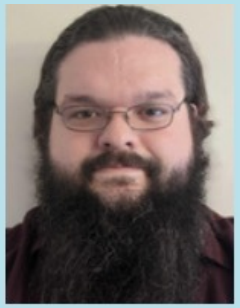
The Fourth AWS MeetUp!

Kevin
Lewis



Nicholaus
Lawson

Your MeetUp organizers

- Welcome
- Our vision for this group
  - Monthly gatherings
  - Building community
  - Gain skills
  - Connecting talent to opportunity
  - Learn from each other
  - Member driven

John
Rice

Tonight's speaker

After this: West Sixth – First round is on AWS!

- Our Next Series of Discussions
- August: ML/AI Sage Maker
- September: Getting started with CDK
- October: Security
- Kubernetes EKS – November
- December: Round Table Discussion
- **January: Data Lake with John Rice**

- **Upcoming**
- February: Networking Dan Powers
- March: VM Ware Cloud on AWS Marshall Radwin

- **Agenda for tonight**
- Around the room intros (and make a name tag)
    - who you are
    - where you work
    - are you working on an exciting project
- John Rice takes the floor
- SWAG giveaway

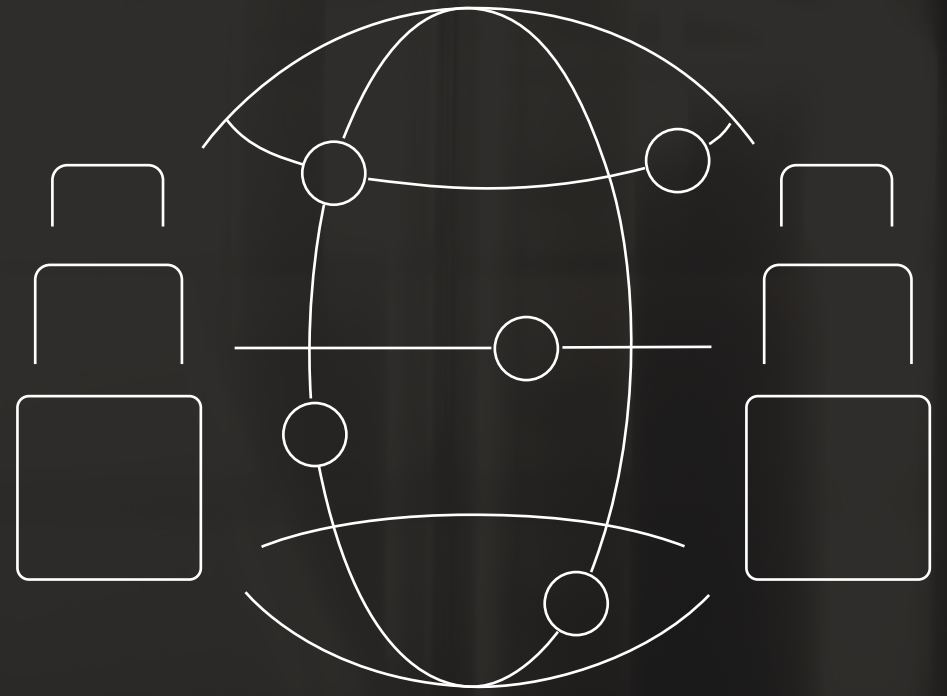
Build secure data lakes with AWS Lake Formation

John Rice, AWS Senior Solutions Architect

jkrice@amazon.com

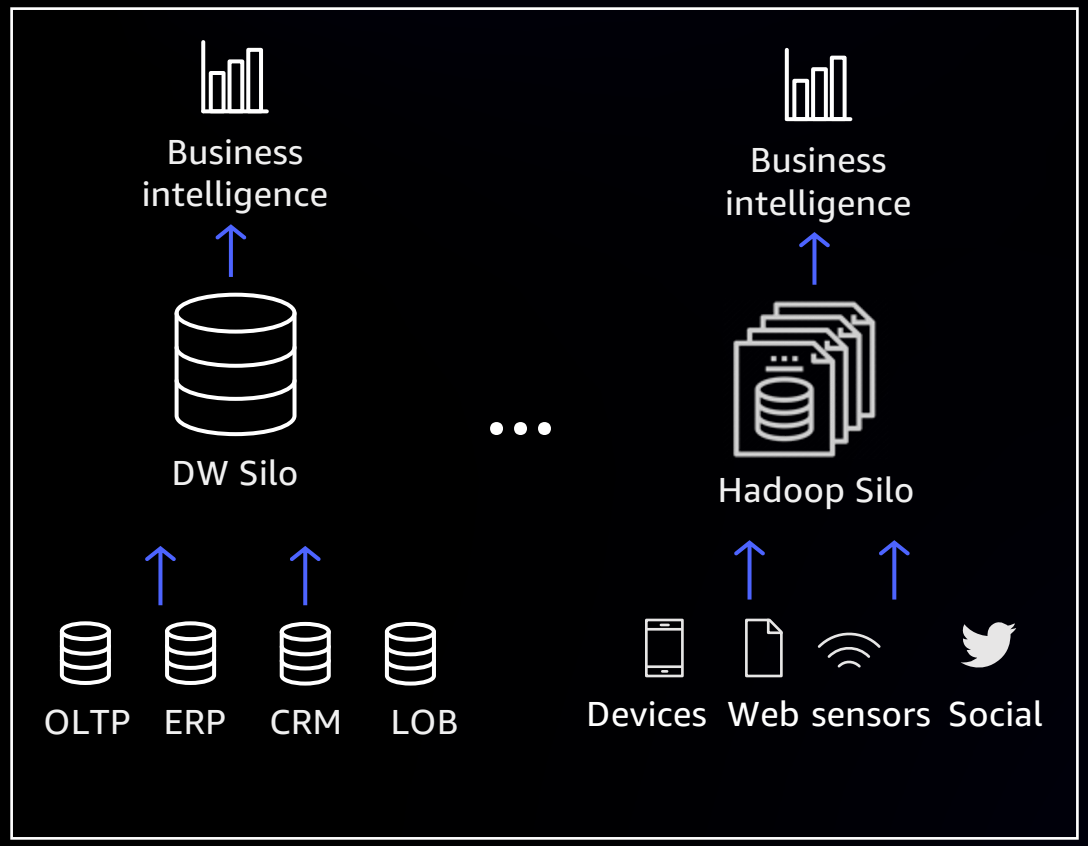
Why AWS Lake Formation?

Our strategy, features & use cases

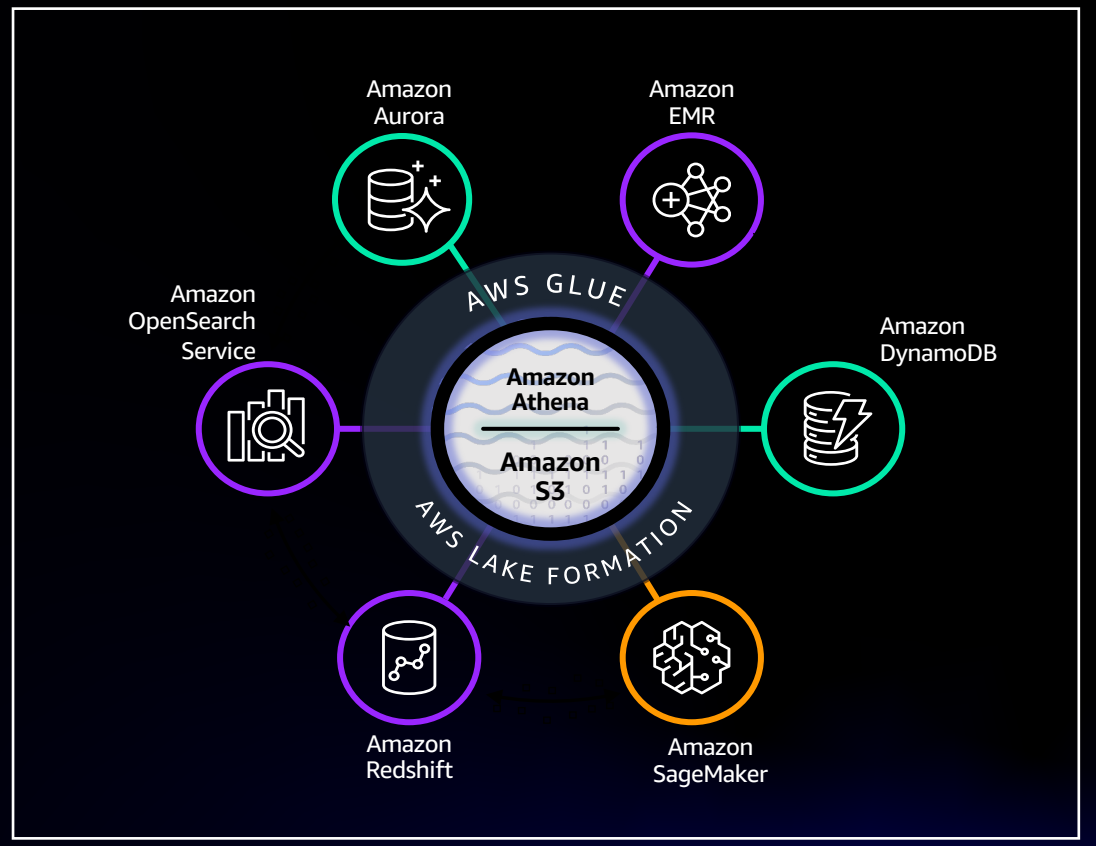


Customers are moving from...

