
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AMZ201

Integrating Amazon's Andes data lake with AWS Lake Formation

Pradeep Misra (he/him)

Sr. Analytics Solutions Architect
Amazon

Albert Quiroga Bertolin (he/him)

Sr. Solutions Architect
Amazon

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Learning objectives

Understand the complexities of fine-grained access controls

Understand AWS Lake Formation

Understand how Lake Formation simplifies permissions

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Introduction: Amazon Andes data lake

Challenges with Andes data lake

AWS Lake Formation

Integrating Andes data lake with AWS Lake Formation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Put data to work



Make better decisions



Improve efficiencies



Respond faster

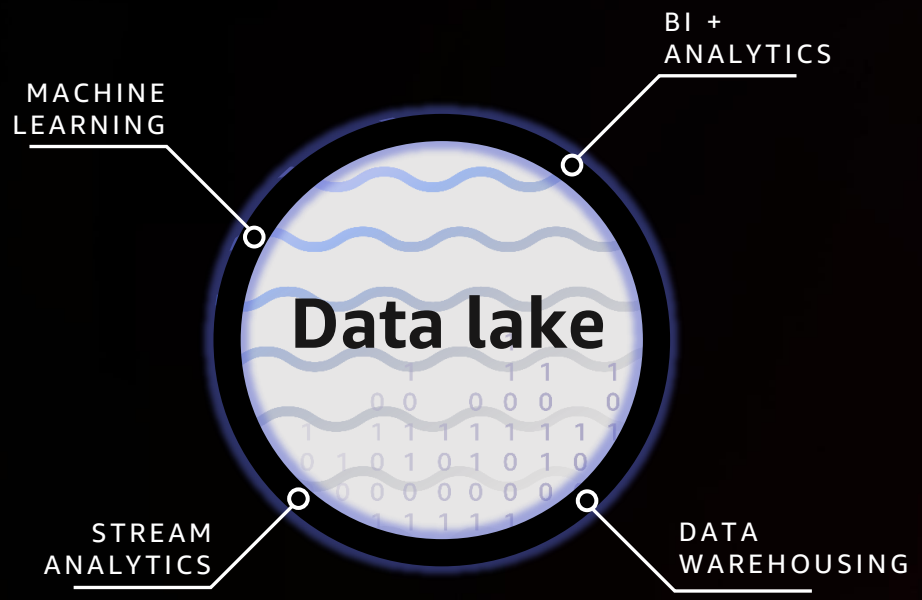


Uncover opportunities

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

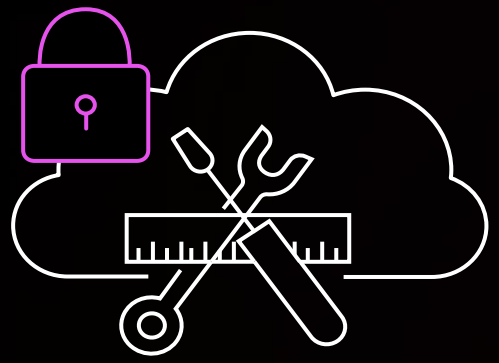
Data lake: Scalable and open



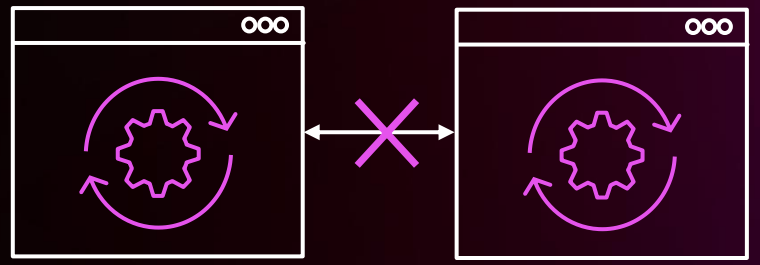
A **flexible**, **secure** repository that enables you to **govern**, **discover**, **share**, and **analyze structured** and **unstructured** data at any scale with the **tool of your choice**

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common operational challenges



Security and governance is hard



Complex self-service analytics, data sharing, external integrations

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Andes Data Lake



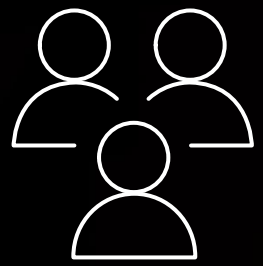
Andes

- Centralized Data Catalog for all of Amazon
- Support Amazon large-scale analytics and ML use cases
- Exabyte-scale data volume
- Searchable
- Strong permissions control

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Andes

USERS CAN PRODUCE DATA IN ANDES, WHICH CAN BE CONSUMED BY ANY NUMBER OF CONSUMERS

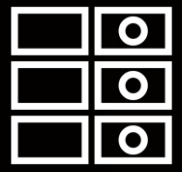


Data producers

Data lake and marketplace



Data consumers



Data input



Andes



Compute

aws

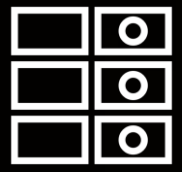
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Andes

amazon

Finance

Data producers ⟷ **Data lake and marketplace** ⟷ **Data consumers**



Data input



Andes

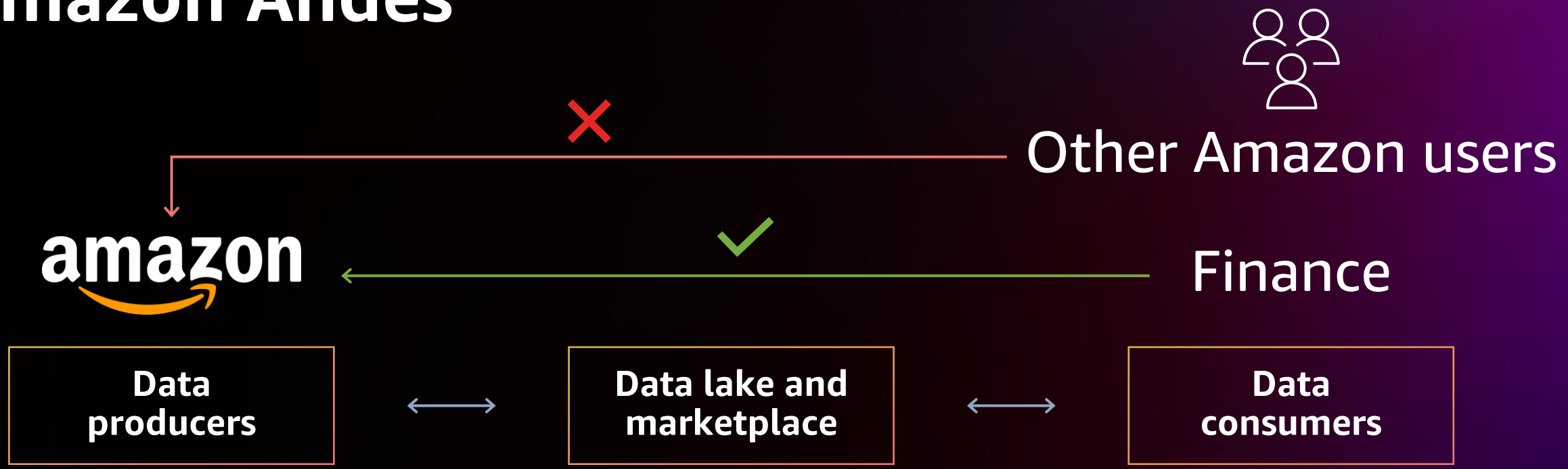


Compute

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Andes



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Principle of least privilege

DOES EVERYONE REALLY NEED ACCESS TO EVERY COLUMN AND EVERY ROW?

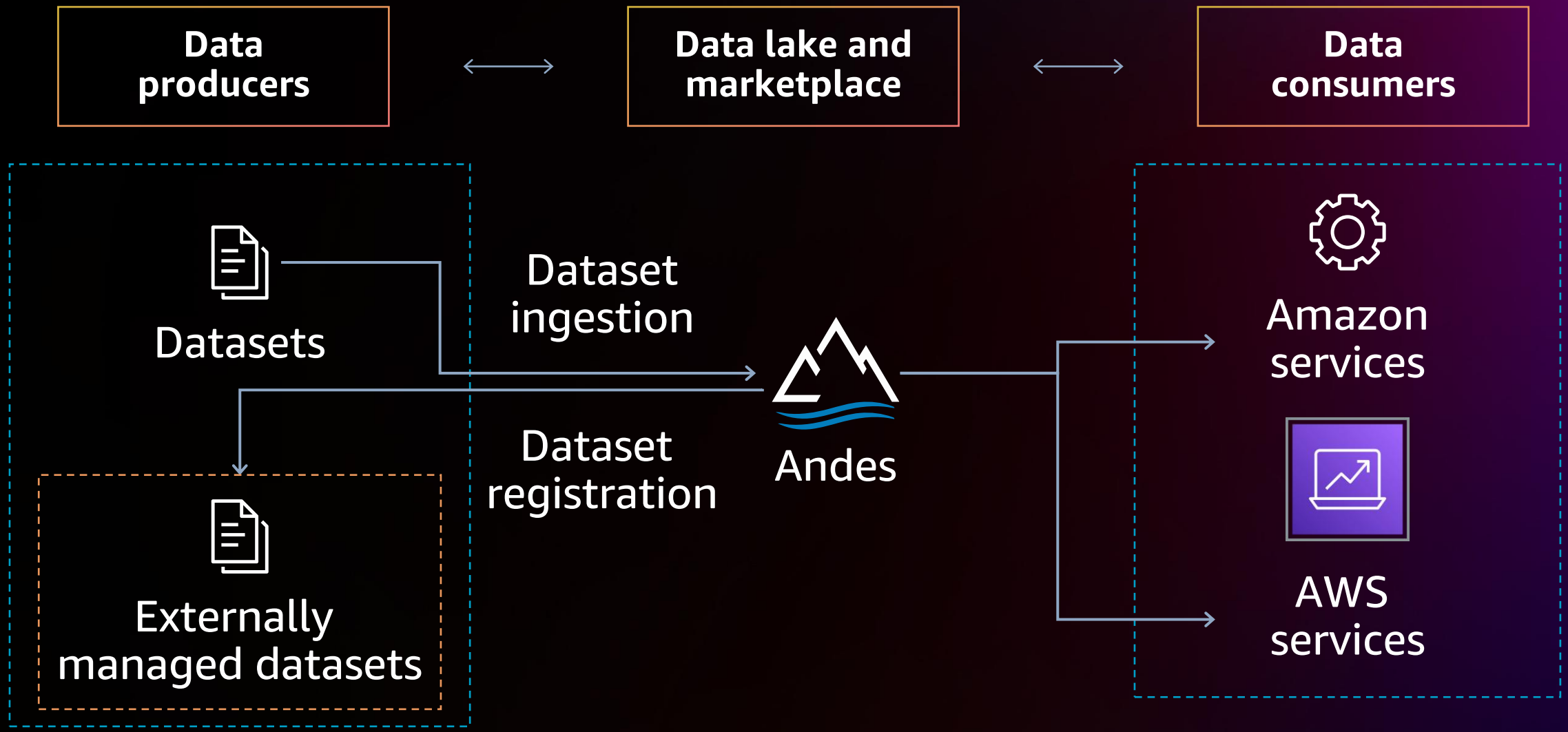
DATACENTRAL | Hoot Applications Jobs Datasets Data Stores Preferences Feedback

legal_entity_id	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Data Warehouse construct used mostly for partitioning, roughly related to region...read more
marketplace_id	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Unique identifier for a marketplace. A replacement for OrgUnit.See https://w.ama...read more
merchant_customer_id	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Unique Identifier for a merchant in Amazon platform.
order_id	True	String	MAX_LENGTH: 19	Identifier assigned to each customer's order to uniquely identify it.
order_type	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Classifies the type of order https://w.amazon.com/?OrderTypes
promised_arrival_day	True	Timestamp	TIMESTAMP_MASK: yyyy-MM-dd'T'HH:mm:ss	The arrival Day we promised to the customer.
promised_ship_day	True	Timestamp	TIMESTAMP_MASK: yyyy-MM-dd'T'HH:mm:ss	The Ship Day we promised.
tender_day	True	Timestamp	TIMESTAMP_MASK: yyyy-MM-dd'T'HH:mm:ss	The day the Shipment packages were ready. Referenced from O_COMPLETED_SHIPMENT_P...read more
seller_of_record_vendor_id	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Vendor that this shipment is tied to for financial reporting purposes. https://w...read more
ship_flags	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Completed, Whenever, Unknown, etc.
ship_method	True	String	MAX_LENGTH: 90	Air, Ground, 2day etc. A SHIP_METHOD of -1 indicates that the shipment was fulfi...read more
shipment_ship_option	True	String	MAX_LENGTH: 90	Ship option (eg. Prime / 2 day / standard /etc.) .
customer_ship_option	True	String	MAX_LENGTH: 90	4 options are available: Holiday Premium Standard SSS: slow ship
shipment_id	True	Int	MIN_VALUE: -1×10^{38} MAX_VALUE: 1×10^{38}	Unique identifier for the shipment.

