Serdica J. Computing **11** (2017), No 2, 73–114

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

BIG DATA RESEARCH AND APPLICATION— A SYSTEMATIC LITERATURE REVIEW*

Dessislava Petrova-Antonova, Sylvia Ilieva, Irena Pavlova

Abstract. In the recent years Big Data has become a research topic for both academia and industry. Given the data value for applications in different domains, as well as the business value of the data per se, there is an urgent need for solid end-to-end, data-driven and data-oriented solutions to guide strategic decisions. Such solutions should include a set of mechanisms for runtime adaptations across the complete data lifecycle of Big Data Value Chain. Thus, advanced data functions enabling data to be structured, cleaned, stored, aggregated, modelled, processed, and analyzed are needed.

Considering the significant value of Big Data, this paper presents a systematic literature review. Its main goal is to provide a holistic view of Big Data challenges as a result of a thorough analysis of state-of-the-art research and applications.

ACM Computing Classification System (1998): Y.1.0, Z.2.1.

Key words: Big Data, Big Data Value Chain, State-of-the-Art.

*This work was supported by the European Commission under grant agreement No 763566, by the National Science Fund, Bulgarian Ministry of Education and Science, within project No DN 02/11, and by the Science Fund of the St. Kliment Ohridski University of Sofia within project 80-10-192/24.04.2017.

1. Introduction. The rapid growth of data, which leads to the so-called "Big Data" phenomenon, provokes the interest of EU organizations and institutions. The European Council emphasizes on the importance of digital economy, innovation and services as drivers for growth and jobs and calls for EU action to provide the right framework conditions for a Big Data and cloud computing single market. In response, the European Commission has launched in 2014 the Digital Europe Strategy, defining Big Data as a basic instrument for economic development.

The results of the data explosion are apparent in all domains of daily life with user-generated content of around 2.5 quintillion bytes every day [1]. Big Data Value Association (BDVA) [2] surveys show that gains from Big Data Value are expected across all sectors, from industry and production to public services. Human activities, industrial processes and research all lead to data collection and processing on an unprecedented scale, spurring new products and services, as well as new business processes and scientific methodologies [3]. By 2020, the world will generate 50 times today's information, creating the "Digital Universe" [4] that grows exponentially since, according to OECD [4], the data creation growth rate is between 40% and 60%.

The purpose of the paper is to explore, synthesize and present a systematic analysis of Big Data to identify the challenges for future research and further development of Big Data applications. In contrast to the surveys using *ad hoc* literature selection, this paper follows the "Evidence-based Software Engineering" concept proposed by Kitchenham et al. [5] and applied to survey the reviews in the field of software engineering [6]. The evidence is defined as "a synthesis of best quality scientific studies on a specific topic or research question", whose primary method is a systematic literature review (SLR). Therefore, the main goal of the paper is not to collect the evidence on Big Data state-of-the-art, but to provide an evidence-based framework for Big Data research and application. It is fully aligned with the research objective of "Bi**G** D**A**ta for Smar**T** Soci**E**ty" (GATE) project to advance the state-of-the-art in the whole Big Data Value Chain.

The GATE project's vision is oriented towards the establishment and long term sustainability of a Big Data Centre that will produce excellent science by seamlessly integrating connected fields and associating complemen-

tary skills. GATE will play a dynamic role in the surrounding innovation system by adding value to knowledge, boost the next generation of early-stage researchers and achieve a high level of international visibility and scientific and industrial connectivity.

The rest of the paper is organized as follows. The applied research methodology is described in Section 2. The application of the methodology and the obtained results are presented in Section 3. Section 4 analyzes the results and answers the research questions addressed by the study. Conclusions and directions for future work are outlined in Section 5.

2. Research methodology. The applied research methodology adopts the guideline for systematic reviews that covers three phases of a systematic review: (1) planning the review, (2) conducting the review and (3) reporting the review as proposed in the technical report by Kitchenham [7]. The phases of our research methodology, shown in Fig. 1, are aligned with the procedures prescribed in this report and are described further in the current section.

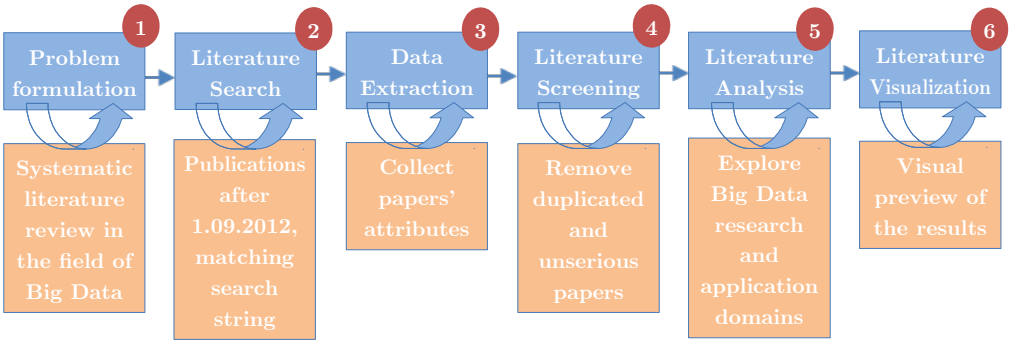

Fig. 1. Research methodology

2.1. Problem formulation. The reasons for performing a SLR varies depending on the specificity of the research field and expected research results, but can be summarized as follows [8]:

- resume the existing evidence related to a given research topic;
- identify the gaps in the current research efforts and give directions for further exploration;
- provide a context for positioning new research achievements.

The reasons for performing a SLR leads to the formulation of a research problem—the academic goal of the review itself, described as a set of research questions. The research questions addressed by this study are as follows:

RQ1: How many SLRs on Big Data have been published in last five years?

RQ2: What Big Data research areas are covered?

RQ3: Which are the primary domains of Big Data applications?

RQ4: What are the challenges for future Big Data research?

2.2. Literature search. An initial search is done covering four major electronic reference databases provided by IEEE, ACM, Elsevier, Taylor & Francis and Springer. They index a large range of research as well as technical papers. Additionally, a manual web search is performed using the Google and Bing search engines. All full accessed papers written in English and published no later than 1 July 2012 were included in the initial result set. A combination of terms was defined to guarantee that relevant information would not be excluded when querying different search engines and databases. As a result, the following research string was created:

"Big Data" AND "Survey" AND "State-of-the-Art"

The automated search in the electronic databases was limited to the title, abstract and key words of the papers. All collected papers are stored with the following format of the file name:

Number from the list—Title of the paper—Year of publication

2.3. Data extraction. The collection of surveys was followed by data extraction according to the attributes presented in Table 1.

2.4. Literature screening. The literature screening was performed in two steps: 1) Filtering according to inclusion and exclusion criteria and 2) Quality assessment using preliminary defined quality questions.

The inclusion criteria are as follows:

Table 1. Attributes of Data Extraction

	Attribute name	**Attribute description**
1	Title	Title of the paper.
2	Institution	Institution of leading author.
3	Country	Country of leading author.
4	Type	Type of the paper, namely Research paper, Technical report, Project deliverable or White paper.
5	Year	Year of publication.
6	Source	Source of publication, namely IEEE, ACM, Elsevier, Springer or Web.
7	Research methods	Research methods covered, if any.
8	Application domains	Application domains covered, if any.
9	Technology stack	Technologies and platforms covered, if any.
10	Summary	Brief description of the goal and main research questions and answers.

- full research papers;
- projects' deliverables;
- papers focused on Big Data research areas and application domains.

The exclusion criteria are as follows:

- white papers, short research papers, technical reports;
- duplicated papers;
- papers covering only technological aspects of Big Data.

The quality assessment questions were defined as follows:

QQ1: Were the methods for data gathering and systematic literature analysis correctly used and described?

QQ2: Were the articles covered in the survey adequately described and analyzed?

QQ3: Are the Big Data gaps, opportunities and challenges identified or the study proposes some reference architecture?

The quality of the publications is calculated according to this equation:

$$Quality = \frac{QQ1 \times 40 + QQ2 \times 30 + QQ3 \times 30}{100}$$

2.5. Literature analysis. For the papers that were selected during the literature screening phase, an additional detailed analysis was performed. The analysis included exploration of the algorithms, approaches and techniques related to Big Data Value Chain activities as well as their application to concrete domains such as energy, healthcare, manufacturing, transportation, education, etc. In addition, the gaps, opportunities and challenges found in each paper are considered.

2.6. Literature visualization. The visualization enables both scientists and professionals to understand data more quickly and thus make reasonable conclusions and better decisions, respectively. Therefore, in order to gain a comprehensive insight into current research and industry trends, the literature studies should provide visual perception of the core findings.

3. Literature results. This section outlines the results of the study.

Table 2 shows the number of studies found in each database together with the number of those found on the web. The initial set of surveys included 145 research papers and 40 projects' deliverables.

Table 2. Literature Search Results

Source	**Number of Studies**
IEEE	34
ACM	13
Elsevier	10
Taylor & Francis	5
Springer	18
SCITEPRESS	10
Web	44
Project deliverables	40

At the first step of the literature screening phase 4 duplicated papers, 1 poster paper, 1 book chapter, 1 student report, 8 white papers and technical

reports, 30 projects' deliverables, and 42 irrelevant research papers – a total of 87 publications – were removed from the initial set.

At the second step, the rest of the publications were assessed according to the quality questions defined. A summary of the quality assessment results is presented in Table 3, while the distribution of the publications by countries is shown in Fig 2.

