
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT201

What's new with Amazon Redshift

Neema Raphael

Chief Data Officer
Goldman Sachs

Eugene Kawamoto

Director, Product Management
Amazon Redshift

Agenda

- Business value of data and customer challenges
- Amazon Redshift and its evolution
- Investments and announcements around key use cases
- Modernizing data warehousing at Goldman Sachs with Amazon Redshift
- Get started

The growth of data brings its own set of unique challenges

In 2020, 64.2 ZB of data was created or replicated, and the amount of digital data created over the next five years will be greater than twice the amount of data created since the advent of digital storage

Source: IDC, Worldwide Global DataSphere Forecast, 2021–2025, Doc # US46410421, March 2021

WHAT CUSTOMERS ARE TELLING US:

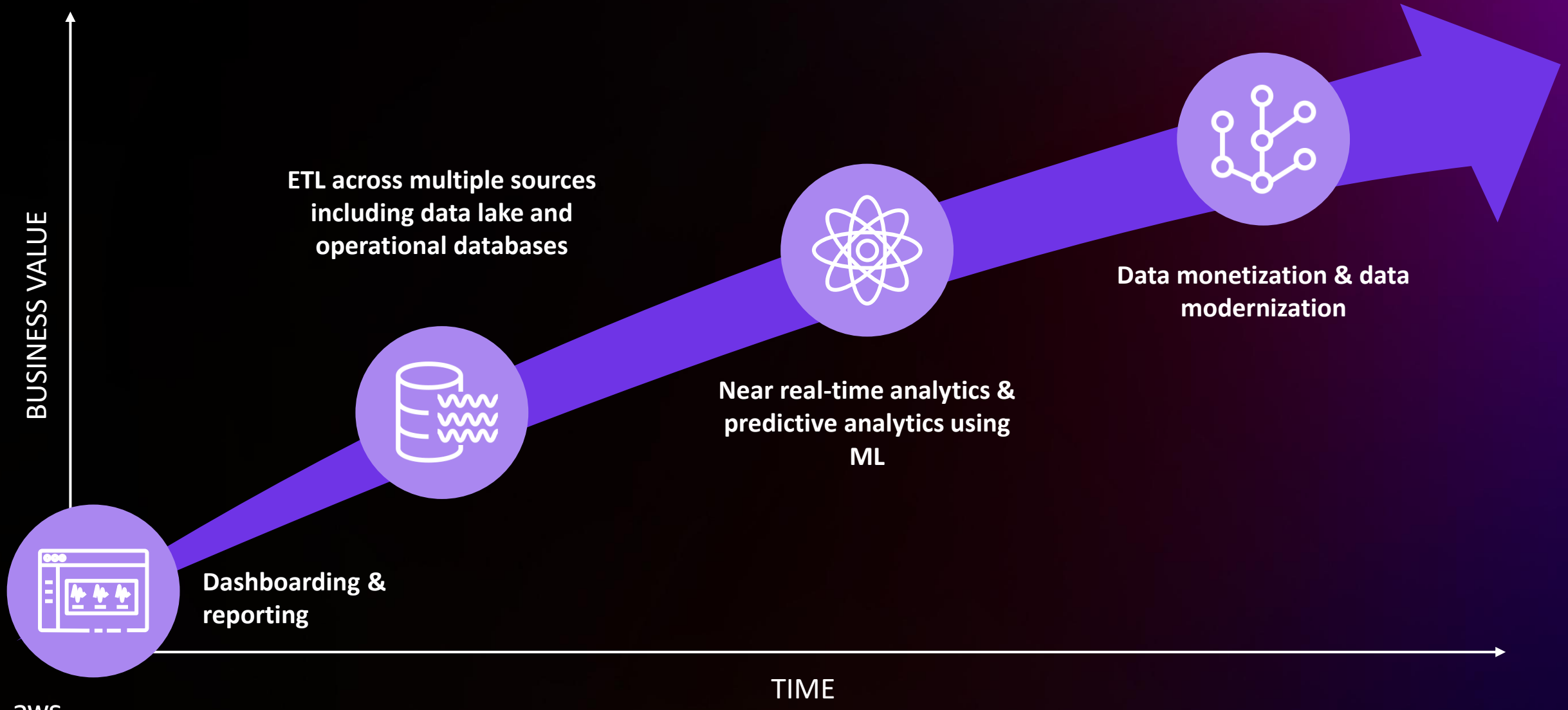
"Everyone's a data user and our data is everywhere"

"Our analytics workloads are mission-critical and run 24/7"

"We're looking for a secure, well-governed, and scalable architecture"

"We want consistent high performance for any amount of data or usage, without costs getting out of control"

Customers are evolving their data platforms



Listening to our customers for the last 10 years

INVENTING AND REINVENTING CLOUD DATA WAREHOUSING WITH AMAZON REDSHIFT

2012 — Terabytes to petabytes to exabytes — 2022

Customer needs

A cost-effective way to run fast analytic queries on large volumes of data

- Data lake analytics and query of open data formats
- Scale storage independently of compute
- Elastic scaling
- Query operational data
- Query without data movement, data sharing, and collaboration
- Data science needs beyond highly skilled data scientists
- Simplified query interface
- Minimal user involvement in performance acceleration, concurrent scaling
- Third-party data analytics
- More data users, needing to abstract complexity
- Real-time analytics in the warehouse
- Granular governance and security controls
- No-code data pipelines, ingestion
- Apache Spark

Amazon Redshift

Amazon Redshift launched as the first MPP cloud data warehouse at a price performance unattainable in on-premises MPP DWs

Price performance at any scale

Concurrency Scaling
Materialized Views
Workload Management
Short Query Acceleration
Vectorized Scans

Analyze all your data

Data Sharing
ADX Integration
Federated Query
Data Lake Query / Spectrum
Streaming Ingestion
Redshift ML
SUPER Data Type
Geospatial Analysis

Easy, secure, and reliable

Serverless
Query Editor V2
RBAC
Column-level Security
Row-level Security

Tens of thousands of customers analyze exabytes of data every day on Amazon Redshift

Amazon Redshift | Analyze all your data with the best price performance cloud data warehouse

Media & entertainment

Financial & professional services

Healthcare

Consumer services

Web and technology

aws

We continue to deliver on your most important requirements

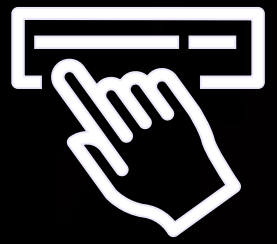
Price-performance at any scale

Analyze all your data

Easy, secure, and reliable

- Self-service analytics
- Easy data ingestion
- Data sharing & collaboration
- Rich analytics with data science & machine learning
- Secure and reliable analytics
- Best price-performance analytics

Self-service analytics



Amazon Redshift Serverless

NEW! Support for tagging, additional monitoring views & Query Monitoring Rules, and new AWS Region availability (US West (Northern California), Europe (Paris))

