LiveDataLab: A Cloud-based Open Lab for Integrating Big Data Research, Education, and Applications

ChengXiang Zhai
Siebel School of Computing and Data Science
Grainger College of Engineering
University of Illinois at Urbana-Champaign, USA
czhai@illinois.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825402

***Abstract*—We present the vision of LiveDataLab and discuss the new research directions and application opportunities it opens up. LiveDataLab is envisioned to be a cloud-based open lab infrastructure where research, education, and application development in big data can be integrated in one unified platform, thus accelerating research, technology transfer, and workforce development in big data.**

***Index Terms*—data lab, reproducible research, project-based learning, continuous integration, academic-industry collaboration**

I. Motivation

Despite significant growth of big data research, commercial products in big data analytics have not kept pace, and many innovative algorithms developed by researchers, including those developed many years ago, have not yet made into a commercial product. While it is not unusual that industry adoption lags behind research, there are also multiple specific barriers that have hindered the growth of big data industry, which, if not addressed, would continue limiting the growth of industry and impact of research in big data:

Challenges in data science research: The difficulty in access to large real-world data sets makes it hard for researchers to align their problem formulation with the actual needs in the industry. The complexity of big data experiments makes it hard to describe all the details of an experiment in a paper to enable replication of an experiment. However, reproducibility is crucial for making scientific progress since unreliable conclusions may mislead research into a non-productive research direction. The lack of reproducibility further hinders research productivity as it would require more effort to compare a new method with the existing ones.

Challenges in data science technology transfer: The non-reproducibility of research results also hinders technology transfer. In order for a new algorithm to be used in industry, the algorithm must be implemented correctly with the right configuration and optimal parameter setting. Unfortunately, only some of the published algorithms are available in the form of software code. Even when the researchers have made the software code open source, it is often still quite challenging for someone else to learn how to use it correctly and set all the parameters correctly, creating another gap between data science research and applications.

Challenges in data science education: No current training program seems able to support hands-on assignments with large real-world datasets. The challenge here is not just because the sheer size of a dataset would make it infeasible for learners to download the dataset to a local machine, but also because data sharing creates serious concerns of privacy infringement; even training providers generally cannot obtain real world datasets either, forcing learners to work on assignments involving small and/or artificial datasets. Unfortunately, the observed behaviors of algorithms on small artificial datasets generally do not reflect their behaviors on large real world datasets, creating a gap between what learners have learned and the skills required to solve a real-world problem (i.e. gap between education and application). This gap makes it difficult for companies to recruit competent data science workforce with exactly the required skills for their jobs. The shortage of competent big data workforce directly hinders the growth of the big data industry.

We propose a novel cloud-based open lab infrastructure that can address all these challenges simultaneously. The proposed infrastructure, called LiveDataLab, can seamlessly integrate big data research, education, and development so that researchers would be able to work on realistic problem formulations based on real world data sets, the learners would be able to learn precisely the hands-on skills required for working on an application project, and the new algorithms developed by researchers would be easily integrated, tested, and adopted in a commercial product. LiveDataLab is a self-sustainable big data innovation ecosystem that brings together multiple stakeholders, including learners, educators, researchers, and industry application developers, where everyone is simultaneously a contributor and beneficiary. Learners use it to obtain project-based training using cutting-edge toolkits and real-world big data sets with an intelligent training agent to provide just-in-time support and help, while also contributing new annotated data sets and new algorithms; educators use it to deploy realistic data science assignments with leaderboard competition, while also contributing training projects and

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

materials; researchers leverage the new data sets created to conduct novel data science research and advance state of the art technology by contributing open-source implementation of new techniques; industry partners leverage it to train and recruit relevant workforce precisely and acquire the newest technology developed by researchers and learners directly for use in their products, while also contributing realistic real-world data sets and tasks. As a result, LiveDataLab would enable data science education at large scale with low cost, minimize the gap between education, research, and applications of data science, and accelerate data science research as well as technology transfer.

II. Vision of LiveDataLab

The key insight that shaped The vision of LiveDataLab is the observation that a root cause of the multiple challenges discussed in the previous section is that researchers and learners cannot experiment with real-world data sets, thus a good solution would be to address this root cause. Since the data sets cannot be downloaded (accessed directly) by researchers or learners, the solution logically would have to build a data laboratory (i.e., LiveDataLab) in the place where the real world data sets are generally stored, i.e,. a cloud computing platform. This way, researchers and learners would be able to use such a cloud-based lab to experiment with algorithms on very big real-world data sets and test their ideas for improving the algorithms in the same way as using a biology or chemistry lab to perform biology or chemistry experiments (in biology and chemistry, researchers and learners must go to a lab to do the experiments because they cannot take the experiment device and sample home, which is similar to the situation of big data experiments, in which the researchers and learners cannot ”take the datasets home” either).

