Protection and control of personal identifiable information: The PoSeID-on approach

Received: 30th of September, 2019

Rui Casaleiro
is an MSc student enrolled in the master's programme in software engineering at the University of Coimbra, where he works as a researcher for the Center for Informatics and Systems of University of Coimbra (CISUC). He holds a BSc in computer science. His main research interests are blockchain technologies and machine learning, with a focus on outlier detection.

University of Coimbra, Coimbra, Portugal
Tel: +351 239 790 000; E-mail: rjfc@dei.uc.pt

Paulo Silva
is a PhD student enrolled in the doctoral programme in information science and technology at the University of Coimbra, where he works as a researcher for the Center for Informatics and Systems of the University of Coimbra (CISUC). He holds an MSc in communications, services and infrastructures. His main research interests are data privacy protection and security services for cloud computing.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: pmgsilva@dei.uc.pt

Paulo Simões
is an Assistant Professor at the University of Coimbra. He has more than 15 years of research experience in the fields of network and infrastructure management, security, critical infrastructure protection and virtualisation of networking and computing resources. He has published over 150 publications in refereed journals and conferences. He was founding partner of two technological spin-off companies and regularly leads technology transfer projects for industry partners such as telecommunications operators and energy utilities. He has also been involved in several European research projects with technical and managerial duties.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: psimoes@dei.uc.pt

Fernando Boavida
is Full Professor in the Department of Informatics Engineering (DEI) in the Faculty of Sciences and Technology at the University of Coimbra. His main research interests are people-centric internet of things, security, mobility and quality of service. He is a senior member of the IEEE (Institute of Electrical and Electronics Engineers) and a licensed professional engineer. He has published over 200 international papers and is/was member of several programme committees. His webpage can be found at http://www.uc.pt/go/boavida.

University of Coimbra, Coimbra, Portugal
E-mail: boavida@dei.uc.pt

Edmundo Monteiro
is Full Professor at the University of Coimbra, from where he got a PhD in electrical engineering and the habilitation in informatics engineering in 1996 and 2007, respectively. His research interests are computer networks, wireless communications, quality of service and experience, service-oriented infrastructures and security. He is author of over 200 publications in books, journals, book chapters and international conferences. He is also co-author of nine international patents.

University of Coimbra, Coimbra, Portugal
Tel.: : +351 239 790000; E-mail: edmundo@dei.uc.pt

Marília Curado

is an Associate Professor with habilitation in the Department of Informatics Engineering at the University of Coimbra, from where she got a PhD in informatics engineering in the subject of quality of service routing in 2005. Her research interests are quality of service, quality of experience, energy efficiency, internet of things, mobility, cloud systems and software-defined networks.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: marilia@dei.uc.pt

Tiago Cruz

has been Auxiliary Professor in the Department of Informatics Engineering at the University of Coimbra since December 2013, from where he obtained his PhD in computer science in 2012. He is a Senior Researcher at the Center for Informatics and Systems of University of Coimbra, having started his research activity in 2001. His research interests cover areas such as management systems for communications infrastructure and services (operator and data centre environments), embedded computing, critical infrastructure security, broadband access network devices and service management.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: tjcruz@dei.uc.pt

Nuno Antunes

is an Assistant Professor in the Department of Informatics Engineering at the University of Coimbra. His research interests include security and dependability assessments of software systems. Nuno received a PhD in information science and technology from the University of Coimbra. He is a member of the IEEE Computer Society.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: nmsa@dei.uc.pt

Marco Vieira

is a Full Professor in the Department of Informatics Engineering at the University of Coimbra. His research interests include dependability and security benchmarking, experimental dependability evaluation, fault injection, software development processes and software quality assurance. Marco received a PhD in computer engineering from the University of Coimbra. He is a member of the IEEE Computer Society.

University of Coimbra, Coimbra, Portugal
Tel.: +351 239 790 000; E-mail: mvieira@dei.uc.pt

Giovanni M. Riccio

is a partner at e-Lex and Professor of comparative media law, copyright law and cultural heritage law at the Universitá di Salerno/DPO.

ELEX — e-Lex Studio Legale, Rome, Italy
Tel: +39 06 8775 0524; E-mail: mverzillo@e-lex.it

Dario Reccia

is an attorney-at-law. He holds a master's degree in 'Diritto e Impresa' from the Business School of 'Il Sole24Ore'. He is Correspondent for the law reviews *Leggi Oggi* and *Medialaws*.

ELEX — e-Lex Studio Legale, Rome, Italy
Tel; +39 06 8775 0524; E-mail: mverzillo@e-lex.it

Maria P. Verzillo
is a partner at e-Lex.

ELEX — e-Lex Studio Legale, Rome, Italy
Tel: +39 06 8775 0524; E-mail: mverzillo@e-lex.it

Philipp Marek
is a Senior Architect and Forensics Specialist at BRZ.

BRZ — Bundesrechenzentrum GmbH, Vienna, Austria
Tel: +43171123883750; E-mail: philipp.marek@brz.gv.at

Laurent Goncalves
is the e-Citiz Research Manager and has an MSc in fundamental computer science. He has 15-plus years of experience in product R&D, engineering and development. Besides, Laurent was Professor at Paul Sabatier University (Toulouse, France), teaching Java enterprise and graphical Java, and at Institute of Technology (Blagnac, France), teaching Java enterprise too.

Softeam, Paris, France
Tel: +33 6 78 87 39 41; E-mail: laurent@softeam.fr

Alessandra Bagnato
is a Research Scientist and the Head of the Research Unit within the Softeam R&D Department. She has served on the technical programme committee of several international events such as the ECMFA or MISE@ICSE. She was also co-organiser of SEC-MDA 2009 and 2010 at the ECMFA, DeCPS at AdaEurope since 2015, and the MeGSuS workshops at the ESEIW since 2015. Her main research interests include software engineering in the context of big data, cyber physical systems design, security and data privacy.

Softeam, Paris, France
Tel.: +33 638 817 652 ; E-mail: alessandra.bagnato@softeam.fr

Alessia Valentini
is Security Manager at Accenture, working as an information security consultant. She has 24 years of professional experience in ICT (information and communications technology) companies and roles. She worked as WP (work packages) leader for the FP7 CyberROAD project and as Dissemination WP leader for the FP7 PREEMPTIVE project. In 2017, she obtained the CISA/ISACA certification for auditing. From 2014 to 2016, she was the Director in Armed Forces Electronic Association (Afcea).

Accenture S.p.A., Rome, Italy
Tel: +39 06 515 9281; E-mail: alessia.valentini@accenture.com

Barbara Intonti
is an Experienced Information Technology System Architect with a demonstrated history of working in the telecommunications industry. She is skilled in Oracle Database, ODSE, Unix/Linux and security governance.

Accenture S.p.A., Rome, Italy
Tel: +39 334 680 4799; E-mail: barbara.intonti@accenture.com

Rita Manzo
is a Security Consulting Analyst at Accenture, specialized in innovation topics.

Accenture S.p.A., Rome, Italy
Tel: +39 348 905 4695; E-mail: rita.manzo@accenture.com

Veronica Della Posta
is a Security Analyst at Accenture, specialized in Business Development, Security Risk Assessment/Analysis/Treatment and Compliance matters as well as Risk mitigation.

Accenture S.p.A., Rome, Italy
Tel: +39 3666638801; E-mail: veronica.della.posta@accenture.com

Livia Zampolini
is a Security Consulting Analyst at Accenture, experienced in Governance, Risk Management, and Compliance matters. She has experience in assessing the security posture of businesses and Public Sector entities while aligning their approaches with regulatory requirements in order to achieve compliance and prepare for market volatility.

Accenture S.p.A., Rome, Italy
Tel: +39 3482341054; E-mail: livia.zampolini@accenture.com

Joris van Rooij
is a lead engineer at Jibe Company, working from Eindhoven, the Netherlands. He specialises in software engineering and architecture with a focus on operational security. He is cofounder of a number of innovative startups and an active participant in civil society.

JIBE — Smartfeedz B.V., Eindhoven, Netherlands
Tel: +31 40 767 6001; E-mail: joris.van.rooij@jibecompany.com

Rick Houf
is cofounder and CTO of JIBE and is a technical/scientific-oriented architect/project manager and is specialised in translating business concepts and requirements into IT solutions, combining common technologies with innovations to find the best match between cost and functionality. He has over 20 years of experience in creating and running software development teams and more than 10 years of experience in running a software business, focusing on web development, big data processing and analytics and delivering cloud solutions.

JIBE — Smartfeedz B.V., Eindhoven, Netherlands
Tel: +31 40 767 6001; E-mail: rick.houf@jibecompany.com

Erkuden Rios
is a senior scientist of the Cybersecurity research team of the TRUSTECH unit in Tecnalia. She is currently the coordinator of the Security WP in the H2020 SPEAR project on Secure Smart Grids as well as in the H2020 ENACT project on Secure and Privacy-aware Smart IoT Systems. Previously, she was the coordinator of the H2020 MUSA project on Multi-cloud Security, successfully ended in 2017, as well as the chair of the Data Protection, Security and Privacy in Cloud Cluster of EU-funded research projects, launched by DG-CNECT in April 2015 (https://eucloudclusters.wordpress.com/

data-protection-security-and-privacy-in-the-cloud/). Furthermore, she has worked in multiple large European and Spanish projects on Cybersecurity and trust such as PDP4E, TACIT, RISC, ANIKETOS, SWEPT, CIPHER and SHIELDS. Erkuden collaborates with Technology Platforms and Forums such as Cybersecurity PPP ECSO, AIOTI WG4 Policy and Privacy and the Spanish Technology Pole on Cybersecurity.

TECN — Fundacion Tecnalia Research & Innovation, Derio, Spain
Tel: +34 664 100 348; E-mail: erkuden.rios@tecnalia.com

Eider Iturbe
is a senior scientist of Cybersecurity research team within Tecnalia. She is specialised in trust and security engineering technologies such as IT and network cyber security, cyber security and risk management on multiple environments (IT systems, Smart grids and IoT systems) and security assessment and monitoring on multiple environments (IT systems, Smart grids and IoT systems) among others. She has broadened her expertise through her participation in different European and Spanish national projects. Eider graduated in Telecommunication Engineering from the University of the Basque Country (Spain) and in the European Master in Project Management at the same university. Before joining Tecnalia in 2009, she worked for software consultancy firms where she acquired management skills and a great technical expertise in the security field.

TECN — Fundacion Tecnalia Research & Innovation, Derio, Spain
Tel: +34 667108806; E-mail: eider.iturbe@tecnalia.com

Iván Gutierrez, MSc
is a software engineer and a Cryptography and Cybersecurity enthusiast. He regularly takes part in R&D projects to apply his experience on the distributed ledger technologies (DLTs) for the team that coordinates the Blockchain Innovation Center by Tecnalia. Iván has gained technical skills by cooperating with two main global blockchain consortia (Enterprise Ethereum Alliance and Hyperledger Corporate Membership) and the Spanish multi-sectoral network (Alastria). With the aim of technology knowledge sharing, Iván is actively contributing to its dissemination through specialised learning programmes and congresses.

TECN — Fundacion Tecnalia Research & Innovation, Derio, Spain
Tel: +34 663 994 554; E-mail: ivan.gutierrez@tecnalia.com

Sergio Anguita
is a graduate in computer engineering at Tecnalia R&D. During his professional career, he has participated in projects related to web infrastructures, implementation and securing of near field communication (NFC) solutions, analysis of massive data for anomaly detection and investigation on Linux and embedded systems. In the field of mobile cybersecurity, he has done potential risk detection and privacy enhancement. He is currently focused on the applicability of blockchain solutions for industrial systems and processes.

TECN — Fundacion Tecnalia Research & Innovation, Derio, Spain
Tel: +34 667 178 827; E-mail: sergio.anguita@tecnalia.com

Celia Gomez
is a Technical Project Manager in the innovation area at Santander City Council.

Ayuntamiento de Santander (Santander Municipality), Santander, Spain
Tel: (+34) 942 200 600 Ext: 60491; E-mail: cgilsanz@ayto-santander.es

Juan Echevarria

is a Technical Project Manager in the innovation area at Santander City Council.

Ayuntamiento de Santander (Santander Municipality), Santander, Spain
Tel: +34 200 600 Ext 350; E-mail: cgilsanz@ayto-santander.es

Hans Houf

is a Senior Business and Innovation Consultant at Jibe.

