aws INNOVATE
DATA EDITION

23 August, 2022

Creating a modern data strategy to enable data sharing in regulated environment

Blair Layton

Principal Business Development Manager
Public Sector, APJ
Amazon Web Services

aws

Agenda

- What is driving change in data strategies?

- Challenges of centralized data lakes in complex organizations

- Data sharing – A population health example

  - ➤ Hospital, hospital clusters, state and national government requirements

- End result - A Data Mesh

- Resources

What is driving change in data strategies?

Customers want more value from their data

GROWING EXPONENTIALLY

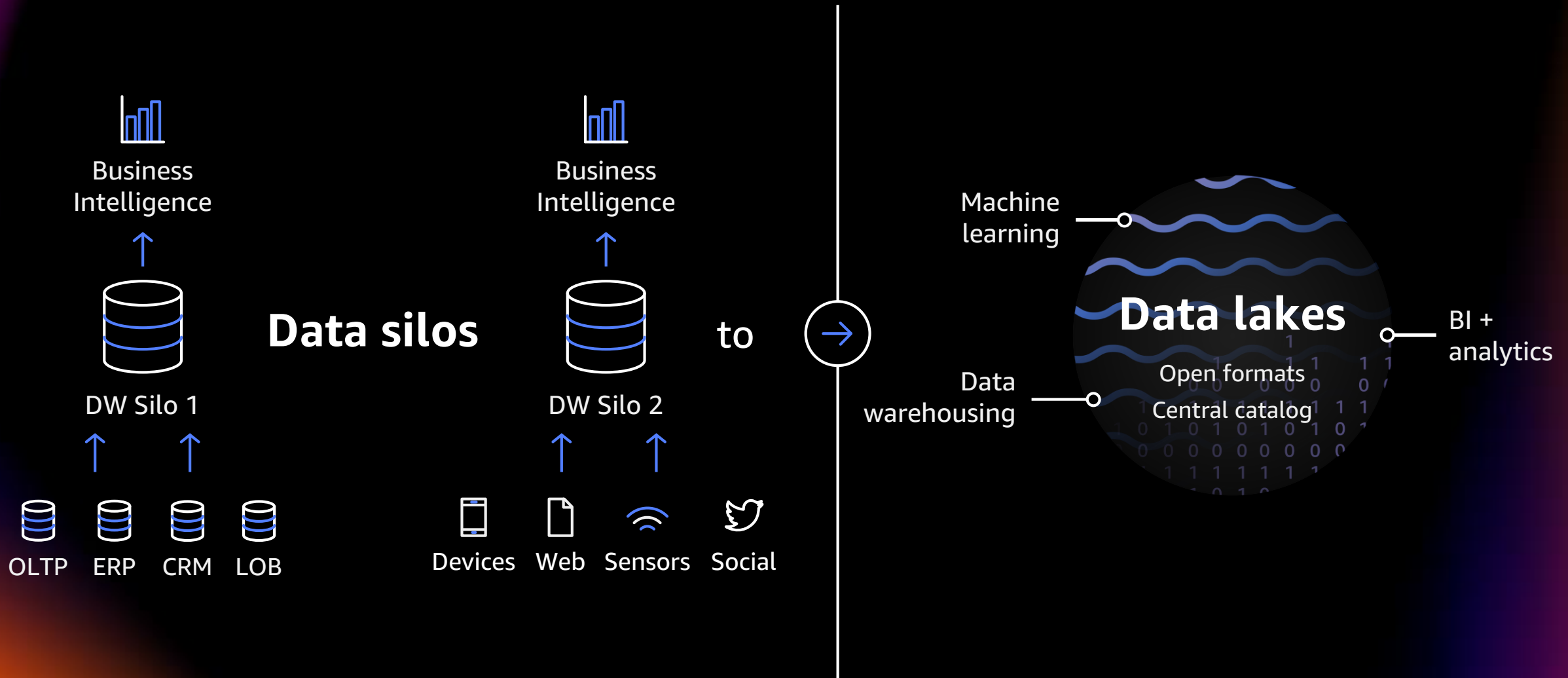**FROM NEW SOURCES**

INCREASINGLY DIVERSE
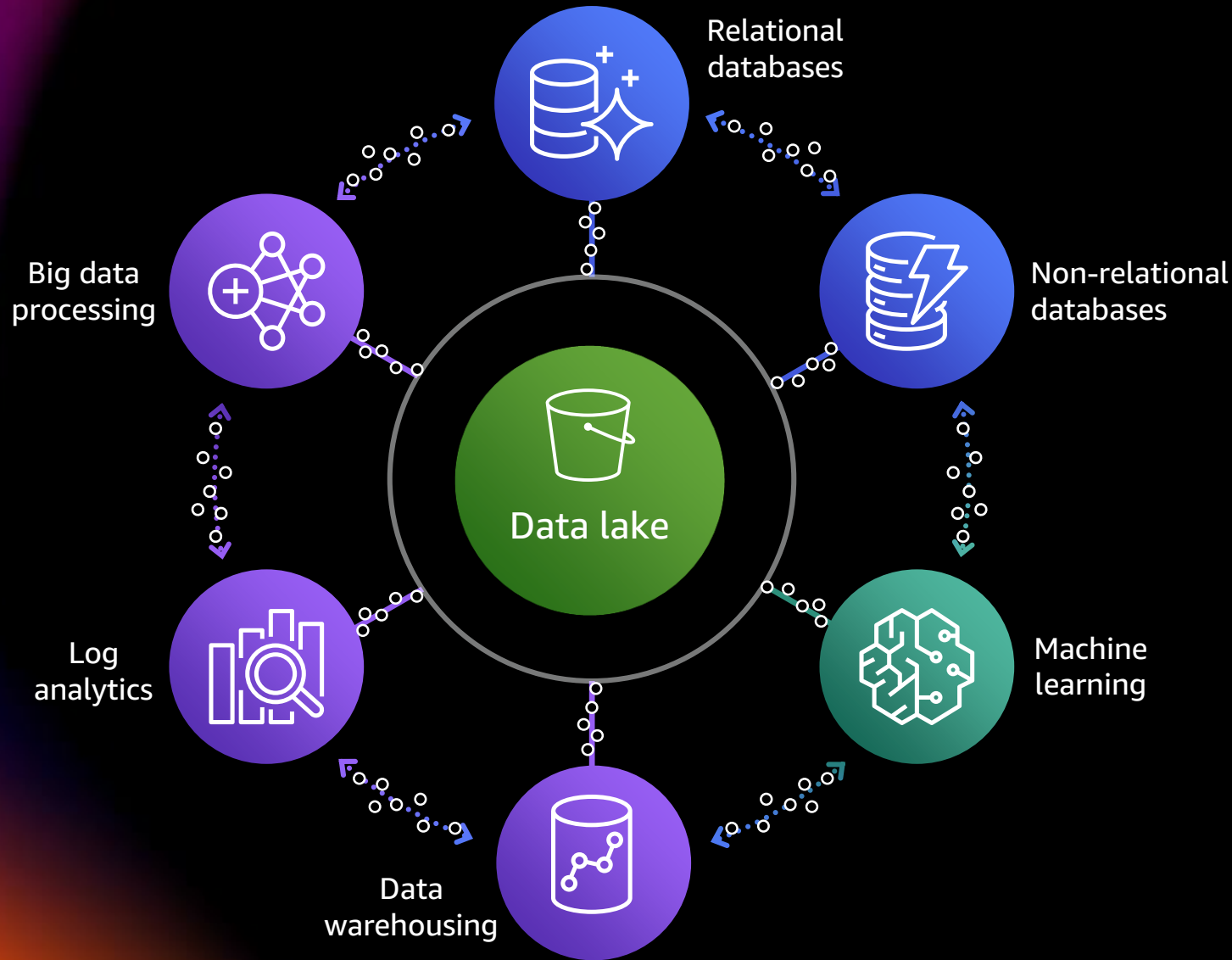
USED BY MANY PEOPLE

ANALYZED BY MANY APPLICATIONS

aws

Traditional data warehousing approaches don't scale

Business Intelligence

Business Intelligence

Data silos

DW Silo 1

DW Silo 2

to

OLTP ERP CRM LOB

Devices Web Sensors Social

Machine learning

Data lakes

BI + analytics

Data warehousing

Open formats
Central catalog

Modern data architecture

Relational databases

Non-relational databases

Machine learning

Data warehousing

Log analytics

Big data processing

Data lake

SCALABLE DATA LAKES

PURPOSE-BUILT DATA SERVICES

SEAMLESS DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

aws

Common data lake challenges we hear from customers

"Sharing and searching data is difficult"

"I just want to get access to the data I need"

"Difficult to meet all requirements across differing business units"

"I wish to focus on innovating with data, not on maintaining and administering a data lake"