Table 3. Quality Assessment Results

Quality Score	**Number of Studies**
2	23
1.7	3
1.6	7
1.4	3
1.2	25
1	1
0.9	10
0.6	16

From Fig. 2 it can be seen that China and India are the countries with the largest number of state-of-the-art research in the field of Big Data.

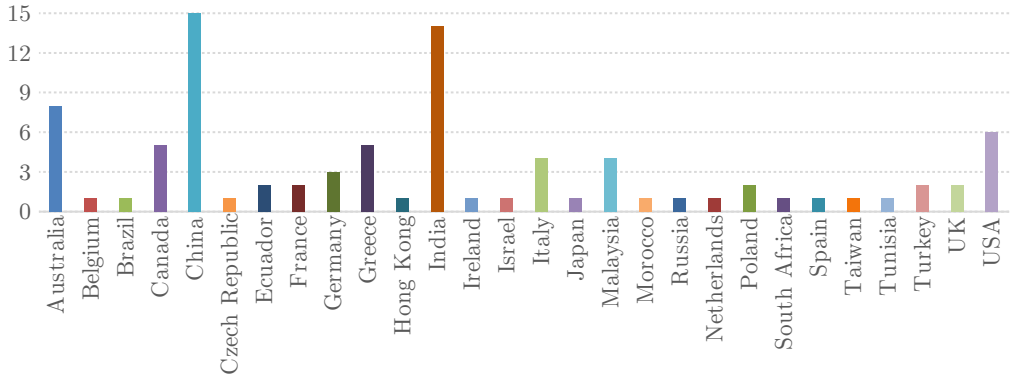

Fig. 2. Distribution of publications by countries

4. Literature analysis and visualization. This section presents the answers to the research questions.

4.1. How many SLRs on Big Data have been published in last five years? The quality assessment results show that there is a lack of surveys providing a systematic literature review in the field of Big Data. Only 26% of the studies obtained at the first step of the literature screening phase have received a quality score of 2, meaning that they are aligned with an existing research methodology or follow a systematic defined research process. The rest present an *ad hoc* literature review that narrows down the research findings and consequently limits the overall Big Data insight.

4.2. What Big Data research areas are covered? The survey's answer to the RQ2 follows the Big Data Value Chain activities shown in Fig. 3.

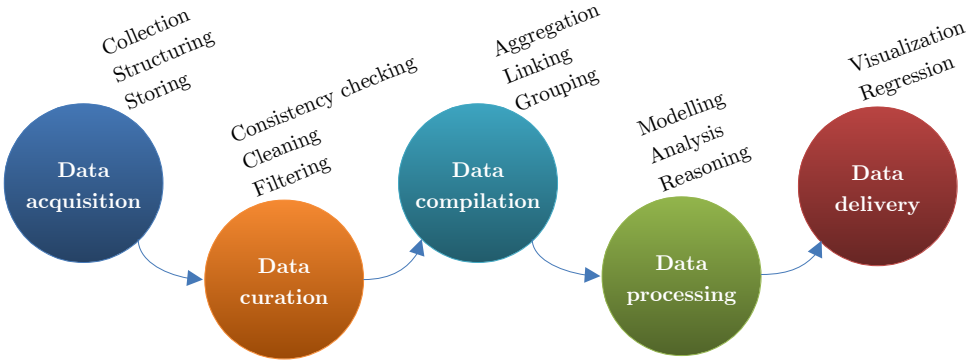

Fig. 3. Big Data Value Chain

Fig. 4 shows a percentage distribution of research efforts according to Big Data Value chain activities, while Fig. 5 presents the distribution per year. As was expected, the largest group of algorithms, approaches and techniques is related to Big Data processing.

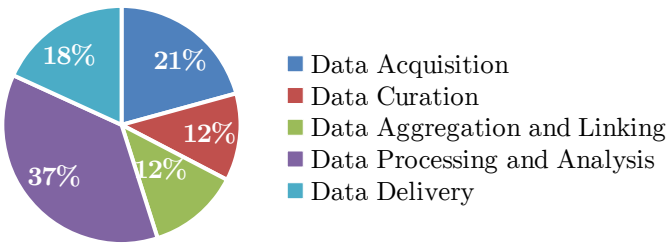

Fig. 4. Big Data Value Chain results

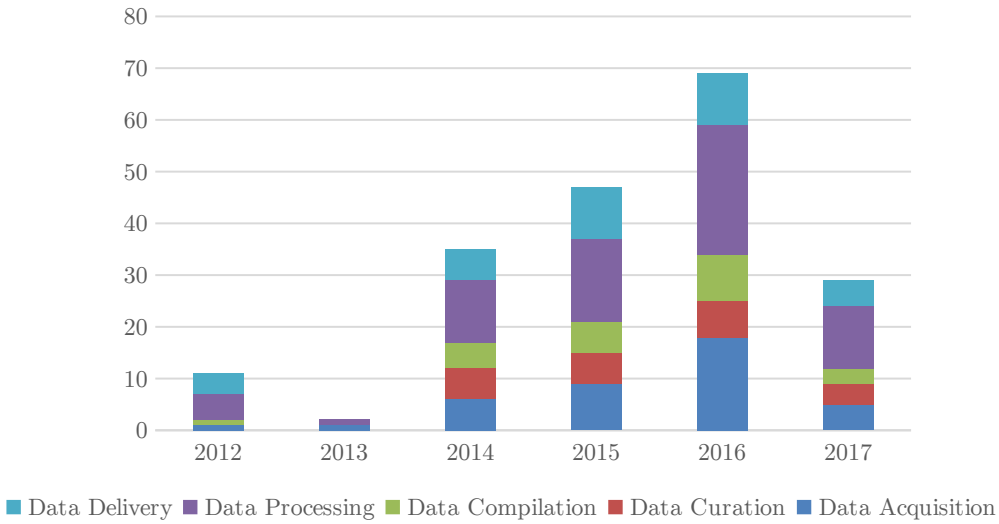

Fig. 5. Distribution of Big Data Value Chain activities per year

4.2.1. Data acquisition. The *Data acquisition* includes data collection, structuring and storing in a scalable data storage, as well as data generalization through digitalization of media, monitoring of activities both online and offline in the real world, as well as through sensors. To achieve this, protocols are required that allow gathering information for distributed data sources of any type (unstructured, semi-structured, structured) and technologies that allow the persistent storage of the data retrieved.

Open state-of-the-art frameworks and protocols for big data acquisition and the current approaches used for data acquisition in the different domains are presented in [9]. The commonly used open protocols for data acquisition are: AMQP (Advanced Message Queuing Protocol), supported by 23 companies, that became an OASIS standard in 2012; Java Message Service (JMS), the de-facto standard for message passing in the Java world; The Memcached protocol, a memory caching system used by many sites (such as Wikipedia, Flickr, Twitter, Youtube) to decrease database load, and hence speed-up the dynamic websites that work on the top of databases. Some Data Acquisition research questions such as data fragment selection, data sampling and scalability are discussed in [10].

The current state of the art in data storage technologies that are capable of handling large amounts of data is assessed in [9], where the following types of storage systems are distinguished:

- *Distributed File Systems* such as the Hadoop File System (HDFS) which offer the capability to store large amounts of unstructured data in a reliable way on commodity hardware. HDFS stores data as replicated blocks in files on the distributed file system for providing fault-tolerance and high availability.
- *NoSQL Databases* use data models other than the relational model and according to the data model used they are divided into Key-value stores: Columnar Stores and Document databases.
- *Graph databases* [11] are Neo4J (disc based), Titan1 (distributed) persistent storage mechanisms that store all information as nodes or edges. They store data in graph structures, which makes them suitable for storing highly associative data such as social network graphs. A particular type of graph databases are triple stores such as AllegroGraph and Virtuoso.
- *NewSQL Databases* are new forms of relational databases that aim at comparable scalability as NoSQL databases while maintaining the transactional properties (atomicity, consistency, isolation and durability) of traditional database systems.
- *Big Data Querying Platforms* are technologies that provide query façades in front of Big Data stores such as distributed file systems or NoSQL databases. Some representatives are Hive, Impala, Shark and Drill.

The HDFS allows for the efficient, distributed, and fault-tolerant storage of massive scale tabular data, as well as being the basis for a large number of other processing solutions, for example Mahout, Giraph, and GraphLab. There are also the graph database solutions that are extremely efficient at storing highly structured data.

Several database management systems are discussed in [12]. The parallel working databases are described as a powerful performance alternative of traditional relational databases. The current NoSQL approaches such as key

value, column oriented and document based databases are compared. The functionality of distributed databases related to analysis of huge amount of data is highlighted. Two horizontal scaling Big Data platforms, namely ApacheTM Hadoop and Berkeley Big Data Analytics Stack (BDAS), are presented in [13]. An emerging Big Data platform at the core of BDAS is the Spark, which is the next generation programming model for large scale data processing. Due to the pertinent performance gains of BDAS over Hadoop, it is getting more attention. However, it is in its infancy with limited support and supporting tools, whereas Hadoop is still widely adopted and has become the de-facto framework for Big Data applications with strong support and variety of supporting data processing tools.

