3535 Thursday 12:45

Harnessing cloud computing for high capacity analysis of neuroimaging data from NDAR

Daniel Clark[1], Christian Haselgrove[2], David Kennedy[2], Zhizhong Liu[3], Michael Milham[1], Petros Petrosyan[4], Carinna Torgerson[3], John Van Horn[3], Cameron Craddock[1]

[1]Child Mind Institute, New York, NY, [2] University of Massachuttes Medical School, Worcester, MA, [3]University of Southern California, Los Angeles, CA, [4]UCLA, Los Angeles, CA, [5]Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY

Introduction

- The National Database for Autism Research (NDAR) hosts a vast collection of neuroimaging datasets that can be processed and utilized to yield significant scientific discoveries.
- This amount of resources necessitates a high-performance computing (HPC) infrastructure, which is not always readily available for researchers in-house.
- Amazon Web Services (AWS) Elastic Compute Cloud (EC2) computing service offers a "pay as you go" model that allows researchers to utilize HPC performance without the up-front captial costs and maintenance of an in-house solution.
- The developers of the Laboratory of Neuro Imaging (LONI) Pipeline, the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) Computational Environment (CE) and the Configurable Pipeline for the Analysis of Connectomes (C-PAC) have implemented pipelines in EC2 that interface with NDAR

Methods

LONI Pipeline

- The LONI Pipeline software was extended to include new pipeline modules to access data from the NDAR database, transfer input data out of Amazon S3 (Simple Storage Service), and to load results back into S3[1]
- A pipeline was constructed to extract cortical thickness and subcortical region volume data from structural MRI images in the NDAR database, which included:
  1. Reorient images to standard orientation using FSL's reorient2std module
  2. Extract cortical thickness using FreeSurfer recon-all
  3. Calculate volumes of subcortical regions using FSL's BET and FIRST all
- The resulting pipeline was used to process 780 T1-weighted structural images and return the results to NDAR

