AWS re:Invent

ANT308

Building Serverless Analytics Pipelines with AWS Glue

Mehul A. Shah
Senior Software Development Manager
AWS Glue

Arup Ray
Vice President, Engineering
Realtor.com

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Today's Agenda

AWS Glue overview

What's new: improvements and features

Realtor.com – customer success

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue overview

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

A year ago …

Fully-managed, serverless extract-transform-load (ETL) service

for developers, built by developers

1000s of customers and jobs

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

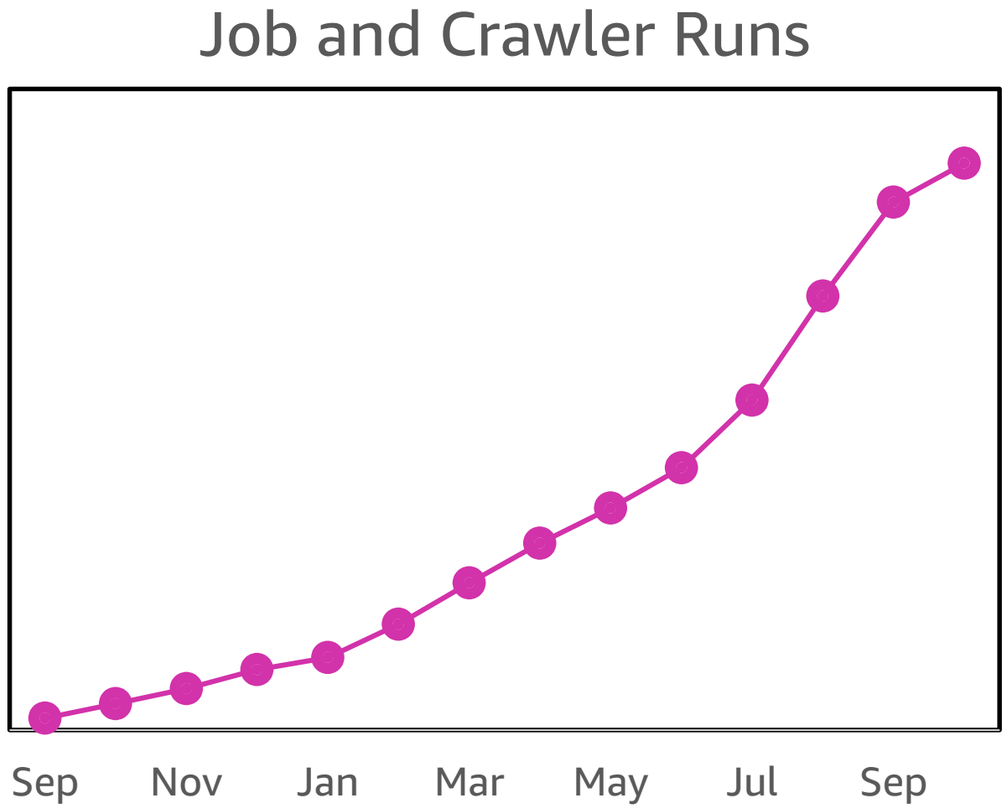
A year ago …

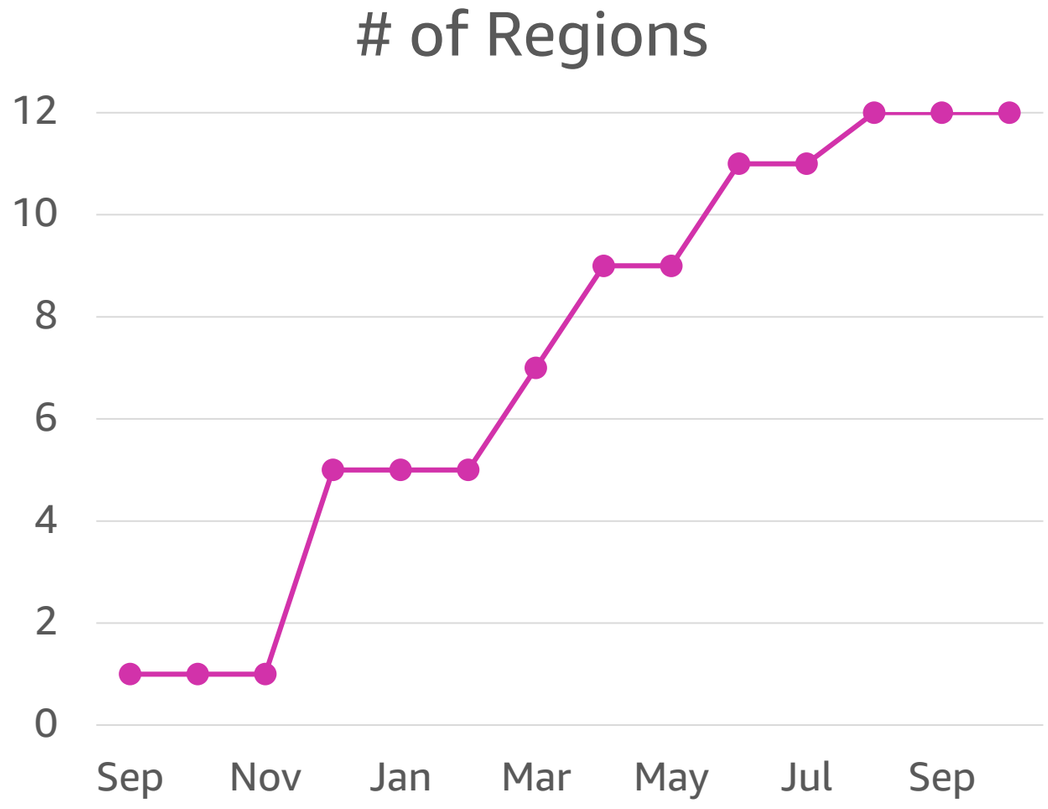
© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Select AWS Glue customers …

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

You helped us grow …





© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

We've been busy…

Access policies for AWS Glue Data Catalog

Amazon SageMaker notebooks

Encryption at rest

Canada central region

Crawler combine compatible schemas

ETL job metrics

Amazon DynamoDB integration

Job delay notification

London region

Seoul region

Crawler merge new columns

Mumbai region

Support Apache Spark 2.2.1

ETL job timeout

Sydney region

Frankfurt region

Readers support JSONPath expressions

New Job Events Types

Support for Scala scripts

Tokyo region

Singapore region

Crawler CWE notifications

XML support

AWS CloudTrail support

Amazon CloudFormation templates

Crawler exclusion patterns

Per-second billing

DynamicFrame Filter and Map

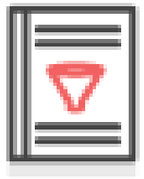
GDPR, HIPAA, and BAA compliance

Ireland, Oregon, Ohio region

30+ major features

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue components



Data Catalog

Discover

Automatic crawling

Apache Hive Metastore compatible

Integrated with AWS analytic services

Serverless Jobs

Develop

Apache Spark core

Python and Scala

Auto-generates ETL code

Orchestration

Deploy

Flexible scheduling

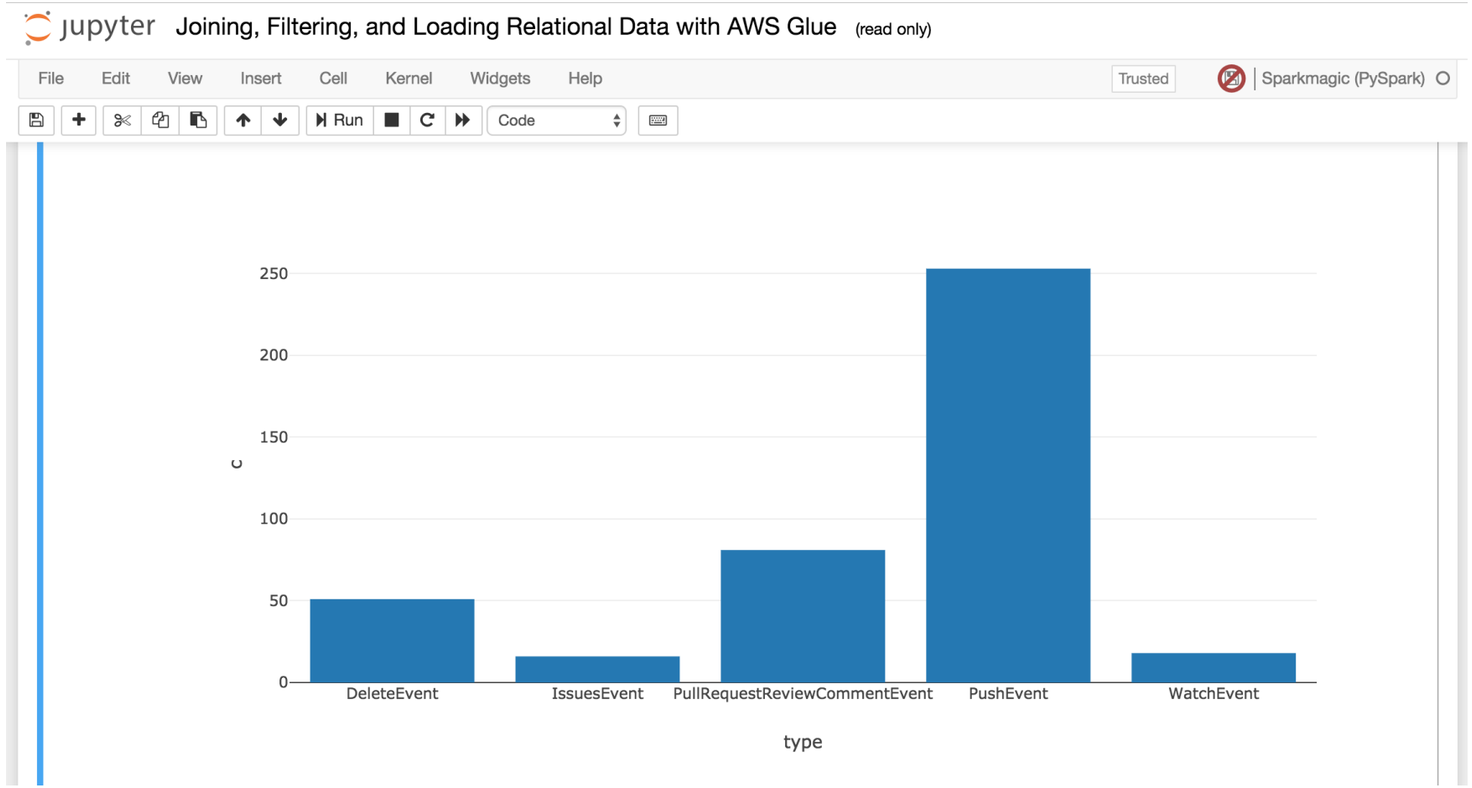
Monitoring and alerting

External integrations

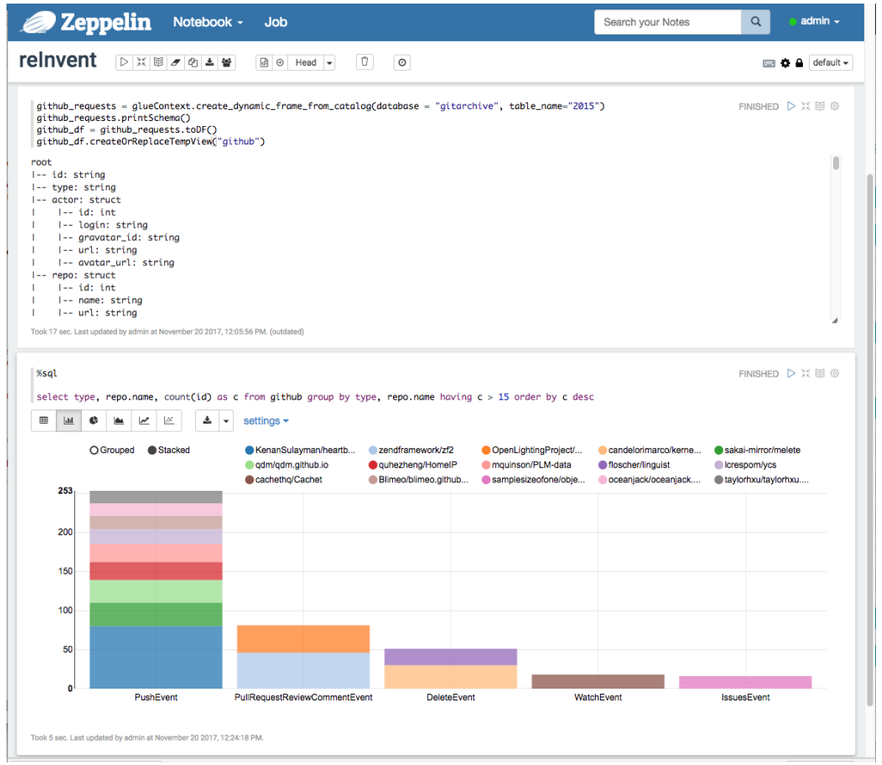
© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond ETL: data science and exploration

Amazon SageMaker integration



Other notebooks



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What you're saying …

"Beeswax uses Amazon S3 and AWS Glue Data Catalog to build a highly reliable **data lake** that is fully managed by AWS. Our platform leverages the … integration with **Amazon EMR** in **Hive** and **SparkSQL** …"

Ram Kumar Regnaswamy, CTO, Beeswax

"… ETL-ing data from our **data lake** to our **Redshift** warehouse is just one of use case examples of AWS Glue. … Being cost-effective is essential. … AWS Glue has enabled our small team of data engineers to run the whole data infrastructure in our company."

