BIG DATA ANALYTICS: A TEXT MINING-BASED LITERATURE ANALYSIS

Ahmed Elragal, Department of Computer Science, Electrical and Space Engineering
Luleå Tekniska Universitet, Luleå, Sweden, ahmed.elragal@ltu.se

Moutaz Haddara, Department of Computer Science, Electrical and Space Engineering
Luleå Tekniska Universitet, Luleå, Sweden, moutaz.haddara@ltu.se
Westerdals- Oslo School of Arts, Communication and Technology, Oslo, Norway.

Abstract

This literature review paper summarizes the state-of-the-art research on big data analytics. Due to massive amount of data exchanged everyday and the increased need for better data-based decision, businesses nowadays are looking for ways to efficiently manage, and optimize these huge datasets. Moreover, because of globalization, partnerships, value networks, emergence of social networks, and the huge information flow across and within enterprises, more and more businesses are interested in utilizing big data analytics. The main focus of this paper is to elucidate knowledge on the characteristics of big data analytics literature as well as explore the areas that lack sufficient research within the big data analytics domain, suggest future research avenues, as well as, present the current research findings that could aid practitioners, researchers, and vendors when embarking on big data analytics projects. Towards that end, we have reviewed 24 publications between 2010 and 2014. Results of text mining the papers revealed that they belong to three clusters with both common as well as distinct characteristics. The reviewed papers were clustered into three main themes, 1) technical algorithsms; 2) processing, cloud computing, opportunities & challenges; and 3) performance, prediction, and distributed systems.

Keywords: *big data analytics, text mining, literature review analysis*

1. INTRODUCTION

The interest in big data (analytics) is on the increase, exponentially. Indeed, Google's adoption of the MapReduce was definitely a catalyst, which has led to a lot of developments in the big data arena. Additionally, the development and deployment of Apache Hadoop has also opened the doors for organizations to process extremely large datasets that has never been possible before due to restrictions on traditional DBMS capacities (Agneeswaran, 2012). Slowly but surely, big data is penetrating various sectors e.g., governments, e-commerce, health, retail, insurance, etc. This penetration is supported by the overwhelming amount of data available from different sources e.g., web applications, trajectory data, streaming data, RFID, etc. which build-up at a progressively growing scale (Chen, Chiang, & Storey, 2012).

Big Data Analytics (BDA) is the use of advanced techniques, mostly data mining and statistical, to find (hidden) patterns in (big) data. BDA is where advanced techniques operate on big data sets (Russom, 2011). The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems (Elgendy & Elragal, 2014). A significant amount of these techniques rely on commercial tools such as relational DBMS, data warehousing, ETL, OLAP, and business analytics tools. During the IEEE 2006 International Conference on Data Mining (ICDM), the top-ten data mining algorithms were defined based on expert nominations, citation counts, and a community survey. In order, those algorithms are: C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbours), Naïve Bayes, and CART. They cover classification, clustering, regression, association analysis, and network analysis. Indeed, mostly, they have been incorporated in commercial as well as open source tools. Furthermore, multivariate analysis techniques such as regression, factor analysis, clustering, and discriminant analysis have been associated with many business applications (Chen, Chiang, & Storey, 2012).

As a matter of fact, it is not only organizations and governments that generate data. Actually, each and every one of us now is a data generator (McAfee & Brynjolfsson, 2012). We produce data using our mobile phones, social networks interactions, GPS, etc. Most of such data, however, is not structured in a way so as to be stored and/or processed in traditional DBMS. This rather calls for (big) data analytics techniques in order to make sense out of such data.

The objective of this research is to elucidate knowledge with regard to big data analytics' state-of-the-art. That is, we will be investigating the current research and well as draft future research agenda based on literature and related work, which has been written or exerted in the area of big data analytics. The analysis will undergo a lens on the three different stages in big data analytics. Those steps are: storage, processing, and analytics. However, more focus will be put into analytics, subject of research, in terms of the techniques uses, tools, applications, opportunities, challenges, etc. Furthermore, we will apply text-mining techniques on the corpus of papers and derive what is common as from the writing (text) perspective. Last, but not least, we will predict future DBA trends. A total number of 24 papers were selected as the corpus of our research. An explanation of how and why we have selected such number of papers as a prime source of our investigation is discussed in the next section.

The rest of the paper is organized as follows: section 2 provides an overview of the research methodology employed in this study. The literature findings are presented in section 3. Section 4 provides a discussion of the findings and the current research gaps. Finally a research conclusion is presented in section 5.

2. SURVEYING THE LITERATURE: RESEARCH METHODOLOGY

In general, the literature search was done in two main stages. First, we have searched in the top 5 conferences and journals in data mining. Second, we have identified the top 10 cited papers in Google Scholar. A detailed description of the process is provided below.

To review the relevant big data analytics literature, we have decided to search using the term "Big Data Analytics", in the last five years (2010-2014), in peer-reviewed articles. Specifically, in the top 5 leading conferences and journals in data mining. As for the conference and journal ranking, we have adopted academic.research.micosoft.com ranking. As per the time of writing the paper (February, 2014), the top five conferences in data mining are: ACM Conference on Knowledge Discovery and Data Mining (KDD), IEEE International Conference on Data Engineering (ICDE), International Conference on Information and Knowledge Management (CIKM), IEEE International Conference on Data Mining (ICDM), and SIAM International Conference on Data Mining (SDM-SIAM). Additionally, the top five journals in data mining are: IEEE Transactions on Knowledge and Data Engineering (TKDE), Information Processing Letters (IPL), The International Journal on Very Large Data Bases (VLDB), Data Mining & Knowledge Discovery (DATAMINE), and ACM's Special Interest Group on Knowledge Discovery and Data Mining Explorations (SIGKDD Explorations). Consequently, the search keyword together with the remaining search conditions have been used in two search databases; IEEE Xplore and ACM DL. Furthermore, and to give the business dimension to big data analytics as well as avoid focusing only on the technical aspect of analytics, the same search term and conditions were applied to EBSCO research database. In EBSCO, two prime sources were selected: Business and Information.

