International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 8, No.3 (May-2019)

http://dx.doi.org/10.12785/ijcds/080306

Ontological Practice for Big Data Management

Kamal Uddin Sarker[1], Aziz Bin Deraman [2], Raza Hasan[3] and Ali Abbas[4]

[1] *School of Informatics and Applied Mathematics, University Malaysia Terengganu, Terengganu, Malaysia.*
[2] *School of Informatics and Applied Mathematics, University Malaysia Terengganu, Terengganu, Malaysia.*
[3]*Department of Information Technology, Malaysia University of Science and Technology, Selangor, Malaysia.*
[4]*Department of Computing, Middle East College, Muscat, Oman.*

Received 22 Jan 2019, Revised 18 Apr. 2019, Accepted 28 Apr. 2019, Published 1 May 2019

Abstract: A common inward research area *big data* is being introduced due to the rapidly increasing of information systems in all social and business organizations. The global connectivity in the virtual environment become smarter and specifying physical objects' status identity by the internet of things that leads to generate vast amount of information in the cloud. On the other side, the traditional database applications faced problem in data integration processes from multiple sources, where individual source has unique platform and architecture; hence big data management technique become complex. The higher dimensionality of big data till now not yet defined with measurable technique. A mathematical model is developed in this paper to define big data which overcome the existence of conceptual definition and ontological approach is being developed for managing the big data challenges. Finally, the ontology is theoretically evaluated based on the defined mathematical model and it proposes integrity based big data architecture for enterprise.

Keywords: Big Data, Big Data Definition, Big Data Complexity, Ontology, Big Data Modeling, Big Data Management.

1. INTRODUCTION

Till now big data is a conceptual and refers to a complex and huge amount of data that fails to handle by systematic database applications due to the high progressive rate of the internet users in individual and organization level. The common sources of big data is website, social media, sensor data and internet of things [1]. But this paper consider only for the problem that faced by organization when they have to marge big data from multiple source and big data generation is illustrated in figure-1. Figure-1 has shown the scenario of big data that faces by enterprise level. Enterprise Resource Planning (ERP) system includes a business process in an organization for cost-effectiveness [2] and mostly they ignore large dataset about products: product picture, marketing description and complementary product data [3]. Generally, this data is managed by others type of information systems: product information systems [4] that includes classification of products, translation data management and media of data (e.g. catalogue, brochure and technical dataset for knowledge generation). Same way individual system is required for accessing data like customer relationship management system [5] and content management system [6] in database approach. When different systems are required to access from an ERP it is become more complex and challenging which increases cost for data analysis. Purposes of the information system in an organization is to extract knowledge from current and historical data for improving the strategy of an organization besides ensuring effective working environment. Statistical report creation and report summary visualization by dashboard presentation is an important function. Data transforming from multiple sources and multiple formats make the system more complicated too.

In a traditional relational database landscape system separates data analysis and data transaction system [7] because both works in separate functionalities, requirements and characteristics. Moreover, a relational database works on a single system and depend on predefined structure with an expected amount of data that is predicted during design, so the performance is evaluated during testing and execution; also upgraded systems' complexity comparatively easier than big data applications. The understanding of big data analysis is described by figure-1 (n number of sources having different data types and file format: text, tabular, image, website, tabular, doc, pdf, etc.) and its' complexity is summarized in figure-2 with v-characteristics.

E-mail: ku_sarker@yahoo.com, a.d@umt.edu.my, raza.hasan@pg.must.edu.my, aabbas@mec.edu.om

Organization structure and functionality must be increased due to inclusion of data from multiple companies in a single platform that is commonly tend in this globalization world of business. So now a days the organization that is accessing traditional database system there is no guarantee that in near future it will not face big data challenges.

Big data is measured with certain Vs which is illustrated in figure-2; where big data referees to a collection of huge amount complex data that featured with the dimensions: volume, verity and velocity [8]; where "volume" adjudicators the steadily increasing of the data in a system; the types of data and its format is indicated by the terminology "verity" like: tabular, hierarchy,

