



Building Analytics at Scale with Amazon Athena

Suresh Akena

Principal WW GTM leader, Athena

Naresh Gautam

Principal Specialist SA, Analytics

Put data to work



Make better
decisions



Improve
efficiencies



Respond
faster



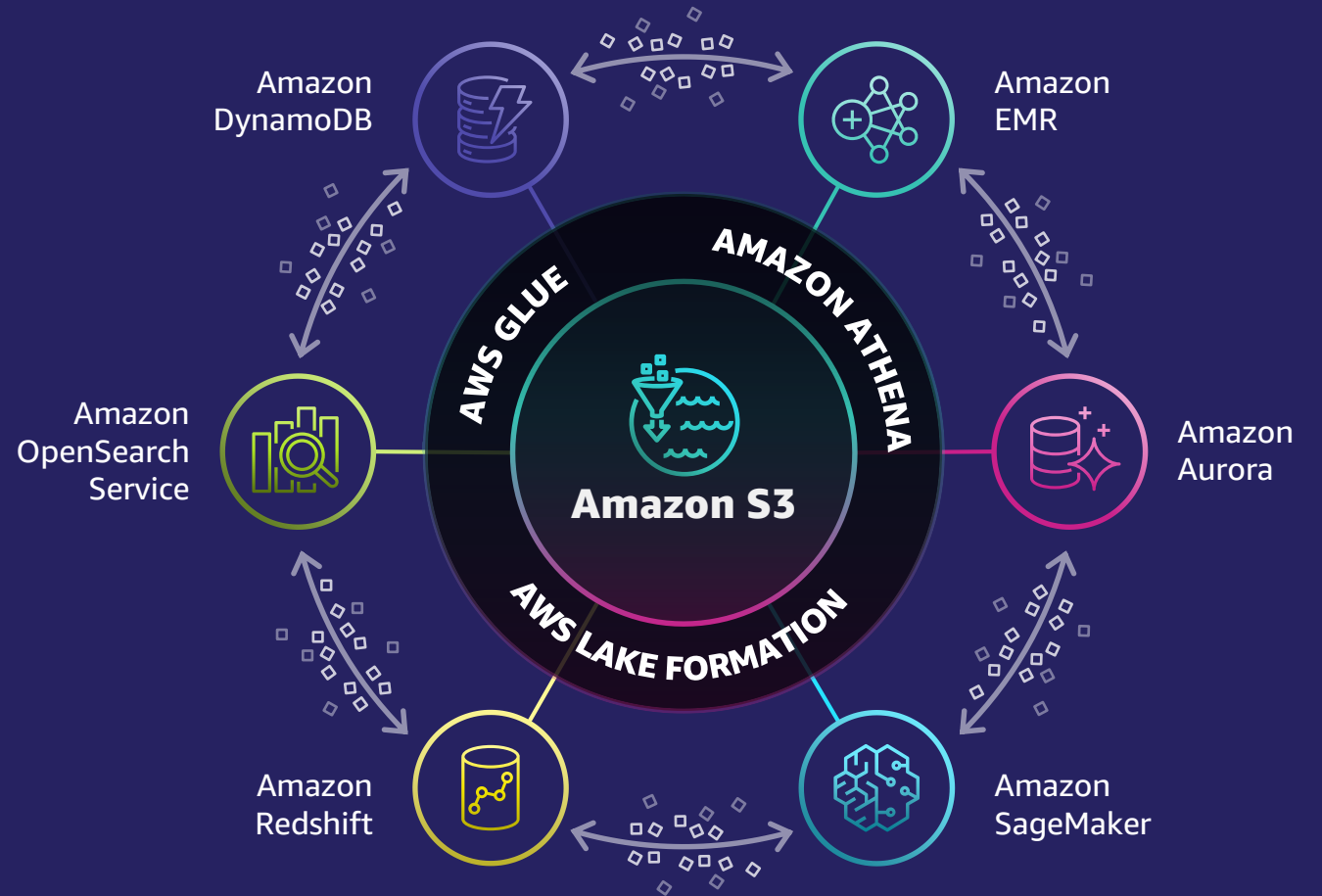
Uncover
opportunities

Traditional approaches don't scale





Modern data strategy on AWS



Amazon Athena

Query instantly



No cluster setup

Point to S3 and start analyzing

Cost-effective



Pay only for data scanned

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

Easy



No servers to manage

Access data in AWS and other data stores

Open & Standard



Use ANSI SQL for querying
Supports Parquet, ORC, Avro and other standard data formats



Caltech

J Sainsbury plc

zendesk



wejo

workiva

stripe

WaltDisneyCompany



acxiom.



Allianz

amazon.com

asurion

AUTODESK

AXONIX

Bristol-Myers Squibb

Carbon Black.

careerbuilder

dataxu.



edmunds



FICO



move

GitHub

happyfresh



HermanMiller

HONDA

HubSpot

IAC

IDEXX

illumina

J.D. POWER

JABIL



make my trip

Mobvista.



Movable Ink

moz://a

Nextdoor

News Corp



Plume

protegrity

punchh.

RavenPack

RICOH

Robinhood

ROVIO



Segment

shutterstock

SIEMENS

SKECHERS

slack

Snap Inc.

STRAVA

Sysco

+ a b l e a u

TRUECar.

Uber



Pricing

\$5/TB

Pay only for queries run

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

Free of charge

- ✓ DDL operations
- ✓ Query concurrency
- ✓ Data connectors

Athena Use Cases



Data analytics in Data lake

Your Data



Amazon RDS



Amazon S3

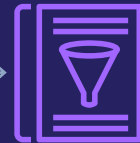


Amazon Aurora

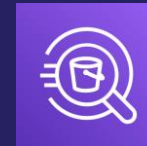


On premises
data

*Discover your data and
store associated
metadata*

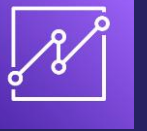


AWS Glue Data
Catalog

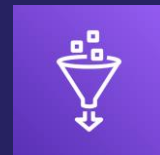


Amazon Athena

SQL
Clients



Amazon QuickSight



AWS Glue ETL

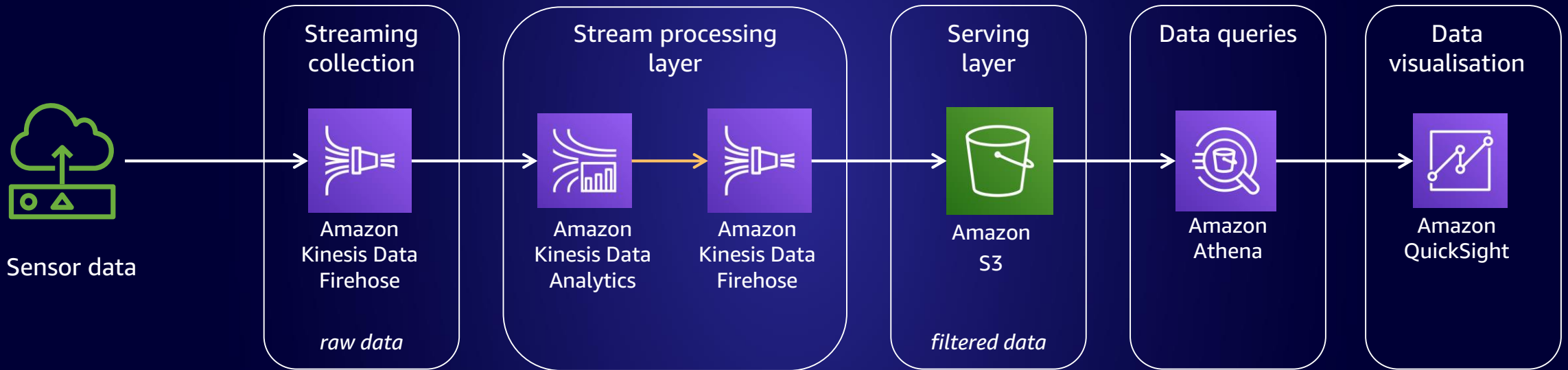
*Read and write
metadata to Glue
Catalog*



