ULTIMO

Advancing Sustainable User-centric
Mobility with Automated Vehicles

DELIVERABLE

D5.5 – Big Data Collection Platform and Data Management Portal - First version

Co-funded by
the European Union

Project co-funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

Disclaimer

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 CINEA. Neither the European Union nor the granting authority can be held responsible for them.

This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Document Information

Grant Agreement	**101077587**
Full Title	Advancing Sustainable User-centric Mobility with Automated Vehicles
Acronym	ULTIMO
Deliverable	D5.5 – Big Data Collection Platform and Data Management Portal - First version
Deliverable Due Date	30.09.2023
Work Package	WP5
Lead Partner	CERTH
Leading Author	Alexandros Papadopoulos (CERTH/ITI), Georgios Nikoletos (CERTH/ITI), Georgios Chondrogiannis (CERTH/ITI), Konstantinos Giapantzis (CERTH/ITI), Georgios Agrafiotis (CERTH/ITI), Athanasios kalfoutzos (CERTH/ITI), Georgios Spanos (CERTH/ITI), Antonios Lalas (CERTH/ITI), Konstantinos Votis (CERTH/ITI)
Dissemination Level	Public
Reviewer	ArgYou

Document History

Version	**Date**	**Author**	**Description of change**
0.1	24.08.2023	Alexandros Papadopoulos, (CERTH/ITI)	ToC proposed
0.2	25.08.2023	Antonios Lalas, (CERTH/ITI)	ToC changed
0.3	31.08.2023	Georgios Nikoletos, (CERTH/ITI)	Contributed to Chapter 1

0.4	08.09.2023	Alexandros Papadopoulos (CERTH/ITI), Georgios Nikoletos (CERTH/ITI), Georgios Chondrogiannis (CERTH/ITI), Konstantinos Giapantzis (CERTH/ITI), Georgios Agrafiotis (CERTH/ITI), Athanasios kalfoutzos (CERTH/ITI), Georgios Spanos (CERTH/ITI), Antonios Lalas (CERTH/ITI), Konstantinos Votis (CERTH/ITI)	Contributions to Chapters 2,3,4,5
0.5	15.09.2023	Alexandros Papadopoulos, (CERTH/ITI)	Cross-check for syntax errors
0.6	20.09.2023	Antonios Lalas (CERTH/ITI)	Review version
0.7	21.09.2023	Christoph Glauser (ArgYou)	Review
0.8	27.09.2023	Alexandros Papadopoulos (CERTH/ITI), Georgios Nikoletos (CERTH/ITI), Antonios Lalas (CERTH/ITI), Konstantinos Votis (CERTH/ITI)	Review comments addressed
0.9	28.09.2023	Antonios Lalas (CERTH/ITI), Konstantinos Votis (CERTH/ITI)	Version ready for submission
1.0	29.09.2023	Lars Abeler (DB)	Deliverable submission

Table of Contents

List of Figures

List of Tables

Acronyms

ADS	Automated Driving Systems
AI	Artificial Intelligence
AM	Automated Mobility
API	Application Protocol Interface
AV	Automated Vehicle
BM	Bestmile
BMM	Business Modelling Manager
CAV	Connected and Automated Vehicles
CB	Consortium Body
CERN	European Organization for Nuclear Research
D7.1	Deliverable 7.1
DC	Demonstration Coordinator
DI	The department of infrastructure (Swiss Canton of Geneva)
DMP	Data Management Plan
DSES	Department of Security and Economy - Traffic Police (Swiss Canton of Geneva)
DTU test track	Technical University of Denmark test track
EAB	External Advisory Board
EC	European Commission
ECSEL	Electronic Components and Systems for European Leadership
EM	Exploitation Manager
EU	European Union
EUCAD	European Conference on Connected and Automated Driving
F2F	Face to face meeting
FEDRO	(Swiss) Federal Roads Office
FM	Fleet Management
FOT	(Swiss) Federal Office of Transport
GDPR	General Data Protection Regulation
GIMS	Geneva International Motor Show
GNSS	Global Navigation Satellite System
HARA	Hazard Analysis and Risk Assessment
IPR	Intellectual Property Rights
IT	Information Technology
ITU	International Telecommunications Union
LA	Leading Author
LIDAR	Light Detection And Ranging
MEM	Monitoring and Evaluation Manager
MT	MobileThinking
OCT	General Transport Directorate of the Canton of Geneva
ODD	Operational Domain Design
OEDR	Object And Event Detection And Response
OFCOM	(Swiss) Federal Office of Communications
PC	Project Coordinator
PEB	Project Executive Board
PGA	Project General Assembly
PRM	Persons with Reduced Mobility
PSA	Group PSA (PSA Peugeot Citroën)
PTO	Public Transportation Operator
PTS	Public Transportation Services
QRM	Quality and Risk Manager
QRMB	Quality and Risk Management Board
RN	Risk Number
SA	Scientific Advisor
SAE Level	Society of Automotive Engineers Level (Vehicle Autonomy Level)
SAN	(Swiss) Cantonal Vehicle Service
SDK	Software Development Kit
SLA	Sales Lentz Autocars
SMB	Site Management Board
SoA	State of the Art
SOTIF	Safety Of The Intended Functionality

LaaS	Logistics as a Service
MaaS	Mobility as a Service
SWOT	Strengths, Weaknesses, Opportunities, and Threats.
T7.1	Task 7.1
TM	Technical Manager
TPG	Transport Publics Genevois
UITP	Union Internationale des Transports Publics (International Transport Union)
V2I	Vehicle to Infrastructure communication
WP	Work Package
WPL	Work Package Leader

Executive Summary

This document elaborates on the “Big Data Collection Platform and Data Management Portal - First version” associated to “T5.4 Deployment Data Collection”. This task aims to analyse the data project requirements taking into account the envisioned services of WP4, along with the evaluation requirements and related KPIs from the pilot sites. Moreover, this task provides a foundation of services (e.g. Kafka, Apache Spark, machine learning libraries, etc.). Specifically, this deliverable provides the design and implementation of the ULTIMO storage platform and the big data execution engines that will be used. In order to fulfil these objectives, the SHOW data management platform is utilised as a design baseline, taking into account the data collected and stored there, along with the data models derived. Then, the ULTIMO Data Registry is presented, along with the ULTIMO Big Data Collection System and Open Data Portal. A detailed ULTIMO Data List is also presented that will form the base for data collection activities within the development of services (WP4) and pilot site activities (WP5).

The development of the Big Data Management Platform and the Open Data portal of the ULTIMO infrastructure is an iterative procedure, which is going to be further enhanced during the next period, taking into account the envisioned architecture, the services development within WP4, as well as the related KPIs. The final version of this deliverable will be provided in “D5.6 Big Data Collection Platform and Data Management Portal - Final version”, where all components will be described in more detail in terms of data formats and functionalities, as well as data transfer functions, taking into account the developed services and the related data.

1 Introduction

During the past few years, many projects and initiatives were undertaken deploying and testing Automated Vehicles (AVs) for public transportation and logistics. However, in spite of their ambition, all of these projects stayed on the level of elaborated experimentation and never reached the level of a large-scale commercial deployment of transport services. The reasons for this are many, the most important being the lack of economically viable and commercially realistic models, the lack of scalability of the business and operating models, and the lack of user-oriented services required for large end-user adoption of the solutions.

The ULTIMO project will create the very first economically feasible and sustainable integration of AVs for MaaS public transportation and LaaS urban goods transportation. ULTIMO aims to deploy in three sites in Europe 15 or more multi-vendor SAE L4 AVs per site. A user centric holistic approach, applied throughout the project, will ensure that all elements in a cross-sector business environment are incorporated to deliver large-scale on-demand, door-to-door, well-accepted, shared, seamless-integrated and economically viable CCAM services. We target the operation without safety driver on-board, in a fully automated and mission management mode with the support of innovative user centric passenger services.

