



How to leverage AWS Modern Data Architecture to Accelerate your Data Strategy

Bob Maus

Head, WW Data & Streaming GTM Team

Ryan Shevchik

Sr Solution Architect, Data & Analytics

Put data to work



Make better
decisions



Improve
efficiencies



Respond
faster



Uncover
opportunities

Most comprehensive set of services for the entire end-to-end data, analytics, and ML journey for all workloads and all types of data

Ingest + store

Analyze + visualize

Predict



**Security +
Governance +
Access control**

Common barriers to driving business outcomes with data

- Support Collaborative Model of Data Producer and Consumers
- Data as a Platform vs. Product Mechanism of Driving Business Outcomes
- Data Governance & Compliance
- Common Access across the Enterprise
- Trust and Confidence in Data

THE ASPIRATION Becoming a data-driven organization

An organization that harnesses data as an asset, to drive sustained innovation and create actionable insights to supercharge the experience for their customers so they demand more...

99% of businesses want to be data driven

24% have been successful

Source: Forbes Online; New Vantage Partners Survey

Common Themes We Hear From Our Customers

“Sharing and searching data is difficult”

“Difficult to meet all requirements across differing business units”

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

“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”

“If I share data, I’ve lost control”

“Why doesn’t our organization treat data as a product?”

“There is a mis-match between executive leadership goals and business line deliverables and incentives”

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

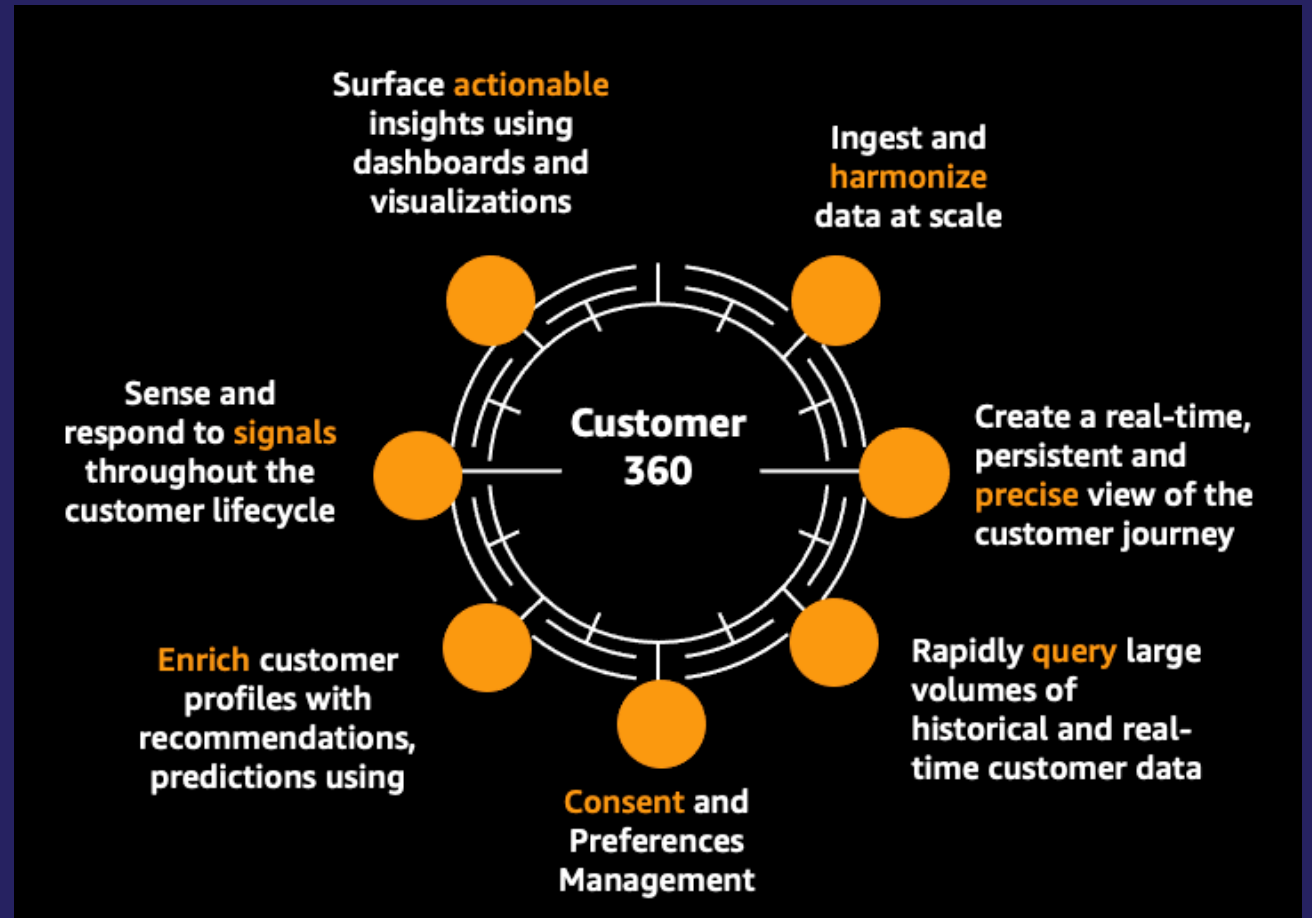
“Current data architecture is complex and monolithic and slow to change”

“Our internal policies on what can be shared unclear and there is lack of incentive to share

“I just want to get access to the data I need”

