Personal Database Security and Trusted Execution Environments: A Tutorial at the Crossroads

Nicolas Anciaux, Luc Bouganim, Philippe Pucheral, Iulian Sandu Popa, Guillaume Scerri

Personal Database Security and Trusted Execution Environments: A Tutorial at the Crossroads

Nicolas Anciaux
Inria Saclay, U. Versailles Saint-Quentin, Université Paris-Saclay, France
Nicolas.Anciaux@inria.fr

Luc Bouganim
Inria Saclay, U. Versailles Saint-Quentin, Université Paris-Saclay, France
Luc.Bouganim@inria.fr

Philippe Pucheral
U. Versailles Saint-Quentin, Inria Saclay, Université Paris-Saclay, France
Philippe.Pucheral@uvsq.fr

Iulian Sandu Popa
U. Versailles Saint-Quentin, Inria Saclay, Université Paris-Saclay, France
iulian.sandu-popa@uvsq.fr

Guillaume Scerri
U. Versailles Saint-Quentin, Inria Saclay, Université Paris-Saclay, France
guillaume.scerri@uvsq.fr

ABSTRACT

Smart disclosure initiatives and new regulations such as GDPR in the EU increase the interest for Personal Data Management Systems (PDMS) being provided to individuals to preserve their entire digital life. Consequently, the thorny issue of data security becomes more and more prominent, but highly differs from traditional privacy issues in outsourced corporate databases. Concurrently, the emergence of Trusted Execution Environments (TEE) changes the game in privacy-preserving data management with novel security models. This tutorial offers a global perspective of the current state of work at the confluence of these two rapidly growing areas. The goal is threefold: (1) review and categorize PDMS solutions and identify existing privacy threats and countermeasures; (2) review new security models capitalizing on TEEs and related privacy-preserving data management solutions relevant to the personal context; (3) discuss new challenges at the intersection of PDMS security and TEE-based data management.

1. INTRODUCTION

As Tim Berners Lee himself advocates [40], time has come *"to restore the power and agency of individuals on the web*". Smart disclosure initiatives (e.g., Blue Button [49] in the US, MiData [50] in UK, MesInfos [48] in France) and new privacy-protection regulations (e.g., GDPR in Europe [15]) allow individuals to freely retrieve their personal data from companies and administrations hosting them. Hence, individuals can now gather their full digital environment in a Personal Data Management Systems (PDMS) [4], also called Personal Cloud or Personal Information Management System [1], Personal Data Server [2] or Personal Data Store [12]. A PDMS not only stores data from many (previously) isolated information silos (e.g., secondary copies of data issued by their bank, employer, supermarket) but also primary data (e.g., produced by quantified-self devices and smart meters, photos or documents). This unprecedented concentration of personal data opens the way for new value-added services when crossing multiple data of a given person (e.g., crossing bank statements with shopping history) or crossing data of multiple individuals (e.g., conducting an epidemiological study), under the concerned individuals' control. These perspectives should not eclipse the security issues raised by the PDMS paradigm given the sensitivity and quantity of managed personal data. Several products (e.g., [39–50]) and research initiatives on PDMS (e.g., [1, 2, 3, 12, 17, 25, 37]) are riding this wave. While PDMS have been studied and developed for more than a decade, the proposed solutions provide diverse sets of functionalities and consider diverse threat models.

An important point is that in a PDMS-like context the threat models differ significantly from the traditional corporate DBMS context. In particular, one must consider partially compromised user systems and a large number of individuals collaborating for performing distributed computations, not all of them honest. Moreover, one cannot rely on security conscious, expert administrators to set up the system and ensure that only trustworthy computations are performed. While using secure hardware to secure databases is hardly a new direction [7, 9], most approaches do not tackle such threats. However, with the recent democratization of TEEs, several works have focused on new security models where trust lies on the hardware rather than on the user environment [13, 29, 34] and solutions [16, 28] have been designed for secure data management against new threat models (e.g., distributing trust between many computation nodes and considering corrupted user environments). Although most of these solutions do not focus on personal data management, the proposed trust models and associated solution represent an important step towards suitable threat models for the PDMS.

