AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT301

Democratizing your organization's data analytics experience

Imtiaz (Taz) Sayed (he/him)

WW Analytics Tech Leader
AWS

Adam Driver (he/him)

WW Analytics SA Leader
AWS

Agenda

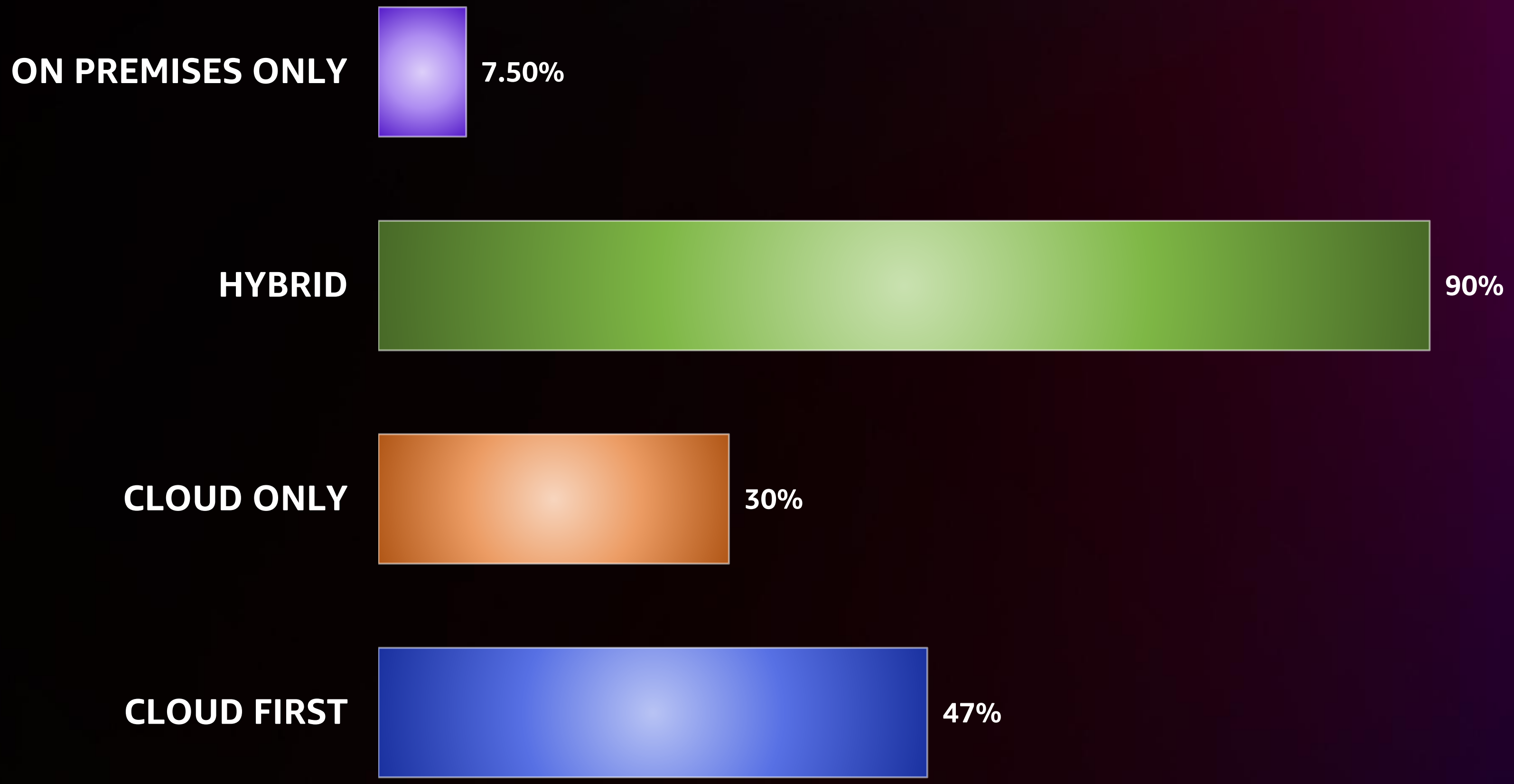
Cloud strategies and data gravity

Democratizing analytics

Ease of use

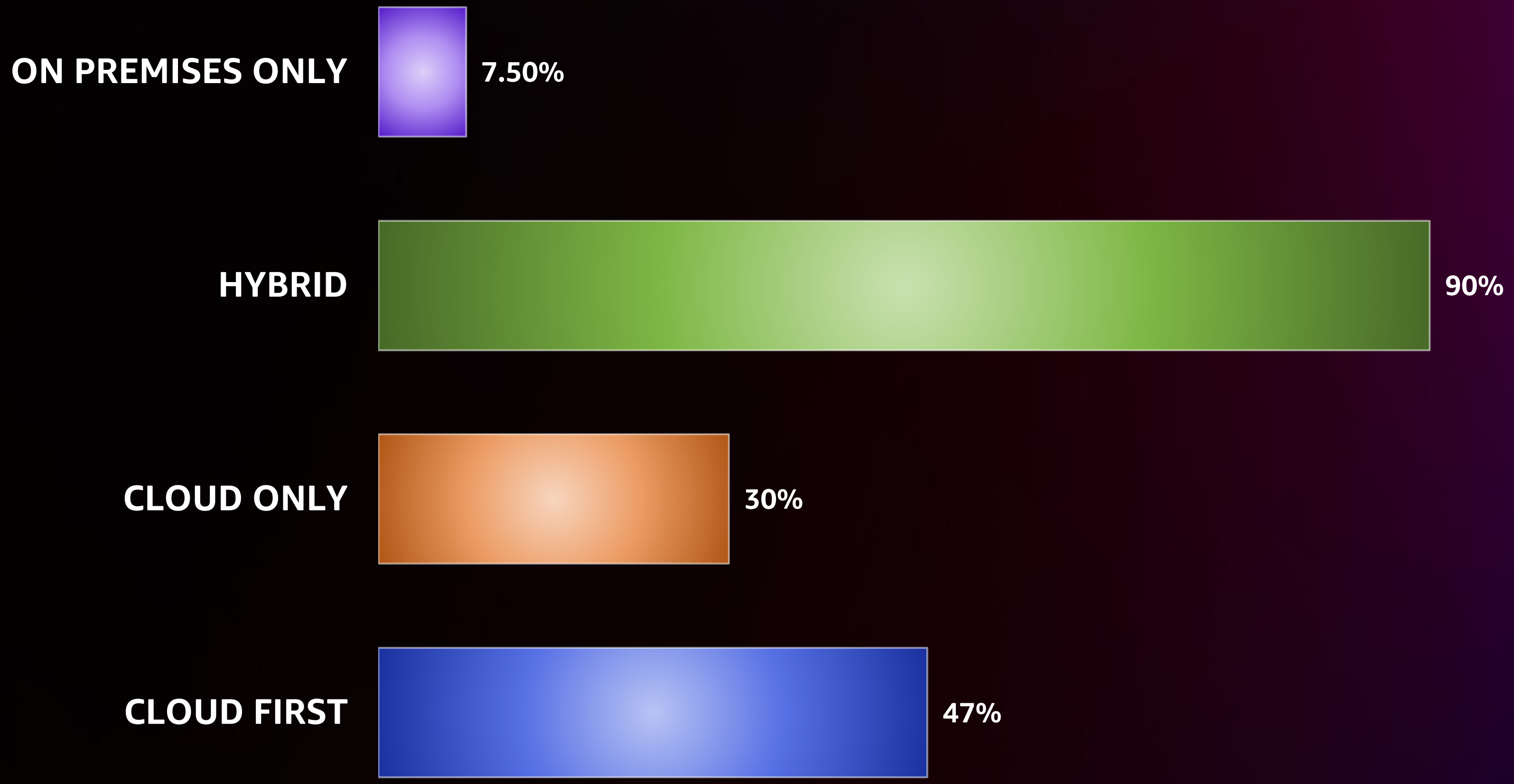
Price performance

Cloud strategies



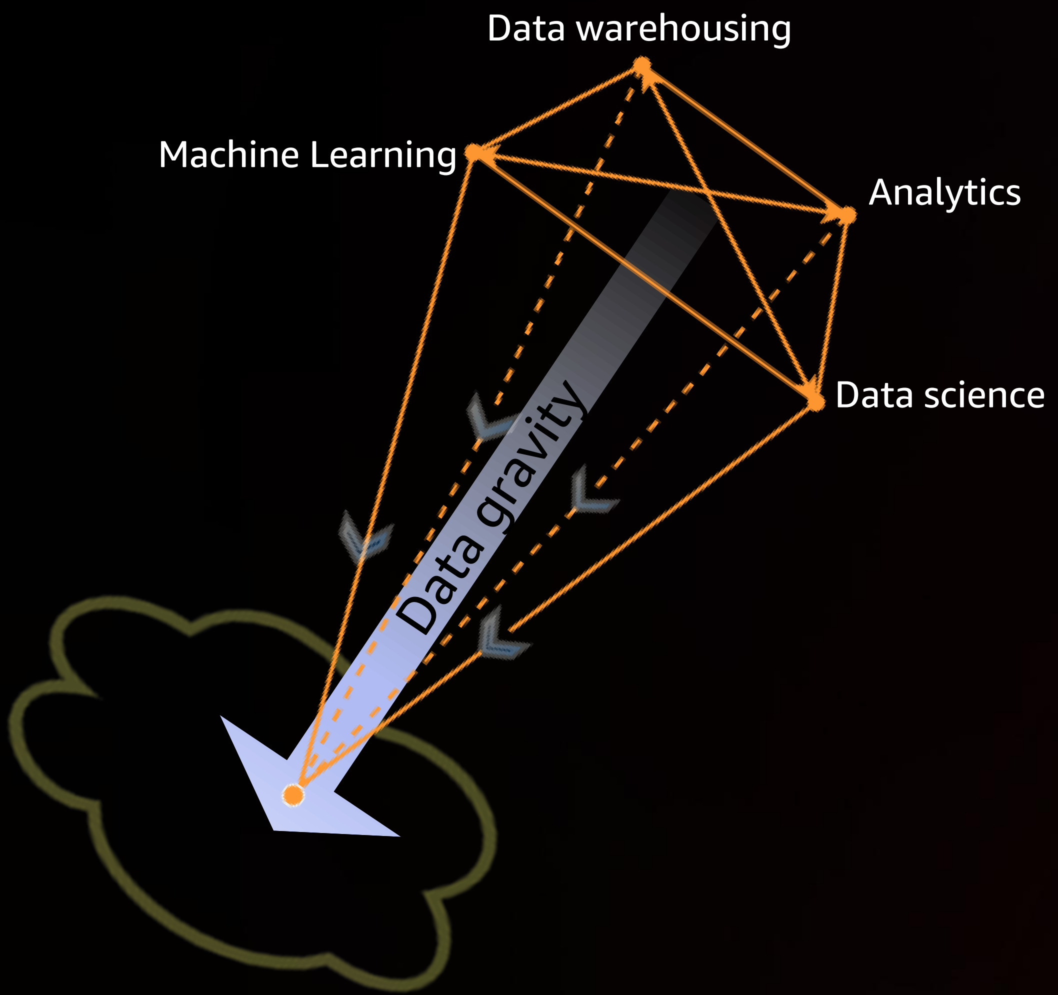
Source: O'Reilly cloud adoption

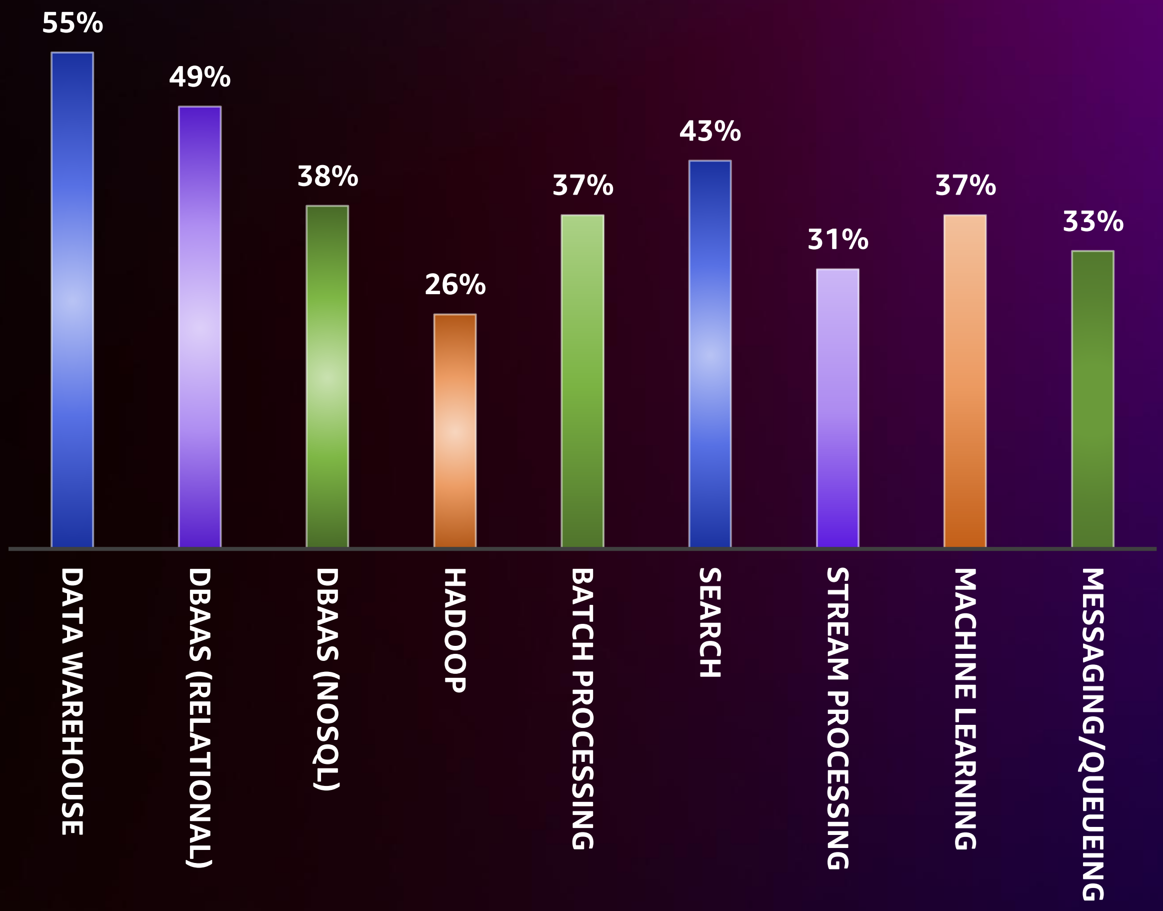
Cloud strategies



Source: O'Reilly cloud adoption

Data gravity





Source: Flexera cloud computing trends

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

The Big Data and Analytics software and cloud services has reached $90.4B spend in 2021, with <u>44%</u> 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

