CYBERNETICA

ESTAT 2019.0232 Solution Analysis

Hendrik Eerikson, Armin Daniel Kisand, Baldur Kubo, Kalev Mändmaa, Jaak Randmets, Angela Sahk, Triin Siil, Ville Sokk, Riivo Talviste, Toivo Vajakas

Version 2.11

24.11.2021

54 pages

Doc. Y-1440-1

Disclaimer. This document was prepared by Cybernetica AS as part of a procured project under Service Contract No ESTAT 2019.0232 (Ref. Ares(2020)2309804 - 30/04/2020). The opinions expressed in this document are those of the authors. They do not purport to reflect the opinions, views or official positions of the European Commission or its members.

Table of Contents

1 Glossary

BPMN	Business Process Model and Notation
DPA	Data Protection Authority
Eurostat	European Commission
GDPR	General Data Protection Regulation of the European Union
IAS	Intel® Attestation Service
IMSI	international mobile subscriber identity
MNO	Mobile Network Operators
MNO-AUDIT	MNO internal audit unit
MNO data	Location data from the mobile network infrastructure
MNO-ND	MNO network department
MNO-VAD	MNO value-added services department
MPC	multi-party computation
NSI	National Statistics Institute
PETs	privacy-enhancing technologies
PLEAK	Privacy LEAKage analysis tool for privacy audit
PoC	Proof of Concept
Project	joint project ESTAT 2019.0232 code named “Sharemind-Eurostat”
SDC	Statistical Disclosure Control
SGX	Intel Software Guard Extensions technology
Sharemind HI	Sharemind Hardware Isolation development platform
Solution	An innovative privacy-preserving solution being envisaged as a result of this Project that meets the needs of both the MNO and the NSI, whilst reducing risks to privacy
Subscriber	The user of the services of the MNO; owner of the positioned device
TEE	Trusted Execution Environment
TLS	Transport Layer Security (cryptographic protocol designed to provide communications security over a computer network)

2 Introduction

In this document, the problem background, solution overview and business process are introduced. This document is intended to be read before the rest of the delivery documents.

Further details about the solution requirements, privacy enhancing technology of choice, and technical architecture are presented in the Solution Architecture document. The DPIA Evaluation Report and the DPIA Scoping Report contain the risk analysis and legal analysis.

The full list of delivery documents:

- Solution Analysis
- Solution Architecture
- DPIA Evaluation Report
- User Guide for NSI
- User Guide for MNO-VAD
- User Guide for Auditors
- User Guide for MNO-ND
- Sharemind HI Documentation
- Sharemind HI ToS
- Sharemind HI License
- Synthetic Test Data Generation

2.1 Background

European Commission ("**Eurostat**") issued a tender for development of proof-of-concept technical solution in the field of Privacy Enhancing Technologies for the processing of mobile positioning data collected from the Mobile Network Operator infrastructure ("MNO data" for short) for the production of future official statistics. The tender was awarded to Cybernetica AS, an IT company that takes fundamental technologies from early stage of research to global markets. Cybernetica's core expertise is in mission-critical systems, operational technology, information security, cryptography, and protocol analysis. Following the tender award, Cybernetica and Eurostat have started a joint project (**ESTAT 2019.0232**) that covers the development of a proof-of-concept technical solution and associated legal documentation ("**Project**").

The reference scenario was defined by Eurostat and assumes a collaboration between a single MNO and a single statistical office. This document describes the business process for the solution that is being developed by Cybernetica for adoption in the reference scenario.

Along with this document, separate companion documents are under preparation for describing the architecture of the proof-of-concept technical solution and its impact on the protection of personal data (as required under Article 35 of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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)).

2.2 Context

MNO data, produced during the use of mobile devices and recorded by a mobile network operator, is useful for different types of analysis. As a large share of European population is equipped with mobile devices, statistical offices are considering to leverage MNO data for the production of statistics about population distribution, travel patterns, etc. Historically, there have been other data sources and methods to compile such statistics (e.g., surveys), but MNO data can be considered a useful complementary source and its secondary use by NSIs is attracting increasing interest around the world[1].

MNO data represent highly sensitive personal information, especially when considered over long periods[2]. Having direct access to this data, one could potentially identify personal information, even secrets of individuals and groups (e.g. which religious establishments someone is visiting, when someone is or is not at home or work). At the same time, the NSI holds some fine-grain data that could improve the quality of the final statistics based on MNO data, e.g. in the sense of a better calibration. However, such data are confidential and cannot be passed to the MNO. In this document, we are seeking to reconcile the conflicting needs of *processing* and at the same time *protecting* the data across the two organisations.

The following elements are considered for the design of the reference scenario:

- Mobile Network Operators ("**MNO**") implement preventive measures to comply with existing regulations and mitigate risks for the privacy of their customers. In the reference scenario defined for this project, it is assumed that the MNO pseudonymises MNO data before conducting any kind of analytics. In order to reduce the risk and potential impact of privacy breaches, it is assumed that in the reference scenario the MNO is obliged to change the pseudonyms periodically, in order to prevent long-term tracking of the same (pseudonymised) device.
- National Statistics Institutes ("**NSI**") require access to MNO data to compile official statistics. The statistical methodology adopted by an NSI requires processing of individual trajectories over long periods, larger or much larger than the pseudonymisation refresh period adopted by MNOs.
- NSIs hold fine-grain auxiliary data (e.g. from census or administrative registers) that could help to better calibrate the statistics produced from MNO data. However, such auxiliary data is confidential and cannot be shared with MNOs.

[1] UN Big Data GWG Task Teams: https://unstats.un.org/bigdata/taskteams/mobilephone/

[2] Twelve Million Phones, One Dataset, Zero Privacy: https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html

2.3 Reference Scenario

For this Project, we assume the following reference scenario involving one MNO and one NSI (extensions to multi-MNO scenarios are not part of this Project and are regarded as direction for future study). The reference MNO data have the following structure:

- international mobile subscriber identity ("**IMSI**") - SIM card identifier,
- timestamp,
- position, typically at the level of individual radio cells.

NSIs want to compile statistics from MNO data.

The reference MNO has different departments with different permissions for accessing this data. More specifically, the two departments that are involved in the reference scenarios are:

- The MNO network department ("**MNO-ND**"): this is the department collecting raw data and using it for the *primary purpose*, namely analysis of service quality, network operation troubleshooting, and alike.
- The MNO value-added services department ("**MNO-VAD**"): this can be the MNO marketing department or another department in charge of producing aggregate analytics based on MNO data (*secondary use*). The MNO-VAD is not allowed to have access to the raw data. For this reason, the IMSI shared with the MNO-VAD is pseudonymised. In this Project we assume the pseudonyms are periodically changed, i.e., they are (re)computed at every pseudonymisation refresh cycle based on a new pseudonymisation key.

The NSI would like to cooperate with the MNO-VAD, but the MNO-VAD cannot conduct longitudinal analyses over longer periods, beyond a single pseudonymisation refresh cycle.

Figure 1, below illustrates the current process of sharing and analysing MNO data between MNO-ND and MNO-VAD.

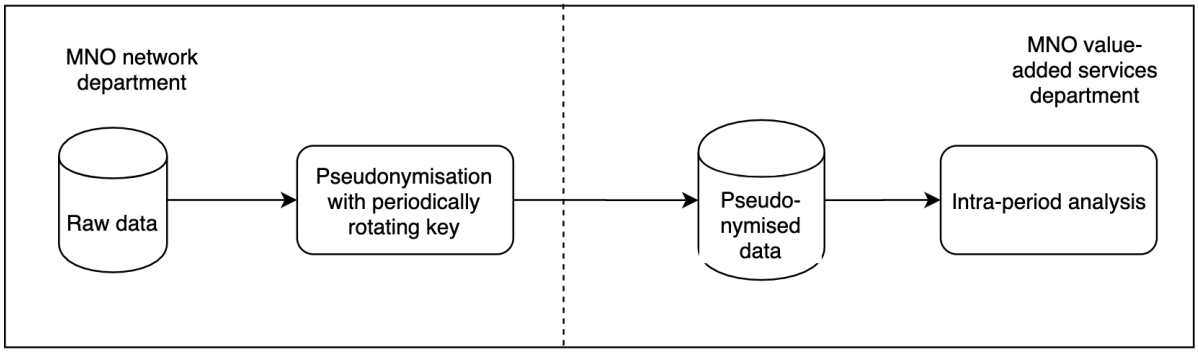

Figure 1: The current process of sharing and analysing MNO data

The goal of this Project is to develop a solution based on some privacy-enhancing technology ("**PET")** which (i) preserves the existing analysis process by the MNO-VAD, based on periodically (re)pseudonymised data, but at the same time (ii) enables the NSIs to calculate statistics based on the processing of individual subscribers' data over longer periods and (iii) allows the joint processing of MNO data and auxiliary NSI data for a better calibration of the final statistics. All that should require **no exchange of confidential data between the two organisations**. The solution would be deployed at the premises of the MNO and only final aggregate (non-personal) data would be passed to the NSI.

2.4 **Goals**

As a result of this Project, we are envisaging an innovative privacy-preserving solution that meets the needs of both MNO and NSI, whilst reducing risks to privacy ("**Solution**").

In the context of the identified needs, the Solution would allow:

- Continuing the required and already existing pseudonymisation process: the MNO-VAD will continue accessing the pseudonymised MNO data for their commercial analytics as before, i.e. IMSI pseudonyms can be linked within a short period of time by the MNO-VAD, but not across longer timespans. At the same time, the process of IMSI pseudonymisation is carried out in a way that the changing pseudonyms can be securely reversed *exclusively for the purposes of compiling statistics for the NSI* and only with a previously agreed algorithm implementation (code).
- Producing new input for statistical analysis in a privacy-preserving manner: NSIs will get access to a new source of data for compiling statistics, without inducing any reduction of the level of data protection for the MNO data. The approach honours very strictly the principles of **data minimisation**, **privacy-by-design** and **purpose specification** that lie at the foundation of the General Data Protection Regulation[3] ("**GDPR**") of the European Union.
- Joint processing (fusion) of the MNO data with the confidential auxiliary data held by the NSI for a better calibration of the final statistics, without exposing the NSI data.

The proposed solution aims at raising the level of data protection up to the highest standard allowed today by state-of-the-art technology, that is also one of the requirements mandated by the GDPR.

In addition to serving the purpose of producing official statistics, the proposed solution could in principle be extended to support the delivery of new additional analytic services by the MNO while keeping strong privacy guarantees for the mobile customers.

The solution is designed and developed with the explicit requirement of facilitating future extensions and implementation of additional processing logics. However, the actual implementation of such extensions is regarded as a point for future work and is not included in the scope of the present Project.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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). OJ L 119, 4.5.2016, p. 1–88.

3 Analysis of Technological Alternatives

3.1 Comparison of Privacy Enhancing Technologies

There are three main technology options to improve computation privacy beyond traditional privilege-based security: Secure Multiparty Computation (MPC), Fully homomorphic encryption (FHE) and Trusted Execution Environments (TEE)[4].

MPC and FHE provide somewhat different security guarantees from TEE. They both depend on algebraic schemes to ensure that data remains encrypted even during calculations. TEE is based on specific hardware that restricts access to the data being processed; in order to mitigate risks that stem from potential side-channel attacks it is desirable that the software is written according to certain rules[5].

MPC and FHE may be several orders of magnitude slower than TEE[6]. Considering the data volume envisioned by Eurostat[7], Cybernetica PET experts have decided to opt for TEE over MPC. For the sample use case of MNO data of 75 million mobile devices we estimated that the report over 6 months of raw data would take approximately 1 week of single node CPU time[8] on TEE. If we assume that MPC is 1000 times and FHE 10 000 times slower as compared to ordinary computations on plaintext data, it would take 10 cores years for MPC and tens of years for FHE to perform the same computation on encrypted data. These figures are not feasible for the project's planned use cases. Optimisation has been shown to reduce MPC computational overhead by two orders of magnitude[9]. Parallelisation has been used to reduce calendar time of the computation from a week down to two hours[10]. For the current Project Cybernetica experts considered these optimization needs unfeasible within the tender's time and budget constraints.

Another consideration was that TEE requires only one server hosted in a single organisation to run the computations, thus reducing the complexity of the overall setup also from an organisational point of view. Security of a MPC setup relies on non-collusion of participants. This means that computation nodes must be distributed over several isolated sites and servers must be all individually managed by different organizations. The servers also need to communicate with the other servers with minimal latency and maximal network throughput to follow cryptographic protocols while computing on encrypted data. This would add complexity to the general setup and technical infrastructure.

