Big Data Analytics as-a-Service: Issues and challenges

Claudio A. Ardagna, Paolo Ceravolo
DI, Università degli Studi di Milano
Crema, 26013, Italy
Email: {firstname.lastname}@unimi.it

Ernesto Damiani
Consorzio Interuniversitario Nazionale per l'Informatica
Roma, 00198, Italy
EBTIC, Khalifa University
Abu Dhabi, UAE
DI, Università degli Studi di Milano
Crema, 26013, Italy
Email: ernesto.damiani@kustar.ac.ae

***Abstract*—Big Data domain is one of the most promising ICT sectors with substantial expectations both on the side of market growing and design shift in the area of data storage managment and analytics. However, today, the level of complexity achieved and the lack of standardisation of Big Data management architectures represent a huge barrier towards the adoption and execution of analytics especially for those organizations and SMEs not including a sufficient amount of competences and knowledge. The full potential of Big Data Analytics (BDA) can be unleashed only through the definition of approaches that accomplish Big Data users' expectations and requirements, also when the latter are fuzzy and ambiguous. Under these premises, we propose Big Data Analytics-as-a-Service (BDAaaS) as the next-generation Big Data Analytics paradigm and we discuss issues and challenges from the BDAaaS design and development perspective.**

***Index Terms*—Big Data; Big Data analytics; Issue and challenges.**

1. Introduction

Big Data has recently become a major trend attracting both academia, research institutions, and industries, with a potential market of 187 billion dollar by 2019 and an increasing rate of 50% over five years [1]. Today pervasive and interconnected world, in fact, make people at the center of a continuous sensing process, where an enormous amount of data are generated and collected every minute. In particular, according to [2], every human in the world is producing over 6 megabytes for minute, a total of 1.7 million billion bytes of data. Also, the Compliance, Governance and Oversight Council claimed that information volume doubles every 18-24 months for most of organizations, while 90% of data have been collected in a couple of years [3].

Many organizations, in all domains, have discovered that, to become or remain competitive, they have to deal with business cases where the volume of data reaches terabytes and even petabytes, often with a rich variety of datatypes to be considered. Clearly, low latency access to this huge amount of distributed data represents a competitive advantage in the market, especially for business intelligent applications [4]. Different IT companies then propose to their customers to manage Big Data challenges using a mix of technologies going from NoSQL ("notonlySQL") databases like Cassandra or HBase, data preparation utilities like Paxata, and distributed, parallel computing systems like Hadoop or Stark. However, the level of complexity achieved and the lack of standardisation of Big Data management architectures represents a huge barrier towards the adoption and execution of analytics. Another major hindering factor to Big Data Analytics (BDA) adoption is the so-called "regulatory barrier:" concerns about violating data access, sharing and custody regulations when using BDA, and the high cost of obtaining legal clearance for their specific scenario are discouraging companies, particularly SMEs, from taking over BDA. Finally, the scarcity of skills makes Big Data scientists, architects, and developers experienced with Big Data projects costly and in high demand internationally. Even outsourcing BDA to a service provider and/or engaging consultants do not eliminate the need of costly in-house skills; the supply of data-related competences internationally will not be able to satisfy the demand in the short and medium term.

Following the above discussion, Big Data introduces two conflicting requirements that need to be reconciled in order to unleash its full potential: *i)* the need of combining complex skills related to data analytics and system architectures; *ii)* improving the acceptance level and the usability of Big Data technologies. Evidently, the ever-increasing complexity of analytics, managing high-dynamics and heterogeneous data and multiple data types, contrasts with the need to reach customers with low Big Data skills. The Big Data community is therefore at a turning point: *how to keep complexity under control? How can Big Data Analytics be granted to any users? Is it possible to make next-generation Big Data Analytics simpler? Which are the issues and challenges introduced by the next generation of Big Data Analytics?*

To answer the above questions, we introduce the concept of *Big Data Analytics-as-a-Service* (BDAaaS) and discuss

issues and challenges introduced by this new concept. To this aim, we first give an overview of Big Data concepts (Section 2); we then define and compare BDA and BDAaaS providing some relevant application scenarios (Section 3); we finally discuss issues and challenges related to BDAaaS. In particular we focus our attention on: quality and diversity, security and privacy, configurability and negotiation, SLAs and assurance, societal and organizational aspects (Section 4).