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

Data challenges

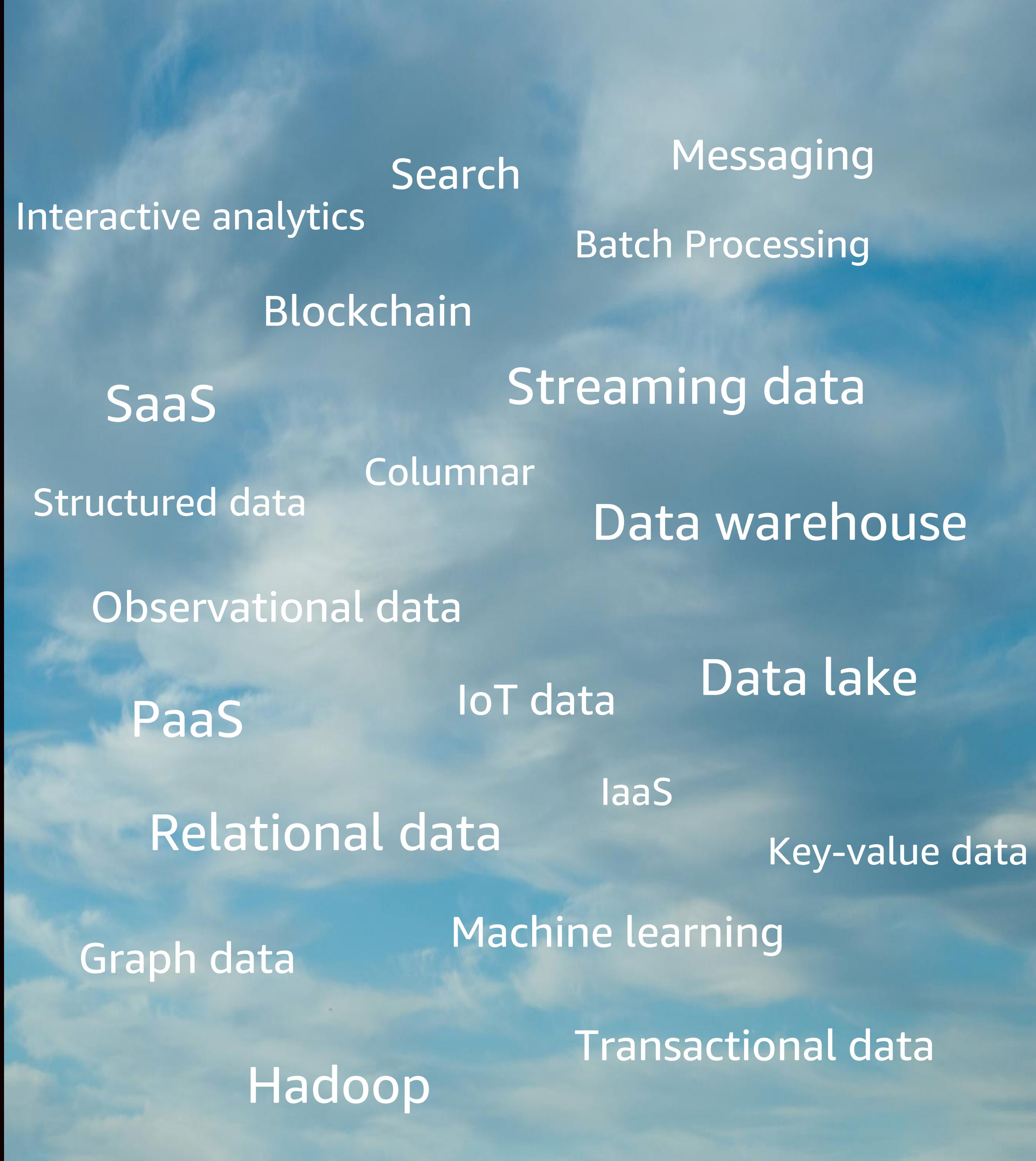
Cost of data management

Interoperability

Operational freedom

Scale-at-speed

Data driven



Messaging

Search

Interactive analytics

Batch Processing

Blockchain

Streaming data

SaaS

Columnar

Structured data

Data warehouse

Observational data

Data lake

IoT data

PaaS

IaaS

Relational data

Key-value data

Machine learning

Graph data

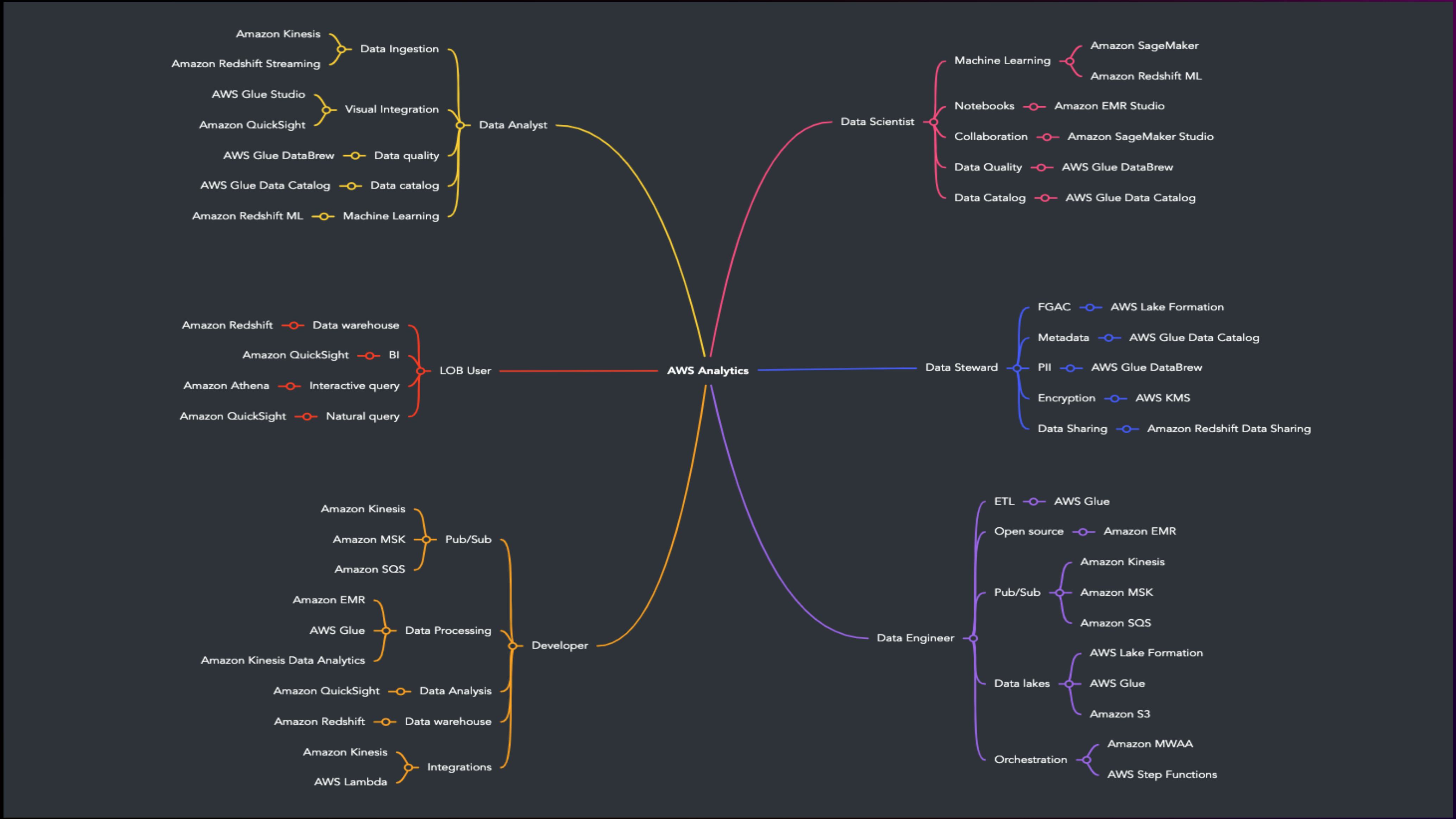
Transactional data

Hadoop

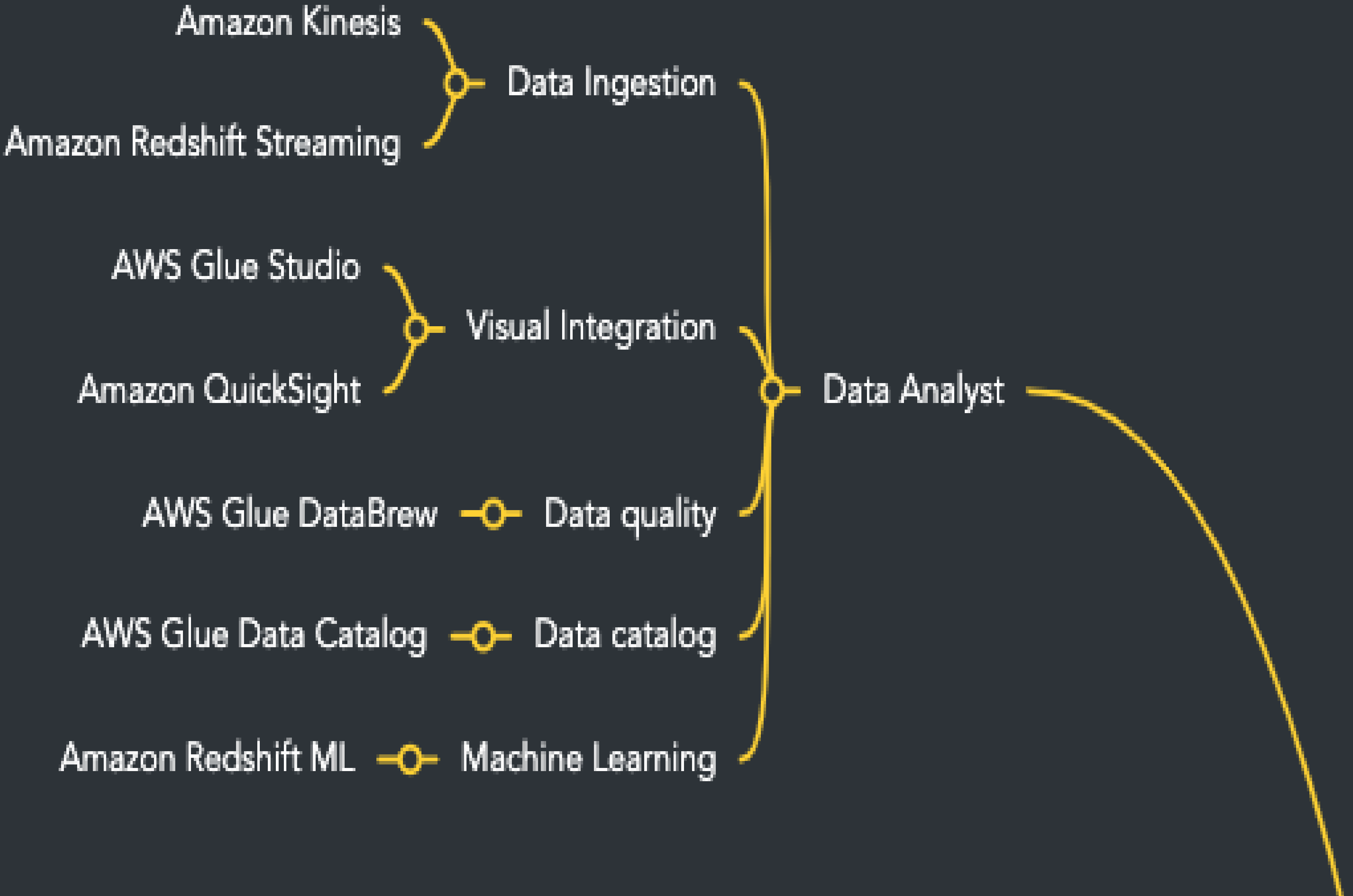
Democratizing analytics

Make analytics available, accessible and affordable

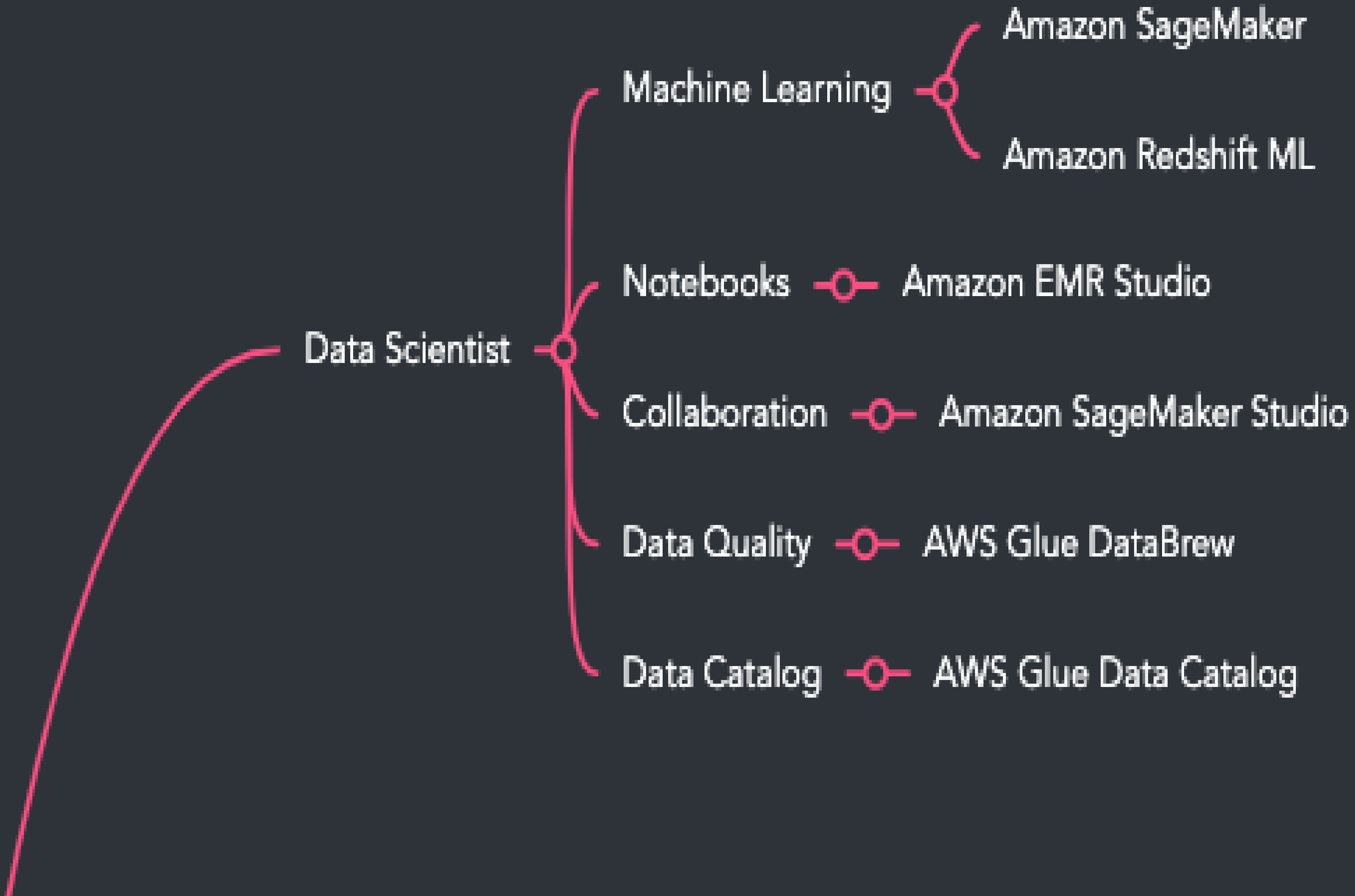
AWS analytics mind-map



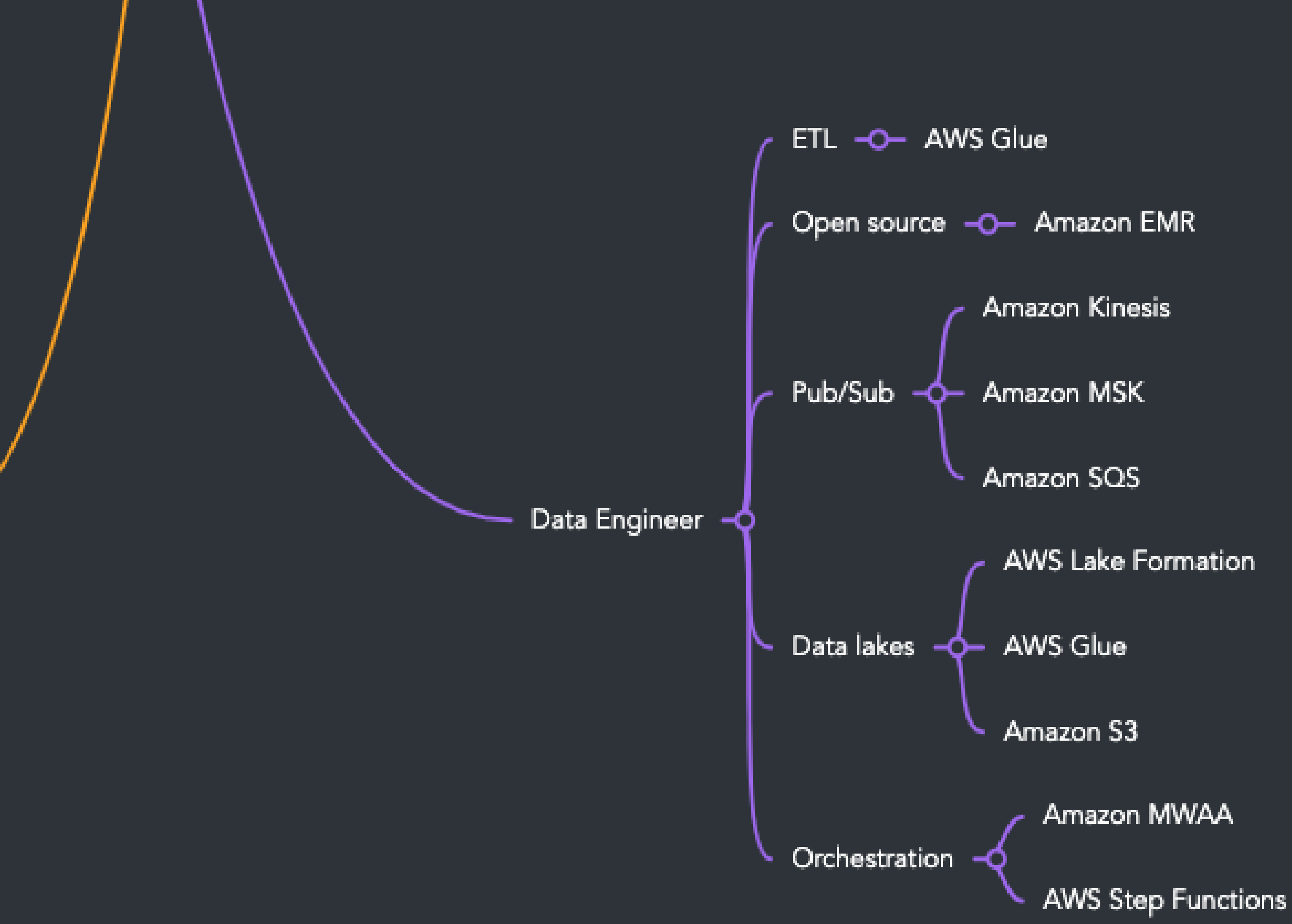
- AWS Analytics
  - Data Analyst
    - Data Ingestion
      - Amazon Kinesis
      - Amazon Redshift Streaming
    - Visual Integration
      - AWS Glue Studio
      - Amazon QuickSight
    - Data quality
      - AWS Glue DataBrew
    - Data catalog
      - AWS Glue Data Catalog
    - Machine Learning
      - Amazon Redshift ML
  - LOB User
    - Data warehouse
      - Amazon Redshift
    - BI
      - Amazon QuickSight
    - Interactive query
      - Amazon Athena
    - Natural query
      - Amazon QuickSight
  - Developer
    - Pub/Sub
      - Amazon Kinesis
      - Amazon MSK
      - Amazon SQS
    - Data Processing
      - Amazon EMR
      - AWS Glue
      - Amazon Kinesis Data Analytics
    - Data Analysis
      - Amazon QuickSight
    - Data warehouse
      - Amazon Redshift
    - Integrations
      - Amazon Kinesis
      - AWS Lambda
  - 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 Steward
    - FGAC
      - AWS Lake Formation
    - Metadata
      - AWS Glue Data Catalog
    - PII
      - AWS Glue DataBrew
    - Encryption
      - AWS KMS
    - Data Sharing
      - Amazon Redshift Data Sharing
  - 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



