ModStore: An Instructional HPC-based Platform for National Science Experiment Big Data Challenge

Weiwen Zhang, Yong Liu, Long Wang, Jiuyu Zhou, Jiawei Du and Rick Siow Mong Goh
Institute of High Performance Computing, Singapore
Email: {zhangww, liuyong, wangl, zhoujiy, dujw, gohsm}@ihpc.a-star.edu.sg

***Abstract*—In this paper, we present an instructional HPC-based platform, ModStore, for Singapore students to learn data science through National Science Experiment (NSE) big data challenge. Before the challenge, Singapore students, ranging from secondary school, high school to junior college, polytechnics and ITE, carry specially designed sensors to collect data on their daily travel and the surrounding environment during the experiment. However, most of the students do not have any programming background to do data science for this big data challenge. To address this issue, we develop ModStore, powered by National Supercomputing Centre, Singapore, which serves as a powerful tool to enable students to learn about the Internet of Things and Big Data in a drag-and-drop manner by building the workflow, analyzing the information, creating graphs, interpreting visualizations, and comparing trends. 180 students were using ModStore to participate this challenge from October to December, 2016. It has been touted as a promising instructional platform for learning Mathematics, statistics and data science.**

***Index Terms*—ModStore, Big data, Data science, Supercomputing, National Science Experiment**

I. Introduction

National Science Experiment (NSE), organized by the National Research Foundation (NRF) Singapore and Ministry of Education (MOE), is a nation-wide science experiment carried out by Singapore students, ranging from secondary school, high school to junior college, polytechnics and ITE in 2015 and 2016 [1]. Its aim is to track the students' carbon footprint, travel mobility patterns, and the environmental data. Special sensors are designed to collect those data during their daily life. The collected data are transferred wirelessly to a central database for students to obtain the data and do further analysis [2], thus meeting one of the smart nation programme's objectives [3], i.e.,"collect and comprehend" data.

The NSE big data challenge, held in 2016, is a competetion among students to find out meaningful results by analzing the data. Through this challenge, students are supposed to learn about the Internet of Things [4] and Big Data [5]. However, most of the students do not have any programming background; rather, they would rely on scientific calculators or Microsoft Excel. These simple tools are unable to handle a large amount of data or flexible enough to cater for complicated modeling and visualization, which restricts them to interact with the data to solve real-world problems. Even though some students can have some knowledge and make use of relational databases (e.g., MySQL) or analytic tools (e.g., Tableau Desktop), the dataset is huge to be analyzed on their single computers. In addition, DataCamp [6] and Microsoft Azure Machine Learning [7] are too complicated for students to learn the fundamentals of data literacy. Therefore, there still lacks of an interactive tool for students to support data science learning.

We develop ModStore as an instructional platform to enable students to do data science for the NSE big data challenge [8]. It is customized to incorporate relevant MOE math syllabus. Given a set of statistical tools, students can formulate a workflow in a browser in a drag-and-drop manner with appropriate configuration. The complexity of data science is hidden by the well-designed modules in ModStore. Students will only need to focus on solving their real-world problems for meaningful findings by running the workflow, e.g., the relationship between the home-school distance and the transport mode of students going to school. With ModStore, students can read and analyze the information, create graphs, interpret visualizations, and compare trends.

ModStore is powered and supported by HPC backend, which is scalable to support hundreds of students doing data science. It is orchestrated by a web server in the front end and hundreds of compute nodes from National Supercomputing Center (NSCC) [9]. Upon receiving computing job requests from students, the web server will send them to a job dispatcher. The jobs are then distributed to the compute nodes in a scalable manner, which can process a large amount of job requests simultaneously to achieve low latency using high performance computing infrastructure.

ModStore had been used as the platform for the NSE big data challenge that was held from October to December, 2016. With ModStore, students can understand abstract math syllabus much better with real life examples. In addition, they can learn and practice data science concept and skills. More importantly, it can stimulate their great interests in data science.

