Preproceedings of the Federated Conference on
Computer Science and Information Systems pp. 893–898

Big Data Techniques, Systems, Applications, and Platforms: Case Studies from Academia

Atanas Radenski*, Todor Gurov†, Kalinka Kaloyanova‡‖, Nikolay Kirov§‖
Maria Nisheva‡‖, Peter Stanchev¶‖, and Eugenia Stoimenova†‖

* Schmid College of Science and Technology, Chapman University
Orange, CA, USA, Email: radenski@chapman.edu
† Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Sofia, Bulgaria
Email: (gurov,jeni)@parallel.bas.bg
‡ Faculty of Mathematics and Informatics, Sofia University, Sofia, Bulgaria
Emails: (kkaloyanova,marian)@fmi.uni-sofia.bg
§ Department of Informatics, New Bulgarian University, Sofia, Bulgaria
Email: nkirov@nbu.bg
¶ Department of Computer Science, Kettering University, Flint, MI, USA
Email: pstanche@kettering.edu
‖ Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
Emails: (nkirov,stanchev)@math.bas.bg

***Abstract*—Big data is a broad term with numerous dimensions, most notably: big data characteristics, techniques, software systems, application domains, computing platforms, and big data milieu (industry, government, and academia). In this paper we briefly introduce fundamental big data characteristics and then present seven case studies of big data techniques, systems, applications, and platforms, as seen from academic perspective (industry and government perspectives are not subject of this publication). While we feel that it is difficult, if at all possible, to encapsulate all of the important big data dimensions in a strict and uniform, yet comprehendible language, we believe that a set of diverse case studies – like the one that is offered in this paper – a set that spreads over the principal big data dimensions can indeed be beneficial to the broad big data community by helping experts in one realm to better understand currents trends in the other realms.**

***Index Terms*—Metadata, semantic annotations, Spark, NoSQL, data-intensive applications.**

I. Introduction

The principle dimensions of big data include its defining characteristics (such as volume, velocity, variety, and veracity), techniques (such as data mining, machine learning, natural language processing, neural networks, clustering, pattern recognition, sentiment analysis, predictive modeling, supervised learning, time series analysis, to mention a few), software systems (such as Hadoop, Spark, NoSQL DBMSs), applications (such as business analytics, marketing, healthcare, research, performance optimization, security, law enforcement, transportation, and many others), computing platforms (such as clusters, NUMA in-memory database servers, and cloud computing platforms), and big data milieu (such as industry, government, academia).

The big data dimensions are not only broad but also in perpetual change. This is why the task of compiling and maintaining a specification that is rigorous yet comprehensible seems impractical. Instead, we believe that reports like this one, presenting case studies with broad coverage of the big data realm can be beneficial for the broad big data community. Our broad collection of case studies can potentially help experts in one big data dimension expand their understanding of other dimensions.

The technical core of this paper comprises seven case studies. In section II, Techniques, we illustrate the potential of ontologies to semantically enhance data (subsection II.A) and metadata to facilitate image data mining (subsection II.B). In section III, Software Systems, we focus on some of the forces behind the transition from relational to NoSQL DBMS (subsection III.A) and from Hadoop MapReduce to Spark. In section IV, Application Domains, we discuss applications in astronomy and earth science (subsection IV.A) and in biomedical research (subsection IV.B). Finally, section V, Platforms, highlight the convergence of HPC and big data, as seen in the Avitohol computers system (subsection V.B).

All case studies are based on the authors' own research projects.

II. Techniques

A. Semantic Enhancement of Data with Ontologies

The characteristics of big data discussed above create a number of challenges to the methods and tools for their utilization. For example, the volume of data to be processed requires an ability to abstract the data in a form that summarizes the situation and is actionable from the point of view of humans and

decision-making software systems [16]. This requirement for *semantic scalability* is also important in the context of variety of data formats. The latter implies an additional requirement for an ability to integrate and interoperate with heterogeneous data – "to bridge syntactic diversity, local vocabularies and models, and multimodality" [19]. The velocity, i.e. the rapid appearance and change of data, requires the ability to focus on the relevant data and to process it quickly. Data veracity requires the capability of finding anomalies in it and making some types of reasoning based on proper domain knowledge. Extracting value using data analytics methods on various kinds of data creates the need of ability to extract knowledge from data and integrate it with existing knowledge bases.

