How to do Large Scale Data Research on a Slurm HPC Cluster with OmniSci

OmniSci Virtual Summit, May 2020

Devika Kakkar and Ben Lewis

Harvard Center for Geographic Analysis

Raminder Singh

Harvard Research Computing

Center for
Geographic Analysis
Harvard University

Spatiotemporal Thinking, Computing and Applications
STC

NSF

omni
sci

FAS RC

FAS RESEARCH COMPUTING
HARVARD UNIVERSITY
FACULTY OF ARTS & SCIENCES

+

omni
sci

Spatiotemporal Thinking, Computing and Applications
STC

Center for
Geographic Analysis
Harvard University

IQSS
The Institute for Quantitative Social Science

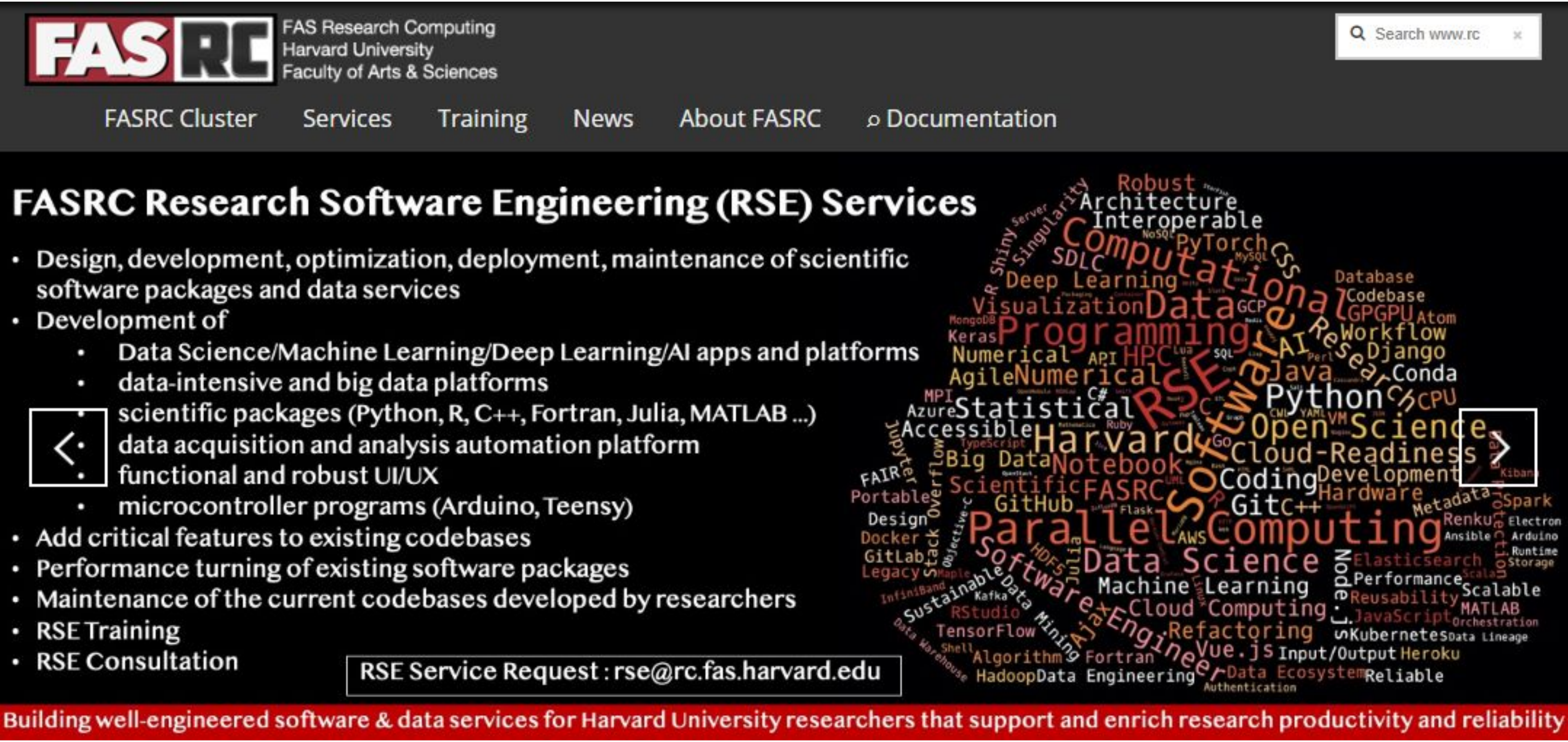
NSF Collaboration between OmniSci and CGA

Industry & University Cooperative Research Center (I/UCRC)

Center for Spatiotemporal Thinking, Computing, and Applications

Harvard University

George Mason University



FASRC Research Software Engineering (RSE) Services

- Design, development, optimization, deployment, maintenance of scientific software packages and data services
- Development of
  - Data Science/Machine Learning/Deep Learning/AI apps and platforms
  - data-intensive and big data platforms
  - scientific packages (Python, R, C++, Fortran, Julia, MATLAB ...)
  - data acquisition and analysis automation platform
  - functional and robust UI/UX
  - microcontroller programs (Arduino, Teensy)
- Add critical features to existing codebases
- Performance turning of existing software packages
- Maintenance of the current codebases developed by researchers
- RSE Training
- RSE Consultation

RSE Service Request : rse@rc.fas.harvard.edu

Building well-engineered software & data services for Harvard University researchers that support and enrich research productivity and reliability

Geotweet Archive:
A global social media record spanning time, geography, and language:

