
AWS
re:Invent

ANT291-L

Harness the power of data with AWS analytics

Rahul Pathak
Vice President Analytics
Amazon Web Services
@rahulpathak
rapathak@amazon.com

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What to expect

01

UNDERSTAND

our strategy
and portfolio

02

HEAR

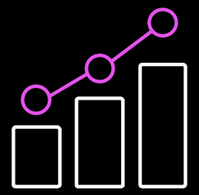
about the
new announcements

03

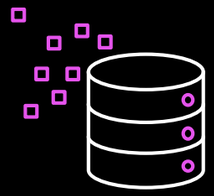
LEARN

how you can
get started today

Customers want more value from their data



GROWING EXPONENTIALLY



FROM NEW SOURCES



INCREASINGLY DIVERSE



USED BY MANY PEOPLE



ANALYZED BY MANY APPLICATIONS

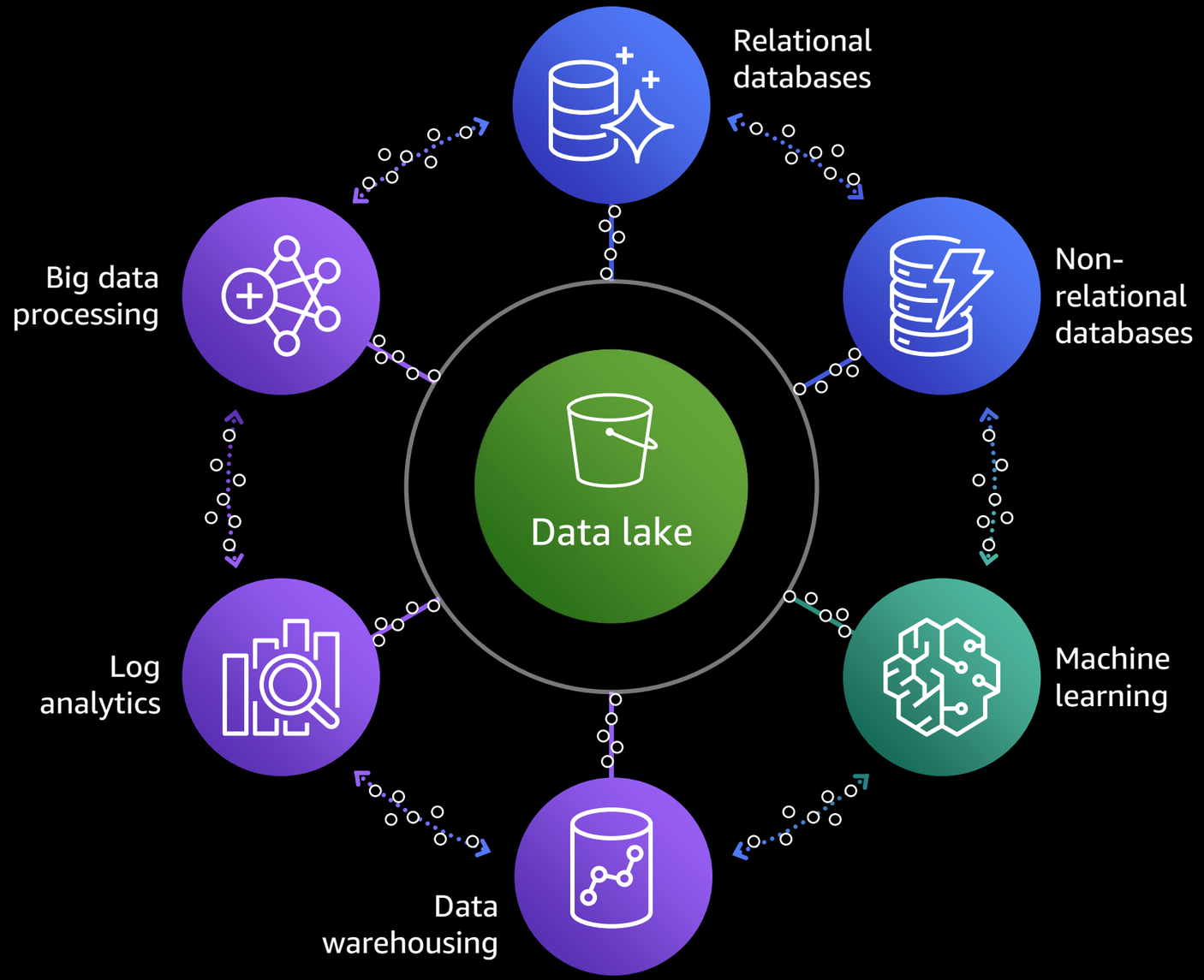
Agility is more important than ever

vyaire™ MEDICAL

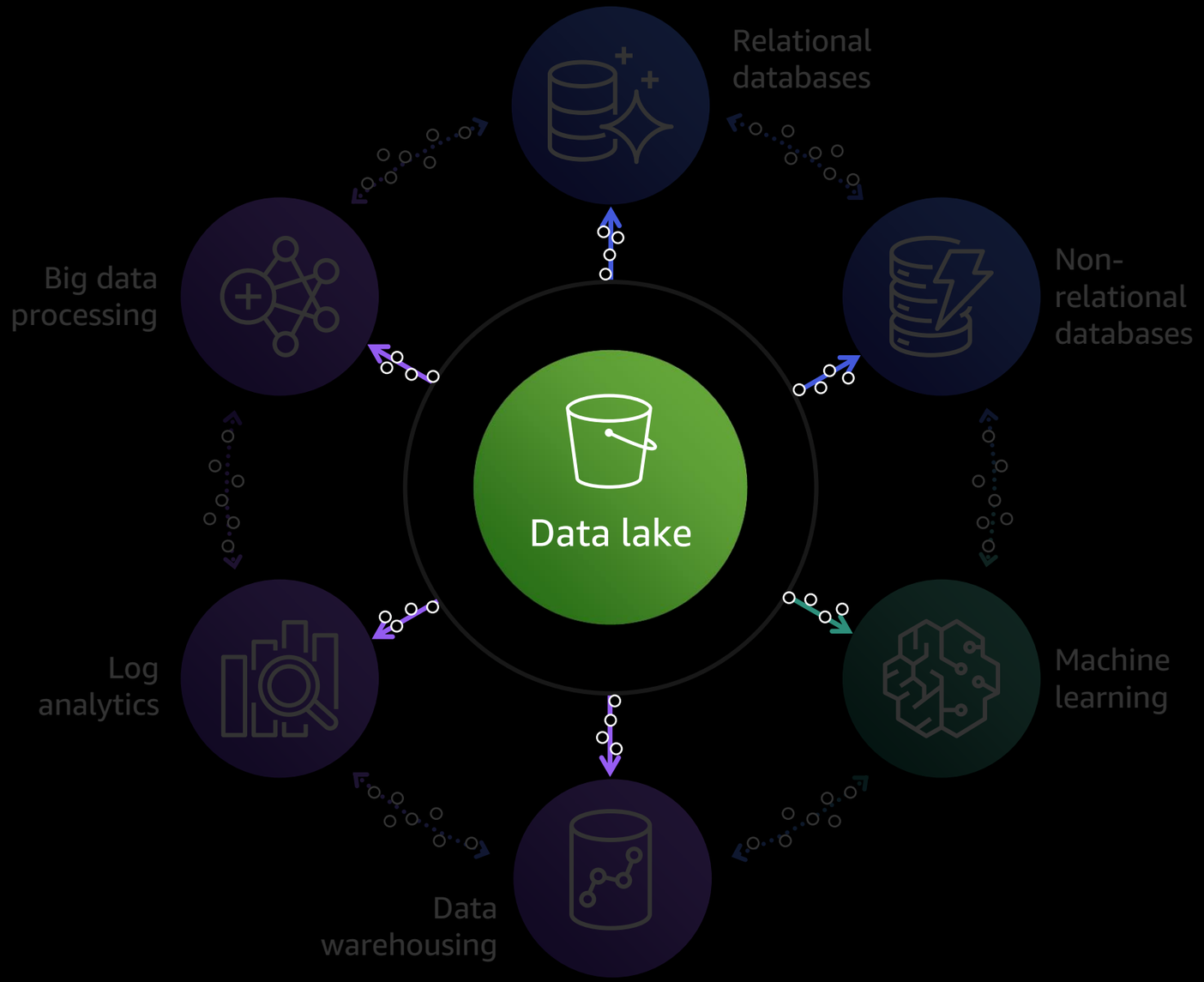
moderna



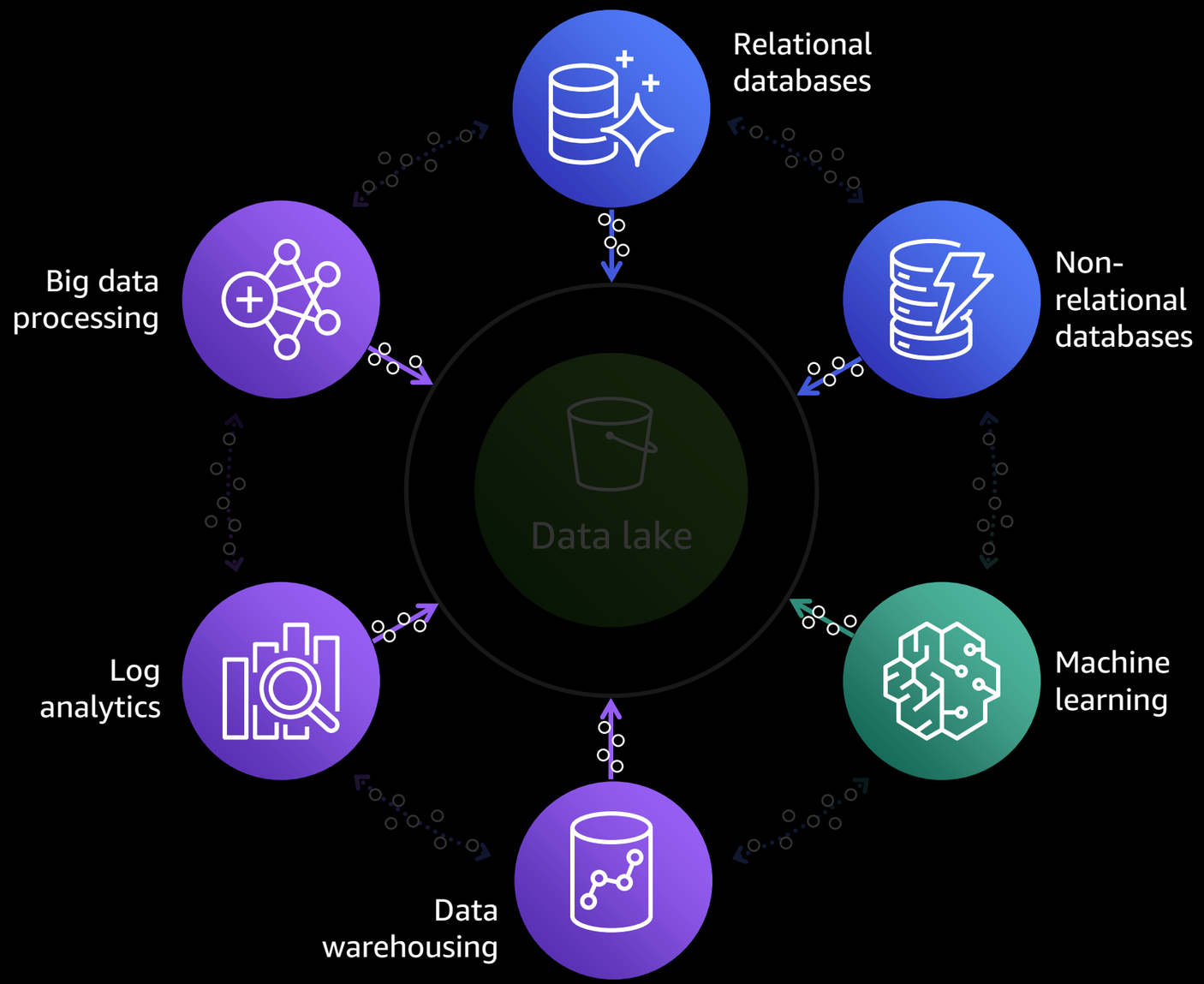




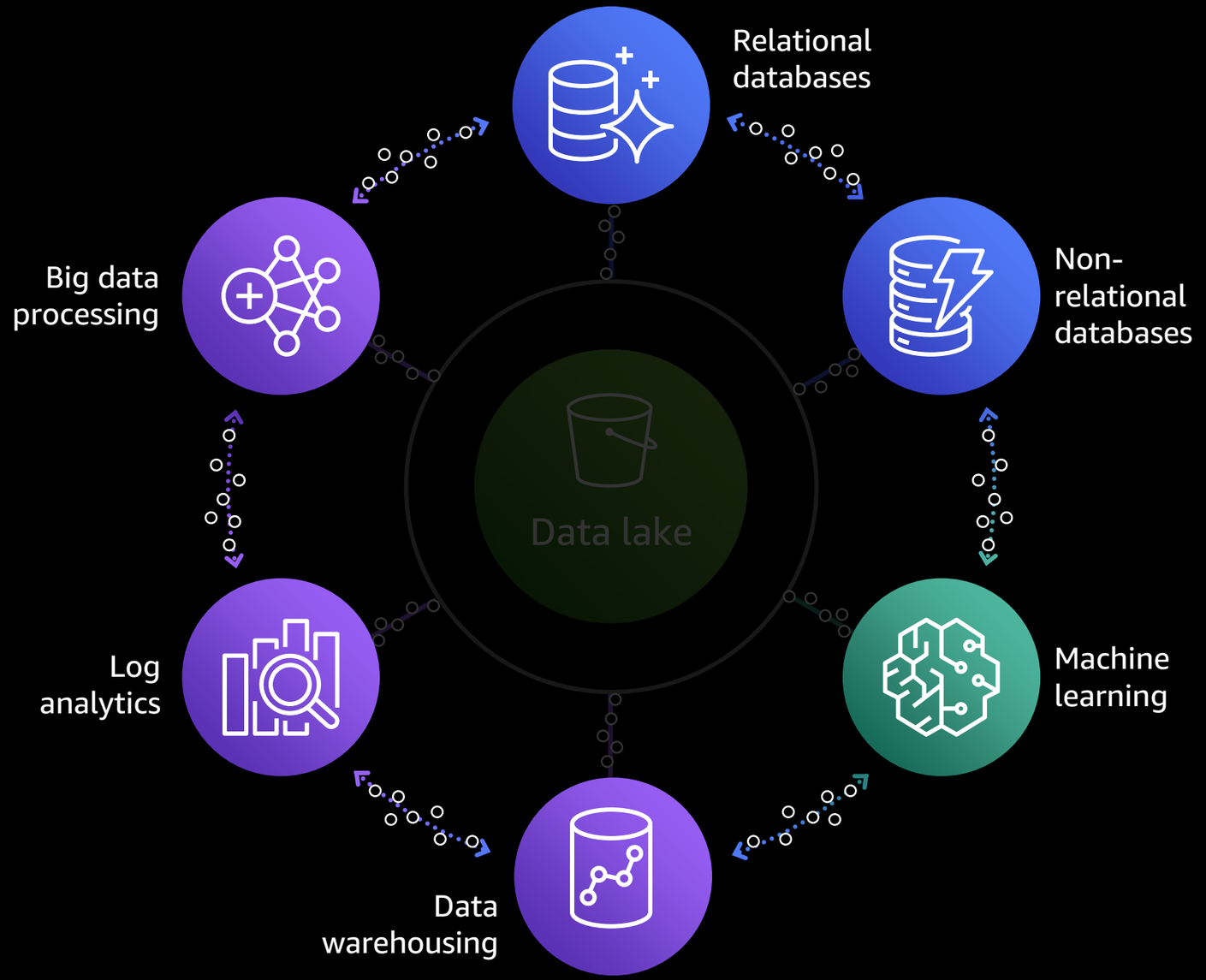
Customer data in the real world


Relational databases
Non-relational databases
Machine learning
Data warehousing
Log analytics
Big data processing
Data lake

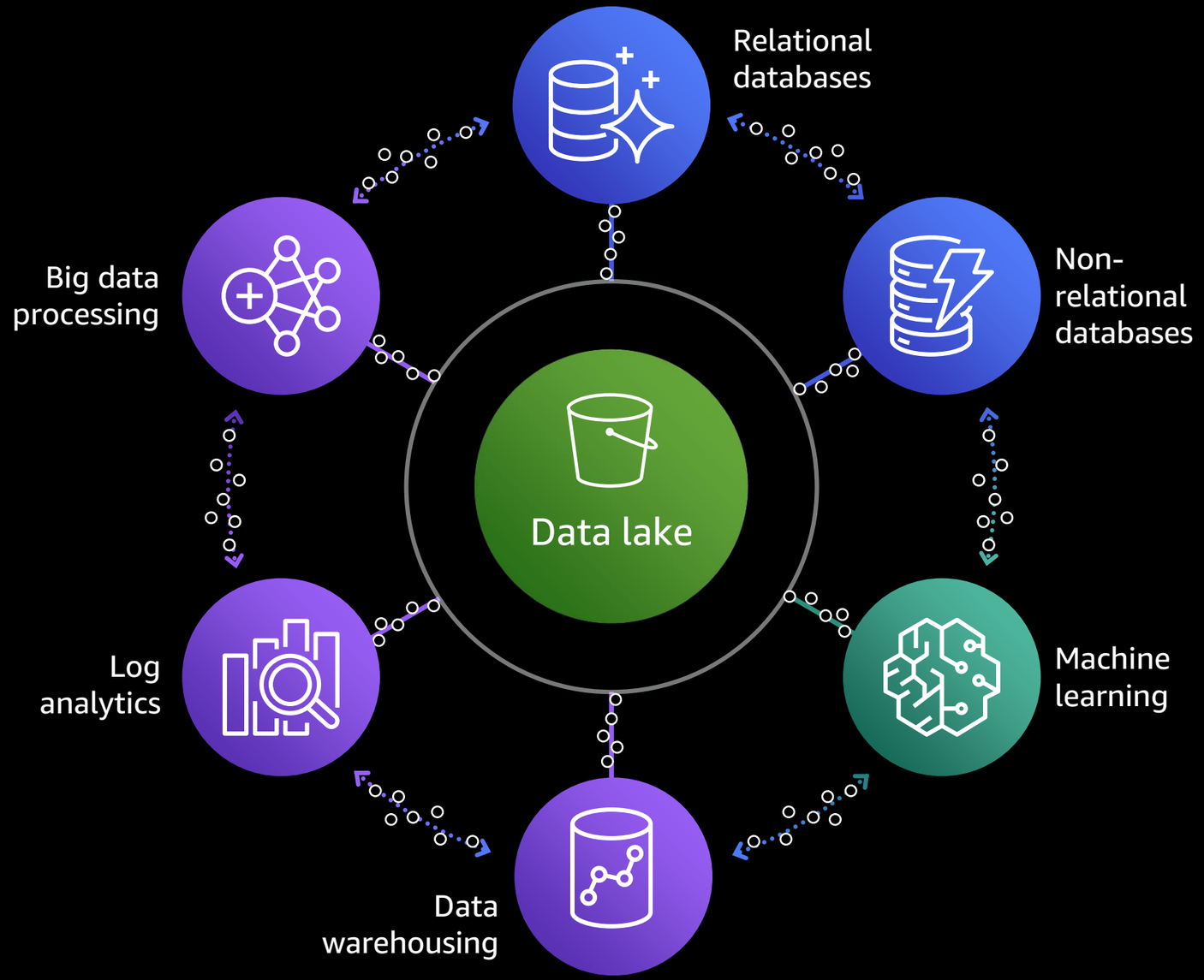
Inside out


Relational databases
Non-relational databases
Machine learning
Data warehousing
Log analytics
Big data processing
Data lake