Essentially the LiveDataLab would enable us to move algorithms to where the data set is so that we can protect the privacy (if needed) and test algorithms on the cloud using potentially very large data sets that the researchers and learners do not have direct access to; this is in contrast with the conventional way of having the data moved to the machines where students would develop and test their algorithms. Specifically, when working on an experiment, a researcher or learner would first check out the source code of a software toolkit, work on modifying the code locally with a small data set for testing, and then submit the revised code (the revision could range from changing parameter configurations to adding completely different, but compatible new algorithms) to the lab where the software code would be compiled and tested on some big data sets on the cloud. A leaderboard can be further used to publish the performance of algorithms developed by researcher or learner for solving a real-world data science problem/task; when a new data set or a new data science challenge is deployed on LiveDataLab, it would enable researchers to work on a new research problem and advance the state of the art of big data research. Since learners and researchers can both use the LiveDataLab to perform experiments in exactly the same way, LiveDataLab naturally integrates research and education. With the recent progress in Generative AI, it would be also feasible to deploy an intelligent tutoring agent on LiveDataLab to provide any user of LiveDataLab just-in-time support (e.g., to help them with background knowledge, recommend relevant online tutorials, or answer their questions), enabling a novel project-based learning paradigm where a data project is a unit of learning (a learner can choose the most relevant project/skill to learn on LiveDataLab). From application perspective, since all the submitted code would be archived on LiveDataLab as open-source tools, application developers can easily access the best-performing algorithms on relevant leaderboards along with the exact parameter configuration that has led to the best performance and use them in their application systems, making technology transfer highly efficient. It is in this sense, LiveDataLab also integrates Application with Research and Education. While LiveDataLab is similar to some existing infrastructure, notably Kaggle, there is a key difference in that LiveDataLab emphasizes continuous integration of code, which is key to achieving many benefits that none of the existing infrastructures have.

We illustrate the architecture of LiveDataLab in Figure 1, where we see that the cloud-based infrastructure (in the middle) consists of the following modules: 1) Open source toolkits, which are meant to have all the implementations of all kinds of big data algorithms, which would be continuously improved as learners contribute potentially better algorithms. 2) Datasets, which may include all kinds of realistic datasets publicly available and any synthetic datasets that we can create to simulate real-world big data sets. As needed, the datasets can be protected by setting appropriate permissions to restrict its access to only authenticated users. Note that a major benefit of LiveDataLab is that even though a learner or researcher has no permission to access a dataset, they can still upload their algorithms to evaluate their algorithms on all the datasets, thus enabling privacy-preserving research and experimentation with real-world proprietary datasets. 3) A collect of data projects that each is designed based on a particular data set and a particular open source toolkit (we view a single software tool as a special case of a toolkit) along with evaluation metrics that can generate quantitative performance numbers for an algorithm to be shown in a leaderboard. 4) Leaderboards that serve as a detailed documentation of the results of all the submitted big data experiments, which are generally organized based on specific projects. 5) Experiment manager, which processes all the experiment requests, general experiment results, and update the related leaderboards. Specifically, when a learner submits an experiment request, including a modified open source toolkit of a project (either with a parameter variation or a new algorithm), the Experiment manager would dynamically create a virtual machine on the cloud to compile the submitted code, and execute it on the specified dataset(s) to generate performance numbers, which would then be added to the relevant leaderboards. The Experiment Manager would then delete the virtual machine. Such an elastic design would minimize the cost of using virtual machines and ensure robustness against any attack of malicious code submission (since

