D1. ‘STATE OF THE ART OVERVIEW, USE CASE ANALYSIS AND PRELIMINARY TECHNICAL SPECIFICATION OF THE SOLUTION’

UTIP-DAM

21/03/2024

Funded by the European Union

TrustChain Project. Funded by the European Union under GA No 101093274. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Grant Agreement No.: 101093274
Call: HORIZON-CL4-2022-HUMAN-01
Topic: HORIZON-CL4-2022-HUMAN-01-03
Type of action: RIA

D1. 'STATE OF THE ART OVERVIEW, USE CASE ANALYSIS AND PRELIMINARY TECHNICAL SPECIFICATION OF THE SOLUTION'

>UTIP-DAM<

Due date	22/03/2024
Submission date	21/03/2024
Team	Erel Rosenberg Clara Lee Iris Desbrousses Thanda Oo Janine Son Hyunsoo Gang Minseong Kim Bitna Gu Clea Rozenblum
Version	V1.0
Authors	Erel Rosenberg Clara Lee Iris Desbrousses Clea Rozenblum

EXECUTIVE SUMMARY

In today's data-driven landscape, understanding human mobility is pivotal for various stakeholders, including government entities addressing climate change, businesses refining urban experiences, and healthcare institutions monitoring disease outbreaks. However, the acquisition and analysis of mobility data present significant privacy challenges, with conventional anonymization techniques often failing to adequately protect individuals' privacy rights.

To tackle this issue, the UtiP-DAM project introduces an innovative approach to anonymizing mobility data. Departing from centralized methods where individuals entrust their data to a single entity, UtiP-DAM employs a decentralized and verifiable method, ensuring privacy preservation while empowering both individuals and organizations.

Central to UtiP-DAM is its decentralized anonymization process, conducted at the data collection point rather than by a central authority. This strategy minimizes the risk of re-identification, safeguarding individuals' privacy while maintaining the data's utility for research and planning endeavors.

Additionally, UtiP-DAM offers advanced verification tools for individuals and organizations alike. Individuals can verify whether their personal mobility data has been included in public datasets, enhancing transparency and accountability. Meanwhile, organizations can audit their datasets for de-anonymization risks and utilize the UtiP-DAM algorithm to re-anonymize them effectively.

The project's web-based functionalities, such as anonymity verification and anonymization tools, will be accessible through the NGI TrustChain > UtiP-DAM GitHub repository and a dedicated webpage hosted by Correlation Systems. This dual approach ensures accessibility for all stakeholders, combining open-source utilization with a pay-per-use model.

UtiP-DAM not only addresses existing anonymization risks associated with centralized control but also fosters a more privacy-enabling, and open environment for sharing mobility data. By decentralizing anonymization and promoting open-source tools, the project bridges the gap between privacy concerns and innovative solutions, ultimately enhancing privacy rights protection and instilling trust in data-driven solutions.

In essence, UtiP-DAM signifies a transformative initiative that empowers individuals and organizations to engage in mobility research and planning while safeguarding privacy rights. With its decentralized methodology and advanced tools, UtiP-DAM sets a new standard for data anonymization, enabling secure and ethical data utilization for the betterment of society.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

IDS	International Data Space
IoT	Internet of Things

1 INTRODUCTION

In today's interconnected world, understanding human mobility is of paramount importance in various industry verticals, ranging from urban planning and public health to commercial enterprises. However, this necessity is juxtaposed with the pressing concern of safeguarding individual privacy rights. Traditional approaches to data anonymization fall short, as even seemingly anonymized data can be re-identified, posing significant privacy risks.

In this context, the UtiP-DAM project proposes a groundbreaking initiative aimed at revolutionizing the anonymization of mobility data. Developed by Correlation Systems, an Israeli SME with a rich history in AI and IoT technologies, UtiP-DAM offers a decentralized approach to data anonymization, empowering individuals and organizations alike. By shifting the anonymization process to the point of data collection, UtiP-DAM ensures privacy protection without compromising data utility.

Moreover, UtiP-DAM provides innovative verification tools for individuals to ascertain if their mobility data has been included in public datasets. For organizations, UtiP-DAM offers anonymity auditing and anonymization tools, ensuring compliance with privacy laws and mitigating de-anonymization risks.

UtiP-DAM represents a paradigm shift in data privacy and empowerment, enabling the seamless integration of mobility data into research and decision-making processes while upholding individual rights. It's a manifestation of trust and responsibility in the digital age, where data-driven insights coexist harmoniously with privacy protection.

2 MOTIVATION AND PLANNED FUNCTIONALITIES

Correlation Systems is developing UtiP-DAM with a core focus on customer needs in the context of privacy-compliant crowd data management. Our innovative solution addresses the challenges associated with adhering to strict local privacy laws when collecting, storing, and analyzing crowd-sourced data.

Notably, UtiP-DAM will empower users, including its existing customers, to:

- Distribute data openly, securely and in compliance with data privacy laws: Through UtiP-DAM, dataset owners will be able to share anonymized data in open-source frameworks such as the International Data Space, and private frameworks such as the UtiP-DAM marketplace, fostering collaboration and innovation for mobility and location- based solutions.
- Unlock new revenue streams: the UtiP-DAM marketplace will enable the resell of anonymized data while protecting individual privacy, potentially generating additional income for Correlation Systems and dataset owners.

By addressing these market needs, Correlation Systems aims to achieve the following benefits:

- Enhanced customer satisfaction: UtiP-DAM will provide customers with the tools and confidence to navigate the complexities of data privacy regulations.
- Increased market potential for our IoT sensors and crowd analytics solutions: the UtiP-DAM project can attract new customers hesitant about data privacy concerns, expanding market possibilities for Correlation Systems.
- Stronger public trust: UtiP-DAM empowers citizens with tools to verify that their personal information is not exposed by the system. This fosters public trust and transparency in the use of IoT data collection technologies.

To realize these benefits, a set of features will be implemented during the project:

- Functionality 1 - Core auditing and k-anonymization
- Functionality 2 - Mobility Raw Data Marketplace
- Functionality 3 - Anonymization Auditing Tool
- Functionality 4 - Anonymization tool for companies
- Functionality 5 - Anonymized data sharing
- Functionality 6 - User accounts
- Functionality 7 - Dataset purchase
- Functionality 8 - API for the marketplace
- Functionality 9 - Distributed EDGE network
- Functionality 10 - Add download raw data function to the existing dashboard

For a detailed description of each functionality, see below.

2.1 Functionality 1 - Core auditing and k-anonymization

Utip-DAM will actively contribute to the developer community by providing a suite of open-source K-anonymity tools. These tools address a critical need for developers working with mobility data: ensuring privacy while preserving data utility.

Below is a breakdown of the functionalities offered and their significance:

- General auditing tool: this tool analyzes a mobility data file (CSV format) and outputs the lowest K value within the data. This provides a quick assessment of the overall anonymization strength of the dataset.
- Detailed auditing tool: this tool delves deeper, providing a K value for each individual mobility sequence within the data. This granular analysis enables developers to pinpoint specific areas within the dataset that might require additional anonymization measures.
- K-anonymization tool: This tool addresses the core challenge of anonymizing mobility data. It replaces identifiers of mobility patterns below K using random IDs. This anonymization technique ensures that individual movements cannot be linked with another, while preserving valuable data insights for analysis.

These open-source tools will empower developers in several ways:

- Developers can leverage these tools to easily integrate K-anonymization into their own applications without building them from scratch. This will help save development time and resources.
- By incorporating these tools into their applications, developers can ensure they comply with data privacy regulations like GDPR and the upcoming Data Act. This builds trust with users and allows responsible data analysis.
- The K-anonymization tool strikes a balance between privacy and data usability. Anonymization occurs only for patterns with low occurrence, minimizing the impact on data insights.

While these tools are valuable to the developer community, they also play a crucial role within Utip-DAM itself.

The auditing functionalities will be integrated within the marketplace, enabling data owners, data consumers and individuals to leverage the outputs of the UtiP-DAM project.

By offering open-source K-anonymity tools, Utip-DAM will foster a collaborative environment within the open-source community, not only benefitting individual developers but also promoting responsible data anonymization practices within the broader mobility data landscape.

2.2 Functionality 2 - Raw Mobility Data Marketplace

2.2.1 General Overview

The UtiP-DAM marketplace provides a secure platform for the general public, as well as private and public institutions, to download anonymized mobility data sets collected by various data owners. This data can be made available at cost, or offered for a fee, as determined by the dataset owner.

There are two main reasons why Correlation Systems has chosen to use the marketplace medium as one way to distribute mobility dataset (bear in mind that dataset owners will also be able to distribute their data via open repositories and frameworks, such as the International Data Space):

- New revenue streams: the marketplace can enable data owners to generate additional revenue by selling anonymized data sets to interested parties. This can incentivize data sharing and contribution by third-parties to the platform.
- Enabling innovation: mobility data is currently a scarce resource, hindering research and development in various fields. In this contect, the UtiP-DAM marketplace can be a game-changer in facilitating the secure sharing of anonymized data, fostering collaboration and innovation within the research community and private sector.

The marketplace will be deployed at ngi.cs.co.il (in fact, a prototype version of the marketplace is already available) and will provide the following landing page:

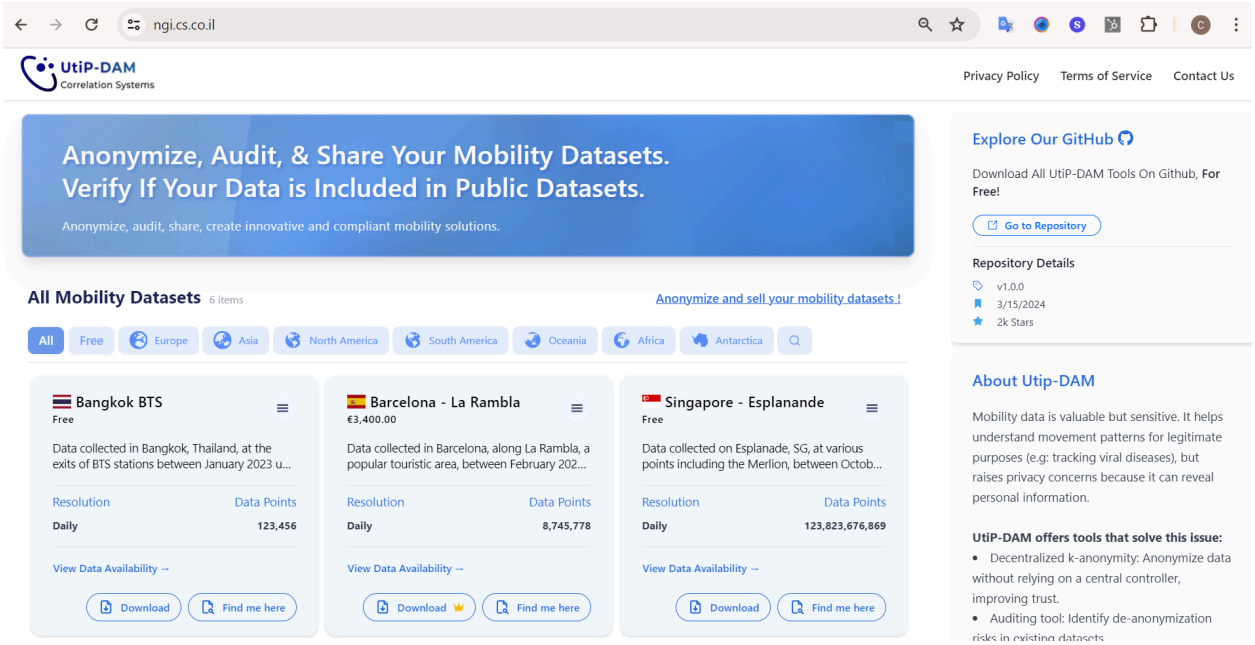

Figure 1 - Landing Page of the UtiP-DAM marketplace (*under development*)

The landing page is divided into three main sections:

- The header: where the UtiP-DAM logo and menu is available,
- The left-side panel, which is the main part of the page, showcasing the available datasets and available features,
- The right-side panel, which links to the GitHub repository of the project, and below that, provides additional information on the project (including the mention of NGI funding).

While the header and right-side panel are self-explanatory, it is worth delving into the left-side panel and the functionalities offered within.

2.2.2 The Left-side panel

The left-side panel includes the list of datasets that are available on the marketplace, with a possibility to filter datasets based on location, and price (free versus premium), as well as a search function. A short description of the dataset, number of available data and the cost of downloading a dataset will be provided.

The user will be able to visualize available days in the dataset (i.e: days for which data is provided - we assume that data may not always be provided continuously, as it is the case for some of our customers, who use their Correlation Systems' devices during the workweek, or for special events) by clicking on "View Data Availability".

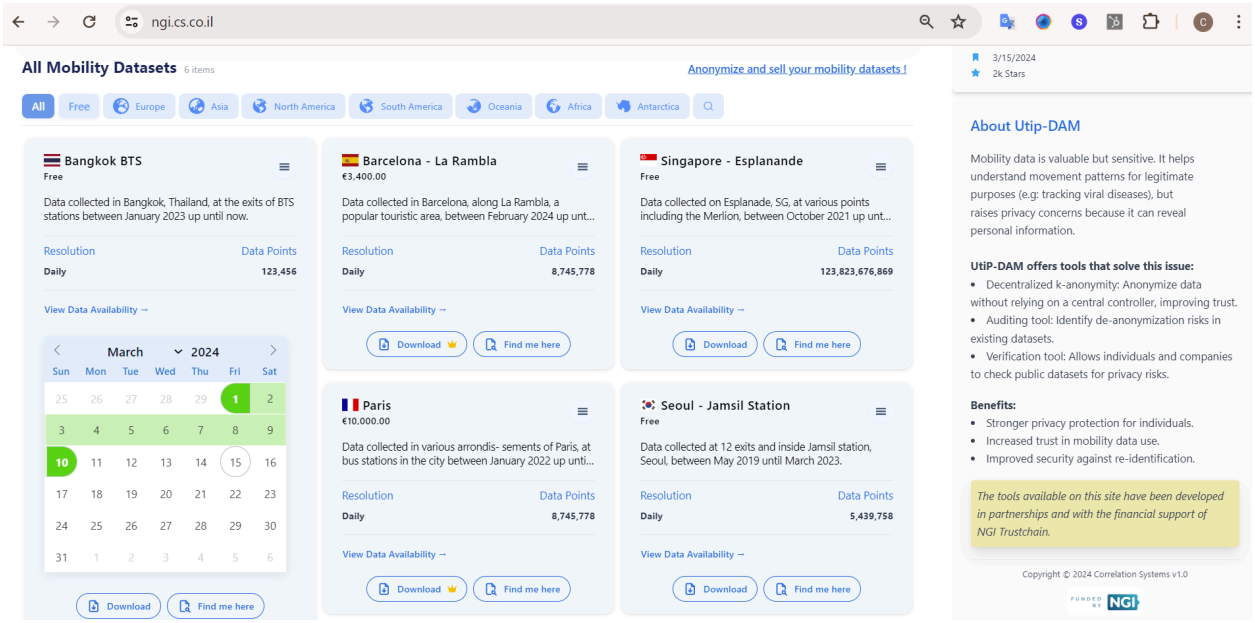

Figure 2 - Landing Page with Available Date View (*under development*)

The landing page of the marketplace (https://ngi.cs.co.il) will provide access to two two key features of the UtiP-DAM

- The anonymization auditing tool, available via the “Find me here” button, under each dataset: enables anyone to check if their personal data (i.e: their personal trajectories are included in a dataset)
- The k-anonymizer tool (+ sharing tool), available via the “Anonymize and share your mobility datasets !” link above the filters: enables dataset owners to anonymize their data, and optionally: share their data, via the marketplace.

The following sections detail each feature.

2.3 Functionality 3 - Anonymization Auditing Tool

The UtiP-DAM marketplace offers a free auditing tool that empowers individuals to check if their personal trajectory might be present in a specific dataset.

The tool represents a significant advancement in data privacy for individuals. This innovative feature empowers individuals, in a user-friendly manner, to verify the potential presence of their data within specific mobility datasets. By doing so, the tool offers key benefits for the general public:

- Unprecedented User Control: For the first time, individuals will gain the ability to assess the presence of their journeys within specific datasets. This fosters a sense of agency and promotes trust in the data anonymization process.
- Enhanced Transparency: The marketplace, and the auditing tool itself, foster transparency by providing users with the ability to understand how their data might be leveraged for research or other legitimate purposes.
- A Unique Service: To the best of our knowledge, no comparable tools exist that empower individuals to verify their data's inclusion within public datasets.

Here is how it works:

1. User selects a dataset: User chooses the specific dataset they believe their data may be included in, among the available options on the marketplace.
2. User specifies date and mobility pattern: User selects a day their data may have been captured, and selects among the data points available, their trajectory (e.g: if 5 data points are shown on a map - meaning that the dataset includes location data only for these 5 data points - the user should select, in the order of their trajectory, each data point that matches their journey)
3. Risk Assessment: The system analyzes the user’s input and provides a risk level for potential de-anonymization based on a K-anonymity score. This score indicates the likelihood of someone identifying them from the anonymized data.

The tool prioritizes user privacy by enabling users to audit datasets without requiring any form of login or personal identification – they simply select data points among those already available in the dataset.

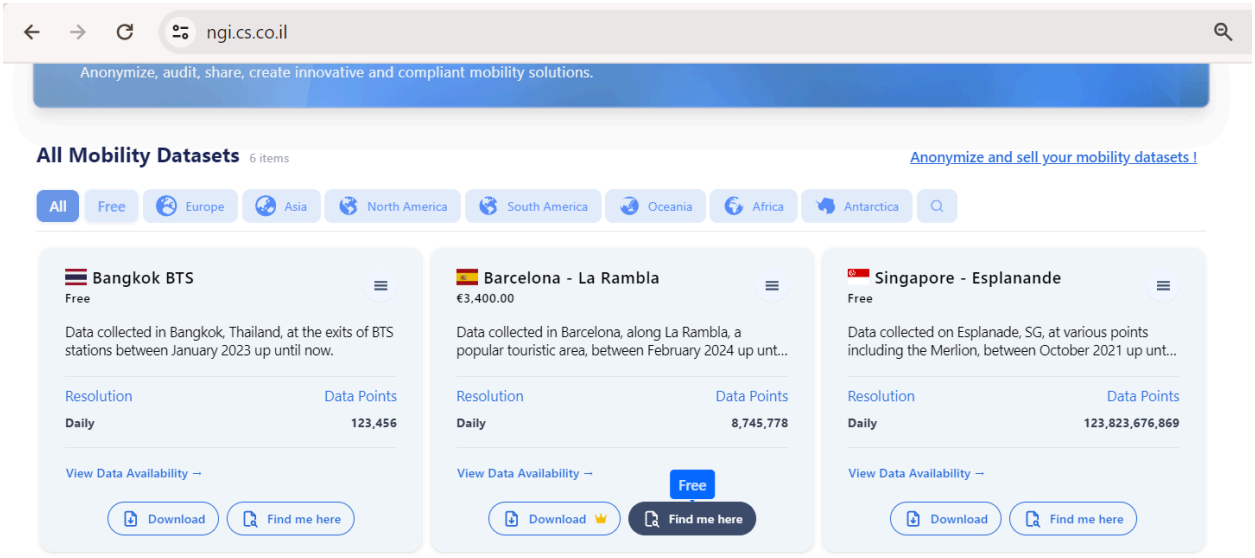

Figure 3 - Landing Page with Find me here button (*under development*)

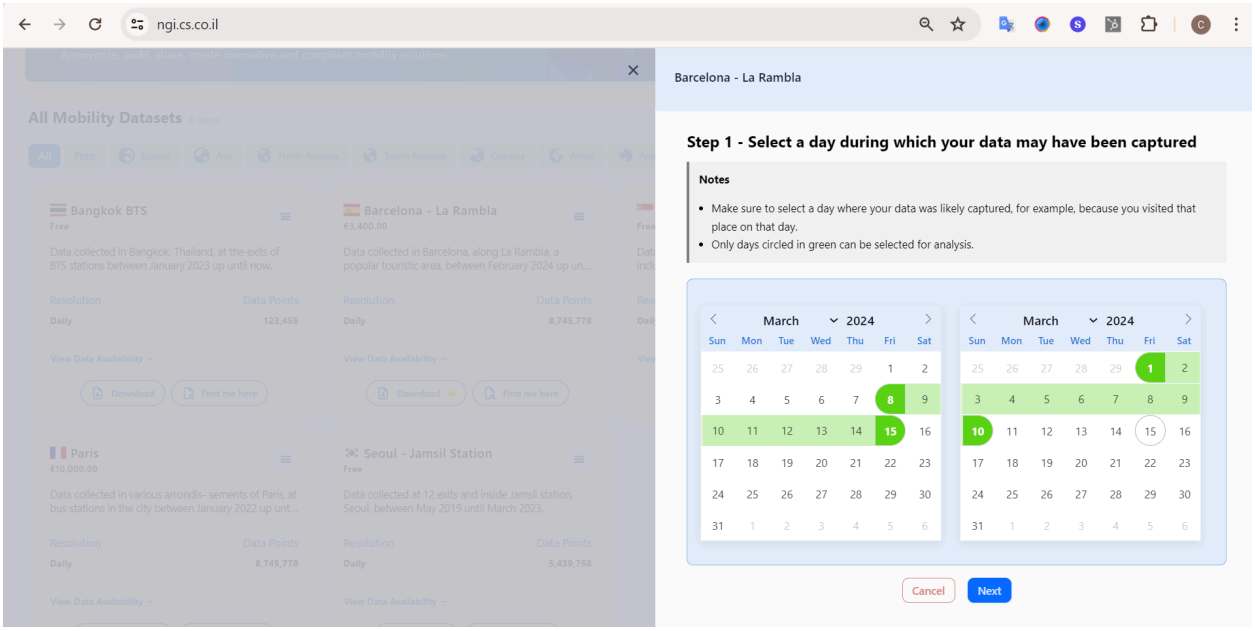

Figure 4 - Find me here - step 1 (*under development*)

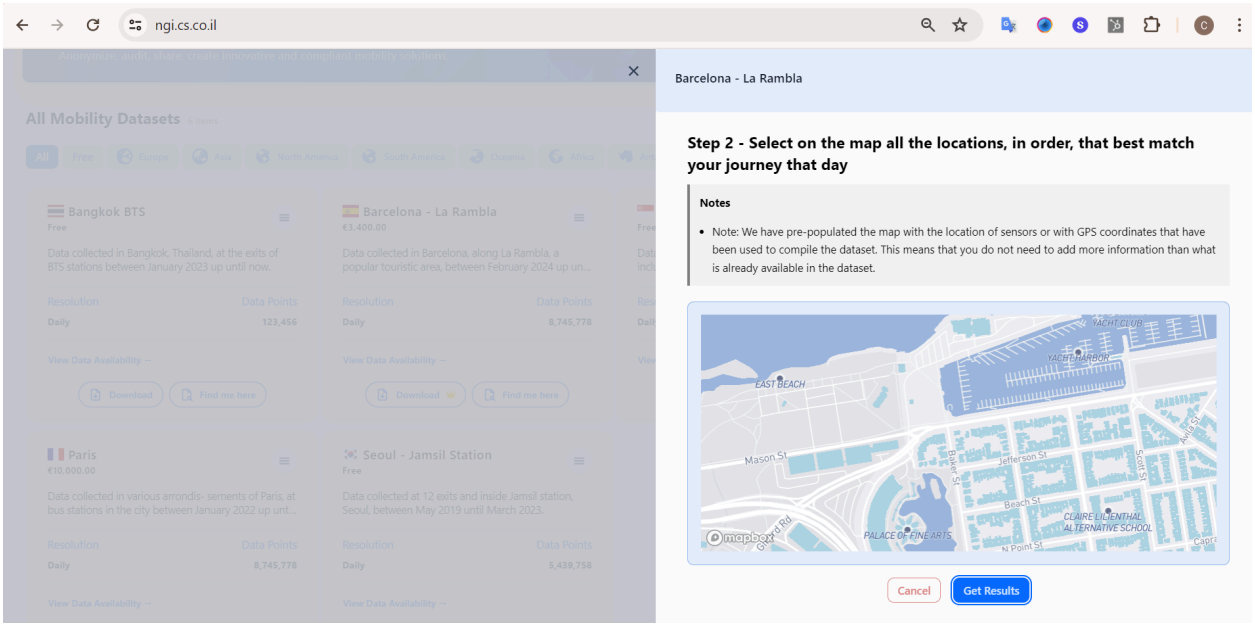

Figure 5 - Find me here - step 2 (*under development*)

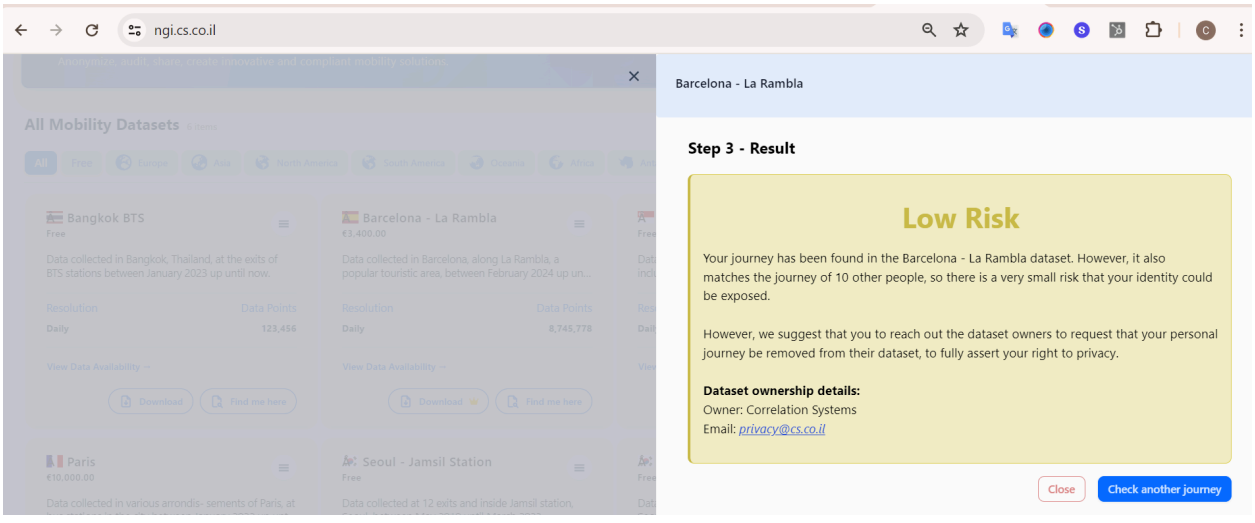

Figure 6 - Find me here - step 3 (*under development*)

2.4 Functionality 4 - Anonymization tool for companies

The UtiP-DAM anonymization tool offers a user-friendly web interface designed to ease the data anonymization process for public and private location dataset owners. This innovative tool addresses critical needs for dataset owners:

- Legal compliance: Complex privacy laws, like GDPR, mandate stringent data protection measures. Anonymizing data helps companies comply with these regulations by removing personally identifiable information (PII), reducing the risk of hefty fines and reputational damage.
- Minimizing data storage risks: Storing personal data exposes businesses to potential security breaches and data leaks. Utilizing an anonymization tool minimizes the risk of leaking personal data (hence, shielding these businesses from potential fines) by enabling the removal of sensitive information while preserving the analytical value of the data.
- Enabling innovation: Valuable insights can be obtained from exchanging data between organizations and with the general public. However, privacy concerns often hinder such collaborations. Anonymization tools enable secure data sharing, fostering innovation, research, and citizen participation in innovative projects.

