BIG DATA ANALYTICS TOOLS: A BIBLIOMETRIC LITERATURE REVIEW

by

Holden Jones

A Senior Honors Project Presented to the

Honors College

East Carolina University

In Partial Fulfillment of the

Requirements for

Graduation with Honors

by

Holden Jones

Greenville, NC

May 2016

Approved by:

Dr. John Drake

Dept. of Management Information Systems

Abstract

Analytics as a field is rapidly growing because of businesses need to tackle increasingly large and complex datasets to gain a competitive edge in the business environment. Currently there are two distinct types of analytics technologies: proprietary and open-source. Each has their distinct strengths and weaknesses, but which is relevant in today business environment? Questions like this are important for businesses wanting utilize these technologies in order to become larger and more efficient. This research answered this question and it was found that both proprietary and open-source technologies are equally relevant in current analytics research. Because of this, businesses should be more aware of the analytics field and how both types of technologies could benefit their current operations and should strategically utilize both types to meet their specific data needs.

Table of Contents

Introduction

As a field, Analytics is rapidly expanding to meet the demands of the business sector. Because of this growth, the McKinsey Global Institute reported that there will be a shortage of 140,000 to 190,000 analytics professionals (Chen, Chiang, Storey 1165). Research from SAS and Bloomberg showed that 97% of businesses with a revenue greater than or equal to $100 million use analytics to make business decisions (Bloomberg 2). Because of such a high percentage, it would seem that analytics tools are becoming a norm in today's business practices and decision making.

This growth is also creating competition within the technology industry. Technology companies are quickly expanding their current business models to include their own proprietary analytics options (Liberatore and Luo, 313). By doing this, they hope to gain a competitive advantage within this upcoming field. Increasingly, the data collected in businesses is becoming larger and more complex. Because of these changes, new open-source technologies are also being created to manage and facilitate the complex data analysis (Batrinca and Treleaven 102-103). In order to gain a foothold in the market, companies must first understand how the analytics field has changed over time and which technologies available are relevant to today's business environment.

The purpose of this study is to explore what types of analytics technologies have been mentioned within information systems (IS) research over the past decade. In order to meet this objective, four questions must first be addressed.

- Which publications within the IS field are the most relevant?
- What information from journal articles is needed?
- What key-terms related to the analytics field should be used to find relevant articles?
- How should analytics technologies be segmented for this study?

These guiding questions will help create a framework for this study and future researchers should be able to follow this framework or adapt it to meet their specific needs.

Currently, no studies have been completed to understand the types of analytics technologies mentioned in scholarly literature and how the volume of literature about these technologies has

changed over the last decade. Similar studies have been completed on the entire field of analytics which will be a general framework for this study, but this research will go beyond their original methods and look deeper into the analytics field. By completing a bibliometric literature review, some understanding of how analytics technologies and the field is growing could be used in a business setting to make decisions about which analytics technologies are relevant and which are not.

Background

Within the business world, phrases like "dig data" and "data science" are becoming increasingly common as businesses in general are becoming more reliant on technology. These expressions go hand-in-hand with analytics or business intelligence. Business analytics is the

> "… the study of data through statistical and operations analysis, the formation of predictive models, application of optimization techniques and the communication of these results to customers, business partners and colleague executives." (Leonard)

Although the use of these terms are a relatively recent phenomena, similar techniques have been in business for decades under the name Operations Research (Liberatore and Luo 315). According to Liberatore and Luo, business analytics and operations research share many similarities but there are distinct drivers that are fueling this newer movement into analytics (315). These four drivers include the rapidly increasing amount of data stored within businesses, the development of advanced software packages designed for analytical analysis, more structured business processes, and executives who embrace advanced technology and expect its use within the business environment (Liberatore and Luo 315-316).