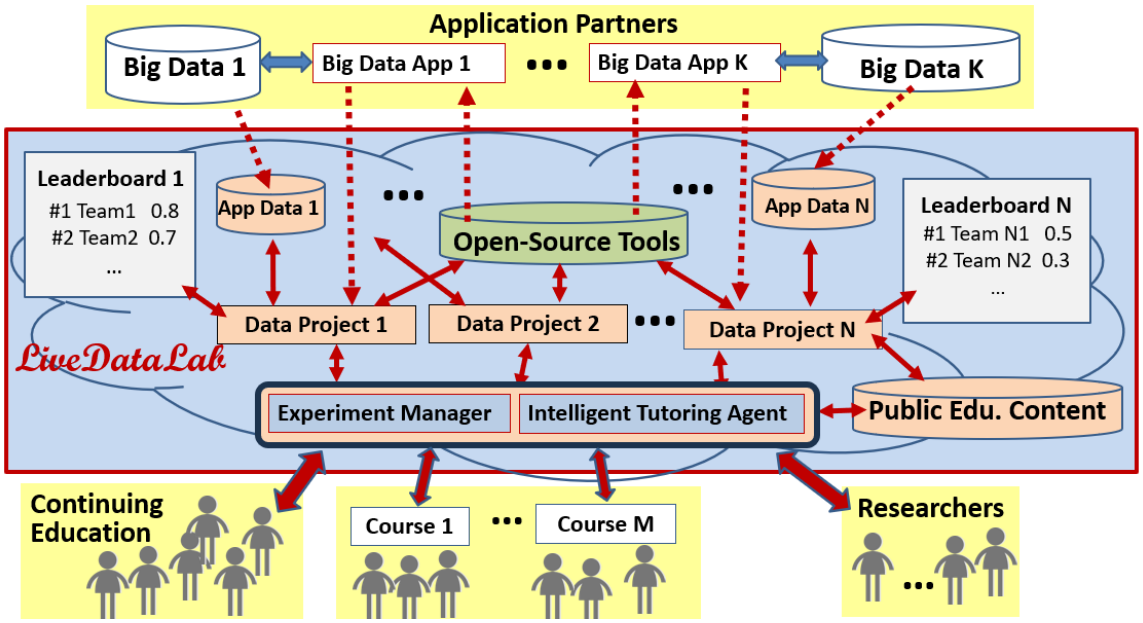

Fig. 1. Illustration of LiveDataLab as an infrastructure and ecosystem for integrating research, education, and application development in big data.

the virtual machine is disconnected with other machines, the worst case is just a malicious code submission would crash the virtual machine alone). The experiment submitter would also receive feedback about the experiment results without being able to see the details of the data sets, which may help the submitter improve experiments. 6) Project-based learning support module, which mainly consists of an AI-powered personalized intelligent tutoring agent that can provide a learner just-in-time tutoring, support, and feedback as needed. It leverages all the publicly available educational content and uses the observed learner interaction data with LiveDataLab to infer a learner's knowledge state and provide personalized recommendation of relevant tutorial readings or videos for the learner to study as needed. This enables learners with variable backgrounds to start working on a project for learning skills of big data analytics immediately without necessarily having all the needed background, i.e., enabling a new paradigm of workforce training with a project, instead of a course, as a training module, which would be especially for upskilling since it would enable a learner to acquire precisely a skill needed for a target big data job with minimum amount of time and financial investment.

The bottom part of the figure shows that multiple types of users, including skill upgraders (continuing education), students of university courses or degree training programs, and researchers, can all benefit from using LiveDataLab to learn or do research. From the industry's perspective (shown on the top), LiveDataLab not only trains workforce with highly relevant skills immediately useful for working in the industry, but also facilitates recruiting of the best learners based on their performances on the most relevant leaderboards. Furthermore, the improved algorithms developed by the learners and researchers can be continuously integrated into an open source toolkit, which can directly power an application system, creating a self-sustainable ecosystem where new technology is integrated into an application system continuously. This gives industry incentives to support the running of LiveDataLab financially, which further helps reduce the cost for learners, creating a self-sustaining talent ecosystem.

LiveDataLab offers two benefits for research in data science. First, researchers would not need to reproduce any baseline results since they are already all available and archived on LiveDataLab, thus improving productivity of research. Second, new research results are guaranteed to be reproducible since the data sets, the implementation of any new methods and any configuration of parameters in the experiments are all well archived on LiveDataLab. If we place the data sets behind a firewall, LiveDataLab can also enable privacy-preserving research in data science by sending algorithms to the place of data sets and only publishing the experiment results on the protected data sets without details of the data sets themselves.

With LiveDataLab, learners, educators, researchers, and application developers can collaborate efficiently in a potentially self-sustaining ecosystem where everyone is simultaneously a contributor and beneficiary. Industry would partner with academic institutions to continuously create new data projects on LiveDataLab to achieve their shared goal of improving big data education and workforce training. New annotations of data sets can be created continuously by crowdsourcing annotations using annotation assignments that enable learners to learn how to evaluate algorithms. The new data sets enable new research in data science and new leaderboard-based competition tasks, facilitating invention of new technology, which would be immediately available for the industry to adopt via the continuous integration mechanism of LiveDataLab, giving additional incentives to industry for contributing to LiveDataLab. LiveDataLab also enables industry to recruit more precisely the "right" workforce for their jobs based on

the performance of learners on the most relevant leaderboards (in some sense, LiveDataLab pre-trains the workforce for the industry for free), adding even more incentives for industry to participate in such an ecosystem and even provide financial support for sustaining the LiveDataLab cloud infrastructure. For example, an industry alliance may potentially share all the cost of sustaining the cloud computing infrastructure of LiveDataLab, thus enabling all learners and researchers to use the infrastructure for free, which would attract more learners and researchers to use the infrastructure, who would make LiveDataLab more attractive to industry for recruiting and technology transfer.