Amazon S3

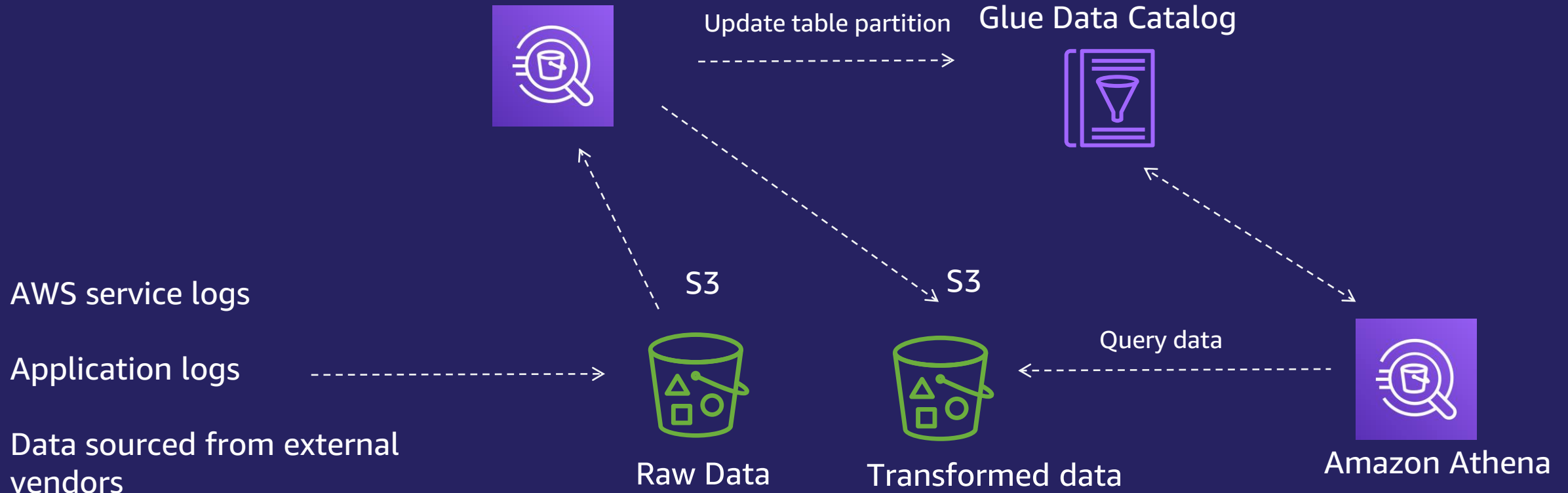


Stream processing

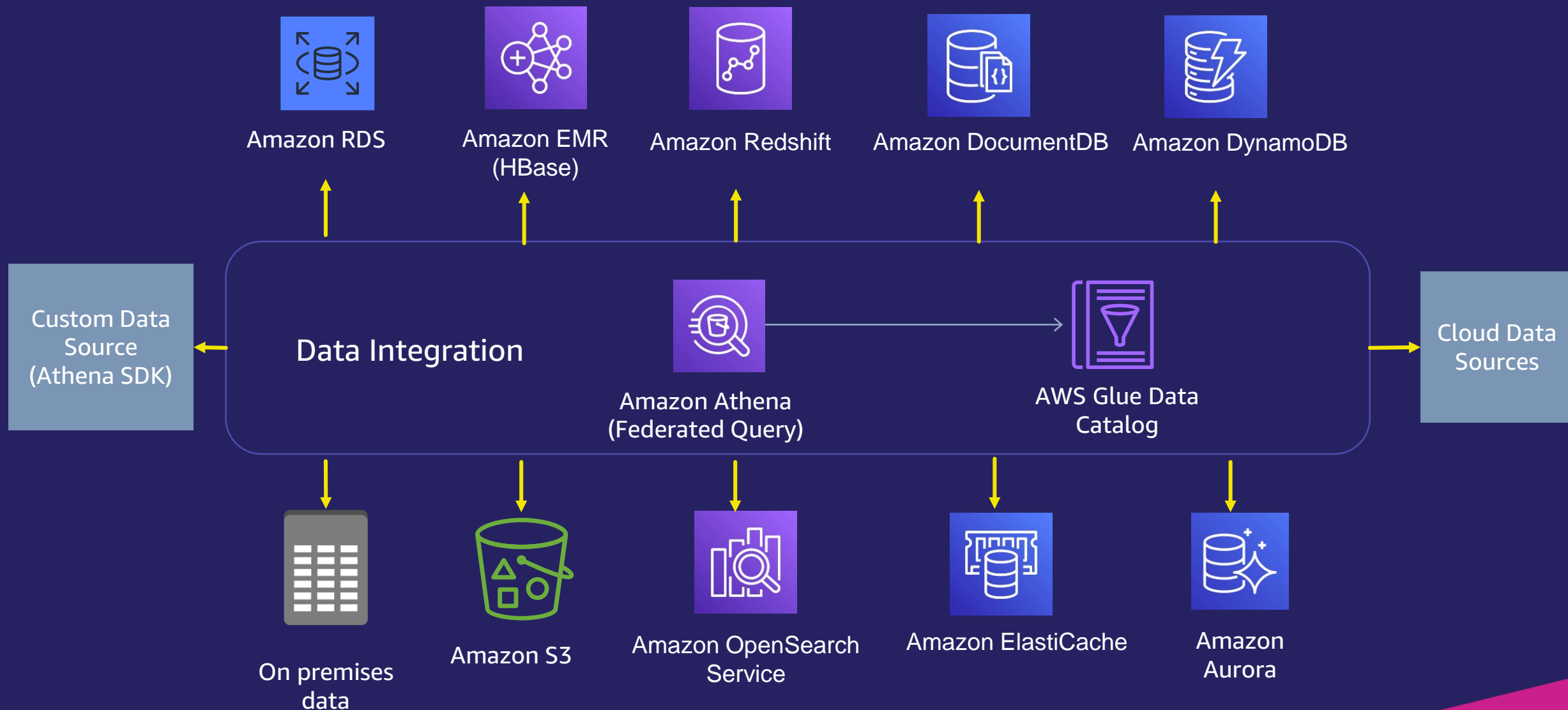


ETL and query

Athena CTAS and INSERT INTO
to ETL



Federated Query with different sources



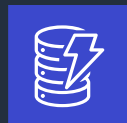
Operationalize ML Models with Athena

Orders/Customer/Supplier



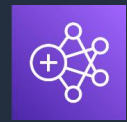
Amazon RDS

Parts/PartsSupplier





Amazon DynamoDB

LinItem




HBase on Amazon EMR

Notebook
Data
Preparation




Amazon Athena
(Federated Query)



Richard - the Engineer

Notebook


Build → Train → Deploy



Amazon SageMaker





Maria - the Scientist



Amazon S3
(model artifacts)

Notebook
Inference



Amazon Athena
(ML Federated
Query)



Ana - the Analyst

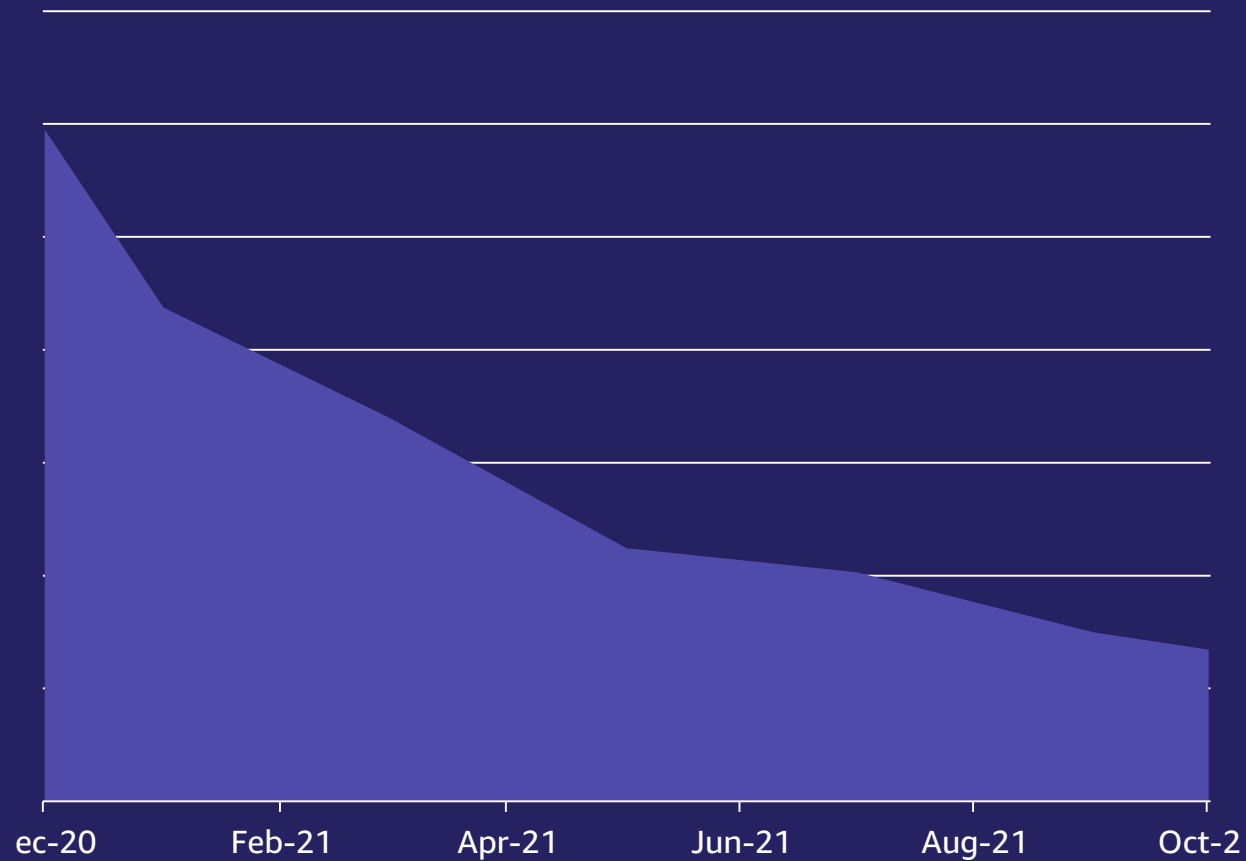


Amazon S3
(inference
results)

