Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Research Analysis using Big Data

Ms. Sheetal and Dr. Deepa.S
Department of MCA, Presidency College, Bangalore, India

Abstract: Big data refers to data that are not only large, but also more in variety and velocity, which is very difficult to handle the data using traditional methods. The large amounts of data are transacted in internet due to different number of internet applications. Big data analytics is used to optimize manage and utilize the huge data and it can be used by business firms, researchers and extra .This paper focuses on review of research on the field of recent trends in big data analytics. The findings of this paper will help the researchers to find the problem statement and to continue the research in Big data domain and it can be used for the business organization to invest on different big data projects .The content of this paper is based on review of different publications from recent years and based on results this paper can be classified into different main themes.

Keywords: *Big data analytics, literature review*

I. INTRODUCTION

In the past, this challenge was mitigated by processors getting faster, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

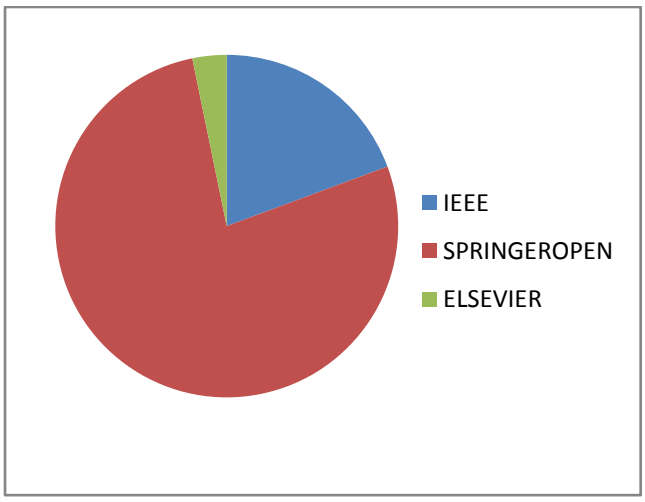
II. SURVEYING THE LITERATURE: RESEARCH METHODOLOGY

The literature search was done based on different Publishers from different journals. A detailed description of the process is provided below.

To review the relevant big data analytics literature, we have decided to search using specific term "Big Data", in the recent three years (2015-2017).We have collected 62 journals. The classification of the journal is described in the following tables.

PHASE I:

IEEE	SPRINGEROPEN	ELSEVIER	Total
12	48	2	62



PHASE II:

IEEE		SPRINGEROPEN			ELSEVIER		Total
General	Technical	General	Technical	Survey	General	Technical	62
03	09	22	18	08	1	1	

PHASE III:

IEEE		SPRINGEROPEN			ELSEVIER		Total
General	Technical	General	Technical	Survey	General	Technical	33
3	9	6	8	5	1	1	

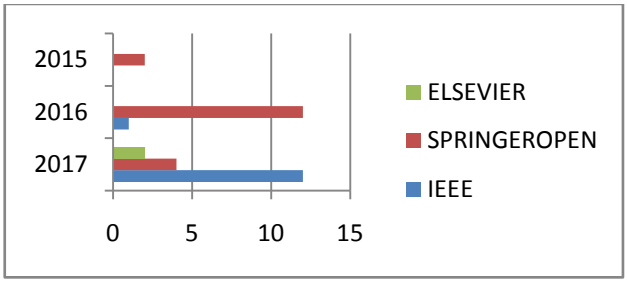
PHASE IV:

Technical	General	Survey	Total
18	**10**	**5**	**33**

Approach Diagram & Paper Analysis(2015-2017)

Paper Analysis

Paper analysis going to provide some analysis of the 33 selected papers. The results shows that china is taking the leadership role in big data analytics. That is, two-thirds of authors are working either in India or companies e.g., Google, Microsoft, Facebook, etc. The other one-third has USA , Finland, Qatar, Iran and UK affiliation, with some relation to US academic institute or business.