2. Big Data: Overview

Big Data refers to datasets whose characteristics make typical database approaches unable to store, analyze, and manage them [5]. Big Data techniques and technologies are often seen as techniques and technologies able to work on huge amount of data with good performance. However, Big Data are much more and recently have been defined using the 5V model: Volume, Variety, Velocity, Value, Veracity [6], [7]. In short, this model highlight how Big Data technologies are oriented to the implementation of distributed analytics, that is, scalable analytics handled in architectures that adapt computational resources to the volume, variety and velocity of data, accepting the increasing difficulty in controlling data quality and trustworthiness (veracity), and making possible to radically improve the value generated by offering results at runtime.

Big Data technology and services market represents a fast-growing multibillion-dollar worldwide opportunity and is expanding rapidly [1]. This rise of data and data analytics offers huge opportunities for existing organizations as well as for new start-ups, and both public and private organizations have been instrumental in accelerating the development of a competitive Big Data market. A first result of this rise has been the definition of a variety of technological stacks that, if on one side, increase the possibilities for final users, on the other side, make the selection of the proper solution difficult also for users with good-level ICT competences. Big Data technologies deal with important issues such as task partitioning, scalability, and data normalization, which traditionally represented the main barriers towards the adoption of traditional database approaches.

In other words, without a vigorous research and innovation effort, Big Data solutions may well continue to fall short of widespread adoption due to their costs, usability problems, and proliferation of technological solutions. In the following, we summarise the general hurdles that Big Data technology is facing today.

The technology opacity hurdle. While Big Data analytics can in principle support existing or new value propositions in a number of business domains, choosing and deploying the "right" analytics on the "right" computational infrastructure is still more an art that a science or an engineering practice [8], [9]. Today, only large organizations with deep pockets can afford going trial-and-error for weeks on failure-prone, resource intensive Big Data projects. If SMEs and other limited budget actors like start-ups and no-profit organization have to join the Big Data ecosystem, provisioning a Big Data analysis process must become fast, transparent, affordable, repeatable and robust.

The data diversity hurdle. According to the current Big Data hype, the world is awash in readily accessible Big Data having common time, location, and identity references. Reality is very different. Over-The-Top (OTT) operators, like Google, have proprietary, semantically rich application and query data; they can conceivably expand their scope, creating uniform location and identity systems. In general, data diversity is much higher [10], [11], [12]: utility companies own sensor, management and billing information, telecommunication operators provide a number of location and identity systems, while public administrations of different member states offer different open data on their territory and urban environments. With respect to relatively uniform OTT data, these multi-owner data are highly diverse: they differ in volume (involving small giga-scale and large peta-scale data sizes), granularity, veracity as they are collected and stored by different owners on different platforms. Also, their processing exhibits diverse characteristics such as different data access patterns and different locality.

The security/privacy compliance hurdle. Many domains where Big Data can make a real difference (healthcare, transportation, energy, even entertainment) are highly regulated for security and privacy [13], [14]. The peculiarities of judicial space cannot be addressed on a project-by-project basis. Rather, certified compliance of each Big Data analysis process (e.g., in the form of a Privacy Impact Analysis and privacy controls) must be made available from the outset to all actors that use Big Data in their business model.

The legal hurdle. According to current trends in data management, businesses are increasingly interested in BDA and governments around the world are committed to make data publicly available and usable. Data management however comes with legal issues that need to be accomplished by a proper approach to BDA. How to account intellectual property and how to shape the economical exploitation of data in distributed environments, especially when third parties are involved [15]? How to provide evidence that data processing is compliant to norms and directives [16]? Those are among the questions that still require mature and reliable solutions.

As a result, Big Data acceptance rate is slower than expected [17] [18], and points to the need of automatic approaches that instil the competences of data scientists and data technologist in a single framework. For example, IDC [19] reports that 60% of organizations are hampered by too little business intelligence and only 10% of employees are satisfied with the Big Data technology resources available [19]. Big Data Analytics-as-a-service can play a role in bringing Big Data to the mass, representing the entry point also for companies lacking Big Data skills and competences.

3. Big Data Analytics-as-a-service

The BDAaaS paradigm represents the next evolution step of Big Data to accomplish the hurdles discussed in Section 2. It consists of as a set of automatic tools and methodologies that allows customers lacking Big Data expertise to manage BDA and deploy a full Big Data pipeline addressing their goals. BDAaaS can be seen as a function that takes as input users' Big Data goals and preferences and returns as output a ready-to-be-executed Big Data pipeline.

