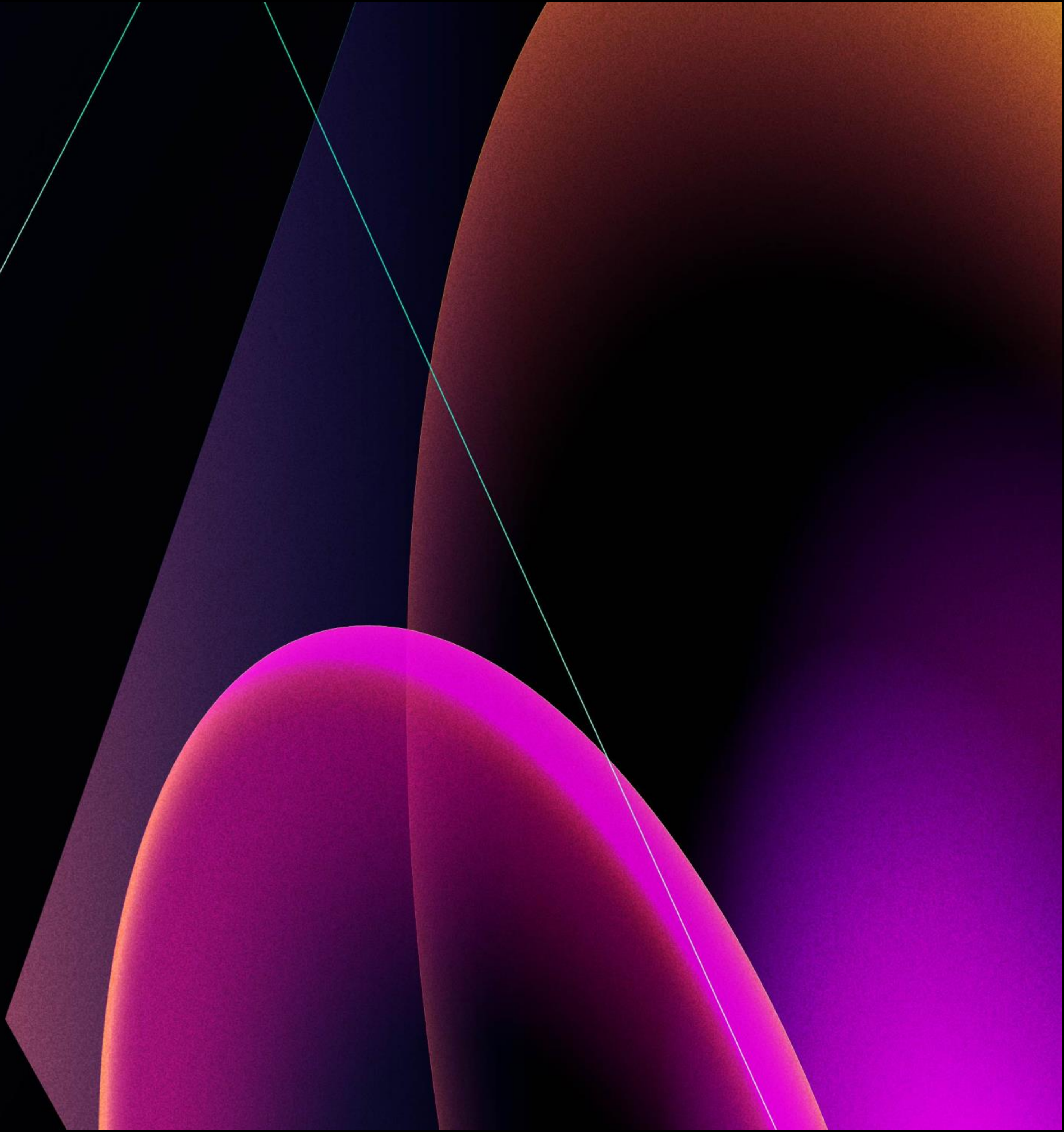
AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV



ANT348

Innovations in AWS analytics: Zero-ETL and data integrations

Paul Van Liew
(he/him)
Director of Engineering
Motive

Jyoti Aggarwal
(she/her)
Senior Product Manager
AWS

Harshida Patel
(she/her)
Principal Specialist SA
AWS

Agenda

01 What is operational analytics?

02 Challenges in traditional data integration

03 AWS analytics innovations

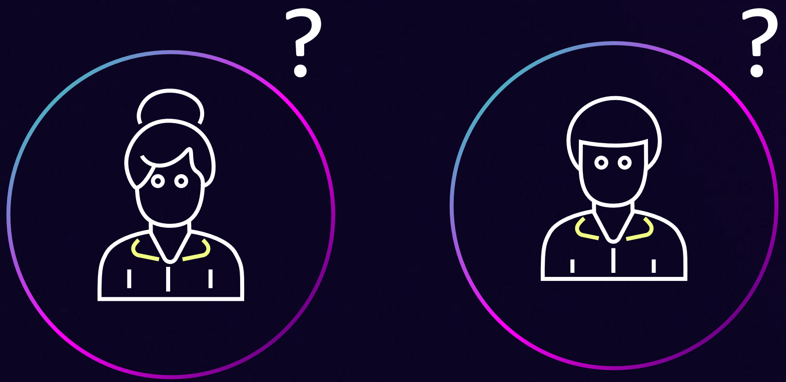
04 Patterns

05 Customer story: Motive

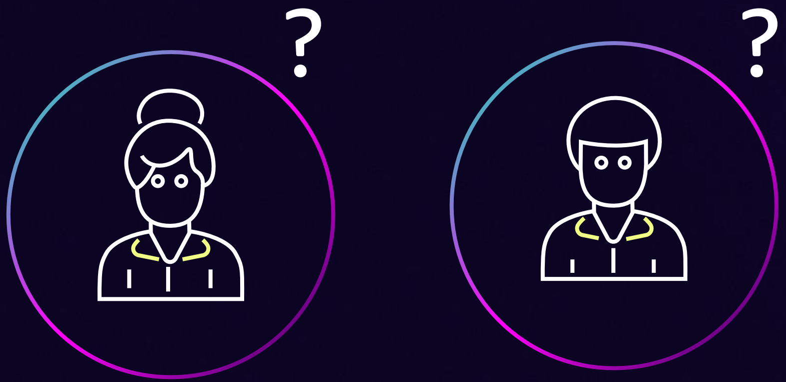
Data drives innovation

- Personalization
- Location optimization
- Internet of Things (IoT)
- Fraud detection
- Inventory optimization
- Anomaly monitoring
- Product insights
- Churn prevention
- Customer relationship management
- Sales operations
- More…
- Gaming leaderboards
- Scoring
- Marketing optimization

By a show of hands . . . who needs to analyze data from one or more operational databases or applications?



By a show of hands . . . who builds, operates, or maintains data pipelines?



Connecting data often requires **complex ETL pipelines**

AWS offers purpose-built databases to meet your demands

Broadest and deepest set of relational and purpose-built databases

Relational

Amazon Aurora

Amazon RDS

Purpose-built

Key-value

Amazon DynamoDB

Caching

Amazon ElastiCache

Document

Amazon DocumentDB

Graph

Amazon Neptune

Memory

Amazon MemoryDB

Wide-column

Amazon Keyspaces

Time-series

Amazon Timestream

Purpose-built data warehouse for analytics at scale

Amazon Redshift

FULLY MANAGED, AI-POWERED CLOUD DATA WAREHOUSING

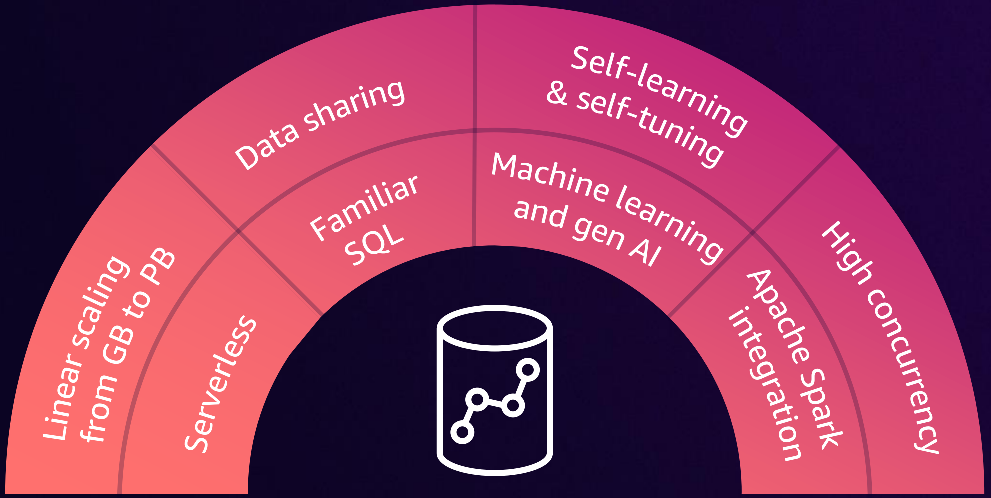
Data

Transactional data

Clickstream

IoT telemetry

Application logs


Linear scaling from GB to PB
Serverless
Data sharing
Familiar SQL
Self-learning & self-tuning
Machine learning and gen AI
High concurrency
Apache Spark integration

Amazon Redshift

Unify data across databases, data lakes, and data warehouses with a **zero-ETL** approach

Best-in-class security, governance, and compliance

Insights

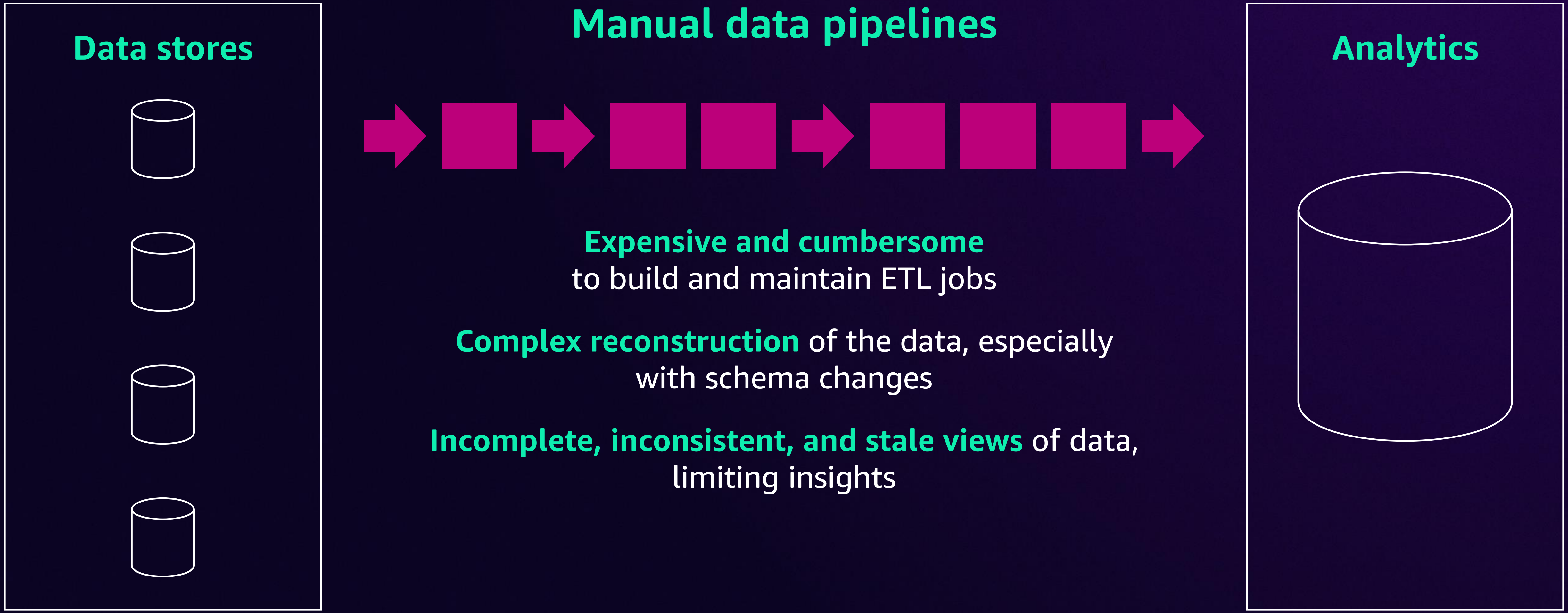
Analyze and visualize data

Deliver real-time & predictive analytics

Build data-driven applications

Existing solutions can be hard!

