GDPR Compliance For Databases

Put together by Primal Pappachan
Data Privacy is important!
What is GDPR and what does it do?

- General Data Protection Regulation
  - EU Privacy Law
  - Proposed on April 14, 2016 and came into effect on May 25, 2018
  - Applies to all EU Members
- **Worldwide scope:** Applies to all companies that collect, store, and process data belonging to EU citizens
- Similar laws in other parts of the world
  - California Consumer Privacy Act (CCPA – Jan 2020)
  - Brazil’s Lei Geral de Proteção de Dados (LGPD – Sept 2020)
  - India’s Personal Data Protection Bill (Proposed in 2019)
What does it do?

- Establishes privacy and protection of **personal data** as a fundamental right
- 99 legal articles + 173 Recitals
  - Regulate the **collection, processing, protection, transfer, and deletion** of personal data
- Grants Rights to People
  - For protection and privacy of their data
- Assigns Responsibilities to Companies
  - For safe and responsible collection and processing
- Risks for serious consequences for non-compliance
  - Max Penalty of 4% of global revenue or €20 million, whichever is greater
GDPR has been doing ‘fine’ so far

**TOP 5 BIGGEST GDPR FINES**

*Only includes final & binding fines*

<table>
<thead>
<tr>
<th>Country</th>
<th>Fine (€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Inc.</td>
<td>50,000,000</td>
</tr>
<tr>
<td>TIM - Telecom Provider</td>
<td>27,800,000</td>
</tr>
<tr>
<td>Austrian Post</td>
<td>18,000,000</td>
</tr>
<tr>
<td>Wind Tre S.p.A.</td>
<td>16,700,000</td>
</tr>
<tr>
<td>Deutsche Wohnen SE</td>
<td>14,500,000</td>
</tr>
</tbody>
</table>

**Total reported GDPR fines imposed***

<table>
<thead>
<tr>
<th>Country</th>
<th>Fine (€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>51.1m</td>
</tr>
<tr>
<td>Germany</td>
<td>24.6m</td>
</tr>
<tr>
<td>Austria</td>
<td>18.1m</td>
</tr>
<tr>
<td>Italy</td>
<td>11.6m</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>3.2m</td>
</tr>
</tbody>
</table>

*From 25 May 2018 to 17 January 2020

Source: https://www.enforcementtracker.com/?insights
GDPR Overview

- Can be broadly categorized into five categories
  - Articles 1-11 layout definitions and **principles of data processing**
  - Articles 12-23 establish **rights of the people (data subjects)**
  - Articles 24-50 mandate **responsibilities of the data controllers and processors**
  - Articles 50–76 describe roles and tasks of supervisory authorities
  - Rest cover liabilities, penalties, and specific situations

- Out of the 99 GDPR articles, **31 relate to behavior of data storage systems**
  compared to **11 that relate to compute and network infrastructure** (Shastri et al.)
GDPR Roles

- Data Subject
- Controller
- Processor
  - Processes data on behalf of controller
- Supervisory Authority
  - Public authorities of the controller or data subject location and responsible for monitoring application of regulation

Art. 4 Definitions
# (6 + 1) Principles of Personal Data Processing

1. Processed lawfully, fairly, and in a transparent manner (lawfulness, fairness, and transparency)
2. Collected for specific and legitimate purposes; data cannot be used for anything other stated purposes (Purpose limitation)
3. Relevant and limited to requirements of processing (Data minimisation)
4. Kept up to date and inaccuracies fixed or removed (Accuracy)
5. Stored for as long as specified in the retention policy (Storage limitation)
6. Protected against unauthorised access, accidental loss, or damage (Integrity and confidentiality)
7. Able to demonstrate compliance with above principles (Accountability)

Art. 5 GDPR Principles relating to processing of personal data
What is personal anyway?

