# Data Analytics Program in Community Colleges in Preparation for STEM and HPC Careers

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

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

## ABSTRACT

Students in community colleges are either interested in a quick degree or a skill that allows them to hop onto a career area while minimizing debt. Attending a four-year university can be a challenge for financial costs or academic reasons, and acceptance can be competitive. Today's job market is challenging in hiring and retaining diverse staff. More so within the High Performance Computing (HPC) or a government laboratory. Industry offers higher salaries, potentially better benefits, or opportunities for remote work, factors that contribute to the challenge of attracting talent. At the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory, site reliability engineers manage the HPC data center onsite 24x7. The facility is a unique and complex ecosystem that needs to be monitored in addition to the normal areas such as the computational systems, the three-tier storage, the supporting infrastructure, the network and cybersecurity. Effective monitoring requires the understanding of data collected from the heterogeneous sources produced by the systems and facility. With so much data, it is much easier to view the data in graphic format and NERSC uses Grafana to display their data. To encourage interest in HPC, NERSC partnered with Laney College to create a Data Analytics Program. Once Laney faculty learns how to teach the classes toward a certificate program, they fill a need for their students to build the skill in data analytics toward a career or to continue toward a four-year degree as transfer students. This also fills a gap where the nearby four-year university has a long waitlist. This paper describes how NERSC partners with to create a pipeline toward a data analytics career. This is the follow-up program to creating a pathway into HPC and Science, Technology, Engineering and Mathematics (STEM) [3].

## KEYWORDS

Site Reliability Engineer, HPC Education, HPC Training, Diversity, Inclusion, STEM, community college, data analytics.

## 1 INTRODUCTION

Everything today requires data or a visual representation of the data. According to a 2020 study at the Massachusetts Institute of Technology, the demand for "data scientists" is expected to grow in the future as more data is being accumulated. However, even academia is confused because there is no standard on how to train a workforce in this area. Is Data Science a series of principles, a skillset or an "umbrella term" that encompasses a series of required expertise? Is it a specific discipline? What are the jobs associated with this type of degree [7].

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. Although the position does not necessarily require a college degree or certifications, the job description does require 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.

Further, one of the Laboratory's mission statements is widening Diversity and Inclusion in these areas and has programs that support internship programs in underserved communities as well as recruit staff from a wide area of disciplines with the idea that these skills can be transferable into science research. However, according to a 2022 Lab study of our workforce demographics, we continue to see a small percentage of staff who are underrepresented such as Black/African American, American Indian/Alaska Native, Asian, Hispanic or Latino, etc. We see an even smaller percentage of women in these demographics especially within the Lab Senior Leadership roles. The numbers demonstrate a compelling story [5, 9].

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 [1]. As NERSC moves toward exascale, how will we increase diversity and inclusion percentages in this area and also continue to recruit the much needed talent to fill positions that are also in high demand in our neighboring Silicon Valley?

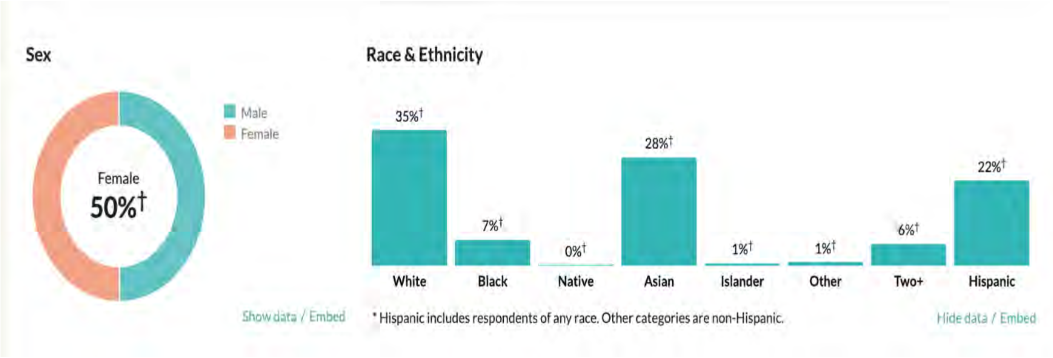
Part of the solution was to partner with the local community college in the neighborhood to create a potential pipeline by influencing students' curriculum and education with much needed support for these underserved students as well as providing a place for them to practice their skills hands-on.