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

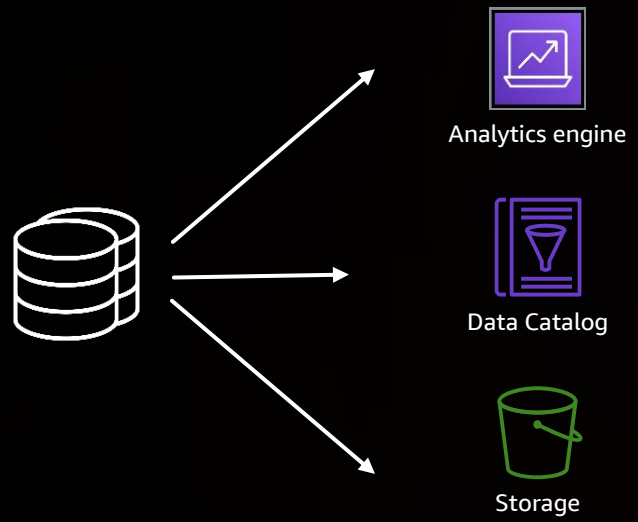
Supporting all use cases



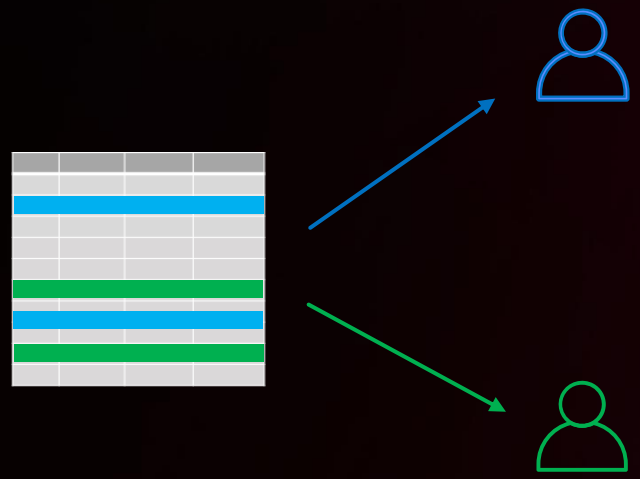
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why managing data lake permissions is hard

Unifying permissions across the data lake stack



Enforcing fine-grained permissions to restrict access



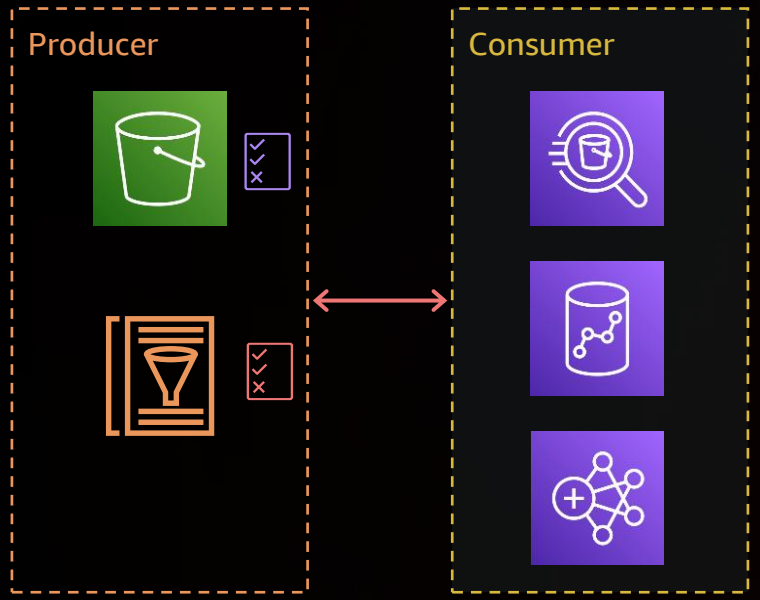
Ensuring data access complies with regulations



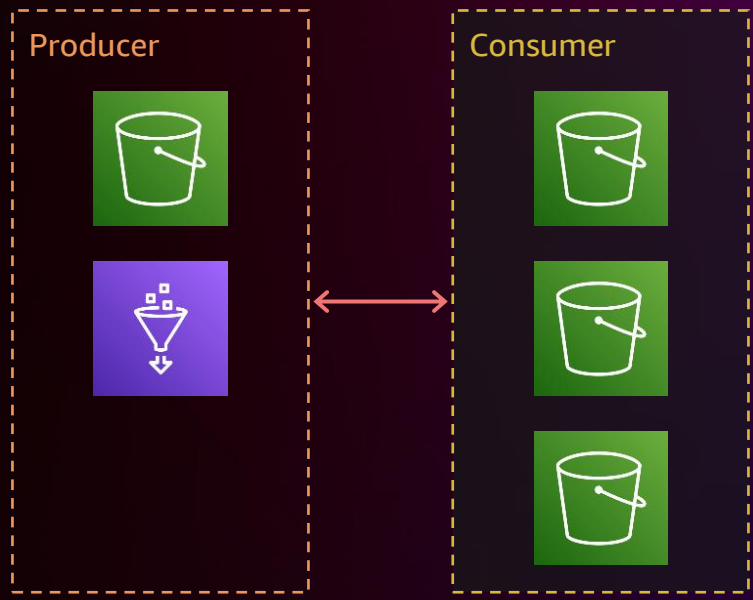
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why is sharing data across accounts hard?

To share data …



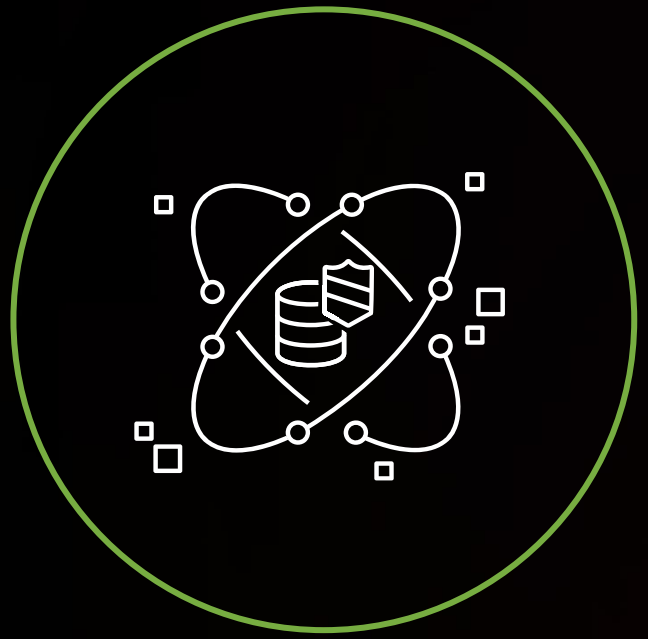
Duplicating data



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



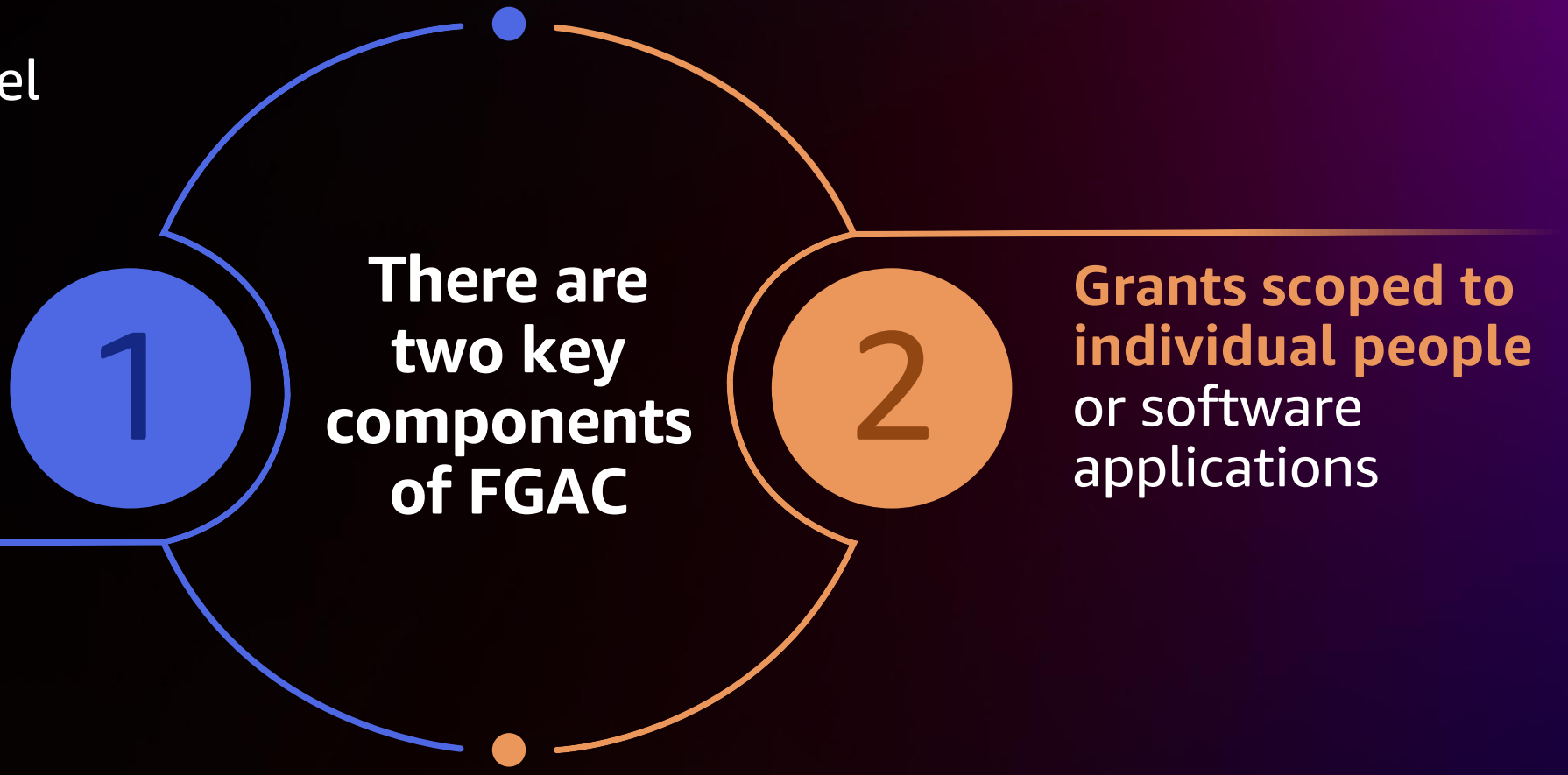
Unified data access, security, and governance

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is fine-grained access control anyway?

Row and column level permissions grants

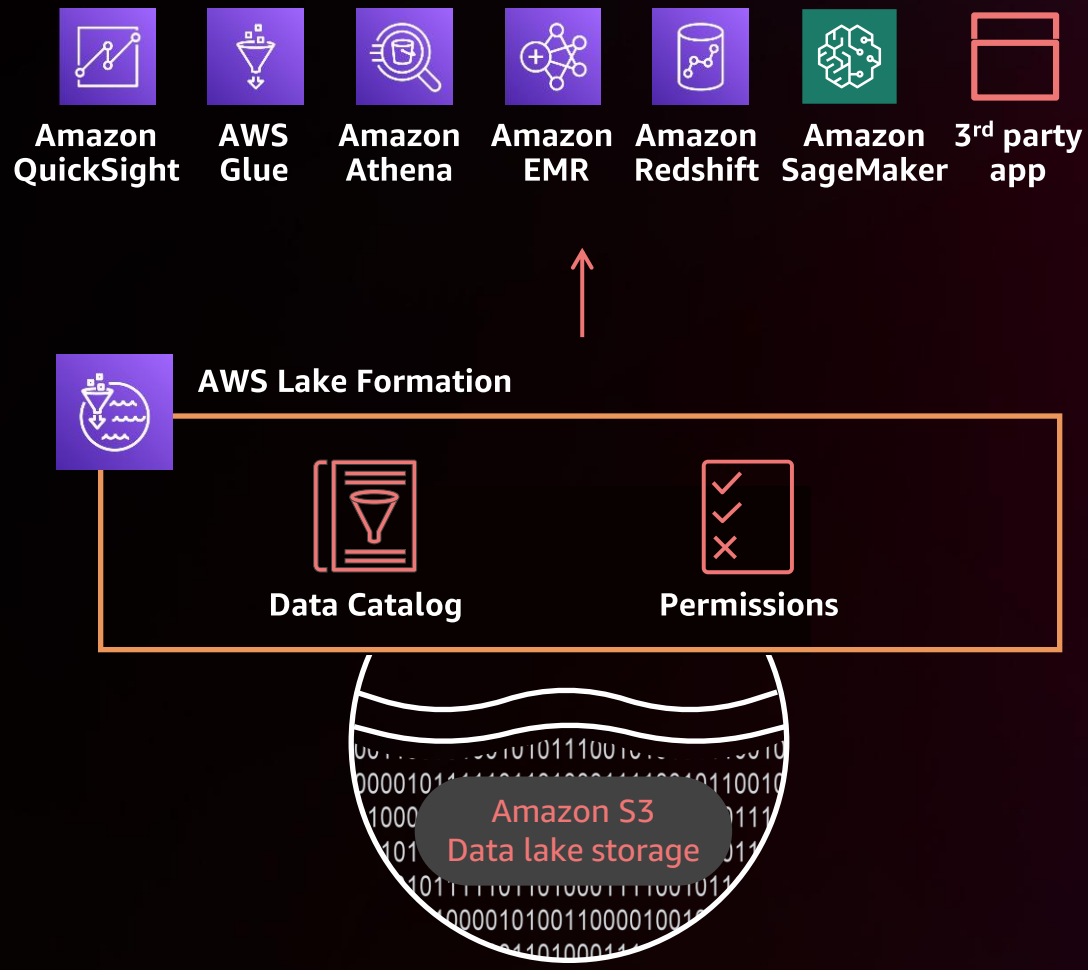
We need to be able to **grant access at the individual column or row level**



