A collaborative, multinational cyberinfrastructure for big data analytics

Prof. Raymond A Hansen, Purdue University
Dr. Tomasz Wiktor Wlodarczyk, University of Stavanger

Dr Tomasz Wiktor Wlodarczyk, is an Associate Professor at the Department of Electrical and Computer Engineering at University of Stavanger, Norway. His work focuses on analysis, storage and communication in data intensive computing. His particular interest is time series storage and analysis. He is currently working on these areas in several research projects including: SEEDS (EU FP7), Safer@Home (RCN), A4Cloud (EU FP7), BigDataCom-PU-UiS (SIU), SCC-Computing (EU FP7). He has also been the Program Committee Chair of IEEE CloudCom – International Conference on Cloud Computing Technology and Science for 2011 and 2012.

Prof. Thomas J. Hacker, Purdue University, West Lafayette

Thomas J. Hacker is an Associate Professor of Computer and Information Technology at Purdue University in West Lafayette, Indiana. His research interests include cyberinfrastructure systems, high performance computing, and the reliability of large-scale supercomputing systems. He holds a PhD in Computer Science and Engineering from the University of Michigan, Ann Arbor. He is a member of IEEE, the ACM, and ASEE.

A collaborative, multinational cyberinfrastructure for big data analytics

Introduction

The emergence of Big Data and Data Intensive Systems as specialized fields within computing has seen the creation and delivery of curricula to provide education in the techniques and technologies needed to distill knowledge from datasets where traditional methods, like relational databases, do not suffice. Within the current literature and these new curricula, there is a seeming lack of a thorough and coherent method for teaching Data Intensive Systems so that students understand the theory and the practice of these systems, allowing them to be effective in the laboratory and, ultimately, as data analysts/scientists [1][2][3][4]. One paradigm that has been widely adopted in industry is MapReduce as implemented in open-source tool, Hadoop [3]. Although these systems are based on many years of research work, the conceptual framework on which these systems were built differs largely from what could be found in the earlier research work and education curricula. University courses available today are largely focused around various areas (some from repackaged content) that cover some selected parts of Big Data spectrum, mostly: data mining, distributed systems and most recently data science [5][6][7].

We believe that the pedagogical approach used by related education programs today lacks focused intended learning outcomes built on the use of current technology and is not coherently mapped into teaching/learning activities and assessment tasks. Perhaps one of the biggest challenges for creating Big Data and Data Intensive Systems curricula is to define coherent and stable learning objectives in a highly dynamic field. One of the reasons that courses offered at different institutions are not clear in this regard is because they are anchored deeply in the detailed research areas of lecturers, as opposed to industry needs. While this may not be bad in principle for an *advanced* course, a significant shared curriculum is necessary to facilitate knowledge transfer and increase quality of education for an introductory course.

Current training from organizations (e.g. Cloudera) and from textbooks (e.g. Hadoop in Action [8] or Mining of Massive Datasets [9]) has been built around technical referencing or singular engineering problems, and does not offer a firm theoretical basis to guide students or advanced practitioners in their exploration of the field. Some good materials are available, but they are spread throughout various university or professional courses. As a result, current curricula and industrial training programs suffer from a fragmentation of knowledge and the lack of a strong link between theory and practice. Additionally, many of these focus on relatively few of the concepts that are essential to data analytics and data intensive systems. To resolve this perceived gap we specifically designed our course to provide the necessary theoretical framework and then bridge the gap to application. The following is a cursory glance at existing texts and courses that cover varying aspects of big data analytics:

Related Works

A. Books and Textbooks

We divided the readily available books into five categories: Big Data, MapReduce, NOSQL, Hadoop and Data Science. We include a short overview of each book in this order. While this list is not to be considered exhaustive, it does cover a significant percentage of widely available books. Also, many of these books could be applied to several categories at the same time. However, we chose only a single category for each book in order to provide basic classification, but different categorizations could also be argued.

Big Data Books

The text *Big Data* [10] is a work-in-progress that focuses on real-time Big Data systems. It present a general architecture for hybrid approaches based on real-life applications.