Such challenges need be overcome to permit big data's full-scale potential for the user. Most traditional utilization approaches do not work in a satisfactory way for big data, so more agile utilization paradigms are needed in this case.

The main idea of the so-called *semantic utilization of big data* is to provide a kind of *semantic enhancement of data* that can be realized with the help of proper ontologies used to annotate data.

An *annotation* is a form of metadata attached to a specific dataset, to a particular database field or to a particular section of a document content. An annotation provides additional information (metadata) about an existing piece of data. Compared to tagging, which speeds up searching and helps one to find relevant and precise information, *a semantic annotation* goes one level deeper: It enriches the unstructured or semi-structured data with a context that is further linked to the available structured domain knowledge and makes it possible to process complex filter and search operations expecting results that are not explicitly related to the original search queries.

Ontologies [6] are the only widely accepted paradigm for the representation and management of open, sharable, and reusable knowledge in a way that allows automatic interpretation and inference. They provide semantic enhancement of data suggesting controlled vocabulary for annotations and thus permitting agile integration and semantic interoperability.

This kind of semantic enhancement of data may be characterized as an "arm's length approach" [15] – it presumes no change of data but association of each database field with an entire knowledge base. Data should be leaved as they are but incrementally tagged with terms from a consistent and non-redundant set of ontologies.

The successful implementation of the discussed approach depends on the creation of a shared resource (for example, a shared repository) of ontologies that could be used for annotation purposes. Moreover, it will be necessary to build an agile methodology for dynamic creation, application and extension of ontologies to annotate new sources of streaming data [15]. Such methodology should define a simple, repeatable process for ontology development and change management as well as an unambiguous process for data annotation using available ontologies.

B. Metadata in Image Data Mining

A most commonly accepted definition of "data mining" is the discovery of "models" for data. A "model", however, can be one of several things [12]. There are different approaches to modeling data. For thousands of years science was empirical. It was only in the last few hundred years that the theoretical paradigm emerged. The data-driven scientific inquiry came with data mining. The typical feature-based model looks for the most extreme examples of a phenomenon and represents the data by these examples. Some of the important kinds of feature extraction from large-scale data are:

1) *Frequent Itemsets* – a model makes sense for data that consists of "baskets" of small sets of items;
2) *Similar Items* – data looks like a collection of sets, and the objective is to find pairs of sets that have a relatively large fraction of their elements in common.

Data mining can be viewed as a result of the natural evolution of information technology. Data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high-performance computing, and many application domains. In the field of image data mining, we developed an approach for extending the learning set of a classification algorithm with additional metadata. It is used as a base for the assignment of appropriate names to found regularities. The analysis of the correspondence between connections established in the attribute space and existing links between concepts can be used as a test for the creation of an adequate model of the observed world. Meta-PGN classifier is suggested as a possible tool for establishing these connections. We applied this approach to the field of content-based image retrieval of art paintings by designing system architecture for the extraction of specific feature combinations, which represent different sides of artists' styles, periods and movements [8]. Technically, we provide the system with a description of the real world and the systems then follows our mental model to generate appropriate names of detected concepts. The system interacts with the user, displaying those parts of the mental model that are utilized in the name generation process. This interaction is used to further improve and extend the mental model.

III. Software Systems

A. NoSQL versus RDBMS

The huge amounts of data that needs to be stored and processed on multiple servers is a recognized challenge of big data.

To manage the integrity of data, the classical database decisions (mostly relational databases) are based on transactions and support the main transactional characteristics – **A**tomicity, **C**onsistency, **I**solation, **D**urability, also known as ACID property.

But relational databases are difficult to scale, and they cannot guarantee increasing expectations for the performance

and availability when it comes to managing huge volumes of data on different servers.

Published by Eric Brewer in 2000, the CAP Theorem sets the basic requirements for a distributed system – Consistency, Availability, and Partition Tolerance [5]:

- Consistency – all the servers in the system have the same data;
- Availability – the system always responds to a request;
- Partition Tolerance – the system continues to operate as a whole even if an individual server fails.

The CAP Theorem postulates that only two of the three different aspects of scaling out can be fully achieved at the same time.

Traditional relational databases (Oracle, MS SQL, IBM DB2, etc.) are architected to run on a single machine and use strong, schema-based approach to modeling data that rely on consistency. So they represent the first group – Consistent, Available (CA) systems. Some column stores like Vertica, etc., also belong in that group.