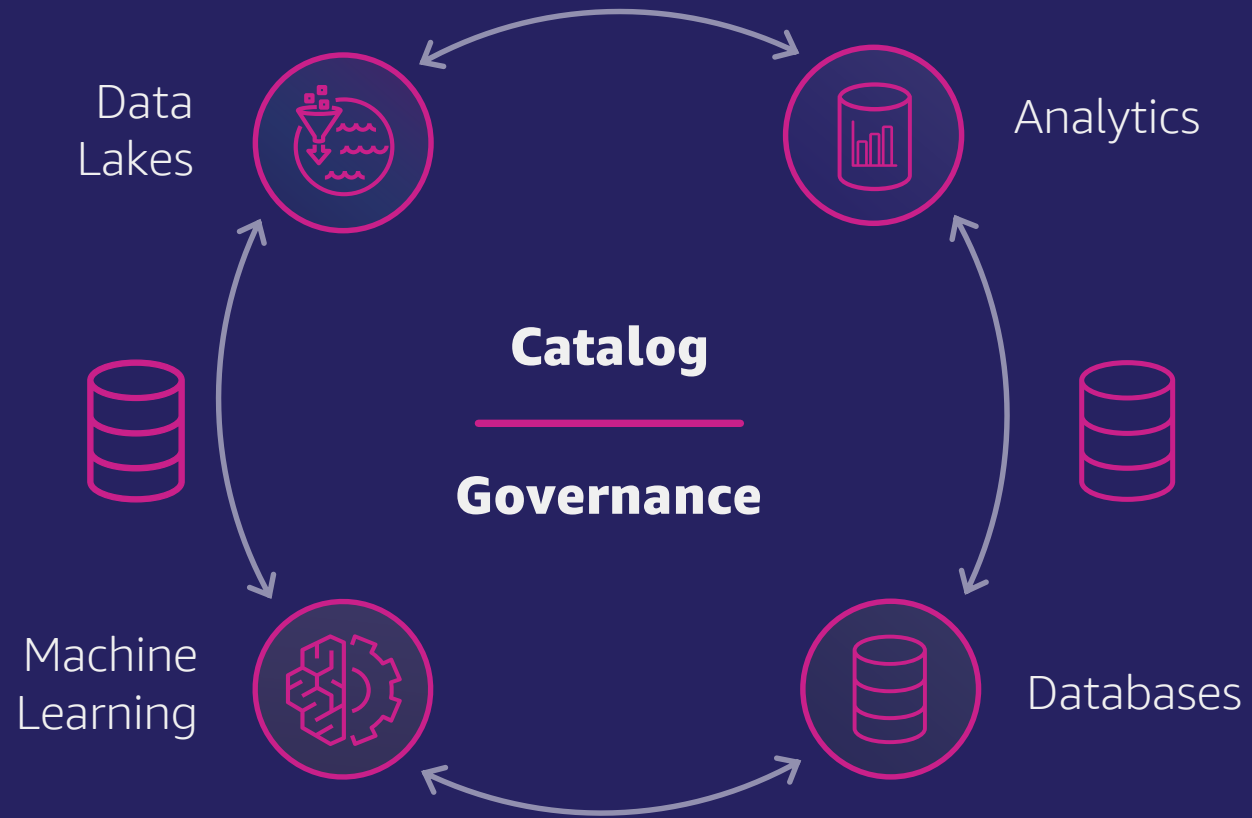
What is a Data Product?

- Broad, cohesive collections of related data aligned to **business use cases and goals**
- **Federated** governance of data created, provided, stored, transformed in, or consumed from the data product
- **Diverse** data types sourced from distributed internal and external sources.



Traditional approaches don't scale



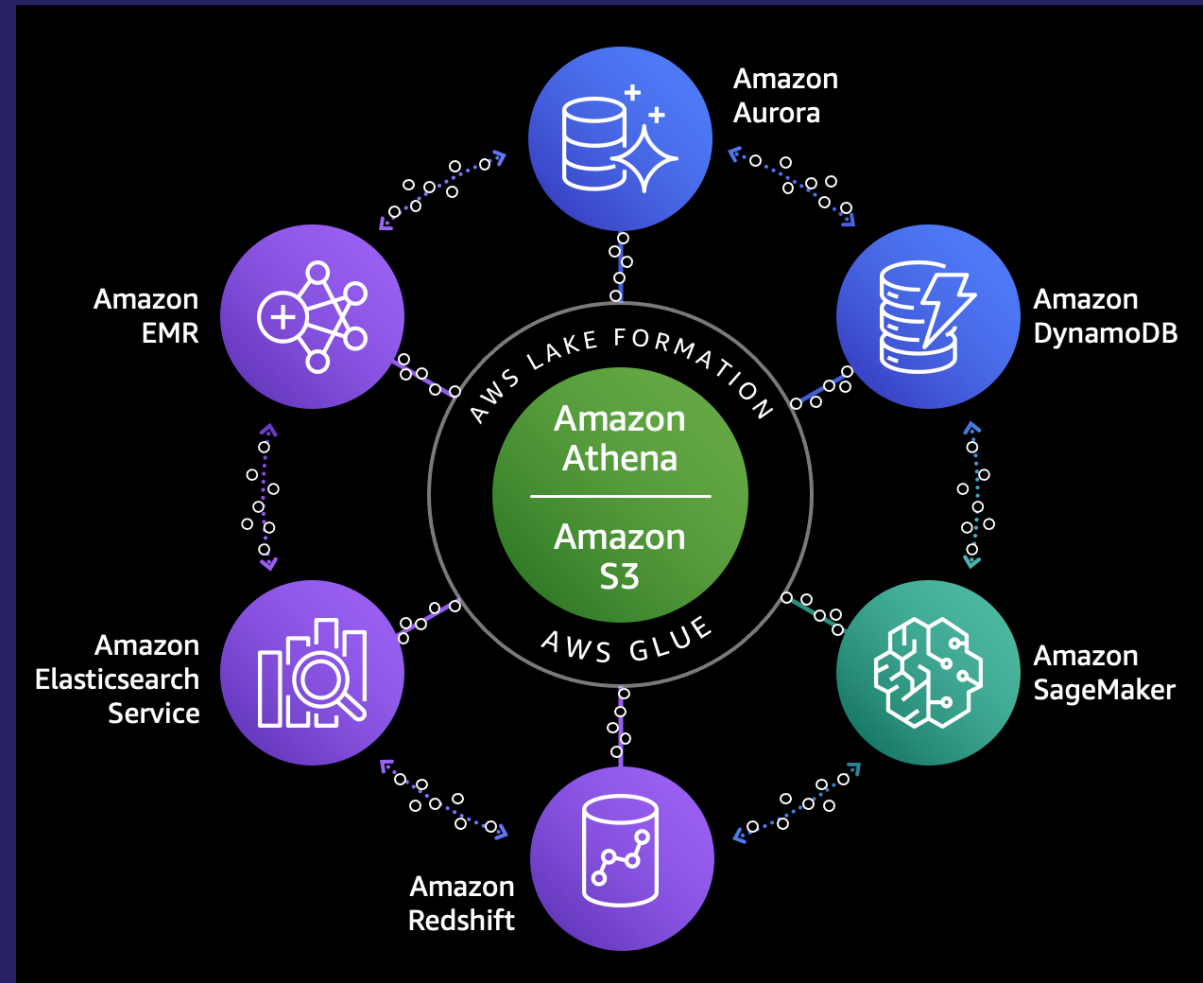


What is a Modern Data Architecture?

With a Modern Data Architecture (MDA) on AWS, customers can store **domain-specific** data in a data lake and use a ring of purpose-built data services around the lake allowing them to make decisions with speed and agility, and scale their systems at a low cost without compromising performance.

