Democratizing your organization’s data analytics experience

Randy Staton
Solutions Architect

Agenda

Cloud strategies and data gravity

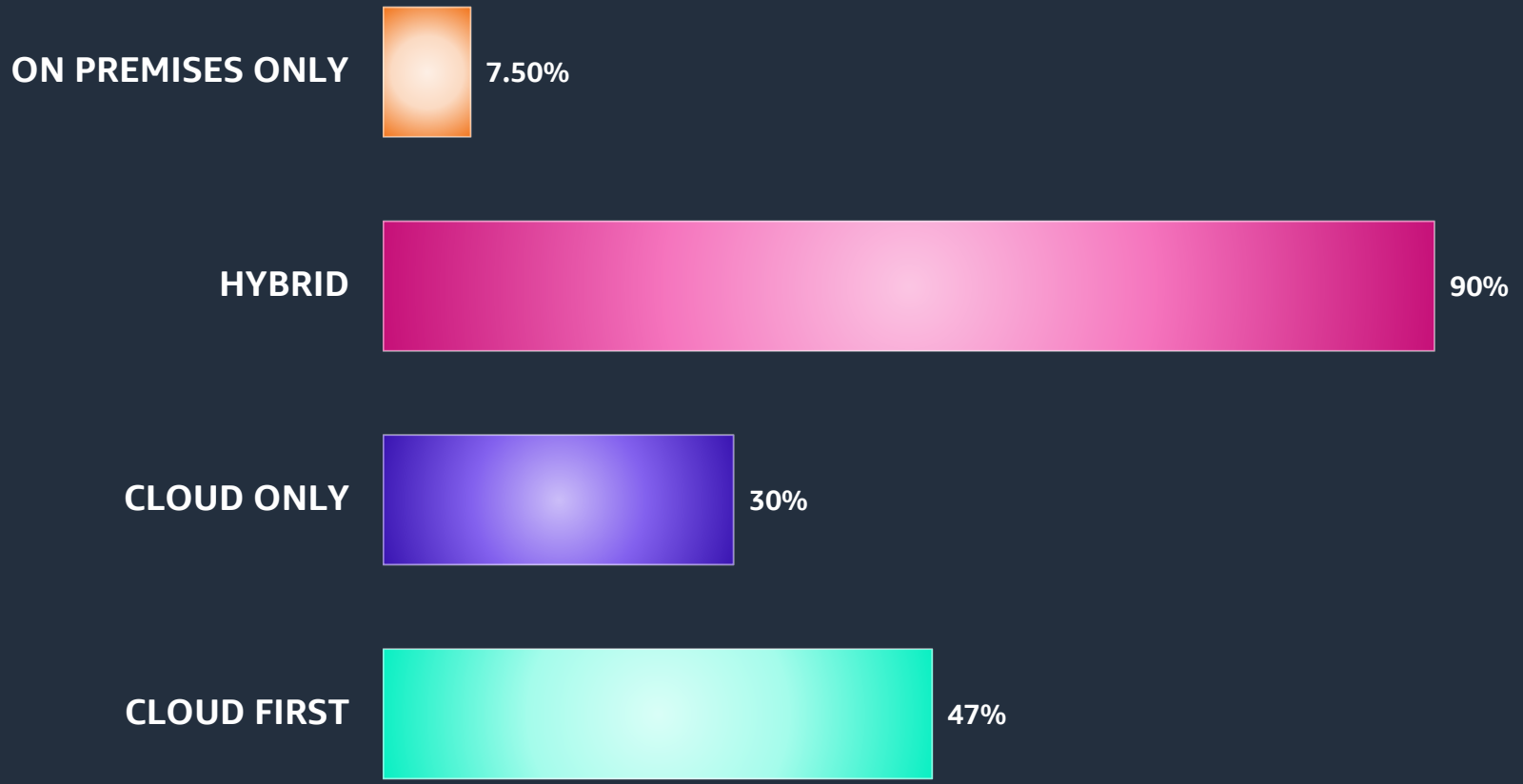
Modern Data Strategy – People, Process, Technology

