Quality Issues of Open Big Data Ecosystems: Toward Solution Development

Guma Lakshen[*],Valentina Janev**, Sanja Vraneš***
[*] School of Electrical Engineering, University of Belgrade, Serbia
,* Mihajlo Pupin Institute, University of Belgrade, Serbia
e-mail: jlackshen65@yahoo.com, e-mail: valentina.janev@pupin.rs, sanja.vranes@institutepupin.com

***Abstract*— Open Big Data is steadily gaining momentum and rapidly becoming a new technology hub in industry and science that motivates technology trends to data-centric architecture and operational models. Therefore, the definition of basic information/semantic/operational models and architectural components that encompasses what is called an *Open Big Data Ecosystem* is becoming a necessity. The purpose of this study is to present our vision of a Big Data Ecosystem and to discuss issues observed with quality of open datasets relevant for the healthcare industry. Furthermore, the paper will propose a solution that will help the healthcare industry to take full advantage of the emerging trends in building pharmaceutical data lakes.**

Keywords: *Linked Data, Big Data, Open Big data Ecosystems, metadata, quality metrics, Best Practices*.

I. Introduction

Data in the Web is growing at a tremendous rate according to IBM [1], Gartner [2], Laney [3], Manyika et al. [4]; this data represents 2.5 quintillion bytes (Exabyte (EB) = 10^{18} bytes). More than 800,000 Petabyte (1 PB= 10^{15} bytes) of data were stored in the year 2000. By end of 2019, this volume will reach 35 Zettabytes (1 ZB= 10^{21} bytes) [5], and is expected to grow 61% and exceeds 175 zettabytes by 2025 as per International Data Corporation IDC expectations [6].

There is no single definition of the term “Big Data”. Big Data to Amazon or Google is very different than Big Data to a medium-sized insurance or telecommunications organization. Many different definitions emerged over time [7], but in general, it refers to “data management challenges” for enterprises and “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” [8].

Big Data is usually described by its characteristics. Laney [9], proposed three dimensions that characterize the challenges and opportunities of increasingly large data volumes: *volume*, *velocity*, and *variety*, known as the (3V's) of data, seeFigure 1 that shows the original 3V's as proposed by Laney. Additional V's of data have been added over the years.

While 3V's have been continuously used to describe big data, the additional dimensions of *veracity* and *value* have been added to describe data integrity and quality to become what is known as 5 V's of big data [10], more V’s were introduced such as validity, vulnerability, volatility, and visualization which sums up to the 10 V’s of big data [11].

In our analysis, we will take into consideration that Big Data is also associated to a new generation of software, applications, system, storage and architecture, all designed to derive business value from unstructured data. Thus this field has become an indispensable area of research as it is expected to add big value to enterprises in the real world.

Regarding the term “Ecosystem”, in ICT literature, it is defined as a complex network or interconnected systems. The main function of a data ecosystem is to capture data and to produce useful insights. In the past, data ecosystems were designed to be relatively centralized and static [12]. However, the birth of the web and cloud services has changed that and thus the infrastructure and services that are used to collect data in organizations adapt and change constantly.

In this paper, the authors present a study related to quality issues in “Big Data Ecosystems”. In particular, we are interested in the concept of data quality assessment when a company is integrating open data in the existing data services. Our initial exploration showed that a simple Web search of the terms “Data quality” through any search engine returns over twelve millions pages, which clearly indicates the importance of data quality and its issues.

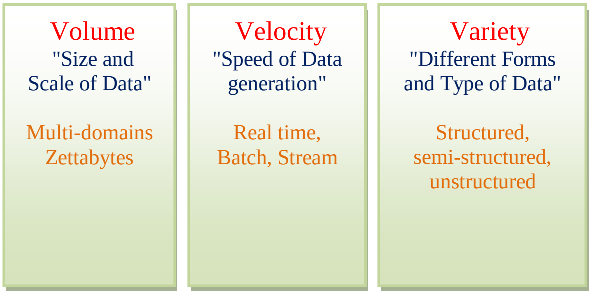
Volume
"Size and Scale of Data"
Multi-domains Zettabytes
Velocity
"Speed of Data generation"
Real time, Batch, Stream
Variety
"Different Forms and Type of Data"
Structured, semi-structured, unstructured

Figure 1. Original 3V's of Big Data

Data quality issues evolved from traditional structured data managed relational databases to Big Data. Advanced tools, software, and systems are required to capture, store, manage, and analyze the data sets, all in a timeframe that preserves the intrinsic value of the data.

