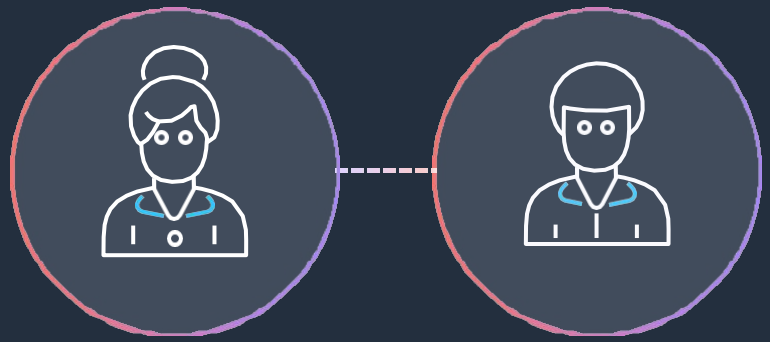
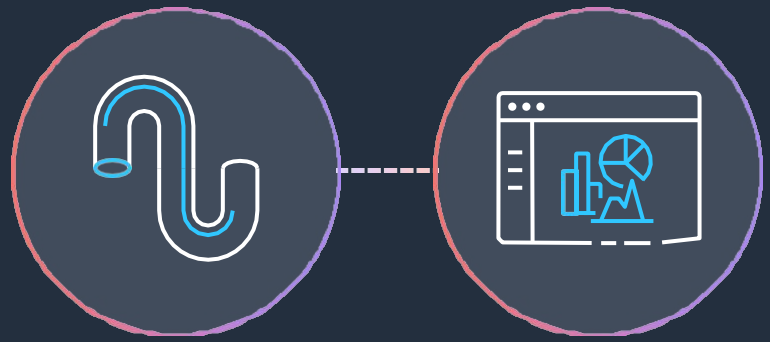
Zero ETL on AWS

Accelerating Insights, Eliminating Complexity



Are you a **data engineer** or **data analyst**?



Do you worry about **data pipelines** or analytics on operational data?

Agenda

Data pipelines: overview and challenges

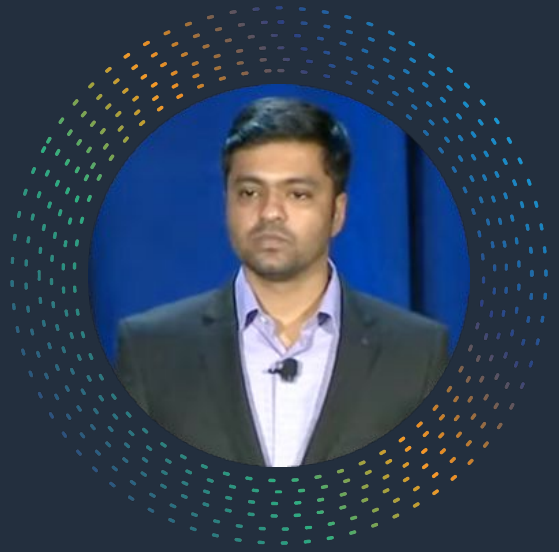
AWS helps simplify data pipelines

AWS Zero-ETL features

Demo

Resources

Speaker



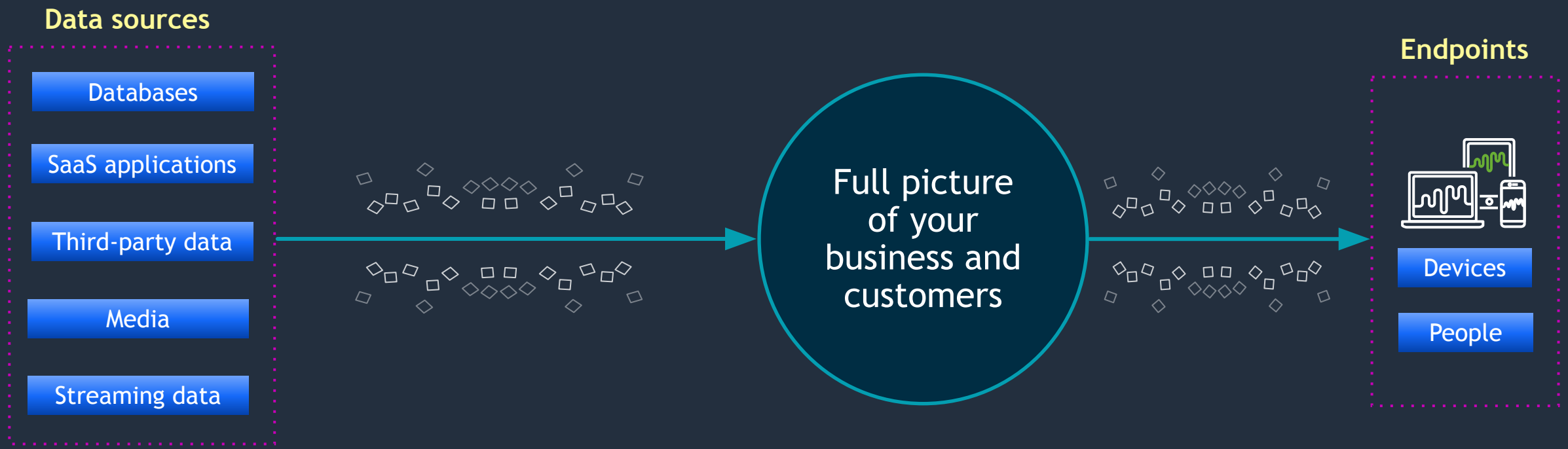
Sanchit Jain

FOLLOW ME

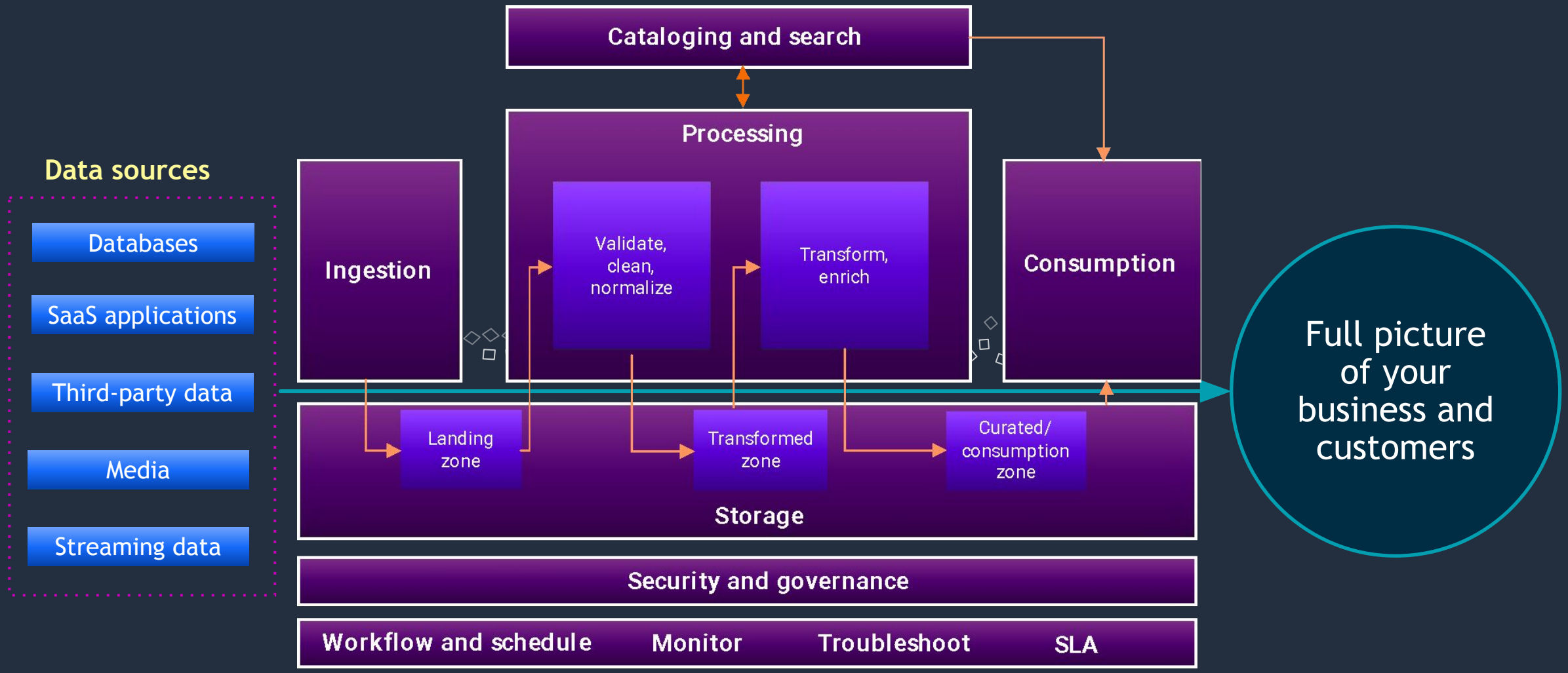
- Working as Data Analytics & Cloud Practice Lead - AWS Alliance at Quantiphi
- Recognition - AWS Hero & AWS Ambassador, MVP by Multicloud4u & AWS User Group Mumbai Lead
- Speaker at various events like AWS re:Invent, AWS Summit, AWS User Groups, AWS Community Days, and various educational institutes
- Follow me @ LinkedIn, Twitter or Github