Only English literature was selected and papers which have full text electronically available. Additionally, the publication type where papers published in business intelligence, data mining, computer science, information systems, or business related journals and conferences. This led to the exclusion of many types of papers such as whitepapers, working papers, books and book chapters.

The search for articles has gone through four different phases.

- **The Top 5 Criterion**

In phase one, the search term applied and only articles published in the top 5 data mining conferences and journals were filtered. Results are in the below table:

IEEE Xplore	ACM DL	EBSCO	Total
27	5	21	53

Table 1. Phase I outcome

The 53 articles were then downloaded and based on their abstract, only relevant papers were included forward and irrelevant papers were excluded. This has resulted in a total number of articles equal 28. Below is the table showing the papers in phase II:

IEEE Xplore	ACM DL	EBSCO	Total
8	5	15	28

Table 2. Phase II outcome

The 28 articles in Phase two were then filtered based on content where papers not directly discussing Big Data analytics were removed, together with duplicates, and too surface-level papers. This has resulted in a total of 14 papers to be reviewed. The below table shows the total number of papers selected in this phase.

IEEE Xplore	ACM DL	EBSCO	Total
8	3	3	14

Table 3. Phase III outcome

- **Google Scholar: The Citation Criterion**

Relying only on the outcome of the top-5 ranked conferences and journals as shown above, ended up with 14 articles. However, and to reach a better understanding of big data analytics literature, in phase IV, we have undertaken another round of paper selection from Google Scholar. But, this time the criteria have been the number of citations. Same search term and conditions were also used in Google Scholar. That is, big data analytics was the search term and peer-reviewed publications between 2010-2014 were filtered. Accordingly, we have added to our corpus of 14 articles, the top 10 cited big data analytics papers as per Google scholar. In total, we now have 24 articles (see table 4). For us, that was sufficient to start the analytics journey. The 10 articles added from Google scholar have almost the same characteristics of the previously collected corpus. That is, 80% of the authors are either affiliated solely to the US companies or academic institutions. Once again, it shows the competitive edge of the US companies and universities in the area of big data analytics. Lastly, the majority (90%) of the publications belong to year 2011 and 2012.

IEEE Xplore	ACM DL	EBSCO	G. Scholar	Total
8	3	3	10	24

Table 4. Phase IV outcome

The below figure explains the process which we have followed in order to reach the final list of papers to be analyzed in our literature review analysis research.

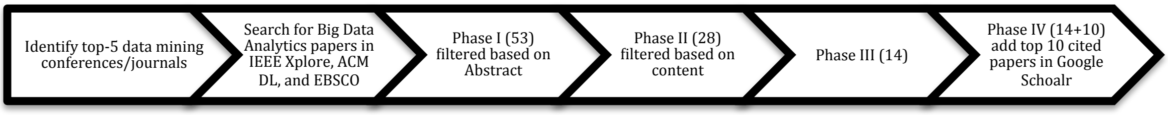
Identify top-5 data mining conferences/journals
Search for Big Data Analytics papers in IEEE Xplore, ACM DL, and EBSCO
Phase I (53) filtered based on Abstract
Phase II (28) filtered based on content
Phase III (14)
Phase IV (14+10) add top 10 cited papers in Google Schoalr

Figure 1: Our approach

2.1 Papers meta-analysis

In this section we are going to provide some meta-analysis of the 24 selected-for-analysis papers. Papers' meta-analysis shows that US is taking the leadership role in big data analytics. That is, two-thirds of authors are working either in US-universities or companies e.g., Google, Microsoft, Facebook, etc. The other one-third has China and India affiliation, with some relation to US academic institute or business.

On the other hand, the publications distribution over the years of analysis – see fig. 2 – shows increase in 2012 and 2013, compared to 2010.

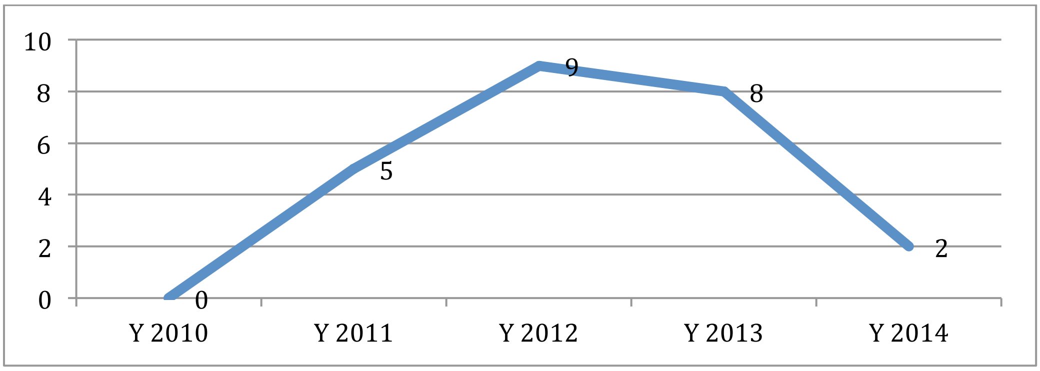

Figure 2. Publications over the 5-years

Another meta-analysis is related to the outlets where the papers have been published. Results show that IEEE and its affiliated conferences and journals take leadership with regards to publications in the area of big data analytics, the past 5 years. See below figure.