The paper is structured as follows. Section II defines the open big data ecosystem architecture for a modern organization from the pharmaceutical domain. Section III defines the open big data quality issues including definitions, dimensions, metrics, and evaluation. Section IV introduces the Data Lake concept and a methodology for building such a repository based on Linked Data principles. Section V concludes the paper with a general discussion and outlines some future directions.

II. Open Big Data Ecosystem Architecture

Big Data is not regarded as a traditional database or Hadoop[1] dilemma; even they contain the essential components and technologies for large scale data processing and data analytics [13, 14, and 15]. Big Data refers to the complete process of storing, processing, visualizing and delivering results to the intended users and applications.

This process, a lump of complex interrelated components, can be defined as the Big Data Ecosystem which deals with the evolving data, models and supporting infrastructure during the whole Big Data lifecycle.

A. Open Big Data Life-cycle

Open Big Data ecosystem is organized in a systematic sequence of operations that should be followed in order to achieve its goals. Figure 2 depicts the most important stages that must be carried out till it reaches its intended purpose and uses as follow:

1. *Data Generation:* data is generated from different data sources such as signals, sensors, devices, sites posts, videos, records, etc.
2. *Data Acquisition & Storage:* this stage consists of 4 operations [16, 17]:
   - *Data Collection:* data is gathered in specific data formats from different sources using a specifically designed script to crawl the web.
   - *Data Transmission:* collected data transmitted using interconnected networks into storage data centers.
   - *Data Pre-Processing:* activities like *Data Integration, Enrichment, Transformation, Reduction, and Cleansing* are performed in this stage.
   - *Data Storage:* represents the infrastructure data center where the data is stored and distributed among several clusters.
3. *Data Analyze & Computations:* application of Data Artificial Intelligence, Mining algorithms, Machine Learning, and Deep Learning to process the data and extract useful knowledge for better decision making.
4. *Data Visualization:* processed data value is assessed by visual examination and taking correct decisions accordingly.

In the past, corporations were used to dealing with static, stored data which could be collected from various sources before they were analyzed and interpreted for visual results.

However, the rapid onset of large data volumes –log files, social media sentiments, click stream information (“customer clicks on the website”) means datasets are no longer expected to reside within a central server or within a fixed place in the cloud.

Also, traditional methods to analyze these information patterns are not at all adequate to handle the copious amount of data, which in turn demands the rise of the advanced analytical tools which can process and store billions of bytes of real-time data, with hundreds of thousands of transactions per second.

Additionally, the development of business intelligence services is a rather simple affair when all data sources collect information based on unified file formats and the data is uploaded to a data warehouse. However, the biggest challenge facing enterprises is the undefined and unpredictable nature of data emerging in multiple formats. Hence, we can conclude that some of the characteristics of a modern data ecosystem are:

- Data is originating from internal systems, cloud-based systems, as well as external data provided from partners and third parties
- Acquisition of data via near real-time data streams in addition to batch loads
- Delivery of analytics to traditional platforms such as data marts and semantic layers, as well as specialty databases such as graphing or mapping
- Analytics use cases ranging from operational and corporate BI to advanced analytics and data science
- Support for the needs of all types of users, ranging from casual consumers to data analysts to data scientists

B. Data Ecosystem Architecture

Roughly, a Big Data Ecosystem can be divided into 3 layers

- Data sources layer
- Data storage and semantic processing layer
- Artificial intelligence technologies and business Intelligence layer

As presented in Figure 3, *the data layer* is composed of both private and public data sources. The *data storage and semantic processing layer* in modern data ecosystem are composed of data lakes and data warehouses. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

The four key differences between a data lake and a data warehouse are given in Table 1. The two types of data storage are often confused but are much more different than they are alike. In fact, the only real similarity between them is the high-level purpose of storing data. In our research, the authors have studied the quality issues in the semantic layer where data from different sources is expected to be integrated.