Users with different skills and expertise can benefit by using a BDAaaS paradigm. Users lacking expertise proper of data scientists (e.g., modeling, analysis, problem solving) can use a BDAaaS solution for preparing the real analytics, reason on data to find out hidden patterns and information, and solve business problems. Users lacking expertise proper of data engineers (e.g., builds a robust and fault-tolerant data pipeline, install a Big Data system) can use a BDAaaS to automatically identify and deploy the proper set of technologies that accomplish their requirements. Users lacking both expertise can still use BDAaaS solutions for a proper initiation in the Big Data realm.

Users' requirements are in the form of platform-independent declarative goals, which are then transformed in low-level platform-dependent configurations of the Big Data pipeline. Requirements can be defined in five different conceptual areas as follows:

- **Data preparation** specifies all activities aimed to prepare data for analytics. For instance, it defines how to guarantee data owner privacy.
- **Data representation** specifies how data are represented and expresses representation choices for each analysis process. For instance, it defines the data model and data structure.
- **Data analytics** specifies the analytics to be computed. For instance, it defines the expected outcome and the type of analytics.
- **Data processing** specifies how data are routed and parallelized. For instance, it defines the processing type and the parameters driving a map-reduce processing.
- **Data visualization and reporting** specifies an abstract representation of how the results of analytics are organized for display and reporting. For instance, it defines visualization type and visual density.

BDAaaS paradigm applies to Big Data scenarios involving enterprises that, for different reason, cannot rely on the adequate level of Big Data competences and/or on skilled data scientists and engineers. In the following, we discuss the issues and challenges introduced by the BDAaaS paradigm.

4. Issues and Challenges

4.1. Data Quality and Diversity

4.1.1. Entity reconciliation. Entity reconciliation, also known as record linkage, concerns the identification of records that refers to the same entity across different data sources. This task is crucial in data integration where different sources may adopt different identifiers for the same entity or more generally may refer to related data but no explicit linkage is available. The techniques traditionally proposed are based on a probabilistic evaluation of the matching intensity between two records belonging to different data sources.

Issues and Challenges: In the context of a parallelised computation the techniques today adopted must be revised as the matching algorithm too must be reorganised according to a parallelised approach. This also implies developing fault-tolerant evaluation of the results acquired from the matching functions, to avoid noisy data is altering the reconciliation process. Top-down or up-front approaches can also be considered to manage disambiguation between entries. However, as illustrated in [20], semantic heterogeneity may results from applying a common model to data sources that was not originated from a same process.

Possible Solutions: BDA should consider entity reconciliation as a relevant step in data preparation offering to the user a pallet of techniques for adressing it ranging from probabilistic to up-front methods.

4.1.2. Data diversity. The data handled via Big Data technology can present variety of formats with different constraints on their structure. Centralized systems, such as for instance Enterprise Resource Planning (ERP), enforce a predefined structure on data (structured data). In distributed environments no enforcement is done on the content level then content is usually encapsulated in meta-data. When data is generated to be consumed by humans, such as in forms, emails, text, videos, audio, and images, the format is unstructured, except for a metadata level that can be applied at data preparation stage. Today, the rise of cloud, IoT, and Big Data paradigms has resulted in the proliferation of sensors and probes of any kind. Moreover, data are often multi-owner and therefore show different format, volumes, granularities. In this scenario, collected data have an even higher level of diversity and heterogeneity.

Issues and Challenges: The problem of data diversity is even exacerbated in a BDAaaS scenario, where the addition of a new and diverse dataset must be accomplished in an automatic way. Traditional normalization approaches must be supported by techniques that adapt the deployed Big Data pipeline to the new source of data. Pipeline adaptation should keep the complexity under control and reduce the points of variability touched by its activities.

Possible Solutions: Given the high level of automation promised by BDAaaS, traditional techniques should be extended to accomplish the intrinsic data variety. Normalization extended with adaptation techniques can also be used as a possible approach. Predefined patterns for on-the-fly integration of new and diverse datasets should also be prepared.

