PANTHEON

Community-Based Smart City Digital Twin Platform

for Optimised DRM operations and Enhanced Community

Disaster Resilience

D4.1

INTEGRATED DATA MODEL FOR THE SCDT

The project has received funding from the European Union's Horizon Europe programme under Grant Agreement N°101074008.

DOCUMENT INFO

Deliverable Number	D4.1
Work Package Number and Title	WP4: Design and Development of a Smart City Digital Twin for Community DRM
Lead Beneficiary	FINT
Due date of deliverable	31/08/2024 (M20)
Deliverable type[1]	R
Dissemination level[2]	PU
Author(s)	Fanourios Fakoukakis (FINT), Myrto Koletti (FINT), Harisios Zafeiris (FINT), Andreas Tsarouchas (FINT)
Internal reviewer(s)	Mike Karamousadakis (THL), Marc Bonazountas (EPSILON), Kiril Shtefchyk (ISPC)
Version - Status	3.0

TASK ABSTRACT

This deliverable addresses the definition of the characteristics and the development of an integrated data model to be used as part of the Smart City Digital Twin (SCDT). This includes data transfer requirements and communication protocols, formatting issues and data interoperability and storage needs, along with the relevant and necessary preprocessing functions. The necessity and ability of the SCDT to harvest and process big volumes of data from a multiplicity of sources and devices, to support the simulations and the decision making process, constitutes one of its major features. Therefore, the work performed in this task corresponds to all the activities performed towards the development of the SCDT data model and the herein report describes them in detail.

[1] **Please indicate the type of the deliverable using one of the following codes**:
R = Document, report
DEM = Demonstrator, pilot, prototype, plan designs
DEC = Websites, patents filing, press & media actions, videos
DATA = data sets, microdata
DMP = Data Management Plan
ETHICS: Deliverables related to ethics issues.
OTHER: Software, technical diagram, algorithms, models, etc.

[2] **Please indicate the dissemination level using one of the following codes**:
PU = Public
SEN = Sensitive

REVIEW HISTORY

Version	**Date**	**Modifications**	**Editor(s)**
1.0	09/07/2024	ToC	Fanourios Fakoukakis (FINT)
1.1	26/07/2024	First Draft	Fanourios Fakoukakis (FINT)
1.2	02/09/2024	Second Draft	Fanourios Fakoukakis (FINT)
2.0	03/09/2024	Final Draft	Fanourios Fakoukakis (FINT)
2.0	04/9/2024	Final Draft	Marc Bonazountas (EPSILON)
2.0	09/09/2024	Review	Mike Karamousadakis (THL), Kiril Shtefchyk (ISPC)
3.0	12/09/2024	Final version	Fanourios Fakoukakis (FINT)

DISCLAIMER

The document is proprietary of the PANTHEON consortium members. No copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights.

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

TABLE OF CONTENTS

Table OF Contents 3
LIST OF FIGURES 5
LIST OF TABLES 6
LIST OF ABBREVIATIONS 7
EXECUTIVE SUMMARY 8
1. Introduction 9
2.1 SCDT Architecture and General Overview 9
2.2 Data Integration 10
2.3 Technical Specifications and Requirements 11
3. Big Data Harvesting 12
3.1 SCDT Data Types and Sources 12
3.1.1 In-Situ Weather Station Data 12
3.1.2 Aerial Data 12
3.1.3 Satellite Data (Copernicus) 12
3.1.4 Third Party Data 13
3.2 Disaster Scenarios Utilised Data 13
3.2.1 Athens Disaster Scenarios 13
3.2.2 Vienna Disaster Scenarios 13
3.3 SCDT Network Devices and Data Aggregation 14
4. Data Aggregating and Preprocessing Component 16
4.1 General Architecture 16
4.2 Basic Functionalities Overview 16
4.3 Component's User Interface 17
4.4 Formatting, Preprocessing and Normalisation 18
4.4.1 CSV Formatting 18
4.4.2 Image Formatting 18
4.4.3 JSON Structuring 18
4.4.4 Ensuring Data Quality and Consistency 19
4.4.5 Benefits of Standardised Data Formatting 19
4.5 Detailed Preprocessing Workflow 19
4.5.1 In-Situ Weather Data PreProcessing 20

4.5.2 Aerial Data Preprocessing 20
4.5.3 Satellite Data Preprocessing 21
4.5.4 Traffic Data Preprocessing 21
4.5.5 Benefits of Comprehensive Preprocessing 23
5. Data Transfer Communication Protocols 24
5.1 Overview 24
5.2 Rest Protocol for Data Ingestion and Preprocessing 24
5.3 Kafka for Internal Data Transfers and Communication 24
5.3 Security and Performance Considerations 25
6. Data Storage 26
6.1 MinIO Storage Solution 26
6.1.1 Scalability and High Availability 26
6.1.2 Object Storage 26
6.1.3 Security and Access Control 26
6.2 Storage in MinIO 27
6.2.1 Storing Preprocessed Data 27
7. Data Interoperability 28
7.1 Integration of Diverse Data Formats 28
7.1.1 Standardised Data Formats 28
7.1.2 Metadata Standards 28
7.1.3 Interoperability Frameworks 28
7.2 Benefits of Data Interoperability 29
8. Conclusions 30
9. APPENDIX 31

LIST OF FIGURES

Figure 1 - SCDT general concept and architecture. 10
Figure 2 - Overview of the data aggregating and preprocessing procedure. 16
Figure 3 - Login page of the component's user interface. 17
Figure 4 - Data upload page of the preprocessing component. 17
Figure 5 - Example of in-situ weather data. 20
Figure 6 - Production of RGB image from the Red, Green and Blue Bands. 21
Figure 7 - Example of the Copernicus satellite data. 21
Figure 8 - Example of preprocessing of Athens traffic data for a single device. Outliers are removed utilising z-score filtering and missing values are addressed with Exponential Smoothing. 22
Figure 9 - Example of the Athens traffic dataset. The longitude and latitude of each counting point are derived from the location address (road_name) and direction (road_info). 22
Figure 10 - Example of the Vienna traffic dataset. The longitude and latitude of each counting point are derived from the location address (RINAME) and direction (STRNR). 23

LIST OF TABLES

Table 1: List of the main data used and their basic features. 14

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ACL	Access Control Lists
API	Application Programming Interface
CSV	Comma-Separated Value
GDPR	General Data Protection Regulation
HTTPS	HyperText Transfer Protocol Secure
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
ML	Machine Learning
NGSI	Next Generation Service Interfaces
PNG	Portable Network Graphics
REST	Representational State Transfer
SASL	Simple Authentication and Security Layer
SCDT	Smart City Digital Twin
SSL	Secure Sockets Layer
TIFF	Tag Image File Format
TLS	Transport Layer Security
UAV	Unmanned Aerial Vehicles
UI	User Interface

EXECUTIVE SUMMARY

This document constitutes the result of the work performed in T4.1 - "Big data generation and harvesting", which addresses the development of an integrated data model for the Smart City Digital Twin (SCDT) data aggregation. Furthermore, it concerns the development of a SCDT pre-processing component, which is responsible for the initial curation of the SCDT incoming big data. Therefore, the objective of this deliverable is to describe and present the work that was performed towards the aforementioned developments.

PANTHEON's core technology, the SCDT platform, is based on the exploitation of data from a multiplicity of sensors/sources in order to efficiently and accurately represent/simulate the physical environment, phenomena and disaster scenarios into the digital world. These data are mainly derived from four (4) basic sources: i) in-situ sensors, ii) aerial (UAVs), iii) satellite and iv) third party databases. Thus, there is the need for a common process and an integrated model for all this data to be gathered and appropriately pre-processed, before being forwarded to the next stages of the SCDT.

The deliverable commences with an introduction and a high-level overview/description of its sections. Initially, a general overview of the SCDT architecture is provided, followed by a description of the data harvesting process with respect to the specific disaster scenarios. Moreover, a non-trivial part of the document is dedicated to a detailed overview of the preprocessing component. The document continues with the data transfer and storage procedures that are followed during and after the aggregation of the data. Data interoperability and formatting are also discussed, before the final conclusions.