Outside in

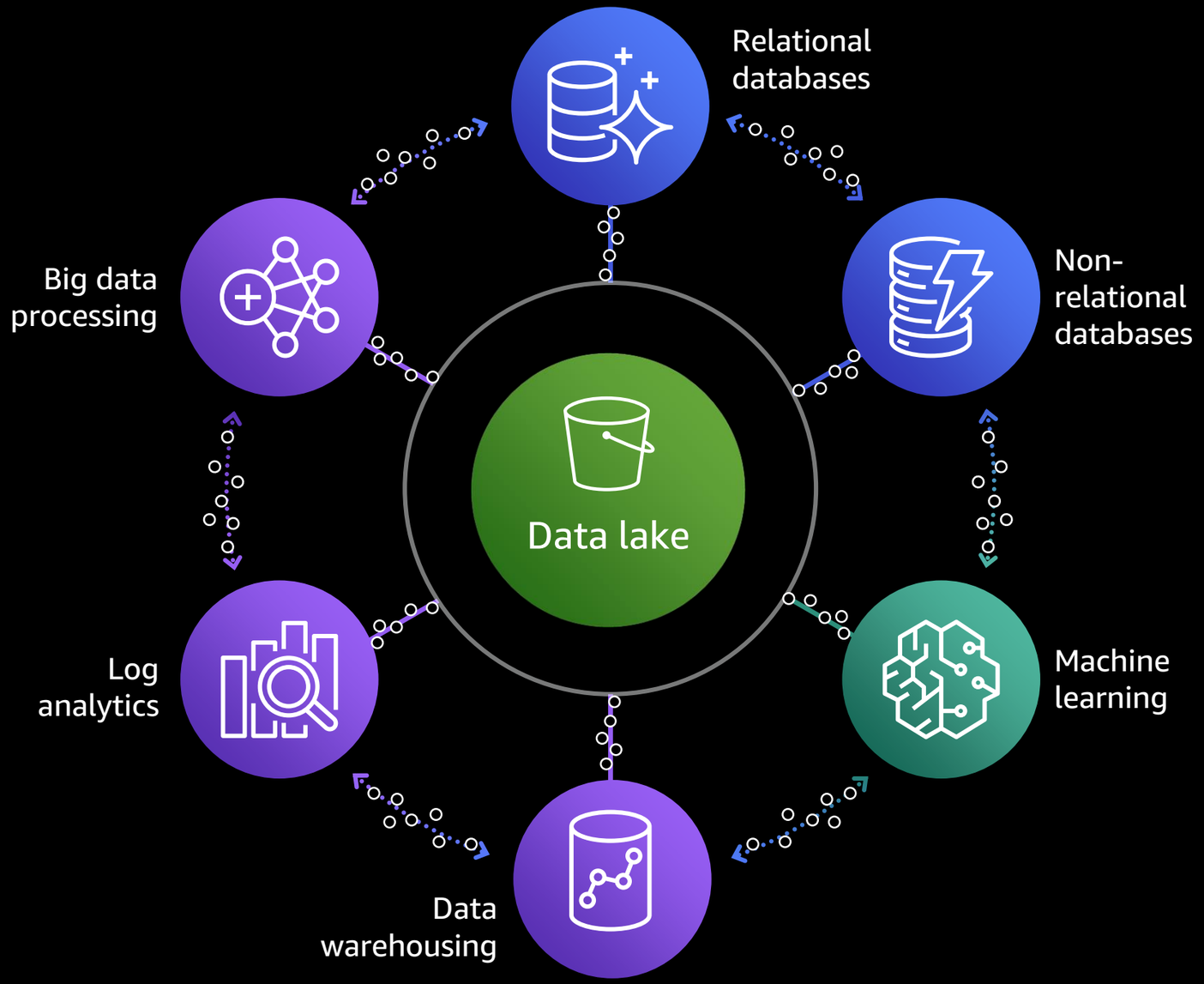


Around the perimeter



Data gravity

Lake House architecture



SCALABLE DATA LAKES

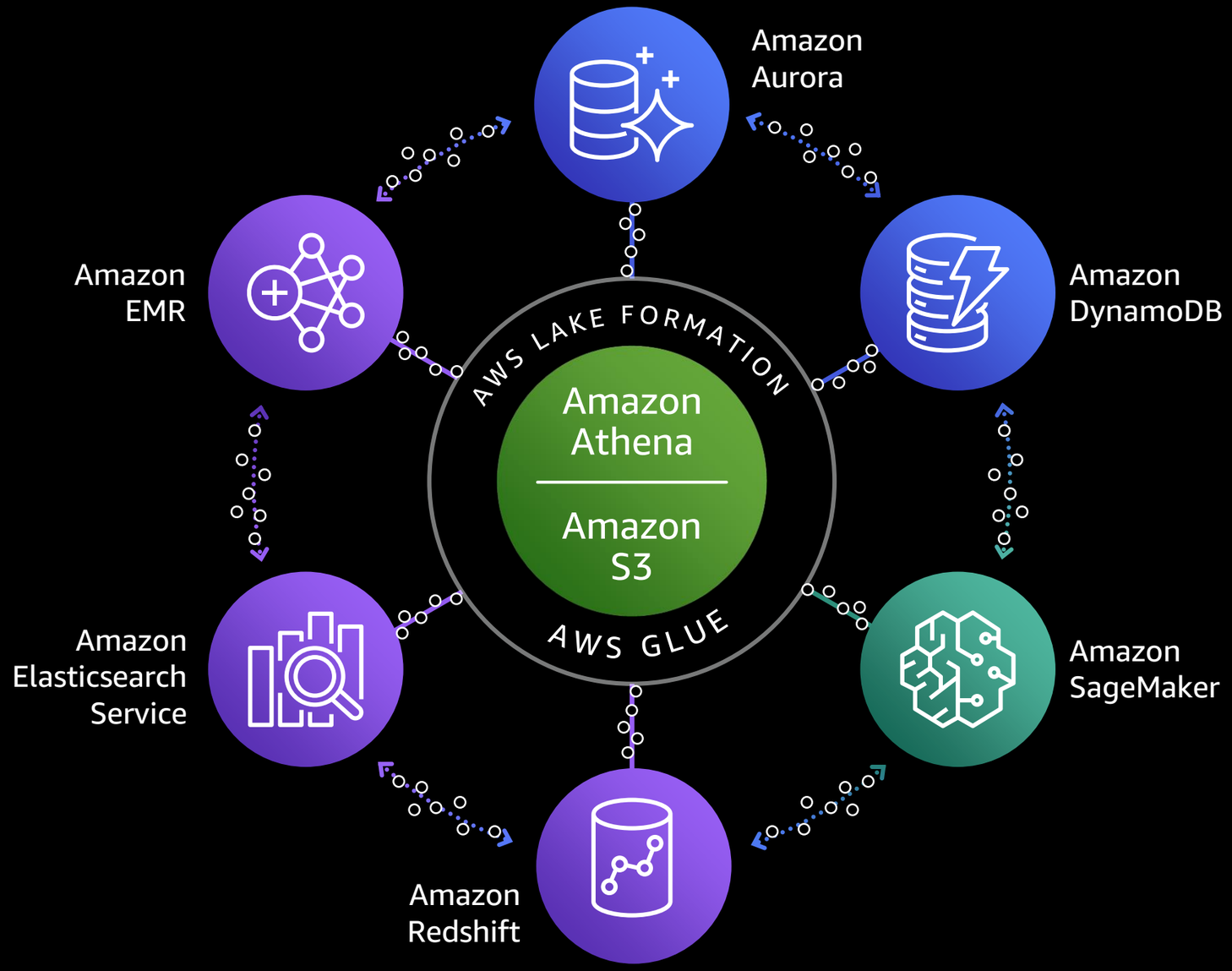
PURPOSE-BUILT DATA SERVICES

SEAMLESS DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

Lake House architecture on AWS



SCALABLE DATA LAKES

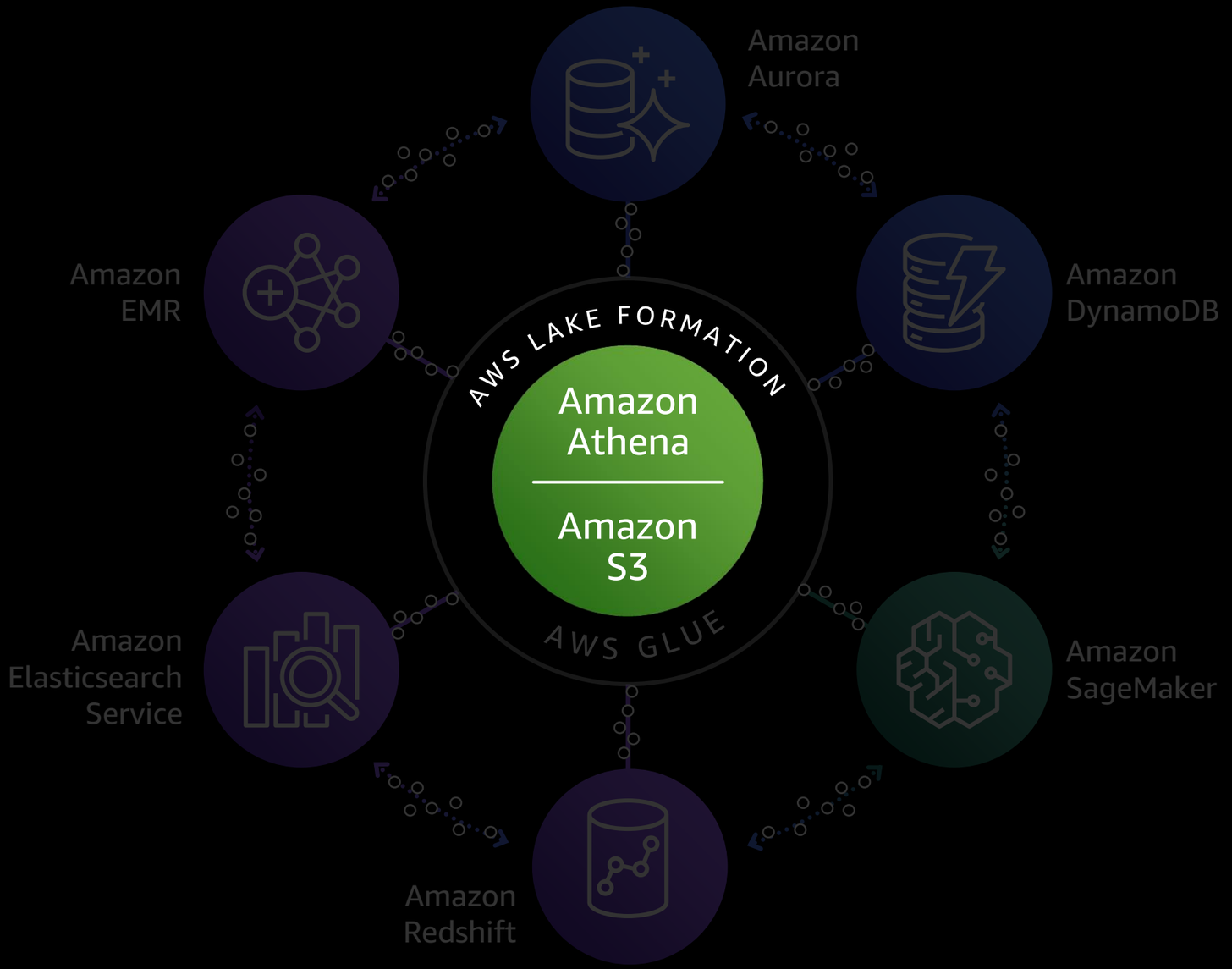
PURPOSE-BUILT
DATA SERVICES

SEAMLESS
DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND
COST-EFFECTIVE

Scalable data lakes



Amazon S3 is the most popular choice for data lakes

Unmatched **durability, availability,** and **scalability**

Easiest to use with **cost optimization:** Intelligent tiering

Most ways to **get data in**



Most **object-level controls**

Broadest portfolio of analytics tools

Best **security, compliance,** and **audit** capabilities

Cold storage and archive capabilities

AWS Lake Formation

BUILD A SECURE DATA LAKE IN DAYS



Build data lakes quickly