4.1.3. Accuracy. Being able to provide accurate results as the output of analytics is fundamental to avoid completely missing the analytical goal. However, in general, this implies to be aware of the accuracy level of a dataset before processing it. This problem may be even stronger with non-stationary data series.

Issues and Challenges: Big Data accuracy is strongly connected to the need of measuring the distance between the user's expectations and the provided Big Data analytics. The accuracy strongly depends on the specificity of the analytics requirements proposed by the users. The more the requirements are detailed, the higher is the accuracy. In case of abstract requirements, different Big Data pipelines can be deployed by BDAaaS. Then, the choice of the proper pipeline points back to the traditional scenario where choosing and deploying the "right" analytics on the "right" computational infrastructure is more an art than a science. To get rid of such scenario, a BDAaaS solution should correctly represent the knowledge that in current solutions are owned by Big Data experts.

Possible Solutions: Adaptive optimisation based on feedback on the performance of previous BDAaaS deployments can play a role to increase accuracy of the computation. Alternative deployments for a single request might be weighted on the basis of performance observed in similar cases (i.e., for a similar set of requirements).

4.1.4. Usability. Usability represents a measure of the effectiveness and efficiency in achieving a goal. When speaking about data usability we can consider two aspects. On one side, data usability may increase with the availability of a more compact description of the dataset, offering to the users a clear understanding of the dimensions of data that are affecting a phenomenon. On the other side, the mapping between data and corresponding analytics would become cumbersome when data are not stored using human-friendly representations.

Issues and Challenges: As a BDAaaS solution is expected to be configured from high level goals it is very relevant to offer to the user an understanding of the data to be processed. A clear understanding about the interconnections between data format, data integration issues and analytics is also required when issuing a Big Data pipeline.

Possible Solutions: Summarization is a key data mining concept which involves methods for finding a compact description of a dataset. Summarization can be viewed as compressing a given set of transactions to a smaller set of patterns while retaining the maximum possible information. The capability of previewing a BDAaaS solution based on a trial that test and demonstrate the feasibility of a solution is also an important line of action to be implmented.

4.1.5. Trust and provenance. Being able to verify data provenance is fundamental for a proper Big Data management and at the basis of a trustworthy Big Data analytics. Accuracy and usefulness of the results in fact depend on the quality and origin of data, which in turn contribute to increase the trust of the final users in the process producing such results. In addition, when outsourcing Big Data computation to third parties, trust in third parties processes and activities is mandatory. Third parties need to provide evidence of their behavior and proper data management.

Issues and Challenges: Similar to a traditional Big Data scenario, BDAaaS should provide means to verify the data and their origin. When trust is concerned, BDAaaS requires stronger means for trust establishment. In fact, Big Data users' are not only outsourcing their data and Big Data computation, but also all choices regarding the deployment of the Big Data pipeline. Monitoring and testing components should be deployed as well to verify the behavior of BDAaaS solutions.

Possible Solutions: Traditional approaches to data integrity and non-repudiation should be applied in conjunction with assurance and SLA verification techniques. Trust and provenance in BDAaaS is not only related to data verification, but it also needs to verify the BDA pipeline behavior, as well as the provenance of retrieved results.

4.2. Security and Privacy

Security and privacy are play an important role in hindering paradigms based on data outsourcing. Moving not only data but also computation to external infrastructures further increases the concerns of users about their security and privacy; data loss, data breach, and data theft become critical threats to organization assets. Moreover, the distributed and multi-source nature of Big Data environments introduce several challenges. Compliance with owners' requirements and legal aspects become then paramount to support Big Data outsourcing, especially in critical security and privacy scenarios. In other words, certified compliance of Big Data analytics must be made available from the outset to all actors that will use Big Data in their business model.

Issues and Challenges: The advent of BDAaaS further strengthen the security and privacy problem. In addition to traditional issues related to protection of data integrity, availability, and confidentiality, new issues emerge related to the impact a completely outsourced infrastructure and deployment plan would have on security and privacy. For instance, BDAaaS gives to an attacker the possibility of implementing inference attacks at a much lower cost.

Possible Solutions: In addition to traditional assurance verification of data security and privacy, in a BDAaaS scenario, it becomes fundamental to guarantee the compliance of the BDAaaS process again existing standards and users' requirements for privacy and security. Also, private data that can be inferred by the process configurations must be protected.