"My data science team should easily find the datasets they seek and have the ability to share them with others"

"My team needs to own datasets, pipelines and repositories that are isolated from other teams"

"Why doesn't our organization treat data as a product?"

"Current data architecture is complex and monolithic and slow to change"

"If I share data, I've lost control"

"There is a mis-match between executive leadership goals and business line deliverables and incentives"
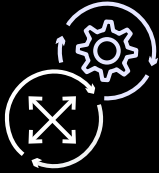
"Our internal policies on what can be shared are unclear and there is lack of incentive to share"

"Need to create a model to support sharing from both producers and consumers of data"

What if I can democratize and share data?

Improve operational efficiency

Make more informed decisions

Accelerate innovation

Healthcare

- Population health refers to the health status and health outcomes within a group of people rather than considering the health of one person at a time. For public health practitioners, improving population health involves understanding and optimizing the health of a population broadly defined by geography.

- Sharing data for vaccine information from a population health perspective in Australia has the following structure:

  - Australian Government Department of Health

  - State/Territory Department of Health

  - Local district hospital clusters

  - Hospitals

- Then there are GPs and pharmacies that can deliver vaccines outside of government facilities
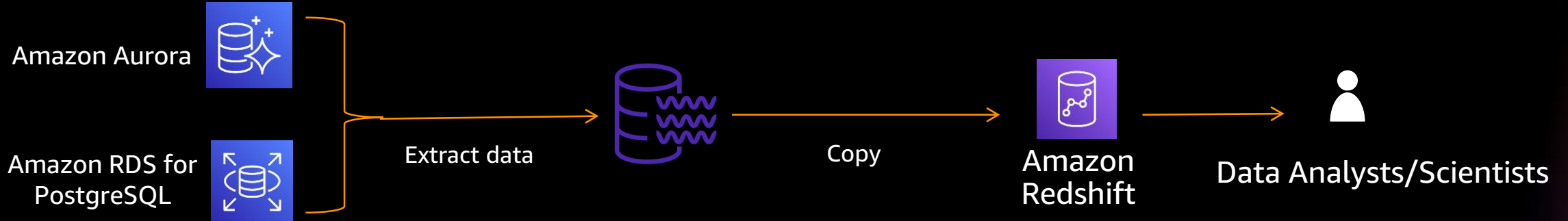
Sharing data in a hospital

