big data, big science, advanced manufacturing

Enrico G. CALDAROLA[*], *Marco SACCO*[**], *Walter TERKAJ*[**]

BIG DATA: THE CURRENT WAVE FRONT OF THE TSUNAMI

Abstract

In recent years, a real tsunami has flooded many human activities. Genomics, Astronomy, Particle Physics and Social Sciences are just a few examples of fields which have been intensively invaded by a massive amount of data coming from simulation, experiments or exploration. This huge pile of data requires a new way to deal with, a real paradigmatic shift respect to the past as for theories, technologies or approaches in data management. This work outlines the current wave front of Big Data, starting from a possible characterization of this new paradigm to its most compelling applications and tools, with an exploratory research of Big Data challenges in manufacturing engineering.

1. INTRODUCTION

Post or late-modern societies are going through a new revolution in these years. Current economic, social and technological trends recognize a predominant role of information and the emergence of new information-related activities, challenges and opportunities to an extent never seen before. As these new activities increase, the ICT infrastructures supporting them explode in turn. According to Hilbert and Lopez [1], in 2007, humankind was able to store 2.9×10^{20} optimally compressed bytes, communicate almost 2×10^{21} bytes, and carry out 6.4×10^{18} instructions per second on general-purpose computers. This computing capacity grew at an annual rate of 58% and the majority of our technological memory has been in digital format since the early 2000s. We are living through an age in which the generation of wealth, the exercise of power, and the creation of cultural codes depend on the societies and individuals

[*] Institute of Industrial Technologies and Automation – National Research Council, Via Lembo, 38, Bari, Italy, enrico.caldarola@itia.cnr.it
Dipartimento di Ingegneria Elettrica e Tecnologie Dell'Informazione – Università di Napoli "Federico II", Via Claudio, 21, Napoli, Italy, enricogiacinto.caldarola@unina.it
[**] Institute of Industrial Technologies and Automation – National Research Council, Via E. Bassini, 15, Milano, Italy, marco.sacco@itia.cnr.it, walter.terkaj@itia.cnr.it

attitude towards technologies [2]. The ubiquitous of ICTs in all human activities and the increasing digitalization of the world have led to a great availability of data for organizations and individuals. This has generated a real tsunami that requires a paradigmatic shift respect to the past as for theories, technologies or approaches in data management and more attention to survive it. The data explosion comes from many sources falling under the new term Big Data that is receiving a lot of buzz [3]. What does this term means is not so obvious and is still under debate. There is not a single definition encompassing all its facets and although it has become a catchy term that retains some mystique and persuasive impact in use [4], it still remains elusive. Furthermore, we can define it according to different perspectives, which emphasize some aspects more than others. From a technological point of view, Big Data "refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" [5]. It may also refers to data, which "exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its users" [6]. The previous definitions imply that what qualifies as big data will change over time with technology progress [3]. For this reason, VP and analyst at Gartner's Intelligence and Information Management Group Donald Feinberg stated that the "bigness" of Big Data is a "moving target" and what was Big Data historically or what is Big Data today won't be Big Data tomorrow. This is true to the extent that the term Big Data itself is going to disappear in the next years, to become just Data or Any Data [7]. Taking into account the variability of the definition over the time, Adam Jacob provided the following statement: "Big Data should be defined at any point in time as data whose size force us to look beyond the tried-and-true methods that are prevalent at that time" [8]. From a marketers point of view Big Data is an organizational and decision problem. It is not a technology problem but a business problem [3]. Finally, from a user point of view Big Data can be understood as new exciting, advanced software tools which replace the existing ones. Perspectives aside, the authors define Big Data as a new time-variant paradigm in data management whose raison d'être comes from the enormous availability of data in every human activities that needs to be acknowledged according to different points of view: technological, economic, social, scientific, etc…