Grants scoped to individual people or software applications

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

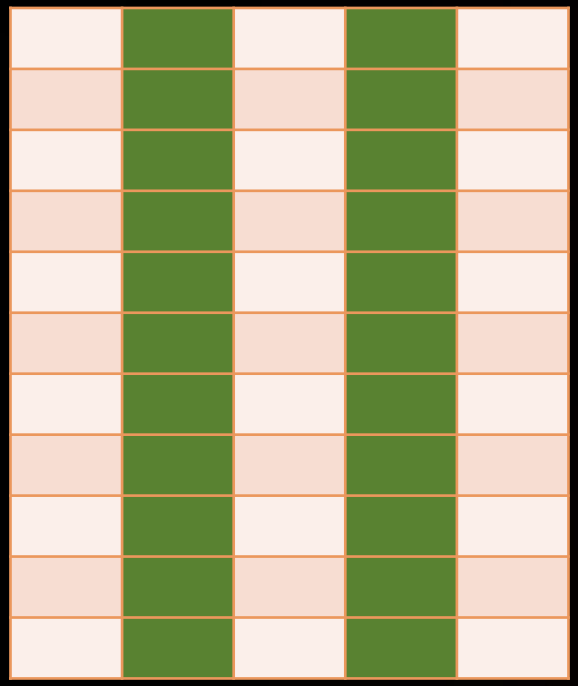
AWS Lake Formation permissions model



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

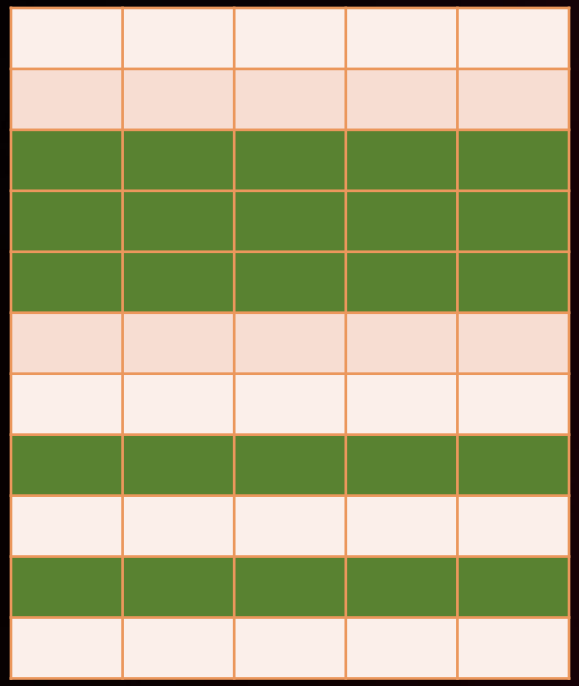
AWS Lake Formation permissions on …

Columns



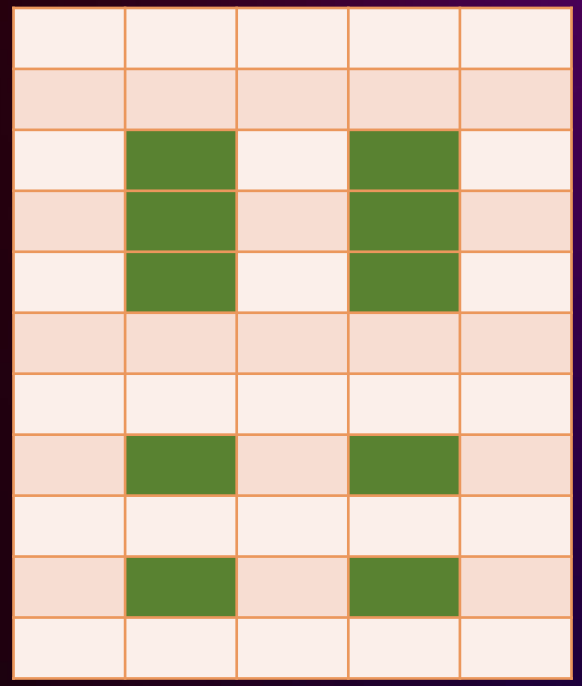
Specify include or exclude list

Rows



Specify row filter with PartiQL

Cells

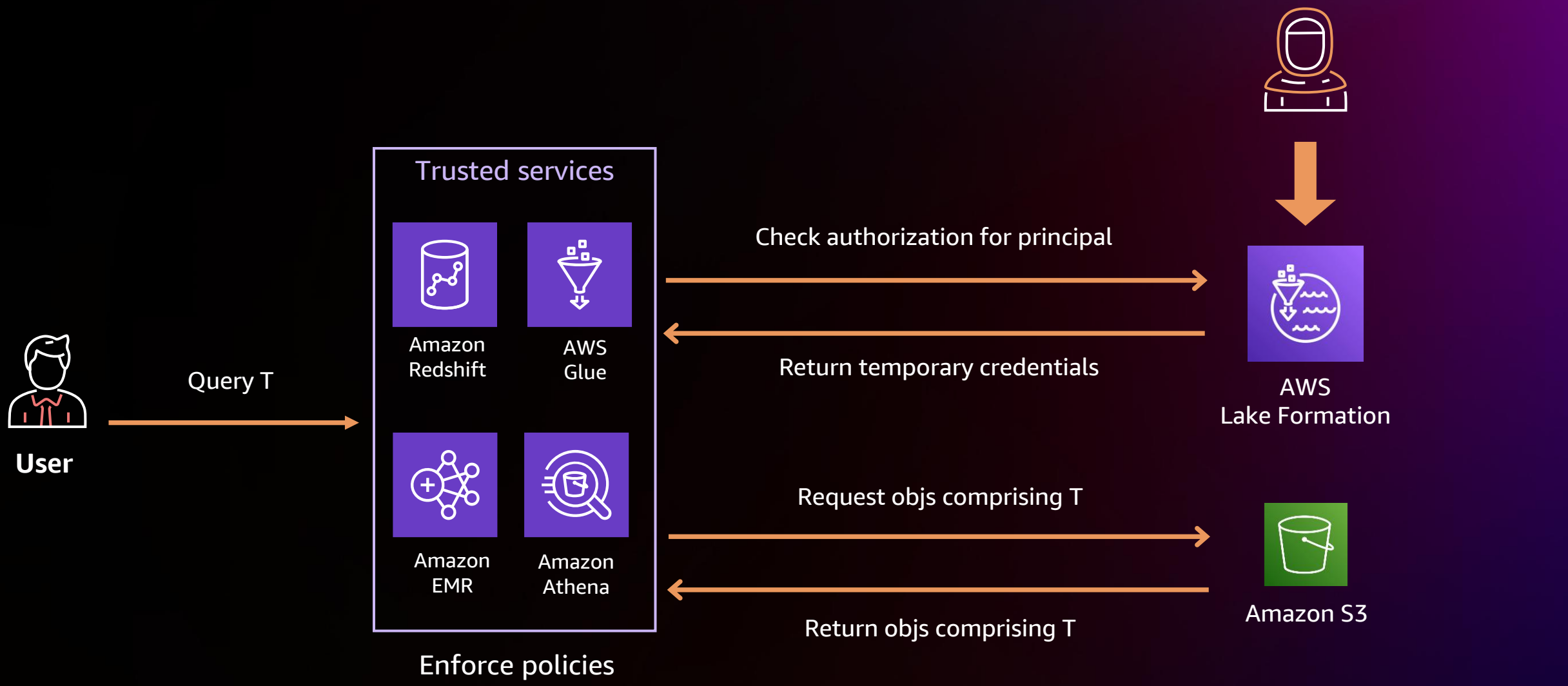


Combine column and row filters

aws

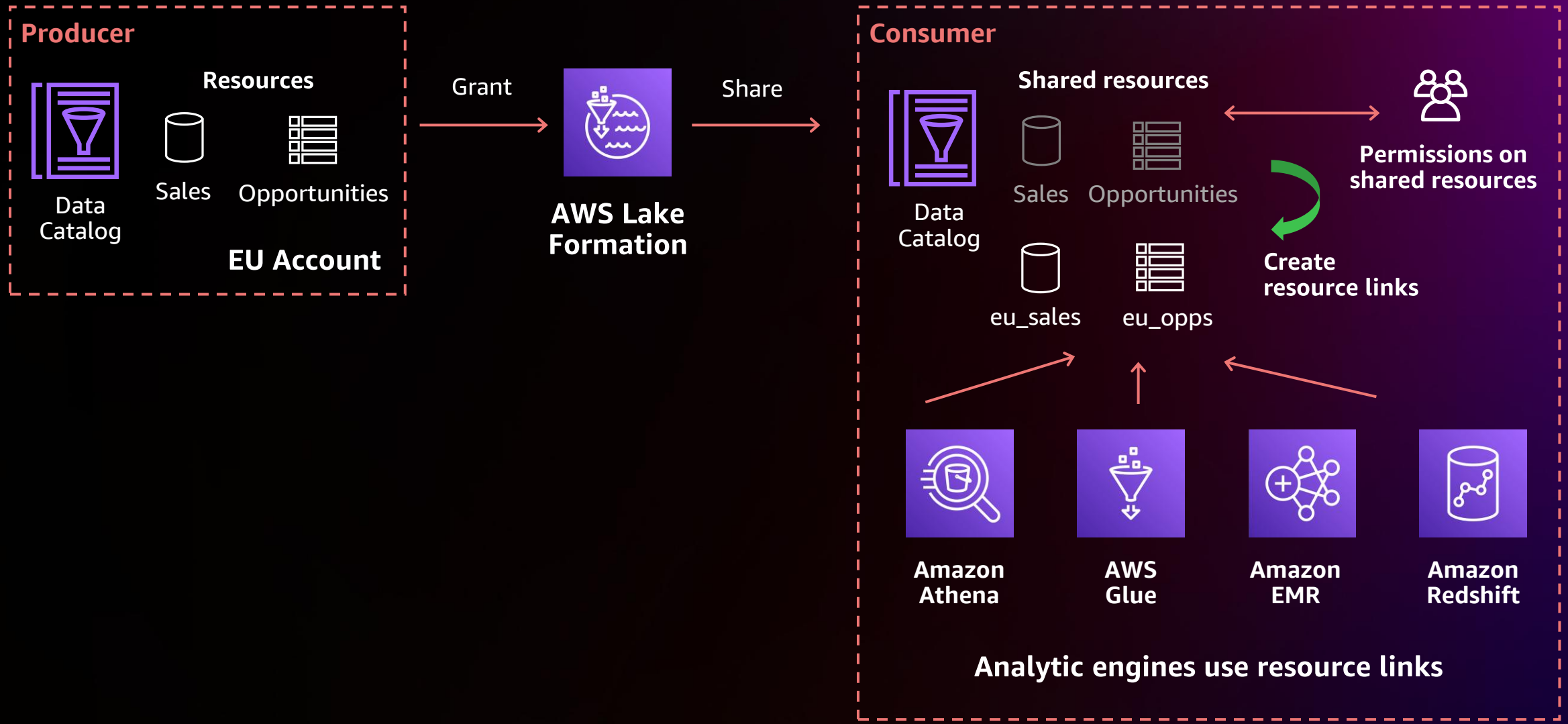
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

LF query flow



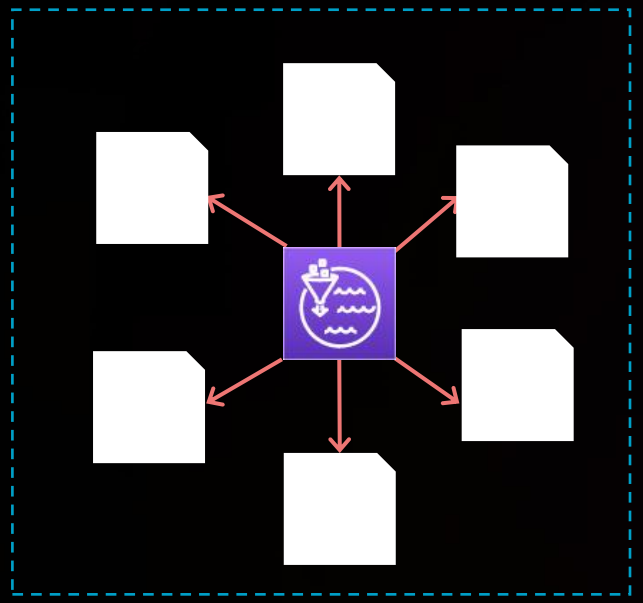
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation cross-account sharing

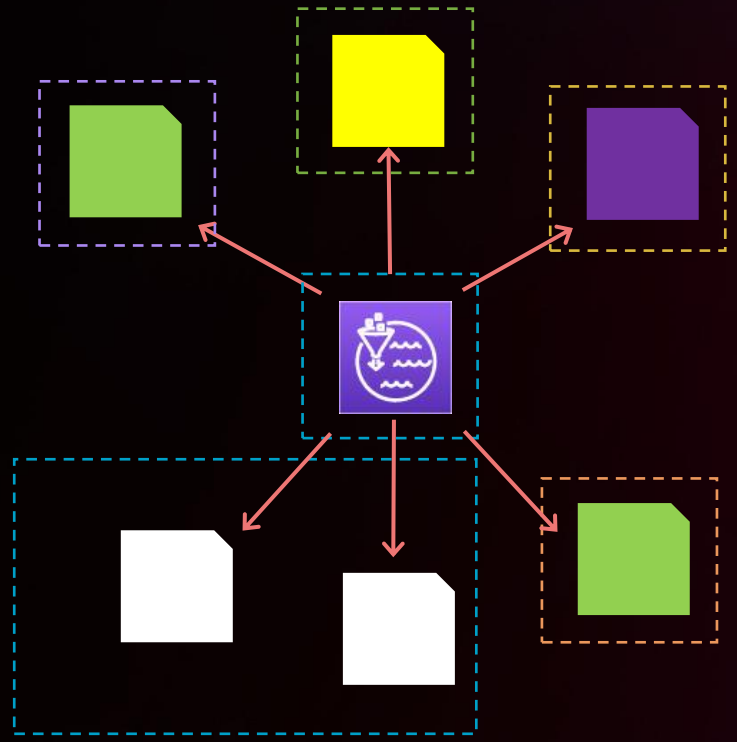


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

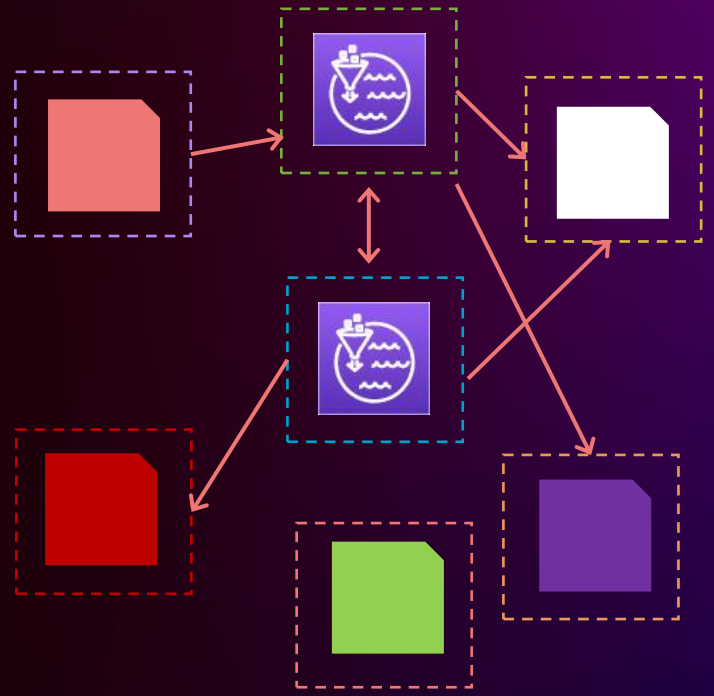
LF supports common data sharing topologies



