.

PANTHEON

Community-Based Smart City Digital Twin Platform for optimised DRM operations and Enhanced Community Disaster Resilience

D3.4

DATA DELIVERY SCHEME FOR COMMUNITY BASED DISASTER RESILIENT MANAGEMENT (CBDRM)

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

DOCUMENT INFO

Deliverable Number	D3.4
Work Package Number and Title	WP3: PANTHEON Requirements, Participatory Design Process & Pilot Use Cases Specifications
Lead Beneficiary	THL
Due date of deliverable	31/12/2023
Deliverable type[1]	Report
Dissemination level[2]	Public
Author(s)	Marc Bonazountas, Dimitris Petridis, Mike Karamousadakis, Cristina Barrado, John Tsaloukidis, Thanos Kyritsis, Jim Sharples
Internal reviewer(s)	THL, Simon Bittner (JOAFG), Constanze Geyer (JOAFG), SIMAVI
Version - Status	1.0 – Final version

TASK ABSTRACT

The objective of the PANTHEON Data Delivery Scheme for CBRM (Community-Based Risk Management) is six-fold: (1) it aims to delineate the diverse types of data and their respective sources that will be utilized throughout the project's duration, (2) it seeks to provide comprehensive descriptions of the data's attributes, encompassing aspects such as quality, security, privacy, availability, and integrity in detail, (3) the scheme is designed to concentrate on delineating the most appropriate methodologies for acquiring and managing the data effectively, (4) it aims to elucidate the various types of data processing methods, tools, frameworks, and methodologies that align optimally with the project's objectives and requirements, (5) it is geared towards outlining the recommended approaches for conducting data analysis, ensuring that it is conducted rigorously and in accordance with the project's objectives, (6) the scheme will expound upon the most suitable data delivery schemes tailored specifically to the two selected pilot cases, namely Athens and Vienna.

[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)**
0.1	10/09/2023	First Populated Draft	Marc Bonazountas (EPSILON)
0.2	02/11/2023	Second draft on new TOC	Dimitris Petridis (EPSILON)
0.3	16/11/2023	Third & Agreed Draft on TOC	Dimitris Petridis
0.4	07/12/2023	Fourth Populated Draft on Agreed TOC	Dimitris Petridis (EPSILON) Cristina Barrado (UPC) John Tsaloukidis (KEMEA) Mike Karamousadakis (THL) Thanos Kyritsis (INTEROPT)
0.5	23/01/2024	Contributions & Corrections by Partners	Mike Karamousadakis Esther Salami (UPC) John Tsaloukidis
0.6	06/02/2024	Contributions by ENAC and final wrap-up	Jim Sharples (ENAC) Dimitris Petridis
0.7	07/02/2024	Internal Review	Simon Bittner (JOAFG) Constanze Geyer (JOAFG)
1.0	29/02/2024	Final Document	Marc Bonazountas

DISCLAIMER

The document is proprietary to 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

List of Figures

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BIM	Building Information Modeling
CCC	Change Data Capture
CI	Critical Infrastructure
C4I	Command-Control-Communications-Computers & Intelligence
CSV	Comma-Separated Values
DRM	Disaster Risk Management
DOCX	Microsoft Word Text Document
EMS	Emergency Management Services
EO	Earth Observation
ESA	European Space Agency
GIS	Geographic Information System
IoT	Internet of Things
JSON	Java Script Object Notation
KML	Keyhole Markup Language
ML	Machine Learning
NASA	National Aeronautics & Space Administration
NOAA	National Oceanic & Atmospheric Administration
NLP	Natural Language Processing
PDF	Portable Document Format
PII	Personal Identifiable Information
RGB	Red, Green & Blue
SQL	Standard Query Language
UAV	Unmanned Air Vehicle
WMS	Web Map Services
XML	Extensive Markup Language
3D	Three Dimension (al)

SUMMARY

This document D3.4 provides an extensive overview of the data processing, integration, analysis, and delivery strategies essential for the PANTHEON project's objective of community-based disaster risk management in the pilot cities of Athens and Vienna. Here is a summary overview:

Data Processing Technologies

1. Emphasizes the significance of cloud-based big data processing for efficiently managing large datasets, highlighting technologies like distributed storage (e.g., HDFS, Cassandra), cloud computing services (e.g., Amazon EMR, Google Cloud Dataproc), and serverless computing (e.g., AWS Lambda, Azure Functions).
2. Explores emerging technologies such as edge computing and fog computing, illustrating their role in decentralizing computational tasks and reducing latency by processing data closer to its source. The impact of 5G/B5G technology on edge data processing capabilities is also discussed.

Data Integration & Fusion

3. Discusses the criticality of integrating diverse data from various sources into a unified representation within the digital twin to facilitate informed decision-making and enhance urban sustainability.
4. Introduces approaches like the data warehouse and mediator approaches, along with data fusion techniques and preprocessing methods aimed at harmonizing heterogeneous data.
5. Explores fusion algorithms such as semantic reasoning, fusion through correlation, and decision support mechanisms to synthesize heterogeneous data into actionable insights.

Data Analysiss

6. Highlights the importance of visualization elements like interactive dashboards, geospatial maps, time-series charts, and heat maps for transforming data into actionable insights.
7. Discusses the role of notifications in event detection and real-time alerting to stakeholders.
8. Emphasizes the significance of logs for simulation replay and post-processing analysis, enabling a thorough examination of events contributing to disaster scenarios and facilitating scenario comparisons and decision support.

Data Delivery Schemes

9. Defines disaster scenarios for Athens and Vienna, outlining associated technologies and data delivery requirements.
10. Indicates a preference for the JavaScript Object Notation (JSON) format for data delivery, ensuring compatibility and ease of use across various systems and applications.
11. Highlights ongoing efforts to establish high-performance, automated, and responsive data delivery schemes capable of integrating data from satellites, in-situ sources, infrastructure, traffic, UAVs, and community inputs.

Conclusion

12. Concludes by underlining the importance of establishing optimal data delivery schemes to support community-based disaster risk management in urban environments.
13. Notes a consensus on using JSON format for data delivery, reflecting discussions during project meetings and ensuring compatibility and ease of use across diverse systems and applications.
14. Provides a comprehensive framework for leveraging advanced data processing, integration, analysis, and delivery technologies within the PANTHEON project, aimed at enhancing disaster risk management in urban environments.

1 INTRODUCTION

A significant **challenge** for PANTHEON lies into the integration of big data generated from six discrete sources (Satellite & Copernicus, In-Situ Io, Infrastructure, Traffic, UAVs, Community Data) into the Smart City Digital Twin. Data integration is essential and involves aggregating data from diverse sources into a cohesive view.

Data integration tools play pivotal role, sifting through heterogeneous data to extract relevant information from vast structured and unstructured datasets. Selecting the most suitable tool for PANTHEON is crucial. An evaluation of ten leading data integration tools was recently conducted based on core features, ease of use, customer support, and annual price. The following tools were assessed: Fivetran, Microsoft SQL Server, Apache Airflow, Informatica PowerCenter, Pentaho, Talent, MuleSoft AnyPoint Platform, IBM InfoSphere DataStage, Boomi, and Oracle Data Integrator. **For efficiency**, it is logical to seek tools that offers high performance, low maintenance, responsive service, and full automation, whist projects like PANTHEON leverage existing resources to meet tight deadlines and avoid additional expenditures. Thus, PANTHEON utilizes the Microsoft SQL Server and Apache Airflow, for data integration, leveraging the expertise and familiarity of its consortium partners (PhoenixNap, 2024)[1].

Data delivery methods encompass the various approaches used to transfer data from one system to another. Effective data delivery is crucial for analytics and extracting value from Big Data. However, not all data delivery solutions offer the same advantages. Data is essential to support analytics and ensure consistency in information systems, facilitating decisions based on a unified version of reality. Data must flow seamlessly between systems and stakeholders within organizations, as well as across end users and enterprise boundaries. To address these challenges, numerous data delivery alternatives exist, yet an optimal solution should embody six key characteristics: Performance, Automation, Ease of Use, Low Maintenance, Readiness to Use, and Responsive Service (ABT, 2024)[2].

Data Delivery Scheme should aim to streamline the often-laborious process of loading data from source databases into target databases and data warehouses. It should support both full-load and partial uploads through real-time Change Data Capture (CDC). The design should empower users to replicate vast amounts of data effortlessly. Furthermore, the architecture should translate into minimal total cost of ownership for stakeholders beyond those defined in PANTHEON, that possess the capability to identify and capture critical events in real-time, thereby enhancing insight, agility, and overall competitiveness (ICICE 2024)[3].

2 SCOPE

The PANTHEON Data Delivery Scheme for Community-Based Disaster Relief Management (CBDRM) is tasked with providing a comprehensive analysis and outlining the requirements of PANTHEON technologies concerning communication, data, and applications. The aim is to gather and process relevant information from various community sources, thereby enhancing the speed and accuracy of disaster impact assessments.

PANTHEON aims to furnish updated and valid information during disaster management operations, by utilizing resources such as Earth observations from Unmanned Aerial Vehicle (UAV) Swarming systems, the Copernicus system, in-situ observations from Smart City Internet of Things (IoT), satellite imagery, real-time data from Geographical Information System (GIS) sources, social media, mobile apps, and crowdsourcing applications, The goal is to translate these components into functional and non-functional requirements for the PANTHEON architecture.

As per the findings of D3.2, reference models should be applied across five main use-case categories, including two targeting scenarios before the disaster, two during the disaster, and one after the disaster. In that respect, PANTHEON can be utilized for:

1. Planning early warning according to simulations (Prevention phase).
2. Training and exercises (Preparedness phase).
3. Situational awareness during the disaster (Response phase).
4. Cross-organization communication during the disaster (Response phase).
5. Documentation and evaluation after the disaster (Recovery phase).

The reference models will primarily emphasize **prevention & preparedness** (P&P) with secondary recommendations for **response & recovery** (R&R) measures to be customized by stakeholders and end-user communities to meet specific requirements. Implementation of measures will be controlled via a **Command & Control (C2)** platform depending on the level of automation employed. PATHEON does not deliver a C4I (command, control, communications, computers, intelligence) platform (Bonazountas et al, 2022)[4].

3 DATA DEFINITION

In the realm of scientific discourse, data encompasses a range of values that convey information, including quantities, qualities, facts, statistics, and symbols subject to interpretation. Each datum represents a single value within this collection, often organized into structured formats like tables for clarity and relevance. These structures themselves can serve as data within larger frameworks, functioning as variables in computational processes to represent abstract concepts or tangible measurements. Data finds extensive use across scientific research, economics, and various organizational domains. Examples of datasets pertinent to *wildfires, floods, earthquakes* include statistics on weather patterns, land use, population demographics, and disaster response efforts. In this context, data serves as the fundamental building blocks from which valuable insights and informed decisions can be gleaned (Purdue, 20219)[5]

Data pertaining to wildfires, floods and earthquakes is collected through diverse methods such as databases, measurement, observation, querying, analysis. Typically represented as numerical or character-based formats for further processing, this data originates from uncontrolled real-world environments (field data) or controlled scientific experiments (experimental data). Analysis of such data involves several techniques including remote sensing processing, calculation, reasoning, discussion, presentation, visualization, and post-analysis methodologies. Prior to analysis, raw data undergoes cleaning procedures to address outliers and rectify instrument or data entry errors (Bonazountas et al, 2005)[6]

Data serves as the cornerstone for calculation, reasoning, and discussion within scientific discourse. Ranging from abstract concepts to tangible measurements, including statistical figures, data gains significance when organized thematically within relevant contexts, transforming into actionable information. Interconnected pieces of contextual information culminate in data insights or intelligence. The accumulation of insights and intelligence over time, derived from synthesizing data into information, is often termed knowledge.

In the contemporary digital landscape, the proliferation of computing technologies led to an advent of big data, characteristically representing vast quantities of information, often at the petabyte scale. *Conventional* data analysis methods and computing infrastructures encounter challenges in handling such massive and expanding datasets. Theoretically, infinite data would yield infinite information, rendering the extraction of insights or intelligence impractical. Consequently, the emerging field of data science harnesses machine learning (ML) and other artificial intelligence (AI) techniques to facilitate the efficient application of analytical methods to big data. PANTHEON is utilising data that make sense and assist to deliver tangible results, valuable to stakeholders (Lynggaard, K. 2019)[7].

3.1 TYPES OF DATA

3.1.1 TYPES OF DATA SOURCES

Data sources refer to physical or digital locations where information is stored in formats such as data tables, data objects, or other storage formats. The most prevalent types of data include (Bonazountas, Woldbak, Hellenic Civil Protection, 2023)[8]:

1. **Structured, unstructured or semi-structured data**
    a) Structured data is a standardised format to providing information about a page and classifying the page content, i.e., Excel files with the names of Consortium Partners, their postal and email addresses or Standard Query Languages (SQL) databases.
    b) Unstructured data is the compilation of many various types of data that are stored in their native formats, i.e., social media, images, audio, video and text files.

c) Semi-Structured data sources include emails, CSVs, XML and other markup languages, binary executables, zipped files or webpages. This category was introduced because it is easier to analyse than unstructured data.

2. **Big data** like the three types of data (i.e., structured, unstructured, or semi-structured data) is a collection of data from different sources and characteristics as the 5Vs (volume, value, variety, velocity, veracity).
3. **Internal data** comprises facts and information originating from an organization's systems. In many instances, external parties access/analyse internal data without explicit permission of the owning entity.
4. **External data** originates outside an organisation and is readily available to the public.
5. **Third party analytics data** is collected and managed by organisations that do not directly interact with its customers (e.g., data sets compiled from governmental, non-profit or academic sources.
6. **Open data** is openly accessible to all, including companies, citizens, media, consumers. Content can be freely used, modified, and shared by anyone (e.g., environmental data).

In the context of PANTHEON, data sources for Disaster Risk Management (DRM) are delineated to the six (6) streams: (1) Satellite, (2) In-Situ, (3) Infrastructure, (4) Traffic, (5) Unmanned Aerial Vehicle (UAV), (6) Community. Each of these sources constitutes Open and External data, with some potentially falling under the category of big data, contingent upon the complexity of the pilot cases being utilized. Additionally, several pertinent sub-types of data sources, likely to be employed in the PANTHEON project, are identified below:

3.1.2 Geospatial Data

Geospatial Data[9] encompasses vast collections of spatial data sourced from various outlets in diverse formats. It may include census data, satellite imagery, weather data, cell phone data, drawn images, and social media data. Geospatial data proves invaluable when discoverable, shareable, analyzable, and amalgamated with conventional business data. Consequently, it plays a pivotal role in disaster impact assessments (Bonazountas, 2017)[10].

Formats such as Geographic JavaScript Object Notation (GeoJSON), Vector/Shape geodata, or Raster geodata (GeoTIFFs) are commonly utilized for representing geographic information. These formats facilitate the integration of various data types such as terrain elevation, land use, and infrastructure due to their inclusion of geographic information, including maps, coordinates, and spatial attributes.

Geographic Information Systems (GIS) are processing and mapping data via “layered” visual representations. For instance, when overlaying a hurricane map (depicting location and time) with another layer illustrating potential areas for lightning strikes, GIS functionality becomes evident. Consequently, geospatial data enables the visualization and analysis of the spatial aspects of disasters. ArcGIS/ESRI is a leading s/w technology.

3.1.3 Time Series & Sensors Data

This category encompasses data sourced from diverse outlets such as weather stations, remote sensing satellites, and Internet of Things (IoT) devices, offering real-time information. Common data formats include databases like Standard Query Language (SQL) or NoSQL, Java Script Object Notation (JSON), Comma Separated Values (CSV), or specialized formats tailored for specific sensor types (e.g., NetCDF for climate data). Standards like MQTT and Extensible Markup Language (XML) are frequently employed for transmitting IoT sensor data (Bonazountas, Arc/FIRE, 2007)[11]

3.1.4 Remote Sensing Data

Formats such as satellite imagery, aerial photographs, and drone-captured data offer visual insights into disaster-affected areas (Bonazountas, 2022)[12]

3.1.5 3D Models Data

For simulating and visualizing disaster scenarios in 3D space, formats like COLLADA (DAE) or GeoTIFFs are utilized to represent 3D models of buildings, terrain, and infrastructure (Bonazountas, SEIS, 2015)[13]

3.1.6 PANTHEON Dashboard

To overcome the difficulty of integrating data from various sources such as Building Information Modelling (BIM), Geographic Information System (GIS) and Internet of Things (IoT), PANTHEON is to expand and upgrade data import and integration capabilities, by providing the ability to import data types into a common dashboard.

3.1.7 Static, Dynamic, Real, Synthetic Data

3.1.7.1 Static Data

Static data refers to information that remains unchanged or infrequently updated over time. It serves as a reference or guideline for other data and typically does not require frequent updates or alterations. Examples of static data in the PANTHEON context include infrastructure data, reports or records generated by human resources, and information about past incidents such as wildfires, earthquakes, heatwaves, floods, terrorist attacks, or cyberattacks.

3.1.7.2 Dynamic Data

In data management, dynamic or transactional data refers to information that undergoes periodic updates, evolving asynchronously over time as new information emerges. This concept holds significance in data management, as the temporal nature of the data dictates its processing and storage methods.

3.1.7.3 Real & Synthetic Data

Real data is collected from genuine events, while synthetic data is artificially generated by computer algorithms. In recent years, there has been a burgeoning interest in employing synthetic data for diverse applications, including machine learning and data analytics. According to Gartner, by 2030, the utilization of synthetic data in AI (Artificial Intelligence) models is projected to surpass that of real data.

3.1.7.4 Synthetic Data Creation

While various methods exist for generating synthetic data, AI-generated synthetic data is crafted by AI models trained on intricate real-world datasets, leveraging the capabilities of deep learning algorithms. The advantage of employing generative AI lies in its ability to autonomously discern patterns, structures, correlations, and other complex relationships within real data. Subsequently, the AI model learns to generate entirely new data instances while preserving the inherent patterns observed in the original dataset. This structural similarity is visually apparent (TONIC 2024, Figure 1)[14,15].

A prevalent technique involves generating data through computer algorithms that emulate the behavior observed in real-world datasets. This method enables the creation of synthetic datasets that closely resemble real datasets in terms of their distribution and variability. Another commonly employed approach for generating synthetic data is utilizing a random number generator to produce data adhering to specific statistical distributions, devoid of any inherent correlations.