SEPARATE SYSTEMS REQUIRE COMPLEX DATA MANAGEMENT PIPELINES



Data stores

Manual data pipelines

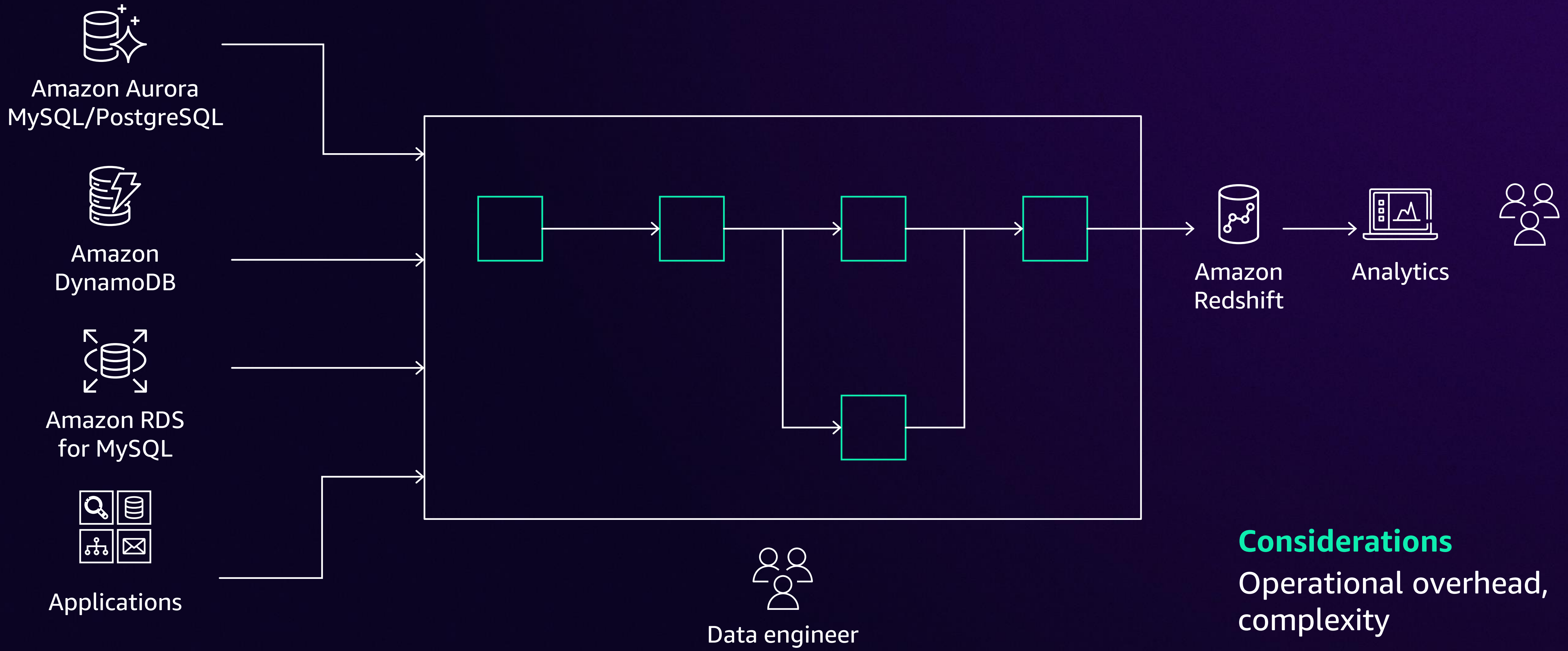
Analytics

Expensive and cumbersome to build and maintain ETL jobs

Complex reconstruction of the data, especially with schema changes

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

There are many reasons to build a data pipeline . . .



Considerations

Operational overhead, complexity

Zero-ETL is fully managed by AWS

Secure

Data is encrypted at rest and in transit

Accurate

Comprehensive data type mapping and DDL replay

Reliable

Resilient processing, with checkpointing and failure mode handling

Efficient

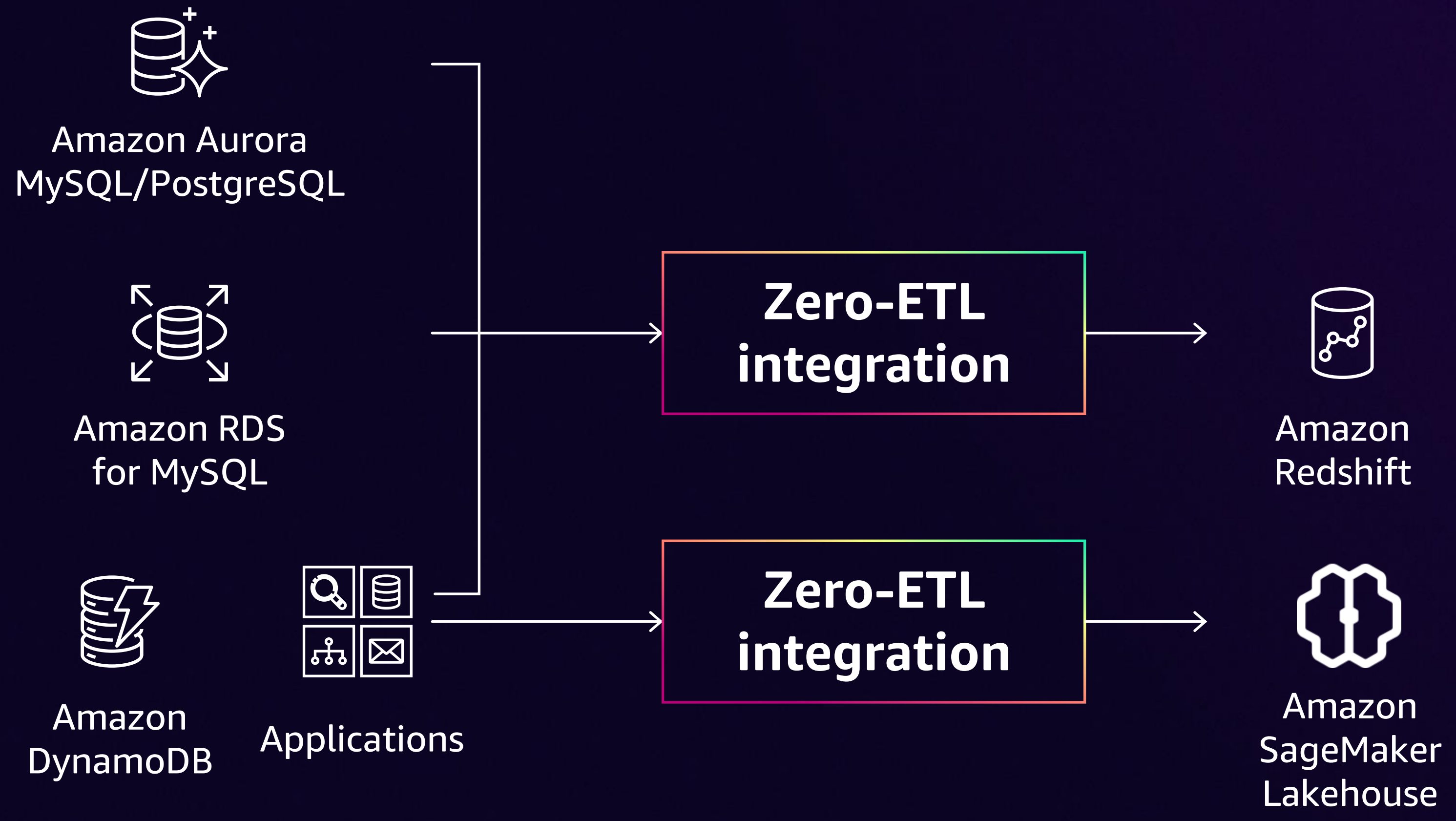
Minimal performance impact to source and destination

Performant

Updates typically reflected in Amazon Redshift within seconds for Aurora and Amazon RDS and in minutes on Amazon DynamoDB

Amazon Redshift supports 12 zero-ETL sources

Simple to set up → Simple to manage → Powerful analytics

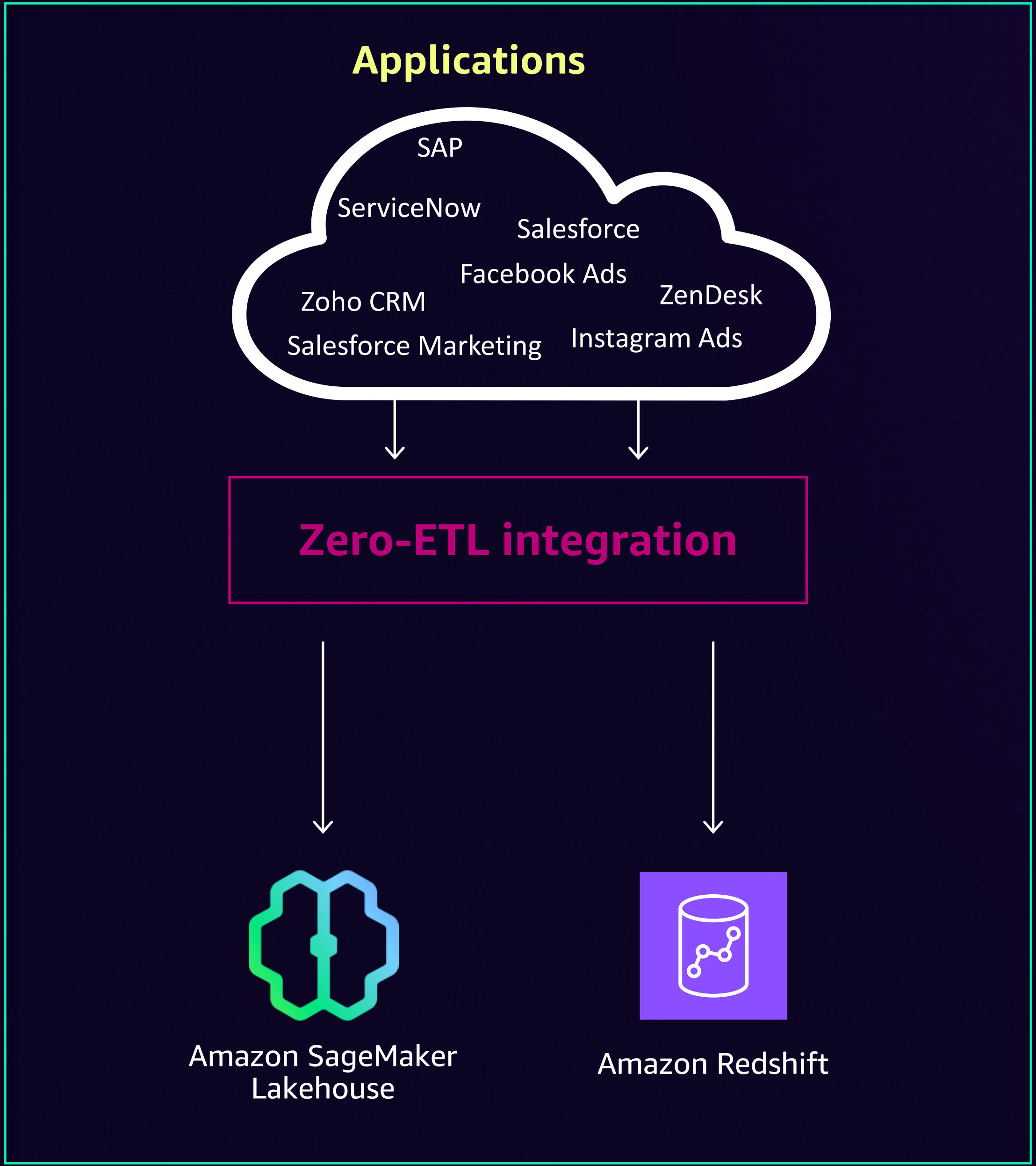


Combine data from many databases and applications into a single data warehouse

Amazon databases – Aurora MySQL, Aurora PostgreSQL, RDS for MySQL, DynamoDB

NEW!

Applications – Salesforce, Zendesk, ServiceNow, SAP, Facebook Ads, Instagram Ads, Salesforce Marketing, Zoho CRM



New

Amazon SageMaker Lakehouse and Amazon Redshift support zero-ETL integrations from eight applications

Simplifies data replication and ingestion from applications to your lakehouse and data warehouse

Accelerates insights by unifying data from applications

Removes undifferentiated, heavy lifting involved in building connectors in-house or using 3rd party services

Reduces costs by removing complex pipeline management and enabling faster decision-making

“

We faced significant data transfer delays, high cost, and complex pipeline management before adopting AWS. With Amazon Aurora zero-ETL integration with Amazon Redshift, we’ve dramatically improved our data processing capabilities. Our reports now execute in just **120 milliseconds, down from 15 seconds previously.** This significant performance boost allows us to scale our operations efficiently while providing better experiences for our customers. The **zero-ETL feature has been particularly transformative**, streamlining our data ingestion and enabling us to gain valuable insights more rapidly.

Sumit Kumar

Sr. Director, Cloud Engineering

firstcry.com

> We have dashboards built on top of our transactional data in Redshift. Earlier, we used our **homegrown solution** to move data from DynamoDB to Redshift but those jobs would often **time out** and lead to **a lot of operational burden** and **missed insights on Redshift**. Using DynamoDB zero-ETL integration with Redshift, we no longer run into such issues, and the integration **seamlessly and continuously replicates data to Redshift.**

Keith McDuffee

Director of DevOps Verisk

How does it work?

