HAL
open science

Privacy-Aware personal Information Discovery model based on the cloud

Thiago Moreira da Coasta, Hervé Martin, Nazim Agoulmine

To cite this version:

Thiago Moreira da Coasta, Hervé Martin, Nazim Agoulmine. Privacy-Aware personal Information Discovery model based on the cloud. 8th Latin American Network Operations and Management Symposium (LANOMS 2015), Oct 2015, Joao Pessoa, Brazil. pp.35–40, 10.1109/LANOMS.2015.7332667 . hal-01260103

HAL Id: hal-01260103

https://hal.science/hal-01260103v1

Submitted on 25 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Privacy-Aware Personal Information Discovery Model based on the cloud

Thiago Moreira da Costa
and Hervé Martin
Laboratoire d'Informatique de Grenoble
Université de Grenoble
Saint Martin d'Heres
{thiago.moreira, herve}@imag.fr

Nazim Agoulmine
IBISC
Université d'Evry
Evry
nazim.agoulmine@ibisc.fr

***Abstract*—Data collection, storage and manipulation have become more critical due to the growth of magnitude of their misuse or mismanagement impact in business and political scenarios nowadays. While research has pushed technology to deliver more powerful information discovery algorithms, and responsive on-demanding storage and processing capacity through data analysis and distributed cloud infrastructure, concerns about privacy have globally raised several discussions involving different sectors of the society. In particular, individual rights are highly impacted by privacy issues due to nowadays geographic distribution of sensitive information and its discovery.**

In this work, we present a model for privacy awareness during the data analytics process in a context of scalable computing using the cloud. Our approach addresses privacy issues both in data analytics process and in the infrastructure resource allocation according to privacy regulation in Service Level Agreements (SLA). The proposed model for Privacy-Aware Information Discovery (PAID-M) provides privacy awareness by executing data analytics algorithms encapsulated with privacy preserving techniques. The model also presents how it intends to address the privacy issue in the cloud deployment process by considering differences in privacy regulations and jurisdictions.

I. INTRODUCTION

The pace on which information has modeled the world where we live has amazingly increased. Indeed, technology has changed the way organizations and people produce and treat data, making possible to produce and share information unprecedentedly faster. Furthermore, Data Science has allowed us to consume information quicker and more efficiently by aggregating and selecting relevant information from large amount of data using data mining techniques. This myriad of aspects of data, and its production, analysis and dissemination techniques have also led to the change of how individual's privacy has been compromised.

A recent report of the Execute Office of the President (USA)[1] about big data and privacy highlighted several problems about personal privacy, the society is facing now or in a near future: intrusion upon seclusion, public disclosure of private facts, disclosure of inferred private facts (sometime false positives), defamation using inferred private facts, invasion of unencrypted private communication, invasion in personal virtual space, stalking and violation of locational privacy, foreclosure of self-determination, lost of autonomy, and others.

While some of these issues are related to information leak, and consequently involve broader governance and systematic to minimize the collection and analyze of data by third-parties, some of privacy threats ramifications are related to new big data technologies that are able to extract valuable information from large volumes[2]. In fact, the continuous ratio growth of data volume, velocity, variety and veracity in the big data scenario is its main challenged and distributed computing are one of the strategy to overcome these technical constraints [3].

The modern technology used to mining semantic enriched data and the computational power provided by nowadays distributed and cloud computing to process huge data sets are permitting fast investigation/correlation of relevant information (decreasing time to discovery private personal facts) and high significance level of inferred information (minimizing the ratio of false positives). Big data storage and analytics technology made it possible to store data produced from web and social interaction, machine logs, sensing, transactions and Internet of Things[4], not limiting in some selective relevant information only. The services provided by modern applications in portable devices that are context-sensitive, social and location-based depend on this data to deliver smart feature and customize users' experience, behaving accordingly to the individual's context and situation. However, the indiscriminate storage of personal digital data has led to a practice of *life logging* [5] which has the potential to be intentionally exploited to extract private personal information that was not initially intended by individuals who use these services.