Improved scale and performance



3.2x faster performance at the same price



TPC-DS benchmark

Comprehensive areas of improvement

Open Source engine

Dynamic filtering / partition pruning

Hash algorithm for splits

Nested type pruning

TopN rank() functions

Athena Platform

Spill to disk for operators

S3 cost-based optimization

Query plan optimization rules

Drivers & API

Improved input stream processing

Increased output stream throughput

Expanded user experience

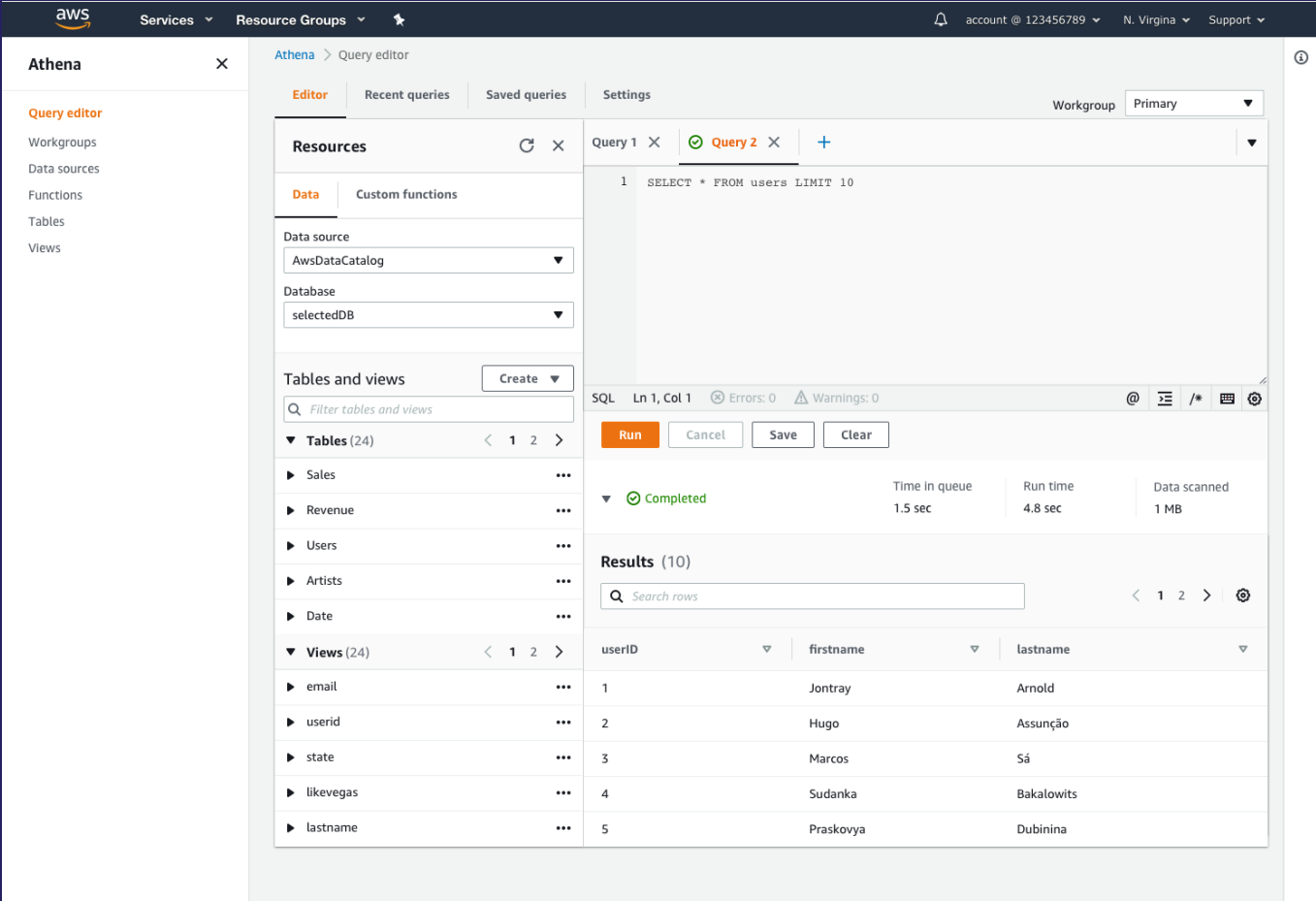


Console upgrade

Refreshed UI framework

Improved workflows

Preferences & personalization



Introducing: Step Functions integration

Jobs

▼ **Create Athena jobs with Step Functions workflows** [Info](#)

AWS Step Functions helps you orchestrate your Amazon Athena queries using serverless workflows, called *state machines*. You can use Step Functions to handle errors, execute multiple queries, and query large datasets.

Send query status
Run a query, and send a notification on success or failure.

[Get started](#)

Execute multiple queries
Execute queries in sequence and parallel, with error handling and notifications.

[Get started](#)

Query large datasets
Use a Glue Crawler to ingest data to S3, then submit a query with emailed status updates.

[Get started](#)

State machines (2)
State machines which orchestrate your Athena queries. [Info](#)

[Refresh](#) [Create state machine](#)

Name	Type	Created
FindFraudCandidatesWorkflow	Standard	2 weeks ago
PullAndProcessDsWorkflow	Standard	1 week ago

Embedded console experience

Use case templates

State machine access

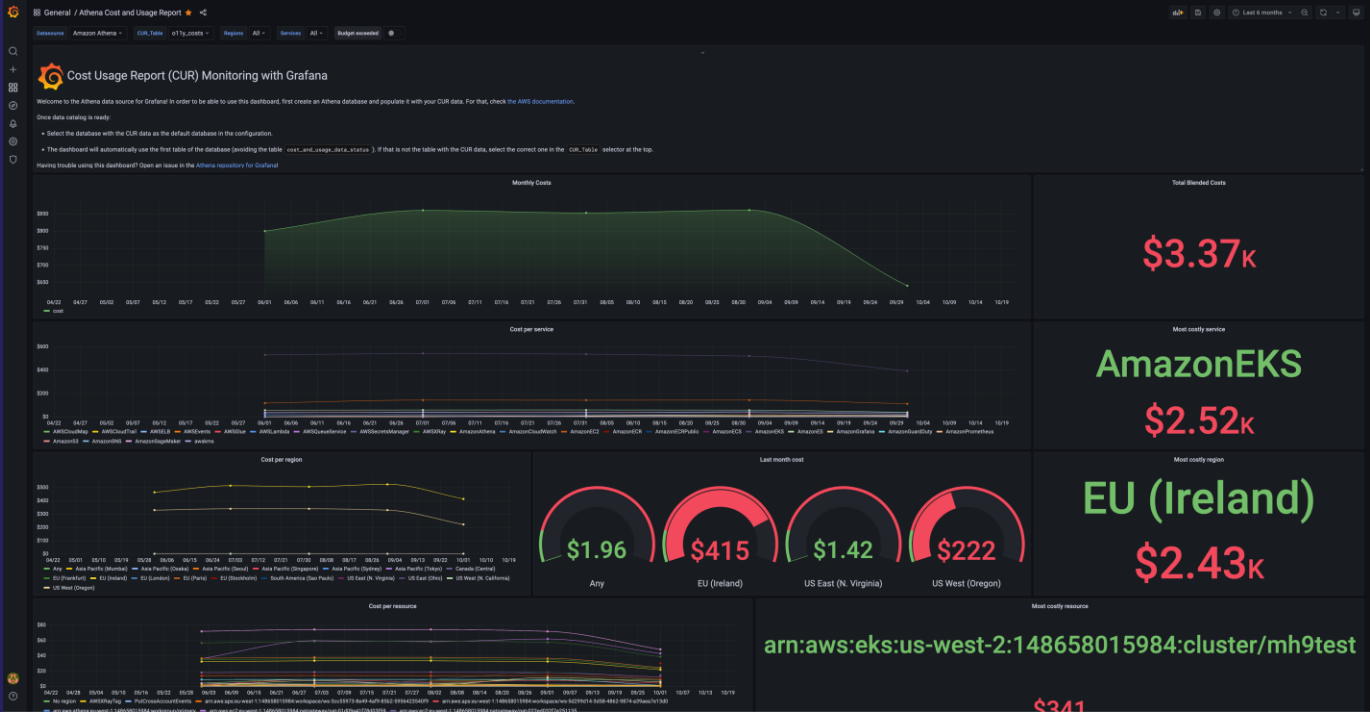
Introducing: Athena Grafana Plug-in

Simple configuration setup

Built-in dashboards

SQL query input

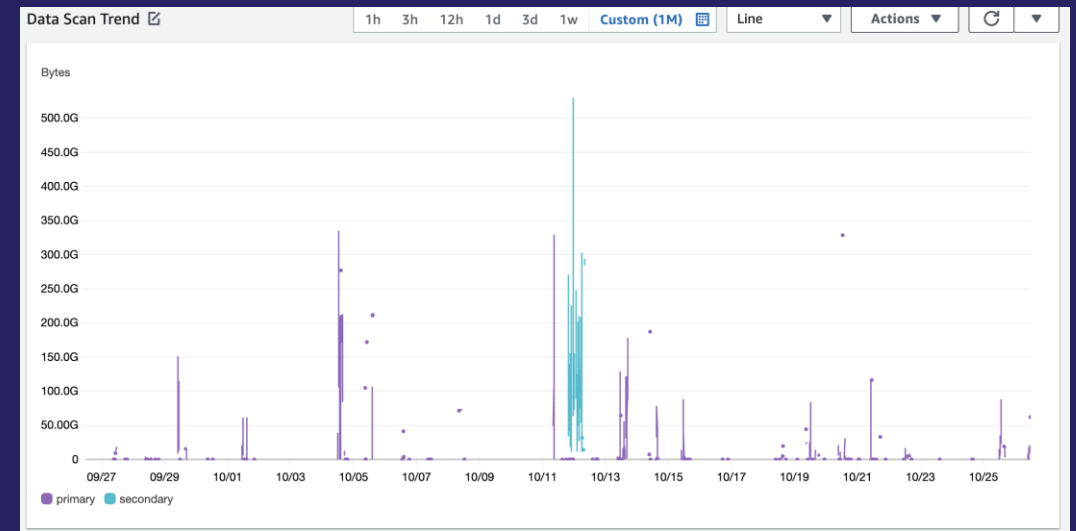
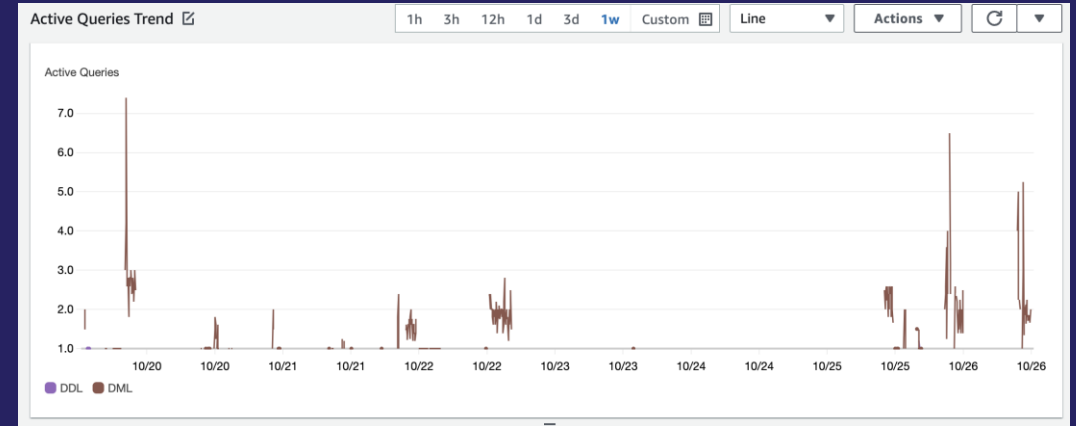
Visualize Athena results



