Creating Guidelines to Supplement the Data Analytics Program in Community College toward Preparation of STEM and HPC Careers

Elizabeth Bautista
NERSC, Lawrence Berkeley National Laboratory
ejbautista@lbl.gov

Nitin Sukhija
Slippery Rock University of Pennsylvania
nitin.sukhija@sru.edu

ABSTRACT

Data science continues to create opportunities in the technology and HPC industry resulting from growing data sets, the need for more insights, the necessity of automation, the evolving roles and changes in job descriptions as those positions are needed and the shortage in the workforce with this talent. However, despite the growing demand, not enough students are learning the basic skills or being able to be given opportunities for hands-on work. In the Northern California Community College system, many of the students return to school after having graduated with a bachelor's degree or find the need to gain new skills to enhance their resume or to change careers altogether. Unfortunately, in the community colleges, there are not enough classes or instructors who are trained in data science to teach the class. In the four-year university, the program is usually waitlisted for transfer students from the community college. This paper is a continuation of the work after the National Energy Research Scientific Computing Center (NERSC) partnered with Laney College to start a Data Analytics program. After two years, they are challenged with not enough instructors to the number of students that are interested in the program. Further, approximately 40% of students are struggling to continue the rigorous material they need to learn. These students may have to work to support families and are unable to put in the 20-40 hours of work to earn a living as well as the 20-40 hours of study and homework that the program requires. Therefore, Laney partnered with Codefinity, an online education program that has a track for Python Data Analysis and Visualization. In the prior year, students who learned how to visualize data with Grafana [8] were able to master it in two months and create new visualizations for NERSC Operations staff [3]. Codefinity [6] provides a similar program where in six courses they learn about the data and learn a visualization program at their own pace of either a three-month plan or an annual plan. This paper describes the early results of using these guidelines to aid students who are not able to adhere to a rigorous study program in a community college as well as for students who want this skill in addition to the concentration they are currently taking.

KEYWORDS

1 INTRODUCTION

As new data sets are created and data volume grows, organizations are looking for more insights than ever before because they want to leverage the collected data. With the improvement of technologies like machine learning and artificial intelligence being much more accessible, organizations not only want their routine work to be automated but the job itself enhanced so that the human is doing something smarter and more effectively.

At NERSC, the Operations Technology Group (OTG) staff are the 24x7 onsite site reliability engineers who are the first responders to anything that occurs in the data center. This particular job description requires knowledge of system administration of HPC systems, local and wide area networking, a three-tier onsite storage and data center facility management at minimum for staff to be successful. NERSC is one of the largest facilities in the world devoted to providing computational resources and expertise for basic scientific research. NERSC currently supports close to 10,000 users globally across almost 1,000 scientific projects.

As NERSC moves toward an exascale system, one of the OTG's strategic plan is to use the data collected by the Operations Monitoring and Notification Infrastructure (OMNI) system to create a new paradigm called Artificial Intelligence Operations (AIOps) [1]. This means, the analysis of the data should be able to predict usual problems before they occur and mitigate them without human intervention.

But in today's shortage of talent, how will they be able to hire staff who can do this work unless they participate in the educational system to help create a pipeline of staff? The ongoing solutions reported in prior papers have been effective but this industry needs more and with the constraints of not enough faculty to many more students interested in learning these skills, NERSC partnered again with Laney College who partnered with Codefinity to create guidelines for not only their students but all students in general to help supplement the current shortage in faculty for these classes [5, 11].

This paper documents the process for creating a pipeline where data analytics students can supplement their education, non-data analytics students can learn these skills, and students in the program who cannot adhere to the rigorous program have an alternative. Section 2 will explain how the partnership with Codefinity was established. Section 3 shows how the program was designed by faculty and NERSC staff with Codefinity. Section 4 provides the logistics of funding for this program. Section 5 will provide case

studies of positive outcomes from the trial year of the program. Section 6 will provide lessons learned, future work and final thoughts to be able to continue the program.

2 BACKGROUND

This paper is the continuation of work done with Laney College over the last two to three years [2–4]. We started with creating pathways for disadvantaged students toward getting them educated and trained for STEM and HPC careers. Following that, we strengthened the community college's computer science program by creating a data analytics program that can be offered to students to give them an advantage in getting a job in this field and also to provide them leverage for when they transferred to a four year university [9].