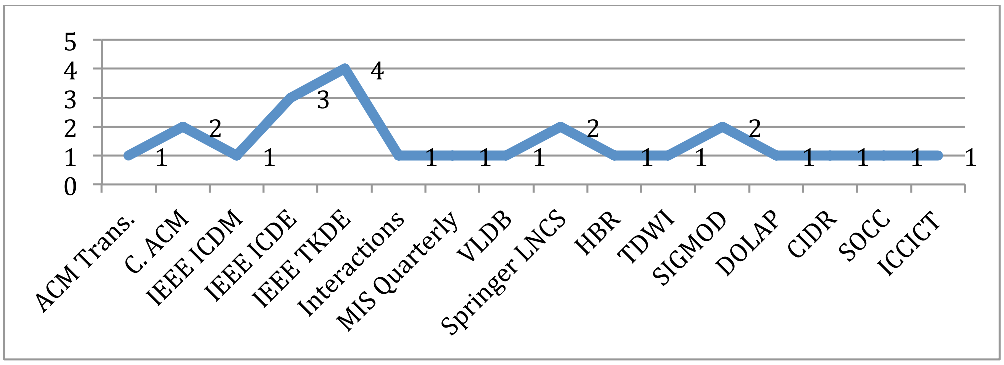

Figure 3. Publication outlets

2.2 Text mining

Text mining is the process of deriving knowledge or quality information form text. Text mining, AKA text analytics, is the process of deducing qualitative or unstructured pattern from text data set. As a matter of fact, we have been thinking about the framework to use, or theory, to analyze the text documents but we could not find. Therefore, we have followed this text-mining based approach to analyze literature. Therefore, and in an attempt to find hidden patterns in text of documents, the content of our corpus has been text-mined. The 24 documents were processed in Rapid Miner. The mining process is found in the following figure:

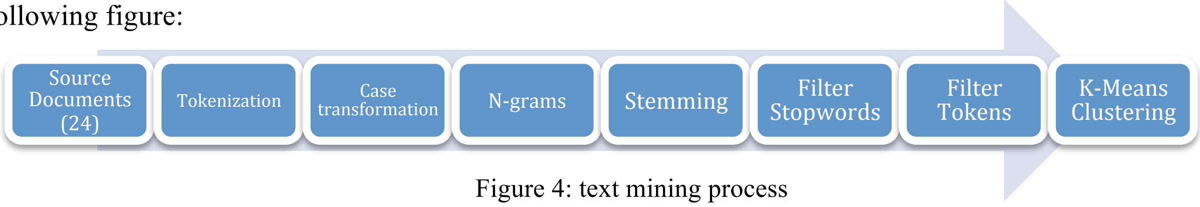

Figure 4: text mining process

As presented in figure 4, the process is –broadly speaking – split into two main sub processes; document processing and k-means clustering. Below is a description of both processes.

2.2.1 Document processing

Document processing is a complex text mining process; hence, it is broken down into sub processes as explained below:

- Tokenization: tokenizing transforms each and every word or character in documents' text into a token;
- Change case: Change case transforms the case of all the tokens to upper case or lower case to avoid treating them like two tokens;
- N-grams: once tokens have the same case, then n-grams operator takes place and this operator merge different tokens together to form structures that have a meaning. To illustrate, assumingly

we have 2 tokens: Big and Data. The n-gram operator will fuse them into “Big_Data” and that normally improves accuracy of later classification;

- Stemming: the process of stemming brings the word to its stem. That is, to make sure that the same words, which have various tenses, adjectives, nouns, etc. are considered the same. Also, this enhance classification accuracy;
- Filter Stopwords: the filter stopwords operator removes stopwords e.g., an, a, or, etc. This ensures that the words being matched together will be keywords not just simple stopwords;
- Filter tokens: the filter tokens operator takes the minimum and maximum length as a parameter.

2.2.2 Clustering

Grouping of similar records into separate, or overlapping, clusters is one of the fundamental tasks in data mining. So basically clustering is classifying unclassified data. Depending on the type of observations, or cases, clustering can be executed using central or pairwise techniques. Central clustering minimizes the average distance between an observation and its cluster center. Thus, clustering solution can be described by means of cluster centroids. Clustering can be used to achieve both objectives; description and prediction. In our case, we have used K-means, central clustering, in order to cluster the corpus (24 documents) into separate clusters. Mixed Euclidean measure was adopted. Results show that the documents fall into 3 clusters with different themes and focus.

The k-means algorithm is one of the most frequently used central clustering techniques. Data, in our case these are papers/documents, is divided iteratively into k clusters by minimizing average squared Euclidean Distance between the cases, or observations, and its cluster center.

2.2.3 Pre-processing

Distances between the 24 documents were calculated using Euclidean measure. This is the first step in the clustering process, which will be followed by fusing the nearest neighbours (documents with least distances and hence start forming k-clusters. The below figure shows the similarity matrix as a circle.

