Equity, Scalability, and Sustainability of Data Science Infrastructure

Anthony Suen[§∗], Laura Norén[‡], Alan Liang[§], Andrea Tu[§]

Abstract—We seek to understand the current state of equity, scalability, and sustainability of data science education infrastructure in both the U.S. and Canada. Our analysis of the technological, funding, and organizational structure of four types of institutions shows an increasing divergence in the ability of universities across the United States to provide students with accessible data science education infrastructure, primarily JupyterHub. We observe that generally liberal arts colleges, community colleges, and other institutions with limited IT staff and experience have greater difficulty setting up and maintaining JupyterHub, compared to well-funded private institutions or large public research universities with a deep technical bench of IT staff. However, by leveraging existing public-private partnerships and the experience of Canada's national JupyterHub (Syzygy), the U.S. has an opportunity to provide a wider range of institutions and students access to JupyterHub.

Index Terms—data science education, Jupyter, Jupyterhub, higher education

Introduction

Data science education has experienced great demand over the past five years, with increasing numbers of programs and majors being developed. This demand has fueled the growth of JupyterHubs, which create on-demand, cloud based Jupyter notebooks for students and researchers. Compared to local environments that run Jupyter, a cloud based JupyterHub provides many conveniences including not requiring any installation, quicker access to course content, and computing flexibility, such that users even on Chromebooks or iPads are able to run Jupyter notebooks.

Additional benefits include the ability to quickly deploy customizations for different use cases, authentication, autograding, and providing campus-wide computing and storage. Overall, universities have found that utilizing JupyterHubs increases accessibility to data science tools, improves the scaling of data science and computing courses into many other domains, and provides a cohesive learning and research platform.

However, little was known about the barriers universities face when attempting to deploy JupyterHub, which has only been in use since 2015.

This paper aims to understand how JupyterHub is affecting the equity, scalability, and sustainability of data science education by providing four cases studies of how JupyterHubs are being deployed in varying academic institutions across the United States and Canada. We will look at the barriers to deploy, maintain, and grow JupyterHub from the technical staffing and financial perspectives of each institution. The four case studies include large and technical universities such as UC Berkeley, small liberal arts colleges, private universities with large endowments like Harvard, and the Canadian National JupyterHub Model.

We conducted over 10 qualitative interviews with university faculty and IT staff from around the U.S. and Canada. We also reviewed documentation found on Github and websites of 20 institutions regarding their JupyterHub deployments. We structured our analysis by first trying to understand the institution's educational goals and how it drives funding and decision/structure. We then delve into the infrastructural costs, capabilities, along with team size. We lastly measured educational impact, such as the number of students served and the number of classes provided. We conclude with a summary of the findings and potential ways to improve equity, scalability, and sustainability of current existing JupyterHub infrastructure.

Case Study 1: UC Berkeley

In Spring 2015, UC Berkeley became one of the first universities to adopt JupyterHub[1]. Initially set up for 100 students in the new Foundations of Data Science Course *Data 8*, the JupyterHub instance has quickly expanded to now support over 1,000 students in Data 8 each semester and more than 3,000 students in Berkeley's Data Science connectors, modules, and upper division courses. An additional 45,000 students utilize the JupyterHub in Data 8's free online EdX version.

UC Berkeley aims to serve large portions of its 30,000 undergraduates with data science tools, thus creating the motivation for it to build one of the largest JupyterHub deployments in the world. This cross campus pedagogical vision is assisted by the presence of a large technical team, which consists of many members of the core Jupyter team. UC Berkeley's JupyterHub runs on the Kubernetes platform, which allows for easily scalable clusters that can support many thousands of users. Furthermore, Berkeley's JupyterHub infrastructure, which subsists on cloud credits, is supported by long running industry relations and partnerships with cloud vendors like Microsoft and Google.

The UC Berkeley infrastructure team in charge of running Berkeley's instance of JupyterHub, known as "Datahub", consists of the Dean of the Division of Data Sciences, one tenured teaching faculty, one full-time staff member, ~10 postdocs and graduate students who can help troubleshoot–many of which are from

∗ *Corresponding author: anthonysuen@berkeley.edu*
§ *University of California, Berkeley*
‡ *New York University*

the core Jupyter team–along with a large, technically proficient undergraduate support staff[2].

UC Berkeley's model faces sustainability challenges given its heavy reliance on undergraduates, graduate students and postdoc staff and donated computing credits from cloud vendors. Student and postdoc staff generally move on and have other priorities to advance their careers as they typically do not advance their careers by doing SysAdmin work, leading to a lack of consistent support staff and a consequent lack of consistent expertise. The reliance on free cloud credits is further not guaranteed forever and requires regular negotiations with public cloud vendors.

