A Comparative Study of Implementation Challenges in Various Systems of Data Science

[1]Ch.Neelima, [2]P.Sai Srujana, [3]O.Kiran Kumar
[1,2,3]Department of Computer Science, St.Joseph's Degree College, Kurnool,
Andhra Pradesh, India.

Received: May 20, 2018 Accepted: July 15, 2018

ABSTRACT *s the software and hardware technologies has increased dramatically in its scope of managing huge amount of data. The term Data Science refers to an area where statistical, domain-related and technical knowledge is required to perform a proper data analysis. Data Science provides a broader area of research and analysis for people not only belongs to technology but also for mathematicians. A major part of Data Science is performing data analysis by applying a required range of skills. Such skills are further used in designing and implementing related areas of architecture, acquisition, analysis and achieving of data.*

In this research article, the authors have explored some of the different issues to be considered while implementing data science.

Keywords:*Data Science, Domain-Related Knowledge, Software, Statistical Data Analysis.*

I. Introduction

Data Science is an analytical process of extracting unstructured data from huge amount of data. The concept of data science also involves the concept of data or information mining by perspective investigation otherwise known as information disclosure. Here unstructured data can be a message, photograph and any other substance produced by the client. Data Science deals with a frequent oblique dealing of measuring data and performing calculations to extract bits of knowledge from an information.

The application of combining some logic with data does not ensure a reasonable outcome from a given amount of data. It is a process of understanding, extracting, visualizing and communication a very important skill. The major complimentary factor is understanding the data and extracting a valuable outcome from it.

Data Science is a field of using planning and research issues in various spaces of improvement, discovery, applying strategies and so on. Data Science is an area where researchers utilize it to discover and translate data from various sources by standing equipment, perfect programming and forward speed imperatives.

The process of applying data science can be summed as raw data collection, processing, cleaning, exploring data analysis, applying models and algorithms, communicate with visual reports and data production.

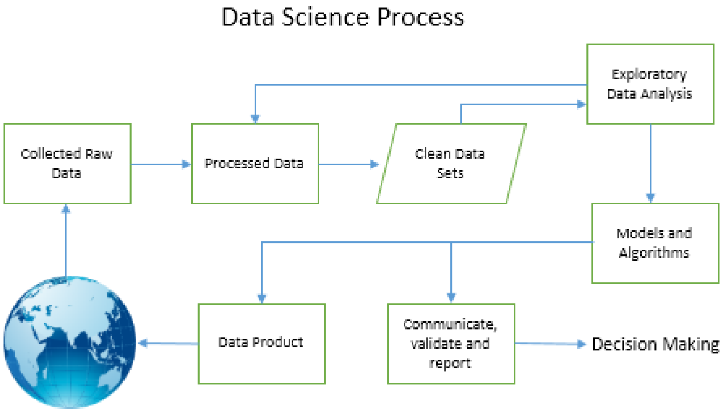

Figure 1: Architecture of Data Science Process

1. Data Wrangling: It is a manual process of data collection, conversion or mapping raw data into more convenient forms by using semi-automated tools. This process also includes sampling, sorting and merging.
2. Data Analysis: It is a process of assessing and demonstrating data by finding useful data, suggested the conclusions and allowing decision-making. Finally, certain statistical algorithms are applied and implements machine learning to extract useful and meaningful information from the huge amount of data.

3. Conveying Data: It is a method of transforming mathematical and statistical drawing of data from one form to another so that it can be understood and can able to interpret easily. Conveying data is another but empowering the development from one perspective to another like a beginner to an expert. It is also a process of making new technologies appear as an integral part of a system.

II. Literature Review

In [2]. W.Tukey specified the term 'bit' which is used to make an argument that emphasis the need of data analysis and confirmation side-by-side.

In [3], Peter Naur has made a survey on contemporary information processing models on wide range of different applications. He also suggested a definition to data science saying that it is a process of dealing with data and its relationship between the other areas of science.

In 1977, IASC was established with an objective of linking traditional methods of statistics, conventional computer technologies and domain-specific knowledge to convert data into various forms of information and knowledge.

In [4], Gregory Piatetsky has made some research on knowledge discovery and information mining for better understanding of data science and its applications.

In [5], Usama Fayyad has demonstrated the data relevance and discovery of data in databases.

In [6], William S.Cleveland proposed an approach to enlarge the main areas of work in the field of technology and statistics.

In [7], Leo Breiman proposed a statistical modelling of data in two different cultures as stochastic and algorithmic data models. These models made an evolution and gaining popularity in modern world of information by a data science.