This paper documents the process for creating a pipeline where data analytics students can get an education, are provided training and hands-on projects [2]. Section 2 will explain the academic program itself and the creation of the internship program. Section 3 provides the background on the type of training they will be provided by the internship program, in addition to the program at school. Section 4 will provide logistics such as sourcing funding streams and our experiences in the first year of the program. Section 5 will provide case studies of positive outcomes. Section 6 will provide lessons learned, future work and final thoughts to continue the program.

## 2 BACKGROUND

With so much information and a list of the different type of programs in academia, we decided to focus on the type of training needed by the employers in the San Francisco, Oakland, Berkeley, and Alameda areas, the areas surrounding the school. A survey from the school's industry partners resulted in a need for more training in these skill areas: Data Analyst, Data Engineers, and Database Administrator. Although the survey included other skill areas such as Machine Learning Engineer, Data Scientist, Data

© 2024 Journal of Computational Science Education
ttps://doi.org/10.22369/issn.2153-4136/15/1/12



* Hispanic includes respondents of any race. Other categories are non-Hispanic.

**Figure 1: Workforce demographics survey, Lawrence Berkeley National Lab, 2022**

Architect, Statistician, Business Analyst, and Data and Analytics Manager, the school felt that the latter required more foundation than they could provide, therefore, they focused on the three skillsets first mentioned.

A second survey found that although these jobs are in high demand by the area employers, it requires a very specific skill. For students, this could be daunting even just imagining undertaking such training, especially when they have low confidence in their science and math skills. However, these skills are obtainable with the right education, hands-on training and support of the students from peers and faculty. After all, we are training the next generation of data analysts.

Because we also had a focus on diversity and inclusion, we wanted to recruit students into the program who are underrepresented in the data science career areas. In the case of NERSC, the area they wanted to serve was the San Francisco, Oakland, Berkeley, Alameda Metro area with an underrepresented population of 4,579,599. According to a workforce survey by the Human Relations department at LBNL in 2022, less than 10% of the Bay Area workforce are within the underrepresented population of Black, Pacific Islander, Other, Native American and two plus who work in STEM fields. Within these groups, there is still only 50% or less women, especially in the senior management levels [5] See Figure 1.

We were also hoping to fill a gap in the closest four-year university offering a data science degree had a wait list of at least 200 students in the last five years that includes the pandemic. Students who are accepted as freshmen into the four-year university could be accepted into the data science program; however, due to the wait list, a community college transfer student would also have to join the waitlist. Filling this gap through teaching the classes at the community college level could give these transfer students a chance to enter the program in their junior year.

## 2.1 Train the Teacher

One of the early steps is to train the faculty on how to teach various classes in the program. Through various contacts with the University of California, Berkeley (UCB), Electrical Engineering and Computer Science (EECS) department, we decided that the best training to get is from the trainer themselves. Therefore, the faculty who have committed to teach at the community college, entered various UCB extension programs to go through the classes themselves.

Three faculty members started the program in the prior spring semester 2022 before we launched the classes in the fall, September 2023. By the fall program start, they would have taken two semesters, spring and summer, of classes and should be able to teach a series of classes in the fall.

## 2.2 Create the Program

The program itself would not be finalized until the middle of the summer prior to the launch. This is due to funding which I will cover in Section 4. Parts of the program that we would submit toward a certificate are already being taught and we will just integrate the data science classes.

