
AWS
re:Invent

WPS302

Building analytics workloads in highly regulated environments

Ben Snively

Specialist Solutions Architect

AWS

AWS re:Invent

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

Building analytics workloads in highly regulated environments

Introduction to highly regulated environments

Core components to analytics workloads



AWS re:Invent

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

Highly regulated industries

Financial services

Healthcare

Legal

Government

. . . and many others



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

Grand River Hospital builds data lake on AWS, achieves “seamless business continuity”

Challenge

Build a data lake to house the hospital’s sensitive patient and administrative data while retiring its legacy hospital information systems composed of electronic patient record and other administrative systems

Solution

Amazon Simple Storage Service (Amazon S3) as the heart of the data lake, Amazon S3 Glacier for archival, AWS Glue for data cataloging and transformation, and AWS Database Migration Service (AWS DMS) for migrating data from legacy commercial databases

Benefits

- Provide a platform that contains not only legacy data but also new data from multiple systems and sources
- Working with AWS and Deloitte, we were able to meet our deadlines on budget

“Through the procurement process, we shortlisted a number of options for cloud providers and ultimately determined we needed a provider who could give us the best value in terms of technology capabilities, scalability, cost, service, and relevant experience to the Canadian health marketplace. We have legal and ethical obligations as it relates to our data. We had to pick a provider we could trust.”

Young Lee, Vice President – Planning, Innovation, and Transformation at Grand River Hospital

GRAND RIVER HOSPITAL
Advancing Exceptional Care

Company: Grand River Hospital
Industry: Healthcare
Country: Canada
Employees: 3,000
Website: www.grhosp.on.ca

About Grand River Hospital

Grand River Hospital provides exceptional care through inspired people, innovative initiatives, and strong partnerships in the communities we serve.

aws

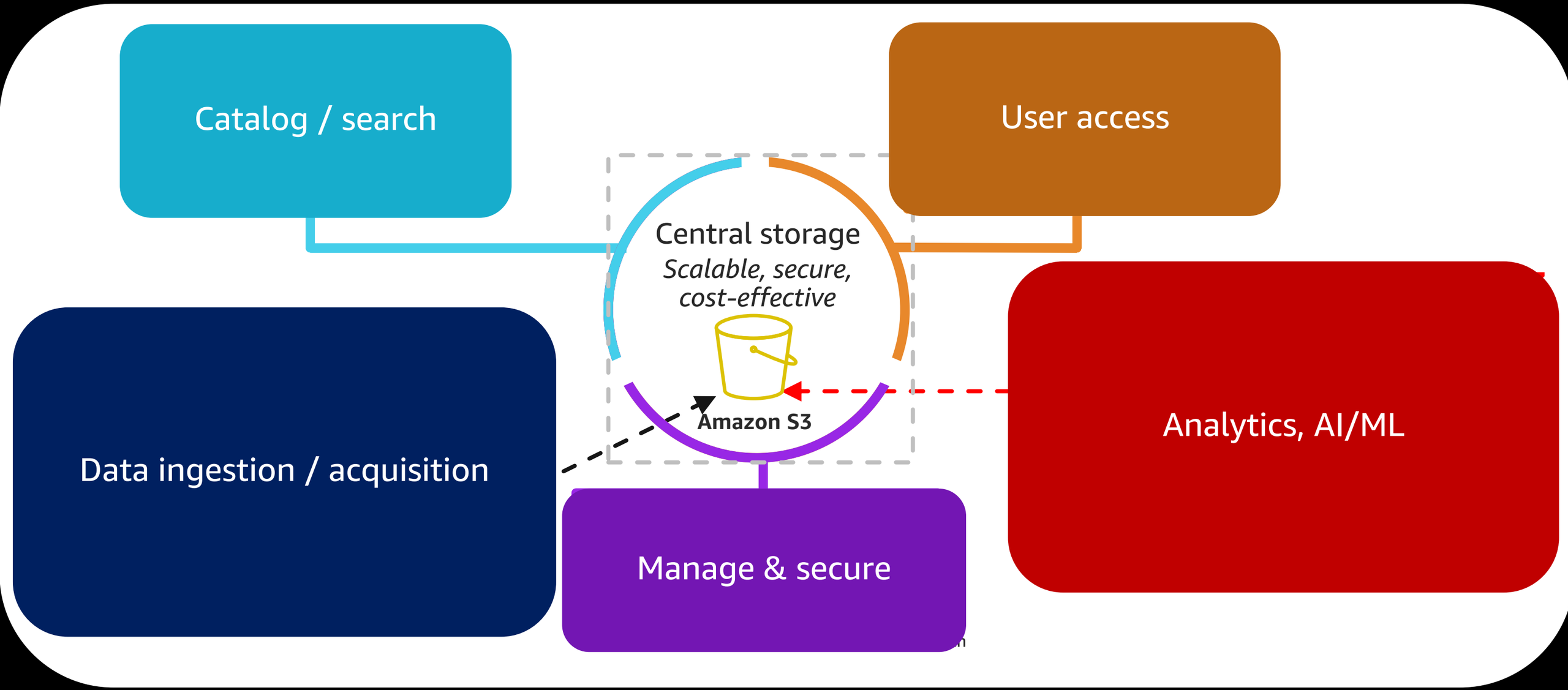
Core components of an analytics workload



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

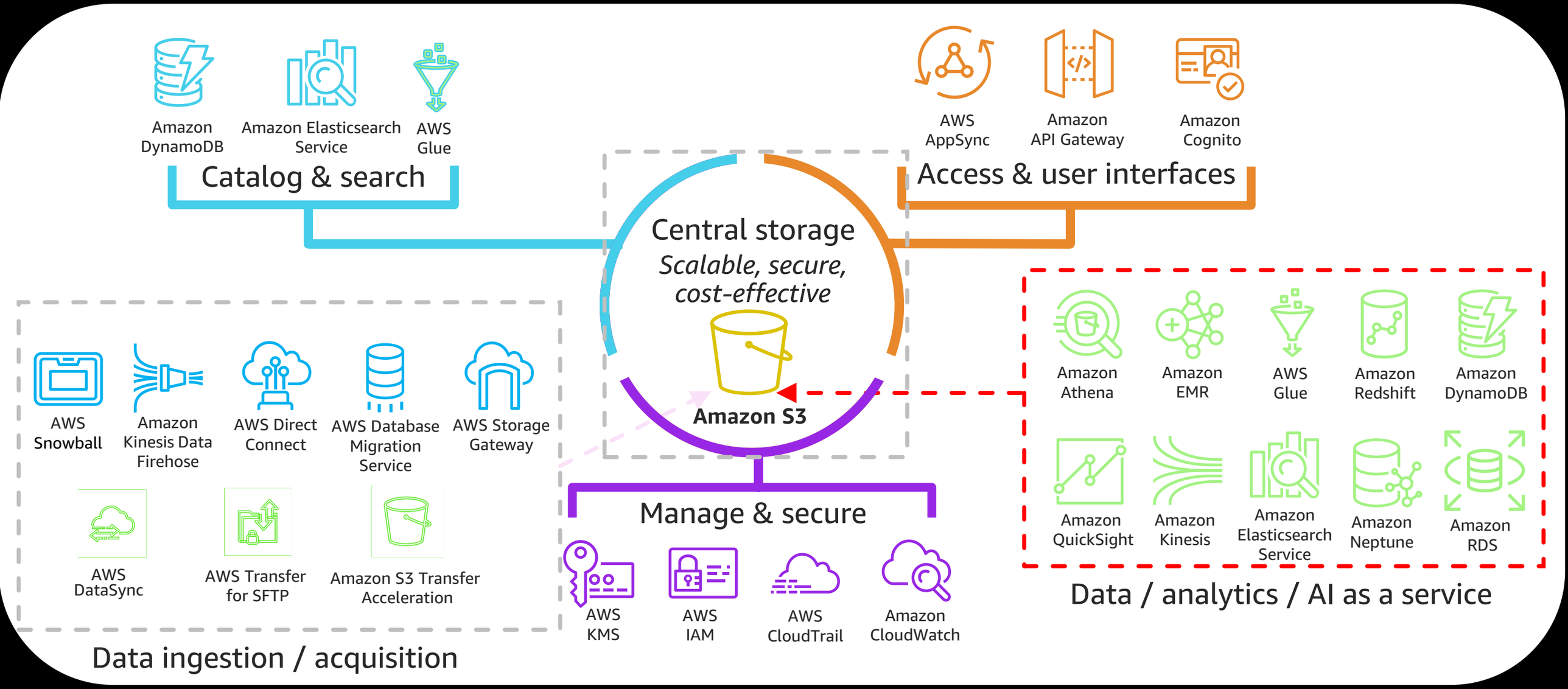
Layers of an analytics workload





Layers of an analytics workload





Data storage – Amazon S3

RIGHT STORAGE TIER FOR YOUR REQUIREMENTS





Standard

- Active, frequently accessed data
- Milliseconds access
- ≥ 3 AZ



Intelligent-Tiering

- Data with changing access patterns
- Milliseconds access
- ≥ 3 AZ
- Monitoring fee per object
- Min. storage duration



Standard-IA

- Infrequently accessed data
- Milliseconds access
- ≥ 3 AZ
- Retrieval fee per GB
- Min. storage duration
- Min. object size



One Zone-IA

- Recreatable, less accessed data
- Milliseconds access
- 1 AZ
- Retrieval fee per GB
- Min. storage duration
- Min. object size



Amazon S3 Glacier

- Archive data
- Select minutes or hours
- ≥ 3 AZ
- Retrieval fee per GB
- Min. storage duration
- Min. object size



S3 Glacier Deep Archive

- Long-term archive data
- Select hours
- ≥ 3 AZ
- Retrieval fee per GB
- Min. storage duration
- Min. object size

Frequent *Access frequency* ***Archive***

Data storage



Common concerns

Data should never be deleted

Ensure data integrity

Able to find and use the data

Maintain fine-grained permissions

Audit use of the data

Data storage



Common concerns

- Data should never be deleted
- Ensure data integrity
- Able to find and use the data
- Maintain fine-grained permissions
- Audit use of the data

To answer these questions

Let’s discuss the security controls

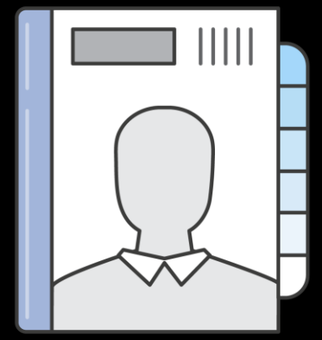


© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Identity and Access Management overview

Principals (identities)

- Identity of actors, known as principals (users, groups, roles)
- IAM / Directory can manage cloud identities, or you can federate through your own identity provider



API actions

- What does the API do, or how does it respond?