The text *Mining of Massive Datasets* [9] is used for the CS345A course at Stanford University. It offers a more theoretical background than other available books. It focuses on a set of algorithms for a few key problems in data mining e.g. link analysis or clustering.

Understanding Big Data [11] provides a general overview of the Big Data landscape from IBM's perspective. This bias is noted throughout the book with noticeable influence in the content.

Big Data Glossary [12], as the title suggests, provides a short overview of Big Data and machine learning terminology without particular applicability for education or classroom/laboratory environments.

The *Little Book of DATA SCIENCE* [13] and its re-release as *A Simple Introduction to DATA SCIENCE* [14] provides basic information on Big Data, Hadoop, and an overview of Cassandra with Data Science applications. It has a noticeable academic focus, however, as its title suggests it is a primer to aid further exploration.

MapReduce Books

Data-Intensive Text Processing with MapReduce [15] addresses different MapReduce algorithm design techniques with a narrow focus on language processing.

MapReduce Design Patterns [16] is an advanced topics book that is focused on MapReduce patterns. The text is a very useful source for users and students who are already familiar with basic MapReduce concepts.

NOSQL books

Mahout in Action [17] describes a framework for machine learning implemented using Hadoop. It is focused on the technical details of different algorithms and methods.

Cassandra: The Definitive Guide [18] and *Cassandra High Performance Cookbook* [19] describe the Cassandra database management system. The first book gives a detailed overview of Cassandra, and the second text provides practical solutions to common tasks and problems.

NoSQL Distilled [20] is primarily a concept book that (in our opinion) may be too general and, therefore, limited for educational purposes.

MongoDB: The Definitive Guide [21] and *MongoDB in Action* [22] describe a document-oriented database called MongoDB. The first book provides a detailed overview of MongoDB, and the second text provides a practical user guide with little perspective towards its application towards big data analytics.

Hadoop Books

Hadoop: The Definitive Guide 3rd edition [23] is probably the most well-known and most complete reference book for Hadoop, but it might be difficult to follow if it is used as the introductory book for students on the subject.

Hadoop in Action [8] covers the basic technical aspects of using Hadoop. The text focuses on key functionalities, while not seeking to cover the entirety of Hadoop. In our opinion, the text could be stronger in the theoretical aspects of data analytics, but provides a sufficient introduction to Hadoop.

Hadoop in Practice [24] focuses on practical techniques applied in Hadoop, and is a good reference book for practical implementations. The text has high-quality diagrams that provide clarity to help with understanding the content.

Hadoop Operations [25] is well-suited for use of Hadoop in practice and it contains up-to-date information. It is recommended for the operational aspects of Hadoop cluster management.

Hadoop Essentials [26] is one of the few books written with purpose of being a textbook. It covers the basics of Hadoop well, but could be stronger on providing a more general context and overview of the area.

Pro Hadoop [27] is a practice based book, shorter than most of the other books. The text covers the basics from a practical point of view.

Data Science Books

Data Intensive Science [28] and *Scientific Data Management: Challenges, Technology, and Deployment* [29] contain a series of chapters by different authors describing "big data" projects.

Our list of Data Science books could have been longer, but we chose to exclude many texts that clearly (both by content and title) had limited relevance.

Summary of Books and Textbooks

After the book review we arrived at the following conclusions:

- There are many high quality technical sources
- There are few sources that could be used directly in the classroom (textbook)
- A few textbooks available are very specialized in particular analytic domains (e.g. web analytics, language processing)
- There are currently no textbooks that offer a full package including related slides, labs and sample exams

B. Courses

The following lists courses taught at either the undergraduate or graduate level, or as continuing education/outreach courses at respected universities.

For the course *Mining Massive Data Sets* [6] at Stanford University refer to the book of the same title in the previous section.

Analytics from Big Data [5] at Stanford University is an advanced first-year MBA course in data-mining, machine learning, and cloud computing. It is focused around Matlab and R. The course does not cover Hadoop or MapReduce. The course mostly focuses on statistics without significant focus on programming or implementation of a big data analytics environment.