III. DISCUSSION AND FINDINGS

A. Technical algorithms

1. Scalable Uncertainty-Aware Truth Discovery in Big Data Social Sensing Applications for Cyber-Physical Systems

(Chao Huang & et.al, 2017) proposed a Scalable Uncertainty-Aware Truth Discovery (SUTD) scheme to address scalablity change. The SUTD scheme solves a constraint estimation problem to jointly estimate the correctness of reported data and the reliability of data sources while explicitly considering the uncertainty on the reported data. To address the scalability challenge, the SUTD is designed to run a Graphic Processing Unit (GPU) with thousands of cores, which is shown to run two

to three orders of magnitude faster than the sequential truth discovery solutions.

2. A Distributed Stream Library for Java 8

(Yu Chan & et.al, 2017) has explained an overview of the Java 8 Streams API and proposes extensions to allow its use in Big Data systems. It also shows how the API can be used to implement a range of standard Big Data paradigms. Finally, it compares performance with that of Hadoop and Spark. Despite being a proof-of-concept implementation, results indicate that it is a lightweight and efficient framework, comparable in performance to Hadoop and Spark, and is up to 5 times faster for the largest input sizes tested.

3. A Secure and Verifiable Access Control Scheme for Big Data Storage in Clouds

(Chunqiang Hu & et.al, 2017) proposed a secure and verifiable access control scheme based on the NTRU cryptosystem for big data storage in clouds. In the proposed NTRU decryption algorithm to overcome the decryption failures existing NTRU, and analyze its correctness, security strengths, and computational efficiency. Our scheme allows the cloud server to efficiently update the cipher text when a new access policy is specified by the data owner, who is also able to validate the update to counter against cheating behaviors of the cloud. It also enables (i) the data owner and eligible users to effectively verify the legitimacy of a user for accessing the data, and (ii) a user to validate the information provided by other users for correct plaintext recovery. Rigorous analysis indicates that the proposed scheme can prevent eligible users from cheating and resist various attacks such as the collusion attack.

4. Exploiting Efficient Densest Subgraph Discovering Methods for BigData

(Bo Wu & Haiying Shen, 2017) proposed the study the densest subgraph problem by designing two different algorithms based on different features that natural graphs have. First, by analyzing the features of natural graphs, the proposed heuristic algorithm for discovering the connected densest subgraph for massive undirected graphs in a MapReduce framework by taking advantage of the features of natural graphs. Second, the proposed exact algorithm for big data for the problem of discovering the densest subgraph.

5. Ring: Real-Time Emerging Anomaly Monitoring System over Text Streams

(Weiren Yu & et.al, 2017) proposed real-time emerging anomaly monitoring system over microblog text streams. RING integrates our efforts on both emerging anomaly monitoring research and system research. From the anomaly monitoring perspective, RING proposes a graph analytic approach such that (1) RING is able to detect emerging anomalies at an earlier stage compared to the existing methods, (2) RING is among the first to discover emerging anomalies correlations in a streaming fashion, (3) RING is able to monitor anomaly evolutions in real-time at different time scales from minutes to months. From the system research perspective, RING (1) optimizes time-ranged keyword query performance of a full-text search engine to improve the efficiency of monitoring anomaly evolution, (2) improves the dynamic graph processing performance of Spark and implements our graph stream model on it, As a result, RING is able to process big data to the entire Weibo or Twitter text stream with linear horizontal scalability. The system clearly presents its advantages over existing systems and methods from both the event monitoring perspective and the system perspective for the emerging event monitoring task.

6. Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud

(Hui Cui & et.al, 2017) proposed an attribute-based storage system with secure deduplication in a hybrid cloud setting, where a private cloud is responsible for duplicate detection and a public cloud manages the storage. Compared with the prior data deduplication systems, prposed system has two advantages. Firstly, it can be used to confidentially share data with users by specifying access policies rather than sharing decryption keys. Secondly, it achieves the standard notion of semantic secaurity for data confidentiality while existing systems only achieve it by defining a weaker security notion. In addition, the forth methodology to modify a ciphertext over one access policy into ciphertexts of the same plaintext but under other access policies without revealing the underlying plaintext.