Democratizing analytics

Ease of use

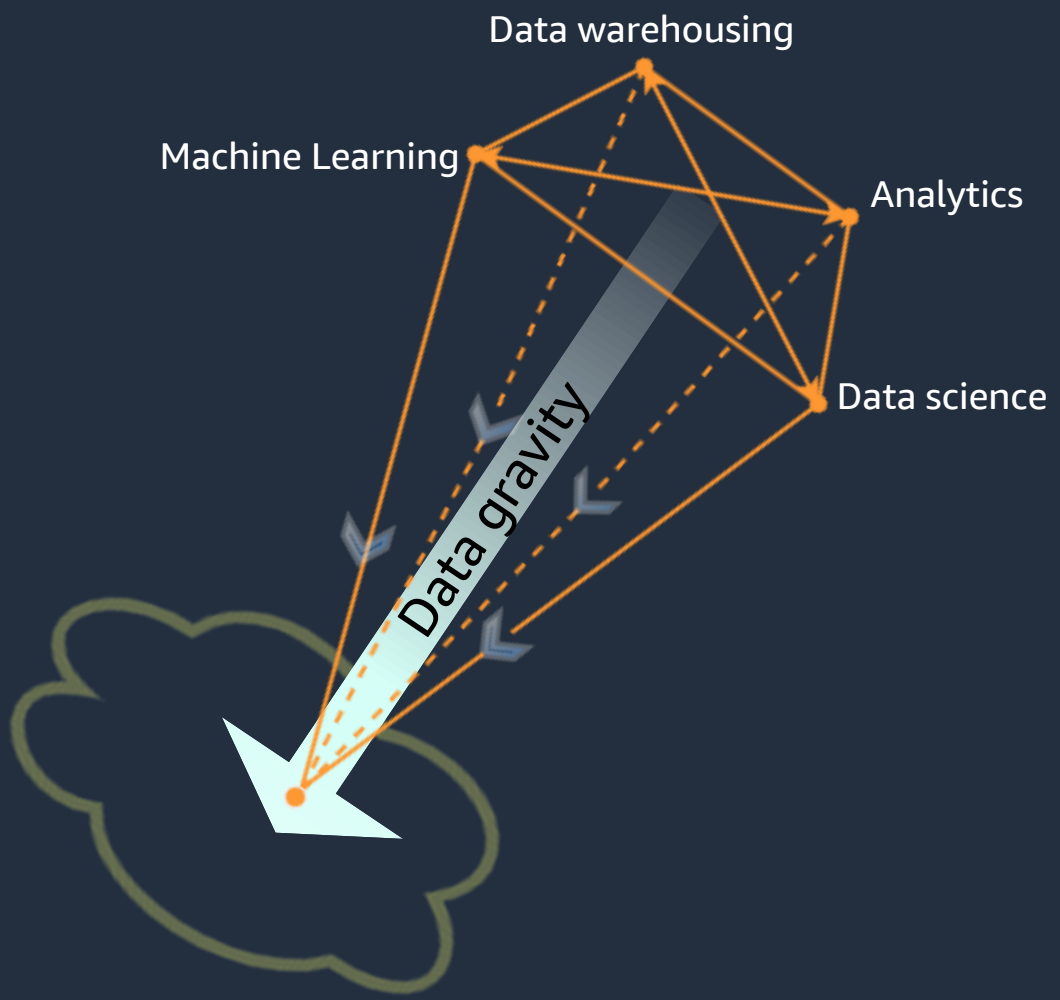
Price performance

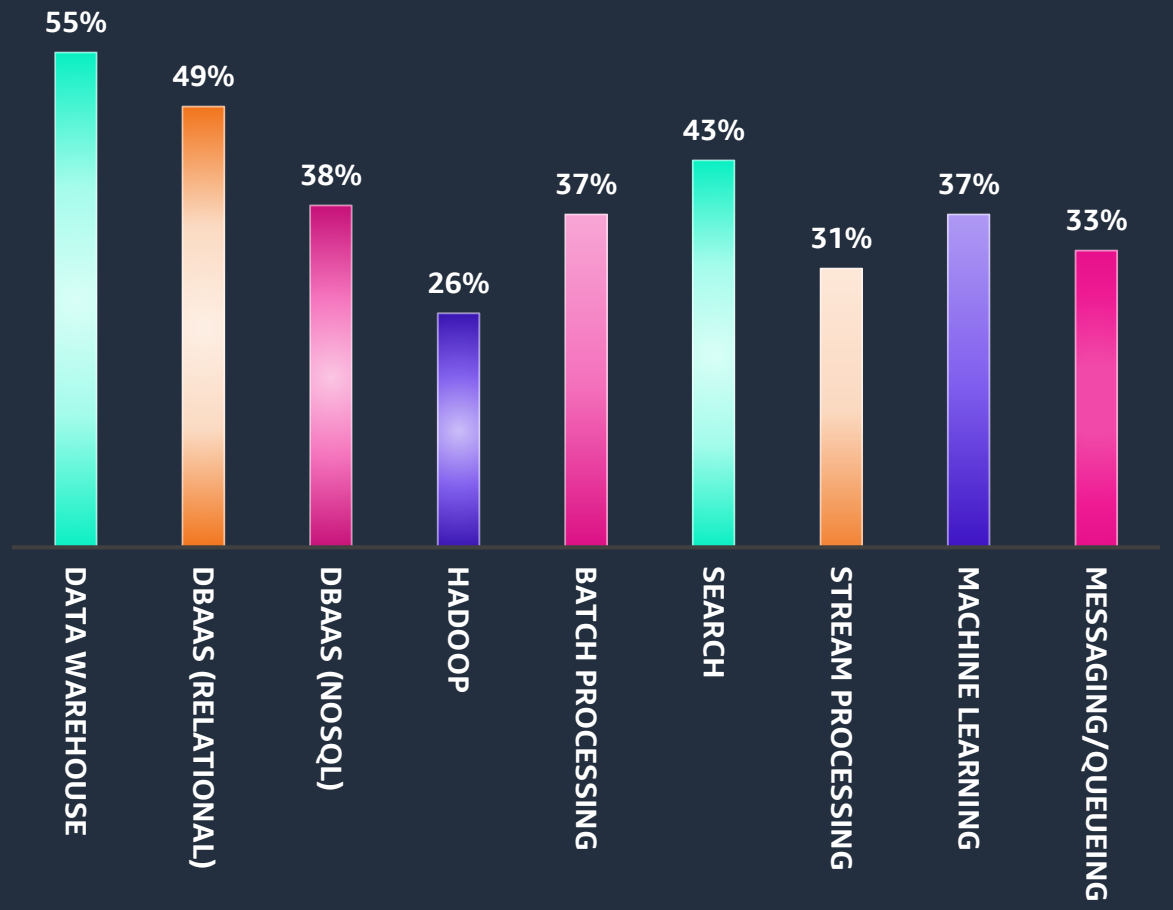
Cloud strategies



Source: O'Reilly cloud adoption

Data gravity





Source: Flexera cloud computing trends

Continued Analytics Cloud Growth

The Big Data and Analytics software and cloud services has reached $90.4B spend in 2021, with 44% deployed in the cloud and the remaining 56% on-premises.

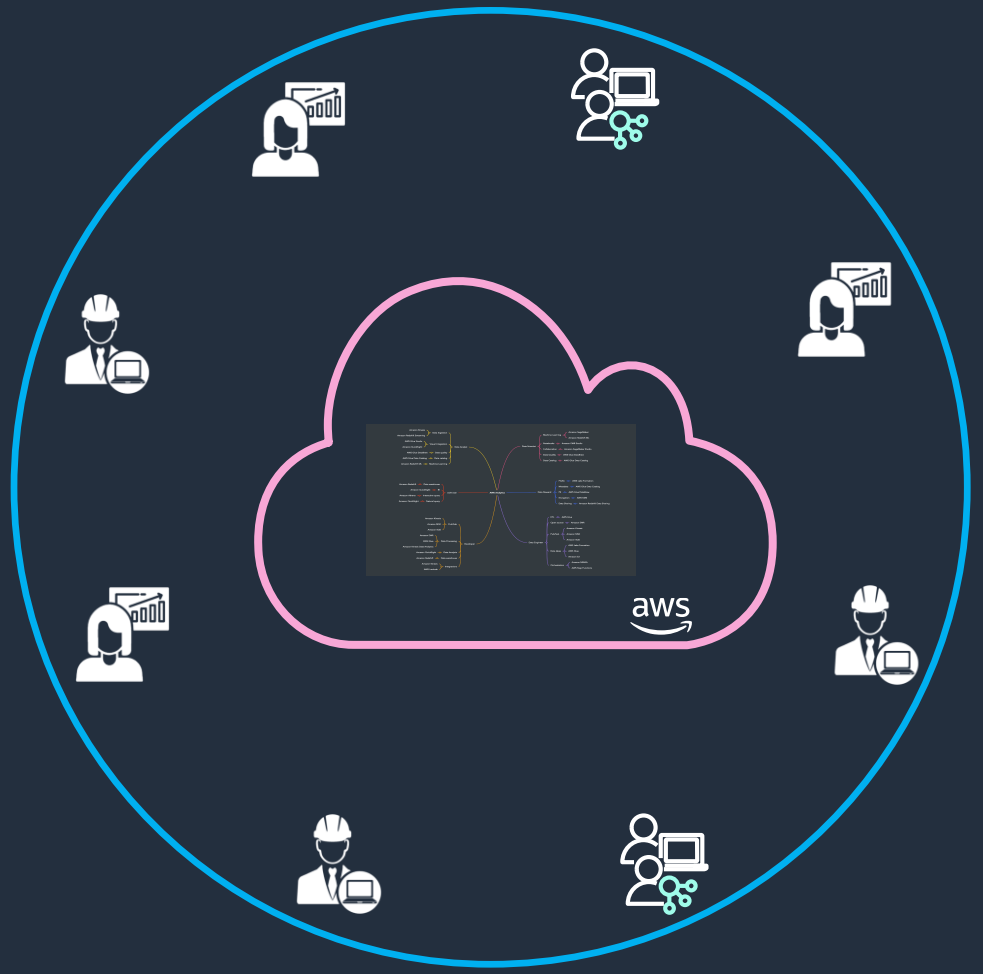
-IDC

Organizations will move more than 70% of their advanced analytics (enriched with AI/ML) to the cloud by 2024.

-Gartner

Modern Data Strategy

People, Process, Technology



Challenges we are hearing from customers

BUSINESS CHALLENGES: Struggle to answer important Qs timely, limited transparency and auditability erodes trust, limited analytics capacity and capabilities to data share

DATA CHALLENGES: No defined governance models, poor data quality, no metadata for audit/compliance/provenance, no data cataloguing

OPERATIONAL CHALLENGES: Manual/ad hoc processes consume resources, inconsistent data manipulation, no 2-way pipelines, variation in data extraction and identity resolution, PII and PHI protection untraceable and vulnerable

FUNCTIONAL CHALLENGES: Roles and responsibilities not codified, not scalable

What is a Data-Driven Organization?

- Data is valued as a **strategic asset** not a by-product of business processes
- Data is **FAIR** (Findable, Accessible, Interoperable, and Reusable)
- Stakeholders throughout the organization are **empowered** to make actionable decisions

What is a Modern Data Strategy?

An **agile** plan of **aligned actions** spanning mindset, people, process, and technology that **accelerates creating value using data** in direct support of strategic objectives

(Mindset + People + Process)
x Technology

Best practices to combat fears

Reasonable Fear	Best Practice	Benefit
Vendor lock-in	Microservices	Agility & extensible
Cool new technology	Component Architect.	Safe to change
Dependent monolith	Decoupling	Operational efficiency
Policy change	Loose Configuration	Flexibility & queuing
Staff size	Event Driven Automation	Cost management
Expanded adoption		Machine learning & automation
Replicability	Containers	Standard practices
Traceability	Infrastructure as Code	Logging & auditing
	Metadata	

Best practices for governance

Reasonable Fear	Best Practice	Benefit
Master Data Mgmt	Data Documentation Attribute Assignment	• Standardization → Normalization • Hierarchy for resolution & truth • Security automation • Ease of update
Security	Data Protection Resilience	• Multiple layers of encryption (at rest, in movement) • ML logging – flag anomalies, PII/PHI, avoid error • No single dependencies • Hardening, replicated data stores • Automated scaling built in elasticity

Best practices for governance

Reasonable Fear	Best Practice	Benefit
Master Data Mgmt	Data Documentation Attribute Assignment	• Standardization → Normalization • Hierarchy for resolution & truth • Security automation • Ease of update
Security	Data Protection Resilience	• Multiple layers of encryption (at rest, in movement) • ML logging – flag anomalies, PII/PHI, avoid error • No single dependencies • Hardening, replicated data stores • Automated scaling built in elasticity

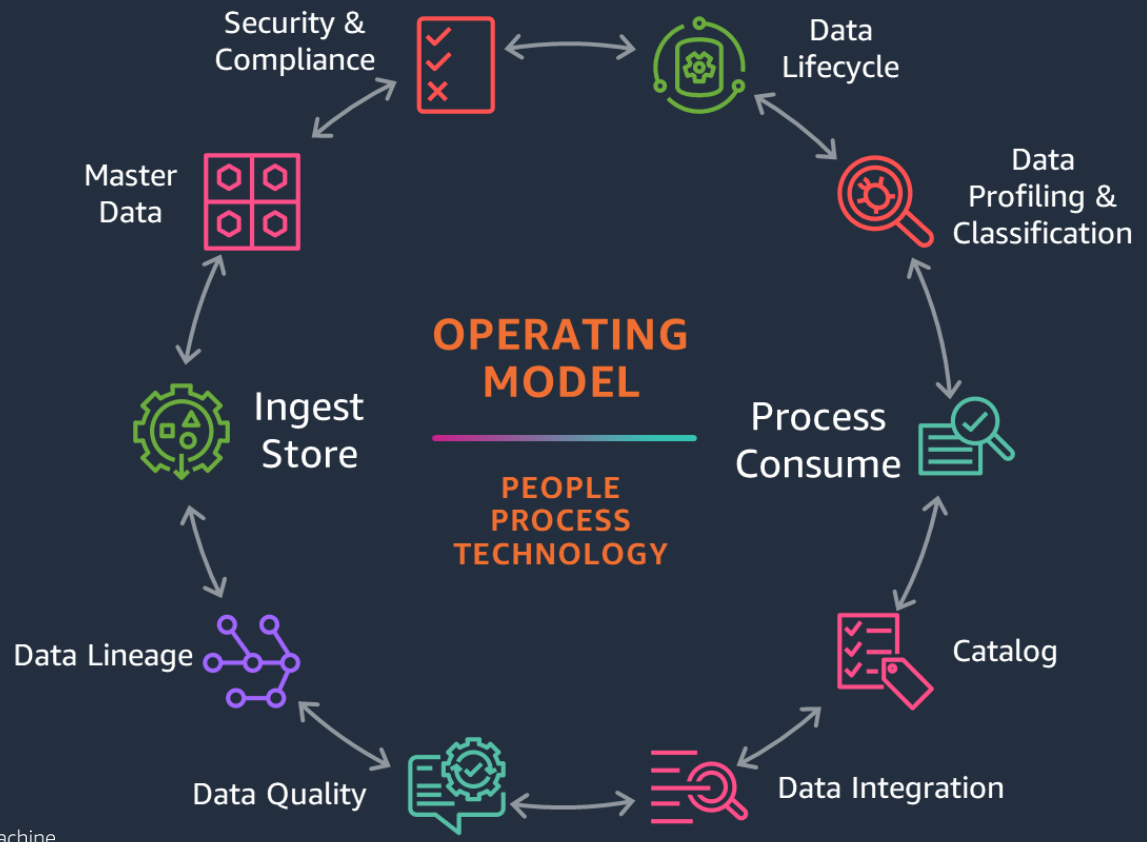
What is a Data Lake? (aka layered & flexible storage)

