Multi-dimensional and Customizable Open-Source Labware for Promoting Big Data Analytical Skills in STEM Education

Ying Xie, Kai Qian, Jing (Selena) He

Department of Computer Science
Kennesaw State University
Kennesaw, Georgia, USA
{yxie2, kqian, jhe4}@kennesaw.edu

***Abstract*—In order to remove resource barriers and smooth the learning curve for education on big data analytics in STEM disciplines, we develop an portable open source labware that is called STEM-BD for promoting education on big data analytics. STEM-BD integrates the following four critical components, big data platform, big data sets, data analytics algorithms and hands-on lab exercises in a multi-dimensional and customizable way. In this paper, we provide a detailed description of the design goal of STEM-BD, its prototype, preliminary evaluation results, and future development.**

Keywords-component; big data; labware; STEM education;

I. INTRODUCTION

Big data is becoming a pervasive characteristic of our society, leading to a revolution in the way we understand the world and make decisions [1, 2, 3]. Qualified professionals for big data analysis are in a critical shortage. However, big data analytics skills are currently underrepresented in traditional STEM education. Especially for less-resourced universities and colleges, the paucity of dedicated faculty and lab facilities constitutes a substantial challenge for their participation in promoting big data education.

To meet the emerging workforce and address the above challenges, this project advocates the "learning by doing" approach and aims to develop an innovative open-source labware, which is denoted as STEM-BD, to enhance students' learning of big data analytical skills in STEM disciplines. Four key components have been identified in the design of STEM-BD, including the big data analytics platform, the big data sets, the analytical algorithms, and the hands-on labs. For the computing platform, the labware provides Linux based virtual machine (VM) images for three outstanding big data computing platforms, namely Hadoop, Spark, and HPCC. Students can create an instant big data analytics platform that is ready for analytical tasks by downloading and launching VM image of interest. For the data component, a group of real-world datasets with reasonable volume and complexity have been selected from different STEM disciplines to be incorporated in the labware. In the algorithm components, a wide variety of machine learning algorithms have been chosen from the open-source packages designed for each of the three big data analytics platforms. In the lab component, hands-on practices that integrate different datasets and algorithms for well-defined learning objectives of analytical skills have been constructed. The goal of STEM-BD is to promote big data analytical skills in STEM education with the following intellectual merits. (1) Portable open source labware. The project will provide portable computing environment with hands-on learning resources based on virtual machine technology, removing the barrier of setting up high cost distributed infrastructure for learning big data analytical skills. (2) Multi-dimensional and customizable design. Learning materials and resources in different components can be customized flexibly via a multi-dimensional approach, satisfying different learning objectives and learning levels. (3) Smooth learning curve. The easy-to-adopt and customizable characteristics of STEM-BD will provide a smooth learning curve for computing students' understanding the principals of big data analytical skills

II. RELATED WORK

STEM-BD is built upon existing efforts on big data research and educational projects, curriculum development, as well as data repository construction in related areas. Kafura et al. [4] proposed crafting authentic and engaging learning approach using "big data" with a focus on visual interaction interface to explore an approach that weaves together the curriculum, pedagogy, and tools to engage learning using "big data". Their work focused on using the data, rather than on enhancing the analysis skills for real world problem solving. Several big data related courses have been offered in computing curricula in recent two years, such as [5, 6]. Nevertheless, most of them were dedicated CS graduate courses, which might suggest high requirements for learning big data analytical skills. Instead of offering another version of dedicated big data courses, STEM-BD promotes education on big data analytics in STEM disciplines by using the "learning by doing" approach. In other words, STEM-BD tries to provide an easy-to-access learning platform for non-computing majors to practice big data analytics.

III. STEM-BD FOR BIG DATA EDUCATION

One unique aspect of STEM-BD is that it integrates multiple necessary components for Big Data analytics in multi-dimensional and customizable way. Those components include big data sets from different problem domains, big Data analytic platforms, and big data analytic algorithms. The STEM-BD provides a multi-dimensional approach for students to form a big data analytics lab

exercise by selecting components on different dimensions from the pail. As illustrated in Fig. 1, a basic lab exercise on big data analytics can be a simple combination of a big dataset from a domain of interest, an analytics algorithm and a suitable big data platform. A more advanced lab exercise may involve multiple components from one dimension. For examples, a real-world related lab exercises may require a pipeline of analytical processes that involve multiple algorithms; comparative studies of big data analytics platforms may analyze a group of selected data sets using a group of selected algorithms on multiple platforms; a more complex analysis may need to link multiple data sets cross different domains. In the following, we will describe the components on each dimension in details.