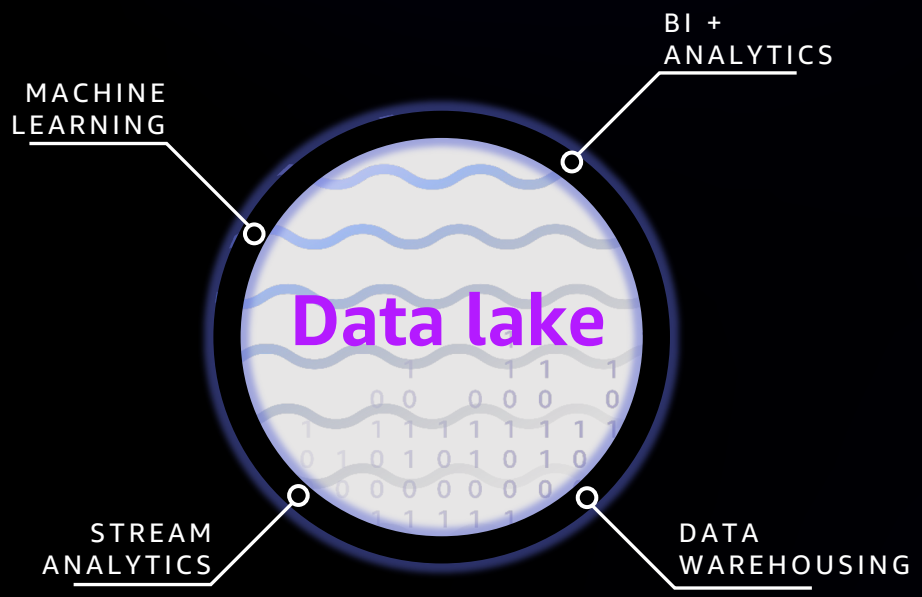
Siloed data



Modern data architectures



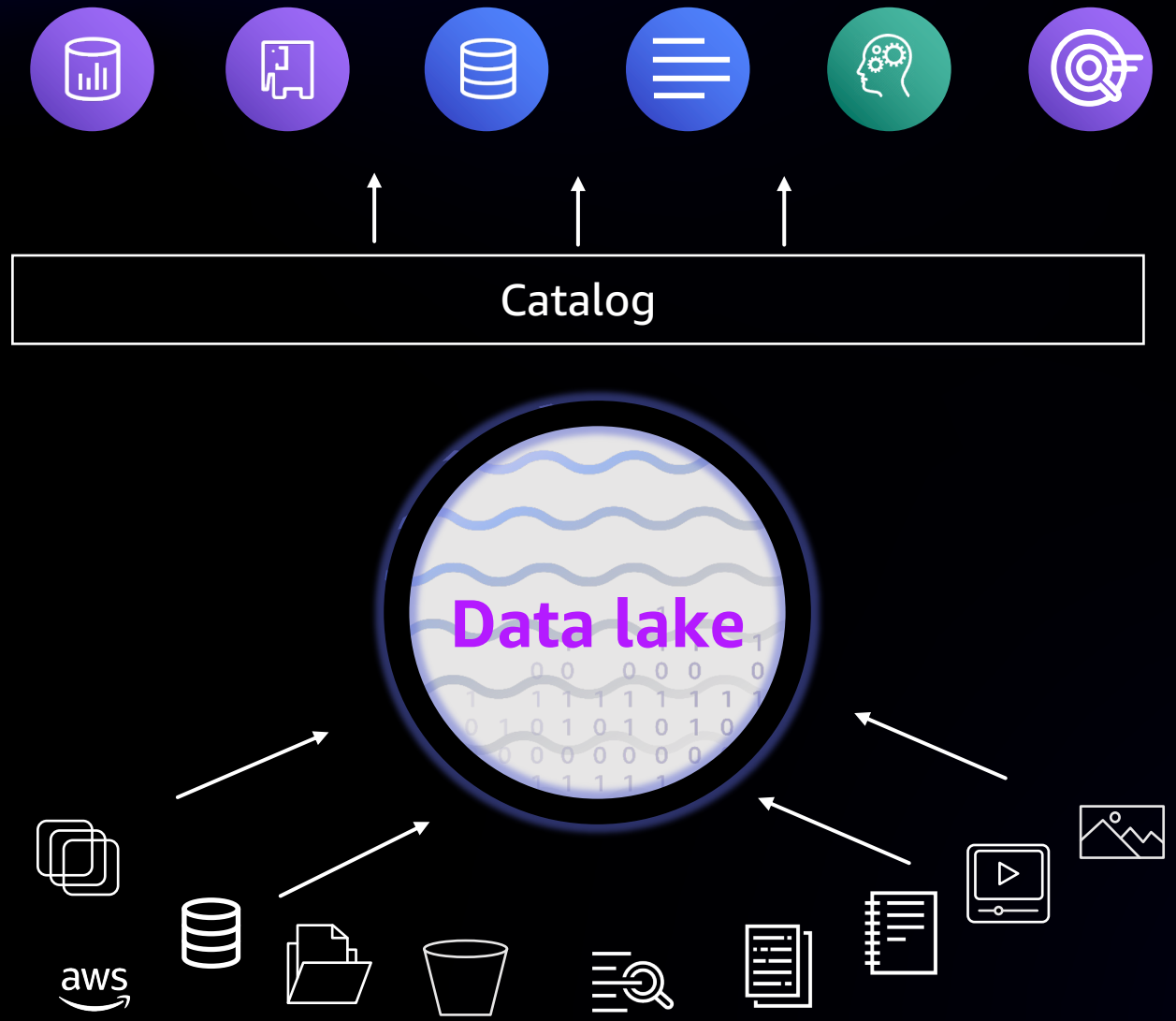
Data lake: Scalable and open



A **flexible**, **secure** repository that enables you to **govern**, **discover**, **share**, and **analyze structured** and **unstructured** data at any scale with the **tool of your choice**.

aws

The benefits of modern data lakes:



Store all your data

Cost effectively scale storage to EBs

Decouple storage from compute

Pay-as-you-go analytical and ML engines

Process data in-place

Amazon S3: most popular choice to build data lakes

Unmatched **durability, availability,** and **scalability**

Easiest to use with **cost optimization:** Intelligent tiering

Most ways to **get data in**

Most **object-level controls**

Broadest portfolio of analytics tools

Best **security, compliance,** and **audit** capabilities

Cold storage and archive capabilities

aws

Challenges in building data lakes

Challenge #1: Data Ingestion

Building reliable ingestion pipelines is complex.

Challenge #2: Data management

Managing how data is stored and optimized in S3 is time consuming.

Challenge #3: Security & governance

Managing permissions at scale is difficult and error-prone.