[4] UN Handbook on Privacy-Preserving Computation Techniques
http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf

[5] https://software.intel.com/content/www/us/en/develop/articles/intel-sgx-and-side-channels.html

[6] Maturity and Performance of Programmable Secure Computation. IEEE Security and Privacy. 2016.

[7] see "ESTAT 2019.0232 Solution architecture" 2.2.3.1 Assumptions for performance criteria

[8] see architecture description of "ESTAT 2019.0232 Solution architecture"

[9] https://academic.oup.com/view-large/125254164 (https://doi.org/10.1093/comjnl/bxy090)

[10] https://academic.oup.com/view-large/125254170 (https://doi.org/10.1093/comjnl/bxy090)

3.2 Comparison of Trusted Execution Environment Platforms

There are TEE solutions available for various hardware platforms – Intel, AMD, ARM. Compared to Intel, AMD provides larger directly accessible memory but somewhat lower security guarantees. Unlike Intel SGX, AMD SME and SEV do not provide memory integrity protection[11]. Several products are offered for different platforms. Most popular is Intel Software Guard Extensions (Intel SGX). Intel SGX is being thoroughly investigated for weaknesses by the academia[12]. In 2021 the market of TEE solutions is not fully developed yet and majority of the offerings are not very mature. Therefore we describe only the most relevant options.

A. One of the most complete offerings is provided by Fortanix. Fortanix offers **Fortanix Confidential Computing Enclave Manager and SDK** which provide "the flexibility to run and manage the broadest set of applications, including existing applications, new enclave-native applications, and pre-packaged applications".[13] Due to this ultimate flexibility there is larger attack surface and also less support for the specific focus of the current project -- data computations held by different parties with strong data-sharing security guarantees.

B. Cybernetica researches and develops an SGX-based solution for secure computations called **Sharemind HI[14].** This product belongs to the Sharemind family oriented towards shared computations on secret data where multiple parties provide data that remains private during the whole data usage lifecycle. Only pre-agreed analysis results are delivered to the authorized parties. For this purpose, the concept of "data flow configuration" is implemented so that data passing and handling rules can be implemented in easy-to-understand manner. It is possible to declare some calculations secret, i.e., none of the participants can see these intermediate results. Dataflow configuration makes it easier to implement the solutions for shared secret computations in modular and flexible way. This Project benefits directly from this approach.

C. Most similar to Cybernetica Sharemind HI approach is project Opaque[15], but it is not commercialised and is in status of an academic project. Furthermore, the system is not currently intended for analysis of joint data and targets computation outsourcing.

3.2.1 Note on Side-channel Attacks.

Intel is actively mitigating software-based attacks against SGX[16]. All published critical vulnerabilities have been promptly mitigated. Sharemind HI clients verify that the server is hosting enclaves on a fully mitigated platform. The Solution will be hosted by MNO-VAD on premise. The complexity of a successful side-channel attack of the proposed system is significantly higher than hacking an existing component of other systems (e. g. the raw data storage at MNO-ND). As an example the recent hardware-based fault-injection attack

[11] A Comparison Study of Intel SGX and AMD Memory Encryption Technology

[12] https://github.com/vschiavoni/sgx-papers

[13] https://resources.fortanix.com/confidential-computing-enclave-manager-data-sheet

[14] see "ESTAT 2019.0232 Solution architecture" appendix Sharemind HI overview document

[15] Opaque: An Oblivious and Encrypted Distributed Analytics Platform

[16] https://github.com/intel/Intel-Linux-Processor-Microcode-Data-Files/blob/main/releasenote.md

VoltPillager[17] requires access to the physical server, needs hardware and environment modifications, and must observe extensive numbers of enclave invocations (between 1000 and 100000) to be successful. The source systems in MNO-ND with traditional hardware technology would be a more appealing target for an attacker. In other words, the deployment of the proposed solution will not increase the attack exposure with respect to the legacy situation before its deployment.

3.3 Sharemind Hardware Isolation

Sharemind Hardware Isolation (HI) is a development platform for confidential analysis of data from multiple parties on a centralized server with full control over exposure of data and results to the involved parties. This enables building privacy-preserving applications.

Sharemind HI relies on a Trusted Execution Environment (TEE) technology to provide security guarantees for confidentiality and integrity. A TEE isolates security sensitive parts of an application from the rest of the system with the help of trusted hardware. The TEE technology used in Sharemind HI to implement the privacy-preserving data processing is Intel® Software Guard Extensions[18] (SGX) which is available in modern Intel® processors.

3.3.1 Intel® Software Guard Extensions

Applications often have some private information like passwords, cryptographic keys or secret data that only specific recipients should be able to see. The operating system hosting the application provides some level of protection for the application. However, these protections are not sufficient when the application host is itself malicious or compromised. Intel SGX is a technology to provide an extra layer of protection for the private information. The three key concepts that SGX provides to protect data are enclaves, attestation and data sealing:

- **Enclaves** - SGX is a set of CPU instructions for creating and operating with enclaves. An enclave is a program and its encrypted memory running on trusted hardware. When an application creates an enclave, it provides a protected memory area with confidentiality and integrity guarantees. These guarantees hold even if privileged malware is present in the system, meaning that the enclave is protected even from the operating system that is running the enclave. Using enclaves, it is possible to significantly reduce the attack surface of an application.
- **Attestation** - Attestation is a mechanism for cryptographically proving that an enclave with a specific fingerprint and attributes was created and is running on a trusted platform. Essentially, attestation makes sure that only the expected code runs in the enclave. SGX offers two variants of attestation: local attestation and remote attestation.

  Local attestation is performed between two enclaves on the same machine. Each enclave verifies that the other is the expected enclave. Local attestation is essential for building applications consisting of multiple enclaves where communication between the enclaves is required.

  Remote attestation is used to prove to an external party that the expected enclave was created on a remote machine. During remote attestation, the enclave generates a

[17] https://www.usenix.org/system/files/sec21summer_chen-zitai.pdf

[18] Intel® Software Guard Extensions: https://software.intel.com/content/www/us/en/develop/topics/software-guard-extensions.html

report that can be remotely verified with support by the Intel® Attestation Service (IAS). Using remote attestation, an application can verify that a server is running trusted software before private information is uploaded.

- **Data Sealing** - allows enclaves to store data outside the enclave without compromising confidentiality and integrity of the data. The sealing is achieved by encrypting the data before it exits the enclave. The encryption key is derived in a way that only the specific enclave on that platform can later decrypt it.

3.3.2 Sharemind HI Concept

Sharemind HI is built as a client-server service. The Solution is based on tasks that run inside SGX. Each task resides in a separate SGX enclave. The client is an application that calls operations on the server, encrypts data and performs remote attestation on the server.

The Sharemind HI server does the bulk of the work and is responsible for the following:

- Checking if a user has the right to access the system (authentication).
- Checking if a user has proper roles and data access permissions to perform an operation (authorization)
- Managing the encryption keys of the data.
- Managing the secure data transport in the solution between tasks and external stakeholders including data upload and download
- Storing a log of the operations performed in the server.
- Scheduling the solution tasks to run.

3.3.3 Dataflow configuration and roles

Task inputs are either stakeholder uploads or outputs from another task. Task outputs might be accessible to certain stakeholders, and might also be available to another task as input. In the current version of Sharemind HI the whole dataflow must reside in one processor.

The solution specific access rules are described in a *dataflow configuration* file. The central concept in this configuration is the dataflow graph, where stakeholders and computing tasks are nodes and the edges are topics of the messages that are passed between them. The dataflow configuration contains the following:

- A collection of tasks which as a whole form the solution.
- The fingerprints of the tasks that perform the analysis.
- Approvals given by the enforcers.
- User identities (certificates) of the related stakeholders as well as their capabilities.
- A dataflow graph serving as an access control list for data provided by clients and data created by tasks.
- Task running permissions

This way the dataflow configuration defines also three roles

- **Input Provider** – can upload inputs for at least one topic.

- **Output Consumer** – can download outputs from at least one topic.
- **Runner** – can start execution of at least one task.

A typical dataflow for deployment of this project is depicted on Figure 4.

In addition to the previously mentioned roles, that are directly related to interacting with the tasks and data (runner, producer, consumer), a stakeholder in the Sharemind HI based system can have an additional set of responsibilities and capabilities through the following roles:

- **Coordinator** - The Coordinator is responsible for coordinating any setup / deployment related activities for stakeholders involved in the solution.
- **Auditor** - The Auditor is involved in Solution development and setup to verify ex-ante and ex-post the correctness of the application. The Auditor has access to the application source code and verifies ex-ante the fulfilment of privacy requirements (including e.g. the non-personal nature of the final output). The Auditor has also access to the system audit logs to verify ex-post that the data processing with the Solution was legitimate (in accordance to the law), in accordance with the agreement of the Parties and that the solution had not been tampered with (identification of potential attacks against the solution). There can be multiple Auditors, if needed.
- **Enforcer** - An agreement ("approval") from all the Enforcers on the content of the dataflow configuration is required ex-ante, before the data collection or the analysis itself can be run. This role has the responsibility to check ex-ante that the specified dataflow configuration holds the security objectives that the parties want to achieve.

The main difference between the **Auditor** and **Enforcer** is that the Auditor has access to the application source code ex-ante and application logs ex-post, while the Enforcer(s) rely on the work of the Auditor for code checking and focus on the dataflow configuration ex-ante.

The assignment of roles in this project is discussed in chapter 4.3 Roles in the Solution.

3.3.4 Security Model

The security model of Sharemind HI relies on the security guarantees provided by SGX.

The data encryption model of Sharemind HI is illustrated below. The input data, shown in red, is encrypted at the client side and sent to the server. The input data encryption keys are securely transferred to the SGX protected enclaves. Likewise, the output data, shown in green, is encrypted inside of the enclave and stored on the server. When requested, the enclave securely transfers the output data encryption keys to the authorized clients.

It is the obligation of the Enforcers to verify that a task is configured as expected. Input Providers and Output Consumers specify which Enforcers they trust with this task. Sharemind HI ensures that they can only upload data to and download data from tasks which have been approved by their trusted Enforcers. This link of trust prevents clients from sending data to or receiving data from a wrong task enclave.

At any point during the deployment, a client can request a cryptographic proof of what analysis code is running in the server, shown in blue on the figure. This proof can be compared against a previously generated proof by an Auditor who has validated the code to be secure and privacy-preserving as a part of the Solution setup.

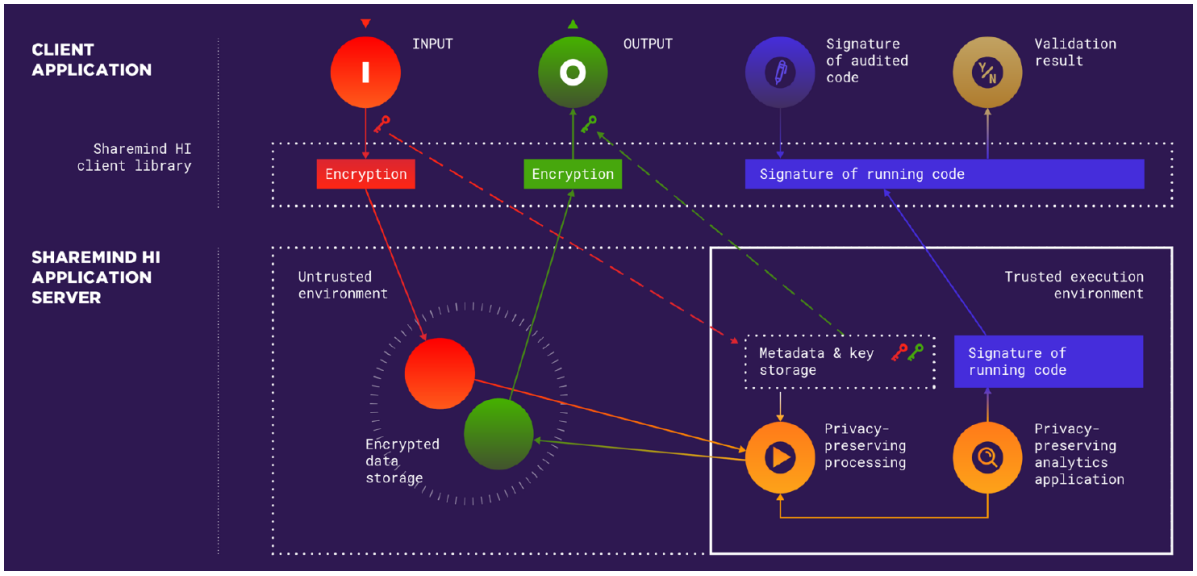

Figure 2 Sharemind HI security model

For each Sharemind HI deployment, a deployment Coordinator has to generate a deployment specific private key and public key certificate signed by the Cybernetica Deployment Root CA for Sharemind HI. This private key is used to sign all the client keys that want to communicate with the Sharemind HI server. The signed deployment certificate is loaded into the server at start-up and is used to authenticate clients in remote attestation. The Cybernetica Deployment Root CA certificate is embedded into the server and verifies the validity of the deployment certificate.

