Cloud Scalable Omics Data Analysis Pipeline using Serverless Task Infrastructure

Hugo Kitano*, Iman Mohtashemi, Mathew Ellenberger, Jessica Chu, Gabriel Castro, Taher Elgierari, Marwin Ko, Theo Platt, Asim Siddiqui and Serafim Batzoglou

An automated, scalable proteomics data analysis workflow

Liquid Chromatography coupled with Mass Spectrometry (LCMS) is the premier detection technology for comprehensive proteomics analysis of complex samples, yet its commonly-used data analysis tools are not built for scalability into the future.

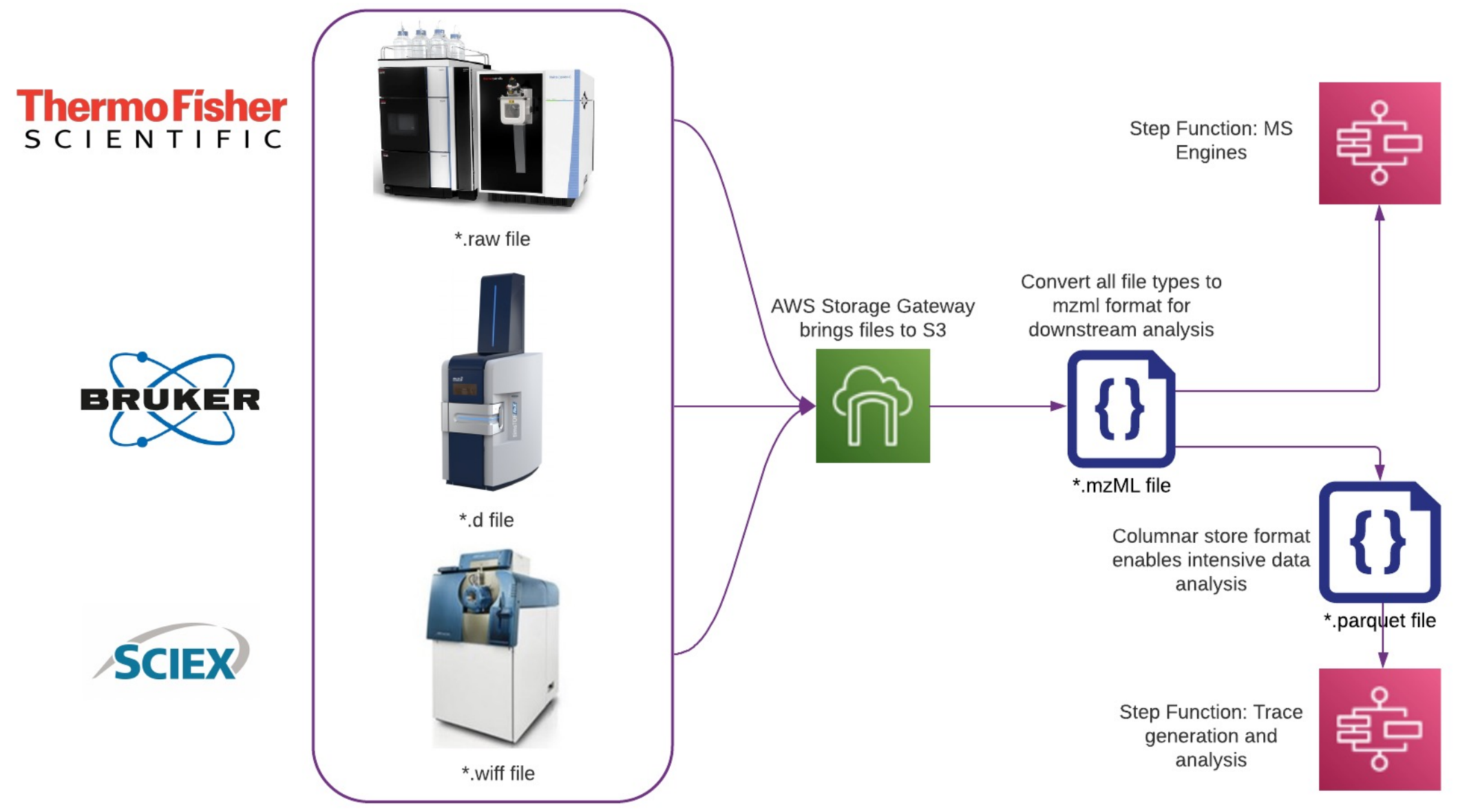
Challenges

- Most LCMS applications are built for desktop environments, some even only work on Windows environments
- Different vendors use different filetypes that require different processing modules
- Applications are not designed for increasing compute and memory
- There is a need to modularize the ever-growing collection of applications for both DDA- and DIA- acquired LCMS proteomics data