Move, store, and catalog your data faster

Simplify security management

Centrally define security, governance, and auditing policies in one place

Provide self-service access to data

Build a catalog of datasets

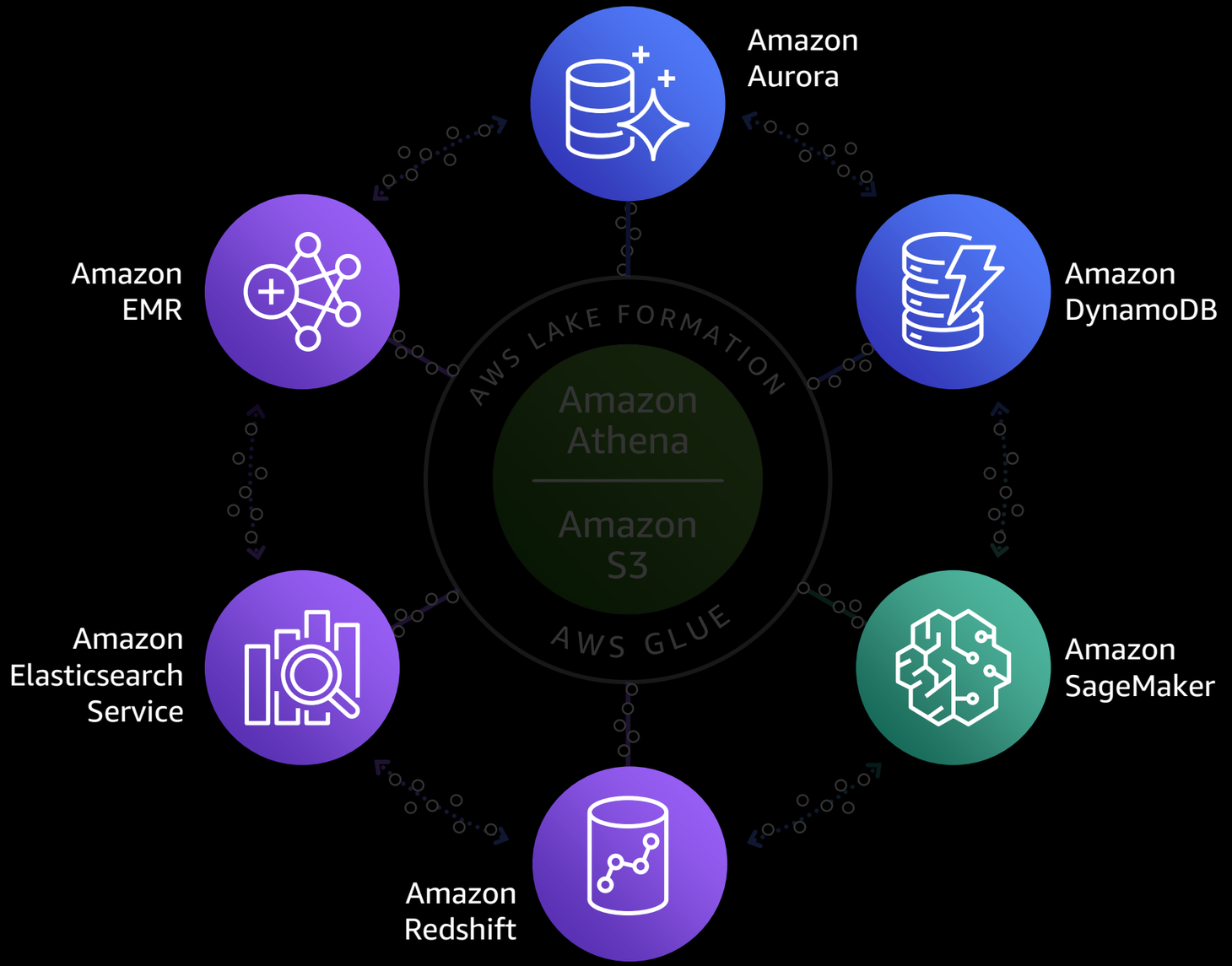
Define user access policies

More data lakes run on AWS than anywhere else

TENS OF THOUSANDS OF DATA LAKES RUN ON AWS ACROSS ALL INDUSTRIES



Purpose-built data services



Purpose-built analytics services

OPTIMIZE PERFORMANCE, COST, AND SCALE FOR YOUR USE CASES



Amazon Athena

Interactive query



Amazon EMR

Big data processing



Amazon Elasticsearch Service

Log and search analytics



Amazon Kinesis

Real-time analytics

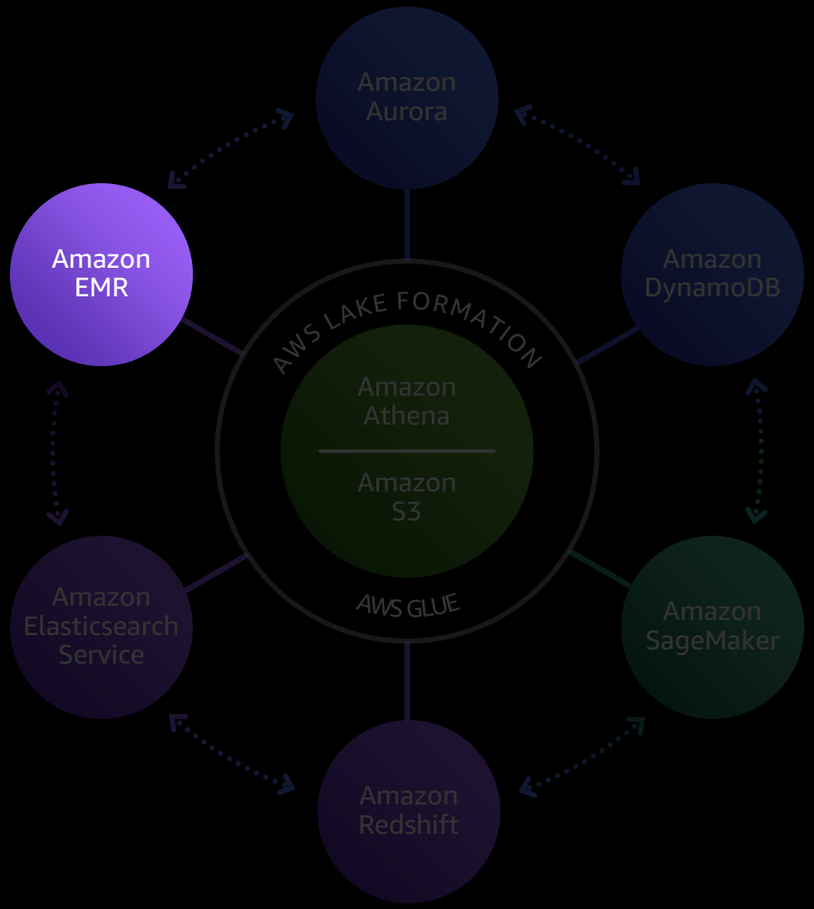


Amazon Redshift

Data warehousing

Amazon EMR

EASILY RUN SPARK, HADOOP, HIVE, PRESTO, HBASE, AND OTHER BIG DATA FRAMEWORKS



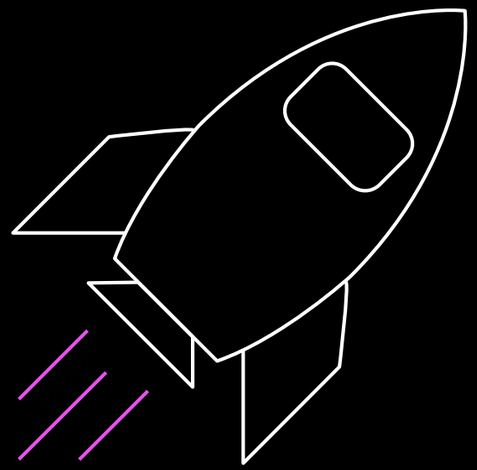
Automate provisioning, configuring, and tuning
Easy setup, management, and monitoring