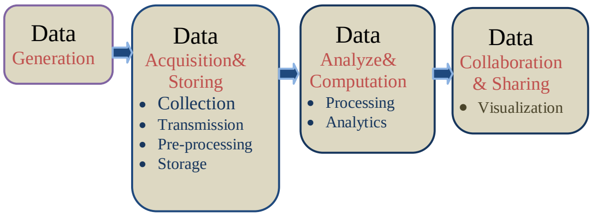

Figure 2. Open Big Data Life-cycle

[1] http://www.hadoop.apache.org/

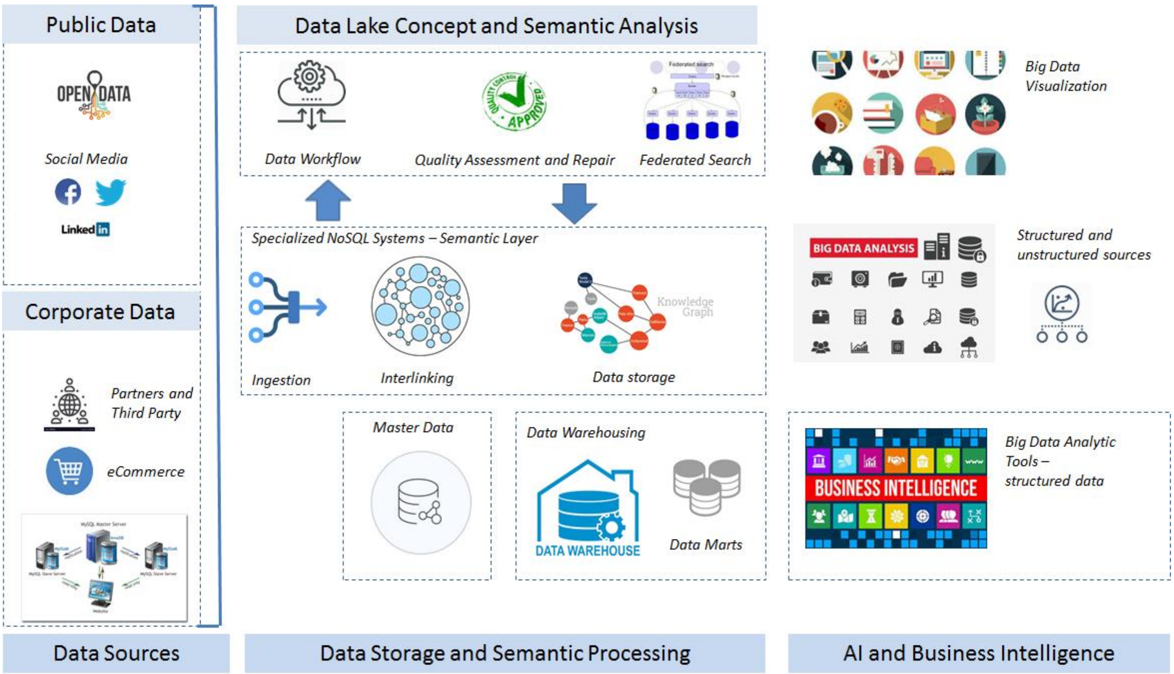

Figure 3. Modern data ecosystem

One of the main functions of the semantic layer is to ensure interoperability. Interoperability is the ability to exchange information between systems in a meaningful way. Hence, the data ingestion step and interlinking steps are often implemented by using open standards such as the standards recommended by the W3C consortium.