The differences between various Big Data storage platforms are explained in [14]. They are examined and compared in the context of main features, benefits and limitations. MongoDB is an open-source, cross-platform document oriented NoSQL database. After Google's BigTable, Apache Foundation provided an open-source, non-relational distributed database named HBase. Facebook developed Cassandra, which is an open-source DDBMS. It is designed to operate across multiple commodity servers. The RDBMS of Microsoft (SQL Server) provides extensive support for data warehousing. The processing of geospatial data is supported by the open-source relational database PostgreSQL. VoltDB is an open-source ACID-compliant in-memory RDBMS. BlazeGraph is an open-source graph database with support for big data. Apache Accumulo is an open-source distributed key-value store based on Google's BigTable.

4.2.2. Data curation. With the emergence of big data, the data sources appear to be of many different origins, some not well-known and not verified. Low quality data has become a serious problem for many organizations, as data is often collected from such different sources.

Data curation is related to data consistency checking and cleaning, including identification of the data records that are out of range, logically inconsistent or have extreme values, as well as treatment of missing responses to minimize their adverse effects by assigning a suitable value or discarding them methodically. Thus, data quality improves for further processing such as data analysis [15]. Therefore, the data cleaning approach must be adjusted in

terms of context and domain, complexity and according to the enhancement required to support the data analysis to be performed further. A sample cleansing technique for e-commerce data is described in [15], in which crawlers are detected and regular de-duping of customers and accounts is performed. In two other outlined cases, a probabilistic model for missing data in mobile environments and a system with application-defined global integrity constraints to correct input data errors automatically are presented; yet another example is the framework called BIOAJAX that standardizes biological data for additional computation and eliminates errors and duplicates, so common data mining techniques can operate more effectively and searching quality is improved. Two other methods also discuss the robust and computationally efficient multivariate technique called Minimum Covariance Determinant (MCD) that computes the subset of h points from the data that minimizes the covariance matrix determinant, which is very important for multivariate statistical methods, and Conditional Functional Dependencies (CFD) in which the accuracy of data cleaning totally relies on the superiority of the dependencies used in data cleaning.

BIG project provides a very comprehensive analysis of the state-of-the-art, future requirements and emerging trends in the field of data curation, based on literature research, interviews with domain experts, surveys and case studies [9]. The main methods for data curation presented are:

- Master Data Management (MDM) that supports a single point of reference for the data of an organization and can be used to remove duplicates, standardize data syntax and as an authoritative source of master data;
- Collaboration Spaces such as Wikis (that scale to very large user bases) and Content Management Systems (CMSs) (that focus on smaller and more restricted groups to collaboratively edit and publish online content) allow users to collaboratively create and curate unstructured and structured data;
- Crowdsourcing is based on the so-called "wisdom of crowds" and advocates that potentially large groups of non-experts can solve complex problems usually thought to be solvable only by experts.

The effectiveness of crowdsourcing has been demonstrated through websites such as Wikipedia, Amazon Mechanical Turk, and Kaggle.

Several Data Curation Models are also presented in [9] such as Minimum information models, Curating Nanopublications: coping with the long tail of science and Investigation of theoretical principles and Domain Specific Models.

The technologies for data curation can be presented in the context of the specific application domains [10]. For example, Provenance management is cited as the key enabler of trust for health data curation. Human-data interaction technologies such as natural language interfaces or schema-agnostic query formulation are outlined as promising research directions in the area, as well as NLP pipelines, entity and relation algorithms and image segmentation algorithms to address the specific characteristics of health text and image data. There are basic infrastructures in place to support data curation in the domain of Telecommunication, Media and Entertainment. Data curation frameworks such as Open Refine and Data Tamer are cited as the main early stage players in this space, as well as Karma as a research-level data curation framework. These approaches are to be integrated with crowdsourcing-level platforms such as CrowdFlower. The most commonly used technologies for data cleaning are spike removal, integrity checks, faulty data labelling, as well as tracking provenance during the chain of executions [16].

It is apparent that Data Curation is very important step in the whole Big Data value chain. The research efforts should be focused on more automated and effective approaches for consistency checking and cleaning to ensure higher quality of the data.

4.2.3. Compilation (data aggregation and linking). *Data compilation* covers data aggregation and linking, including converting large quantities of data to linked data through semantic web technologies, in order to provide more sophisticated queries. Linked data technologies are the baseline of the so-called Web of Data, a web with a large amount of interconnected data that enables large-scale data interoperability [9]. Exploiting the principles of Linked Data, all relevant data sources are grouped through a dataspace into a unified shared repository. This provides an effective

mechanism to cover the heterogeneity of the Web (large-scale integration) and deal with broad and specific types of data. Semantic technologies such as SPARQL, OWL and RDF allow for management of these data.

The important challenges related to Big Data aggregation and linking are efficient indexing, entity extraction and classification and support search over data found on the Web [9]. The widely used approaches are as follows:

- *Entity summarization*—"... provides keyword-based search for Semantic Web entities", as well as concept search, ontology and class recommendation, and a popularity-based approach for ranking statements an entity is involved in;
- *Google's Knowledge Graph*—provides entity disambiguation ("Find the right thing") and exploratory search ("Go deeper and broader"), together with summaries of entities, i. e., "Get the best summary".
- *Social properties*—considered as "patterns that represent knowledge grounded in the social sciences about motivation, behavior, organization, interaction..." combined with the generic work flow patterns, and highly relevant to the materialization of the communication patterns.
- *Entity recognition and linking* through relation extraction, entity recognition and ontology extraction.
- *Open data to integrate structured & unstructured data*—used as a common-sense knowledge base for entities and able to be extended with domain specific entities inside organizational environments. Named entity recognition and linking tools such as DBpedia Spotlight can be used to link structured and unstructured data.
- *Natural language processing pipelines (NLP)*—deal with unstructured data. Open Source projects such as Apache UIMA support the integration of NLP into other systems. A powerful industry-developed tool is IBM Watson.
- *Retrieval of work flows and semantic annotation paradigm*, including user and behavior modelling methods in certain domains, design and implementation of work flow patterns tailored for communication in the Social Web, and context-aware adaptation and evolution of the patterns.

The approaches and technologies for data aggregation and linkage are mapped to concrete requirements from a number of application domains in [10]. For example, accessing the whole volume of medical big data, such as medical text or medical images, requires the additional enrichment of unstructured data with structured, semantic labels that represent the content. Algorithms for Automated Detection of anatomical structures and Abnormal Structures (including automated measuring) for image processing are also mentioned in this context. Further, in the Energy and transport sectors the authors conclude that lightweight semantic data models that represent the multiple links within various data sources, such as Linked Data, may provide cost-efficient abstractions, especially when the data domain is so complex and highly interlinked. In the Government domain, data abstraction based on ontologies and communication workflow is outlined as a solution for Predictive policing using open data. Text Analytics, Annotation Frameworks, Context representation for data repositories and semantic patterns are other technologies mentioned as very relevant.

Data processing can be classified as stream processing that includes filtering and annotation or as batch processing that focuses on cleaning, combining and replication [17]. Classification, entity recognition, relationship extraction and structure extraction are listed as the most common approaches for information extraction and data fusion, entity recognition and schema integration as the basic data integration ones.

The Linked Data Oriented Architecture (LOA) is a logical, distributed data representation model that represents the data as a collection of links, navigable via Uniform Resource Identifiers (URIs) [18]. This type of architecture facilitates the process of knowledge discovery by enabling the machine or human user to navigate the graph of links and discover new relationships and facts embedded in the network of links. Further, they state that linked data based architectures offer greater flexibility of knowledge representation and ease of navigation of the knowledge graph than the alternatives. Recent implementations of LOA-like architecture presented are Google's Knowledge Graph [8], DBPedia and Freebase. The Lambda architecture is described as the most prominent Big Data software architecture pattern in [19]. The semantification of the architecture is realized through

RDF data processing from source till result, in order to support provenance information collecting, data interoperability and data dissemination. In the above context, it is obvious that Linked and Open Data and semantic approaches have a significant role to support the Big Data Analysis.

4.2.4. Data processing. *Data processing* is related to data modelling, analysis and reasoning including semantic and knowledge-based analysis, scalable and incremental reasoning, linked data mining and cognitive computing. Since Big Data itself is usefulness, insights are required in order to support reasonable decision-making process. Hence, Big Data analytics is a powerful tool to gain value out of the information of data. Four types of Big Data analytics are described in [20]: Prescriptive Analytics, Predictive Analytics, Diagnostic Analytics and Descriptive Analytics. The Prescriptive Analytics aims to set up the structure, relationships and meaning of data. The Predictive Analytics forecasts the outcome using the collected data. The Diagnostic Analytics uses the past data to explain the what and why something has happened. The Descriptive Analytics prescribes actions using the collected data.

Undoubtedly, Data Mining is the major technique used for predictive analytics. It identifies hidden patterns with the help of classification, regression, association rule and cluster analysis on huge datasets. Data Mining incorporates techniques from statistics and machine learning with database management. Machine Learning provides algorithms which allow learning automatically from data and enhance through experience [21]. Natural Language Processing is an example of a machine learning algorithm for human language analysis. One of its applications is using sentiment analysis on social media to obtain information about customers' reactions during product campaigns [20]. Supervised learning is based on the machine learning techniques that reason on relationships using training data. In contrast, unsupervised learning identifies the hidden trends and patterns in data without using prior knowledge.