Nonetheless, Berkeley's model benefits from its campus-wide scale, setting the ground for a large and diverse array of data science courses to be setup with minimum infrastructure overhead[3]. The infrastructure can also support very large courses, like quantitative gateway courses for many departments. The Berkeley Datahub has a workflow with unique features like interactive links and Ok.py for large scale autograding of thousands of assignments. Finally, it provides a common suite of tools that are widely accessible, allowing students a productive and cohesive environment for both learning and research.

Case Study 2: Small Liberal Arts Universities

The team interviewed several small liberal arts colleges to see how they utilized Jupyter in their data science or computer science curricula. We learned that lack of funding, insufficient technical knowledge, limited relationships and experiences dealing with cloud vendors, and a shortage of time from busy instructors seem to be the major hurdles to deploying a successfully running JupyterHub.

At liberal arts colleges, deployments are usually designed for small classes consisting of ~20-30 students and maintained by one or two professors. There exists little IT help for the professor, as compared to the vast number of support staff at institutions like UC Berkeley. Some smaller institutions have even asked public institutions like UC Berkeley for support. The lack of proper guidance and departmental resources, along with overburdened faculty, often may dissuade efforts to set up JupyterHub altogether. Generally, paying for such technology is also tough and ad hoc for smaller institutions.

One of the exceptions is Bryn Mawr College; its JupyterHub deployment currently hosts and allows access to a wide range of courses. Some courses such as *Introduction to Computing* (introductory computer science course) have migrated to the JupyterHub environment, while new courses such as *Computing in Biology* have been introduced specifically utilizing Jupyter. Bryn Mawr has emphasized using JupyterHub due to its accessibility for biology students who have limited experienced with programming, while also making it useful for CS students who are interested in biological applications for CS. The *Bio/CS 115: Computing Through Biology* course[4], which was developed based on the Jupyter environment, serves as an alternative CS intro course and a 2nd semester Biology intro course. This option reduces the prerequisite barriers of entry to both domains and allows students to learn both in a well-integrated manner, especially given the amount of intro courses that compete for their schedules.

Case Study 3: Wealthy Private Universities

Compared to smaller liberal arts universities, well-funded private universities often have a rich suite of IT resources. Even if internal IT staff encounter limitations, well funded private universities often pay third-party vendors to help deploy and maintain JupyterHubs and all related support infrastructure. Harvard has said that they "hired a firm to help us implement JupyterHub on AWS". Compared to smaller liberal arts colleges, the experience is relatively free of frustration since the university covers all costs. Nonetheless, Harvard has noted that using JupyterHub has increased flexibility and hence decreased setup costs for both users and instructors, and has further claimed that this solution is much more cost effective compared to traditional solutions.

Most of the classes that have deployed JupyterHub are still relatively small, with most having 12-50 students. At Harvard, JupyterHub was deployed on AWS for two classes in the School of Engineering, which provided significant customization. The Signal Processing class used a Docker-based JupyterHub, where each user was provisioned with a docker container notebook. For the Decision Theory class, JupyterHub used a dedicated EC2 instance per user's notebook, providing better scalability, reliability and cost efficiency[5]. Harvard's School of Engineering and Applied Science (SEAS) further announced in October 2017 for a schoolwide JupyterHub deployment[6]. In addition to SEAS's JupyterHub, the Harvard Medical School has its own JupyterHub deployment.

Instead of deploying and maintaining their own JupyterHubs, other universities have found success by contracting a third-party vendor to deploy JupyterHub. Vocareum[7], an example of one company specializing in this space, helps to set up and manage environments like Jupyter and hosts labs for students to access. Currently, their data sciences lab is used by many wealthy private universities including Cornell, Columbia, and the University of Notre Dame. Others firms that provide similar services include CoCalc and Gryd.

However, the majority of universities generally have less experience with cloud computing and experienced IT staff, thus limiting the replicability of the model. Furthermore, most universities' data science initiatives cannot rely on their university's operating budget to support this type of teaching expense, especially if classes are relatively small (12-50 students), hindering scalability of the model. If done in an uncoordinated way, the costs can skyrocket if departments independently contract with cloud providers and IT consultants to set up their own JupyterHubs.

Case Study 4: Canadian Federation (PIMS)

In 2017, an initiative in Canada led by the Pacific Institute of Mathematics and Sciences (PIMS) and hosted by Compute Canada started a new national model for JupyterHub that provides access to numerous institutions across Canada[8]. With data privacy laws removing the option of using cloud service providers, Syzygy grew to become the largest federally funded JupyterHub and is utilized by more than 8,000 students across 15 universities in Canada. Syzygy is run and supported by one full-time system network manager based at PIMS who oversees installations and collaborates with IT staff at Compute Canada. Any Canadian University can simply ask Syzygy for a JupyterHub and a new cluster will be set up. The system manager is paid for by Compute Canada, and further grants from the Canadian federal government ($4.5m) and Alberta ($1m) support professors and teachers. There is also time donation from professors at 10 different institutions.