C. *Research Questions*

The semantic layer is Built upon the Linked Data principles. The Linked Data principles defined back in 2006 by Tim Berners-Lee [18], enable linking of datasets through references to common concepts. HTTP URIs (Uniform Resource Identifiers) serves to name the entities and concepts, as well as relations (links) to other related URIs. The Resource Description Framework (RDF) is the model used for the representation of the information (entities and concepts), as well as a model that enables exchange and reuse of structured metadata.

Very relevant for Linked Datasets is the variety dimension. This dimension makes the Linked Datasets part of Big Data. Linked Data overlaps with Big Data in 1) LD is considered as a whole as Big Data, 2) No rigid up-front (e.g., relational) data model, 3) Big Data technologies (e.g., Hadoop) are used to handle LD, 4) LD can represent knowledge extracted from big unstructured data.

Taking the drug management domain as an example, the development of Big Linked Data solutions for the pharmaceutical industry raises questions such as:

- How Linked Data technologies can be used to improve the existing business value chain?
- How Big Linked Data Analysis can be implemented that take into consideration unstructured data that is freely available on the Web?
- How a newly establish semantic data lake can coexist with the implemented data warehouse?
- What is the relation between Big Data dimensions (veracity, variability, validity, vulnerability, volatility) and data quality?
- What will be the impact of integrating freely available data sources in the Data Lake?

III. Big Data Quality Issues

The idiom Data Quality DQ is a well-known connotation in the database community and researchers dealt with it for a long time. However, there are additional new characteristics uncovered by Big Data that make it's quality assessment very challenging. Big data variety and massive volume, for instance, introduce new difficulty of its quality evaluation.

TABLE 1. Data Lake vs. Data Warehouse

	Data Lake	Data Warehouse
Data Structure	Raw	Processed
Purpose of Date	Not Yet Determined	Currently In Use
Users	Data Scientists	Business Professionals
Accessibility	Highly accessible and quick to update	More complicated and costly to make changes

In addition, variability, velocity, and volatility features bring new challenges in managing, storing, and assessing the quality of Big Data. As per the author's knowledge, a standard quality management framework for Big Data has not emerged yet.

According to [19], DQ is difficult to define; its definitions are data domain aware. In general, there is unanimity that data quality is always dependent on the quality of the data source [20]. Most of the data quality issues have been addressed heavily in the research community and yet it is still not adapted in Big Data.

A. Big Data Quality definition

DQ definition is inconsistent and related to the specific domain, context area intended for [21], [22]. In [23], the authors summarized data quality from the well-known and used definitions from ISO 25012 Standard. In the literature, data quality is defined as “fitness for use”. In [24], data quality is defined as the appropriateness for use or meeting user needs.

B. Big Data Quality Dimensions(BDQD’s)

Different combinations of dimensions have been studied by several authors. According to [21, 22, and 23], a BDQD provides a path to measure and manage data quality. There are many quality dimensions every quality dimension is linked to specific metrics. According to Zaveri et al. the identified dimensions according to the classification introduced in [20] are:

- *Accessibility*: Availability, licensing, interlinking, security, and performance
- *Intrinsic*: Syntactic validity, semantic accuracy, consistency, conciseness, and completeness
- *Contextual*: Relevancy, trustworthiness, understandability, and timeliness
- *Representational*: Representational conciseness, interoperability, interpretability, and versatility

C. Big Data Quality Metrics

For every mentioned dimension a quantification and measurement are needed. The metrics illustrate the required steps to evaluate a particular dimension. Usually, most metrics used for measurement of data quality are mostly within a range from 0 to 1, with 0 representing incorrect value and 1 representing a correct value [20].

Dimensions such as accuracy, completeness, and consistency amongst others are calculated by the function:

$$M = 1 – (Ni/Nt)$$

Where M is the metric for a given dimension, Ni is the count of incorrect values and Nt is the total values for the dimension concerned.

