aws INNOVATE

DATA EDITION

Innovate faster: Reinvent your organization with data

Olivier Klein
Chief Technologist Asia-Pacific

Kris Howard
Head of Developer Relations Asia-Pacific
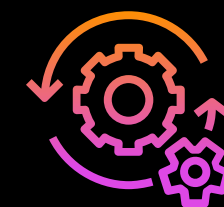
aws

Every type of app

aws

Every type of data

aws

Put data
to work
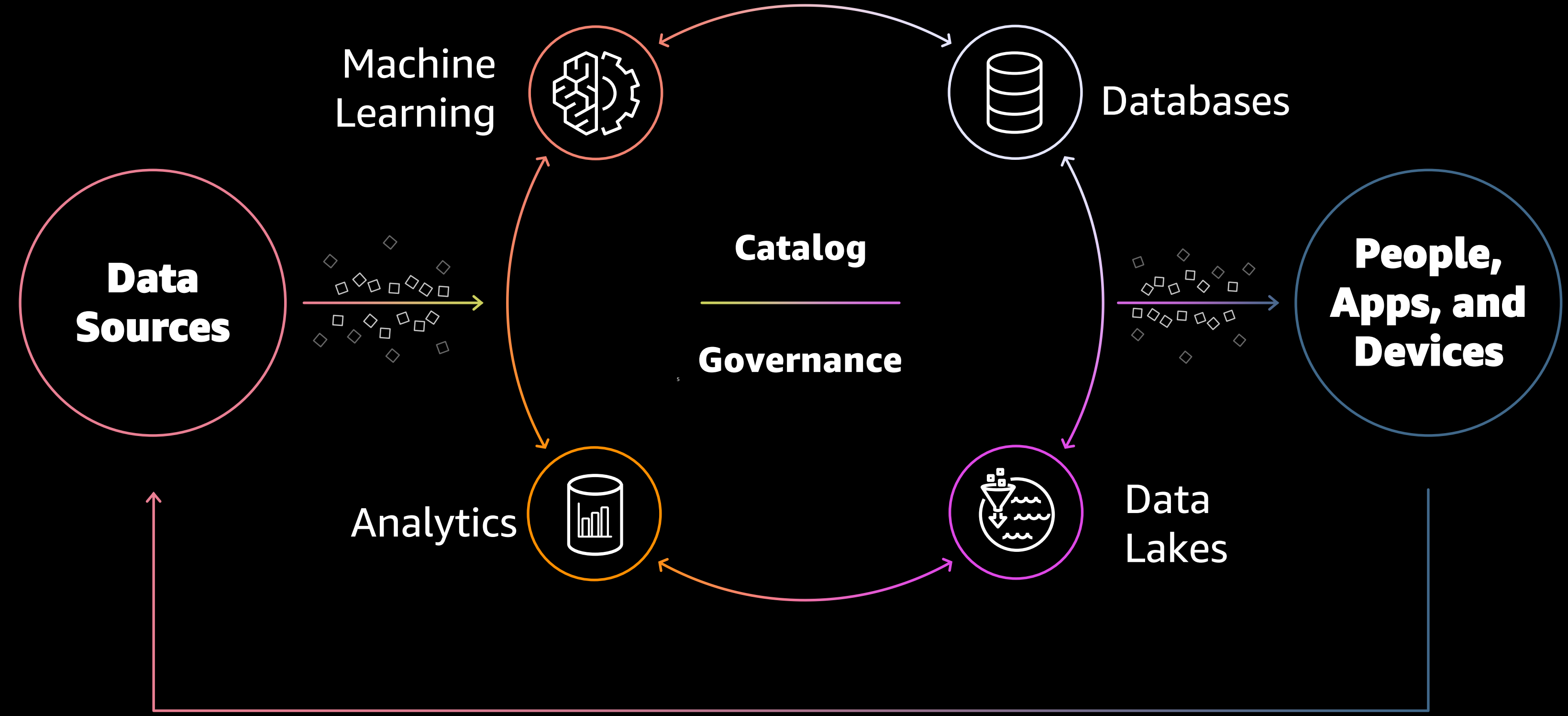
- Make better decisions
- Improve efficiencies
- Respond faster
- Uncover opportunities

aws

Modern data strategy in action

FINra

Ingest **250 billion** records daily using Amazon EMR

Handle **20 TB** of data daily
Finish **90% of jobs in under one minute**

aws

Pinterest

Scale daily log search and analytics from **500 GB to 1.7 TB**

Reduce operational costs by **30%**

Nasdaq

30 billion to 70 billion records a day with no disruption – **peak volume 113 billion during COVID**

Time-to-insight – **up to five hours faster**
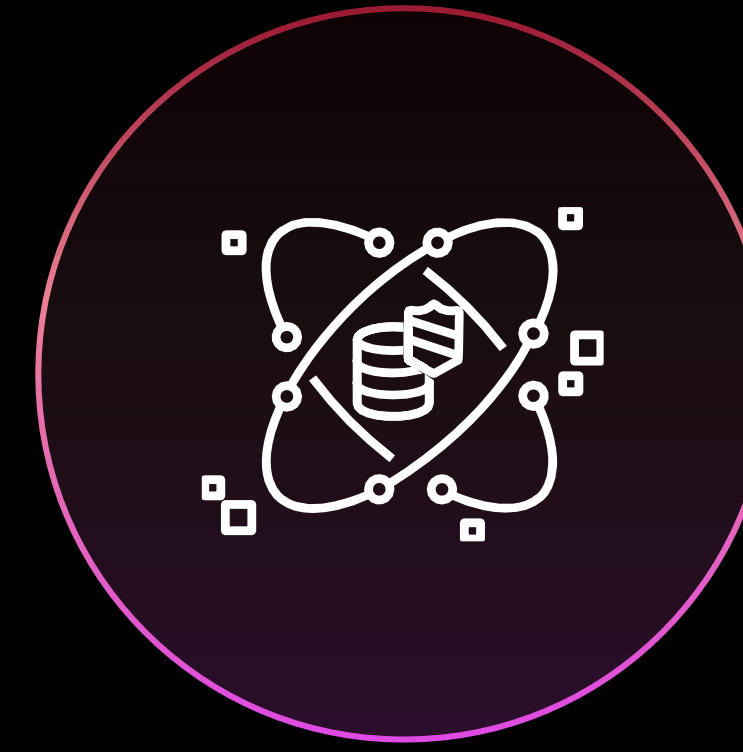
AWS Analytics Pillars

Scalable data
lakes

Purpose-built for
performance and
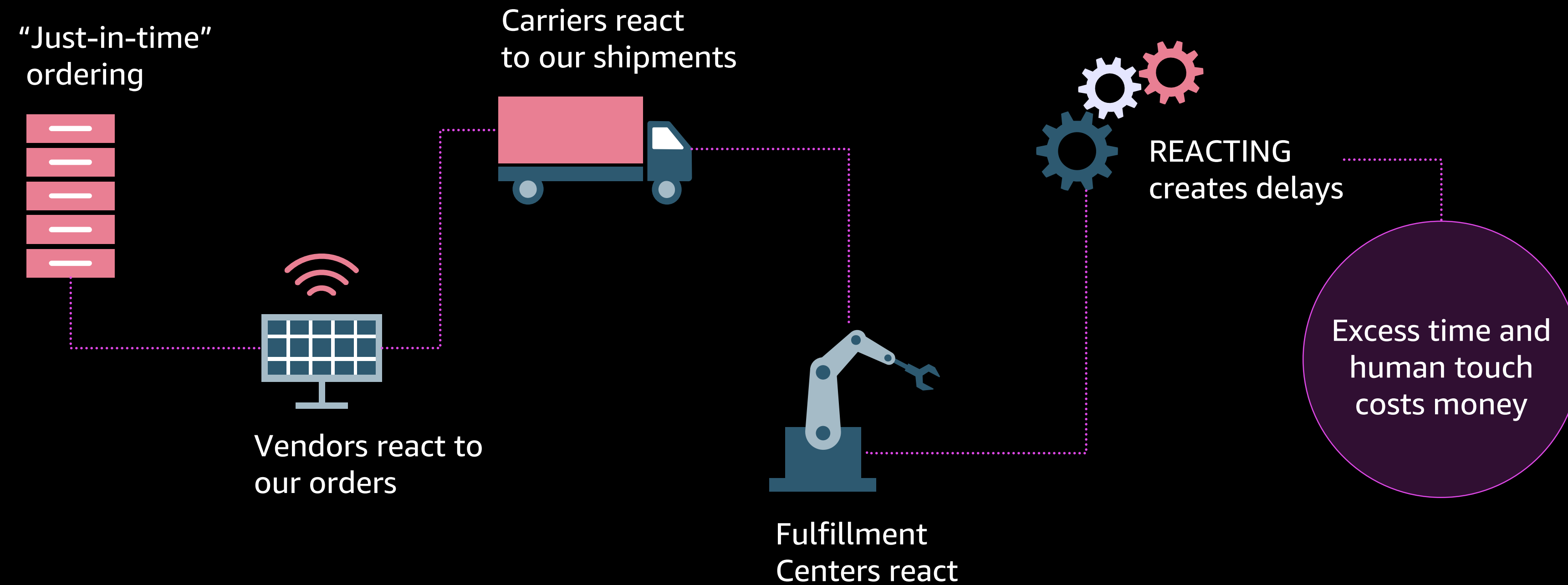cost

Serverless and
easy to use

Unified data
access, security,
and governance

Built-in machine
learning

aws

Driving Business Outcomes through Data

Traditional Supply Chain Flow

"Just-in-time" ordering

Carriers react to our shipments

REACTING creates delays

Vendors react to our orders

Fulfillment Centers react

Excess time and human touch costs money

Amazon's Supply Chain - **Automation End-to-end**

Forecasting   **Capacity**   **Buying**   **Placement**   **Fulfillment**
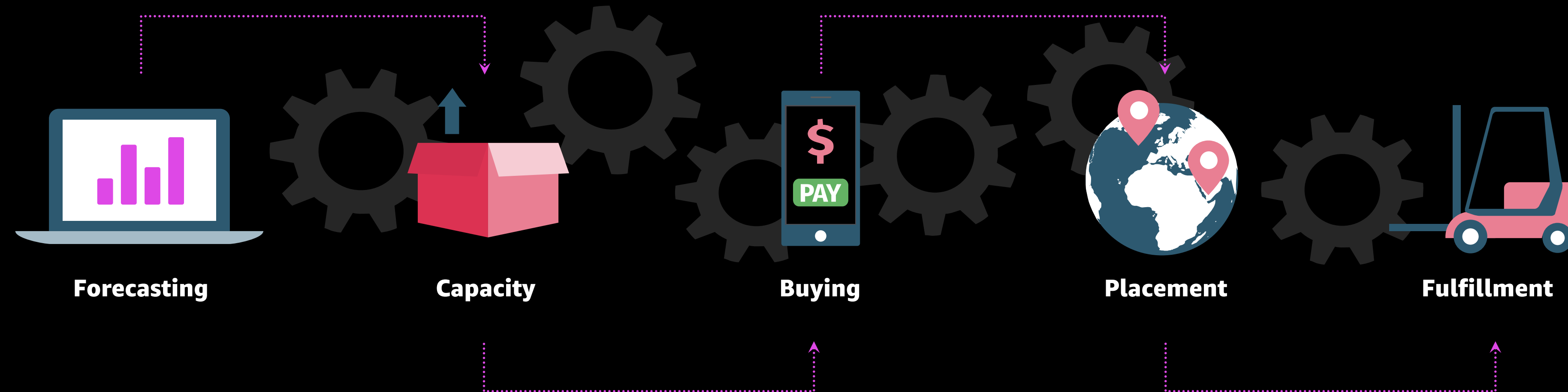
Demand forecasting for over **400MM products daily**

Amazon – **Over 99%** of inventory is ordered automatically

Adjust for regional demand by **redistribution of items** in the network

sayurbox

Keeping costs low – customers can expect to **save 30% over supermarkets**, and farmers **earn up to 20% more for harvest**

Scaling – handle up to **300% spike in demand** while maintaining **24h delivery time for produce**

aws

sportsbet

Migrated corporate data warehouse to **data lake house architecture** powered by **Amazon RedShift and Delta Lake**

Innovation – developer productivity **increase by 40%**

paytm

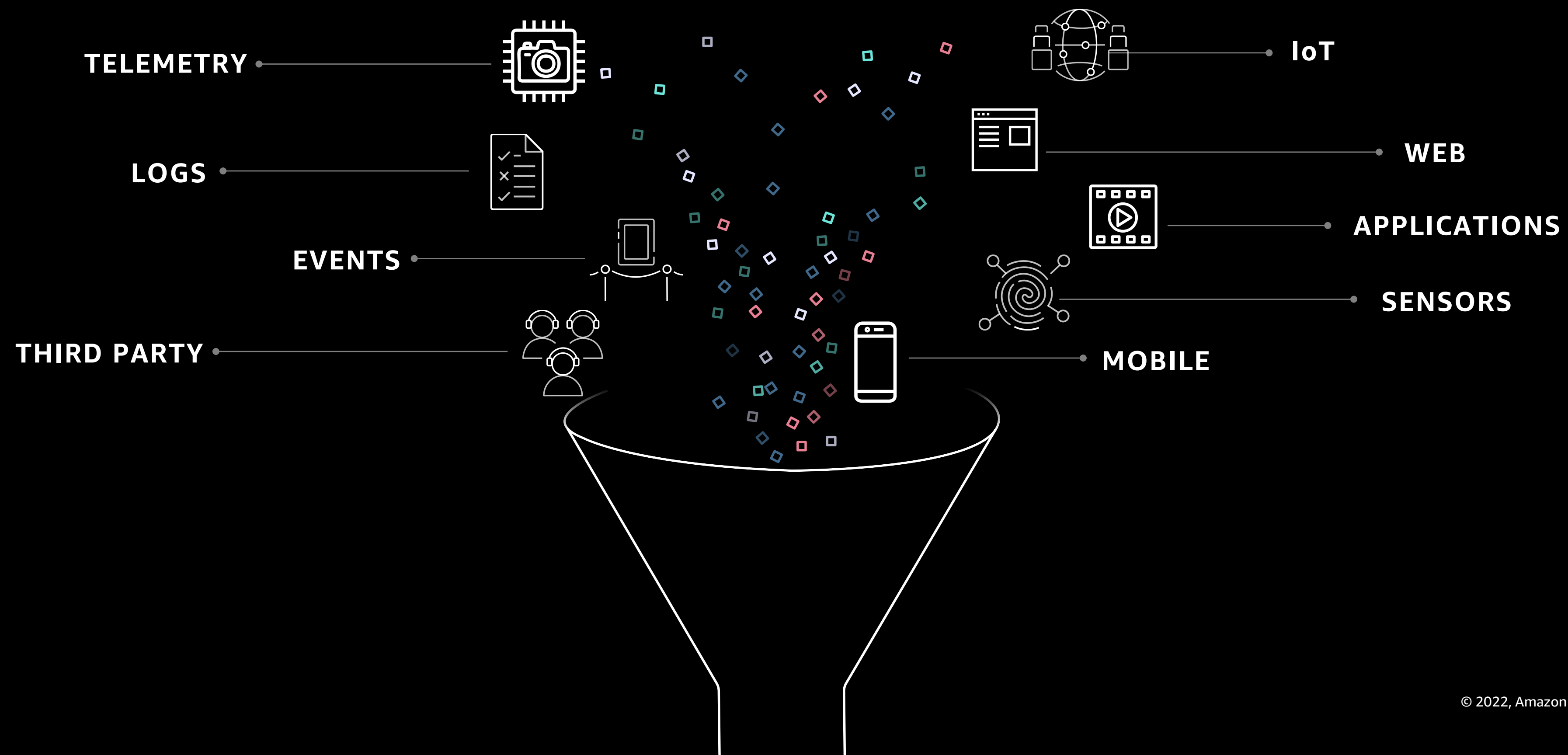Deliver data to business users **30% faster** and at **70% the cost of on-premises solution**