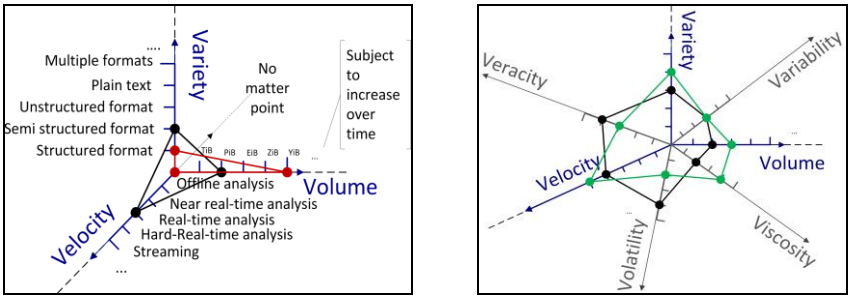
Big Data received a lot of attention in few years: a look at Google Trends shows that, starting from 2011, the term Big Data has been increasingly growing in popularity over time. This eventually demonstrates a global interest not only by research groups but also by numerous individuals with different background: managers, marketers, scientists, public or private organisms. Fields of application of Big Data are great in number: astronomy, physics, social sciences, manufacturing, e-health, genomic and a lot of others [9]. The McKinsey Global Institute [10] also specified the potential of big data in five main topics: healthcare, public sector, retail, personal location data and manufacturing.

Taking into account all the previous consideration, this work aims at reviewing the most compelling current scenarios for Big Data, highlighting the main issues to be tackle according to different dimensions, while also briefing a technological outline of its solutions.

The reminder of this paper is structured as follows. After a deep characterization of Big Data dimensions in the next section, the third section shows four case studies of Big Data in astronomy, particle physics, genomics and social science. The forth section explores the topic in advanced manufacturing and the last section presents an overview of software tools in the Big Data landscape.

2. BIG DATA DIMENSIONS

The concept of Big Data has different dimensions since the term Big not refer only to the quantity of data but also to the heterogeneity of data sources and to the velocity in analyzing data. A widely spread model to characterize Big Data is that of the 3Vs [11, 12] (Fig. 1.a), which shows the three fundamental dimensions of Big Data: Volume, Velocity and Variety.



a) 3Vs Dimensions **b) 7Vs Dimensions**

Fig. 1. Big data dimensions [source: own study]

Along the Volume axis, current scenarios involve technological solutions dealing with data sets with an order of magnitude equal to pebibyte (2^{50} bytes), exbibyte (2^{60} bytes) or higher. Along the velocity dimension, it is possible to distinguish the following typology of analysis: offline analysis (without time constraints over responses), near real-time analysis (must guarantee response within tolerant time constraints), real-time analysis (must guarantee response within strict time constraints), hard-real time (must guarantee response within very strict time constraints) and streaming that refers to data stream mining [13]. Along the Variety axis, the following data formats can be mentioned: structured formats (e.g. relational database data), semi-structured formats (XML grammars-based data, JSON-based, etc.), unstructured formats (data expressed in a no

standard representation language), plain text and multiple format (which combines more data formats). Each dimensions in Fig. 1.a may have a greater or lesser weight than the others and in some cases may not exist at all, nevertheless we keep using the term Big Data. The same figure, for example, shows a scenario, marked by the red edges off, and characterized by large structured data in the order of yobibyte representing a Big Data case anyhow, even if the velocity dimension does not constitute a problem. The triangle defined by the black edges is a more complete scenario involving the need to cope with all the three challenges of Big Data. It requires technological solutions to deal with great storage capacity (PiB), real-time analytics over large data sets and mining or extracting knowledge from very heterogeneous and unstructured data. In addition to the dimensions previously described, some works in the literature provide other Vs: viscosity, variability, veracity and volatility [14, 15] (Fig. 1.b). They measure respectively the resistance to flow of data, the unpredictable rate of flow and types, the biases, noise, abnormality, and reliability in datasets and finally how long data are available and if they should be stored. The green and black irregular polygons in Fig. 1.b identify two scenarios in which different strategies are needed to face the challenges that each dimension places. Although all the seven Vs are increasing, they are not equal as well as their importance depends on the particular case studied: namely, the variety poses a great challenge for many organizations in finding economical ways of integrating newer heterogeneous data sources within existing systems. Veracity is also critical today, since the proliferation of social networks and social media requires much attention in analyzing data before decision-making, as the data can be easily manipulated.

3. BIG DATA SCENARIO

This section analyzes four well-known case studies from different fields highlighting the critical issues related to the previously described dimensions.

3.1. Data-Intensive Science

An important field of application of Big Data is Big Science and Data-Intensive Science. The availability of massive data sets from simulation, exploration or experiments has determined a new fourth paradigm for science based on data intensive computing [16]. This new paradigm has led to a new stage in history of science in which scientific discoveries happen by analyzing large sets of data in parallel computing systems instead of looking throughout telescopes. The next paragraphs describe two of the most compelling and state-of-the-art examples of intensive data science from two fascinating fields: astronomy and particle physics.