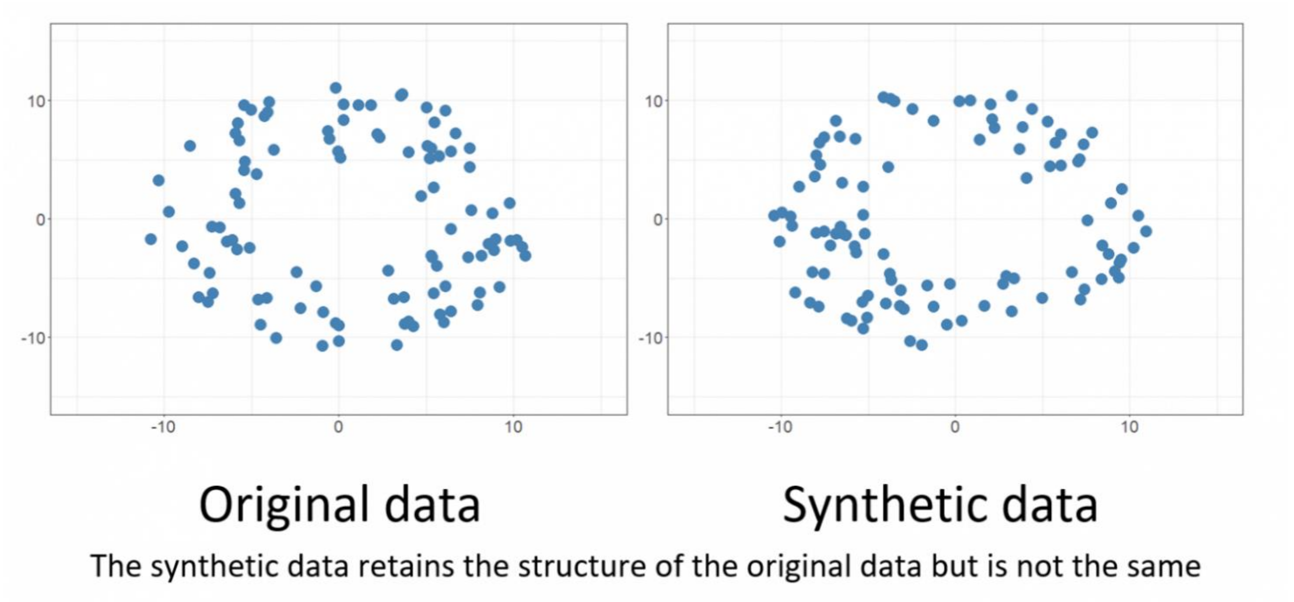


Figure 1: Original and Synthetic Data (TONIC, UK Government, 2024)

3.1.7.5 *Benefit of synthetic data vs real time data*

There are several advantages to using synthetic data over real data[16]:

1. **Overcoming regulatory restrictions:** Synthetic data bypasses regulatory constraints associated with real data, as it replicates essential statistical properties without compromising privacy. This eliminates concerns regarding privacy regulations and facilitates:
2. **Privacy preservation**: Traditional anonymization methods often compromise data utility while protecting privacy; this privacy/utility trade-off by eliminating the need to safeguard real data.
3. **Resistance to reidentification**: Real data, even after anonymization, can still be reidentified. Synthetic data mitigates this risk, as it does not contain identifiable information.
4. **Aptitude for innovation and monetization**: Synthetic data can be shared with third parties for research and innovation purposes without privacy concerns, offering opportunities for monetization.
5. **Streamlines simulation**: Synthetic data enables the generation of data to simulate scenarios not yet encountered. In instances where real data is unavailable, synthetic data provides the only viable solution. For example, automotive companies may use synthetic data to train smart cars for situations not captured in real-world data.
6. **Statistical issues**: Synthetic data is immune to common statistical problems such as item non-response and skip patterns. By carefully designing rules for data generation, synthetic data can be created to avoid these pitfalls, ensuring robust statistical analyses.
7. **Speeds up the process**: Synthetic data can be generated much faster than real data can be collected, saving time and enhancing agility and competitiveness in the market.
8. **Achieves higher consistency**: Synthetic data is more uniform and consistent than real data, which may exhibit variability due to its natural origins. This consistency facilitates accurate analyses on synthetic datasets.
9. **Ensures easy manipulation**: Synthetic data can be manipulated more easily than real data in a controlled manner, enabling precise testing and training of machine learning models. It can be generated in large quantities with specific characteristics and biases, enhancing performance in various applications.
10. **Increases cost-effectiveness**: Synthetic data can be more cost-effective than real data. While there are upfront costs associated with building simulations for synthetic data generation, the recurring costs of collecting and revising real data are avoided.
11. **Facilitates AI/ML training**: Synthetic data is valuable for training AI/ML models, as it is not subject to the regulations governing real data and can be generated in abundance. This enriches model training and enhances learning capacity.

3.1.7.6 Challenges with using synthetic data against real data

In addition to the array of benefits the utilization of synthetic data presents several challenges as (Datomize, 2024)[17]:

1. **Biased or deceptive results**: Synthetic data may yield misleading, constrained, or discriminatory outcomes due to its limited variability and correlations.
2. **Lack of accuracy**: Synthetic data is often generated using computer algorithms, which may not always produce accurate representations. Consequently, synthetic data runs the risk of occasionally generating inaccurate results.
3. **Time-consuming verification steps**: Synthetic data necessitates additional verification procedures, such as comparing model outputs with human-annotated real-world data. These verification efforts are time-intensive and can prolong project timelines.
4. **Loss of outliers**: Synthetic data may fail to encompass outliers present in the original dataset, as it can only mimic but not perfectly replicate real-world data. However, outliers may hold relevance for certain research endeavors.
5. **Dependency on real data**: The quality of synthetic data is often contingent upon the quality of the real dataset and model utilized for its creation. Without a robust and high-quality real dataset, synthetic datasets generated from it may prove ineffective or even erroneous.
6. **Consumer scepticism**: As the use of synthetic data becomes more prevalent, businesses may encounter consumer scepticism regarding the credibility of data-driven conclusions and products. Consumers may demand transparency regarding data generation techniques and assurance of their data privacy.

3.1.8 Data Format & Delivery Modes

3.1.8.1 Data Format

Data format plays crucial role in structuring and representing information, especially in disaster impact assessments. Several common data formats include (FEMA, 2024)[18]:

1. **Numerical data** encompasses two primary types: (a) discrete and (b) continuous. Discrete data comprises specific or fixed values, such as the number of people in an institution, and can be displayed via Bar Charts. On the other hand, continuous data falls within a given range of values, is measurable, and is graphically depicted using Histograms.
2. **Textual data** refers to information conveyed in written or printed form, encompassing sources like books, articles, surveys, social media posts, chat conversations, and emails.
3. **Image data** is generated by scanning a surface with an optical or electronic device and includes examples like scanned documents, remotely sensed data (e.g., satellite images), and aerial photographs. Images are stored as raster datasets of binary or integer values representing the intensity of reflected light, heat, or other values on the electromagnetic spectrum.

Above formats can be further categorized as structured, unstructured, or geospatial:

1. **Structured Data:** Structured data, organized in rows and columns, can be employed to store information such as historical wildfire occurrence records, fire weather data, and fuel moisture content. These structured datasets can aid in analysing trends, identifying high-risk areas, and developing predictive models for wildfire behaviour.
2. **Unstructured Data:** Unstructured textual data, including reports from fire incident commanders, news articles, and social media posts, can provide valuable contextual information about ongoing wildfires. Natural language processing techniques can be utilized to extract relevant insights from these sources, helping emergency responders gain a comprehensive understanding of the current wildfire situation.
3. **Geospatial Data:** Geospatial data, including satellite imagery, aerial photographs, and maps, is instrumental in wildfire management. Satellite imagery can be used to detect active fire hotspots, monitor fire spread, and assess burn severity. Aerial photographs provide detailed views of wildfire-affected areas, aiding in damage assessment and post-fire recovery efforts. Maps, containing geographic

features such as vegetation types, road networks, and water sources, facilitate strategic planning and resource allocation during wildfire response operations (Zhou, G., et al. 2023)[19].

The mentioned data formats are instrumental in wildfire, flood, and earthquake management by facilitating the integration of critical geographic information. This integration encompasses various data types such as terrain elevation, land use, and infrastructure, which include essential geographic details like maps, coordinates, and spatial attributes. Geographic Information Systems (GIS) utilize geospatial data to visualize and analyse the spatial dimensions of these natural disasters, and for (Bonazountas, 2015)[20]:

- **wildfires**, geospatial data enables the mapping of fire-prone areas, identification of fuel types, and assessment of fire behavior based on terrain elevation. It also aids in planning evacuation routes, locating fire suppression resources, and assessing post-fire impacts on landscapes and ecosystems.
- **floods**, geospatial data assists in delineating floodplains, identifying vulnerable infrastructure such as bridges and roads, and predicting flood extents based on terrain elevation and land use patterns. GIS can also be utilized to model flood scenarios, assess flood risk, and develop floodplain management strategies.
- **earthquake** management, geospatial data plays a crucial role in identifying seismic hazard zones, assessing building vulnerability, and planning emergency response routes. GIS enables the visualization of earthquake shaking intensity maps, identification of critical infrastructure at risk, and prioritization of retrofitting efforts to enhance community resilience.
- **overall** by leveraging these data formats within GIS frameworks, emergency managers and decision-makers can gain valuable insights into the spatial characteristics of wildfires, floods, and earthquakes. This enables proactive planning, effective response coordination, and informed decision-making to mitigate the impacts of these disasters on communities and infrastructure.

3.1.8.2 Data delivery mode

Efficient data delivery modes play a pivotal role in ensuring that pertinent scientific insights reach decision-makers promptly and efficiently, especially in the context of wildfires, floods, and earthquakes. Several key delivery modes are indispensable in disaster management. Efficient data delivery methods are critical for effective management of wildfires, floods, and earthquakes. Here are tailored delivery modes for these disaster scenarios (Cao, L. et al 2023)[21]:

1. **Real-Time Data Streaming**: Continuous data streams from sensors, satellites, and monitoring networks facilitate the instantaneous flow of real-time information. Protocols like MQTT or Web Socket enable swift transmission of critical data, empowering decision-makers to respond promptly to evolving disaster situations.
2. **Cloud-based Storage/Data Integration Platforms/APIs & Web Services**: Leveraging cloud-based storage solutions, data integration platforms, and APIs ensures standardized access to data with scalable storage and processing capabilities. This enables seamless integration and retrieval of data across various systems and applications, facilitating efficient management and sharing of large datasets among stakeholders and end-users. Cloud storage solutions provide scalability and accessibility, ensuring critical data remains readily available during disaster events.
3. **Mobile Data Delivery**: Mobile technology and applications enable field personnel to access and contribute to real-time data delivery. This facilitates communication between remote areas and central command, enabling informed decision-making during disaster response efforts. RESTful APIs or GraphQL endpoints can be utilized to deliver data from diverse sources to digital twin platforms, enhancing situational awareness and decision support capabilities.
4. **Batch Processing**: For historical analysis or batch updates of data, batch processing modes such as data pipelines are invaluable. Tools like Apache Kafka or Apache Spark streamline batch processing workflows, enabling comprehensive analysis and reporting. Recent reports from KEMEA indicate the maintenance

of historical data spanning the last decade concerning wildfires and floods in the Athens region, highlighting the importance of batch processing for historical analysis and trend identification in disaster management.

3.1.9 Data on Earth Observation, UAV, in-situ, IoT, and Copernicus Services

Earth observation (EO) data involves collecting information about the Earth's surface, atmosphere, and oceans using remote sensing technologies like satellites, drones, and ground-based sensors. This data is crucial for various applications, including disaster management. Here are key points about EO data (EEA, 2023)[22]:

1. **Remote Sensing Technologies**: (a) Satellites: Capture images and data from space, equipped with sensors such as optical, radar, and specialized instruments, (b) Unmanned Aerial Vehicles (UAVs): Used for data collection, offering various types and capabilities depending on weight, propulsion, and sensors.
2. **Types of Earth Observation Data**: (a) Optical Imagery: Visible and infrared light for land cover classification, agriculture monitoring., (b) Radar Imagery: Microwave frequencies for all-weather imaging, useful for land cover changes and terrain mapping, (c) Thermal Infrared Imagery: Measures heat radiation, beneficial for detecting wildfires and urban heat islands, (d) Hyperspectral Imagery: Captures various wavelengths for detailed material analysis, useful for mineral exploration
3. **Applications of EO Data**: (a) Environmental Monitoring: Track changes in land use, deforestation, and air quality, (b) Natural Resource Management: Manage forests, water, and minerals efficiently, (c) Disaster Response: Assess impact and support response efforts for disasters like hurricanes, earthquakes, and floods, (d) Agriculture: Monitor crops, predict yields, and practice precision agriculture, (e) Climate Studies: Contribute to understanding climate patterns and changes in sea levels, ice cover, and atmospheric conditions.
4. **Data Providers:** (a) Organizations like NASA, ESA, NOAA, and private companies operate satellites and provide EO data to the public.
5. **Challenges**: (a) Data Volume: Large datasets require advanced storage and processing capabilities, (b) Data Access and Sharing: Ensuring global access and promoting data sharing remains a challenge, (c) Data Integration: Integrating EO data with ground-based measurements is essential for comprehensive analysis.

In PANTHEON, EO data primarily comes from satellites, including services like Copernicus, Landsat, and VIIRS (Bonazountas, 2023)[23]:

1. **Copernicus Services**: Provides various data formats like GeoTIFF, GeoJSON, and NetCDF. Offers structured metadata for data details like acquisition date, sensor specifications. Provides access through Web Map Services (WMS) and APIs for data retrieval
2. **Landsat Services**: Provide images of Earth's surface, distributed by USGS EROS. Sensors acquire data in different frequency ranges with varying spatial resolutions.
3. **UAVs**: Used for detailed data collection, offering high-resolution imagery and LiDAR scanning. Data can be transmitted in real-time or stored onboard for later processing.
4. **In-Situ Sensors**: Provide real-time data for continuous monitoring of environmental conditions.
5. **IoT-based weather stations** offer detailed weather information transmitted in JSON format.
6. **PANTHEON**: Efficiently managing and utilizing these diverse data sources is essential for informed decision-making and effective disaster management in PANTHEON.

3.1.10 Data themes for use cases during the phases of disaster management

Based on D2.2 this chapter contains considerations about the use cases in the pilot regions.

3.1.10.1 Athenian Demonstrator

Greece, particularly the Attica region where the capital Athens is located, is prone to both seismic activity and wildfires, presenting significant challenges for disaster management efforts.

The Attica region has a history of devastating **seismic events**, with notable earthquakes occurring in 1938, 1953, 1981, and 1999. These earthquakes have resulted in loss of life, damage to critical infrastructure, and the collapse of buildings, alongside secondary effects like landslides and soil liquefaction. Given that nearly half of Greece's population resides in the Athens metropolitan area and the broader Attica region, earthquake scenarios are of paramount importance in disaster preparedness and response planning.

In addition to seismic risks, **wildfires** have emerged as a major concern in recent decades, particularly during the summer months. The mountains surrounding Athens, including Parnitha, Penteli, and Imittos, have been significantly impacted by wildfires. The devastating August 2018 wildfire, which swept through the suburb of Mati, resulted in the tragic loss of 103 lives and underscored the urgency of addressing wildfire scenarios in disaster management plans.

While earthquakes and wildfires are primary hazards, other scenarios such as **floods**, exemplified by the catastrophic events in the suburb of Mandra in 2017, and the potential for terrorist attacks, further complicate disaster preparedness efforts. The likelihood of such events occurring is influenced by the complex geopolitical dynamics of the South-East Mediterranean region.

Given the complexity of simulating these hazards and the limited time available of PANTHEON, the selection of scenarios for exercise is crucial. The aim is to produce actionable insights for first responders and stakeholders based on current experiences and the most likely scenarios. For first responders, this may involve updating action checklists, while stakeholders may need to revise preparedness plans to enhance resilience and response capabilities. By focusing on the most probable and impactful scenarios, disaster management efforts can better mitigate risks and protect communities in the Athens region.

3.1.11 Vienna Demonstrator

In the context of Vienna, two significant disaster scenarios emerge as focal points for preparedness and response efforts: heatwaves and wildfires, each presenting unique challenges and considerations.

Heatwaves are a prevalent threat during the summer months in Central and Northern Europe, including Vienna. The absence of widespread air conditioning in residential areas, coupled with limited availability of ambulances and medical personnel due to summer vacations, exacerbates the vulnerability of elderly and at-risk individuals during extreme heat events. Therefore, preparing for and responding to heatwaves is crucial to safeguarding public health and reducing heat-related illnesses and fatalities.

Vienna also faces the potential risk of **wildfires** triggered by man-made events, such as a cyber-attack targeting critical infrastructure like a power plant. A cyber-attack on a power plant could ignite a fire capable of spreading rapidly through the surrounding forested areas, including the outskirts of Vienna. This scenario highlights the interconnectedness of infrastructure vulnerabilities and environmental risks, underscoring the need for comprehensive disaster preparedness and response strategies.

By selecting these scenarios for simulation exercises, stakeholders can evaluate and enhance their readiness to address both natural and man-made disasters effectively. The choice of a cyber-terrorism trigger for the wildfire scenario allows for the exploration of cascading effects and interagency coordination in response to unexpected events. Through collaborative training exercises involving first responders, healthcare providers, law enforcement agencies, municipal authorities, infrastructure providers, and media representatives, stakeholders can improve their coordination and cooperation in managing complex disaster scenarios.

While the initial focus may be on heatwaves and cyber-triggered wildfires, the development of final scenarios will be contingent on the complexity of the simulations and the available project timeline. Deliberate consideration of cascading effects and the potential exclusion of certain elements may be necessary to ensure the feasibility of the pilot project in Vienna within the project's timeframe. Ultimately, the aim is to leverage the digital twin platform to strengthen preparedness, response, and resilience across various disaster scenarios in Vienna and beyond.