Streamlined data processing for majority workloads **by 98%**

Work backwards from business outcomes
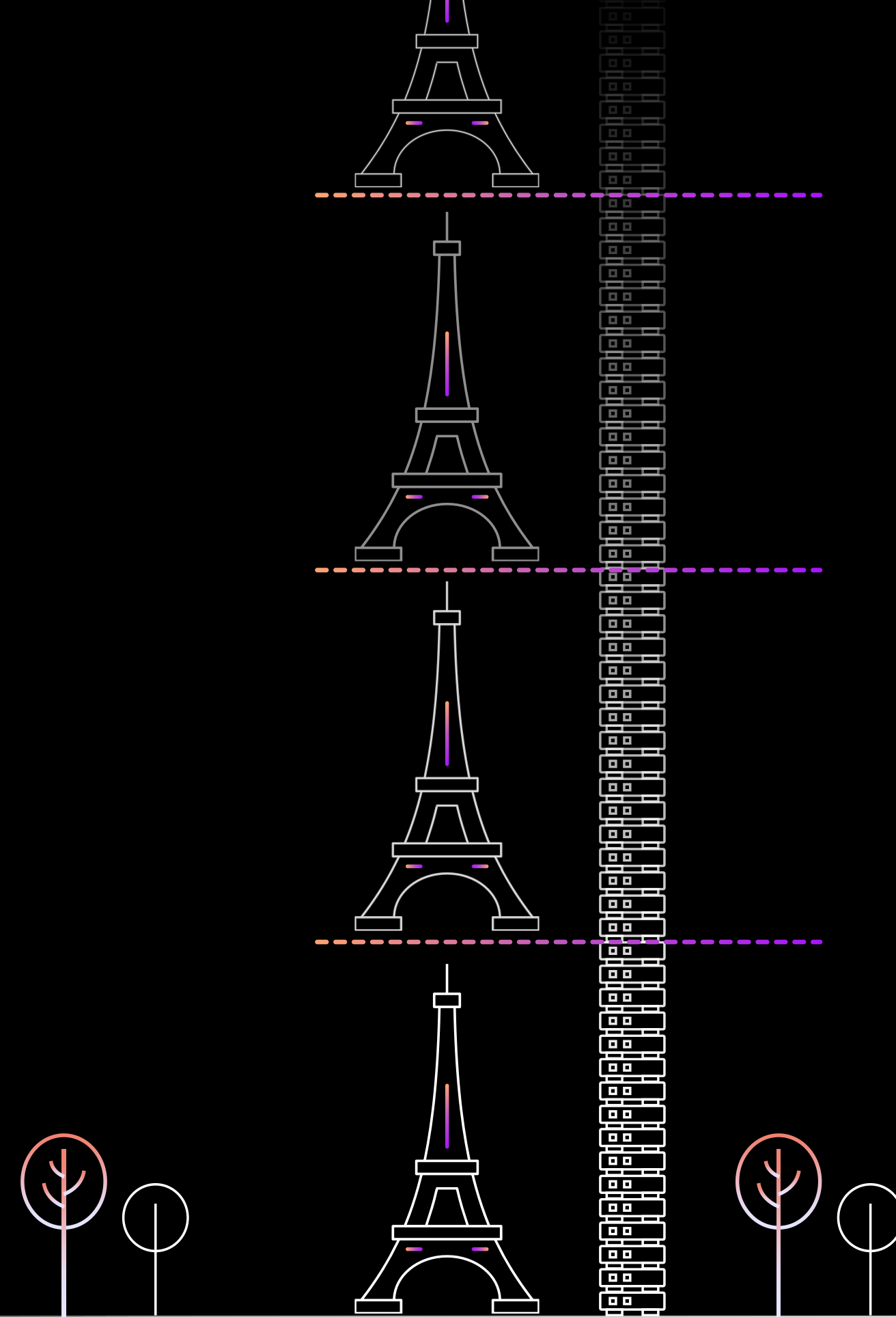
Have data strategy that doesn't throw data away

aws

Modern apps store massive amounts of data

TELEMETRY

LOGS

EVENTS

THIRD PARTY

IoT

WEB

APPLICATIONS

SENSORS

MOBILE

aws

97 ZB

of data created
in 2022

aws

over
25 Million
times

Data-driven organizations
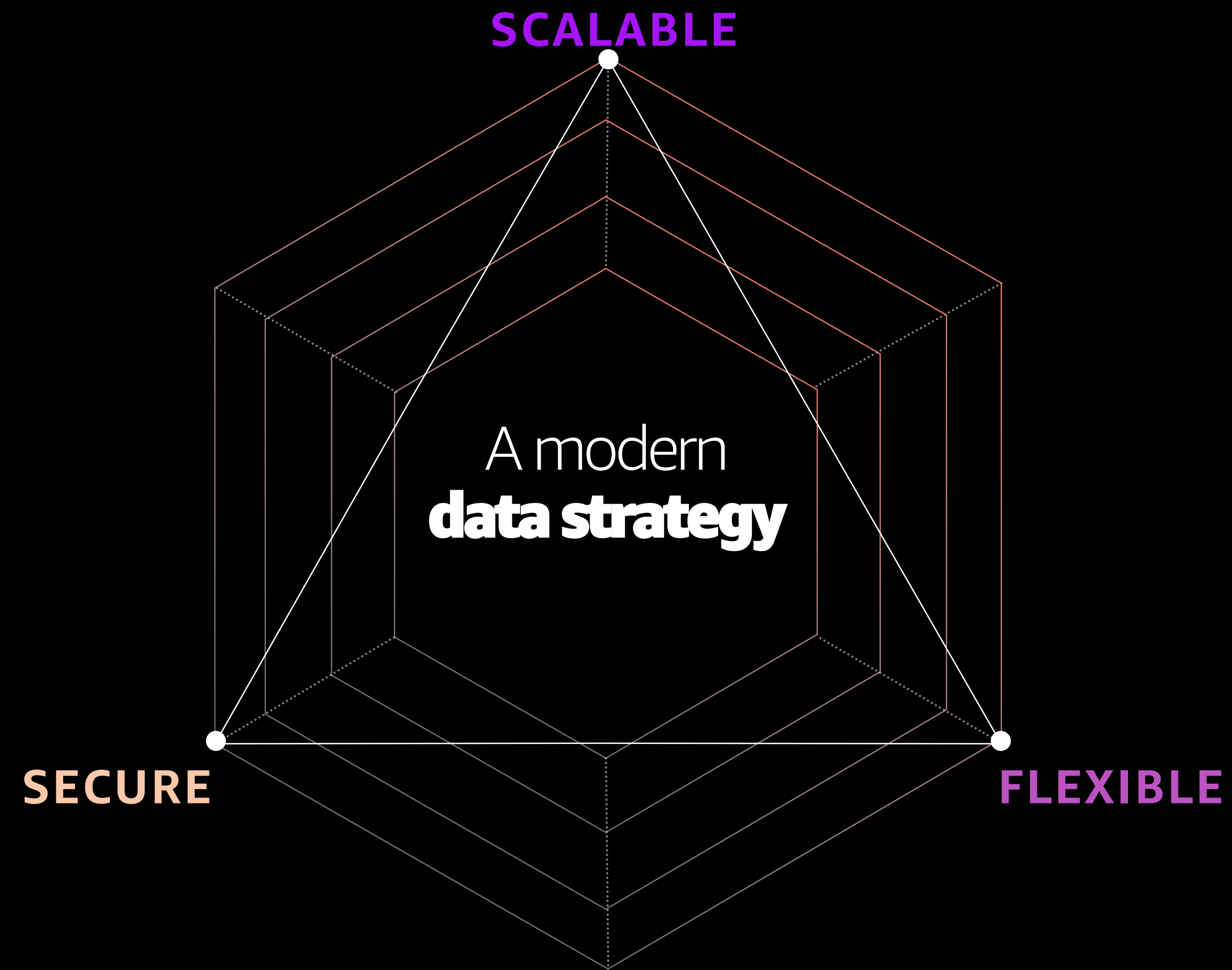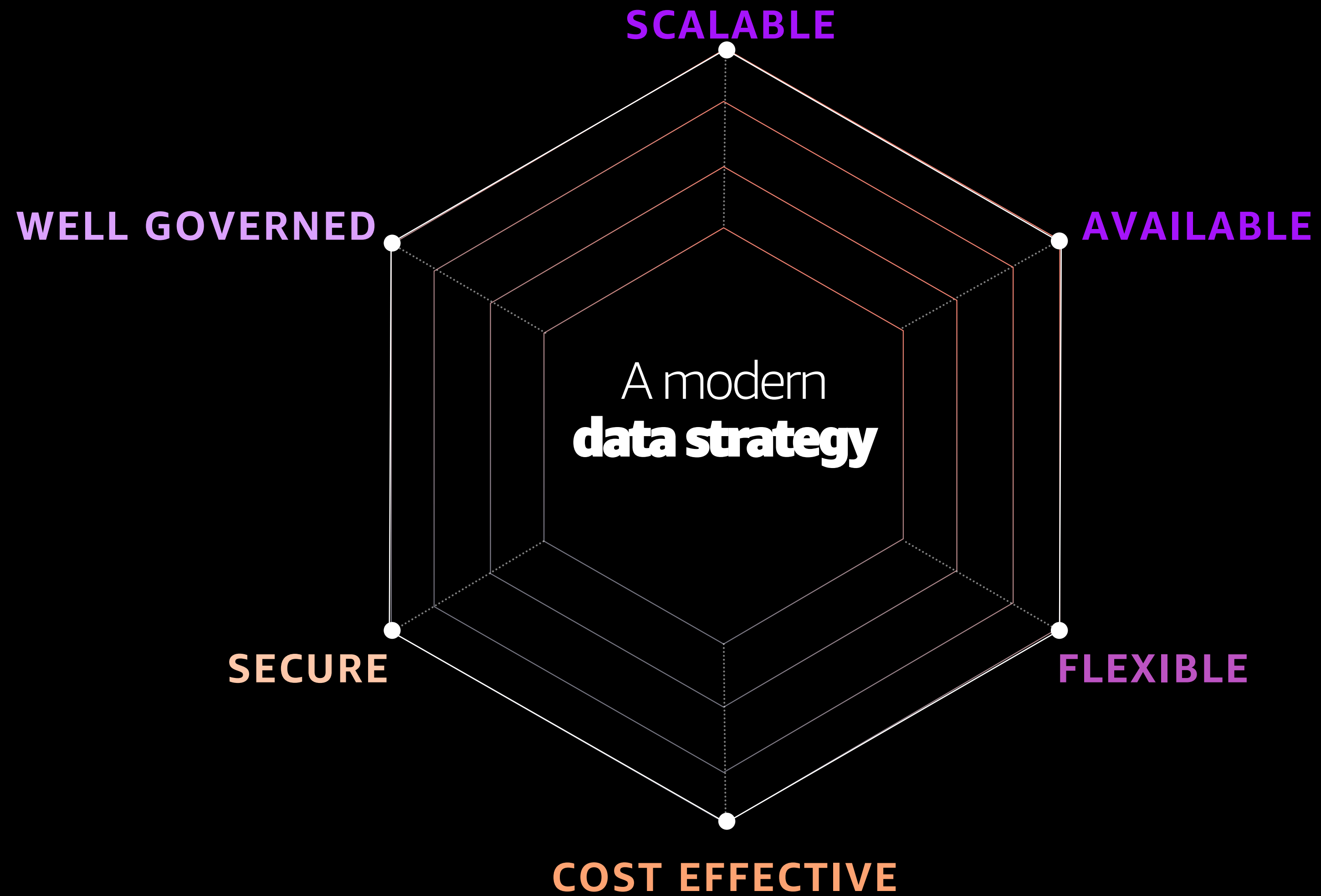are growing an average of

30%+

Statistic provided by Forrester study

aws

68%

of organizations reported that they are
still unable to realize value from data

Statistic provided by Accenture study

aws

SCALABLE

SECURE

FLEXIBLE

A modern
data strategy

SCALABLE

WELL GOVERNED

AVAILABLE

A modern
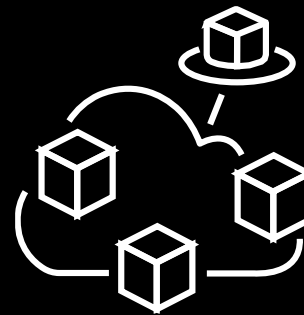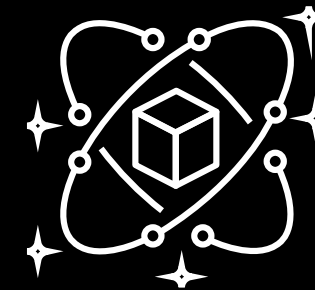**data strategy**

SECURE

FLEXIBLE

COST EFFECTIVE

More than 1.5 million AWS customers
use database, analytics, or machine learning AWS services

Building a
modern data strategy
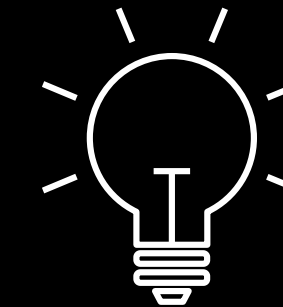
Modernize

Unify

Innovate

Building a
modern data strategy
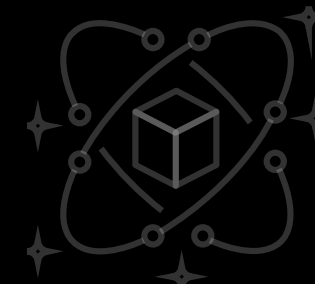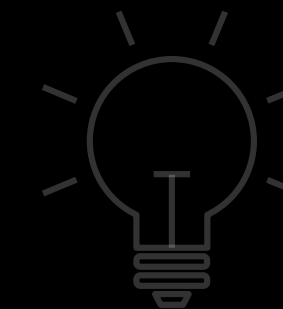
Modernize   Unify   Innovate

Broadest portfolio of
purpose-built databases

RELATIONAL

KEY-VALUE

GRAPH

MEMORY

DOCUMENT

CACHING

TIME-SERIES

LEDGER

WIDE COLUMN

AMAZON
AURORA

AMAZON
RDS

AMAZON
REDSHIFT

AMAZON
DYNAMODB

AMAZON
NEPTUNE

AMAZON
MEMORYDB

AMAZON
DOCUMENTDB

AMAZON
ELASTICACHE

AMAZON
TIMESTREAM

AMAZON
QLDB

AMAZON
KEYSPACES

aws

Amazon
Aurora

MySQL / PostgreSQL-compatible
Relational database service

Security, performance, and availability of high-end commercial databases

Simplicity and cost-effectiveness of open-source databases

5x performance over MySQL

3x performance over PostgreSQL

1/10th the cost of commercial databases

10x

Growth of Amazon Aurora customer usage since 2017
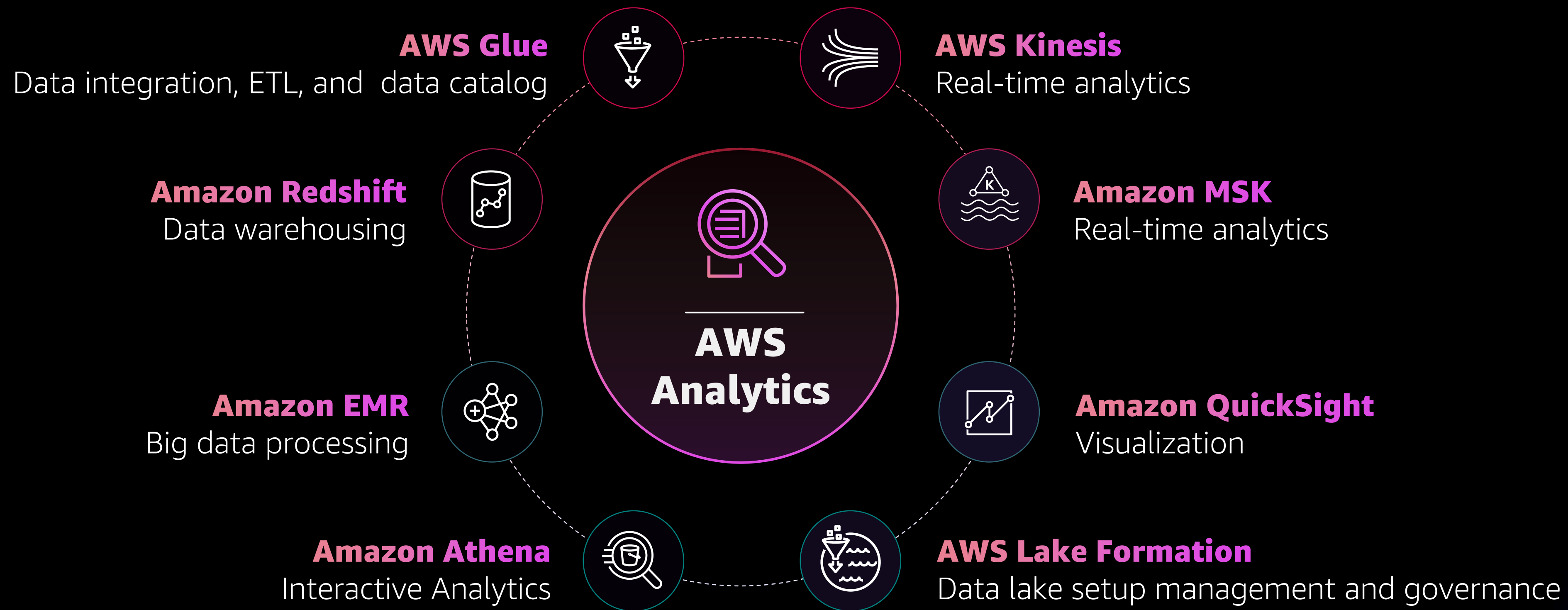
Amazon **Aurora Serverless v2**

Scales to hundreds of thousands of transactions in a fraction of a second

Scale capacity up and down in fine-grained increment for just the right amount the application requires