Umang Rustagi, Co-founder and COO, FinAccel"

"Glue ETL performed **200% faster** than traditional ETL tools with no operational overhead due to the serverless nature of Glue ETL and at a fraction of the cost."

Miki Hardisty, CTO, Jack-in-the-box

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What's new: improvements and features

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Innovation areas

Scalability and performance

Crawlers and ETL jobs

Profiling and debugging

ETL metrics

Interoperability

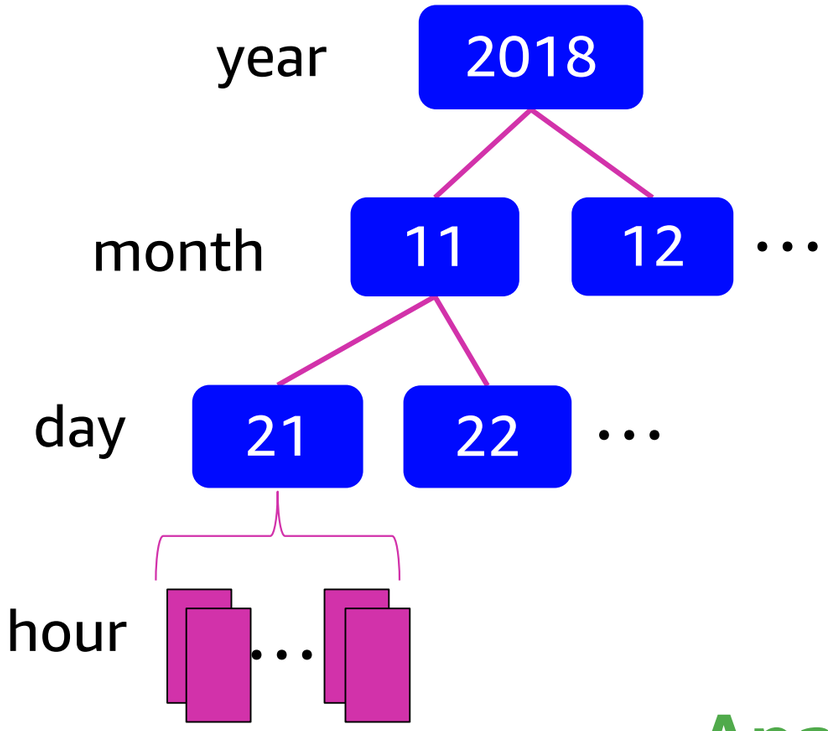
Orchestration features

Announcing a new job type: Python shell

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example data organization in:
Amazon Simple Storage Service (Amazon S3)

source Amazon S3 bucket

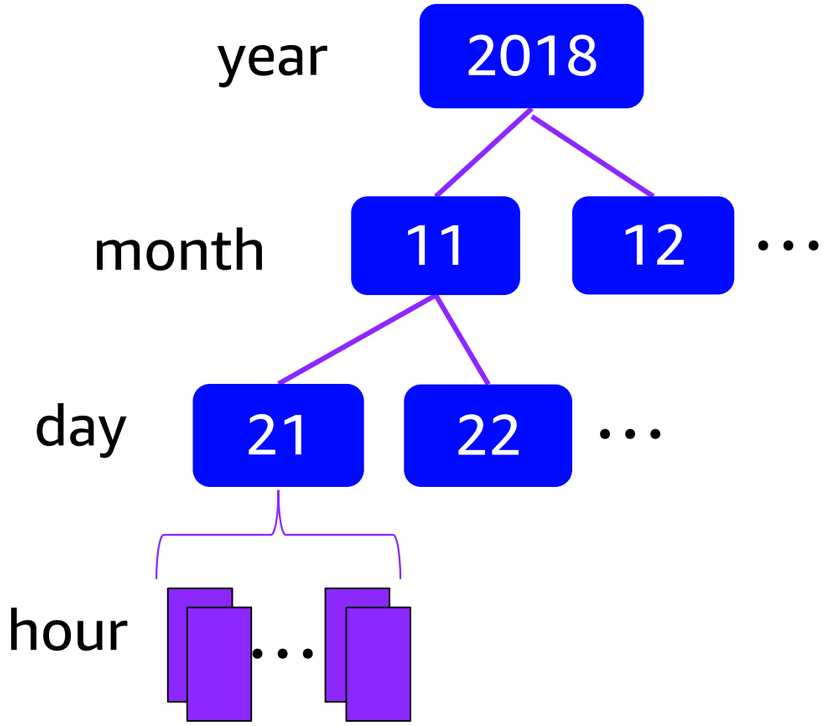


JSON

transform

Apache Hive-style partitions

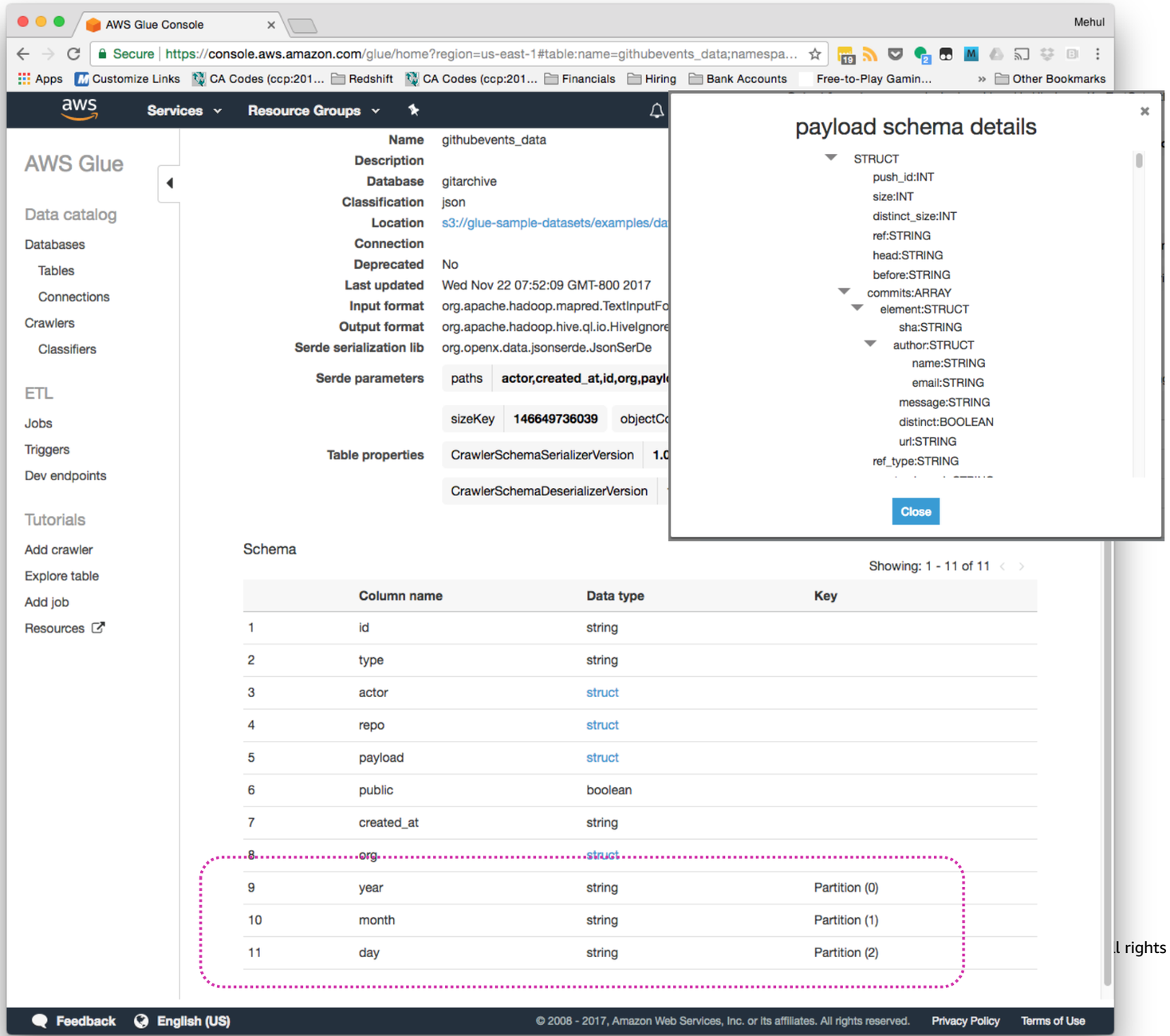
target Amazon S3 bucket



Parquet

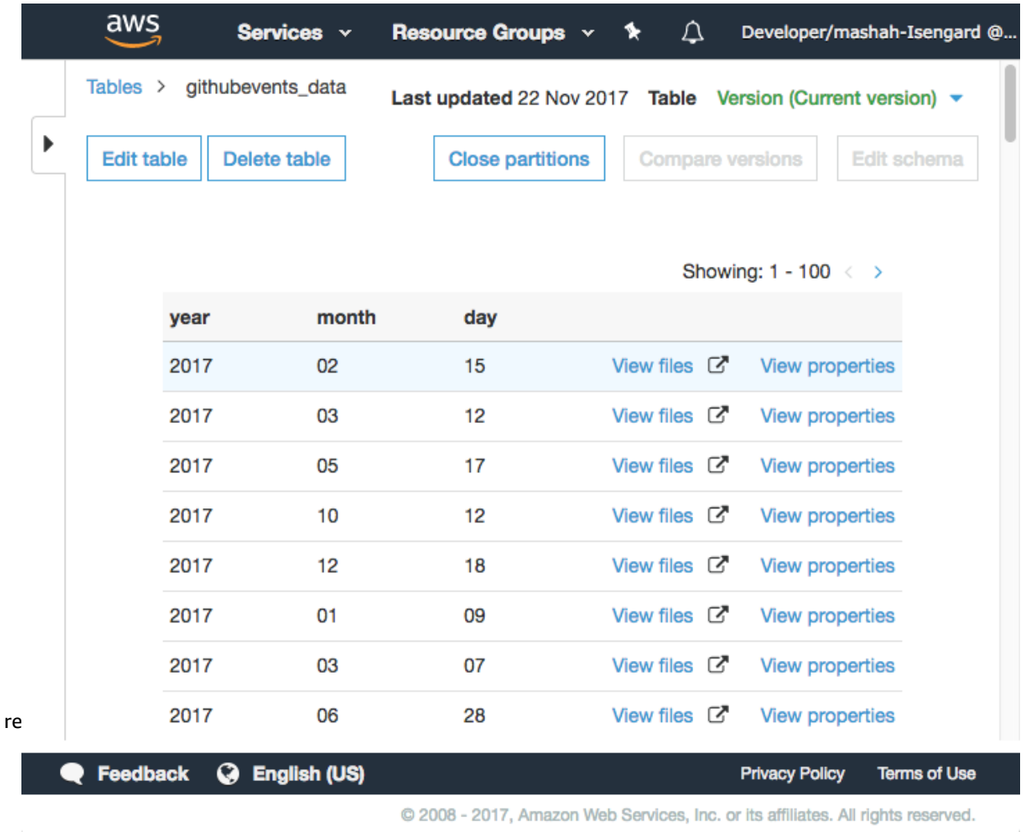
© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Crawler results …