7. JouleMR: Towards Cost-Effective and Green-Aware Data Processing Frameworks

(Zhaojie Niu ;& et.al,2017) proposed JouleMR, a cost-effective and green-aware data processing framework. Specifically, it investigate how to exploit such joule efficiency to maximize the benefits of renewable energy as well as dynamic pricing schemes for MapReduce framework. The proposed scheme develop job/task scheduling algorithms with a particular focus on the factors on joule efficiency in the data center, including the energy efficiency of MapReduce workloads, renewable energy supply, dynamic pricing and the battery usage. The proposed method is implemented JouleMR on top of Hadoop YARN. The experiments demonstrate the accuracy of our models, and the effectiveness of proposed cost-effective and green-aware optimizations outperform the state-of-the-art implementations over Hadoop YARN.

8. Towards Max-Min Fair Resource Allocation for Stream Big Data Analytics in Shared Clouds

(Zhaojie Niu xho&et.al,2016) proposed an efficient resource allocation scheme for a heterogeneous shared stream big data analytics cluster shared by multiple topologies, in order to achieve max-min fairness in the utilities of the throughput for all the topologies. Proposed method first formulate a novel model resource allocation problem, which is a mixed 0-1 integer program. The NP-hardness of the problem is rigorously proven. To tackle this problem, in this method transform the non-convex constraint to several linear constraints using linearization and reformulation techniques. Based on the analysis of the problem-specific structure and characteristics, The proposed approach iteratively solves the continuous problem with a fixed set of discrete variables optimally, and updates the discrete variables heuristically. Simulations show that our proposed resource allocation scheme remarkably improves the max-min fairness in utilities of the topology throughput, and is low in computational complexity.

9. Toward Efficient and Flexible Metadata Indexing of Big Data Systems

(Yuxuan Jiang et.al, 2017) proposed design Dindex, a distributed indexing service for metadata. Dindex incorporates a hierarchy of coarse-grained aggregation and horizontal key-coalition. Theoretical analysis shows that the overhead of building Dindex is compensated by only two or three queries. Dindex has been implemented by a lightweight distributed key-value store and integrated to a fully-fledged distributed filesystem. Experiments demonstrated that Dindex accelerated metadata queries by up to 60% with a negligible overhead.

10. d2o: a distributed data object for parallel high‐performance computing in Python

(Steininger et.al, 2016)Proposed design is Python module for cluster-distributed multi-dimensional numericalarrays. It acts as a layer of abstraction between the algorithm code and the datadistribution logic. The main goal is to achieve usability without losing numerical performance and scalability. d2o is written in pure Python which makes it portable and easy to use and modify.

11. Limited random walk algorithm for big graph data clustering

(Zhang et.al, 2016) The proposed method restricts the reach of the walking agent using an inflation function and a normalization function. The proposed algorithm tackles the problem in an entirely different manner. We use the limited random walk procedure to find attractor vertices in a graph and use them as features to cluster the vertices. Since the proposed method uses the embarrassingly parallel paradigm, it can be efficiently implemented and embedded in any parallel computing environment such as a MapReduce framework. Given enough computing resources, we are capable of clustering graphs with millions of vertices and hundreds millions of edges in a reasonable time.

Published in: Springeropen Journal(Zhang *et al. J Big Data (2016) 3:26)*

12. Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud

(Singh and Kaur, 2016)The proposed design is Apache Hadoop is the good option and it has many components that worked together to make the hadoop ecosystem robust and efficient.

Apache Pig is the core component of hadoop ecosystem and it accepts the tasks in the form of scripts. To run these scripts Apache Pig may use MapReduce or Apache Tez framework. Try to perform the analysis on multinode cluster which is installed at Amazon cloud.

Published in: Springeropen Journal(Singh and Kaur *J Big Data (2016) 3:19)*

13. An efficient strategy for the collection and storage of large volumes of data for computation

(Suthakar et.al, *2016)* this paper proposes a number of models are explored to understand what should be the best approach for collecting and storing Big Data for analytics. An evaluation of the performance of full execution cycles of these approaches on the monitoring of the Worldwide

LHC Computing Grid (WLCG) infrastructure for collecting, storing and analyzing data is presented. Moreover, the models discussed are applied to a community driven software solution, Apache Flume, to show how they can be integrated, seamlessly.