Simple to set up

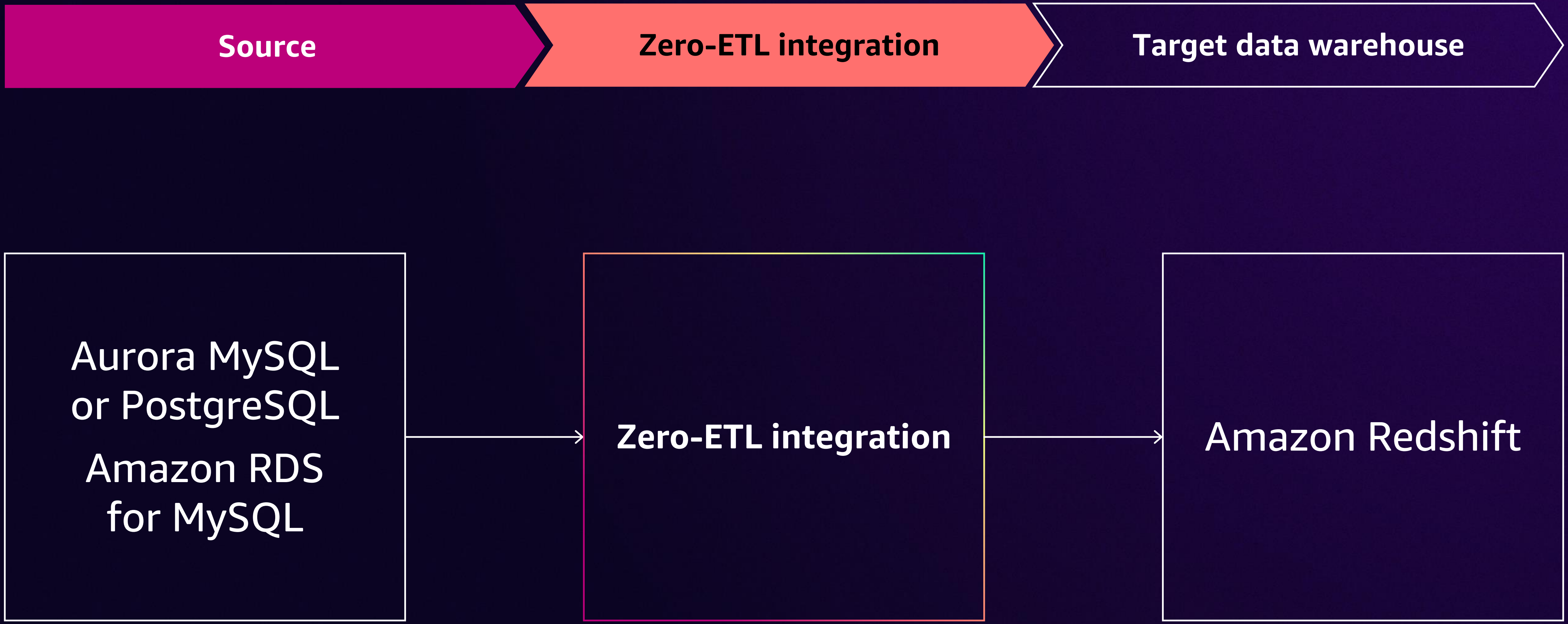
Simple to manage

Powerful analytics

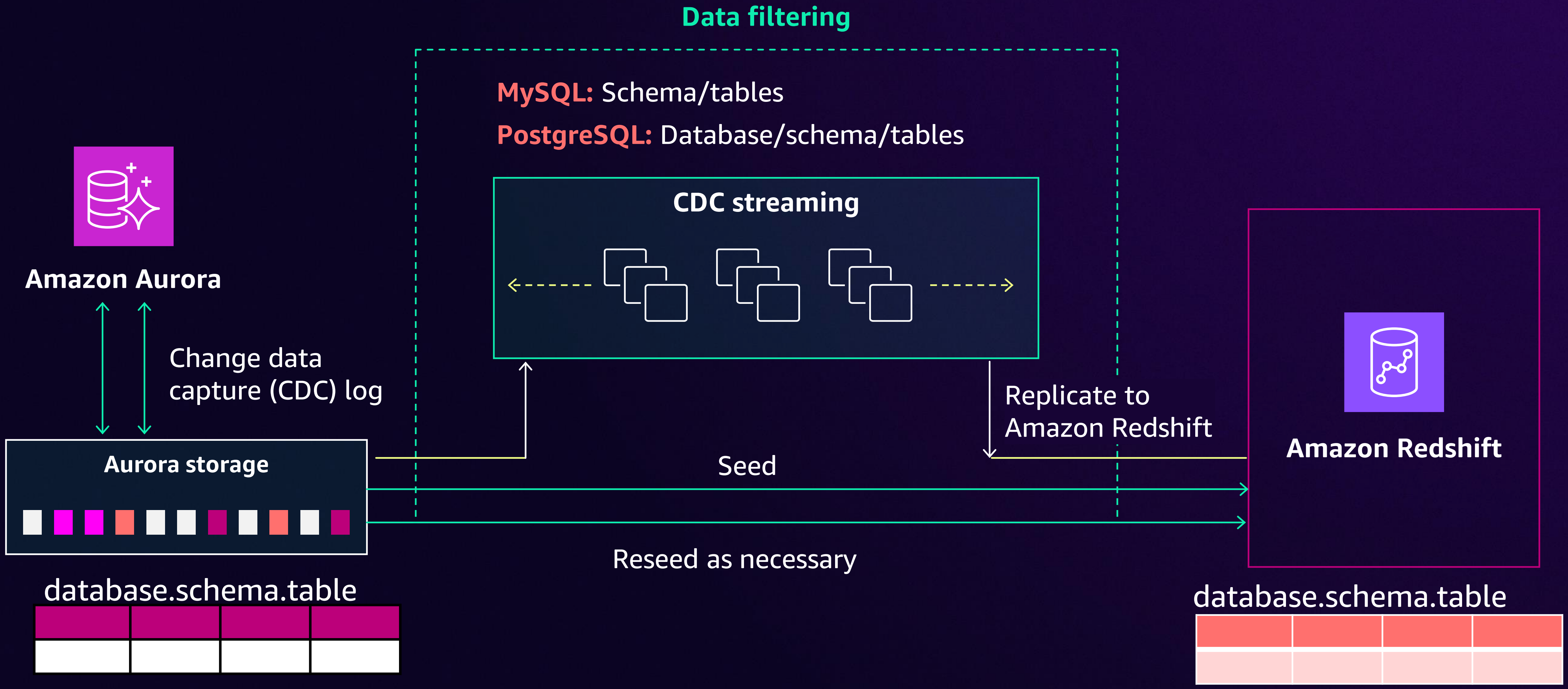
IAM policy-based authorization (flexible policy and cross-account)

Simple setup

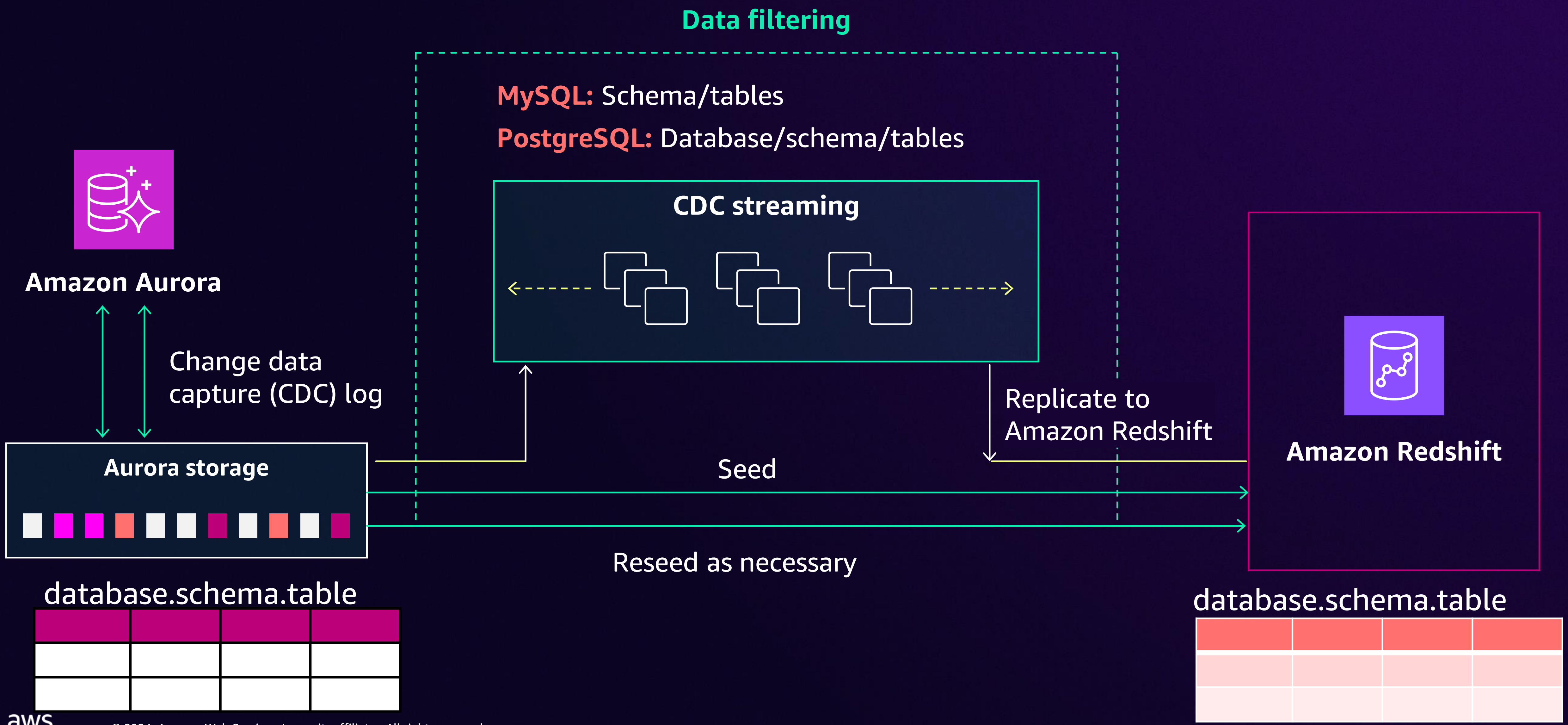
Ongoing monitoring, error reporting, and resiliency


Source
Zero-ETL integration
Target data warehouse
Aurora MySQL or PostgreSQL
Amazon RDS for MySQL
Zero-ETL integration
Amazon Redshift