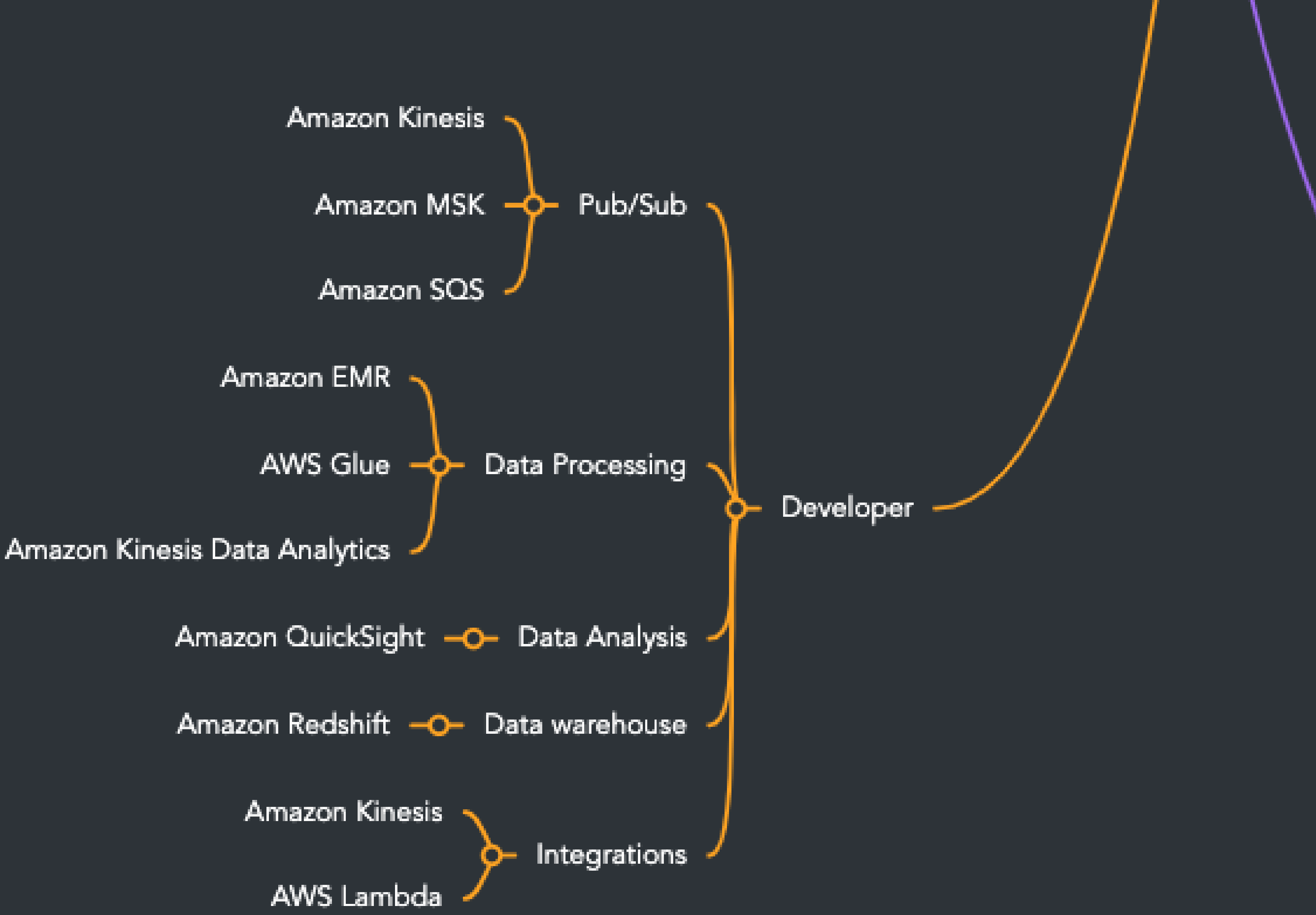
- Data Analyst
  - Data Ingestion
    - Amazon Kinesis
    - Amazon Redshift Streaming
  - Visual Integration
    - AWS Glue Studio
    - Amazon QuickSight
  - Data quality
    - AWS Glue DataBrew
  - Data catalog
    - AWS Glue Data Catalog
  - Machine Learning
    - Amazon Redshift ML



- 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



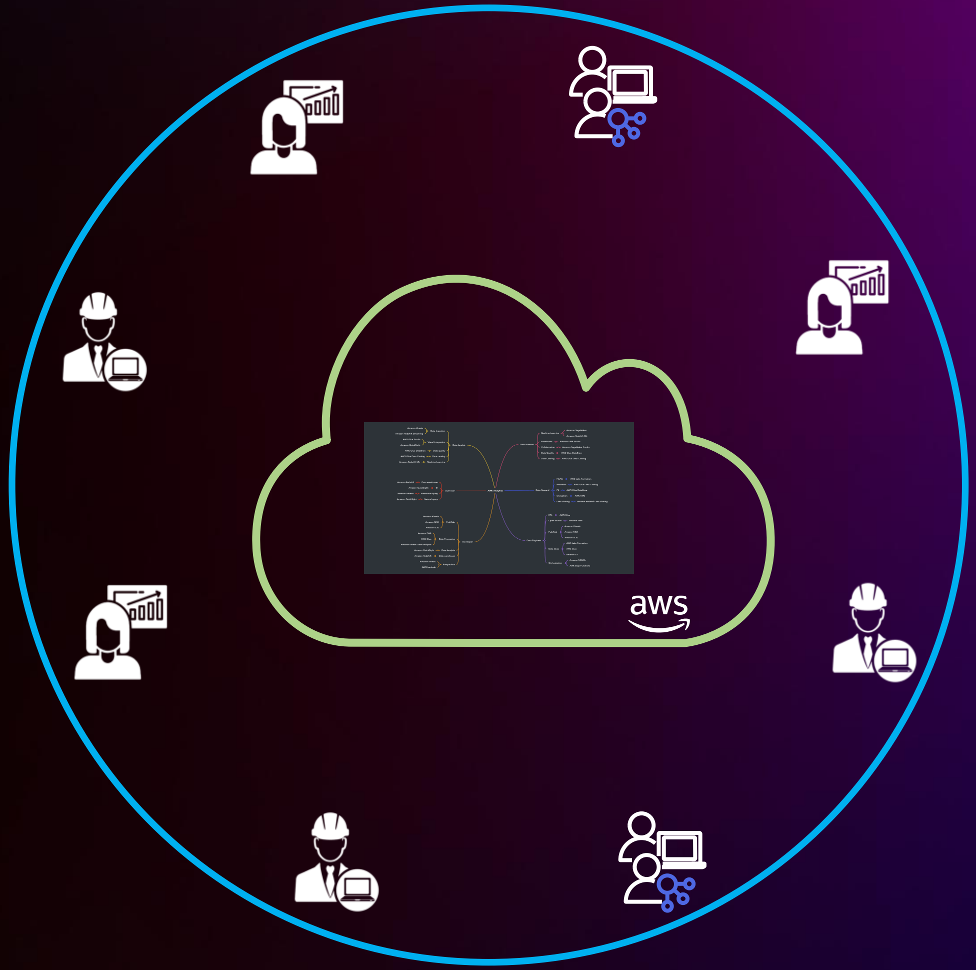
- Data Engineer
  - ETL
    - AWS Glue
  - Open source
    - Amazon EMR
  - Pub/Sub
    - Amazon Kinesis
    - Amazon MSK
    - Amazon SQS
  - Data lakes
    - AWS Lake Formation
    - AWS Glue
    - Amazon S3
  - Orchestration
    - Amazon MWAA
    - AWS Step Functions



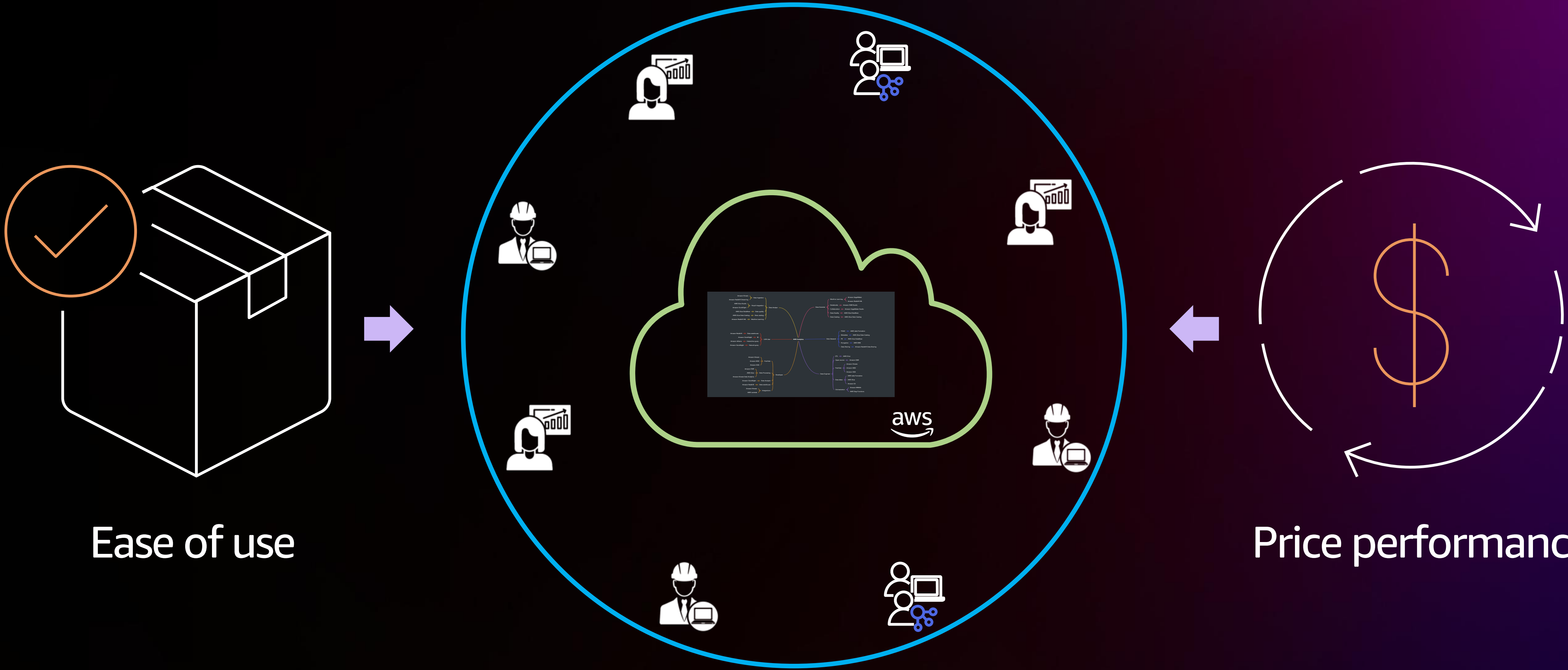
- Developer
  - Pub/Sub
    - Amazon Kinesis
    - Amazon MSK
    - Amazon SQS
  - Data Processing
    - Amazon EMR
    - AWS Glue
    - Amazon Kinesis Data Analytics
  - Data Analysis
    - Amazon QuickSight
  - Data warehouse
    - Amazon Redshift
  - Integrations
    - Amazon Kinesis
    - AWS Lambda