4 High-level Solution Overview

4.1 Proposed Solution

In this Project, the Solution is implemented using Sharemind HI briefly described above[19]. Figure 3 below illustrates the architecture of the proposed Solution.

Components and interactions highlighted in green are trusted by cryptographic verification. We consider two departments in the MNO: MNO-ND with access to raw MNO data and MNO-VAD with access to (periodically refreshed) pseudonymised MNO data.

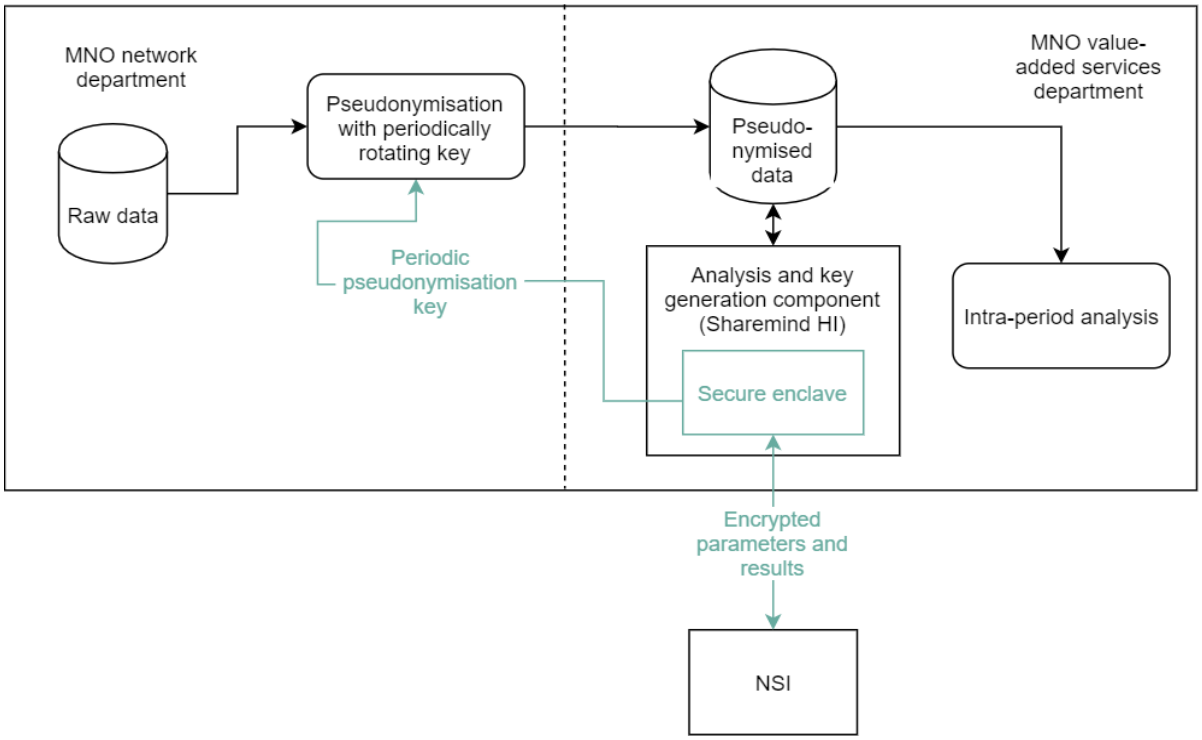

Figure 3: Proposed secure architecture of sharing and analysing MNO data

Please note that it is not in the scope of this Project to analyse how MNOVAD uses and processes the MNO data. We designed the add-on process so that deploying the Solution will leave the legacy process of pseudonymisation at the MNO substantially unaltered (the only modification concerning the periodic key generation process).

4.2 Stakeholders

In general terms, stakeholders are defined as organisations or individuals with a relationship to the change, the need, or the solution. In this section we cover the list of stakeholders that are relevant for the case at hand. Each stakeholder may be tied to one or more roles of the Solution,

[19] For a more detailed description refer to appendix “Sharemind HI overview” in the "ESTAT 2019.0232 Solution architecture" document.

List of stakeholders involved in the Solution proposed in this document:

- **National Statistics Institute (“NSI”)**

  The public institution in charge of producing and publishing official statistics. NSI might not have a direct access to the Solution, input data or pseudonymisation keys. The limits of NSI access are subject to agreement between the NSI and the MNO, considering legal regulations, and technically implemented by the Solution.

- **Mobile network operator network department (“MNO-ND”)**

  The data being analysed in the Solution are produced by the mobile network infrastructure operated by the MNO network department.

- **Mobile network operator value-added department (“MNO-VAD”)**

  Any MNO value-added department (for instance, the marketing department or a business intelligence branch) could also order statistical analysis based on MNO data for (private) business interests. MNO-VAD might not have direct access to personal data collected by MNO-ND.

- **Cybernetica AS**

  In the preliminary prototype solution (and possibly in the field-testing phase if needed) Cybernetica AS will be mediating the Intel Attestation Service provided by Intel. Specifically, Cybernetica AS will take the role of an attestation service proxy in order to avoid the need to register the new solution with Intel. The Cybernetica AS registration with Intel will be used for the Intel Attestation Service. This will simplify the setup during prototyping and initial field testing of the solution.

  For any future live solutions (production ready), Cybernetica AS would be replaced by Intel and **registration to Intel Attestation Service would be required**.

- **Optional external auditor.**

  In addition to MNO-VAD and NSI who would fill the auditor roles, external auditor(s) may be involved to ensure trust of the Solution deployment. The external auditor reviews the Solution including application source code ex-ante and logs ex-post. The auditor acts as one of the Enforcers who will sign-off the Solution. In practice, the auditing task may be carried out by an MNO internal audit department, by the national Data Protection Agency, or by an accredited auditing company.

4.3 Roles in the Solution

In this Project, the Solution is implemented using Sharemind HI. We provide an estimation of which stakeholder might be best suited to perform each role, with the understanding that different configurations may be flexibly implemented if desired. A stakeholder is a human, device, or system that performs a role in interacting with the Solution. Each stakeholder in the Sharemind HI system has one or more roles.

4.3.1 Sharemind HI Server Host

The Host is someone who hosts the Sharemind HI deployment (solution), which receives data from the MNO-ND, processes data using the Solution and issues results to the Output Consumer for further processing. The largest input data volume for Sharemind HI server is

MNO data exported from an existing Data Warehouse infrastructure in the MNO-VAD. This large dataflow is encrypted for the enclave of the Solution by MNO-VAD.

Privacy and confidentiality of data is protected during the entire analysis process by the Solution. Conducted actions are auditable by a designated Auditor, and therefore in principle the Sharemind HI server Host ("**Host**") could be any institution who owns or operates the necessary infrastructure (a computer using Intel SGX technology or cloud computing service). However, in order to reduce the data transmission load across different premises, in practice it is preferable to assign the Host role to MNO-VAD so as to allow co-location of the Solution with the source Data Warehouse.

Potential stakeholders:

In the proposed Solution, the Host role is carried out by MNO-VAD. Any other MNO department could also act as the Host, depending on the specific MNO organisational structure and IT infrastructure. For example, there could be a dedicated cloud services provision unit or department in the MNO, owning and managing their own cloud infrastructure.

4.3.2 Coordinator

The Coordinator is responsible for coordinating any setup / deployment related activities for stakeholders involved in the solution. They are responsible for coordinating the agreement on the analysis that will be run. The Coordinator ensures that the required processes are followed in order to guarantee the security of the solution.

Potential stakeholders:

In the proposed Solution, we assume the Coordinator role is carried out by the NSI. This because the NSI has the interest of adding new analyses to the Solution and the knowledge of the required analytics to be added. Note however that the Coordinator role could also be carried out by MNO, if that configuration is preferred.

4.3.3 Developer

In order to run a new or updated analytic application (e.g. a new statistics) a new Sharemind HI solution version will need to be developed and configured. The development produces a deployment package for the target operation system. The Developer is responsible for developing or ordering[20] the development of the new/updated (statistical) analysis application.

Potential stakeholders:

In the proposed Solution, the stakeholder responsible for the development (in-house or through external procurement) of new analytics is the NSI. At the moment we do not foresee the MNO needing to add new analysis applications to the solution.

4.3.4 Auditor

The Auditor validates critical code components ex-ante, before deployment, including that the implemented algorithm complies with privacy requirements. Upon successful code review the Auditor issues the application fingerprint.

[20] The party carrying out development of the software should have sufficient knowledge and expertise in cryptography, Intel SGX, Sharemind HI and secure/side-channel safe development techniques in order to develop secure analytics and adhere to the solution security requirements. Cybernetica is capable of providing training and consulting for privacy-preserving design and development of Sharemind HI solutions for the preferred developer or providing development services as needed.

Later this fingerprint will be used to verify the correctness of the deployed application in the Host's system. The Enforcers will rely on this fingerprint to approve the deployment. During setup and maintenance of the solution the Auditor will check ex-ante, if the Solution uses the approved code.

The Auditor verifies that the data processing within the Solution

- is legitimate - done in accordance to the law,
- is in accordance with the agreement of the Parties and
- has not been tampered with (identification of attacks against the solution).

The Auditor role has also access to the system audit logs for ex-post analysis.

Potential stakeholders:

In addition to NSI and MNO-VAD the Auditor of the proposed Solution should be an External Auditor (external to the entities acting as Input and Output parties). It can be for example the audit unit internal to the MNO ("**MNO-AUDIT**"), or an entity appointed by the MNO or Data Protection Authority ("**DPA**"). There may be multiple entities taking the Auditor role in parallel.

4.3.5 Enforcer

The Enforcers are required to provide approval on the contents of the analysis (dataflow configuration) before the data collection or the analysis can take place. Stakeholders in this role have the responsibility to check that the specified dataflow configuration, consisting of the analysis code's fingerprint and the assignment of roles to the parties, complies with the confidentiality requirements.

In general, there might be multiple Enforcers, and any party that is interested in some aspect of the analysis may take the role of Enforcer together with others. For example, the Input Providers are interested in protecting their input data throughout the lifetime of the Solution, while the Output Consumers are interested in correct analysis results. Enforcement allows parties to refuse new analytics to be executed if any such party has doubt about the kind of analysis, roles setup, deployment that do not / no longer meet the security or privacy requirements.

The main difference between the Auditor and Enforcer is that the Auditor has access to the application source code ex-ante and application logs ex-post, while the Enforcer(s) rely on the work of the Auditor for code checking and focus instead on the dataflow configuration ex-ante. In other words, the Auditor and the Enforcer perform different layers of ex-ante verification.

Potential stakeholders:

In the proposed Solution, the Enforcer role should be carried out by all Input Providers and Output Consumers, plus any additional stakeholder interested in making sure that the analysis does what it is supposed to. This includes NSI, MNO department responsible for MNO data (MNO-ND) and possibly, MNO-VAD. Optionally, an External Auditor could also act as an Enforcer.

4.3.6 Attestation Service Provider

Remote attestation (chapter Intel® Software Guard Extensions) is used to prove that the expected solution/deployment (enclave) was created on a remote machine (Host) using Intel SGX technology with latest security patches. Before secret data is uploaded, by using remote

attestation, an application can verify that a server is running trusted software in the trusted hardware.

Potential stakeholders:

In the proof-of-concept, Attestation Service is provided by Cybernetica AS to simplify deployment of the PoC solution for MNO.

4.4 Dataflow Related Permissions and Roles

The Solution is based on tasks that run inside TEE. The task inputs are either stakeholder uploads or outputs from another task. The task outputs can be configured to be accessible to certain stakeholders, and be available to another task as input.

Multiple usage scenarios can be implemented using the approach. A stakeholder can be configured to receive output and be able to decrypt it to insert into ordinary analytics, reporting or publishing process. On the other hand, the stakeholder can provide the public key of a different enclave, implementing another Dataflow, and securely consume the secret output without ever seeing the contents.

The solution specific access rules are described in the *dataflow configuration* file. This configuration defines the dataflow graph, where stakeholders and computing tasks (enclaves) are nodes and the edges are topics (i.e. pseudonymisation key generation or analysis) of the messages that are passed between them, see Figure 4. Task running permissions are also defined in this file.

