AWS State, Local, and Education Learning Days

Philadelphia

aws Learning Days
State, Local, and Education

Building a Modern Data Strategy

Sid Joshi (he/him)

Solutions Architect
AWS
joshisj@amazon.com

Learning Days
State, Local, and Education

Agenda

- Why modern data architecture
- Modern data strategy
- Building Modern Data Architecture
- Reference architectures for common scenarios
- Getting started

Why modern data architecture

“If we have data, let’s look at the data.
If all we have are opinions, let’s go with mine”

Jim Barksdale

CEO of Netscape

Data is just the Building Blocks

Data ▶ Information ▶ Insights

Without structure, tools and processes,
Data has very little value

The data challenge

Availability of electronic data is growing exponentially

Data coming from new, disconnected sources

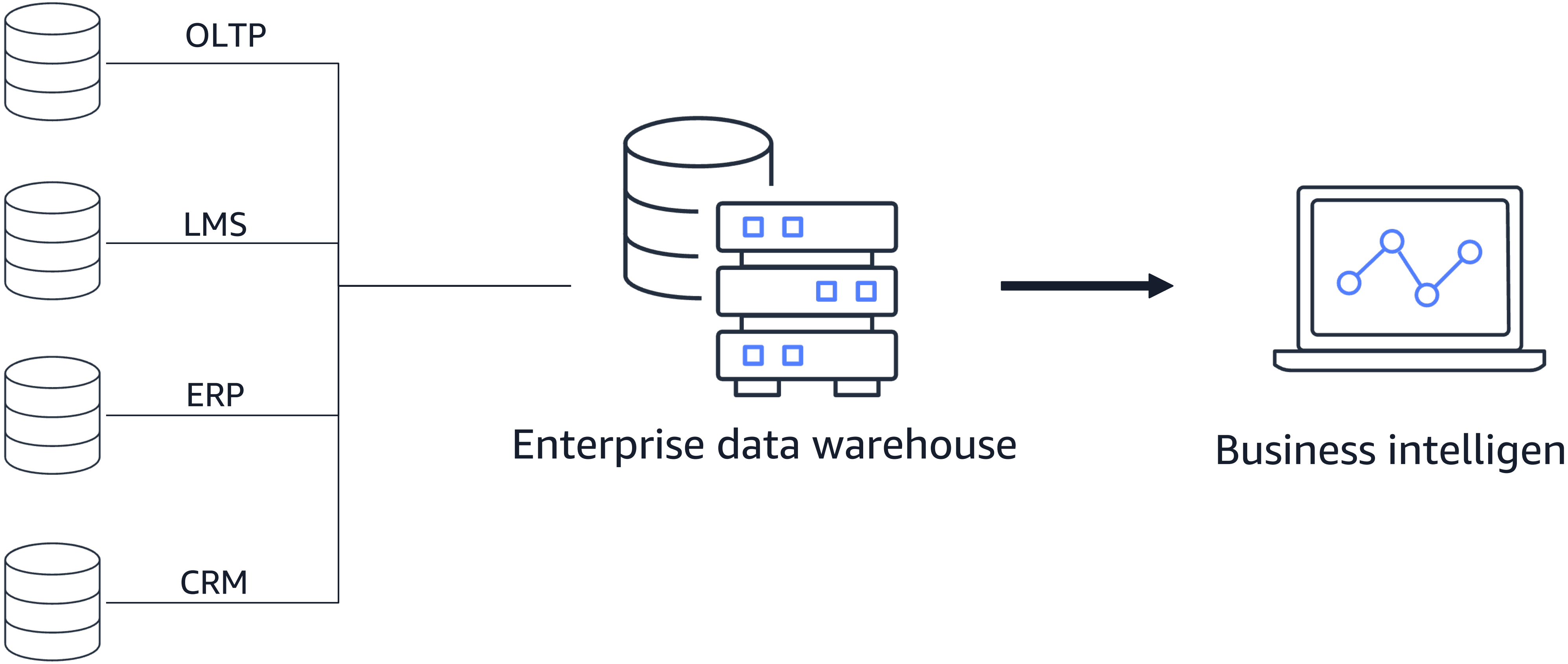
Increasingly diverse in file type and volume

Used by many people (e.g. policy makers, researchers, etc.)

Analyzed by many applications

Current state

Currently, decision-making revolve around the **enterprise data warehouse**



Data no longer scales

There is more data and more diversity of data than people think

Data growth

>10x

every 5 years

Data platforms need

To live for

15+

years

To scale

1,000x

Accessibility of data

Data scientists

Business users

Analysts

Applications

Machine learning

SQL analytics

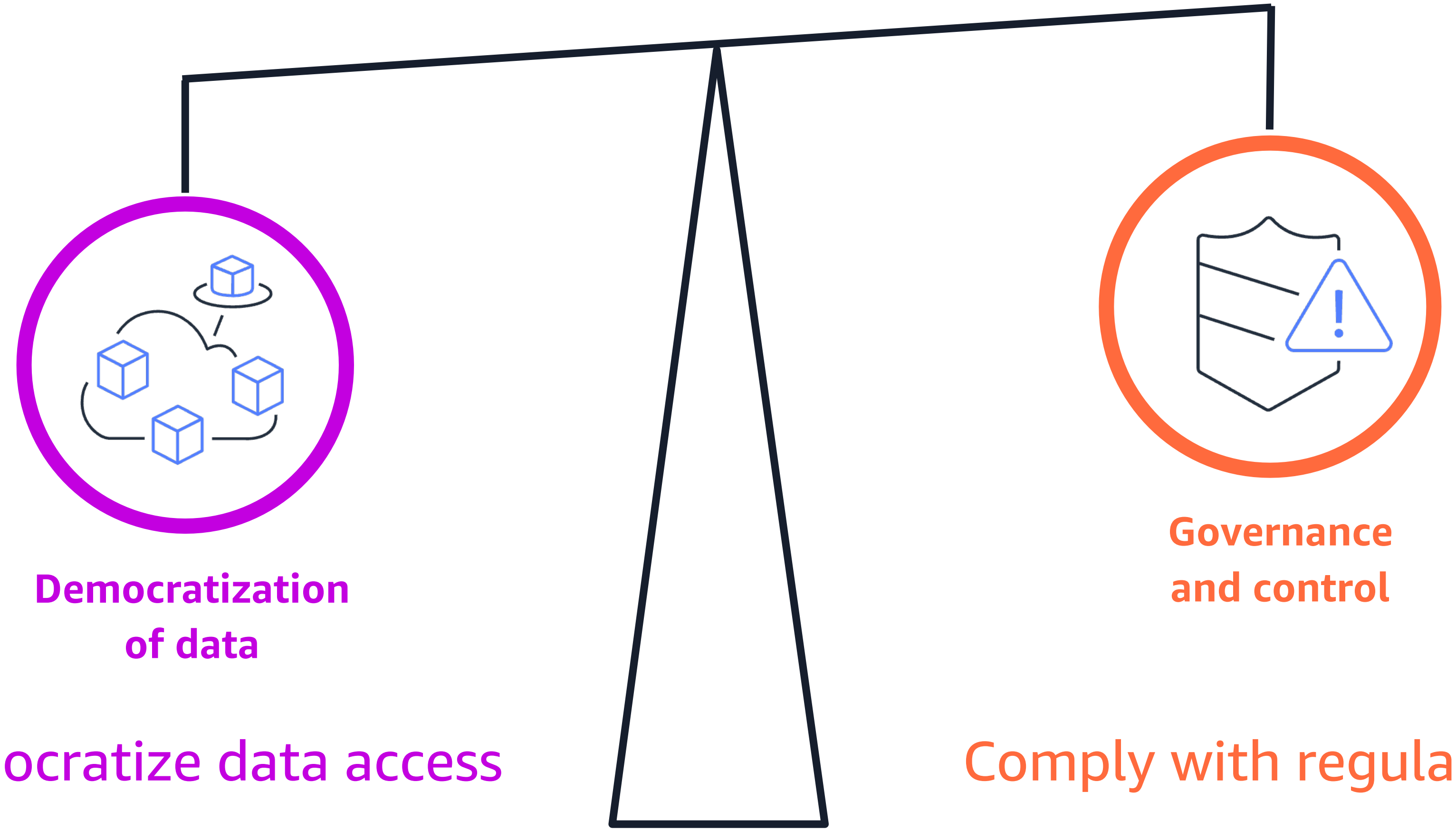
Scientific

Real-time, streaming

There are **more people** accessing data

And in **different ways**

More regulatory pressure



Democratization of data

Governance and control

Democratize data access

Comply with regulations

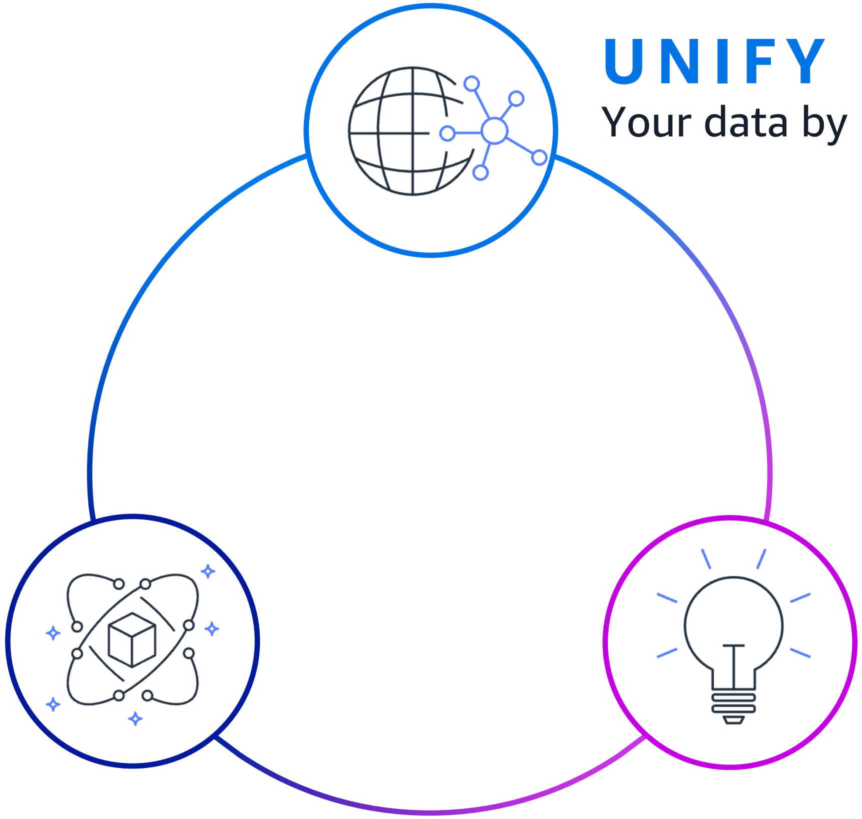
What now? Let’s rethink everything

Raw Data

Insights

Modern Data Strategy

Modern data strategy for better business outcomes



UNIFY

Your data by breaking down silos

MODERNIZE

Your data infrastructure to a scalable, trusted, and secure cloud provider

INNOVATE

By inventing new experiences and reimagining existing processes

Modernize

MODERNIZE DATA INFRASTRUCTURE FROM A LEGACY SOLUTION TO A SCALABLE, TRUSTED, AND SECURE CLOUD PROVIDER

- Reduce operational overhead with purpose-built, cloud-based databases
- Modernize analytics tools to handle structured, unstructured, and streaming data – at scale
- Standardize on a modern ML infrastructure to harness the ML benefits at scale

