Cumulus: An introduction

July 17, 2019

Mark Boyd
boyd@developmentseed.org

Who am I?

developmentSEED

- Cloud Engineer at Development Seed
- Member of the Cumulus core team
- Lover of dogs, hiking, and the Baltimore Orioles

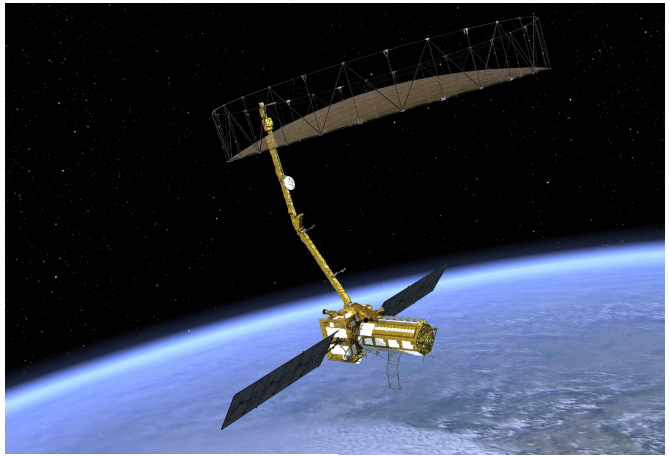
Outline

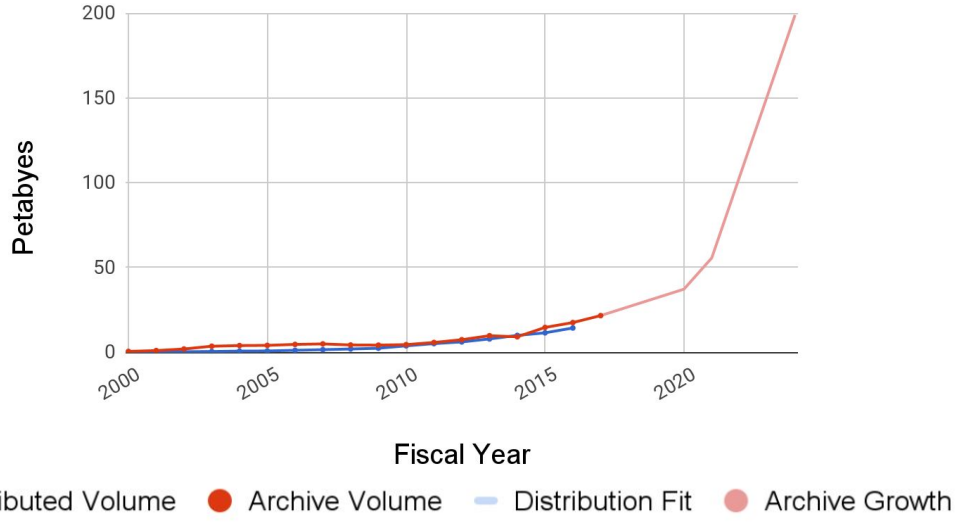
- Why Cumulus?
- What is the value?
- What is it?
- How does it work?
- What is next for Cumulus?
- Where is Cumulus?
- Demo
- Questions

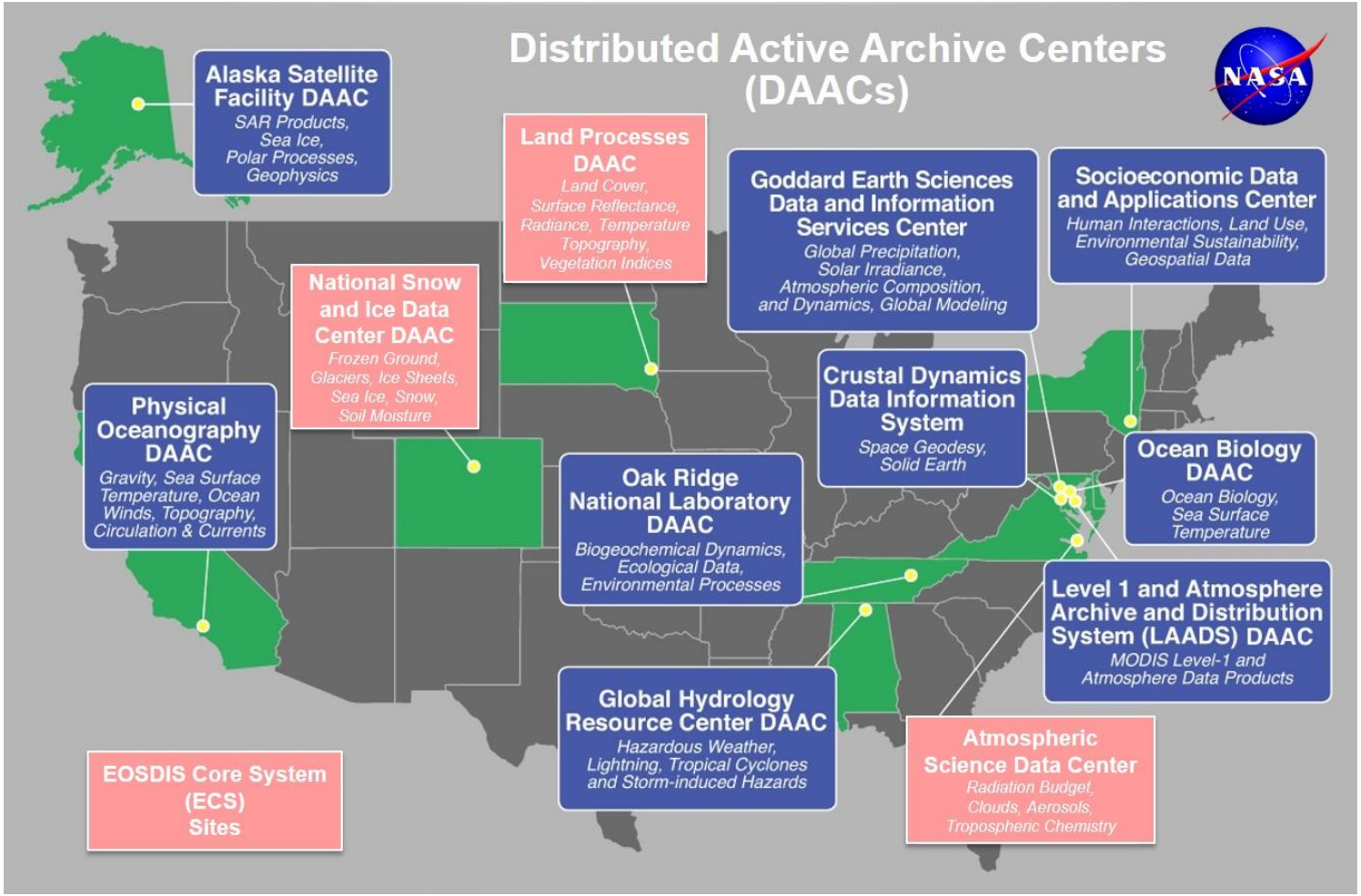
Why Cumulus?

