Integrating Big Data Analytics into an Undergraduate Information Systems Program using Hadoop

Justin DeBo
jadebo@millikin.edu

RJ Podeschi
rpodeschi@millikin.edu

Tabor School of Business
Millikin University
Decatur, IL 62522

Abstract

With the emergence of big data as a strategic weapon in business, the need for hands-on activities in undergraduate courses is essential for preparing the next wave of technical talent. As the availability of programs in data analytics and data science grows based on market demands, the need for foundational technical skills is important to equip graduates for readily available entry level jobs in the field. While the available literature contains elements of application of big data into the classroom, mainstream tools like Apache Hadoop have not been readily addressed. This paper evaluates two different methods of providing students exposure to Hadoop through either an on-premise cluster or virtual machines. A curriculum is provided for students to gain hands-on experience through lab exercises, assessed through pre- and post-quizzes to test understanding. In addition, student work is assessed for application and analysis in a Business Intelligence and Big Data undergraduate course. This work contributes to the information systems (I.S.) community by providing foundational elements essential for integrating software tools such as Hadoop, Hive, and Spark into coursework.

Keywords: big data, Hadoop, business intelligence, curricula, pedagogy, data analytics

1. INTRODUCTION

With the brisk pace of technological advancement, information systems (I.S.) curricula could be rewritten annually and still not represent the most up-to-date standards of the information technology industry. The last published guidelines for model undergraduate I.S. curricula were in 2010 (Topi, et al.). It is a well-researched observation in the I.S. field that there is a gap between the skills that employers want and the skills that college students are actually learning (Wixom et al., 2014). In the data-filled world that exists today, companies seek qualified individuals who can analyze and manage large and complex data sets. According to the Bureau of Labor Statistics (2018), the job market for Computer and Information Research Scientists, on of their closest equivalents to data analytics experts, is positioned to grow by nearly 20 percent by 2026. Many schools have been slow to adapt and incorporate advanced database concepts like big data analytics and NoSQL into their curricula. Traditionally in I.S. curricula the focus remains on skills like infrastructure, programming, systems analysis, and database management.

A midwestern private liberal arts institution has positioned itself to form a collaboration between

mathematics and information systems through a mathematics-focused data science major that infuses the technical data-related courses in I.S. Both data science students and I.S. students need applied skills to enter the workforce prepared to tackle the next set of technical challenges. The knowledge of theories and concepts is beneficial and necessary, but hands-on work with industry standard technology is necessary to differentiate oneself in the workplace. To gain a competitive advantage, emerging technologies should be continually introduced into I.S. curricula. This paper explores the technology behind the Hadoop open source framework for managing big data. This work also includes an evaluation of different methods of delivering Hadoop to students in a course entitled Business Intelligence and Big Data, either through a departmental server cluster or through individual virtual machines on lab computers. A methodology is proposed to assess both learning and value from the hands-on lab exercises. Ultimately, the purpose of this proposed study is to determine effective means for students to gain real experience using big data platforms and software (e.g. Hadoop, MapReduce, etc.). This research contributes to the I.S. community by aiding educators on how best to incorporate Big Data and, more specifically, Hadoop into appropriate courses.

2. REVIEW OF LITERATURE

"We are awash in a flood of data today" (Agrawal et al., 2012). Data is changing the way that entire industries function. Astronomers who used to simply take pictures of the sky are now able to analyze thousands of those pictures every day searching for new stars and galaxies. In healthcare, doctors are being aided by algorithms that sift through research data to predict illnesses and recommend treatments in time to fight previously fatal illnesses. With data from sensors, social media, and transactions, more data than ever is available to store and, more importantly, analyze. Now the challenge is how to do it. Researchers have agreed on three characteristics, often referred to as the "3 V's," that make these problems unfit for traditional relational database management systems: volume, velocity, and variety. Volume refers to the quantity of data, usually in terms of terabytes and petabytes, that companies are taking in every day. Velocity reflects the need for this data to be analyzed efficiently to provide actionable insights in real-time. The last V, variety, represents the different forms that data can take: structured, semi-structured, or unstructured. The combination of these three characteristics is what will often categorize a data set as "big data" (Coronel, 2017).