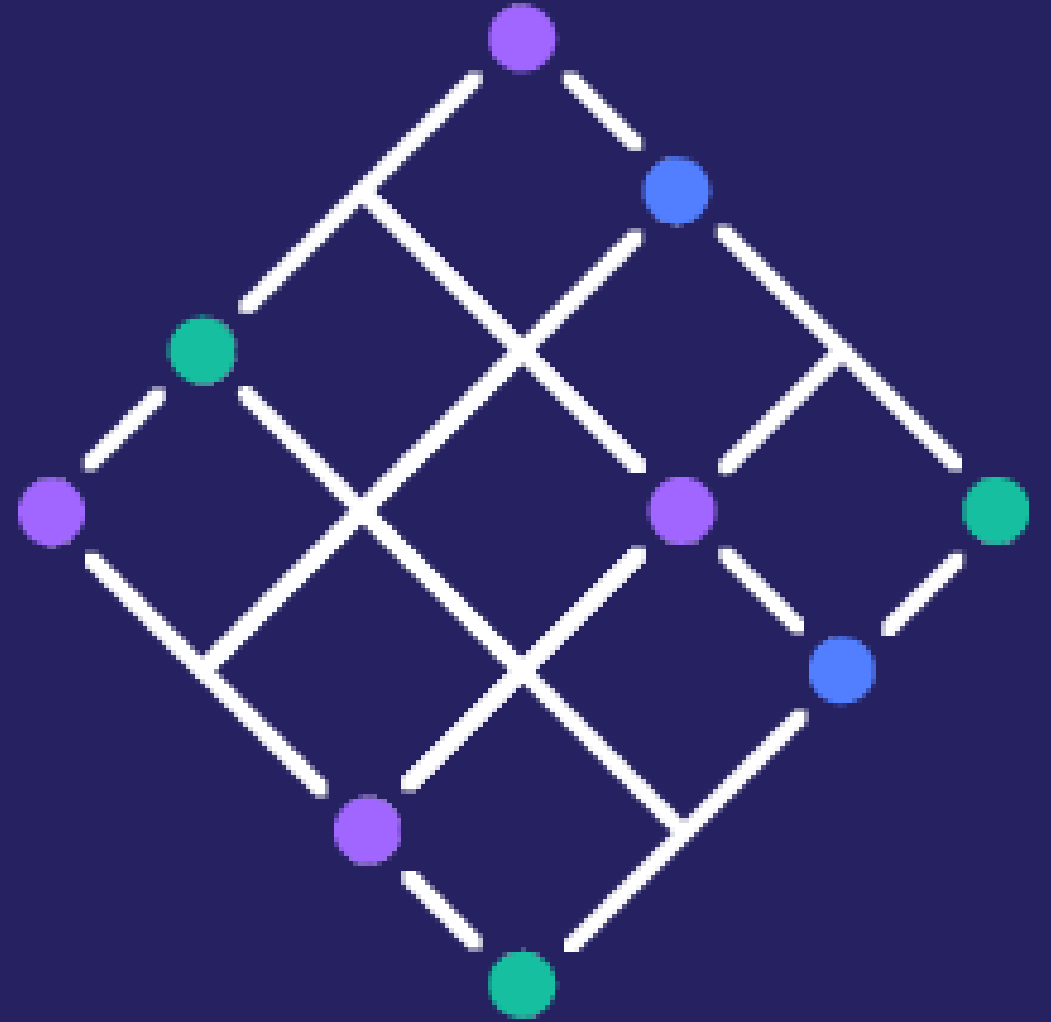
The MDA enables customers to:

- **Deliver** a collection of domain-specific data products
- **Rapidly** build scalable data lakes
- Use a broad and deep collection of **purpose-built data services**
- **Ensure compliance** via a unified way to secure, monitor, catalog, and manage access to your data
- **Scale your systems at a low cost** without compromising performance

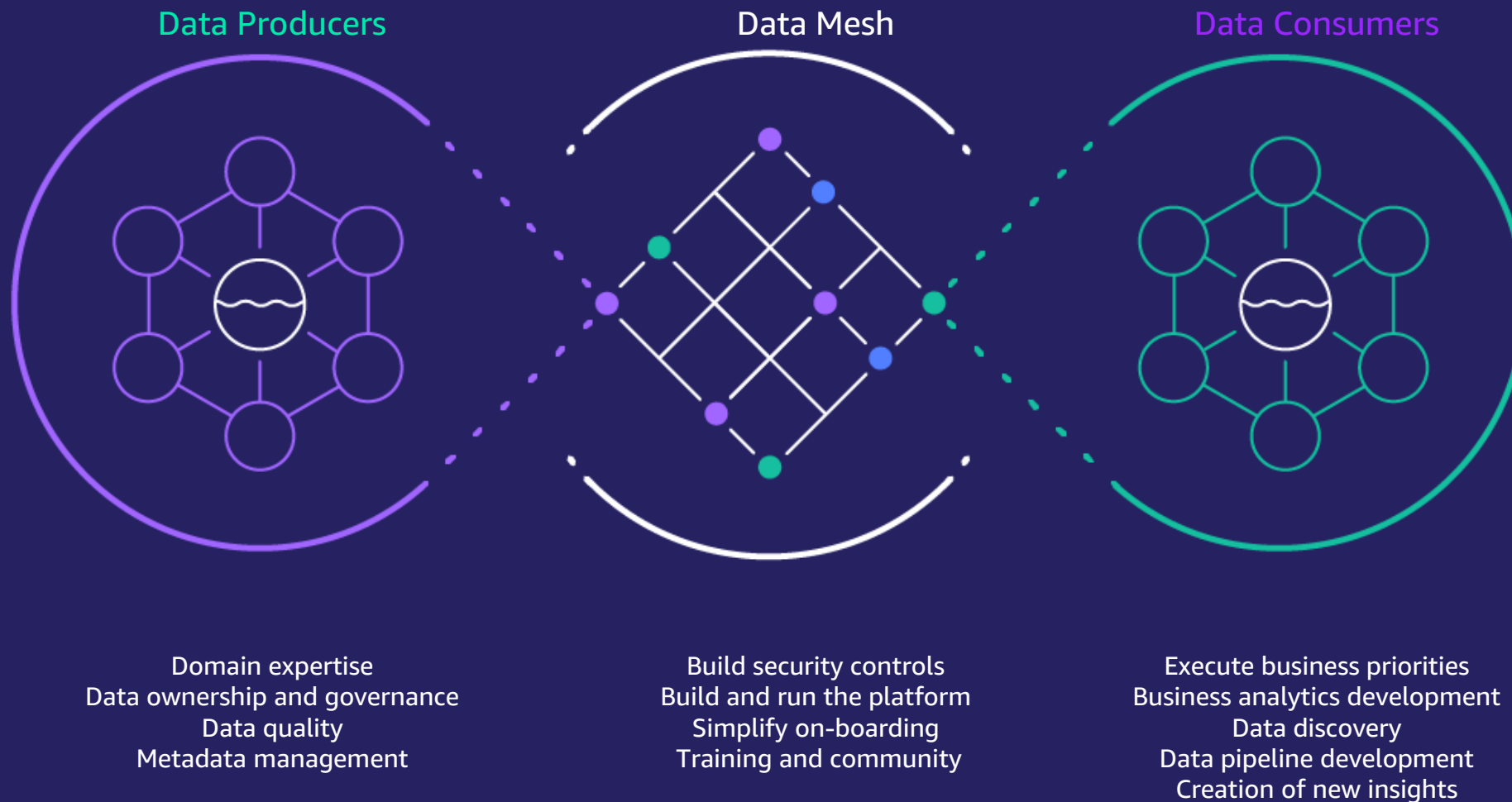


What is a Data Mesh?

