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

ANT204

Enabling agility with data governance on AWS

Jason Berkowitz (he/him)

Sr. Manager, AWS Lake Formation
AWS

Shihas Vamanjoor (he/him)

Vice President, Product Owner - Enterprise Data Platforms
Prudential

Agenda

How does data governance help you become data driven?

Data governance patterns with AWS analytics services

Prudential Financial Services data journey

Data driven themes we hear from customers

THE TRANSFORMATION IS CHALLENGING, REQUIRING A STRONG VISION, NEW CULTURE, SKILLS, AND TECHNOLOGY

Understanding what “great looks like”

Identifying and prioritizing use cases

Creating sponsorship & business case

Creating a data-driven culture

Gaps in skills and technologies

Data privacy, security, compliance, and governance

Data governance is essential to being data driven

85% of businesses want to be data driven

"Data governance is **no longer optional** for enterprise organizations. They are finally realizing the value of data as an asset that needs to be protected, managed, and maintained to increase asset value."

IDC

Only

37% have been successful

"Organizations lack data knowledge for efficient and effective data governance activities**; 30% of time spent on data governance is wasted**."

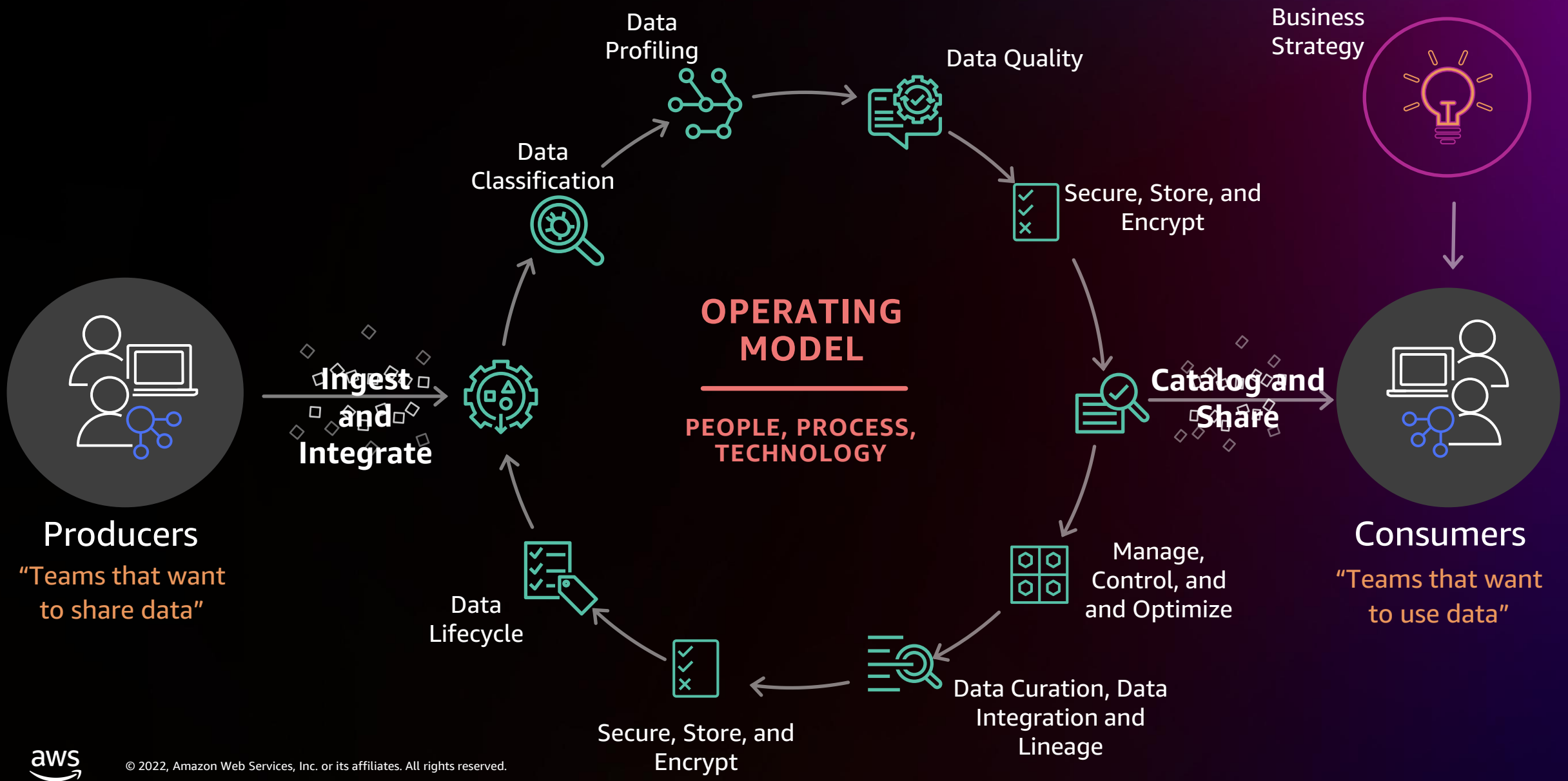
IDC

Stat Source: Forbes Online; New Vantage Partners – Big Data Executive Survey

Definition

Data governance is the collection of policies, processes, and systems that organizations use to ensure the quality and appropriate handling of their data throughout its lifecycle for the purpose of generating business value

Data governance starts with business



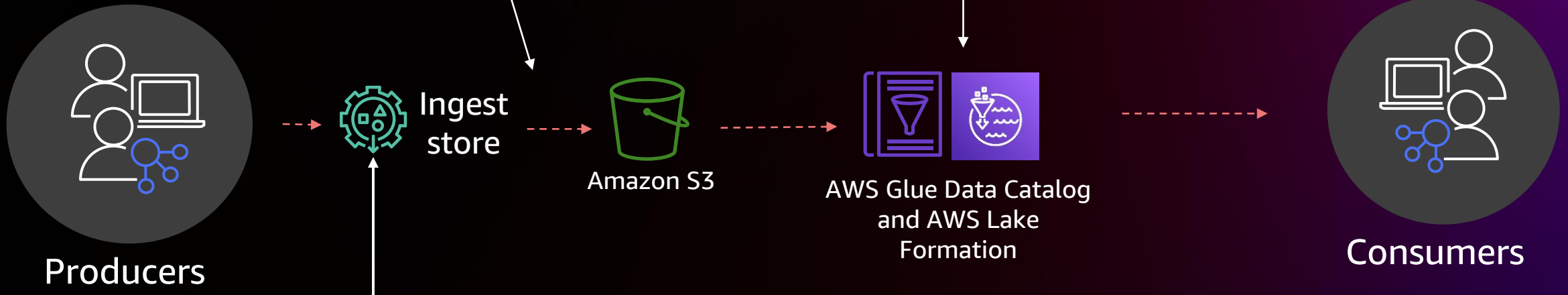
How do AWS services help?