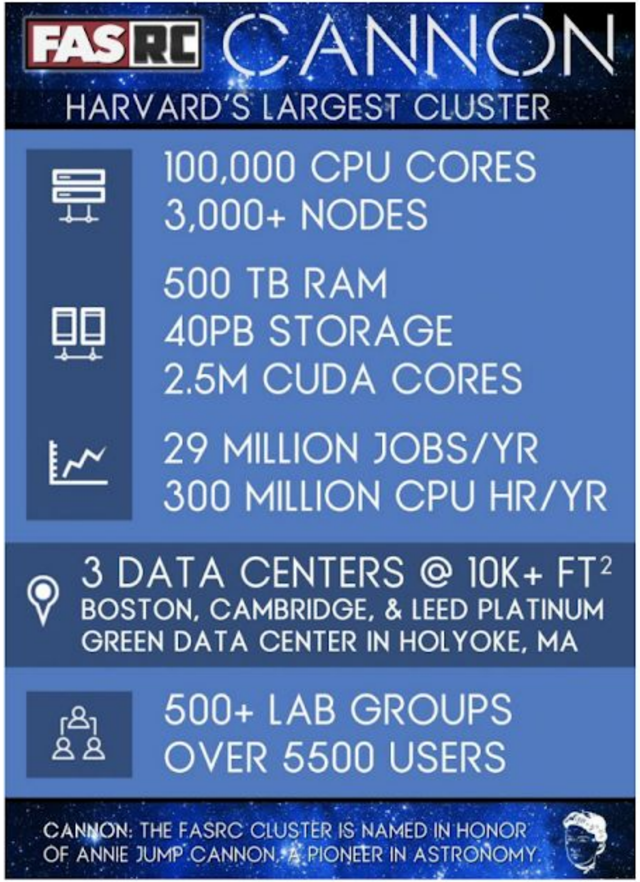
- Developed in collaboration with the University of Salzburg Department of Geoinformatics
- Extends from 2010 to the present and updated daily
- Geotagged by GPS or user designated place name

for more information:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F3NCMB6

Introduction to Harvard’s FAS Research Computing Cluster

- Compute: 100,000 compute nodes, 8-64 cores/node, 12Gb to 512Gb memory/node, 2,500,000 NVIDIA GPU cores
- Software: CentOS 7 operating system, Slurm job manager, Singularity, 1000+ scientific tools and programs
- Storage: 100 GB (Home dir), 4TB+ (Lab storage), 70Gb/node (Local scratch), 2.4PB (Global scratch), 3PB (Persistent Research data)
- #144 in TOP500 Supercomputers in world



FASRC CANNON

HARVARD’S LARGEST CLUSTER

100,000 CPU CORES
3,000+ NODES

500 TB RAM
40PB STORAGE
2.5M CUDA CORES

29 MILLION JOBS/YR
300 MILLION CPU HR/YR

3 DATA CENTERS @ 10K+ FT^2
BOSTON, CAMBRIDGE, & LEED PLATINUM
GREEN DATA CENTER IN HOLYOKE, MA

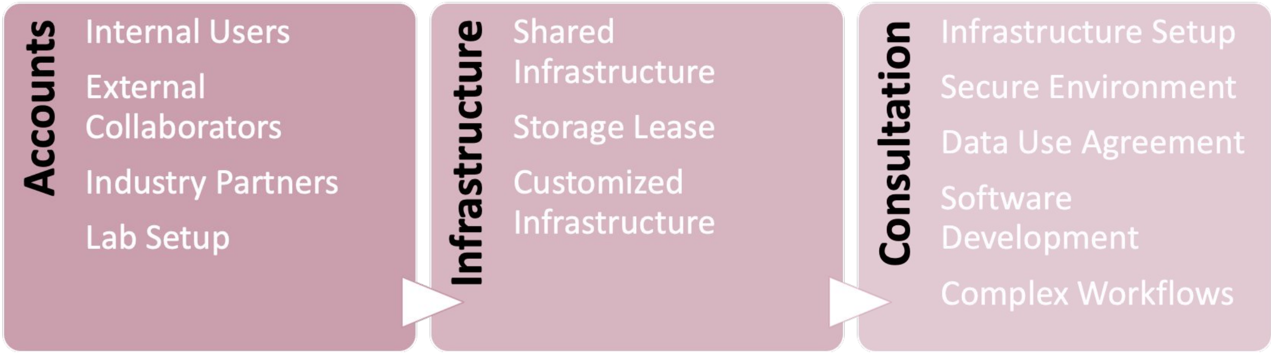
500+ LAB GROUPS
OVER 5500 USERS

CANNON: THE FASRC CLUSTER IS NAMED IN HONOR OF ANNIE JUMP CANNON, A PIONEER IN ASTRONOMY.

FASRC Services

Software:

- Operating System CentOS 7
- 1,000+ scientific tools and programs
  - https://portal.rc.fas.harvard.edu/apps/modules
- C, C++, Fortran and Intel compilers available
- Languages like Python, R and Julia etc. can be used
- Databases like MySQL, Postgres and MongoDB



Accounts

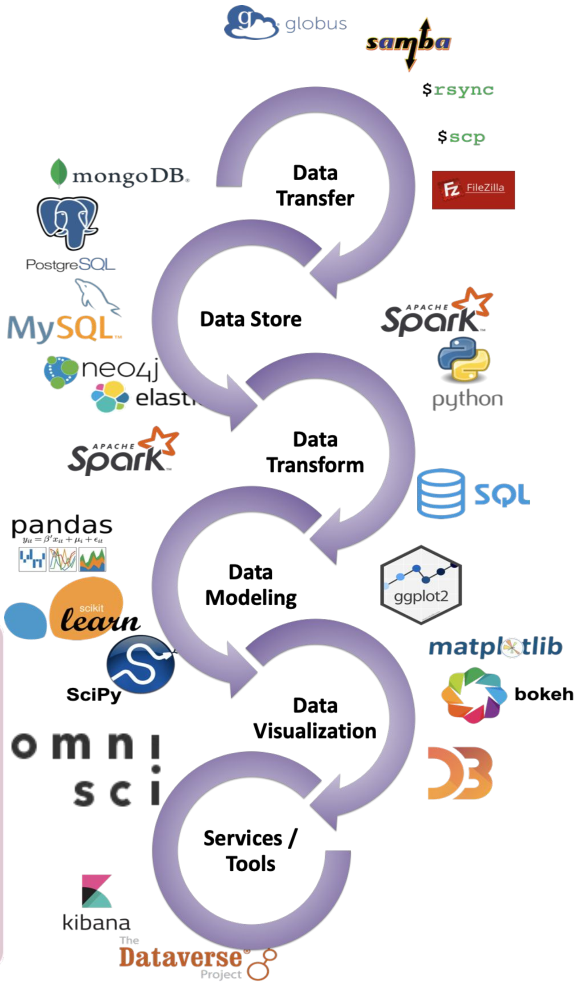
- Internal Users
- External Collaborators
- Industry Partners
- Lab Setup