1. INTRODUCTION

Big data are generally defined as large and diverse collections of structured and/or unstructured datasets[3], especially used in Machine Learning (ML) and other analytics techniques to support problem solving and decision making.

The PANTHEON SCDT operation is founded on the effective aggregation and processing of big data to efficiently function as a disaster management preparedness and training tool. In order for all this amount of big data to be gathered and processed, a data model scheme must be developed, along with the proper data aggregation and preprocessing component.

The data model defines all the needed requirements, specifications and protocols that are demanded for all the big data to be transmitted from the respective data sources and subsequently gathered and pre-processed by the corresponding SCDT component.

Furthermore, the data aggregation and preprocessing component constitutes the first component of the SCDT, which, apart from gathering the incoming data from the various sources, is responsible for the initial preprocessing imposed on them, to prepare them for the next stages of the SCDT and especially for the simulations part.

Therefore, the document sections can be summarised as:

- **Section 2:** Short overview of the SCDT concept and architecture that is implemented in the PANTHEON project.
- **Section 3:** Analysis of the data and their harvesting process with respect to the four (4) basic disaster scenarios that are addressed in the project.
- **Section 4:** Detailed description of the data aggregation and preprocessing component, constituting the first part/stage of the SCDT.
- **Section 5:** Data transfer communication protocols, explaining the methodologies and technologies used for the transmission of the data to the preprocessing component.
- **Section 6:** Data storage techniques and repositories, including MinIO.
- **Section 7:** Data interoperability discussion, detailing the data and metadata formatting.
- **Section 8:** Final concluding remarks.

2.1 SCDT ARCHITECTURE AND GENERAL OVERVIEW

Figure 1 depicts the basic schematic diagram of the SCDT architecture, as it was designed and presented in D3.7 - “Overall Architecture and High Level Functionalities”.

[3] https://cloud.google.com/learn/what-is-big-data

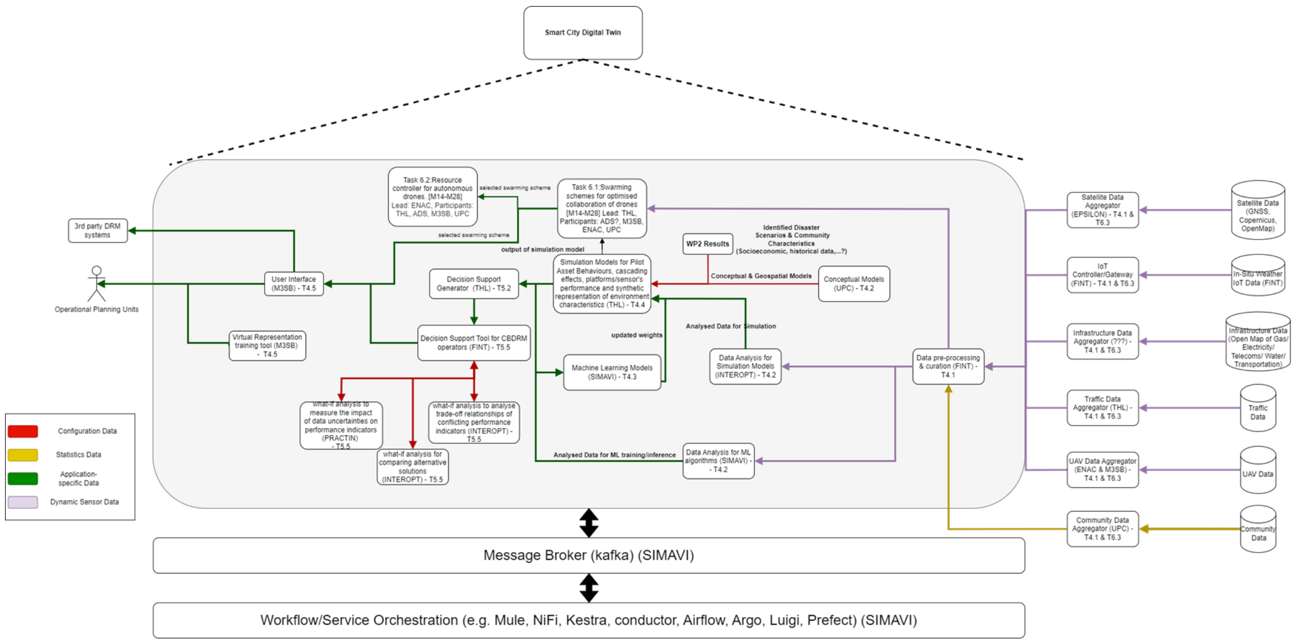


Figure 1 - SCDT general concept and architecture.

As can be seen from the above figure, the various data (created or already existing in the sources) will be gathered/acquired from the responsible partner(s) through a relative aggregator and subsequently sent/aggregated to the SCDT preprocessing component. For the latter to be achieved, a REST API[4] (Application Programming Interface) has been developed, to be used by the partners in order to upload the data to the PANTHEON platform, where the SCDT is hosted and operated. In that way, the SCDT data aggregation is data agnostic, while there are no major restrictions in the dataset file size.

Once received, the data undergo a certain "cleaning" and preprocessing procedure in order to prepare them for the next stages of the SCDT. These procedures will be described in the next sections.

2.2 DATA INTEGRATION

In the PANTHEON project, the process of data integration focuses on the efficient management and storage of various data types collected from the multiplicity of sources/aggregators. Given that the project deals with diverse datasets, including CSV files, JSON records, images etc, the integration process is streamlined to ensure that all data is properly stored organised and accessible within the MinIO[5] object storage system.

The primary goal of data integration in this context is to standardise the storage of these varied data formats in a way that facilitates easy access and retrieval for analysis and application use. Each data type—whether it's CSV files containing in-situ measurements, JSON files representing UAV or satellite data or image files captured by aerial sensors—is uploaded to a specific bucket in the MinIO storage system with its metadata. These metadata are given information according to the data source and type, ensuring that all related data is grouped together and are easily searchable and navigable.

To ensure consistency across the system, the data is catalogued with metadata that describes its origin, type and relevance to the specific disaster scenarios. This metadata is crucial for enabling efficient searches and

[4] https://www.ibm.com/topics/rest-apis#:~:text=A%20REST%20API%20(also%20called,transfer%20(REST)%20architectural%20style.
[5] https://min.io/

retrievals within the MinIO storage, allowing users and/or the other components (or even applications) to quickly find and use the data they need.

Data integration also involves maintaining data quality and consistency. While the raw data is stored in its original format, the component/system employs scripts or middleware to validate data upon ingestion, checking for errors, missing values or format inconsistencies. This step ensures that the data is reliable and ready for downstream applications, such as disaster modelling, planning and response simulations.

Furthermore, the integration process ensures that data from different sources can be accessed simultaneously, providing a unified view of the areas/conditions included in the disaster scenarios. This is particularly important in applications where multiple data types need to be correlated, such as combining UAV imagery with in-situ weather data to analyse the impact of a heatwave, for example.

Overall, the data integration strategy in the PANTHEON project is designed to leverage the capabilities of the MinIO storage system, ensuring that all collected data is securely stored, easily accessible and ready for analysis, thereby supporting specifically the project's disaster scenario simulations and more generally the disaster management objectives.

2.3 TECHNICAL SPECIFICATIONS AND REQUIREMENTS

The component's technical specifications and requirements are centred around the effective aggregation, preprocessing, storage, management and access of diverse data types within the MinIO object storage system. This section outlines the key technical components and configurations necessary to support the project's data integration and overall functionality.

Storage Infrastructure: The project utilises MinIO, a high-performance, distributed object storage system, to handle the various data formats collected from in-situ sensors, UAVs, satellites and third-party sources. MinIO's S3-compatible API enables seamless integration with the project's existing data pipelines, ensuring that all data—whether CSV, JSON or image files—is efficiently stored and managed within a unified platform. The system is configured to ensure high availability and durability of data, with automatic replication across multiple nodes to prevent data loss.