Furthermore, big data technology is increasing the complexity in privacy policy implementation, since it differs from the traditional privacy management where private information was meaningful for the individual who owns it, and it could be classified as sensitive or not, such as race, health status records and salary. Big data technology can infer information from raw data that are not clear for those who produce it, for instance, work place location or risk of loan eligibility.

Besides that, algorithms to prepare data, reduce sensor noises, remove outliers, compress data, integrate heteroge-

neous data are continuously proposed, leveraging the quality of data stored. In the field of semantic trajectory analysis, for instance, algorithms for outlier removal, kernel smoothing, and compression prepare spatiotemporal data (GPS points) for posterior processing[6]. Other techniques for normalization, integration and filtering may be applied to prepare data for analytics algorithms as well. Furthermore, the best practices for publishing and connecting structured data on the Web, called Linked Data[7], are changing the way data is stored, retrieved, and integrated by using semantic web technology and adding contextual information to them. This allows data analytics algorithm to take into account several aspects of context on which these data were produced, leveraging the level of significance of information extracted from personal data of individuals[8]. Recent cases in Facebook perceiving individuals' moods have disclosure the level of detail currently achieved by big data technology in social networks using history of annotated digital traces [9][10]. This threat to personal privacy become more complex when added to the fact that distributed infrastructure, such as the cloud, is used to achieve this analytics results.

Cloud Computing has a particular role in this scenario specially because its distributed infrastructure bring complex privacy issues [11], [12]. Hashem et al. [4] describe several security and privacy concerns in the cloud that needs to be accounted according to the recent privacy regulations, such as encryption, privacy-safe query, data protection architecture, social network sensitive information publishing, and statistical privacy attack. However, it is in privacy jurisdictions that most regulation decision have focused, pushing personal privacy relevance to a critical level, similar to what was formerly found in business multi-tenant requirements for the cloud [13].

The same characteristics that are required by multi-tenant environments, such as privacy governance and accountability, compliance to regulations, jurisdictional clearance, and service-level agreements also apply to software and services that storage and process personal data of individuals. Therefore, jurisdiction remains a problem to be solved in cloud computing. Restrictions for cross-border data flow for storage or processing may applied in order to guarantee that personal privacy conditions are respected despite of the service provider's location, as currently states in the Russian Statute on *Roskomnadzor* (Russian Federal Service for Supervision of Communications, Information Technology, and Mass Media) and in the European Data Protection Regulation, for instance.

Nonetheless, cloud's technology competitiveness relies heavily on its delivery and deployment models [11] that allow cost reduction and efficient resource allocation due to the cloud elasticity across multiple IaaS providers. How cloud's server vendors can provide such elasticity and still be compliant to multi-jurisdiction privacy regulations?

In this context, three aspects of privacy must be addressed to propose a privacy-safe environment for personal information discovery: i) policies that take into account the direct or indirect relation between raw data and information semantically relevant; ii) privacy-aware programming platform that provides *composable* and extensible data mining modules; iii) privacy sensitive distributed platform that provide elastic storage and processing capacity.

For this propose, we introduce in this work a Privacy-Aware Information Discovery Model (PAID-M) to deploy a reliable, scalable and privacy sensitive data mining system. The main contributions of our work are:

- A privacy policy management that is able to preserve data that can be used to infer sensitive information from being published. By applying privacy policy based on taxonomies, this approach uses semantic reasoning to find direct and indirect relations between privacy policy data.
- *Privacy encapsulation* of traditional data mining algorithms by modularizing them while verifying their inputs and outputs. We define solution for reuse traditional data mining algorithms by supplying a preparation phase and a post-processing phase to verify privacy. Furthermore, a concept of data mining module allows combining an array of data mining algorithms, providing a solution to build information discovery requests.
- Scalable processing and storage using the cloud and interacting with MOSt, a Multisite Orchestration System[14], in order to provide privacy sensitive cloud elasticity.

The rest of this paper is organized as follows: Section 2 describe the privacy-aware information discovery model, describing step by step the rational used to address the privacy problem in the two levels of the solution (programming and infrastructure abstractions). In Section 3, the global architecture of the solution is explained, each components and the technology that is expected to be implemented. Finally, conclusions and next steps of the work are presented.