CloudWatch usage metrics

ActiveQueryCount

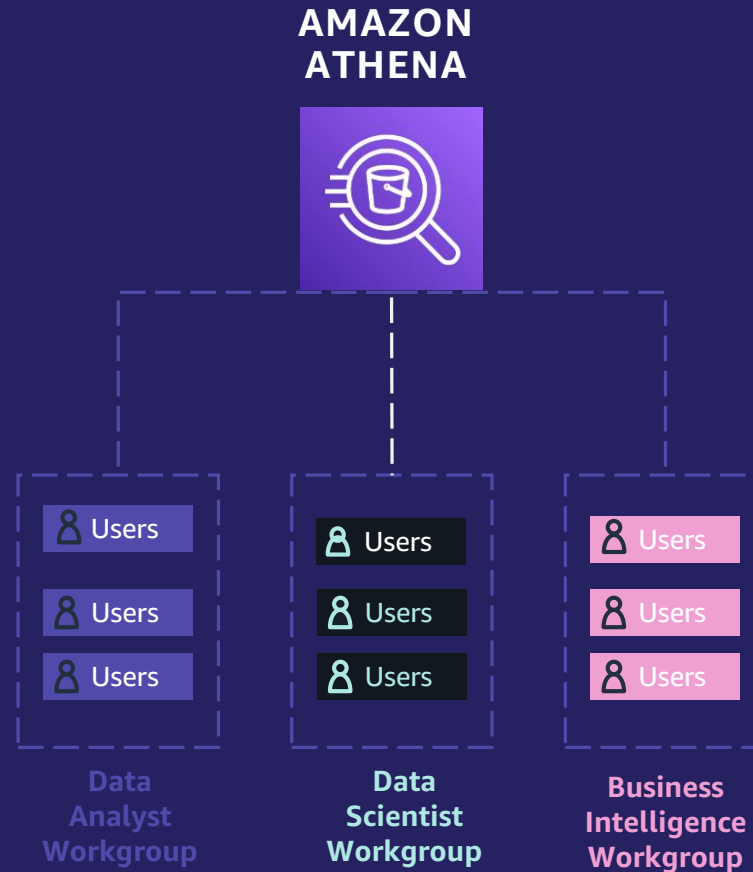
Dimension	Description
Service	The name of the AWS service: Athena
Type	The type of entity: Resource
Resource	The type of resource: ActiveQueryCount
Class	The class of resource: DML or DDL



Multitenant architecture and Cost controls

Workgroups and Cost controls

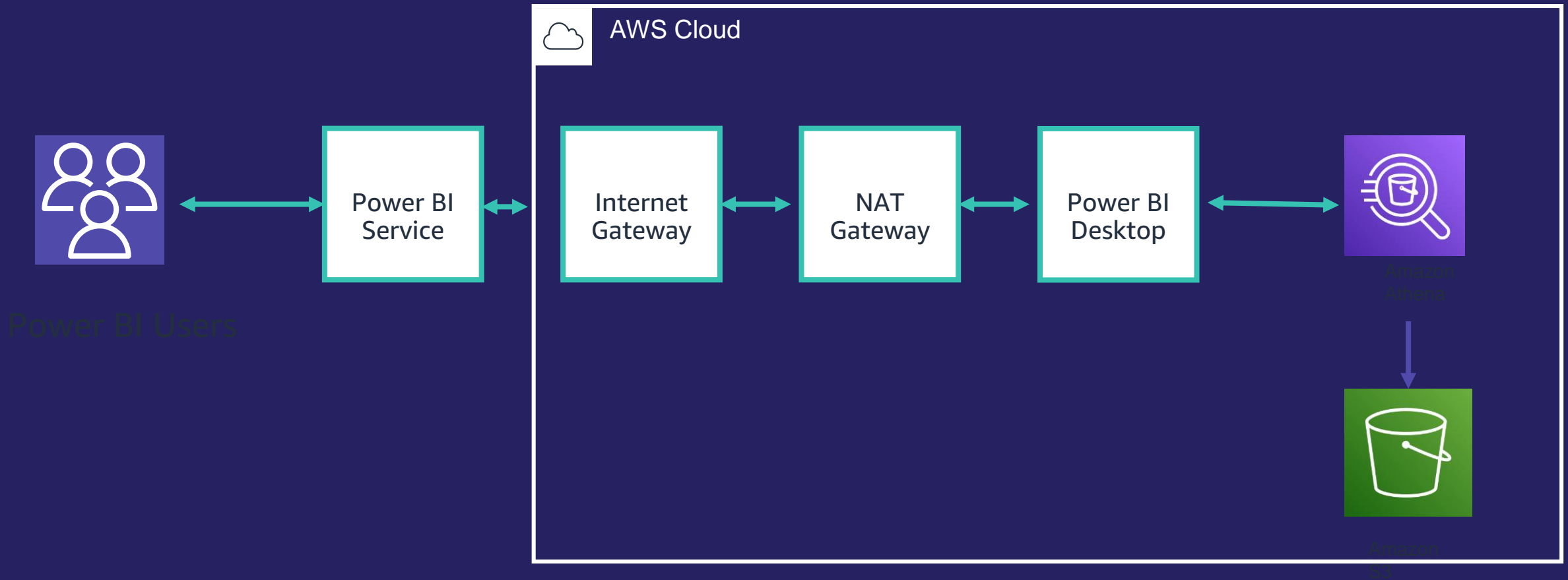
- Isolate users, teams, applications or workloads within an account using workgroups.
- Establish IAM based policies controls who can run queries within a workgroup.
- Workgroups allows you to set data usage controls or thresholds at a query and workgroup level.
- Athena publishes workgroup query metrics to Cloudwatch. In Cloud watch, you can create custom dashboards, and set thresholds and alarms on these metrics.



Amazon Athena Power BI Connector



Power BI Connector



Simplifying data operations



Common DataOps Needs

Timeliness – fresh data is available promptly

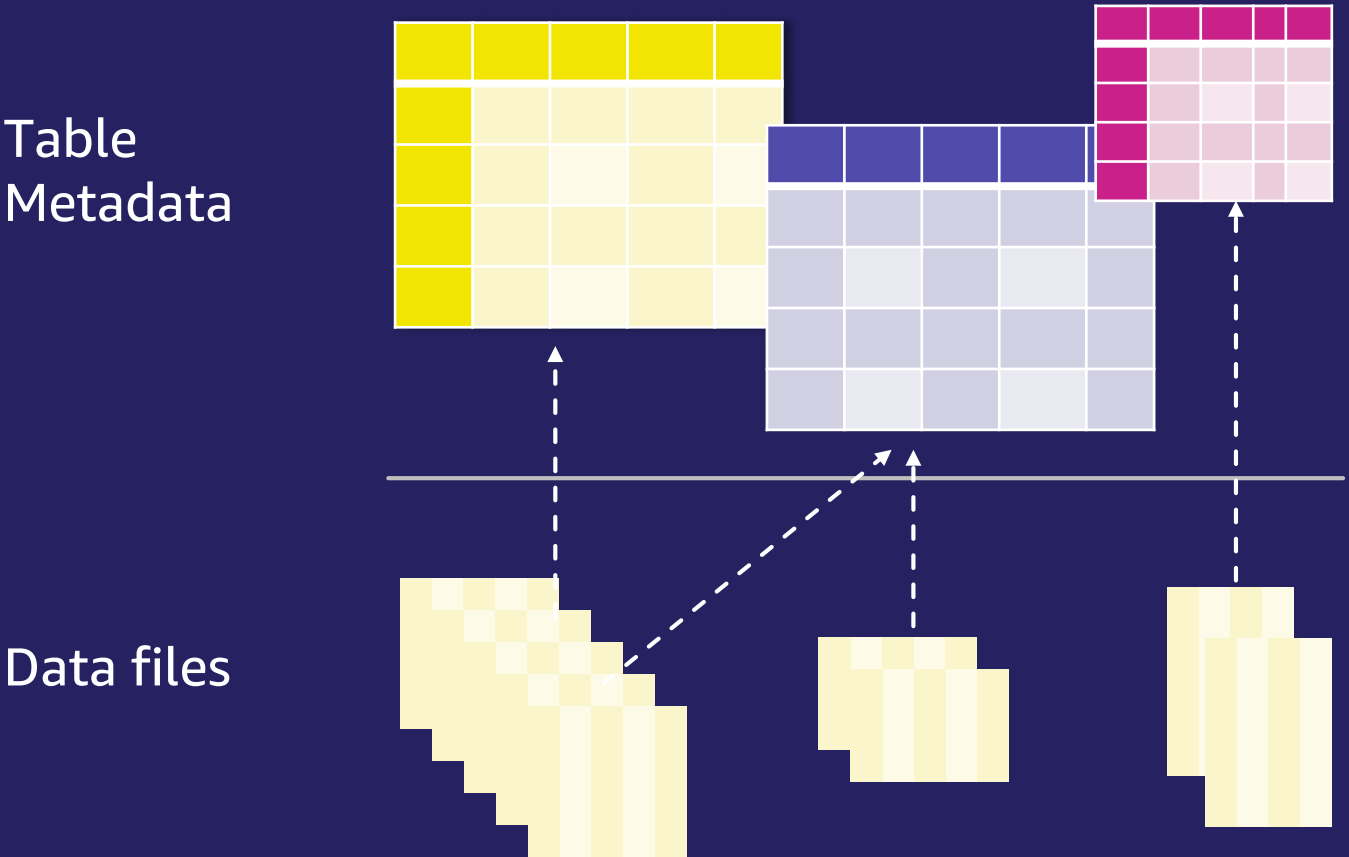
Evolution – data fields and types are free to evolve

Reliability – queries are resilient to updates and deletes

Security – confidentiality is balanced against access

Table-based solutions

Data lake table formats



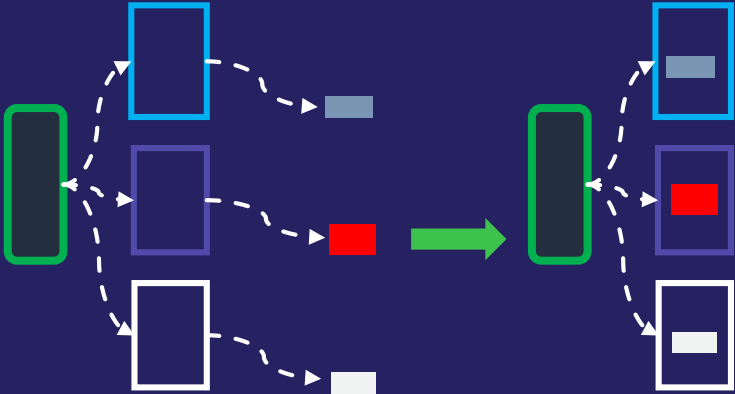
Lake Formation Governed Tables



ACID transactions

Consistent across tasks
Insert, update, delete
Converge batch & real-time

Reliable



Storage optimization

Auto-compact small files
Push-down filters
Reduce data scan

Performant

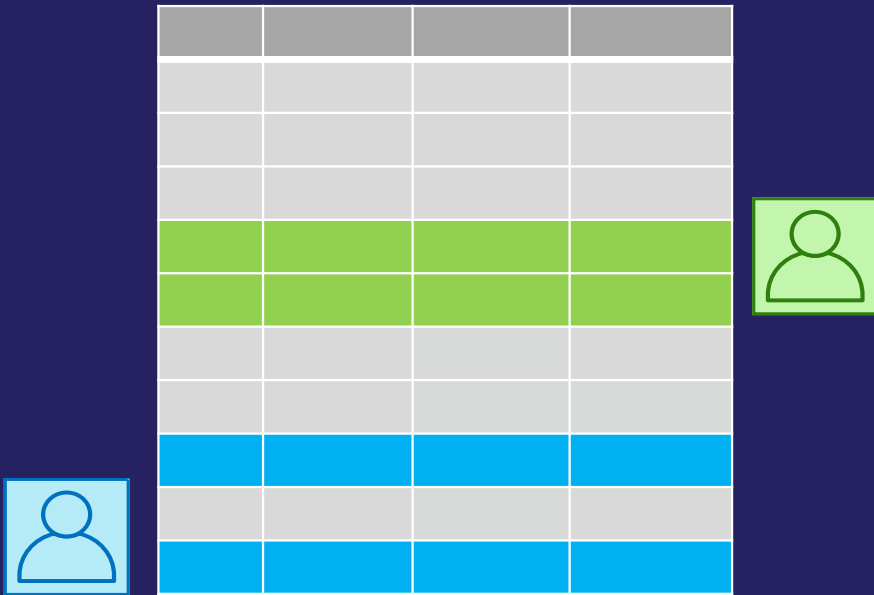


Time travel

Data history
Reproduce experiments
Audit changed data

Versioned

Lake Formation Fine-grained security



Table, column, row and cell-level security policies

Row filter expressions are
"WHERE" clause in "PartiQL"

