
AWS
re:Invent

ANT328

The lake house approach to data warehousing with Amazon Redshift

Vinay Shukla

Principal Product Manager, Amazon Redshift

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Trends driving lake house architecture

Intro to lake house architecture and Amazon Redshift Spectrum

Use cases enhanced with lake house architecture

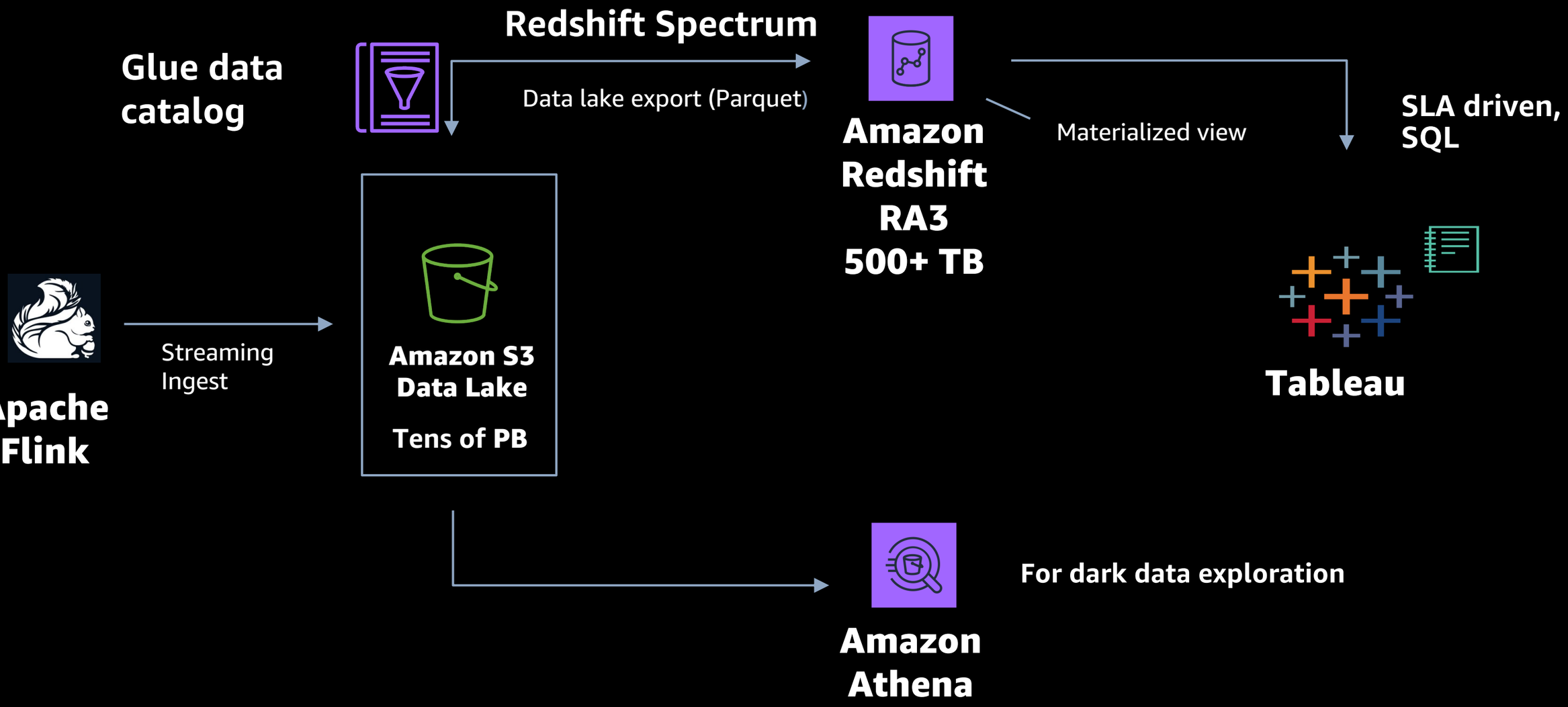
New features of Amazon Redshift Spectrum



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

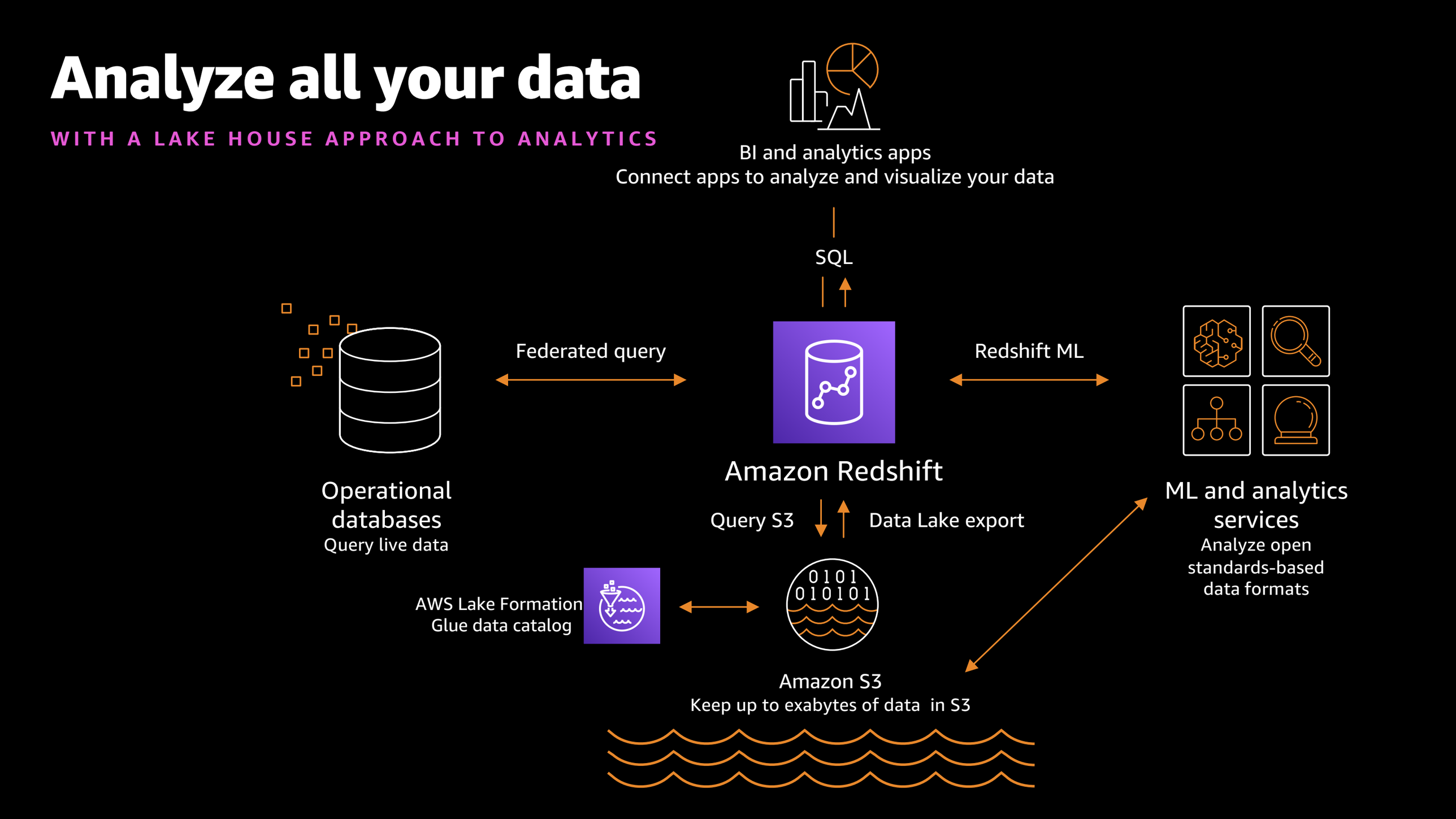
Yelp chose lake house architecture

HOW DOES THIS ARCHITECTURE HELP?