The rest of the paper is organized as follows. In Section II, we present the background of NSE. In Section III, we present an overview of ModStore. Section IV provides the functional architecture of ModStore for NSE big data challenge. Performance verification is given in Section V. Section VI concludes the paper and suggests future work.

II. Background of National Science Experiment

NSE was conducted from September to November, 2015 and from April to August, 2016. Over 43,000 students from

978-1-5386-1075-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCCRI.2017.11

128 schools were engaged in the experiment. During the experiment, SENSg was the IoT device to measure and store environmental data, including temperature, humidity, atmospheric pressure, light intensity and sound pressure levels, which are correlated to the sensor's location [10]. In addition, transport mode and indoor/outdoor time were inferred from the raw data [11], [12]. Moreover, CO2, travel distance, elevation are also recorded into the dataset. In total, over 700 million lines of data were received during the experiment in 2015 and 2016. More details about NSE can be found in [1], [13].

A big data challenge was held from October to December, 2016, for students to find out meaningful results by analzing the data collected during the NSE. We want to teach students data science by means of ModStore as the tool for the challenge. Students are formed as a team for the competition. At the end of the challenge, all the teams are required to write a report for the findings. They can make good use of ModStore for the challenge, which is the focus of this paper.

III. Overview of ModStore

In this section, we present an overview of ModStore for the NSE big data challenge.

A. Browser-based System

ModStore allows students to start data science project with just a browser to explore, visualize and find insight from data, by constructing, configuring and executing a workflow. Figure 1 shows the web interface of ModStore. It consists of three panels, i.e., dataset & analytic tools panel, workflow panel, and configuration & result panel:

- For the dataset & analytic tools panel, there are data section and function section. Students can get access of the data from the data section. The function section provides a set of statistical tools and graphs: polyfit, histogram, boxplot, stem and leaf diagram, etc. It enables students to practise what they have learnt in classroom.
- From the workflow panel, students can construct a workflow by connecting the dataset to certain functions. It is easy to use due to the drag-and-drop style for quick data science model building and data visualization. The students can save the workflow as a draft. After configuring the workflow, the students can click to run the workflow.
- From the configuration & result panel, students can get the information of the functions on how to use them and configure the functions accordingly before running the workflow. After the workflow execution is completed, students can download the results and put them into their final reports for the NSE big data challenge.

B. HPC-based Infrastructure

ModStore is powered by NSCC. NSCC is a national Petascale computing facility for high performance computing in Singapore.

Specifically, ModStore is orchestrated by a web server at the front end and hundreds of compute nodes from NSCC at the back end. Whenever a student designs and runs a workflow

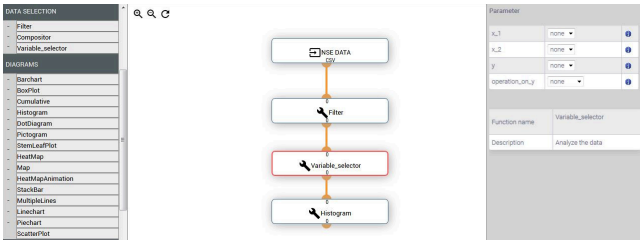


Fig. 1. Interface of ModStore for NSE big data challenge. It consists of three panels, i.e., dataset & analytic tools panel (left), workflow panel (middle) and configuration & result panel (right).

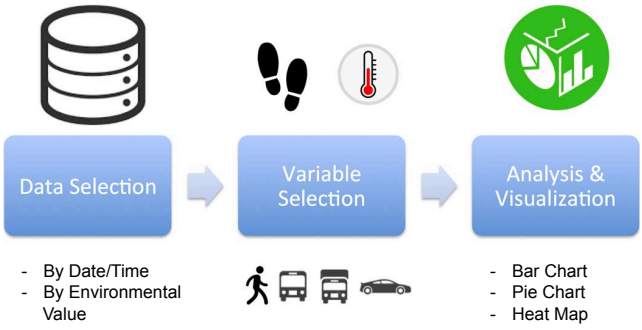


Fig. 2. Modules of ModStore for NSE big data challenge. It consists of three modules, i.e., data selection, variable selection, and analysis & visualization.

through the web interface at the front end, a job will be submitted to one of the compute nodes at the back end.