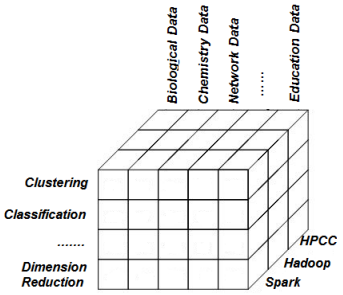

Fig. 1. Multi-dimensional design of STEM-BD

A. Big Data Analytics Platform

The STEM-BD project produced Linux based VM images for three outstanding big data analytics platforms including Hadoop, Spark, and HPCC. Students can easily create an experimental environment of any of these platforms by downloading and playing the corresponding VM image on a local computer. With its unique strengths, each of these platforms has been favored by different user communities. Therefore, the STEM-BD provides maximum flexibility on comparative studies of big data analytics platforms. These three big data platforms are briefly introduced below.

- Hadoop: Hadoop framework was developed as an open-source product and widely adopted for big data analytics by the academic and industrial communities. Hadoop Distributed File System (HDFS) and MapReduce are two main building blocks of Hadoop. MapReduce engine takes care of executing the application by moving binaries to the machines that have the related data. Hadoop cluster can be deployed on any number of commodity computers and is robust against failures in a distributed environment.
- Spark: Spark is a more recent big data platform developed at UC Berkeley. This platform offers a general-purpose programming interface for interactive, in-memory data analytics of large datasets on a cluster. Spark has proved to be 20X faster than Hadoop for iterative applications, was shown to speed up a real-world data analytics report by 40X, and has been used interactively to scan a 1 TB dataset with 57 seconds latency [7].
- HPCC: High Performance Computing Cluster (HPCC) is an open source big data analytics platform developed by LexisNexis Risk Solutions. HPCC can be created by using any number of commodity computers. Major components of an HPCC system include a data refinery cluster called Thor and a data query cluster called Roxie. A performance comparison of HPCC with Hadoop shows that, on a test cluster with 400 processing nodes, HPCC is 3.95 faster than Hadoop on the Terabyte Sort benchmark test [8].

B. Big Data from Different Problem Domains

We carefully selected a group of real-world data sets cross different problem domains to be used in STEM-BD. The data sets we chose are not over large or too complex, yet not small or simple in any sense. In other words, the complexity of the data is reasonable so that this portable labware is suitable for not only learning a specific data analytics method such as classification or clustering, but also experimenting explorative data analysis with adjustable complexities. STEM-BD provides URL links for students to download selected data sets from their original websites. We will keep adding new data sets across different domains to the STEM-BD. STEM-BD is also open for learning communities to contribute or suggest new data. The initial data sets that are included in the first version of STEM-BD are described as below.

- Biology Domain:
  - Yeast Protein-Protein Interaction Data with 162 derived features [9]
  - Tobacco Smoke Effect on Maternal & Fetal Cells [10]
- Physics Domain:
  - Higgs Boson Competition Data (800,000 Events with 30 features) [11]
  - SDSS Quasar Catalog (46,420 records with 23 features) [12]
- Engineering Domain:
  - Network Intrusion Data (8,050,290 records, 41 attributes) [13]
  - Sensorless Drive Diagnosis Data Set (58,509 records, 49 attributes) [14]
- Text Data Set:
  - The REUTERS-21578 TEXT DATASET (21,578 documents) [15]
  - 20 News Groups Dataset [16]
- Data Set for Recommendation
  - Audioscrobbler Data Set [17]
  - MovieLens Dataset (100,000 ratings) [18]

C. Big Data Analytics Algorithms

Each selected big data platform is associated with some well-developed open source machine learning packages. The VM image that STEM-BD provides for each big data platform contains an IDE configured to be ready-to-use of the corresponding machine learning package. STEM-BD

associates each of the data set included in the labware with one or multiple machine learning algorithms that are suitable for this data set and further provides code samples on applying each of those algorithms to that data set. Therefore, a lab exercise can be easily created for students to learn a particular analytical task on the selected data set. The open-source machine learning package that STEM-BD uses on each big data platform is shortly described below

- Hadoop platform (MapReduce): Apache Mahout [19] is configured for Hadoop VM. The following algorithms are chosen to be associated with one or multiple data sets in the labware. Classification algorithms include Naïve Bayes and Randome Forest; Clustering algorithms include K-means, fuzzy K-means, and Spectral Clustering; Dimensionality Reduction includes SVD, PCA; Topic Modeling includes: Latent Dirichlet Allocation. Collaborative Filtering includes: User-based Collaborative Filtering, Item-Based Collaborative Filtering, and Matrix Factorization with ALS. The major programming language is Java
- Spark platform: MLlib [20] is configured for Spark VM. The following algorithms are chosen to be associated with one or multiple data sets in the labware. Classification algorithms include SVMs, logistic regression, naïve bayes, decision trees, and random forest; Clustering algorithms include k-means and Gaussian mixture; Recommendation via Alternating Least Squares (ALS); Dimensionality Reduction includes SVD, PCA; Topic Modeling includes: Latent Dirichlet Allocation. The major programming language is Scala.
- HPCC platform: ECL-ML[21] is configured for HPCC VM. The following algorithms are chosen to be associated with one or multiple data sets in the labware. Classification algorithms include logistic regression, naïve bayes, perceptron, decision trees and random forest; Clustering algorithms include k-means; Dimensionality Reduction includes Stacked Auto-Encoder and Deep Belief Network. The major programming language is ECL.