Data pipelines: overview and challenges

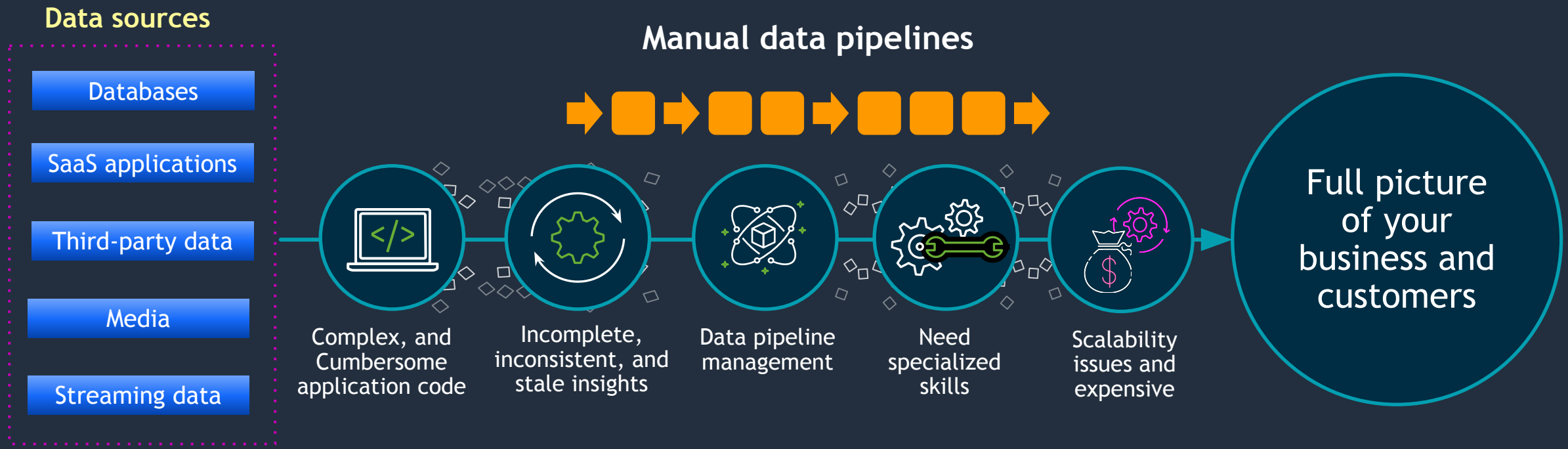
Data integration gives you a full picture of your business



Data integration and processing overview



Data integration and processing challenges



AWS helps simplify data pipelines

Easily connect to all your data

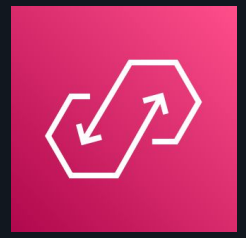
Services for
connecting to 100s of data sources