Save up to 90% vs. provisioning capacity for peak load

Multi-AZ support, Global Database, read replicas

Serverless options for data analytics in the cloud

AWS Glue
Data integration, ETL, and data catalog

AWS Kinesis
Real-time analytics

Amazon Redshift
Data warehousing

Amazon MSK
Real-time analytics

AWS Analytics

Amazon EMR
Big data processing

Amazon QuickSight
Visualization

Amazon Athena
Interactive Analytics

AWS Lake Formation
Data lake setup management and governance

Building a
modern data strategy

Modernize | **Unify** | **Innovate**

CATALOG

LOGS

DATA LAKE

APPS

DATABASES

ON-PREM

BIG DATA ENGINES

IoT

APPLICATIONS

E-COMMERCE

EDGE

DEVICES

DATA LAKE

ANALYTICS

APPLICATIONS

VISUALIZATION

MACHINE LEARNING

aws

Hundreds of thousands of data lakes
running on AWS today

aws

Amazon S3

Unmatched durability, availability, and scalability

Best security, compliance, and audit capabilities

ML-enabled continuous cost optimization: Intelligent tiering

Features to fit any enterprise's object-level control needs

Adapts to customer data lake needs

Broadest portfolio of analytics tools

Cold storage, replication, and archive capabilities

AWS **Lake Formation**

BUILD A SECURE DATA LAKE IN DAYS

Move, store, catalog, and clean your data faster with ML

Manage access control from a single location

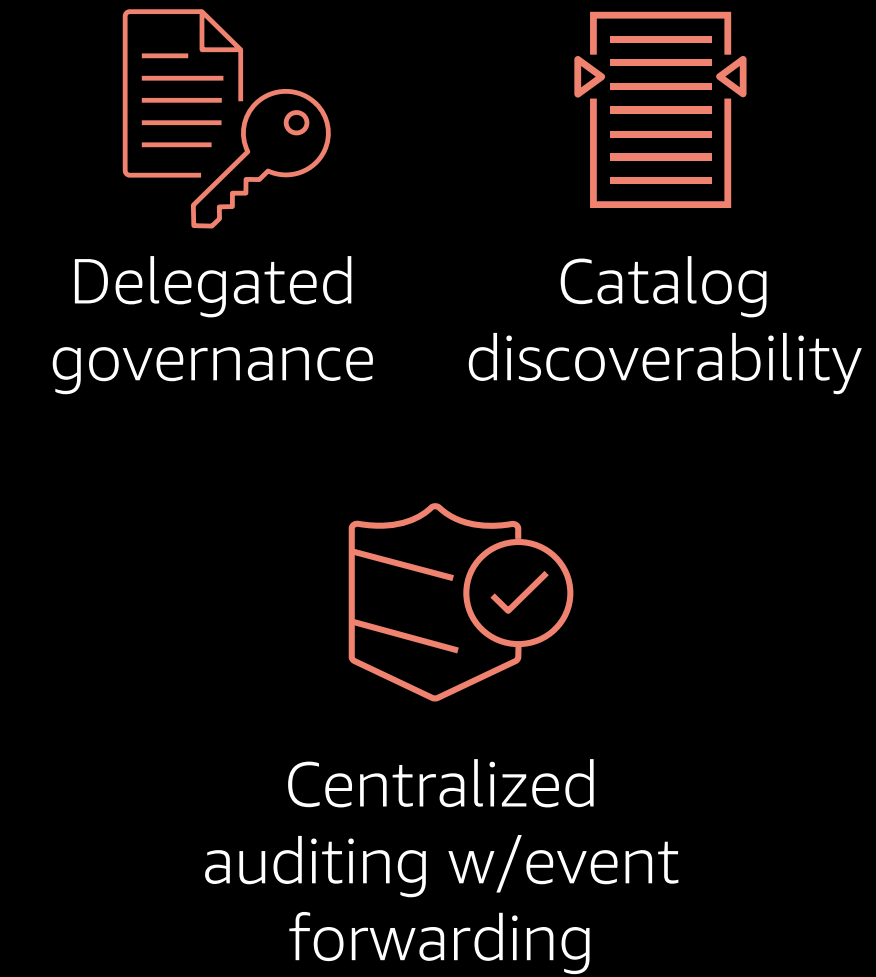Enforce security policies across multiple services

Build

Acid transactions

Centralized catalog

ETL workflow creation

ML transform

Storage optimization
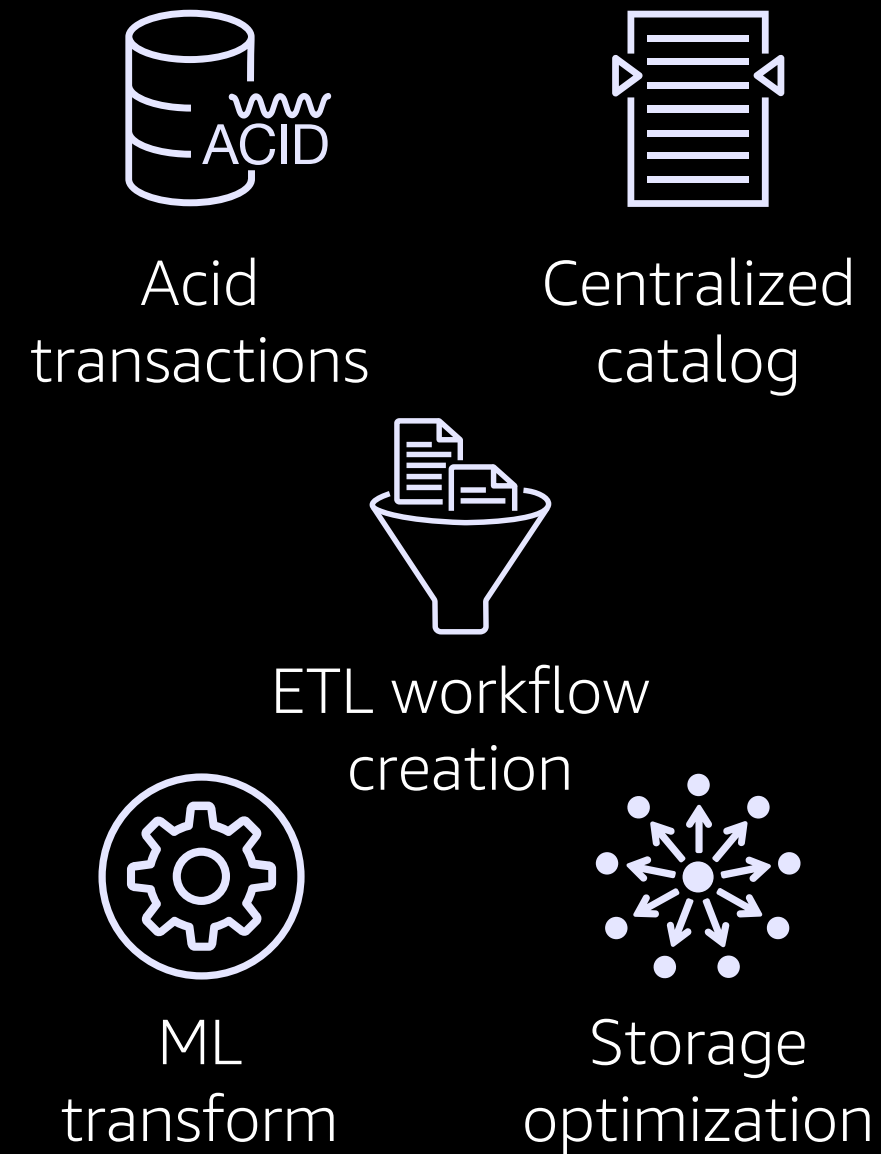
Simplify data ingestion and data management

Secure

DB-style permissions

S3 access managed by resource permission

Fine-grained catalog permissions

Tag-based access to scale permissions

Centrally define and manage security and governance

Share

Delegated governance

Catalog discoverability

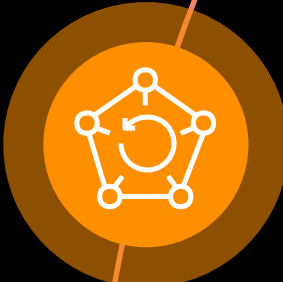**Centralized auditing w/event forwarding**

Enable cross-account data sharing

Seamless Data Movement

AWS Storage Portfolio

Disaster Recovery & Backup

- AWS BACKUP
- AWS EBS SNAPSHOTS
- AWS TRANSFER FAMILY

BLOCK
AMAZON EBS

OBJECT
AMAZON S3 & S3 GLACIER

FILE
AMAZON FSX FAMILY
AMAZON EFS

Data Movement Services
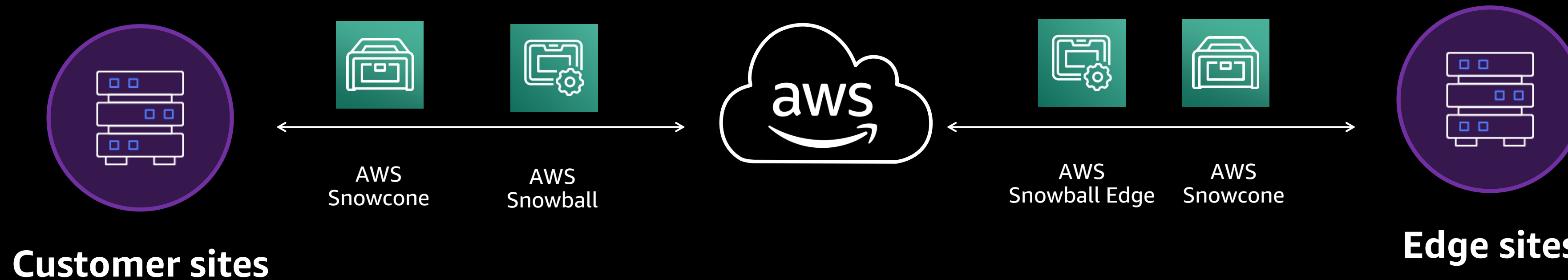
- AWS DATASYNC
- AWS STORAGE GATEWAY
- AWS SNOW FAMILY
- AWS OUTPOSTS

Evolution of the AWS **Snow Family**

Data movement

Edge computing

Customer sites

AWS
Snowcone

AWS
Snowball

aws

AWS
Snowball Edge

AWS
Snowcone

Edge sites

Set of services and functionality on the partially to fully disconnected and secure edge

Allows data collection, data processing, and data movement from the edge to AWS

AWS **Snowcone**

SMALL, PORTABLE, RUGGED, AND
SECURE EDGE COMPUTING AND
DATA TRANSFER DEVICE

Offline and Online Import to Amazon S3, Amazon EFS
and Amazon FSx for Windows File Server

8TB of usable HDD storage and 14TB of SSD storage

Military-grade security @ 2.1kg with battery operation
supported for mobility

Amazon EC2 computing or AWS IoT Greengrass function
(2CPU x86 @ 2.22 Ghz and 4GB memory)

625,000+ databases
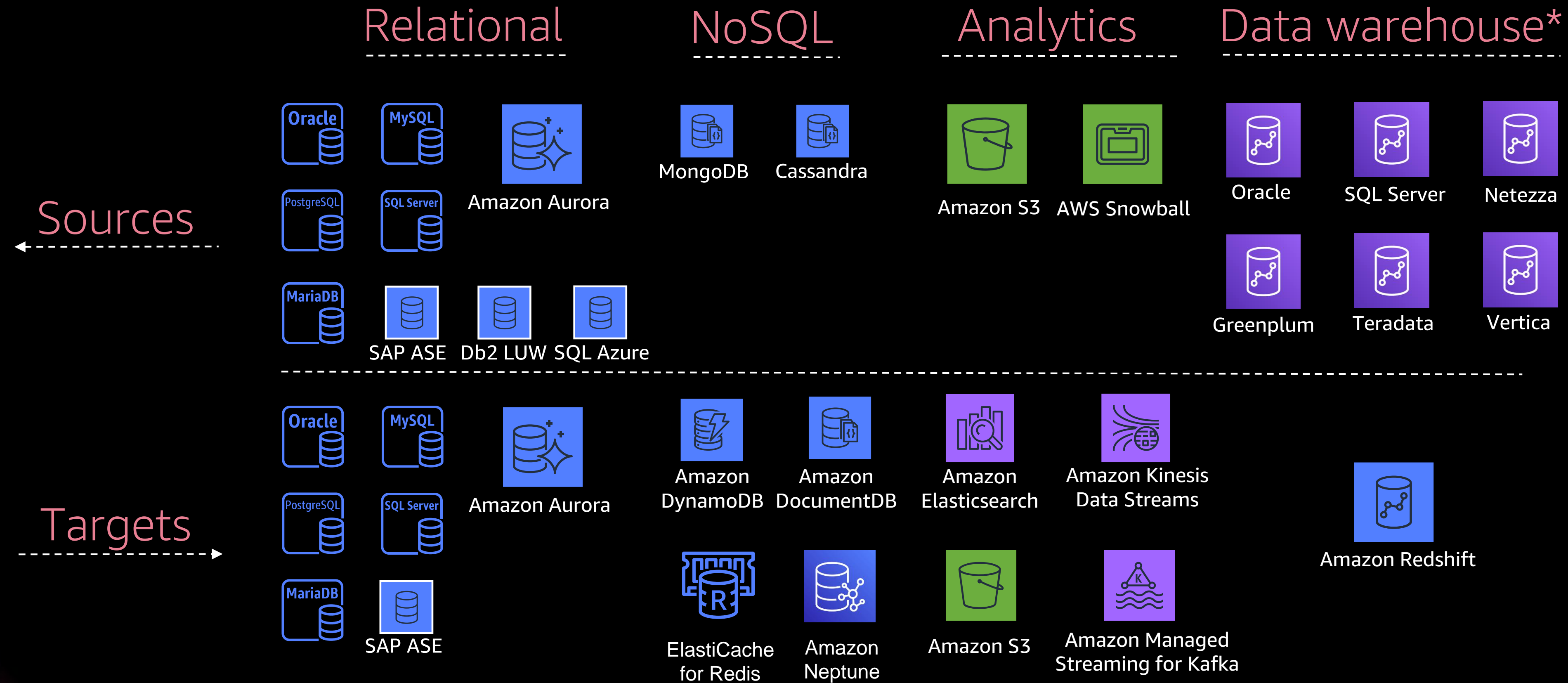migrated to AWS using
AWS Database Migration Service

AWS **Database Migration Service (DMS)**

Migrate data to and from most widely used commercial and open-source databases (on-premises and AWS)
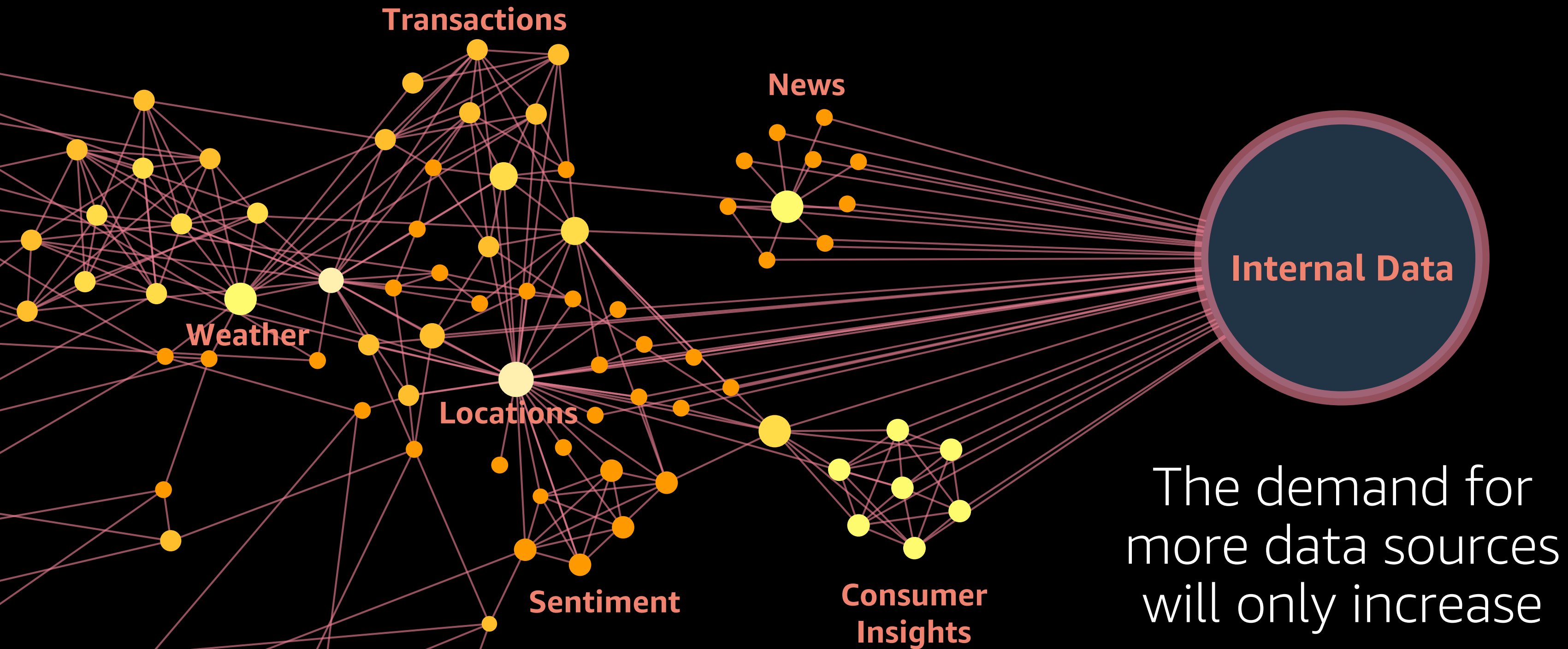
Automated schema conversion