Centralized, hub and spoke



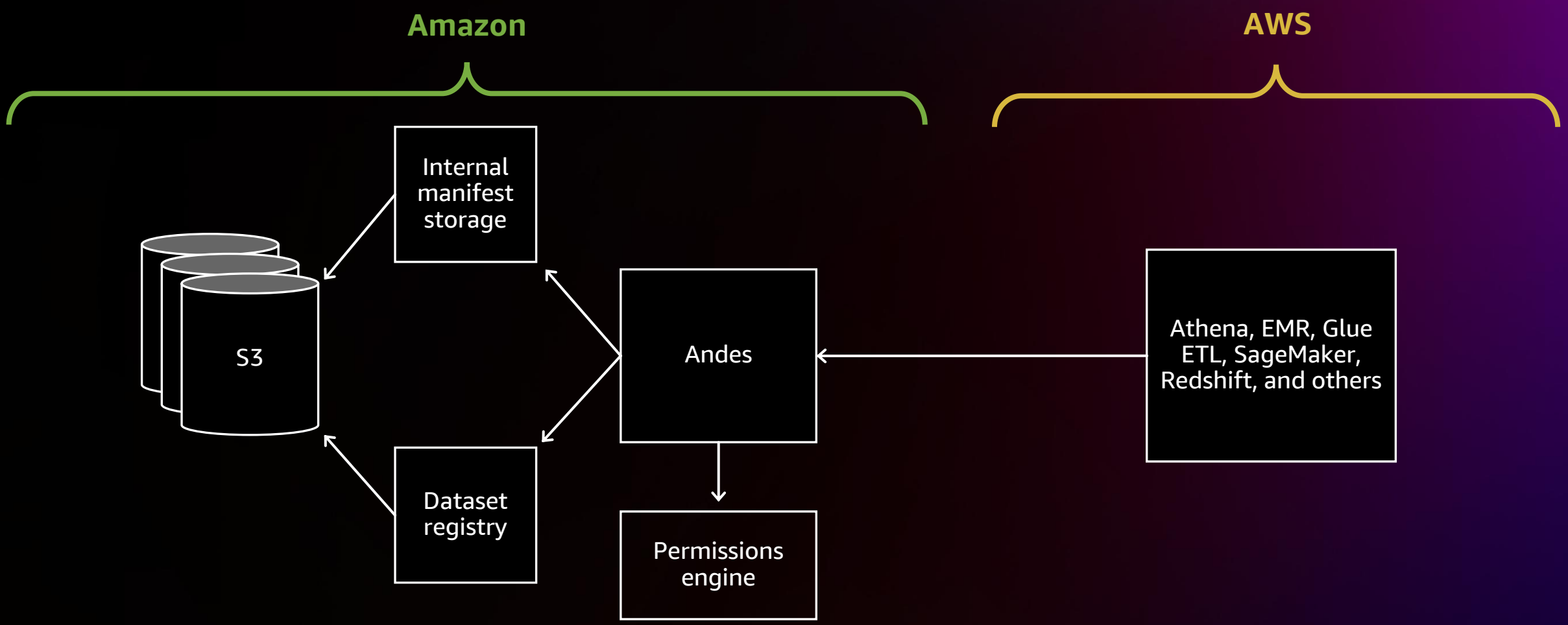
Across orgs and companies such as vendors, suppliers, aggregators, distributors



Data mesh

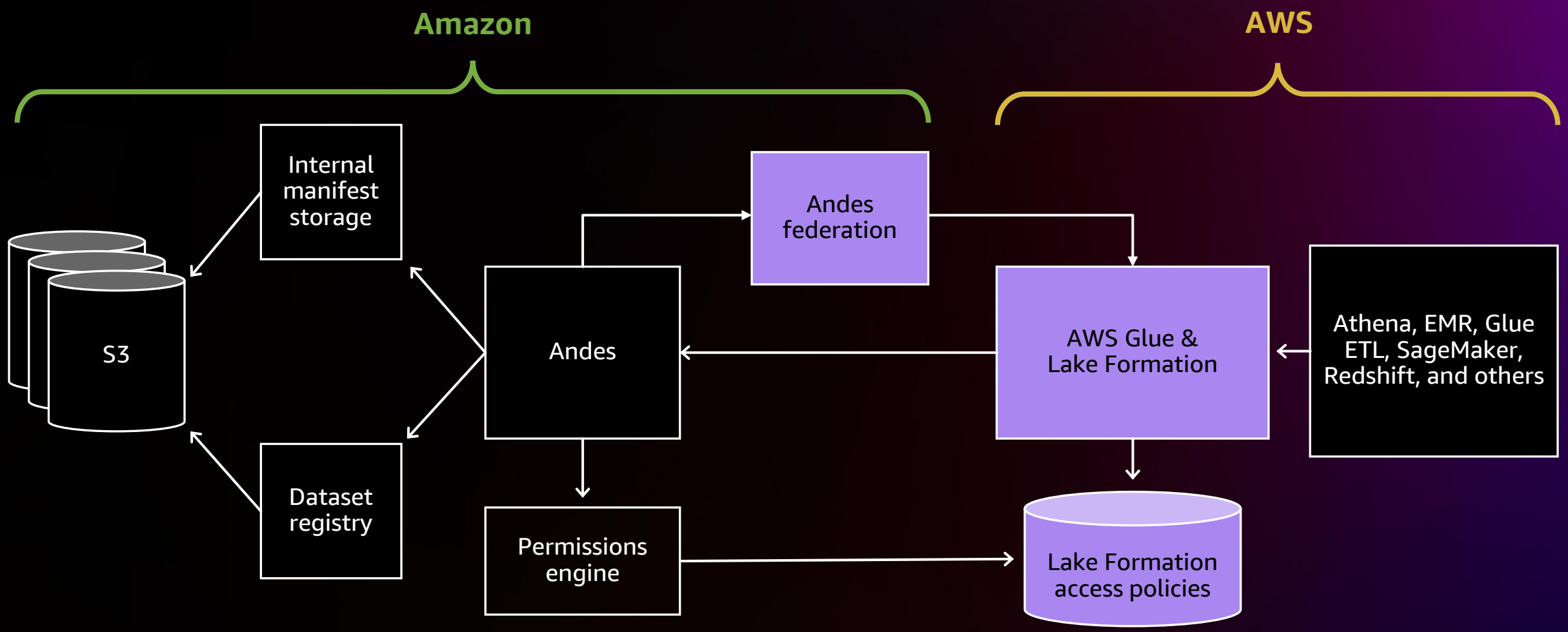
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Previous architecture



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

So how does it actually work?



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Now, let’s have a chalk talk!

- Permissions at scale
- Performance implications
- Integration with other systems
- Build data mesh architecture

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Pradeep Misra
pkmisra@amazon.com

Albert Quiroga Bertolin
bertolb@amazon.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Integrating Amazon Andes with AWS Lake Formation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using Amazon Andes data



Andes

Working with data at Amazon's scale almost always involves making trade-offs between completeness, precision, performance, and cost

JOINs tend to be slow and expensive, so we rely heavily on wide denormalized datasets

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using Amazon Andes data



Andes

with principle of least privilege

Fine-grained access controls also allow us to limit the breadth of fields exposed to a given consumer without sacrificing the performance and cost benefits of wide denormalized tables!

In addition, minimizing the number of rows or columns that a given query needs to process can even yield performance improvements!

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Andes



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Possible solution – encryption

Andes

consumers

id	Name	Surname	Orders	Credit Card #
1	Farzad	Khandan	3	
2	Vince	Marchillo	5	
3	Dhairya	Jindani	4	

Need to manage a key

I need to manage who can access the key

I need to make sure all reading clients support using a key

I need to manage performance to de/encrypt

Doesn’t scale for multiple row and column combinations

Operational overhead

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Possible solution – duplication

Andes

consumers

id	Name	Surname	Orders	Credit Card #
1	Arnav	Desai	3	111 …
2	Paulo	Santos	5	222 …
3	Saanvi	Sarkar	4	333 …



consumers_masked

id	Name	Surname	Orders
1	Arnav	Desai	3
2	Paulo	Santos	5
3	Saanvi	Sarkar	4

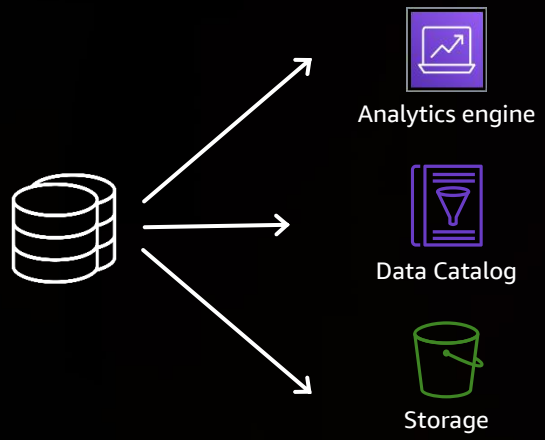
Data replication, doesn’t scale

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why managing data lake permissions is hard

Unifying permissions across the data lake stack

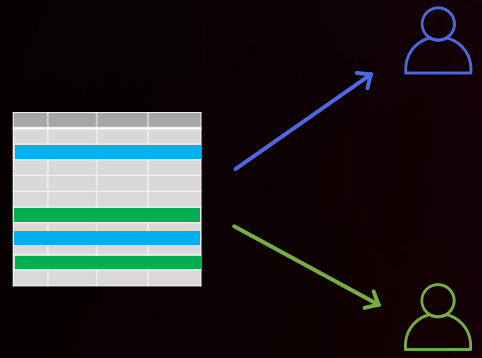


Split storage, metadata, and compute

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 access only portions

1,000s of resources and 10,000s of users

Ensuring data access complies with regulations



Democratize data access

Regulations and governance

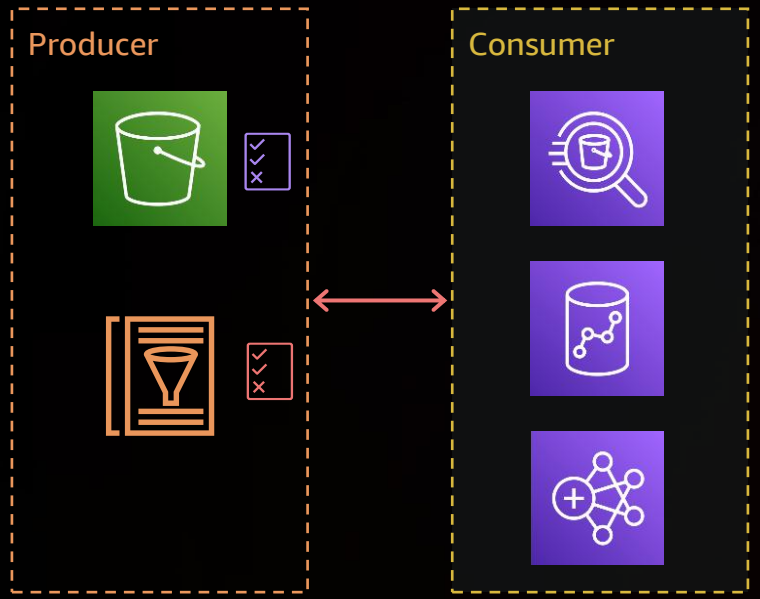
Monitor and audit data access

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

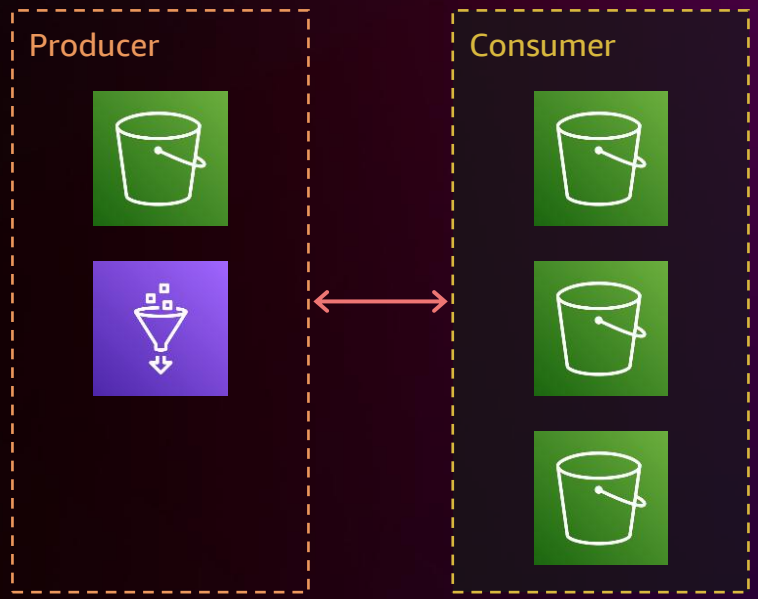
Why is sharing data across accounts hard?

To share data …



- S3 and IAM policies
- Limited by service support
- Lacks discoverability
- Policy size limits (coarse grained)

Duplicating data



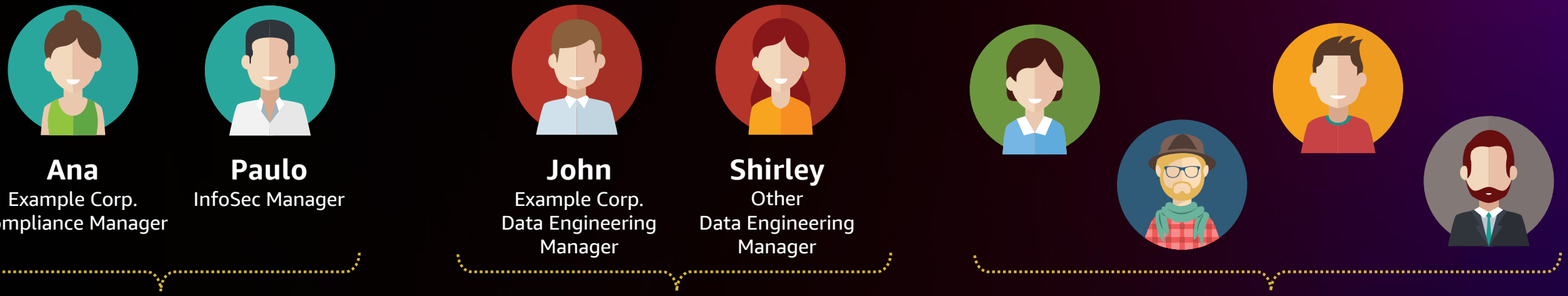
- ETL pipelines
- Multiple redacted copies
- Expensive, brittle, and error prone

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consider the following example

Example Corp. is a supply chain and logistics services provider for merchants who do not sell on our amazon.com platform

The Example Corp. team needs to protect their customer data in any SDO datasets



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data stewards and dataset owners together define the access control for a given dataset

1



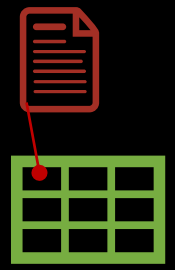
Data stewards create tag(s) to represent the data that they are responsible to protect and define access control (data access policy) against the tag