3.2 Characteristics of Data

3.2.1 Data Quality

Given the diverse array of data involved in PANTHEON, ensuring high-quality data is essential for its success. Establishing a robust data quality framework is paramount to certify the reliability and usefulness of the data collected and utilized throughout the project. A comprehensive data quality framework comprises various components, each playing a crucial role in maintaining and enhancing data quality. Here are the key components of such a framework (Taleb, 2021)[24]:

1. **Data Governance**: Data governance involves establishing policies, procedures, and responsibilities for managing and safeguarding data assets. It ensures that data is handled consistently, securely, and in compliance with relevant regulations and standards. For example, in the context of **wildfires**, data governance may dictate protocols for sharing satellite imagery data among stakeholders to facilitate timely response and decision-making.
2. **Data Profiling**: Data profiling entails analysing the structure, content, and quality of data to identify anomalies, inconsistencies, or inaccuracies. For instance, in the case of **heatwaves**, data profiling may involve examining historical temperature records to detect outliers or data entry errors that could affect the accuracy of heatwave predictions.
3. **Data Quality Rules**: Data quality rules define criteria or standards for acceptable data quality. These rules establish benchmarks against which data quality can be assessed and monitored. In the context of **floods**, data quality rules may specify thresholds for river water levels or rainfall intensity, beyond which data is flagged for review or cleaning.
4. **Data Quality Assessment**: Data quality assessment involves evaluating the adherence of data to predefined quality standards and rules. This process may include automated checks, statistical analyses, and manual reviews to identify and address data quality issues. For example, in the aftermath of a **flood**, data quality assessment may involve verifying the accuracy of damage reports submitted by field teams.
5. **Data Cleaning**: Data cleaning encompasses the process of correcting errors, removing duplicates, and standardizing data to improve its quality and consistency. In the case of **wildfires**, data cleaning may involve reconciling discrepancies between satellite imagery and ground observations to produce accurate fire perimeter maps for emergency response teams.
6. **Data Monitoring**: Data monitoring involves ongoing surveillance of data quality metrics and indicators to detect deviations or anomalies. Continuous monitoring ensures that data remains accurate, timely, and relevant over time. For instance, in the context of **heatwaves**, data monitoring may involve tracking temperature trends and heat stress indices to anticipate and mitigate health risks.
7. **Data Issue Management**: Data issue management refers to the process of documenting, tracking, and resolving data quality issues and discrepancies. This includes assigning responsibilities, prioritizing issues,

and implementing corrective actions to address root causes. During a **flood** event, data issue management may involve coordinating with data providers to resolve inconsistencies in flood inundation maps used for evacuation planning.

8. **Data Reporting**: Data reporting involves communicating data quality status, findings, and insights to stakeholders through various reports and dashboards. Clear and transparent reporting facilitates informed decision-making and accountability. For example, in the context of **wildfires**, data reporting may involve disseminating real-time fire behaviour forecasts and evacuation orders to emergency responders and the public.
9. **Continuous Improvement**: Continuous improvement is an ongoing process of refining and enhancing the data quality framework based on feedback, lessons learned, and evolving requirements. Regular evaluations and adjustments ensure that the framework remains effective and adaptable to changing needs and circumstances. In the aftermath of a disaster, such as a **heatwave**, continuous improvement may involve conducting post-event reviews to identify opportunities for optimizing data collection, analysis, and dissemination processes.
10. **Implementation** of a comprehensive data quality framework encompassing these key components, PANTHEON can ensure that the data used for **disaster** management, including wildfires, heatwaves, and floods, is accurate, reliable, and actionable, ultimately enhancing the effectiveness of response and mitigation efforts.

3.2.2 DATA INTEGRITY/SECURITY

Ensuring data security (and integrity) is crucial for disaster management, where timely and accurate information can make a significant difference in response and recovery efforts. Let's enrich the discussion by incorporating examples from wildfires, floods, and earthquakes (Bonazountas ey al., 2017)[25]:

1. **Wildfires**: In the case of wildfires, data integrity is essential for assessing the extent of the fire, predicting its behaviour, and coordinating evacuation efforts. Satellite imagery, weather data, and ground observations are integrated to map the fire perimeter and identify areas at risk. However, ensuring the integrity of this data is challenging due to factors like smoke interference with satellite sensors or inaccurate ground reports. Data security measures must be in place to prevent unauthorized access to sensitive information, such as evacuation routes or firefighter locations, which could compromise response efforts.
2. **Floods**: During floods, accurate data is critical for assessing flood extent, depth, and velocity to predict flood behaviour and plan emergency responses. Data sources such as river gauges, rainfall measurements, and hydraulic models are integrated to create flood inundation maps. However, data integrity can be compromised by factors like sensor malfunctions, human error in data collection, or cyber-attacks targeting flood monitoring systems. Implementing robust data security measures is essential to prevent tampering with flood data, which could lead to inaccurate flood forecasts and inadequate response measures.
3. **Earthquakes**: In earthquake-prone areas, data integrity is vital for assessing seismic activity, identifying affected areas, and estimating damage severity. Seismic sensors, geospatial data, and building inventory databases are integrated to create earthquake hazard maps and prioritize response efforts. However, ensuring the integrity of earthquake data is challenging due to factors like sensor calibration errors, data transmission delays, or data manipulation by malicious actors. Data security protocols must be implemented to safeguard seismic data integrity and prevent misinformation that could hamper emergency response and recovery operations.
4. **Heatwaves**: In regions prone to heatwaves, accurate data is essential for assessing temperature trends, predicting heatwave intensity and duration, and implementing effective heatwave response measures. Temperature data from weather stations, satellite observations, and urban heat island mapping are

integrated to identify areas susceptible to extreme heat events and plan appropriate interventions. However, data integrity can be compromised by factors like sensor calibration errors, data transmission glitches, or manipulation of temperature records. Maintaining data integrity is crucial to prevent misinformation that could impact public health and safety during heatwaves. Implementing robust data security measures is essential to safeguard temperature data from unauthorized access or tampering, ensuring the reliability of information used for heatwave risk assessments and mitigation strategies. This involves encrypting data transmissions, implementing access controls, and conducting regular security audits to detect and address vulnerabilities in heatwave monitoring systems. Additionally, educating stakeholders on the importance of data integrity and security measures is essential for building trust in heatwave data and promoting effective response measures.

5. **PANTHEON overall**: By addressing data security challenges and complying with security requirements, disaster management agencies can enhance data integrity and ensure the reliability of information used for impact assessments and decision-making. This involves implementing encryption measures to protect data in transit and at rest, establishing access controls to limit unauthorized access, and conducting regular audits to detect and mitigate security vulnerabilities. Additionally, training personnel on data security best practices and fostering a culture of security awareness are essential for maintaining data integrity in disaster management contexts.

3.2.3 Data Security challenges & Considerations in Disaster Management

Selected issues on data security natural disaster (FIAU 2024, Bonazountas 2022)[26,27]:

1. **Compliance with Regulations**
   a) Wildfires: Agencies responsible for wildfire management, such as the Mechanism of the European Commission's Emergency Response Coordination Centre (ERCC, The Centre monitors wildfire risks and emergencies across Europe, supported by national and European monitoring services such as the European Forest Fire Information System EFFIS) must comply with regulations like the European Health Observatory from Smoke and Wildfires (ADAPT, 2024)[28], to protect sensitive health data collected during wildfire evacuations and medical response efforts.
   b) Floods: Flood response agencies, such as European Flood Observatory (EEA, 2024)[29] must adhere to regulations like the European Union's General Data Protection Regulation (GDPR) when handling personal data of flood-affected individuals during disaster relief operations.
   c) Heatwaves: Health departments monitoring heatwave impacts on vulnerable populations must comply with regulations such as the European Heatwave Observatory (ADAPT, 2024)[30] or the US Health Information Portability and Accountability Act (HIPAA) to safeguard patient data collected from heat-related illnesses.
   d) Earthquakes: Seismological agencies such as the The European-Mediterranean Seismological Centre (EMSC)[31] or the US Geological Survey (USGS) must comply with regulations such as the European Union's General Data Protection Regulation (GDPR) when sharing seismic data with international partners.

2. **Complexity**
   a) Wildfires: Integrating data from satellite imagery, weather forecasts, and ground sensors to predict wildfire behavior and assess its impact on communities requires sophisticated data management and analysis tools.
   b) Floods: Assessing flood risk and coordinating emergency response efforts involve analyzing complex datasets, including topographic maps, rainfall forecasts, and floodplain models.
   c) Heatwaves: Monitoring heatwave intensity and duration across urban areas requires integrating data from temperature sensors, satellite observations, and urban heat island mapping.

d) Earthquakes: Analyzing seismic data from multiple monitoring stations to detect earthquake signals and assess ground shaking intensity demands advanced computational techniques and data processing algorithms.

3. **Resource Constraints**
   a) Wildfires: Investing in cybersecurity measures to protect wildfire data from unauthorized access and cyber threats requires significant financial resources.
   b) Floods: Deploying secure data storage systems capable of handling large flood-related datasets may strain agency budgets.
   c) Heatwaves: Training personnel to handle heatwave data securely and comply with data protection regulations necessitates ongoing investments in workforce development.
   d) Earthquakes: Upgrading seismic monitoring networks and implementing encryption protocols to safeguard earthquake data may require additional funding allocations.

4. **Interoperability**
   a) Wildfires: Ensuring interoperability between wildfire management systems used by federal, state, and local agencies to share real-time fire incident data and coordinate response efforts.
   b) Floods: Integrating flood risk assessment tools with Geographic Information Systems (GIS) platforms to visualize flood hazard maps and evacuation routes for at-risk communities.
   c) Heatwaves: Connecting temperature monitoring networks with public health databases to identify heatwave hotspots and target interventions for vulnerable populations.
   d) Earthquakes: Establishing data-sharing protocols between international seismological agencies to facilitate real-time earthquake monitoring and early warning systems.

5. **Training**
   a) Wildfires: Training fire personnel on secure data handling practices and cybersecurity protocols to prevent data breaches and ensure the integrity of wildfire incident reports.
   b) Floods: Educating emergency responders on GDPR compliance and data protection principles when collecting and sharing flood-related information with government agencies and relief organizations.
   c) Heatwaves: Providing healthcare professionals with training on HIPAA regulations and patient confidentiality when accessing and analyzing heatwave health data.
   d) Earthquakes: Conducting workshops and seminars for seismologists and data scientists on best practices for safeguarding seismic data and preventing unauthorized access.

6. **Implementation of Robust Data Security Measures**
   a) Wildfires: Implementing encryption protocols and access controls to protect sensitive wildfire suppression plans and firefighter deployment strategies from cyber threats.
   b) Floods: Deploying secure cloud-based platforms with multi-factor authentication to store and share flood risk assessment data with government agencies and emergency responders.
   c) Heatwaves: Utilizing secure data transmission protocols and encryption algorithms to safeguard heatwave vulnerability assessments and public health data collected from temperature monitoring stations.
   d) Earthquakes: Establishing secure data servers with role-based access controls to protect seismic event catalogs and ground motion records from unauthorized tampering or manipulation.

3.2.4 Data Security & Compliance Requirements

Security and compliance requirements are indispensable in disaster impact assessments to ensure data integrity, privacy, and reliability. In an era marked by an increasing frequency of disasters, the ability to map relevant scientific knowledge while safeguarding sensitive information is paramount. By addressing these requirements, disaster management Agencies can expedite their response efforts, minimise damage, and ultimately save lives. Furthermore, compliance with data protection regulations enhances public trust and demonstrates a commitment to responsible data management. Balancing the need for security and compliance with the demands of rapid and accurate disaster assessments is a critical step toward building a more resilient and prepared world (IMMUTA, 2024)[32]. Examples related to wildfires, floods, heatwaves, and earthquakes:

1. **Data Encryption:** All sensitive data should be encrypted both in transit and at rest to protect it against data breaches and from unauthorised access. Example: In **wildfire** management, sensitive data such as evacuation plans and infrastructure maps are often transmitted between agencies. Implementing end-to-end encryption ensures that this data remains secure during transit, protecting it from interception by unauthorized parties.
2. **Access Control**: Implement robust access control mechanisms to ensure that only authorised personnel can view or modify sensitive data. Example: During **flood** risk assessments, access to detailed floodplain maps and vulnerability assessments should be restricted to authorized personnel within disaster management agencies. Role-based access control mechanisms can be implemented to ensure that only individuals with the appropriate clearance can view or modify sensitive flood-related data.
3. **Data Backups, Redundancy, and Recovery**: Regular data backups and redundancy systems along with the establishment of Recovery procedures, guarantee data availability even in the case of infrastructure failures, and prevention of data loss due to cyber-attacks or other disasters. Example: In regions prone to **earthquakes**, seismic sensor networks continuously collect data on ground motion and seismic activity. Implementing robust backup systems and redundancy measures ensures that this critical seismic data is preserved, even in the event of infrastructure damage caused by earthquakes or other disasters.
4. **Data Masking & Anonymization:** Personally Identifiable Information (PII) and other sensitive data should be masked or anonymized to preserve privacy while still enabling analysis Example: In **heatwave** planning, demographic data such as age and health conditions may be used to identify vulnerable populations. However, to protect individual privacy, this data can be anonymized by replacing identifiable information with pseudonyms or aggregated into broader categories before analysis.
5. **Compliance Audit:** Regular audits and assessments should be conducted to ensure adherence to data protection regulations and industry standards Example: In **earthquake** risk assessments, geological survey data and building infrastructure information are essential for predicting seismic vulnerability. Regular compliance audits ensure that data collection and storage practices adhere to relevant seismic safety standards and regulations, such as building codes and seismic zoning ordinances.

3.2.5 Benefits from Mapping Scientific Knowledge

Security and compliance requirements are fundamental considerations when mapping relevant scientific knowledge for disaster impact assessments, yielding numerous benefits, including (Nature 2017)[33]:

1) **Enhanced Data Trustworthiness**: Secure and compliant data instills confidence among decision-makers, ensuring that the information utilized for assessments is dependable and accurate.
2) **Regulatory Compliance Assurance**: Adhering to data protection regulations not only mitigates legal complexities but also fosters public trust in disaster management endeavors, showcasing a commitment to accountability and transparency.

3) **Resilience Against Cyber Threats**: By implementing robust encryption and access controls, disaster management teams fortify their defence against cyber threats, safeguarding critical data and infrastructure from potential breaches and disruptions.

Example for **Wildfire** Management: In the realm of wildfire management, ensuring the security and compliance of data utilized for assessing *fire risks and devising response strategies* is paramount. By adhering to stringent data protection regulations and implementing encryption measures, wildfire management agencies can inspire confidence in decision-makers and the public regarding the *reliability of the information* used. Moreover, by proactively addressing cybersecurity threats through access controls and encryption protocols, these agencies bolster their resilience against potential cyberattacks targeting critical wildfire management data and systems.

3.2.6 DATA AVAILABILITY

Uninterrupted access to data is indispensable during disaster response and recovery, underscoring the importance of resilience against cyberattacks and infrastructure failures. This resilience is particularly critical in scenarios such as wildfires, floods, heatwaves, and earthquakes, where timely access to accurate data can mean the difference between effective response and widespread devastation (Taylor & Fransis, 2024)[34]. For example, for

- **Wildfires**, access to real-time data on fire behaviour, weather patterns, and evacuation routes is essential for firefighting agencies to make informed decisions and allocate resources effectively. In the event of a flood, access to flood maps, water level sensors, and evacuation orders is crucial for authorities to coordinate evacuations and deploy emergency services to affected areas promptly.
- **Heatwaves**, access to data on temperature forecasts, vulnerable populations, and cooling centres is vital for public health agencies to implement heatwave preparedness measures and prevent heat-related illnesses and fatalities. In earthquakes, access to seismic activity data, building vulnerability assessments, and emergency response plans is indispensable for local authorities to assess damage, prioritize rescue efforts, and coordinate disaster response activities.
- **Overall** to ensure uninterrupted access to data during such disasters, redundancy and backup systems are essential. By maintaining redundant data storage systems and implementing backup protocols, organizations can mitigate the risk of data loss due to cyberattacks, infrastructure failures, or natural disasters. Whether storing data on-premises or in the cloud, establishing secure and accessible data repositories is crucial for ensuring that critical information remains available when needed most, enabling effective disaster response and recovery efforts

4 DATA LIFE CYCLE

The PANTHEON Architecture was recently produced (Figure 2. In a similar way the Logical Architecture is also presented in the following Figure 2. The presented architecture is not the final and may be updated depending on the availability of data, thus is considered a living one to be finalised by delivery T3.7 and WP4. The layered architecture is presented in Figure 3