Access and Retrieval: The SCDT employs an API and web interfaces to facilitate access to the stored data. These interfaces are designed to support both automated data retrieval by applications and manual access by users, providing flexibility in how the data is used. The use of standardised data formats like CSV and JSON, combined with the S3-compatible API of MinIO, ensures that data can be easily integrated into various analytical tools and platforms used within the project.

Integration with External Systems: MinIO's compatibility with the S3 API allows for seamless integration with other cloud services and data processing platforms. This capability is essential for the project and the SCDT, as it enables the easy transfer of data between MinIO and external systems used for advanced analytics, machine learning and disaster modelling, along with data and decisions exchange.

Security and Privacy: The whole SCDT infrastructure, along with the data aggregation, integration and processing procedures, are governed by the appropriate security and privacy protocols and directives. No personal data are being used, while fully complying with GDPR regulations.

3. BIG DATA HARVESTING

3.1 SCDT DATA TYPES AND SOURCES

As presented in Section 2, the SCDT system integrates data from various sources to provide comprehensive insights and support smart city operations for dealing with natural disasters. The utilised data mainly fall under four (4) general categories: in-situ, aerial, satellite and 3rd party databases. The six (6) specific sources and data shown in Figure 1 will be used according to the specific disaster scenarios that are taken into account in the PANTHEON project. These specific data use cases are presented in the next subsections and were analysed in detail in D3.6 - Use Case Scenarios.

3.1.1 IN-SITU WEATHER STATION DATA

Meteorological data from local weather stations is crucial for real-time weather monitoring and forecasting. This data includes temperature, humidity, wind speed and precipitation levels and is typically structured in CSV or JSON NGSI format. It must be again mentioned that this data does not contain or involve any personal data, rather than only weather/environmental parameters.

- Sources: Ground-based weather stations in the city.
- Usage: Local environmental data for comprehensive analysis.
- Metadata: Timestamp, location and file type.

3.1.2 AERIAL DATA

UAVs/drones are used to capture high-resolution aerial imagery, providing a bird's-eye view of the disaster phenomenon and of the area under monitoring. This data is particularly valuable for both the disaster preparedness phase (planning, training, early warning) and the post-disaster management/response actions, offering valuable information about urban framework, infrastructure inspection and conditions (pre- and post-disaster) and even human rescue. The imagery is stored in formats such as JPEG, PNG or TIFF. Similarly to the above mentioned data, no personal data will be collected and/or processed, therefore there is no ethical risk in this procedure, also.

- Sources: UAVs equipped with high-resolution cameras and weather parameter sensors.
- Usage: Monitoring construction sites, natural disasters and surveying large areas quickly.
- Metadata: Timestamp, location and file type.

3.1.3 SATELLITE DATA (COPERNICUS)

The Copernicus program provides extensive satellite data of various types. This data is invaluable for a range of applications such as urban planning, environmental monitoring, land cover and risk assessment. Again, no GDPR issues are involved, since no personal data will be collected/processed.

- Sources: Copernicus satellites.
- Usage: Land use monitoring, disaster and environmental impact assessments and emergency response.
- Metadata: Timestamp, location and file type.

3.1.4 THIRD PARTY DATA

A characteristic example of third party data that will be utilised is the traffic data from the two main disaster scenario areas (i.e. Athens and Vienna). Traffic data is essential for monitoring and managing urban mobility effectively. It is collected from roadside sensors deployed throughout the city. This real-time data provides insights into traffic flow, vehicle speed, congestion levels and incidents, crucial for optimising transportation networks and enhancing public safety. The datasets contain the following information: timestamp, location, direction, sensor ID and timestamp, location, direction, sensor ID and number of vehicles. It is important to mention that no personal data is involved (i.e vehicle plate number).

- Sources: Roadside sensors
- Usage: Congestion analysis for route optimization.
- Metadata: Timestamp, location and file type.

3.2 DISASTER SCENARIOS UTILISED DATA

3.2.1 ATHENS DISASTER SCENARIOS

The Athens disaster scenarios include an urban fire in the city (DS-ATH-B) and an earthquake (DS-ATH-A), both majorly affecting the city's urban fabric.

As listed in D3.6, the available/needed technologies and data/sources for the Athens disaster scenarios, can be summarised as:

- Athens earthquake (DS-ATH-A):
  - Aerial (UAV data/imagery)
  - Satellite (satellite imagery/data)
- Athens fire (DS-ATH-B):
  - In-situ (weather station data)
  - Aerial (UAV data/imagery)
  - Satellite (satellite imagery/data)

The above listing means that for the earthquake scenario, utilised in the planning and early warning application (A), only aerial and satellite data/sensors/sources are of importance to be exploited, in order to assist the end-users and civil authorities to prepare, plan and effectively react in an upcoming earthquake. Moreover, for the urban fire scenario, covering the training and exercises application (B), in-situ measurements/data of the area is also needed to enhance the efficient monitoring of the area and to support the effective training process of the first responders and other relevant stakeholders being involved in the disaster management.

Furthermore, third party databases will be also utilised in order to get access to additional data, such as traffic and community data, which can offer valuable information about the area/community that will be affected by the disasters. This community data will be strictly and exclusively statistical, concerning the general population features and trends, and will not contain any personal information whatsoever.

3.2.2 VIENNA DISASTER SCENARIOS

Correspondingly, Vienna disaster scenarios include a heatwave (DS-VIE-A) and a man-made disaster, such as a cyberattack/cyber-terrorism (DS-VIE-B).

- Vienna heatwave (DS-VIE-A):
    - In-situ (weather station data)
    - Satellite (satellite imagery/data)
- Vienna man-made disaster (DS-VIE-B):
    - In-situ (weather station data)
    - Aerial (UAV data/imagery)
    - Satellite (satellite imagery/data)

Similarly to the previous subsection analysis, the above listing suggests that for the heatwave scenario (planning and early warning application) only the in-situ and satellite data/sources are needed, whereas for the man-made disaster (training and exercise application), in-situ, satellite and aerial data can be utilised to support the training activities of the involved stakeholders and especially the first responders and civil protection authorities. Lastly, additional third party data will also be exploited in the Vienna disaster scenario use cases.

3.3 SCDT NETWORK DEVICES AND DATA AGGREGATION

Regarding the devices to acquire and/or send the various datasets to the SCDT, these can be classified as:

- In-situ: weather/environmental parameter IoT-based sensors (wind, temperature, solar radiation, etc)
- Aerial: UAV cameras and weather parameters sensors (e.g. wind)
- Satellite: optical/infrared payload for atmosphere/land monitoring
- Third party: existing databases

With respect to the aggregation of the data to the SCDT, this will also be performed in a variety of ways, which can be summarised as:

- The in-situ environmental measurement data will be firstly sent and stored to the FINoT cloud platform and then sent/uploaded to the PANTHEON platform through the user interface
- The aerial data will be initially stored in local storages and subsequently sent to the PANTHEON platform through the UI
- The satellite data will be downloaded and stored in a Zenodo[6] storage space, from where can be accessed and sent to the PANTHEON platform through the UI
- The data coming from third party databases (traffic, community, etc.) will be initially stored locally and then sent to the PANTHEON platform through the UI

Table 1 lists the main data and their basic features that are being used in the PANTHEON project, in order to obtain a complete view of the areas under disaster.

Table 1: List of the main data used and their basic features.

Provider	Type	Format	Storage	Provision
FINT (IoT)	Measurements	JSON NGSI	Timeseries databases	API
THL (traffic)	measurements	CSV/JSON	files	API

[6] https://zenodo.org/

ENAC (drone)	Measurements, images (position, UAV status, payload data)	Postlog: CSV and custom log files; live: JSON	Log files and images	API
M3SB (community data)	Text file	JSON/GeoJSON	File directory	API
EPSILON (satellite)	files	JSON	File directory	API

4. DATA AGGREGATING AND PREPROCESSING COMPONENT

4.1 GENERAL ARCHITECTURE

This subsection describes the general architecture and the basic functionalities of the data aggregating and preprocessing component, as mentioned in the previous sections of this document. An overview of the procedure involved in the operation of the component is depicted in Figure 2.