UtiP-DAM develops a web interface serving multiple purposes:

- Simplified K-anonymity: K-anonymity, the chosen anonymization technique for the project, can be complex to implement. The web interface provides a user-friendly platform for companies to anonymize their data sets using centralized k-anonymity without requiring extensive programming expertise.
- Demonstrating the full solution before committing: Potential users of the anonymization tool can experience the full functionality of UtiP-DAM through the web interface, allowing them to test-drive the anonymization process and gain a clear understanding of its capabilities. If they are satisfied with the tool, they can then deploy the software code, which will be available via github, in their company.
- Effective Marketing Tool: The web interface for the anonymization tool will be directly connected to the data sharing feature (see screenshots below), enabling dataset owners to share their anonymized data in one click, making data sharing a smooth process.

Here is how it works:

1. User downloads the sample file (csv file), allowing him to understand the type of data the system expects to receive, mapping their data to expected data. The User then fills in the sample file, save and upload it on Utip-DAM

2. User selects K: this is a key step in the anonymization process, as the value chosen by the user will impact the anonymity versus utility trade-off of the anonymization process. At this step, Correlation Systems will provide some guidance as to what K represents, and offer a default value of 20, which our experience shows as being a good starting point for k-anonymity.
3. The system anonymizes the dataset using the algorithm developed by our data scientists.
4. User downloads their anonymized dataset: at this step, we will require the user to agree to the terms and conditions of using the system, prior to being allowed to download the dataset. This enables Correlation Systems, and to some extent, NGI, to be protected from potential legal actions instigated by end-users.
5. Option: user shares their dataset on the marketplace and International Data Space (more on this in part 2.5).

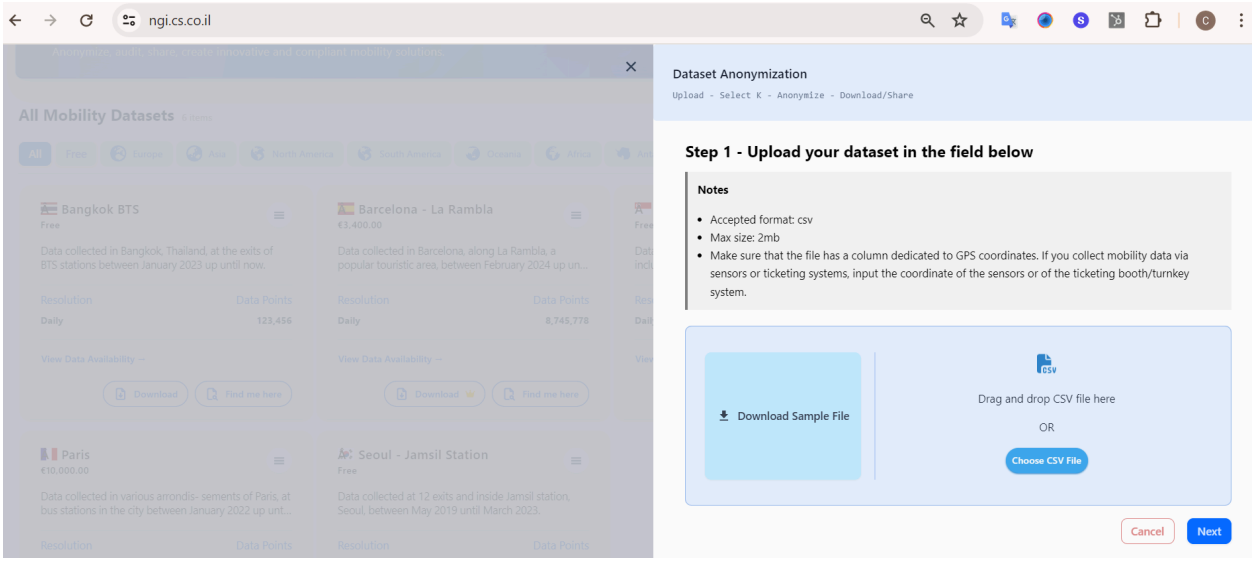

Figure 7 - Anonymization tool - step 1 (*under development*)

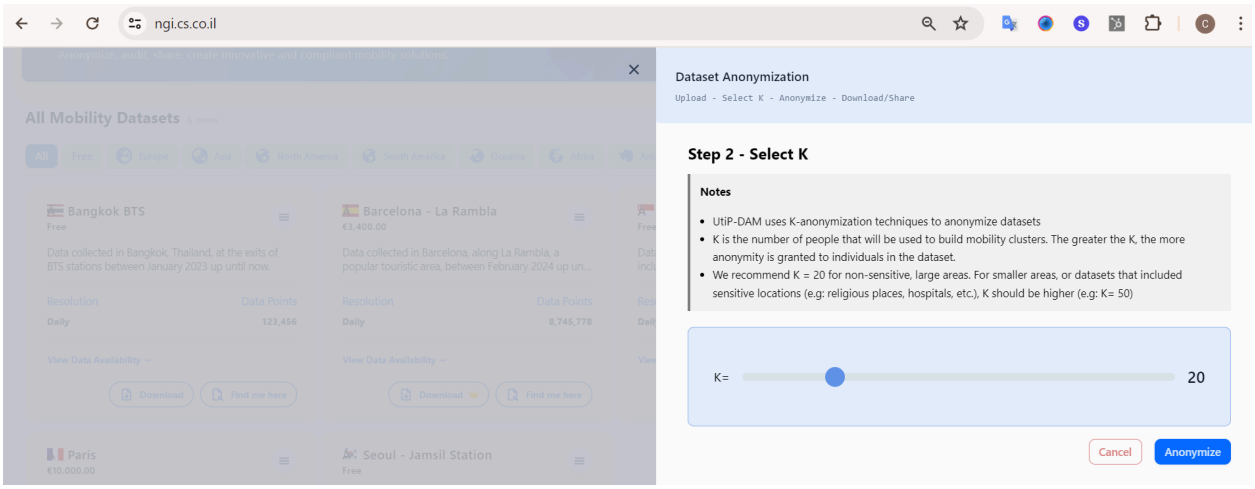

Figure 8 - Anonymization tool - step 2 (*under development*)

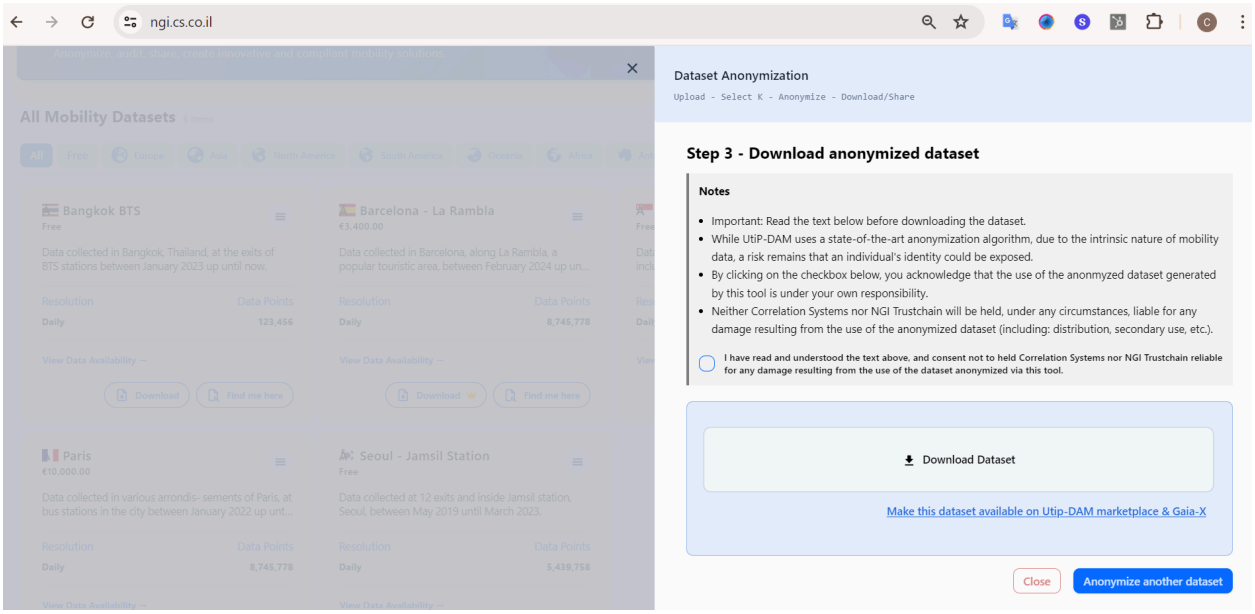

Figure 9 - Anonymization tool - step 3 (*under development*)

2.5 Functionality 5 - Anonymized data sharing

The UtiP-DAM marketplace goes beyond just providing anonymized data; it fosters a secure and responsible environment for data sharing.

A core feature of the platform is the dataset sharing functionality, specifically designed for, and accessible to, users who have anonymized their data using the UtiP-DAM anonymization tool.

By making anonymization via UtiP-DAM a prerequisite to dataset sharing on the marketplace, we aim to:

- Guarantee privacy protection: Requiring anonymization through UtiP-DAM ensures that all data shared on the marketplace adheres to the highest privacy standards. This further safeguards individual privacy and builds trust within the data ecosystem.
- Compliance with upcoming regulations: Data regulations are constantly evolving, with initiatives like the EU's Data Act emphasizing responsible data sharing practices. By mandating UtiP-DAM anonymization and making it easy to share user-generated data, the marketplace proactively aligns with these regulations, providing a future-proof solution for data owners.

The data sharing feature offers the following benefits:

- Fueling Innovation: Data sharing creates the potential for innovative research and development across diverse fields. Anonymized datasets on the UtiP-DAM marketplace can serve as valuable resources for researchers, enabling them to gain new insights and develop innovative solutions.
- Fostering collaboration: The UtiP-DAM marketplace facilitates secure mobility secondary use, enabling companies to exchange data, even working together, and with private citizens, to address world challenges.
- Maximizing data value: much valuable data often remains underutilized within individual organizations. Data sharing allows companies to access the full potential of their anonymized data sets by making them available, and even profiting from sharing them with a wider audience.

Here is how it works:

1. After anonymizing their datasets, users can choose to share their data with the public
2. If they choose to do so, they must fill in mandatory information (see figure 10 below): among other things, this ensures that users who have identified their

journeys in a dataset can reach out to the dataset owner to request the removal of this specific journey.

3. User must agree to terms and conditions related to sharing their dataset on the Utip-DAM marketplace

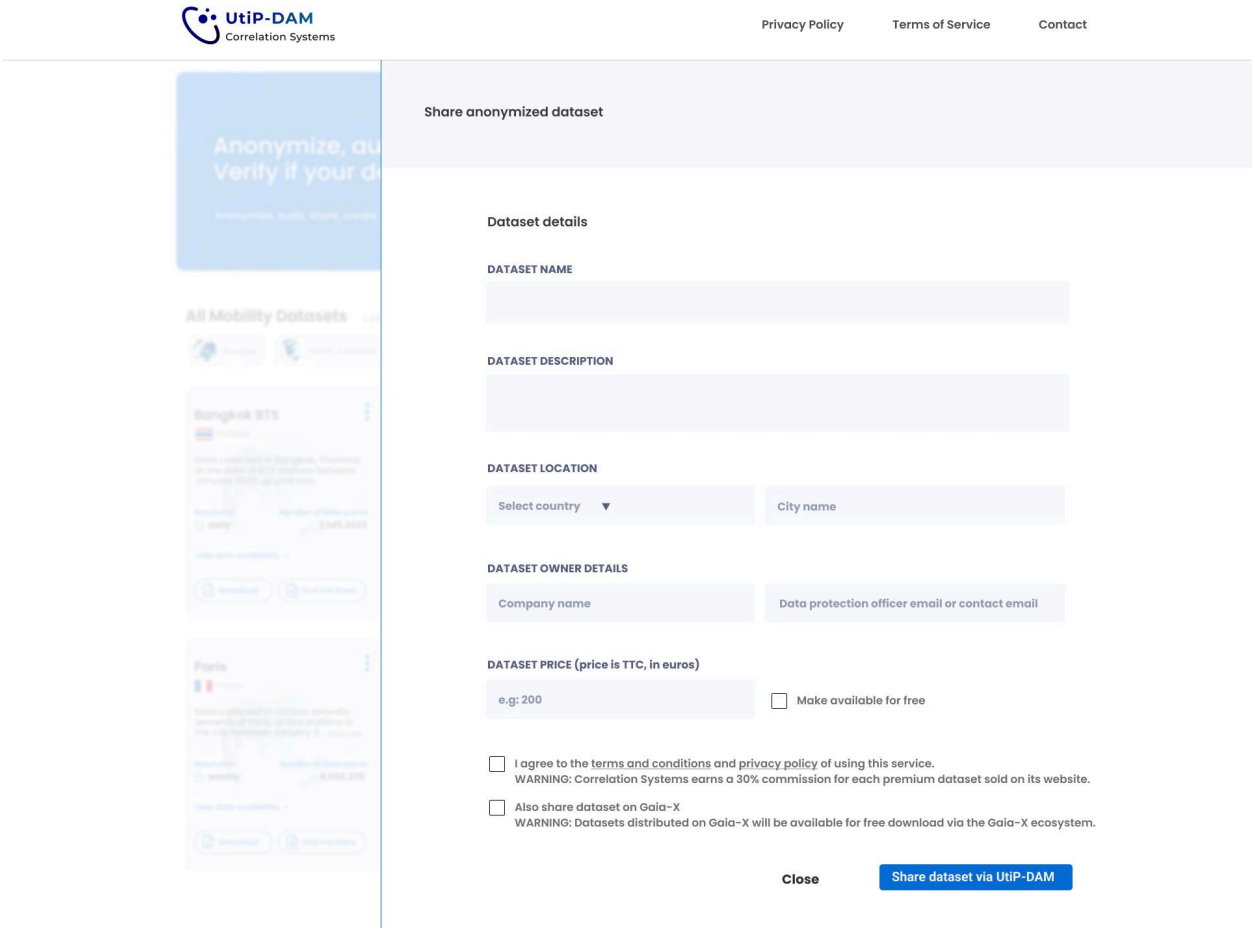

Figure 10 - Data Sharing (mockup)

2.6 Functionality 6 - User accounts

The UtiP-DAM marketplace utilizes a user account system to provide a more secure, convenient, and transparent experience for all participants. Here's how user accounts benefit both data downloaders and distributors:

- Improved User Experience: User accounts make the data download process a smoother experience. Once datasets are downloaded, users can easily track their history, manage their files, and even re-download them if needed. This eliminates the need to remember complex download links or worry about losing access to valuable data sets.
- Combating spam and misuse: User accounts deter illegitimate download requests and potential data misuse. By requiring account creation, UtiP-DAM aims to safeguard the integrity of its platform and ensure that datasets are downloaded by authorized users with genuine research or analytic purposes.
- Facilitating Data Distribution (For Distributors): For users who choose to distribute their anonymized mobility datasets through the marketplace, user accounts provide crucial functionalities:
    - Managing distributor information: Users can maintain their contact details and other relevant information within their account. This allows potential data buyers to easily reach out to dataset owners for inquiries or clarifications about the data sets, including for requesting that their personal data be deleted (see part 2.3). Dataset owners can also modify or delete this information at any time, ensuring data privacy control.
    - Revenue Tracking and Transparency: Distributors who choose to sell their anonymized datasets can gain valuable insights via their user accounts. They will be able to track download statistics, monitor their revenue generated from data sales, and gain a clear understanding of the value of using UtiP-DAM. Transparency fosters trust and incentivizes responsible data sharing within the marketplace.

Here, it bears reminding that individuals wishing to use the auditing tools (see part 2.3) will **not** be required to create user accounts.

Aligned with the GDPR's data minimization principle, UtiP-DAM only collects and stores the essential data required for operating the marketplace. This commitment extends to dataset auditing, where user data collection is not required.

Mockups for this feature are currently under development.

2.7 Functionality 7 - Dataset purchase

The UtiP-DAM marketplace offers dataset owners the flexibility to choose between free and premium distribution models. As previously mentioned, Correlation Systems' own mobility data will be made available for free under an open-source license, adhering to our original proposal to NGI Trustchain. However, enabling paid data distribution for third parties offers several important benefits:

- Incenting Data Sharing: The opportunity to earn revenue incentivizes third-party dataset owners to contribute their anonymized mobility data to the marketplace. This enriches the data ecosystem and fosters a wider variety of datasets for downloaders.
- Sustainable Marketplace Development: Maintaining the UtiP-DAM marketplace requires ongoing investment in server infrastructure, feature development, and security measures. By allowing premium data sales, from which Correlation Systems earns a small commission, we generate revenue to support these costs, ensuring the marketplace's long-term sustainability and continued value to all users.
- Reflecting Data Value: Certain anonymized mobility datasets may hold significant value for specific research or commercial purposes. The ability to charge a premium reflects the inherent value of the data and compensates data owners for the time and resources invested in data collection and anonymization.

The UtiP-DAM marketplace will use Stripe, a secure and trusted payment gateway, to facilitate seamless transactions for premium data purchases.

Here are a few reasons as to why Stripe is the ideal choice for our marketplace:

- Security: Stripe prioritizes security, adhering to stringent industry standards to safeguard financial information. This fosters trust among users conducting transactions within the marketplace.
- Ease of integration: Stripe offers a user-friendly and well-documented API, simplifying the integration process for the UtiP-DAM platform. This helps ensure a smooth payment experience for all users.
- Global reach: Stripe supports a vast array of payment methods and currencies, enabling data purchases from around the world. This broadens the marketplace's reach and user base.
- Competitive fees: Stripe's transparent and competitive commission fees minimize transaction costs for both data buyers and distributors. This ensures that a larger portion of the revenue reaches data owners.

- Marketplace functionality: Stripe is specifically designed to facilitate transactions within marketplaces. This allows Correlation Systems to efficiently manage payouts to dataset owners, ensuring they receive a fair share of the revenue generated from their data sales.

By choosing Stripe, the UtiP-DAM marketplace prioritizes security, ease of use, global accessibility, and cost-effectiveness for all users. This fosters a healthy marketplace ecosystem where data owners are incentivized to contribute and buyers can securely access valuable datasets.

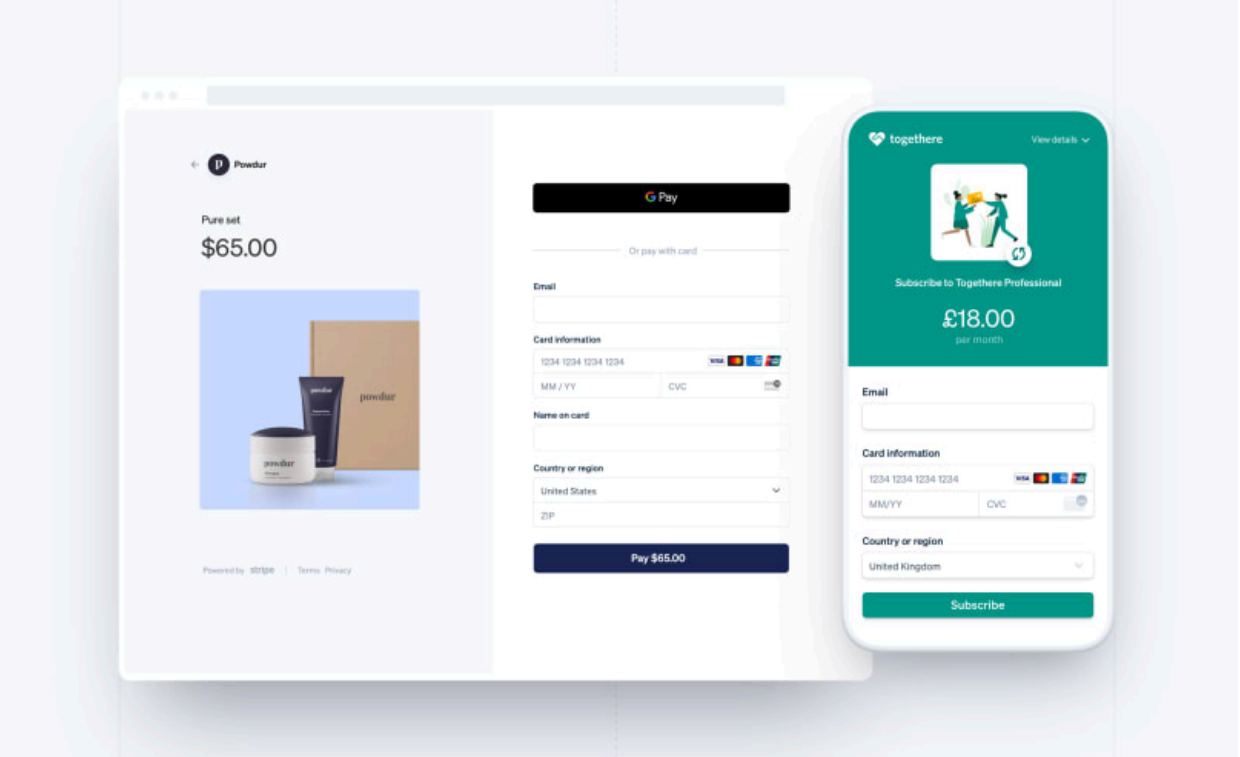

Figure 11 - Stripe Checkout (example)

2.8 Functionality 8 - API for the marketplace

In addition to the graphical user interface (GUI) offered through the UtiP-DAM marketplace, we will provide APIs (Application Programming Interfaces). These APIs empower users to interact with UtiP-DAM programmatically, allowing them to automate data retrieval from the marketplace directly within their applications.

Specifically, developers will be able to leverage the UtiP-DAM APIs to:

- Register a new dataset within the marketplace
- Add a specific day's data to an existing dataset
- Download a specific day's data from an existing dataset

Developing APIs is an important part of the UtiP-DAM project as APIs can play a critical role in fostering innovation within the UtiP-DAM ecosystem.

By offering open and well-documented APIs, we hope to enable the following:

- Streamlined data integration: Developers can easily integrate anonymized mobility data from the UtiP-DAM marketplace into their applications, accelerating development cycles and enriching their offerings.
- Enhanced user experience: Developers can create innovative applications and services that leverage the rich data available through UtiP-DAM, ultimately improving the user experience.
- Expanded marketplace reach: APIs broaden the reach of the UtiP-DAM marketplace, attracting developers from diverse backgrounds and expertise who can contribute to the creation of novel data-driven solutions.

Further, UtiP-DAM commits to making its APIs compatible with the International Data Spaces (IDS) initiative. This decision creates significant benefits for the project:

- Future-proofing the marketplace: By adhering to IDS standards, UtiP-DAM ensures seamless integration with the emerging mobility data space. This approach could position the marketplace as a key player in the data market.
- Increased data sharing opportunities: IDS compatibility facilitates data exchange with other compliant data spaces, potentially leading to a wider range of anonymized mobility data becoming available within the UtiP-DAM marketplace.
- Enhanced data security: The IDS framework emphasizes data security and privacy. Compatibility ensures the secure exchange of data within the UtiP-DAM marketplace, fostering trust among data providers and consumers.