NASA's Commitment to open data

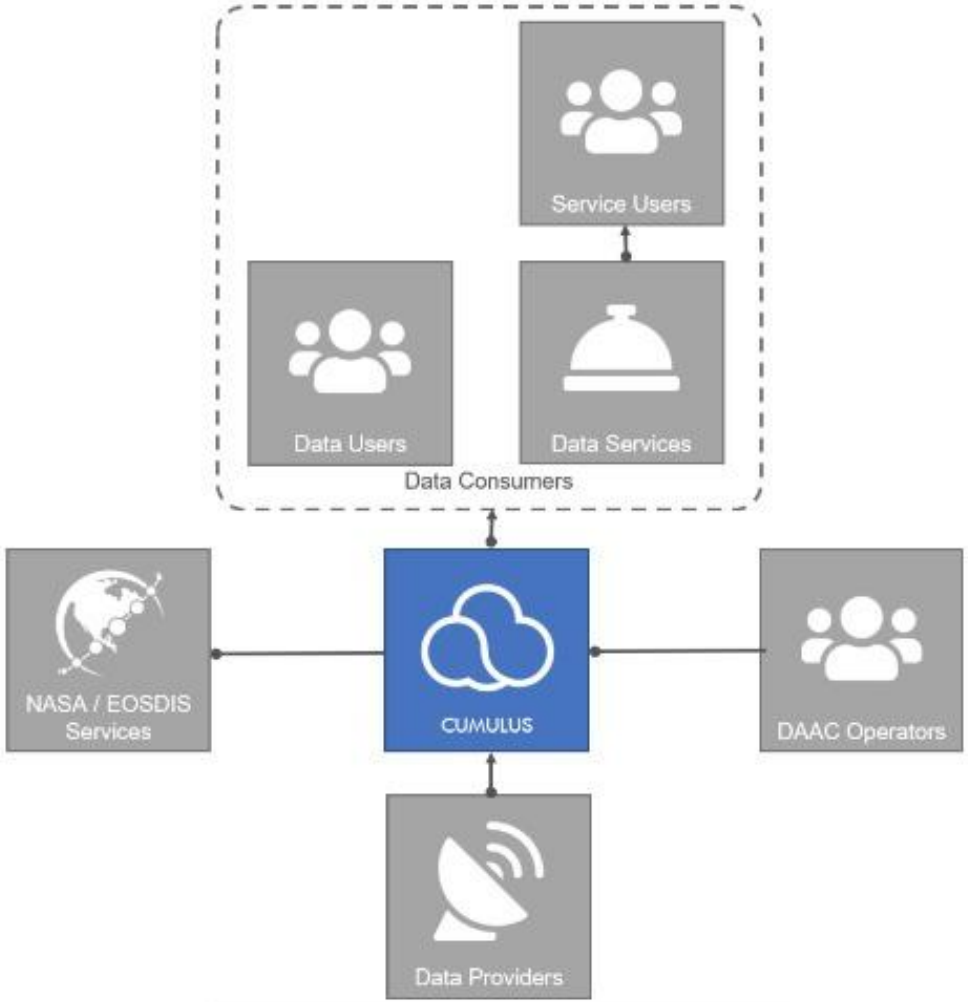
Since 1994, the ESDS Program has committed to the full and open sharing of Earth science data obtained from NASA instruments to all users.




200
150
100
50
0
Petabyes
2000
2005
2010
2015
2020
Fiscal Year
Distributed Volume
Archive Volume
Distribution Fit
Archive Growth



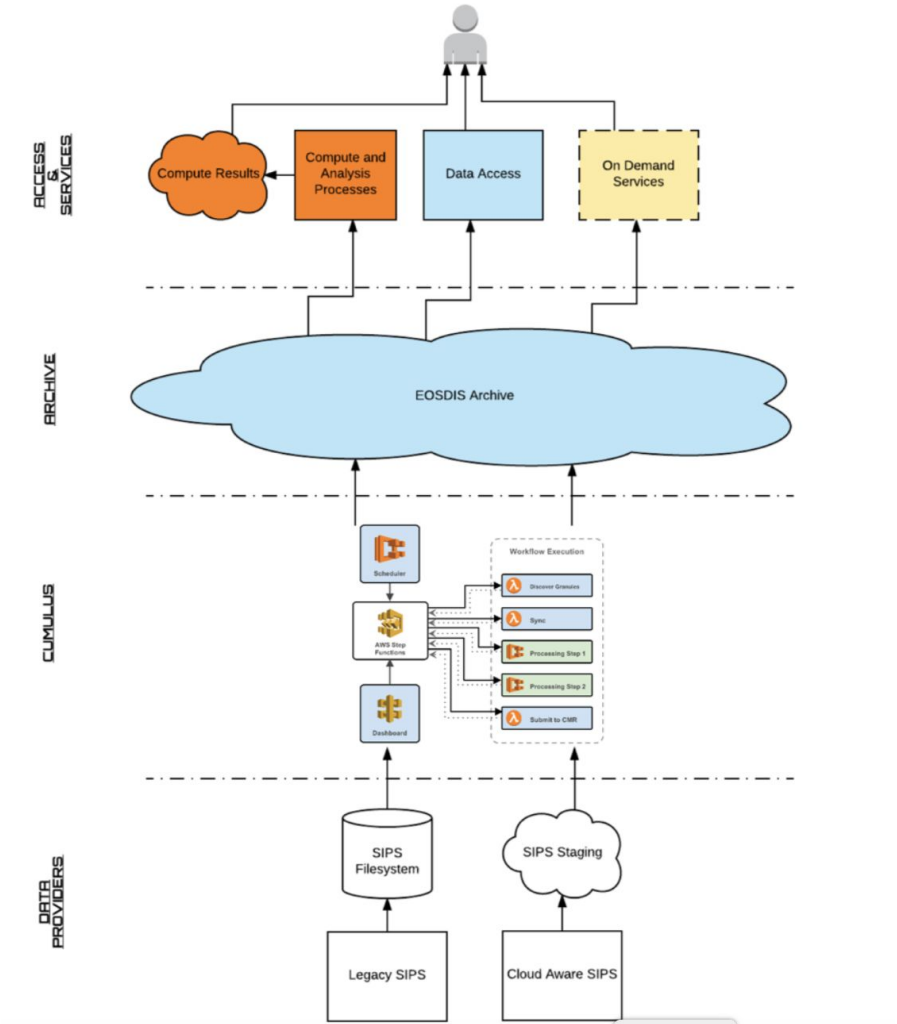
From https://earthdata.nasa.gov/eosdis/daacs