Published in: Springeropen Journal(Suthakar *et al. J Big Data (2016) 3:21)*

14. Spatial data extension for Cassandra NoSQL database

(Ben Brahim et.al, 2017) The proposed design is emergence of the NoSQL databases, like Cassandra, with their massive scalability and high availability encourages us to investigate the

management of the stored data within such storage system. They have harness the geohashing technique to enable spatial queries as extension to Cassandra query language capabilities while preserving the native syntax. The developed framework showed the feasibility of this approach where basic spatial queries are underpinned and the query response time is reduced by up to 70 times for a fairly large area.

Published in: Springeropen Journal(Ben Brahim *et al. J Big Data (2016) 3:11)*

15. Towards shortest path identification on large networks

(Selim and Zhan J, 2016) This paper is focused and proposes on application of the techniques of data reduction based on data nodes in large networks datasets by computing data similarity computation, maximum similarity clique (MSC) and then finding the shortest path in a quick manner due to the data reduction in the graph. As the number of vertices and edges tend to increase on large networks the aim of this article is to make the reduction of the network that will cause an impact on calculating the shortest path for a faster analysis in a shortest time.

Published in: Springeropen Journal(Selim and Zhan *J Big Data (2016) 3:10)*

16. Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream‐based processing

(Pekka Pääkkönen, 2016) The contribution of this paper is feasibility analysis of technologies for

stream-based processing of semi-structured data. Particularly, feasibility of a Big Data management system for semi-structured data (AsterixDB) will be compared to Spark streaming, which has been integrated with Cassandra NoSQL database for persistence. The study focuses on stream processing in a simulated social media use case (tweet analysis), which has been implemented to Eucalyptus cloud computing environment on a distributed shared memory multiprocessor platform. The results indicate that AsterixDB is able to provide significantly better performance both in terms of throughput and latency, when data feed functionality of AsterixDB is used, and stream processing has been implemented with Java. AsterixDB also scaled on the same level or better, when the amount of nodes on the cloud platform was increased. However, stream processing in AsterixDB was delayed by batching of data, when tweets were streamed into the database with data feeds.

17. Data stream clustering by divide and conquer approach based on vector model

(Khalilian et al, 2016) DCSTREAM method is proposed with regard to the mentioned issues to cluster big datasets using the vector model and k-Means divide and conquer approach. Experimental results show that DCSTREAM can achieve superior quality and performance as compare to STREAM and ConStream methods for abrupt and gradual real world datasets. Results show that the usage of batch processing in DCSTREAM and ConStream is time consuming compared to STREAM but it avoids further analysis for detecting outliers and novel micro-clusters.

18. Smart4Job: A Big Data Framework for Intelligent Job Offers Broadcasting Using Time Series Forecasting and Semantic Classification

(SidahmedBenabderrahmane et.al, 2016) This paper proposes the adequate job boards for the dissemination of a new job offer. Our system is based on a hybrid representation on a big data platform, which includes both a domain knowledge analysis and a temporal prediction model. The semantic classification of job boards requires a textual analysis using a controlled vocabulary. The time series analysis module is used to predict the best job

board for a given offer, using the history of the clicks. The answers of these modules are combined during the decision making process. The proposed system has been evaluated on real data, and preliminary results seem very promising.

B. General Papers

1. Mining Chinese social media UGC: a big‐data framework for analyzing Douban movie reviews

(Yang and Yecies, 2016) This paper is designed as an improved Apriori algorithm based on MapReduce is proposed for content-mining functions. An exploratory simulation of results demonstrates the flexibility and applicability of the proposed framework for extracting relevant information from complex social media data, knowledge which can in turn be extended beyond this niche dataset and used to inform producers and distributors of films, television shows, and other digital media content.