Infrastructure

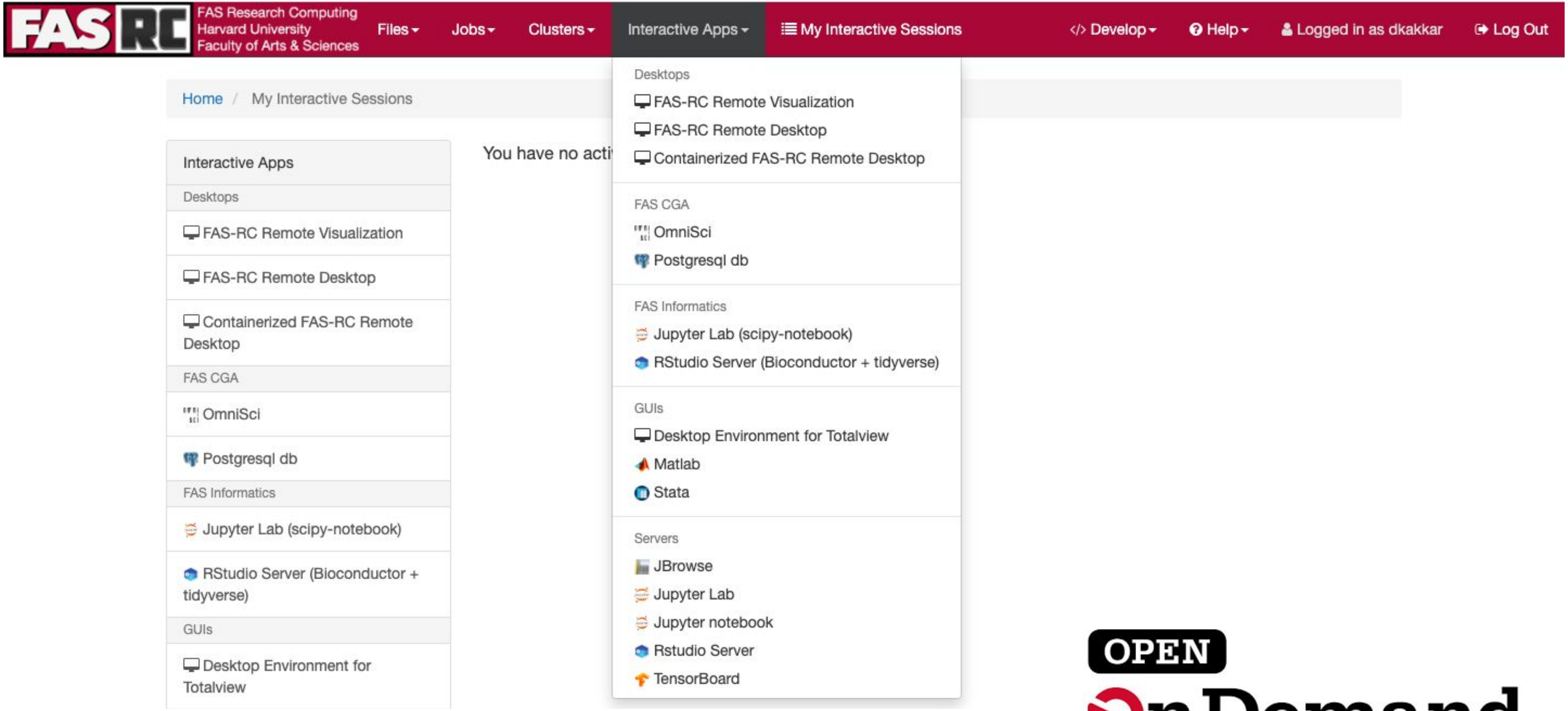
- Shared Infrastructure
- Storage Lease
- Customized Infrastructure

Consultation

- Infrastructure Setup
- Secure Environment
- Data Use Agreement
- Software Development
- Complex Workflows



Data Science Virtual Desktop Apps

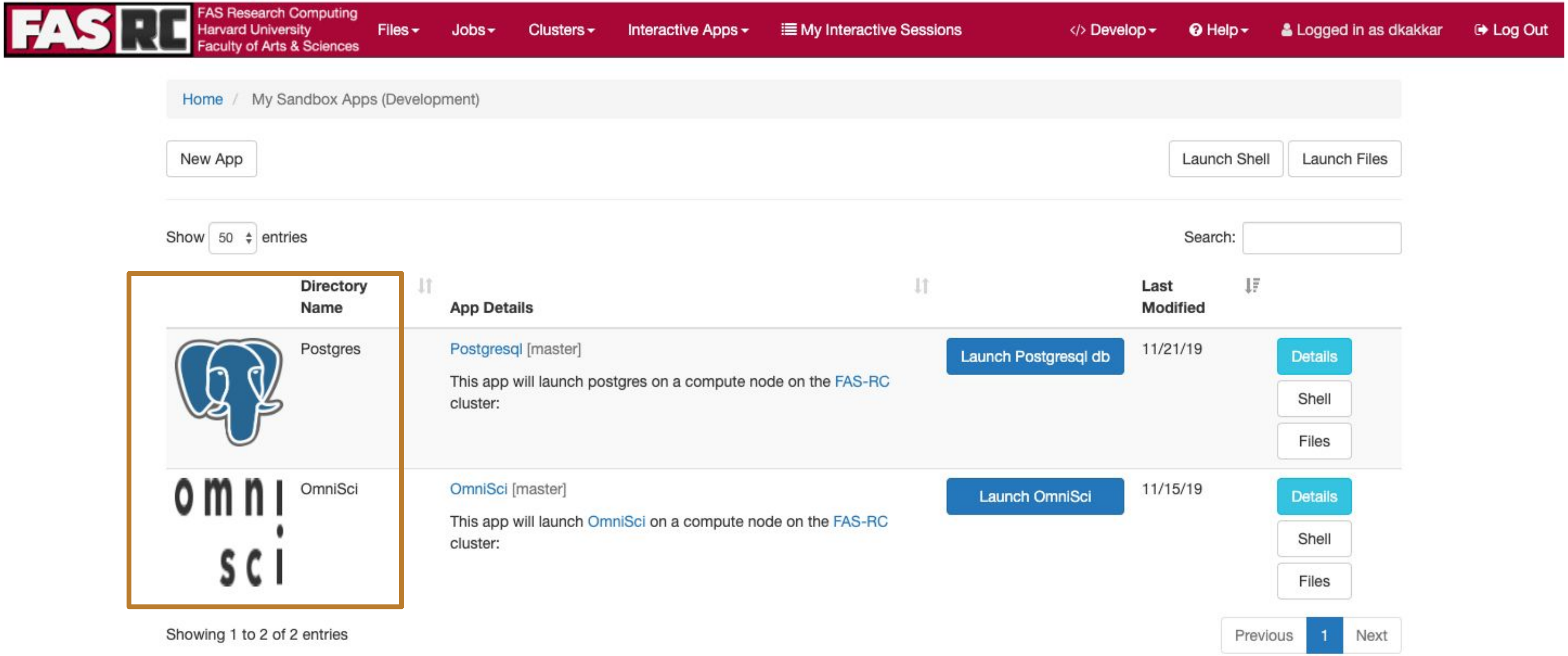


OPEN
OnDemand

GIS Databases for Big Data

- **PostGreSQL**: Powerful, open source object-relational database system
- **PostGIS**: Provides spatial objects for the PostgreSQL database, allowing storage and query of information about location and mapping
- **OmniSci**:
  - Designed to overcome the scalability and performance limitations of legacy analytics tools
  - Super fast queries/analytics (including machine learning) of unindexed data (open source)
  - Super fast interactive rendering (free for educational use) of millions or billions of features, on-the-fly on a map
  - Leverages the massively parallel processing of GPUs alongside traditional CPU compute