Continuous replication and monitoring for minimal downtime
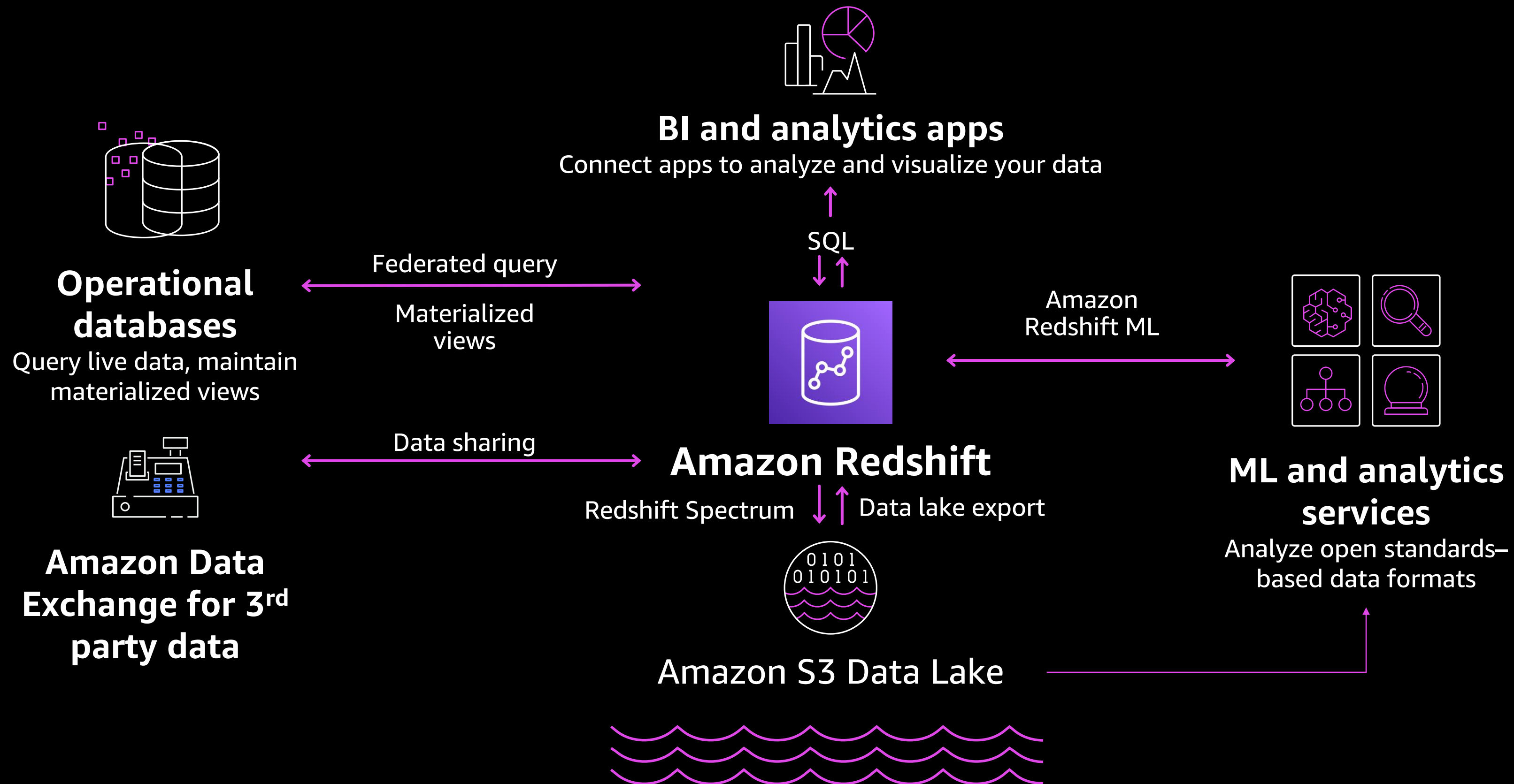
AWS **Database Migration Service**

Relational

NoSQL

Analytics

Data warehouse*

Sources

Oracle MySQL Amazon Aurora

PostgreSQL SQL Server

MariaDB SAP ASE Db2 LUW SQL Azure

MongoDB Cassandra

Amazon S3 AWS Snowball

Oracle SQL Server Netezza

Greenplum Teradata Vertica

Targets

Oracle MySQL Amazon Aurora

PostgreSQL SQL Server

MariaDB SAP ASE

Amazon DynamoDB Amazon DocumentDB Amazon Elasticsearch Amazon Kinesis Data Streams

ElastiCache for Redis Amazon Neptune Amazon S3 Amazon Managed Streaming for Kafka

Amazon Redshift

Breaking down the data silos

External Data is transforming how businesses derive insights

Transactions

News

Internal Data

Weather

Locations

Sentiment

Consumer Insights

The demand for more data sources will only increase

aws

Analyze external data with **AWS Data Exchange**

BI and analytics apps
Connect apps to analyze and visualize your data

SQL

Operational databases
Query live data, maintain materialized views

Federated query

Materialized views

Amazon Redshift ML

Amazon Redshift

Data sharing

Redshift Spectrum   Data lake export

Amazon Data Exchange for 3rd party data

ML and analytics services
Analyze open standards–based data formats

Amazon S3 Data Lake

Vast range of **data providers** across industries

HEALTHCARE AND LIFE SCIENCES	FINANCIAL SERVICES	LOCATION INTELLIGENCE	MEDIA & ENTERTAINMENT	DEMOGRAPHICS AND CONSUMER INSIGHTS	SUSTAINABILITY AND ESG
CHANGE HEALTHCARE	BondCliQ	FOURSQUARE	VBI Variety Business Intelligence	acxiom	arabesque
EVERSANA	CoreLogic	KOCHAVA	IMDb	apptopia	CSRHUB
IBM Watson Health	dun & bradstreet	mapbox	Meltwater	ENVESTNET Yodlee	SASB
IHME Measuring what matters	FACTSET	PlaceIQ	REUTERS	EPSILON	teralytic
ovation	Tradition	precisely	SAMBA TV	EQUIFAX	Verisk Maplecroft
QIAGEN	Tradeweb	SAFEGRAPH	STATS PERFORM	experian	weathertrends360 better business in any weather

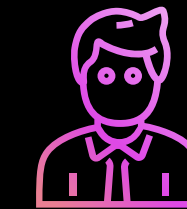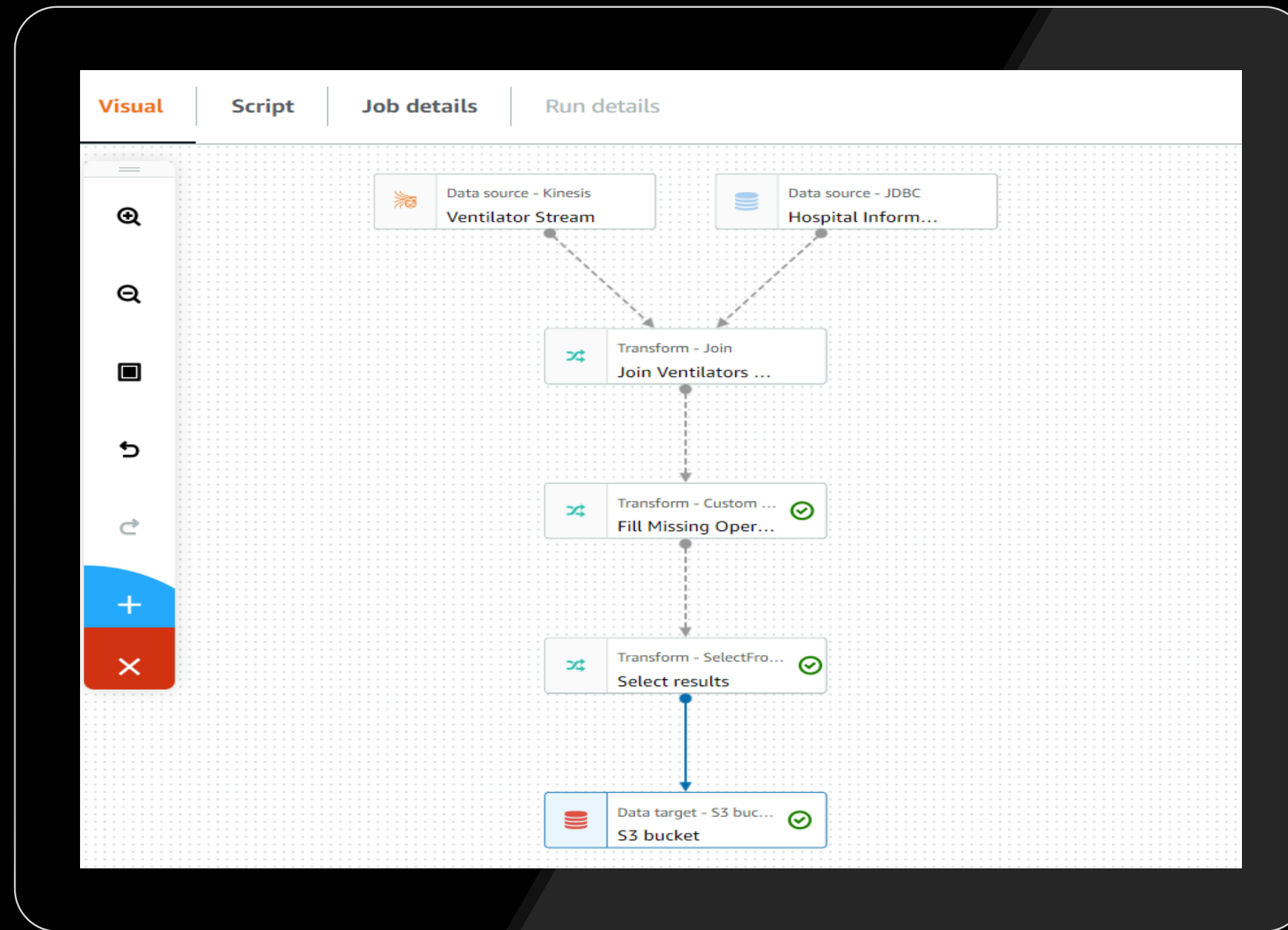# Break down data silos with **Amazon AppFlow**
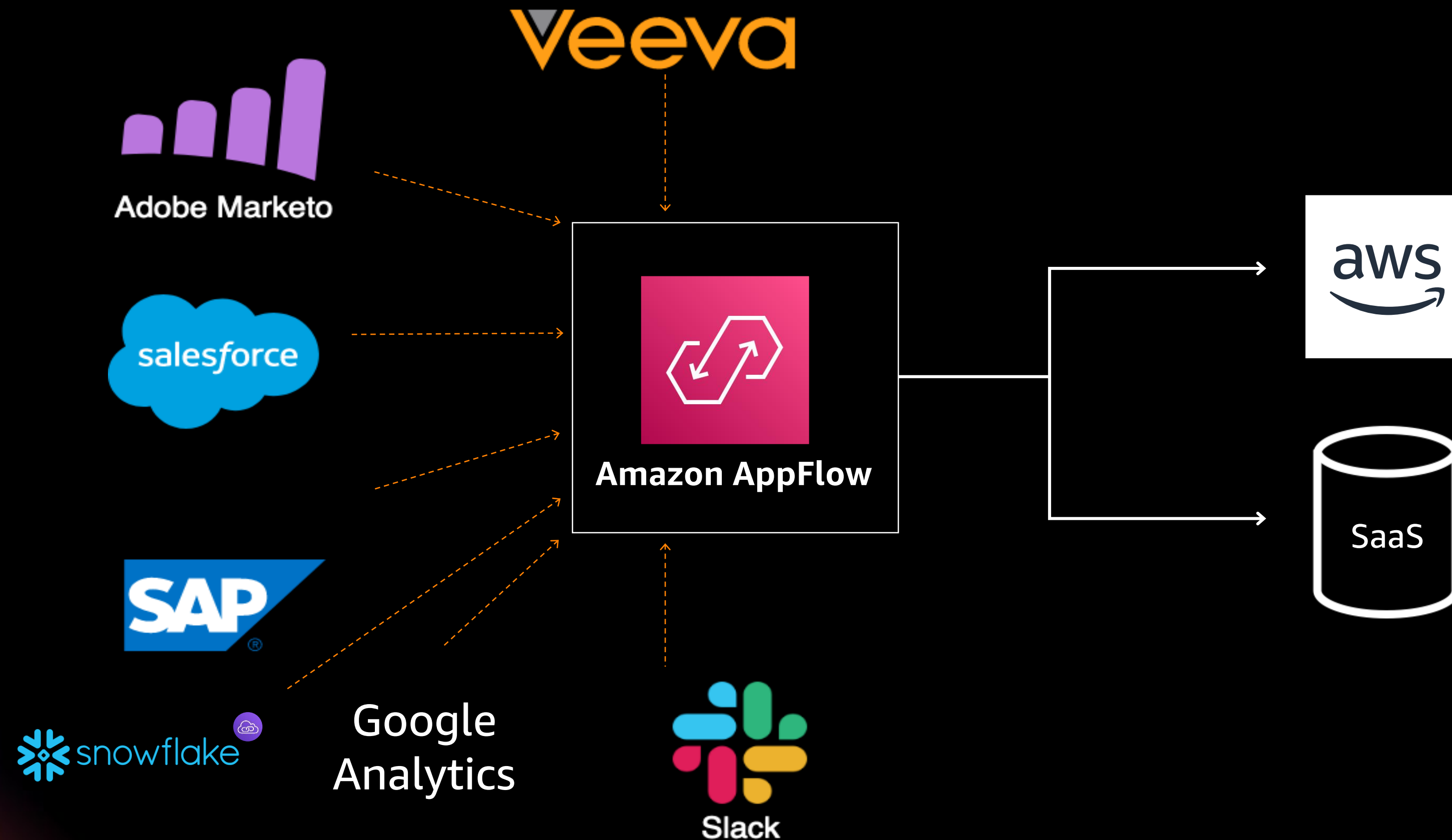
Automate bi-directional data flows between SaaS applications and AWS Services in just a few clicks

Seamlessly connect your data flows with your **SaaS providers**
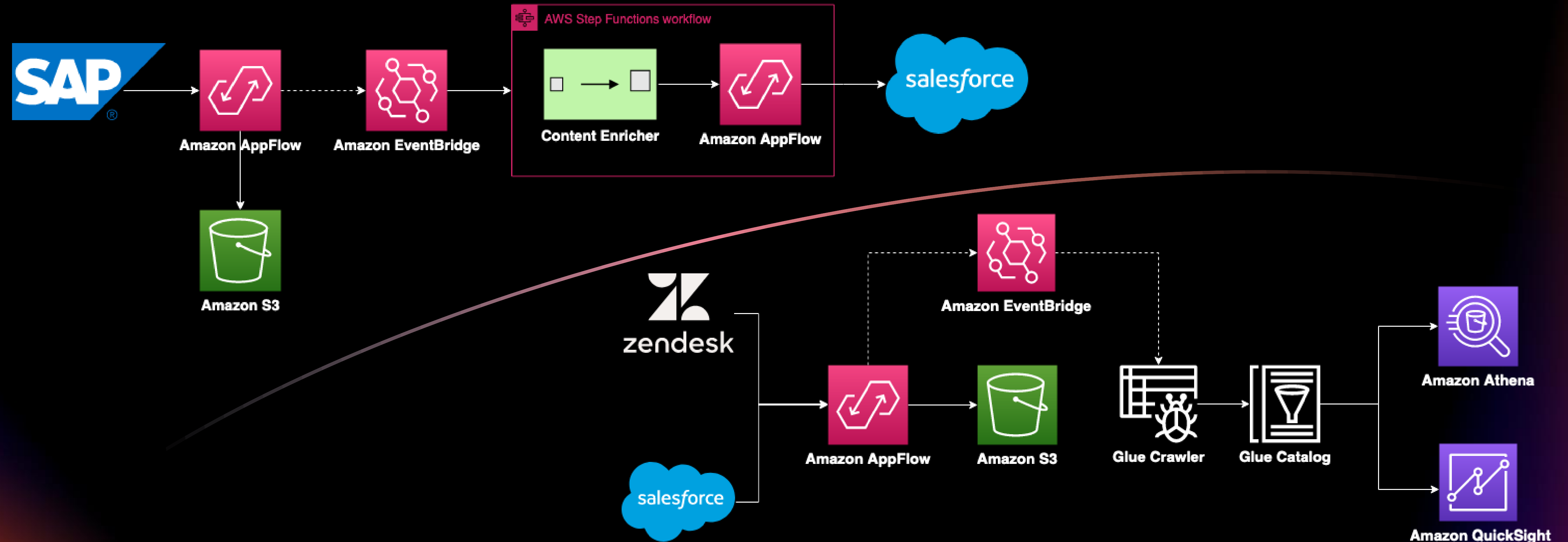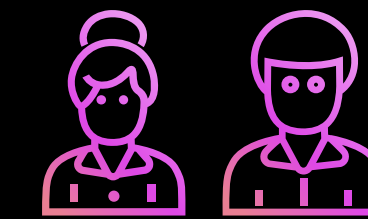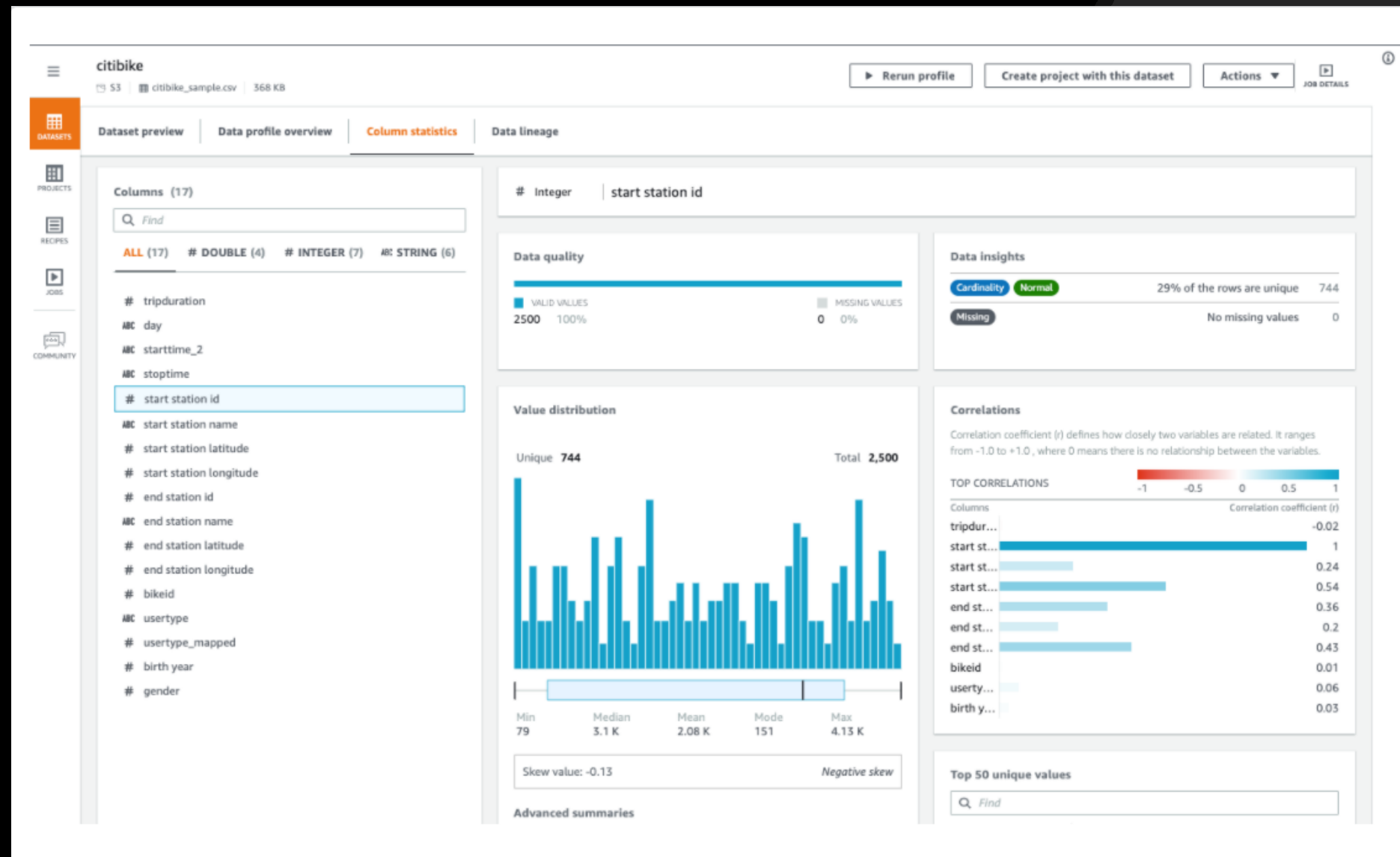
AWS Glue Studio

