Big data and analytics programs in the US: A pedagogical analysis

Someswar Kesh
University of Central Missouri

Prasad Rudramuniyaiah
University of Central Missouri

Sam Ramanujan
University of Central Missouri

ABSTRACT

Recent surveys and studies have consistently projected a growing demand for IT professionals with big data, business analytics, and machine-learning skills. This demand is expected to further increase with more organizations implementing business analytics efforts within their organizations. To fulfill the expected shortage of skilled professionals, many institutions of higher learning have developed degree programs at both graduate and undergraduate levels. These programs however, are not similar. This paper therefore systematically analyzes the degree programs offered at various levels, the differences between these programs, and the curriculum content of various programs offered within the United States. The findings indicate five broad themes, or areas, of business analytics among these programs. This research contributes to pedagogy in designing new programs and to improve the pedagogical content of current programs.

Keywords: Big Data, Predictive Analytics, Machine Learning, Data Science

INTRODUCTION

Data is considered to be an important asset for organizations (Qu and Jiang, 2019; Goes, 2014) and organizations expect to realize significant impacts by analyzing big data (Rai, 2016). Big Data is defined using the three vectors: volume, velocity and variety, commonly referred to as the 3Vs. Due to the increased awareness of the importance of big data, organizations are generating and capturing large volumes of data for analysis. The analysis of big data however, requires the use of several techniques and tools due to the complexity of big data (Baker et al, 2019; Powers, 2014; Katal et al, 2013). Rapid developments in these areas has spawned the entire discipline of Big Data Analytics, having applications in a wide variety of business and engineering fields. In marketing, for example, it has been used to predict consumer choice, target and find new customers, position products etc. (Miller 2015). In healthcare, it has been applied in clinical operations, evidence-based medicine, patient profile analytics etc. (Raghupathi & Raghupathi 2014). It has also been applied to logistics and supply chain management (Wang et.al 2016) resulting in increased efficiencies.

Due to the diverse set of application areas, the demand for experts in big data and analytics has grown significantly in recent years. A search using the popular search engine Google using the term “demand for big data professionals” returned close to 8,000 results with virtually every result expecting more than a 25% growth in demand. The 2019 World Economic Forum Report predicts that data analytics will be the “fastest growing profession”. Statistics compiled by IBM predicts that around 700,000 new jobs will be created in this area during 2020 and Morning Future predicts that a degree in Big Data Analytics is “the degree of the future” (Morning Future 2018). Educational institutions, including universities have responded to the needs of industry by creating new undergraduate and graduate degree programs and, offering shorter certification programs as well.

A key problem for an individual interested in pursuing these programs however is to decide which program suits them best considering the numerous academic disciplines (e.g. mathematics, computer science, computer engineering, social sciences and business) grouped under the umbrella term “Big Data Analytics”. An online search for “big data programs in the USA” yielded 875,000,000 results with results including courses from several disciplines, including those mentioned earlier which can be intimidating to and possibly confusing, to both, prospective students and parents.

To address this important issue, this paper systematically analyses term Big Data and Analytics programs in the United States based on several criteria and also provides a thematic structure for developing degree programs in the field of big data analytics. The structure has been developed by analyzing the coursework of current degree programs in the USA, considering the level of these programs: certificate, undergraduate, Master’s and Doctoral programs. In addition to catering the prospective analytics professional, educational institutions interested in developing new programs in Big Data Analytics can use this information to develop their programs or, to make changes to their programs. This paper therefore contributes to pedagogy, academia and practice by providing clarity on these programs.

WHAT IS BIG DATA ANALYTICS

As a field, Big Data makes use of two components, Big Data and Predictive Analytics (including machine learning). The merging of these fields has created a powerhouse of knowledge and information that businesses can use to optimize decision-making. Big Data and Predictive Analytics are first discussed separately, and subsequently the combination of both the fields.

Big Data

It is now widely accepted that the term big data comprises of three components; volume, velocity, and variety (SAS Insights, 2018). Volume signifies the amount of data that generated. For example, credit card processing systems handle millions of transactions per second all around the globe. Credit cards are not the only ones responsible for data generation, there are many other sources of data as well. For example; emails, corporate transactions, also generate massive volumes of data. The combination of data generated from all these sources generates millions of bytes of data per second. This rate is the velocity of data. Finally, the data sources themselves are varied in nature. Data can originate from financial transactions, social media sites etc. Some of the data can be numerical, others textual, and some others multimedia data image, audio and video (from sources like Therefore, hardware and software should be able to deal with all these varied data sources.