Companies all over the world are scrambling to hop on the big data train before they get left behind (Hurwitz, 2013). Previously, companies relied on large storage area networks (SAN) to store data and expensive mainframes and supercomputers for analysis. This process where all the computation is done by one machine is referred to as centralized computing. One of the biggest disadvantages of centralized computing is that it takes time to transfer data, especially when sizes reach into the petabytes. When the data are being used for real-time decisions, extra query processing time can make or break applications. Housing large quantities of data presents another challenge, and can be cost prohibitive for many companies. Increasing the storage capacity for centralized systems often means upgrading hardware, which could deplete project budgets rapidly.

Because of time and costs, businesses needed a more efficient way of storing and processing large data sets. Instead of transferring the data to centralized computers to process, in distributed processing the computing is done on the same machines where the data are stored (Hurwitz, 2013). This approach to data processing has led to the development of the Hadoop framework for big data processing. "Apache Hadoop technology is transforming the economics and dynamics of big data initiatives by supporting new processes and architectures that can help cut costs, increase revenue and create competitive advantage" (IBM, 2014).

As Prajapati (2013) states, Hadoop is made up of its storage system, Hadoop Distributed File System (HDFS), and its distributed processing framework (MapReduce). "Scalability and availability are key traits of HDFS, achieved in part due to data replication and fault tolerance" (Holmes, 2015) as seen in Figure 1 below. HDFS stores data sets by breaking them up into "blocks" and replicating these blocks to be stored on a cluster of servers. HDFS functions on the principle of "Horizontal Scaling" (Warden, 2011). As opposed to traditional vertical scaling, where a company would simply buy bigger servers to store more data, in horizontal scaling when storage needs increase, companies instead buy more commodity computers and add on to the Hadoop clusters. This process, combined with the data replication, makes HDFS incredibly fault tolerant, since anytime one of the computers in the cluster (a "node"), fails, it can be easily replaced and the data that are stored on it can be

ISSN: 2473-3857
v4 n4632

copied from the replicated blocks back onto the new node.

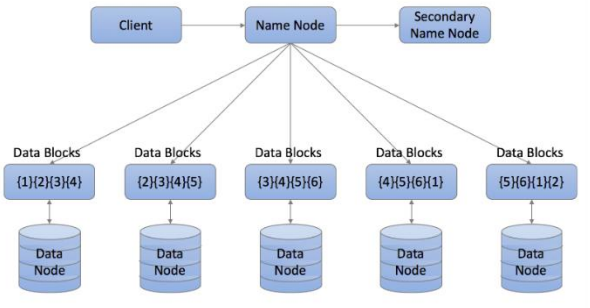

Figure 1. HDFS model

The MapReduce framework partners with HDFS to allow companies to analyze data from different sources that "could take days or longer using conventional serial programming techniques…" (Holmes, 2015). MapReduce is made up of two operations, a map function and a reduce function. The mapping determines how the cluster is going to divide the work and delegate data processing, per block, to computers with available capacity. The reduce function reassembles the data after processing is complete as seen in Figure 2 below. With its ability to store and analyze large data sets on commodity hardware, Hadoop has become an integral part of data analytics at many industry leading companies.