2. Vaccination allocation in large dynamic networks

(Zhan et.al, 2017) This paper presents a dynamic node vaccination solution that seeks to take advantage of these local recalculations. Network infections that are already in progress cause challenges to those officers trying to preserve those nodes not yet infected. Static solutions can take advantage of global knowledge of the network to produce quick and approximate answers for those members who should be vaccinated. In dynamic situations however, small changes can severely alter those static solutions making them irrelevant. Yet in dynamic situation sit cannot be known with certainty which small changes will affect the solution and those that will not. Computational resources are wasted recalculating a global solution for the entire network, when a local recalculation may be enough.

3. A data mining framework to analyze road accident data

(Kumar and Toshniwal, 2015)This paper proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further a trend analysis have also been performed for each clusters and EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

4 Big data, Big bang?

(Bughin, 2016) The paper proposes the performance test relies on a so-called trans-logarithmic production function, allowing for a more direct test of the complementarity between big data capital and big data labour investments; further, we have used a Heckman correction to adjust for the fact that companies investing in big data are generally more productive than their peers. We confirm and extend early results of a productivity impact from big data. We find that for the average of our sample, more productive firms are also faster adopters of big data than their industry peers (this explains 2.5% of productivity difference). Big data investments in labour and IT architecture are complements, with a total productivity growth effect of about 5.9%. Big data projects targeting customers and competitive intelligence domains bring slightly more performance than big data projects aimed at supply chain improvements.

5. Understanding big data themes from scientific biomedical literature through topic modeling

van Altena *et al. J,2016)* In this paper we pursue a better understanding of the term big data through a data-driven systematic approach using text analysis of scientific (bio)medical literature. We attempt to find how existing big data definitions are expressed within the chosen application domain. The resulting top-20 words per topic were annotated with the twelve big data themes by seven observers. The analysis of these annotations show that the themes proposed by De Mauro et al. are strongly expressed in the corpus. Furthermore, several of the most popular big data V's (i.e., volume, velocity, and value) also have a relatively high presence. Other V's introduced more recently (e.g. variability) were however hardly found in the 25 topics. These findings show that the current understanding of big data within the (bio)medical domain is in agreement with more general definitions of the term.

6. A new method of large‐scale short‐term forecasting of agricultural commodity prices: illustrated by the case of agricultural markets in Beijing

(Wu *et al. J, 2017)* This paper proposes a mixed model, which combines ARIMA model and PLS regression method based on time and space factors. This mixed model is able to obtain the forecasting results of weekly prices of agricultural commodities in different markets. Meanwhile, this paper sets up variables to measure the price changing trend based on the change of exogenous variables and prices, thus achieves the warning of daily price changes using neural networks. The model is tested with the data of several types of agricultural commodities and error analysis is made. The result shows that the mixed model is more accurate in forecasting agricultural commodity prices than each single model does, and has better accuracy in warning values. The mixed model, to some extent, forecasts the daily price changes of agricultural commodities.

7. Producing Linked Data for Smart Cities: The Case of Catania

(SergioConsoli et.al, 2016) This paper proposes a comprehensive data model for smart cities that integrates several data sources, including, geo-referenced data, public transportation, urban fault reporting, road maintenance and municipal waste collection. We show some novel ontology design patterns for modeling public transportation, urban fault reporting and road maintenance. Domain practitioners and general members of the public have been asked to play with the prototype, and fill out a survey with questions and feedbacks. A computational experiment has been also conducted to evaluate the performance of our data model in terms of practical scalability over increasing data and efficiency under complex queries. All produced data, models, prototype and questionnaire results are publicly accessible online.

8. Visual Analysis of Multiple Route Choices based on General GPS Trajectories

We develop a visual analytic system to help users handle the large-scale trajectory data, compare different route choices, and explore the underlying reasons. Specifically, the system consists of: 1. the interactive trajectory filtering which supports graphical

trajectory query; 2. the spatial view which gives an overview of all feasible routes extracted from filtered trajectories; 3. the factor visualizations which provide the exploration and hypothesis construction of different factors' impact on route choice behaviour, and the verification with an integrated route choice model. Applying to real taxi GPS dataset, we report the system's performance and demonstrate its effectiveness with three cases.

