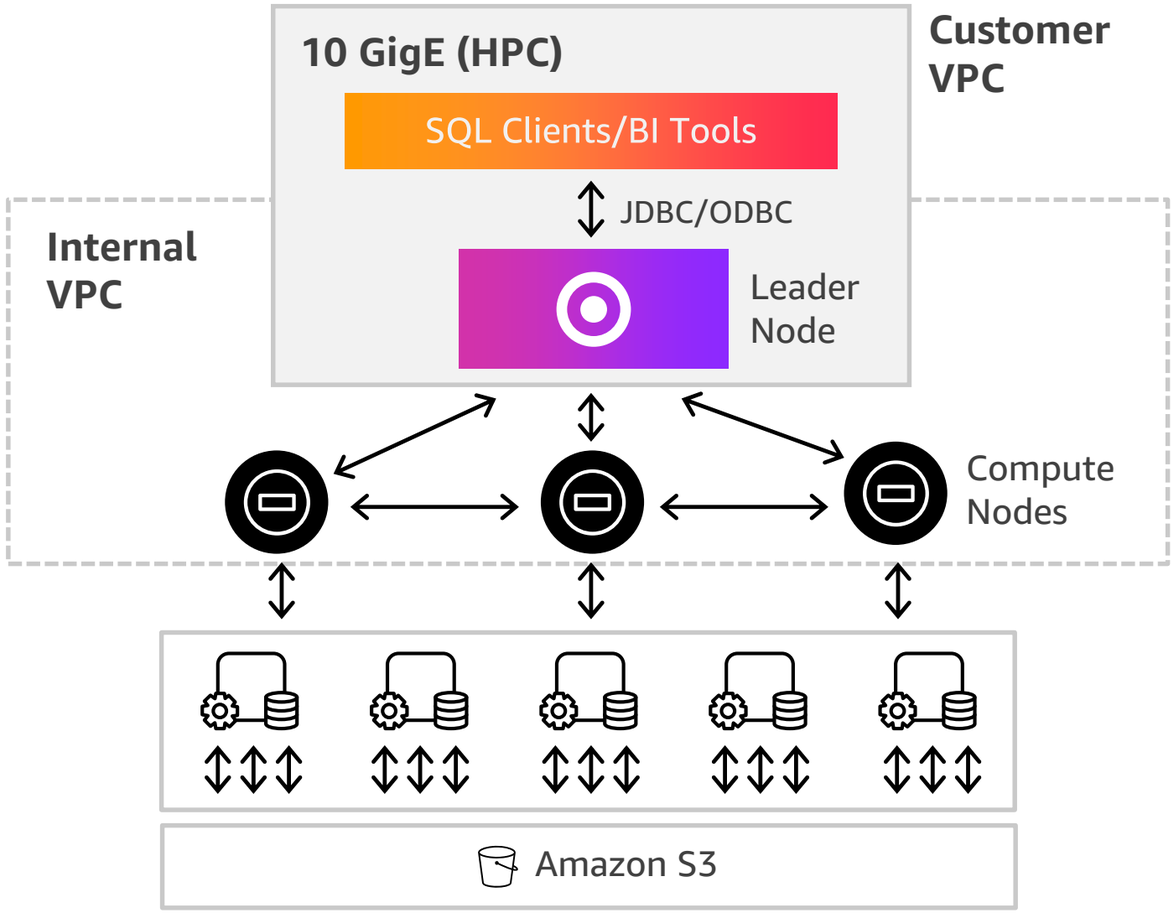
End-to-end encryption

Integration with AWS Key Management Service

Select compliance certifications*

AICPA SOC · FedRAMP · DEPARTMENT OF DEFENSE · PCI DSS COMPLIANT · HIPAA COMPLIANT · ISO

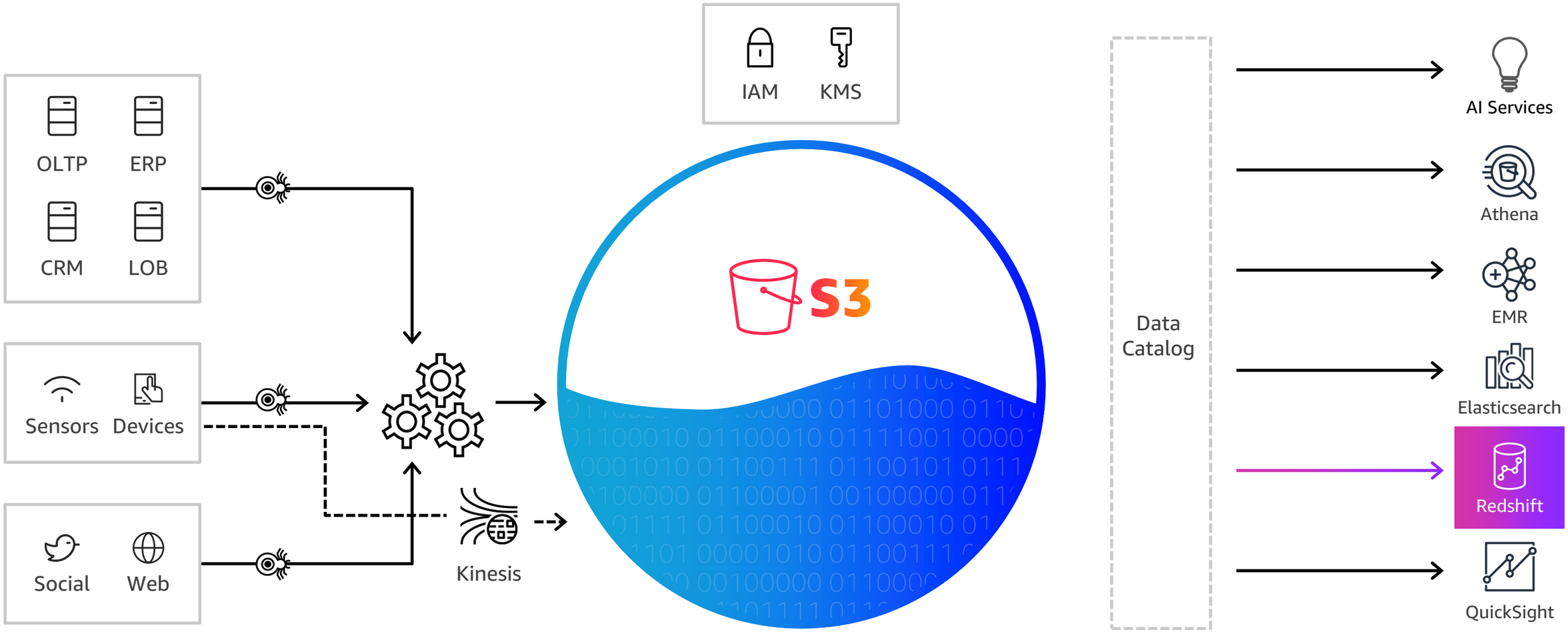
Network Isolation



*Full list of compliance certifications is available here: https://aws.amazon.com/compliance/

Integration with Amazon Lake Formation

Coming Soon!



Amazon Redshift

New features

Speed

- Improving short query acceleration
- NEW! Auto data distribution
- NEW! Elastic resize
- COMING SOON Concurrency Scaling

Scale

- COMING SOON Spectrum Request Accelerator
- COMING SOON Unload to Parquet

Simplicity

- COMING SOON WLM Concurrency Setting
- COMING SOON Auto-Vacuum & Auto-Analyze
- COMING SOON Support for stored procedures
- COMING SOON Auto Data Distribution
- NEW! Deferred Maintenance
- COMING SOON Snapshot Scheduler

Security

- COMING SOON Amazon Lake Formation integration

aws re:Invent

aws

More places to learn about Amazon Redshift

Try it out for yourself:
aws.amazon.com/redshift/

Modern Data Warehousing on AWS ebook

Sign up for **Concurrency Scaling**

Amazon Redshift **customer use cases**

Blog on performance matters:

Amazon Redshift is now up to 3.5x faster for real-world workloads

Amazon Redshift and the **art of performance optimization** in the cloud

by Werner Vogels

Building a Proof of Concept for Amazon Redshift

Thank you!

Vidhya Srinivasan
Vid@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.

AWS re:Invent

aws