

AWS
re:Invent

ANT 308

A Deep dive Into running Apache Spark on Amazon EMR

Mert Hocanin

Principal Big Data Architect
Amazon Web Services

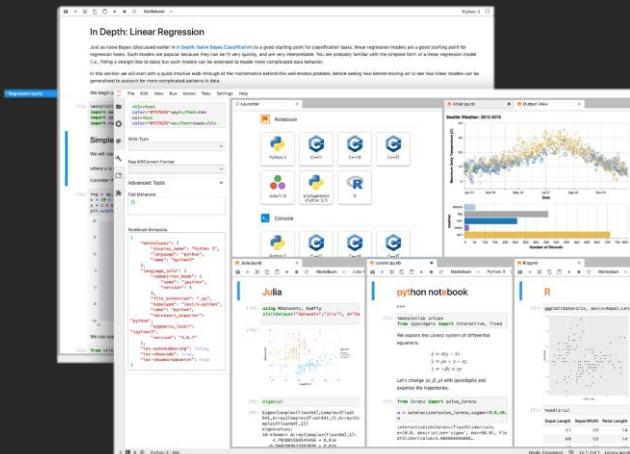
Abhishek Sinha

Principal Product Manager
Amazon Web Services

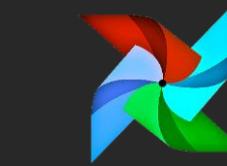
How will we spend our time today

1. Apache Spark Improvements through the year
2. Common architectural patterns
3. Demos

Data Engineering Platforms are evolving



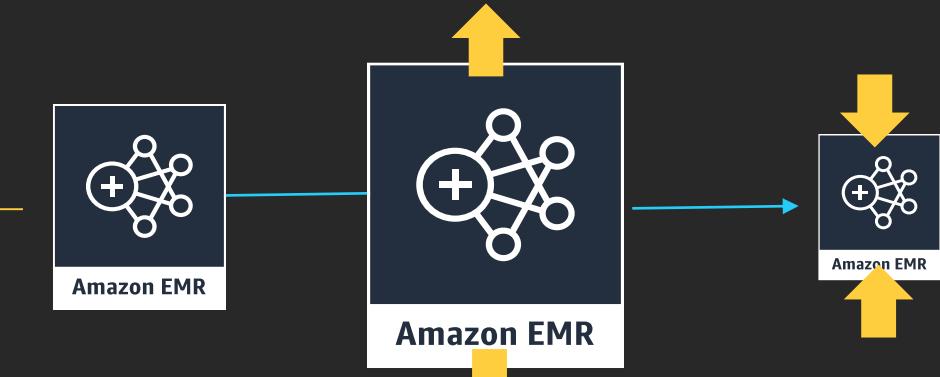
Notebooks



Apache Airflow



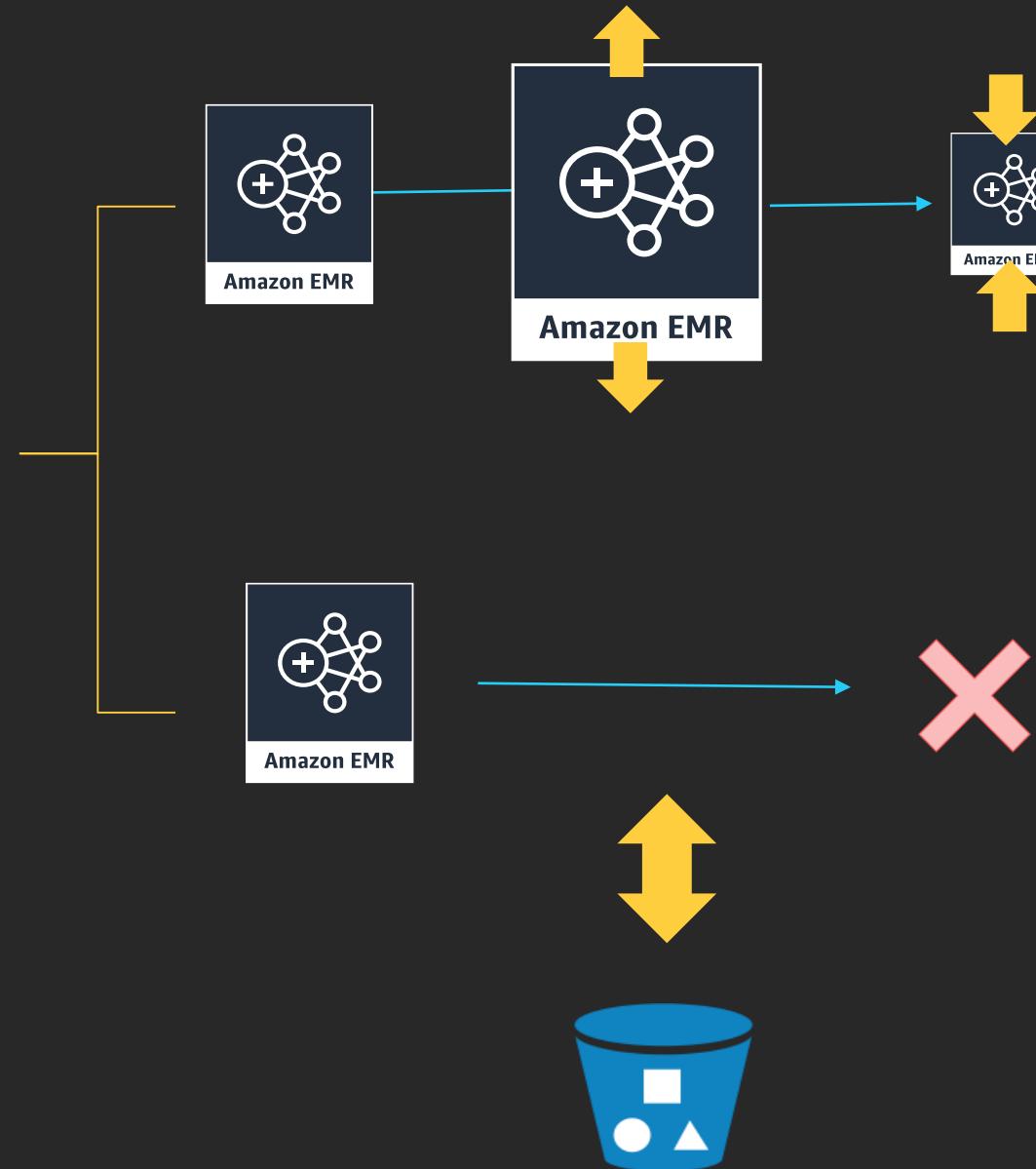
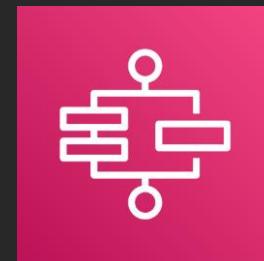
AWS Step Functions



Amazon EMR



Customers have started to adopt cloud-native patterns



Long running & auto scaling cluster

Purpose-built, job-scoped clusters

How do you decide

Long-running and auto scaling

1. Great for lines of business clusters
2. Great for short-running jobs
3. Ideal to save costs for multi-tenanted data science and data engineering jobs

Transient and job scoped

1. Work well for long-running, job-scoped pipelines
2. Separating production pipelines into job-scoped clusters reduces blast radius

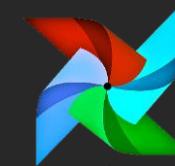
A common architectural pattern for data platforms

Clients



Notebooks

Orchestration

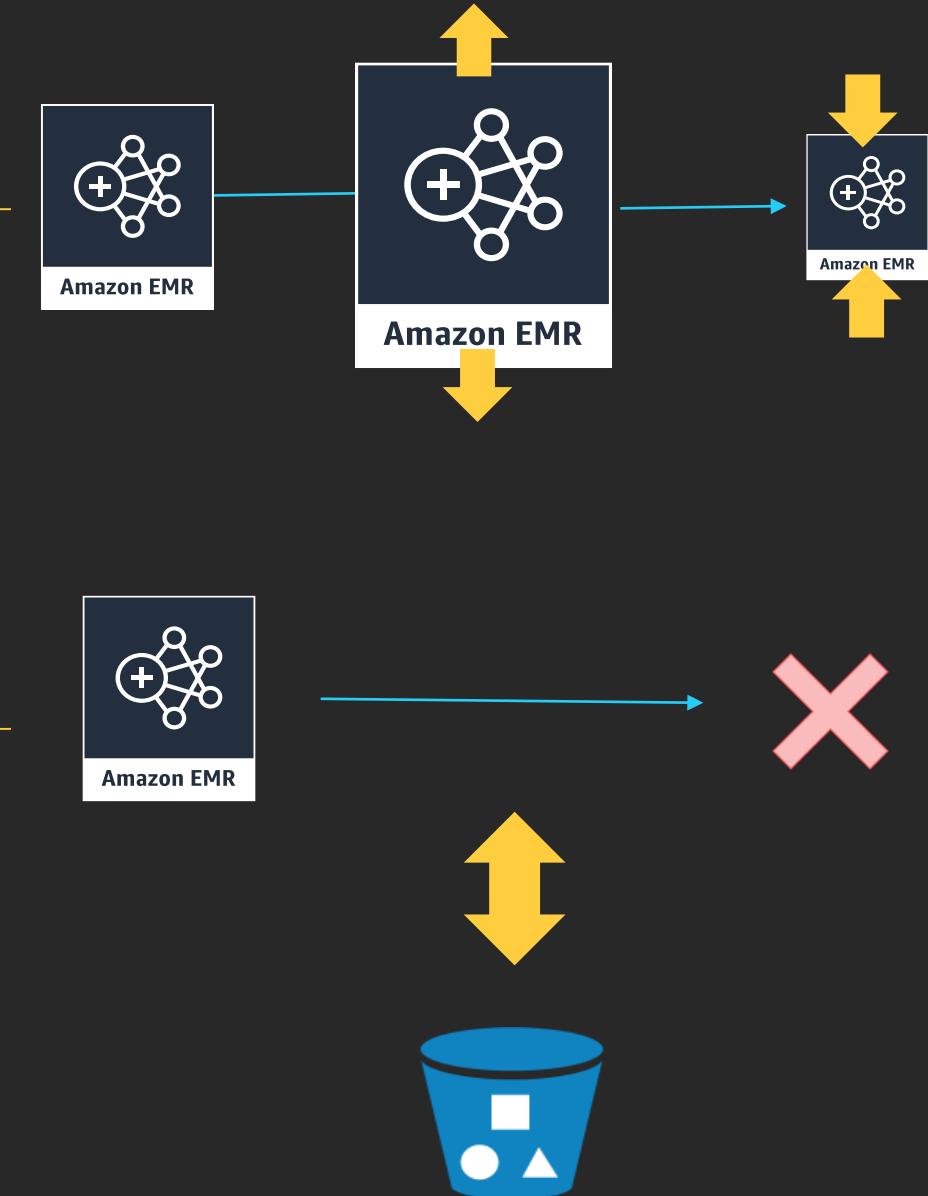


Apache Airflow



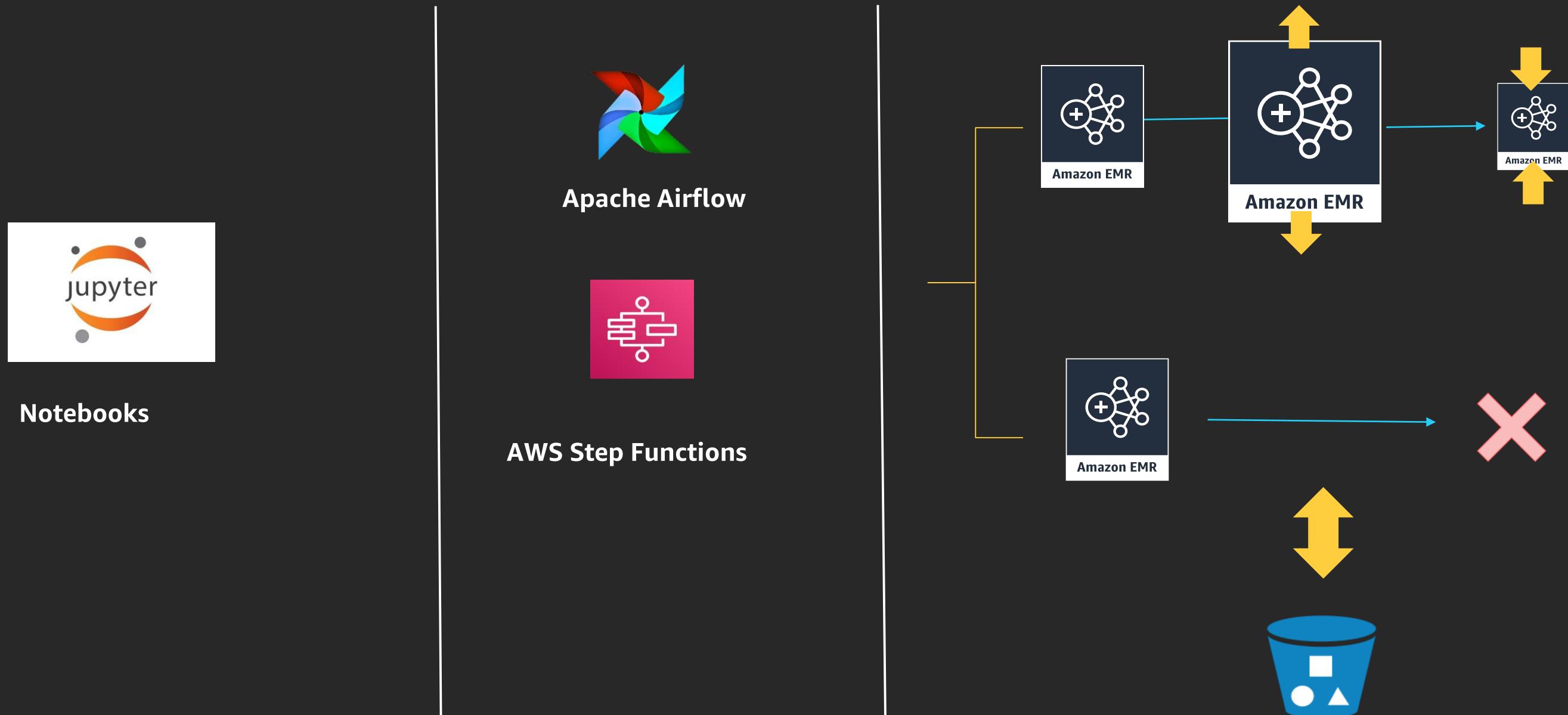
AWS Step Functions

Execution



A common architectural pattern for data platforms

Improvements



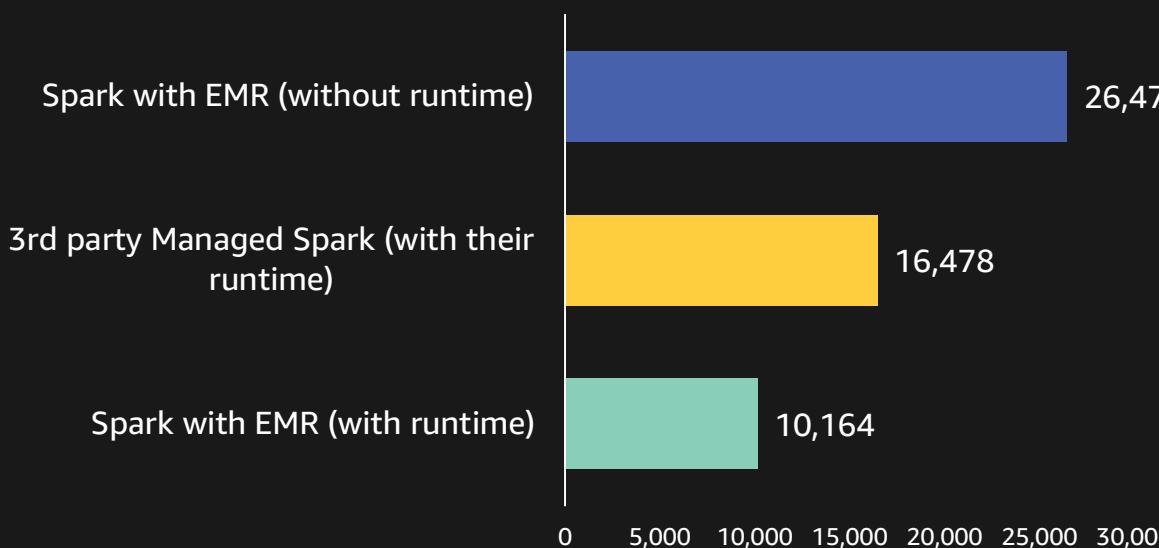
1

EMR Runtime For Apache Spark

Performance improvements in Spark for Amazon EMR NEW!

Performance-optimized runtime for Apache Spark, 2.6x faster performance at 1/10th the cost

Runtime total on 104 queries (seconds - lower is better)



*Based on TPC-DS 3TB Benchmarking running 6 node C4x8 extra large clusters and EMR 5.28, Spark 2.4

Runtime built on a optimized version of Apache Spark

Best performance

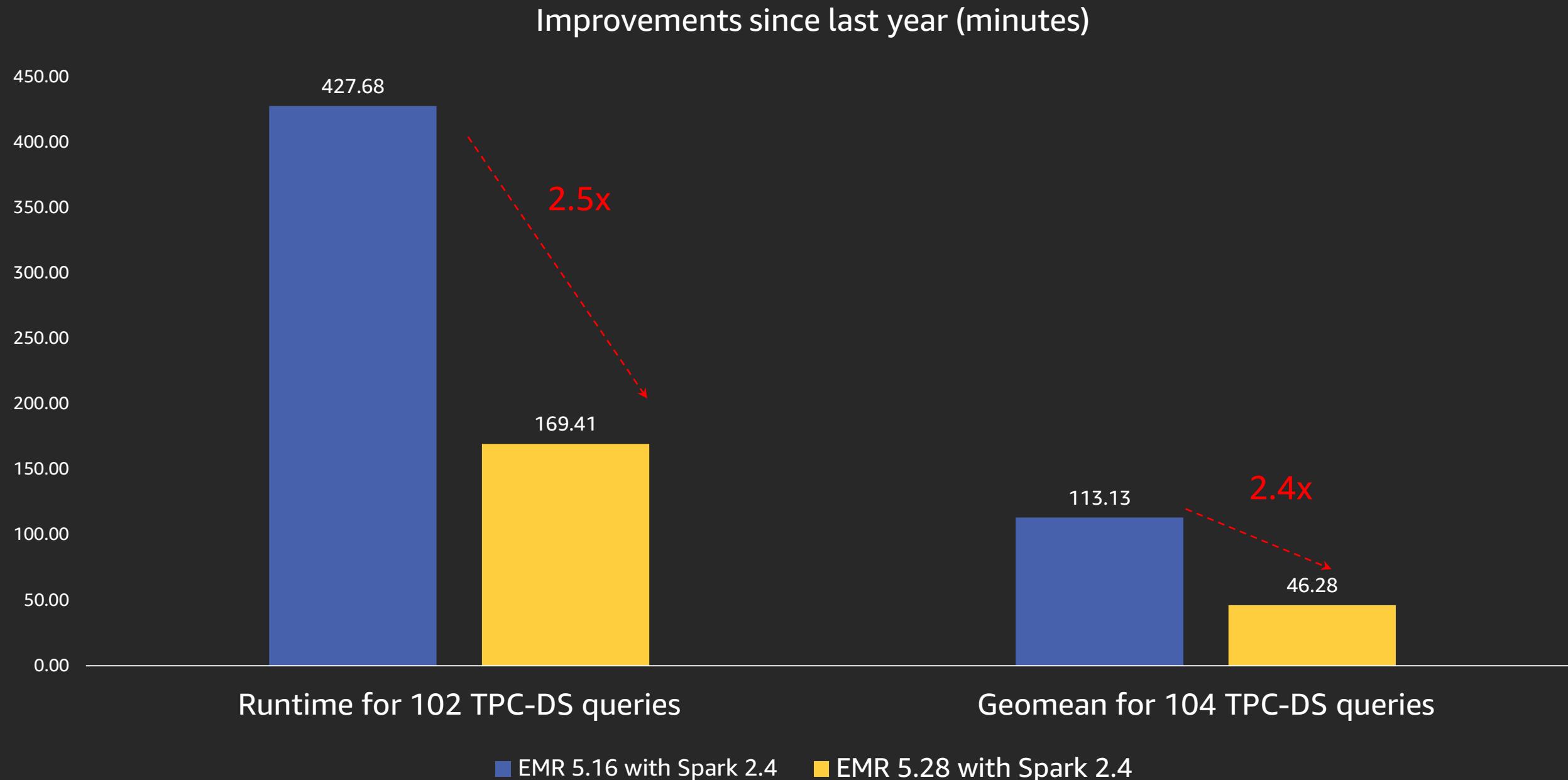
- **2.6x faster** than Spark with EMR without runtime
- **1.6x faster** than 3rd party Managed Spark (with their runtime)