Zero-ETL replication



Zero-ETL replication



Data filtering

MySQL: Schema/tables

PostgreSQL: Database/schema/tables

CDC streaming

Amazon Aurora

Change data capture (CDC) log

Aurora storage

database.schema.table

Replicate to Amazon Redshift

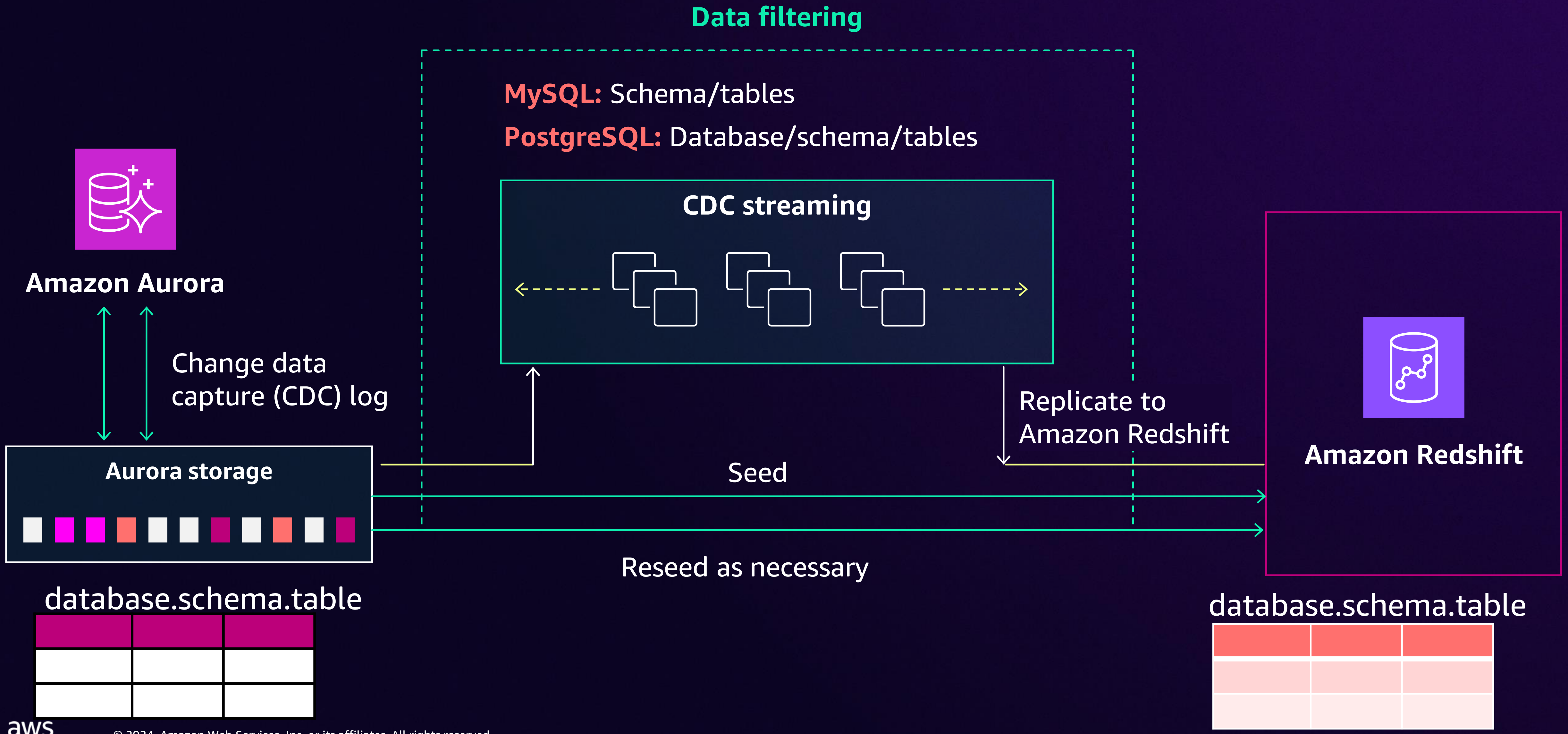
Seed

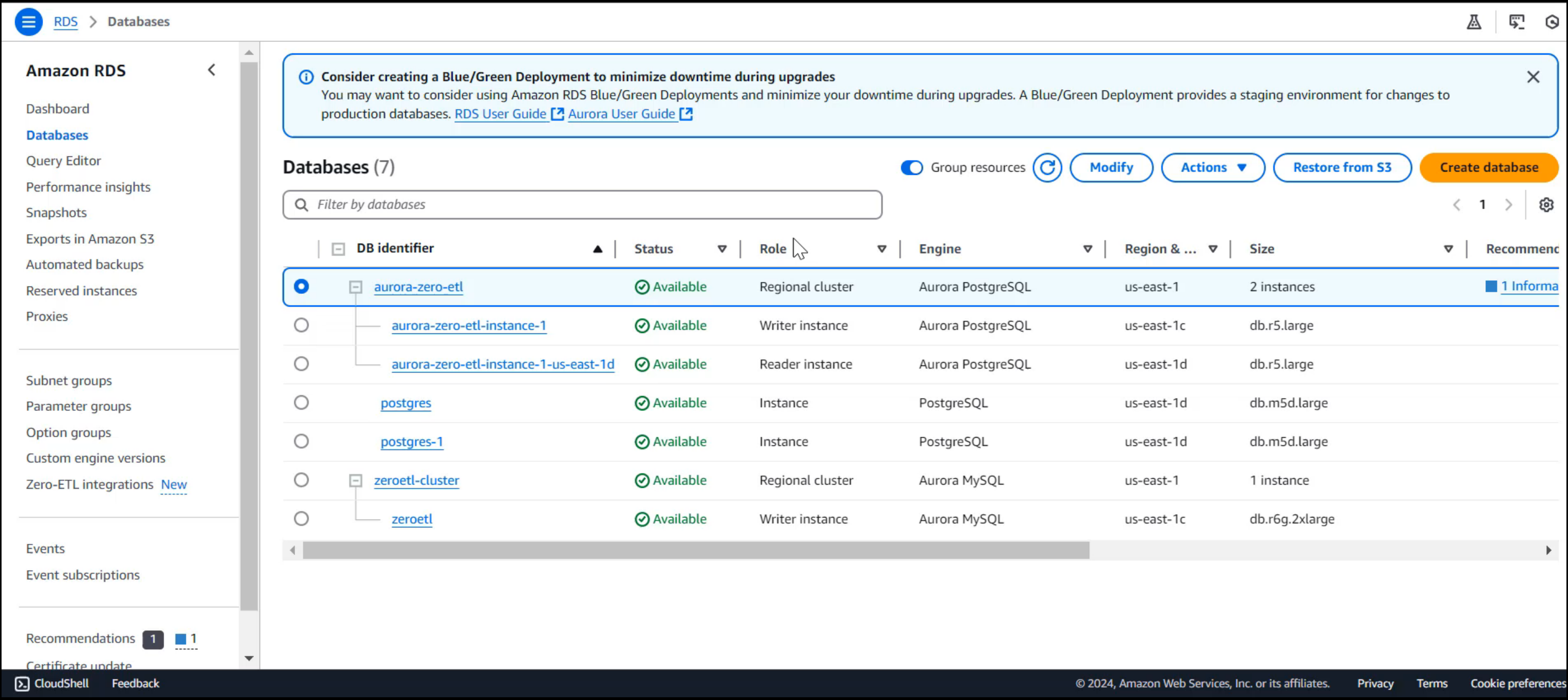
Reseed as necessary

Amazon Redshift

database.schema.table

Zero-ETL replication





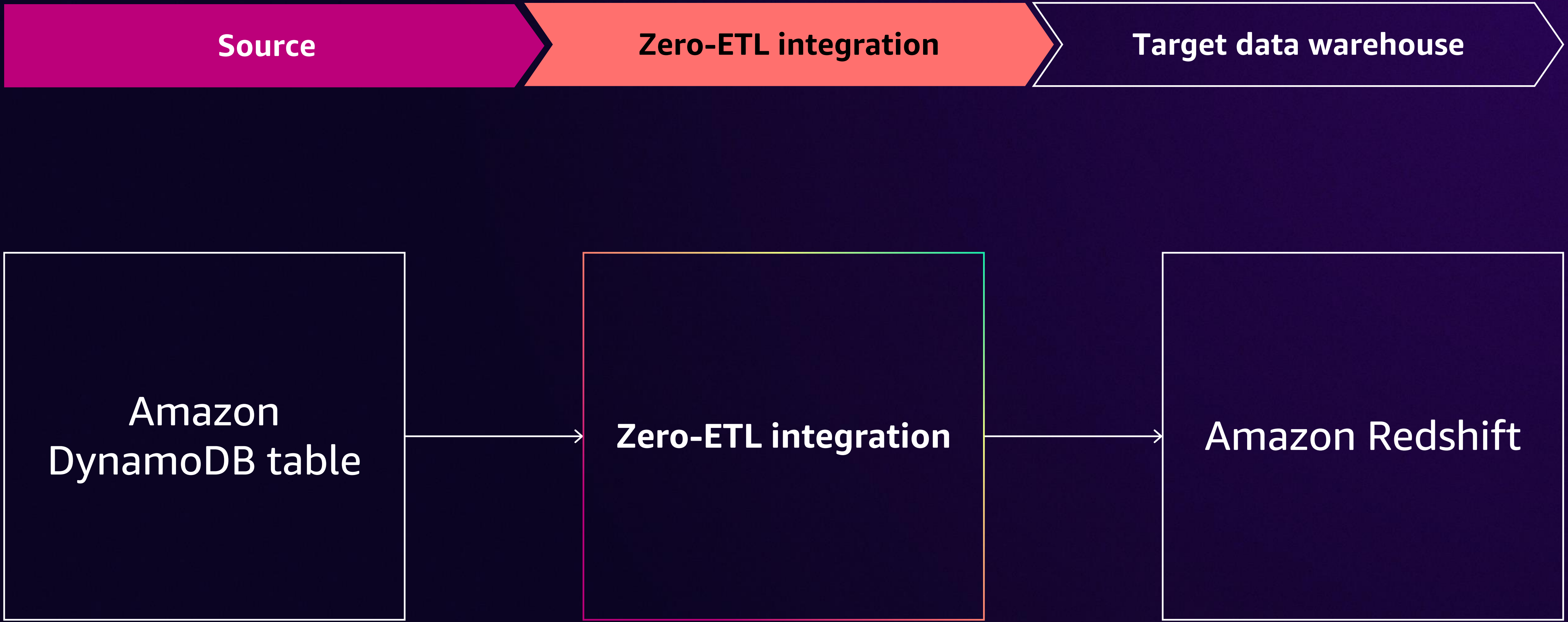
Consider creating a Blue/Green Deployment to minimize downtime during upgrades

You may want to consider using Amazon RDS Blue/Green Deployments and minimize your downtime during upgrades. A Blue/Green Deployment provides a staging environment for changes to production databases. RDS User Guide Aurora User Guide

Databases (7)

	DB identifier	Status	Role	Engine	Region & ...	Size	Recommenc
●	aurora-zero-etl	✓ Available	Regional cluster	Aurora PostgreSQL	us-east-1	2 instances	1 Informa
○	aurora-zero-etl-instance-1	✓ Available	Writer instance	Aurora PostgreSQL	us-east-1c	db.r5.large	
○	aurora-zero-etl-instance-1-us-east-1d	✓ Available	Reader instance	Aurora PostgreSQL	us-east-1d	db.r5.large	
○	postgres	✓ Available	Instance	PostgreSQL	us-east-1d	db.m5d.large	
○	postgres-1	✓ Available	Instance	PostgreSQL	us-east-1d	db.m5d.large	
○	zeroetl-cluster	✓ Available	Regional cluster	Aurora MySQL	us-east-1	1 instance	
○	zeroetl	✓ Available	Writer instance	Aurora MySQL	us-east-1c	db.r6g.2xlarge	

© 2024, Amazon Web Services, Inc. or its affiliates.



Source

Zero-ETL integration

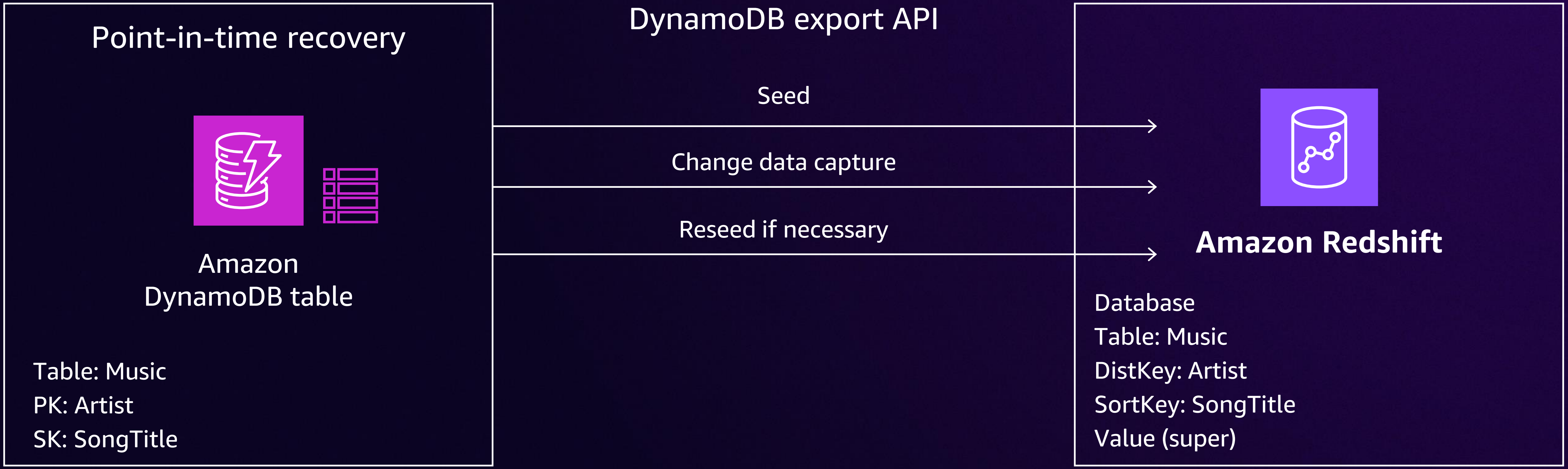
Target data warehouse

Amazon DynamoDB table

Zero-ETL integration

Amazon Redshift

Zero-ETL replication



Point-in-time recovery

Amazon DynamoDB table

Table: Music

PK: Artist

SK: SongTitle

DynamoDB export API

Seed

Change data capture

Reseed if necessary

Amazon Redshift

Database

Table: Music

DistKey: Artist

SortKey: SongTitle

Value (super)

Item

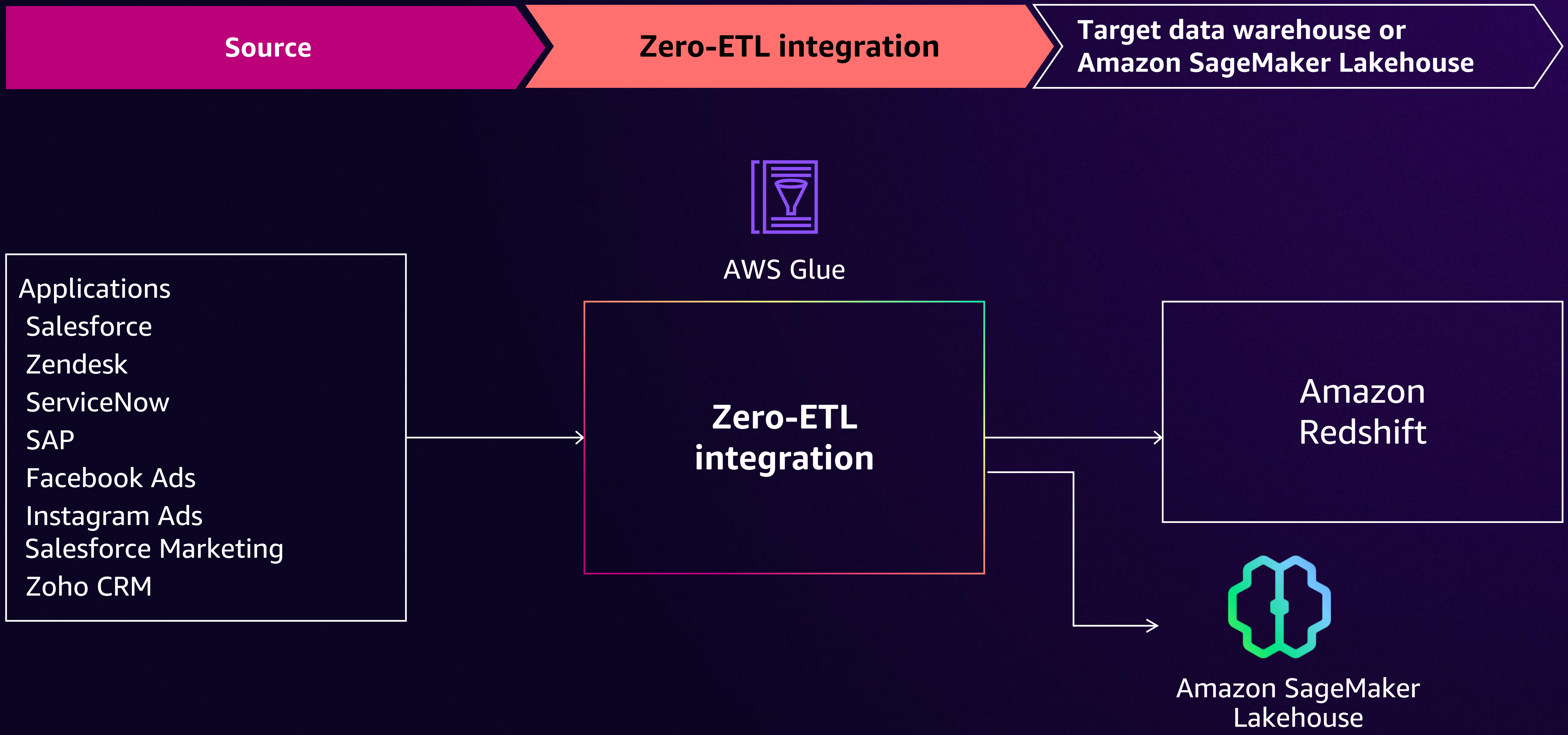
Artist	Song Title	Genre	Rating
Yohani	Jab we met	Folk	1