Get the latest, stable, open-source releases
Latest open-source framework updates within 30 days

Automatically scale up and down
Manage cluster size based on utilization to reduce costs

Simple and predictable pricing
Per-second pricing, and save 50%–80% with Amazon EC2 Spot and Reserved Instances

Amazon EMR performance improvements



1.7x faster performance than standard Apache Spark 3.0 at 40% of the cost

---

Up to 2.6x faster performance than open-source Presto 0.238 at 80% of the cost

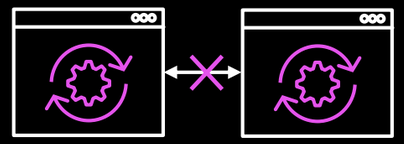
---

11.5% average performance improvement with Graviton2

25.7% average cost reduction with Graviton2

Until now, you had to choose between

Fully managed Amazon EMR on Amazon EC2



DEDICATED RESOURCES CAN'T BE SHARED

Self-managed Spark on Amazon EKS



COMPLEXITY AND OPERATIONAL OVERHEAD



CUSTOMERS WANT THE BEST OF BOTH WORLDS

Announcing

GENERALLY AVAILABLE



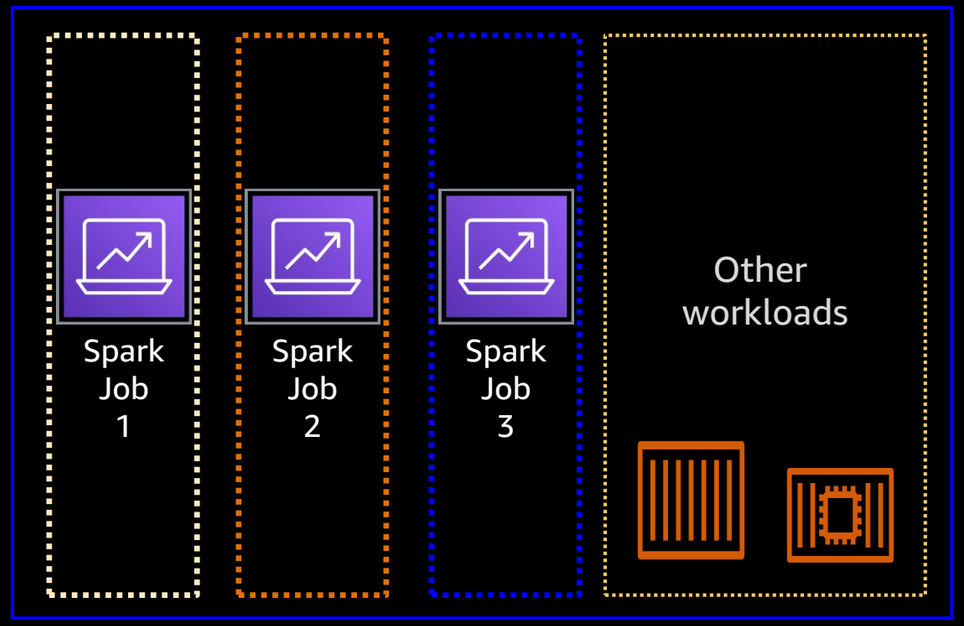
Amazon EMR for Amazon EKS

RUN APACHE SPARK JOBS ON AMAZON ELASTIC KUBERNETES SERVICE (AMAZON EKS)

Amazon EMR on Amazon EKS

INTRODUCING
GA

RUN APACHE SPARK JOBS ON AMAZON EKS



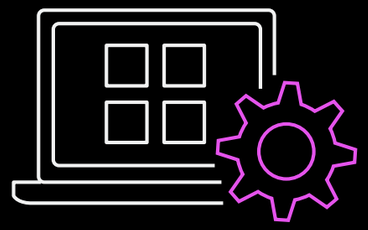
Amazon EKS

Eliminate underutilized and overprovisioned resources

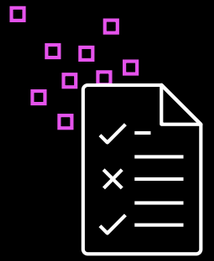
Automate management of big data frameworks

Optimize Apache Spark performance

Self-managed notebooks have challenges



CONSOLE-ONLY ACCESS



DISJOINTED WORKFLOW



DIFFICULT TO DEBUG

Announcing

PREVIEW



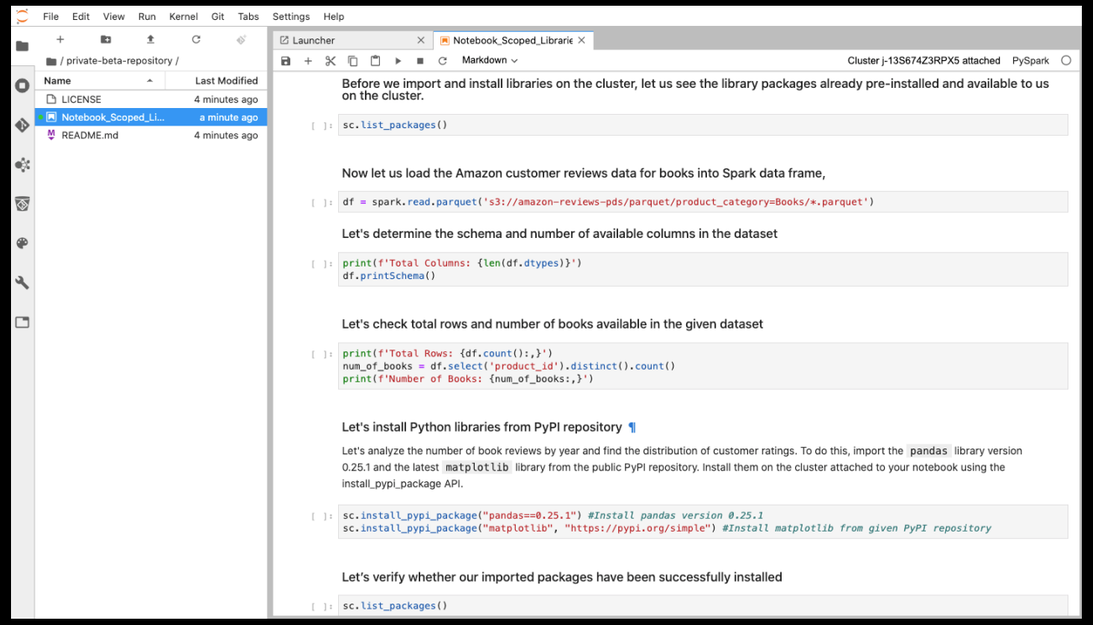
Amazon EMR Studio

FULLY MANAGED NOTEBOOKS WITH SSO INTEGRATION

Amazon EMR Studio

INTRODUCING
PREVIEW

FULLY MANAGED NOTEBOOKS WITH SSO INTEGRATION



Use corporate identity to log directly into notebooks

Develop, visualize, debug, and optimize analytics and ML apps

Collaborate with peers by sharing notebooks via GitHub

Build pipelines using orchestration services like Apache Airflow



Amazon Kinesis

EASILY COLLECT, PROCESS,
AND ANALYZE DATA AND
VIDEO STREAMS IN REAL TIME



Kinesis Data Streams

Capture, process, and store data streams



Kinesis Data Firehose

Load data streams into AWS data stores



Kinesis Data Analytics

Analyze data streams with SQL or Apache Flink



Kinesis Video Streams

Capture, process, and store video streams



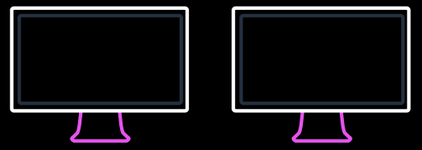
Managed streaming for Kafka

Fully managed, highly available, and secure Apache Kafka service

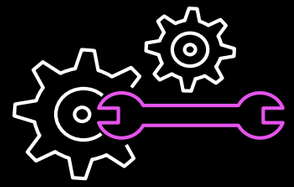
Processing and retaining real-time and historical data is increasingly common



DATA IS STORED IN TWO DIFFERENT SYSTEMS



NEED TWO SETS OF PROCESSING LOGIC



ADDS COMPLEXITY AND OPERATIONAL OVERHEAD

Announcing

GENERALLY AVAILABLE



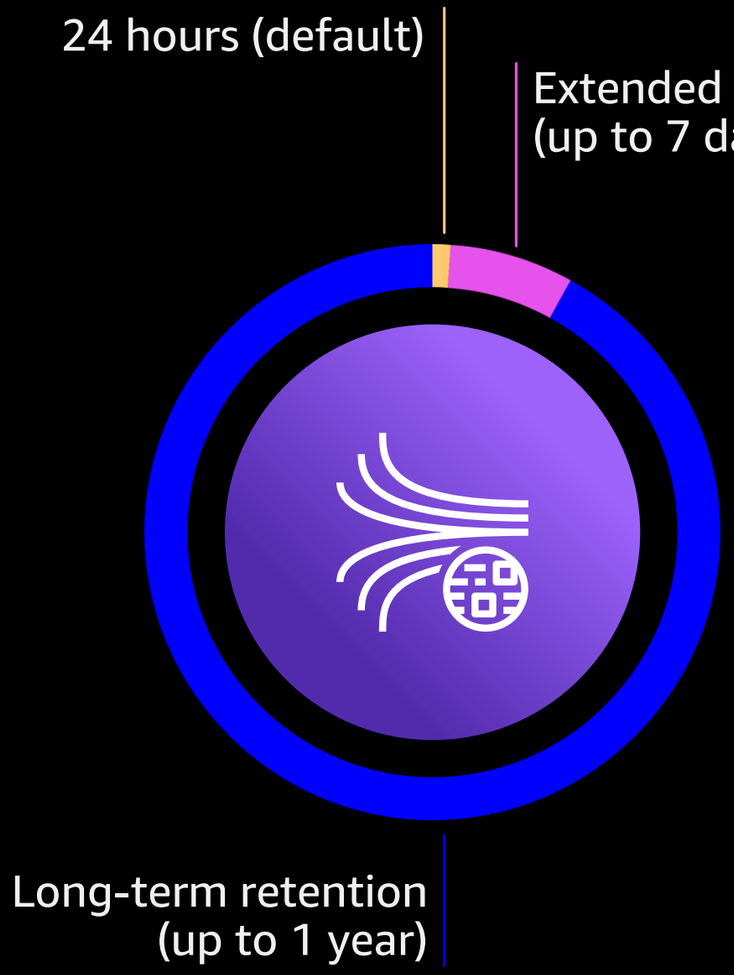
Amazon Kinesis Data Streams: Long-term retention

RETAIN DATA STREAMS FOR UP TO 1 YEAR

Amazon Kinesis Data Streams: Long-term retention

INTRODUCING
GA

RETAIN DATA STREAMS FOR UP TO 1 YEAR



Satisfy compliance requirements without moving data

Same processing code for real-time and historical data

Simplify analytical processing

Amazon Redshift

ANALYZE ALL YOUR DATA WITH THE FASTEST AND MOST WIDELY USED CLOUD DATA WAREHOUSE



Analyze all your data

Deepest integration with your data lake

Performance at any scale

Up to 3x better price performance than other cloud DW

Lower your costs

At least 50% less expensive than other cloud DW

Amazon Redshift innovates to meet your needs



Analyze all your data

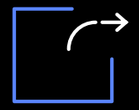
Lake House with AWS integration



Amazon Redshift Spectrum + Lake Formation



Data lake export



Federated query

NEW!



Data sharing

NEW!



Amazon Redshift ML



Performance & scale

Fast and self-tuning



Concurrency scaling



RA3 nodes & managed storage



AQUA



Materialized views

NEW!



Automated perf. tuning



Low cost & best value

Predictable costs



On-demand and RIs

NEW!



