Big Data and Analytics in the age of the GDPR

Piero A. Bonatti Università di Napoli Federico II

Sabrina Kirrane WU Vienna

Joint work with all SPECIAL's partners

IEEE SERVICES - Big Data Congress, 8/7/2019

SPECIAL

European Commission

Horizon 2020
European Union funding
for Research & Innovation

Outline

- Impact of the GDPR (the new European General Data Protection Regulation) on Big Data Processing
  - Especially Analytics
  - Strategic role of consent
  - Difficulties related to anonymization

- A brief summary of approach to compliance with the GDPR of SPECIAL
  - An H2020 project funded under the Big Data PPP call

Big Data and Personal Data Processing

- Some of the most interesting big data are personal information
- A trivial example: location data
- Useful for the public good *and* business
- Links to data subjects:
  - Explicit (phone numbers, device IDs, account names, …)
  - Implicit (e.g. through location data mining)
  - “Fingerprints” based on location data are particularly precise

Constraints on Personal Data Processing – the GDPR

- The GDPR (the new European Data Protection Regulation) significantly restricts personal data processing
- It applies to all organizations that track or provide services to European citizens (Art. 3)
- Infringements have severe consequences
  - On reputation
  - Sanctions of up to 4% of worldwide annual turnover (but not less than 20 million €)
- *Data controllers* (the entities that process personal data) are looking for methodological and technological means to comply with the GDPR

The Role of Consent in GDPR compliance

- By default, the GDPR forbids personal data processing
- Then, in Art. 6, it provides a list of exceptions (legal bases for personal data processing), for example
  - Public interest, Vital interest of the data subject,
  - Legitimate interest of the controller, Contracts, …
  - Explicit consent of the data subject
- Consent is the mainstream approach to personal data processing
  - The other legal bases are restricted by provisos & caveats
  - Incompatible with many application domains

The Role of Anonymous Data in GDPR compliance

- The GDPR states that anonymous data are not personal data
  - So anonymous data can be freely used
- On the one hand, the GDPR encourages the use of anonymous data
- On the other hand, technical difficulties arise due to GDPR’s strict definition of anonymity…

What is (not) Anonymous

- *Personal data* means any information relating to an identified or identifiable natural person [the *data subject*] (Art. 4(1))
- Anonymous data are not personal and can be freely used
- Identification can be
  - Direct or indirect
  - Via names, IDs, location data,
  - Any factors related to the physical, physiological, genetic, mental, economic, cultural or social identity
- To determine whether a person is identifiable the controller shall consider
  - all the means reasonably likely to be used to identify the person
  - by the controller or any other entity (Recital 26)