As the European Data Space concept gains momentum, adhering to interoperability standards becomes increasingly important for open projects such as UtiP-DAM. Data relevant to tourism and mobility, which forms the core of the UtiP-DAM marketplace, will be highly valuable within this ecosystem.

By being interoperable with IDS, UtiP-DAM positions itself to play a vital role in facilitating data exchange and unlocking the potential of this data for various stakeholders.

2.9 Functionality 9 - Distributed EDGE network

The UtiP-DAM project takes a significant step forward in privacy-preserving data collection by introducing k-anonymity functionality for Edge IoT devices.

A key innovation within this functionality is the key management system. It enables Edge devices to exchange anonymized data securely without exposing local data to remote nodes. This ensures that data privacy is maintained even in use-case of an IoT network operated by a central data controller.

This innovation offers several key advantages:

- Decentralized data processing: K-anonymity on Edge devices eliminates the need to transmit raw data to a central data controller. This minimizes privacy risks and reduces reliance on centralized databases that may house vast amounts of personal information.
- Enhanced data security: By processing data directly on Edge devices, the k-anonymity functionality minimizes the attack surface for potential data breaches. This fosters a more secure data processing environment.
- Reduced network traffic and costs: Processing data locally on Edge devices reduces network traffic and associated costs. This is particularly beneficial for large-scale deployments where data transmission can be a significant expense.

Correlation Systems is committed to promoting secure and privacy-conscious data practices within the Edge IoT landscape. We will be implementing this k-anonymity functionality on our own Edge devices, with the code readily available on GitHub and distributed under open-source license. This allows other IoT system integrators to leverage and integrate this functionality within their own deployments, fostering broader adoption of privacy-preserving Edge processing solutions.

While we currently don't foresee direct integration of this k-anonymity functionality within the UtiP-DAM marketplace itself, the open-source code paves the way for third-party exploitation of the project results.

2.10 Functionality 10 - Add download raw data function and Share data function to the existing dashboard

The UtiP-DAM project offers two valuable new features for Correlation Systems' existing customers: the ability to download anonymized raw data directly from their current dashboard and share their data to UtiP-DAM and the IDS.

These functionalities unlocks several key benefits:

- Enhanced customer satisfaction: Existing customers rely on Correlation Systems' mobility module for visitor data insights. By offering in-dashboard downloads of anonymized raw data, we empower them to conduct deeper, more granular analyses using their preferred tools. This caters to customers who require a level of detail beyond aggregated data.
- Additional revenue streams: With the shared raw data via UtiP-DAM, our customers will be able to generate new revenue streams from the distribution of their data via the marketplace. This will enable Correlation Systems to move away from its position as a costly service provider, and towards a revenue-generating partner.
- Improved brand visibility: By leveraging UtiP-DAM and International Data Space for data sharing, our customers can enhance their online brand presence. This increased visibility can potentially lead to higher brand value and attract new customers.
- Improved service offering: These features strengthen Correlation Systems' service offering and differentiate us from competitors.
- Attracting new customers: The ability to download and share anonymized raw data increases our appeal to potential clients seeking in-depth data analysis capabilities without compromising privacy. This can become a significant advantage in attracting new business.

One key motivation for this functionality is to address the needs of customers to use their crowd data to the fullest; who may not be satisfied with aggregated mobility data (Correlation Systems' current display of mobility data) and wish to conduct more in-depth analysis of the trajectories of individuals on their sites.

This in-dashboard solution provides them with a secure and compliant way to access anonymized raw data for internal analysis, meeting their business needs while adhering to privacy regulations.

This feature will leverage the existing dashboard environment familiar to our customers. They can seamlessly access and download k-anonymized raw data using a "download CSV function," reducing the need to learn new platforms or procedures. This fosters a smoother user experience.

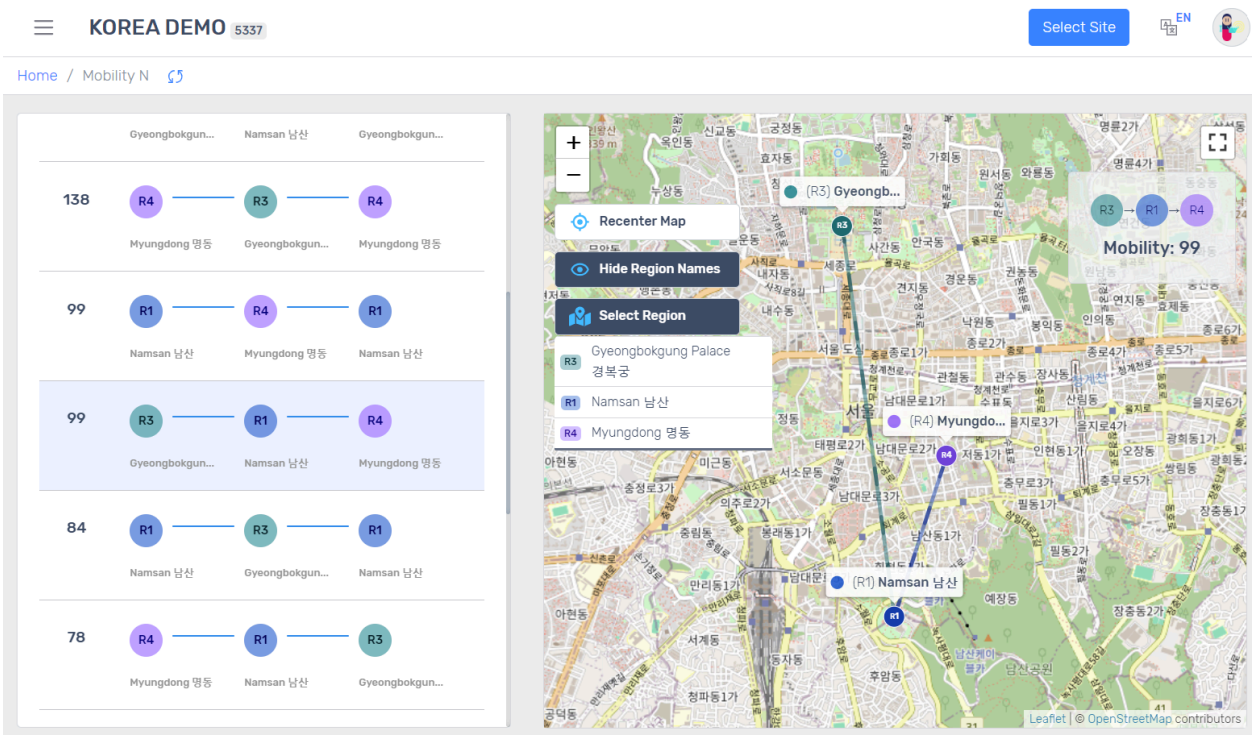

Figure 12 - Current view for the Mobility module (will be modified to include the Download CSV and Share Data Button)

3 USER NEEDS ASSESSMENT

3.1 RESEARCH ON USER NEEDS

The UtiP-DAM project prioritizes user centricity.

Our core user group comprises Correlation Systems' existing customers who utilize our sensor systems and generate mobility data. The first step involved identifying the most relevant customer segment. Customers with limited deployments or purely interested in basic visitor analytics were excluded. This resulted in a focused group of 21 customers actively using the mobility feature.

To gain deeper insights into customer needs, we used a two-pronged approach:

1. Interviews with internal stakeholders: We engaged with the internal customer success staff (typically, an external distributor, who acts as the contact point between Correlation Systems and the customer) responsible for each of the identified 21 mobility-using customers. These interviews explored:
    - Customer Use Cases: We investigated how customers are currently leveraging mobility data and their specific needs.
    - Raw Data Interest: We assessed customer interest in accessing raw mobility data and their internal/external data processing capabilities.
    - Data Sharing Policies: We explored customer policies regarding data sharing and their openness to sharing anonymized data publicly.
    - Prototype Feedback: We showcased the UtiP-DAM prototype to gauge its perceived usefulness and potential feature additions.
2. In-depth customer interviews: We conducted one-hour online interviews with three key customers who will closely collaborate with Correlation Systems throughout the project:
    - La Rambla Tourist Association: This customer's anonymized dataset will be available for public download via the UtiP-DAM marketplace. The interview explored their data sharing goals in the context of the data sharing agreement established with Correlation Systems (see Appendix A of D1). Appendix B details the in-depth findings of this interview.
    - MIE, System Integrator SME (Bangkok Public Transport Network): MIE operates 12 sensors in the Bangkok network and will contribute their anonymized data to the marketplace. The interview focused on their data sharing experiences and expectations. With MIE, we also reviewed their goals as potential dataset buyers, as they mentioned their interest in purchasing external datasets generated by public transport operators to compare network performance and user experience with theirs. Appendix C provides an in-depth analysis of this interview.
    - BatYam mall (Israel): This customer operates 12 sensors, deployed in key areas of a large mall. During the interview, we detailed the benefits of the anonymization algorithm, as Israel's data privacy laws are even more stringent than the GDPR; and received positive feedback on the possibility to sell their data on the UtiP-DAM marketplace. Appendix D provides greater details on this interview.

Recognizing the importance of researchers as both data users and providers for the UtiP-DAM marketplace and open-source tools, we conducted separate one-hour online interview with European researchers from Bruno University (Czech Republic), known for their expertise in IoT, and from the Luxembourg Institute of Science and Technology, known for their expertise in network security, with whom Correlation Systems has a long experience of collaboration on innovative projects.

These discussions explored:

- Project concept review: We presented the UtiP-DAM concept, highlighting its benefits (increased data availability, innovative IoT solution) and potential challenges due to technical complexities.
- Researcher perspective: Gaining feedback from a research perspective proved valuable in shaping the platform's functionality for both data access and contribution.

Key Findings:

- Data privacy is paramount: Unsurprisingly, data privacy emerged as a primary concern for all customers. Fear of negative publicity that could occur from divulging private information, even when due diligence is done, is a significant deterrent to data sharing.
- Internal data utilization: Most customers manage "big data" projects, but the results are often used internally. Data sharing is limited, primarily due to the aforementioned privacy concerns.
- Open data initiatives limited: While some customers participate in open data initiatives, these typically involve processed data. Raw data sharing must be contingent on robust privacy guarantees.
- Commercialization potential: Semi-government and PPP projects showed a deep interest in data commercialization. However, data privacy and public sentiment remain crucial considerations before launching such initiatives.

The detailed analysis of these use cases is presented in the next sections. However, these initial findings underscore the importance of building a platform that prioritizes data privacy while facilitating secure and valuable data exchange for both researchers and businesses.

3.2 User Stories and the UtiP-DAM Functionalities

We consider the following users:

- Correlation Systems' Customers: Existing customers who utilize Correlation Systems' sensor technology and generate mobility data via this infrastructure.
- Third-Party Dataset Owners: This group includes individuals, public entities, and private organizations (not Correlation Systems' customers) who collect, process and/or own mobility data sets.

- Data Privacy Officers (DPOs): These individuals are responsible for ensuring data compliance with privacy regulations. They may be staff of third-party dataset owners, or act as both data owner and DPO for smaller organizations.
- Data Consumers: This broad category includes anyone who utilizes the anonymized data available through the UtiP-DAM marketplace or shared with the International Data Space. This could encompass researchers, businesses, or any entity seeking insights from mobility data.
- Citizens: This group of individuals include anyone who is interested in finding out if their personal data is included in a public mobility dataset shared on the UtiP-DAM marketplace.
- Developers: This group includes individuals, employees from public or private organizations, as well as researchers, who will utilize the UtiP-DAM APIs and open-source code to integrate the project's functionalities within their own infrastructure or project.
- Researchers: A specific sub-group of data owners, consumers, and developers. Researchers, typically affiliated with universities or research institutions, may have distinct data usage goals compared to non-researchers; specifically: test and validate the decentralized K-anonymity algorithm and UtiP-DAM software code developed by Correlation Systems; create innovative research projects.

Now, let’s map these users and their needs to UtiP-DAM functionalities.

Functionality 1: Core auditing and K-Anonymization algorithm

- As a DPO, I want to be able to audit my proprietary location/mobility dataset so that I can ensure that it does not contain any risk for re-identification.
- As a researcher, I want to be able to test the K-anonymity algorithm developed by Correlation Systems, so that I can validate it (or not).
- As a researcher, I want to be able to iterate upon the K-anonymity algorithm developed by Correlation Systems so that I can generate new insights..
- As a developer, I want to be able to integrate the K-anonymity algorithm into my company or my project’s mobility application so that I can ensure it does not generate personal information.
- As a data consumer, I want to be able to audit mobility datasets that I have acquired through the years, so that I can verify that they do not contain private data, before using it for my own purpose.

Functionality 2: Raw Mobility Data Marketplace

- As a data consumer, I want to be able to search and browse the UtiP-DAM marketplace for anonymized mobility datasets, so that I can find datasets relevant to my research or analysis needs.

- As a data owner, I want to be able to upload and publish my anonymized mobility dataset on the UtiP-DAM marketplace, so that I can share my data with others and potentially generate revenue.
- As a Correlation Systems' customer, I want to be able to share my mobility dataset on the UtiP-DAM marketplace, so that I can potentially generate revenue and increase brand awareness.
- As a researcher, I want to be able to distribute my anonymized dataset openly on an online marketplace, so that I can align with my funded research's open distribution requirements and/or enable other researchers to use my data for validation purposes.

Functionality 3: Anonymization Auditing Tool

- As a citizen, I want to be able to use the UtiP-DAM anonymization auditing tool, so that I can verify that none of the datasets distributed on the internet, via the UtiP-DAM marketplace, includes my personal data.

Functionality 4: Anonymization Tool for Companies

- As a DPO, I want to be able to anonymize mobility datasets generated by my company so that I can ensure that data stored on my company's servers does not contain personal information.
- As a third-party dataset owner, I want to be able to anonymize my mobility datasets so that I can ensure that I do not store personal data.
- As a researcher, I want to be able to anonymize the mobility data generated by my research project prior to distributing it or making it accessible to other researchers, so that I can ensure that it fulfills privacy laws requirements.

Functionality 5: Anonymized Data Sharing

- As a researcher, I want to be able to share my research data on the UtiP-DAM marketplace and/or the International Data Space so that I can enable other researchers and data consumers to utilize the data for their own purposes.
- As a third-party dataset owner, I want to be able to share my mobility datasets so that I may create new revenue streams and/or increase brand visibility.
- As a DPO, I want to be able to share anonymized mobility data so that I can comply with upcoming regulations such as the Data Act, which will make it mandatory for me to provide access to my users with the data they generate.
- As a Correlation Systems' customer, I want to be able to share my mobility data so that I may create new revenue streams and/or increase brand visibility.

Functionality 6: User Accounts

- As a data consumer, I want to be able to create a user account, so that I can manage my downloads, uploads, and other activities within the marketplace.
- As a third-party dataset owner (who distributes my data on the UtiP-DAM marketplace), I want to be able to create a user account, so that I can track dataset downloads, as well as payouts, revenues, and manage my information.
- As a researcher, I want to be able to create a user account, so that I can track dataset downloads and manage my information.

Functionality 7: Dataset Purchase

- As a data consumer, I want to be able to securely purchase premium datasets from the UtiP-DAM marketplace using a variety of payment methods, so that I can access valuable data for my specific needs.

Functionality 8: API for the Marketplace

- As a developer, I want to be able to integrate with the UtiP-DAM marketplace through an API, so that I can automate data download or upload processes within my own applications.

Functionality 9: Distributed Edge network

- As a Correlation Systems' customer, I want to be able to deploy the distributed Edge network algorithm on my Correlation Systems IoT infrastructure so that I can ensure that Correlation Systems does not have access to raw mobility data generated by my visitors.
- As a DPO, I want to be able to deploy the distributed Edge network algorithm on my company's IoT infrastructure, so that I can ensure that even I do not have access to raw mobility data, making it easier for me to comply with privacy laws.
- As a researcher who uses IoT networks in their research, I want to be able to deploy the distributed Edge network algorithm on my IoT infrastructure, to ensure that even I do not have access to raw mobility data, potentially helping me increase trust from my research participants.

Functionality 10: Add Download Raw Data and Share data Functions to Dashboard

- As a Correlation Systems customer who already uses the mobility module, I want to be able to download anonymized raw data directly from my existing dashboard, so that I can leverage the anonymization capabilities of UtiP-DAM without needing to switch platforms.
- As a Correlation Systems customer, I want to be able to share my data on the UtiP-DAM marketplace and/or IDS, so that I can distribute my data in a click, potentially adding new revenue streams and increasing brand awareness.

4 STATE OF THE ART ANALYSIS, BACKGROUND, AND INNOVATION

4.1 State of the Art Analysis

4.1.1 Anonymization techniques

Our era is defined by data-driven decision-making, but also by increasingly stringent privacy regulations. This necessitates a balance between safeguarding individuals' privacy on the one hand and preserving the usefulness of the data on the other hand.

This challenge is particularly relevant for entities that collect, own, and distribute personal datasets.

In this context, data anonymization emerges as a critical tool, offering a variety of techniques that obfuscate sensitive details within datasets while attempting to preserve their analytical value: the data anonymization umbrella encompasses a variety of techniques designed to obfuscate sensitive details within datasets, thereby severing the connection between data points and identifiable individuals.

Below is a comprehensive exploration of prominent anonymization techniques, emphasizing their efficacy for location and mobility datasets:

4.1.1.1 Data Pseudonymization

Data pseudonymization is usually considered as a preliminary step towards anonymization, focusing on replacing directly identifiable information with fictitious identifiers, often referred to as pseudonyms (Ohm, 2010). Replacing direct identifiers makes it more challenging to link anonymized data points back to specific individuals, thereby mitigating the risk of re-identification.

Pseudonymization is often a less computationally intensive process compared to other anonymization techniques, making it potentially easier, faster and more cost-effective to implement.

In certain scenarios, pseudonymization also allows for the preservation of relationships between data points, which can be beneficial for specific use cases, such as mobility use-cases. In this way, pseudonymization is a great way to preserve the utility of datasets.

An example of pseudonymization consists in substituting email addresses with numerical codes or replacing full names with initials.

While pseudonymization offers some degree of obfuscation, it has some noteworthy weaknesses:

- Reversibility: Unlike true anonymization, pseudonymization is often reversible. With the appropriate key or additional information, it's possible to potentially link pseudonyms back to their original identifiers, especially if the pseudonymization method is not sufficiently robust.
- Limited protection against linkage attacks: Pseudonymization primarily focuses on direct identifiers. It doesn't address the potential presence of indirect identifiers within the data. These indirect identifiers, such as zip codes combined with birthdates, could still be exploited by attackers to re-identify individuals, particularly when combined with information from other sources (Ohm, 2010).
- Limited utility for highly sensitive data: For this sort of data, pseudonymization might not provide a sufficient level of privacy protection. In such cases, stronger anonymization techniques are necessary.

4.1.1.2 Data Masking

Data masking encompasses a diverse set of techniques designed to obscure or alter data values within a dataset, hindering the reconstruction of original values and safeguarding individual privacy (Ohm, 2010). This approach proves particularly valuable when working with sensitive datasets containing personally identifiable information (PII).

- This static approach involves the creation of a separate, anonymized copy of the original dataset. This technique offers a straightforward and efficient means of anonymization, but it's crucial to select appropriate masking methods that maintain data utility for the intended purpose. Common methods include:
    - Substitution: Replacing sensitive attributes with fictitious values. Examples include replacing zip codes with broader city names, birthdates with age ranges, or full names with initials (Ohm, 2010).
    - Redaction: Completely removing sensitive data elements from the dataset. This approach is suitable for highly sensitive information that is not crucial for analysis.

    - Shuffling: Rearranging data points within a dataset to disrupt the original order and connection between specific data points (Domingo-Ferrer & Torra, 2005).
    - Aggregation: Grouping specific values into broader categories. For instance, converting timestamps into time windows or income ranges into broader salary brackets (Domingo-Ferrer & Torra, 2005).

- Dynamic Masking: In contrast to static masking, dynamic masking alters data "on the fly" during queries or transfers. Imagine a continuously morphing disguise, where the level of anonymization can be adjusted based on the specific query or user (Li et al., 2007). This approach offers enhanced protection against specific attack vectors, particularly those that exploit patterns within the data. Here are some examples of dynamic masking techniques:
    - Tokenization: Sensitive data is replaced with a non-reversible token, a random string of characters that holds no meaning on its own. The original value can only be retrieved if the corresponding tokenization key is available.
    - Data Redaction: Specific data elements are completely removed from the dataset based on pre-defined rules or user permissions (El Emam et al., 2008).
    - Format-preserving encryption: Sensitive data is encrypted while preserving its original format, allowing for some basic operations to be performed on the encrypted data without decryption (Agrawal et al., 2002).

The selection of the most appropriate data masking technique depends on various factors, including the sensitivity of the data, the level of anonymity required, and the intended use of the anonymized dataset.

K-anonymity, encryption, and differential privacy are all frequently employed masking techniques for location data. We will elaborate on these techniques in subsequent sections.

4.1.1.3 K-Anonymity

K-anonymity is a data anonymization technique that aims to anonymize individual data points by ensuring they are indistinguishable from at least k-1 other data points within a specific dataset (Sweeney et al., 2002). It follows the principle of "anonymity

by obscurity," essentially hiding an individual's data amongst a group of similar individuals.

For example, we may consider a dataset containing location data with attributes like zip code and age and gender. K-anonymity might anonymize this data by replacing zip codes with broader city regions, age with age ranges. This ensures that any record for a specific individual cannot be distinguished from the records of at least k-1 other individuals within the same city, age group and gender.

Various factors affect the strength of K-Anonymity-based anonymization:

- K Value: The higher the k value, the stronger the anonymization. With a larger k, it becomes increasingly difficult to pinpoint a specific individual's data. However, increasing k can also lead to significant data utility loss due to the broader generalization required.
- Identifying Attributes: The selection of identifying attributes significantly impacts k-anonymity's effectiveness. Including more granular attributes like zip code, age, and income level creates more robust anonymity guarantees. However, it also reduces the number of records that can potentially fall into the same k-anonymous group.

Consequently, K-anonymity can offer a relatively straightforward approach to data anonymization, making it easier to understand and implement.

By carefully choosing k and identifying attributes, k-anonymity allows dataset owners to achieve a reasonable level of privacy protection without sacrificing excessive data utility.

There are nonetheless, two main disadvantages of K-Anonymity:

- Vulnerability to attackers with additional information: If attackers possess supplementary information about individuals, such as data from social media profiles, they might be able to re-identify individuals even within a k-anonymous dataset (Machanavajjhala et al., 2007).
- Data utility loss: Achieving strong k-anonymity often necessitates significant data generalization, which can lead to a loss of data granularity and accuracy.

4.1.1.4 Differential Privacy

Differential privacy is another robust data anonymization technique that prioritizes the privacy of individuals within a dataset, even when revealing statistical insights generated from processing the data.

Unlike traditional anonymization methods that focus on obscuring the raw data itself, differential privacy injects noise into the outcomes of statistical queries executed on the data (Dwork et al., 2014). This approach ensures that even an observer with some knowledge over the raw and anonymized results cannot determine the exact contribution of any single individual to the query's outcome.

For example, consider a statistician analyzing the average income of a city's population. Differential privacy adds a controlled amount of statistical noise to the calculated average. While the overall trend (average income) remains discernible, it becomes impossible to pinpoint the exact salary of any specific individual.

The core strength of differential privacy lies in its demonstrably strong privacy guarantees. It quantifies the level of risk associated with re-identification, mathematically formalizing the protection offered to individuals within the dataset.

However, the high level of privacy offered by differential privacy comes at a cost. The injected noise may introduce a degree of inaccuracy into the results of statistical queries. Consequently, utilizing differential privacy is a balancing act between ensuring a robust level of individual privacy protection while maintaining the utility and accuracy of the anonymized data for analysis.

4.1.1.5 Data Generalization

Data generalization is a data anonymization technique that focuses on obscuring individual characteristics within a dataset by coarsening the granularity of the data (Domingo-Ferrer & Torra, 2005). This approach essentially broadens the view of the data, making it difficult to distinguish between specific data points. Imagine looking at a detailed map versus a simplified one – the general trends remain visible, but the finer details are blurred.

We can distinguish between two main types of data generalization:

- Automated Generalization: This approach leverages algorithms to automatically determine the optimal level of data coarsening. These algorithms typically strive to strike a balance between maximizing privacy protection and minimizing the loss of data utility for analysis (Verykios et al., 2004).
- Declarative Generalization: In contrast, declarative generalization involves manually defining the specific level of distortion for each attribute within the dataset. This approach offers greater control over the anonymization process but requires a deeper understanding of the data and the intended use case.

Data generalization is a widely used technique for data anonymization, but it requires carefully considering the level of coarsening and the potential impact on data utility to achieve a reasonable balance between privacy protection and data usability.

4.1.1.6 Data Perturbation

Data perturbation is a data anonymization technique that injects meticulously calculated randomness into specific data points within a dataset (Lipton et al., 2002).