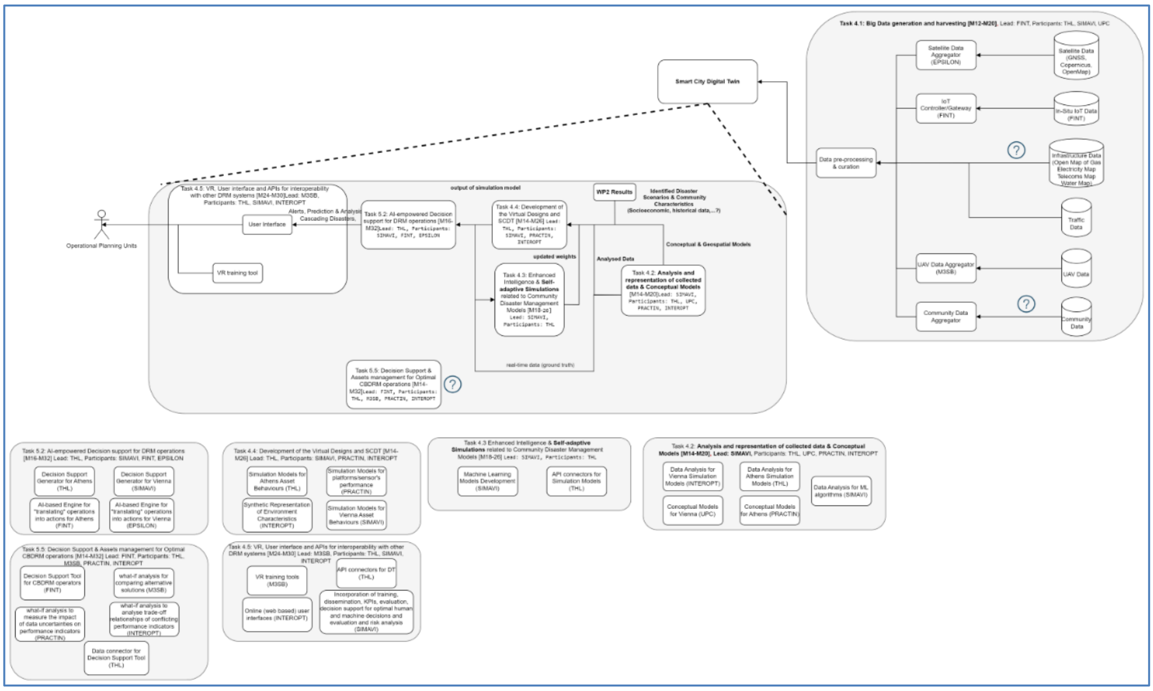

Task 4.1: Big Data generation and harvesting [M12-M20], Lead: FINT, Participants: THL, SIMAVI, UPC
Satellite Data Aggregator (EPSILON)
Satellite Data (GNSS, Copernicus, OpenMap)
IoT Controller/Gateway (FINT)
In-Situ IoT Data (FINT)
Infrastructure Data (Open Map of Gas Electricity Map Telecoms Map Water Map)
Traffic Data
UAV Data Aggregator (M3SB)
UAV Data
Community Data Aggregator
Community Data
Data pre-processing & curation
Smart City Digital Twin
?
Task 4.5: VR, User interface and APIs for interoperability with other DRM systems [M24-M30]Lead: M3SB, Participants: THL, SIMAVI, INTEROPT
User Interface
VR training tool
Operational Planning Units
Alerts, Prediction & Analysis Cascading Disasters,
Task 5.2: AI-empowered Decision support for DRM operations [M16-M32]Lead: THL, Participants: SIMAVI, FINT, EPSILON
output of simulation model
Task 4.4: Development of the Virtual Designs and SCDT [M14-M26] Lead: THL, Participants: SIMAVI, PRACTIN, INTEROPT
updated weights
Task 4.3: Enhanced Intelligence & Self-adaptive Simulations related to Community Disaster Management Models [M18-26] Lead: SIMAVI, Participants: THL
WP2 Results
Identified Disaster Scenarios & Community Characteristics (Socioeconomic, historical data,...?)
Conceptual & Geospatial Models
Analysed Data
Task 4.2: Analysis and representation of collected data & Conceptual Models [M14-M20]Lead: SIMAVI, Participants: THL, UPC, PRACTIN, INTEROPT
real-time data (ground truth)
Task 5.5: Decision Support & Assets management for Optimal CBDRM operations [M14-M32]Lead: FINT, Participants: THL, M3SB, PRACTIN, INTEROPT
Task 5.2: AI-empowered Decision support for DRM operations [M16-M32] Lead: THL, Participants: SIMAVI, FINT, EPSILON
Decision Support Generator for Athens (THL)
Decision Support Generator for Vienna (SIMAVI)
AI-based Engine for "translating" operations into actions for Athens (FINT)
AI-based Engine for "translating" operations into actions for Vienna (EPSILON)
Task 4.4: Development of the Virtual Designs and SCDT [M14-M26] Lead: THL, Participants: SIMAVI, PRACTIN, INTEROPT
Simulation Models for Athens Asset Behaviours (THL)
Simulation Models for platforms/sensor's performance (PRACTIN)
Synthetic Representation of Environment Characteristics (INTEROPT)
Simulation Models for Vienna Asset Behaviours (SIMAVI)
Task 4.3 Enhanced Intelligence & Self-adaptive Simulations related to Community Disaster Management Models [M18-26] Lead: SIMAVI, Participants: THL
Machine Learning Models Development (SIMAVI)
API connectors for Simulation Models (THL)
Task 4.2: Analysis and representation of collected data & Conceptual Models [M14-M20], Lead: SIMAVI, Participants: THL, UPC, PRACTIN, INTEROPT
Data Analysis for Vienna Simulation Models (INTEROPT)
Data Analysis for Athens Simulation Models (THL)
Data Analysis for ML algorithms (SIMAVI)
Conceptual Models for Vienna (UPC)
Conceptual Models for Athens (PRACTIN)
Task 5.5: Decision Support & Assets management for Optimal CBDRM operations [M14-M32] Lead: FINT, Participants: THL, M3SB, PRACTIN, INTEROPT
Decision Support Tool for CBDRM operators (FINT)
what-if analysis for comparing alternative solutions (M3SB)
what-if analysis to measure the impact of data uncertainties on performance indicators (PRACTIN)
what-if analysis to analyse trade-off relationships of conflicting performance indicators (INTEROPT)
Data connector for Decision Support Tool (THL)
Task 4.5: VR, User interface and APIs for interoperability with other DRM systems [M24-M30] Lead: M3SB, Participants: THL, SIMAVI, INTEROPT
VR training tools (M3SB)
API connectors for DT (THL)
Online (web based) user interfaces (INTEROPT)
Incorporation of training, dissemination, KPIs, evaluation, decision support for optimal human and machine decisions and evaluation and risk analysis (SIMAVI)

Figure 2: The PANTHEON Architecture (THL, 2023)

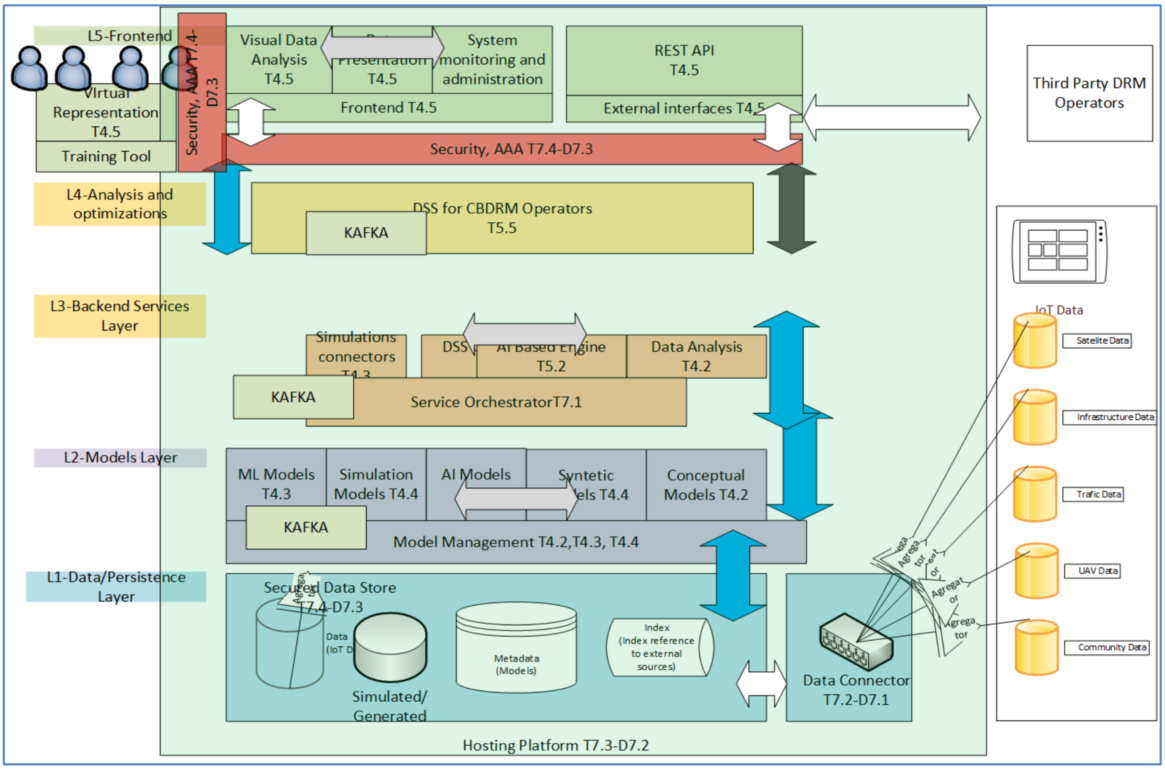


Figure 3: The Logical Architecture of PANTHEON (SIMAVI, 2023)

4.1 Data acquisition & Administration

Data is obtained from six distinct sources: (1) Satellites and Copernicus, (2) In-Situ IoT, (3) Infrastructure (e.g., electricity, telecommunications, and water maps), (4) Traffic, (5) UAVs, and (6) Community input. The management of data originating from multiple sources poses several challenges that must be addressed to ensure the quality of the outcomes required during the various phases of administration. These phases typically encompass Curation, Pre-processing, Post-processing, and Storage, each of which is described below:

1. **Curation** involves organizing and integrating data collected from various sources related to wildfires, floods, heatwaves, and earthquakes. For example, in the case of **wildfires**, data collected from satellites, UAVs, and ground-based sensors, it involves annotating wildfire boundaries, identifying fire severity levels, and publishing this information in a format that can be easily accessed and interpreted by stakeholders such as emergency responders and policymakers. Similarly, for **floods**, curation may involve organizing data on precipitation, river levels, and flood extents obtained from remote sensing platforms and in-situ sensors. This curated data helps in understanding flood dynamics and assessing potential impacts on communities. In the context of **heatwaves**, curation may entail compiling data on temperature, humidity, and heat indices from weather stations and IoT devices. This curated information assists in identifying heat-prone areas and implementing targeted interventions to protect vulnerable populations. For **earthquakes**, curation may involve integrating seismic data, ground motion recordings, and building vulnerability assessments. This curated data helps in assessing earthquake hazards and developing strategies for mitigating risks to infrastructure and communities. **Overall**, curation ensures that data relevant to wildfires, floods, heatwaves, and earthquakes is organized, annotated, and presented in a manner that facilitates its effective use for decision-making and disaster response.

2. **Pre-processing** is crucial for preparing raw data into a clean dataset suitable for analysis. This step ensures that the data is free of errors, inconsistencies, and missing values before it is fed into algorithms. There are four main steps involved in data preprocessing:
   a. Data Quality Management: This involves maintaining high-quality information from data acquisition to distribution. In the context of disaster management, this could mean ensuring that **wildfire** data collected from satellites, ground sensors, and UAVs is accurate and reliable, or verifying the integrity of **flood** data obtained from river gauges and weather stations.
   b. Data Cleansing/Validation: This step ensures that the data has undergone cleansing processes to confirm its quality and usefulness. For example, in the case of **heatwaves**, validation routines may check temperature data for accuracy and consistency, ensuring that it reflects actual environmental conditions.
   c. Data Transformation: This process involves converting, cleaning, and structuring data into a format suitable for analysis. In the context of **earthquakes**, data transformation may involve converting seismic readings and ground motion data into standardized formats that can be integrated with other datasets for analysis.
   d. Data Reduction: This optimization technique involves simplifying data to free up storage capacity. For instance, in **flood** management, data reduction may involve aggregating detailed rainfall data into summary statistics to reduce storage requirements while retaining essential information.
   e. Overall, dcuration and preprocessing are critical steps in PANTHEON, particularly in ensuring that data from various sources are *harmonized and standardized* for integration into the Smart-City Digital Twin. This facilitates the development of simulation models for disaster scenarios considered: wildfires, floods, heatwaves, and earthquakes.

3. **Processing** involves manipulating data using computer systems, encompassing tasks such as converting raw data into a machine-readable format, managing data flow through the CPU and memory, and formatting or transforming output. In the context of the PANTHEON project, two prominent processing methods are utilized: Batch Processing and Stream Processing.
    a. Batch Processing: This method involves the execution of data processing tasks in batches, where data is collected, processed, and outputted in discrete units. For example, in **wildfire** management, batch processing may be employed to analyze historical fire data collected over specific time intervals to identify trends and patterns in fire occurrence.
    b. Stream Processing: Stream processing involves the real-time analysis of data streams as they are generated. In **flood** monitoring, stream processing can be used to continuously analyze sensor data from river gauges to detect sudden changes in water levels indicative of potential flood events.

4. **Post-processing** occurs after primary data processing stages are completed and involves refining data to derive objective indicators and measures. In PANTHEON, post-processing enables the extraction of actionable insights and scenario results, as outlined in paragraph 4.3.3. For instance, after analysing **earthquake** simulation data, post-processing techniques may be applied to derive seismic risk assessments and vulnerability maps for affected areas.

5. **Storage** refers to the retention of data using various recording media and devices. Due to the substantial volume of data generated and utilized in PANTHEON, multiple repositories are employed:
    a. The first repository is the Partners repository (BOX), established by the coordinator for storing project deliverables.
    b. The second repository is the Zenodo repository of the CERN EOS Service, which boasts an 18 petabytes disk cluster with redundant copies of each file stored on different disk servers. This redundancy ensures data security and integrity, facilitating seamless retrieval and reuse for the project's duration and beyond.

4.1.1 Athens Demonstrator

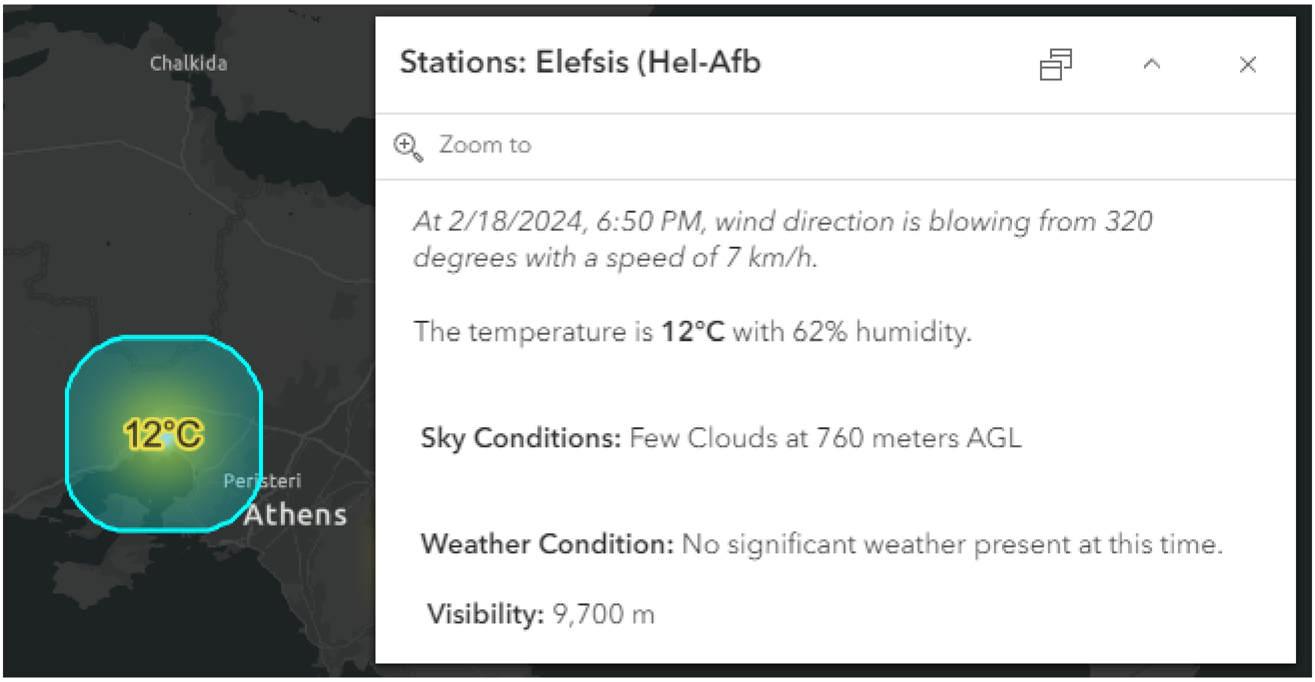
The PANTHEON tool will undergo deployment, testing and evaluation through scenarios: wildfire occurrence and earthquake events. As delineated in D3.2 *Report on Participatory Design,* two distinct applications labelled focusing on the pre-catastrophic phases of the disaster cycle: prevention and preparedness. The system will undergo testing across two primary dimensions (Spiteri, N.; Epsilon Malta, 2024):

1. Planning and Early Warning through Simulations: This aspect involves utilizing PANTHEON to simulate real-life events and assess planning and early warning mechanisms.
2. Training and Exercises: PANTHEON will be utilized for training and conducting exercises to enhance stakeholders' preparedness and response capabilities.

4.1.1.1 Wildfire Demonstrator

The **wildfire** scenario will serve as a basis to simulating real-life events through training and exercises. PANTHEON is designed to integrate various data sources to support stakeholders, improve implemented procedures, and enhance capabilities in wildfire prevention and preparedness, corresponding to the initial phase of wildfire management.

Of particular interest for the prevention phase are data sources such as weather. This information can be obtained remotely using weather satellites or from in-situ sources like meteorological stations. Numerous weather stations in Greece, including those in the Attica Region, are managed by organizations such as the Hellenic National Meteorological Service and the National Observatory of Athens. These stations provide real-time measurements of critical meteorological parameters essential for wildfire risk assessment, including wind direction and speed, humidity, and temperature. **Error! Reference source not found.** i llustrates an example of near real-time weather data for Athens, showcasing the type of information available for wildfire prevention efforts.



Stations: Elefsis (Hel-Afb

At 2/18/2024, 6:50 PM, wind direction is blowing from 320 degrees with a speed of 7 km/h.

The temperature is **12°C** with 62% humidity.

Sky Conditions: Few Clouds at 760 meters AGL

Weather Condition: No significant weather present at this time.

Visibility: 9,700 m

Figure 4: Near real-time weather data for Athens

Vegetation and land cover data play a critical role in wildfire risk assessment as they offer insights into the vegetation type and **fuel** availability within a given area. Information on vegetation type, density, and moisture content aids in estimating potential fuel loads and the rate of fire spread. By combining vegetation data with meteo-info, areas with a high likelihood of wildfire occurrence can be identified. Satellite data form the Copernicus Global Land Service provide valuable evidence on vegetation, dry matter productivity, and Other land cover and land use factors.

Fire propagation models utilize various data sources and algorithms to simulate and predict the behaviour and spread of fires under different conditions. Key data sources include fuel-soil characteristics, weather conditions, topography, and ignition points. By integrating these data, fire propagation models can effectively forecast the potential evolution of a fire, enabling stakeholders to prepare and take pre-emptive actions. These models facilitate proper allocation of operational resources.

Landscape visualization is crucial for understanding and mitigating wildfires, as terrain features such as slope, elevation, and topography significantly influence firefighting strategies and fire behaviour. A digital elevation model (DEM) provides essential information for assessing terrain characteristics. Figure 6 illustrates a three-dimensional example in Athens, where satellite imagery is overlaid with a DEM to visualize the landscape.