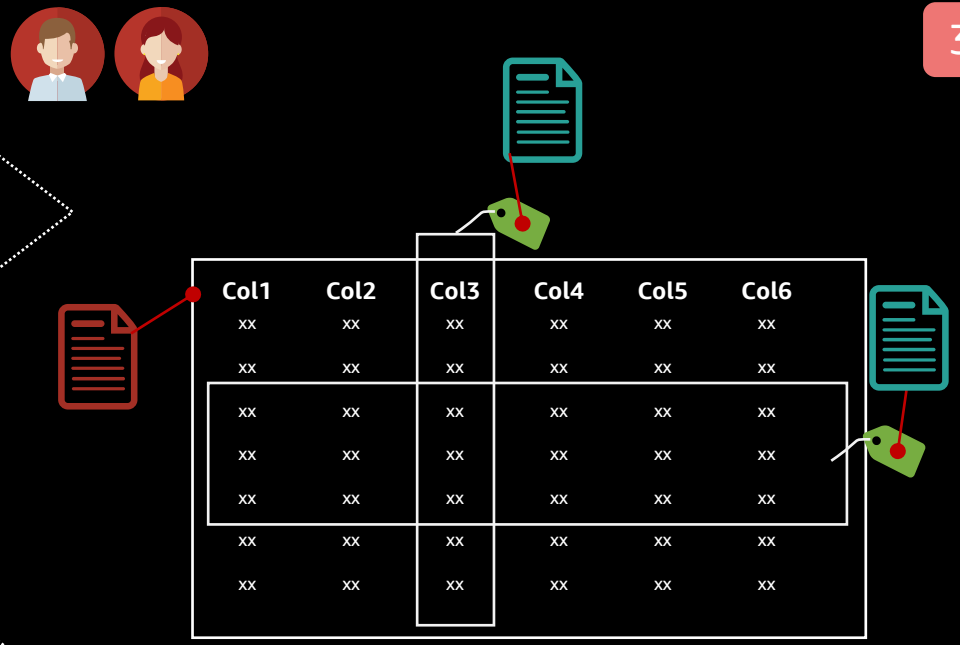
2



Dataset owners define the access control for any data in their dataset



3



Dataset owners classify the columns and rows in the dataset by associating appropriate tags

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data consumers set up endpoint to consume Andes dataset and get permissions to read the dataset

4

Data consumers select a dataset and submit access request for specific identity and data scope

5

Dataset owner as well as the **approvers** review the request and decide to either approve or deny the request, each independently

6

Data consumers set up their endpoint tools (such as Athena) to access Andes dataset by configuring corresponding resources (such as Glue catalog)

7


Redshift


QuickSight


EMR

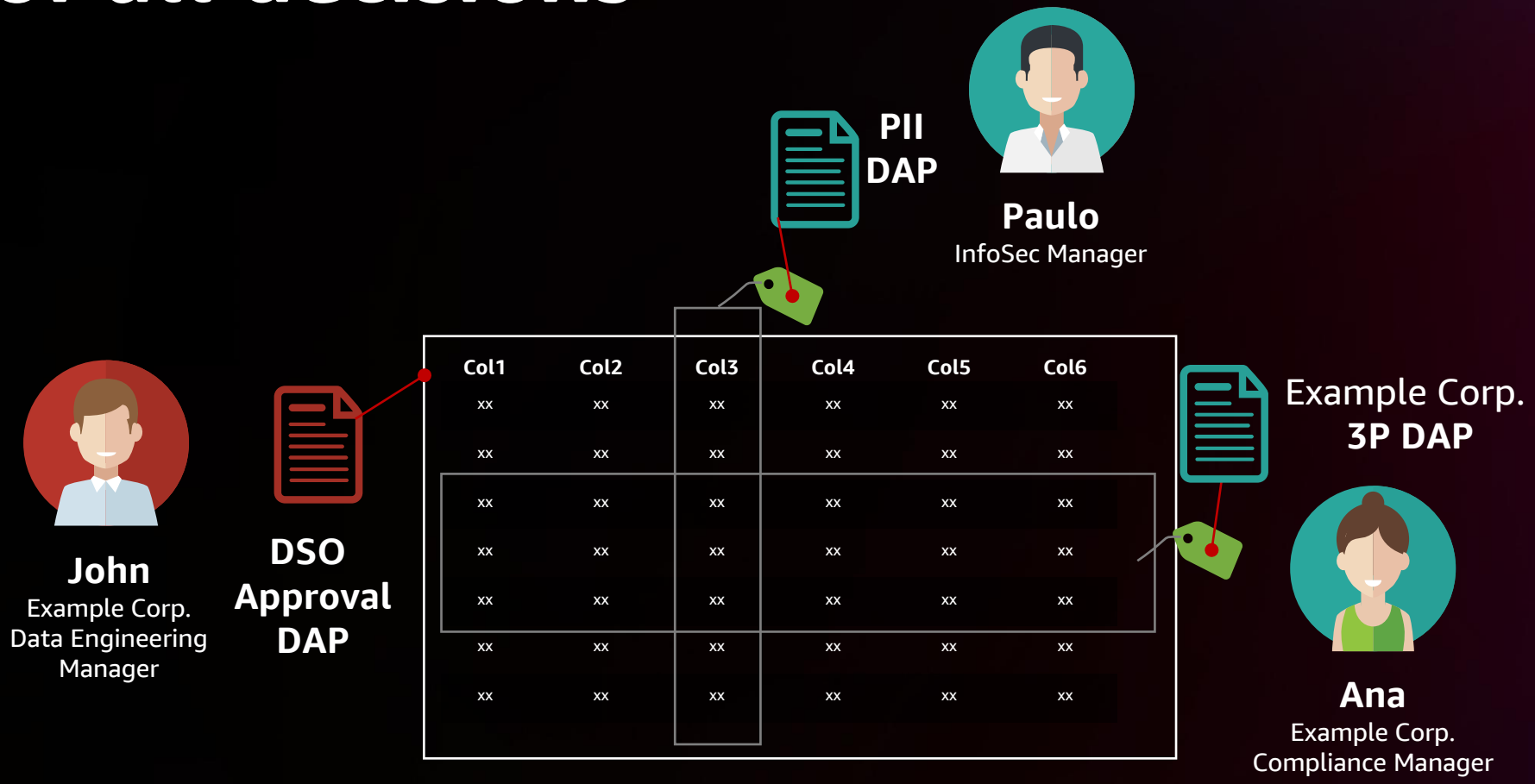

Athena

Data consumers can query the Andes dataset approved for access using the endpoint configured to access the Andes dataset with the identity approved to read the dataset

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

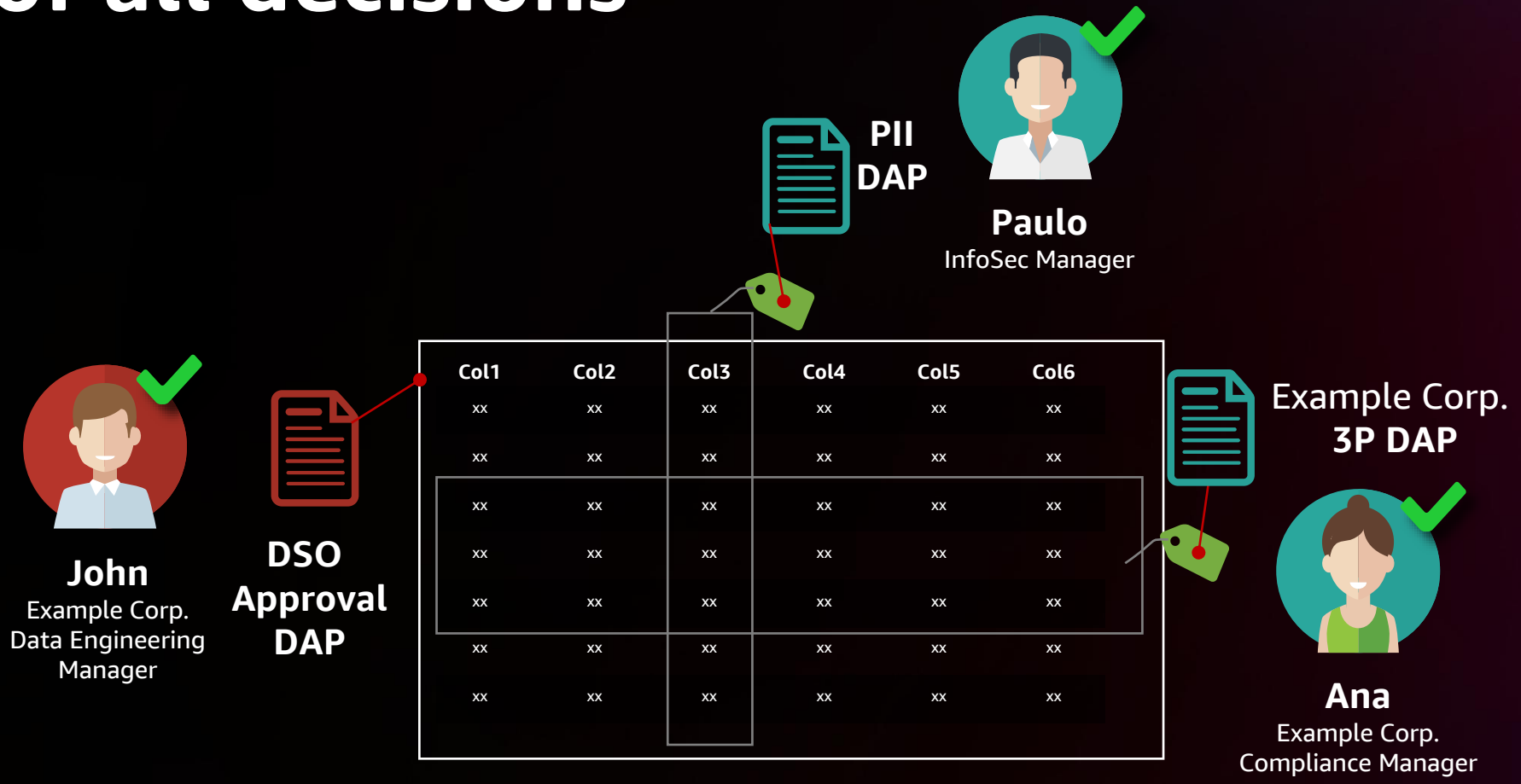
Consumers are granted access based on union of all decisions



When multiple DAPs overlap, requires approval from all to have access

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

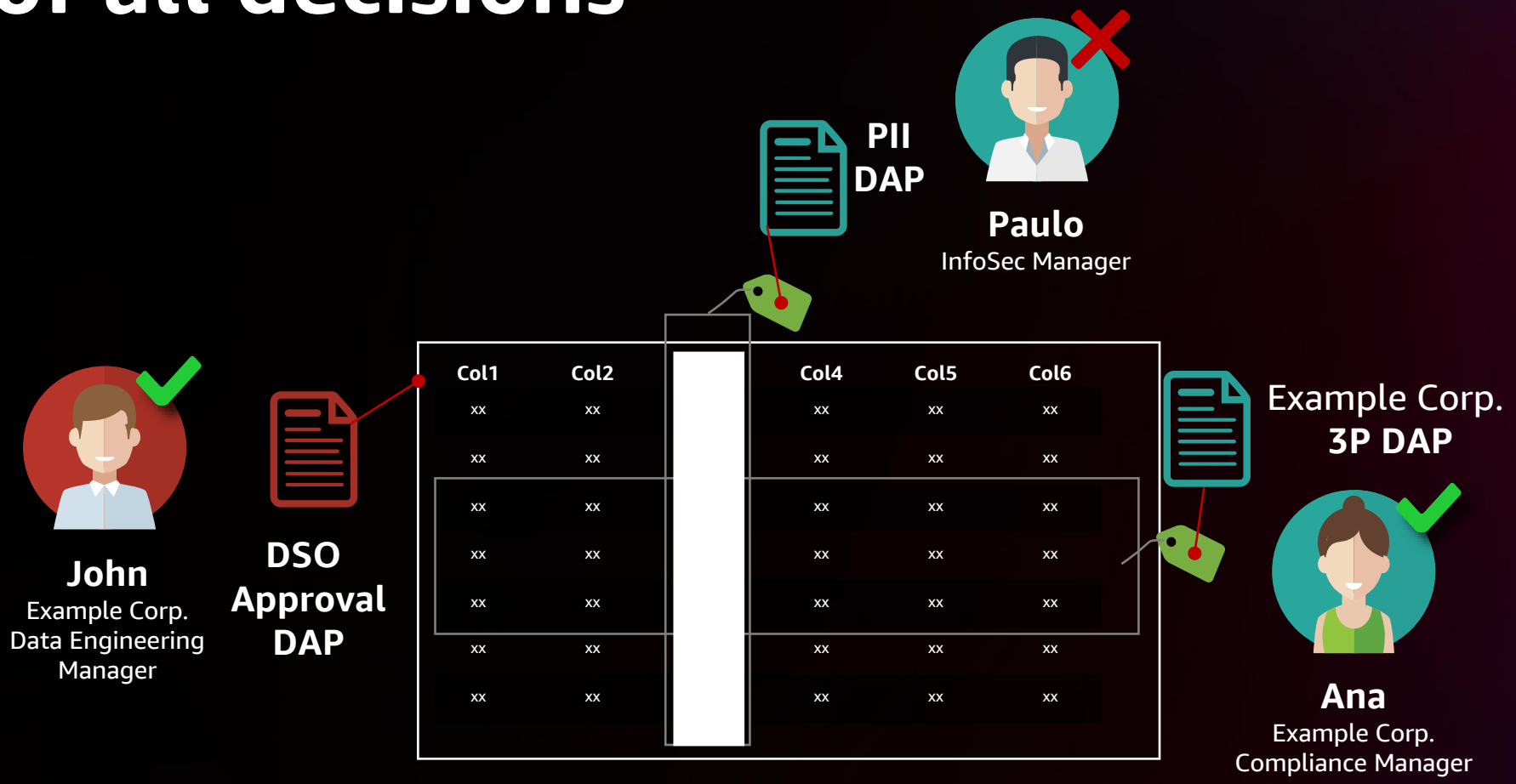
Consumers are granted access based on union of all decisions