ULTIMO’s innovative transportation models are designed for a long-term sustainable impact on automated transportation in Europe, around the globe and on society. The composition of the consortium ensures the interoperability between multiple stakeholders by making adoption of new technology at minimum costs and maximum safety. The integration of the ongoing experiments of previous AV-demonstrator projects ensures highest possible technical, environmental and societal impacts from the very beginning of the project, as well as during the project lifetime and even long after its completion.

1.1 On-demand Mobility

Public transportation is a key element of a region's economic development andits citizens the quality of life. Governments around the world are defining strategies for the development of efficient public transport based on different criteria of importance to their regions, such as topography, citizens' needs, social and economic barriers, environmental concerns, and historical development. However, new technologies, modes of transportation and services are appearing, which seem very promising to the support of regional strategies for the development of public transport.

On-demand transportation is a public transport service that only works when a reservation has been recorded and will be a relevant solution for areas where the demand for transport is diffuse and regular transport is still inefficient.

On-demand transportation differs from other public transport services since vehicles do not follow a fixed route and do not use a predefined timetable. Unlike taxis, on-demand public transport is usually also not individual. An operator or an automated system takes care of the booking, planning and organization.

It is recognized that the use and integration of on-demand automated vehicles has the potential to significantly improve services and provide solutions to many of the problems encountered today in the development of sustainable and efficient public transport services.

1.1.1 Fully Automated Vehicles

A self-driving car, referred in the ULTIMO project as a **Fully Automated Vehicle** (**AV**), is a vehicle that can sense its environment and moving safely with no direct human input.

The terms *automated vehicles* and *autonomous vehicles* are often used inter changably. The Regulation 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles defines "automated vehicle" and "fully automated vehicle" based on their autonomous capacity:

An "automated vehicle" means a motor vehicle designed and constructed to move autonomously for certain time periods without continuous driver supervision but in respect of which driver intervention is still expected or required.

"Fully automated vehicle" means a motor vehicle that has been designed and constructed to move autonomously without any driver supervision.

In ULTIMO we operate ***Fully Automated mini and micro shuttles for shared public transportation***. In relation to the SAE levels, the ULTIMO project will operate SAE Level 4 vehicles.

Table 1: SAE Driving Automation levels (©2020 SAE International)

SAE INTERNATIONAL

SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	**You are driving** whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			**You are not driving** when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering **OR** brake/ acceleration support to the driver	These features provide steering **AND** brake/ acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering **OR** • adaptive cruise control	• lane centering **AND** • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/ steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

1.2 The ULTIMO dream for public transportation

Our city has a population of 700.000, and the local PTO (Public Transport Operator) deploys 3.000 kilometres of public transportation services' routes, with 4 backbone tram lines, 8 backbone large bus/trolley lines, and 340 fully automated multi-manufacturer mini and micro-shuttles in an on-demand, door-to-door shared public transport service. The fleet of the 340 automated mini-buses is composed of different vehicle generations and models from different manufacturers, as it was purchased in different batches during the last 12 years (it all started with the first 15 bought by a European project in 2023/24). It is today composed of the first 15 twelve-year old vehicles from manufacturer A, 160 eight-year old vehicles from manufacturer B and 160 two-year old vehicles from manufacturer A, which are the new generation of manufacturer's A vehicles. These last 160 vehicles replaced the last 70 old, driver based, thermic buses that reached the end of their life after 18 years of operation. The change from driver large buses to automated mini-buses and from fixed line service to on-demand, door-to-door, for the last area of the city, was decided 6 years ago, and was based to the great success of the already deployed AV-based services in the other parts of the city.

The integration of new vehicles in the PTO fleet was done within just a few days. Thanks to the now standardised APIs of the automate vehicles' connection with the AV fleet management and orchestration service, that allows a plug-and-operate functionality, while also enabling the combination of AVs with all other means of transport, a seamless citizen centric journey is provided to the users. Similarly, the now standardized High Definition (HD) maps between the vehicle manufacturers, allows the reuse of maps created in the past for the 3.000 kilometres of routes of the city, with small adaptations for the new vehicle capabilities. Concurrently with the purchase of the last 160 AVs, we also switched to a new AV fleet management and orchestration service, from another provider, which offers more precise trip time estimations, considering also traffic jams in the city and weather conditions (that affect traffic in the city). The replacement was made within just two weeks, thanks to the new standardised interfaces between AV fleet management and PTO MaaS services.

Passengers are now able to reserve, via their mobile phone app, on the fly or with pre-reservation trips from one side of the city to another, using the optimal transport mode available for them. This is accomplished with a combination of different transport modes, and having the automated vehicle arriving next (or almost) to their door entrance and bringing them to their final destination or to a backbone transport connection. All trips are offered with highly reliable and accurate transport and arrival times, based on real time, dynamic route changes of the vehicles, responding to the passenger needs. Many passenger services are available in the AVs via a mobile phone app. These services adapted to all passengers' needs, from special needs passengers (visually impaired, wheel-chair users, pregnant women, elderly, children), to persons transporting suitcases or a baby-stroller. They provide adapted information and communication of their trip (where to board off, PTO general information, and even infotainment, like city related information and historical facts and anecdotes, related to their current position).

With the full suppression of on-board operators, automated in-vehicle monitoring provides a multitude of passenger services, which were, and are developed with a continuous collaboration and innovation between users and the PTO. An on-board unit identifies locally (that is, in the bus) incidents and events, like aggressions, vandalism, forgotten objects, littering, accidents, etc., and raises incident alarms to the PTO back-office. The PTO officer at the back-office evaluates the incident and decides, based on the operator policies, what action to take: dismiss the incident, or stop the vehicle and call one of the intervention teams to go the vehicle and take the required action. The back-office operator, based on the number of incidents and intervention time legal obligations, can deploy any of the 15 intervention teams available in the city, ready to reach within a few minutes any AV. Service operators are able, with one set of screens linked to the fleet management, to monitor a number of variables, like: The vehicle status, trip reservations, routes chosen, open and closed road segments (due to works, accidents, etc. thus adapting dynamically the trip routings). Meanwhile in a second set of screens, the back-office operators can follow the passenger services' incidents and the status and location of the intervention teams. These two systems, although independent, are linked so that the operator can transfer incidents to intervention teams and turn-on, when possible and under the local legal restrictions, video connection to inspect the incident.

The public transport services are operating 24/7 with different active vehicles, depending on passenger transport demands. The PTO, being eco-sensible, have defined that a vehicle must have an average occupancy of at least 3 passengers. This of course has implications, especially in off-peak hours, in the transport delays, but it is fully justified for an optimal and eco-responsible operation. Exceptions are made for people with special needs or when the passenger is ready to pay an additional price.

While at peak-hours the full 340 strong fleet of AVs is in service for public transportation, in the off-peak and night hours, due to the reduced transportation needs, part of the fleet is not serving passenger transportation. However, because having a vehicle waiting idly is a waste of resources, the PTO has signed an agreement with Logistics as a Service (LaaS) Service providers. During the off-peak hours, the unused vehicles are dynamically chartered by the different LaaS providers to be used for other types of incidental transport of goods or even non-public transportation of passengers. Local shops and enterprises use the service, on demand with dynamic reservation, to offer home delivery of on-line purchases, or B2B delivery of goods from an urban consolidation center to the distribution center (hub, many to one or one to many [7], or even transport their VIP customers directly to their offices or exposition centers. This way they can deliver the goods to their customers in the city, with a delay of just 60 minutes with affordable prices, as the last mile is the most inefficient part of the goods delivery chain. The goods send can include books from local bookshops or libraries, supermarket bags or even documents exchanged between companies or waste. The local shops are even using the service in the early hours (5 o'clock in the morning) for re-provisioning their shops for the day. The so chartered AVs are "returned" to the PTO for public transport at the peak-hours automatically, at the end of the chartered time window.