ETL developer

Rich **visual interface**

250+ built-in transformations

Profile data to understand data patterns and anomalies

Work on large datasets **at scale**
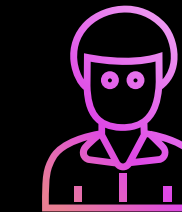
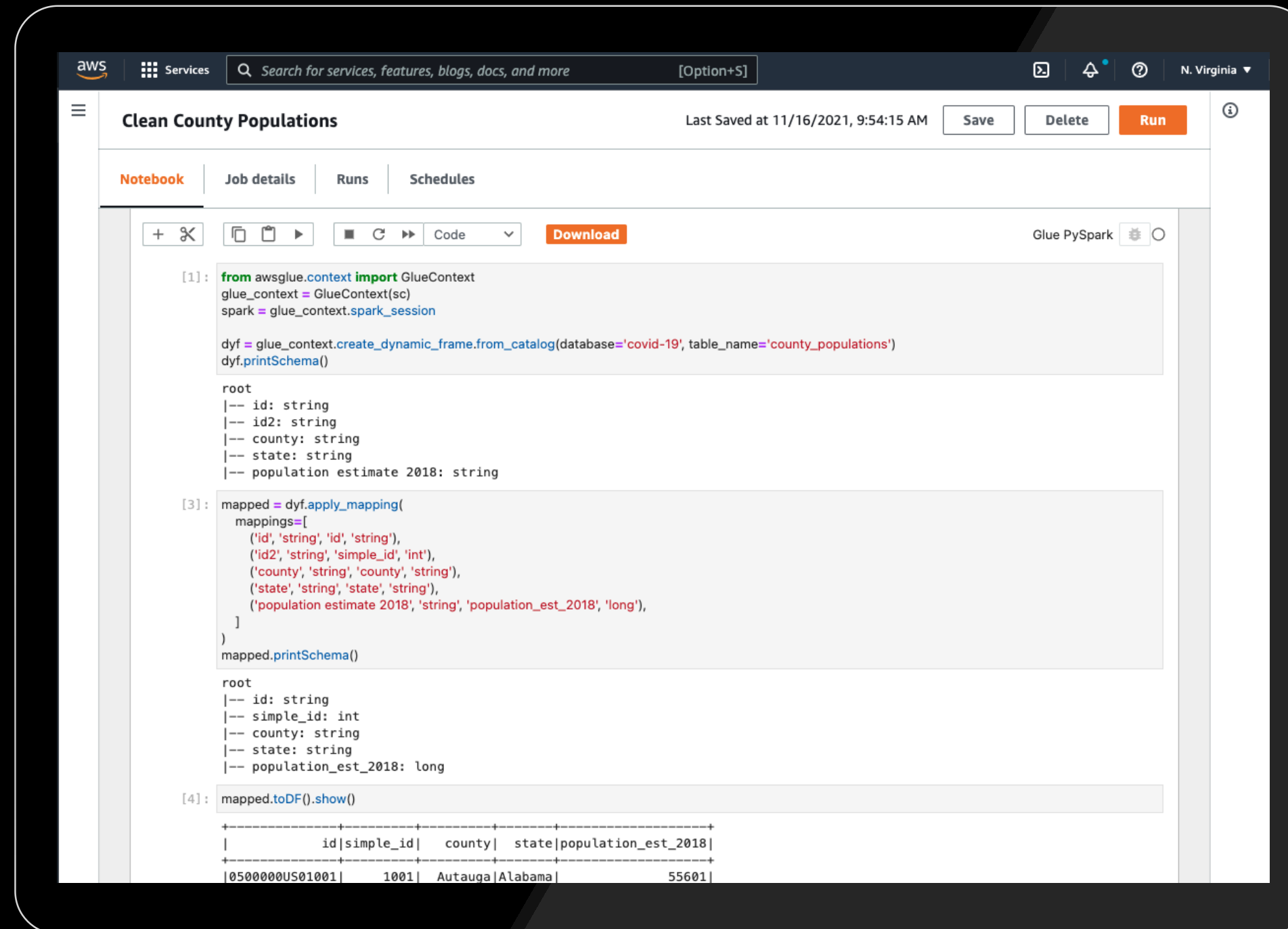# AWS Glue DataBrew

Business Analyst
Data Scientist

Run ETL jobs **without writing code**

Monitor thousands of jobs through a single pane of glass

Distributed processes

**Advanced transforms** through code snippets

AWS Glue Notebooks

Clean County Populations — Last Saved at 11/16/2021, 9:54:15 AM

Notebook   Job details   Runs   Schedules

```
[1]:  from awsglue.context import GlueContext
      glue_context = GlueContext(sc)
      spark = glue_context.spark_session

      dyf = glue_context.create_dynamic_frame.from_catalog(database='covid-19', table_name='county_populations')
      dyf.printSchema()

      root
      |-- id: string
      |-- id2: string
      |-- county: string
      |-- state: string
      |-- population estimate 2018: string

[3]:  mapped = dyf.apply_mapping(
          mappings=[
              ('id', 'string', 'id', 'string'),
              ('id2', 'string', 'simple_id', 'int'),
              ('county', 'string', 'county', 'string'),
              ('state', 'string', 'state', 'string'),
              ('population estimate 2018', 'string', 'population_est_2018', 'long'),
          ]
      )
      mapped.printSchema()

      root
      |-- id: string
      |-- simple_id: int
      |-- county: string
      |-- state: string
      |-- population_est_2018: long

[4]:  mapped.toDF().show()

      +--------------+---------+-------+-------+-------------------+
      |            id|simple_id| county|  state|population_est_2018|
      +--------------+---------+-------+-------+-------------------+
      |0500000US01001|     1001| Autauga|Alabama|             55601|
```

Glue PySpark

Data engineer

Clean and normalize data with a rich **visual interface**

Choose from **250+ built-in transformations** to automate tasks
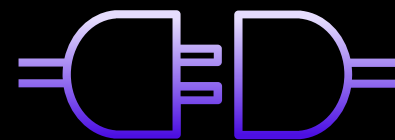
Profile data to understand data patterns and anomalies

Data Movement and Transformation

AWS **Glue**

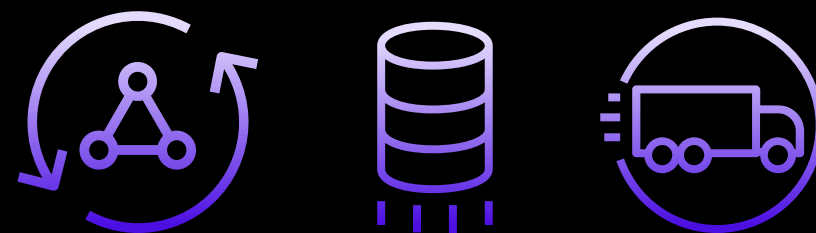Data Integration,
Cataloging, and ETL
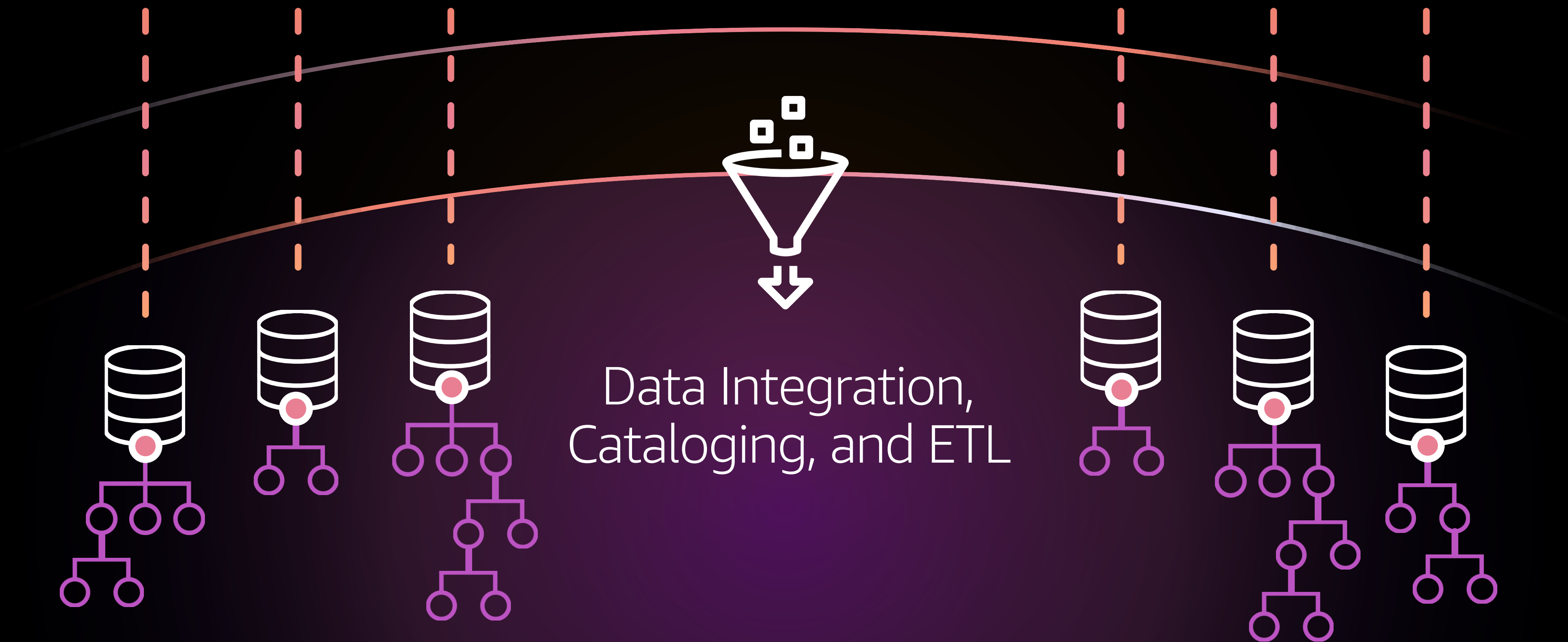
AWS **Glue**

Connect and ingest data

AWS Glue connectors

AWS Glue connector marketplace
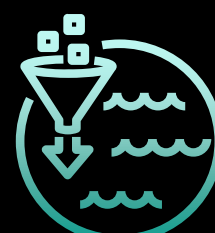
Variety of API and UI interfaces

Centralized and unified data governance
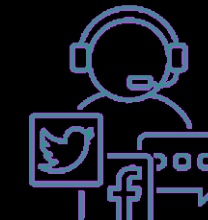
AWS Glue Data Catalog:
Metadata management

AWS Glue Crawlers:
Schema discovery

AWS Lake Formation:
Data lake permissions and management
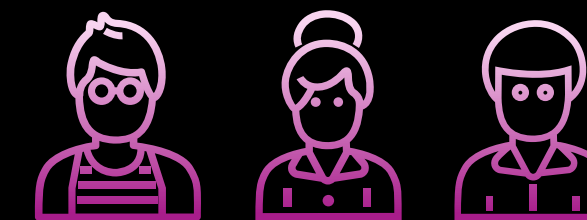
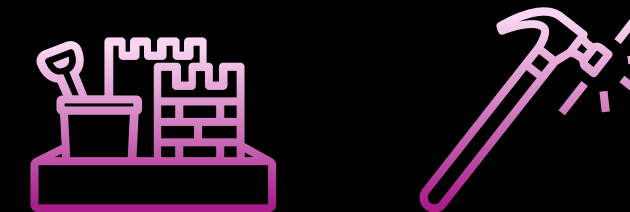## Data management and intelligence

Sensitive data detection

Data cleansing

Data quality

User productivity and data ops

Skillset-specific job authoring

Deployment and orchestration tools

Monitoring and observability tools

Predicting compute demands is difficult for data engineers

Resource demands change

Data volume

Time of day

Seasonality

Needs tuning

Calculation prediction

Right-sizing resources

Ongoing monitoring

Trade-offs
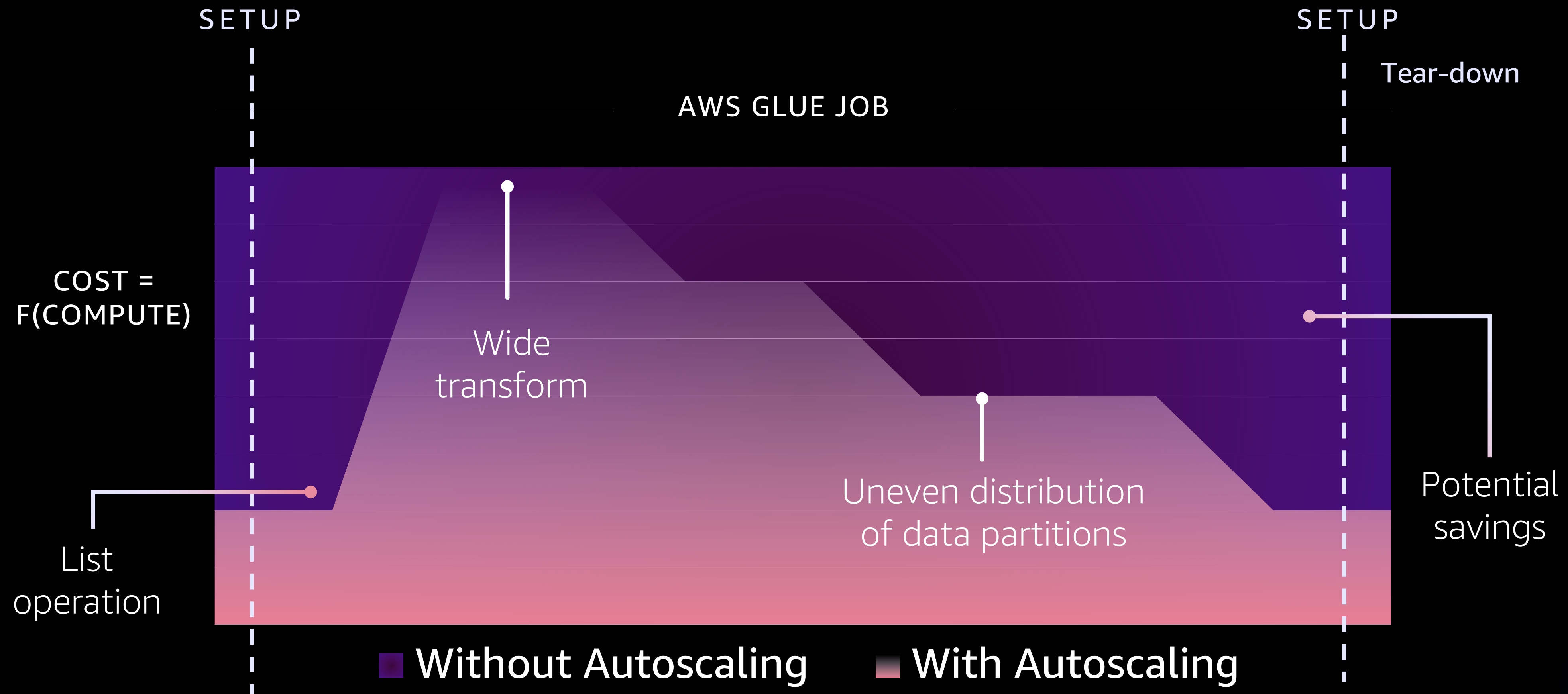
Over-provisioning

Performance variation

AWS **Glue Autoscaling**

- Dynamically scale workers based on workload
- Automatically optimize job resources
- Reduce over-provisioning related costs

aws

SETUP

SETUP

Tear-down

AWS GLUE JOB

COST =
F(COMPUTE)