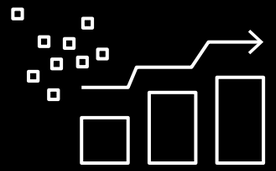


Analyze all your data

WITH A LAKE HOUSE APPROACH TO ANALYTICS



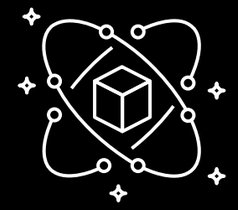
Trends driving the lake house architecture



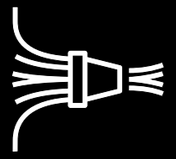
New data sources, sensors drive data growth



Empower personas – use best-fit analytical services



Growth in open data formats



Streaming data – analyze without loading it first



Single source of truth – unify analytics



Unified security, compliance

Benefits of lake house approach



No data loading, any type of data

Terabyte–exabyte scale, with low cost

Security, compliance, and audit capabilities across data lake

Empower all personas – use best-fit analytic services

Democratize ML with SQL; no ETL needed

Unify analytics across operational, data warehouse, and data lake

Amazon Redshift Spectrum

EXTEND THE DATA WAREHOUSE TO EXABYTES OF DATA IN AMAZON S3 DATA LAKES

No data loading required, bring firehose of data

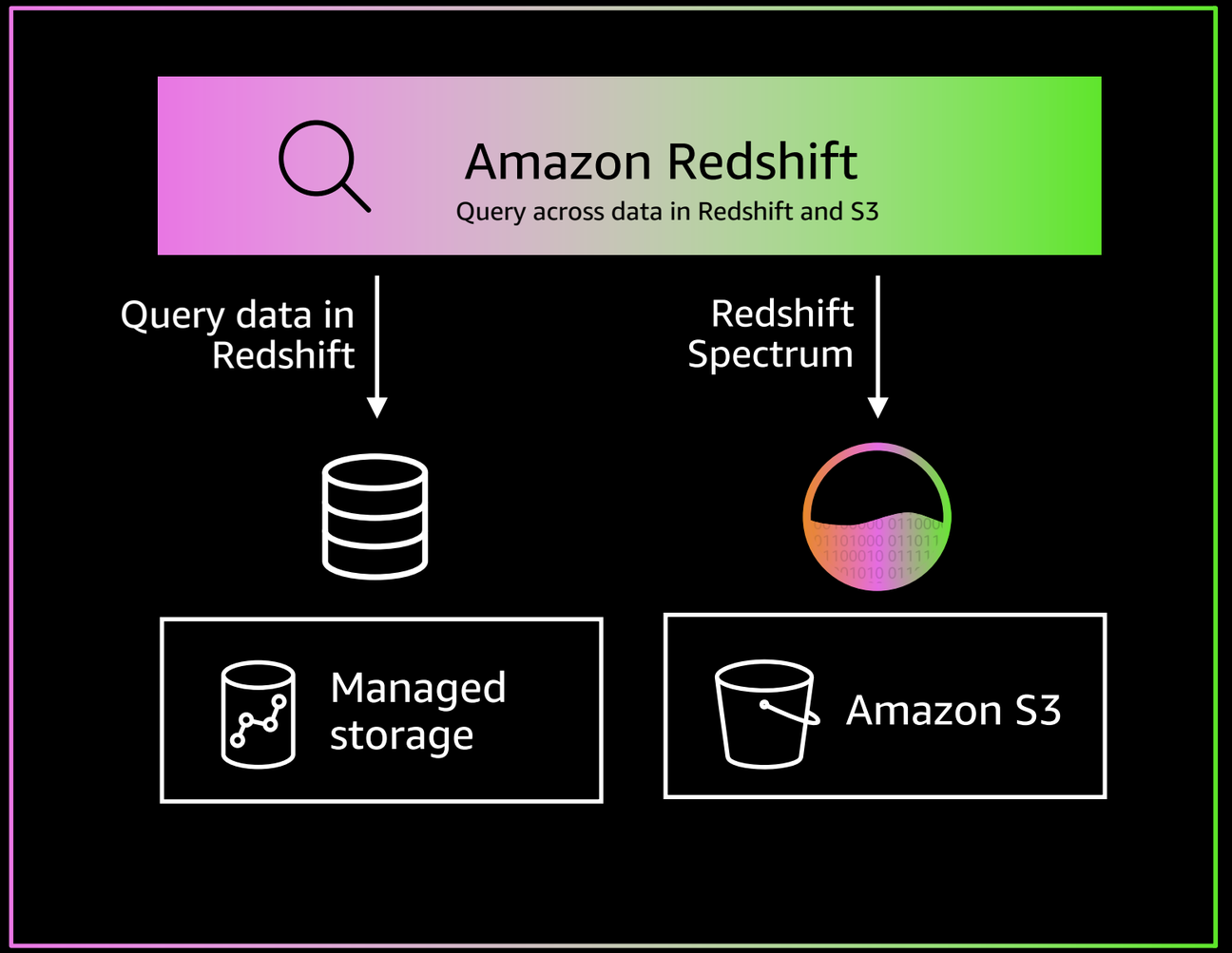
Empowered to use analytics engines of choice

Directly query data stored in Amazon S3

Apache Parquet, ORC, Apache Avro, JSON, and CSV data formats

Any scale of data, pay for what you use

Built-in integration with AWS Lake Formation



Lake house use cases

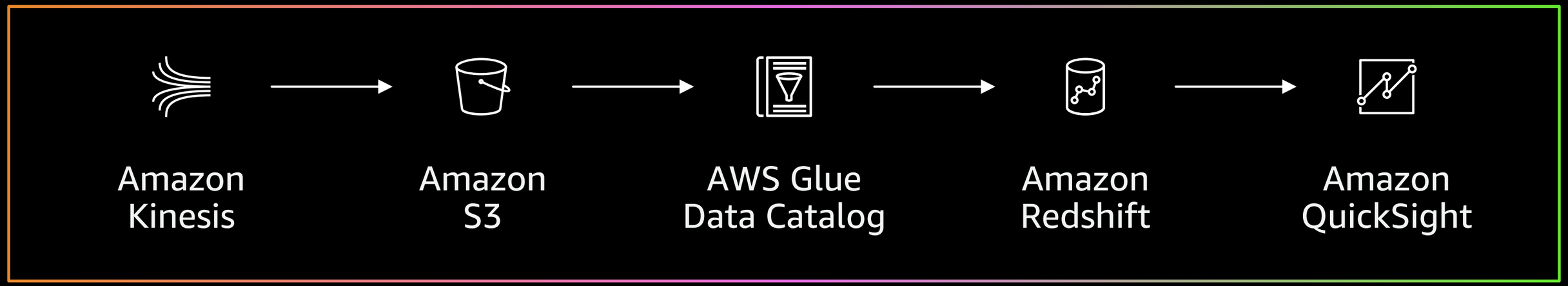


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use case: Handle explosive growth and velocity of data

