The CTE Journal ISSN 2327-0160 (Online)
Volume 3. Number 1.

Addressing the Challenges of Teaching Big Data in Technical Education

Christopher B. Davison

Ball State University

cbdavison@bsu.edu

Abstract

There is a growing demand for professionals with knowledge of big data and data science. This knowledge can range from data storage, systems infrastructure, to data analytics. The problem in technical education is two-fold: big data is a relatively recent phenomenon, and the infrastructure required for big data is prohibitively expensive. The purpose of this paper is to present strategies that will assist educators in teaching big data on a little classroom budget.

Introduction

Big data is everywhere. From the choices found at login to Amazon (i.e., Amazon's recommendations) to the coupons received in the mail. Big data and the concomitant data analytics provide useful information to any organization. The issue addressed in this paper is teaching technology students big data concepts (e.g., storage, retrieval, and analytics) on a little classroom budget. Servers and storage are expensive and big data requires massive storage, retrieval, and CPU cycles. The networks that transport big data are high speed and expensive as well. The analytics software required to process massive amounts of data can be expensive and complex. All of these issues create a complex scenario with regard to teaching big data on a little classroom budget.

This paper begins by providing definitions specific to big data and data sciences. This will set the context for the subsequent sections and the discussions therein. Following that, the importance of big data education is discussed. The next section explores specific challenges faced by educators when attempting to teach big data concepts and practices. This is followed by a presentation of tools that can be utilized by educators when teaching big data. Finally, three classroom exercises are presented that will engage students in learning about big data.

Definitions

As it is a recent phenomenon, big data's definition is a moving target. The size and scope of what constitutes big data can vary. In the general sense, the size of big data is an *N* where traditional Relational Database Management Systems (RDBMS) cannot scale to process the enormity of the data. In today's sense, this can be measured in petabytes of information. However, as that number is a variable and data management software undergoes continual updating and improving, it could be much larger tomorrow. According to Hashem, Yaqoob, Anuar, Mokhtar, Gani and Khan (2015), big data is more than size but it is encompasses the integration of techniques and technologies to uncover hidden values in complex, massive, and heterogeneous data sets.

Laney (2001), provides a more holistic and three dimensional definition for the growth and challenges associated with big data. His definition addresses not only the size but the scope of big data: The Three Vs (Volume, Velocity, and Variety). Volume refers to sheer size (quantity) of the data. Volume is perhaps the most important characteristic of the definition as the term big data itself implies volume and is relative to size. See Figure 1 for a graphical representation of Laney's work.

Velocity refers to the speed at which big data is generated and processed. An example of this is the streams of data generated in multi-modal sensing environments. Large sensor arrays generating massive streams of low-level data can potentially overwhelm networks and data management systems. As these sensor streams converge to their destination, the sheer velocity of the data presents interesting challenges.

Finally, variety refers to the heterogeneous nature of big data. The data could be structured in the typical RDBMS sense or quite unstructured such as combinations of video, text, and various file formats streaming into a big data system. Managing, cataloging, indexing, and providing retrievals and analytics on top of a variety of data pose unique challenges to system designers.

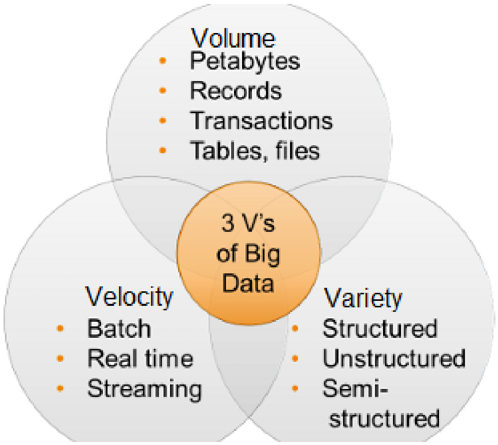


Figure 1. The three V's of big data. This figure illustrates the three elements that define big data.

Another facet of big data requiring definition is data analytics. Data analytics includes the technologies that make sense of big data and provide meaning from it. While business intelligence models and analyzes current and past periods, data science is more forwarding looking and produces predictive models (EMC, 2015). In the aggregate, data analytics encompasses the tools, technologies, and techniques that transform big data into useful, actionable information. Data analytics often makes use of sophisticated software for data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, complex event processing and prescriptive analytics. Algorithms to perform this level of knowledge creation and management are complex and quite sophisticated. The software expense reflects this level of sophistication.

The concepts above and their concomitant definitions are normally derived from the Information Systems (IS) domain. Conversely, the systems and networking required to host, transmit, and process big data are found in the Information Technology (IT) domain and will be discussed further in this paper.