This tutorial is at the crossroads of personal database security and TEE-based data management. It first reviews and categorizes the various PDMS alternatives in terms of provided functionalities, targeted threat models, trust models and security countermeasures (see Figure 1, Part 1). It then presents existing approaches for secure (distributed) data management focusing on TEE-based solutions (Part 2). Conclusions drawn from Parts 1 and 2 allow us to define an abstract architecture for an *extensive* and *secure* PDMS, and sketch important open research issues.

2. DETAILED TUTORIAL OUTLINE

This tutorial consists of three parts having similar length detailed in the next subsections.

2.1 Personal Data Management Systems

In the first part of the tutorial, we review, compare and categorize, academic PDMS proposals [1, 2, 3, 12, 17, 19, 25, 32, 33, 37] and industrial products representative of the current PDMS offer (CozyCloud [39], Inrupt [40], MyDex [41], Digi.me [42], Meeco [43], BitsAbout.Me [44], Perkeep [45], CloudLocker [46], MyCloud [47]). We also consider products targeting data storage and synchronization for personal applications like SpiderOak [11] which are related to PDMS. This review is driven by the analysis of provided *functionalities* with respect to the targeted *trust and security* models. Using this analysis, we classify the solutions according to their architecture (and hence threat models and security solutions) in five categories covering different functionalities with sometimes irreconcilable security approaches: *online PDMS* (e.g., [39, 43]), *zero-knowledge* solutions (e.g., [11, 41]), *home-cloud software* (e.g., [12, 17, 37]), *home cloud plugs* (e.g., [46, 47]) and *tamper-resistant* versions (e.g., [22]).

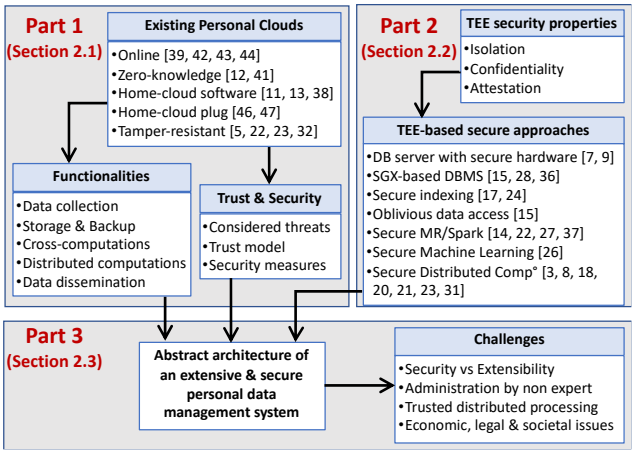

Figure 1: Organization of the Tutorial

Interesting conclusions can be drawn from this state-of-the-art analysis. Regarding the *functionality* aspect, **the whole life-cycle of personal data is targeted** overall. However, taken individually, the solutions only tackle some stages of the life-cycle: *data collection*, *storage*, *backup*, *cross-computations* or *data dissemination*. In particular, **distributed computations functionalities are only supported by very few proposals** even though such functionalities pave the way for Big Data computations with many applications in this context (recommendations, participative studies, training a neural network in patient communities, etc.). Notably, works addressing this step investigate solutions based on privacy preserving home cloud [28, 37] where data is under users' control at the edge of the network.