4.3. Configurability and negotiation

4.3.1. Technological variety (tools, products). The rise of Big Data paradigm resulted in the development of several tools and products managing different aspects of the Big Data pipeline. Among these products, we can find NoSQL databases (e.g., MongoDB, Cassandra), parallel processing frameworks (e.g., Spark, Storm, Apache Tez), workflow management and execution framework (e.g., Apache Oozie, Azkaban, Luigi). The variety of Big Data products make it difficult for users to select the approach that best suits their goals. Only Big Data experts have the competences to compare similar products on the basis of the requirements of the analytics they have in mind.

Issues and Challenges: BDAaaS can represents a suitable approach to help Big Data users in finding a clear way through the jungle of Big Data technologies. However, BDAaaS paradigm assumes a detailed knowledge of the characteristics of Big Data tools and products, on one side, and of the users' requirements, on the other side. When this information is less precise, there is a need of compensation techniques driving the selection of the best set of products. Feedback about the performance of previous deployments can play a role in increasing the quality of the selection.

Possible Solutions: BDAaaS requires a detailed definition of tools and products characteristics, with a priori mapping to users' requirements. Proper taxonomies and vocabularies must be defined following technological evolution. A binding between pre-defined Big Data pipelines and scenarios of applicability can also increase the quality of the analytics and of the retrieved results.

4.3.2. Analytics requirements. The selection and configuration of analytics also requests the definition of a complex set of parameters that depend on the data types and analytics requirements. Requirements driving an analytics process can in some cases interfere or even be incompatible. These inconsistencies can be easily solved manually by expert users, which build on their expertise to take over on conflicting requirements.

Issues and Challenges: BDAaaS further exacerbates the issues introduced by incompatibilies in analytics requirements, since they are entirely in the hands of end users and sometimes are defined at an abstract level, increasing the probability of conflicts. In addition, BDAaaS introduces the need of evaluating interferences between requirements that can only materialize at deployment time. For instance, when data are anonymized, we may find out that requirements on data visualization cannot be satisfied anymore.

Possible Solutions: A first approach to address the above issues consists of a conflict management solution based on interferences modeling, which drives users in a consistent selection of requirements. Two types of interference rules can be foreseen: *a priori interferences* modeling aspects that are common to any Big Data analytics, *data-dependent interferences* modeling aspects that are analytics dependent. An additional layer can specify mix of requirements that are discouraged.

4.3.3. SLAs and Assurance. The monitoring and control of a Big Data analytics process is fundamental to increase adoption of Big Data facilities. Definition and verification of SLAs are important to guarantee that Big Data pipeline and its processes are behaving as expected, for instance, exhibiting the required performance, guaranteeing a minimum precision, and preserving privacy and security properties. Assurance techniques support SLA verification by providing "*the way to gain justifiable confidence that infrastructure and/or applications will consistently demonstrate one or more security properties, and operationally behave as expected despite failures and attacks*" [21]. They provide a means to obtain providers' guarantees on the viability of the executed process and to support a posteriori auditing processes. Thanks to SLAs and assurance, organizations wishing to put BDA in their governance and/or command&control chains can compute clear ROIs and risk assessments.

Issues and Challenges: To put in contact organizational needs and technological solutions, BDAaaS needs to provide organizations with measurable factors at two different levels. On one side, features at the execution environment and architecture level must be mapped with specific measurable factors. On the other side, the achievement of the requirements specified by the users must be measurable as well. In this context, assurance techniques measuring the distance between the user's expectations and the provided Big Data pipeline become a pressing need.

Possible Solutions: A first approach to address the above issues consists of the definition of multi-layer assurance agents that collect information out of Big Data pipeline execution, verify the support of users' requirements, and verify the correct behavior of the Big Data architecture. Also, inference and reverse engineering on analytics results can help in further strengthening assurance evaluation.

4.4. Societal and organisational challenges

4.4.1. Competences and learning curve. Big Data offers unprecedented opportunities for organizations. It supports advanced business intelligent applications and allows to extract value out from huge amount of diverse data. The complexity of Big Data and the proliferation of Big Data solutions raise the bar on the need of costly in-house skills, which is hampering its wide adoption. Outsourcing part of Big Data management to the outside is not reducing the need of in-house skills and, most importantly, precludes the adoption of Big Data solutions by SMEs.