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

Demo walkthrough

Scenario

Build a secure, scalable, reliable and available 3P data pipeline to

1. Ingest data from a SaaS source
2. Perform transformations on the data
3. Catalog and store for upstream analysis

AWS services used

Amazon AppFlow

Visual automation of 3P data pipelines

Built-in monitoring and auditing

High scale data transfer

Encryption and fine-grained permissions

AWS Glue DataBrew

Visual data preparation at scale

Advanced data profiling

Fully reusable configurations

AWS Glue Crawler

AWS Glue Data Catalog

Automatic schema discovery

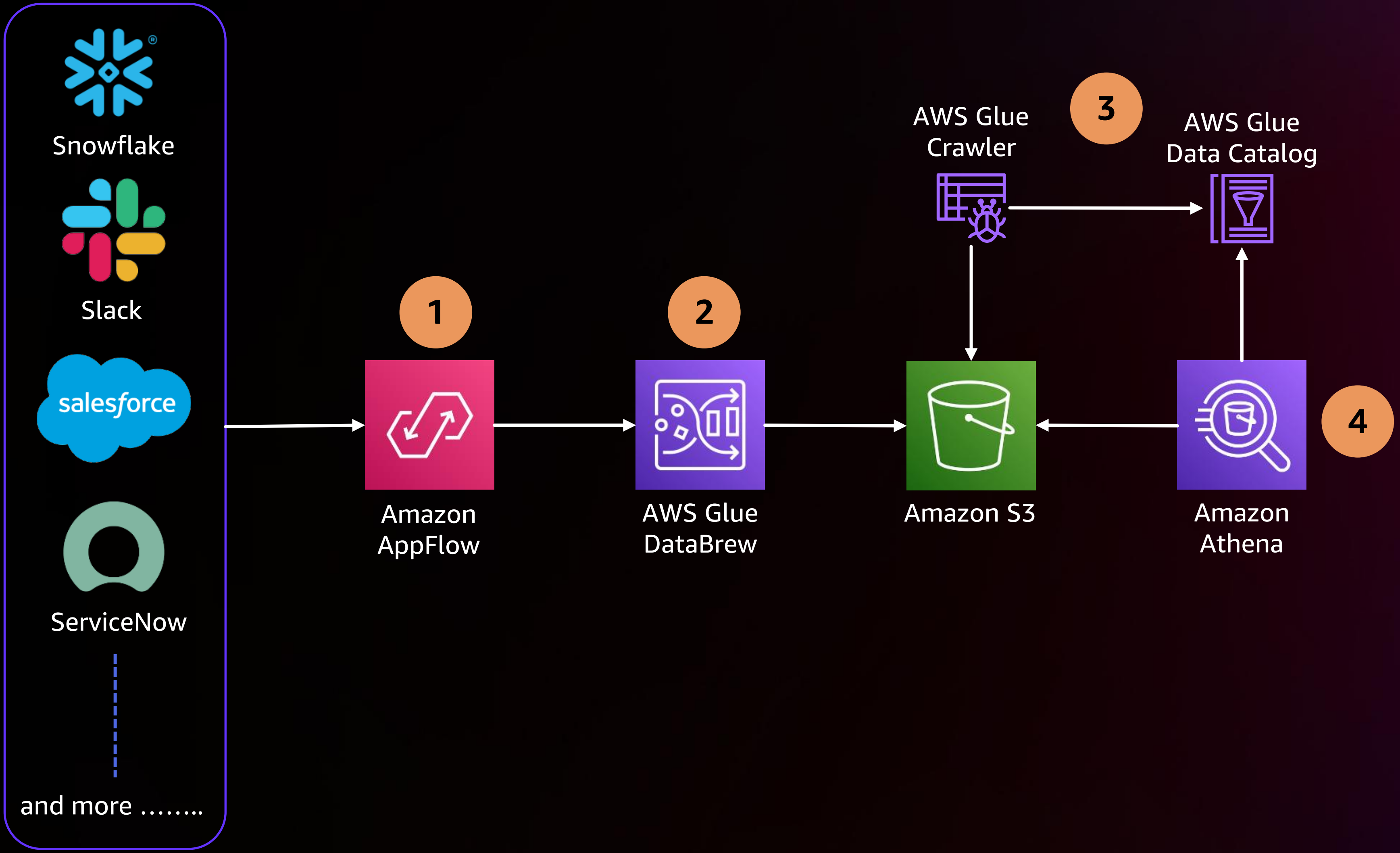
Persistent metadata store

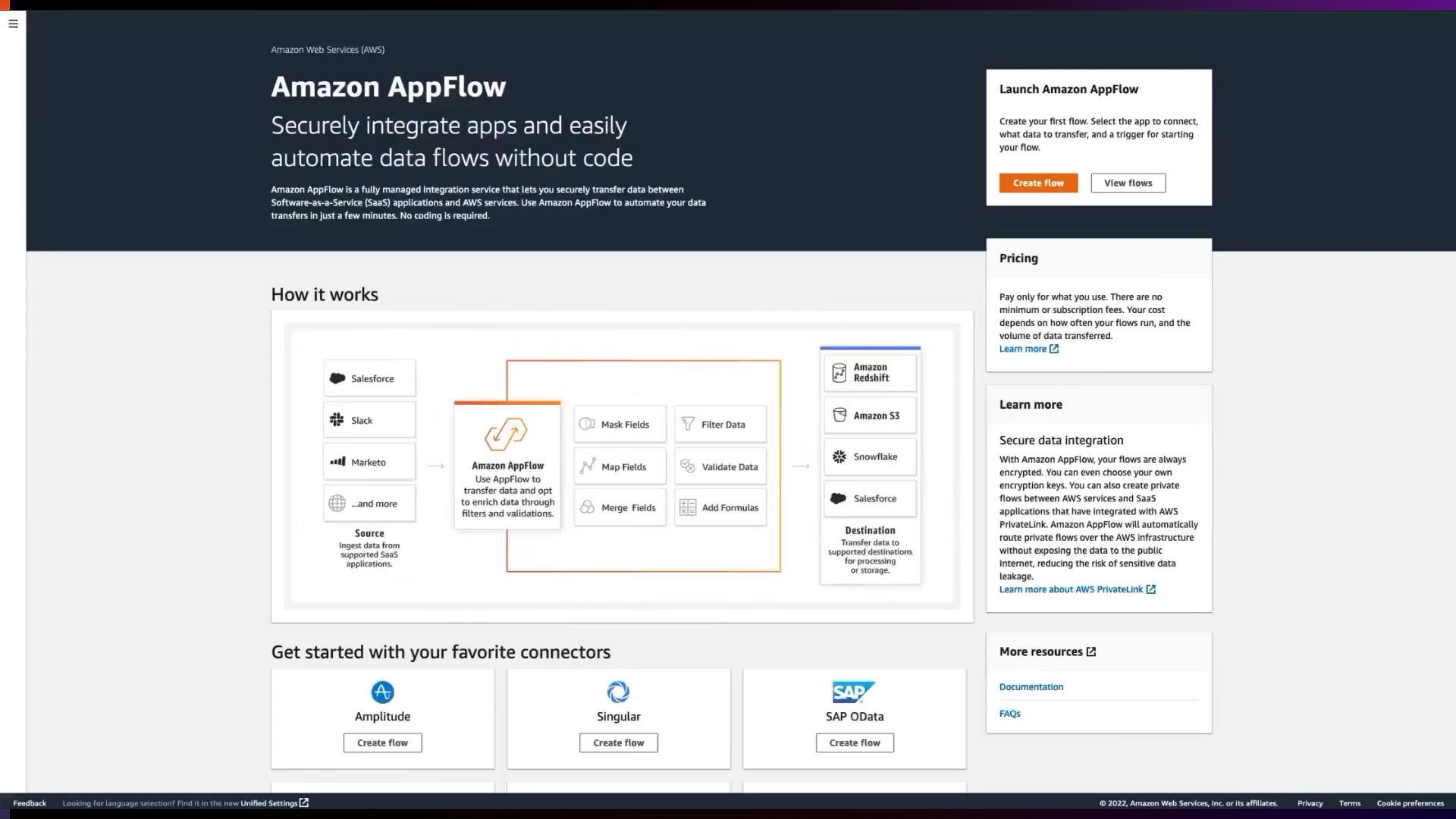
Amazon Athena

Interactive query service

AWS Glue native integration

Demo architecture





Amazon Web Services (AWS)

Amazon AppFlow

Securely integrate apps and easily automate data flows without code

Amazon AppFlow is a fully managed integration service that lets you securely transfer data between Software-as-a-Service (SaaS) applications and AWS services. Use Amazon AppFlow to automate your data transfers in just a few minutes. No coding is required.

Launch Amazon AppFlow

Create your first flow. Select the app to connect, what data to transfer, and a trigger for starting your flow.

How it works

Salesforce

Slack

Marketo

...and more

Source
Ingest data from supported SaaS applications.

Amazon AppFlow
Use AppFlow to transfer data and opt to enrich data through filters and validations.

Mask Fields

Filter Data

Map Fields

Validate Data

Merge Fields

Add Formulas

Amazon Redshift

Amazon S3

Snowflake

Salesforce

Destination
Transfer data to supported destinations for processing or storage.

Pricing

Pay only for what you use. There are no minimum or subscription fees. Your cost depends on how often your flows run, and the volume of data transferred.
Learn more

Learn more

Secure data integration

With Amazon AppFlow, your flows are always encrypted. You can even choose your own encryption keys. You can also create private flows between AWS services and SaaS applications that have integrated with AWS PrivateLink. Amazon AppFlow will automatically route private flows over the AWS infrastructure without exposing the data to the public internet, reducing the risk of sensitive data leakage.
Learn more about AWS PrivateLink

Get started with your favorite connectors

Amazon AppFlow

Amplitude

Singular

SAP OData

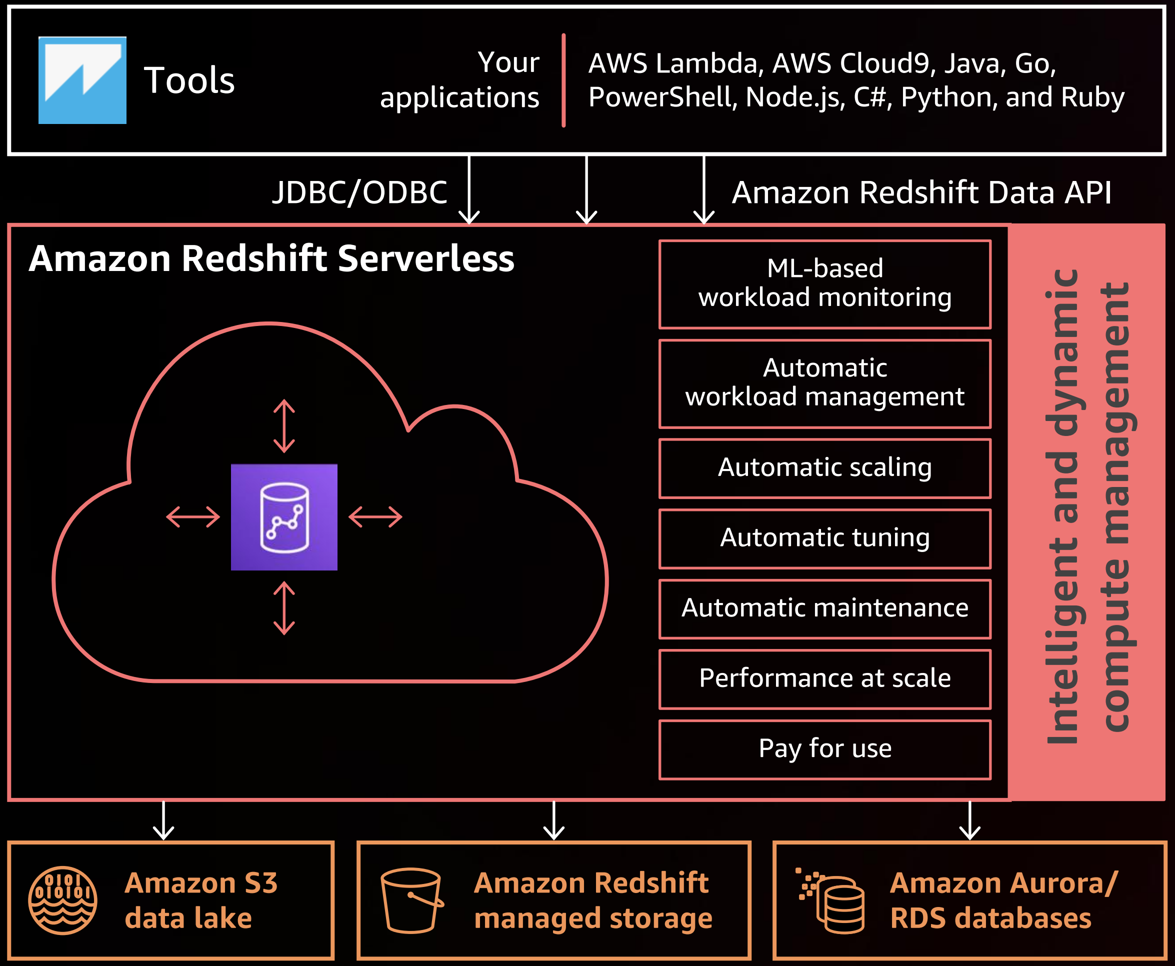
More resources

Documentation

FAQs

Looking for language selection? Find it in the new Unified Settings

Amazon Redshift Serverless



Tools

Your applications | AWS Lambda, AWS Cloud9, Java, Go, PowerShell, Node.js, C#, Python, and Ruby