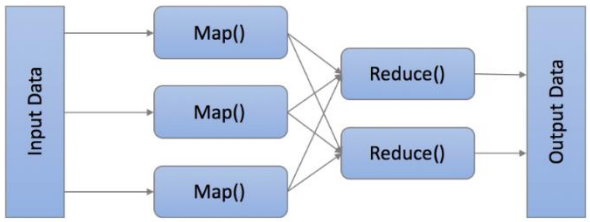

Figure 2. MapReduce model

The current research question is, how can this technology be applied in a classroom setting to prepare students for work in the real world? According to the findings of the Business Intelligence Congress 3, as described in Wixom et al. (2014), "Employers are not satisfied with the practical experience of university graduates" and "Demand for BI/BA students continues to outpace supply." As early as 2003 (Lu, Bettine), the need for I.S curriculum modernization in the areas of business intelligence and business analytics has been recognized. I.S departments have been working hard and fast to incorporate business intelligence into their curricula, to try and satisfy the industry demand for experienced graduates. Mrdalj (2007) writes about the creation of an applied, graduate level, business intelligence course at Eastern Michigan University. More recently, Podeschi (2015) describes the use of an industry software, QlikView, to construct an experiential learning environment in an undergraduate level business intelligence class. The use of industry standard technology is vital to successful hands on learning. Molluzzo and Lawler (2015) proposed a "Concentration Curriculum Design for Big Data Analytics for Information Systems Students". In this paper, the authors discuss the different technologies they plan to incorporate into their data science concentration, including Hadoop and MapReduce. They again discuss the industry need for experts who have the knowledge to work with big data projects, and the lack of supply of such graduates. While previous research includes the discussion of these big data topics in the classroom, there is still a lack of evidence that demonstrates the incorporation of hands-on learning related to tools such as Hadoop, MapReduce, and so on. The research does agree, however, on the need to incorporate industry standard technology to prepare students for the real world.

3. COURSE OBJECTIVES

Students taking the Business Intelligence and Big Data course are part of the I.S. major, data management certificate, or data science concentration from mathematics. Students in the I.S. program are grounded in the theoretical areas of programming, system analysis and design, relational databases, and computer networks. Through the major, students learn hands-on skills in both business and technology that prepare them for jobs managing data, developing web applications, securing systems, and analyzing systems. Prior to this course, it is expected that students have an intermediate knowledge of spreadsheets, an introductory knowledge of SQL, a broad understanding of the fundamentals of information systems, and have taken at least one statistics course. This Business Intelligence (BI) course introduces and builds on the concept of data warehousing and modeling data for analysis rather than for transactional and operational processes. This course provides students with hands-on experience in data warehousing, data analytics, and executive dashboards through real-world data sets and applications. Students are exposed to a variety of software tools throughout the course including: Oracle Database, QlikView, R, and as introduced in this paper, Hadoop and MapReduce. The primary learning objectives of the course are:

- Understand the strategic importance of business intelligence and data analytics.
- Understand the difference between descriptive, prescriptive, and predictive analytics.
- Design and develop a data warehouse based on data needs and user requirements.
- Extract, transform, and load operational data into a data warehouse.
- Build a business intelligence application for dashboarding, analysis, and reporting.
- Interpret data into informed decisions for recommendation.

4. LAB ENVIRONMENT

Prior to developing lab exercises for students, two different platforms were evaluated: departmental servers using manual installation and configuration (on-premises) or virtual machine appliances provided by Cloudera for students to gain hands-on experience. The first approach involved installation and configuration of Hadoop and its associated tools on two used HP servers. Cloudera virtual machine appliances, with Hadoop and tools pre-installed, were also evaluated. These two approaches provide different benefits and costs to the institution, the faculty, and student learning.

On-premises Hadoop Cluster

To build the on-premises option, the lack of computing resources needed to be addressed by acquiring sufficient hardware to support a cluster-computing environment. Through corporate donations and departmental funds, the program acquired two used HP servers and sufficient hard disk storage and memory. Using multiple servers on-premises allows students to observe how distributed computing works in a small environment with one server as the name node and the other as a data node.

Hadoop is built to run on a Linux operating system, so the decision on which operating system to use was first. Red Hat Enterprise Linux (RHEL) is the industry standard for large scale server infrastructures (Gillen, 2017). A significant number of industry companies use RHEL for their server architecture. Because funding was neither available nor warranted for RHEL licenses, CentOS 4.8.5 was selected for its price (free), stability, active support community, and close code base to RHEL. Hadoop 2.7.2 was the latest stable version at the beginning of the project.