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale

You can store your data as-is, without having to first structure the data, and then easily run different types of analytics or transformations

AWS approach to data modernization

Data governance is the combination of people, processes, and technology that organizations use to ensure the quality and security of their data throughout its lifecycle



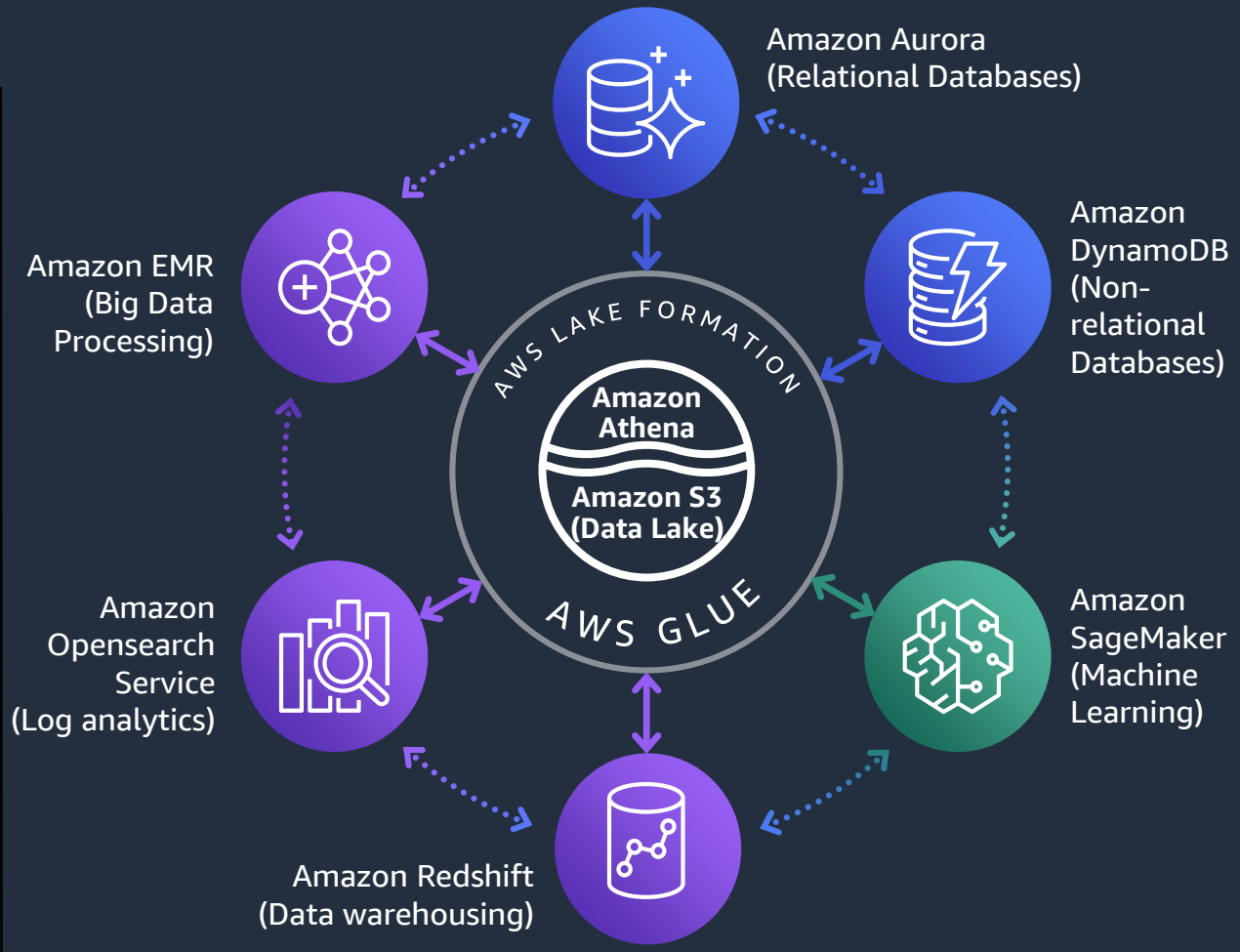
THINK BIG, START SMALL, SCALE FAST

1. Architect data governance to support the wider data strategy
2. Implement incrementally based on business initiatives and use cases that drive the data strategy
3. Further evolve data governance capabilities over time

aws machine learning

Modern Data Capabilities on AWS

- Customers are increasingly moving to data lake architectures
- Amazon Redshift allows you to extend your data warehouse, to your data lake - a *Modern Data Architecture*
- Flexibility to store highly structured, frequently accessed data in Redshift, keep in-frequently used data in S3
- Query seamlessly across both to provide unique insights
- Amazon Redshift is the only data warehouse that extends your queries to your Amazon S3 data lake without moving data



Customers moving to modern data architectures

Redshift enables you to have a modern data architecture

Modern Data Layers

THE STEPS TO ESTABLISH A MODERN DATA ARCHITECTURE WITH AWS LAKE FORMATION

Collection

Ingest data through multiple methods

Lambda, Serverless, API, SFTP, Transfer Family, Streaming

Land all data in raw S3 buckets

Raw buckets secured via LakeFormation

Discovery

Identify and standardize the data

Cataloguing, tagging, standardization, metadata creation

Land tagged and prepared data in clean S3 buckets

Use Lake-Formation and Glue Crawlers for discovery

Glue and Glue Studio for Data Discovery and identification

Governance

Use Clean data

2nd Round of tagging- consider attribute > roles based access

Glue for ongoing mastering and governance

Lambda to automate

Integrate access with AD or external identity providers (Okta)

Use LF-Tags for governance

Consider layers/complexity of governance rules for partner selection

Transform & Store

Create pathways for merges, transformations, and curation of datasets (Cleansed data lake)

Use Glue Databrew, Glue Studio for visual data transformations or integrate with your code

Create repeatable transformation recipes for consistency

Create curated data products stored in different structures or file types for intentional consumption

Consume

Purpose-built datasets and data products for consumption

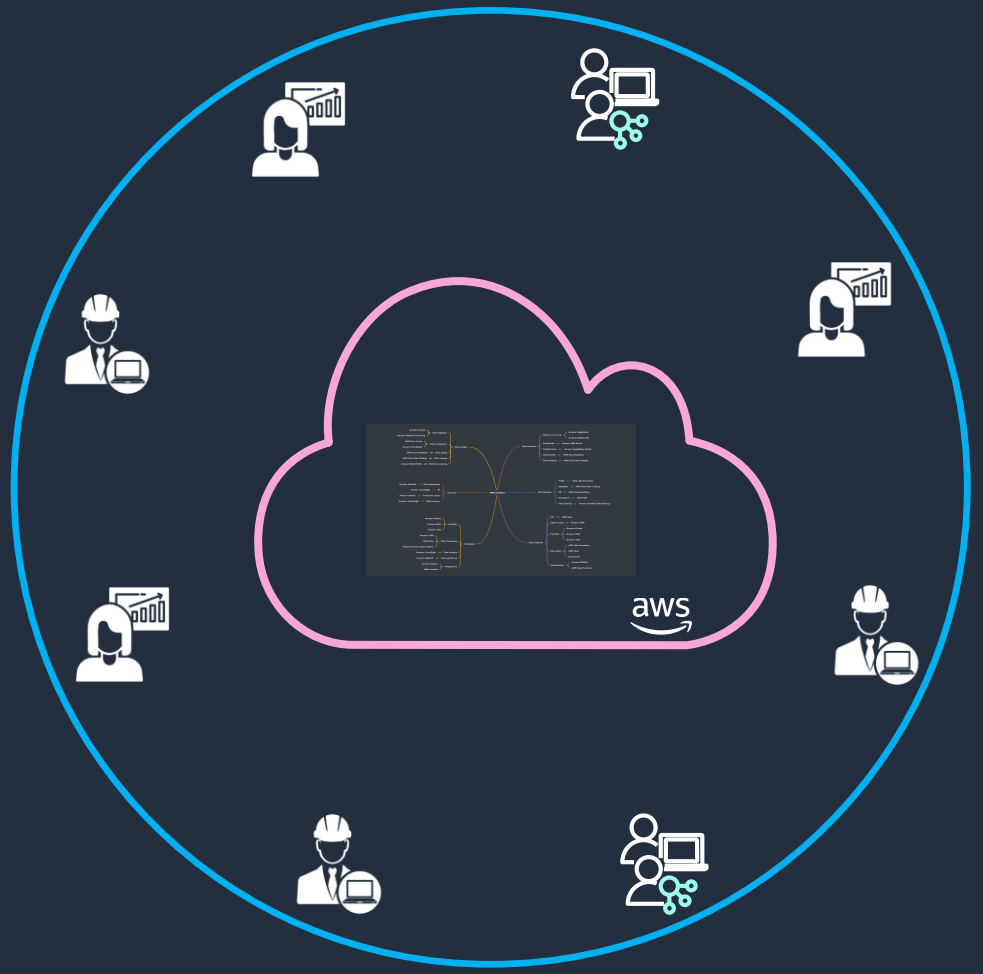
Native, external, AI, visualization & reporting tools, direct query tools, etc.

