LiveDataLab: A Cloud-Based Platform to Facilitate Hands-on Data Science Education at Scale

Aaron Green

University of Illinois at Urbana-Champaign
aaronng2@illinois.edu

ChengXiang Zhai

University of Illinois at Urbana-Champaign
czhai@illinois.edu

ABSTRACT

We present LiveDataLab, a novel general cloud-based platform that facilitates data science education at scale by enabling instructors to offer hands-on data science assignments using large real-world datasets. Using real course assignments as examples, our demonstration will walk attendees through the process of an instructor deploying an assignment, students working on and submitting assignments, and leaderboard-based competition and automated grading to demonstrate the major functions and benefits of LiveDataLab.

CCS CONCEPTS

• **Social and professional topics** → **Information systems education; Information science education;** • **Applied Computing** → *Interactive learning environments;* • **Software and its engineering** → Cloud computing;

KEYWORDS

data science education, cloud computing, virtual lab

ACM Reference format:

Aaron Green and ChengXiang Zhai. 2019. LiveDataLab: A Cloud-Based Platform to Facilitate Hands-on Data Science Education at Scale. In *Proceedings of L@S '19, Chicago, IL, June 24–25, 2019 (L@S '19)*, 3 pages.
<https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

The growth of the "big data" industry creates an urgent need for educating a large number of data scientists and engineers. Hands-on experience using large real-world datasets allows

learners to gain relevant knowledge and skills useful for solving real-world problems, and is thus an essential component of any serious data science education. However, due to privacy restrictions and the sheer size of these datasets, it is impractical to expect all learners to download them to their local machines. The ability to deploy effective data science assignments at large scale with low cost is thus an urgent challenge that has been hindering data science education.

To address this challenge, we present LiveDataLab, a novel cloud-based general platform that enables flexible data science assignments using large real-world datasets at scale, with very low cost, and for learners of all backgrounds. LiveDataLab is built on top of CLaDS [1], an infrastructure that leverages version control and continuous integration to dynamically create virtual machines to run code that uses remote datasets. It extends CLaDS with a novel dataset annotation module that enables students to collaborate in the curation of new real-world datasets. These new datasets can then be directly used to support not only new leaderboard-style competition assignments but also new research in data science, effectively integrating education and research in data science. Ultimately, by engaging industry partners to provide datasets, LiveDataLab would seamlessly integrate data science education, research, and applications in a unified self-sustaining big data ecosystem at scale where students are simultaneously contributors (contributing new datasets and new algorithms) and beneficiaries (learning from hands-on assignments using industry datasets). The student interaction log on LiveDataLab further offers great opportunities for automating assessment of student performance and providing feedback.

2 DESIGN AND ADVANTAGES

The backbone technology of LiveDataLab is described in the original CLaDS paper [1]. On top of CLaDS, LiveDataLab further introduces a dataset annotation module within the same cluster as the original CLaDS system, allowing students to collaborate in creating new, labelled, real-world datasets. The datasets can then be used in assignments for the course. This enables students to experience many facets of data science, from collecting the data by writing scripts to parse webpages, to helping annotate the dataset, to utilizing the complete dataset and participating in a search engine

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
L@S '19, June 24–25, 2019, Chicago, IL

© 2019 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-9999-9/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

Figure 1: Functional overview of the LiveDataLab system. In practice, there are many students working on a variety of assignments and an autoscaling number of grading servers.

competition. Student-sourced creation and annotation of a dataset at this scale is otherwise not possible without the LiveDataLab system. This provides students with an end-to-end experience from curation to competition in a hands-on, collaborative data science ecosystem (see Fig 1). The generality of the LiveDataLab system allows for integration with a variety of existing learning infrastructures, including MOOCs, to provide hands-on data science education to a more diverse demographic than traditionally reached by the university systems, enabling data science education at the largest and most accessible scale.

3 USE CASE

LiveDataLab is currently supporting over three hundred students at the University of Illinois at Urbana-Champaign in two courses: We have leveraged this technology to support text retrieval algorithm analyses and search engine competitions using real-world datasets for CS 410: Text Information Systems, and we have simultaneously used it as an autograding environment in ECE 418: Image and Video Processing. In CS 410, we leveraged the dataset annotation module to deliver a multi-stage assignment, enabling students to gain end-to-end experience building and improving a search engine for finding experts. Initially, they wrote scripts to crawl homepages of computer science faculty across hundreds of universities. The crawled pages were then aggregated to form a large collection of documents deployed in the LiveDataLab cloud infrastructure. Students were then given an annotation task where they would create a realistic query (e.g., 'uiuc text mining') and mark whether or not the returned faculty homepages were relevant to the query. Finally, we

used the newly labelled dataset in a search engine competition where students wrote custom search engine algorithms, and we utilized the student-created annotations to quantitatively evaluate their submissions using a leaderboard. Overall, we created 4,720 annotations for 283 unique search engine queries with 6,598 total documents in the dataset. This new dataset will be released to facilitate future research.

In its current state, LiveDataLab provides a general infrastructure enabling many types of assignments in a cost-effective, customizable, and autoscaling environment, which can be leveraged for many future use cases.

4 FUTURE DIRECTIONS

We are currently working on LTI integration to allow courses offered through Coursera and other LMS to utilize the LiveDataLab infrastructure to provide hands-on data science education to an even larger number of learners through MOOCs and other online courses. Along with LTI integration, students can use the Gitlab Web IDE to transform the LiveDataLab into a complete workspace where they can edit and run code on the autoscaling platform with real-world datasets all in the browser, thereby increasing productivity by not having to worry about any local setup as well as providing an accessible environment to users of all backgrounds and experience levels. We are also developing a user-friendly interface around the Gitlab system to provide a streamlined process for new instructors to get onboarded to the system, deploy assignments, and upload datasets. This will allow LiveDataLab to easily support many more courses and provide learners hands-on data science education at the largest scale possible.

5 DEMONSTRATION

Our demonstration will allow attendees to experience the process of deploying new assignments, working on and creating submissions for an assignment, automated grading for assignments, and participating in leaderboard-style search engine competitions. Attendees will also be able to use the annotation module to submit annotations for our new computer science faculty dataset. Additionally, we will seek feedback from attendees about the system and components they found valuable, as well as features they wish the system had or any inconveniences they experienced throughout their usage, allowing us to continually improve the system to provide the best scalable hands-on data science education possible.

REFERENCES

- [1] Chase Geigle, Ismini Lourentzou, Hari Sundaram, and ChengXiang Zhai. 2018. CLaDS: a cloud-based virtual lab for the delivery of scalable hands-on assignments for practical data science education. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3197091.3197135>