USE NEWLY BORN DATA FOR EXPLORATORY ANALYTICS

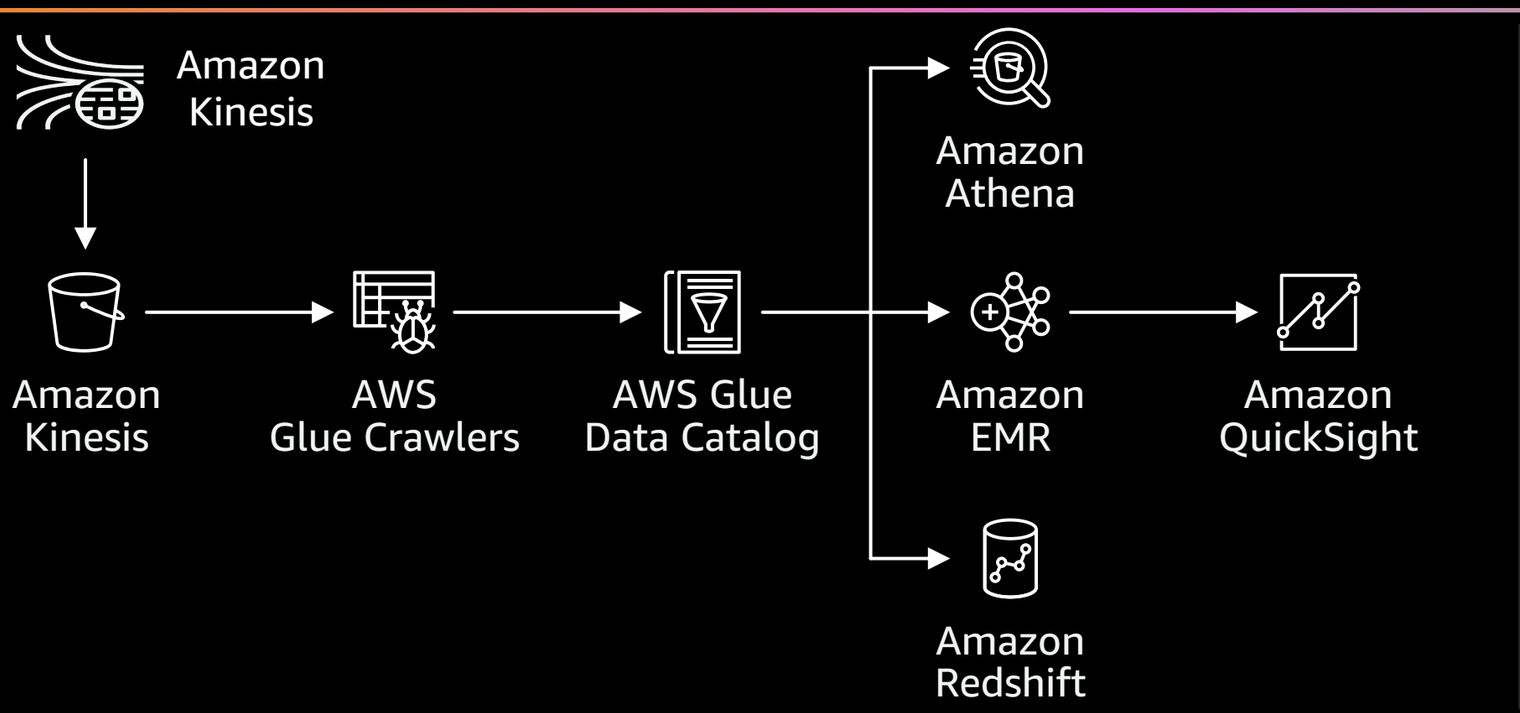
- Handle explosive data growth with Amazon S3 capacity
- Run queries on any data without data load
- Ingest only highly valuable data into Amazon Redshift
- Query unstructured, structured, and open data formats



Use case: Allow personas to use choice of analytics engines

RUN MULTIPLE TYPES OF ANALYTICS ON SINGLE SOURCE OF TRUTH

- Share the data in Amazon S3 with other analytic services
- Analyze data in-place, stored in various formats in Amazon S3
- Offload “cold” data to Amazon S3



AWS Glue
ETL

Amazon Athena
Serverless SQL

Amazon EMR
Data processing

Amazon Redshift
Petabyte scale data warehouse

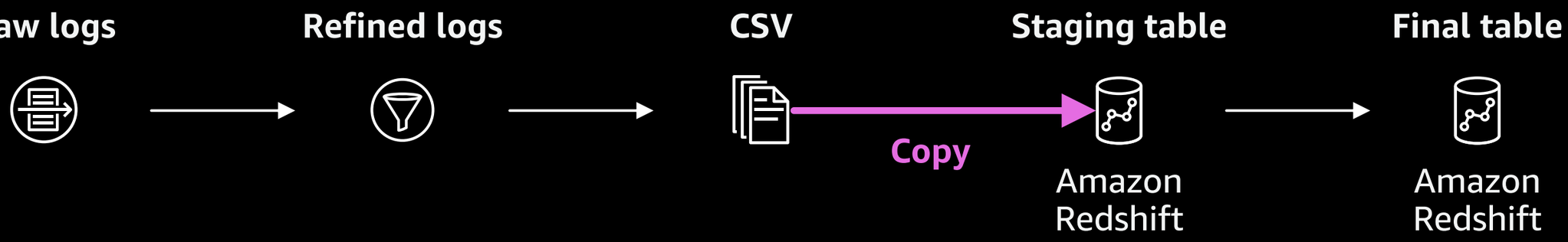
Amazon Redshift Spectrum
Extends Amazon Redshift to query data in Amazon S3

Amazon QuickSight
Visualization

Use case: Simple and fast data ingestion

MOVE FROM ETL TO ELT; USE FAMILIAR SQL FOR ETL

ETL – clean and transform logs before and after **copy** into Amazon Redshift



ELT – clean and transform after **copy** into Amazon Redshift

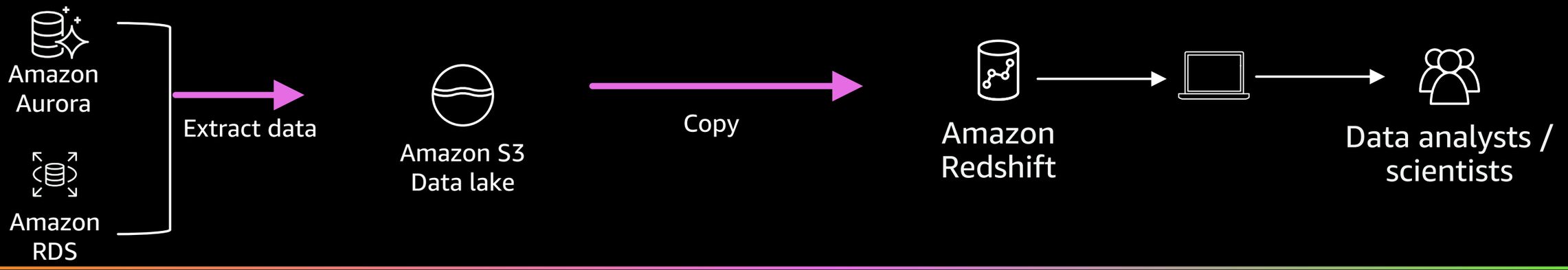



```
INSERT INTO TABLE AS or
(CREATE TABLE AS )
SELECT col1,sum(col2)..
FROM S3.xxx
WHERE …
GROUP BY …
```