Data perturbation achieves its anonymization goals by introducing controlled noise into sensitive data elements. There are two primary ways this is achieved:

- Numeric Perturbation: This approach focuses on numerical data attributes, such as age, income, or timestamps. It injects carefully calibrated noise into these values, either by adding or subtracting a small random value within a predefined range. This ensures that the overall trends and statistical properties of the data remain largely intact, while making it significantly more challenging to re-identify specific individuals.
- Categorical Perturbation: For categorical variables, such as zip code or occupation, data perturbation might involve randomly swapping category labels within certain constraints. This disrupts the linkage between specific data points and their original categories, offering a layer of privacy protection.

Data perturbation has the great advantage of introducing minimal distortion to the data, ensuring that the anonymized dataset remains suitable for most analytical purposes. Further, data perturbation techniques are relatively fast and computationally inexpensive, making them suitable for large datasets.

Like any other anonymization technique, they do have a few drawbacks. Namely:

- Potential for Bias: Depending on the specific implementation, data perturbation may introduce a slight bias into the anonymized data. Careful calibration is crucial to minimize this risk.
- Limited Protection Against Certain Attacks: While effective against basic re-identification attempts, data perturbation might be vulnerable to more

sophisticated attacks, especially when combined with information from other sources.

4.1.1.7 Data Swapping

Data swapping, also known as data shuffling or permutation, is a data anonymization technique that focuses on rearranging the order or linkage between attribute values within a dataset (Drineas et al., 2006).

Data swapping operates by randomly reordering specific data values within a dataset. This disrupts the original relationship between these values and the corresponding data points. There are two primary ways data swapping is implemented:

- Swapping within Attribute Values: This approach involves randomly shuffling the values within a single attribute column. For instance, swapping zip codes or birth dates within a dataset disrupts the linkage between these specific values and the individuals they originally belonged to.
- Swapping Between Data Points: Here, entire rows or records within the dataset are swapped, further disrupting the original association between individual data points and their attribute values.

Consequently, unlike some anonymization techniques that alter the data itself, data swapping primarily affects the order and linkage of data points, minimizing the risk of significant data distortion.

However, while data swapping offers some degree of anonymization by disrupting data linkage, it doesn't guarantee complete anonymity, particularly for smaller datasets or in combination with other information.

Further, for mobility datasets, which are built upon sequential movements (trajectories), swapping data points can generate nonsensical and potentially physically impossible movement patterns.

4.1.1.8 Synthetic Data

Synthetic data has emerged as a game-changer for artificial intelligence (AI) and machine learning (ML) applications.

This type of data, created by algorithms, closely resembles real-world data but is entirely artificial. This characteristic makes it an invaluable tool for training and

validating machine learning models without relying on vast quantities of sensitive personal information (King et al., 2020).

In fact, the potential of synthetic data is widely recognized, and its adoption is rapidly expanding. Predictions from Gartner Research suggest that in 2024, synthetic data will constitute a staggering 60% of the data used in AI development and analytics projects (Gartner, 2023).

Despite its advantages, synthetic data also presents some challenges. Here are a few key considerations:

- Data Realism: Generating synthetic data that accurately reflects the complexities and nuances of real-world data remains an ongoing challenge. Sophisticated algorithms are necessary to ensure that synthetic data effectively captures the underlying statistical properties and relationships present in real data.
- Domain Expertise: Creating realistic synthetic data often requires a deep understanding of the specific domain and the intended use case. Collaboration between data scientists and domain experts is crucial for successful synthetic data generation.
- Security Concerns: While synthetic data protects real-world identities, it's still essential to ensure the security of the underlying algorithms and generation processes to prevent potential vulnerabilities.

Unlike projects that rely on synthetic data, the UtiP-DAM project focuses on enabling the release of raw mobility data while guaranteeing both anonymity and utility.

4.1.2 The Evolving Landscape of Linking and Delinking ("Profiling") in Data Anonymization

While the anonymization techniques described above aim to protect individual identities, recent advancements have revealed the potential for "re-identification" in “anonymized” datasets through linking attacks. This highlights the need for robust de-linking strategies to balance data utility with privacy concerns.

What do we mean by “linking”?

Linking methods combine anonymized datasets from various sources to potentially identify individuals. For example, combining anonymized location data with social

media accounts can create a detailed profile for a specific datapoint, potentially revealing someone's identity. Research explores several key linking techniques:

- Machine Learning: Studies by Fredriksen et al. (2014) demonstrate the ability of supervised and unsupervised learning models to uncover hidden patterns and correlations within anonymized data, potentially revealing identifiers.
- Network Analysis: Narayanan and Shmatikov (2009) investigate how graph-based methods can analyze relationships between data points to uncover social connections and behavior patterns that could lead to individual identification.
- Homogeneity Attacks: Backes et al. (2008) explore exploiting the lack of diversity within specific groups in anonymized datasets. By isolating unique characteristics, attackers can potentially identify individuals.

Delinking Strategies for Enhanced Privacy

Delinking techniques aim to disconnect an individual's identity from potentially identifying data while preserving its usefulness for analysis. Here are some key de-linking approaches:

- Differential Privacy: Dwork et al. (2014) propose introducing controlled noise into data analysis. This adds a layer of obfuscation to individual details while maintaining usability for statistical purposes.
- Federated Learning: Kairouz et al. (2019) discuss distributing computations across multiple data holders. This makes it difficult for any single entity to access complete datasets and hinders linkage attempts.
- Homomorphic Encryption: Gentry (2009) explores enabling computations on encrypted data. This protects identifiers while allowing analysis.

Striking a balance between data analysis needs and individual privacy concerns is key for responsible data usage. Ohm (2010) emphasizes the importance of transparency and accountability throughout the data lifecycle, in other words: even after an anonymized dataset has been published.

4.1.3 Comparing anonymization methods and their impact on linking attacks

Method	Impact on Mobility	Impact on Linking attempts
Data Masking	Low impact: data anonymized but location data may still be identifiable.	Moderate impact: noise may disrupt linking attempts, but additional information can often reveal individuals.
K-Anonymity	Moderate impact: location may be generalized (e.g., city instead of street), reducing precision.	Low impact: profiles can still be built based on other attributes if K is not chosen properly.
Differential Privacy	High impact: location data obfuscated with noise, reducing accuracy for both mobility analysis and profiling.	High impact: noise makes profiling highly inaccurate, but information from other sources could still be used.
Data Pseudonymization	Limited impact: location data remains identifiable by the pseudonymization key.	No impact: pseudonyms don't anonymize, profiles can be built using the linked data.
Data Generalization	Moderate impact: location generalized (e.g., regional area), reducing precision for mobility analysis.	Moderate impact: profiles can still be built based on other attributes with generalization.
Data Perturbation	Moderate impact: location data slightly randomized, affecting mobility analysis accuracy.	Moderate impact: noise disrupts profiling but additional information can be used to overcome it.
Data Swapping	Low impact: location data remains largely unchanged.	No impact: data swapping doesn't anonymize, profiles can be built using the original data.
Synthetic Data	High impact: simulated location data used, preserving privacy but potentially lacking real-world accuracy.	High impact: profiles built with synthetic data are not real and therefore, cannot reveal identities.

Table 1 - comparison of anonymization methods

4.1.4 Data anonymization in practical applications

A variety of software projects offer practical implementations of anonymization techniques. AnonyMiser, for example, utilizes k-anonymity to anonymize online user activity. However, limitations such as scalability or a lack of emphasis on decentralization, a core innovation of UtiP-DAM, still exist within these tools.

The competitive landscape extends beyond mere anonymization techniques. Industry leaders, such as Google, Apple and IBM, are exploring a broader range of privacy-preserving approaches. Their investments in differential privacy and federated learning aim to protect user privacy while enabling collaborative data analysis. Initiatives like Google's Federated Learning provide innovative methods for training machine learning models on decentralized datasets, further contributing to privacy-preserving data analysis.

Blockchain also plays a significant role in the privacy-preserving landscape. Projects like Ocean Protocol and Oasis Labs offer secure data exchange and analysis on decentralized platforms. While not directly providing anonymization solutions, these platforms pose potential competition in facilitating the exchange and analysis of anonymized mobility data.

Differentiations of UtiP-DAM

Despite the existence of various privacy-preserving techniques offered by open initiatives and market competitors, UtiP-DAM's focus on decentralized k-anonymity offers distinct advantages:

- Enhanced user control: End-users can retain greater privacy over their data as anonymization occurs at the point of collection on their devices, meaning that the data controller does not need to centralize raw data before anonymization.
- Transparency and auditability: UtiP-DAM's open-source nature fosters transparency in the anonymization process and allows for easier auditing compared to potentially opaque solutions offered by some competitors.
- Focus on k-Anonymity: k-anonymity guarantees that a data point cannot be uniquely re-identified by linking it with a group of at least k-1 similar data points. This offers a strong privacy guarantee compared to some differential privacy approaches that introduce statistical noise, potentially impacting data accuracy.

The competitive landscape constantly evolves. It is important to acknowledge that these companies might develop their own decentralized solutions or collaborate with blockchain platforms in the future. Maintaining a close watch on industry trends and ongoing research will be crucial for UtiP-DAM to secure and maintain its competitive edge.

4.2 Correlation Systems' Background

Correlation Systems is a manufacturer of IoT sensors that track anonymized mobile devices data on a city wide scale.

Those sensors are used for multiple purposes such as outdoor people counters, crowd management, monitoring the effectiveness of DOOH and OOH advertising and more.

4.2.1 Top Level System Architecture

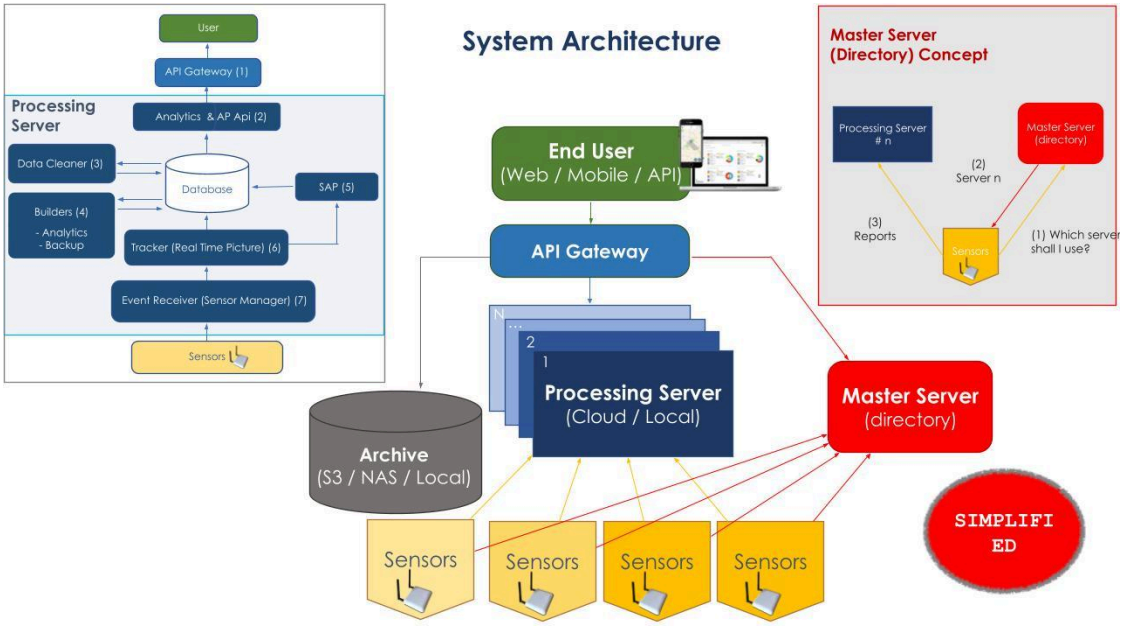

Figure 13 - graph of LBASense architecture

How does the system work?

1. When a sensor starts - it will connect to the master server (a management server) which will inform the sensor where to find its production server.
2. The sensor will detect data (WiFi headers) using a software defined radio (SDR), encrypt the MAC address in the server and report the data to the local production server (according to the allocation received before from the master server)

3. The processing server will associate between data received from multiple sensors, perform a process called “de-randomization” which is clustering of multiple messages generated by the same phone and store the processing results in a local database
4. In addition, every day the raw data will be stored on an archive as CSV files, the archive can be local disk or external storage.
5. A single dashboard is using the master server in order to locate the correct processing server and show the detections to the customer using a dashboard.
6. An API server is using the same method as the dashboard and allow customers to retrieve data from the system database via APIs

In addition to the main system architecture, we are also able to install the whole system on a single server without connection to the internet in order to support customers that do not want to provide internet connection. This requires a special version of the sensor.

EDGE sensor - it is also possible to install the processing server software on the sensor and to have an EDGE device that can process and present the results from a single sensor. The functionality of this device is similar to the full system with the exception that functions that require multiple sensors (for example mobility) are not currently available under this configuration.

4.2.2 The dashboard

4.2.2.1 LBASense Dashboard Overview

Overview

LBASense is Correlation Systems’ crowd data analysis solution, based on Wi-Fi sensors and big data collection and analysis.

LBASense cloud dashboard provides a variety of functions such as real-time visitor monitoring, movement pattern analysis (revisit rate, stay time, movement route), report function, equipment condition, etc.

It also provides Open APIs that can be linked to existing ERP. Applications provide big data-based real-time statistical information in various fields such as smart cities, events, advertising-marketing, disasters/safety/security, etc.

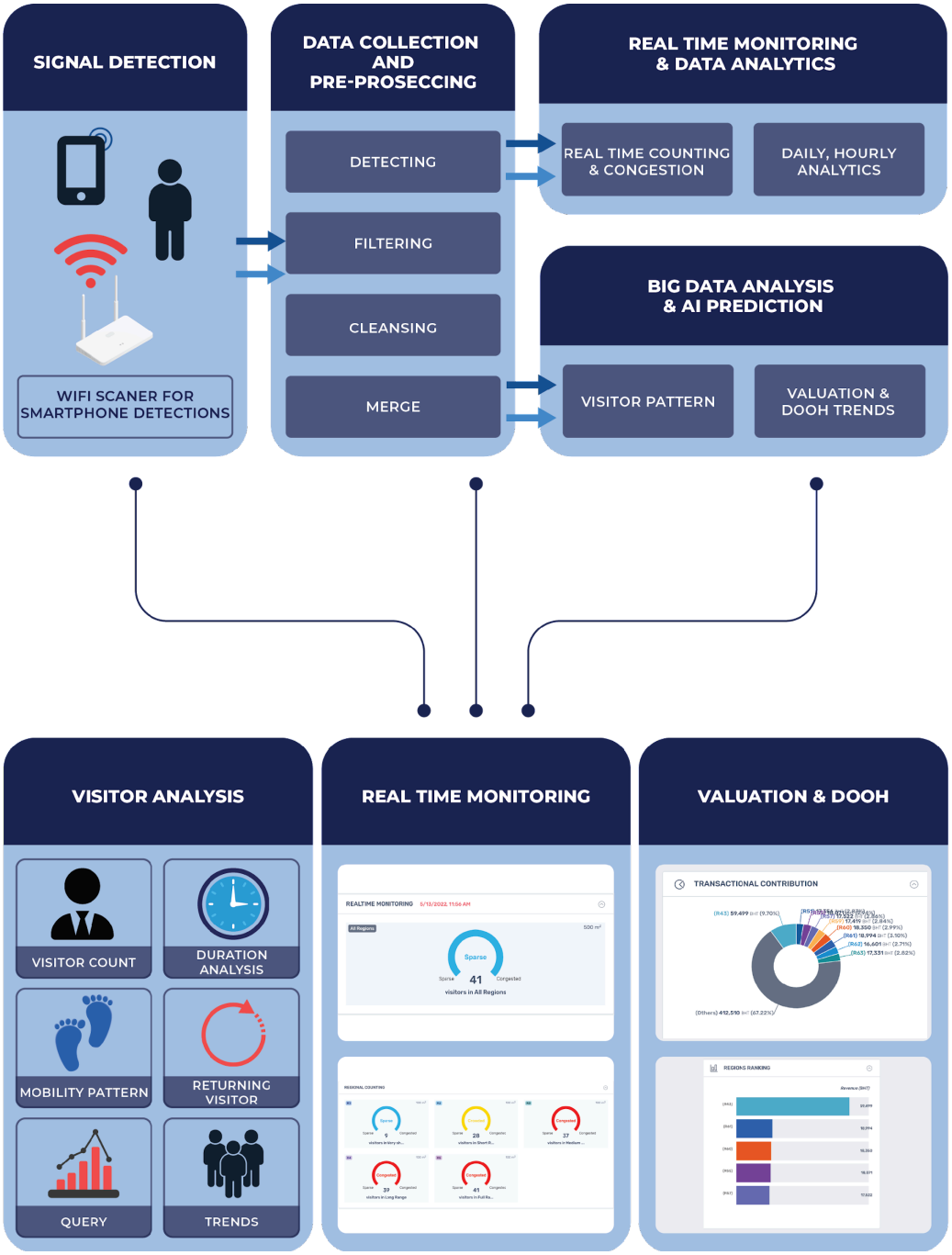

Figure 14 - Data flow chart

Composition

Division	Content
IoT Sensor Software	Software that collects Wi-Fi signals transmitted from smartphones in real time, anonymizes the collected data, and transmits them to a big data analysis server.
Analysis Software	Software that collects and preprocesses data transmitted in real time, and provides monitoring and statistical analysis data in real time.

Table 2 - Software Versions for the Dashboard and IoT sensors

4.2.2.2 Access to the Dashboard

Link to access LBASense dashboard:

- Web site : https://dashboard.lbasense.com

User Information

Users can login using the user information issued by the Correlation Systems customer success representative.

After logging, users can search for and select their site ID in the Select Site popup.

- ✔ Site Id:
- ✔ User information: Username + Password

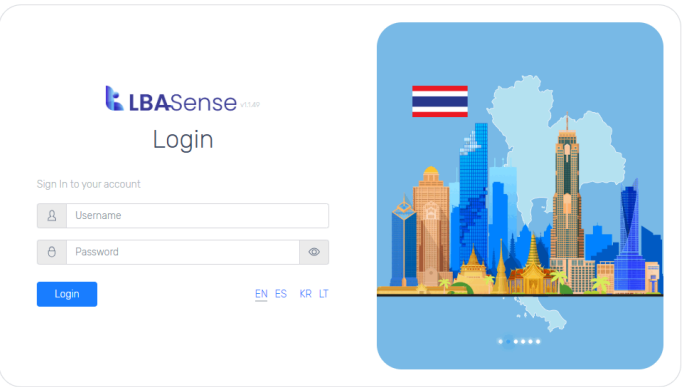

Figure 15 - LBASense login screen

The LBAsense dashboard provides multilingual support (English, Korean, Spanish and Lithuanian).

4.2.2.2 Dashboard menu description

The LBASense dashboard is largely divided into five categories: ① data visualization, ② mobility, ③ valuation, ④ system, and ⑤ reports. It analyzes big data collected from sensors and provides a variety of menus for users to easily analyze data using various visualization charts or graphs.

1st level	2d level	3rd level	Functional explanation
LBASense Dashboard	Login/ Logout	–	The ability to log in or out of the product.
	Data Visualization	Insights Map	Interactive map that allows users to click and view statistical data and heatmap
		Site Overview	Provides overall site visitor statistics such as current visitor counts, daily trend, monthly trend, and annual trend.

		Regions Overview	Provides regional visitor statistics such as current visitor counts, returning visitors, cumulative counts for daily, monthly, and annually.
		Overview Total	Provides overall site visitor counts and predictions of the day, week and month.
		Analytics	Provides regional trends of visitor counts, returning visitors and average durations as line graph by date & day & hour.
		Analytics Duration	Provides duration analysis of visitors per region.
		Geospatial	Provides cumulative visitor count of the day, represented on a map.
		Real time Monitoring	Real-time monitoring system that provides current visitor count and congestion level per regions, and current weather of the site.
		Asset Tracking	Real-time monitoring and alert system of assets
	Mobility	Mobility Map	Interactive map that provides all visitor movements between regions.
		Mobility Table	Shows visitor movements between two regions (e.g. R1→R2) as both values and percentages.
		Mobility N	Shows visitors movements of multiple routes among three regions (e.g. R1→R2→R3) represented in a map.
	Valuation	Daily OOH	Provides statistics of transactional contributions of each region based on the daily visitor counts and values.
		Daily OOH Trends	Provides daily transactional contribution of each regions as trends
		Hourly Valuation	Provides hourly transactional contribution of each region.

		DOOH	Provides daily valuation of digital billboards based on number of people exposed to each “blip”
		Hourly DOOH	Provides hourly valuation of digital billboards based on number of people exposed to each “blip”
	System	Sensor health	Provides sensor-related information (position, state, connection time, S/W version, operation time) management and inquiry functions.
		Region setup	It provides inquiry, registration, and modification functions for regional management.
	Alert	Alert Monitor	Provides alert monitoring service to the site. All alerts generated by the system are stored and viewed here.
		Alert Rules	Users can manage thresholds of alerts.
	Report	Monthly report	Generating reports by region and month
		Daily report	Generating reports by region and date and providing printing functions.
		Summary report	Generating reports by region
		CSV download	Download function is provided as a CSV file for the number of visitors per day, week, month, year, and hour.
API	Interlinked API	–	API uses JSON format GET type. Access API for analysis using username and password used when logging in on another platform.

Table 3 - Dashboard menu description

4.2.2.3 Data visualization

Real time monitoring

The Realtime Monitoring module displays the real time population data of the area where the sensor is installed in real time, and allows users to access minute by minute data for a specific day with a Playback History feature

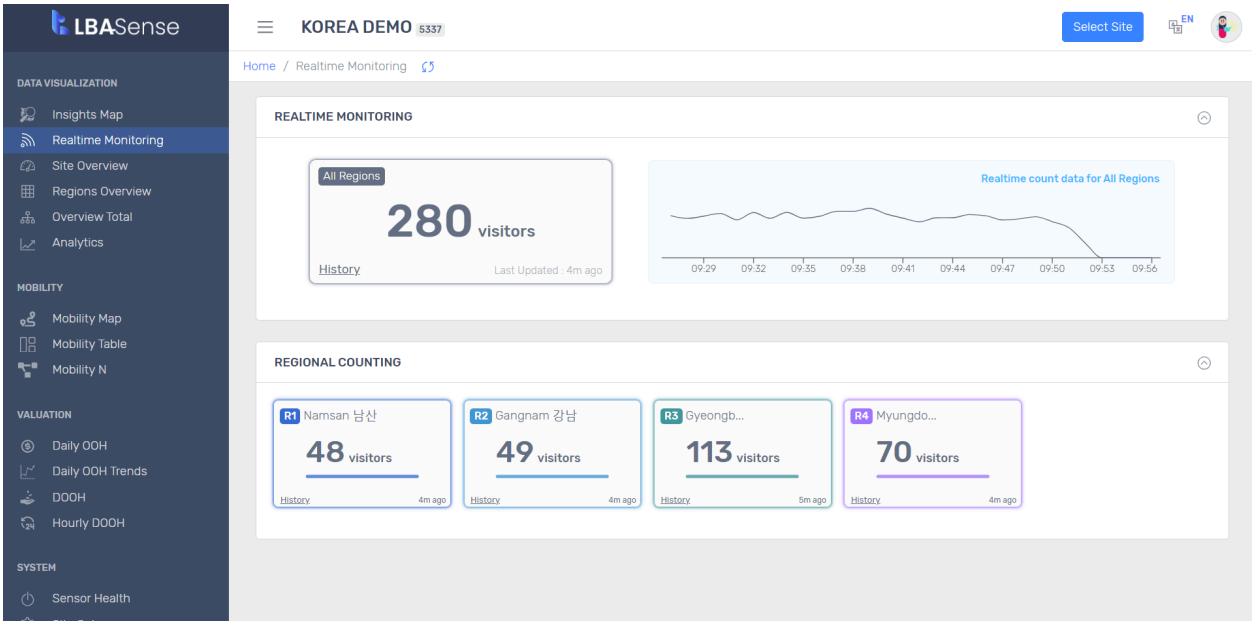

Figure 16 - Real time monitoring

The real time number of people detected by each sensor is displayed in one data card, as well as the congestion level in the region.

R1 Namsan... → ID and name of the region

48 visitors → number of visitors right now

4m ago → last time the data was updated

History → access playback history for the region

Figure 17 - Namsan Realtime data

Insights map

The Insights Map menu displays the population data of the area where the sensor is installed in real time (or on a daily basis), and allows users to check the sensor location and regional population analysis data on the map.