III. Preliminary implementation experience

To study the feasibility of LiveDataLab, we implemented a basic infrastructure of LiveDataLab on Azure cloud computing platform by leveraging Gitlab to build the experiment manager, called CLaDS [2]. The CLaDS infrastructure was successfully deployed to host all the programming assignments with leaderboard competitions and automated grading for a course (CS410 Text Information Systems) that the author taught at UIUC with about 200 on-campus students and over 200 online students (who are mostly professionals with a full-time job). Historically, the course could only use very small "unreal" data sets for those assignments, but CLaDS, for the first time, enabled the course to use much larger real-world data sets for all the assignments, thus enabling students to learn skills that can be directly useful for solving real world problems. Moreover, thanks to the elastic virtual machine creation strategy used in CLaDS, the estimated cost was as little as $7.40 per student per semester [2], demonstrating the potential for scaling up the support of many more learners. Indeed, a major benefit of a cloud-based infrastructure such as LiveDataLab for learning is that all the deployed assignments on LiveDataLab can be easily made available to any learners around the world that have an Internet connection to allow them to compete on the same leaderboard, thus enabling collaborative learning at scale.

To support data annotations, we extended CLaDS with an additional data annotation system, called COLDS [4], [5] so that students can use a Web interface to annotate data sets used for evaluating a search engine. We also added a project-based learning module and made CLaDS more general by leveraging a Jenkins cluster for managing virtual machines on demand, a Flask API for submitting and managing jobs to be executed on the Jenkins cluster, and a core web application with a React front-end and Flask API, which operates as the project-based learning platform. The details are described in a Master Thesis [3]. The new version of LiveDataLab has been used successfully by the author in teaching CS410 every year since then with multiple additional programming assignments deployed every year, further confirming the feasibility of LiveDataLab even with our basic implementation. Student feedback shows that many of them were able to leverage LiveDataLab to experiment with multiple algorithms on realistic data sets and improve over baselines by proposing their own ideas. For example, the following is an excerpt from the notes of a top-performing student on a leaderboard, where we can see clearly the student was doing interesting research on a realistic real-world dataset on LiveDataLab: *" ... [Testing the idea of] Pseudo feedback: since we have the best ranking for BM25 and MPtf2ln, Can we combine the ranking output of these two functions? New Ranking merge this two ranking function's output, ... With above methods, I received MAP 0.6962 on the Phase 1 Validation Leaderboard, by far the highest score on the leader board."*

We also explored the capability of LiveDataLab in supporting integration of education and application using a comprehensive search engine assignment of CS410, where the students have collectively built an expert search engine with algorithms designed by themselves via multiple modularized assignments. Specifically, we first asked the students to work on a data crawling task, where each of them crawled the computer science faculty homepages from a different university in the United States. We then combined their crawled data into one test collection, which can be potentially used to build an Expert Search Engine (e.g., finding an expert working on security and machine learning). We then asked each student to come up with a sample query for expert search and judge the top-ranked documents from the test collection that they created to annotate relevance of each of them, and combined all their annotations to obtain a labelled test collection for evaluating search engine algorithms. Finally, we deployed a search engine toolkit and the dataset that they created on LiveDataLab to allow them to compete in designing effective ranking algorithms. The best algorithm from the student as shown on the leaderboard was then used to power a working Expert Search Engine (with a search interface added). This experience has been described in an ACM SIGCSE paper [1]. In the following semesters, we used the same assignment but asked students to crawl additional faculty homepages (outside United States) and they continued to add more annotations as well as further improved the ranking algorithm, thus enabling us to continuously expand the collection and improve the ranking algorithm for the Expert Search Engine developed solely by collaborative effort of students from different classes. This experience demonstrated the feasibility of leveraging LiveDataLab to continuously improve a big data product by opening some components of a complex system via LiveDataLab for learners and researchers to work on improving. The new (better) algorithms developed by them would be available on LiveDataLab and can thus be immediately adopted to improve the product, without any gap in technology transfer.

In sum, our overall experience with the preliminary implementation of LiveDataLab was quite positive and confirmed the great potential of LiveDataLab for integrating research, education, and application development.

IV. Future Applications Opportunities and Research Challenges

The vision of LiveDataLab opens up many new interesting application opportunities and research challenges; a thorough

discussion of them is out of the scope of this vision paper, but we will briefly discuss a few major ones here.