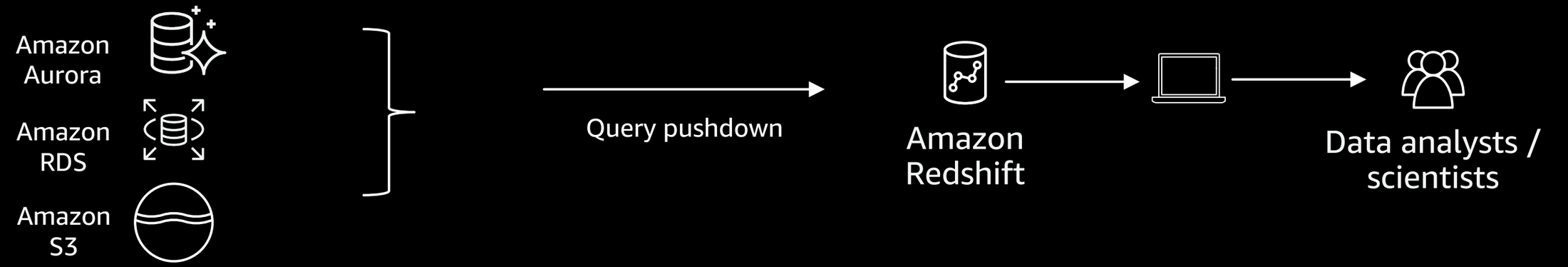
Use case: Query across data sources

QUERY ACROSS OPERATIONAL, WAREHOUSE, AND DATA LAKE; NO ETL REQUIRED

Before: Run ETL jobs to extract data from various sources and copy into Amazon Redshift

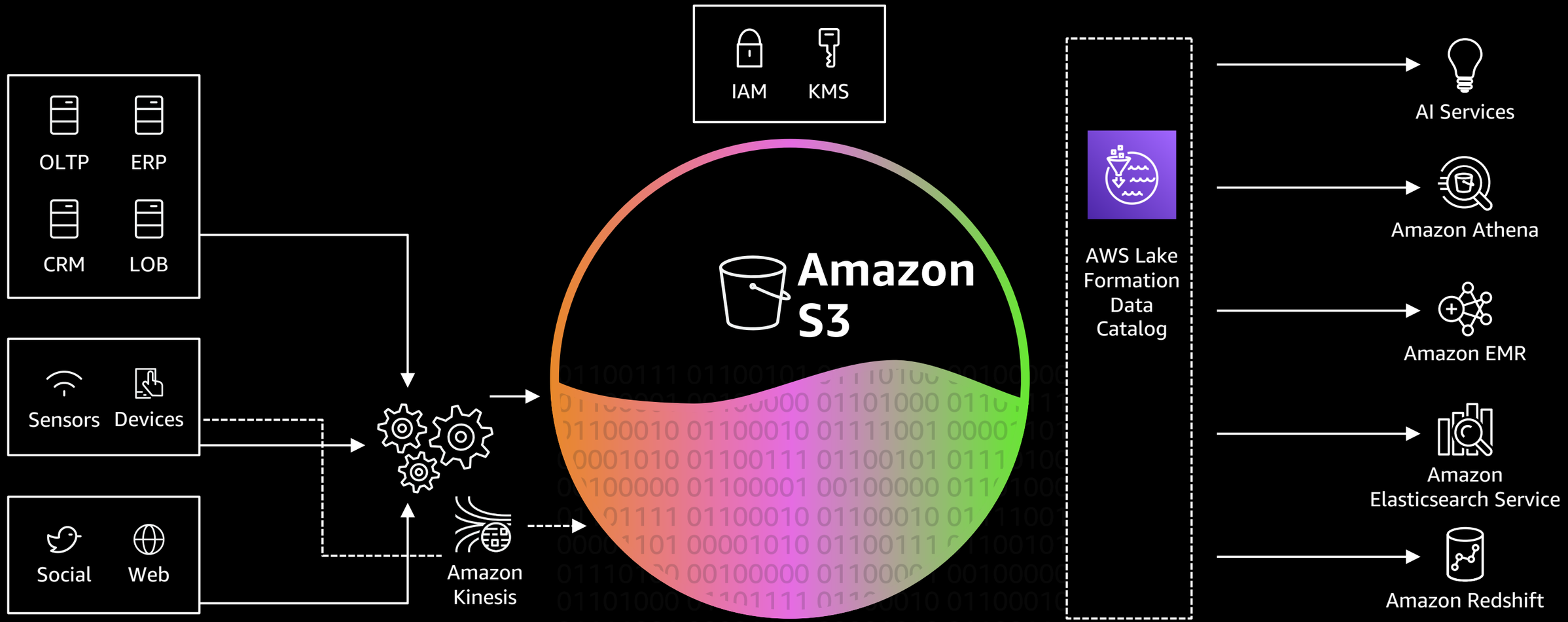


Now: Query data from operational, warehouse, and data lake from Amazon Redshift without ETL



Use case: Apply common security across the data lake

DEFINE SECURITY POLICY ONCE, USED ACROSS DATA LAKE SERVICES



New features



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Redshift Spectrum and AWS Lake Formation integration