JDBC/ODBC

Amazon Redshift Data API

Amazon Redshift Serverless

- ML-based workload monitoring
- Automatic workload management
- Automatic scaling
- Automatic tuning
- Automatic maintenance
- Performance at scale
- Pay for use

Intelligent and dynamic compute management

Amazon S3 data lake

Amazon Redshift managed storage

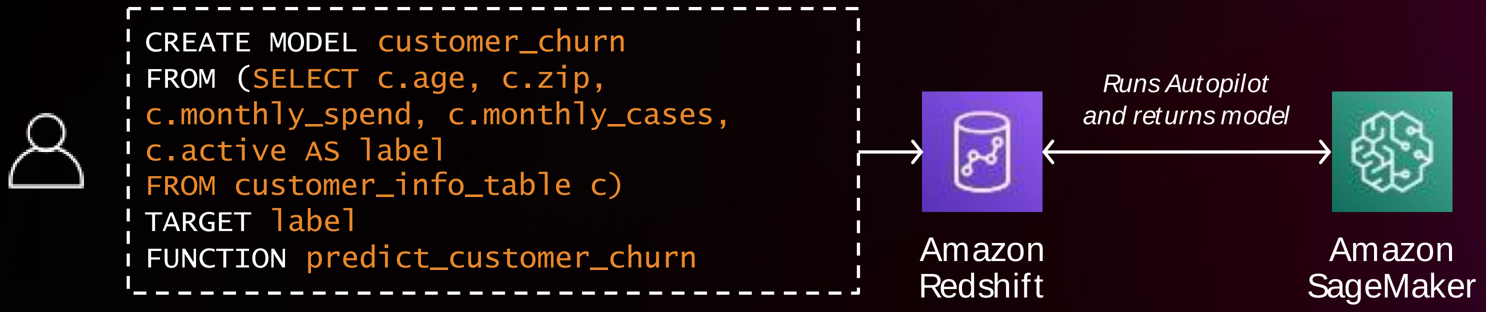
Amazon Aurora/ RDS databases

All Amazon Redshift SQL functionality applies

- Security and user management
- Semi-structured data
- Data sharing
- Machine learning functions
- Data lake queries
- Federated query
- Durability and transactional guarantees
- JDBC/ODBC and Data API

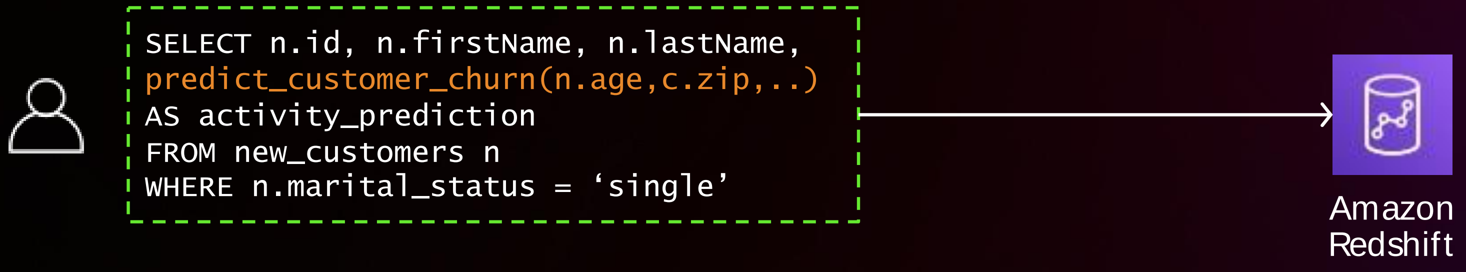
Amazon Redshift ML

TRAIN



```
CREATE MODEL customer_churn
FROM (SELECT c.age, c.zip,
c.monthly_spend, c.monthly_cases,
c.active AS label
FROM customer_info_table c)
TARGET label
FUNCTION predict_customer_churn
```

PREDICT



```
SELECT n.id, n.firstName, n.lastName,
predict_customer_churn(n.age,c.zip,..)
AS activity_prediction
FROM new_customers n
WHERE n.marital_status = ‘single’
```

Demo walkthrough

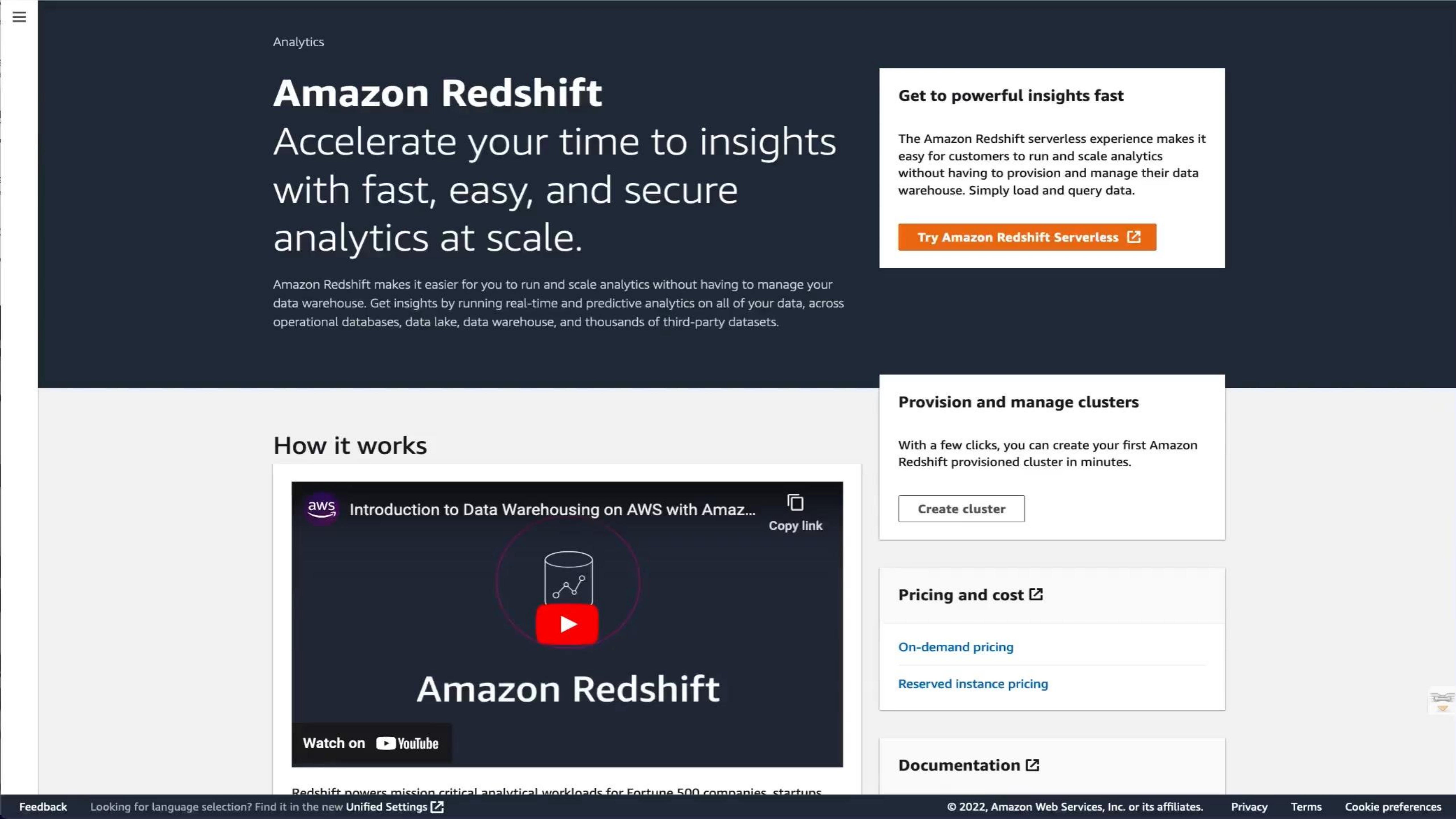
Scenario

Analyze the Abalone dataset and determine the relationship between the physical measurements, and use that to determine the age of the abalone.
The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope.

Outcome

Predict the age using different physical measurements, which is easier to measure. The age of abalone is (number of rings + 1.5) years.

Abalone (Snail)



Analytics

Amazon Redshift

Accelerate your time to insights with fast, easy, and secure analytics at scale.

Amazon Redshift makes it easier for you to run and scale analytics without having to manage your data warehouse. Get insights by running real-time and predictive analytics on all of your data, across operational databases, data lake, data warehouse, and thousands of third-party datasets.

Get to powerful insights fast

The Amazon Redshift serverless experience makes it easy for customers to run and scale analytics without having to provision and manage their data warehouse. Simply load and query data.

Provision and manage clusters

With a few clicks, you can create your first Amazon Redshift provisioned cluster in minutes.

