Teaching Big Data: Experiences, Lessons Learned, and Future Directions

by Betsy Page Sigman, William Garr, Robert Pongsajapan, Marie Selvanadin, Kristin Bolling, Greg Marsh, Georgetown University

What Is Big Data?

There is disagreement about what constitutes the field of Big Data. For the purposes of this paper, Big Data can be defined by these two characteristics: (1) Massive quantities of data, which may consist of flows of real-time data; and (2) various degrees of structure in the data, which may include anything from personal information stored on social networking sites to remote sensor data.

When talking about the field of Big Data, one must also talk about the tools that have arisen to analyze it. These tools include massive, inexpensive server farms and open-source software for both parallel processing on multiple servers and analyzing the data to develop insights. Words—word “clouds” in which the size of the word reflects its prevalence—show certain words such as Hadoop, unstructured, structured, SQL, Hive, and H-Base as prevalent in the common usage of the term “Big Data.” Although there is no single agreed-upon definition, Gartner’s interpretation seems to be widely accepted. It states: “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” (Gartner, 2013).

There is some question as to whether the term “Big Data” will stick. Gartner predicts that the term “Big Data” will become obsolete and subsumed into the term “data” by 2020 (Gartner, 2012). Large amounts of data are becoming easier and less expensive to store. Storage racks, used to hold massive amounts of data, are easier to set up and use today. And by 2020, it is estimated that a petabyte of data will cost only $4 to store (Siegele, 2011). Big Data, regardless of what it is called, is more than just a buzzword and will continue to grow in importance in the world.

Making a Case for Big Data Education

As more businesses, government agencies, non-profit organizations, and academia race to harness the much-publicized power of Big Data analytics, they look for people to fill the jobs that this analysis requires. Recently, Gartner noted in its 2013 Big Data study that 64 percent of organizations surveyed were either investing or planning to invest in Big Data technology (Gartner, 2013). The U.S. Department of Labor made the same sort of forecast, predicting that there will be a growth of 25 percent in data analytics jobs by 2018 (Bertolucci, 2013).

These organizations will need employees and/or contractors to take advantage of Big Data technology. McKinsey has famously predicted that the U.S. will be seriously short of the people needed: “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” (McKinsey, 2011).

To find the right employees is not easy. Numerous sources have pointed out the variety of skills needed by data analysts and data scientists. The complete data scientist today has many chal-
Challenges. He or she needs to be able to not just understand statistics, data modeling, and analytics but also information technology, computer science and machine learning. And the data scientist must also understand how to get results quickly, understand the data’s business or organizational importance, and perhaps most importantly, be able to convey the results clearly to those who can use them for decision making (IBM, 2012).

Further, the need for analytical training is increasingly moving from the realm of back-office analysts to higher levels of management. Leaders increasingly need to be able to understand data and then know how to ask the right questions to get results that will enable them to get useful answers. They need to be versed in data visualization as well, so that they can quickly understand the results that managers bring to them daily. While these abilities may not always require deep technical expertise, they do require an understanding of data and how it can be used to answer questions about business and organizational processes, customers, and metrics.

The Need for Employees Who Understand Big Data: Businesses, Government, Non-Profit Organizations, and Universities

Businesses, with their bottom lines at stake, express the most immediate concerns for training and finding employees to fill Big Data positions. Gartner (2013) recently predicted that “through 2015, 85 percent of Fortune 500 organizations will be unable to exploit big data for competitive advantage.” Conceivably, companies could address this gap internally, through training or hiring more employees. However, training can be expensive and take away from other concerns of the business. And the time and resources used to hire more employees can also impact a company’s bottom line. Respondents at organizations “currently using or planning to use data analytics, BI or statistical analysis software” agreed that “Big Data expertise is scarce and expensive” (Information Week, 2013). To help address the gap in available talent, IBM, Microsoft and other technology companies have begun to develop specific and detailed programs for reaching out and supporting academics incorporating Big Data in the classroom (Gardner, 2013; Fischman, 2011).

Government has always been the most prolific collector, generator, and analyzer of big data, and the most data-intensive agencies (e.g., the Bureau of Labor Statistics and the Bureau of the Census) have many computer scientists, data analysts and statisticians on staff, but they will need many more skilled employees in the near future to deal with the geometrically expanding volume of data.

Nonprofits generally have less money to spend than other similar-sized organizations, but they should also take advantage of opportunities Big Data analysis gives for better decision making and learn how to be more effective with the data resources they possess. Big Data analysis will help them with this decision-making process.

