An Extensive Literature Review on Big Data & Quality Problem

Prof. Manoj Niwariya, Prof. Piyush Dubey

Abstract: Data intensive computing deals with computational methods and architectures to analyze and discover intelligence in huge volumes of data generated in many application domains. Cloud computing is a major aim for data intensive computing, as it allows scalable pro cessing of massive amount of data. It is a promising next-generation computing paradigm given its many advantages such as scalability, reliability, elasticity, high availability, and low cost. Here the problem is huge data and is growing exp potentially. The data used by various organizations and institutions is increasing at a rapid rate. In today's world, organizations have to pro cess pet bytes of data. The traditional database management system fails to process such large amount of data. So, need to find out an effective way of approach for handling and pro cessing such large amount of data which gives rise to big data problem. Size or volume of the data is not only the single criteria to classify big data, To keep in mind the type of the data that is whether data is structured or semi structured or unstructured.

Key words: Big Data, Data Quality, Returns to Scale

I. INTRODUCTION

The rapid increase in size of data has now emerged as the biggest problem for many large organization and institutions. Let us take an example of Go ogle or Face book which deals with peta byte of data daily. In Face book, starting from creating profile to like, comment, sending friend request, adding visited places, messages every details has to b e stored in database so that whenever needed by the user, it can b e available. Same is the case with Go ogle which has to store all the information regarding the user mail account, the search items done by the user, the last visited sites by the user etc. Similarly all companies deals with vast amount of data which given rise to big data problem.

Big data has emerged as a challenge for the researchers. Their main focus is how to handle such vast amount of data, how to pro cess those large amount of data within an acceptable time limit. The traditional database management system also can b e used but this is not a feasible option to tackle the big data problem. Generally the traditional database management system has a size cap as how much data it can pro cess. If this is the case, then it is absolutely impossible to overcome the problems faced by the big data which grows exp potentially. Moreover, some database supp orts infinite data but the data accessing time and pro cessing time is extremely large which is not feasible. So have to find an efficient solution for these problems or challenges faced by big data problem.

An incredible "data deluge" is currently drowning the world. Data sources are everywhere, from Web 2.0 and user-generated content to large scientific experiments, from social networks to wireless sensor networks. This massive amount of data is a valuable asset in the information society.

Data analysis is the process of inspecting data in order to extract useful information. Decision makers commonly use this information to drive their choices. The quality of the information extracted by this process greatly benefits from the availability of extensive datasets.

The Web is the biggest and fastest growing data repository in the world. Its size and diversity make it the ideal resource to mine for useful information. Data on the Web is very diverse in both content and format. Consequently, algorithms for Web mining need to take into account the specific characteristics of the data to be efficient.

As we enter the "petabyte age", traditional approaches for data analysis begin to show their limits. Commonly available data analysis tools are unable to keep up with the increase in size, diversity and rate of change of the Web. Data Intensive Scalable Computing is an emerging alternative technology for large scale data analysis. DISC systems combine both storage and computing in a distributed and virtualized manner. These systems are built to scale to thousands of computers, and focus on fault tolerance, cost effectiveness and ease of use.

Data on the Web is often produced as a by product of online activity of the users, and is sometimes referred to as data exhaust. This data is silently collected while the users are pursuing their own goal online, e.g. query logs from search engines, co-buying and co-visiting statistics from online shops, click through rates from news and advertisings, and so on.

This process of collecting data automatically can scale much further than traditional methods like polls and surveys. For example it is possible to monitor public interest and public opinion by analyzing collective click

behaviour in news portals, references and sentiments in blogs and micro-blogs or query terms in search engines.

Let us now more precisely define Web mining. Web mining is the application of data mining techniques to discover patterns from the Web. According to the target of the analysis at hand, Web mining can be categorised into three different types: Web structure mining, Web content mining and Web usage mining.

Web structure mining mines the hyperlink structure of the Web using graph theory. For example, links are used by search engines to find important Web pages, or in social networks to discover communities of users who share common interests.

Web content mining analyzes Web page contents. Web content mining differs from traditional data and text mining mainly because of the semi-structured and multimedia nature of Web pages. For example, it is possible to automatically classify and cluster Web pages according to their topics but it is also possible to mine customer product reviews to discover consumer sentiments.

Web usage mining extracts information from user access patterns found in Web server logs, which record the pages visited by each user, and from search patterns found in query logs, which record the terms searched by each user. Web usage mining investigates what users are interested in on the Web.

Mining the Web is typically deemed highly promising and rewarding. However, it is by no means an easy task and there is a flip side of the coin: data found on the Web is extremely noisy.