IV. SAMPLE LAB PROJECTS FOR EXPLORATIVE DATA ANALYTICS

Students who use STEM-BD can starts with applying a single analytic algorithm to a selected data set on the platform of interest. Sample labs to learn a single analytic algorithms provided by STEM-BD are shown in TABLE 1.

TABLE I. SAMPLE LABS IN THE PROTOTYPE OF STEM-BD

Lab on Clustering Analysis
Learning Objectives • Understand the concepts of clustering analysis • (Optional)learn clustering algorithms: K-Means clustering • Understand the basic concept of network instrusion • Apply clustering algorithms to network intrusion data • Understand distributed platforms for big data analytics
Algorithm • Clustering Algorithms: K-Means clustering
Data • Network intrusion data
Big Data Platform • Hadoop (with Apache Mahout) • Spark (with MLlib) • HPCC
Lab on Classifying Higgs Boson Competition Data
Learning Objectives • Understand the concept of classification • Understand distributed platforms for big data analytics • (Optional) Learn Logistic Regression based classification Algorithm • Apply Logistic Regression based Classifier to Higgs Boson Data • Understand how to evaluate the results of classfication
Algorithm • Logistic Regression based classification algorithm
Data • Higgs Boson Competition Data
Big Data Platform • Apache Spark (with MLlib) • HPCC
Lab on Recommender System
Learning Objectives • Understand the concept of collaborative filtering • (Optional) Learn the Alternating Least Squares (ALS) recommender algorithm • Apply ALS algorithm for movie and music recommendation • Understand how to evaluate the results of recommendation • Understand distributed platforms for big data analytics
Algorithm • ALS Recommender Algorithm
Data • Audioscrobbler Data Set [17] • MovieLens Dataset
Big Data Platform • Apache Spark (with MLlib)
Lab on Topic Modeling
Learning Objectives • Understand the concept of topic modeling • (Optional) Learn the Latent Dirichlet Allocation (LDA) algorithm • Apply LDA algorithm to topic extraction on text data • Understand how to evaluate the results of topic modeling • Understand distributed platforms for big data analytics
Algorithm • LDA algorithm for topic modeling
Data • The REUTERS-21578 TEXT DATASET • 20 News Groups Dataset
Big Data Platform • Apache Spark (with MLlib)

After gaining knowledge on different analytics tasks, students can experiment with more comprehensive explorative data analysis that involves multiple tasks in a pipeline. A sample explorative lab project of visualizing big high dimensional data can be shortly described as follows. For the network intrusion data, students can first use k-means clustering algorithms to generate different clustering results with different K values. Then design a strategy to detect an optimal K value, such as calculating average entropy for each K value if utilizing the class labels available for this data set, or detecting a sharp change in average distance to cluster center in a sequence of continues K values if not considering class labels. Once an optimal clustering result is produced, students can further apply dimension reduction technique such as PCA, SVD, Stacked Auto Encoder or Deep Belief Network to reduce the dimension of the data to three. Afterwards, apply K-means algorithm again on the data after dimension reduction to generate new clusters. Compare clustering results generated before and after dimension reduction as a means to assess the quality of the dimension reduction. If the quality is satisfactory, students can visualize the data in a 3 dimensional space to gain an intuitive feeling on the distribution of the data and the distribution of different intrusion types

V. Preliminary Evaluation of STEM-BD Project

The prototype of STEM-BD has been released for evaluation by undergraduate CS majors selected from different CS classes at the authors' institute. This version of STEM-BD contains the VM images of the three big data computing platforms that are configured with open-source machine learning packages. STEM-BD also associates one or more suitable analytics algorithms with each of the selected data sets and provides sample codes that apply selected analytics algorithms on that data set. Furthermore, lab manuals for the lab projects described in Table 1 have be developed for this version of STEM-BD. Each lab manual includes pre-lab learning materials, hands-on activities, and post-lab reviews and quizzes.