Data governance across the data pipelines

#1: Automate Data Ingestion, Classification and Quality
Automate the creation of reusable data pipelines within the enterprise, define data quality rules, and establish data classification tagging that can drive permissions

#3 Share Data to Data Consumers

Share data so that data consumers can build new insights

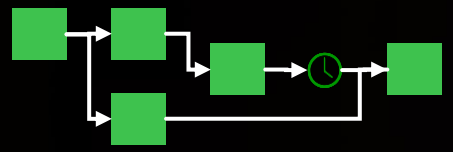


#2 Catalog Data to Make It Useable
Manage a data catalog so that data can be found easily, access can be requested and used to make decisions

Publish new data products

Automate data governance on ingestion

Automate Ingestion Pipelines

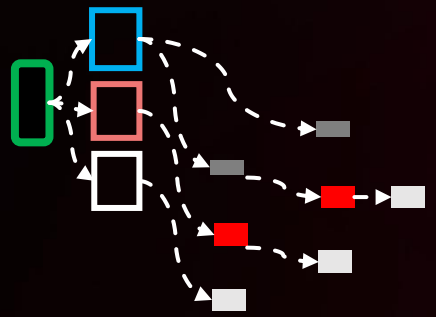


User and use case driven

Automate CI/CD pipelines

Many data sources (RDBMS, files, streams, SaaS)

Inconsistent Performance, Reusability, and Quality



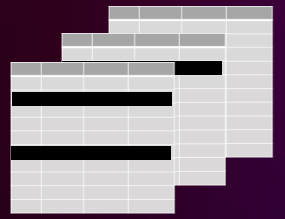
Data stored how it arrived

Inconsistent data formats

Leverage data profiling

Standardized data quality rules

Complying with Regulations



Classify and tag PII data on the way in

Build data lifecycle policies

Example automation data ingestion & store

Data Classification

Data Profiling

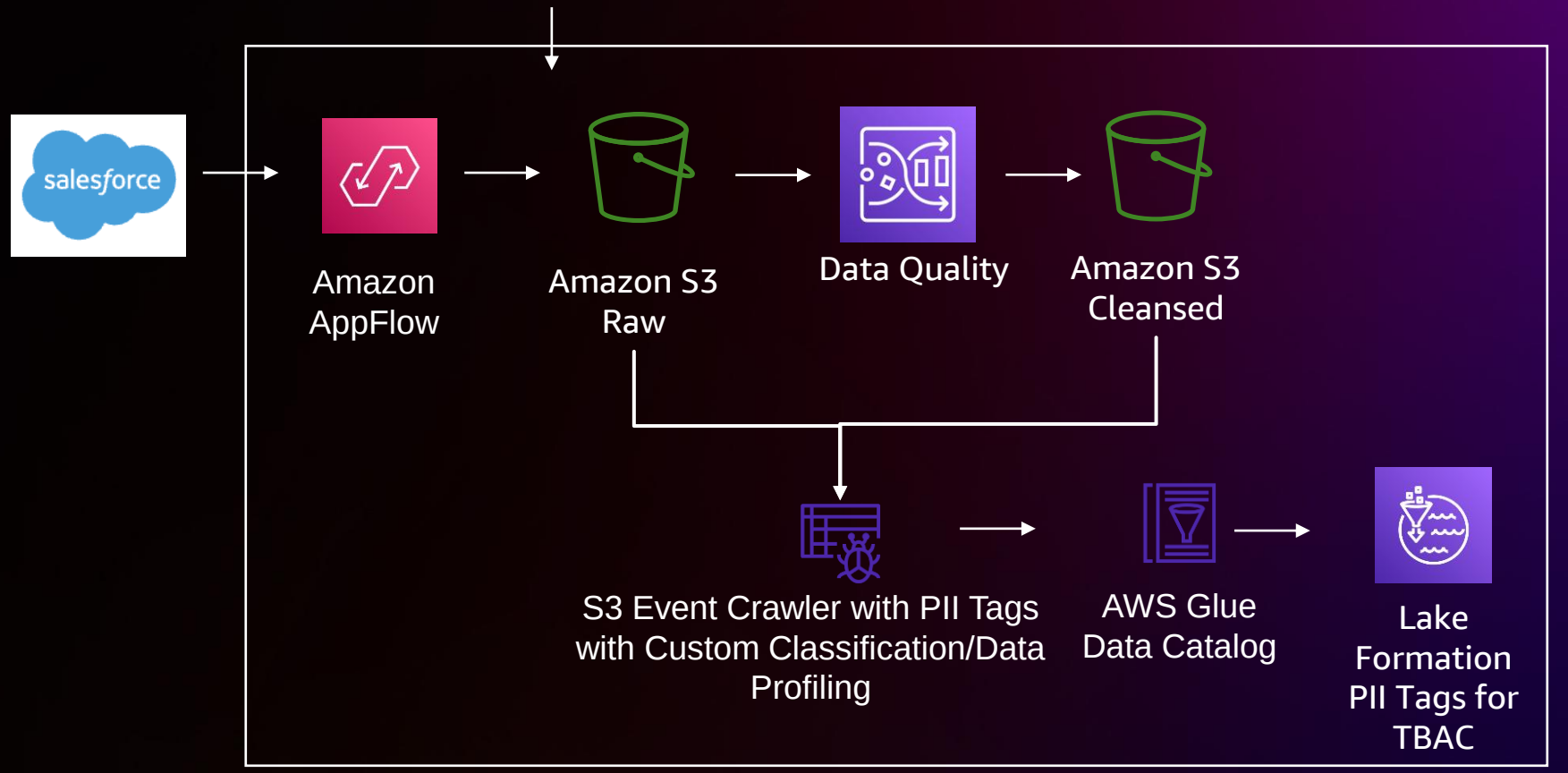
Data Quality

Secure, Store, and Encrypt

Catalog

AWS Data Ops Development Kit (AWS CDK)

#1: Automate Data Ingestion, Classification, and Quality
Automate the creation of reusable data pipelines within the enterprise, define data quality rules, and establish data classification tagging that can drive permissions



Catalog your data for findability

#1: Automate Data Ingestion, Classification, and Quality
Automate the creation of reusable data pipelines within the enterprise, define data quality rules, and establish data classification tagging that can drive permissions

#3 Share Data to Data Consumers

Share data so that data consumers can build new insights

Producers

Ingest
Store

Amazon S3

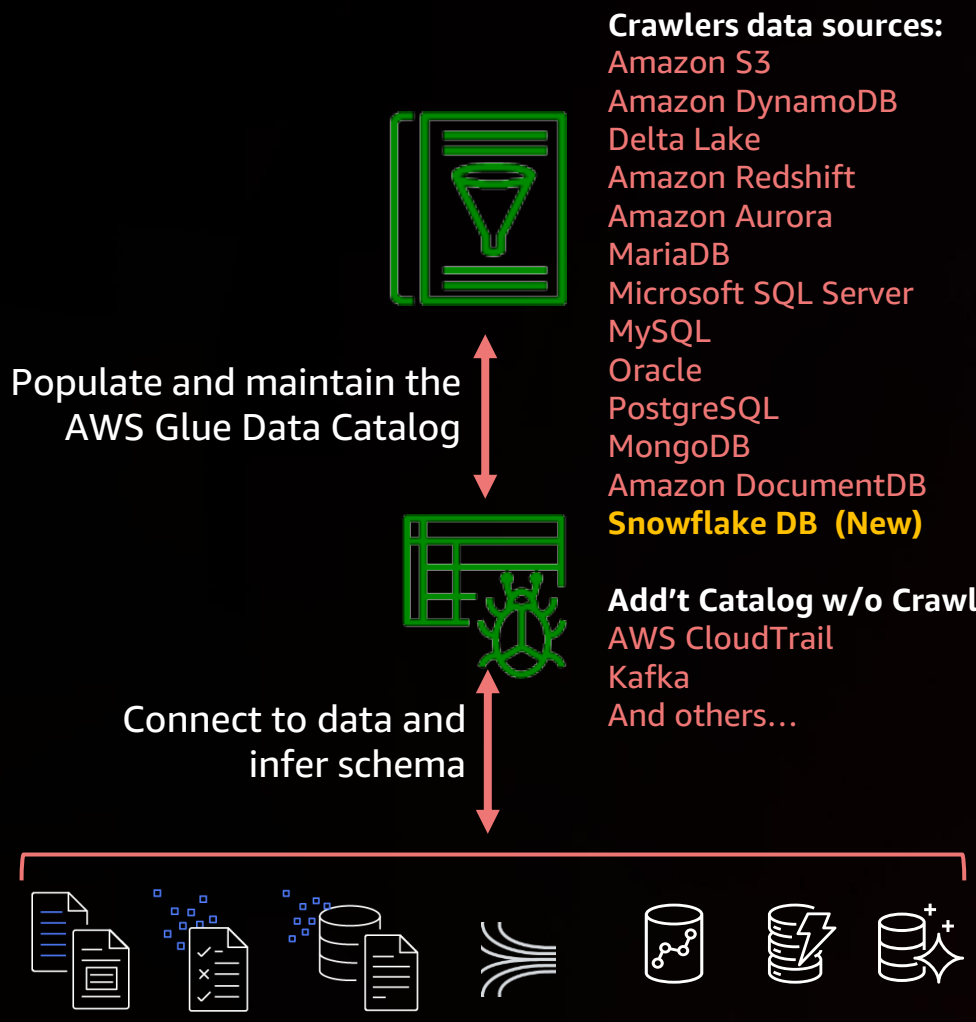
AWS Glue Data Catalog
and Lake Formation