Big Data Problem

Big Data has developed on the grounds that there are using much kind of technologies which uses very large amount of data. One present challenge associated with big data is the trouble we face while working with it by utilizing traditional statistics/visualization packages and relational databases. So, need a “greatly parallel programming running on many numb er of servers”. The other different difficulties confronted in big data administration include adaptability, unstructured data, availability, accessibility, constant monitoring, fault tolerance and many more. Notwithstanding varieties in the measure of data put away in different sectors, the sorts of data produced and put away i.e., whether the data encodes pictures, audio, video, or content/numeric data additionally vary notably from industry to industry. So, need a mechanism to overcome such big data problem.

- Volume: Hadoop gives system to scale out evenly to self-assertively substantial information sets to address such huge volume of information.
- Velocity: Hadoop handles incensed rate of approaching information from expansive framework.
- Variety: Hadoop bolsters complex occupations to handle any mixture of unstructured information.
- Variability: Hadoop handles data which is complex and varies according to size.

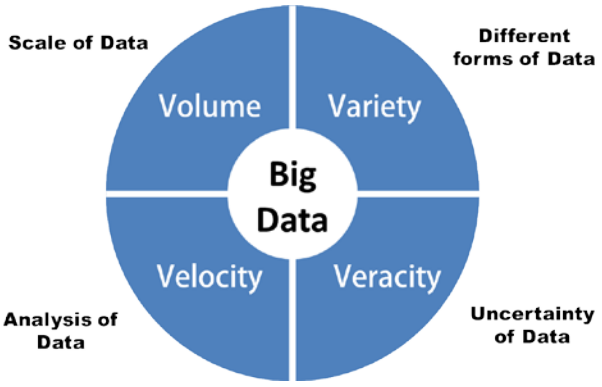

Fig. 1: 4 Vs of Big Data

The key attributes of Hadoop are that it is redundant and reliable, that is if you lose a machine due to some failure, it automatically replicates your data immediately without the operator having to do anything, it is extremely powerful in terms of data access and is preliminary batch processing centric and makes it easier to distributed applications using MapReduce software paradigm. Moreover it runs on commodity hardware which cuts off the cost of buying special expensive hardware and RAID systems.

HDFS

HDFS stands for Hadoop Distributed File System which is a distributed file system. It is designed so that very large size files can b e stored with streaming data access patterns, running on clusters of commodity hardware.

Streaming data access

HDFS is designed with keeping in mind that it can use it like reading many times but writing only once. After accepting a dataset, have to analyze it thoroughly to understand the hidden pattern. So they are dealing here how to read the entire data set not just part of it.

Commodity

Hadoop can be made run on low cost, easily available systems. Usually run it on no des of commodity system. As, cluster failure can happen, Hadoop is designed in such a way that it can overcome such failure easily.

II. LITERATURE SURVEY

SR. NO.	TITLE	AUTHORS	YEAR	METHODOLOGY
1	Big data, big data quality problem	D. Becker, T. D. King and B. McMullen	2015	Key findings of this study reinforce that the primary factors affecting Big Data reside in the limitations and complexities involved with handling Big Data while maintaining its integrity
2	High quality clustering of big data and solving empty-clustering problem with an evolutionary hybrid algorithm	J. Karimov and M. Ozbayoglu	2015	The model is implemented using a Hadoop Mapreduce algorithm for achieving scalability when faced with a Big Data clustering problem.
3	Next Big Thing in Big Data: The Security of the ICT Supply Chain	T. Lu, X. Guo, B. Xu, L. Zhao, Y. Peng and H. Yang	2013	Introduces several famous international models both on physical supply chain and ICT supply chain.
4	Data quality: The other face of Big Data	B. Saha and D. Srivastava	2014	This tutorial presents recent results that are relevant to big data quality management, focusing on the two major dimensions of (i) discovering quality issues from the data itself, and (ii) trading-off accuracy vs efficiency, and identifies a range of open problems for the community.
5	Cross-platform aviation analytics using big-data methods	T. Larsen	2013	Key aviation data sets for operational analytics, presents a methodology for application of big-data analysis methods to operational problems, and offers examples of analytical solutions using an integrated aviation data warehouse.