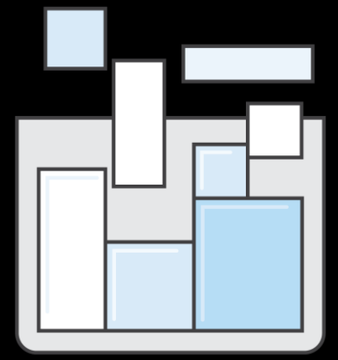


Resources

- AWS resources that IAM actions apply to
- Some resource types support resource policies that further define allowed actions

Conditions

- Precisely define required characteristics of identities, actions, resources, and other requests parameters to allow an action



IAM policy structure

```
{
 "Statement":[{
   "Effect":"effect",

"Principal":"principal",
   "Action":"action",
   "Resource":"arn",
   "Condition":{
     "condition":{
       "key":"value" }
      }
    }
   ]
 }
```

Principal: The entity that is allowed or denied access

"Principal":"AWS":"arn:aws:iam::123456789012:user/username"

Action: The type of action that is allowed or denied access

"Action":"s3:GetObject"

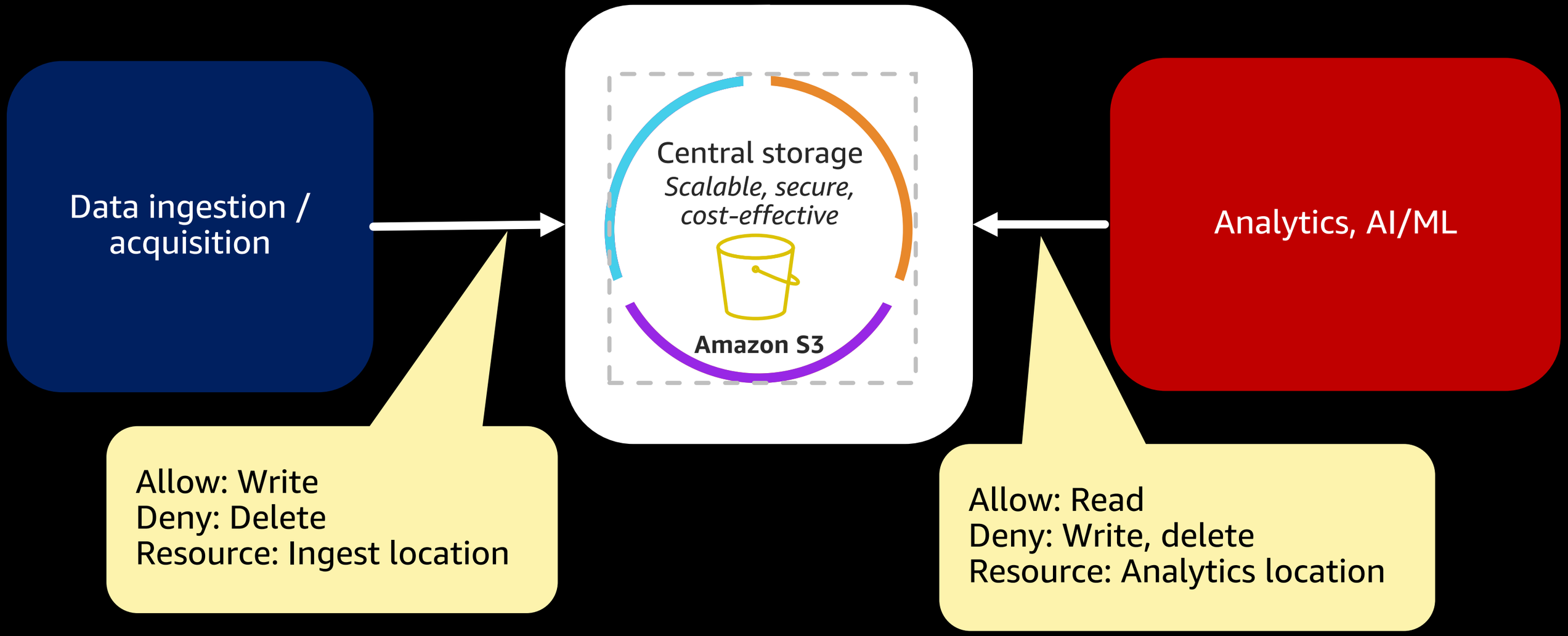
Resource: The Amazon resource(s) that the action will act on

"Resource":"arn:aws:sqs:us-west-2:123456789012:queue1"

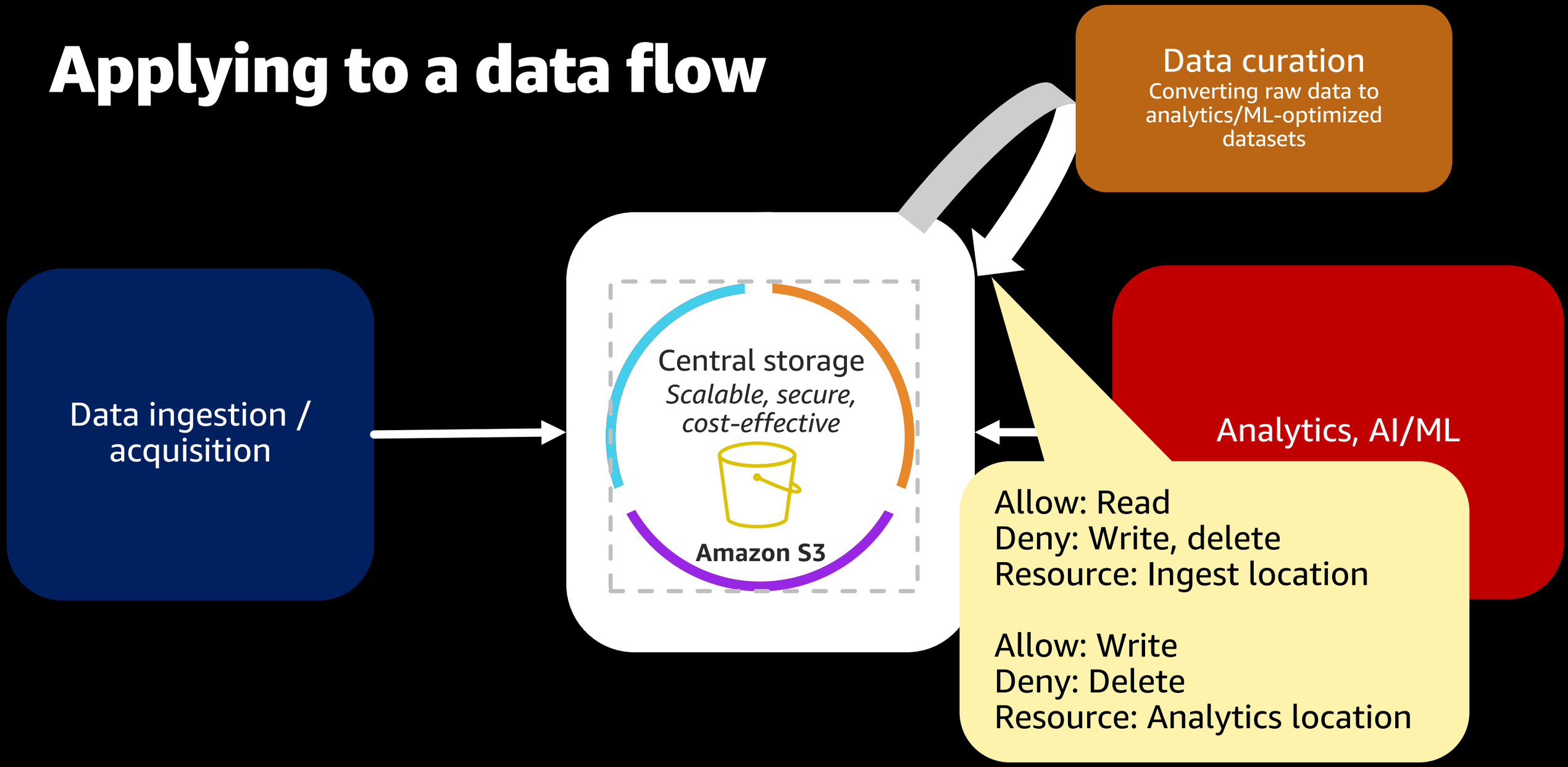
Condition: The conditions under the access defined as valid

