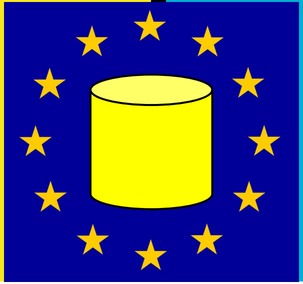
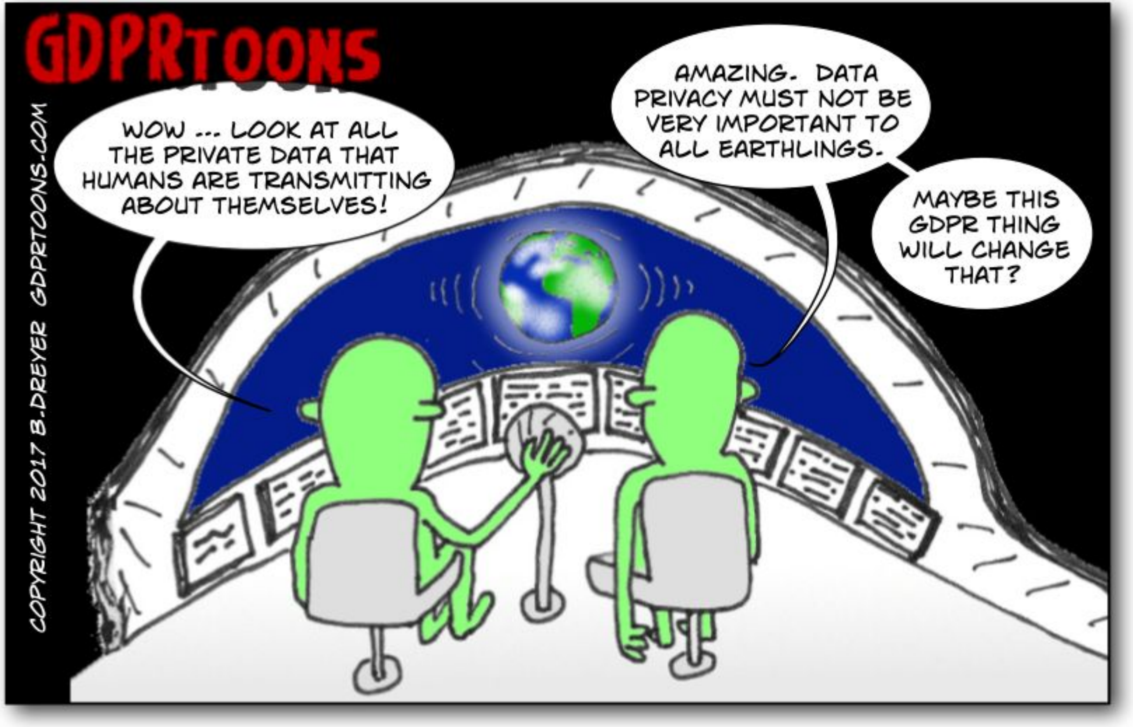


GDPR Compliance For Databases

Put together by
Primal Pappachan

Data Privacy is important!


GDPRTOONS
COPYRIGHT 2017 B.DREYER GDPRTOONS.COM
WOW ... LOOK AT ALL THE PRIVATE DATA THAT HUMANS ARE TRANSMITTING ABOUT THEMSELVES!
AMAZING. DATA PRIVACY MUST NOT BE VERY IMPORTANT TO ALL EARTHLINGS.
MAYBE THIS GDPR THING WILL CHANGE THAT?

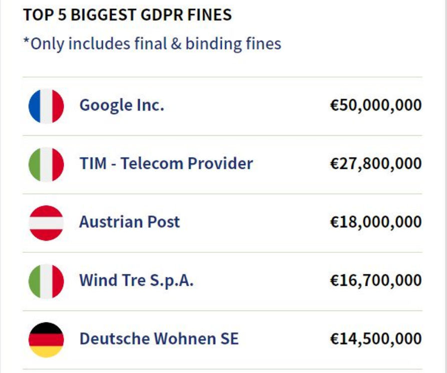
What is GDPR and what does it do?