Get insights from data in seconds

Experience consistently high performance

Get started with no modifications

Pay for what you use

YOU
focus on insights

Automatic compute provisioning

Automatic scaling

Automated patching

Automatic workload management

Advanced monitoring

Backup and recovery

Performance at scale

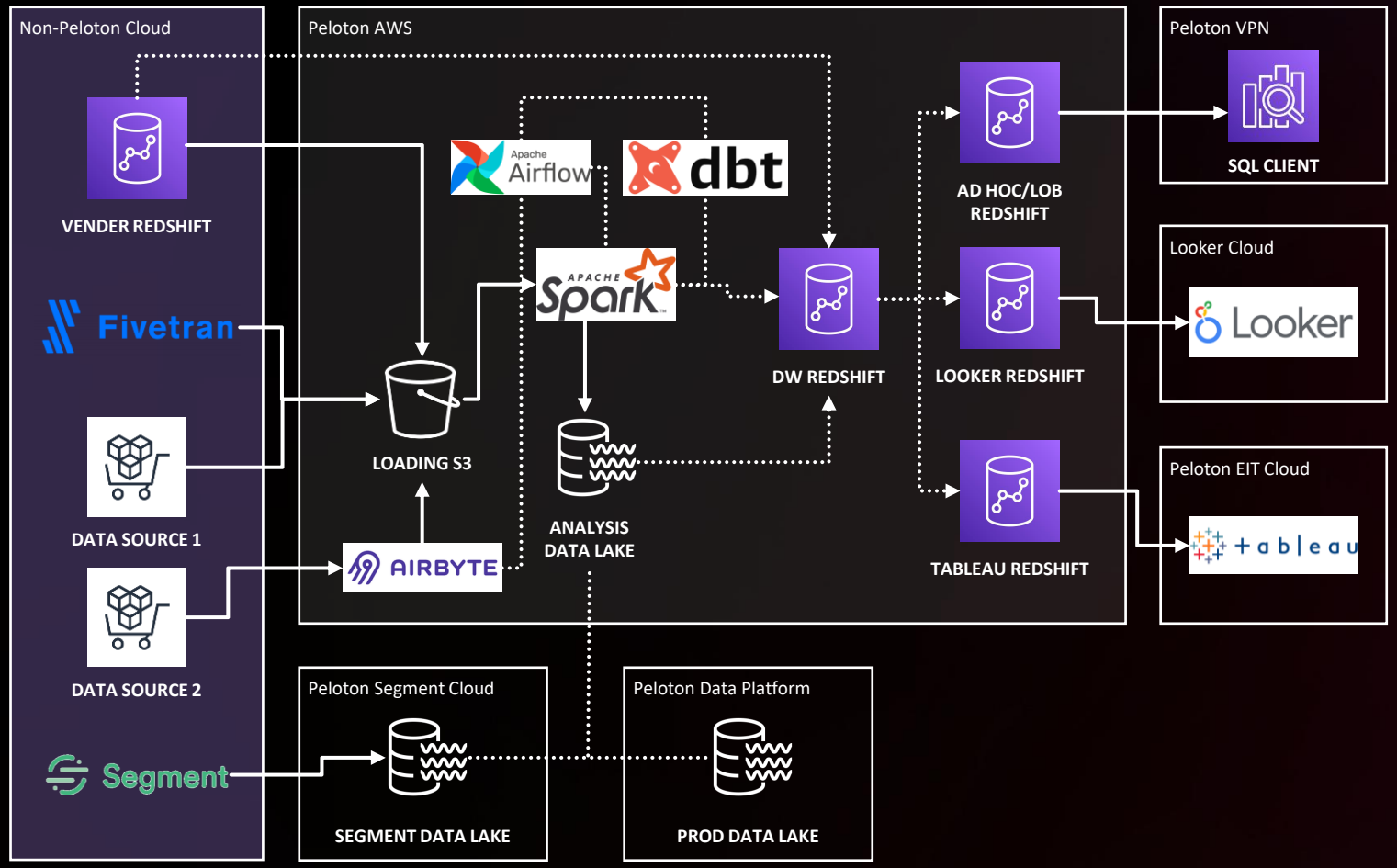
Security and industry compliance

aws

takes care of the rest

aws

Easier, faster, and cost-effective analytics at Peloton



PELOTON

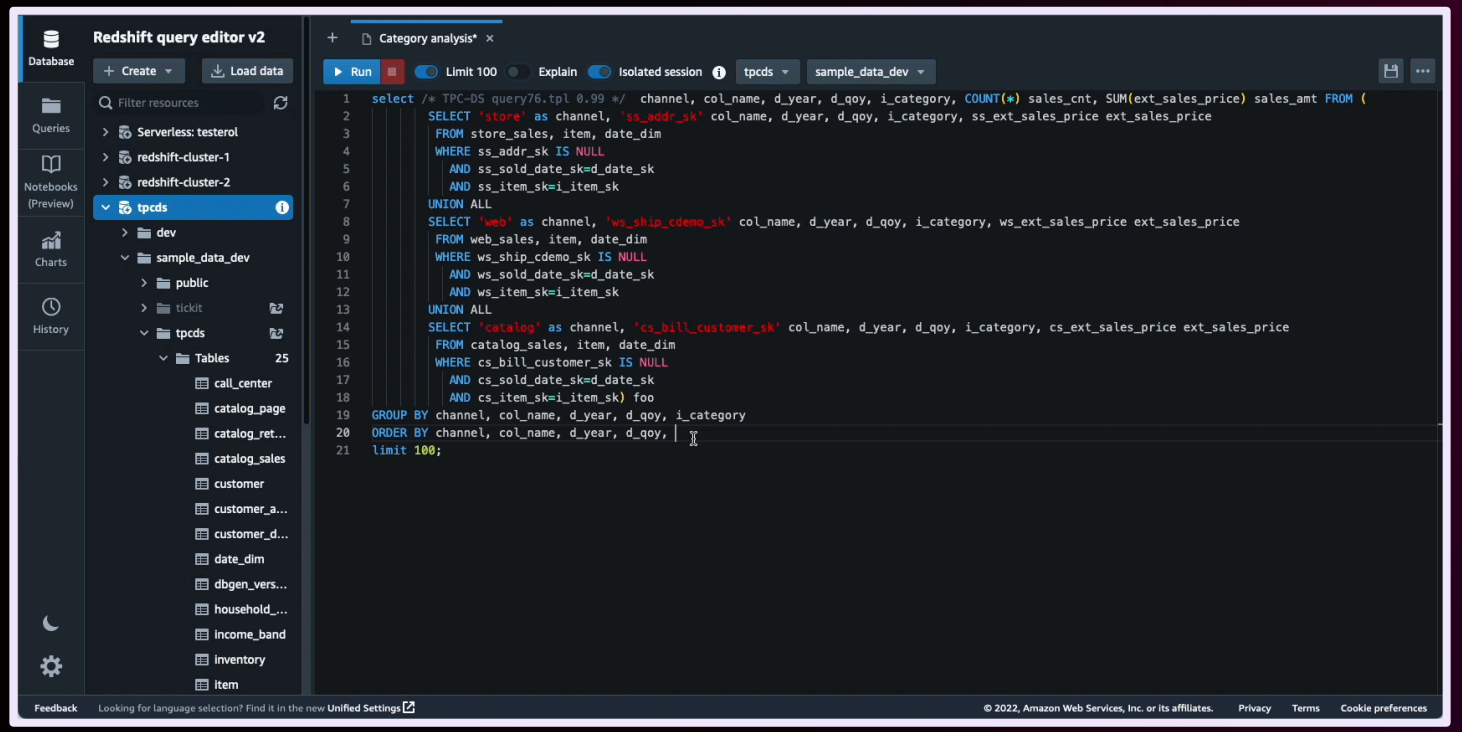
Hub-spoke architecture with serverless to quickly & easily start new DWs

Data sharing to share data across teams, vendors, and other systems

$300K savings annually with Amazon Redshift Serverless & moving manual snapshots

Amazon Redshift Query Editor V2

FREE WEB-BASED TOOL FOR DATA EXPLORATION AND ANALYSIS USING SQL



Browse, create schema & tables, load data, write SQL queries & stored procedures, visualize query results with charts, and track query changes

NEW! Improved collaboration with Notebooks to author, organize, and annotate queries

Allows users to access with Identity Provider (IdP) credentials

Autonomics in the data warehouse

ML-BASED OPTIMIZATIONS TO GET STARTED EASILY AND GET OPTIMIZED PERFORMANCE QUICKLY

"By adopting Auto WLM, our Amazon Redshift cluster throughput increased by **at least 15%** on the same hardware footprint. Our average concurrency **increased by 20%,** allowing approximately 15,000 more queries per week now. All this with marginal impact to the rest of the query buckets or customers."

—Alex Ignatius, Director, Electronic Arts

Automated physical data distribution and schema design

Automatic workload management for peak performance on critical workloads

System & query optimization with MVs, Amazon Redshift Advisor

ATO: automatic sort & distribution keys

Auto analyze, vacuum delete, column encoding

Auto workload management

ATO: smart defaults

Auto MVs, auto refresh, & query rewrite

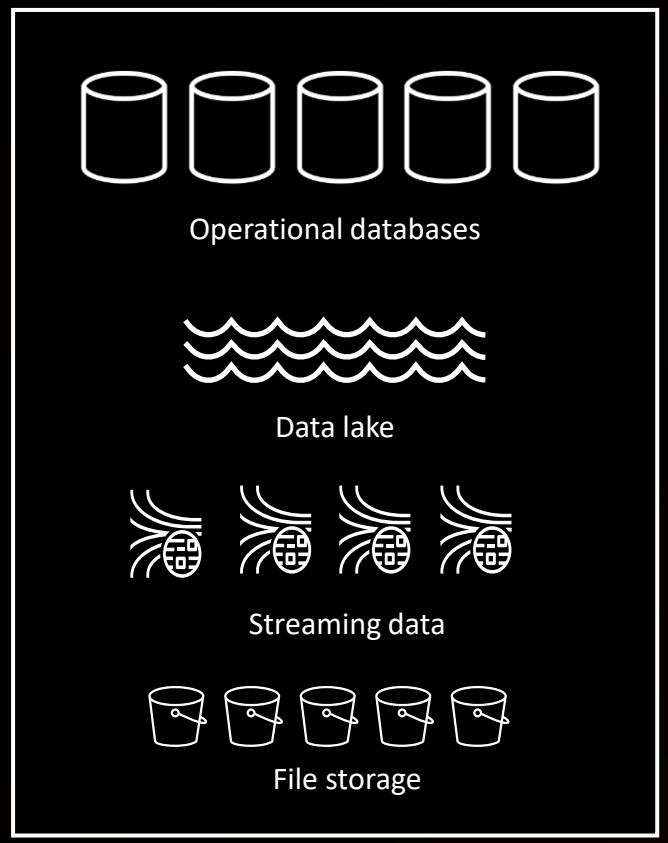
Amazon Redshift Advisor

Easy data ingestion

Ingesting data into analytics systems is complex

MULTIPLE SOURCE SYSTEMS REQUIRE SEPARATE & COMPLEX MANUAL DATA PIPELINES

Data Sources



Manual data pipelines

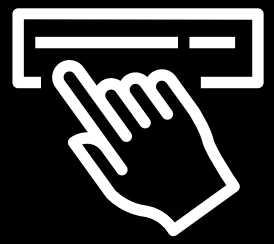
X **Expensive and cumbersome** to build and maintain ETL jobs

X **Complex reconstruction** of the data especially with schema changes

X **Incomplete, inconsistent, and stale views** of data, limiting insights

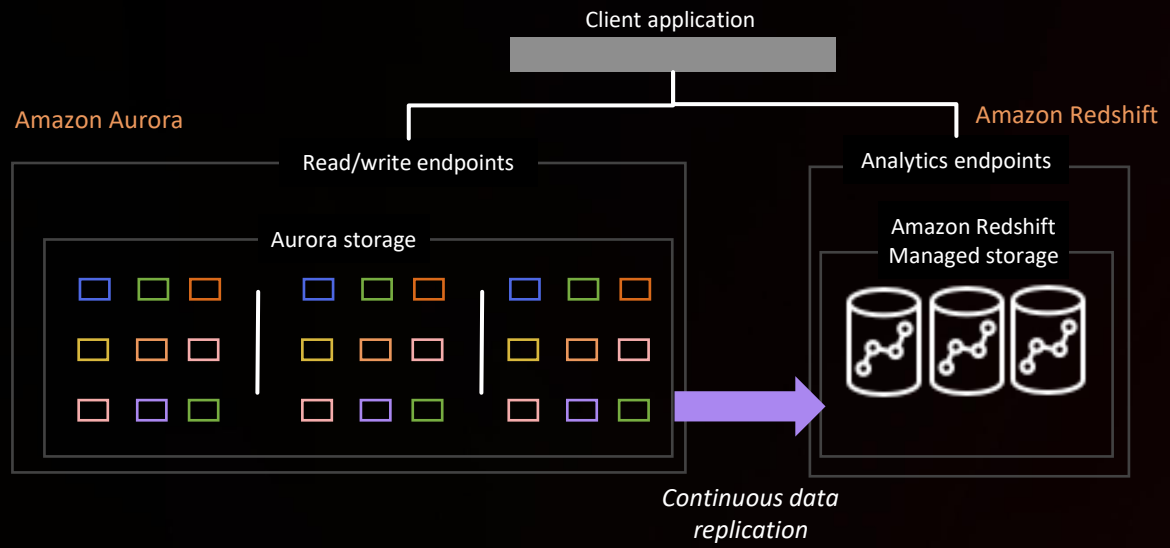
Analytics

NEW



NEW [PREVIEW]