9. An Enhanced Visualization Method to Aid Behavioral Trajectory Pattern Recognition Infrastructure for Big Longitudinal Data

Big longitudinal data provide more reliable information for decision making and are common in all kinds of fields. Trajectory pattern recognition is in an urgent need to discover important structures for such data. Developing better and more computationally-efficient visualization tool is crucial to guide this technique. This paper proposes an enhanced projection pursuit (EPP) method to better project and visualize the structures (e.g. clusters) of big high-dimensional (HD) longitudinal data on a lower-dimensional plane. Unlike classic PP methods potentially useful for longitudinal data, EPP is built upon nonlinear mapping algorithms to compute its stress (error) function by balancing the paired weights for between and within structure stress while preserving original structure membership in the high-dimensional space. Specifically, EPP solves an NP hard optimization problem by integrating gradual optimization and non-linear mapping algorithms, and automates the searching of an optimal number of iterations to display a stable structure for varying sample sizes and dimensions. Using publicized UCI and real longitudinal clinical trial datasets as well as simulation, EPP demonstrates its better performance in visualizing big HD longitudinal data.

10. Large-Scale Data Pollution with Apache Spark

Because of the increasing volume of autonomously collected data objects, duplicate detection is an important challenge in today's data management. To evaluate the efficiency of duplicate detection algorithms with respect to big data, large test data sets are required. Existing test data generation tools, however, are either not able to produce large test data sets or are domain-dependent which limits their usefulness to a few cases. In this paper, we describe a new framework that can be used to pollute a clean, homogeneous and large data set from an arbitrary domain with duplicates, errors and in homogeneities. To prove its concept, we implemented a prototype which is built upon the cluster computing framework Apache Spark and evaluate its performance in several experiments.

C. Survey Paper

1. A survey of transfer learning

(Weiss et al, 2016) This survey paper formally defines transfer learning, presents information on current solutions, and reviews applications applied to transfer learning. Lastly, there is information listed on software downloads for various transfer learning solutions and a discussion of possible future research work. The transfer learning solutions surveyed are independent of data size and can be applied to big data environments.

2. Conceptualizing Big Social Data

(Olshannikova *et al. J,2017)*This paper proposed that emerging research field around these concepts would benefit from understanding about the very substance of the concept and the different viewpoints to it. With our review of earlier research, we highlight various perspectives to this multi-disciplinary field and point out conceptual gaps, the diversity of perspectives and lack of consensus in what Big Social Data means. Based on detailed analysis of related work and earlier conceptualizations, we propose a synthesized definition of the term, as well as outline the types of data that Big Social Data covers. With this, we aim to foster future research activities around this intriguing, yet untapped type of Big Data.

3. Big data privacy: a technological perspective and review

(Jain *et al. ,2016)*The goal of this paper is to provide a major review of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. This paper also presents recent techniques of privacy preserving in big data like hiding a needle in a haystack, identity based anonymization, differential privacy, privacypreserving big data publishing and fast anonymization of big data streams. This paper refer privacy and security aspects healthcare in big data. Comparative study between various recent techniques of big data privacy is also done as well.

4. Visualizing Big Data with augmented and virtual reality: challenges and research agenda

(Olshannikova et al,2015) This paper proposes a non-traditional approach is proposed: we discuss how the capabilities of Augmented Reality and Virtual Reality could be applied to the field of Big Data Visualization. We discuss the promising utility of Mixed Reality technology integration with applications in Big Data Visualization. Placing the most essential data in the central area of the human visual field in Mixed Reality would allow one to obtain the presented information in a short period of time without significant data losses due to human perceptual issues. Furthermore, we discuss the impacts of new technologies, such as Virtual Reality displays and Augmented Reality helmets on the Big Data visualization as well as to the classification of the main challenges of integrating the technology.