Lowest price

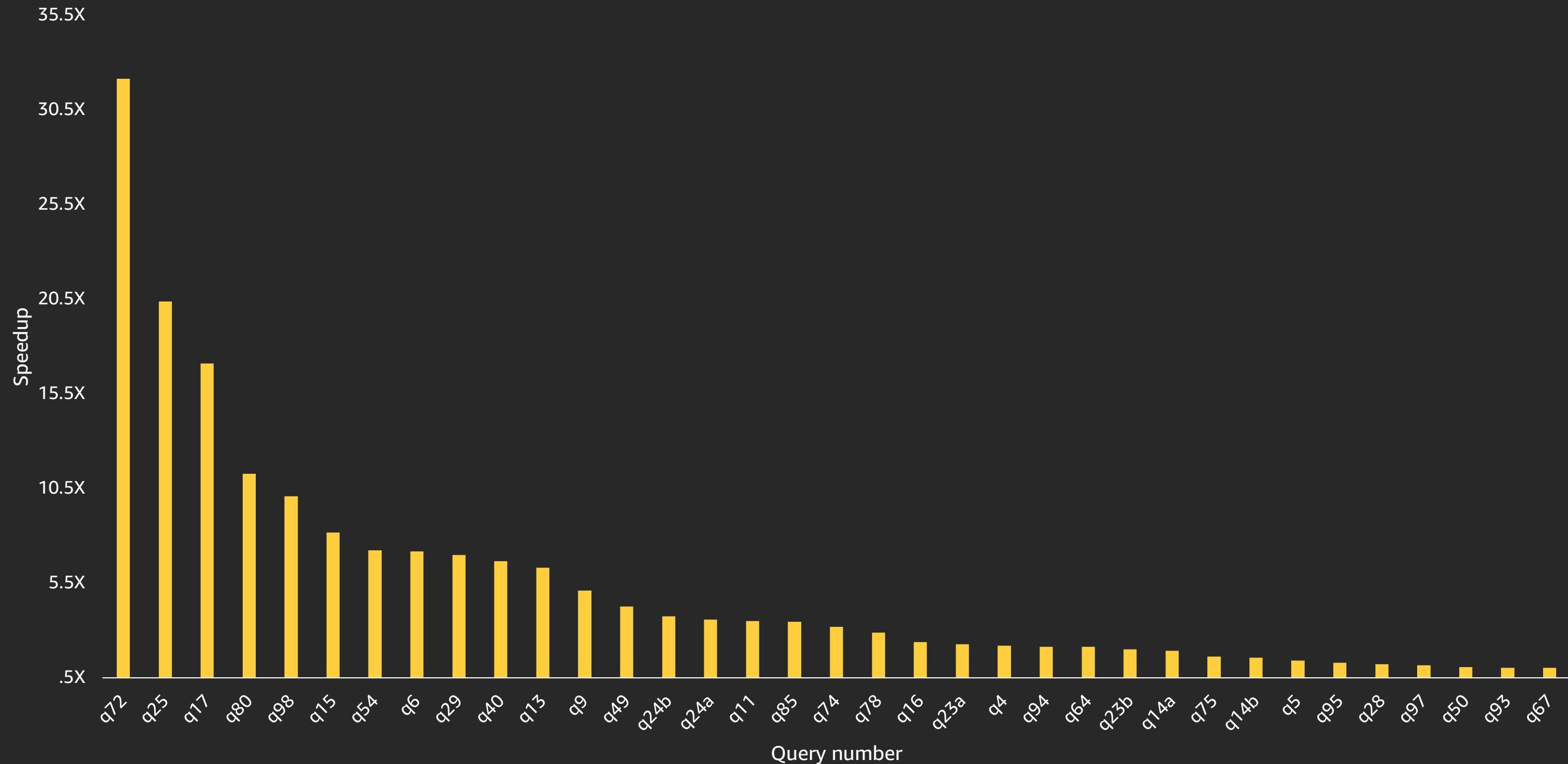
- 1/10th the cost of 3rd party Managed Spark (with their runtime)

100% compliant with Apache Spark API's

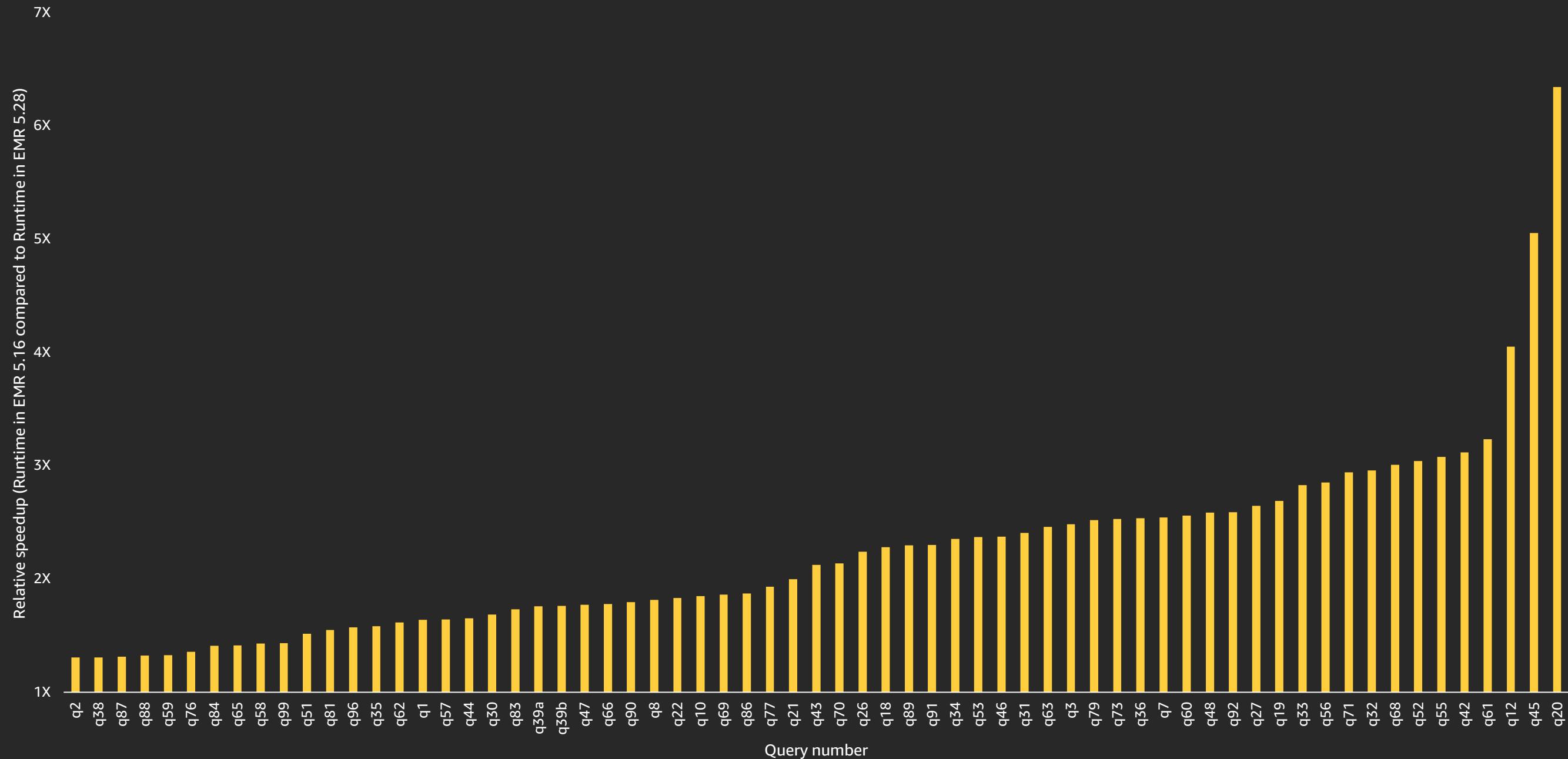
Upgrade to the latest version helps you save costs



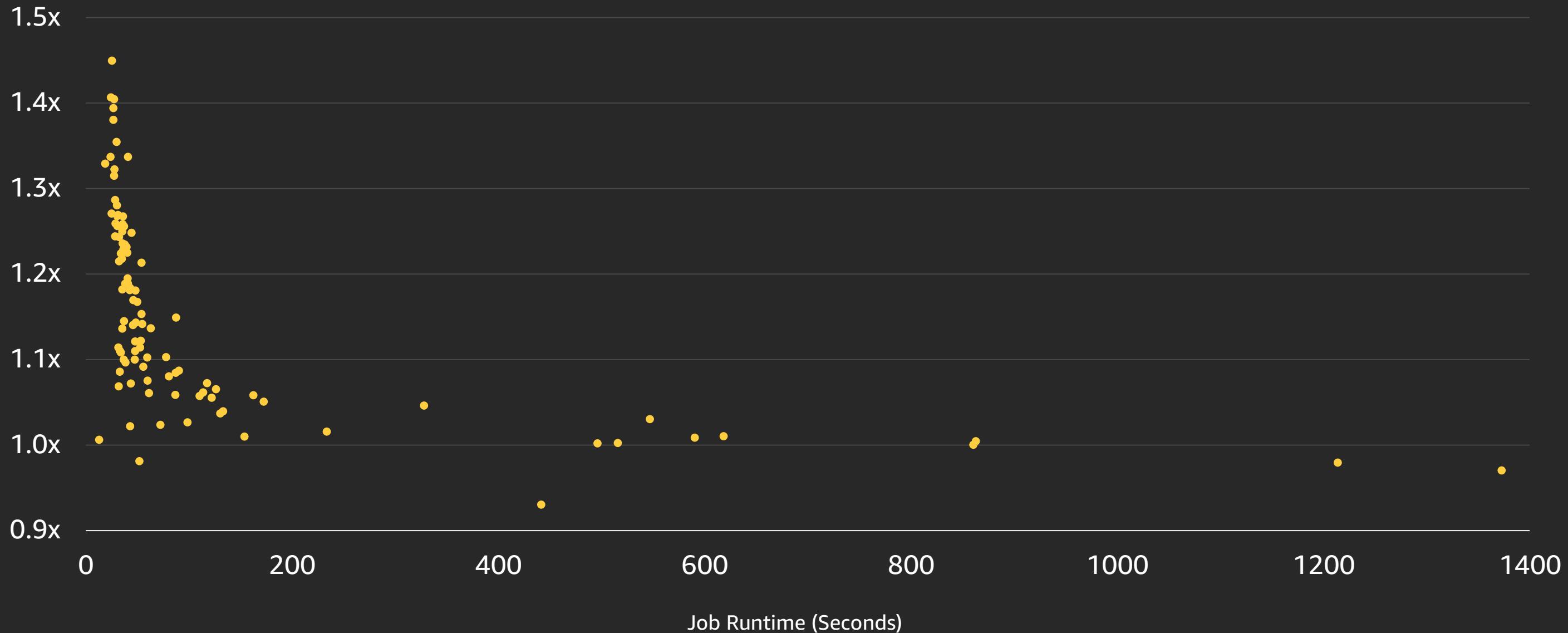
Average of 5x speed up on long-running queries



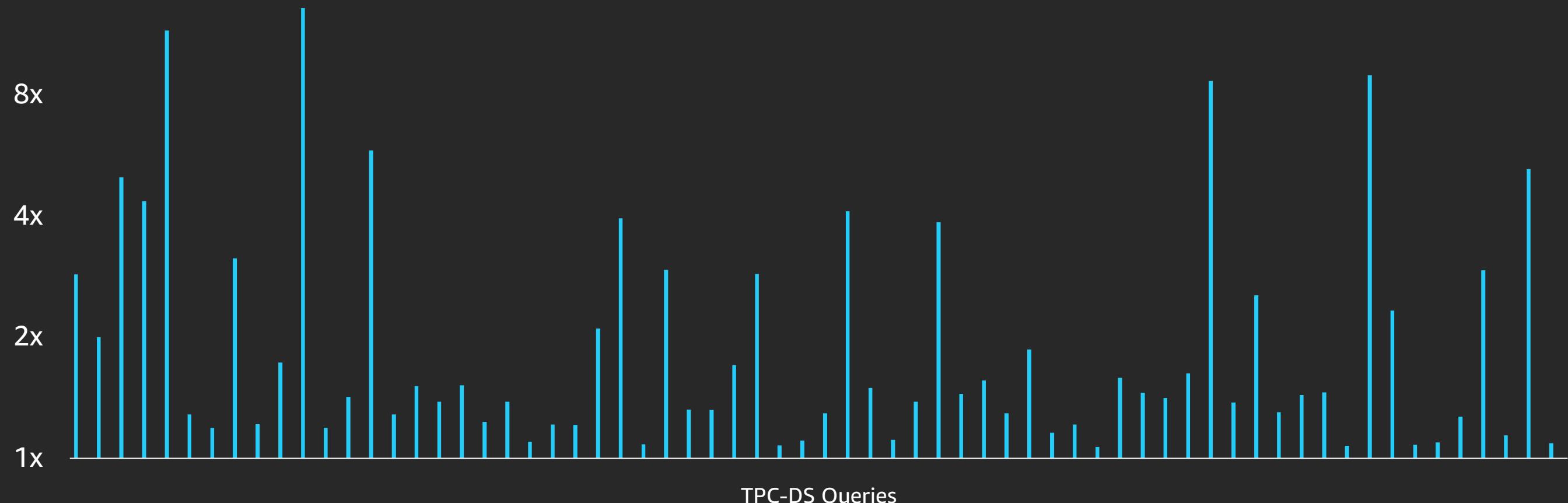
Average of 2x speed up on short-running queries



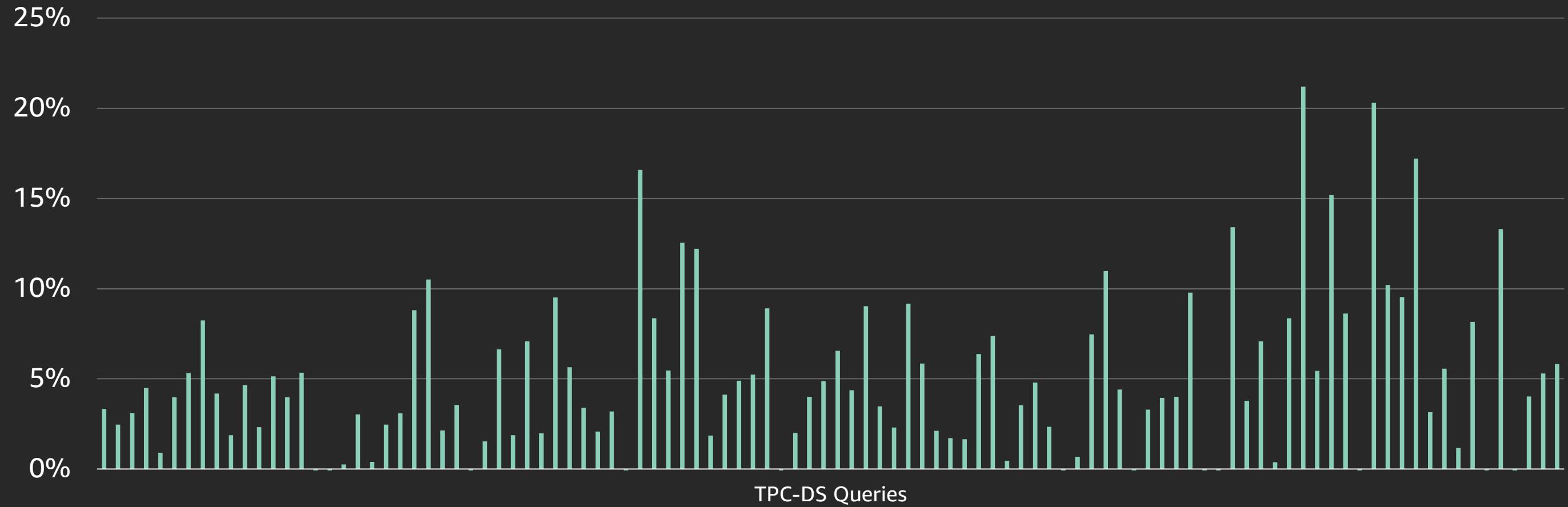
Job startup—eager worker allocation



Planning/optimization—dynamic partition pruning



Query execution—data prefetch



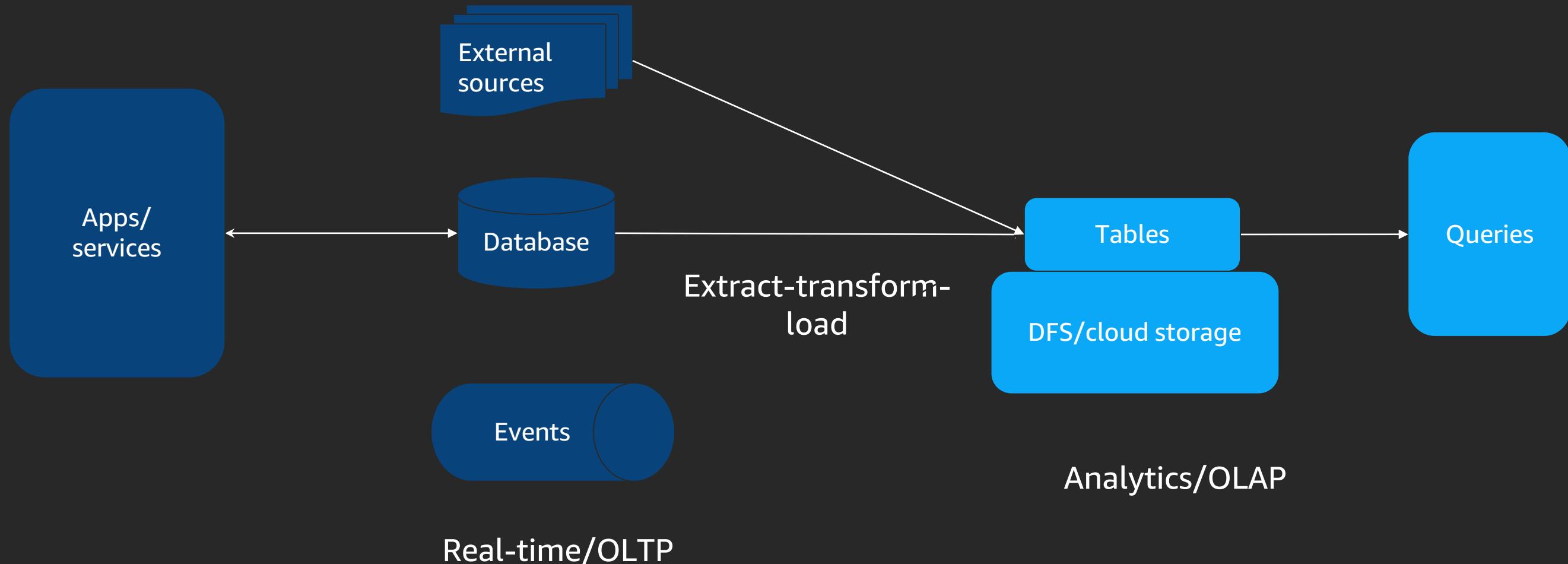
Optimization deep dive

- Configuration
 - CPU/disk ratios, driver/executor conf, heap/GC, native overheads, instance defaults
- Planning/optimization
 - **Dynamic partition pruning**, join reordering w/o stats and more
- Query execution
 - **Data pre-fetch** and more
- Job startup
 - **Eager executor allocation**, and more

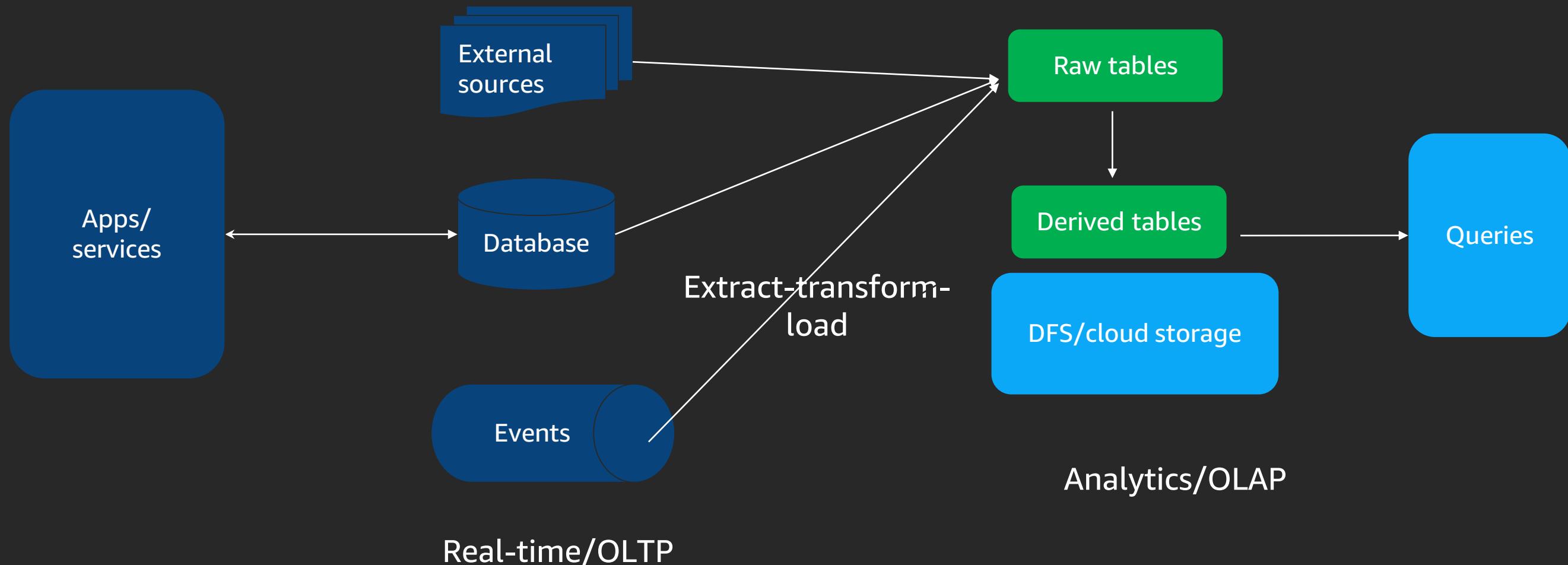
2

Record-level insert, update, and delete with Apache Hudi

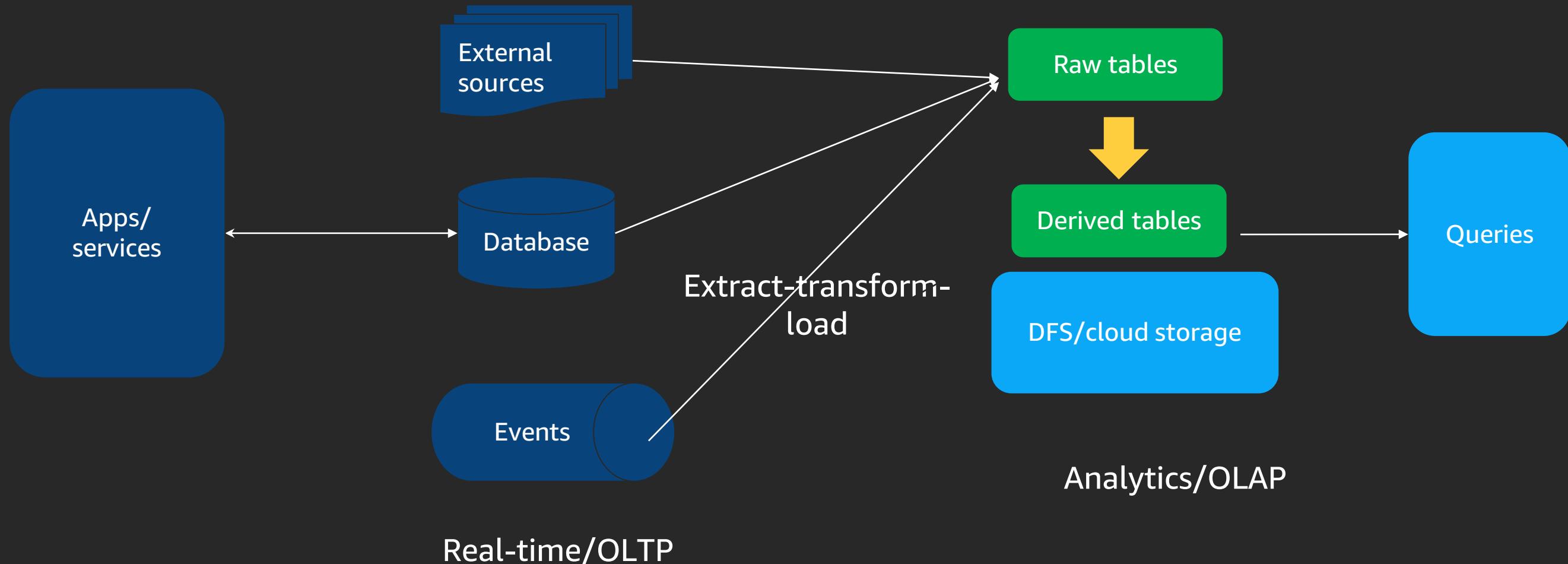
Building a data lake is simple



Typical lakes



Typical lakes



Common use cases are still difficult

Data deletions: Privacy regulation

Enforcing data privacy laws

Delete data within a specific time frame

Delete data across all data sets

Identifying files & partitions with specific data

Lack of indexes on data lake storage

Change Data Capture & apply

Applying change streams to Amazon Simple Storage Service (Amazon S3) data

Bulk loads don't scale

No support for upserts

Data quality is a serious concern

Need similar guarantees as a database

Inserts, updates, deletes



Users	
Column	Type
userID	int
country	string
last_modified	long
...	...