1.3 Preamble – ULTIMO Big Data Collection Platform and Data Management Portal

The present deliverable is dedicated to the data collection within the ULTIMO project. It encompasses a comprehensive description of the corresponding platform and portal, as well as the mechanisms by which they are integrated. The primary objective of this task is to conduct an analysis of the project's data requirements while considering the envisioned services of WP4. Additionally, it takes into account the evaluation criteria and key performance indicators associated with the pilot sites.

Furthermore, this task establishes a foundational framework of services, including but not limited to Kafka, Apache Spark, machine learning libraries, etc. These services are configured in accordance with the guidelines outlined in the architecture of WP3to ensure uniform levels of quality, encompassing performance, security, and high availability, among others. Specifically, this infrastructure encompasses the design and implementation of the ULTIMO storage platform and the big data execution engines that will be utilized. Notably, this platform will be constructed based on the SHOW data management platform developed by CERTH, while accommodating data sourced from various vehicles, fleet management solutions, and passenger services. All data will be collected and stored in a standardized, compatible format in a dedicated database for each operating site All the databases will be hosted in a common server deployed and maintained by the CERTH/ITI’s technical team.

This approach facilitates the daily uploading of vehicle and service-related data to a secure repository, thus facilitating the creation of an extensive dataset capable of enhancing Cooperative, Connected, and Automated Mobility (CCAM) adherence through associated services. The storage platform's implementation relies on a scalable, fault-tolerant, and distributed file system such as Hadoop Distributed File System (HDFS). It supports the storage of large datasets, incorporates data compression and encryption features, and can scale to encompass hundreds of physical nodes.

Additionally, this task will deliver the ULTIMO open data portal, wherein all pertinent internal and external data assets will be consistently collected and cataloged alongside their respective metadata. Consequently, the ULTIMO data management portal will be constructed utilizing existing open-source technologies, aiming to maximize interoperability with open data federators and streamline deployment. It is poised to become a valuable resource within the project and extend its utility to external institutions seeking to leverage the data assets generated by the vehicles and the operating centers.

2 Related work

Data Management Platform (DMP) refers to a software platform for collection, processing and management of large and complex datasets created from various sources. Such platforms are utilized in several activities (e.g., industry, healthcare, environmental monitoring, logistics) that require immediate processing of generated data for decision making, creating statistics, drawing conclusions and so on.

2.1.1 Event Processing

Event Processing is a method of monitoring, storing, analysing and processing data flows, which are relevant to various events that have taken place, with the purpose of extracting useful conclusions and responding with proper actions. Event Processing based methods that DMPs utilize for data collection and processing in real time are listed below.

Complex Event Processing (CEP)

Complex Event Processing (CEP) is a computational method of processing continuous data flows, which signify specific events, from several distributed sources. The advantage of such methods, compared to conventional event processing, is that large amounts of data can be processed in real time or near-real time, allowing for immediate detection of significant events or sequences of events (e.g. through pattern recognition techniques) and in time decision making capabilities[1], [2], [3]. More particularly, systems that apply CEP are enhanced with the ability of detecting events through relevant input data and, with proper processing, extract their relationships and patterns in order to identify event combinations and sequences. Thus, simple and complex high importance events (e.g. accidents, dangerous situations, emergencies) can be immediately detected, allowing for fast appropriate response by the system.

The CEP techniques are mainly utilized in applications with requirements such as:

• **Complex Patterns**: The requirements for event detection include complex patterns of events with temporal or spatial relationships, or both.

• **Low Latency**: the completion time of an event processing since relevant data are generated, is expected to be in the order of a few milliseconds or less.

• **High Event Rate**: the incoming events rate can be hundreds or thousands of events per second.

To summarize, CEP refers to a set of resources, which are provided with data from system parts or distributed sources and through processing, extract meaningful results for decision makers. Commonly, CEP is a useful tool for high level for system organizing and design. Indicatively, areas with CEP utilization may be:

• **Business monitoring**: CEP methods are used to detect opportunities, misoperation, frauds, processing data for risk analysis and so on.

• **Autonomous vehicles**: CEP manages large amounts of data (images, video streaming, sensor measurements) which are exchanged between the vehicles, pedestrians, road infrastructure etc.

• **Power and SmartGrids:** CEP analyses sensor data and consumer demands, develops optimized strategies for power distribution.

• **Ad Targeting**: CEP collects and analyses data from users' behavior in real time and respond with viewing related advertising content.

• **Healthcare:** Uses CEP mechanisms to monitor patients' vital signs, detect emergencies and inform healthcare professionals. Moreover, CEP may be used to detect disease outbreaks.

Event Stream Processing (ESP)

CEP is a useful computational tool for extraction of recurrent correlations between low-level events, while Event Stream Processing (ESP) refers to a range of computational methods, which are employed to process the stream and apply analysis. In contrast with CEP, ESP handles data from spatially and temporally continuous streams of events, while CEP performs management of data originating from seemingly unrelated events from various sources and distinguishes the patterns that connect them. Therefore, a set of subsequent events cannot be an input of a CEP model[4].

ESP algorithms are considered faster since they perform stateless processing, that is, the state of event patterns is stored for short time periods, which allows for more efficient memory and processing resources usage. Despite that this computational method focuses on processing streams of events in general, without the feature of detecting complex events, recent ESP solutions are designed to address such cases on a basic level. This is achieved by making use of the incident time (i.e., the time the event occurs) rather than processing time (i.e., the time that the relevant to the event data arrive at the system for processing)[5].

2.1.2 Massive Scaling

Data scalability refers to the ability of a system to adapt the usage of computational resources (CPU frequency, RAM, disk) to rearrangements in the extent of the application's use scale. The expected expansion of Public Transport systems in the upcoming years, in terms of fleet of autonomous vehicles and number passengers-users of the relevant applications, makes data scalability crucial for the implementation and operation of the aforementioned services.

Two types of scaling are commonly used: Vertical scaling, known as "scaling up" and horizontal scaling, known as "scaling out". Vertical scaling can be described as a straightforward increase or decrease of computation resources, such as CPU, RAM and storage, with respect to the changes in the application requirements. Its main benefits are simplicity and better maintenance. However, environmental expansion requires larger machines. Horizontal scaling refers to data sharding, that is, storing and processing data across multiple machines.

MongoDB[6] is a popular software solution that performs Horizontal scaling, using two different methods. The first method is called sharding and it is based on dividing the database into smaller and more manageable parts (shards) which are distributed across multiple machines. This method requires additional complexity to ensure that information is not duplicated or lost. The second method is replication and it is based on full copies of databases in multiple machines. Obviously, this method adds extra traffic in the systems. Both methods can ensure efficient management and up-to-date databases.

2.2 Commercial Solutions

In this subsection, commonly used commercial solutions for big data collection and management are presented and compared.

Apache Hadoop[7] is the most common software solution for big data processing. Hadoop uses a batch processing scheme by default for most of its development. MapReduce [8] is the batch processing paradigm used by Hadoop. In the "map phase", data is divided and distributed across several nodes. The data are grouped and processed at reducer nodes, where the final results from data processing are extracted. This process is similar to the extract-transform-load (ETL) processing. The leverage of Apache Hadoop is widespread as Hadoop is capable of distributed processing.

Despite Hadoop in terms of efficiently handling big data, there are use cases that need data processing closer to real-time, rather than batch processing. Processing of streams is a method focusing on real-time and near real-time processing. Hadoop has been enhanced with the introduction of new computation tools to address that requirement for real-time processing. Stream processing tools are embedded in systems like Apache Spark, Apache Storm, Apache Samza, Apache Flume etc. Opposed to batch processing of MapReduce, these solutions allow the continuous process of data flows.

Apache Spark[9] is a flexible cluster computing platform offering high performance and scalability. It is highly accessible, providing rich libraries and user-friendly environments in Python, Java, Scala and SQL.

- **Optimized for speed**: Spark extends the MapReduce model to support computing methods such as interactive queries and stream data processing. One of the main features that Spark applies for speed enhancement is the ability to use memory for calculations.