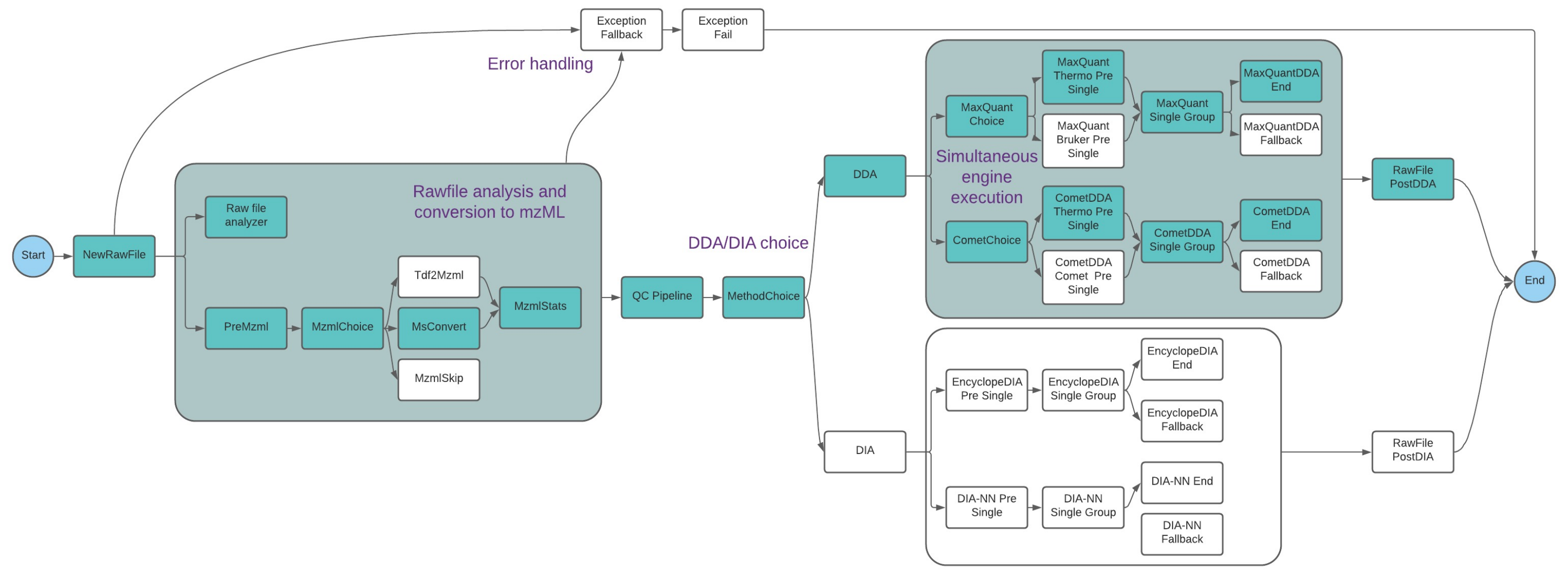
Solution

A carefully curated AWS proteomics data analysis workflow with choices, error handling, and exception fallbacks including:

- **Automated file transfer** to the cloud and **conversion** to the standard mzML filetype
- **Automate single file analysis** for every injection upon raw data file arrival
- User-specified **group run analyses** with pre-defined recipes and settings (possible with 100s of files)


ThermoFisher SCIENTIFIC
BRUKER
SCIEX
*.raw file
*.d file
*.wiff file
AWS Storage Gateway brings files to S3
Convert all file types to mzml format for downstream analysis
*.mzML file
Step Function: MS Engines
Columnar store format enables intensive data analysis
*.parquet file
Step Function: Trace generation and analysis