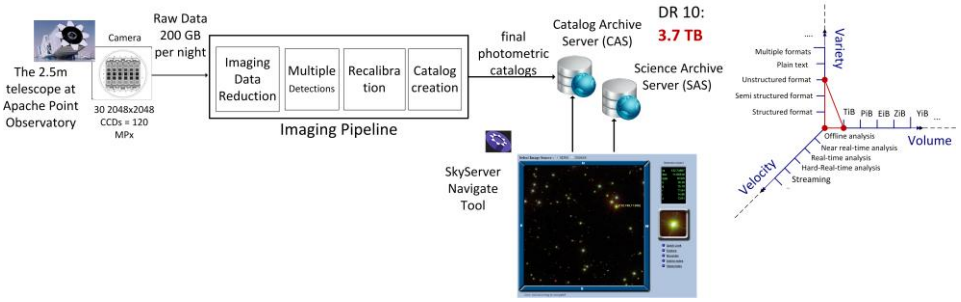


Fig. 2. SDSS overview [source: own study]

Astronomy has been among the first disciplines to undergo the paradigm shift to data intensive science. The Sloan Digital Sky Survey (SDSS) was the first example of publicly available large dataset of three-dimensional maps of the Universe and spectra for millions of astronomical objects. The telescope's camera collects photometric imaging data using an array of thirty 2048 by 2048 pixel CCDs, totaling approximately 120 Megapixels. Every night the telescope produces about 200 GB of data. These raw data pass through an imaging pipeline, which processes them acting a data reduction, a multiple data detection and a recalibration before complex algorithms perform astronomical objects recognition to produce FITS files and catalogs of imaging parameters (Fig. 2). To have an idea of the archive dimension, the photometric parameters for objects in each imaging field in the current data release (DR10) are distributed in around 938,000 fits files, each around 3.5 MB, so the total data set is about 3.7 TB. Going from pixels on the camera to robust catalog information of sky objects is a long and complicated process. Thus, the SDSS project poses important challenges not only in data volume but also along the variety and veracity dimension due the unstructured and noisy nature of raw data acquired from the CCDs. The diagram on the top right corner of Fig. 2 characterizes this scenario in the previously explained three dimensions.

A similar transformation toward data-intensive computing is happening in particle physics. The Large Hadron Collider (LHC) at CERN (Fig. 3) is set to create an integrated data system resembling the SDSS. Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they are made to collide at four locations around the accelerator ring, corresponding to the positions of four particle detectors. Every second 600 millions of collisions happen inside the LHC and each of them is recognized by real-time electronic signals detectors and sent to the CERN Data Center (DC) for digital reconstruction. The DC produces about 30 PB of *collision events* every year and processes them in a distributed computing infrastructure arranged in hierarchical tiers. A part from the dimension of the data to be processed, important challenges in LHC project consist in digital reconstructing of raw electronics

signals coming from particle detectors and in filtering collision events through increasingly refined processing algorithms to detect interesting events. This scenario implies big volume, big variety of data, big veracity and big velocity as shown in the right part of Fig. 3.

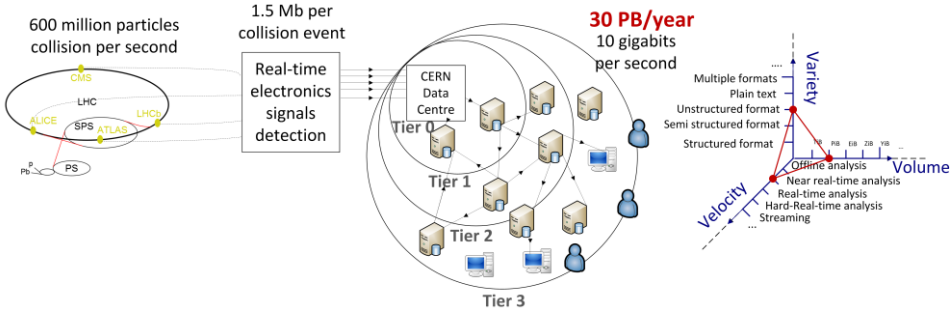


Fig. 3. LHC Overview [source: own study]

3.2 An example from Genomics: GenBank

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is created and maintained by the National Center for Biotechnology Information (NCBI).