Cross-AZ cluster recovery



Pause and resume



Cost controls



Built-in security features

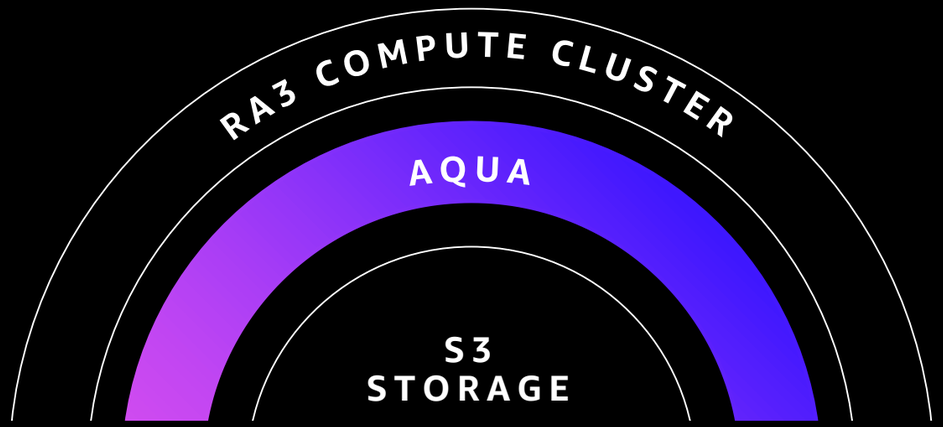


Automatic workload manager

Amazon Redshift: Advanced Query Accelerator

COMING SOON
GA

UP TO 10X BETTER QUERY PERFORMANCE THAN OTHER CLOUD DATA WAREHOUSES
(WITHOUT INCREASING COSTS)

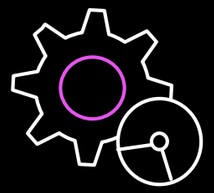


High-speed cache on top of Amazon S3 can scale out to process data in parallel across many nodes

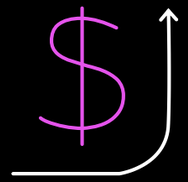
Brings the compute to the storage layer

AWS-designed analytics processors accelerate data compression and encryption

Sharing data across your organization and externally can be complex



TIME CONSUMING TO MANUALLY UNLOAD AND COPY



EXPENSIVE WITH DELAYS AS USERS GROW



SECURITY AND GOVERNANCE IS COMPLEX

Announcing

PREVIEW



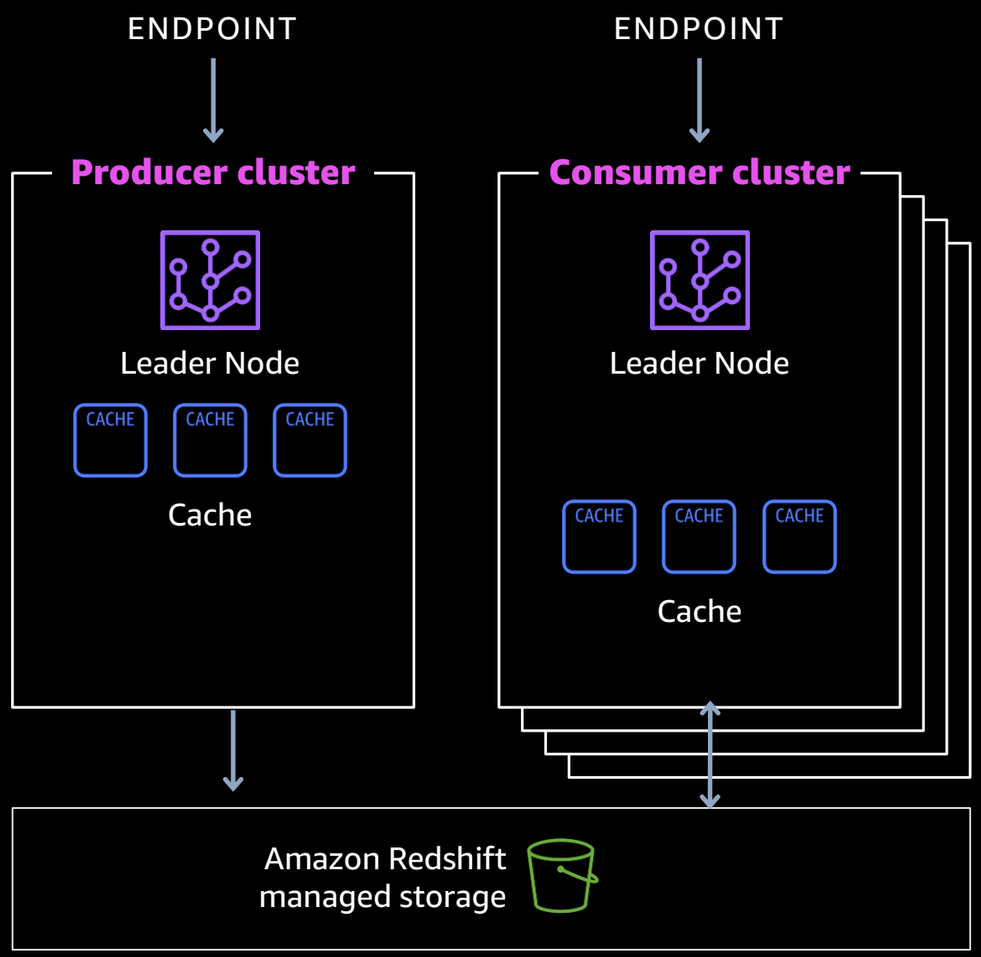
Amazon Redshift: Data sharing

SHARE LIVE AND CONSISTENT DATA WITHIN AND ACROSS YOUR ORGANIZATION AND EXTERNAL PARTIES

Amazon Redshift: Data sharing

INTRODUCING
PREVIEW

SHARE LIVE AND CONSISTENT DATA WITHIN AND ACROSS YOUR ORGANIZATION AND EXTERNAL PARTIES



Simple and direct way to share live and consistent data

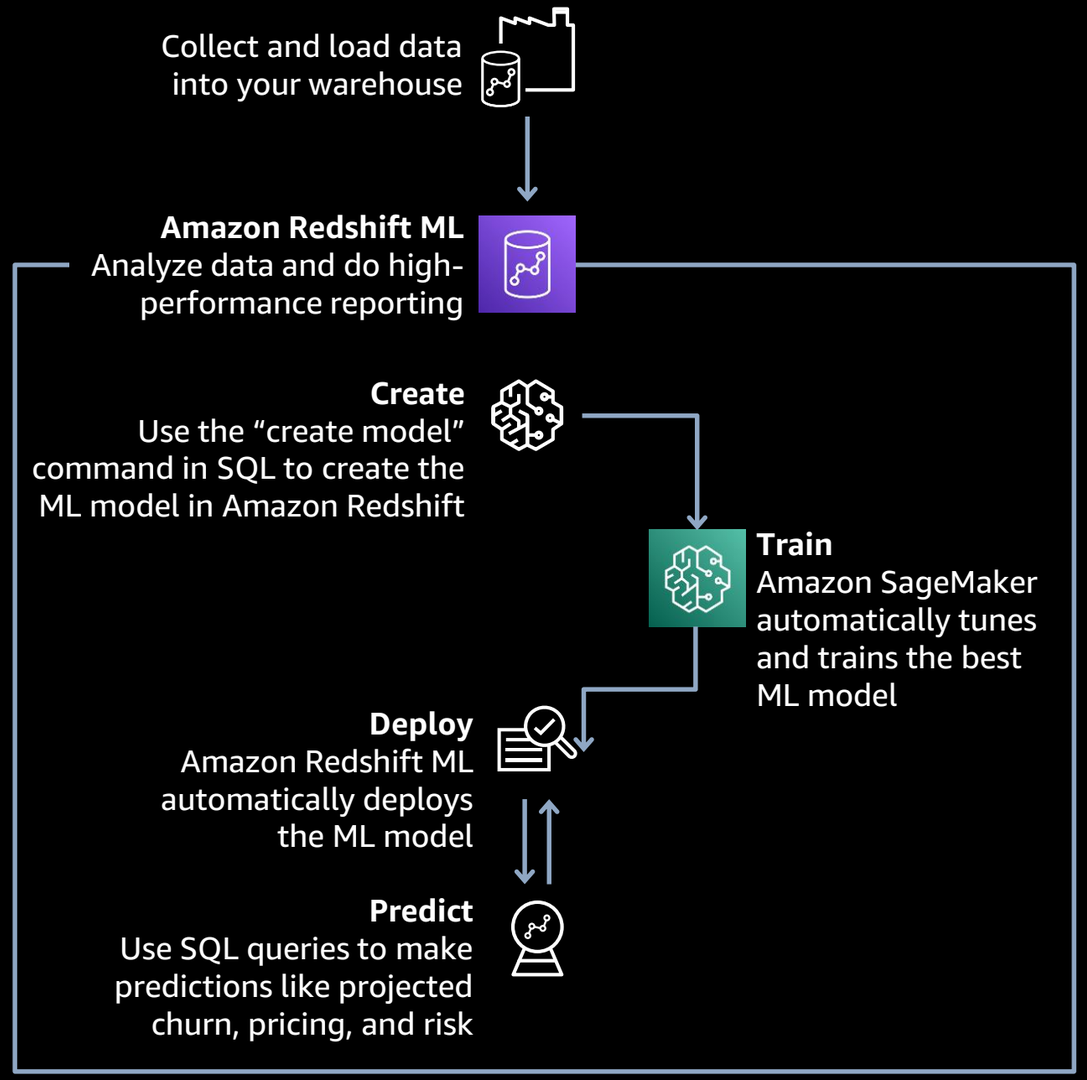
Set up individual cluster configurations to optimize for performance, cost, or both

Share data across multiple groups in a secured and governed way

Amazon Redshift ML

NEW
PREVIEW

SQL USERS CAN CREATE MODELS AND MAKE PREDICTIONS ON DATA IN THEIR DATA WAREHOUSE

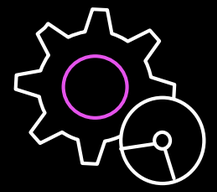


Automatically create and train ML models

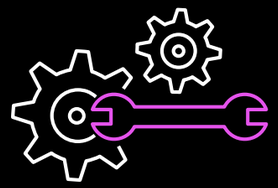
Apply ML models on Amazon Redshift data using standard SQL