D. Becker, T. D. King and B. McMullen, [1] A USAF sponsored MITRE research team undertook four separate, domain-specific case studies about Big Data applications. Those case studies were initial investigations into the question of whether or not data quality issues encountered in Big Data collections are substantially different in cause, manifestation, or detection than those data quality issues encountered in more traditionally sized data collections. The study addresses several factors affecting Big Data Quality at multiple levels, including collection, processing, and storage. Though not unexpected, the key findings of this study reinforce that the primary factors affecting Big Data reside in the limitations and complexities involved with handling Big Data while maintaining its integrity. These concerns are of a higher magnitude than the provenance of the data, the processing, and the tools used to prepare, manipulate, and store the data. Data quality is extremely important for all data analytics problems. From the study's findings, the "truth about Big Data" is there are no fundamentally new DQ issues in Big Data analytics projects. Some DQ issues exhibit return-s-to-scale effects, and become more or less pronounced in Big Data analytics, though. Big Data Quality varies from one type of Big Data to another and from one Big Data technology to another.

J. Karimov and M. Ozbayoglu, [2] Achieving high quality clustering is one of the most well-known problems in data mining. k-means is by far the most commonly used clustering algorithm. It converges fairly quickly, but achieving a good solution is not guaranteed. The clustering quality is highly dependent on the selection of the initial centred selections. Moreover, when the number of clusters increases, it starts to suffer from "empty clustering". The motivation in this study is two-fold. Authors not only aim at improving the k-means clustering quality, but at the same time not being affected by the empty cluster issue. For achieving this purpose, authors developed a hybrid model, H(EC)2S, Hybrid Evolutionary Clustering with Empty Clustering Solution. Firstly, it selects representative points to eliminate Empty Clustering problem. Then, the hybrid algorithm uses only these points during centroid selection. The proposed model combines Fireworks and Cuckoo-search based evolutionary algorithm with some centroid-calculation heuristics. The model is implemented using a Hadoop Mapreduce algorithm for achieving scalability when faced with a Big Data clustering problem. The advantages of the developed model are particularly attractive when the amount, dimensionality and number of cluster parameters tend to increase. The results indicate that considerable clustering quality performance improvement is achieved using the proposed model.

T. Lu, X. Guo, B. Xu, L. Zhao, Y. Peng and H. Yang,[3] In contemporary society, with supply chains becoming more

and more complex, the data in supply chains increases by means of volume, variety and velocity. Big data rise in response to the proper time and conditions to offer advantages for the nodes in supply chains to solve previously difficult problems. For any big data project to succeed, it must first depend on high-quality data but not merely on quantity. Further, it will become increasingly important in many big data projects to add external data to the mix and companies will eventually turn from only looking inward to also looking outward into the market, which means the use of big data must be broadened considerably. Hence the data supply chains, both internally and externally, become of prime importance. ICT (Information and Telecommunication) supply chain management is especially important as supply chain link the world closely and ICT supply chain is the base of all supply chains in today's world. Though many initiatives to supply chain security have been developed and taken into practice, most of them are emphasized in physical supply chain which is addressed in transporting cargos. The research on ICT supply chain security is still in preliminary stage. The use of big data can promote the normal operation of ICT supply chain as it greatly improve the data collecting and processing capacity and in turn, ICT supply chain is a necessary carrier of big data as it produces all the software, hardware and infrastructures for big data's collection, storage and application. The close relationship between big data and ICT supply chain make it an effective way to do research on big data security through analysis on ICT supply chain security. This paper first analyzes the security problems that the ICT supply chain is facing in information management, system integrity and cyberspace, and then introduces several famous international models both on physical supply chain and ICT supply chain. After that the authors describe a case of communication equipment with big data in ICT supply chain and propose a series of recommendations conducive to developing secure big data supply chain from five dimensions.

B. Saha and D. Srivastava, [4] in the Big Data era, data is being generated, collected and analyzed at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Recent studies have shown that poor quality data is prevalent in large databases and on the Web. Since poor quality data can have serious consequences on the results of data analyses, the importance of veracity, the fourth `V' of big data is increasingly being recognized. In this tutorial, highlight the substantial challenges that the first three `V's, volume, velocity and variety, bring to dealing with veracity in big data. Due to the sheer volume and velocity of data, one needs to understand and (possibly) repair erroneous data in a scalable and timely manner. With the variety of data, often from a diversity of sources, data quality rules cannot be specified a priori; one needs to let the “data to speak for itself” in order to discover the semantics of the data. This tutorial presents recent results that are relevant to big data quality management, focusing on the two major dimensions of (i) discovering quality issues from the data itself, and (ii) trading-off accuracy vs efficiency, and identifies a range of open problems for the community.