Within analytics, the skills desired usually include the understanding of how to collect, store, clean, analyze, and model the data (Chen, Chiang, Storey 1166). Most of these practices can be done using many different software packages. Currently, there are two distinct paths when it comes to

analytics software: enterprise and open-source. The most common enterprise analytics software packages used by industry leaders are created by SAS (Bloomberg 2). SAS has been creating these software packages for four decades and is constantly innovating to meet the ever-changing demands of customers (SAS). The second path, open-source analytics solutions, is a much newer alternative when it comes to the analytics world, but have their own distinct advantages in the business environment.

Because of growing, changing, and more complex data needs, open-source software platforms are also being used in business (Batrinca and Treleaven 102). Examples of where open-source analytics are being utilized include network, security, and text analytics. Many of these advancements have come with the growth of social network sites and the vast amounts of data they are constantly producing (Batrinca and Treleaven 102). Some common tools created for this social media realm include noSQL databases like Cassandra/hive, distributed data computing frameworks like Hadoop, and programs like Mahout that facilitate the use of complex analytics algorithms for this data mining (Batrinca and Treleaven 102-103). All of these elements allow social media companies like Facebook and Twitter to analyze categories of data that were impossible with traditional analytics (Batrinca and Treleaven 89). Another advantage these open-source technologies have on enterprise analytics packages is their cost. The programs and platforms themselves are free. Because of this, the businesses utilizing these technologies are avoiding the expensive licenses associated with packaged analytics products.

Currently, there is a lack of physical information related to businesses use of specific types of analytics technologies. This creates the need to analyze this area of the field and serves as the motivation for this study. This bibliometric study will give a view of the amount of academic literature being produced about analytics and what technology types, proprietary or open-source, are relevant. Hopefully, this will identify a clear leader within field and can provide meaningful knowledge to businesses about relevant technologies. Details of this study are contained in Research Methodology.

Research Methodology

The methodology for this literature review will closely mirror the techniques used by Chen, Chiang, and Storey in "Business Intelligence and Analytics: From Big Data To Big Impact" (1179). For this study, eight high-impact publications were chosen to represent the IS field. Because analytics spans so many disciplines, there are many academic journals with publications containing information about the subject. This entire collection of journals would be impractical to use for this study because of the amount of information that would be needed to collect and the limited time frame available to compete this study. For these reasons, the list of publications will come from the top Information Systems (IS) journals according to the article in MIS Quarterly by Chen, Chiang, and Storey (1181). They were as follows:

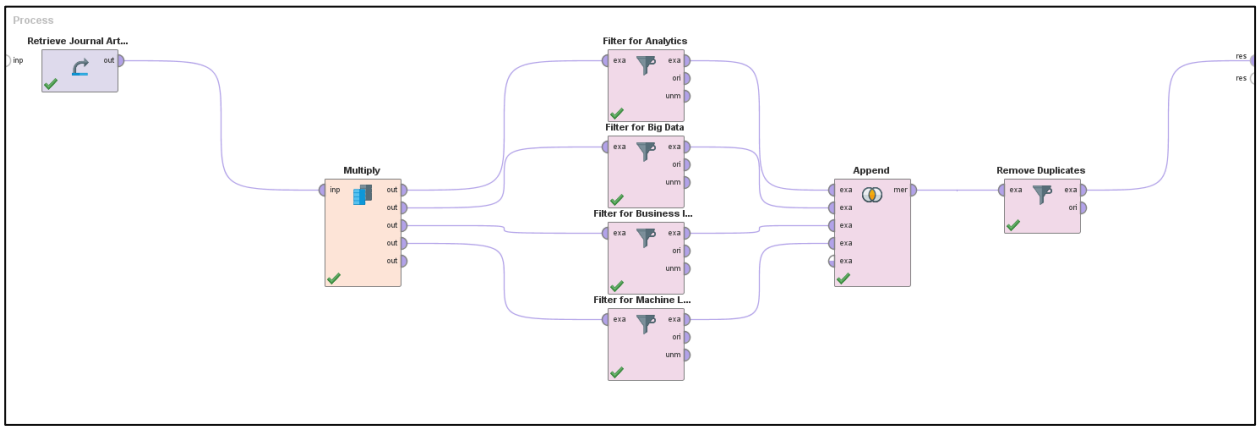
- Communications of the Association for Information Systems
- Decision Sciences
- Decision Support Systems
- Information Systems Research
- Journal of Management Information Systems
- Journal of the Association for Information Systems
- Management Science
- MIS Quarterly