Jibe.Company, Belgium
Tel: +35 989 845 5955; E-mail: hans.houf@jibecompany.com

Luca Nicoletti

is an IT Systems Architect at SOGEI. He holds a degree in physics from La Sapienza University, Rome. He worked for CONSIP between 2000 and 2013 (prior to its merger with SOGEI), where he designed many MEF IT (information technology) infrastructures such as the HR management system, the Access Management and Single Sign On system, the J2EE application platform, the 'Cloud DT' platform and others. This allowed him to obtain significant experience in project management and IT solutions development.

SOGEI — Societa Generale D'Informatica SPA, Rome, Italy
Tel: +39 320 431 1898; E-mail: lnicoletti@sogei.it

Roberta Lotti

has been working in the public sector since 1990, and currently she is the Director of the Project Management Office at the Information Systems and Innovation Directorate of the Ministry of Economics and Finance. She manages the promotion and supervision of technological innovation projects, with a focus on integration and cooperation between national and European institutions. Moreover, she is responsible for the definition of new policies and rules to increase privacy protection and data and information security management at the Ministry of Economics and Finance.

MEF — Ministero dell'Economia e delle Finanze, Rome, Italy
Tel: +39 06 4761 5942; E-mail: roberta.lotti@mef.gov.it

Domenico Natale

is an Administrative Officer in the Department for the General Affairs (DAG) of the Italian Ministry of Economy and Finance (MEF). He has been working in the Project Management Office at the Directorate of Informative Systems and Innovation, mainly dealing with European projects and privacy. He graduated with honours in economics and focused his postgraduate studies and work experience on the public sector.

MEF — Ministero dell'Economia e delle Finanze, Rome, Italy
Tel: +39 064 761 5971; E-mail: domenico.natale@mef.gov.it

Luca del Pizzo

is an IT (information technology) official in the Department for the General Affairs (DAG) of the Italian Ministry of Economy and Finance (MEF). He works in the Project Management Office (PMO) at the Directorate of Informative Systems and Innovation (DSII), and his activities deal mainly with project management, security and privacy. He graduated in computer engineering in 2012 and completed his

PhD in 2017 from the University of Salerno (Italy). He has published international papers on computer vision and artificial intelligence.

MEF — Ministero dell'Economia e delle Finanze, Rome, Italy
Tel: +39 064 761 5702; E-mail: luca.delpizzo@mef.gov.it

Francesco Pane
is an IT (information technology) official in the Department for the General Affairs (DAG) of the Italian Ministry of Economy and Finance (MEF). He works in the Project Management Office (PMO) at the Directorate of Informative Systems and Innovation (DSII), and his activities deal mainly with project management, security and privacy. He also worked for the University of Naples 'Federico II', and his main research interests are computer vision and neural networks.

MEF — Ministero dell'Economia e delle Finanze, Rome, Italy
Tel: +39 064 761 5601; E-mail: francesco.pane@mef.gov.it

Francesco Schiavo
has been the Director of Directorate of Informative Systems and Innovation (DSII) within the Ministry of Finance since 2009. He is responsible for the management of IT (information technology) infrastructure and systems that serve both the Ministry of Finance and the Italian public sector more widely. He leads the Departmentâ€™s Spending Review actions and is in charge of relations with SOGEI, the Italian government's in-house IT services firm. Prior to his appointment, he gained significant experience in the public sector in a variety of roles, including the Director of Italy's Higher School of Economics and Finance, Vice Director of a Customs Agency and Director of IT Systems for the Tax System, all within the Ministry of Finance.

MEF — Ministero dell'Economia e delle Finanze, Rome, Italy
Tel: +39 0647615744; E-mail: francescopaolo.schiavo@mef.gov.it

Abstract Personal data is currently being used in countless applications in a vast number of areas. Despite national and international legislation, the fact is that users have little or no control over who uses their data and for what purposes, and data protection is still, in many cases, a theoretical possibility only. In this paper, we present an approach to solve the problem of user data protection and control, currently being developed in the scope of the Protection and control of Secured Information by means of a privacy enhanced Dashboard (PoSeID-on) H2020 European project. In addition to an overall view of the project goals, the paper provides information on the requirements, challenges, conceptual architecture and functional description of the main modules. The presented solution complies with the European Union's General Data Protection Regulation and explores the use of the blockchain technology to provide data transactions protection and accountability as well as offer users full control over their personal information.

KEYWORDS: blockchain, GDPR, personal identifiable information, data protection, privacy

I. INTRODUCTION

The widespread use of digital services has led to user concerns on privacy and on the processing of their personal information by data processors and third parties. This, in turn, led to national and/or regional legislation, such as the European Union's (EU) General Data Protection Regulation (GDPR),[1] that aim at providing legal assurances in what concerns the protection of personal identifiable information (PII). On the contrary, technological development continues to deliver frameworks, tools and applications that demand PII user data in order to fulfil user needs in a large variety of areas, from public administration to sensitive individual health data. In this context, the demand for ways of protecting and controlling PII information has never been as high as now.

Given the preceding context, keeping track of who does what with each piece of personal information is essential. For this purpose, owing to its immutability and trustworthiness characteristics, blockchain[2,3] appears as a promising technology, although, at the same time, it poses considerable challenges, as PII cannot be stored in blockchains for obvious reasons.[4]

Several European projects address the topical areas of privacy, data protection and digital identities. Some of them are PDP4E,[5] BPR4GDPR,[6] DEFeND,[7] SMOOTH,[8] PAPAYA[9] and PoSeID-on.[10]

In this paper, we provide an overview of the approach taken by the Protection and control of Secured Information by means of a privacy enhanced Dashboard (PoSeID-on) European project for ensuring protection of and user control over PII, in environments where multiple data processor entities deal with data pertaining to a potentially high number of data subjects. In addition to proposing solutions to the highly demanding challenges of PII protection and control, the PoSeID-on proposes and explores innovative ways of using blockchains in contexts dealing with personal information. The list of contributions made by this paper is as follows:

i. A conceptual architecture for protecting and controlling PII using blockchains.
ii. An identification of requirements and challenges when dealing with personal data.
iii. A functional view of the main system modules currently being developed, namely blockchain, user dashboard, risk management, personal data analyser (PDA) and internal communication.

The remainder of this paper is organised as follows. Visual and Initial Guidelines as the name indicates, provides the PoSeID-on vision and initial project guidelines. Conceptual Architecture presents a high-level view of the system architecture, describing its main building blocks and respective roles. This section also provides examples of typical use cases. Personal Data Requirements and Challenges provides an insight into how PoSeID-on deals with PII related requirements and challenges. Permission Handling briefly identifies the solution that was adopted for handling PII transaction permissions in a way that does not disclose data subjects and data processors relationship mappings. Functional View provides a functional view of the main PoSeID-on modules. Lastly, the final section presents the conclusions and guidelines for future work.

II. VISION AND INITIAL GUIDELINES

This section starts by providing a brief introduction to the goals of the PoSeID-on project and its envisaged platform, allowing readers to understand the rationale behind the system requirements and architectural design decisions presented later in this paper. This section also presents a set of guidelines that were discussed and agreed upon between the project partners, as a common understanding of the intended nature of the PoSeID-on platform, and shared across the project partners, pilot owners and end users. These guidelines were informally produced in multiple interactions involving all the project partners and are reproduced here in order to reflect the process that led to the definition of the PoSeID-on system requirements and system architecture.

A. Goal of the PoSeID-on project

The goal of the PoSeID-on Project is to develop a transparent ecosystem for personal data protection, in line with the EU's GDPR, with respect to digital security.

As a matter of fact, in the current scenario of widespread use of digital services, and despite the overall awareness of legal requirements by the entities that process data, end users remain worried about data privacy, data protection, digital identities and data ethics.

The PoSeID-on solution is based on innovative technologies, such as blockchain, smart contracts and cloud computing, that provide targeted benefits for end users, potentially enabling them to manage personal data and data access authorisations in an easy, secure and auditable way. Additionally, it helps both public and private entities to identify new business opportunities, to be compliant with GDPR while processing personal data, as well as to undergo a substantial information and communications technology (ICT)-driven transformation, which will ensure higher security of end users' data. The PoSeID-on also impacts society as a whole, as it leads to increased trust in the digital market, in addition to supporting fundamental rights in the digital society.

The platform developed by the project is now being assessed in four different pilot deployments (in Italy, France, Spain and Malta) in public, private and mixed contexts. Specifically, the Italian pilot aims at enhancing e-services for public officials, the Spanish pilot aims to improve e-government services for the citizens of Santander, the Maltese pilot focuses on helping businesses to better sponsor and offer their services to customers, and the French pilot is aimed at simplifying e-services for French citizens. Initially, pilots involve a basic, limited set of users, to be enlarged during the evaluation phase. The pilots run in a controlled environment in order to simulate real life services and conditions.

B. Envisaged platform

According to its description of work (DoW), the PoSeID-on project is supposed to design, implement and validate a privacy-enhancing dashboard for personal data protection, a platform that manages all the personal data transactions between a data subject (owner of personal data) and private or public entities acting as data controllers or data processors. All relevant information shall be made available to users via a user-friendly web dashboard that allows them to track PII, manage PII access permissions and view the risk level stemming from their data exposure. In order to reduce identity fraud and protect the privacy of users, access to the dashboard is to be made available through electronic identification (eID) accounts only, in line with the eIDAS (electronic identification and trust services) regulation.[11]

The PoSeID-on project considers a solution based on smart contracts and permissioned blockchain. Through smart contracts, the project aims to meet the need of data confidentiality, inviolability and access control for data subjects. Through the blockchain technology, references to PII shall be managed and exchanged securely. In compliance with the data minimisation principle, smart contracts are only supposed to contain the reference to the users' personal data necessary for the specific transaction. In order to restrict the number of third parties who can access PII and give/revoke the authorisation to process it, the PoSeID-on considers the adoption of a fully private, permissioned blockchain. This means that not only write permissions are limited to PoSeID-on participants but also read permissions. In this way, once all the permissions and references to personal data are inserted into the blockchain, they are only visible to authorised third parties. In the context of this paper, the term 'permissioned blockchain' refers to a blockchain for which read/write permissions are restricted to the PoSeID-on participants.

The blockchain technology was selected due to two main reasons. First and foremost, there was the need to maintain an irrevocable record of PII transactions,

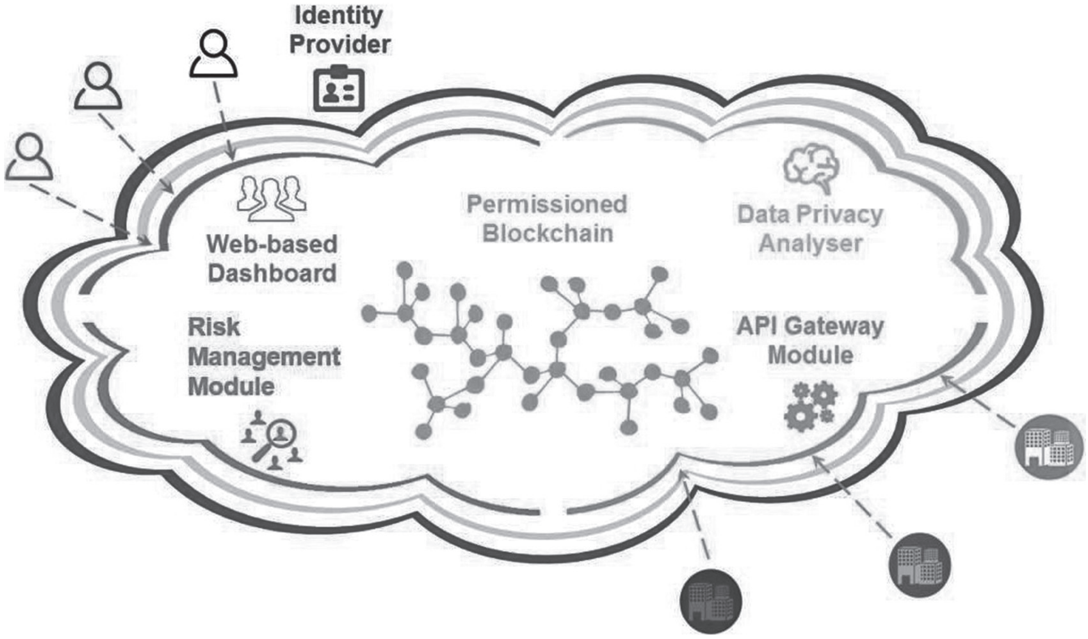

Figure 1: High-level view of the envisaged PoSeID-on platform
Note: API, application programming interface; PoSeID-on, Protection and control of Secured Information by means of a privacy enhanced Dashboard.

including permissions handling and all kinds of operations involving PII processing, for providing full control to PII owners, for accountability and for legal assurances. On the contrary, there was the need to allow multiple entities to share data and to contribute to data processing without relinquishing control over their own databases or without relying on a central datastore. By agreeing to participate in the PoSeID-on system, users benefit from full control over their PII and third parties can provide an auditable ledger of all their PII-related operations to users and regulators. Moreover, it should be highlighted that no PII is ever stored in the blockchain, which only stores information on permissions and on PII handling.