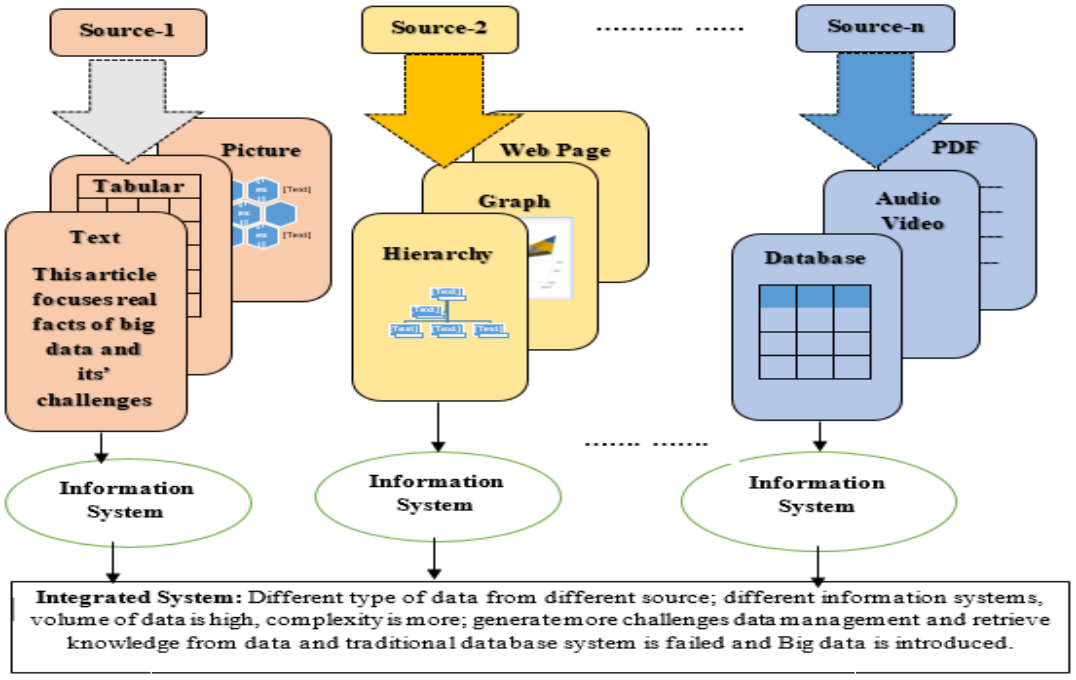

Figure.1. Big data architecture

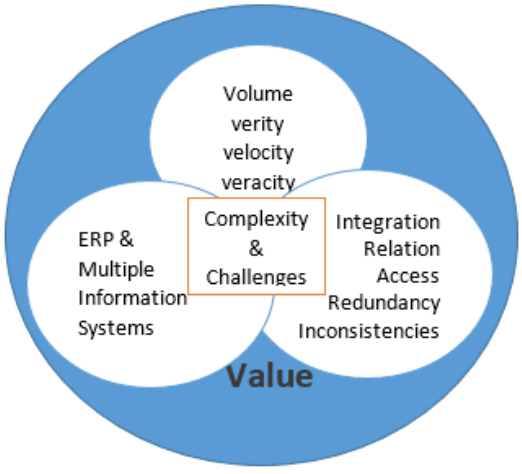

Figure.2. Big data-driven issues (Cause-effect)

while velocity describes the ways and techniques for accessing data and effectiveness of handling it, for example: how fast data can be retrieved?

Recently International Business Machines Cooperation (IBM) focuses on measuring the reliability of data science by the terminology "veracity" [9] when data arrives from multiple sources. Most of the case the sets are incomplete and not in a unique standard, so traditional database applications are failed to process the dataset. When the dimensions and complexity of the dataset is increased and it faces new challenges for developing a new method to manage, handle, analyze, store, query, share, visualize and so on operations [10]. Data has value when it can be used to make decision and it is possible when we can make it meaningful after processing. Features of data, operations and applications are represented by three circles where intersection considered the complexity and challenges in figure-2 and finally if you can manage this data the possible knowledge is valuable for decision making.

$$f(\text{volume})=\begin{cases}0\ ; coverage\ by\ traditional\ database\ applications\\ 1\ ;\ low\ performance\ or\ not\ possible\ by\ traditional\ approach\end{cases} \quad \text{[i]}$$

$$f(\text{verity})=\begin{cases}0;\ unique\ structure\ for\ same\ field\ in\ multiple\ data\ set\\ number\ \geq 1; degree\ of\ structure\end{cases} \quad \text{[ii]}$$

$$f(\text{velocity})=\begin{cases}0; reasonable\ time\ for\ accessing\ information\\ 1; higher\ performance\ machine\ required\end{cases} \quad \text{[iii]}$$

$$f(\text{veracity})=\begin{cases}0;\ complete\ and\ reliable\\ medium;\ incomplete\ and\ reliable\\ higher;\ complete\ or\ incomplete\ and\ not\ reliable\end{cases} \quad \text{[iv]}$$

$$F(\text{complexity}(f(\text{volume})+f(\text{verity})+f(\text{velocity})+f(\text{veracity}))=\begin{cases}0; traditional\ database\ approach\\ >0; big\ data\ handling\ approach\end{cases} \quad \text{[v]}$$