Row

Artist	Song Title	Value
Yohani	Jab we met	{"Artist":{"S":"Yohani"},"genre":{"S":"folk music"},"rating":{"N":"1"},"songTitle":{"S":"jab we met"}}

Innovations in zero-ETL

	Aurora MySQL/PostgreSQL RDS for MySQL	Amazon DynamoDB
Data filtering	✓	
Refresh interval	0 seconds to 5 days	15 minutes to 5 days
Incremental materialized view with auto refresh	✓	✓
Customer managed sort key	✓	✓
Lag	Seconds	15 to 30 minutes
Integrations per source	5 for Aurora MySQL 1 for Aurora PostgreSQL 1 for RDS for MySQL	1 table



Source

Zero-ETL integration

Target data warehouse or Amazon SageMaker Lakehouse

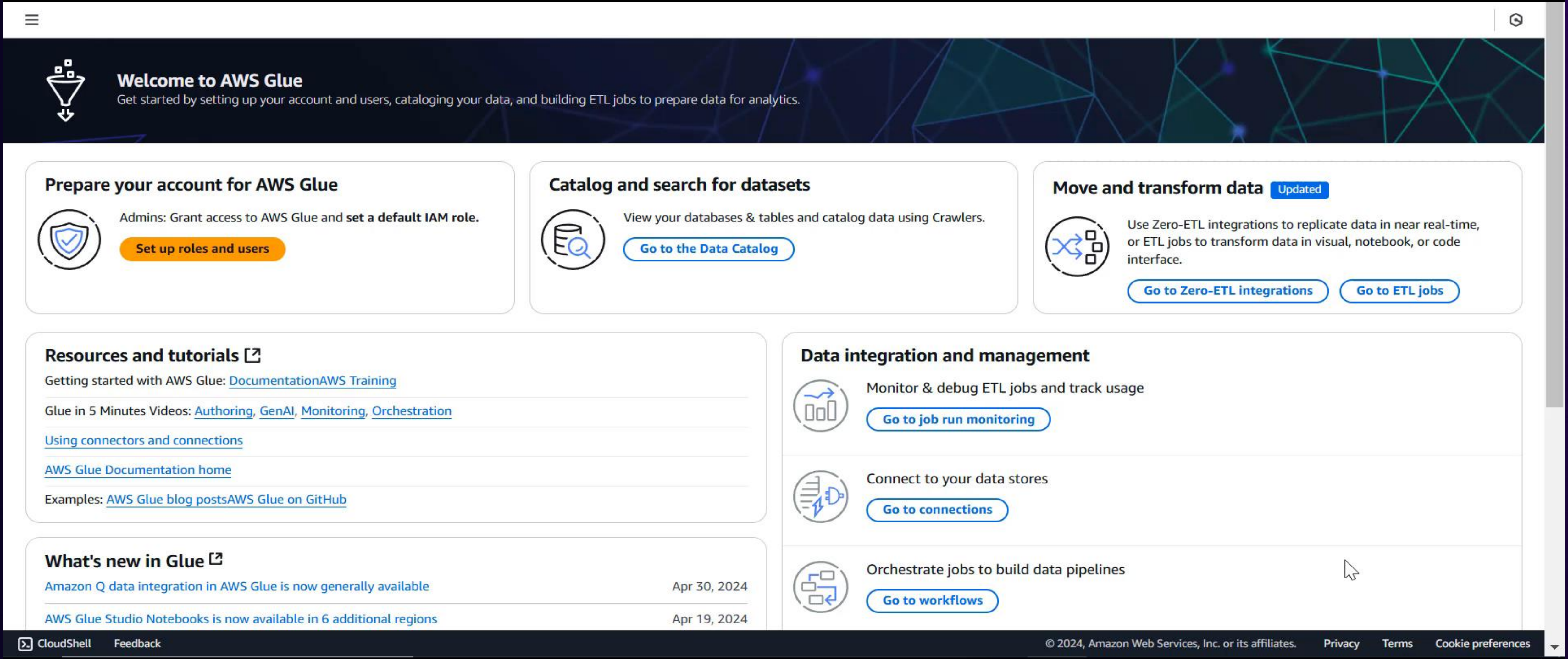
Applications
- Salesforce
- Zendesk
- ServiceNow
- SAP
- Facebook Ads
- Instagram Ads
- Salesforce Marketing
- Zoho CRM

AWS Glue

Zero-ETL integration

Amazon Redshift

Amazon SageMaker Lakehouse



Welcome to AWS Glue

Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Prepare your account for AWS Glue

Admins: Grant access to AWS Glue and **set a default IAM role.**

Catalog and search for datasets

View your databases & tables and catalog data using Crawlers.

Move and transform data Updated

Use Zero-ETL integrations to replicate data in near real-time, or ETL jobs to transform data in visual, notebook, or code interface.

Resources and tutorials

Getting started with AWS Glue: DocumentationAWS Training

Glue in 5 Minutes Videos: Authoring, GenAI, Monitoring, Orchestration

Using connectors and connections

AWS Glue Documentation home

Examples: AWS Glue blog postsAWS Glue on GitHub

What's new in Glue

Amazon Q data integration in AWS Glue is now generally available	Apr 30, 2024
AWS Glue Studio Notebooks is now available in 6 additional regions	Apr 19, 2024

Data integration and management

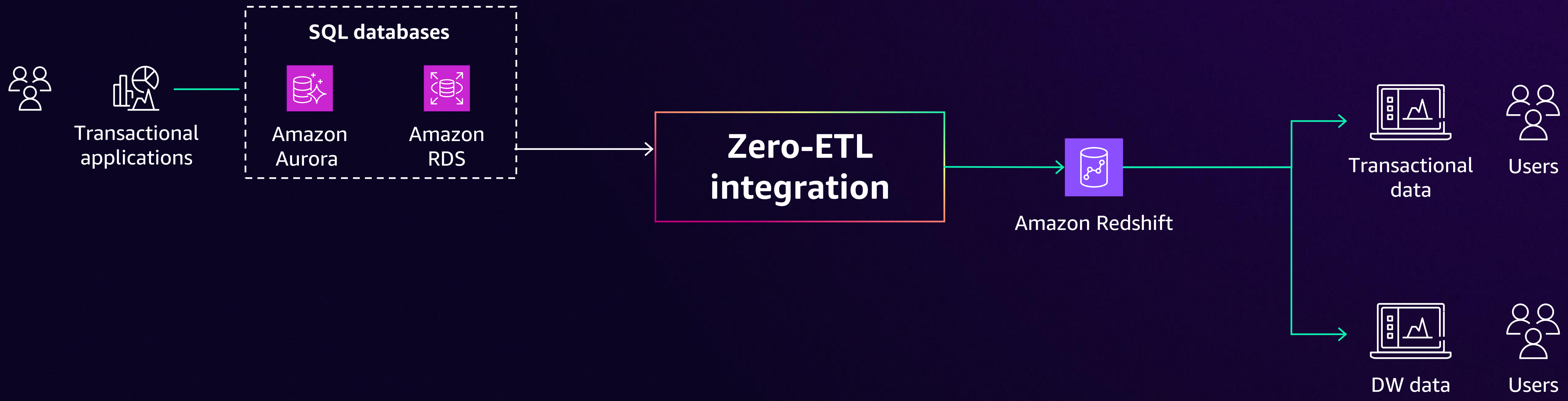
Monitor & debug ETL jobs and track usage

Connect to your data stores

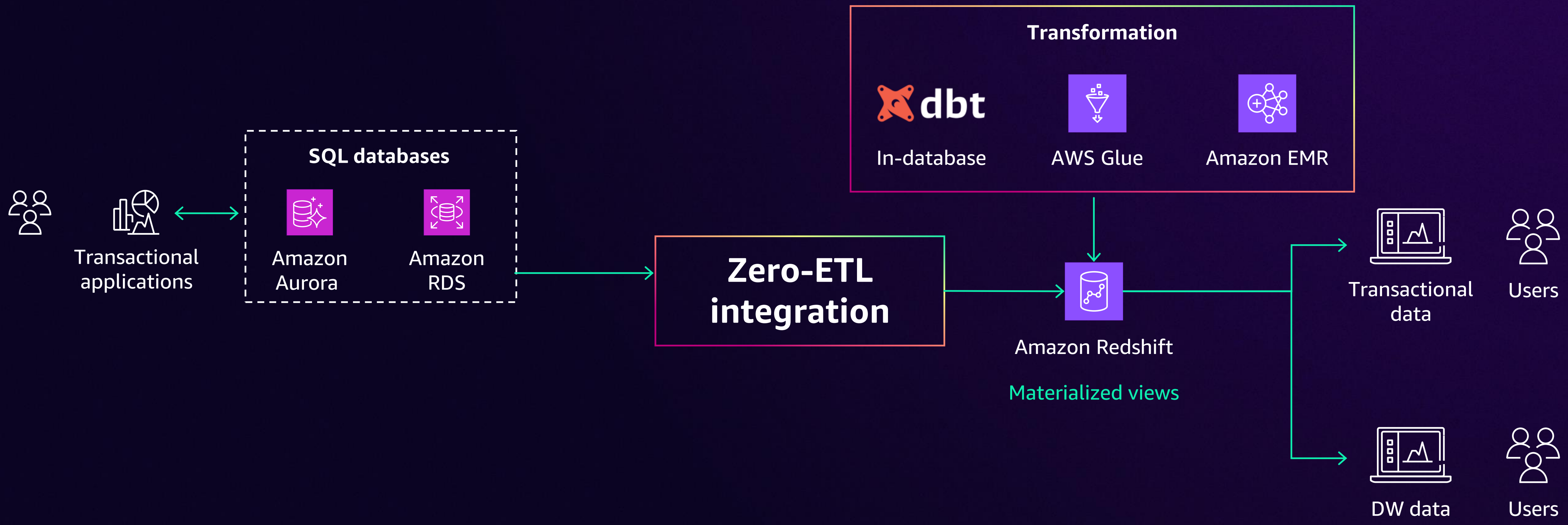
Orchestrate jobs to build data pipelines