AWS Glue for
value-add data
transformations and more

Zero-ETL future enabled
by service integrations

Services for connecting to 100s of data sources

50+



Amazon Appflow

30+

Amazon Kinesis Data Firehose

25+

Amazon Athena

40+

Amazon SageMaker

30+

Amazon QuickSight

AWS DataSync

300+

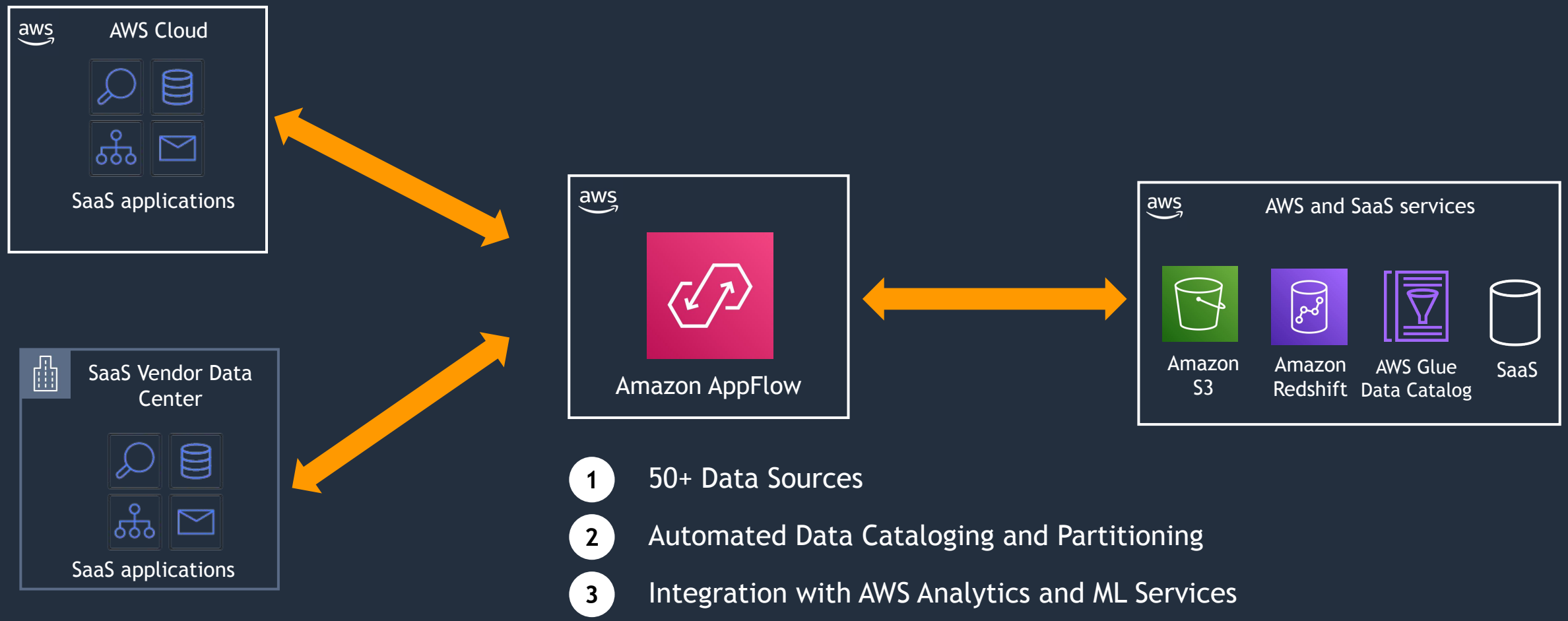
AWS Data Exchange

70+

AWS Glue

Integrating SaaS applications data

Amazon AppFlow makes it easy to ingest SaaS applications data into storage layer

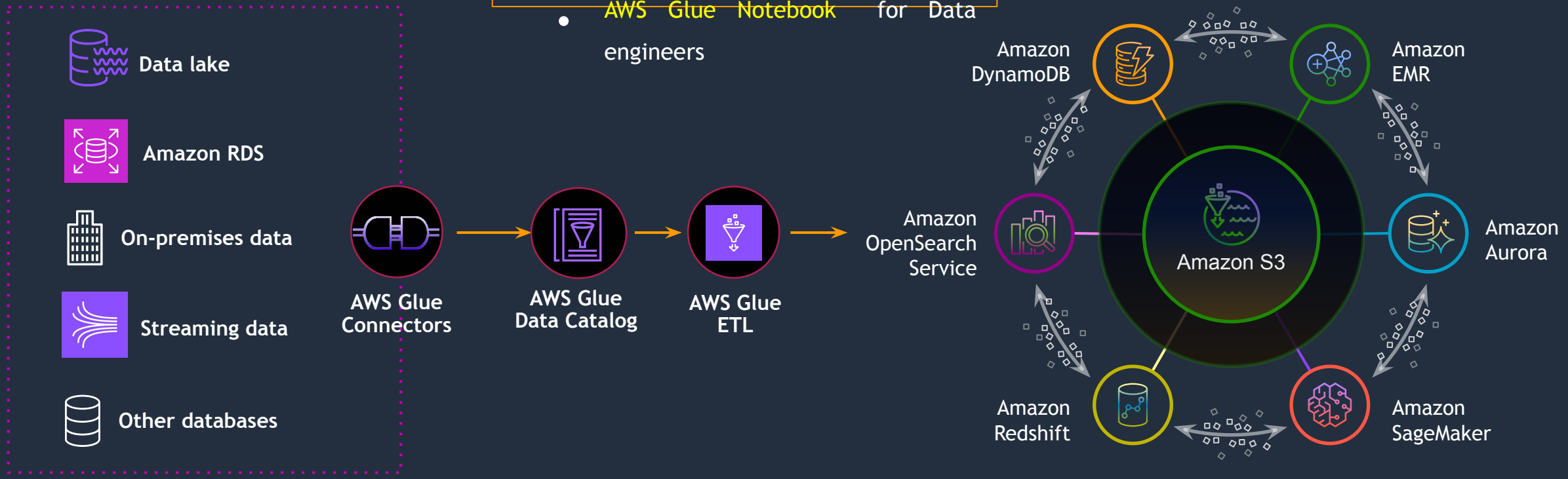


1. 50+ Data Sources
2. Automated Data Cataloging and Partitioning
3. Integration with AWS Analytics and ML Services

Seamless data integration with AWS Glue

Persona-specific tools

- AWS Glue Studio for ETL developers
- AWS Glue DataBrew for Business analysts
- AWS Glue Notebook for Data engineers



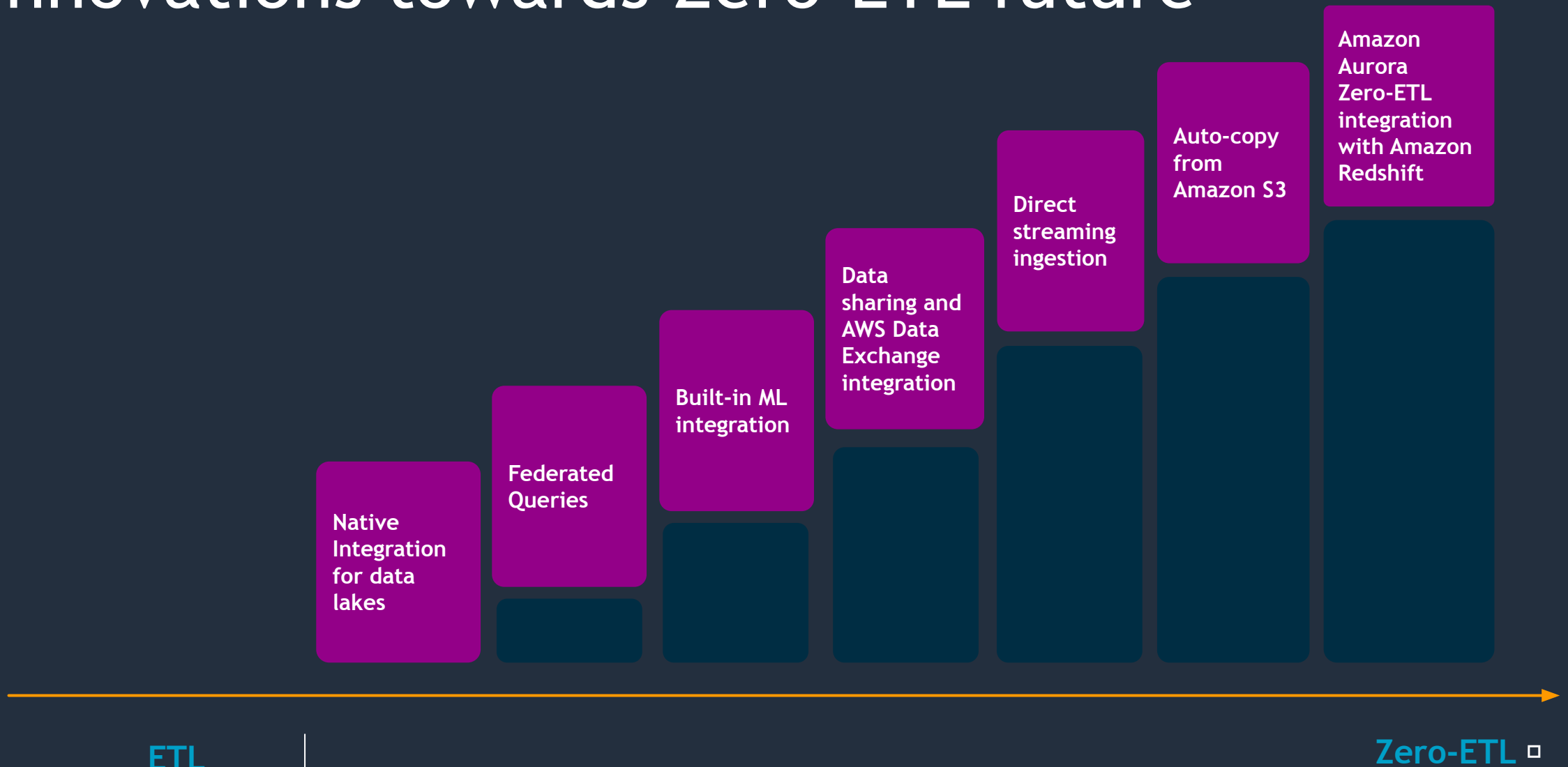
Easily connect to all your data

Services for
connecting to 100s
of data sources

AWS Glue for
value-add data
transformations and more

Zero-ETL
future enabled
by service integrations

Innovations towards Zero-ETL future



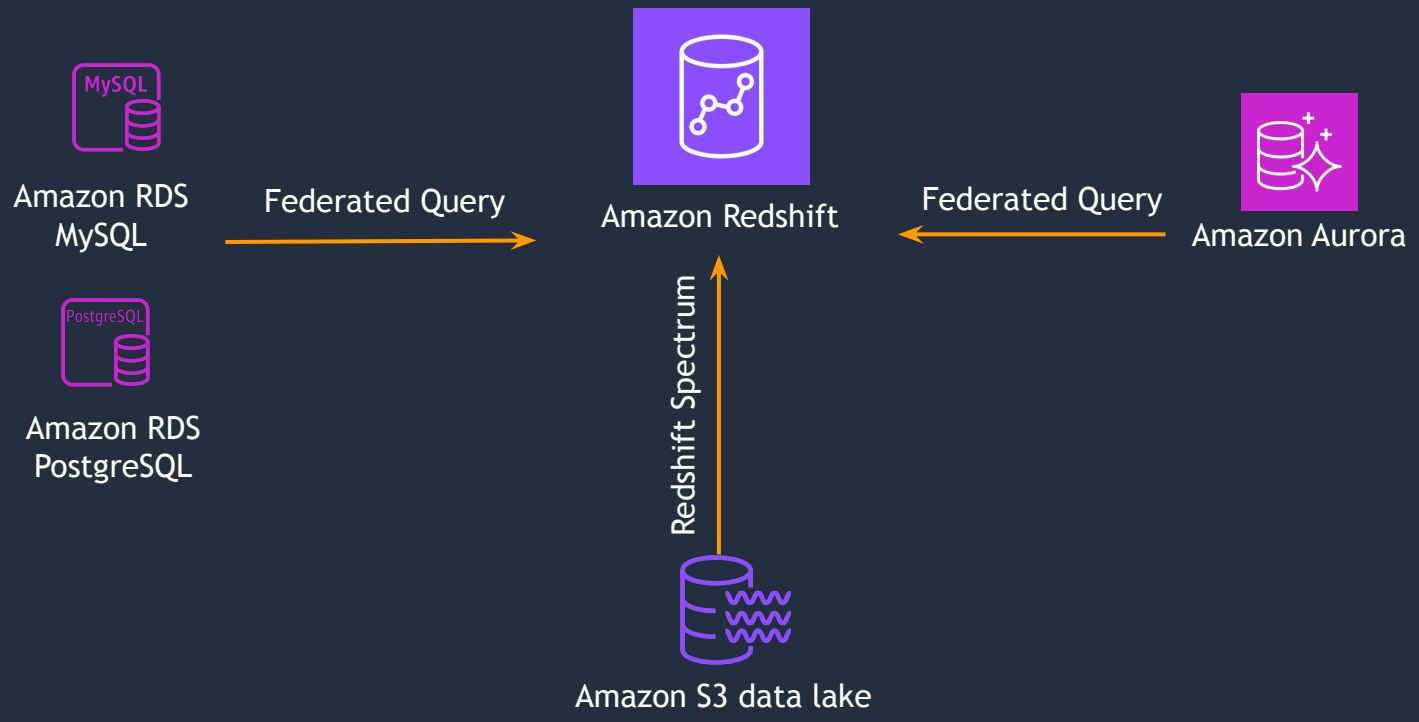
Unified analysis: Amazon Athena federated queries