- 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

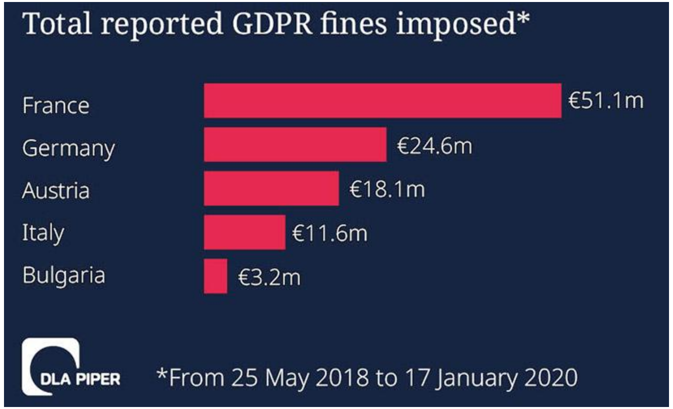
GDPR has been doing ‘fine’ so far



TOP 5 BIGGEST GDPR FINES

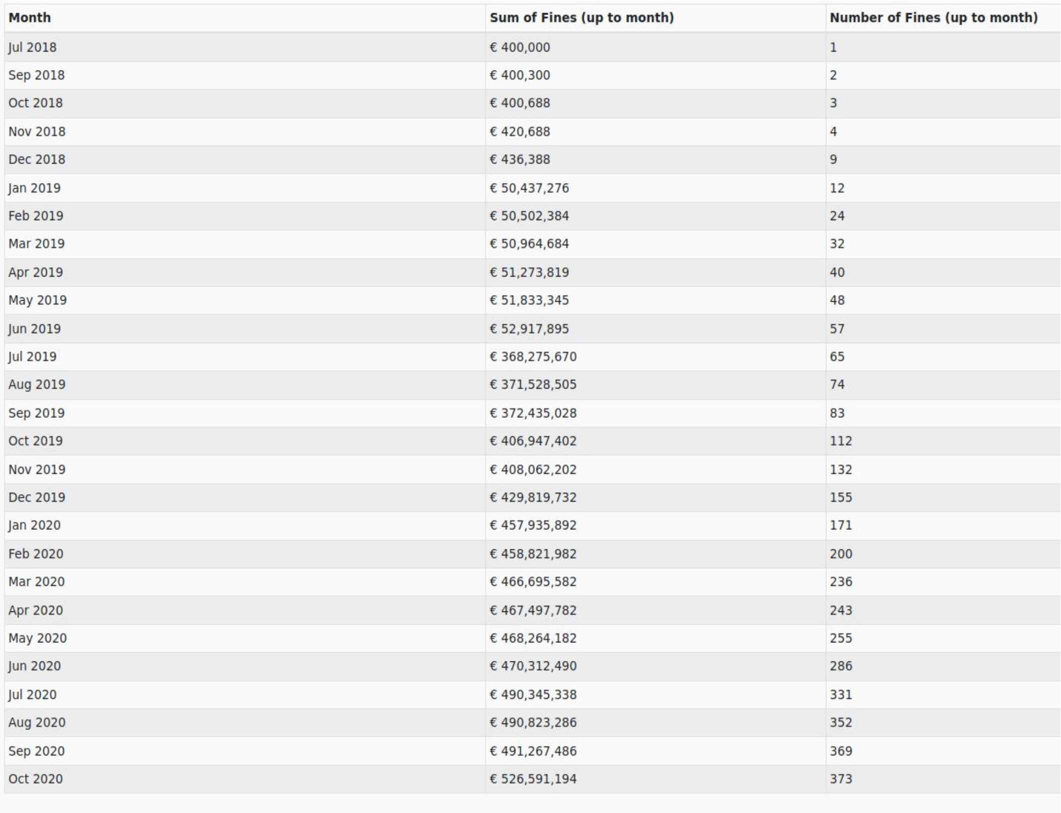
*Only includes final & binding fines

Google Inc.	€50,000,000
TIM - Telecom Provider	€27,800,000
Austrian Post	€18,000,000
Wind Tre S.p.A.	€16,700,000
Deutsche Wohnen SE	€14,500,000



Total reported GDPR fines imposed*

*From 25 May 2018 to 17 January 2020



Month	Sum of Fines (up to month)	Number of Fines (up to month)
Jul 2018	€ 400,000	1
Sep 2018	€ 400,300	2
Oct 2018	€ 400,688	3
Nov 2018	€ 420,688	4
Dec 2018	€ 436,388	9
Jan 2019	€ 50,437,276	12
Feb 2019	€ 50,502,384	24
Mar 2019	€ 50,964,684	32
Apr 2019	€ 51,273,819	40
May 2019	€ 51,833,345	48
Jun 2019	€ 52,917,895	57
Jul 2019	€ 368,275,670	65
Aug 2019	€ 371,528,505	74
Sep 2019	€ 372,435,028	83
Oct 2019	€ 406,947,402	112
Nov 2019	€ 408,062,202	132
Dec 2019	€ 429,819,732	155
Jan 2020	€ 457,935,892	171
Feb 2020	€ 458,821,982	200
Mar 2020	€ 466,695,582	236
Apr 2020	€ 467,497,782	243
May 2020	€ 468,264,182	255
Jun 2020	€ 470,312,490	286
Jul 2020	€ 490,345,338	331
Aug 2020	€ 490,823,286	352
Sep 2020	€ 491,267,486	369
Oct 2020	€ 526,591,194	373

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

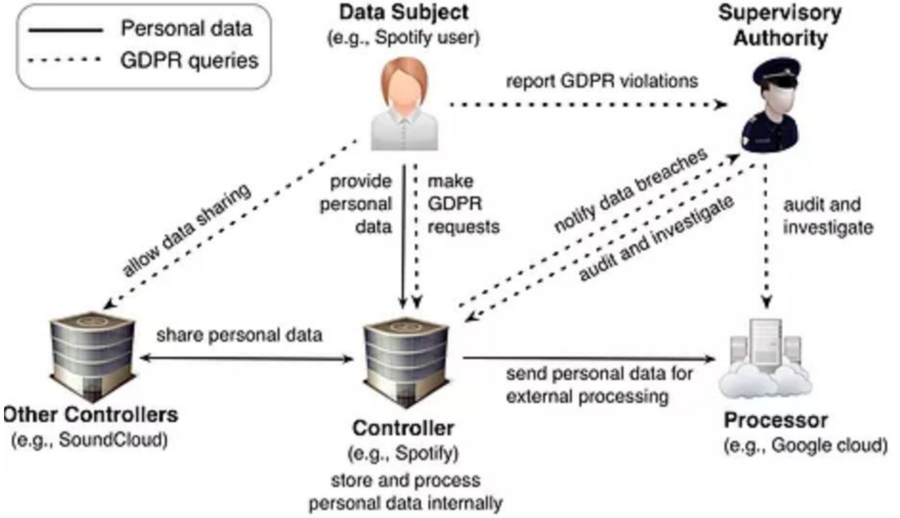


Image Reference

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

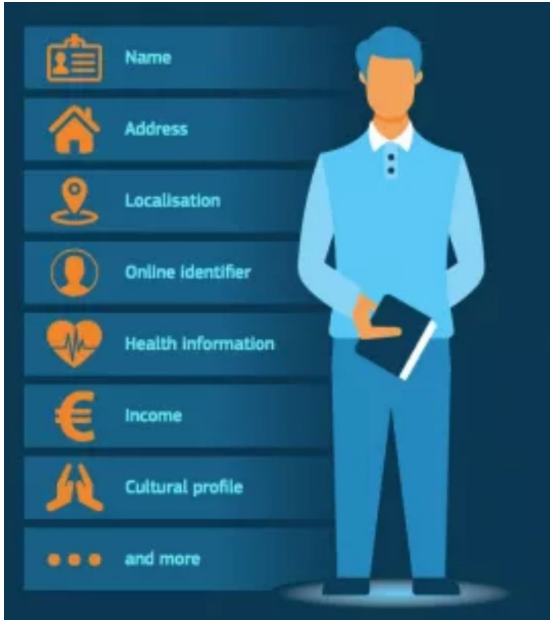
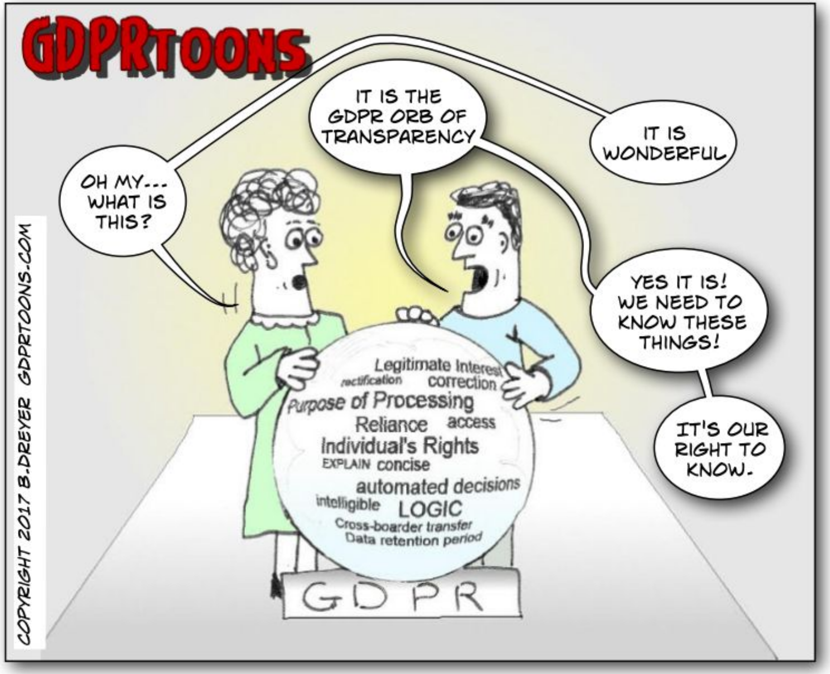


Image Reference:

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

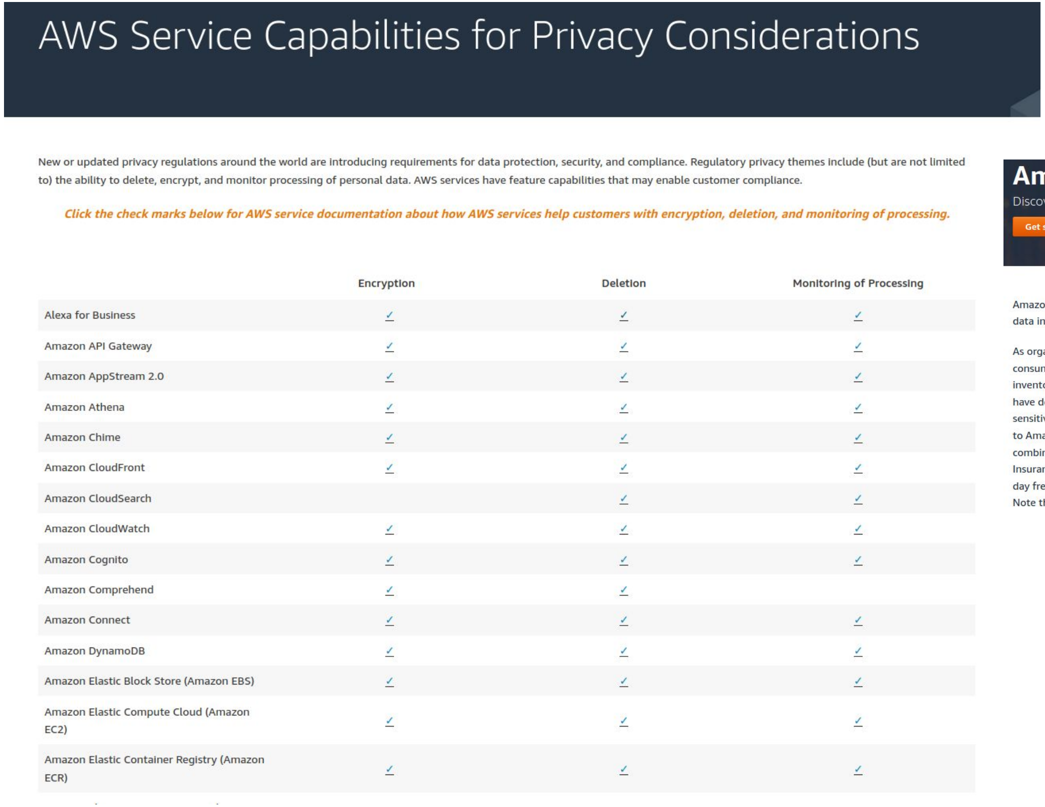


Responsibilities of data controllers

24, 25	Designing secure infrastructure
30	Maintain records of processing
33, 34	Notify data breaches within 72 hours
35, 36	Analyze risks prior to processing large amounts of personal data
37-39	Designate a Data Protection Officer
44	Controlling location of data
_	Create interfaces for users to exercise their GDPR rights

GDPRTOONS
ADEQUATE RELEVANT LIMITED
ACCURATE
SPECIFIED PURPOSE
TIME LIMITED
LAWFULL FAIR TRANSPARENT
ADEQUATE SECURITY
COPYRIGHT 2017 B.DREYER GDPRTOONS.COM
INSPIRED BY: H.HADDON - L.KENDALL

Examples of Compliance? - Amazon

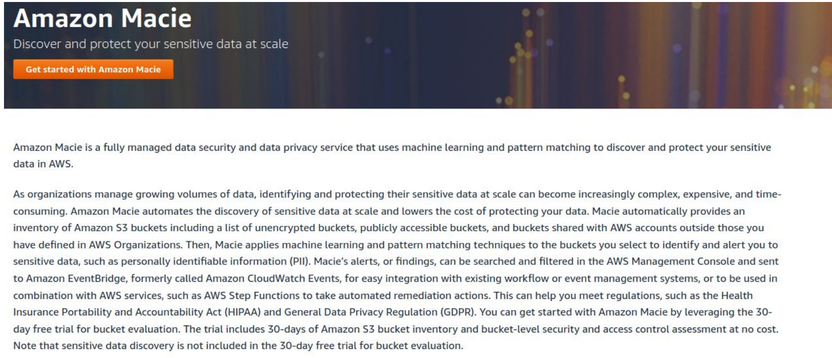


AWS Service Capabilities for Privacy Considerations

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 feature 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.

	Encryption	Deletion	Monitoring of Processing
Alexa for Business	✓	✓	✓
Amazon API Gateway	✓	✓	✓
Amazon AppStream 2.0	✓	✓	✓
Amazon Athena	✓	✓	✓
Amazon Chime	✓	✓	✓
Amazon CloudFront	✓	✓	✓
Amazon CloudSearch		✓	✓
Amazon CloudWatch	✓	✓	✓
Amazon Cognito	✓	✓	✓
Amazon Comprehend	✓	✓	
Amazon Connect	✓	✓	✓
Amazon DynamoDB	✓	✓	✓
Amazon Elastic Block Store (Amazon EBS)	✓	✓	✓
Amazon Elastic Compute Cloud (Amazon EC2)	✓	✓	✓
Amazon Elastic Container Registry (Amazon ECR)	✓	✓	✓



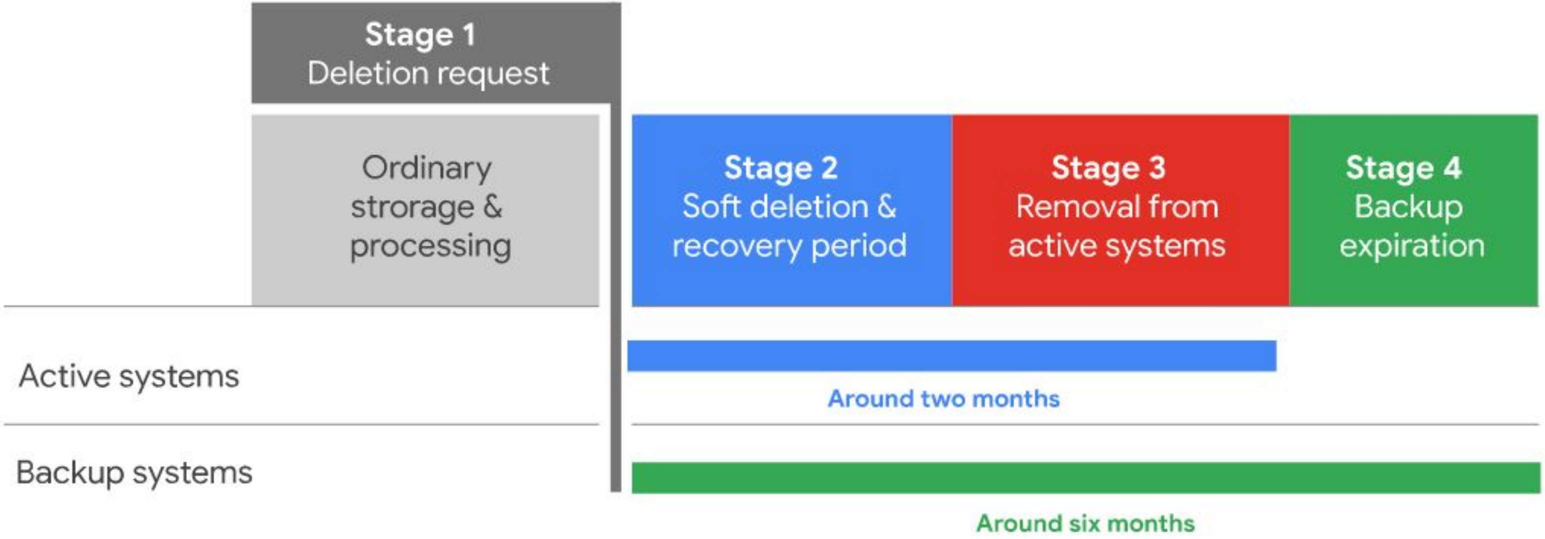
Amazon Macie

Discover and protect your sensitive data at scale

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.