Another common features "value" used to represent the significance of data; it will be only valuable when we can convert to usable information; moreover, visualization considered most complexity of big data to represent information that focuses on actual value of the processing data. But we considered that visualization is not common and it is only for on demand of a particular application. So visualization is not considered as a major complexity factor for defining (in equations) but it is an important accessing factor (figure- 2).

2. BIG DATA DEFINITION

Though the big data concept is explained with the figure-1 and 2 its' definition is not unique. There are huge numbers of definition available for big data and a Research work [11] showed comparative revised definitions. Those definitions are developed until 2014 and all of them focusing on the complexity and the functionalities in qualitative manner. Most interesting is that, no one has denied others one and all definitions are conceptual. So our definition is developed by mathematical model (equation i-v) based on the realization of complexity and measurement is concreted by the mentioned equations (i to v). The complexity measure of big data depends on the four criteria: volume, verity, velocity and veracity which are briefly explained in the previous section and by the equations. Equation [i] measures the volume and the complexity is 0 only when it can be handle by traditional database system. Similarly equation [ii][iii] and [iv] measures the complexity of verity, velocity and veracity respectively. And equation [v] measures the complexity by summarizing previous four equations by calculation according to F(complexity)=f(volume)+f(verity)+f(velocity)+f(veracity) ; and Big data is the considered from the [v] when any one of the functions is satisfied except 0 for lower complexity and carries medium or highest complexity for each when individually true for nonzero.

Beside that only 0 outcome recommends for traditional database approach. And equation [v] is the general formula to distinguish a big data and traditional database application. It is commonly seen that when data are gathered from multiple sources, it created the issues on integration, redundancy, inconsistencies, relations, visualization, retrieve knowledge and so on that leads for big data (figure-1, figure-2 and equation [v]).

Big data challenges can be simplified by the ontological approach and our ontology is designed, then theoretically evaluated base on the foundation of given definition. This research systematically showed the way of complexity reduction by the ontological model in the following sections with figure, logical relation and tabular presentation. Actually ontology is flexible for tools but domain oriented.

3. ONTOLOGY AND BIG DATA

Ontology has a rich historical background that was starting from the philosophy and coverage language, applied physics, industrial science and management. In computing, it is introduced in semantic web technology for explicitly specify the concepts of domain knowledge [12], and its functionalities are summarized in "defining the concepts" and "specification of the relationship among the concepts". Ontologies are designed and developed for the specification of domain knowledge like genome ontology [13], it is used for specification uniformity of genome data that leads on research work in that field; agriculture data ontology [14] adapted for specification "agriculture data" for shared content development.

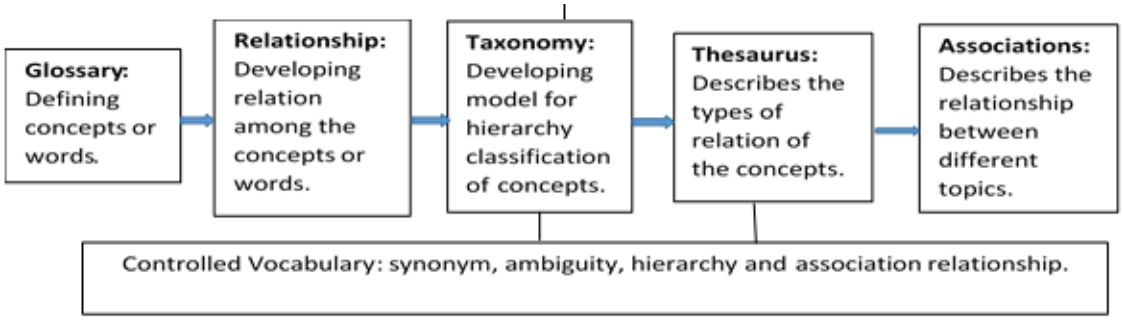

Figure-3. Data ontology

Beside specification, it also works for explicit design in software fields [15], requirement analysis with structural development [16], data analysis and data mining applications [17]. It also brings sustainability in soft computing [18] by the reusability of design and code of software industries. In this article, ontology is being used for design specification of big data enterprises. So that they can easily and effectively handle on big data operations.