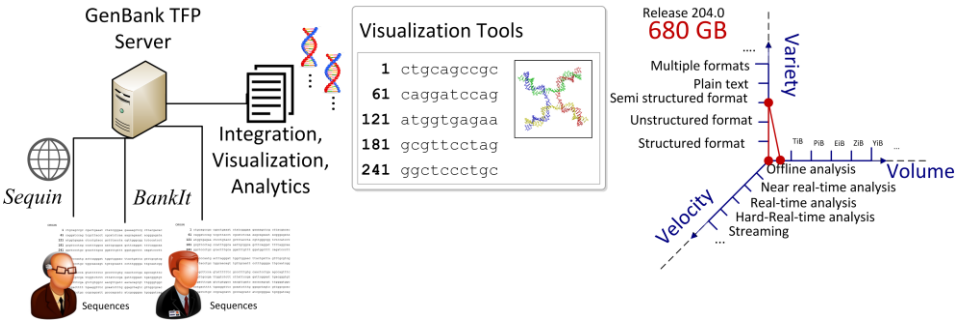


Fig. 4. GenBank Overview

The current database release (204.0) includes a set of 2216 ASCII text files most of which contain sequence data and require roughly 680 GB of disk space for the uncompressed version of them. Each file includes the nucleotide sequence and a header providing the following information: locus, definition, keywords, organism (formal scientific name of the organism and taxonomic classification levels), reference and so on. All sequences are accessible via the GenBank ftp server. This Big Data scenario is characterized by a quite data volume and highly structured data. The velocity and the veracity do not

constitute a problem since GenBank rely on direct authors submission of data to ensure that they achieves its goals of completeness, accuracy, and timeliness. Generally, major challenges in bioinformatics and computational biology regards data analytics, display and integration because computational tools are quickly becoming inadequate for analyzing the amount and the heterogeneity of genomic data that can now be generated (Fig. 4).

3.3 Social Big Data: Twitter case study

With its 284 million monthly active users and 500 million tweets sent per day, Twitter it is not just a social networking service enabling users to send and read short 140-character messages, but it represents a vast array of ideas and opinions, a real global consciousness [17]. Analyzing billions of tweets can help researchers to discover new insights about public health issues and the way disease is spread, can help people during moments of crisis like natural disasters, or help in sentiment analysis form marketing researches [18]. From a Big Data perspective, Twitter poses challenges not only in volume, but also in velocity, since in some cases, e.g. epidemic and disaster prevention and control, strict time deadlines are required for a prompt intervention, and in variety due the complexity of some text analysis and computational linguistics algorithms to identify and extract subjective information from tweets.

4. BIG DATA IN ADVANCED MANUFACTURING

The globalization and the ICT revolution have made this world "flat" [19], i.e. a *level playing field*, where competitors from everywhere have equal opportunities and access to the same global market. To be competitive in this scenario, companies must continuously strive for excellence reducing the design-to-market time, providing customized products at the lowest prices and innovating products and processes. The information and communication technologies become more and more crucial to achieve these goals. In the contemporary scenario, most of the components inside the factories (machines, robots, product lines, plants, etc.) are turning into cyber-physical systems (CPSs): large scale interconnected and heterogeneous systems able to integrate computation with physical process and to communicate with each other [20]. CPSs have a virtual counterpart in the cyber-space accessible by Virtual Reality frameworks, which synchronize the real and virtual representations of the factory [21-24]. They also become great real-time data producers with the widespread presence of small smart sensors at a shop-floor level. The high availability of data collected inside factories from each component and in every product's life cycle phases makes relevant the Big Data problem inside modern

factories. Real-time data produced during the manufacturing process, production planning and control data [25], the fault information data coming from machine fails, all the resource data coming from raw material information and the extra data brought by the market, policy and environmental changing contribute to increase the volume dimension (Fig. 5). Their heterogeneity also poses challenges along the variety dimension. As a matter of facts, data can be physical variable measures over time, statistics, structured or unstructured data coming from legacy or new storage systems, users' activities logs, and so on. Semantic-web oriented solutions, in this case, may help to improve the interoperability between software tools or data sources by adopting extensible and shared data models for the representation of production systems objects, resources, processes and products [21, 26]. Furthermore, discrete event simulation [27, 28] and High-Level Architecture (HLA) for distributed computer simulation systems poses challenges in the velocity dimension, as a lot of run-time data must be transmitted and synchronized between the various nodes of a network [29]. Finally, analytics over Big Data in advanced manufacturing represents a great challenge trying to mine the correlation hidden behind data that seem uncorrelated. By means of analytics tools, marketers or managers are able to find strategic factors, which can be used as guidelines in decision-making. Forecast fluctuating sales orders across a specific region, decide if buy or do not a new machinery, start or do not the production of a new product based on customer surveys, are a few examples of decision-making problems that Big Data analytics might help to resolve.