Patterns

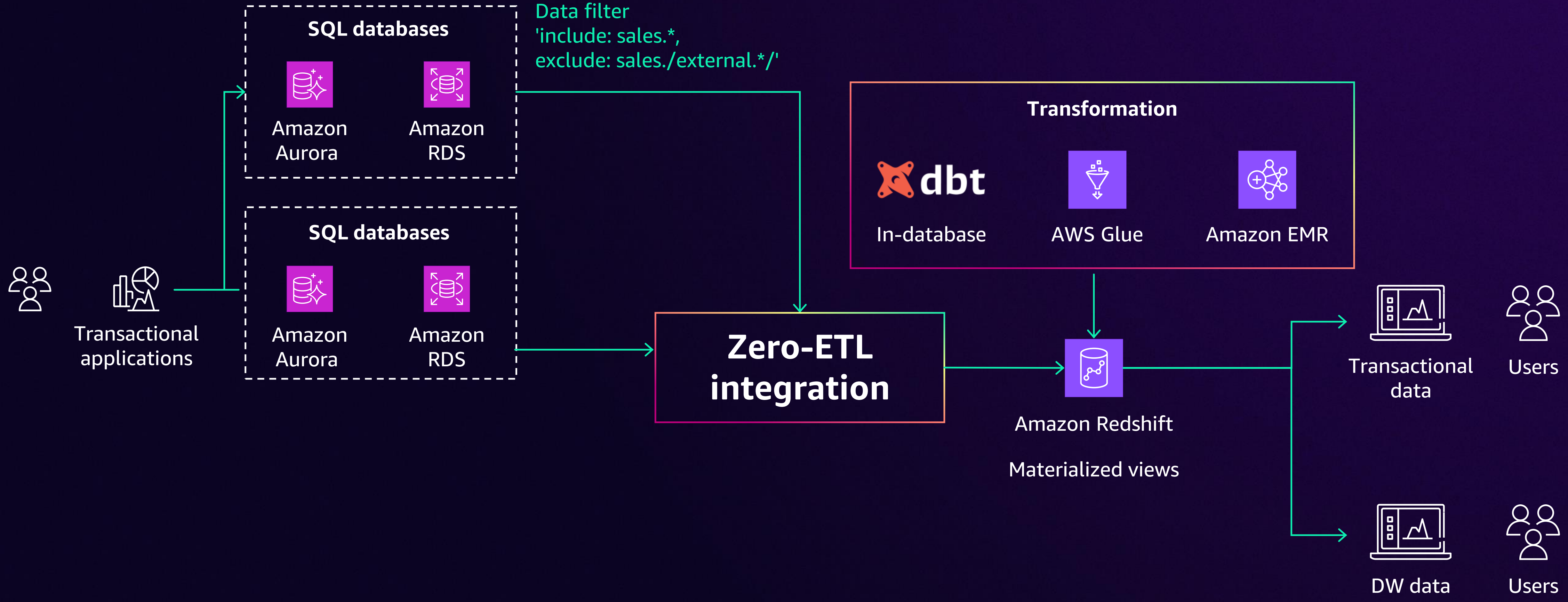
Starting your journey



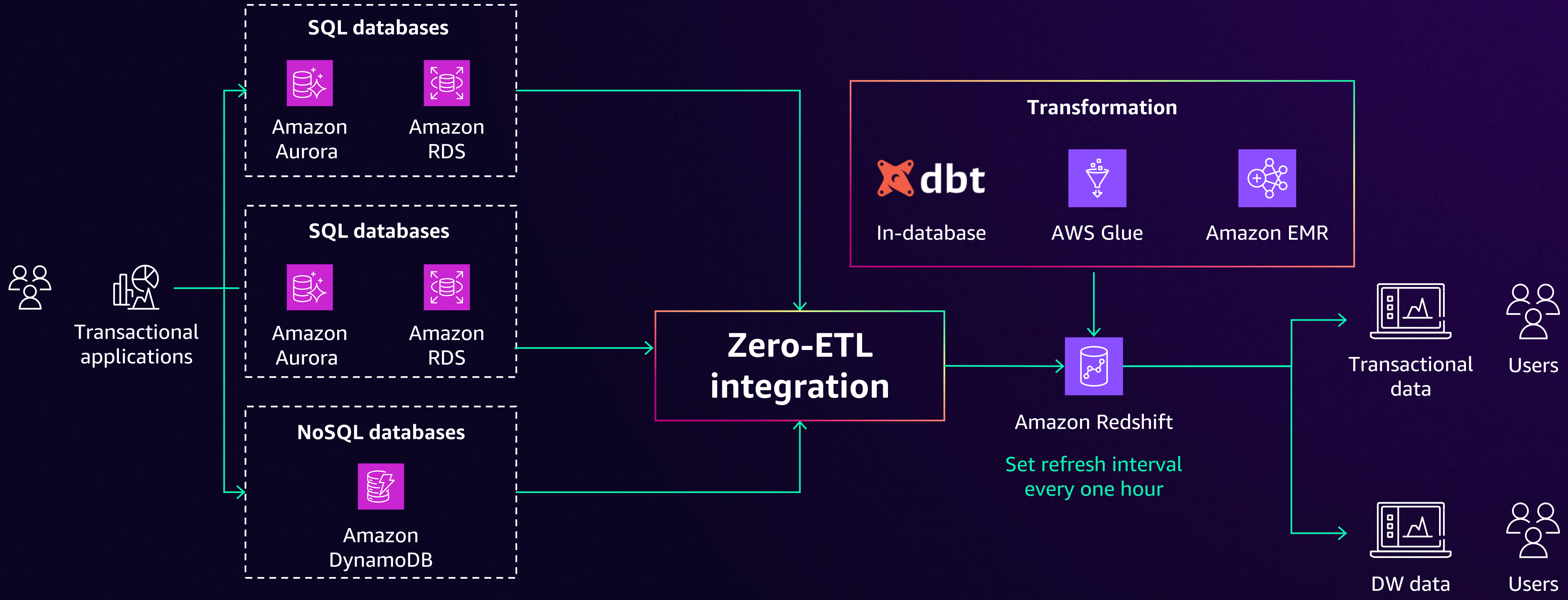
Starting your journey – Transformation



Add new sources – Be selective

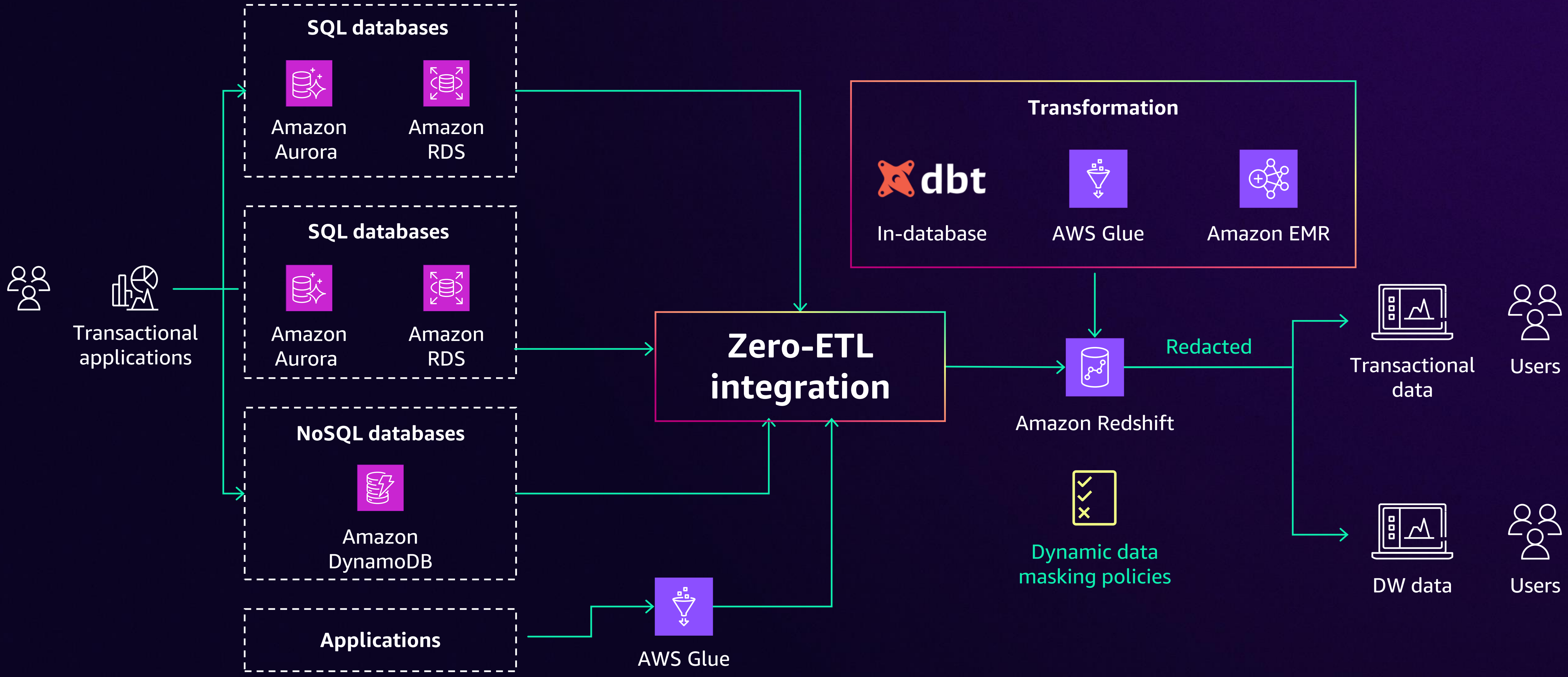


Add new sources – Refresh at same interval

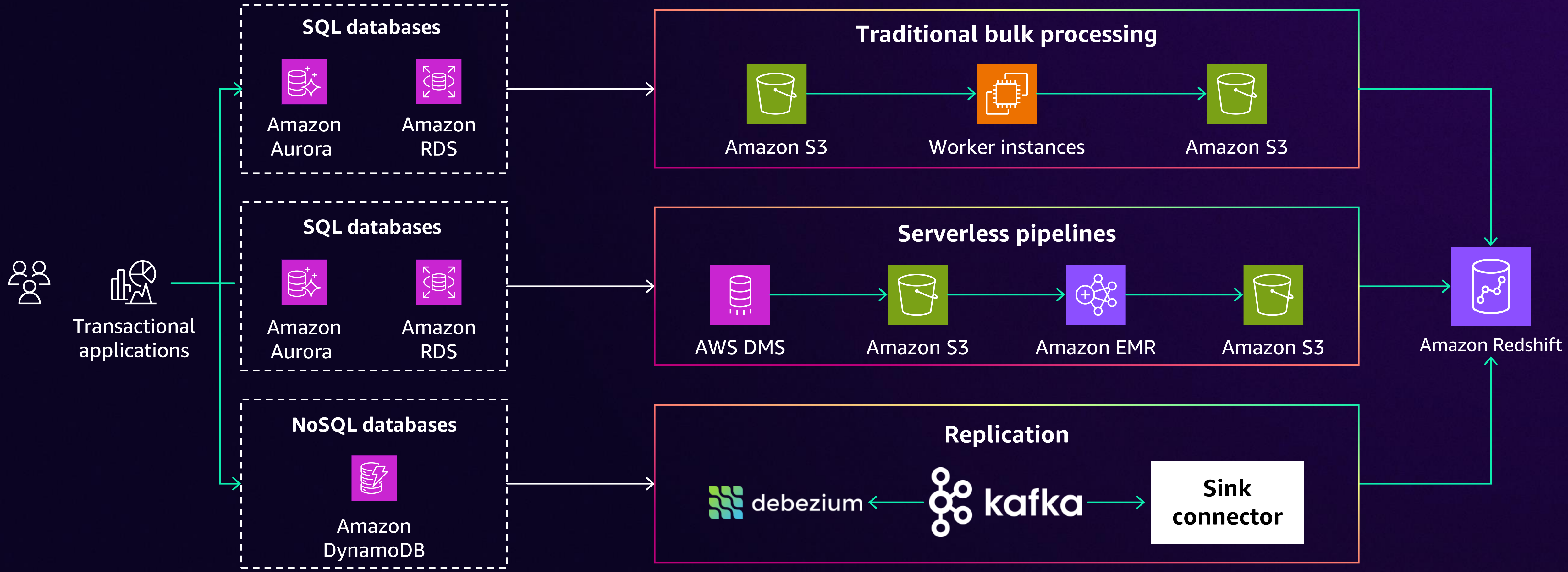


Set refresh interval every one hour

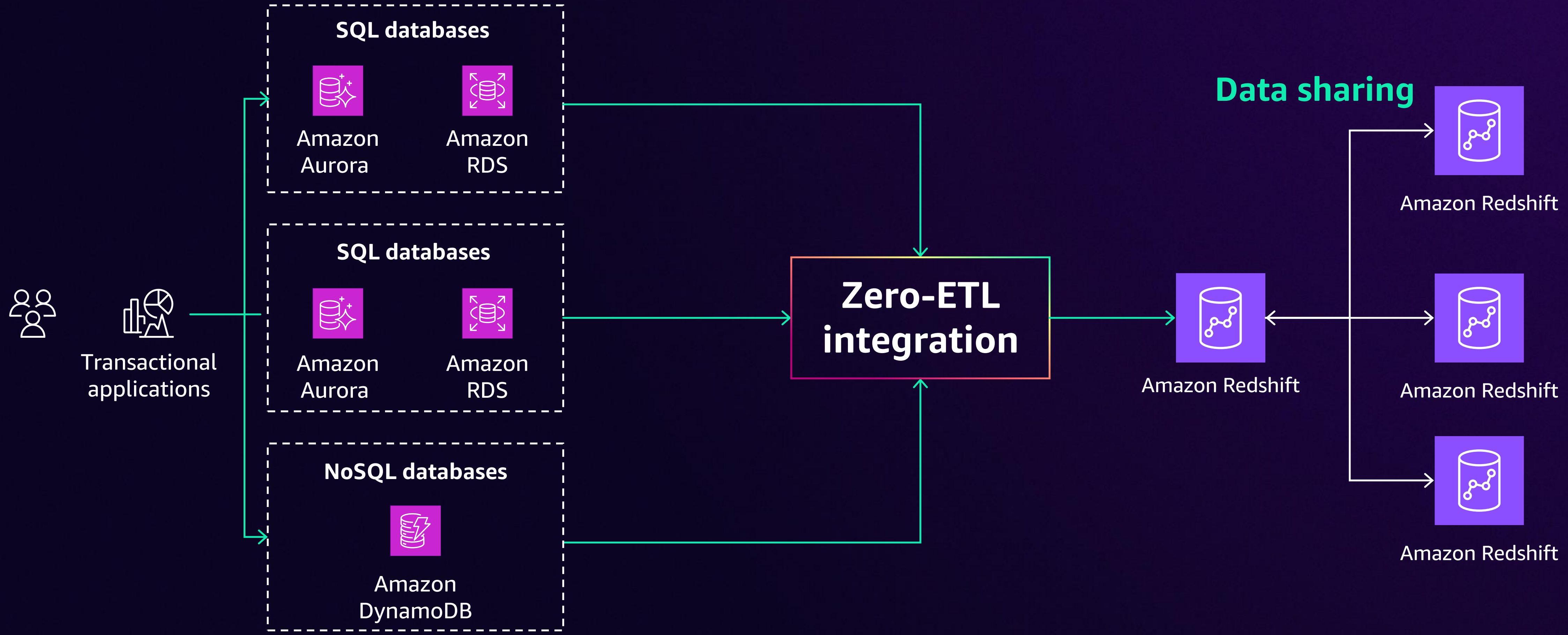
Mask sensitive columns



Existing legacy ETL pipeline

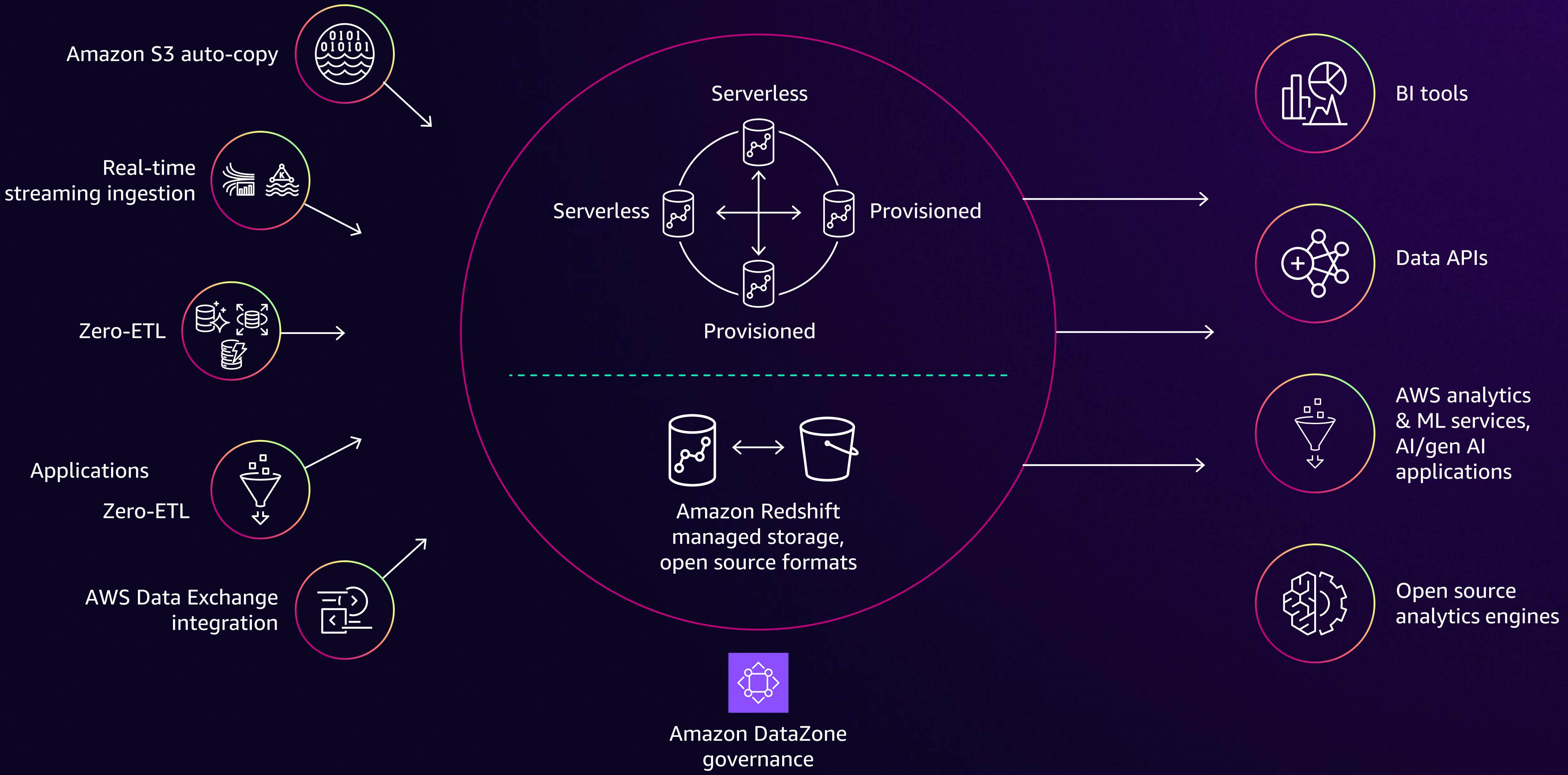


Simplify using zero-ETL



Analytics on all data

Analytics for all users



Amazon S3 auto-copy

Real-time streaming ingestion

Zero-ETL

Applications Zero-ETL

AWS Data Exchange integration

Serverless

Serverless

Provisioned

Provisioned

Amazon Redshift managed storage, open source formats

Amazon DataZone governance

BI tools

Data APIs

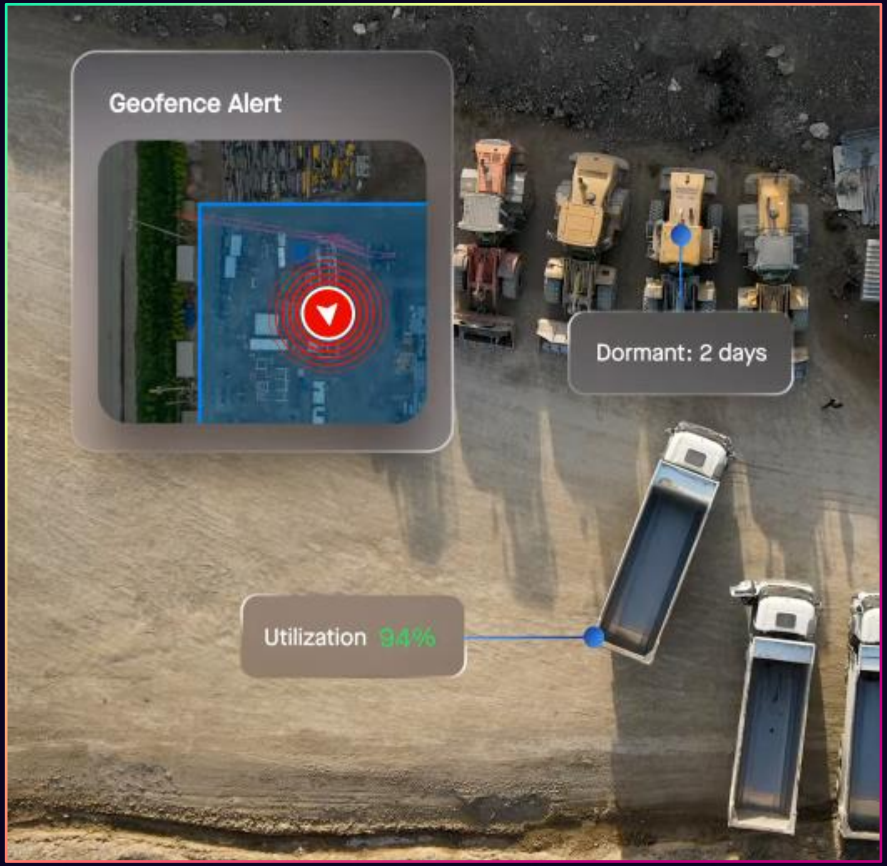
