Bootstrapping GDPR: Technical infrastructure requirements and architectures to implement GDPR

Yuri Demchenko (y.demchenko@uva.nl), Fatih Turkmen, Cees de Laat
University of Amsterdam

Abstract

The new European GDPR introduce many technically challenging requirements to data processing applications and platforms that are heavily based on the Big Data Infrastructure which is commonly cloud based and often tightly integrated with the cloud infrastructure itself. There is not many research on addressing technical aspects of the GDPR for distributed services and applications processing large amount of heterogeneous data typical for modern business applications. New technologies to address required by the GDPR "Privacy by Design" (PbD) principle yet need to be proposed and research and technology community still to come with the innovative solutions that would protect data privacy while not inhibiting technology development.

1. Introduction

The EU General Data Protection Regulation (GDPR) [1] is expected to imply many changes into the practice of organisations dealing with personal data, which may virally spread over most of data processed by the organisations if they use, for example, social network data or data collected on user activities over the services provided by the organisations. Majority of discussions so far related to GDPR implementation were focused on legal and regulatory aspects with a particular focus on the preparation of the organisations for the official GDPR launch from 25 May 2018. This included not only policy and procedures development, and relevant services deployment into organisational procedures but also transforming all former/historical data stored and maintained by the organisation that are also required to be compliant with new requirements.

There is not many research on addressing technical aspects of the GDPR for distributed services and applications processing large amount of heterogeneous data typical for modern business applications. Although the GDPR document require implementing "Privacy by Design" (PbD) principle, there is no new technologies proposed in the context of PbD. Research and technology community still to come with the innovative solutions that would protect data privacy while not inhibiting technology development. Recent case with the Facebook personal data misuse by Cambridge Analytica [2, 3] is a signature case illustrating the past and current controversy between requirements of the personal data protection and a way how modern data driven industry operates and offers services.

In this paper we will analyse technical aspects of the GDPR implementation and attempt to formulate technical requirements to GDPR aware services design and operation for modern data driven applications and services. We will also identify critical research areas that would allow developing GDPR aware general data protection models for data intensive applications, big data analytics and allow emerging data market. The proposed research is based on the long time research by authors in the area of Big Data and cloud security and compliance [4, 5] for emerging data intensive and data driven technologies.

2. Agile Data Driven Enterprise (ADDE) and GDPR

Companies are increasingly adopting new Agile Data Driven Enterprise (ADDE) business and operational model that incorporates company wide data management and governance that is documented in the DAMA Body of Knowledge (DM-BOK) [6] and Data Management Maturity (DMM) model by CMMI (Capability Maturity Model Integration) [7].

Effective and market oriented GDPR implementation and services (re-)design must be based on best privacy practices in data protection and privacy enhancement that are available in many industries that are working with personal data and making decisions based on data that affect services provided or delivered to individuals. This is especially related to modern Data Science and Analytics (DSA) services and methods [8].

The following GDPR requirements will affect and need to be addressed when designing and operating DSA based services, including corresponding testing and compliance verification procedures to be developed:

- data processing and profiling;
- right to explanation;
- right to be forgotten;
- bias and discrimination.

Key question for companies is how to design their Big Data Infrastructure (BDI) and corresponding data management and governance model and DSA services to comply with the GDPR requirements and create competitive advantage [9]. Companies must deal not only with the PbD as a technical aspect/dimension but also with other dimensions such as the general company’s code of operation and culture that includes additional dimensions such as organisation and governance, business functions and processes, personnel training and skills.

3. Role of Data Infrastructure and challenges

Essential role in implementing GDPR and PbD principles will belong to data infrastructure that is increasingly becoming distributed and extensively using cloud technologies and cloud resources offered by public Cloud Service Providers (CSP) in a typical hybrid cloud deployment scenario. The following challenging aspects need to be addressed in the GDPR aware data infrastructure design and operation:

- Requirements and compliance of the intended infrastructure, platforms and applications
- Manageable security services, including data protection (confidence, integrity, non-repudiation) and access control
- Fine-grained activities and events logging supported by effective data mining services supporting activity reporting
- New security models to allow using data as digital goods (consumable, transferrable/owned, usable, recyclable)

- Well-defined data governance policy and data management plan (supported corresponding tools and machine readable templates)
- Data provenance (lineage) services and mechanisms that allow the whole data processing lifecycle/activity documenting

The GDPR mandated “right to explanation” and none-discrimination are regarded as specifically challenging to modern DSA algorithms that are statistically based. Although GDPR explicitly prohibits using personal characteristics (age, ethnicity, race, and others) for automated decision making, a number of proxy characteristics may still have effect. Additional research and study of best practices in such industries and banking on credit risk assessment are required.