Figure 1 shows a high-level view of the envisaged PoSeID-on platform, as originally presented in the project DoW, illustrating the different roles (end users, data processors) and components (dashboard, permissioned blockchain, identity provider, data processor application programming interface [API] module, risk management module [RMM] and data privacy analyser). These will be detailed in the remaining sections of this paper.

It should be noted that from the perspective of the PoSeID-on, data controllers (entities that determine the purposes, conditions and means for the processing of PII) and data processors (the entities that process PII on behalf of data controllers) are treated in exactly the same way, as these functionalities often reside in the same system. Thus, hereafter, the term 'data processor' indistinguishably applies to both entities.

C. Project guidelines

In order to detail and clarify the PoSeID-on high-level vision presented in the previous section, during the initial stages of the project, the PoSeID-on partners identified and agreed upon a set of guidelines to be used as a common understanding of the PoSeID-on platform. These are briefly referred here.

Dataset

The dataset of PII types serviced by the PoSeID-on must cover not only the pilot use cases but also be extensible in order to cover (all) future use cases. The pilots are used for defining reasonable datasets that, nevertheless, must not limit the PoSeID-on system operation.

Blockchain implementation

The blockchain will only be used to store permissions. PII will not be stored in the blockchain. The right to be forgotten, as expressed in Article 17 of the EU GDPR, requires to not store data (PII) into the non-editable blockchain.

PII storage

The PII will be stored either with authoritative data processors or, for a small subset of PII that is owned only by the data subject, on the PoSeID-on platform itself. According to the permissions given by the data subject, the PII will be synchronised with other data processors. Data subjects will always be in full control of their own PII.

Since the release of these guidelines, the approach to PII storage in the PoSeID-on platform itself has been adjusted. While from a functional point of view, the architecture still supports this notion of having the data subject storing and managing some sort of 'master PII' used to synchronise with data processors (eg to update a phone number or an address at all data processors), from a conceptual point of view, this function has been moved to an external data processor able to provide these services and to act on behalf of the data subject when requesting PII updates to other data processors. Having the storage of PII within the PoSeID-on platform itself would create unnecessary technical restrictions and could create potential loopholes in the principle of protecting data subjects' privacy even from PoSeID-on operators. Therefore, this functionality has been externalised to an autonomous data processor which, depending on deployment options, may be operated by the PoSeID-on authorities or by third parties trusted by the data subject. A proof-of-concept data processor providing this functionality is being developed during the project to demonstrate the concept and to support the pilots.

Right to be forgotten

As PII is not stored in the blockchain, but shared and synchronised with third-party systems, a procedure needs to be in place to assure the right to be forgotten as defined in the GDPR. First of all, third parties are legally forced to destroy PII that is no longer covered by the permissions given by the data subject. Secondly, the PoSeID-on will notify the data processors automatically, via an API call, when their permission over PII has been revoked. It will be up to the data processors to destroy all copies of that PII in order to stay within the confines of the GDPR. It should be noted that there are exceptions to the right to be forgotten and that data processors may not be able to delete some PII due to legal obligations, defence of legal claims, public interest, research purposes or even statistical purposes, for example. The verification and/or assurance of the actual erasure (or not) of the PII is outside the scope of PoSeID-on, whose role is limited to notifying data processors.

Sharing of PII with data processors

Data subjects share PII with data processors. PII will flow through the PoSeID-on system. If a data processor cannot be reached, for whatever reason, the PoSeID-on will hold onto the PII in-transit, acting as an intermediary cache.

Limits will be in place to make sure the PoSeID-on will discard the PII in-transit after a configurable amount of time has elapsed during which the data processor was offline. While the PII is in-transit, it will be encrypted in such a way that only the recipient can decrypt it. The PoSeID-on system will transparently take care of this encryption. The platform shall provide non-repudiation of access to the PII by data processors. The authorisation management shall be implemented in a granular way, also providing users segregation.

Publishing updates of PII

Data processors will accept any update on the PII they receive from the PoSeID-on, as it is guaranteed to be an update initiated by the data subject. All changes to the PII will be automatically published to the data processors, which have been allowed access according to the permissions as stored in the blockchain.

PII safeguarding: encryption

Using the public key infrastructure that is in place, all the PII (at rest and in-transit) will be encrypted so that only the recipient can decrypt it. The sending party (which can be the PoSeID-on or a data processor) is responsible for encrypting the PII before submitting it.

A central index of all public keys is needed. Every participant in both the blockchain and the PII data exchanges will need to be identified with a public key in order to safely transport the PII and do blockchain operations. This central index can improve security by storing all the keys needed for encrypting transactions. It will be accessible only to authorised super users for management activities and log analysis.

PII safeguarding: data processors and third parties

All data processors need to comply with the GDPR. This means that before integrating with the PoSeID-on, data processors must have their operational security and data safeguarding certified by an accredited certification body in accordance to Article 43 of GDPR. Similarly, third parties are required by the GDPR to safeguard PII with proportionate security measures.

III. CONCEPTUAL ARCHITECTURE

This section describes the PoSeID-on's system architecture, introducing the core components and actors of the platform. For the sake of overall readability, the conceptual architecture is provided before the discussion of system requirements (cf. next section), to offer the reader a first overview of the platform components.

As already mentioned in Visual and Initial Guidelines , the PoSeID-on aims at providing a privacy-enhancing dashboard for personal data protection, a platform that manages all the personal data transactions between a data subject (owner of personal data) and private or public entities acting as data processors (including direct data transactions between data processors). All relevant information shall be made available to users via a web dashboard. Access to the dashboard is made available through eID accounts.

Figure 2 illustrates the overall PoSeID-on architecture, identifying the various system components. Table 1 lists the conceptual components and the respective short description. The following subsections provide additional details regarding each of the components.

A. Data subjects, data processors and administrators

The PoSeID-on considers three different types of actors: data subjects, data processors and administrators.

Data subjects are natural persons that represent the primary target of the GDPR. Data subjects own PII that constitutes a valuable resource for third parties and represents a privacy risk. Nevertheless, in several situations, data subjects need to share

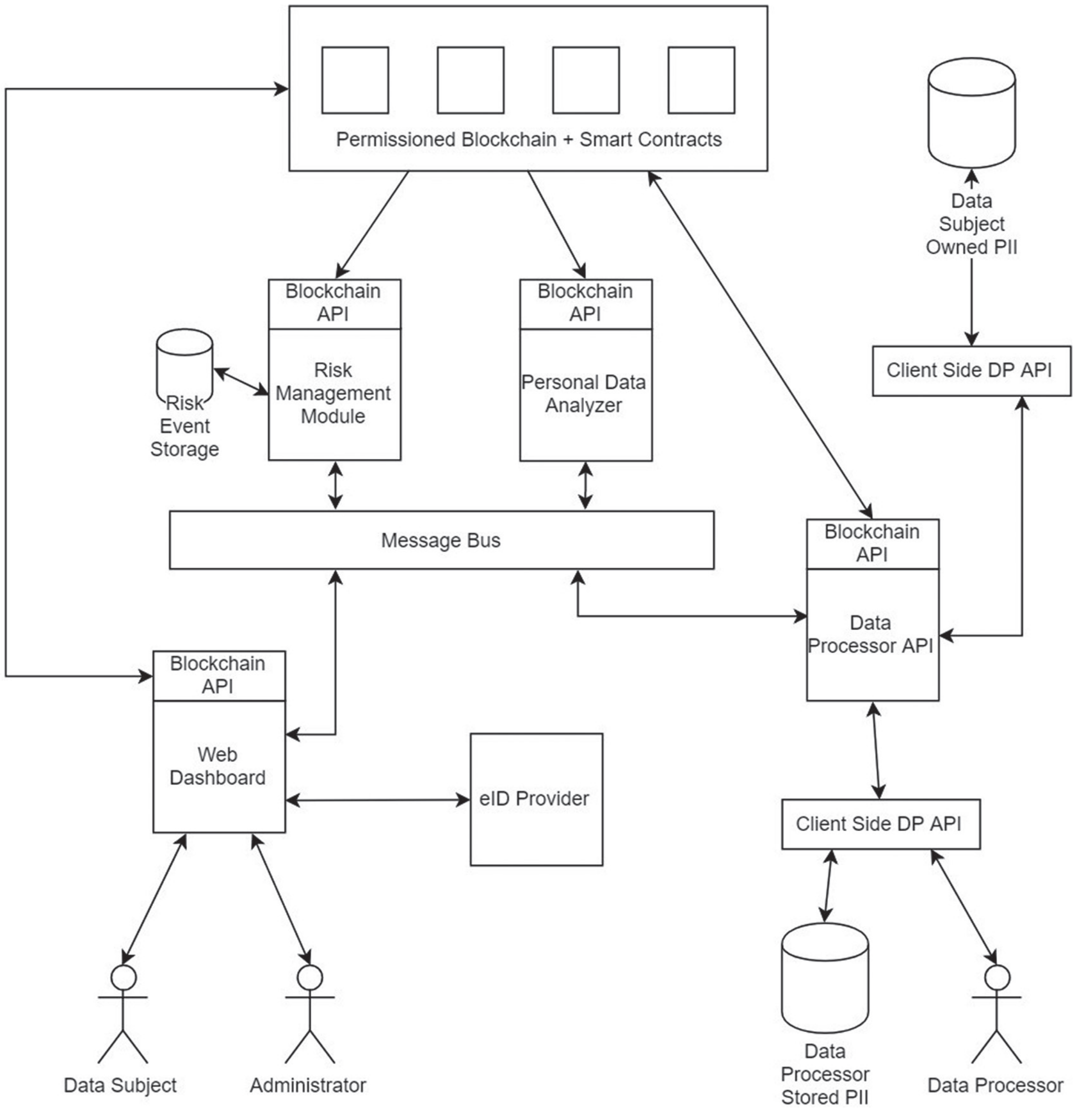

Figure 2: PoSeID-on general architecture
Note: API, application programming interface; DP API, data protection API; eID, electronic identification; PII, personal identifiable information; PoSeID-on, Protection and control of Secured Information by means of a privacy enhanced Dashboard.

this PII with third parties (such as insurance companies, health institutions, tax services, other public services, banks and so forth), either directly (data subject provides the PII) or indirectly (a third party that already has some PII shares this PII with another third party, after having been given permission by the data subject — this is the case, for instance, of a tax service getting revenue reports directly from a bank).

Third parties exchanging PII with the data subject or between themselves are designated as data processors. Depending on the granted permissions and on the nature of the exchanged PII, several models of exchange may be used, such as single or multiple use with limited time and scope or use for an unlimited period of time. Depending on the nature of the permissions granted by the data subject,

TABLE 1: Conceptual architecture components

PoSeID-on conceptual architecture components	Brief description
Data subjects, data processors and administrators	Primary target of PoSeID-on platform end users
Dashboard	Interface for data subjects and administrators
Data processor API	Access point for data processors to send/receive requests
Client-side data processor API	Connector to data processor's internal information systems
Permissioned blockchain and smart contracts	Blockchain implementation where only authorised parties can propose changes. Serves as a backend for PII access management within the PoSeID-on platform
Blockchain API	Abstraction layer that allows modules to access and interact with the blockchain
Risk management module	Detects operational anomalies that may translate to security and privacy risks
Personal data analyser	Detect and evaluate privacy risks within PII transactions
eID provider	Authenticates users in the PoSeID-on platform
Data subjects' PII repository	PoSeID-on's storage for PII owned by the data subject (eg not belonging to a data processor, introduced manually by the data subject into the PoSeID-on)
Message bus	Messaging module for the PoSeID-on's components communication

Note: API, application programming interface; eID, electronic identification; PII, personal identifiable information; PoSeID-on, Protection and control of Secured Information by means of a privacy enhanced Dashboard.

data processors may store PII and share this PII with other data processors within the limits established by the permissions. The PII storage mechanisms actually employed by each data processor fall outside the scope of the PoSeID-on. The data processor is supposed to comply with data subject requests and permissions regarding retention, deletion and processing of their PII due to regulatory and legal obligations.

