A Requirements Specification Framework for Big Data Collection and Capture

Noufa Al-Najran and Ajantha Dahanayake(✉)

Prince Sultan University – College for Women, King Abdullah Road,
Riyadh 11586 Saudi Arabia
it.nouf@hotmail.com, adahanayake@psu.edu.sa

Abstract. The ad hoc processes of data gathering used by most organizations nowadays are proving to be inadequate in a world that is expanding with infinite information. As a consequence, users are often unable to obtain relevant information from large-scale data collections. The current practice tends to collect bulks of data that most often: (1) containing large portions of useless data; (2) leading to longer analysis time frames and thus, longer time to insights. The premise of this paper is; that big data analytics can only be successful when they are able to digest captured data and deliver valuable information. Therefore, this paper introduces 'big data scenarios' to the domain of data collection. It contributes to a paradigm shift of big data collection through the development of a conceptual model. In time of mass content creation, this model aids in a structured approach to gathering scenario-relevant information from various domain contexts.

Keywords: Big data scenarios · Information filtering · Big data collection · Big data analytics · Scenario-based data collection · Software requirements engineering

1 Introduction

Today, for business and many other purposes, everyone is dealing with data in one way or another. People communicate through social networks and generate content such as blog posts, photos and videos. Wireless sensors and RFID readers create signals, and servers continuously create log messages. Scientists make scientific experiments and create detailed measurements and marketers' record information about sales, suppliers, customers, and etc. This rapid growth of data is the reason behind the evolution of big data [1]. However, these huge volumes of data are evolving in a great pace, making the process of retrieving relevant and valuable information for decision making very difficult [2].

In the past, excessive data volume was a storage issue, but with decreasing storage costs, organizations tend to acquire and store all the available data through data streaming, whether it matches their organizational needs or not [3]. This leads to the size of datasets getting so huge that efficiency becomes a big challenge for current data

© Springer International Publishing Switzerland 2015
T. Morzy et al. (Eds): ADBIS 2015, CCIS 539, pp. 12–19, 2015.
DOI: 10.1007/978-3-319-23201-0_2

analytical technologies [4]. This is unfortunate because analytics consume a lot of time trying to figure out matching patterns in the data and may not come up with answers to important questions in a timely manner. Organizations are stuck with the ever-growing volume of data, and miss out opportunities to take actions for critical business decisions [2]. Therefore, it is imperative that businesses, organizations, and associations find better approaches for information filtering, which would effectively decrease the information overload and improve the precision of results [5]. They need to separate the meaningful information from the chatter and focus on what counts. Thus, the real issue behind big data value does not only include the acquisition and storage of the massive volumes of data; rather it lies in the process of acquiring only what is suspected of being relevant for further analysis [6]. When the amount of data to be analysed is reduced, the managing of their storage, merging, analysing, and governing different varieties of the data is expected to be simpler and more controllable [5].

Big Data analytics can only be effective when the underlying data collection processes are able to leverage the relevant information to a particular scenario [7]. Thus by improving the usefulness of the analysis results. Therefore, this research looks into the question: "*How can we improve the ad hoc process of data collection that hinders the efficiency of extracting value from large datasets in a timely manner*?" This research endeavours to answer this question through the introduction of a Requirements Specification Framework, for collecting the requirements of a data collection process in order to assist users in understanding what they require to know before attempting to collect the data.

2 Related Works

In terms of 'big data collection', much research is conducted in this field but there is no clear and sufficient information on how to determine relevancy within structured, semi structured and unstructured data in all the available universes of information [1].

In [8], the authors emphasized that data analysis based on spatial and temporal relationships yields new knowledge discovery in multi-database environments. They have developed a novel approach to data analysis by turning topsy-turvy the analysis task. This approach provides the data collector concept, that the analysis task drives the features of data collectors. These collectors are small databases which collect the data of interest.

Nakanishi emphasizes in [9] that most current data analytics and data mining methods are insufficient for big data environments. Therefore, they have designed and proposed a model that creates axes for correlation measurement on big data analytics. This model maps the Bayesian network to measure correlation mutually in the coordination axes. It contributes to a shift in the domain of big data analytics.

The authors in [10], studied different big data types and problems. They developed a conceptual framework that classifies big data problems according to the format of the data that must be processed. It maps the big data types with the appropriate combinations of data processing components. These components are the processing and analytic tools in order to generate useful patterns from this type of data.

IBM in [11] provides a means of classifying big data business problems according to a specified criteria. They have provided a pattern-based approach to facilitate the

task of defining an overall big data architecture. Their idea of classifying data is to map each problem with its suitable solution pattern and provides an understanding of how a structured classification approach can lead to an analysis of the need and a clear vision of what needs to be captured.