II. PAID-M - Privacy-Aware Information Discovery Model

[15] classifies technology required for Data Analytic in two categories: *Programming Abstraction* and *Infrastructure Abstraction*. The *Programming Abstraction* aggregates those algorithms that make sense of data, extract non obvious patterns, and predict future trends and behaviors. The *Infrastructure Abstraction* is the technology responsible to provide a scalable, fault-tolerant, and safe environment for data analytics algorithms. In this context, we propose a Privacy-Aware Information Discovery Model as a platform-independent model to address the problem of privacy of personal data in a scenario of big data information discovery focusing in the programming and infrastructure abstractions. Our model address privacy in the level of *Programming Abstraction* by supporting the data analyst through the process of building an analytic process in order to be able to intermediate the execution of analytic processes and verify privacy according to the individual's policy. For this matter, it is necessary to cover what are the types of data analytics algorithms and understand how these algorithms can be concatenated. Furthermore, for the *Infrastructure Abstraction* our model propose an infrastructure that provide parallelized, scalable, fault-tolerance and safe using the cloud. In this level of infrastructure, privacy is

implemented by intervening deployment plans and resource allocation in a way that SLAs for privacy and regulation are accounted. Different components in our model can be deployed according to the individual's privacy policy and SLA. In the next subsections, different aspects of our models are discussed.

A. Programming Abstraction

The *Programming Abstraction* is part of the technology intended to provide an accessible, small and composable interface to reuse techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions, as described in Big Data Analytics[16]. Several works have studies the common characteristics of analytics algorithms and the Data Analytic Workflow (DAW)[17], [18], [19], [20], [21], [1], [22], [23], [16], [24]. The definition of a high-level class of data analytics algorithms on top of big data allows algorithms reuse, modularization and possibility to express different kinds of relationships among classes. As an example, the work described in [24] proposes mega-modules as super classes of analytics algorithms capable of compose any DAW using these modules. Figure 1 depicts the DAW overview for a single step of data analysis and what is expected in each step. This strategy for encapsulating the analytic steps in modules containing preparation, analytics and result phases has been successfully implemented in several projects, such as GeoDeepDive[22], M-Atlas [25], and Apache Flink (former Stratosphere[18]).

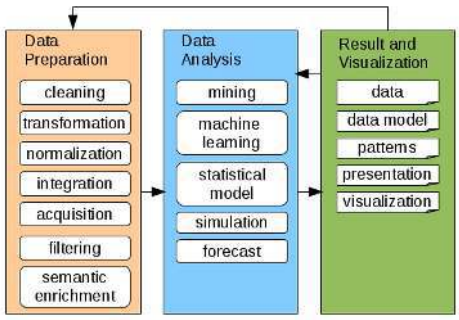

Fig. 1. Data Analytic Workflow Overview

Although some of these projects has partially covered privacy concerns, *privacy enforcement* is not formerly considered into their DAW. As an example, M-Atlas, which works with spatiotemporal data, provides a high-level query language called DMQL [26] that can be used to express DAW's. Nonetheless, DMQL lacks expressiveness to incorporate other privacy preserving techniques than *k*-anonymization[27]. Besides that, the privacy preserving techniques is only used when expressed by the data analyst. Another issue concerning the privacy in the process of data analysis is interpreting its result. Sometime the results are not intelligible or not conclusive. In a traditional Data Analytic scenario, the data analyst concludes if the result is meaningful or if it needs to be re-executed, discarded, validated or adjusted. The output of data analytics algorithms are mainly data, data models and patterns[20]. Data and patterns are mostly result of data mining techniques, while data models can be output from machine learning, statistical model algorithms, simulation and forecast. For the matter of understandability of the results, visualization and presentation can be necessary as well.