Amazon Aurora zero-ETL integration with Amazon Redshift



No need for customers to build and maintain complex ETL pipelines

Run near real-time analytics and machine learning on petabytes of transactional data from Amazon Aurora

Derive insights using advanced analytics in Amazon Redshift from data consolidated from multiple Amazon Aurora databases

NEW

NEW [PREVIEW]

Support for auto-copy from Amazon S3

Simplified & automated file ingestion from Amazon S3 into Amazon Redshift

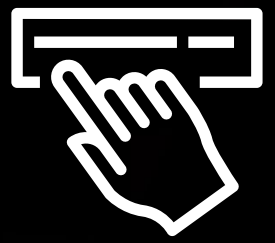
Simple, low-code data ingestion

Avoid re-ingestion and manual tracking of loaded files

Easily convert your existing COPY statements into automatic ingestion jobs

Automatic ingestion of new data from Amazon S3 based on user defined configurations

NEW



NEW [GENERAL AVAILABILITY]

Amazon Redshift streaming ingestion support

Ingest streaming data into your data warehouse for real-time analytics

Directly ingest streaming data into your data warehouse from Amazon Kinesis Data Streams (KDS) and Amazon Managed Streaming for Apache Kafka (Amazon MSK) without staging in Amazon S3

Perform rich analytics using familiar SQL, and easily create and manage ELT pipelines

Process large volumes of streaming data from multiple sources to derive insights in seconds

Adobe

LiveMe

NEW

NEW [GENERAL AVAILABILITY]

Informatica Data Loader integration

Quickly build data pipelines with seamless integration between Informatica Data Loader & Amazon Redshift

30+ sources available

Amazon Redshift

Simply pick Informatica Data Loader from the Amazon Redshift console navigation menu

Run high-speed, high-volume data loading

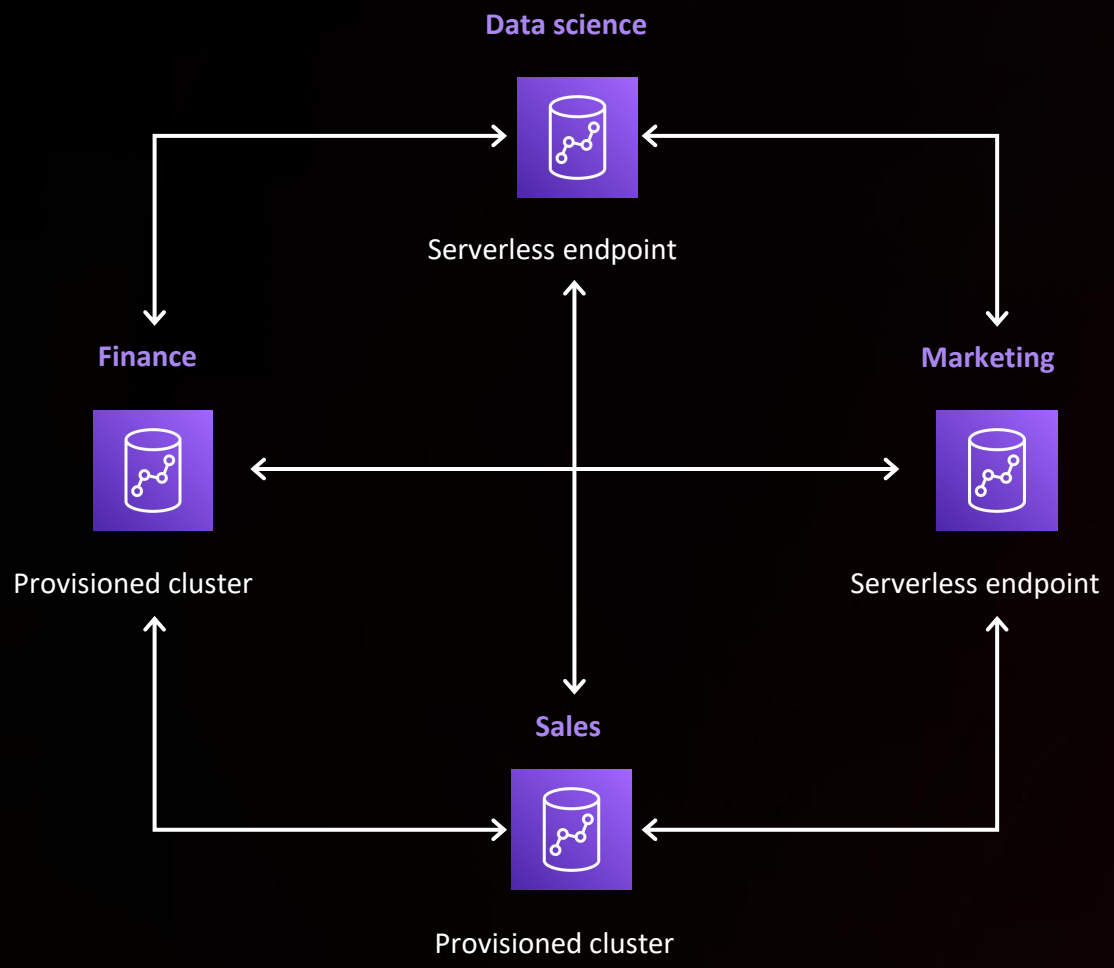
Load data of practically any type

Move your data in minutes

Data sharing & collaboration

Activate data sharing and collaboration

WITH AMAZON REDSHIFT DATA SHARING



Instant, secure & live data access without copies

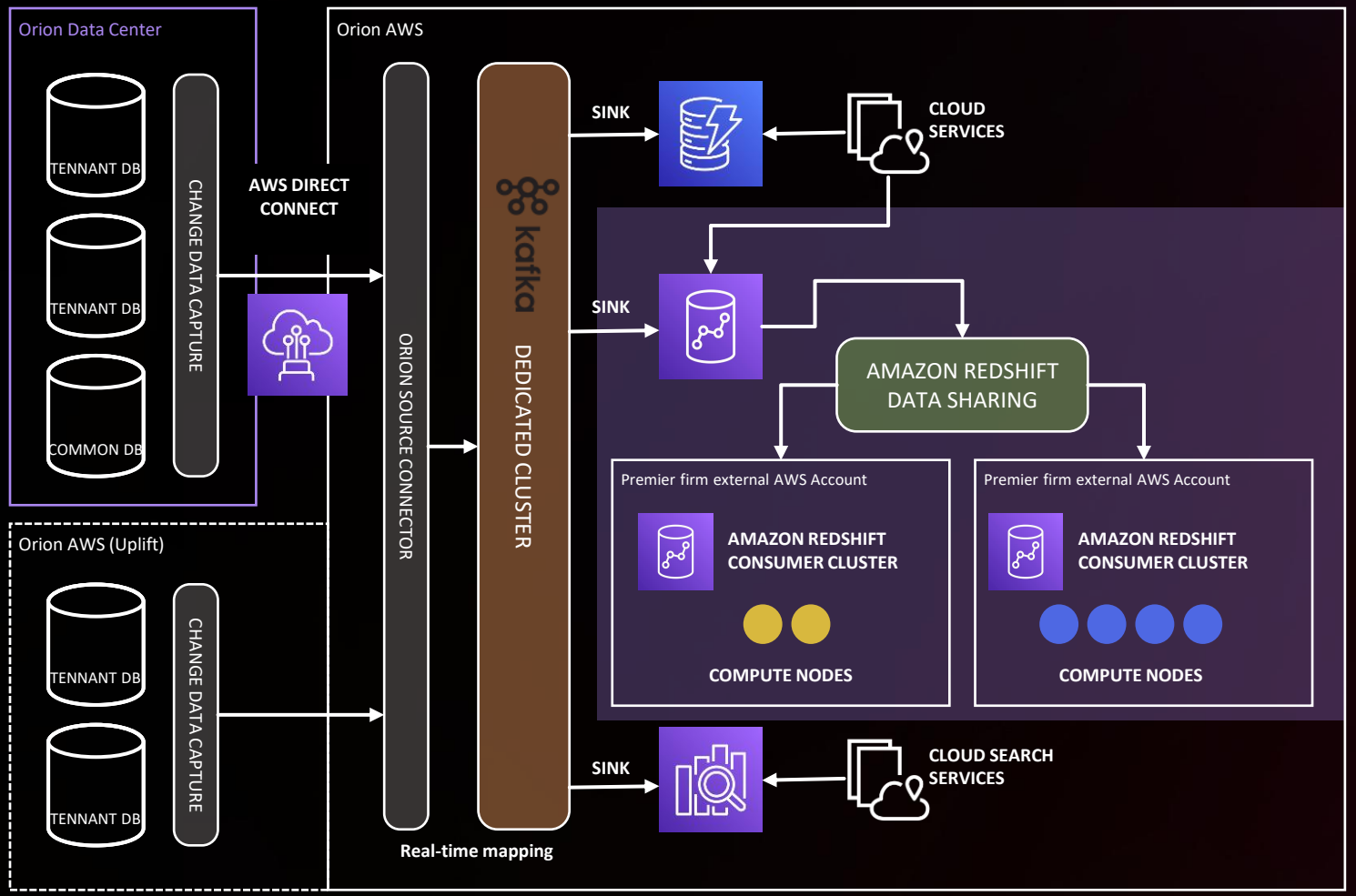
Transactional consistency across data

Workload independence and chargeback

Cross-account and cross-Region data sharing

Data marketplaces with AWS Data Exchange integration

Orion delivers real-time data-as-a-service



ORION

Source events from 2,500+ SQL Server databases both on premises and AWS

Purpose-built CDC, source processors, and Kafka connectors

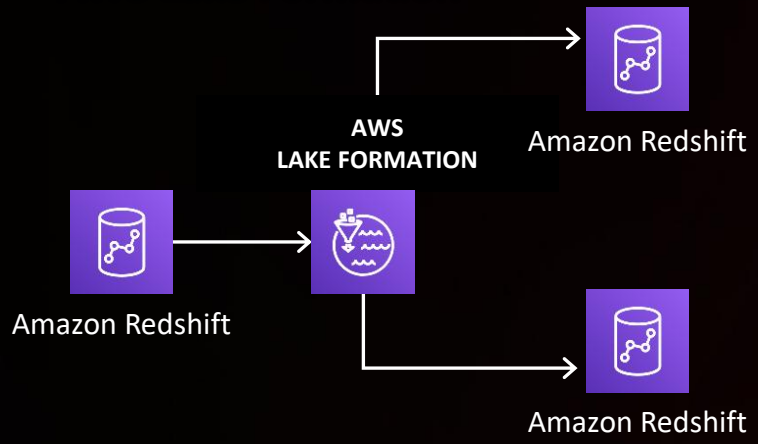
Real-time data sharing & collaboration through Amazon Redshift Data Sharing