Consumers

#2 Catalog Data to make it useable
Manage a data catalog so that data can be found easily; access can be requested and used to make decisions

Publish new data products

Crawl datasets to remove heavy lifting



Crawlers data sources:
Amazon S3
Amazon DynamoDB
Delta Lake
Amazon Redshift
Amazon Aurora
MariaDB
Microsoft SQL Server
MySQL
Oracle
PostgreSQL
MongoDB
Amazon DocumentDB
Snowflake DB (New)

Add't Catalog w/o Crawlers
AWS CloudTrail
Kafka
And others…

Crawlers - Automatically discover new data and extract schema definitions

Detect schema changes and maintain tables, determine partitions on Amazon S3

Use built-in classifiers for popular data types such as PII or create your own custom classifier using Grok expressions

Profile data to share table statistics through a single catalog (New)

Run on demand, incrementally, on a schedule, on an event, or catalog data built in AWS Glue or Amazon AppFlow (New)

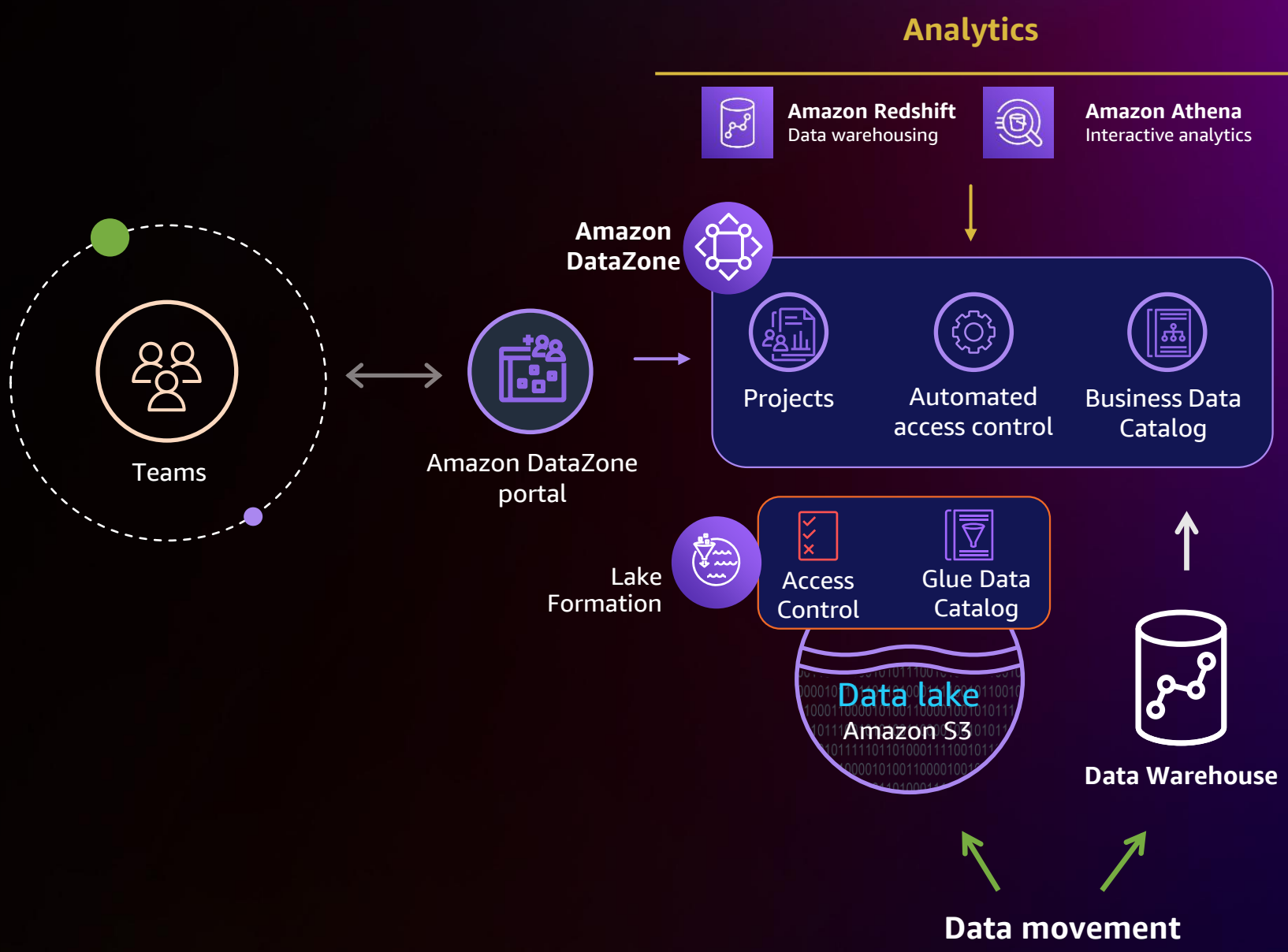
Amazon DataZone extends the AWS Analytics stack

ORGANIZATION-WIDE BUSINESS DATA CATALOG

GOVERNANCE AND ACCESS CONTROL

SIMPLIFIED ACCESS TO ANALYTICS

DATA PORTAL



Share your data easily

#1: Automate Data Ingestion, Classification, and Quality
Automate the creation of reusable data pipelines within the enterprise, define data quality rules, and establish data classification tagging that can drive permissions

#3 Share Data to Data Consumers

Share data so that data consumers can build new insights

Producers

Ingest
Store

Amazon S3

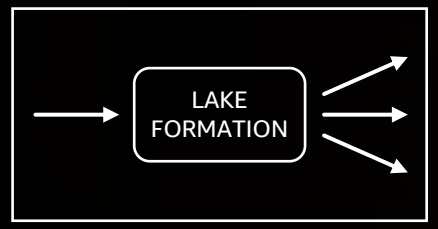
AWS Glue Data Catalog
and Lake Formation

Consumers

#2 Catalog Data to Make It Useable
Manage a data catalog so that data can be found easily; access can be requested and used to make decisions