Geospatial on Harvard VDI

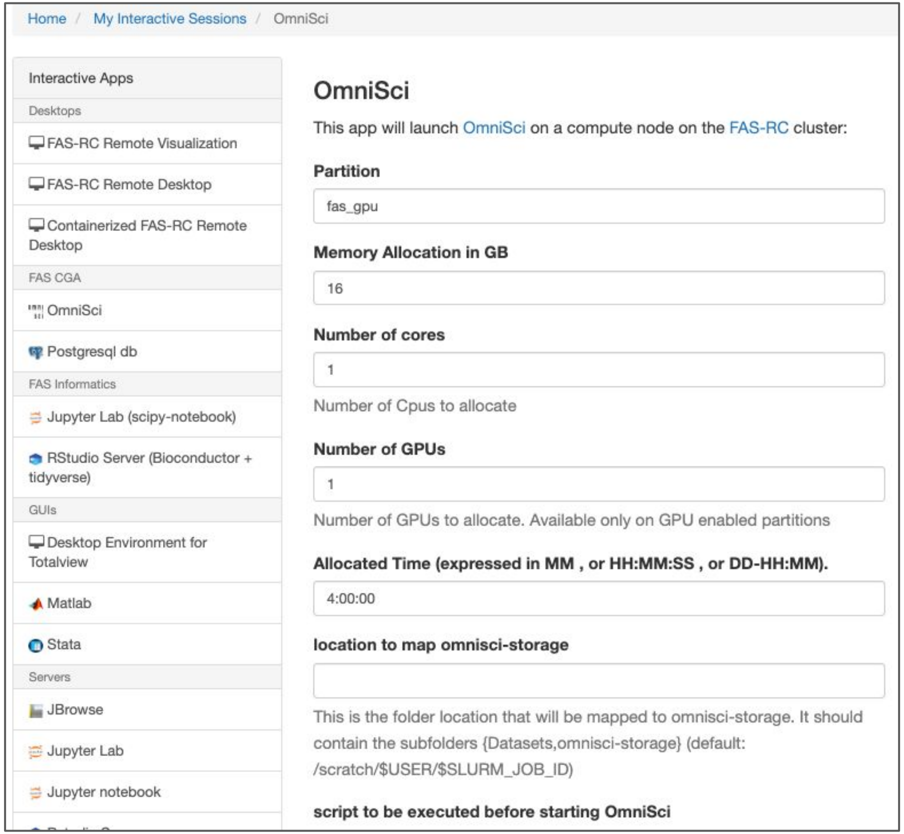


Show 50 entries

	Directory Name	App Details		Last Modified	
	Postgres	Postgresql [master] This app will launch postgres on a compute node on the FAS-RC cluster:	Launch Postgresql db	11/21/19	Details Shell Files
omni sci	OmniSci	OmniSci [master] This app will launch OmniSci on a compute node on the FAS-RC cluster:	Launch OmniSci	11/15/19	Details Shell Files

Showing 1 to 2 of 2 entries

OmniSci on FASRC


Home / My Interactive Sessions / OmniSci
Interactive Apps
Desktops
FAS-RC Remote Visualization
FAS-RC Remote Desktop
Containerized FAS-RC Remote Desktop
FAS CGA
OmniSci
Postgresql db
FAS Informatics
Jupyter Lab (scipy-notebook)
RStudio Server (Bioconductor + tidyverse)
GUIs
Desktop Environment for Totalview
Matlab
Stata
Servers
JBrowse
Jupyter Lab
Jupyter notebook

OmniSci

This app will launch OmniSci on a compute node on the FAS-RC cluster:

Partition

fas_gpu

Memory Allocation in GB

16

Number of cores

1

Number of Cpus to allocate

Number of GPUs

1

Number of GPUs to allocate. Available only on GPU enabled partitions

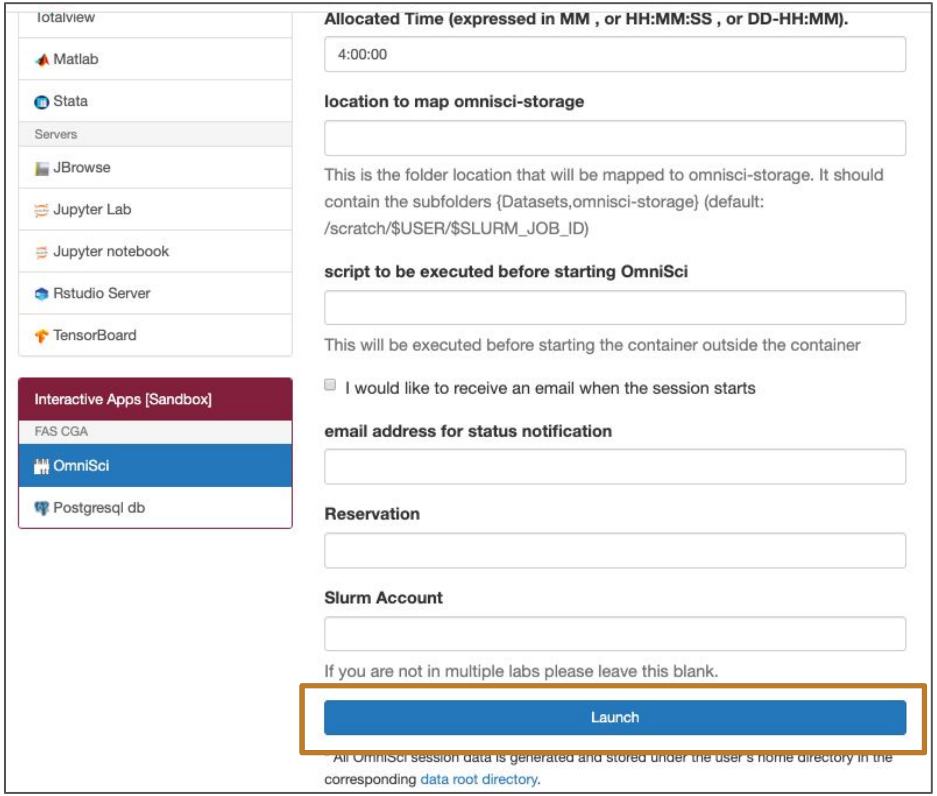
Allocated Time (expressed in MM , or HH:MM:SS , or DD-HH:MM).