Pricing and cost

On-demand pricing

Reserved instance pricing

Documentation

How it works

Redshift powers mission critical analytical workloads for Fortune 500 companies, startups

Feedback
Looking for language selection? Find it in the new Unified Settings

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

Apache Spark on Amazon EMR

Dynamic sized executors

Adaptive join selection

Dynamic pruning of data columns

Operator Optimization

Early worker allocation

Intelligent filtering

Parallel/async initialization

Redundant scan elimination

Data pre-fetch

Broadcast join w/o statistics

Stats inference

Optimized metadata fetch

Apache Spark on Amazon EMR

Dynamic sized executors

Adaptive join selection

Dynamic pruning of data columns

Operator Optimization

Early worker allocation

Intelligent filtering

Parallel/async initialization

Redundant scan elimination

Data pre-fetch

Broadcast join w/o statistics

Stats inference

Optimized metadata fetch

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

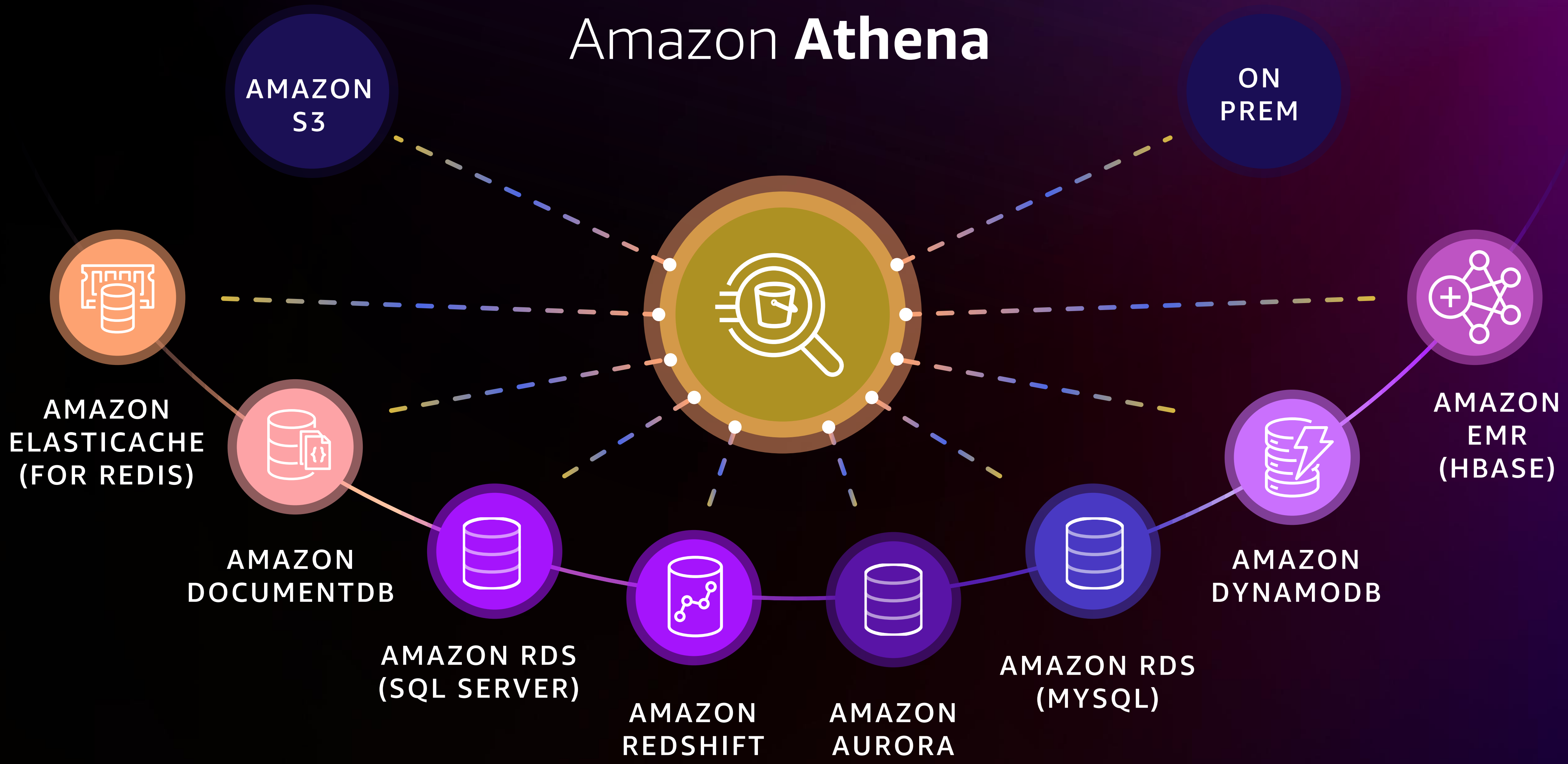
Point to S3 and start querying

DDL operations

Query concurrency

Integrated data connectors

Amazon **Athena**



New data source connectors

- SAP HANA
- Teradata
- Cloudera
- Hortonworks
- Snowflake
- Microsoft SQL Server
- Oracle
- Google BigQuery
- Azure Data Lake Storage Gen2
- Azure Synapse

No per-connector costs – pay only for the queries you run

Easy to configure from Athena console

Configure once and share across accounts

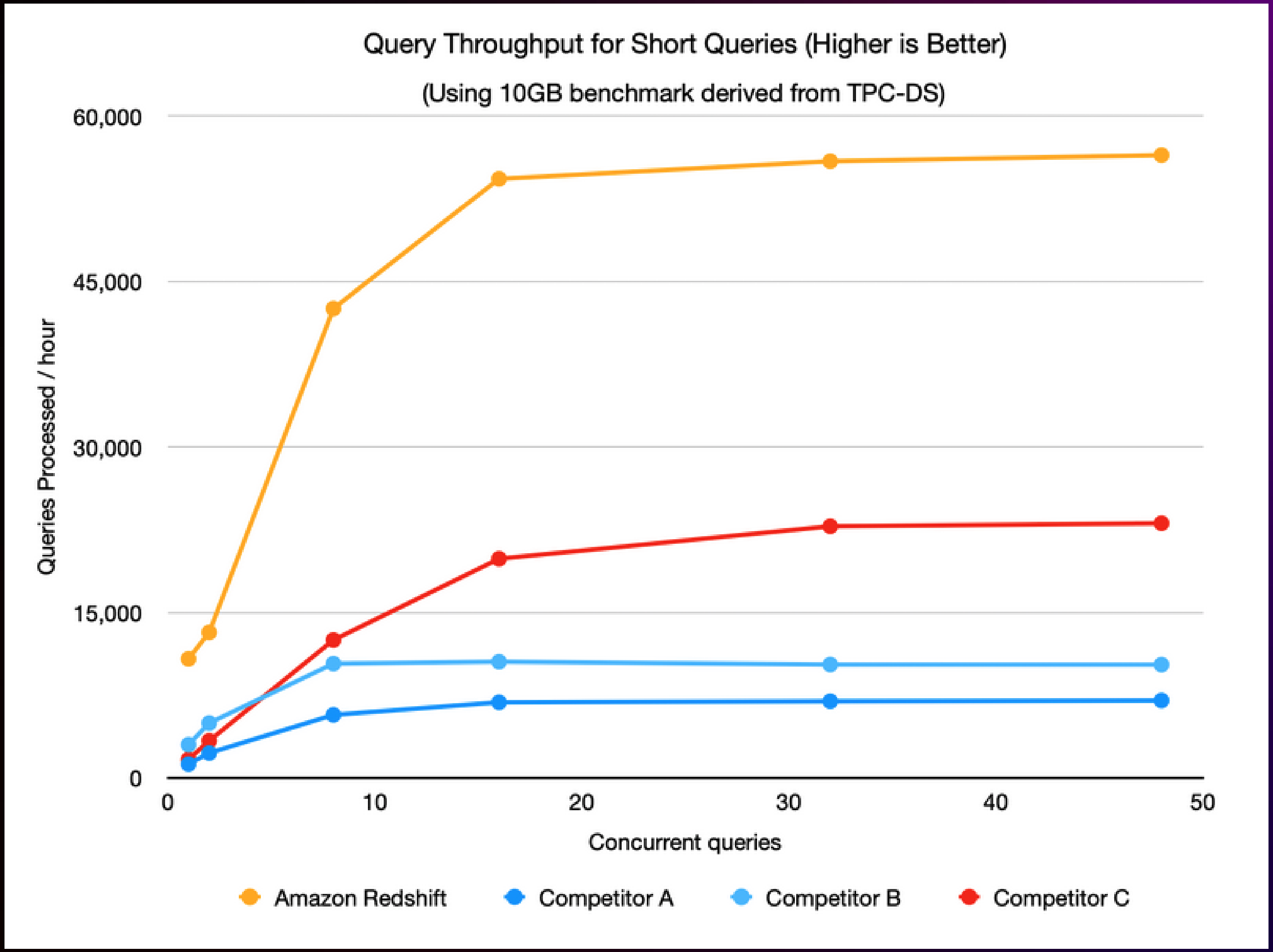
Open source and fully supported by AWS

Amazon Redshift

ML-BASED OPTIMIZATIONS TO GET STARTED EASILY AND GET THE FASTEST PERFORMANCE QUICKLY

- Automatic vacuum delete
- ATO: Automatic distribution keys
- ATO: Automatic sort keys
- Auto workload manager
- Automatic table sort
- ATO: Automatic column encoding
- Auto Analyze
- Auto refresh & re-write Materialized Views