"StringEqualsIfExists": {"aws:RequestTag/project": [“CodeA“]}

Applying to a data flow

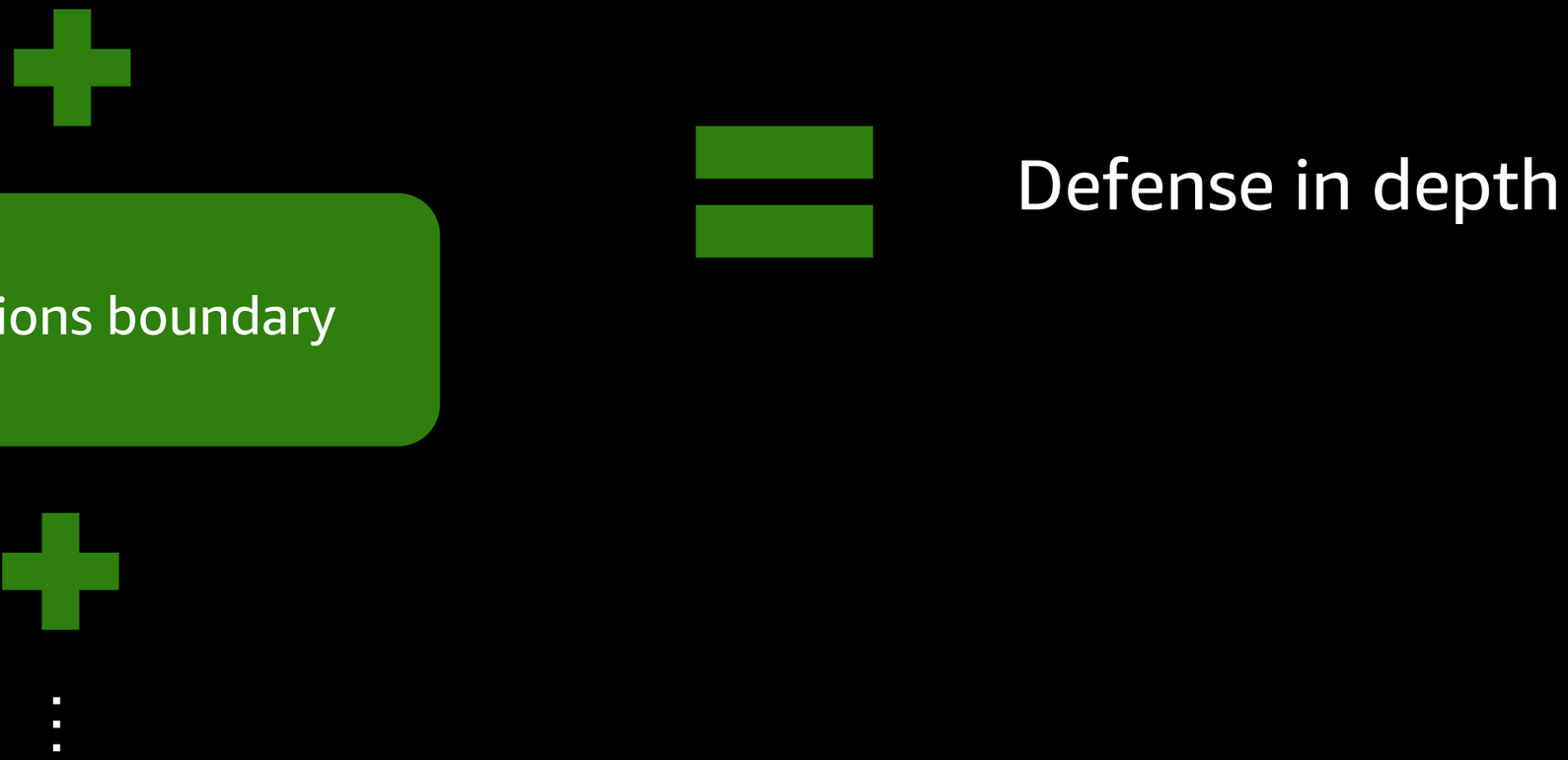


Applying to a data flow



Answering the question, “How do we preserve data?”

IAM policies only allowing the action for each process and user

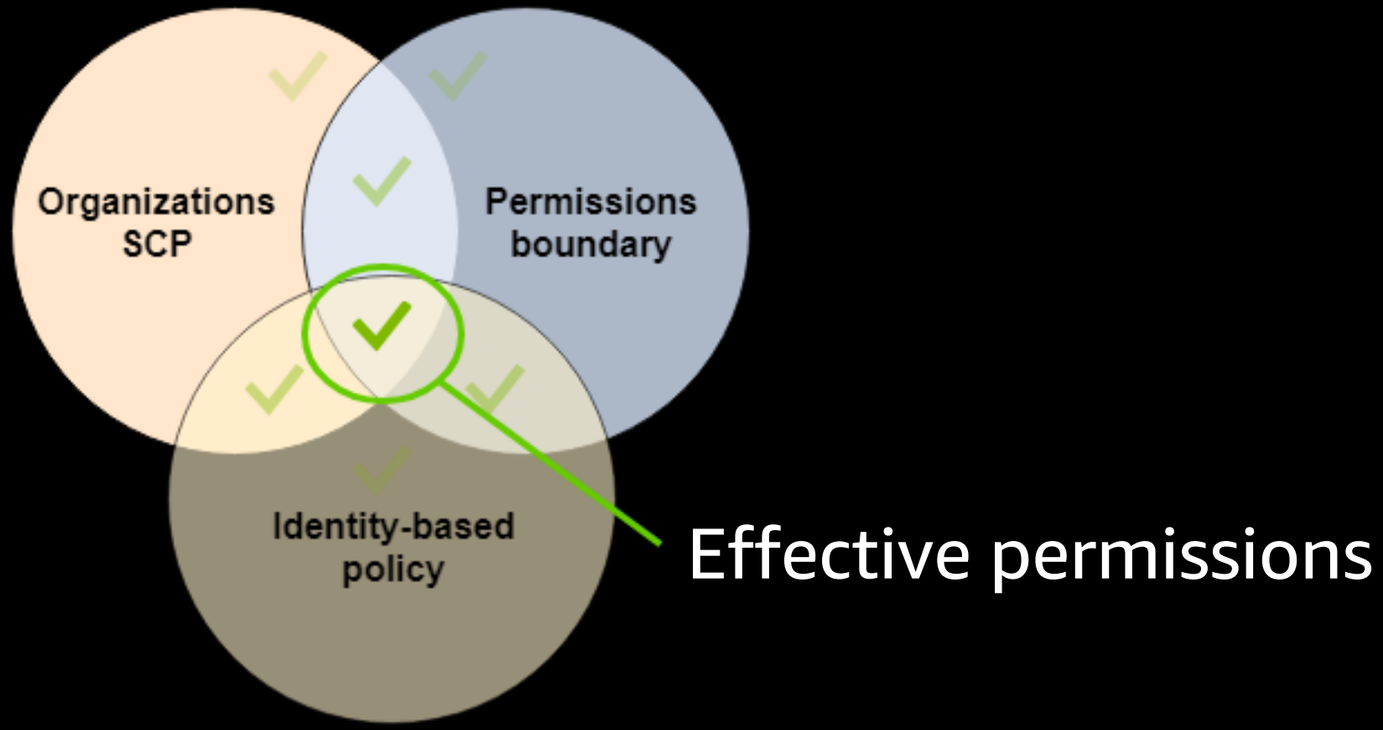


Defense in depth

Permissions boundaries

Scale and delegate permissions management to developers safely

Control the maximum permissions that employees can grant



Data storage



Common concerns

Data should never be deleted

Ensure data integrity

Able to find and use the data

Maintain fine-grained permissions

Audit use of the data

Layers of an analytics workload



Amazon DynamoDB
Amazon Elasticsearch Service
AWS Glue
Catalog & search

AWS AppSync
Amazon API Gateway
Amazon Cognito
Access & user interfaces

Central storage
Scalable, secure, cost-effective
Amazon S3

AWS Snowball
Amazon Kinesis Data Firehose
AWS Direct Connect
AWS Database Migration Service
AWS Storage Gateway
AWS DataSync
AWS Transfer for SFTP
Amazon S3 Transfer Acceleration
Data ingestion / acquisition

Manage & secure
AWS KMS
AWS IAM
AWS CloudTrail
Amazon CloudWatch

Amazon Athena
Amazon EMR
AWS Glue
Amazon Redshift
Amazon DynamoDB
Amazon QuickSight
Amazon Kinesis
Amazon Elasticsearch Service
Amazon Neptune
Amazon RDS
Data / analytics / AI as a service

