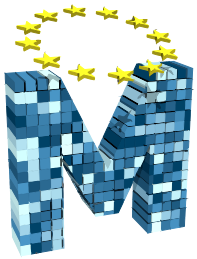

Project title: Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control

Project acronym: MOSAICrOWN

Funding scheme: H2020-ICT-2018-2

Topic: ICT-13-2018-2019

Project duration: January 2019 – December 2021

D2.2

Report on Requirements, Research Alignment and Deployment Plan

Editors: Megan Wolf (MC)
Michele Mazzola (MC)
Reviewers: Stefano Paraboschi (UNIBG)
Pierangela Samarati (UNIMI)

Abstract

This deliverable builds on the requirements defined in D2.1 to provide an overall status update on progress made towards meeting the needs of MOSAICrOWN's strategic use cases. It will provide high-level updates using a Use Case lens on progress made in Work Packages 3 through 5, determine any gaps that need fulfillment before moving forward, and potential solutions or creative ideas to further the progress of the requirements. This document will serve as a roadmap for MOSAICrOWN's progress moving forward in the coming months.

Type	Identifier	Dissemination	Date
Deliverable	D2.2	Public	2020.06.30

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825333.

MOSAICrOWN Consortium

1.	Università degli Studi di Milano	UNIMI	Italy
2.	EMC Information Systems International	EISI	Ireland
3.	Mastercard Europe	MC	Belgium
4.	SAP SE	SAP SE	Germany
5.	Università degli Studi di Bergamo	UNIBG	Italy
6.	GEIE ERCIM (Host of the W3C)	W3C	France

Versions

Version	Date	Description
0.1	2020.06.04	Initial Release
0.2	2020.06.25	Second Release
1.0	2020.06.30	Final Release

List of Contributors

This document contains contributions from different MOSAICrOWN partners. Contributors for the chapters of this deliverable are presented in the following table.

Chapter	Author(s)
Executive Summary	Megan Wolf (MC)
Chapter 1: Introduction	Megan Wolf (MC)
Chapter 2: Use Case 1 (EISI)	Aidan O'Mahony (EISI)
Chapter 3: Use Case 2 (MC)	Megan Wolf (MC)
Chapter 4: Use Case 3 (SAP SE)	Jonas Boehler (SAP SE), Benjamin Weggenmann (SAP SE)
Chapter 5: Conclusions	Megan Wolf (MC)

Contents

List of Figures

Executive Summary

With an increasing focus and need for securitized data processes around the world, the creation of a comprehensive, multi-owner, data protection focused platform would greatly impact and improve the state of data analytics. MOSAICrOWN, a coalition of academic and business partners collaborating to create a broadly relevant and applicable data protection process, has made significant progress towards creating such a platform since this deliverable's counterpart, "Requirements from the Use Case" (D2.1).

This deliverable focuses mainly on progress made towards meeting the requirements outlined in D2.1. As the document progresses, specific updates for each use case will be provided by the use cases' business owner. Then, using the use case lens, a clear progress report will be provided detailing the steps taken to meet the requirements (previously defined in D2.1) in each Work Package. Each Work Package focuses on a different, key area of MOSAICrOWN: data governance, data wrapping, and data sanitization. Without clear requirements in these three areas, MOSAICrOWN will be unable to move forward in creating an effective and successful data protection platform.

Deliverable D2.2 builds on what was presented in D2.1. The three use cases, defined and overseen by their respective business owners, provide a wide variety of scenarios for MOSAICrOWN's data protection platform. Use Case 1, defined by EISI, focuses on Intelligent Connected Vehicles (ICV) and, more specifically, how sensitive data is passed from a charging station to the data market. Use Case 2, lead by MC, focuses on transaction-level financial data and the importance of data wrapping techniques. Finally, Use Case 3, overseen by SAP SE, takes on a more broad set of operational and experience data used for consumer analytics while considering a cloud-based data storage approach.

While each use case varies from the next, MOSAICrOWN's previous research has determined that the underlying processes to ensure data protection for each is very similar - the key components described above (governance, wrapping, sanitization) are critical for each use case's overall success. This deliverable will provide general updates on the requirements and determine key next steps in creating a platform, namely the gaps necessary to close before beginning the research prototype phase.

1. Introduction

MOSAICrOWN sits at the intersection of academia and industry to provide effective data protection techniques for data markets. Through providing effective techniques for data governance, wrapping, and sanitization, MOSAICrOWN will enable an increase in adoption of data protection concepts and strengthen the competitive position of European industry and the European leadership in promoting and sustaining data protection. Thus, to meet the objectives above and requirements previously detailed in "Requirements from the Use Cases" (D2.1), MOSAICrOWN plans to create a collection of policies, procedures and products that will ensure protection to data in storage and processing, via data wrapping and sanitization, regardless of the data owner status or size. By doing this, MOSAICrOWN will ensure that all data, no matter the source, are properly controlled and maintained, thus significantly decreasing the threat of data breach.

1.1 Structure of this Document

This document begins with this chapter, which will outline the overall purpose of the document as well as a review of the Work Packages and Requirements. Chapters 2 through 4 cover updates regarding each use case, respectively. Each chapter is structured the same - it will begin with a technical overview of the Use Case, move to a discussion of specific and relevant aspects of WPs 3-5 as they relate to the use case, followed by suggested technologies or tools to fill any gaps remaining from the WPs, and concluding with a review of requirements and their statuses. The deliverable will conclude with a chapter summarizing the general findings and providing a status update on what is to come.

1.2 Purpose of this Document

The goal of Work Package (WP) 2 is to coordinate the use cases considered in the project, provide requirements, deployment and validation of MOSAICrOWN solutions, and enable direct exploitation by the industrial partners. This specific document's goal is to provide updates on requirement progress for each use case and detail any further needs or gaps that have not yet been addressed by WPs 3 through 5. For the original list of requirements, along with a summary of previous progress, please see D2.1. This document will build on D2.1 and provide an update on each individual Use Case. Figure 1.1 visualizes the interaction between WP2 and WPs 3 through 5 in MOSAICrOWN, which may be useful to review before continuing with the document.