Performance with Amazon Redshift

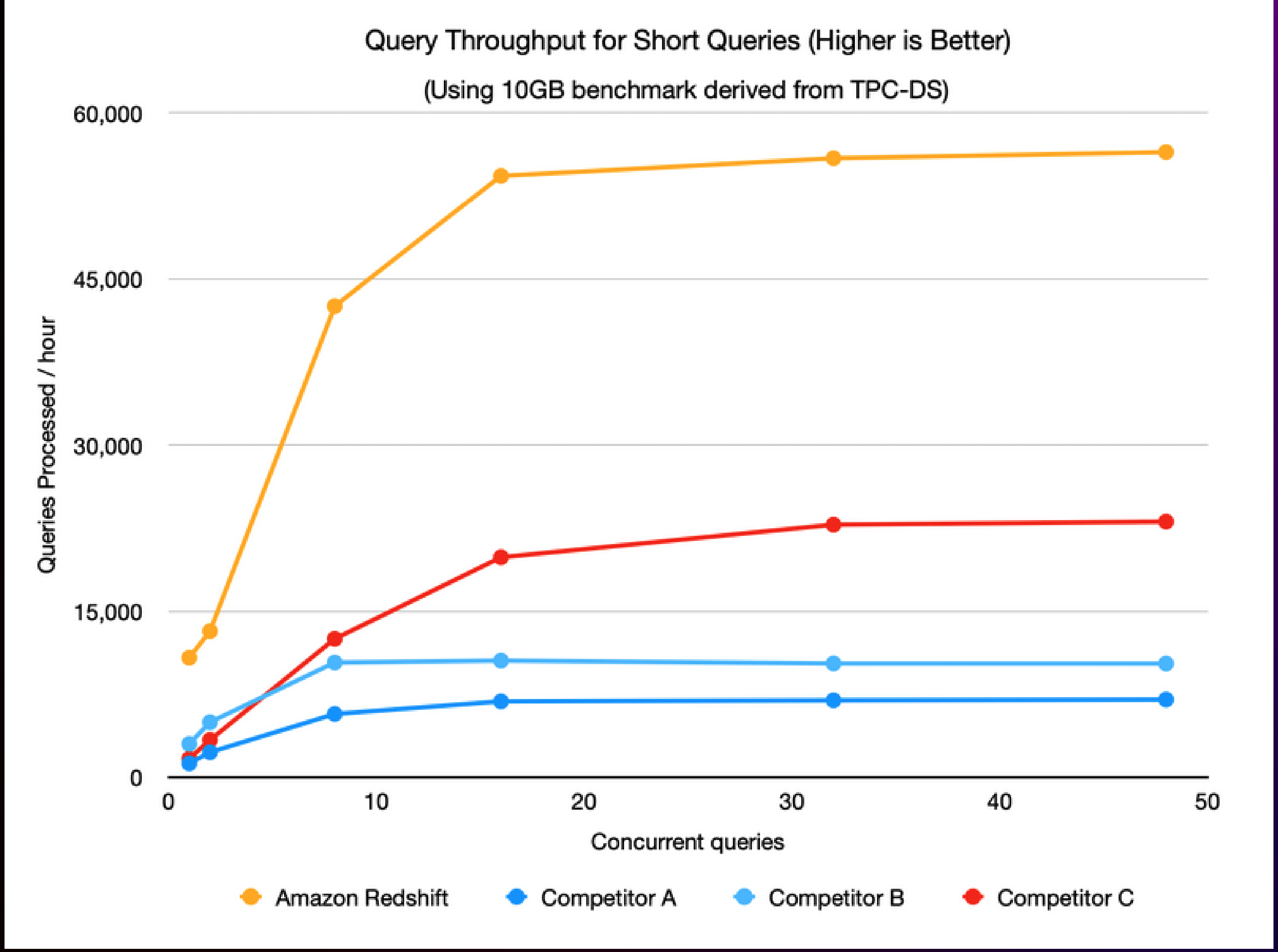


Performance with Amazon Redshift

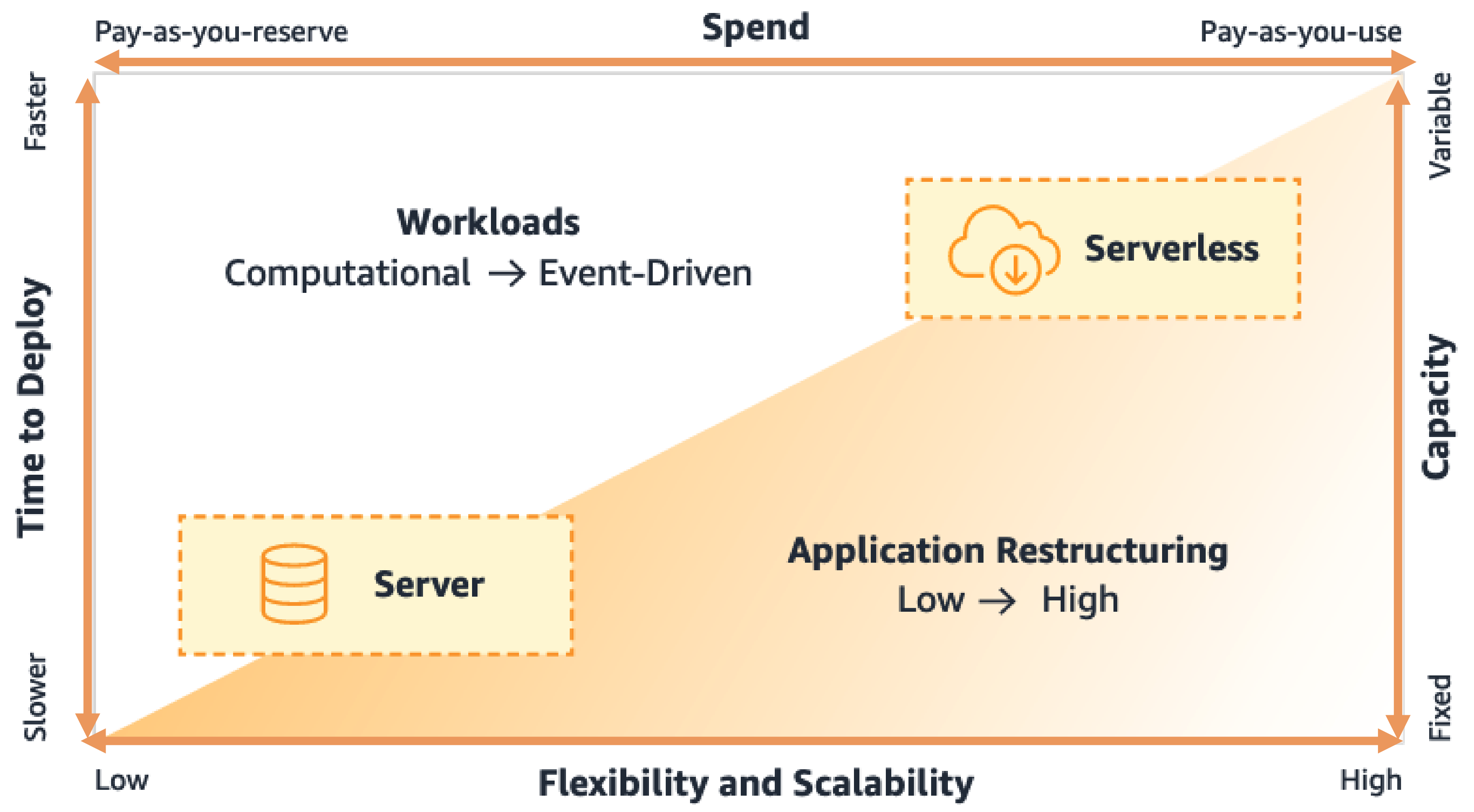
Reduced query planning overhead

Concurrent process optimization

Improved query parallelism

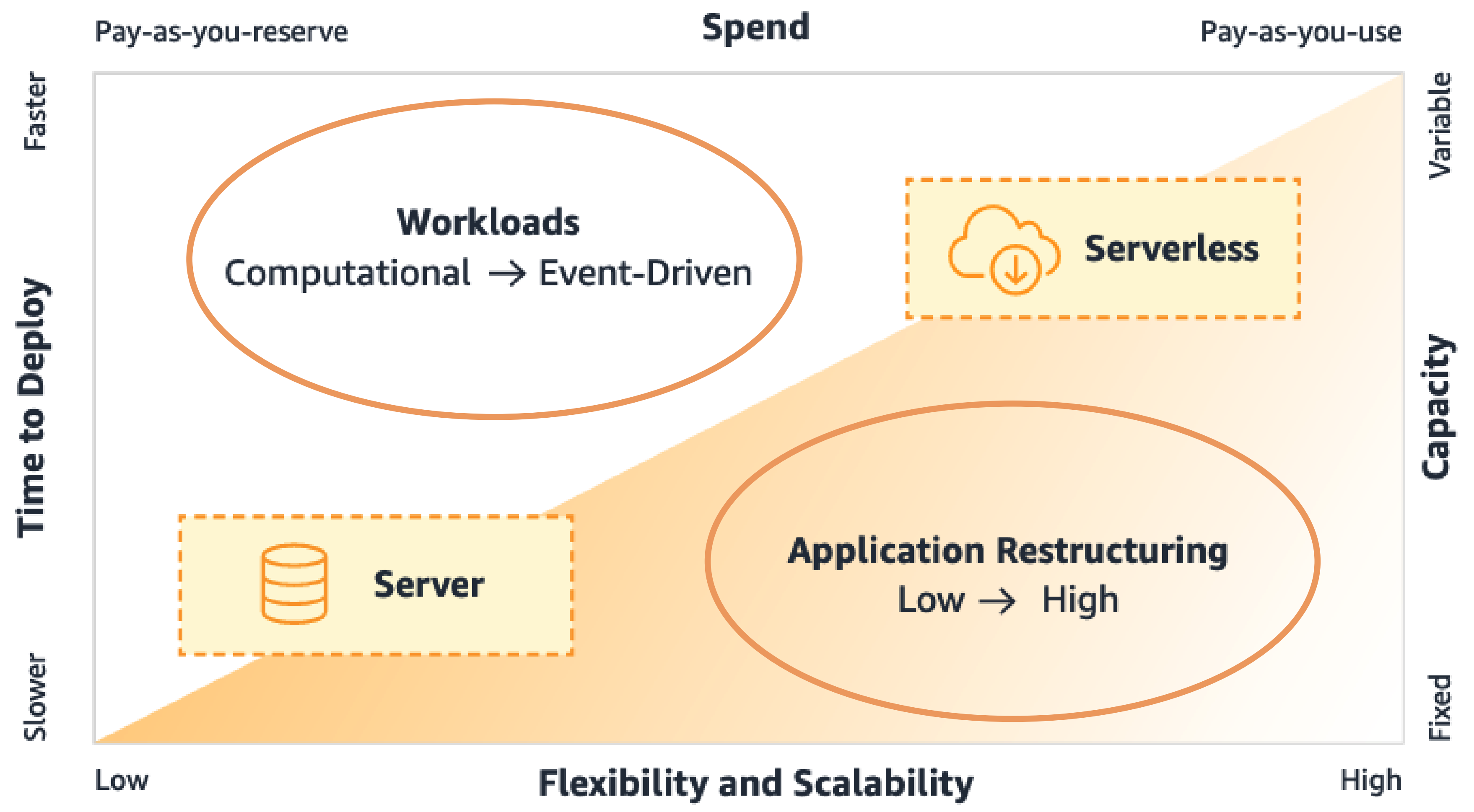


Serverless TCO



Source: Deloitte

Serverless TCO



Source: Deloitte

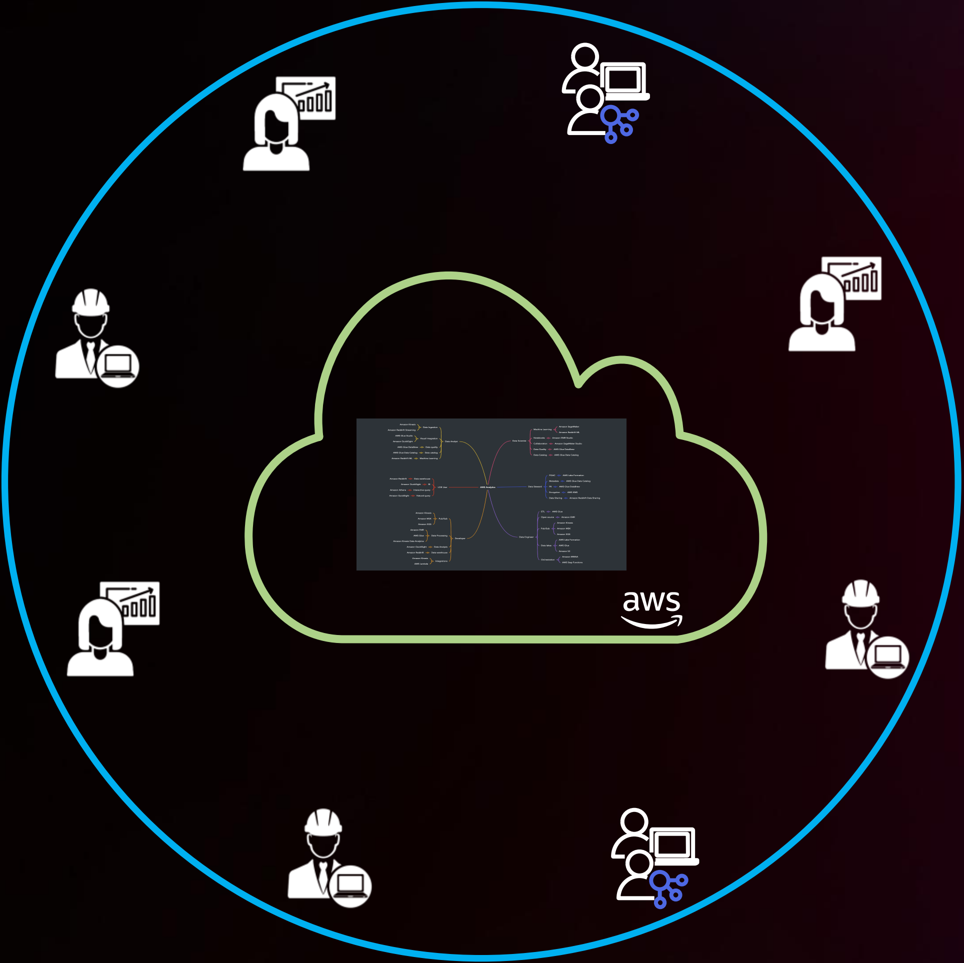
Serverless data analytics on AWS

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

aws data lab

Work backwards from big ideas

Focused, real-world solution building

Accelerate path to production by months

Come with an idea, leave with a solution.

Embedded Analytics Data Lab (EADL)

EADL is a no-cost collaborative engagement that helps development teams decrease the time required to launch applications with embedded analytics from **Amazon QuickSight** by providing hands-on guidance and architectural best practices.

Create differentiated, analytics-driven experiences that empower end-users to make more informed decisions by embedding rich analytics directly into applications:

- Interactive visuals
- Dashboards
- Machine learning-powered natural language query using **Amazon QuickSight Q**

Build skills to unlock the value of your data with AWS Training and Certification

Explore 180+ relevant trainings including:

Building Modern Data Analytics Solutions on AWS (new collection of Classroom Trainings)

Data Analytics Fundamentals

Get AWS Certified:

aws certified Data Analytics SPECIALTY

aws certified Database SPECIALTY

aws certified Machine Learning SPECIALTY

aws certified Security SPECIALTY

Talk to your AWS account team to learn more!

Thank you!

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