Service Users
Data Users
Data Services
Data Consumers
NASA / EOSDIS
Services
CUMULUS
DAAC Operators
Data Providers

What is the value of Cumulus?

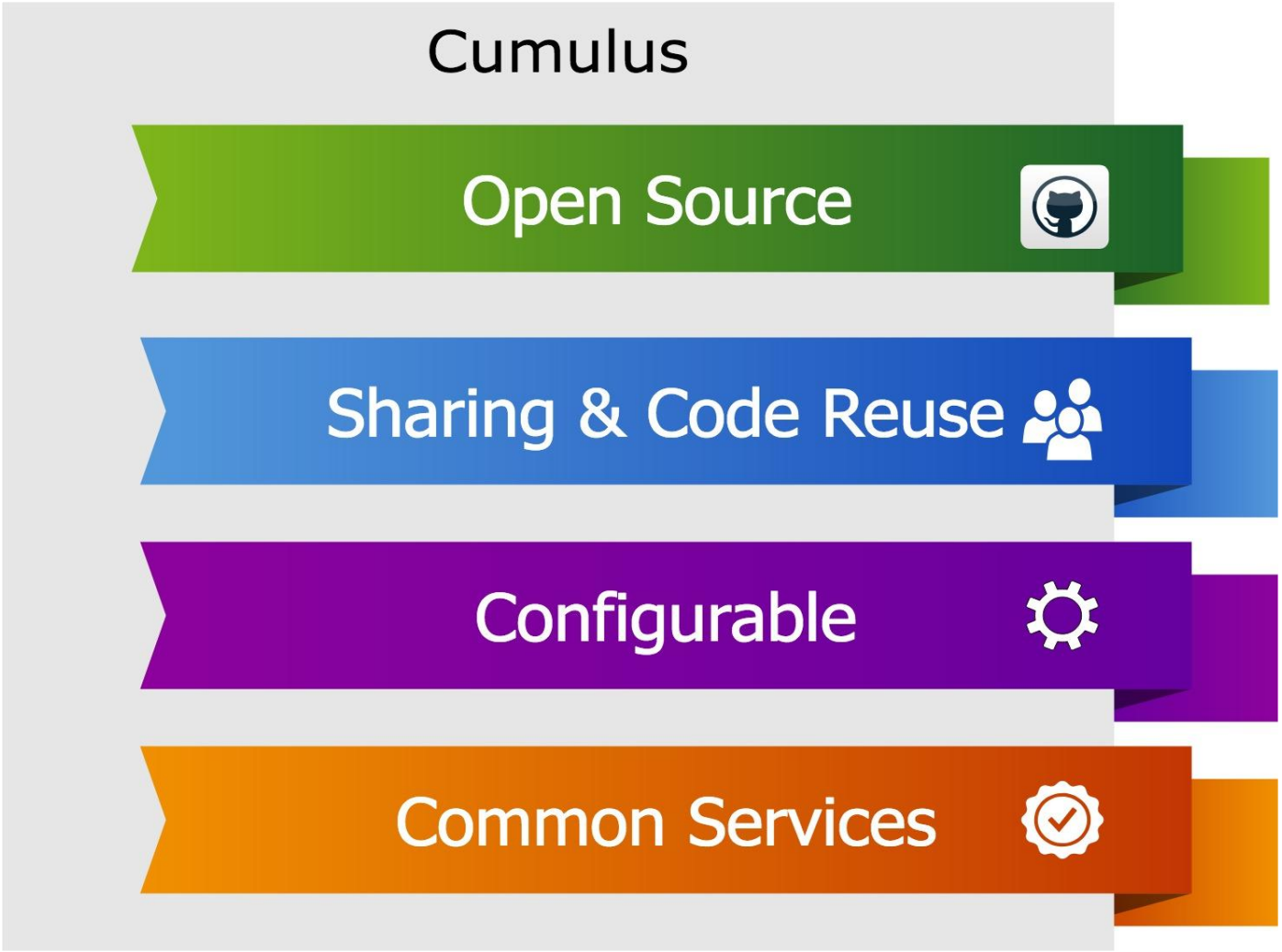
Cumulus enables science to happen in the cloud


SECURITY &
COMPLIANCE
SERVICES & METRICS
BACKUP & RECOVERY
STEWARDSHIP
SIMPLE ACCESS
ERROR HANDLING
PROVENANCE


ACCESS & SERVICES
Compute Results
Compute and Analysis Processes
Data Access
On Demand Services
ARCHIVE
EOSDIS Archive
CUMULUS
Scheduler
AWS Step Functions
Dashboard
Workflow Execution
Discover Granules
Sync
Processing Step 1
Processing Step 2
Submit to CMR
DATA PROVIDERS
SIPS Filesystem
SIPS Staging
Legacy SIPS
Cloud Aware SIPS