Issues and Challenges: BDAaaS introduces a rethinking of existing BDA approaches, requiring users to specify their analytics in a declarative and abstract way, and leaving to the analytics providers the responsibility of deploying the right pipeline and analytics. However, the quality of

BDAaaS strongly relies on the quality and precision of the requirements specified by the users. The more BDAaaS users have a clear understanding of their goals and Big Data technologies, the higher the BDAaaS performance.

Possible Solutions: Model-based, declarative specification of Big Data analytics represents a first possible approach. Users should be supported by a wizard in the selection of declarative requirements, minimizing the risks of conflicting specifications (see Section 4.3.2).

4.4.2. Standardisation of methodological approaches. "*Until very recently, the global IT community has been looking at Big Data in the same way that the six blind men in the fable inspected the elephant. That is, each member of the community considered the subject (Big Data) from only one perspective, at most*" [22]. For these reasons, today, only few standardization initiatives have emerged. Well-defined and internationally recognized standards can potentially reduce the possible controversy due to the different legal and regulatory compliance requirements existing in different countries.

Issues and Challenges: As stated in the Strategic Research and Innovation Agenda (SRIA) published by the European Big Data Value Partnership[1], the lack of standards represents a major barrier to the diffusion of Big Data technologies.

Possible Solutions: Some standardization initiatives can be taken as reference. IEEE Big Data Technical Community held in 2015 the 1st IEEE Big Data Initiative (BDI) Standards Workshop[2]. During the workshop several topics for standardization have been identified, among which *Metadata Standard for Big Data Management* and *Data Representation in Big Data Management* are important for BDAaaS. ITU-T also started an effort towards the definition of Big Data standards[3], and in 2015 approved the first standard on Big Data *Big Data - Cloud computing based requirements and capabilities* [23]. Finally, in 2016, ISO started the effort towards the definition of standard ISO/IEC CD 20546 Information Technology – Big Data – Definition and Vocabulary [24].

4.4.3. Regulatory barrier. Concerns about violating data access, sharing and custody regulations when using BDA, and the high cost of obtaining legal clearance for their specific scenario are discouraging companies, particularly SMEs, from taking over BDA.

Issues and Challenges: As already discussed, data management comes with legal issues that need to be accomplished by a proper approach to Big Data analytics. This is especially true when, in addition to data, also the entire Big Data pipeline is outsourced.

1. http://www.bigdatavalue.eu/
2. http://bigdata.ieee.org/standards
3. http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx

Possible Solutions: A first approach should rethink existing regulations to achieve peculiarities of BDAaaS scenario. Specifically, it should provide a consistent way of managing data and infrastructures across different countries and regulations.

4.5. Big Data modeling

Until now, Big Data application developers have devoted little attention to modeling [10]. Traditional data modeling, which focused on resolving the complexity of relationships among schema-enabled data, has been discarded as no longer applicable to Big Data scenarios. Recent ideas have started from potential Big Data users' expectations and requirements to develop the idea that achieving the full potential of Big Data analytics needs a model comeback.

Issues and Challenges: BDAaaS adds a layer of complexity to traditional BDA that must be managed by proper modeling techniques. Not only data modeling will be key for BDAaaS, but also process execution and architecture deployment will require a model-based approach.

Possible Solutions: A suitable approach to Big Data modeling should provide a model-driven architecture for BDAaaS.

The TOREADOR project[4] is a H2020 project aimed at providing a specification of a fully declarative framework and a model set supporting Big Data Analytics. TOREADOR will enable users to (i) specify their goals at business level (ii) describe and manage data and process diversity (iii) have a single learning curve for diverse analytics- and simulation- driven applications in different domains.

5. Conclusions

In this paper, we defined the Big Data Analytics-as-a-Service (BDAaaS) paradigm as the next evolution step of Big Data domain. We then presented the main issues and challenges introduced by BDAaaS and outlined possible solutions to address them. BDAaaS can represent a suitable driver bringing Big Data to those organizations and SMEs lacking sufficient in-house competences.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the TOREADOR project, grant agreement No 688797. It was also partly supported by the program "piano sostegno alla ricerca 2015' funded by Università degli Studi di Milano.

References

[1] IDC, *Worldwide Big Data and Business Analytics Revenues Forecast to Reach $187 Billion in 2019*, May 2016, https://www.idc.com/getdoc.jsp?containerId=prUS41306516.