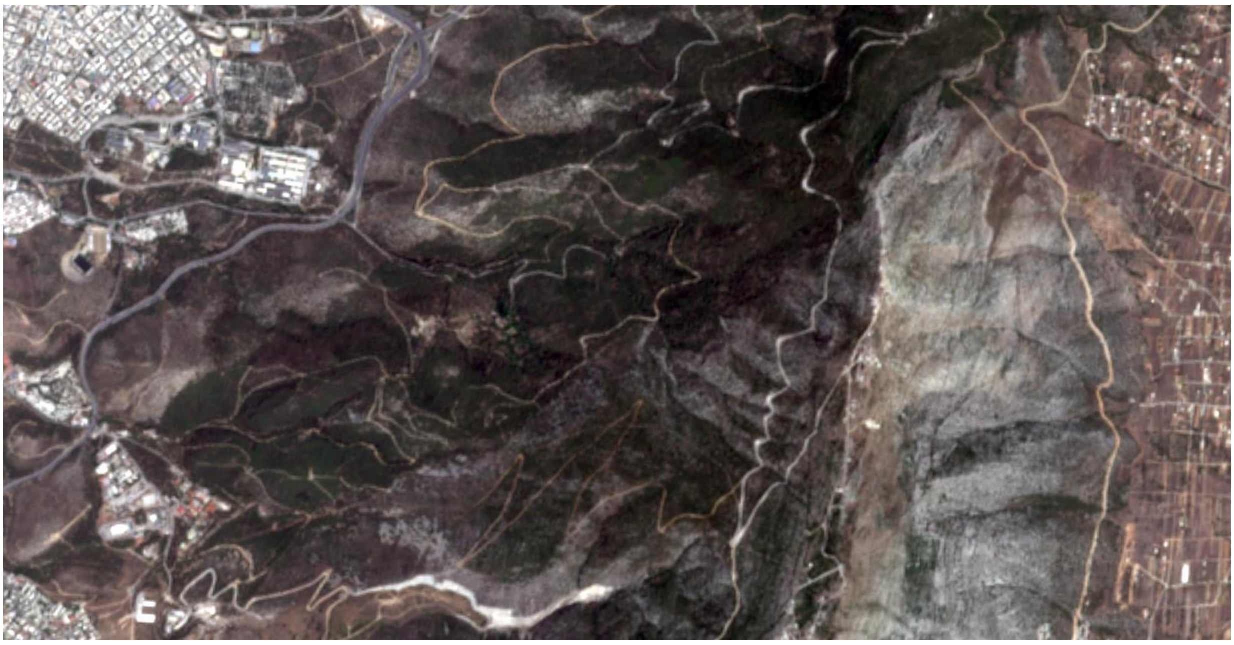


Figure 5: Satellite imagery from Landsat on vegetation and land cover of Attica (Epsilon, 2024)

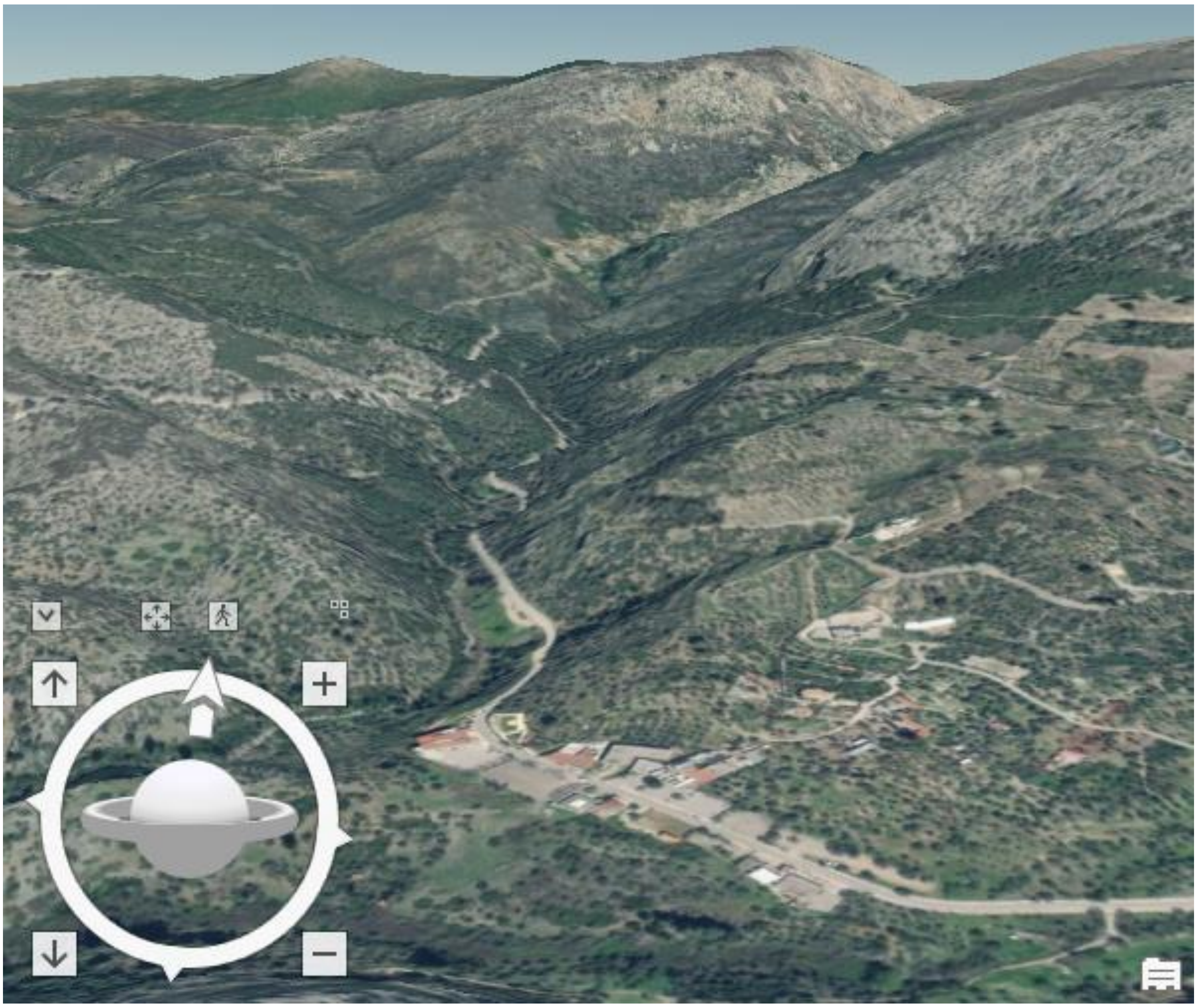


Figure 6: Athens DEM from satellite imagery (Epsilon, 2024)

Historical Fire Data: Historical weather data are crucial as they provide insight into past fire behaviour under specific meteorological conditions. Data on past wildfire events, including occurrence, size, and behaviour, can offer valuable insights into fire patterns and aid in identifying high-risk areas.

In-situ Sensor Data: Ground-based sensors, such as fire danger rating systems, provide real-time information on fire conditions, aiding in early detection and warning.

Info on Critical Infrastructures (CI): In addition to factors directly related to wildfire occurrence, stakeholders require access to information about critical infrastructures (CIs), including their exact location, interdependency with other infrastructures, and potential impact in case of disruption. Data related to transportation, energy, traffic, and mapping of natural gas pipelines or power lines are important considerations.

Integration with Legacy Systems/Platform: PANTHEON SCDT should seamlessly communicate and exchange real-time information with existing tools used by first responders. For example, the Fire Hub service used by the Hellenic Fire Service provides real-time wildfire information and a GIS-based platform with historical event data. The system should facilitate bidirectional communication with other tools, such as the ENGAGE IMS/CAD solution used for incident management. Additionally, social media crawling and monitoring citizen posts on webpages and social media platforms can serve as a significant source of real-time information about ongoing events.

Information on Operational Resources: Stakeholders, particularly first responders in Command & Control (C2) centers, require data on operational personnel, vehicles, and resources available at all times. This information enables accurate decision-making regarding the mobilization of forces to respond to wildfire events.

Infrared Data: UAVs equipped with infrared cameras can detect and map active hotspots during nighttime firefighting operations without disrupting daytime efforts. These drones should capture GPS metadata and have known camera specifications, enabling the use of GIS software to generate orthophotos and digital surface models for extended analysis capabilities.

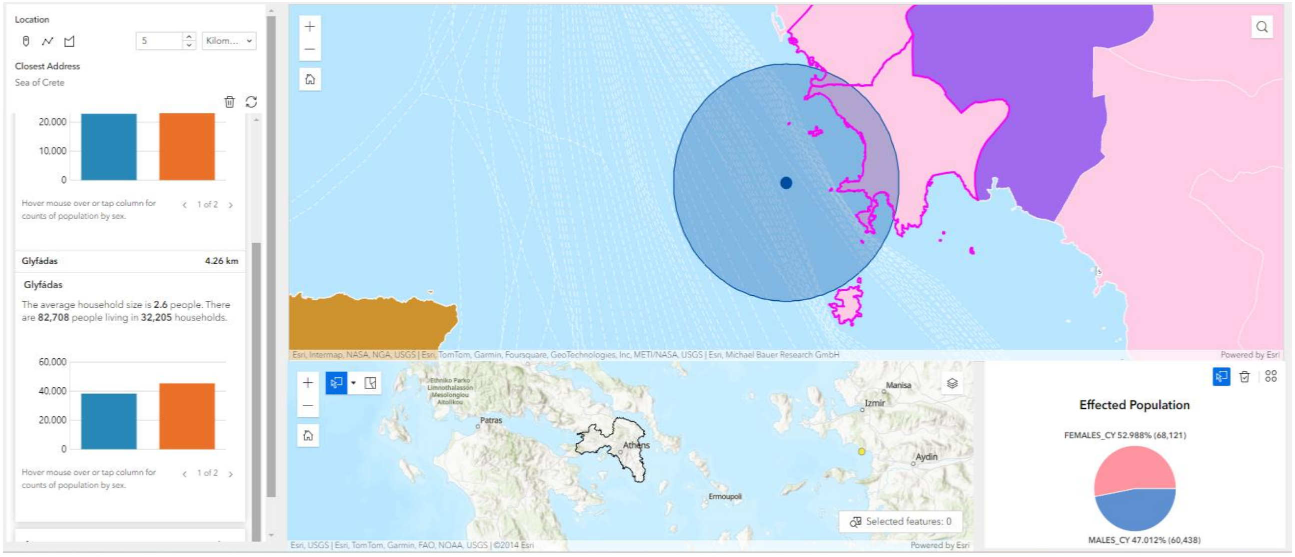
Other Data: Demographic and urban planning data, such as population density, vulnerable citizens, spatial plans, evacuation routes, and open spaces, are crucial sources of information to be considered and integrated within the system. This ensures that disaster management stakeholders are better prepared in the event of a large wildfire.

4.1.1.2 Earthquake Demonstrator

The earthquake demonstrator serves as the foundation to evaluate PANTHEON as an early warning tool for aid stakeholders in *planning & preparedness* to managing incidents. Earthquakes pose significant challenge, particularly during the *prevention & preparedness* phases of the disaster management cycle, primarily due to the limited competence for seismic prediction. While maps indicating probability of earthquake occurrence and timeframes exist, the ability for precise prediction remains limited. Below are outlined data sources that should be integrated into PANTHEON to enhance earthquake planning:

Statistical data on buildings and their attributes: The *Hellenic Statistical Authority* serves as the primary source of building-related data in Greece, offering information on factors such as the number of buildings in specific areas. This data includes details on structural characteristics like building materials, number of floors, roofing materials, year of construction, and usage. Additionally, upon request, the Authority can provide supplementary data such as building density per city block.

Demographics & spatial data: Similar to the wildfire scenario, demographic data plays a vital role in informing disaster management experts, first responders, and urban planners. Demographic data primarily pertains to population figures, including numbers and densities, while spatial data encompasses information on open spaces and evacuation routes. This information aids civil protection experts in developing effective evacuation and sheltering plans.

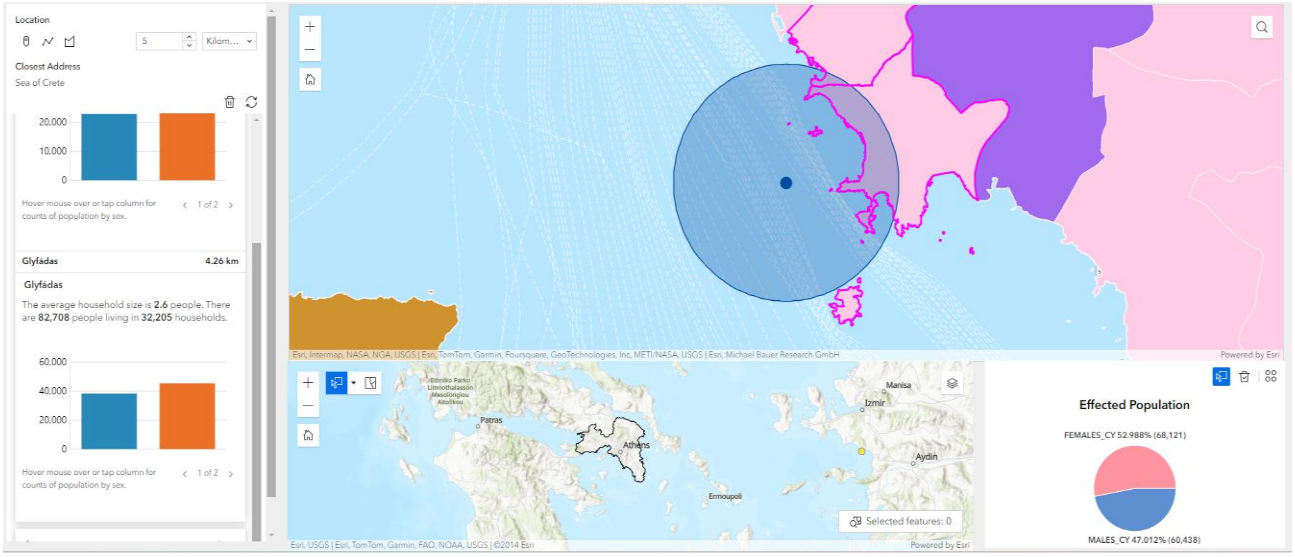

Location
5
Kilom...
Closest Address
Sea of Crete
20,000
10,000
0
1 of 2
Glyfádas
4.26 km
Glyfádas
60,000
40,000
20,000
0
1 of 2
Esri, Intermap, NASA, NGA, USGS | Esri, TomTom, Garmin, Foursquare, GeoTechnologies, Inc, METI/NASA, USGS | Esri, Michael Bauer Research GmbH
Powered by Esri
Ethniko Parko Limnothalasson Mesolongiou Aitolikou
Patras
Athens
Manisa
Izmir
Aydin
Ermoupoli
Selected features: 0
Esri, USGS | Esri, TomTom, Garmin, FAO, NOAA, USGS | ©2014 Esri
Powered by Esri
Effected Population
FEMALES_CY 52.988% (68,121)
MALES_CY 47.012% (60,438)

Hover mouse over or tap column for counts of population by sex.

The average household size is 2.6 people. There are 82,708 people living in 32,205 households.

Hover mouse over or tap column for counts of population by sex.

Figure 7 illustrates population impacted by an earthquake in Athens.


Location
5
Kilom...
Closest Address
Sea of Crete
20,000
10,000
0
1 of 2
Glyfádas
4.26 km
Glyfádas
60,000
40,000
20,000
0
1 of 2
Esri, Intermap, NASA, NGA, USGS | Esri, TomTom, Garmin, Foursquare, GeoTechnologies, Inc, METI/NASA, USGS | Esri, Michael Bauer Research GmbH
Powered by Esri
Ethniko Parko Limnothalasson Mesolongiou Aitolikou
Patras
Athens
Manisa
Izmir
Aydin
Ermoupoli
Selected features: 0
Esri, USGS | Esri, TomTom, Garmin, FAO, NOAA, USGS | ©2014 Esri
Powered by Esri
Effected Population
FEMALES_CY 52.988% (68,121)
MALES_CY 47.012% (60,438)

Hover mouse over or tap column for counts of population by sex.

The average household size is 2.6 people. There are 82,708 people living in 32,205 households.

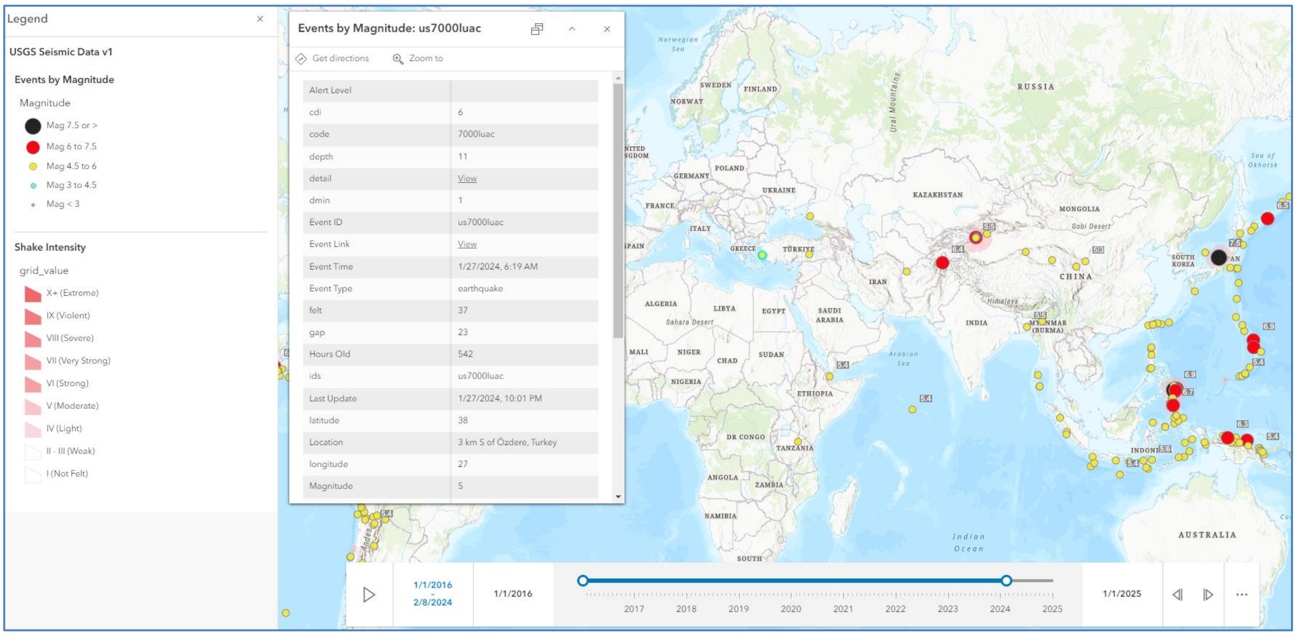
Hover mouse over or tap column for counts of population by sex.