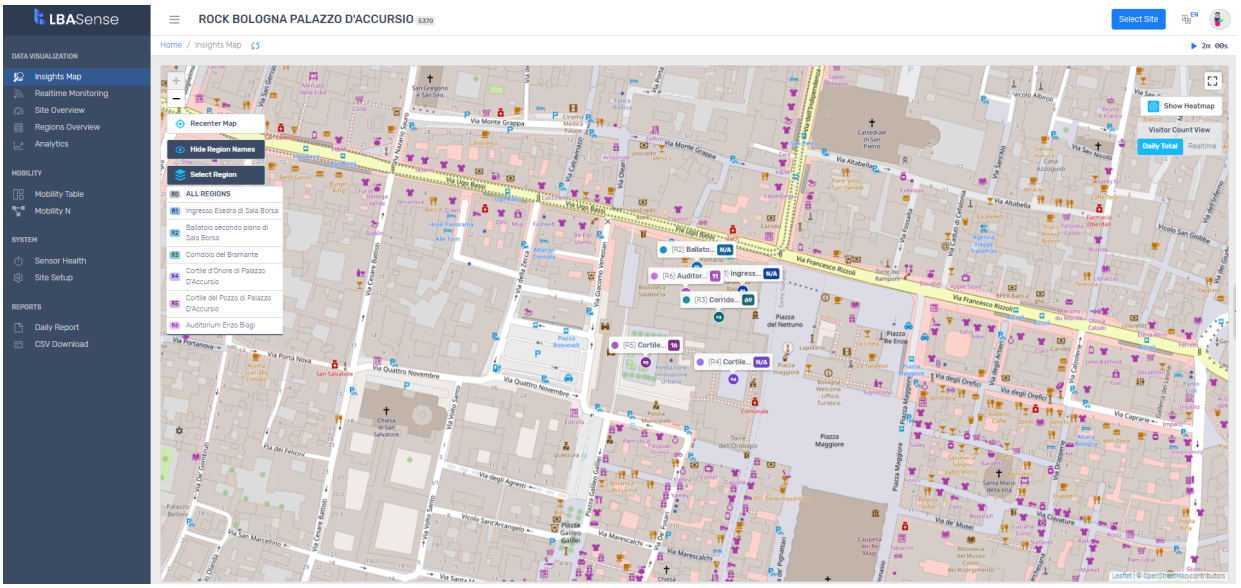

Figure 18 - Insights map

Users can select a region on the map, to view the details of the selected region (number of visitors by hour, status of visits, weekly visits, monthly and annual visits).

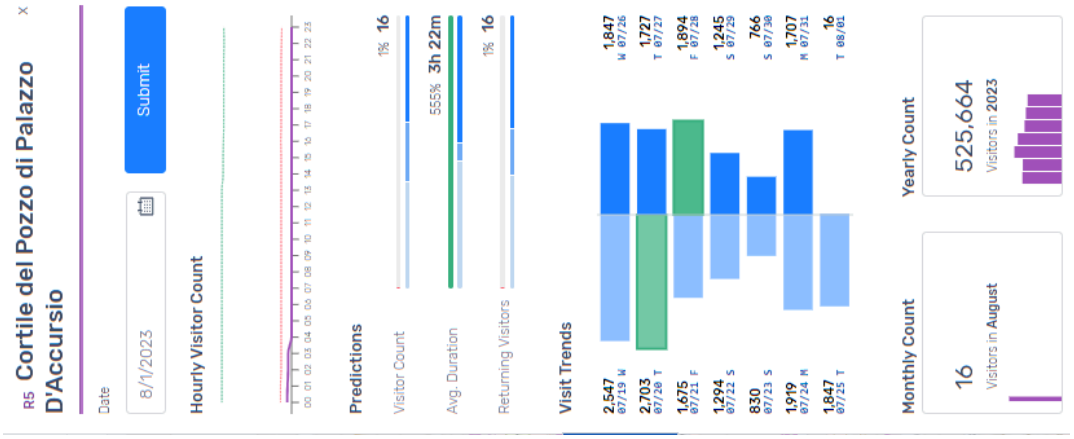

Figure 19 - Insights map - detailed region view

Site overview

The site overview module displays data by statistically analyzing population data of the entire area where sensors are installed, and provides data such as visit status, visit trend, and regional status.

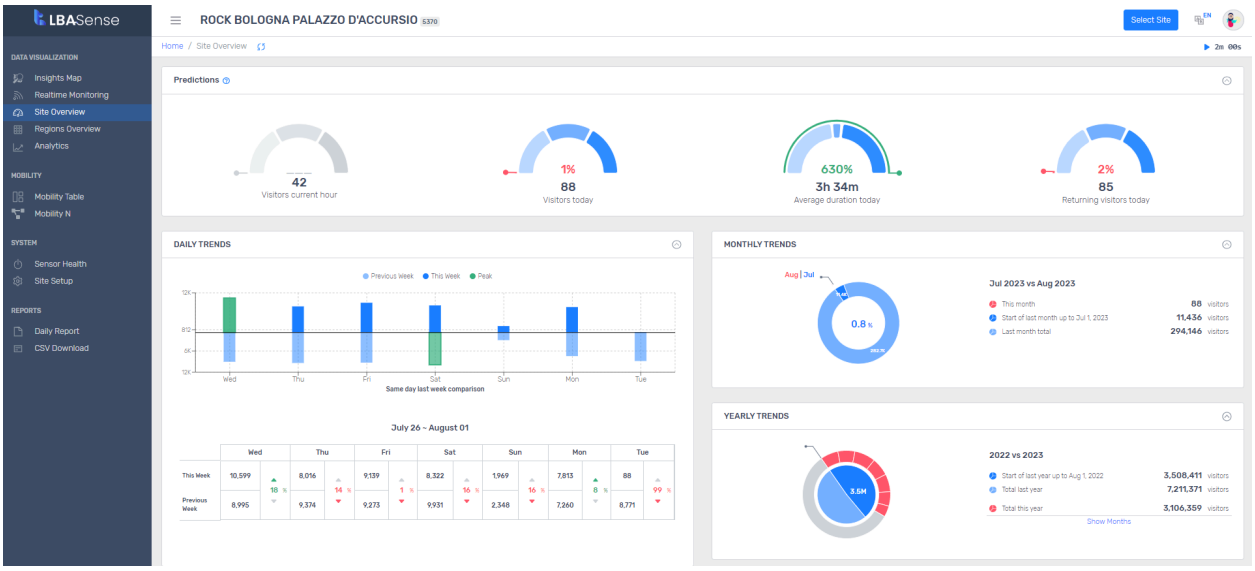

Figure 20 - Site overview

- By analyzing the data of the entire floating population of the site, the data is displayed through the chart with visitors within an hour, today's visitor, today's average residence time, and today's revisit.

- The visit status represents the range of the number of people measured today compared to the existing measurements, so you can check the information on the graph only when more than 10 matches of data are collected.

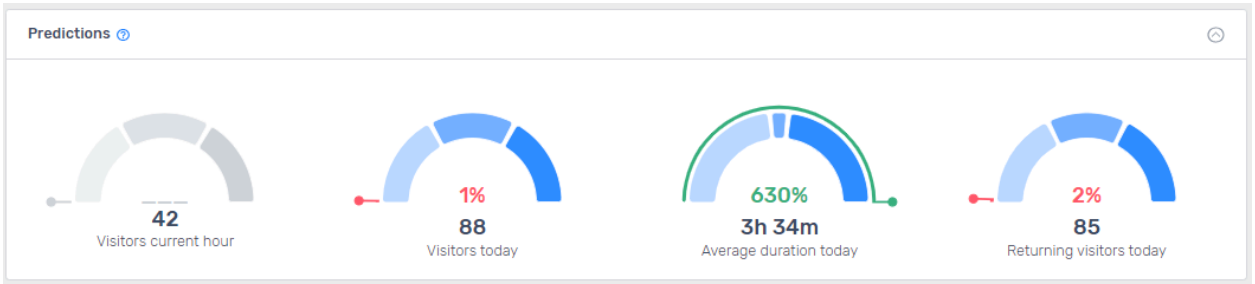

Figure 21 - Site overview - Predictions

Legend: upper prediction lower prediction

Weekly, Monthly and Yearly comparisons:

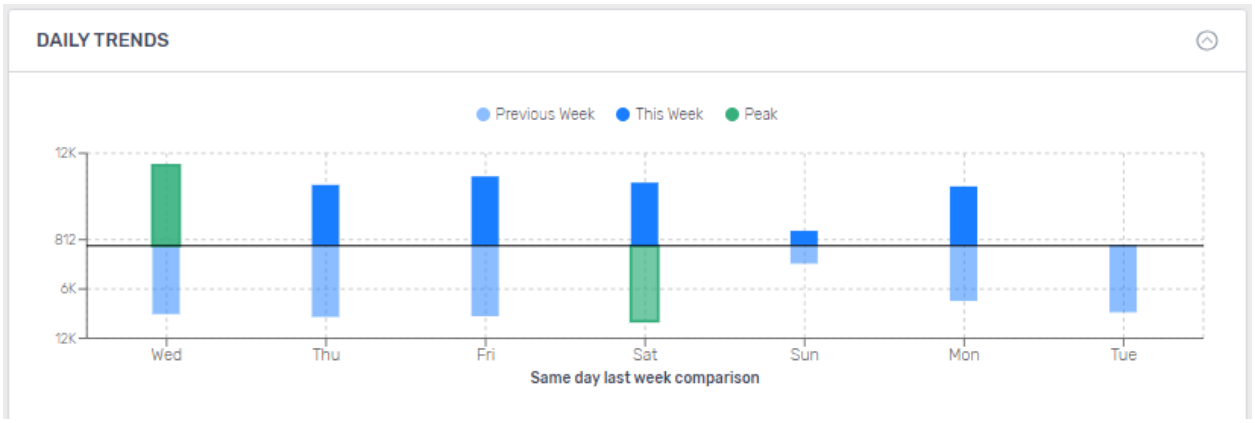

Figure 22 - This week versus last week comparison

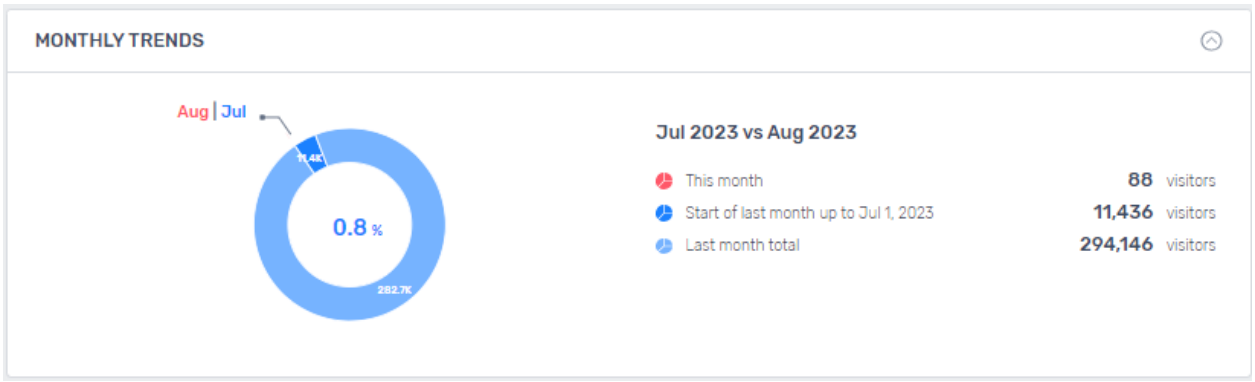

Figure 23 - This month versus Last month comparison

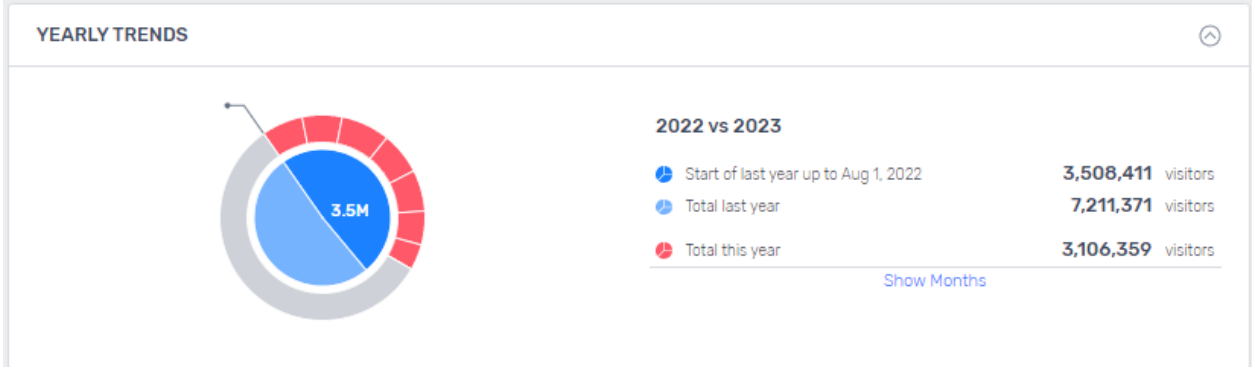

Figure 24 - This year versus last year comparison

Regions overview

The Regions Overview menu displays data by statistics and analysis of population data of the entire area where sensors are installed by region, and provides data such as visit status, visit trends, and regional status.

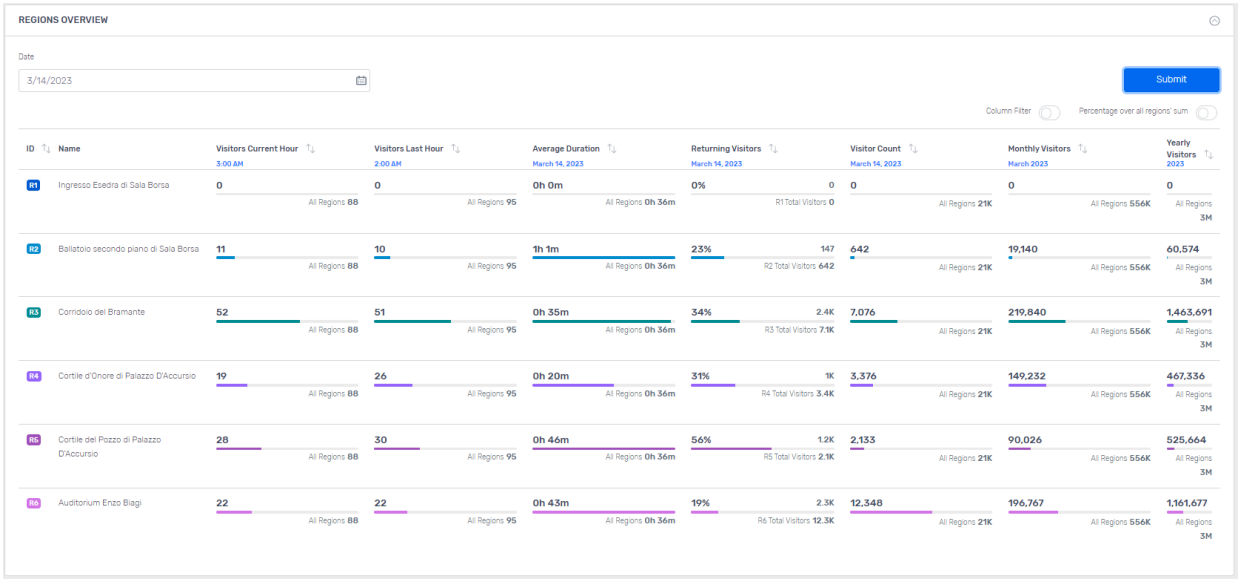

Figure 25 - Regions overview

This module displays the population data for all regions where sensors are installed. The analyzed data is displayed in the form of a table grid to enable users to visualize the data based on the region ID, region name, last hour visitor, today visitor, monthly visitor, and annual visitor items.

Users can select the “Percentage over regions sum” option (top-right, above the table) so that the regional population data becomes displayed as counts or ratios between each region and all regions.
Users can search for the region by region ID and region name by enabling the column filter option (on the left-side of the Percentage over regions sum”).

Finally, users can look for another day’s data by clicking on the calendar displayed above the table.

Analytics menu

The Analytics menu provides the ability to analyze data by inquiring data collected from areas where sensors are installed under various search conditions.

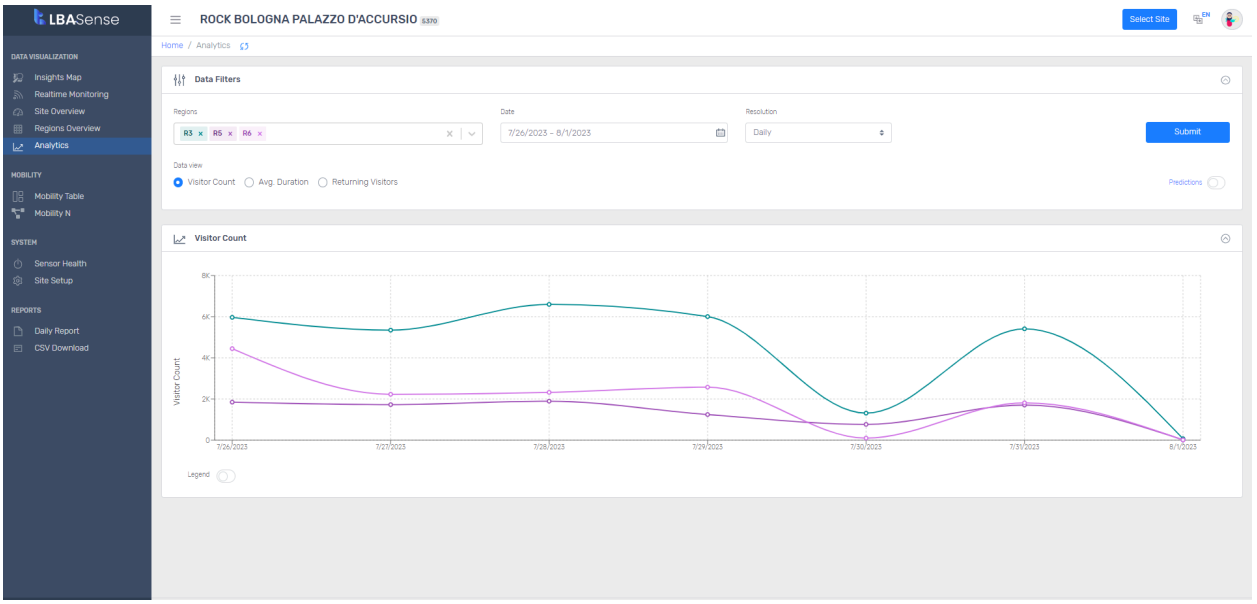

Figure 26 - Analytics

Data Filters:

- Select Regions: Users can select a specific region to display,
- Select Date: Users can query a specific time period,
- Select Resolution: Users can select a time resolution (monthly / daily / hourly)
- Select Data View : Users can select to view data as Visitor Count / Average Duration / Returning Visitors

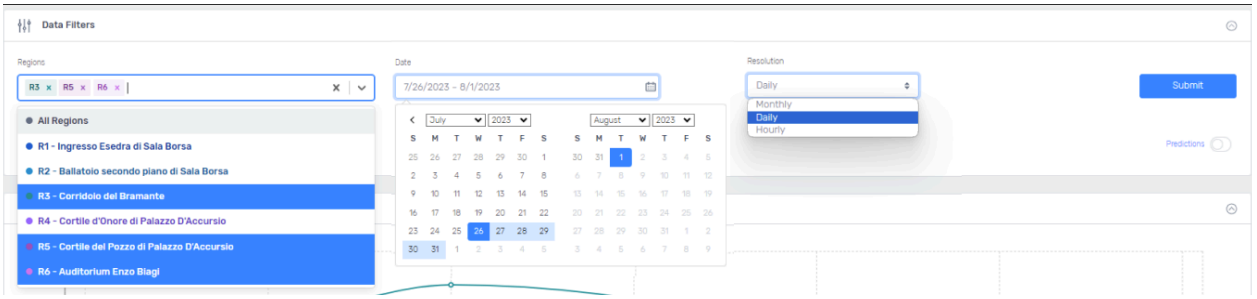

Figure 27 - Analytics - data filters

Analytics Duration

The Analytics Duration menu provides the ability to analyze visitors duration for an entire site or a specific region, on daily basis

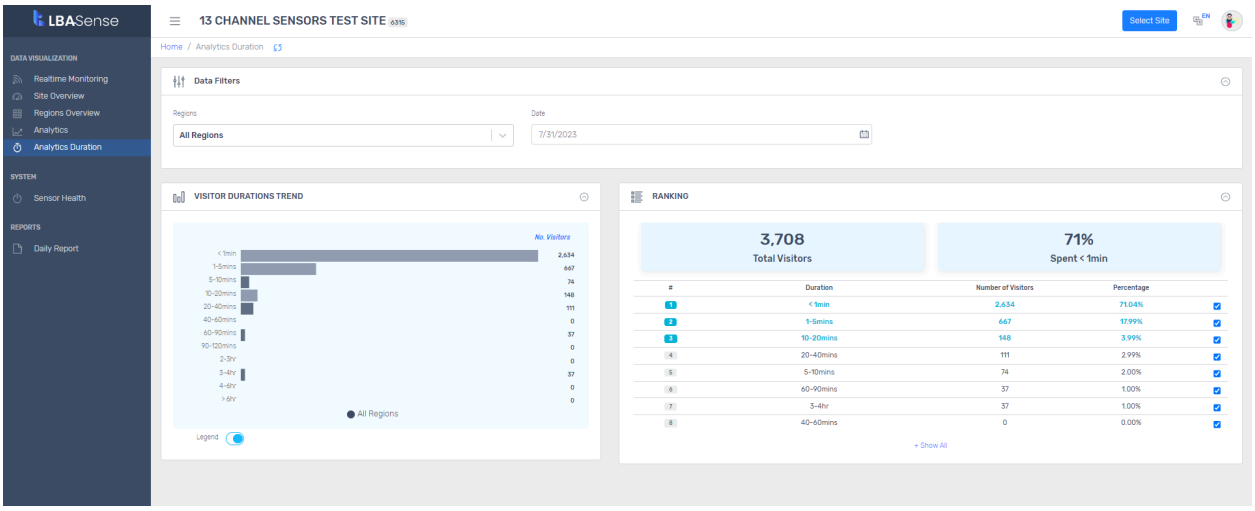

Figure 28 - Analytics Duration

How to understand Analytics Duration?
Here is how to read the table below: “on the 31st of July 2023, 71% of all visitors (or 2,634 visitors) stayed less than 1 minute on site; 17.99% of all visitors (or 667 visitors) started between 1 and 5 minutes on site, etc.”

Figure 29 - Analytics Duration - Ranking

Mobility N

The MobilityN menu provides a visual representation of journeys undertaken by visitors in a site.

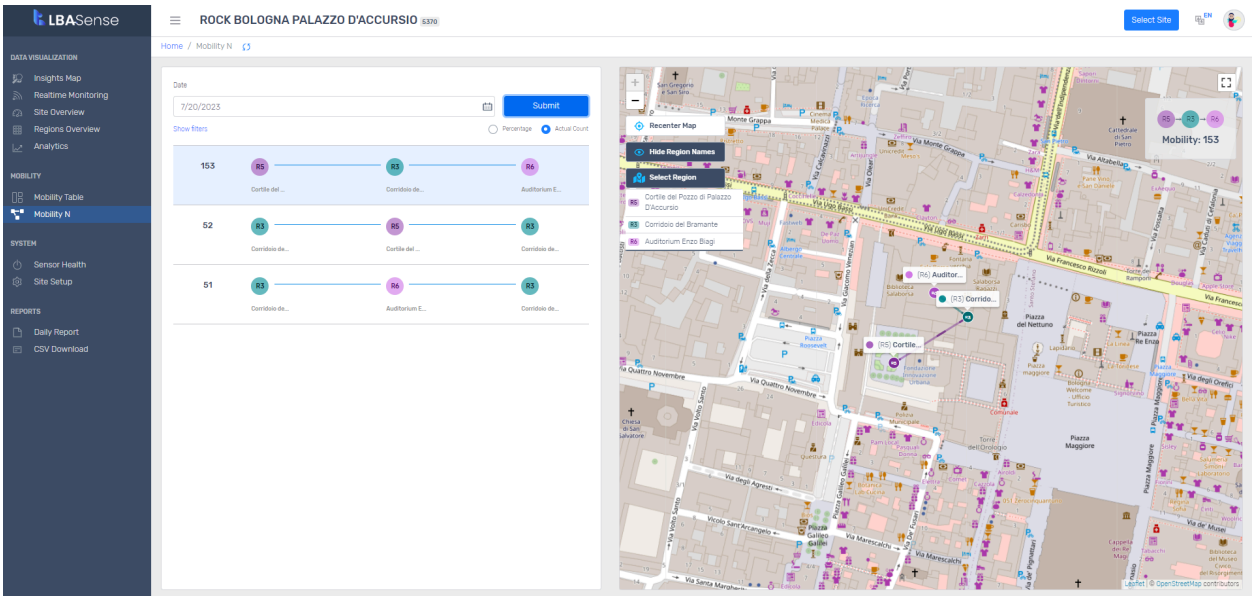

Figure 30 - Mobility N

The MobilityN analysis is based on the number of regions selected by the user.

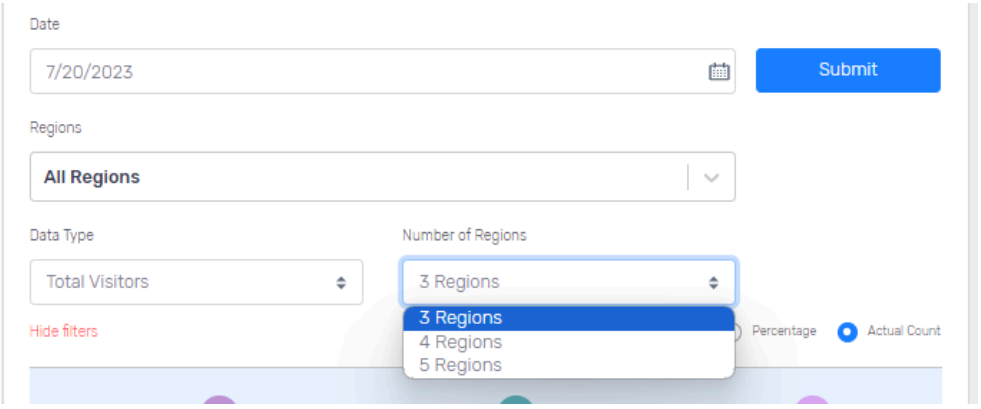

Figure 31 - Mobility N - Data filters

Note: sites with small amounts of visitors (< 500 daily visitors) may not see any data displayed when selecting more than 3 regions because not enough people would have traveled to all four or five regions in the site to conduct the mobility analysis.