Massive Data Analysis [30] at New York University Poly appears to be based primarily on the two books discussed above: *Mining of Massive Data Sets* and *Data-Intensive Text Processing with MapReduce*. The course provides a good coverage of the general topics, but seems to be a collage of related talks from other authors, and its implementation seems to lack a coherent framework. The course is good at addressing applications, with some underlying computer science related to the technologies, but does not address Big Data in Science.

Precision Practice with Big Data [31] at Stanford University is an application survey course. The course does not cover Hadoop or MapReduce. From our assessment from available material, there does not seem to be sufficiently detailed information for applied programming. However, the course has been taught since 2008 and considers information policy issues, so it has had the ability to mature as a course.

Parallel and Distributed Data Management [32] at Stanford University is primarily a database course. A couple of the later lectures in the course deal with topics like MapReduce, but not to a depth that would provide any sufficient level of proficiency in big data analytics.

Analyzing Big Data with Twitter [7] at the University of California - Berkeley focuses on algorithms to mine sentiment analysis and trend detection in social streams. The course seems to be strongly focused around Twitter applications.

Big Data: Making Complex Things Simpler [33] at MIT is a short course aiming to provide a general overview of big data.

Introduction to Data Science [34] at Columbia University provides a general data science overview that includes topics like Hadoop and related programming languages.

Applied Data Science [35] at Columbia University is mostly focused on statistics.

Curriculum and Course Design

The analysis of available books and courses in the sections above provided us with the target and intent for a course that would address both the theory and hands-on application of big data analytics. The theory was introduced primarily through reading assignments and lectures, while the hands-on application was presented to students through a physical infrastructure for projects and research assignments. This paper addresses the cyberinfrastucture for a graduate-level, synchronous, distance course between two universities. *Cyberinfrastructure for Big Data Analytics* (~~X~~ Purdue University) and *Data Intensive Systems* (University of ~~Y~~Stavanger) stress the universal applicability of the covered topics to the big data and data intensive system domains that use data analytics. In addition, since the course is an entry-level graduate course in data intensive topics, it is applicable to significant extent to any program that includes the use of parallel, high-performance or distributed computing. Based on this entry-level expectation of students' skillset, this course was attended by students from Computer & Information Technology, Computer Science, Industrial Engineering, Mechanical Engineering, and Agronomy.

Taking into account the expected knowledge, skills, and abilities that a professional in the field has to have, as well as the tasks and responsibilities they are expected to perform, we defined the following intended learning outcomes:

LO1. design, construct, test, and benchmark a small data processing cluster (based on Hadoop);

LO2. characterize Hadoop job tracker, task tracker, scheduling issues, communications, and resource management;

LO3. describe elements of Hadoop ecosystem and identify their applicability;

LO4. analyze real-life problems and propose suitable solutions;

LO5. describe and compare RDBMS, data warehouse, unstructured big data, and keyed files, and show how to apply them to typical data processing problems;

LO6. construct programs based directly on MapReduce paradigm for typical problems;

LO7. construct programs based on high-level tools (for MapReduce paradigm) for typical problems;

LO8. understand algorithmic complexity of the worst case, expected case, and best case running time, and the orders of complexity; apply the analysis to real life algorithms;

LO9. analyze influence of peak and sustained bandwidth rate on system performance;

LO10. evaluate, communicate and defend a solution w.r.t. relevant criteria.

In order to achieve each of these desired learning outcomes, several different assignments were given in order to accurately assess student performance and competency against the learning outcomes. These assignments were broken down into three categories: presentations, projects, and examinations. For this paper, only the projects portions are of importance. These projects required the cyberinfrastucture that was constructed specifically for this course. Overall, each of the Learning Outcomes were addressed through the assignments, but not all learning outcomes were addressed by the projects. However, it is desired that each learning outcome will have a project-based assessment in the future offerings of this course.

A tutorial of Project 1 was provided to students that detailed the basic installation and configuration of Hadoop within the cyberinfrastucture environment.

Project 2 was assigned to students to have them ingest data into the Hadoop HDFS. Students were able to create their own use case and to choose their own datasets and scenarios to support that use case for this project. As such, there was a significant amount of flexibility for them to determine specific requirements. For example, students needed to determine the amount of data required for their scenario and how many replications of that data they should use to ensure adequate performance. If they chose too few replications, then their MapReduce jobs (Project 3) would require a significant amount of time to complete (especially given the finite resources of the computing environment. Conversely, if they chose too many replications of their data, then they may not have been able to store their full data set, as there was limited storage within the HDFS.