4. www.TOREADOR-project.eu

[2] European Commission, *Helping SMEs Fish the Big Data Ocean*, July 2014, http://ec.europa.eu/digital-agenda/en/news/helping-smes-fish-big-data-ocean.

[3] D. Austin, *eDiscovery Trends: CGOCs Information Lifecycle Governance Leader Reference Guide*, May 2012, http://www.ediscoverydaily.com.

[4] C. Ardagna and E. Damiani, "Network and storage latency attacks to online trading protocols in the cloud," in *Proc. of the International Conference on Cloud Computing, Trusted Computing and Secure Virtual Infrastructures*, Amantea, Italy, October 2014.

[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity*, 2011. [Online]. Available: http://tinyurl.com/z9wjhuw

[6] R. K. Lomotey and R. Deters, "Analytics-as-a-service framework for terms association mining in unstructured data," *International Journal of Business Process Integration and Management (IJBPIM)*, vol. 7, no. 1, pp. 49–61, 2014.

[7] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How big data can make big impact: Findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234 – 246, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925527314004253

[8] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communication of the ACM*, vol. 57, no. 7, pp. 86–94, July 2014.

[9] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto, "Big data, bigger dilemmas: A critical review," *Journal of the Association for Information Science and Technology*, vol. 66, no. 8, pp. 1523–1545, 2015.

[10] V. Markl, "Breaking the chains: On declarative data analysis and data independence in the big data era," *Proc. of VLDB Endowment*, vol. 7, no. 13, pp. 1730–1733, August 2014.

[11] D. Abadi, R. Agrawal, A. Ailamaki, M. Balazinska, P. A. Bernstein, M. J. Carey, S. Chaudhuri, J. Dean, A. Doan, M. J. Franklin, J. Gehrke, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. V. Jagadish, D. Kossmann, S. Madden, S. Mehrotra, T. Milo, J. F. Naughton, R. Ramakrishnan, V. Markl, C. Olston, B. C. Ooi, C. Ré, D. Suciu, M. Stonebraker, T. Walter, and J. Widom, "The beckman report on database research," *ACM SIGMOD Record*, vol. 43, no. 3, pp. 61–70, December 2014.

[12] P. Russom, *Big Data Analytics*, TDWI best practices report, TDWI Research, 2014, http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1821.

[13] D. Eckhoff and C. Sommer, "Driving for big data? privacy concerns in vehicular networking," *IEEE Security Privacy*, vol. 12, no. 1, pp. 77–79, January 2014.

[14] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," *IEEE Network*, vol. 28, no. 4, pp. 46–50, July 2014.

[15] D. Wu, M. J. Greer, D. W. Rosen, and D. Schaefer, "Cloud manufacturing: Strategic vision and state-of-the-art," *Journal of Manufacturing Systems*, vol. 32, no. 4, pp. 564–579, 2013.

[16] K. E. Martin, "Ethical issues in the big data industry," *MIS Quarterly Executive*, vol. 14, p. 2, 2015.

[17] K. A. Salleh and L. Janczewski, "Adoption of big data solutions: A study on its security determinants using sec-toe framework," in *CONF-IRM 2016 Proceedings*, 2016. [Online]. Available: http://aisel.aisnet.org/conrm2016/66

[18] N. Rahman, "Factors affecting big data technology adoption," http://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1099, 2016.

[19] IDC, "Six patterns of big data and analytics adoption," 3 2016, http://www.oracle.com/us/technologies/big-data/six-patterns-big-data-infographic-2956541.pdf.

[20] A. Azzini and P. Ceravolo, "Consistent process mining over big data triple stores," in *2013 IEEE International Congress on Big Data*. IEEE, 2013, pp. 54–61.

[21] C. Ardagna, R. Asal, E. Damiani, and Q. Vu, "From security to assurance in the cloud: A survey," *ACM Computing Surveys (CSUR)*, vol. 48, no. 1, pp. 2:1–2:50, August 2015.

[22] E. Gasiorowski-Denis, *Big plans for big data*, March 2014, http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1821.

[23] IUT-T, *Big data – Cloud computing based requirements and capabilities*, November 2015, http://www.itu.int/rec/T-REC-Y.3600-201511-I.

[24] ISO/IEC, *ISO/IEC CD 20546: Information Technology – Big Data – Definition and Vocabulary*, 2016, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=68305.