Sharing data in a hospital

- Sharing of data from operational systems

- Easy access for analytics requirements

- Up-to-date information

- Minimal ETL required to solve business outcomes

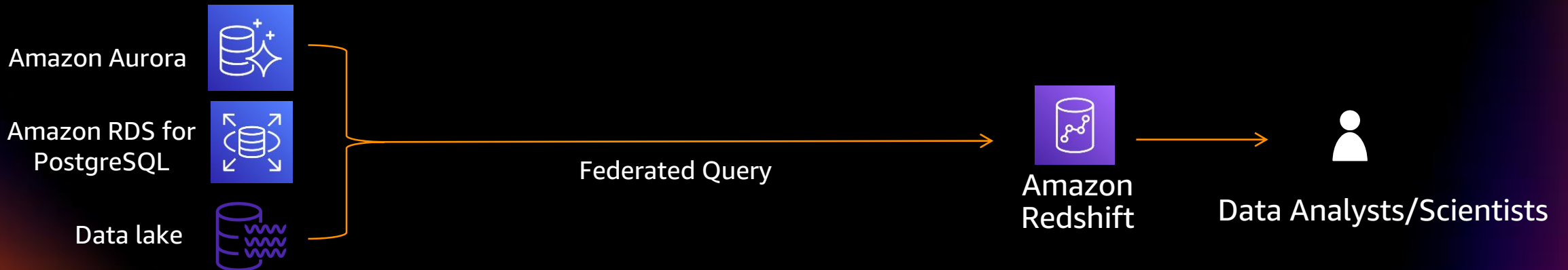- Could require a data warehouse or basic data lake