Data lake
Amazon S3

Streaming data ingestion

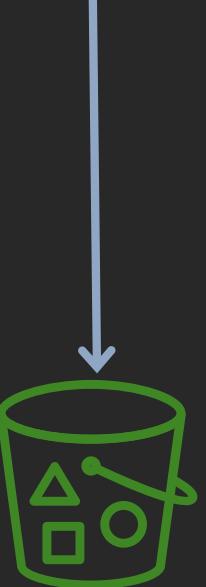
High-volume, time-ordered data

Duplicate events mess up analytics

Need fast ingestion to Amazon S3

Schema management, checkpointing

Produce events
in Apache Kafka



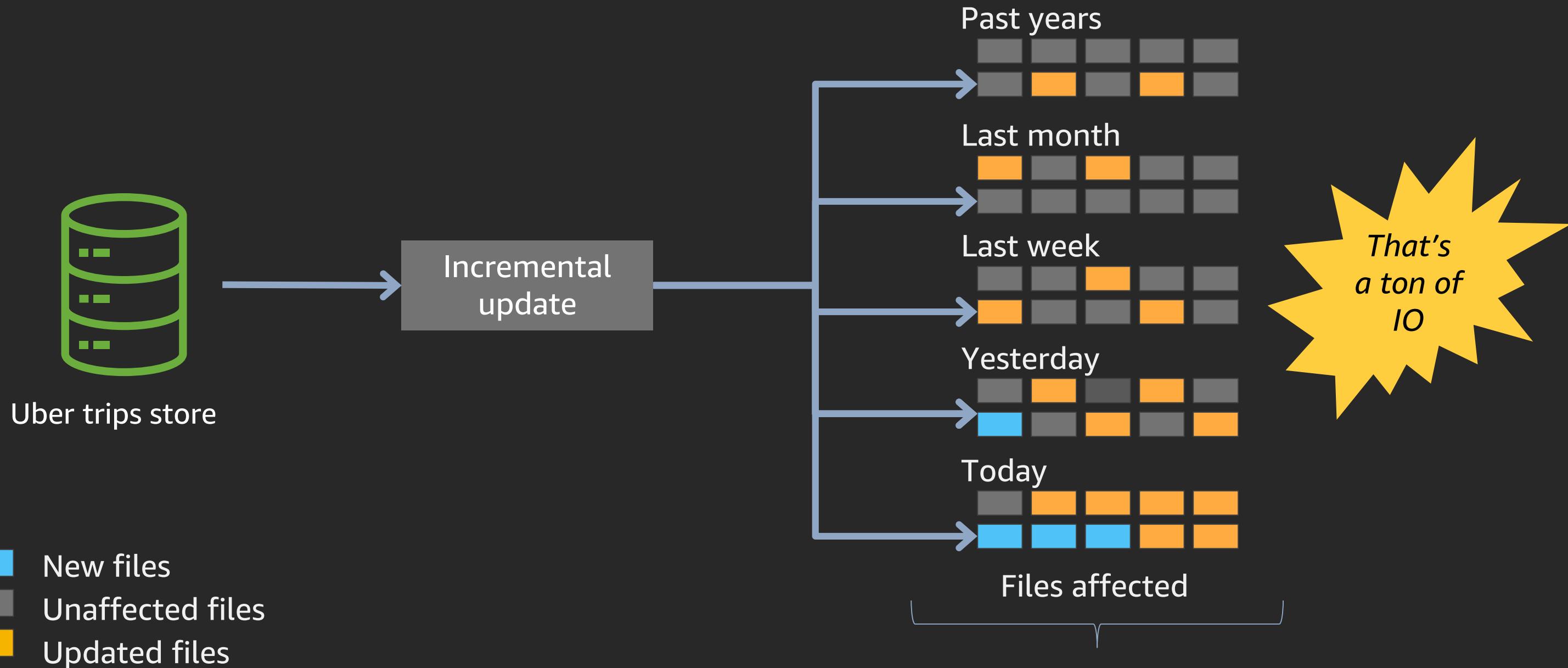
Amazon S3

Impressions

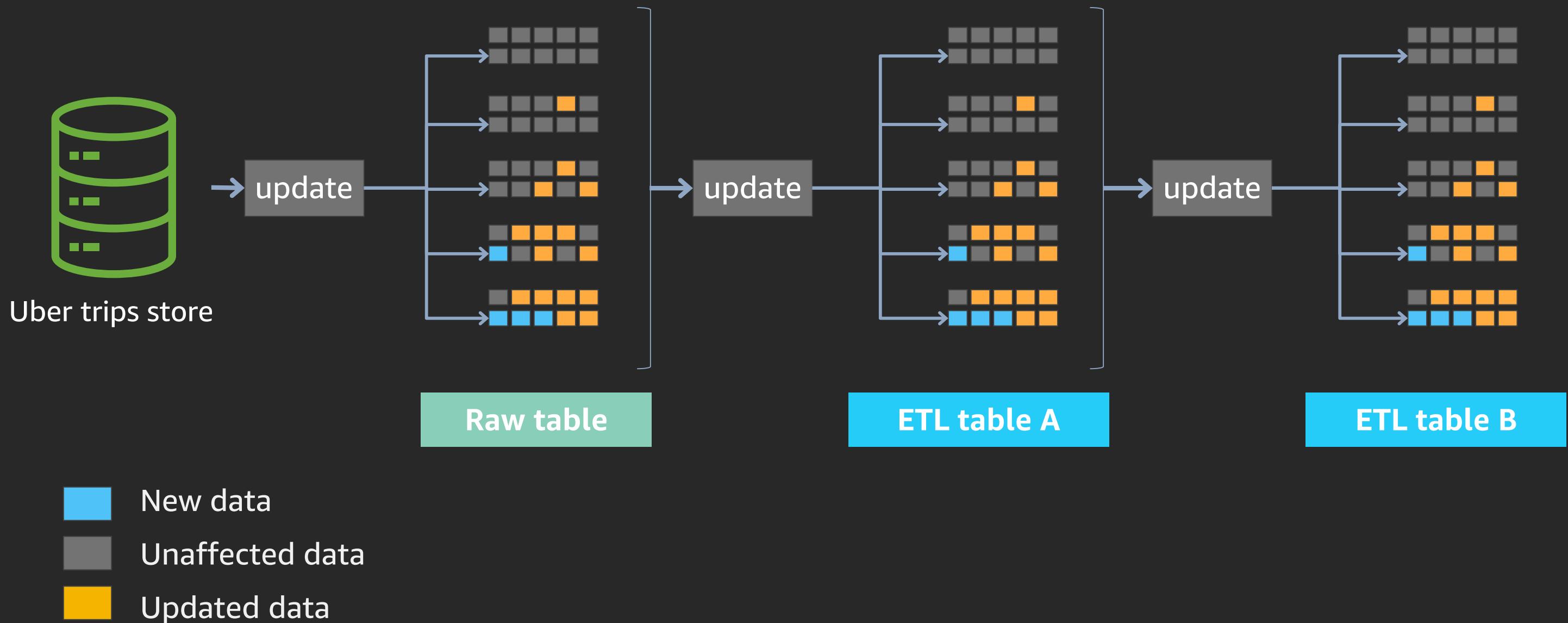
Field Name	Type
event_id	string
datestr	string
time	long

Uber's Motivating Use Case

Motivating use case

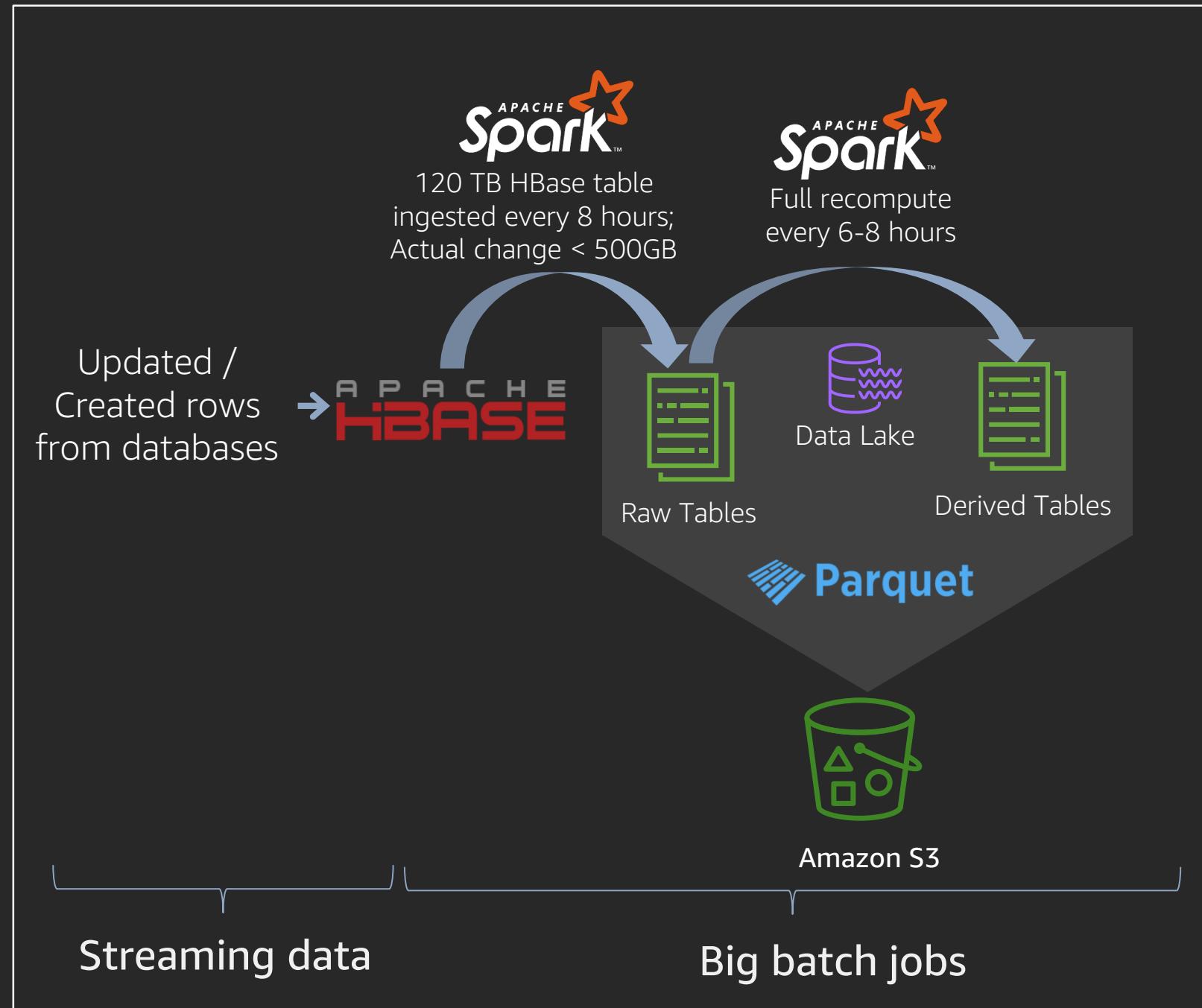


Cascading effects



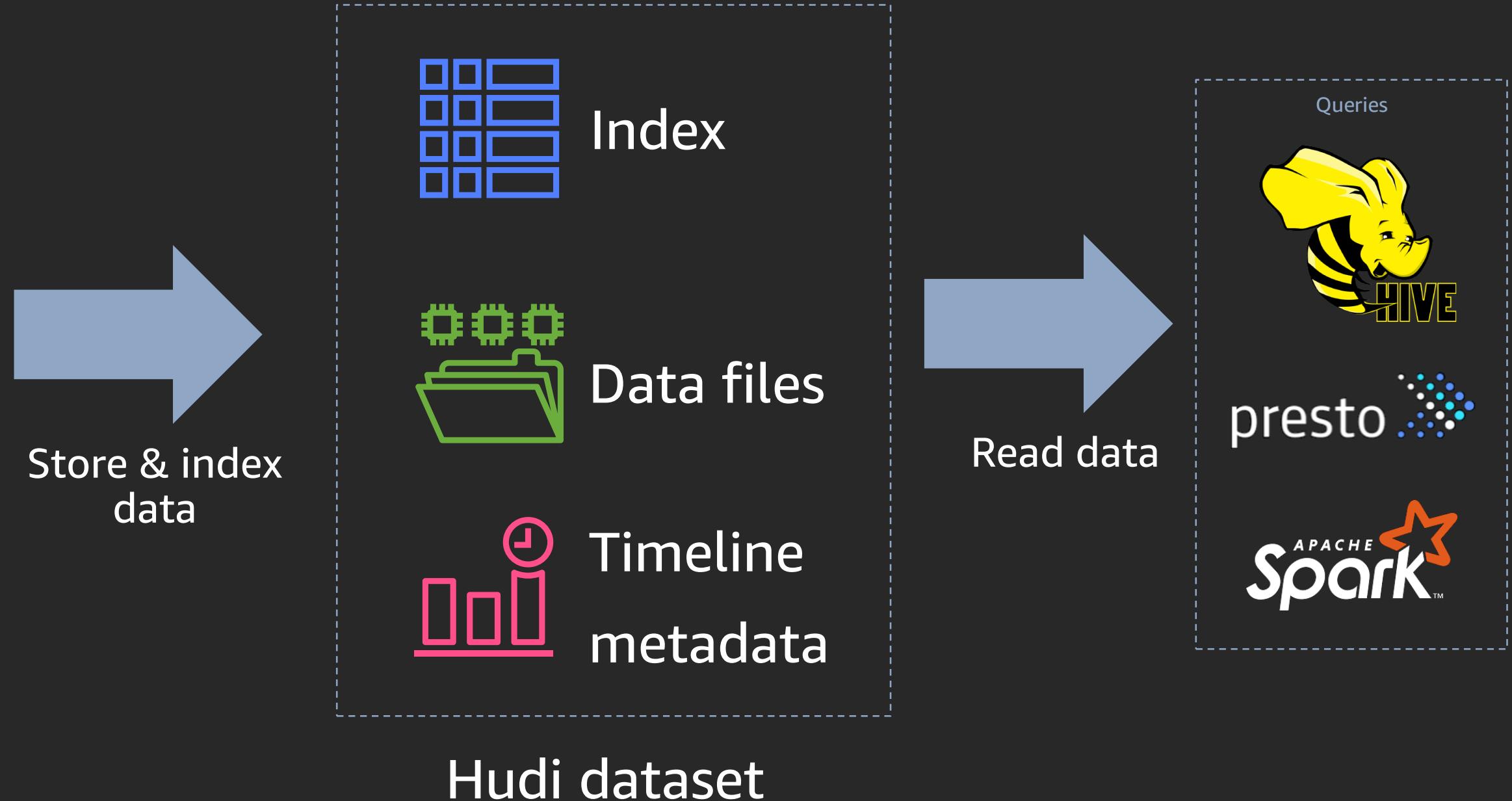
Slow data lake

120 TB Apache Hbase table ingested every
8 hours to account for less than 500GB of
data



Introducing Apache Hudi (incubating)


Hudi Spark Data
Source



Why You Should Care

- Near real-time data ingestion
- Batch jobs on steroids
- Unified, optimized analytical storage
- Row-level deletions to simplify data privacy
- Building block for great data lakes
- All open source, open formats

Primitives supported

- `upsert()` support with fast, pluggable indexing
- Atomic publish with rollback, save points
- Snapshot isolation between writer & queries
- Manages file sizes, layout using statistics
- Async compaction of row & columnar data
- Timeline metadata to track lineage

Storage types

Copy On Write
Read heavy



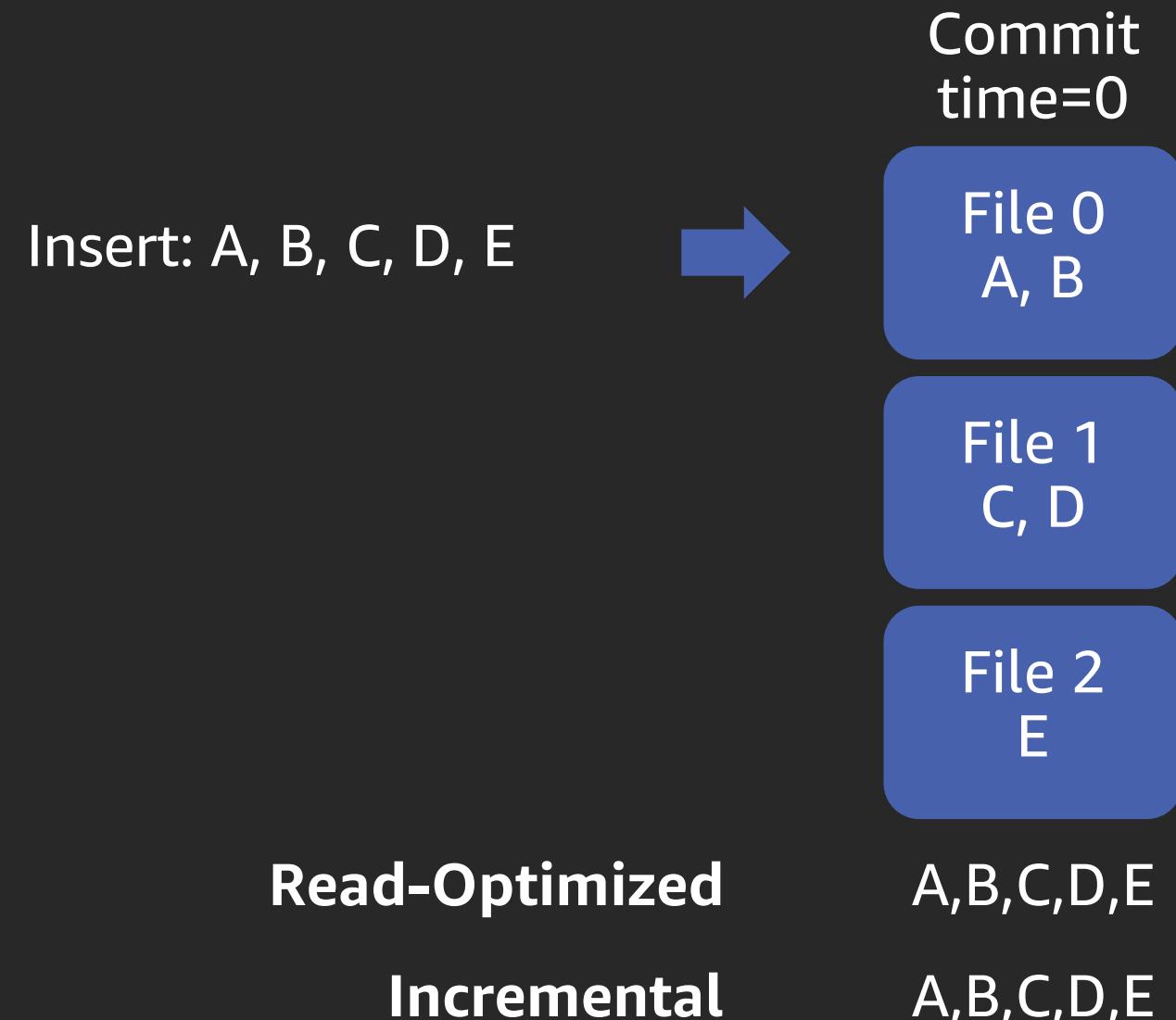
Hudi
Dataset

Merge On Read
Write heavy

Storage types & Views

Storage Type: Copy On Write

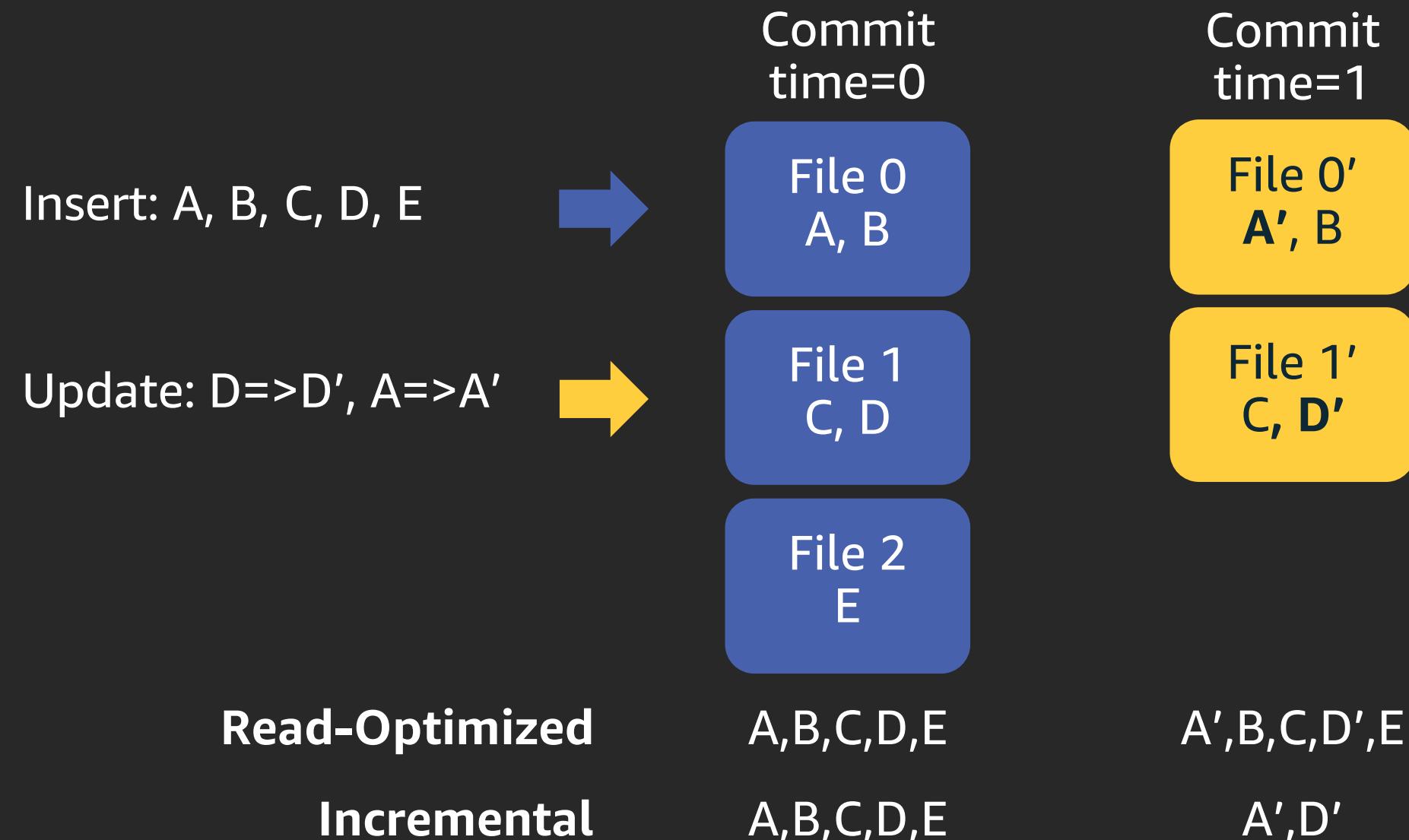
Views/Queries: Read-Optimized, Incremental