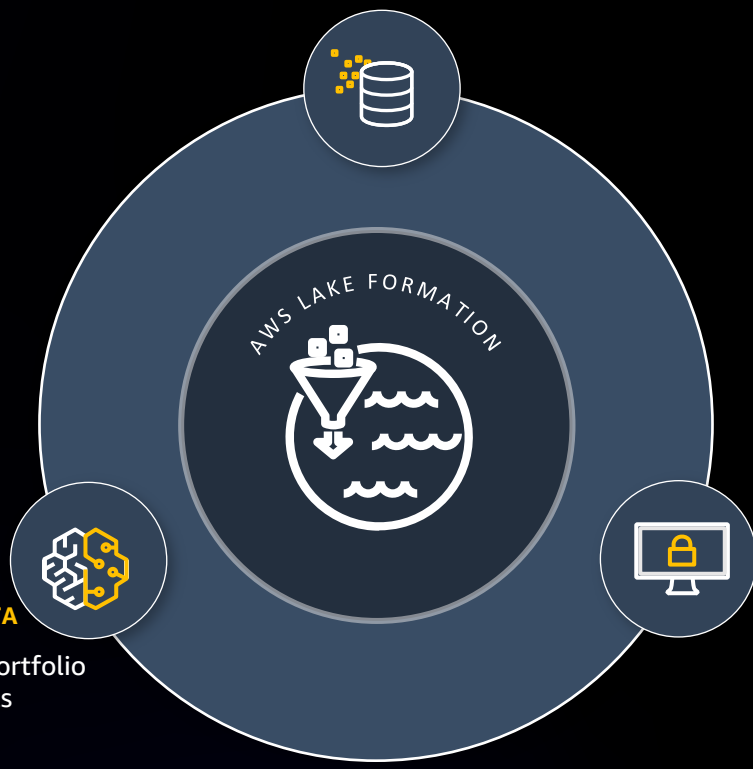
Challenge #4: Data sharing

Sharing across accounts and organizations is cumbersome.

Challenge #5: Integrations

A choice of integrated services is critical for productivity.

AWS Lake Formation is a fully managed serverless service that allows you to build clean and secure data lakes in days



INGESTION AND IDENTIFICATION

Blueprints simplify ingest, store and catalog faster
ML transforms for data cleaning

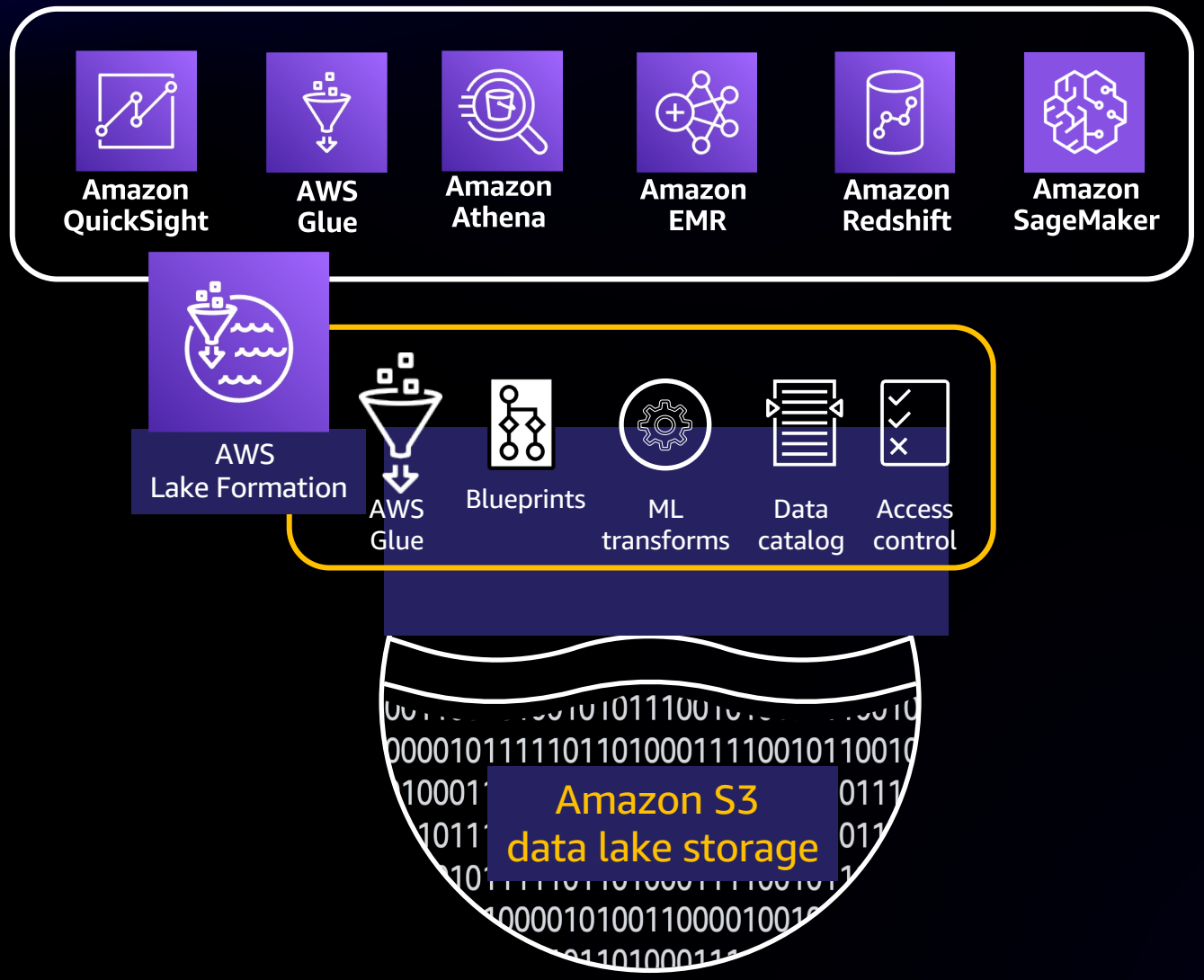
SIMPLIFIED SECURITY

Centralized permissions table and column
Real-time monitoring and integrated auditing

SELF-SERVICE ACCESS TO DATA

Comprehensive portfolio of integrated tools

AWS Lake Formation: The foundation for data lakes



Single place to manage access

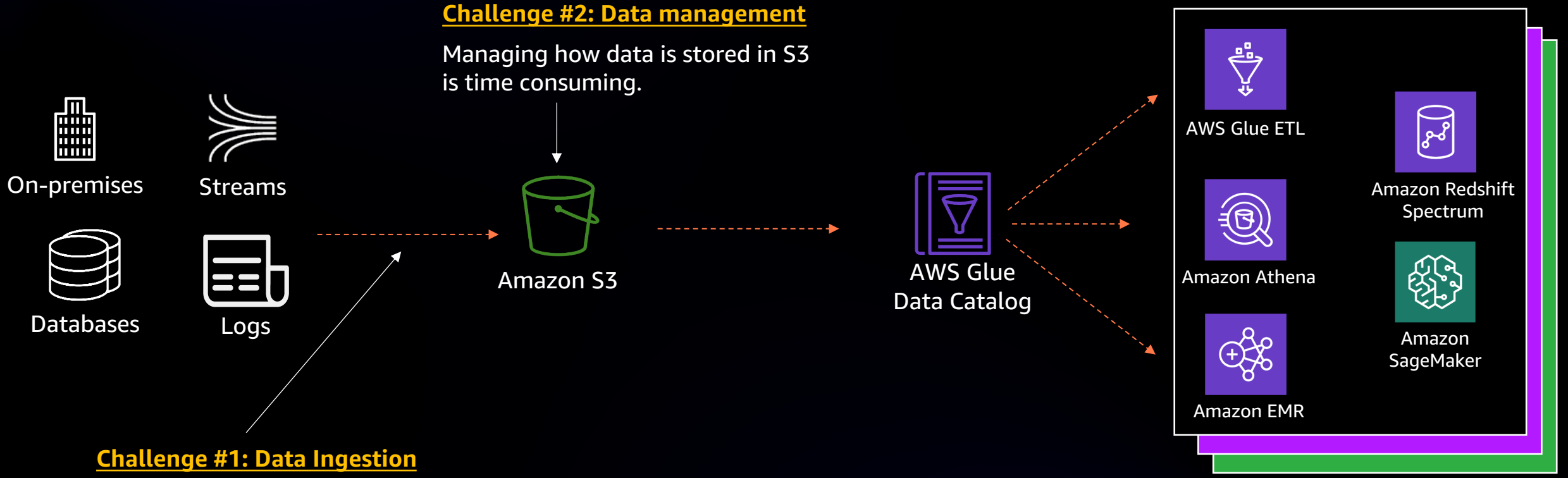
Open file formats

Efficient sharing

Ecosystem of integrated tools

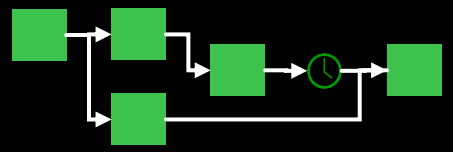
Cost effective

Challenge: Data ingestion & management



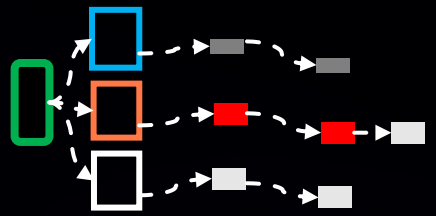
Why is data ingestion and management hard?