Pattern: Query across data sources

Query across operational, warehouse, data lake; No ETL required

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

Amazon Aurora

Amazon RDS for PostgreSQL

Extract data

Copy

Amazon Redshift

Data Analysts/Scientists

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

Amazon Aurora

Amazon RDS for PostgreSQL

Data lake

Federated Query

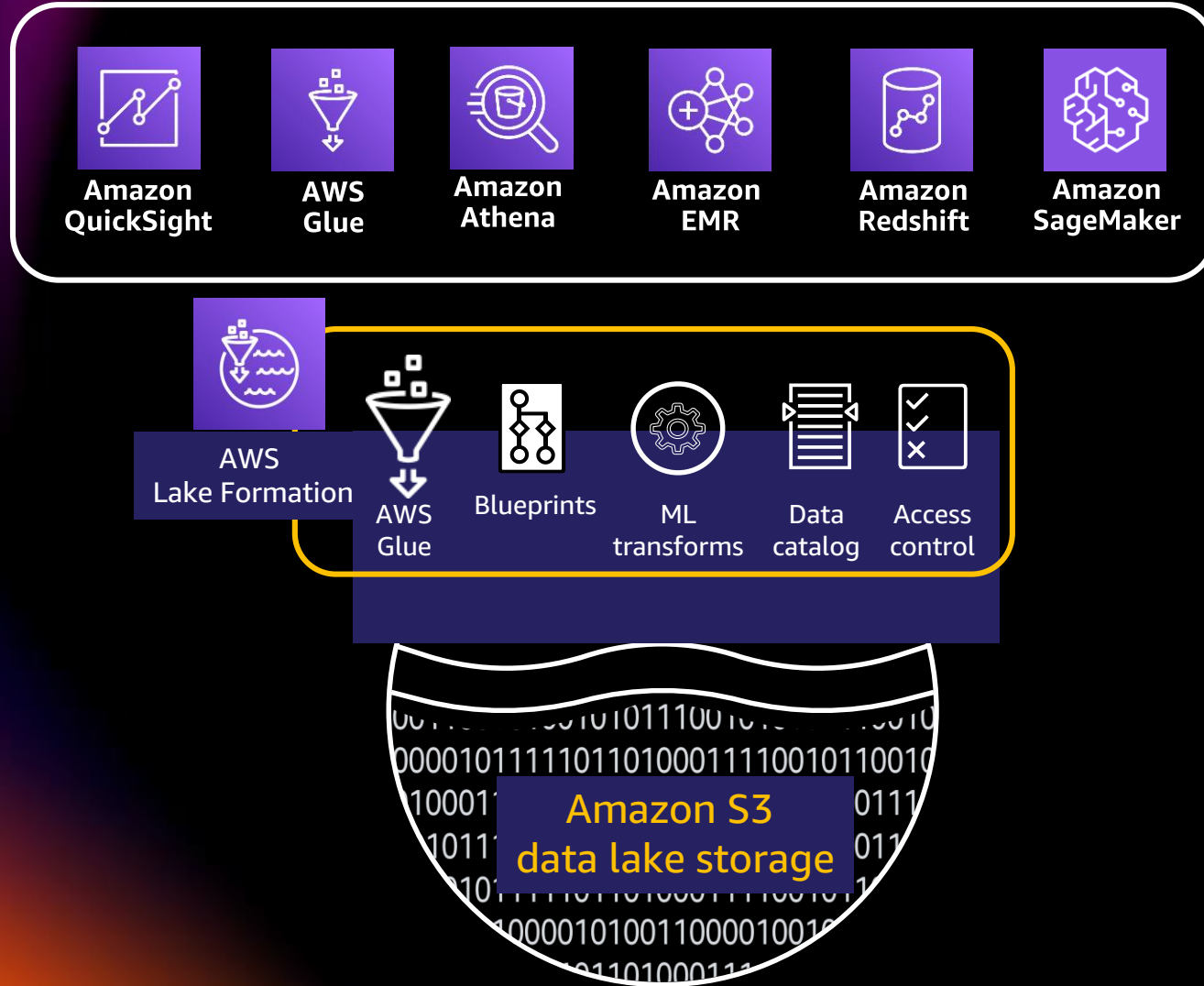Amazon Redshift

Data Analysts/Scientists

Sharing data – Hospital cluster

Sharing data – Hospital cluster

- Need to combine data from multiple hospitals

- Data pipelines will most likely already exist

- Need to ensure data is secure

- Consumers only get access to the data they need

- Need to share datasets with the state's Department of Health

- Requires a data lake with good security and governance

AWS Lake Formation: The foundation for data lakes

Amazon QuickSight
AWS Glue
Amazon Athena
Amazon EMR
Amazon Redshift
Amazon SageMaker

AWS Lake Formation

AWS Glue
Blueprints
ML transforms
Data catalog
Access control

Amazon S3 data lake storage

Single place to manage access

Open file formats

Efficient sharing

Ecosystem of integrated tools

Cost effective

aws

AWS Lake Formation permissions model

Amazon QuickSight · AWS Glue · Amazon Athena · Amazon EMR · Amazon Redshift · Amazon SageMaker · 3rd party app

AWS Lake Formation

Permissions

Data Catalog

Amazon S3 data lake storage
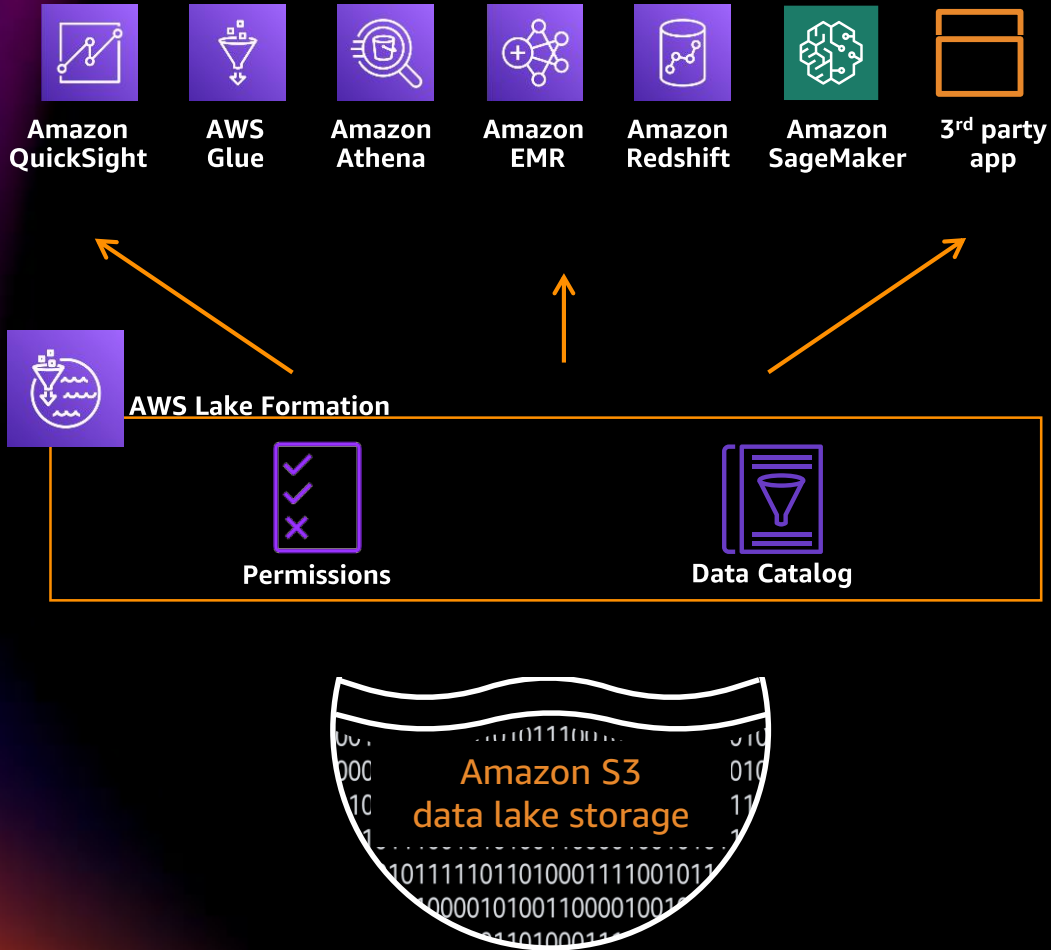
DB style fine-grained permissions

Fine-grained permissions on catalog resources

Amazon 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

AWS Lake Formation permissions

Columns

Rows

Cells

Specify include or exclude list

Specify row filter with PartiQL

Combine column and row filters

Data governance

Data governance includes:

processes,
roles,
policies,
standards and
metrics

that drive accurate and meaningful use of information enabling an organization to achieve its goals.