Predictive Analytics and Machine Learning

Predictive analytics involves a wide variety of statistical models and empirical methods to create empirical predictions (Shmueli, Koppius 2011) and for discovering patterns. Clustering techniques are prime examples of tools used in discovering patterns. Various other classification methods include linear regression, decision trees and forests, logistic regression etc (Babcock, 2016). Other predictive analytics tools include text analytics and sentiment analytics (Miller 2015).

Machine Learning

Even though there is some overlap between predictive analytics and machine learning, machine learning techniques allow systems to generate new knowledge from data. This includes development of machine learning classifiers, like Decision trees, K-nearest neighbors (Raschka, S., 2015). Figure 1 shows the various components of big data and machine learning tool. These tools and techniques are applied to a wide variety of business and engineering problems.

Programs may select to be in one of the areas of big data skills and knowledge or predictive analytics, machine learning, and AI skills and knowledge or they can be in any of intersection areas. For example, if a program decided to choose the Big Data area only, they might focus on data engineering. However, the choice may be just to focus on predictive analytics and machine learning. This is associated mostly with data science. Domain knowledge defines the application area of the technologies and tools of the other two areas. Big data analytics have been applied to a myriad of domain areas, like information security, criminal justice, marketing decision making, and health care as well.

Programs can also focus on intersection areas. An intersection between big data tools and technologies, and predictive analytics represents an area where students can learn about big data like Hadoop, Apache Spark, NoSQL; as well the architecture of big data databases, and then use the storage and retrieval techniques to feed the data to predictive analytics tools. The intersection between big data and domain knowledge will represent big data applications to a particular domain like health care (Celesti A., et. al 2016). Similarly, the intersection of predictive analytics and machine learning with domain knowledge will be the application of technologies like neural networks or model-based reasoning to domains like healthcare. The intersection of all three areas is probably the most interesting, where big data tools and technologies are used for storage and retrieval of data, machine-learning tools and technologies are then used to analyze the data, and they are applied to a domain. Academic programs may focus on the applications in a single domain or multiple domains. For example, some programs are focused only on the health care industry (Celesti, A., et. al). and other are focused on business (Sirignano and Cont 2019)

ANALYSIS OF CURRENT PROGRAMS

An Internet based search was conducted to identify big data and analytics programs using several websites publishing information on degree programs in the USA (e.g. College Choice) and the ensuing results were analyzed at three criteria related to these programs; the location, level and structure of the program.

Program Location

The program location identifies the college or department that houses the program. Determining the location of the program is important because it drives the nature and orientation of the course content the type of employers and students it may attract. Of the fifty programs that were analyzed by us, sixteen programs were housed in the college of business and five were in the department of Computer Science. There was only one program in the College of Engineering. Two programs were in the college of Information Sciences. Two programs were housed in the department of Statistics, and one in mathematics, and one in the school of professional studies. Some of these programs were the effort of a collaboration between multiple departments. Common collaborative efforts included collaboration between Computer Science and the School of Business; Mathematics and Statistics and Mathematics and Computer Science.

Given that the field of big data analytics has significant business applications, it is no surprise to see that business schools lead in offering degree programs in this field. Also, because of the diversity of the field collaboration between various disciplines is not uncommon.

Structure of the Coursework

In order to analyze the structure of the coursework, an analysis of the coursework of the fifty programs listed under the link was first performed. Most of the programs were at master's level, thereby providing some consistency. The coursework were grouped into themes based on the authors' knowledge of the domain as shown in Table 1. Similarly, the names were also selected by the authors. For example, the theme business/decision making included courses like economic analysis, customer relationship management etc. The numbers in the parenthesis also tells how many programs included that coursework or a very similar coursework.

Theme 1: Business/Decision Making

Many programs offering degrees in big data analytics integrate courses related to business decision making especially if the program is housed in the college of business. Courses with a business orientation included management strategy, economic analysis, and marketing management. For example, customer relationship management and ERP systems courses are included in some programs. Some of the programs had a course in project management as well as courses that integrated supply chain courses and project management. Given its importance to any program, it is not surprising that project management is the most widely represented course in this category.

Theme2: Information Systems Technology

Not surprisingly, the most common information technology course in big data degrees is database management. In some cases, a database management course is included in the perquisites. Related to database management, some programs also include advanced database classes like database warehousing in the program. Courses related to database management was also offered in information retrieval and analysis, which included web search and data mining courses.