Syzygy has some potential bottlenecks. Firstly, there is only one dedicated staff member conducting core management and operations for 15 different institutions. Some scaling issues also currently exist as any institution's JupyterHub is at most able to handle ~2 classes of students concurrently (around 200-300 students). Nonetheless, this is a functional model in terms of scale and sustainability based on the number of universities involved, Canada's population size, and strong governmental support.

The leaders of the effort believe that there are multiple benefits to the strategy. Firstly, it can accommodate small classes, modules, and even high schools across the country. Secondly, it allows instructors to focus more on course development, instead of operating a JupyterHub. Thirdly, it fosters better cross university collaboration by sharing experiences and course modules through a common network.

Conclusion - A Path Forward to a National Jupyterhub

While the grassroots efforts across the U.S. have sparked significant innovation in the realm of data science education infrastructure, it has also created a growing chasm of capabilities between institutions. To equitably increase the access to JupyterHub requires a new model to support many smaller institutions.

Today, only large public or wealthy private universities in the U.S. can provide JupyterHub for many undergraduates. At smaller resource-constrained institutions, deploying a JupyterHub instance for a single class possesses nontrivial costs and may be daunting for one instructor or their university IT staff. Unfortunately, if there is no alternative way to access JupyterHub for data science education, smaller less well-funded institutions and underrepresented communities cannot utilize JupyterHub.

When considering the future of JupyterHub in higher data science education, we see four potential pathways:

- **Status Quo** - Continuing the current grassroots and uncoordinated JupyterHub deployments across institutions would mean smaller or less resource rich institutions would likely continue to face existing barriers. For smaller and resource constrained institutions, JupyterHub would continue to experience very low slow rates of adoption.
- **Institutional Grants** - Increasing foundational or governmental funding for individual universities to set up their JupyterHubs is another option. Funding can enable individual institutions to hire IT staff or pay third-party vendors to create a JupyterHub environment. Based on Berkeley's and Harvard's experiences, we've concluded that grants to hire staff to deploy Jupyterhub is non-scalable given the high costs of hiring IT staff with such specialized experience. Funding third-party vendors like CoCalc, Gryd, Vocareum and public cloud providers like Google or Microsoft to help set up individual JupyterHubs is conceivable, but the individual nature of these transactions may end up being more costly than potential coordinated national or regional models.
- **A National JupyterHub** - A national JupyterHub would offer cost benefits such as utilizing existing federally funded national supercomputing centers. However, a single national hub is difficult to realize due to high coordination costs with thousands of universities.
- **Regional Hubs Model** - Given the number of universities in the U.S., establishing several regional hubs can reduce the burden of deployment and maintenance costs that individual universities experience today. For each regional network, by deploying a large Kubernetes cluster that can support many thousands of users, individual universities can then deploy their own JupyterHubs on the cluster.

The West Big Data Innovation Hub, UC Berkeley, and Microsoft will be launching a pilot program by setting up a Kubernetes cluster using Azure for a small group of Western U.S. universities to pilot their JupyterHubs starting in the Summer of 2018. This will lower the administrative burden while providing a free scalable infrastructure solution for many small or resource constrained universities. Further integration of regional computing facilities at major research universities should be investigated.

[1] Kim, A. (2018, May 2). The Jupyterhub Journey: Starting Small and Scaling Up. Retrieved July 5, 2018, from https://data.berkeley.edu/news/jupyterhub-journey-starting-small-and-scaling

[2] Suen, A. (2018, March 15). People. Retrieved July 5, 2018, from https://data.berkeley.edu/about/people

[3] Kim, A. (2018, February 20). Modules: Data Made Accessible to Many. Retrieved July 5, 2018, from https://data.berkeley.edu/news/modules-data-made-accessible-many

[4] Shapiro, J. (2017, May 20). Computing Through Biology with Jupyter. Speech presented at Jupyter Day Philly, Philadelphia. Retrieved May 24, 2018, from https://github.com/BrynMawrCollege/TIDES/blob/master/JupyterDayPhilly/JAShapiro_JupyterDayPhilly_2017-05-19.pdf

[5] Harvard. (2018). cloudJHub. Retrieved May 24, 2018, from https://github.com/harvard/cloudJHub

[6] Ba, D. (2017, October 23). SEAS Computing and Academic Technology for FAS Launch JupyterHub Canvas Integration. Retrieved July 6, 2018, from https://atg.fas.harvard.edu/news/seas-computing-and-academic-technology-fas-launch-jupyterhub-canvas-integration

[7] DATA SCIENCES LAB @ VOCAREUM. (n.d.). Retrieved July 6, 2018, from https://www.vocareum.com/home/data-sciences-lab/

[8] Canadians Land on Jupyter. (2017, July 11). Retrieved May 24, 2018, from https://www.pims.math.ca/news/canadians-land-jupyter

[9] Mandava, V. (2017, June 8). NSF Big Data Innovation Hubs collaboration - looking back after one year - Microsoft Research. Retrieved May 24, 2018, from https://www.microsoft.com/en-us/research/blog/nsf-big-data-innovation-hubs-collaboration/