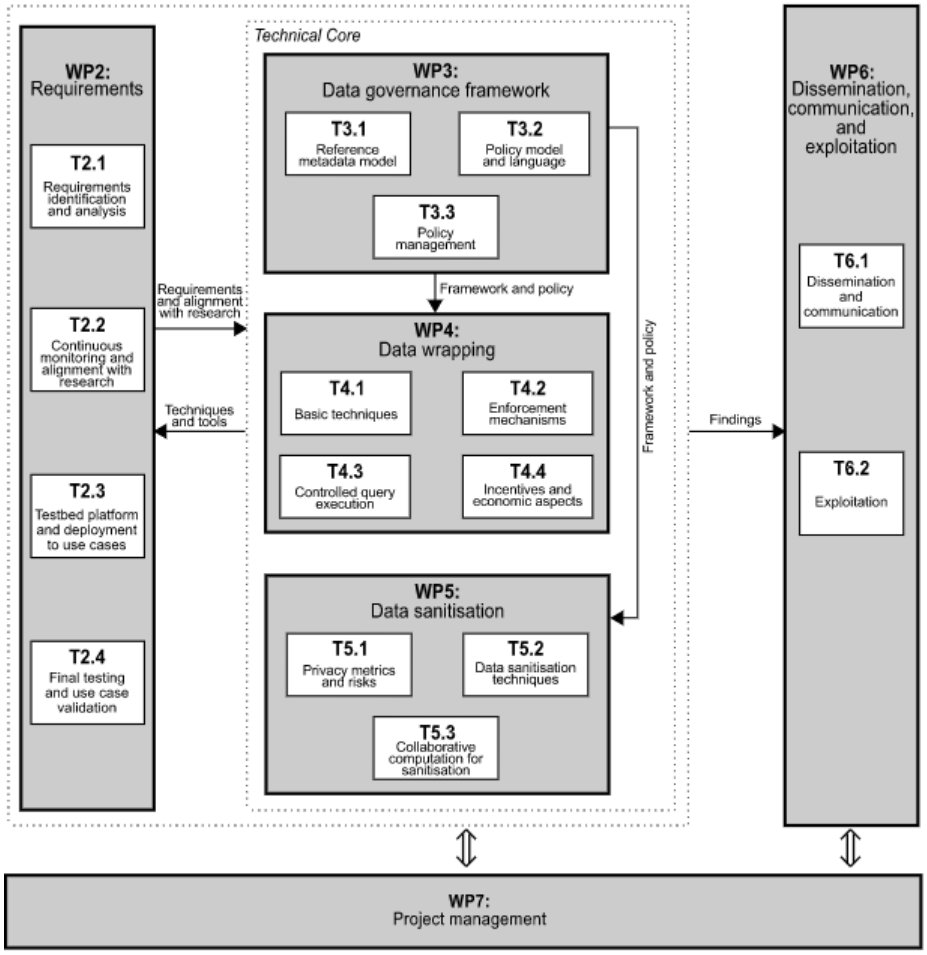

Figure 1.1: Interactions between WP2 and other WPs within MOSAICrOWN

1.3 Work Package Summaries

Before moving into the following chapters, it is important to provide a brief summary of each Work Package, as they will be referenced frequently throughout. While these summaries will be short, further information can be found regarding the Work Packages in either D2.1 or in the Work Packages themselves.

- **Work Package 3 - Data Governance Framework** WP3, led by UNIMI, focuses mainly on creating a data governance framework. This includes a policy model and language, as well as the overall structures and processes necessary to ensure proper data storage and access.
- **Work Package 4 - Data Wrapping** WP4, led by UNIBG, focuses on creating data wrapping solutions to ensure effective protection of data stored/processed in the data market. A key feature of the working of WP4 is the flexibility of applying wrapping techniques without sacrificing policy adherence.
- **Work Package 5 - Data Sanitization** WP5, led by SAP SE, focuses on data sanitization techniques. WP5 aims at designing techniques for producing sanitized versions of the data for the privacy-preserving use, sharing, and computation with other parties.

1.4 Requirements Analysis Methodology

As a brief refresh from information covered in D2.1, requirements for each Use Case were determined with specific business scenarios and actors in mind. Each requirement is uniquely identified with a meaningful tag allowing an overall requirements catalogue to be derived. A tag has the structure REQ-UC*x*-*yn*, where $x \in \{1, 2, 3\}$ identifies the use case (UC), *y* groups the requirements by dimension, phase or topic (e.g., DI for data ingestion) and *n* is a number (as more than one requirement per group *y* is possible). It is this deliverable's goal to provide an update on the alignment of the WPs with each requirement and provide a clear path to completion, highlighting any gaps or areas of improvement that need to be addressed in the coming months before research prototyping begins.

These status updates are provided by the business owner of each use case, and thus are not standardized. The statuses are set based on the business owners' understanding of their use case, the progression made towards addressing each requirement in WPs 3 through 5, and the overall progress they would expect in the following months. MOSAICrOWN can use this deliverable to continually improve their findings and ensure that all requirements are met in a timely and effective manner.

2. Use Case 1 (EISI)

Use Case 1 (UC1) is focused on the protection of sensitive data in an Intelligent Connected Vehicle (ICV) with a focus on how to exchange data in such a fashion that ICV electric vehicle fleet owners and Electric Vehicle (EV) charging infrastructure owners can derive mutually beneficial insights into the status of the EV charging infrastructure in a privacy preserving fashion.

The primary focus of UC1 is that of data governance, i.e, application of policy, metadata management, data management, access control management, data wrapping, data sanitization. The goals of UC1 are to develop tools for:

- Oversight of both streaming and batch ingestion both of single sources and multiple concurrent sources
- Sanitizing certain key fields
- Allowing data owners to describe in a clear language how MOSAICrOWN should govern their data (e.g., access control, data tracking, data removal)
- Facilitation of electric vehicle fleet analytics

At this stage of MOSAICrOWN we discuss the alignment with the requirements, research, and deployment in terms of the deliverables up to this point.