CONTINUOUS UPDATES



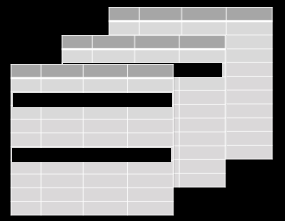
COMPLEX ETL
DELAYS IN DATA FRESHNESS
EXPENSIVE, BRITTLE & ERROR-PRONE

INCONSISTENT PERFORMANCE



DATA STORED HOW IT ARRIVED
LOTS OF SMALL FILES
PARTITION UPDATES

COMPLYING WITH REGULATIONS



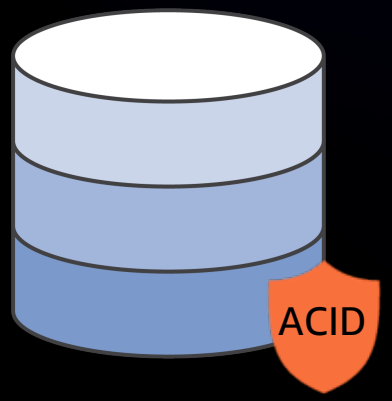
DIFFICULT TO FIND NEEDLE IN VERY LARGE HAYSTACK

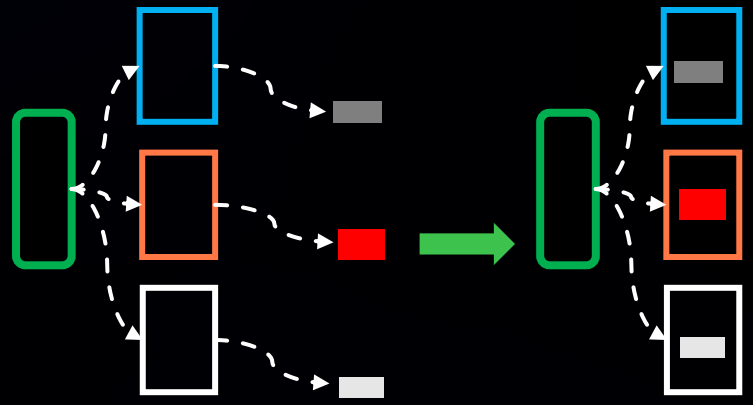
aws

NEW!

AWS Lake Formation Governed Tables

SIMPLIFY DATA INGESTION AND DATA MANAGEMENT





ACID transactions

Consistent across tasks
Insert, update, delete
Converge batch & real-time

Storage optimization

Auto-compact small files
Push-down filters
Reduce data scan

Time travel

Data history
Reproduce experiments
Audit changed data

Reliable

Performant

Versioned

aws

Preview

AWS Lake Formation: Transactions, row-level security, and acceleration

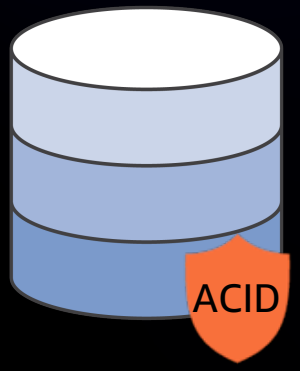
ACCELERATE AND GOVERN ACCESS TO YOUR AMAZON S3 DATA LAKE

https://aws.amazon.com/lake-formation/preview/

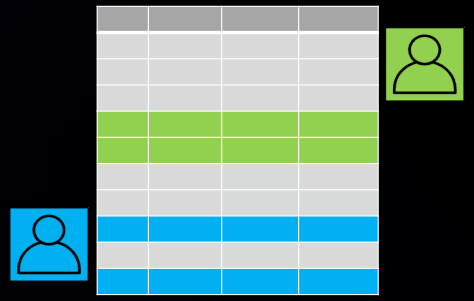
aws

New AWS Lake Formation update and access APIs

- Open and public: Build your own application



Atomic, consistent, isolated, and durable (ACID) transactions



Row-level filtering for security

Optimizations for fast analytics on S3 data lakes

AWS Glue

Amazon Athena

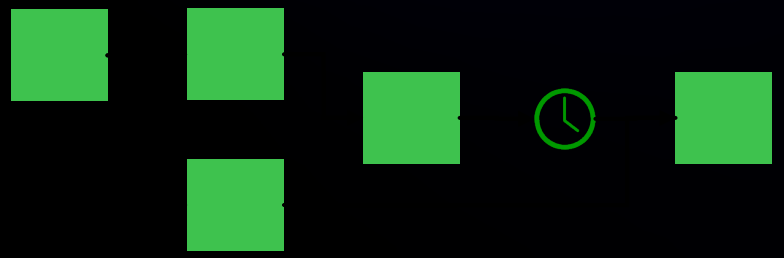
Amazon EMR

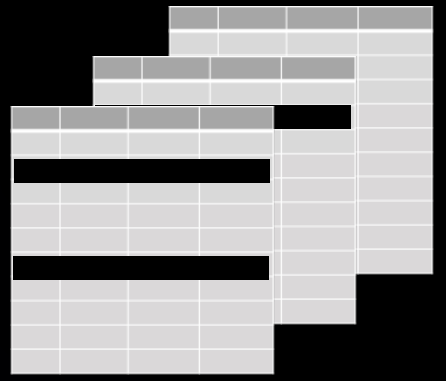
Amazon Redshift

Amazon QuickSight

Why transactions?

Continuous ingest and updates



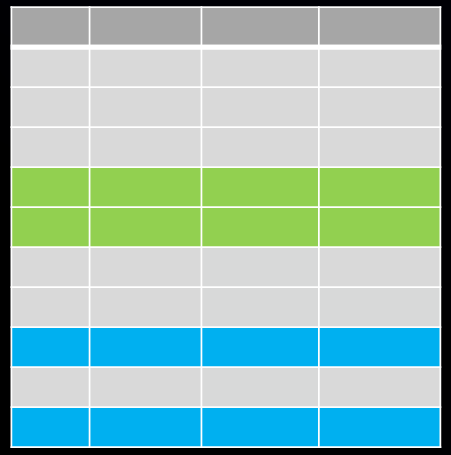


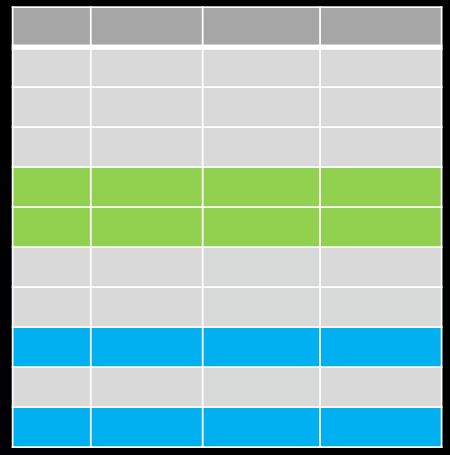
Easy to build applications on Amazon S3 data lakes

Allow for concurrent inserts, updates, and deletes that appear immediately

aws

Why row-level security?