Curated data products restricted by user, group, role

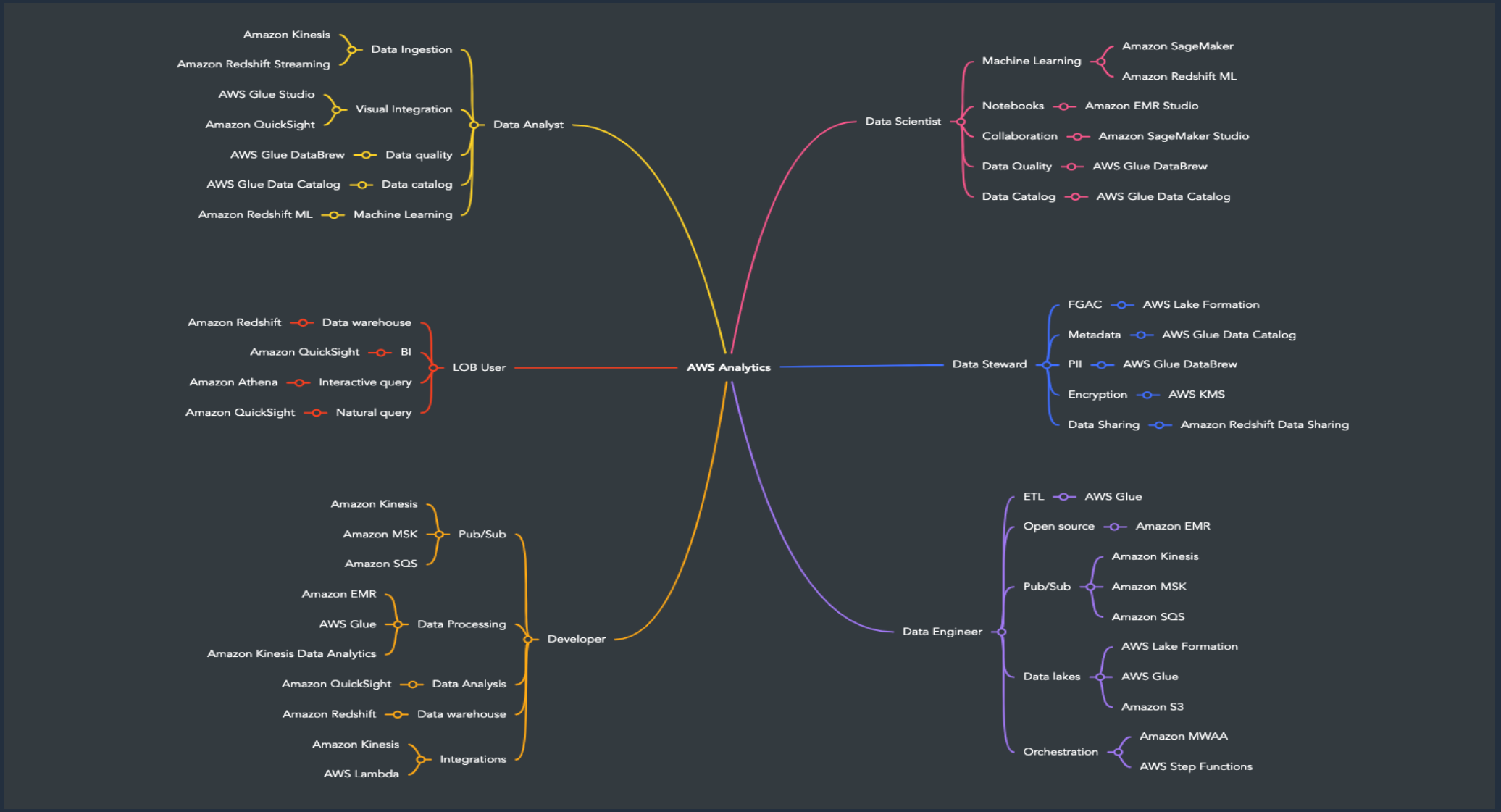
For lower readiness or staff skill- consider partner offerings to help with the stages of modern data architecture