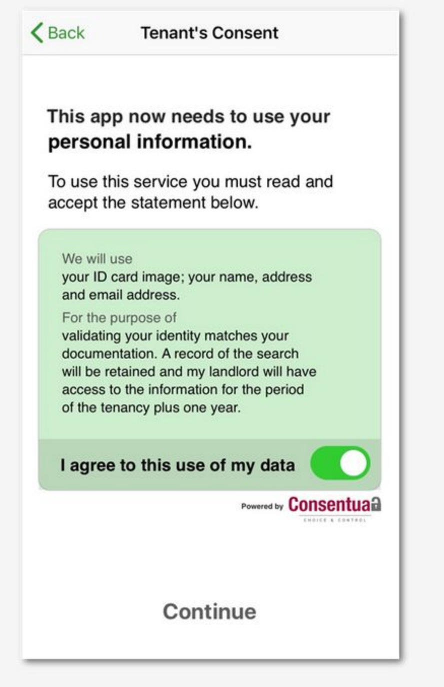
As organizations manage growing volumes of data, identifying and protecting their sensitive data at scale can become increasingly complex, expensive, and time-consuming. Amazon Macie automates the discovery of sensitive data at scale and lowers the cost of protecting your data. Macie automatically provides 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 Regulation (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



https://cloud.google.com/security/deletion

Example of Compliance? Consentua



This app now needs to use your personal information.

To use this service you must read and accept the statement below.

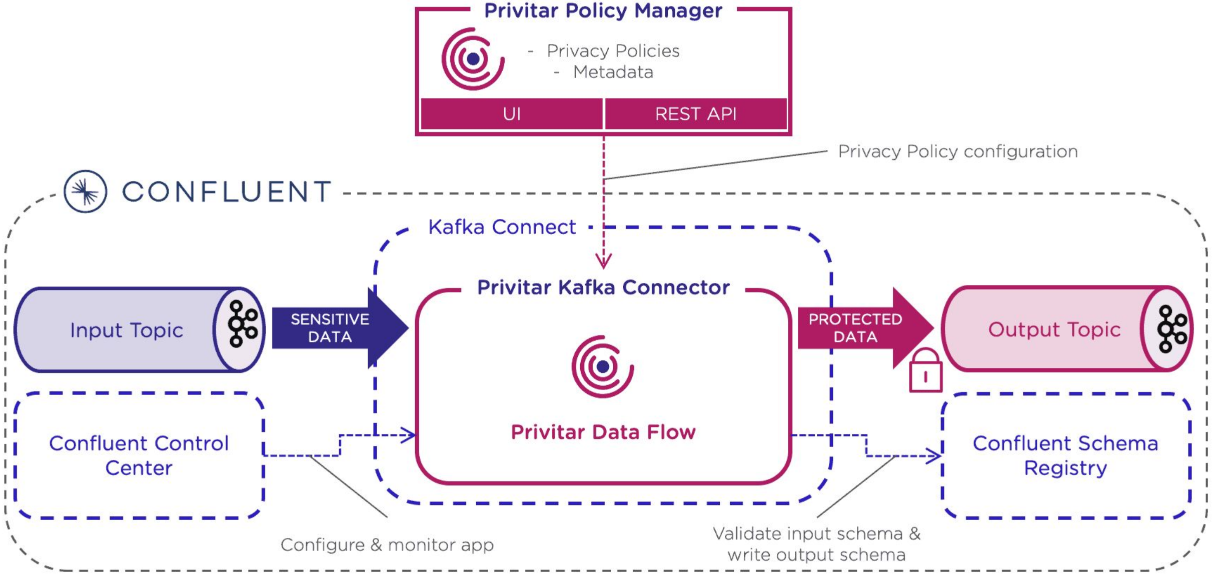
We will use
your ID card image; your name, address and email address.

For the purpose of
validating your identity matches your documentation. A record of the search will be retained and my landlord will have access to the information for the period of the tenancy plus one year.

I agree to this use of my data ●

https://consentua.com/

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

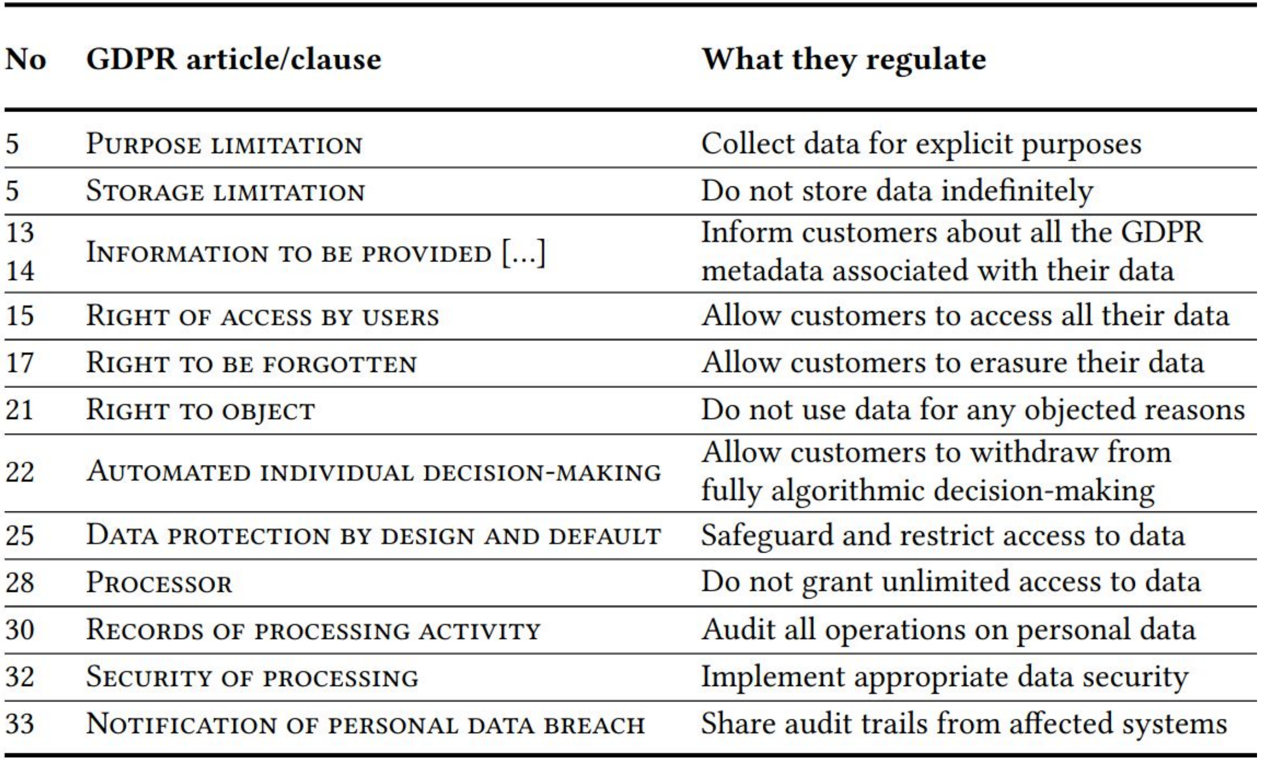
Art. 30 (Records of Processing Activity) and Art. 33 (Notification of Data Breach)