Complex, nested fields

Groups files into
Apache Hive-style partitions



l rights re

Crawler performance

7x improvement YoY

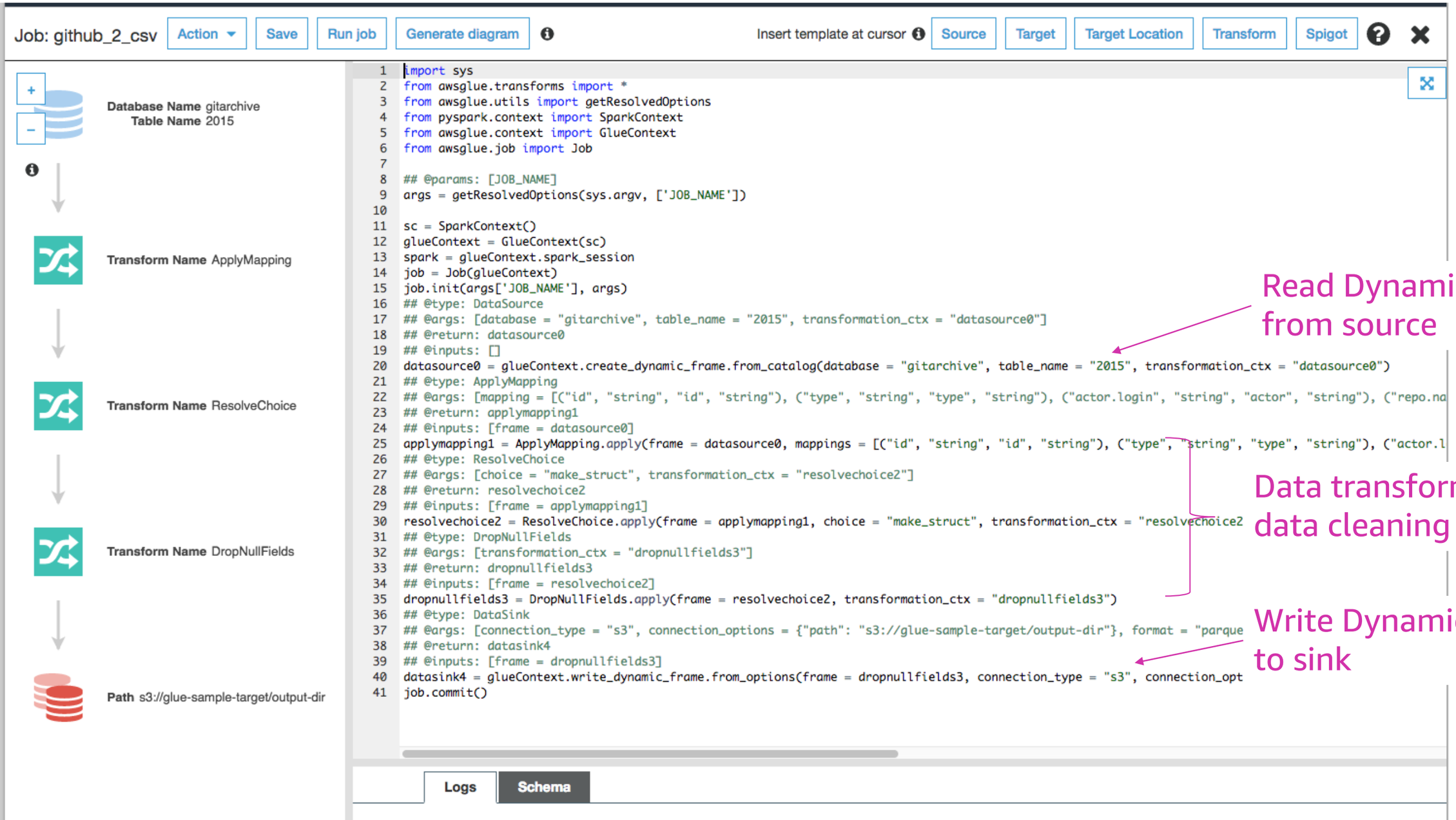
90M+ files per day

Millions of partitions

YMMV with partitions
size and **complexity**

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue ETL scripts

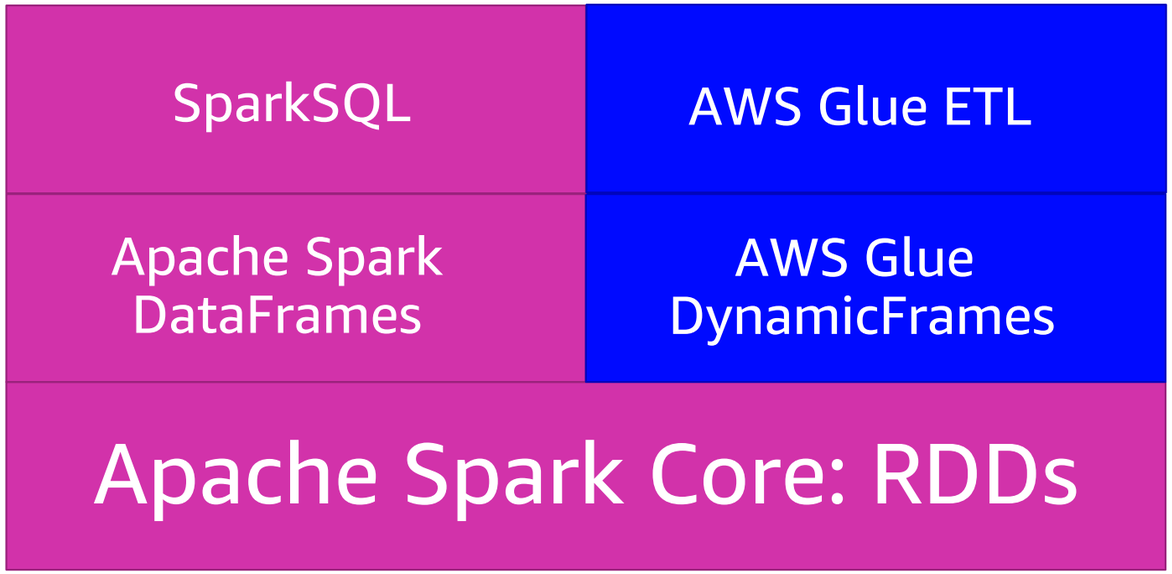


Read Dynamic Frame from source

Data transformation + data cleaning functions

Write Dynamic Frame to sink

Review: Apache Spark and AWS Glue ETL



Application

Data Structure

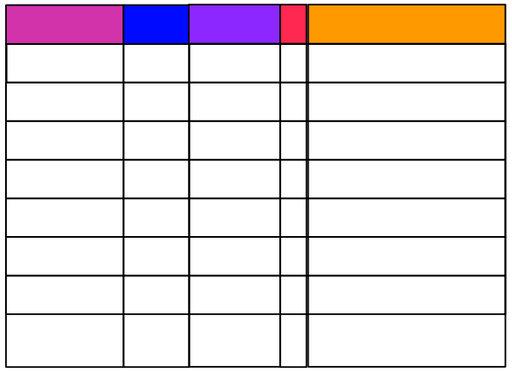
Execution

Apache Spark is a distributed data processing engine for complex analytics.

AWS Glue builds on the Apache Spark to offer ETL specific functionality.

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DataFrames and DynamicFrames



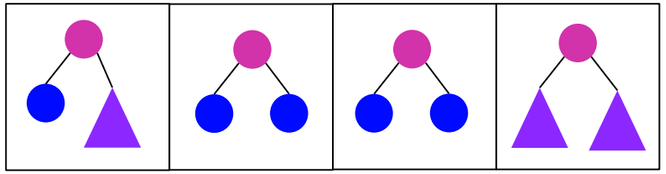
DataFrames

Core data structure for SparkSQL

Like structured tables

- Need schema up-front
- Each row has same structure

Suited for SQL-like analytics



DynamicFrames

Like DataFrames for ETL

Designed for processing **semi-structured** data, e.g. JSON, Avro, Apache logs …

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DynamicFrame performance

Raw conversion speed

4x improvement YoY

1TB in under 1.5 hours
with 10 DPU

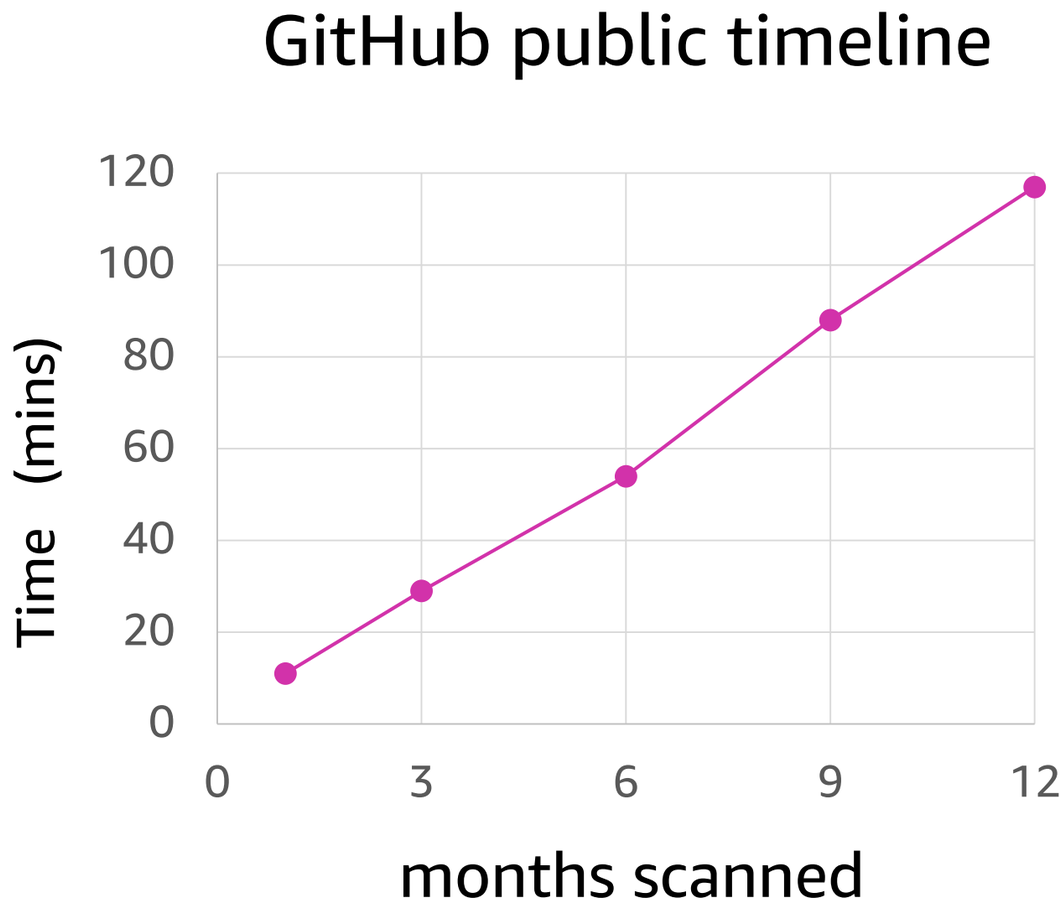
YMMV with **data format**
and script **complexity**

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DynamicFrame predicate pushdown