D. Big Data Quality Evaluation

The aim of Big Data Quality Evaluation (BDQ) Scheme is to address the data quality before starting data analytics. BDQ is carried out by estimating the quality of data attributes or features by applying a BDQD metric to measure the quality characterized by its accuracy, completeness or/and consistency. The expected result is data quality assessment suggestions indicating the quality constraints that will increase or decrease the data quality. The authors believe also that data quality must be handled at many other phases of the big data lifecycle. BDQ issues exist due to several factors or processes occur at different levels:

- *Data sources level:* unreliability, trust, data copying, inconsistency, multi-sources, and data domain.
- *Generation level*: human data entry, sensors devices readings, social media, unstructured data, and missing values.
- *Process and/or application level* (acquisition, collection, transmission).

The data pre-processing improves data quality through executing many tasks and activities such as data transformation, integration, fusion, and normalization.

Big Data problems convene with Data quality issues is justifiable due to the strong bonds between these two concepts. Many authors in literature such as [25 - 33] have stressed that it is very important to discover quality issues and map them with big data problems in the lifecycle as early as possible because any data quality issues will be reflected in the analytics. This will help isolate and adapt the processes that must handle both concerns.

IV. CASE STUDY

The pharmaceutical/drug industry was leading others in expressing interest in validating the approach for publishing and integrating open data [34].

The objective of this study is to propose a methodology for linking open drug data from Arabic countries and define a quality assessment approach for building Linked Data application taking into consideration the possibility of reusing published datasets including the DrugBank[2]and DBpedia[3].

Four drug datasets were selected from four different Arab countries namely; Iraq, Saudi Arabia, Syria, and Lebanon. The quality of the data in selected datasets was so poor, and clearly lacks homogeneity which strengths our proposal of having quality issues resolved at every stage of drug big data lifecycle.

In Big Data, the process of data management is impartial to its quality management. Hence, quality issues and requirements need to be identified at every stage of the big data lifecycle. To ensure a high-quality value chain, a quality assessment procedure is a must and should be executed at each stage of the lifecycle. The goal is that Quality management activities and procedures should be undertaken without adding extra communication, processing and cost overhead on the different Big Data ecosystem layers.

In Figure 4, the authors explore a methodology of a quality management model that captures important quality aspects and proposes how to deal with Big Data quality throughout its lifecycle. The authors identified the processes that must handles and address data quality problems and provide quality assessment schemes to ensure its effective management. The most important modules in the methodology are:

- *Data source module*: in this module, the targeted data is selected from the Arabic drug datasets (Iraq, Saudi Arabia, Syria, and Lebanon) along with the public datasets (DrugBank and DBpedia). In addition to corporate data.
- *Data storage and semantic processing module*: in this module data lake concepts and semantic processing is performed, which include all the

[2] https://www.drugbank.ca/
[3] https://wiki.dbpedia.org/

stages of data preparing, modeling, and conversion is performed. At each stage quality issues are revised and if the quality is not satisfied a return back to the appropriate stage is performed. After the transformation is performed master data is stored for subsequent use.

- *Artificial intelligence and business intelligence module*: in this module, big data visualization is performed using big data analytic tools.

Big Data lifecycle stages and their related quality information’s and processes must be applied to achieve an end-to-end Big Data quality management driven lifecycle.