Preview

Cumulus + your code = science


Cumulus
Open Source
Sharing & Code Reuse
Configurable
Common Services

Collaboration

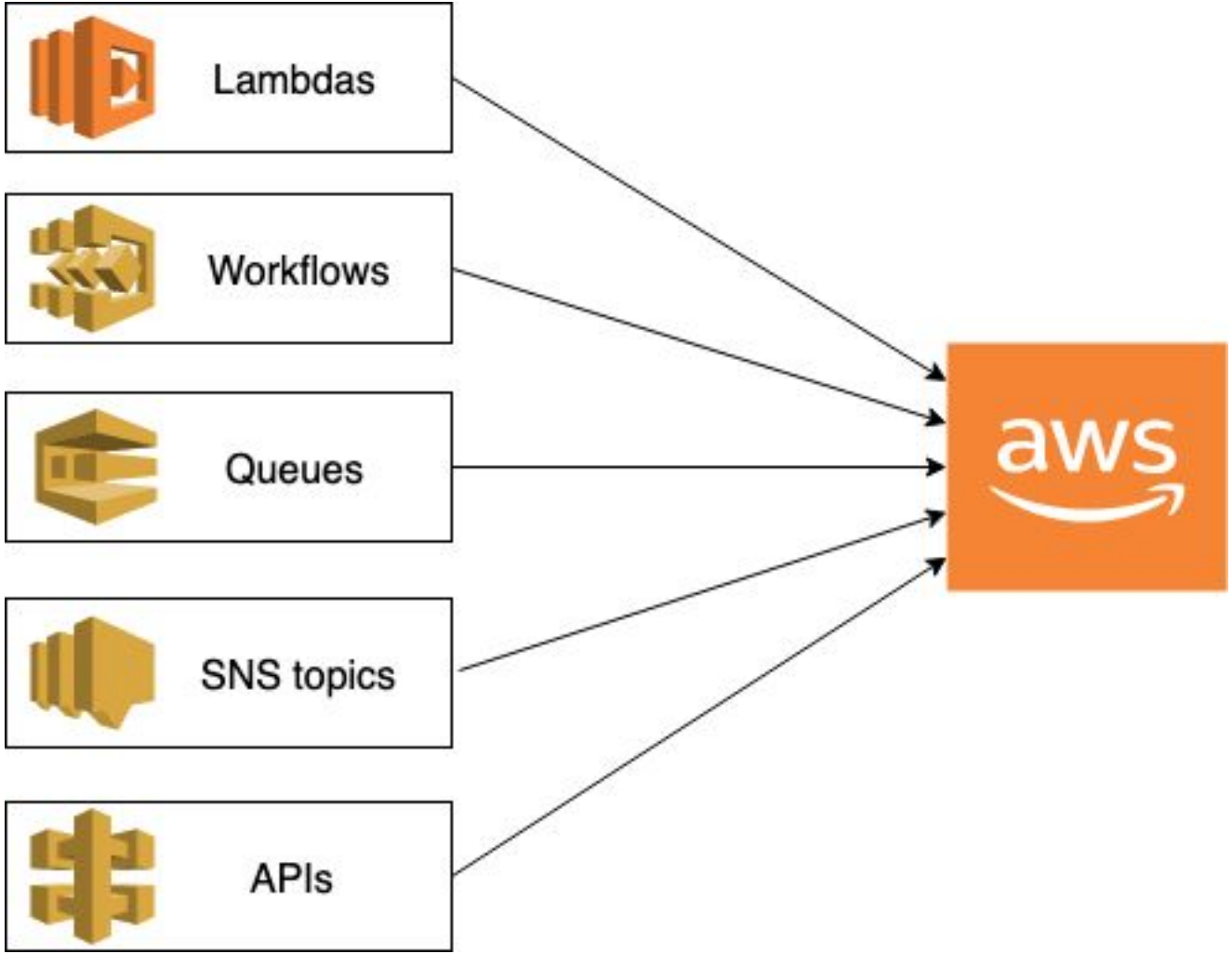
developmentSEED

Raytheon

Element 84

What is Cumulus?

Cumulus is a collection of resources for deploying and configuring a data pipeline in the cloud.


Lambdas
Workflows
Queues
SNS topics
APIs
aws

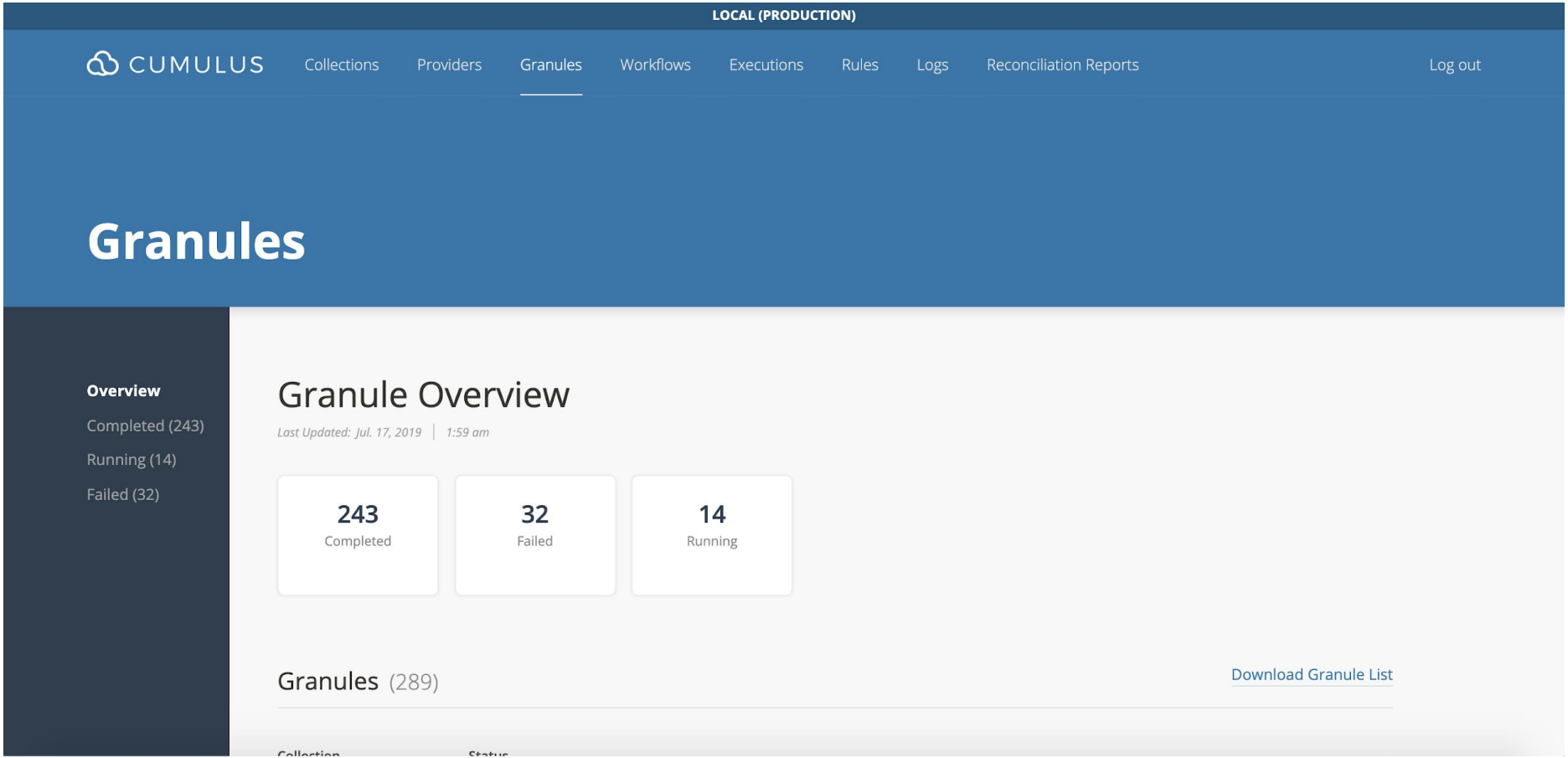
There is not one “Cumulus”