On the *trust and security* side, a first conclusion is that, **all the privacy threats addressed in the state-of-the-art solutions make sense in the PDMS context** to protect user's privacy and security in a meaningful way, from data snooping and secondary uses performed by cloud providers (e.g., data monetization) to corrupted applications or client devices (e.g., ransomware) to cite only a few. However, these threats are unequally covered by the considered PDMS architectures. More, a second (negative) conclusion is that **unifying the proposed countermeasures does not lead to a secure PDMS architecture**, mainly because building the union of the proposals would undeniably face irreconcilable architectural choices, with different and sometimes contradictory security measures and functionalities depending on the considered threat model. For example, combining a zero-knowledge encrypted storage with an online PDMS offering data-oriented computation functionalities would require returning all the individual's data to the client side and hence put it at risk.

To conclude this part, we derive the *extensible* set of functionalities to be implemented in a PDMS to cover the complete data life-cycle and list the privacy threats the PDMS must circumvent to be deemed *secure*. For each functionality, we identify its main specificities and highlight the way it differs from corporate data management systems from a security point of view.

2.2 TEE-based Data Management

A majority of the works on secure database computations focus on outsourcing corporate databases to honest-but-curious cloud services to manage large sets of sensitive data. In this part of the tutorial, we first state the problem of extensible and secure PDMS as a 'mutual trust' problem (i.e., where data and query confidentiality and integrity should be guaranteed to both the individuals managing their own data and third parties accessing shared data or query results) and position the main privacy-preserving database technologies in relation to this issue.

We then briefly introduce the TEE technology and the three main security properties it provides: (1) *isolation*: TEEs isolate the environment (including the OS) for the code it runs, which means that an attacker controlling the environment (even with root access) cannot influence the behavior of code executed within the TEE; (2) *confidentiality*: TEEs provide confidentiality against a compromised user/OS, ensuring that private data handled by the TEE may never be observed, except through explicit input/output of the code running inside a TEE; and (3) *attestation*: TEEs provide an attestation mechanism that allows the code running inside a TEE to prove its identity and provide guarantees that the result it produced was indeed obtained with that code. While TEEs aim at providing unconditional protection to the executed programs, a motivated attacker may still be able to perform so called side-channel attacks to gain information on the data processed inside TEEs [30, 36], leading to consider counter measures to protect –to some extent– against such attacks.

In this context, the first crucial threat for personal data management is a potentially compromised user system or mismanagement of the said system by a non-expert user. To this end, we investigate the TEE based existing solutions for securing a DBMS, as protecting a user 'from himself' is closely related to protecting a system against a corrupt system administrator. We first review the database computing techniques deploying secure hardware at the database server side [6, 7, 9]. Most techniques basically split the query processing in one part executed directly on the encrypted data and the other executed inside the secure hardware on cleartext data and make the processing oblivious to prevent adversary learning anything from the data access pattern. Besides, we analyze recent Intel SGX-based database initiatives, with a focus on works where the security relies on a unique DBMS controller running in a single SGX enclave. We start from the implementation of very simple DBMS stores in SGX such as a key-value store [31], and then review proposal for performing advanced database operations in SGX, from secure indexing like HardIDX and Oblix [16, 24], to proposal dedicated to protect data access patterns analysis like ObliDB [14], up to the execution of an entire DBMS engine in SGX like EnclaveDB [28].

The second crucial issue is enabling mutually trusted data management between PDMS users and third parties in a distributed database setting. Emerging TEE-based solutions address the problem of outsourcing distributed computations in the cloud, leveraging TEEs for integrity and privacy guarantees. For example, [13, 27, 29] offer secure map-reduce frameworks using Intel's SGX, [26] proposes a machine learning framework and [38] offers a Spark SQL framework. These works do not tackle the problem of distributing trust between users, but they (at least partially) address the problem of obtaining integrity and security of computations over multiple TEEs, and can be viewed as a necessary step in this direction. We then focus on an initial line of work [8, 18] which consists in leveraging TEEs for distributing trust between actors during a distributed computation.