If all approve, access is granted for the entire dataset

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

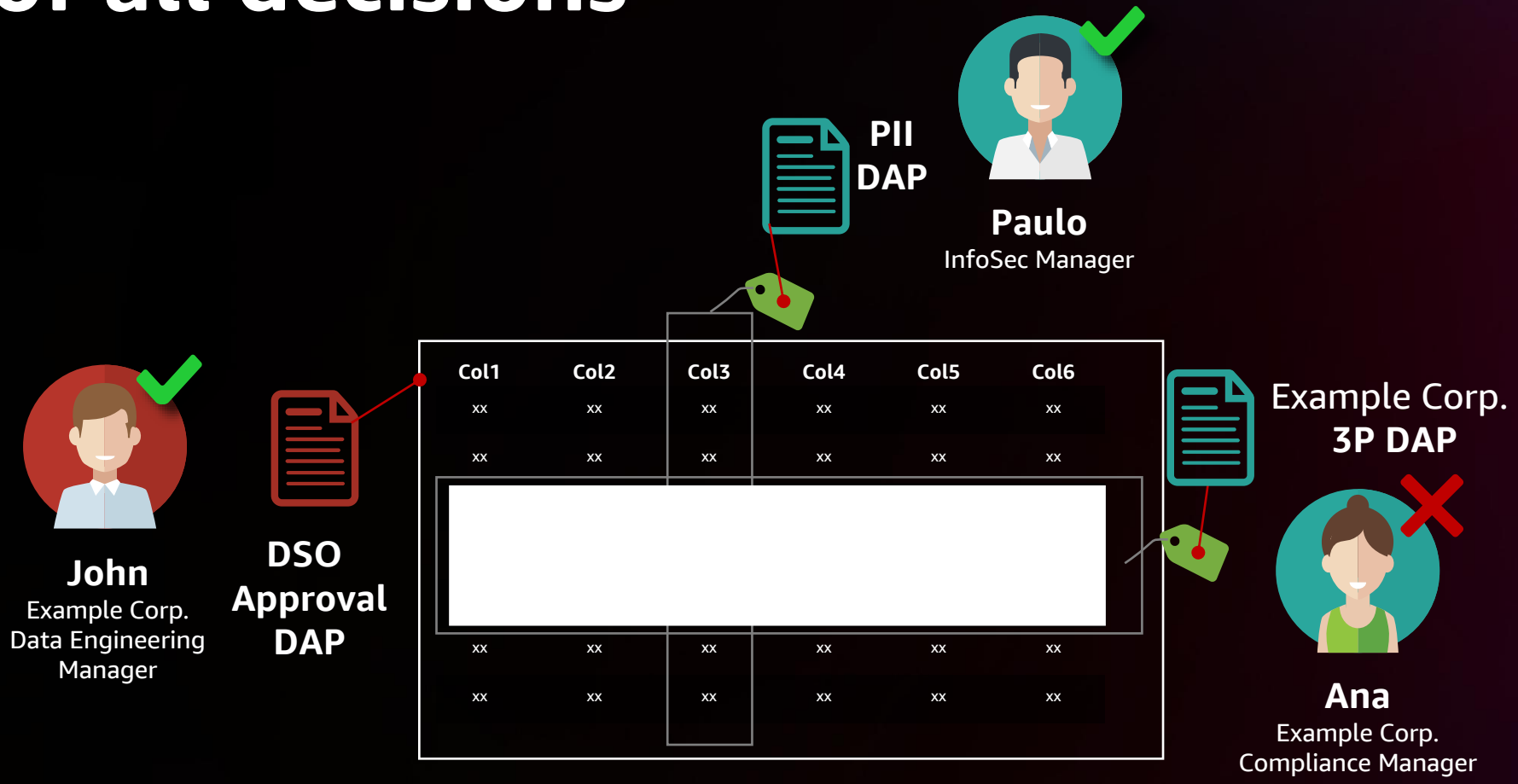
Consumers are granted access based on union of all decisions



Any deny decision removes its scope from the allowed data scope – even those approved by another overlapping DAP

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

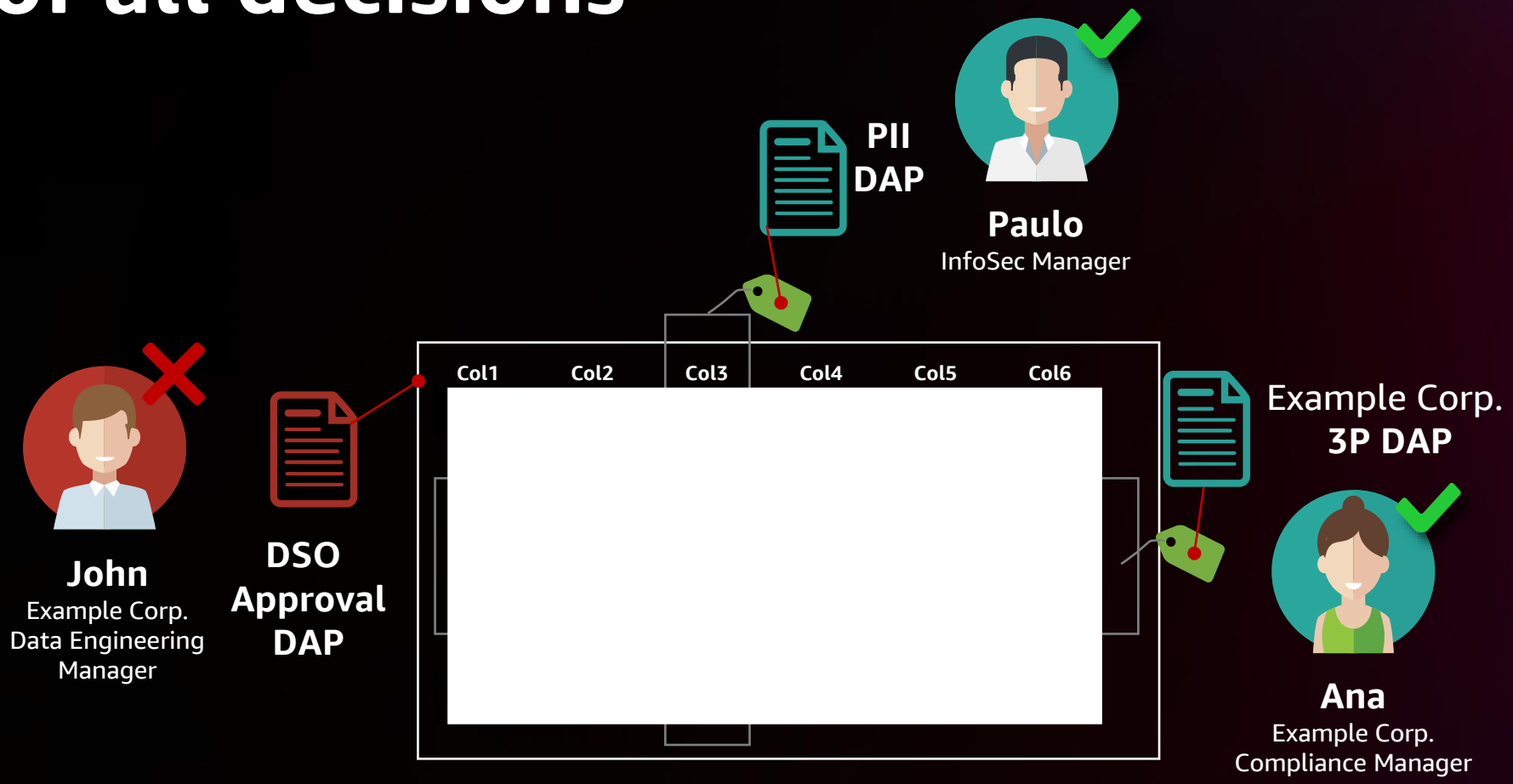
Consumers are granted access based on union of all decisions



Any deny decision removes its scope from the allowed data scope – even those approved by another overlapping DAP

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consumers are granted access based on union of all decisions



If denied by DSO, no access is granted, even if approved by all others

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DB-style fine-grained permissions on resources

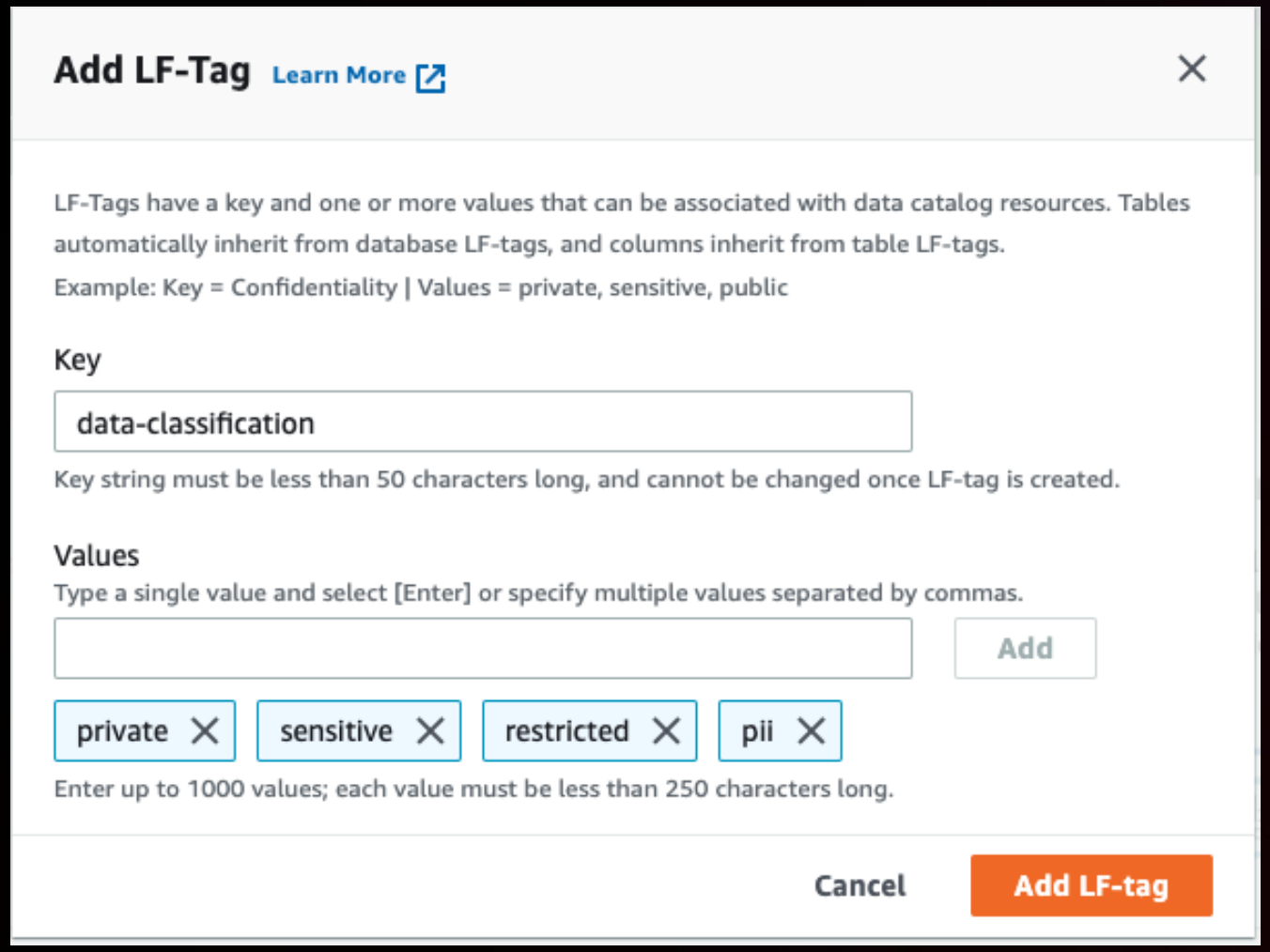
DB-style GRANT/REVOKE permissions
Confer capabilities to principals on resources

Principals	Resources	Metadata permissions	Data permissions
IAM users and roles	Catalog	CREATE_DATABASE	NA
SAML users and groups	Databases	CREATE_TABLE, ALTER, DROP	NA
	Table	DESCRIBE, ALTER, DROP	INSERT, DELETE, SELECT
	Data filters (row & cells)	DESCRIBE, DROP	SELECT
	Columns	NA	SELECT

Delegate permissions management with the GRANTABLE option
Effective permissions for a principal are the union of all GRANTS

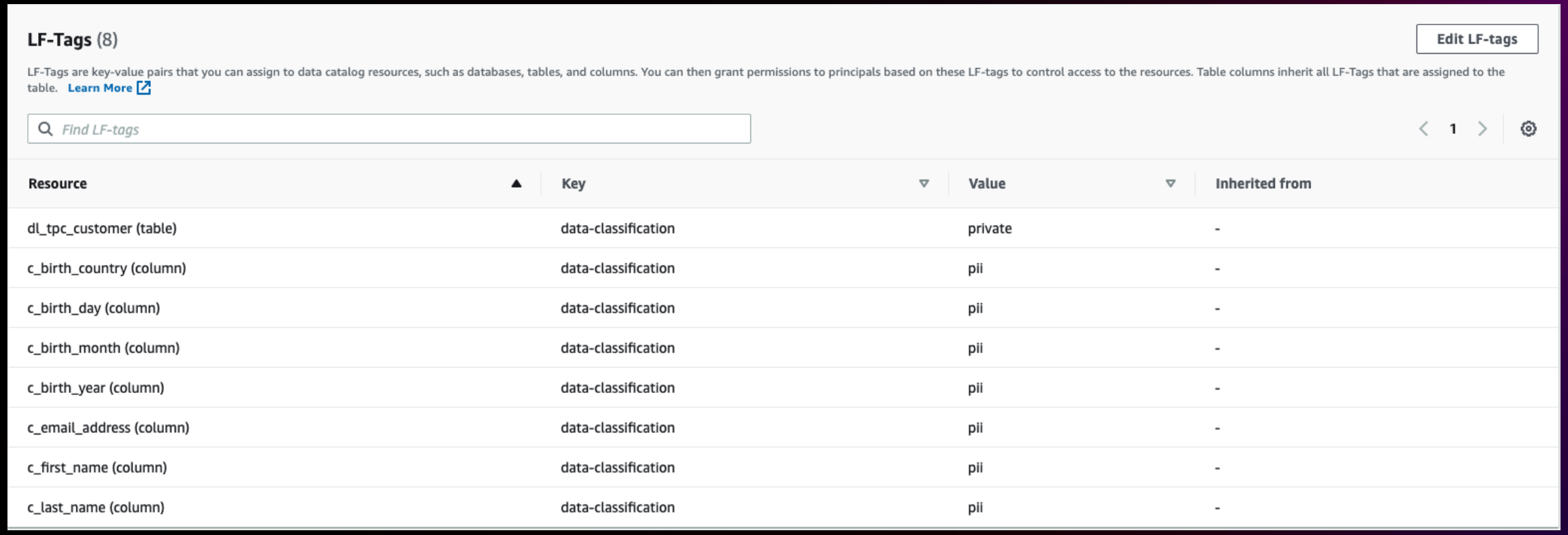
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Defining LF-Tag ontology



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Associating LF-Tags with resources



LF-Tags (8)

Edit LF-tags

LF-Tags are key-value pairs that you can assign to data catalog resources, such as databases, tables, and columns. You can then grant permissions to principals based on these LF-tags to control access to the resources. Table columns inherit all LF-Tags that are assigned to the table. Learn More