An ontology can be represented by graphical presentation, logical view and formal textual way according to the understanding capability of the stakeholders. This paper is developed for the software technologist who has desired knowledge of technical terminologies, working procedure and technical aspect of the working sequence, and the procedure is represent by algorithm (table-I). That categorized with three situations: *new application* that may be member of next big data application, *ontological designed application* forwarded in big data application and big data application for the applications *those are developed without ontology*.

A. *New Application*

For scalability and extension facility need to incorporate by ontological design so that in future the data from the current system can be easily adapt with big data application.

B. *Ontologycal Designed System*

When the ontological design is available just do extension ontology for keeping and managing the existed data that brings further improvement in the system for big data processing.

C. *Existed System Without Ontology*

Need to understand the system and the features of data and try to retrieve the ontology for future action. This is mostly complicated and currently existed as well faced by ERP system.

TABLE I. ALGORITHMS

New Application	Existing applications with ontology	Existing applications without ontology
Step-1: Apply individual ontology for each domain of work Requirement, data, design ontology *Step-2:* Keep portability ontology for each system • Upgrading scope • Specified the way of upgrading • Constraints of upgrading	*Step-1:* Understand the existing ontology for all subdomain *Step-2:* Analysis of an integrated issue *Step-3:* Select the easiest integrated procedure *Step-4:* As much as possible common platform develop *Step-5:* Apply integration *Step-6:* Keep own ontology for further differentiation or more integration	Understand the current system in details and keep an ontological record and apply: *Step-1:* Analysis of the integrated issue *Step-2:* Select the easiest integrated procedure *Step-3:* As much as possible common platform develop *Step-4:* Apply integration *Step-5:* Keep own ontology for further differentiation or more integration

Data understanding and demonstration includes few processes those are mentioned in figure-3 and description as follows for better clarification as well as actions:

I. Glossary

Defining the data and their meanings in the certain domain of knowledge that helps to reduce confusion among stakeholders as well as efficient uses of the dataset for knowledge extraction. A meaningful description of fields is required.

II. Relationship

The internal data items has a relationship in the system and when it required to make relation with different source then knowledge of internal relation improve the productivity for applying external relationship.

III. Taxonomy

The defined concept classified by the hierarchy structure for more clarity and informatics way that also helps for internal and external relationship. It will develop a clear *has a* and *is a* relationship that helps to get inheritance features.

IV. Thesaurus

Identification synonym for knowledge and similar knowledge so that is can be simplify and helps in searching domain oriented knowledge. This will effective for data cleansing and making complete set. It will reduce data redundancy and extra operations.

V. Association

Developing rules among multiple data sources and their relations that is possible for setting logical relation and can be apply mathematical operations. This is become critical when data and their format is unknown because of the lack of ontological practice, especially when source is totally remote and few or less previous record available.

4. KNOWLEDGE EXTRACT ONTOLOGY FOR BIG DATA

Retrieval knowledge from multiple sources of information systems is not only passion but also require tactic and experience. Sometimes need to apply a series of operations sequentially to get desired knowledge and the operations have logical linking relationship.