Running parameters do not contain secrets and are not considered part of the dataflow.

These roles are used on Figure 4 to describe the process.

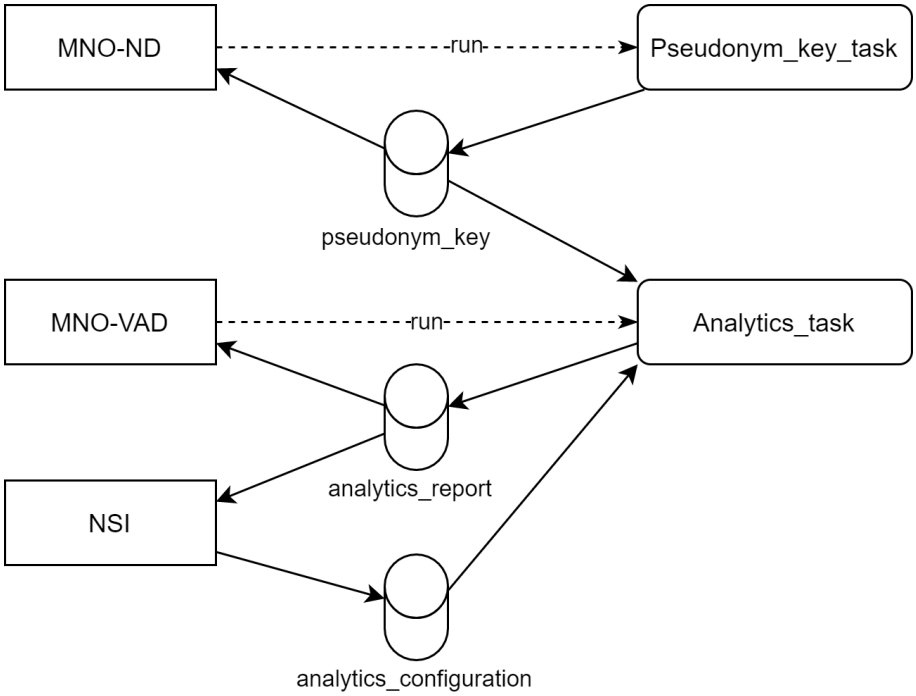

Figure 4 Depiction of the planned dataflow

In the proposed Solution, the pseudonymisation key generation task is run by MNO-ND (department responsible for MNO data storage). Analysis task can be run by MNO-VAD who hosts the servers. Being a Runner does not imply access to inputs or outputs.

NSI can upload the data needed for reports. This data is not visible to other stakeholders.

The graph defines explicitly which parties can download which data (topics). In the proposed Solution, the periodic pseudonymisation keys are accessed by MNO-ND (used to perform one-way pseudonymisation). The produced statistical analysis results can be downloaded by NSI and MNO-VAD.

It is easy to configure different dataflow endpoints, e.g. MNO-VAD as runner instead of MNO-ND for Pseudonym_key_task, and NSI instead of MNO-VAD for Analytics_task), if such alternative configuration is preferred.

4.5 Certificate management

To access the Solution, each stakeholder in the client role must have a public key cryptography key pair that allows them to create an encrypted and mutually authenticated communication channel with the Sharemind HI server. The key pair is generated by each such stakeholder itself and their public key (certificate) is then signed by the Solution Coordinator. Each stakeholder needs to select a secure channel for certificate transfers. Coordinator's signing key itself is signed by the Solution developer (in the PoC solution, Cybernetica AS) and embedded in the Sharemind HI server. This makes it possible to root the Solution trust in a specific hardware deployment.

4.6 Stakeholders and Roles Matrix

The stakeholders and roles matrix proposes a distribution of Project stakeholders and their roles in the Sharemind HI Solution. Assigning of roles depends on availability of resources and competences of the stakeholders and may be renegotiated upon need.

See Table 1 Roles of Stakeholders in the Proposed Solution for the summary and

		Stakeholders				
		MNO-ND	**NSI**	**MNO-VAD**	**External Auditor**	**Intel via Cybernetica proxy**
Roles	**Sharemind HI server Host**			+		
	Coordinator		+			
	Enforcer	+	+	+	+	
	Input Provider	PT[21]	AT	AT[22]		
	Output Consumer	PT	AT	PT[21] AT[23]		
	Runner	PT		AT		
	Developer		+			
	Auditor		+	+	+	
	Attestation Service Provider					+

Table 1 Roles of Stakeholders in the Proposed Solution

note: some roles are task-related. The involved tasks are marked: PT -- pseudonym key process, AT -- analysis process. Other roles are not task-specific and are marked with “+”.

[21] MNO-ND provides MNO data to be received by MNO-VAD, but this data transfer is done entirely outside Trusted Execution Environment and therefore not configured via Dataflow Configuration.

[22] For optimization reasons, the access to daily footprints data provided by MNO-VAD are accessed directly by Sharemind HI task bypassing the default Dataflow Configuration mechanism. Therefore that role of MNO-VAD is not described in Dataflow Configuration.

[23] For the PoC we recommend the MNO-VAD to also see the outputs of the NSI analytics in order to improve confidence in the process.

5 Description of Work Process

The following process model (using PE-BPMN notation and the PLEAK process design tool[24]) depicts stakeholders using the Solution, their actions and data elements that are being used or created. The detailed descriptions are in the following separate subparagraphs. The process models include security methods used to protect secret data. This allows assessing the visibility of data elements for each stakeholder involved in the process.

Process diagrams depict Solution stakeholders (not roles) on separate swim lanes for better readability and understandability. Processes are detailed below in tables covering tasks and data elements. In order to provide the link between stakeholders and their roles, tasks have been connected to their associated roles in the tables below.

5.1 P0 General Process Overview

In this subparagraph the general process of the Solution is being described. Each subprocess in this general work process is being detailed in the following subparagraphs.

The processes are combined in a process map depicted in Figure 5Figure 5: illustrating the value chain of the privacy-preserving solution. For each activity on the map, a detailed business process model is provided with the involved stakeholders and their activities. Data items used or created during the process are defined and specified in the detailed business processes.

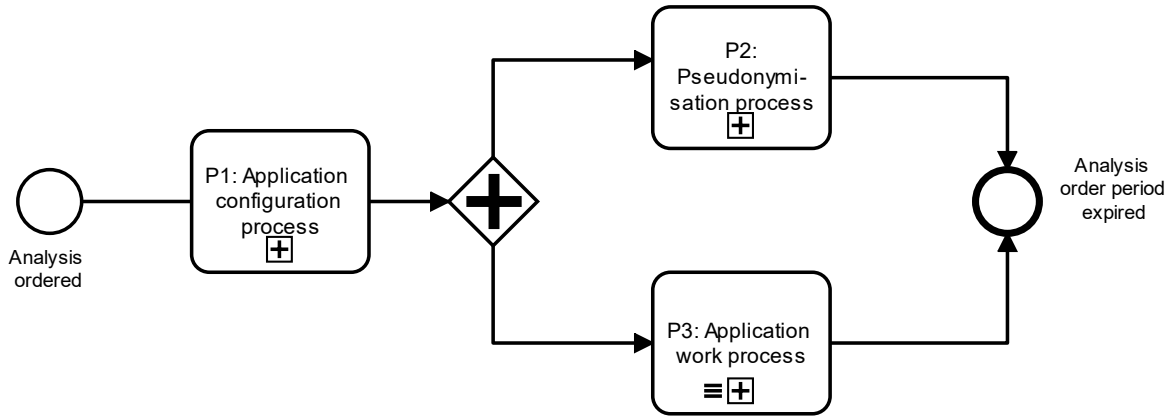

Figure 5: Solution general process

The general process starts with the configuration of the Sharemind HI solution P1.

Process P2 runs continuously and prepares the pseudonymised data that P3 uses as input. Process P3 runs concurrently with P2, the Solution accepts multiple sequential runs of P3 process.

[24] PLEAK (Privacy LEAKage). General overview of PLEAK tool is presented in Appendix 3 – PLEAK Open-source Process Analysis Software. For more information: https://pleak.io/wiki/pleak

5.2 **P1 Application Configuration Process**

In this subparagraph the initial configuration of the analysis application is being described. The ultimate goal of this step is to ensure that the analysis will not leak secret data.

For readability, this model only shows MNO-ND in the Enforcer role. However, the same role is also taken by MNO-VAD, NSI and External Auditor.

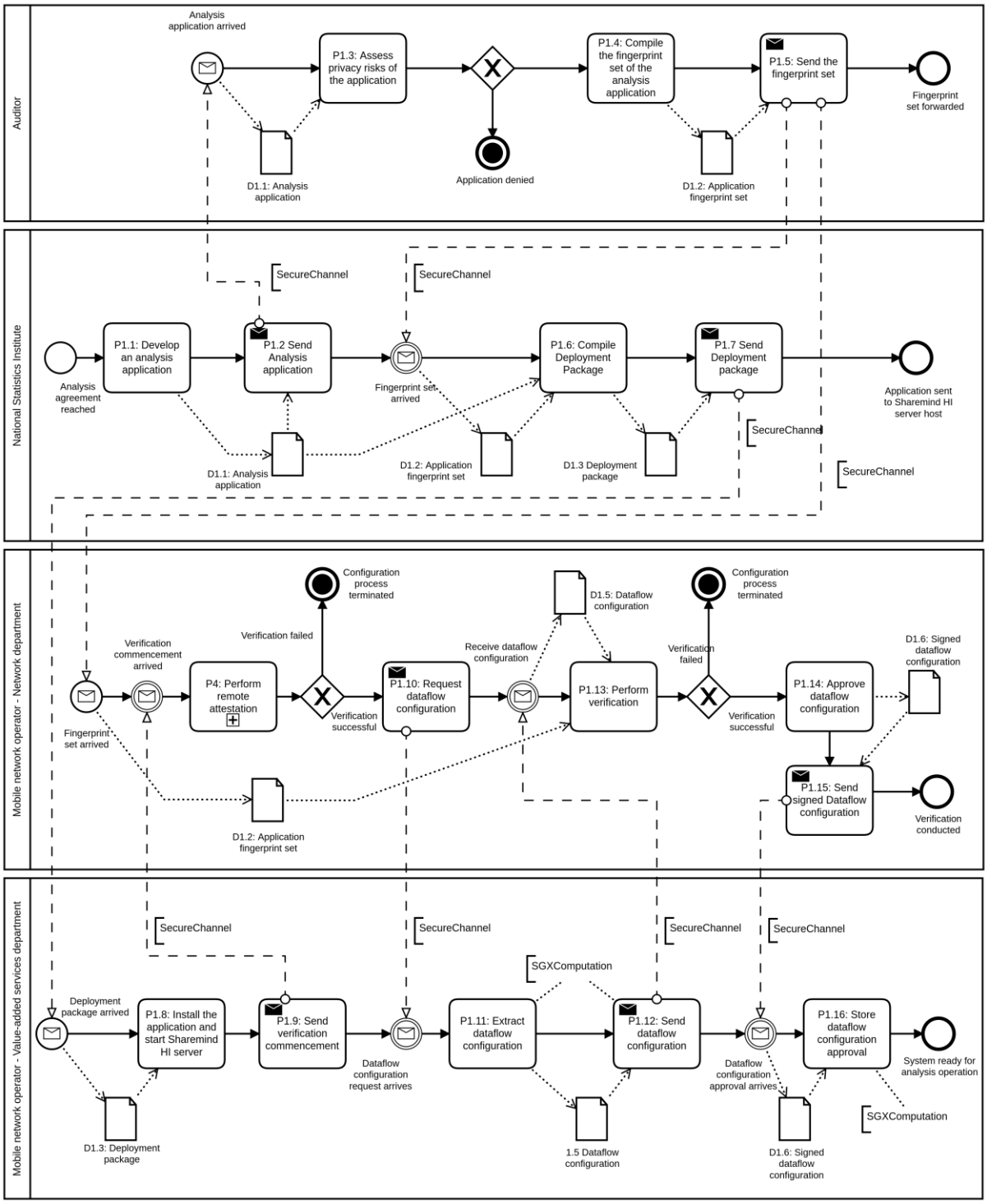

Figure 6: Application configuration process

5.2.1 Actions

In this subparagraph the actions, their triggers, goals and stakeholders while configuring the Solution are described in detail. Each activity has the following information: a unique identification number referenced from the process diagrams, the name of the process, the name of the role who performs the activity (with the reference to proposed stakeholder in PoC), a brief description of the activity, the trigger and the goal of the activity.