4.2.2.3 System management

Sensor Health

The Sensor Health module displays sensor statuses of the entire site. Sensor status data is provided using a table grid and a map for additional context.

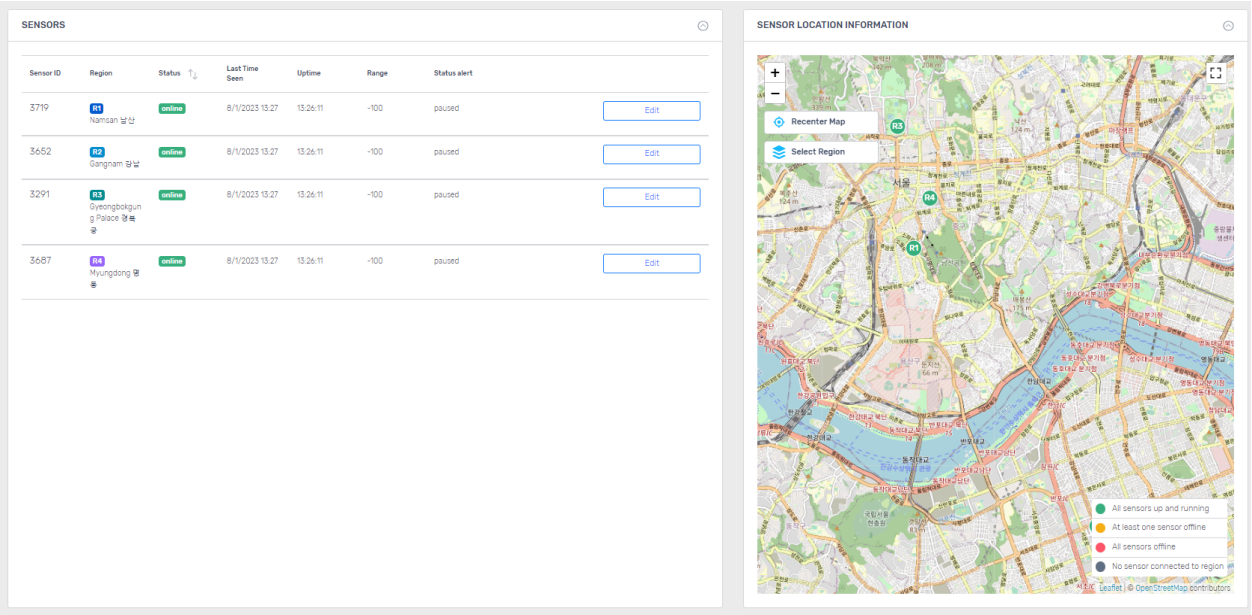

Figure 32 - Sensor Health

Notes:

- The sensor status is displayed based on sensor ID, region name, sensor state, most recent connection time, operating time, sensor range and status alert.
- Users can click on the sensor on the map, to access the sensor's information, such as sensor ID, status, recent connection time, and operating time will show up.

Sensor configuration capabilities:

Users can edit sensor range and activate/pause a sensor status alert by clicking on the Edit button next to the relevant sensor.

(see below)

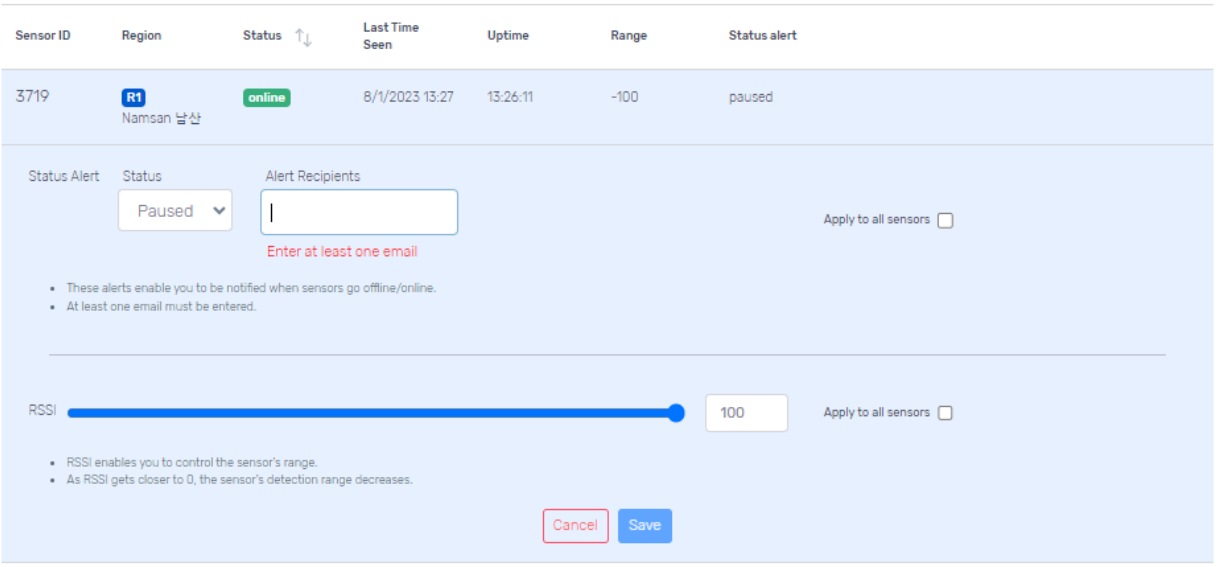

Figure 33 - Sensor Health - Edit sensor

What are Status Alerts (shown in Figure 33 above) ?

Sensors may go offline during their deployment. Status alerts enable users to receive an email notification, informing them when a sensor goes offline. The system sends another email notification to the email recipients when the sensor goes back online.

What is RSSI ?

RSSI defines the detection range of a sensor. The closer the range is to 100, the larger the detection range is.

In an open space, without any obstacles, RSSI = 100 is about equal to a 100m radius.

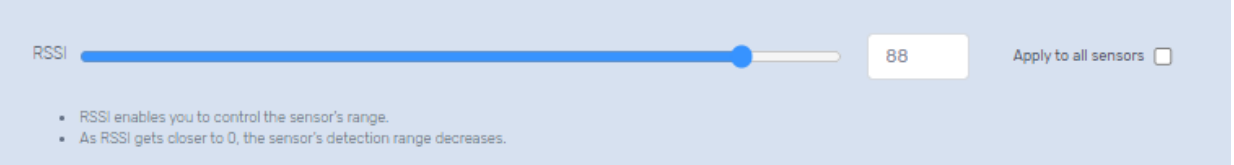

Figure 34 - Sensor Health - Edit RSSI

Note:

Users may select the “Apply to all sensors” checkbox in order to force all sensors in a site to use the same RSSI value

Site Setup

The Site Setup menu displays general site information and additional information on where the sensors are installed. In this menu, users can edit region information (region name, region description, GPS coordinates) and site information..

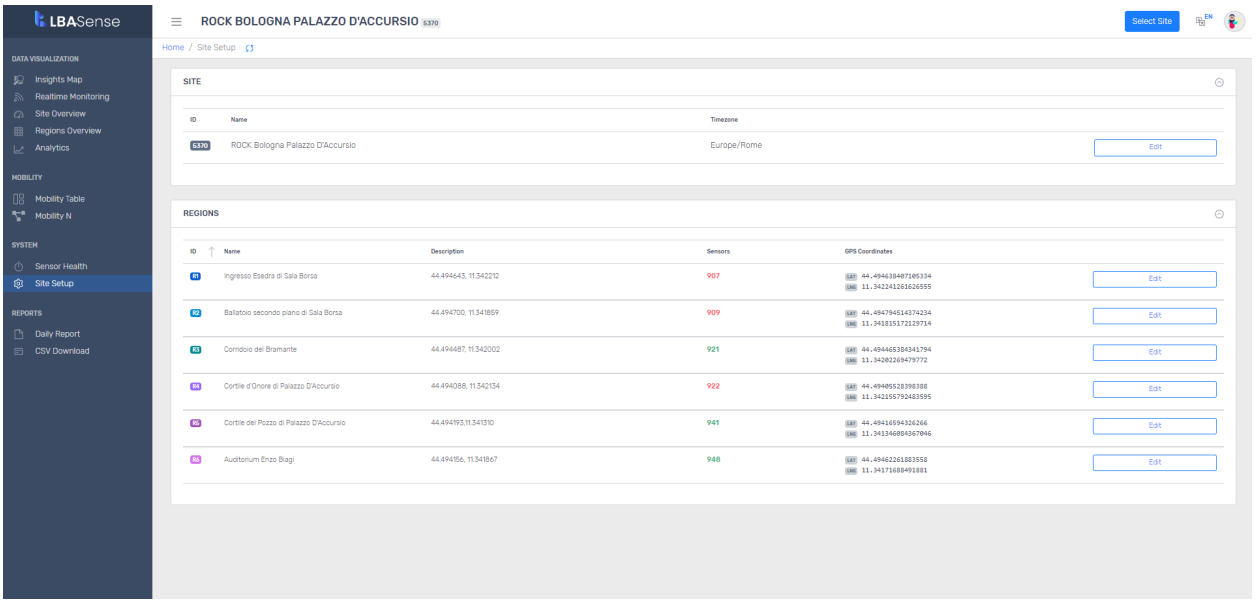

Figure 35 - Site Setup

- Users can edit information (region name, region description, GPS coordinates) using the modification button of each region information.
- Users can edit Site Name and/or Site Timezone: The choice of timezone is very important, as it is used to determine when days start and end, which is information used to calculate daily visitors, daily MobilityN, Duration, etc.

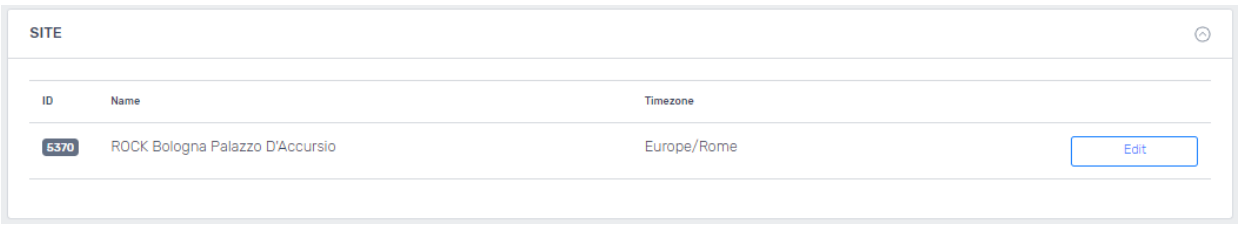

Figure 36 - Site Setup - Site name and Site timezone

4.2.2.3 Reports

Daily Report

The Daily Report menu provides the analyzed population data in the form of a daily report. Report data items display data such as visit status, visit trend, and regional status.

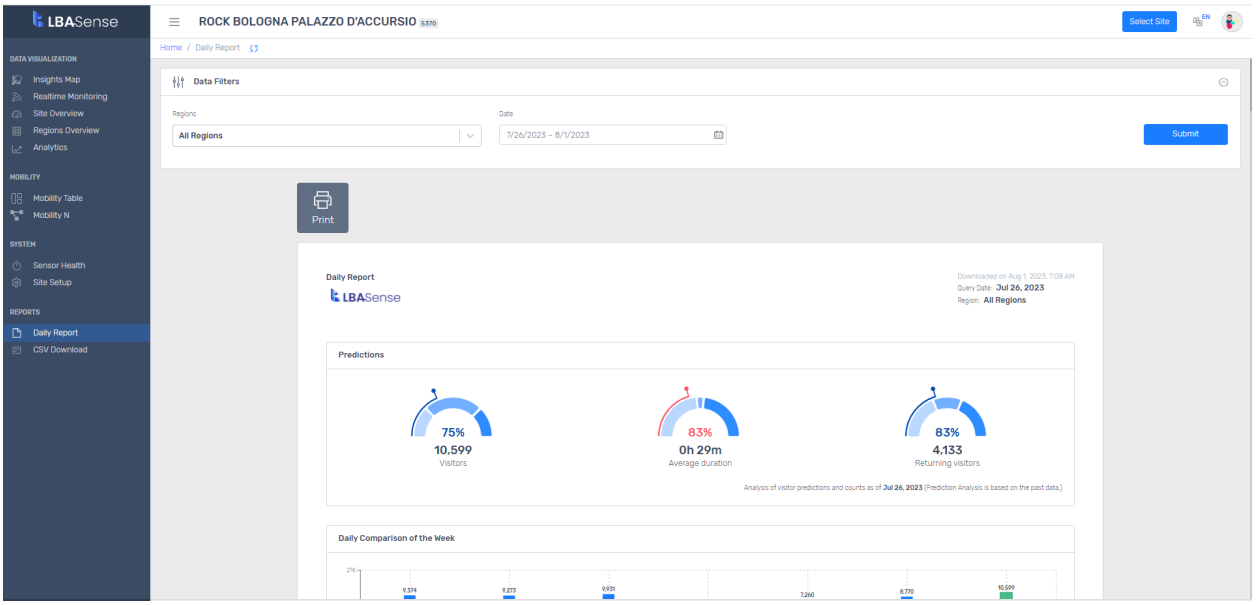

Figure 37 - Daily Report

Notes:

- Regional status selection: Users can select to display data for a specific sensor.
- Date Selection : Users can select the date they want to create the report for.

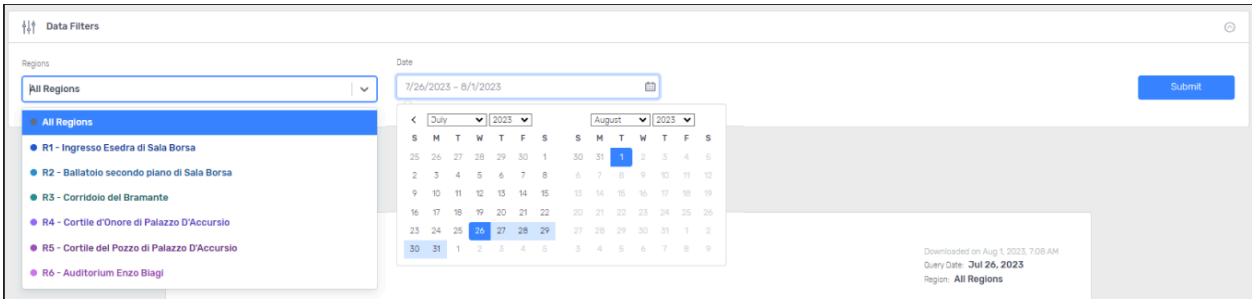

Figure 38 - Daily Report - filters

CSV download

The CSV download menu provides the analyzed floating population data as a CSV file. File items include daily visitors, weekly visitors, monthly visitors, annual number of visitors, hourly visitors, average duration, returning visitors, and visitors' mobilities.

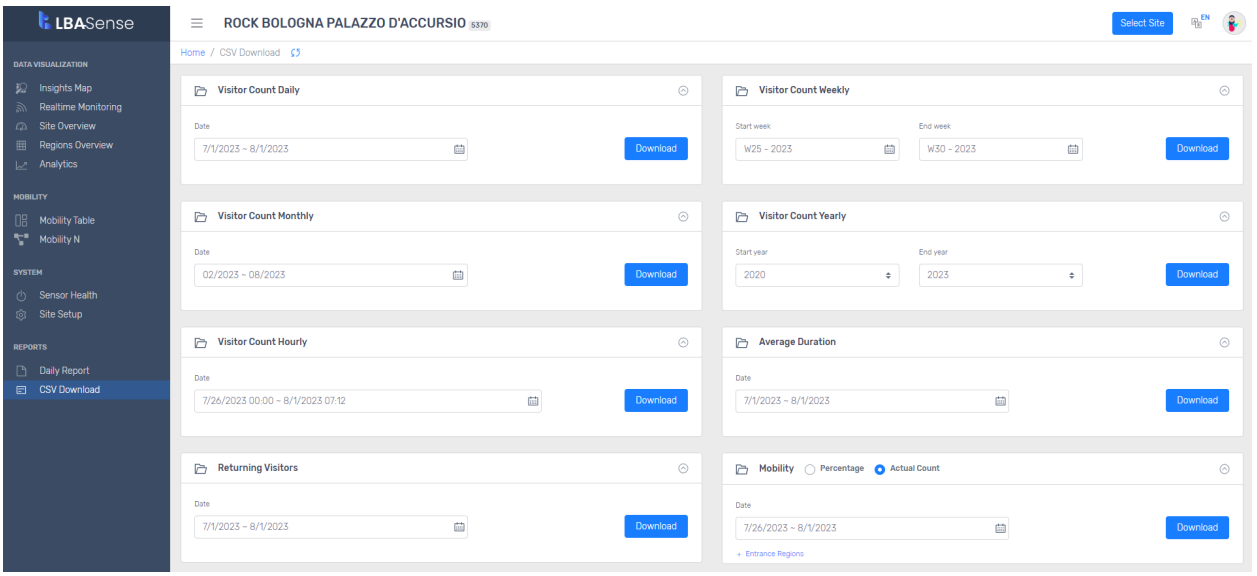

Figure 39 - CSV Download

Note:
- Date Selection: Users can select the date they want to create the CSV file for.

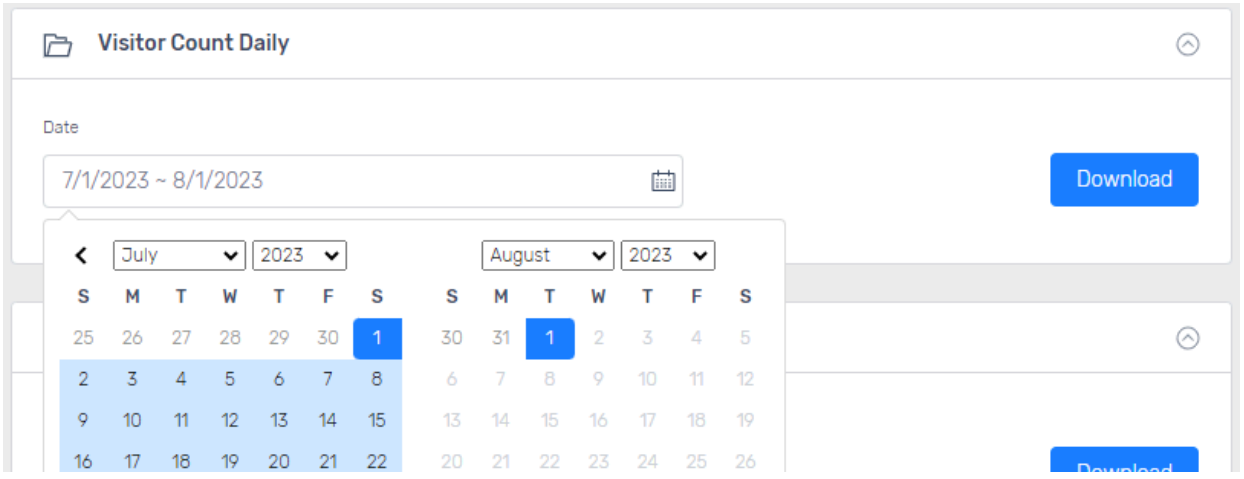

Figure 40 - CSV Download - filters

4.2.3 The APIs

LBASense provides many APIs that enable users to harness the power of LBASense solutions. In this section, we will provide an overview of the system, and describe some of the APIs that are available on the platform, and how to use them.

4.2.3.1 Available APIs

The LBASense developer tools and endpoints are grouped into the following APIs.

API name	Description
VisitorCount	These APIs let you query your historical data by defining a time frame, resolution and more.
RealtimeVisitor	This API lets you query your realtime data and run a playback
SensorHealth	This API lets you query the current state of your sensor (online/offline, software version,etc.)
RSSI	These APIs enable you to query the current RSSI value for a sensor and update to a new value.

Table 4 - LBASense APIs

4.2.3.2 Basic Requirements

All APIs described above require the user to provide their user credentials (username + password) to authenticate the request.

4.2.3.3 Concept

Parameter	Explanation
site	site is the logical collection of sensors where all data collected is processed and fused into a single Situation Awareness Picture. Usually, a site includes multiple regions

region	region is the logical area with geographic boundaries and some logical similarities. Regions usually reflect the detection ranges of the sensors; they extend according to customers' requirements
sensor	sensor is the physical device, uniquely identified by a numerical ID assigned by the LBASense system, that detects anonymous signals

Table 5 - LBASense APIs - Concept

4.2.3.4 Details for Visitor Count API

Sample call for Visitor Count

```
https://{domain}/api/Analytics/VisitorCount/?format=json&site_id={site_id}&user={username}&pass={password}&start_time={start_time}&end_time={end_time}&resolution={resolution}
```

Sample response for Visitor Count

```
{
    "data": [
   {
            "date": "2023-03-21",
            "region_data": [
        {
          "region": 0,
          "count": 478
        },
                {
          "region": 1,
          "count": 478
                }
    ]
   }
 ]
}
```

Sample call for Mobility

```
https://{domain}/api/Analytics/AnalyticsMobility/?format=json&site_id={site_id}&user={username}&pass={password}&start_time={start_time}&end_time={end_time}
```

Sample response for Mobility


```
{
    "data": [
   {
          "date": "2023-03-21T00:00:00",
          "mobility_data": [
        {
          "region_from": 1,
          "region_to": 2,
          "count": 478,
          "percentage": 50,
          "valid": true,
        },
              {
          "region_from": 1,
          "region_to": 3,
          "count": 124,
          "percentage": 50,
          "valid": true,
              }
    ]
   }
 ]
}
```

4.2.4 Focus on mobility

Overview of the Mobility feature

One of the key features of Correlation Systems' IoT sensors and the LBAsense system is mobility analysis. This feature empowers system owners to analyze movement patterns across entire cities.

Mobility analysis provides insights into the number of people who travel from specific points (e.g., Sensor A) to other destinations (e.g., Sensor B and then Sensor C). This information, referred to as Mobility N (where N represents the number of waypoints in a journey), can be generated for any combination of sensors.

Mobility N analysis is derived by processing raw sensor data. Whenever a mobile device is detected in a new location (i.e: by a sensor), a movement pattern is created. The system then aggregates these mobility points over a time period to generate the Mobility N results.

Our system offers two methods for generating mobility patterns:

- Tracking: Provides real-time insights by continuously processing data as devices move.
- Rebuild: Enables on-demand analysis of historical data by reprocessing past information.

The system processes data daily. After this process, users can choose to delete it or retain it for future analysis. Historical data is stored in CSV format.