The metadata for all peer-reviewed articles contained within these journals was retrieved from the ABI/INFORM academic database. Because this study was limited to information from the las decade, only publications from the year 2005–2015 were included. The metadata retrieved was then cleaned, formatted, and combined in Microsoft Excel to contain article Title, Abstract, Authors, Publication Title, Publication Year, and Subject Terms. These fields contained the information needed to find the articles relevant to the analytics research. In total, there were 3,811 total journal articles from these publications over the last decade.

In order to find articles related to analytics from the metadata, Chen et al. chose to screen the articles' title, abstract, and subject terms for specific key words that related to the analytics field. The

same was done for this study using the following key words: analytics, big data, business intelligence, and machine learning. This filtering process was completed using an open-source analytics platform called RapidMiner with the Text Processing extension installed. Figure 1 shows the process model for this filtering as seen on the RapisMiner GUI.

Figure 1



Running this process produced 51 journal articles that were relevant to the analytics field as defined by my keywords. These articles were them hand reviewed and categorized into four groups by publication date. These four categories were as follows:

- Proprietary Analytics Technologies
- Open-source Analytics Technologies
- Both (contained information about both proprietary and open-source technologies)
- None/Not Specified (research didn't apply to specific technologies or was not specified)

By separating the articles in this fashion, one can understand how literature about analytics has changed over the past decade and gives a glimpse of the types of technologies that are mentioned or involved with this research.

Results

Before analyzing each of the 51 articles for content, they were broken down by publication title. Interestingly, a single title contained over 50% of relevant analytics articles. Because of this, Decision Support Systems would be a good place for businesses to start if they wanted to gain an understanding of the current analytics environment. Decision Support Systems was the clear leader in regards to publishing content about analytics. The closest alternative was the journal Management Science with 7 relevant articles or 13.73%.

Table 1 shows the breakdown of relevant journal articles by publication title.