As such, this is the program that was created and presented beginning September 2023, noted by Figure 2. Though the data science foundation classes would not be taught until the end of their degree program, most students who added the data science program would have already completed most of the pre-requisite classes in the prior year. Thus, they were second year community college students. Most of these students came from the cohort of students from the previous year who entered another program to prepare them for STEM classes.

Database Programming with SQL
Introduction to Microsoft Excel for Business
Introduction to Computer Programming
Introduction to Computational Thinking with Data
Introduction for Computer Science
Microcomputer Assembly Language
Introduction to Artificial Intelligence and Machine Learning
Object Oriented Programming Using C++
Java Programming or Python Programming
Data Structures and Algorithms
Structure and Interpretation of Computer Programs
Foundations in Data Science
Introduction to Statistics

**Figure 2. Data Science certificate program**

These classes vary from three to five semester units. The program requires a minimum completion of 27 units - 31 units to acquire an Associate of Science (AS) degree with a specialization in data science. These classes are also transferable into the University of California system as credits toward their freshman and sophomore years. In the University of California, Berkeley, it would be acceptable as pre-requisites to classes they would need to take to earn a Bachelor of Science (BS) degree with a concentration in data science.

## 2.3 Support of Students

Understanding that some of the students who enter the program may have been challenged in math and science through the high school level, the college decided that we need to provide some tutorial support for the students. Thus, one faculty member was transferred from the previous STEM program into this program to assist in supporting the students with their homework, concepts explanation, any lab they need and generally with their homework. In addition, upper classmen students from the math and science departments were recruited to help during the tutorial lab. Industry partners were also asked to volunteer time to help the students at their discretion. Due to the wide variety of volunteers who assist in the tutorial lab, the support program can be available from 9:00 a.m. through 9:00 p.m. Monday through Friday, depending on the needs of the student and availability of tutors. After 5:00 p.m., tutorial times are on a scheduled basis.

## 2.4 Internship Program

The internship program was modeled after the prior program for STEM preparation. We used most of the same industry partner employers. The students work 40 hours during a thirteen-week summer program that required for the employer to provide a hands-on training class for at minimum forty hours. Further, they provide a project where the student can do the work that they just learned to do in the eleven weeks after. In the last week, the students need to prepare a poster presentation that will be explained to several of the company staff. In lieu of the poster, they will submit a paper for review to a conference within one year. The preparation of the submission will be approved by the student's supervisor and faculty advisor. The employer commits to send the student to the conference if the paper is accepted for presentation. More on how this travel will be funded in Section 4.

# 3 TRAINING BY THE EMPLOYER

Each employer we recruited committed to providing the necessary training, a very specific skill in the data science area that a student can learn within forty hours and put that training into practice on a project that they will work on for the next eleven weeks in their organization.

For NERSC, they participated in the programs pilot cohort of students. In the summer of 2023, the students from the program took a class on Grafana [6], an open-source software that visualizes data for a time-series database into graphs and visualizations that can be queried, alerted on and explore metrics and logs. NERSC uses Grafana to do exactly this from the heterogeneous data in their data [10] warehouse called the Operations Monitoring and Notification Infrastructure (OMNI) [4].

NERSC hired two trainers from Grafana Enterprise to teach the students how to create visualizations from the data. Each student was assigned a mentor and had a project that would create