Programming languages were the next category for IT courses. Python and R were the most common programming classes. In some cases, Java was also used as a programming language, and for programs with a computer science orientation, data structures and computational mathematics courses were also included. Other IT courses that were included but not common, were courses in data communications and systems analysis and design. For programs with an IT orientation, IT strategy courses were included. Programs offered by the computer science department sometimes included theoretical courses like computational mathematics.

Theme 3: Big Data Tools and Technologies

While there may not be a general agreement on big data tools and technologies and some of them overlap and are integrated with predictive analytics tools, there are some tools that are more widely used than others. The most common group of big data tools that is used is the Apache Hadoop ecosystem. The other most popular tool is Apache Spark. Some however consider Apache Spark to be part of the Hadoop ecosystem, even though it has carved a niche of its own. Other big data tools that are popular in the market are, MongoDB, Hive, Hbase, Cassandra, Kafka (AcadGild, 2018) and various NoSQL databases.

Theme 4: Predictive Analytics and Statistics

All programs, irrespective of the level or location, have courses in statistics and predictive analytics. Courses in statistics were in time series analysis and regression analysis. Courses in predictive analysis were offered under different names, like business intelligence, which covered SAP. Some predictive analytics courses were also domain specific. For example, a predictive analytics course could be focused on marketing. Other examples of such specialized courses would be the use of analytics in government.

Theme 5: Machine and Deep Learning and Artificial Intelligence

There is less emphasis on the broader field of artificial intelligence but greater emphasis on both machine and deep learning. Machine learning courses on both supervised and unsupervised learning. Examples of supervised learning include various linear models, decision trees and vector machines. Various unsupervised learning algorithms taught include dimensionality reduction like principal component analysis and clustering. Various other topics covered in unsupervised learning algorithms include how to evaluate and improve models, develop pipelines and grid searches. (Muller and Guido, 2016). Deep learning courses include the mathematical building blocks of neural networks, including data representation of scalars, and tensors, the various building blocks of neural networks, and applications of deep learning to various fields like computer vision and applications to texts and sequences (Chollet, 2018).

Theme 6: Other Skills

Successful professionals need a wide variety of personal skills like professionalism, communication and leadership skills. Many schools ensure that these skills are developed, while some others do not. A critical element of successful professionals is to develop research skills, both from a technological viewpoint as well as a business integration viewpoint. Depending on the program itself, various mathematical skills are also deemed important. Some have considered advanced mathematical skills while others have considered elementary mathematical knowledge.

CONCLUSIONS

This paper has explored the structure of the programs in big data analytics in the US. The research has shown that the focus of the coursework is more on analytics and machine learning. Many of these programs were in the business school, and therefore added some business component to their program or made the courses business oriented. Irrespective of where the degree program was located, or the focus of the program (more business oriented or more technical), there was wide agreement about the content of the degree programs. The fundamentals of database management systems and data warehousing, programming in Python and R, predictive analytic tools and techniques, AI, deep and machine learning tools and techniques were all common to almost all the programs. Another finding was that the emphasis was greater on the analytics skills rather than the big data engineering skills. This is an area that many institutions may offer courses in. A major weakness identified is that there is a significant demand for data engineering skills but, there wasn’t any program that addressed this significant need.

It is hoped that this paper will provide a roadmap for institutes of higher education considering developing degree programs in big data analytics or considering making changes to their current programs. For future research, the authors hope to correlate the findings of this study with requirements for employers by analyzing job advertisements in this field and using qualitative research methodologies like content analysis to validate the work done.

References:

A. Katal, M. Wazid, R. H. Goudar, (2013), "Big data: Issues challenges tools and good practices", Proc. Int. Conf. Contemp. Computing., pp. 404-409.

AcadGild (2018), 7 Trending Big Data Tools and Technologies. Retrieved Jan 30, 2020: https://acadgild.com/blog/7-trending-big-data-tools-technologies

Babcock, J. *Mastering Predictive Analytics with Python*, , 2016, Packt Publishing Ltd., UK.

Celesti, A., et. al, "A Hospital Cloud Based Archival Information System for the Efficient Management of Big Data", 39th International Convention on Information and Communication Technology, Electronics, and Microelectronics. (page numbers not given).

Chollet, F., "*Deep Learning with Python*", 2016 Manning Publications, NY.

Goes, P. 2014. "Editor's Comments: Big Data and IS Research," MIS Quarterly (38:3), pp. iii-viii.