Wide
transform

Uneven distribution
of data partitions

Potential
savings

List
operation

■ Without Autoscaling   ■ With Autoscaling

Security is **job zero**

AWS **Glue Sensitive Data Detection**

Define sensitive data types

Detect sensitive data at scale

Remediate: delete, redact, replace, or report

Data integration
Federated Query for Athena and Redshift

AGGREGATION

DATA SOURCES

SINGLE REQUEST
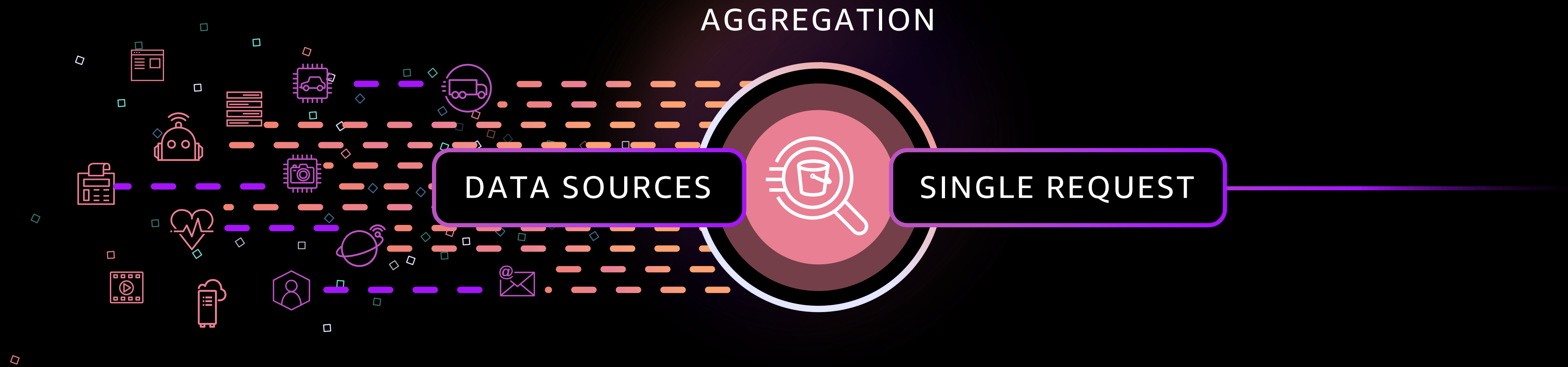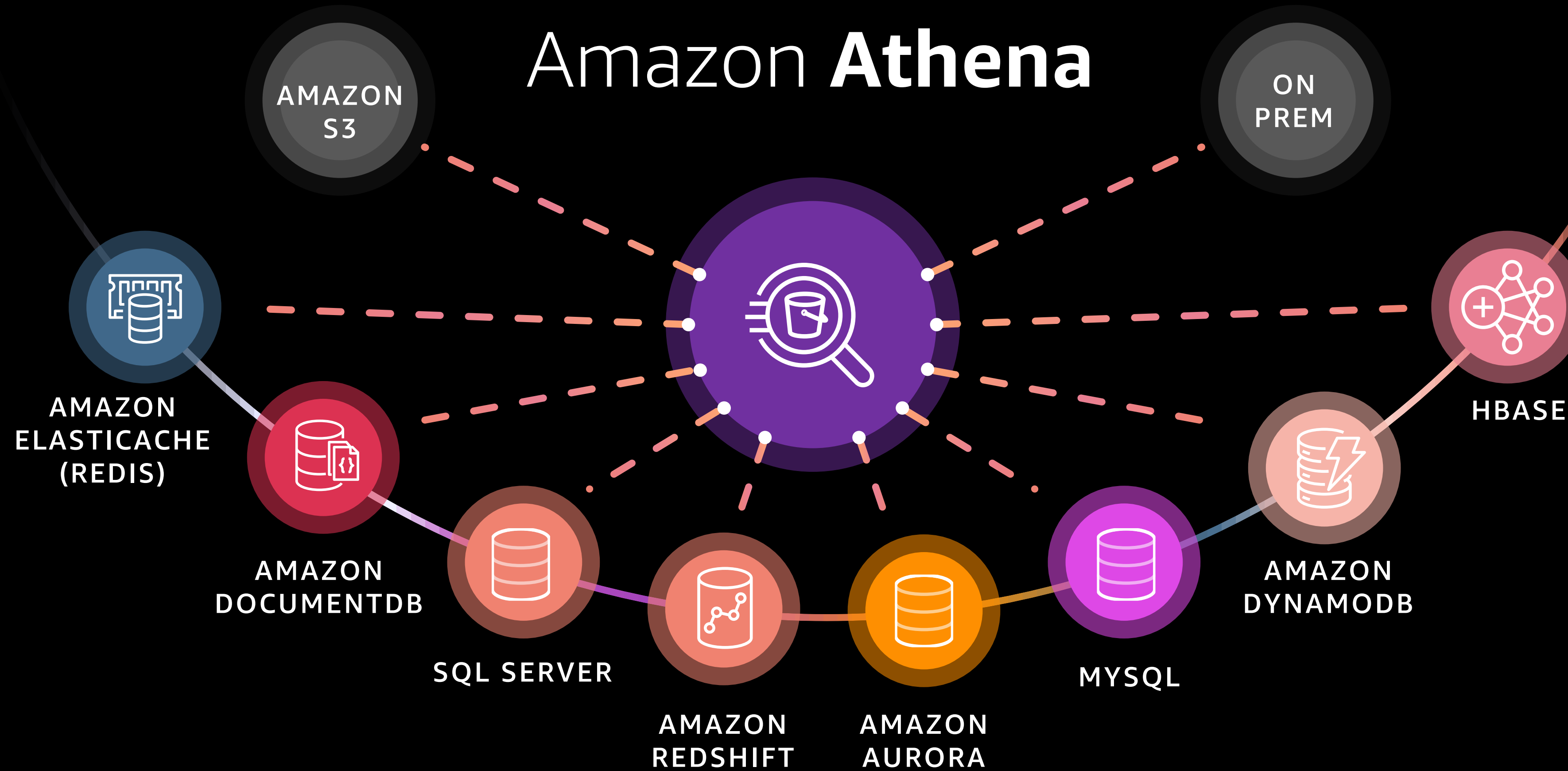
10 New Data Source Connectors for Amazon Athena

Now supporting SAP HANA, Snowflake, Teradata, and more

Connect multiple data sources in a single interface

No need for data movement or copying

Easy setup to popular data stores

aws

Federated Query in Amazon Redshift and Athena

Amazon Redshift

Amazon Athena

Operational databases (i.e., Aurora, RDS)
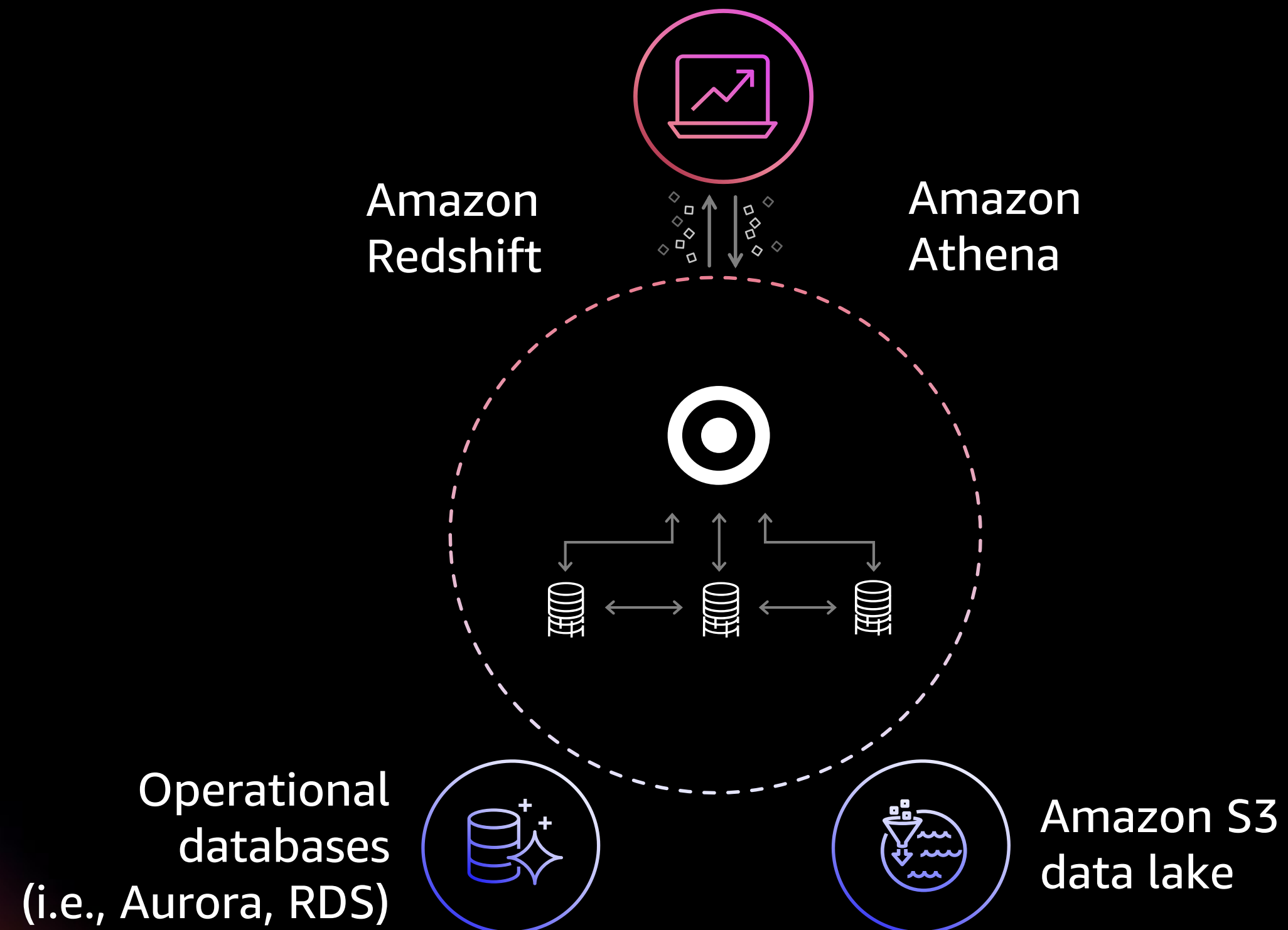
Amazon S3 data lake

Integrate operational database with data warehouse and Amazon S3 data lake

Analytics on operational data **without data movement and ETL delays**

Flexible and easy way to ingest data, avoiding complex ETL pipelines
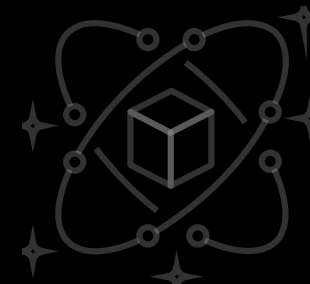
AWS Glue Elastic Views
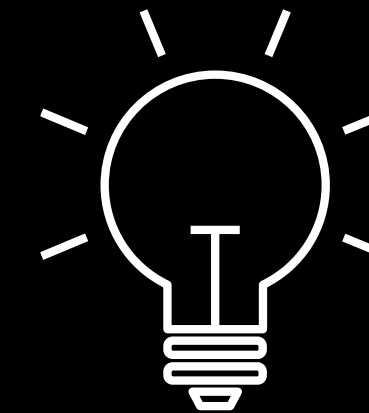
EASILY COMBINE AND REPLICATE DATA ACROSS MULTIPLE DATA STORES

Amazon DynamoDB

Amazon Aurora

Amazon RDS

Amazon Glue Elastic Views

Amazon RDS

Amazon Redshift

Amazon OpenSearch Service

Amazon Aurora

Amazon DynamoDB
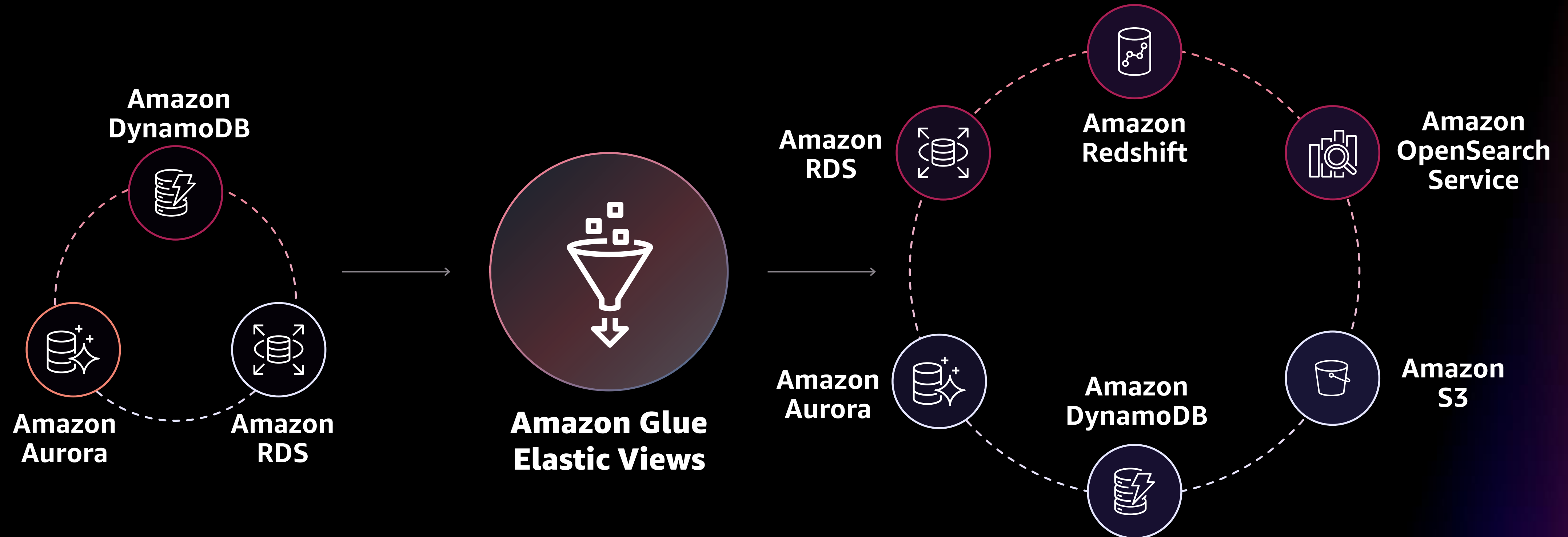
Amazon S3

Building a
modern data strategy

Modernize

Unify

Innovate

AWS brings
ML closer to data

Databases

+

Data warehouses
+ data lakes

+

Business
intelligence tools

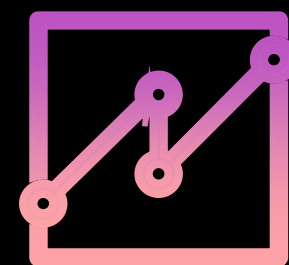
AMAZON
AURORA ML

AMAZON
NEPTUNE ML

AMAZON
REDSHIFT ML

AMAZON
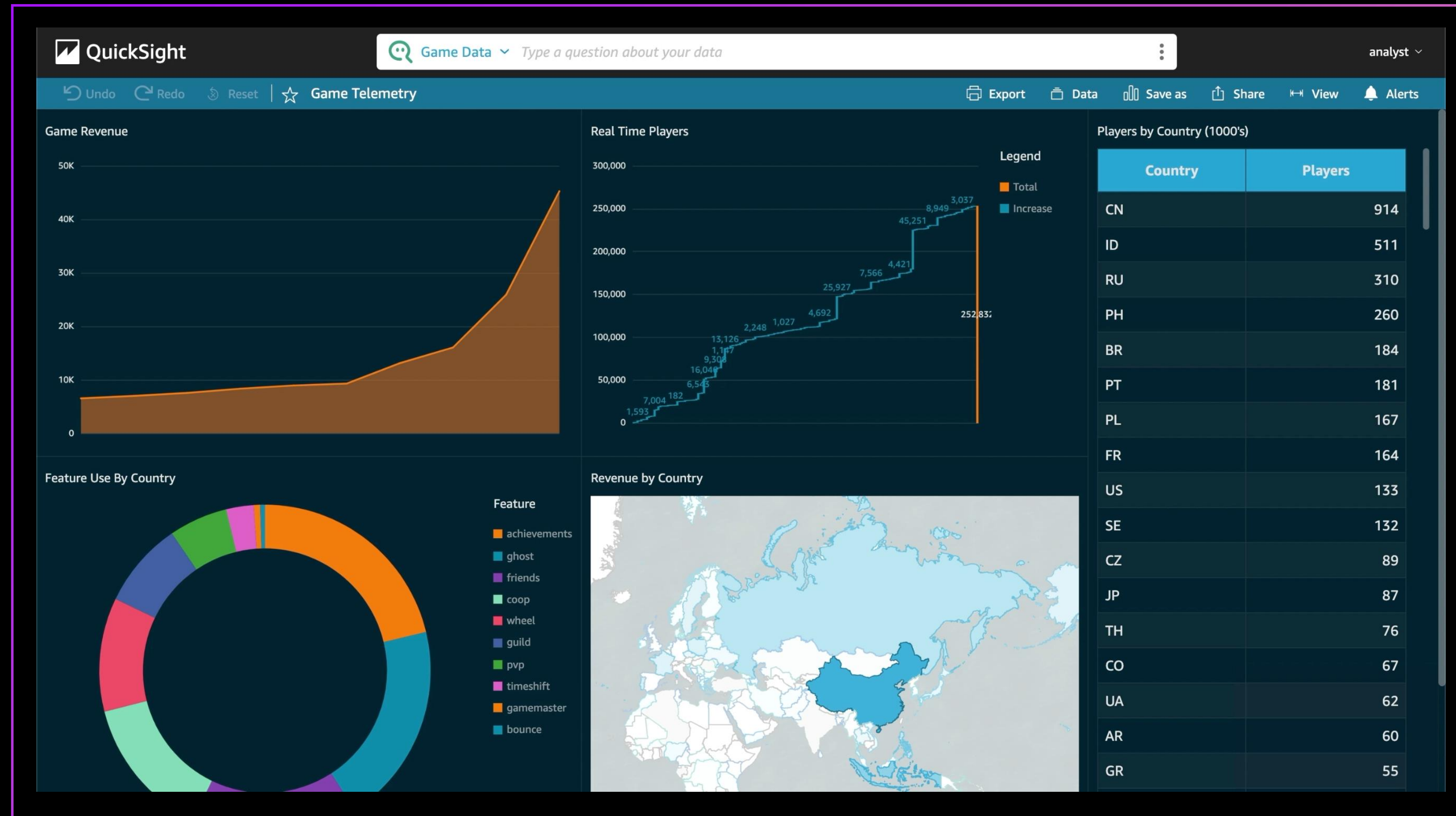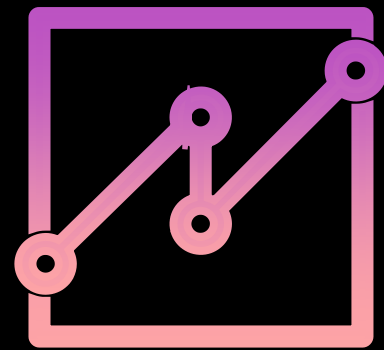ATHENA ML

AMAZON
QUICKSIGHT ML
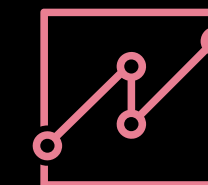
Unlocking business intelligence with

Amazon **QuickSight**

# Accurate results in seconds with **QuickSight Q**

QuickSight

Game Data | Type a question about your data

analyst

Undo | Redo | Reset | ☆ Game Telemetry

Export | Data | Save as | Share | View | Alerts

Game Revenue

Real Time Players

Legend
- Total
- Increase

3,037
8,949
45,251
4,421
7,566
25,927
4,692
1,027
2,248
13,126
9,30
16,04
6,54
7,004 182
1,593

252,83

Players by Country (1000's)

Country	Players
CN	914
ID	511
RU	310
PH	260
BR	184
PT	181
PL	167
FR	164
US	133
SE	132
CZ	89
JP	87
TH	76
CO	67
UA	62
AR	60
GR	55

Feature Use By Country

Feature
- achievements
- ghost
- friends
- coop
- wheel
- guild
- pvp
- timeshift
- gamemaster
- bounce

Revenue by Country

aws

How could I bring analytics into every app I already use?