Unify

BREAK DOWN SILOS, SO DATA CAN BE PUT TO WORK ACROSS DATABASES, DATA LAKES, ANALYTICS, AND ML SERVICES

- Unify your data and make data accessible and shared in a secure way
- Ensure that data can easily get to wherever it’s needed, with the right controls
- Enable analysis and insights through analytics, visualization, and ML tools

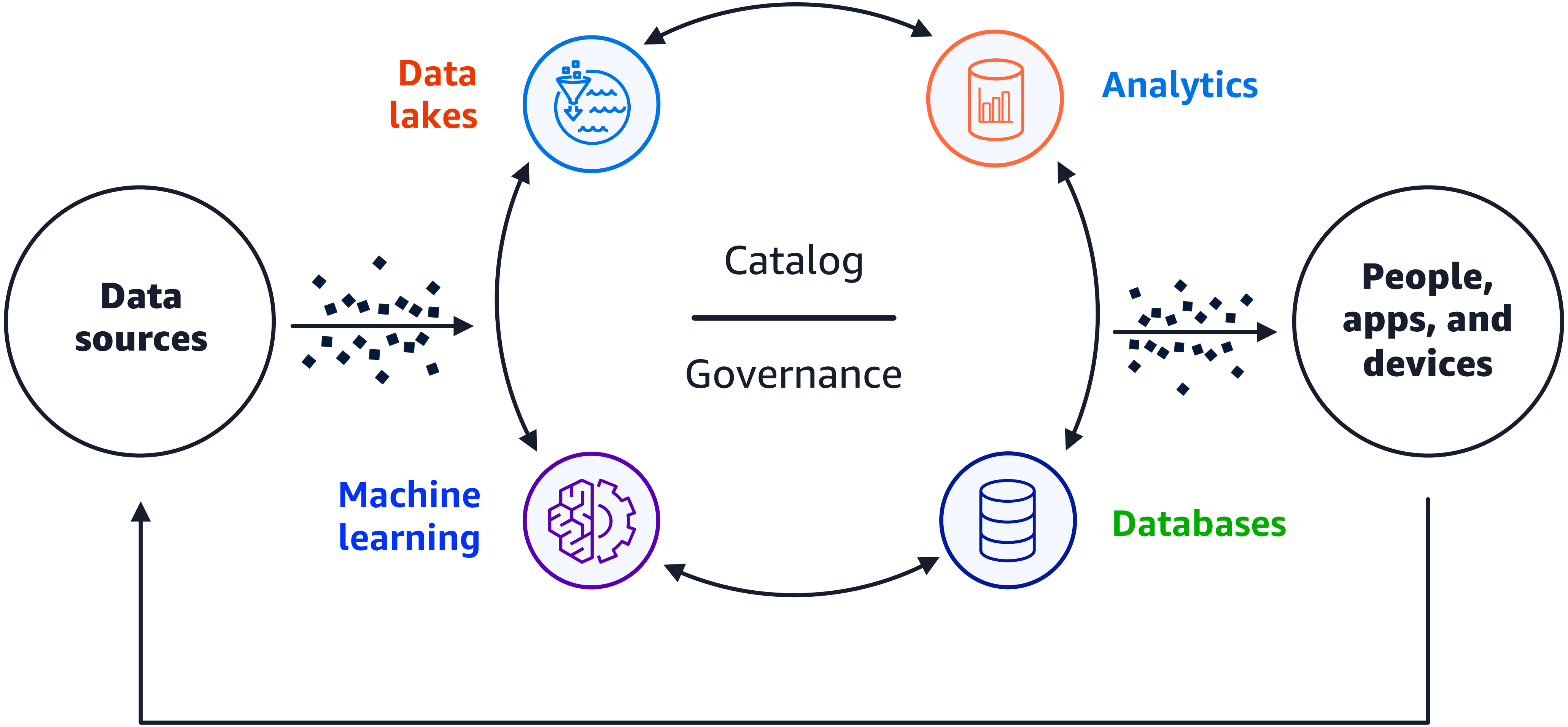
Innovate

INVENT NEW EXPERIENCES AND REIMAGINE PROCESSES WITH PURPOSE-BUILT DATABASES, ADVANCED ANALYTICS AND ML

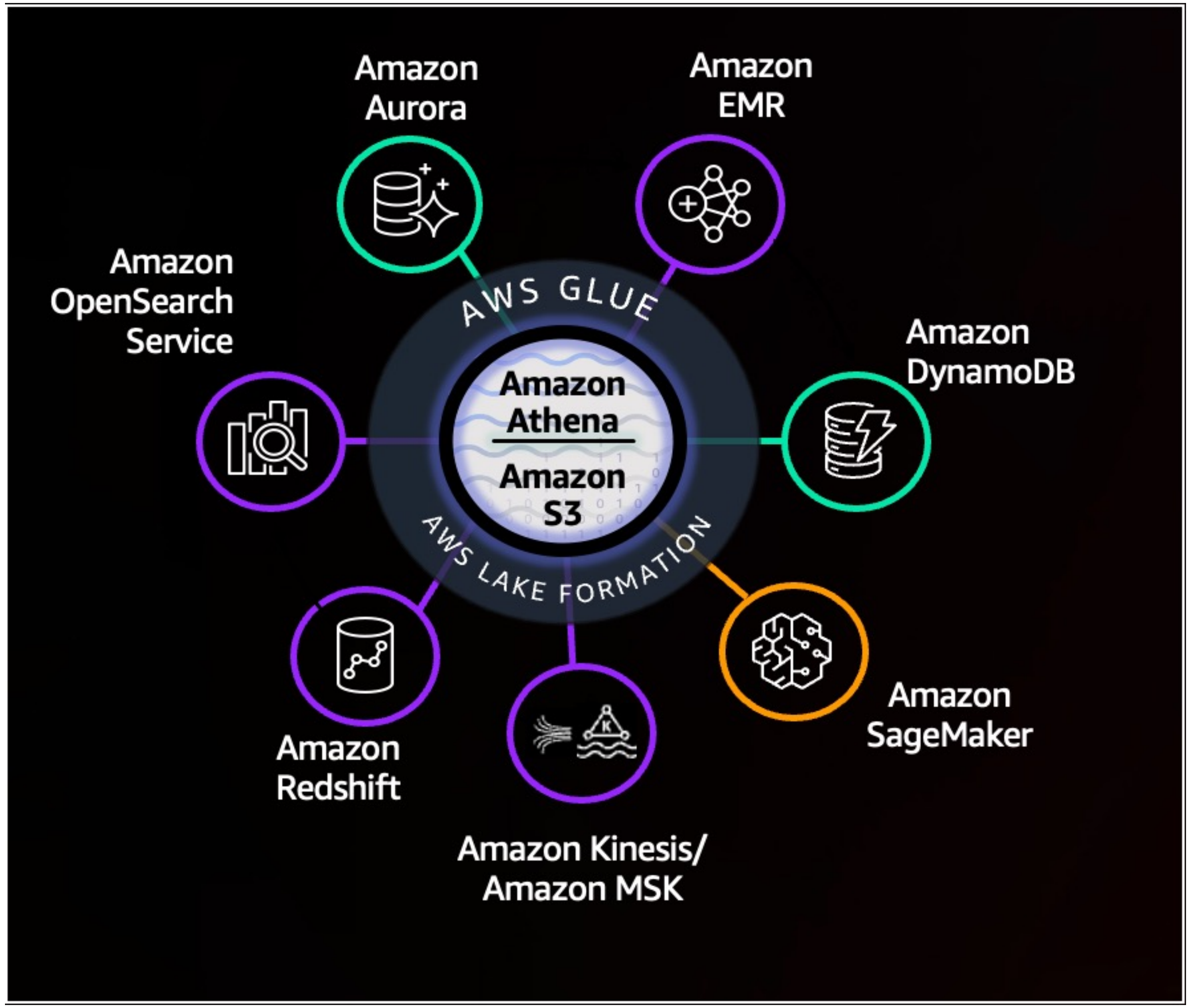
- As the types of data and workloads evolve, the databases, analytics tools, and ML services need to evolve
- ML is driving unprecedented levels of innovation
- Create better customer experiences with insights and predictions enabled by ML