- The web server is equipped with 16 cores and 64 GB memory to serve the requests from students. Upon receiving requests of workflow execution from students, the web server will send them to a job dispatcher powered by PBS Pro [14]. The job submissions will then be distributed to the compute nodes in a scalable manner.
- Each compute node is equipped with 24 cores and 128 GB memory to support the execution of workflows.

Using high performance computing infrastructure, ModStore can process a large amount of job submissions simultaneously to achieve low latency.

IV. System Functional Architecture

In this section, we present the system functional architecture of ModStore for NSE big data challenge.

As shown in Figure 2, it consists of three modules, i.e., data selection, variable selection, and analysis & visualization. Essentially, the data selection module allows students to filter the data, e.g., by time, date and environment. The variable selection module allows students to select the variables for analysis, e.g., transport mode. The analysis and visualization module allows students to select appropriate functions to analyze and visualize the data, e.g., bar chart, pie chart and heat map.

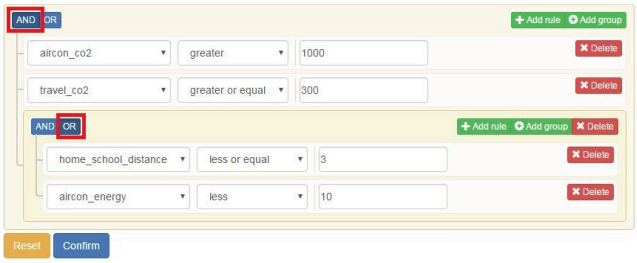
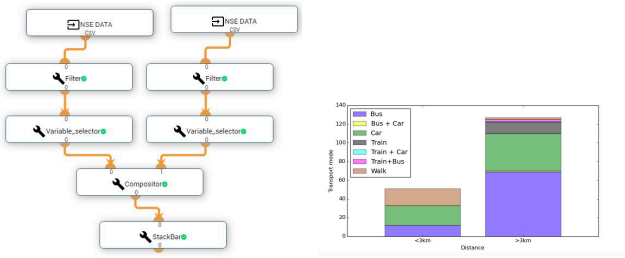


Fig. 3. Module of data filter. Students can set the rules to filter the original data.



(a) Compositor (b) Stackbar

Fig. 4. An example of compositor. The compositor can combine two datasets for comparison.

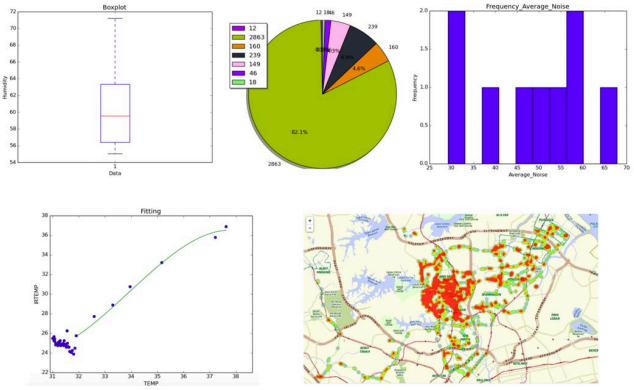


Fig. 5. Some visualization tools for NSE big data challenge: box plot, pie chart, bar chart, polyfit and heat map.

A. Data Selection Module

The data selection module is enabled by the operator "Filter". The filter is based on QueryBuilder [15], which is a web UI component to filter the given data by boolean logic. Figure 3 illustrates an example of data filter on how to perform the filtering. Students can simply click "Add rule" to add more filter rules. Each rule consists of a data field that is dynamically loaded from the connected dataset, an operator (e.g., greater than, equal to, less than), and a value field (i.e., a number for the threshold). All the rules in the same group are based on AND logic by default. Students can change the group logic by clicking "OR". For example, the filtering condition on the original dataset in Figure 3 is

$$aircon\ co2 \geq 1000\ \&\ travel\ co2 \geq 300\ \&\ (home\ school\ distance \leq 3\ ||\ aircon\ energy < 10).$$

With the operator filter, students can easily design the filtering condition to filter a subset of data from the original dataset.