AWS analytics & ML services, AI/gen AI applications

Open source analytics engines

Fraud detection with zero-ETL at Motive

Motive

EMPOWERING THE PEOPLE WHO RUN THE PHYSICAL ECONOMY



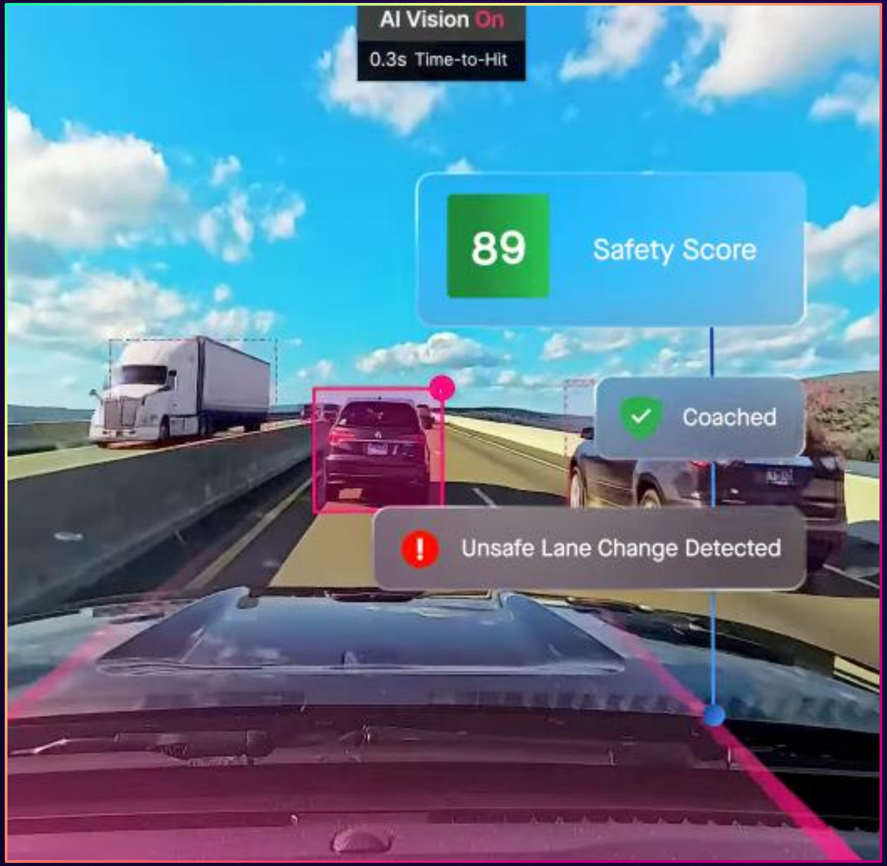
AI Dashcam

Fleet management

Safety & compliance

Tracking & telematics

Spend management



Platform at Motive

Paul Van Liew

Director, Platform Engineering

Data Platform

ML Platform

App Arch

DevProd

IoT

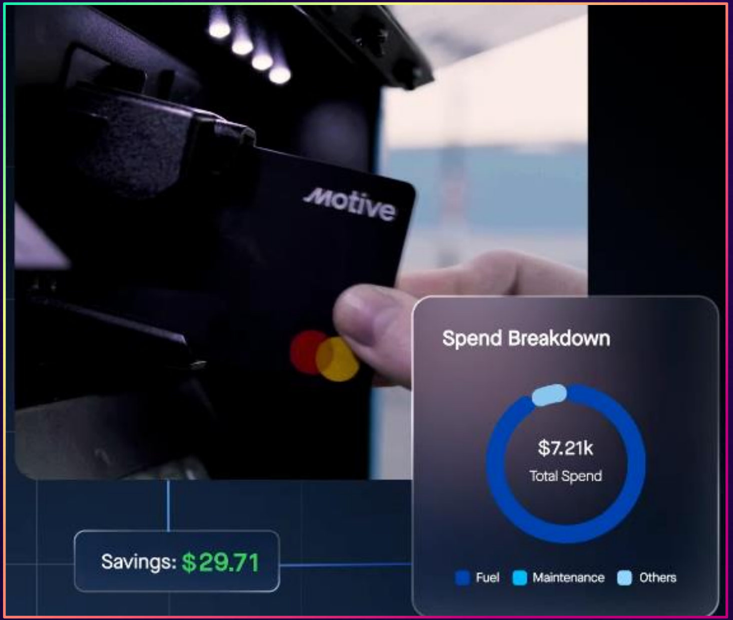
DBA

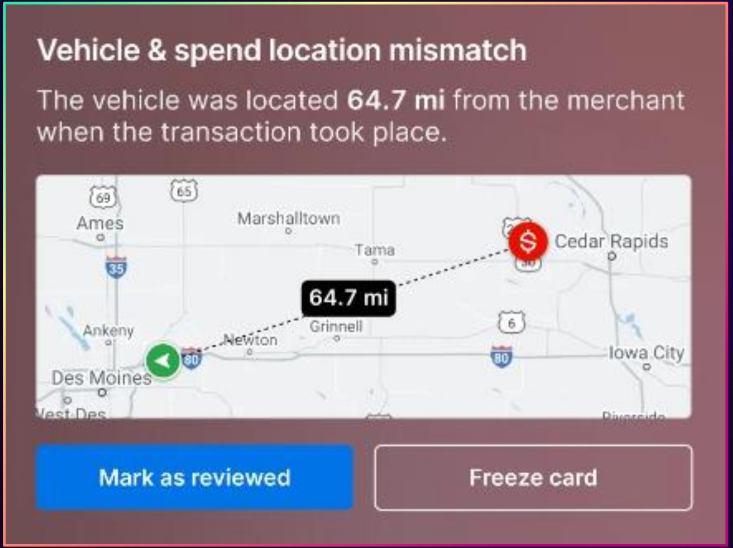
Security

Infra/SRE

Fleet card fraud detection

- Based on fuel events, location, and others
- Real-time transaction blocking
- Background detection (6 hours+)
- ML and analytics to advance our strategy





Vehicle & spend location mismatch

The vehicle was located **64.7 mi** from the merchant when the transaction took place.