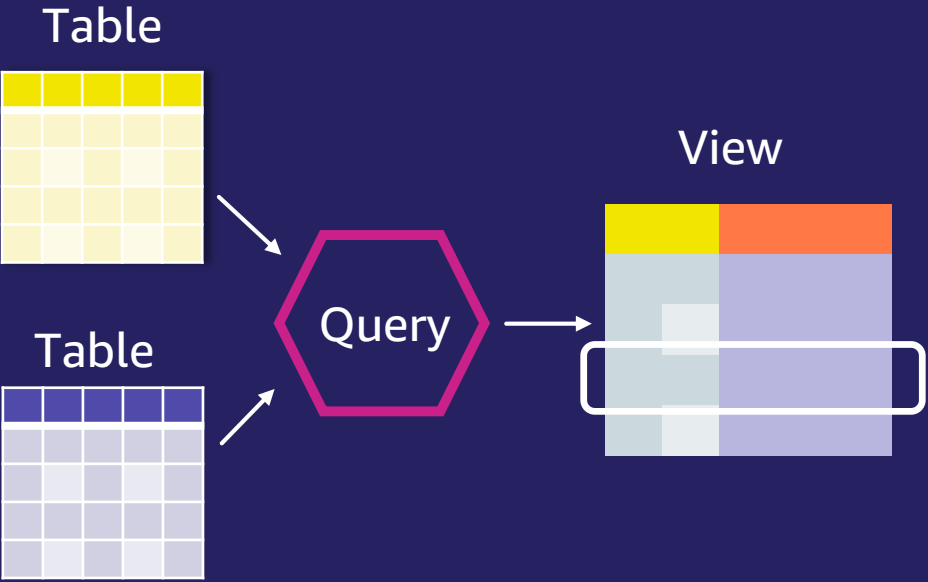
Governed and non-governed table support

Easy-to-audit permissions and access

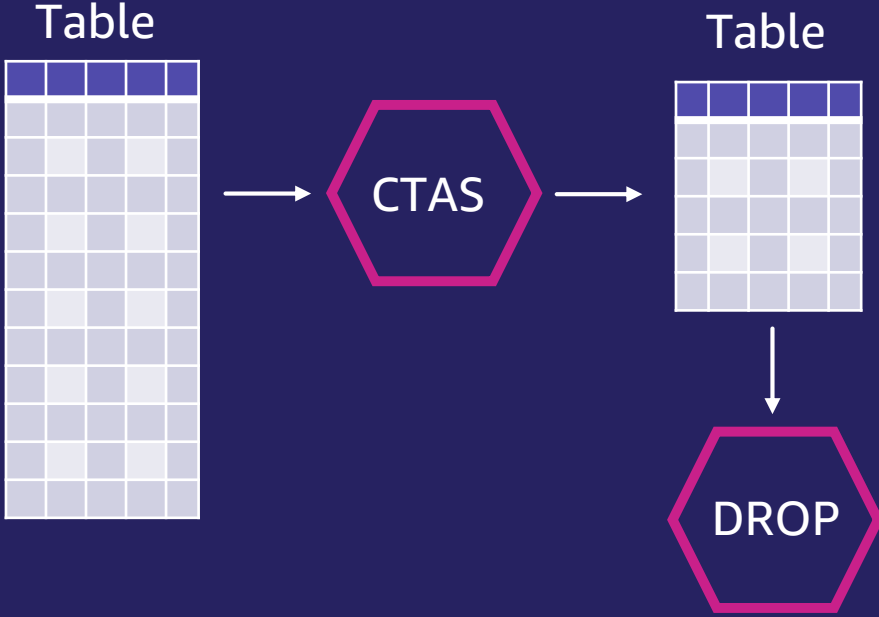
Hive: Improved compatibility

Option 1
(recommended)