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P1.1	Develop an analysis application	Developer (NSI)	The Developer develops an application for statistical analysis.	The statistical analysis of MNO data has been ordered.	The new statistical analysis application will be deployed and run on the Solution.
P1.2	Send the analysis application	Developer (NSI)	The Developer sends the analysis application to Auditor	The application development has completed.	Analysis application reaches the auditor
P1.3	Assess privacy risks of the application	Auditor (MNO-VAD, NSI and External Auditor)	Assess privacy risks of the application and estimate the possibility for a private data leak.	Analysis application arrived.	Auditor has validated whether processing is suitable.
P1.4	Compile the fingerprint set of the analysis application	Auditor (MNO-VAD, NSI and External Auditor)	Compile the fingerprint set of the analysis application which is being verified on the launch of the application.	Privacy risks of the application have been assessed and the application has been approved.	Compile a fingerprint set of the audited solution for later validations.
P1.5	Send the fingerprint set	Auditor (MNO-VAD, NSI and External Auditor)	Send the fingerprint set of the analysis application to MNO-ND and NSI.	Analysis of the application has been compiled.	Fingerprint set of the analysis application has been sent to the MNO-ND and the NSI.
P1.6	Compile deployment package	Coordinator (NSI)	Compile a package including the application and its fingerprint set.	Application has been developed and its fingerprint set has arrived.	Deployment package ready to be sent.

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P1.7	Send deployment package to MNO-VAD	Coordinator (NSI)	Send the package to the Sharemind HI server Host to be installed.	Compilation of the package is completed.	Application setup processes initiated.
P1.8	Install the application and start Sharemind HI server.	Host (MNO-VAD)	Install the application developed and sent by Developer (NSI) on the Sharemind HI server	Deployment package has arrived.	Application has been installed.
P1.9	Send verification commence ment to enforcers	Host (MNO-VAD)	Send message of installation completion to Enforcers	Application is installed.	Message has been sent to Enforcers.
P4	Perform remote attestation.	Coordinator (NSI)	Perform the remote attestation process to prove to an external party that the expected enclave was created on a remote machine. Remote attestation is described as separate subprocess P4 (see subparagraph 5.5).	Installation of the application has been launched and an enclave attestation is needed.	Attestation performed successfully (attestation report received). Trust of the application deployment verified.
P1.10	Request Dataflow Configu-ration.	Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	Enforcer asks for the Dataflow Configuration from the Host in order to verify that tasks and data flows are configured as agreed upon.	Attestation has been performed successfully.	Dataflow Configuration request has been sent.
P1.11	Extract Dataflow Configuratio n	Host (MNO-VAD)	Host exports actual dataflow configuration from server.	Dataflow configuration request has arrived.	Actual dataflow configuration is exported.

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P1.12	Send Dataflow Configu-ration.	Host (MNO-VAD)	Host sends the Dataflow Configuration to the Enforcer	Dataflow configuration is extracted.	Dataflow configuration has been sent to the Enforcer.
P1.13	Perform verification.	Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	The enforcer verifies the data-flow configuration which contains the stakeholder cer-tificates, the roles of each stake-holder, task fingerprints and the dataflow. The correct certify-cates, roles, fingerprints, dataflow is provided by the Coordinator and/or the Auditor. In case the verification fails the process is terminated.	Dataflow configuration has arrived.	Verification is successful.
P1.14	Approve Dataflow Configu-ration.	Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	Approve Dataflow Configuration, sign its approval.	Verification has been successful.	Dataflow Configuration is approved.
P1.15	Send signed Dataflow configuratio n	Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	Send approved Dataflow Configuration to Host (MNO-VAD).	Approval was successful	Approval forwarded to Host.
P1.16	Store Dataflow Configu-ration approval.	Host (MNO-VAD)	Store Dataflow Configuration approval in Sharemind HI server SGX enclave.	Dataflow Configuration approval has arrived.	Dataflow Configuration approval has been stored. System is ready for analysis operation

Table 2 Process 1 Actions

5.2.2 Data Elements

This section describes the data elements related to the process covered by the previous section. Each data element has the following information: a unique identification number referenced from the process diagrams, name of the data element (a BPMN term), name of the holder of the element and a brief description of it.

ID	Name	Holder	Description
D1.1	Analysis application	Developer (NSI) Coordinator (NSI) Auditor (MNO-VAD, NSI and External Auditor)	Analysis application is a software program that is composed of algorithms processing statistical data out of raw MNO data.
D1.2	Application fingerprint set	Auditor (MNO-VAD, NSI and External Auditor) Coordinator (NSI) Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	A short unique combination of symbols (numbers and letters) that provides the possibility of verifying whether the application in use has the same functionality than the one that was audited.
D1.3	Deployment Package	Coordinator (NSI) Host (MNO-VAD)	Software package that contains dataflow configuration with the application fingerprint set, Sharemind HI server, analysis application, and all other necessary software (services, scripts, etc) and various dependencies required for the functioning of the Sharemind HI server.
D1.5	Dataflow Configuration	Coordinator (NSI) Host (MNO-VAD) Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	A document binding together the fingerprints of programs running in the enclaves, identities (public keys) of authorised users and the dataflow graph.
D1.6	Signed Dataflow Configuration approval	Enforcer (MNO-ND, MNO-VAD, NSI, External Auditor)	Digitally signed statement by the Enforcer to show that they have verified the Dataflow Configuration against an authentic copy or agreement.

Table 3 Process 1 Data Elements

5.2.3 Security Controls

In this subparagraph the security controls depicted on the configuration diagram are being described.

5.2.3.1 SecureChannel

Data transmission with the SecureChannel annotation on the PE-BPMN diagram means that a secure audited and encrypted channel is being used. Technically, the use of up-to-date standard solutions (for instance, TLS, Transport Layer Security) is recommended.

A communications channel that is set up this way won't leak data to the communication service provider or third parties.

5.2.3.2 SGXComputation

SGXComputation annotation on the PE-BPMN diagram denotes that secure enclave computation is being applied. All computations are being conducted inside of the enclave and it is impossible to read the data in open form from the computer (server).

5.3 P2 Pseudonymisation Process

In this subparagraph the pseudonymisation process of the MNO data is described: generation of pseudonymisation key, performing pseudonymisation and storing pseudonymised data.

The pseudonymisation process illustrates how the Solution is a drop-in replacement for the MNO-ND's existing process.

When the Sharemind HI solution is available, the MNO-ND retrieves a new periodic pseudonymisation key from the pseudonymisation key enclave, instead of generating the key itself. The key leaves the TEE in encrypted form and is only decrypted by the MNO-ND. As the unprotected key never leaves the MNO-ND's premises and it is only used to pseudonymise data that the MNO-ND has access to, the decryption of the key is not significant to protecting MNO data. Similarly, the MNO-ND is required to delete the key after completing pseudonymisation but this deletion is also not necessary for protecting MNO data.

The pseudonymisation process uses a symmetric encryption scheme. Alternatively, a public key encryption scheme could be used where the enclave provides a new public key to the MNO-ND for every pseudonymisation period. The symmetric scheme was chosen because of the smaller ciphertext, faster key generation and faster pseudonymisation.

Moreover, as periodic pseudonymisation keys are generated in the enclave, the Solution facilitates a possible future extension scenario where a single Key Task Enclave provides periodic pseudonymisation keys to several MNOs.

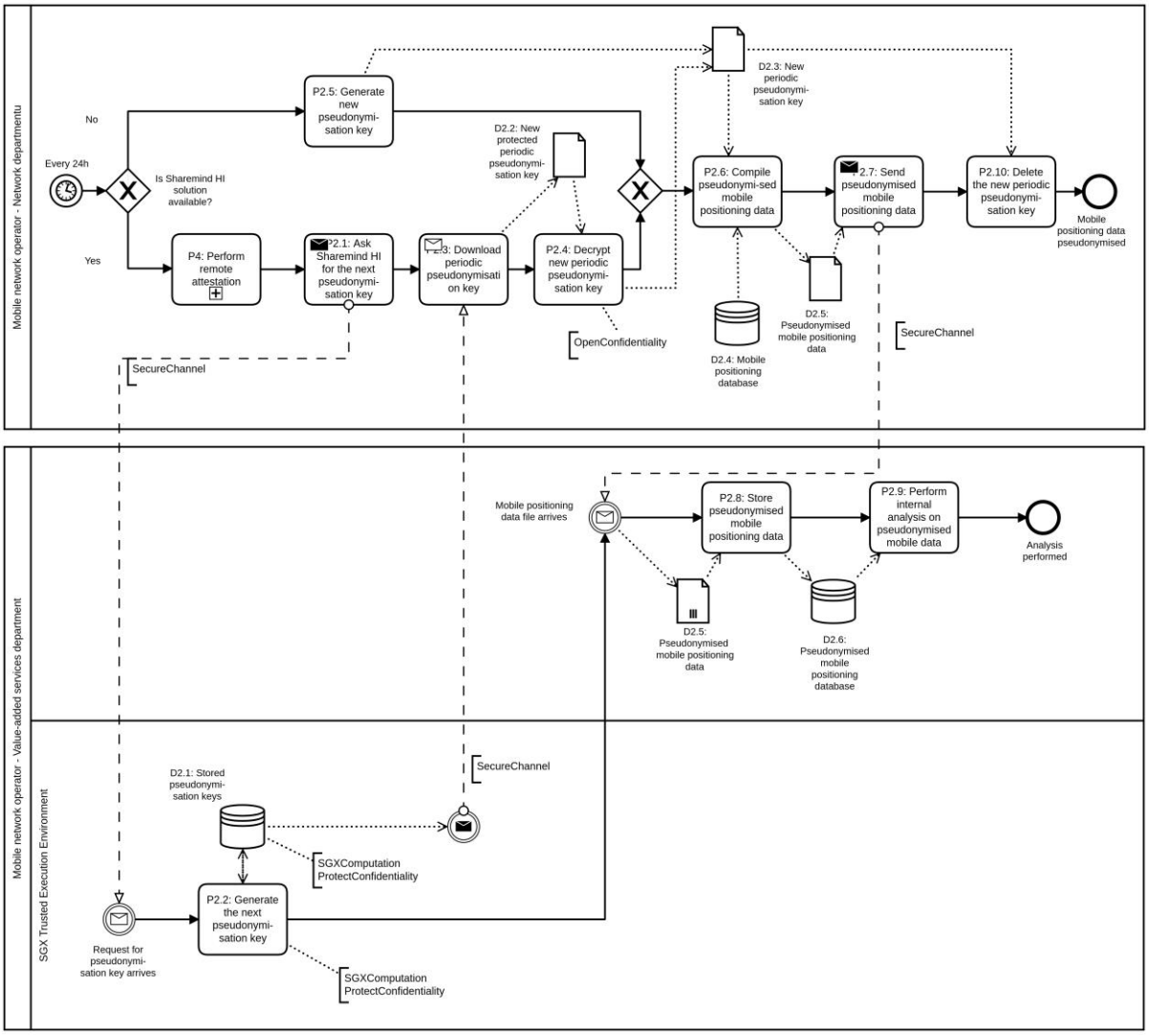

Figure 7: MNO data pseudonymisation process

5.3.1 Actions

In this subparagraph, the actions, their triggers, goals and stakeholders in the pseudonymisation process are described in detail. Each activity has the following information: process, the name of the role who performs the activity (with the reference to a possible concrete stakeholder in the PoC), a brief description of the activity, the trigger, and the goal of the activity.

ID	Name	Role (Stakeholder)	Description	Trigger	Goal
P4	Perform remote attestation.	Runner (MNO-ND)	In case Sharemind HI solution is available, remote attestation process will be conducted. Remote attestation is described in the separate subprocess P4 (see subparagraph 5.5).	Pseudonymisation process has started and Sharemind HI solution is available	Remote attestation performed successfully and trust of the application deployment verified.
P2.1	Ask Sharemind HI for the next pseudonymi-sation key.	Runner (MNO-ND)	Send request to Host (MNO-VAD) for pseudonymisation key.	Remote attestation performed successfully	Request for next pseudonymi-sation key is sent to Host.
P2.2	Generate the next pseudonymi-sation key.	Host (MNO-VAD)	A new periodic pseudonymisation key is generated. This key is protected by the Sharemind HI enclave, i.e. not accessible by the Host.	Request for the next pseudonymisation key has been received	A protected pseudonymi-sation key is generated for the next period
P2.3	Download the periodic pseudonymi-sation key.	Output Consumer (MNO-ND)	The protected pseudonymisation key of the next period is downloaded to be used in the data pseudonymisation process.	The next protected periodic pseudonymisation key has been asked from the Sharemind HI server.	The protected pseudonymi-sation key for the next period is downloaded.
P2.4	Decrypt the new periodic pseudonymi-sation key.	Runner, Output Consumer (MNO-ND)	The new periodic pseudonymisation key will be decrypted after arrival.	The protected next periodic pseudonymi-sation key has arrived.	Protected next periodic pseudonymisation key has been decrypted.