The prototype of the STEM-BD was been released for preliminary evaluation by undergraduate CS majors selected from different CS classes in the authors' institute. All students participated in the preliminary evaluation have not taken any data mining, machine learning, or big data courses. Participating students were divided into 4 groups, Students in each group were asked to work on one of the lab project described in Table 1 independently by following the corresponding lab manuals. After self-studying the pre-lab learning materials on the concept of big data analytics related to the lab project, students went through hands-on activities step by step to complete the required data analysis and evaluation, and finally conduct the post lab reviews and quizzes. Preliminary evaluation of STEM-BD was then conducted via a survey with the following questions

- The pre-lab learning material of STEM-BD is clear for me to understand the concept of big data analytics with respect to the required lab project.
- The lab manual for the hands-on activities is easy to follow.
- The hands-on activities reinforce my learning on the pre-lab learning material.
- The STEM-BD helps me to learn big data analytics quickly.
- Big data analytics is fun to me.

Each of the above questions was ranked using a scale of 5 to 1, with 5 being strongly agree and 1 being strongly disagree. All questions received either Strongly Agree or Agree responses. All of the students enjoyed the hands-on labs and felt that STEM-BD helps them to quickly understand big data analytics. We plan to further expand this preliminary evaluation to a wide range of students in other STEM disciplines.

Furthermore, we plan to offer a Big Data Analytics course that is suitable for all majors in STEM disciplines. STEM-BD will be the major learning and lab materials of this course. In this course, students first work on samples labs on single data analytics tasks to learn basic concepts of big data analytics; then work on sample explorative labs that use two or more components from one or multiple dimensions of STEM-BD; finally students will be required to customize their own explorative labs according to their disciplines for their course project.

VI. Conclusion and Future Work

This paper presents our ongoing work on developing an innovative labware called STEM-BD to facilitate students' hands-on learning of big data analysis in STEM education. STEM-BD is composed of four components that are essential for learning big data analytical skills, including big data computing platforms, big data, big data analytics algorithm, and hands-on labs. A prototype of STEM-BD has been implemented for a preliminary evaluation and a positive feedback was obtained from students. We are currently working towards adding to the next version of STEM-BIG a lightweight software component that automatically converts users' configuration of an analytic pipeline to source code that runs on the corresponding big data platform. We also plan to incorporate more types of data into the next version STEM-BIG, such as sequence data, time series data and image data. The next version of STEM-BIG will be open to public for evaluation and usages in December 2016. The ultimate goal of STEM-BIG is that students and professionals in STEM disciplines can easily conduct big data analytics without the need to learn much prerequisite knowledge in computing.

References

[1] Hshinchun Chen, Roger H. L. Chiang & Veda C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, vol 36(4), pp. 1165-1188, Dec. 2012

[2] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. "Big data, analytics and the path from insights to value."MIT sloan management review 21 (2013).

[3] Wei Fan and Albert Bifet. "Mining big data: current status, and forecast to the future." ACM SIGKDD Explorations Newsletter 14.2 (2013): 1-5.

[4] D. Kafura, E. Tilevich, and C. Shaffer, "TUES: EAGER: Scaffolding Big Data for Authentic Learning of Computing," Award Number 1444094, Division Of Undergraduate Education, National Science Foundation, 2014.

[5] S.L. Pallickara, "Syllabus of CS535 Big Data," 2015, https://www.cs.colostate.edu/~cs535/Syllabus.html, retrieved: November 10, 2015.

[6] L. Khan, "Course Syllabus of CS 6301 BIG DATA ANALYTICS/MANAGEMENT," 2013, http://dox.utdallas.edu/syl33611, retrieved: November 10, 2015.

[7] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy Mc-Cauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, NSDI'12, pages 2–12

[8] Anthony M Middleton, HPCC Systems : Data Intensive Supercomputing Solutions, white paper, LexisNexis Risk Solutions, 2011.

http://cdn.hpccsystems.com/whitepapers/wp_data_intensive_computing_solutions.pdf

[9] Yanjun Qi, Ziv Bar-Joseph, & Judith Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction http://www3.interscience.wiley.com/cgi-bin/fulltext/112392432/HTMLSTART

[10] Votavova H, Dostalova Merkerova M, Fejglova K, Vasikova A et al.Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. Placenta 2011 Oct;32(10):763-70. http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3929

[11] https://www.kaggle.com/c/higgs-boson/data

[12] http://www.iiap.res.in/astrostat/School07/datasets/SDSS_quasar.html

[13] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[14] https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis#

[15] Fabian Paschke, Christian Bayer, Martyna Bator, Uwe Mönks, Alexander Dicks, Olaf Enge-Rosenblatt, Volker Lohweg, Sensorlose Zustandsüberwachung an Synchronmotoren. In: 23. Workshop Computational Intelligence, 05.-06.12.2013, Dortmund VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA), Düsseldorf Dec 2013

[16] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html

[17] http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html

[18] http://grouplens.org/datasets/movielens/

[19] http://mahout.apache.org/

[20] http://spark.apache.org/docs/latest/mllib-guide.html

[21] https://hpccsystems.com/download/free-modules/ecl-ml