AWS re:Invent

A N T 3 7 2

Building Advanced Workflows with AWS Glue

Santosh Chandrachood
SDM
AWS Glue

AWS re:Invent

aws

Agenda

Overview

Building Blocks

Building a usecase

Event driven workflows

Workflow considerations

Monitoring and Tuning

Bring your own workflow engine

Breakout repeats

Monday, Nov 26

ANT 372 – [CT] Building Advanced Workflows with AWS Glue
10:00 – 11:00 | Aria East, Plaza Level, Orovada 3

Tuesday, Nov 27

ANT 333 – [BS] Building Advanced Workflows with AWS Glue
2:30 – 3:30 | Mirage, Grand Ballroom D, Table 4

Wednesday, Nov 28

ANT 381 – [BS] Building Advanced Workflows with AWS Glue
5:30 – 6:30 | Aria West, Level 3, Starvine 10, Table 5

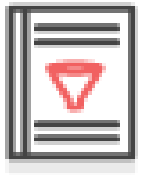
AWS Glue

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

for developers, built by developers

1000s of customers and jobs

AWS Glue Components



Data Catalog

Discover

- Automatic crawling
- Apache Hive Metastore compatible
- Integrated with AWS analytic services

ETL Engine

Develop

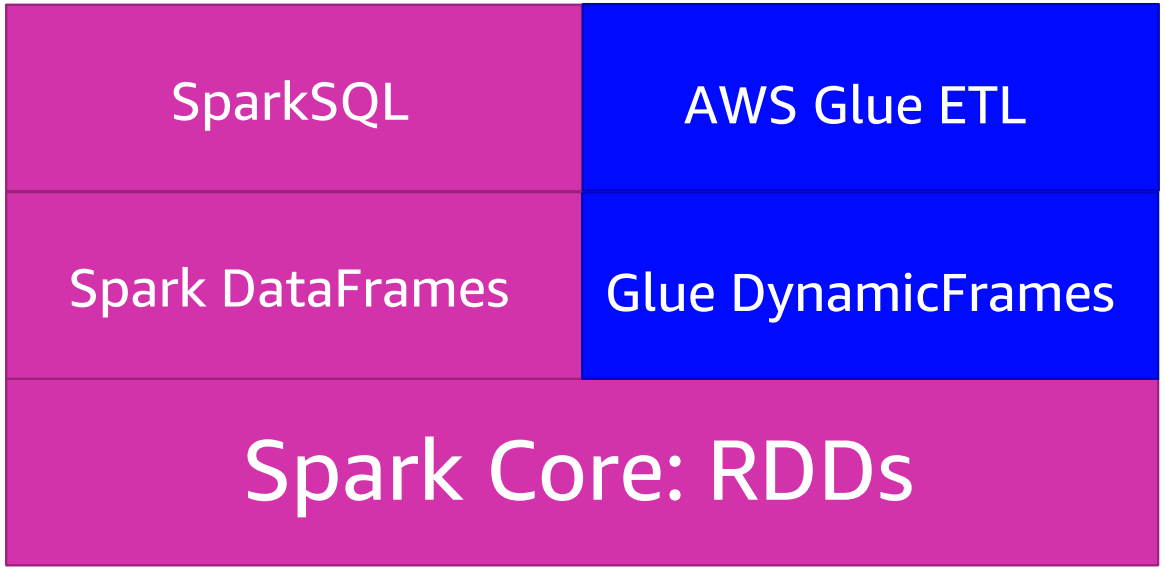
- Apache Spark core
- Python and Scala
- Auto-generates ETL code