Importance of Learning Big Data

The U.S. White House has announced several Big Data partnerships in what the government refers to as Data to Knowledge to Action partnerships. John Holdren (2013), White House Office of Science and Technology Policy Director, calls Big Data a "big deal" (para. 4). From a government perspective, big data is important to not only the business of running the U.S., but it is also important to the economic future and well-being of the country.

According to the USA Today (2013), the sexiest job of the 21st century is a data analyst. The newspaper quantifies the starting salary in the $125,000 per year range. Additionally, USA Today finds the demand for such skills far exceeding the supply: roughly 20 percent of the demand is met. From an employment perspective, high salaries and a plentiful employment market makes big data analytics an enticing area.

The future for the data science job market appears healthy. The McKinsey Global Institute (2011) report is projecting that by 2018, "the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (para. 8).

Pedagogical Issues Educators Encounter with Big Data

While learning about big data is important and the career outlook for big data employment is optimistic, educators face challenges in teaching data science. First there is the lack of technology infrastructure to teach the subject. Secondly, since the concept of big data is a relatively recent phenomenon, the instructor's knowledge of data sciences and available tools might be lacking. In order to address these issues, there are a number of tools and partnerships available to educators and academic institutions. Many of these tools are free, open-source, or web-based tools.

Consider the expense and complexity of the eBay.com 40 petabyte (PB) data warehouse; the company has two of these in separate clusters. Furthermore, eBay has an additional Teradata 7.5PB data warehouse (Tay, 2013). For educators to attempt to duplicate the smaller 7.5PB warehouse would require thousands of dollars in hardware expenses for storage arrays, networking, and clustered computational systems. The software (operating systems, database and data analytics) will add thousands of dollars more to the cost. Finally, there is the question of populating the database with information. The analytics software must process data and these data have to originate from somewhere and be meaningful.

The IT infrastructure required to manage big data can be quite complex and expensive. Considering the velocity aspect of big data, the networking infrastructure must be fast and the corollary to fast is expensive. The volume and variety aspects require massive storage, retrieval, and CPU cycle architectures. Again, this can be an expensive proposition; prohibitively so for educators and the educational environment. As an alternative to engineering the IT infrastructure, educators can turn to cloud-based computing which would provide massive computational capabilities without the need to maintain expensive hardware and software (Hashem, Yaqoob, Anuar, Mokhtar, Gani, & Khan, 2015). However, cloud computing and storage is often expensive even for educators with academic pricing.

Big data requires not only complex hardware and software but also new techniques for computational analysis. MapReduce is one such framework for processing and generating large data sets (Dean & Ghemawat, 2004). The MapReduce framework is designed to work on a distributed system (commodity PCs) and it parallelizes (i.e., adapts for concurrency) and distributes the computations and data across the system. The framework analyzes geographical and network costs and processes data locally to the storage while also providing fault tolerance from node crashes. The Map function processes data in parallel across assigned nodes then reports results to the Reduce function that merges those results. The Apache Hadoop system has an implementation of MapReduce that is free and open source. Hadoop is available at: http://hadoop.apache.org/ with a number of release versions available.

Even if the aforementioned technology challenges in hardware and software are overcome, educators must be trained in big data science. As big data analytics are complex and new, many instructors may not have the requisite knowledge which draws from a variety of domains. A strong quantitative background and knowledge of statistics is necessary. There is often a great deal of IS and IT infrastructure knowledge and troubleshooting that comes with big data education. Furthermore, it requires a good deal of effort, expense, and training to become proficient in just one aspect of big data.

While there are a number of challenges for big data educators, there are a number of tools, many of which are free, available to educators. These tools are discussed in the next section.

Tools for Educators

There are a number of free tools available for educators and data scientists. Tools such as R (The R Project for Statistical Computing: http://www.r-project.org/Goo) is open source and can be downloaded free of charge. While there will not be terabytes of data for students to practice analytics, the R software is relatively straightforward to download and setup. It supports Windows, Mac, and Linux. Furthermore, R is extensively utilized in the data analytics community (Tippmann, 2015) so this can provide the fundamentals for students. There is a distinct learning curve for R as it is a programming language. R is a command line, interpreted language and as with any command line programming language, practice and effort must be invested.

To assist with learning and using R, another free tool, RStudio, is available. RStudio is an open source, graphical user interface (GUI) integrated development environment (IDE). RStudio is available at: http://www.rstudio.com/ and fully integrates with R. The R package (version 2.11.1 or higher) must be installed first before loading RStudio.