The third project that was assigned (Project 3) was designed for students to use the Hadoop environment that they installed, along with the data that was ingested and replicated as part of Project 2 and then define the necessary mapper and reducer functions in order to understand MapReduce. Additionally, a secondary set of requirements were given to evaluate the use, necessity, and performance impacts of adding combiners to the job. Some students chose to show this impact via `top`, `mpstat`, and `sar`, while other students evaluated the amount of ingress/egress data flowing between nodes and the total amount of time required to complete the jobs.

Additional projects were assigned that required the use of the cyberinfrastucture, but the details of those projects are not pertinent to this paper.

Cyberinfrastucture Design

The foundation of the cyberinfrastructure to be used by the students was fifty-six Dell OptiPlex GX620 PCs. Each PC had an Intel Pentium D dual-core CPU operating at 2.8 GHz with 3-4 GB RAM, and 2x 160 GB hard disc drive, with gigabit Ethernet NICs. This provided the environment for VMWare's ESXi 5 hypervisor to host three virtual machines for each physical PC. Each virtual machine was created from a basic image of a fully patched Fedora 19. This can be seen in Figure 1~~Figure 1~~. We ensured that the current JVM and Perl Data Language, along with their dependencies, were installed prior to the creation of the base image. Hadoop 2.1.0-beta was downloaded from a mirror of the source code available at apache.org but was not installed. This was due to the students installing and configuring Hadoop as part of their first laboratory project.

Formatte

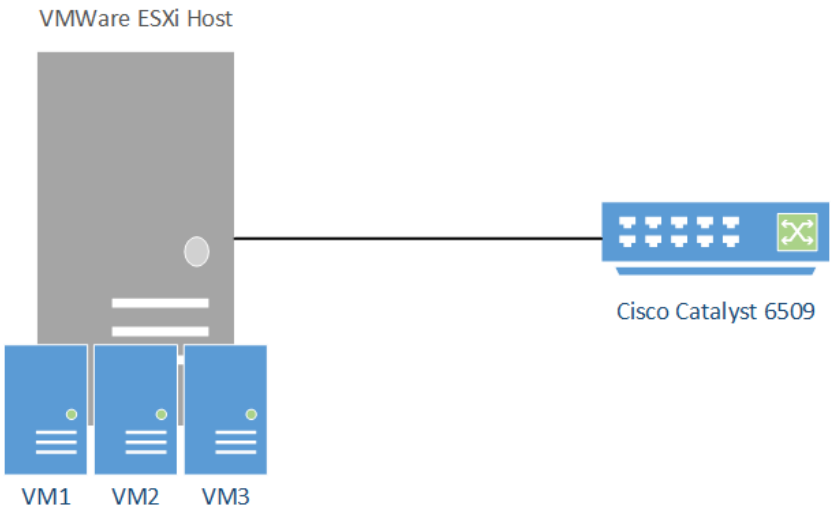

Figure 1 - Generic VMWare ESXi Architecture

Each student was assigned four VMs from the pool of all machines. In order to maximize the performance of each VM group, the VMs were distributed across the physical machines so that the Hadoop Head Node and two Data Nodes all reside on different physical machines. This also reduced the impact that a failure of a single physical machine would have on a student's ability to complete their project assignments. The underlying ESXi host was assigned an IP address in a specific /24 range that would allow remote administration via VMWare vSphere and could have an ACL applied to limit student access to the management segment of the network. Table 1~~Table 1~~ shows a portion of the IP address allocation scheme for the virtualize hosts.