Cloudera Virtual Machine

The other architecture evaluated was a virtual machine appliance provided by Cloudera, a value-added Hadoop vendor. In this scenario, students would be able to have their own Hadoop cluster in order to follow prescriptive labs, allow for iteration, and even experience failures. While Cloudera does not provide free licenses for their full enterprise distribution of Hadoop for academic use, they do offer a free sandbox virtual machine called Cloudera QuickStart. Cloudera QuickStart allows users to download virtual machines (VMs) or Docker images that come pre-installed with Hadoop and a set of applications that can be built into the Cloudera Distributed Hadoop (CDH) platform as seen in Image 1 in Appendix A. This platform gives students the ability to build their own isolated Hadoop environment to use in class, as well as develop and test applications they wish to eventually run on the full cluster environment.

Both environments allow students to see how the Hadoop ecosystem functions. Building an on-premises cluster for student-use takes a significant amount of effort for system and network configuration along with creating a cluster. This environment would be beneficial for students who are interested in pursuing careers as server or system administrators as more of the individual components are exposed. However, a department should go through a proper analysis of available resources for not only installation and configuration, but also maintenance and support. While beneficial to understand the architecture, the cost of time may be too burdensome for departments with fewer resources. Cloudera QuickStart, however, provides a quick and easy way for students to deploy an environment with the appropriate tools for understanding the fundamentals. Exercises and labs, in this instance, are likely to be smaller in scale, as hard disk space and memory capacity become concerns. For purposes of getting students early exposure to Hadoop and its tools, Cloudera is a better suited platform for designing introductory labs. While more resource intensive, an on-premises server cluster is likely more appropriate for independent research or client projects where additional computing resources like CPU, memory, and disk space are necessary.

6. PROPOSED METHODOLOGY

This research proposes that Cloudera QuickStart be used for introductory labs for students to gain initial exposure to Hadoop and its associated tools such as Hive and Apache Spark. Learning the Hadoop ecosystem, the available tools, and its terminology (as depicted in Figure 3 below, is essential to understanding. It is planned for these lab exercises to take place during the last third of the semester. By this point in the course,

ISSN: 2473-3857
v4 n4632

students will have been exposed to traditional forms of business intelligence through data warehousing concepts along with descriptive, prescriptive, and predictive analytics.

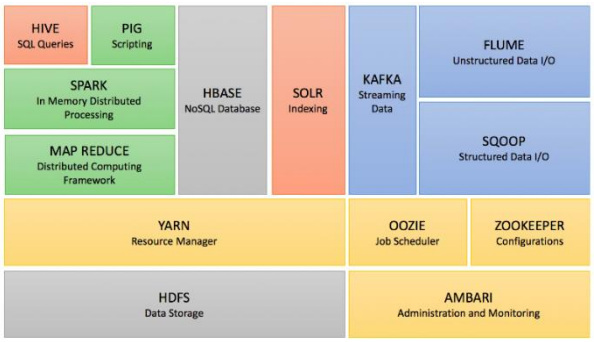

Figure 3. Hadoop ecosystem

The primary learning objectives for this module that support the course objectives are to: 1) recognize and recall definitions and terminology related to big data and Hadoop; 2) understand the key difference between traditional and NoSQL databases and the use cases for each; 3) understand the key advantages of distributed data storage and processing; and 4) gain an introductory exposure to Hadoop, HDFS, Hive, and Spark. These learning objectives intentionally map to Bloom's Revised Taxonomy as seen in Table 1 (Krathwohl, 2002). As a result, students move from lower to higher order thinking skills throughout the course of the module. From the knowledge dimension, students should be able to gain both factual and conceptual knowledge about these elements, and through the lab exercises, gain procedural knowledge of what steps to take and when, specifically, to take them based on a given problem.

This module of the course will begin with an overview of big data, Hadoop and its associated tools, along with appropriate terminology and definitions. Student can gain further understanding through cases, and in-class demonstrations. More detailed Hadoop, HDFS, Hive, and Spark material can be supplemented through additional videos and content produced by Cloudera's OnDemand training, such as the Cloudera Essentials for Apache Hadoop course. Once a solid foundation is achieved, students will then go through lab exercises, as described, to gain an exposure to Hadoop and its associated tools.