Building modern data architectures

Modern data architecture



Modern data architecture on AWS



Modern data architecture pillars

Data at any scale

The best price performance

Seamless data access

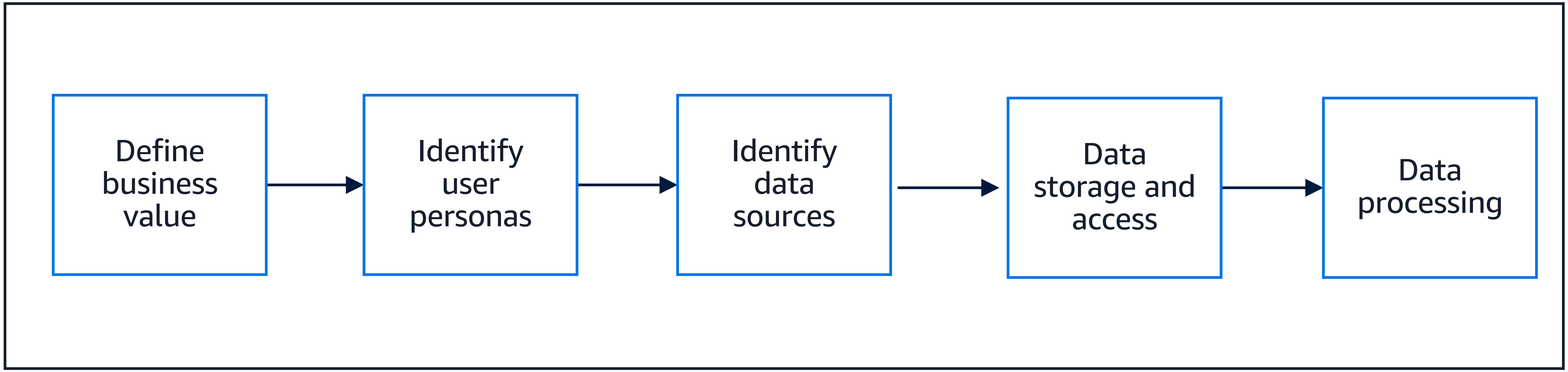
Unified governance

AI and ML to solve business challenges

Data discovery

FIRST STEP IN BUILDING MODERN DATA ARCHITECTURES

The data discovery process consists of a number of interactive sessions with various stakeholders within an organization



Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion

Bring the data into your data platform

2. Data storage

Store your structured and unstructured data

3. Data cataloging

Store your metadata

4. Data processing

Create data processing pipelines

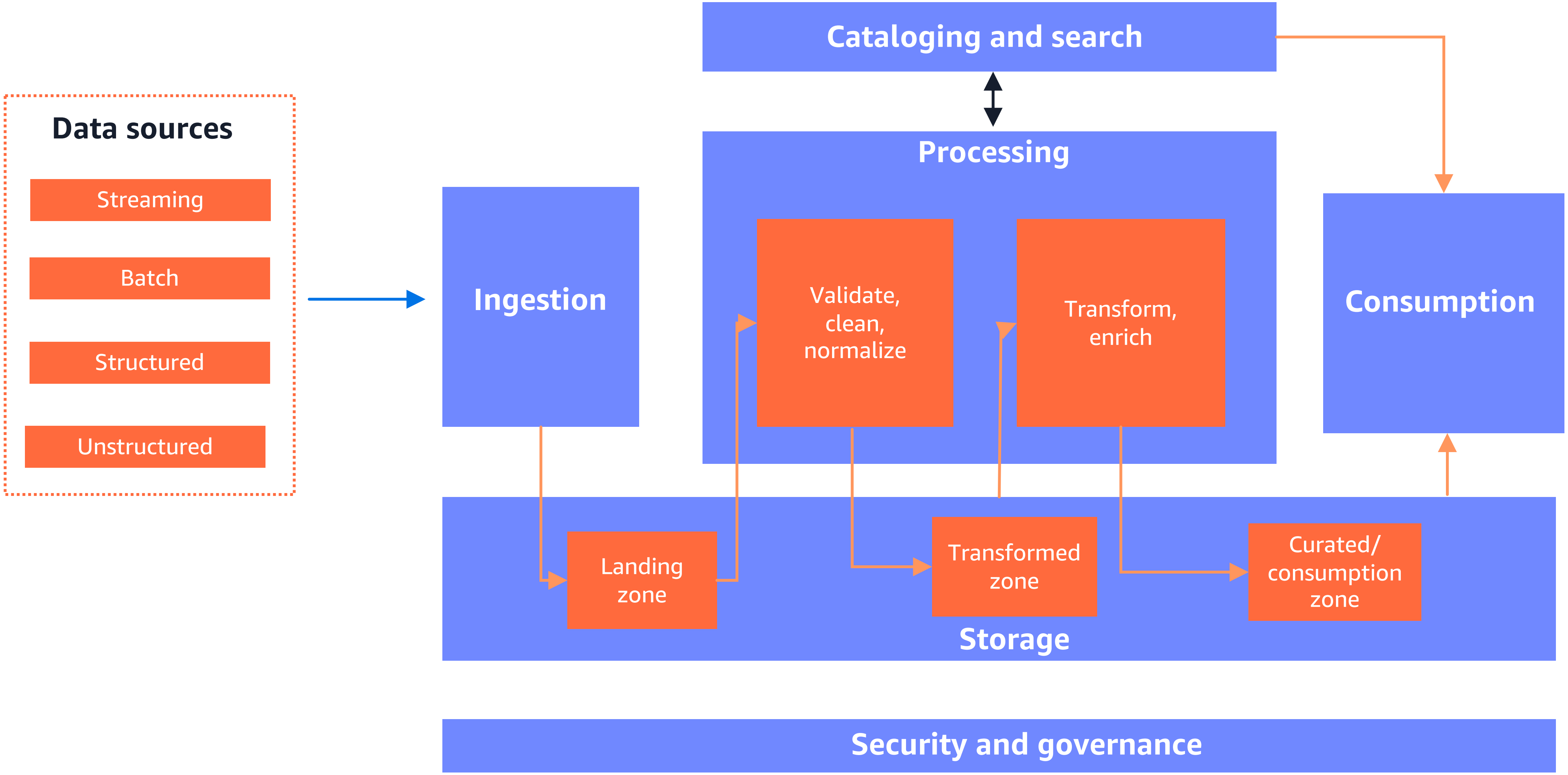
5. Data consumption

Enable your user personas for purpose-built analytics and machine learning

6. Security and governance

Protect your data across the layers and data access management

Layered modern data architecture



Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion	2. Data storage	3. Data cataloging	4. Data processing	5. Data consumption	6. Security and Governance
Bring the data into your data platform	Store your structured and unstructured data	Store your metadata	Create data processing pipelines	Enable your user personas for purpose-built analytics and machine learning	Protect your data across the layers and data access management

Data ingestion layer

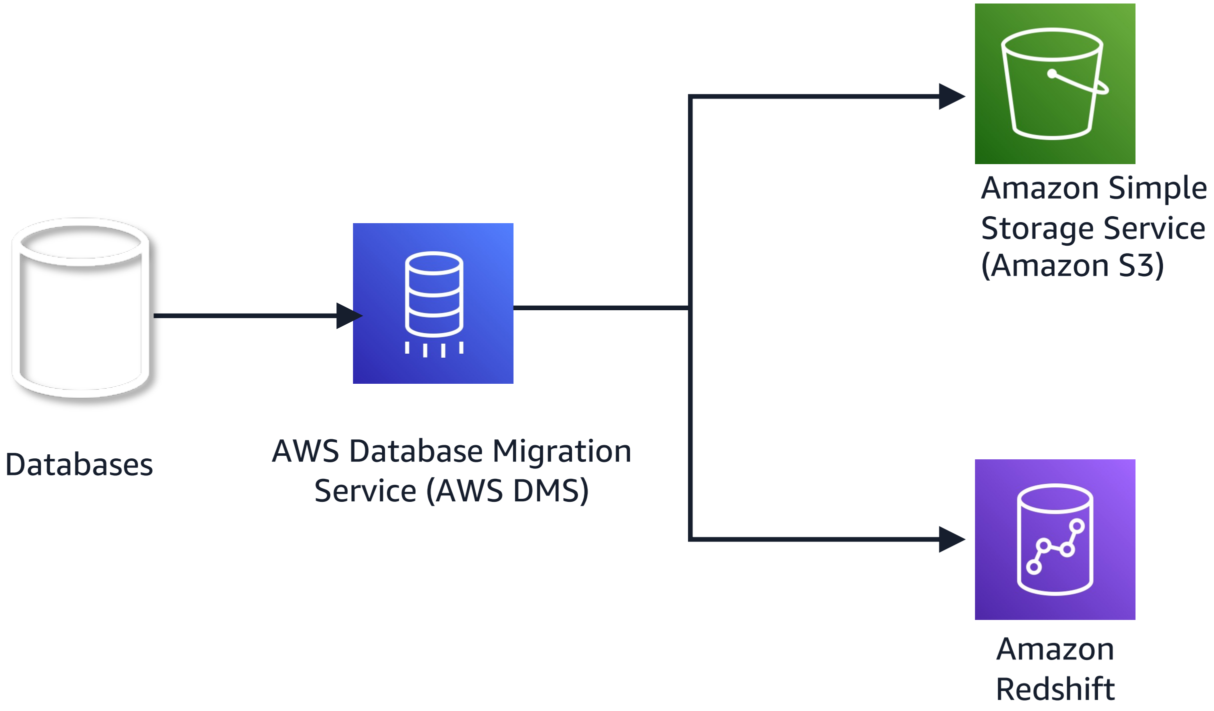
Ingest data from a wide variety of data sources to support unique data sources and data types

The typical list of data sources

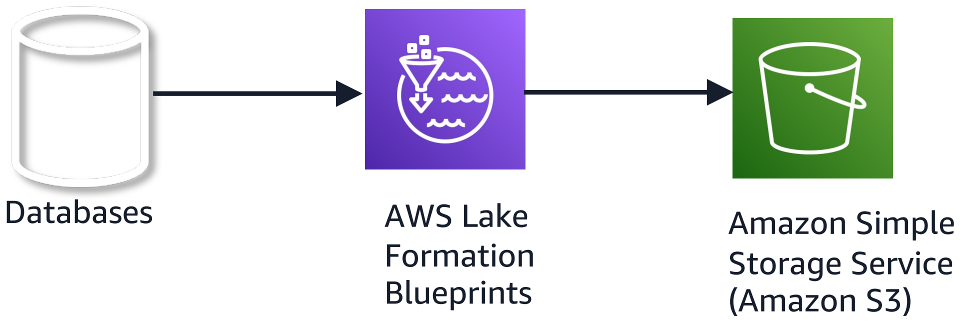
- Database data sources
- Files shares
- SaaS applications
- Partner data feeds
- Third-party data products
- Custom data sources
- Streaming data sources

Database data sources

We provide AWS Database Migration Service (AWS DMS) and AWS Lake Formation blueprints by generating AWS Glue crawlers, jobs, and triggers that discover and ingest database data into storage layer