Run query across relational, non-relational, object, or custom data sources

Integrate external sources with your data lake without ETL

Run query across on-premises or cloud data sources

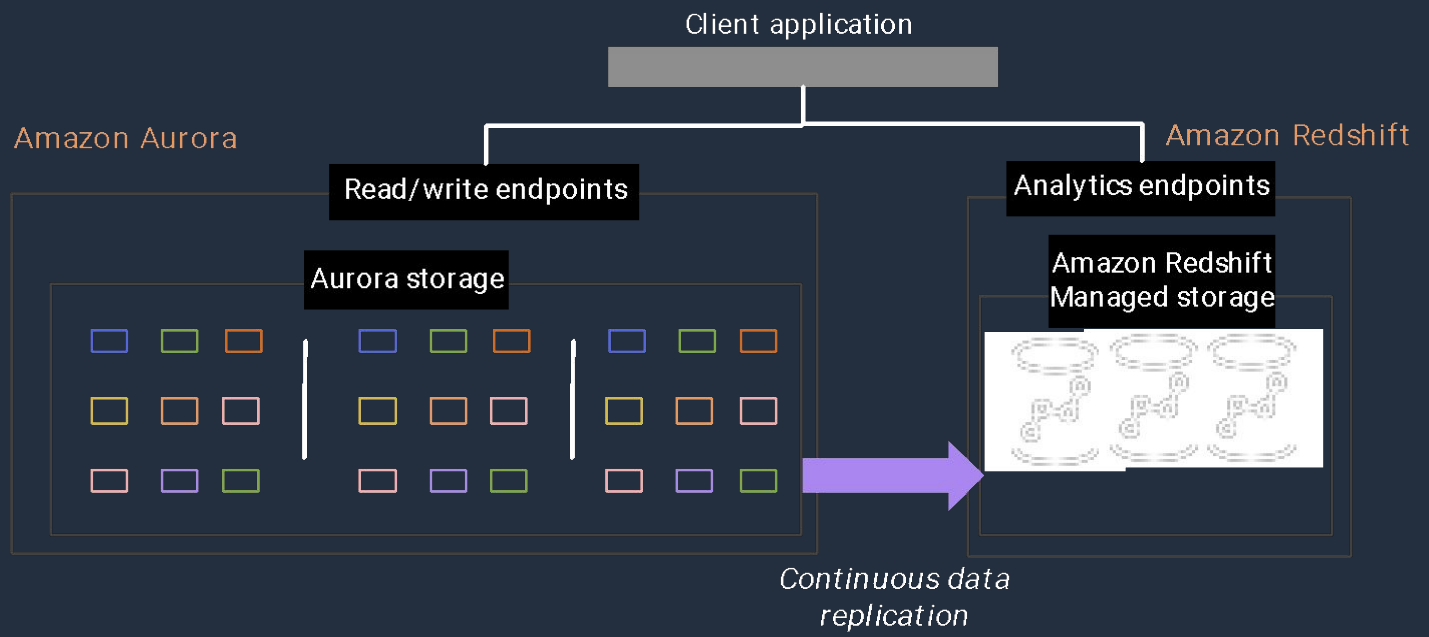
Redshift federated queries and Redshift Spectrum



- Integrate operational databases (Amazon RDS, Amazon Aurora) with data warehouse and data lake for real-time analytics
- Analytics on operational data without data movement and ETL delays
- **Flexible and easy** way to ingest data, avoiding complex ETL pipelines
- Support for open table formats like Apache Hudi and Delta Lake

Amazon Aurora zero-ETL integration with Amazon Redshift

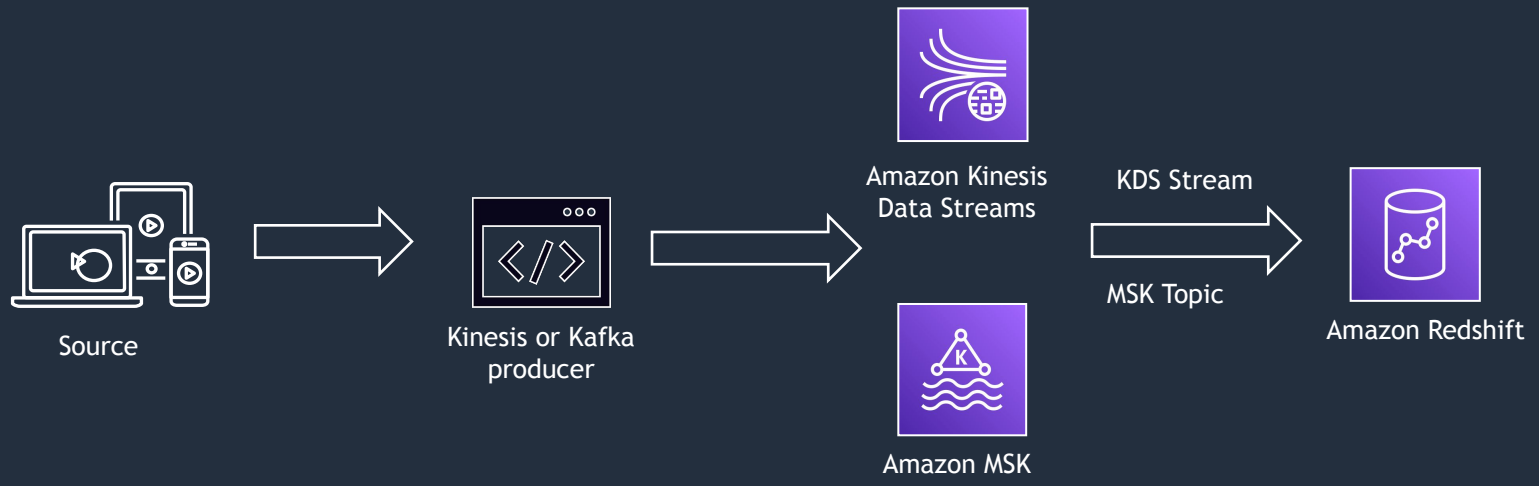
BOOST ANALYTICAL QUERY PERFORMANCE



- No ETL pipelines
- Real-time analytics and machine learning on transactional data
- Consolidated insights from multiple Amazon Aurora databases