Ensemble Learning is a supervised learning technique providing better predictions than a simple constituent model [20]. One of the most used supervised learning technique is Classification. Taking data as input, it builds a classification model and maps a target class to the classifier. The Decision Tree is simple for the implementation classifier, but due to its space limitation and overfitting problem it is not applicable for large data sets [21]. Another

classification technique is the Support Vector Machine, but it has slow training and expensive computational time. The Neural Network algorithm has similar drawbacks due to its black-box nature and difficulty in interpretation. Its disadvantages are avoided by Deep Learning approaches providing enhanced performance. The main benefits of Deep Learning as an analytics technique for Big Data are summarized in [22] as follows:

- hierarchical layer appropriate for processing unstructured data;
- high-level abstraction providing complex representations of data that make the machines independent of human knowledge;
- process high volume of data based on a large dimensional raw data input;
- universal, model meaning that the physical models are used to characterize the universal phenomena;
- does not overfit the training data while finding complex dependencies between different dimensions of the data.

Clustering is an unsupervised learning approach that classifies objects into groups of similar objects, whose similarities are not known in advance. One of the fastest clustering algorithms capable of handling large datasets are the Partitioning algorithms. K-means is a known partitioning clustering algorithm used in data mining and computer vision. Another popular clustering technique is provided by the graph-partitioning-based algorithm that is appropriate for partitioning a graph G into subcomponents with specific properties. The second group of Clustering methods is presented by the Hierarchical algorithms that follows two main approaches, namely Agglomerative (bottom-up) and Divisive (top-down). Density-based algorithms form a third group of clustering algorithms that search for clusters of nonlinear and arbitrary shapes. They are widely adopted in medical datasets such as biomedical images.

In contrast to the Classification techniques that are based on a discrete target, the Regression techniques use a continuous target. Their main purpose is to identify how the value of an independent variable is changed according to the value of one or more independent variables. Thus, such techniques are suitable for predicting sales volumes taking into account different economic and market variables.

Evolutionary Computation is recognized as a new technique for finding optimal solutions and solving real world problems [23]. Since it enhances the performance of information retrieval, the Evolutionary algorithms are applicable in the field of Big Data. There are four main methods of Evolutionary Computation, as follows: Genetic Algorithms, Evolutionary Strategies, Evolutionary Programming and Genetic Programming. Such methods are suitable for solving nonlinear problems, such as improving job scheduling in manufacturing and optimizing the performance of an investment portfolio [20].

A recent technique for collecting data as well as metadata to obtain and enhance the semantic of data is Crowdsourcing. Together with Text analytics, this technique allows creating a new knowledge repository based on mass collaboration. Due to the characteristics of Big Data, the algorithms and techniques that are applicable to Big Data analytics are evolving continuously. Some of them, such as A/B testing and regression analysis, can be applied effectively to smaller datasets. Nevertheless, all analytics approaches discussed in this section can be applied to Big Data to obtain more numerous and insightful results than smaller, less varied ones.

4.2.5. Data delivery. *Data delivery* provides both accessibility and proper visualization, including advanced visualization techniques that consider a variety of Big Data (i. e., graphs, geospatial, sensor, mobile, etc.) available from diverse domains.

The most used and widespread visualization tools and techniques for large data sets focusing on their main functional and non-functional characteristics are surveyed in [24]. 36 software tools for data visualization are evaluated according to scope, software category, visualization structure, operating system, license, scalability, extendibility and latest release version. The selected tools are grouped in 4 subgroups: *Data Visualization* (the main goal is to communicate information clearly and efficiently to users, involving the creation and study of the visual representation of data), *Information Visualization* (the main task is the study of [interactive] visual representations of abstract data [numerical and non-numerical data, such as text and geographic information] to reinforce human cognition). *Scientific Visualization* (for spatial representation) and *Business Intelligent and Visualization tools*. A functional comparison of commercial Visual Analytics tools, such as Tableau,

Spotfire, QlikView, JMP (SAS), Jaspersoft, ADVIZOR Solutions, Board, Centrifuge, Visual Analytics, and Visual Mining, is presented in [25].

The *Interactive visualization* attracts increasing interest according to the published research papers in the last 3 years. The interactive visualization of Big Data in a management perspective is addressed in [26].

The commercial Visual Analytics frameworks are explored in [25], complementary to an existing survey on open source visual analytics tools. The visualization techniques are divided into graphical representations of data and interaction techniques. On a high level, the visualization techniques are classified by the type of visualized data: 1) numerical data (bar chart, line chart, pie chart and scatter plots, parallel coordinates, heatmaps, and scatter plot matrix); 2) text/web (word cloud and theme river); 3) geo-related data (projection on map); and 4) network data (graph) (Treemap, Other graphs).

Techniques such as 1D, 2D, 3D, multidimensional, temporal, tree, and network are important to get deep insights about the large volume of data [13]. Among different types of *chart visualization*, the bar chart type has better usage compared to the pie and line type chart. *Maps type of visualization* is used to analyze multidimensional type of data. Heat types of maps are widely used by companies to analyze data in an effective way, followed by Geo, Tree and Spatial. Also, there are other miscellaneous tools used by big data to analyze data such as tables, spreadsheets, Counts, Data sheets, Histograms, Scatterplots, Statistics.

4.3. Which are the primary domains of Big Data applications? Big Data now affects every aspect of our life. Nevertheless, there are some domains in which it has huge impact and can lead to tremendous change of the current practice and business models. From the analysis results, summarized in Fig. 6, it becomes clear that Healthcare, Business, Government (Public services) and the more crosscutting Security and Privacy domain are the ones in which Big Data arouses special interest and to which the current research efforts are targeted.

This is especially the case for the last one, as the application and adoption of Big Data raises many security and privacy concerns (see challenges below) that do not yet have effective enough solutions to overcome them.

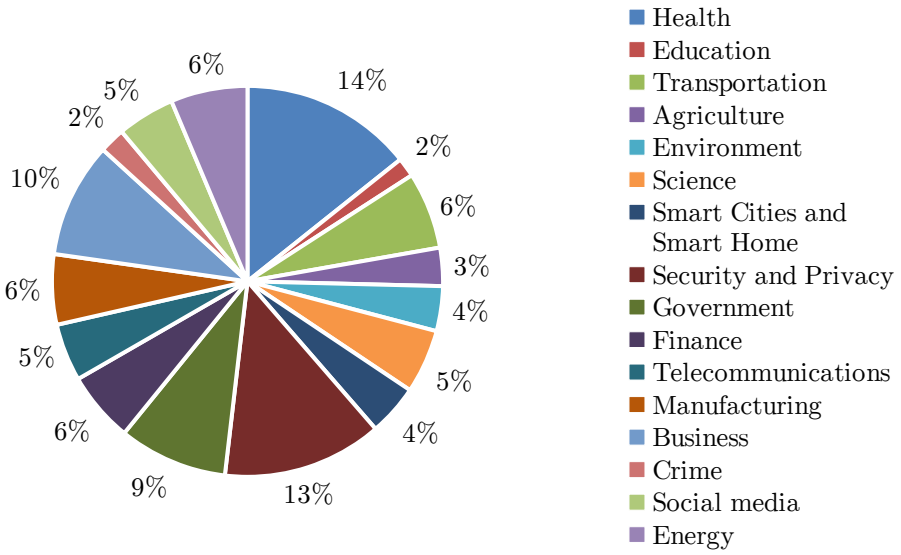

Fig. 6. Big Data Application Domains Results

The distribution of the research efforts per year is presented in Fig. 7.

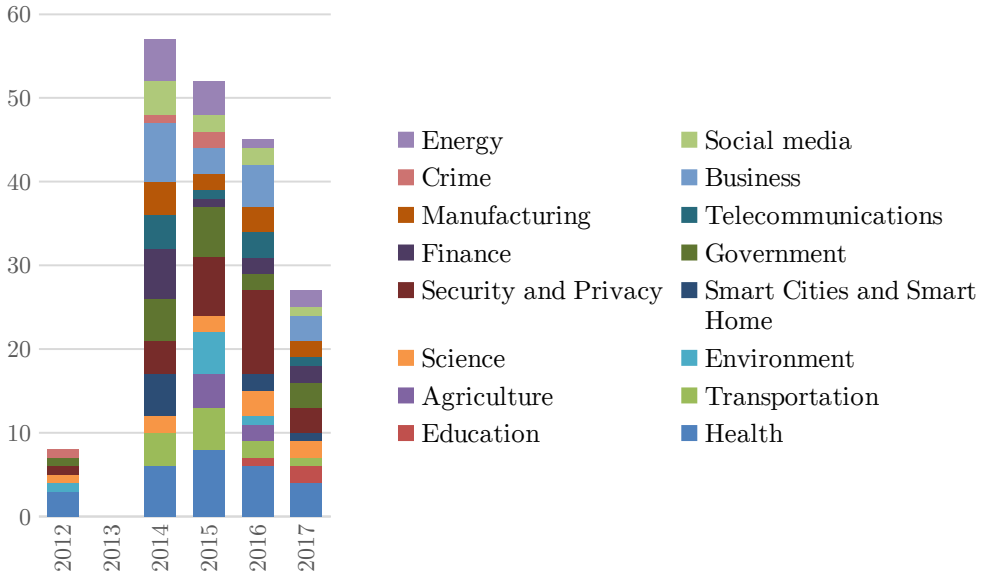

Fig. 7. Distribution of Big Data application domains per year

Healthcare is a domain with huge potential for application of Big Data [12], since data from hospitals, clinics, medical governing bodies and even insurance providers can be mined to study disease affliction rates, patterns and susceptibility trends [27]. The benefits from analysis of health data are in more precise and faster diagnoses and in reduction in the costs of medical systems [28]. Big Health Data (technology) aims to establish a holistic and broader concept whereby clinical, financial and administrative data, as well as patient behavioral data, population data, medical device data, and any other related health data are combined and used for retrospective, real-time and predictive analysis [9]. From an organizational point of view, the storage, processing, access and protection of Big Data has to be regulated on several different levels: institutional, regional, national, and international. Therefore a proper and consistent legal framework or guidelines for all those four levels is needed.