NEW

NEW [PREVIEW]

Data sharing access control with AWS Lake Formation

Centrally manage data sharing with AWS Lake Formation



Centrally manage granular access to data across all consuming data services

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

No manual scripting or complex querying

Define policies once and enforce those consistently for multiple consumers

Cloudifying our Data Stack at Goldman Sachs

via AWS, Legend, and GS Financial Cloud for Data

Neema Raphael (he/him)

Chief Data Officer
Goldman Sachs

The Rundown

1 Who // What // Why

Introduction to Goldman Sachs and Data Engineering

2 Problems, What Problems?

SQL ocean + lack of data governance = Paralysis!

3 #FTW

AWS + Legend + GS Financial Cloud For Data = WIN!

4 A Little Deep Dive

How we provide scale, precision, self-service, and rigor

What does Goldman Sachs do?

- A leading global investment bank and financial services company
- Customers include institutions, corporations, consumers
- Increasingly tech forward platforms and engineering-led businesses
- Data is the lifeblood of everything we do

$2T+
Assets Under Supervision

12,000+
Engineers

~100
Locations

Goldman Sachs

Lighting Round

Me => Data Nerd

- ✓ 19 years at GS as a strat, software engineer, data engineer
- ✓ Global head of Data Platforms, Data Engineering, & Chief Data Officer

But first, the data!

- ✓ True real-time (ns/ms), near real time streaming, and batch
- ✓ Hundreds of thousands of datasets also in many forms - time-series, relational, hierarchical, graph
- ✓ Information has to be “right” (e.g., in order, accurate) not just directionally/statistically correct

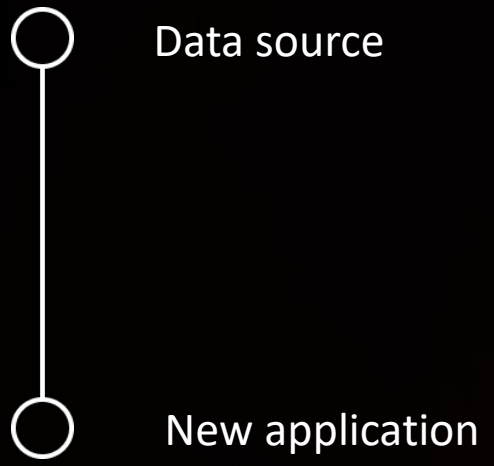
How we roll

- ✓ Open Source Legend Platform (github.com/finos/legend)
- ✓ GS Financial Cloud for Data (developer.gs.com/discover/data)
- ✓ AWS – AWS Data Exchange, Amazon S3, AWS Glue, Amazon Redshift ...

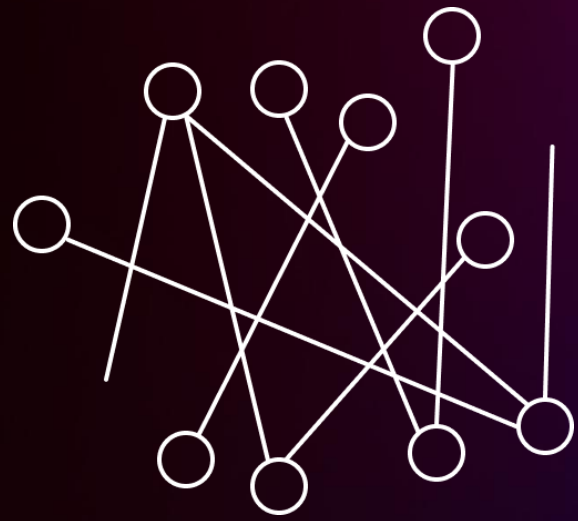
Part 1:
If You Build It ...

A Costly Mistake (Application Centric)

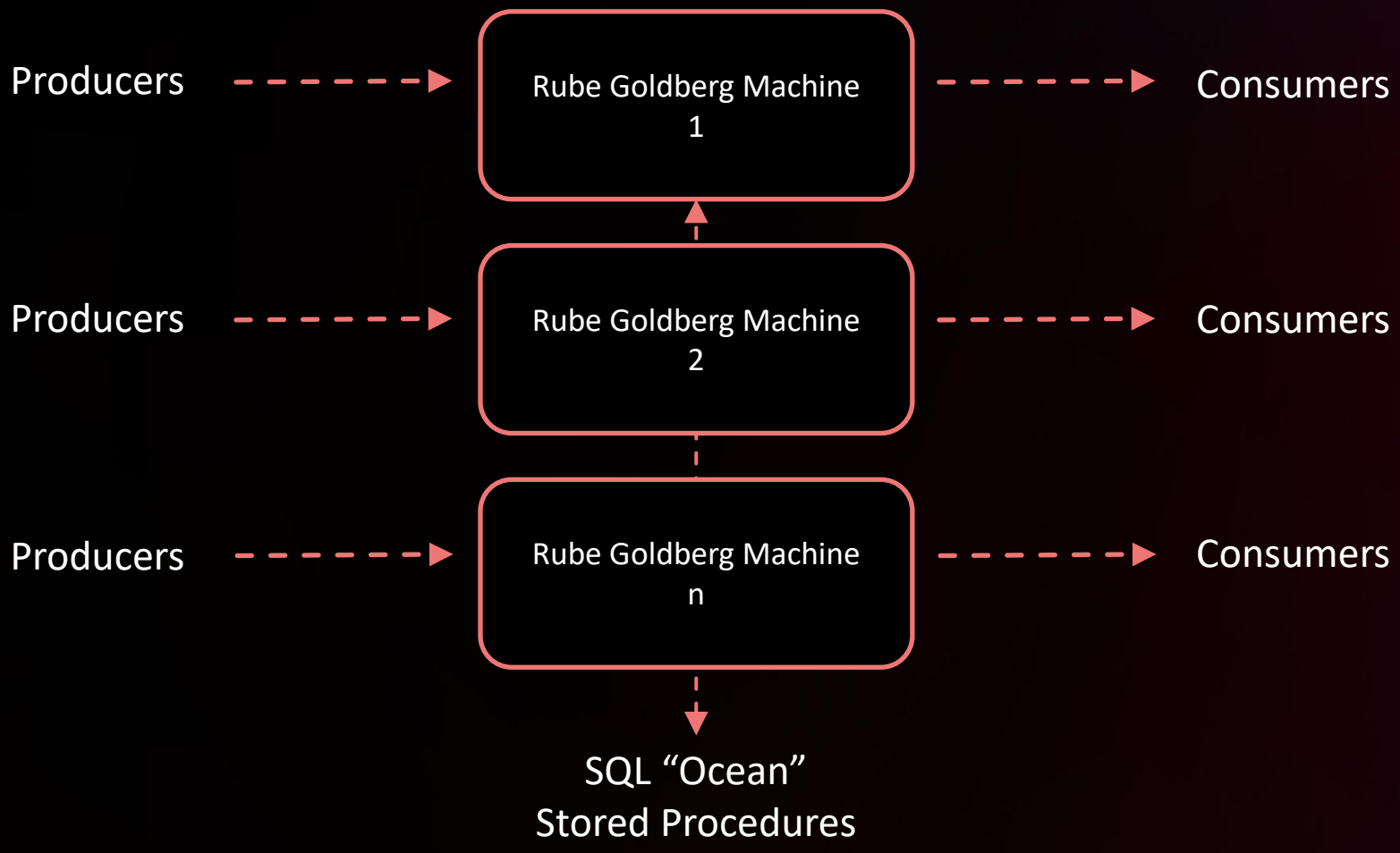
“I’ll just onboard this new feed real fast” – Eager, entrepreneurial engineer



Some time later. . .



