Understanding and Benchmarking the Impact of GDPR on Database Systems

Supreeth Shastri  Vinay Banakar  Melissa Wasserman  Arun Kumar  Vijay Chidambaram

TExAS  Hewlett Packard Enterprise  UC San Diego  VMware
General Data Protection Regulation (GDPR)

Privacy and protection of personal data is a fundamental right of natural persons
General Data Protection Regulation (GDPR)

Privacy and protection of personal data is a fundamental right of natural persons

99 Legal Articles
Regulate the collection, processing, protection, transfer and deletion of personal data
General Data Protection Regulation (GDPR)

*Privacy and protection of personal data is a fundamental right of natural persons*

**99 Legal Articles**
Regulate the collection, processing, protection, transfer and deletion of personal data

**Grants Rights to People**
Grants all European people a right to protection and privacy of their personal data
General Data Protection Regulation (GDPR)

*Privacy and protection of personal data is a fundamental right of natural persons*

**99 Legal Articles**
Regulate the collection, processing, protection, transfer and deletion of personal data

**Grants Rights to People**
Grants all European people a right to protection and privacy of their personal data

**Assigns Responsibilities to Companies**
Those who collect and process personal data are solely responsible for its privacy and protection
General Data Protection Regulation (GDPR)

Privacy and protection of personal data is a fundamental right of natural persons.

99 Legal Articles
Regulate the collection, processing, protection, transfer and deletion of personal data.

Assigns Responsibilities to Companies
Those who collect and process personal data are solely responsible for its privacy and protection.

Grants Rights to People
Grants all European people a right to protection and privacy of their personal data.

Hefty Penalty
Max penalty of 4% of global revenue or €20 million, whichever is greater.
Complying with GDPR has been a challenge
Complying with GDPR has been a challenge.

- Google: €50M, French Data Protection Authority, Jan 2019
- Marriott: $123M, UK Data Protection Agency, Jun 2019
- British Airways: £183M, UK Data Protection Agency, Jun 2019
Complying with GDPR has been a challenge.

<table>
<thead>
<tr>
<th>Company</th>
<th>Fine (in Millions)</th>
<th>Authority/Agency</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>€50M</td>
<td>French Data Protection Authority</td>
<td>Jan 2019</td>
</tr>
<tr>
<td>Marriott Hotels &amp; Resorts</td>
<td>$123M</td>
<td>UK Data Protection Agency</td>
<td>Jun 2019</td>
</tr>
<tr>
<td>British Airways</td>
<td>£183M</td>
<td>UK Data Protection Agency</td>
<td>Jun 2019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Public Complaints</td>
<td>144,376</td>
</tr>
</tbody>
</table>

EU-wide (Year 1)
Personal Data

any information relating to an identified or identifiable natural person

GDPR §4(1)
Personal Data

any information relating to an identified or identifiable natural person

GDPR §4(1)
Personal Data

any information relating to an identified or identifiable natural person

GDPR §4(1)
Personal Data

any information relating to an identified or identifiable natural person

Prof. Albus Dumbledore

- Has a phoenix as pet
- Drinks coffee at 8am
- Published a paper at VLDB 2020

GDPR §4(1)
I have **eight rights!**

*Right to know, access, rectify, erase, object, port, restrict processing, and withdraw from automated processing*

---

**Personal Data**

*any information relating to an identified or identifiable natural person*

---

**Prof. Albus Dumbledore**

- Has a phoenix as pet
- Drinks coffee at 8am
- Published a paper at VLDB 2020
Personal Data

any information relating to an identified or identifiable natural person

Prof. Albus Dumbledore
- Has a phoenix as pet
- Drinks coffee at 8am
- Published a paper at VLDB 2020

I have eight rights!
Right to know, access, rectify, erase, object, port, restrict processing, and withdraw from automated processing

I have responsibilities
To obtain consent, track data usage, keep it secure, notify breaches etc.

GDPR §4(1)
Personal Data

any information relating to an identified or identifiable natural person

Prof. Albus Dumbledore
- Has a phoenix as pet
- Drinks coffee at 8am
- Published a paper at VLDB 2020

I have eight rights!
Right to know, access, rectify, erase, object, port, restrict processing, and withdraw from automated processing

I have responsibilities
To obtain consent, track data usage, keep it secure, notify breaches etc.
How to build a **GDPR-compliant** database system for storing personal-data?

**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
Store Data with a Timeline for Deletion
Store Data with a Timeline for Deletion

§ 5(1)(E): Storage Limitation

"[...] kept for no longer than is necessary for the purposes for which the personal data are processed [...]"
§ 5(1)(E): Storage Limitation

"[...] kept for no longer than is necessary for the purposes for which the personal data are processed [...]"