How does LBASense currently handles data privacy

To safeguard user privacy, the system anonymizes detected MAC addresses through a strong encryption process. Here's how it works:

- 2048-bit Hashing Algorithm: A robust 2048-bit hashing algorithm is used for encryption. This algorithm transforms the MAC address into a unique, seemingly random string of characters.
- Private Key Mechanism: Each customer receives a unique private key. This key acts as a secret ingredient in the encryption process, ensuring the anonymized output (hashed MAC address) is different for each customer, even if they share the same original MAC address.
- One-Way Encryption: The encryption process is one-way. This means that decrypting the anonymized data back to the original MAC address is mathematically infeasible.

erel — ubuntu@LUCKY: /data/sites/site5201/realtime/tracks — ssh LUCKY — 163×24

```
site_id,region_id,visitor_id,population_type,global_id,first_time_seen,last_time_seen,rssi_avg,n_events,manufacturer,ssid_1,length
5201,5,16300924937,0,1,2023-03-14 13:42:30,2023-03-14 13:52:58,-82,5,"Samsung Electronics Co.,Ltd",NOSSID,73
5201,5,16300924937,0,1,2023-03-14 17:54:47,2023-03-14 17:54:47,-83,1,"Samsung Electronics Co.,Ltd",NOSSID,28
5201,7,20473088318,0,1,2023-03-14 07:08:48,2023-03-14 07:12:32,-83,2,Xiaomi Communications Co Ltd,NOSSID,4006
5201,6,28072931170,0,1,2023-03-14 08:36:01,2023-03-14 08:36:01,-77,1,"vivo Mobile Communication Co.,Ltd.",NOSSID,116
5201,1,29228591224,0,1,2023-03-14 14:15:59,2023-03-14 14:15:59,-76,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,127
5201,6,31312429953,0,1,2023-03-14 08:58:17,2023-03-14 08:58:17,-87,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,45
5201,1,34120225413,0,1,2023-03-14 08:00:42,2023-03-14 08:04:38,-75,2,"vivo Mobile Communication Co.,Ltd.",NOSSID,82
5201,1,34120225413,0,1,2023-03-14 16:56:59,2023-03-14 16:58:15,-83,2,"vivo Mobile Communication Co.,Ltd.",NOSSID,28
5201,3,34120225435,0,1,2023-03-14 08:17:18,2023-03-14 08:17:18,-90,1,"vivo Mobile Communication Co.,Ltd.",NOSSID,4040
5201,7,37775429168,0,1,2023-03-14 14:07:25,2023-03-14 14:09:09,-78,8,"Four systems Co.,Ltd.",NOSSID,4071
5201,2,39973669916,0,1,2023-03-14 10:14:01,2023-03-14 10:14:01,-78,1,"Samsung Electronics Co.,Ltd",NOSSID,58
5201,4,40035823296,0,1,2023-03-14 17:55:31,2023-03-14 17:55:39,-82,2,Xiaomi Communications Co Ltd,NOSSID,34
5201,8,40382905618,0,1,2023-03-14 08:59:11,2023-03-14 08:59:11,-82,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,127
5201,2,41437330744,0,1,2023-03-14 06:56:53,2023-03-14 06:56:53,-76,1,"SIMCom Wireless Solutions Co.,Ltd.",NOSSID,86
5201,3,44079180380,0,1,2023-03-14 01:06:44,2023-03-14 01:06:44,-83,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,86
5201,3,44079180380,0,1,2023-03-14 02:31:12,2023-03-14 02:31:12,-96,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,86
5201,4,44079180380,0,1,2023-03-14 03:22:56,2023-03-14 03:22:56,-86,1,"HUAWEI TECHNOLOGIES CO.,LTD",NOSSID,86
5201,6,46505467646,0,1,2023-03-14 14:56:00,2023-03-14 14:56:00,-77,1,SAMSUNG ELECTRO-MECHANICS(THAILAND),NOSSID,65
5201,2,49695995117,0,1,2023-03-14 05:32:57,2023-03-14 05:32:57,-80,1,"Samsung Electronics Co.,Ltd",NOSSID,86
```

Figure 41 - Raw data - Real time

times_from	times_to	trace_counts	unique_trace_counts
2022-09-14 00:00:00	2022-09-14 03:00:00	22932	13
2022-09-14 03:00:00	2022-09-14 06:00:00	2581	10
2022-09-14 06:00:00	2022-09-14 09:00:00	10521	34
2022-09-14 09:00:00	2022-09-14 12:00:00	32710	29
2022-09-14 12:00:00	2022-09-14 15:00:00	43433	44
2022-09-14 15:00:00	2022-09-14 18:00:00	40296	28
2022-09-14 18:00:00	2022-09-14 21:00:00	37503	21
2022-09-14 21:00:00	2022-09-15 00:00:00	31043	7

Figure 42 - Number of “K”

4.2.5 The sensors

Correlation Systems sensors are based on a Software Defined Radio which is able to listen passively to the WiFi 2.4 GHz (some versions are capable of morning the 5GHz band in addition to the 2.4GHz band) and to extract the WiFi headers that are transmitted over this frequency band.

OM2P	
Picture	
WLAN Standard	801.11n (1x1)
Main Antenna	3dBi RP-SMA Omni
RF Power	MCS0: 400mw (26dBm) / MCS7: 127mw (21dBm)
Receive Sensitivity	MCS0: -95dBm / MCS7: -73dBm
Ethernet	2 (WAN & LAN)
POE	12-18v (non-802.3af)
Power Supply	12vdc, 110vac
LEDs	Power, Ethernet(2), Mesh
Temperature	0-50 C
Dimensions	3.75” x 2.75” x 1”
Certification	FCC / IC / CE

Table 6 - OM2P sensor specs

S100	
Picture	
Processor	650 MHz
Antenna	Dual external antenna
WLAN standards	IEEE 802.11b/g/n 2.4 GHz (scanning interface, not for connectivity)
Memory	64 MB DDR
Ethernet	2 x 100 Mbps
LEDs	4 (Power, Ethernet (2), WiFi)
Certifications	KC, CE, RoHS
Operating temperature	0-50 C
Dimensions	11.22 cm x 2.53 cm x 8.68 cm (excluded antenna)
Weight	200 g
Enclosure dimensions	13,2 cm x 15.0 cm x 4.8 cm
DC Input	12 V / 1 A
PoE Input	Passive 36-57 V (standard 802.3af)
Supported ranges	1, 3, 5, 7, 10, 25, 100 meters.

Table 7 - S100 sensor specs

Indoor sensor	
Picture	
Communication Protocol	LAN
Processor	1.5G-Z 64-bit quad-core ARM Cortex-A72
Memory/Storage	4GB / 32GB
Antenna	Dipole or Directional
Power Supply	5V (via USB type-C)
Power Consumption	Max 10W Average 4W
Operating Temperature	-30 ~ 75 C
Dimension	103mm x 144 x 36mm
LAN	1X 10/100/1000M LAN port (RJ45 interface)
WIFI	2.4GHz and 5DHz 802.11a/b/g/n/ac
Enclosure Box	Aluminium IPo
Mounting Option	Available
Supported Detection Ranges	Software controlled up to 100 meters (Open Space)

Table 8 - Indoor sensor specs

Outdoor Sensor	
Picture	
Communication Protocol	LAN
Processor	1.5G-Z 64-bit quad-core ARM Cortex-A72
Memory/Storage	4GB / 32GB
Antenna	1 antenna (external)
Power Supply	802.11af-PoE
Power Consumption	Max 4.7W Average 3 37W
Operating Temperature	-30 ~ 75 C
Dimension	212mm x 220 mm x 84mm
LAN	1X 10/100/1000M LAN port (RJ45 interface)
WIFI	2.4GHz and 5DHz 802.11a/b/g/n/ac
Enclosure Box	Aluminium IPo/
Mounting Option	Available
Supported Detection Ranges	Software controlled up to 100 meters (Open Space)

Table 9 - Outdoor sensor specs

Outdoor sensor with GPS	
Picture	
Communication Protocol	LAN
Processor	1.5G-Z 64-bit quad-core ARM Cortex-A72
Memory/Storage	4GB / 32GB
Antenna	GPS antenna, omnidirectional antenna
Power Supply	Power cables
Power Consumption	Max 4.7W Average 3 37W
Operating Temperature	-15 ~ 50 C
Dimension	197mm x 147 x 86mm
LAN	1X 10/100/1000M LAN port (RJ45 interface)
WIFI	2.4GHz and 5DHz 802.11a/b/g/n/ac
Enclosure Box	Plastic
Supported Detection Ranges	Software controlled up to 100 meters (Open Space)

Table 10 - Outdoor sensor with GPS specs

4.3 INNOVATION COMPARED TO THE STATE OF THE ART

4.3.1 End User Access to Datasets

A key innovation in our solution lies in enabling individuals to verify if their data is included within anonymized datasets, all the while preserving their anonymity. This aligns with GDPR's Article 15, granting individuals the "right to obtain confirmation as to whether or not personal data concerning him or her are being processed."

4.3.1.1 The limitations of the current SotA.

Current methods for verifying data inclusion in anonymized datasets often create a conflict between user privacy and data controller efficiency.

In fact, Correlation Systems' very own processes when answering third-party's data privacy requests (in the context of the GDPR), were less than satisfactory as they would force the third-party to disclose some private information.

In detail, below are the two methods currently used by Correlation Systems (and, to our knowledge, most data controllers) in verifying if an individual's personal data is included in their datasets:

- Proof of identity and device ownership: This method, while providing a definitive answer for the data controller, requires users to disclose additional information that could potentially compromise their right to privacy (e.g: the individual must disclose their name, email or their device's MAC address to enable the data controller to query their datasets using these identifiers). Indeed, by revealing private data to the controller, users essentially sacrifice some anonymity in exchange for verification.
- Captive Portal Hardware: This approach aims to protect user anonymity by utilizing a dedicated hardware device. The captive portal is able to capture the device's MAC address, and a software can then use this information to query available datasets for a matching MAC address. While this process can be considered more privacy-friendly than the above-mentioned one, as it can be fully automated hence removing the need for direct intervention from the data controller, it introduces a security concern: the potential for unauthorized individuals to exploit the verification process and gain access to data.

4.3.1.2 UtiP-DAM: an improvement over the SoTA

Our solution breaks this trade-off between anonymity and verification by offering a more elegant approach:

Enabling anonymous verification with quantitative exposure information:

- Our system grants users access to a secure platform without requiring them to reveal any personal data, as even in the process of auditing a dataset, users choose among location points that were included in the original dataset. This ensures anonymity throughout the verification process.
- The dataset auditing tool provides users with quantitative information (i..e: the “K” value) indicating the level of anonymity of their personal data (i.e: the higher the K, the more anonymity they can preserve). This information enables users to understand how exposed their private data is, without compromising the anonymization itself.

By providing this anonymous verification mechanism, our solution aligns well with the spirit of Article 15 of the GDPR. Users can exercise their right to information regarding their data's inclusion in anonymized datasets while maintaining complete anonymity.

4.3.2 Sharing of raw mobility data

Another main innovation in our solution lies in enabling dataset owners to share their anonymized datasets on the UtiP-DAM marketplace. This aligns very well with the upcoming EU’s Data Act:

- Focus on data sharing: A core principle of the EU Data Act is the promotion of FAIR (Findable, Accessible, Interoperable, Reusable) data sharing practices. UtiP-DAM's user-friendly marketplace directly aligns with this objective by providing a secure platform for the exchange of anonymized mobility data. This data type is particularly relevant, as it often falls under the category of user-generated data, a key target for open sharing initiatives within the EU.
- Interoperability: The Data Act promotes interoperability standards for data sharing across different market actors. UtiP-DAM adheres to interoperability values, specifically by promoting the sharing of datasets on the IDS.

4.3.2.1 The limitations of the current SotA.

Traditionally, sharing raw mobility data, particularly in privacy-conscious regions like Europe, has been limited due to concerns about re-identification. Public releases of raw mobility data (e.g., (e.g: researchers were able to re-identify individuals in a poorly anonymized dataset released by Transport for London in 2014) have unfortunately demonstrated the potential for individuals to be identified.

4.3.2.2 UtiP-DAM: an improvement over the SoTA

Our solution introduces a novel approach to this challenge: leveraging k-anonymity to enable the secure sharing of raw mobility data. While k-anonymity itself is not a new technique, its application in this context represents a significant advancement.

Here are two of the main reasons why Correlation Systems believe that this solution is innovative:

- Balancing data utility with privacy: K-anonymity ensures that location data, and the journeys derived from their analysis, cannot be uniquely linked to individuals by guaranteeing that each data point appears within a group of at least k-1 similar data points. This offers a robust privacy guarantee.
  Indeed, consider a scenario where location data is k-anonymized using k= 10. This means any given data point will be indistinguishable from at least 9 other data points within the anonymized dataset. This is what we call “anonymity in the crowd”. This approach strikes a good balance between data utility and privacy, allowing valuable insights to be extracted from mobility patterns while safeguarding individual anonymity.
- Facilitates collaboration and innovation: Enabling the sharing of anonymized raw mobility data, unlocks new possibilities for collaboration and innovation. Researchers, businesses, and other stakeholders can access this anonymized data to develop novel applications and gain deeper insights into mobility patterns. Imagine researchers studying traffic congestion patterns in a city. Traditionally, such studies might rely on aggregated data that offers limited granularity. With access to anonymized raw mobility data, researchers could potentially analyze individual movements within the anonymized dataset, while still protecting individual privacy, to develop more effective traffic management strategies. This anonymized data could also be valuable for businesses, allowing them to understand customer behavior and develop location-based services or optimize delivery routes.

4.3.3 EDGE distributed mobility

4.3.3.1 The limitations of the current SotA.

Traditionally, mobility calculations rely on a central database architecture. Sensor data converges at this central point, where mobility patterns are then calculated. This approach raises security concerns, as a single point of failure can compromise the entire system.

4.3.3.2 UtiP-DAM: an improvement over the SoTA

Moving towards edge computing, where processing occurs closer to the data source, presents a challenge for mobility calculations. In such decentralized systems, sensors often operate with isolated databases, hindering the calculation of mobility patterns that require combined data from multiple sources.

Sharing data directly between all edge devices is technically feasible but comes with drawbacks as this approach essentially replicates the central database at the edge, creating numerous smaller points of vulnerability compared to a secure cloud-based central system.

The UtiP-DAAM solution

Our project proposes a novel approach that overcomes these limitations:

- Each edge device (i.e: IoT sensor) maintains its own database, minimizing the risk of deanonymization compared to a shared data server. This approach ensures data privacy by limiting the exposure of individual sensor data.
- Each Edge device can share relevant information with nearby Edge devices in a privacy-preserving manner. To promote “privacy-preserving”, we will use different "visitor ids" for each Edge device and will allow Edge devices to share raw data to a predefined list of Edge devices using the "visitor id" of the target device (see Figure 41 for a screenshot of Correlation Systems’ raw sensor data where “visitor id” are visible. This anonymized data exchange facilitates the calculation of mobility patterns even in a decentralized environment.

Our decentralized approach offers several advantages:

- Enhanced security: Distributing data storage and processing across multiple Edge devices reduces the risk of a single point of failure and minimizes the impact of potential security breaches.
- Improved scalability: The system can easily accommodate additional Edge devices without requiring significant changes to the infrastructure, making it highly scalable.
- Increased privacy: By limiting data exposure and leveraging privacy-preserving communication techniques, our solution prioritizes user privacy.

This innovative approach to decentralized mobility calculations paves the way for secure and scalable processing of location data at the edge, while safeguarding user privacy.

5 SOFTWARE DESIGN AND ANALYSIS, COMPONENT SPECIFICATION

5.1 SOFTWARE MODULES

5.1.1 Understanding the system Architecture of UtiP-DAM

5.1.1.1 Overall System Architecture

The proposed architecture is based on Correlation Systems' existing architecture, as we detailed in part 4.2.1 Top Level System Architecture of this document.
The components that will be developed and integrated for the UtiP-DAM project are indicated in red in the figure below.

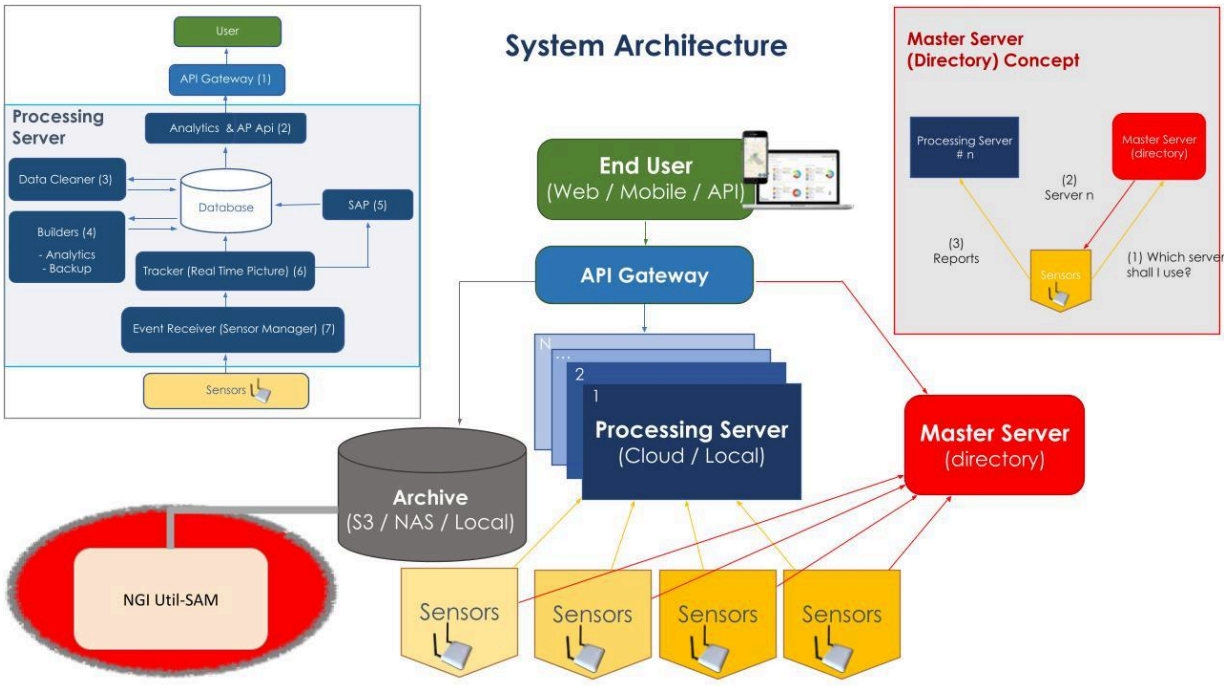

Figure 43 - Overall System Architecture

5.1.1.2 Focus on: EDGE Architecture

Since a fully operational system cannot be created from scratch for this project, we will leverage a simulated environment replicating a large edge site: specifically, we will use the Bangkok site to simulate the distributed version for the UtiP-DAM tools. This approach allows us to test and validate the functionality of the distributed system without compromising the integrity of the live production environment.

Here is how we will proceed:

1. We will set up a mirror site replicating the Bangkok BTS site which collects around 500,000 data points daily. This mirror site will function as a testbed for the distributed Utip-DAM system.
2. To accurately simulate the behavior of this large edge site, we will utilize real data collected by sensors of the BTS site. This real-world data will be fed into the mirror site, giving us a realistic representation of a large-scale deployment.
3. Correlation Systems' sensors have a built-in feature that enables duplicating reporting to multiple servers. We will leverage this feature to direct sensor data to both the Bangkok BTS site and the mirror site simultaneously.
4. A custom utility will be developed to replace the standard sensor software on the mirror site. This utility will receive the data from the duplicated channel, ensuring the mirror site receives the necessary sensor data for the project.

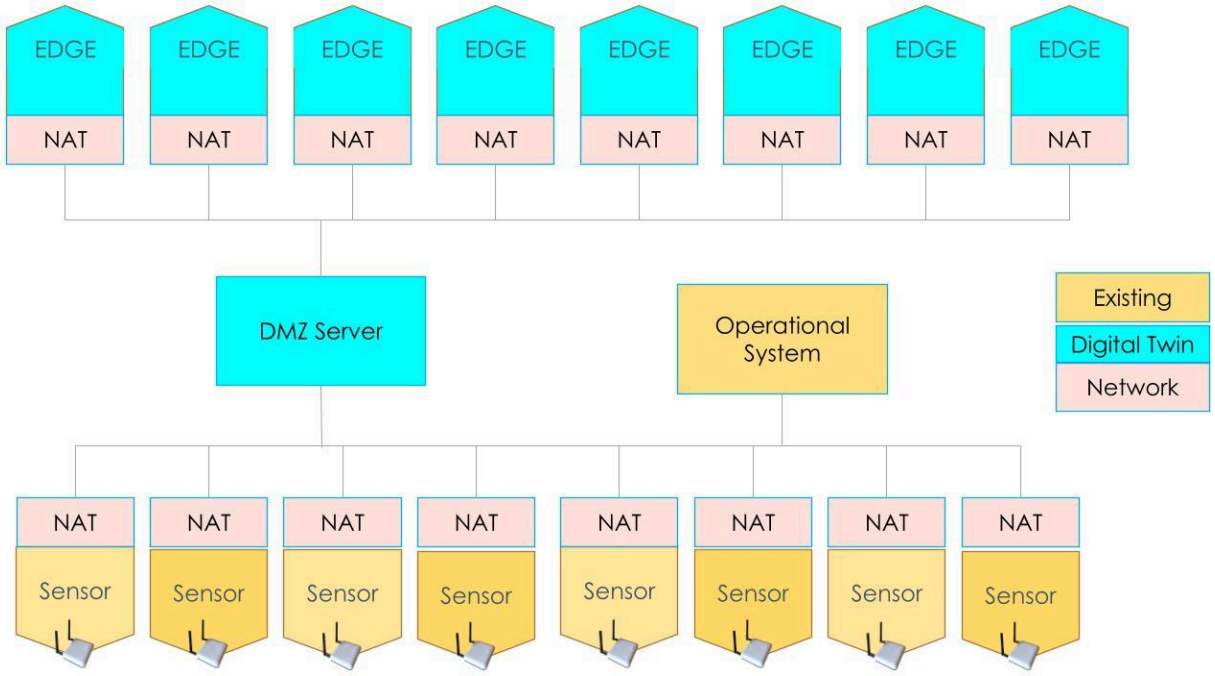

Figure 44 - Edge Architecture

Networking considerations

Typically, an EDGE sensor will be connected to the internet via a modem. In such a case. by default, the sensor will be located behind NAT and will not be accessed from outside the network. A similar architecture is also common in private installations where the local firewall prevents access from the local network to the sensor.

As this is the most common situation for the deployment of sensors, our prototype will be operated behind NAT and access to sensors will be available only via polling.

5.1.2 Software Design

5.1.2.1 Design Considerations

Our vision for Utip-DAM extends to various deployment scenarios:

- Integration with existing systems: Utip-DAM should seamlessly integrate with existing mobility data collection and management systems (e.g: Correlation Systems' LBASense software), enriching these systems' functionalities.
- Auditing tool for legacy datasets: Utip-DAM serves as an external tool for auditing and analyzing existing mobility datasets within external systems.
- Interactive testing and auditing platform: Utip-DAM offers an interactive user interface specifically designed for testing and auditing purposes, facilitating user interaction with the system.

To cater to these diverse deployment needs, Utip-DAM adopts a modular design approach. This approach offers several advantages:

- The Utip-DAM's core functionalities are encapsulated within standalone, self-contained packages. This modularity simplifies integration with third-party libraries and external components.
- The modular design allows for efficient code reuse across different Utip-DAM deployments.
- A modular architecture simplifies maintenance and updates, as changes can be implemented within specific modules without affecting the entire system.

5.1.2.2 Top-level software design

As we have mentioned before, Utip-DAM is designed with a modular architecture, allowing for flexibility and integration with various user needs.

Core Functionalities

The core functionalities of Utip-DAM, including anonymization and auditing of mobility datasets, are packaged as standalone modules. Again, the purpose for this modular design is to facilitate integration with third-party applications, minimizing dependency requirements.

The “core functionalities” of UtiP-DAM are defined as:

- Mobility Data Set Management: This module provides tools to handle and manipulate mobility data sets effectively.
- Data Auditing: This module analyzes mobility datasets, identifies potentially risky patterns, and suggests solutions for mitigating those risks.
- Anonymization: This module processes anonymization requests of mobility data sets, preserving user privacy.
- Anonymity auditing: This module allows users to search for specific patterns within mobility datasets.

Distributed mobility

The distributed mobility module is a software deployed on the Edge device that arranges:

- A database where raw data captured by the device is stored,
- A software that anonymizes raw data locally by replacing an identifier (i.e: MAC addresses) by a randomized “Visitor ID”,
- A software that enables an Edge device to communicate its raw, anonymized data with a predefined list of Edge devices using the “Visitor ID” of the target device.
- A software that enables an Edge device to anonymize its data (includes both the data received from other devices and the data captured by the device) locally by running the k-anonymity algorithm and then sends to the cloud server the result (i.e: the dataset) of this process.

Other Utilities

- Open Market Interface: This utility enables users to upload mobility datasets to the International Data Space (IDS) and potentially other data marketplaces, facilitating data sharing and collaboration.
- Mobility Sandbox: This web-based graphical user interface serves as a sandbox environment for users to explore and test the core functionalities of Utip-DAM.

5.1.2.3 Interface with the International Data Space

Utip-DAM, whose overarching goal is to facilitate the sharing of anonymized mobility data, will interface with the International Data Space initiative.

Standardized Metadata Descriptions

Utip-DAM will leverage the IDS Connector framework for easy communication with the IDS infrastructure. These connectors act as bridges, ensuring secure and compliant data exchange based on IDS specifications.

Standardized Metadata Descriptions

Utip-DAM will use standardized metadata descriptions within the IDS framework. This "common language" allows data providers such as UtiP-DAM, to accurately describe their anonymized mobility datasets. This facilitates efficient discovery and retrieval by potential data consumers within the IDS ecosystem.

Here are some key metadata parameters:

- Title: A clear and descriptive title for the mobility data resource.
- Description: A detailed description of the data content, including anonymization techniques used and geographical coverage.
- Keywords: Relevant keywords for easier data discovery (e.g., "pedestrian movement," "traffic patterns").
- Usage Policy: Description of the terms and conditions for accessing and using the data resource.

By adhering to these standardized metadata practices, Utip-DAM ensures its anonymized mobility data is discoverable and easily integrated into various IDS-compliant applications and data analysis workflows.