Simple data sharing with Lake Formation

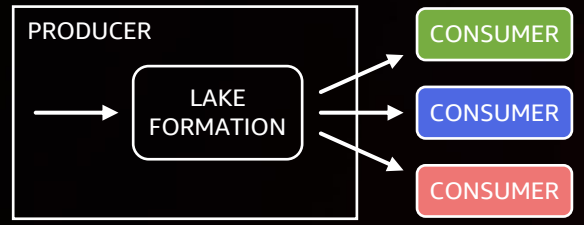
Single Account



Centralized
Single Account

Simple to get started

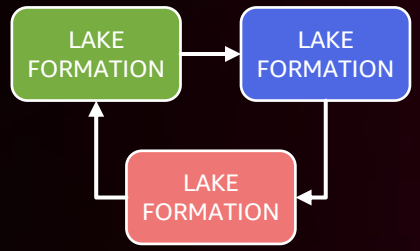
Hub and Spoke



Hub and Spoke
Multi-Account

Cross-organization

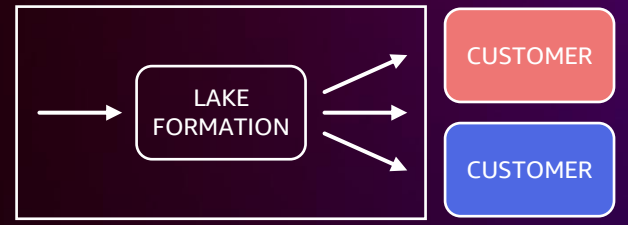
Data Mesh



Data Mesh
Central Governance

Organizational autonomy

Business to Business

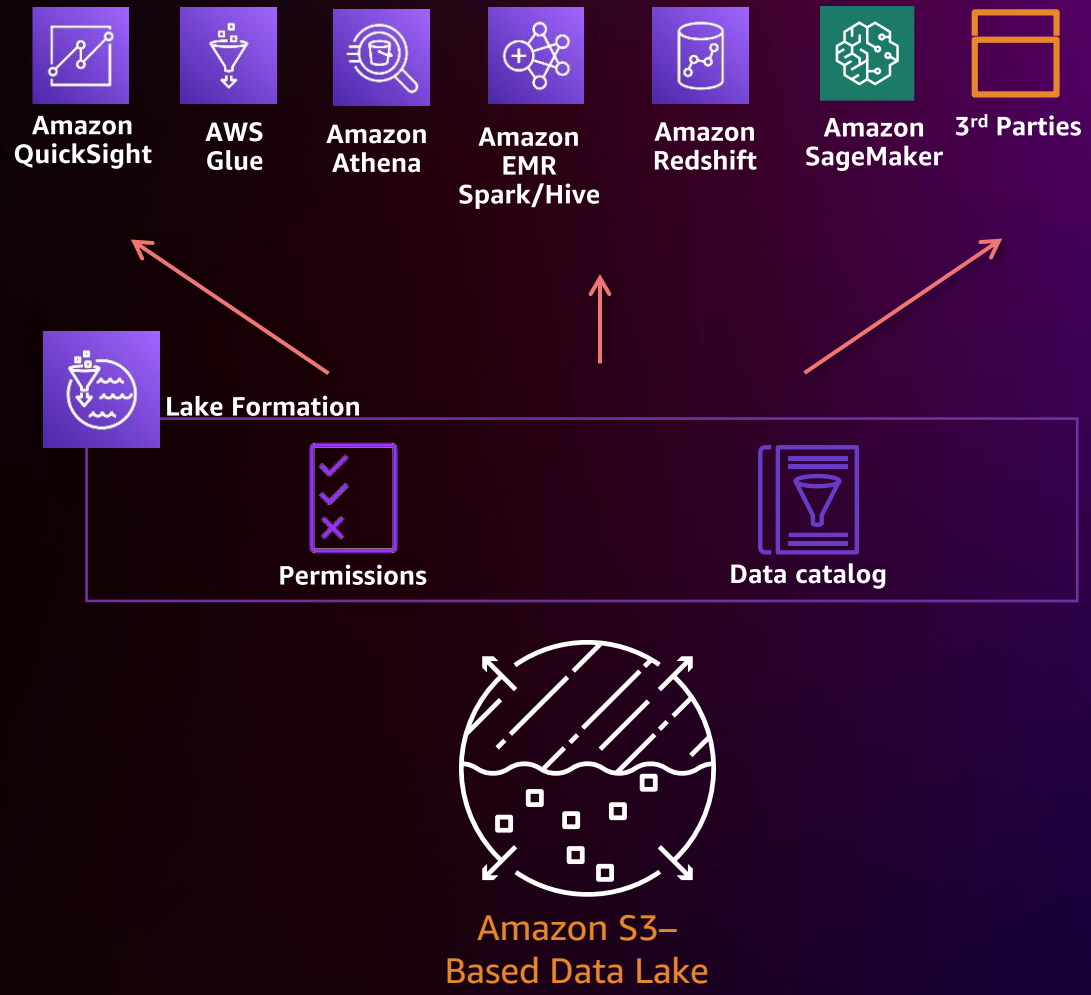


Multi-Customer

Cross-organization

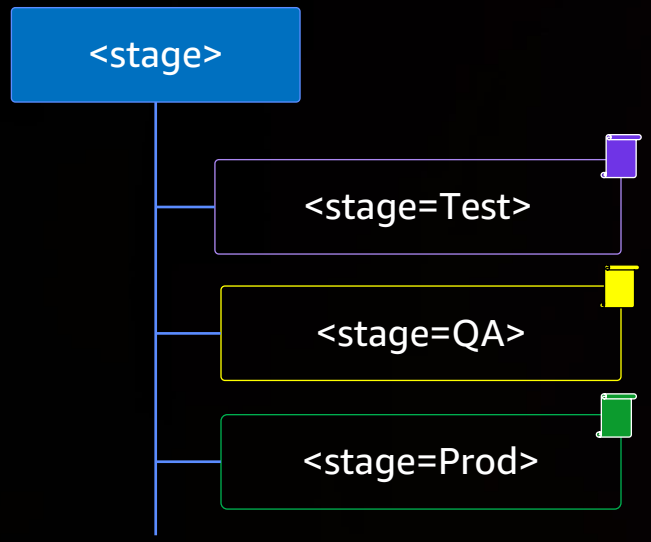
Lake Formation permissions model

- DB-style fine-grained permissions on resources
- Scale permissions management Lake Formation Tag-Based Access Control (LF-TBAC)
- Unified Amazon S3 permissions
- Integrated with services and tools
- Easy to audit permissions and access



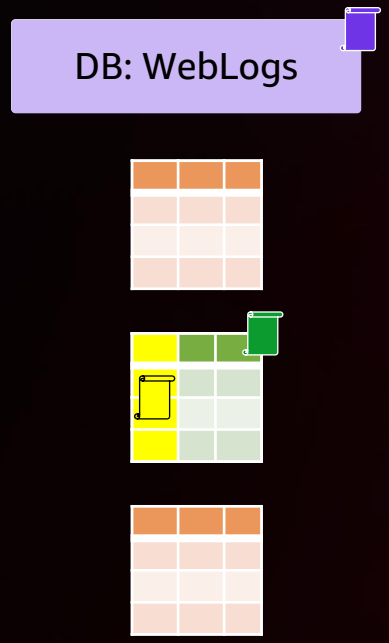
Leverage Lake Formation TBAC to scale permissions