Databases
AWS Database Migration Service (AWS DMS)
Amazon Simple Storage Service (Amazon S3)
Amazon Redshift

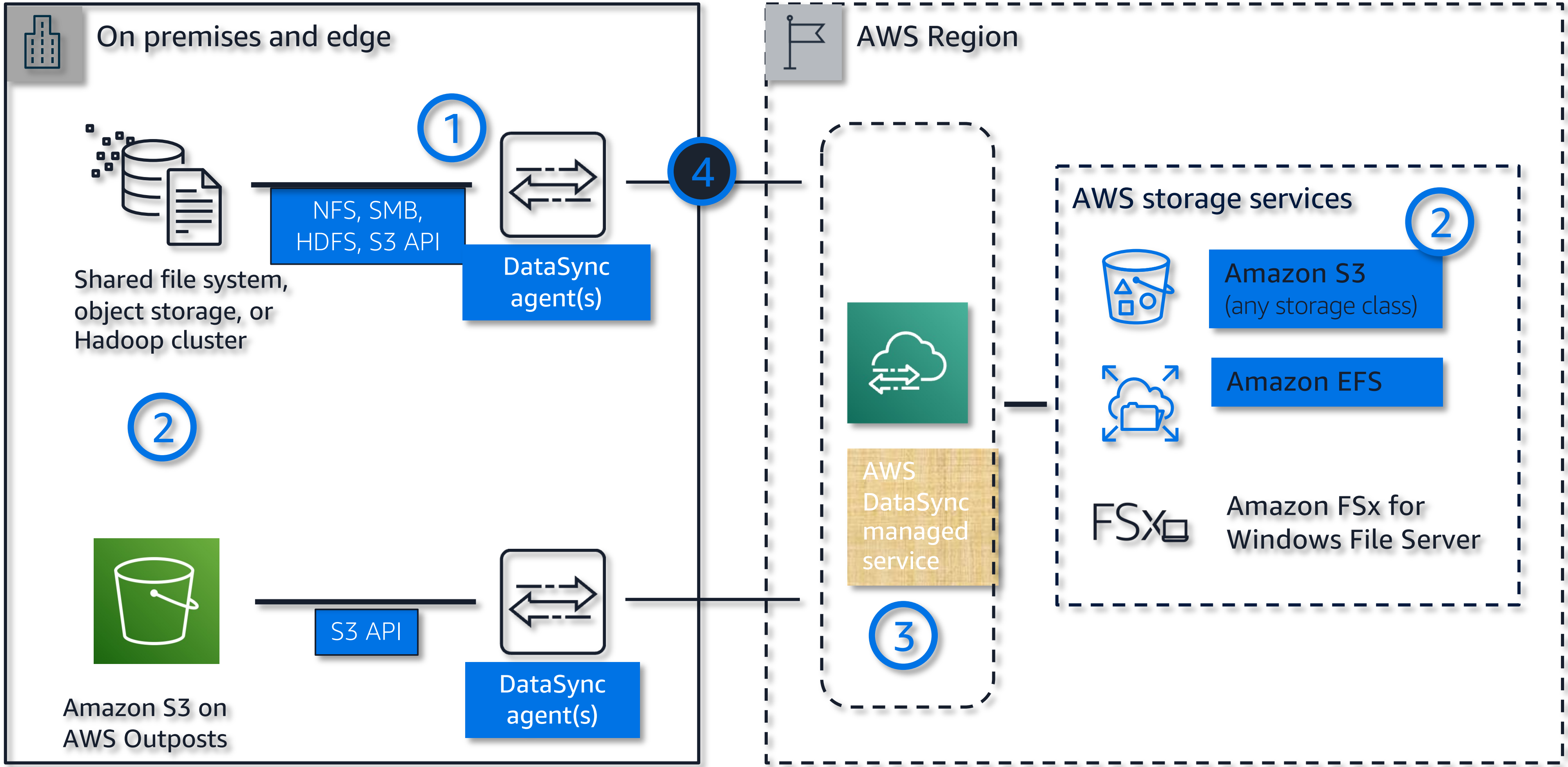
https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Source.html
https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Target.html


Databases
AWS Lake Formation Blueprints
Amazon Simple Storage Service (Amazon S3)

https://docs.aws.amazon.com/lake-formation/latest/dg/workflows-about.html

File shares

AWS DataSync makes it simple and fast to move large amounts of files from Network File System (NFS) shares, Server Message Block (SMB) shares, Hadoop Distributed File Systems (HDFS) into Amazon S3 data lake

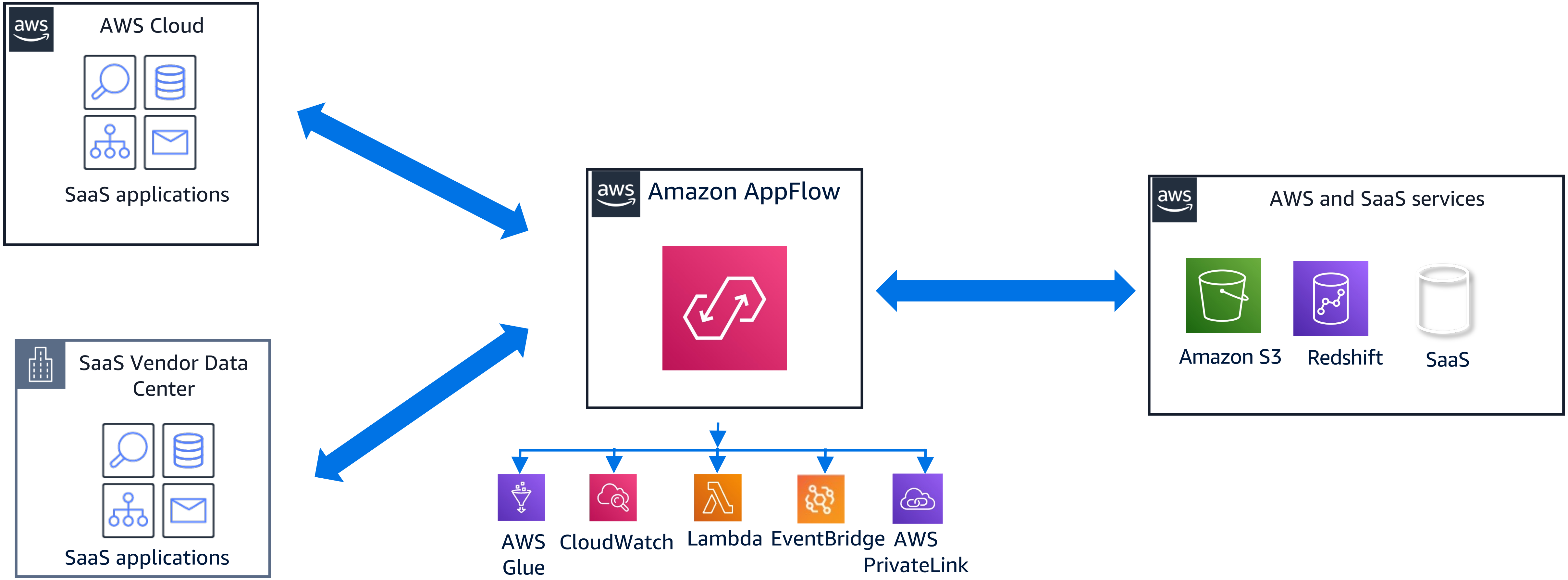


On premises and edge
1
Shared file system, object storage, or Hadoop cluster
NFS, SMB, HDFS, S3 API
DataSync agent(s)
4
2
Amazon S3 on AWS Outposts
S3 API
DataSync agent(s)
AWS Region
AWS DataSync managed service
3
AWS storage services
2
Amazon S3 (any storage class)
Amazon EFS
FSx
Amazon FSx for Windows File Server

1. Agents are deployed to connect to on-premises storage
2. Locations control how AWS DataSync connects to storage
3. AWS DataSync managed service connects to AWS storage and coordinates the transfer
4. For transfers between on premises and AWS, internet, AWS Direct Connect, and AWS Virtual Private Network (VPN) are supported

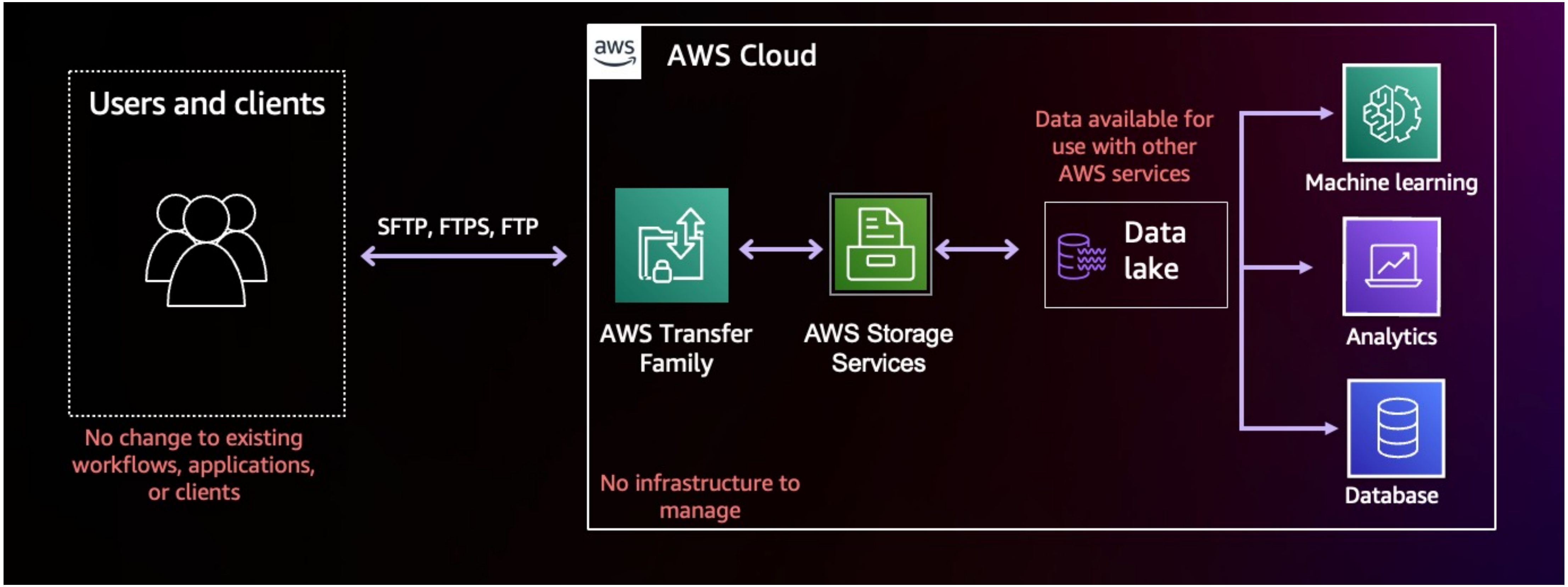
SaaS applications data

Amazon AppFlow makes it easy to ingest SaaS applications data into storage layer



Partner data feeds

AWS Transfer Family is a serverless service that provides secure FTP endpoints and integrates with Amazon S3 and it stores partner data feeds as S3 objects in the landing zone of the data lake