Embed predictions like fraud detection and risk scoring

Challenges with manual performance tuning



TIME CONSUMING



REQUIRES SPECIAL SKILLS



CAN BE DONE INCORRECTLY

Announcing

GENERALLY AVAILABLE



Amazon Redshift: Performance tuning – Automated

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

Amazon Redshift: Performance tuning – Automated

INTRODUCING
GA

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



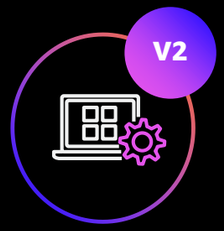
Automatic vacuum delete



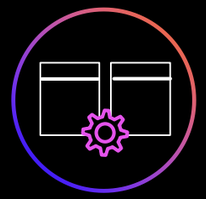
Automatic distribution keys



Automatic sort keys



Workload management



Automatic table sort



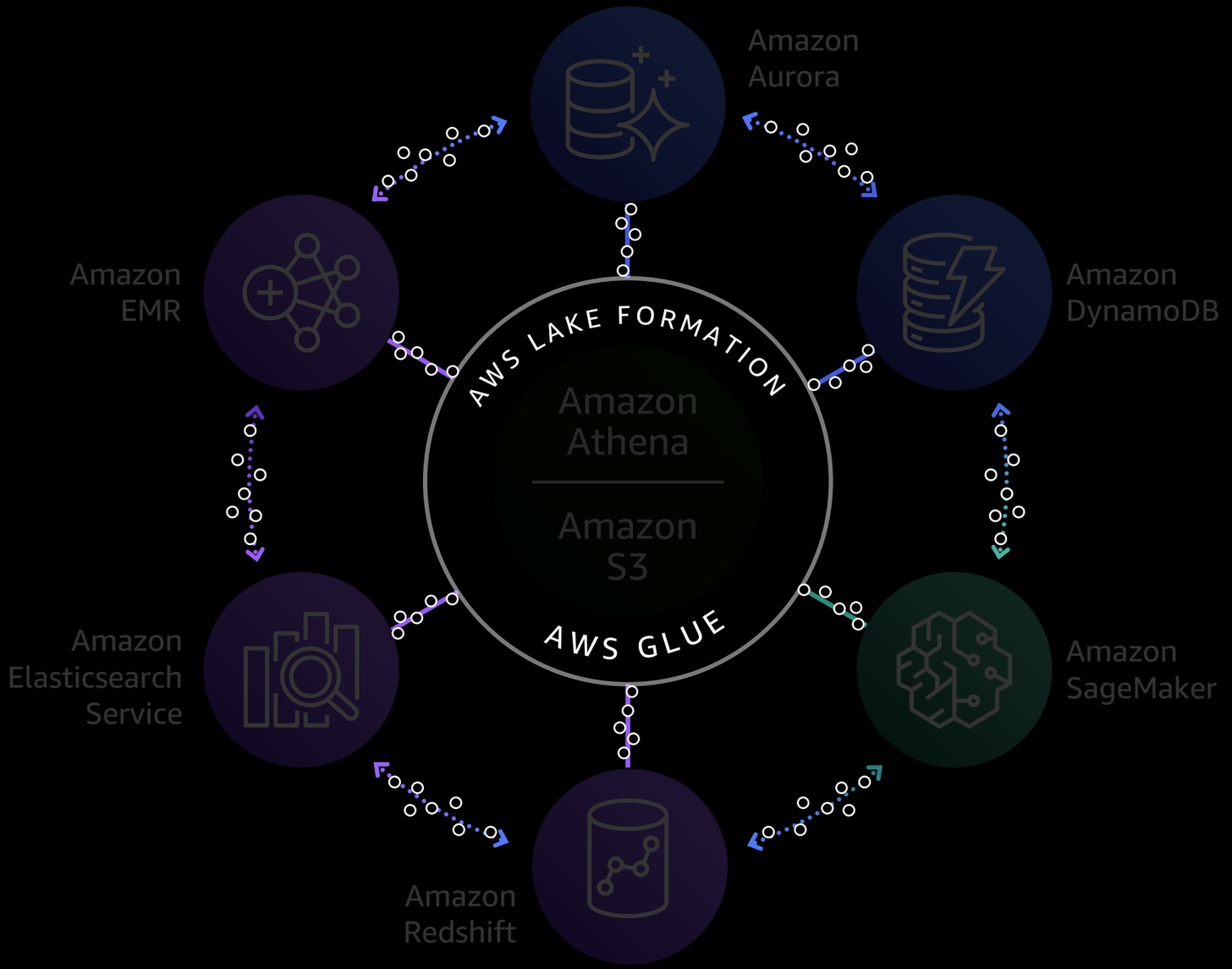
MV autorefresh and rewrite

Automates physical data design and optimization

Optimizes for peak performance as data and workloads scale

Leverages ML to adapt to shifting workloads

Seamless data movement



AWS Glue

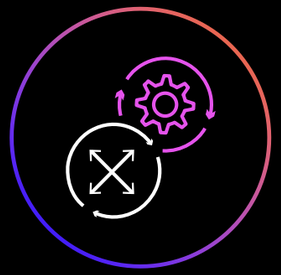
SIMPLE, SCALABLE, AND SERVERLESS DATA INTEGRATION





Integrate data faster

Prepare and put your data to use from months to minutes



Automate at scale

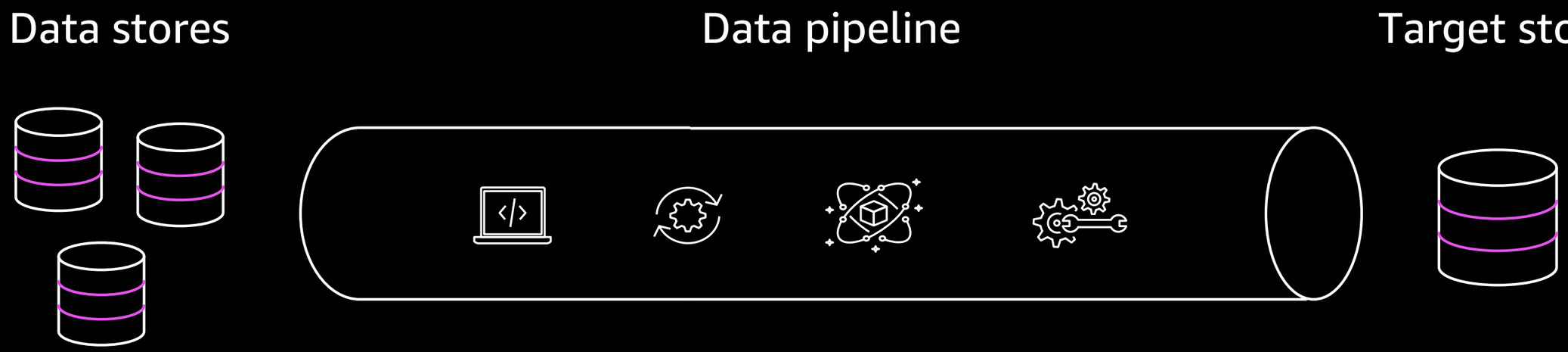
Easily run and manage thousands of ETL jobs

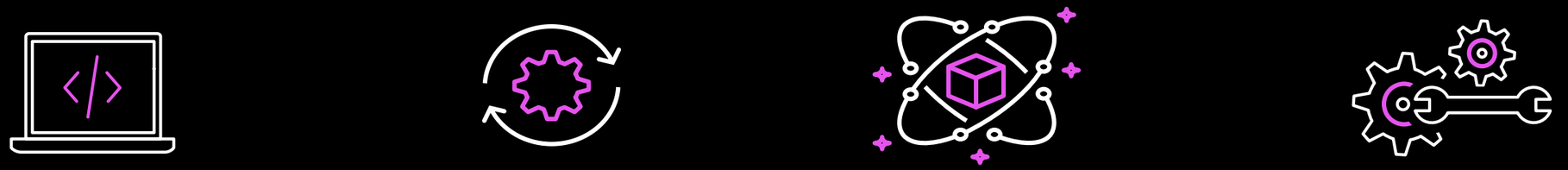


No servers to manage

Pay only for the resources your jobs use while running

Combining and replicating data between multiple stores can be challenging





COMPLEX APPLICATION CODE

CUMBERSOME RETRY LOGIC

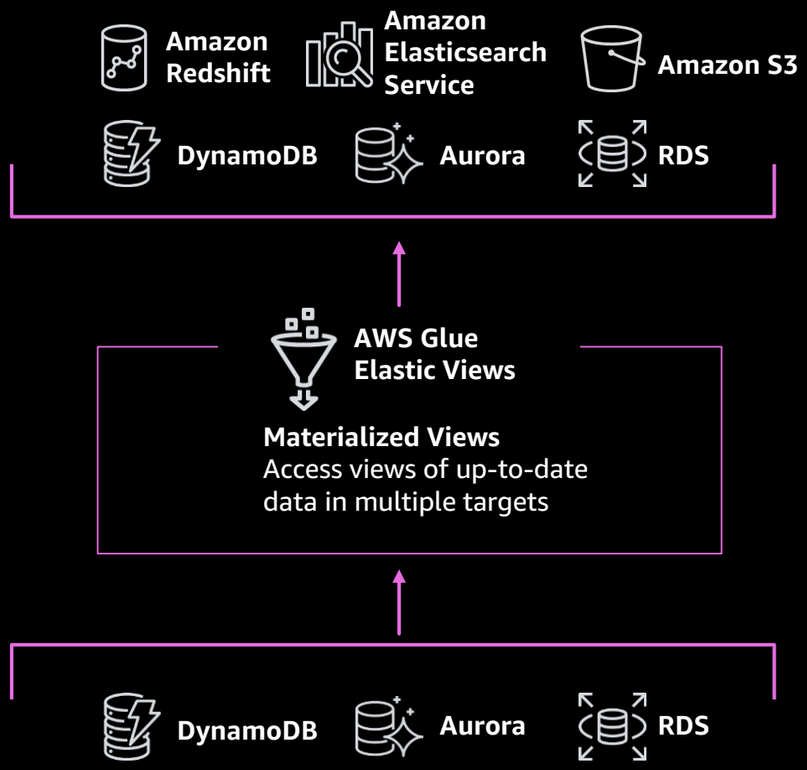
PIPELINE MANAGEMENT

NEED SPECIALIZED ETL SKILLS

AWS Glue Elastic Views

NEW
PREVIEW

EASILY COMBINE AND REPLICATE DATA ACROSS MULTIPLE DATA STORES

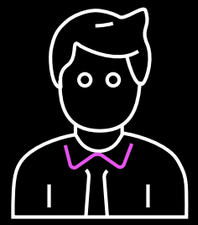


Create materialized views across a wide variety of databases and data stores using familiar SQL

Continually monitors source databases for changes and updates targets within seconds

Serverless and automatically scales capacity up and down to accommodate your workloads

Handles the heavy lifting of copying and combining data without requiring custom code



ETL DEVELOPER

AWS Glue Studio

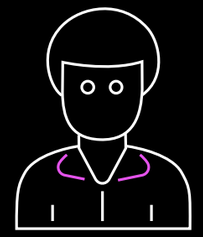
Visual authoring of ETL jobs without writing code

Monitor thousands of jobs through a single pane of glass

Distributed processing without the learning curve

Advanced transforms though code snippets





BUSINESS ANALYST

DATA SCIENTIST

NEW
GA

AWS Glue DataBrew

Clean and normalize data with a rich visual interface

Choose from 250+ built-in transformations to automate tasks

Profile data to understand data patterns and anomalies

Work on large datasets at scale

AWS Glue DataBrew

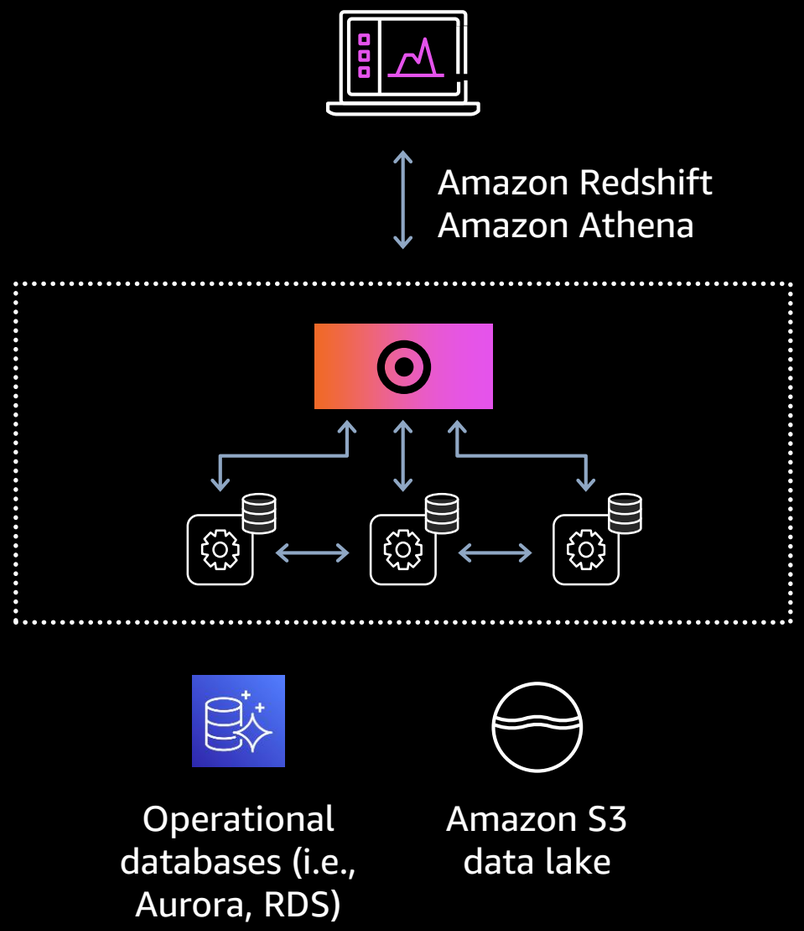


AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Federated query in Amazon Redshift and Athena

UNIFIED ANALYTICS ACROSS DATABASES, DATA WAREHOUSE, AND DATA LAKE



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

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

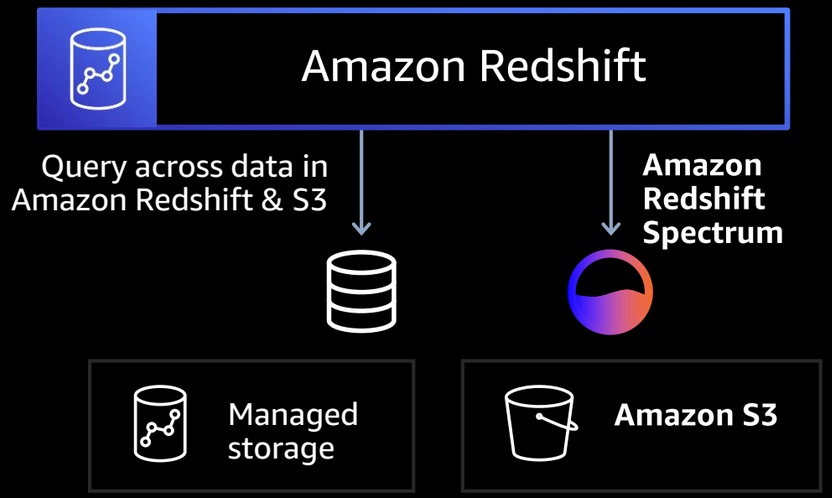
Amazon Redshift Spectrum

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



Amazon Redshift: Data lake export

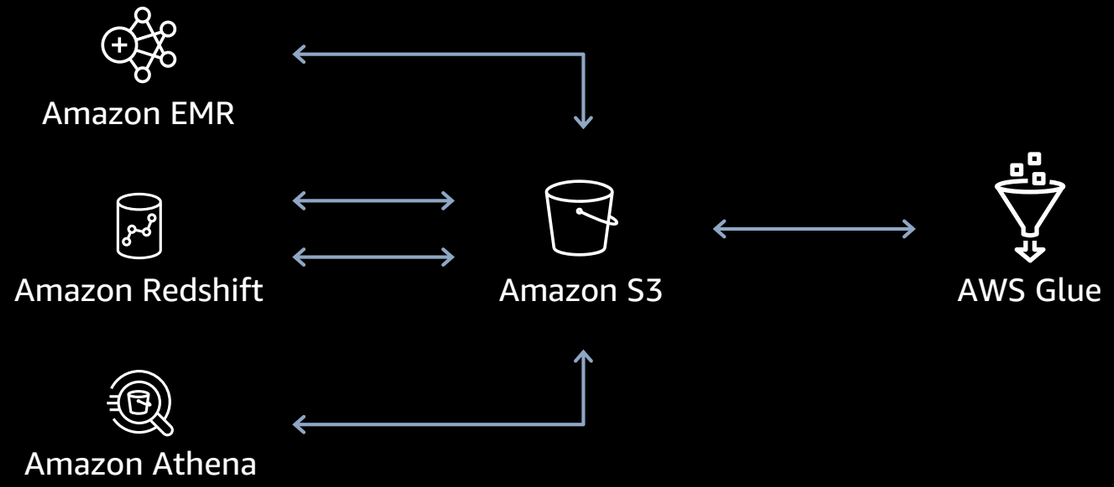
UNLOAD AMAZON REDSHIFT DATA AS PARQUET TO AMAZON S3 DATA LAKES FOR FASTER SHARING AND ANALYTICS