Formatte

Table 1 - IP Address Layout of Virtual Machines

	HeadNode	192.168.65.X	DataNode 1	192.168.65.X	DataNode 2	192.168.65.X
student 1	Head 1.1	10	Data 1	11	Data 1	12

student 2	Head 2.1	15	Data 2	16	Data 2	17
student 3	Head 3.1	20	Data 3	21	Data 3	22
student 4	Head 4.1	25	Data 4	26	Data 4	27

The entire computing environment was interconnected via gigabit Ethernet to a Cisco Catalyst 6509 switch. There were two 48-port line cards (WS-X6148-GE-TX) installed to provide connectivity to the physical hosts as well as provide a 2Gbps aggregate link to the managed data center infrastructure. In order to ensure full line speed capabilities between the nodes (as is desired in a Hadoop installation), knowledge of the backplane allocation of each line card is necessary. Even though these line cards separate ports into combinations of 12 physical interfaces, access to the backplane is divided among eight ports. Specifically, 6 Gbps of throughput to the backplane is allocated to these eight port groupings. This provides six individual groupings of eight ports per line card, for a total of 12 groups across the entire switch, as configured. One of those groups was assigned to the IEEE 802.3ad/LACP group; leaving 11 remaining groups. This allowed for the potential of 66 hosts to be connected, which was more than adequate for the number of physical hosts available. Figure 2~~Figure 2~~ shows a generalized topology of the VMWare ESXi hosts interconnecting to the network. Even though there was sufficient capacity within the switches backplane, and available slots within the chassis, and line cards to provide connections, a single NIC was used in each machine. This was primarily due to the fact that we had seen a very low failure rate in the NICs used in these machines throughout their use in other laboratories and projects. It was determined that the cost-benefit was not significant enough to warrant the efforts to install and configure the additional cards as well as the increase in power budget (as negligible as the increase may be).

Formatte

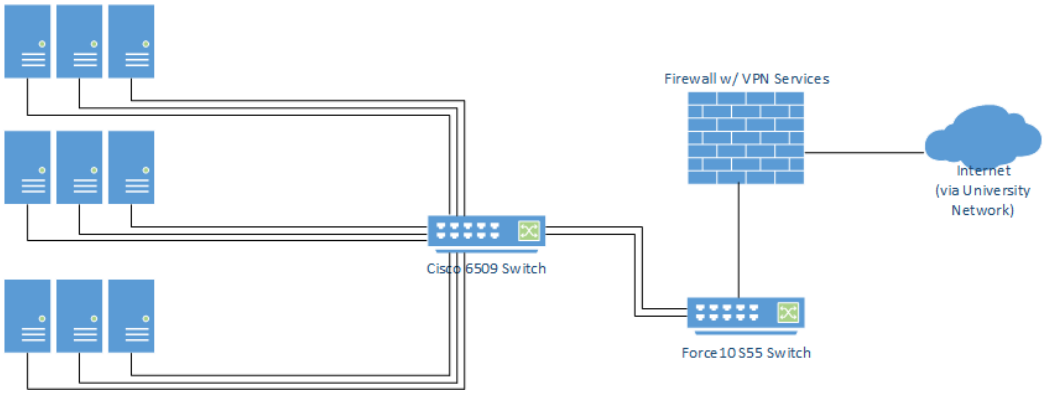

Figure 2 - General Network Topology

In order to provide a certain amount of high availability to this cyberinfrastructure, this environment was constructed in the test data center of the Cyberinfrastructure and High

Performance Computing Lab, which provided multiple 30A circuits and more than adequate cooling for the computing needs. In addition, by placing the course cyberinfrastructure into this data center, it was clear to all users of the data center and laboratory that these machines were serving a specific purpose, and should not be allocated for any other purpose.

It may be noted that no backup service was provided for students by way of NAS or SAN. The onus is place on the student to perform the necessary backup and disaster recovery processes to ensure their data is recoverable. This is done for multiple reasons; the students will not learn to follow the best-practice of off-site/off-system backup and documentation of work if they are never asked/forced to do so. If we provide the service for them within the cyberinfrastructure, then they will lose sight of the necessity and processes required to have an effective recovery mechanism. Additionally, it a cost-related issue within the network. We simply do not have the capital funds, or recurring funds, available to construct a robust storage architecture for all student configuration, workspace, and data storage requirements. As such, we make no attempt to offer this as a service to the students.