GDPR-compliant data store should have support for

- Associating an **audit trail** with data
- **Monitoring/logging** all data accesses

Slide Source

Remember the Articles?



No	GDPR article/clause	What they regulate
5	Purpose limitation	Collect data for explicit purposes
5	Storage limitation	Do not store data indefinitely
13 14	Information to be provided [...]	Inform customers about all the GDPR metadata associated with their data
15	Right of access by users	Allow customers to access all their data
17	Right to be forgotten	Allow customers to erasure their data
21	Right to object	Do not use data for any objected reasons
22	Automated individual decision-making	Allow customers to withdraw from fully algorithmic decision-making
25	Data protection by design and default	Safeguard and restrict access to data
28	Processor	Do not grant unlimited access to data
30	Records of processing activity	Audit all operations on personal data
32	Security of processing	Implement appropriate data security
33	Notification of personal data breach	Share audit trails from affected systems

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

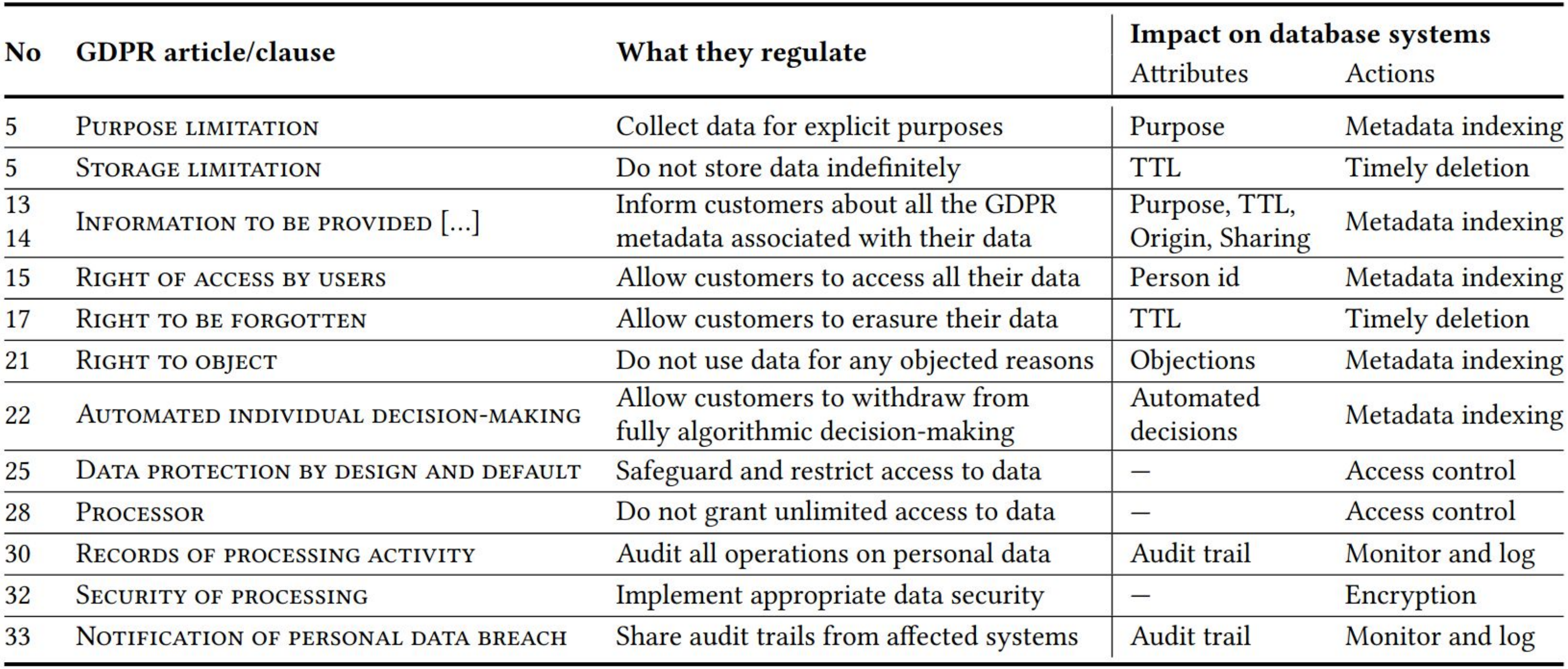
Characterizing Personal Data

- 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



No	GDPR article/clause	What they regulate	Impact on database systems	
			Attributes	Actions
5	Purpose limitation	Collect data for explicit purposes	Purpose	Metadata indexing
5	Storage limitation	Do not store data indefinitely	TTL	Timely deletion
13 14	Information to be provided [...]	Inform customers about all the GDPR metadata associated with their data	Purpose, TTL, Origin, Sharing	Metadata indexing
15	Right of access by users	Allow customers to access all their data	Person id	Metadata indexing
17	Right to be forgotten	Allow customers to erasure their data	TTL	Timely deletion
21	Right to object	Do not use data for any objected reasons	Objections	Metadata indexing
22	Automated individual decision-making	Allow customers to withdraw from fully algorithmic decision-making	Automated decisions	Metadata indexing
25	Data protection by design and default	Safeguard and restrict access to data	—	Access control
28	Processor	Do not grant unlimited access to data	—	Access control
30	Records of processing activity	Audit all operations on personal data	Audit trail	Monitor and log
32	Security of processing	Implement appropriate data security	—	Encryption
33	Notification of personal data breach	Share audit trails from affected systems	Audit trail	Monitor and log

1. Handle metadata explosion 2. Support data protection by design 3. Support GDPR queries

GDPRBench

- 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

Key	Data	Purpose	TTL	User	Objections	Automated Decisions	Third Party Sharing	Originating Source