- Processes and policies underpin the quality and security of the data

- Clearly define the actor that can take what action upon what data, in which situations using what methods

- Business drivers highlight what data needs to be carefully controlled in your data governance strategy and the benefits expected from this effort

- Roles related to data are clearly defined to ensure that responsibility and accountability are agreed upon across the enterprise

Key data governance outcomes:

✓ Create an intelligent data governance plan

✓ Automate policies, standards, and processes

✓ Design for data quality and model for information relevance

✓ Support data governance for legal activities and security

✓ Ensure continuous compliance

✓ Enable your business with value-based data

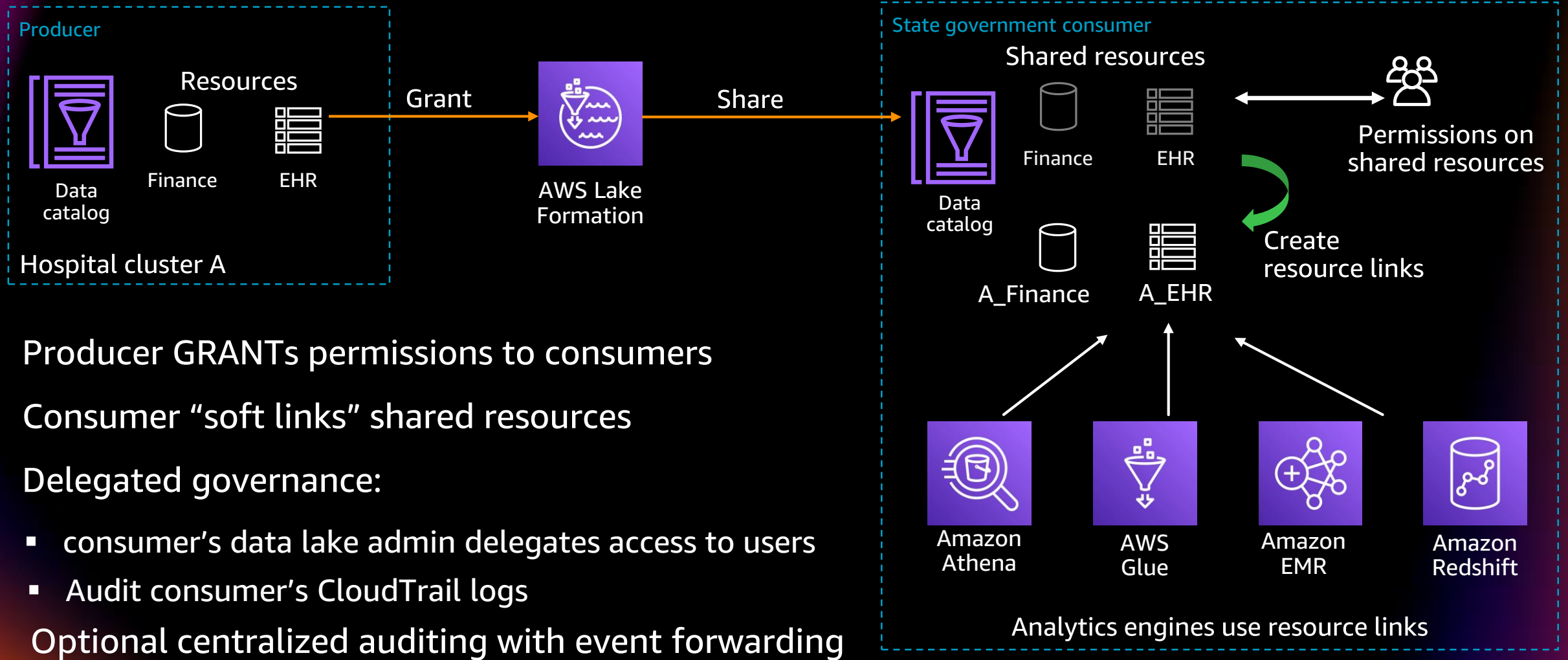## Accelerate compliance leveraging cloud:
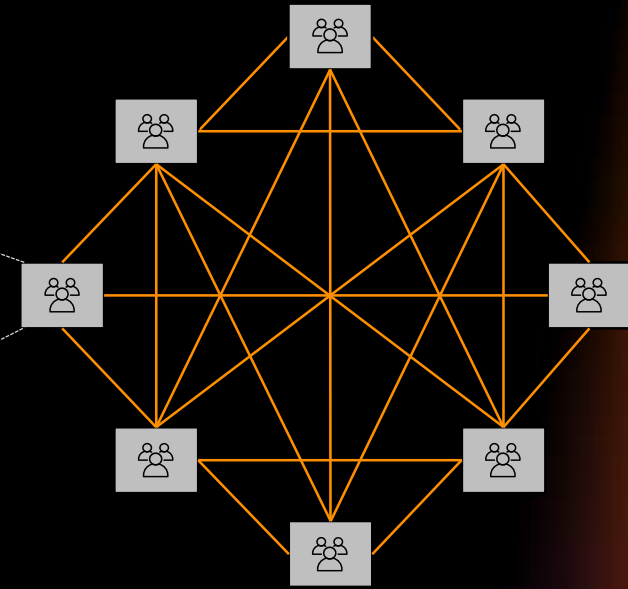