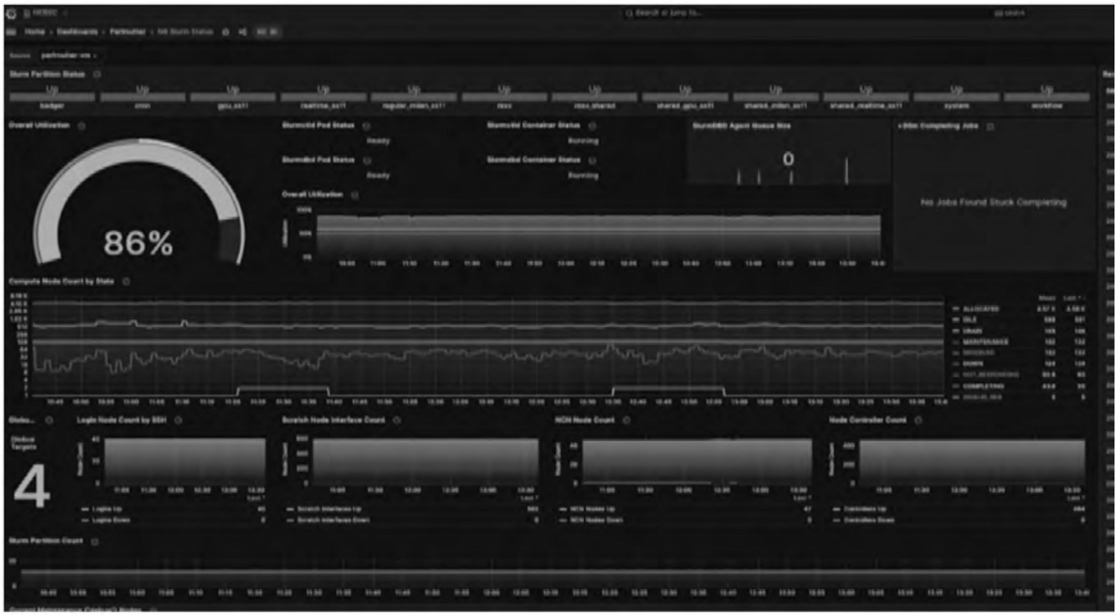


Figure 3: Slurm dashboard for perlmutter

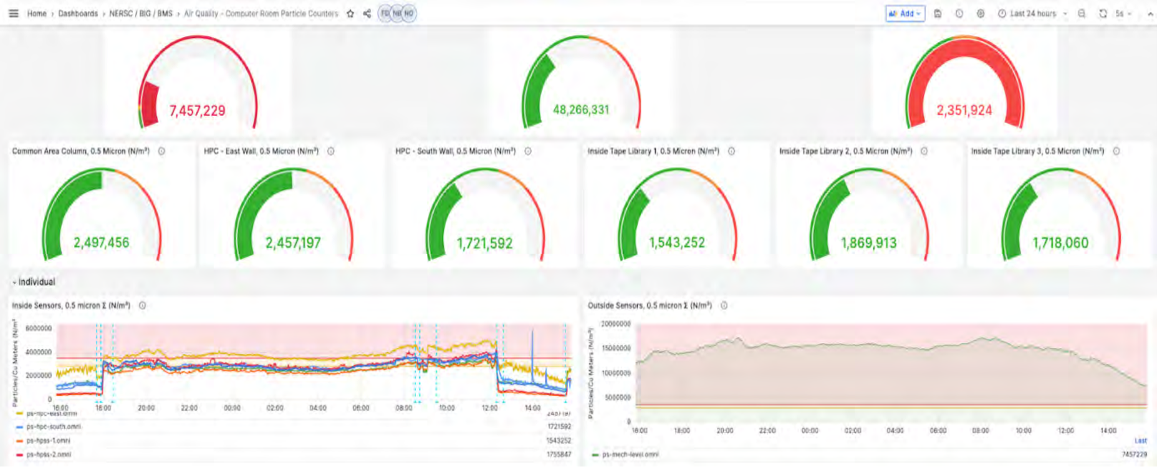

Home › Dashboards › NERSC / BIG / BMS › Air Quality - Computer Room Particle Counters
Add
Last 24 hours
5s
7,457,229
48,266,331
2,351,924
Common Area Column, 0.5 Micron (N/m³)
2,497,456
HPC - East Wall, 0.5 Micron (N/m³)
2,457,197
HPC - South Wall, 0.5 Micron (N/m³)
1,721,592
Inside Tape Library 1, 0.5 Micron (N/m³)
1,543,252
Inside Tape Library 2, 0.5 Micron (N/m³)
1,869,913
Inside Tape Library 3, 0.5 Micron (N/m³)
1,718,060
Individual
Inside Sensors, 0.5 micron Σ (N/m³)
Particles/Cu Meters (N/m³)
6000000
4000000
2000000
0
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
ps-hpc-east.omni
ps-hpc-south.omni
ps-hpss-1.omni
ps-hpss-2.omni
2457197
1721592
1543252
1755847
Outside Sensors, 0.5 micron Σ (N/m³)
Particles/Cu Meters (N/m³)
20000000
15000000
10000000
5000000
0
16:00
18:00
20:00
22:00
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