§17: Right To Be Forgotten

(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay [...]"
Store Data with a Timeline for Deletion

§ 5(1)(E): Storage Limitation
“[... ] kept for no longer than is necessary for the purposes for which the personal data are processed [...]”

§17: Right To Be Forgotten
(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay [...]

GDPR-compliant datastore should:

- Associate a **time-to-live** attribute with all data
- Have support for **timely deletion** of data
§ 5(1)(E): Storage Limitation
“[…] kept for no longer than is necessary for the purposes for which the personal data are processed […]”

§ 17: Right To Be Forgotten
(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay […]”

GDPR-compliant datastore should:

- Associate a **time-to-live** attribute with all data
- Have support for **timely deletion** of data
Store Data with a Timeline for Deletion

§ 17: Right To Be Forgotten

(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay [...] 

GDPR-compliant datastore should:

- Associate a **time-to-live** attribute with all data
- Have support for **timely deletion** of data

Keep Record of Data Processing Activity

§ 30: Records of Processing Activities

(1) Each controller [...] shall maintain a record of processing activities under its responsibility.
Store Data with a Timeline for Deletion

§ 5(1)(E): Storage Limitation
“[…] kept for no longer than is necessary for the purposes for which the personal data are processed […]”

§17: Right To Be Forgotten
(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay […]

GDPR-compliant datastore should:
- Associate a **time-to-live** attribute with all data
- Have support for **timely deletion** of data

Keep Record of Data Processing Activity

§ 30: Records of Processing Activities
(1) Each controller […] shall maintain a record of processing activities under its responsibility.

§ 33: Notification of A Data Breach
(1) the controller shall without undue delay and not later than 72 hours after having become aware of it, notify […] (3) The notification shall at least describe the nature of the personal breach.”
§ 5(1)(E): Storage Limitation
"[...] kept for no longer than is necessary for the purposes for which the personal data are processed [...]"

§17: Right To Be Forgotten
(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay [...]"

Store Data with a Timeline for Deletion

GDPR-compliant datastore should:
- Associate a **time-to-live** attribute with all data
- Have support for **timely deletion** of data

Keep Record of Data Processing Activity

GDPR-compliant datastore should:
- Associate an **audit trail** with all data
- Implement support for **monitoring/logging** of all data accesses

§ 30: Records of Processing Activities
(1) Each controller [...] shall maintain a record of processing activities under its responsibility.

§ 33: Notification of A Data Breach
(1) the controller shall without undue delay and not later than 72 hours after having become aware of it, notify [...] (3) The notification shall at least describe the nature of the personal breach."
Translating GDPR Articles into Systems-Level Attributes and Actions
Translating GDPR Articles into Systems-Level Attributes and Actions

*We analyzed all the 99 articles of GDPR, both individually and collectively…*
Translating GDPR Articles into Systems-Level Attributes and Actions

We analyzed all the 99 articles of GDPR, both individually and collectively…

GDPR Metadata

Associate seven behavioral attributes with personal data
We analyzed all the 99 articles of GDPR, both individually and collectively…

**GDPR Metadata**

Associate *seven behavioral attributes* with personal data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Purpose</td>
</tr>
<tr>
<td>2</td>
<td>TTL</td>
</tr>
<tr>
<td>3</td>
<td>Audit trail</td>
</tr>
<tr>
<td>4</td>
<td>Objections</td>
</tr>
<tr>
<td>5</td>
<td>Origin of data</td>
</tr>
<tr>
<td>6</td>
<td>Externally shared?</td>
</tr>
<tr>
<td>7</td>
<td>Use in automated decision-making?</td>
</tr>
</tbody>
</table>
Translating GDPR Articles into Systems-Level Attributes and Actions

We analyzed all the 99 articles of GDPR, both individually and collectively…

<table>
<thead>
<tr>
<th>GDPR Metadata</th>
<th>GDPR Capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associate <em>seven behavioral attributes</em> with personal data</td>
<td>Implement <em>five features</em> in the database system</td>
</tr>
<tr>
<td>1  Purpose</td>
<td></td>
</tr>
<tr>
<td>2  TTL</td>
<td></td>
</tr>
<tr>
<td>3  Audit trail</td>
<td></td>
</tr>
<tr>
<td>4  Objections</td>
<td></td>
</tr>
<tr>
<td>5  Origin of data</td>
<td></td>
</tr>
<tr>
<td>6  Externally shared?</td>
<td></td>
</tr>
<tr>
<td>7  Use in automated decision-making?</td>
<td></td>
</tr>
</tbody>
</table>
Translating GDPR Articles into Systems-Level Attributes and Actions

We analyzed all the 99 articles of GDPR, both individually and collectively…

**GDPR Metadata**

Associate *seven behavioral attributes* with personal data

1. Purpose
2. TTL
3. Audit trail
4. Objections
5. Origin of data
6. Externally shared?
7. Use in automated decision-making?