- Any information that relates to a person that can be used directly or indirectly to identify them
- Interpreted as broadly as possible
  - Recordings of work times and lunch breaks
  - Written answers from a candidate for a test
  - Tracking IP address and network activity
  - Search terms sent to Google
- Particularly sensitive
  - healthcare, racial, sexual, political, religious, genetic, and biometric data

Art. 4 Definitions
Rights of data subjects

15  Right of access to personal data

16  Right of rectification

17  Right to erasure / to be forgotten

18  Right to restrict processing

20  Right to data portability

21  Right to object

22  Right to withdraw from Automated Decision-making
## Responsibilities of data controllers

<table>
<thead>
<tr>
<th>Page(s)</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>24, 25</td>
<td>Designing secure infrastructure</td>
</tr>
<tr>
<td>30</td>
<td>Maintain records of processing</td>
</tr>
<tr>
<td>33, 34</td>
<td>Notify data breaches within 72 hours</td>
</tr>
<tr>
<td>35, 36</td>
<td>Analyze risks prior to processing large amounts of personal data</td>
</tr>
<tr>
<td>37-39</td>
<td>Designate a Data Protection Officer</td>
</tr>
<tr>
<td>44</td>
<td>Controlling location of data</td>
</tr>
<tr>
<td></td>
<td>Create interfaces for users to exercise their GDPR rights</td>
</tr>
</tbody>
</table>
Examples of Compliance? - Amazon

AWS Service Capabilities for Privacy Considerations

New or updated privacy regulations around the world are introducing requirements for data protection, security, and compliance. Regulatory privacy themes include (but are not limited to) the ability to delete, encrypt, and monitor processing of personal data. AWS services have features capabilities that may enable customer compliance.

Click the check marks below for AWS service documentation about how AWS services help customers with encryption, deletion, and monitoring of processing.

<table>
<thead>
<tr>
<th>Service</th>
<th>Encryption</th>
<th>Deletion</th>
<th>Monitoring of Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa for Business</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon API Gateway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon AppStream 2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Athena</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Chime</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon CloudFront</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon CloudSearch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon CloudWatch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Cognito</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Comprehend</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Connect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon DynamoDB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Elastic Block Store (Amazon EBS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Elastic Compute Cloud (Amazon EC2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Elastic Container Registry (Amazon ECR)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Amazon Macie is a fully managed data security and data privacy service that uses machine learning and pattern matching to discover and protect your sensitive data in AWS.

As organizations manage growing volumes of data, identifying and protecting their sensitive data at scale can become increasingly complex, expensive, and time-consuming. Amazon Macie automates the discovery of sensitive data at scale and lowers the cost of protecting your data. Macie automatically provides an inventory of Amazon S3 buckets including a list of unencrypted buckets, publicly accessible buckets, and buckets shared with AWS accounts outside those you have defined in AWS Organizations. Then, Macie applies machine learning and pattern matching techniques to the buckets you select to identify and alert you to sensitive data, such as personally identifiable information (PII). Macie’s alerts, or findings, can be searched and filtered in the AWS Management Console and sent to Amazon EventBridge, formerly called Amazon CloudWatch Events, for easy integration with existing workflow or event management systems, or to be used in combination with AWS services, such as AWS Step Functions to take automated remediation actions. This can help you meet regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and General Data Privacy Regulations (GDPR). You can get started with Amazon Macie by leveraging the 30-day free trial for bucket evaluation. The trial includes 30 days of Amazon S3 bucket inventory and bucket-level security and access control assessment at no cost. Note that sensitive data discovery is not included in the 30-day free trial for bucket evaluation.
Examples of Compliance? - Google Cloud

[Diagram showing stages of deletion process]

Stage 1: Deletion request
- Ordinary storage & processing

Stage 2: Soft deletion & recovery period
- Active systems: around two months

Stage 3: Removal from active systems
- Active systems: around two months
- Backup systems: around six months