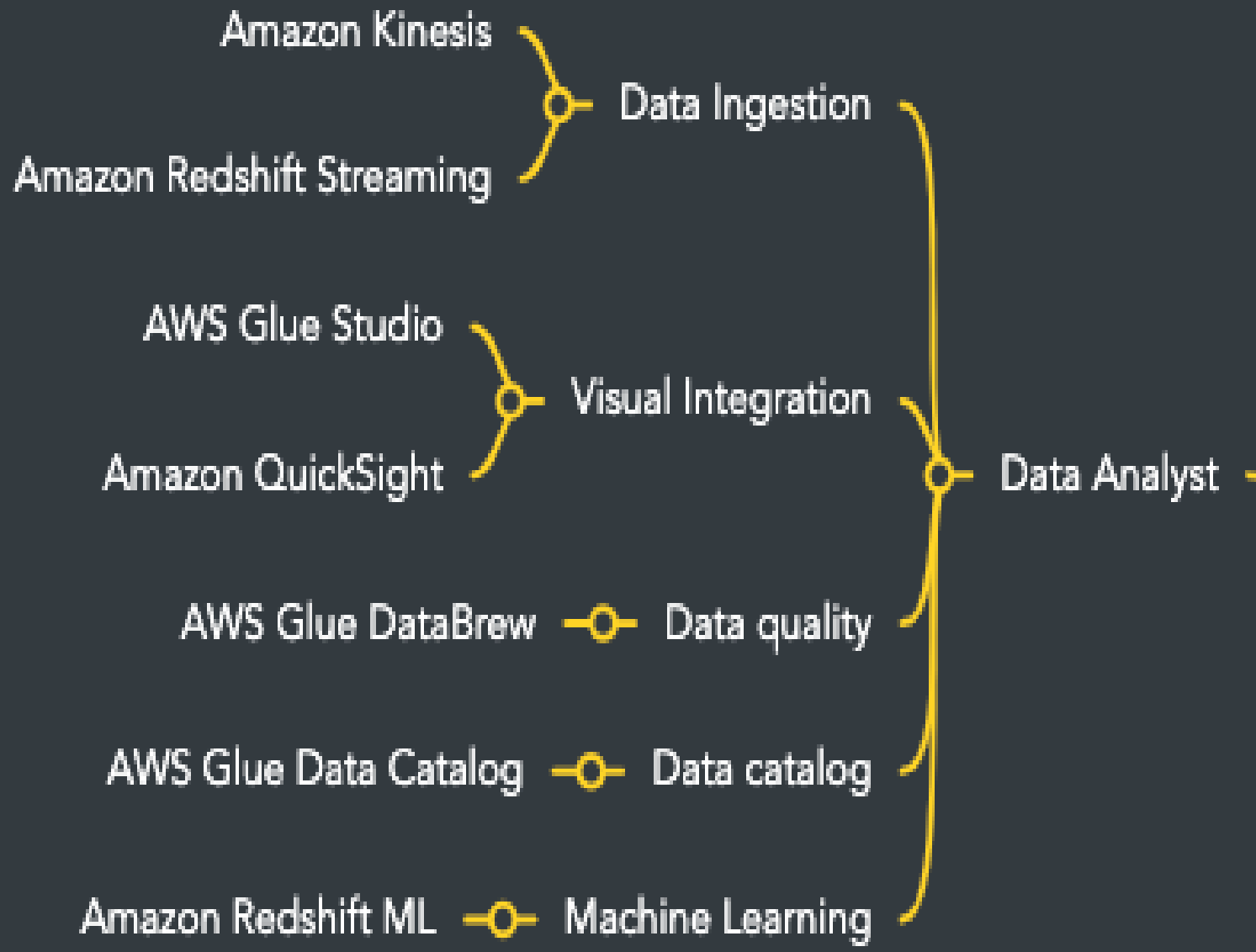
Democratizing analytics

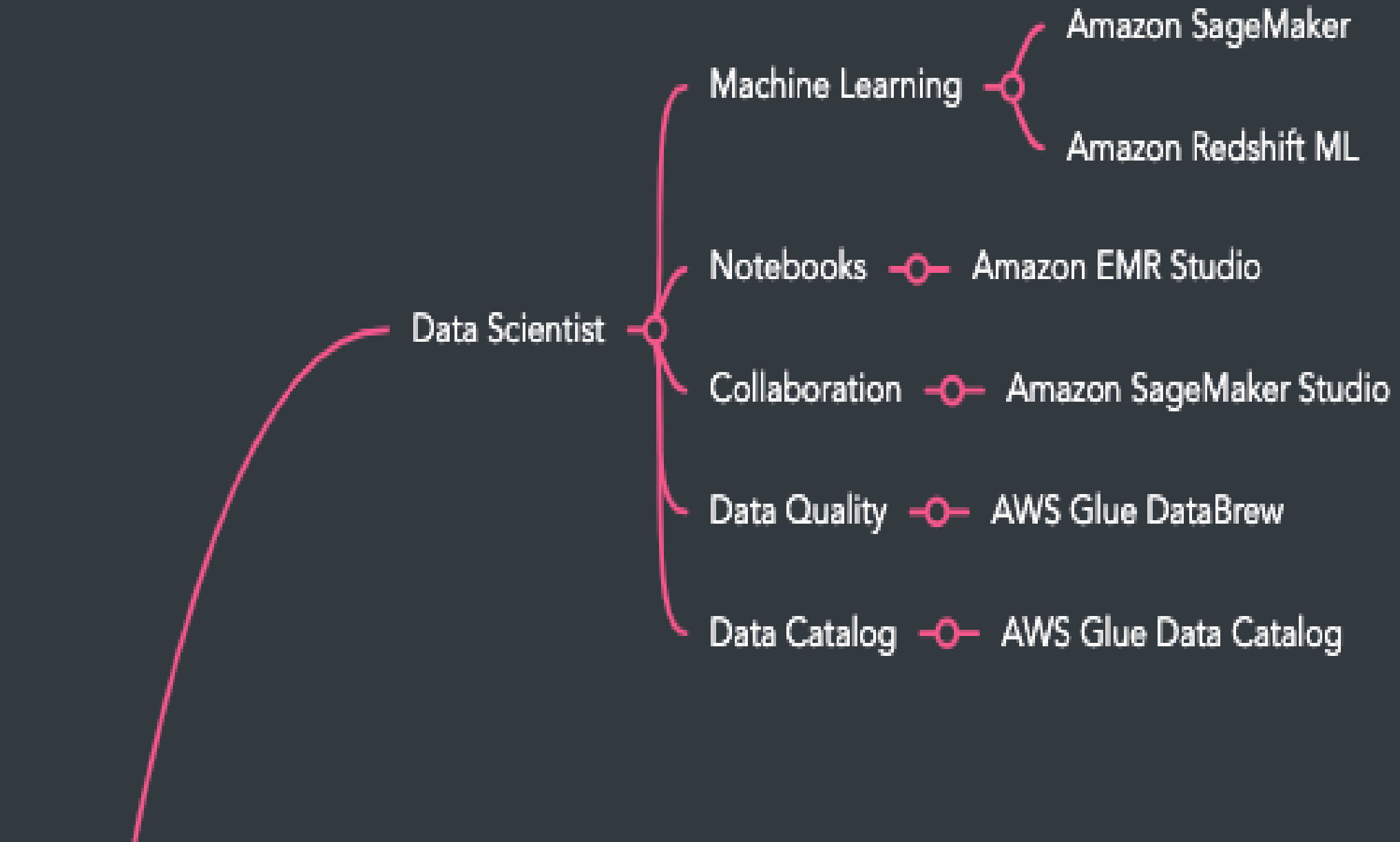
Make analytics available, accessible and affordable

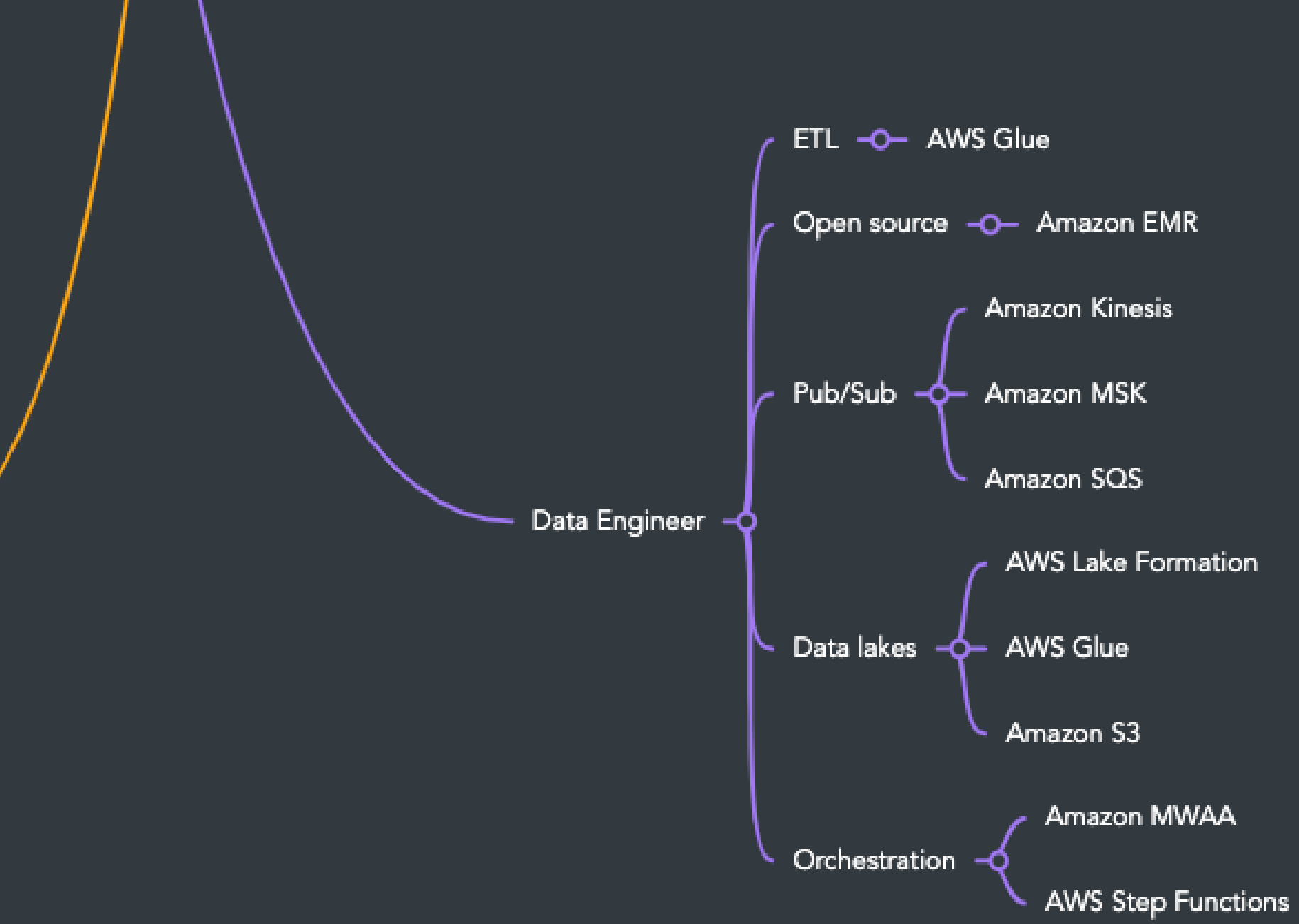


AWS analytics mind-map




Amazon Kinesis
Amazon Redshift Streaming
Data Ingestion
AWS Glue Studio
Amazon QuickSight
Visual Integration
AWS Glue DataBrew
Data quality
AWS Glue Data Catalog
Data catalog
Amazon Redshift ML
Machine Learning
Data Analyst

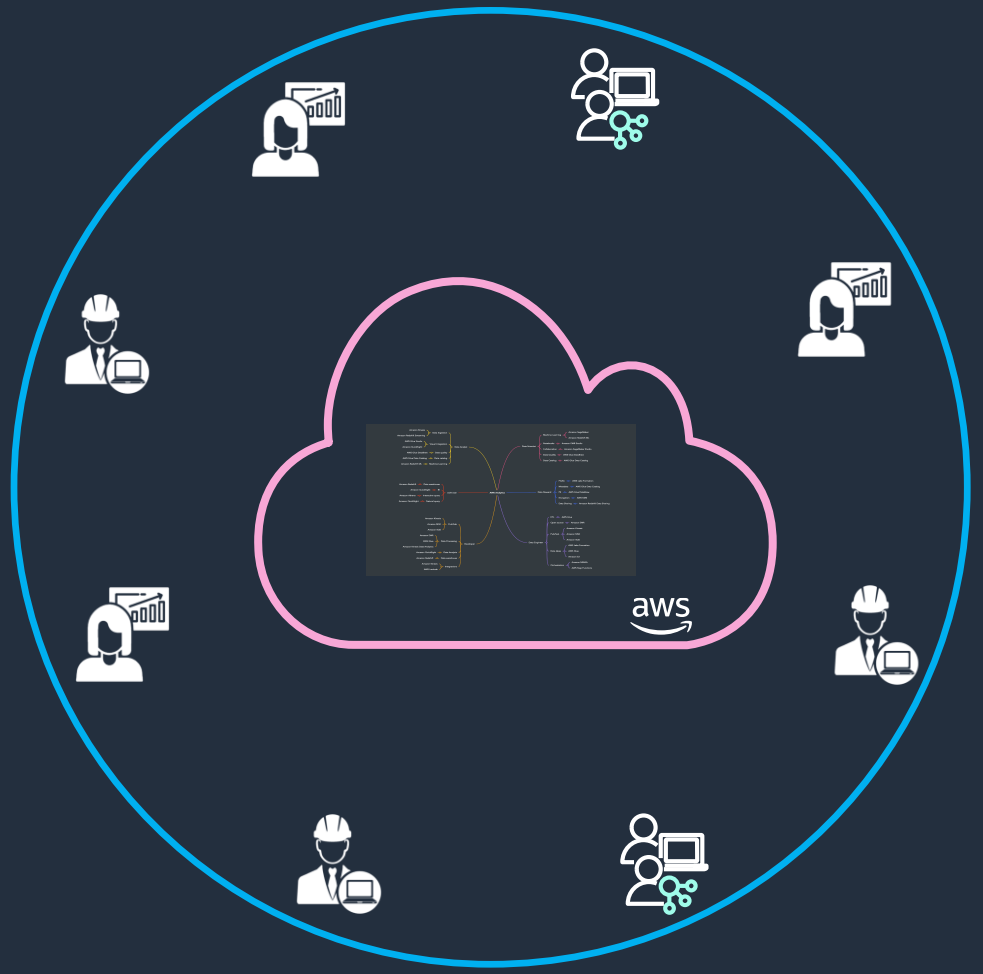

Data Scientist
Machine Learning
Amazon SageMaker
Amazon Redshift ML
Notebooks
Amazon EMR Studio
Collaboration
Amazon SageMaker Studio
Data Quality
AWS Glue DataBrew
Data Catalog
AWS Glue Data Catalog


