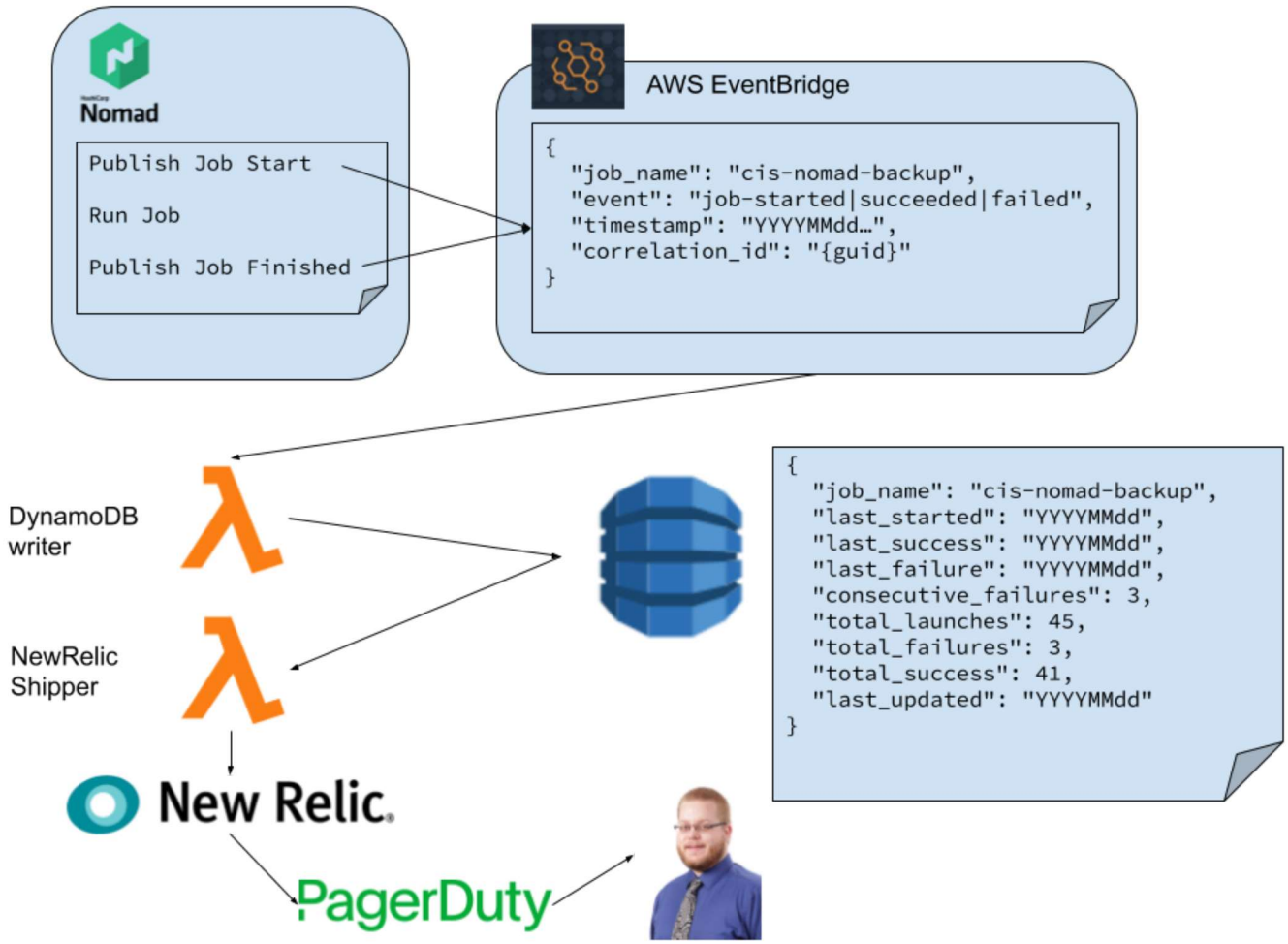
Platypus Operations

Link to various operations related documentation follow:

- Platypus FMEA
- Consul Runbook
- Nomad Runbook

Platypus Monitoring



Platypus monitoring workflow diagram

The Platypus project includes a robust system for monitoring job execution failures and alerting when necessary. This is instrumented by wrapper scripts built into the Platypus batch base Docker images and a publish-subscribe event sourcing mechanism built upon Amazon

EventBridge, DyanamoDB, and Lambda functions. A sample workflow of a job execution with monitoring is as follows:

- Nomad triggers a job on its periodic schedule
- The wrapper script sends a "job started" event to Amazon EventBridge
- Asynchronously, Lambda functions respond to the event by recording the event and updating the "last started" time for that job in DynamoDB
- The wrapper script spawns a process to execute the desired executable
- The wrapper script interprets the process's exit code and either sends a "job succeeded" or "job failed" event on EventBridge
- Similar to the "job started" message, Lambda functions record the event and update the job's status in DynamoDB

Every minute, a third Lambda posts each job's status from DynamoDB to the New Relic Insights API. Dashboards have been created to query this data and report on "unhealthy" (see below) jobs and other useful metrics.

New Relic Dashboards

A New Relic dashboard has been created for each account to display key metrics such as running jobs, failing jobs, and high level cluster healh metrics.

- DEV dashboard
- QA dashboard
- Prod dashboard

Each dashboard contains a `Job Metrics` page to view job run history and can be filtered down to a specific job.

What constitutes an "unhealthy" job?

Due to the flaky nature of several of DS's batch jobs, not every job failure should cause the job to go unhealthy and trigger an alert. The agreed upon definition for an "unhealthy" job is a job that meet any of these conditions:

- failed 2 or more consecutive executions
- failed once when the last successful execution was at least 24 hours ago

Alerting

New Relic alerts are set up to post to Pager Duty for when production jobs become unhealthy. As of April 15, 2020, these alerts are escalated to Team CNAME, who will then escalate with owners of the unhealthy jobs.

Developer usage

Due to the way the Platypus batch base Docker images are built and the Nomad batch job Yeoman generator, jobs built using them will automatically get the aforementioned monitoring capabilities built in. However, some use cases may require additional monitoring/alerting to ensure quality. It is at the developer's discretion to implement additional monitoring checks if needed.