Orchestration

Deploy

- Flexible scheduling
- Monitoring and alerting
- External integrations

Review: Apache Spark and AWS Glue ETL



Application

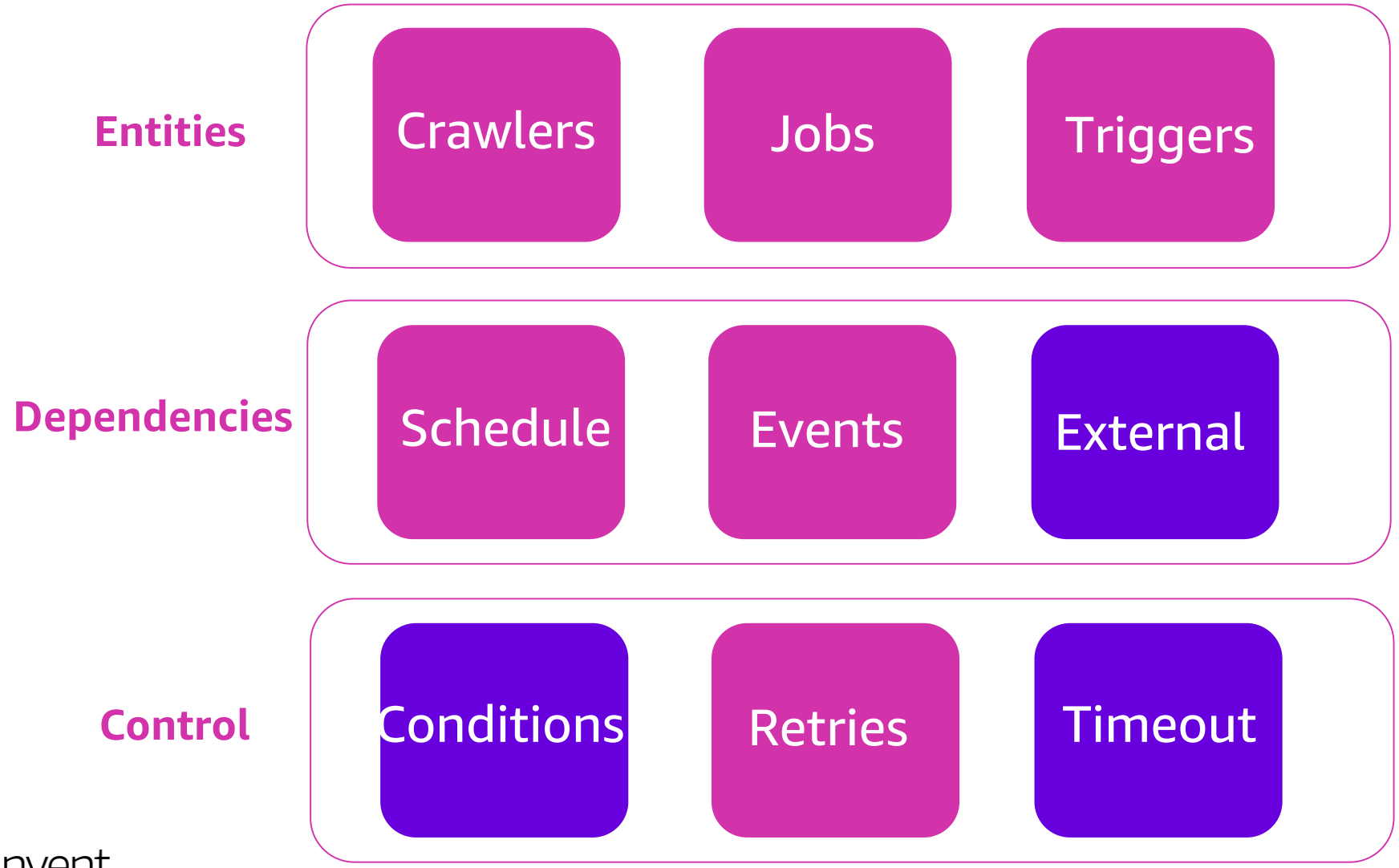
Data Structure

Execution

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

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

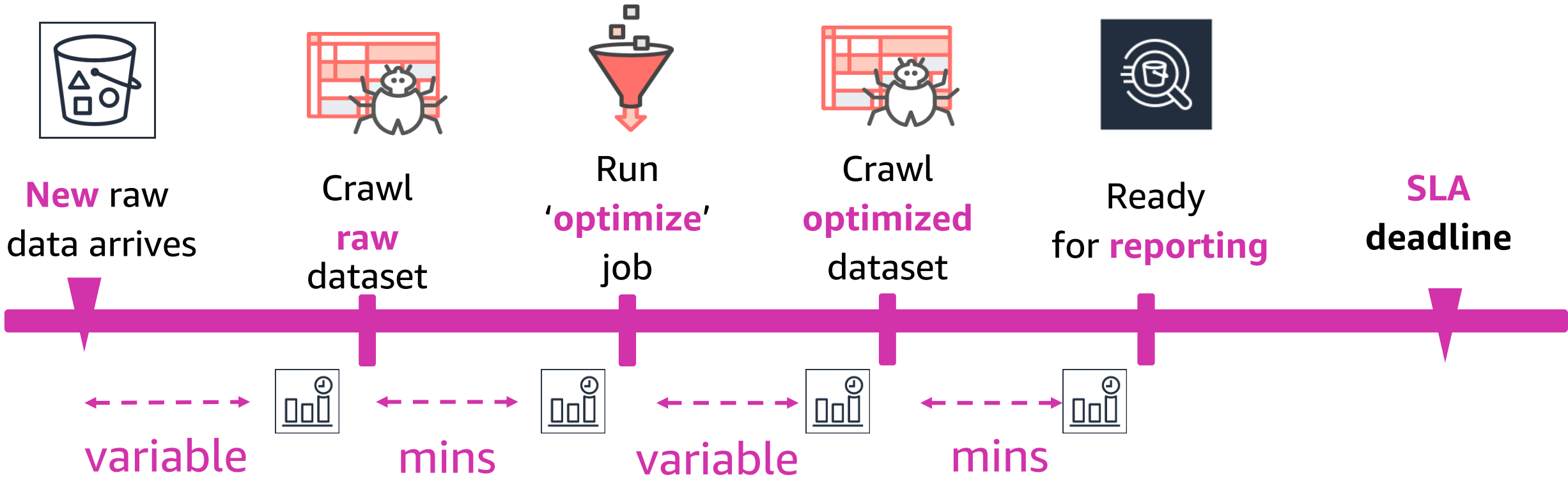
Building Blocks