Lastly, academia has increasingly realized the significance of the role it can play in the Big Data revolution. Around 60 articles in the last few years of the Chronicle of Higher Education contained the phrase “Big Data.” On campuses, there is a new focus on “learning analytics”—essentially Big Data techniques applied to learning—to help schools become more effective and efficient via insights produced by data analysis, as they experience increased pressures to keep from raising costs.

How Universities Are Responding to This Need

Where are organizations going to find people with the technical and analytical skills they need? While argueably a little late to the game, academia has recently begun to respond to the need for data science programs at the bachelor and masters levels (Williams, 2013), largely in part due to the widespread interest in Big Data from students in many different schools and disciplines. More universities are rolling out analytics degrees and certificates and developing data science courses (Williams, 2013). Some of these are taught largely in the classroom (Carnegie Mellon), some are taught almost completely online (Northwestern), and many are hybrids of the two methods.

Where are these programs housed? Data is incomplete and constantly changing, due to the influx of new programs, but of the top 20 programs listed by Doug Henschen (Information Week, 2013), 50 percent (10) are housed in a business or management school. Others are found in engineering and computer science departments/schools (6) and statistics/analytics (2) departments. (See Appendix A.)

Many universities are still wrestling with the question of where an analytics program should be housed, but there is general agreement about the need for cross-campus collaboration when considering data analytics. In fact, that issue was one of the primary topics addressed in the recent conference, “Big Data: Educating the Next Generation,” held at Georgetown University on April 3, 2013. The event brought together faculty and business people across multiple disciplines and industries to collaborate and encourage discussion about how best to educate students about Big Data (Georgetown University, 2013).

Some schools have solved this issue by creating a separate center. DePaul University, for instance, created a Center for Data Mining and Predictive Analytics, a joint venture of the Department of Marketing and the College of Computer Science and Digital Media. And at Georgetown, the McCourt School of Public Policy recently announced the addition of a Massive Data Institute, which will “train the next generation of leaders to critically analyze, extract and use these large sets of data to better inform public policy” according to President DeGioia (Georgetown University, 2013). This institute will have links to several other departments and centers across campus.

Teaching about Big Data: Bridging the Technical with the Understandable

Business professors face additional challenges as to how to address the skill gap as they attempt to fulfill the imperative
of preparing students to take meaningful jobs in business and other sectors. They are asked by the AACSB (the accreditation organization for business schools) and their own colleges and universities to give both a well-rounded business as well as liberal arts education at the undergraduate level. At the MBA level, students concentrate more exclusively on the business fields, but professors must teach students about a wide variety of business topics (finance, marketing research, operations, information systems, accounting, etc.) in a limited amount of class time.

To incorporate Big Data in the classroom, professors first must decide how to teach it in the context of the definition they have decided on. What should the goals of Big Data education be? The desire is to give our students knowledge that will allow them to understand big data and how to use it. But how deep do we want them to go?

For example, in a business school, how much statistics should be required? Most business schools require statistics, so we can usually assume some statistical knowledge, at least in an upper-level undergraduate or MBA class. But how should we further develop students’ statistical skills so that they can become proficient at developing insights from data that will be useful to their employers? Students are eager to learn about those skills and insights that they can immediately apply in their professional careers, and are sometimes less enthusiastic about learning deeper technical skills.

What about data modeling? Introductory level statistics touch on this, but may not delve very deeply into it, nor give students the expertise in forecasting they need. Statistical tools, although important building blocks for working in Big Data analysis, have been sometimes lessened or “watered-down,” and statistics professors may struggle to make a case for the role of these courses in the business curriculum. Despite the fact that statistics and/or data science is now, much to the astonishment of career statisticians, increasingly referred to as “sexy” in the literature (McKinsey Quarterly, 2009), hard skills can still sometimes be a hard sell (Griffith et al., 2012).

And what of other skills, such as an understanding of information systems and machine learning? Not all schools would include this and certainly, the information systems knowledge needed for Big Data differs from the material covered in an introductory systems analysis class. Yet, to understand Big Data analysis, there is a need for some understanding of the information systems that make the analysis possible. And should Big Data education address even more complicated and technical topics—such as Hadoop programming and machine learning or is this asking too much of undergraduate students?