2.1 Technical overview

The automotive industry is beginning to realize the potential of the ongoing period of transformation regarding connected technologies and the volume, variety, and velocity of data being generated by ICVs and also autonomous vehicles. There are many potential benefits to using a privacy-preserving data market including improved traffic management, monitoring the state of the road infrastructure, and efficient usage of EV charging point stations. Figure 2.1 illustrates how UC1 will interact with MOSAICrOWN. The stages presented in Figure 2.1 are as follows:

1. Ingestion of ICV data into MOSAICrOWN ingestion interface
2. Ingestion of EV charging station data into MOSAICrOWN ingestion interface
3. Presentation of data to MOSAICrOWN storage phase
4. Analytics within MOSAICrOWN framework
5. Presentation of data and analytics to user

6. Ingestion of analytics data into MOSAICrOWN ingestion interface

7. Repetition of cycle through user interface

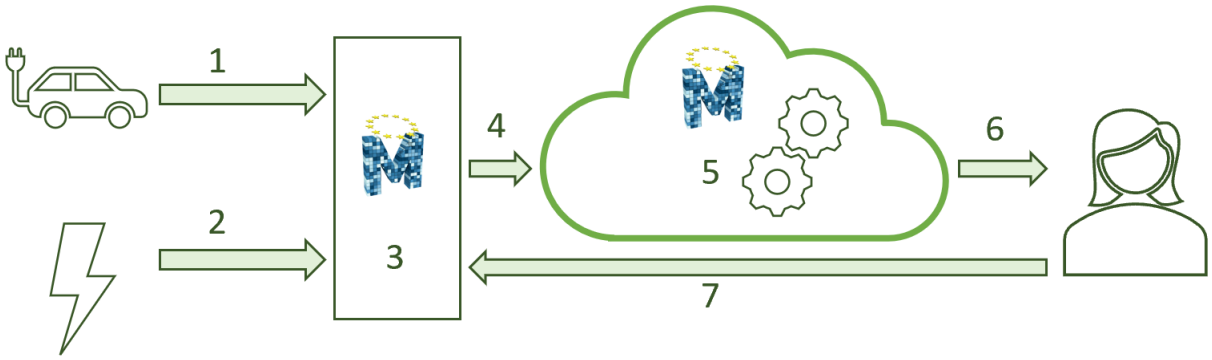

Figure 2.1: UC1 Steps

As described in D3.2 (“Preliminary version of tools for the governance framework”), the ICV data is presented to MOSAICrOWN in JavaScript Object Notation (JSON) format and as such, we have been able to take advantage of the semantic web technologies [BLHL01] and “Linked Data” technologies. There are two methods of data being presented to MOSAICrOWN, batch and streaming. UC1 leverages Ireland’s Open Data Portal [Cou18] to provide the details of electric charging points. In this instance data is being provided in a batch form in comma separated value format.

The application of policy to data is highly coupled with UC1. The policy provided by UC1 relates to enforcement of access control, storage restrictions, data sanitization, data wrapping, quality annotations, and treatment of metadata.

MOSAICrOWN is expected to provide data ingestion functionality as a service. The requirements stipulate both batch and streaming as ICVs generate significant volumes of data which ideally would not be stored in the vehicle. The data ingestion service should allow the ingestion of multiple different data types and formats in both structured and unstructured form.

UC1 requires different wrapping as well as data sanitization techniques to implement policy-based protection mechanisms. These requirements include preventing unauthorized access to properties such as the Vehicle Identification Number (VIN) and preventing linkage attacks. Also, important to note is the fact that insights gained from the data through analytics are valuable and should therefore also have data protection mechanisms for wrapping and sanitization applied to protect against unauthorized access.

With regard to sanitization, UC1 requires techniques which provide strong security and privacy guarantees where special care must be taken with regards to how to sanitize the data and protect the privacy of individuals while allowing statistical analysis.

2.2 Alignment with WPs

In this section we examine how the work packages and tasks related to governance, ingestion, data wrapping, and data sanitization are contributing to UC1. We are specifically interested in WP3

("Data governance framework"), WP4 ("Data wrapping"), and WP5 ("Data sanitization"). Within these work packages T3.1 ("Reference metadata model"), T3.2 ("Policy model and language"), T3.3 ("Policy management"), T4.1 ("Basic techniques"), T4.2 ("Controlled query execution"), T4.4 ("Incentives and economic aspects"), and T5.1 ("Privacy metrics and risk") are of relevance.

2.2.1 WP3 alignment

WP3 is assigned the role of creating the data governance framework. It achieves this by identifying tools and concepts which will satisfy the requirements of UC1, arrive at a policy model and language suitable for describing the concepts put forward by UC1, and finally provide techniques for reasoning and assisting with policy compliance. T3.1 has produced a deliverable D3.1 ("First version of the reference metadata model") which describes the vocabularies required to describe the metadata supplied by the UC1 ICV and EV charging station data. This work package has also produced deliverable D3.2 which is aligned to each of the tasks in this work package. D3.2 aligns very closely to UC1 in terms of providing an approach for specifying policies regulating data access, use, and processing.

2.2.2 WP4 alignment

Data wrapping has significant relevance to UC1 regarding protection of ICV private data. Examples of privacy related ICV data include vehicle registration plate, driver information, location and speed information. The novel technical solutions presented in Deliverable D4.2 ("Report on encryption-based techniques and policy enforcement"), and the preliminary version of the corresponding tools described in Deliverable 4.1 ("First Version of Encryption-based Protection Tools"), offer solutions for protecting sensitive ICV data stored or processed in the data market. The presented techniques as well as the tools are of relevance to UC1. Tasks 4.1 ("Basic techniques"), T4.2 ("Enforcement mechanisms"), T4.3 ("Controlled query execution"), and T4.4 ("Incentives and economic aspects") are all relevant to UC1.

2.2.3 WP5 alignment

WP5 is concerned with data sanitization. The goals regarding data privacy metrics, techniques for sanitization and anonymization, and multi-owner collaborative computation for sanitization are all relevant to UC1. Deliverable D5.2 ("First report on privacy metrics and data sanitization") which delivers methods for sanitization of large data collections operating over Apache Spark complements the data governance framework introduced in D3.2 ("Preliminary version of tools for the governance framework").

2.3 Assessment of Potential Tools and Technologies for UC1

The deliverables of WP3, in particular, appear to align closely with UC1. The requirements for close to source deployment as well as facilitating the ingestion of streaming data are crucial to UC1 and these are in development. With regard to sanitization, the tools related to privacy guarantees

delivered in D5.1 are of relevance and the tools delivered by WP4, i.e., secure query optimization, are of importance to the analytics required by UC1 in a multi-owner scenario.