NoSQL database decisions, on the other hand, are considered as an alternative to relational databases at times when organizations like Google and Amazon recognize that operating at scale is more effective using clusters of servers, and a schema-less data models are a feasible alternative in case of variety of data.

When distributed data stores are used, at the time of network partition, it is not possible to have both Consistency and Availability. So while traditional data-bases focus on Consistency, the NoSQL systems try to focus on Availability. In that case Consistency could be replaced by Eventual Consistency – data is not consistent at all time, but will be at given time or Local Consistency (the consistency is assured only within one single node and not throughout the cluster). Thus most NoSQL databases rely on properties less strict than the ACID ones, which are called BASE – Basic Availability, Soft state and Eventual consistency [13].

Consistent and Partition-Tolerant (CP) systems like HBase, MongoDB, and Big Table have difficulty to achieve data consistency over partitioned nodes, while Cassandra, Couch DB and Dynamo that support AP (Availability, Partition-Tolerance) achieve ”eventual consistency” through replication and verification.

Furthermore, the first group contains systems that support even ACID properties – like Dynamo, but most systems conform to BASE properties.

In the majority of cases non-traditional systems yield a better performance when ordinary data operations are measured. Our experiments on Oracle, Vertica, and Mongo DB platforms present particular results that confirm this thesis [10], [11].

B. From Hadoop MapReduce to Spark

Spark, initiated at the University of California, Berkley in 2009, was donated in 2013 to Apache. As an Apache project, Spark has gained popularity as a flexible and efficient in-memory implementation of the map-reduce distributed computing model and has already emerged as a faster substitute for the original Hadoop MR in-disk engine. Early Apache projects, such as Hive and Mahout, which originally compiled into Hadoop MR [27], have now been implemented to run on Apache Spark. Besides speed, Apache Spark's advantages to Hadoop MR include capabilities for interactive computing, stream processing, and sensor data processing (which are all lacking from the rigid two-stage Hadoop MR engine). While Apache Spark can be used within Apache Hadoop, it can also run independently, together with its own libraries, such as Spark SQL, Spark Streaming, Spark GraphX, and MLib (machine learning).

It has been broadly acknowledged that Spark has a pronounced efficiency edge over MR, but strict performance comparisons and analysis were scarce before 2014. In late 2014, Databricks, the company founded by the creators of the original Spark, released big data benchmark results that illustrate the speed advantages of Spark over Hadoop MR [26]. Spark was reported to perform three times faster than Hadoop MR on a 100 TB sort workload, and four times faster on a 1 PB workload, using – in both cases – significantly less hardware.

In early 2015, [14] reported performance experiments with codon count algorithms on nucleotide sequence data on AWS (the Amazon cloud computing platform). To do so, the authors measured the performance of a basic Spark codon count algorithm and compared it to a couple of Hadoop MR algorithms: a basic Hadoop MR codon count algorithm and an optimized ”local aggregation” Hadoop MR algorithm. The experiments confirmed that basic codon count with Spark is 15 times faster than basic codon count with MapReduce, a result that is unsurprising. Unexpectedly, however, basic codon-count with Spark remains about two times slower than optimized ”local aggregation” codon count with Hadoop MR. This result hints that properly optimized Hadoop MR code can be faster than the same analysis with Spark. The authors therefore suggest that available optimization techniques, such as local aggregation, be considered for speeding-up of legacy Hadoop MR applications in place of their eventual re-implementation in Spark for performance gains.

IV. Application Domains

A. Astronomy and Earth Science

With the current emergence of terabyte– and very soon petabyte-scale astronomical and Earth observation systems, the traditional approach to basic functions such as data searching, analytics and visualization are becoming increasingly difficult to handle. Simple database queries can result now in data subsets so large that they are incomprehensible, slow to handle, and impossible to visualize with commodity visualization tools. Astronomy and remote sensing complement each other, as they are on the quest for new big data interpretation capabilities: both disciplines have peculiar data, typical data processing and analysis chains, and specific models to be fed with data. However, both disciplines lack the capabilities for easily accessible semantics-oriented browsing in large data archives. Therefore, joint efforts to design and develop

innovative big data tools should help users in many different fields and set new standards for many communities. Several broad challenges to this line of reasoning that demand a multidisciplinary approach through international networking of experts and professionals have been identified. These challenges would then be channeled into COST Action TD1403 Objectives [1]:

- Digital curation and data access;
- New frontiers in visualization;
- Adaptation to new high performance computing technologies;
- New generation of interdisciplinary scientists.