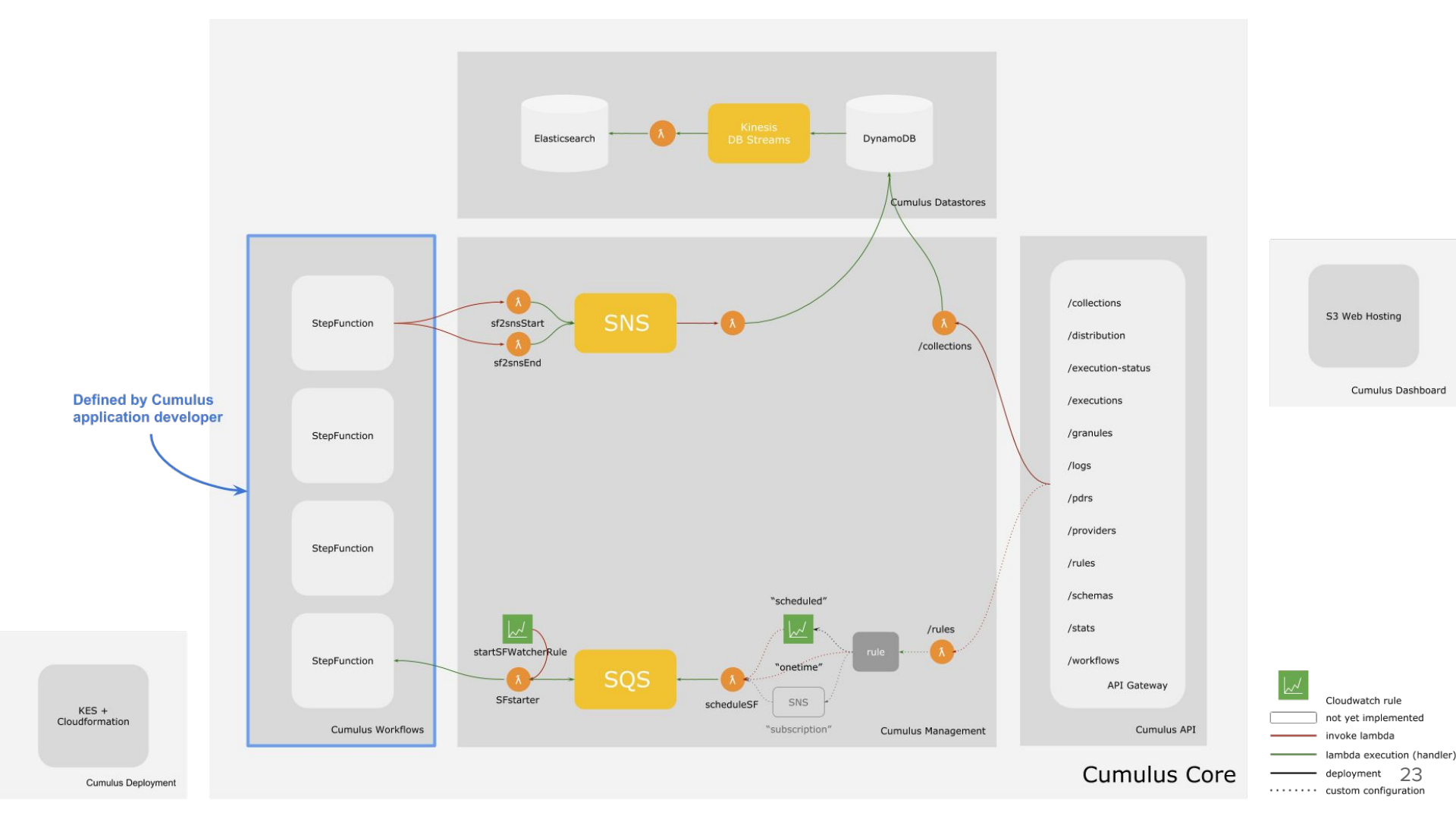
Core functionality

- Features a rich API for triggering, scheduling and monitoring workflows
- A set of core tasks for workflows
- AWS Serverless architecture
- Data persistence and API

Configurable features

- VPC support
- ECS autoscaling
- Cloudwatch rule triggers for Lambda
- Execution throttling


LOCAL (PRODUCTION)
CUMULUS
Collections
Providers
Granules
Workflows
Executions
Rules
Logs
Reconciliation Reports
Log out
Granules
Overview
Completed (243)
Running (14)
Failed (32)
Granule Overview
Last Updated: Jul. 17, 2019 | 1:59 am
243
Completed
32
Failed
14
Running
Granules (289)
Download Granule List