- **For general purpose**: Spark is designed to be flexible, supporting a wide range of different kinds of computations that require different distributed systems. Spark supports execution on a single machine, which makes it easy and cost effective to achieve combinations of several processing modes. In addition, it reduces the overhead of maintaining separate tools.

Spark has been used in frameworks for data flow and real time data processing optimization within systems with interconnected AVs. In[10] an AV framework is presented based on Spark streaming [11], for traffic prediction with a maximum error of 4%. The proposed method can maintain stability and low latency in extreme changes of data rates. The solution is deployed in an edge computing node for sensor data processing.

Apache Storm[12] is a distributed real-time open-source solution. Storm is capable of reliably processing unbounded data streams, doing for real-time processing. It is simple and compatible with any programming language.

Some of Storm's use cases include online machine learning, real-time analytics, social media analysis, network traffic analysis, ETL and so on. Some of its benefits include:

- **Fast performance**

- **Scalability**
- **Fault-tolerance**
- **Reliability**
- **User friendliness**

Storm design uses CEP, as other solutions, focusing on handling complex events in real time with low latency, making it suitable for applications involving time critical decisions. In [13], a big data architecture is proposed that collects data from a vehicle, and offload that data at cloud nodes for processing. The results for the vehicle's behaviour are visualized at the user's mobile device application. The mobile device collects information of fuel consumption, temperature and maintenance indicators and transmits the data to the cloud for further processing including big data analysis and machine learning.

Apache Samza[14]is an open source, near-real-time framework for stream processing. It is compatible with Scala and Java. Samza [15] is used to develop stateful applications for real-time data processing from multiple sources including Apache Kafka. Its benefits include:

- **Fault tolerance**
- **Isolation**
- **Stateful processing**

Unlike batch systems such as Apache Hadoop, it applies continuous data processing methods resulting in sub-second[15] response times. In [16] an animal health surveillance platform is presented for beehives monitoring. Data from several sources are offloaded to the cloud for processing. The system applies lambda architecture that includes the utilization of LPWA and 3GPP protocols. Data are transmitted for real time processing in order to produce inferences and analytics. The system utilizes the messaging system Apache Kafka asan interface between the data producing nodes and the processors running the Apache Samza algorithms.

Apache Flume[17] is a secure distributed open-source software for storing, aggregating and transferring log data in a large scale. It is focused on streaming data flows with a simple and scalable design. It applies several methods for tuneable stability, failover and recovery making it is stable and fault tolerant. It uses an expandable model, suitable for online applications. In [18] a big data solution on AVs that utilizes Flume is presented. The architecture's building block is the data sources that generate data. The systems architecture is based on several applications: Flume, Kafka, Zookeeper, HDFS, and Spark. The data transmitted from sources to HDFS is handled by Flume and Kafka. In [19], Flume is also used in a proposed Cloud-Based IoT data processing pipeline. In this work, authors aim to provide a solution for uninterrupted QoS and cost efficiency fulfilment by proper resource allocation strategy.

Apache Flink [20] is an open-source software, combining both stream and batch processing. It is also a distributed processing tool for stateful computations over unbounded and bounded data streams. Flink algorithms can be executed in all common cluster environments and perform computations at in-memory speed and at any scale. Compared to other solutions with high throughput and low latency, Flink's framework provides event time processing and state management. In terms of data storage, Flink is compatible to systems that provide this feature, such as Amazon Kinesis, Kafka and HDFS. Apache Flink is applicable in Cloud and Edge/Fog Computing[21],and in IoT infrastructure [22].

Apache Kafka [23] is an open-source stream-processing software framework, written in Scala and Java, however it is designed to be language-agnostic, i.e., compatible with a large number of languages. Kafka provides a single, low-latency, high-throughput network to manage real-time data flows. Kafka Connect is the interface for integration to external systems (for data import/export). Kafka Streams is a stream processing library for Java. Kafka uses an efficiency-optimized binary TCP-based protocol that groups messages together for reduction of the network's round trip overhead.

This TCP-protocol enables:

- The **transmission** of **larger network packets**,
- **Larger** sequential **disk operations**,
- Continuous memory blocks.

The aforementioned method enables Kafka to convert streams of random message writes to linear writes that are send to the consumers.

In Table 1, afore mentioned software solutions are summarized.

Table 2: Big Data Management Software Solutions

Name	Company	Supported Languages	Processing Type
Hadoop[7]	Apache Foundation	Java, Perl, Python, Ruby, Php	Batch
Spark[9]	Apache Foundation	Scala, Java, R, Python	Stream/Batch
Storm[12]	Apache Foundation	Java, Perl, Python, Ruby	Stream
Flink[20]	Apache Foundation	Scala, Java, Python	Stream/Batch
Kafka[23]	Apache Foundation	Java, C/C++, Go, .NET, Python, Scala, Erlang, Groovy, Haskell, Kotlin, Lua, Node.js, Ocaml, Php, Ruby, Rust, Tcl, Swift and more	Stream
Samza[14]	Apache Foundation	Java, Scala, Python, Go	Stream
Flume[17]	Apache Foundation	Java	Batch
Amazon Kinesis[24]	Amazon	Java, Python, Ruby, Node.js	Stream

2.3 SHOW Data Management Platform

The Shared Automation Operating Models for Worldwide Adoption (SHOW[1]) is a Horizon 2020 European initiative with a specific focus on autonomous public transport fleets. Comprising a consortium of 20 operators, with more than 70 vehicles, spanning across Europe, these operators employ a variety of original equipment manufacturers and systems. The primary objective of this endeavor is the effective integration and deployment of autonomous public transport fleets.

2.3.1 SHOW Data Management Platform at a glance

A main component of SHOW is the Data Management Platform (DMP). SHOW DMP is the entity that includes all the mechanisms, dedicated to the collection, the manipulation and the storing of the data within SHOW Project. As it has already been mentioned, SHOW project composes a very challenging ecosystem due to the existence of different and numerous partners relevant to Automated Mobility. For this reason, apart from the mechanisms, respective models and formats are also part of the DMP architecture. In Figure 1, the basic workflow of SHOW DMP is depicted.

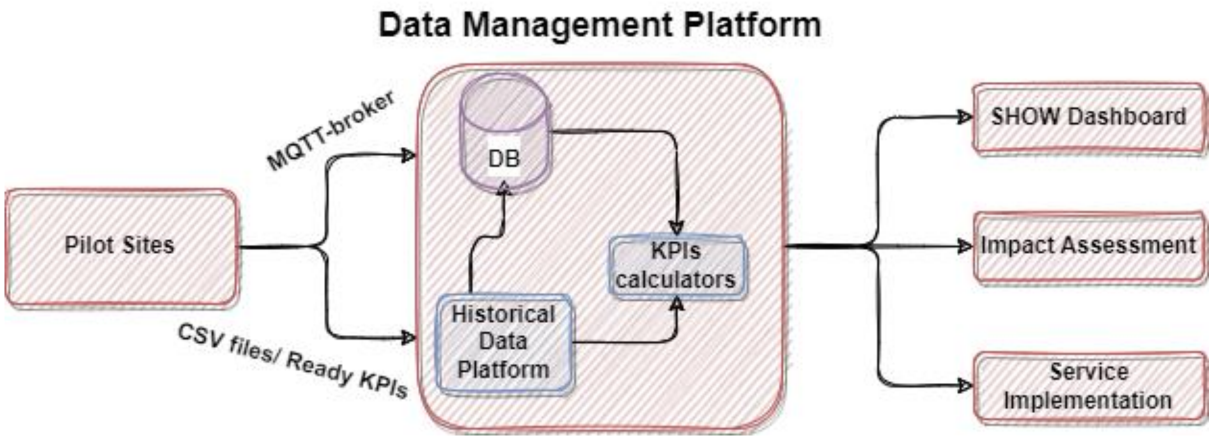

Figure 1: Basic Workflow of SHOW Data Management Platform

As an initial phase, the participating sites are required to establish the method of connectivity with the SHOW Data Management Platform. Three principal alternatives exist for this connection, with the primary determinant being the method employed for the calculation of Key Performance Indicators (KPIs). The KPIs within the SHOW project are primarily centered around metrics derived from performance data originating from the sensors on vehicles and infrastructure. These metrics serve to portray the extent of integration of autonomous fleets within the operational ecosystem of the city. Illustrative examples of these KPIs encompass the average operational speed, variance in acceleration, frequency of hard braking events, total kilometers traveled, and passenger count. The KPIs

[1]https://show-project.eu/

are presented in a publicly accessible SHOW Dashboard, offering insights on a per-site basis and delineating data on a daily, weekly, and monthly basis pertaining to the operation of its vehicles.