The COST Action TD1403 was launched in April 2015 and will last for two years with a possible extension of 2 more. Now it involves 26 European countries. BigSkyEarth COST Action is organizing meetings, workshops, training schools and conferences. Also it supports Short Term Scientific Missions – exchange visits for individual mobility, strengthening existing networks and fostering collaboration between researchers.

The Bulgarian participation in the Action is in connection with a group of experts in astro-informatics – astronomers, mathematicians and computer science specialists [9]. Our fields are digitization of widefield (larger than or equal to 1 square degree) astronomical photographic plates, image processing and image compression.

The Wide-Field Plate Database [24], established in 1991, is the basic source of data for the wide-field plates obtained with professional telescopes world-wide [22]. It consists of four parts:

- Catalogue of Wide-Field Plate Archives with data for over 500 instruments (telescopes, cameras, etc.);
- Catalogue of Wide-Field Plate Indexes with descriptions of about 600 000 plates;
- Data Bank of digitized plate images (at low resolution for quick plate visualization and easy on-line access, and at high resolution intended for photometric and astrometric measurements);
- Links to online services and cross-correlation with other existing catalogues and journals.

We have identified more than 2 400 000 wide-field plates [21]. They allow one to obtain information of celestial objects over the past 133 years (1872-2005). At present over 300 000 wide-field plates have been digitized with total data about 30 TB.

Digitized photographic plates are irreplaceable sources for:

- Studies of the stellar long term brightness changes, as a result of observations conducted in different observatories;
- Studies of the long term variability of active galaxies;
- Searching and identification of potentially hazardous asteroids and comets which might cause catastrophic events by their collision with Earth.

B. Biomedical Research

Stimulated by the progress in computer technology and electronics data acquisition, recent decades have seen the growth of huge databases in biomedical sciences. For instance, Next generation sequencing (NGS) is a significant technological advance in biomedical sciences. It generates massive genomic datasets that play a key role in the big data phenomenon that surrounds us today. Advancing machine learning, data mining and statistical techniques for processing of big data are the key to transforming big data into actionable knowledge. One major problem with big data is that the standard methods of applied statistics are not really relevant for big data analysis. To extract information from high-dimensional data sets and make valid statistical inferences and predictions, novel data analytic and statistical techniques are needed. Here are some modern issues that we focus on.

Current advances in biomedical research technology, expression and SNP microarrays yield big data sets for many thousands of transcripts, genes or SNPs. Researchers are often interested in finding differences among these features between two separate groups, e.g. patients and controls, treatment and control groups; different strands; different tissues etc. Due to the differences of the underlying technologies and their biophysical and biochemical processes, scientists need to use statistical data analysis methods designed specifically for the particular technology. These tests often employ multiple comparison designs, where each gene, transcript or SNP is separately tested for significance and in many situations these tests are conservative. In complex multiple testing hypotheses, the classic statistical tests overestimate the p-values, leading to both loss of statistical power and increased experimental costs. One really common choice for correcting for multiple testing is to use the false discovery rate to control the rate at which things you call significant are false discoveries. There has been recent interest in developing efficient algorithms for multiple comparison to increase the statistical power and reduces the experimental costs. A computationally efficient technique has been proposed recently [4] that increases the statistical power, while controlling the False Discovery Rate of the statistical tests. This technique is applied to DESeq – a popular method for finding differentially expressed genes using RNA-sequence data. The statistical power increase is particularly high in small sample size experiments, often used in preliminary experiments and funding applications.