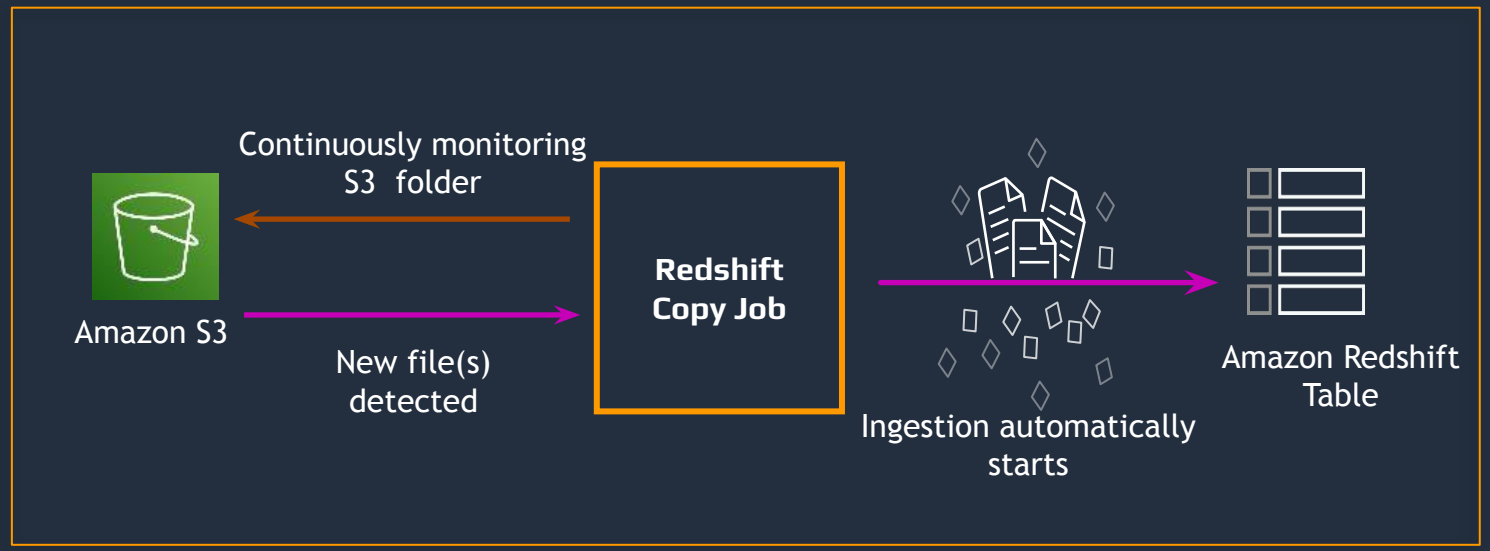
Amazon Redshift Streaming Ingestion support

NATIVE INTEGRATION WITH STREAMING ENGINES FOR REAL-TIME ANALYTICS



- Directly ingest streaming data from Amazon Kinesis Data Streams (KDS) and Amazon Managed Streaming for Apache Kafka (Amazon MSK) without staging in Amazon S3
- Uses Redshift materialized view
- Flexible schema: Define a schema or ingest semi-structured data with the SUPER data type
- Speed: Support 300K rows per second (2 KB size) with <30s latency
- Near real-time analytics – IOT devices, system telemetry, or clickstream data from a busy website

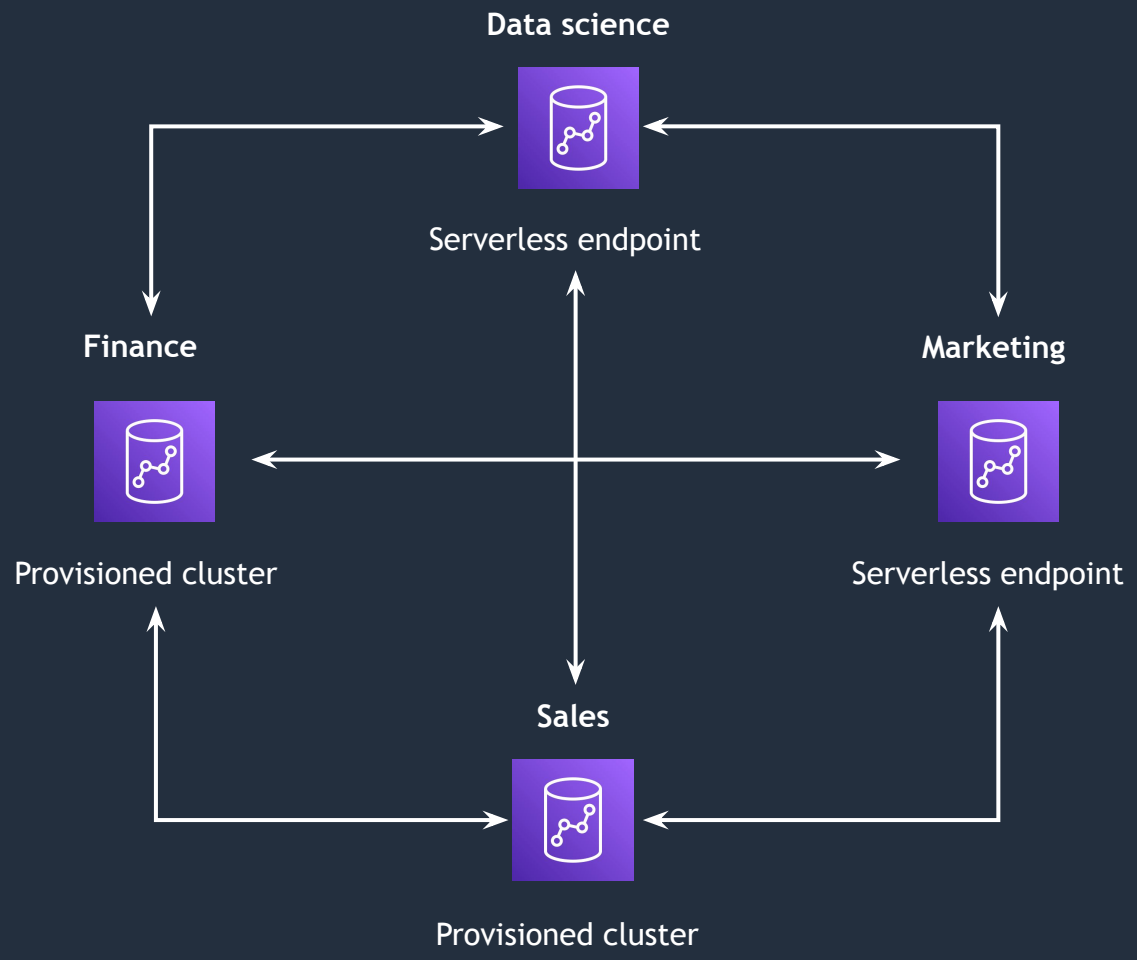
Amazon Redshift auto copy from Amazon S3



- Simplify and automate file ingestion into Amazon Redshift
- Continuously ingest data as soon as new files are created in Amazon S3
- Easily create and maintain simple data ingestion pipelines
- Automate data loading without engineering resources

Amazon Redshift Data sharing

A SECURE AND EASY WAY TO SHARE DATA ACROSS AMAZON REDSHIFT CLUSTERS



- Instant, secure & live data access without copies
- Workload independence and chargeback
- Peer-to-peer or centrally manage access to Amazon Redshift datashares with AWS Lake Formation
- Secure and governed collaboration within and across organizations (Data Mesh pattern)
- Integration with AWS Data Exchange allows you to easily find, subscribe to & use third-party data in the cloud