Stage 4: Backup expiration

https://cloud.google.com/security/deletion
Example of Compliance? Consentua
Example of Compliance? Kafka

https://www.privitar.com/
GDPRBench Approach to building Compliance

**Analyze**
Translate GDPR articles into system-level capabilities and characteristics

**Build**
Implement GDPR requirements in Redis and PostgreSQL

**Measure**
Benchmark compliant systems against GDPR workloads

Supreeth Shastri, Vinay Banakar, Melissa Wasserman, Arun Kumar, and Vijay Chidambaram.
*Understanding and Benchmarking the Impact of GDPR on Database Systems* VLDB 2020
An Example of Compliance

**Store Data with a Timeline for Deletion**

Art. 5 (Storage Limitation) and Art. 17 (Right to be forgotten)

*GDPR-compliant data store should have support for*
- Associating **time-to live** with data
- **Timely deletion** of data

**Keep Record of Data Processing Activity**

Art. 30 (Records of Processing Activity) and Art. 33 (Notification of Data Breach)

*GDPR-compliant data store should have support for*
- Associating an **audit trail** with data
- **Monitoring/logging** all data accesses

[Slide Source]
<table>
<thead>
<tr>
<th>No</th>
<th>GDPR article/clause</th>
<th>What they regulate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>PURPOSE LIMITATION</td>
<td>Collect data for explicit purposes</td>
</tr>
<tr>
<td>5</td>
<td>STORAGE LIMITATION</td>
<td>Do not store data indefinitely</td>
</tr>
<tr>
<td>13</td>
<td>INFORMATION TO BE PROVIDED [...]</td>
<td>Inform customers about all the GDPR metadata associated with their data</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>RIGHT OF ACCESS BY USERS</td>
<td>Allow customers to access all their data</td>
</tr>
<tr>
<td>17</td>
<td>RIGHT TO BE FORGOTTEN</td>
<td>Allow customers to erasure their data</td>
</tr>
<tr>
<td>21</td>
<td>RIGHT TO OBJECT</td>
<td>Do not use data for any objected reasons</td>
</tr>
<tr>
<td>22</td>
<td>AUTOMATED INDIVIDUAL DECISION-MAKING</td>
<td>Allow customers to withdraw from fully algorithmic decision-making</td>
</tr>
<tr>
<td>25</td>
<td>DATA PROTECTION BY DESIGN AND DEFAULT</td>
<td>Safeguard and restrict access to data</td>
</tr>
<tr>
<td>28</td>
<td>PROCESSOR</td>
<td>Do not grant unlimited access to data</td>
</tr>
<tr>
<td>30</td>
<td>RECORDS OF PROCESSING ACTIVITY</td>
<td>Audit all operations on personal data</td>
</tr>
<tr>
<td>32</td>
<td>SECURITY OF PROCESSING</td>
<td>Implement appropriate data security</td>
</tr>
<tr>
<td>33</td>
<td>NOTIFICATION OF PERSONAL DATA BREACH</td>
<td>Share audit trails from affected systems</td>
</tr>
</tbody>
</table>
Articles to Attributes and Actions

GDPR Metadata

1. Purpose
2. Time to Live
3. Objections
4. Audit Trail
5. Origin and sharing
6. Automated Decision Making
7. Associated Person

GDPR Capabilities

1. Encryption
2. Monitoring
3. Access Control
4. Timely Deletion
5. Metadata-based querying
Characterizing Personal Data

● **Purpose**
  ○ Collected and processed based on purposes; No purpose bundling

● **Time to Live**
  ○ As long as necessary to serve the purpose; Should be provided to customer at the time of collection

● **Objections**
  ○ Right to object for any purpose

● **Audit Trail**
  ○ Maintain Records of processing activities for every personal data item; In event of data breach use this to report number and details of records exposed
Characterizing Personal Data

- **Origin and sharing**
  - Origin of data and external entities with whom the data has been shared (Data Provenance)

- **Automated Decision Making**
  - Allows users to ask which of their records were used in ADS and request that their records not be used