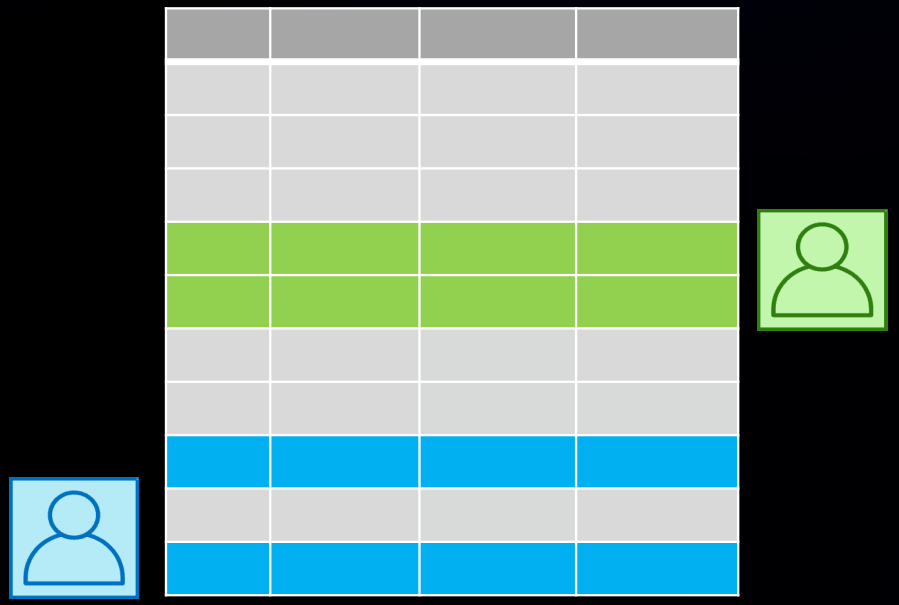




Today, requires multiple redacted datasets

AWS Lake Formation row-level security

- Read APIs uniformly enforce granular compliance policies



row-level

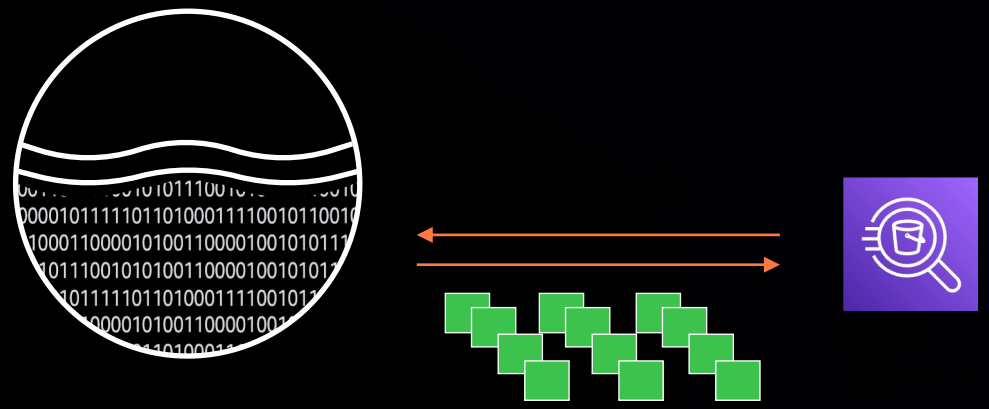
many formats

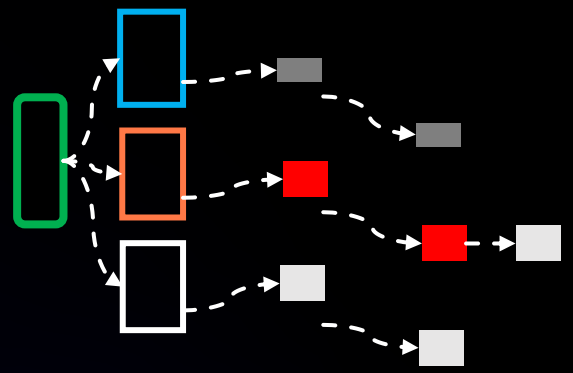
Open managed

easy to audit permissions and access

Why acceleration?

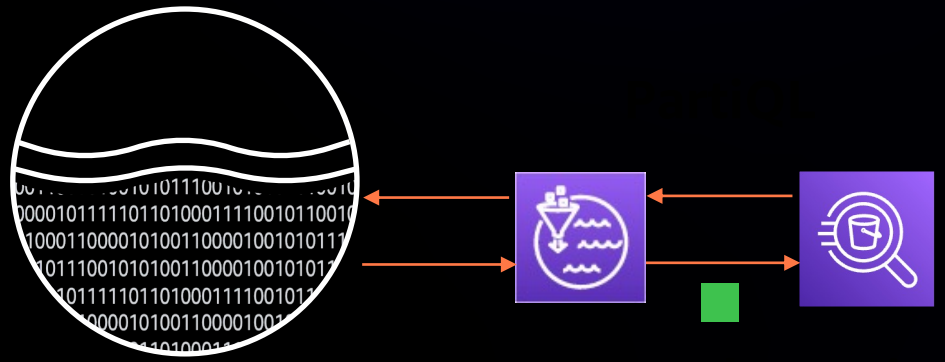
- Fast access for analytics



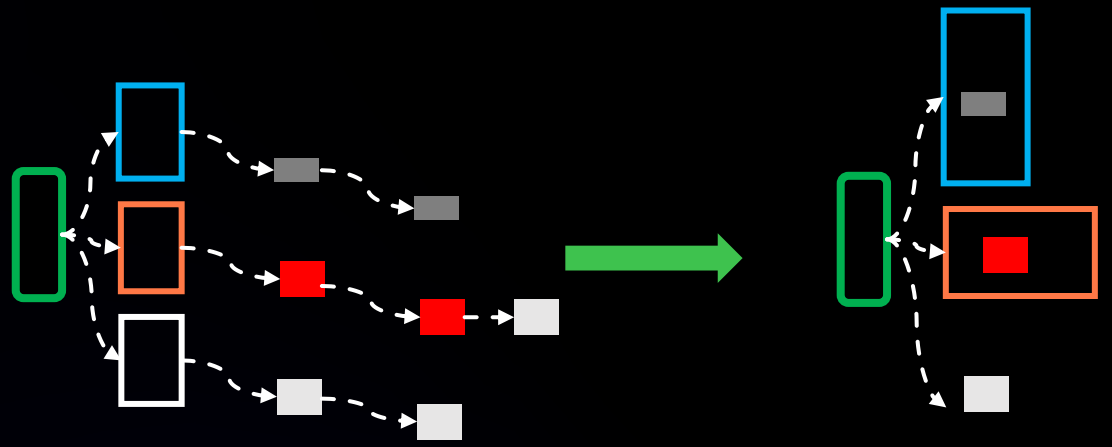


Acceleration capabilities

- Pushdowns and storage optimizer

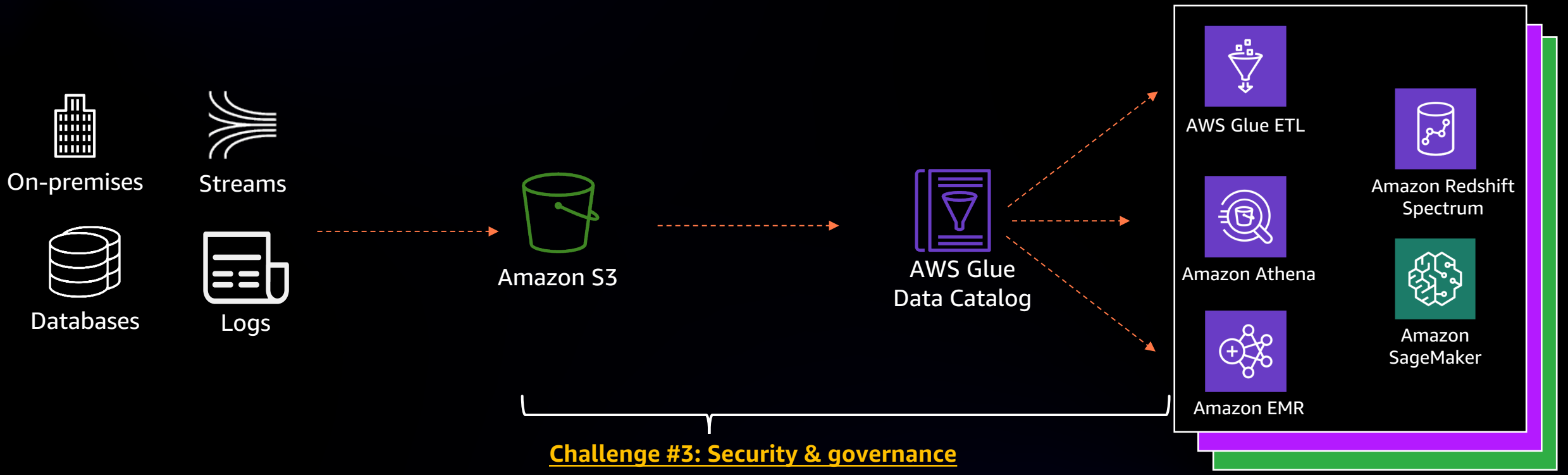


PartiQL



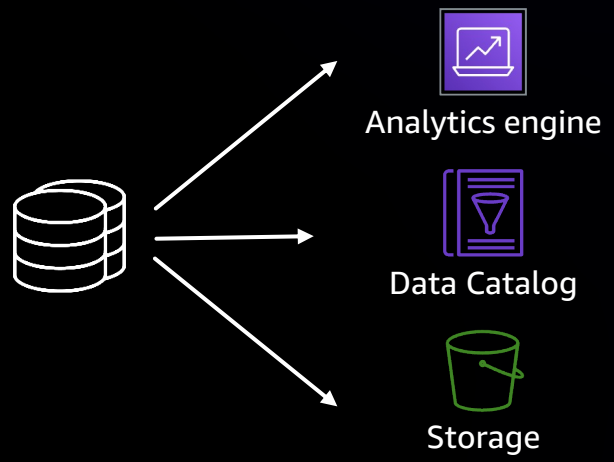
compacts

Challenge: Security and governance



Why is securing data lakes hard?