Find LF-tags

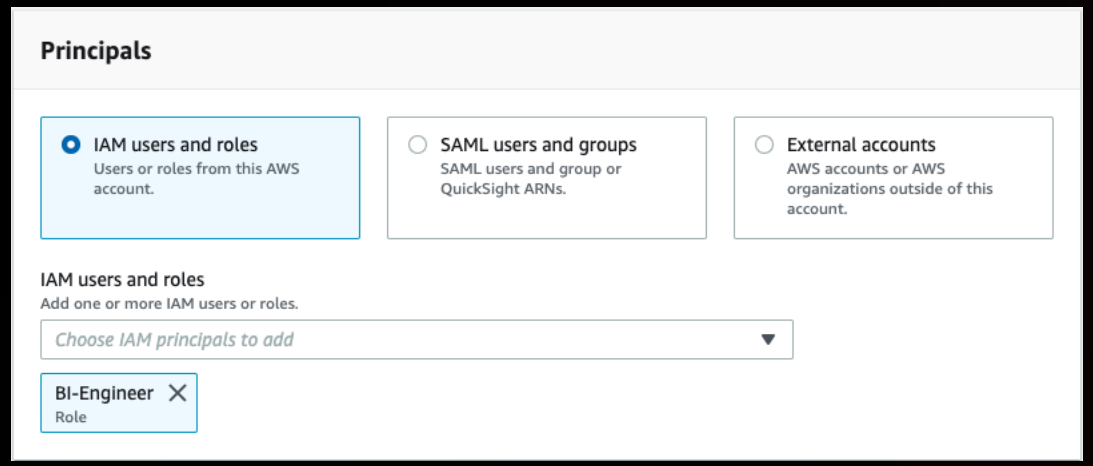
1

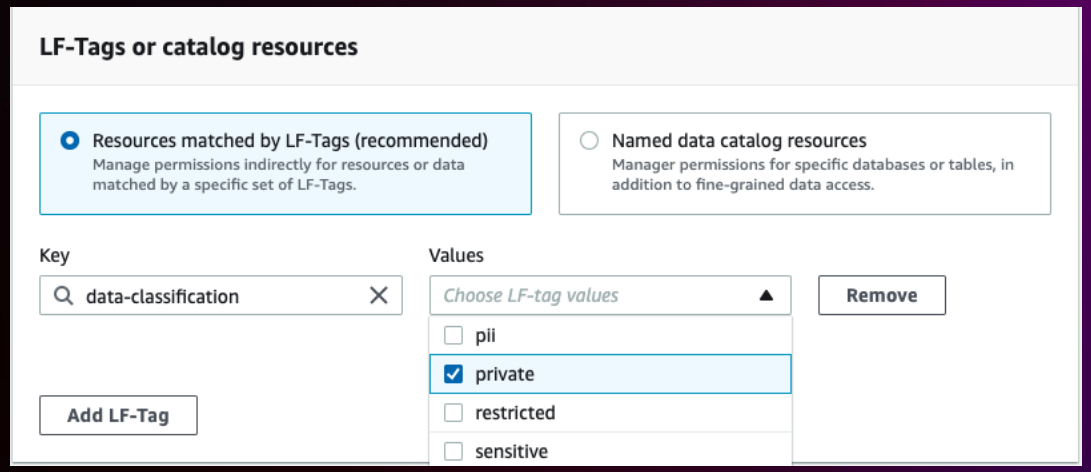
Resource	Key	Value	Inherited from
dl_tpc_customer (table)	data-classification	private	-
c_birth_country (column)	data-classification	pii	-
c_birth_day (column)	data-classification	pii	-
c_birth_month (column)	data-classification	pii	-
c_birth_year (column)	data-classification	pii	-
c_email_address (column)	data-classification	pii	-
c_first_name (column)	data-classification	pii	-
c_last_name (column)	data-classification	pii	-

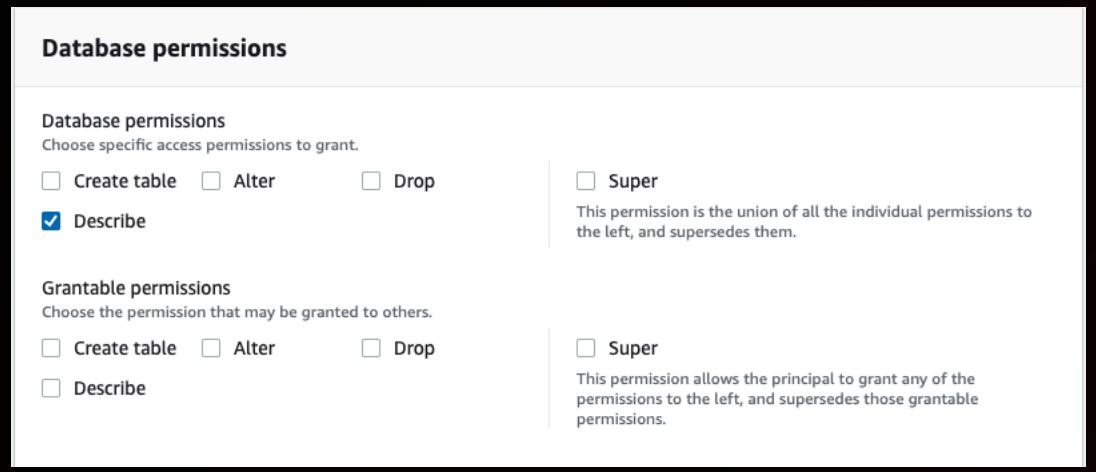
aws

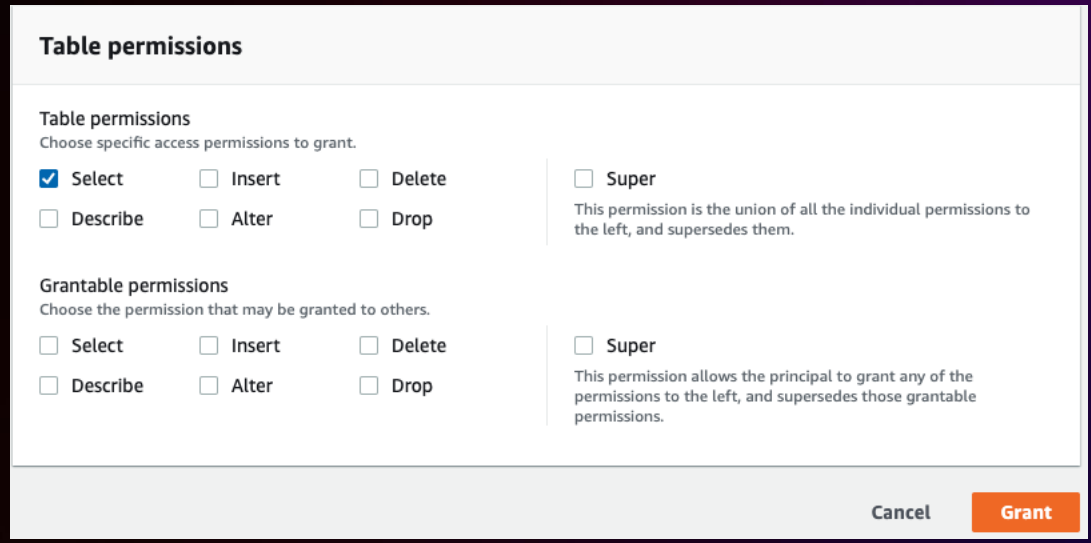
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Granting permissions based on LF-tags



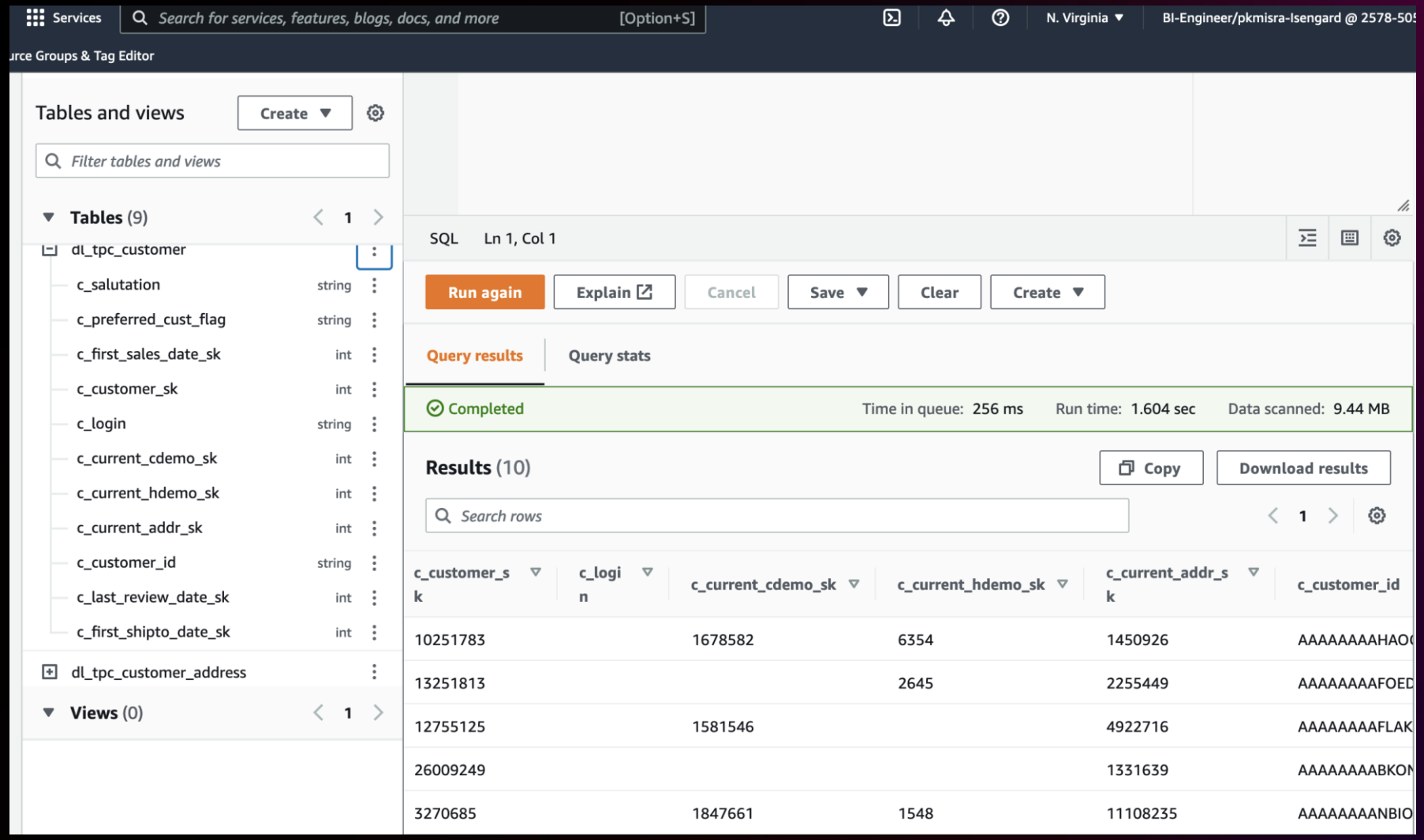






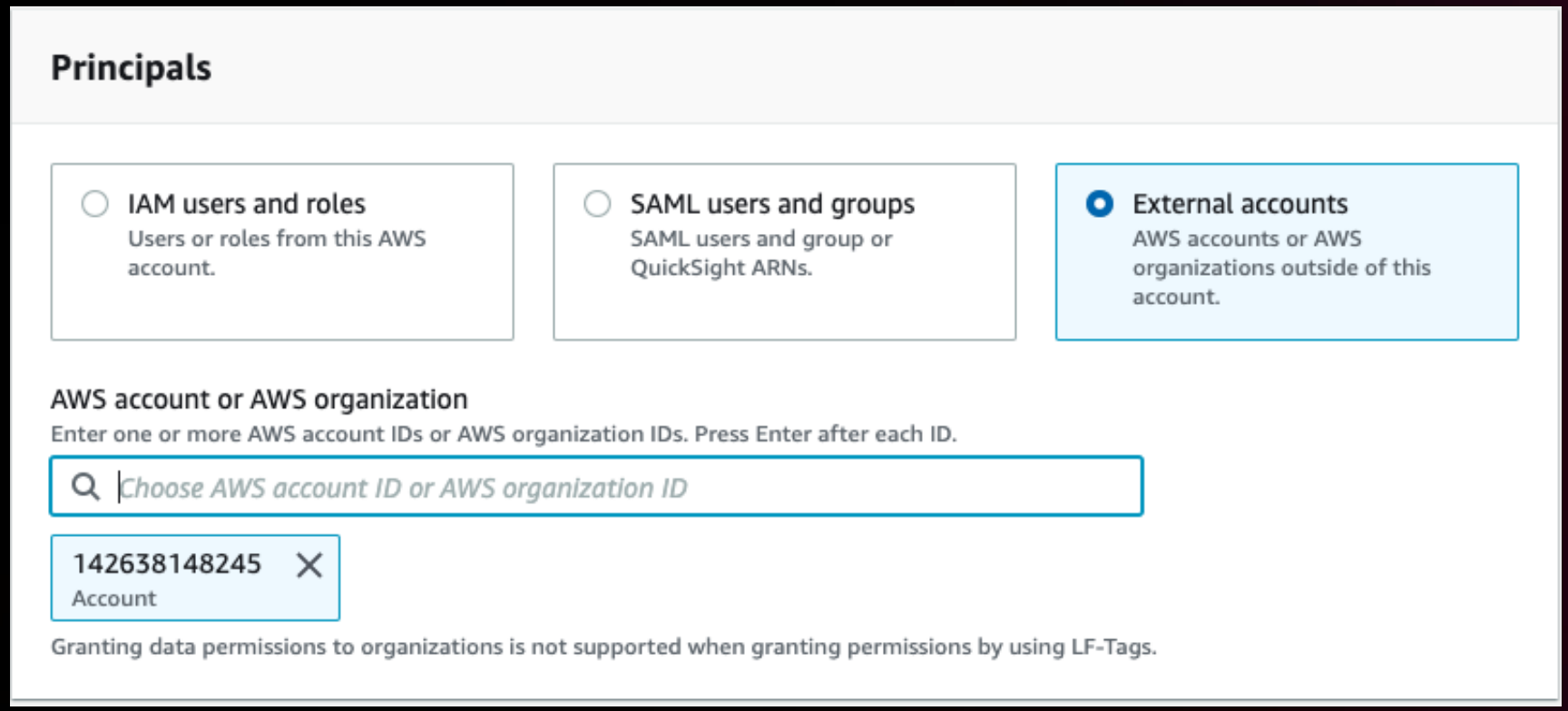
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Accessing LF-Tag data – BI engineer