Reproducible Research: The continuous code integration enables LiveDataLab to naturally ensure reproducible research since all the algorithms and configurations are all archived by design. The availability of all the experimental data further enables additional analysis of those results by anyone to further our understanding of relative strengths and weaknesses of different algorithms. The archiving of all experiment results further enables researchers to focus their effort entirely on inventing new algorithms without having to reproduce baselines. The main technical challenge in realizing this goal is how to develop an efficient leaderboards manage system to support fast access to all the experimental data, which may require intelligent algorithms for compressing the data without compromising performance of real-time access. The system also needs to deal with parallel submissions potentially by many researchers at the same time. There will be significant challenges in developing a scalble efficient research data management system. Once the system is built, there will be many interesting new research opportunities on designing research data analysis algorithms to digest massive amounts of experiment results. Joint analysis of experiment results of related algorithms would be especially interesting (e.g, enabling comparison of different families of algorithms that solve the same problem).

New research via data simulation: LiveDataLab enables industry partners to set up many new data challenges, thus accelerating research. However, data privacy remains a challenge that needs to be addressed. Although we can protect a large real-world data set with a firewall, this is not the most efficient solution since a researcher would not be able to see any detail about the data (they would only see performance), which makes it hard to improve their algorithms. Running every algorithm on a large real data set is also inefficient. A better solution would be to use a generative model trained on the real data to generate synthetic data, which can be seen by researchers. The synthetic data can also be made much smaller than the original data, enabling many experiments to be run efficiently. However, an important challenge here is how to ensure that the synthetic data does not leak any critical information in the original data set, thus requiring new research in security and data privacy protection.

New project-based learning paradigm: With our basic implementation, LiveDataLab can already be extended to serve as a general learning platform for supporting many hands-on assignments in Data Science with realistic data projects reflecting real-world needs. The current generation of online learning platforms such as Coursera, EdX and Utdacity, have "moved" the lectures online. As an extension of such platforms, LiveDataLab can be regarded as a general infrastructure to further "move" a laboratory online to enable all kinds of learners to acquire hands-on experience with working on realistic data sets. LiveDataLab naturally enables all the learners to compete on the same leaderboard and collaborate on learning relevant knowledge and skills with help from an intelligent tutoring agent, regardless where the learners are from (they can be from different university or different countries). While standard degree programs are useful for "formal training," for upskilling, a degree program is both expensive and taking an unnecessarily long time to complete. Employees of a company may only need to periodically learn new skills relevant to their tasks. For such people, project-based learning or skill-based learning would be much more useful and would enable them to acquire precisely the needed skill quickly without spending too much time or money. LiveDataLab can be turned into such a project-based learning platform by using each leaderboard-based assignment as a "learning module" and recommend related materials such as lecture videos and readings to learners in the context of a specific project. To make this happen, a major research challenge is how to develop a general intelligent tutoring agent by leveraging recent advancement in Large Language Models (LLMs) that can both understand the specific tasks of a data project and provide just-in-time support in a personalized manner, which further requires modeling a learner's knowledge state in handling a data science problem.

Continuous improvement of big data applications: LiveDataLab can be used to "open up" any real-world big data application system by exporting any component of the system and setting up a leaderboard-based challenge to engage researchers and learners in improving the component, which not only facilitates education and new research in big data but also enables the real world application system to be continuously improved by adopting the newest algorithms developed by big data and AI researchers. At the same time, the end users of the application system can also naturally contribute annotations to enable new algorithm research in data science while interacting with the application system. The major challenge here is how to design an effective mechanism to facilitate industry and academic institutions in collaborating on designing new data projects compatible with a real world application. How to generate synthetic data to simulate real data with privacy protection is a major technical challenge as discussed above.

V. Acknowledgments

This work is supported in part by NSF under Award number 2229612 and by the SRI Program and IIDAI at the University of Illinois at Urbana-Champaign..

References

[1] Bhavya, A. Boughoula, A. Green, and C. Zhai. Collective development of large scale data science products via modularized assignments: An experience report. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 1200–1206, 2020.

[2] C. Geigle, I. Lourentzou, H. Sundaram, and C. Zhai. 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*, pages 176–181, 2018.

[3] A. Green. Livedatalab: A cloud-based platform for data science education. Master's thesis, University of Illinois at Urbana-Champaign, 2020.

[4] A. Green and C. Zhai. Livedatalab: A cloud-based platform to facilitate hands-on data science education at scale. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–2, 2019.

[5] X. Yu. Design and implementation of the search engine module in colds. Master's thesis, University of Illinois at Urbana-Champaign, 2018.