Difficulties in Establishing Anonymity

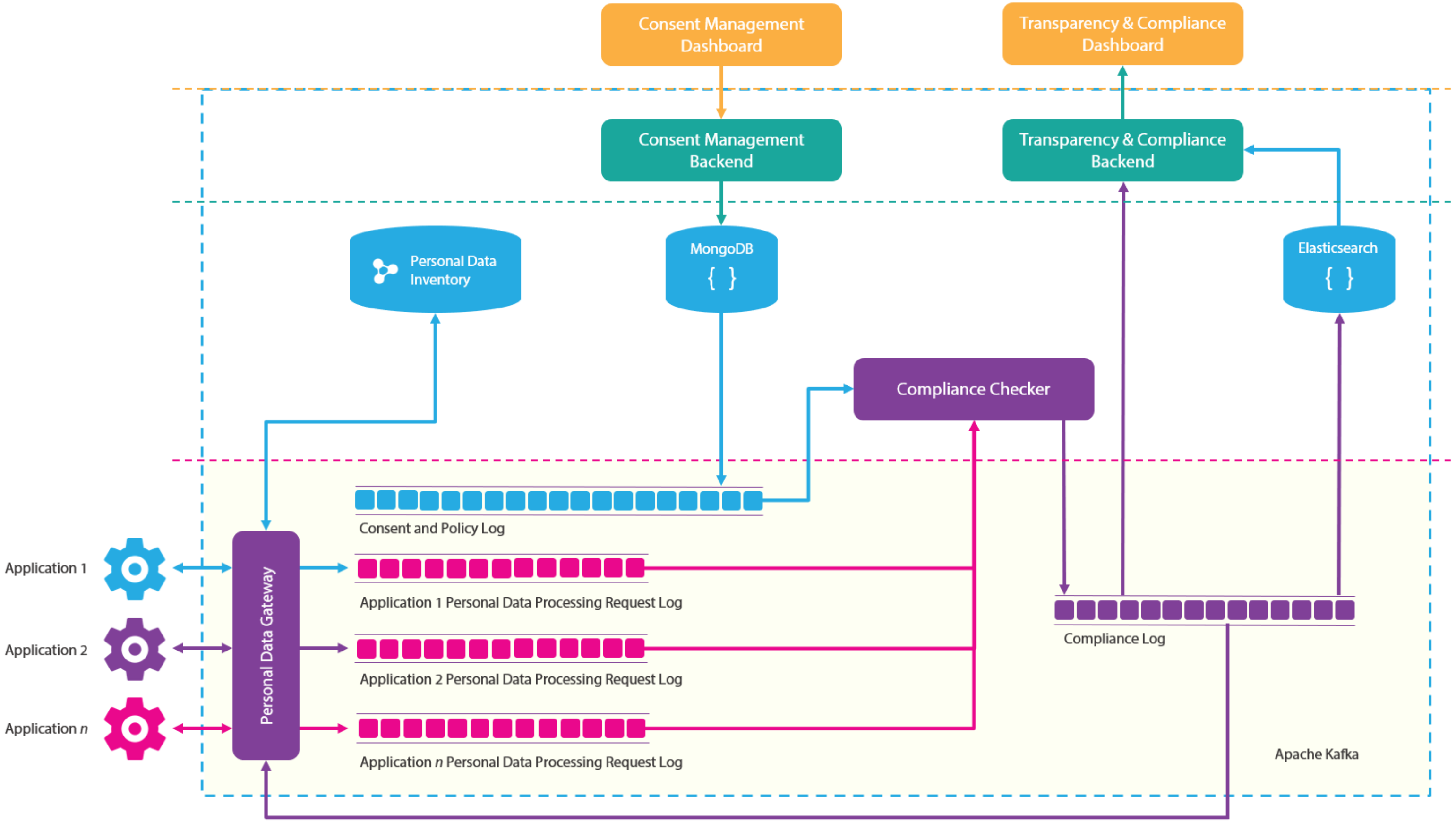
- Increasingly effective and scalable tools for analytics [indirect identification]
- Mismatch between legal and technical anonymity
- Examples of technical guarantees:
  - Number of indistinguishable individuals in the data source [*k*-anonymity]
  - Variety of their properties [*l*-diversity, *t*-closeness]
  - Probabilistic indistinguishability of sources with/without a given data subject [ε-differential privacy]
- All sensitive to attacks based on background knowledge
- Which parameters yield *legally anonymous* output ?
- Which background knowledge is available to attackers?

Data Anonymization as Risk Management or Consent Fostering

- So, in practice, anonymization involves risks
  - Benefits of analytics vs Risk of reputation loss and sanctions
  - What if tomorrow the controller is sued by a re-identified data subject?
  - We observed different companies adopting different strategies
- Legislators not likely to establish standard parameters that guarantee “legal anonymity”
  - How to reconcile the different notions of anonymity ?
  - How to estimate background knowledge ?
- Anonymization + Consent
  - Anonymization may encourage consent to processing
  - The legal basis for processing is consent (no risks)

Consent Management: SPECIAL's approach

- SPECIAL is an H2020 project funded under the Big Data PPP call
- Main goal: Supporting GDPR compliance, with a particular focus on consent management [given its strategic role]



Consent Management Dashboard
Transparency & Compliance Dashboard
Consent Management Backend
Transparency & Compliance Backend
Personal Data Inventory
MongoDB { }
Elasticsearch { }
Compliance Checker
Consent and Policy Log
Application 1
Application 2
Application *n*
Personal Data Gateway
Application 1 Personal Data Processing Request Log
Application 2 Personal Data Processing Request Log
Application *n* Personal Data Processing Request Log
Compliance Log
Apache Kafka