Unifying permissions across the data lake stack

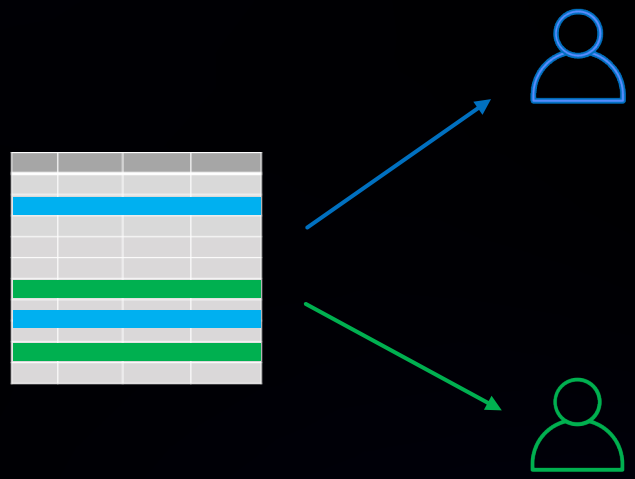


Split storage, metadata, & compute

Each system has different permissions

Syncing permissions is error-prone

Enforcing fine-grained permissions to restrict access

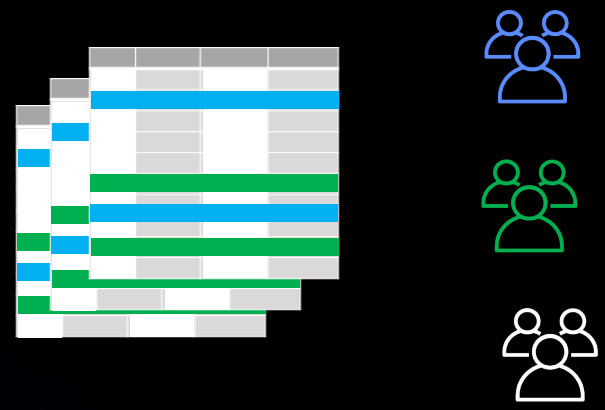


Data lakes contain a lot of data

Users should only access portions

Slice and dice the data into portions

Scalable permissions to manage data and users

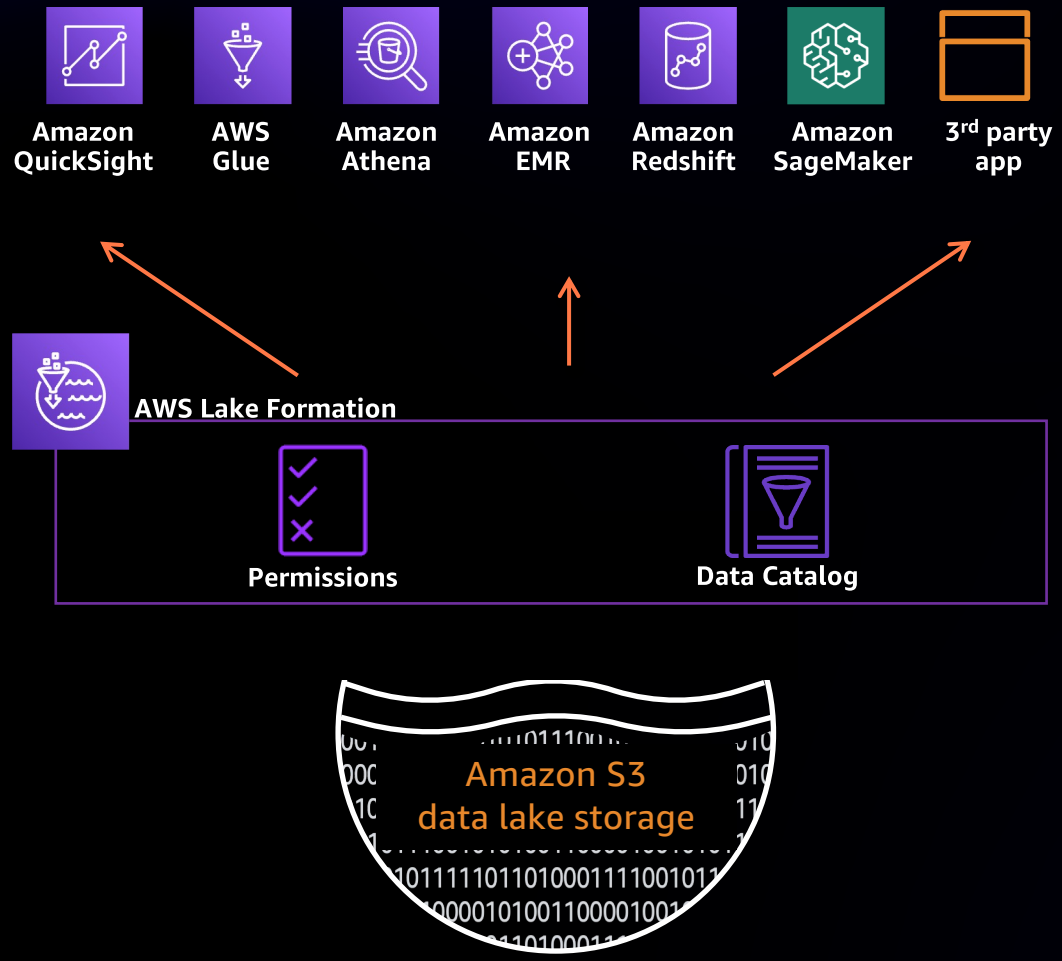


1,000s of DBs and tables

10,000s of users

New data sets added constantly

AWS Lake Formation permissions model

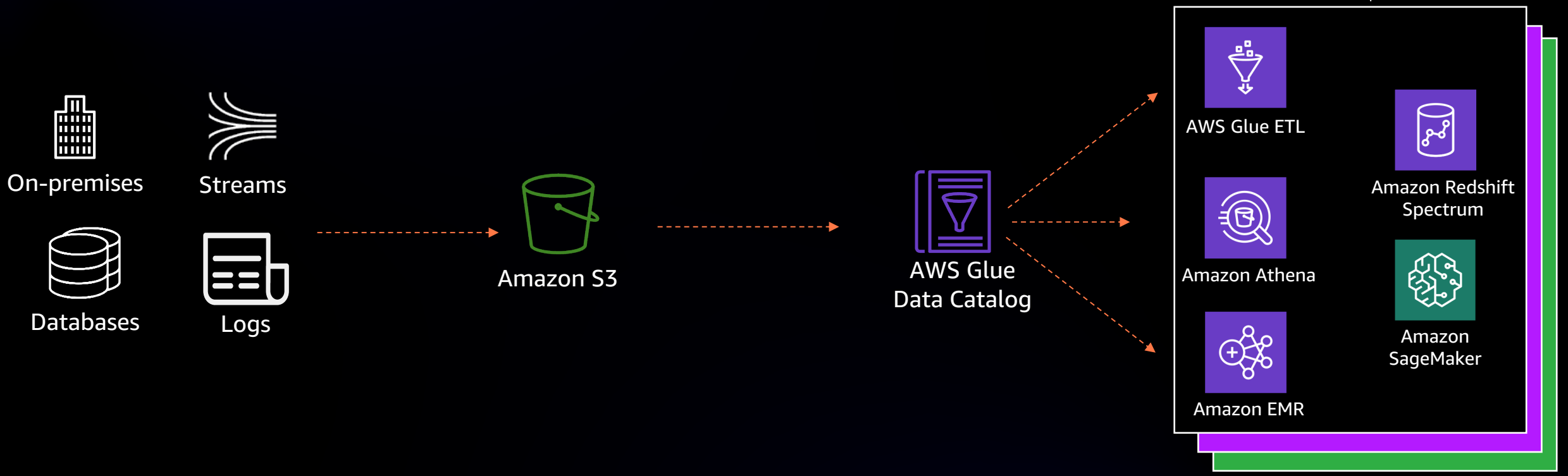


- DB style fine-grained permissions
- Fine-grained permissions on catalog resources
- S3 access managed by permission on resources
- LF-Tag based access control (LF-TBAC) to scale
- Integrated with services and tools
- Easy to audit permissions and access