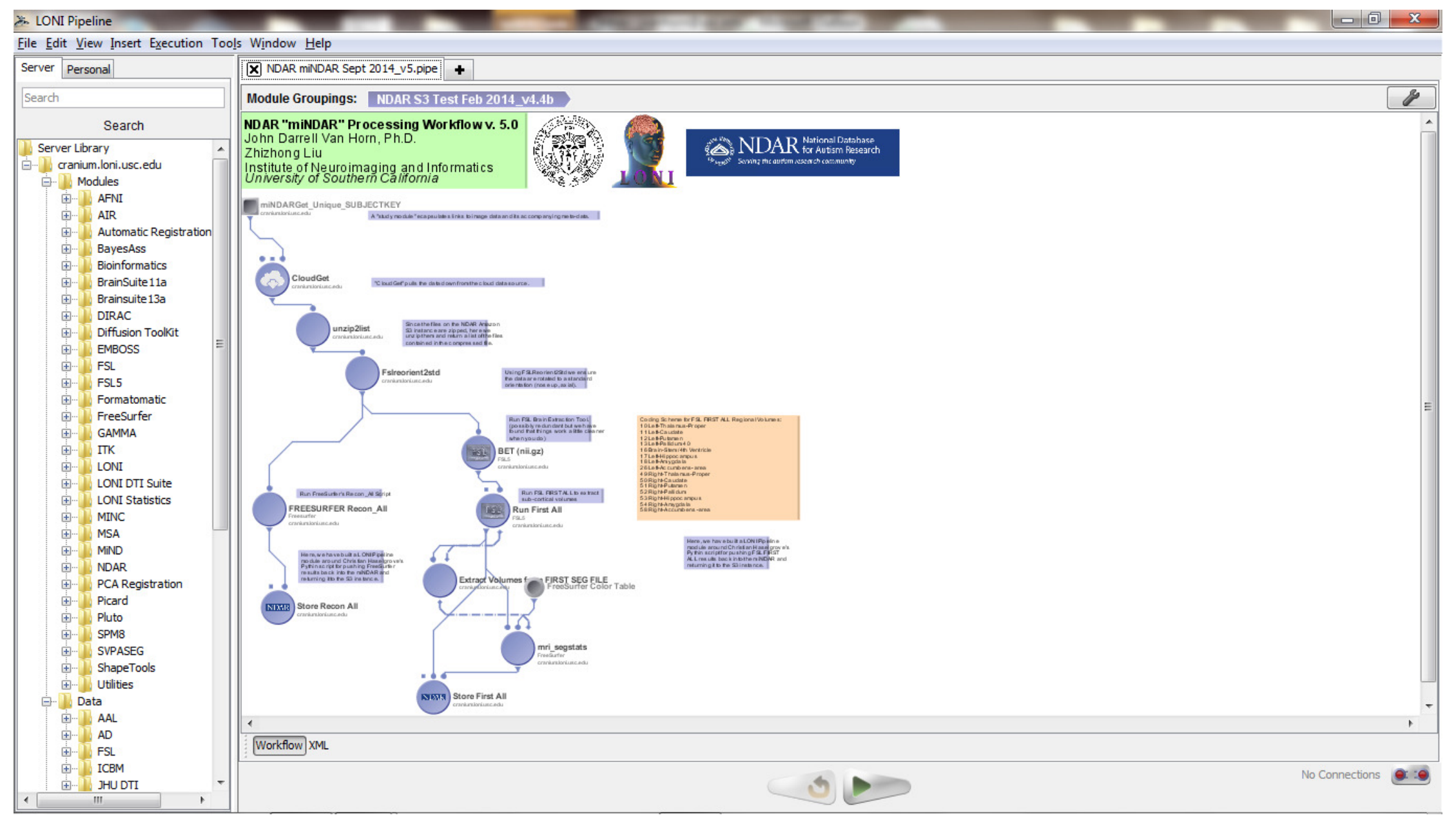


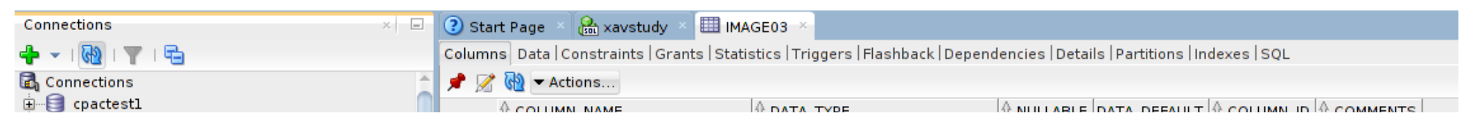
Figure 1 : Graphical layout of the constructed pipeline

Configurable Pipeline For the Analysis of Connectomes (C-PAC)

- C-PAC modules were written in Python to build input data lists by querying NDAR, read input data from S3, write processed results to S3 and write values back to the NDAR database[2]
- New pipelines were created to perform the ANTS cortical thickness[3] procedure and the Preprocessed Connectomes Project's Quality Assessment Protocol (`http://preprocessed-connectomes-project.github.io/quality-assessment-pipeline`)[4]
  1. Reorient images to standard orientation using FSL's reorient2std module
  2. Extract cortical thickness using FreeSurfer recon-all
  3. Calculate volumes of subcortical regions using FSL's BET and FIRST all
- The resulting pipeline was used to process 2,085 T1-weighted structural images and return the results to NDAR

Neuroimaging Informatics Tools and Resources Clearinghouse Computational Environment (NITRC-CE)

- The NITRC pipeline processed data using three primary utilities
  1. Extract anatomical and surface-base measures with Freesurfer recon-all
  2. Segment subcortical structures using FSL's FIRST to produce volumetric and mesh outputs
  3. Time series QA measures using the fmriqa_generate.pl utility from the BXH/XCEDE Tools suite, including mean intensty, center of mass, per-slice variation, and others
- Python modules were created to query and download data from NDAR as well as to store results back to their database
- The recon-all and FIRST tools processed 986 and 1,247 T1-weighted anatomical scans, respectively; the fmriqa_generate QA measures were generated from 1,349 functional scans.



Results

Table 1 : Processing completed as a part of the effort. Nodes corresponds to the number of hosts used in the calculation. PF is parallelization factor and corresponds to the number of jobs ran in parallel on each node. On demand instances were used for the master node and spot instances were used for all computation nodes in the cluster. CPU Time is the total amount of time required to perform the computation and Wall Time is the amount of time that passed. # DS: Number of datasets. CPD: Cost Per Dataset. C-PAC: Configurable Pipeline for the Analysis of Connectomes. NITRC-CE: NITRC Computational Environment