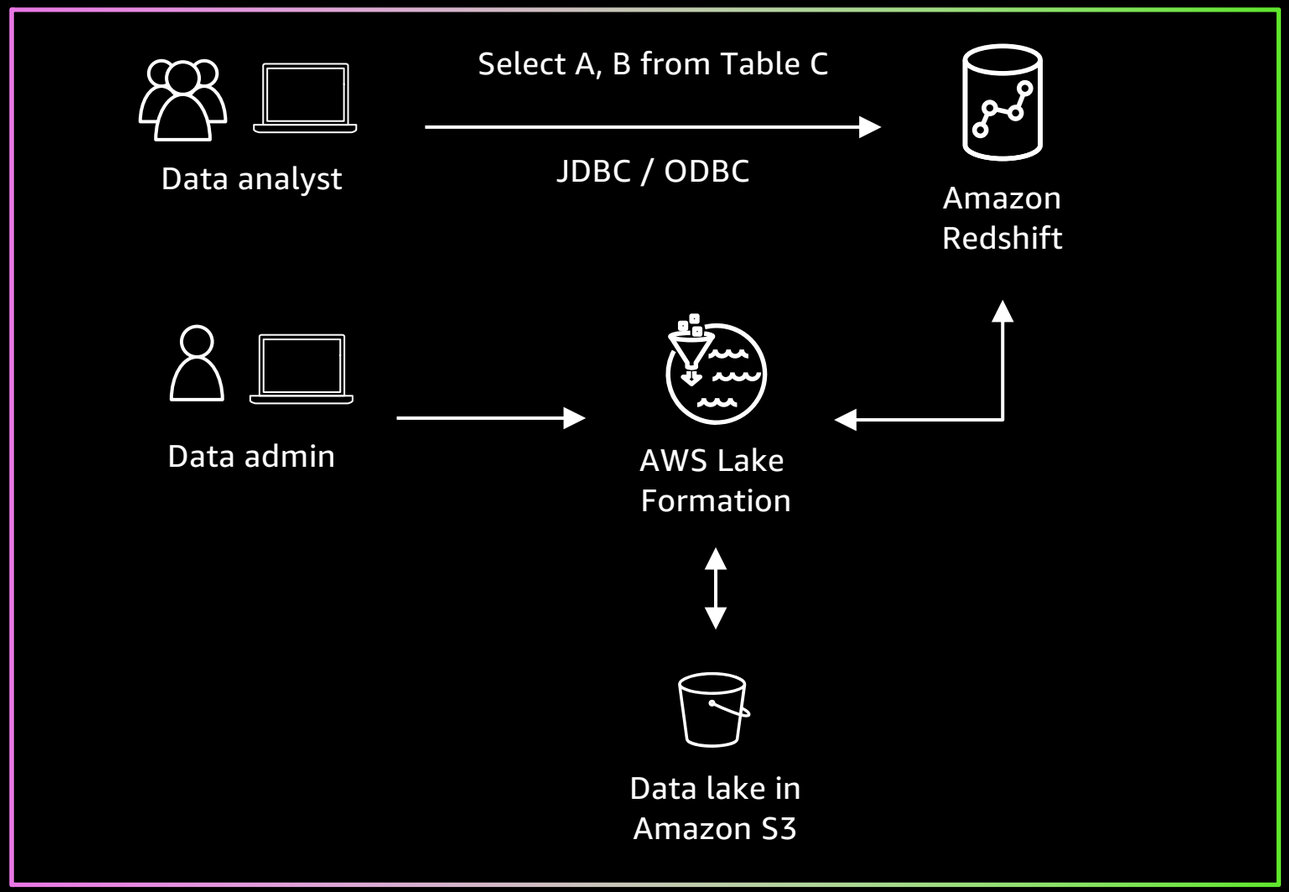
TABLE- AND COLUMN-LEVEL ACCESS CONTROL FOR FINE-GRAINED SECURITY

Benefits

Fine-grained security – only allow access to minimum set of data

Amazon Redshift Spectrum integrates with Lake Formation catalog

Admin creates an IAM role using Lake Formation and grants access to table and column via Lake Formation Console



Data lake export

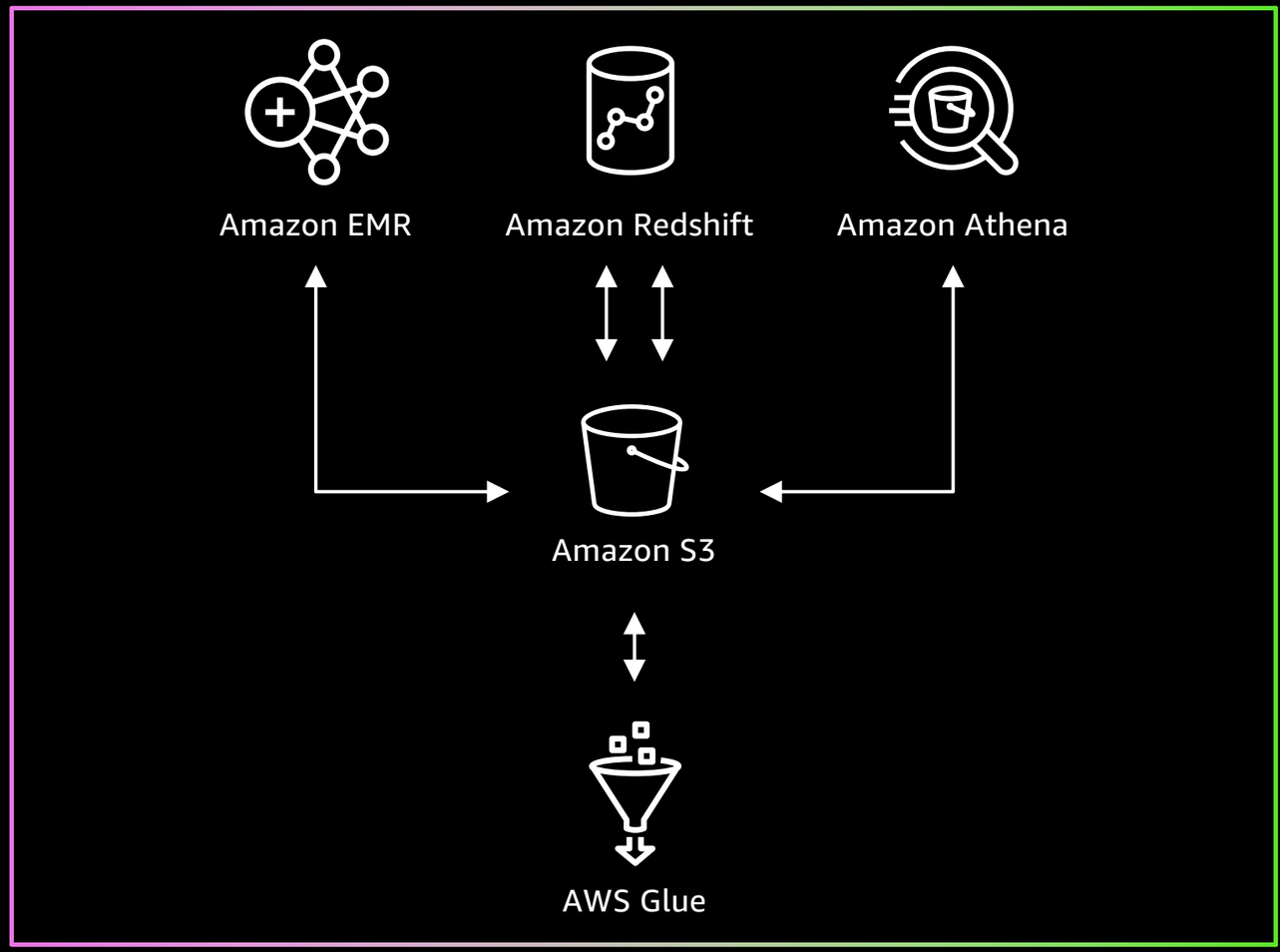
UNLOAD AMAZON REDSHIFT DATA AS PARQUET TO AMAZON S3 DATA LAKES FOR FASTER SHARING AND ANALYTICS

Parquet is an open data format supported by Amazon EMR, Athena, and Amazon Redshift

Amazon Redshift now supports exporting data to Amazon S3 in Parquet format

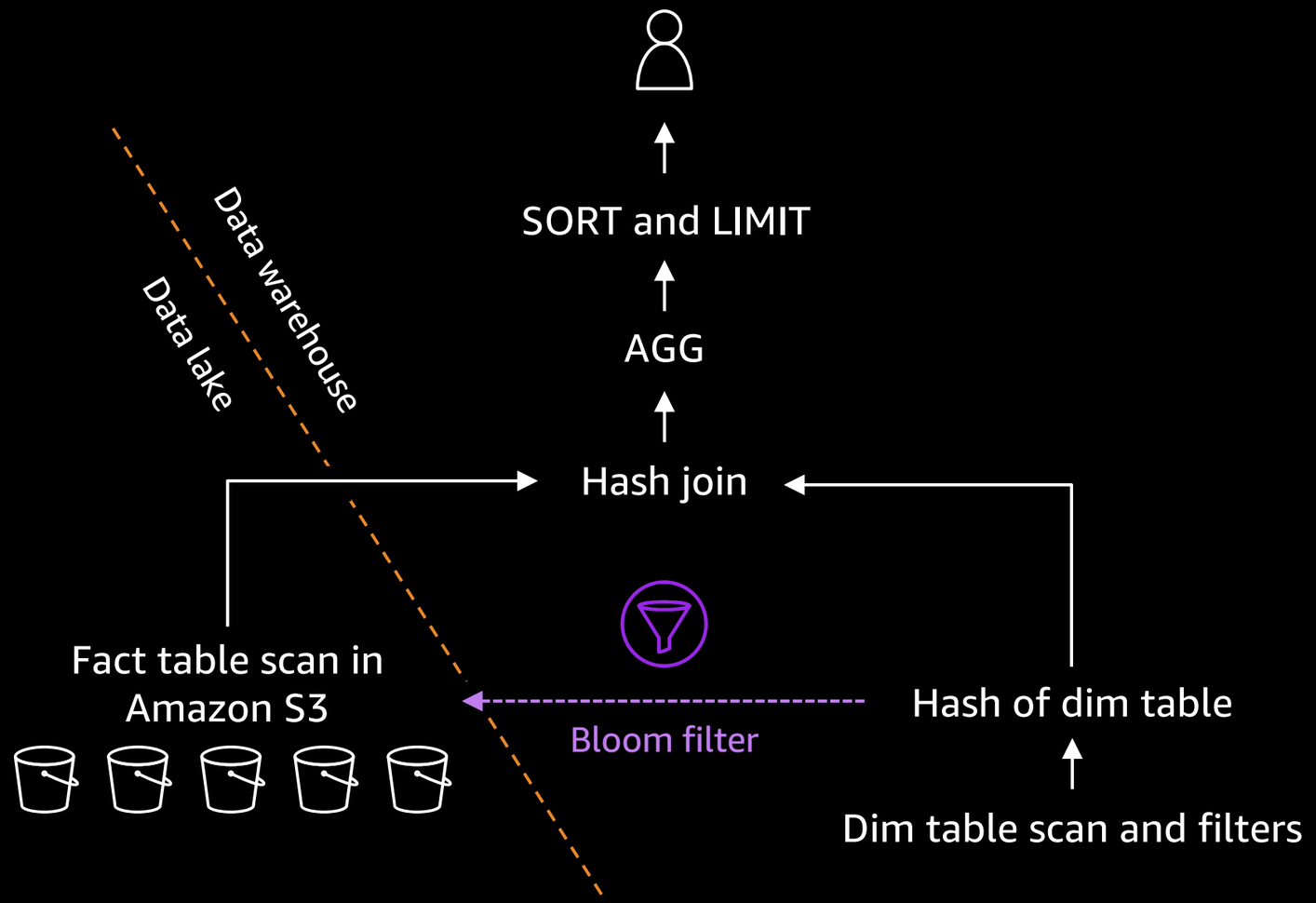
Use SQL with Redshift's UNLOAD command to export data in Parquet format