aws
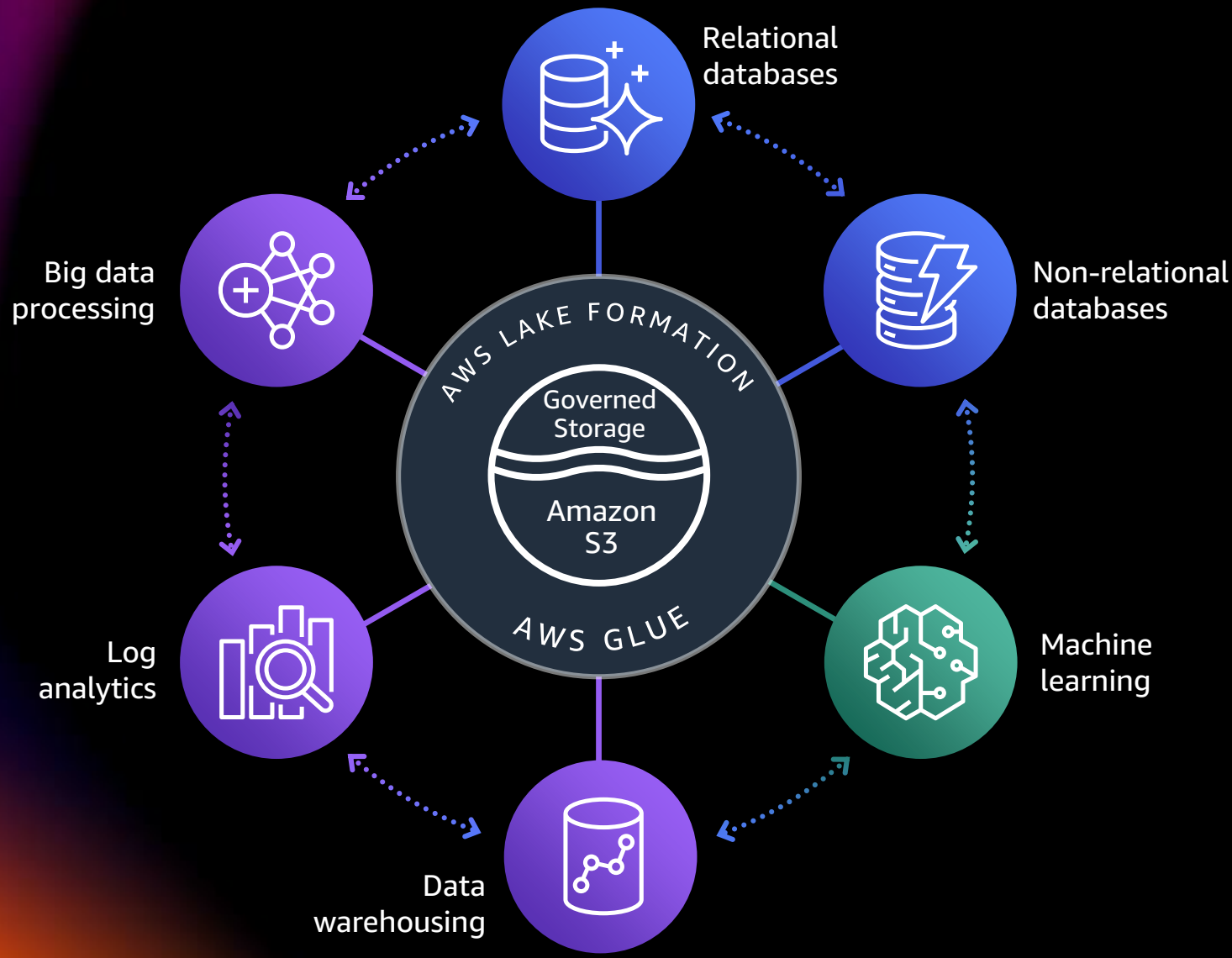
Sharing health data – State level

Sharing health data – State level

- Combine data from multiple hospital districts/clusters

- Share data with other states and national government agencies

- Publish public data sets

- Need a simple and secure method to share data from producers (hospitals, GPs, pharmacies) and to consumers (other agencies)

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

Evolution to a Data Mesh on AWS

Relational databases

Non-relational databases

Big data processing

AWS LAKE FORMATION

Governed Storage

Amazon S3

AWS GLUE

Machine learning

Log analytics

Data warehousing

- Common tech stack
- Scalable, durable and available
- Secure and compliant
- Simple to manage, standard ops
- Common skillset / quicker ramp up
- Cost effective

aws

Data mesh pattern

A data mesh architecture empowers business units (organized into domains) to have high ownership and autonomy for the technologies they use, while providing technology that enforces data security policies both within and between domains through data sharing.