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

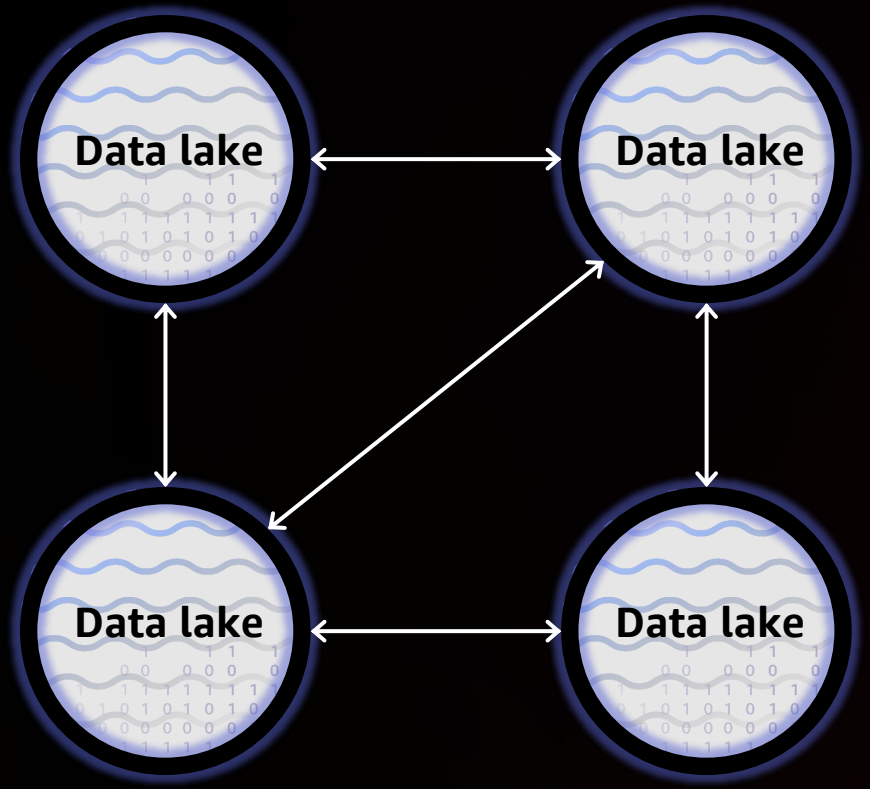
How to set up LF cross-account sharing



Share with accounts, organization, organization unit, or external accounts

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data mesh: Ownership and governed sharing

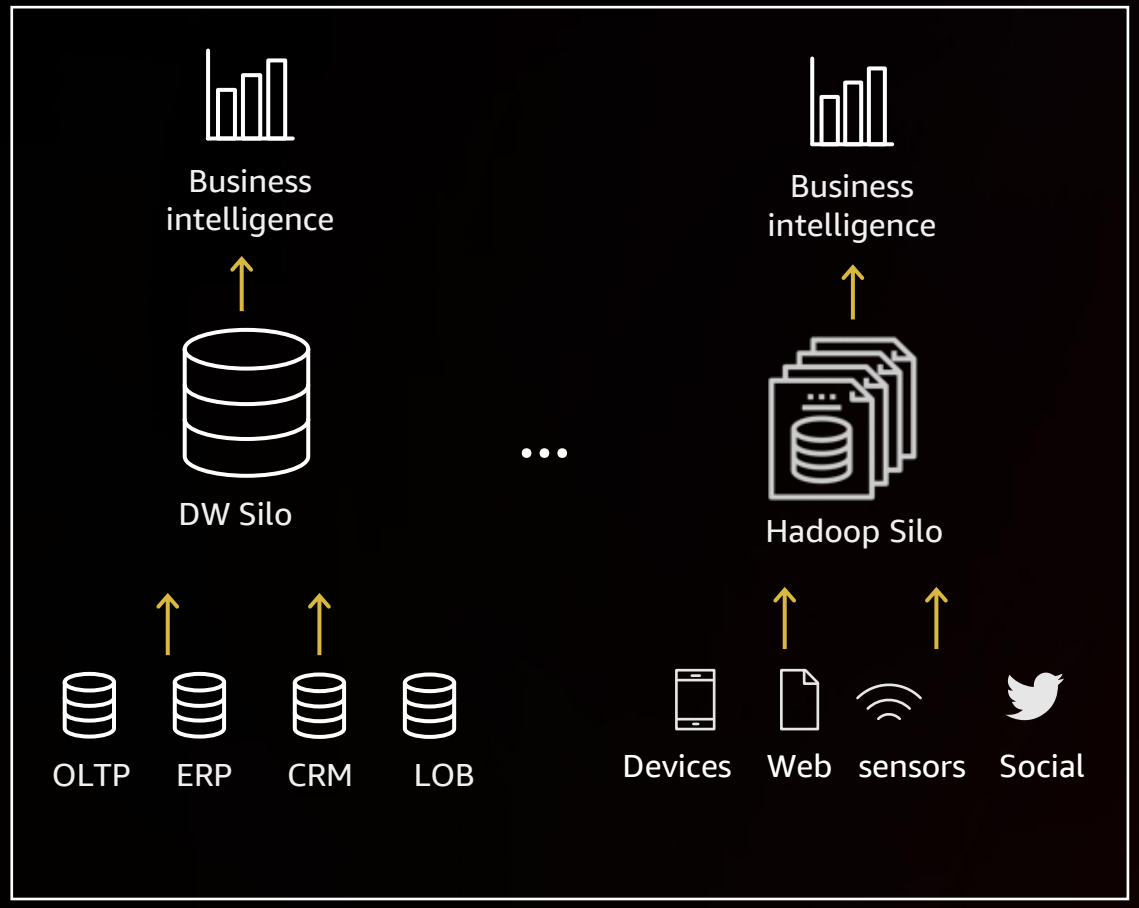


A **decentralized, domain-oriented** data architecture to drive **governed sharing** of **data products**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

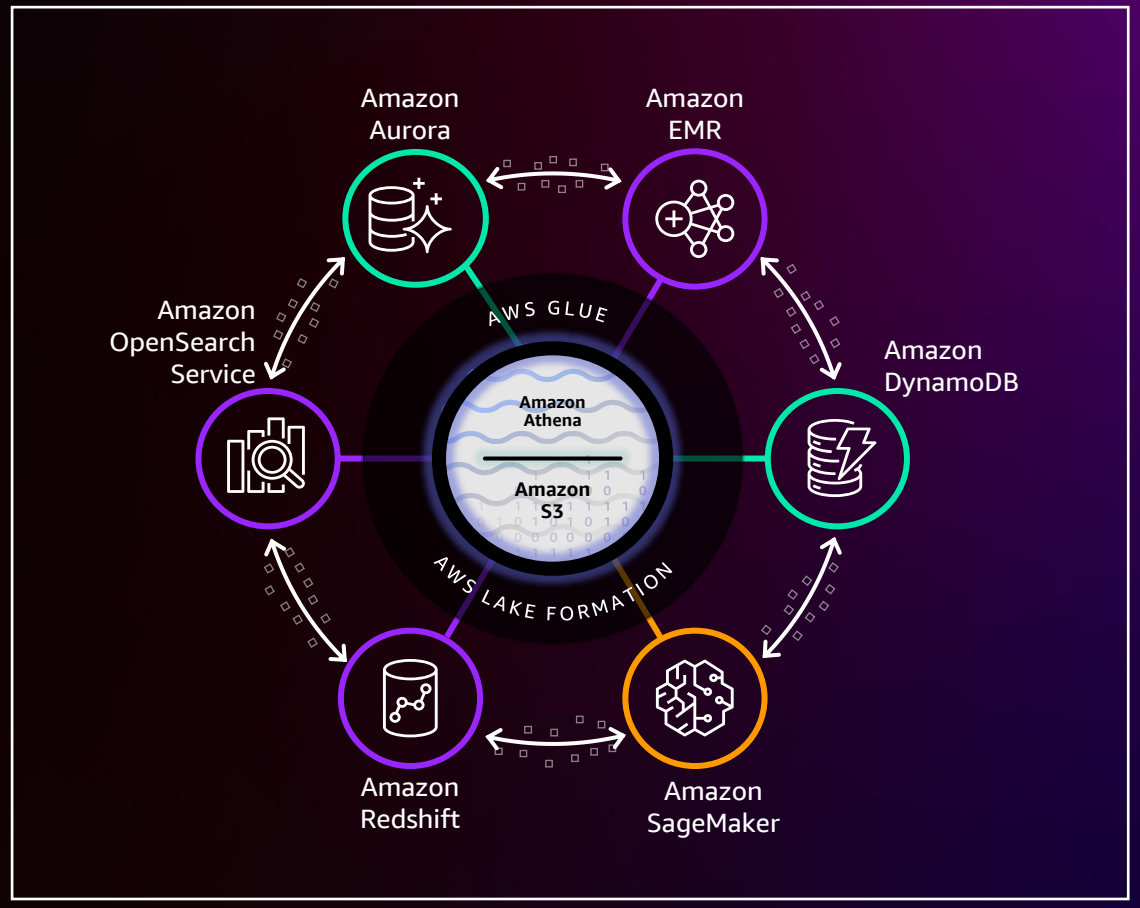
Customers moving from

Siloed data



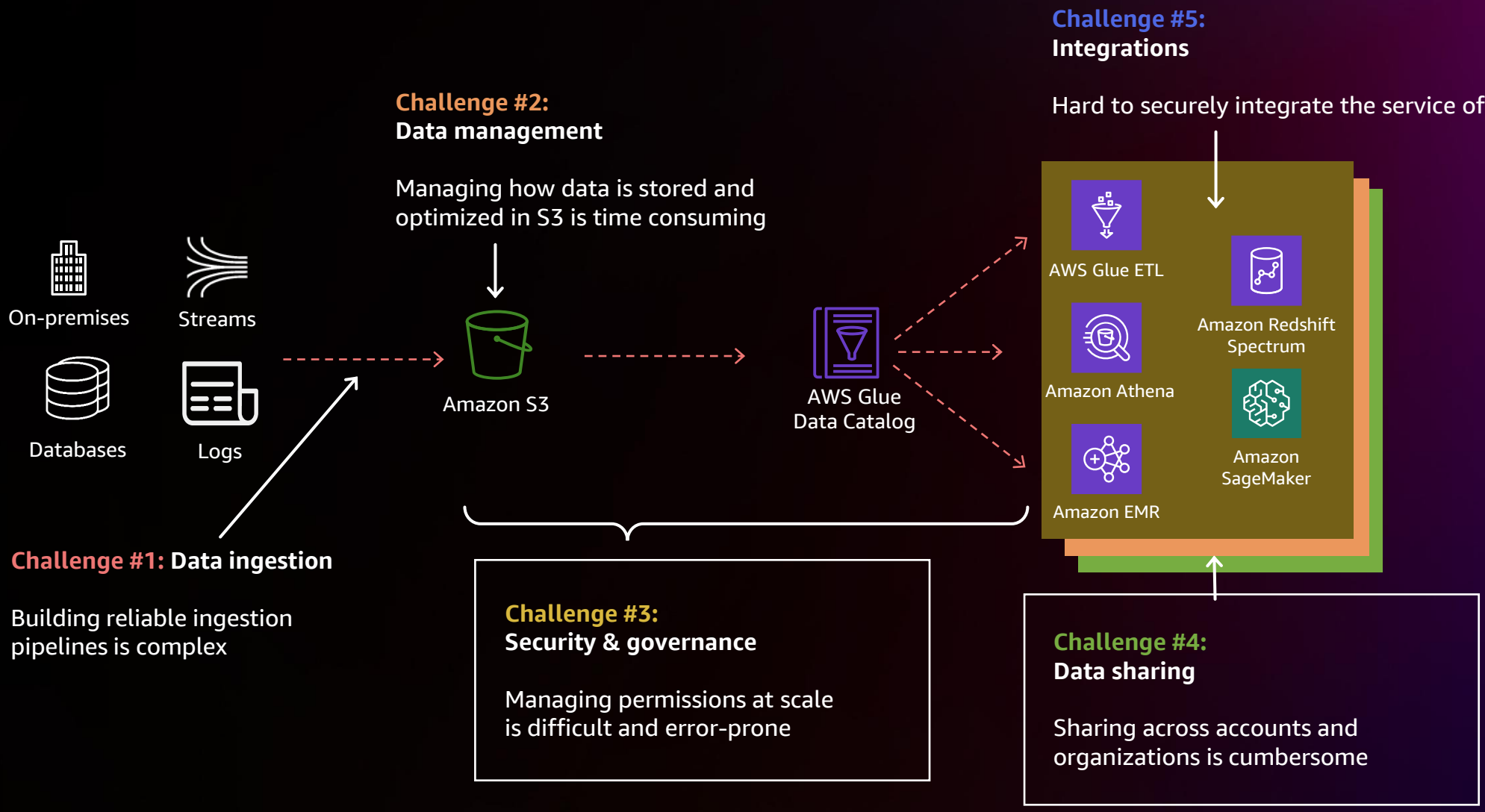
to

Modern data architectures



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

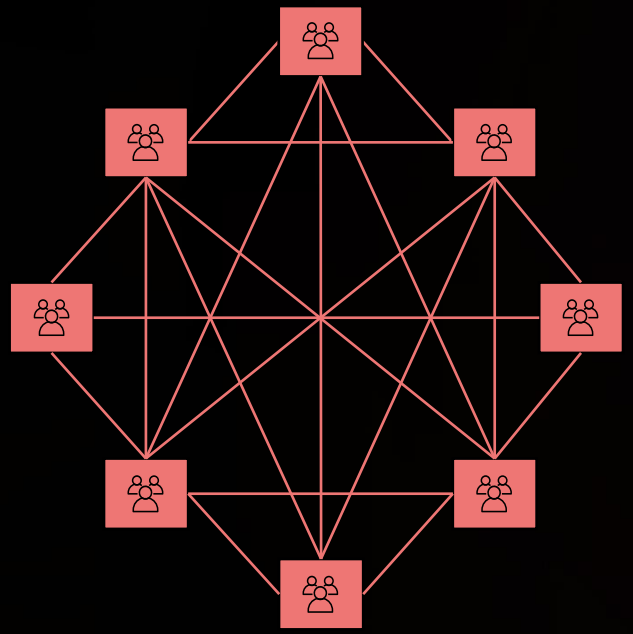
Challenges with building data lakes



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next evolution of our data lake

EXTENDING THE DATA LAKE FOLLOWING THE DATA MESH DESIGN PATTERN





AWS
Lake Formation



Consumer Bank

Corporate Technology

Corporate Investment Bank

Line of business autonomy

Enterprise data products

Federated governance

Central data discovery

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.