Some other issues arise from the method for finding differentially expressed gene. Apart from the DESeq method there are a few more like edgeR and limma frequently used by biomedical researchers [17]. These methods often produce similar, but not identical results. At the same time, due to randomness, even the same method can produce slightly different results on a data simulated from the same model. Therefore we are interested whether the slightly different results produced by different models can be attributed to randomness or to an underlying difference in the methods. Since the gene sequence is very long, we might not be interested in the full ranking

of the p-values but in some incomplete or partial orderings of them. Consequently, we want to compare such partial orderings. For example we can split the genes into several groups according to the size of their p-values lying in the sub-intervals $[0, 0.001]$, $[0.001, 0.01]$, $[0.01, 0.05]$, $[0.05, 0.1]$, $[0.1, 1]$. Then we construct a distance measure to compare how similar/dissimilar two incomplete rankings are based on the number of items present in the same ordered groups in both rankings [18]. Based on simulated large number of rankings and computed distance between any two of them, we can make inferences about the significance of a particular distance. That is to estimate the similarity between the corresponding incomplete rankings. Scientific computing is involved in all of these steps: simulating incomplete rankings, applying the method for finding differentially ex-pressed genes, computing all distances and estimating the distribution of the distance. We use the advanced computing resources at the Institute of Information and Communication Technologies (IICT) [7].

V. Platforms

Academic organizations are already moving towards unifying their HPC and big data processes within integrated HPC/big-data platforms, as observed in the development of the Avitohol platform at the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences.

As owner and manager of the Advanced Computing and Data Centre [7], IICT provides advanced computing resources and expertise thus helping Bulgarian science to come at the forefront of development worldwide.

The new multifunctional High Performance Computing system – Avitohol, forms the core of the computing infrastructure in the Institute. It consists of 150 computational servers HP SL250s Gen8, equipped with two Intel Xeon E5-2650v2 CPUs and two Intel Xeon Phi 7120P coprocessors, 64 GB RAM, two 500 GB hard drives, interconnected with non-blocking FDR InfiniBand running at 56 Gbp/s line speed. The total number of cores is 20700 and the total RAM is 9600 GB, respectively. The servers are deployed in 4 dual racks HP MCS 200, which have water cooling and can deliver up to 50 kW of power per rack. A central rack contains most of the storage, management servers and the central communication switches.

The system is controlled by two management (head) nodes and 4 I/O nodes. All those nodes are of the type HP ProLiant DL380p Gen8 with 2 Intel Xeon E5 2650v2 CPUs and 64 GB RAM. The I/O nodes provide access to 96 TB of raw storage capacity (24 disks of 4 TB each), which is provided by a SAN system.

The theoretical peak performance of the system is estimated at 412.3 TFlop/s in double precision while the RMAX Performance according the LINPACK benchmark is 264.2 TFlop/s. The Avitohol HPC system has been operational since June 2015 and it is ranked on 388th place according the 46th T0P500 list [20].

The second advanced computing system at IICT is the heterogeneous High Performance Computing Grid (HPCG) cluster [7]. It has been operational since 2010 and consists of HP Cluster Platform Express 7000 enclosures with 36 blades BL 280c (Total 576 CPU cores), 24 GB RAM per blade; 8 controlling nodes HP DL 380 G6 with dual Intel X5560 2.8 GHz, 32 GB RAM (total 128 CPU cores); 2 HP ProLiant SL390s G7 4U servers with 16 NVIDIA Tesla M2090 graphic cards (total 8192 GPU cores); 2 HP SL270s Gen8 4U servers with 8 Intel Xeon Phi 5110P Coprocessors each (total 480 cores, 1920 threads); 3 SAN storage systems with 132 TB total storage. All servers are interconnected with FDR InfiniBand running at 56 Gbps line speed. The theoretical peak performance of the system is estimated at 21.92 TFlops/s in double precision.

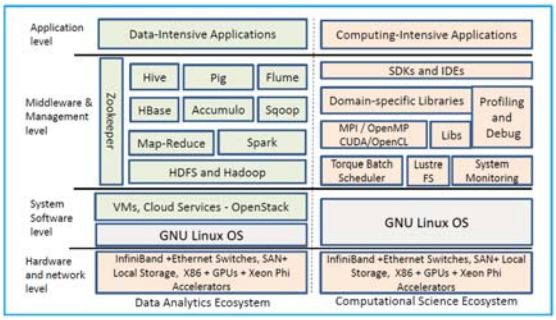

Fig. 1. Data analytics and computational ecosystems.

A dedicated optical network link has been established between the two systems, as well as between the Avitohol system and the main node of the Bulgarian Research and Educational Network (BREN) [2].

The existing computing facilities at IICT are involved in two computational infrastructures: the European Grid Infrastructure [3] and regional VI-SEEM infrastructure [23].