Example code

```
glueContext
  .getCatalogSource(
    database  = "gitarchive",
    tableName = "githubevents_data",
    pushDownPredicate =
      "(year=='2017' and month=='04')")
  .getDynamicFrame()
```



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Innovation areas

Scalability and performance

Crawlers and ETL jobs

Profiling and debugging

ETL metrics

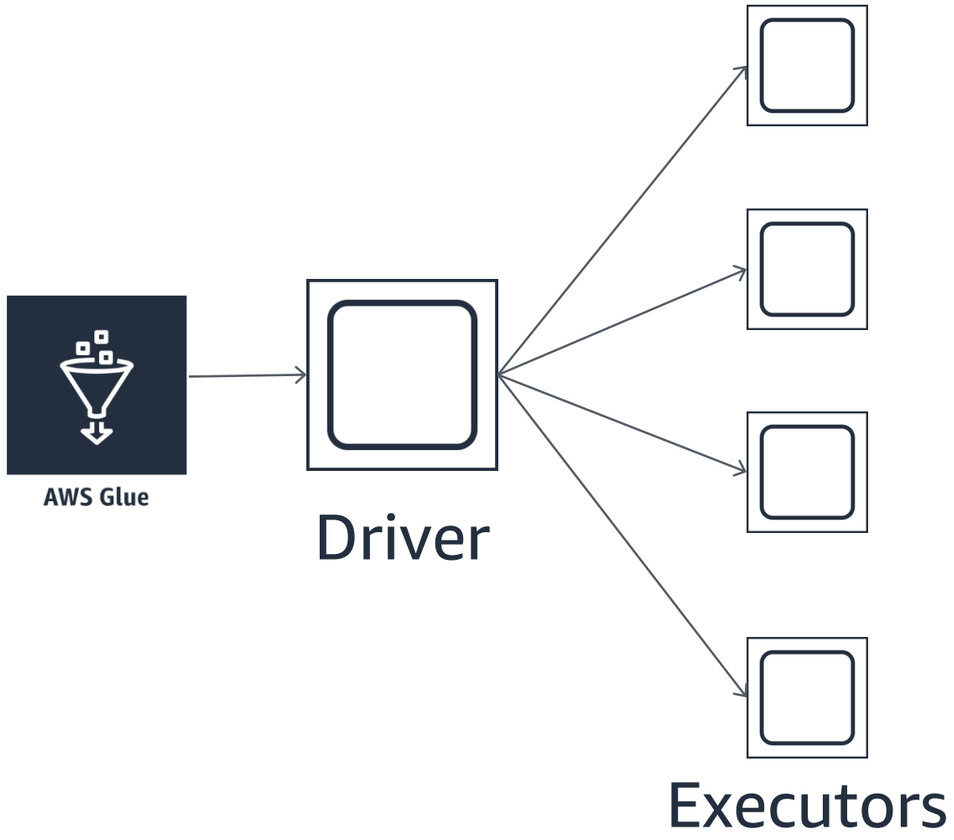
Interoperability

Orchestration features

Announcing a new job type: Python shell

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue execution model



Apache Spark and AWS Glue are *data parallel*. Data is divided into *partitions* (shards) that are processed concurrently.

Jobs are divided into *stages*

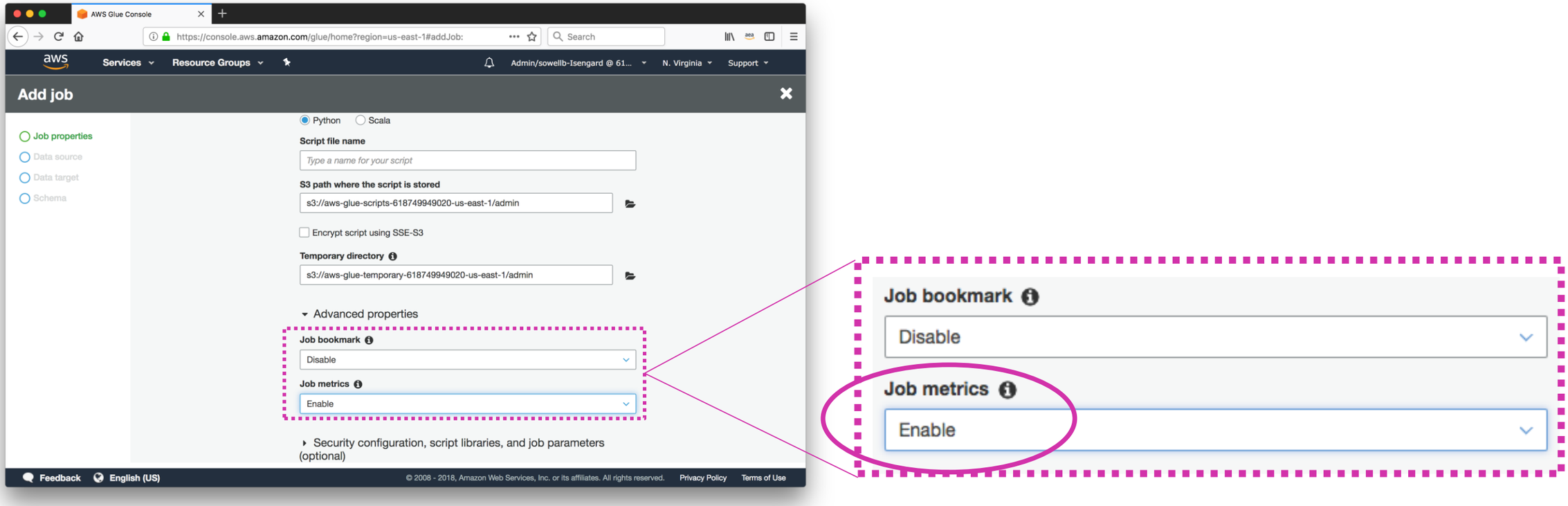
1 stage x 1 partition = 1 *task*

Driver schedules tasks on *executors.*
2 executors per DPU

Overall throughput is limited by the number of partitions (shards)

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue job metrics



- Metrics can be enabled in the AWS Command Line Interface (AWS CLI) and AWS SDK by passing `--enable-metrics` as a job parameter key.

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Profile jobs using ETL metrics

Derived from the underlying Apache Spark metrics

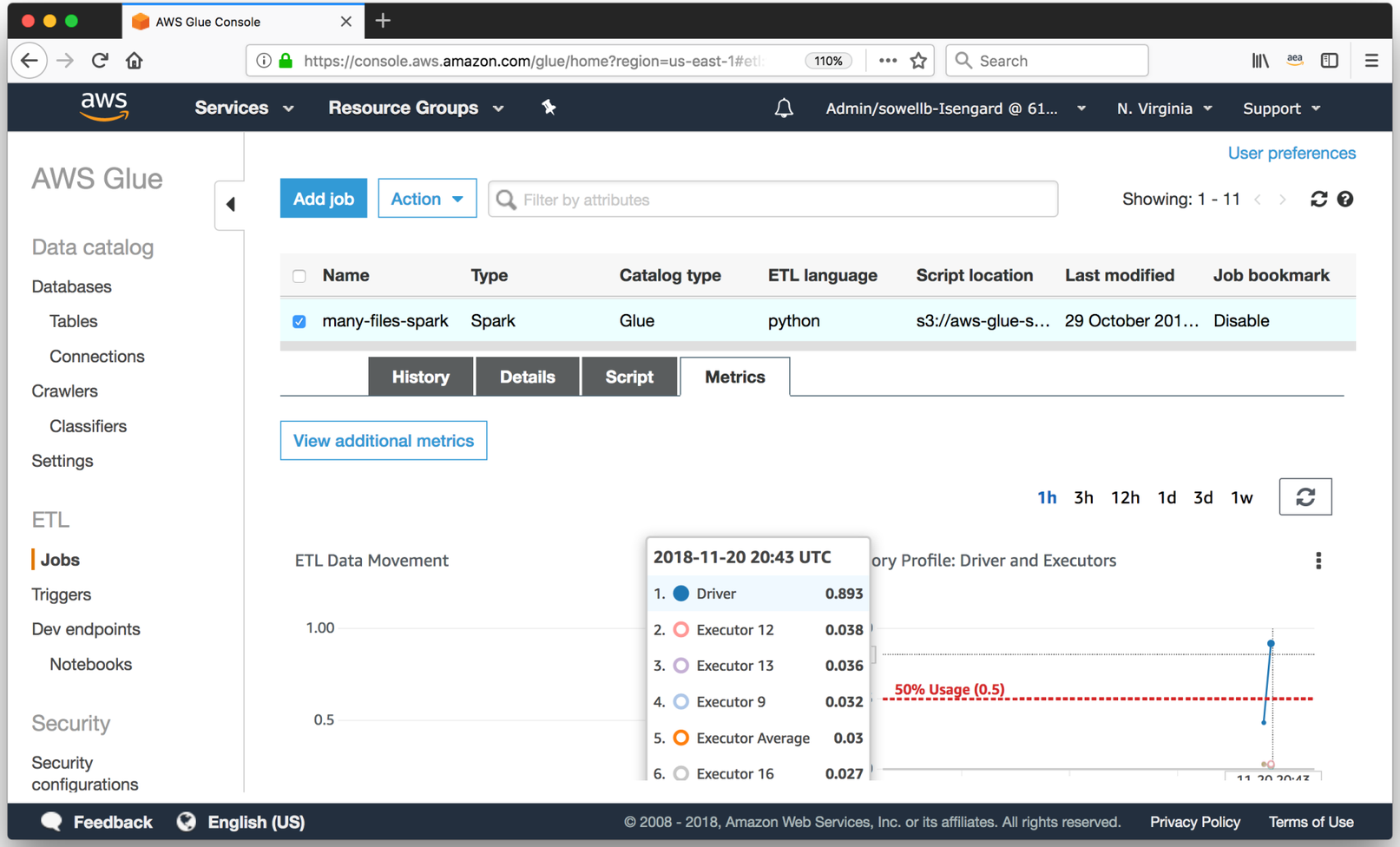
Driver and per executor

aggregates and instantaneous
reports to Amazon CloudWatch metrics every 30 sec

Metrics: memory usage, bytes read and written,
cpu load, bytes shuffled, needed executors,
and more

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example: profiling memory usage

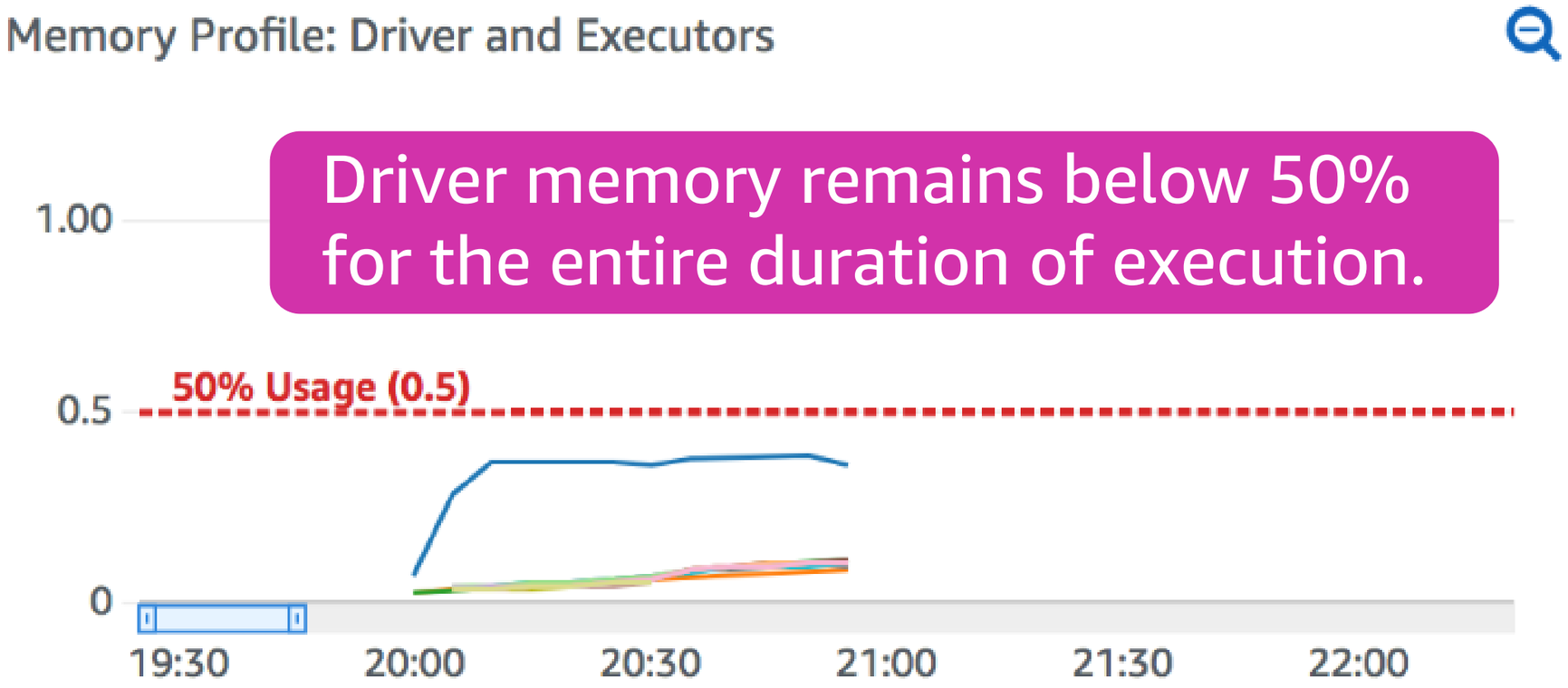


Processing lots of small files with DataFrames

Task metadata overwhelms driver

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example: AWS Glue small-file handling