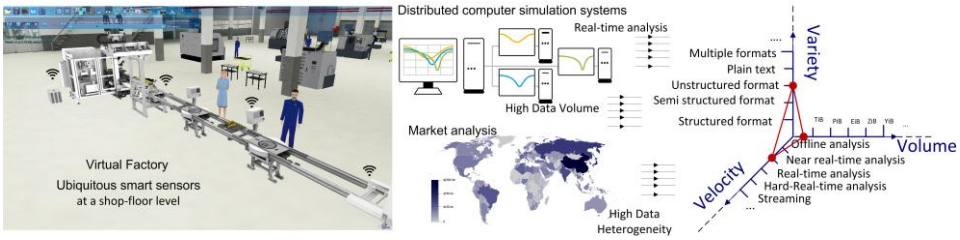


Fig. 5. Advanced Manufacturing Scenario [source: own study]

5. A TECHNOLOGICAL OVERVIEW

Most Data Base Management Systems (DBMSs) are designed for efficient transaction processing: adding, updating, searching for, and retrieving a small amount of information in a large database. Typically, these data sets grow little by little to eventually become Big Data. At this point, difficulties arise when we want to analyze large pile of accumulated data to learn something from them. In this case, relational DBMSs, Data Warehouses and OLAP (OnLine Analytical

Processing) turn out to be too slow or inadequate to face the Big Data analytics challenges [8]. This has led to a proliferation of commercial and open source tools so fare that try to overcome the traditional limitations as for data storage mechanisms or efficient and effective data analysis. As regards to data persistence, for example, No-SQL (Not Only-SQL) storage solutions such as Oracle NoSQL, Big Table, Neo4J, and MongoDB, are increasingly used in Big Data landscape. These solutions are modelled on data structures such as key-value, graph and document, which allow operations that are more efficient over large data sets than traditional databases. Form a distributed computing perspective, the Map-Reduce [30] programming paradigm remains the most used for processing and generating large data set. This paradigm was pioneered by Google and is based on two functions: Map() and Reduce() which orchestrate the data processing and generation across a cluster of machines, by providing fault tolerance and redundancy. The most used implementation of Map-Reduce paradigm is Apache Hadoop but other solutions like Couchdb, MongoDB and MapR are available. Most of Big Data solutions are increasingly available as online services. Amazon Elastic Compute Cloud (EC2) [31], for example, is a web service that provides resizable compute capacity in the cloud, while Amazon S3 provides a fully redundant data storage infrastructure for storing and retrieving great amount of data. In the analytics landscape, various tools are available: from statistical computing frameworks such as R, Matlab, etc. to machine learning algorithms GUIs like WEKA. Finally, it is possible to mention Gephi and GraphViz as data visualization tools and JSON and BSON formats for data serialization.

6. CONCLUSION

As for Big Sciences or Advanced Manufacturing, tools and strategies are needed to get insights and value from the current high availability of data. This work has provided an overview of motivations, applications and solutions in different scenario with a characterization of major challenges posed by each of them. A detailed analysis of Big Data in the Virtual Factory context will be the aim of future works. In particular, the main causes leading to an increase of data available at a shop-floor level in modern manufacturing companies will be further investigated together with the issue of heterogeneity of data sources and the need for run-time analysis in the context of distributed simulation. In addition to this, a qualitative analysis based on functional and technical characteristics of Big Data technologies will be also subject of further researches.