Installing these tools in a Windows workstation environment requires certain operating system permissions that may not be granted to instructors or students (Davison, 2015). Institutional technical support policies often prevent instructors from installing software. Many classroom environments reset the laboratory computer systems' state (e.g. Deep Freeze system restore) to a base level installation. An effective compromise is to install the software in a virtual machine environment where the instructors and students have administrator privileges on the virtual machine and the technical support personnel retain administrator control of the hosting machine. This creates a fully functional testbed for students including machine setups and operation (Gonzales, Romney, Bane, & Jeneau, 2013). Virtual machine environments are common place in online educational environments as well (Chao, Hung, & Chen, 2012)

As discussed earlier, mining big data can be problematic from a logistical sense. Most educators do not have large data sets from which to mine, but there are free and web-based tools to mine Google and Twitter data.

Free and web-based data analysis tools are provided by Google with their Correlate (http://www.google.com/trends/correlate) and Trends (http://www.google.com/trends/) tools. Correlate was originally designed to analyze flu-related searches (https://www.google.org/flutrends/) compared to flu activity (Ginsberg, Mohebbi, Patel, Brammer, Smolinski & Brilliant, 2009). The tool was expanded to allow users to correlate search activity (Vanderkam, Schonberger, Rowley, & Kumar, 2013). Additionally, users may upload their own data for analysis. Google Trends allows users to analyze search terms and search activity as well as filter results based on locations and date ranges. Google Correlate can create heat maps based on US State maps (comparing search activity by state) as well as provide line and scatter plot diagrams. Both tools allow users to download their data as a CSV file, if they have a Google account.

A free web-based tool to analyze Twitter content (http://maps.iscience.deusto.es/) is provided by iScience Maps (Reips & Garaizar, 2011). Twitter activity can be analyzed from a worldwide or local search. The search can be fine-tuned with time periods and radius of search area. The count (number of hits) that the search term is found in the parameterized Twitter activity is provided along with the ability to download the data in a CSV file.

Twitter provides an Application Programming Interface (API) for programmers and scientists to sample their data. The sample set is a random sample of 1 to 10 percent of total content. While programmers are free to design their own customized Twitter Activity data miner, the iScience Maps team provides a web interface to mine the Twitter activity for users with no programming knowledge.

Classroom Activities

The following three classroom activities are designed to engage students in big data concepts and practices. While there is no *a priori* knowledge of big data concepts assumed; basic computer, mathematics, and Windows proficiency is required. These activities are meant to build upon each other. The first is an exercise designed to assist students in beginning to quantify data size (volume). The second is an activity that will examine the speed at which big data can be transmitted (velocity). The third exercise is a data analytics exercise examining Google queries (variety) and the difference between causation and correlation.

Learning Objectives

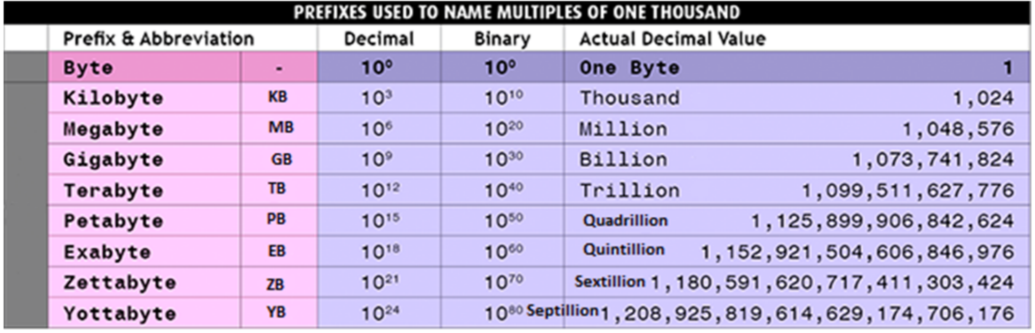
1. Students will demonstrate knowledge of quantification of data.
2. Students will mathematically analyze and calculate transmission speed of big data.
3. Students will synthesize and evaluate Google queries in terms of frequency, distribution, correlation, causation, and geography.

Required Materials

1. Computer with Web Browser
2. Internet Connectivity
3. Calculator

Procedures

1. Quantification of Big Data
    a. The instructor begins the discussion on "How big is big data?" and shares information regarding large, commercial data warehouses such as eBay, Google, Twitter, and Amazon. Big data classification issues such as volume, velocity and variety are discussed.
    b. The following figure (Figure 2) can be used to explain byte quantification prefixes and sizes.