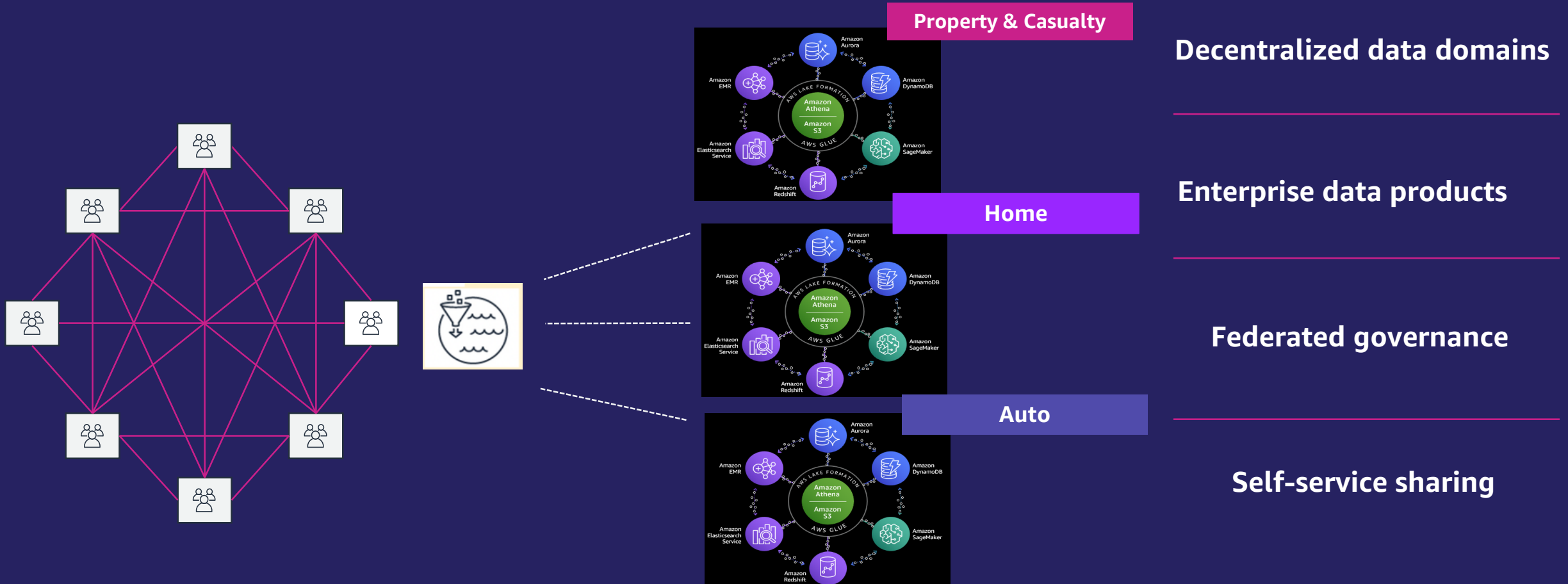
- **Data mesh** is a pattern for defining how organizations can organize around **one or more data domains** with a focus on delivering **enterprise data products**.
- Supports **data producers** and **consumers** and provides federated **governance** through lightweight centralized policy



Decentralized Data Domains – One of More Data Domains Supporting Lake House Architecture



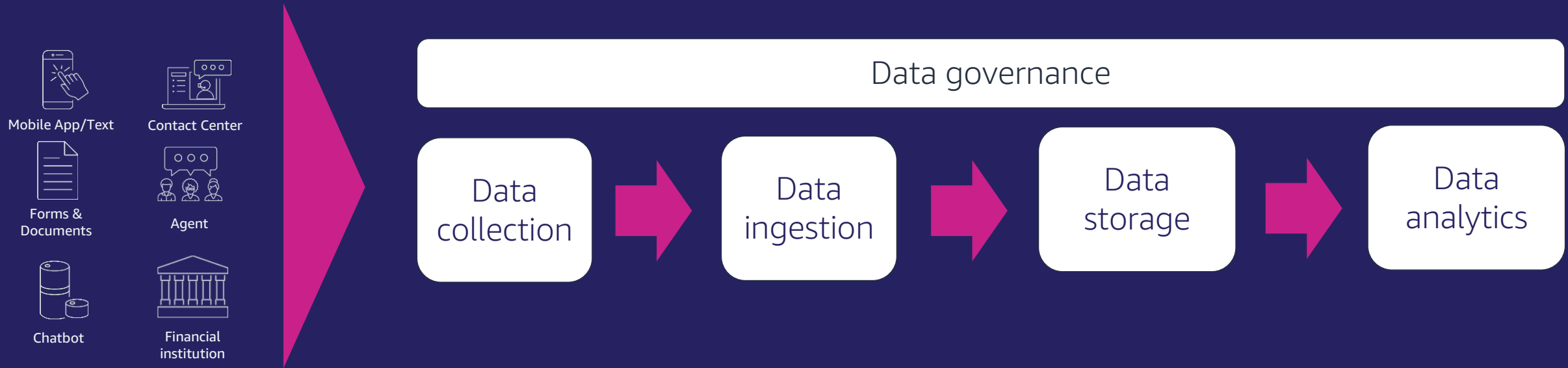
Data Mesh is a Collection of Domain-Specific Lake Houses



Enterprise View Across Distributed Data **VS** a Centralized Data Store

AWS's Modern Data Architecture

Key components of a modern data architecture



Key considerations:

1

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

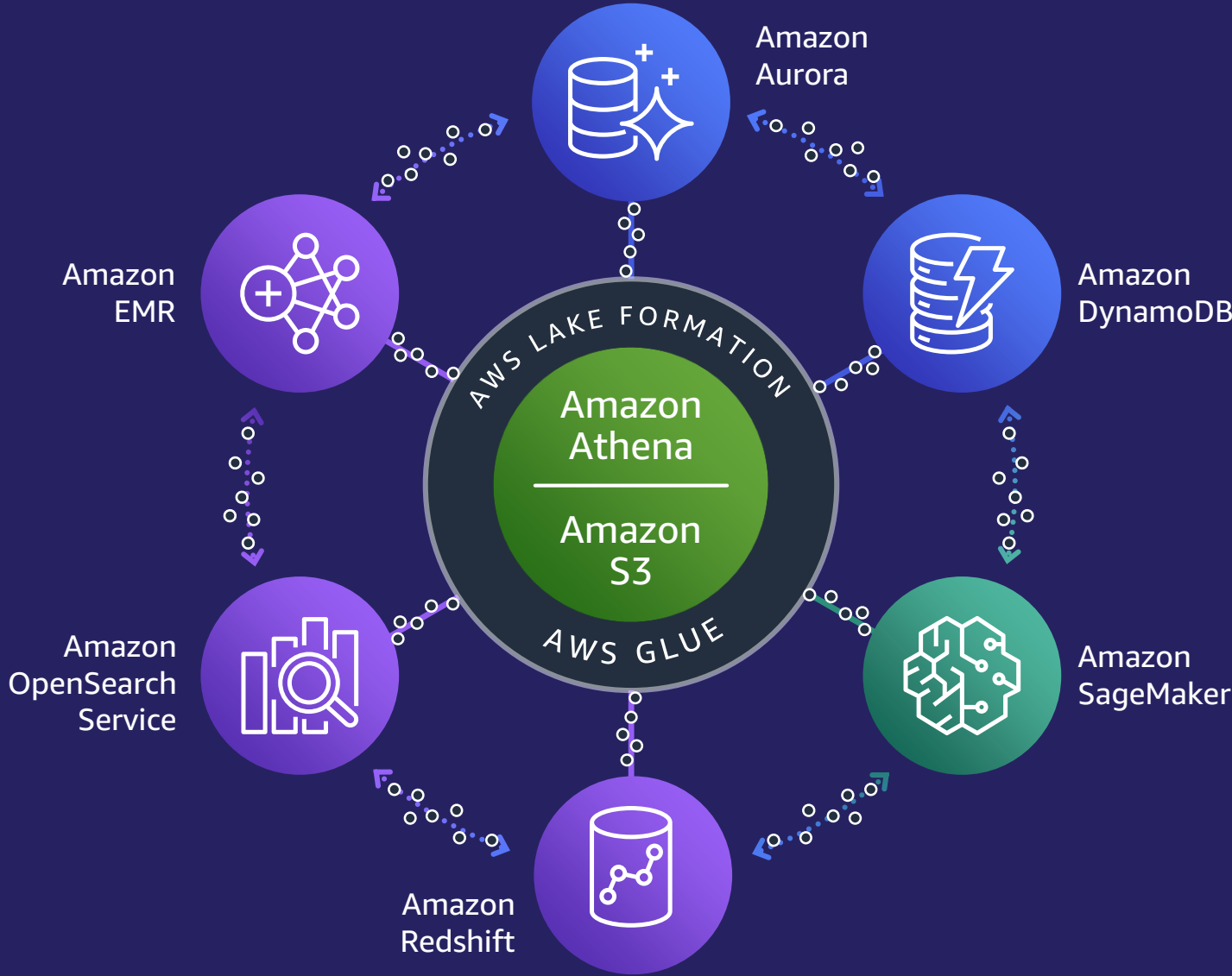
2

Each component should be independently scalable

3

Make data easily accessible and sharable

Modern Data Architecture (MDA) on AWS



SCALABLE DATA LAKES

PURPOSE-BUILT
DATA SERVICES

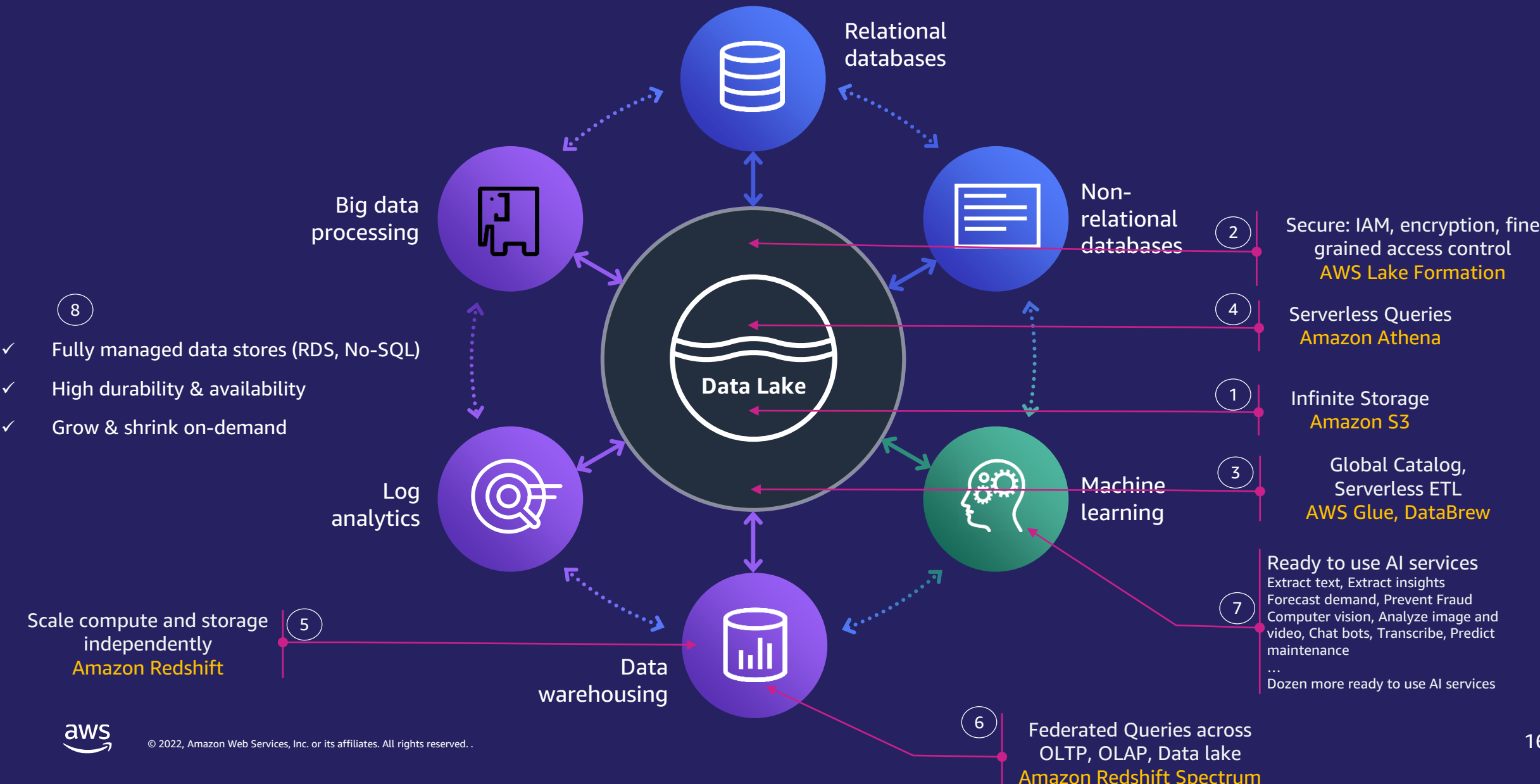
UNIFIED DATA ACCESS

UNIFIED GOVERNANCE

PERFORMANT AND
COST-EFFECTIVE



MDA Critical Capabilities



Amazon Fully Managed Services



BENEFITS

Increase
topline
growth

Optimize
operational
costs

Reduce
risk

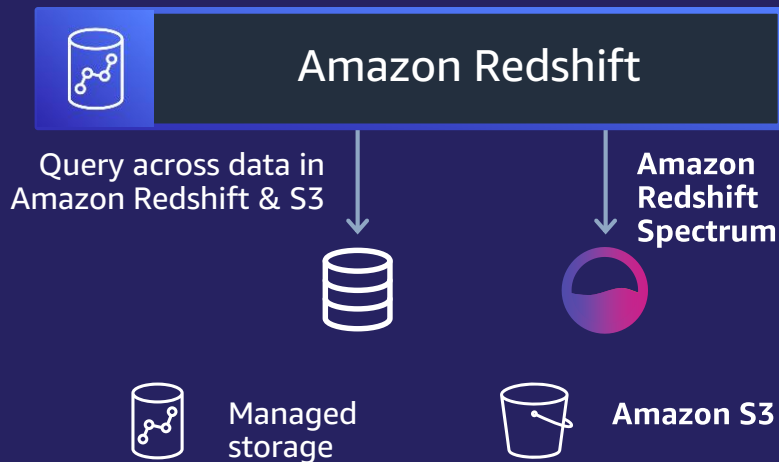
Moving data to and from the data lake