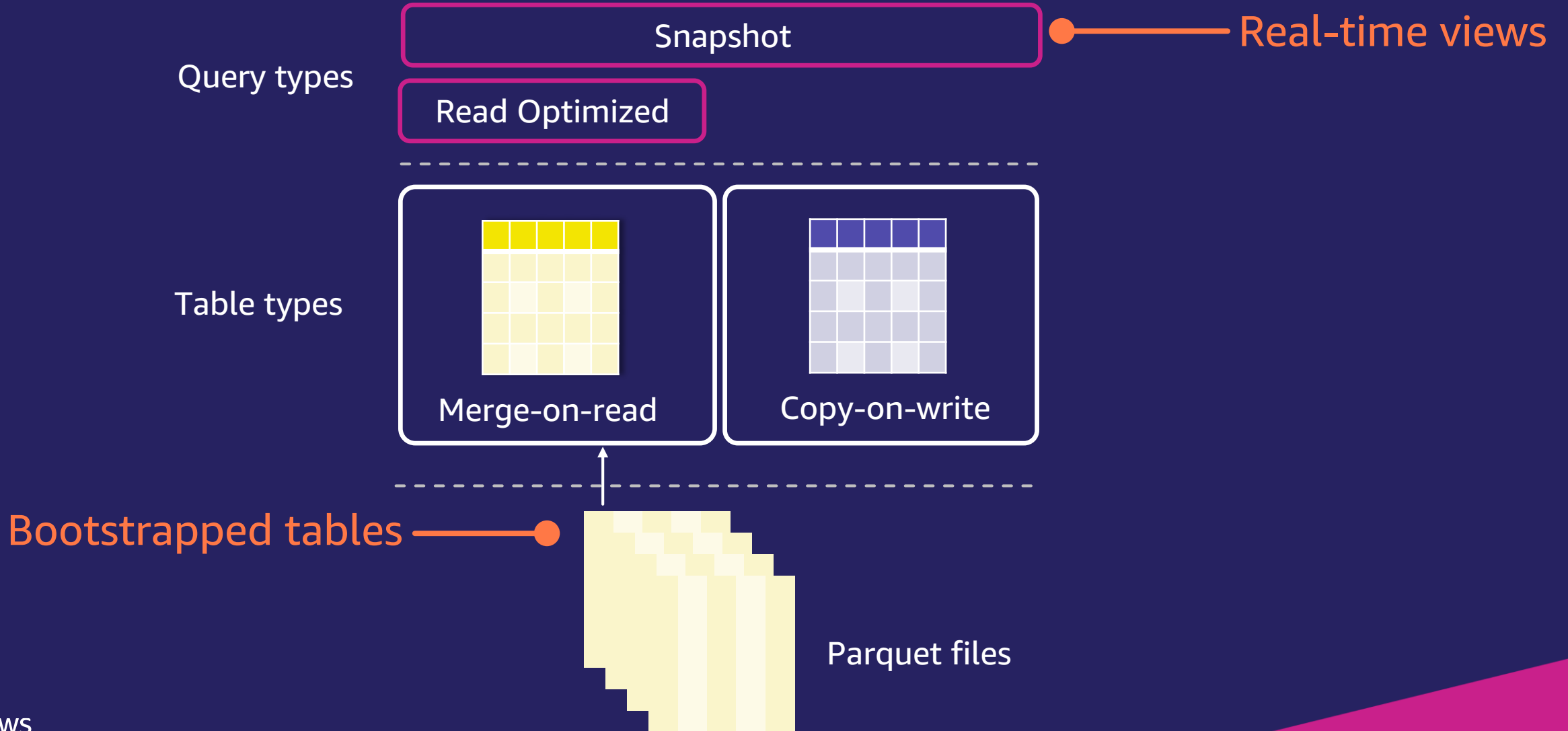
Hive Views



CREATE TABLE AS SELECT



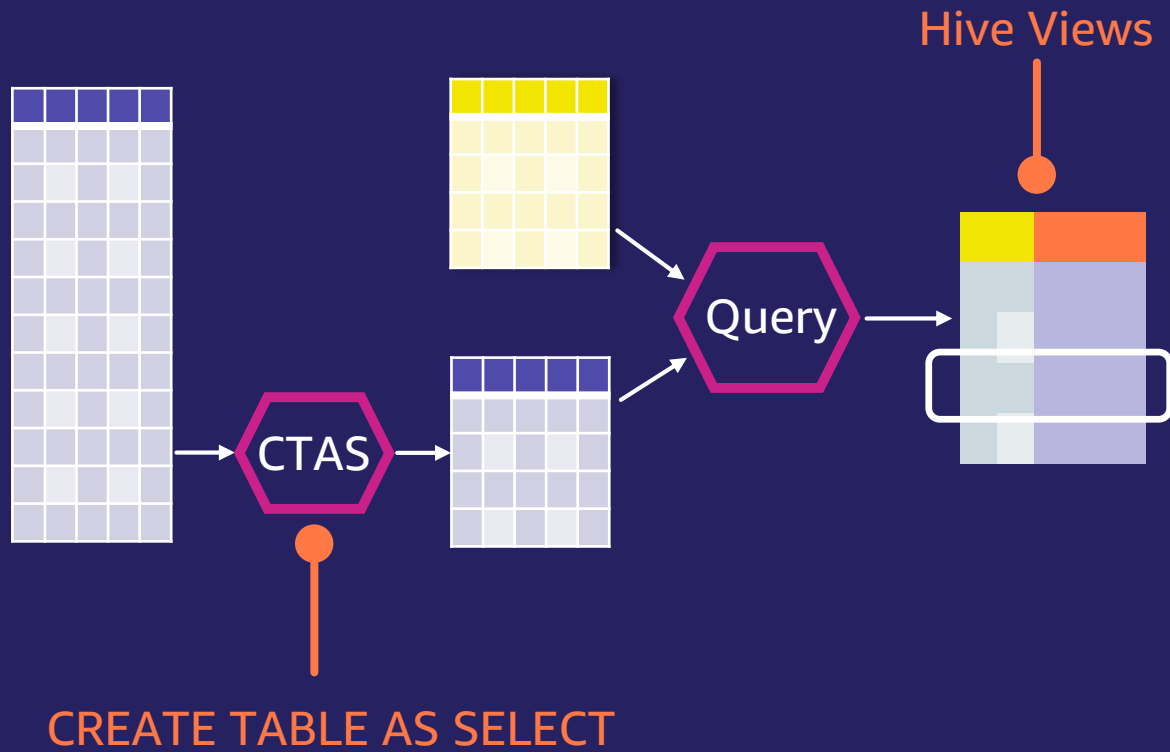
Hudi: Improved compatibility



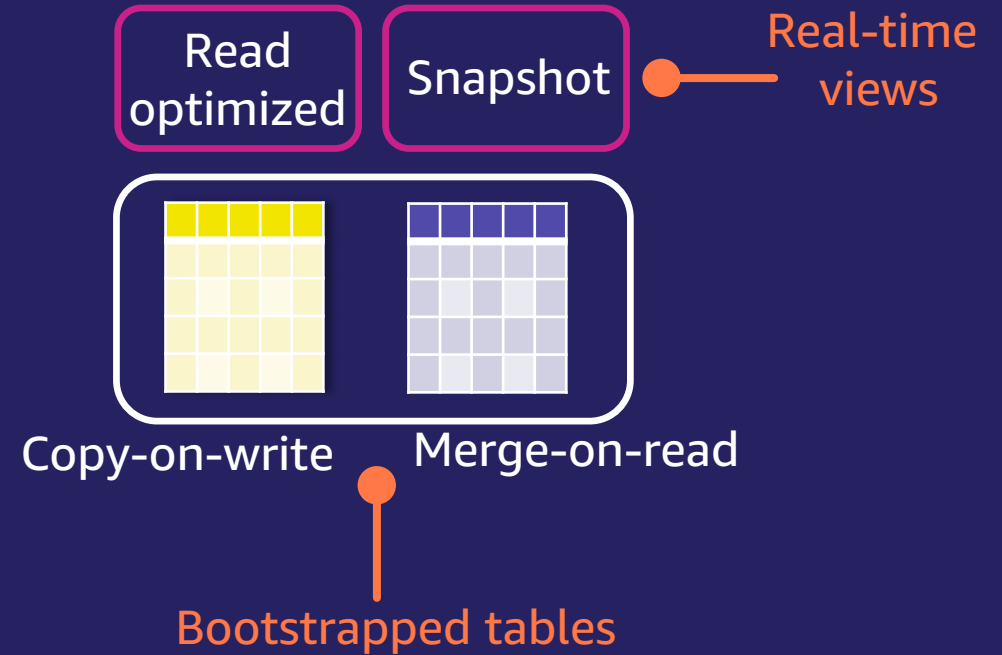
Hive and Hudi: Improved compatibility

Option 2
(not recommended)

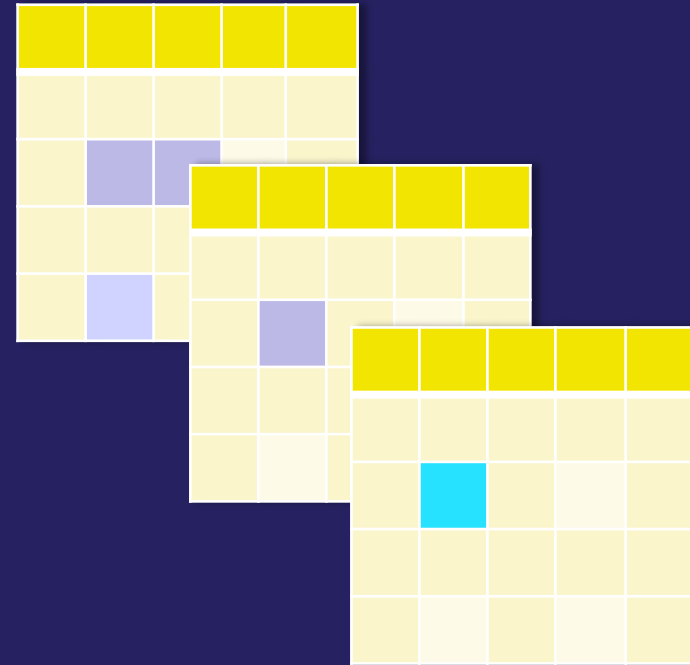
Hive Metastore



Hudi Tables



Announcing: Apache *Iceberg* Integration (Preview)



Open specification table format for huge analytic datasets

Companies writing about Iceberg

Netflix

“Iceberg is widely adopted in Netflix as a data warehouse table format that addresses many of the usability and performance problems with Hive tables.”

LinkedIn

“Gobblin now supports the ORC file format and Iceberg metadata catalog to provide our data customers with high-query performance.”

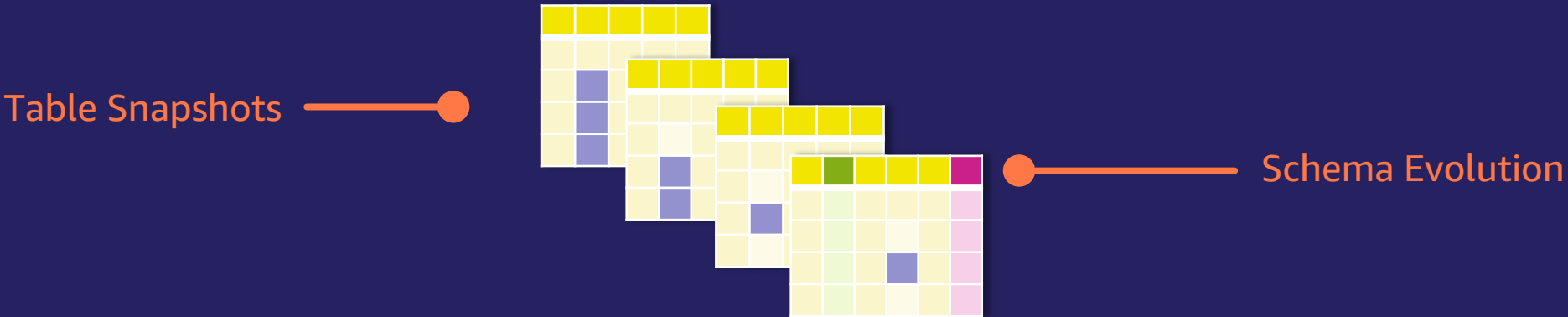
Adobe

“A design point we deeply appreciated with Iceberg was that it’s simply a Java library that could easily be embedded inside of our current applications.”

Expedia

“Iceberg has excellent, inbuilt support for schema evolution that provides guarantees against committing breaking changes to your table.”

Apache Iceberg table features

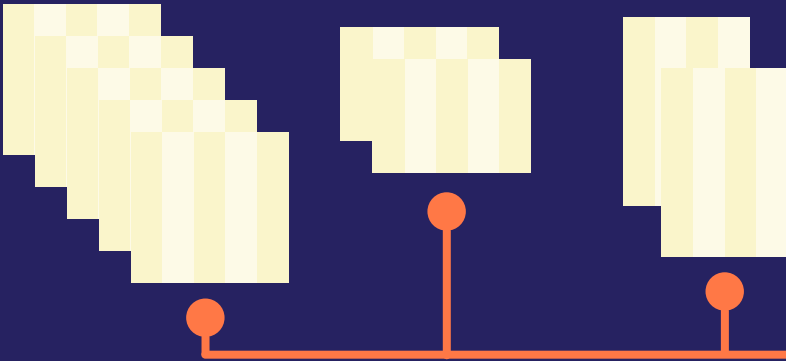


Metadata layer

Data layer

"hidden partitioning"

Partition Evolution



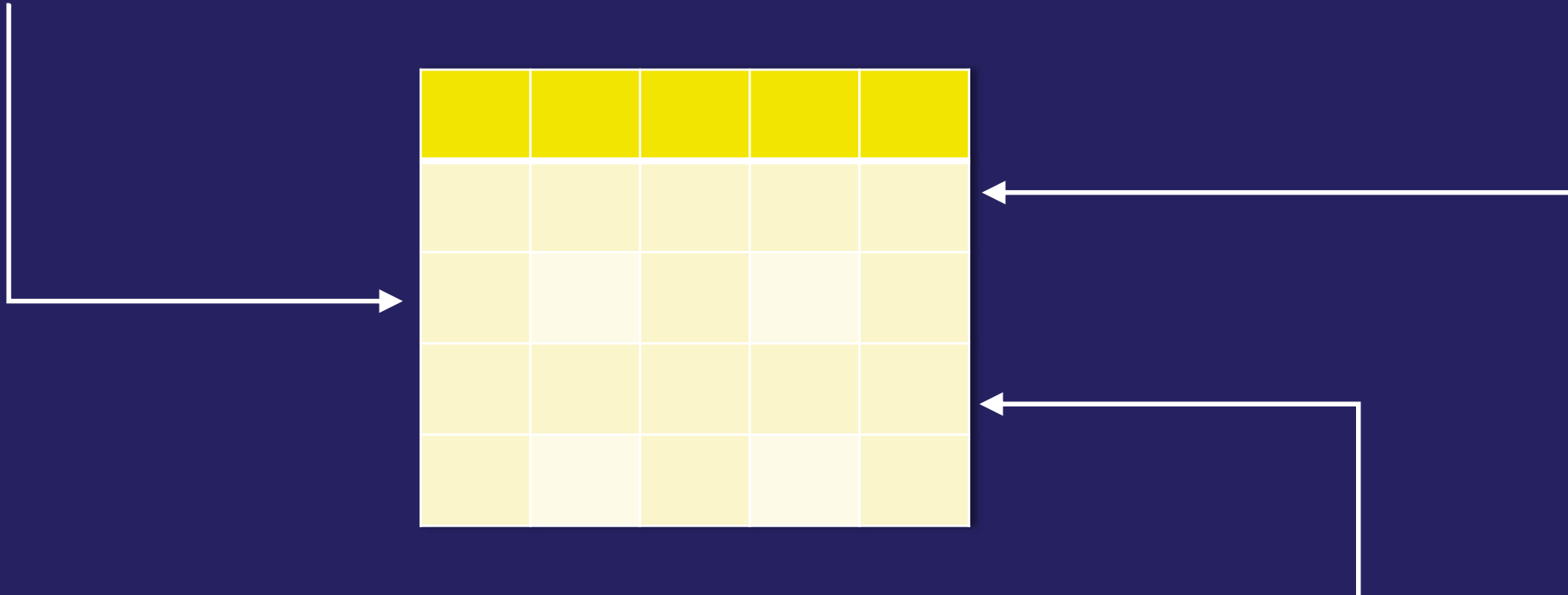
Multiple file formats
(Parquet, ORC, Avro)