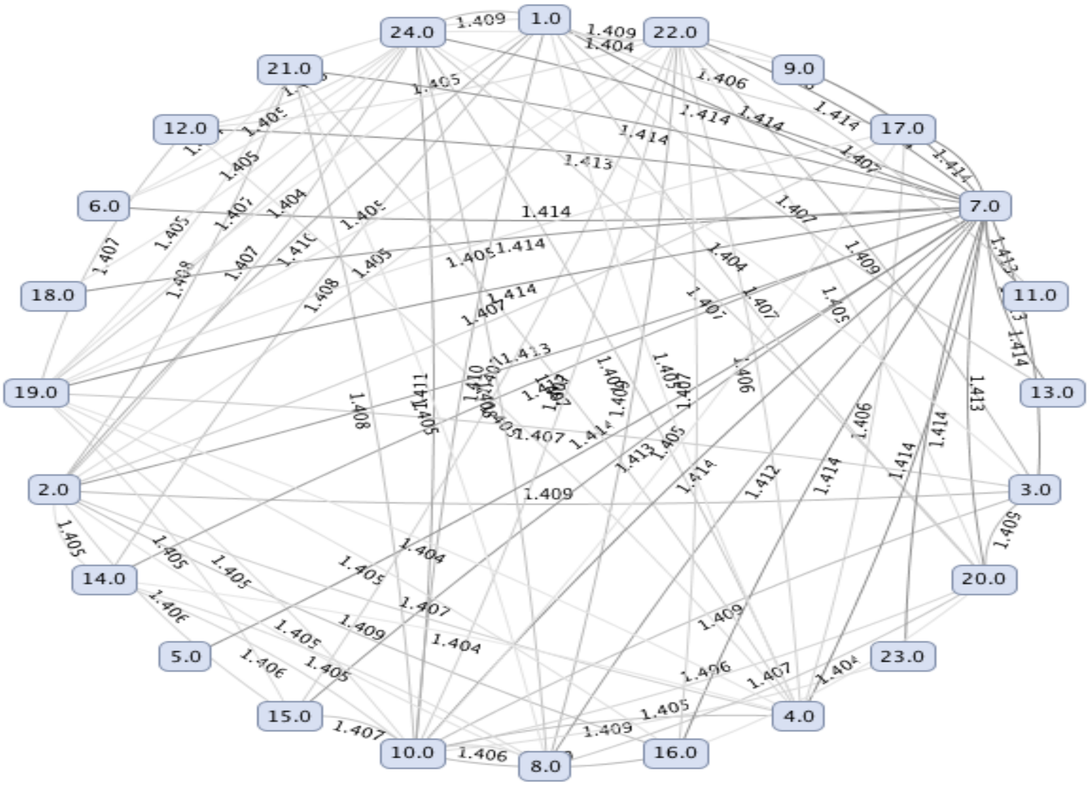

Figure 5: Similarity Circle

K-Means Results showed that the 24 documents are classified into 3 clusters. Cluster 0 contains 6 documents (25%); cluster 1 contains 9 documents (37.5%), which is the same size of cluster 2. This is, in principal, good result as indicates fairly good representation of documents into the three clusters. Details of membership is described as below:

- Cluster 0: consists of 6 documents (Wu, Zhu, Wu, & Ding, 2014), (Laptev, Zeng, & Zaniolo, 2013), (Lu & Li, 2013), (Wang, Zhao, Hoi, & Jin, 2014), (Chandramouli, Goldstein, & Quamar, Scalable Progressive Analytics on Big Data in the Cloud, 2013), (Han, Li, Yang, & Wang, 2013)
- Cluster 1: consists of 9 documents (Agneeswaran, 2012), (Russom, 2011), (Gupta, Gupta, & Mohania, 2012), (Singh & Singh, 2012), (McAfee & Brynjolfsson, 2012), (Herodotou, et al., 2011), (Cuzzocrea, Song, & Davis, 2011), (Ghazal, et al., 2013), (He, et al., 2011)
- Cluster 2: consists of 9 documents (Huai, Lee, Zhang, Xia, & Zhang, 2011), (Cheng, Qin, & Rusu, 2012), (Mozafari, Zeng, D'Antoni, & Zaniolo, 2013), (Narang, Srivastava, & Katta, 2012), (Chandramouli, Goldstein, & Duan, Temporal Analytics on Big Data for Web Advertizing, 2012), (Kumar, Niu, & Ré, 2013), (Dhar, 2013), (Chen, Chiang, & Storey, 2012), (Fisher, DeLine, Czerwinski, & S., 2012).

3. DISCUSSION OF FINDINGS

One thing that was common amongst the three clusters is the fact that each and every cluster included documents (papers) from different outlets. That is, cluster 0 has papers from VLDB, TKDE, SIGMOD, and IDCE. Cluster 1 has papers from ACM Interactions, MISQ, IKDE, ICDE, HBR, Springer chapters, and other conferences. As for cluster 2 it has papers from SIGMOD, TDWI, CACM, ICDM, ICDE, and others. As a matter of fact, one can learn here that membership of clusters indicates that there is no journal nor conference dominance. That is, as per their published text, different journals and conferences are classified together and that includes no strong text-publisher relationship. In the next subsections, we will be shedding some lights on the characteristics and contents of each cluster.

3.1 Characteristics of Cluster 0: Technical algorithms

The first finding is that the cluster is the least in size amongst the three; 25% (6 papers). Additionally, the cluster has obtained papers from VLDB, ICDE, TKDE, and SIGMOD. That is, the 6 papers are sort of technical papers. The papers are discussing or suggesting (new) algorithms and, techniques, and infrastructures. In contrary, there is almost nothing about business models, big data value to business, return on investment (ROI), challenges, security, cloud-based big data, etc. Hence, this cluster could be swiftly named “Technical algorithms”. Surprisingly, this cluster has none of the top cited papers as per Google Scholar. One reason could be, that vastly specialized technical and algorithmic papers are not usually highly cited when compared to papers that discuss broad business ideas or issues.