The increasing healthcare cost and the need for *healthcare* coverage force the demand for Big Data technology [29]. The availability and access of health data is continuously improving. The Big Data technology like advanced data integration and analytics technologies, are already used in healthcare. The trend towards value-based healthcare delivery will trigger the collaboration of stakeholders in order to enhance the value of the patient's treatment. Thus, the Big Data applications are very much needed.

The following technical requirements have to be addressed in the *healthcare* sector [10]:

1. the content of unstructured health data is enriched by semantic annotation;
2. data silos are conquered in terms of efficient technologies for semantic data storage and exchange;
3. technical tools of legal frameworks have to provide the transparent sharing and exchange of health data;
4. tools are needed for improving and assessing the data quality.

The Practice Fusion Medical Research Data contains a sample of 15,000 de-identified health records, including information about patients, diagnosis, medications, prescriptions, allergies, immunizations, and vitals respec-

tively [25]. It is used as a viable Use Case to test the data handling capability of 15 commercial Visual Analytics frameworks, as well as some basic analytical capability for answering simple analysis questions and visualizing related information. The requirements and a pilot developed in the Health domain by the Open PHACTS Foundation are presented in [16] and [19]. The emphasis is on the availability of Big Data relating to biological/medical questions at the early stages of drug discovery idea generation and target validation.

Big Health Data technologies help in taking existing healthcare Business Intelligence (BI), Health Data Analytics, and Clinical Decision Support (CDS) as well as health data management applications to the next level by providing means for the efficient handling and analysis of complex and large healthcare data by relying on data integration (multiple, heterogeneous data sources instead of single data sources); real-time analysis (instead of benchmarking along predefined key performance indicators (KPIs)) and predictive analysis [30].

Security and privacy are very important for all aspects of Big Data and are a major requirement of all domains [29, 16, 19, 31]. If they are not appropriately addressed, the phenomenon of Big Data will not receive much acceptance globally [32]. The privacy and security of big data is considered one of the most prominent challenges as it directly impacts individuals [33]. Through big data, individuals lose control over how their data are used and are unable to protect them. Since today legal frameworks defining data access, security and privacy issues and strategies are missing, the sharing and exchange of data is hindered. Further, as the involved parties lack procedures for sharing and communicating relevant findings, important data and information often remains siloed within one department, group or organization.

The global requirements for Security and Privacy in a number of domains and across the Big Data value chain are analyzed and mapped to concrete approaches and technologies in [10]. The AEGIS project is especially focused on the problem of providing systems support for Public Safety and Personal Security (PSPS), and the approach involves supporting both Big Data and Linked Data, as a means to integrate a wide variety of potential data sources [30].

Large-scale *enterprise systems* are seen in [34] as one of the major

sources for Big Data, including enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), and others. Further, the online-to-offline (O2O) commerce that has become a booming business trend in China has been used to illustrate innovative uses of Big Data. The Location-based services employ real-time Big Data and analytics to enable companies to understand how their customers' needs and interests change as the customers move locations.

The tools used for data acquisition in Retail sector can be grouped by the two types of collected data: Sales data from accounting and controlling departments and Data from the marketing departments [9]. The increasing amount of data (currently over 2 petabytes for a full-range retailer) needs to be collected from different sources, processed and stored. In order to be successful in future, retailers must have the ability to extract the right information out of huge data collections acquired [29]. Existing Business Intelligence for retail analytics must be reorganized to understand customer behavior and be able to build more context-sensitive, consumer- and task-oriented recommendation tools for retailer-consumer dialog marketing.

The digital services for customers provided by smart systems are essential for the success in the future *business* [10]. The store of the future will not only provide services for the retailers. In particular, the retail domain will especially be focused on highly efficient and personalized customer assistance services that make use of Big Data and allow a new level of personalized and high-quality Efficient Consumer Response.

Government is another domain with a huge potential for Big Data [12], where the most research benefits of big data technologies are for both citizens and public authorities [28]. Open data and government initiatives are seen to provide more transparency in government spending, as well as allowing innovative solutions to be the developed due to the availability of city data. The types of benefits of Big Data in public sector are advanced analytics, improvements in effectiveness and efficiency, where better services can be provided, and learning from the performance of such services [30].

The *Public sector (Government)* is centered around the activities of the citizens [9]. Data acquisition includes the following areas: tax collection, crime statistics, water and air pollution data, weather reports, energy consumption,

internet business regulation: online gaming, online casinos, intellectual property protection and others.

The *Government* domain is analyzed with regard to its requirements to Big Data [29]. The foreseen benefits include Open Government and Data sharing. The Information from both traditional and new social media can help policy makers to prioritize services and be aware of citizens' real interests and opinions. The segmenting and tailoring government services to individuals can increase effectiveness, citizen satisfaction, etc.

The *Public Sector* needs to exploit the full potential of large scale data analysis and the integration and sharing of existing information to address society's demands for: (I) providing more transparency and fighting against fraud without spending additional public resources; (II) improving its operative efficiency through the analysis of historical operational data through predictive, modelling and simulation tools providing better services to citizens [10].

The *Scientific research* also has been revolutionized by Big Data. Practical examples on how Big Data changed the approach to science are the Sloan Digital Sky Survey in the field of Astronomy and public biological data repositories available for scientists [30]. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially and the researchers and funders recognize the value of integrating clinical research networks. The Social Sciences stand out as one of the areas in which Big Data has a disruptive impact [12, 9]. The availability of micro-level behavioral data creates collaboration opportunities for company researchers, social scientists, and data scientists. Further, Big Data makes innovative projects possible and opens opportunities for investigations that can yield deeper insights into and understanding of human motivation, consumer choices, social phenomena, and the micro-level impact of *business* activities.

Big Data has the potential to also revolutionize *Education* [30]. The educational activities generate an increasingly large amount of data about students' performance. The formal as well as informal learning within the big enterprises provides a potential area for benefitting from Big Data analysis and technologies [35].

Big Data provides benefits for complex *agricultural* and *environmental*

modelling [12, 16]. It is recognized as a solution to the issues of identification, registration and description of big datasets for agriculture, food and environment. Big Data in *Agriculture* is associated mostly with information collected by sensors, satellites or drones combined with genomic information or climate data, which can all help farmers to optimize their farms' operations. A number of cases where Big Data is used in the *Agriculture* and *food industry* can be found in [19, 31]. Meteorologists could use Big Data from the global weather sensors to make more accurate weather predictions and provide timely natural disaster alerts, in order to fight global warming and the serious damage it is causing to the environment and wildlife [27]. The requirements, technical details and realization results are discussed in [16, 19] with respect to a Use Case that employs Big Data for supporting data-intensive climate research. The Use Case considers the huge rise in the availability of large-data sets and argues that the climate research and impact assessment communities need to interface with useful data resources to satisfy the requirement of extracting data efficiently and timely.

The Big Data technologies are proposed for use to process the large data sets that cities generate to impact society and businesses and especially for applications in *traffic* and *emergency* response [9]. Using data analytics technologies, it is possible to monitor what is happening in urban environments and optimize existing infrastructure, to increase collaboration and integration among economic actors [30]. A deep analysis of several sectors' need and requirements towards Big Data including *Smart Cities* is elaborated in [29]. The *Smart Cities* domain requirements are mapped to concrete Big Data approaches and techniques in [10]. The metropolitan authorities responsible for traffic management and building infrastructure facilities could use current as well as historical data to build smart cities and green buildings that use minimum amounts of energy and water for heating and maintenance [27].

The banking and *Financial industry* are huge data producers and consumers as well [27]. The analysis of the tremendous amount of transactional data that commercial banks handle on a daily basis could be used to better understand the spending patterns of their customers. Furthermore, the use of credit cards, patterns of use of mobile banking application, mortgage and credit history of customers could provide banks with greater insight into the

needs of its customers and help tailor their products accordingly. Using all available data for building accurate models can help the financial sector to better manage financial risks [9]. Data quality is one of the most important requirements for Financial Services [10]. The more timely, accurate and relevant the data (along with good analytics), the better the assessment of the current financial state.

Telco, Media & Entertainment are centered on knowledge included in the media files [9]. Since mass media files and metadata about them have been increasing rapidly due to the evolution of the Internet and the social web, data acquisition in this sector has become a substantial challenge. The combination of benefits within marketing and offer management, customer relationship, service deployment and operations is summarized as the achievement of the operational excellence for telco players [29].

The *manufacturing* industry is undergoing radical changes with the introduction of IT technology on a large scale [29]. The developments under "Industry 4.0" include a growing number of sensors and connectivity in all aspects of the production process. Thus, data acquisition is concerned with making the already available data manageable, i. e., standardization and data integration are the biggest requirements. Furthermore, data analysis is already applied in intra-mural applications and will be required for more integrated applications that cover complete logistics chains across factories in the production chain and even into the post-sale usage of (smart) products. Production planning needs to be supported by data-based simulation of complete environments.

The tools for data acquisition have to obtain sensor data that can be incompatible with other sensor data. Therefore, the data integration challenges need to be addressed, especially when sensor data is passed through multiple companies [9]. Other tools need to handle the integration of data produced by sensors in a production environment. In most cases it is achieved when tools operate with standardized meta-data formats. The Big Data Driven Waste Analytics is useful for significantly reducing the construction waste through design [13].