Finding a Platform to Use

In addition to finding the right balance between business and technical material, choosing a Big Data platform to use can be difficult. As Baru et al. (2013) put it, “New platforms with increasingly more features for managing big datasets are being announced almost on a weekly basis. Yet, there is currently a lack of any means of comparability among such platforms.” For this reason, Baru et al. have supported the “Big Data 100 List,” which seeks to find ways to benchmark big data platforms. Academics have an even tougher job, for they must find an appropriate platform to use to perform analytics on huge datasets, but it must also be very inexpensive or free, easy to use, and accessible by many students. The importance of minimizing the time that the instructor must allot to helping the students get up and running on the system should be underscored. Professors are too time-constrained with other duties to enable them to spend much time troubleshooting technical issues.

How Can Universities Work with Companies?

To further enhance these analytics programs and curriculum, many universities are working together with data companies to develop Big Data curricula. These companies can help to keep universities informed on the latest techniques by pointing them towards systems and methods that might not have appeared yet in academic literature, and also by sharing training manuals and tutorials that the companies use to train developers or their own employees. This is in the best interest of the companies, for they can gain converts to their particular methods and software, but it can also be in the best interest of the students, who can conceivably use software experience to help gain technical understanding, improve performance in interviews, and lessen the time it takes for them to learn job skills once they are employed.

Companies can also band together to help support Big Data education in academia today. By pooling resources, they can help support foundations and institutions to learn the best ways to efficiently inform people about Big Data analysis. They can also sponsor research that will enable academic institutions to make better decisions as to how to best educate students about Big Data. This type of collaboration will, we hope, become more frequently seen in the years ahead.

In a company-academic partnership, it is imperative that companies work closely with professors. Companies can help professors refine their own technical abilities and can save professors hours of troubleshooting time. Since Big Data skills are not yet widespread among researchers and companies can act as valuable resources and colleagues for professors who are often teaching themselves about Big Data. One caveat is that the academics should always stay in control of the education process itself, to ensure focus on student learning and making sure materials are well suited for an academic setting.

Our Approach

In an effort to enhance the Big Data course offerings at Georgetown University, we decided to introduce a three-day module on Big Data in the fall of 2012 as part of a pre-existing undergraduate database class (OPIM 257) at Georgetown’s McDonough School of Business. Plan-
ning for this class began in August 2012, three months before the scheduled class sessions in mid-November. We initially learned the basics of Hadoop and HDFS using self-guided modules at Big Data University (bigdatauuniversity.com), and IBM’s Academic Initiative site (www-03.ibm.com/ibm/university/academic/pub/page/academic_initiative).

We decided to largely work with IBM InfoSphere BigInsights, which involved command-line tools. We set up three instances on the IBM Academic Skills Cloud that included a basic installation of Hadoop. The labs we created had the students connect to one of the instances, start Hadoop and its associated services, run a simple word count MapReduce job on Twitter data we obtained from Infochimps, and review/interpret the results of the job.

We used the first day of the module to discuss Big Data terminologies and case studies, and also introduce some technical information about MapReduce and other topics. The next two class periods were hands-on sessions where we had students run the aforementioned MapReduce jobs. This seemed in alignment with our goals for the class, which were to introduce students to the concepts of Big Data and to give them hands-on experience using Big Data technologies.

We decided to conduct pre- and post-tests to gauge students’ experience prior to entering the Big Data module and to obtain a measure of how our efforts were received. The results of the pre-tests for the undergraduate class show that students began the module having heard of Big Data in the news, but having little idea of what it was. After the module (which took up three class periods), they had a better idea of what it was, how it might be useful and also expressed a desire to learn more. A few of them said they would like to pursue a job in the area.

In summer 2013, we taught an Intensive Learning Experience (ILE) to MBAs for a week. This was a module that counted as a half-semester course (1.5 credits). The class included guest speakers, lectures, discussions, and hands-on exercises, which served to introduce the students to the definitions and language of Big Data. In this session, we also used IBM’s Infosphere Big Insights to run through some labs, achieving our goal of giving students hands-on experience with Big Data technologies.

For both of these modules, we relied upon freely available articles and e-books. The e-books used were Understanding Big Data (IBM, 2012) and Harnessing the Power of Big Data (IBM, 2013).

Lessons Learned

The first takeaway from our experience is that students seeking an education in Big Data will often come from a wide variety of backgrounds and experiences. In our MBA class, for example, we had a student with a Ph.D. in mechanical engineering, someone who had virtually no understanding of databases, and many students whose experience was somewhere in between these two extremes. We used a survey to assess technical knowledge before the class began, but recognized later that we should have reviewed some of the basic technical database principles during the course introduction. This would have helped ensure that students had a solid foundation before getting into the rest of the course material. Students could also have been asked to complete self-guided tutorials ahead of time to help build the foundation. And we learned that if there is a mix of students with varying ability levels, the instructor should consider grouping those with more technical experience with those who have less during group activities to better balance the skill sets.