B. Variable Selection Module

The variable selection module is enabled by the operator "Variable selector". The variable selector takes the filtered data as the input. Over the input, it enables students to specify x column (e.g., date) for grouping data and y column (e.g., temperature) as the target for further analysis and visualization. In addition, students can specify the aggregated operation on the y column. The aggregated operations include sum, average, count, standard deviation, maximum and minimum. If no aggregated operation is selected, the plain y will be plotted with the x. With the feature of variable selector, students do not need to write codes to analyze the filtered data by themselves.

In addition, ModStore also provides an operator "Compositor". The compositor can combine two subsets of data. Figure 4 demonstrates the relationship between home-school distance and transport mode using the compositor and the stackbar. The compositor connects with two variable selectors. It is also connected with a visualizer, with its output as the input of the stackbar for the visualization. In this example, there are two subsets of data, one with am travel distance less than 3 km on the left branch of the workflow and the other with am travel distance greater than 3 km on the right branch of the workflow (Figure 4(a)). In this case, the compositor combines two data on the two branches for comparison. Figure 4(b) indicates the result of the workflow, i.e., the number of different transport modes (i.e., bus, bus + car, car, train, train + car, train + bus, walk) for the travel distances less and greater than 3 km, respectively. As a result, the compositor can bring more freedom and flexibility for students to construct the workflows for the data analytics.

C. Analysis and Visualization Module

The analysis and visualization module is enabled by a set of statistical operators and visualizers. It supports a variety of functions that are in accordance with math syllabus under O-level and Pre-U H1 & H2 in Singapore. They are classified into three categories:

- Diagrams: bar chart, stacked bar, pie chart, line chart, scatter plot, box plot, stem leaf plot, dot diagram, pictogram, histogram, map, heat map and cumulative frequency;
- Descriptive statistics: quantile, mean, median, mode, standard deviation and variance;
- Inferential statistics: T-test, Z-test, polyfit and correlation.

Figure 5 shows some of the graphs generated by the functions in ModStore.

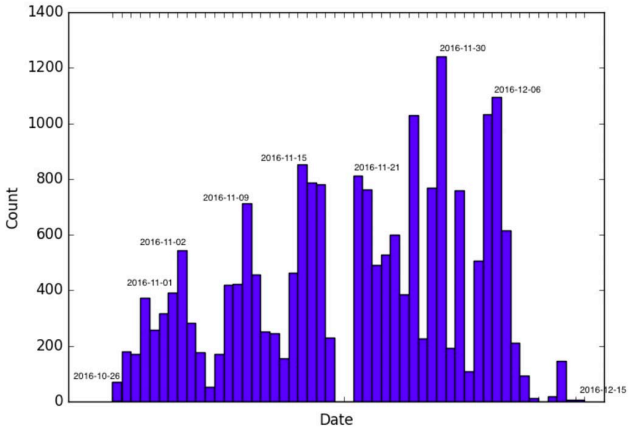


Fig. 6. Job submission during the NSE big data challenge from October to December, 2016.

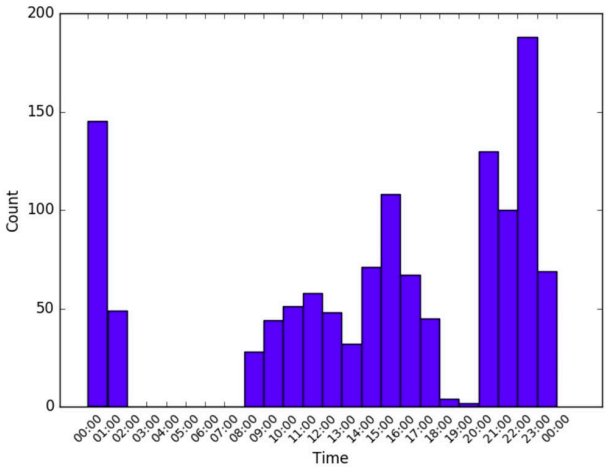


Fig. 7. Job submission throughout a day on November 30, 2016.