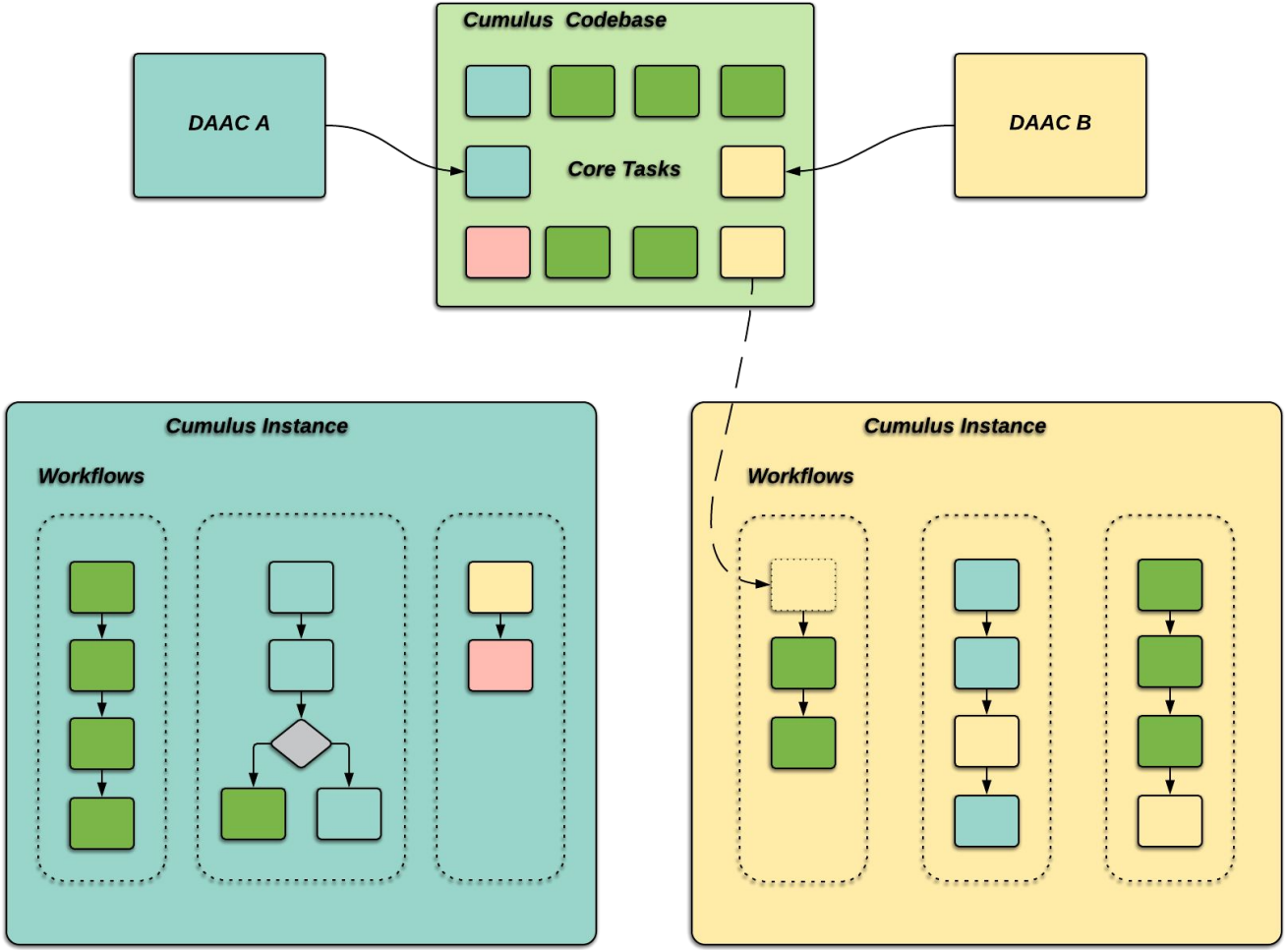

Elasticsearch
Kinesis
DB Streams
DynamoDB
Cumulus Datastores
StepFunction
StepFunction
StepFunction
StepFunction
Cumulus Workflows
Defined by Cumulus
application developer
sf2snsStart
sf2snsEnd
SNS
/collections
/distribution
/execution-status
/executions
/granules
/logs
/pdrs
/providers
/rules
/schemas
/stats
/workflows
API Gateway
Cumulus API
S3 Web Hosting
Cumulus Dashboard
startSFWatcherRule
SFstarter
SQS
scheduleSF
"scheduled"
"onetime"
SNS
"subscription"
rule
/rules
Cumulus Management
KES +
Cloudformation
Cumulus Deployment
Cumulus Core
Cloudwatch rule
not yet implemented
invoke lambda
lambda execution (handler)
deployment
custom configuration

Core resources

- **@cumulus/deployment:** A Node.js module for configuring and creating a Cumulus deployment.
- **@cumulus/api:** A Node.js module for deploying the Cumulus API and other AWS resources required to run Cumulus workflows.
- Node.js modules for common tasks to be run as part of Cumulus Workflows, for example **@cumulus/discover-granules**
- **cumulus-dashboard:** Code to generate and deploy the dashboard for the Cumulus API.