There are highly relevant products in deployment of crime prevention software based on Big Data analytics [9]. Spatio-temporal crime patterns can

be discovered by segmentation and can be very useful for crime analysis [12]. Crime detection and prevention can use the analysis and recognition of suspicious behavior patterns from social media networks and advanced image recognition and matching from surveillance cameras and other video sources, as well as integration with other related existing sources like police and criminal records [29].

Online social graphs are one of the main sources of Big Data [34]. This includes the major social networks such as Facebook, Twitter, Weibo, and WeChat, having close to two billion people that leave a digital trail that can be tracked, graphed, and analyzed. The managing and sharing content can be a challenge, especially for media and entertainment industries [9]. With the need to access video footage, audio files, high resolution images, and other content, a reliable and effective data sharing solution is required.

Media players are more connected with their customers and competitors than ever before and thanks to the impact of disintermediation, content can be generated, shared, curated and republished by literally anyone [29]. The ability of Big Data technology to process various data sources, and in real-time, brings value to the companies willing to invest in it.

In the *Energy* sector, Smart Grid and Smart Meter management is an area that promises both high economic and environmental benefits [9]. The installation and detailed analysis of high-resolution smart meter data can help grid operators to increase the stability of the electrical grid by forecasting energy demand and controlling electrical consumers. The measurement structures together with data extraction technologies facilitate the control and the efficient distribution of water and energy. They support the constant monitoring of distribution networks in search for flaws and planning of infrastructure [28]. These technologies provide a number of advantages, including lower measurement costs, resource waste reduction for customers, theft detection, increased reliability of supply methods and the possibility of custom pricing strategies.

4.4. What are the challenges for future research? The survey's findings regarding the RQ4 are summarized by adopting the three dimensions of Big Data challenges proposed in [20], namely data, process, and management. Data challenges are related to characteristics of Big Data itself such as

volume, velocity, variety, variability, veracity and value (6Vs). The process challenges cover the activities in Big Data Value Chain. The management challenges are primary in security, privacy, governance and ethical context.

4.4.1. Data challenges. The 6Vs of Big Data are crosscutting and pose challenges to every activity within the Value Chain. This section summarizes and presents these challenges in a concrete activity context.

The main characteristic that makes data "big" is the sheer ***Volume***. The total amount of information is growing exponentially every year (it is estimated that around 2.5 quintillion bytes are produced every day) and the amount of information being collected is huge. Because of the Volume, traditional relational database management systems fail to handle Big Data. It actually poses challenges to the whole Big Data value chain, as traditional methods are unable to handle large amounts of data. Volume places scalability in the center of all data processing steps and this is further discussed in the following section. Large-scale reasoning, semantic processing, data mining, machine learning and information extraction are some of the technologies required to cope with large volumes of data. Volume also brings more complex issues of cost, reliability, long query times, and their inability to handle new sources of unstructured or semi-structured data.

Variety is one the most interesting, but at the same time challenging aspects, as more and more information is digitized in a non-structured manner. Traditional data types (structured data) such as date, amount, and time on a bank statement for example, are well defined in a set of rules and fit very well in a relational database. On the other hand, there is also unstructured data that comes from social media feeds, audio files, images, web pages, web log, i. e., anything that can be captured and stored but doesn't comply to a set of rules to frame a concept or idea, e. g., a meta model to define it. A picture, a voice recording, a tweet—they all can be different but express ideas and thoughts based on human understanding. Unstructured data is a fundamental concept in Big Data and one of the main goals is to use technology to take and make sense of them. Variety poses challenges as data varies not just in formats, but in its origin. This brings challenges to the hardware and software requirements of systems processing Big Data. Moreover, the algorithms at present like Bayesian, Random Forest, Back Propagation and so on require

homogenous data, in contrast to the heterogeneous nature of data today.

Velocity is defined as frequency of incoming data that needs to be processed. Big Data is characterized with extremely high velocity, as it is often created in real-time. Such velocity is as a result of huge amount of data sent or the so called "Data-in-motion" (SMS messages, Facebook status updates, etc.) for example from smart phones and mobile applications. With the emergence of Internet of Things and the introduction of sensors everywhere around, the challenges related to Velocity are increasing. Consistency and completeness of fast moving streams of data are one concern. Matching them to specific outcome events is another. In addition, Velocity refers to characteristics such as timeliness or latency. It deals with data capturing at a rate or with a lag time that makes it useful. A challenge to Velocity and data lifetime is also to predict for how long the data will be valuable as well as how to discard outdated data, which is no longer meaningful. Real Time Big Data Analytics also creates challenges in terms of the speed with which data must be stored, processed and retrieved.

Variability refers to the meaning of data which is constantly changing over time, depending on the context for example. This is particularly the case when gathering data relies on language processing. Sophisticated algorithms, which can 'understand' context and decode the precise meaning of words through it, are needed. There are several other meanings for Variability related to whether the data is consistent in terms of availability or interval of reporting, or whether it accurately portrays the event reported. For example, in cases when data contains many extreme values, it poses a statistical problem to determine whether these values contain a new and important signal or are just noisy data.

Veracity is related to the trustworthiness of the data and its source. How this is measured and ensured is seen as a challenge in terms of what the provenance of the data is, whether it comes from a reliable source or whether it is accurate and complete. Big Data is worthless if it's not accurate. This is particularly true for automated decision-making, or feeding the data into an unsupervised machine learning algorithm. The results of such tools are only as good as the data they're working with. By its nature Big Data is heterogeneous, mostly unstructured, "messy and noisy" and it requires a

significant amount of work that goes in to produce an accurate dataset to allow for a meaningful analysis.

In addition to and probably on top of the above five main aspects, Big Data ***Value*** is also considered as an ultimate objective to these five V's. Value refers to the final product created, based on the processing of the data and its importance to the end users. The Big Data Value Chain comprises a multidisciplinary process for collecting, aggregating, cleaning, analyzing, interpreting and visualizing data from different types, origins and volumes. The goal of this process is the discovery of models, correlations, deviations and other facts and events that are hidden in the data itself but represent Value to the end user. Value also places a challenge, as there are interdisciplinary collaboration needs to ensure that Value can be extracted from the Big Data scale, complexity, heterogeneity and incompleteness in a timely and cost-effective manner. The opportunities to learn and generate Value from Big Data depends on the statistically valid use of the information, including interpretation, propensity, and correlations.

4.4.2. Process challenges. Big Data Acquisition deals with high-velocity, variety and real-time data. The data can come from multiple sources, in different formats, structured or unstructured, and at a very high pace. Traditional technologies have limitations as they were meant for homogenous domains and tasks and sometimes have a high computational cost for a large volume of data. Moreover, applying classical methods for accessing Big Data is not suitable for dealing with Big Data provenance. Thus, the main challenge in acquiring Big Data is to provide effective approaches and technologies that will ensure the required throughput without losing any data in the process. In this context, some of the most important challenges to deal with in the future are:

- ability to deal with a wide range of different implementations for acquiring data from different sources;
- mechanisms to connect data acquisition with data pre-processing (curation) and post-processing (analysis) and storage;
- self-adaptive data detection and acquisition algorithms, including pattern detection for "finding the right data", in order to reflect the constant evolution and motion of data;

- synchronization of data by constantly updating extracted knowledge bases if/when sources are changing.

Digitalization of data is an important pre-requisite for Data acquisition. At the same time, a significant amount of data in different domains is not yet digitized. There is a substantial potential to create value, if these pools of data can be digitized, combined, and used effectively.

In the context of the Big Data 6Vs, Data Storage main challenge is about Volume, i. e., related to storing and managing large amounts of data, but storage solutions are also impacted by velocity, variety and value. Velocity is concerned with query latencies, i. e., how long does it take to get a response for a query. Variety is also an important challenge, as there is a strong requirement to manage both structured and unstructured data and handle a multitude of heterogeneous data sources. The importance of the above aspects may vary depending on the domain and the concrete application, but the degree to which the volume, velocity and variety are addressed determines to a great extent how Big Data storage adds value.

Three areas can be outlined as drivers for future Data Storage research: (1) standardization of query interfaces—as currently no standards exist for the individual NoSQL storage types beyond the *de facto* graph database related Blueprint standard and the triple store's data manipulation language SPARQL; (2) increasing support for data security and protection of users' privacy, including legal framework, transparency, data tracing and provenance, etc.; and (3) support for large-scale storage and management of semantic data models.

Big Data Curation is a complex collaborative process that depends on the interaction between automated curation platforms and large pools of data curators, as well as on integration of the provenance representation standards and provenance capture mechanisms into existing data curation, ETL (Extract, Transform and Load) and data management tools. In this context, the main challenge of Big Data curation comprises dealing with huge volumes of data and improving scalability, reducing the cost of data curation, while increasing the number of data curators and performing data curation tasks under limited time constraints.

Big Data solutions often lack data quality due to inexistent or

insufficient validation (curation). This is a challenge, since data validation is difficult when the volume is very large and there is also huge variety of structured and unstructured data. Therefore, devising *ad hoc* data models is needed to support provenance and data sources transformation among different data formats, as the sources may vary significantly. Future research should focus on the integration of structured and especially of unstructured data and its transformation into structured or semi-structured models valid for data analysis. Computational overheads and latencies for which the technology has not been yet specifically optimized during Data Curation also present an open problem to be considered in future efforts. Data curation is also affected by some management-related challenges such as the already mentioned lack of interoperability, privacy and security issues. The available tools are difficult to use and require special skills.