ID	Name	Role (Stakeholder)	Description	Trigger	Goal
P2.5	Generate new pseudonymi-sation key.	Runner (MNO-ND)	In case Sharemind HI solution is not available, the new periodic pseudonymisation key is randomly generated by the MNO-ND.	Pseudonymisation process has started and the Sharemind HI solution is not available	Protected next periodic pseudonymisation key is generated
P2.6	Compile pseudonymised mobile positioning data.	Output Consumer Input Provider (MNO-ND)	The MNO-ND will compile the MNO data file (acquiring data from MNO-ND database) and pseudonymises the data using the new periodic pseudonymisation key	Protected next periodic pseudonymisation key has been generated	MNO data file has been composed and pseudonymised
P2.7	Send pseudonymised MNO data.	Input Provider (MNO-ND)	Input Provider sends pseudonymised MNO data to MNO-VAD.	MNO data file has been composed and pseudonymised	Pseudonymised MNO data has been sent to the MNO-VAD.
P2.8	Store pseudonymised MNO data.	Host (MNO-VAD)	Pseudonymised MNO data is added to an already existing database of pseudonymised MNO data.	Pseudonymised MNO data file has arrived	Pseudonymised MNO data has been stored
P2.9	Perform internal analysis on pseudonymised mobile data.	Host (MNO-VAD)	After storing the Pseudonymised MNO data, it is possible to conduct stakeholder internal analysis on it (that span only one period).	Pseudonymised MNO data has been stored in the database	Analysis has been performed.
P2.10	Delete the new periodic pseudonymi-sation key.	Input Provider (MNO-ND)	The periodic pseudonymisation key will be deleted in the Input Provider's environment.	Pseudonymised MNO data has been sent to Sharemind HI Host.	The periodic pseudonymisation key has been deleted.

Table 4 Process 2 Actions

5.3.2 Data Elements

This section describes the data elements related to the process covered by the previous section. Each data element has the following information: a unique identification number referenced from the process diagrams, name of the data element (a BPMN term), name of the holder of the element and a brief description of it.

ID	Name	Holder	Description
D2.1	Stored pseudonymisation keys	Host (SGX TEE)	Pseudonymisation keys that have been generated inside the Sharemind HI enclave upon request from MNO-ND. Each key is an independent random value valid for one period only. Each key has an assigned period id. These keys are **protected** from the Sharemind HI **Host (MNO-VAD)**. The individual keys can be downloaded by the **MNO-ND** as D2.2.
D2.2	Protected new periodic pseudonymisation key	Host (SGX TEE) Input Provider (MNO-ND)	A new pseudonymisation key used to pseudonymise data belonging to one specified period. This key is **protected** from the Sharemind HI **Host (MNO-VAD)**. It is securely sent to the **MNO-ND** who **can decrypt, see, and use** the key in plaintext as D2.3.
D2.3	New periodic pseudonymisation key	Input Provider (MNO-ND)	Same as D2.2, but at its intended user and with the protection removed (decrypted).
D2.4	MNO database	Input Provider (MNO-ND)	Database composed by the MNO-ND during its normal operations and under its control. The following information is being collected and stored in the database: IMSI (International Mobile Subscriber Identity), timestamp and position.
D2.5	Pseudonymised MNO data	Input Provider (MNO-ND) Host (MNO-VAD)	Pseudonymised version of an extract from the data element D2.4 that is used as input for statistical analysis.
D2.6	Pseudonymised MNO database	Host (MNO-VAD)	Database composed of data elements D2.5 over several periods.

Table 5 Process 2 Data Elements

5.3.3 Security Controls

In this subparagraph the security controls depicted on the operation diagram are being described.

5.3.3.1 SecureChannel

See subparagraph 5.2.3.1.

5.3.3.2 SGXComputation

See subparagraph 5.2.3.2

5.3.3.3 ProtectConfidentiality

ProtectConfidentiality is a generic PE-BPMN action stereotype provided by PLEAK to indicate that some protection mechanism is applied on the input data to protect its confidentiality until action with the OpenConfidentiality stereotype is triggered.

In the context of Sharemind HI and this document, the ProtectConfidentiality stereotype is implemented as follows:

- Sharemind HI enclave generates a new symmetric cryptographic key and uses this to encrypt the data. Encrypted data is stored on disk outside the enclave, but the encryption key stays in the enclave.
- In one of the subsequent actions, a client application can connect to the enclave to retrieve this data. Sharemind HI verifies whether this client is authorised to get this data. If yes, the enclave passes the encrypted data along with the encryption key directly to the client application over a secure authenticated channel (see SecureChannel).

5.3.3.4 OpenConfidentiality

OpenConfidentiality is a generic PE-BPMN action stereotype provided by PLEAK which indicates that a confidentiality protection mechanism is removed from a data element. It works in tandem with ProtectConfidentiality.

In the context of Sharemind HI and this document, the OpenConfidentiality stereotype is implemented as follows:

- It is assumed that a stakeholder has received the encrypted data and the encryption key from the enclave in one of the previous actions (see ProtectConfidentiality).
- This key is used to decrypt data downloaded from the enclave.

`

5.4 P3 Application work process

In this subparagraph the operation of the analysis application is being described: forwarding the MNO data to the Host that runs the analysis application, processing/analysis of the input data and output of the analysis results. Note that MNO-VAD is not depicted as an Output consumer for simplifying the process figure for the Project.

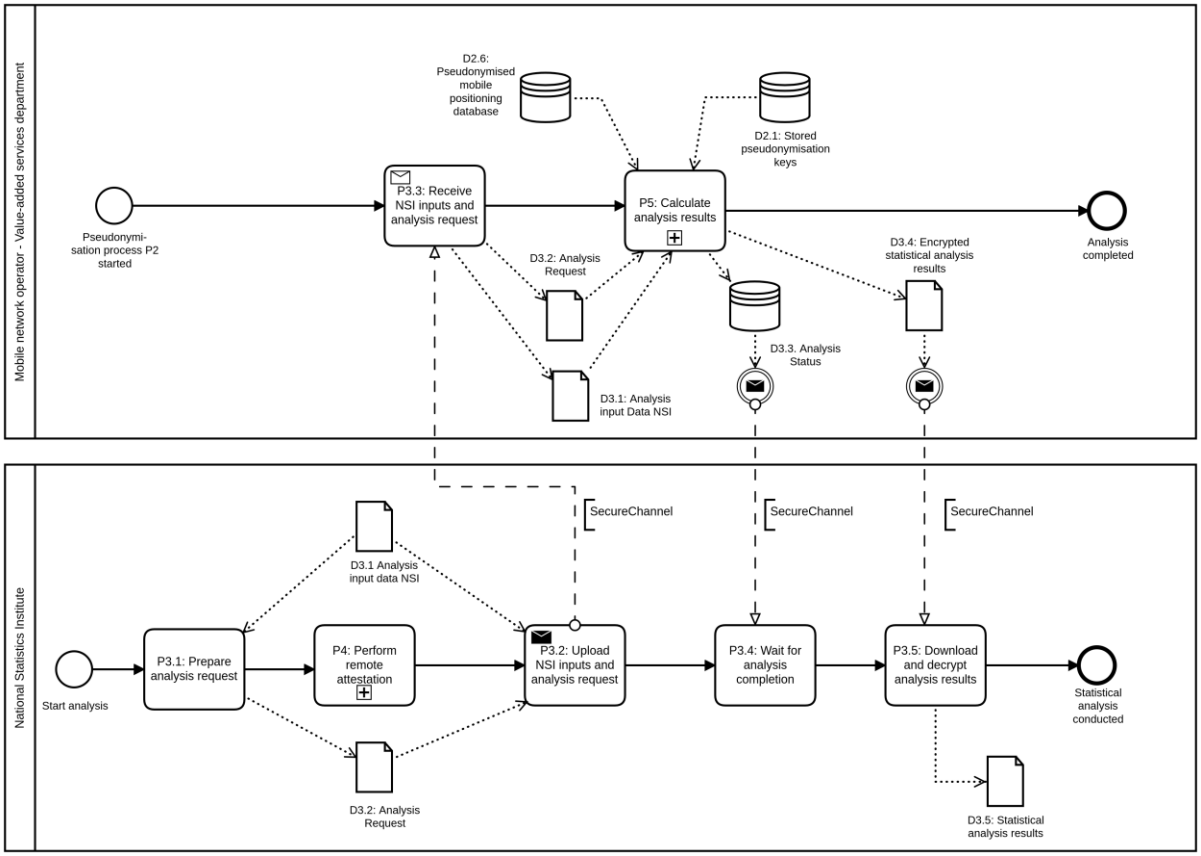

Figure 8: Application work process

5.4.1 Actions

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P3.1	Prepare the analysis request	Input Provider (NSI)	Put together the request.	The Solution is ready for operation and remote attestation performed successfully.	Uploadable contents prepared.

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P4	Perform remote attestation	Input Provider (NSI)	The remote attestation process will be conducted. Remote attestation is described as separate subprocess P4 (see subparagraph 5.5).	Analysis to be started	Remote attestation is performed successfully and trust of the application deployment is verified.
P3.2	Upload NSI inputs and analysis request	Input Provider (NSI)	Send the secret NSI data to Host and a request to Host to run the analysis application with given parameters	Remote attestation succeeded	NSI data and analysis request sent to Host
P3.3	Receive NSI inputs and analysis request	Host (MNO-VAD)	Receive NSI data and analysis request	NSI started upload of NSI inputs and analysis request.	NSI data and analysis request available for analysis process P5
P5	Calculate analysis results	Host (MNO-VAD)	Run the analysis application to fulfil the given analysis request and compose encrypted file with statistical analysis results. Calculation of analysis results is described in more detail through use cases as subprocess P5 (see subparagraph 5.6)	Analysis request has arrived and data has been made available.	Encrypted statistical analysis results have been composed. Analysis Status has been updated.
P3.4	Wait for analysis completion	Output Consumer (NSI)	Poll Analysis status D3.4 until P5 is complete	Analysis has been conducted and the encrypted results have been composed.	End of analysis task P5 is detected

ID	Name	Role (Stake-holder)	Description	Trigger	Goal
P3.5	Download and decrypt analysis results	Output Consumer (NSI)	Download analysis results files, decrypt the files	Analysis completion detected.	Statistical analysis results have been downloaded and decrypted.

Table 6 Process 3 Actions

5.4.2 Data Elements

This section describes the data elements related to the process covered by the previous section. The elements common with P2 are described above in description of P2. Each data element has the following information: a unique identification number referenced from the process diagrams, name of the data element (a BPMN term), name of the holder of the element and a brief description of it.

ID	Name	Holder	Description
D2.1	Stored pseudonymisation keys	Host (SGX TEE)	Pseudonymisation keys that have been generated inside the Sharemind HI enclave upon request from MNO-ND. Each key is independent random value valid for one period only. These keys are **protected** from the Sharemind HI **Host (MNO-VAD)**. The individual keys can be downloaded by **MNO-ND** as D2.2.
D2.6	Pseudonymised MNO database	Host (MNO-VAD)	Database composed of data elements D2.5 over several periods.
D3.1	Analysis input data NSI	Input Provider (NSI) Host (SGX TEE)	Necessary input files for analysis from NSI. **Note that only NSI has access to secret NSI input data!**
D3.2	Analysis request	Input Provider (NSI)	File containing the parameters of the request.
D3.3	Analysis status	Host (MNO-VAD) Output Consumer (NSI)	Analysis status reflects current progress in the Analysis task. It is updated by the analysis task and it can be read by both the **MNO-VAD** and the **NSI**

ID	Name	Holder	Description
D3.4	Encrypted statistical analysis results	Host (MNO-VAD) Output Consumer (NSI)	The statistical report that is produced as a result of the operation of the analysis application and is encrypted to be transmitted to the Output Consumer.
D3.5	Statistical analysis results	Output Consumer (NSI)	Decrypted statistical analysis report.

Table 7 Process 3 Data Elements

5.4.3 Security Controls

Application work process doesn't incur new security controls that need to be described. All necessary security controls have been described previously.

Any additional security controls required for any statistical analysis will need to be assessed case-by-case.