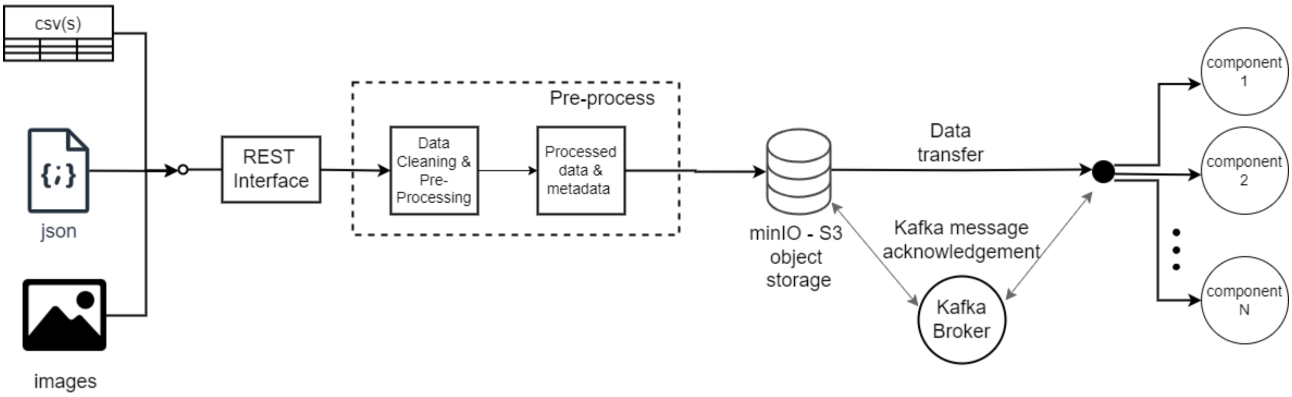


Figure 2 - Overview of the data aggregating and preprocessing procedure.

Initially, the various data are gathered/aggregated from the respective sources. These data include a multiplicity of file formats and/or types (such as CSV and JSON) and/or images, which are gathered through the obtained access to the REST interface, developed as a part of the preprocessing component. The next step is the actual cleaning, curating and preprocessing of the data, in order to prepare them for the next stages of the SCDT. It must be noted that the resulting files also contain the proper metadata, useful for the storage requirements and for the next SCDT components and the data exchange between them. Finally, the data are stored in a S3 compatible MinIO object storage cloud space, from where they can be accessed by any other system component, through a Kafka[7] broker messaging process.

4.2 BASIC FUNCTIONALITIES OVERVIEW

Starting with the REST interface, it is based on OpenAPI[8] (see Appendix) specifications and it enables the users to easily subscribe and upload the data files on the PANTHEON SCDT cloud platform. The uploading is followed by the actual curating and preprocessing, performed by the component. This process includes cleaning of the raw data, geo-referencing and time-stamping them and other procedures, which are described in detail in the next sections. Thereafter, the data files are stored as objects in MinIO buckets. Whenever a system component needs access to the specific data, it can send a message through the Kafka broker, where it will be acknowledged and informed on where the needed data are stored in the MinIO cloud space. As soon as the former procedure is concluded, the system component can directly access the data in the dedicated MinIO bucket.

[7] https://kafka.apache.org/documentation/
[8] https://www.openapis.org/

4.3 COMPONENT'S USER INTERFACE

The component's user interface constitutes the means through which the users can upload the respective data on the PANTHEON platform and be ready to be processed by the preprocessing component. Figure 3 shows the first (initial) login page of the component, with its UI. The process is quite simple and intuitive. The user, by using his assigned credentials (Username and Password), can login to the component/platform and consequently access the upload form page, shown in Figure 4.

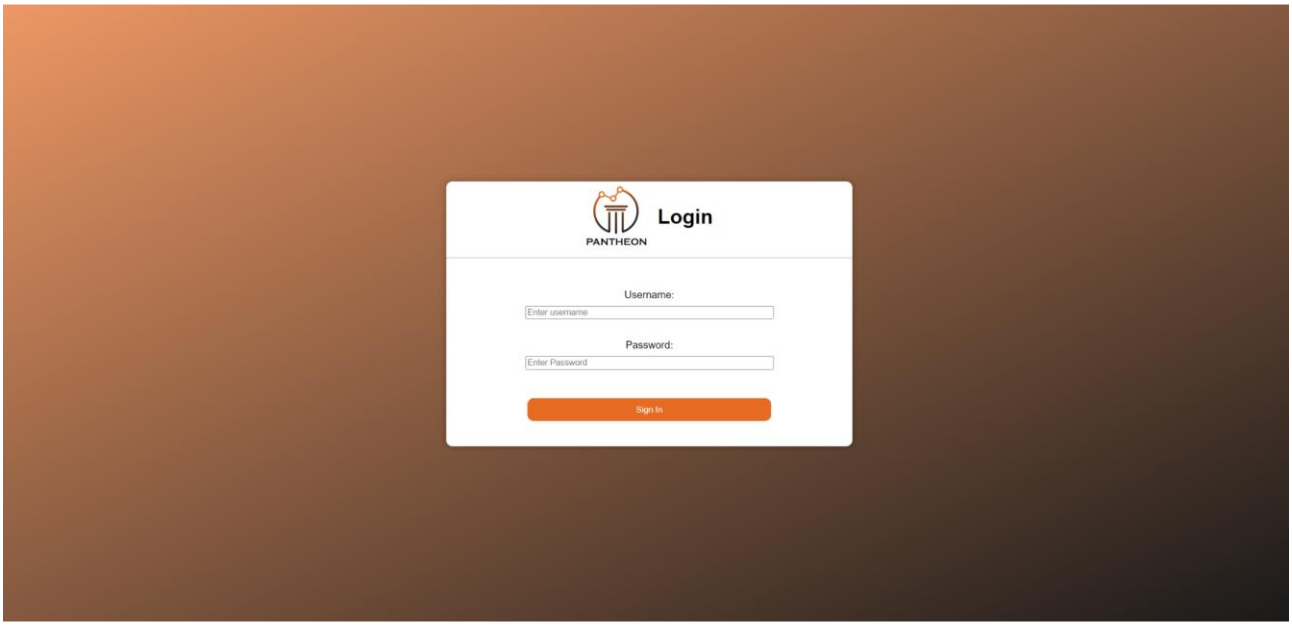


Figure 3 - Login page of the component's user interface.

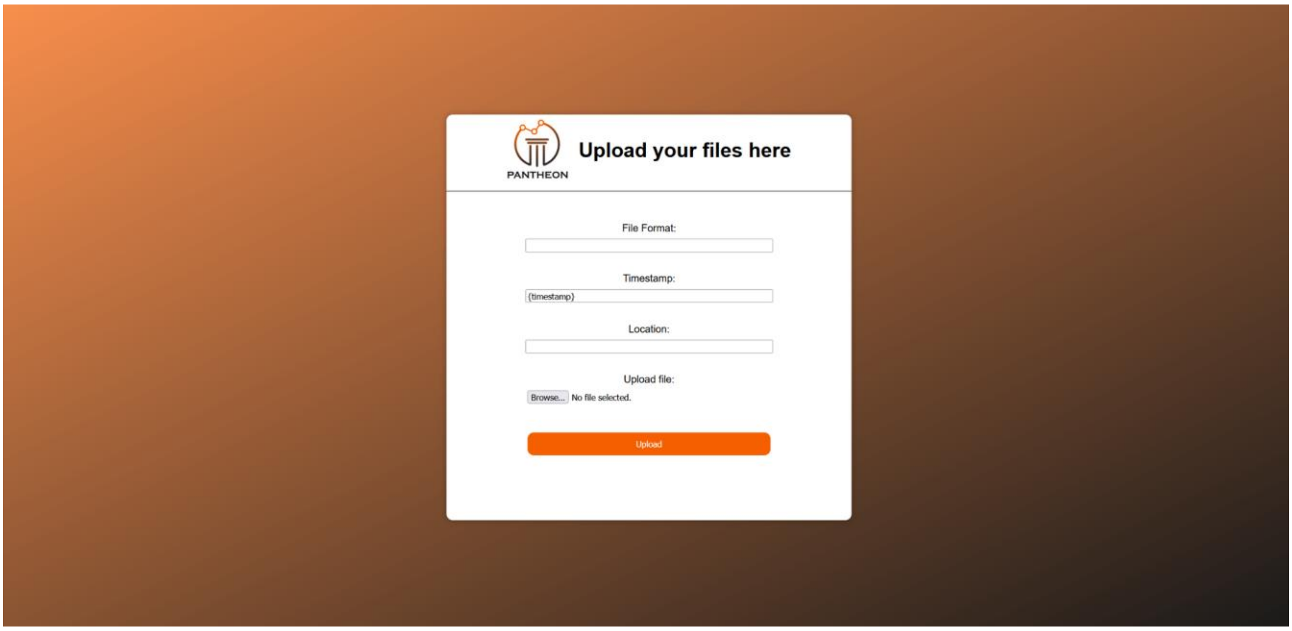


Figure 4 - Data upload page of the preprocessing component.