The administrators are the operators managing the PoSeID-on platform. The platform is operated by a pre-assigned entity — typically a public or semipublic organisation with the mission of providing a PII management framework such as PoSeID-on — usually considered a trusted third party. Nevertheless, the whole platform design minimises the risk of intentional or accidental breaches of privacy originating on the management entity. Administrators manage the PoSeID-on platform on behalf of this entity.

B. Dashboard

The interface provided to data subjects, to access the PoSeID-on platform, is a web-based application that provides access to the various types of operations performed by data subjects, such as granting, modifying and revoking permissions for a specific data processor, checking the history of exchanges of their PII and receiving alarms due to high privacy exposure risks. This human-oriented dashboard is the primary interface for data subjects, although, in some situations, other communication channels may be used according to the data subject preferences (eg receiving urgent alarms of privacy exposure via e-mail or SMS [short message system]). Access to the dashboard is based on logging with the data subject's national eID (or similar credentials, depending on the specific PoSeID-on instance).

The dashboard also provides an interface for PoSeID-on administrators, although with a different set of functionalities and user interfaces.

C. Data processor API

The access point for data processors, for communicating with PoSeID-on, is the data processor API. While the interface for data subjects and administrators (the dashboard) is primarily human-oriented, the interface for data processors is focused on interconnecting the data processor's already-existing information systems with the PoSeID-on so that their business logic can integrate with the PoSeID-on functionality. This application provides an authenticated API over which data processors can send requests for/about PII. The application is also used for sending data processor messages through their client-side data processor API, about changes in PII values and PII deletion requests.

D. Client-side data processor API

In order to interconnect to the PoSeID-on platform, data processors can use a client-side data processor API, which is made available to data processors to interface with their PII store. It manages the transport of PII and the revocation of permissions.

The PoSeID-on itself will also develop one client-side data processor API to manage communication with the data processor API, which will be implemented for the PoSeID-on platform to store PII that has no central authority otherwise (PII such as personal phone numbers). From the rest of the system's point of view, this is just another data processor. This PoSeID-on local data processor can also act as a reference implementation for the data processor API specification.

E. Permissioned blockchain and smart contracts

The PoSeID-on permissioned blockchain consists of a special-purpose blockchain implementation built from standards and open source, configured and customised to work within the PoSeID-on system. It will be permissioned, meaning that only known parties will be allowed to propose changes to the ledger. This will also allow the PoSeID-on network to be faster, avoiding slow consensus mechanisms such as proof-of-work and proof-of-stake. Smart contracts in the blockchain rule the management of the requests and permissions to grant, deny and check PII access. The blockchain nodes are hosted by the PoSeID-on administration entity and the participating data processors. It is also possible to allow the inclusion of nodes hosted by other GDPR-certified entities, which might make sense in some scenarios, eg scenarios in which there is a small number of participating data processors.

F. Blockchain API

The blockchain API abstracts all blockchain operations into a high-level API suitable for integration into other applications. As the use of blockchain carries some important implications on how the clients and the servers behave, actions like account management (data processor and data subject identity on the blockchain) or system functionality (smart contract functions usage) change drastically. For instance, users shall sign every call to a smart contract, but other components in the overall PoSeID-on architecture, like the dashboard module, might function normally despite the existence of a blockchain network behind it.

Modules writing information to the blockchain ledger must have a mechanism to access the network. This is the blockchain API that also allows the intervening modules to feed the risk management and personal data analyser modules. Thus, interaction with the blockchain is limited to two main components: a blockchain communication module (the blockchain API) and the distributed ledger nodes. The user authenticates using eID through the dashboard. The eID provides the public key for identification of the user. A private key is then created within the PoSeID-on based on a user-introduced password. This key is then bound to the eID of the user, which is then associated to a burnable user id, managed by the blockchain API, to be used for identification in the ledger. From the dashboard and data processors perspective, the interaction with the blockchain or smart contracts happens through the blockchain API, which abstracts all blockchain operations into a high-level API suitable for integration into other applications. This API is not a web service, but a wrapper used in the dashboard, providing direct connection to one of the nodes. This API takes care of actions like account management (data processor and data subject identity on the blockchain) and system functionality (smart contract functions usage). The calls to smart contracts are signed using the private key generated when the user first logged in on the platform.

G. Risk management module

The RMM is responsible for monitoring the PoSeID-on, both from a system-wide perspective and from the point of view of individual data subjects' operations.

The RMM is expected to detect and evaluate possible security and privacy risks, such as anomalous behaviour from data processors (eg a specific data processor suddenly starts collecting much more data than usual from a large number of data subjects, which may mean the data processor was hacked and is being used to syphon PII to external attackers) or risks associated with a specific data subject (such as successive attempts to login with his or her credentials). Risk detection is performed by combining machine learning (ML) algorithms, which analyse multiple sources of information about transactions, user-level behaviour and system-level behaviour, in the form of component logs. When the data subject provides explicit consent, transaction-specific data and PII will also be sent to the RMM and used to complement this analysis. High risk levels may trigger alerts to PoSeID-on administrators and data subjects, depending on the RMM settings.

H. Personal data analyser

The PDA will be used to monitor personal data transactions and related warnings generated by the blockchain platform, in order to detect and prevent anomalies and misbehaved transactions. A warning is generated each time a transaction is not received and approved by all the interested parties.

This component is used to control personal data in a transaction, with the aim of discovering all previously non-identified personal data, such as personal data for which there is no data subject authorisation. Due to the sensitive nature of the analysed data, the PDA acts only when explicit consent is provided by the user. Furthermore, all input data is discarded after each analysis. In case the user does not provide explicit consent, the PDA will simply not operate for the data from that specific data subject. While this scenario is not the most desirable, it does not affect other data subjects.

I. EID provider

An identity provider authenticates persons and organisations on behalf of the PoSeID-on platform, in line with the European eIDAS regulations and ecosystem.

J. Data subject's PII repository

Data processors may receive PII directly from the data subject or from (allowed) transactions with other data processors. In the former case, it is important to provide an easy-to-use mechanism for the data subjects to insert, store and update this PII. Therefore, it is considered that data subjects always use a data processor for storing their PII, even when this data processor has no other specific purpose (in practice, in which case, the data processor just acts as a storage point for the data subject's PII). This component will be responsible for storing all the PII considered private to a data subject for which no authoritative data processor exists. It will interface with the rest of the PoSeID-on system by using the same data processor API as all other data processors. The only difference is that this component acts as a data processor with the data subject's credentials. All data stored within this component is encrypted with the data subject's public key instead of a data processor–specific key.

Depending on deployment options, storage itself may take place

- on premises with the rest of the PoSeID-on components, although, due to the adopted encryption mechanisms, the PoSeID-on authority actually has no access to the data subject's PII — therefore meeting the privacy requirements previously discussed (whenever the same authority manages the PoSeID-on platform and the data subject's PII repository, as will happen in the planned pilots) or
- in the infrastructure of third-party entities, when this service is (hypothetically) provided by third parties selected by the data subject. The adopted encryption mechanisms still prevent unauthorised access to the data subject's PII in this case. Also, it should be noted that this deployment model is supported by the conceptual architecture but will not be tested and/or adopted by any of the planned PoSeID-on pilots.

A side benefit of this approach is the fact that, in this way, a sample data processor will be developed and made available to the (real) data processors participating in the pilots, as a showcase of how to use the data processor API to access PoSeID-on services. Those data processors can use this example as a starting point to interconnect their own information systems with the PoSeID-on.

K. Message bus

This component provides a messaging infrastructure for PoSeID-on components to communicate with each other in a controlled but decoupled fashion, which will allow for the easy addition and removal of components without affecting the overall system operations and asynchronous communication, thereby facilitating scalability and fault tolerance.

The different components of the PoSeID-on can be seen as individual applications. Messages are passed between them, through a queue mechanism, with each and every one signed by the sender and encrypted for the recipient. When a component is unreachable (for instance, a data processor is offline), the message will be kept until either a pre-defined timeout (say, one week) passes or the recipient comes back online.

In this way, the PoSeID-on can stay easy to reason about, easy to test, easy to develop and maintain by many distributed parties, and easy to scale up when the need arises.

L. Examples of typical use cases

In order to complement the description of the PoSeID-on conceptual architecture provided earlier, next we briefly present a few representative use cases for the PoSeID-on system.

Use cases initiated by the data processor

Through the data processor API, a data processor can request the PII values pertaining to a given data subject. The PoSeID-on will check the blockchain in order to determine if access has been granted by the data subject and will fetch the PII values from the authoritative data processors responsible for storing this data. The act of accessing (permitted or not) will be stored in the blockchain. The data will be encrypted using the requesting data processor's public key, making the PII secured in-transit and only readable by the requesting data processor.

Through the data processor API, a data processor can request access to one or more PII types of a given data subject. The data processor needs to tell PoSeID-on, for each PII type, whether the corresponding PII value will be stored by the data processor itself or if it is being delegated to an authoritative data processor for that PII type. The authoritative data processor for each PII type can be configured by the PoSeID-on administrator. For example, a passport number is supplied by the ministry of internal affairs.

For each requested PII type, the data processor also needs to supply information about whether the PII value can be updated by the data subject, whether the PII value is mandatory, and what this PII type will be used for. This request will be converted into a smart contract and stored into the blockchain. The data processor will now appear in the data subject's dashboard.

Through the data processor API, a data processor can also request access to one or more PII types of all data subjects. This is the same procedure as asking it for a specific data subject, apart from the fact that every data subject will now see the data processor in his or her dashboard instance. This is, for instance, the case of a governmental organisation. Requesting the PII of all data subjects at once can be very invasive and, therefore, should be restricted to pre-authorised data processors only. The PoSeID-on administrator can configure which data processors can do blanket requests like these.

Through the data processor API, a data processor can send PII data to the PDA for analysis and detection of possible privacy issues.

Use cases initiated by the data subject

Through the web-based dashboard, a data subject can grant access for a data processor to a PII type, after the data processor has requested access to that PII type. The data subject can see what this PII type will be used for, as supplied by the data processor. Access is automatically granted if the data processor has deemed this PII type to be mandatory. The act of granting access will be stored in the blockchain.

Through the dashboard, a data subject can revoke access to a PII type for a given data processor. This will only be possible if the data processor has not made this PII type mandatory. The reason for making a given PII type mandatory, as supplied by the data processor, will be visible to the data subject. If this PII type is read-only, access cannot be revoked either. Typically, this is the case for the PII generated by the data processor itself (eg a driver's license number). After revoking access to certain types of PII, the data processor will be notified through an API call of this effect. It is up to the data processor to destroy all copies of this PII. The act of revoking access to the PII will be stored in the blockchain.

Through the dashboard, a data subject can view all PII types and values known to any given data processor. The PII values will be requested from either the data processor or from the associated authoritative data processors. The PII will be encrypted using the data subject's public key so it will be both secure in-transit and only readable by the data subject.

Through the dashboard, a data subject can update PII values, if they are not specified as read-only by the data processor. Making specific checks based on a set of pre-defined rules, the dashboard will not accept malformed data. After the data subject has successfully updated his or her PII, the PoSeID-on will send an API request to the data processor informing it of the new value. All data processors that use this supplied PII, by using the primary data processor as an authoritative source for this PII type, will also be notified of the new value. The updated PII values will be encrypted separately for each receiving data processor, by using the data processor's public key. The act of updating the PII value (but not the value itself) will be stored in the blockchain. In some cases, PII can be supplied to data processors by a data subject in a one-shot operation. The PII will be used to execute a process specific to that data processor and later discarded. So, instead of having read/write access to the PII stored in the data processor's system, the data subject can only write. A successive write can either trigger an amendment or a restart of that process.

IV. PERSONAL DATA REQUIREMENTS AND CHALLENGES

PoSeID-on architectural components are hereafter analysed in Table 2 with reference to their potential impact on final users' personal data, posing the following challenges for their implementation.

In compliance with the GDPR and according to the challenges identified in Table 2, the most relevant legal requirements to be considered in conceiving and developing the PoSeID-on platform following the privacy by the design approach are hereafter listed. It should be noticed that these requirements refer to the entire platform development.