4:00:00

location to map omnisci-storage

This is the folder location that will be mapped to omnisci-storage. It should contain the subfolders {Datasets,omnisci-storage} (default: /scratch/$USER/$SLURM_JOB_ID)

script to be executed before starting OmniSci



Allocated Time (expressed in MM , or HH:MM:SS , or DD-HH:MM).

4:00:00

location to map omnisci-storage

This is the folder location that will be mapped to omnisci-storage. It should contain the subfolders {Datasets,omnisci-storage} (default: /scratch/$USER/$SLURM_JOB_ID)

script to be executed before starting OmniSci

This will be executed before starting the container outside the container

□ I would like to receive an email when the session starts

email address for status notification

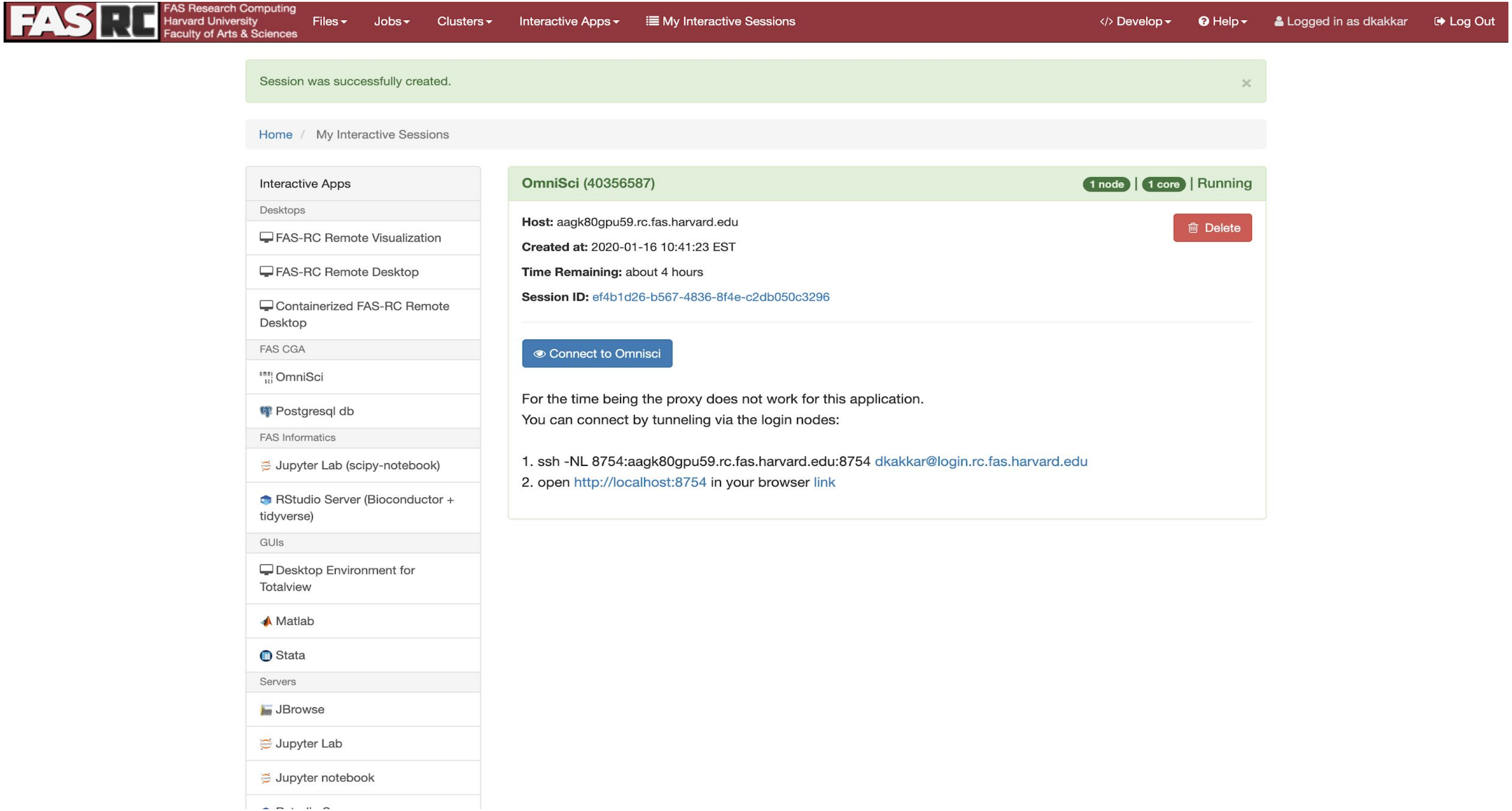
Reservation

Slurm Account

If you are not in multiple labs please leave this blank.

* All OmniSci session data is generated and stored under the user's home directory in the corresponding data root directory.

OmniSci on FASRC



Session was successfully created.