Storage types & Views

Storage Type: Copy On Write

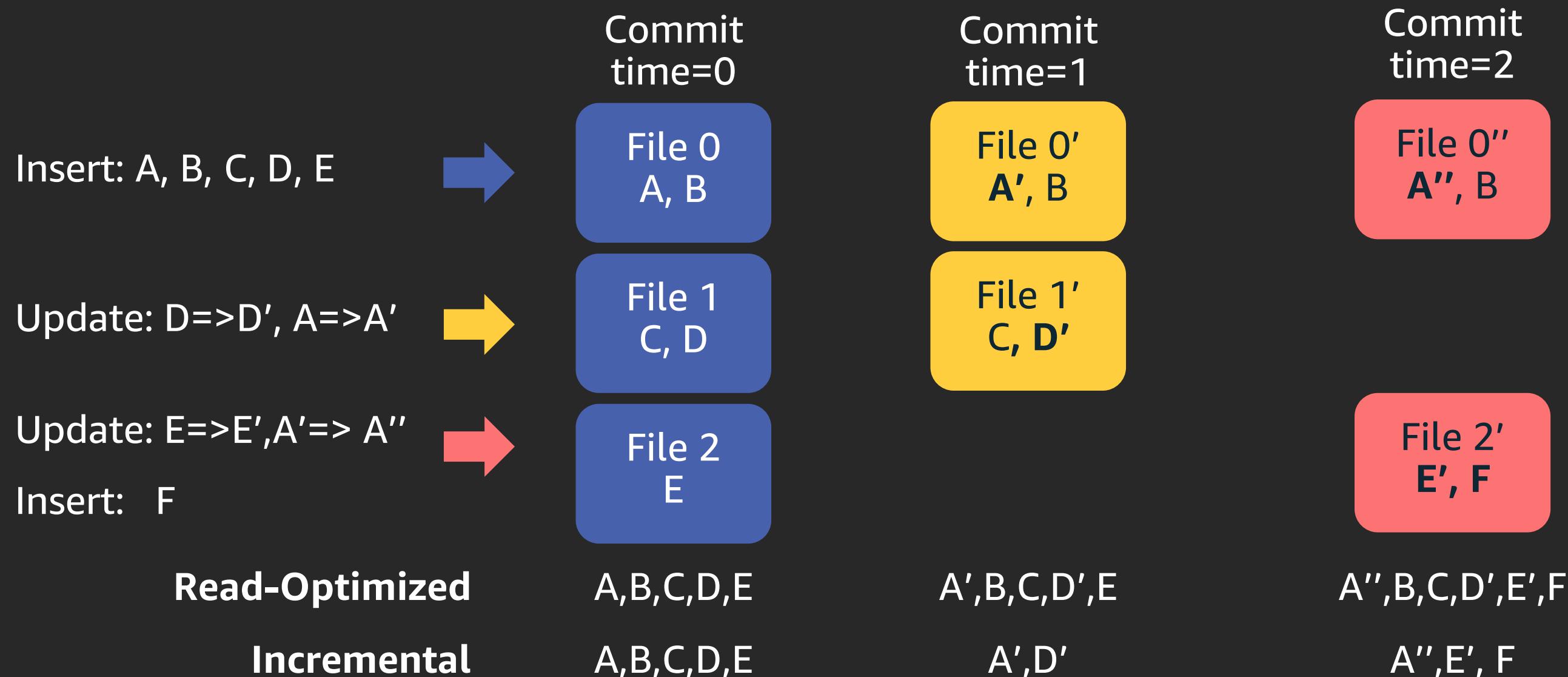
Views/Queries: Read-Optimized, Incremental



Storage types & Views

Storage Type: Copy On Write

Views/Queries: Read-Optimized, Incremental



Storage types & Views

Storage type: Copy On Write

Views: Read Optimized, Incremental

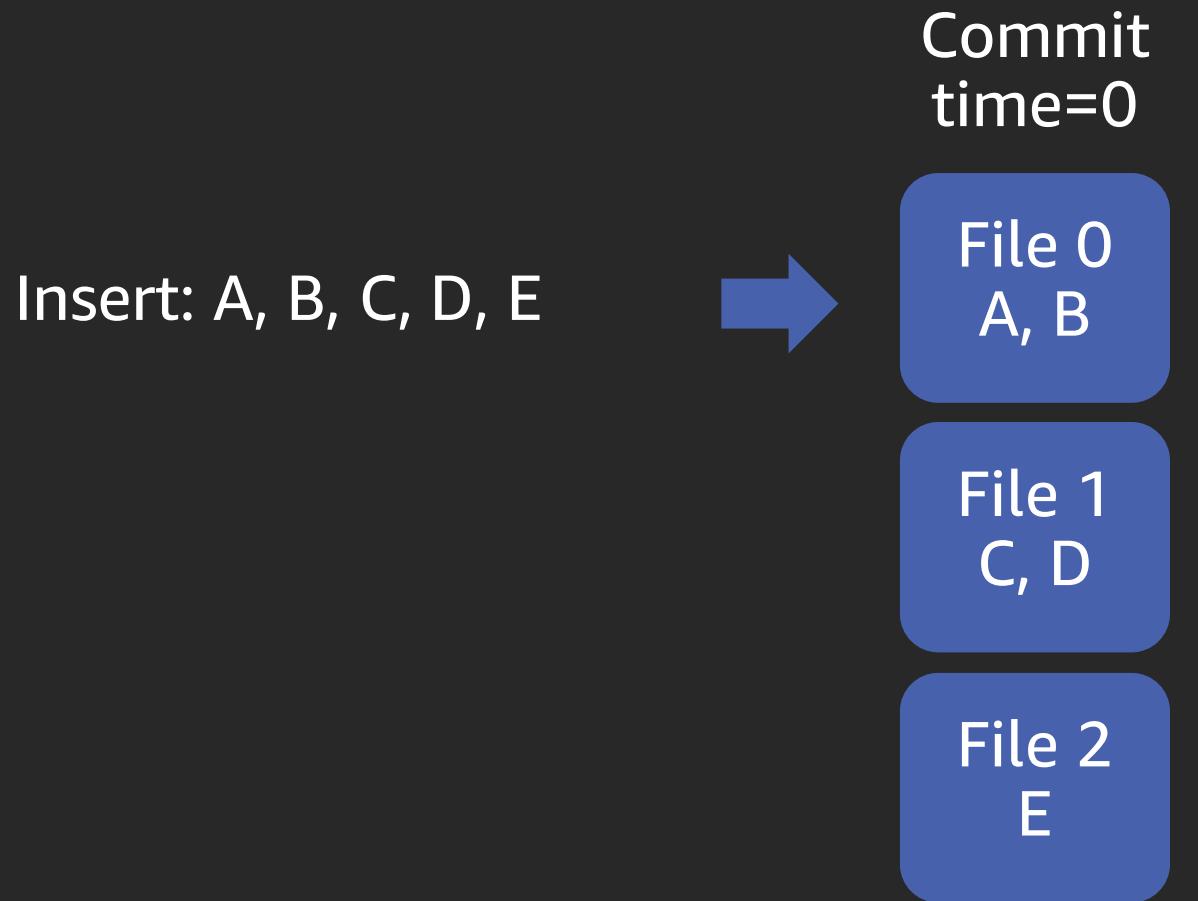
When to Use

- Your current job is rewriting entire table/partition to deal with updates
- Your workload is fairly well understood and does not have sudden bursts
- You're already using Parquet files for your tables
- You want to keep things operationally simple

Storage types & Views

Storage type: Merge On Read

Views/Queries: Read Optimized, Incremental, Real Time



Real time

A,B,C,D,E

Incremental

A,B,C,D,E

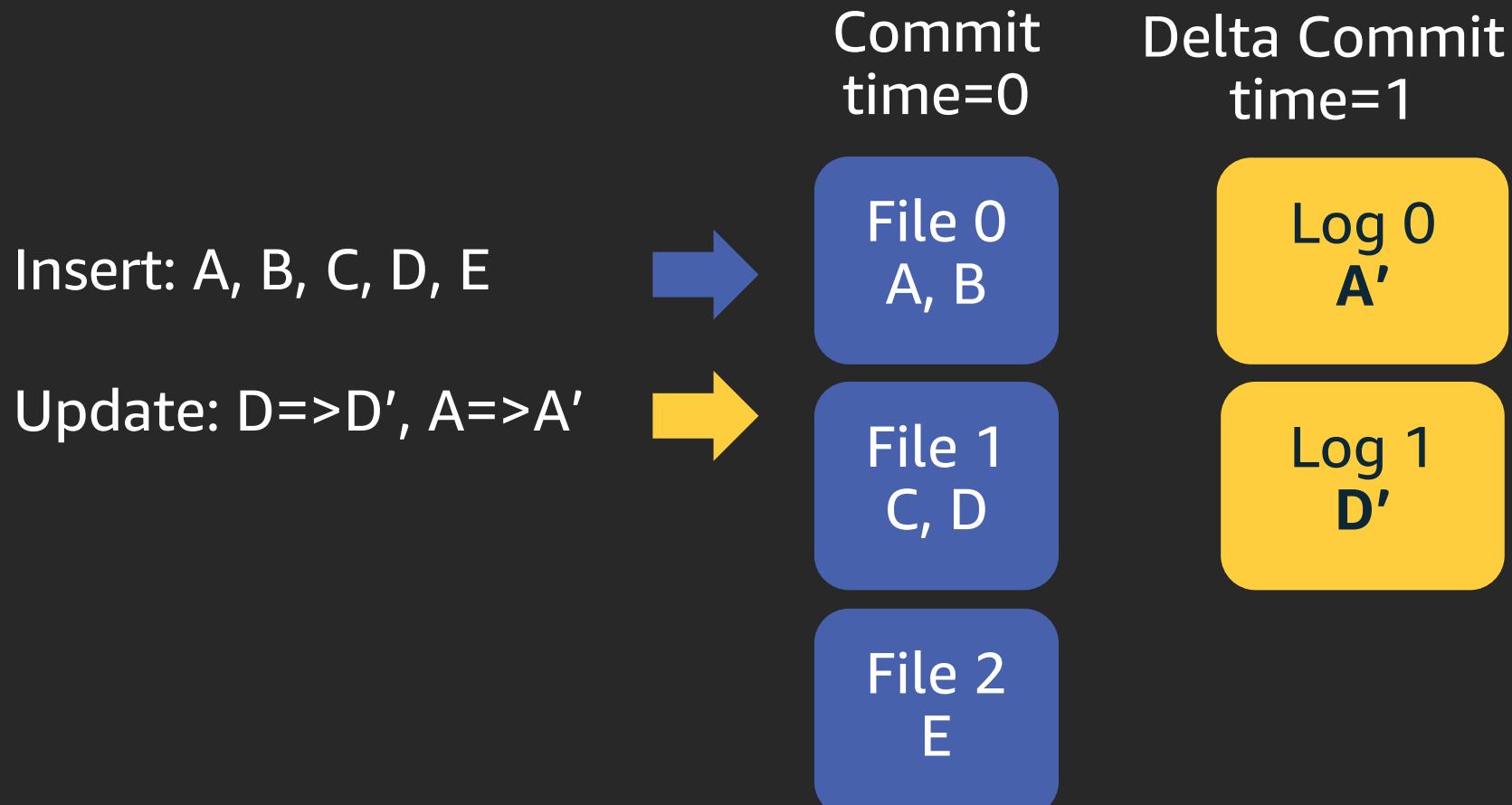
Read-Optimized

A,B,C,D,E

Storage types & Views

Storage type: Merge On Read

Views/Queries: Read Optimized, Incremental, Real time



Real time

A, B, C, D, E

A', B, C, D', E

Incremental

A, B, C, D, E

A', D'

Read-Optimized

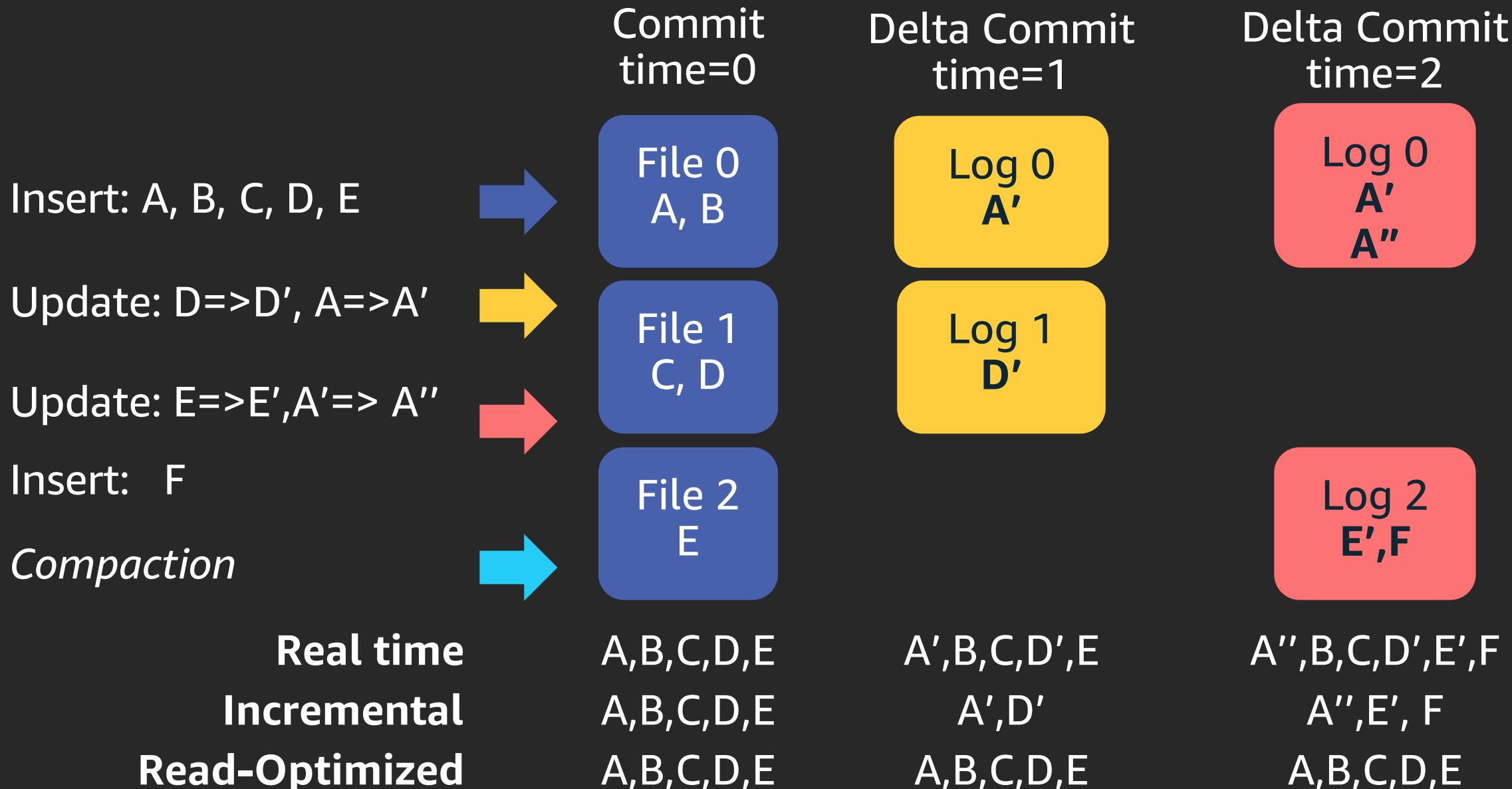
A, B, C, D, E

A, B, C, D, E

Storage types & Views

Storage type: Merge On Read

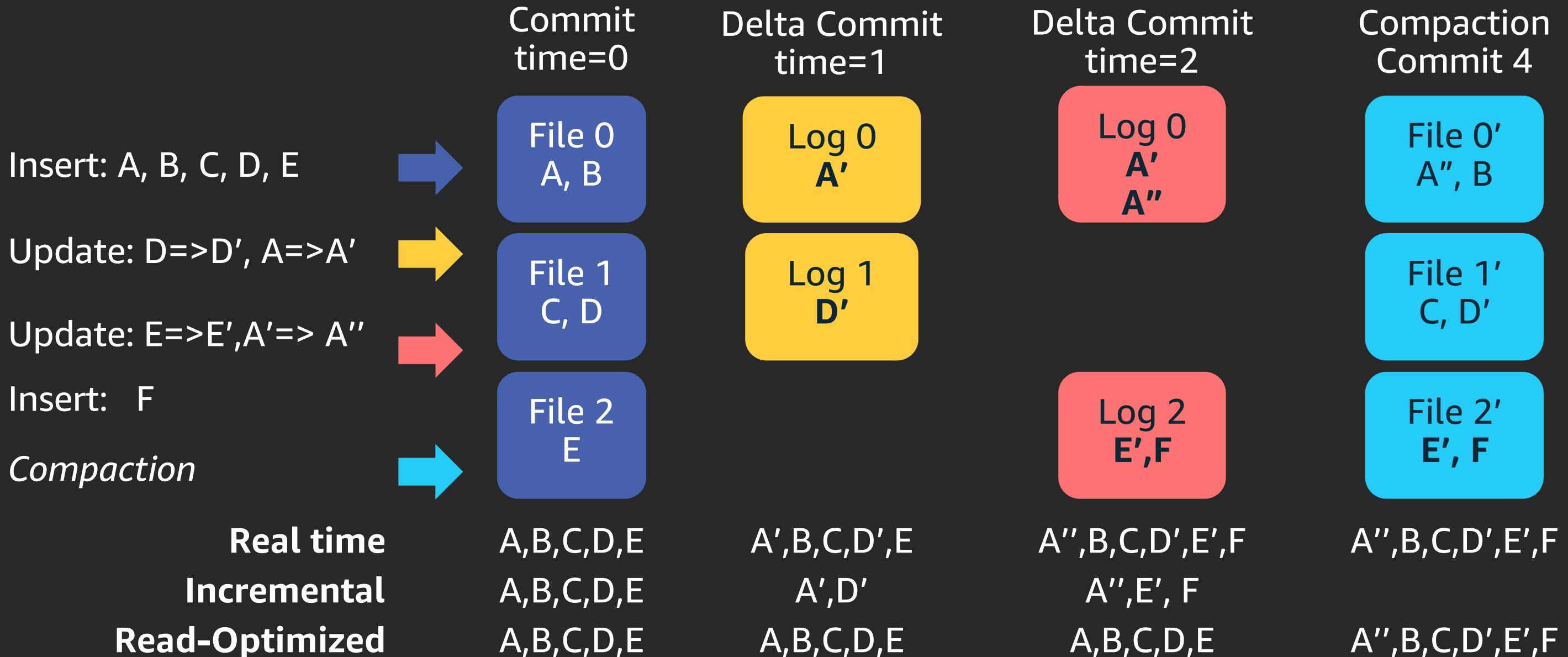
Views/Queries: Read Optimized, Incremental, Real time