Detection source data

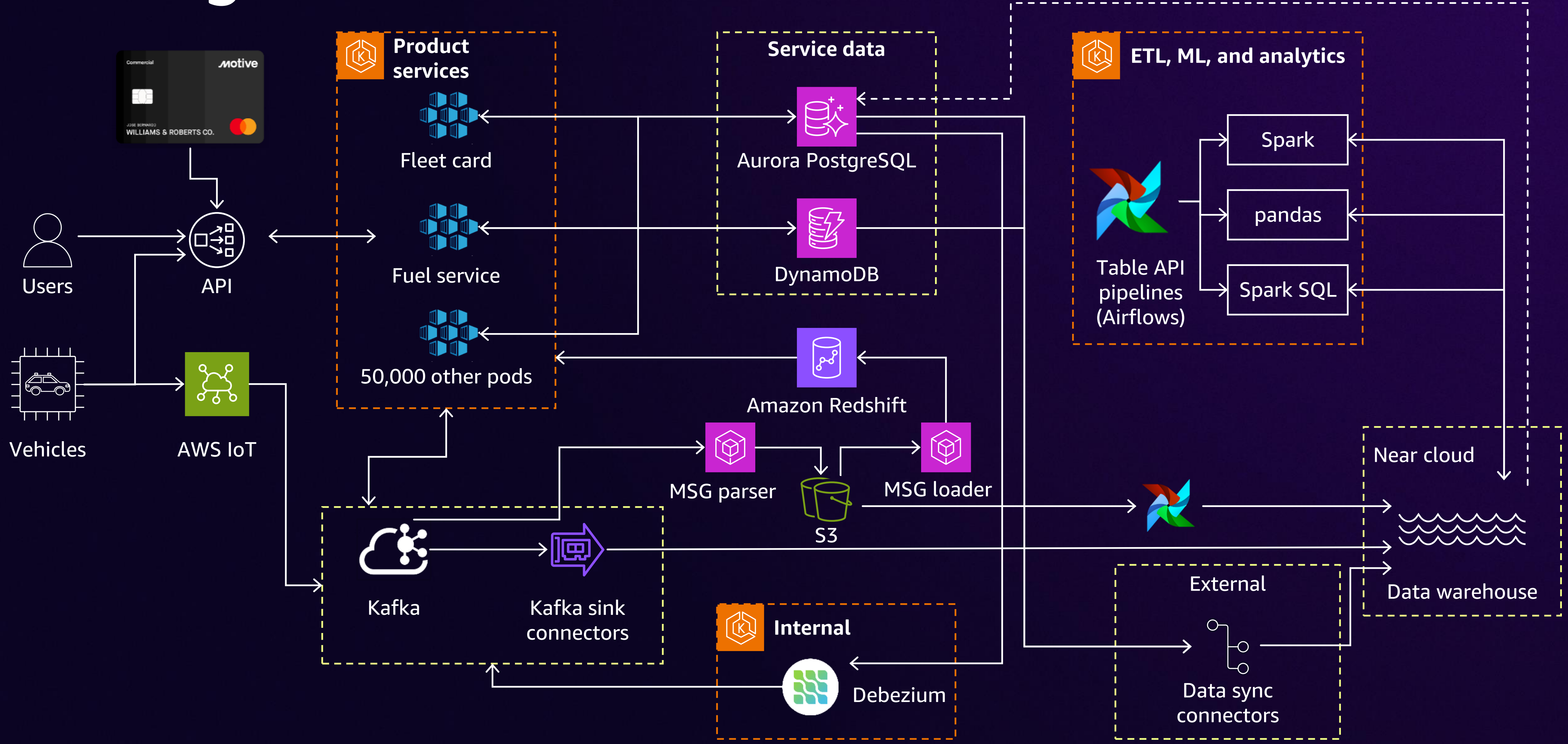
Aurora PostgreSQL – Transaction and fuel events – 11 TB – 1 GB/day

- VARCHAR, TIMESTAMPS, and VARIANTS, oh my!

DynamoDB – Vehicle locations, metadata, fraud profiles

- 9 GB of JSON, geolocation, metadata, etc.
- Main location table is 250 TB – too much for now

Existing architecture – Cards and fuel events



Challenges

Cost and complexity

- Data duplication, multiple sync methods, network cost

Speed

- Cadenced ETL processes and syncs – 45 minutes, 3 hours, 6 hours

Maintenance

- Multiple teams: Reliability, replication slots, provisioning, visibility

Goals

Simplify – Remove multiple sync methods and work

Stream – Achieve sub-minute delays

Support – Direct app usage of internal data warehouse

Save – Cost savings are always a nice bonus!

Zero-ETL to the rescue

Easy choice

- Tight integrations, easy compliance, all internal

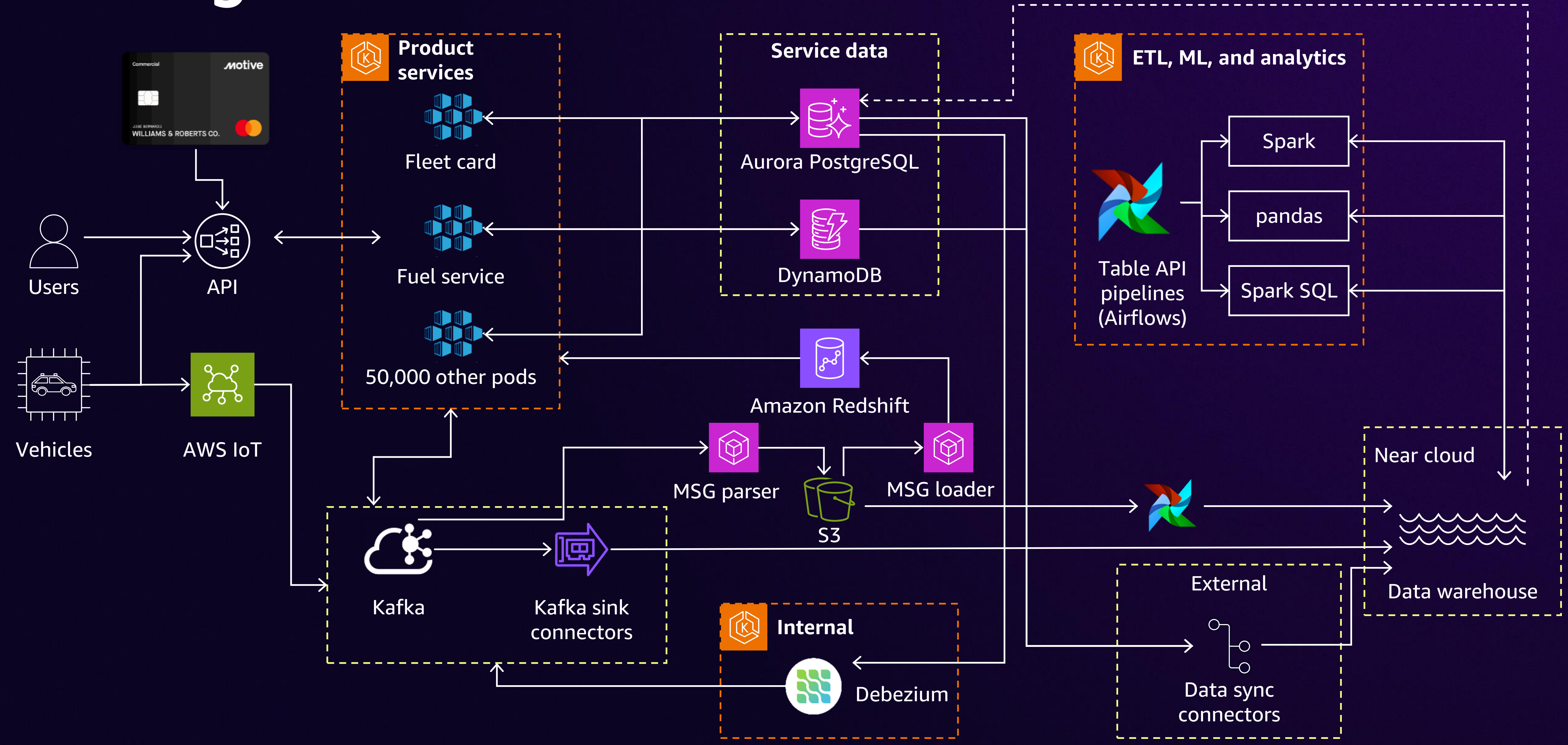
Simple setup

- Provisioning is a breeze
  - Guided info – DynamoDB PITR, permissions
    - Automation helpers: Fix it for you, Amazon Redshift DB creation

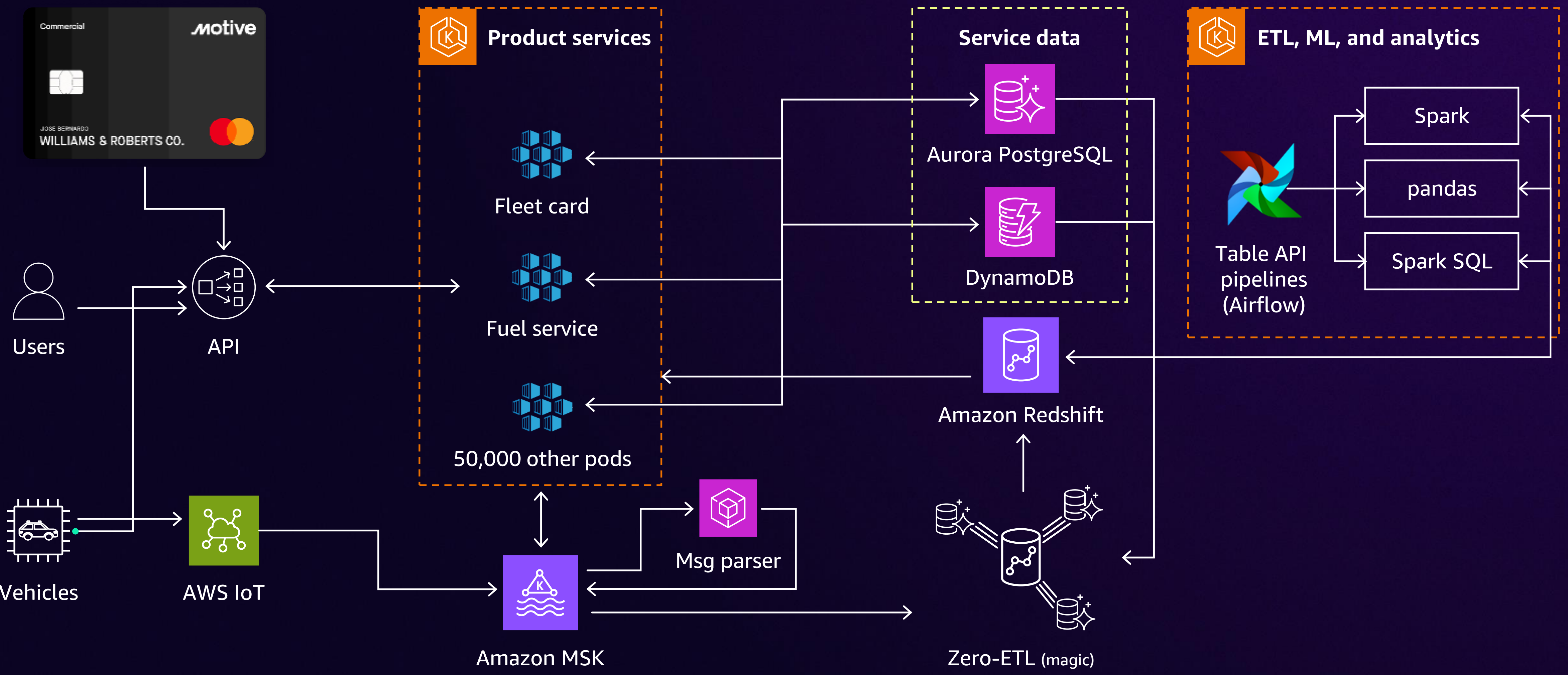
Filtering

- Including future tables is huge

Existing architecture – Cards and fuel events



New architecture with zero-ETL



Results

- From 4 sync methods to 1
- Latency – 15s and 15m (vs. 45m and 3h+)
- Effort reduced – No maintenance, less provisioning
- Better visibility
- Reduced costs
  - $120K/year removed of connectors, ETL and DWH compute, Kafka

What’s next?

- Zero-ETL all the things!
  - Amazon MSK for all & DB upgrades to PostgreSQL 16.4+
- Finish main vehicle ELD message pipeline (30 TB Kafka topic – 200 MB/s peak)
  - Amazon Managed Service for Apache Flink → Amazon MSK → Zero-ETL
    - >$750K/year savings estimated
- Amazon SageMaker Lakehouse Apache Iceberg API to ease DWH migration while zero-ETL-ing

“It takes a team, and a bit of magic.

Thanks to the Motive and AWS teams – Data Platform

Tianyao Zhang, Burhan Ateeq, Pushkar Pande, Angelica Heeney, Uzair Ahmad

Thank you!

Please complete the session survey in the mobile app

Paul Van Liew

Jyoti Aggarwal

Harshida Patel