The (Data) Chase



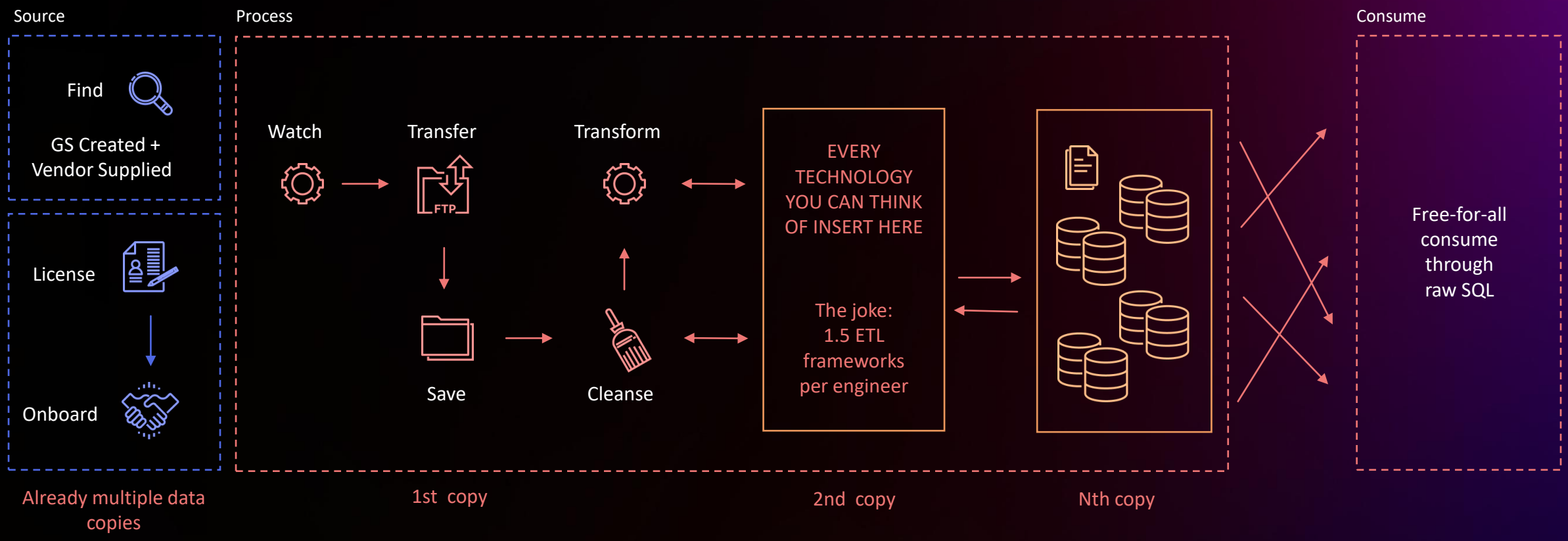
Months
to onboard new data

1+ MM
ungoverned SQL queries

1000s
Consumers

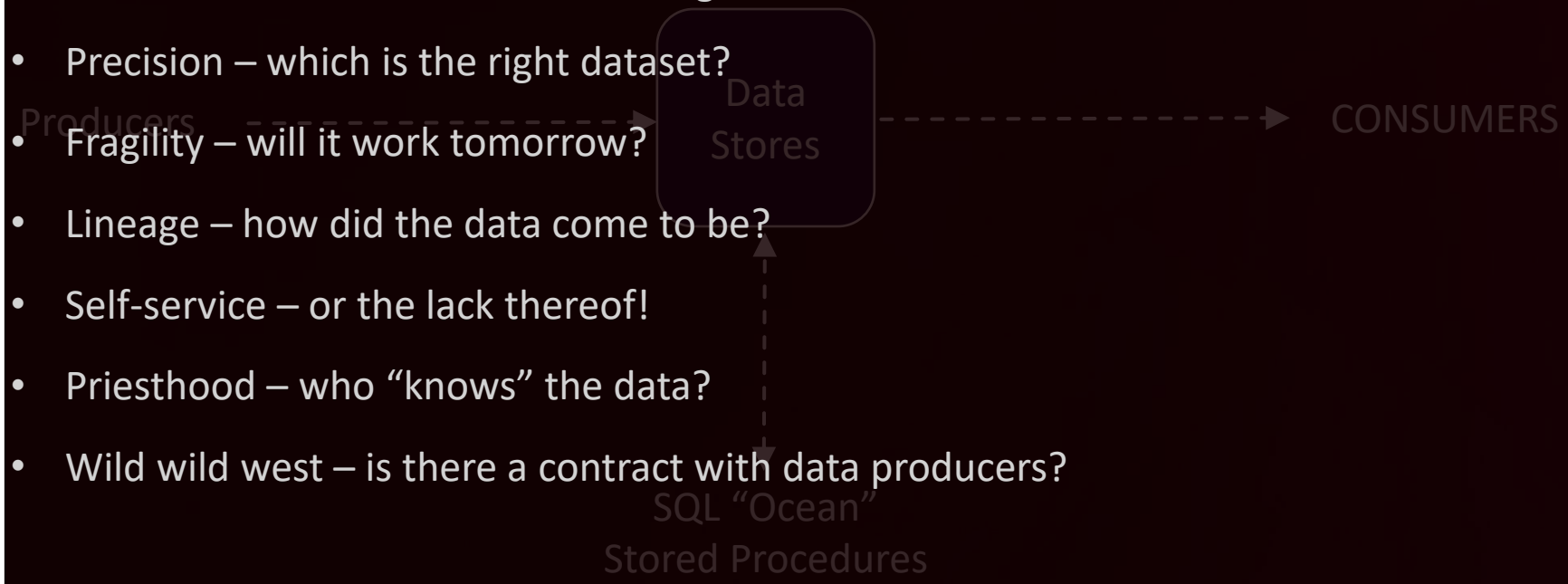
Thousands
complex data stores

A Day in the Life of … BEFORE!



Questions We Asked Ourselves

- CI/CD – can I trust what I am running?
- Precision – which is the right dataset?
- Fragility – will it work tomorrow?
- Lineage – how did the data come to be?
- Self-service – or the lack thereof!
- Priesthood – who “knows” the data?
- Wild wild west – is there a contract with data producers?



Months

To scale systems

1 MM

sql queries

10K

databases

Part 2:
Revenge of the (Data) Nerdz

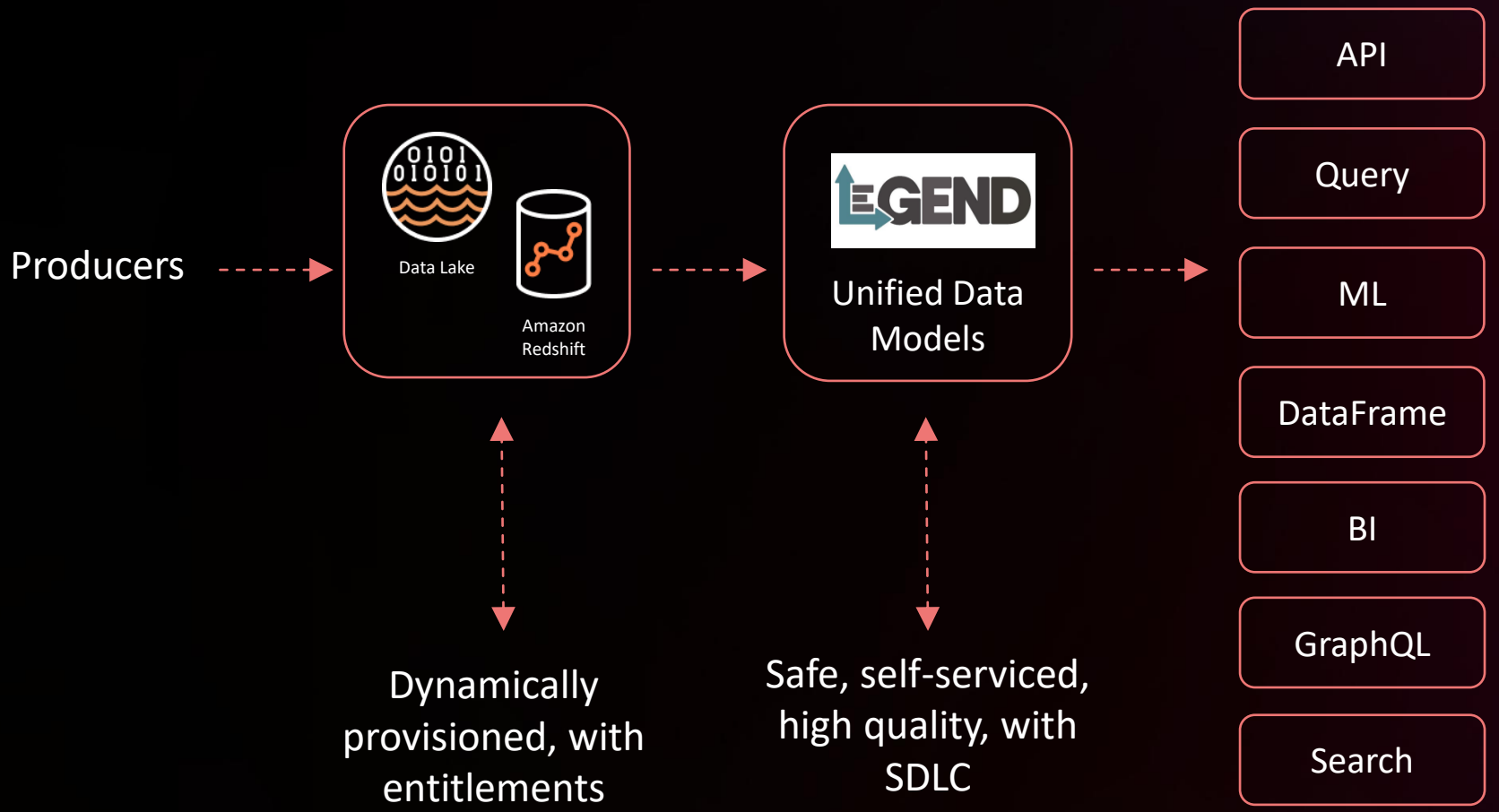
GS [Legend // GS Financial Cloud for Data]

x

AWS [AWS Data Exchange // Amazon S3 // AWS Glue // Amazon Redshift]

Collab

The Data Answer



Minutes
to locate and consume

Thousands
curated data models

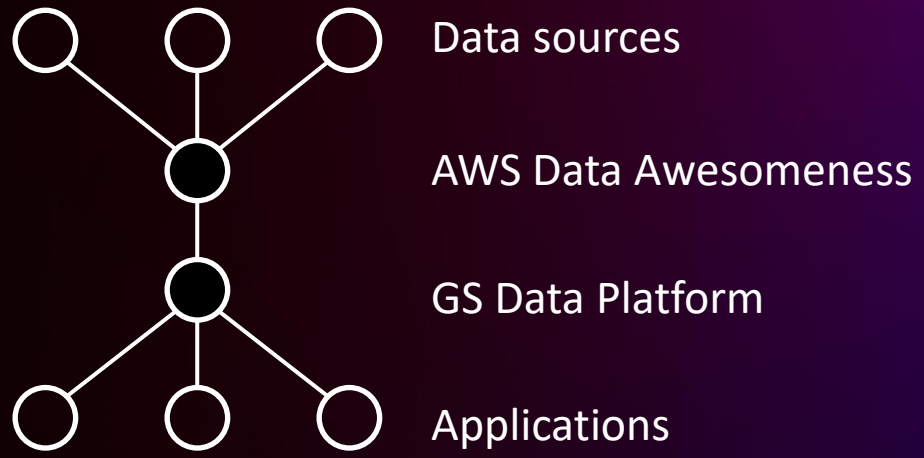
1
Query API

Goldman Sachs

The Right Stuff (Data Centric)

Make it natural and easy to be disciplined up-front

Later …



Legend + Amazon Redshift

API | Query | ML | DataFrame | BI | GraphQL | Search

LEGEND

- ✓ Data models, APIs, discovery, lineage, quality

Amazon Redshift

- ✓ Vertical + horizontal scaling
- ✓ Semi-structured data
- ✓ No devops ETL
- ✓ Serverless!!