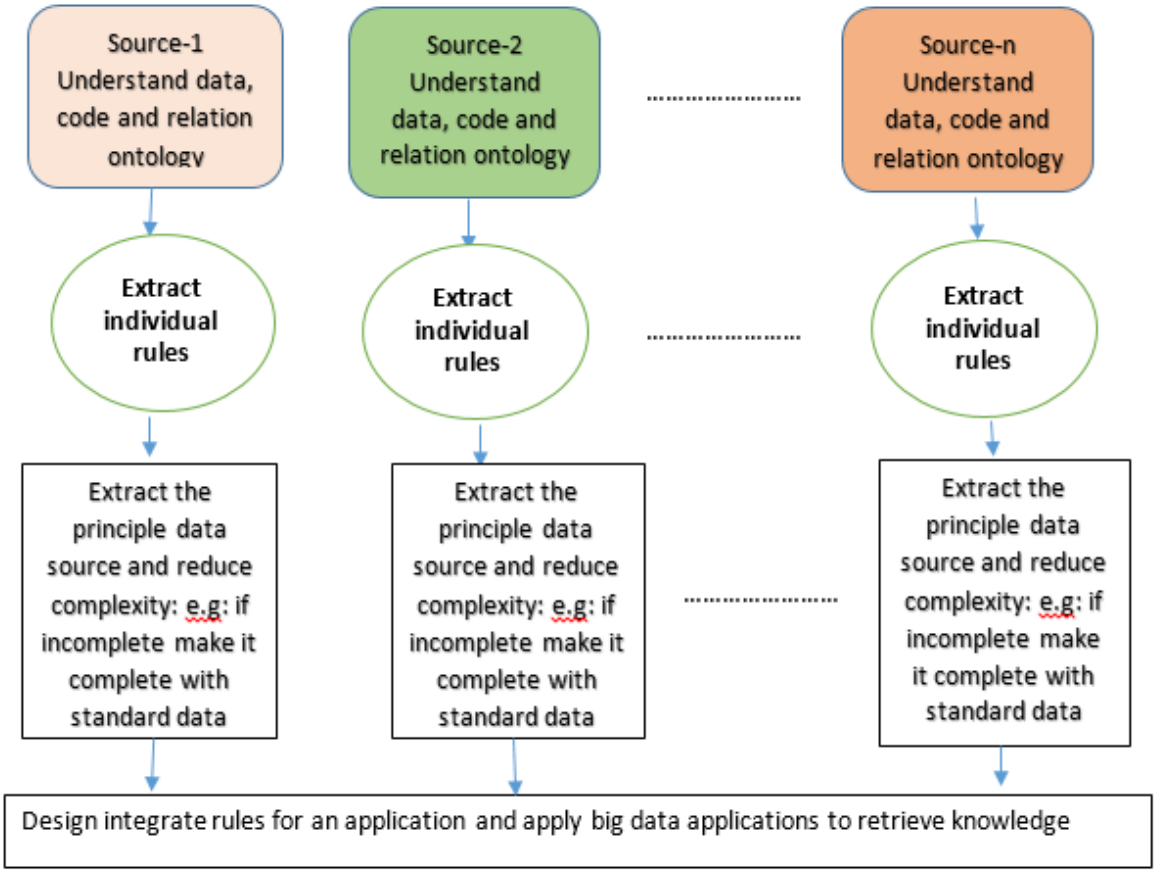

Figure.4. Knowledge Extract Ontology from Multiple Source by Using Big Data Applications

The query formulation for transformation of data and analysis of data are distinguish functions. The comparison of query formulation tools of ontology [19] showed that there is no general ontology to apply. Hence multiple ontology tools are required for knowledge recovery. Our generalization approach showed the integration technique from multiple individual domains to a single integrated system (figure-4) that can be reverse order too.

Inter transferable mapping is proposed by the paper [19] within ontology and database schema for knowledge extraction by three steps: ontology generation from the current database schema, identify ontological knowledge and extract knowledge from a database according to the ontological schema. Those sequential actions will keep the formal records and reusability facility for further actions. The integration process (figure-1) and knowledge extract process (figure-4) are aligned. The developing team can utilize the ontology for professionalism and very suitable in open source domain of knowledge and wisdom extraction.

It includes sequential activities in order like: understanding data type, format, constraints and code; internal relationship and linking with another sources and finally integration for desired knowledge by applications. This is the generalization and adaptable approach so that individual domain data can be specified and retrieve. Technique, technology and approach is individual for a specific domain and application. Currently, most of the case eXtensible Markup Language (XML) based ontology languages developed for semantic web application and those are machine interpretable [20] and the agent markup language called DARPA that includes DAML + OIL(Ontology Inference Language) [21]. Specification did by Resource Description Framework (RDF) [22]. Due to the progress of semantic technology and knowledge presentation, the dimensionality of ontology focuses on a web-based application for specification knowledge search. Web Ontology Language (OWL) and extended by OWL2 [23] for developing domain oriented ontology by W3C. An organization can also develop their own ontology and application for regular analysis. Few domain oriented ontology are mentioned in introduction section and our research is generalization so we prepared descriptive logic [24] for ontology that is reflected in figure-5.