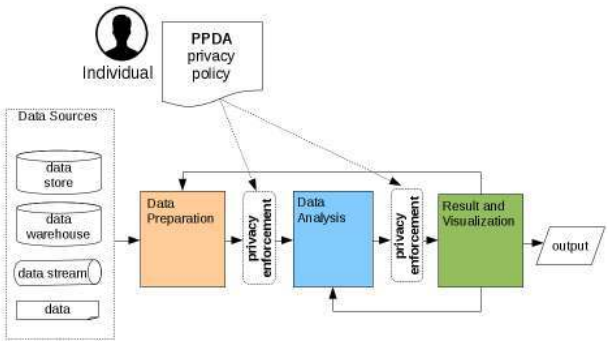

Fig. 2. Privacy Enforcement in Data Analytic Workflow

In PAID-M, we propose enforce privacy in the DAW by incorporating privacy preserving techniques according to the personal privacy policy, as seen in Figure 2.

Among the myriad of existing techniques to be applied to the data during the preparation and post-analysis steps, data characteristics and data types are properly accounted by the data scientist during the Data Analytic Process (DAP[1]) definition. It is no different when considering the privacy preserving techniques. According to [28], privacy-preserving data mining and algorithms can be classified in randomization methods, group-based anonymization, distributed privacy-preserving data mining, privacy-preservation of application results. Some examples of such techniques are as follows:

- Randomization method by adding noise in order to mask attribute values;
- *k*-anonymity and *l*-diversity methods reduce the possibility of indirect identification of individual who produced by generalizing (reducing granularity of the data representation) and suppression;
- Horizontal and vertical partitioning of data sets cross multiple sites in the data mining process, sharing partitions in a way that minimize the possibility of an individual entity;
- Downgrading of application effectiveness, such as rule mining hiding, classifier downgrade, and query auditing to minimize privacy violation by reducing algorithm effectiveness.

Both data analytics algorithms and privacy preserving techniques have a semantic in the DAP that must be observed. That means these techniques are context-sensitive and its application varies according to the data (and its characteristics)

[1]Data Analytic Process is an instance of data analysis following the Data Analytic Workflow

to be processed and to the result analysis. For the PAID-M, this semantic and meta-data (of techniques and data), must be mapped in order to provide adequate techniques during the process of composition of the DAP. Furthermore, this semantic annotation and meta-data can trigger the automatic privacy enforcement by using semantic web technology for reasoning which privacy preserving techniques could be used according to each step of the DAP.

Concerning the privacy policy, the solution proposed in our model is composed by two parts which are implemented in different abstractions of the process. In the *Programming Abstraction*, the Privacy Policy for Data Analytic (PPDA), as depicted in Figure 2, describes the individual's intention for data use, indicating in terms of what aspects of her life should be considered private. In the execution of a DAP, for each step, semantic annotation of data analytics algorithm will guide the system to apply privacy preserving techniques according to PPDA. By doing so, the strategy of PAID-M for privacy preservation goes beyond the all-or-nothing paradigm of policy data access control, taking advantage of privacy preserving techniques to prepare data and to verify the result of data analytics algorithms.

B. Infrastructure Abstration

The *Infrastructure Abstraction* is the technology that provides the solution for the big data challenges using cloud computing. Cloud computing provide many features such as parallelization and distributed computing that are extremely important to overcome big data barriers. Although, while some of the concepts behind cloud computing are not new, from the point of view of scalability, virtualization technology aggregate the most convenient on-demand capacity for data analytic processes. The elasticity property of the cloud allows a very fast increment or decrease of resources, being transparent for the Virtual Machines (VMs). This is done by the virtualization technology (hypervisor) that intermediates physical machines and operating systems (OS), as shown in Figure 3, in order to manage and allocate resources.

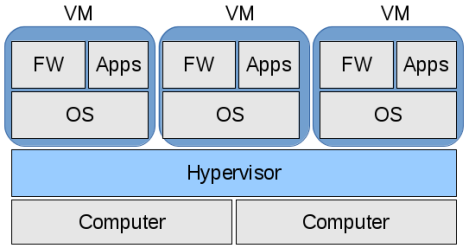

Fig. 3. Example of cloud virtualization

In this context, infrastructure are offered as a service (IaaS[2]), where cloud providers can provide a physical infrastructure (mainframe, servers, network, storage). In order to make this physical infrastructure available, cloud providers must use a IaaS cloud platform to manage the physical resources, such as OpenStack[3] or Amazon Web Service[4].