- **Lawfulness of collected personal data:** An explicit consent for sensitive data should be requested. The consent should be recorded and a clear record of the agreement of each data subject should be in place. Functional mechanisms should also be set up for consent withdrawal, implying the capability to locate and remove the personal data during the process and also from backups and archives (even in cloud).
- **Adequateness, relevance and proportionality:** The collection of personal data must be limited to what is directly relevant and necessary to accomplish the PoSeID-on platform-specified purpose. The purpose has to be legitimate, and it has to be specified and made explicit before collecting personal data.
- **Accuracy of the collected personal data:** Procedures to keep data up-to-date should be implemented, such as the final user validation of data. Functional mechanisms should be in place to check, edit and extend stored data, with various controls concerning secure and reliable identification, authentication, access, validation, etc. These mechanisms may also affect backup and archives copies.
- **Storage limitation:** Functional mechanisms should be able to erase specific stored data, with various controls concerning secure and reliable identification, authentication, access, validation, etc. These mechanisms may also affect backup and archives copies.
- **Transparency and openness:** The PoSeID-on platform should provide appropriate information to individuals to exercise their rights, data controllers to evaluate their processors and data protection authorities to monitor according to their responsibilities.
- **Individual rights:** Secure and reliable identification, authentication and data access should be ensured. A withdrawing form should be available in the platform. A mechanism should be implemented to identify the specific data that is to be blocked or restricted. Extracted data should be limited to the identified and authenticated person concerned and communicated securely (eg encrypted). All these mechanisms may also affect backup and archives copies.
- **Automatic processing:** Mechanisms allowing manual review should be implemented.

TABLE 2: Personal data protection challenges in PoSeID-on

PoSeID-on architecture component/technology	Component brief description referred to personal data management	Data protection challenges
Web dashboard	It will allow the collection of subjects' data and the exercise of their rights, including access to personal data	• Adequateness, relevance and limitation of data collection • Lawfulness, fairness and transparency in collecting the data • Collecting of data for specified, explicit and legitimate purposes • Accuracy of data • Collection of sensitive data • Concise, transparent, intelligible and easy access to personal data by the subject • Information on how data is processed expressed in a clear and plain language
Digital identity connector	It will allow the use of the electronic identification (eID) accounts by data subjects, to access the PoSeID-on platform	• Reliability of the authentication process
Data processor API	It acts as an API front-end, receives API requests, enforces throttling and security policies, passes requests to the back-end service (blockchain) and then passes the response back to the requesters. It will be available to external entities, to interact with the PoSeID-on platform	• Trustworthiness and reliability of the authentication process • Adequateness, relevance and limitation of data collection • Lawfulness, fairness and transparency in collecting the data; processing of data for specified, explicit and legitimate purposes • Accuracy of data • Processing of sensitive data
Risk management module, machine-learning algorithms and techniques	It will have historical data logs — including network and authentication logs collected and stored in log management systems — to identify patterns of traffic caused by user behaviours, both normal and malicious. These activities, based on the user behavioural analysis (UBA) approach, will not act based on their findings, but are intended to provide security teams with actionable insights	• Ensuring rights and freedoms of subjects while using automated processing of personal data to evaluate subjects' behaviour
Personal data analyser, natural language processing and understanding	It will monitor privacy risks, notifying privacy threats to data subjects during data transactions and discovering all previously non-identified personal data, such as personal data for which there is no data subject authorisation	• Discovered data, matched with the current data, could identify the subject
Permissioned blockchain, smart contracts	It will contain references to personal data and will deploy smart contracts for personal data management	• Privacy incursions due to the presumed blockchain transparency • Ensuring the right to be forgotten and to rectification of data
Cloud	Personal data is saved on the cloud	• Uncontrolled distribution of personal data • Data leakage • Geographical distribution of data and consequent difficulty to determine applicable law due to data volatility in the cloud • Exercising of rights of data subjects may be subject to different conditions depending on data geographical distribution • Data retention period should be granted in multiple locations • Breach notification obligations and protocols • Ensuring data portability • Ensuring data security

Note: API, application programming interface; PoSeID-on, Protection and control of Secured Information by means of a privacy enhanced Dashboard.

- **Accountability:** Examples of accountability measures are related to tracking of personal data access and of communications with external systems, documenting and recording all processing activities, and mapping data flow. Moreover, appropriate data breaches reporting, response, assessment and information security should be developed.
- **Data security:** From a privacy and data protection perspective, there is need for a set of rules for limiting access to authorised people only and to ensure that the data is trustworthy and accurate. Therefore, data should be kept secure by applying privacy enhancing technologies (eg encryption, pseudonymisation, anonymisation, identity and access management), preventing accidental disclosure of personal data and securing communications with external stakeholders (such as, for instance, external systems).

Given the requirements and challenges presented earlier, Table 3 provides an insight into how the PoSeID-on addresses them.

V. PERMISSION HANDLING

The PoSeID-on architecture provides ample support for encryption (and privacy) between the various actors involved (data subjects, data processors and administrators). The PII data exchanges (eg between data processors or between data subjects and data processors) are made through the message bus, but PII is encrypted in such a way that its content is not accessible, not even to the PoSeID-on administrations. The PII in clear text form pertaining to non-public transactions must not be stored in the PoSeID-on platform. This includes transaction permissions, as these are PII, because they represent relationships and, thus, should not be public knowledge. In short, no single party is allowed to have access to relationship mappings outside its own memberships.

In order to achieve this, PoSeID-on permission handling explores the use of ’constellations’, provided by the Quorum technology.[12] Quorum is an Ethereum-based open-source distributed ledger protocol developed by J.P. Morgan, which allows private contracts and transactions between parties. The constellation is a Java implementation of a general-purpose system for enabling private transactions of information in a secure way.

A Quorum node delegates all functionality directly related to secure data exchange to its corresponding constellation, with which it communicates. The constellation consists of two independent sub-modules:

- Transaction manager is responsible for transaction privacy. It stores and allows access to encrypted transaction data, but it does not have access to any sensitive private key, as it delegates all the cryptographic functions to the ’enclave’.
- Enclave performs all cryptographic operations, such as symmetric key generation or encryption/decryption, and holds the related private keys, acting as a virtual security module.

With these components, the features of the proposed solution can be summarised as follows:

- Only the public state ledger is public knowledge and stores public transactions.
- All nodes store in the public state ledger all transactions and private transactions contain only the hash of the message between the sender and recipient.
- Only the two parties involved in a data transaction relationship can learn of the contents of the private transaction, ie only the recipient receives the message.
- The constellations (transaction managers) transfer the messages and they are not part of the immutable ledger.
- Messages are transmitted encrypted with the pub key of the recipient.

TABLE 3: The PoSeID-on approach to GDPR requirements

GDPR provision		GDPR reference	PoSeID-on guarantees
Lawfulness of collected personal data	Lawfulness implies having legitimate grounds for collecting and using the personal data, including having the unambiguously given consent of the individual whose personal data is being processed and not using the data in ways that have unjustified adverse effects on the individuals concerned, and each purpose is consented separately unless it is appropriate to merge them. Sensitive data concerning a person's race, political opinions, religion, sexuality, genetic information and other biometrics should be prohibited by default, unless consent is explicitly given and processing is necessary.	Art. 5, 6	The consent is recorded and a clear record of the agreement of each data subject is kept in the blockchain ledger. An explicit consent for sensitive data is always requested and recorded before any data is transferred. Consent withdrawal is also possible, and although PoSeID-on does not directly enforce the deletion of information previously transferred to a data processor, data processors are informed of the revocation and must abide to the law. In this situation, the PoSeID-on acts as an auditing tool.
Adequateness, relevance and proportionality	For the purpose limitation principle, collected data should be adequate and relevant to the objectives of the system of collected personal data. Therefore, for the data minimisation principle, only the minimum amount of personal data that is needed to achieve the specific PoSeID-on platform purpose must be collected, used and retained.	Art. 5	The collection of personal data is limited to what is directly relevant and necessary to accomplish the PoSeID-on platform-specified purpose. This purpose is specified when the user first uses the platform through a set of policies, terms and conditions. In addition, data processors requesting data must specify their purpose and this purpose is analysed by the PDA module and cross-checked against the requested PII fields.
Accuracy of the collected personal data	Data has to be the right value; it has to precisely represent the value in consistent form, and it must be up to date.	Art. 5	The dashboard allows for the rectification of PII stored by data processors. An update request is sent to the data processors holding the updated PII through the data processor API.
Storage limitation	Personal data must be retained only as long as necessary to fulfill the declared purpose. It must be erased or effectively anonymized as soon as it is not needed anymore for the given purpose.	Art. 5	Permissions have an associated timespan, at the end of which each data processor is informed of the fact and its permissions are revoked and a data deletion request is issued.
Transparency and openness	Being transparent about the purpose to use the data.	Art. 12	By using the blockchain to maintain a ledger of permissions and data access, it is possible to audit and confirm the lawful use of the data.

(continued)

TABLE 3: The PoSeID-on approach to GDPR requirements *(continued)*

GDPR provision		GDPR reference	PoSeID-on guarantees
Individual rights	Individuals must have the possibility of effectively and conveniently exercising their rights to access and rectify as well as to block and erase their personal data and to obtain a usable portable electronic copy of their personal data. Furthermore, they have the right to withdraw given consent with effect for the future.	Art. 15, 16, 17, 18, 19, 20, 21	Secure and reliable identification, authentication and data access are ensured through the use of eID and a personal password. Withdrawing from the platform is possible and mechanisms to ensure the connection to previously stored information are in place, through the form of burnable pseudo-identifiers.
Automatic processing	Subjects have the right to insist that key decisions arising from automatic processing of their personal data are manually reviewed/reconsidered.	Art. 22	Automatic processing of personal data is performed by the PDA and RMM modules, but neither of these two modules act on the behalf of the data subject. The PDA and RMM modules provide the data subject with relevant information which he/she can then use to act on his or her own regarding permissions.
Accountability	Data controllers and processors should be able to demonstrate the compliance with privacy and data protection principles and legal requirements.	Art. 24	By recording every PII exchange and issued request by data processors in the ledger, this effectively provides a means to demonstrate compliance with privacy and data protection principles when it comes to traceability and accountability of PII processing. It is out of the scope of the PoSeID-on, thought to control the PII once it has been delivered to a data processor, on the data processor side.
Data security	Data security addresses integrity, confidentiality and availability concerns.	Art. 25	All data is encrypted at rest, and end-to-end encrypted when in transit. Pseudo-anonymisation of identifiers is performed in order to hide the real identity of each data subject while allowing the analysis of each data subject's PII operations history. In addition, the goal of the RMM is to identify anomalous behaviour and the volume of operations in order to warn data subjects when their consented PII might be under unlawful use.

Note: API, application programming interface; eID, electronic identification; GDPR, General Data Protection Regulation; PDA, personal data analyser; PII, personal identifiable information; PoSeID-on, Protection and control of Secured Information by means of a privacy enhanced Dashboard; RMM, risk management module.

- Only the transaction managers in the sender and the recipient store the transaction payload in clear.
- Databases in transaction managers manage private messages and can be deleted.
- Each node has a private state ledger, owned and accessible only by it. The message can optionally be stored there.
- Enclaves manage the private keys to encrypt/decrypt the transactions in the transaction manager.

This is a smart contract–based solution that guarantees that all public transactions and only hashes of private transactions are stored in the public ledger. Moreover, the solution ensures that permission message types and protocol (grant/update/revoke) can be defined as wanted and that all transactions are auditable and the code (smart contracts) is verifiable by external auditors. Moreover, as the solution is based on the extensively used Ethereum[13] technology, it is mature and easily maintainable.

VI. FUNCTIONAL VIEW

This section provides an overview of the main PoSeID-on components from a functional perspective.

A. Blockchain and smart contracts

The need of the PoSeID-on system to provide a distributed, auditable and secured PII permission management solution is addressed by the blockchain module, whose architecture is shown in Figure 3.

With the application of blockchain, permissions over data subjects' PII and transactions of data subjects' PII between data processors will be stored in a distributed ledger that allows participant organisations to share data subjects' information and, yet, to share a mechanism to protect that data from being flowed to nonparticipating organisations. Each of the participant organisations will be a peer participating in the distributed ledger by keeping a private ledger synchronised with the rest.

As PII permissions and transactions are both a form of PII as well, there is need for protecting their confidentiality and integrity. Therefore, a blockchain platform implementation that allows private communications between parties is necessary. To this aim, the PoSeID-on solution relies on a permissioned blockchain implementation. This means that each deployed peer will belong to a well-known party who can only participate in the consensus of data validation for the operations that are privately shared with it. These will be separate ledger (data) access permissions not only virtually but also physically, thus leading to added security of private information.