Figure 7: Population affected by an earthquake in Athens (Epsilon,. 2024)

Earthquake data. Seismicity can be classified into two categories: (1) historical seismicity pertains to events predating 1900 and relies on historical sources, primarily comprising macro-seismic intensity data; (2) instrumental seismicity encompasses seismic events occurring after 1900, leveraging measurements from seismographs and accelerographs. These instruments provide crucial seismic indices and factors, including macro-seismic intensity data. Various organizations, such as the National Observatory of Athens, the Departments of Geophysics at the National and Kapodistrian University of Athens, the Aristotle University of

Thessaloniki, the University of Patras, and the Earthquake Planning and Protection Organisation, furnish data on past events. This data encompasses essential factors such as magnitude, focal depth, epicenter location, earthquake generation mechanisms, and shake maps detailing indices like peak ground acceleration, peak ground velocity, instrumental intensity, and macro-seismic intensity. Armed with this information, stakeholders can gain a comprehensive understanding of the seismic history of a particular area, evaluate the potential for similar or even more intense future events, and devise appropriate plans accordingly.

Seismicity can be divided in two categories: (1) historical seismicity[35] which refers to past events prior to 1900 and relies on historical sources and consists mainly of macro seismic intensity data, and instrumental seismicity[36]. Data that refers to seismic events after 1900, rely on measurements from seismographs and accelerographs, which provide data regarding crucial seismic indices and factors. Macro-seismic intensity data are also available. Different entities, such as the National Observatory of Athens, the Departments of Geophysics of the National and Kapodistrian University of Athens, the Aristotle University of Thessaloniki, the University of Patras, and the Earthquake Planning and Protection Organisation, provide data related to past events. These data include crucial factors i.e., magnitude, focal depth, epicentre location, earthquake generation mechanisms, shake maps that include indices such as peak ground acceleration, peak ground velocity, instrumental intensity, and macro-seismic intensity. Through this information, stakeholders can have a clear picture and understanding of the seismic past of a specific area, assess the potential for similar or even stronger future events, and apply their plans accordingly. Above data are mapped on a GIS to provide a visual understanding of the areas which are more prone to seismic activity. A time slider can be added to the map to give users the possibility to filter data by time easily and efficiently (Figure 8).



Events by Magnitude: us7000luac

Alert Level	
cdi	6
code	7000luac
depth	11
detail	View
dmin	1
Event ID	us7000luac
Event Link	View
Event Time	1/27/2024, 6:19 AM
Event Type	earthquake
felt	37
gap	23
Hours Old	542
ids	us7000luac
Last Update	1/27/2024, 10:01 PM
latitude	38
Location	3 km S of Özdere, Turkey
longitude	27
Magnitude	5

Figure 8: Seismic Centres on GIS (Epsilon, 2024)

Geological Data: The Hellenic Authority of Geological and Mineral Surveys, along with other relevant organizations, can furnish geological maps detailing the geological and tectonic characteristics of a region. Additionally, national-level maps depicting active seismicity provide insights into geological formations and active faults capable of triggering earthquakes.

Information on Critical Infrastructures (CI): Analogous to scenarios involving wildfires, data pertaining to CI are of importance and should be integrated into the Scenario Coordination and PANTHEON Decision Tool

(SCDT). Precise CI locations, interdependencies among different infrastructures, and estimates regarding the number of households affected in case of disruptions are crucial for all stakeholders involved.

Geo-tagged Aerial Cartography: Unmanned Aerial Vehicles (UAVs) can be deployed to synchronize current aerial imagery with geographical coordinates, facilitating precise mapping of land features.

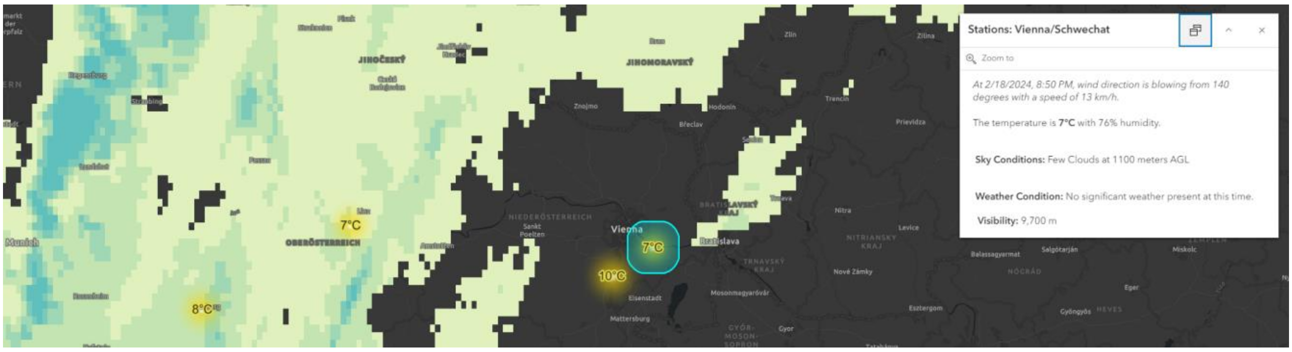
Information on Operational Resources: Comprehensive data concerning available operational vehicles and resources such as UAVs and ambulances play a pivotal role in enhancing coordination during response operations.

4.1.2 VIENNA DEMONSTRATOR

The Vienna Demonstrator considers: (1) Heatwaves, and (2) City fires.

4.1.2.1 Vienna Heatwaves

Heatwaves occur during the summertime, and a significant part of the resources is going to be either unavailable (personnel vacations) or practically inactive (ambulances which cannot be moved due to lack of the appropriate personnel). In this case the simulation of a disaster scenario is rather straight forward because all the involved parameters are known in advance. As such, we can make the scenario as strict as possible, based on the worst-case consequences anticipated, to measure the reaction of the first responders. Again, GIS can help in this scenario by providing all the necessary information available at near to real time, as weather data from satellite imagery (Figure 9).



At 2/18/2024, 8:50 PM, wind direction is blowing from 140 degrees with a speed of 13 km/h.

The temperature is **7°C** with 76% humidity.

Sky Conditions: Few Clouds at 1100 meters AGL

Weather Condition: No significant weather present at this time.

Visibility: 9,700 m

Figure 9: Weather data from satellite imagery, Vienna

4.1.2.2 Vienna wildfire caused by a cyber-terrorism attack

Vienna also faces the potential risk of wildfires triggered by man-made events, such as a "cyber-attack" targeting critical infrastructure like a power plant. A cyber-attack on a power plant could ignite a fire capable of spreading rapidly through the surrounding forested areas, including the outskirts of Vienna. This scenario highlights the interconnectedness of infrastructure vulnerabilities and environmental risks, underscoring the need for comprehensive disaster preparedness and response strategies.

By selecting this scenario for simulation exercises, stakeholders can evaluate and enhance their readiness to address both natural and man-made disasters effectively. The choice of a cyber-terrorism allows for the exploration of cascading effects and interagency coordination in response to unexpected events. Through collaborative training exercises involving first responders, healthcare providers, law enforcement agencies, municipal authorities, infrastructure providers, and media representatives, stakeholders can improve their coordination and cooperation in managing complex disaster scenarios.

While the initial focus may be on heatwaves and cyber-triggered wildfires, the development of final scenarios will be contingent on the complexity of the simulations and the available project timeline. Deliberate consideration of cascading effects and the potential exclusion of certain elements may be necessary to ensure the feasibility of the pilot project in Vienna within the project's timeframe. Ultimately, the aim is to leverage the digital twin platform to strengthen preparedness, response, and resilience across various disaster scenarios in Vienna and beyond.

Note: As of today, there haven't been any reported instances of a cyber-attack directly triggering a wildfire or fire in Vienna. However, the scenario might be plausible, highlighting potential risks associated with attacks on critical infrastructure and their cascading effects on the environment and public safety. Thus, the risk of such an event cannot be completely discounted and will be analysed evaluated by PANTHEON.

4.2 Data Processing

4.2.1 Types of data

PANTHEON can leverage two primary methodologies on data processing: (1) batch and (2) stream. Each methodology possesses distinct characteristics and is suited for handling different types of data, such as volume and velocity considerations. For instance, PANTHEON may employ batch processing tools, methodologies, and workflows to execute machine learning (ML) algorithms to analysing historical data in batches. This approach could involve GIS, weather data, or other input streams, allowing for the identification of trends and patterns that can support disaster simulation. Conversely, PANTHEON could utilize stream data processing tools and workflows to handle real-time data streams, such as those from Copernicus or UAV sources. In this scenario, ML algorithms will analyse data in real-time, providing valuable insights to feed into the Scenario Coordination and Decision Tool (SCDT) simulation models and decision support components.

4.2.1.1 Batch Processing

Batch processing serves as a foundational technique in data processing, particularly for efficiently managing large datasets. The method involves executing repetitive data tasks in bulk on a scheduled basis. Tasks such as backups, filtering, and sorting can be computationally intensive if processed individually. Therefore, data systems handle these tasks collectively in batches, typically during periods of reduced computational demand such as during off-peak hours or overnight. The typical flow of batch processing is illustrated in Figure 10.

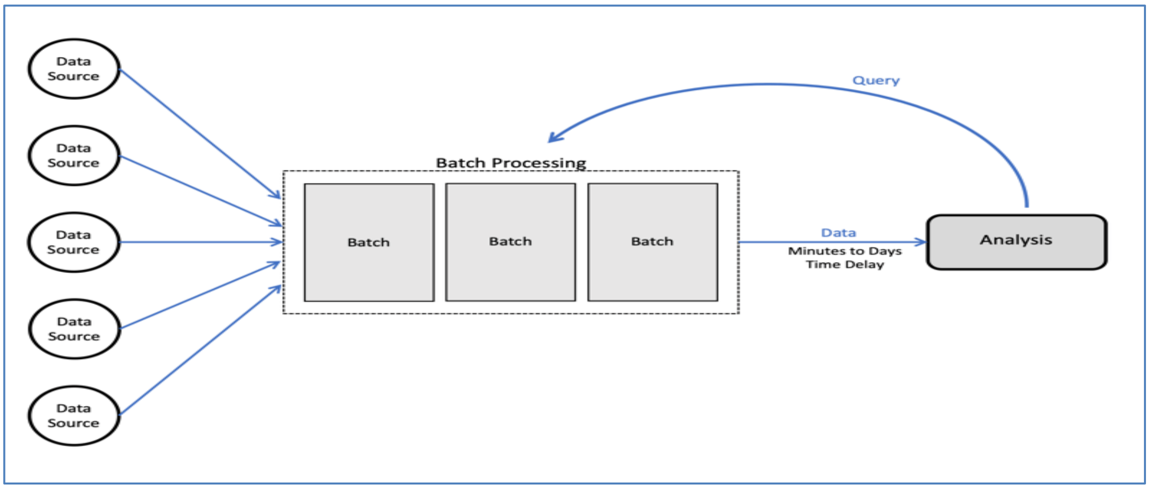


Figure 10: Batch Processing flow [37]

4.2.1.2 Tools & Frameworks

In recent years, batch processing tools evolved significantly delivering a range of functionalities from distributed computing frameworks to workflow orchestration platforms. **Apache Hadoop**[38] is considered a prominent tool in batch processing since its distributed file system (HDFS) and its MapReduce programming model enables the processing of large datasets across a cluster of computers. It has been widely adopted in several industries for its scalability and fault tolerance[39].

Another prominent tool is **Apache Spark[40]** which supports in-memory processing and facilitates the development of complex data processing workflows[41]. **Apache Flink[42]** is another framework that provides features such as event-time processing and exactly once semantics. It can handle both real-time (see next section) and batch data processing tasks[43]. **Apache Airflow[44]** is an open-source platform used to configure complex business processes. The defined workflows need to be Directed Acyclic Graphs (DAG) which enables the definition and execution of workflows with dependencies and scheduling capabilities. Thus, has gained popularity for its flexibility and extensibility[45].

Another tool known for its distributed SQL query engine, **Presto[46]** is employed for interactive analytics and batch processing. It efficiently processes large datasets stored in various data sources, including Hadoop Distributed File System (HDFS), **Apache Cassandra**. Presto's ability to perform federated queries across multiple data stores enhances its versatility[47]. **Apache Storm[48]** has evolved as a distributed stream processing framework to support both stream (see next section) and batch processing. Recognized for its low-latency capabilities, Storm is ideal for workloads requiring real-time data processing[49]. It provides a fault-tolerant and scalable solution, making it suitable for diverse use cases[50].

Apache Beam[51] takes a unified approach to batch and stream processing, by offering portability across different runtimes and simplifying the development of data processing pipelines. It provides a high-level API, enabling users to express their data processing logic concisely[52]. **Apache NiFi[53]** focuses on data integration and workflow automation and excels in orchestrating and managing data flows. Its user-friendly interface facilitates the design of complex data workflows, making it a valuable tool in modern data architectures[54]. NiFi's capabilities extend to both batch and streaming scenarios, providing flexibility in handling diverse data processing requirements.

4.2.1.3 Batch data processing methodologies

Batch data processing methodologies evolved significantly, benefiting from new techniques such as Extract, Transform, Load (ETL) and MapReduce to increase data processing efficiency and scalability. These two techniques are briefly described below.

- The **ETL (Extract, Transform, Load)** process has three main steps; (1) data is collected from a variety of sources, (2) data is modified to meet the needs of the research. The modified data is transferred to the specified system. ETL processes have been shown to improve data quality and accuracy as noted in [3]. Several works [4] have highlighted the importance of productivity in ETL processes which reduces the need for manual intervention, improves accuracy and increases overall productivity. Additionally, one of the most important aspects of ETL is data quality assurance. In [5] the author emphasises that ETL processes use reliable techniques to clean and validate data, ensuring accuracy and reliability for subsequent analysis.

- Google introduced **MapReduce**, a programming model for processing and generating extensive datasets. By breaking down tasks into smaller tasks, processing concurrently, and combining results, MapReduce dramatically increases data processing speed and scalability [6]. Notably, MapReduce has fault tolerance as highlighted in [7]. To ensure the reliability of the quantitative data processing, the MapReduce algorithm provides a fault-tolerant reconfiguration in case of node failures. A well-known framework that uses MapReduce is Hadoop.

4.2.1.4 Best Practices

Batch data processing incorporates a variety of best practices designed to improve performance, accuracy, and scalability across large amounts of data. An important factor is the adoption of **automated data processing workflows**, which increase the overall efficiency of batch processes [3]. Automation reduces risk associated with manual processes, reduces errors, and increases repeatable capabilities [4]. It is considered important to ensure **data quality**, with best practices encouraging comprehensive data cleaning and validation processes during the extraction, transformation, and loading (ETL) process [5]. Maintaining high levels of data quality facilitates downstream analysis and decision making. In addition, **parallel processing** is considered a good practice, especially in the case of MapReduce methods [6].

Parallel processing, achieved through mapping and step reduction, makes it easier to distribute computations across clusters, greatly increasing processing speed and scalability. **Fault tolerance** is an important factor in batch data processing, especially the use of MapReduce techniques [7]. MapReduce design principles include fault-tolerant mechanisms, which ensure that the process is repeated in the event of node failure. This approach increases the reliability of large data processing systems. Ongoing research focuses on continuously optimising and improving batch data processing techniques, solving performance challenges, and exploring algorithmic improvements [8].

4.2.1.5 Stream Processing

This subsection offers a thorough overview of tools, methodologies, best practices, frameworks, and other pertinent aspects associated with real-time processing. Stream processing entails the immediate analysis and management of data as it is generated, as depicted in Figure 11. Stream processing implies operating within predefined and non-negotiable time constraints, analysing data upon its arrival, and being characterized as live analysis. This live analysis can be performed in real-time to avoid jeopardizing other processes by consuming excessive computing resources.