Challenge: Data sharing

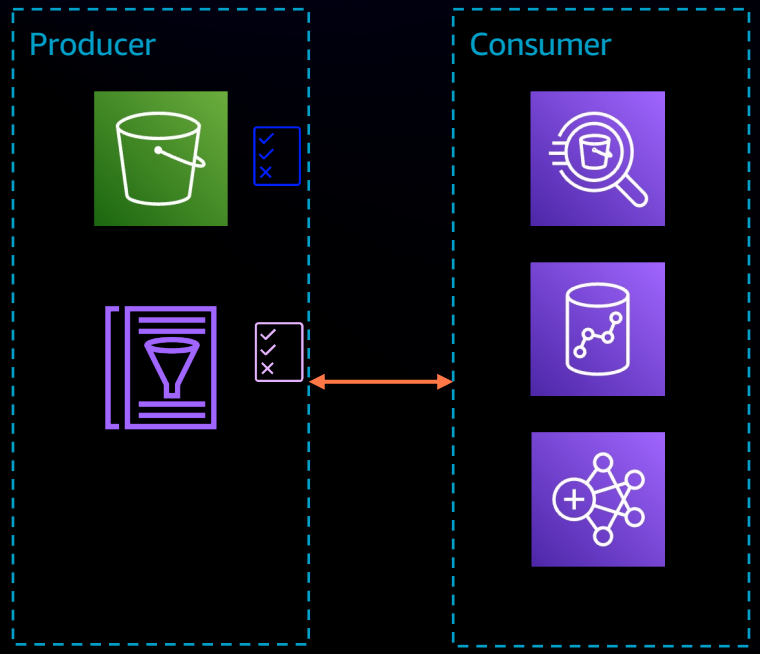
Challenge #4: Data sharing

Sharing across accounts and organizations is cumbersome.



Why is sharing data across accounts hard?

To share data…



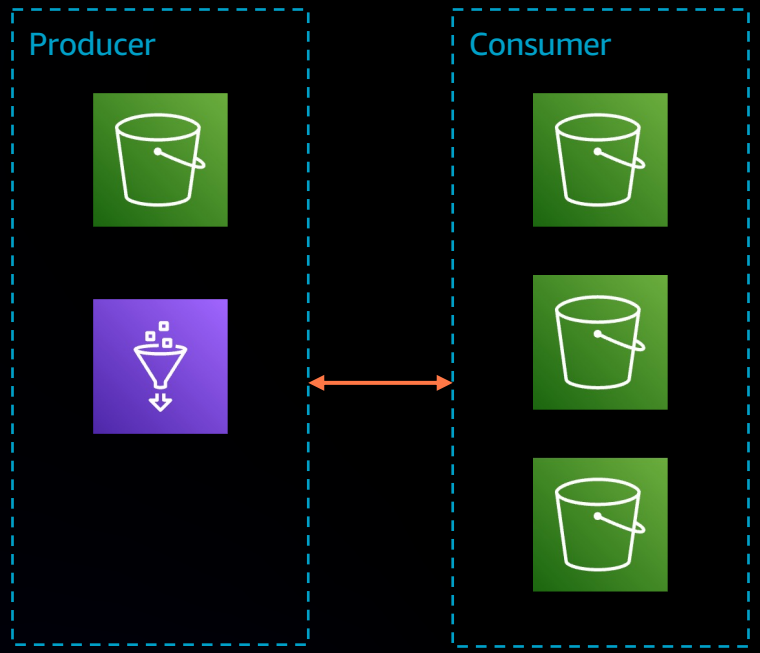
S3 and IAM policies

Limited by service support

Lacks discoverability

Policy size limits (coarse grained)