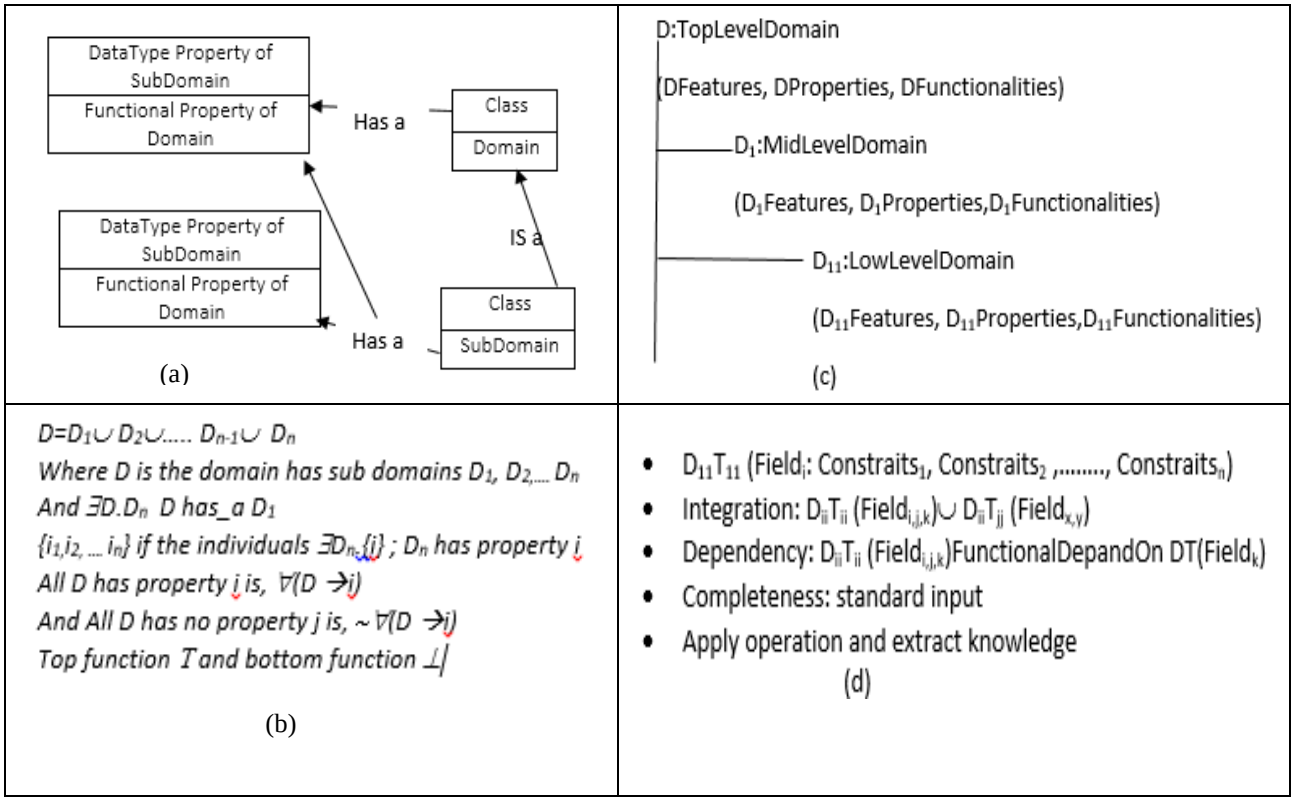

Figure.5. (a) Derived Class hierarchy and "has a", "is a" relation, (b) Domain Knowledge representation by Descriptive Knowledge, (c) Hierarchical presentation of domain and (d) Data specification and operation example

5. FRAMERWORK ANALYSIS

This research commences from the definition of the big data and finally defined by the mathematical model (equation i-v), and how the big data is introduced has illustrated by figure-1. It also mentioned integration complexity for multiple sources, format and lack of formal design (ontology). Followed figure (figure-2) that highlighted the complexity of the big data with issues and value. The big data will be meaningful and useful when we can retrieve knowledge from those data. The specific and measurable definition with complexity analysis and their applications exposed the importance of the study. This research has introduced a specific theory in this area.

Section 3 has discussed the understanding of ontology and what is the roles in big data. Moreover literature coverage few applications of ontology in computing that showed the deepness and dimensionality of ontology. The variation of ontology tools applied: algorithm, tabular, figure and descriptive logic (figure-5). Action based on the application is specified by table-I that includes three different conditions. It also prescribes possible solution for respective situation by individual algorithm.

Figure-5(a) is a relational ontology that has showed the object oriented *has a* and *is a* relationship, a sub class is connected with base class by *is a* relationship and database application an associative has property of base entity. The functionalities and features can be denoted with *has a* relation. Figure-5 (b) has showed the descriptive logic relationship among domain and subdomain of a process. The logical expressions represented *has a* relation, *is a* relation, individual properties, and top/bottom functions. The figure-5(a) and figure-5(b) are the different presentation of same activity. Figure-5(c) has showed the hierarchy of domains and their properties to easily capture which one belong to whom and how they are generated. A big domain can be easily divided to subdomain and can represented based on the hierarchy architecture that easier to understand by non-technical person. Last one (figure-5(d)) has set upped rules by descriptive logic to make clearer about internal data, constrains, dependency and so on.