Data Lake

- ✓ Managed, scalable, resilient
- ✓ Versioning
- ✓ Publishing APIs
- ✓ End-to-end reproducible

98M+

Production SLO API calls/month

140K

self-service *modeled* query templates

1

Unified data model

Goldman Sachs

aws

Back To The Future

Legend

Modelled + Queryable
Data Mesh

Amazon Redshift Serverless

Look ma, no hands!

Zero ETL

Make DataOps
a non-thing

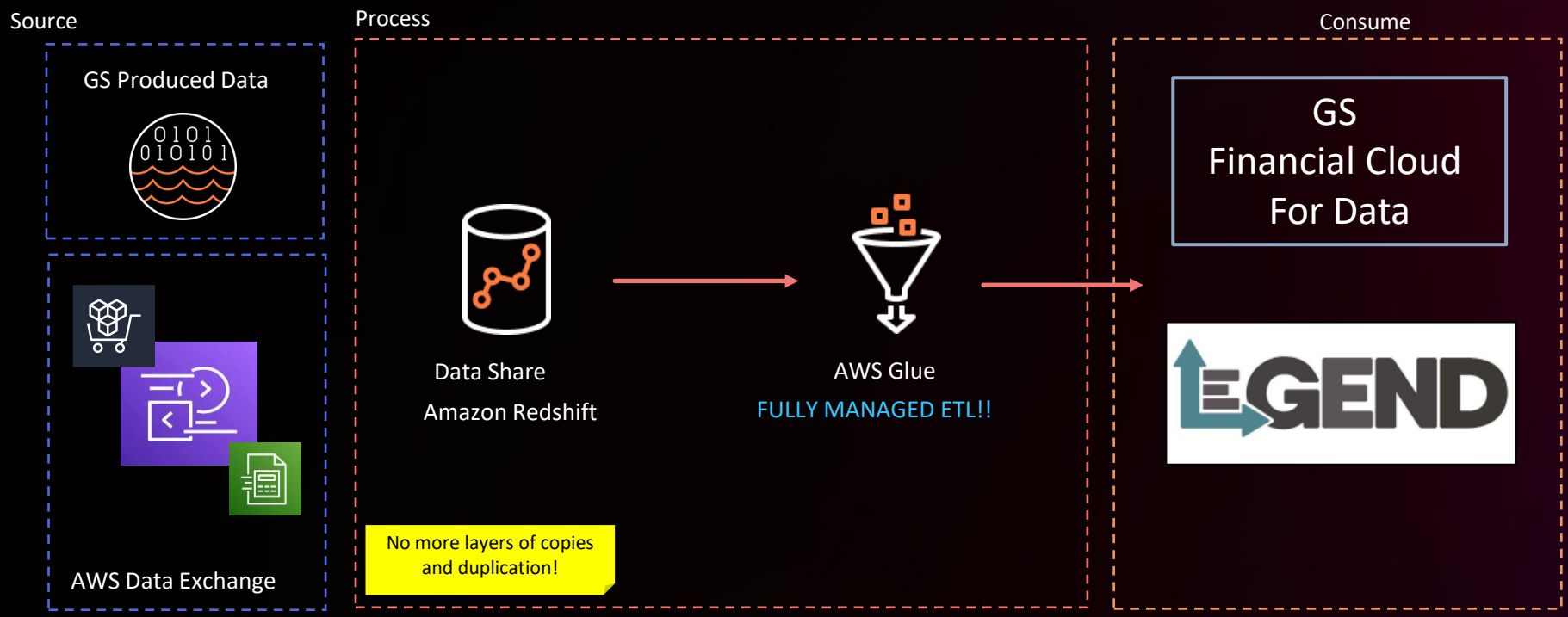
Amazon Redshift ML

Democratize predictive
analytics

Goldman Sachs

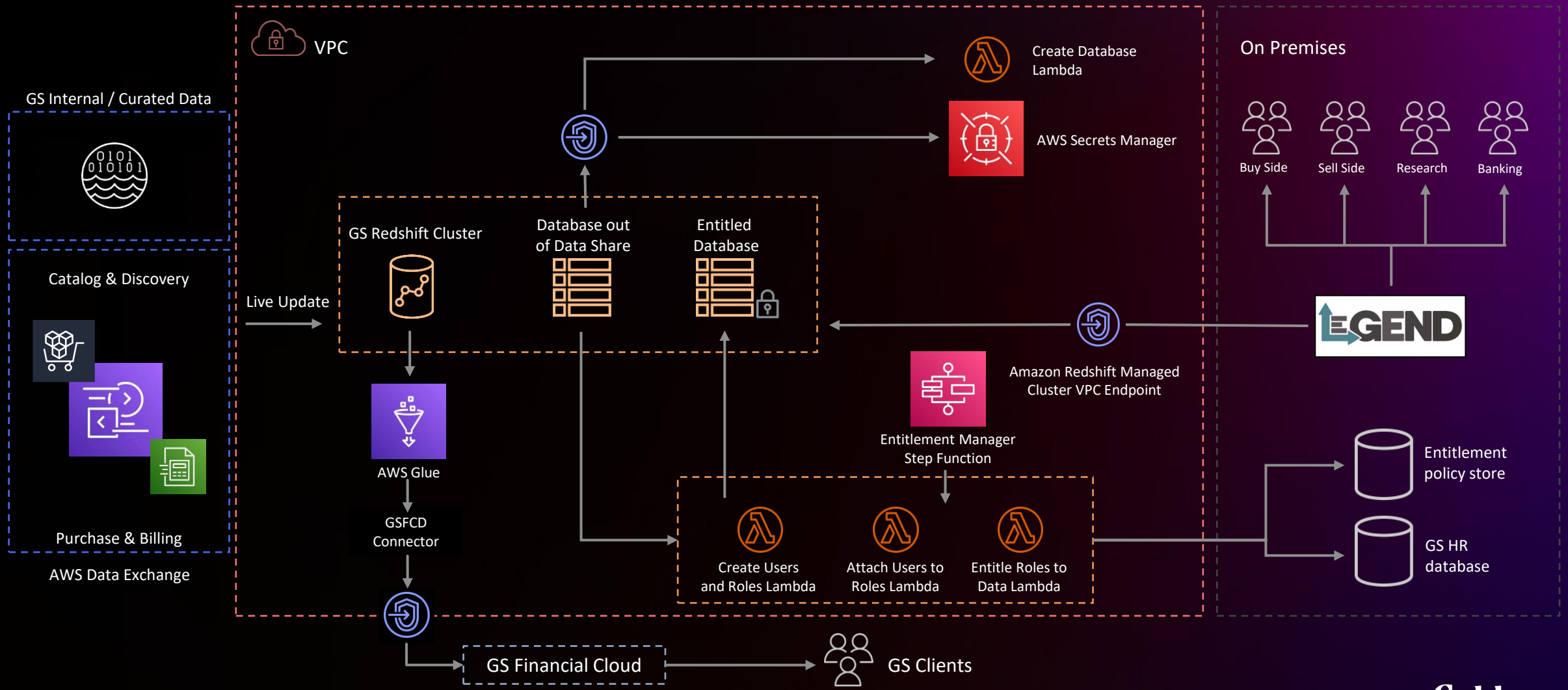
Part 3: #BehindTheScenes

A Day in the Life of … AFTER!