AWS Glue components



Data catalog

Discover

Apache Hive Metastore compatible

Integrated with AWS services

Automatic crawling



Job authoring

Develop

Auto-generates ETL code

Python and Apache Spark

Edit, debug, and share



Job execution

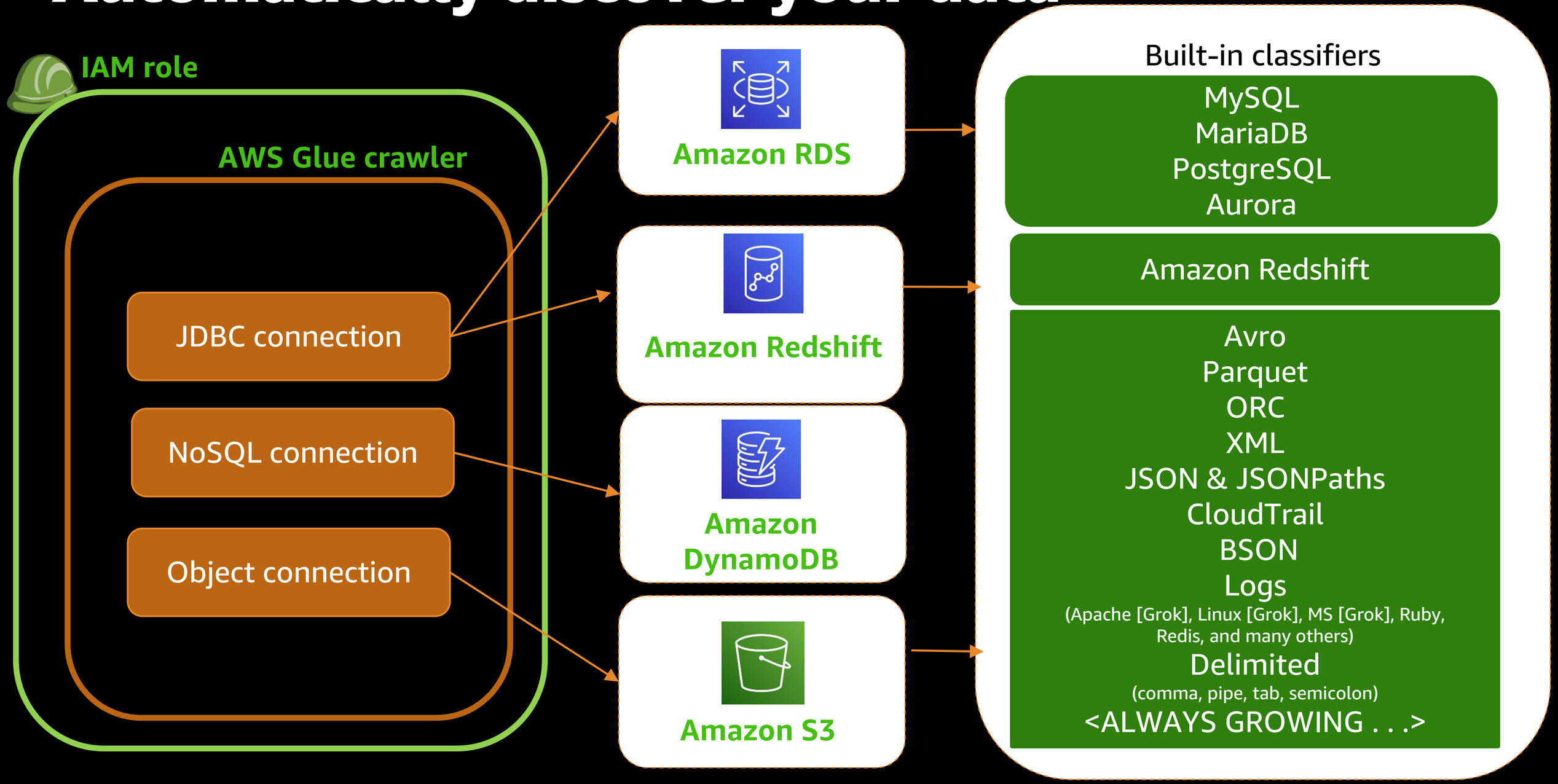
Deploy

Serverless execution

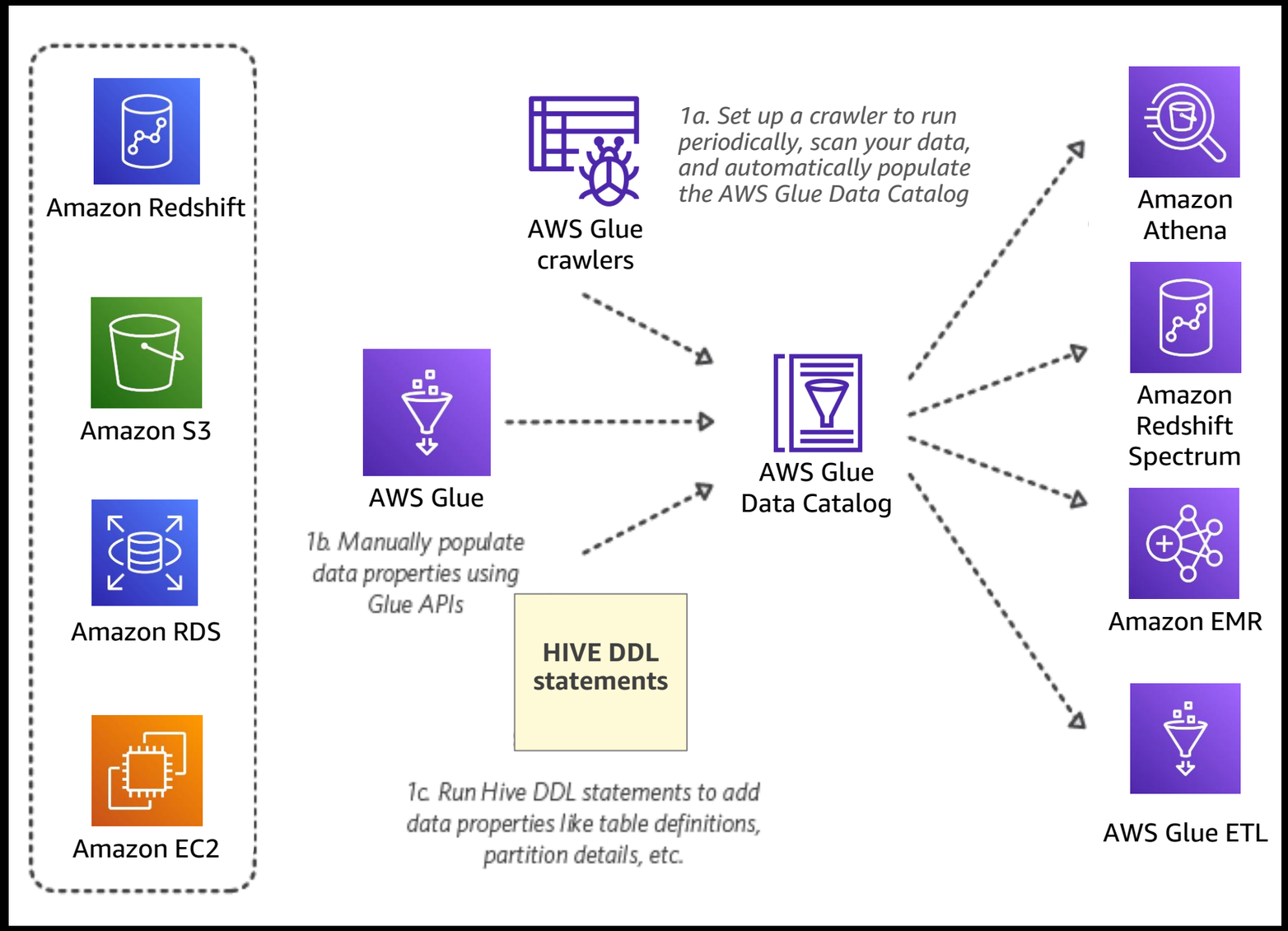
Flexible scheduling

Monitoring and alerting

Automatically discover your data



Future-proofing your catalog



Data storage



Common concerns

Data should never be deleted

Ensure data integrity

Able to find and use the data