Student access to this environment was provided via three distinct manners. The first, and least used by students was to physically connect a monitor and keyboard to the physical PCs within the data center and directly interact with their assigned machines in this way. However, since their virtual machines were placed on multiple physical PCs, this was typically the most error prone approach as it was relatively easy for students to connect to the incorrect PC and then had difficulties arising from that. For students that were local to the campus, the department's wireless network provided authenticated and authorized access into the data center network. Also, there were a specified set of network drops that would provide connectivity to the desired network segment. The third, and most used, manner to access this network was via a virtual private network connection (VPN). Two distinct types of VPNs were created: a client access PPTP VPN and a site-to-site IPSec VPN. The client access VPN was created for X University students to use. The IPSec VPN was created between a pfSense firewall within the data center (Figure 2~~Figure 2~~), and another pfSense firewall that was installed in the laboratory at University of Y. The general architecture of this access method is shown in Figure 3~~Figure 3~~.

Formatte

Formatte

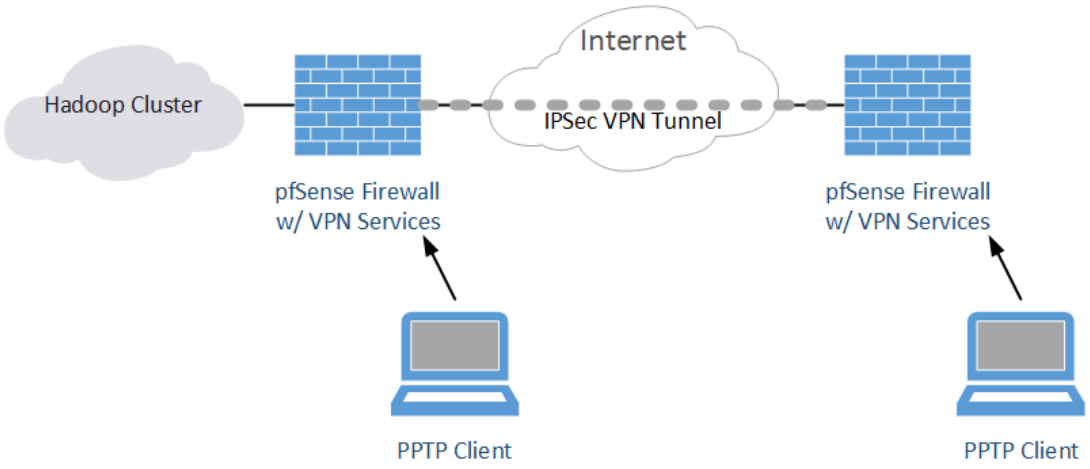

Figure 3 - VPN Architecture

Findings & Challenges

During the construction and maintenance of this cyberinfrastructure and its usage by students as part of these courses, several challenges arose and, as a result, many findings were identified. These can be broken down into three major sections: personnel, hardware and environment, and policy.

As for the students, they were adept at defining a real-world case study problem and subsequently finding datasets to support their projects (ex. U.S. census data, Twitter data, USGS earthquake data, supercomputer log data). Because of these diverse datasets, generalized discussions of theory were able to be focused onto a specific type of problem for students to gain a deeper understanding of the topics at hand. This was true for students at both institutions. In this way, the cyberinfrastructure was valuable and beneficial for student learning.

There were also issues with the student knowledge. Due to the varied background and skillsets of the students accepted into this course, many students had a lack of familiarity with the Linux platform and its CLI. As such, interacting with Hadoop in a meaningful way was a foreign concept to many students, as they didn't understand the environment in which they were being asked to operate. A selection of tutorials, guides, and help sessions allowed most of the students to overcome this deficiency in relatively short order. The students were by no means experts in either Linux or Hadoop. But they could function within that environment to a degree necessary to be successful within the confines of this course.

A major personnel issue occurred at the onset of the semester in that significant faculty time was spent in the installation, maintenance, and troubleshooting of the cyberinfrastructure. The acquisition and provisioning of all the PCs and network hardware were performed by the faculty during the weeks prior to the beginning of the semester, as well as the first few weeks of the semester. Over 100 hours of time was spent by the two faculty members ensuring that this environment (computing, communications, filtering, VPNs, etc.) was operational for students and that the assigned projects were able to be completed with the tools and resources provided. As problems arose, additional time was committed to their resolution. Some of these issues are unavoidable with operating any laboratory environment, while others are a result of the following two findings and challenges.