Unloaded data is automatically registered in AWS Glue Data Catalog



Runtime Bloom filters for fast queries on external tables

EARLY AND EFFECTIVE DATA FILTERING FOR SIGNIFICANTLY FASTER QUERY PERFORMANCE



Cost control for concurrency scaling and Amazon Redshift Spectrum

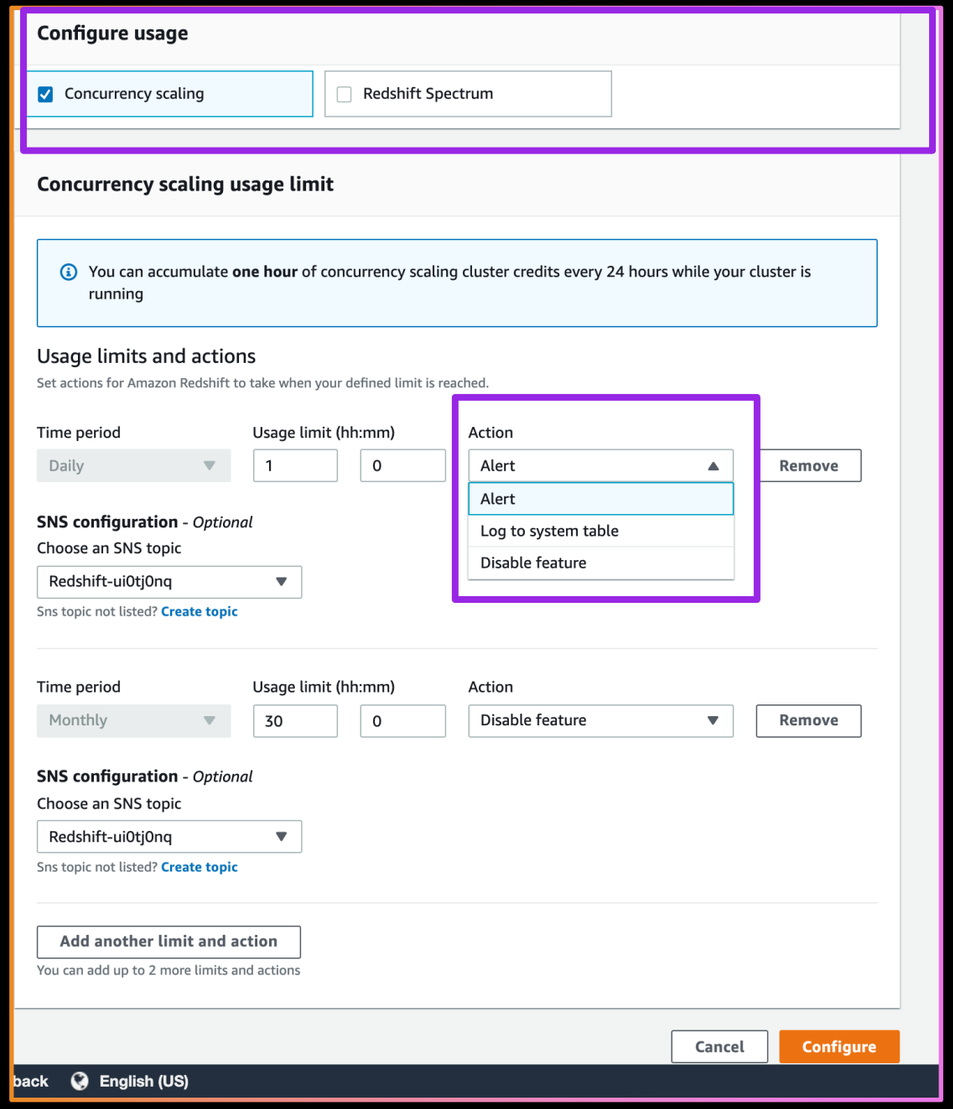
CONTROL YOUR SPENDING BY CREATING USAGE LIMITS

Create daily, weekly, and monthly limits

Up to four limits per feature

Log to system table, alert, disable feature

Automatically generate
Amazon CloudWatch alarm



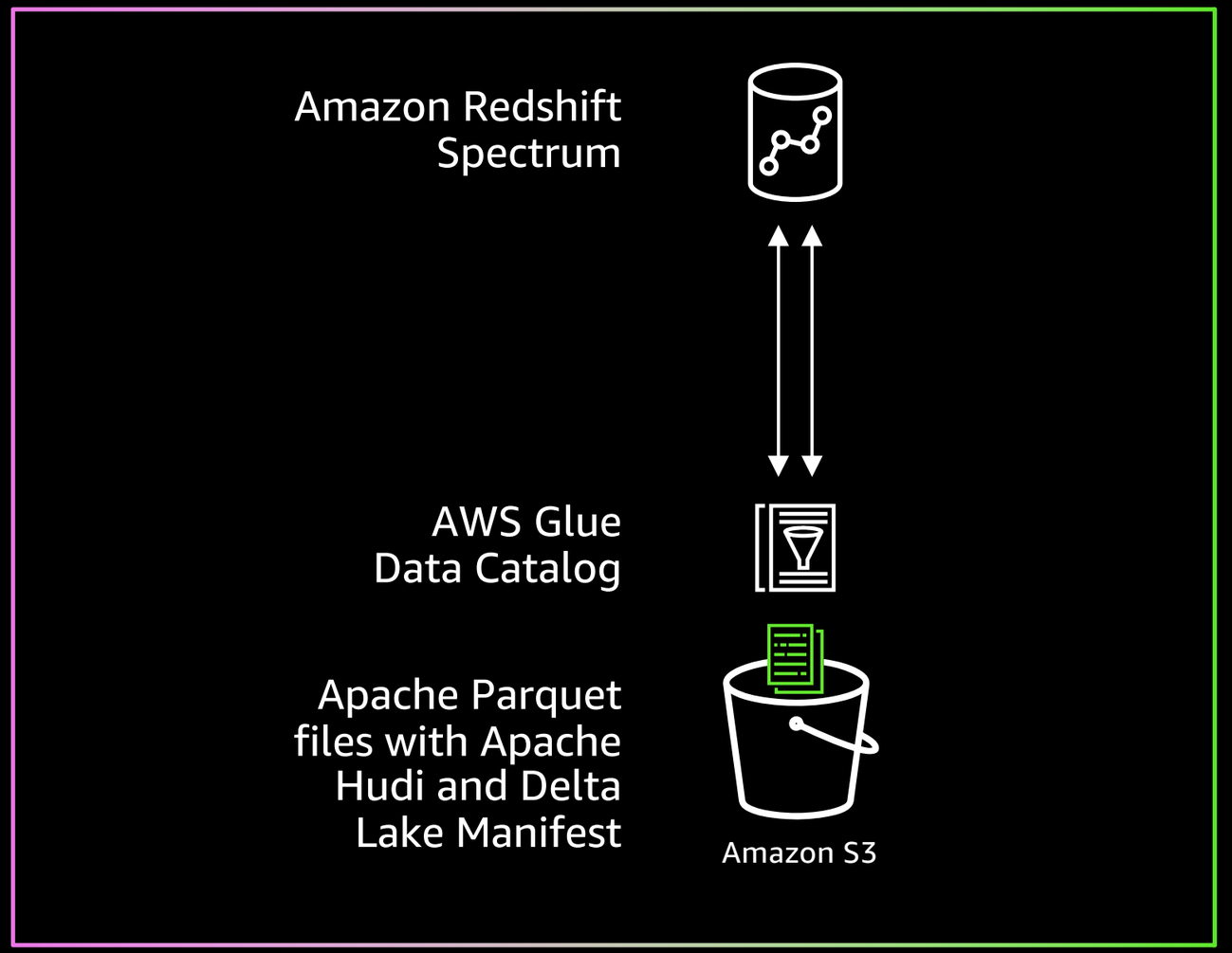