Building a usecase

Goal: compose jobs in DAG through dependencies

In-practice: time-based workflows



New features

External

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

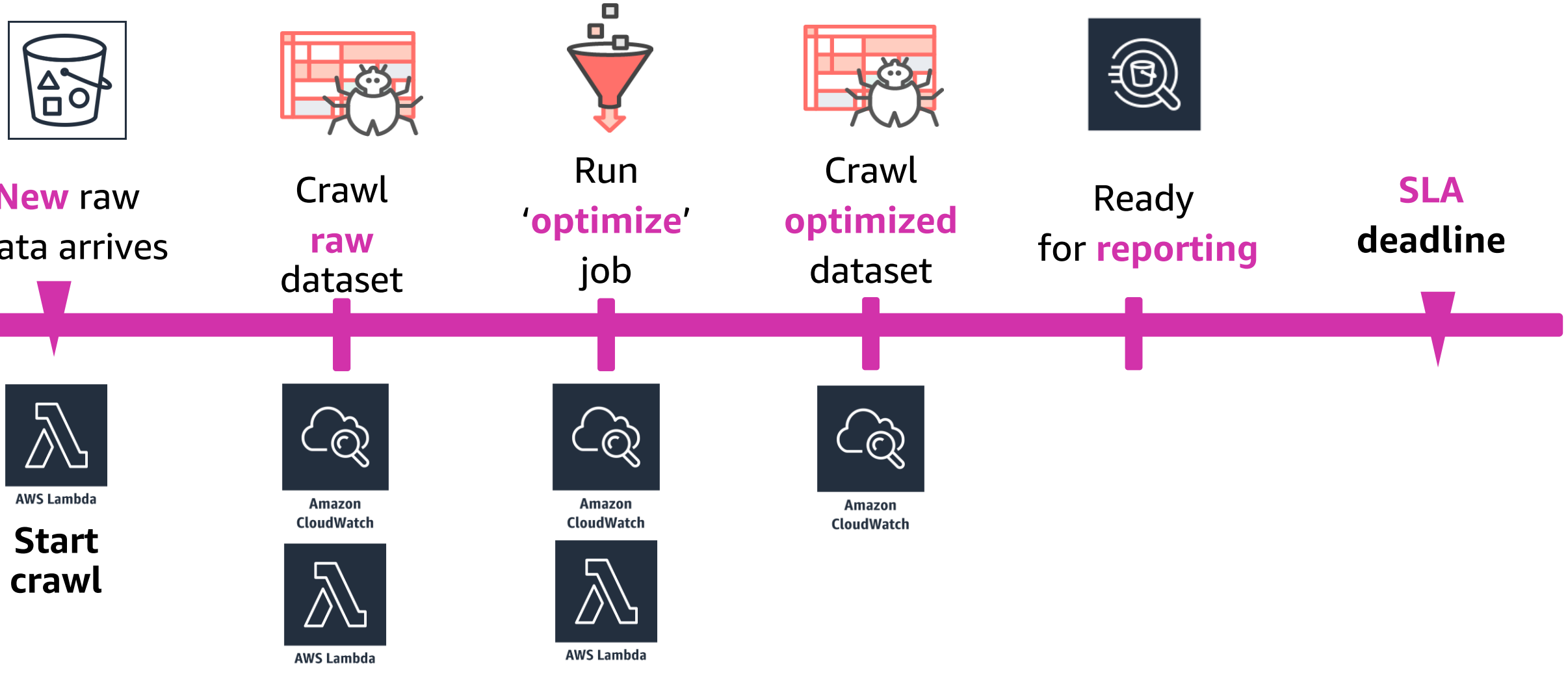
Conditions

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

Control

Configure job timeout
Job delay notifications
On-demand cancel