Define LF-Tags



Specify who can assign LF-Tags and values

Assign LF-Tags to resources



Tag databases, tables, columns

LF-Tags are hierarchical and may be overridden

Create policies on LF-Tags

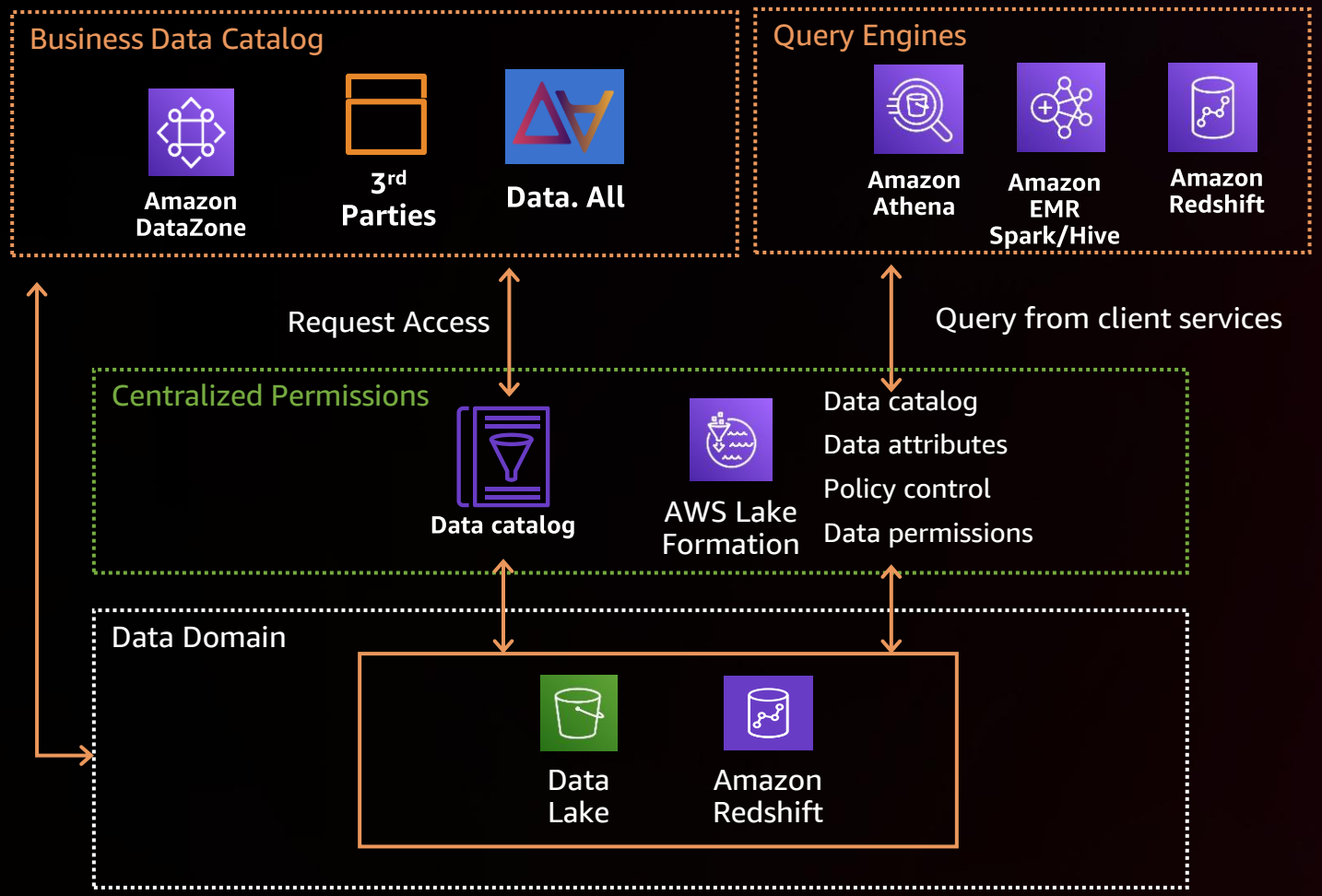
IAM user/role

AD users and groups (SAML assertion)

Scale by applying permission on LF-Tags

Automate data sharing

AN OPEN APPROACH TO DATA SHARING



Data Consumer

- Enable personas to discover, understand, consume, and request access
- Abstract complexity through automation
- Enforce governance through data classification built into LF Tags
- Extend data marketplace capabilities to third-party catalogs

Key takeaways

- Automate ingestion pipelines and include compliance controls, standard data quality rules so data engineers can move at the speed of business
- Automate cataloging, classifying, profiling your data through crawlers and make that data available via a business catalog
- Automate the management of tag-based access control to ensure data is protected through the data lifecycle
- Automate sharing of data so users in one interface can find the data they want and immediately start working with it

Prudential enterprise data platform

Shihas Vamanjoor

Vice-President, Enterprise Data Platforms
Prudential

Agenda

About Prudential

Challenges

Solution – concept, journey, engineering, features, and experience

Lessons and business impact

Prudential

- Founded in 1875 with headquarters in Newark, NJ
- Provide insurance, retirement planning, investment management, and other products and services to retail and institutional customers
- $1.7 trillion in assets under management
- 50 million customers in over 40 countries

About me

Shihas Vamanjoor

Product Owner
Enterprise Data Platforms

Delivering data innovation at scale

- Programmer, engineer & executor
- Data platforms, data portals, marketplaces, lake houses, analytics
- Financial services, telecom, manufacturing, media & entertainment

Data athletes

VALUE CREATOR CHALLENGES

Hard to locate data

+

Long time to access

+

Lots of human engineering

+

Complex governance

+

Tedious & Repetitive work

=

Long time-to-value

Desired experiences

Simplified
Data Discovery

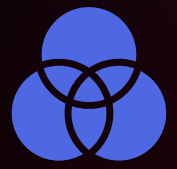
What? Make finding data for insights as easy as shopping at Amazon.com with ability to comment/rate datasets

Why? Find data in seconds versus days/weeks

Automated
Data Onboarding

What? Automated data onboarding for various ingestion patterns with governance & controls

Why? Data made available within 1 business day post-approval versus weeks/months



Optimized
Human Engineering

What? Remove yak shaving to make curators more productive on data transformation to meet business needs (ML, analytics, BI)

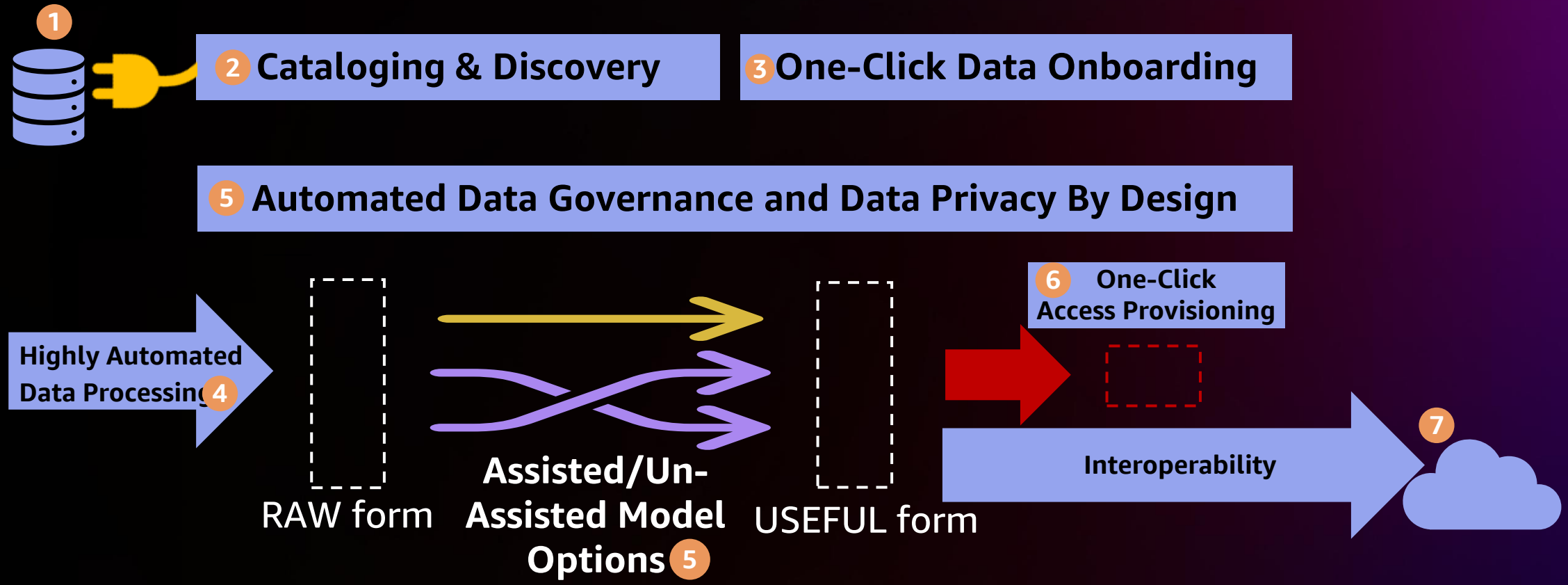
Why? Reduce data curation taxes with maximal auto-generated standardization