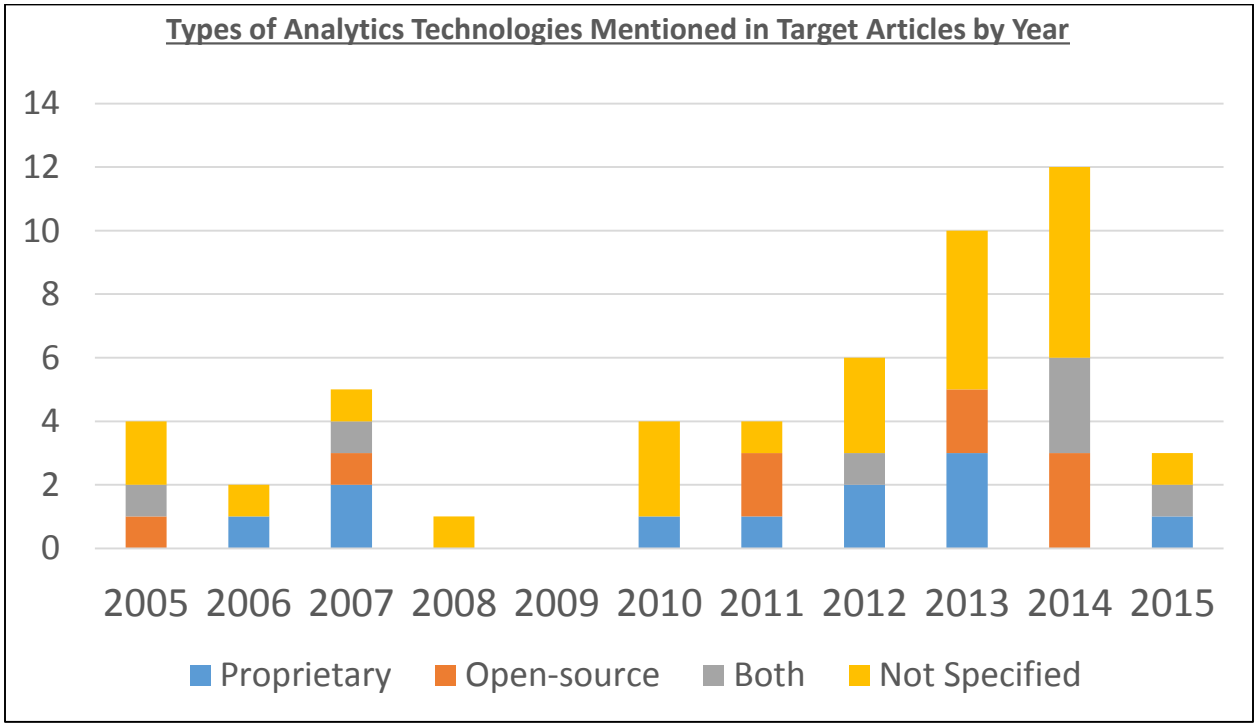
Table 1

Publication Title	Articles Count	Percentage
Communications of the AIS	0	0.00%
Decision Sciences	0	0.00%
Decision Support Systems	27	52.94%
Information Systems Research	6	11.76%
Journal of MIS	5	9.80%
Journal of the AIS	3	5.88%
Management Science	7	13.73%
MIS Quarterly	3	5.88%
Total	51	100.00%

As shown below, proprietary and open-source analytics tools are roughly equally popular in current IS research. More surprisingly, a large portion of the journal articles did not specify specific analytics technologies. This was because many of the "not specified" articles were using analytics to create frameworks for many different business areas and applications. These frameworks were process models for how do complete specific tasks, but were not specific to any one type of technology.

Figure 2 shows the breakdown of relevant journal articles by analytics technology and year.

Figure 2



Out of 3,811 journal articles, only 51 were found to be related to analytics. This is a very small percentage, but the field is relatively new. As seen in Figure 2, there is an obvious growth over the past decade. If the publication list were or filtering words used were expanded, it is expected that this trend would become even clearer.

Conclusions

Over the past decade, analytics research has shown an upward trend in the number of articles published. According to this sample, there is approximately the same volume of literature published about proprietary and open-source analytics technologies. Because of this, businesses need to be cognizant of both types of technologies and how they can benefit their business as a whole. A good way for these businesses to begin understanding this changing environment would be subscribing to Decision Support Systems. Being the clear leader in regards to publishing works related to analytics, Decision Support Systems could provide businesses with relevant research related to the field and give them an idea of how the field is evolving over time.

Going into this study, it was expected that proprietary options would be much more prevalent in IS literature because of their strong presence in business, but this shows that open-source technologies are just as important to analytics research and the field as a whole. Because of the growth in popularity of open-source tools, proprietary analytics software producers like SAS and IBM are designing software that is compatible with open-source options like Hadoop and R. Businesses need to learn about each of these technology types and how they each can meet their business needs. No one technology can meet all of the analytics needs within very large corporations. Because of this, the technologies used in a specific business need to be strategically chosen to need their specific needs, but while still being compatible.

This study provided a glimpse of the current field of analytics. Going forward, this study could adapted and further refined to produce more meaningful insight into IS publications and the analytics field.