The mainstream of big data is nothing but the amount of data: specification and understanding data is the main goal of the big data study and the technique is mentioned in figure-3. Figure-4 is the replica for integration process that is based on the figure-1; while figure-1 created the problems and figure-4 follows the steps to resolve the problems of big data integration and knowledge management. To lead on the figure-4, required knowledge and analysis tools are mentioned by the figure-5.

6. CONCLUSION AND FUTURE WORKS

This effort assimilate with fundamental basic study and it has developed own definition of big data that represented by the mathematical model is unique and resolve the conceptual definition of big data. A common scenario of big data creation and their complexity illustrated with multiple figures that will give a clear understanding on big data and it will lead to enrich research too. And followed by different algorithms, technique and required analysis to solve big data issues. This work focuses on a common platform and which will be a little bit far from a specific domain of application. But this generalization ontology helps to establish or select appropriate tools for ontology development for a certain domain. This ontology is developed such a way that technologist can adapt for his desire information. It is for a common environment (descriptive logic) and easily converted by programming. So it increases the scope of the study and works for individual domain knowledge, tools, data specification, relation, rules setting, and so on. This paper has coverage related terminology and the systematic way of problem solving for big data. The logical presentation and ontology clearly describes the way of handling and presenting big data.

This article has developed vital scope of work on each diagram for adapting on individual situation. Researcher can work on the specific domain, data format to inaugurate individual algorithm. Management issues are created by this research that leads to develop management procedure and policy for big data. Details descriptive logic expression can be developed for each and individual section of this article. Control language and software designing tools can be introduced to explain management policy. A novel work can be done by developing ontology development tools for big data. Based on the algorithm we will do implementation and performance evaluation in our next article.

REFERENCES

[1] Rabl T., Jacobsen HA. (2014) Big Data Generation. In: Rabl T., Poess M., Baru C., Jacobsen HA. (eds) Specifying Big Data Benchmarks. WBDB 2012, WBDB 2012. Lecture Notes in Computer Science, vol 8163. Springer, Berlin, Heidelberg

[2] Sumner, M. Enterprise Resource Planning; Prentice Hall: Upper Saddle River, NJ, USA, 2005.

[3] Bastian E et al "Ontology-Based Big Data Management" Systems 2017,5,45.

[4] Lucas-Nülle, T. Product Information Management in Deutschland—Marktstudie; Pro Literatur Verlag: Mammendorf, Germany, 2005.

[5] Ngai, E.W.T. Customer relationship management research (1992–2002): An academic literature review and classification. Market. Intell. Plan. 2005, 23, 582–605.

[6] Boiko, B. Content Management Bible; John Wiley & Sons, Inc.: New York, NY, USA, 2001.

[7] Agrawal, D., El Abbadi, A., Ooi, B.C., Das, S., Elmore, A.J.: The evolving landscape of data management in the cloud. IJCSE 7(1), 2–16 (2012). http://dx.doi.org/10.1504/IJCSE.2012.046177

[8] Beyer, M. Gartner Says Solving "Big Data" Challenge Involves More Than Just Managing Volumes of Data. 2011. Available online: http://www.gartner.com/newsroom/id/1731916 (accessed on 10 May 2017).

[9] Zikopoulos, P.C.; deRoos, D.; Parasuraman, K.; Deutsch, T.; Corrigan, D.; Giles, J.; Melnyk, R.B. Harness the Power of Big Data—The IBM Big Data Platform. 2011. Available online: http://www-01.ibm.com/software/data/bigdata (accessed on 10 August 2016)

[10] Labrinidis, A.; Jagadish, H.V. Challenges and Opportunities with Big Data. Proc. VLDB Endow. 2012, 5, 2032–2033.

[11] Andrea De MauroMarco, GrecoMarco GrecoMichele and GrimaldiMichele Grimaldi, A formal definition of Big Data based on its essential features,Published on Library Review, Vol. 65 Iss: 3, pp.122 – 135, DOI: 10.1108/LR-06-2015-0061

[12] Gruber, T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int. J. Hum. Comp. Stud. 1995, 43, 907–928.

[13] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. Nat. Gen. 2000, 25, 25–29.