2.3 Reference Architecture and Challenges

Finally, we draw conclusions from the first two parts and show how one could leverage existing TEE-based solutions at various levels to address the specific threats associated with personal data management. We then derive a reference architecture for a TEE based PDMS. A first important challenge linked to the design of such a PDMS architecture is rooted in the tension between PDMS security, rich data processing and mutual trust. Few recent works go into this direction. [4] envisions a minimal Secure Core implementing a basic set of operations, extended with richer but untrusted data processing operators implemented in TEEs. We further illustrate the database challenges in the case of collective or distributed queries and focus on the problem of concretely implementing distributed data processing without leaks and using TEEs to propagate trust between a large number of users contributing to a common processing task. We review here a few initial proposals applied to specific privacy preserving database computations with dataflow control [3, 20, 21, 23, 32].

Several other challenges relate to the control tools enabling non-expert users to effectively regulate the dissemination and usage of their data. While preliminary works exist regarding access control enforcement with TEEs [10, 35], much is left to do to tackle usage control. Finally, the reciprocal entanglements between economic, legal, societal and technological aspects of personal data management also constitute major concerns.

3. TUTORIAL OUTCOME

This tutorial is devoted to a large audience, from industrials involved in personal data management, to researchers and computer scientists working on data management and data security issues. We expect the audience to gain a better perception of the fundamental security properties needed at each step of the personal data life-cycle, helping to identify hidden flaws linked to architectural choices. The public can also get a broad view of the PDMS landscape and obtain an insight of the main research challenges linked to secure personal data management and the impact of recent TEE research on these challenges. Finally, we hope this tutorial will help the audience perceive how architectural choices in the personal data domain disrupt established principles of accountability between platform providers, data controllers and individuals.

4. PRESENTERS BIOGRAPHIES

All presenters are members of PETRUS (Personal and Trusted Cloud) group at Inria and UVSQ. PETRUS conducts research on secure personal cloud architectures, privacy preserving administration models and enforcement, global distributed processing and economic, legal and societal issues of the personal cloud. PETRUS has recently launched an Inria Innovation Lab named OwnCare, which aims at building a TEE-based secure medical-social personal cloud facilitating the coordination of home care for elderly people. This secure personal cloud is being deployed over 10.000 patients in the Yvelines district in France.

Nicolas Anciaux is a research director at Inria Saclay-Ile de France and heads the PETRUS team. His main research interest is in secure database processing using trusted hardware, data sharing models and large-scale distributed processing on personal data.

Luc Bouganim is a research director at Inria Saclay-Ile de France and technical coordinator of the OwnCare Inria Innovation Lab. His research interests are on secure (personal) data management using TEEs, decentralized query execution and modern storage.

Philippe Pucheral is full Professor at UVSQ and head of the OwnCare Inria Innovation Lab. His main research interest is on data management embedded in secure hardware and TEE-based decentralized querying protocols.

Iulian Sandu Popa is an Associate Professor at UVSQ. His research interests include secure distributed data management, spatiotemporal databases and mobile data management.

Guillaume Scerri is an Associate Professor at UVSQ. His research interests include provable security of systems based on TEEs and distributed cryptographic protocols.

5. ACKNOWLEDGEMENTS

This research is partially supported by the ANR PerSoCloud grant no ANR-16-CE39-0014.

6. REFERENCES

[1] S. Abiteboul, and A. Marian. Personal information management systems. *EDBT Tutorial*, 2015.

[2] T. Allard, N. Anciaux, L. Bouganim, Y. Guo, L. L. Folgoc, B. Nguyen, P. Pucheral, I. Ray, and S. Yin. Secure personal data servers: a vision paper. *PVLDB*, 3(1), 25-35, 2010.

[3] N. Anciaux, P. Bonnet, L. Bouganim, B. Nguyen, I. S. Popa, and P. Pucheral. Trusted cells: A sea change for personal data services. *CIDR*, 2013.

[4] N. Anciaux, P. Bonnet, L. Bouganim, B. Nguyen, P. Pucheral, I. S. Popa, and G. Scerri. Personal data management systems: The security and functionality standpoint. *Inf. Syst.*, 80:13–35, 2019.