The options are the following:

1. **The site provides real-time data**. The data is led to the database directly via a Message Queuing Telemetry Transport (MQTT) broker. The KPIs are calculated automatically. In the following steps, they are uploaded to the dedicated Application Programming Interfaces (APIs) in order the connection with the SHOW Dashboard to be feasible.
2. **The site provides CSV files with historical data.** The files are uploaded once per day, week or, in the worst-case scenario, once per month. The data is led to the database by technical team and the procedure is completed with the same way as in the previous option. The provision of historical data encompasses the upload of prior data, which can be leveraged for the development and training of new services.
3. **The site provides Ready KPIs.** The respective files are uploaded also in the historical data platform. After this, they are loaded in the database and, overcoming the calculation step, they are led to the APIs in order to be visualized within the SHOW Dashboard.

Apart from the visualization of the KPIs in the SHOW Dashboard, the collected data and the produced KPIs are utilized also for other purposes. The Impact Assessment needs to have access both in the raw data and the KPIs in order it to check the calculation methods and create new ones. Furthermore, the deployment of services mandates access to historical data as well as real-time workflow information. These data exchanges are facilitated through specialized APIs among the collaborating entities, thereby guaranteeing both the ease of connectivity and the security of data transmission.

2.3.2 Mechanisms of SHOW DMP

As previously mentioned, the SHOW DMP comprises two primary components: mechanisms dedicated to data collection and commonly utilized data models and formats across all participating sites. Focusing on the data collection workflow, a MQTT/Kafka broker is employed to establish real-time connections with the fleets. The technical team responsible for the DMP has implemented a visualization middleware tool called Lenses to ensure the proper functioning of all connections, allowing for effective management.

The generated data is stored in a Mongo DB database. The delivery of KPIs, also stored in separate tables within Mongo DB, is accomplished through the utilization of APIs and an authentication/authorization mechanism supported by the KEYCLOAK IDM platform. Beyond its connection with the SHOW Dashboard, additional APIs have been designed to provide data to both the impact assessment and service implementation teams.

Regarding the provision of historical data, the dedicated platform employed is CKAN. Pilot sites have the capability to seamlessly upload their datasets through a user-friendly interface. The data imported in this manner is then directed to MongoDB. CKAN also facilitates data provisioning through its own APIs. To safeguard the data within CKAN, an authorization mechanism, KEYCLOAK, is in place to ensure protection.

Given that the SHOW Dashboard is publicly accessible, the ultimate verification of KPI values before their release is conducted through a private tool developed in Grafana. The method of connection between this internal dashboard and the official one mirrors that of the entire system. After the final inspection is completed, the values are made accessible to the public through SHOW Dashboard. The mechanisms of the SHOW Data Management Platform are visually represented in Figure 2.

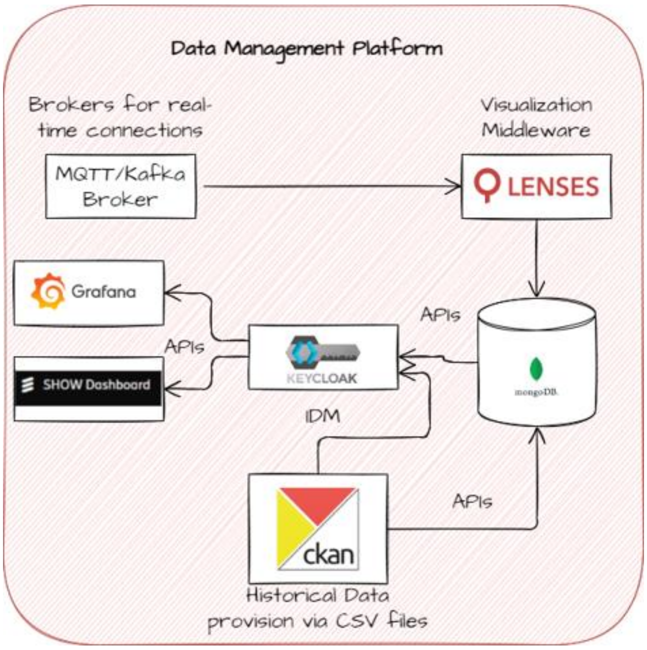

Figure 2: Mechanisms of SHOW DMP

2.3.3 Data Model and Formats

The consistent utilization of standardized definitions and formats within the SHOW project has proven to be immensely valuable. This practice ensures that various fleets and manufacturers align with the project's documentation, thereby avoiding substantial time losses associated with translating between different models. At the heart of this standardization effort lies the Data Registry, which comprehensively describes all data attributes provided by the participating sites. This serves as a guiding reference for both real-time and historical data provision.

In addition to establishing common definitions, the project has predefined the format for each data approach. In the case of real-time data connection, the JavaScript Object Notation (JSON) message format for each attribute has been disseminated, facilitating the technical team's consistent calculation of KPIs. Moreover, the data provision frequency has been standardized to approximately 1-2 Hz to optimize computational resources.

For historical data provision, specific formats for the uploaded CSV files have been meticulously designed. Additionally, a corresponding format for Ready KPIs has also been included. Sites adhere to an agreed-upon timeframe for data uploads, collaboratively established with the technical team of SHOW DMP. The structure and data model format for both data connection methods are illustrated inFigure 3. More details are included in the next chapter.

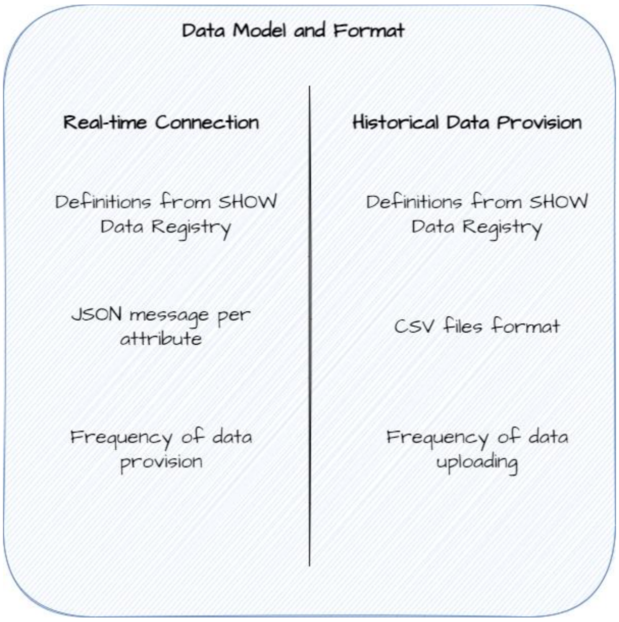

Figure 3: Data Model and Formats of SHOW DMP

3 Data models

An essential requirement for the development of ULTIMO services is the establishment of a common data model. A comprehensive overview of 19 different data models is already provided in SHOW Deliverable 5.1[27]. The ULTIMO project intends to address the growing demand of information exchange between various stakeholders (e.g., partners, fleet operators and pilot sites). Without the adoption of a common data model, the process of integration could be proven exceedingly challenging from the potential lack of understanding regarding the semantics of exchanged data. Therefore, achieving seamless interoperability between systems is going to be attainable if all involved parts utilize similar definitions, structures and interpretations of their data.

Broadly, a data model consists of three components:

1. A collection of types of **data structure**.
2. **A compilation of operators** or inferencing rules that can be employed on valid instances of the data types to extract data from various segments of those structures.
3. An **assortment of general integrity rules**, which directly or indirectly define the set of coherent database states or alterations in states, or both.