Built-in ML integration

Databases

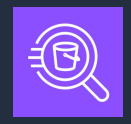
Data warehouses and data lakes

Business intelligence tools

Amazon Aurora ML

Amazon Neptune ML

Amazon Redshift ML



Amazon Athena ML

Amazon QuickSight ML

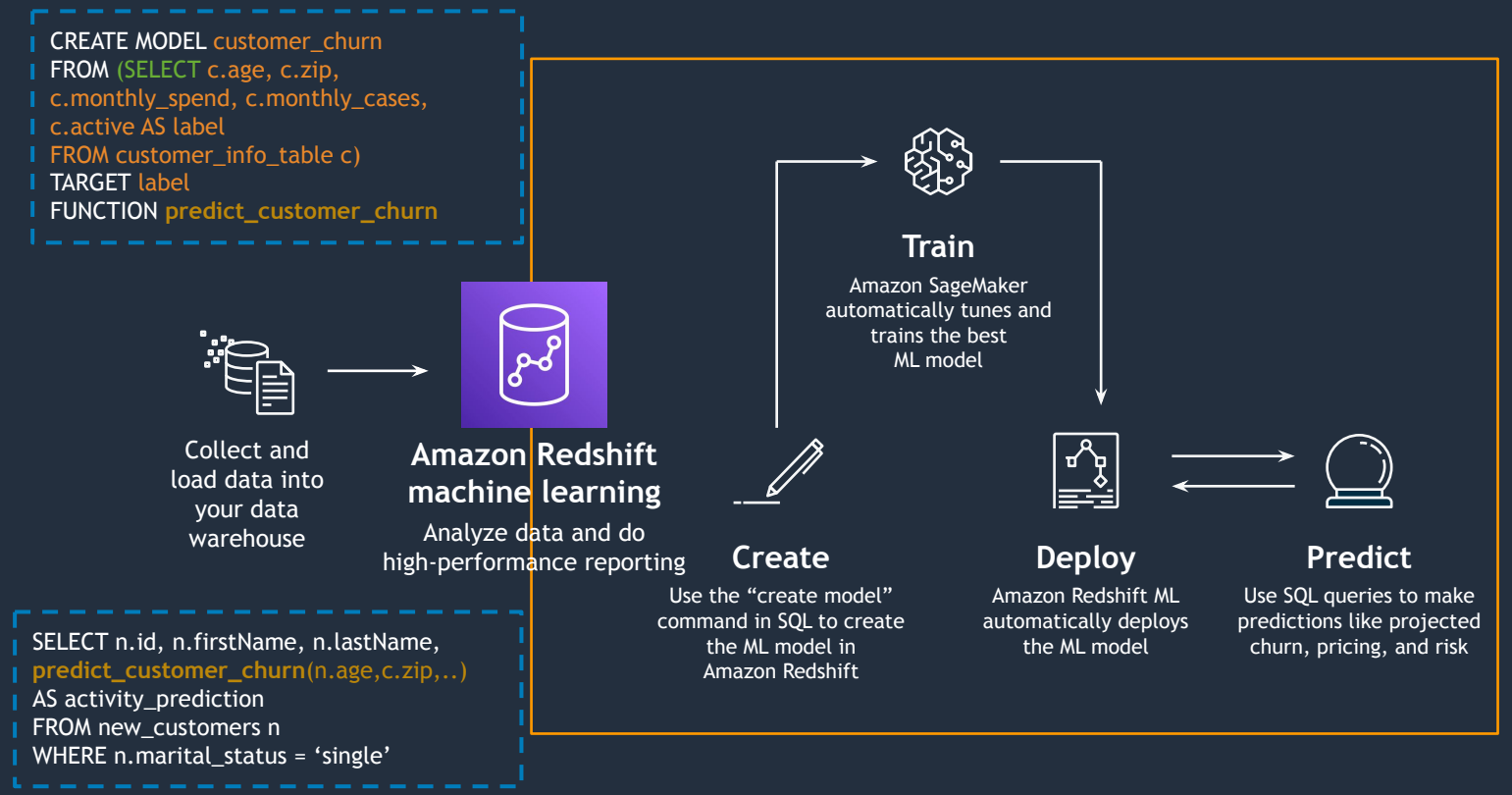
Easily create and train ML models using SQL queries with Redshift ML

Out-of-the-box integration with amazon forecast and SageMaker for predictions/forecasting with SQL

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

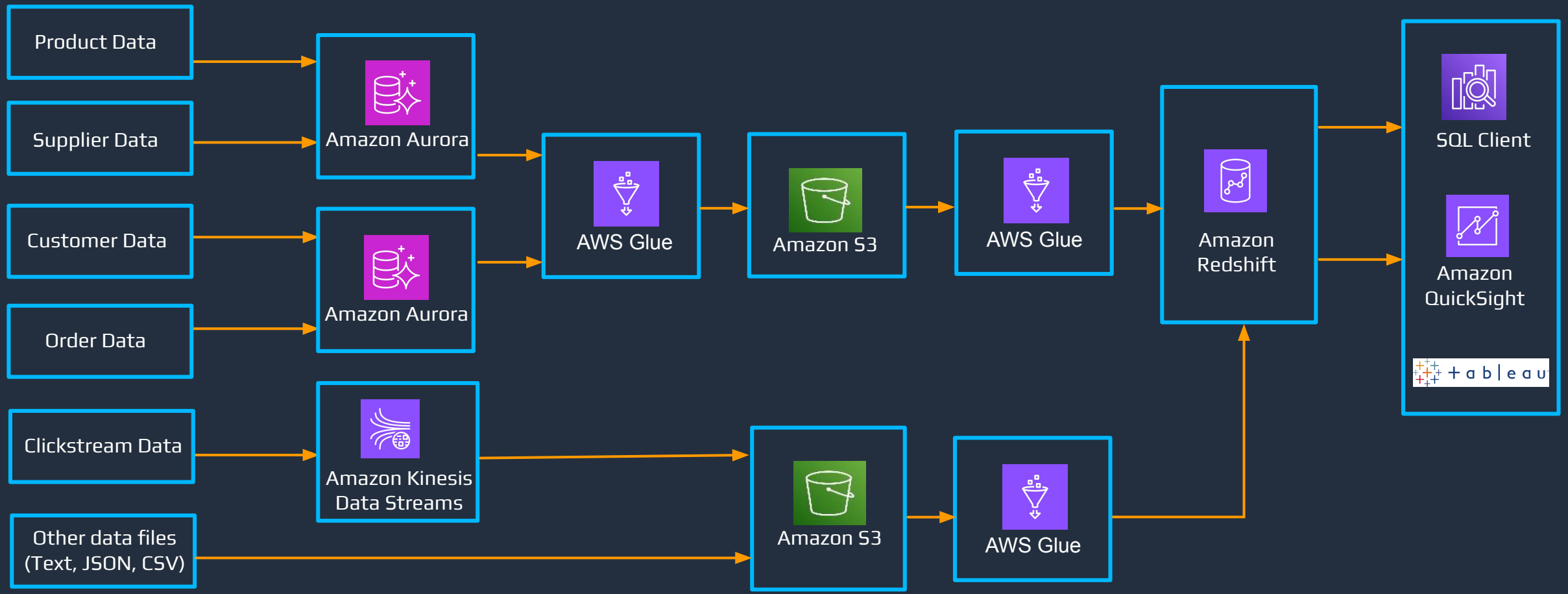
SageMaker models for in-database or remote inference

Supervised and unsupervised trainings



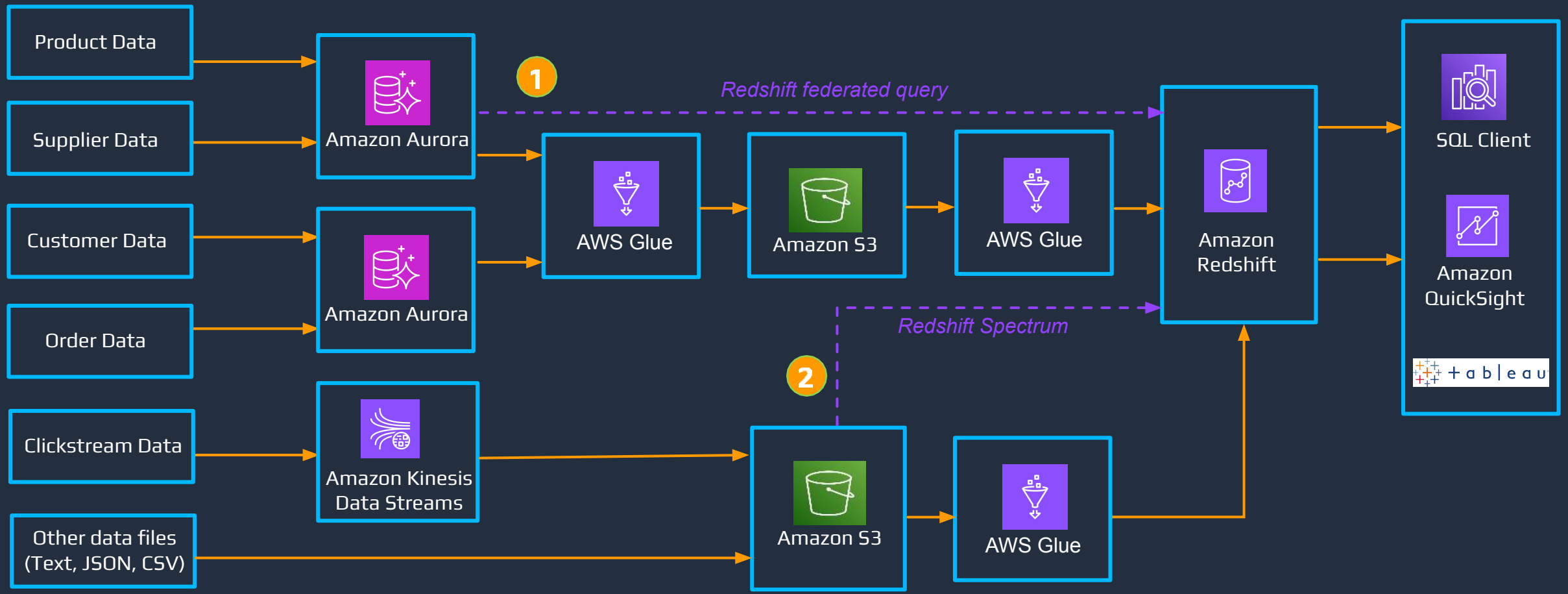
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