REFERENCES

[1] HILBERT M., LOPEZ P.: *The World's Technological Capacity to Store, Communicate, and Compute Information*. Science 332, 60 (2011).
[2] CASTELLS M.: *End of Millennium, The Information Age: Economy, Society and Culture*. vol. III, Wiley-Blackwell, Malden, MA, 2000.
[3] FRANKS B.: *Taming the Big Data Tidal Wave*. John Wiley & Sons, Inc., 2012.
[4] WEINBERG B.D. DAVIS L., BERGER P.D.: *Perspectives on Big Data*. Journal of Marketing Analytics Vol. 1, 4, pp. 187-201.
[5] McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, May 2011.
[6] MERV A.: *"Big Data,"* Teradata Magazine Online, Q1, 2011.
[7] Kerry Butters, *Big Data Will Die Within Two Years*, OneStopoClick Article, 15 May 2014.
[8] JACOBS A.: *The pathologies of big data.* Communication of the ACM, August 2009, Vol. 52, N. 8.
[9] SAGIROGLU S. and SINANC D.: *Big Data: A Review*. International Conference on Collaboration Technologies and Systems (CTS), 2013.
[10] MANYIKA J., CHUI M., BROWN B. et al.: *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.
[11] MOHANTY S., JAGADEESH M., SRIVATSA H.: *Big Data Imperatives.* Apress.
[12] JAGADISH H.V. et al.: *Big Data and its Technical Challenge*. COMMUNICATIONS OF THE ACM, July 2014, Vol. 57, n. 7.
[13] GABER M. M., ZASLAVSKY A., KRISHNASWAMY S.: *Mining Data Streams: A Review*, ACM SIGMOD Record, Vol. 34, No. 2, June 2005, pp. 18-26.
[14] DESOUZA K. C., SMITH K. L.: *Big Data for Social Innovation*, Stanford Social Innovation Review, 2014.
[15] RIJMENAM M. VAN, Think Bigger. Developing a successful big data strategy for your business, American Management Association, 2014.
[16] HEY T., TANSLEY S., TOLLE K.: *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Microsoft Research, 2009.
[17] SMOLAN R., ERWITT J.: *The Human Face Of Big Data*. Against All Odds Productions, 2012.
[18] PAK A., PAROUBEK P.: *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* LREC, 2010.
[19] FRIEDMAN T.L.: *The World is Flat – A Brief History of the Twenty-first Century*. Farrar, Straus, & Giroux, 2005.
[20] LEE E.A.: *Cyber Physical Systems: Design Challenges*. Electrical Engineering and Computer Sciences University of California at Berkeley, 2008.
[21] KÁDÁR B., TERKAJ W., SACCO M.: *Semantic Virtual Factory supporting interoperable modelling and evaluation of production systems*. CIRP Annals Manufacturing Technology, 62(1), 2013, pp. 443-446.
[22] GHIELMINI G. et al.: *Virtual Factory Manager for semantic data handling. CIRP Journal of Manufacturing Science and Technology*, 2013, 6(4), pp. 281-291.
[23] PALAJOVÁ S., FIGA Š., GREGOR M.: *Simulation of manufacturing and logistics systems for the 21th century*, Applied Computer Science, V. 8, N. 1, 2012.
[24] CHEN D., KJELLBERG T., EULER A. von: *Software Tools for the Digital Factory – An Evaluation and Discussion.* Proceedings of the 6th CIRP-Sponsored International Conference on Digital Enterprise Technology Advances in Intelligent and Soft Computing Volume 66, 2010, pp. 803-812, V. 4, N. 2, 2008.
[25] KOZŁOWSKI E., GOLA A., ŚWIĆ A.: *Model of Production Control in Just-in-Time Delivery System Conditions*. Advances in Manufacturing Science and Technology, Vol. 38, No. 1, 2014, pp. 77-88.

[26] MODONI GE, SACCO M., TERKAJ W.: *A survey of RDF store solutions*. Engineering, Technology and Innovation (ICE), International ICE Conference 2014.
[27] GOLA A., ŚWIĆ A.: *Design of storage subsystem of flexible manufacturing system using the computer simulation method*, Actual Problems of Economics, No. 4 (142) 2013, pp. 312-318.
[28] TERKAJ W., URGO M.: Ontology-based modeling of production systems for design and performance evaluation. Proceedings of 12th IEEE INDIN, 2014, pp 748-753.
[29] PEDRIELLI G., SACCO M., TERKAJ W., TOLIO T.: *An HLA-based distributed simulation for networked manufacturing systems analysis.* Journal of Simulation, 2012, 6(4): 237-252.
[30] DEAN J., GHEMAWAT S.: *MapReduce: simplified data processing on large clusters,* Communications of the ACM, 2008.
[31] Amazon Web Services: *Amazon Elastic Compute Cloud. User Guide for Microsoft Windows*. API Version 2014-09-01, 2014.