College Choice 2018: Best Big Data Degrees, Retrieved Jan 30, 2020: https://www.collegechoice.net/rankings/best-big-data-degrees/

Miller, Thomas, W. *Marketing Data Science, Modeling Techniques in Predictive Analytics with R and Python ,* 2015, Pearson Education Inc., New Jersey.

Miller, Thomas, W. *Modeling Techniques in Predictive Analytics with Python and R A Guide to Data Science*, Pearson Education Inc., New Jersey.

Morning Future Newsroom (2018): Data Analyst, the most in-demand job of the coming years Retrieved Jan 30, 2020: https://www.morningfuture.com/en/article/2018/02/21/data-analyst-data-scientist-big-data-work/235/

Power, D.J. (2014). Using 'Big Data' for analytics and decision support. Journal of Decision Systems, 23, 222-228.

Qu, X., and Jiang, Z. (2019), "A Time-based Dynamic Synchronization Policy for Consolidated Database Systems," MIS Quarterly, 43:4, pp. 1041-1057

Raghupathi, W. and Raghupathi, V. *Big Data Analytics in Healthcare: Promise and Potential, Health Information Science and Systems*, Vol. 2:3, 2014.

Rai, A. (2016), "Synergies Between Big Data and Theory", MIS Quarterly, 40:2, pp. iii-ix

Raschka, S., Python Machine Learning, 2015, Packt Publishing Limited, UK.

SAS Insights: Big Data: What it is and why it matters, Retrieved Jan 30, 2020, from: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Shmueli, G., and Koppius, O.R., *Predictive Analytics in Information Systems Research*, MIS Quarterly, Vol. 35, No. 3, pp. 553-572, September 2011.

Sirignano, J. and Cont, R., "Universal Features of Price Formation in Financial Markets: Perspectives From Deep Learning",

Wang et. al. *Big Data Analytics in Logistics and Supply Chain Management*, Certain Investigations for Research and Applications, Vol. 176, June 2016, pp. 98-110.

APPENDIX

Table 1: Themes for Big Data Analytics

Theme	Theme Area	Skills
1	Business/Decision Making	1. Management Strategy 2. Economic Analysis 3. Marketing Management (including CRM) 4. Project Management 5. Supply Chain Analytics
2	Information Systems and Technology	1. Database management Systems 2. Advanced Database Technologies 3. Information Retrieval and Analysis 4. Programming Languages (Python and R) 5. Data Communications 6. Systems Analysis and Design 7. Data Structures 8. Computational Mathematics
3	Big Data Tools and Technologies	1. Hadoop 2. Apache Spark 3. Hive 4. PIG 5. mongoDB 6. Cassandra 7. Kafka 8. Various noSQL databases
4	Predictive Analytics and Statistics	1. Exploratory Data Analysis 2. Data Visualization 3. Clustering and Unsupervised Learning 4. Classification Methods
5	Machine and Deep Learning and Artificial Intelligence	1. Data Mining 2. Artificial Intelligence Concepts and Applications 3. Machine Learning 4. Deep Learning using Neural Networks and Bayesian Statistics 5. Deep Learning Applications in Computer Vision, Text Sequencing etc.
6	Other Skills	1. Personal skills 2. Research Skills 3. Mathematical skills

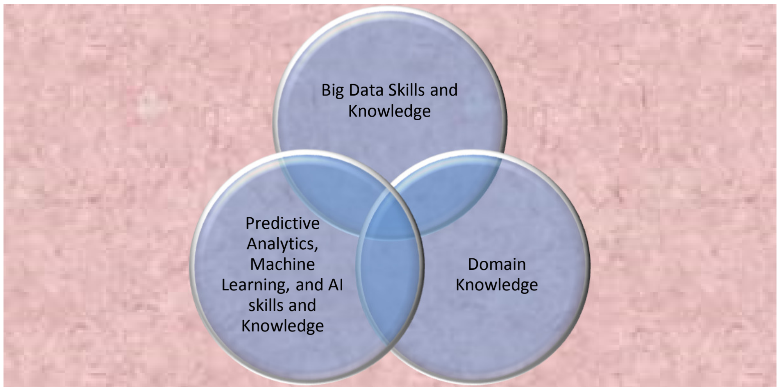
Big Data Skills and Knowledge
Predictive Analytics, Machine Learning, and AI skills and Knowledge
Domain Knowledge

Figure 1: Components of Big Data and Predictive Analytics Programs