Key pillars of a data mesh

- Domain-oriented, decentralized data ownership and architecture
  - Organizational autonomy

- Creating data products with true data owners
  - Single-threaded owner of data, treated as a product

- A self-service data infrastructure platform
  - Simple to use data tools using a common infrastructure framework

- Federated computational governance
  - Virtualize access to data in a secure and governed way

AWS blog: Build a modern data architecture and data mesh pattern at scale using AWS Lake Formation tag-based access control

aws

# Sharing health data – National level

Sharing health data – National level

- Consolidating data from multiple states
- Sharing data to other national agencies
- Publishing national public data sets
- Charging for specific and customized data requests

Amazon Redshift data sharing enables consistent data access for all the consumers

Simple and direct way to share data
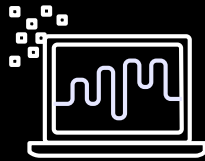
Instant, granular, and high-performance

Live and transactionally consistent views

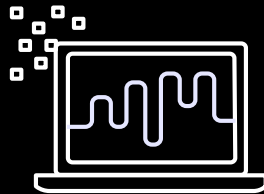**Secure and governed collaboration**

Monetize data products with AWS Data Exchange for Amazon Redshift

Data subscribers

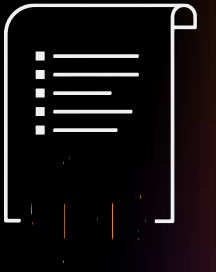Query data in minutes, no ETL

Always use the latest data

Data providers

Automate data entitlements, billing, and payment management

Multiple data delivery methods on a single product

APIs

DBs
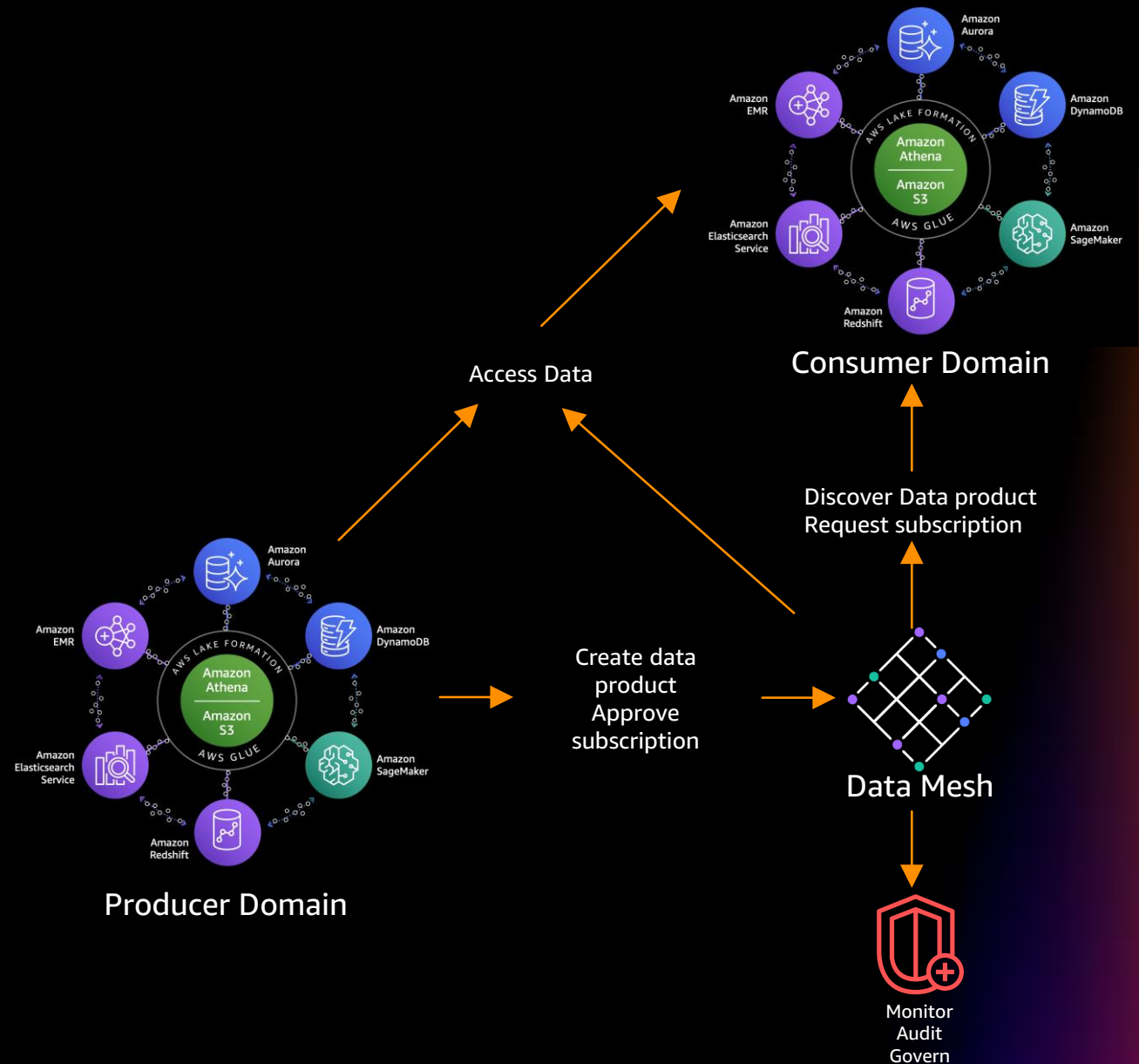
Files

One subscription

The end result

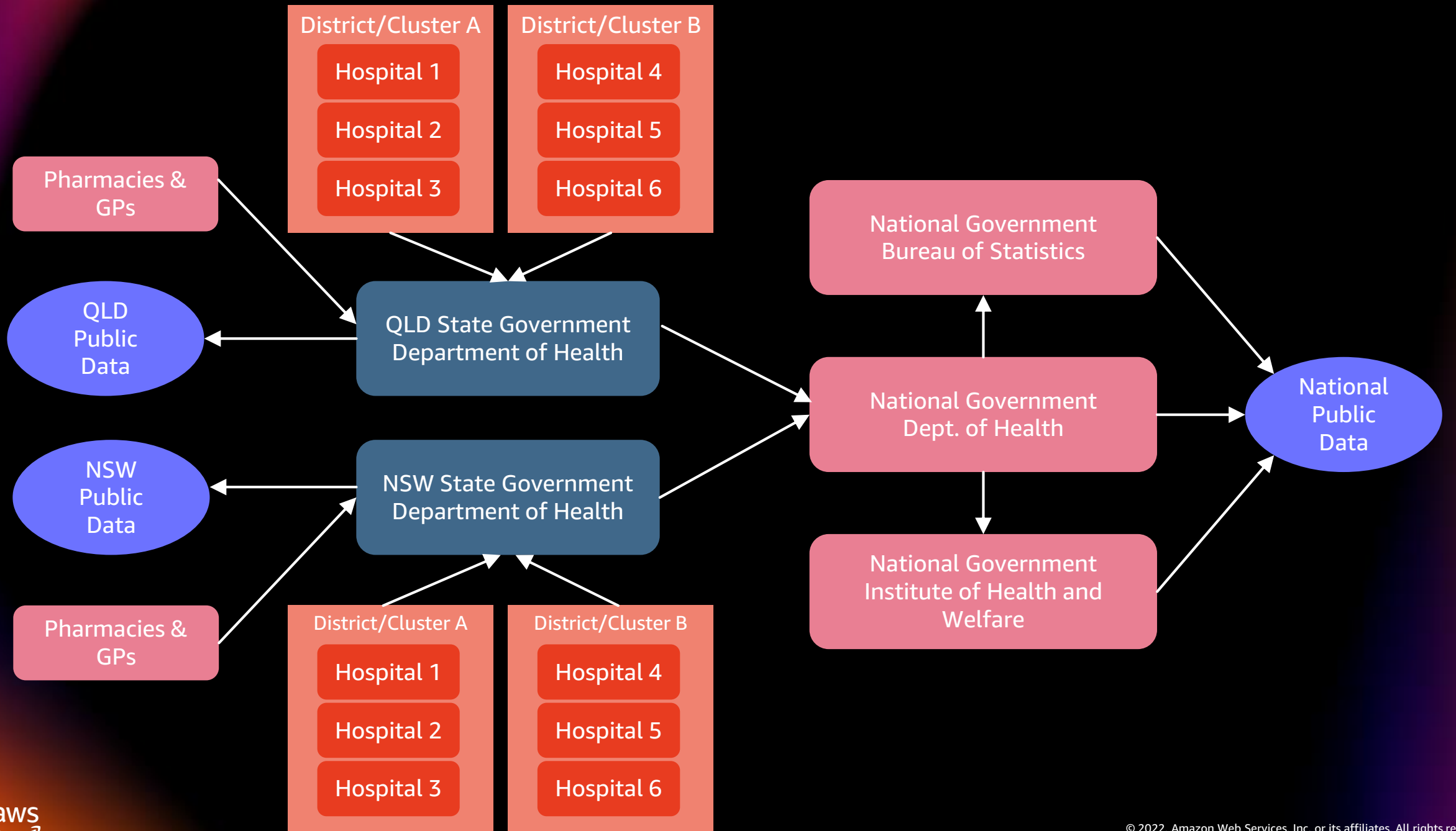With a Data Mesh Architecture, we aim to:
- **Accelerate** the **independent** delivery of business domain data lakes
- Increase data security and compliance within domains