In a market that starts to become more heterogeneous, with the advantage of private clouds and the increase of cloud vendors, choosing which IaaS to contract is not an easy task. SLA4Cloud[14] is a project that aims to dynamically provide optimal and composition placement of virtual machines delivering better network capabilities and performance trades according to the specified SLA. The Multisite Orchestration System (MOSt) intermediate the process of deployment serving as a hub for several IaaS cloud providers. However, some SLA requirements are still difficult to achieve, due in part to the lack of mechanisms and tools to ascertain if the cloud vendor maintains compliance with the contracted SLA, in part because there are still different SLA requirements that is not easily implemented and monitored due its multidisciplinary characteristics. Privacy can be considered as such a challenge [11], [12] that involves policy, regulation and technology[1]. Its implementation is not trivial, in view of the nowadays constant changing in regulation in the world. This question issues the cloud computing directly, because the cloud elasticity is directly affected and restricted by regulation of personal privacy in some countries that restricts cross-border data flow, for instance. In our model, we propose to cover privacy regulations in the *Infrastructure Abstraction* by extending MOSt[14] with the capability to identify which IaaS cloud vendors are not eligible due privacy regulation restrictions. The PPDP will contain SLA that will specify privacy requirements, such as jurisdiction restrictions and other constraints related to Privacy Regulation. Figure 4 present the interaction between the system and the cloud.

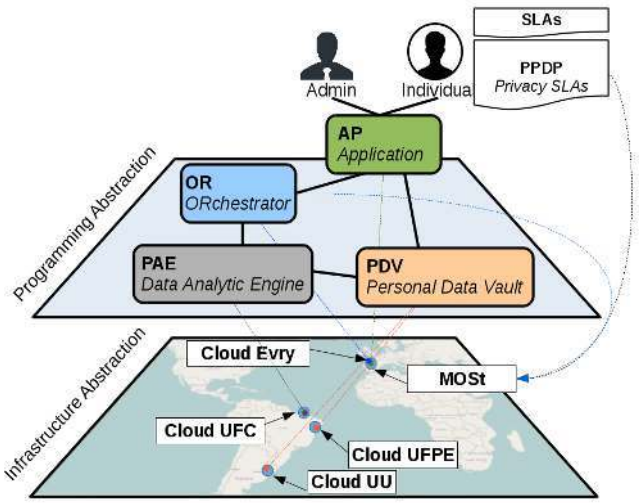

Fig. 4. Example of privacy enforcement in the cloud

[2]Infrastructure as a Service

[3]http://www.openstack.org/

[4]http://aws.amazon.com/

III. Global Architecture

The system to be implemented using the PAID-M will be based in open-source technology and the infrastructure of cloud federation currently maintained by universities in Brazil, France and Uruguay. The *Infrastructure Abstraction* implementation of this work intend to use the infrastructure set up for the SLA4Cloud project, using the IaaS provided by the universities cited above and noted in Figure 4.

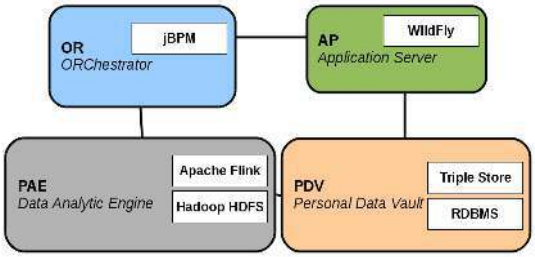

Fig. 5. Architecture Overview

For the *Programming Abstraction* implementation, the four main components, presented in Figure 5, are the followings:

- The ORchestrator (OR) component is the responsible for starting and stopping the other three components. It is the first to be deployed and also controls the queue for data analysis. The Orchestrator is intended to be implemented using Wildfly[5] and jBPM[6] technology;
- The Application Server will be responsible for interacting with administrator, data producers and data analysts. Interfaces to upload data, submitting DAP and visualizing results is expected to be implemented in AP using WildFly;
- The Personal Data Vault (PDV) will be in charge of storage of personal data. For each data producer, one account will be created in order to protect and isolate private data. Two type of storage will be available to support different type of data, a Triple Store and a Relational Database Management System (RDBMS);
- The Data Analytic Engine (DAE) is responsible for processing the DAPs. In order to do that, the Distributed FileSystem HDFS[7] should be implemented to provide cache and support for the Apache Flink.