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

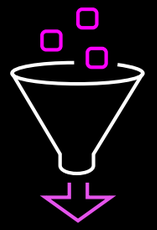
Use SQL with Amazon Redshift's Unload command to export data in Parquet format

Unloaded data is automatically registered in AWS Glue Data Catalog

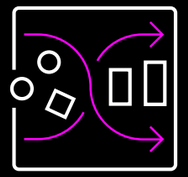


Seamless data movement

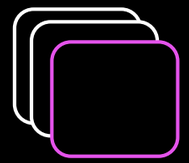
MOVE YOUR DATA, AT SCALE, TO WHERE YOU NEED IT THE MOST



EXTRACT, TRANSFORM, LOAD



VISUAL DATA PREPARATION



DATA REPLICATION

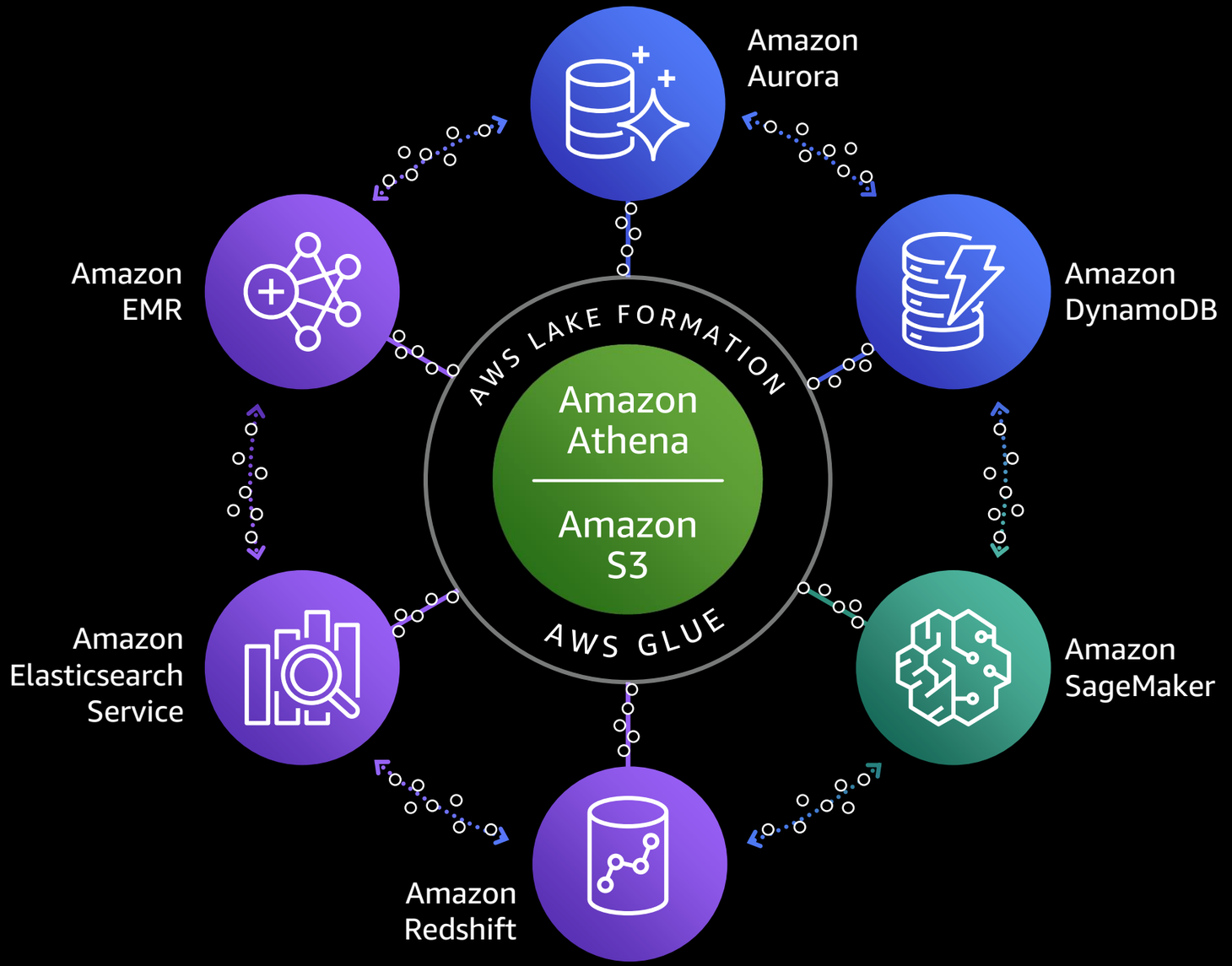


DATA WAREHOUSE TO/FROM DATA LAKE



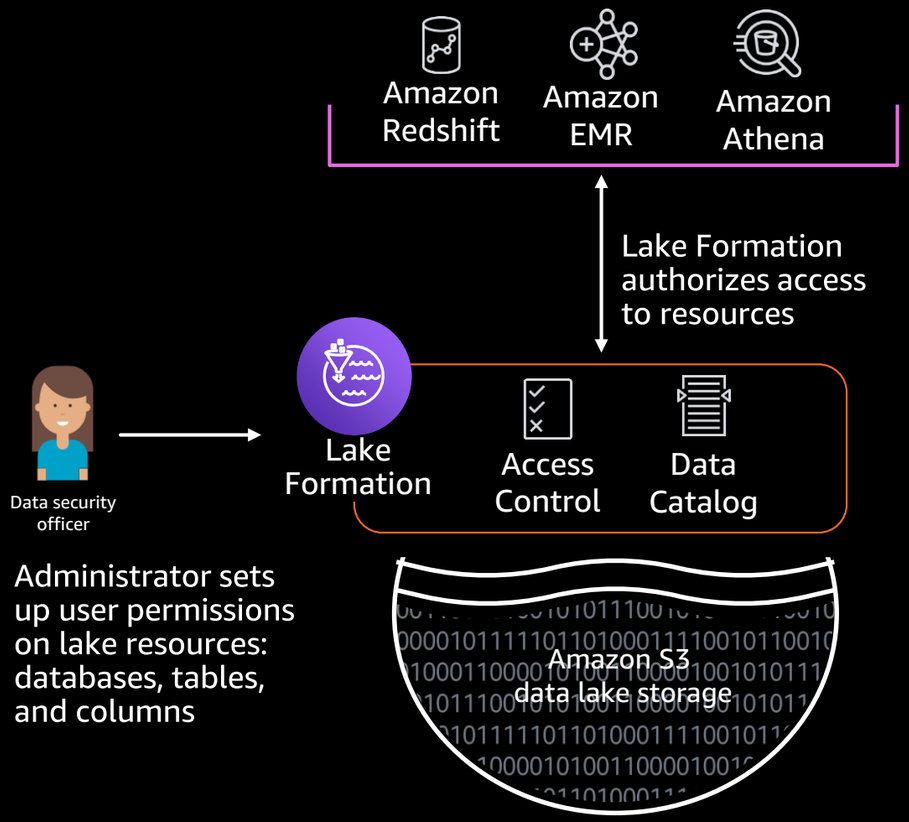
FEDERATED QUERY

Unified governance



Simplify security management with Lake Formation

CENTRALLY DEFINE SECURITY, GOVERNANCE, AND AUDITING POLICIES IN ONE PLACE



Centrally define security, governance, and auditing

Policies are consistently enforced

Integrated with security, storage, analytics, and machine learning services

Permissions on databases, tables, and columns

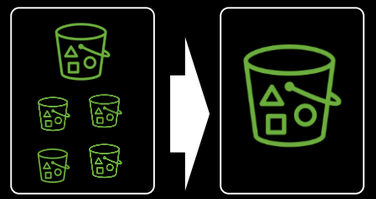
Challenges of building and securing modern data lakes



SUPPORT UPDATES AND DELETES FOR STREAMING DATA, DATABASE REPLICATION, AND PRIVACY REGULATIONS



ROW-LEVEL FINE-GRAINED SECURE SHARING



AUTOMATIC STORAGE OPTIMIZATION

Announcing

PREVIEW



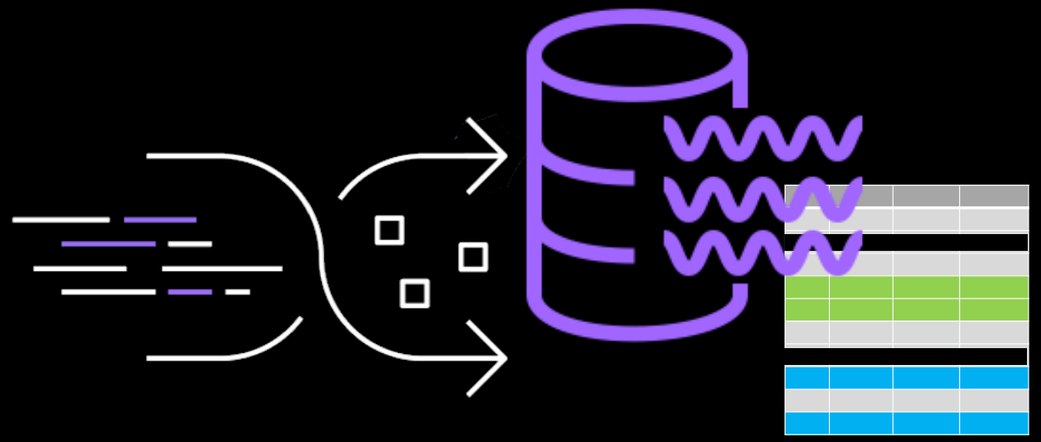
AWS Lake Formation: Transactions, row-level security, and acceleration

ACCELERATE AND GOVERN ACCESS TO YOUR AMAZON S3 DATA LAKE

AWS Lake Formation

INTRODUCING
PREVIEW

TRANSACTIONS, ROW-LEVEL SECURITY, AND ACCELERATION



ACID transactions for concurrent updates

Security and governance with granular row-level filtering

Improve performance by optimizing data layout and data retrieval operations

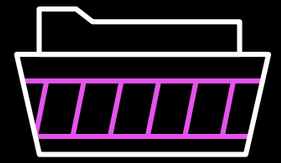
Unified governance across your data lakes and data stores