Storage types & Views

Storage type: Merge On Read

Views/Queries: Read Optimized, Incremental, Real time



Storage types & Views

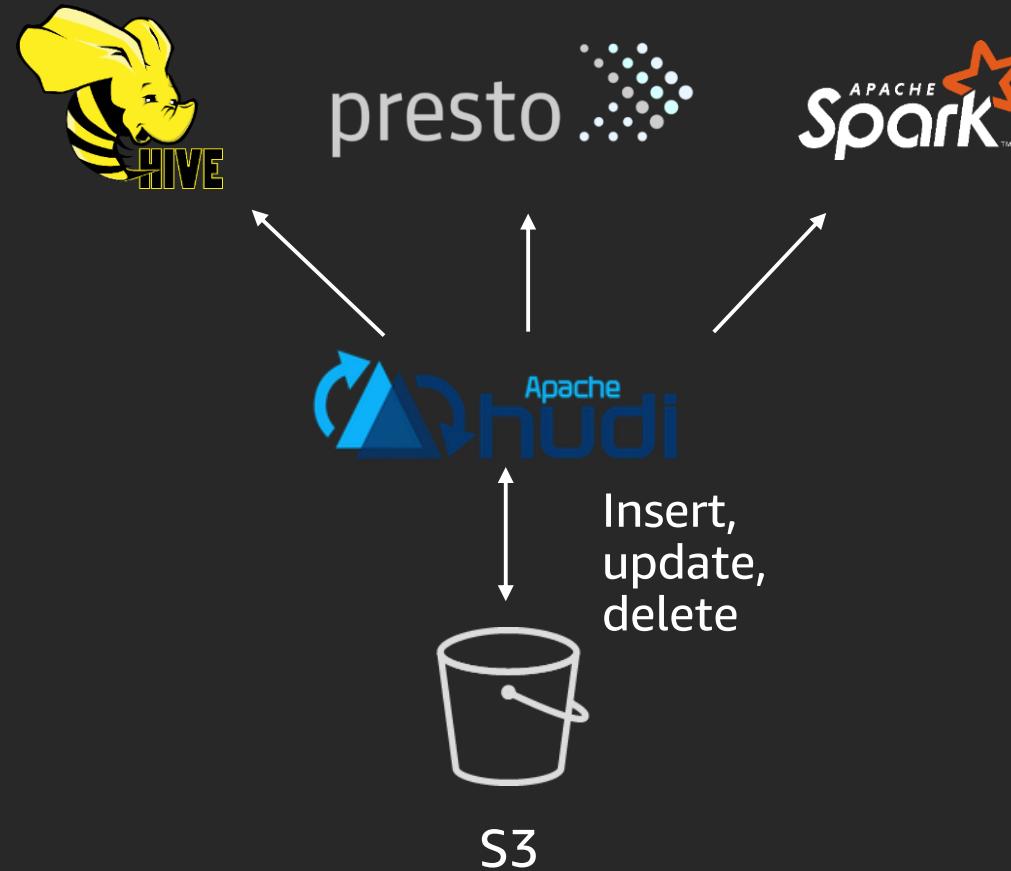
Storage type: Merge On Read

Views: Read Optimized, Incremental, Real time

When to Use

- You want ingested data available for query as fast as possible
- Your workload can have sudden spikes or changes in pattern
 - **Example:** Bulk updates to older transactions in upstream database cause updates to old partitions in Amazon S3

So why should you use Apache Hudi



Open source, open format, vendor neutral with Apache Hudi

Support for Spark, Hive, and Presto

Enables data lakes to

- a) Comply with data privacy laws
- b) Consume real-time streams and change data captures
- c) Reinstate late arriving data
- d) Track change history and rollback

Learn more

ANT239 - Insert, upsert, and delete data in Amazon S3 using Amazon EMR – Vinoth Chandar (Lead for Apache Hudi), Paul Coddington (Amazon EMR) - Wednesday, Dec 4, 1:45 PM - 2:45 PM– Mirage, Events Center B3 Green

ANT301-R1 - [REPEAT 1] Build & optimize Spark data pipelines for incremental data processing (Workshop)

Demo – Apache Hudi with EMR

Simplified view of data engineering platforms



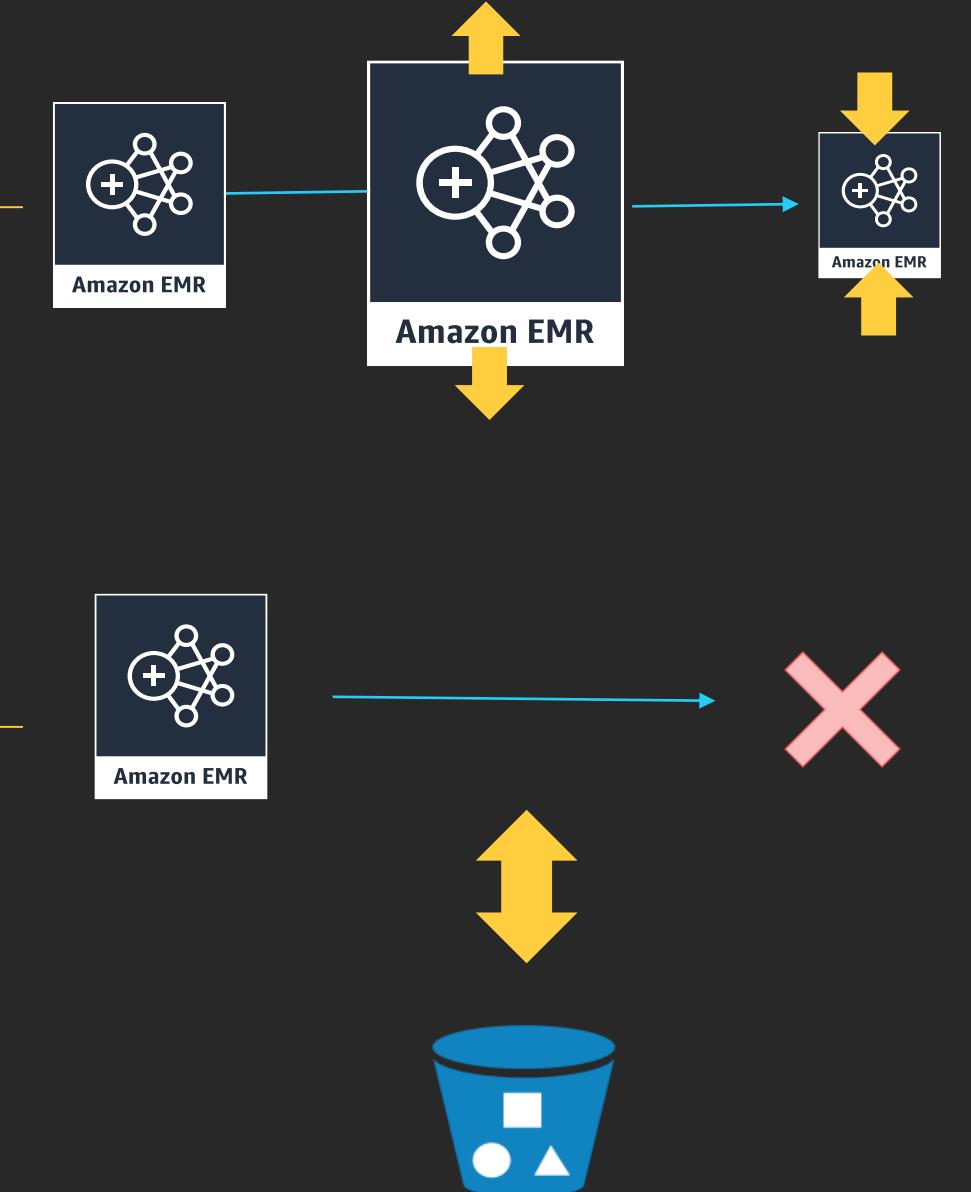
Notebooks



Apache Airflow



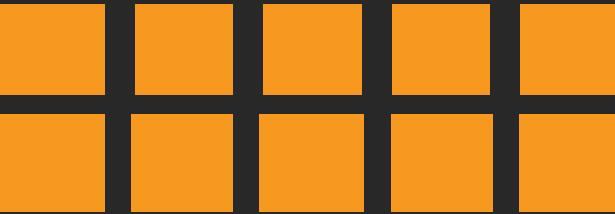
AWS Step Functions



3

Reduced cost with capacity-aware Spot provisioning

Scale up with Spot Instances



10 node cluster running for 14 hours
Cost = $1.0 * 10 * 14 = \$140$

Scale up cluster with Spot Instances



Add 10 more nodes on Spot

Scale up cluster with Spot Instances



20 node cluster running for 7 hours

$$\begin{aligned}\text{Cost} &= 1.0 * 10 * 7 = \$70 \\ &= 0.5 * 10 * 7 = \$35\end{aligned}$$

Total \$105

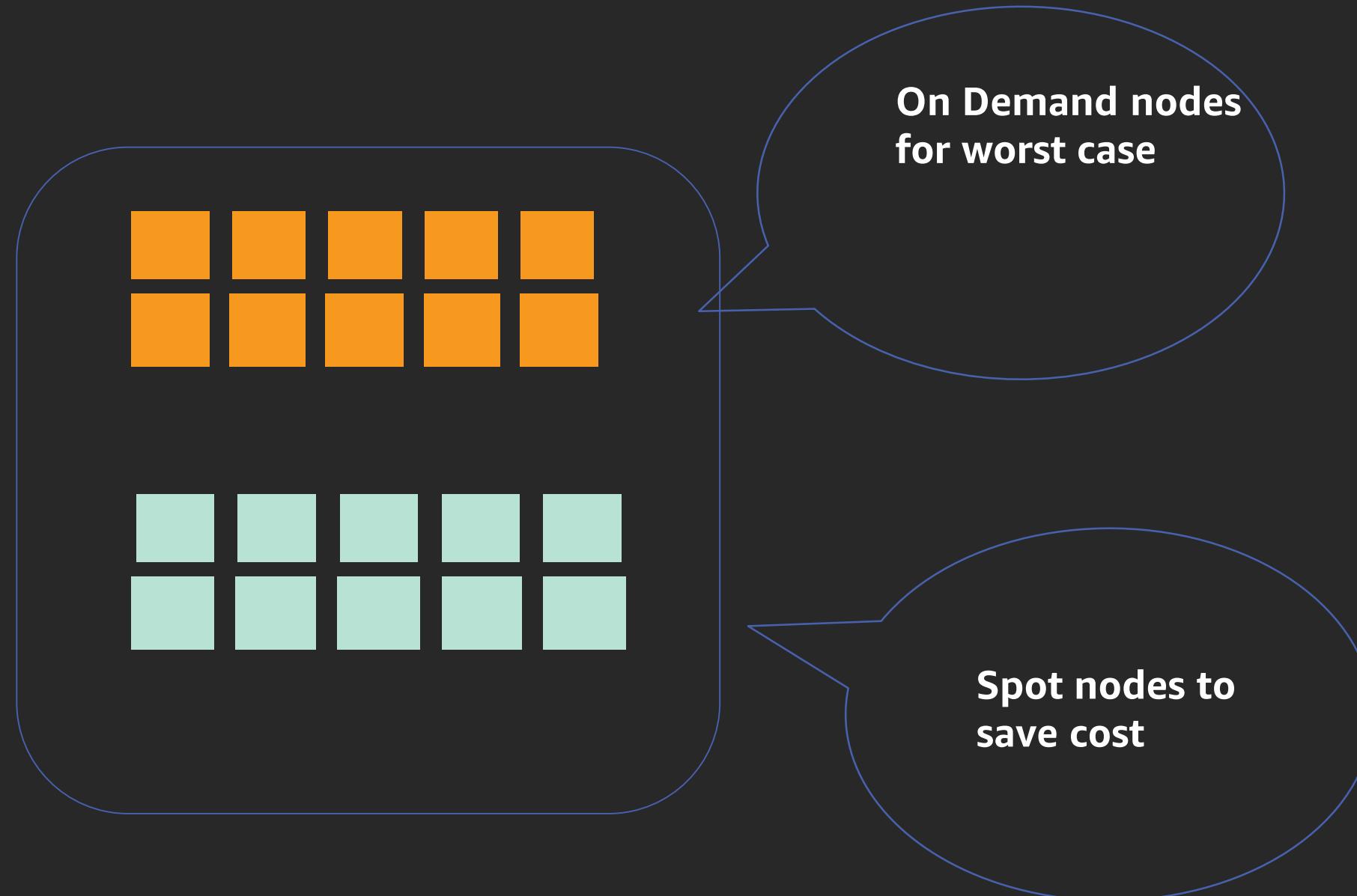
Scale up cluster with Spot Instances



50 % less run-time (14 → 7)

25% less cost (140 → 105)

Scale up cluster with Spot Instances



Scale up cluster with Spot Instances

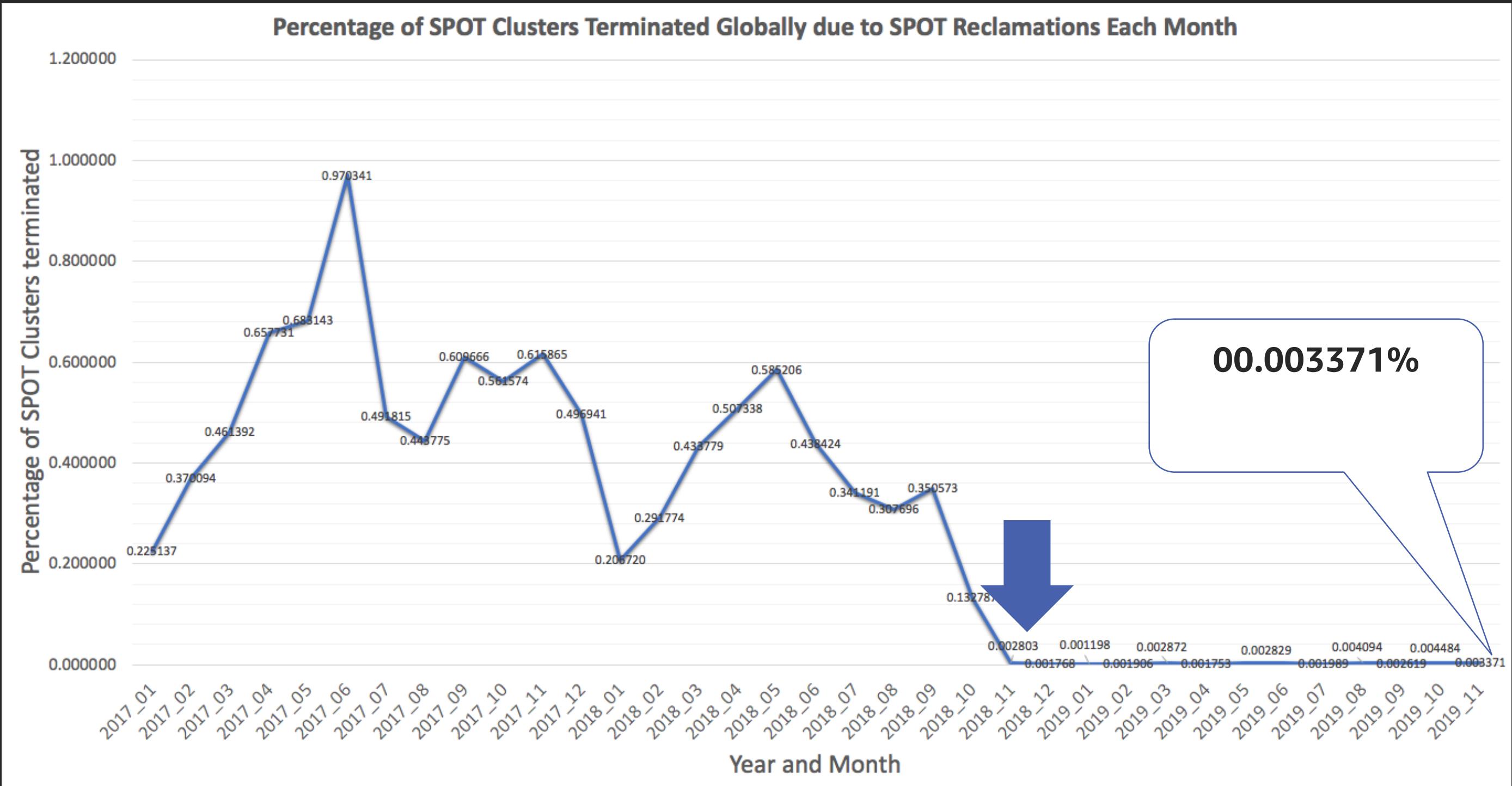


**On-demand nodes
for worst case**

**Spot nodes to
save cost**

**Up to 60% lower
cost with compute
savings plan**

"Spot" the interruption



4

Managed resize (private beta)

Managed resize (beta)

Completely managed environment for automatically resizing clusters

- No configurations required except min/max cost constraints
- More data points and faster reaction time
- Can save 20-60% costs depending on the workload pattern

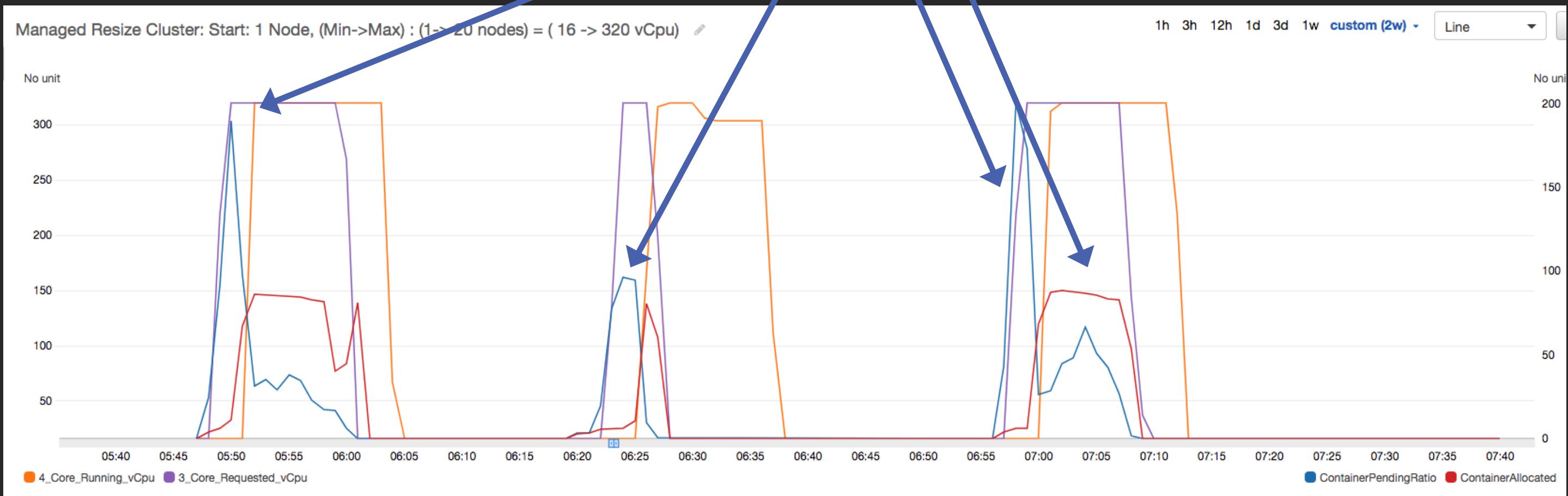
Similar to autoscaling Amazon EMR Clusters

- Use autoscaling for DIY scaling with custom metrics
- Managed resize for completely managed option