Limitations

Limitations of this study include:

- Relevant articles could have been omitted because the filtering keywords used
- Limited scope of Publication field (IS field)
- Publication from conferences and Industry periodicals were not included

Appendix

List of relevant analytics article used in this study

Title	**Authors**
A complexity theory approach to IT-enabled services (IESs) and service innovation: Business analytics as an illustration of IES	Chae, Bongsug
A Machine Learning Approach to Improving Dynamic Decision Making	Meyer, Georg G.; Adomavicius, Gediminas G.; Johnson, Paul E.; Elidrisi, Mohamed M.; Rush, William A; Sperl-Hillen, JoAnn M; O'Connor, Patrick J
A perspective on applications of in-memory analytics in supply chain management	Hahn, GJ; Packowski, J J.
A relative patterns discovery for enhancing outlier detection in categorical data	Pai, Hao-Ting H.; Wu, Fan; Hsueh, Pei-Yun Sabrina
A social network-empowered research analytics framework for project selection	Silva, Thushari; Guo, Zhiling; Ma, Jian; Jiang, Hongbing; Chen, Huaping
A unified foundation for business analytics	Holsapple, Clyde; Lee-Post, Anita; Pakath, Ram
A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration	Chung, Wingyan; Chen, Hsinchun; Nunamaker, Jay F
An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study	Kowalczyk, Martin; Buxmann, Peter
An Empirical Analysis of Digital Music Bundling Strategies	Danaher, Brett; Huang, Yan; Smith, Michael D; Telang, Rahul
Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing	Ryan, Nelson; Todd, Peter A; Wixom, Barbara H
Big Data Investment, Skills, and Firm Value	Tambe, Prasanna
Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research	Agarwal, Ritu; Dhar, Vasant

Business Intelligence in Blogs: Understanding Consumer Interactions and Communities	Chau, Michael; Xu, Jennifer
Class-Restricted Clustering and Microperturbation for Data Privacy	Li, Xiao-Bai; Sarkar, Sumit
Combining Information Seeking Services into a Meta Supply Chain of Facts *	Roussinov, Dmitri; Chau, Michael
Constraint-centric workflow change analytics	Wang, Harry Jiannan; Zhao, J Leon
Data quality: Setting organizational policies	Storey, Veda C; Dewan, Rajiv M; Freimer, Marshall
Data, information and analytics as services	Delen, Dursun; Demirkan, Haluk
Efficient identity matching using static pruning q-gram indexing approach	Khairul, Nizam B; Shahrul, Azman
Estimating Demand for Mobile Applications in the New Economy	Ghose, Anindya; Han, Sang Pil
Formal workflow design analytics using data flow modeling	Sun, Sherry X; Zhao, J Leon
From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data	Zheng, Zhiqiang (Eric); Fader, Peter; Padmanabhan, Balaji
From clicking to consideration: A business intelligence approach to estimating consumers' consideration probabilities	Wang, Hao; Wei, Qiang; Chen, Guoqing
Heuristic Theorizing: Proactively Generating Design Theories	Gregory, Robert Wayne; Muntermann, Jan
How To Build Enterprise Data Models To Achieve Compliance To Standards Or Regulatory Requirements (and share data)	Kim, Henry M; Fox, Mark S; Sengupta, Arijit
Integrated decision support systems: A data warehousing perspective	March, Salvatore T; Hevner, Alan R
IT Capabilities, Process-Oriented Dynamic Capabilities, and Firm Financial Performance*	Kim, Gimun; Shin, Bongsik; Kim, Kyung Kyu; Lee, Ho Geun