5. A survey of open source tools for machine learning with big data in the Hadoop ecosystem

(Landset *et al,2015)* This paper is intended to aid the researcher or professional who understands machine learning but is inexperienced with big data. In order to evaluate tools, one should have a thorough understanding of what to look for. To that end, this paper provides a list of criteria for making selections along with an analysis of the advantages and drawbacks of each. We do this by starting from the beginning, and looking at what exactly the term "big data" means. From there, we go on to the Hadoop ecosystem for a look at many of the projects that are part of a typical machine learning architecture and an understanding of how everything might fit together. We discuss the advantages and disadvantages of three different processing paradigms along with a comparison of engines that implement them, including Map Reduce, Spark, Flink, Storm, and H2O. We then look at machine learning libraries and frameworks including Mahout, MLlib, SAMOA, and evaluate them based on criteria such as scalability, ease of use, and extensibility. There is no single toolkit that truly embodies a one-size fits- all solution, so this paper aims to help make decisions smoother by providing as much information as possible and quantifying what the tradeoffs will be. Additionally, throughout this paper, we review recent research in the field using these tools and talk about possible future directions for toolkit-based learning.

IV. FUTURE RESEARCH

In general, 33 articles across 3 years period is relatively a high 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 methodology, algorithms experiments conducted on actual data, like amazon and douban Movie Review 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.

CONCLUSION

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.

References

[1] Chao Huang & et.al,2017) **.Scalable Uncertainty-Aware Truth Discovery in Big Data Social Sensing Applications for Cyber-Physical Systems** IEEE Transactions on Big Data Volume: PP, Issue: 99

[2] Yu Chan ; Andy Wellings ; Ian Gray ; Neil Audsley.(2017). A Distributed Stream Library for Java 8 IEEE Transactions on Big Data Volume: PP, Issue: 99

[3] Chunqiang Hu ; Wei Li ; Xiuzhen Cheng ; Jiguo Yu ; Shengling Wang ; Rongfang Bie. (2017).A Secure and Verifiable Access Control Scheme for Big Data Storage in Clouds IEEE Transactions on Big Data Volume: PP, Issue: 99

[4] Bo Wu ; Haiying Shen. (2017).**Exploiting Efficient Densest Subgraph Discovering Methods for BigData** IEEE Transactions on Big Data Volume: PP, Issue: 99

[5] (lWeiren Yu ; Jianxin Li ; Md Zakirul Alam Bhuiyan ; Richong Zhang ;Jinpeng Huai,2017) **Ring: Real-Time Emerging Anomaly Monitoring System over Text Streams** IEEE Transactions on Big Data Volume: PP, Issue: 99

[6] Hui Cui ; Robert H. Deng ; Yingjiu Li ; Guowei Wu. (2017). Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud IEEE Transactions on Big Data Volume: PP, Issue: 99

[7] Zhaojie Niu ; Bingsheng He ; Fangming Liu. (2017).JouleMR: Towards Cost-Effective and Green-Aware Data Processing Frameworks IEEE Transactions on Big Data Volume: PP, Issue: 99

[8] Yuxuan Jiang ; Zhe Huang ; Danny H. K. Tsang(2017). Towards Max-Min Fair Resource Allocation for Stream Big Data Analytics in Shared Clouds IEEE 9Transactions on Big Data Volume: PP, Issue: 99

[9] Dongfang Zhao ; Kan Qiao ; Zhou Zhou ; Tonglin Li ; Zhihan Lu ;Xiaohua Xu(2016)Toward Efficient and Flexible Metadata Indexing of Big Data Systems IEEE Transactions on Big Data Volume: PP, Issue: 99

[10] Theo Steininger1, Maksim Greiner, Frederik Beaujean, and Torsten Enlin. (2016). d2o: a distributed data object for parallel high‐performance computing in Python Springeropen Journal

[11] Honglei Zhang1, Jenni Raitoharju1, Serkan Kiranyaz and Moncef Gabbouj. (2016).Limited random walk algorithm for big graph data clustering Springeropen Journal(Zhang *et al. J Big Data (2016) 3:26)*

[12] Rupinder Singh & Puneet Jai Kaur. (2016). Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud Springer open Journal(Singh and Kaur *J Big Data (2016) 3:19)*

[13] 13Uthayanath Suthakar, Luca Magnoni, David Ryan Smith, Akram Khan and Julia Andreeva. (2016). An efficient strategy for the collection and storage of large volumes of data for computation Springeropen Journal