The form consists of 4 main fields:

- File format
- Timestamp
- Location
- Upload File

Before uploading the data file, the user must fill in some information about it, such as the file format and the timestamp and location of the acquired/measured data. Lastly, the data file must be uploaded.

In that way, an easy-to-use procedure is provided for the partners/users to upload the needed data that will contribute to the simulations of the SCDT.

4.4 FORMATTING, PREPROCESSING AND NORMALISATION

Data formatting is a critical step to ensure that all incoming data adheres to the required standards and formats before being stored in MinIO. Proper formatting not only facilitates seamless integration and interoperability but also enhances the efficiency of data processing, retrieval and analysis.

Preprocessing and normalisation are essential to prepare raw data for storage and subsequent use. This involves cleaning, transforming and standardising data to meet predefined criteria, ensuring consistency and quality. As presented in Section 3, the three (3) basic data formats that are being used are CSV files, Json files and images. The next subsections provide further details on these data formats and their features. Specific examples for each data type are showcased in subsection 4.5.

4.4.1 CSV FORMATTING

Tabular data from traffic sensors and weather stations is formatted into CSV files with consistent column headers and standardised data types. This process involves several steps:

- **Data Cleaning:** Removing any anomalies, duplicates or erroneous entries to ensure data accuracy.
- **Column Standardisation:** Ensuring that column headers are consistent and descriptive, facilitating easier interpretation and analysis.
- **Data Type Standardisation:** Converting all entries to standardised data types (e.g., integers, floats, strings) to maintain uniformity across datasets.
- **Delimiter Uniformity:** Using a consistent delimiter (typically a comma) to separate values, ensuring compatibility with most data processing tools.

By adhering to these standards, the system ensures that CSV data is easily readable, processable and compatible with analytical tools.

4.4.2 IMAGE FORMATTING

Drone imagery is provided in TIFF format. Depending on the specific requirements, conversion to standardised formats such as JPEG or PNG is performed.

4.4.3 JSON STRUCTURING

Structured and semi-structured data, such as metadata and configuration files, are formatted as JSON. JSON (JavaScript Object Notation) is a versatile format that ensures flexibility and ease of integration with other systems. Its basic features include:

- **Schema Definition:** Defining a clear and consistent schema for JSON data to ensure uniformity across different datasets.
- **Data Validation:** Implementing validation mechanisms to check the correctness and completeness of JSON data against the defined schema.
- **Hierarchical Structuring:** Utilising JSON's hierarchical structure to represent complex data relationships and nested information.

By structuring data in JSON format, the system ensures that it is easily exchangeable, interpretable and integrable with various applications and platforms.

4.4.4 ENSURING DATA QUALITY AND CONSISTENCY

The PANTHEON SCDT system places a strong emphasis on ensuring data quality and consistency throughout the formatting process. This is achieved through:

- **Automated Preprocessing Pipelines:** Utilising automated pipelines to handle data cleaning, transformation and formatting, reducing manual errors and ensuring consistency.
- **Validation and Verification:** Implementing rigorous validation and verification steps to ensure that formatted data meets the required standards and is free from errors.
- **Standard Operating Procedures (SOPs):** Establishing SOPs for data formatting to guide the process and ensure adherence to best practices.
- **Continuous Monitoring:** Continuously monitoring the data formatting process to identify and address any issues promptly, ensuring ongoing data quality.

4.4.5 BENEFITS OF STANDARDISED DATA FORMATTING

Standardised data formatting provides several key benefits for the PANTHEON SCDT system:

- **Improved Data Integration:** Consistent data formats facilitate seamless integration of diverse data sources, enhancing the system's ability to provide comprehensive insights.
- **Enhanced Data Usability:** Standardised data is easier to process, analyse and visualise, supporting more effective decision-making and operational efficiency.
- **Reduced Processing Time:** Automation and standardisation streamline the data preparation process, reducing the time and resources required for data handling.
- **Increased Interoperability:** Standardised formats and metadata ensure that data can be easily exchanged and utilised across different platforms and applications, promoting greater interoperability.

4.5 DETAILED PREPROCESSING WORKFLOW

Data preprocessing is a critical stage in the data management workflow of the PANTHEON SCDT system. It involves cleaning, normalising and transforming raw data to ensure that it meets the necessary data quality standards for the Pantheon project before being stored in MinIO. This section details the comprehensive preprocessing workflow implemented for various data types, ensuring data integrity, consistency and usability.

The preprocessing workflow is tailored to the specific requirements of different data sources, encompassing a series of steps designed to enhance data quality and prepare it for effective storage and analysis.

4.5.1 IN-SITU WEATHER DATA PREPROCESSING

In-situ weather data is collected from ground-based weather stations, providing real-time meteorological/environmental information and monitoring. An example of the data collected is shown in Figure 5. This data is complementing each other with the related satellite ones. The preprocessing steps for weather data include:

- **Data Cleaning:** Erroneous or missing data points are identified and corrected or removed. This involves checking for outliers, filling gaps with interpolated values and ensuring data consistency.
- **Normalisation:** Weather data is standardised to uniform units of measurement (e.g., temperature in Celsius, wind speed in metres per second etc) and consistent formats, facilitating seamless integration and analysis.
- **Temporal Aggregation:** Weather data is aggregated hourly to support various analytical needs. This procedure helps in capturing significant trends and patterns, such as temporal temperature variations and precipitation levels.

Timestamp	Air Relative Humidity (%) (2023)	Air Temperature (°C)	Average Wind Speed (m/s)	Daily Precipitation (mm)	Gust Wind Speed (m/s)	Precipitation Rate (mm/h)	Solar Radiation (W/m²)	Wind Direction (°)
2024-07-15T09:00:00+03:00	34.52	28.59	19.69	0	37.59	0.8	251.93	232
2024-07-15T10:00:00+03:00	31.81	29.12	24.6	0	40.08	0	708.5	87
2024-07-15T11:00:00+03:00	28.26	29.87	22.86	0	42.66	0	861.08	22

Figure 5 - Example of in-situ weather data.

4.5.2 AERIAL DATA PREPROCESSING

Aerial (UAV) data consists mainly of high-resolution aerial images captured by UAVs, which are valuable for urban planning, infrastructure monitoring and disaster response. The preprocessing workflow for drone data includes:

- **Georeferencing**: Each drone image is associated with precise geographic coordinates included in the relevant metadata. This georeferencing ensures that the spatial context of the data is preserved, enabling accurate mapping and analysis.
- **Image Enhancement**: Techniques such as contrast adjustment, noise reduction and sharpening can be applied to improve the quality and clarity of drone images. Enhanced images provide better visibility of features and details, which is crucial for detailed analysis.
- **RGB Image Extraction**: Extracting the RGB channels is important for obtaining a true-colour image, which is essential for realistic visualisation and precise interpretation. The process is shown schematically in Figure 6.