There are several traditional approaches and technologies that may possibly lead to have a control on limiting or reducing unwanted data such as:

- *Visualization and manual Data Collection* [13]. However, several challenges emerged as a result of this process. These include the possibility for correct misses/false alarms and errors in categorizing the data and can be very time consuming.
- *Machine Learning and Data Mining techniques* [14]. However, data mining can only be applied to structured data that can be stored in a relational database.
- *Collaborative Filtering (CF)* is a common web technique for providing personalized recommendations, such as the ones generated by Amazon (based on the user profile and transaction history). In spite of the technique's effectiveness, it rises privacy issues as some customers don't prefer to have their preferences or habits widely known, along with other associated challenges such as data sparsity, scalability, and synonymy [15].
- *Contextual Approach* uses semantic technologies such as an NLP, annotation, and classification to handle information integration (depending on the context of the web page at that moment in time) and querying of distributed data. For query representation, SPARQL language is specifically designed for the semantic technology and enables constructing sophisticated queries to search for different types of data [16]. This approach is efficient in terms of its high precision in controlling unwanted data, as it takes into account the important factors such as keywords, synonyms and antonyms. However, it requires a different infrastructure and highly skilled experts to deal with the complicated technology.

Backward Analysis

According to [17], "Backward analysis is the process of defining the properties of the input, given or based on the context and properties of the output". This concept is utilized in optimizing the process of data collection. Analysing the properties of the scenario at hand and determining the relevant elements that, when collected, will probably reveal hidden value, should be done prior to the data collection process. Comprehensive backward analysis will eliminate the chance of being overwhelmed by bulks of irrelevant data. This will help users and businesses to generate fast management decisions and answer mission critical questions. Therefore, collecting data upon prior analysis needs to a particular business scenario eliminates the presence of unrelated data. Therefore, the effectiveness of the final insights derived from the analytics depends on the quality more than the quantity of the data that will form the foundation for the analytic techniques [1].

Much research is conducted around big data scenarios and around data collection [12]. However, there is no clear and sufficient information that links the two fields together. Therefore, the innovativeness of this research lies in the development of a scenario-based big data collection framework that performs as the Requirements Engineering phase for big data capturing. The framework links the two or more aspects together to provide a well-defined approach for identifying the properties of the scenario context in which the data collection process will take place. This research

studies the requirements specification of the big data collection process and makes it more tailored to the business needs, in order to decreases the analysis time and increases the value of the results by making faster management decisions.

3 The Big Data

According to the Gartner group [19]: "Big Data are high-*Volume*, high-*Velocity*, and/or high-*Variety* information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". Yet there are two more equally important characteristics to consider, which are *Veracity* and *Value* [20].

Characteristics of Big Data spans across five dimensions:

- *Volume*: How big the data is growing.
- *Velocity*: How fast the data is being generated.
- *Variety*: Variations of data types and structures.
- *Veracity*: Trustworthiness, validity and quality of the data.
- *Value*: The success of big data drives businesses in terms of better and faster management decisions and financial performance.

Following are some companies which are taking advantage of big data to leverage their performance:

- **Amazon** uses big data to build and power their recommender system that suggests products to people through their purchase history and clickstream data.
- **Samsung Inc.** uses big data on its new smart TVs to enhance their content recommendation engine, and thus, provide the customer with more accurate and user specific recommendations.
- **Progressive Insurance Inc.** relies on big data to decide on competitive pricing and capture customer driving behaviour.
- **LexisNexis Risk Solutions Inc.** uses big data to help financial organizations and other clients detect and reduce fraud through identifying individuals, including family relationships.

4 Data Collection

According to [21], the process of data collection is defined as: "The process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes". Clearly, those huge volumes of continuously generated data are more than what conventional technologies can sustain. Hence, the lack of effective processes for information collection and management in organizations adapting to big data solutions, can result in a negative impact.

Acquiring the data that holds useful information from tremendous amounts of available data with the rapid increase of online information is a non-trivial task. Collecting scenario-based relevant data from all the available information sources, poses several challenges:

- Integrating multi-disciplinary methods aiming to locate useful data in the large volume, messy, often schema-less, and complex world of big data.
- Understanding big data analyzing techniques as well as big data capturing techniques to be able to select the right one for the scenario being processed. And consolidating the possible factors that can have a control over reducing the unwanted data.
- The ability to develop a simple yet comprehensible and powerful approach to guide and streamline the data collection process, based on the properties of the given scenario.
- Collecting big data requires experts with technical knowledge who can map the right data to the right analytical technique, and execute complex data queries.