Maintain fine-grained permissions

Audit use of the data

AWS Lake Formation

BUILD A SECURE DATA LAKE IN DAYS



Identify, ingest, clean, and transform data



Enforce security policies across multiple services



Gain and manage new insights

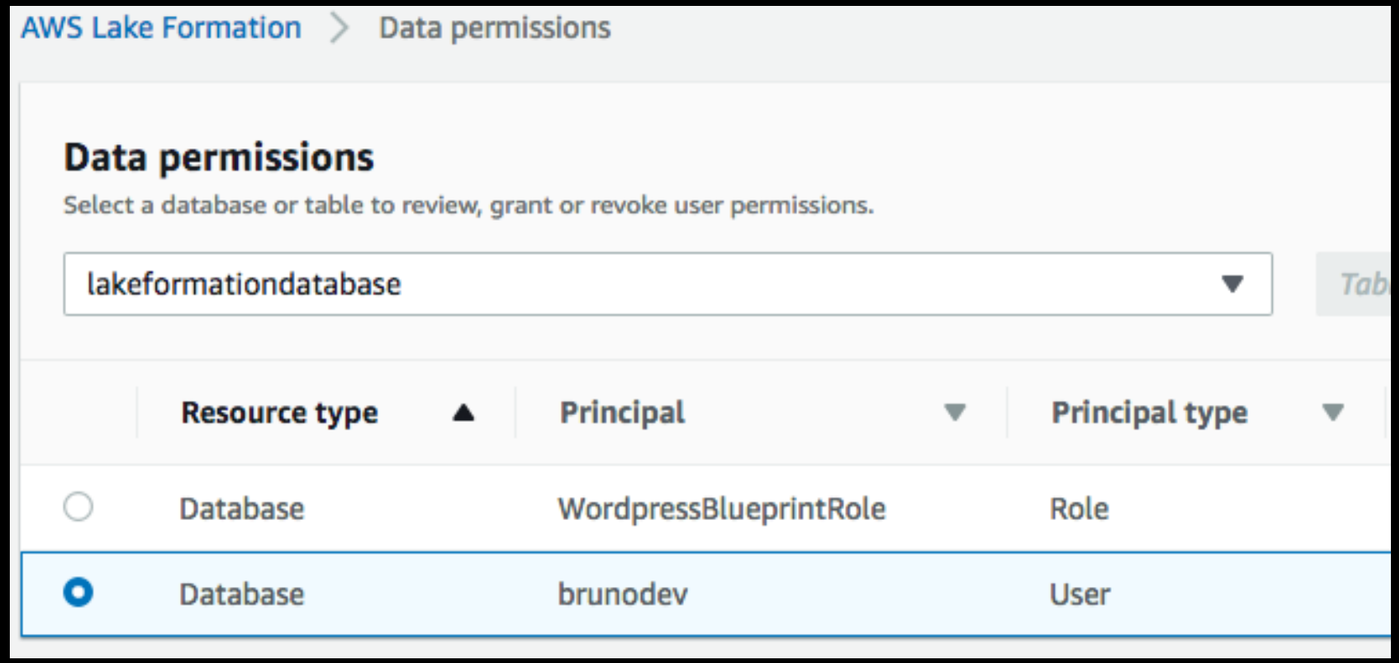
Fine-grained access control

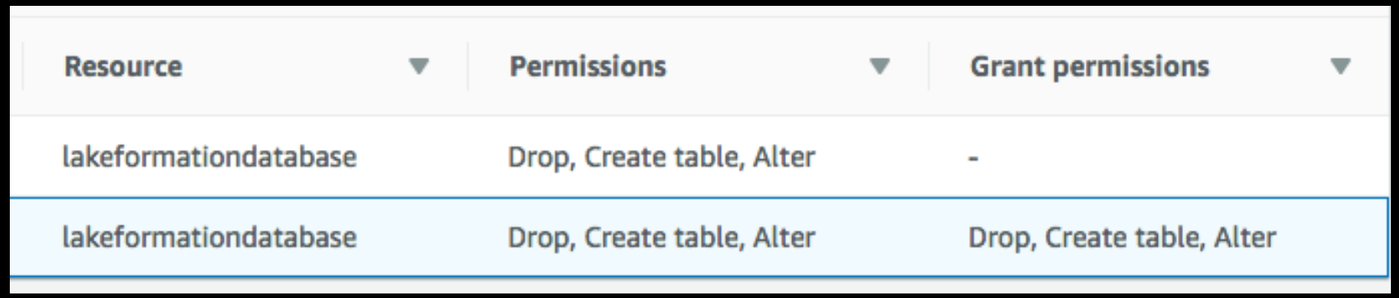
Control data access with simple grant and revoke permissions

Specify permissions on tables and columns rather than on buckets and objects

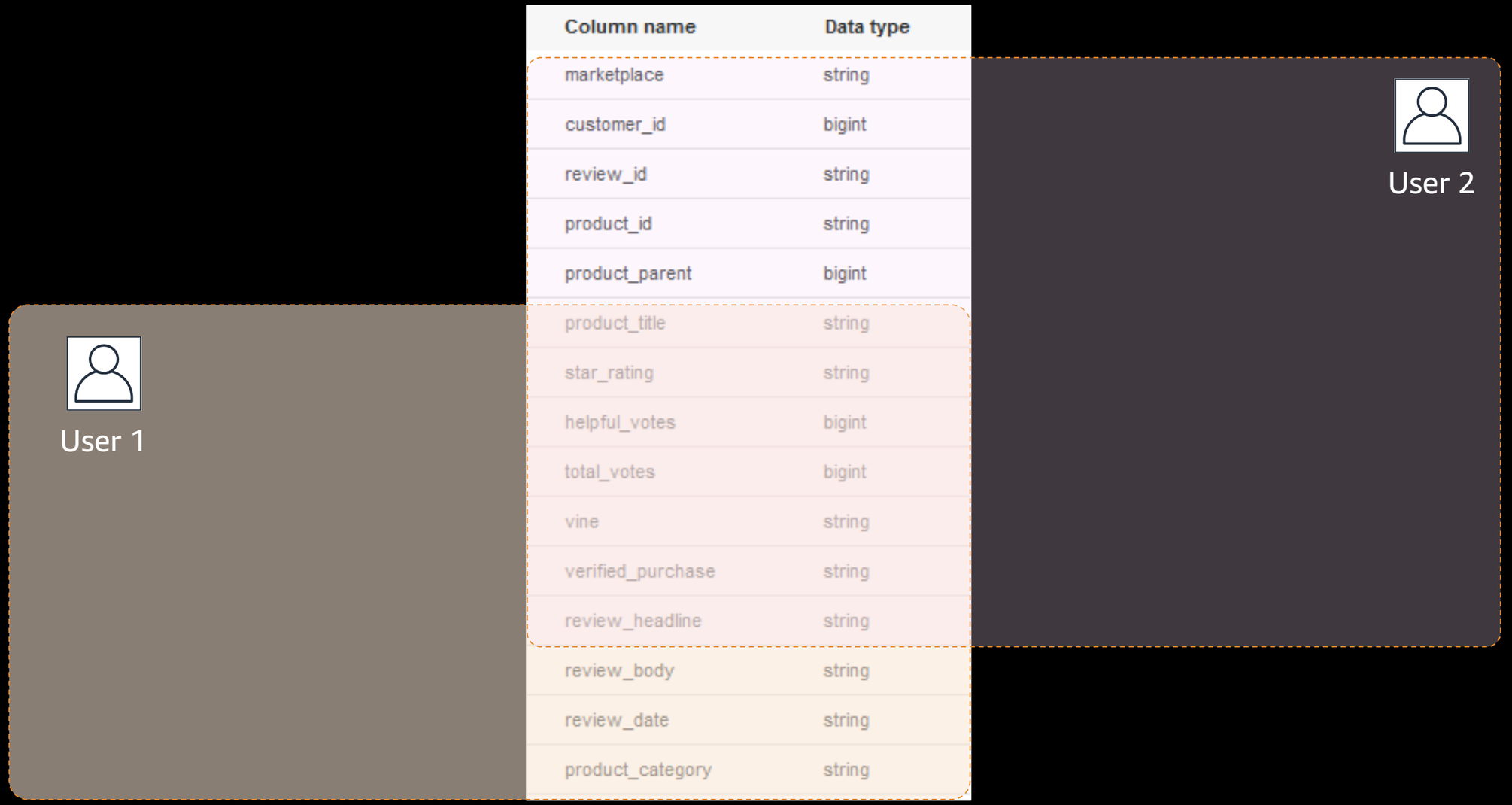
Easily view policies granted to a particular user