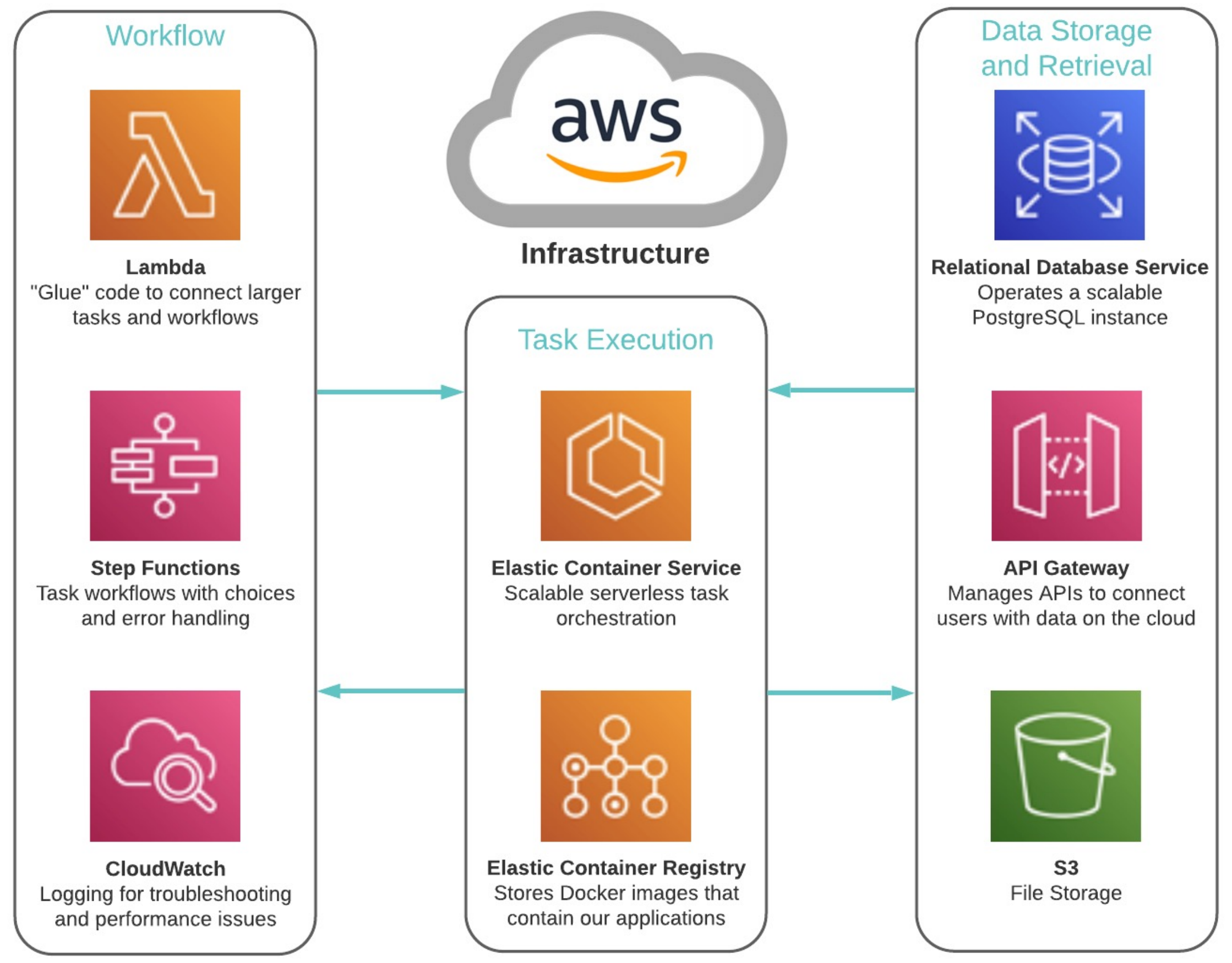
Step Function workflow for the automated analysis of raw files