Extend the data warehouse to exabytes of data in Amazon S3 data lakes

Directly query data stored in Amazon S3

Parquet, ORC, Avro, JSON, and CSV data formats

Any scale of data; pay for what you use



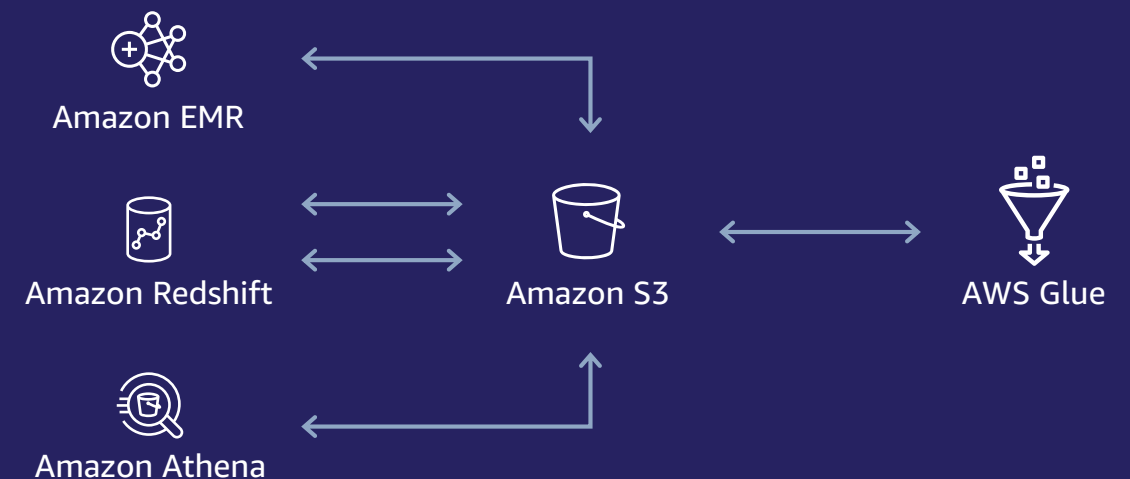
Unload Amazon Redshift data as Parquet to Amazon S3 data lakes for faster sharing and analytics

Parquet is an open data format supported by Amazon EMR, Athena, and Amazon Redshift

Amazon Redshift now supports exporting data to Amazon S3 in Parquet format

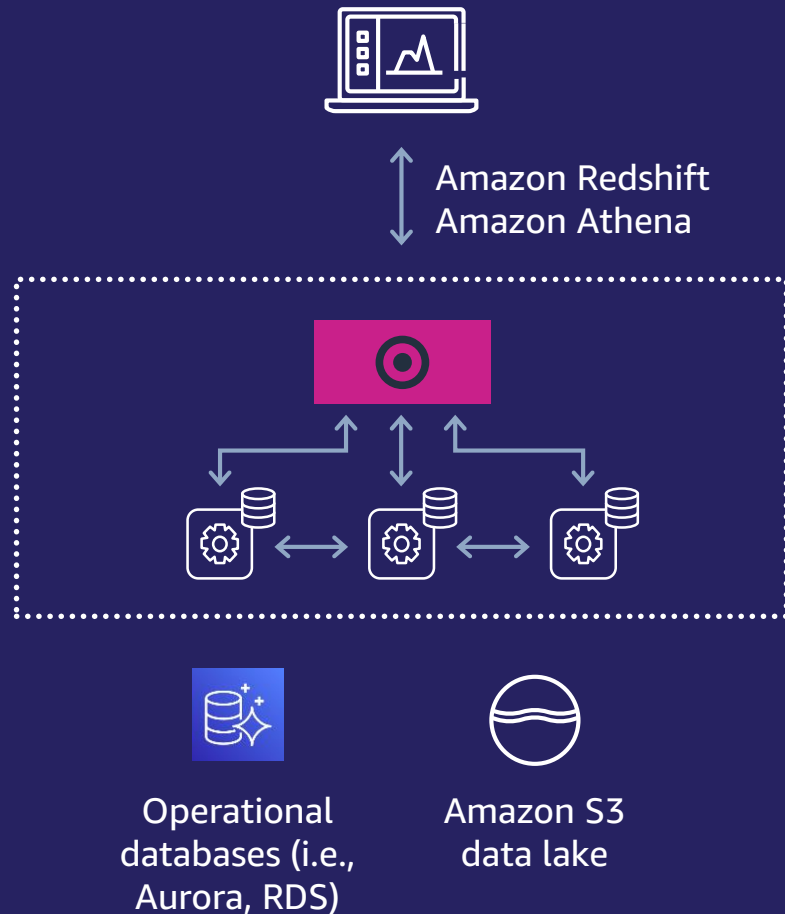
Use SQL with Amazon Redshift's Unload command to export data in Parquet format