GDPR Workload

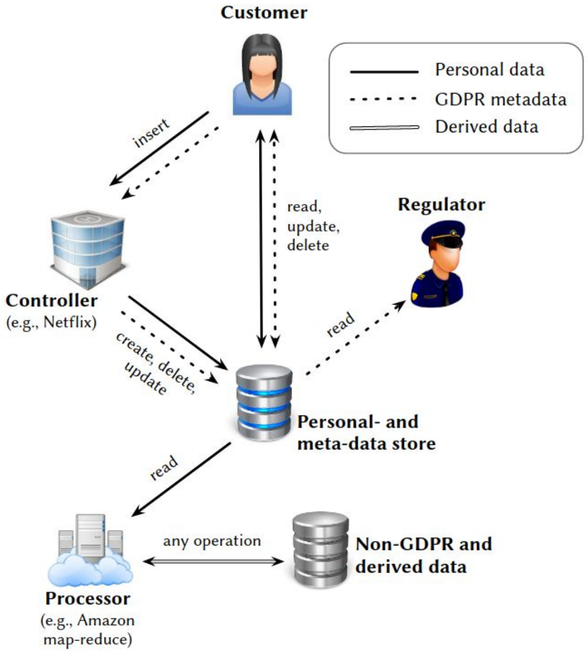
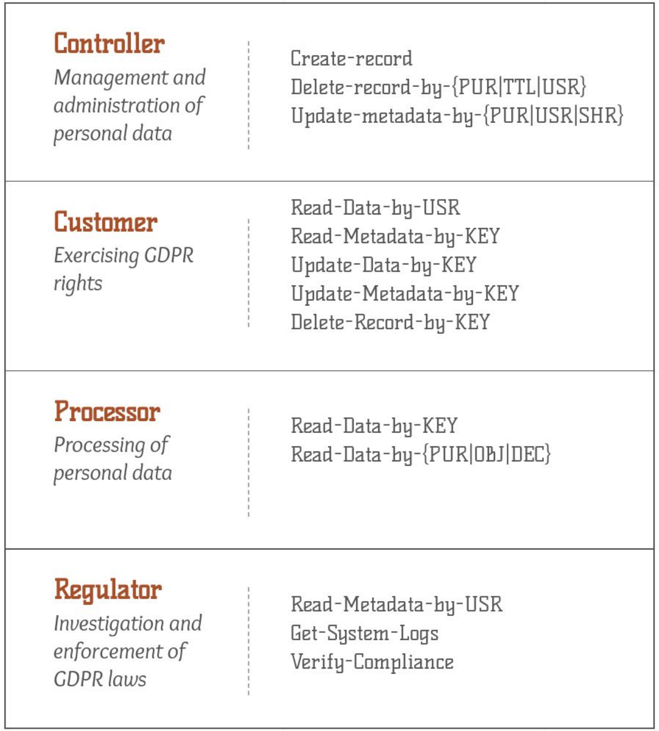


Image Reference



Controller
Management and administration of personal data
	Create-record
	Delete-record-by-{PUR|TTL|USR}
	Update-metadata-by-{PUR|USR|SHR}

Customer *Exercising GDPR rights*	Read-Data-by-USR Read-Metadata-by-KEY Update-Data-by-KEY Update-Metadata-by-KEY Delete-Record-by-KEY
Processor *Processing of personal data*	Read-Data-by-KEY Read-Data-by-{PUR|OBJ|DEC}
Regulator *Investigation and enforcement of GDPR laws*	Read-Metadata-by-USR Get-System-Logs Verify-Compliance

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%)

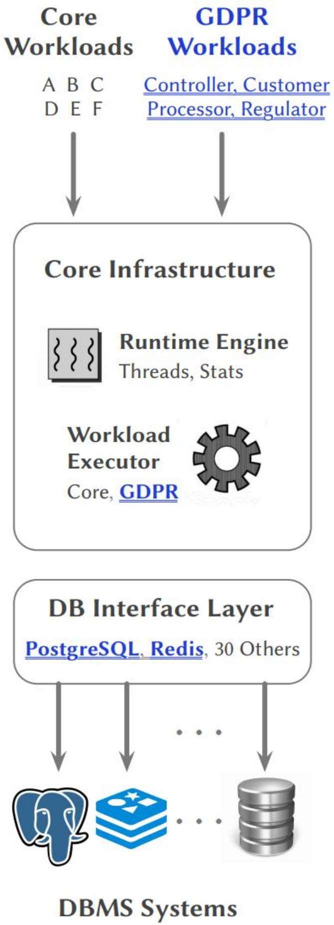
* Based on European's Data Board summary of first 9 months of roll out
* Zipf and uniform distribution

Benchmark Metrics

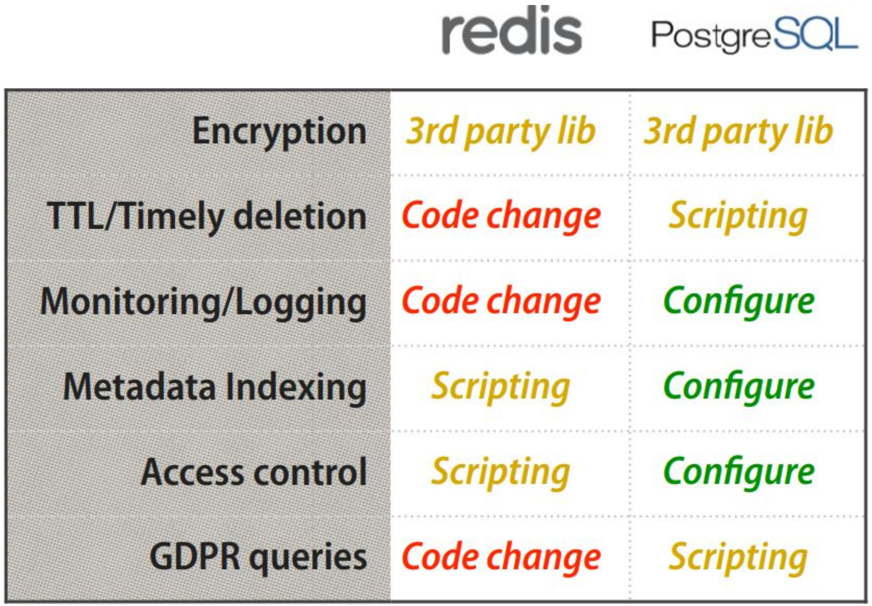
- **Correctness**
  - Validation of metadata-based access control
  - Percentage of query responses that match the results
  - Cumulative across 4 workloads
- **Completion Time**
  - Separately for each workload
  - More important than latency as utility depends upon completion of operation
  - E.g., 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



	redis	PostgreSQL
Encryption	3rd party lib	3rd party lib
TTL/Timely deletion	Code change	Scripting
Monitoring/Logging	Code change	Configure
Metadata Indexing	Scripting	Configure
Access control	Scripting	Configure
GDPR queries	Code change	Scripting

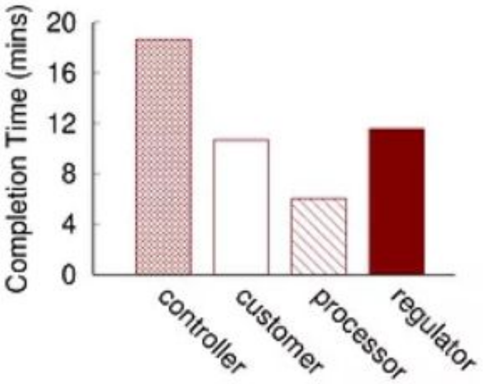
Implementation details

redis	PostgreSQL
LUKS and TLS	LUKS and SSL
Probabilistic algorithm with progressive delay	Modify INSERT queries and periodic checking (1s)
Append-Only-File with code to log all actions	csv-log with row level security policies
None	Secondary indices
External Client	External Client
–	–