Data Engineer
ETL
AWS Glue
Open source
Amazon EMR
Pub/Sub
Amazon Kinesis
Amazon MSK
Amazon SQS
Data lakes
AWS Lake Formation
AWS Glue
Amazon S3
Orchestration
Amazon MWAA
AWS Step Functions

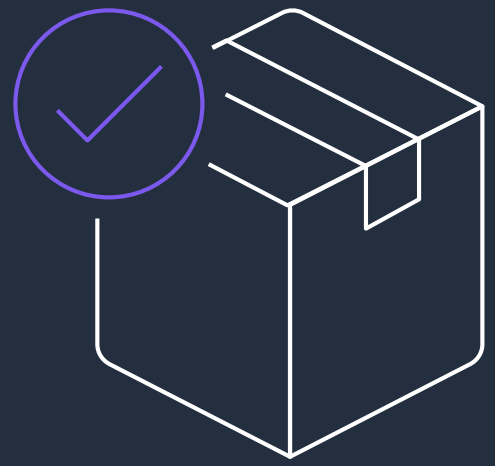

Amazon Kinesis
Amazon MSK
Amazon SQS
Pub/Sub
Amazon EMR
AWS Glue
Amazon Kinesis Data Analytics
Data Processing
Amazon QuickSight
Data Analysis
Amazon Redshift
Data warehouse
Amazon Kinesis
AWS Lambda
Integrations
Developer

Democratizing analytics

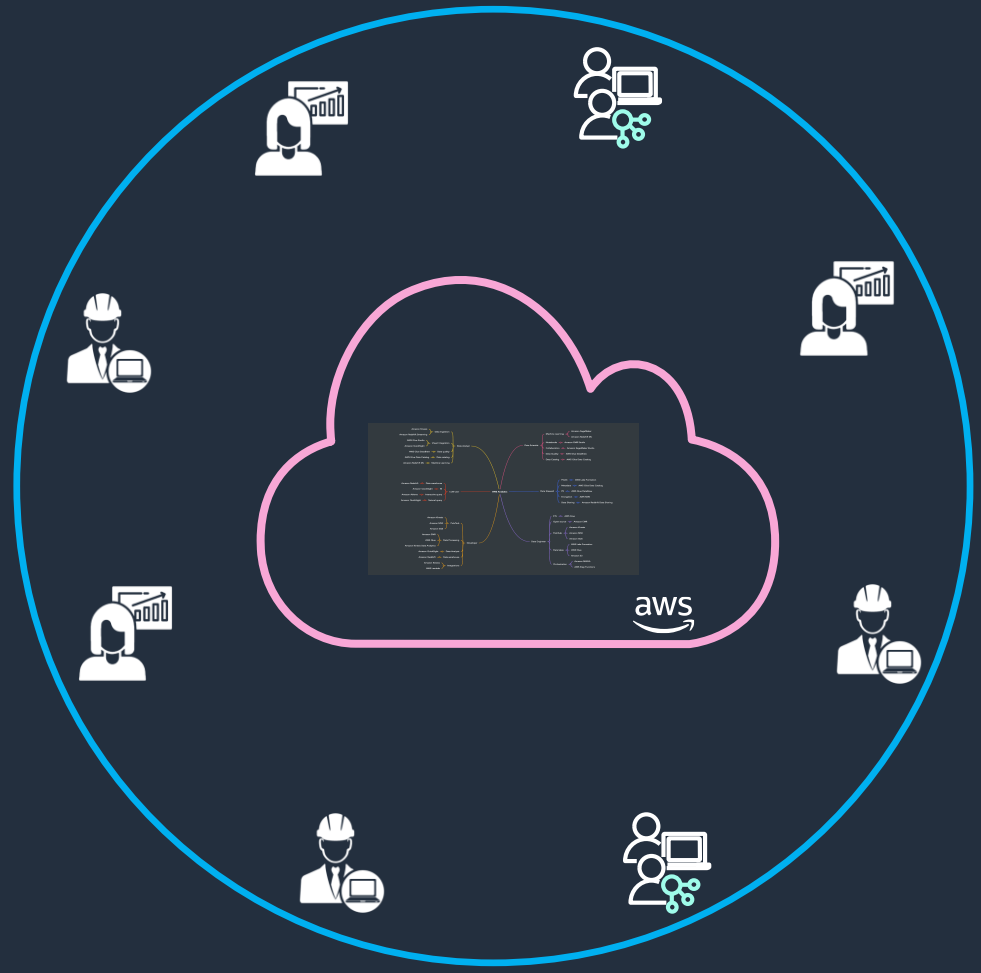
Make analytics available, accessible and affordable



AWS differentiators



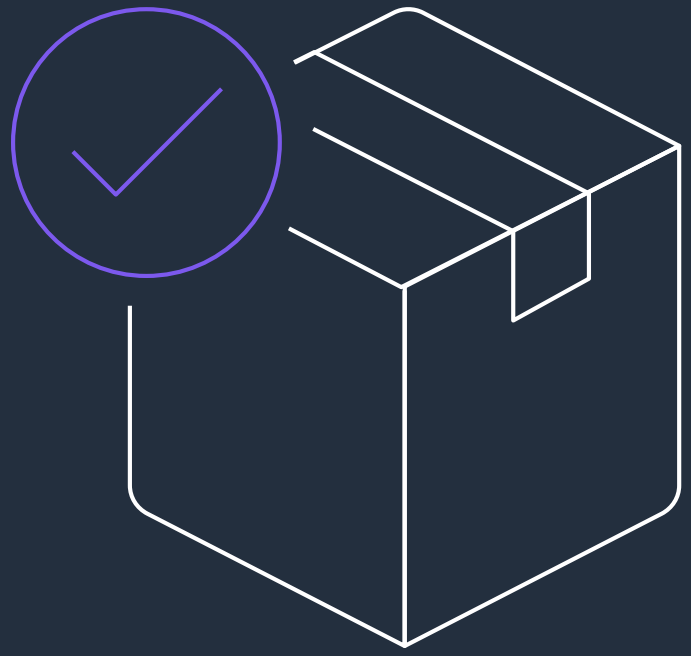
Ease of use



Price performance

Ease of use

Ease of use



- Low barrier to entry
- Reduced operational burden
- Low code / No code experience

Ease of use by AWS

Low barrier
to entry

Intuitive

Start quick / Fail fast

Open to a wider audience

Ease of use by AWS

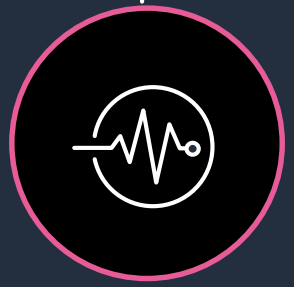
Low barrier to entry

Intuitive

Start quick / Fail fast

Open to a wider audience

Reduced operational burden



Automation

Monitoring

Operations

Ease of use by AWS

Low barrier to entry

Intuitive

Start quick / Fail fast

Open to a wider audience

Reduced operational burden

Automation

Monitoring

Operations

Low code / No code experience

Increased business agility

Rapid development / higher productivity

Reduced OpEx

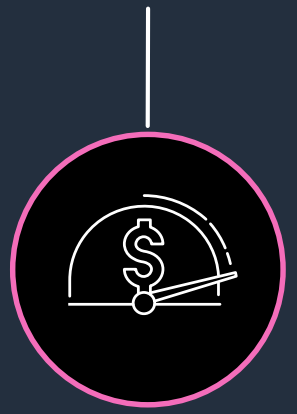
Price performance

Price performance

- Performance pricing
- Do more with less
- Best fit

Price performance by AWS

Performance pricing



Consumption based pricing models

Continuous performance improvements

Price performance by AWS

Performance pricing

Consumption based pricing models

Continuous performance improvements

Do more with less

Iterative feature development

3P and native integration support

Price performance by AWS

Performance pricing

Consumption based pricing models

Continuous performance improvements

Do more with less

Iterative feature development

3P and native integration support

Best fit

Deployment choices

Amazon EMR

BIG DATA ANALYTICS USING OPEN-SOURCE FRAMEWORKS: APACHE SPARK, PRESTO, TRINO, HIVE, HBASE, HUDI AND FLINK

Differentiated performance for Runtimes

Performance optimized runtime for popular frameworks like Spark, Hive, Presto, and Flink with 100% open source API compatibility

Latest open source features

New open source features available within 30 days of release in open source

Best price performance for big data analytics

Reduce cost using EC2 Spot, EMR Managed Scaling and per-second billing

Self service data science

Data Science IDE with EMR Studio and Deep integration with Sagemaker Studio provides ability to use open source UX and frameworks to build, visualize and debug applications

Run workloads on EC2, EKS or on-premises

EMR provides flexibility to run big data workloads on EC2, EKS, and on-premises with Outpost

S3 Data Lake Integration

Fine grained access controls with AWS Lake Formation and Apache Ranger, and Integrations with Apache HUDI and Apache Iceberg to enable S3 data lake use cases