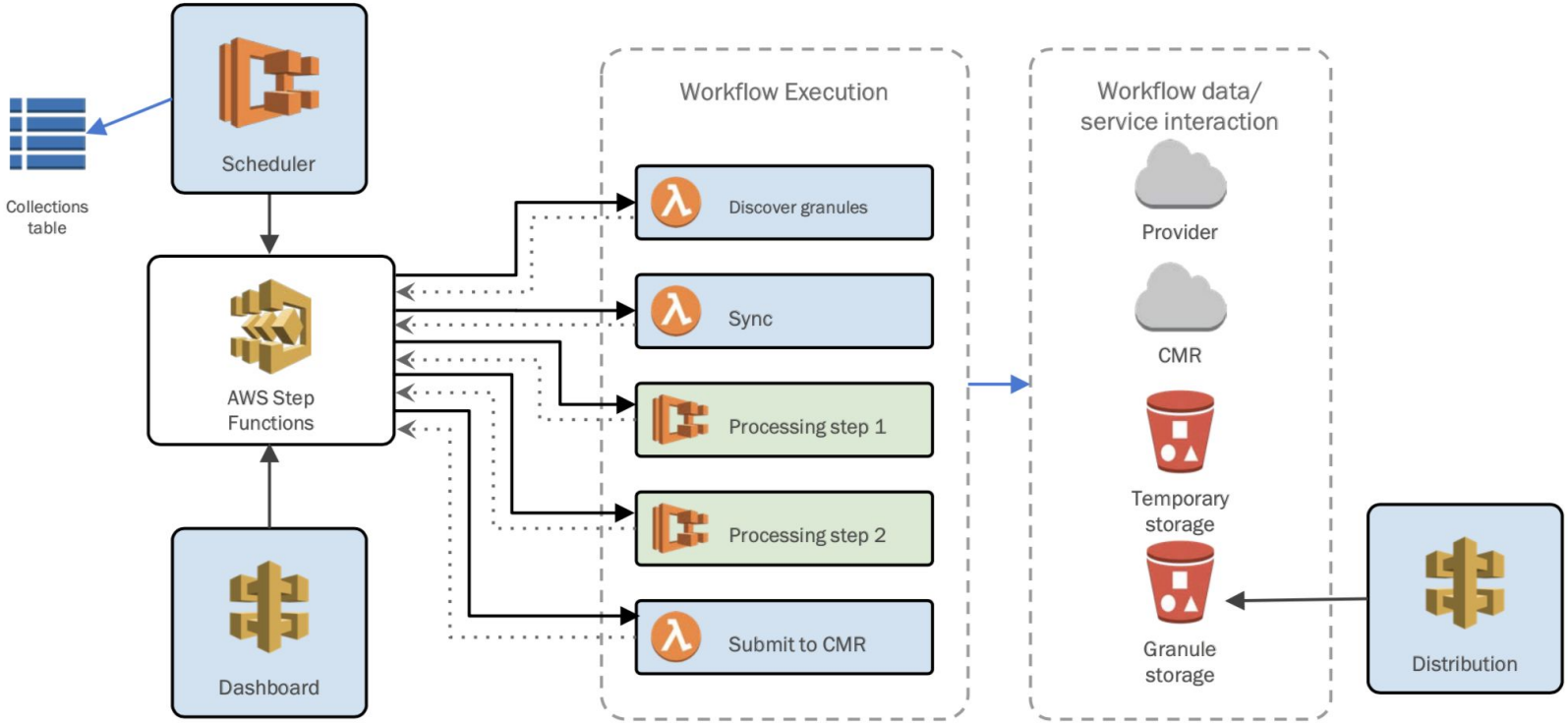
Contributed resources

- Thin Egress App (https://github.com/asfadmin/thin-egress-app/)
  - Provides an API to access files stored in the cloud
- CNM to Granule Lambda
- <insert_your_contribution_here>


Cumulus Codebase
DAAC A
DAAC B
Core Tasks
Cumulus Instance
Workflows
Cumulus Instance
Workflows

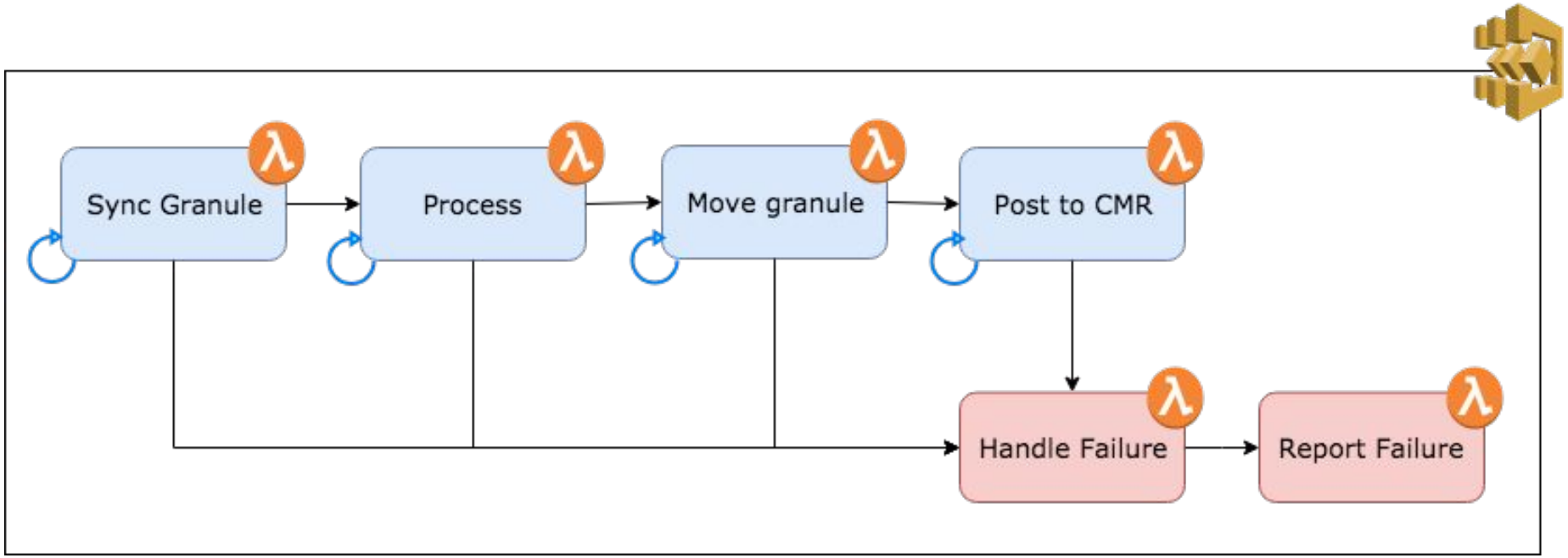
How does Cumulus work?