AWS Cloud
Users and clients
SFTP, FTPS, FTP
AWS Transfer Family
AWS Storage Services
Data available for use with other AWS services
Data lake
Machine learning
Analytics
Database
No change to existing workflows, applications, or clients
No infrastructure to manage

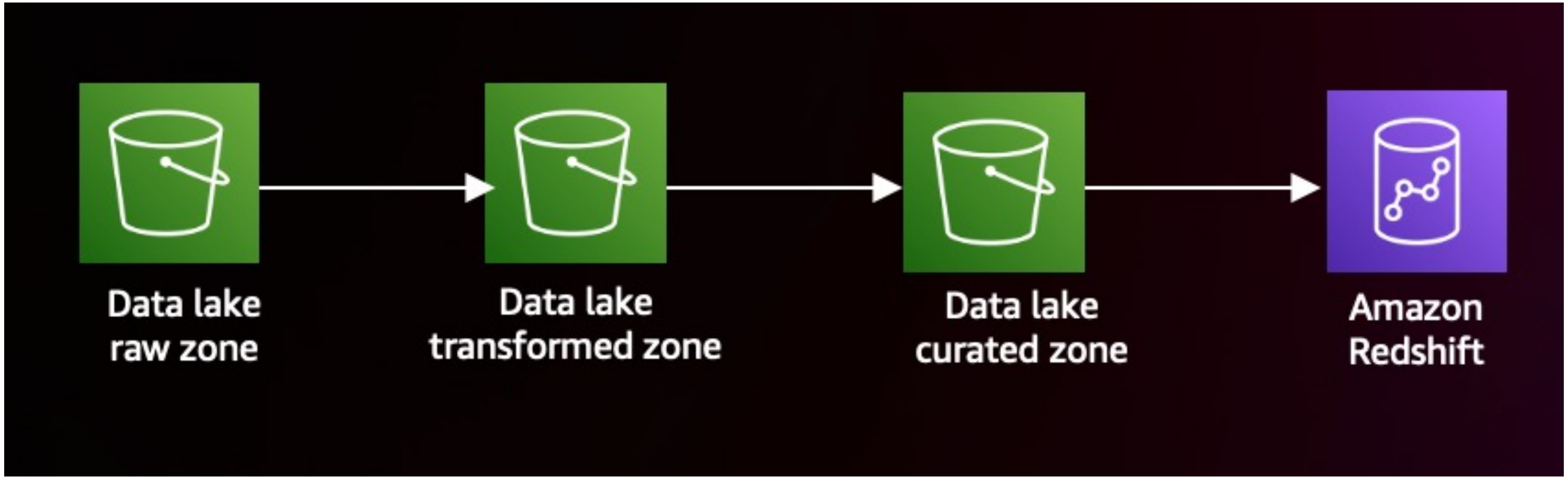
Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

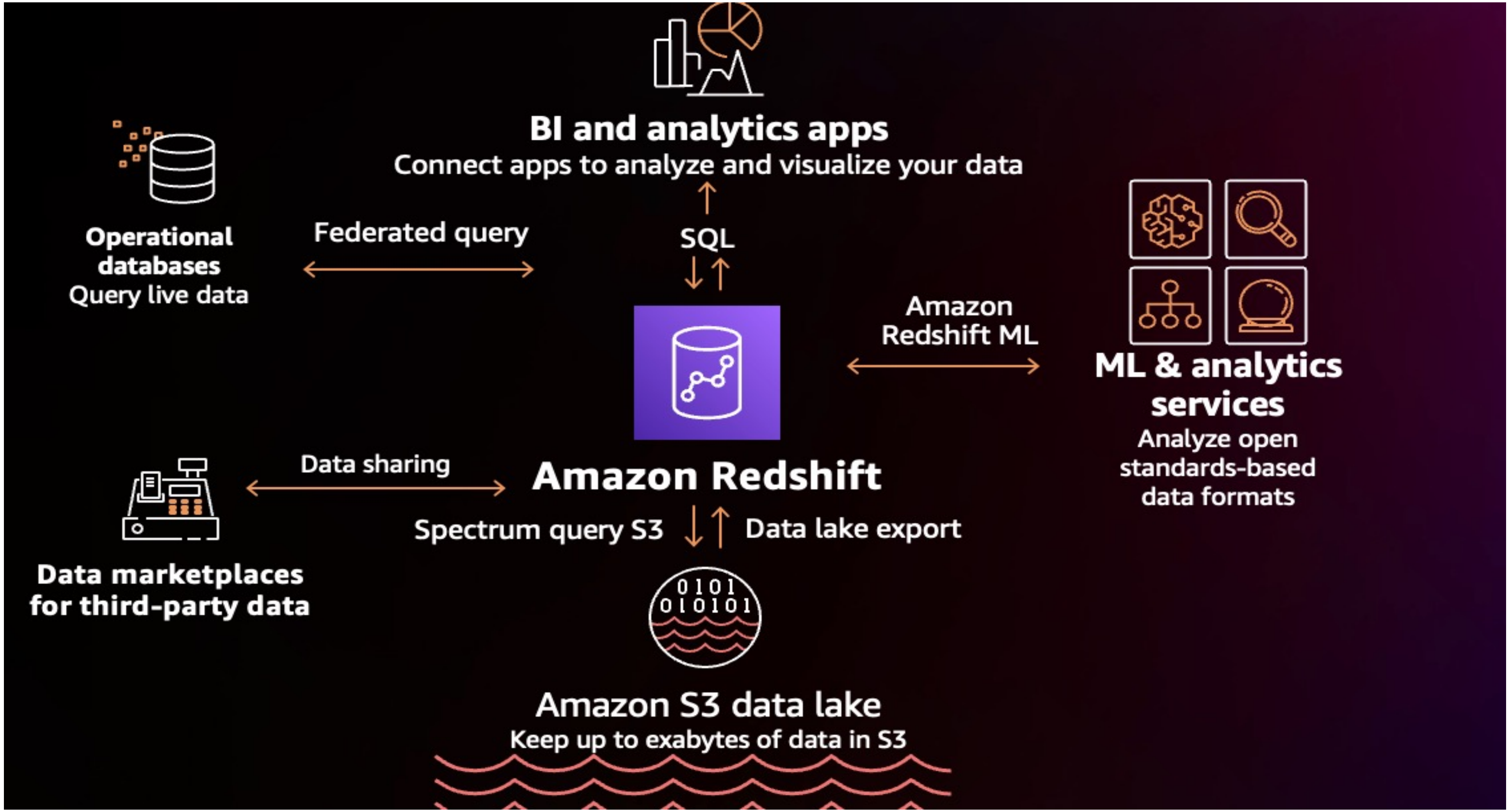
1. Data ingestion	2. Data storage	3. Data cataloging	4. Data processing	5. Data consumption	6. Security and Governance
Bring the data into your data platform	Store your structured and unstructured data	Store your metadata	Create data processing pipelines	Enable your user personas for purpose-built analytics and machine learning	Protect your data across the layers and data access management

Data storage layer

- The storage layer consists of Amazon S3 and Amazon Redshift, an integrated storage layer for the modern data architectures on AWS. You can put datasets into three different areas in S3 data lake: raw zone, cleaned or transformed zone, and curated zone



Modern data architecture storage layer integrates Amazon S3 data lake and Amazon Redshift data warehouse



Building modern data architecture

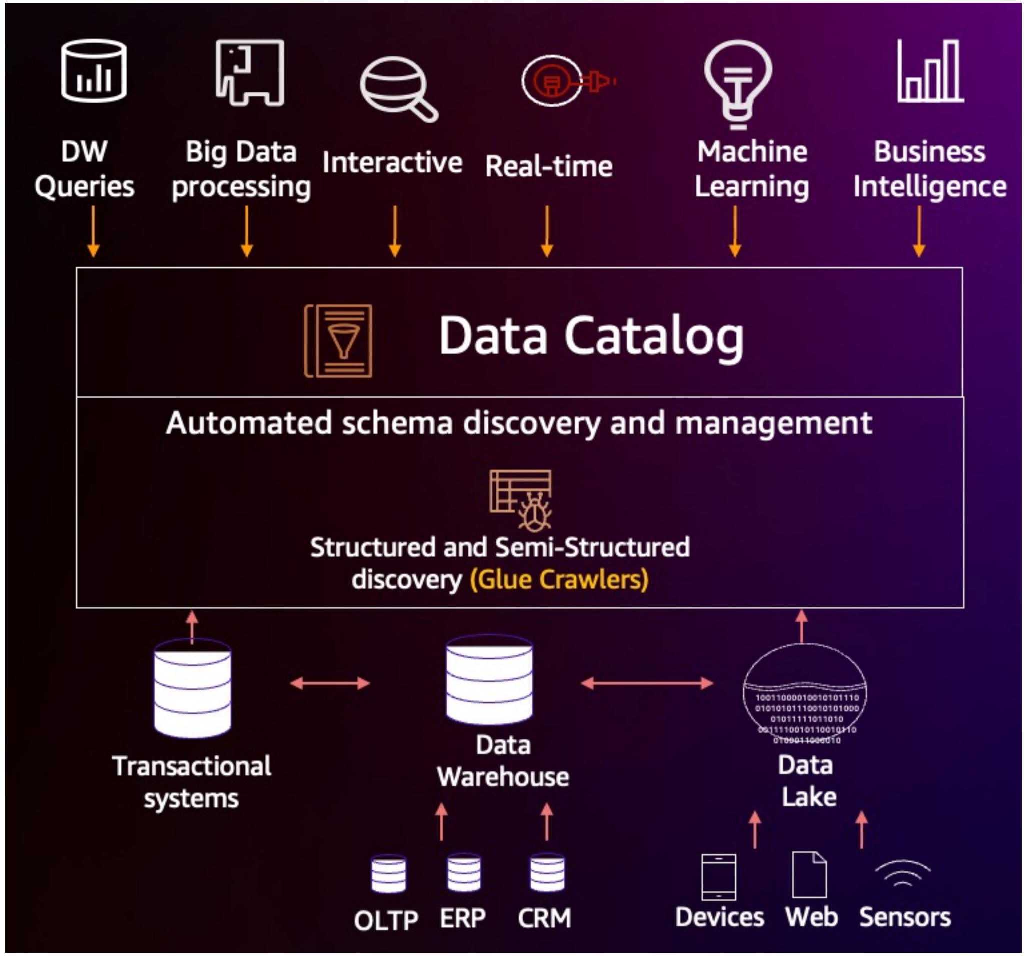
ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion	2. Data storage	3. Data cataloging	4. Data processing	5. Data consumption	6. Security and Governance
Bring the data into your data platform	Store your structured and unstructured data	Store your metadata	Create data processing pipelines	Enable your user personas for purpose-built analytics and machine learning	Protect your data across the layers and data access management