Given the complexity and actual distribution of data over many sources or sites, managing and mining big data are non-trivial yet very attractive tasks (Wu, Zhu, Wu, & Ding, 2014). While the notion of big data primarily concerns about data volumes, however Wu et al. (2014) argue that size is not the main characteristic or challenge of big data. They also discuss the technical challenges related to data samples, structures, heterogeneity of sources, mining models and algorithms, and systems infrastructures that would support data analytics. In addition, Wu et al. (2014) proposed the HACE theory, in which they explained the characteristics of big data through being 1) huge with Heterogeneous and various data sources, 2) Autonomous with dispersed and decentralized control, and 3) Complicated and Evolving in data and knowledge associations. They conclude that in order to create value through big data analytics, high-performance computing platforms are required and a standardized and reliable information sharing protocol is needed. In addition, there is a need to design global models that are able to fuse and form a unified view of data from multiple sources. Also, there is a persistent need for carefully-designed data mining algorithms that are able analyze model correlations between scattered sites, and are able fuse decisions from multiple sources to gain the best value of the big data (Wu et al., 2014).

Laptev, Zeng, & Zaniolo (2013) argue about the efficiency of mining in smaller samples of data in contrast to mining a whole data set, specifically large data sets (e.g. big data). Advanced analytical applications lack speed and efficient computational resource management, specifically when drawing samples from large databases. Existing systems do not provide accurate estimate of incremental results and are best accommodating for batch processing. Thus, they have introduced the early accurate result library (EARL), which is a bootstrapping-based framework that estimates results and errors for the different mining techniques. Generally, EARL works on data samples through predicting the learning curve and aiding in choosing the appropriate sample size for achieving the desired data mining error bound specified by users. They concluded that it is mostly unnecessary to mine a large dataset in bulk, and it is usually sufficient to sample only up to 1% of the dataset for mining. In addition, their results show that with the proper sampling, increasing the sample size doesn't specifically improve the error ratio (Laptev et al., 2013). While Laptev et al. (2013) focused on reducing the sample size of online big data to optimize mining, (Lu & Li, 2013) focused on reducing bias in small samples and datasets. Lu and Li have developed an algorithm for bias correction in small samples drawn from online big data (e.g. Twitter) (Lu & Li, 2013). They argue that bias mainly depends on the expected number of collisions in the sample. The evaluation results of their algorithm show that it is able to accommodate both uniform random sampling and random walk sampling.

Other studies (Wang, Zhao, Hoi, & Jin, 2014), have focused on enhancing the feature selection methods to enhance the performance of mining algorithms (e.g. classification) in big data. While, the evaluation of their proposed algorithms shows that they are moderately effective for feature or variable selection jobs on online applications, still they are considerably more efficient and scalable than some of the existing state-of-the-art batch feature selection algorithms.

In order to optimize querying samples from big data in the cloud, Chandramouli, Goldstein, & Quamar (2013) proposed a progressive model called Prism. Prism allows users to transfer progressive samples from the cloud to the system, and aid in more efficient and deterministic query processing over those samples. In addition, Prism provides repeatable semantics and attribution to data scientists. Their results show that Prism provides optimized speed and progressive SQL support over big data in Microsoft's cloud Azure. Likewise, other researchers focused on query optimization techniques (Han, Li, Yang, & Wang, 2013). Specifically, the optimization of skyline queries over big data. Han et al. (2013) argue that the current algorithms are not efficient when performing skyline on big data. Thus, they proposed a new skyline algorithm (SSPL), which employs sorted positional index lists of low space overhead in order to reduce the input/output costs. Their results show that proposed SSPL algorithm has a substantial improvement over the current skyline query algorithms.

3.2 Characteristics of Cluster 1: Processing, cloud computing, opportunities & challenges

This cluster contains 9 papers, which represents 37.5% of the total corpus. This cluster, also, obtained papers from various published e.g., ACM Interactions, MISQ, HBR, Springer (book chapters), TKDE, ICDE, and other conferences. Here we notice the mix of publishers also supporting that there is no strict (single) publisher-to-(single) cluster association.

In this cluster, the topics discussed in the papers and articles focused mainly on big data processing, Hadoop, Map-reduce, No-SQL, big data applications in business, cloud computing, and big data challenges and opportunities. These topics characterize this cluster, as these topics were not covered in the other two clusters.

Trends, challenges, and opportunities of big data analytics to businesses have been well covered in literature. The evolving of social networks, and the increasing speed of networks and processing power had enabled businesses to exploit big data in several ways. Thus, several researchers explored the fundamental trends that led big data into the spot light as well as discuss the analytics, engineering and theoretical trends in this domain (Agneeswaran, 2012; Russom, 2011). The studies concluded that several technologies and infrastructures aided in the emergence and utilization of big data. For example the emergence of powerful systems that enable video analytics, ad placements and the software defined networks (SDNs) have contributed to the exploitation of big data (Agneeswaran, 2012). Additionally, the emergence of systems that support multi-connectors e.g. Hadoop aided in wide range analytics on large datasets (Agneeswaran, 2012). Moreover, researchers and practitioners are working on enhancing several data mining algorithms that would aid in large-scale complex analytics, like those in video and real-time analytics (Agneeswaran, 2012). Also, these genres of studies advocate on evaluating and comparing the existing systems and solutions before investing in big data analytics projects (Russom, 2011). Similarly, Gupta, Gupta, & Mohania (2012) have analyzed the current trends of big data analytics and cloud computing from a database perspective. While there are systems and technologies that support distributed processing and analytics of large unstructured data, yet there is still a persistent need by enterprises for faster and more powerful systems (Gupta, Gupta, & Mohania, 2012). In addition, Gupta et al. (2012)

stressed on the advantages of moving to the cloud over having dedicated data-management resources. They concluded that more work and research are needed to enhance cloud security and ability for real-time processing and analytics, as businesses and governments are still skeptical to move their critical data to the cloud.