Unloaded data is automatically registered in AWS Glue Data Catalog



Federated query in Amazon Redshift and Athena

Unified analytics across databases, data warehouse, and data lake



Integrate operational database with data warehouse and Amazon S3 data lake

Analytics on operational data without data movement and ETL delays

Flexible and easy way to ingest data, avoiding complex ETL pipelines

*Other sources available in Amazon Athena: Amazon ElastiCache for Redis, Amazon DocumentDB, Amazon DynamoDB, HBase in Amazon EMR

AWS provides end-end capabilities for data analytics



Data catalog
Crawls and catalogs your data; discover, prepare, and combine data for analytics and ML

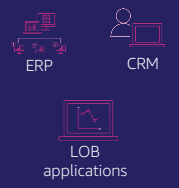
AWS Glue



Data governance
Centralized authorization layer to define data sources and data access & security policies

AWS Lake Formation

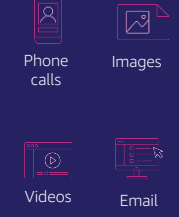
Structured data
Data that are highly normalized with common schema and stored in relational databases, powering transactional line-of-business applications



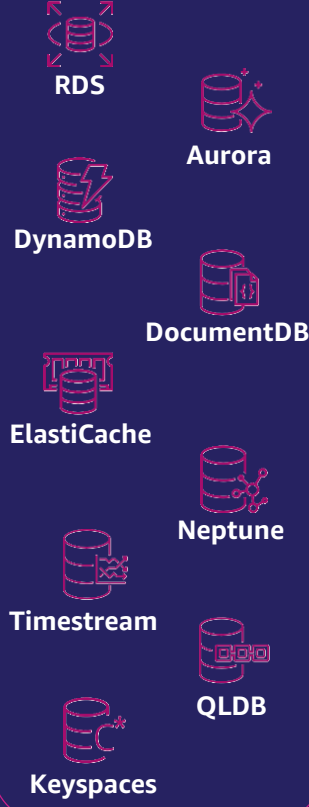
Semistructured data
Data that contain identifiers without conforming to a predefined schema




Unstructured data
Data that do not conform to a data model and are typically stored as individual files




Data collection
Purpose-built databases for different workloads



Batch load
Extracts data from various data sources at periodic intervals and moves them to the data lake



Streaming
Ingests data that are generated from multiple sources such as log files, telemetry, mobile applications, and social networks



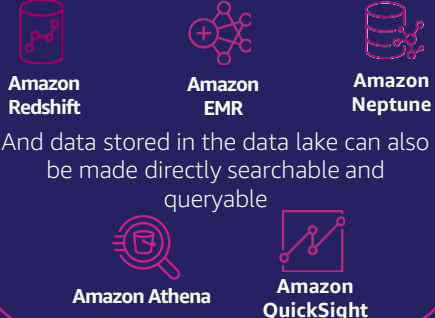
Amazon S3 data lake
Cloud-scale centralized and scalable architecture that enables enterprise data science




Amazon S3

Analytics
Leverage data warehouses, Spark, and graph databases to gain insights from your data

And data stored in the data lake can also be made directly searchable and queryable



Machine Learning
Storing data in an Amazon S3 data lake enables customers to leverage predictive or prescriptive analytics; perform ad-hoc analyses; and use AI/ML for automation and efficiency



Customer Examples

FINRA built an AWS data lake to enable markets surveillance at scale



“ We got some huge pleasant surprises out of [going all in on AWS] that we weren’t expecting at all. First of those is amazing performance improvements. On average, 400 times improvement to interactive queries. The investigative capacity to our surveillance team has expanded dramatically ”

– Steve Randich, CIO, FINRA



FINRA needed a platform that could ingest, process, and store 36 billion market events on an average day and dynamically scale up to handle 100 billion events on a peak day.



FINRA built a data lake on AWS using Amazon S3 and EMR to store and analyze data from 3,700 broker dealers and 12 exchanges.

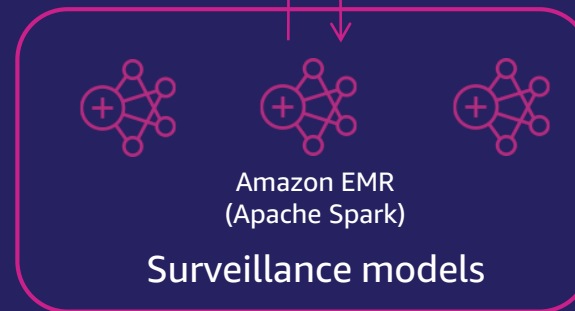
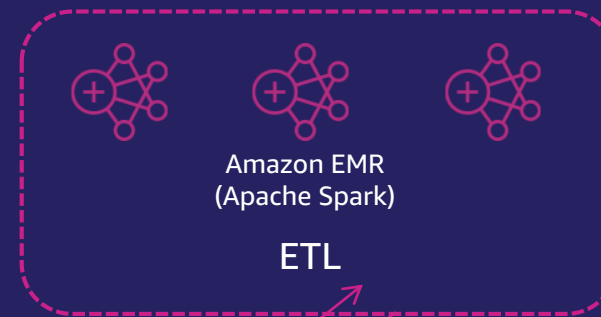


FINRA’s flexible platform can adapt to changing market dynamics while providing analysts with the tools needed to query the data set.

Sample reference architecture for markets surveillance

Data validation

- Use serverless computing for event-driven data validation as files come in to reduce cost and processing times



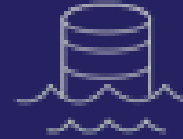
Data ingestion

- Leverage fully managed and scalable services such as Amazon S3 and AWS Transfer for SFTP which automatically adjusts to highly variable market activity

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

Fannie Mae enables machine learning with its data lake on AWS



- Fannie Mae built its enterprise data lake with 100% native AWS services to support more than 3,000 data sets and more than 100 applications.