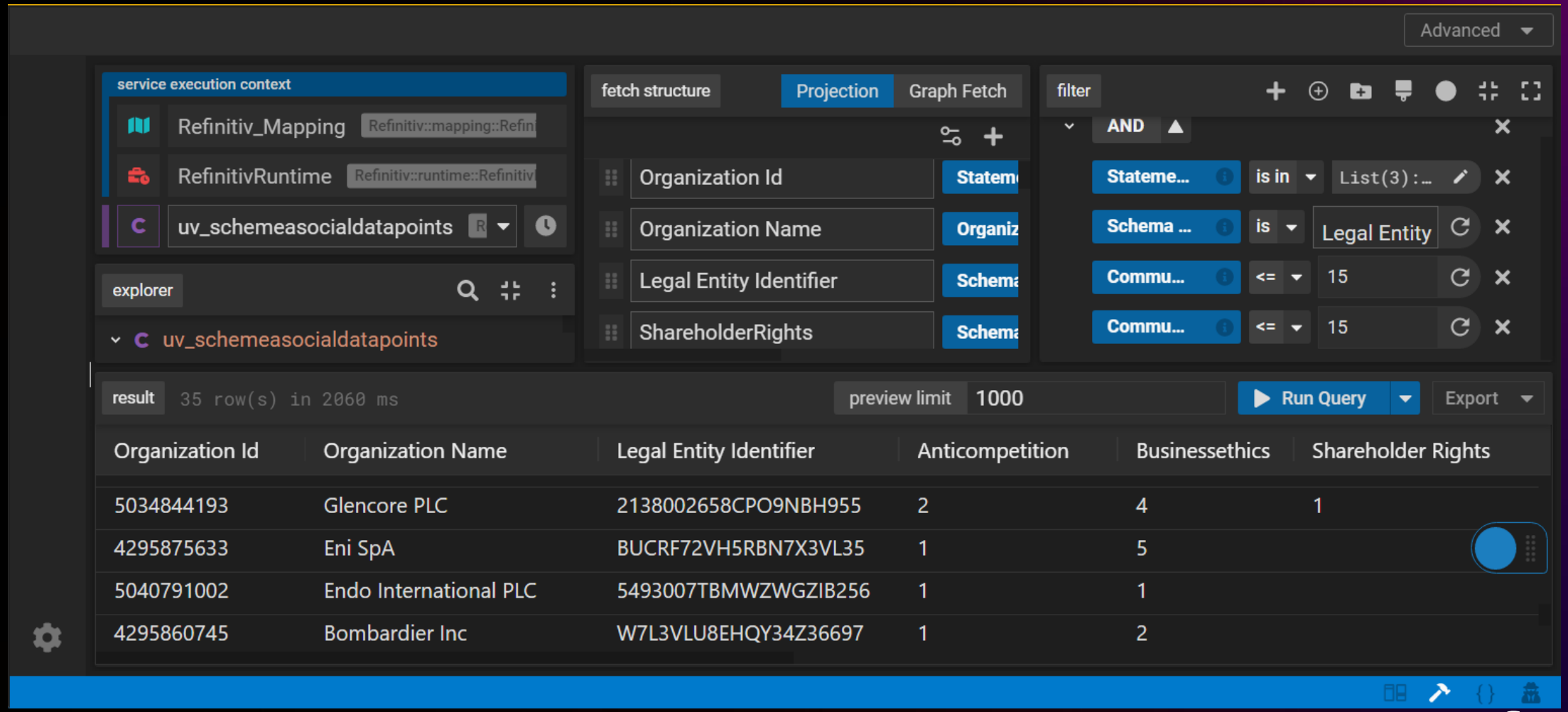


- PlotTool Pro
- API
- Query
- ML
- Frame
- BI
- GraphQL
- Search

Cloud Market Data Platform - Deep Dive



Legend Query – Self Service for All!



Rich analytics with data science & machine learning

Amazon Redshift ML

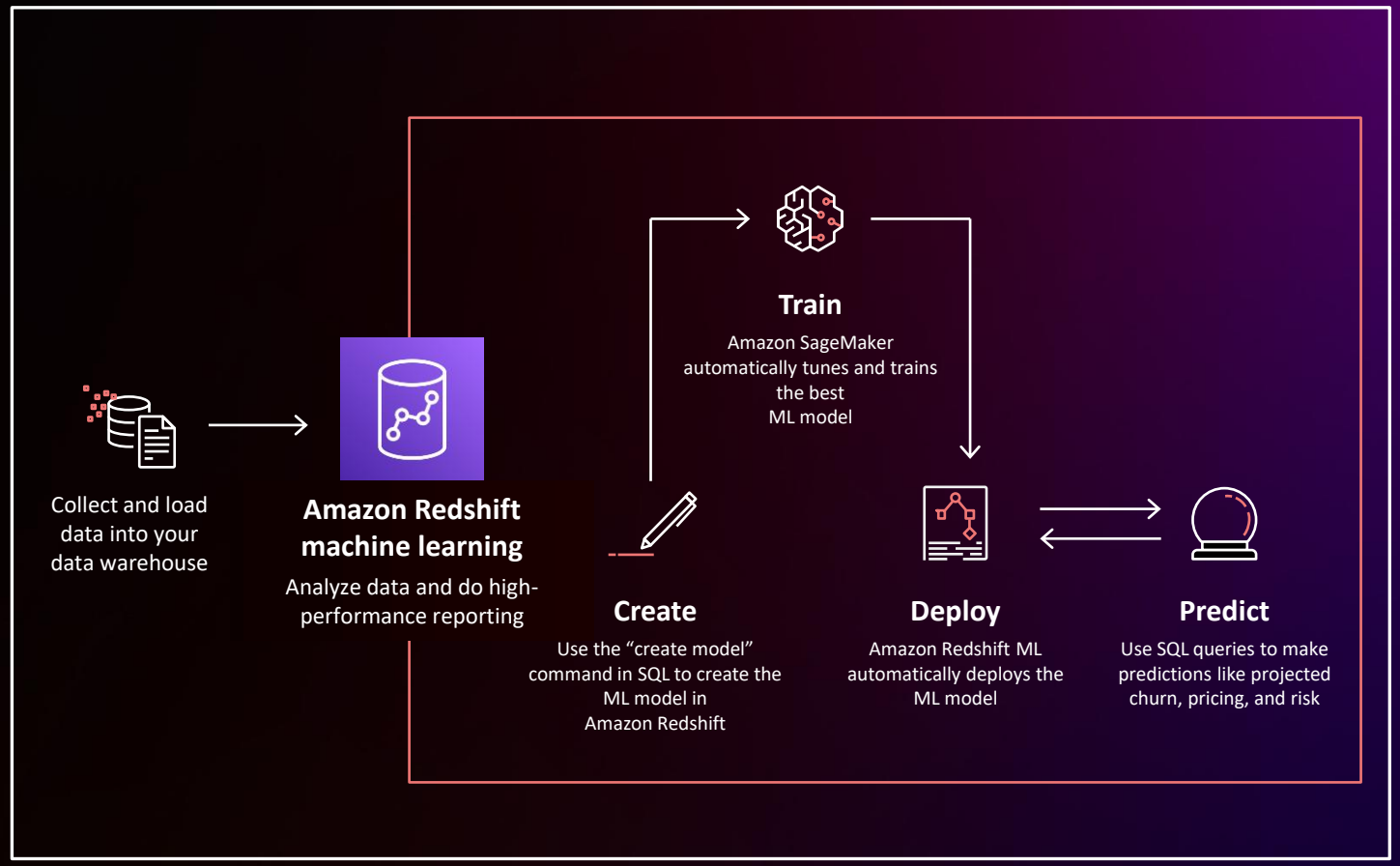
EASILY CREATE AND TRAIN ML MODELS USING SQL QUERIES WITH AMAZON SAGEMAKER

Train and create ML models using SQL

Automatic pre-processing, creation, training, deployment, and inferencing of models

SageMaker models for in-database or remote inference

Supervised and unsupervised trainings



jobcase

MagellanRx MANAGEMENT

aws

Scales ML workflows to support billions of daily events using Amazon Redshift ML

> "With Redshift ML, we have evolved to model architectures that generate a 5%–10% improvement in revenue and member engagement rates across several different email template types, with no increase in inference costs."

Mike Griffin

EVP Optimization & Analytics

NEW

NEW [GENERAL AVAILABILITY]

Amazon Redshift integration for Apache Spark

Simplify and speed up Apache Spark applications accessing Amazon Redshift data from AWS analytics services

Author Apache Spark applications using Java, Python, Scala, with access to rich, curated data in your data warehouse

No manual setup and maintenance of uncertified versions of Spark-Redshift open-source connectors

Improved performance with only relevant data moved from Amazon Redshift to consuming applications

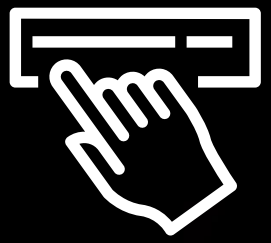
Improved security with IAM-based credentials

Security and reliability for mission-critical analytics

Built-in security and compliance

SECURITY AND COMPLIANCE FEATURES WITH NO EXTRA COSTS WITH AMAZON REDSHIFT

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



NEW [PREVIEW]

Amazon Redshift supports dynamic data masking

baffle

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

Ameesh Divatia

CEO & Co-Founder, Baffle.io

NEW

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

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

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

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

aws

NEW

NEW [PREVIEW]

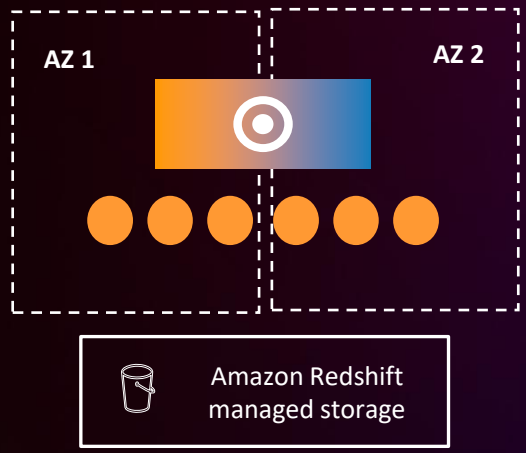
Amazon Redshift Multi-AZ

Highly resilient data warehouse

Auto-failover with zero data loss and no manual intervention

Easy management through a single endpoint

Workload processing across AZs



NEW

NEW [General Availability]

AWS Backup integration

Simplify data protection for Amazon Redshift resources through seamless integration between AWS Backup & Amazon Redshift

Centralize data protection for all Amazon Redshift resources

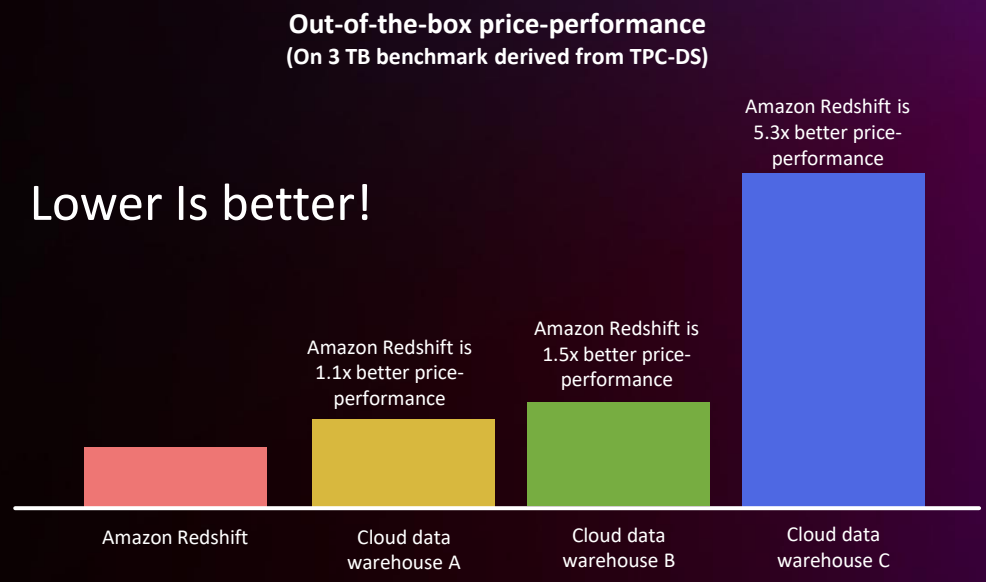
Automate backup scheduling and retention by configuring backup plans