[5] N. Anciaux, L. Bouganim, P. Pucheral, Y. Guo, L. L. Folgoc, and S. Yin. Milo-db: a personal, secure and portable database machine. *Distributed and Parallel Databases*, 32(1):37–63, 2014.

[6] A. Arasu, K. Eguro, M. Joglekar, R. Kaushik, D. Kossmann, and R. Ramamurthy. Transaction processing on confidential data using cipherbase. In *ICDE*, 435–446, 2015.

[7] A. Arasu, K. Eguro, R. Kaushik, and R. Ramamurthy. Querying encrypted data. In *SIGMOD Conference*, 1259–1261, 2014.

[8] S. Arnautov, B. Trach, F. Gregor, T. Knauth, A. Martin, C. Priebe, J. Lind, D. Muthukumaran, D. O'Keeffe, M. Stillwell, D. Goltzsche, D. M. Eyers, R. Kapitza, P. R. Pietzuch, and C. Fetzer. SCONE: secure linux containers with intel SGX. In *OSDI*, 689–703, 2016.

[9] S. Bajaj and R. Sion. TrustedDB: A trusted hardware-based database with privacy and data confidentiality. *IEEE Trans. Knowl. Data Eng.*, 26(3):752–765, 2014.

[10] E. Birrell, A. T. Gjerdrum, R. van Renesse, H. D. Johansen, D. Johansen, and F. B. Schneider. SGX enforcement of use-based privacy. In *WPES@CCS*, 155–167, 2018.

[11] A. P. K. Dalskov and C. Orlandi. Can you trust your encrypted cloud? An assessment of spideroakone's security. In *AsiaCCS*, 343–355, 2018.

[12] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland. OpenPDS: Protecting the privacy of metadata through safeanswers. *PloS one*, 9(7):1–9, 2014.

[13] T. T. A. Dinh, P. Saxena, E. Chang, B. C. Ooi, and C. Zhang. M2R: enabling stronger privacy in mapreduce computation. In *USENIX Security Symp.*, 447–462, 2015.

[14] S. Eskandarian and M. Zaharia. An oblivious general-purpose SQL database for the cloud. *CoRR*, abs/1710.00458, 2017.

[15] European Parliament. General Data Protection Regulation. Law. (27 April 2016).

[16] B. Fuhry, R. Bahmani, F. Brasser, F. Hahn, F. Kerschbaum, and A. Sadeghi. Hardidx: Practical and secure index with SGX in a malicious environment. *Journal of Computer Security*, 26(5):677–706, 2018.

[17] H. Haddadi, H. Howard, A. Chaudhry, J. Crowcroft, A. Madhavapeddy, and R. Mortier. Personal data: thinking inside the box. *Aarhus conf. on critical alternatives*, 2015.

[18] T. Hunt, Z. Zhu, Y. Xu, S. Peter, and E. Witchel. Ryoan: A distributed sandbox for untrusted computation on secret data. *ACM Trans. Comput. Syst.*, 35(4):13:1–13:32, 2018.

[19] D. Koll, D. Lechler, and X. Fu. Socialgate: Managing large-scale social data on home gateways. In *ICNP*, 1–6, 2017.

[20] R. Ladjel, N. Anciaux, P. Pucheral and G. Scerri. Trustworthy Distributed Computations on Personal Data Using Trusted Execution Environments. *TrustCom*, 2019.

[21] R. Ladjel, N. Anciaux, P. Pucheral and G. Scerri. A manifest-based framework for organizing the management of personal data at the edge of the network. *ISD*, 2019.

[22] S. Lallali, N. Anciaux, I. S. Popa, and P. Pucheral. Supporting secure keyword search in the personal cloud. *Inf. Syst.*, 72:1–26, 2017.

[23] J. Loudet, I. S. Popa, and L. Bouganim. SEP2P: secure and efficient P2P personal data processing. In *EDBT*, 145–156, 2019.