As the need and popularity of the program grew, funding for the program was more challenging to obtain. Therefore, a decision was made to help supplement the program through other means to assist students who are not able to continue with the rigorous study required to complete the educational portion and obtain the hands-on experience they needed to get a job. We decided to leverage online training programs and investigated several. One of the potential programs we investigated was Merit America [10]. This program was online, provided various tech programs including coding and data analytics. They also provided a low cost onramp into the program with a stipulation that pay back isn't required until the student is earning at least 40K a year in a job. Cost is a limiting factor in this instance. Although there is practically zero start up cost, the student would go into debt again. This did not work for returning students who already have a four year degree, college debt and yet do not have a job which is why they are back in a community college.

Another program we investigated is Coursera [7]. This program had many choices and the courses seemed like a duplicate of what students already took in the community college. They were also too fundamental for the more advanced students. And though they provided a programming segment, it does not specifically have courses in data analytics. Course was not as prohibitive even though you pay as you go along but sometimes, the decision is not only about the cost.

There were many bootcamps online that were also provided that promised to gain a student a skill in a short amount of time. Again, these courses are very expensive and would be prohibitive for students already in debt. Further, while they present much marketing that they will help students get a job, interviews with prior students show that although they have a job, it took them more than one year after taking the program and it is not the career they necessarily envisioned for themselves.

We decided to use Codefinity is because of the following reasons:

- The cost was affordable for students and did not get them into additional debt.
- The programs were what we were looking for including data science, data analytics, machine learning and artificial intelligence, which could supplement the Laney curriculum.
- There is a hands-on component where students complete projects in the software's workspace or students and faculty

can design a project that groups of students can complete. We hope it's the next best thing to an internship.
- The timeline to complete a course is up to the students up to one year before the course needs to be redone and paid for again. Further, the quicker the students complete a course, the faster they could complete the program.

Negotiation to partner with Codefinity included the potential for Laney to subsidize the cost for students, allowing the cost to be even less. The program costs 25% less as long as students complete the program. The 25% is credited to students when they successfully complete the program and applied toward the next course. The courses cost 50% less as they complete 70% of the program therefore, there is an incentive for students to complete courses as well as the program.

3 PROGRAM DESIGN

Codefinity, the Laney faculty and NERSC staff collaborated to "pick and choose" various classes in their programs to create a new online course of study that would allow students to do a course online if they cannot participate in school and extend the course of study to more than a semester should a student need it. Further, there were stand alone courses not part of the data analytics track that could supplement Laney's classes in programming such as a python course for beginners and Pandas for more advanced students. For students in the data analytics track, a course called Business Analysis Fundamentals is being used to help students get hands-on practice in looking at business data and creating analytics graphs for them to answer those insights needed by companies. For more advanced programmers in the analytics track is a class that uses python with visualization software and linear regression with python. The addition of these more advanced courses will help students planning on applying to the four year university data analytics program. The advantage of doing the courses online is that they complete them in their own time. As long as they can set aside time each day or each week to complete a series of modules, life does not have to compete with their education and training.

For example, in the beginning python class, the course has 192 chapters which should take a total of 16 hours for the average students who will need to do homework to complete 168 tasks. While the 16 hours may seem daunting, it is a total 16 hours that can be done across a semester or one month, as long as they complete all the tasks. All that's needed is for the student to commit to the time, a device with an approved browser and the motivation to learn.

4 COST AND FUNDING

Codefinity agreed to the following plans for our students:

- An annual plan that can be billed monthly or annually. This would be the best plan for a single course. The billing cycle renews if the student does not complete the course within in this period.
- A three month plan for a single course that can be billed every three months or monthly. The billing cycle renews if the student does not complete the course within this period.
- A monthly plan. This is the most expensive of the plans and this plan does not qualify for the school discount. As with

the above two plans, if the student does not complete the course in this time frame, the billing cycle renews.

- An annual professional plan where a student can start up to five courses and can take up to one year to complete them. While this costs more, it is less expensive than to pay for a course at a time.

As previously mentioned, Laney will subsidize 25% of the annual cost or the three month plan cost, and is credited to the student as they complete each course. When they are about 70% completed, they can get up to a 50% discount. A cost comparison to Laney's current tuition cost which is currently free as long as the student lives within the area served by the school, while more expensive, it is less expensive than a semester at the current state university and is subsidized by the school. Currently, the school has funding for three years. We currently offered this program in the spring of 2024. We will continue to offer this program throughout the year as long as we can secure funding. Financial support for community colleges is always dire. It is our hope that we have expanded the Computing Science Department with this program will continue to get us yearly funding from the state. We also have a few grants that we hope to complete before the fall to give us additional funding.