3.1 Data models in Public Transport Domain

Numerous data models exist for Public Transport, and the following are some of the most notable:

1. **Transmodel [28]:**Transmodel serves as the standard reference data model for European public transport, providing reference data structures that encapsulate the semantics of the public transport domain.
2. **NeTEx [29]:** NeTEx, which stands for Network and Timetable Exchange, is a data exchange format for public transport networks, timetables, and fares. It is based on the Transmodel framework and utilizes XML for data representation.
3. **INSPIRE [30]:** The emphasis of INSPIRE lies upon spatial information infrastructures operated by the European Union Member States. The Directive covers 34 spatial data topics, including Transport Networks.
4. **TAP- TSI [31]:** Commission Regulation (EU) 454/2011, also known as TAP-TSI, is a regulation aimed at standardizing data exchange among railway undertakings, railway infrastructure managers, and ticket vendors.
5. **DATEX II [32]:** With the objective of facilitating the realization of a interconnected cross-border sharing of traffic and travel information, encompassing state, urban, and interurban road administration, infrastructure providers, and service providers, DATEX II encompasses a standardized framework for data exchange requirements.

It is noteworthy that Transmodel maintains connections with numerous EU standards, including:

• **DVC** [33] (Data Communication on Vehicles)

• **IFOPT** [34] (Identification of Fixed Objects in Public Transport)

• **SIRI** [35] (Standard Interface for Real-Time Information)

• **DJP/OJP** [36] (Open API for distributed journey planning)

• **NeTEx**[29] (Network Timetable Exchange)

• **OpRa** [37] (Operating Raw Data and statistics exchange).

3.1.1 Transmodel

Transmodel serves as the abbreviated term for the European Standard known as the "Public Transport Reference Data Model" and stands as the dominant choice, particularly within Europe, concerning public transportation networks research [38]-[45]. It has been incorporated across most of the countries hosting **ULTIMO pilot sites**. This ensures a seamless data transition to the DMP with minimal conversion efforts. By employing Transmodel, the forthcoming cutting-edge European Public Transport Systems ensure both interoperability and compatibility. Notably, Transmodel transcends the limitations posed by different Transport Modes.

3.1.1.1 Transmodel features

Undoubtedly, Europe showcases the most intricate and densely interconnected public transit networks worldwide. European nations have prioritized the development of systematic information models to support the advancement of transportation information systems and the effective management of a wide variety of distributed data sets [46]. The primary European initiative in this regard is **Transmodel V6.0**[46], [47].

Public transport services are becoming increasingly dependent on information systems to achieve the following objectives [28]:

- Enhancing **accessibility** for a broader audience.
- Ensuring **reliable** and **efficientoperation**.
- Delivering **precise passenger information**.

These systems serve a variety of specific functions, which include [28]:

- Creating **timetables**.
- Establishing **schedules**.
- **Overseeing** vehicle fleets.
- Issuing **tickets** and **receipts**.
- Offering **real-time updates** on service status.

Transmodel has successfully created a well-defined terminology, a notable accomplishment. This is particularly significant in the realm of Public Transport, where common language usage of terms like route, journey, trip, or stop often encompasses a broad spectrum of interconnected ideas. Additionally, Transmodel distinctly distinguishes between the description of the movement of a person, from the movement of a vehicle in public transport [28]. In general, the Transmodel standard offers a structured approach for defining and agreeing on data models. Utilizing this standard, along with data models derived from it, will facilitate seamless collaboration among ULTIMO partners and achieve success in integrated systems. Attempting to comprehend and convert the unique data formats provided for each site without Transmodel would be a challenging and futile process.

3.1.1.2 Transmodel packages

Transmodel V6.0 is organized into primary packages, each corresponding to distinct sections that contain sub-packages, or models, which group classes based on shared functional objectives. In various sections, specific packages referred to as 'Explicit Frames' are established, and the details of these models will be explored in their respective sections.

The main packages include:

1. **Common Concepts [48]**
2. **Public Transport Network [48]**
3. **Timing Information and Vehicle Scheduling [48]**
4. **Operations Monitoring and Control [49]**
5. **Fare Management [50]**
6. **Passenger Information [51]**
7. **Driver Management [52]**
8. **Management Information and Statistics [53]**

3.1.1.3 Data definition of Transmodel

In reference to [54], an extensive collection of data definitions is available for use within Transmodel. This data list can provide comprehensive descriptions of the previously mentioned packages as well as every operation within a Public Transport system.

3.1.1.4 Basic elements of Transmodel vocabulary

In this section, we will introduce a limited portion of Transmodel's **vocabulary**. The selection of the meanings to be examined has been made to illustrate the significance of adopting a standardized data format that all SHOW partners should utilize.

The initial phase involves distinguishing between the motion of a vehicle and that of a person. To achieve this distinction, Transmodel introduces the term **'service journey'** for the vehicle's movement and **'PT trip'** for the person's movement. For instance, when an individual intends to travel from point A to point B, this defines the PT trip. Subsequently, it becomes essential to consider the service journeys of various transportation modes to enable the transportation process.

These definitions are provided as follows:

- The **trajectory** of a public transport vehicle is determined by its scheduled stopping points. The arrangement of these points defines the **journey pattern**.It isimportant to emphasize that a journey pattern includes a distinct direction. For instance, a journey from point A to point B represents a different journey pattern than a journey from B to A. Journey patterns occur frequently throughout an operational day (further elaborated in the next section)."

- Considering the concept of journey patterns and their repetitive nature, we introduce a new definition: the **'vehicle journey'**. This term not only signifies movement from point A to point B but also includes a specific timeframe for this movement. Consequently, journey patterns are applicable only once per Operating Day. A 'service journey,' which is designed to transport passengers, either exclusively passengers or passengers along with cargo, is the term used for such a journey."
- A **'dead run'** refers to a vehicle journey in which passengers are not permitted. The concept of a dead run holds significant importance as it allows for the focused analysis of various functions of Public Transport vehicles. For instance, the evaluation of service utilization by the public can be determined based on the time when the vehicle is not empty. To ensure accurate calculations, the time spent in dead runs must be excluded from these computations.
- To attain **optimal Fleet Management efficiency**, it is highly advantageous to have precise daily records of each vehicle's movements. To facilitate this, we employ the concept of a **'block'.** A block depicts the entire journey of a vehicle, starting from when it departs the depot until it returns. Normally, each block begins and ends with a Dead Run, with none of this time considered as part of the Operation Day. It is important to note that this description is not obligatory.
- A **route** outlines the course followed by regular public transport services. In the context of a road network, a route is defined by a **starting point**, an **ending point**, and potentially one or more intermediary route points. Using a technique called projection, the actual path through the road network can be deduced from this information. In some cases, it may be necessary to include additional intermediary route points to clarify the intended physical path. The subsequent step involves defining the locations where passengers can board and disembark, such as a bus stop, platform, gate, or pier. These locations are typically represented as **scheduled stop points** within the planning software.

3.1.2 NeTEx

The Network and Timetables Exchange (NeTEx), under the purview of CEN/TC 278 WG3/SG9, is a pre-CEN/technical standard presently in the process of development. NeTEx seeks to establish itself as an effective European standard for sharing public transport schedules and associated data. NeTEx is built upon Transmodel and shares common characteristics with IFOPT and SIRI standards. It serves as a valuable complement to the SIRI standard, offering support for enhancing SIRI services related to network, timetables, and fare exchanges. To facilitate the effective and current exchange of complex transport data among distributed systems, NeTEx utilizes a comprehensive model that encompasses key public transport concepts. It also aims to encourage the utilization of data in the development of digital web services and to provide assistance for passenger information and operational applications. It provides a means to share data among diverse computer systems for passenger information, encompassing details such as stops, route timetables, and fares, as well as pertinent operational data. This functionality can be harnessed by multiple stakeholders to gather, implement, and seamlessly integrate data as it evolves across successive versions. In Public Transport projects, unquestionably, the most widely employed standard is NeTEX [29]. In conclusion, this standard encompasses all the vital domains of Public Transport services, including **Network Topology**, **Timing Information**, **Vehicle Scheduling**, and **Fare Information**.