Last
ps-mech-level.omni
7457229

Figure 4: Dashboard of particle count in tape libraries

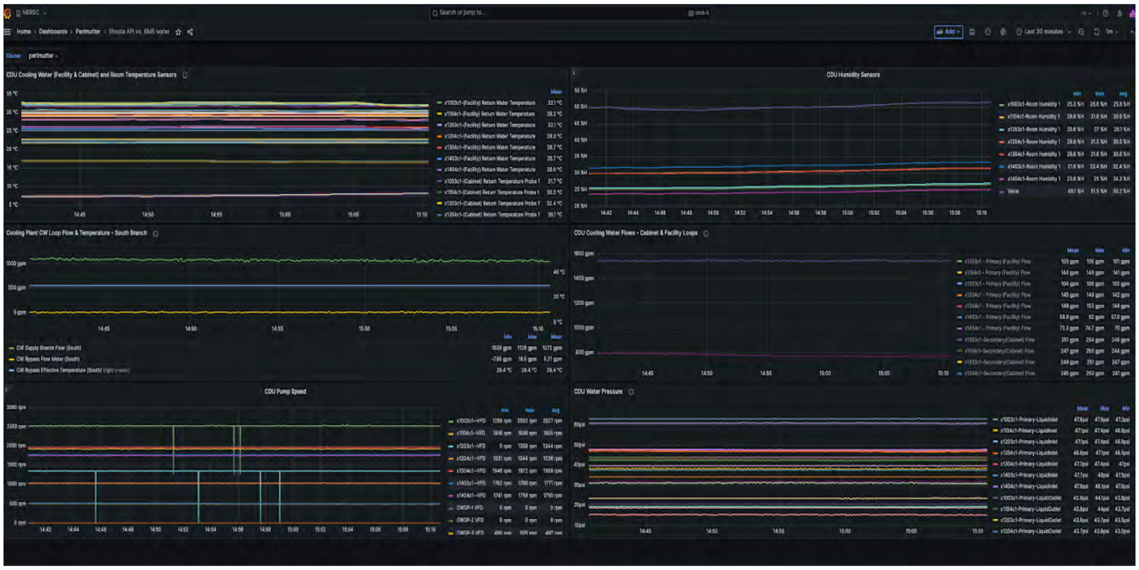

NERSC
Search or jump to...
Home › Dashboards › Perlmutter › Shasta API vs. BMS water
Cluster perlmutter
CDU Cooling Water (Facility & Cabinet) and Room Temperature Sensors
CDU Humidity Sensors
Cooling Plant CW Loop Flow & Temperature - South Branch
CDU Cooling Water Flows - Cabinet & Facility Loops
CDU Pump Speed
CDU Water Pressure
Last 30 minutes

**Figure 5. Pipe water pressure**

a series of graphs to showcase the data, to "have the data tell a story". This supplemented hands-on training as well as their academic background formed a foundation for a successful experience for the students.

## 4 FUNDING AND OTHER CHALLENGES

The program needed to be funded. Though some activities already had their own funding streams, such as the classes already being taught, the new classes needed basic funding, especially for the tutorial lab.