- **Associated Person**
  - Association of the data subject with a personal data item
Mechanisms for Protection

- **Timely deletion**
  - TTL and Right to Forget

- **Monitoring**
  - Compliance and Notification in the event of data breaches

- **Indexing via Metadata**
  - Access based on and modify metadata fields

- **Encryption**
  - At rest and in transit

- **Access Control**
  - Limited access based on purposes, for specific entities, for a predefined duration of time
## Blueprint for GDPR compliant database systems

<table>
<thead>
<tr>
<th>No</th>
<th>GDPR article/clause</th>
<th>What they regulate</th>
<th>Impact on database systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>PURPOSE LIMITATION</td>
<td>Collect data for explicit purposes</td>
<td>Purpose, Metadata indexing</td>
</tr>
<tr>
<td>5</td>
<td>STORAGE LIMITATION</td>
<td>Do not store data indefinitely</td>
<td>TTL, Timely deletion</td>
</tr>
<tr>
<td>13</td>
<td>INFORMATION TO BE PROVIDED [...]</td>
<td>Inform customers about all the GDPR metadata associated with their data</td>
<td>Purpose, TTL, Origin, Sharing, Metadata indexing</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>RIGHT OF ACCESS BY USERS</td>
<td>Allow customers to access all their data</td>
<td>Person id, Metadata indexing</td>
</tr>
<tr>
<td>17</td>
<td>RIGHT TO BE FORGOTTEN</td>
<td>Allow customers to erase their data</td>
<td>TTL, Timely deletion</td>
</tr>
<tr>
<td>21</td>
<td>RIGHT TO OBJECT</td>
<td>Do not use data for any objected reasons</td>
<td>Objections, Metadata indexing</td>
</tr>
<tr>
<td>22</td>
<td>AUTOMATED INDIVIDUAL DECISION-MAKING</td>
<td>Allow customers to withdraw from fully algorithmic decision-making</td>
<td>Automated decisions, Metadata indexing</td>
</tr>
<tr>
<td>25</td>
<td>DATA PROTECTION BY DESIGN AND DEFAULT</td>
<td>Safeguard and restrict access to data</td>
<td>Access control</td>
</tr>
<tr>
<td>28</td>
<td>PROCESSOR</td>
<td>Do not grant unlimited access to data</td>
<td>Access control</td>
</tr>
<tr>
<td>30</td>
<td>RECORDS OF PROCESSING ACTIVITY</td>
<td>Audit all operations on personal data</td>
<td>Audit trail, Monitor and log</td>
</tr>
<tr>
<td>32</td>
<td>SECURITY OF PROCESSING</td>
<td>Implement appropriate data security</td>
<td>Encryption</td>
</tr>
<tr>
<td>33</td>
<td>NOTIFICATION OF PERSONAL DATA BREACH</td>
<td>Share audit trails from affected systems</td>
<td>Audit trail, Monitor and log</td>
</tr>
</tbody>
</table>

1. Handle metadata explosion  
2. Support data protection by design  
3. Support GDPR queries
- Existing benchmarks do not recognize abstraction of personal data
- Diversity of roles makes it complex to benchmark one thing
- Currently impossible to compare compliance levels or performance of today’s systems supporting GDPR

### Data Record

<table>
<thead>
<tr>
<th>Key</th>
<th>Data</th>
<th>Purpose</th>
<th>TTL</th>
<th>User</th>
<th>Objections</th>
<th>Automated Decisions</th>
<th>Third Party Sharing</th>
<th>Originating Source</th>
</tr>
</thead>
</table>
GDPR Workload

**Workload characteristics**

* Twice number of updates as creates and deletes
  * Uniform distribution

* Based on Google’s implementation of RTBF
  * Zipf distribution

* Based on workloads from existing benchmarks
  * Metadata operations based on GDPR analysis (20%)

* Based on European’s Data Board summary of first 9 months of roll out
  * Zipf and uniform distribution
Benchmark Metrics