Proteograph analysis suite with a smart cloud infrastructure

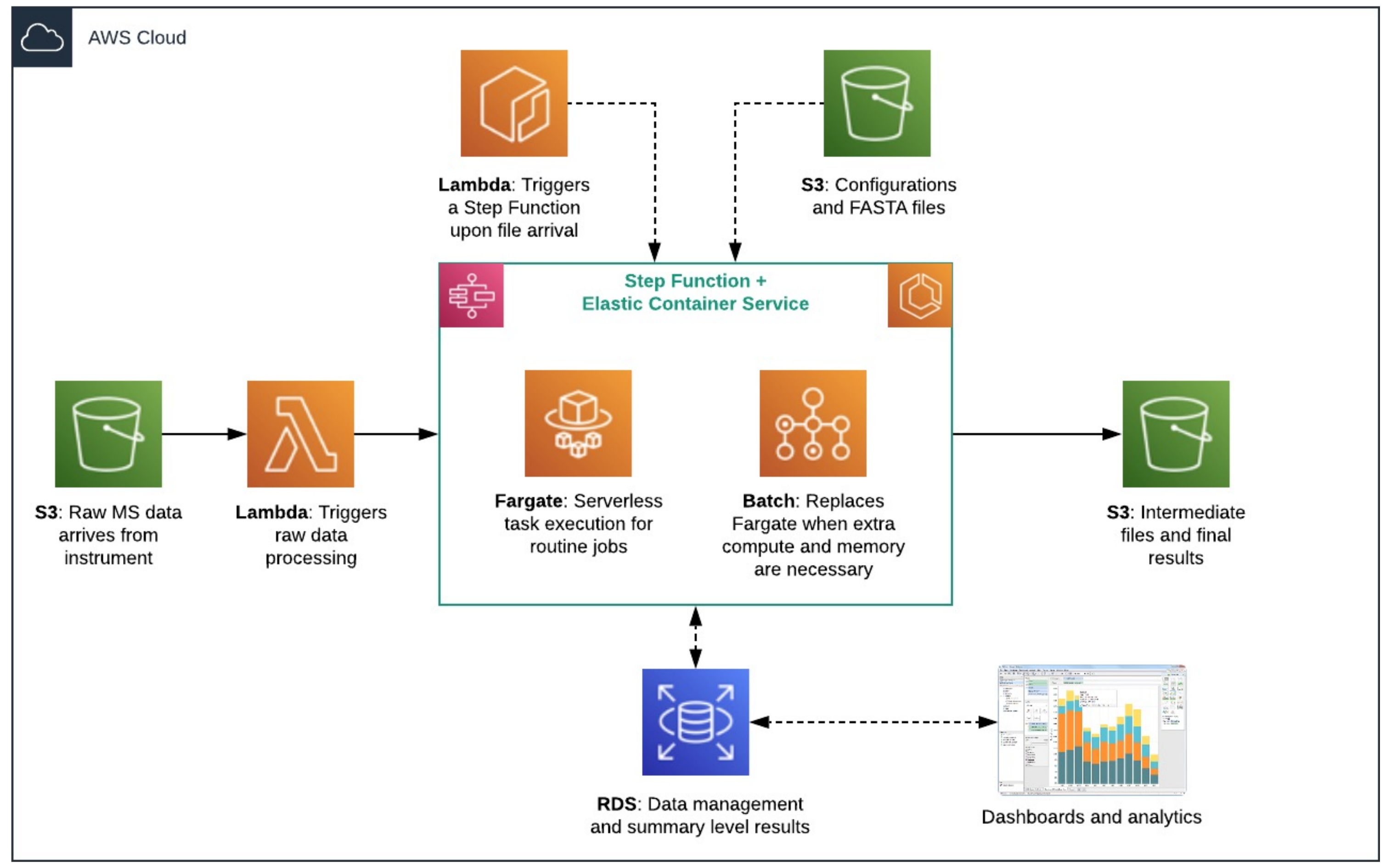
A combination of AWS services to process, store, and retrieve data

The AWS ecosystem at Seer


Workflow
aws
Infrastructure
Data Storage and Retrieval
Task Execution
Lambda
"Glue" code to connect larger tasks and workflows
Step Functions
Task workflows with choices and error handling
CloudWatch
Logging for troubleshooting and performance issues
Elastic Container Service
Scalable serverless task orchestration
Elastic Container Registry
Stores Docker images that contain our applications
Relational Database Service
Operates a scalable PostgreSQL instance
API Gateway
Manages APIs to connect users with data on the cloud
S3
File Storage