Based on their experiences in the last 10 years, the scientific and support staff in the center for HPC and Data computing at IICT [7] is dedicated to providing full support for the development and deployment of innovative scientific and industrial applications with substantial needs for computing power and data storage and transfer. The system can be considered to have two equally important sides. On the one hand it allows for state-of-the-art high performance computations, providing a full stack from base operating system software and libraries up to specially configured and deployed applications that make full use of the special capabilities of the systems, e.g. the Xeon Phi accelerators and CUDA GPGPU devices and the InfiniBand network. On the other hand, the storage systems provide access to data using various base protocols. The Lustre file system is the most frequently used for HPC workloads, while protocols like iSCSI are used for Cloud provisioning and other types of data processing. The data processing capabilities of the new Avitohol system are currently under configuration and testing, with the aim to build-up the full ecosystem with Hadoop and HDFS as the base layer and the components like Hive, Spark, Pig, etc., working on top of it. We plan to allow these components access to the Xeon Phi coprocessors

for advanced deep learning and related algorithms. Such new data processing capabilities will foster the development of integrated scientific applications that are based on realtime data, coming from local and international sources. Fig. 1 shows a schematic representation of the hardware and software components that are available or planned for deployment. System architects and engineers are developing a mixed HPC/Big data cluster, to provide for the convergence of HPC and Big data computing into a "single" environment.

The current goal of IICT is to achieve petaflop and petabyte-level of computing and storage capability, coupled with a developed software and middleware stack and services, opening the way to new forms of scientific research, more directly connected with the national industry and the societal challenges.

VI. Conclusion

This paper offers seven case studies that span several dimensions of big data: techniques, systems, applications, and platforms. All case studies are extracted from current research projects of the authors themselves. The case studies cover various aspects of big data and can, hopefully, be beneficial to the broad big data community by helping experts in one realm get acquainted with current cases in other realms.

The main contributions of the individual authors can be described as follows. M. Nisheva presented the potential of ontologies to semantically enhance data (section II.A). P. Stanchev discussed the use of metadata in image data mining (section II.B). K. Kaloyanova reviewed the capabilities of NoSQL databases as opposed to RDBMS (section III.A). N. Kirov described utilization of big data in Astronomy and Earth Science (section IV.A), while E. Stoimenova highlighted the specifics of statistical applications in biomedical research (section IV.B). T. Gurov focused on the convergence of HPC and big data, as realized with the Avitohol platform (section V.B). A. Radenski discussed the transition from Hadoop MapReduce to Spark (section III.B). A. Radenski also planned the overall structure of this publication and drafted the abstract, the introductory section I (minus the specification of the big data characteristics), the background section V.A, and the concluding section VI.

Acknowledgment

The work of the first author (A.R.) was supported by an AWS in Education 2014-2015 award from Amazon. The work of the second author (T.G.) was partly supported by the National Science Fund of Bulgaria under Grant DFNI-I02/8 and by EC Programme Horizon 2020 under project VI-SEEM project, Grant Agreement No: 675121. The work of the last author (E.S.) was supported by the National Science Fund of Bulgaria under Grant DFNI-I02/19.