There are several open-source projects for large data sets processing, such as Apache Spark, Apache Storm, Apache Flink, Apache Tez, and the traditional Apache Hadoop-MapReduce [8] and all variation based on market's implementation and requirements. Apache Hadoop 2 and its YARN (Yet Another Resource Negotiator - Resource Manager) predominates in the large data set processing engines currently available. For a matter of succinctness and objectiveness, the parameters used to compare the available engines are not going to be presented. Although, it is important to highlight what makes Apache Flink[9] the best candidate for the DAE.

Formerly known as Stratosphere, Apache Flink is a project sponsored by public and private sectors in Europe Union to empower data scientists to conduct complex data analysis while providing parallelization, query optimization, broad infrastructure and source connectivity, flexible composition of analytics process, automatic optimization in different stages of the process, high performance run-time, rich programming languages to interact along the stages of the analytics process [29]. Besides that, Apache Flink is a data analytic platform that enables the extraction, analysis, and integration of heterogeneous data sets, being capable of performing information extraction and integration, traditional data warehousing analysis, model training (and machine learning), and graph analysis using a programming model based on second order functions[18]. The platform can run standalone, natively in computer clusters, or Hadoop clusters via YARN(see Figure 6).

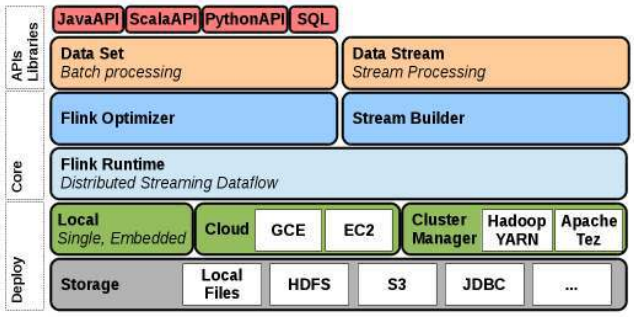

Fig. 6. The Flink Stack

IV. Conclusions

The model presented in this work aims to provide a systematic that will support data consumer to control privacy in the era of big data. Two challenges related to privacy are faced by the Data Analytic in large data sets. The first related to the extraction of private information from data using data analytics algorithms (not clear to the individual who produced). The second related to the cloud computing elasticity, that poses a problem of privacy jurisdiction and regulations. In the next phase, we aspire to implement the PAID-M and run case studies using spatiotemporal data sets, evaluating if individual's privacy policy were satisfied.

Acknowledgment

This research is partially funded by the Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education (CAPES/MEC) and the EU EASI-CLOUDS project (ITEA 2 #10014) and STIC-AmSud SLA4CLOUD project(14 STIC #11).

[5]http://wildfly.org/
[6]http://www.jbpm.org/
[7]http://hadoop.apache.org/hdfs/
[8]https://hadoop.apache.org/
[9]http://flink.apache.org/

Thanks to all the partners of the project who have helped with their discussions to improve the research work presented in this paper.

References

[1] P. C. o. A. o. S. PCAST and Technology, "Big data and privacy: a technological perspective," Tech. Rep. May, 2014.

[2] A. Cavoukian and J. Jonas, *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada, 2012.

[3] A. Jacobs, "The Pathologies of Big Data," *ACM Queue*, vol. 7, no. 6, p. 10, 2009.

[4] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of Big Data on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306437914001288

[5] K. O'Hara, M. Tuffield, and N. Shadbolt, "Lifelogging: Privacy and Empowerment with Memories for Life," *Identity in the Information Society*, no. 2008, pp. 155–172, 2009. [Online]. Available: http://eprints.soton.ac.uk/267123/

[6] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "Semantic Trajectories: Mobility Data Computation and Annotation," vol. 9, no. 4, 2012.