2.4 Status of UC1 Requirements

Regarding UC1, there is strong emphasis on data governance-related requirements. As such, the deployment status of MOSAICrOWN is evaluated with the data ingestion, data governance, access control management, data management, and data processing requirements in mind.

Data ingestion Figure 2.2 presents, at a high level, the architecture of the data ingestion features delivered in D3.2. The ingestion is facilitated with several open source tools augmented with use case specific configurations and data transforms. This allows the ingestion of heterogeneous data formats. The transform language provided as a feature of Apache NiFi is called "JsOn Language for Transform" (JOLT). Furthermore, this approach facilitates close-to-source deployment and batch-mode ingestion.

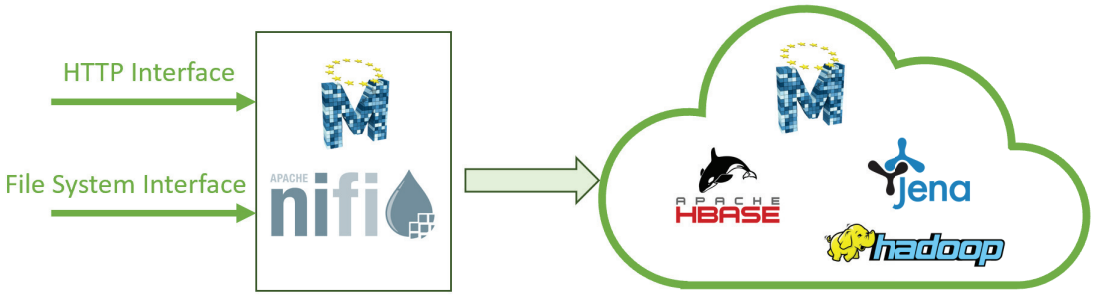

Figure 2.2: UC1 ingestion deployment

Data governance MOSAICrOWN provides an initial data governance policy language in the form of D3.3 ("First version of policy specification language and model") which results from T3.2. The approach described in D3.2 ("Preliminary version of tools for the governance framework") facilitates data owners in providing their own data governance models via the ingestion of policies represented in OWL. The Apache Jena component of Figure 2.2 provides this functionality in the deployment via an OWL API. These tasks and deliverables address a large number of UC1 requirements.

Access control management The access control management of the existing deployment enables configuration of policies for data sets by data providers via the metadata and policy ingestion. Other aspects of access control can be configured in NiFi via "Transport Layer Security" (TLS). Centralized key management is also supported by NiFi. Data sharing can be enabled in the data lake when using the database technologies deployed in D3.1. Also, the policy language described in D3.3 ("First Version of Policy Specification Language and Model") specifically addresses the requirements for supporting the definition and enforcement of access control rules with varying levels of granularity.

Data management The current deployment facilitates several data management features. Also, the current deployment facilitates extension to account for the remaining requirements, such as data at rest and data in transfer protection and data tracking. These features will be incorporated in the final platform.

Data processing This category of requirements is concerned primarily with treatment of outputs of the analytics process. While this phase of the data life-cycle is not accounted for in the current deployment, there is both a plan to enable data processing in the final platform and there are tools and techniques in WP5 concerned with the sanitization of the output of the analytics process.

2.5 Findings

Regarding UC1, the research conducted to this point in WP3 is of particular relevance and the preliminary tools delivered in D3.2 satisfy a significant number of the requirements presented in D2.1. The metadata model presented in D3.1 alongside the vocabularies provide a solid basis from which to further development in WP2 and WP3 with a goal of delivering a prototype for UC1 as required for D2.4 (“Use Case prototypes”). It is also expected that D3.3 will further facilitate the implementation of a prototype which will satisfy the requirements for UC1. Finally, with reference to WP3, the findings of D3.2 provide a foundation for delivering D3.4 (“Final tools for the governance framework”).

The tools provided by WP4 also facilitate UC1 regarding access control management in terms of allowing queries and analytics to be carried out on shared data with with users have partial access. The efficient “All-Or-Nothing Transform“ delivered in D4.1 also provides a route to satisfying requirements in the category of access control management, specifically related to access revocation to specific data sets.

The tools delivered in D5.1 (“First version of data sanitization tools”), specifically those related to sanitization in Apache Spark, complement the tools produced in WP3 and the requirements in the data processing section of the UC1 requirements. Also, the membership inference attack (MIA) library provides a solution for monitoring privacy risk and data leakage when using differential privacy, which also complements the requirements of UC1.

3. Use Case 2 (MC)

Use Case 2 (UC2) is the consideration of financial institutions in the context of confidentially and confidently sharing data. While it may seem obvious that financial data includes varying levels of personally identifiable information (PII), the less obvious aspect of UC2 is determining how to define, store, wrap and, eventually, analyze this PII. Without the ability to securely store, access and utilize this data, financial institutions lose a key part of their revenue and innovation streams.

Data Governance, Wrapping and Sanitization are key components of UC2 - without any of these three steps, the data financial institutions interact with on a daily basis would be unusable and potentially pose a threat to their clients and cardholders. As each of the WPs detail, UC2 poses interesting caveats to the stereotypical and usual approach to each component. Given the high visibility of transaction-level data and the amount of PII tied to every transaction, it truly is critical to ensure the data resulting from every swipe, tap, or click is securitized and treated with the importance it demands.

3.1 Technical Overview

As the regulatory and compliance landscapes change and data is continuously shared amongst multiple parties, the usage of personal data has increasingly narrowed and shifted power to the customer. Businesses now need to obtain explicit consent for specific usage and access of personal data. In the advent of GDPR and these new restrictions, governments, commercial enterprises and charitable organizations have been reluctant to share data.