III. Traditional and Data Science Systems

Data Science systems are more traditional and conventional information analysis and processing. These systems can interact as an ultimate source for data analysis. Data Science systems share few common aspects with different areas of traditional system. Therefore, it is possible to apply the quality view of a product and can consider the traditional issues of data quality in new data science systems. Simultaneously, these systems might provide an insight into facing the new challenges by data science systems.

A. Relational Database Management System

Data processing in traditional systems can be implemented using relational database management systems. These systems process the data based on entity relationship model and Codd's suggested relational model. RDBMS can provide platforms for various applications which need the data transaction processing.

B. Distributed Database Management System

The integrated view of RDBMS and other similar data systems leads to a concept of distributed databases and its handing as management system. A distributed DBMS uses several alternative architectures to provide simple and uniform interface for data accessed from various traditional database management system.

C. Data Warehousing

Data Warehousing is a parallel approach of implementing distributed DBMS approach. It defines an approach of data processing as a culmination of evolution series of processing information and decision supporting systems. Data Warehousing defines its approach as subject-oriented, non-volatile, integrated and time-variant data collection which supports decision making.

IV. Quality Issues of Data Science and Solutions

Data Science deals with the data extracted from various and different sources. Because of this data merging and integration is a big issue. Many technologies dealing with data sciences support no-schema databases. That means the schema information will be embedded in a specific format and every record basis.

A. Duplication: As acquiring of data is a regular process there is a chance of data duplication. This may occur due to the regular extraction of data from the same sources for more than once.

B. Fragmentation: Unlike a distributed DBMS, there may be a little fragmentation of different pieces of data is available with only visual form of documentation. It is complicated to reconstruct the plans and programs to fill the gap or overlaps between the fragments.

C. Location Issues: Location transparency is also one of the major issue of data science implementation because of standards for data distribution and specifying the naming details of data sets as a part of deployment.

D. Integration: As data science systems does not have a life cycle model to process consistent integration. This situation cannot be hidden due to lack of communication and coordination among data sets. The integration issue may also lead to suffering of quality in analysis when it is combined with duplication and location issues.

Strategy based potential solution for solving duplication, fragmentation, location and integration issues is to maintain meta data integration. This is becoming difficult to manage complexities which can degrade the performance of data science quality. Some of the strategy based potential solutions are addressing the issues like location, fragmentation, duplication/replication, and integration of data.

V.Future Scope

Every issue of quality cannot be addressed within a single scope. There is a need of detailed study to design the meta data models in the strategies where data science has not been created. Exploring various tools to enable the meta data capturing, loading, and to analysis. The more in-depth research is needed to address the real world data science issue with some potential strategy point.

VI. Conclusion

This article has considered for potential quality issues of data sciences by focusing on data input sets, processes used for transformation and its integration. The strategies presented in this article has begun by considering traditional quality models and the primary issue identified through this research is incoordination of meta data integration and its management. Solutions to the issue were suggested which requires few modifications of development processes and management of meta data.

REFERENCES

1. Data Mining and Knowledge Discovery. ISSN: 1384-5810 (print version). ISSN: 1573-756X (electronic version). Journal no. 10618
2. Mining Data for Nuggets of Knowledge Dec 10, 1999 Mining Data for Nuggets of Knowledge. http://knowledge.wharton.upenn.edu/article/mining-data-for-nuggets-of-knowledge/
3. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics William S. Cleveland Statistics Research, Bell Labs.
4. http://www.stat.purdue.edu/~wsc/papers/datascience.pdf
5. Statistical Modelling: the two cultures Leo Breiman. Statistical Science. Vol. 16 No.3 (August 2001) 199-215.
6. Davenport, Thomas H. (January 1, 2006). "Competing on Analytics". Harvard Business Review
7. Glossary of Terms. Machine Learning - Special issue on applications of machine learning and the knowledge discovery process archive. Volume 30 Issue 2-3, Feb/March, 1998. Pages 271-274
8. Bolton, R. & Hand, D. (2002). Statistical Fraud Detection: A Review (With Discussion). Statistical Science 17(3): 235–255.
9. Neural data mining for credit card fraud detection. Brause, R.; Langsdorf, T.; Hepp, M. Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on. Publication Year: 1999, Pages: 103-106
10. ReinhardKlette (2014). Concise Computer Vision. Springer. ISBN 978-1-4471-6320-6.
11. Hutchins, W. John; Somers, Harold L. (1992). An Introduction to Machine Translation. London: Academic Press. ISBN 0-12-362830-X.
12. Akshi Kumar, Teeja Mary Sebastian. Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012. ISSN (Online): 1694-0814.
13. Raul Isea. The Present-Day Meaning of the Word Bioinformatics, Global Journal of Advanced Research, 2015. Vol-2, Issue-1 PP. 70-73. ISSN: 2394-5788.