The Data Science Journal Club

Using today's buzzword to train tomorrow's HPC scientists

William S. Monroe
IT Research Computing, University of Alabama at Birmingham

Ravi Tripathi
IT Research Computing, University of Alabama at Birmingham

ABSTRACT

In 2012, the Harvard Business Review termed data science "The Sexiest Job of the 21st century." [1]Since that time, with the rise of GPU computation and the concurrent rise deep learning, "data science" skills are sought after more than ever. In early 2019, job postings for data science on Indeed.com have increased by 256% since December 2013. As such, many academic disciplines are searching for ways of integrating data science techniques into their programs. As many data science techniques are computationally intensive, it is natural to pair data science training with training in proper use of HPC resources. At UAB, Research Computing has initiated a weekly Data Science Journal club. By training members in data science, we naturally touch on HPC topics as well, making this an ideal frontier for HPC education. Similarly to traditional academic journal clubs, new topics are discussed each week; however, rather than reading papers, the club focuses on working through practical data science applications with published code.

ACM Reference Format:
William S. Monroe and Ravi Tripathi. 2019. The Data Science Journal Club : Using today's buzzword to train tomorrow's HPC scientists. In *Proceedings of Supercomputing Conference Series (SC19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

At University of Alabama Birmingham (UAB), Research Computing (RC) has implemented the Ohio Supercomputing's Open OnDemand (OOD) [2]. Open OnDemand is a web-based client for accessing high performance computing (HPC) resources. In the year prior, RC had initiated two user training initiatives, the first being monthly user sessions for in depth training on HPC topics. These have historically been well attended and gradually grown to the point that the venue for these trainings is typically maxed out. Training topics have varied from python virtual environments to appropriately requesting and using resources on HPC.

The second training initiative was to open weekly office hours to aid our users on a more individual basis with their HPC use. At the beginning of 2018, these office hours occurred once per week. By the end of 2018, RC had increased the offered office hours to cover three separate days of the week. Much of the work done during office hours had to do with what could be categorized as "barrier to entry" issues. This includes but is not limited to using the job scheduler, ssh, and VNC desktops. As HPC users discovered these office hours, we naturally had a growing set of users showing up for troubleshooting help.

After the implementation of OOD, the number of visits to RC office hours dropped off; however, attendance at monthly user training sessions continued to grow.

At the same time, requests to RC for help with data science and computational science applications has likewise continued to grow. The Data Science Journal Club (DSJC) was created by RC at the coinciding of these trends. The design of the DSJC is to live stream a practical data science example once per week and archive the videos on YouTube. Each example is chosen because it includes freely available code. Participants internal to UAB can always run the applications in the exact same way as is done during the stream. During the live streams HPC topics such as requesting resources, using GPUs on HPC, creating Jupyter notebooks, job submission, basic Git use, and python virtual environments have been covered. Within the context of the DSJC, these HPC related issues are a means to an end, but still highly important to the workflow. Data Science topics covered have ranged from data manipulation to deep learning for noise reduction.

The DSJC has registered participants from the UAB departments of library, neurobiology, biochemistry, biology, pathology, materials engineering, and computer science. As a baseline metric, the RC YouTube channel, which was started for this purpose, rose from zero to 130 subscribers in less than four months.

Creating an archive of these videos has enabled RC to maintain a list of videos which essentially contain step by step instructions for several workflows with HPC best practices.

2 DISTRIBUTION

2.1 Official Credited Course

The DSJC has been registered as an official journal club with the the Graduate Biomedical Sciences program at UAB. As such, graduate students can sign up and receive their journal club credit through the DSJC weekly tutorial sessions. As no RC staff is active faculty, it is offered in partnership with a sponsoring neuroscience faculty member

2.2 Live Streaming

There are several different streaming services and software which could be deemed appropriate for such an application. Live streaming to websites such as YouTube, Twitch, and Facebook require two primary elements, a streaming software, and a user account. For DSJC purposes, the Open Broadcasting Software (OBS) Studio

SC19, November 17-22, 2019, Denver, CO

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

was chosen. OBS is a fully functional, open source, free option for streaming media to remote hosts.

Initial streaming was performed primarily to YouTube, which will freely host both the live stream and archive the videos after production. After gaining some additional expertise, RC added simultaneous streaming to Twitch. Twitch is a streaming service primarily used for videogame streaming, but may be more accessible to a portion of the intended audience. In order to stream simultaneously, the Restream.io platform is used to mirror the stream to all selected streaming services.

2.3 Links to services used to create the DSJC

- RC YouTube: https://youtu.be/tfGJlO9AeXU
- RC Twitch: https://www.twitch.tv/wsmonroe
- UAB RC OnDemand: https://rc.uab.edu
- OBS: https://obsproject.com
- Restream.io: https://restream.io/marketplace?ref=Ev3p4
- OOD: https://openondemand.org/

2.4 Streaming vs Pre-recording

RC determined to use streaming for two primary reasons.

First, the time requirement of semi-professionally editing content every week seemed overwhelming at the outset of the project. By utilizing streaming, the time commitment for creating the content is kept to a minimum.

Second, what are sometimes viewed as "mistakes" during a demonstration in programming are often extremely valuable moments for novice students who may experience similar issues. By leaving the troubleshooting process in the content, novice students of both data science and HPC are provided with a low-level tutorial not only in the chosen topic, but in practical work within HPC systems and data science topics.

3 ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Kristina Visscher, for her collaboration and support as the official course director for the DSJC, Dr. Curt Carver for his encouragement to explore Twitch and other alternative streaming platforms, and Dr. Ralph Zottola and John-Paul Robinson for their support in the creation and operation of the DSJC.

REFERENCES

[1] Thomas H Davenport and DJ Patil. Data scientist. *Harvard business review*, 90(5):70–76, 2012.

[2] David E Hudak, Douglas Johnson, Alan Chalker, Jeremy W Nicklas, Eric Franz, Trey Dockendorf, and Brian McMichael. Open ondemand: A web-based client portal for hpc centers. *J. Open Source Software*, 3(25):622, 2018.