5.5 P4 Remote Attestation Process

In order to establish trust and ensure a proper and secure setup of the Solution, all parties connecting to the Solution (all stakeholders in the roles of Coordinator, Enforcer, Input Provider, Output Consumer, Runner, Auditor) will need to perform the remote attestation process. Without this the client application will not be able to communicate with the Solution (Sharemind HI server) and carry out further activities. Even if one of the verifications carried out as a part of this process fails, the Sharemind HI client application will drop the process and refuse to establish a connection.

Remote attestation may need to be repeated upon need. For instance:

- In case of the Sharemind HI server restart process;
- A stakeholder has misplaced their session key;
- In case the Solution, for some reason, is stopped and restarted.

Using remote attestation, the client application can verify the following before proceeding with any further interaction with the service (e.g., upload of secret data):

- The expected enclave was created on a remote machine that supports Intel SGX instructions. This ensures that the safety guarantees provided by enclaves (data protection from the external environment) are met.
- The Solution is running a particular version of the Sharemind HI application server. Software versions are being compared using cryptographic fingerprints. It ensures that Sharemind HI safety procedures are followed.
- The Sharemind HI application server is running trusted software (in this case, the statistical analysis application that is running in an enclave). The Sharemind HI application server will compare the application's cryptographic fingerprint with its actual fingerprint. This ensures that analysis is being conducted on a statistical analysis application that was audited, verified and agreed upon.

For further, more technical, information on remote attestation please see section *Intel® Software Guard Extensions.*

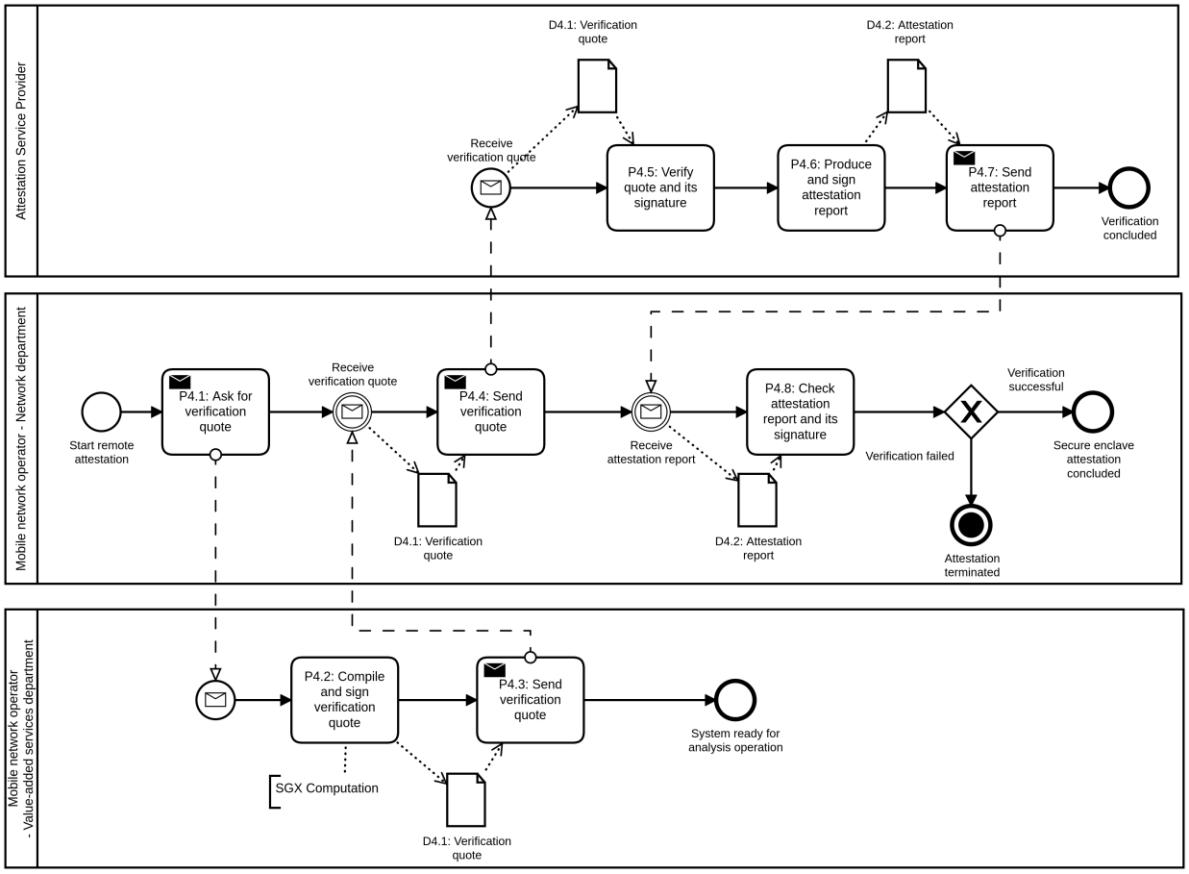

Figure 9: Remote attestation process

Please note that the above figure only covers the process from the MNO-ND viewpoint. In general, all stakeholders using client application to communicate with the Solution (i.e. those in Coordinator, Enforcer, Input Provider, Output Consumer, Runner or Auditor role) are required to carry out the remote attestation process in order to take part in the Solution (for instance, subprocess P4 on Figure 8: Application work process). In the tables below, such stakeholders are shown with the role **"Verifier"** for brevity.

5.5.1 Actions

In this subparagraph the actions, their triggers, goals and stakeholders in remote attestation process are described in detail. Each activity has the following information: a unique identification number referenced from the process diagrams, the name of the process, the name of the role who performs the activity (with the reference to a possible concrete stakeholder in PoC), a brief description of the activity, the trigger and the goal of the activity.

ID	Name	Role	Description	Trigger	Goal
P4.1	Ask for verification quote.	Verifier (MNO-ND)	The client application asks for verification quote from the Host.	Remote attestation has been started.	Verification quote, (attestation evidence), has been asked from the Host.

ID	Name	Role	Description	Trigger	Goal
P4.2	Compile and sign verification quote.	Host (MNO-VAD)	Host compiles and signs the verification quote.	Verification quote has been asked.	Verification quote has been compiled and signed.
P4.3	Send verification quote.	Host (MNO-VAD)	Host sends the signed verification quote to client application.	Signed verification quote is ready to be sent.	Verification quote is sent to client application.
P4.4	Receive and send verification quote.	Verifier (MNO-ND)	The client application receives the signed verification quote and re-sends it to the Attestation Service Provider.	Verification quote has been received.	Verification quote is sent to Attestation Service Provider.
P4.5	Verify quote and its signature.	Attestation Service Provider (Intel via Cybernetica proxy)	Attestation Service Provider verifies the quote and its signature.	Verification quote has been received.	Verification procedures have been concluded.
P4.6	Produce and sign attestation report.	Attestation Service Provider (Intel via Cybernetica proxy)	Attestation Service Provider produces a verification report and signs it.	Verification procedures have been concluded.	Producing a signed verification report.
P4.7	Send attestation report.	Attestation Service Provider (Intel via Cybernetica proxy)	Attestation Service Provider sends the signed verification report to the client application.	A signed verification report has been produced.	A signed verification report has been sent to the client application.
P4.8	Check attestation report and its signature.	Verifier (MNO-ND)	The client application checks the verification report and its signature. In case the verification was declined, the attestation is terminated.	A signed verification report has been received.	Secure enclave attestation concluded.

Table 8 Process 4 Actions

5.5.2 Data Elements

This section describes the data elements related to the process covered by the previous section. Each data element has the following information: a unique identification number referenced from the process diagrams, name of the data element (a BPMN term), name of the holder of the element and a brief description of it.

ID	Name	Holder	Description
D4.1	Verification quote	Host (MNO-VAD) Verifier (MNO-ND) Attestation Service Provider (Intel via Cybernetica proxy)	A cryptographic measurement of the current running application enclave which is generated by the Intel SGX platform and signed with the platform's EPID key. Only the Intel Attestation Service can verify this signature. During remote attestation the verification quote (D3.1) is delivered to the Verifier who needs to decide whether the enclave can be trusted.
D4.2	Attestation report	Attestation Service Provider (Intel via Cybernetica proxy)	A report issued by the Intel Attestation Service containing the attestation status for the verification quote. The report indicates whether the enclave being verified has been tampered with, whether it is running on the genuine platform with Intel SGX enabled, and whether it is running at the latest security level. The report is digitally signed by the Intel Attestation Service using the Report Signing Key and verified by the client using the Attestation Report Signing CA Certificate that is publicly available on the IAS portal.

Table 9 Process 4 Data Elements

5.5.3 Security controls

Remote attestation process doesn't incur new security controls that need to be described. All necessary security controls have been described previously.

5.6 P5 Specified Use Cases

This chapter describes the use cases of the Solution specified and proposed by Eurostat.

The business process of these use cases is a specification of the process task P3.3 “Calculate analysis results”. The Sharemind HI solution conducting computations is hosted by MNO-VAD in the role of Host (see subparagraph 4.3.1).

5.6.1 Use Case Description

Eurostat has defined two use cases #1 and #2 (see Annex “ESTAT 2019.0232 Use case description”), where use case #1 is a subset of use case #2. Further below we only address the more general 2nd use case and refer to it as the use case. The use case #1 is also covered by this solution.

The use case goal is to calculate aggregated statistics for specified time periods (e.g. 3 months) for each tile (ca 1Cybernetica x 1Cybernetica area). The 24h daily cycle is divided into IT = 3 daily sub-periods, not necessarily disjoint, corresponding to night time, working time and evening.

Aggregated statistics is calculated separately for each daily sub-period

1. 24h full day
2. Night time
3. Working time
4. Eventing

The goal of the use case is to calculate two separate spatial demography reports:

a) total footprint per tile aggregated over report period, separately for in each daily sub-period;

b) Functional Urban Footprint over report period (FUF). FUF is a statistic inspired by the concept of Functional Urban Area.

c) Urban Area

The contents of the reports and the calculation algorithms are described in Annex “ESTAT 2019.0232 Use case description”.

The use case input is footprint of each single person within one day as a set of tiles visited, for each daily sub-period separately.

5.6.1.1 Statistical Disclosure Control

The computations are performed in Trusted Execution Environment (TEE) so intermediate results are not accessible to anyone. This is guaranteed by task code structure and is validated during attestation. All reports include Statistical Disclosure Control (“**SDC**”) step before finalising.

For SDC reasons, all elements below a given threshold parameter are either omitted or merged with neighbouring elements so that the combined sum exceeds the threshold in the reported output. SDC procedure is closely related to the processing algorithm and specified together with the algorithm in Annex “ESTAT 2019.0232 Use case description”.

The threshold parameters are fixed in task and cannot be changed after attestation. This guarantees that SDC mechanism is not switched off.

5.6.2 Business Process Description

First the chapter describes the business process of the use case and following sections elaborate on the elements of the process (actions and data elements).

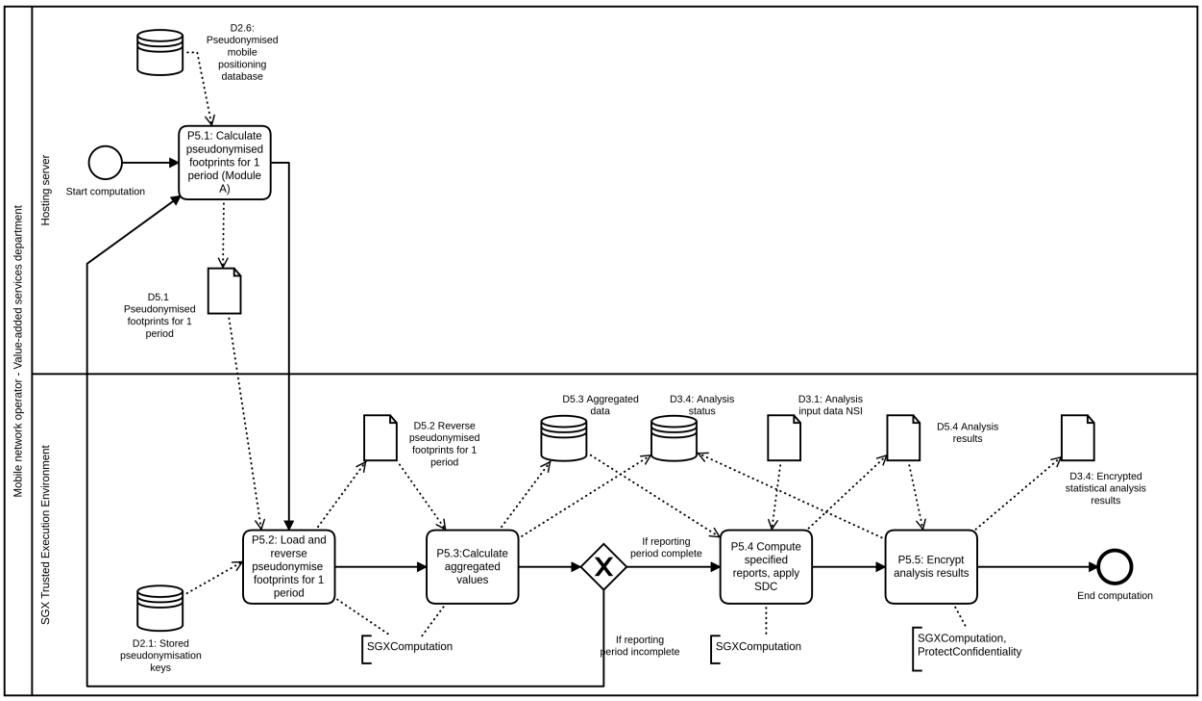

Figure 10: Use Case Process

5.6.2.1 Actions

In this subparagraph the actions, their triggers, goals and stakeholders in remote attestation process are described in detail. Each activity has the following information: a unique identification number referenced from the process diagrams, the name of the process, the name of the role who performs the activity (with the reference to possible concrete stakeholder in PoC), a brief description of the activity, the trigger and the goal of the activity.