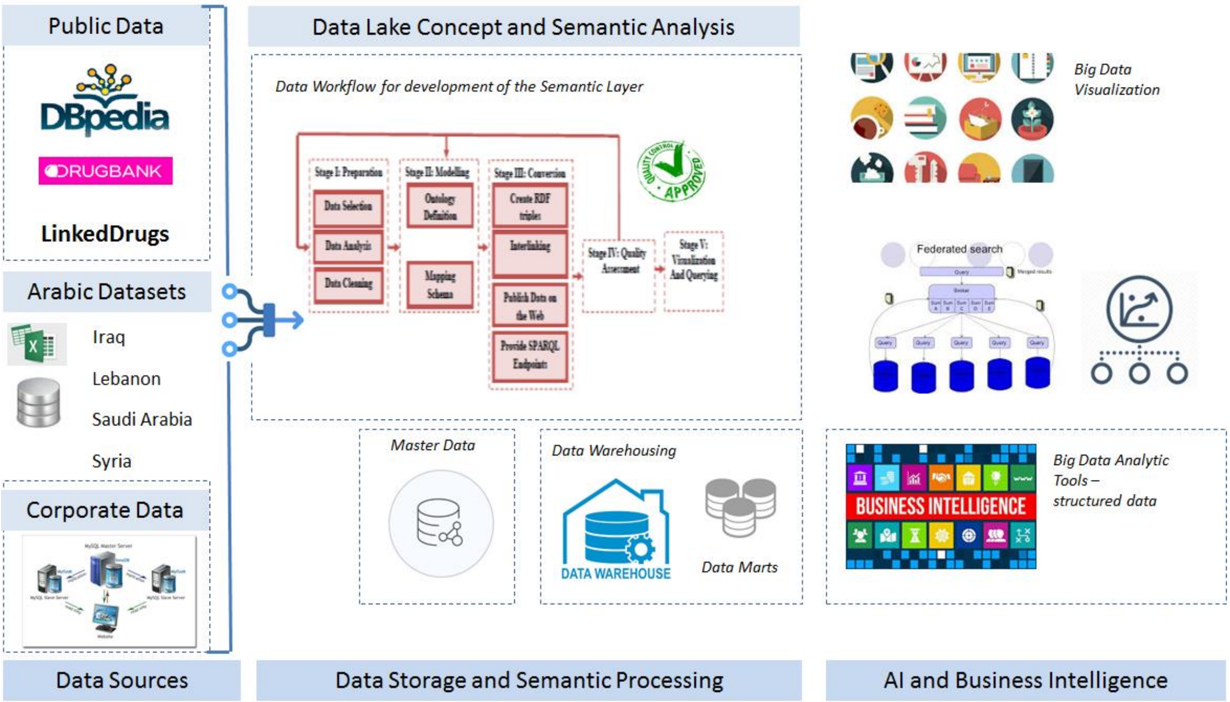

Figure 4. Building a pharmaceutical data lake for Arabic datasets

V. Discussion And Future Directions

Big Open Data has steadily emerged as a new pattern for handling big, continuous, varying, and complex data. Its quality is regarded as the key for its acceptance and usefulness.

The utmost challenge in open big data theme is quality assurance. A number of solutions and approaches proposed in both industry and research to confront a quality issue, but none proposed a comprehensive method for quality assurance solutions.

Using conventional approaches and techniques to manage big open data proven to be not appropriate, this necessitates the need to design of new approach to managing quality.

The authors have identified the key research challenges in Big Data quality and highlighted their importance.

The paper emphasizes the advantages of using the Open Data approach in the pharmaceutical/drug industry and for the first time discusses the issues with drug data from Arabic countries (authors selected four drug data files from four different Arabic countries, Iraq, Syria, Saudi Arabia, and Lebanon).

The authors believe that quality issues in drug domain in the Arab countries are still wide open for further study and evaluation.

Future work will include implementation of a stable and open-source version of a Java web application that will allow the end-user to fully explore and assess the quality of the consolidated dataset, and if possible, to repair the errors observed in the Arabic Linked Drug dataset.

Acknowledgment

The research presented in this paper is partly financed by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010) and partly by the EU project LAMBDA (GA No. 809965).

References

[1] IBM - What is big data?” [Online]. Available: http://www01.ibm.com/software/data/bigdata/what-is-big-data.html . [Accessed: 22-April-2019].

[2] “What Is Big Data? - Gartner IT Glossary - Big Data,” Gartner IT Glossary, 25-May-2012. [Online]. Available: http://www.gartner.com/it-glossary/big-data/ . [Accessed: 23-April 2019].

[3] D. Laney, “The importance of’Big Data’: A definition,” Gart. Retrieved, vol. 21, pp. 2014–2018, 2012.

[4] J. Manyika et al., “Big data: The next frontier for innovation,

competition, and productivity," McKinsey Glob. Inst., pp. 1–137, 2011.

[5] P. Zikopoulos and C. Eaton, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data," 2011.