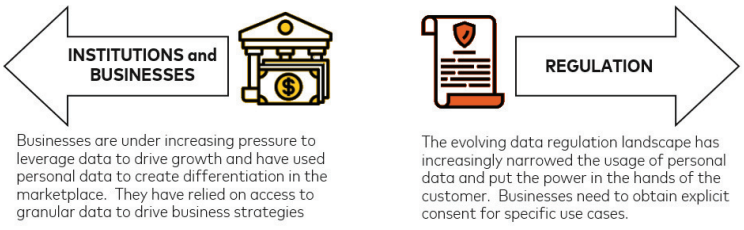

Figure 3.1: Opposing forces in the market place

While Mastercard, business owner of UC2, has a business need to operate on combined data provided by multiple parties to analyze at both microeconomic and macroeconomic levels, all financial institutions have an increasing need to use data to leverage their business. Furthermore, with the advent of open banking, there is a core need to pool data from all players in the ecosystem to enable new and improved product offerings, analysis on customer needs, and utilization of current products.

It is important to note that it is paramount to protect the privacy of individual data contributor's information, but it is also critical that the combined data asset be protected with the same rigor as well, since the access to combined information can be used to gain undue market level advantage. There is further need to track the origins and lineage of the data in the data market life-cycle. The development of novel techniques for providing effective data protection will enable better and enriched data analytics.

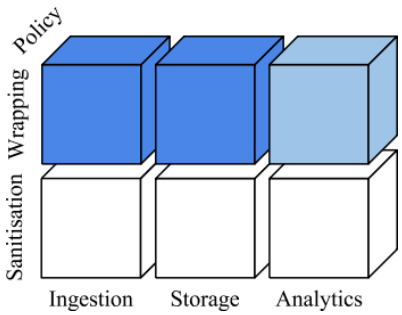

Figure 3.2: Overview of the Wrapping Dimensions

Figure 3.2 above describes the MOSAICrOWN dimensions that apply to Use Case 2. MOSAICrOWN can provide solutions for enabling the sharing and processing of microdata in respect of privacy regulations and on possible privacy/confidentiality constraints holding over the data. These scenarios will require techniques, provided by data wrapping, which are first operated by the data owner (ingestion phase), and by the data analyst and data processor at the storage and analytics phase (in case of internal processing) or at the storage phase only (in case of data extraction for external analytics).

3.2 Alignment with WPs

WPs 3 through 5 touch on the varying components of storing, securitizing and sanitizing financial data such that it is safe and ready for analysis.

3.2.1 WP3 alignment

WP3 provides multiple clear examples as they pertain to the data governance framework and a policy language for regulating access and use of data for UC2 data. Section 2.2 details the overall definition and discussion surrounding sensitive PII versus non-sensitive, transaction-level data in addition to a sample of suggested vocabulary to be used for transaction-level metadata. A key component covered in WP3 is the discussion around data transformations in support of wrapping and sanitizing sensitive PII before and after it enters the data lake.

WP3 covers some more granular topics with regards to data governance and data sharing in sections 3.5 and 4.2 of D3.1 ("First version of the reference metadata model"). Both sections provide samples of vocabulary and detailed examples of potential use cases of each within UC2. As mentioned in Section 3.1, the topic matter covered in Sections 3.5 and 4.2 are critical in not only maintaining security of the financial data, but also ensuring financial institutions are able to

utilize the securitized data for analysis. Without data governance or sharing, the transaction level data would be rendered useless.

3.2.2 WP4 alignment

WP4, the main research effort of MC for MOSAICrOWN, outlines the various forms of data wrapping that are available for use. UC2 is uniquely positioned to potentially use multiple forms of data wrapping based on the use case presented. Different classifications of data could and should require different wrapping techniques. While highly sensitive PII should require multiple techniques, like Mix&Slice or OAEP, as described in D4.1, there are certain situations where applying those techniques might not be economically or logically necessary. As D4.1 discusses, in this scenario the idea of a two-phase optimization process is necessary - with a two-phased approach, it makes determining which wrapping technique is best for the situation simpler.

3.2.3 WP5 alignment

WP5 focuses mostly on data sanitization in partnership with the wrapping techniques investigated in WP4. The key component of WP5 that is relevant to UC2 is the ability to guarantee security – regardless of access level, the model described will enforce a multi-tiered access structure that best fits UC2. This means that, depending on one's role, they will be able to access varying amounts of securitized data for analysis, successfully allowing financial institutions to continue to utilize transaction level data without putting the data at risk of breach. Furthermore, the work completed in WP5 allows for this model to be utilized in parallel with the data wrapping techniques described in WP4, making wrapping and sanitization synchronous in the MOSAICrOWN workflow, an added benefit for UC2.

3.3 Assessment of Potential Tools and Technologies for UC2

For UC2, the most critical and potentially problematic component is storage – with financial data and the level of wrapping necessary to ensure security, storage space for actual data, metadata, tokens and keys is a limiting factor for many financial institutions. Given this, determining the best tools and technologies to house this data is necessary for UC2. While a cloud-based storage system would be ideal, constraints and concerns around data privacy and maintenance make consideration of non-cloud based storage systems necessary [CW16]. As [CW16] covers, it becomes clear that given the potential downsides to a non-cloud system (i.e. network connectivity issues, inability to scale, etc.), a cloud-based storage system is preferred. SAP SE's coverage of potential solutions for a cloud-based storage system would meet the requirements of UC2.

3.4 Status of UC2 Requirements

In the following paragraphs, we will list out the requirements detailed in D2.1. For each group of requirements, we will summarize overall progress towards completion and, afterward, describe plans regarding requirement fulfillment in future work.

In D2.1 we distinguished requirements (denoted with prefix REQ-UC2-) for data ingestion (DI), access control (AC), data management (DM), data wrapping (W), sanitization (S), data controller (DC) and data analytics (DA).

Data Ingestion which focuses on details regarding how data can and will be integrated into the platform.

WP3 covers data ingestion and its proposed processes at length, and EISI is underway on prototype development.

Access Control management focuses on ensuring the proper level of access is provided to users. Again, these requirements are being discussed and covered throughout WP3. Important progress has been and will continue to be made as the WP and prototyping progresses.

Data Management covers the details with regards to data combination and manipulation.

Similar to DI and AC, these requirements are being fulfilled through the work in WP3. For more information on specific of these, see D3.3, ("First version of policy specification language and model").

Data Wrapping focuses on the various techniques and procedures around necessary data wrapping.

Certain requirements, such as analytical functions with high utility on anonymized data and the pre-wrapping functionalities previously discussed, are ongoing research activities of WP4. The vast majority of these requirements are a focal point of work being completed in WP4 and WP5, and thus are well on their way to being met. Mastercard currently has processes in place for most of these, and will continually work towards successful implementation through MOSAICrOWN

Sanitization focuses on the elements necessary for UC2 with regards to sanitizing the data.