Example event-driven workflow



Workflow considerations

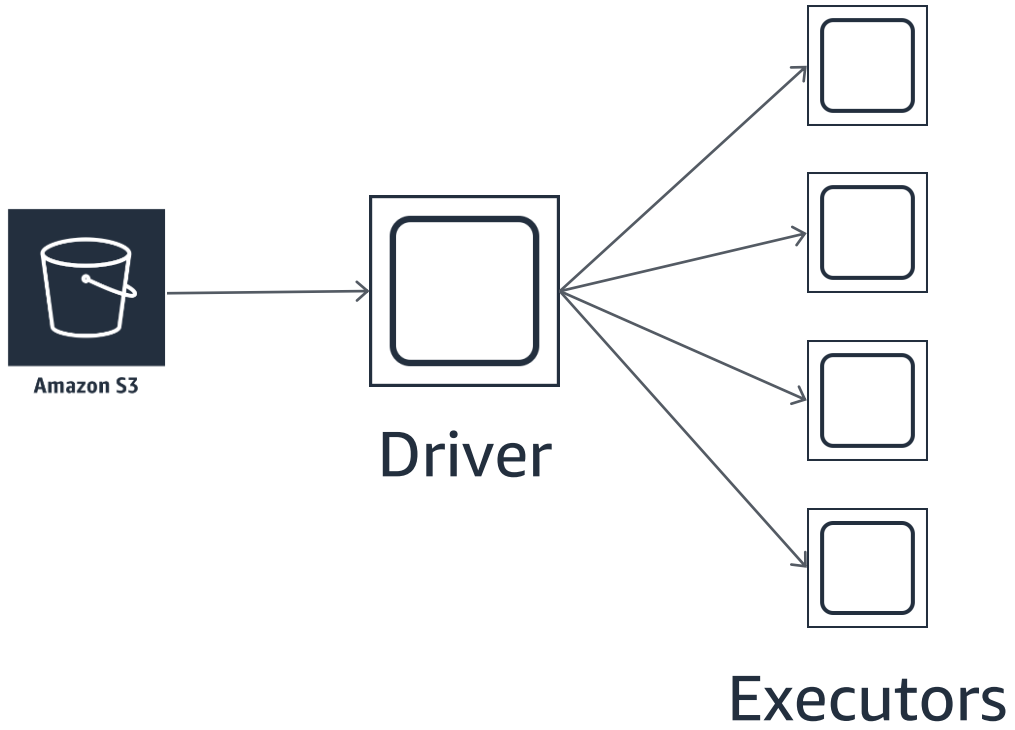
- Incremental data processing
  - Job bookmarks to keep state
  - Job parameters to select new datasets
- Job size
  - Unique versus One job per logical units of work
  - Multiple small jobs or one big job
- Job parameters
  - Initial, Global, In-between jobs
  - Use Amazon S3 to pass parameters