ID	Name	Role	Description	Trigger	Goal
P5.1	Calculate pseudonymised footprints for 1 period (Module A)	Host (MNO -VAD)	Hosting Server uses pseudonymised mobile data from MNO-VAD database to calculate pseudonymised footprints (H in algorithm description) and stores result to hosting server file system. This functionality “Module A” is not part of Cybernetica	Analysis launch order has arrived.	Pseudony-mised footprints data for 1 period have been exported to file in predefined format.

ID	Name	Role	Description	Trigger	Goal
			solution and is implemented in another Eurostat project.		
P5.2	Load and reverse pseudonymised footprints for 1 period	Host (SGX TEE)	Sharemind HI Server loads data from file and performs reverse pseudonymisation on MNO footprint data.	Pseudonymised mobile footprint data file is present in file system	MNO footprint data is converted into a format where each subscriber's footprint data can be analysed over a longer time period.
P5.3	Calculate aggregated values	Host (SGX TEE)	Sharemind HI server computes the aggregations needed to produce specified reports. Update Analysis status to reflect progress.	Reverse pseudonymisation has been conducted	All available data is accounted into aggregation results
P5.4	Compute specified reports, apply SDC	Host (SGX TEE)	Sharemind HI server computes the specified reports (FUF and total footprint)	All available data over report period has been aggregated	Reports are ready and saved inside SGX TEE
P5.5	Encrypt analysis results	Host (SGX TEE)	Sharemind HI Server encrypts and stores analysis results. Update Analysis status to reflect completion.	The reports of FUF and total footprint have been computed	Analysis results have been encrypted and stored, Analysis status is updated.

Table 10 Process 6 Actions

5.6.2.2 Data Elements

Some data elements used in this process are already described under processes P2 and P3.

ID	Name	Holder	Description
D2.6	Pseudonymised MNO database	Host (MNO-VAD)	Database composed of data elements D2.5 over several periods.
D5.1	Pseudonymised footprints for 1 period	Input Provider (MNO-VAD)	Footprint data for one period (usually 1 day), using the frequently changed pseudonym
D5.2	Reverse pseudonymised footprints for 1 period	Host (SGX TEE)	Footprint data for one period, using the long-term pseudonym. No one can see the footprint data, it remains inside TEE
D5.3	Aggregated data	Host (SGX TEE)	Aggregated for one report, using the long-term pseudonym. No one can see the data, it remains inside TEE
D5.4	Analysis results	Host (SGX TEE)	Analysis results, not encrypted. No one can see the data directly; it remains inside TEE. (same data will be available for download after encryption)
D3.4	Encrypted statistical analysis results	(SGX TEE) Host (MNO-VAD)	The statistical report that is produced as a result of the operation of the analysis application and is encrypted to be transmitted to the Output Consumer.

Table 11 Process 5 Data Elements

5.6.2.3 Security Controls

Use Case process does not add new security controls. All necessary security controls have been described previously.

6 Data Elements Visibility

To verify that solution security requirements around data visibility are met, it is important to verify who sees which data elements in the solution. Note, that the data elements can be data, keys, application code, requests moving between stakeholders or results of calculations. The following table presents the visibility analysis of data elements in the Pseudonymisation process (P2) and Application work process (P3). The visibility analysis focuses on the inputs and computation process of the analysis since SDC guarantees the statistical output does not represent personal data.

The table below lists all stakeholders and shows, whether and how the stakeholder sees a data element. Each cell is marked with:

- V (visible): the contents of the data element are fully visible to the stakeholder. V* denotes planned visibility of results for the MNO-VAD to support trust of the MNO in the Project, while for the sake of simplicity, the MNO-VAD was not shown on the P3 Application work process BPMN diagram and can technically be restricted from seeing the process output.
- H (hidden): the stakeholder has the data element but this element is protected with security measures. In effect, the stakeholder cannot see the content of the data element (in cryptographic terminology, the stakeholder sees the ciphertext, but not the plaintext).
- O (owner): the stakeholder is the creator/owner of the data element and it is fully visible to the stakeholder.
- "-": the stakeholder does not see the data element in any way in the process.
- S (SecureChannel): the data element is only transmitted over secure communication channels.

The preliminary analysis was conducted by the analyst using the Pleak.io analysis tool. The final elements chosen for the table are selected to provide clearer understanding for the reader. (i.e., key - plaintext instead of the technical element: protected key - cyphertext) and reflect the workflows represented on Figure 7: MNO data pseudonymisation process and Figure 8: Application work process.

The visibility of different data elements depends on the designed and agreed workflow.

For traceability Appendix 2 contains technical visibility tables generated from the individual process BPMN diagrams above.

	MNO - ND	MNO-VAD	NSI	Shared over
Analysis input data NSI	-	H	O	S
Analysis request	-	V	O	S
Periodic pseudonymisation key	V	H	-	S
MNO database	O	-	-	-
Pseudonymised MNO data	O	V	-	S
Pseudonymised MNO database	-	O	-	-

Reverse pseudonymised data	-	H	-	-
Statistical analysis results	-	V*	O	-

Table 12 Summary Data Elements Visibility

V* marks planned visibility of results for the MNO-VAD to support trust of the MNO in the Project.

The visibility analysis shows the following:

1) The processing of data on Sharemind HI server Host's (MNO-VAD in this Project) incurs minimal risks since the pseudonymisation keys are generated inside the Sharemind HI enclave and are only visible to the MNO-ND.
2) Reverse pseudonymised data remains hidden from the host of the environment. The Host coordinates the analysis process and applies analysis methods (runs the applications) but is unable to extract protected data. Their access is limited to measuring the progress of the analysis application (for instance, knowing the amount of data in various stages of analysis).

Appendix 1 – Details of Data Visibility

- V (visible): the contents of the data element are fully visible to the stakeholder. V* denotes planned visibility of results for the MNO-VAD, while for the sake of simplicity the visibility to the MNO-VAD was not shown on BPMN diagrams.
- H (hidden): the stakeholder has the data element but this element is protected with security measures. In effect, the stakeholder cannot see the content of the data element (in cryptographic terminology, the stakeholder sees the ciphertext, but not the plaintext).
- O (owner): the stakeholder is the creator/owner of the data element and it is fully visible to the stakeholder. The owner is marked only on global process
- "-": the stakeholder does not see the data element in any way in the process.
- S (SecureChannel): the data element is only transmitted over secure communication channels.

6.1 Global Process P2+P3+P4+P5 Visibility (excluding configuration P1 and attestation P4)

ID	Data	MNO - ND	MNO-VAD	NSI	Shared over
D2.1	Stored pseudonymisation keys	-	H	-	-
D2.2	Protected New periodic pseudonymisation key	V	H	-	S
D2.3	New periodic pseudonymisation key	V	-	-	-
D2.4	MNO database	O	-	-	-
D2.5	Pseudonymised MNO data	O	V	-	S
D2.6	Pseudonymised MNO database	-	O	-	-

D3.1	Analysis input data NSI	-	H	O	S
D3.2	Analysis request	-	V	O	S
D5.1	Pseudonymised footprints for 1 period	-	O	-	-
D5.2	Reverse pseudonymised footprint data for 1 period	-	H	-	-
D5.3	Aggregated data	-	H	-	-
D3.4	Encrypted statistical analysis results	-	V*	O	S
D3.5 D5.4	Analysis results	-	V*	O	S

Table 13 Global process P2+P3+P4+P5 Visibility (excluding configuration P1 and attestation P4)

6.2 Pseudonymisation Process P2 visibility

ID	Data	MNO - ND	MNO-VAD	NSI	Shared over
D2.1	Stored pseudonymisation keys	-	H	-	-
D2.2	Protected new periodic pseudonymisation key	V	V	-	S
D2.3	New periodic pseudonymisation key	V	H	-	-
D2.4	MNO database	O	-	-	-
D2.5	Pseudonymised MNO data	O	V	-	S
D2.6	Pseudonymised MNO database	-	O	-	-

Table 14 Pseudonymisation Process P2 visibility

6.3 Application Process P3 and sub-process P5 visibility

ID	Data	MNO - ND	MNO-VAD	NSI	Shared over
D2.1	Stored pseudonymisation keys	-	H	-	-
D2.6	Pseudonymised MNO database	-	V	-	-
D3.1	Analysis input data NSI	-	H	O	S
D3.2	Analysis request	-	V	O	S
D3.4	Encrypted statistical analysis results	-	V*	V	S
D3.5	Statistical analysis results	-	V*	V	S
D5.1	Pseudonymised footprints for 1 period	-	O	-	-
D5.2	Reverse pseudonymised footprints for 1 period	-	H	-	-
D5.3	Aggregated data	-	H	-	-
D5.4	Analysis results	-	H	-	-

Table 15 Application Process P3 visibility

Appendix 2 – Stakeholder-Role Matrix by Task

6.4 Stakeholders and Roles Matrix for Pseudonymisation Task (by Dataflow Configuration)

		Stakeholders				
		MNO-ND	**NSI**	**MNO-VAD**	**External Auditor**	**Intel via Cybernetica proxy**
Roles	**Sharemind HI server Host**			+		
	Coordinator		+			
	Enforcer	+	+	+	+	
	Input Provider					
	Output Consumer	+[25]				
	Runner	+				
	Developer		+			
	Auditor		+	+	+	
	Attestation Service Provider					+

Table 16 Roles of Stakeholders in the Pseudonymisation task

[25] MNO-ND is in role of Output Consumer as it reads the keys produced for short-term pseudonymisation.

6.5 Stakeholders and Roles Matrix for Analytics Task (by Dataflow Configuration)

		Stakeholders				
		MNO-ND	**NSI**	**MNO-VAD**	**External Auditor**	**Intel via Cybernetica proxy**
Roles	**Sharemind HI server Host**			+		
	Coordinator		+			
	Enforcer	+	+	+	+	
	Input Provider		+	+		
	Output Consumer		+	+[26]		
	Runner			+		
	Developer		+			
	Auditor		+	+	+	
	Attestation Service Provider					+

Table 17 Roles of Stakeholders in the Analytics Task

[26] For the PoC we recommend the MNO-VAD to also see the outputs of analytics too in order to improve confidence in the process.

Appendix 3 – PLEAK Open-source Process Analysis Software

In this Project, we use open-source software PLEAK (Privacy LEAKage) for designing and analysing privacy-aware processes. Cybernetica researches and develops PLEAK together with Tartu University.

PLEAK is a next-generation analysis tool for the privacy audit of an existing system and the design of new privacy-aware systems. PLEAK lets analysts model business process using the Business Process Model Notation (BPMN) and privacy-preserving algorithms using the SecreC privacy-preserving programming language. PLEAK can then analyse the data flows that use cryptographic privacy and differential privacy. PLEAK supports the inclusion of Privacy Enhancing Technologies (PETs) in the business process models to reduce leakage of private information.

PLEAK has been developed by Cybernetica AS and the University of Tartu under the DARPA Brandeis privacy technology development program[27] since 2015. The tool is part of the Cybernetica privacy-aware system analysis methodology and used in every Cybernetica project needing privacy-aware system analysis or design.

In the current Project PLEAK is used for modelling, analysing and publishing business processes. See example of a process with a live analytical model in PLEAK[28].

[27] Press release related to start of Pleak development: https://cyber.ee/news/2015/11-23/

[28] Example process with a live model https://cyber.ee/blog/2020/04-20/