Data catalog layer

AWS Glue Data Catalog provides the central catalog to store metadata for all datasets hosted in the storage layer

- No movement of data = Low Costs/Admin
- All metadata centrally available for search and query = Productivity
- Unify structured, semi-structured data = Speed to Insight
- Automate data discovery = Productivity



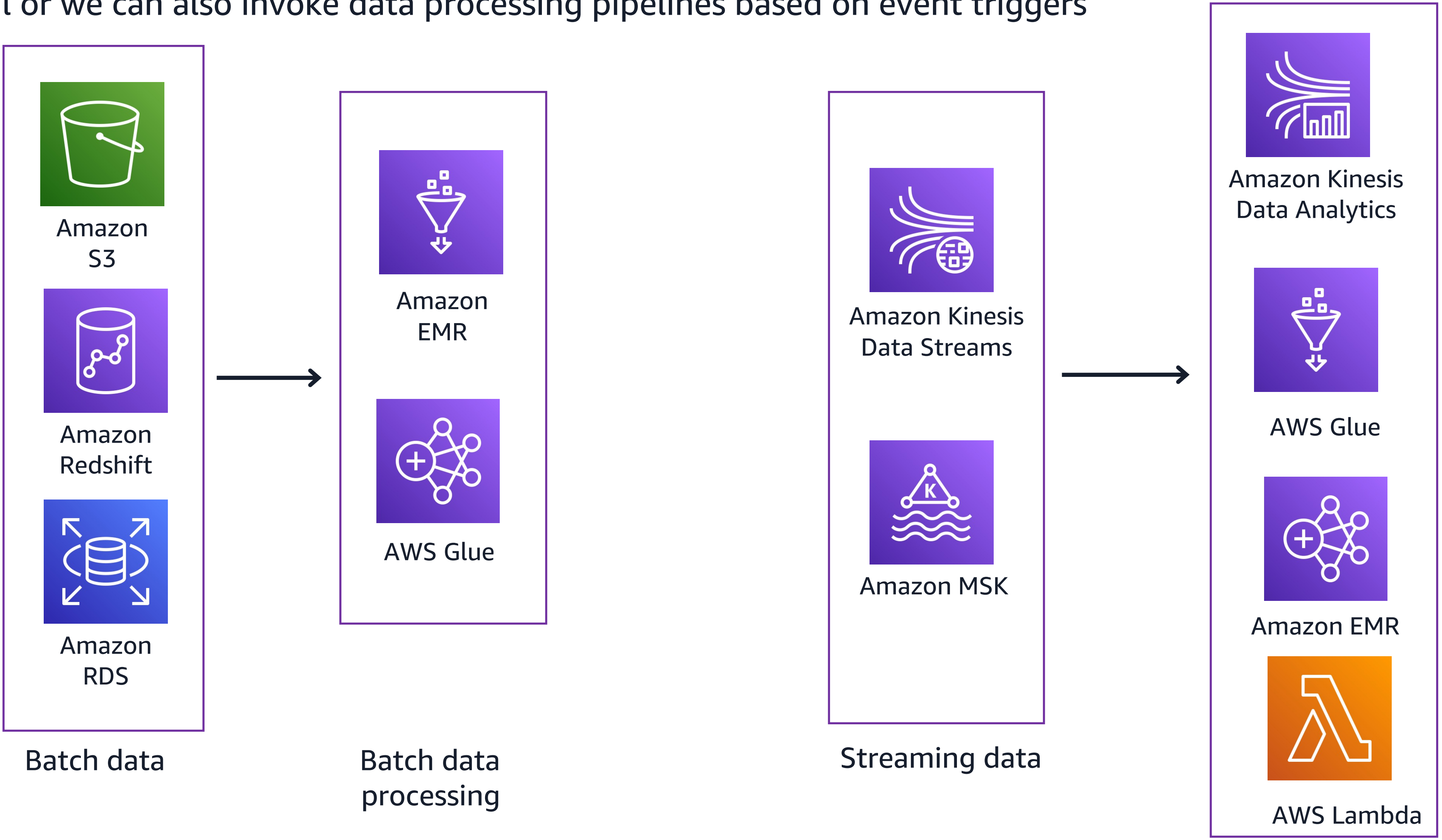
Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion	2. Data storage	3. Data cataloging	4. Data processing	5. Data consumption	6. Security and Governance
Bring the data into your data platform	Store your structured and unstructured data	Store your metadata	Create data processing pipelines	Enable your user personas for purpose-built analytics and machine learning	Protect your data across the layers and data access management

Data processing layer

Data processing pipelines can be multistep data processing pipelines or scheduled data processing pipelines on a regular interval or we can also invoke data processing pipelines based on event triggers



Amazon S3

Amazon Redshift

Amazon RDS

Batch data

Amazon EMR

AWS Glue

Batch data processing

Amazon Kinesis Data Streams

Amazon MSK

Streaming data

Amazon Kinesis Data Analytics

AWS Glue

Amazon EMR

AWS Lambda

Streaming data processing

Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion

Bring the data into your data platform

2. Data storage

Store your structured and unstructured data

3. Data cataloging

Store your metadata

4. Data processing

Create data processing pipelines

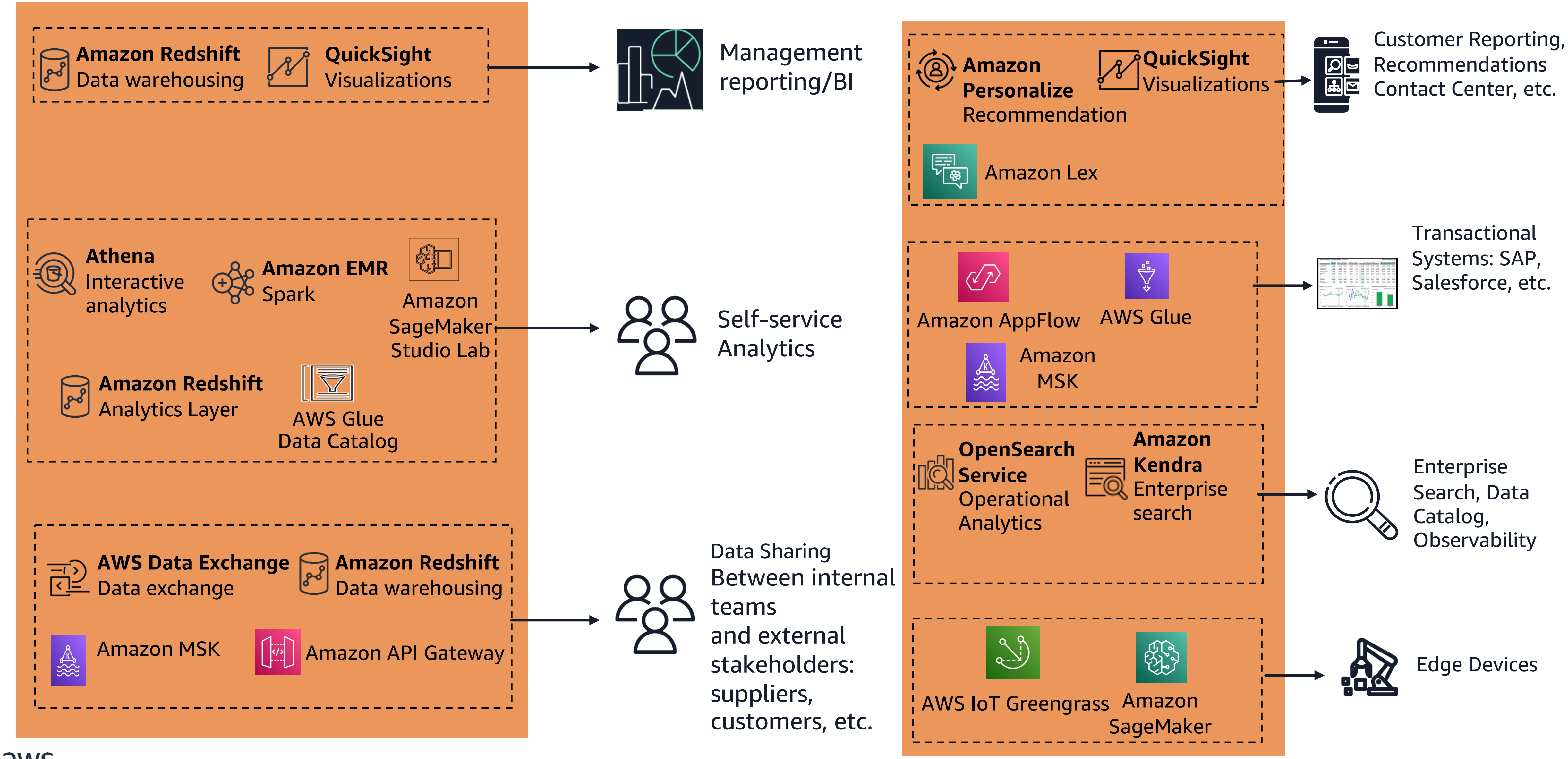
5. Data consumption

Enable your user personas for purpose-built analytics and machine learning

6. Security and Governance

Protect your data across the layers and data access management

Data consumption patterns



Amazon Redshift Data warehousing

QuickSight Visualizations

Management reporting/BI

Athena Interactive analytics

Amazon EMR Spark

Amazon SageMaker Studio Lab

Amazon Redshift Analytics Layer

AWS Glue Data Catalog

Self-service Analytics

AWS Data Exchange Data exchange

Amazon Redshift Data warehousing

Amazon MSK

Amazon API Gateway

Data Sharing
Between internal teams
and external stakeholders: suppliers, customers, etc.