Both requirements are addressed through policies discussed in WP5, specifically D5.2 ("First report on privacy metrics and data sanitisation").

Data Controller focuses on specific requirements with the data repository.

Both requirements are met through discussions and prototyping based on D3.3 ("First version of policy specification language and model").

Data Analytics focuses on details around client specific requirements.

Both of these are requirements that should be met in later stages of the MOSAICrOWN process, namely, when piloting and testing of the product begins. Since these will be unique to each client, Mastercard will address these individually, when the time comes.

Right now, the vast majority of these requirements are partially completed, which is to be expected given MOSAICrOWN is at the halfway mark. The ideas and overall layout of the majority of these requirements have been fulfilled, with a few not yet touched upon in WPs 3 through 5. With development on the product to begin shortly, the progress made thus far is on track, if not

ahead of schedule. Overall, the progress of the research for UC2 has been very strong and will continue to proceed accordingly.

3.5 Findings

Supporting financial institutions in protecting and securitizing their data, the central tenant of UC2, is progressing as expected. The progression of UC2 is based on ongoing research efforts in all three WPs, specifically WPs 4 and 5. UC2 will continually benefit from more focus on the detailed and granular requirements for wrapping, along with more formalized instructions on the suggested implementation of data procedures from MOSAICrOWN.

UC2 is a strong example of every facet of MOSAICrOWN - without one component (governance, wrapping, sanitzation), UC2 would not be successfully accounted for. As stated previously, the vast majority of requirements surrounding UC2 have been met and work is progressing to fully complete the remainder. The next step in successful implementation will be continued creation of research prototypes to fulfill the suggested written requirements in WPs 3 through 5. While there may be some challenges with regards to accessing previously securitized data for analysis, those challenges should not pose any significant threats to meeting deadlines.

4. Use Case 3 (SAP SE)

The objective of Use Case 3 (UC3) is to enable privacy-preserving consumer analytics via a cloud-based data market. UC3 considers business-to-business (B2B) data sharing, potentially containing personal (customer) information as well as sensitive business data. There are valuable, holistic insights that can be gained from combining the distributed data hidden in data vaults of different companies. However, to ensure that the data can be shared adequate protections are required.

The technological challenges of this use case are concerned precisely with these protections. Different sanitization techniques are investigated and implemented in the course of MOSAICrOWN, as well as privacy metrics to quantify and bound the risks for such data sharing to an acceptable level while still providing a meaningful utility.

In the first phase of MOSAICrOWN we focus on how to design and build a cloud-based service that could be integrated in SAP SE's software ecosystem. This service prototype, a proof-of-concept with at first limited features, will be expanded in the next phase to fulfill all required capabilities.

4.1 Technical Overview

Figure 4.1 shows an overview of the actors in UC3, as described in D2.1, and the ingestion, storage and analytics phases indicated by numbered arrows. We consider two data owners, a

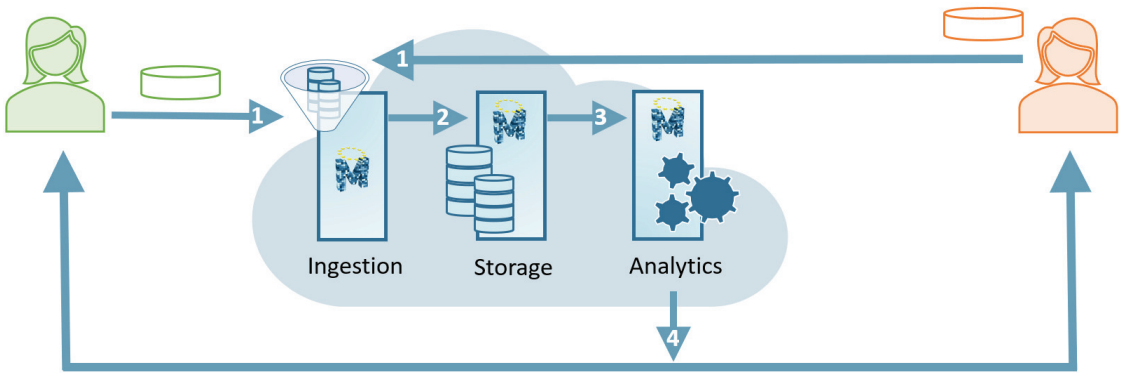

Figure 4.1: UC3 Steps

retailer and a producer of goods, and a cloud platform provider supporting privacy-preserving analytics. The retailer and producer want to compute analytics on their joint data to gain new insights, e.g., redirect marketing budget, find bottlenecks in their supply chain. In a first step, the data is ingestion into the cloud platform, i.e., arrow 1 in Figure 4.1. The data can be sanitized

during the ingestion step (before storage), i.e., in the so called *local model*, or in the analytics phase, i.e., the *central model*. Note that the *hybrid model*, based on cryptographic tools as detailed in Section 2.3 of deliverable D2.1 ("Requirements from the Use Cases"), skips the storage phase and lets parties securely compute privacy-preserving statistics. After the data has been ingested into the cloud-based data market, it can be persisted (arrow 2) or directly processed by an analytical function (arrow 3). Finally, the sanitized result is released to the data owners (arrow 4).

The required meta-data, i.e., to control the sanitization process and select the analytical function of interest, are detailed in D3.1 ("First version of the reference metadata model"). In the remainder of this section we focus on technical aspects to support privacy and security in a cloud-based environment.

4.1.1 Cloud Deployment

SAP SE supports flexible deployment models of their business services in cloud infrastructures of hyperscalers (e.g., Alibaba Cloud Computing Limited, Amazon Web Services, Google Cloud Platform,...). We aim to support the same flexibility by using broadly supported software with minimal or easily met dependencies. To ensure rapid prototyping of research results we mainly use Java and Python, which is increasingly being used for many machine learning and data analytics tasks, as programming languages. A simple web-based front-end acts as the interface between UC3 actors, i.e., producers/retailers and the cloud-based service, as is commonly found in cloud-based solutions. Further details for a first version of data sanitization tools are given in D5.1. ("First version of data sanitization tools"

4.1.2 Policy & Security Mechanisms

Meta-data, describing sanitization and protection policies, is defined and developed in WP3. D3.1 ("First version of the reference metadata model") describes the meta-data required to manage the ingestion, storage and analytics phases for UC3. Also, the data can be encrypted at rest, i.e., in the storage phase with mechanisms from WP4, see D4.1 ("First version of encryption-based protection tools") for an overview. How UC3 aligns with WP3 and WP4 is detailed in Sections 4.2.1 and 4.2.2.