Second, because we were teaching a business class, we realized that we should have reviewed topics that were more business related as opposed to spending time on more technical principles. There is both a challenge and an opportunity to housing Big Data courses in the business school. The challenge is to find the right balance between teaching about how you actually get hands-on with Big Data, which requires statistical, computer science, and data modeling skills, and teaching about how managers should ask the appropriate questions and be able to understand the answers they receive at a more conceptual level.

At the outset, we felt a responsibility to give students at least some familiarity with the technical aspects of Big Data systems so they would have a better understanding of how to use them. But, based on feedback, we need to revisit the balance that we struck between our desire to give hands-on understanding and an approach geared to the managerial decision-making level, particularly with the MBA-level class. Focusing on the business problems that need to be solved would allow students to think of ways to use Big Data to produce better outcomes. This could help to ensure that students understand the managerial implications and how Big Data can be used as a means to make better decisions. The technology will be constantly evolving, and such an approach would better prepare students to understand the basic principles for using data for decision making.

The third key recommendation, and one that is closely related to the second, is that the professor should focus course material on specific learning objectives related to Big Data. The term can mean many different things to different people, so it is important to narrow the focus for the course and clearly state the learning objectives so students’ expectations are properly set. Of course, this is true of any course topic, but it is especially true for Big Data, where there are such a wide variety of applications.

Fourth and last, the development of Big Data curricula should align with the type of careers the students are pursuing. One goal of education is to prepare students (at least to some extent) for the jobs they will take, and jobs in Big Data may be quite diverse. McKinsey did a taxonomy of career paths involving Big Data, in which they note three main categories: deep analytical, big data savvy, and supporting technology. As shown in the table below, deep analytical skills include those math, statistical, and IT skills needed for jobs as database analysts, actuaries, statisticians and the like. Big Data savvy skills are those needed by business managers, analysts, and others who need to understand how to make decisions using Big Data, but from a higher level,
decision-making perspective. Supporting technology skills are needed by those who make Big Data systems work, so results can be produced. These persons include the computer science programmers and staff to help collect, store, and develop systems that others will use to analyze Big Data and use it for decision making. In the same way instructors are well-advised to look at students’ experience coming into the classroom; they are also well-advised to consider their future career goals. This will help professors better focus the material for them and help start them down an appropriate career path related to Big Data.

**Recommendations for the Future**

In this last section, we want to address the way we would ideally like to be able to teach a class on Big Data.

To begin with, we would like the implementation of a system to show students how to work with Big Data to be simple and clear. Big Data software is rapidly developing, in terms of more intuitive graphic user interfaces and ease of implementation, but it is still lacking in some areas. Small companies keep emerging to help satisfy the demand for new, more facile analytic tools. Some of these are acquired quickly by larger data companies trying to update the ways analysts can work with data. However, the process of getting a class to experience hands-on labs with Big Data is still nowhere near as easy to implement as it should be. Ideally, the process should involve a simple login procedure, a well-structured process to connect to streaming data or large datasets, an intuitive graphic user interface, and clear tutorials and instructions. This process should be straightforward and able to be completed in a reasonable amount of time by a class of students, with little technical support needed from the professor.

For those students who are more technically advanced, they should have opportunities to go further with the system on their own, either following online tutorials or exploring on their own. These more advanced capabilities should be clearly explained, with examples and step-by-step instructions.

Companies that produce tutorials tend to err on the side of producing point-and-click labs versus labs with robust explanations for why steps are being taken and how those steps fit the big picture of data analysis. Ideally, if academics are going to make use of these labs, they need to contain more explanations about the purpose or point of each step taken. Students need to have a deeper understanding for how to apply these tools so that when they are better understood, students can use them more intelligently and in more varying and complex ways.

The field of Big Data is growing and changing rapidly. Academics need to be prepared to incorporate these changes into their classes, so that students are better prepared for their future careers. Thus coursework needs to be designed to allow for flexibility as tools for Big Data evolve. Additionally, coursework should provide students with frameworks that will allow for problem solving using Big Data, regardless of information systems architecture. This will help them be better prepared for a variety of positions in business, government or nonprofits.