Modelling Consent, Business Policies and the GDPR

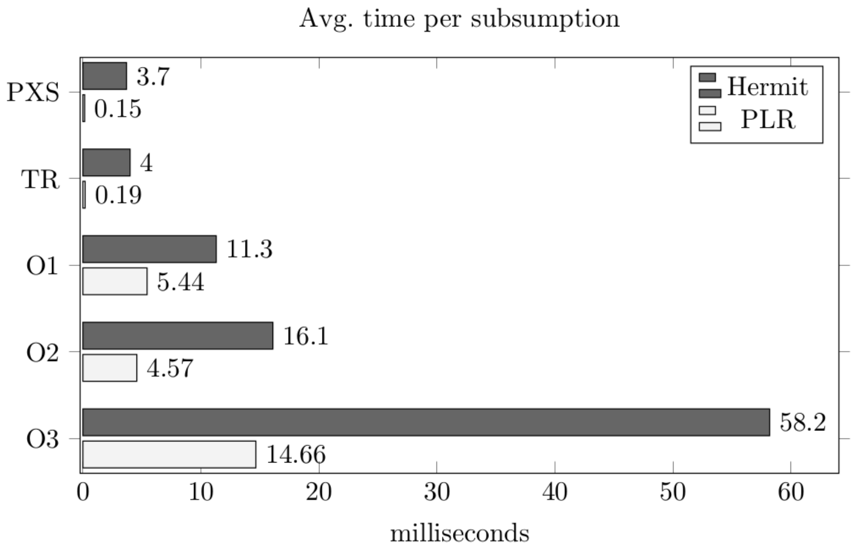
- SPECIAL’s data usage policy model, derived from the GDPR:
  - Purpose of the processing
  - Data categories involved in the processing
  - Recipients
  - Transfers to other countries
  - Time limits for erasure
- Extensions for business policies & GDPR
  - Duties, Legal bases
- The vocabularies/ontologies for purposes, data categories etc. are being defined by W3C’s DPVCG
  - Data Privacy Vocabularies and Controls Community Group
  - Promoted by SPECIAL
  - Wider range of stakeholders

SPECIAL's Policy Language

- The policy model can be encoded with an extension of Jason or a new profile of OWL2
- Some features:
  - Standard encoding
  - Extensibility (expressiveness) → to accommodate DPVCG's work
    - ➢ Without changing algorithms
  - Formal semantics → algorithm "certification" & interoperability
    - ➢ Correctness / completeness guarantees
    - ➢ Coherent compliance checking, explanations, policy validation, auditing …
    - ➢ Shared interpretation of *sticky policies*
  - Class-oriented → obtain & model *general consent*
    - ➢ Leverage "similar purposes", avoid repeated, similar consent requests

Scalability of Compliance Checking in PL

- $\mathcal{PL}$ is the new *policy logic* profile of OWL2 [IJCAI’18]
- Each compliance check takes 150-190 μ-sec in Java without resorting to parallelism
- By embedding our checker PLR in the BD architecture we can check compliance in real time, in hard telco use cases



Other Big Data Aspects in SPECIAL

- Volume: The transparency log keeps the history of all personal data processing events.
- Variety: Due mainly to:
  - The variety of personal data involved
  - The integration in existing systems
  - Interoperability [data transfers]

  SPECIAL leverages linked data, semantics, and DPVCG’s work
- Veracity: Faithfulness of policies & logged events

  [work in progress]

Conclusions & Challenges

- Consent is the mainstream approach to personal data processing under the GDPR
- Anonymization is not generally applicable and involves legal risks
  - Anonymous ≠ Anonymized
- *However it is not clear how to do exploratory analytics with consent*
  - Consent requests should specify the purpose
  - The purpose is not known a priori
  - Currently exploratory analytics only possible on anonymous data
- *Anonymization decreases the utility of data*
  - SPECIAL is studying *natively private data mining methods*
  - Goal: introduce no additional noise to protect the data

Conclusions & Challenges (II)

- *Usability*
  - Data subjects awareness / understanding of privacy & consent [dashboards, explanations]
  - Managing large histories of data usage events [dashboards]
  - Asking for consent without annoying the user
    - Monolithic requests are too large & complex
    - Pointwise requests are too frequent
    - SPECIAL is experimenting with a novel *dynamic strategy*

Conclusions & Challenges (II)

- *Usability*
  - Data subjects awareness / understanding of privacy & consent [dashboards, explanations]
  - Managing large histories of data usage events [dashboards]
  - Asking for consent without annoying the user
    - Monolithic requests are too large & complex
    - Pointwise requests are too frequent
    - SPECIAL is experimenting with a novel *dynamic strategy*

QUESTIONS?