Next Generation NOTA Data Flow and Archiving Pipelines

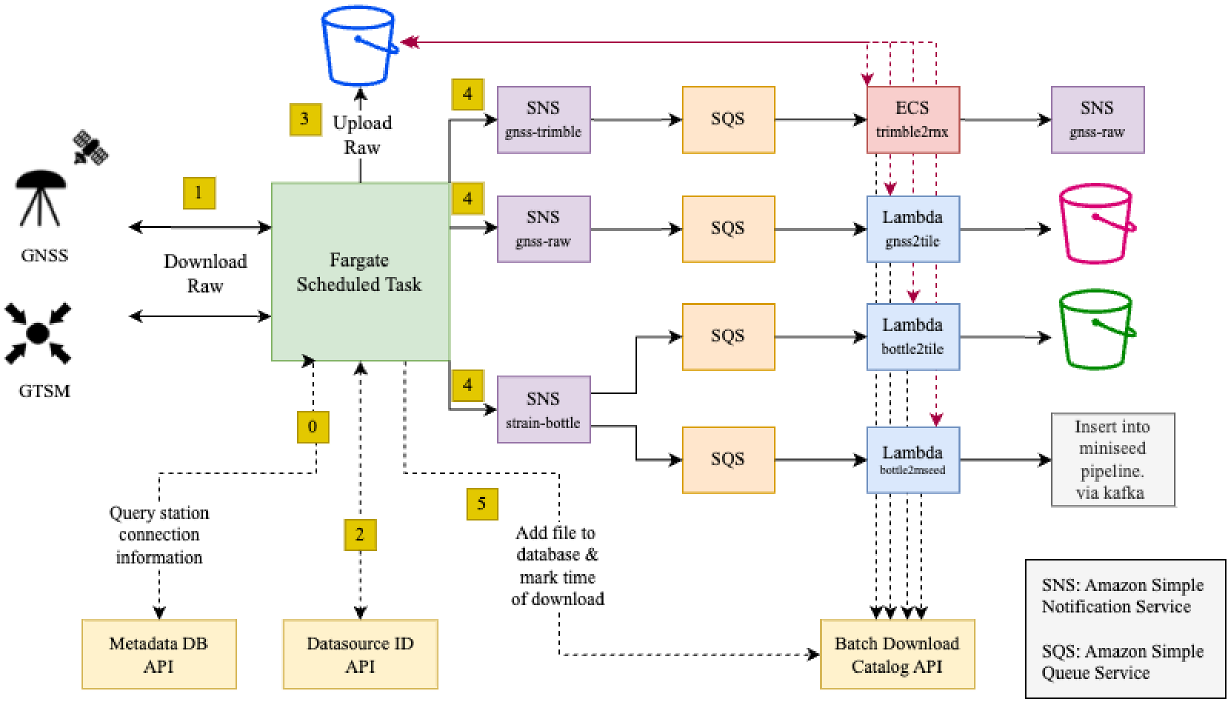
Rachel Terry (rachel.terry@earthscope.org), Michael Gottlieb, Henry Berglund, Rowan Gaffney,
Charlie Seivers, Douglas Ertz, Alex Hamilton, Mike Rost, David Maggert, Warren Gallaher
Affiliation: Earthscope

The National Science Foundation (NSF) funded Network of the Americas (NOTA) — managed by Earthscope — currently has more than 1,300 sensors, including GNSS (majority), strainmeters, tiltmeters, seismometers, and meteorological sensors. As Earthscope transitions from an on-premise data center to the cloud, the hourly and daily batch datafile download applications are being significantly refactored.

Containerized applications will run in AWS as fargate scheduled tasks, only using compute resources while running. The first step in the download process involves retrieving station connection metadata from a database. Previous downloads are tracked using a postgres database behind an API. Datafiles are downloaded via stream to an S3 bucket. File metadata and S3 location is pushed via Simple Notification Service (SNS), which is then read by one or more Simple Queue Service (SQS) queues.

We use SQS to accommodate back-pressure and support concurrent file processing with lambda functions. A lambda function is invoked for each message received in the SQS queue. Once invoked, the function reads a file from S3, deserializes the content, and writes the data to a TileDB array in our archive. Files downloaded from Trimble GNSS stations require an additional preprocessing step. Trimble data are submitted to a separate queue for conversion into a non-proprietary format (e.g. RINEX) before being submitted to the archival queue.

In this poster, we present the state of Earthscope's prototype batch download system and preliminary designs for community data and event submissions. From remote station downloads through data archival, our new prototype is currently running in AWS with a subset of stations. We believe that this new system will allow us to accommodate a variety of use cases, including internal downloads, and that the process for all submissions will follow the same SNS, SQS, lambda to TileDB archive path.


GNSS
GTSM
1
Download
Raw
3
Upload
Raw
Fargate
Scheduled Task
0
Query station
connection
information
2
5
Add file to
database &
mark time
of download
Metadata DB
API
Datasource ID
API
4
SNS
gnss-trimble
4
SNS
gnss-raw
4
SNS
strain-bottle
SQS
SQS
SQS
SQS
ECS
trimble2rnx
Lambda
gnss2tile
Lambda
bottle2tile
Lambda
bottle2mseed
SNS
gnss-raw
Insert into
miniseed
pipeline.
via kafka
Batch Download
Catalog API
SNS: Amazon Simple
Notification Service
SQS: Amazon Simple
Queue Service