Another study presented the current big data adoption trends, vendor and product selection criteria, best practices, and benefits of big data analytics (Singh & Singh, 2012). The study concluded that the cooperation with customers and insights from user-generated online contents are critical key enablers for the success in the era of social media. In addition, the study stressed that there is a considerable gap of skilled managers that can make decisions based on data analysis. Thus, constructing teams that are able to make analytic-based decisions is of paramount importance to grasp the real value of big data (Singh & Singh, 2012). Similarly, MacAfee & Brynjolfsson (2012) argue about the importance of data-based decision making. They presented a study in which they concluded that companies who conduct big data analytics are 5% more productive, and 6% more profitable than those who don't. They argue that there are five management challenges that businesses should manage and change effectively in order to realize benefits from big data. The five challenges are leadership, talent management, technology, decision-making, and company culture. Finally, they concluded that managers who don't base their decisions upon data analytics would have no place in future businesses due to the fierce market competition (McAfee & Brynjolfsson, 2012). The MapReduce based system Hadoop provides opportunities for optimizing big data analytics. Nevertheless, optimizing and tuning a system like Hadoop are not easy tasks for the average user. In their research, (Herodotou, et al., 2011) have introduced Starfish. Starfish is a self-tuning system that works on top of Hadoop in order to aid in automatically tuning Hadoop with little effort and expertise from the user's side. One of the main objectives of Starfish is to, with almost no intervention from the user side, optimize the speed of Hadoop throughout the data lifecycle in analytics. While there were previous similar projects/systems like Hive (Thusoo, 2010), however, the authors argue that the novelty of Starfish is in its ability to optimize and reduce job costs simultaneously on the different scheduling and workload granularities and levels (Herodotou, et al., 2011). Cuzzocrea et al. (2011) studied the research trends and state-of-the-art issues in big data analytics. They have presented the current research gaps, which mainly address the need for online analytical processing (OLAP)-like systems that enable big data analytics on multi-dimensional data models, as star schema based data warehouses.

(Ghazal, et al., 2013) proposed a novel big data analytics benchmarking system called BigBench. Through a real-world retailer's big data, BigBench was implemented and evaluated on Teradata Aster DMBS (TAD) system. The evaluation process included a list of workload queries that focus on the different types of processing in big data. The evaluation results show that BigBench was proven feasible and applicable for metrics evaluation of big data analytics and queries when implemented on TAD. Moreover, the authors have introduced a novel technique for producing and integrating unstructured text data with a structured data generator (Ghazal, et al., 2013).

Given the high cost of data placement into conventional large databases (e.g. data warehouse), a study (He, et al., 2011) developed RCFile, a novel big data placement architecture. In general, data placement structures have four important requirements; 1) fast data loading, 2) fast query processing, 3) efficient storage space management, 4) adaptivity to highly dynamic workload patterns (He, et al., 2011). The authors claim that the current conventional data placement structures, like rowstores, column-stores, and hybrid-stores don't satisfy all of these considerations and requirements in distributed MapReduce big data environments. Thus, they developed RCFile (Record Columnar File). The RCFile system was implemented over Hadoop and tested extensively on Facebook's data. The experiments show that RCFile satisfies the four above-mentioned requirements, and has been adopted by Facebook and integrated in their data warehouse system. In addition, it has been adopted by Hive and Pig big data ecosystems (He, et al., 2011).

3.3 Characteristics of Cluster 2: Performance, prediction, and distributed systems

This cluster – as well - has 9 papers, which represents 37.5% of the total corpus. This cluster, also, obtained papers from various outlets e.g., MISQ, SIGMOD, CACM, TDWI, ICDE, ICDM, and other conferences. Once again, we also notice the mix of publishers also supporting that there is no strict (single) publisher-to-(single) cluster association.

This cluster has been dominated by papers focusing on suggesting methods or architecture to measure big data performance, applications of big data analytics in predictions, and design of big data in distributed environment. Unlike the first cluster, cluster 0, this cluster enjoys some of the highly- cited papers. Probably this is attributable to the generic models and methods suggested by those papers.

In their paper, Huai et al. (2011) have addressed performance issues of big data analytics in distributed environments. They mainly argue that the current data analytics systems and applications are not efficient when analyzing distributed large data sets. So, they proposed the DOT framework, which extends the big data ecosystems in order to optimize handling large distributed data sets for analytics, concurrency control, and interaction between users and those data sets in real-time manner. Tests on DOT show its effectiveness, scalability, and fault-tolerance ability on complex analytic queries over MapReduce and Dryad (Huai, Lee, Zhang, Xia, & Zhang, 2011). Likewise, another study focused on optimizing data analytics operations over distributed big data environments (Cheng, Qin, & Rusu, 2012). The study presented GLADE, which is a scalable distributed system for big data analytics. GLADE mainly takes analytical functions and executes them efficiently on the inputted data. The proposed system attempts to take full advantage of the parallelism available inside single machines as well as across a cluster of distributed computing nodes. The system demonstrations show that GLADE can outperform Hadoop in several querying and analytical operation scenarios, namely Average, k-means, Group by, and Top-K. This is mainly because GLADE uses columnar storage and reads only the data needed for query execution while Hadoop normally reads all the data in the relation. In addition, GLADE employs point-to-point communication between the distributed nodes while Hadoop requires all-to-all communication among the nodes (Huai, et al., 2011).