Luckily, there was a grant that the school was able to obtain that would cover three years of the program from inception. Some of the funding also provided for salaries, travel and housing costs for students who are out of the area for the internship program. This only used up less than 10% of the funding in the first year since the main focus was to assist the metro area students. However, beyond three years of funding, it will be a challenge to continue unless they can secure more funding which is the challenge because, it was easy to acquire startup funding, but without proof of success, funding can be a challenge.

However, first we have the challenge of our program being accepted for certification. At the time of the conference presentation of this work, we were only in the first semester after launch. To acquire certification for the program, we must show a success rate of at least two semesters and have a plan for the second year of the program. We are now currently in the second semester and it looks like it could be successful and there is a plan for the third and fourth semesters.

Some of the faculty teaching the program had a difficult time helping the students understand the concepts and they were not quite confident that the students will successfully pass. After a few weeks, the faculty members were encouraged that the tutorial program seems to be helping a lot, therefore, they felt much better by the end of the fall semester to see that they had an 85% pass rate of the classes.

Though the program had some faculty, as the program expands, the college will need more faculty to teach classes. At this point, the administration is currently challenged to recruit other faculty to commit to taking the extension classes and be able to teach in the program. They may have to hire from external to the college however, funding will be an issue. There is a challenge of recruiting more employers to participate in the internship program. While the program was able to support the salaries, travel and housing, that part of the funding was only available in year one and not for year two or three. Though they still have a commitment from the existing employers, as the program expands, more employers need to participate in providing internships. It is a continuing challenge to entice an employer to hire from the community college instead of the four-year university. Further, with the Metro Area having a reputation for being a tech area, most employers expect to hire four-year university students. The pandemic has changed that for the Metro Area, and we are attempting to educate the industry that investing in a community college student can be far rewarding with potential longevity especially in the underrepresented groups who are from the area.

In a positive note, the students who participated in the pilot program have said that it was the most rewarding experience that they've ever had. Participating in the internship was most valuable and they feel they now have a skill that they offer to employers.

## 5 CASE STUDIES

This section will discuss the positive outcomes of three students in the program at NERSC.

Student #1 is the first student who has entered college from an immigrant family. He has taken 2 years of courses in a community college and took classes that were part of the pilot program. After Grafana training, he was able to create various dashboards, one of which was used during the acceptance of NERSC's Perlmutter system, as noted in Figure 3.

While it may be difficult to see the dashboard clearly because it is so full of information, there are two graphics that show system utilization, noted by 86% and where you see the 4, these graphs show that jobs were running on the nodes. There are other details but this is a good summary of what the SRE staff need to see on the system.

Student #2 was a student who was also in his third year of community college and was preparing to get a transfer to a four-year university. Like the first student, he was also the first college student in the family from immigrant parents.

The Metro Area can experience high smoke levels from fires during the summer but can also extend into early fall. Because of the design of the data center, we have no chillers, instead use evaporative cooling system, hot air recycling and air movement to keep the data center cool, therefore, any pollutants external to the facility can impact the air within the facility. Such as the case in late September 2023 where there were a series of fires more than fifty miles away, and high winds blew smoke toward the data center impacting our tape libraries, as noted in Figure 4.

You can see the top level of the dashboard that shows three circles and two are red. This shows that two out of three libraries have a high particle count and is in the critical stage. At this point, library 3, to the right, was shut down to save moving parts, since smoke particles can potentially scratch the moving parts and void the warranty. That said, we needed to find out when this high particle count occurred because it happened very quickly.

The student created the third panel and fourth panel of graphs. The left one shows when the spike of high particles occurred by time in comparison to the panel on the right, that shows high particle count external to the facility. In this way, we can see when the high particle count occurs, put an alert on it and be able to monitor the graphs so we can mitigate it in the data center.

Student #3 was a female student who previously graduated from a four-year university with a marketing degree. However, after one year, she has not been successful in finding a job in her area due to the intense competition post pandemic. Therefore, she decided to enter a community college data science program to learn new skills but minimize her debt.

After the training program she created these dashboards that show the speed of the water in the pipes that help cool the facility on two loops that cool the ambient air around the equipment, as noted in Figure 5.