Email emr-pm@amazon.com to participate in the beta

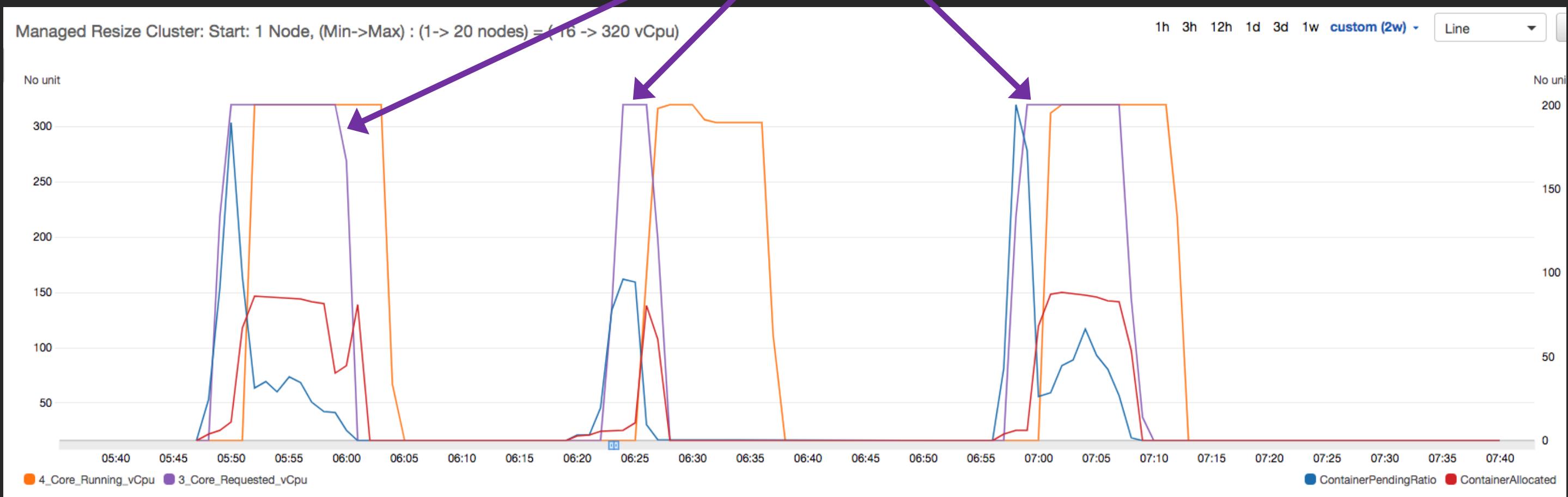
Managed resize

Load patterns



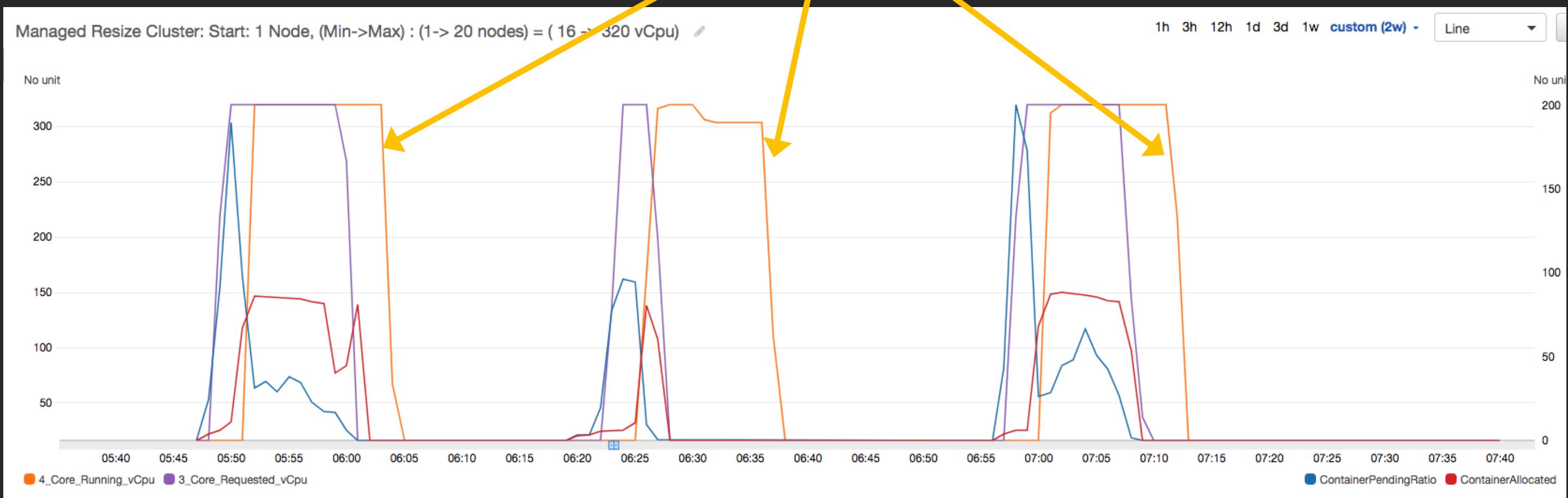
Managed resize

Requested scale



Managed resize

Scale down



Managed resize

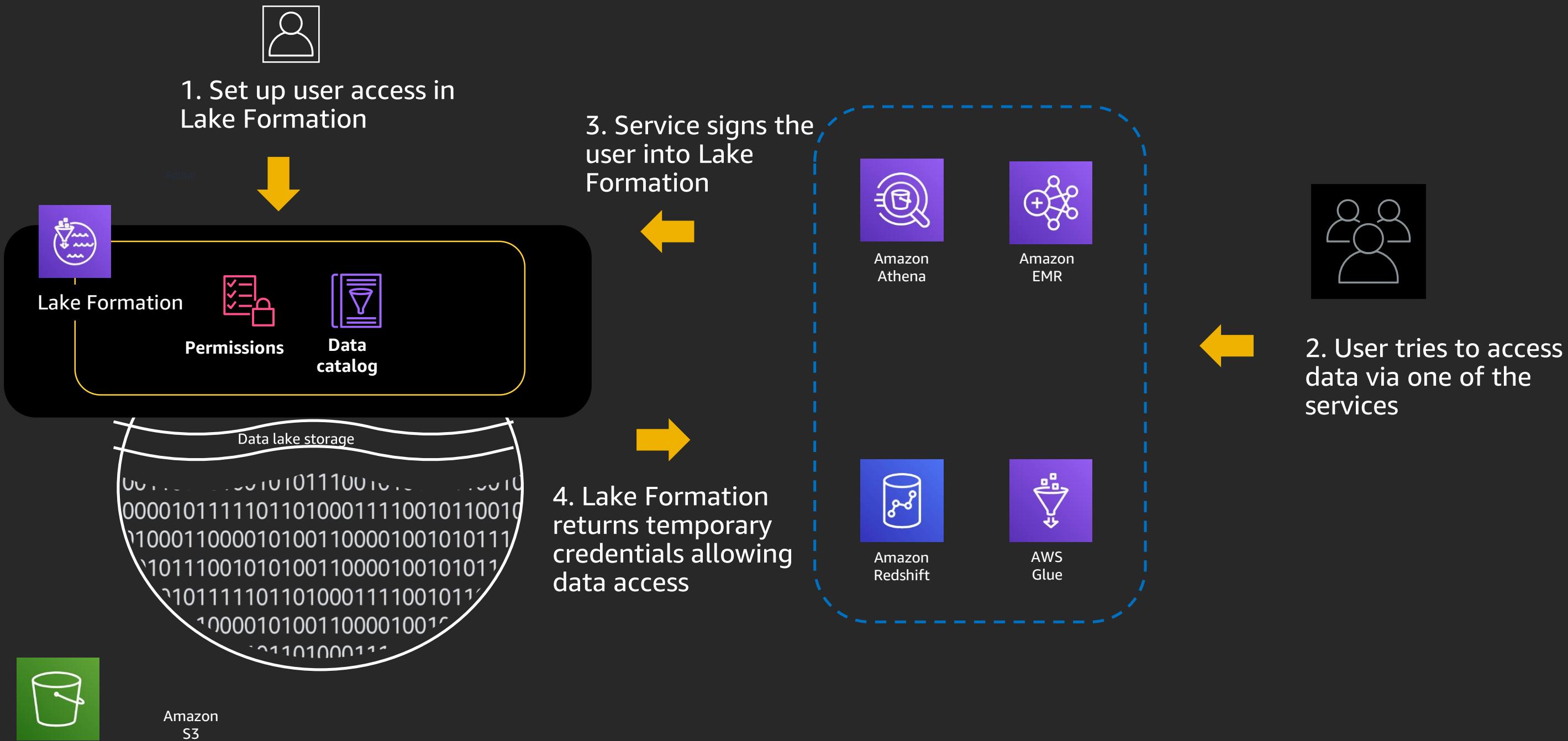
63% savings compared to fixed 20-node cluster



5

Fine-grained access control with AWS Lake Formation

Lake Formation: Secure once, access in many ways



Control data access with GRANT and REVOKE

AWS Lake Formation > Tables

Tables (2)

Name	Database	Location
sales	enterprise	s3://enterprise-data
elb_logs	sampledb	s3://athena-example

Actions ▾ Create table using a crawler ▾ Create table

Table Edit Last updated
 Mon, Oct 14, 2019, 9:46 PM UTC

Drop View data Permissions
 Grant Revoke Verify permissions View permissions

Grant permissions sales

Choose the access permissions to grant. IAM permissions must also allow access.

IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add

Active Directory users and groups (EMR beta only)
Enter one or more Active Directory users or groups.
Ex: arn:aws:iam::<AccountID>:saml-provider/<SamlProviderName>:user/<UserName>

Column - optional
Choose filter type
Exclude columns

Exclude columns - optional
Grant permissions to access all but the selected columns.
Choose columns

total_revenue X string total_cost X string total_profit X string

Table permissions
Choose the specific access permissions to grant.
 Alter Insert Drop Delete Select

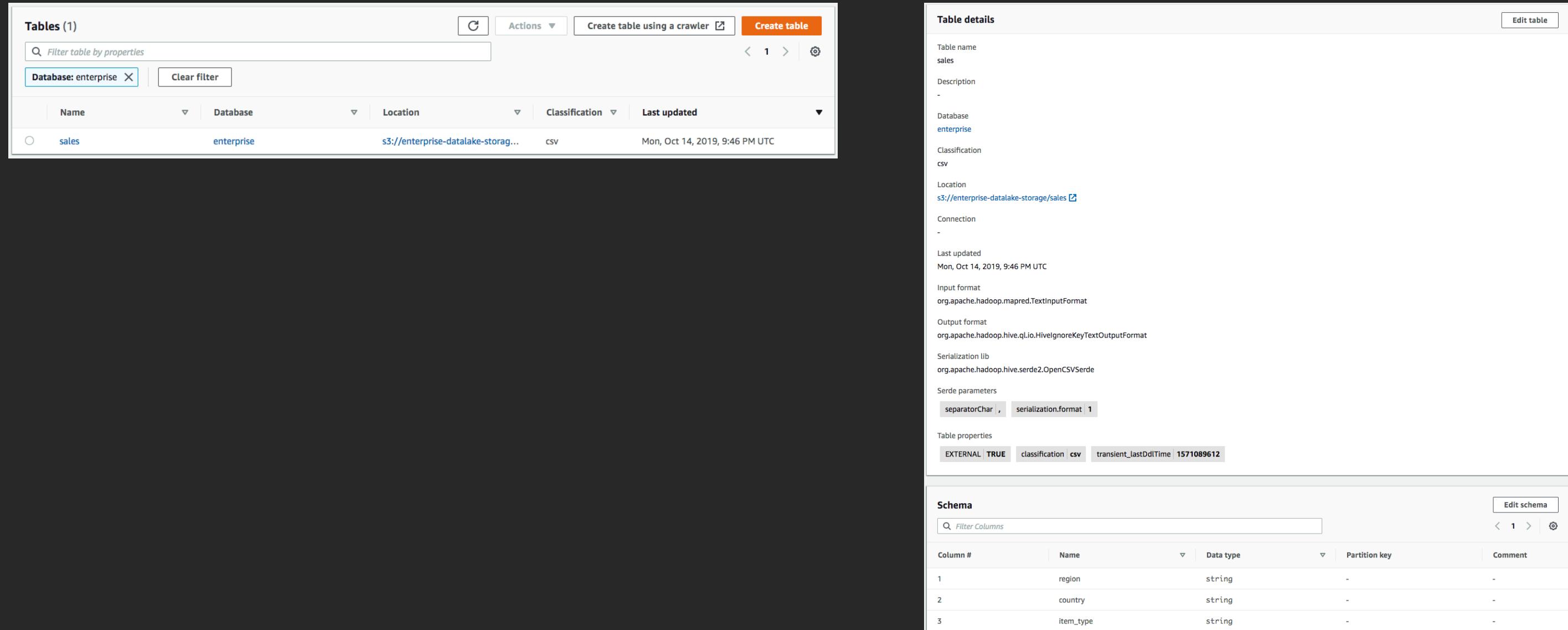
Super
This permission is the union of the individual permissions above and supersedes them. [See here](#)

Grantable permissions
Choose the permissions that may be granted to others.
 Alter Insert Drop Delete Select

Super
This permission allows the principal to grant any of the above permissions and supersedes those grantable permissions.

Cancel **Grant**

Permissions on tables and columns, not Amazon Simple Storage Service (Amazon S3)



Tables (1)

Actions ▾ Create table using a crawler ▾ Create table

Filter table by properties

Database: enterprise X Clear filter

Name	Database	Location	Classification	Last updated
sales	enterprise	s3://enterprise-datalake-storage/sales	csv	Mon, Oct 14, 2019, 9:46 PM UTC

Table details

Table name: sales

Description: -

Database: enterprise

Classification: csv

Location: s3://enterprise-datalake-storage/sales

Connection: -

Last updated: Mon, Oct 14, 2019, 9:46 PM UTC

Input format: org.apache.hadoop.mapred.TextInputFormat

Output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

Serialization lib: org.apache.hadoop.hive.serde2.OpenCSVSerde

Serde parameters: separatorChar , serialization.format 1

Table properties: EXTERNAL, TRUE, classification: csv, transient_lastDdlTime: 1571089612

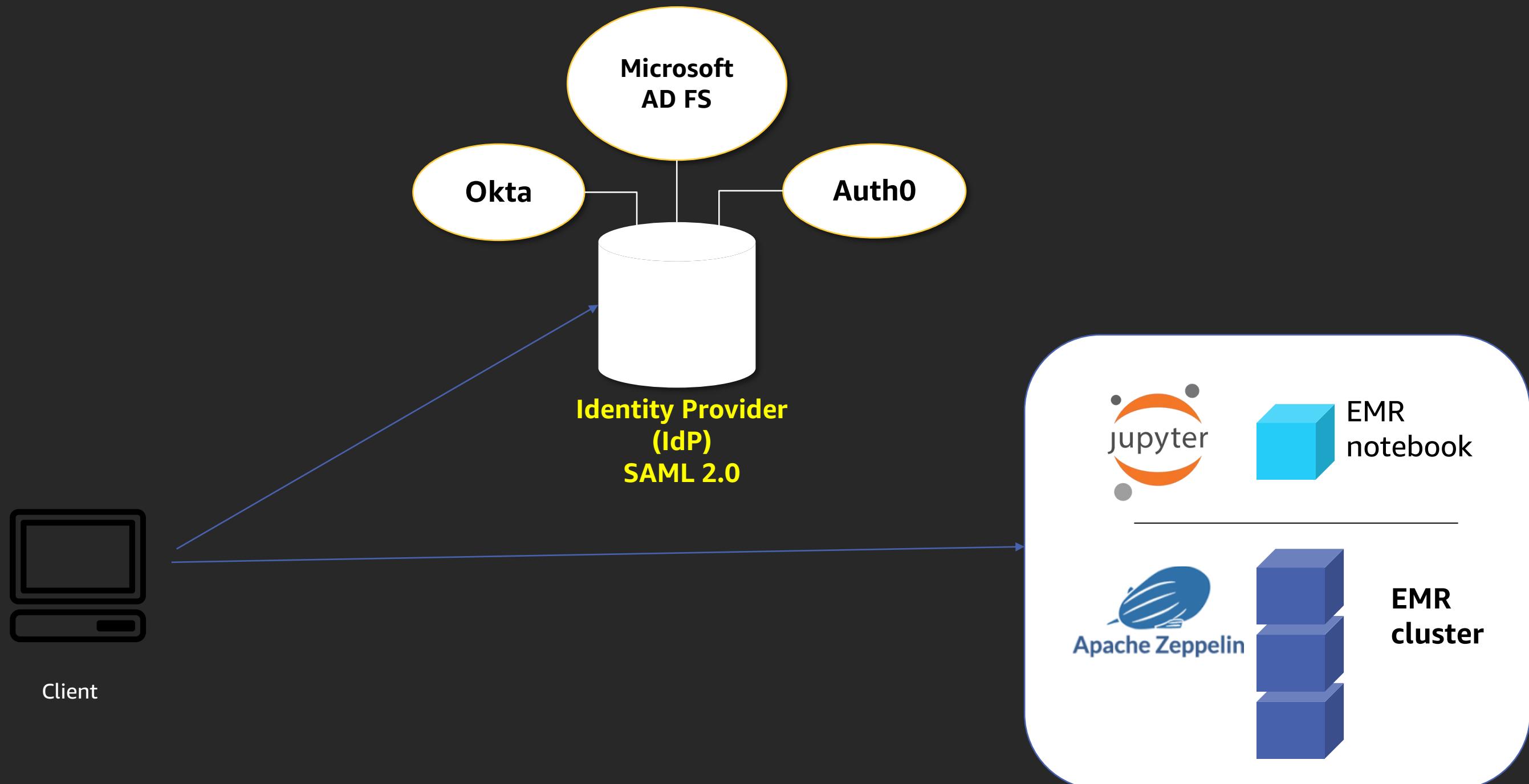
Schema

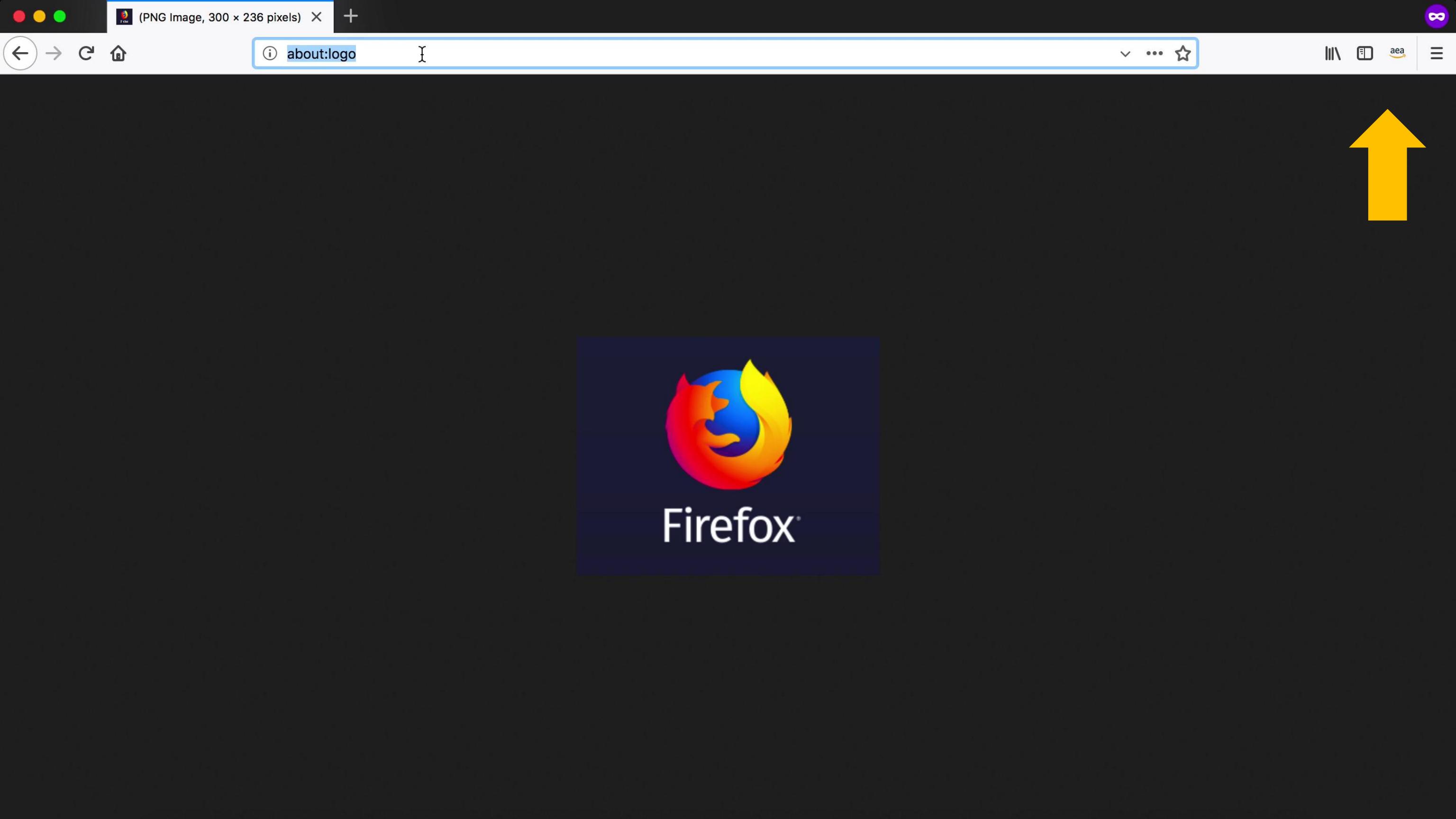
Column #	Name	Data type	Partition key	Comment
1	region	string	-	-
2	country	string	-	-
3	item_type	string	-	-

Why integrate Amazon EMR with Lake Formation?