[24] P. Mishra, R. Poddar, J. Chen, A. Chiesa, and R. A. Popa. Oblix: An efficient oblivious search index. In *S&P*, 279–296, 2018.

[25] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, and J. Crowcroft. Human-data interaction: The human face of the data-driven society. *CoRR*, abs/1412.6159, 2014.

[26] O. Ohrimenko, F. Schuster, C. Fournet, A. Mehta, S. Nowozin, K. Vaswani, and M. Costa. Oblivious multi-party machine learning on trusted processors. In *USENIX Security Symposium*, 619–636, 2016.

[27] R. Pires, D. Gavril, P. Felber, E. Onica, and M. Pasin. A lightweight mapreduce framework for secure processing with SGX. In *CCGrid*, 1100–1107, 2017.

[28] C. Priebe, K. Vaswani, and M. Costa. EnclaveDB: A secure database using SGX. In *S&P*, 264–278, 2018.

[29] F. Schuster, M. Costa, C. Fournet, C. Gkantsidis, M. Peinado, G. Mainar-Ruiz, and M. Russinovich. VC3: trustworthy data analytics in the cloud using SGX. In *S&P*, 38–54, 2015.

[30] M. Shih, S. Lee, T. Kim, and M. Peinado. T-SGX: eradicating controlled-channel attacks against enclave programs. In *NDSS*, 2017.

[31] Y. Tang, J. Chen, K. Li, J. Xu, and Q. Zhang. Authenticated key-value stores with hardware enclaves. *CoRR*, abs/1904.12068, 2019.

[32] D. H. T. That, I. S. Popa, K. Zeitouni, and C. Borcea. PAMPAS: privacy-aware mobile participatory sensing using secure probes. In *SSDBM*, 4:1–4:12, 2016.

[33] Q. To, B. Nguyen, and P. Pucheral. Private and scalable execution of SQL aggregates on a secure decentralized architecture. *ACM TODS.*, 41(3):16:1–16:43, 2016.

[34] F. Tramèr, F. Zhang, H. Lin, J. Hubaux, A. Juels, and E. Shi. Sealed-glass proofs: Using transparent enclaves to prove and sell knowledge. In *EuroS&P*, 19–34, 2017.

[35] P. Tran-Van, N. Anciaux, and P. Pucheral. SWYSWYK: A privacy-by-design paradigm for personal information management systems. *ISD*, 2017.

[36] W. Wang, G. Chen, X. Pan, Y. Zhang, X. Wang, V. Bindschaedler, H. Tang, and C. A. Gunter. Leaky cauldron on the dark land: Understanding memory side-channel hazards in SGX. In *CCS*, 2421–2434, 2017.

[37] J. R. Zhao, R. Mortier, J. Crowcroft, and L. Wang. Privacy-preserving machine learning based data analytics on edge devices. In *AIES*, 341–346,2018.

[38] W. Zheng, A. Dave, J. G. Beekman, R. A. Popa, J. E. Gonzalez, and I. Stoica. Opaque: An oblivious and encrypted distributed analytics platform. In *NSDI*, 283–298, 2017.

[39] Cozy Cloud. Your digital home. https://cozy.io/en

[40] Inrupt.com. https://tinyurl.com/TBL-inrupt.

[41] Mydex. https://mydex.org

[42] Digi.me Private Sharing. https://digi.me/

[43] Meeco.me. https://meeco.me

[44] Bitsabout.me. https://bitsabout.me

[45] Perkeep.org. https://perkeep.org

[46] CloudLocker. www.cloudlocker.eu

[47] MyCloud. My Cloud Home. https://mycloud.com

[48] Fing. The MesInfos project. mesinfos.fing.org/english

[49] Blue Button. Find Your Health Data. www.healthit.gov/topic/health-it-initiatives/blue-button

[50] MiData. The midata project. https://www.midata.coop