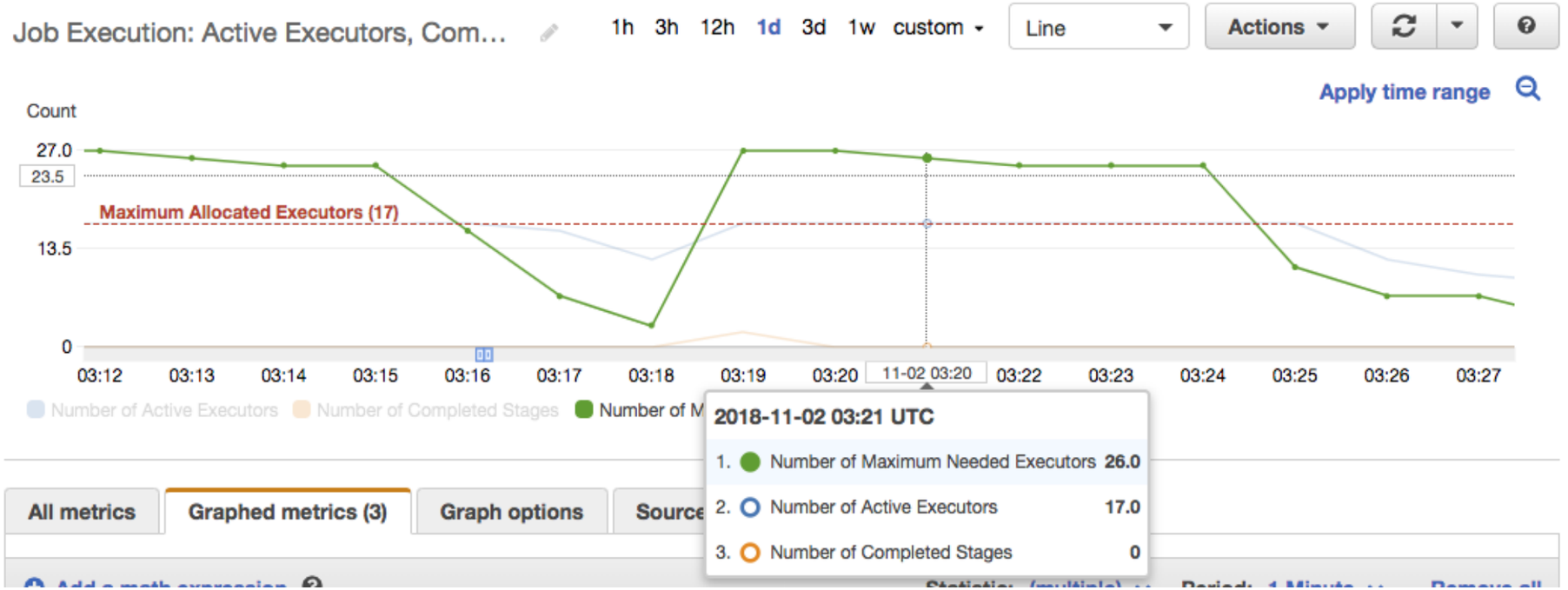


DynamicFrames automatically group files into fewer tasks

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example: optimizing parallelism

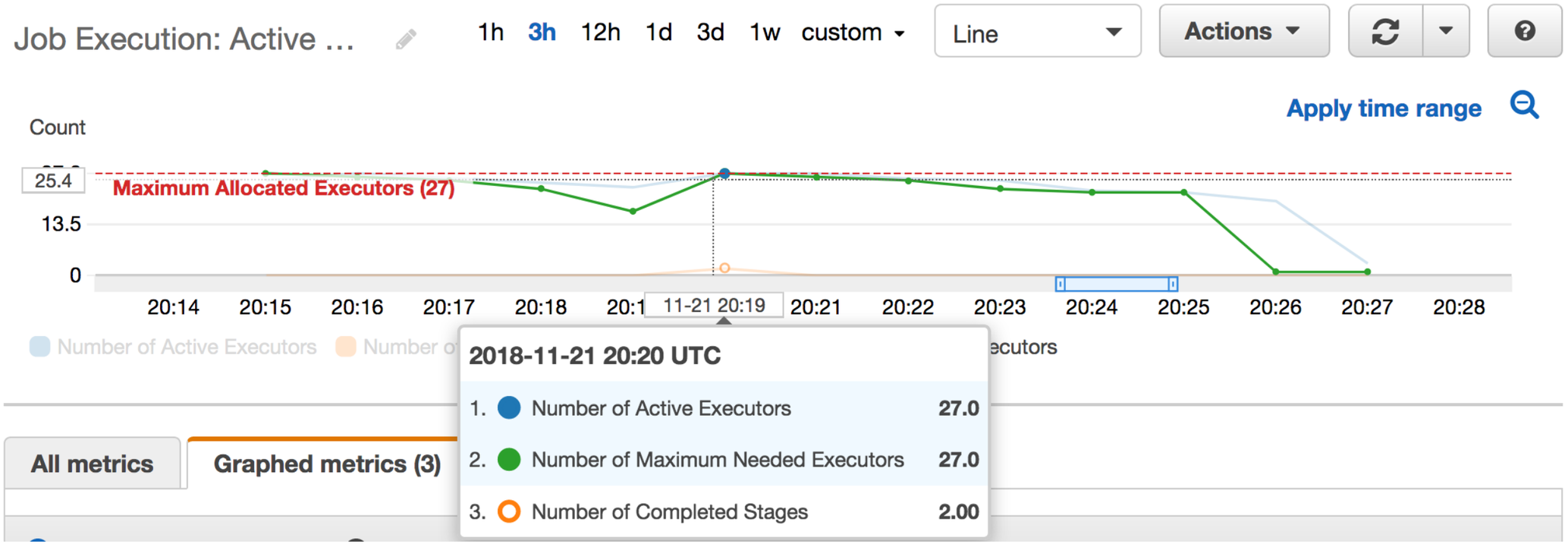
Processing large, split-able bzip2 files.
With 10 DPU, metric *maximum needed executors* shows room for scaling.



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example: optimizing parallelism

With 15 DPU, *active executors* closely tracks *maximum needed executors*.