Processing	# DS	Platform	Nodes	PF	CPU Time	Wall Time	Cost	CPD
ANTS Cortical Thickness	3197		20	8	23,018	147	$760.24	$0.24
Resting state fMRI processing w/ 4 strategies	1112	C-PAC	20	3	834	22	$80.54	$0.07
Quality Assessment Protocol	1112		20	4	380	14	$19.02	$0.02
Freesurfer recon-all	986		4	32	23,664	193	$211.44	$0.21
FSL FIRST	1247	NITRC-CE	4	32	208	3	$2.19	> $0.01
Temporal QA	1349		4	32	450	13	$4.69	> $0.01
Freesurfer recon-all and FSL FIRST	780	LONI Pipeline	20	32	18,720	49	$252.36	$0.32

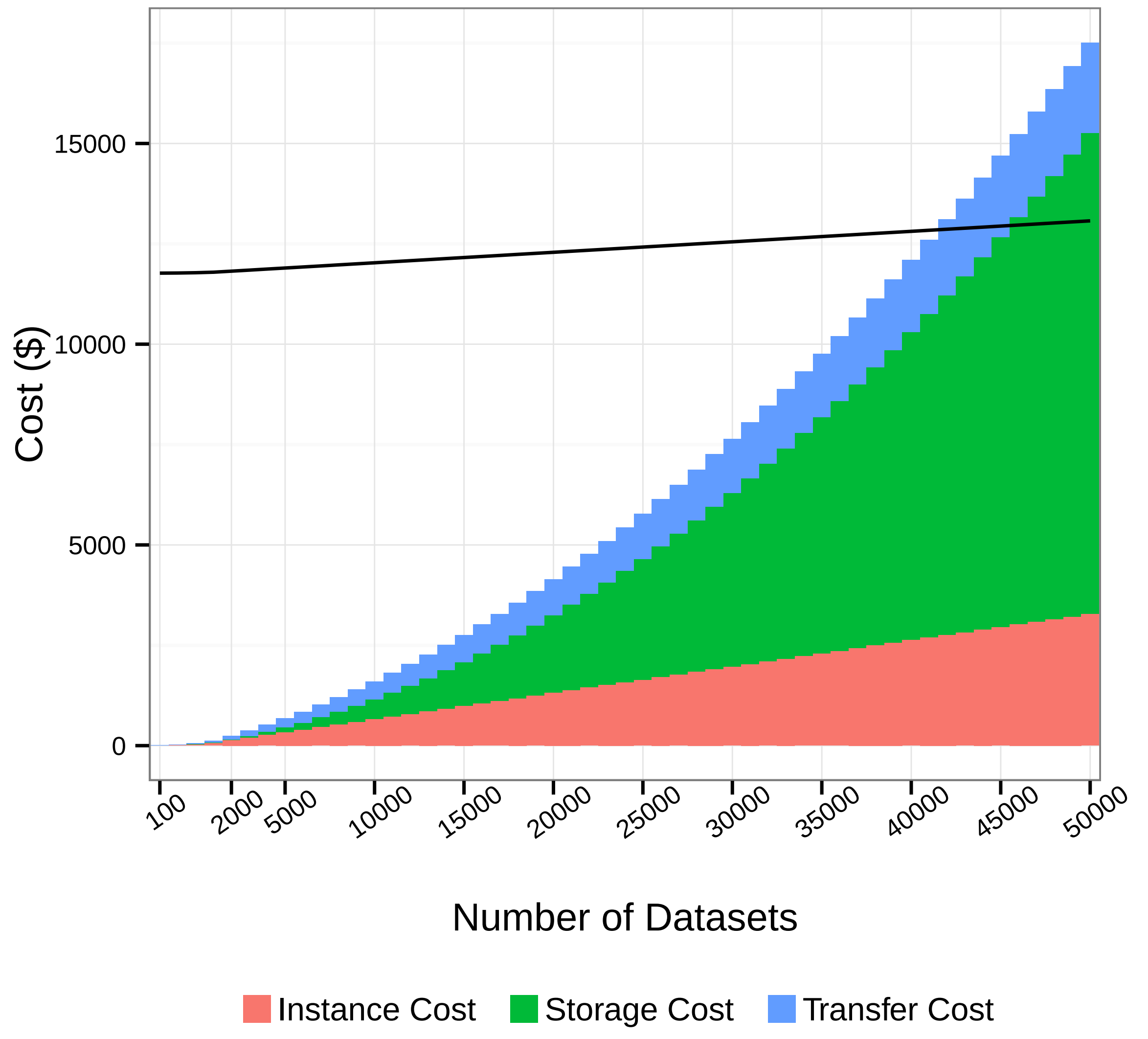
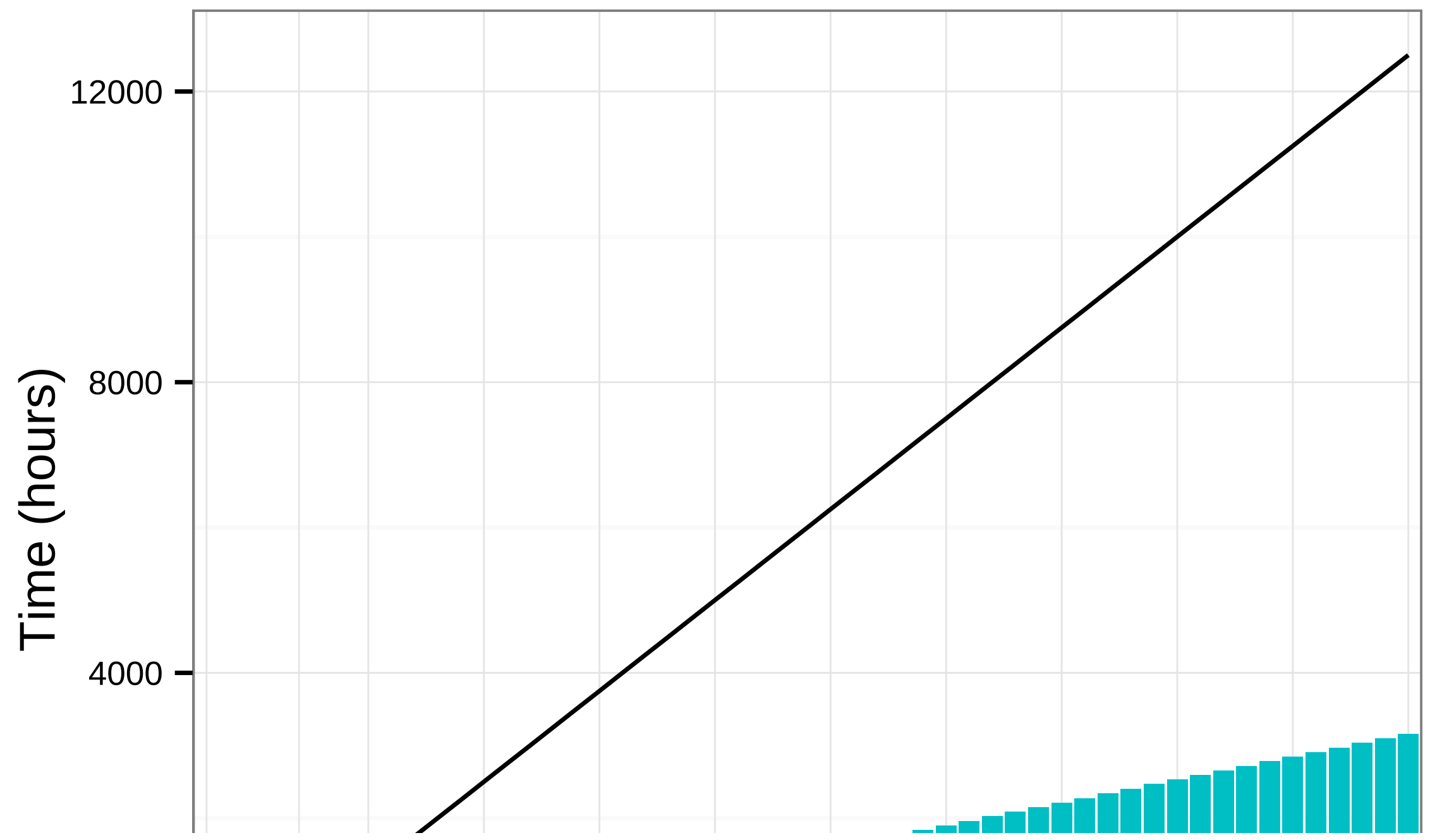


Figure 4 : AWS EC2 costs, grouped by cost type, for a typical C-PAC pipeline for different sized datasets versus owning and maintaining own server

Benefits of the cloud

- The scalability and payment model for processing this data on a cloud platform allowed for an efficient processing of the NDAR datasets, both in time and cost
- Cloud computing services like AWS EC2 are viable solutions for the processing and analysis of large amounts of data
- Cost and time models were produced to demonstrate the benefits of using cloud computing for increasingly large datasets



cameron.craddock@childmind.org