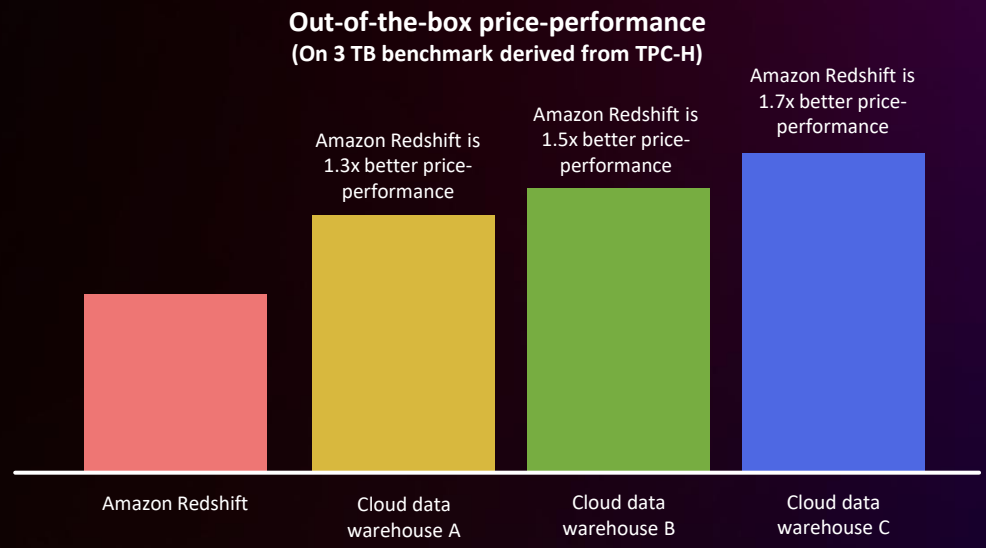
Restore an entire cluster or a table to a desired point in time from backups

Best price-performance analytics

Up to 5X better price-performance

TPC-DS and TPC-H out-of-the-box 3 TB benchmark

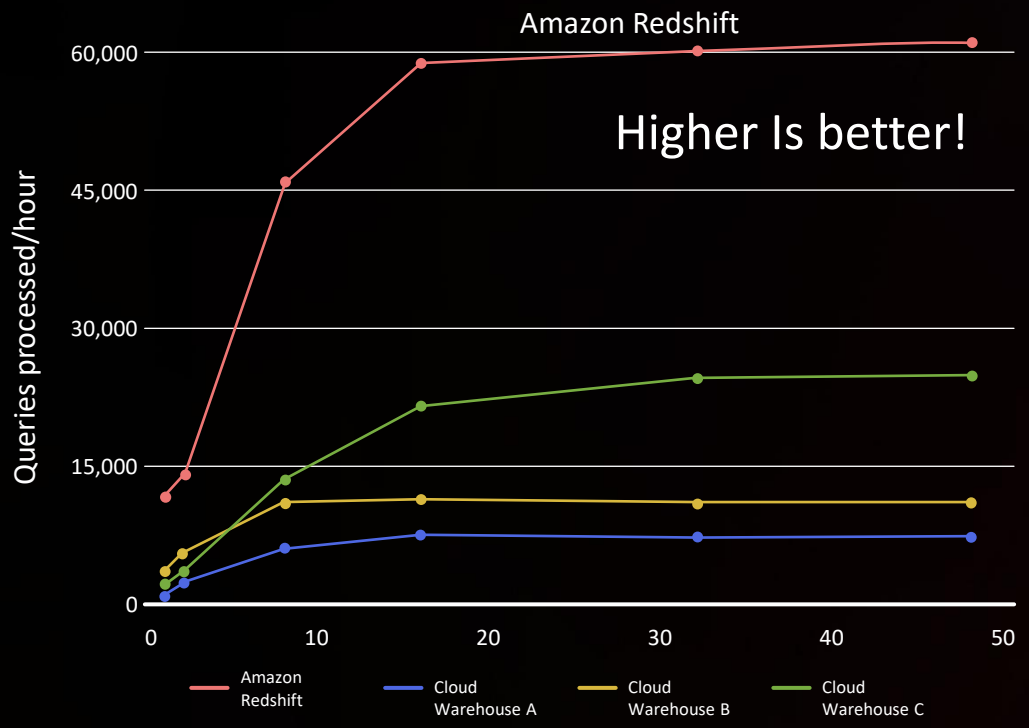




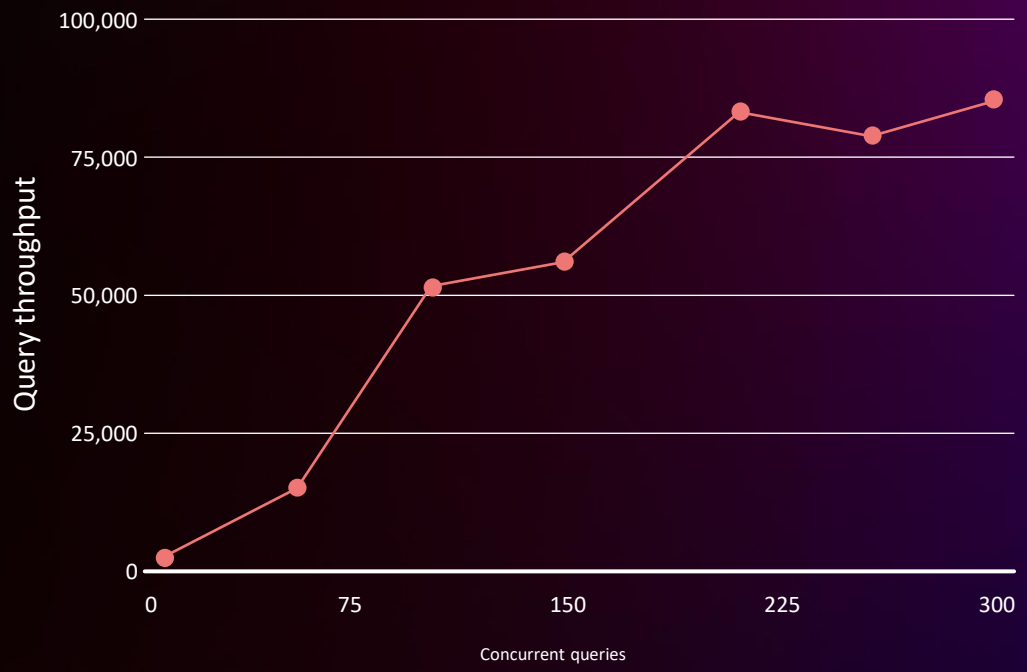
Up to 7x better price-performance for BI dashboards

10 GB AND 100 GB BENCHMARK DERIVED FROM TPC-DS

Query throughput for short queries
(On 10 GB benchmark derived from TPC-DS)



Scaling Amazon Redshift with concurrency scaling
(On 100 GB TPC-DS benchmark derived from TPC-DS)



Out-of-the-box performance improvements

CONTINUOUS IMPROVEMENTS FOR BETTER PRICE-PERFORMANCE AT ANY SCALE

Compute

- NEW — Vectorized scans for Amazon Redshift tables
- NEW — Write/commit performance
- NEW — Concurrency scaling writes (GA)

System

- CaaS Region expansion
- Snapshot isolation

Autonomics

- NEW — Auto WLM enhancements
- NEW — ATO enhancements
- NEW — Advisor enhancements

SQL enhancements & migration support

ACCELERATE MIGRATIONS FROM LEGACY DATA WAREHOUSES

SQL Syntax

- PIVOT/UNPIVOT
- NEW MERGE (preview)
- NEW Enhanced Identity support
- Unload to JSON
- NEW CONNECT BY
- NEW GROUPING SETS, ROLLUP, CUBE (all in preview)

Types

- SUPER
- GEOGRAPHY
- VARBYTE

Limits

- 100,000 Tables
  10,000 SPs
- NEW SUPER data type
  16 MB Support (preview)

Tooling & connectivity

- **RSQL -** command line tool with improved control flow and BTEQ support
- SQL Alchemy, Apache Airflow, and QueryBook support
- NEW LAMBDA UDF optimizations

KEY customers moving from legacy data warehouses to Amazon Redshift

LexisNexis · intuit · moderna · TOYOTA · zynga

Get started today

New to Amazon Redshift?

- ✓ IDC study on Business value of Amazon Redshift
- ✓ Explore demos, customer stories, and the latest features on aws.amazon.com/redshift
- ✓ Get trained
- ✓ Ask your account team for a 10-minute demo

Already using Amazon Redshift?

- ✓ Ask your account team for a free Amazon Redshift optimization session
- ✓ Learn more about what's new https://aws.amazon.com/redshift/whats-new/

Thinking of migration?

AWS teams and programs can help you with

Building your business case, creating migration plan

Deciding on budget, scope, etc.

Conduct a proof of concept or architect your data strategy

Get Help

aws.amazon.com/redshift

Thank you!

Neema Raphael
linkedin.com/in/neema-raphael/

Eugene Kawamoto
linkedin.com/in/kawamoto/

Drop a star on GitHub!
https://github.com/finos/legend

Please complete the session survey in the **mobile app**