Two other aspects of the modern data infrastructure and data processing environment will have influence and need to addressed are

(a) distributed nature of data processing when data may move between systems and data processors/controllers,
(b) growing use of cloud-based data storage and processing environment.

4. Big Data Infrastructure compliance

Most of BDI platforms and tools are cloud based and are tightly integrated with the hosting cloud platform or provider. We will refer to the modern cloud based infrastructure as CBDI [10]. There is a growing acceptance between security experts that clouds provide much more advanced security services and compliance support than the ones provided by individual companies, in particular SMEs. Although leading CSPs have already declared the compliance of their infrastructure with GDPR and other general compliance standards, the responsibility for data security and compliance will remain with the customers. Actual data processing services in cloud need to ensure trusted, secure and consistent integration with the related cloud services and benefit from the advanced cloud security infrastructure, including security data storage, backup, transfer and deep logging and monitoring services.

4.1. Cloud and Big Data Infrastructure compliance

Compliance and security are related and in some cases interchangeable terms and concepts. Security is commonly defined as a set of technical, physical, and administrative controls in order to ensure normal operation of a system or application. Compliance is a certification or confirmation that the system or an organization meets the requirements of the specified standards, established legislation, regulatory guidelines or industry best practices that can be jointly defined as compliance framework.

Why compliance is important for cloud? When moving to cloud, the organization moves from internal security and operational environment (that may not be formally defined) to external operational security that will become a part of SLA (or business requirement) with CSP. Compliance in this case will define the expected level of security and assurance. When developing cloud based applications, the applications developer must analyse and ensure compliance of the end user applications with the industry related compliance requirements.

It is a common practice in cloud security that Cloud Service Provider (CSP) implements Shared Responsibility Model that splits responsibility for the security of different layers and components between CSP and a customer that can be cloud based application developer, or end user, or both.

As an example, Amazon Web Services (AWS) as an IaaS cloud provider ensures the security of the cloud infrastructure and cloud platform services: facilities, physical security of datacenter, network infrastructure, virtualisation platform and infrastructure. While the customer is responsible for security of the following components: Amazon Machine Instances (AMI), OS, and applications, data in transit, data at rest, and data stores, credentials, policies and configurations. The customer is specifically responsible to comply with the Acceptable Use Policy (AUP), ensure correct use of the cloud platform, and for security update and patching of the guest OS and installed applications.

Data security and protection in CBDI are also shared responsibilities that involve:

(1) Cloud provider responsibility to ensure secure data storage, processing and transfer and well as provide necessary security mechanisms to enable application level security;

(2) Application developer responsibility to correctly implement the application security in the cloud multi-tenant virtualised environment (often referred to as Security Development Lifecycle and defined by a number of industry standards and guidelines) to protect user data and personal information, integrate applications security with the provided cloud platform security services and mechanisms, and provide necessary and easy usable security services for end user to correctly use application security;

(3) End user responsibility to ensure security of their application access client (typically browser with hosting OS), access credentials and data.

4.2. Compliance standards

Cloud compliance is generally defined by the Cloud Security Alliance Guidance for Critical Area of Focus in Cloud Computing (CSA3.0) [11] that define 13 domains of the security concerns for Cloud Computing that are divided into two broad categories that define corresponding security controls for cloud governance and operation. The CSA GRC Stack (Governance, Risk Management and Compliance) [12] includes the Cloud Controls Matrix (CCM) and Consensus Assessments Initiative Questionnaire (CAIQ) [13] and other documents. CAIQ provides comprehensive tool that maps general IT and data protection controls and different industry specific requirements to CCM. In particular CAIQ includes mentioned above ISO/IEC 27001:2005, PCI DSS, SOC1-SOC3, FISMA, FedRAMP, EU GDPR, HIPAA/HITECH and in total up to 32 different documents.