- Enable **self-service** data product creation, discovery, and subscription
- Allow consumers to **transparently** access data products
- **Monitor, audit, and govern** access over time

…with high security
…with a broad and deep set of technology choices
…without high operational costs

Consumer Domain

Access Data

Producer Domain

Create data product Approve subscription

Discover Data product Request subscription

Data Mesh

Monitor Audit Govern

Visit the AWS Data resource hub

A modern data strategy can help you manage, act on, and react to your data so you can make better decisions, respond faster, and uncover new opportunities. Dive deeper with these resources today.

- Harness data to reinvent your organization
- In unpredictable times, a data strategy is key
- Make data a strategic asset
- Rewiring your culture to be data-driven
- Put your data to work with a modern analytics approach
- … and more!

https://tinyurl.com/data-hub-aws

Visit resource hub

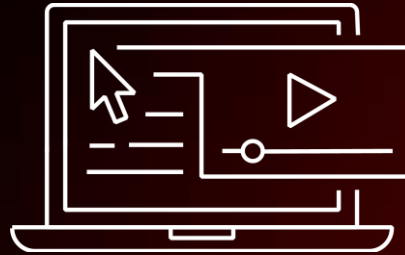# AWS Training and Certification for Data and Analytics

AWS Data & Analytics FREE Training Resources

Discover how to harness data, one of the world's most valuable resources, and innovate at scale.

https://bit.ly/3Ntlhy7

AWS Data Analytics Learning Plan

This learning plan expose you to the fastest way to get answers from all your data to all your users. It can also help prepare you for the AWS Certified Data Analytics - Specialty certification exam.

https://bit.ly/3wBVjD1

AWS Certified Data Analytics - Specialty

Earning AWS Certified Data Analytics – Specialty validates expertise in using AWS data lakes and analytics services.

https://go.aws/3lwF0RR

aws

Thank you for attending AWS Innovate – Data Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

Thank you!