Duplicating data

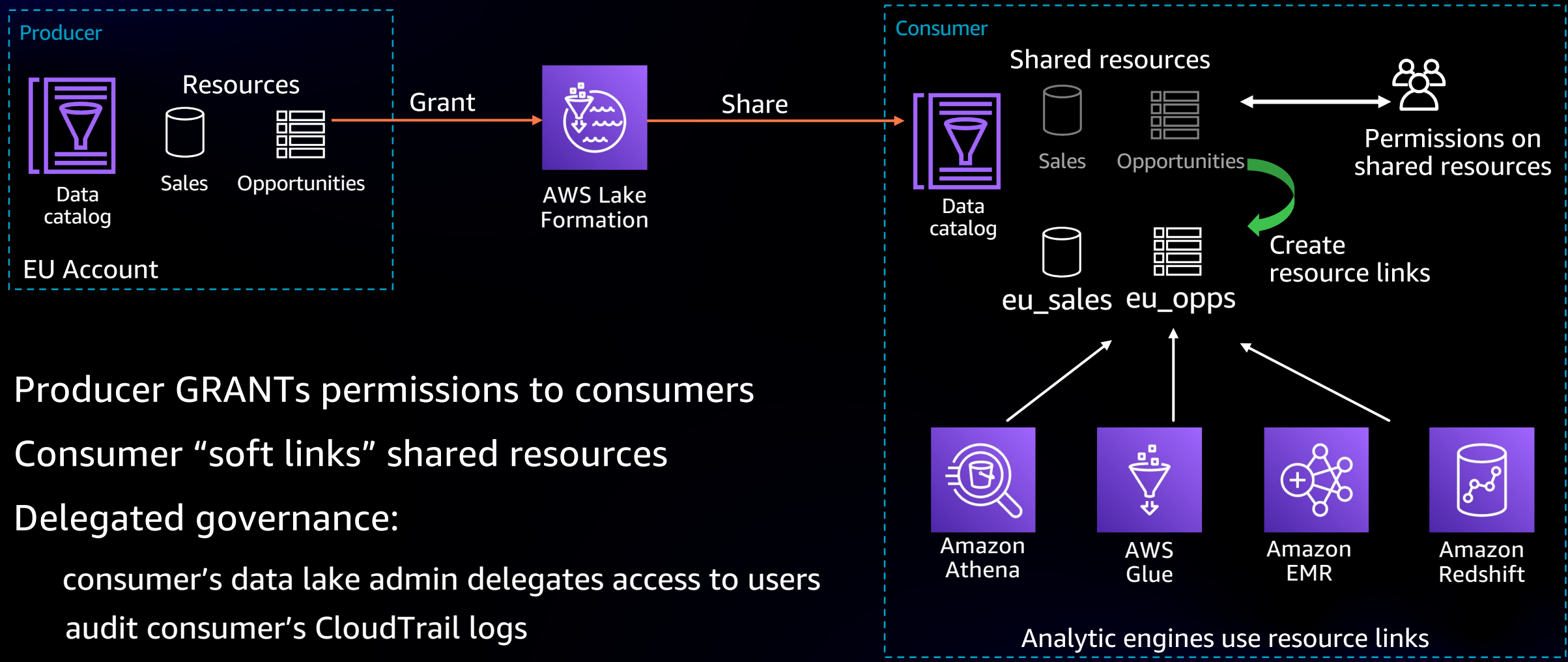


ETL pipelines

Multiple redacted copies

Expensive, brittle, and error-prone

AWS Lake Formation cross-account sharing



Producer GRANTs permissions to consumers

Consumer “soft links” shared resources

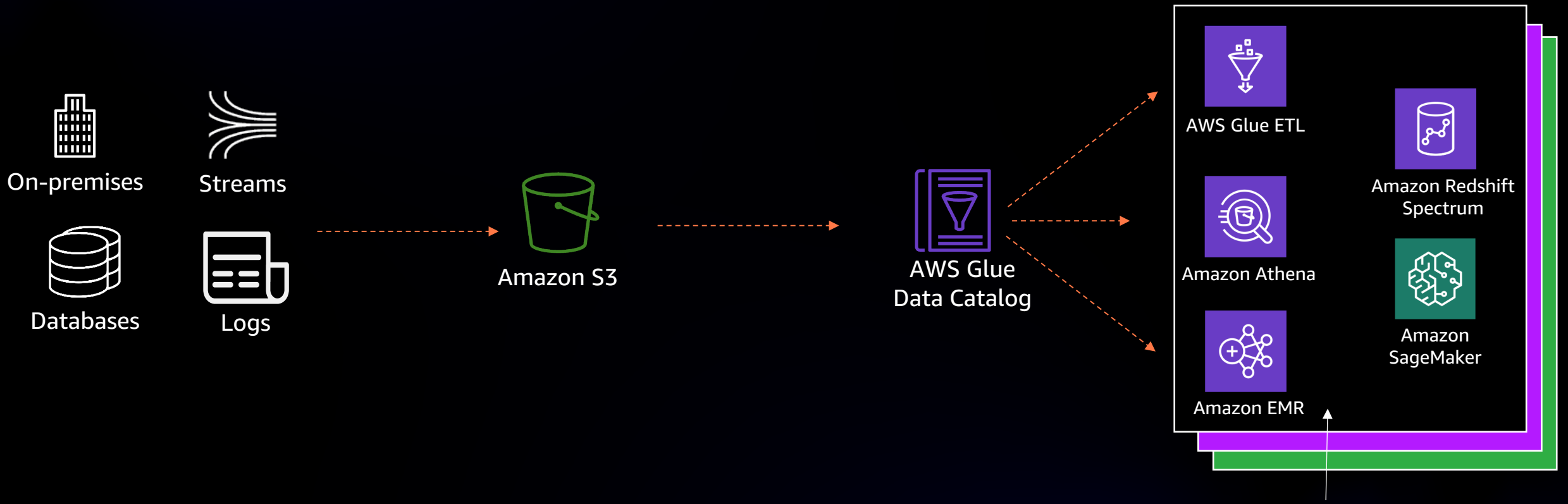
Delegated governance:

- consumer’s data lake admin delegates access to users
- audit consumer’s CloudTrail logs

Optional centralized auditing with event forwarding

aws

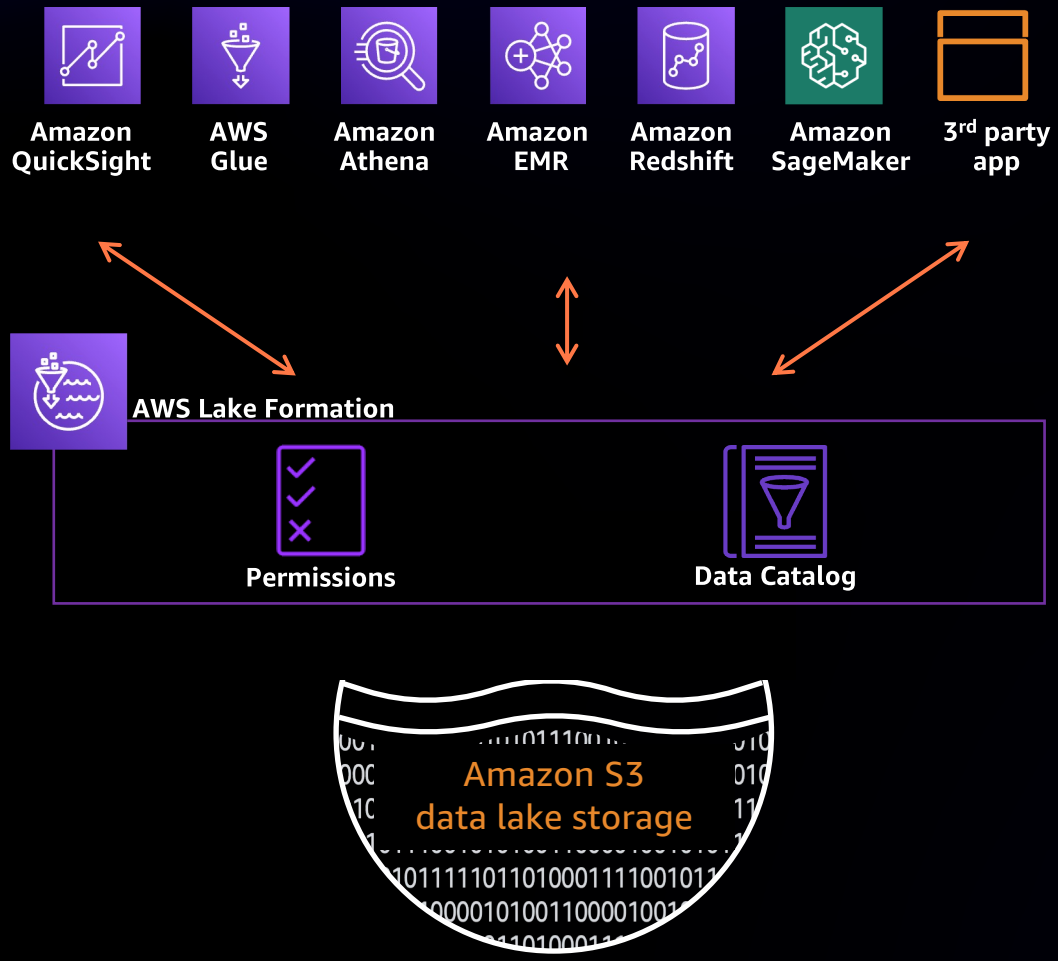
Challenge: Integrations



Challenge #5: Integrations

A large set of integrated services is critical for productivity.

AWS Lake Formation integrations



Multiple integration options to

Credential vending APIs

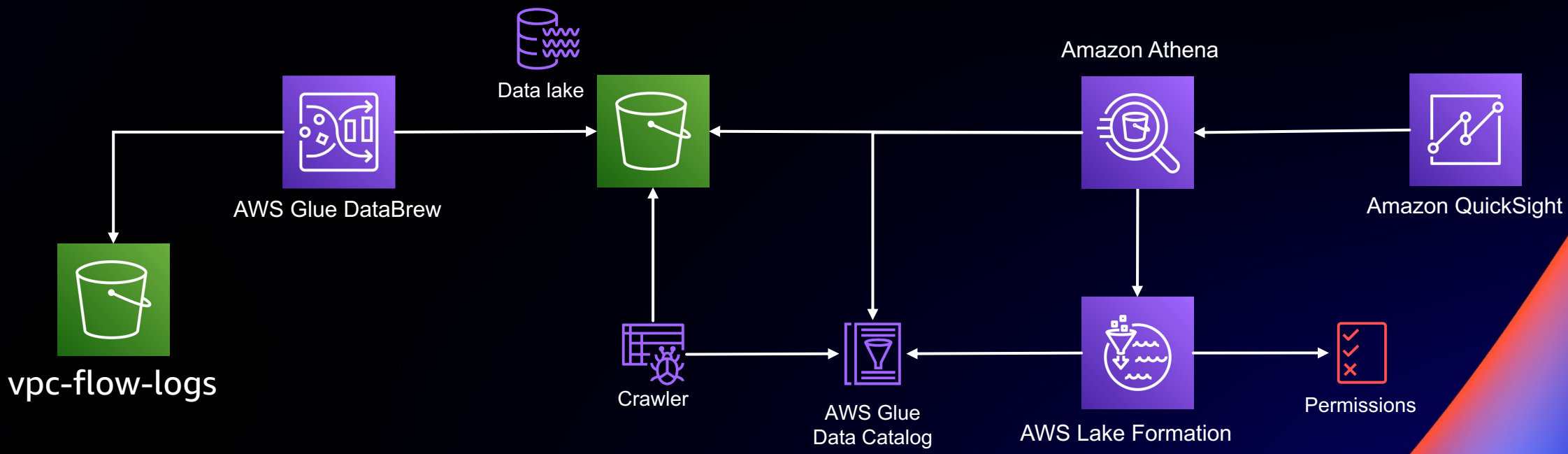
distributed enforcement w/ fail close

Universal Data Access APIs

centralized enforcement
for simplified integrations

Demo

Demo Architecture



Thank you!

John Rice

jkrice@amazon.com