Big Data Aggregation and Linking are also impacted by the data volume as they target converting large quantities of data to linked data through semantic web technologies, in order to provide more sophisticated queries.

Other problems that strongly hamper the Data aggregation and linkage task and pose challenges are the lack of processable linked data and of combined data, as well as the lack of seamless data access, interconnectivity and low levels of interoperability. The data is often in silos, its ownerships are fragmented and data sharing is difficult due to a lack of standardized formats and semantics. Hence, the respective technology should be provided, capable of coping with heterogeneous data models, coming from different sources/owners and supporting user annotation over Big Data and enrichment of data with semantics, in order to correctly and effectively merge data from different sources while processing.

In this respect, much work is required to enable real time insights at large scale with semantic (Linked Data) technologies, where Big Data sets must be transformed into high-level information before meaningful visualization. The data linkage should be (semi-)automatically extracted to increase current scalability.

Efficient indexing, entity extraction and classification and support for search over data found on the Web pose other important challenges to Big

Data aggregation and linking. Important prerequisites for Data aggregation and linking are:

- Data Sharing and Integration, which aims to establish a basis for the seamless integration of multiple and diverse data sources, and
- Increased usage of Common Semantics for describing the meaning of data items, i. e., the extent to which the involved parties and data sources rely on the same or aligned standards for describing the semantics of collected data items (semantic labelling).

The major challenge of ***Big Data Processing (Analysis)*** is related to the scalability of big data, which can be really explosive. Furthermore, the problem of overwhelming amount (or volume), speed (or velocity) and heterogeneity (or variety) of data is also considered as very challenging. These three attributes make traditional data analysis models obsolete or insufficient. Querying Big Data and processing this data in a timely manner is a crucial open problem now, due to the inefficient technologies applied individually or in combination in the new Big Data context.

Besides scale, different technologies should be integrated and applied simultaneously, which is not a trivial task. For example, not all algorithms can be processed in a parallel way, as valuable information might get lost during processing. Further, there are issues occurring in the context of multidimensional structures that have to build up on top of distributed file structures and also in query languages that are missing capabilities to effectively support complex analytics. In addition, integration concepts are needed for heterogeneous data sources, as well as potential combination efforts of NoSQL and relational databases. Accuracy and trust, privacy, interactivity, distributed mining, time evolving data, hidden Big Data and garbage mining are other aspects that should be also considered for holistic data analysis.

Future research directions in data analysis include Stream data mining —to handle high volumes that will come from sensors or from a high number of online users. 'Good' data discovery requires crawlers to find Big Data sets, metadata for Big Data, meaningful links between related data sets, data set ranking mechanisms, etc. "Dealing with both very broad and very specific data" raises the question how to go deeper into a specific domain whilst

maintaining the broader context.

Big Data Delivery is focused on providing access and proper visualization of Big Data, considering a variety of available sources from diverse domains. The main goal is to present the results of data analysis including trends and other predictions by adequate visualization tools. Visual presentation can be crucial for using the results of data analytics in Data Delivery for allowing data scientists or business decision makers to draw conclusions, and for making large result sets manageable and effective. At the same time, depending on the complexity of the visualizations they can be computationally costly and hinder the interactive usage of the visualization.

In this respect, the main technical challenge that can be derived is Usability, related to the extent to which the data delivered is understandable and fits the purposes of the users. There is a need for intelligent tools to visualize actual results, to adapt the visualization to the user and to support next-step decisions. This is possible through reducing the complexity of data and interrelations, as well as the results of analysis. Furthermore, this calls for techniques for customization, situation adaptivity, context-awareness and personalization. The Visual scalability brings another challenge related to the capability of the visualization technology to effectively display large datasets, in terms of either the number or the dimension of individual data elements.

Visual analytics is a new multidisciplinary area combining visualization, data analysis and statistics, human-computer interaction, data management, geo-spatial data processing, spatial decision support, etc. The Big Data 6Vs affect visual analytics in various ways. The volume of Big Data creates the need to visualize multidimensional data and the results of their analysis and to display multiple, linked graphs. In many cases Interactive Visualization – an emerging trend, which enables the display and intuitive understanding of multidimensional data – provides a variety of visualization chart types, and enables users to accomplish traditional data exploration tasks by making charts interactive, and analysis environments are needed that include dynamically linked visualizations. Data velocity and the dynamic nature of Big Data call for correspondingly dynamic visualizations that are updated much more often than previous, static reporting tools. Data variety also presents new challenges for multiple graphs, for cockpits and dashboards.

Complementarity of skills is yet another challenge that creates a discrepancy between technical know-how necessary to execute data analysis (by technical staff) and usage in business decisions (by non-technical staff).

4.4.3. Management challenges. Apart from the research and technological challenges that are presented above, a number of management challenges exist. ***Cost*** is one of the factors that hamper the application of Big Data to a larger scale. This is due to several reasons. Today, business data are held on numerous legacy information systems with a variety of interfaces and different information standards or formats. The cost of migrating these data to new systems is significant, and the risk of loss, damage, or destruction is high. Additionally, foreseeable future storage cost is more of a limiting factor than any shortage in storage capacities. The costs of implementing Big Data Analytics and especially Big Data Curation are still high and are a business barrier for Big Data technology adoption. Coping with data variety can be costly even for smaller amounts of data, the cost of publishing high quality data is also not negligible and there are costs related to data maintenance as well. This calls for development of appropriate approaches to quantify the economic impact, value creation and associated costs behind data resources is a fundamental element for justifying private and public investments in data infrastructures.

Unclear and unsynchronized ***regulations over Big Data*** appear to be a primary external influencer on Big Data. Increased regulatory uncertainty, regulatory pressures, and global business demands are forcing organizations to rethink the value of the data technologies and data management in business processes they use to operate effectively, compete, and manage risk. Rules and regulations are fragmented across Europe. On one hand there are high security demands which can be difficult to address and legislative restrictions on data sharing decreases availability across Europe and makes European-focused initiatives addressing these issues more difficult. On the other hand, Data-driven services are not tied to a particular location. On contrary, conducting cross-domain and cross-border correlations is where data adds value the most, but these are subject to different legislation in different countries. where respective policies are often too connected to the 'old data' world. The existence of a clear and synchronized regulatory framework is an essential prerequisite for the adoption of Big Data and for motivating the busi-

nesses to share and exploit information with and from their service providers.

Some sectors are more resistant to the adoptions of new technologies or of change in general. For example, this is often the case in government and public organizations where people are accustomed to bureaucracy and to certain traditional ways of doing things. Many organizations still depend on old rigid IT infrastructure, with data siloes and a great many legacy systems. Big Data, therefore, is seen as an add-on, rather than as a completely new standalone initiative. These organizations will need to extract the legacy data out, streamline it, build the traceability and lineage. Lack of organizational agility and skills training may also affect negatively the adoption. Culture is an even bigger barrier to Big Data deployment. Many organizations fail to adopt Big Data because they are unable to appreciate how data analytics can improve their core business. Thus, a new challenge rises for companies, namely acceptance of change or keeping old culture and infrastructures**.**

The ***Privacy and Security of Big Data*** can certainly be highlighted as the most important challenge and the one with the strongest impact that was found most often in the literature analysis. This is especially the case when the privacy and security of Big Data is related to how an individual's privacy can be maintained and preserved. Most solutions use personal or sensitive data of people or businesses, but these are not properly handled by the existing security techniques that work with a limited volume of traditional data sources. Privacy attacks and counter attacks constitute serious risks in the perspective Big Data arena, and there are many problems to explore in the field. The distributed nature of the Big Data platforms presents yet another vulnerability. When large data sets are being distributed geographically, the physical security controls should be standardized and secure and privacy-preserving Big Data provenance techniques should be implemented. Another challenge that Big Data adoption faces is that the current tools for managing the Big Data applications were not built with a concern for security. Security, privacy and data protection, including IPR, are cross-cutting challenges highlighted all over the Big Data Value Chain and if they are not appropriately addressed the phenomenon of Big Data will not receive much acceptance globally.

There are many ***Standardization*** problems in the Big Data area, as

data is often fragmented or generated in IT systems with incompatible formats. The lack of standards in gathering and provisioning of data and lack of standards in solution models also makes its reuse difficult. As a result, solutions become very specific, which can hinder the development in the field. In addition, ***Interoperability*** can also be considered as one of the main obstacles for the application of Big Data solutions, because there are no standardized data schemas. Many of the data assets are implemented using unique data formats, proprietary interfaces, with multiple technologies. This makes it difficult to integrate with other systems and to carry out changes. This obstacle is further complicated by the fragmentation of data ownership that leads to the data silo problem. It is an issue that can only be solved from the concrete domain itself at an EU scale with a willingness to harmonize and integrate.

Another common challenge that can be noted in the literature is related to access to the right level of ***Big Data skills***. Even though many organizations recognize the Big Data and its potential, they lack human capital with the right level of skills to be able to bridge the gap between data and potential business opportunity. The skills of a data scientist which are 'missing' can be viewed from three complementary directions:

- commercial skills, or the ability to translate business problems into technology solutions;
- analytical skills that are related to strong statistical skills (that differ however from the skills of a statistician) with a background particularly targeted towards unstructured data mining;
- strong scientific or technical skills for example to be able to write scripts and really extract the core value from data.