In the PoSeID-on, as the dashboard and data processor components (and their related API components) belong to well-known parties, it is possible to find a scheme of blockchain participants. Each of these participants will be aware of a certain set of the data subject's PII permissions and then will be able to check their states.

This architecture includes the required design modules according to the selected blockchain base platform, Quorum, which fits the PoSeID-on requirements. While public contracts and transactions work practically in the same way as in the Ethereum public network, private contracts and transactions (and, of course, their resulting state) are private to the specified parties (the one that sourced the transaction and the destination ones). No other party can disclose the transactions or their resulting state, as all the information exchanged is cyphered in a way only the involved parties can decrypt. In order to observe the compatibility and basic operation of Ethereum, a hash of the transaction payload is registered in a block of the common chain, but no information can be disclosed by other parties from that hash; it is only useful to the authorised parties for indexing

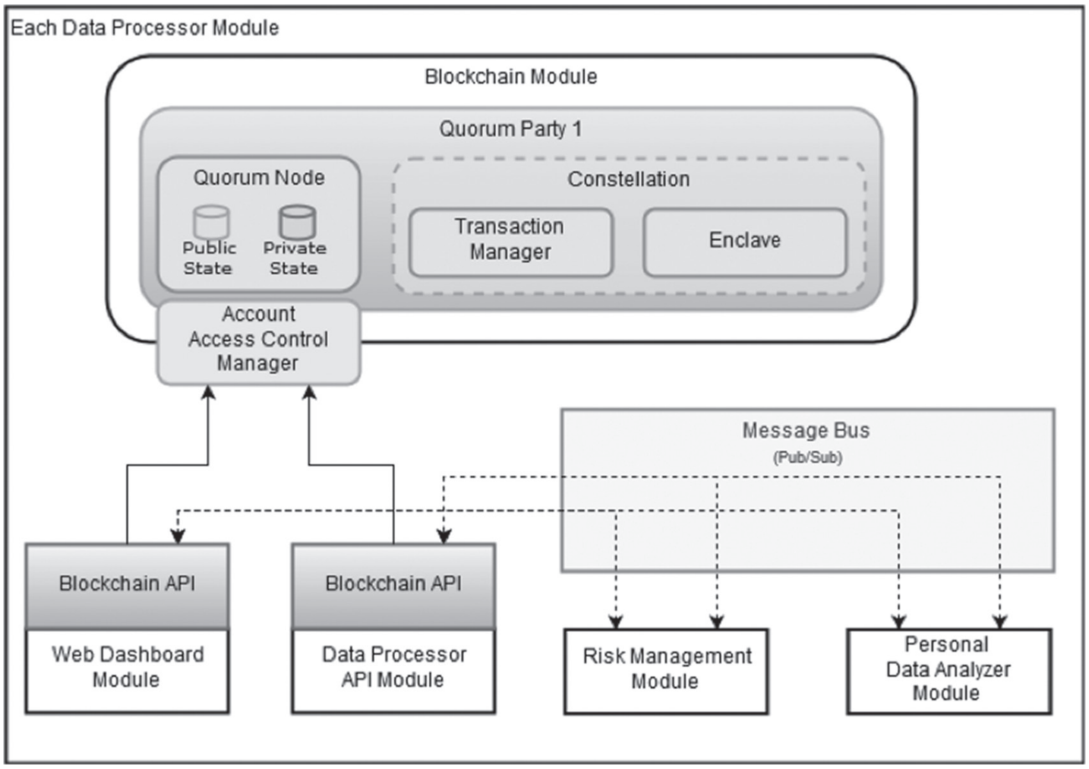

Figure 3: Blockchain module architecture
Note: API, application programming interface.

the transaction and checking its integrity. In this way, no private information is (or can be) registered on the common chain or the public state. A separate private database is available in every Quorum node to store the private state that refers to the private information that pertains to that party.

Private contracts and transactions refer to a set of parties. A party can be understood as an organisation and its related infrastructure (basically a Quorum node and a Constellation in charge of the process of the private transaction). By restricting the user accounts that can access the party (that is, by defining and guarding an organisation), privacy is transferred to final users.

Organisations participating in the PoSeID-on's fully private permissioned blockchain network are conceptual entities who have permission to maintain pieces of information and interact with other organisations through the distributed ledger.

These organisations are responsible for maintaining their Quorum party infrastructure, which executes the smart contracts, maintains the blockchain and updates the public and private states.

Quorum manages the privacy of transactions between parties, but the association between parties and user accounts is out of the scope of the Quorum itself. Thus, in order to support the concept of organisation as referred to a group of authorised participants, a module that associates user accounts to organisations and manages the access of the account to the Quorum party is needed. This module is named account (access control) manager in the blockchain API architecture.

Smart contracts are used for permission management, checking and lookup. Every operation involving PII is confirmed first against the permissions present in the ledger, through the call of a smart contract. Any requests or performed operations involving PII are also recorded in the ledger. The implemented smart contracts allow for, and record, the following operations:

- PII Permission Request: Contains the requested PII type, the time of the permission expiration, a field describing the intended use for the requested PII, and the ID of the issuer of the request.
- PII Permission Grant: Contains the PII type being granted access, the ID of the data processor requesting the information, and the ID of the PII owner.
- PII Permission Revocation: Contains the revoked PII type, the ID of the data processor which had its access revoked, and the ID of the owner of the PII.
- PII Access: Contains the accessed PII type, the ID of the data processor that accessed the information, and a timestamp.
- PII Permission Check: Contains the PII type being checked for permissions, the ID of the data processor consulting the permissions, the ID of the PII owner, and returns whether the requesting data processor has access to the concerned PII.
- PII Permissions List: Contains the data subject's ID whose permissions were listed.

The smart contract implementation in combination with the blockchain API supports the consultation and writing of these ledger entries. As the identity data (external account address) is stored in the ledger in every user interaction (eg change a permission status in the PoSeID-on system), it is not possible to use a single user ID because it would result in breaking GDPR compliance, as the PII about that user would be traceable. To solve this issue, in PoSeID-on we have implemented a mechanism to create a pool of pseudo-identities for each data subject user that can be erased on request. When the pseudo-identities are deleted by the data subject, the link between two given permissions ceases to exist, a permission history over the time cannot be created, the data controller owning the data stored on the ledger is not known and the data subject addresses are forgotten by the PoSeID-on.

B. Dashboard

The web-based privacy-enhancing dashboard (from here on, PED), whose architecture is presented in Figure 4, is the application responsible for the data subject's user interface to the PoSeID-on system. Secondly, the PED is also an information source to the platform administrator.

As the name suggests, the PED is completely web based. This means that in order to use the application, the user will visit the application's URL (uniform resource locator) with a supported user agent. At the time of writing, these user agents include Mozilla Firefox 64.0, Google Chrome 71.x and Apple Safari 12.0.2. For maximum usability, the web application will be designed in such a way that it supports the same functionality across multiple devices with a mobile-first approach. In addition, in accordance with EU directive 2016/2102,[14] the application needs to follow strict accessibility guidelines to further increase usability across the EU population.

The data subject and the administrator use their user agent to connect to the PED. This connection uses transport layer security (TLS), and all non-TLS connections will be redirected to TLS. Through a reverse proxy, the user agent loads all the static content needed for the web application from the web application front-end file server. The client-side code will then start to communicate with the web application back-end through the reverse proxy. The web application back-end is responsible for authentication and authorisation, offering compatibility with eIDAS.

All of the business logic of the PED resides in the web application back-end. This includes communication with the other PoSeID-on components. The web

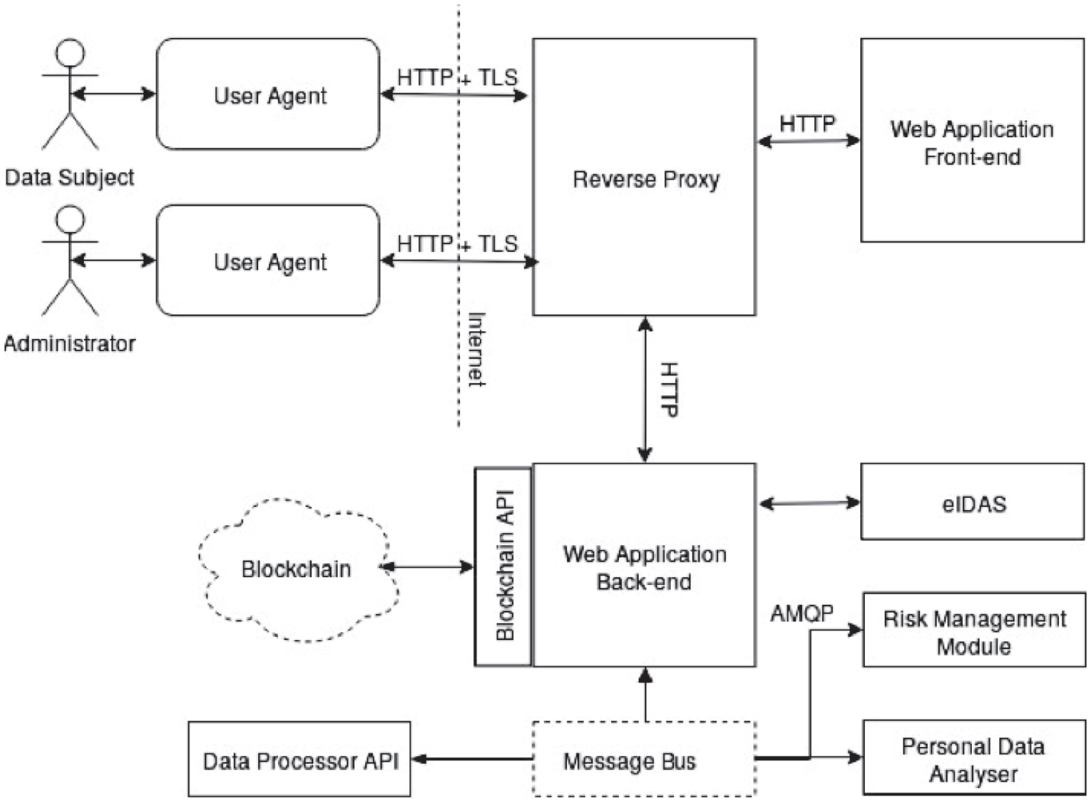

Figure 4: Dashboard architecture
Note: AMQP, advanced message queuing protocol; API, application programming interface; eIDAS, electronic identification and trust services; HTTP, hypertext transfer protocol; TLS, transport layer security.

application back-end is a service, written in the Python programming language, which exposes an authenticated HTTP GraphQL[15] API for consumption in the web application front-end. This API allows the web application front-end to use the back-end's functionality.

The web application front-end is the part of the PED that is sent to the user agent and runs on the client side. It consists of a graphical user interface built with web technology (HTML [hypertext markup language], CSS [cascading style sheet], images) and application code suitable for running in the user agent (primarily JavaScript). The application code contains a client to the authenticated GraphQL API exposed by the web application back-end.

The HTTP (hypertext transfer protocol) reverse proxy sits in between the services that make up the PED and the internet, on which the platform's users connect to the platform. The web application back-end and the web application front-end do not directly communicate with the users over the internet and are, thus, not directly available.

C. Risk management module

As mentioned in Conceptual Architecture, the RMM module is used to evaluate and manage privacy and operational risks within the PoSeID-on system, through analysis of operational logs and PII exchanges, and manages a risk score associated with each data processor. This can be used to advise on which service should eventually be disabled in case of anomalies and high exposure to privacy risks.

The anomaly detection approach taken is based on system log analysis[16–18] and follows the well-described framework by M. Lyu et al.[19] This framework is comprised of four main steps — log collection, log parsing, feature extraction and anomaly detection.