Home / My Interactive Sessions
Interactive Apps
Desktops
FAS-RC Remote Visualization
FAS-RC Remote Desktop
Containerized FAS-RC Remote Desktop
FAS CGA
OmniSci
Postgresql db
FAS Informatics
Jupyter Lab (scipy-notebook)
RStudio Server (Bioconductor + tidyverse)
GUIs
Desktop Environment for Totalview
Matlab
Stata
Servers
JBrowse
Jupyter Lab
Jupyter notebook

OmniSci (40356587) 1 node | 1 core | Running

Host: aagk80gpu59.rc.fas.harvard.edu

Created at: 2020-01-16 10:41:23 EST

Time Remaining: about 4 hours

Session ID: ef4b1d26-b567-4836-8f4e-c2db050c3296

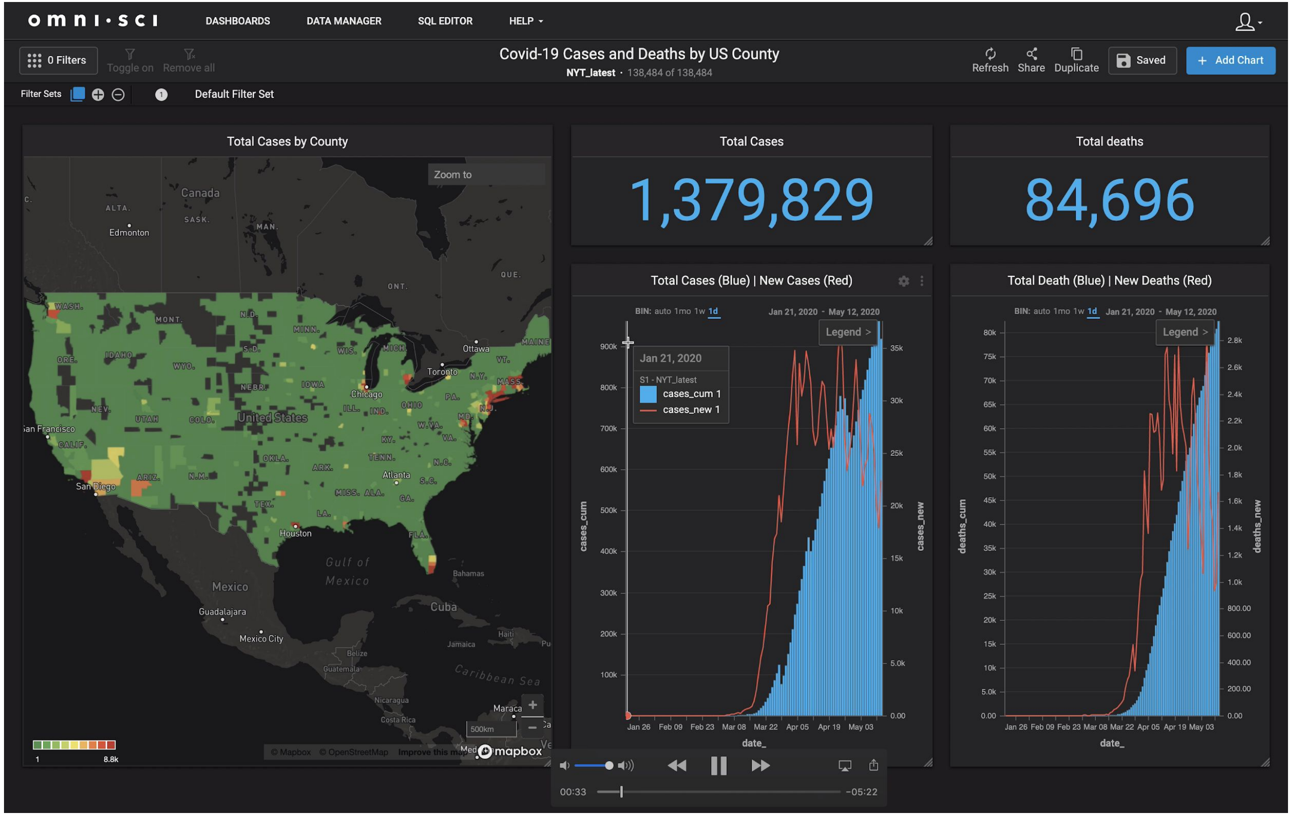
For the time being the proxy does not work for this application.
You can connect by tunneling via the login nodes:

1. ssh -NL 8754:aagk80gpu59.rc.fas.harvard.edu:8754 dkakkar@login.rc.fas.harvard.edu
2. open http://localhost:8754 in your browser link