These skills do exist in isolation in the industry, but in-depth re-skilling is required to produce the human capital that can really extract Value from Big Data.

A fundamental problem of Big Data research is ***terminology ambiguity.*** Different definitions of the common terms as well as various interpretations of Big Data related concepts and paradigms exist. For example, the number of Big Data Vs varies from 3 to 10, depending on the authors.

With regard to the Big Data Value Chain the borders between Data Acquisition and Analysis are blurred in the preprocessing stage. Some authors argue that pre-processing is part of processing, and therefore of Data Analysis, while others believe that Data Acquisition does not end with the actual gathering, but also with cleaning the data and providing a minimal set of coherence and metadata on top of it. Data cleaning (Curation) usually takes several steps, such as boilerplate removal, language detection and named entities recognition (for textual resources), and providing extra metadata such as timestamp, provenance information, etc. A unification of terminology and common understanding of Big Data concepts is needed.

5. Conclusions. Based on the findings from existing surveys, this paper has synthesized and presented a comprehensive structured analysis on Big Data to provide an evidence-based framework for Big Data research and application. The methodology for systematic literature review adopted has proved as an adequate tool for conducting an exhaustive exploration of the current Big Data studies and synthesis of their core implications. The challenges identified in the paper can assist both industry and academia to develop and advance Big Data approaches across a variety of domains such as healthcare, energy, education, manufacturing, etc. The significant value of Big Data in all aspects of people's lives leading to implementation of data-driven applications and services make it a topical research area for scientific studies.

Future work includes further exploration of Big Data research and applications. The results from the current survey will be extended with investigation based on questionnaires with representatives from both science and business, giving rise to more deep analysis of the role of Big Data for spreading excellence across all sectors, from industry and production, to public services and society.

REFERENCES

[1] Bringing big data to the enterprise. `https://www-01.ibm.com/software/in/data/bigdata/`, 18 February 2018. Big Data Value Association, `http://www.bdva.eu/`, 18 February 2018.

[2] Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Towards a thriving data-driven economy. http://eur-lex.europa.eu/procedure/EN/1042141, 18 February 2018.

[3] The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf, 18 February 2018.

[4] New sources of growth: Knowledge-based capital. http://www.oecd.org/sti/inno/newsourcesofgrowthknowledge-basedcapital.htm, 19 February 2018.

[5] Kitchenham B. A., O. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman. Systematic literature reviews in software engineering—A systematic literature review. *Information and Software Technology*, **51** (2009), No 1, 7–15.

[6] Kitchenham B. A. Procedures for Undertaking Systematic Reviews. Joint Technical Report. Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd. (0400011T.1), 2004.

[7] Kitchenham B. A., T. Dybå, M. Jørgensen. Evidence-based software engineering. In: Proceedings of the 26th International Conference on Software Engineering ICSE'04, IEEE Computer Society, Washington, DC, USA, 2004, 273–281.

[8] Zicari R. V. Big data: Challenges and opportunities. In: R. Akerkar (ed.) Big Data Computing. CRC Press, 2014, 103–128.

[9] Big Data Public Private Forum (BIG). D2.2.2 Final Version of Technical White Paper, 2014. https://big-project.eu/sites/default/files/BIG_D2_2_2.pdf, 19 February 2018.

[10] Big Data Public Private Forum (BIG). D2.4.2 Final version of Sector's Roadmap, 2014. https://big-project.eu/sites/default/files/BIG_D2_4_2_FINAL_v0_851.pdf, 19 February 2018.

[11] Khan A., K. Turowski. A Perspective on Industry 4.0: From Challenges to Opportunities in Production Systems. In: Proceedings of

the International Conference on Internet of Things and Big Data (IoTBD 2016), 441–448. doi: 10.5220/0005929704410448.

[12] Pospiech M., C. Felden. Big Data—A State-of-the-Art. In: Proceedings of the Eighteenth Americas Conference on Information Systems, Seattle, Washington, 2012, 1–11.

[13] Bilal M., L. O. Oyedele, J. Qadir, K. Munir, S. O. Ajayi, O. O. Akinade, H. A. Owolabi, H. A. Alaka, M. Pasha. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, **30** (2016), No 3, 500–521.

[14] Exploiting Oceans of Data for Maritime Applications (BigDataOcean). D2.1—Analysis Report on Big Data Components, Tools and Methodologies, 2017. http://www.bigdataocean.eu/site/wp-content/uploads/2017/04/BigDataOcean_Analysis_Report_on_Big_Data_Components_Tools_and_Methodologies-v1.00.pdf, 19 February 2018.

[15] Siddiqa A., I. A. T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband, A. Gani, F. Nasaruddin. A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, **71** (2016), 151–166.

[16] Big Data Europe. Deliverable 2.12: Community-Driven Big Data Requirements II, 2015. https://www.big-data-europe.eu/wp-content/uploads/D2.12_Community-Driven_Big_Data_Requirements_II.pdf, 19 February 2018.

[17] Salma C. A., B. Tekinerdogan, I. N. Athanasiadis. Feature Driven Survey of Big Data Systems. In: Proceedings of the International Conference on Internet of Things and Big Data, 2016, 348–355.

[18] Begoli, E. A short survey on the state of the art in architectures and platforms for large scale data analysis and knowledge discovery from data. In: Proceedings of the WICSA/ECSA 2012 Companion Volume, 177–183.

[19] Big Data Europe Deliverable 3.1: Assessment on Application of Generic Data Management Technologies I, 2015. https://www.big-data-europe.eu/wp-content/uploads/D3.1-final.pdf, 19 February 2018.

[20] Venkatram K., M. A. Geetha. Review on Big Data & Analytics—

Concepts, Philosophy, Process and Applications. *Cybernetics and Information Technologies*, **17** (2017), No 2, 3–27.

[21] Fang R., S. Pouyanfar, Y. Yang, S.-C. Chen, S. S. Iyengar. Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys*, **49** (2016), No 1, Article No 12.

[22] Hordri N. F., A. Samar, S. S. Yuhaniz, S. M. Shamsuddin. A Systematic Literature Review on Features of Deep Learning in Big Data Analytics. *Int. Journal of Advance Soft Computing Applications*, **9** (2017), No 1, 32–49.

[23] Irfan S., B. V. Babu. Information retrieval in big data using evolutionary computation: A survey. In: Proceedings of IEEE International Conference on Computing, Communication and Automation, ICCCA 2016, 208–213.

[24] Caldarola E., A. M. Rinaldi. Big Data Visualization Tools: A Survey: The New Paradigms, Methodologies and Tools for Large Data Sets Visualization. In: Proceedings of the 6th International Conference on Data Science, Technology and Applications, 2017, 296–305.

[25] Zhang L., A. Stoffel, M. Behrisch, S. Mittelstädt, T. Schreck, R. Pompl, S. H. Weber, H. Last, D. Keim. Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology, 2012, 172–182.

[26] Plank T., M. Helfert. Interactive Visualization and Big Data: A Management Perspective. In: Proceedings of the 12th International Conference on Web Information Systems and Technologies, 2016, 42–47.

[27] Kumar V. D., P. Alencar. Software Engineering for Big Data Projects: Domains, Methodologies and Gaps. In: IEEE International Conference on Big Data, 2016, 2886–2895.

[28] Ribeiro F., F Ferraz, G Alexandre. Big Data Solutions for Urban Environments. In: The First International Conference on Big Data, Small Data, Linked Data and Open Data, 2015, 22–28.

[29] Big Data Public Private Forum (BIG). D2.3.2 Final Version of Sector's

Requisites, 2014. https://big-project.eu/sites/default/files/BIG_D2_3_2.pdf, 19 February 2018.

[30] AEGIS. D2.2—AEGIS Data Value Chain Bus Definition and Data Analysis Methods, 2017. https://www.aegis-bigdata.eu/wp-content/uploads/2017/03/AEGIS-D2.2-AEGIS-Data-Value-Chain-Bus-Definition-and-Data-Analysis-Methods-v1.0.pdf, 19 February 2018.

[31] Big Data Europe. Deliverable 3.4: Assessment on Application of Generic Data Management Technologies II, 2015. https://www.big-data-europe.eu/wp-content/uploads/D3.4_Assessment_on_application_of_generic_data_management_technologies_II.pdf, 19 February 2018.

[32] Sivarajah U., M. Kamal, Z. Irani, V. Weerakkody. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, **70** (2017), 263–286.

[33] Shozi N., J. Mtsweni. Big data privacy and security: A systematic analysis of current and future challenges. In: Proceedings of the 11th International Conference on Cyber Warfare and Security, 2016, 296–303.

[34] Baesens B., R. Bapna, J. R. Marsden, J. Vanthienen, J. L. Zhao. Transformational issues of Big Data and Analytics in networked business. *Journal of MIS Quarterly*, **40** (2016), No 4, 807–818.

[35] Bagriyanik S., A. Karahoca. Personal learning environments: A Big Data perspective. *Global Journal of Computer Sciences: Theory and Research,* **6** (2016), No 2, 36–46.

Dessislava Petrova-Antonova, Sylvia Ilieva, Irena Pavlova
Department of Software Engineering
Faculty of Mathematics and Informatics
St. Kliment Ohridski University of Sofia
5, J. Baurchier Blvd
1164 Sofia, Bulgaria
e-mail: {d.petrova, sylvia, irena_pavlova}@fmi.uni-sofia.bg

Received October 10, 2017
Final Accepted November 27, 2017