**GDPR Capabilities**

Implement *five features* in the database system

- Encryption
- Monitoring
- Access control
- Timely deletion
- Metadata-based querying
Translating GDPR Articles into Systems-Level Attributes and Actions

We analyzed all the 99 articles of GDPR, both individually and collectively…

**GDPR Metadata**

Associate *seven behavioral attributes* with personal data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Purpose</td>
</tr>
<tr>
<td>2</td>
<td>TTL</td>
</tr>
<tr>
<td>3</td>
<td>Audit trail</td>
</tr>
<tr>
<td>4</td>
<td>Objections</td>
</tr>
<tr>
<td>5</td>
<td>Origin of data</td>
</tr>
<tr>
<td>6</td>
<td>Externally shared?</td>
</tr>
<tr>
<td>7</td>
<td>Use in automated decision-making?</td>
</tr>
</tbody>
</table>

**GDPR Capabilities**

Implement *five features* in the database system

- Encryption
- Monitoring
- Access control
- Timely deletion
- Metadata-based querying

*Storage overhead*
We analyzed all the 99 articles of GDPR, both individually and collectively…

Translating GDPR Articles into Systems-Level Attributes and Actions

**GDPR Metadata**

Associate *seven behavioral attributes* with personal data

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Purpose</td>
</tr>
<tr>
<td>2</td>
<td>TTL</td>
</tr>
<tr>
<td>3</td>
<td>Origin of data</td>
</tr>
<tr>
<td>4</td>
<td>Externally shared?</td>
</tr>
<tr>
<td>5</td>
<td>Use in automated decision-making?</td>
</tr>
</tbody>
</table>

**GDPR Capabilities**

Implement *five features* in the database system

- Timely deletion
- Monitoring
- Access control
- Encryption
- Metadata-based querying

**Obstacles**

- Storage overhead
- Performance overhead
GDPR-Compliant Storage Systems
GDPR-Compliant Storage Systems

Goal: Introduce GDPR-compliance into two representative storage systems and measure its impact
GDPR-Compliant Storage Systems

Goal: Introduce GDPR-compliance into two representative storage systems and measure its impact

redis  PostgreSQL
# GDPR-Compliant Storage Systems

*Goal: Introduce GDPR-compliance into two representative storage systems and measure its impact*

<table>
<thead>
<tr>
<th>Encryption</th>
<th>PostgresQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTL/Timely deletion</td>
<td>Monitoring/Logging</td>
</tr>
<tr>
<td>Metadata Indexing</td>
<td>Access control</td>
</tr>
<tr>
<td>GDPR queries</td>
<td></td>
</tr>
</tbody>
</table>
## GDPR-Compliant Storage Systems

**Goal:** Introduce GDPR-compliance into two representative storage systems and measure its impact

<table>
<thead>
<tr>
<th>Component</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>
# GDPR-Compliant Storage Systems

Goal: Introduce GDPR-compliance into two representative storage systems and measure its impact

<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>

Performance overhead in Yahoo! Cloud Serving Benchmark (YCSB)
## GDPR-Compliant Storage Systems

**Goal:** Introduce GDPR-compliance into two representative storage systems and measure its impact

<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>

**Performance overhead in Yahoo! Cloud Serving Benchmark (YCSB):**

- redis: 80% ↓
- PostgreSQL: 50% ↓
How to benchmark database systems against GDPR workloads?
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench

GDPR Queries
Control- and data-path operations performed on GDPR datastores
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench

GDPR Queries
Control- and data-path operations performed on GDPR datastores

manage & administer
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench.

**GDPR Queries**

Control- and data-path operations performed on GDPR datastores.
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called **GDPRbench**

**GDPR Queries**

Control- and data-path operations performed on GDPR datastores

![Diagram showing GDPR queries and their operations]

- **Manage & administer**
- **Exercise rights**
- **Audit**
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench

GDPR Queries

Control- and data-path operations performed on GDPR datastores
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench

GDPR Queries
Control- and data-path operations performed on GDPR datastores
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench.

GDPR Queries
Control- and data-path operations performed on GDPR datastores

Benchmark Metrics
How to benchmark database systems against GDPR workloads?