Amazon EMR

3.9x

Faster than standard Apache Spark 3.0 in TPC-DS 3 TB benchmark

4.2x

Faster than standard OSS Trino 388 in TPC-DS 3TB benchmarks

11-16%

Performance improvement with Graviton2 at 20%+ reduced cost

100%

Open-source API compliant

Amazon EMR deployment options

Feature	
Multi-AZ Availability	
OSS frameworks	
Ability to choose OSS version	
Automatic resource scaling	
Ability to choose instance type	
Ability to use EC2 Spot	
Pricing	
Ability to allocate costs	

Amazon EMR deployment options

Feature	Amazon EMR on EC2
Multi-AZ Availability	No (clusters run in a single AZ)
OSS frameworks	Spark, Hive, Presto, Trino, Flink
Ability to choose OSS version	Yes
Automatic resource scaling	Yes
Ability to choose instance type	Yes
Ability to use EC2 Spot	Yes
Pricing	By instance type used
Ability to allocate costs	Per cluster

Amazon EMR deployment options

Feature	Amazon EMR on EC2	Amazon EMR on EKS
Multi-AZ Availability	No (clusters run in a single AZ)	Yes (with multi-AZ EKS clusters)
OSS frameworks	Spark, Hive, Presto, Trino, Flink	Spark
Ability to choose OSS version	Yes	Yes
Automatic resource scaling	Yes	Yes
Ability to choose instance type	Yes	Optional (use EC2 instances or AWS Fargate)
Ability to use EC2 Spot	Yes	Yes
Pricing	By instance type used	By vCPU and memory used
Ability to allocate costs	Per cluster	Per application

Amazon EMR deployment options

Feature	Amazon EMR on EC2	Amazon EMR on EKS	Amazon EMR Serverless
Multi-AZ Availability	No (clusters run in a single AZ)	Yes (with multi-AZ EKS clusters)	Yes (automated job redirection)
OSS frameworks	Spark, Hive, Presto, Trino, Flink	Spark	Spark, Hive
Ability to choose OSS version	Yes	Yes	Yes
Automatic resource scaling	Yes	Yes	Yes
Ability to choose instance type	Yes	Optional (use EC2 instances or AWS Fargate)	No
Ability to use EC2 Spot	Yes	Yes	No
Pricing	By instance type used	By vCPU and memory used	By vCPU and memory used
Ability to allocate costs	Per cluster	Per application	Per application or per job