NeTEx is organized into three distinct sections, each addressing a specific functional aspect of Transmodel for Public Transport Information:

- Part 1: Explains the Public Transport Network topology.
- Part 2: Provides details on Scheduled Timetables.
- Part 3:Addresses Fare information.

NeTEx is designed as a versatile XML format, aiming to facilitate the efficient and up-to-date exchange of intricate transport data among distributed systems. This capability allows the data to be effectively utilized within contemporary web service architectures and to support a broad range of operational applications and passenger information needs.

All three sections covered by NeTEx utilize a consistent framework comprising reusable components, versioning mechanisms, validity conditions, global identification mechanisms, among others, as defined within the NeTEx framework. Additionally, NeTEx incorporates container elements referred to as "version frames" for the purpose of organizing data into coherent sets, thereby facilitating efficient data exchange. The NeTEx schema serves as a versatile tool for the exchange of a wide range of Public Transport information, including but not limited to:

- **Public Transport schedules**, featuring fundamental elements of a Public Transport System such as stops, routes, departure times, frequency, operational details, and map coordinates.
- **Routes**, providing the capability to describe intricate network topologies, including circular routes, cloverleaves, lollipops, as well as complex operational scenarios such as short workings and express patterns. It also enables the description of connections with other services.
- NeTEx includes information regarding the **days on which services operate**. This data is instrumental for adapting the Public Transport System to public holidays and differentiating service levels between weekdays and weekends, among other considerations.
- **Composite journeys**, such as train journeys that involve the merging or splitting of trains.
- **Operator information**, detailing the entities responsible for providing the services.
- **Supplementary operational data**, including positioning runs, garage-related information, layovers, duty crews, which prove valuable for AVL (Automatic Vehicle Location) and on-board ticketing systems.
- Data related to the **accessibility of services** for passengers with restricted mobility.
- Data is versioned and enriched with **management metadata**, enabling updates across distributed systems.

- **Fare structures, fare products**, and specific fare prices applicable on particular dates.

3.2 The ULTIMO Data Model

In the context of European projects spanning multiple sites characterized by varying maturity levels, diverse systems, and protocols, it becomes imperative to establish stringent requirements for data models. The harmonization of data models across these heterogeneous environments is critical to ensure seamless data interoperability and effective collaboration among project stakeholders. The specific requirements governing these data models are multi-faceted and must encompass factors such as simplicity, reusability, and extensibility. Among the myriad of available options, there is a compelling need to adopt the SHOW Data Model for EU projects.

The SHOW Data Model shines in this context because of its simplicity. It is user-friendly and easily adoptable by a wide range of operators working in Europe. This simplicity not only makes the model accessible but also boosts overall efficiency. Additionally, its ability to support reusability and extensibility is vital. It allows project teams to make the most of existing data structures while accommodating future data needs seamlessly. What's more, the fact that over 20 operators across Europe are already familiar with this model speaks to its credibility and reliability. By opting for the SHOW Data Model, European projects can navigate the challenges posed by diverse sites and different data systems, ultimately enhancing their chances of success.

The ULTIMO Data Model is fundamentally rooted in the previously described Transmodel framework and it is an extension of SHOW Data Model. It primarily comprises the ULTIMO Data Registry, along with the data formats and metadata associated with it. The ULTIMO Data Registry, as elaborated in the subsequent section, serves as the foundational vocabulary for defining data names. Within this registry, the included Data Types encompass all the requisite information necessary for implementing cutting-edge services that can comprehensively address the spectrum of needs within the realm of autonomous public transport.

Moreover, the ULTIMO Data Registry extends its purview to encompass data attributes originating from all conceivable data generators within a smart city context. The ULTIMO Data Registry, in turn, draws inspiration from the SHOW Data Registry and adopts its underlying philosophy. In scenarios where the implementation of services requires such actions, such as the delineation of specific Use Cases, or the formulation of KPIs, the augmentation of new data attributes will be considered. This approach ensures that the data model remains agile and responsive to the evolving demands and complexities inherent to smart city initiatives and the broader domain of public transportation. Regarding data formats, the ULTIMO Data Model employs JSON as the foundational structure for real-time data provisioning. An illustrative message format is presented in Figure 4. In this format, each operational site, fleet, and vehicle is uniquely identified, facilitating seamless data provisioning. Within the example message, critical information such as position, speed, and acceleration values are conveyed. In instances where a vehicle is unable to provide acceleration data within the same message, the corresponding data point is left empty.

Furthermore, the message incorporates essential data components, including timestamp, journey ID, and the name of the upcoming stop. Additionally, comprehensive information regarding the prevailing city traffic conditions is included, offering valuable insights into the operational environment. Lastly, the message features a flag denoting whether the vehicle is operating in manual or automated mode. This standardized JSON-based format not only streamlines data exchange but also enhances the overall comprehensibility and utility of real-time data within the context of autonomous public transport, promoting efficient decision-making and system performance optimization.

```
2.  Attribute Type: GPS

    - MQTT Topic Name: `show/<fleet-id>/<vehicle-id>/GPS`
    - Vehicle Role: `PUBLISHER`
    - SHOW Dashboard Role: `SUBSCRIBER`
    - Sample Message:

```
```
{
        "vehicleId": "<vehicle-id>",
        "fleetId": "<fleet-id>",
        "siteId": "<site-id>",
        "timestamp": "2020-11-07T13:50:11.326860Z",
        "switchToManual": false,
        "position": {
            "geoPosition": {
                "latitude": "59.408259",
                "longitude": "17.95298",
                "altitude": ""
            },
            "orientation": "160.9",
            "orientationUnt": "radians"
        },
        "journey": {
            "journeyId": "",
            "next_stop": ""
        },
        "speed": {
            "speedValue": "20.0",
            "speedUnit": "KMPH",
            "accelerationValue": "",
            "accelerationUnit": "KMP(H^2)",
        },
        "cityTraffic": "Moderate"
    }
```

Figure 4: An indicative example of JSON message format

Within the framework of the ULTIMO Data Model, a well-defined format for historical data provisioning has been established. This format involves the submission of data files in CSV format, which are then uploaded to a dedicated platform for processing and analysis. It is noteworthy that the attribute names utilized in these CSV files conform to the standardized and agreed-upon nomenclature as outlined in the ULTIMO Data Registry.

4 ULTIMO Data Registry

This section is responsible for the presentation of the ULTIMO data list, liable for the project requirements. The compilation of the data list requires cooperation and contribution from the pilot sites, as it is a crucial procedure that defines the realization of the project. It is important to note that the data list provided in this deliverable is not definite, as revisions and finalization is happening in the next months, following the overall progress of the ULTIMO project.

The ULTIMO data list derives from the data necessary for the accomplishment of the ULTIMO provisional services, as defined for the use cases in D4.4. The existing analysis will be affected by the hardware limitations of the autonomous shuttles, as well as the development of extra services that might be proposed in the duration of the project, along with the needs of the ULTIMO Data Management Portal and the according Dashboard. The GDPR requirements play a pivotal role in the creation of the aforementioned data list, as it affects the Privacy Policy of the ULTIMO ecosystem, as described in Appendix I.

4.1 Data Types

The provided ULTIMO data list consists of 7 (seven) Data Types. The **Static Data** are all the data that define the specific shuttle. This data type also includes the available in-vehicle infrastructure that addresses specific needs, such as accessibility etc.

The second data type Is **Dynamic Data**, including all the data a shuttle's sensors can produce, therefore being paramount for the success of the services and the project as a whole. Dynamic data portray the shuttles' state at a given timestamp or in a specific time window. This data type includes, but is not limited to, vehicle data such as speed, acceleration, location, occupancy, in-vehicle environmental attributes, along with attributes needed for better shuttle supervision from the operators' side.