Learning Goal	Bloom's Revised Taxonomy
Recognize and recall definitions and terminology related to big data and Hadoop	Remember, Understand
Understand the key differences between traditional and NoSQL databases, and the use cases for each.	Understand, Analyze
Understand the key advantages of distributed data processing through HDFS	Remember, Understand
Gain an introductory exposure to Hadoop, HDFS, Hive, and Spark	Apply, Create

Table 1. Module learning goals mapped to Bloom's Revised Taxonomy

Lab Exercises

With every student having their own full environment through Cloudera QuickStart, it became possible to spend more time on labs and less time getting students an environment ready. With Cloudera QuickStart, instead of narrowing the curriculum to look at just Spark and HDFS, it could incorporate a holistic education on the Hadoop ecosystem. Three main labs were developed to teach these tools in the Business Intelligence and Big Data class. The first of these labs focuses on the use of HDFS, simply inserting files, moving files, and then reading files from the command prompt. This lab will be combined with the setup of the VirtualBox Cloudera QuickStart VM. The second lab involves building a database schema in Hive through the command prompt as seen in Image 2 in Appendix A. This lab is a review of SQL, and will pair with the data warehousing component of the Business Intelligence and Big Data class, allowing students to build a database they will have developed earlier in Oracle Database, and then rebuild it in HiveQL. The last lab focuses on using the PySpark or SparkR API's through the command line as seen in Image 3 in Appendix A. Both the R, and Python programming languages are taught, so the student can use whichever language they are more familiar with to write a Spark job to do a word frequency analysis on selected text.

Assessment

To evaluate the effectiveness of the first implementation of the curriculum, students will be assessed on the first two domains of Bloom's Revised Taxonomy, remembering and

understanding, through administering pre- and post-quizzes in the fall 2018 semester. The quiz is designed to measure the comprehension and retention of the material at a conceptual level. Examples of tasks students will be asked in the quiz are: differentiating between properties of NoSQL and traditional databases, recalling key benefits of the MapReduce paradigm, matching Hadoop ecosystem applications to their respective functions, and differentiating between distributed storage and centralized storage. The quiz will be administered at the beginning of the big data unit to establish a baseline of the students' preexisting subject knowledge. The same quiz will be administered after the unit to provide evidence of the impact of the coursework.

Application and analysis of key tools (HDFS, Spark, Hive and HBase) will be assessed through the hands-on labs as described in the methodology. During key points of the labs, students will be asked short essay-style and reflective questions requiring them to connect the class content to the lab exercises, and ideally, application. The rubric for the labs will focus on the successful completion of the prescribed exercises, as well as the students' thorough responses to the prompts.

Expected Outcomes

The goal of these hands-on lab exercises is for students to walk away with more than just exposure to current tools in the big data space. Ideally, after comparing results from the pre-assessment to the post-assessment, students should perform better on the post-assessment. Reflective responses should demonstrate a connection between the knowledge and the application through the hands-on labs. The quantitative and qualitative analysis of the assessment instruments will provide valuable information on how to structure hands-on labs related to big data in the future, with implications for contributing to best practice pedagogy in big data. Moreover, information systems educators have wrestled with similar issues in the past such as how to provide students with hands-on experience in areas of programming, database, and so on. It is anticipated that the results of this research study will help educators better align applied big data and related courses with other hallmark information systems courses such as system analysis and design, database, and enterprise architecture.

7. DISCUSSION AND CONCLUSIONS

After assessing the outcomes of the first implementation of this classroom module, the labs will be revised in response to students' feedback. As big data technology continues to mature, the tools emphasized in the curriculum will need to be reviewed on an ongoing basis to assess industry relevance and currency. Furthermore, the engagement of industry experts, some of them alumni, and their input through such mechanisms as advisory boards will be beneficial to keeping the content relevant.