Frictionless
Data Consumption

What? Automated data access and data sharing via embedded governance

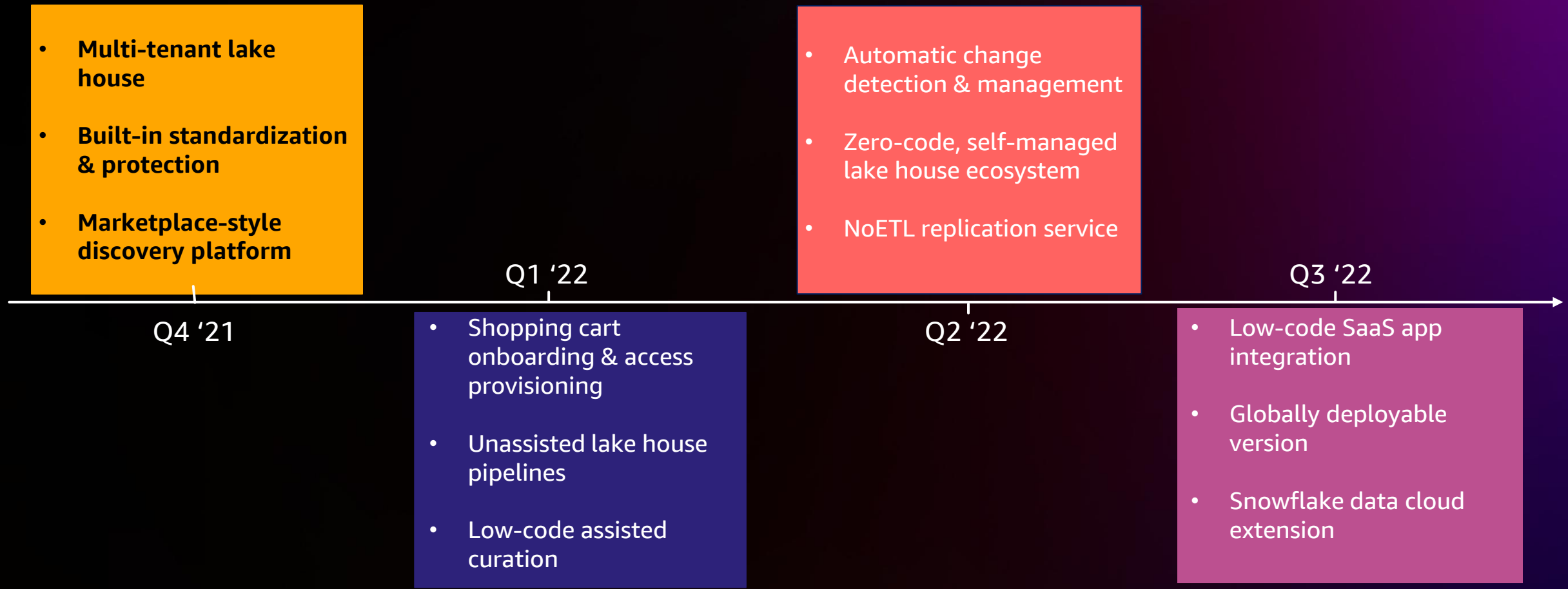
Why? Rich data ecosystem to enable collaboration and innovation at much faster clip

Data platform: Concept



Our journey

9 MONTHS AND 4 RELEASES LATER:



Team

Product Owner

ENGAGE

- Business SME
- Project Manager

- Engagement Manager
- Sr. Practice Manager

ENGINEER

- Data Engineer
- DevOps Engineer

- Lead Data Architect
- DevOps Engineer
- Data Engineer

SUPPORT

- Cloud Engineering
- Database Engineering
- InfoSec
- Informatica
- Cloud Security
- Data Analytics

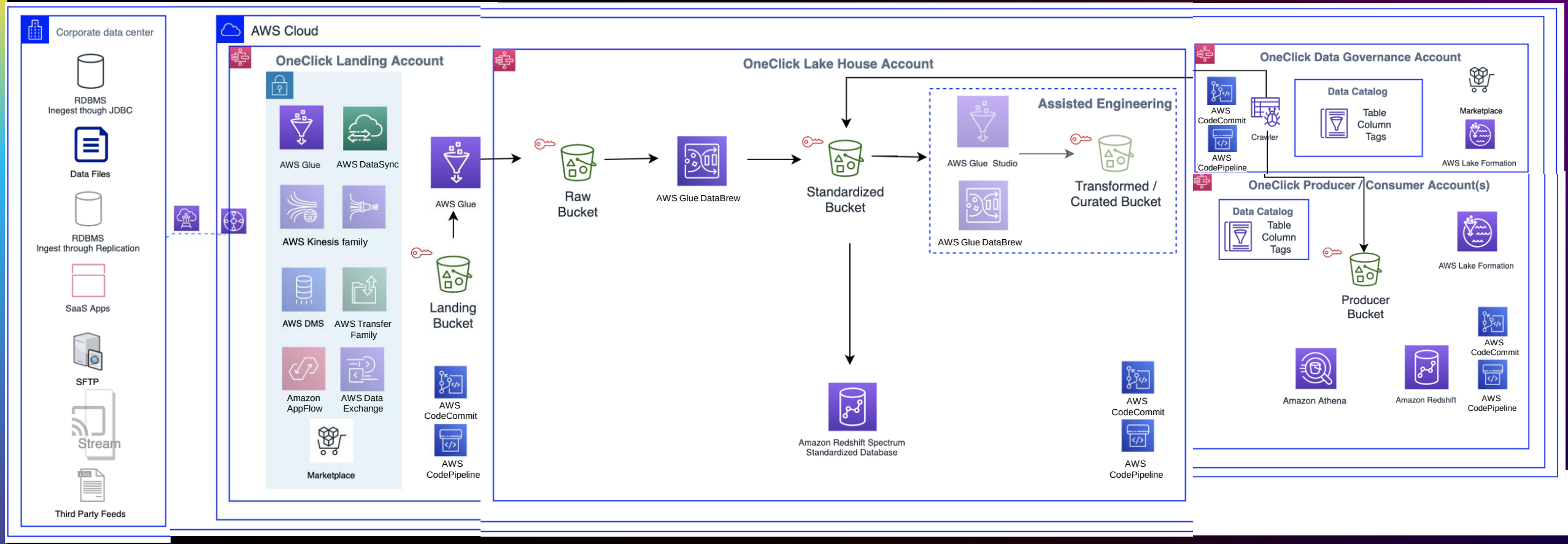
- Cloud Infra Architect
- Security Consultant