[7] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009. [Online]. Available: http://eprints.soton.ac.uk/271285/

[8] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The meaningful use of big data: Four perspectives - Four challenges," *SIGMOD Record*, vol. 40, no. 4, pp. 56–60, 2011. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84857148573{\&}partnerID=40{\&}md5=4662920b41e2753d23919fb73ee5c6dc

[9] E. Bakshy, S. Messing, and L. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science Press*, 2015.

[10] E. C. Tandoc, P. Ferrucci, and M. Duffy, "Facebook use, envy, and depression among college students: Is facebooking depressing?" *Computers in Human Behavior*, vol. 43, pp. 139–146, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0747563214005767

[11] H. Takabi, J. B. Joshi, and G.-J. Ahn, "Security and Privacy Challenges in Cloud Computing Environments Cloud," *Cloud Computing*, no. December, 2010.

[12] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, and A. V. Vasilakos, "Security and privacy for storage and computation in cloud computing," *Information Sciences*, vol. 258, pp. 371–386, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2013.04.028

[13] European Network and Information Security Agency (ENISA), "Cloud Computing: Benefits, Risks and Recommendations for Information Security," Tech. Rep., 2009. [Online]. Available: www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment/at{_}download/fullReport

[14] E. H. Cherkaoui, E. Rachkidy, M. Santos, P. A. L. Rego, J. Baliosian, and J. N. De, "SLA4CLOUD : Measurement and SLA Management of Heterogeneous Cloud Infrastructures Testbeds," *3rd International Workshop on Advances in ICT*, pp. 1–6, 2014.

[15] A. Kumar, F. Niu, and C. Ré, "Hazy: Making it easier to build and maintain big-data analytics," *Communications of the ACM*, vol. 56, no. 3, pp. 40–49, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2428570

[16] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. a.S. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing*, 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0743731514001452

[17] R. L. Grossman, "What is analytic infrastructure and why should you care?" *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 5, 2009.

[18] A. Alexander, B. Rico, E. Stephan, F. Johann-Christoph, H. Fabian, H. Arvid, K. Odej, L. Marcus, L. Ulf, M. Volker, N. Felix, P. Mathias, R. Astrid, J. S. Matthias, S. Sebastian, H. Mareike, T. Kostas, and W. Daniel, "The Stratosphere platform for big data analytics," *VLDB*, 2014.

[19] R. Bordawekar, B. Blainey, and C. Apte, "Analyzing analytics," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 17–28, feb 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2590989.2590993

[20] S. Ceri, E. D. Valle, D. Pedreschi, R. Trasarti, and L. B. Pontecorvo, "Mega-modeling for Big Data Analytics," pp. 1–15, 2012.

[21] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[22] C. Zhang and C. Ré, "GeoDeepDive : Statistical Inference using Familiar Data-Processing Languages," pp. 993–996, 2013.

[23] P. Atzeni, D. Cheung, and S. Ram, "Conceptual Modeling," in *ER*, 2012. [Online]. Available: http://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:No+Title{\#}0

[24] S. Ceri, T. Palpanas, and E. Valle, "Towards mega-modeling: a walk through data analysis experiences," *SIGMOD*, vol. 42, no. 3, pp. 19–27, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2536673

[25] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *VLDB*, vol. 20, no. 5, pp. 695–719, jul 2011. [Online]. Available: http://link.springer.com/10.1007/s00778-011-0244-8

[26] R. Trasarti, S. Rinzivillo, M. Nanni, and F. Giannotti, "Types and Operators in M-Atlas system," no. i, pp. 1–12, 2011.

[27] R. J. Bayardo and R. Agrawal, "Data Privacy Through Optimal k-Anonymization," *ICDE*, no. Icde, 2005.

[28] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," 2008.

[29] V. Markl, "Breaking the Chains: On Declarative Data Analysis and Data Independence in the Big Data Era," *Proceedings of the VLDB Endowment*, pp. 1730–1733, 2014. [Online]. Available: http://www.vldb.org/pvldb/vol7/p1730-markl.pdf