Amazon
QuickSight
1-Click Public Embedding
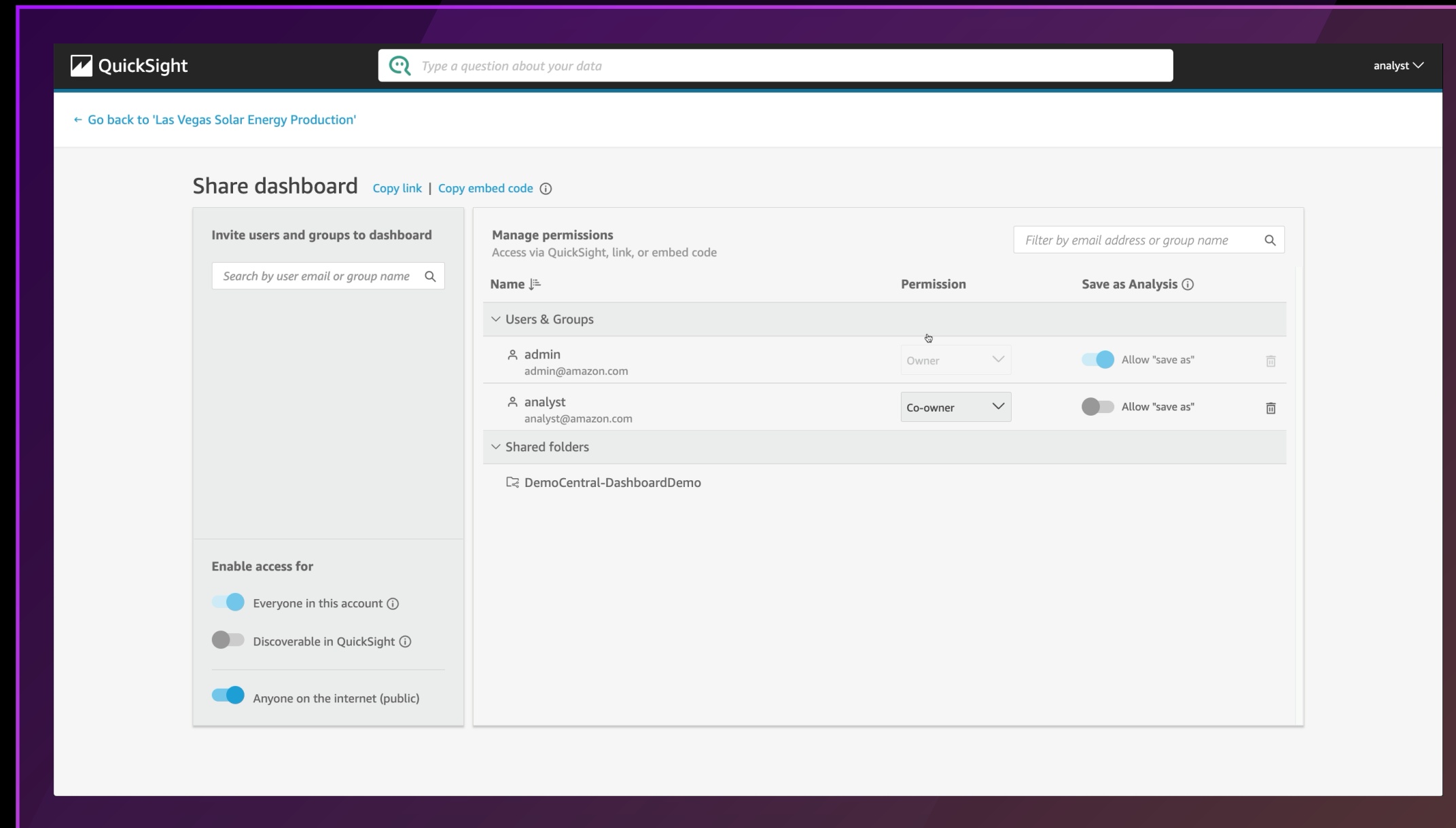
Data Mesh – "Data as a product"

aws

MACHINE LEARNING

BI + ANALYTICS

Data lake

Data lake
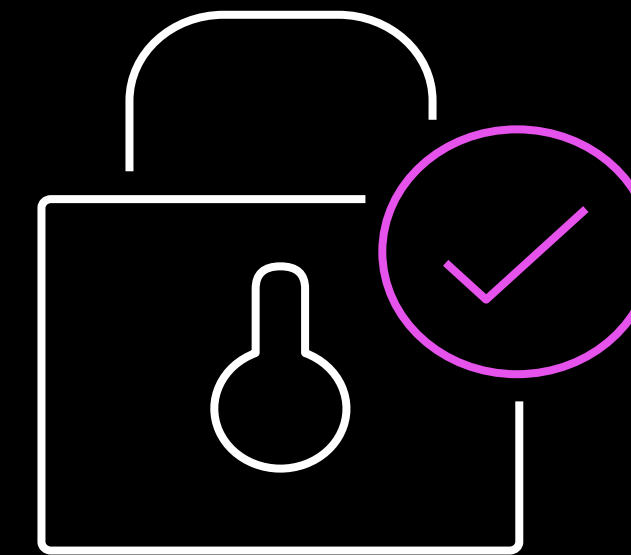
DATA WAREHOUSING

Data lake
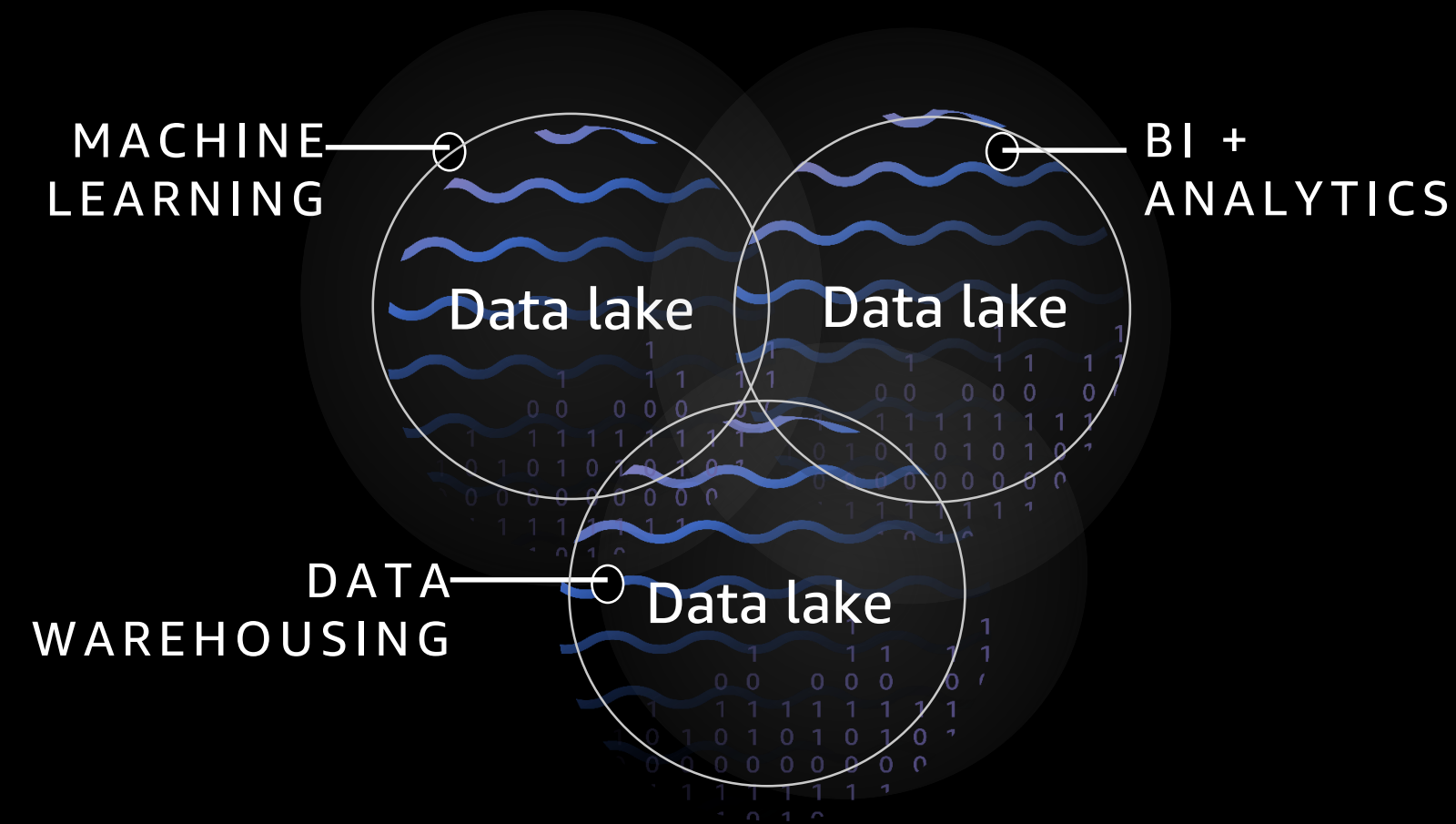
Organizational autonomy & self-service
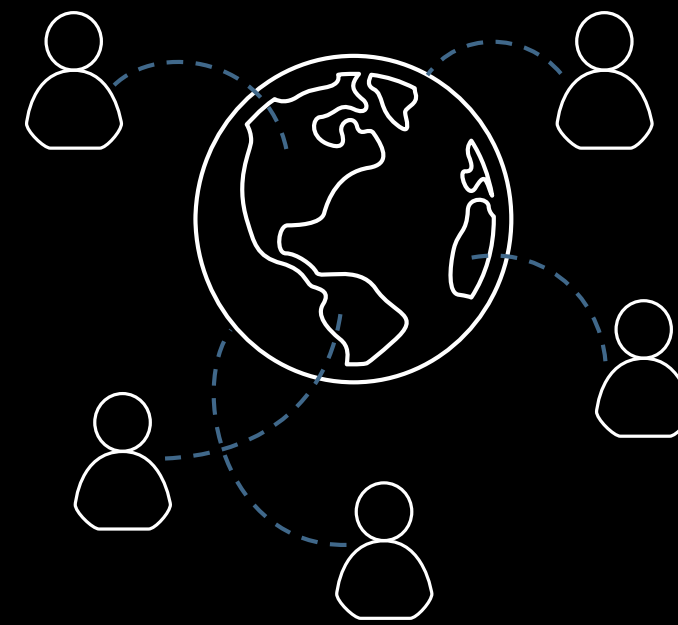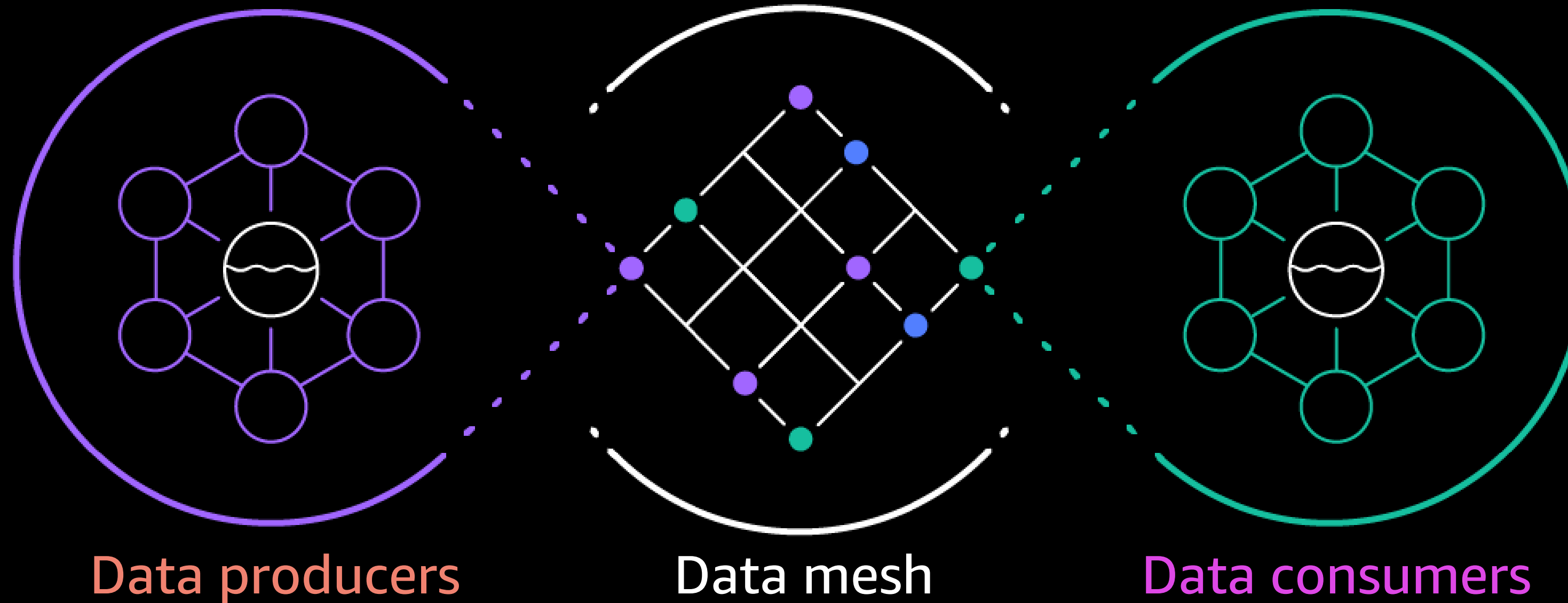
Increasingly diverse consumers of data

Data privacy and compliance

Secure collaboration and sharing
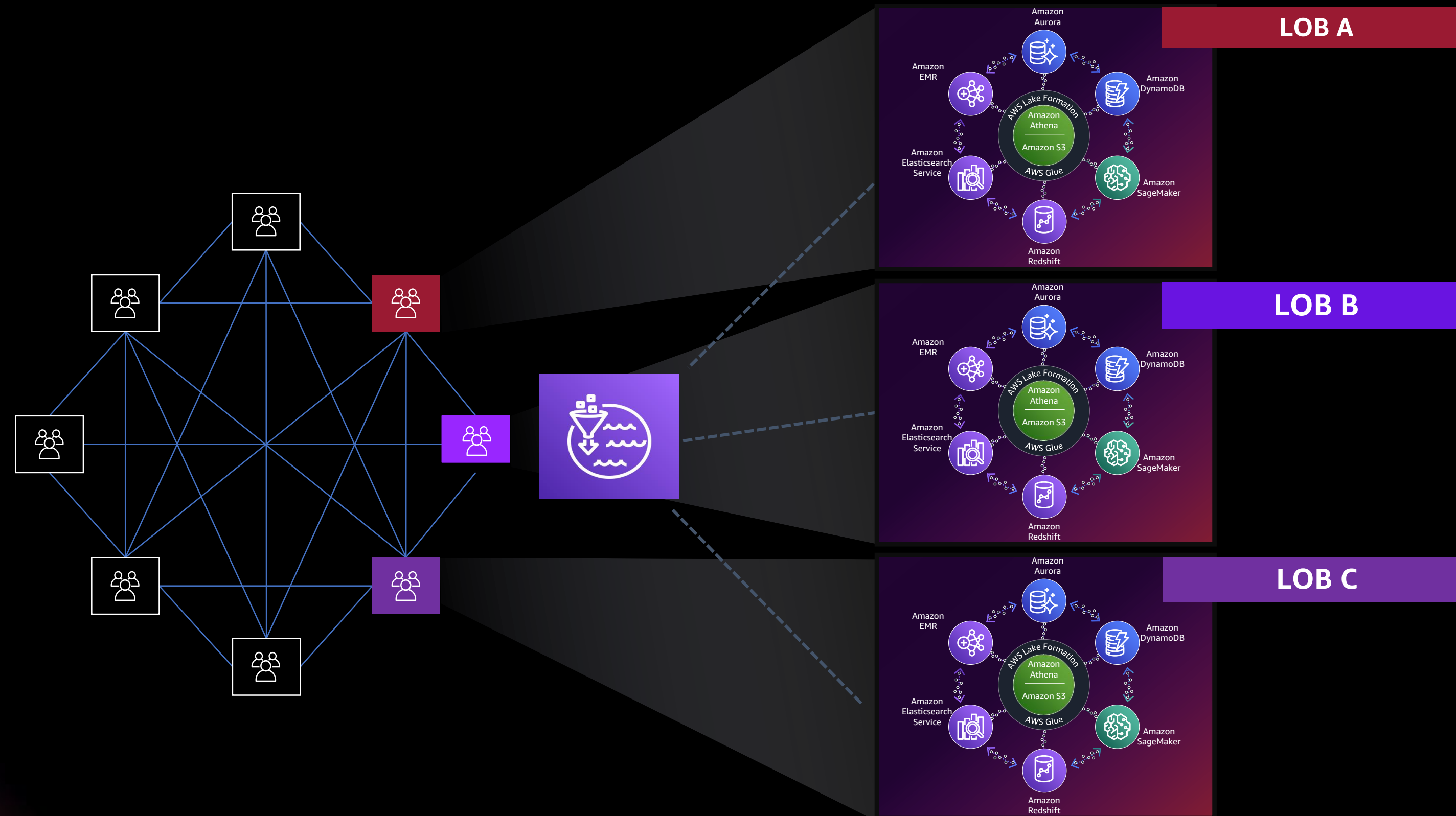
aws

Data Mesh: **Governed sharing of data products**

Architecture that enables the **governed sharing** of **data products** across one-or-more data lakes