Multiple cloud services working in harmony

The coordination of automated file analysis from instrument to data storage

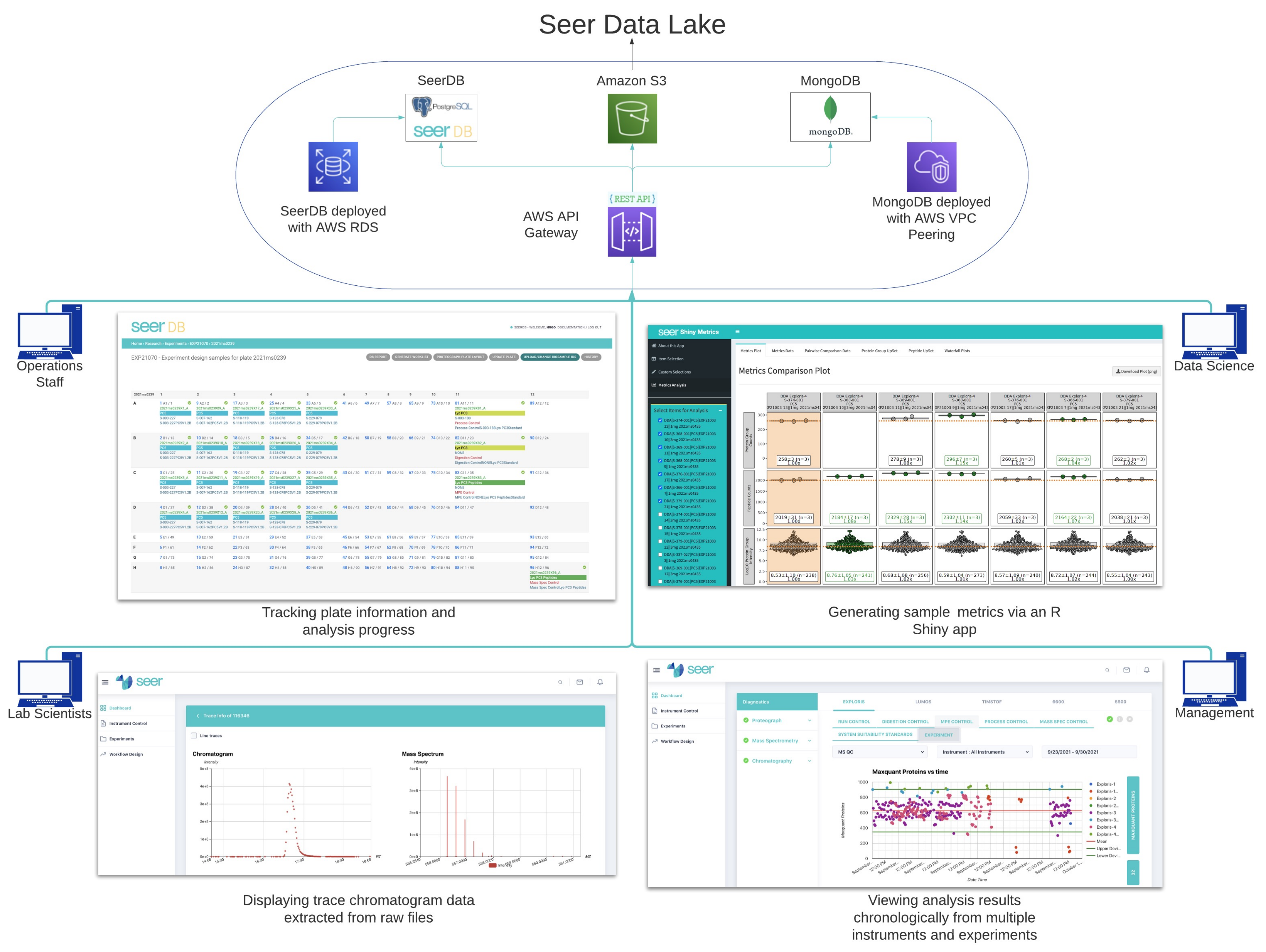

AWS Cloud
Lambda: Triggers a Step Function upon file arrival
S3: Configurations and FASTA files
Step Function + Elastic Container Service
S3: Raw MS data arrives from instrument
Lambda: Triggers raw data processing
Fargate: Serverless task execution for routine jobs
Batch: Replaces Fargate when extra compute and memory are necessary
S3: Intermediate files and final results
RDS: Data management and summary level results
Dashboards and analytics

Seer data lake, an actionable hub for robust Large scale proteomics analysis

To organize our results, we adhere to the principle of polyglot persistence, where differently structured data is stored in different types of databases to best suit our needs.

- Highly structured experimental data in a relational PostgreSQL database (SeerDB)
- Instrument readings and quality control data (largely unstructured) in non-relational MongoDB
- APIs and various internal apps to query both datastores and return information collectively

API design: connecting teams to the data



Tracking plate information and analysis progress

Generating sample metrics via an R Shiny app

Displaying trace chromatogram data extracted from raw files

Viewing analysis results chronologically from multiple instruments and experiments

Results

A next-generation platform capable of analyzing thousands of samples in hours supporting fleets of LCMS instruments

- Supporting hundreds of terabytes of incoming LCMS data annually
- 150 files with 140 AWS Batch jobs, and 2600 AWS Fargate tasks currently analyzed per day
- Future infrastructure will support massively parallel group run contexts

Publications

Seer, Inc., Redwood City, CA – hkitano@seer.bio