[14] Mohamed Ben Brahim, Wassim Drira, Fethi Filali and Noureddine Hamd.(2016). Spatial data extension for Cassandra NoSQL database Springeropen Journal*)*

[15] Haysam Selim and Justin Zhan. (2016). Towards shortest path identification on large networks Springer open Journal(Selim and Zhan *J Big Data (2016) 3:10)*

[16] Pääkkönen *J Big Data (2016)* Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream‐based processing Springer open Journal (Pääkkönen *J Big Data (2016) 3:6)*

[17] Madjid Khalilian , Norwati Mustapha and Nasir Sulaiman. (2016).Data stream clustering by divide and conquer approach based on vector model Springeropen Journal

[18] Sidahmed Benabderrahmanea Nedra Melloulia, Myriam Lamollea, Patrick Paroubek(2017)Smart4Job: A Big Data Framework for Intelligent Job Offers Broadcasting Using Time Series Forecasting and Semantic Classification 2017 Elsevier Inc.

[19] Jie Yang and Brian Yecies. (2016). Mining Chinese social media UGC: a big‐data framework for analyzing Douban movie reviews Springeropen Journal

[20] Justin Zhan, Timothy Rafalski, Gennady Stashkevich and Edward Verenich. (2017). Vaccination allocation in large dynamic networks Springeropen Journal(Zhan *et al. J Big Data (2017) 4:2)*

[21] Sachin Kumar and Durga Toshniwal. (2015). *A* data mining framework to analyze road accident data Springer open Journal(Kumar and Toshniwal *Journal of Big Data)*

[22] Jacques Bughin(2016) Big data, Big bang?Springeropen Journal

[23] Allard J. van Altena*, Perry D. Moerland, Aeilko H. Zwinderman and Sílvia D. Olabarriaga(2016) Understanding big data themes from scientific biomedical literature through topic modeling Springer open Journal(van Altena *et al. J Big Data (2016) 3:23)*

[24] Haoyang Wu1* , Huaili Wu1, Minfeng Zhu1, Weifeng Chen2 and Wei Chen1(2017)A new method of large‐scale short‐term forecasting of agricultural commodity prices: illustrated by the case of agricultural markets in Beijing Springeropen Journal

[25] Karl Weiss, Taghi M. Khoshgoftaar and DingDing Wang. (2016). A survey of transfer learning Springeropen Journal

[26] Ekaterina Olshannikova1 , Thomas Olsson, Jukka Huhtamäki and Hannu Kärkkäinen. (2017). Conceptualizing Big Social Data Springeropen Journal*)*

[27] Priyank Jain , Manasi Gyanchandani and Nilay Khare. (2016).Big data privacy: a technological perspective and review Springeropen Journal
[28] Haoyang Wu1* , Huaili Wu1, Minfeng Zhu1, Weifeng Chen2 and Wei Chen.(2017)Visualizing Big Data with augmented and virtual reality: challenges and research Agenda Springeropen Journal
[29] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter* and Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem Springeropen Journal
[30] Sergio Consolia,b,, ValentinaPresuttia, Diego Reforgiato Recuperoa, c, AndreaG. Nuzzolesea, SilvioPeronia, d, Misael Mongiovi'a, Aldo Gangemia (2017)Producing Linked Data for Smart Cities: The Case of Catania Elsevier Inc.
[31] Min Lu ; Chufan Lai ; Tangzhi Ye ; Jie Liang ; Xiaoru Yuan(2017)Visual Analysis of Multiple Route Choices based on General GPS Trajectories IEEE Transactions on Big Data (Volume: PP, Issue: 99)
[32] Hua Fang ; Zhaoyang Zhang(2017)An Enhanced Visualization Method to Aid Behavioral Trajectory Pattern Recognition Infrastructure for Big Longitudinal Data. IEEE Transactions on Big Data (Volume: PP, Issue: 99)
[33] Kai Hildebrandt ; Fabian Panse ; Niklas Wilcke ; Norbert Ritter(2017)Large-Scale Data Pollution with Apache Spark IEEE Transactions on Big Data (Volume: PP, Issue: 99)
[34] Big data analytics: a text mining-based literature analysis