Amazon Personalize Recommendation

QuickSight Visualizations

Amazon Lex

Customer Reporting, Recommendations Contact Center, etc.

Amazon AppFlow

AWS Glue

Amazon MSK

Transactional Systems: SAP, Salesforce, etc.

OpenSearch Service Operational Analytics

Amazon Kendra Enterprise search

Enterprise Search, Data Catalog, Observability

AWS IoT Greengrass

Amazon SageMaker

Edge Devices

Data consumption layer – Machine learning

Amazon SageMaker is a complete, end-to-end service for machine learning

PREPARE →

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store
Store, update, retrieve, and share features

SageMaker Clarify
Detect bias and understand model predictions

BUILD →

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring your-own Algorithms
Dozens of optimized algorithms or bring your own

Local Mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart NEW
Pre-built solutions for common use cases

TRAIN & TUNE →

Managed Training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare
every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries NEW
Training for large datasets and models

SageMaker Debugger NEW
Debug and profile training runs

Managed Spot Training
Reduce training cost by 90%

DEPLOY & MANAGE →

Managed Deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager NEW
Manage and monitor models on edge devices

SageMaker Pipelines NEW
Workflow orchestration and automation

SageMaker Studio
Integrated development environment ((IDE) for ML

Building modern data architecture

ENVISION A MODERAN DATA ARCHITECTURE AS A STACK OF SIX LAYERS

1. Data ingestion

Bring the data into your data platform

2. Data storage

Store your structured and unstructured data

3. Data cataloging

Store your metadata

4. Data processing

Create data processing pipelines

5. Data consumption

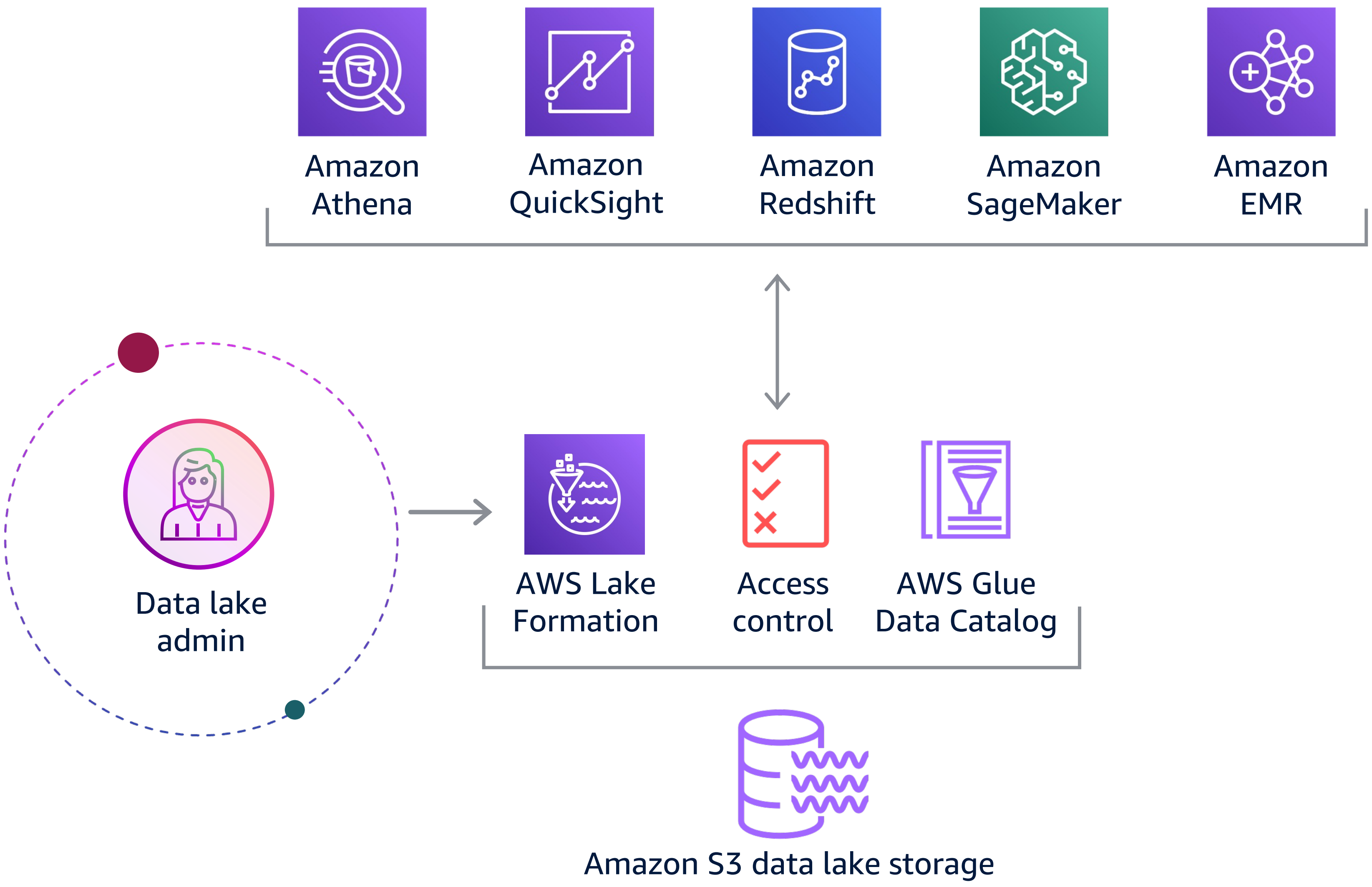
Enable your user personas for purpose-built analytics and machine learning

6. Security and Governance

Protect your data across the layers and data access management

AWS Lake Formation: Unified data governance

Simplified and unified security management



Amazon Athena

Amazon QuickSight

Amazon Redshift

Amazon SageMaker

Amazon EMR

Data lake admin

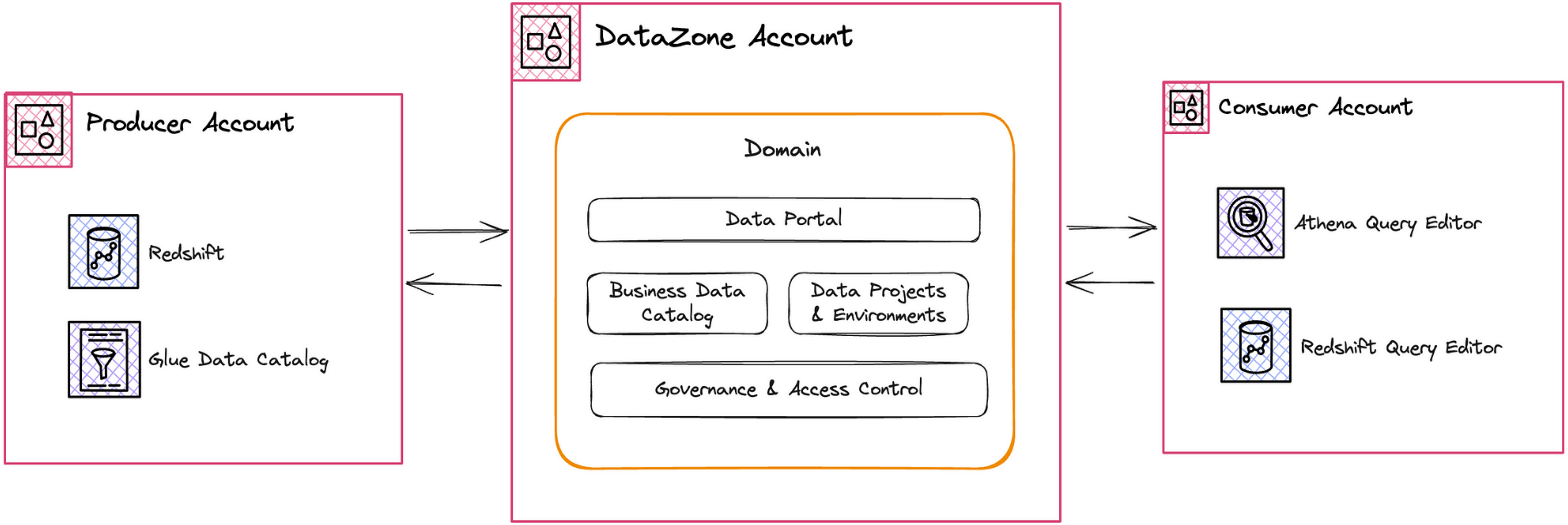
AWS Lake Formation

Access control

AWS Glue Data Catalog

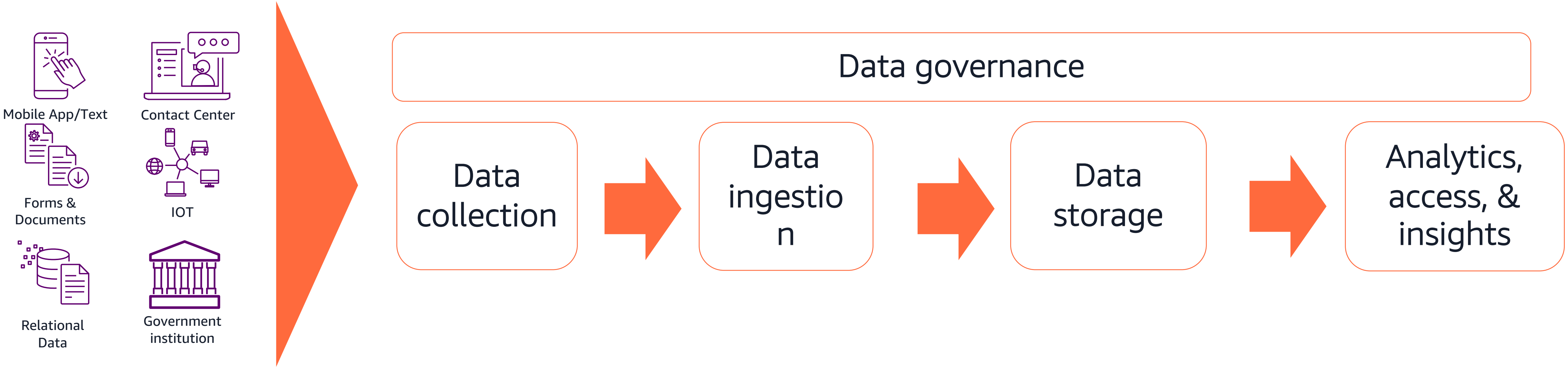
Amazon S3 data lake storage

DataZone for Data Mesh Architecture



Putting it all together

Key components of modern data architecture



Security – Reliability – Operational Excellence – Performance Efficiency – Cost Optimization – Sustainability