5   EARLY RESULTS

In the spring of 2024, 30 pilot students entered the Codefinity program. Fifty percent of the students are part of our two previous programs and have opted to take the one year professional plan. The other twenty-five percent are newly enrolled fall students in the computer science program and the last twenty-five percent are new students not in a computer science program. The second batch of fifty percent of students are taking either an annual plan or a monthly plan per course.

As of the end of May 2024, 10 students competed five courses and when they applied to a summer internship, six students were accepted to a Port of Oakland Authority internship, two students at the San Francisco Airport Commission and the other two are in the summer program at LBNL. The other students have completed a minimum of one semester of work in their programs. One student was accepted to New York University's Data Science program as a junior and a second student was able to transfer to California State East Bay's Computer Science Program. A third student was able to be accepted to the University of California, Davis's Computer Science Program. Two students earned a summer internship at Stanford Linear Accelerator Laboratory.

Although not every student can say that they earned an internship or were able to complete a course, none of the 30 students in the pilot program left the program because they could not complete the courses. In fact, those students who earned a transfer to the four year universities plan to continue with the online program.

6   LESSONS LEARNED AND FUTURE WORK

An important lesson for us is that students need flexibility and an educational institution needs to provide this to ensure success for their students. The tutoring and mentoring program continues to help the students. We continue to create partnerships with Bay Area organizations to provide internships for our students. Our big issue is continued funding especially now that normal classes are back and more students are enrolling due to the results of lack of employment from COVID. A Bureau of Labor Statistics survey in 2023 noted that in the US 2.5 million people lost work because the employer closed or lost their business. Many of them went into another industry. The impact of the pandemic today continues as companies who may have survived the closure now find that they need to shut down. In Oakland, where Laney College is located, 25-30 businesses close every month in multiple industries. This is in addition to the layoffs that continue to happen.

As we lose these companies, people will look for alternatives and will want to gain a new skill. The tech industry is a good possibility for them because our salaries tend to be higher even when the jobs are entry level. However, educational organizations need to provide the training and hands on experience to update skills.

At NERSC, we have experienced that students can quickly understand how to analyze and visualize data. As we continue to gather data, we need these skill sets so that Site Reliability Engineers can support their data center. The future work involves being able to secure funding so that the program can grow as enrollment grows and to continue recruiting more employers to participate in our internship program.

REFERENCES

[1] Elizabeth Bautista, Melissa Romanus, Thomas Davis, Cary Whitney, and Theodore Kubaska. 2019. Collecting, monitoring, and analyzing facility and systems data at the national energy research scientific computing center. In *Workshop Proceedings of the 48th International Conference on Parallel Processing*. ACM, Kyoto Japan, 1–9. https://doi.org/10.1145/3339186.3339213

[2] Elizabeth Bautista and Nitin Sukhija. 2021. Employing directed internship and apprenticeship for fostering HPC training and education. *JOCSE* 12, 2 (2021). https://doi.org/10.22369/issn.2153-4136/12/2/8

[3] Elizabeth Bautista and Nitin Sukhija. 2023. Creating pathways in disadvantaged communities towards STEM and HPC. *JOCSE* 14, 2 (2023), 2–5. https://doi.org/10.22369/issn.2153-4136/14/2/1

[4] Elizabeth Bautista and Nitin Sukhija. 2024. Data analytics program in community colleges in preparation for STEM and HPC careers. *JOCSE* 15, 1 (2024), 59–63. https://doi.org/10.22369/issn.2153-4136/15/1/12

[5] Berkeley Lab. 2022. *Berkeley lab demographics*. https://diversity.lbl.gov/berkeley-lab-workforce-demographics-fy2022/

[6] Codefinity. 2024. *Boost your career...* https://codefinity.com

[7] Coursera. 2024. *Your path to career growth*. https://www.coursera.org

[8] Grafana Labs. 2024. . https://grafana.com/

[9] Rafael A. Irizarry. 2020. The role of academia in data science education. *Harvard Data Science Review* 2 (2020).

[10] Merit America. 2024. *Start your tech career...* https://meritamerica.org/

[11] Public Policy Institute of California. 2014. *California's need for skilled workers*. https://www.ppic.org/publication/californias-need-for-skilled-workers/