IT innovation adoption by enterprises: Knowledge discovery through text analytics	Basole, Rahul C; Seuss, David C; Rouse, William B
Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming	Cui, Geng; Wong, Man Leung; Wong , Hon-Kwong Lui
Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model	Lee, Hyeseon; Cho, Hyunbo; Im, Kwanyoung; Kim, Yong Seog
Motivational Differences Across Post-Acceptance Information System Usage Behaviors: An Investigation in the Business Intelligence Systems Context	Li, Xixi; Hsieh, J J Po-An; Rai, Arun
Multitask Agency, Modular Architecture, and Task Disaggregation in SaaS	Susarla, Anjana; Barua, Anitesh; Whinston, Andrew B
Newsmap: a knowledge map for online news	Thian-Huat, Ong; Chen, Hsinchun; Wai-ki, Sung Zhu, Bin
Predicting crime using Twitter and kernel density estimation	Gerber, Matthew S
Predicting Web Page Status	Pant, Gautam; Srinivasan, Padmini
Pricing and Resource Allocation in Caching Services with Multiple Levels of Quality of Service	Hosanagar, Kartik; Ramayya, Krishnan; Chuang, John; Choudhary, Vidyanand
Secure federation of semantic information services	Fabian, Benjamin; Kunz, Steffen; Müller, Sebastian; Günther, Oliver
SIMBA: A simulator for business education and research	Borrajo, Fernando; Bueno, Yolanda; De Pablo, Isidro; Santos, Begona; Fernandez, Fernando; Garcia, Javier; Sagredo, Ismael
Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis	Lau, Raymond YK; Li, Chunping; Liao, Stephen
Stochastic programming and scenario generation within a simulation framework: An information systems perspective	Di Domenica, Nico; Mitra, Gautam; Valente, Patrick; Birbilis, George

The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence*	Tseng, Frank S; Chou, Annie
The Dynamic Structure of Management Support Systems: Theory Development, Research Focus, and Direction	Clark, Thomas D, Jr; Jones, Mary C; Armstrong, Curtis P
The evolution of web-based optimisation: From ASP to e-Services	Valente, Patrick; Mitra, Gautam
The impact of business analytics on supply chain performance	Trkman, Peter; McCormack, Kevin; de Oliveira, Marcos Paulo Valadares; Ladeira, Marcelo Bronzo
The Impact of Cloud Computing: Should the IT Department Be Organized as a Cost Center or a Profit Center?	Choudhary, Vidyanand; Vithayathil, Joseph
Three-Way Complementarities: Performance Pay, Human Resource Analytics, and Information Technology	Aral, Sinan; Brynjolfsson, Erik; Wu, Lynn
Towards business intelligence systems success: Effects of maturity and culture on analytical decision making	Popovic, Ales; Hackney, Ray; Coelho, Pedro Simões; Jaklic, Jurij
Understanding Postadoptive Behaviors in Information Systems Use: A Longitudinal Analysis of System Use Problems in the Business Intelligence Context	Deng, Xuefei (Nancy); Chi, Lei
Understanding the paradigm shift to computational social science in the presence of big data	Chang, Ray M; Kauffman, Robert J; Kwon, YoungOk
Vehicle defect discovery from social media	Abrahams, Alan S; Jiao, Jian; Wang, G Alan; Fan, Weiguo
Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Mergers and Acquisitions	Lau, Raymond Y; Liao, Stephen; Wong, K F; Chiu, Dickson K

References

Batrinca, Bogdan, and Phillip Treleaven. "*Social media analytics: a survey of techniques, tools and platforms,"* AI & SOCIETY 30.1 (2015): 89-116. Web. 12 Dec. 2015.

Bloomberg Businessweek. "*The Current State of Business Analytics: Where Do We Go from Here*?." Bloomberg Businessweek Research Services, 2011. PDF File. 12 Dec. 2015. <http://www.sas.com/resources/asset/busanalyticsstudy_wp_08232011.pdf>

Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." MIS Quarterly 36.4 (2012): 1165. Web. 13 Dec. 2015

Leonard N. Stern College of Business. Masters of Science in Business Analytics: What is Business Analytics. New York University, 2015. Web. 13 Dec. 2015.

Liberatore, M. J., and W. Luo. "The Analytics Movement: Implications for Operations Research." Interfaces 40.4 (2010): 313-324. Web. 12 Dec. 2015.

"*SAS Analytics in Action*." SAS. SAS Institute INC, 2016. Web. 24 April 2016.