Key considerations:

1. Ability to handle the increasing volume, velocity, and variety of data

2. Each component should be independently scalable

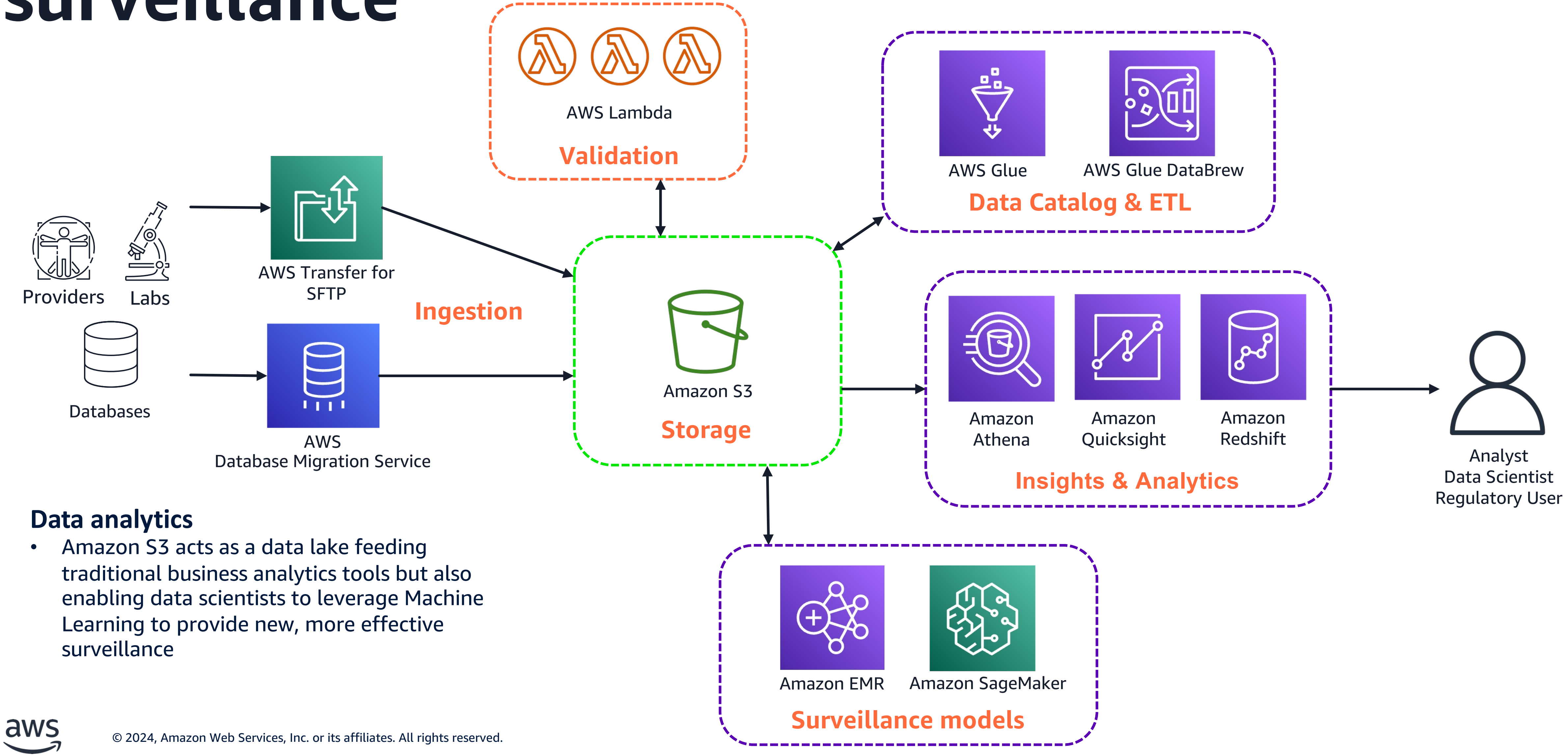
3. Make data easily accessible and sharable

Reference Architectures

Public health Organization

1. Pandemic brings 1000% increase in disease surveillance data
2. Legacy management systems
3. Limited capabilities to consolidate data sets from multiple systems
4. Difficulty mandating data formats from various partner organizations

Sample reference architecture for disease surveillance



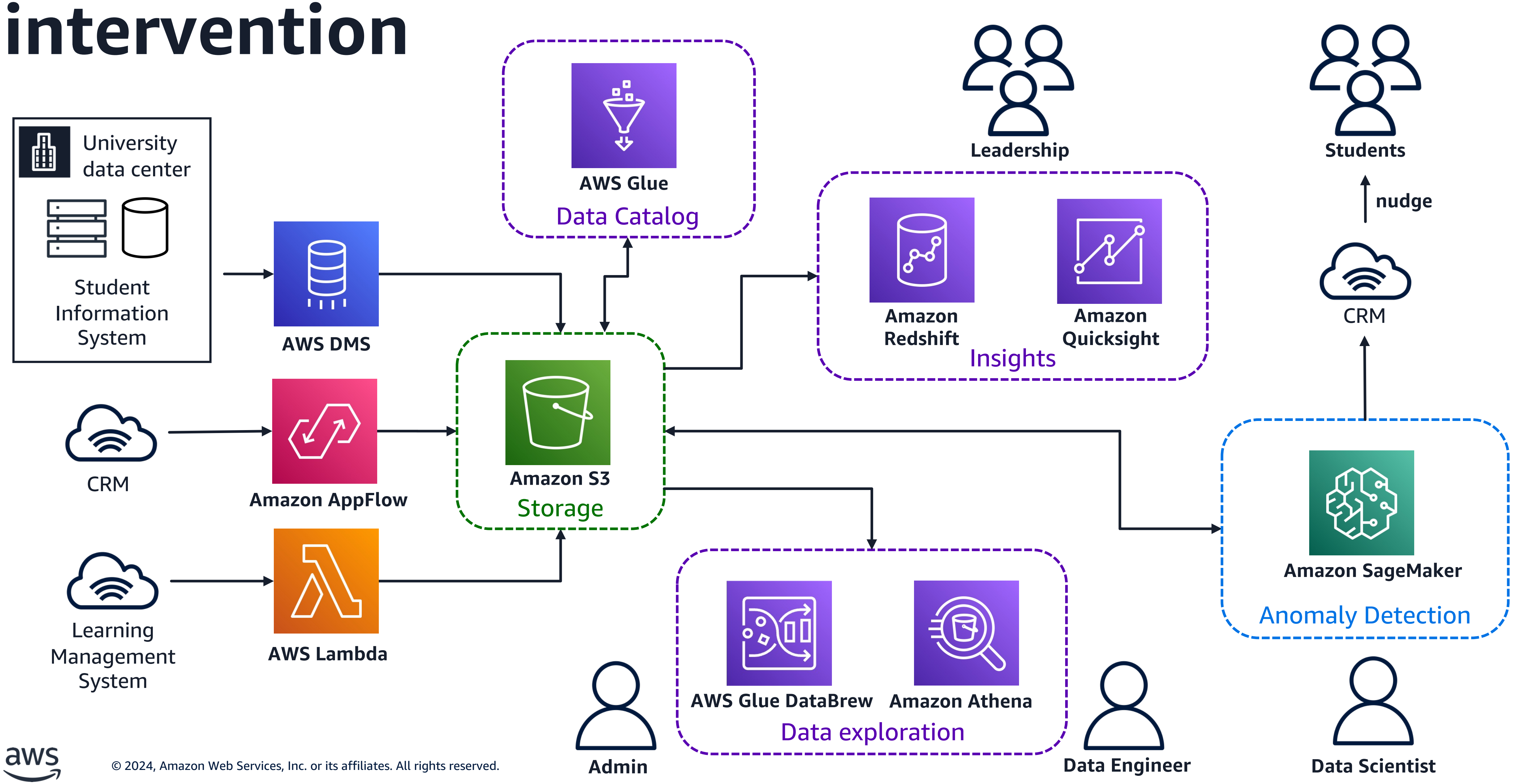
Data analytics

- Amazon S3 acts as a data lake feeding traditional business analytics tools but also enabling data scientists to leverage Machine Learning to provide new, more effective surveillance

Improving student outcomes - Retention

1. **Identify at-risk students** from behaviors
2. Aggregated student touchpoint data from the **SIS, LMS, and CRM**
3. Feed insights into communication platform for **early intervention and nudging**

Sample reference architecture for student intervention



Moving the needle on retention

MARYVILLE UNIVERSITY

1. IT staff participated in data lake and modern data architecture **skills development**
2. Aggregated student touchpoint data from the **SIS, LMS, and CRM** into a data lake in 6 weeks
3. Automated processing and machine learning to **identify at-risk students** from behaviors
4. Fed insights into communication platform for **early intervention and nudging**

Get started

BUILD WITH US

ML Solutions Lab

AWS Professional Services

AWS Immersion Day

Data-Driven Everything

Migration Assistance Program

BUILD WITH PARTNERS

AWS Partner Network—
100,000+ partners

AWS Marketplace (ISVs)

UPSKILL YOUR TEAMS

AWS Training and Certification

ML Embark Program