ACID Transactions: row-level *Write, Delete, Update*

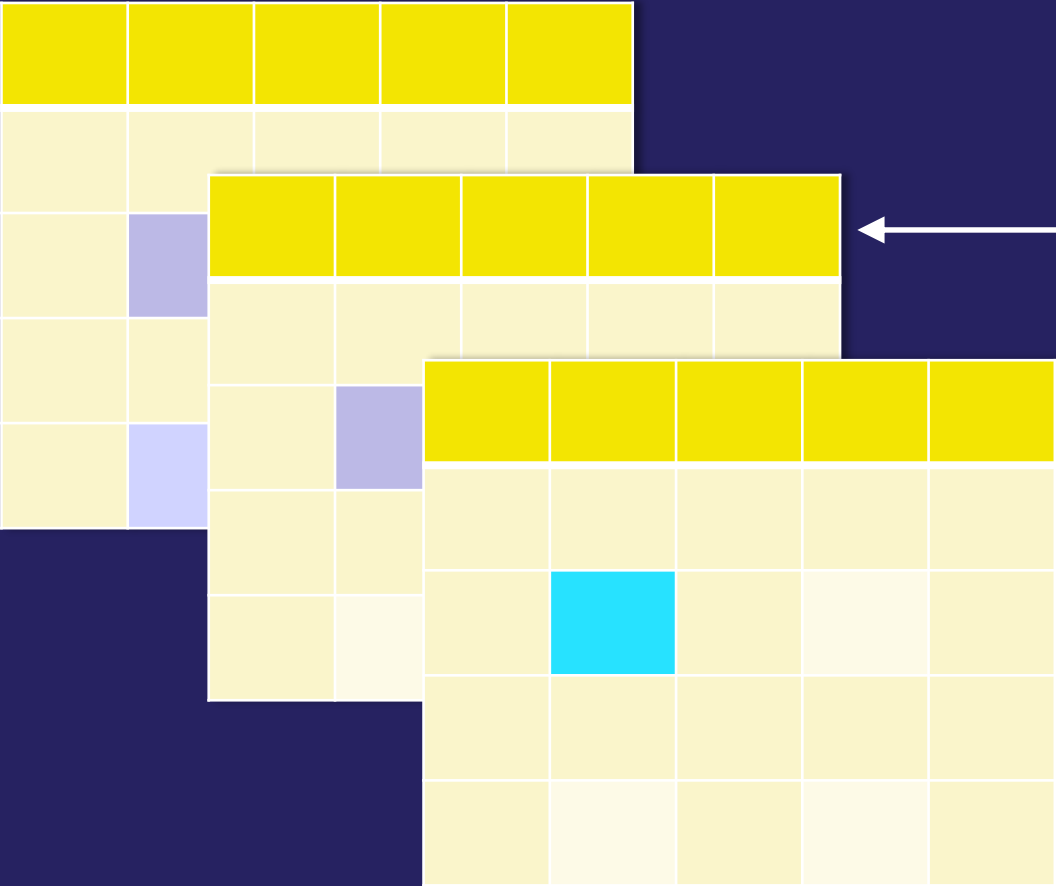
INSERT INTO *tablename* VALUES (v1, v2,...)...

UPDATE *tablename* SET xx=yy[,...] [WHERE *predicate*]



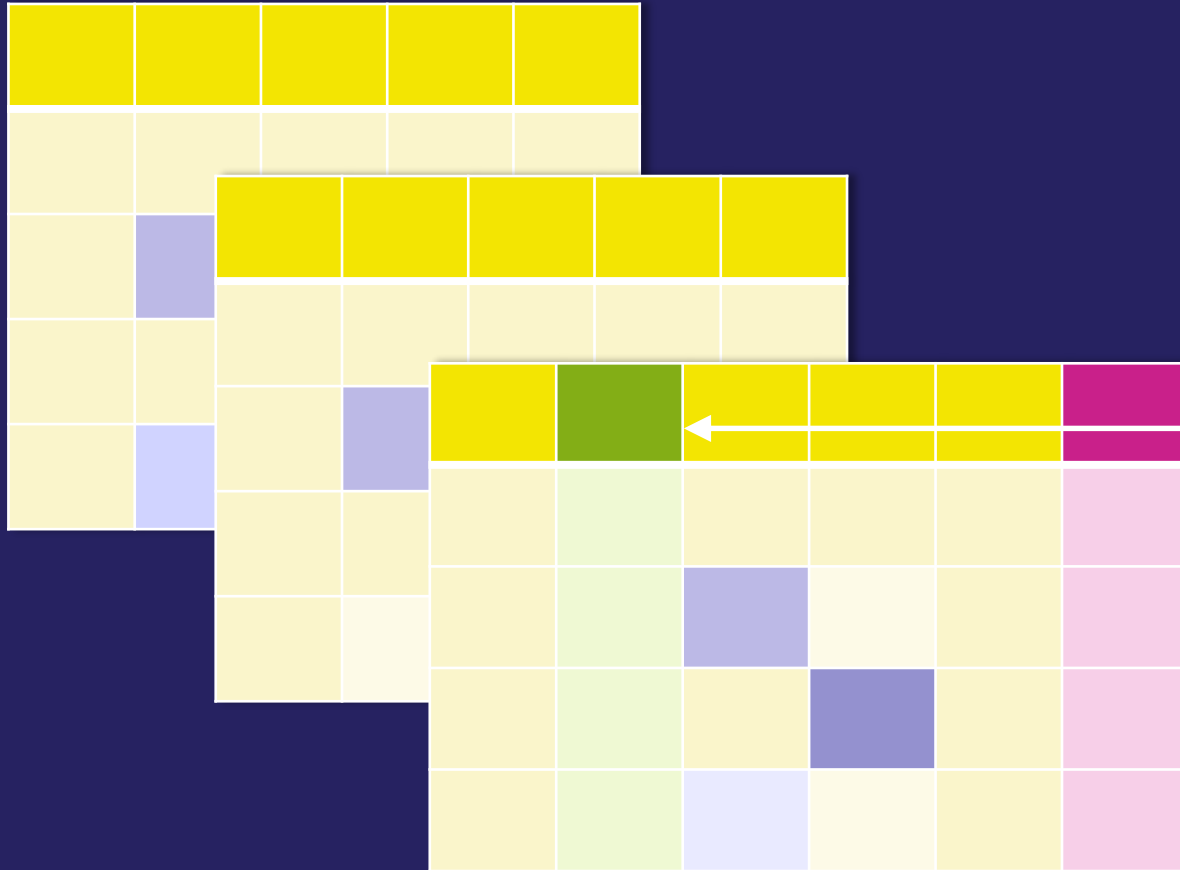
DELETE FROM *tablename* [WHERE *predicate*]

ACID Transactions: *version and time travel* queries



```
SELECT * FROM database.table  
FOR [SYSTEM_TIME | SYSTEM_VERSION]  
AS OF [timestamp | version]
```