Amazon Athena

SERVERLESS

ZERO setup cost

Serverless: zero infrastructure, zero administration

PAY PER QUERY

Pay only for queries run

$5/TB

Save **30%–90%** on per-query costs through compression

OPEN AND FLEXIBLE

ANSI SQL

JDBC/ODBC drivers

Multiple formats, compression types, and complex joins and data types

EASY TO USE

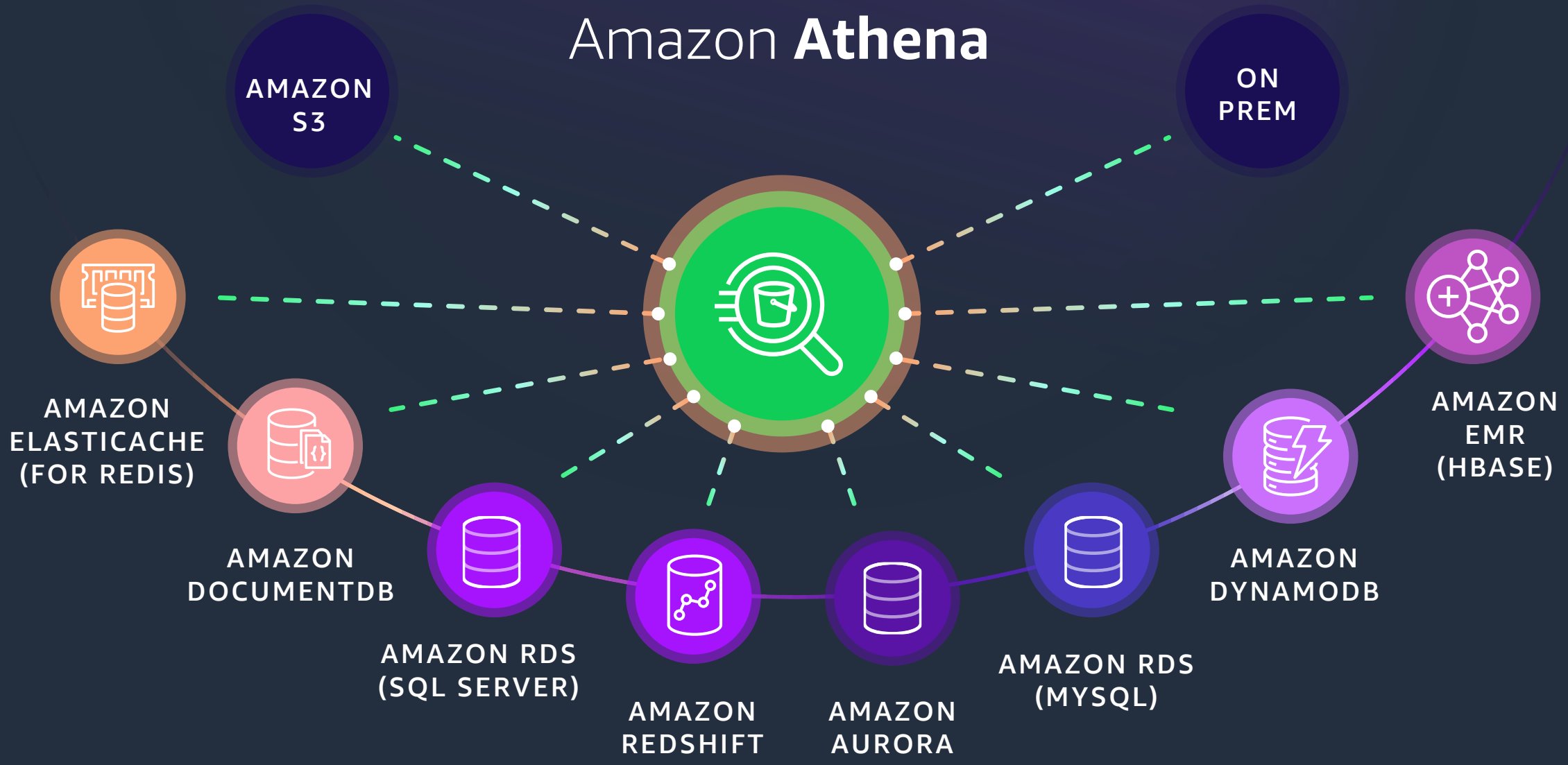
Point to S3 and start querying

DDL operations

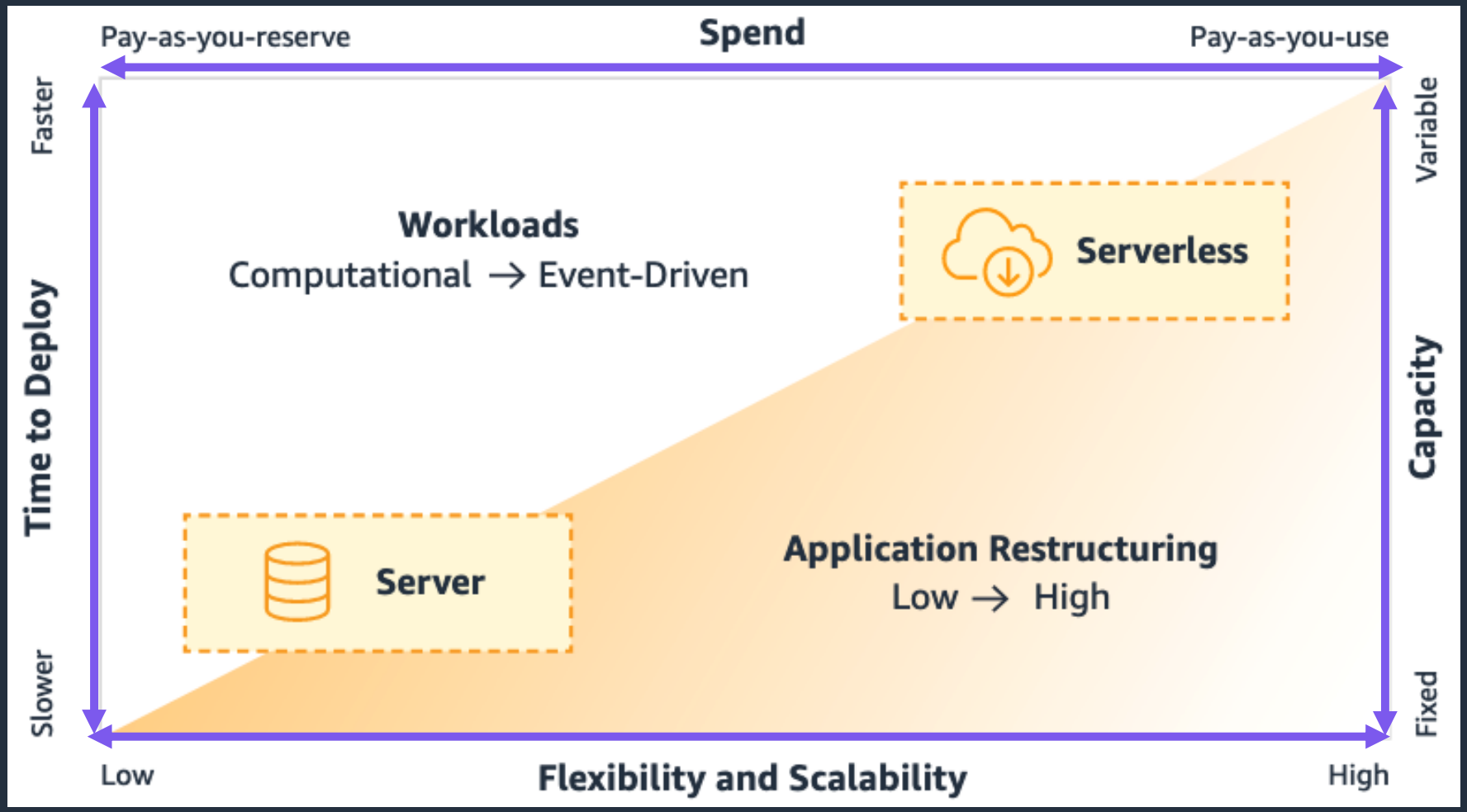
Query concurrency

Integrated data connectors

Amazon **Athena**

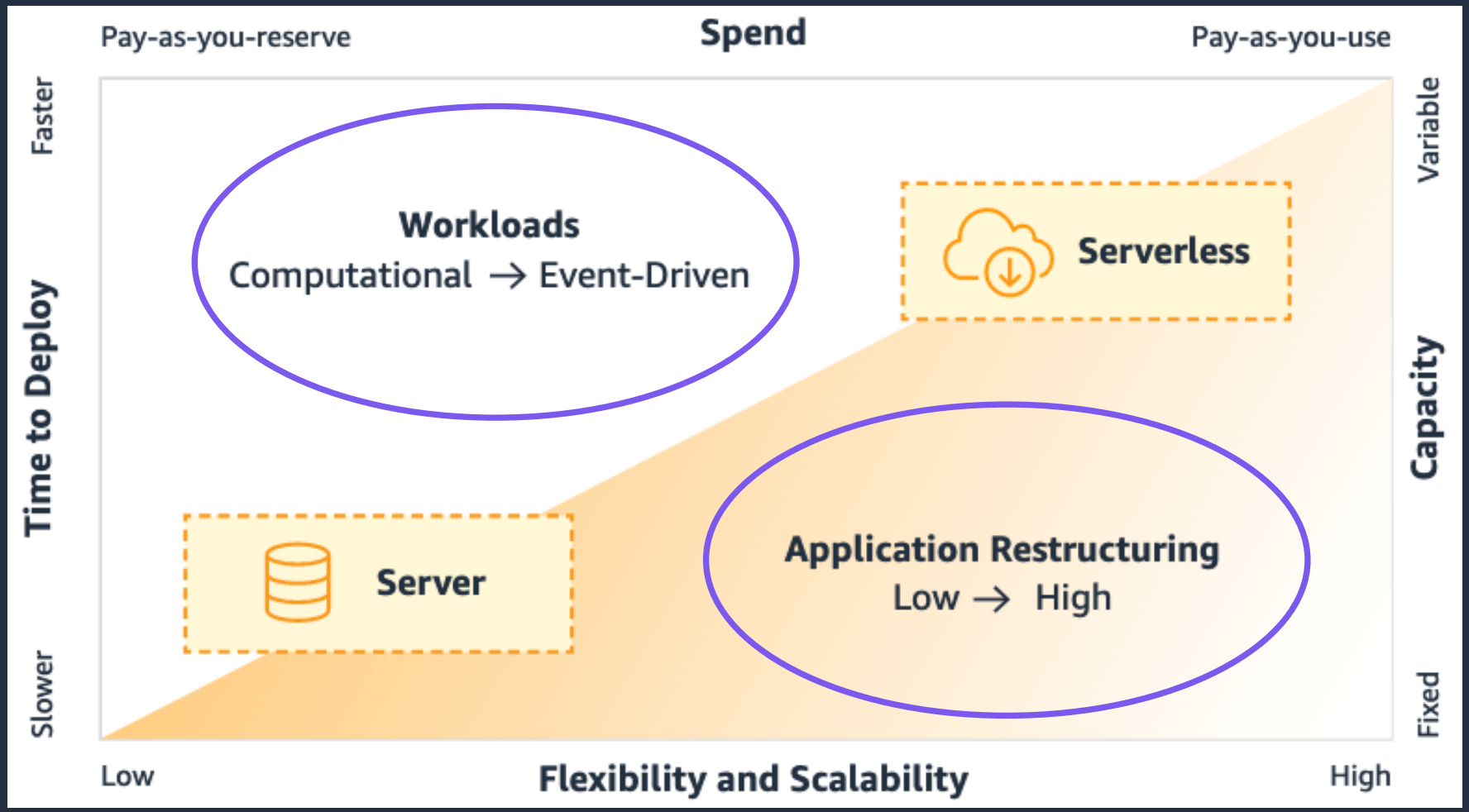


Serverless TCO



Source: Deloitte

Serverless TCO



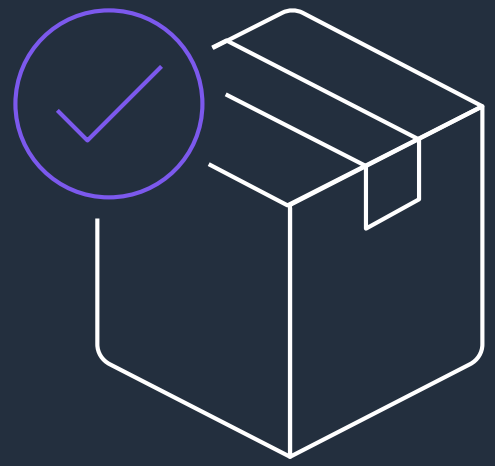
Source: Deloitte

Serverless data analytics on AWS

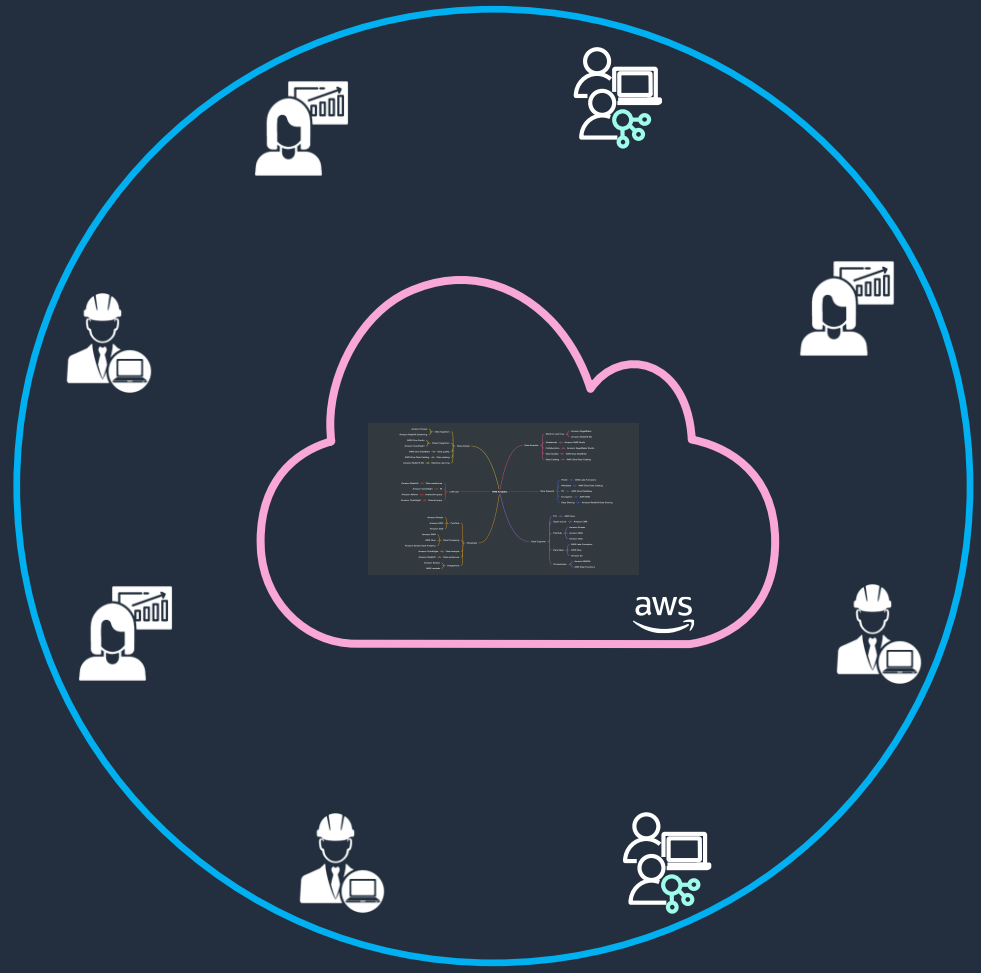
AWS has the **most serverless options** for data analytics in the cloud

INTERACTIVE QUERY	BIG DATA PROCESSING	REAL-TIME ANALYTICS	REAL-TIME ANALYTICS	DATA WAREHOUSING	DATA INTEGRATION	DATA VISUALIZATION	DATA LAKE SETUP MANAGEMENT AND GOVERNANCE
							
AMAZON ATHENA	AMAZON EMR	AMAZON MSK	AMAZON KINESIS	AMAZON REDSHIFT	AWS GLUE	AMAZON QUICKSIGHT	AWS LAKE FORMATION

AWS differentiators



Ease of use



Price performance

Request for Survey

Track: Data and Analysis Track

Topic: Democratizing your organization's data analytics experience