T. Larsen,[5] This paper identifies key aviation data sets for operational analytics, presents a methodology for application of big-data analysis methods to operational problems, and offers examples of analytical solutions using an integrated aviation data warehouse. Big-data analysis methods have revolutionized how both government and commercial researchers can analyze massive aviation databases that were previously too cumbersome, inconsistent or irregular to drive high-quality output. Traditional data-mining methods are effective on uniform data sets such as flight tracking data or weather. Integrating heterogeneous data sets introduces complexity in data standardization, normalization, and scalability. The variability of underlying data warehouse can be leveraged using virtualized cloud infrastructure for scalability to identify trends and create actionable information. The applications for big-data analysis in airspace system performance and safety optimization have high potential because of the availability and diversity of airspace related data. Analytical applications to quantitatively review airspace performance, operational efficiency and aviation safety require a broad data set. Individual information sets such as radar tracking data or weather reports provide slices of relevant data, but do not provide the required context, perspective and detail on their own to create actionable knowledge. These data sets are published by diverse sources and do not have the standardization, uniformity or defect controls required for simple integration and analysis. At a minimum, aviation big-data research requires the fusion of airline, aircraft, flight, radar, crew, and weather data in a uniform taxonomy, organized so that queries can be automated by flight, by fleet, or across the airspace system.

X. L. Dong and D. Srivastava, [6] The Big Data era is upon us: data is being generated, collected and analyzed at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of Big Data. BDI differs from traditional data integration in many dimensions: (i) the number of data sources, even for a single domain, has grown to be in the tens of thousands, (ii) many of the data sources are very dynamic, as a huge amount of newly collected data are continuously made available, (iii) the data sources are extremely heterogeneous in their structure, with considerable variety even for substantially similar

entities, and (iv) the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This seminar explores the progress that has been made by the data integration community on the topics of schema mapping, record linkage and data fusion in addressing these novel challenges faced by big data integration, and identifies a range of open problems for the community.

III. CONCLUSION

Today there has been an enormous data explosion due to the new trend and paradigm such as social networks and cloud computing. Moreover, data has been getting more diverse, more complex, and less structured and it also needs to be processed rapidly. This situation has caused a new challenge for the traditional technologies such as relational databases and scale-up infrastructures

REFRENCES

[1] D. Becker, T. D. King and B. McMullen, "Big data, big data quality problem," Big Data (Big Data), 2015 IEEE International Conference on, Santa Clara, CA, 2015, pp. 2644-2653.

[2] J. Karimov and M. Ozbayoglu, "High quality clustering of big data and solving empty-clustering problem with an evolutionary hybrid algorithm," Big Data (Big Data), 2015 IEEE International Conference on, Santa

[3] T. Lu, X. Guo, B. Xu, L. Zhao, Y. Peng and H. Yang, "Next Big Thing in Big Data: The Security of the ICT Supply Chain," Social Computing (SocialCom), 2013 International Conference on, Alexandria, VA, 2013, pp. 1066-1073.

[4] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," 2014 IEEE 30th International Conference on Data Engineering, Chicago, IL, 2014, pp. 1294-1297.

[5] T. Larsen, "Cross-platform aviation analytics using big-data methods," Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013, Herndon, VA, 2013, pp. 1-9.

[6] X. L. Dong and D. Srivastava, "Big data integration," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, Brisbane, QLD, 2013, pp. 1245-1248.

[7] C. F. Olson, “Parallel algorithms for hierarchical clustering,” Parallel computing, vol. 21, no. 8, pp. 1313–1325, 1995.

[8] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[9] R. Lammel, “Google’s mapreduce programming model—¨revisited,” Science of computer programming, vol. 70, no. 1, pp. 1–30, 2008.

[10] T. White, Hadoop: the definitive guide: the definitive guide. ” O’Reilly Media, Inc.”, 2009.

[11] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst, “Haloop: Efficient iterative data processing on large clusters,” Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 285–296, 2010.

[12] S. Owen, R. Anil, T. Dunning, and E. Friedman, Mahout in action. Manning Shelter Island, 2011.

[13] R. M. Esteves, R. Pais, and C. Rong, “K-means clustering in the cloud–a mahout test,” in Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on. IEEE, 2011, pp. 514 519.

[14] R. M. Esteves and C. Rong, “Using mahout for clustering wikipedia’s latest articles: a comparison between k-means and fuzzy c-means in the cloud,” in Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on. IEEE, 2011, pp. 565–569.

[15] Y. Tan and Y. Zhu, “Fireworks algorithm for optimization,” in Advances in Swarm Intelligence. Springer, 2010, pp. 355– 364.