

## **WIP: Pods: Privacy Compliant** Scalable Decentralized Data Services

#### Jonas Spenger<sup>12</sup> < jspenger@kth.se>, Paris Carbone<sup>12</sup>, Philipp Haller<sup>2</sup>

<sup>1</sup>RISE Research Institutes of Sweden, Stockholm, Sweden <sup>2</sup>Digital Futures and EECS, KTH Royal Institute of Technology, Stockholm, Sweden

Presented at Poly'21: Polystore systems for heterogeneous data in multiple databases with privacy and security assurances, August 20, 2021



## 1. Problem Scope 2. Pods: Our Approach 3. Open Questions / Research Directions

# Outline

## **1. Problem Scope** 2. Pods: Our Approach 3. Open Questions / Research Directions

# Outline



- Easy to program

data services, our work

# **Privacy Regulations**

#### **GDPR: General Data Protection Regulation** CCPA: California Consumer Privacy Act



#### https://gdpr-info.eu/

	DRC .	skip to content home	accessibility FAQ feedback sitemap login			
	California LEGISLA	TIVE INFORMATION	Quick Search:         Bill Number         AB1 or ab 1 or ABX1-			
me Bill Info	rmation California Law	Publications Other Resource	As My Subscriptions My Favorites			
ornia Law >> >> (	Code Section Group	_				
		Code: Se	elect Code v Section: 1 or 2 or 1001 Search C			
Code Search	Text Search					
		Up^ Add To My Favorites				
DIVISIO PAI TITLE 1.81 55, Sec. 3	<ul> <li>N 3. OBLIGATIONS [1427 - 32</li> <li>₹T 4. OBLIGATIONS ARISING I</li> <li>.5. California Consumer Priva</li> <li>.9</li> </ul>	<ul> <li>(73.16) (Heading of Division 3 amende FROM PARTICULAR TRANSACTION cy Act of 2018 [1798.100 - 1798.199.1</li> </ul>	ed by Stats. 1988, Ch. 160, Sec. 14. ) S [1738 - 3273.16] (Part 4 enacted 1872.) 100] ( Title 1.81.5 added by Stats. 2018, Ch.			
1798.100. informatio collected.	(a) A consumer shall have t on disclose to that consumer	the right to request that a busines: the categories and specific pieces	s that collects a consumer's personal s of personal information the business has			
(b) A busi consumer personal i personal i section.	ness that collects a consume s as to the categories of per- nformation shall be used. A nformation collected for add	er's personal information shall, at sonal information to be collected a business shall not collect additiona litional purposes without providing	or before the point of collection, inform and the purposes for which the categories of al categories of personal information or use the consumer with notice consistent with this			
(c) A busi consumer	ness shall provide the inform request.	nation specified in subdivision (a)	to a consumer only upon receipt of a verifiable			
(d) A busi promptly section. T be in a po this inforr any time, period.	ness that receives a verifiab take steps to disclose and de he information may be delivi- rtable and, to the extent tec nation to another entity with but shall not be required to	le consumer request from a consu eliver, free of charge to the consu- ered by mail or electronically, and chnically feasible, readily useable f nout hindrance. A business may pr provide personal information to a	umer to access personal information shall mer, the personal information required by this if provided electronically, the information shall format that allows the consumer to transmit rovide personal information to a consumer at consumer more than twice in a 12-month			
(e) This s transactio that is no	ection shall not require a bus n, if such information is not t maintained in a manner the	siness to retain any personal infor sold or retained by the business o at would be considered personal ir	mation collected for a single, one-time or to reidentify or otherwise link information nformation.			
(Amended by Proposit	by Stats. 2019, Ch. 757, Sec. 1 tion 24.)	1. (AB 1355) Effective January 1, 2020	0. Superseded on January 1, 2023; see amendment			
<u>1798.100.</u>	General Duties of Businesse	es that Collect Personal Informatio	n			
(a) A busi collection	(a) A business that controls the collection of a consumer's personal information shall, at or before the point of collection, inform consumers of the following:					
(1) The ca information categories with the co notice cor	(1) The categories of personal information to be collected and the purposes for which the categories of personal information are collected or used and whether that information is sold or shared. A business shall not collect additional categories of personal information or use personal information collected for additional purposes that are incompatible with the disclosed purpose for which the personal information was collected without providing the consumer with notice consistent with this section.					
(2) If the	business collects sensitive p	organal information the categorie	s of sensitive personal information to be			

https://leginfo.legislature.ca.gov/faces/codes\_displayText.xhtml? division=3.&part=4.&lawCode=CIV&title=1.81.5



# **Privacy Regulations**

#### **GDPR: General Data Protection Regulation**



#### https://gdpr-info.eu/

#### Chapter 3 Rights of the data subject

	Section 1	-	Transparency and modalities
	Article 12	-	Transparent information, communication and modalities for the exercise of the rights of the data subject
	Section 2	-	Information and access to personal data
	Article 13	-	Information to be provided where personal data are collected from the data subject
	Article 14	-	Information to be provided where personal data have not been obtained from the data subject
	Article 15	-	Right of access by the data subject
	Section 3	_	Rectification and erasure
	Article 16	-	Right to rectification
	Article 17	-	Right to erasure ('right to be forgotten')
	Article 18	-	Right to restriction of processing
	Article 19	-	Notification obligation regarding rectification or erasure of personal data or restriction of processing
	Article 20	-	Right to data portability
	Section 4	-	Right to object and automated individual decision-making
	Article 21	-	Right to object
	Article 22	-	Automated individual decision-making, including profiling
	Section 5	_	Restrictions
	Article 23	_	Restrictions