Audit all data access at one place

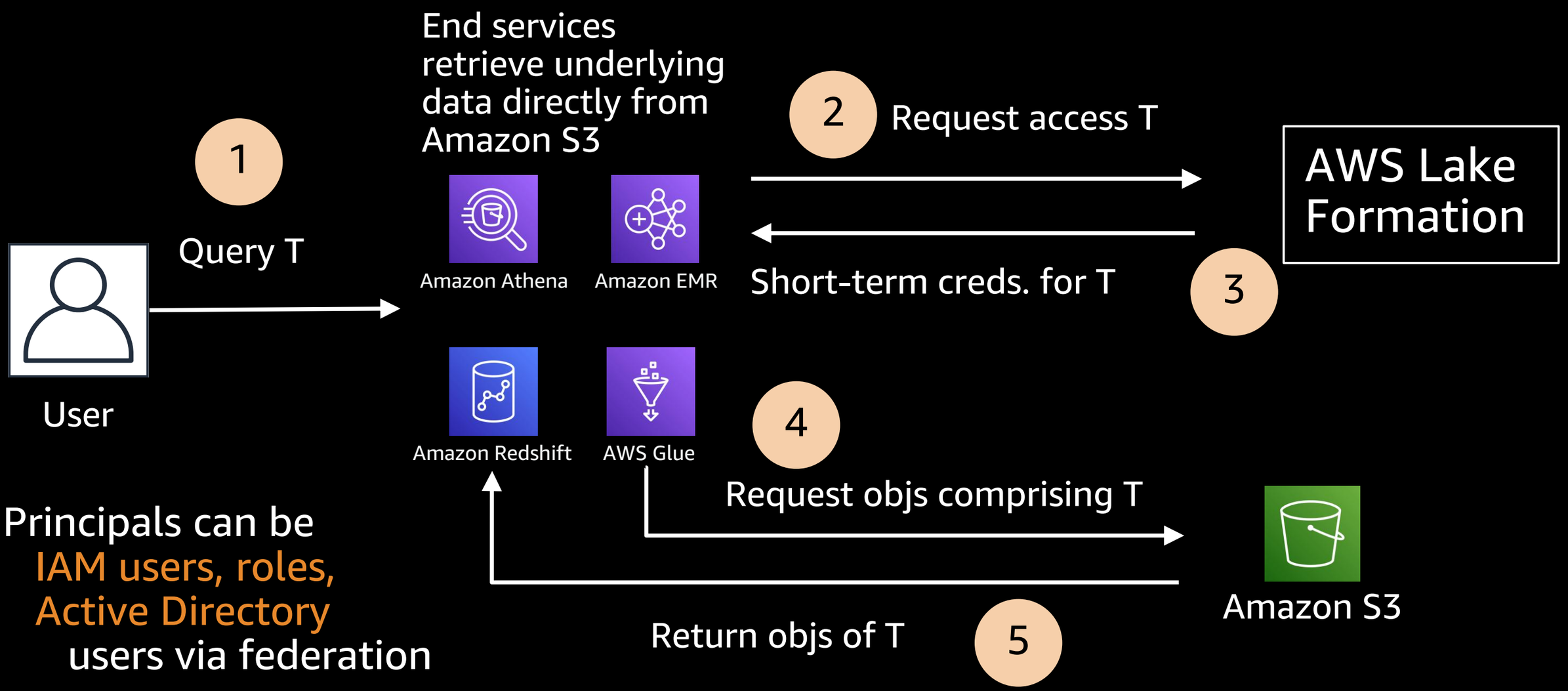




Grant table- and column-level permissions



Security – Deep dive



Principals can be
IAM users, roles,
Active Directory
users via federation

Data storage



Common concerns

Data should never be deleted

Ensure data integrity

Able to find and use the data

Maintain fine-grained permissions

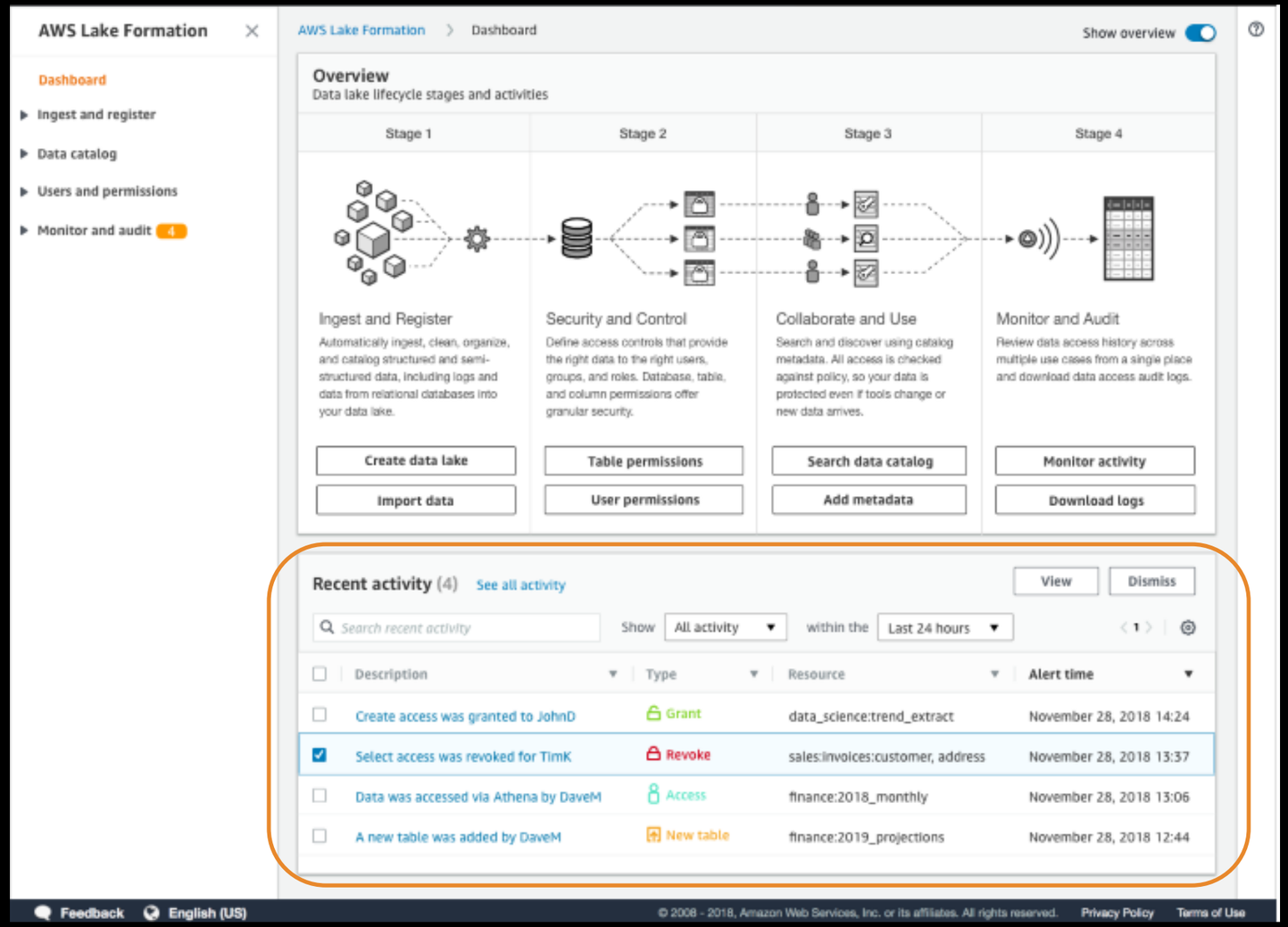
Audit use of the data

Audit and monitor in real time

See detailed alerts in the console

Download audit logs for further analytics

Data ingest and catalog notifications also published to CloudWatch events



Data should never be deleted

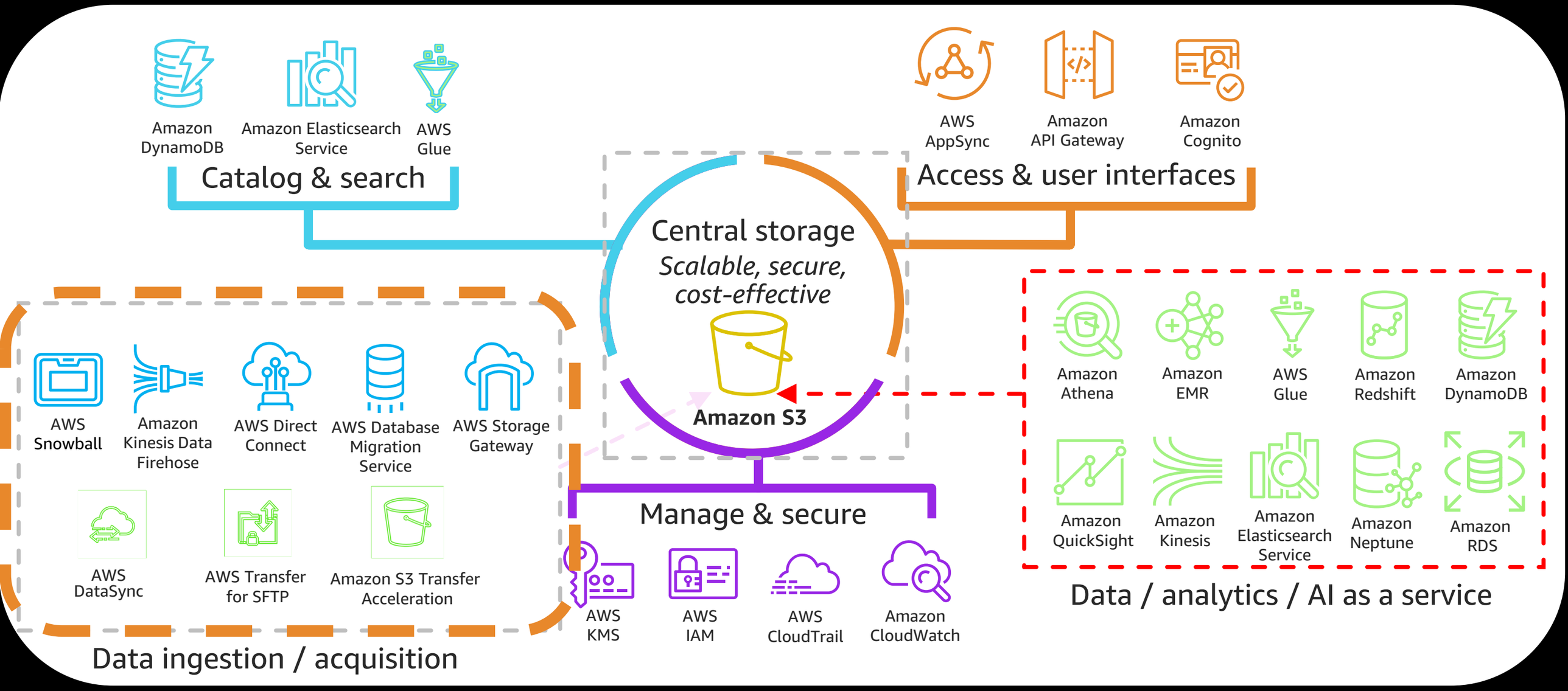
How does this impact data ingestion?