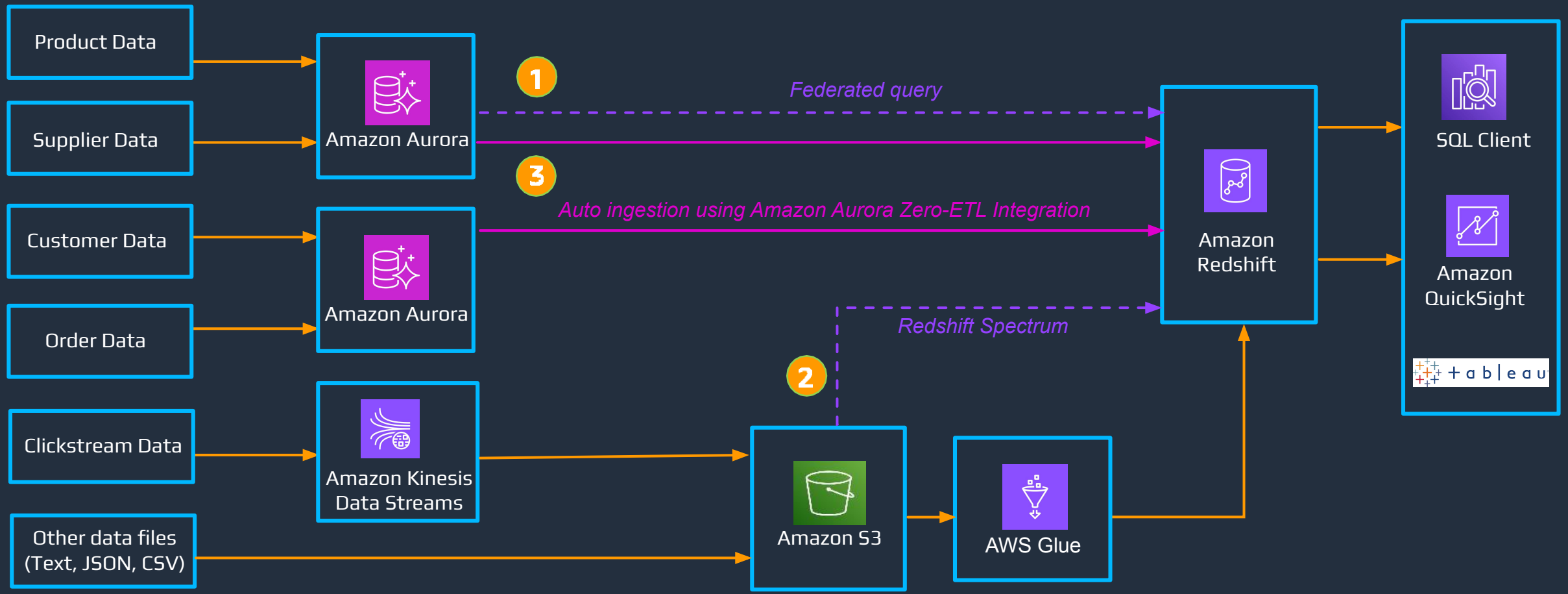
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



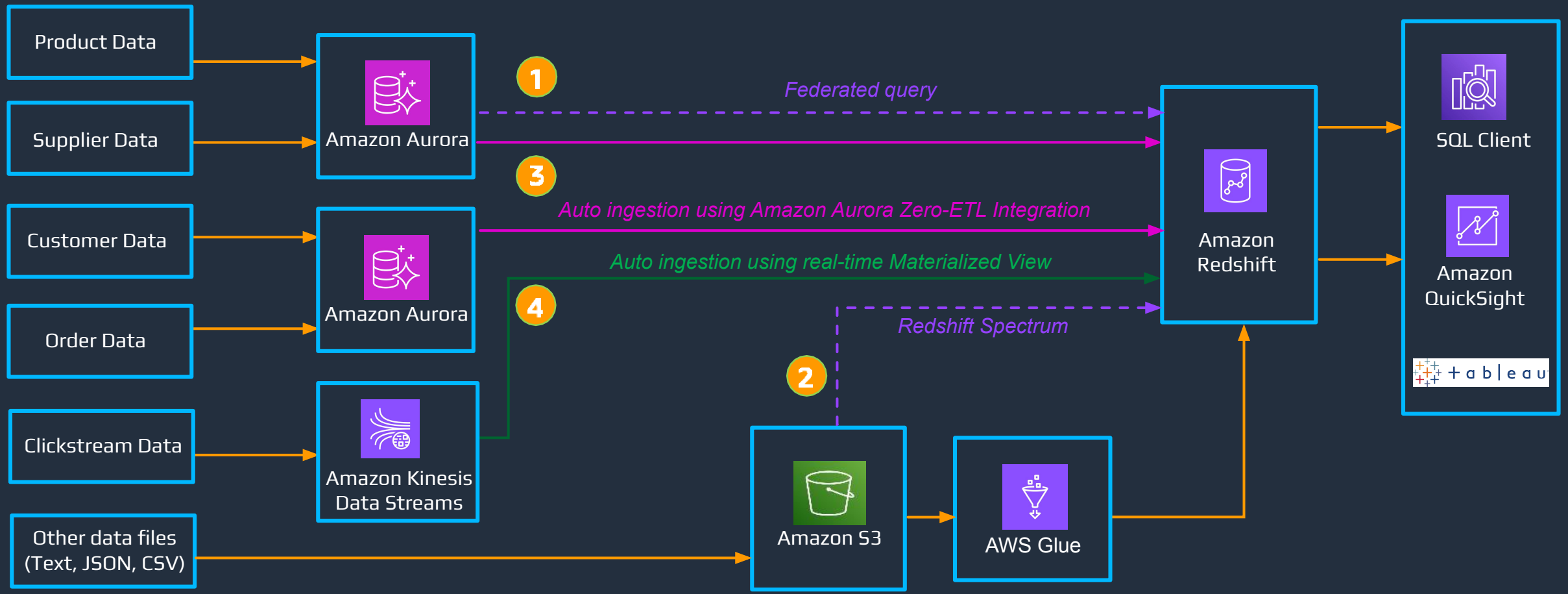
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