- **Fine-grained, column-level access to databases and tables**
 - Allows shared multi-tenant clusters to securely access data
 - Uses the AWS Glue Data Catalog as the metadata store
- **Federated single sign-on from your enterprise identity system**
 - Active Directory Federation Services (AD FS), Auth0, Okta, and many others
 - Uses Security Assertion Markup Language (SAML) 2.0

Enterprise single sign-on access to notebooks





Zeppelin

https://ec2-35-170-74-144.compute-1.amazonaws.com:8442/gateway/default/zeppelin/#/

Search bob

Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

Notebook

- [Import note](#)
- [Create new note](#)

Filter

- [Zeppelin Tutorial](#)
- [Trash](#)

Help

[Get started with Zeppelin documentation](#)

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

- [Mailing list](#)
- [Issues tracking](#)
- [Github](#)



Tables (7)



Actions ▾

Create table using a crawler

Create table

Filter by tags and attributes or search by keyword

< 1 >



Database : record_server_basic X

	Name	Database	Location	Classification	Last updated
<input type="radio"/>	csv	record_server_basic	s3://record-server-int...	-	Tue, Jul 2, 2019, 11:34 PM UTC
<input type="radio"/>	avro_decimal	record_server_basic	s3://record-server-int...	-	Mon, Jul 1, 2019, 9:04 PM UTC
<input type="radio"/>	avro_catalog_sales	record_server_basic	s3://record-server-int...	-	Mon, Jul 1, 2019, 9:02 PM UTC
<input type="radio"/>	avro	record_server_basic	s3://record-server-int...	avro	Wed, Jun 19, 2019, 5:02 PM UTC
<input type="radio"/>	parquet	record_server_basic	s3://record-server-int...	parquet	Wed, Jun 19, 2019, 4:56 PM UTC
<input type="radio"/>	orc	record_server_basic	s3://record-server-int...	orc	Wed, Jun 19, 2019, 4:56 PM UTC
<input type="radio"/>	json	record_server_basic	s3://record-server-int...	json	Wed, Jun 19, 2019, 4:42 PM UTC

Dashboard

▼ Data catalog

Databases

Tables

Settings

▼ Register and ingest

Data lake locations

Blueprints

Crawlers Jobs 

▼ Permissions

Admins and database creators

Data permissions

Data locations

Tables (7)

New recording

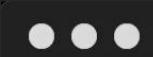
Actions Create table using a crawler 

Create table

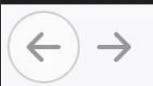
Filter by tags and attributes or search by keyword

< 1 > Database : record_server_basic 

	Name	Database	Location	Classification	Last updated
	csv	record_server_basic	s3://record-server-int...	-	Tue, Jul 2, 2019, 11:34 PM UTC
	avro_decimal	record_server_basic	s3://record-server-int...	-	Mon, Jul 1, 2019, 9:04 PM UTC
	avro_catalog_sales	record_server_basic	s3://record-server-int...	-	Mon, Jul 1, 2019, 9:02 PM UTC
	avro	record_server_basic	s3://record-server-int...	avro	Wed, Jun 19, 2019, 5:02 PM UTC
	parquet	record_server_basic	s3://record-server-int...	parquet	Wed, Jun 19, 2019, 4:56 PM UTC
	orc	record_server_basic	s3://record-server-int...	orc	Wed, Jun 19, 2019, 4:56 PM UTC
	json	record_server_basic	s3://record-server-int...	json	Wed, Jun 19, 2019, 4:42 PM UTC



(PNG Image, 300 x 236 pixels) X +



https://ec2-35-170-74-144.compute-1.amazonaws.com:8442/gateway/default/zeppelin/#/



New recording



Spark SQL - Zeppelin
All Applications
+

Back
Forward
Home
ec2-35-170-74-144.compute-1.amazonaws.com:8088/cluster
...
Star
Minimize
Maximize
Close



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores
3	0	1	2	1	1.38 GB	12 GB	0 B	1

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Recovering Nodes
1	0	0	0	0	0

Scheduler Metrics

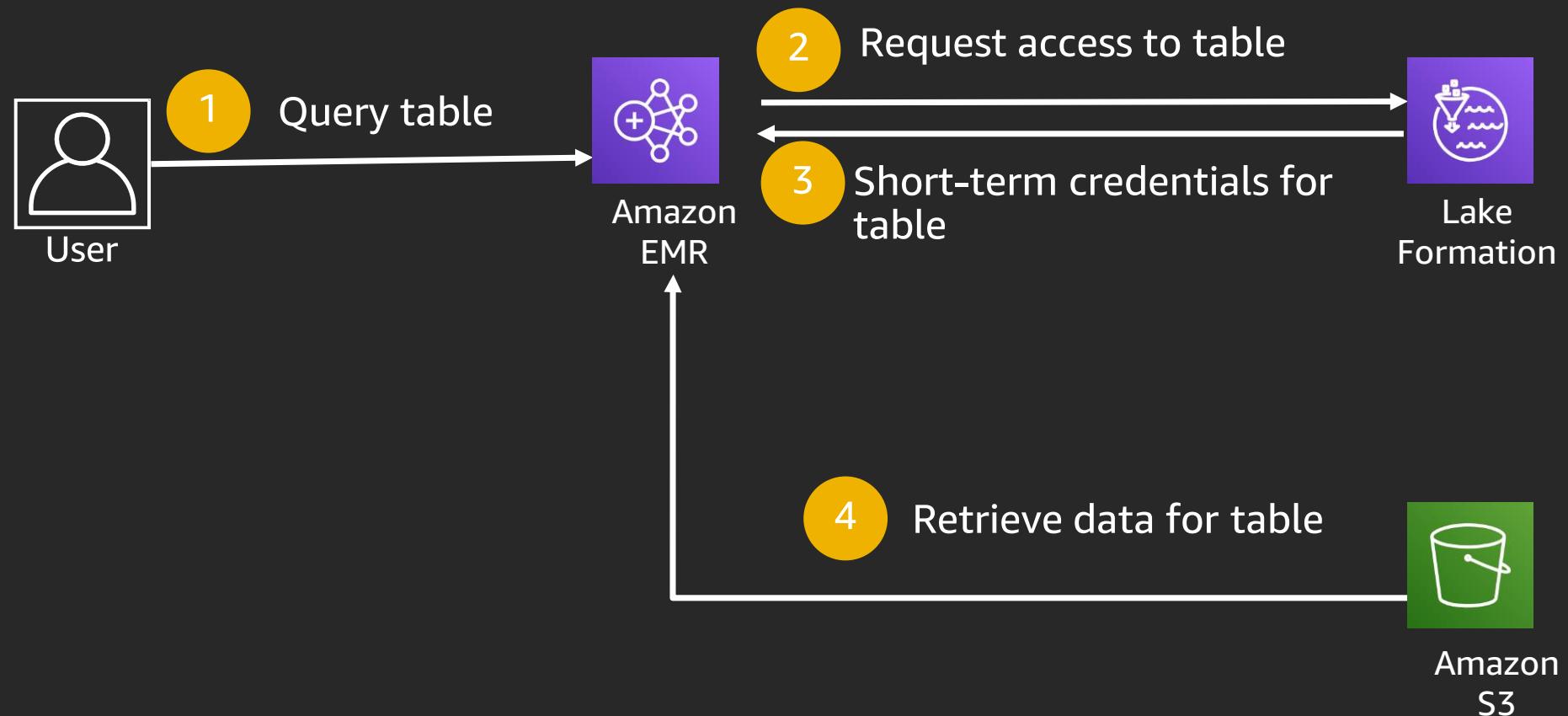
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	
Capacity Scheduler	[MEMORY]	<memory:32, vCores:1>	<memory:12288, vCores:8>	0

Show 20 entries

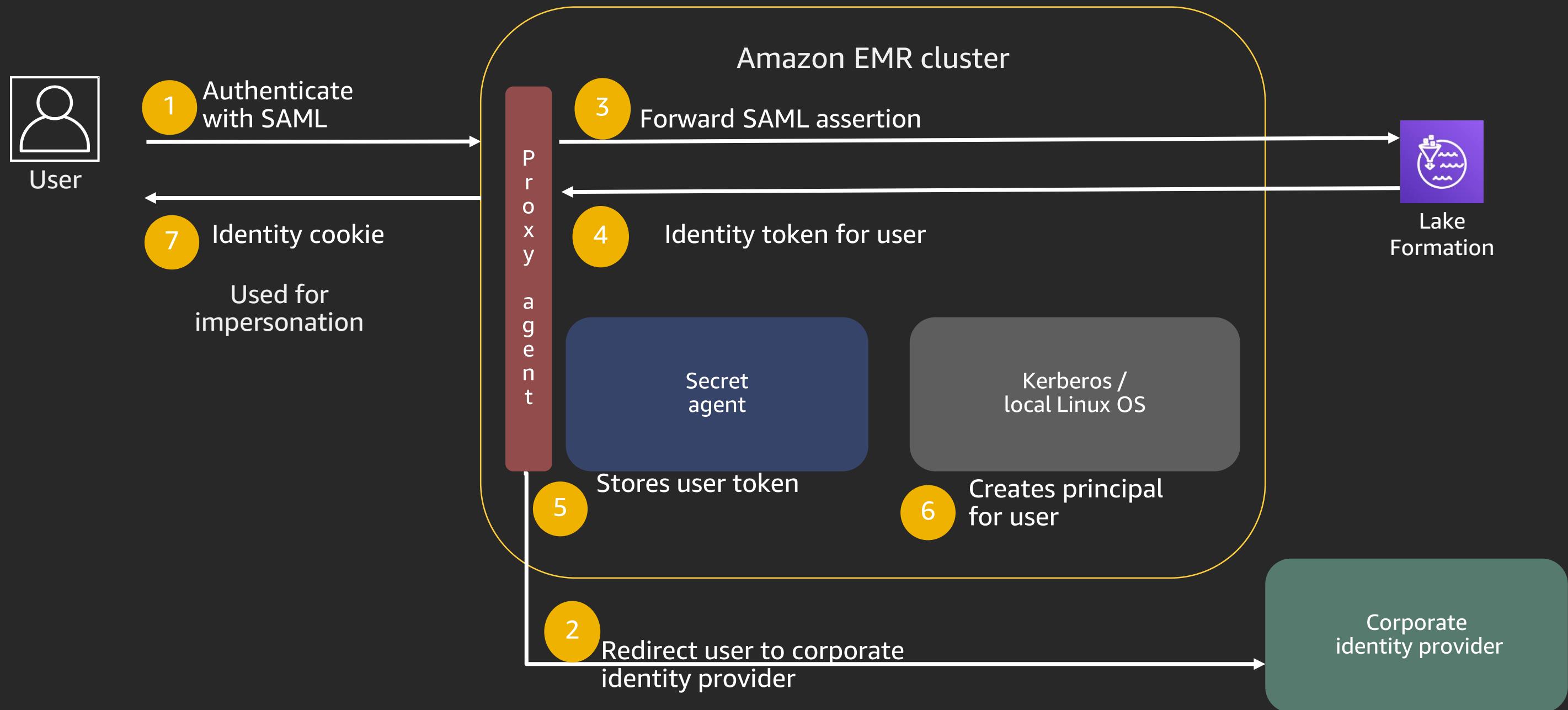
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	% of Queue
application_1564091049305_0003	paul	livy-session-2	SPARK	default	0	Fri Jul 26 08:33:56 -0700 2019	N/A	RUNNING	UNDEFINED	1	1	1408	11.5
application_1564091049305_0002	bob	livy-session-1	SPARK	default	0	Fri Jul 26 07:09:21 -0700 2019	Fri Jul 26 08:16:33 -0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0
application_1564091049305_0001	bob	livy-session-0	SPARK	default	0	Thu Jul 25 15:01:06 -0700 2019	Thu Jul 25 16:07:32 -0700 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0

Showing 1 to 3 of 3 entries

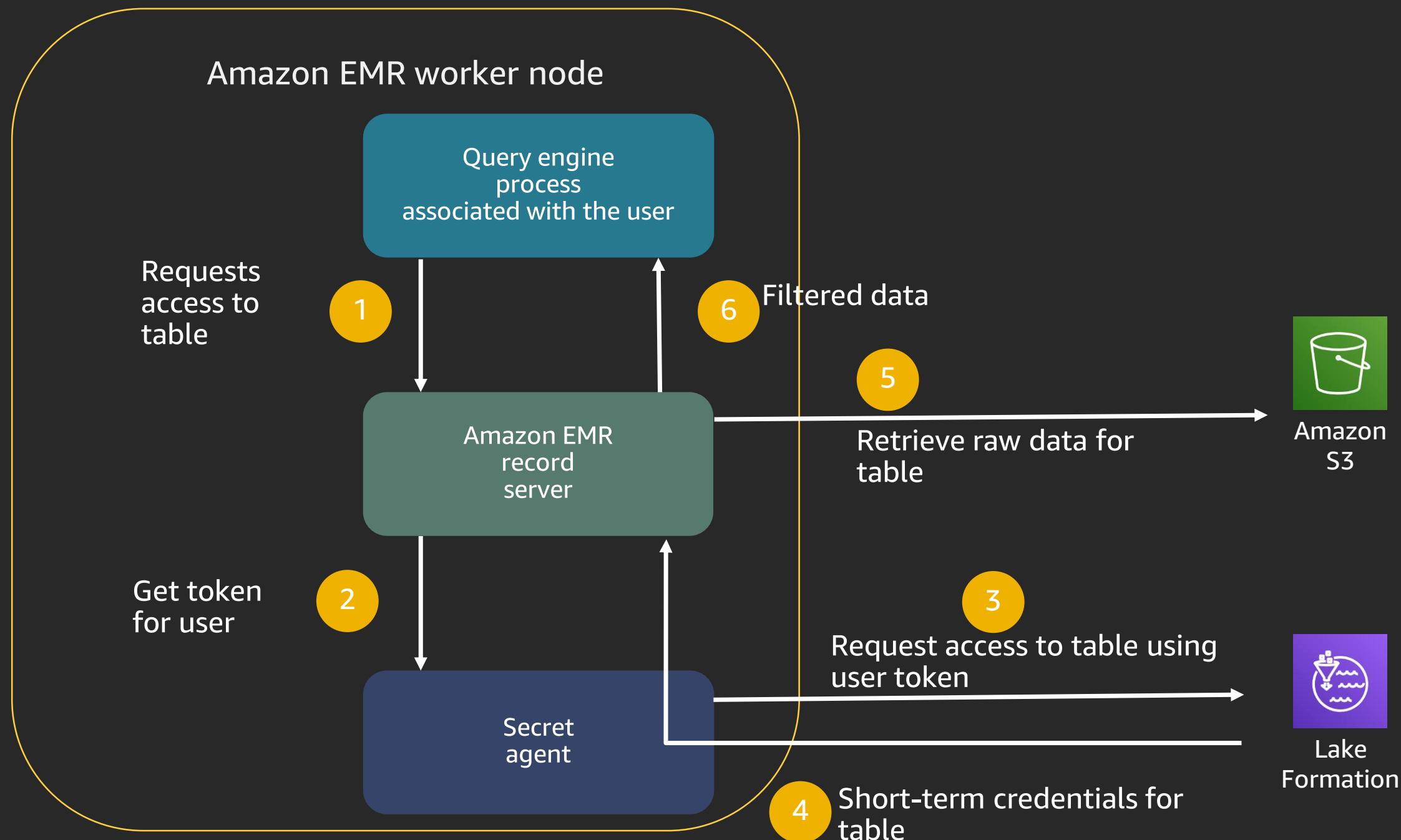
Query execution overview



Amazon EMR authentication



Query execution under the hood



Amazon EMR: Supported applications

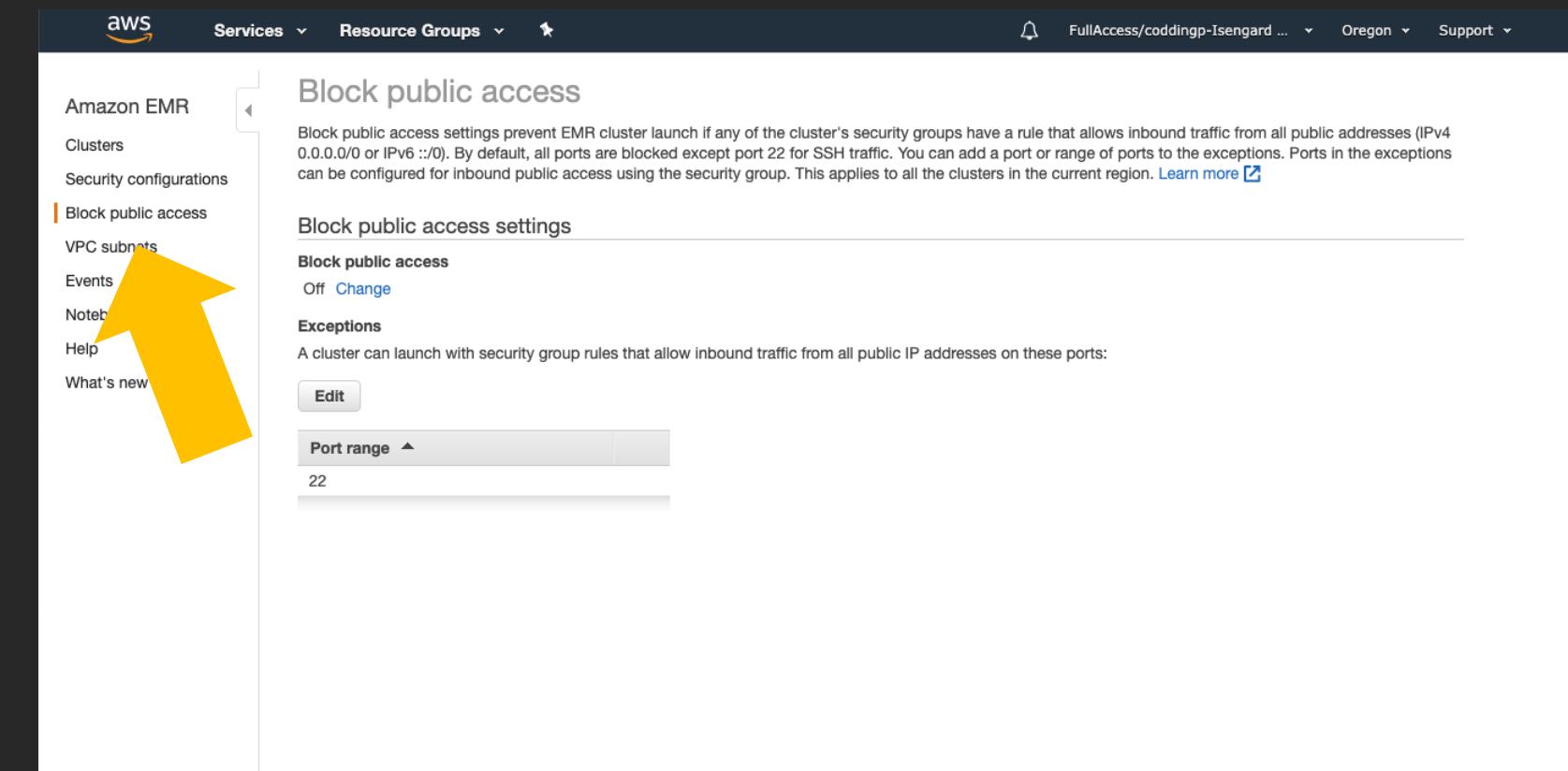
- AWS Glue Data Catalog
- Identity providers with support for SAML
- Applications
 - Spark SQL
 - Amazon EMR notebooks and Zeppelin with Livy

6

Block public access to open ports

Block unintended network exposure

- Prevent users from launching clusters with security groups that allow access from any IP address (IPv4 0.0.0.0/0 or IPv6 ::/0)
- Policy set at the account level, with region-specific configuration
- Allows exceptions to for public access to a single port or port range



7

Native EBS encryption



Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

S3 encryption

 Enable at-rest encryption for EMRFS data in Amazon S3

Amazon S3 encryption works with EMR File System (EMRFS) objects read from and written to Amazon S3. Specify server-side encryption (SSE) or client-side encryption (CSE). [Learn more](#)

Local disk encryption

 Enable at-rest encryption for local disks

Amazon EC2 instance store volumes and the attached Amazon Elastic Block Store (EBS) storage volumes are encrypted using Linux Unified Key Setup (LUKS). Alternatively, when using AWS KMS as your key provider, you can choose to turn on EBS encryption to encrypt EBS root device and storage volumes. AWS KMS customer master keys (CMKs) require additional permissions for EBS encryption. [Learn more](#)

Key provider type AWS KMS customer master key

- Encrypt EBS volumes with EBS encryption
Recommended for compliance with AWS Config
Managed rules. Requires adding IAM role for EMR to the KMS CMK.
- Encrypt EBS volumes with LUKS encryption

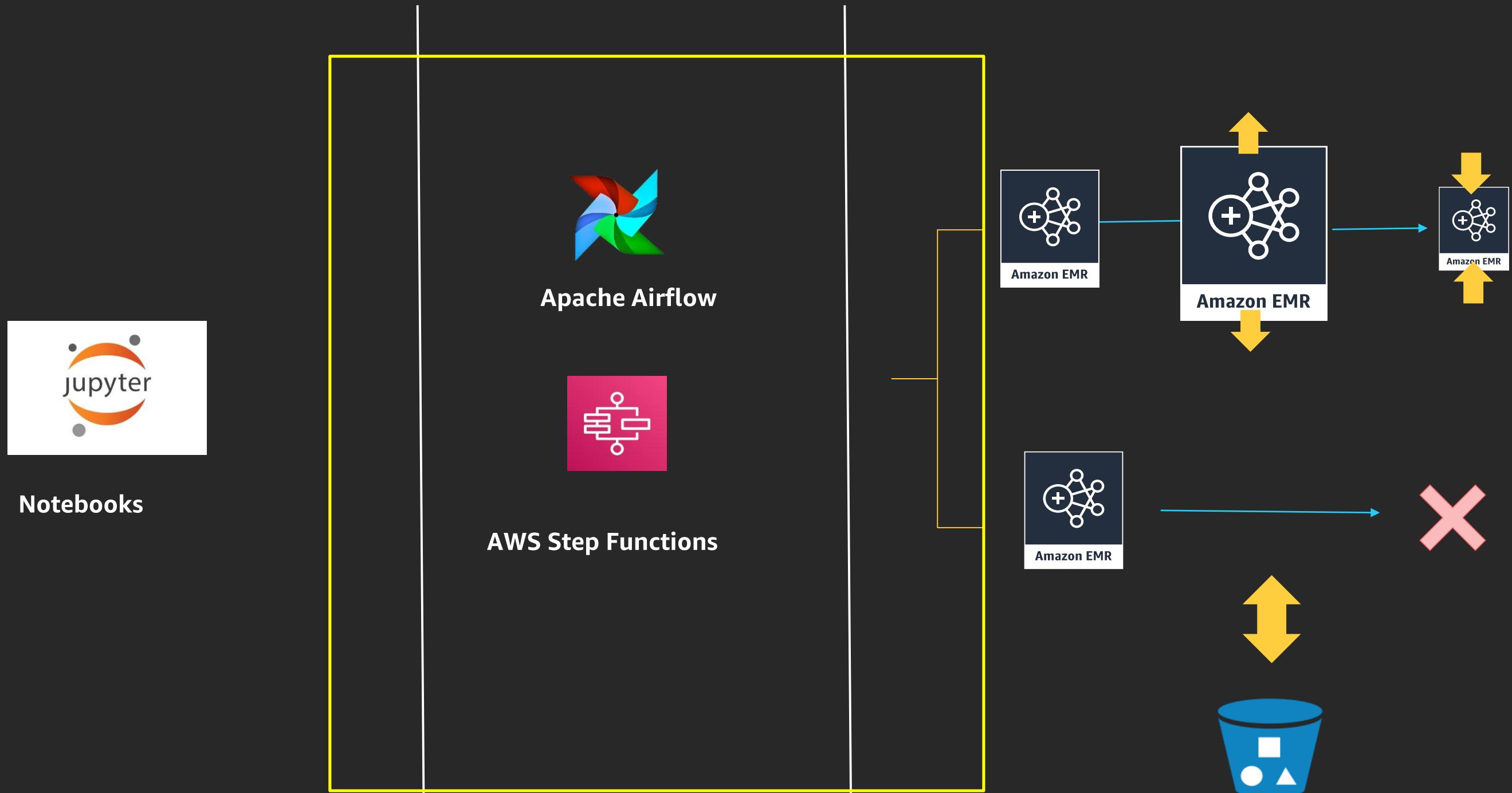
Data in transit encryption

 Enable in-transit encryption

Transport Layer Security (TLS) is essential for encrypting information that is exchanged on the internet. Turn on open-source TLS encryption features for in-transit data and choose a certificate provider type. [Learn more](#)