# **GDPR: Rights of the Data Subject**

### **Right of Access**

### **Right to Erasure**

Grant access to "what"/"how" personal data is being processed
Reply within one month Erase all personal data concerning the data subject
Within one month

Operations/Privacy Requests:

## Access-Request <Paccess, datasubject>



### **Right to Object**

 Object to certain types of processing
 Within one month

Within one month

### **Erasure-Request**

<Perase, datasubject>

## ! Data service should implement these operations

#### **Objection-Request** <Pobject, datasubject, purpose>



## **Consistent Privacy Requests**

### **Ideal Scenario**



#### Pros: Privacy requests are regular operations, ACID => consistent execution



## **Consistent Privacy Requests**

**Pros: Simple,** decentralized storage, distributed data centres, high-performance

erase object

Cons: **Eventual consistency**, relational, coarse-grained



update

### **Real-World Scenario**

Concepts from Schwarzkopf et al. "Position: GDPR compliance by construction" [16, 11]

## **Example of Issues**

#### Eventually consistent dataflow system



! Failure -> rollback failure recovery









write request is propagated

! Erase on all operators atomically









## **Our Approach to Consistent Privacy Requests**

#### Causally consistent reads from materialized views Ο

• Subsequent reads observe same or more updated state

#### • Serializable privacy requests

Ο no concurrent operations on the system

#### • Executing privacy requests

- 0
- UDFs 0
- Fine-grained 0

Effect of privacy request is as if it was executed atomically, with

Materialize and execute the request according to specification



## 1. Problem Scope 2. Pods: Our Approach 3. Open Questions / Research Directions

# Outline

#### Goal:

- Scalable, failure-resilient, high-performance, dataflow composition
- Privacy compliance

#### **High-level:**

- User-shards + dataflow composition [16]

Couple data with policy-metadata ("Data Capsule") [22], fine-grained

### User shards

- User data is ingested in "user shards" [16]
- Persistently logged, replicated
- Emits events on update



#### Pod tasks

- Subscribe to input streams, execute operations, generate output streams
- Snapshotted consistent state is externally queryable



# Pod tasks Subscribe to input streams, execut generate output streams Snapshotted consistent state is ext



### **Causally Consistent Reads from Materialized Views**





Using asynchronous epoch commit from [6], reproduced/adapted









information flow [21, 15]

Event e1 = <a, metadata={Alice}> Event e2 = <b, metadata={Bob }> e1 + e2 = <a+b, metadata={Alice, Bob}>

=> Compute the correct privacy policy of derived data

## **Fine-Grained Information Flow**

Couple data with policy-metadata ("Data Capsule") [22], use fine-grained

## **Pods: Executing privacy requests**

#### **Execution Strategies for Updates:**

- - **1** Differential updates
  - 2 Recompute state from replaying events
  - (Ordering does not affect results)
- Fine-grained metadata [22], user-defined functions =>
  - **3** Apply request on state
  - Pro: always works
  - Con: may "erase" state that cannot be recomputed, e.g. joint state

Static dataflow data dependencies, relational operators [16] =>

## 1. Problem Scope 2. Pods: Our Approach **3. Open Questions / Research Directions**

# Outline

## **Open Questions / Research Directions**

#### **Efficient information flow tracking**

into declassification [14]

#### **Executing privacy requests**

specification may uncover more issues

#### **Consistent integration with external services**

#### A more flexible programming model

(sources and sinks)

• Fine-grained information flow => efficiency challenge for aggregate data; look

• Semantics of privacy requests are unclear [19]; UDFs; implement full

Propagate privacy requests to external services with atomic consistency

Supporting cycles; dynamic deployments; actors; and push-based updates



## References

[16] Schwarzkopf, Malte, et al. "Position: Gdpr compliance by construction." Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Springer, Cham, 2019. 39-53.

[11] Gjengset, Jon, et al. "Noria: dynamic, partially-stateful data-flow for high-performance web applications." 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18). 2018.

[22] Wang, Lun, et al. "Data capsule: A new paradigm for automatic compliance with data privacy regulations." Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Springer, Cham, 2019. 3-23.

[19] Stonebraker, Michael, et al. "Poly'19 Workshop Summary: GDPR." ACM SIGMOD Record 49.3 (2020): 55-58.

[6] Carbone, Paris, et al. "State management in Apache Flink®: consistent stateful distributed stream processing." Proceedings of the VLDB Endowment 10.12 (2017): 1718-1729.

[15] Salvaneschi, Guido, et al. "Language-integrated privacy-aware distributed queries." Proceedings of the ACM on Programming Languages OOPSLA (2019): 1-30.

[21] Volpano, Dennis, Cynthia Irvine, and Geoffrey Smith. "A sound type system for secure flow analysis." Journal of computer security 4.2-3 (1996): 167-187.



## **WIP: Pods: Privacy Compliant** Scalable Decentralized Data Services

#### Jonas Spenger<sup>12</sup> < jspenger@kth.se>, Paris Carbone<sup>12</sup>, Philipp Haller<sup>2</sup>

<sup>1</sup>RISE Research Institutes of Sweden, Stockholm, Sweden <sup>2</sup>Digital Futures and EECS, KTH Royal Institute of Technology, Stockholm, Sweden

Presented at Poly'21: Polystore systems for heterogeneous data in multiple databases with privacy and security assurances, August 20, 2021



## Pod Task

#### Pod task

- Dataflow composition, pods connected via channels
- Separate state and logic => serverless, elastic scaling
- Context handles privacy request "transparently"
- Distributed snapshotting => resilience to failures [6]
- External applications access snapshot consistent state
- Fine-grained information flow





dependencies;