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

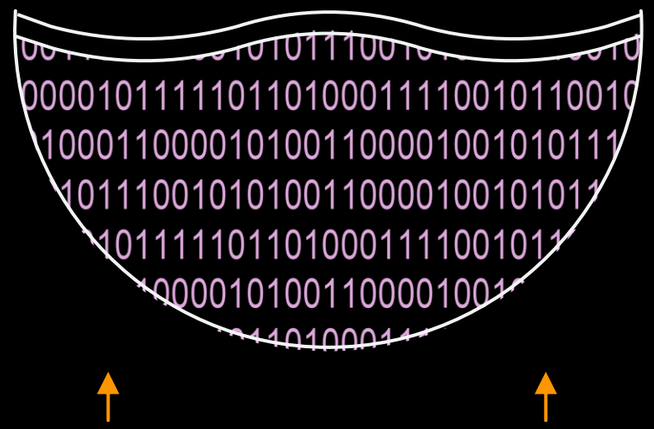
Layers of an analytics workload





Data ingestion / integration

MOST WAYS TO MOVE DATA



Data movement from your on-premises data centers

Data movement from real-time sources

Synchronizing data across environments

Data movement from on-premises data centers

Dedicated network connection

Secure appliances

Ruggedized shipping containers

Database migration

Gateway that lets applications write to the cloud

Data movement from real-time sources

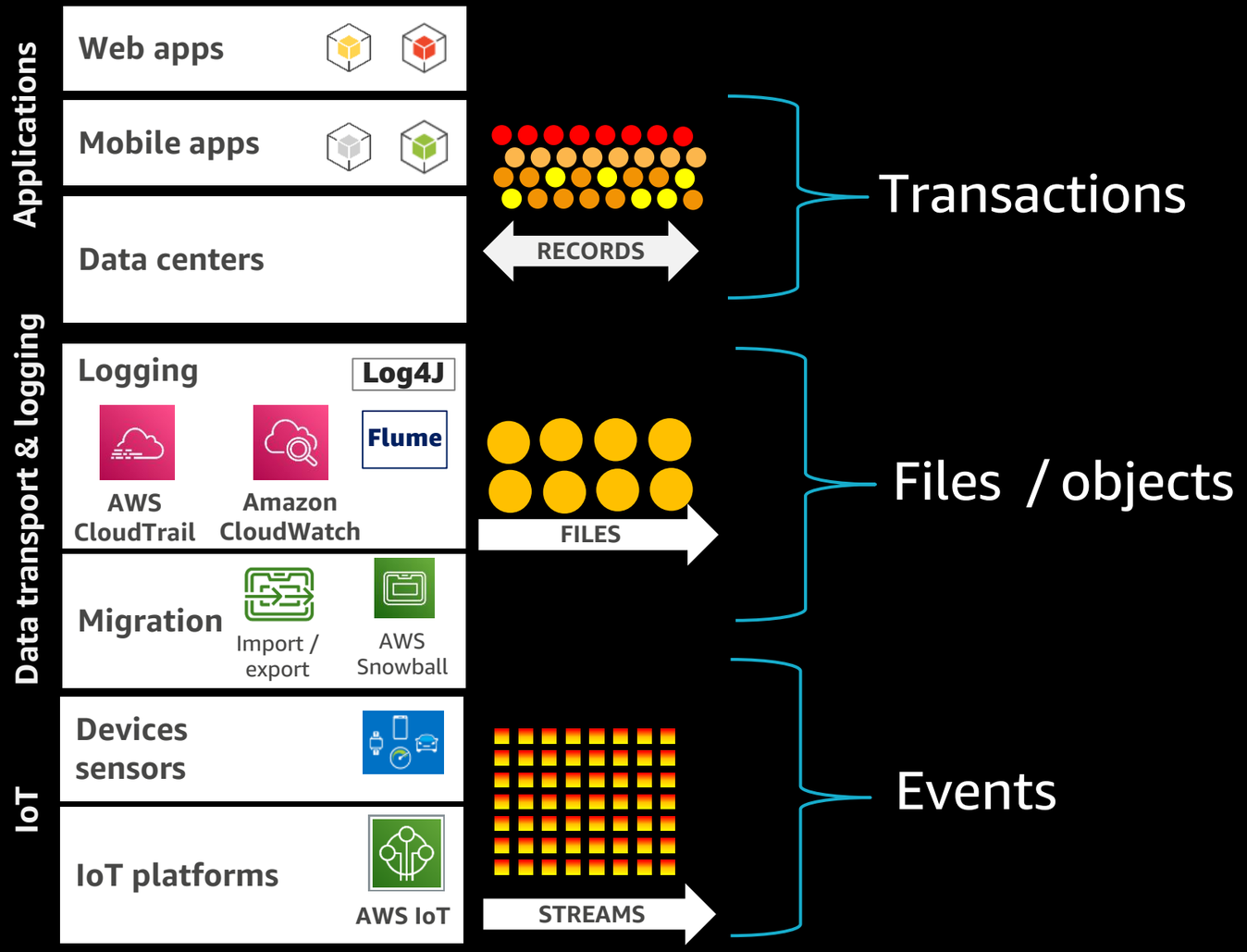
Connect devices to AWS

Real-time data streams

Real-time video streams

Data ingestion / integration

INGEST DATA FROM ANY SOURCE



Data durability built into each service

AWS Snowball

Amazon Kinesis Data Firehose

AWS Direct Connect

AWS Database Migration Service

AWS Storage Gateway

AWS DataSync

AWS Transfer for SFTP

Amazon S3 Transfer Acceleration

...

Data ingestion / acquisition

Other common questions



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

Common security questions / considerations

Compute and network isolation

- Can all traffic transfer over private and secure network links?
- Can we block ingress and egress internet access?
- What is the tenancy model, and are segregated user spaces available?

Authentication

- Is all access compliant with our authentication standards?
- Can all user interfaces integrate with our active directory?

Authorization

- How do we authorize access to all resources?
- Can we limit access to data, code, analytics, and training resources by role and job function?

Data protection

- Can we encrypt all data in transit and at rest?
- Does the service support encryption with AWS KMS customer managed keys (CMKs)?

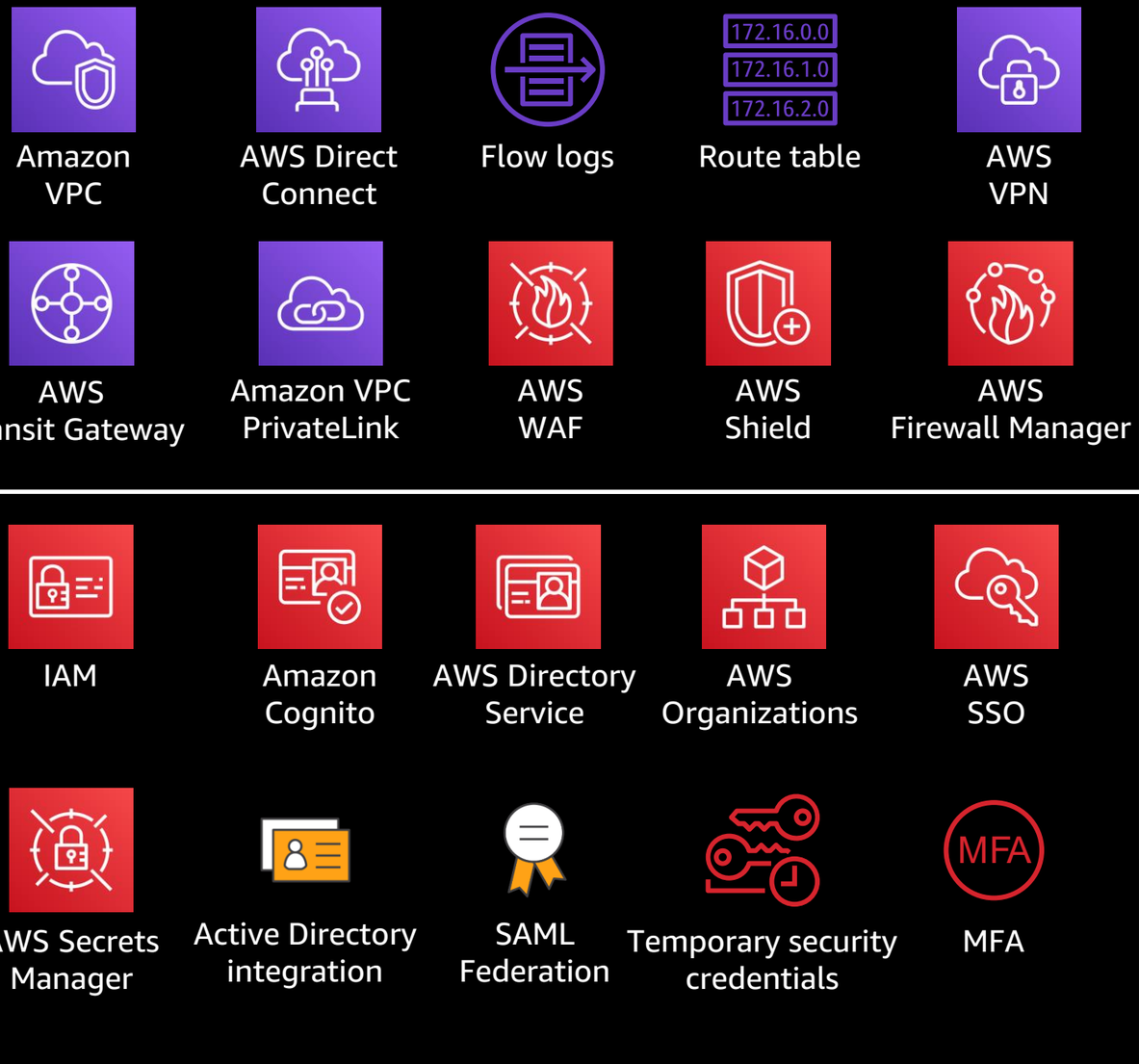
Configuration management

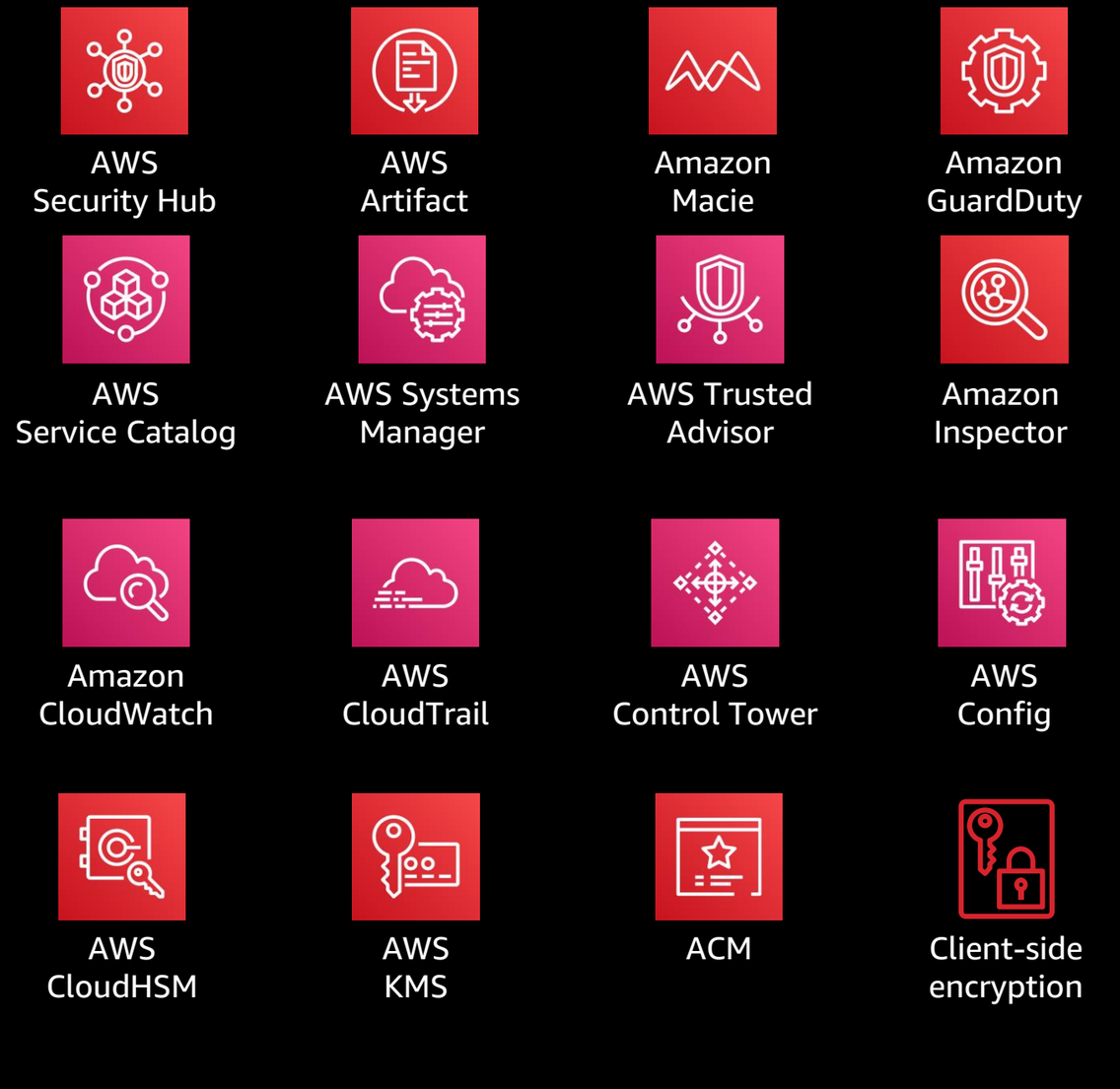
- How do we safely bring in public and private libraries and frameworks?
- How do we securely persist and protect code and model artifacts?