**References**


---

### Deep Analytical, Big Data Savvy, Supporting Technology

| Database administrators and programmers | Business and functional managers | Computer and information scientists |
| Actuaries | Budget, credit, and financial analysts | Computer programmers |
| Mathematicians | Engineers | Computer software engineers for applications |
| Operations research engineers | Life scientists | Computer software engineers for systems software |
| Epidemiologists | Market research analysts | Computer systems analysts |
| Economists | Survey researchers | Database administrators |
| Industrial organizational psychologists | | |

---

**DECISION LINE**

JANUARY 2014
Appendix A. Top Programs for Big Data Analytics Master’s Degrees.

<table>
<thead>
<tr>
<th>College or University</th>
<th>Degree</th>
<th>School</th>
<th>Length of Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bentley</td>
<td>MS in Marketing Analytics</td>
<td>School of Business</td>
<td>1 to 1.5 years</td>
</tr>
<tr>
<td>Carnegie Mellon</td>
<td>Masters of Information Systems Management, Concentrations in Business Intelligence and Analytics</td>
<td>Heinz College (Public Policy &amp; Information Systems)</td>
<td>1 year Masters of Information Systems Mgt, 16 mos. for concentration in BI &amp; Analytics</td>
</tr>
<tr>
<td>Columbia University</td>
<td>MS in Computer Science, Concentration in Machine Learning</td>
<td>The Fu Foundation School of Engineering and Applied Science</td>
<td>2 years</td>
</tr>
<tr>
<td>DePaul University</td>
<td>MS in Predictive Analytics</td>
<td>College of Computing and Digital Media</td>
<td>2 years</td>
</tr>
<tr>
<td>Drexel University</td>
<td>MS in Business Analytics</td>
<td>College of Computing and Digital Media</td>
<td>2 years</td>
</tr>
<tr>
<td>Harvard University</td>
<td>MS in Computational Science and Engineering</td>
<td>School of Engineering and Applied Sciences</td>
<td>1 year</td>
</tr>
<tr>
<td>Louisiana State Univ.</td>
<td>MS in Analytics</td>
<td>College of Business</td>
<td>12 months</td>
</tr>
<tr>
<td>MIT</td>
<td>MBA</td>
<td>Sloan School of Mgt.</td>
<td>2 years</td>
</tr>
<tr>
<td>New York University</td>
<td>MBA in Bus. Analytics</td>
<td>Stern School of Business</td>
<td>2 years</td>
</tr>
<tr>
<td>NC State University</td>
<td>MBA in Analytics</td>
<td>Institute for Advanced Analytics</td>
<td>10 mos.</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>MS in Analytics</td>
<td>McCormick School of Engineering and Applied Science</td>
<td>15 mos.</td>
</tr>
<tr>
<td>Rutgers University</td>
<td>Masters of Business and MS in Operations Research</td>
<td>Professional Science Masters Program</td>
<td>1.5 to 2 years</td>
</tr>
<tr>
<td>Stanford University</td>
<td>MS in Computer Science, Information Management and Analytics</td>
<td>School of Engineering, Computer Science Dept.</td>
<td>2 years</td>
</tr>
<tr>
<td>University of California</td>
<td>Masters of Engineering, Interdisciplinary Program</td>
<td>College of Engineering, Electrical Engineering and Computer Sciences</td>
<td>10 mos.</td>
</tr>
<tr>
<td>University of Cincinnati</td>
<td>MS in Business Analytics</td>
<td>Lindner School of Bus.</td>
<td>9 mos. - 1 year</td>
</tr>
<tr>
<td>Univ. of Connecticut</td>
<td>MS in Business Analytics and Project Management</td>
<td>School of Business</td>
<td>1 year</td>
</tr>
<tr>
<td>University of Illinois</td>
<td>MS in Statistics, Analytics</td>
<td>Graduate College</td>
<td>3 semesters</td>
</tr>
<tr>
<td>University of Ottawa</td>
<td>Masters in Electronic Business Technologies</td>
<td>Telfer School of Mgt, Information Technology &amp; Engineering, Faculty of Law</td>
<td>1 year MA in EBT, 16 mos MS in EBT</td>
</tr>
<tr>
<td>University of Tennessee</td>
<td>MS in Business Analytics</td>
<td>College of Business Admin, Statistics</td>
<td>2 years</td>
</tr>
<tr>
<td>York University, Toronto</td>
<td>MS in Business Analytics</td>
<td>Schulich School of Business</td>
<td>1 year</td>
</tr>
</tbody>
</table>