The cloud providers operating globally need to comply with the different regulations in different countries. In particular, this is important for European Union that has a strict data protection regulation. The new European GDPR will require many data handling processes to be re-designed [14]. The GDPR will be applied to all businesses and companies operated in the European Union, and would prohibit the transfer of personal data to non-European Union countries that do not meet new EU regulation. In order to bridge these differences in approach and provide a streamlined means for U.S. organizations to comply with the new Regulation, the U.S. Department of Commerce in cooperation with the European Commission developed a so-called “Privacy Shield” that comes in place of the former "Safe Harbor" framework [15].

4.3. Compliance analysis and implementation

The compliance of the cloud platform and applications is an important part of setting up and operating cloud based services. It can be also a part of the automated SLA negotiation and monitoring, in particular this functionality should be supported by the Cloud Service Broker. The following sequence can be used for cloud compliance analysis and implementation:

1) Define applications in cloud
2) Identify what data will be moved to the cloud
   - for security and compliance reasons, organisations may decide that some highly confidential data will always remain on an internal network (or private cloud) and will not move to the public cloud
3) For the data moved to cloud, negotiate with the provider about
   - What type of data will reside on the consumer's own/VPC cloud
   - Back up services
   - Possibility to audit
   - Incident report about data incidents
4) Check what compliance documents or industry best practices are used by CSP (see reference list mentioned above)
5) Check what eDiscovery services and tools are available from the cloud provider and develop incident response plan.
6) Define responsibility of all roles involved into data management and have corresponding contacts on the cloud provider side

The cloud based application developer must consider all aspects of the security compliance to ensure that the final application provides consistent security including cloud platform security and application security from the point of view of the application end user. In particular, the ones related to access control, user identity management and user data protection.

5. General Requirements and Design Principles for CBDI Security Services

Referring to earlier authors' work [5], the following general requirements and design principles to CBDI security services can be specified that incorporate and extend current best practices in cloud security [14, 15] (note, enumeration ICSF# is used from the original paper what means InterCloud Security Framework).

ICSF01. CBDI security infrastructure should provide consistent access control, security credentials and security context management for multi-cloud applications deployment, operation and management, in general covering all application lifecycle, including applications secure session management.
ICSF02. CBDI security services should allow users and applications (internally and on behalf of users) to access all distributed multi-cloud resources using single credentials that should be federated with the individual cloud credentials and access control mechanisms.
ICSF03. CBDI security infrastructure should support federated access control and resource management model, allowing integration with the cloud federation services.
ICSF04. Application based access control must be integrated with the cloud based security services and implement in a consistent way the shared security responsibility model that is defined and implemented by cloud services providers as a standard cloud services security model.

ICSF05. CBDI security infrastructure must ensure data protection during the whole data handling lifecycle, including data transfer between different clouds and security domains as well as data storage in-rest.
ICSF06. CBDI security infrastructure should provide secure trust bootstrapping for the provisioned on-demand cloud based security services that should bind the deployed security services to the applications runtime environment and virtualisation platform, to prevent unauthorised virtual environment cloning.
ICSF07. Security Services Lifecycle Management functionality must support the security context management during the whole security services lifecycle, including binding security context to the provisioning session and virtualisation platform.
ICSF08. Security session synchronization mechanisms should implemented to protect the integrity of the remote run-time environment, including secure session fail-over that should rely on the session synchronization mechanism when restoring the session.
ICSF09. CBDI security infrastructure should support Dynamic Security Associations (DSA) to provide fully verifiable chain of trust from the user client/platform to the virtual resource and the cloud provider platform.
ICSF10. SLA and compliance management, including initial SLA negotiation and further SLA enforcement, must be implemented at the planning/design and operation stages. This functionality can outsourced to and implemented as a part of the user controlled or brokered cloud automation platform.
ICSF11. Brokered and third party security services should ensure cloud compliance with general international and applications domain specific security standards; Cloud Service Broker should include explicit compliance assessment stage when provisioning brokered services.