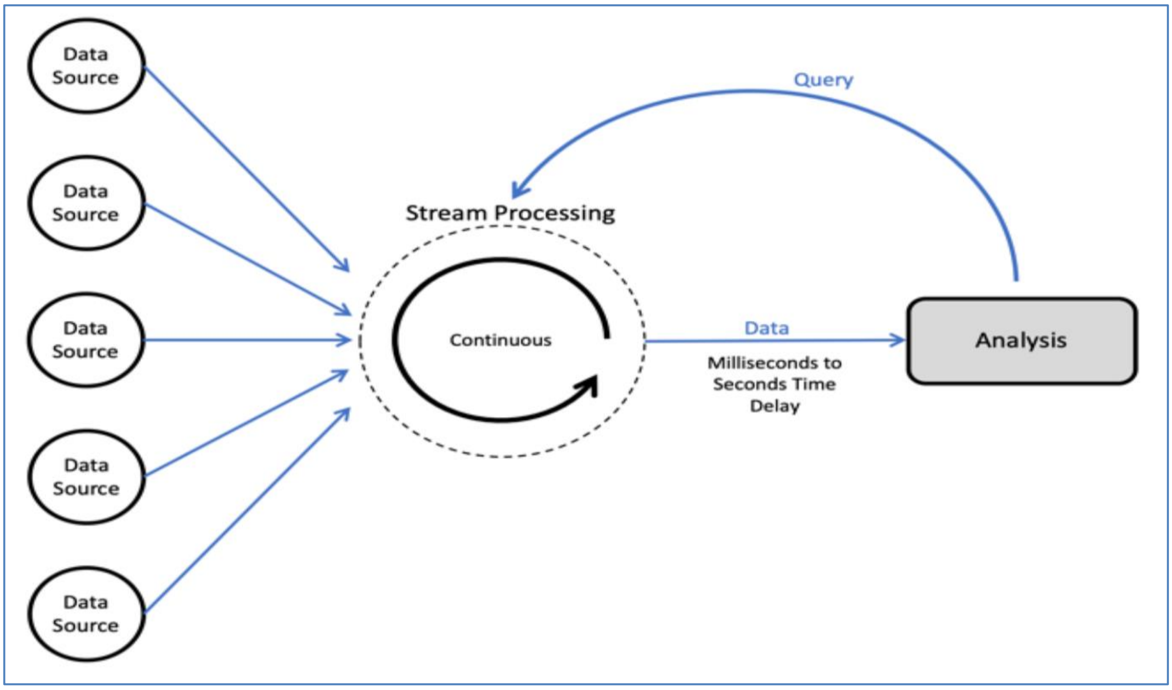


Figure 11: The Stream Processing flow [55]

4.2.1.6 Stream Processing Methodologies

Stream processing methodologies play an important role in real-time data handling:

- **Lambda Architecture** proposed in [14], combines batch and stream processing to ensure accuracy and fault tolerance. Lambda Architecture consists of three layers: Batch Layer, Speed Layer, and Serving Layer. The Batch Layer handles historical data, performing complex computations and generating batch views. The Speed Layer processes real-time data, producing incremental updates. The Serving Layer merges results from both layers, ensuring a unified view. This dual-layer architecture provides fault tolerance and enables analytics at scale [12], [15].
- **Microservices architecture** further enhances real-time capabilities by breaking down applications into modular, independently deployable services. Another notable methodology in stream data processing involves leveraging **Data Stream Processing** (DSP) techniques. DSP focuses on real-time data processing, employing specific methodologies for capturing and analysing relevant data in a streaming fashion [16]. Real-time analytics, an integral part of DSP, involves the immediate extraction of insights from moving data, ensuring timely decision-making.

4.2.1.7 Best Practices in stream processing

Achieving high efficiency and stability in stream processing requires adherence to proven methods. By applying event-driven architectures, where actions are triggered by immediate events, one can significantly improve responsiveness [17]. Additionally, incorporating Continuous Integration and Continuous Deployment (CI/CD) principles guarantees rapid and accurate deployment of real-time applications [18]. The dynamic nature of stream data processing involves continuously examining rapidly produced data, requiring meticulous methodologies to extract valuable insights. In today's stream data processing, the significance of **real-time analytics** cannot be overstressed. Streams demand swift action to support timely decision-making. To tackle the temporal complexities inherent in streaming data analysis, approaches such as event time processing and watermarking have been devised [19-20]. **Machine learning** plays an important role in stream data analysis. An advanced algorithm enables real-time detection of patterns and anomalies, increasing the predictive capability of the system [21]. In addition, integrating edge computing is essential to ensure efficient data handling close to the data source, reduce latency, and for bandwidth utilisation [22]. Robust fault handling mechanisms and inspection techniques improve system stability and reduce the impact of failures [19]. Furthermore, scalability is achieved through parallelization and distributed processing, including accommodating ever-increasing volumes of streaming data. Given the sensitivity of real-time data, security is a key and must be considered in streaming data processing. Encryption and authentication measures comply with modern cybersecurity standards, preserving the integrity and confidentiality of transmitted data.

4.2.2 Data Processing Technologies

4.2.2.1 Cloud-Based Big Data Processing

Within PANTHEON, technical partners will harness cloud resources to expedite the software development and deployment lifecycles. Within the domain of data processing, leveraging cloud-based big data processing has emerged as a fundamental strategy for efficiently and economically managing large datasets. This section delves into the essential technologies, tools, methodologies, techniques, mechanisms, and frameworks deployed in cloud environments for processing big data. These resources could be leveraged by technical partners throughout the development and deployment phases of the PANTHEON platform.

4.2.2.2 Cloud Processing Technologies

In the last decade cloud data processing technologies have rapidly evolved, with the tools and frameworks mentioned in 4.2.1.1 and 4.2.1.2 being incorporated in cloud providers' suites of tools. A common technology found in cloud environments for data management is distributed storage. Technologies such as Apache Hadoop Distributed File System (HDFS) and Apache Cassandra provide fault-tolerant storage at scale [23].

Object storage services such as Amazon S3 and Google Cloud Storage provide storage for large datasets with durability and accessibility. Another technology is cloud computing services.

Cloud providers offer advanced computing services optimised for big data processing. Apache Spark, a distributed computing framework, is widely accepted for in-memory processing and analytics and offers high speed and ease of use [23]. Additionally, cloud-native services such as Amazon EMR, Google Cloud Dataproc, Amazon EC2, Google Compute Engine, or Azure Virtual Machines simplify data system deployment and management and provide dedicated computing services for large data processing in parallel [24].

Another emerging technology in cloud environments is serverless computing. This technology allows developers to focus on code without maintaining the underlying infrastructure. Technologies such as AWS Lambda and Azure Functions enable event-driven computing, automatically scaling resources based on demand [23]. Serverless architectures increase the speed and cost-effectiveness in processing variable workloads [25]. Additionally, common technologies entail data scheduling and workflow management. Apache Airflow and Apache NiFi are key components for orchestrating robust data workflows in cloud-based big data applications. They enable data migration, transformation and scheduling efficiency, ensuring consistency with data sources and destinations [23-25].

Finally, contemporary tools include advanced analytics and machine learning. Cloud providers offer managed services for advanced analytics and machine learning. Google Cloud's BigQuery ML and Amazon SageMaker simplify the development and deployment of machine learning models on large datasets [24]. TensorFlow and PyTorch are widely used frameworks for distributed machine learning training on cloud platforms [25].

4.2.2.3 Cloud Tools & Frameworks

Cloud-based data processing tools essentially encompass those previously mentioned, tailored and integrated for cloud environments. Below is a non-exhaustive list of several tools. One prominent cloud-based data processing tool is Apache Spark on Cloud. Utilizing Apache Spark on cloud infrastructures (e.g., Amazon EMR, Google Dataproc) enables distributed data processing with its powerful in-memory computation capabilities. Additionally, cloud-based Hadoop offerings (e.g., Amazon EMR, Azure HDInsight) facilitate distributed processing of large datasets using tools from the Hadoop ecosystem, including MapReduce, Hive, and Pig.

Furthermore, technologies like Apache Kafka and cloud-native solutions (e.g., AWS Kinesis, Azure Stream Analytics) support real-time stream processing, crucial for time-sensitive applications. Cloud-based machine learning frameworks, such as TensorFlow on platforms like Google AI Platform, enable scalable training and deployment of machine learning models. Lastly, serverless frameworks (e.g., AWS Serverless Application Model, Azure Functions) abstract infrastructure management, allowing developers to focus on writing code, thus enhancing productivity.

4.2.2.4 Edge Data Processing

While PANTHEON's primary focus lies outside of edge data processing, it may prove advantageous for certain PANTHEON partners contributing data to the platform to explore and employ specific techniques, tools, or frameworks tailored for edge data processing. This approach aims to optimize the format, structure, and volume of data transmitted over the Internet, thereby minimizing resource utilization and enhancing transmission rates. Edge processing stands as a pivotal component within modern data processing technologies, decentralizing computational tasks and reducing latency by processing data in close proximity to its source. This section provides an overview of key elements, techniques, and technologies intrinsic to edge data processing, offering insights into their potential applicability within the context of PANTHEON.

Although PANTHEON is not focused on edge data processing, it could be beneficial for specific PANTHEON partners that offer data to the PANTHEON platform, to consider and utilise specific techniques, tools, or frameworks for edge data processing to optimise the format, shape and volume of data transmitted through the Internet, minimising resource utilisation and speeding up transmission rates. Edge processing is a pivotal aspect of modern data processing technologies, decentralising computational tasks, and reducing latency by processing data closer to the source.

4.2.2.5 Edge data processing Technologies, Tools, Frameworks

Several technologies contributed to edge data processing within the last decade, with the technological area witnessing big technological advancements. Notably, the proliferation of Internet of Things (IoT) devices has necessitated green and localised data processing at the network's far edge and numerous technologies contribute to this paradigm shift.

Edge computing leverages the abilities of devices closer to data sources, lowering latency and improving responsiveness. Fog computing extends this concept by introducing intermediary nodes between edge devices and centralised cloud servers, fostering a hierarchical approach to data processing. In this landscape, technologies such as Apache Kafka[56] and MQTT (Message Queuing Telemetry Transport)[57] serve as robust messaging protocols, ensuring seamless communication between edge devices and central servers. These protocols prioritise low latency, a crucial requirement for time-sensitive applications. Another noteworthy technique is edge analytics, which involves processing data on local devices or gateways. This approach reduces the reliance on centralised cloud servers, offering benefits (e.g., lower latency, improved privacy).

Studies indicate the increasing adoption of edge analytics in applications such as video surveillance and industrial IoT. Platforms such as Microsoft Azure IoT Edge and AWS IoT Greengrass empower edge devices to perform analytics locally, optimising bandwidth usage. Microservices architectures, implemented through tools such as Docker[58], k3s[59] and Kubernetes[60], enable modular and scalable edge applications. This facilitates the deployment and management of containerized services, promoting flexibility and ease of integration.

As a commercial example, Microsoft's Azure IoT Edge[61] framework empowers developers to deploy containerized applications to edge devices seamlessly. Machine learning (ML) at the edge (Edge AI) has gained prominence, with frameworks such as TensorFlow Lite[62] and PyTorch[63] providing lightweight implementations for edge devices. This empowers devices to perform inferencing locally, reducing the need for constant communication with central servers. Edge databases, exemplified by Amazon DynamoDB[64] and SQLite[65], meet the unique challenges of constrained environments. These databases are designed for efficient storage and retrieval, ensuring optimal performance on resource-limited edge devices.

Moreover, edge orchestration frameworks such as OpenStack[66] and Apache OpenWhisk[67] streamline the management of edge resources. These frameworks enable the dynamic allocation and deallocation of computing resources, adapting to the fluctuating demands of edge applications. Furthermore, the emergence of 5G/B5G technology has extensively impacted edge data processing capabilities. The high data transfer rates and low latency of 5G/B5G networks enable more efficient communication between edge devices, enhancing overall performance of edge computing solutions. Finally, research has explored the integration of serverless computing at the edge, contributing to resource optimization and efficient execution of functions on edge devices.

4.2.3 Data Integration & Fusion

4.2.3.1 Integrating Multisource Data

In the framework of the PANTHEON Smart City Digital Twin environment, multi-source data integration encompasses the process of amalgamating and harmonizing diverse and heterogeneous data from various input sources within the project. This amalgamation aims to construct a unified and comprehensive representation within the digital twin, mirroring the physical attributes, processes, and behaviors of the target cities, Athens and Vienna, in real-time. The integration of data from multiple sources is pivotal in crafting an accurate and holistic digital twin, thereby facilitating simulations and informed decision-making to augment overall urban sustainability.

Regarding sensor networks and in-situ IoT devices, data integration entails consolidating data from an array of sensors dispersed throughout the target cities. These sensors encompass environmental, temporal, traffic, and infrastructure monitoring sensors. Through data integration, real-time insights into city behavior and operations can be gleaned. Simultaneously, the integration of geospatial data, including maps, satellite imagery, data captured from UAV sensors, and GIS layers, enables the spatial representation of the cities within the digital twin model. This integration furnishes the architecture with precise positioning and movement data, fundamental for comprehending mobility patterns and transportation systems.

Furthermore, augmenting the aforementioned architecture involves the integration of real-time data streams from social sources, emergency and event monitoring public services. This integration empowers the digital twin to capture individual citizen or community engagement, facilitating social planning features and enabling dynamic responses to unfolding events and disasters with heightened transparency.

The integration of multi-source data can be classified into two main approaches: (1) the data warehouse approach and (2) the mediator approach as illustrated in Figure 12 (Xie et al., 2022).

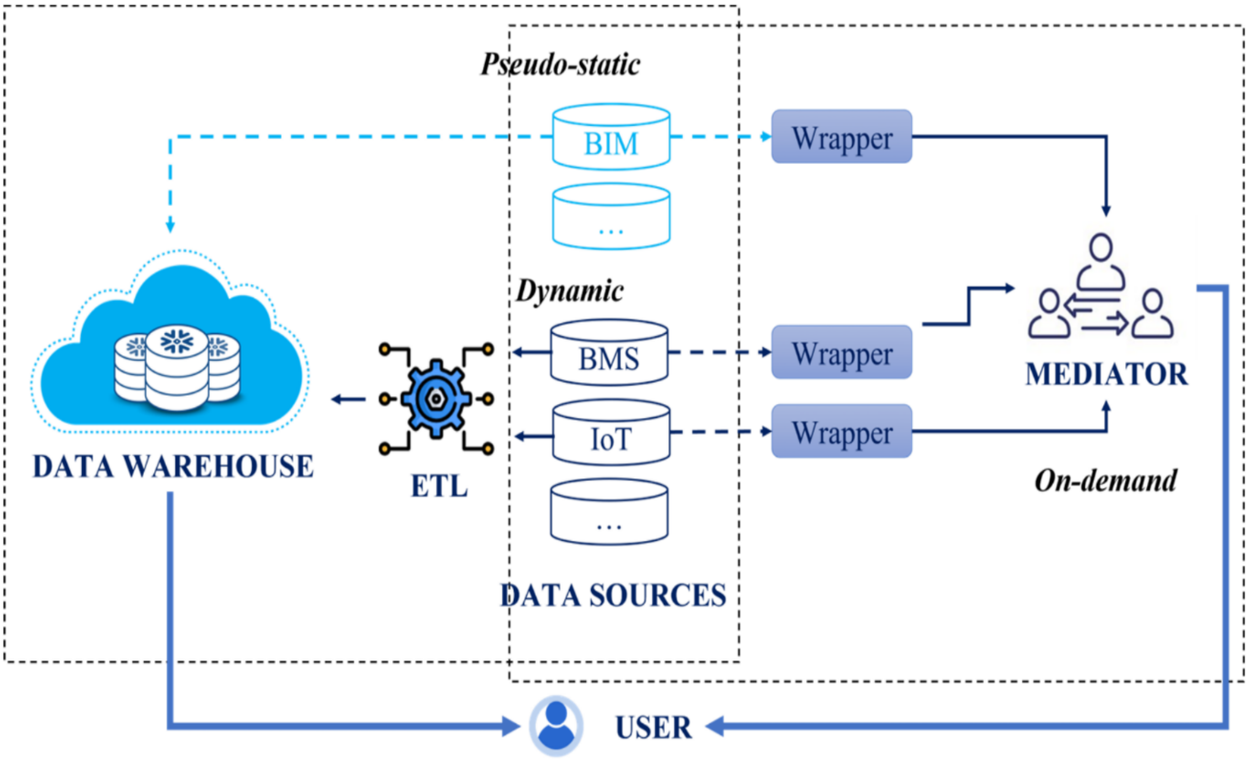


Figure 12: Integration approaches of multi-source data (Xie et al. [26])

The **data warehouse approach** involves consolidating all data sources into a centralized data warehouse while implementing fusion, cleansing, customization, and reformatting operations on the data. This approach offers the advantage of readily accessible and semantically consistent data, thanks to the fusion and cleansing operations performed before storing the data in the warehouse. However, there are instances where ingesting all data into a persistent location is impractical due to factors such as ownership issues, increasing data size, storage overhead, and scaling challenges. To address these obstacles, the mediator approach can be employed.

In contrast to integrating all data into a persistent location, the **mediator approach** enables fetching necessary data directly from their original sources using appropriate wrapper aggregators, as outlined in the design deliverables of the Pantheon project (D3.1 and D3.3). This approach offers the advantage of data autonomy, allowing additional data to be easily incorporated into the digital twin model on-demand.

Research suggests that PANTHEON could adopt a hybrid approach for implementing multi-source data integration, combining the strengths of both the data warehouse and mediator approaches proposed by Xie et al. (2022). It also elucidates the technologies and methods that the proposed integration implementation in Pantheon should encompass. Critical methods include data fusion techniques and preprocessing, particularly data aggregation, normalization, and transformation, to harmonize heterogeneous data into a standardized format, ensuring consistency and interoperability.

Processing data closer to each data source via an aggregator wrapper can help reduce latency and enhance the efficiency of data integration, particularly for time-sensitive applications. Additionally, employing ontologies and standardized data formats ensures semantic interoperability, enabling different systems to comprehend and interpret shared information. Furthermore, data integration through Application Programming Interfaces (APIs) facilitates communication and data exchange between Pantheon's various systems, promoting interoperability among the digital twin's applications and services. As noted by Raes et al. (2021), individual models integrated through APIs within PANTHEON can collectively form a cloud of models capable of conducting what-if analyses related to disaster, risk, and emergency management.

4.2.3.2 Fusion Algorithms for Comprehensive Insights

According to Liggins et al. (2017), information fusion is defined as "the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision-making." In the context of PANTHEON, this definition underscores the critical role of SCDT data fusion algorithms in synthesizing and integrating heterogeneous data from diverse sources to offer comprehensive insights into the functioning of the pilot cities, Athens and Vienna. Overarching objective is to create a unified representation that accurately mirrors the real-time status and dynamics of these urban systems. The effective implementation of SCDT data fusion algorithms necessitates a multidisciplinary approach, leveraging expertise in data science, machine learning, computer vision, and domain-specific knowledge pertaining to the project's pilot cities. These algorithms play a pivotal role in transforming diverse data into actionable insights, thereby enhancing the efficiency, sustainability, and resilience of Pantheon's proposed digital twin model.

According to De et al. (2017), key data fusion techniques that PANTHEON could leverage include semantic reasoning and fusion through correlation. Semantic reasoning can be employed to map relational datasets and monitoring data streams, integrating them with match filters to enable context-aware data fusion. This entails algorithms comprehending the context of sensor readings, user interactions, and environmental conditions, while adapting to changes in the urban environment and accommodating new data sources, technologies, and evolving digital twin dynamics. Fusion through correlation encompasses statistical methods, potentially augmented by machine learning algorithms, to calculate the correlation between numerical data streams. For instance, this technique could involve combining community-related social data streams with sensor data within Pantheon's digital twin model. Optionally, machine learning analytics algorithms could be employed to develop predictive models forecasting future emergency events based on historical and real-time data. Separate match filter and pattern recognition algorithms can detect anomalies or irregular patterns in the data, signalling potential disaster events or associated risks that warrant attention.