Access control provides an additional security layer and prevents unauthorized parties to read the (sanitized) data in the data market. Access control is not the focus of the research activity to enable UC3 and any existing solution for key management / access control or mechanisms already provided by the market place/cloud infrastructure, i.e., as investigated in WP3, will be used within the scope of UC3.

4.1.3 Privacy Mechanisms

Data sanitization (WP5) covers non-reversible data transformation aiming to preserve the privacy of the individuals contained in the data while at the same time allowing meaningful statistics about large groups. In the following we give a brief description of differential privacy, some techniques and how to quantify the privacy guarantees.

Differential privacy (DP), a privacy definition [Dwo06], bounds privacy loss due to participation in a statistical computation, i.e., providing one's data for a data set. Differential privacy is a restriction of the output of an algorithm, often called mechanism in the DP literature, that limits by how much the output can change when the input data is altered in one record. It can be interpreted as a stability requirement, i.e., the analysis is not altered much by single elements (or outliers). Its actual goal is to protect any individual in the data set by ensuring that their input (i.e., participating in the data collection) will only lead to a small change (i.e., a bounded privacy loss). Note that one can easily prevent any privacy loss by ensuring that any input change does not lead to a change in output. However, this naive definition completely destroys any meaningful utility, as no statistical analysis will be possible. As a simple example consider a collection of producers who want to learn the most frequently bought items over all their combined sales. If each input change (i.e., a consumer buys more or less goods) cannot influence the output it is without value. However, if each input change leads to a small, formally bounded, probabilistic change in the overall output, then many participants (i.e., consumers that buy certain items more often that others), can influence the statistic of a *large* group of individuals (i.e., allowing meaningful statistics), while their individual privacy is protected.

To be more precise, a *privacy parameter* ε is used to adjust the privacy loss of differential privacy. The selection of ε can be done by the data owner (the participant in cloud-based data market, i.e., retailer, producer) or a selection from preset values for certain analytical functions (e.g., presets for low, medium, high privacy) is possible to enable non-experts to perform such privacy-preserving computations. Note that the choice and interpretation of ε depends on the concrete dataset. However, in general, small ε values tend to result in high privacy, and vice versa. Informally, with the help of the privacy parameter one adjusts the randomization required during the sanitization process, e.g., a low value (corresponding to high privacy) means one adds more noise to a result. This randomization achieved by additive noise, often called perturbation, also depends on the *sensitivity* of the statistics one aims to compute. It is the maximum impact that a single individual can have on the output of a statistic. Potential privacy losses can be quantified via *membership inference attacks*. Roughly, the goal is to compute the probability that a targeted individual was included or excluded in an (anonymized) data set.

A first version of data sanitization tools w.r.t. membership inference attacks is described in detail in D5.1 ("First version of data sanitization tools"). A detailed description of privacy metrics (regarding privacy loss) will be provided in D5.2 ("First report on privacy metrics and data sanitization").

4.2 Alignment with WPs

Regarding the identified requirements detailed in D2.1, UC3 requires alignment with WP3, WP4 and, most importantly, WP5 as detailed next.

4.2.1 WP3 alignment

The data governance framework, defined and developed in the course of WP3, aims to support sophisticated data protection with a description language usable by non-experts. Ease of use for

non-experts w.r.t. data protection guarantees and techniques is an important goal in UC3 to attract and support business customers, e.g., the retailer and producer mentioned in UC3.

For this, SAP SE actively participated in discussions and descriptions found in a first version of the reference metadata model (T3.1), and will closely follow the ongoing efforts in the creation of the policy model and language (T3.2) and policy management (T3.3). The results of WP3, namely policy specifications for data protection mechanisms, are also aligned with the mechanisms defined in WP4, which we describe next.

4.2.2 WP4 alignment

Data wrapping is a security mechanism that can be used additionally to sanitization (a non-reversible randomized data transformation aiming to preserve statistical insights), which is the main goal of UC3. Security mechanisms identified during WP4 include, e.g., hashing (non-invertible and deterministic, i.e., the same input is always mapped to the same output) and encryption (non-reversible without additional knowledge, i.e., a decryption key). Basic techniques (T4.1) and enforcement mechanisms (T4.2) for such additional data protection can be useful for UC3 actors, as it provided an additional protection layer and might enable them to ingest the data in an encrypted fashion (instead of plaintext as described for the central model) and only allow (partial) decryption during an analytical processing phase. Collaborative analytics in the hybrid model rely on cryptographic tools similar to encryption (namely, secure computation) to allow UC3 actors to jointly compute an analytical function over their data without sharing it with either the cloud service provider or other actors.

4.2.3 WP5 alignment

WP5 investigates data sanitization, the main research effort of SAP SE in the course of MOSAICrOWN. SAP SE is leading this work package and it provides the main components to realize UC3. WP5 covers privacy metrics and risks (T5.1), which enables UC3's business customers to assess and bound potential privacy risks, to make informed decisions and provide guidance for the parameterization of privacy-utility trade-offs. Another major aspect for UC3 are data sanitization techniques (T5.2) for different data types (e.g., simple numbers and unstructured text) that support strong, formal privacy guarantees and yet aim to provide high utility. Additionally, collaborative computation for sanitisation (T5.3) covers techniques to realize the hybrid model via secure computation of sanitization mechanisms. This enables even UC3 actors to participate in a data market that do not wish to persistently store their data or reveal any parts of it to the cloud service provider even in the course of an analytical function evaluation.