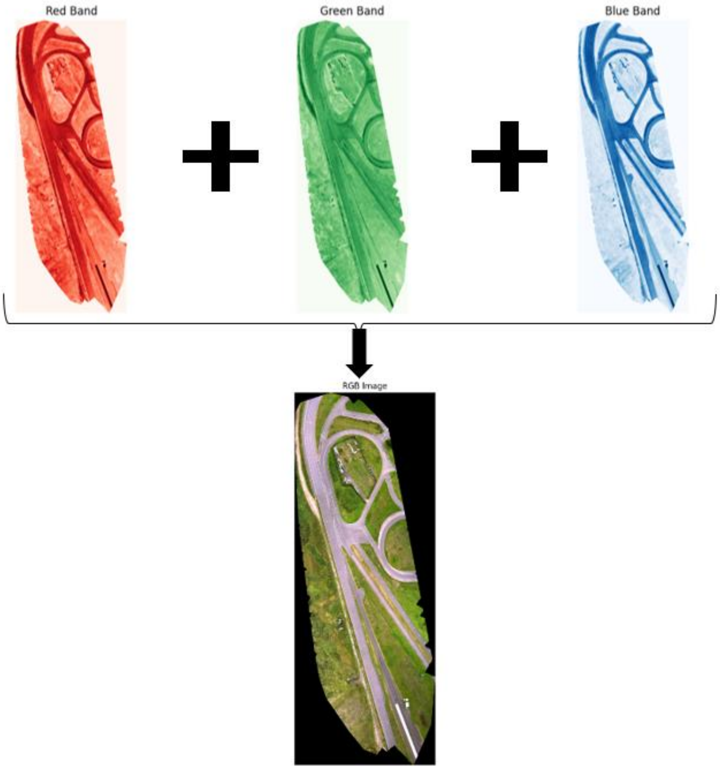


Figure 6 - Production of RGB image from the Red, Green and Blue Bands.

4.5.3 SATELLITE DATA PREPROCESSING

Satellite data from the Copernicus program provides extensive environmental monitoring information. An example of the data collected is shown in Figure 7. The preprocessing of satellite data involves:

- **Temporal Resolution Adjustment:** Satellite data are standardised to an hourly temporal resolution, facilitating integration with the local weather station data.
- **Data Cleaning:** Outliers and missing values are replaced with interpolated values, ensuring data consistency.
- **Normalisation:** Data is standardised to uniform units of measurement and consistent formats, facilitating seamless integration with the in-situ weather data.

timestamp	surface	latitude	longitude	airTemperature(°C)	solarRadiation(W/m^2)
2024-07-15T09:00:00+03:00	0.0	37.9	23.6	28.12	2.546.687
2024-07-15T10:00:00+03:00	0.0	37.9	23.6	28.49	4.277.256
2024-07-15T11:00:00+03:00	0.0	37.9	23.6	29.89	5.860.158

Figure 7 - Example of the Copernicus satellite data.

4.5.4 TRAFFIC DATA PREPROCESSING

Traffic data is collected from various sensors and cameras (not including any personal data and/or vehicle plate numbers) deployed across the city, providing valuable insights into urban mobility patterns. The preprocessing of this data involves several key steps:

- **Noise Reduction:** Traffic data often contains outliers due to sensor errors or environmental factors. Statistical techniques, such as z-score filtering, can be employed to remove these anomalies, ensuring the data's accuracy and reliability. Figure 8 shows an example of outlier removal from the Athens dataset.

- **Data Cleaning:** Data can often contain missing data. Such gaps are replaced with interpolated values in order to ensure data consistency. For larger gaps, more sophisticated techniques are utilised, such as Exponential Smoothing (see Figure 8).
- **Standardisation:** To maintain consistency, traffic data is standardised in terms of format and structure. This includes converting all measurements to uniform units, aligning timestamps to a common format and ensuring consistent data types across datasets.
- **Aggregation:** To manage the volume of data and highlight key metrics, raw traffic data is aggregated hourly. This summarisation process helps in reducing data size, while preserving essential information such as average traffic flow, peak congestion times and incident counts.
- **Coordinates retrieval:** Using an algorithmic method, in order to extract the coordinates (latitude and longitude) of the counting point from the location address and direction. This procedure allows for the precise alignment of the counting point with its corresponding location on a map. The process is shown schematically in Figure 9 and Figure 10.

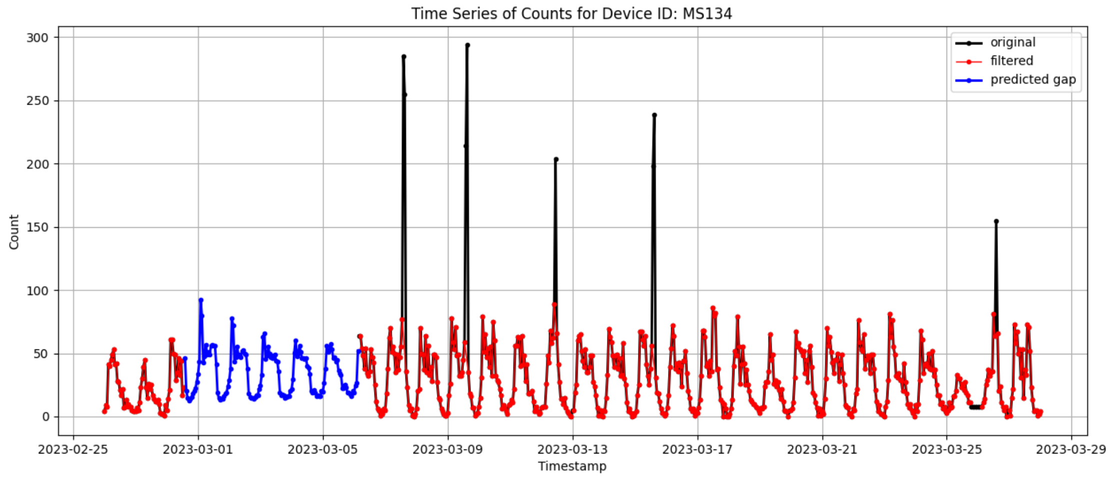


Figure 8 - Example of preprocessing of Athens traffic data for a single device. Outliers are removed utilising z-score filtering and missing values are addressed with Exponential Smoothing.

appprocesstime	average_speed	countedcars	deviceid	road_info	road_name	lon	lat
2023-02-26T00:00:00	37.97560976	136	MS747	ΚΥΡΙΟΣ ΔΡΟΜΟΣ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΠΡΟΣ ΑΚΤΗ ΜΙΑΟΥΛΗ, 120 Μ. ΠΡΙΝ ΑΠΟ ΤΗΝ ΓΟΥΝΑΡΗ	ΑΚΤΗ ΠΟΣΕΙΔΩΝΟΣ	23.64241	37.94629
2023-02-26T00:00:00	34.58536585	124	MS748	ΚΥΡΙΟΣ ΔΡΟΜΟΣ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΠΡΟΣ ΛΙΜΑΝΙ ΠΕΙΡΑΙΑ, 150 Μ. ΠΡΙΝ ΑΠΟ ΤΗΝ ΑΚΤΗ ΠΟΣΕΙΔΩΝΟΣ	ΓΟΥΝΑΡΗ ΔΗΜ.	23.6437	37.94605
2023-02-26T00:00:00	34.7804878	174	MS749	ΚΥΡΙΟΣ ΔΡΟΜΟΣ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΠΡΟΣ ΑΚΤΗ ΚΟΝΔΥΛΗ, 50 Μ. ΠΡΙΝ ΑΠΟ ΤΗΝ ΓΟΥΝΑΡΗ.\n27.09.20:	ΑΚΤΗ ΠΟΣΕΙΔΩΝΟΣ		
2023-02-26T00:00:00	69.82926829	287	MS757	ΚΥΡΙΟΣ ΔΡΟΜΟΣ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΠΡΟΣ ΑΚΤΗ ΠΟΣΕΙΔΩΝΟΣ, 110Μ. ΠΡΙΝ ΤΗΝ ΒΑΣ. ΓΕΩΡΓΙΟΥ Α'	ΑΚΤΗ ΜΙΑΟΥΛΗ	23.64403	37.94352
2023-02-26T00:00:00	65.65853659	375	MS767	ΚΥΡΙΟΣ ΔΡΟΜΟΣ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΠΡΟΣ ΠΕΙΡΑΙΩΣ 150 Μ. ΠΡΙΝ ΑΠΟ ΤΗΝ ΚΟΙΛΗΣ	ΚΑΛΛΙΡΡΟΗΣ	23.71211	37.96384

Figure 9 - Example of the Athens traffic dataset. The longitude and latitude of each counting point are derived from the location address (road_name) and direction (road_info).