[6] A. Patrizio, "IDC: Expect 175 zettabytes of data worldwide by 2025", Network World, December 03, 2018, https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

[7] K. D. Foote, "A Brief History of Big Data, December" 14, 2017, https://www.dataversity.net/brief-history-big-data/

[8] G. Press, "12 Big Data Definitions: What's Yours?" Forbes, 2014/09/03, https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#68edb1cb13ae

[9] D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety", *Application Delivery Strategies*, (February, 6th 2001), https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[10] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning.", Performance Evaluation Review, 41(4), 70-73. DOI:10.1145/2627534.2627557, 2014

[11] G. Firican The 10 Vs of Big Data, February 8, 2017, https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

[12] https://mixpanel.com/topics/what-is-a-data-ecosystem/

[13] J.Layton, "The Top of the Big Data Stack: Database Applications", July 27, 2012. [Online]. Available: http://www.enterprisestorageforum.com/storage-management/the-top-ofthe-big-data-stack-database-applications.html

[14] Explore big data analytics and Hadoop. [Online]. Available: http://www.ibm.com/developerworks/training/kp/os-kp-hadoop/

[15] A. Bloom, "7 Myths on Big Data: Avoiding Bad Hadoop and Cloud Analytics Decisions", April 22, 2013. [Online]. Available: http://blogs.vmware.com/vfabric/2013/04/myths-about-running-hadoopin-a-virtualized-environment.html"

[16] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[17] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[18] T. Berners-Lee. 2006. "Design issues: Linked data". Retrieved August 10, 2017, from http://www.w3.org/DesignIssues/LinkedData.html

[19] O. F. Rodrigues, P. R. Henriques, "A Formal Definition of Data Quality Problems.," in IQ, 2005.

[20] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. "Quality assessment for linked data: A survey". *Semantic Web– Interoperability, Usability, Applicability*, Vol. 7, No. 1 (2016), 63-93.DOI: http://dx.doi.org/10.3233/SW-150175. 2016.

[21] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in 2012 International Conference on Information Retrieval Knowledge Management (CAMP), 2012, pp. 300–304.

[22] I. Caballero and M. Piattini, "CALDEA: a data quality model based on maturity levels," in Third International Conference on Quality Software, 2003. Proceedings, 2003, pp. 380–387.

[23] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in 2012 World Congress on Information and Communication Technologies (WICT), 2012, pp. 1009–1013.

[24] R. Blake, P. Mangiameli, The effects and interactions of Data Quality and Problem Complexity on Classification. ACM Journal of Data and Information Quality, 2(2), 2011.

[25] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, "Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model," in 2014 47th Hawaii International Conference on System Sciences (HICSS), 2014, pp. 4700–4709.

[26] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM Comput. Surv., vol. 41, no. 3, pp. 1–52, Jul. 2009.

[27] C. Fürber and M. Hepp, "Towards a Vocabulary for Data Quality Management in Semantic Web Architectures," in *Proceedings of the 1st International Workshop on Linked Web Data Management*, New York, NY, USA, 2011, pp. 1–8.

[28] M. Chen, M. Song, J. Han, and E. Haihong, "Survey on data quality," in 2012 World Congress on Information and Communication Technologies (WICT), 2012, pp. 1009–1013.

[29] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2015, pp. 179–188.

[30] N. Abdullah, S. A. Ismail, S. Sophiayati, and S. M. Sam, "Data quality in big data: a review," *Int. J. Adv. Soft Comput. Its Appl.*, vol. 7, no. 3, 2015.

[31] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From data quality to big data quality," *J. Database Manag.*, vol. 26, no. 1, pp. 60–82, 2015.

[32] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 134–142, May 2016.

[33] G. A. Lakshen, S. Vraneš, and V. Janev, "Big data and quality: A literature review," in *2016 24th Telecommunications Forum (TELFOR)*, 2016, pp. 1–4.

[34] A. Jentzsch *et* al., Linking Open Drug Data, Triplification Challenge of the International Conference on Semantic Systems, 2009.