Demo of the HPC interface - showing how easy it is to create a large instance

https://www.youtube.com/watch?v=TvqqikT_V58

Demo of a Harvard project using OmniSci, running on the cluster



Covid-19 Cases and Deaths by US County

NYT_latest · 138,484 of 138,484

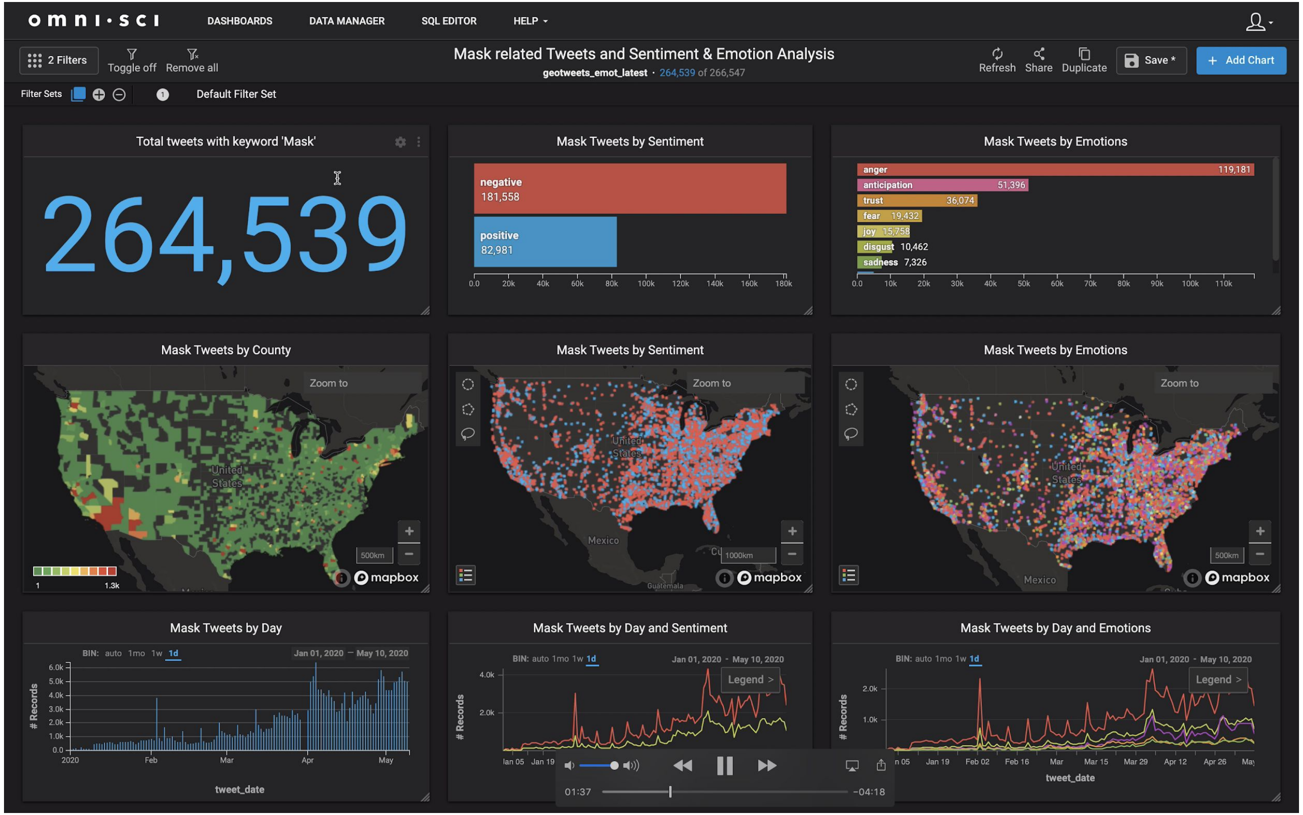
Total Cases

1,379,829

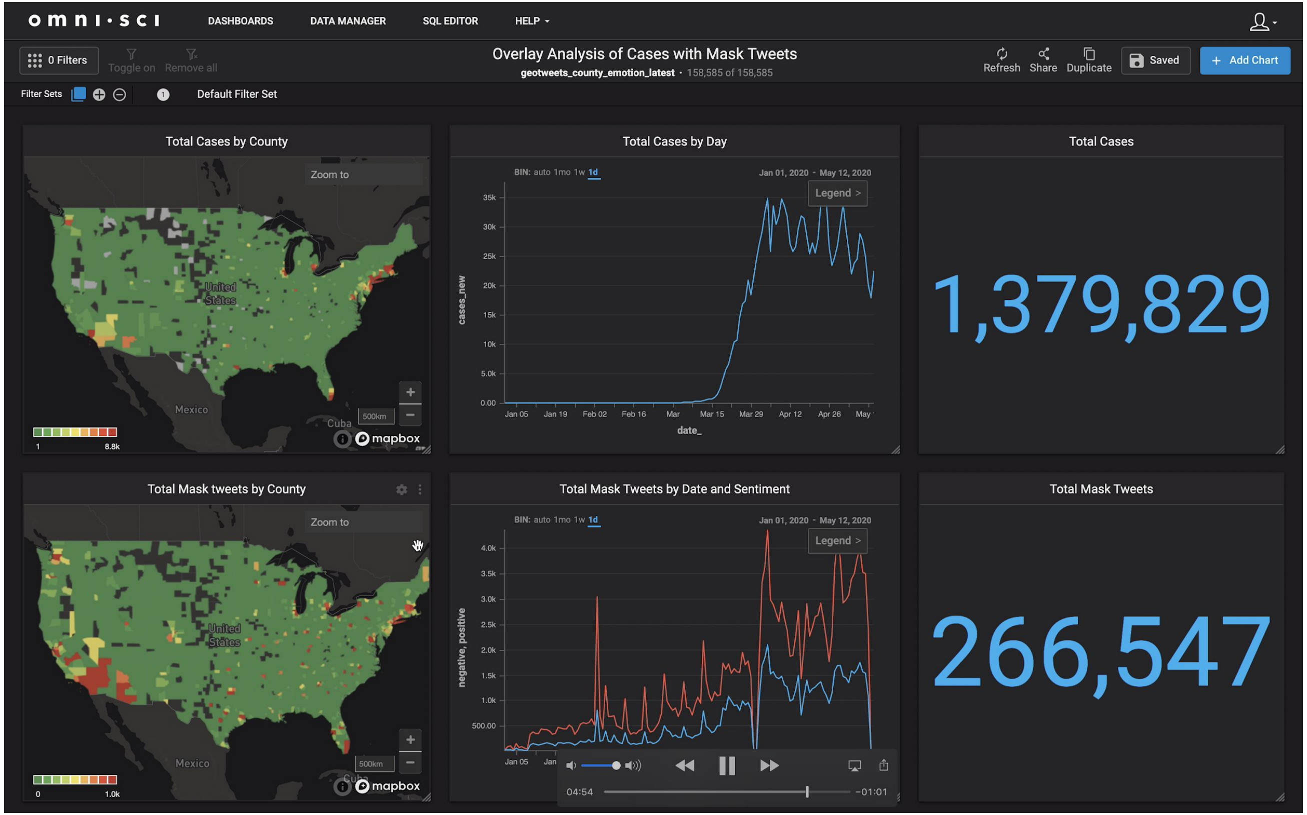
Total deaths

84,696

Demo of a Harvard project using OmniSci, running on the cluster



Demo of a Harvard project using OmniSci, running on the cluster



Demo of a Harvard project using OmniSci, running on the cluster

https://www.youtube.com/watch?v=35nm5I__W-c

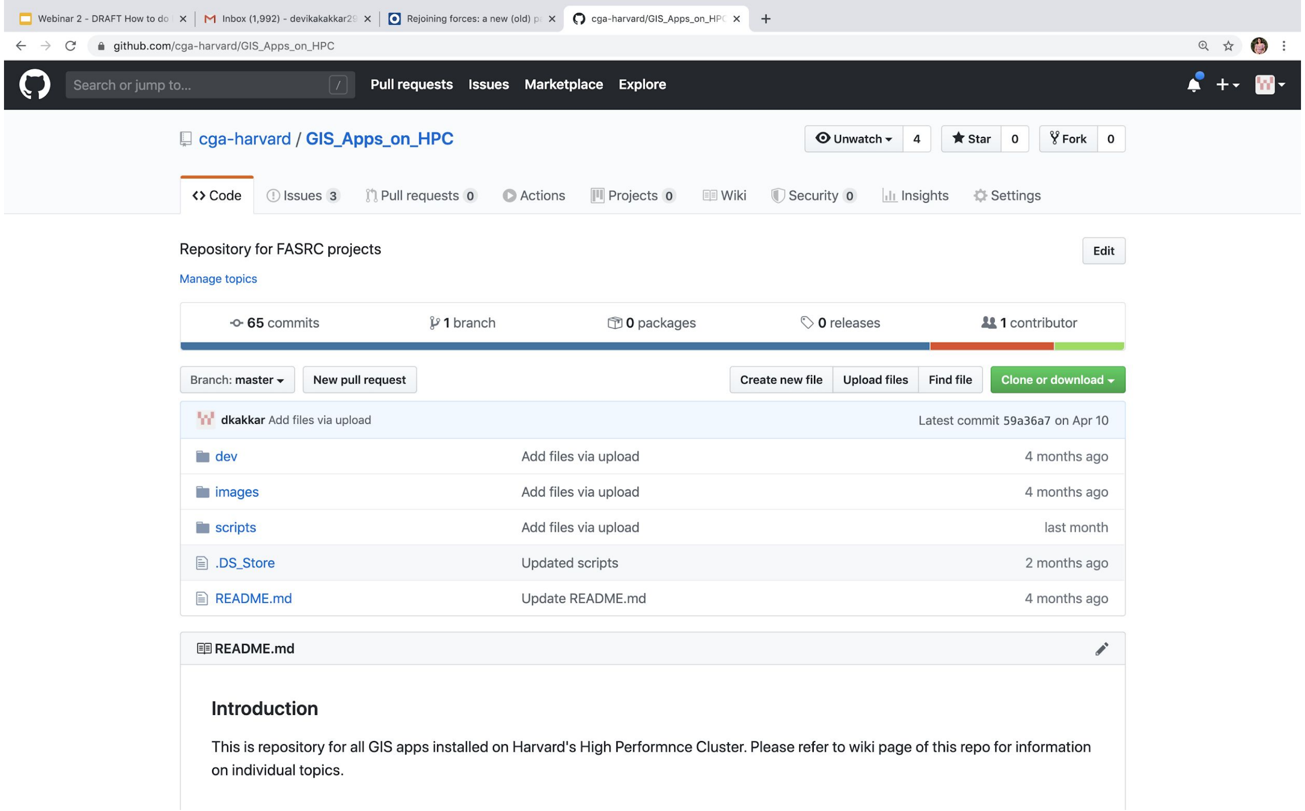
Review of the scripts required to run OmniSci on the slurm cluster

Github Repo:
https://github.com/cga-harvard/GIS_Apps_on_HPC/tree/master/dev/OmniSci

The various script in the repo are run to achieve the following processes:

- Develop a User Interface to request the job parameter(partition, memory, CPUs, GPUs etc)
- Run a slurm job with the parameters specified by the user on Launch
- Install Omnisci using Singularity (Finding port, setting data directory, setting passwords etc.)
- Pass connection parameter to the user (if successful) else display the error

How other slurm clusters can do the same, and get involved in the open source project



cga-harvard / GIS_Apps_on_HPC

Repository for FASRC projects

65 commits | 1 branch | 0 packages | 0 releases | 1 contributor

dkakkar Add files via upload — Latest commit 59a36a7 on Apr 10

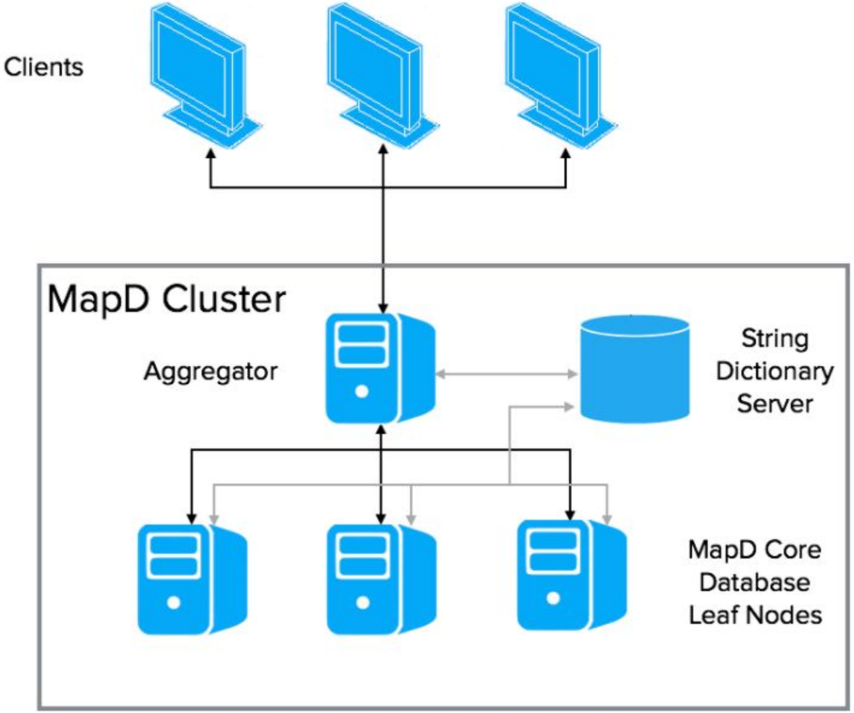
dev	Add files via upload	4 months ago
images	Add files via upload	4 months ago
scripts	Add files via upload	last month
.DS_Store	Updated scripts	2 months ago