[14] ClémentJonquet et al "AgroPortal: A vocabulary and ontology repository for agronomy" Computers and Electronics in Agriculture Volume 144, January 2018, Pages 126-143.

[15] H. HERRE ET AL., General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Technical Report, 8, University of Leipzig, 2006.

[16] Sarker. KU, Derman A. Design Aspects of Near Future Soft Computing. International Journal of Electronics Communication and Computer Engineering, 2017, 8(5):318323.development/ontology101-noy-mcguinness.html

[17] Alexander Vodyaho, Nataly Zhukova," System of Ontologies for Data Processing Applications Based on Implementation of Data Mining Techniques available: http:// ceur-ws.org/Vol1197/paper17.pdf

[18] Sarker K.U, Dr. Aziz Deraman, Raza Hasan, "Green Soft Computing", Conference: ICEL3S2018, Feb 2018, Malacca, Malaysia; Journal of Fundamental and Applied Sciences, Semiannual ISSN: 1112-9867, issue-65, volume-10, page-462-470.

[19] Kamran Mounir and M. Sheraz Anjum, the use of ontologies for effective knowledge modelling and information retrieval, Applied Computing and Informatics 14(2018) 116-126.

[20] K. Munir, M. Waseem Hassan, A. Ali, R. McClatchey, I. Willers, Database independent migration of objects into an object-relational database, in: The 2nd IEEE International Workshop on Autonomous Decentralized System, IEEE, 2002, pp. 132–139, http://dx.doi.org/10.1109/IWADS.2002.1194661.

[21] A. Gómez-Pérez, O. Corcho, Ontology languages for the semantic web, IEEE Intell. Syst. 17 (1) (2002) 54–60.

[22] G. Klyne, J. Carroll, Resource description framework (rdf): concepts and abstract syntax, Ph.D. thesis, 2004.

[23] W3C, OWL 2 web ontology language, world wide web consortium (W3C), 2017.

[24] Sarker K.U, A. B. Deraman and R. Hasan, "Descriptive Logic for Software Engineering Ontology: Aspect Software Quality Control," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2018, pp. 1-5. doi: 10.1109/ICCOINS.2018.8510585

Kamal Uddin Sarker is graduated from RUET (BSC in CSE) & DU (Masters in IT), Bangladesh and PgCert from Coventry University, UK. Currently he is pursuing PhD in Software Quality Control in University Malaysia Terengganu, Malaysia. He has research experience in the area of software quality control, ontology, data mining and wireless mesh network. Moreover, he has one era teaching experience in home country as well as Middle East. Currently he is serving as a senior lecturer in Middle East College, Muscat, Oman.

Dr. Aziz Deraman received his Bachelor from UKM in 1982, Master from Glasgow University in 1984 and PhD from University of Manchester Institute of Science and Technology (UMIST) in 1992 and has served National University of Malaysia (UKM) since 1984 . He has held various academic administrative positions such as head of Computer Science Department (1985-1988), Deputy Dean of IT Faculty (1992-1995), Deputy Director of Computer Centre, UKM (1995-2001) and the Dean of the Faculty of Information Science and Technology (FTSM) (2001-2007). Currently he is the deputy vice chancellor of University Malaysia Terengganu. His research mainly focuses on IT strategic planning, software management and certification, medical computing and community computing. For his research efforts, Prof. Aziz has published more than 100 articles and reports in journals, proceedings, books and other media both local and international and has been awarded the Unesco Fellowship, AIT-Bangkok (1988) and Senior Scientist Visit Program To Usbezkistan - Renong Bhd (1995)

Raza Hasan received his B.S. (Computer Science) from Khadim Ali Shah Bukhari Institute of Technology (KASBIT), Karachi, Pakistan in 2004. He received Masters in Forensic Computing from Staffordshire University UK in 2007. Currently, he is pursuing Doctor of Philosophy in Informatics from Malaysia University of Science and Technology (MUST), Malaysia. He is working as Lecturer in Department of Computing at Middle East College (MEC), Muscat, Oman. His research interests include Learning Analytics, Education Data Mining, BigData and Artificial Intelligence.

Ali Abbas received his M.Sc. degree in Computer Science from Bahira University, Islamabad, Pakistan in 2004, and his M.S. degree in Computer Networks from University of Derby, UK in 2007. Also, He received Ph.D. degree in informatics from the Gyeongsang National University, Korea in 2015. He is with Department of Computing at Middle East College, Oman. His research interests include delay tolerant networks, opportunistic networks and routing protocol design for sensor networks.