The presented requirements and design principles use and leverage the best practices in security design of regular Internet and web applications however extend them necessary security mechanisms to ensure bootstrapping of the virtualised environment to the cloud platform. They also reflect changing security paradigm in complex cloud based applications and infrastructures from formal security models to trust based that is in its own turn based on compliance based built trust. Cloud customers must trust cloud services providers and cloud services providers are interested in complying with the industry verified security design principles and standards.

6. Summary and further research

The presented paper summarises ongoing research by the authors into developing foundation design principles for the cloud based Big Data Infrastructure that is GDPR compliant by design and can enable the future data driven services that will allow data exchange as economic goods and will use such infrastructure and functional components as (open) data markets and data exchange what if defined as a priority research and development areas in the EU Horizon 2020 Framework Programme ICT-13-2018-2019: Supporting the emergence of data markets and the data economy that defines data security and trusted infrastructure as key enabling technologies.

References

[1] General Data Protection Regulation (GDPR). Legal Act, Official Journal of the European Union, 27 April 2016 [online] http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[2] Cambridge Analytica explained: The facts, implications, and open questions, by Stephanie Borg Psaila, 24 April 2018 [online] https://dig.watch/trends/cambridge-analytica

[3] The Facebook and Cambridge Analytica scandal, explained with a simple diagram, By Alvin Chang, May 2, 2018 [online] https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

[4] Demchenko, Y., P.Membrey, C.Ngo, C. de Laat, D.Gordijenko, Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure, Proc. Secure Data Management (SDM'13) Workshop. Part of VLDB2013 conference, 26-30 August 2013, Trento, Italy.

[5] Yuri Demchenko, Fatih Turkmen, Mathias Slawik, Defining Intercloud Security Framework and Architecture Components for Multi-Cloud Data Intensive Applications. Sixth IEEE International Workshop on Cloud Computing Interclouds, Multiclouds, Federations, and Interoperability (Intercloud 2017), In Proc. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. Madrid, Spain, May 14-17, 2017

[6] Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

[7] Data Management Maturity (DMM) model by CMMI (Capability Maturity Model Integration), 2017] [online] https://cmmiinstitute.com/store/data-management-maturity-(dmm)

[8] General Data Protection Regulation (GDPR) and Data Science, By Thomas Dinsmore, July 13, 2017 [online] https://vision.cloudera.com/generaal-data-protection-regulation-gdpr-and-data-science/

[9] Data Governance Target Operating Model: A holistic approach for utilizing data as a competitive advantage, 15. December 2015 [online] https://www.bankinghub.eu/banking/operations/data-governance-target-operating-model

[10] Demchenko, Y., C. de Laat, P. Membrey, Defining architecture components of the Big Data Ecosystem, Collaboration Technologies and Systems (CTS), 2014 International Conference.

[9] Cloud Controls Matrix (CCM) [online] https://cloudsecurityalliance.org/research/ccm/

[10] CSA GRC Stack: Governance, Risk Management and Compliance [online] https://cloudsecurityalliance.org/research/grc-stack/

[11] Consensus Assessments Initiative Questionnaire (CAIQ) [online] https://cloudsecurityalliance.org/research/cai/)

[12] Overview of the EU General Data Protection Regulation, Hunton&Williams [online] https://www.huntonregulationtracker.com/files/Uploads/Documents/EU%20Data%20Protection%20Reg%20Tracker/HuntonWilliams-GDPR-Management-Guide.pdf

[13] EU-U.S. Privacy Shield: stronger protection for transatlantic data flows Adopted 12 July 2016. Repeals former Safe Harbor Framework [online] http://ec.europa.eu/justice/data-protection/international-transfers/eu-us-privacy-shield/index_en.htm

[14] Todorov, D. & Ozkan, Y. (November 2013) 'AWS security best practices', Amazon Web Services [Online]. Available from: http://media.amazonwebservices.com/AWS_Security_Best_Practices.pdf

[15] Dominic Betts, et al, Developing Multi-tenant Applications for the Cloud on Microsoft Windows Azure, Third Edition, Microsoft. 2012. [online] http://download.microsoft.com/download/D/0/6/D0618696-2F91-4F7F-9477-63FC90D4D29E/Developing%20Multi-tenant%20Applications%20for%20the%20Cloud%203rd%20Edition.pdf