Lastly, according to Kaur et al. (2020), decision support mechanisms encompass prediction algorithms that further enhance insights through data fusion. Information flow from raw data to high-level decision-making is facilitated by sensor-to-sensor, sensor-to-model, and model-to-model fusion. For instance, the output of data fusion algorithms can be directly integrated into decision support systems to interpret insights.

4.3 Data Analysis

4.3.1 Visualization & Notifications

Smart City Digital Twins generate vast amounts of data, and effective analysis, visualization, and notifications are crucial for transforming this data into actionable insights. Integrating all these elements into PANTHEON modelled cities empowers stakeholders to make informed decisions, respond to events in real-time, and enhance the overall efficiency and liveability of the proposed architecture.

A variety of visualization elements can be incorporated, primarily interactive dashboards providing a visual summary of key performance indicators. These dashboards enable PANTHEON stakeholders to monitor the pilot cities' status in real-time, enhancing user engagement and understanding. Geospatial maps and overlays represent another essential visual element, offering insights into geographical data within a spatial context and highlighting location-specific patterns and trends concerning PANTHEON pilot cities. Limited 3D visualization can also be included, representing the infrastructure, buildings, and terrain of Pantheon's pilot cities for a more immersive experience.

Time-series charts visualize data changes over time, facilitating the identification of temporal patterns in various metrics. A combination of geospatial and temporal data can inform the creation of heat maps, representing data density in a spatial context and aiding in the identification of areas with high or low activity, congestion, or resource usage.

Complementing the data analysis visualization modules, notifications play a crucial role in creating a dynamic and responsive ecosystem for PANTHEON. Event detection is paramount, with the automated PANTHEON system capable of detecting events or anomalies in the data and triggering notifications when predefined thresholds are crossed. These alerts may include traffic incidents, environmental issues, sensor detections, or security-related alerts, ensuring timely notification of PANTHEON stakeholders and users via email, SMS, or mobile applications. Mobile applications can leverage push notifications to instantly inform users about important updates or emergencies, providing contextual information about the event's location, severity, and potential impact. Customizable notification preferences based on roles, responsibilities, and areas of interest enhance user experience and decision-making.

Furthermore, integrating notifications via APIs with existing communication systems, such as emergency services or public address systems, ensures a coordinated response to critical events and emergencies. Finally, combining notifications with logging enables the analysis of historical notifications, facilitating the identification of recurrent issues, optimization of response strategies, and enhancement of Pantheon's digital twin model's overall resilience.

4.3.2 LOGS

An essential feature of a digital twin is its capability to replay simulations, enabling end-users to gain deep insights into the sequence of events that lead to a particular outcome. Repetition-based understanding is a well-known learning process, and the digital twin serves as a new tool to facilitate this process.

To execute a simulation replay effectively, the digital twin must meticulously track all events contributing to the disaster scenario's impact. These events may originate from data sources or interactions introduced by end-users. Furthermore, the timing of events can significantly influence their effects on the scenario during replay, compared to their impact during the initial simulation.

This necessity underscores a new data requirement for the digital twin: the development of a robust log system capable of storing the chronological sequence of all events affecting the simulation. Fortunately, various tools already exist to support log generation, ranging from simple methods that write text messages to local text files (e.g., using the Java Logger module) to more sophisticated cloud-based solutions capable of storing events in cloud data servers (such as Apache Kafka, Amazon CloudWatch, New Relic, WandB, Dynatrace, Mezmo/LogDNA, etc.).

These tools not only store system logs but also enable real-time monitoring of system health without compromising functionality or performance. Cloud systems offer additional advantages, including full scalability and the ability to distribute and duplicate data, thereby enhancing reliability in the event of partial failures or full-scale disasters.

4.3.2.1 Post – Processing

In addition to enabling simulation replay, logs serve as valuable resources for post-processing, facilitating a comprehensive analysis of key aspects within a scenario. Post-processing involves examining all events to gain a global perspective on significant occurrences during the simulation. For example, analysing data elements from each source, identifying outliers, and assessing variability in data values can all be performed at the conclusion of a scenario simulation. Additionally, post-processing enables scenario comparisons, such as evaluating resource requirements, assessing differences in damages between runs, and determining the duration needed to return to nominal conditions for each scenario option.

By systematizing event log capture and automating post-processing with distributed services, end-users can quickly obtain relevant and useful feedback about scenario runs within minutes. This post-processing phase yields objective indicators and measures that serve as the basis for scenario results. Subsequently, these objective indicators guide post-scenario discussions, complementing subjective assessments. For instance, a thorough analysis of simulation results during post-processing, followed by fruitful discussions, may lead to proposals for updating emergency plans with new actions or altering their prioritization.

Most tools providing cloud log services also include post-processing capabilities. For example, in Apache Kafka, the log system transforms into an event-driven distributed environment, enabling the creation of a client and service ecosystem that operates in parallel with streams of logs organized into different topics. Software for Data Analytics (e.g., Tableau, Data BI, Knime, RapidMiner, etc.), as well as ad-hoc programs, spreadsheets, or data warehouses, are all suitable for post-processing data logs obtained from digital twin simulations.

4.4 Data Delivery Schemes

4.4.1 Athens scenario

For Athens, two disaster scenarios will be considered, each with associated technologies:

1. A-1: The wildfire scenario for which all agreed-upon techniques will be utilized.
2. A-2: The earthquake scenario for which all agreed-upon techniques will be employed, with the exception of Weather Stations. This exception is due to the impracticality of offering any useful data from them, resulting in a simplified simulation.

4.4.2 VIENNA SCENARIO

For Vienna, two disaster scenarios will be considered, each with associated technologies:

1. V-1: The heatwave scenario, for which all agreed-upon techniques will be utilized, except for UAV swarms, which are impractical and provide no assistance in this particular case. Consequently, the simulation for this scenario will be simplified.
2. V-2: The man-made disaster scenario involves a terrorist attack on a power plant, resulting in an explosion and subsequent fire spreading to a nearby wooded area, causing a forest fire that extends to the outskirts of Vienna. This scenario is more complex. All agreed-upon techniques will be employed to effectively depict it, including a cascading effect that encompasses the wildfire scenario simulated for Athens.

4.4.3 ON DEMONSTRATORS DATA

At this stage, it remains premature to delineate the precise data delivery schemes for the two pilot cases, despite finalizing the selection of technologies and tools. The system architecture is still in development, with efforts focused on ensuring that all anticipated data will be obtainable from identified sources. However, in the event that not all expected data becomes available, adjustments may be required to accommodate what is accessible for integration into the final PANTHEON Platform. Nonetheless, based on discussions during the 3rd General Assembly meeting in Malta (Jan 2024), there appears to be consensus that the most probable format for data delivery schemes is JavaScript Object Notation (JSON). JSON is a standard text-based format utilized for representing structured data, based on JavaScript object syntax, commonly employed for transmitting data in web applications, such as sending data from the server to the client for display on a web page.

5 CONCLUSIONS

The objective of Deliverable D3.4, titled "PANTHEON Data Delivery Scheme for Community-Based Disaster Risk Management" is to establish the most suitable format for the Data Delivery Scheme within the two pilot areas, Athens and Vienna. This endeavour requires the acquisition and management of data from six distinct streams of origin: Satellites, In-Situ sources, Infrastructure, Traffic, UAVs, and Community inputs. The aim is to achieve high performance, full automation, ease of use, low maintenance, and responsive service.

To accomplish this goal, various types of data were defined based on their characteristics, processing requirements, integration methods, and analysis needs. This comprehensive assessment aims to determine the optimal data delivery scheme for both pilot cases.

Based on discussions held during the 3rd General Assembly in Malta, it was concluded that the optimal data delivery scheme should utilize the JavaScript Object Notation (JSON) format.

6 REFERENCES & ENDNOTES

[1] Selmy, Hend A., Hoda K. Mohamed, and Walaa Medhat. "Big Data Analytics Deep Learning Techniques and Applications: A survey." *Information Systems* (2023): 102318.

[2] Khalid, Madiha, and Muhammad Murtaza Yousaf. "A comparative analysis of big data frameworks: An adoption perspective." *Applied Sciences* 11.22 (2021): 11033.

[3] Vassiliadis, Panos, et al. "ARKTOS: towards the modelling, design, control and execution of ETL processes." *Information Systems* 26.8 (2001): 537-561.

[4] Kimball, Ralph, and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modelling*. John Wiley & Sons, 2011.

[5] Inmon, William H. *Building the data warehouse*. John Wiley & Sons, 2005.

[6] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

[7] Shvachko, Konstantin, et al. "The Hadoop distributed file system." *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. IEEE, 2010.

[8] Chen, Min, et al. *Big data: related technologies, challenges and future prospects*. Vol. 100. Heidelberg: Springer, 2014.

[9] Kumar, Shashank, Aryan Jadon, and Sachin Sharma. "Global Message Ordering using Distributed Kafka Clusters." *2023 15th International Conference on Innovations in Information Technology (IIT)*. IEEE, 2023.

[10] Carbone, Paris, et al. "Apache flink: Stream and batch processing in a single engine." *The Bulletin of the Technical Committee on Data Engineering* 38.4 (2015).

[11] Zaharia, Matei, et al. "Discretized streams: An efficient and {Fault-Tolerant} model for stream processing on large clusters." *4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 12)*. 2012.

[12] Almeida, Ana, et al. "Time series big data: a survey on data stream frameworks, analysis and algorithms." *Journal of Big Data* 10.1 (2023): 83.

[13] Henning, Sören, and Wilhelm Hasselbring. "Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud." *Journal of Systems and Software* 208 (2024): 111879.

[14] Warren, James, and Nathan Marz. *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster, 2015.

[15] Tantalaki, Nicoleta, Stavros Souravlas, and Manos Roumeliotis. "A review on big data real-time stream processing and its scheduling techniques." *International Journal of Parallel, Emergent and Distributed Systems* 35.5 (2020): 571-601.

[16] Vianna, Alexandre, et al. "A Grey Literature Review on Data Stream Processing applications testing." *Journal of Systems and Software* (2023): 111744.

[17] Newman, Sam. *Building microservices*. " O'Reilly Media, Inc.", 2021.

[18] Humble, Jez, and David Farley. *Continuous delivery: reliable software releases through build, test, and deployment automation*. Pearson Education, 2010.

[19] Kolajo, Taiwo, Olawande Daramola, and Ayodele Adebiyi. "Big data stream analysis: a systematic literature review." *Journal of Big Data* 6.1 (2019): 47.

[20] Miloslavskaya, Natalia. "Stream data analytics for network attacks' prediction." *Procedia Computer Science* 169 (2020): 57-62.

[21] Khalid, Madiha, and Muhammad Murtaza Yousaf. "A comparative analysis of big data frameworks: An adoption perspective." *Applied Sciences* 11.22 (2021): 11033.

[22] Bhatt, Nirav, and Amit Thakkar. "An efficient approach for low latency processing in stream data." *PeerJ Computer Science* 7 (2021): e426.

[23] Jelić, Marko, et al. "A State-of-the-Art Review on Big Data Technologies." (2019).

[24] Marozzo, Fabrizio, and Domenico Talia. "Perspectives on Big Data, Cloud-Based Data Analysis and Machine Learning Systems." *Big Data and Cognitive Computing* 7.2 (2023): 104.

[25] Thayyib, P. V., et al. "State-of-the-Art of Artificial Intelligence and Big Data Analytics Reviews in Five Different Domains: A Bibliometric Summary." *Sustainability* 15.5 (2023): 4026.

[26] Xie, X., Moretti, N., Merino, J., Chang, J.Y., Pauwels, P. and Parlikad, A.K., 2022, November. Enabling building digital twin: Ontology-based information management framework for multi-source data integration. In IOP Conference Series: Earth and Environmental Science (Vol. 1101, No. 9, p. 092010). IOP Publishing.
[27] Raes, L., Michiels, P., Adolphi, T., Tampere, C., Dalianis, A., McAleer, S. and Kogut, P., 2021. DUET: A framework for building interoperable and trusted digital twins of smart cities. IEEE Internet Computing, 26(3), pp.43-50.
[28] Liggins II, M., Hall, D. and Llinas, J. eds.: Handbook of Multisensor Data Fusion: Theory and Practice. CRC press (2017).
[29] De, S., Zhou, Y., Larizgoitia Abad, I., Moessner, K.: Cyber–physical–social frameworks for urban big data systems: a survey. Appl. Sci. 7(10):1017 (2017).
[30] Kaur, M.J., Mishra, V.P. and Maheshwari, P., 2020. The convergence of digital twin, IoT, and machine learning: transforming data into action. Digital twin technologies and smart cities, pp.3-17.

[1] https://phoenixnap.com/kb/data-integration-tools
[2] https://support.sas.com/resources/papers/proceedings16/8301-2016.pdf
[3] https://www.jstage.jst.go.jp/article/comex/6/6/6_2017XBL0020/_article
[4] https://www.witpress.com/elibrary/wit-transactions-on-ecology-and-the-environment/158/23316
[5] https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/in_text_citations_the_basics.html
[6] https://www.researchgate.net/publication/233857788_Forest_Fire_Risk_Analysis
[7] Discourse Analysis, Data and Research Techniques. In Discourse Analysis and European Union Politics (pp. 1796).
[8] https://epiteliki.civilprotection.gov.gr/sites/default/files/PDF/%CE%95%CE%9A%CE%98%CE%95%CE%A3%CE%97%20%CE%91%CE%93%CE%93/en_drm_plan.pdf
[9] https://www.ibm.com/topics/geospatial-data
[10] https://www.academia.edu/68928149/Risk_assessment_supporting_public_policy_in_an_uncertain_world
[11] https://pubmed.ncbi.nlm.nih.gov/16928418/
[12] https://sim4nexus-space.eu/
[13] http://www.epsilon.gr/imported/files/brochures/Bonazountas_SEIS-Malta_Final_020414.pdf
[14] "Synthetic data: Unlocking the power of data and skills for machine learning – Data in government." *Data in government*, 20 August 2020, https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/. Accessed 10 December 2022.
[15] https://www.tonic.ai/features/ai-synthesis?utm_medium=ppc&utm_term=synthetic%20data%20generation&utm_campaign=Capabilities+-+Bottom+Funnel&utm_source=adwords&hsa_kw=synthetic%20data%20generation&hsa_cam=20523542329&hsa_ver=3&hsa_acc=9042438892&hsa_ad=672871772153&hsa_grp=158810611408&hsa_src=g&hsa_mt=p&hsa_tgt=kwd-384396231554&hsa_net=adwords&gad_source=1&gclid=Cj0KCQiAxOauBhCaARIsAEbUSQRq6RuHENGoglRq9aP5BmquSwD3aUp8DrPWBKx9yo5Qo9WFzRrlAgUaAsGnEALw_wcB
[16] https://research.aimultiple.com/synthetic-data-for-deep-learning/
[17] https://www.datomize.com/why-use-synthetic-data-versus-real-data/
[18] https://www.fema.gov/disaster/how-declared/preliminary-damage-assessments/guide
[19] https://www.mdpi.com/2571-6255/6/9/336
[20] http://www.epsilon.gr/imported/files/Qualifications/MaltaRisks.pdf
[21] https://link.springer.com/article/10.1007/s41060-023-00393-w
[22] https://climatedetectives.esa.int/earth-observation-data/
[23] Mils, A., M. Bonazountas (2019). COPERNICUS: National Conference on Copernicus Technology and Applications, European Union Brings together the Copernicus program to the Philippines. Final report, Contract No. 2018/402508, EC Brussels, COWI Coordinator. https://eeas.europa.eu/delegations/philippines/59354/european-union-brings-copernicus-programme-philippines_ar
[24] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00468-0
[25] https://www.um.edu.mt/library/oar/handle/123456789/39441

[26] https://fiaumalta.org/news/maltas-2023-national-risk-assessment-released/
[27] http://www.epsilon.gr/projects/228
[28] https://climate-adapt.eea.europa.eu/en/observatory/evidence/health-effects/wildfires/wildfires
[29] https://climate-adapt.eea.europa.eu/en/observatory/evidence/health-effects/flooding/flooding
[30] https://climate-adapt.eea.europa.eu/en/observatory/evidence/health-effects/heat-and-health/heat-and-health
[31] https://emsc-csem.org/about_us/who_we_are/
[32] https://www.immuta.com/blog/the-complete-guide-to-data-security-compliance-laws-and-regulations/
[33] https://www.nature.com/articles/d41586-019-03558-5
[34] https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/data-availability-statements/
[35] https://www.emidius.eu/AHEAD/query_event/
[36] http://www.geophysics.geol.uoa.gr/ , https://www.gein.noa.gr/ypiresies-proionta/katalogoi-seismon/, https://emsc-csem.org/
[37] https://www.upsolver.com/blog/batch-stream-a-cheat-sheet
[38] https://hadoop.apache.org
[39] https://www.analyticsvidhya.com/blog/2023/02/top-20-big-data-tools-used-by-professionals-in-2023/
[40] https://spark.apache.org
[41] https://www.analyticsvidhya.com/blog/2023/02/top-20-big-data-tools-used-by-professionals-in-2023/
[42] https://flink.apache.org
[43] https://jelvix.com/blog/top-5-big-data-frameworks
[44] https://airflow.apache.org
[45] https://www.integrate.io/blog/data-wrangling-techniques-trends/
[46] https://prestodb.io
[47] https://jelvix.com/blog/top-5-big-data-frameworks
[48] https://storm.apache.org
[49] https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared
[50] https://www.datanami.com/2019/05/30/understanding-your-options-for-stream-processing-frameworks/
[51] https://beam.apache.org
[52] https://thenewstack.io/apache-streaming-projects-exploratory-guide/
[53] https://nifi.apache.org
[54] https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared
[55] https://www.upsolver.com/blog/batch-stream-a-cheat-sheet
[56] https://kafka.apache.org
[57] https://mqtt.org
[58] https://www.docker.com
[59] https://k3s.io
[60] https://kubernetes.io
[61] https://azure.microsoft.com/en-us/products/iot-edge
[62] https://www.tensorflow.org/lite
[63] https://pytorch.org
[64] https://aws.amazon.com/dynamodb
[65] https://www.sqlite.org
[66] https://www.openstack.org
[67] https://openwhisk.apache.org