ACID Transactions: *schema evolution* queries

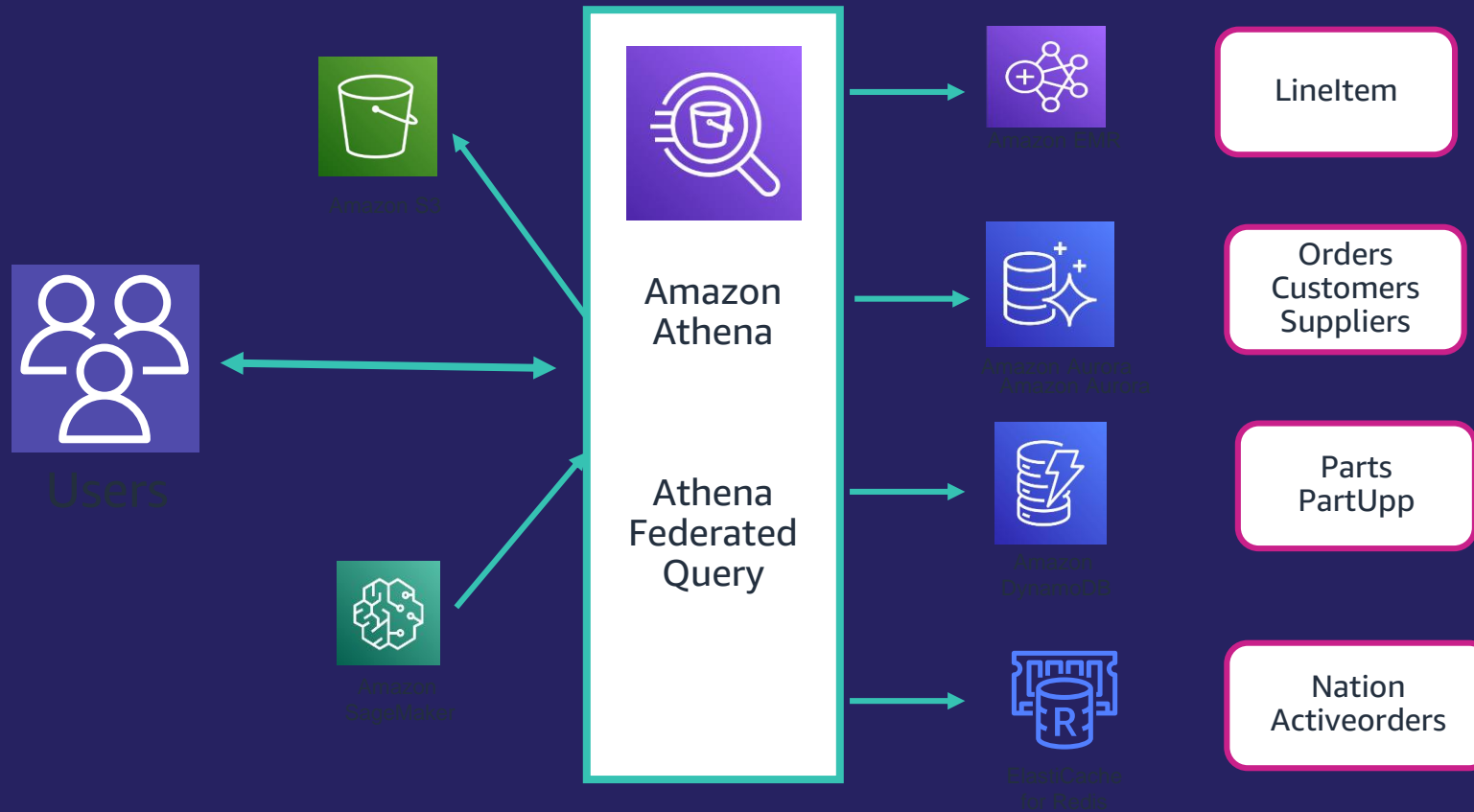


`ALTER TABLE tablename ADD COLUMNS
[column_name data_type [...]]`

Federated Query Capability



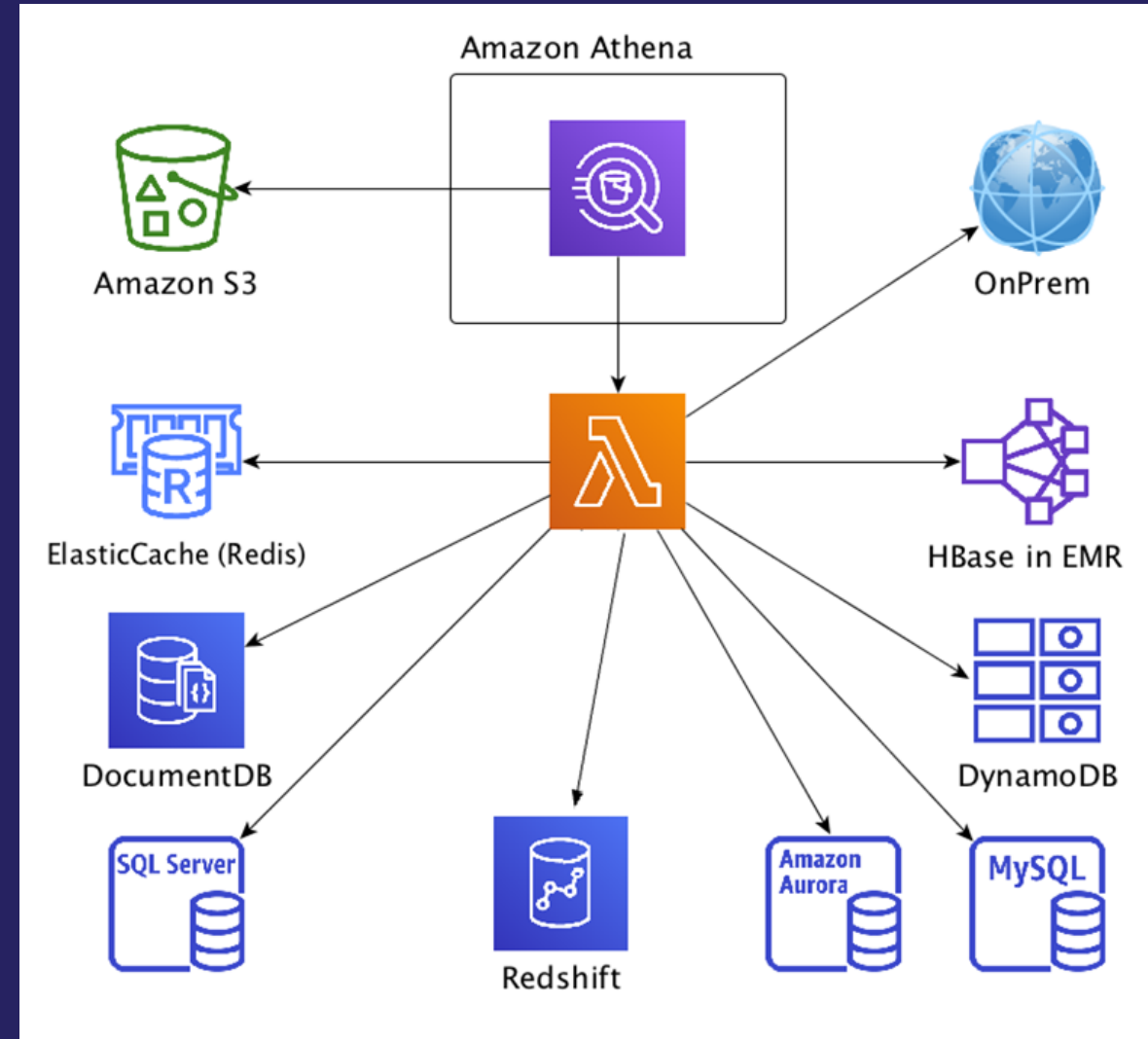
Athena Federated Query



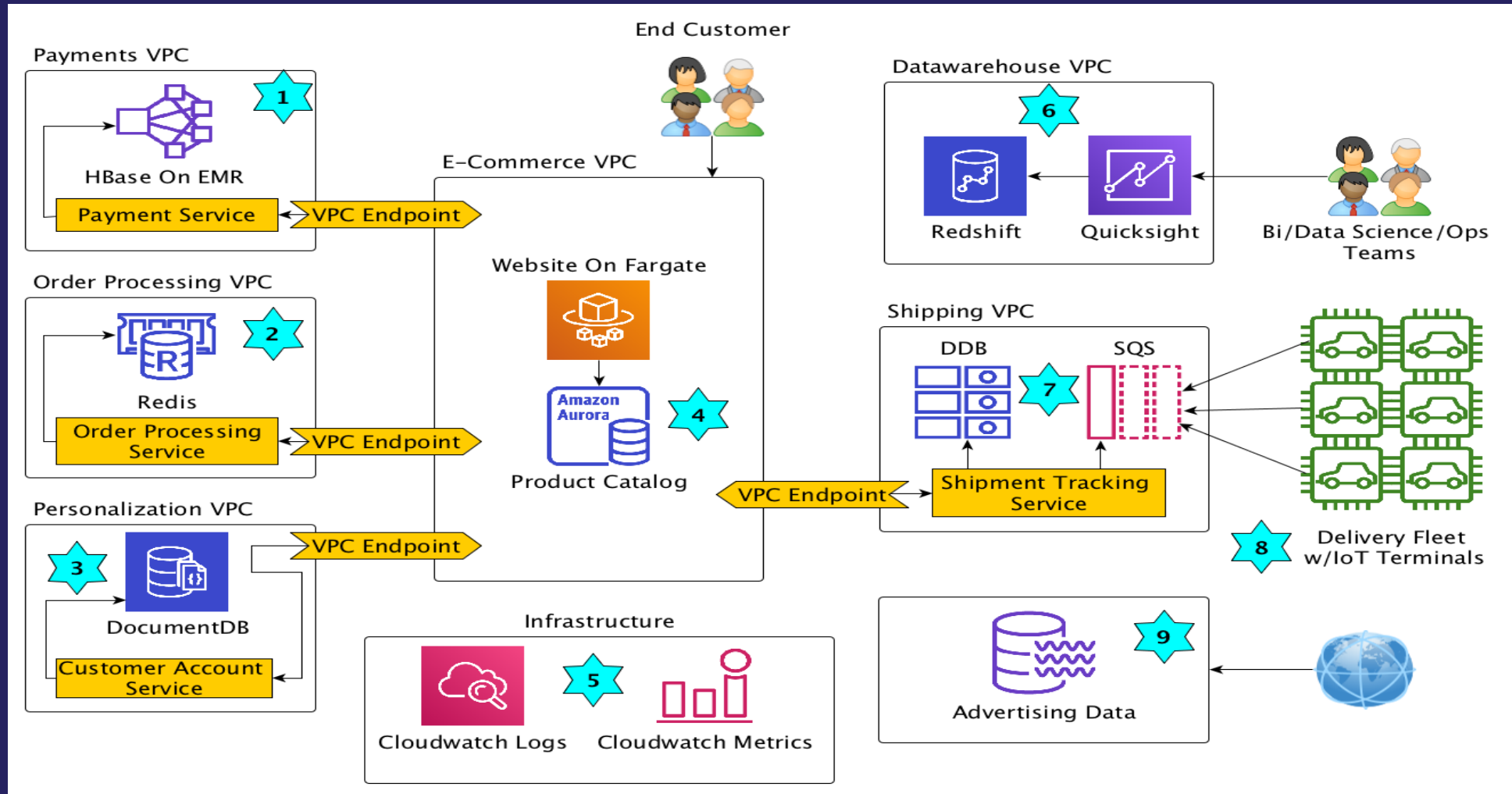
Amazon Athena - Federation

What is federated query?

- Run query across relational, non-relational, object, or custom data sources
- Run query across On-Premises or cloud data sources
- Can be used for ad-hoc investigations, or complex pipelines, or applications



Why do you need federated query?



FINRA operates petabyte-scale analytics on data lakes using Amazon Athena

Challenge

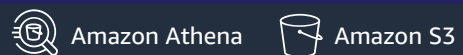
FINRA sought a serverless technology solution for cost-effectively handling data at petabyte scale while maintaining high performance and security under a heavy load.

Solution

FINRA used Amazon Athena to meet its required specifications, including compatibility with Apache Hive Metastore, which FINRA uses for its data catalog.

Result

- Scales automatically
- Provides secure internode encryption and Amazon KMS encryption
- Has a 17-second response time to requests
- Manages over 150 PB of data at \$5 per terabyte of data scanned



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

Migration opportunities



SELF-SERVICE GUIDES

Perform migration on your own using step-by-step guides and reach out to AWS experts for assistance



MIGRATION SUPPORT

Get hands-on support each step of the way through your migration to Amazon Athena

Demo

Amazon Athena Federated Query



Thank you!