5 Data Collection Requirements Modeling

The main inspiration of this research comes from the W*H Conceptual Model for Services [18]. The authors in their research have studied the concept of 'services' as a design artefact. They have aimed to merge the gap between main service design initiatives and their abstraction level interpretations. In order to address their research goal, the authors have developed an inquiry-based conceptual model for service systems designing. This model formulates the right questions that specify service systems innovation, design and development.

By using W*H model, the identification of the main factors that govern the data collection phase of a big data analytics solution is discussed in relation to W*H model. The foundation of the W*H model is based on a set of primary, secondary, and additional questions that support completeness, simplicity, and correctness into service systems innovation, design, and development [18]. In the W*H model, the service is primarily declared by answering the following questions:

- **Wherefore?** (The ends)
  It defines the benefit a potential user may obtain when using the service. This factor is based on the answers for the following questions: *why*, *whereto*, *for when*, *for which reason*.
- **Whereof?** (The sources)
  It defines a general description of the environment for the service.
- **Wherewith?** (The supporting means)
  It identifies the aspects that must be known to potential users in case of utilizing the service.
- **Worthiness?** (The surplus value)
  It defines the value of service utilization for the potential user.

The focus of this paper is identifying the scenario specific to the Data Collection phase for a Big Data Analytics solution. It aims to accelerate the analysis time through data reduction by focusing on retrieving data from the source that meets the scenario. Due to the sheer volume, velocity, and variety of big data, it is challenging to minimize the amount of data to be collected. The framework developed shall be applied during the data collection phase; as the initial process in a big data analytics solution. Figure1 provides a diagram that presents the mapping of the primary phase, which must exist on top of the data collection process.

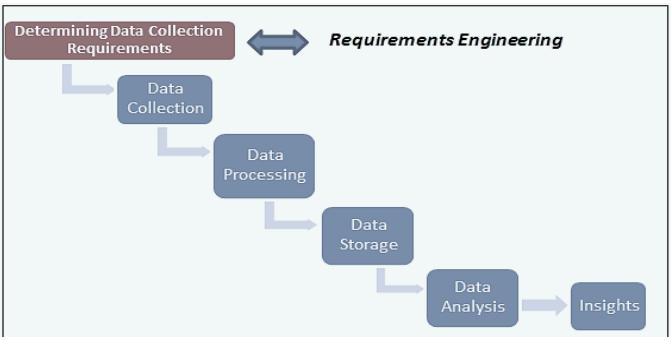

Fig. 1. Requirements Engineering Phase in a Big Data Analytics Life Cycle

The Characteristics of Big Data Scenarios

Scenarios are to be characterized according to their specific properties, the domain that it belongs to, the temporal and spatial factors, the search patterns such as keywords, phrases or named entities, the analysing technique and the capturing technique [11]. They may be composed of other scenarios. In this case, some properties of the scenarios may contradict, while others may be the same. Nested scenarios are beyond the scope of this study. A full description of the scenario-based data collection process is governed by the following factors:

The scenario or purpose - (wherefore) of the data collection and thus the insights a potential domain expert may obtain from analysing the collected data. The scenario description governs the data collection process. It allows to characterize the data collection. This characterization is based on the answers for the following questions: *why, whereto, for when, for which reason*. They define the potential and the capability of the scenario-based data collection process.

The sources - (*whereof*) factor determines the data source that is most likely to generate relevant information about the given scenario among the available data sources. It describes the provider of the data to be collected that is relevant to the given scenario, the consumer of the processed data, the classification of content format and how much data is expected. This classification can help to understand how the data is acquired, and how it will be analysed. These are declared through answering questions such as: *where from, to whom, in what format* and *how much*.

The search patterns - (*by-what*) captures the Supporting needs and Activity factors. It determines which part-of-speech (POS), phrases and keywords correspond to the scenario at hand and must be contained within the data that we want to pull out.

The **Supporting needs** factor describes the analytical technique that can analyse the collected data, the capturing technique that can be utilized to capture the right data, the frequency of the arriving data, the environment at which the data collection process is implemented, whether the data is processed in real-time or batched for later processing. These are declared through answering questions such as: *what, how, whence*, *where and whether.*

The **Activity** factor describes the input and the expected output of the data collection process. These are declared through answering questions such as: *what-in, what-out.*

The value - (*worthiness*) a scenario-based big data collection and analytics is expected to provide for the scenario context and time context, saving time by not collecting garbage but only needed data that is ready to use for more accurate real-time analysis. These are declared by answering the following questions: *where about and when*.