This paper provides a good point of reference for information systems programs looking to incorporate big data skills into their curricula. The details of this learning module should serve as a template for developing new labs based around Hadoop. While the content is subject to individual preference, the keys to making labs like these effective are to build them upon preexisting skill sets and make them easily accessible to students with limited domain knowledge. Utilizing the Cloudera QuickStart VMs removes a number of technical knowledge barriers and provides a no-cost solution to giving students hands-on Hadoop experience. The installation and configuration process for the full cluster took a fair amount of time, and it is plausible to assume that the level of effort, including long-term administration, may not be reasonable for some institutions. The administrative burden would likely be left to faculty, which leads to concerns over re-configuring the environment for each iteration of the class. For this platform to be successful, configuration automation would need to be developed, in addition to clear documentation. For these reasons, Cloudera QuickStart provides the most cost-effective and efficient solution for labs in a classroom setting.

For institutions interested in the server cluster environment, it is important to note the issues involved with sustainability. For the cluster to be sustainable, there needs to be a way in which it can be maintained, updated, reconfigured and, if necessary, rebuilt. It requires dedicated personnel who have some knowledge of the Hadoop ecosystem and thorough documentation for reference to manage the ongoing maintenance, updating, and troubleshooting that would be necessary for an environment like this. A proposed solution is to create a sustainable fund, through a combination of donor funding and third-party business investment, to employ a student to oversee the environment and perform the necessary maintenance. It is, however, unrealistic to expect that every student system

administrator will have the time or knowledge to rebuild the environment in the event of a system failure, even with the best documentation. A necessary extension to this solution is to create an automation script, using Python, that will allow student administrators to easily deploy new servers or to rebuild the system.

8. NOTES

1. Cloudera's QuickStart VM is available at https://www.cloudera.com/downloads/quickstart_vms/5-13.html
2. Hadoop cluster specifications: Two HP DL380 G7 servers, 1.4 terabytes of hard disk storage, and 128 gigabytes of RAM.
3. CentOS is available for free at https://www.centos.org/

9. REFERENCES

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., … Widom, J. (2012). Challenges and opportunities with big data [White Paper]. Retrieved April 17, 2017 from https://www.purdue.edu/discoverypark/cyber/assets/pdfs/BigDataWhitePaper.pdf

Doherty, C., Camina, S., White, K., & Orenstein G. (2016). The path to predictive analytics and machine learning. Sebastopol, California: O'Reilly Media, Inc.

Gillen, A., Marden, M., & Perry, R. (2017). The business value of Red Hat Enterprise Linux [White paper]. Retrieved May 9, 2018, from https://www.redhat.com/en/resources/idc-whitepaper-business-value-red-hat-enterprise-linux

Guttag, J. V. (2016). Introduction to Computation and Programming Using Python. Cambridge, MA: The MIT Press.

Holmes, A. (2015). Hadoop in practice. Shelter Island, NY: Manning Publications Co.

Hurwitz, J., Nugent, A., Halper, Dr. A., & Kaufman, M. (2013). Big data for dummies. Hoboken, NJ; John Wiley & Sons, Inc. IBM. (2014). Big data integration and Hadoop [White paper]. Retrieved March 18, 2017, from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14791USEN

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice, 41(4), 212.

Molluzzo, J.C., & Lawler, J.P. (2015). A Proposed Concentration Curriculum Design for Big Data Analytics for Information Systems Students. Information Systems Education Journal, Volume 13 No. 1, 45-52.

Lu, Y., Bettine, J. (2003). Data Mining: An Experimental Undergraduate Course. Journal of Computing Sciences in Colleges. 18(3), 81-86.

Mrdalj, S. (2007). "Teaching an Applied Business Intelligence Course." Issues in Information Systems, 8(1), 134-138.

Podeschi, R. (2015). Experiential Learning using QlikView Business Intelligence Software. Information Systems Education Journal, 13(4) pp 71-80.

Prajapati, V. (2013). Big Data Analytics with R and Hadoop. Birmingham, UK: Packt Publishing Ltd.