Auditability

- Does the service provide end-to-end auditability?
- Can audit trails be captured at user and file / object level?

Security services integrated with analytics





Let’s look at one more example



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

Common security questions / considerations

Compute and network isolation

- Can all traffic transfer over private and secure network links?
- Can we block ingress and egress Internet access?
- What is the tenancy model, and are segregated user spaces available?

Authentication

- Is all access compliant with our authentication standards?
- Can all user interfaces integrate with our active directory?

Authorization

- How do we authorize access to all resources?
- Can we limit access to data, code, analytics, and training resources by role and job function?

Data protection

- Can we encrypt all data in transit and at rest?
- Does the service support encryption with AWS KMS customer managed keys (CMKs)?

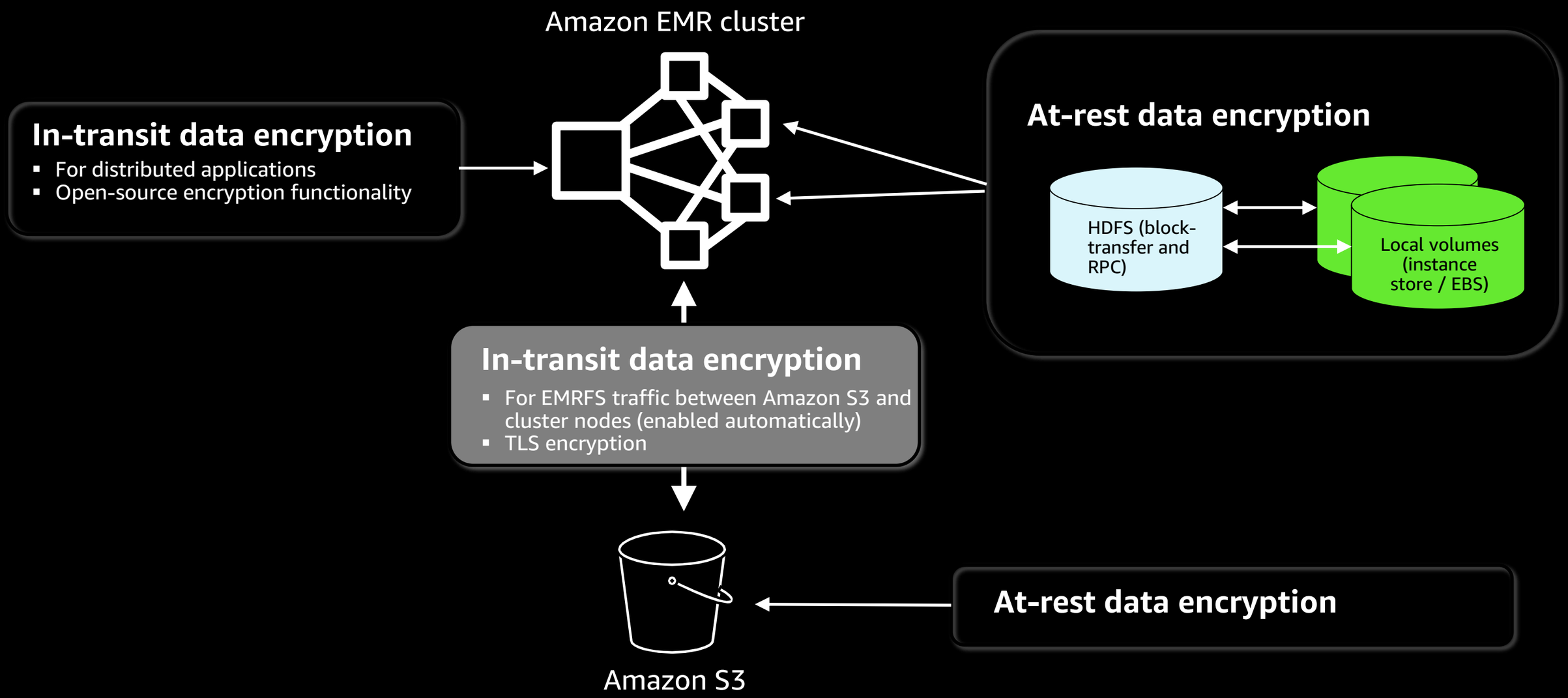
Configuration management

- How do we safely bring in public and private libraries and frameworks?
- How do we securely persist and protect code and model artifacts?

Auditability

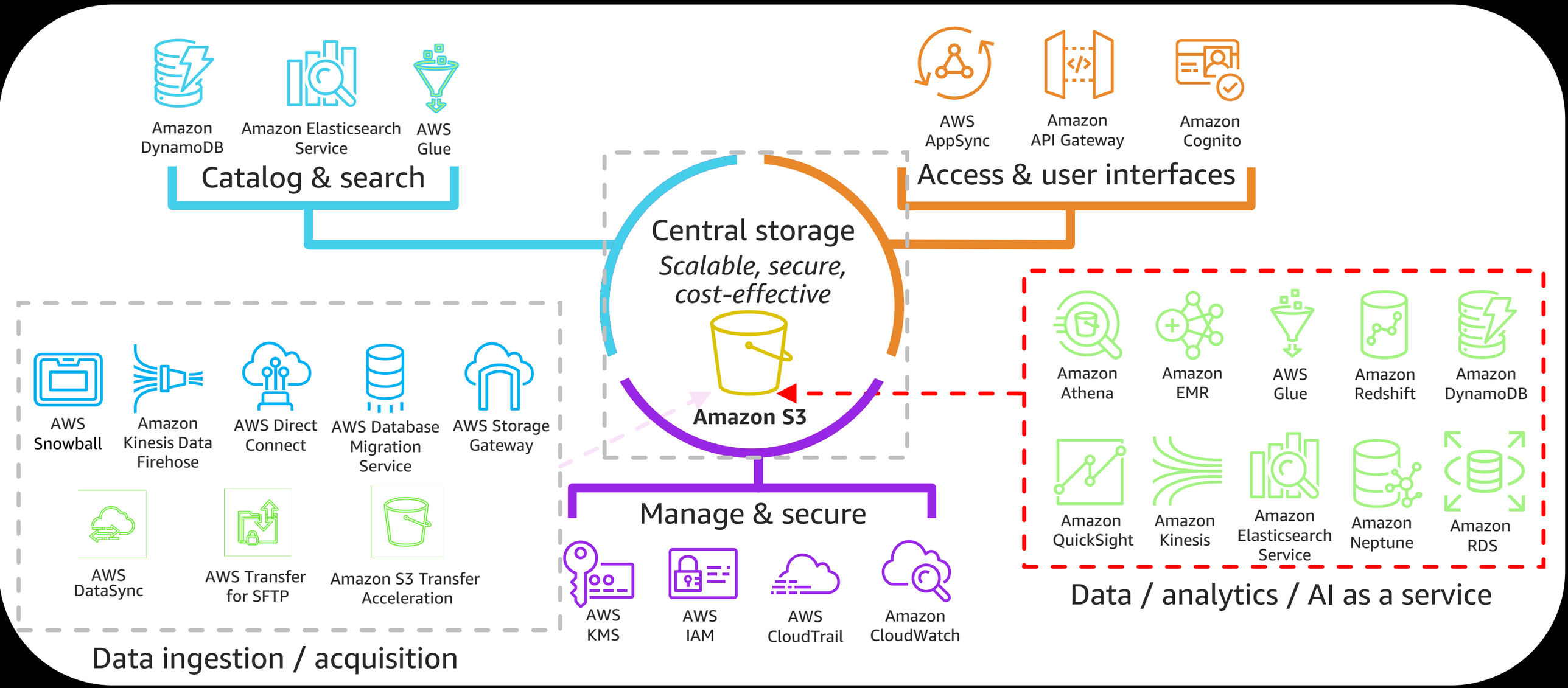
- Does the service provide end-to-end auditability?
- Can audit trails be captured at user and file / object level?

Data protection – Encryption in transit and at rest in Amazon EMR



Building a secure data solution

1,800+ SECURITY CONTROLS TO SECURE YOUR LAKE



Thank you!

Ben Snively

Specialist Solutions Architect

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.