References

[1] Big Data Era in Sky and Earth Observation (Big-SkyEarth) COST Action TD1403, http://bigskyearth.eu/, (Retrieved January, 2016).
[2] Bulg. Research and Educational Network (BREN), http://www.bren.bg/.
[3] EGI, 2016, European Grid Infrastructure, www.egi.eu .
[4] J.P. Ferguson, D. Palejev, "Calibration of p-values for multiple testing problems in genomics", *Stat. Appl. Genet. Molec. Biol.*, vol. 13(6), 2014, pp. 659–73, doi: 10.1515/sagmb-2013-0074.
[5] S. Gilbert, N. Lynch, "Perspectives on the CAP Theorem", *Computer*, vol. 45, no. 2, 2012, pp. 30–36, http://doi.ieeecomputersociety.org/10.1109/MC.2011.389.
[6] T. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", *International Journal of Human-Computer Studies*, Vol. 43, 1995, pp. 907–928, doi:10.1006/ijhc.1995.1081.
[7] Advanced Computing and Data Center, IICT, http://hpc.acad.bg/.
[8] K. Ivanova, I. Mitov, P. Stanchev, Ph. Ein-Dor, K. Vanhoof, "Establishing Correspondences between Attribute Spaces and Complex Concept Spaces Using Meta-PGN Classifier", *Proc. of the 2nd Int. Conf. "Digital Preservation and Presentation of Cultural Heritage"*, V. Tarnovo, Bulgaria, IMI-BAS, Sofia, 2012, ISSN 1314-4006, pp.71–77.
[9] O. Kounchev et al, Astroinformatics: "A Synthesis between Astronomical Imaging and Information & Communication Technologies", *In: Modern Trends in Mathematics and Physics*, ed. S.S. Tinchev, Heron Press, Sofia, 2009, pp. 60–69.
[10] H. Kyurkchiev, K. Kaloyanova, "Performance Study of Analytical Queries of Oracle and Vertica", *Proc. of the 7^{th} International Conference "Information Systems & Grid Technologies"*, Sofia, 2013, pp. 127–139, DOI:10.13140/2.1.3667.0726.
[11] H. Kyurkchiev, E. Mitreva, "Performance Study of SQL and NoSQL Solutions for Analytical Loads", *Proc. of the Doctoral Conference in "Mathematics, Informatics and Education" (MIE2013)*, Sofia, 2014, pp. 49–57, DOI: 10.13140/2.1.1307.7766.
[12] J. Leskovec, A. Rajaraman, J. D. Ullman, *Mining of Massive Datasets*, 3rd Edition, Cambridge University Press, 2014.
[13] E. Mitreva, K. Kaloyanova, "NoSQL solutions to handle big data", *Proc. of the Doctoral Conference in Mathematics, Informatics and Education (MIE 2013)*, Sofia, 2013, pp. 77–85.
[14] A. Radenski, L. Ehwerhemuepha, K. Anderson. "From in-disk to in-memory big data with Hadoop: Performance experiments with nucleotide sequence data", *Proc. ABDA'15, the International Conference on Advances in Big Data Analytics*, CSREA Press (H. Arabnia and M. Yang, Ed.), 2015, pp. 34–40.
[15] D. Salmen, T. Malyuta, A. Hansen, S. Cronen, B. Smith, "Integration of Intelligence Data through Semantic Enhancement", *Proceedings of the Conference on Semantic Technology in Intelligence, Defense and Security STIDS 2011*, CEUR, Vol. 808, 2011, pp. 6–13.
[16] A. Sheth, "Transforming Big Data into Smart Data: Deriving Value via harnessing Volume, Variety and Velocity using semantics and Semantic Web", *Keynote at the 21st Italian Symposium on Advanced Database Systems* 2013. http://j.mp/SmatData, visited on December 23, 2015.
[17] C. Soneson, M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data", *BMC bioinformatics*, vol. 14(1), 2013, 1, DOI: 10.1186/1471-2105-14-91.
[18] E. Stoimenova, D. Palejev, "Comparison of incomplete ranked lists with application to RNA-seq differential expression methods", Preprint, 2016.
[19] K. Thirunarayan, A. Sheth, "Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications", Association for the Advancement of Artificial Intelligence (AAAI) Technical Report FS-13-04, 2013.
[20] TOP500 list, November 2015, http://www.top500.org/site/50586.
[21] M. Tsvetkov, K. Tsvetkova, N. Kirov, "Technology for scanning of astronomical photographic plates", *Serdica Journal of Computing*, vol. 6 (1), 2012, pp. 77–88.
[22] M. Tsvetkov, "Wide-Field Plate Database: a Decade of Development", *In: Observatory: Plate Content Digitization, Archive Mining and Image Sequence*, Processing, iAstro workshop, Sofia, Bulgaria, Eds. M. Tsvetkov, F. Murtagh, R. Molina, 2006.
[23] VI-SEEM, 2016. https://vi-seem.eu/.
[24] Wide-Field Plate Database, http://www.wfpdb.org, (Retrieved January, 2016).
[25] T. White, *Hadoop: The Definitive Guide*, O'Reilly, 2012.
[26] A. Woodie, Spark Smashes MapReduce in Big Data Benchmark. Datanami, October 10, 2014, http://www.datanami.com/2014/10/10/spark-smashes-mapreduce-big-data-benchmark/.
[27] E. Zdravevski, et al, "Parallel computation of information gain using Hadoop and MapReduce", *Proc. of the 2015 Federated Conf. on Comp. Sci. and Inf. Systems" (FedCSIS2015)*, Lodz, Poland, 2015, pp. 181–192, DOI: 10.15439/2015F89.