JAHR	MONAT	ZNR	ZNAME	STRTYP	STRNR	RINAME	FZTYP	DTVMS	DTVMF	DTVMO	DTVDD	DTVFR	DTVSA	DTVSF	TVMAX	TVMAXT
2024	MÄRZ	1075	Reichsbrüc	B	8	Leopoldstadt	Kfz	16758	18451	17714	18884	18000	13951	12453	20269	*Di
2024	MÄRZ	1075	Reichsbrüc	B	8	Leopoldstadt	LkwÄ	640	768	763	796	704	389	354	877	*Mi
2024	MÄRZ	1075	Reichsbrüc	B	8	Donaustadt	Kfz	14099	15042	14410	15089	15435	12991	11246	16824	Fr

Latitude	Longitude
48.2305	16.41391
48.2305	16.41391
48.2305	16.41391

Figure 10 - Example of the Vienna traffic dataset. The longitude and latitude of each counting point are derived from the location address (RINAME) and direction (STRNR).

4.5.5 BENEFITS OF COMPREHENSIVE PREPROCESSING

Implementing a detailed and rigorous preprocessing workflow offers several benefits for the PANTHEON SCDT system.

- **Enhanced Data Quality:** By removing errors, standardising formats and improving clarity, the preprocessing steps ensure that the data is accurate, reliable and ready for analysis.
- **Consistency Across Datasets:** Standardisation and normalisation processes ensure that data from different sources can be seamlessly integrated and compared, facilitating comprehensive analysis and decision-making.
- **Improved Analytical Value:** Pre-processed data is more suitable for advanced analytics, machine learning and visualisation. Enhanced data quality and consistency enable more accurate and insightful analyses, supporting better decision-making and operational efficiency.

5. DATA TRANSFER COMMUNICATION PROTOCOLS

5.1 OVERVIEW

This section outlines the data transfer communication protocols employed in the PANTHEON project, for the purpose of data aggregation to the SCDT. The system utilises REST[9] (Representational State Transfer) architecture for data ingestion and preprocessing, while Kafka is used for internal data transfers and communication between the different system components. These protocols ensure efficient, secure and reliable data flow throughout the system, facilitating disaster management operations.

5.2 REST PROTOCOL FOR DATA INGESTION AND PREPROCESSING

In the PANTHEON project, REST protocol plays a crucial role in both data ingestion and preprocessing before the data is stored in the MinIO object storage system.

- **Data Ingestion:** External data sources, including in-situ sensors, UAVs, satellites and third-party APIs, provide data that is ingested into the system through RESTful API. This API supports various data formats, such as JSON, CSV and image files. Once ingested, this data is immediately passed to the preprocessing middleware.
- **Preprocessing:** Before the data is stored in MinIO, it undergoes preprocessing. The preprocessing tasks might include data validation, cleaning, transformation and enrichment, ensuring that the data is in a consistent and usable format. This step is critical for preparing the data for subsequent storage and analysis.
- **Security and Reliability:** The REST communication is secured using HTTPS, providing encryption for data transfers to protect against unauthorised access. The stateless nature of REST enhances system reliability, ensuring that each request is processed independently, which contributes to the system's scalability.

5.3 KAFKA FOR INTERNAL DATA TRANSFERS AND COMMUNICATION

After preprocessing, Kafka is used within the SCDT to manage internal data transfers and communication between the various components.

- **Decoupling System Components:** Kafka's publish-subscribe architecture allows different parts of the system to communicate without needing direct interaction. For instance, preprocessing services can publish data to Kafka topics, while storage services, analytics tools or other consumers subscribe to these topics to receive data as needed. This decoupling improves the system's resilience and scalability.
- **Data Reliability and Ordering:** Kafka ensures that data records (messages) are reliably delivered and maintain their order, which is essential for processing sequences of events in disaster scenarios.

[9] https://www.techtarget.com/searchapparchitecture/definition/REST-REpresentational-State-Transfer

- **Integration with MinIO Storage:** After data is processed through Kafka, it is eventually stored in MinIO. This integration allows the system to efficiently handle data storage and retrieval, ensuring that processed data is readily available for analysis or further use.

5.3 SECURITY AND PERFORMANCE CONSIDERATIONS

- **Security:** Both REST and Kafka are configured to ensure the security of data transfers. RESTful APIs are secured with HTTPS for encrypted communication, while Kafka can be protected using SSL/TLS for encryption, SASL for authentication and ACLs (Access Control Lists) for managing access permissions. These measures ensure that all data remains secure throughout its journey in the system.
- **Performance Optimization:** Kafka's ability to partition topics and handle parallel processing boosts system performance, allowing it to manage large data volumes efficiently. RESTful services are designed for high availability and fast response times, ensuring smooth data ingestion and preprocessing without creating bottlenecks.

6. DATA STORAGE

6.1 MINIO STORAGE SOLUTION

The PANTHEON SCDT system leverages MinIO, a high-performance, distributed object storage system, to manage its diverse data sets.

6.1.1 SCALABILITY AND HIGH AVAILABILITY

MinIO's architecture is designed to support horizontal scalability and high availability, making it an ideal choice for managing the increasing volumes of data typical of smart city applications. The system can seamlessly expand by adding more nodes, ensuring that performance remains consistent even as data grows. This horizontal scalability allows the PANTHEON SCDT system to handle large datasets efficiently, providing the flexibility needed to accommodate various types of data from different sources.

High availability is another cornerstone of MinIO's architecture. By distributing data across multiple nodes, MinIO ensures that there is no single point of failure. This redundancy guarantees that data remains accessible even in the event of hardware failures or network issues. For real-time applications and decision-making processes within the PANTHEON SCDT system, this reliability is crucial, as it ensures that data is always available when needed.

6.1.2 OBJECT STORAGE

In the PANTHEON SCDT system, data is stored as objects in the MinIO storage solution. This method offers significant advantages over traditional file and block storage by encapsulating the data along with its metadata and a unique identifier in a single object. Object storage, as implemented in MinIO, is particularly beneficial for handling a wide variety of data types, including CSV, TIFF, JPEG, PNG and JSON.

One of the primary strengths of MinIO's object storage is its scalability. As the PANTHEON SCDT system collects more data, MinIO can expand horizontally to accommodate this growth without compromising performance. This scalability is seamlessly integrated with the storage system's ability to manage diverse data types, making it flexible and adaptable to the needs of a smart city.

Moreover, the metadata management capabilities of MinIO's object storage are robust. The name of each object contains custom metadata, in order to enhance data retrieval and management processes. This feature is critical for the PANTHEON SCDT system, where efficient data retrieval and accurate searches across large datasets are essential for effective data utilisation.

High availability is inherently designed into MinIO's architecture, ensuring data durability and accessibility. By replicating data across multiple nodes, MinIO provides a reliable storage solution that guarantees data availability even in the face of hardware or network failures. This reliability is vital for maintaining the continuous operation of smart city applications and services.

6.1.3 SECURITY AND ACCESS CONTROL

MinIO applies automatic encryption to the data, before it is stored on disk, at the REST level. This method guarantees that all the data will be encrypted before stored. Thus, this approach of baseline layer security assures the integrity, privacy and authenticity of the data at the REST level.

Furthermore, MinIO implements the best practices regarding S3 security and access control. It employs the least privilege access principle, allowing to control which applications can access the S3 resources on MinIO. This access is under control by IAM[10] policies, thus directly dependent on the application credentials.

6.2 STORAGE IN MINIO

MinIO's object storage system is designed to handle the diverse data requirements of the PANTHEON SCDT system efficiently. Each type of data, be it traffic data, drone imagery, satellite data or weather information, is processed and stored as an object with comprehensive metadata embedded in their filename. This approach ensures that data is not only stored securely but is also easily retrievable and usable for various analytical and operational needs.