5.2 WORK PLAN FOR DEPLOYMENT

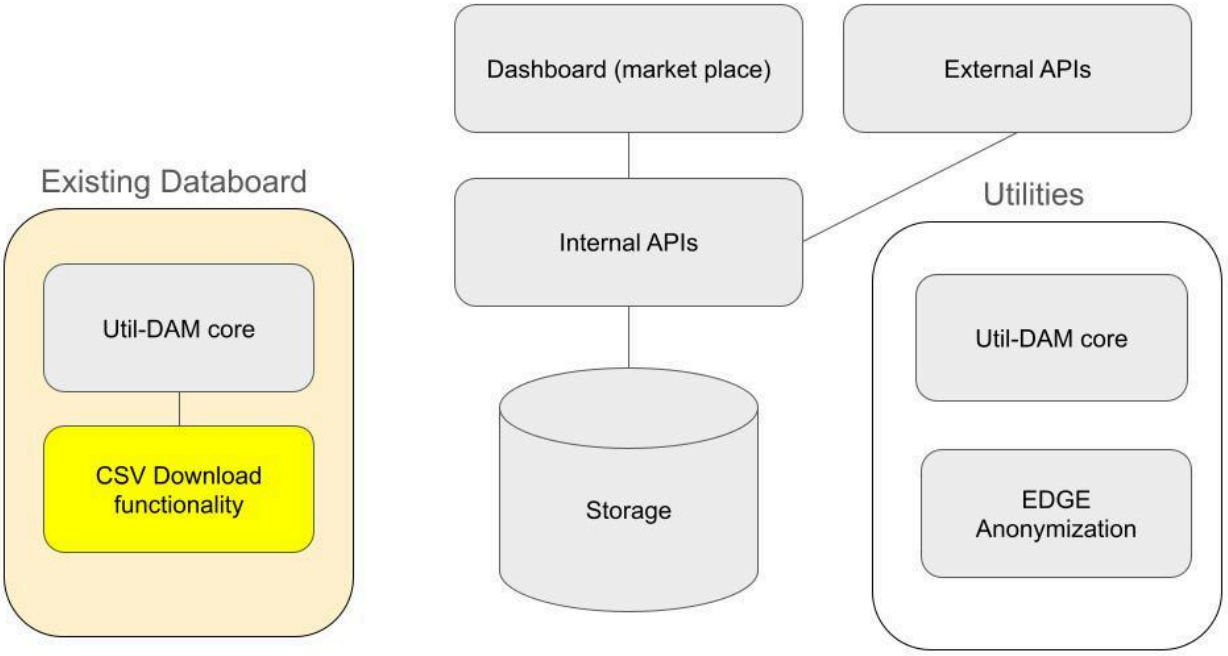

Figure 45 - Top- Level Architecture

Legend:

Gray - new (foreground)

Yellow - existing (background)

Utip-DAM leverages a three-tier architecture to facilitate efficient data management, visualization, and interaction:

5.2.1 User Interface (Marketplace Dashboard):

As described in detail earlier, this layer represents the user interface, also known as the UtiP-DAM marketplace. It serves as the primary point of interaction for users. The dashboard utilizes various functionalities provided by the internal API layer (see below) to perform actions such as data visualization, data exploration, and potentially data upload or management (depending on user permissions).

5.5.2 Internal API Layer (Services):

This middle layer acts as the service layer, providing core functionalities to the user interface. It essentially powers the dashboard functionalities. Here's a breakdown of services offered:

- Access and management: this service facilitates access and management of data stored by the system (see part 5.3.3), including both anonymized mobility datasets within the marketplace and administrative information such as user account information, dataset metadata (e.g: name, price, etc.), etc.
- Data processing and analysis: this service handles interfacing with backend services to process user requests such as anonymization of a dataset, auditing of a dataset, etc.
- User management: this service manages user accounts, authentication, and authorization for accessing different marketplace functionalities.
- Dataset purchase: this service manages interfacing with Stripe APIs for the purchase of premium datasets.

5.5.3 Data Storage

This layer serves as the data repository for the system. It stores:

- Marketplace Data: This includes the anonymized mobility datasets uploaded and shared within the marketplace.
- Administrative Information: This includes data related to user accounts, login credentials, and metadata associated with datasets within the marketplace.

5.5.4 External APIs

In addition to the core three-tier architecture, Utip-DAM also includes an external API layer. This API allows external software applications to interact with the marketplace programmatically. This bi-directional communication enables functionalities such as:

- Data Retrieval: External applications can leverage the UtiP-DAM APIs to access and download anonymized mobility datasets from the marketplace.
- Data Uploads: these APIs will facilitate automated data uploads from external applications, allowing authorized users to contribute their datasets to the marketplace.

6 DETAILED WORK PLAN FOR IMPLEMENTATION AND DEPLOYMENT (PRELIMINARY)

Our development strategy will prioritize user-centricity through rapid prototyping. Our iterative approach will allow us to achieve the following:

- Early functionality delivery: we aim to deliver 80% of the core functionalities (see part 5.1.2.2) as quickly as possible: we envision a beta delivery of these features in early M4. This strategy will enable users to interact with a functional prototype early in the development cycle.
- User feedback integration: by providing a working prototype early on, we can gather valuable user feedback. This feedback will then be incorporated into the remaining time dedicated to development, ensuring the finalized product aligns closely with user needs and expectations.
- Business concept validation: developing a functional prototype allows us to test the core business concept of the UtiP-DAM project in a real-world setting. This process helps us identify potential challenges and opportunities for improvement before pouring more resources in full-fledged development.

Breakdown of the approach

We will start development with the user experience (UX) design, focusing on building a user interface (UI) that is intuitive, easy to navigate, and caters to various user needs, including accessibility requirements.

Following the initial UX and UI development, we will then focus on implementing the backend services and APIs that power those functionalities.

Benefits of this Approach

- Prioritizing core functionalities and gathering user feedback early on through interviews and pilot activities, helps us mitigate development risks and ensure the final product resonates with users.
- Focusing on the most critical functionalities first optimizes development time and resource allocation.
- A user-centric approach, coupled with iterative development, leads to a product that is more likely to meet user needs and expectations.

This fast prototyping strategy with a focus on the quick release of a GUI allows us to deliver value quickly, validate the business concept, and ultimately develop a successful solution that meets the needs of our target audience.

6.1 Work plan for implementation

	Month	1	2	3	4	5	6	7	8	9
1	User Centric Design									
1.1	Interview with end users	X	X							
1.2	Secure datasets for the deployment phase	X								
2	UX									
2.1	Figma design first version	X								
2.2	Figma design 2nd version	X								
2.3	Dashboard mockup (without backend services)		X							
2.4	First version of the dashboard			X	X					
2.4.1	Dataset download			X						
2.4.2	User query (find me function)				X					
2.4.3	Upload function (without encryption)				X					
2.5	2nd version of the dashboard					X	X			
2.5.1	Create a new dataset					X	X			
2.5.2	Payment system integration						X			
2.6	Final version and code documentation							X	X	
3	Internal API and storage		X	X	X	X	X	X	X	

3.1	Initial design		X							
3.2	Services for dashboard			X	X	X				
3.3	Services for external APIs					X	X			
3.4	Final version and code documentation							X	X	
4	External API									
4.1	Selection of dataspace framework	X	X							
4.2	Service design and definition			X	X					
4.3	Implementation and deployment					X	X			
4.4	Final version and code documentation							X	X	
4.5	Sample program - how to use the API								X	X
5	Utilities									
5.1	Core (anonymization and auditing) services	X	X	X	X	X				X
5.2	EDGE anonymization utility					X	X			
5.3	Google Analytics				X		X			
6	Integration with existing system									
6.1	CSV download function				X	X				
6.2	Automatic Dataset upload						X	X		
6.3	Deployment of EDGE system							X	X	

Table 11 - Work plan for implementation

6.2 Work Plan for Deployment

	Month	1	2	3	4	5	6	7	8	9
1	Server									
1.1	Domain registration	X								
1.2	AWS server setup	X								
1.3	NGINX and dashboard deployment		X							
2	Datasets									
2.1	Sample data		X	X						
2.2	Bangkok BTS				X	X	X			
2.3	La Rambla - Barcelona						X	X		
3.	Dataspaces - deployment of services							X	X	
4	Dissemination and promotion									
4.1	First social media campaign						X	X		
4.2	2nd campaign								X	X

Table 12 - Work plan for deployment

7 CONCLUSION

In D1, we have comprehensively explored the design and functionalities of this project: UtiP-DAM is conceived as a platform dedicated to secure and responsible mobility data anonymization and sharing.

At the core of UtiP-DAM lies data privacy, realized through K-anonymization tools and auditing. These features ensure data anonymization adheres to best practices while allowing users to assess the anonymization strength of datasets. Built upon this foundation is the innovative Mobility Raw Data Marketplace, where anonymized data can be securely shared between data providers and consumers.

Beyond the core functionalities, UtiP-DAM offers additional functionalities for various user segments. Companies can leverage a dedicated K-anonymization tool to anonymize their own data before internal use or sharing. A separate anonymization auditing tool provides an in-depth analysis of anonymization effectiveness within mobility datasets. Secure user accounts with different permission levels ensure authorized access and control within the marketplace. Users can seamlessly purchase anonymized datasets that align with their specific needs, facilitating valuable data-driven insights.

Our approach to developing UtiP-DAM

Correlation Systems prioritizes user-centricity. Rapid prototyping methodologies will be employed throughout the development cycle to deliver core functionalities quickly and gather valuable user feedback. This iterative approach helps ensure that the final product closely aligns with user needs and expectations.

The system architecture leverages a three-layer design for optimal performance. The user interface (Correlation Systems' dashboard for existing customers or marketplace interface for external users) provides user-friendly platforms for interacting with the UtiP-DAM tools.. Internal and external API layers facilitate data management, access control, and communication between the user interface and other components of the system. Finally, secure data storage ensures the safekeeping of anonymized mobility data sets, user information, and other system data.

Conclusion

In conclusion, UtiP-DAM presents a compelling solution for secure and responsible mobility data management. By combining a robust feature set, a user-centric development approach, and a well-structured architecture, UtiP-DAM hopes to empower users and unlock the full potential of anonymized mobility data across diverse sectors.

REFERENCES

Agrawal, Rakesh, et al. "Order preserving encryption for numeric data." Proceedings of the 2004 ACM SIGMOD international conference on Management of data. 2004.

Apple Inc. (n.d.). Differential privacy overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Backes, M., Bohme, R., & Hofman, D. (2008). Anonymity vs. linkage in voter data. In Privacy Enhancing Technologies (pp. 103-117). Springer, Berlin, Heidelberg.

Domingo-Ferrer, J., & Torra, V. (2005). Measuring information leakage in microdata based on information theory concepts. In International conference on knowledge and information systems (pp. 557-562). Springer, Berlin, Heidelberg.

Drineas, P., Harpe, A., & Kannan, R. (2006). Fast exact matrix completion via compressed sensing. In 2006 44th annual ieee symposium on foundations of computer science (focs'06) (pp. 451-460). IEEE.

Dwork, C., & Roth, A. (2014). The algorithmic theory of differential privacy. Journal of Privacy and Confidentiality, 7(4), 1-12.

El Emam, Khaled, and Fida Kamal Dankar. "Protecting privacy using k-anonymity." Journal of the American Medical Informatics Association 15.5 (2008): 627-637.

Fredriksen, M., Wachter, S., & Samek, W. (2014). Auditing algorithms: Measuring and mitigating unintended bias. Communications of the ACM, 57(10), 80-88.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In STOC'09-Proceedings of the 41st annual ACM symposium on theory of computing (pp. 169-178).

Kairouz, P., McMahan, H. B., Avent, B., Belanger, M., Bhadra, S., et al. (2019). Federated learning: Defending against adversarial attacks. In International Conference on Learning Representations (ICLR).

King, R., Liu, L., & Lu, X. (2020). An end-to-end data augmentation approach for improving natural language understanding tasks. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 9062-9070).

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In IEEE 23rd International Conference on Data Engineering (pp. 106–115). IEEE.

Lipton, Z. C., Johnson, T., & Niranjan, A. (2002). Secure learning with overfitting protection. In Esop 2002 (pp. 585-597). Springer, Berlin, Heidelberg.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Reingold, M. (2007). L-diversity: Privacy beyond k-anonymity. ACM Transactions on Information Systems (TOIS), 26(2), 1-29.

Ohm, S. B. (2010). Privacy and security for online transactions: Using pseudonymization. IEEE Security & Privacy, 8(6), 70-79.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-574.

Verykios, Vassilios S., et al. "State-of-the-art in privacy preserving data mining." ACM Sigmod Record 33.1 (2004): 50-57.

Appendix A

Barcelona Deployment Data Sharing Contract

Introduction to LBASense

LBASense is a Correlation Systems's patented technology. Its sensors are designed to passively detect anonymous signals transmitted over the WiFi networks by mobile phones.

Data collected is used in aggregated forms to provide the so-called Crowd Analytics, i.e. insights regarding people's counting and flows in the areas where the system is deployed.

Data Collection

LBASense is capable of sensing the device's WiFi MAC address, and uses it to create aggregated statistical figures based on the amount of such identifiers detected (when not randomized).

LBASense sensors do not collect additional information, thus no collection or processing of "sensitive personal data", nor "personality profile data" is performed.

In fact, Correlation SystemsLtd has designed and implemented a data collection protocol used for collection, encryption, statistical elaboration and aggregation of all data acquired for the scope of project experiments and installations.

Particularly, all data collected are anonymised via unidirectional cryptography; hence, the identification of a person's identity is not possible.

Data Retention

In order to be able to calculate the ratio between locals and tourists, raw data is keeped (and processed) for a duration of one year.
Data that is shared as "open data" is keeped forever. (see the following sections)

Data Accessibility

Data acquired by LBASense are securely stored in Amazon AWS databases, located in Ireland, remapped behind NAT, not accessible from the Internet.

Access to the dashboard does not provide access to any personal information (only aggregated information is presented on the dashboard).

Credentials to access the system (username, password, site Id) are personal and assigned by Correlation Systems administrators to selected people, whose access via dashboard (web and mobile) and APIs is recorded in the system's logs.
Aggregated figures may be disclosed as open data in selected sharing platforms.

Data Sharing and access to raw data

The project includes a “data sharing option” which allows Correlation Systems and associated companies to share the data with 3rd parties including sharing of raw data.

Sharing of aggregated data as described before does not include any personal data.

Prior to sharing of raw data, the encrypted MAC address is encrypted again using a different encryption key for each sensor per day (i.e. the identification number of the same device will be different between sensors and days) in order to prevent association between detection that is now part of the system anonymization process.

Mobility patterns, which include the movement of people between sensors, are created using a new ID for each mobility pattern and by removal of the detection time from the mobility pattern.
As a second step, a k-anonymity is activated with k=10, any mobility patterns that are under 10 are removed.

In addition to the aggregated data, raw detections and raw mobility may be shared as open data on a daily basis.

Appendix B - Barcelona Interview Details

Presentation of the use-case

La Rambla Tourist Association mandated Correlation Systems for the purpose of understanding pedestrian flows along La Rambla by distinguishing between those strolling along its length and those merely crossing the street. Their secondary objective is to ascertain the direction of movement.

For these purposes, Correlation Systems has deployed 12 sensors on La Rambla, La Boqueria market and The Port of Barcelona.

A data-sharing agreement has been established with the business association of La Rambla, enabling Correlation Systems to distribute the data publicly (this is applicable only to La Rambla and La Boqueria sites).

Figure 46 - Barcelona - Port Vell - Insights Map

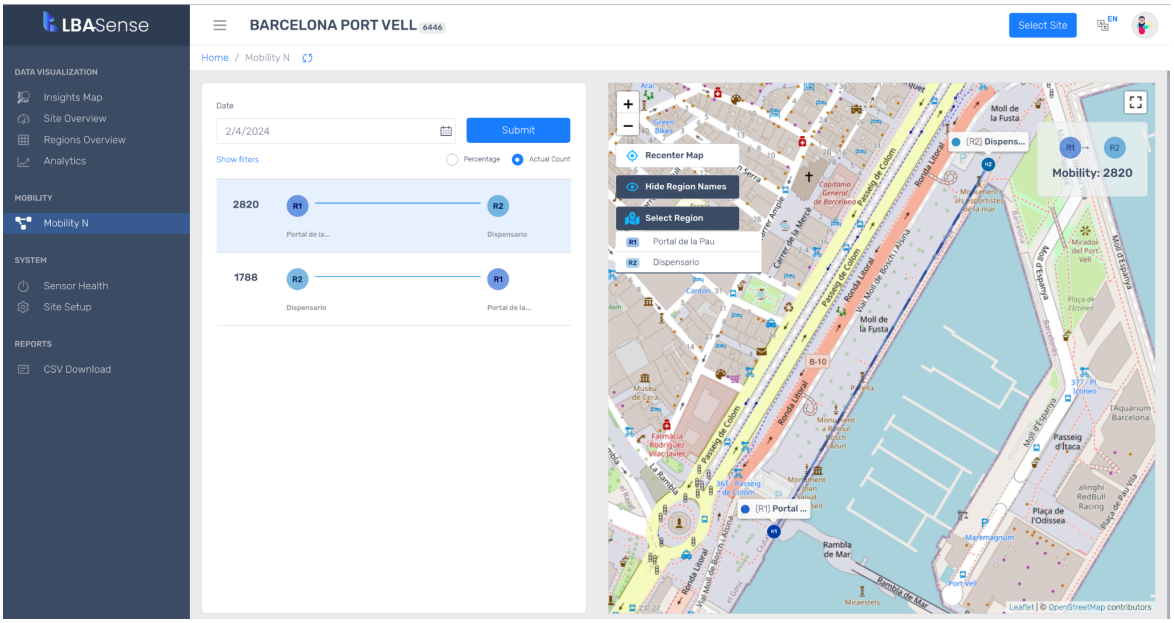

Figure 47 - Barcelona - Port Vell - MobilityN

Interview

During the interview, the Correlation Systems team detailed the purpose of the UtiP-DAM project.

The following decisions were taken:

- The UtiP-DAM tools will be applied internally to anonymize the data collected by the sensors, ensuring that individual trajectories remain private while still providing valuable insights into the mobility patterns along La Rambla.
- The decentralized k-anonymization process aligns with privacy requirements, making it suitable for both internal analysis and public distribution of anonymized mobility data.

Appendix C - MIE Interview Details

Presentation of the use-case

MIE wants to better understand commuter behavior and station usage at the BTS SkyTrain stations in Bangkok. Specifically, by understanding patterns of entry, exit, and dwell times, MIE hopes to optimize station layouts and enhance overall user experience, contributing to the improvement of Bangkok's public transportation experience for its residents and visitors.

For this purpose, MIE has mandated Correlation Systems for the deployment of 10 sensors at the entrance of the BTS SkyTrain station in downtown Bangkok.
Today, this deployment generates approximately 500,000 data points on a daily basis.

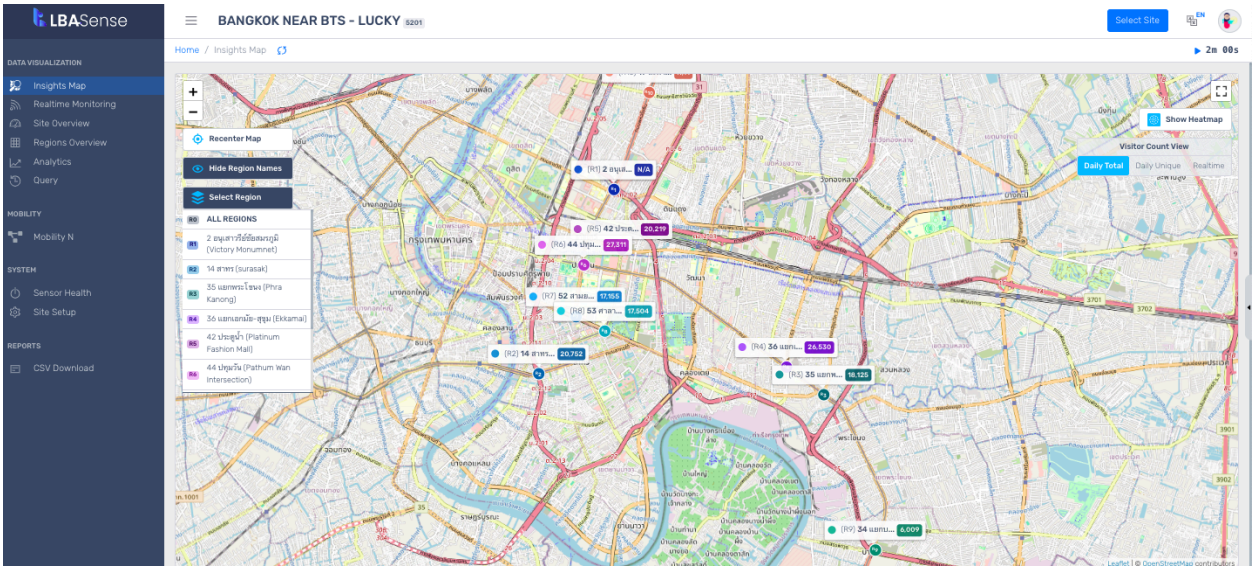

Figure 48 - MIE - Bangkok Near BTS - Insights Map

Figure 49 - MIE - Picture of a sensor deployed near a station entrance

Interview

Together with the MIE team, the following decisions were made:

- UtiP-DAM is set to be implemented for the sensors deployed at the entrances of stations along the Bangkok BTS SkyTrain route in Thailand.
- The data generated by the sensors will be sent to an additional mirror site in order to test the decentralized k-anonymization process (see part 5.1.1.2)
- MIE will provide a report on the usability of the UtiP-DAM marketplace, and will do a short comparative analysis of various public transportation networks, providing that such datasets are made available on the marketplace before the end of the project. The goal will be to generate general recommendations to enhance user experience in public transport.

Appendix C - Bat Yam Mall Interview Details

Presentation of the use-case

This project focuses on analyzing the effectiveness of "anchor stores" within the Bat Yam Mall. Anchor stores are retailers that receive rent reductions due to providing essential services or having brand recognition, attracting customers to the mall. The ideal anchor store also generates significant foot traffic for surrounding retailers within the mall.

The Bat Yam Mall has three anchor stores: a grocery store in the basement, a pharmacy, and a food court. Correlation Systems has deployed 12 sensors within the mall, in order to achieve two key objectives:

1. Quantifying anchor store foot traffic: By analyzing visitor movement patterns, the system determines how many visitors solely visit the anchor stores (particularly the basement grocery store) without exploring other mall areas.
2. Assessing economic impact on other stores: The system analyzes how foot traffic generated by anchor stores translates into visits and potential sales at surrounding retailers.

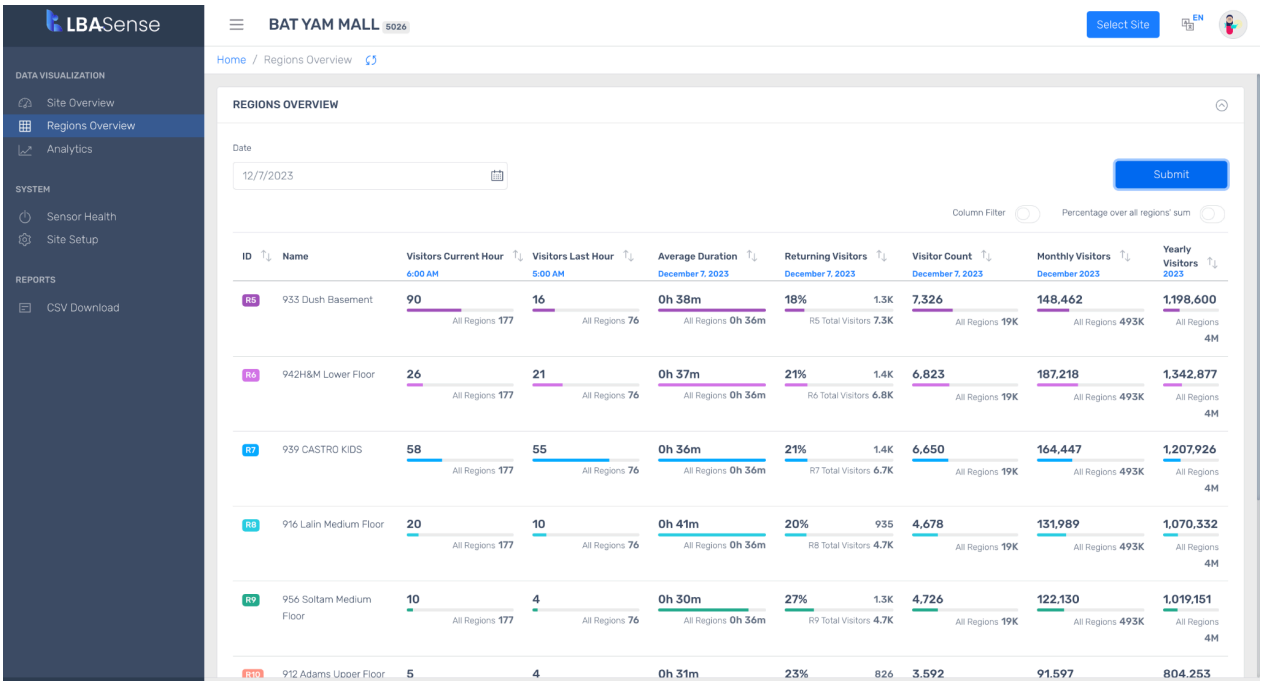

Figure 50 - Bat Yam - Regions Overview

Figure 51 - Bat Yam - Picture of the mall

Interview

Together with the Bat Yam mall team, the following decisions were made:

- UtiP-DAM is set to be implemented for all 12 sensors, which will help the mall comply with Israeli data privacy regulations.
- Bat Yam wishes to use the monetization capabilities of the marketplace. For this, they will utilize the features available on the LBASense dashboard (i.e: a button to push the data to the UtiP-DAM marketplace), and the user account creation feature on the marketplace for managing the published data sets.
- By the end of the project, Bat Yam will provide a report on the usability of the UtiP-DAM marketplace, in the context of data set publishing and monetizing.