Simplified view of data engineering platforms



8

Integration with AWS Step Functions

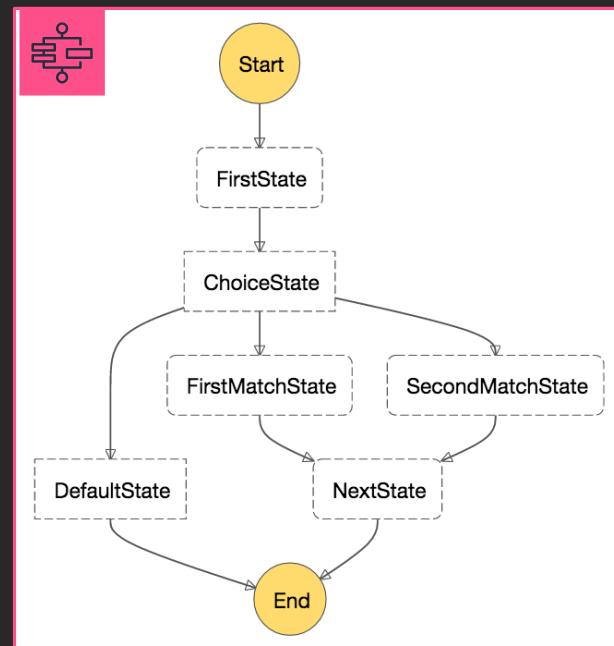
AWS Step Functions: Visual workflows

Define in JSON or Python

```
Code

1 { "Comment": "An AWL example using a choice state.",
2   "StartAt": "FirstState",
3   "States": {
4     "FirstState": {
5       "Type": "Task",
6       "Resource": "arn:aws:lambda:REGION:ACCOUNT_ID:function:FUNCTION_NAME",
7       "Next": "ChoiceState"
8     },
9     "ChoiceState": {
10       "Type" : "Choice",
11       "Choices": [
12         {
13           "When": "选择了第一个选项", "Next": "FirstState"
14         }
15       ]
16     }
17   }
18 }
```

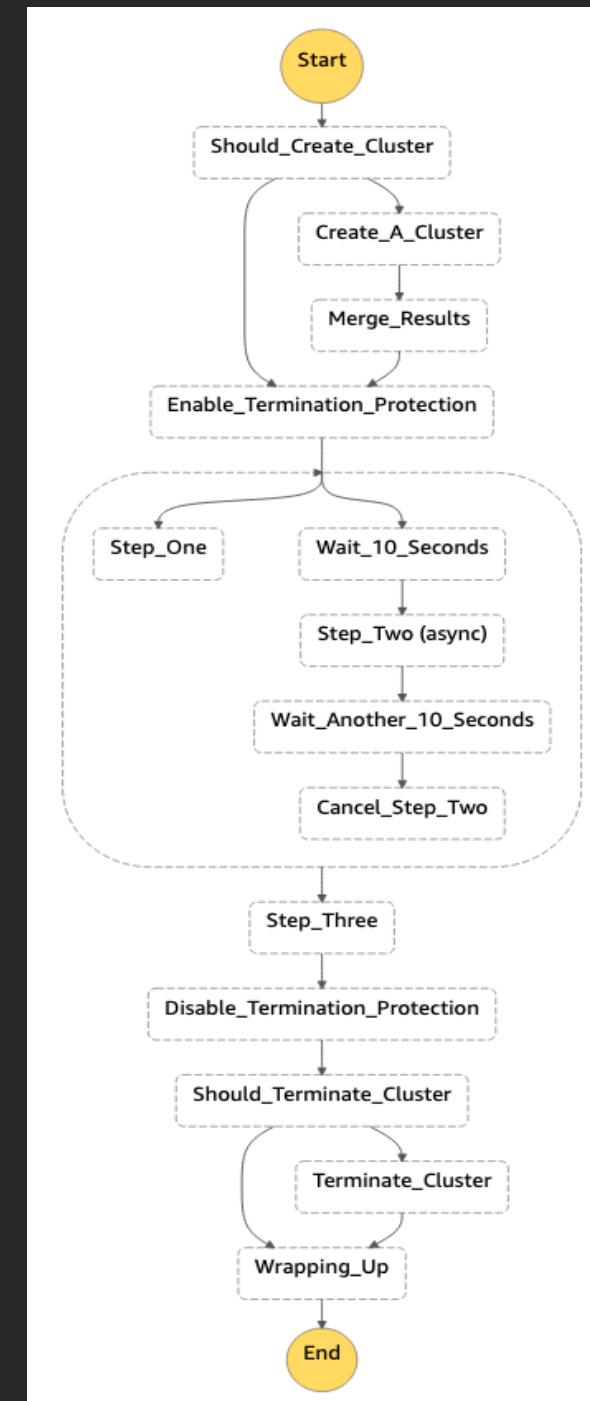
Visualize in the console



Monitor executions

Automate workflows using AWS Step Functions

1. Create, scale, and modify clusters
2. Add, cancel, or run steps in parallel
3. Synchronous and asynchronous steps
4. Handle exceptions/failures
5. Scale clusters up or down
6. Reuse clusters
7. Terminate them



Visual workflow

Code Step details

Name: Create_A_Cluster Type: Task

Status: In Progress

Resource: EMR Cluster j-24OPRHBDAW059

Name	ID	Status	Creation time (UTC+0)	Elapsed time
WorkflowCluster	j-2GV3YEDSFDF52	Waiting Cluster ready	2019-11-13 12:34 (UTC+0)	10 minutes

Summary

Master public ec2-34-244-148-183.eu-west-1.compute.amazonaws.com

Termination protection: Off Change

Tags: -- View All / Edit

Hardware

Master: Running 1 units

Core: Running 1 units

Task: --

Name	Status	Start time (UTC+0)	Elapsed time
The third step	Completed	2019-11-13 12:42 (UTC+0)	52 seconds
The second step	Cancelled	--	--
The first step	Completed	2019-11-13 12:40 (UTC+0)	56 seconds

Simplified view of data engineering platforms



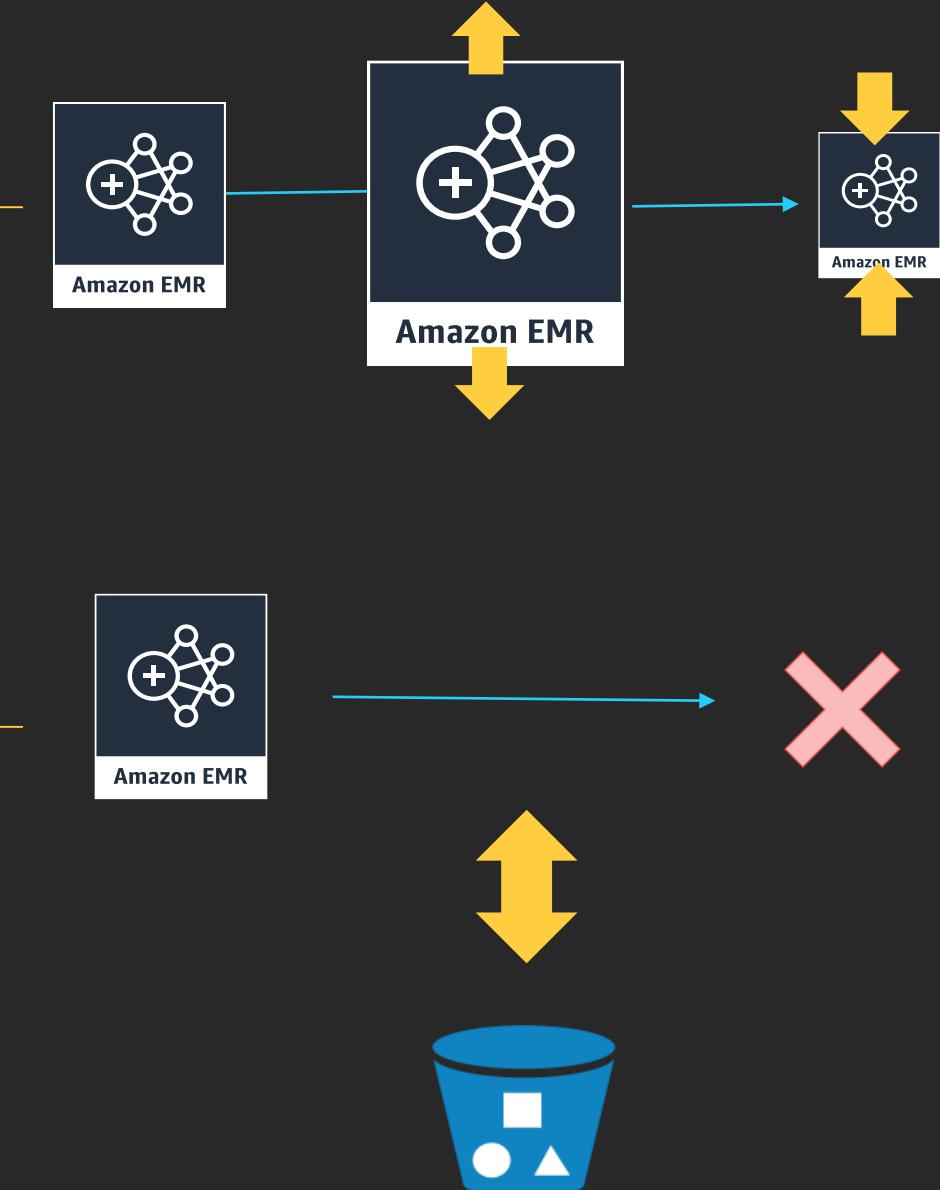
Notebooks



Apache Airflow



AWS Step Functions



Managed notebooks with repository integration and notebook-scoped libraries

Fully managed notebooks for simplified experience

Amazon EMR

- Clusters
- Security configurations
- Block public access
- VPC subnets
- Events
- Notebooks**
- Git repositories

Help

What's new

Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. [Learn more](#)

Notebook name* Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description

256 characters max.

Cluster* Choose an existing cluster [Choose](#) Create a cluster [i](#)

Security groups Use default security groups [i](#) Choose security groups

AWS service role* [i](#)

Notebook location* Choose an S3 location where files for this notebook are saved.

Use the default S3 location
s3://aws-emr-resources-418620260174-us-east-1/notebooks/

Choose an existing S3 location in us-east-1

► **Git repository** Link to a Git repository

► **Tags** [i](#)

* Required

[Cancel](#) [Create notebook](#)

Open in Jupyter or JupyterLab

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Git repositories

Help

What's new

Notebook: sinhaarnotebook Ready Notebook is ready to run jobs on cluster j-31R9ZWRX1HLNQ.

Open in JupyterLab **Open in Jupyter** **Stop** **Delete**

Notebook

Notebook ID: e-1GQ9CX8753M70200ZFUDCY42N

Description: --

Last modified: 6 seconds ago info

Last modified by: ...assumed-role/Admin/sinhaar-Isengard info

Created on: 2019-11-29 08:11 (UTC-8)

Created by: ...assumed-role/Admin/sinhaar-Isengard info

Service IAM role: EMR_Notebooks_DefaultRole edit

Notebook tags: creatorUserId = AROAI6TCY2QJJPZ43WBK:sinhaar-Isengard [View All / Edit](#)

Notebook location: s3://aws-emr-resources-418620260174-us-east-1/notebooks/ file

Cluster

Cluster: DO NOT TERMINATE jnallapa-ats15-test

Cluster Id: j-31R9ZWRX1HLNQ

Cluster status: Waiting Cluster ready after last step completed.

Cluster tags: --

Step logs: s3://aws-logs-418620260174-us-east-1/elasticmapreduce/ file

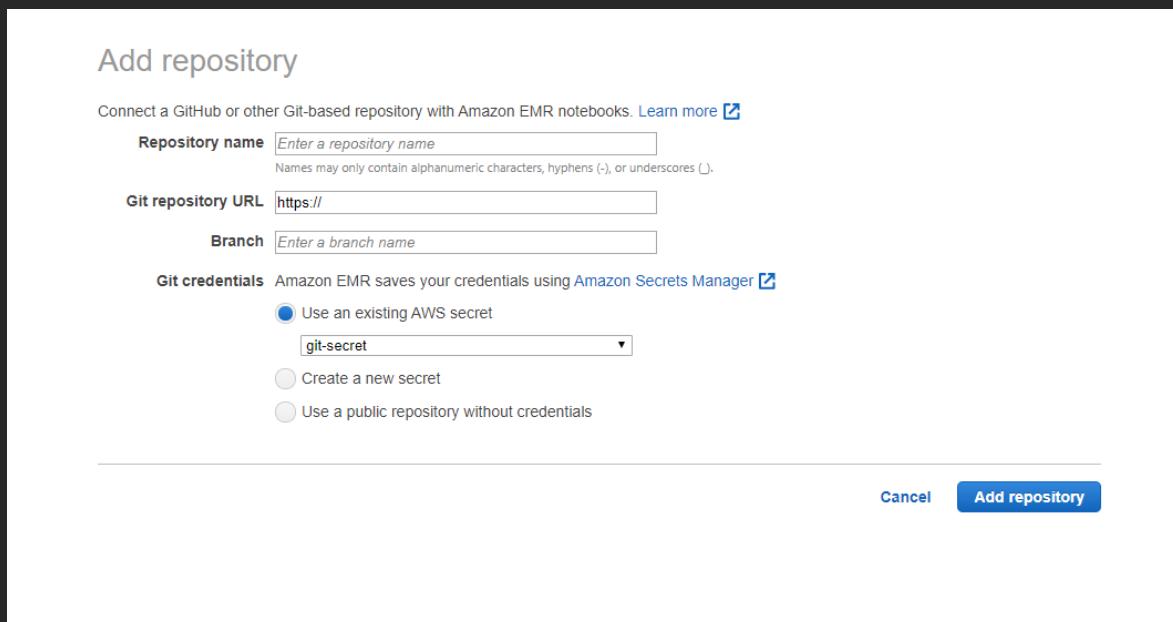
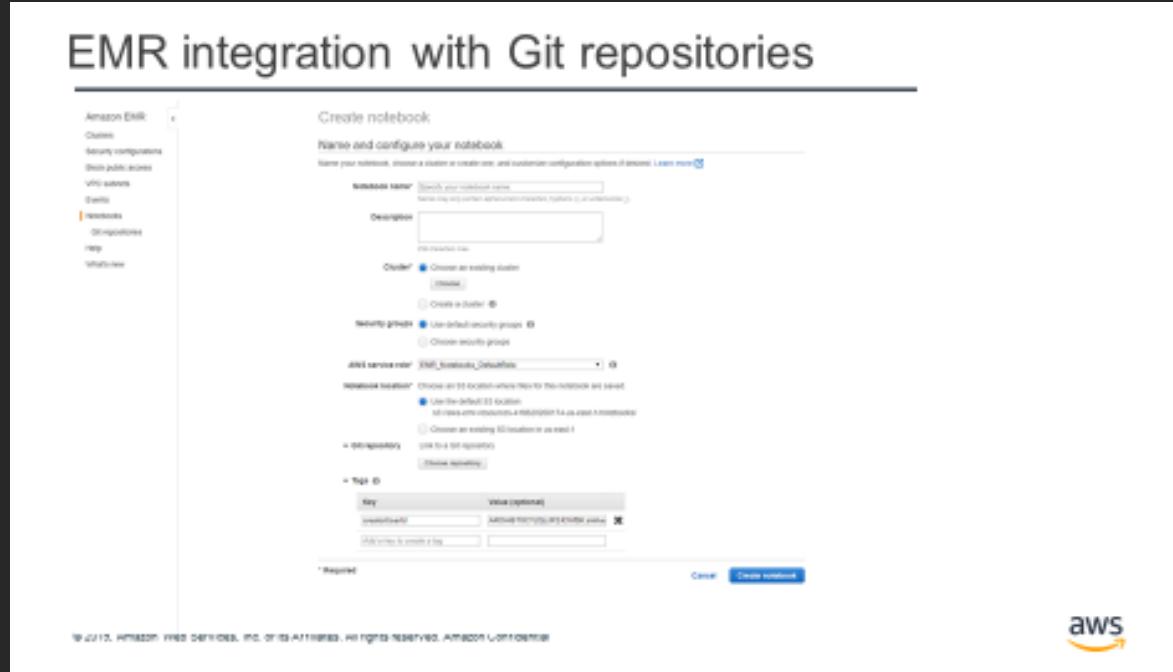
Git repositories

The repository can be linked to a notebook once the notebook is ready. Make sure your cluster, service role and security groups have the required settings. [Learn more](#)

Link new repository **Unlink repository**

Repository name	URL	Branch	Link status	Failure reason
-----------------	-----	--------	-------------	----------------

Associate notebooks with repositories



- Associate notebooks with repositories
- Collaborate and share notebooks
- Integrate with pipelines
- Integration with AWS Secrets Manager

Install notebook-scoped libraries management

- Notebook-scoped libraries are installed in a virtual Python environment
- Scoped to the current notebook session
- Available only during the notebook session
- Libraries are deleted after the session ends

```
sc.install_pypi_package("celery")
import celery sc.range(1,10000,1,100).map(lambda x:celery.__version__).collect()
```

```
sc.install_pypi_package("arrow==0.14.0", "https://pypi.org/simple")
sc.uninstall_package("arrow")
```

10

Off-cluster persistent Spark History Service for simplified debugging

Off-cluster persistent Spark History Service

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Git repositories

Help

What's new

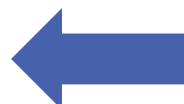
Cluster: DO NOT TERMINATE Parag Test Terminated Terminated by user request

Clone Terminate AWS CLI export

Summary Application history Monitoring Hardware Configurations Events Steps Bootstrap actions

Connections: --

Master public DNS: ec2-3-90-84-213.compute-1.amazonaws.com SSH

History service: Spark history server UI (SSH tunneling not required) 

Tags: --

Summary

ID: j-3ODPTHGROJPUA Release label: emr-5.27.0

Creation date: 2019-10-25 13:30 (UTC-8) Hadoop distribution: Amazon 2.8.5

End date: 2019-11-24 18:17 (UTC-8) Applications: Spark 2.4.4, Livy 0.6.0, Hive 2.3.5, Zeppelin 0.8.1, JupyterHub 1.0.0

Elapsed time: 4 weeks Log URI: s3://aws-logs-418620260174-us-east-1/elasticmapreduce/

After last step Cluster waits completes:

Termination Off protection: EMRFS consistent view: Disabled

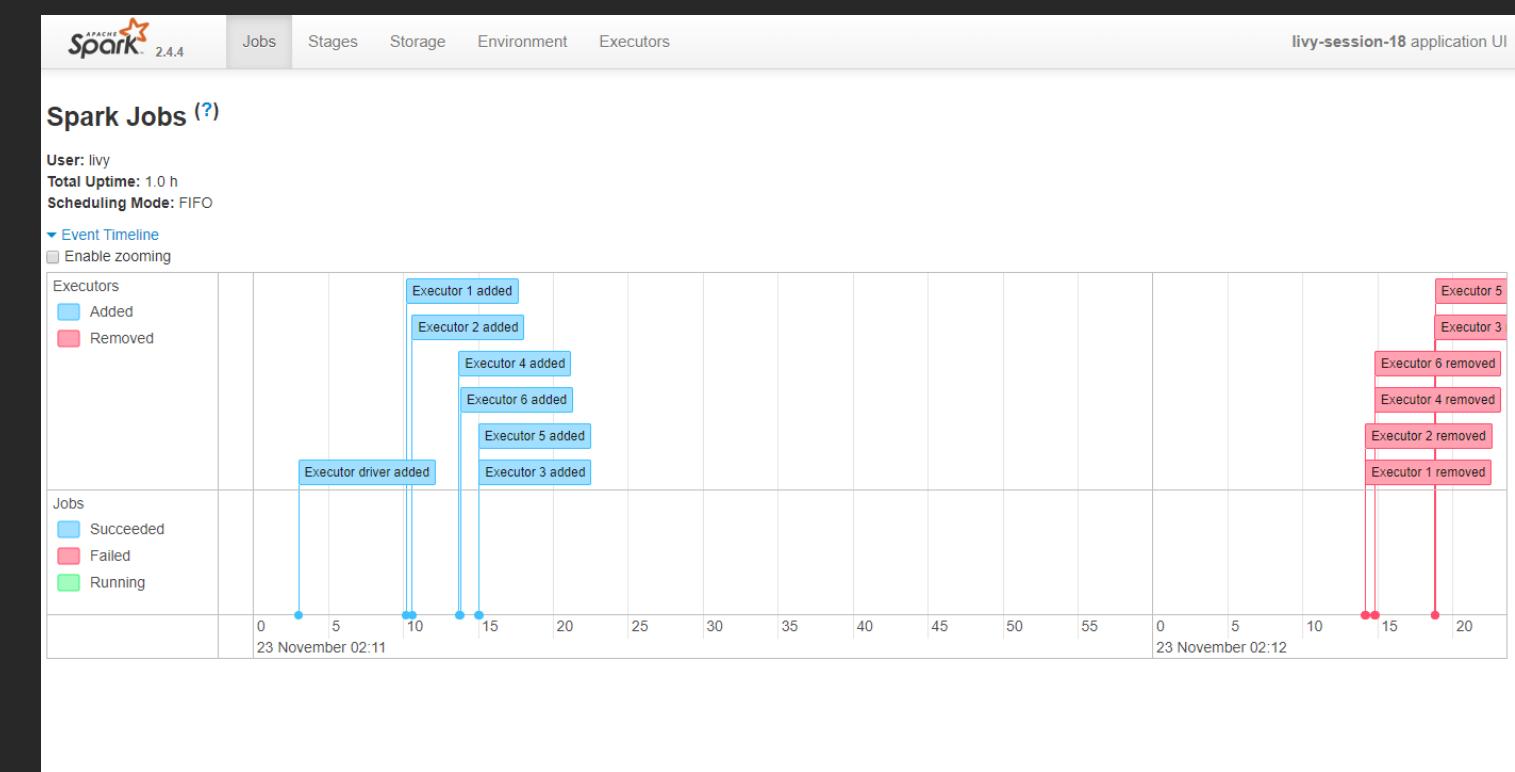
Custom AMI ID: --

Network and hardware

Availability zone: us-east-1e

Security and access

Key name: paragnc.emr.dev.new.us-east-1



Demo

Learn big data with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills



New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, and Amazon Redshift



Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs



Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/pathsspecialty/

Thank you!

Abhishek Sinha

sinhaar@amazon.com

Mert Hocanin

hocanint@amazon.com



Please complete the session
survey in the mobile app.