Workflow considerations

- Data processing unit
  - Number of DPUs
  - Adjusting DPUs
- SLA
  - Job delays notifications
  - Timeouts
- Error handling
  - Retry logic
  - Integration with 3rd party
  - Job re-run

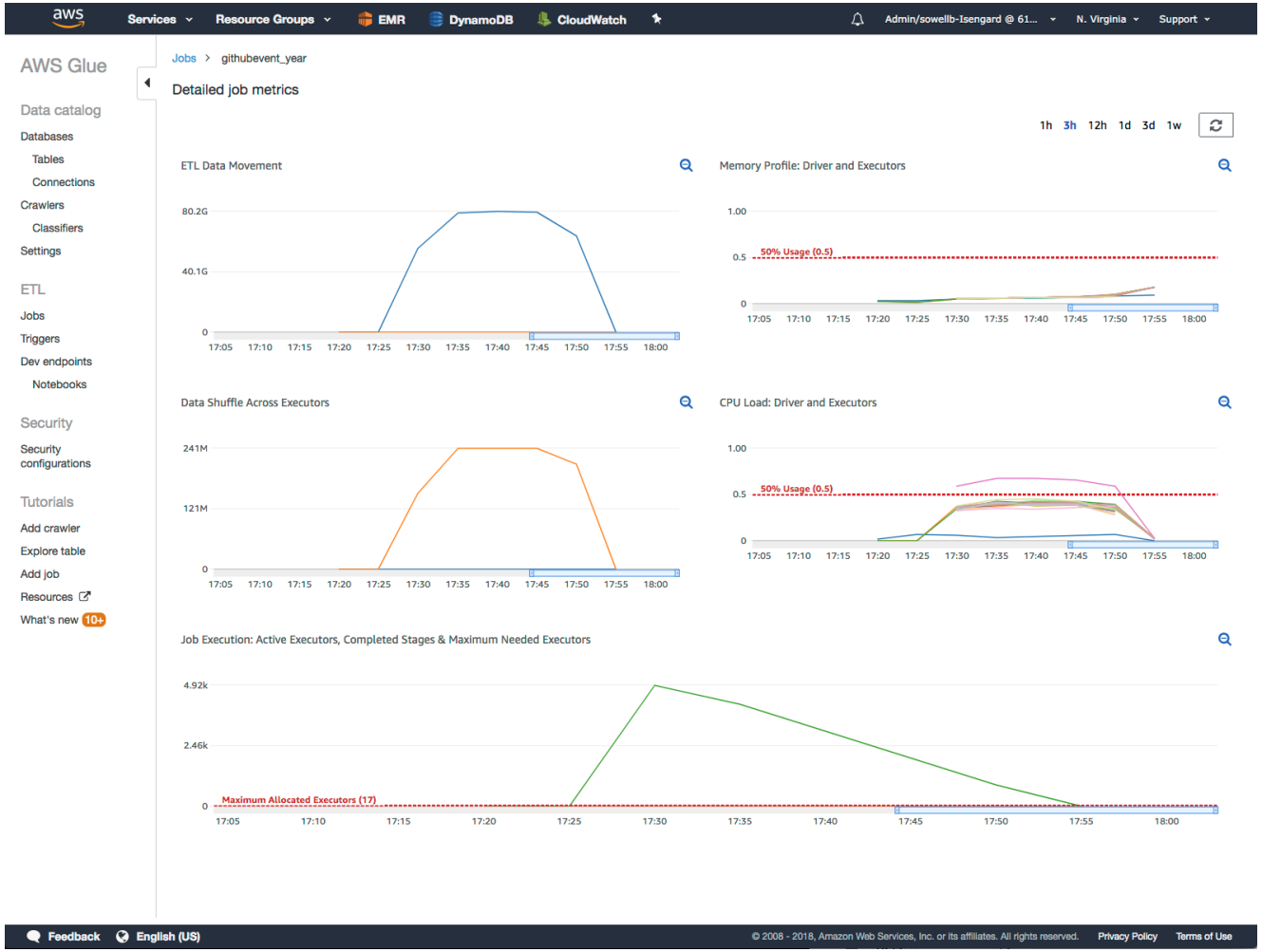
Workflow monitoring—Performance

- How is your dataset *partitioned*?
- How is your application divided into *jobs* and *stages*?
- Data is divided into *partitions* that are processed concurrently