PREFIXES USED TO NAME MULTIPLES OF ONE THOUSAND

Prefix & Abbreviation		Decimal	Binary	Actual Decimal Value	
Byte	-	10^0	10^0	One Byte	1
Kilobyte	KB	10^3	10^{10}	Thousand	1,024
Megabyte	MB	10^6	10^{20}	Million	1,048,576
Gigabyte	GB	10^9	10^{30}	Billion	1,073,741,824
Terabyte	TB	10^{12}	10^{40}	Trillion	1,099,511,627,776
Petabyte	PB	10^{15}	10^{50}	Quadrillion	1,125,899,906,842,624
Exabyte	EB	10^{18}	10^{60}	Quintillion	1,152,921,504,606,846,976
Zettabyte	ZB	10^{21}	10^{70}	Sextillion	1,180,591,620,717,411,303,424
Yottabyte	YB	10^{24}	10^{80}	Septillion	1,208,925,819,614,629,174,706,176

Figure 2. The quantification of data. This figure illustrates large numbers and their representation.

2. Transmission of big data
   a. The instructor explains basic computer networking concepts such as bandwidth, propagation delay, transmission delay, and link speed. Discuss how the speed of light (*c*) and geographic distance impacts data transmission rates.
   b. Students can calculate the transmission delay of 1 byte of data versions the transmission delay of 1PB of data over various link speeds, using the following formula:
   T = L/R, where T = time (seconds), R = link bandwidth (bits per second), L=packet length (bits).
   c. Consider the impact of *c* (propagation delay) on big data. Discuss how propagation delay exceeds transmission delay in implementations such as the Mars Rovers.

3. Analyzing Google Queries
   a. The instructor will guide the discussion to analysis of big data, statistics, and analytics. The instructor will discuss with the students the difference between causation and correlation.
   b. Students will use Google Correlate to find search patterns that correspond to real-world trends. Using the following tool: http://www.google.com/trends/correlate students will enter in a search term and find the closest correlated search terms. Students will then critically analyze the results for both correlation and causation.
   c. Using the results from (b) above, students will further explore the frequencies of their search terms by geography. Google supports analysis of search terms by country and further geographical analysis by U.S. states.

Conclusion

In this paper, the proliferation of big data was discussed. Big data was defined from the perspectives of size (volume), the speed at which it is generated (velocity), and the various forms that it can manifest (variety). The employment gap in the U.S. coupled with the high salary for data scientists should encourage interest in learning about this field. However, educators face a number of issues when teaching big data, notably the lack of tools or training (as big data is a relatively new concept) and the lack of infrastructure. There are a number of tools available to address these issues including free tools such as Google, iScience Maps, and R. Finally, three classroom exercises were offered to assist instructors in teaching big data on a little classroom budget.

References

Chao, K. C., Hung, I. C., & Chen, N. S. (2012). On the design of online synchronous assessments in a synchronous cyber classroom. *Journal of Computer Assisted Learning, 28*(4), 379-395.

Davison, C.B. (2015). Assessing IT Student Performance Using Virtual Machines. *TechDirections, 74*(7), 23-25.

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Larger Clusters.
OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA, December, 2004.

EMC Education Services (2015). *Data Science and Big Data Analytics*. Indianapolis, IN: Wiley.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*.

Gonzales, R.F., Romney, G.A., Bane, C., & Juneau, P. (2013). Virtual education laboratory test bed experimentation. *Journal of Applied Learning Technologies, 3*(1), 6-11.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues, *Information Systems* (47), 98-115.

Holdren, J. (2013, Nov). White House announces big data partnerships. *FCW The Business of Federal Technology*. Retrieved February 18, 2015 from: http://fcw.com/articles/2013/11/12/white-house-big-data-partnerships.aspx

McKinsey Institute (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved February 18, 2015 from: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Reips, U.-D., & Garaizar, P. (2011). Mining Twitter: Microblogging as a source for psychological wisdom of the crowds. *Behavior Research Methods, 43*, 635-642.

Tay, L (2013). Inside eBay’s 90PB data warehouse. *ITNews*. Retrieved February 20, 2015 from: http://www.itnews.com.au/News/342615,inside-ebay8217s-90pb-data-warehouse.aspx

Tippmann, Sylvia (1 Jan 2015). "Programming tools: Adventures with R". *Nature* (517), 109–110.

Vanderkam, D., Schonberger, R., Rowley, H., & Kumar, S. (2013). Nearest Neighbor Search in Google Correlate. Retrieved March 4, 2015 from: http://www.google.com/trends/correlate/nnsearch.pdf