The computing hardware that was provided to students were PCs that had been decommissioned from other laboratories on campus. Some had been desktops within student "open labs" and others had been faculty workstations that were no longer up to par. In both cases, these machines were on the downward side of their productive lifespan. There were several instances of machines requiring specific attention after a power cycle to bring them back to fully operational state. Also, due to the age of these machines, a substantial amount of RAM could not be installed into them to have them host more virtual machines, thereby giving students the opportunity to include a larger number of mappers, combiners, and reducers in their projects. Adding to this, the hard disc drives were relatively small, by current standards, meaning students could not use some of the larger datasets they found.

Even though the cyberinfrastucture was purposefully placed into a data center with more than sufficient power and cooling capacity, several power and cooling outages caused significant issues throughout the length of the course. Over the course of the semester, power was lost to the data center twice due to building-level upgrades to the power infrastructure. One of these outages was unannounced, and the entire Hadoop cluster was ungracefully shutdown. This resulted in us having to physically connect to ~25% of the virtual machines to restart them gracefully. The utilization of UPS's could have prevented some of these issues, but again, the cost of implementing that amount of battery backup was prohibitive to its implementation.

A discussion could be warranted on the usage of other virtualization platforms (Azure, KVM, "the cloud", etc.). However, at this time, we have only fully implemented this in the VMWare environment, with minimal testing using KVM.

Conclusions

Value of infrastructure impacted student learning for the better through the hands-on projects. The assessed projects completed by the students support the desired learning outcomes for the course. As such, it is highly suggested that a deeper understanding of those topics supporting those learning outcomes were possible because of the hands-on nature of the projects.

The location of that infrastructure is not critically important. Due to the nature of the cyberinfrastucture provided to students in the past offerings of this course, the ability to physically and directly interact with the underlying hardware is not necessary. While verifiable student learning occurred, the obstacles and hurdles that occurred throughout the semester potentially detracted from additional learning. The biggest takeaway from the faculty members' experience is to investigate alternative approaches to deliver this hands-on learning via projects. A cursory analysis suggests that Amazon's AWS/EC2 or even Elastic MapReduce would provide many of the needs for the hands-on component of this course. Perhaps an offering from other cloud service provides could be examined as well.

Additional curricula refinement is currently underway for the next offering of this course between the two universities. Based on the curricular refinements, adjustments to the projects and their objectives may also be implemented in order to continually improve the course, the learning outcomes, and student performance and understanding.

References

1. F. Provost, T. Fawcett (2013). Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media.
2. C. O'Neil, R. Schutt (2013). Doing Data Science: Straight talk from the frontline. Sebastopol, CA: O'Reilly Media.