README.md	Update README.md	4 months ago

README.md

Introduction

This is repository for all GIS apps installed on Harvard's High Performnce Cluster. Please refer to wiki page of this repo for information on individual topics.

Future Directions - Distributed OmniSci on FASRC



Distributed Configuration OmniSci [5]

Q and A

References

- Geospatial tools on Harvard Cluster Computing https://gis.harvard.edu/geospatial-data-science-tools-and-data-harvards-high-performance-computing-infrastructure

- Harvard Center for Geographic Analysis / OmniSci Collaboration https://www.omnisci.com/blog/rejoining-forces-a-new-old-partnership-with-the-harvard-center-for-geographic-analysis

- Harvard Center for Geographic Analysis https://gis.harvard.edu/

- Harvard Research Computing (FASRC) https://www.rc.fas.harvard.edu/

- FASRC Cluster Architecture https://www.rc.fas.harvard.edu/about/cluster-architecture/

- Introduction to Cluster Computing: https://www.rc.fas.harvard.edu/wp-content/uploads/2019/12/Intro-to-Cannon.pdf

Thank you

Devika Kakkar (kakkar@fas.harvard.edu)
Ben Lewis (blewis@cga.harvard.edu)
Raminder Singh (r_singh@g.harvard.edu)

Outline

- Introduction to the project
- Harvard’s Center for Geographic Analysis
- Intro to Harvard’s Computation Cluster
- Overview of challenges researchers are facing with data which HPC is designed to address
- Installing OmniSci on Harvard Cluster
- Demo of the HPC interface - showing how easy it is to create a large instance
- Demo of a Harvard project using OmniSci, running on the cluster
- Review of the scripts required to run OmniSci on the slurm cluster
- How other slurm clusters can do the same, and get involved in the open source project
- Q and A