Experimental Results (Workloads)

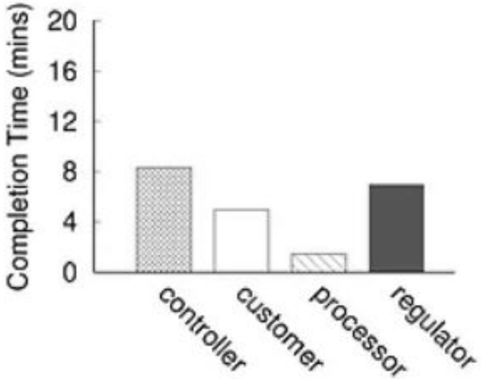
100%
correctness

3.5X
space overhead



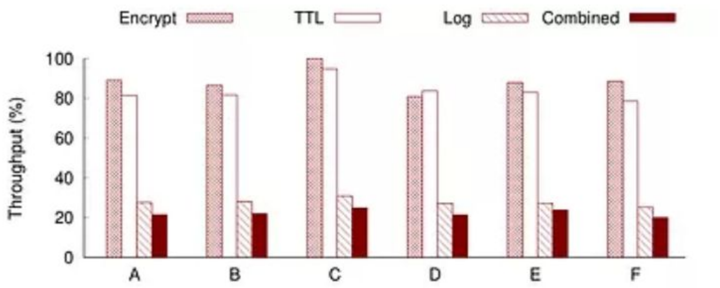
100%
correctness

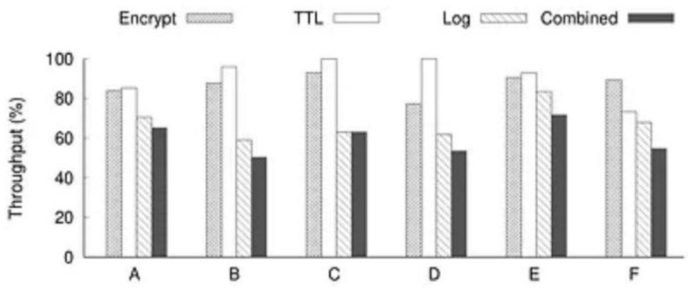
5.95X
space overhead w/
metadata indices



GDPR workloads run faster and scale better on SQL databases due to PostgreSQL's better optimizer and availability of secondary indices

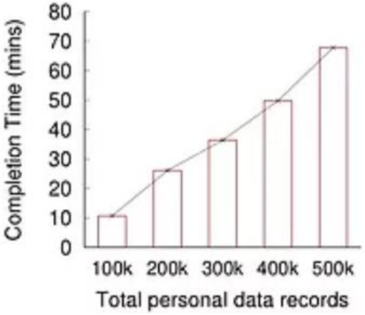
Overhead of security



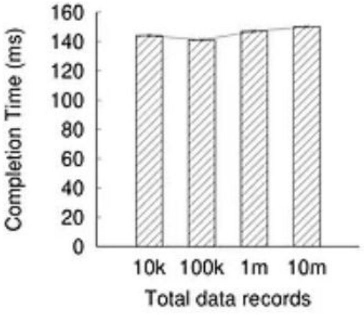


When all features are enabled (solid bar), Redis experiences an overhead of 5×, compare to PostgreSQL's 2× due to significant logging overhead (70% v/s 30%)

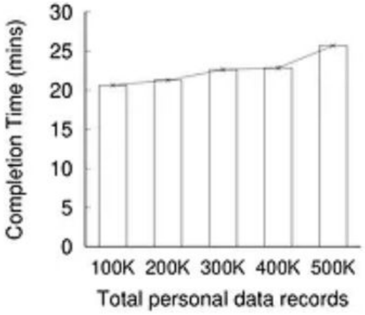
Experimental Results (Effect of Scale)



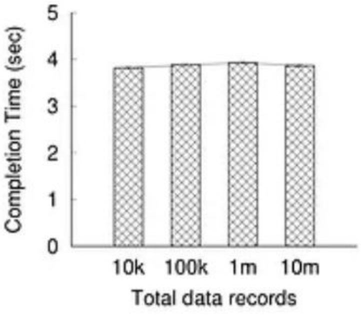
GDPR Customer workload



YCSB workload-C



GDPR Customer workload



YCSB workload-C

- 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

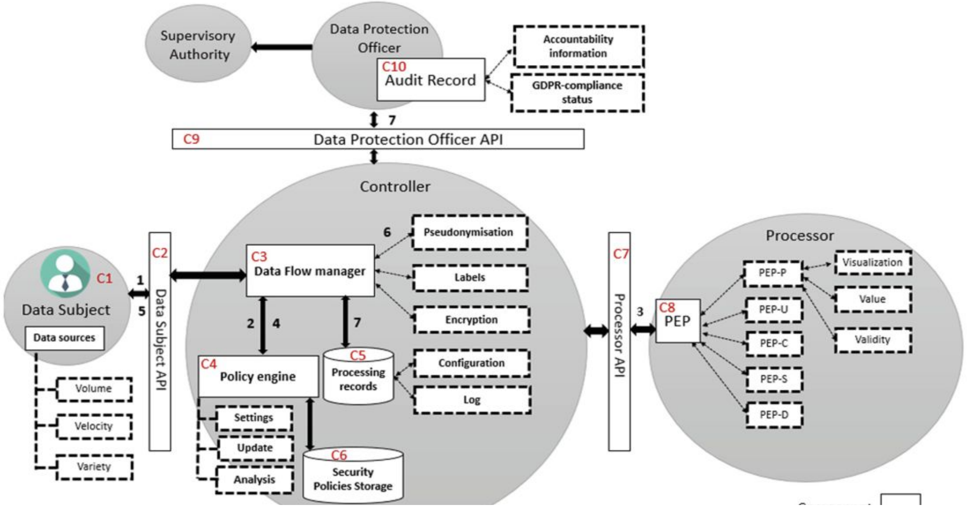
Strengths and Weaknesses

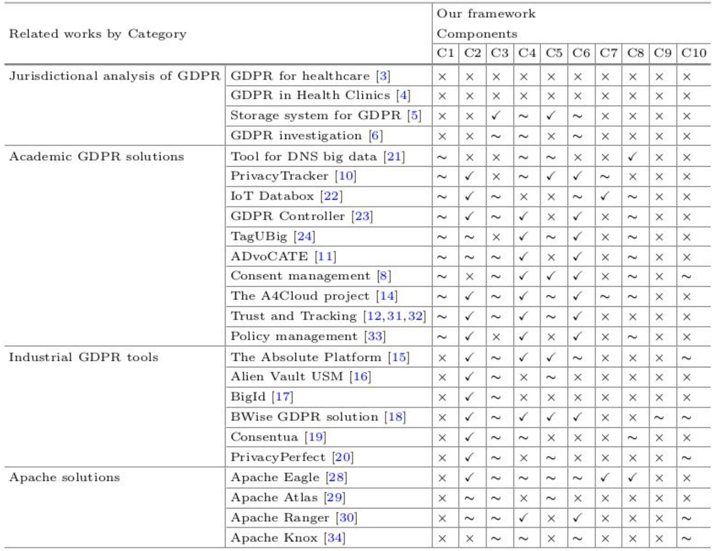
- Through analysis of GDPR Articles
- First characterization of GDPR workload for different roles
- Mapping from legalese to Database System level requirements