A summary of the mapping of W*H primary, secondary, and additional questions to the Scenario-based Big Data Collection Requirements Framework is presented in table 3.

Table 1. Mapping W*H Model Qs to Scenario-based Data Collection Qs

W*H Model Questions		**Scenario-based Data Collection**	
Ends or Purpose	*Why?*	**Scenario**	*Why?*
	Where to?		*Where to?*
	For when?		*For when?*
	For which reason?		*For which reason?*
Supporting means	*Wherein?*	**Search Patterns**	*What?*
	Wherefrom?		*How?*
	For what?		*Whence?*
	Where?		*Whether?*
	Whence?		*What in?*
	What?		*What out?*
	How?		
Sources	*By whom?*	**Sources**	*Where from?*
	To whom?		*To whom?*
	Whichever?		*In what format?*
	What in?		*Where?*
	What out?		*How much?*
Surplus value	*Where at?*	**Value**	*Where about?*
	Where about?		
	Wither?		*When?*
	When?		

6 Conclusion and Further Research Directions

The main contribution of this research focused on improving the current process of big data collection. Based on a study of the scientific research materials and the literature exploration, it has been observed that the concept of "Backward Analysis", which is performing Reverse Engineering, can add a positive impact to the process of data collection. A data collection that considers the properties of the scenario and the required output before attempting to collect any data (input). The research has studied the scenario-based factors that govern the data collection process and organized them in the form of primary, secondary and additional questions. These questions form the

kernel of the Requirements Specification Framework developed as a structured, well-defined approach for scenario-based big data collection process.

References

1. Santovena, Z.A.: Big data: evolution, components, challenges and opportunities. Massachusetts Institute of Technology (2013)
2. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. Capgemini (2012)
3. META: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group (2001)
4. Martin, G.: Profit from Big Data. White paper, HP Corp. (2013)
5. Hermansen, S.W.: Reducing big data to manageable portions. In: SESUG, USA
6. Akerkar, R.: Big Data computing, 1st edn. Chapman and Hall/CRC (2013)
7. EY: Big Data, Changing the way business compete and operate. Insights on Governance, Risk and Compliance (2014)
8. Thalheim, B., Kiyoki, Y.: Analysis-driven data collection, integration and preparation for visualisation. In: EJC 2010, pp. 142–160 (2012)
9. Nakanishi, T.: A data-driven axes creation model for correlation measurement on big data analytics. In: Proceedings of 24th International Conference on Information Modelling and Knowledge Bases (EJC 2014) (2014)
10. Al-Najran, N., Al-Swilmi, M., Dahanayake, A.: Conceptual framework for big data analytics solutions. In: Proceedings of 24th International Conference on Information Modelling and Knowledge Bases (EJC 2014) (2014)
11. Mysore, D., Khupat, S., Jain, S.: Big Data architecture and patterns, Part1: Introduction to Big Data classification and architecture. IBM Corp (2013)
12. Claire, B.B.: Managing semantic big data for intelligence. In: CEUR Workshop Proceedings of the STIDS, vol. 1097, pp. 41–47 (2013)
13. Angela, C.: Challenges of Capturing Relevant Data. Umati Project (2013)
14. Neck, F., Andersen, D.G.: Challenges and Opportunities in Internet Data Mining. Carnegie Mellon University, Pittsburgh, pp. 15213–3890 (2006)
15. Su, X., Khoshgoftaar, T.M.: A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence **6**(4) (2009)
16. Dimitre, D., Roopa, P., Abir, Q., Jeff, H.: ISENS: A System for Information Integration, Exploration, and Querying of Multi-Ontology Data Sources. IEEE Computer Society, ICSC, pp. 330–335 (2009)
17. Backward analysis: The Free On-line Dictionary of Computing (n.d.)
18. Dahanayake, A., Thalheim, B.: W*H: the conceptual model for services. In: ESF 2014 workshop on Correct Software for Web Application. Sringer-Verlage (2014)
19. Regina, C., Beyer, M., Adrian, M., Friedman, T., Logan, D., Buytendijk, F., Pezzini, M., Edjlali, R., White, A., Laney, D.: Top 10 Technology Trends Impacting Information Infrastructure. Gartner publication (2013)
20. Hitzler, P., Janowicz, K.: Linked Data, Big Data, and the 4th Paradigm. Semantic Web Journal **4**(3), 233–235 (2013)
21. Punch, K.F.: Introduction to Social Research: Qualitative and Quantitative Approaches. Sage, Britain (2005)