3. A. Sathi (2013). Big Data Analytics: Disruptive technologies for changing the game. Boise, ID: MCPress Online, LLC.
4. V. Granville (2014). Developing analytic talent: Becoming a data scientist. New York, NY: John Wiley & Sons.
5. "Stanford University Explore Courses." [Online]. Available: http://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&academicYear=&q=OIT+367&collapse=. [Accessed: 09-Aug-2013]
6. "CS9223 - Massive Data Analysis." [Online]. Available: http://vgc.poly.edu/~juliana/courses/cs9223/. [Accessed: 09-Aug-2013]
7. "Info 290. Analyzing Big Data with Twitter | School of Information." [Online]. Available: http://www.ischool.berkeley.edu/courses/i290-abdt. [Accessed: 09-Aug-2013]
8. Lam, *Hadoop in action*. Greenwich, Conn.: Manning Publications, 2011.
9. A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. New York, N.Y.; Cambridge: Cambridge University Press, 2012.
10. N. Marz, *Big data: principles and best practices of scalable realtime data systems.* [S.l.]: O'Reilly Media, 2013.
11. I. P. Zikopoulos, C. Eaton, and P. Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw Hill Professional, 2011.
12. *Big Data Glossary*. [Online]. Available: http://shop.oreilly.com/product/0636920022466.do. [Accessed: 09-Aug-2013]
13. N. Burlingame, *The Little Book of DATA SCIENCE, 2012 Edition*. New Street Communications, LLC, 2012.
14. N. Burlingame and L. Nielsen, *A simple introduction to data science*. Wickford, RI: New Street Communications, 2012.
15. J. Lin, *Data-intensive text processing with MapReduce*. [San Rafael, Calif.]: Morgan & Claypool Publishers, 2010 [Online]. Available: http://www.morganclaypool.com/doi/abs/10.2200/S00274ED1V01Y201006HLT007. [Accessed: 09-Aug-2013]
16. D. Miner and A. Shook, "MapReduce design patterns," 2012. .
17. S. Owen, *Mahout in action*. Shelter Island, N.Y.: Manning Publications Co., 2012.
18. E. Hewitt, *Cassandra: the definitive guide*. Beijing: O'Reilly, 2011.
19. E. Capriolo, *Cassandra high performance cookbook*. Birmingham, UK: Packt Pub., 2011 [Online]. Available: http://proquest.safaribooksonline.com/?fpi=9781849515122. [Accessed: 09-Aug-2013]
20. P. J. Sadalage and M. Fowler, *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*. Upper Saddle River, NJ: Addison-Wesley, 2013.
21. K. Chodorow, *MongoDB: the definitive guide.* Sebastopol: O'Reilly Media, 2013.
22. K. Banker, *MongoDB in action*. Shelter Island, NY: Manning, 2012 [Online]. Available: http://proquest.safaribooksonline.com/?fpi=9781935182870. [Accessed: 09-Aug-2013]
23. T. White, *Hadoop: the definitive guide*. Farnham: O'Reilly, 2012.
24. A. Holmes, "Hadoop in practice," 2012. [Online]. Available: http://proquest.safaribooksonline.com/?fpi=9781617290237. [Accessed: 09-Aug-2013]
25. E. Sammer, "Hadoop operations," 2012. [Online]. Available: http://proquest.safaribooksonline.com/?uiCode=univlaval&xmlId=9781449327279. [Accessed: 09-Aug-2013]
26. H. H. Liu, *Hadoop essentials: a quantitative approach*. [Charleston, S.C.]: PerfMath, 2012.
27. J. Venner, *Pro Hadoop*. Berkeley, CA: Apress, 2009 [Online]. Available: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=478209. [Accessed: 09-Aug-2013]
28. T. Critchlow and K. K. Van Dam, *Data-intensive science*. 2013.
29. A. Shoshani and D. Rotem, *Scientific data management: challenges, technology, and deployment*. Boca Raton: CRC Press, 2010.
30. "Stanford CS246: Mining Massive Data Sets (Winter 2013)." [Online]. Available: http://www.stanford.edu/class/cs246/. [Accessed: 09-Aug-2013]

31. "Stanford University Explore Courses." [Online]. Available: https://explorecourses.stanford.edu/search?view=catalog&filter-coursestatus-Active=on&page=0&catalog=&q=BIOMEDIN+205%3A+Precision+Practice+with+Big+Data&collapse=. [Accessed: 09-Aug-2013]
32. "CS347 Parallel and Distributed Data Management | Stanford University Online." [Online]. Available: http://scpd.stanford.edu/search/publicCourseSearchDetails.do?method=load&courseId=11836. [Accessed: 09-Aug-2013]
33. "Big Data: Making Complex Things Simpler - IT Management Course - MIT Sloan Executive Education." [Online]. Available: http://executive.mit.edu/openenrollment/program/big_data_making_complex_things_simpler/49#/overview. [Accessed: 09-Aug-2013]
34. "Introduction to Data Science," *Introduction to Data Science, Columbia University*. [Online]. Available: http://columbiadatascience.com/about-the-class/. [Accessed: 09-Aug-2013]
35. "Applied Data Science (Statistics W4249)," *Introduction to Data Science, Columbia University*. [Online]. Available: http://columbiadatascience.com/2012/10/19/next-semester-applied-data-science-statistics-w4249/. [Accessed: 09-Aug-2013]