AWS Grant Application - Batch Processing with ndmg

Johns Hopkins University
Department of Biomedical Engineering
Alex Loftus
aloftus2@jhu.edu

June 21, 2019

Brief description of problem to be solved

Independent neuroscience labs around the world regularly generate huge neuroimaging datasets, many of which contain multiple hundreds of scans on the order of 50 MB per scan. Our subfield of neuroimaging, connectomics, concerns itself with generating ”connectomes” from raw neuroimaging datasets: e.g., dividing the brain into a set of subregions, and then creating a map of the connectivity between these subregions. Analyzing these data, however, is prohibitively time-consuming, labor-intensive, and requires a substantial degree of expertise. Current state-of-the-art analysis pipelines typically take on the order of a day to complete, and are not end-to-end: a substantial amount of manual data-wrangling and processing must currently be done to generate connectomes.

To combat these challenges, we created ndmg: NeuroData's MRI to Graphs, an end-to-end pipeline for processing the data contained within a single scan to a fully processed connectome. Our tool processes a scan on the order of 20 minutes, without any need for user intervention. However, analyzing large datasets has proven to be challenging: a great deal of computational power is necessary to process a large amount of scans, and our current resources have proven ineffective.

The code for our tool is stored in the following link: https://github.com/neurodata/ndmg

Proposed AWS Solution (including specific AWS tools, timeline, key milestones)

We have set up the infrastructure to run many scans in parallel using AWS batch. We first pull in data stored on an S3 bucket. We then run a set of Docker containers on EC2 instances, using a Docker image stored on Amazon ECR. Once processing is complete for each scan, we push the processed data back on to S3.

AWS services to be used:

- Batch: for processing scans in parallel
- S3: for storing raw and processed data
- EC2: for performing the computation
- ECR: for storing a Docker container with our processing pipeline

Timeline and key milestones:

- 1 month: set up batch infrastructure for running neuroimaging datasets.

- 3 months: 20+ open-source neuroimaging datasets run on our pipeline using AWS.
- 6 months: open-source service on which other neuroscientists can analyze their neuroimaging datasets.

Plan for sharing outcomes (tools, data, and/or resources) created during project

All of our code is open-source, and will be shared on github and dockerhub at github.com/neurodata/ndmg/ and hub.docker.com/u/neurodata respectively. Additionally, we host both raw datasets and the processed results of our datasets on our website, https://www.neurodata.io/mri-cloud.

Potential future use of AWS beyond grant duration by individual research group or broader community

Researchers in the neuroimaging community will use ndmg for quick and easy analysis of their datasets. This analysis will occur using the AWS infrastructure that we set up. Additionally, our lab operates at the intersection between neuroscience and data science. Many researchers in the broader field of network analysis are interested in connectomes, simply because they constitute an interesting source of real-world data. Therefore, the outputs of our pipeline will be of interest to these data scientists as well.

Any AWS Public Datasets to be used in your research

Currently, all of the datasets we will be using are open neuroimaging datasets shared by the neuroimaging community; we do not make use of any AWS Public Datasets at this time.

Keywords to faciliate proposal review

Neuroscience, Graph Theory, Network Analysis, Connectomics, Machine Learning, Neuroimaging, fMRI, dMRI

Cost Estimate

https://bit.ly/2WTm3K5