Another study by Mozafari, Zeng, D'Antoni, & Zaniolo (2013), worked on extending XML's query language XPath. The study argues that in some scenarios, the well-known complex queries from distributed applications could be troublesome or impossible to be handled by XPath. Thus, the study proposed XSeq, which is an XML query language that extends the conventional XPath expressions and its dialects in order to optimize and ease querying large XML streams and exchanges (Mozafari, et al., 2013). On the other hand, Narang et al. (2012) addressed the challenges of accuracy and speed in real-time co-clustering and collaborative filtering. They have proposed a hierarchical approach for distributed online and offline big data co-clustering and collaborative filtering. The approach has been tested online and offline on Netflix and Yahoo's KDD Cup datasets on a multi-core cluster infrastructure. The test results show that proposed approach, while maintaining high accuracy, outperforms all existing collaborative filtering mechanisms, both online and offline (Narang, et al., 2012). Other authors (Chandramouli, Goldstein, & Duan, 2012) proposed the TiMR framework. As temporal queries easy to specify and naturally real-time-ready, the introduced TiMR framework is based on the use of those queries. TiMR enables temporal queries to scale up to big offline datasets on existing MapReduce infrastructures. The main purpose of this research was to enhance the behavioral targeting (BT) mechanisms used in online-targeted advertisements and optimize it in real-time environments. The research experiments validate TiMR's high scalability and performance in the real-time BT applications domain (Chandramouli, et al., 2012).

The recent success of several big data analytics-driven systems has created a massive interest in bringing such technological potential abilities to a wider variety of business domains (Kumar, Niu, & Ré, 2013). On the other hand, there are still big challenges in making these analytical systems easy to build and maintain. Kumar et al. (2013) have introduced the open-source Hazy project, which targets these challenges through identifying the common patterns across domains that would ease and speed-up building the analytical systems and transferability among domains. The Haze code been employed by four enterprises as well as a research observatory at the South Pole (Kumar, et al., 2013).

Finally, other researchers investigated the meaning of new terms as data science, importance of predictive modeling in big data environments (Dhar, 2013), and how the Information Systems discipline can better support the needs of business managers in light of the emerging business intelligence, analytics technologies and ubiquitous big data infrastructures (Chen, Chiang, & Storey, 2012; Fisher, DeLine, Czerwinski, & S., 2012). The authors predict that data-savvy managers and professionals with deep analytical skills will be much needed for businesses but hard to find (Chen, Chiang, & Storey, 2012; Dhar, 2013). Therefore, universities and IS programs should accommodate the needs of the future job market through providing education combining big data business intelligence and analytics into their curricula (Chen, et al., 2012). On the other hand, it is most likely that the supply of the skilled graduates might be less than required market demand, thus a need for applications that facilitate and allow the data-savvy users to conduct their own analysis and data visualizations (Fisher, et al., 2012).

4. FUTURE RESEARCH AVENUES

The reviewed articles are spread across 16 various outlets. Among the academic outlets, we have recognized only one special journal issue (MISQ) focusing on business intelligence research, which included one paper focusing on big data analytics. As the research interest on big data analytics is increasing, research outlets should pay more attention to this topic.

In general, 24 articles across 5 years period is relatively a low number of publications. Despite the need for research on big data analytics was recognized in previous literature, still the amount of research conducted on this issue is limited and scarce. Thus, more research needs to be carried out in order to gather sufficient knowledge about this phenomenon.

Although some papers presented results of their models and algorithms' lab experiments conducted on actual data, like Twitter, however, more case study research is still in need to be conducted at other various business sectors. In addition, a significant number of researches tackled the subject of social network analytics, but surprisingly, little insights on how to use and integrate social networks analytics into the decision making process in organizations, with the exception of marketing. Also, very few researchers discussed text mining and sentiment analysis algorithms and their application on social data.

From our understanding and analysis of big data analytics papers, we realized that almost all projects, cases, or implementations share one or more of those challenges:

- *Access to source data set*: BDA assumes the availability & access to original (AKA primary or raw) data. Such primary data may not always be available for analytics purposes. Accordingly, we believe that big data analytics should be able to be implemented and run without the luxury of primary data. That is due to (Roddick, Spiliopoulou, Lister, & Ceglar, 2008):
  - Cooperating institutions that are interested in sharing knowledge may not be willing (or allowed) to disclose their primary data;
  - Data in the form of streams are only temporarily available for processing.
  - Even for non-stream data, there are limits on the computation power to be achieved.
- *Understandability of discovered patterns*: while advances in data mining encompass very powerful algorithms, there are fewer advances on driving the knowledge discovery process towards results appropriate for human consumption.
- *Privacy preserving*: The demand for privacy-preservation in data mining emerges in two different related contexts: 1. Personal data must be protected from disclosure towards everyone; & 2. Confidential data must be protected from disclosure towards partners.
- *Algorithm tractability*: Mining techniques are beginning to encounter problems as the volume of data requiring analysis grows disproportionately with the comparatively slower improvements in I/O channel speeds. That is, many mining techniques are becoming heavily I/O bound and this is limiting their benefits. Methods to reduce the amount of data have been presented in the literature including statistical methods e.g., dimension reduction.