6.2.1 STORING PREPROCESSED DATA

Before data is uploaded to MinIO, preprocessing is applied to ensure that they are clean and in standardised format, consistent and ready for analysis.

- Data Upload: Pre-processed data files are transferred to MinIO buckets using secure and efficient protocols.
- Metadata Attachment: The name of each file includes metadata information to facilitate easy identification, search and retrieval.
- Bucket Organization: Data is organised in MinIO buckets based on criteria such as data type, source and collection date, optimising storage management and access.

By implementing this structured approach to data storage, the PANTHEON SCDT system ensures that it can effectively manage vast amounts of diverse data, supporting the project's overarching goals of enhancing effective disaster management, preparedness and response.

[10] https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html

7. DATA INTEROPERABILITY

Ensuring data interoperability is crucial for the PANTHEON SCDT, to integrate and utilise data from different sources effectively. Interoperability enables seamless data exchange and usage across various systems and platforms, facilitating comprehensive data analysis and informed decision-making in smart city operations.

7.1 INTEGRATION OF DIVERSE DATA FORMATS

The integration of diverse data formats is essential to harness the full potential of the data collected from multiple sources. The PANTHEON SCDT system implements several strategies to achieve this, which are analysed in the following sub-sections.

7.1.1 STANDARDISED DATA FORMATS

To ensure seamless data integration, standardised formats that are compatible with the MinIO storage system are required by the SCDT. These formats are CSV, JSON, TIF, JPEG and PNG. This ensures that the system can efficiently store, retrieve and process data, minimising compatibility issues and enhancing data interoperability. The two basic subjects that are associated with the standardisation of the data formats are:

- **Data Conversion:** This process involves validating the data structure, ensuring data integrity and applying necessary transformations.
- **Compatibility:** The standardised formats are chosen to ensure compatibility with MinIO's object storage capabilities, facilitating efficient storage and retrieval operations.

7.1.2 METADATA STANDARDS

Unified metadata standards are employed to describe the data, making it easier to search, retrieve and utilise across different modules and applications of the PANTHEON SCDT platform. Metadata provides essential information about the data, such as its format, timestamp and location.

- **Tagging and Indexing:** Each data object is tagged with metadata upon ingestion into the MinIO storage system. This tagging facilitates efficient indexing and retrieval, allowing users to quickly locate and access the required data.
- **Interoperability:** The use of metadata ensures that data can be easily understood and utilised by various applications and systems within the PANTHEON SCDT ecosystem.

7.1.3 INTEROPERABILITY FRAMEWORKS

The PANTHEON SCDT system employs interoperability frameworks that support data exchange between heterogeneous systems. These frameworks ensure that data can be shared and used across different platforms without compatibility issues, enabling comprehensive analysis and informed decision-making.

API Integration: The system provides APIs (Application Programming Interfaces) that allow external systems and applications to interact with the MinIO storage. These APIs support standard protocols such as REST, ensuring seamless data exchange.

The system uses data formatted with the required formats and structures for different applications. This process ensures that data exchanged between systems/components remains consistent and accurate.

7.2 BENEFITS OF DATA INTEROPERABILITY

The implementation of data interoperability within the PANTHEON SCDT system offers several key benefits. Firstly, by enabling seamless data exchange and integration, the system ensures that all available data can be utilised to its full potential, supporting comprehensive analysis, simulations and decision-making. This interoperability also facilitates collaboration between parts of the system, external partners and stakeholders, enhancing the overall effectiveness and disaster management efficiency. Additionally, the ability to integrate diverse data formats and sources ensures that the system can scale up and adapt to evolving data needs and new technologies, maintaining its relevance and utility over time. Furthermore, standardised data formats and metadata, along with robust interoperability frameworks streamline management processes, reduce manual intervention and minimise errors, thereby improving operational efficiency.

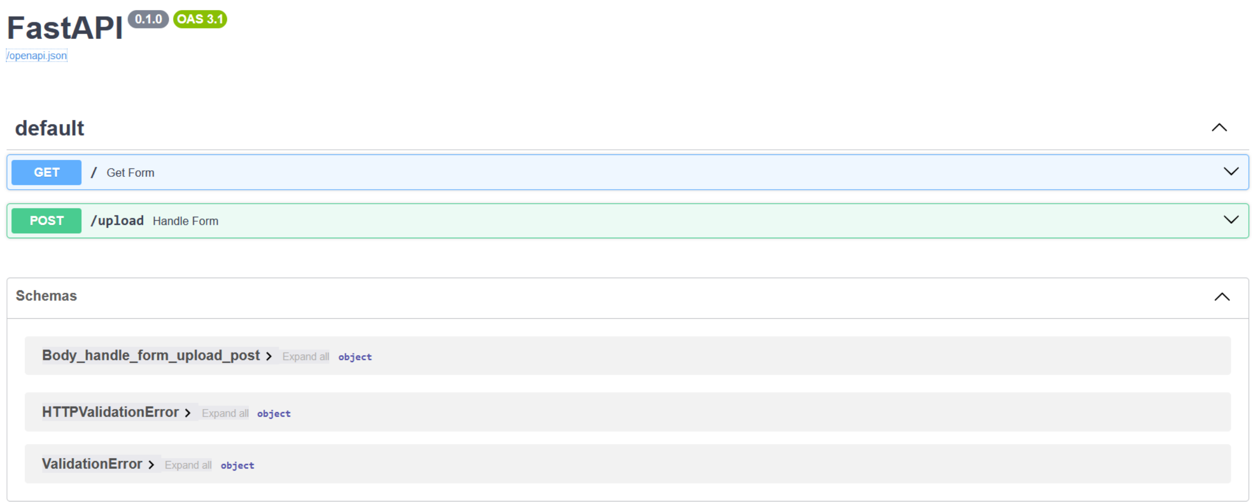
8. CONCLUSIONS

This deliverable provided an overview of the work that has been done with respect to the implementation of SCDT's integrated data model and the development of the preprocessing component, along with all the necessary information. The report includes a short overview of the SCDT architecture and a more detailed description of the preprocessing component itself. The component's UI is presented and the means with which the various datasets will be uploaded to the PANTHEON platform for further actions. The expected/utilised data types are described, along with their formatting, processing and storage in the MinIo storage space of the PANTHEON platform. Lastly, issues about data integration and interoperability are also discussed and addressed.

9. APPENDIX

OpenAPI – FastAPI

End points:



File:

{"openapi":"3.1.0","info":{"title":"FastAPI","version":"0.1.0"},"paths":{"/":{"get":{"summary":"Get Form","operationId":"get_form__get","responses":{"200":{"description":"Successful Response","content":{"text/html":{"schema":{"type":"string"}}}}}}},"/upload":{"post":{"summary":"Handle Form","operationId":"handle_form_upload_post","requestBody":{"content":{"multipart/form-data":{"schema":{"$ref":"#/components/schemas/Body_handle_form_upload_post"}}},"required":true},"responses":{"200":{"description":"Successful Response","content":{"text/html":{"schema":{"type":"string"}}}},"422":{"description":"Validation Error","content":{"application/json":{"schema":{"$ref":"#/components/schemas/HTTPValidationError"}}}}}}}},"components":{"schemas":{"Body_handle_form_upload_post":{"properties":{"file":{"type":"string","format":"binary","title":"File"},"file_format":{"type":"string","title":"File Format"},"timestamp":{"type":"string","title":"Timestamp"},"location":{"type":"string","title":"Location"}},"type":"object","required":["file","file_format","timestamp","location"],"title":"Body_handle_form_upload_post"},"HTTPValidationError":{"properties":{"detail":{"items":{"$ref":"#/components/schemas/ValidationError"},"type":"array","title":"Detail"}},"type":"object","title":"HTTPValidationError"},"ValidationError":{"properties":{"loc":{"items":{"anyOf":[{"type":"string"},{"type":"integer"}]},"type":"array","title":"Location"},"msg":{"type":"string","title":"Message"},"type":{"type":"string","title":"Error Type"}},"type":"object","required":["loc","msg","type"],"title":"ValidationError"}}}}