- **Log Collection:** Logs in the PoSeID-on are delivered through the message bus directly to the RMM, using a custom message protocol. The messages will contain the log, structured according to the Graylog extended log format (GELF)[20] as Graylog is the chosen log management solution. In addition to this, the message will also contain extra fields if it involves operations regarding PII within the PoSeID-on. These fields contain information such as the data subjects and data processors involved and the types of PII requested or exchanged. As this is also considered PII, it will only be added to the message when consent is given in order to enable a more complete analysis.
- **Log Parsing:** As logs consist of free form text, it is necessary to parse them and extract a group of structured event templates. The event templates consist of the most common combination of log parts/segments, which therefore define the constant part of a log. Once a set of event templates is found, each subsequent log can then be considered as an occurrence of one of the derived events or can create a new event template. For the log parsing step, Drain,[21] a state-of-the-art online log parsing approach based on fixed depth trees, was implemented as recent benchmarks have proven its superiority in comparison to other available open source alternatives; it has also been successfully deployed in a production environment at Huawei.[22] In order to integrate Drain into the RMM, a Java implementation of the algorithm was developed, to fully integrate it in the Spark Streaming[23] pipeline.
- **Feature Extraction:** Once logs have been parsed, it is necessary to create numerical feature vectors, which will be provided to the ML models performing the anomaly detection. First, parsed logs are grouped into windows over which we count the number of occurrences of each event. This step is already included in Drain's log parsing algorithm, which provides as a result the set of events and their respective number of occurrences. In the initial phase of the module's development, only temporal windows are considered, but as the development progresses, grouping logs by source or syslog severity levels will be tested in order to evaluate the best approach.
- **Anomaly Detection:** For the anomaly detection step, in the initial phases of the RMM deployment, unsupervised learning models are more favourable due to the lack of operational data, such as K-Means and principal component analysis, which do not require previous labelled data. In order to implement the anomaly detection step MLlib,[24] Spark's ML library is being used. Further testing is necessary in order to select the best unsupervised learning algorithm for the context of the PoSeID-on. Once in production, the RMM can also be extended with other semi-supervised or supervised models based on the results of the unsupervised anomaly detection and manual labelling of the collected datasets. Once anomalies are identified, data subjects and data processors present in the anomalous window will be notified along with the PoSeID-on administrators.

The potential number of data subjects managed by the PoSeID-on is in the order of millions. As such, the risk management module must be designed with performance and scalability in mind. The RMM's architecture (Figure 5) is based on the proposal by Nathan Marz for a generic and scalable data processing architecture, which he called Lambda Architecture.[25] As literature shows,[26–28] adopting the lambda architecture for event processing allows for an efficient processing of high volumes

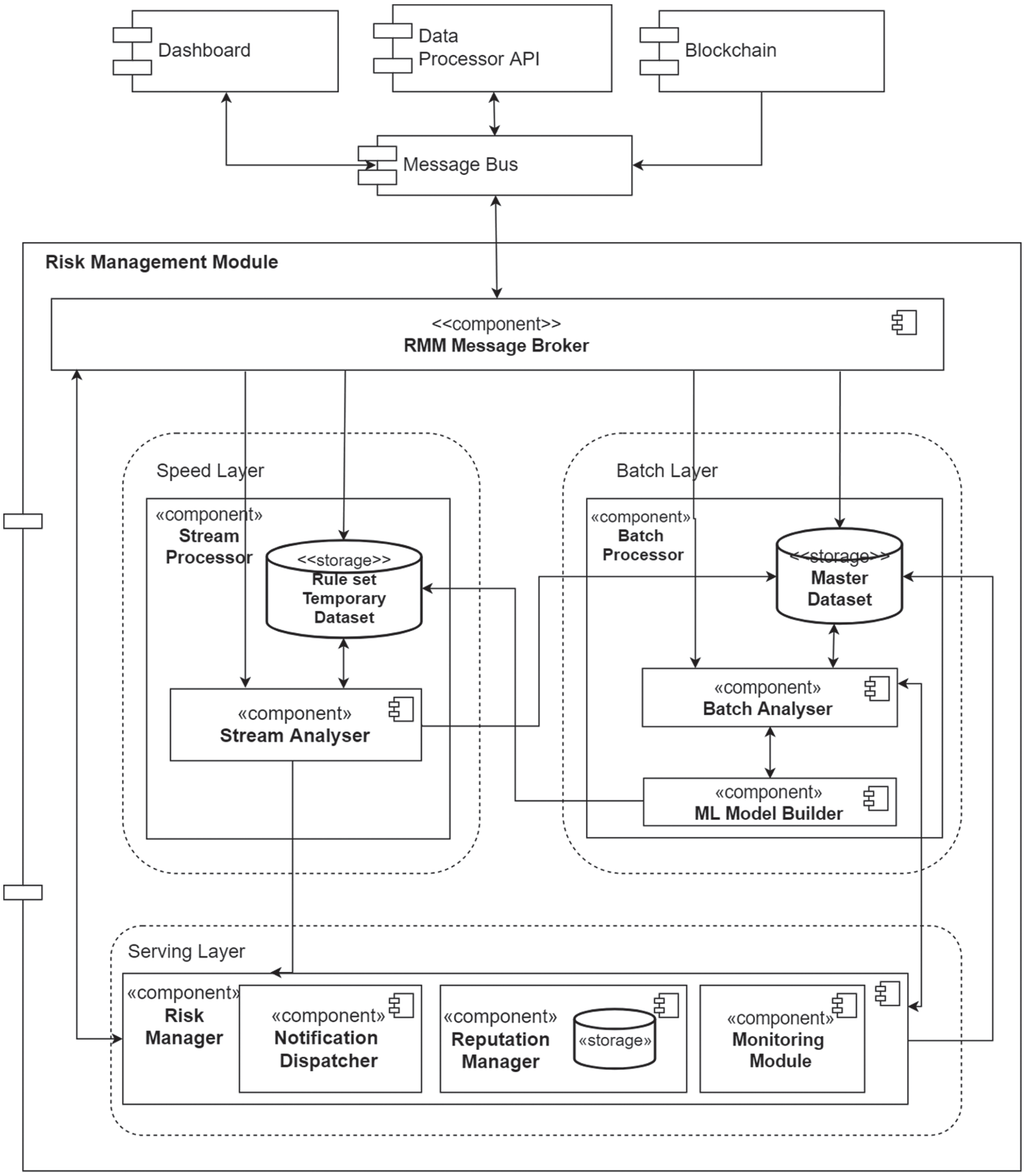

Figure 5: Risk Management Module architecture
Note: API, application programming interface; ML, machine learning.

of information, consistent with the ones expected by the PoSeID-on platform. It allows the analysis of data in near real-time, as it arrives at the RMM, providing data subjects a near instant feedback regarding exposure risks, while also allowing the analysis and extraction of valuable insights from large volumes of data collected over time through more complex or time- and resource-intensive methods.

This architecture can be divided into three layers: a batch layer, a speed layer and a serving layer. As messages containing system operational logs and blockchain transactions arrive at the risk management module, these messages are pre-processed and then dispatched to both the batch and the speed layers.

The batch layer manages a master dataset that is used for historic risk analysis, taking advantage of having a large collection of data, where the time span will depend on how long the RMM is allowed to retain data, provided the data subject has given explicit consent for this purpose. This historic risk analysis takes significant time and is not expected to work in near real-time. Instead, this layer provides warnings in case of risk detection in an 'offline' manner. Having a batch layer also provides the advantage of being able to use ML models that do not have a stream-based counterpart. This layer is also in charge of training and updating ML models for use in near real-time analysis by the speed layer for algorithms that require an offline training step, as is the case with most supervised learning algorithms.

The speed layer deploys the models created by the batch layer in addition to the clustering models and analyses the stream of data in real-time, using the data that arrives between batch analysis. This layer reasons over the data, and in case an anomaly is detected, it dispatches a recommendation action request to the serving layer.

The serving layer is in charge of receiving the results from both the batch and speed layers and notifying the respective entities about risk exposure through the dashboard interface. It is also in charge of receiving feedback from administrators regarding such risk notifications by providing the identification of the log window where the anomaly was identified and allowing them to confirm or deny that a real risk is present within it. According to the administrator's feedback and the history of anomalous logs involving each data processor, the associated risk reputation of data processors is also updated.

The risk management module is deployed using container images compliant with the open containers initiative (OCI)[29] and will be horizontally scalable by leveraging Spark's distributed computing capabilities.

D. Personal data analyser

The PDA monitors PII transactions where explicit consent is provided in order to detect and assess privacy risks. Figure 6 presents the PDA module architecture and displays its inner components and interactions with external modules. The module interacts with the message bus, which, in turn, communicates with the dashboard and the data processor API.

The requests arriving from the message bus to the request processor are processed and dispatched to other components according to the data type (eg direct RPC (remote procedure call) inputs, PDF (portable document format), TXT (text) or other types of structured information). The metadata extractor retrieves all the associated metadata information and feeds it to the next component. The internal parsers are intended to extract information and all data available inside each file or data structure being analysed. The files can be any structured information, PDF, XML (extensible markup language), HTML, URL, CSV (comma-separated values) or TXT.

When all the information is extracted in the previously described components, it is then necessary to derive meaning from it. For that purpose, the natural language reasoning (NLR)/natural language understanding (NLU) processing unit performs semantic and syntactic recognition to allow the identification and classification of named entities such as persons or places. This step is supported by state-of-the-art natural language processing (NLP) tools like Stanford CoreNLP[30] and spaCy,[31] which are then fine-tuned and improved for this particular context. At this stage, in addition to further training and improving the default models offered the spaCy, we are training new and specific models for the named

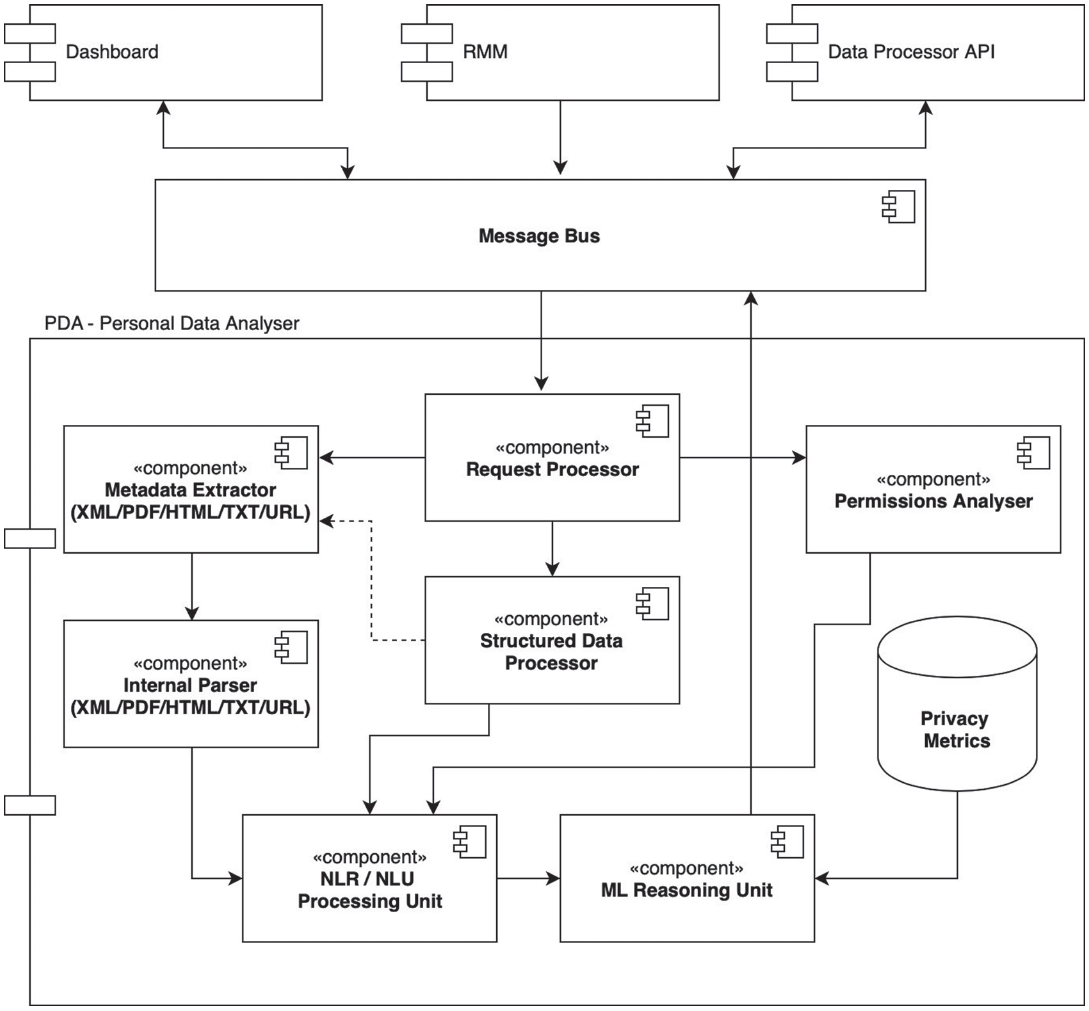

Figure 6: Personal Data Analyser architecture
Note: API, application programming interface; HTML, hypertext markup language; ML, machine learning; NLR, natural language reasoning; NLU, natural language understanding; PDF, portable document format; RMM, risk management module; TXT, text; URL, uniform resource locator; XML, extensible markup language.

entities we aim to analyse. The training can be achieved by feeding the models with intended datasets and respective labels (eg the Groningen Meaning Bank dataset[32]). Assuming that the permissions also contain a description of their purpose, this component also receives as input the descriptions of the permissions associated with each use case.