While the dashboard is difficult to read because it is dense with information, the important graphic is that all these lines are horizontal. Should there be a spike, then the facility will have an issue with their evaporative cooling. Pipes that flow water go through the HPC system and they need to be at constant speed, pressure and temperature. When the speed varies or the pressure varies or the temperature rises, an alarm will alert the SRE on duty who will look at these graphs to find more information to mitigate the situation.

In the three students' experiences, the internship and training helped the grow professionally and develop a new skill. Further, these graphs were put into production and are used daily by the SRE's who work in the data center control room 24x7. The students understand that the work they did has value and gives them confidence that they can do the same for another employer.

## 6 FINAL THOUGHTS AND FUTURE WORK

We've seen the success of just three students at NERSC, however, there were approximately 20 students in the pilot program. The initial cohort who started in September 2023 have 42 students who are looking forward to a summer internship in 2024.

It has been NERSC's experience that learning how to analyze and visualize data is a skill that can be easily learned given that the student has the mathematical or programming foundation. The students are mentored to help understand what the data represents, what is important to show, and eventually create the dashboards that can tell a story to the SRE that helps them mitigate any kind of incidences Funding and finding the right faculty to teach the classes will be continuing challenges but it could be overcome.

Future work involves expanding the program so that we can intake more students, recruiting more employers for internships and exploring multi-semester internships for students who are more senior in the program. NERSC plans to continue to collaborate with Laney College, Oakland, in creating the next step, which is the second year of the Data Science program. For the SRE's at NERSC, OMNI is the center of a monitoring infrastructure that allows them to "see" the health and status of the facility [8].

Visualizing the data allows them to determine issues early, diagnose the problem quickly and come to a resolution as soon as possible in order to continue to serve their global users who use their HPC system and facility. This is the next step in creating a staffing pipeline for new talent and to fulfill the diversity and inclusion mission of the Lab.

## ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DEAC02-05CH11231.

## 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 *ICPP Workshops '19: Workshop Proceedings of the 48th International Conference on Parallel Processing*. August 2019, Kyoto, Japan, 1-9.

[2] Elizabeth Bautista and Nitin Sukhija. 2021. Employing directed internship and apprenticeship for fostering HPC training and education. *Journal of Computational Science 12*, 2, 33-36.

[3] Elizabeth Bautista and Nitin Sukhija. 2023. Creating pathways in disadvantaged communities towards STEM and HPC. *Journal of Computational Science 14*, 2, 2-5.

[4] Elizabeth Bautista, Nitin Sukhija, and Siqi Deng. 2022. Shasta log aggregation, monitoring and alerting in HPC environments with Grafana Loki and ServiceNow. In *Proceedings of the 2022 IEEE International Conference on Cluster Computing (CLUSTER).* September 2022, Heidelberg, Germany. 602-610.

[5] Berkeley Lab. Workforce Demographics. Retrieved from https://diversity.lbl.gov/berkeley-lab-workforce-demographics-fy2022/.

[6] Grafana Labs. Retrieved from https://grafana.com/.

[7] Raphael A. Irizarry. 2020. The role of academia in data science education. *Harvard Data Science Review 2*, 1.

[8] Emad Mushtaha, Saleh Abu Dabous, Imad Alsyouf, Amr Ahmed, and Naglaa Raafat Abdraboh. 2022. The challenges and opportunities of online learning and teaching at engineering and theoretical colleges during the pandemic. *Ain Shams Engineering Journal 13*, 6.

[9] Public Policy Institute of California. California's Need for Skilled Workers. Retrieved from https://www.ppic.org/publication/californias-need-for-skilled-workers/.

[10] Cary Whitney, Thomas Davis, and Elizabeth Bautista. 2016. NERSC Center-wide Data Collect. Retrieved from https://cug.org/proceedings/cug2016_proceedings/includes/files/pap101s2-file1.pdf.