Topi, H., Valacich, J., Wright, R., et al. (2010). IS 2010: Curriculum Guidelines for Undergraduate Degree Programs in Information Systems. Association for Computing Machinery and Association for Information Systems.

United States Bureau of Labor Statistics. (2018). *Computer and Information Technology Occupations*. Available from https://www.bls.gov/ooh/computer-and-information-technology/home.htm

Warden, P. (2011). Big data glossary. Sebastopol, CA: O'Reily Media Inc.

Watson, H. (2008). "Business Schools Need to Change What They Teach." Business Intelligence Journal, 13(4), 4-7.

Wixom, B., Ariyachandra T., David, D., Goul, M., Gupta, B., Iyer, L,. … Turetken, O., (2014). "The Current State of Business Intelligence in Academia: The Arrival of Big Data." Communications of the Association for Information Systems. 34(1), 1-13.

ISSN: 2473-3857
v4 n4632

Appendix A

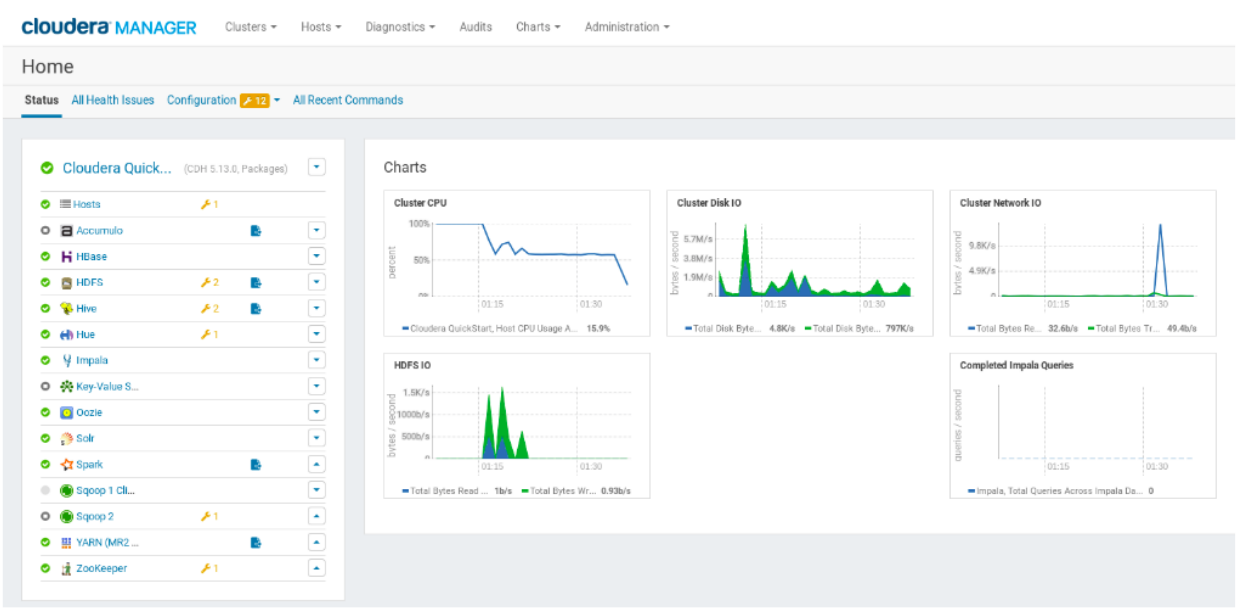

Image 1. Cloudera QuickStart VM

```
hive> SHOW DATABASES;
OK
default
userdb
Time taken: 0.008 seconds, Fetched: 2 row(s)
hive> CREATE TABLE IF NOT EXISTS manager ( eid int, name String, slary String, destination String);
OK
Time taken: 0.111 seconds
hive>
```

Image 2. Hive command prompt

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.6.6 (r266:84292, Jul 23 2015 15:22:56)
SparkContext available as sc, HiveContext available as sqlContext.
>>> print ("Hello World")
Hello World
>>>
```

Image 3. PySpark command prompt