Data producers   **Data mesh**   **Data consumers**

A collection of **domain specific lake houses**

LOB A

LOB B

LOB C

Decentralised data domains

Enterprise data products

Federated governance

Self-service sharing

Access Data

Consumer Domain

Discover Data Product
Request Subscription

Create Data Product
Approve Subscription
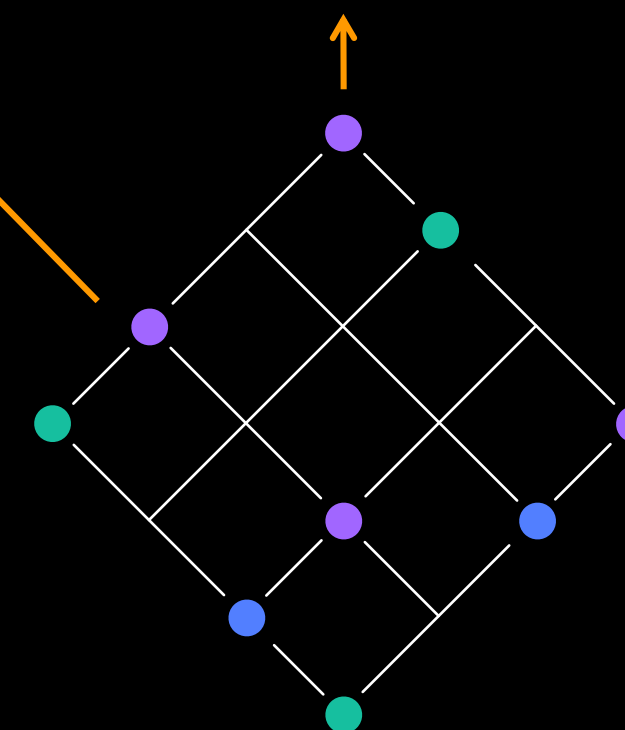
Producer Domain

Monitor
Audit
Govern

Data Mesh **Core Concepts**

Producer Account 1
Raw   Trusted   Refined

Producer Account 2
Raw   Trusted   Refined

Producer Account N
Raw   Trusted   Refined

Federated Data Governance Account
AWS Lake Formation
Register Data Location / Metadata

Consumer Account 1
Amazon Redshift   Amazon EMR Notebook   Amazon Athena

Consumer Account 2
AWS Glue   Amazon EMR Notebook

Data Science Account
Data Wrangler   Managed Notebook   Data Catalog

Data producers share one or more **Data Products** by making them discoverable through a common catalog

Federated data governance enables security and compliance across data domains

Data consumers easily discover and access data using **Resource Shares** or APIs

aws

Federated Data Governance

Federated Access Control to Data Owners

Data Product/ Owner A

Data Product/ Owner B

Databases Owner - A

Databases Owner - B

Databases Owner - N

Security Token Vending

Attribute Based Access Control

Access Auditing

Entitlements

Consumer X

Consumer Y

Data Stewards (Organisation)

PCI /SOX Auditors
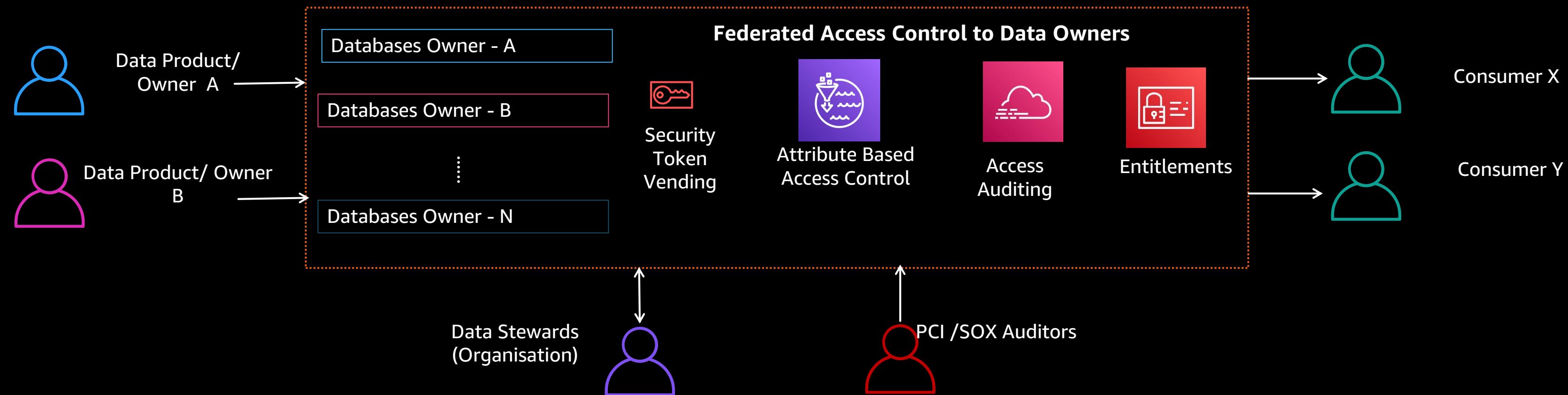
Common security services are core to a governed data mesh

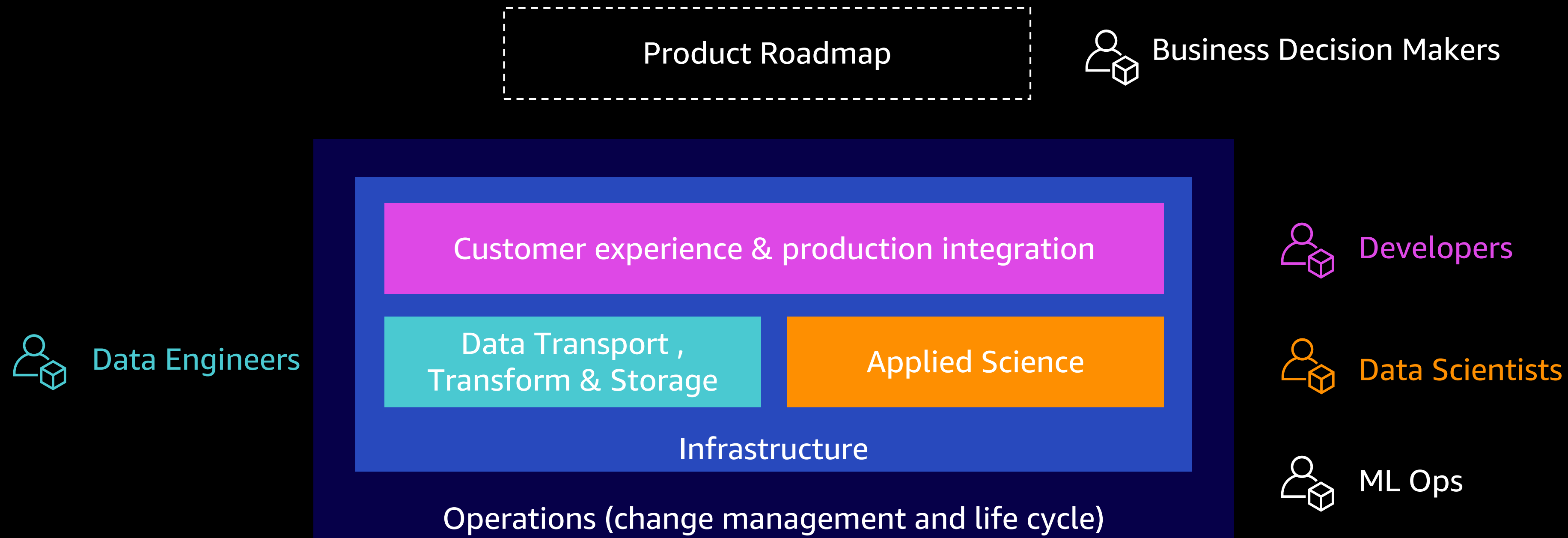Common Identity Provider across producers and consumers

Credential Vending to simplify service integration

Federated metadata search

Fine grained entitlements

Central audit and compliance

aws

Skills and training

It takes a diverse team to build a data product

Product Roadmap

Business Decision Makers

Customer experience & production integration

Developers

Data Engineers

Data Transport, Transform & Storage

Applied Science

Data Scientists

Infrastructure

Operations (change management and life cycle)

ML Ops

- Other important personas in ML projects: BI analysts, Product Managers, Compliance experts
- Processes matter as well (feature prioritization, success metrics, security reviews, etc.)

aws

Developers
- Amazon Rekognition
- Amazon CodeGuru
- AWS Data Pipeline

Business analysts
- Amazon Forecast
- Amazon Athena
- Amazon QuickSight
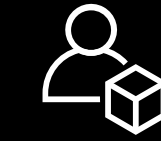
Data scientists
- Amazon SageMaker
- Amazon EMR Serverless
- Amazon Deep Learning AMIs
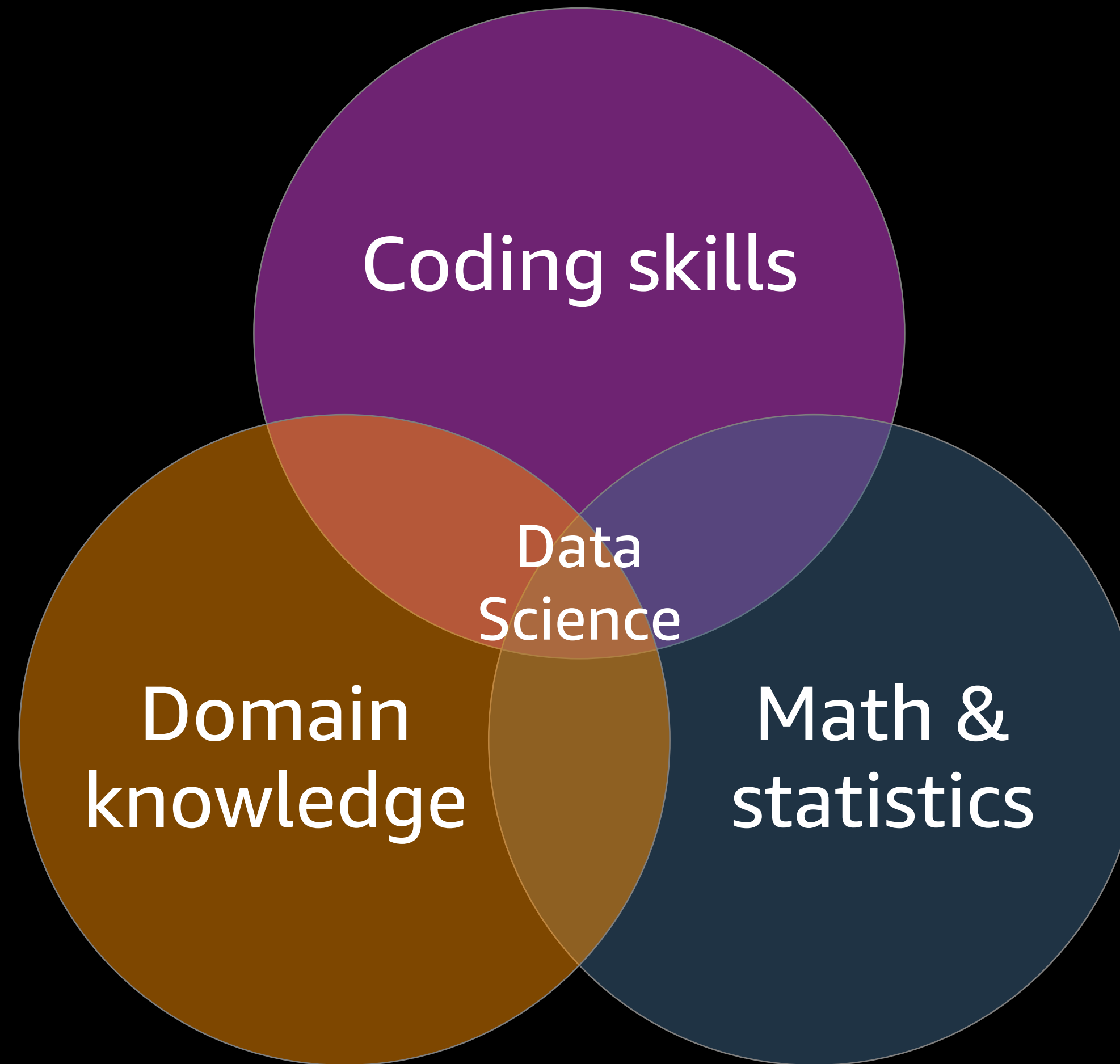
Coding skills

Math & statistics

Domain knowledge

Data Science

AWS
User Groups

➢ **360+ AWS User Groups** ➢ **700+ User Group Leaders**

➢ **In 98 countries** ➢ **560k+ members**

AWS User Groups are community-led meetups which provide builders the opportunity to learn from likeminded peers, participate in AWS training, learn about new AWS services, obtain help in solving technical challenges, and provide networking opportunities for professional growth.

❖ Private Slack community ❖ Interactions with AWS teams

❖ AWS credits ❖ Certification exam voucher

❖ Swag welcome kit ❖ Access to AWS workshops and content

Overview

AWS INNOVATE
DATA EDITION
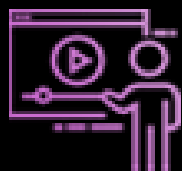
Live Keynote and Closing

FOR BUSINESS
2 TRACKS

Data-driven organizations track 1 & 2

FOR BUILDERS
5 TRACKS

Data movement, processing, management, and governance	Migrate and modernize your databases	Unify your data: Accelerate insights track 1 & 2	Innovate with data and machine learning

3 LOCALIZED
LANGUAGE AGENDA

Japanese	Korean	Bahasa Indonesia

ACTIVITIES

80+ breakout sessions
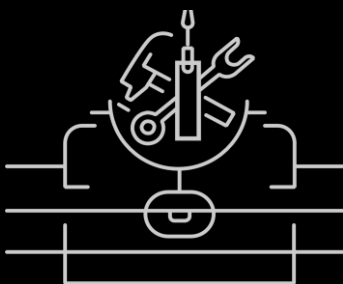
Builders zone

Hands-on demos & labs
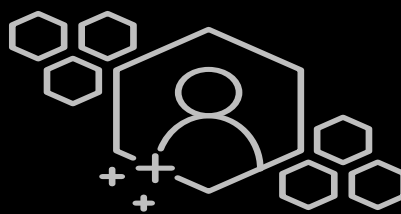
Customer stories

AWS e-books & guides

Ask the experts

Visit the AWS Data resource hub

A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
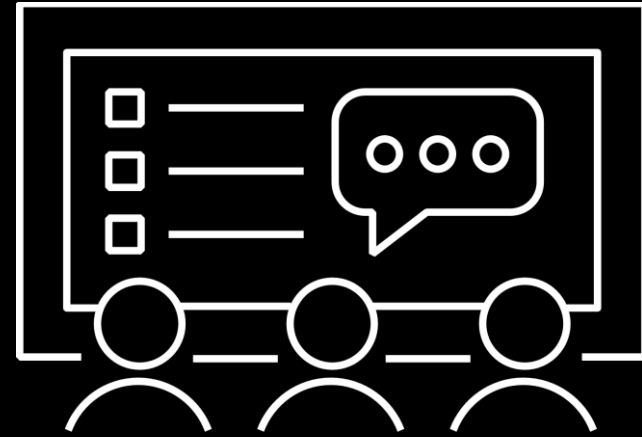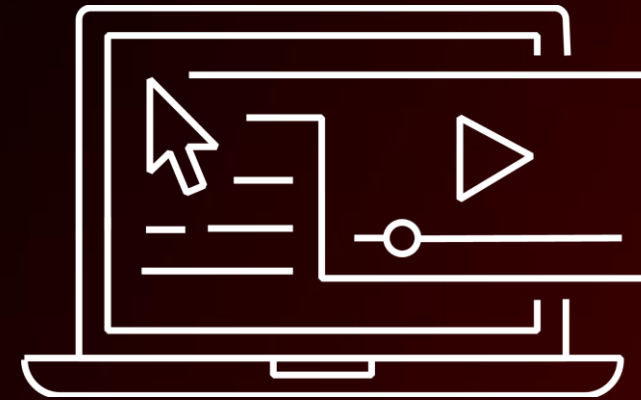- Put your data to work with a modern analytics approach
- … and more!

Visit resource hub

https://tinyurl.com/data-hub-aws

aws

AWS Training and Certification for Data and Analytics

**AWS Data & Analytics
FREE Training Resources**

Discover how to harness
data, one of the world's
most valuable resources,
and innovate at scale.

https://bit.ly/3Ntlhy7

**AWS Data Analytics
Learning Plan**

This learning plan expose you to the
fastest way to get answers from all your
data to all your users. It can also help
prepare you for the AWS Certified Data
Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

**AWS Certified Data
Analytics - Specialty**

Earning AWS Certified Data
Analytics – Specialty
validates expertise in using
AWS data lakes and analytics
services.

https://go.aws/3lwF0RR

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

Thank you!

aws