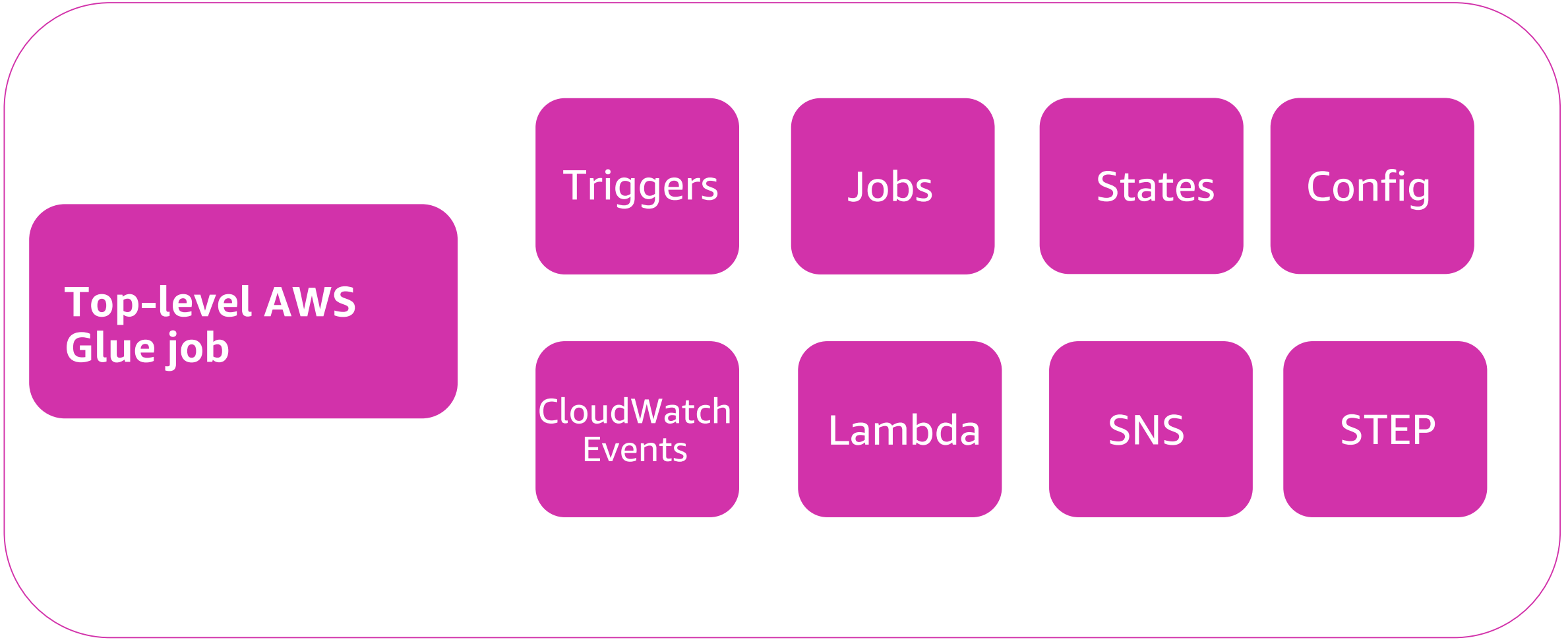


Workflow monitoring—Metrics

- Job metrics
  - CPU
  - Memory
  - Network
  - Executors, stages
  - Data movement
- Use data points to adjust job parameters



Bring your own workflow engine



Thank you!

Santosh Chandrachood
glue-feedback@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.