SECURE YOUR DATA

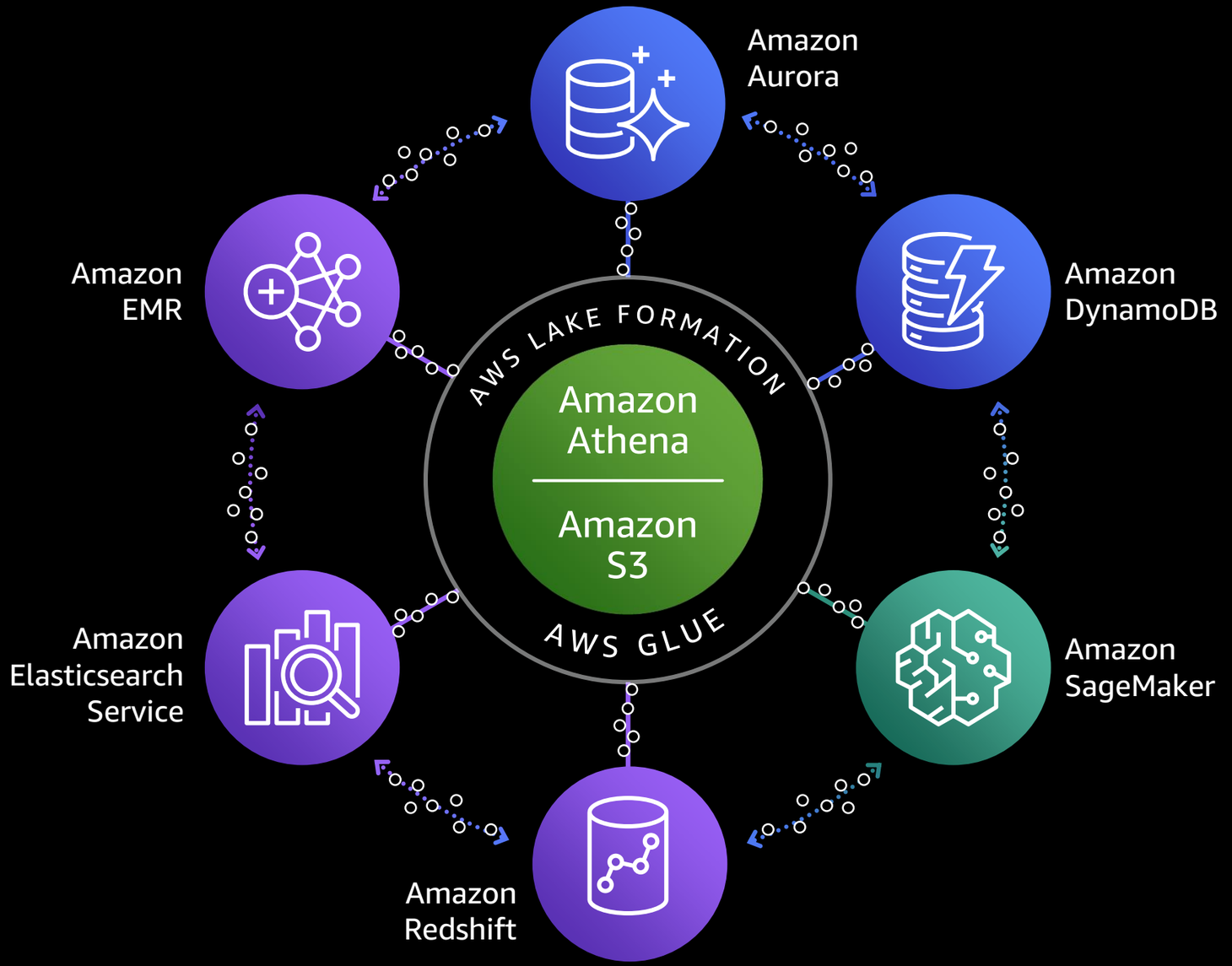


CONTROL ACCESS

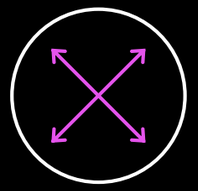


AUDIT TRAIL

Performant and cost-effective



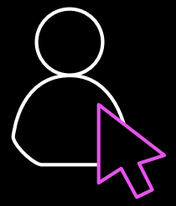
Performant and cost-effective



INDUSTRY-LEADING CHOICE OF 200+ INSTANCE TYPES TO MEET WORKLOAD NEEDS



100 GBPS BANDWIDTH NETWORK INTERFACES FOR PERFORMANCE



ON-DEMAND, RESERVED, AND SPOT INSTANCES TO REDUCE COSTS



FIVE HIGHLY AVAILABLE STORAGE TIERS AND INTELLIGENT TIERING

No compromises on performance, scale, and cost

Amazon Redshift	**3x better price performance** than other cloud data warehouses Automated performance tuning and near-linear scaling
Amazon EMR	Optimized runtimes that provide the best price-performance **1.7x faster** than standard Apache Spark; **2.6x faster** than standard Presto
Amazon Elasticsearch Service	UltraWarm storage tier **reduces costs by 90%** Store **6x more log data** without increasing costs
Amazon S3	Amazon S3 Select retrieves a subset of data leading to queries that run **up to 400% faster** Amazon S3 intelligent tiering **saves up to 70% on storage costs** for data lakes
aws Compute integration	With Graviton2 instance, customers **save 25.7%** for typical workloads

Reinventing business intelligence







Amazon QuickSight

A SCALABLE, EMBEDDABLE, ML-POWERED BI SERVICE BUILT FOR THE CLOUD



BI at scale

No servers to manage; pay per use billing



Embedded analytics

Quickly embed dashboards in your apps



ML-powered insights

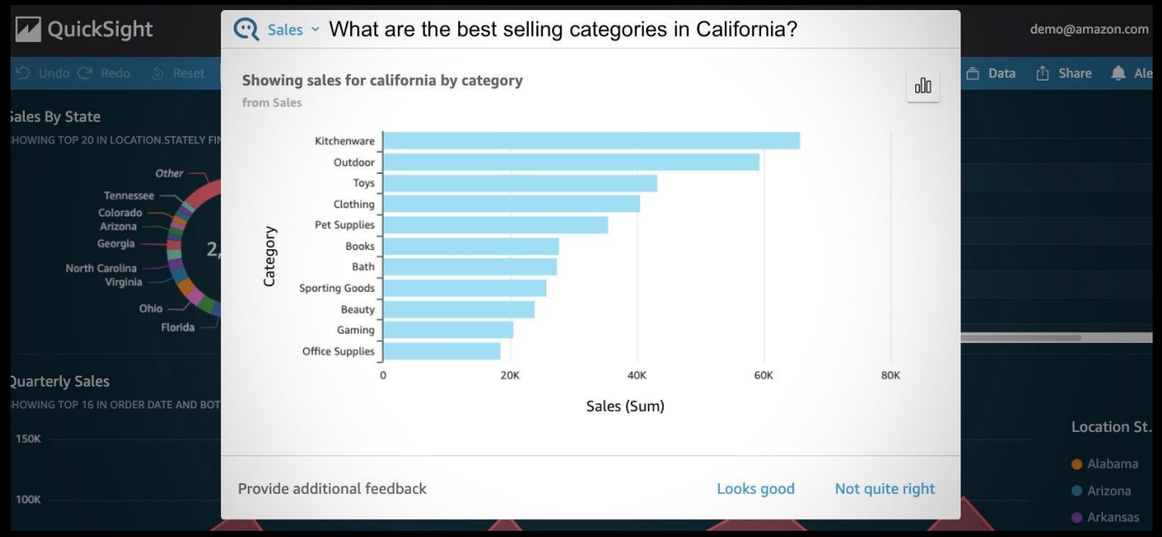
Built-in anomaly detection and forecasting

Written narratives to interpret your data for you

Amazon QuickSight Q

NEW
PREVIEW

ML-POWERED NATURAL LANGUAGE CAPABILITY IN AMAZON QUICKSIGHT



Enter business questions in search bar and get answer in seconds

ML generates data models that automatically understands meanings and relationships

Not limited to only asking a specific set of questions

Amazon QuickSight Q



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Accelerating innovation with AWS analytics

PURPOSE-BUILT DATA SERVICES

- AMAZON EMR ON AMAZON EKS
- AMAZON EMR STUDIO
- AMAZON KINESIS DATA STREAMS LONG-TERM RETENTION
- AMAZON REDSHIFT DATA SHARING
- AMAZON REDSHIFT ML

SEAMLESS DATA MOVEMENT

- AWS GLUE SCHEMA REGISTRY
- AWS GLUE DATABREW
- AWS GLUE ELASTIC VIEWS

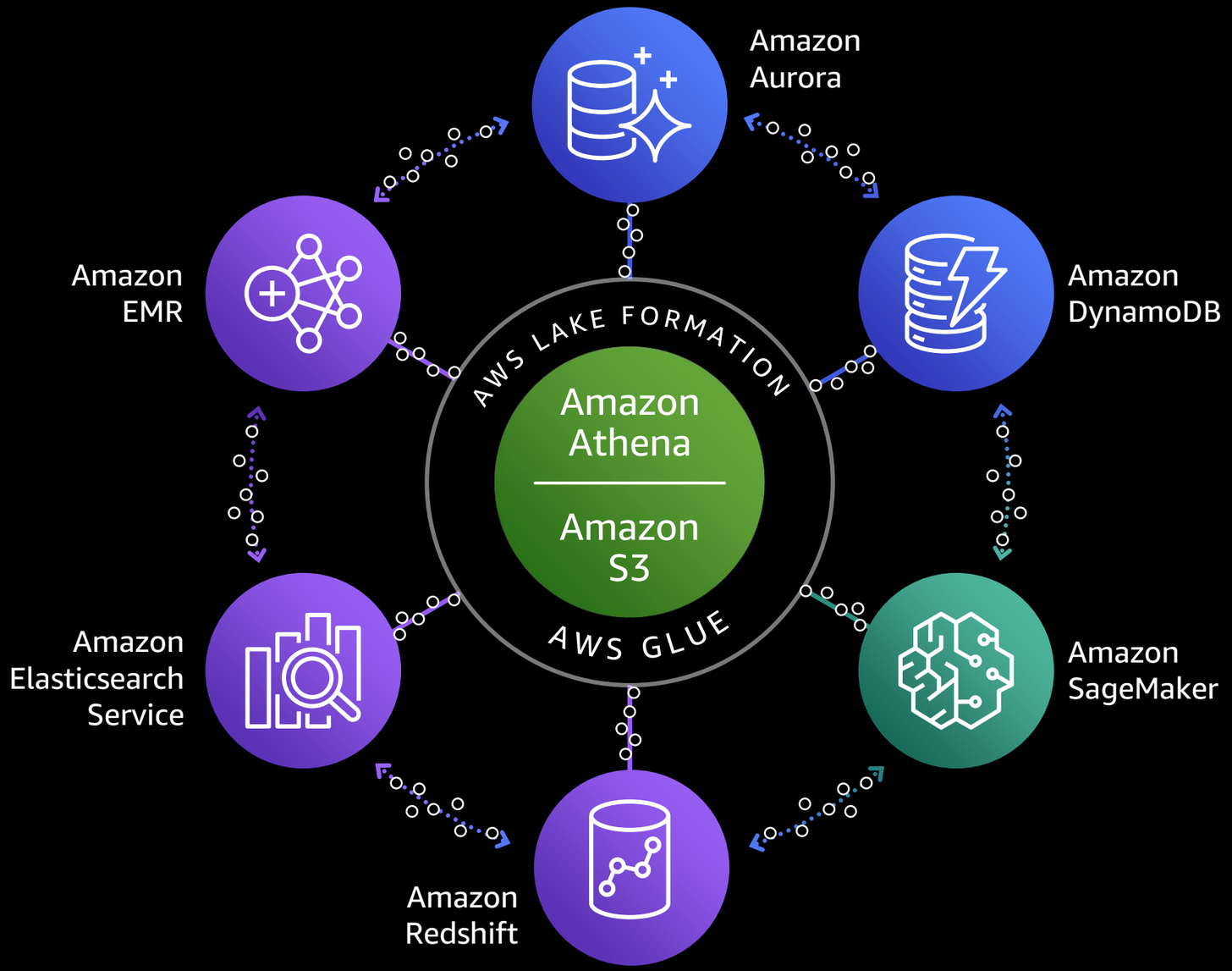
UNIFIED GOVERNANCE

- AWS LAKE FORMATION: TRANSACTIONS, ROW-LEVEL SECURITY, ACCELERATION

PERFORMANT AND COST-EFFECTIVE

- AMAZON REDSHIFT AUTOMATED PERFORMANCE TUNING
- EMR PERFORMANCE IMPROVEMENTS
- AQUA FOR AMAZON REDSHIFT

Lake House architecture on AWS



SCALABLE DATA LAKES

PURPOSE-BUILT
DATA SERVICES

SEAMLESS
DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND
COST-EFFECTIVE

Learn more with these re:Invent sessions

SESSION ID	SESSION TITLE
ANT202	How Amazon Redshift powers large-scale analytics for Amazon.com
ANT203	What's new with Amazon Redshift
ANT302	How Vyaire uses AWS analytics to scale ventilator production & save lives
ANT303	Modernize log analytics with Amazon Elasticsearch Service
ANT304	Migrating a legacy data warehouse to Amazon Redshift
ANT305	How Goldman Sachs uses an Amazon MSK backbone for its Transaction Banking Platform
ANT306	Run big data analytics faster at lower cost with Amazon EMR
ANT307	What's new with Amazon EMR
ANT308	How FINRA operates PB-scale analytics on data lakes with Amazon Athena
ANT309	How Disney+ uses fast data ubiquity to improve the customer experience
ANT311	Under the hood: How Amazon uses AWS for analytics at petabyte scale
ANT312	Serverless data preparation with AWS Glue
ANT318	The right tool for the job: Enabling analytics at scale at Intuit
ANT320	What's new in Amazon Elasticsearch Service
ANT321	Top 5 best practices for data streaming with Amazon Kinesis
ANT322	New use cases for Amazon Redshift

Want to build a data vision and strategy?

aws data-driven everything

- 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?

aws data lab

- 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

Rahul Pathak

Vice President Analytics
Amazon Web Services

@rahulpathak
rapathak@amazon.com

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.


AWS
re:Invent