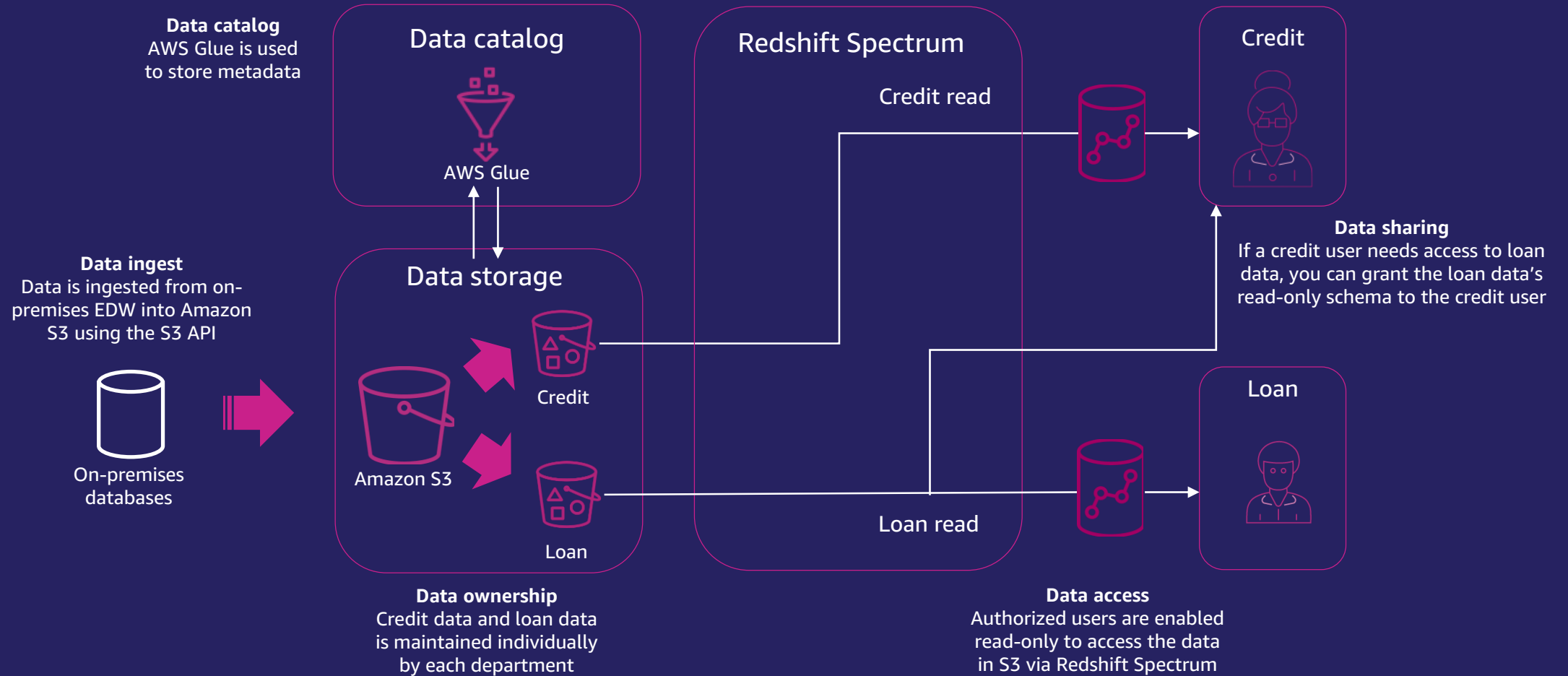


- Fannie Mae's data lake on AWS is the foundation for its machine learning efforts. As a result, Fannie has a highly secure, self-service, and end-to-end traceable ML capability.



- Fannie Mae achieved faster time to market with innovative automated products at a lower cost and has seen improved security and resiliency.

Sample lake house approach using Redshift Spectrum for data sharing



Customer Example – Data Mesh at JPMorgan Chase

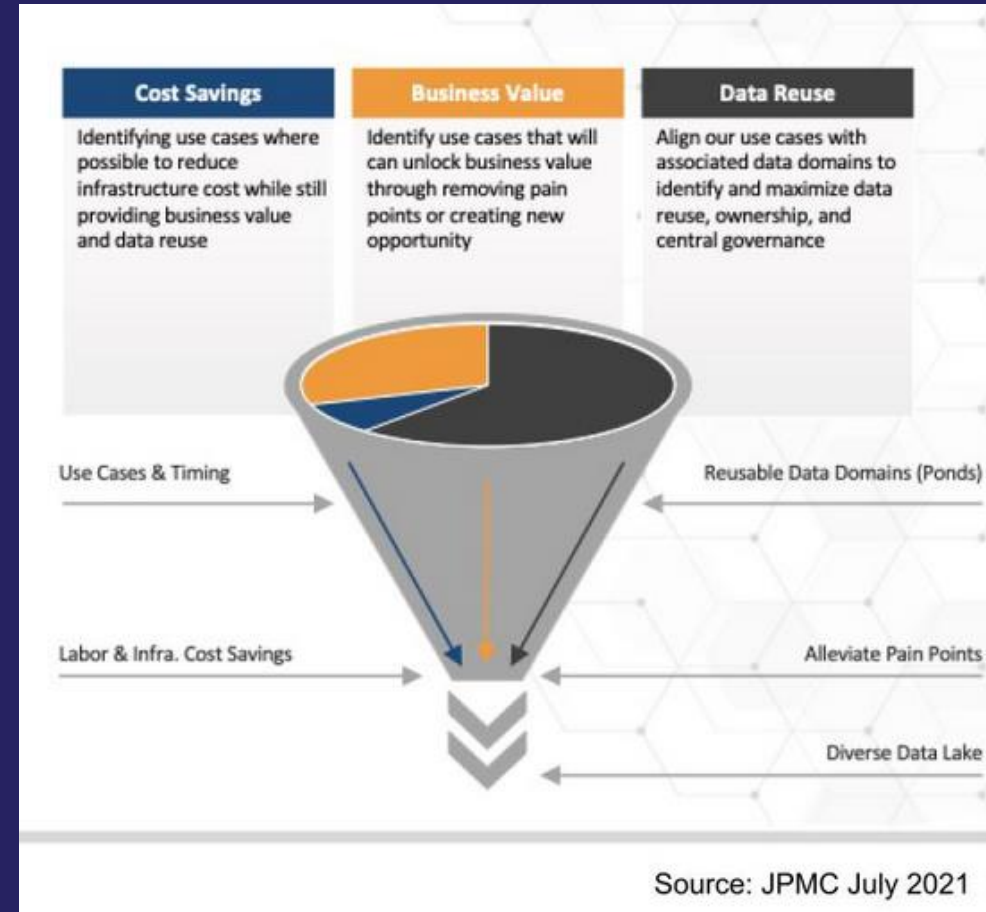
Modernize Data Platform

- Move Beyond Monolithic Data Lake
- Build Loosely Coupled Arch. For Data
- Aggregate fit-for-purpose Data Products
- Distributed Data Pipelines
- Governance and Compliance

Business & Technical Principles:

- Cost Savings
- Business Value – Business (Domain) Use Cases
- Data Reuse (Data Producers/Consumers)

(Source: <https://wikibon.com/breaking-analysis-how-jp-morgan-is-implementing-a-data-mesh-on-the-aws-cloud/>)



Ready to Continue Your Data Journey?

Want to build a data vision and strategy?



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

Jumpstart the data flywheel



Have a strategy and need help executing it?



- ✓ Joint engineering engagements between customers and AWS technical resources
- ✓ Create tangible deliverables to accelerate strategic databases, analytics, and ML initiatives
- ✓ Leave with an architecture, working prototype, path to production, and deeper knowledge of AWS services

Come with an idea, leave with a solution



Thank you!