We build a new open-source benchmark called GDPRbench

GDPR Queries
Control- and data-path operations performed on GDPR datastores

Benchmark Metrics

<table>
<thead>
<tr>
<th></th>
<th>Correctness</th>
<th>Completion Time</th>
<th>Space Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% responses that match the expected results</td>
<td>Time to complete all the workloads</td>
<td>Ratio of total DB size to size of personal data</td>
</tr>
</tbody>
</table>
How do our compliant systems perform against GDPRbench?
How do our compliant systems perform against GDPRbench?
How do our compliant systems perform against GDPRbench?

redis

3.5X
space overhead

PostgreSQL
How do our compliant systems perform against GDPRbench?

redis

3.5X
space overhead

PostgreSQL

5.95X
space overhead w/ metadata indices
How do our compliant systems perform against GDPRbench?

**redis**

- **3.5X** space overhead

**PostgreSQL**

- **5.95X** space overhead w/ metadata indices
How do our compliant systems perform against GDPRbench?

- **redis**
  - Completion Time (mins)
  - 3.5x space overhead

- **PostgreSQL**
  - Completion Time (mins)
  - 5.95x space overhead w/ metadata indices
How do our compliant systems perform against GDPRbench?

**redis**

3.5X space overhead

**PostgreSQL**

5.95X space overhead w/ metadata indices

**GDPR workloads run faster and scale better** on SQL than NoSQL databases
Real-World Implications
Real-World Implications

**Compliance may result in high performance overheads**

*Production system should be carefully analyzed before enabling/claiming compliance*
Real-World Implications

Compliance may result in high performance overheads
Production system should be carefully analyzed before enabling/claiming compliance

Compliance is easier in RDBMS than in NoSQL
Redis needed more involved changes and had much higher overhead
Compliance may result in high performance overheads
Production system should be carefully analyzed before enabling/claiming compliance

Compliance is easier in RDBMS than in NoSQL
Redis needed more involved changes and had much higher overhead

Compliance is a spectrum
Examine tradeoffs b/w strictness vs. efficiency
Need mechanisms for efficient auditing/timely deletion/indexing
We want to hear from you!
We want to hear from you!

Find out more at

https://www.GDPRbench.org/
Our Interpretation of GDPR

“In Law, nothing is certain but the expense” — Samuel Butler
Our Interpretation of GDPR

“In Law, nothing is certain but the expense” — Samuel Butler

Prof. Melissa Wasserman
Law faculty, UT Austin
Our Interpretation of GDPR

“In Law, nothing is certain but the expense” — Samuel Butler

Real-time
Complete GDPR tasks synchronously

Response Time

Eventual
Complete GDPR tasks asynchronously

Prof. Melissa Wasserman
Law faculty, UT Austin
Our Interpretation of GDPR

“In Law, nothing is certain but the expense” — Samuel Butler

Response Time

Real-time
Complete GDPR tasks synchronously

Eventual
Complete GDPR tasks asynchronously

Granularity of Rights

Per data item
Support right for every piece of data

Per service/person
Support rights at the level of service

Prof. Melissa Wasserman
Law faculty, UT Austin
Our Interpretation of GDPR

“In Law, nothing is certain but the expense” — Samuel Butler

- **Real-time**
  - Complete GDPR tasks synchronously

- **Eventual**
  - Complete GDPR tasks asynchronously

- **Per data item**
  - Support right for every piece of data

- **Per service/person**
  - Support rights at the level of service

**Granularity of Rights**

**Response Time**

**Strict interpretation** that will benchmark worst-case performance overhead

Prof. Melissa Wasserman
Law faculty, UT Austin
GDPR-Compliant Storage Systems

redis  Support for TTL and Timely Delete
Redis has built-in support for TTLs but… it internally erases expired keys using a **lazy randomized algorithm**
Redis has built-in support for TTLs but... it internally erases expired keys using a lazy randomized algorithm.
Redis has built-in support for TTLs but... it internally erases expired keys using a lazy randomized algorithm.

### GDPR-Compliant Storage Systems

**Support for TTL and Timely Delete**

- **3 hours**

![Diagram showing time to erase vs. total keys in data store]
Redis has built-in support for TTLs but… it internally erases expired keys using a **lazy randomized algorithm**

![Graph showing time to erase](image)

- **3 hours**

---

**A Key Expired In Redis, You Won't Believe What Happened Next**

27 Mar 2017 - Karan Karnath

At our scale, and assuming >25% expired keys at the beginning of time, it would take at least 110 hours to guarantee no expired keys in our cache.
Redis has built-in support for TTLs but... it internally erases expired keys using a lazy randomized algorithm.
Redis has built-in support for TTLs but... it internally erases expired keys using a **lazy randomized algorithm**

**Code change**: we changed the expiry algorithm to be deterministic

---

**GDPR-Compliant Storage Systems**

6M keys —> 4.5 days