V. Performance Verification

In this section, we present performance verification for the platform. We collect the number of job submissions and also students' feedback on user experience after the NSE big data challenge.

A. Job Submission Statistics

We keep track of logs of job submissions to the NSCC during the NSE big data challenge. Overall, there were 20416 job submissions during 51 days and 400.3 job submissions on average for each day.

Figure 6 shows the number of job submissions from October to December, 2016. We observe that there were more job submissions roughly on November 2, 9, 15, 22, 30 and December 6, which fall on either Tuesday or Wednesday. Particularly, on November 30, there were more than 1200 jobs. However, there were no job submissions on November 19 and 20 during the weekend. This is probably because there was a server breakdown such that no job submissions were captured. In addition, there were not many job submissions during the last week before the deadline of the NSE big data challenge. This can be probably because students had already obtained results from ModStore and focused on writing the final report as the entry for the challenge.

Figure 7 shows the number of job submissions at different time intervals throughout a day on November 30, 2016. We observe that on that day students started to submit jobs at 8 am and worked till 6 pm. After the dinner time (from 6 pm to 8 pm), they continued to submit jobs till the late night. Particularly, from 10 pm to 12 am, it had the largest number of job submissions. Generally speaking, ModStore was experiencing a high volume of requests from students.

B. Students' Experience Feedback

We conducted a survey on the students' experience towards ModStore after the challenge. The number of survey respondents is 70. All of the survey respondents are participants

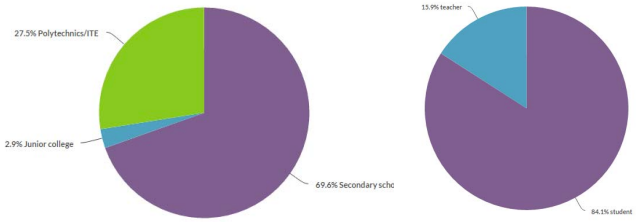


Fig. 8. Respondents for the survey after NSE big data challenge.

of NSE big data challenge 2016. As shown in Figure 8, the majority of respondents are students (84.1%) and are from secondary school (69.6%).

Respondents are satisfied with ModStore overall. Most of them like the platform due to many useful built-in graphs and functions on ModStore. Specifically, those graphs and functions enable them to analyze and visualize the data without programming background. The availability of NSE data on ModStore is another reason for the satisfaction. The respondents also acknowledged that they can leverage the super computers for the big data challenge without installing software and packages on their own computers.

In addition, a large percentage of students agree that they have benefited from using ModStore. As shown in Figure 9, over 70% respondents agree or strongly agree that they have greater interest in data science after using ModStore for the big data challenge and will continue using ModStore. In addition, about 60% respondents agree or strongly agree that they have better understanding of mathematical concepts and computational skills. Therefore, it indicates that ModStore is a promising instructional data science platform.

Moreover, respondents have some suggestions on how to improve ModStore. First, one common feedback from the respondents is that execution time for running workflows can be long. This is primarily due to the large filtered dataset after

	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree	
I have greater interest in data science	0 0.0%	2 3.6%	12 21.8%	28 50.9%	13 23.6%	74.5%
I have better understanding of mathematical concepts	1 1.8%	5 9.1%	16 29.1%	24 43.6%	9 16.4%	60.0%
I have better computational skills	2 3.6%	8 14.5%	13 23.6%	24 43.6%	8 14.5%	58.1%
I would like to continue using ModStore to practice data science concepts and skills to solve real-world problems (assumed that ModStore has made the improvements based on my suggestions)	1 1.8%	7 12.7%	7 12.7%	28 50.9%	12 21.8%	72.7%

Fig. 9. Interest and understanding in data science after using ModStore.

the students set the filtering rule on the original dataset. The estimated running time should be given to students based on the size of dataset and the functions they have chosen. Second, more error messages and guidelines should be given such that students can pinpoint errors for easier troubleshooting when running the workflows. Finally, some respondents would like to define their own functions such that they can do a little bit of programming on the dataset while enjoying the statistical tools and visualizers. By adopting these suggestions, we believe that the user experience of ModStore can be significantly improved.