The final stage of the PDA flow is the ML reasoning unit. In this component, ML algorithms determine whether or not privacy risks are at stake. This is performed with a combination of available data: the validation of transaction PII, the existence or not of previously non-identified PII, the reputation of involved data processors and level of sensitiveness of each PII attribute (eg sensitive or quasi-identifier). Privacy metrics and user permissions are considered in the reasoning process. As privacy metrics rely on large amounts of data, the metrics[33] to be applied depend on the amount of data each transaction holds. With the implementation of the first version of the module, experiments are being performed with the aim of setting privacy thresholds, which can be fine-tuned by administrators. When such privacy thresholds are crossed, response messages are generated and sent to the message bus in order to notify the concerned parties. At the end of each personal data analysis execution, all the data is discarded.

The current version of the PDA (developed in Python version 3) provides a functional NLP pipeline capable of identifying named entities such as persons' names, cities, countries, locations, birth dates and others. It is currently using a hybrid approach where NLTK,[34] Stanford CoreNLP, spaCy and regular expressions are used in conjunction to find PII. The main steps of the NLP pipeline are tokenisation, part of speech tagging, named entity recognition. Recurring to the Python library Pika[35] to interact with RabbitMQ, the PDA is fully integrated with the message bus and successful delivery and validation of messages were achieved between the RMM and PDA. Recent work allowed completing the communication integration with the dashboard and data processor API as well as the release of an OCI-compliant container image of the module.

E. Message bus

The message bus is a middleware that leverages communications between the PoSeID-on system components, in a decoupled fashion, allowing for the easy addition and removal of system components and their communication. In a generic sense, a message bus is a combination of a data model for structuring system messages and communications, a common command set that specifies the available operations on messages and between connected components, and a messaging infrastructure in charge of providing the set of functionalities associated with the command set. Given the preceding description, it is not expected that a message bus is in itself developed by the PoSeID-on, as there are several very good solutions providing message bus functionality already available. In this sense, what is expected is that all components communicating through the message bus implement the necessary APIs provided by one of these solutions in order to communicate through the chosen message bus platform.

One possible solution for providing the message bus functionality for the PoSeID-on system is Apache Kafka.[36] The components described here are the high-level abstractions provided by the Kafka platform. Kafka is a streaming platform that provides the following capabilities:

- Publish and subscribe to streams of messages, which are similar to message queues.
- Store streams of records in a fault-tolerant and persistent way.
- Process streams of records as they occur.
- High availability.

As described in the Kafka documentation, Kafka runs as a cluster on one or more servers, spanning one or more data centres. For high availability, it is recommended to have at least three Kafka servers and three Apache Zookeeper[37] servers. The Kafka cluster stores streams of messages, here called records, in categories called topics, which can act as traditional queues. Besides this, Kafka separates the entities connected to its platform into two different categories: producers and consumers. Thus, Kafka is suitable for the PoSeID-on's needs for inter-module communication, that comprise asynchronous communication, efficiency and dependability.

Another possibility of providing the message bus functionality is RabbitMQ.[38] It provides functionalities similar to Kafka's, which are very well suited to the PoSeID-on. RabbitMQ requires only 3 or more servers to provide high availability. It is also a lossless mechanism and supports RPC. The node performance of the two options are, on average, 20,000 msg/sec for RabbitMQ and 100,000 msg/sec for Apache Kafka. Despite the difference, both options are suitable for

PoSeID-on. Given that both provide the performance that PoSeID-on needs and that RabbitMQ requires less hardware and development complexity, the message bus is currently being developed using RabbitMQ.

All the messages flowing through the message bus are serialised and duly encrypted. The serialisation uses Google's protocol buffer[39] due to its simplicity, speed and efficiency. To guarantee the security and privacy of the messages, the PoSeID-on employs a custom message protocol, which uses Libsodium software library[40] for encryption, decryption and signing. Therefore, all messages are signed with a module-specific ED25519[41] key and encrypted using the respective Curve25519[42] key. This a requirement applied to the messages of all modules integrated in PoSeID-on.

VII. CONCLUSION

In this paper, we have seen the main architectural aspects of an approach to the protection and control of transactions dealing with personal data, in multi-user and multi-processor environments. The approach was developed in the context of the PoSeID-on EU project for GDPR-compliant public and private services delivery and explores the use of blockchains as a base platform. In addition to providing conceptual and functional views, the paper identified several challenges that are being addressed. Besides the implementation of the basic platform functionality and subsequent validation in four pilots, our current work is addressing (i) the development of advanced permission-handling approaches; (ii) technologies, frameworks and algorithms for risk management; and, last but not least (iii) contents analysis, natural language processing, sentence identification and speech tagging for personal data analysis. The innovative features of the proposed architecture and platform open a variety of challenges and opportunities, of which the deployment of the platform itself in real, operating environments is the main one. In addition to finalising its development and to extensive pilot testing, deployment will be the main subject of future work.

AUTHOR'S NOTE

The work presented in this paper was partially carried out in the scope of the PoSeID-on project, grant agreement number: 786713. H2020-DS2016-2017/DS-08-2017.

References

1. Council, E. (2016) 'EU Regulation 2016/679 of the European Parliament and of the Council, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)', *Official Journal of the European Union*,Vol. 59, pp. 1–88.
2. Stallings, W. (2017) 'A blockchain tutorial', *The Internet Protocol Journal*,Vol. 20, pp. 795–820.
3. Underwood, S. (2016) 'Blockchain beyond Bitcoin', *Communications of the ACM*,Vol. 59, pp. 15–17.
4. Yaga, D., Mell, P., Roby, N. and Scarfone, K. (2018) 'NISTIR 8202 — blockchain technology overview', Computer Security Resource Center, available at: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf (accessed 2nd November, 2019).
5. 'PDP4E project: Methods and tools for GDPR compliance through privacy and data protection engineering', available at: https://www.pdp4e-project.eu/ (accessed 2nd November, 2019).
6. BPR4GDPR project, 'Business process re-engineering and functional toolkit for GDPR compliance', available at: https://cordis.europa.eu/project/rcn/214871/factsheet/en (accessed 2nd November, 2019).
7. 'DEFeND project: 'The data governance framework for supporting GDPR', available at: https://www.defendproject.eu (accessed 2nd November, 2019).
8. SMOOTH project, 'GDPR compliance cloud platform for micro enterprises', available at: https://smoothplatform.eu (accessed 2nd November, 2019).
9. 'PAPAYA: PlAtform for PrivAcY preserving data Analytics', available at: https://www.papaya-project.eu (accessed 2nd November, 2019).
10. 'Protection and control of Secured Information by means of a privacy enhanced Dashboard', available at: https://www.poseidon-h2020.eu (accessed 2nd November, 2019).
11. Council, E. (2014) 'EU Regulation No. 910/2014 of the European Parliament and of the Council, on electronic identification and trust services for electronic transactions in the internal market and

repealing Directive 1999/93/EC', *Official Journal of the European Union*, Vol. 57, pp. 73–114.

12. Quorum, 'Advancing blockchain technology', available at: https://github.com/jpmorganchase/quorum/blob/master/docs/Quorum%20 Whitepaper%20v0.2.pdf (accessed 29th November, 2019).
13. Ethereum, 'Blockchain app platform', available at: https://github.com/ethereum/wiki/wiki/White-Paper (accessed 29th November, 2019).
14. Council, E. (2016) 'Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies', *Official Journal of the European Union*.
15. GraphQL, 'A query language for your API', available at: https://graphql.org (accessed 2nd November, 2019).
16. Astekin, M., Zengin, H. and Sözer, H. (2018) 'Evaluation of distributed machine learning algorithms for anomaly detection from large-scale system logs: A case study', *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2071–2077.
17. Lin, Q., Zhang, H,. Lou, J., Zhang, Y. and Chen, X. (2016) 'Log clustering based problem identification for online service systems', *IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)*, May, pp. 102–111.
18. Liu, Z., Qin, T., Guan, X., Jiang, H. and Wang, C. (2018) 'An integrated method for anomaly detection from massive system logs', *IEEE Access*, Vol. 6, pp. 30602–30611.
19. He, S., Zhu, J., He, P. and Lyu, M. R. (2016) 'Experience report: System log analysis for anomaly detection', *IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*, October, pp. 207–218.
20. Google. 'Graylog extended log format — GELF', available at http://docs.graylog.org/en/3.0/pages/gelf.html (accessed 2nd November, 2019).
21. He, P., Zhu, J., Zheng, Z. and Lyu, M. R. (2017) 'Drain: An online log parsing approach with fixed depth tree', *IEEE International Conference on Web Services (ICWS)*, Jun, pp. 33–40.
22. Zhu, J., He, S., Liu, J., He, P., Xie, Q., Zheng, Z., et al. (2018) 'Tools and benchmarks for automated log parsing,' *CoRR*, abs/1811.03509.
23. 'Apache Spark is a unified analytics engine for large-scale data processing', available at https://spark.apache.org/streaming/ (accessed 2nd November, 2019).
24. 'MLlib is Apache Spark's scalable machine learning library', available at https://spark.apache.org/mllib/ (accessed 2nd November, 2019).
25. 'Lambda Architecture', available at http://lambda-architecture.net (accessed 2nd November, 2019).
26. Yamato, Y., Kumazaki, H. and Fukumoto, Y. (2016) 'Proposal of lambda architecture adoption for real time predictive maintenance', *Fourth International Symposium on Computing and Networking (CANDAR)*, November, pp. 713–715.
27. Casas, P., Soro, F., Vanerio, J., Settanni, G. and D'Alconzo, A. (2017) 'Network security and anomaly detection with Big-DAMA, a big data analytics framework,' *IEEE 6th International Conference on Cloud Networking (CloudNet)*, September, pp. 1–7.
28. Rosa, L., Proença, J., Henriques, J., Graveto, V., Cruz, T., Simoes, P, et al. (2017) 'An evolved security architecture for distributed Industrial Automation and Control Systems', available at: https://books.google.pt/books?id=uFA8DwAAQBAJ (accessed 2nd November, 2019).
29. 'Open Container Initiative', available at: https://www.opencontainers.org/ (accessed June 2019).
30. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McClosky, D. (2014) 'The Stanford CoreNLP natural language processing toolkit', *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.
31. spaCy, 'Industrial-strength natural language processing', available at: https://spacy.io (accessed May 2019).
32. Groningen Meaning Bank (GMB) 'A free semantically annotated corpus that anyone can edit!' available at: https://gmb.let.rug.nl/data.php (accessed May 2019).
33. Wagner, I. and Eckhoff, D. (2018) 'Technical privacy metrics: A systematic survey', *ACM Computing Surveys,* Vol. 51, No. 57, pp. 1-57:38.
34. Bird, S., Klein, E. and Loper, E. (2009) 'Natural language processing with Python', USA, O'Reilly Media.
35. 'Pika — a RabbitMQ (AMQP 0-9-1) client library for Python', available at: https://github.com/pika/pika (accessed June 2019).
36. 'Apache Kafka Distributed Streaming Platform', available at: https://kafka.apache.org (accessed January 2019).
37. 'Welcome to Apache ZooKeeper', available at: https://zookeeper.apache.org/ (accessed January 2019).
38. 'RabbitMQ — Open source message broker,' available at: https://www.rabbitmq.com/ (accessed January 2019).
39. 'Google. Protocol Buffers', available at: http://code.google.com/apis/protocolbuffers/ (accessed May 2019).
40. 'Introduction', available at: https://download.libsodium.org/doc/ (accessed May 2019).
41. Bernstein, D. J., Duif, N., Lange, T., Schwabe, P. and Yang, B.Y. (2012) 'High-speed high-security signatures', *Journal of Cryptographic Engineering*, Vol. 2, pp. 77–89.
42. Bernstein, D. J. (2006) 'Curve25519: New Diffie-Hellman speed records', *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*), available at: https://link.springer.com/content/pdf/10.1007%2F11745853_14.pdf (accessed 2nd November, 2019).