So, we feel strong that the future of BDA research should focus on solving those persistent challenges. Some of the problems and challenges of BDA could be potentially solved using higher order mining (HOM). Such approach is absolutely unique and it hasn't been sufficiently covered in information systems literature in general nor widely applied in practice. In addition, based on our review, we couldn't identify any research that discusses HOM in big data context and environments. Higher Order Mining is a form of analytics that is applied over non-primary data or patterns (Roddick, Spiliopoulou, Lister, & Ceglar, 2008).

5. CONCLUSION

Through organizing the literature based on text mining and clustering techniques, this paper contributes to both research and practice through providing a comprehensive literature review of big data analytics. For practice, the paper sheds the light on past and recent issues, challenges, and success stories that can guide consultants, vendors, and clients in their future projects. For research, the organization of literature in the three clusters can aid them in identifying the topics, findings, and gaps discussed in each topic of interest. Finally, we have provided our observations and future research suggestions that would enrich our knowledge in this domain.

Acknowledgments

This research is funded by the Cloudberry Datacenters project, Luleå University of Technology, Luleå, Sweden.

References

Agneeswaran, V. (2012). Big-Data – Theoretical, Engineering and Analytics Perspective. (S. Srinivasa, & V. Bhatnagar, Eds.) LNCS , 7678, 8-15.

Chandramouli, B., Goldstein, J., & Duan, S. (2012). Temporal Analytics on Big Data for Web Advertizing. The 28th International Conference on Data Engineering (ICDE) (pp. 90-101). IEEE.

Chandramouli, B., Goldstein, J., & Quamar, A. (2013). Scalable Progressive Analytics on Big Data in the Cloud. The 39th International Conference on VLDB (pp. 1726-1737). Trento: ACM.

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: from Big Data to Big Impact. MIS Quarterly , 36 (4), 1165-1188.

Cheng, Y., Qin, C., & Rusu, F. (2012). GLADE: Big Data Analytics Made Easy. SIGMOD (pp. 697-700). AR: ACM.

Cuzzocrea, a., Song, I., & Davis, K. (2011). Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! DOLAP (pp. 101-103). Glasgow: ACM.

Dhar, V. (2013). Data Science and Prediction. Communications of the ACM , 56 (12), 64-73.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. The 14th Industrial Conference on Data Mining (ICDM). Petersburg: Springer-LNCS.

Fisher, D., DeLine, R., Czerwinski, M., & S., D. (2012, May-June). Interactions With Big Data Analytics. Interactions , 50-59.

Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., et al. (2013). BigBench: Towards an Industry Standard Benchmark for Big Data Analytics. SIGMOD (pp. 1197-1208). NY: ACM.

Gupta, R., Gupta, H., & Mohania, M. (2012). Cloud Computing and Big Data Analytics: What Is New from a Database Perspective. (S. Srinivasa, & V. Bhatnagar, Eds.) LNCS , 7678, 42-61.

Han, X., Li, J., Yang, D., & Wang, J. (2013). Efficient Skyline Computation on Big Data. TKDE , 25 (11), 2521-2535.

He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., et al. (2011). RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. The 27th International Conference on Data Engineering (ICDE) (pp. 1199-1208). IEEE.

Herodotou, H., Lim, H., Luo, G., Boisov, N., Dong, L., Cetin, F., et al. (2011). Starfish: A Self-tuning System for Big Data Analytics. 5th Biennial Conference on Innovative Data Systems Research (CIDR '11), (pp. 261-272). CA.

Huai, Y., Lee, R., Zhang, S., Xia, C., & Zhang, X. (2011). DOT: A Matrix Model for Analyzing, Optimizing and Deploying Software for Big Data Analytics in Bistributed Systems. SOCC (pp. 1-14). Cascais: ACM.

Kumar, A., Niu, F., & Ré, C. (2013). Hazy: Making It Easier to Build and Maintain Big Data Analytics. Communications of The ACM , 56 (3), 40-49.

Laptev, N., Zeng, K., & Zaniolo, C. (2013). Very Fast Estimation for Result and Accuracy of Big Data Analytics: the Earl System. The 29th International Conference on Data Engineering (ICDE) (pp. 1296-1299). IEEE.

Lu, J., & Li, D. (2013). Bias Correction in a Small Sample From Big Data. TKDE , 25 (11), 2658-2663.

McAfee, A., & Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. HBR , 3-9.

Mozafari, B., Zeng, K., D'Antoni, L., & Zaniolo, C. (2013). High-Performance Complex Event Processing over Hierarchical Data. ACM Transactions on Database Systems , 38 (4), 21-39.

Narang, A., Srivastava, A., & Katta, N. (2012). High Performance Offline & Online Distributed Collaborative Filtering. The 12th International Conference on Data Mining (ICDM) (pp. 549-558). IEEE.

Roddick, J., Spiliopoulou, M., Lister, D., & Ceglar, A. (2008). Higher Order Mining. SIGKDD Explorations , 10 (1), 5-17.

Russom, P. (2011). Big Data Analytics. TDWI , 4th Quarter, 1-38.

Singh, S., & Singh, N. (2012). Big Data Analytics. International Conference on Communication, Information & Computing Technology (ICCICT) (pp. 1-4). Mumbai: IEEE.

Thusoo, A. S. (2010). Hive-a petabyte scale data warehouse using hadoop. 26th International Conference on Data Engineering (ICDE) (pp. 996-1005). IEEE.

Wang, J., Zhao, P., Hoi, S., & Jin, R. (2014). Online Feature Selection and Its Applications. TKDE , 26 (3), 698-710.

Wu, X., Zhu, X., Wu, G., & Ding, W. (2014). Data Mining with Big Data. TKDE , 26 (1), 97-107.