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Innovation areas

Scalability and performance

Crawlers and ETL jobs

Profiling and debugging

ETL metrics

Interoperability

Orchestration features

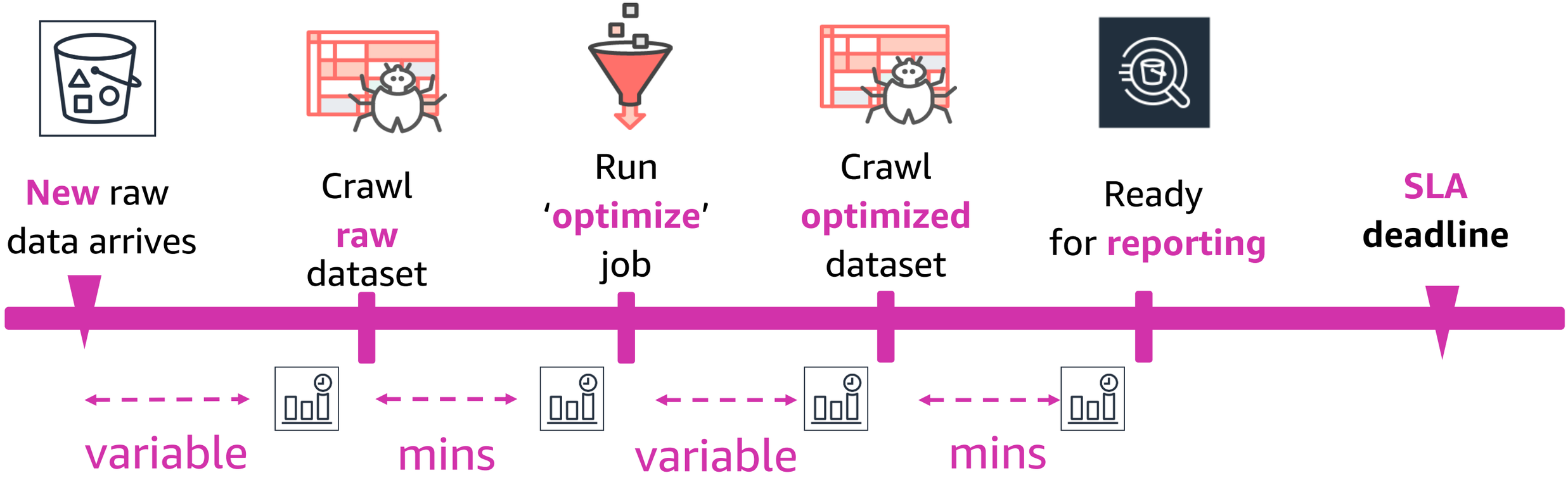
Announcing new job type: Python shell

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Orchestration a year ago …

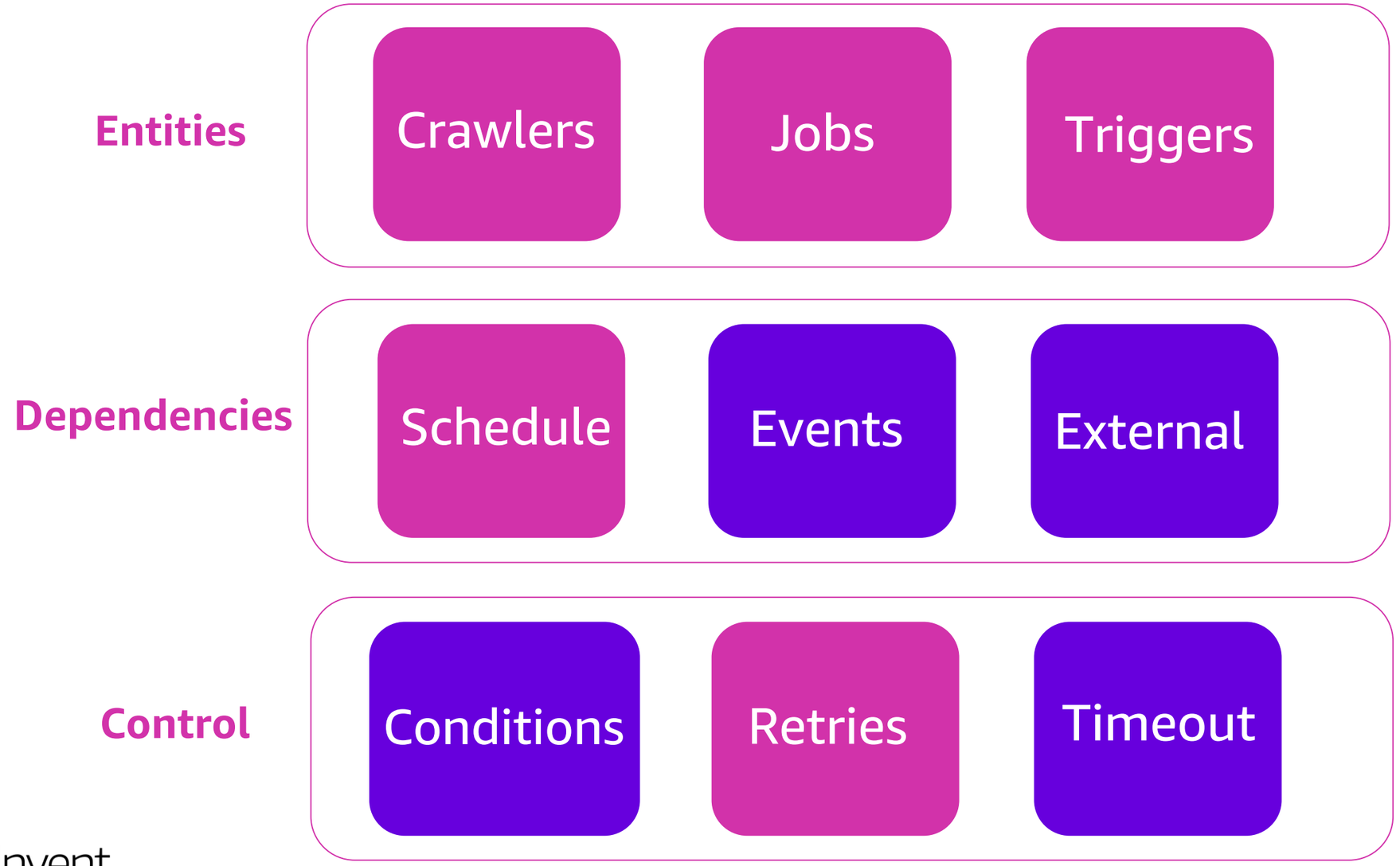
Goal: compose jobs in DAG through event-based dependencies

In-practice: time-based workflows



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Orchestration building blocks



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

New features

External

Publish crawler and job notifications into Amazon CloudWatch
Amazon CloudWatch events to control downstream workflows

Conditions

‘ANY’ and ‘ALL’ operators in Trigger conditions
Additional job states ’failed’, ‘stopped’, or ‘timeout’

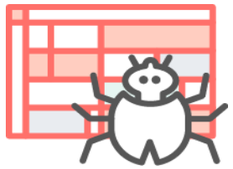
Control

Configure job timeout
Job delay notifications

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

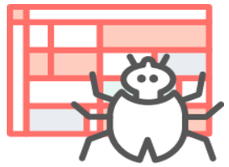
Example event-driven workflow

New raw data arrives



Crawl **raw** dataset

Run ‘**optimize**’ job



Crawl **optimized** dataset

Ready for **reporting**

SLA deadline

AWS Lambda

Start crawl

Amazon CloudWatch

AWS Lambda

On-demand trigger

Amazon CloudWatch

AWS Lambda

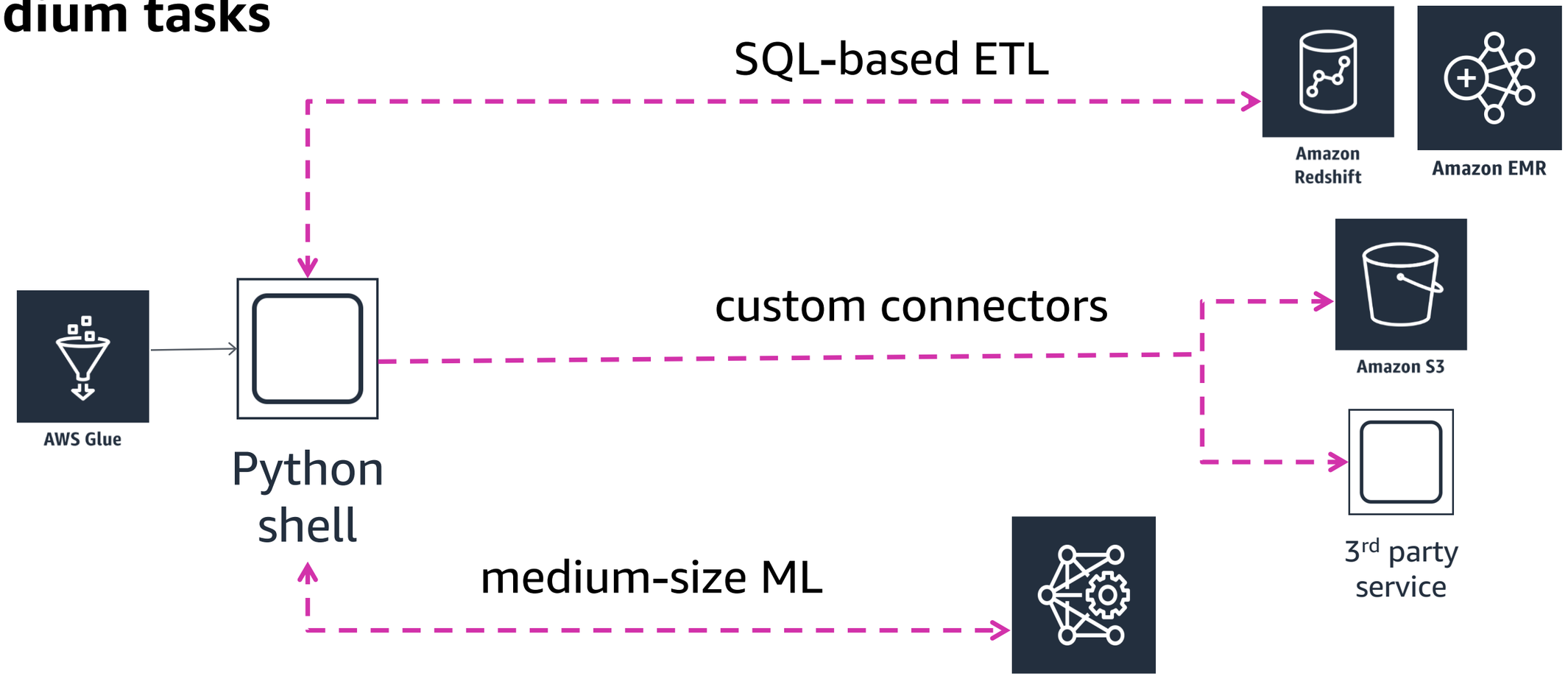
Start crawl

Amazon CloudWatch

eb Services, Inc. or its affiliates. All rights reserved.

Announcing a new job type: Python shell

A new *cost-effective* ETL primitive for small to medium tasks



© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserv

AWS Glue Python shell specs

Python 2.7 environment with
boto3, awscli, numpy, scipy, pandas, scikit-learn, PyGreSQL, …

cold spin-up: < 20 sec, support for VPCs, no runtime limit

sizes: 1 DPU (includes 16GB), and 1/16 DPU (includes 1GB)

pricing: $0.44 per DPU-hour, 1-min minimum, per-second billing

Coming soon (December 2018)

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Python shell collaborative filtering example

Amazon customer reviews dataset (s3://amazon-reviews-pds)

Video category

Compute low-rank approx of (Customer **x** Product) ratings using SVD

uses scipy sparse matrix and SVD library

Step	Time (sec)
Redshift COPY	13
Extract ratings	5
Generate matrix	1552
SVD (k=1000)	2575
Total	4145

1 DPU
matrix: 217K **x** 384K
SVD -- rank = 1000

runtime: 69 min
estimated cost: $0.60

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

And so much more …

Support for Amazon DynamoDB tables in crawlers and AWS Glue ETL
→ Run SQL over DynamoDB tables

```
dyn_frame = create_dynamic_frame.from_options(
  connection_type = "dynamodb",
  connection_options = {"dynamodb.input.tableName" : ”DDB_TABLE"})

dyn_frame.toDF().where(“latency > 12”)
```

Encryption: platform-level, in-transit, and at rest (with AWS KMS keys)

Compliance: GDPR, HIPAA, BAA

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Related breakouts

Tuesday, November 27

ANT326 **METRICS-DRIVEN PERFORMANCE TUNING FOR AWS GLUE ETL JOBS**

8:30AM – 9:30AM | Venetian, Level 2, Veronese 2406

Tuesday, November 27

ANT309 **BUILD AND GOVERN YOUR DATA LAKES WITH AWS GLUE**

9:15AM – 10:15AM | Mirage, Mirage Events Center B

Tuesday, November 27

ANT313 **SERVERLESS DATA PREP WITH AWS GLUE**

7:00PM – 9:15PM | Venetian, Level 2, Venetian H

Wednesday, November 28

ANT342 **WORKING WITH RELATIONAL DATABASES IN AWS GLUE ETL**

2:30PM – 3:30PM | Mirage, Grand Ballroom B, Table 3

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Realtor.com – Customer success

Arup Ray
Vice President, Engineering

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

AWS GLUE @ REALTOR.COM®

Arup Ray
Vice President, Engineering

November 26, 2018

realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

OUR **MISSION**

To empower people by making all things **home** simple, efficient and **enjoyable**

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

REALTOR.COM® MAKES **MEANINGFUL CONNECTIONS**

Engage

- 1.5 times more page views[1]
- 1.3 times longer visits[1]

Attract

- 9 of 10 consumer leads are actively searching or ready to transact[2]
- #1 for finding an agent[3]

Connect

- 63 million unique visitors per month[4]
- 2 billion+ page views per month[4]

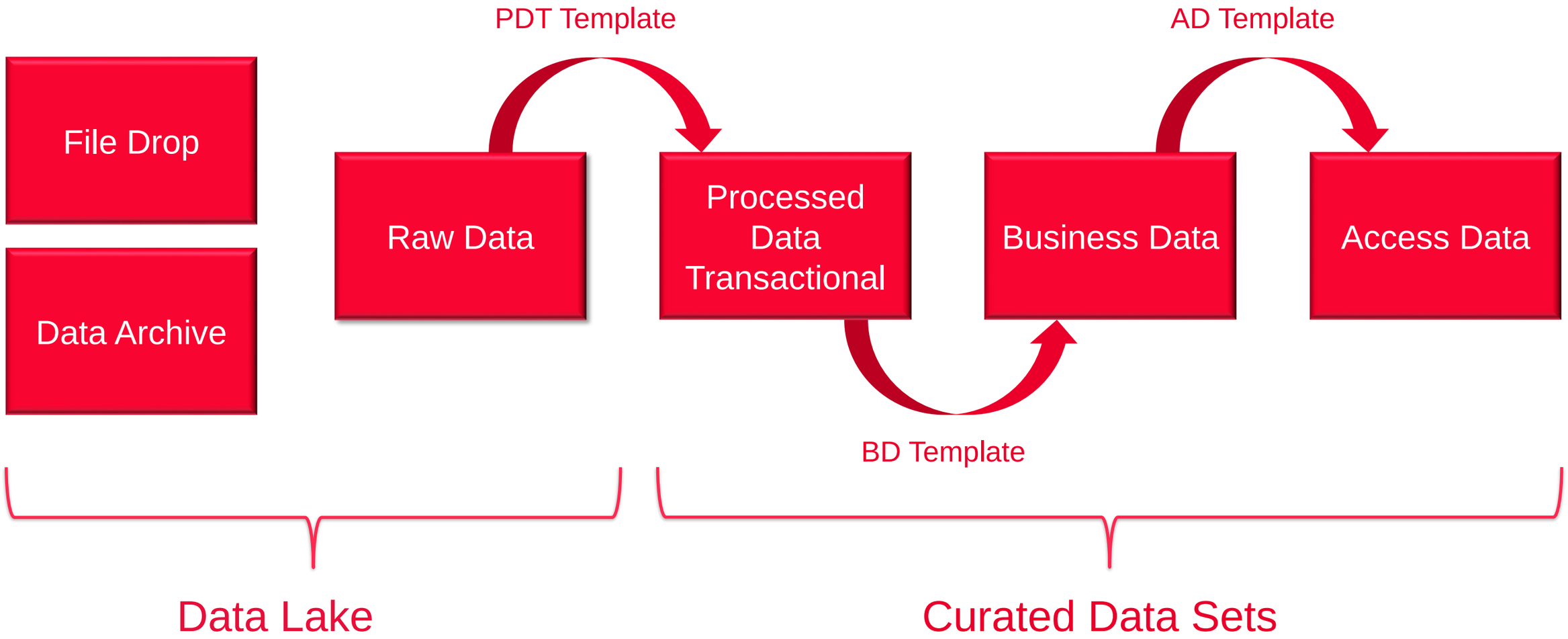
1 comScore Sept. 2018
2 Lead Quality Study, July 2017
3 2017 NAR Profile of Home Buyers and Sellers
4 FY Q418 Internal metrics

realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

DATA ANALYTICS PLATFORM

Template based data transformation



realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

TEMPLATE DRIVEN TRANSFORMATION

PROBLEM

- Generalize the ETL process to handle generic data sets and volumes
- Solve for patterns – like deeply nested JSON attributes, array within JSON attributes
- Make the data easy to query