VI. Conclusion

In this paper, we presented ModStore powered by HPC infrastructure for the NSE big data challenge, catering to students with diverse level of computer and data science competency. Because of the drag-and-drop functionality and a wide range of statistical and visualization tools in ModStore, students with no programming background can easily explore and visualize big data, identify patterns of students' behavior, analyze environmental conditions and obtain interesting insights. Students' feedback from the survey showed that over 70% respondents agree or strongly agree that they have greater interest in data science after using ModStore for the challenge and would like to continue using ModStore, given that improvements have been made. Therefore, ModStore has the potential to be an instructional data science platform to facilitate the learning of data science and to enhance the competency of students in big data and data science.

In the future, we will enhance the functionality of ModStore, by capturing users' behavior. We will keep track of logs of students' behavior on how they construct a workflow and set the configuration, in order to improve the learning experience.

Acknowledgment

The NSE experiment is organised by the National Research Foundation (NRF) Singapore and Ministry of Education (MOE), in partnership with Singapore University of Technology and Design (SUTD), Science Centre Singapore (SCS), Agency for Science, Technology and Research (A*STAR), Singapore Land Authority (SLA) and OneMap, to excite young Singaporeans in science and technology, and to celebrate 50 years of science and technology in Singapore, as part of our Jubilee celebrations. ModStore as the platform for NSE big data challenge is also supported by National Supercomputing Centre (NSCC) for the large-scale computation.

References

[1] "National science experiment (nse)." https://www.nse.sg/.
[2] E. Wilhelm, S. Siby, Y. Zhou, X. J. S. Ashok, M. Jayasuriya, S. Foong, J. Kee, K. L. Wood, and N. O. Tippenhauer, "Wearable environmental sensors and infrastructure for mobile large-scale urban deployment," *IEEE Sensors Journal*, vol. 16, no. 22, pp. 8111–8123, 2016.
[3] "Smart nation." https://www.smartnation.sg/.
[4] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of things," *International Journal of Communication Systems*, vol. 25, no. 9, p. 1101, 2012.
[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.
[6] "Datacamp." https://www.datacamp.com/.
[7] "Microsoft azure machine learning." https://azure.microsoft.com/en-us/services/machine-learning/.
[8] "Modstore." http://modstore.org/.
[9] "National supercomputing centre singapore (nscc)." https://help.nscc.sg/.
[10] N. Z. J. K. G. L. N. T. Erik Wilhelm, Yuren Zhou, "SENSg: Large-scale deployment of wearable sensors for trip and transport mode logging," in *Proceedings of Annual Meeting of the Transportation Research Board (TRB)*, 2016.
[11] E. Wilhelm, D. MacKenzie, Y. Zhou, L. Cheah, and N. Tippenhauer, "Evaluation of transport mode using wearable sensor data from thousands of students," in *Proceedings of Annual Meeting of the Transportation Research Board (TRB)*, 2017.
[12] B. Monnot, E. Wilhelm, G. Piliouras, Y. Zhou, D. Dahlmeier, H. Y. Lu, and W. Jin, "Inferring activities and optimal trips: Lessons from singapores national science experiment," in *Complex Systems Design & Management Asia*, pp. 247–264, Springer, 2016.
[13] "Press release." https://www.nse.sg/wp-content/uploads/2015/05/20151210_NSE2015-Results-Press-Release-Final.pdf.
[14] "Pbs pro." http://www.pbsworks.com/.
[15] "Querybuilder." http://querybuilder.js.org/.