- **Correctness**
  - Validation of metadata-based access control
  - Percentage of query responses that match the results
  - Cumulative across 4 workloads

- **Completion Time**
  - Separately for each workload
  - More important than latency as utility depends upon completion of operation
  - E.g., Google Cloud deletion time of 180 days as we saw earlier

- **Space Overhead**
  - Total size of database/Total size of personal data (always > 1)
  - Tradeoff between reduction of storage versus completion time (e.g., compression)
Implementation - Benchmark

- Adapted YCSB (2010)
  - Added GDPR workloads
  - Modified workload executor to parse GDPR queries
  - Modified the DB interface layer for two different databases
- Redis - NoSQL store
- PostgreSQL - RDBMS
- System-C - Enterprise DBMS with in-built compliance
- Around 2 months of work with lots of scripting/coding
## Making DBMS Compliant

<table>
<thead>
<tr>
<th>Feature</th>
<th>redis</th>
<th>PostgreSQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encryption</td>
<td>3rd party lib</td>
<td>3rd party lib</td>
</tr>
<tr>
<td>TTL/Timely deletion</td>
<td>Code change</td>
<td>Scripting</td>
</tr>
<tr>
<td>Monitoring/Logging</td>
<td>Code change</td>
<td>Configure</td>
</tr>
<tr>
<td>Metadata Indexing</td>
<td>Scripting</td>
<td>Configure</td>
</tr>
<tr>
<td>Access control</td>
<td>Scripting</td>
<td>Configure</td>
</tr>
<tr>
<td>GDPR queries</td>
<td>Code change</td>
<td>Scripting</td>
</tr>
</tbody>
</table>

### Implementation details

<table>
<thead>
<tr>
<th>Feature</th>
<th>redis</th>
<th>PostgreSQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUKS and TLS</td>
<td>Probabilistic algorithm with progressive delay</td>
<td>Modify INSERT queries and periodic checking (1s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Append-Only-File with code to log all actions</td>
<td>csv-log with row level security policies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Secondary indices</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>External Client</td>
<td>External Client</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental Results (Workloads)

100% correctness
3.5X space overhead

100% correctness
5.95X space overhead w/ metadata indices

GDPR workloads run faster and scale better on SQL databases due to PostgreSQL’s better optimizer and availability of secondary indices
When all features are enabled (solid bar), Redis experiences an overhead of 5×, compare to PostgreSQL’s 2× due to significant logging overhead (70% v/s 30%)
Experimental Results (Effect of Scale)

- Time taken for completion of 10K operations as new customers are added
- Neither system scales well for GDPR workloads as completion time linearly scales with size of database
Conclusions and Takeaways

- GDPR compliance requires modification in storage and processing of personal data records
- Today’s DBMSes do not support all the necessary features for achieving compliance
- Proposes a GDPR workload and performance comparison on two different systems
- Compliance is
  - hard and will result in performance overheads
  - easier in RDBMS than in NoSQL
  - a spectrum; allows exploration of tradeoff between strict compliance and high performance
Strengths and Weaknesses

- Through analysis of GDPR Articles
- First characterization of GDPR workload for different roles
- Mapping from legalese to Database System level requirements

- Ad Hoc implementation of compliance mechanisms (e.g., TTL)
- Missing details of implementation of some aspects (e.g., fine grained policy control, auditing)
- Correctness defined only for access control
- Considers compliance as binary with no knobs for adjustment (e.g., logging levels)
- Do not address anything about handling derived data
Related Work

- **DatumDB** – proposes an architectural vision for a database that natively supports guaranteed deletion and consent management (2019)
- **A Framework for GDPR Compliance** in Big Data Systems (2020)
- **Our own Privacy Enhanced IoT (PE-IoT)**

2. GDPRBench https://www.gdprbench.org/

3. GDPRToons http://www.gdprtoons.com/
Thank you!