Prudential Financial

AWS

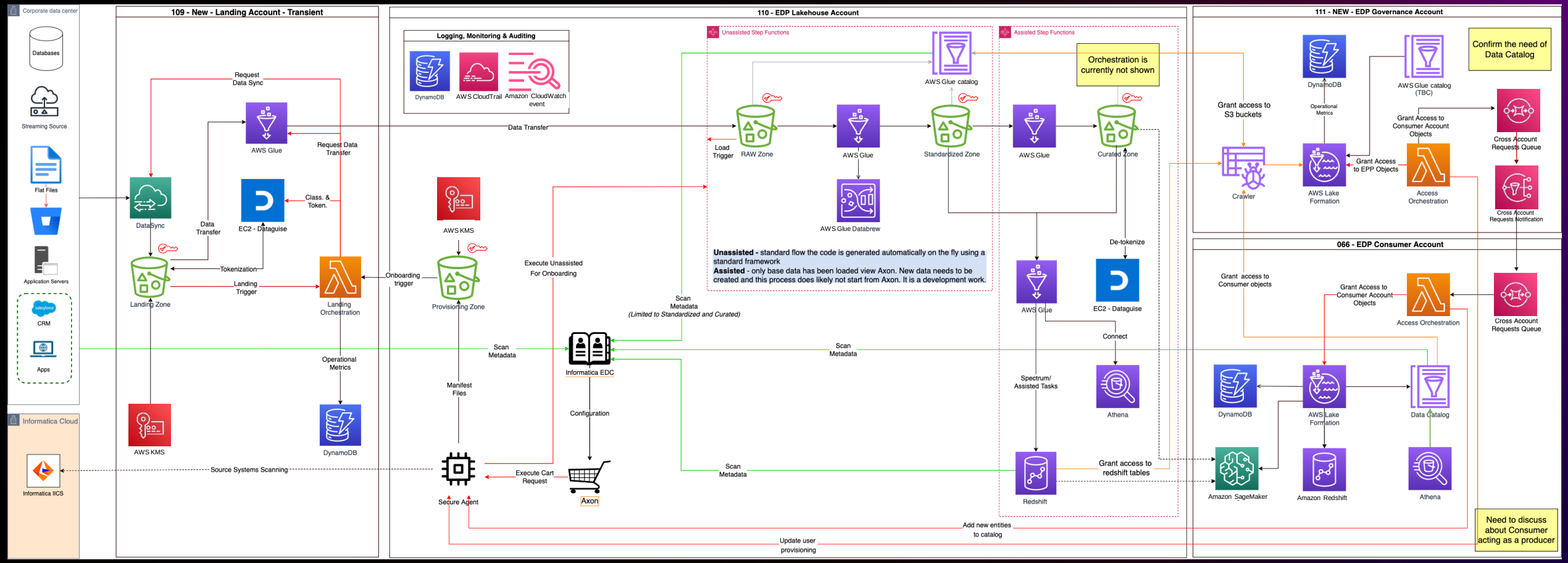
Architecture

THE PLATFORM IS SERVERLESS, MULTI-ACCOUNT, MULTI-TENANT, LOW-CODE/NO-CODE



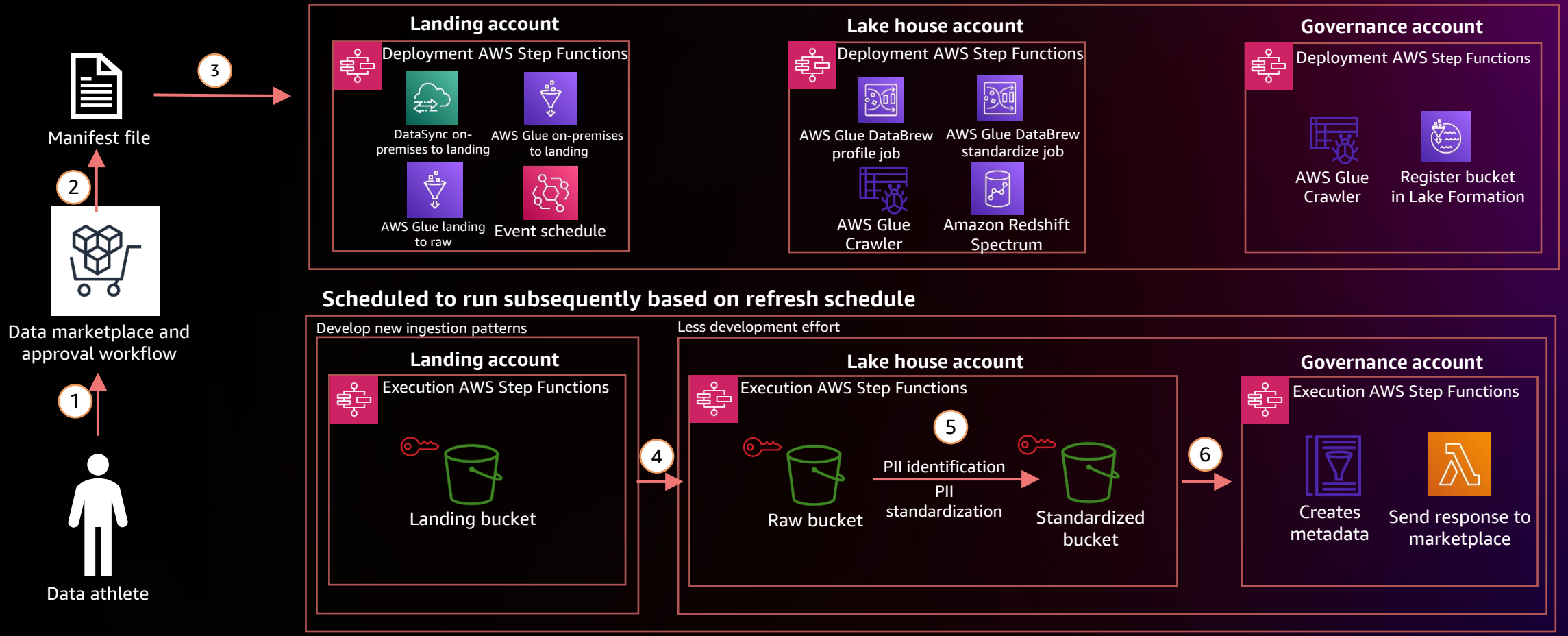
Architecture

THE PLATFORM IS SERVERLESS, MULTI-ACCOUNT, MULTI-TENANT, LOW-CODE/NO-CODE



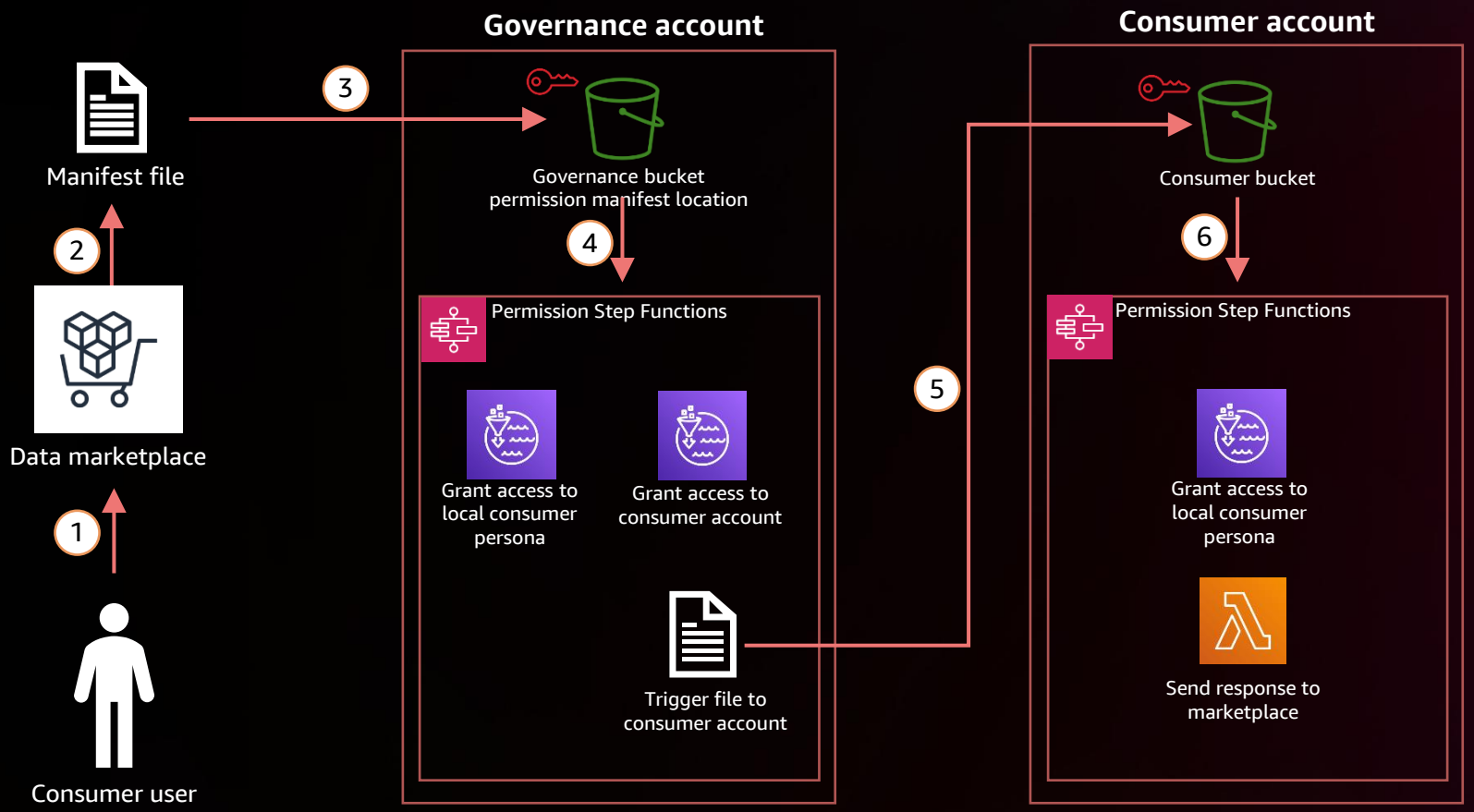
User experience – Data onboarding

ARCHITECTURE TO ON-BOARD DATA INSTANTLY FROM THE DATA MARKETPLACE



User experience – Data lake access

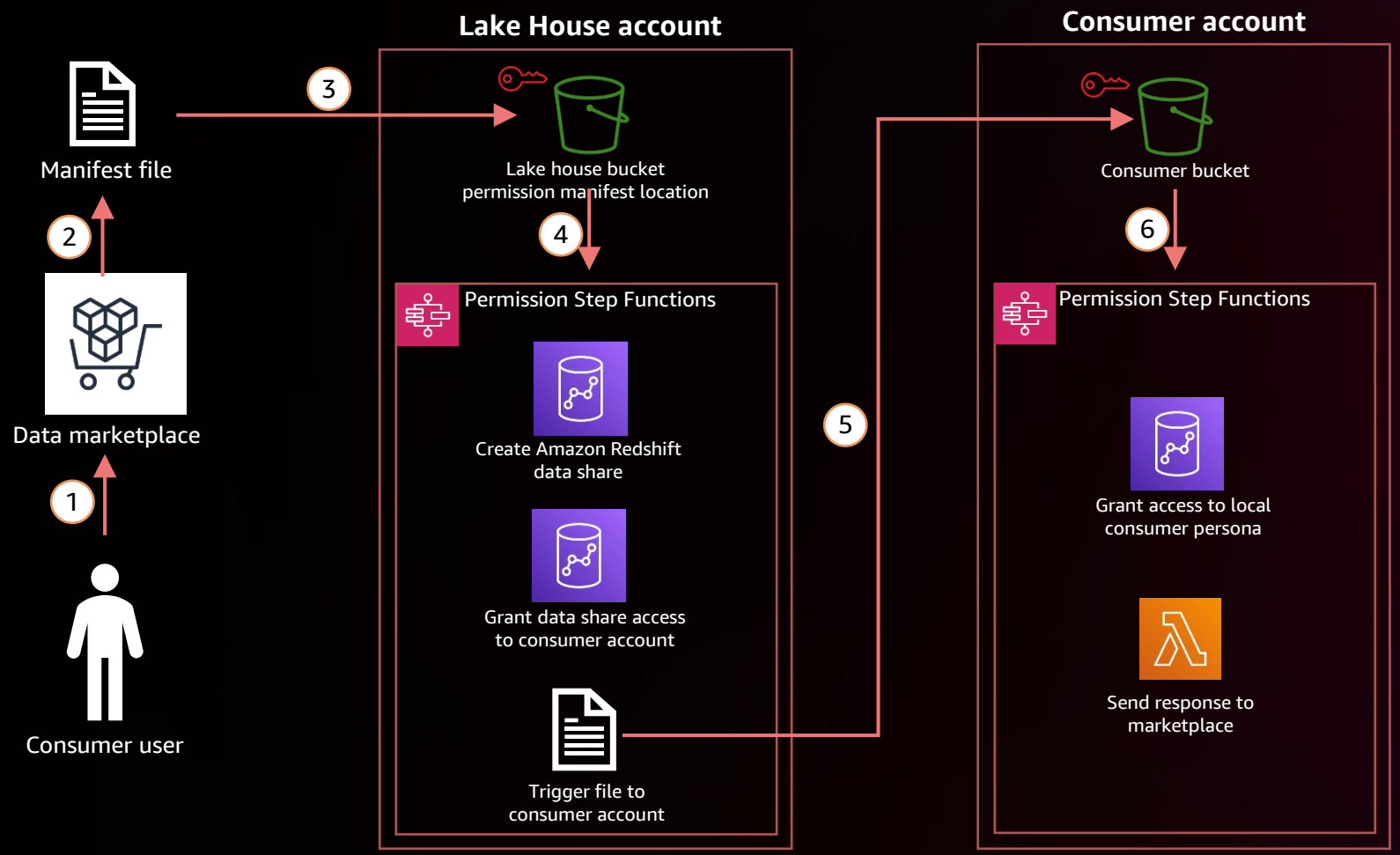
ARCHITECTURE ENABLES INSTANT DATA SHARING TO MULTIPLE CONSUMERS



1. Consumer browses data marketplace and requests access to data lake asset(s)
2. Once the request is approved, a JSON manifest file is created
3. Manifest file(s) is pushed to S3 bucket, which initiates access provisioning pipeline
4. Access is provisioned to the consumer if the consumer is user of the governance account
5. If the consumer is a user of consumer account then the cross-account grants are provisioned
6. Access is provisioned to the consumer in the consumer account

User experience – Data warehouse access

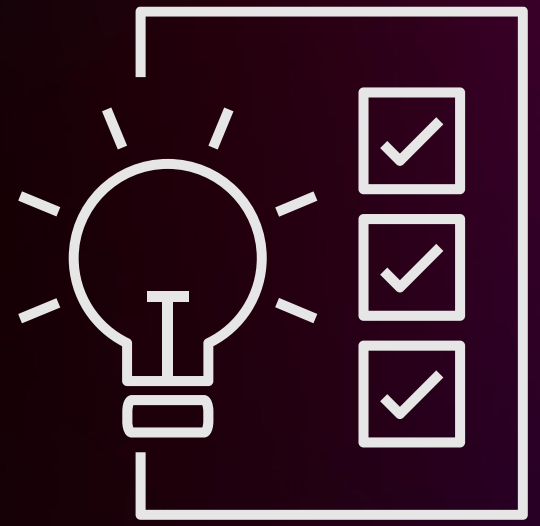
ARCHITECTURE ENABLES INSTANT DATA SHARING TO MULTIPLE CONSUMERS



1. Consumer browses data marketplace and requests access to a data warehouse asset(s)
2. Once the request is approved, a JSON manifest file is created
3. Manifest file is pushed to the S3 bucket, which initiates access provisioning pipeline
4. Amazon Redshift data share is created and shared with the consumer account
5. Trigger file is pushed to the consumer account, which kicks off access provisioning pipeline in the consumer account
6. Access is provisioned to the consumer in the consumer account

Key lessons learned

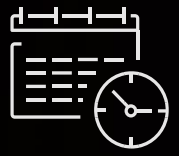
- ✓ Clear vision of end state
- ✓ Automate governance at every step
- ✓ Stay focused on your users
- ✓ Small but talented team
- ✓ Continuous learning