Support for Apache Hudi and Delta Lake

AMAZON REDSHIFT SPECTRUM QUERIES NOW READ TABLES IN APACHE HUDI AND DELTA LAKE; IMPLEMENT USE CASES SUCH AS CHANGE DATA CAPTURE, GDPR

Apache Hudi and Delta Lake are open-source storage framework

Enable use cases (such as change data capture, GDPR) and have data be queried from Amazon Redshift Spectrum

Data stays in Parquet format and stays in Amazon S3



Amazon Redshift federated query

UNIFIED ANALYTICS ACROSS DATABASES, DATA WAREHOUSE, AND DATA LAKE

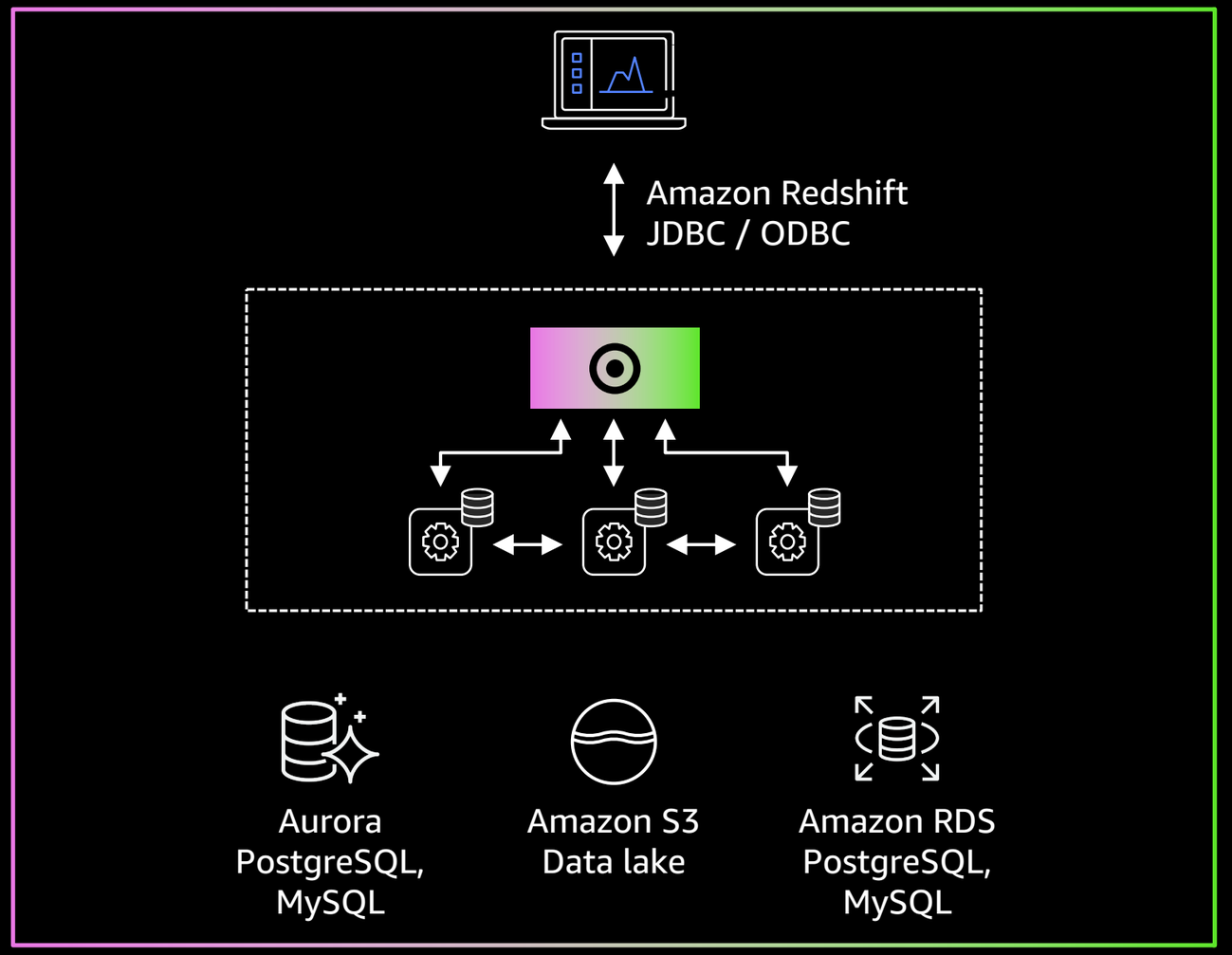
Integrate operational data with data warehouse and Amazon S3 data lake