SOLUTION

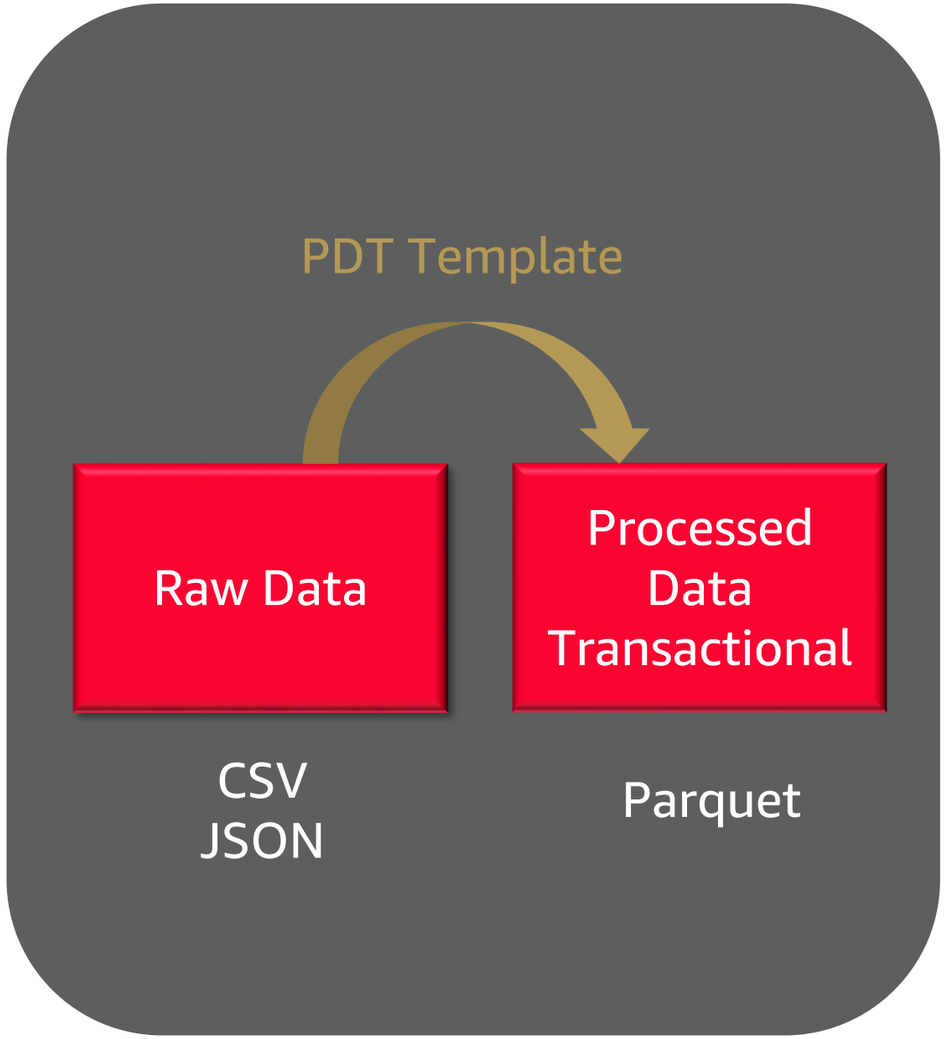
CHALLENGES

- EMR Hive requires managing the cluster
- UDF based functionality requires a lot of code to maintain

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

PDT TEMPLATE SOLUTION

Between RD and PDT layers in Data Analytics Platform



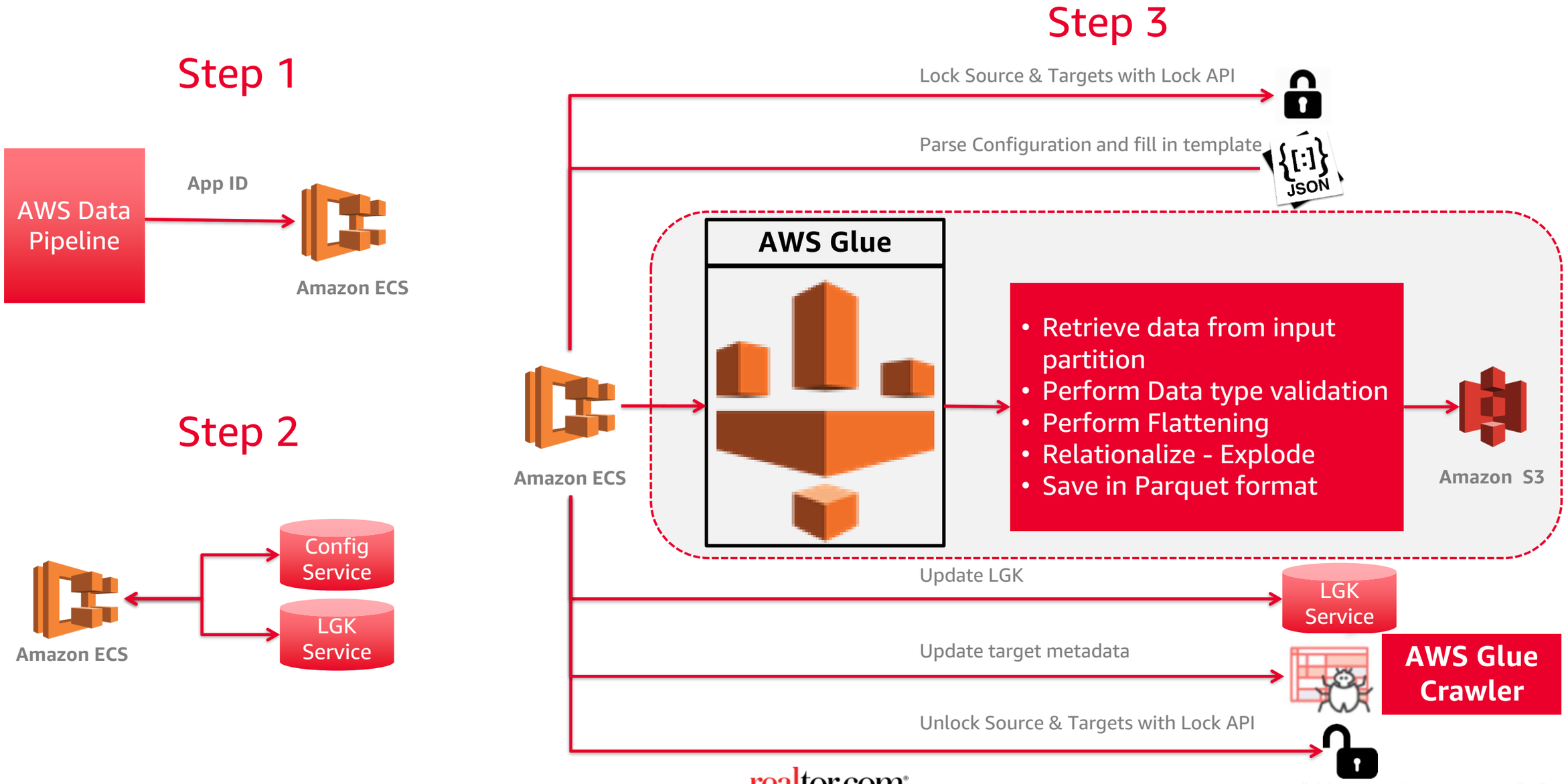
- Template code implemented in Python.
- User defined configuration as JSON
- Supported Raw Data formats – CSV with header, JSON
- These are the template configuration parameters
  - Data location (RD S3 location PDT S3 location)
  - Data structure of input – implicit or header location
  - Input column names that require data type validation
  - Relationalize
    - Input multi-valued columns that requires to be *relationalized*
    - Deeply nested columns that require flattening
  - Duplicate detection of input rows (optional)

realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

PDT TEMPLATE

How AWS Glue performs batch data processing

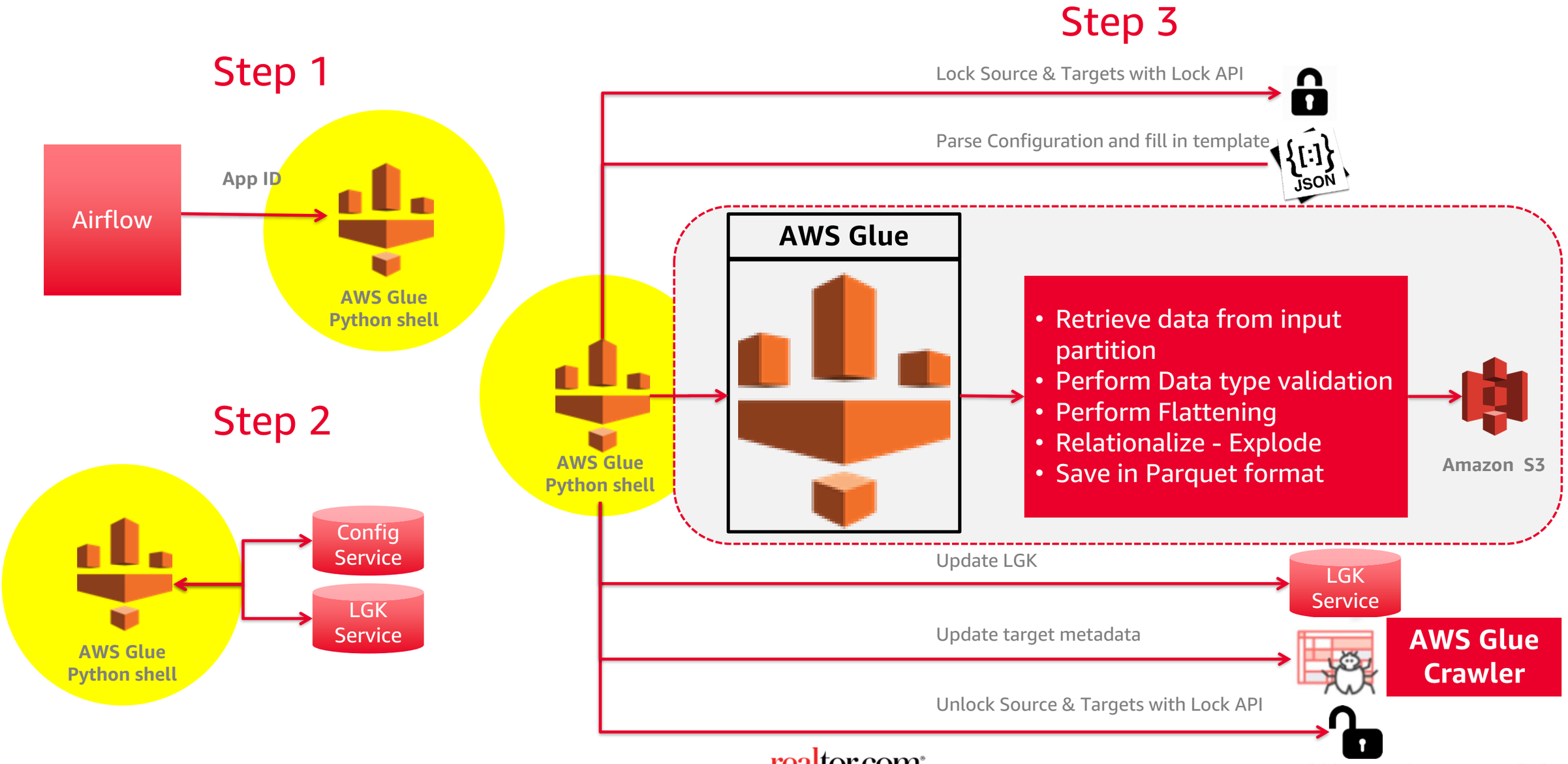


realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

PDT TEMPLATE

How AWS Glue performs batch data processing

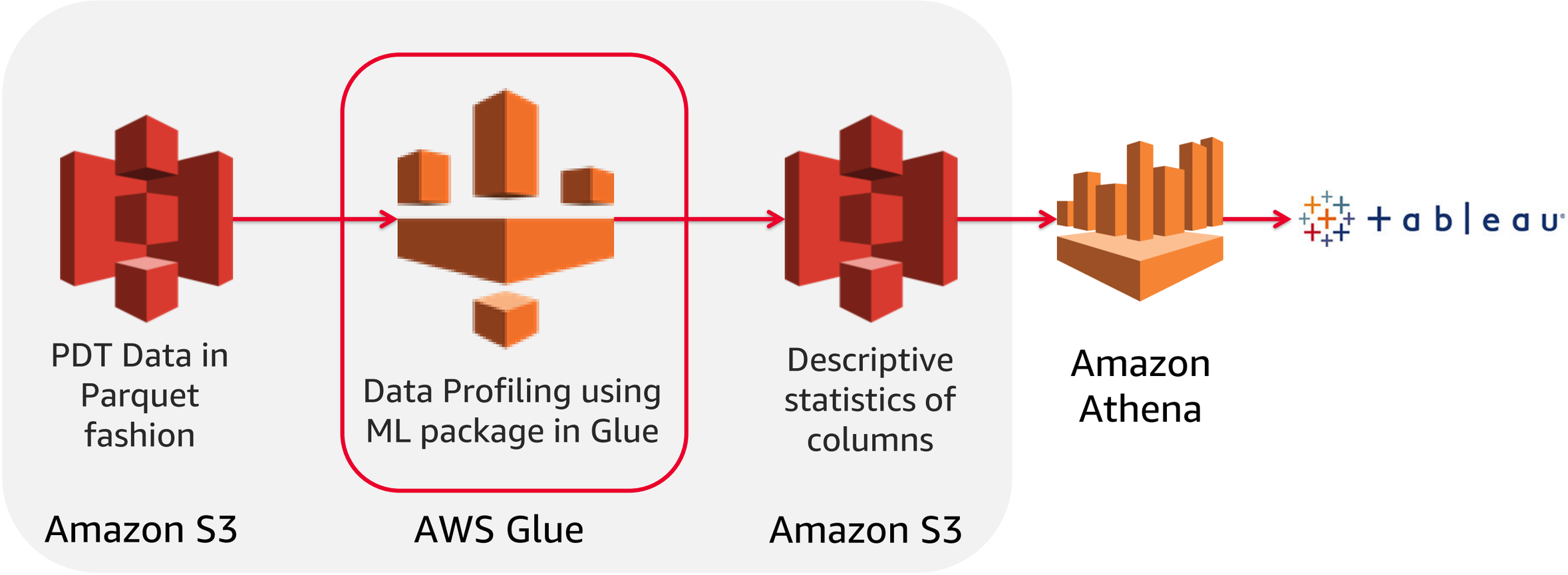


realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

DATA PROFILING

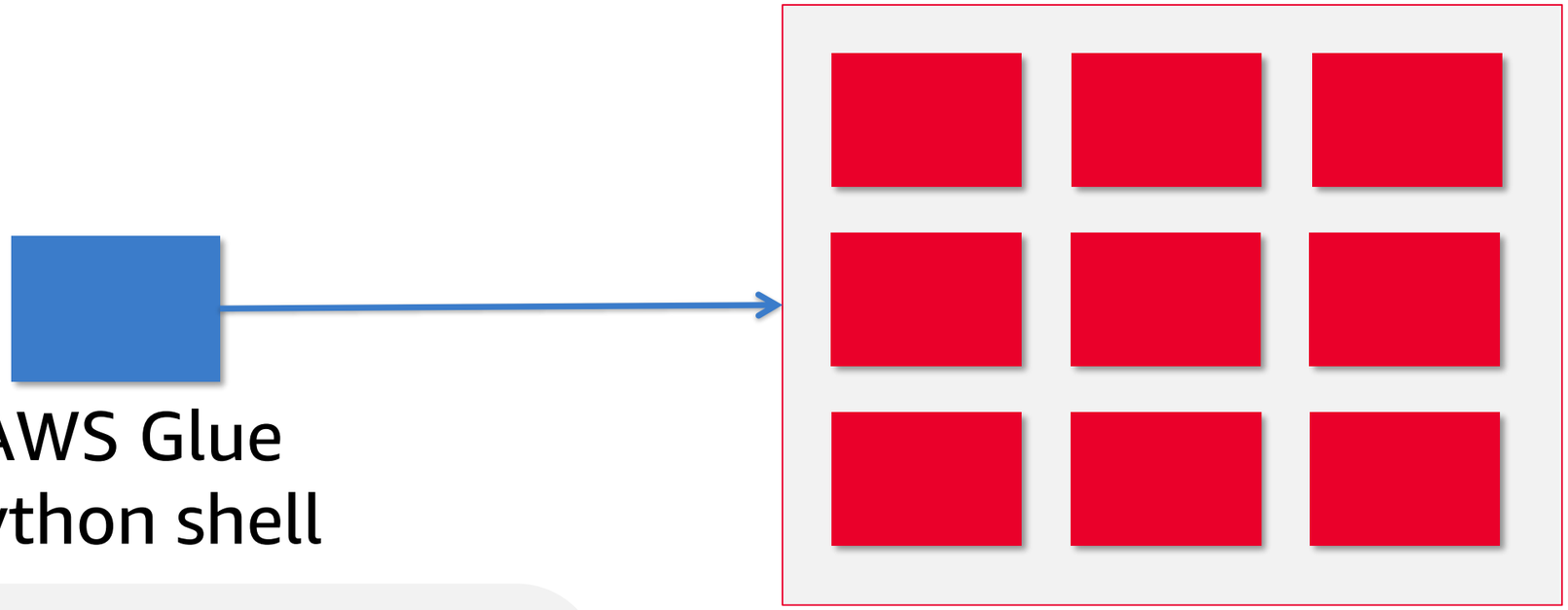
Using Data Validation Framework



© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

AWS GLUE PYTHON SHELL

The way we see it



AWS Glue
Python shell

AWS Glue

Serverless Edge Node for triggering, light transformations, uncompress, tar extract, Parquet conversion

Serverless Distributed Compute Cluster

realtor.com®

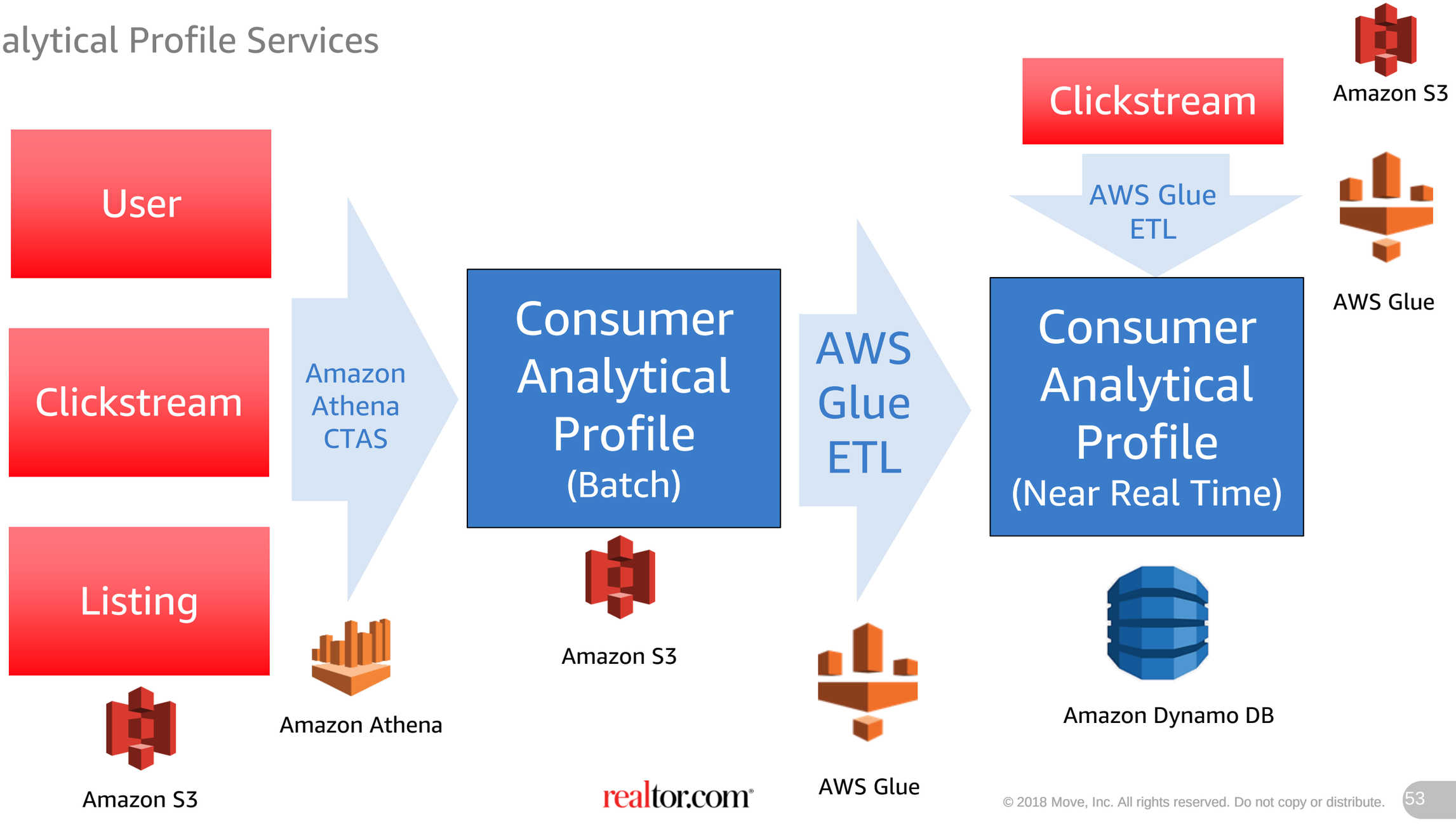
© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

We analyze consumer behavior and expectation to deliver a best-in-class experience for both REALTORS® and consumers

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

BEYOND PDT

Analytical Profile Services



realtor.com

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

BENEFITS OF USING AWS GLUE

Speed of implementation – developer productivity with in-built transforms

Less code to maintain - <10 lines versus 200+ for explosion of array attributes

Operations team loves the server less aspect

Performance for Dynamo DB write using AWS Glue significantly better - about 10X

realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.

Thank you!

Mehul A. Shah
glue-pm@amazon.com

Arup Ray
https://www.linkedin.com/in/arupray

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Please complete the session survey in the mobile app.

AWS re:Invent

© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

THANK YOU

realtor.com®

© 2018 Move, Inc. All rights reserved. Do not copy or distribute.