Ingest & Archive with AWS Step Functions

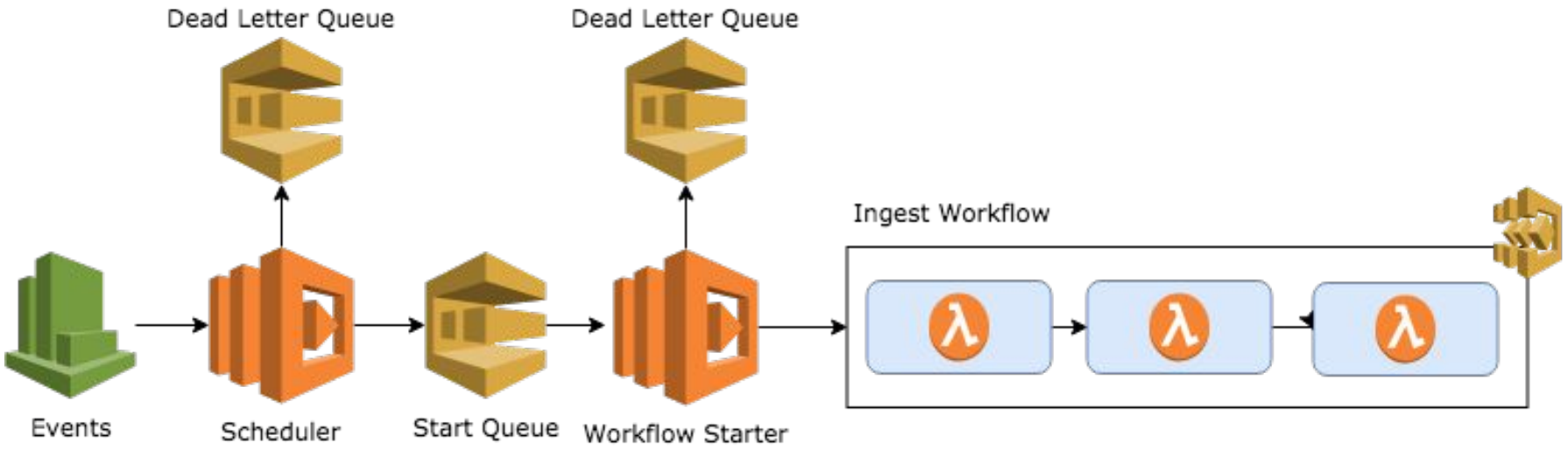


EOSDIS

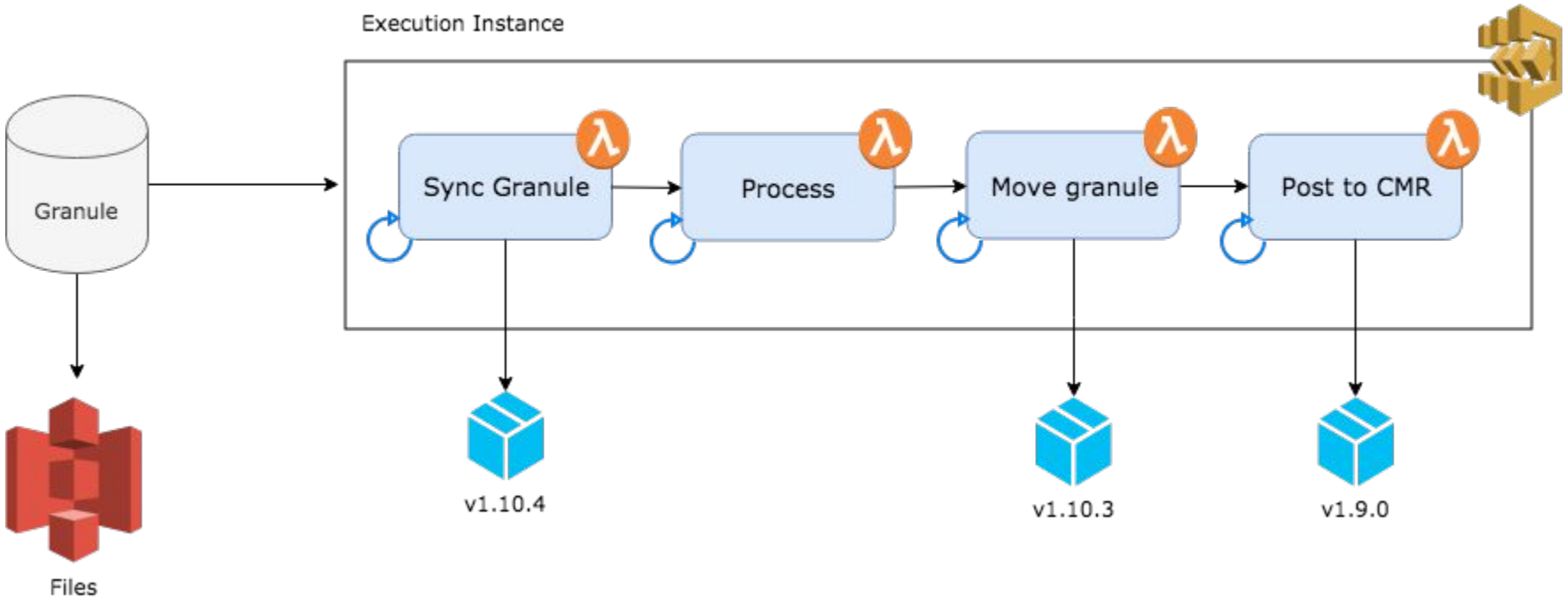
Ingest failures



Failures to trigger ingest



Data provenance



What is next for Cumulus?

Cumulus as an ecosystem

- Better contribution experience
- Module-based deployment via Terraform
- Solution for hosting contributed artifacts

Developer/user experience

- SNS topics for ingest completion
- Reduced dependency footprint
- Dashboard UX and design improvements

Where is Cumulus?

Check us out on Github!

https://github.com/nasa/cumulus

Read the documentation

https://nasa.github.io/cumulus/

Use our packages

https://www.npmjs.com/org/cumulus

Demo!

Key concepts

Provider - Providers generate and distribute input data that Cumulus obtains and sends to workflows

Granule - A granule is the smallest aggregation of data that can be independently managed (described, inventoried, and retrieved). A granule is a grouping of data files.

Collection - Collections are logical sets of data objects of the same data type and version.

Rules - Rules are used by to start processing workflows and the transformation process. Rules can be invoked manually, based on a schedule, or via triggers (Kinesis/SNS).

Questions?

Try it out!

https://github.com/nasa/cumulus-template-deploy

Thanks!

boyd@developmentseed.org