- ✓ Value in trying, value in adjusting
- ✓ Build what you can support



Our business outcomes

Inclusion

- Increased talent pool via lowered entry barrier by 900%



Time savings

- Cut discovery time by 99.75% (400 hours to 60 minutes)
- Cut data project development time by 88.9% (90 days to 10 days)
- Decreased data access time by 90% (10 days to 1 day)

Cost savings

- Decrease costs of bespoke data platforms by 98% (50 to 1)
- Decrease cost of education by 93.3% (15 skills to 1)
- Decrease cost of development by 98% (500K to 10K)

Governance

- Improved data quality from built-in governance
- Improved visibility from data catalog
- Reduced data sprawl via integrated platform

How do I get started?

AWS programs that support data governance

Want to build a data vision and strategy?

aws data-driven everything

- Joint engagements with business and technology stakeholders
- Create an organizational vision for innovation with data to drive business outcomes
- Define the first pilot, learn, and build

Gain business and IT strategic alignment

Have a strategy and need help executing it?

aws data lab

- Joint engineering engagements between customers and AWS
- Create tangible deliverables to accelerate your initiatives
- Delivers an architecture, production ready prototype, and upskilling on AWS services

Come with an idea, leave with a solution

Need help from strategy to implementation?

aws ProServe

- Assess your needs; align data and business strategies
- Work closely with AWS experts to build your data governance framework
- Implement at scale to drive business outcomes

Build your data governance framework

aws

Getting started: Next steps

THINK BIG

Discovery Workshop

Data-Driven Everything

START SMALL

Data Labs

AWS ProServe POC

SCALE FAST

AWS ProServe

AWS Partners

Thank you!

Jason Berkowitz
jberkowi@amazon.com

Shihas Vamanjoor
LinkedIn: shihasvamanjoor

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