4.3 Assessment of Potential Tools and Technologies for UC3

To foster adoption, due to easily usable components, a web-based interface for UC3 is desirable. Commonly used web development tools, e.g., to enable user interaction, include Javascript and design frameworks, e.g., SAPUI5/OpenUI5 [SE]. To enable access control, the data market should

provide intuitive means to define and verify access permissions, e.g., in the form of password-based authentication as commonly found. This is discussed further in WP3. Such means are usually included in the database system, e.g., SAP HANA. To support the hybrid model of differential privacy additional cryptographic tools are required. Sophisticated secure computation frameworks for this goal are, e.g., ABY [DSZ15] and SCALE-MAMBA [AKR+20].

4.4 Status of UC3 Requirements

In D2.1 we distinguished requirements (denoted with prefix REQ-UC3-) for access control (AC), sanitization in the local (SL) and central model (SC), performance (P), extendability (E), and interpretability of the privacy guarantees (I). In the following, we summarize the status of UC3 requirements, afterwards we describe our plans for future work.

Access control (AC) management as investigated in WP3 (or of existing solutions, e.g., from database systems such as SAP HANA) meet our requirements.

Local sanitization (SL) covers parameter selection, storing and sharing of the anonymized result.

SAP SE developed research prototypes to provide anonymization functionalities satisfying these requirements.

Central sanitization (SC) covers parameter selection, input collection and additional cryptographic tools to support the hybrid model.

The research prototypes for SL can fulfill most of the SC requirements as well. While strong progress has been made to meet the requirements for cryptographic tools as used in the hybrid model, additional efforts are required to integrate them in the research prototypes within the context of UC3, as they rely on complex research frameworks for secure computation under active research and development (see Section 4.3).

Performance (P) with regards to utility and scalability of the analytics.

To provide not only privacy-preserving but also meaningful analytics over sensitive data our research focus is providing meaningful utility, hence, our research prototypes inherently fulfill the performance requirement for utility. Scalability is provided for anonymization functions based on carefully selected additive noise, while additional effort is required to further optimize and test additional anonymization techniques.

Extendability (EX) pattern for anonymization functions.

Our prototypes are already designed with extendability in mind, and we aim to consolidate our current interfaces to simplify the extension of existing or creation of new sanitization techniques.

Interpretability (IN) in the form of a simulated adversary.

A first version of a tool providing interpretability of the privacy guarantees (I) is detailed in D5.1 (“ First version of data sanitization tools”).

Future efforts will concentrate on improving the prototypes and integrating them in a common cloud-based analytics framework within the scope of UC3. Analytical functions with high utility on sanitized data (REQ-UC3-SC4, REQ-UC3-P2) and collaborative analytics augmented with cryptographic tools (REQ-UC3-SC3) are ongoing research activities of WP5 and we aim to expand our existing research prototypes developed during MOSAICrOWN.

4.5 Findings

UC3, privacy-preserving consumer analytics via a cloud-based data market, is based on ongoing research efforts regarding sanitization techniques and supported by research prototypes developed and expanded in the course of MOSAICrOWN (WP5). UC3's actors can benefit from additional data wrapping mechanisms (WP4) and protection and policy configuration (WP3) developed in MOSAICrOWN.

Overall, many requirements are already met and improvements in active development. The main research challenges are the design and development of privacy-preserving analytics with a strong privacy guarantee and meaningful utility (T5.2, T5.3) and how to effectively communicate these privacy guarantees with their associated risks to the customers (T5.1). Research prototypes tackling these challenges are being combined into a common cloud-based framework within the scope of UC3 to enable non-experts (UC3 actors) to easily perform collaborative privacy-preserving analytics.

5. Conclusions

This deliverable builds on the updates presented in D2.1 and provides a status update at the mid-year mark of 2020. Each chapter covered a different use case of MOSAICrOWN identified by their assigned industry partners. The use cases span from ICV data protection (UC1) to financial institutions and their transaction-level data (UC2) to a cloud-based consumer-centric data market (UC3).

Through this deliverable, we can identify progress made towards meeting the original requirements laid out for each Use Case and continue to identify any gaps in current and on-going research. Similarly to what was found in D2.1, it becomes ever clearer that, while each Use Case is meant for different industries and general end-consumers, they, more often than not, share the majority of their baseline requirements with regards to each topic covered in WPs 3 through 5. This makes for a strong, collaborative environment and will ensure progress is made in parallel across all Use Cases.

Given this state and the information detailed in previous chapters, we can conclude the following:

- *Varying use cases represent a need for varying levels of protection.* A critically important functionality of the MOSAICrOWN work is its ability to adapt for the various use cases explained above. What is necessary from a data governance, wrapping and sanitization perspective for one differs for another. WPs 3 and 4 go into further depth on how MOSAICrOWN is accounting for differences and proactively allowing for flexibility and choice for each use case.

- *Similarities amongst the use cases result in a potential for shared implementation schemas and technologies.* While there are differences between the Use Cases, it's clear that there is a strong likelihood that many of the core functionalities of MOSAICrOWN can be used across each Use Case. A shared metadata lake, similar data protection techniques, and cloud-based storage systems are all examples of core functionalities that can be repurposed without alteration for each Use Case.

- *Strong progress has been made for all use cases.* Given the great progress made, the next step should be continuing progress of research efforts and development of tools and deployment plans. Despite the current climate and uncertain nature of COVID-19, we do not expect any delays in timelines.

Bibliography

[AKR+20] Abdelrahaman Aly, Marcel Keller, Dragos Rotaru, Peter Scholl, Nigel P. Smart, and Tim Wood. Scale-mamba documentation. `https://homes.esat.kuleuven.be/~nsmart/SCALE/`, 2020.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[Cou18] Cork City Council. Electric vehicle charge points - datasets - data.gov.ie. https://data.gov.ie/dataset/ev-charge-points, Jul 2018. (Accessed on 02/06/2020).

[CW16] V. Chang and G. Wills. A model to compare cloud and non-cloud storage of big data. *Future Generation Computer Systems*, 2016.

[DSZ15] Daniel Demmler, Thomas Schneider, and Michael Zohner. Aby-a framework for efficient mixed-protocol secure two-party computation. In *Network and Distributed Systems Security Symposium*, 2015.

[Dwo06] Cynthia Dwork. Differential Privacy. In *Proc. of Colloquium on Automata, Languages and Programming (ICALP)*, 2006.

[SE] SAP SE. SAPUI5. `https://sapui5.hana.ondemand.com/`.