- 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)**



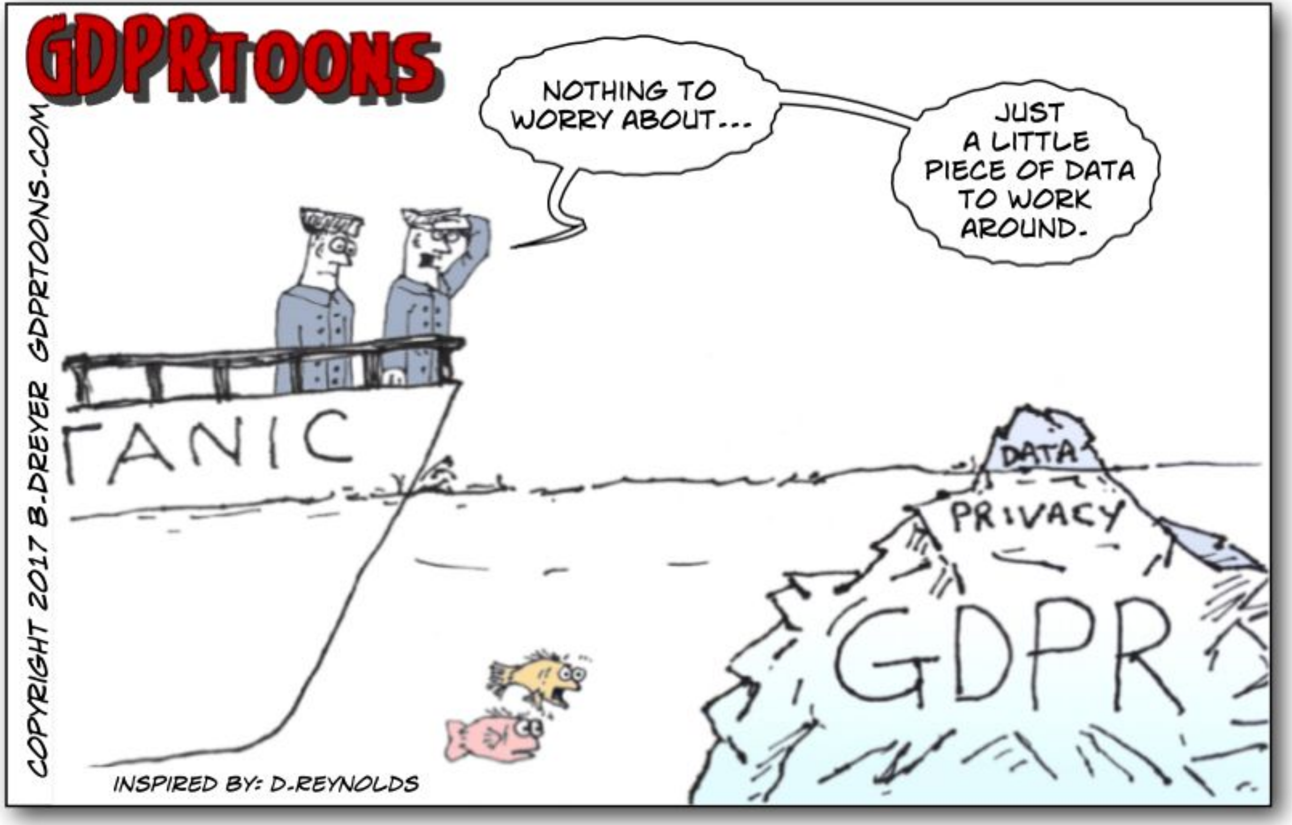


Related works by Category		Our framework Components									
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Jurisdictional analysis of GDPR	GDPR for healthcare [3]	×	×	×	×	×	×	×	×	×	×
	GDPR in Health Clinics [4]	×	×	×	×	×	×	×	×	×	×
	Storage system for GDPR [5]	×	×	✓	∼	✓	∼	×	×	×	×
	GDPR investigation [6]	×	×	∼	∼	×	∼	×	×	×	×
Academic GDPR solutions	Tool for DNS big data [21]	∼	×	×	∼	∼	×	×	✓	×	×
	PrivacyTracker [10]	∼	✓	×	∼	✓	✓	∼	×	×	×
	IoT Databox [22]	∼	✓	∼	×	×	∼	✓	∼	×	×
	GDPR Controller [23]	∼	✓	∼	✓	×	✓	×	∼	×	×
	TagUBig [24]	∼	∼	×	✓	∼	✓	×	∼	×	×
	ADvoCATE [11]	∼	∼	∼	✓	×	✓	×	∼	×	×
	Consent management [8]	∼	×	∼	✓	✓	✓	×	∼	×	∼
	The A4Cloud project [14]	∼	✓	∼	✓	∼	✓	∼	∼	×	×
	Trust and Tracking [12,31,32]	∼	✓	∼	✓	∼	✓	×	×	×	×
	Policy management [33]	∼	✓	×	✓	×	✓	×	∼	×	×
Industrial GDPR tools	The Absolute Platform [15]	×	✓	∼	✓	✓	∼	×	×	×	∼
	Alien Vault USM [16]	×	✓	∼	×	∼	×	×	×	×	×
	BigId [17]	×	✓	∼	×	×	×	×	×	×	×
	BWise GDPR solution [18]	×	✓	∼	✓	✓	✓	×	×	∼	∼
	Consentua [19]	×	✓	∼	∼	×	×	×	∼	×	×
	PrivacyPerfect [20]	×	✓	∼	×	∼	×	×	×	×	∼
Apache solutions	Apache Eagle [28]	×	✓	∼	∼	∼	∼	✓	✓	×	×
	Apache Atlas [29]	×	∼	∼	×	∼	×	×	×	×	×
	Apache Ranger [30]	×	∼	∼	✓	×	✓	×	×	×	∼
	Apache Knox [34]	×	×	∼	∼	×	∼	×	×	×	∼

References

1. European Parliament and Council of European Union (2016) *Regulation (EU) 2016/679*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN (Accessed: 15 October 2020)
2. GDPRBench https://www.gdprbench.org/
3. GDPRToons http://www.gdprtoons.com/

Thank you!



GDPRTOONS
COPYRIGHT 2017 B.DREYER GDPRTOONS.COM
NOTHING TO WORRY ABOUT...
JUST A LITTLE PIECE OF DATA TO WORK AROUND.
TANIC
DATA
PRIVACY
GDPR
INSPIRED BY: D.REYNOLDS