Analytics on operational data without data movement and ETL delays

Query and join data from one or more Amazon RDS and Aurora databases

Flexible and easy way to ingest data, avoiding complex ETL pipelines

Preview — Amazon RDS and Amazon Aurora MySQL support



Materialized views on external tables

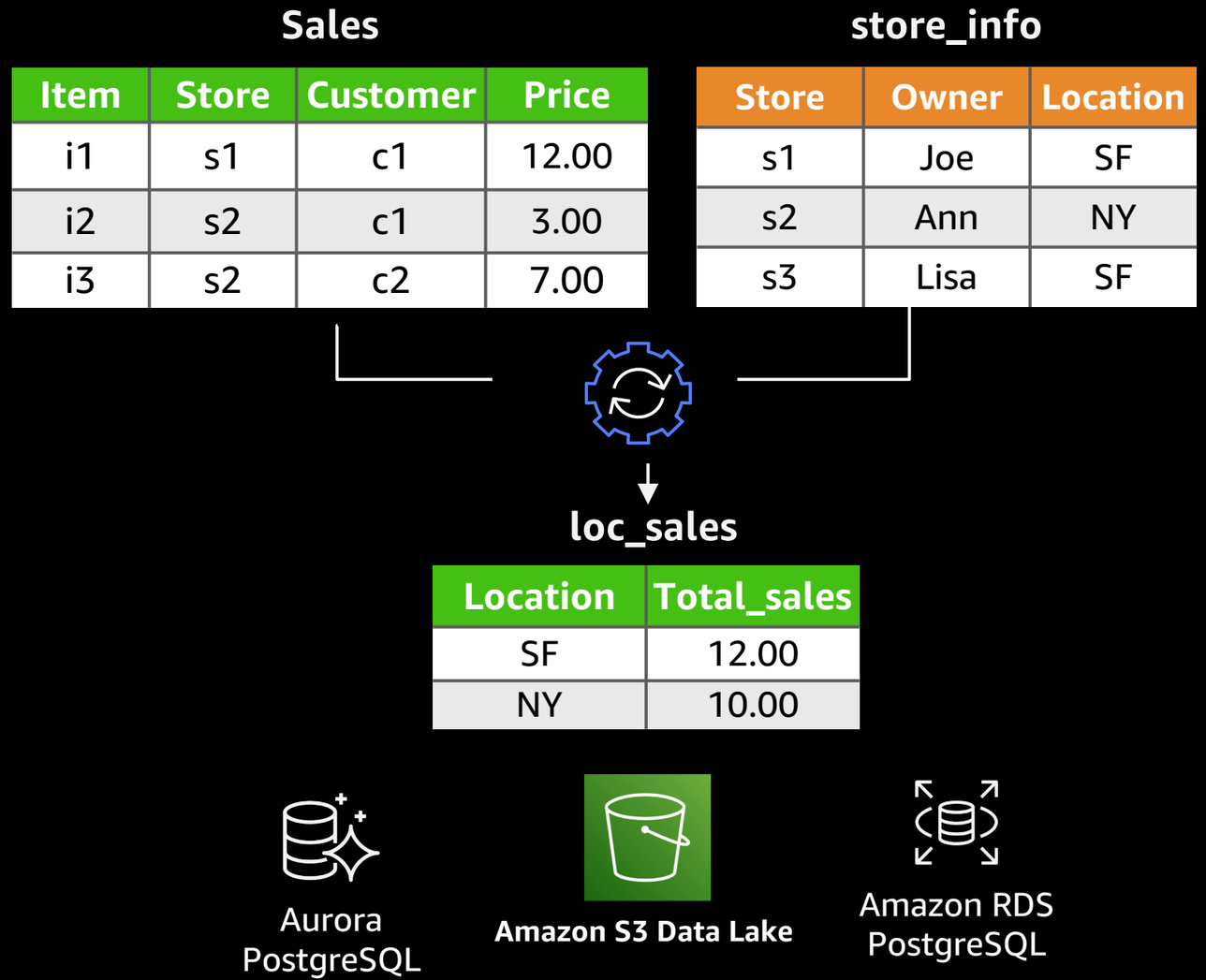
SPEED UP QUERY PERFORMANCE BY ORDERS OF MAGNITUDE WITH PRECOMPUTED RESULTS

Simplify and accelerate ETL, BI, Dashboards Repeated, and predictable analytical workloads

Efficient maintenance manual, scheduled

Leverage MVs for data in data lakes

Easier and faster migration to Amazon Redshift



Amazon Redshift ML

PREVIEW

EASILY TRAIN AND USE ML IN SQL QUERIES WITH AMAZON SAGEMAKER

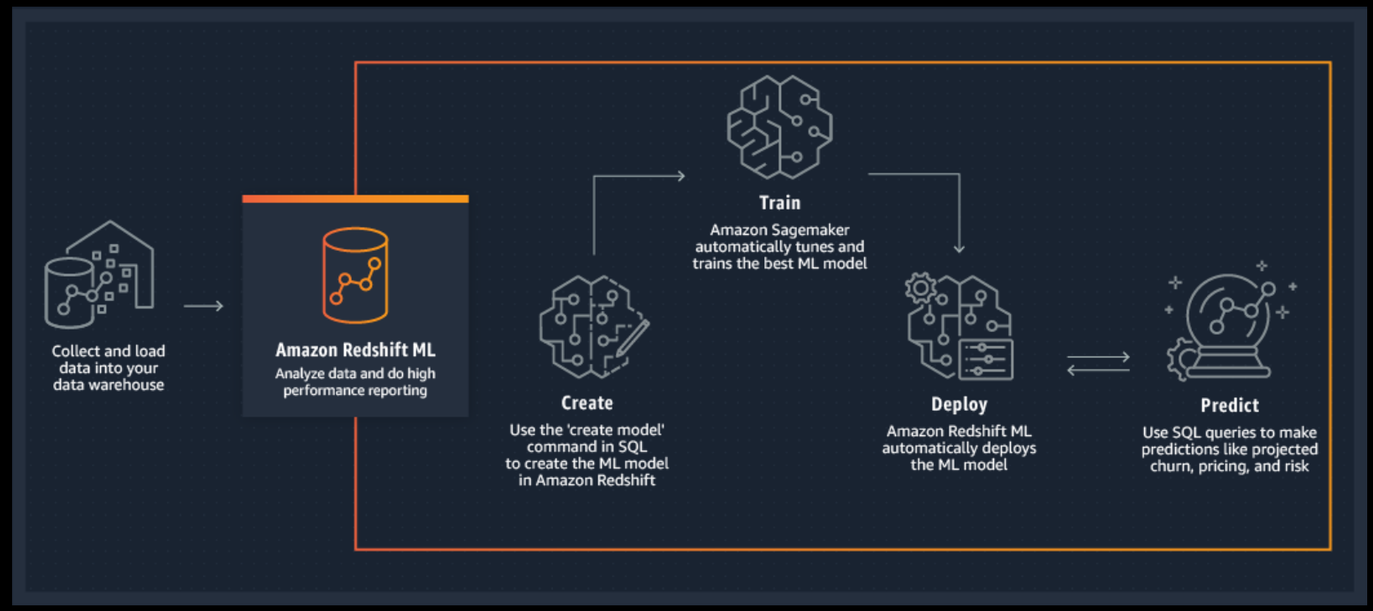
Use case: Product recommendations, fraud prevention, and customer churn reduction

Create, train, and apply ML models using SQL

Automatic selection of ML algorithms or select your algorithm with XGBoost

Automatic preprocessing, creation, training, deployment of your model

Deploy inference models locally in Amazon Redshift; run an inference as invoking a user-defined function as part of SQL statements




```
CREATE MODEL demo_ml.customer_churn
FROM (SELECT c.age, c.zip, c.monthly_spend,
c.monthly_cases, c.active FROM
customer_info_table c)
TARGET c.active;
```

Summary: Benefits of lake house approach



No data loading, any type of data

Terabyte–exabyte scale, with low cost

Security, compliance, and audit capabilities across data lake

Empower all personas – use best-fit analytic services

Democratize ML with SQL; no ETL needed

Unify analytics across operational, data warehouse, and data lake

Next steps

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/

Getting started with Amazon Redshift?

Ask your account team
for a 10-minute demo

Learn more about starting
a proof of concept

https://docs.aws.amazon.com/redshift/proof-of-concept-playbook.html

Or request help from AWS

https://pages.awscloud.com/redshift-proof-of-concept-request

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.