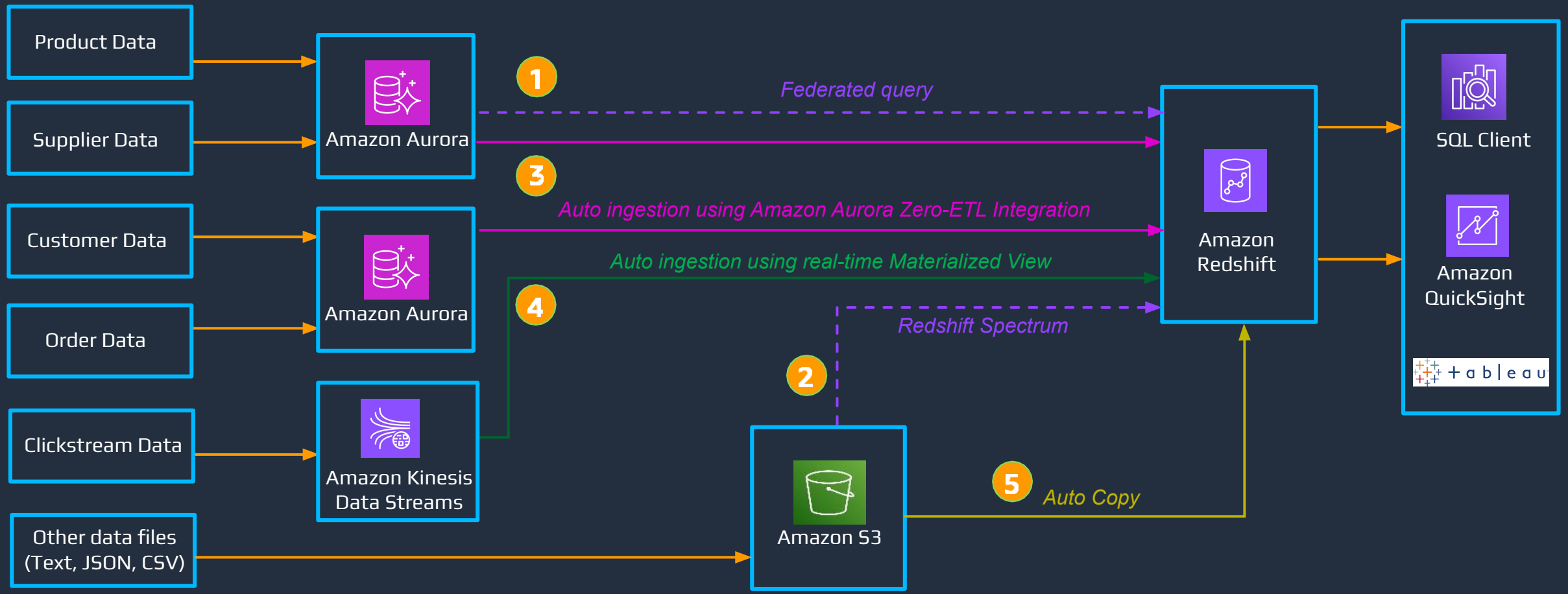
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



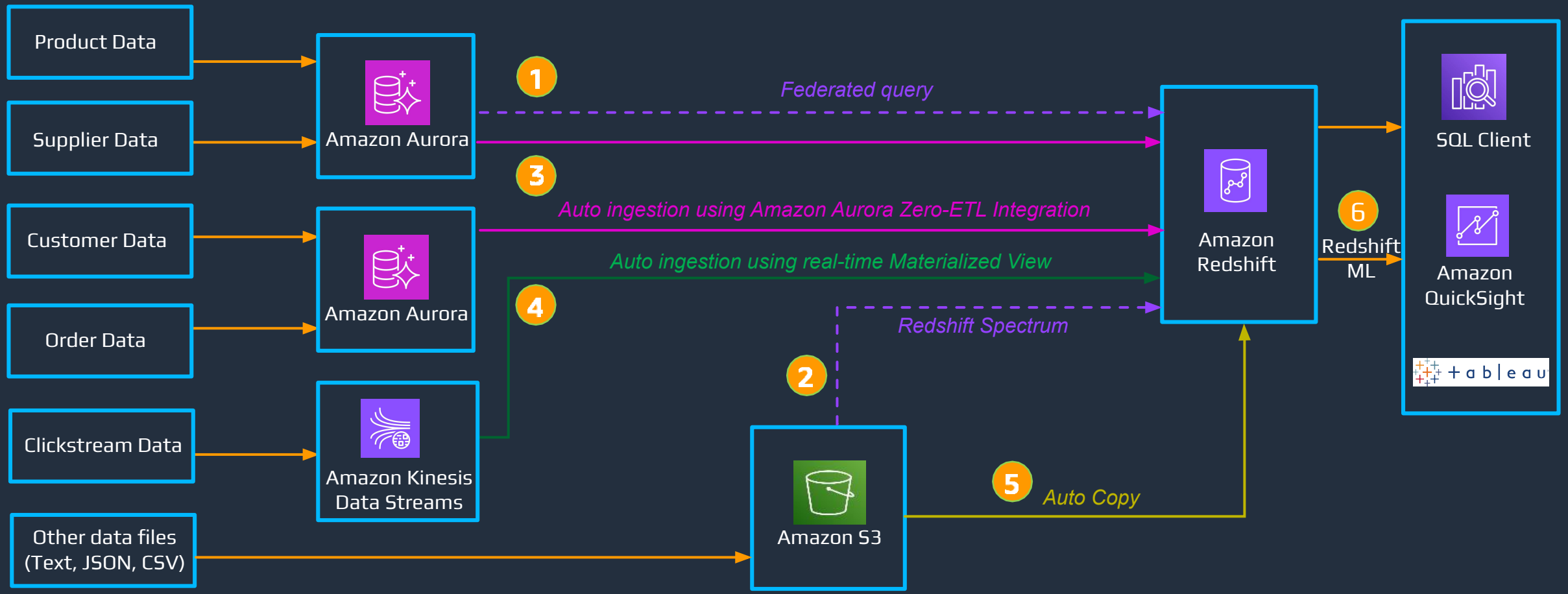
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



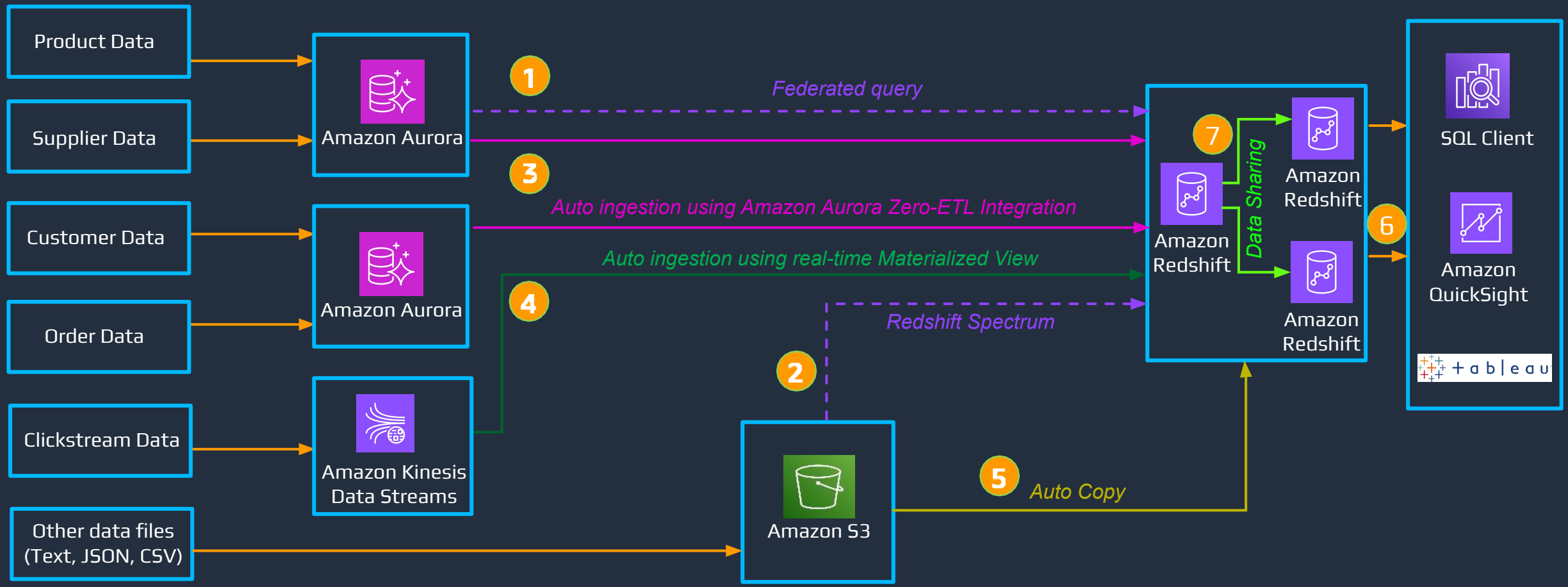
Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



Reference architecture for a Retail customer

UNIFIED REAL-TIME PROCESSING WITH REDSHIFT FEDERATED QUERY AND REDSHIFT SPECTRUM



Demo

Questions?

Thank you!

Sanchit Jain @sanchitdilipjain *(LinkedIn)*