Mobility Data include the data required for the Public Transport topology. Such data could be the itinerary schedules, tools for timetable reconfiguration, along with operation day and day type.

Booking Data include the data from the commuters' actions. The planned, calculated and actual actions can form the elements needed to measure the Quality of Service and the User Experience. The trip reason, lastly, could be included to facilitate a priority policy.

A number of other data can affect the overall quality of the itinerary, such as the outdoor weather, the current traffic situation and **Map Data** APIs. Such data constitute the **External Data**, which strongly affect the success of each transportation service.

Another important section of data is the in-vehicle infrastructure, such as the internal and external cameras, traffic cameras, radar and other sensors, that create the **Infrastructure Data**. Such data are part of a wider smart city infrastructure and can cooperate for seamless integration of the autonomous shuttles in an existing city ecosystem.

Last data type is the **Network Data**, responsible for the interaction of the vehicle with the outside world. Such data include WiFi, 5G, Bluetooth, Radio Frequency data and other data related to communications. Network Metadata are also included in this faction.

4.2 Metadata

The data acquired from the ULTIMO pilot sites, according to the ULTIMO Data List, cannot be properly interpreted without the Metadata that accompany them. A critical parameter, for example, is the frequency of data collection. However, this frequency may vary among vehicles and infrastructure sensors, or the data collection might be inconsistent for some parameters (e.g. Passengers, Traffic, Weather, Maps etc.) where a sparse data collection might be more efficient.

4.3 Personalized Data

The approach for data calculation is also very important. For instance, the shuttle speed could be updated directly from the sensors, or could be calculated from the covered distance and the according timestamps, where the dwell time in bus stops needs to be omitted. Moreover, some data, such as the kilometres travelled, need to have a distinction between itineraries with passengers on board and dead runs.

There are some data that can be tracked back to the user, via the online activity, and reveal specific information. Access to such sensitive and personal data may pose a threat to the end users. This kind of data need to be handled with special care, such as anonymization and encryption techniques. GDPR defines a list of regulations for privacy by design, such as multi-factor authentication and secure storage. Personalized data are mainly data about the commuters, such as:

- **Dynamic Data** (passenger with special needs, wheelchair etc.)
- **Booking Data** (commuter location, destination, onboard and off board location and time, ride distance and duration, etc.)
- **Infrastructure** and **Network Data** (internal/external cameras, WiFi or 5G connections, network traffic metadata). Some network sensors may also transmit personal information, such as username, password, IP and MAC addresses etc. and therefore need exhaustive consideration.

5 ULTIMO Big Data Collection System and Open Data Portal views

The Big Data Collection System and DMP will serve as a valuable resource in the ULTIMO ecosystem as it encompasses the design and deployment of the ULTIMO storage platform and the big data execution engines that will be utilized. This platform will help collect, register, organize and manage data from various internal and external sources (along with their metadata) to enhance interoperability with open data federators and streamline deployment. The mechanisms of ULTIMO Collection system are hosted in dedicated servers accommodated in CERTH/ITI and manipulated by the respective WP5 technical team.

DMPs not only address the complexities of gathering and overseeing diverse data from multiple sources but also conduct data analysis to extract valuable insights. They enable external data procurement and transfer, as well as integration with third-parties that are willing to reuse the data assets gathered and produced by the project.

The ULTIMO Data Management platform will be based in existing open source technology and especially the SHOW DMP that was developed as part of the SHOW EU-funded project, and it will take into account data from various sources, which are the following:

- **Data from the different vehicles**.
- **Human generated data** (e.g., from the passenger services).
- **Infrastructure data** (e.g., data collected from the fleet management solutions, operating centers etc).
- **Third-party data (e.g., weather data from a website)**.

Big Data is a 21st century phenomenon where the exponential growth of the gathered data necessitates the effective collection, storage, management, and analysis of all the data. In the context of the ULTIMO project, the vast majority of data will be generated from the vehicles and their sensors, with each Connected Automated Vehicle (CAV) potentially generating and consuming nearly 40 TB of data every 8 hours[1]. This means that large datasets will be generated and collected during the lifetime of the project, with most of them being processed and stored locally, while a small amount of data will be exchanged through communication networks. This exchanging actions include, mainly, some connection monitoring messages between the fleets and the ULTIMO DMP technical team.

While there are benefits to using a data management platform for handling large volume of data, there are some additional complexities that need to be handled for the system to work properly, which are:

- **Long processing** times, thus need for real-time or near real time processing and analysis.
- **Quality** and **reliability** of enormous amounts of **data streaming** into systems.
- **Diverse** data **sources**.
- **Unstructured** and **incomplete** data.

- **Data noise** caused by external factors.

5.1 Functional and Non-Functional Requirements of ULTIMO DMP

This section presents the functional and non-functional requirements that were defined for the ULTIMO DMP. The terms used within the requirements undertone their importance:

- **"Must"** signifies a requirement of paramount significance for the successful execution of the open data portal. Failure to meet this requirement will have a substantial adverse impact on the site's usability and management.
- **"May"** designates a requirement that is optional but greatly beneficial for achieving success.

Moreover, for describing the functional user requirements, we have adopted the user stories approach[26]. This approach is intuitive and aligns with practices in agile software development, ensuring that all requirements are presented in a straightforward, uniform, and standardized manner. This approach is widely accepted and used in the industry.

5.1.1 Users

The following roles are defined to specify the relevance of each requirement:

- **Reader**: Individuals who visit the portal to discover, access, and utilize data.
- **Editor**: They are typically responsible for the publication or modification of data and other content on the portal.
- **Administrator**: They oversee the technical and configuration aspects of the portal, including visual themes, metadata requirements, and more.

The aforementioned roles do not equate to a permission system, but instead they delineate the user types that are usually fundamental to a Data Management Platform (DMP) and establish the level of authority each user has in accessing and managing data. This categorization is vital for ensuring data security and facilitating the successful implementation of each CCAM-related service.

5.1.2 Functional Requirements

Functional requirements (Table 3) describe the system's capabilities and main functionalities of the ULTIMO DMP and they have been designed by taking into account the ULTIMO framework. The requirements specification is a mandatory and prerequisite procedure for the actions related to the system's architecture, design and implementation.

Table 3: Functional user requirements

Req_Id	Description
Req_001	As a **reader**, I **must** be able to create an account by using an email address, so I can have access to the functionalities of the DMP.
Req_002	As a **reader**, I **must** have the capability to generate, modify, or delete profile information to control my visibility to other DMP users.

Req_003	As a **reader**, I **must** be able to access and search the data catalog to discover available data assets quickly and easily.
Req_004	As a **reader**, I **must** be able to search for datasets using one or more keywords to locate the relevant dataset.
Req_005	As a **reader**, I **must** have read-only access to data assets that I am authorized to view.
Req_006	As a **reader**, I **must** be able to categorize data, in order to explore and find easier the necessary data.
Req_007	As a **reader**, I **must** be able to download the catalog in a format that is machine-readable, facilitating automated use.
Req_008	As a **reader**, I **must** be able to access a summary page for each dataset, allowing me to obtain information about data resources, metadata, and other pertinent documentation.
Req_009	As a **reader**, I **must** be able to download data assets in various formats (e.g., CSV, Excel) in order to get the datasets in a more efficient manner.
Req_010	As a **reader**, I **may** have access to data visualization tools to explore and analyze data in a user-friendly way.
Req_011	As a **reader**, I **may** have the option to preview data resources within their web browser, such as through a grid displaying rows and columns or a map. This allows for a basic understanding of the dataset.
Req_012	As a **reader**, I **may** be able to review (comments, etc.) individual data sets, in order to leave valuable insights for other users of the DMP.
Req_013	As a **reader**, I **may** be able to review (comments, etc.) the data catalog, in order to help other users of the DMP.
Req_014	As a **reader**, I **may** be able to save visualized data with configuration settings, in order to view them again later.
Req_015	As a **reader**, I **may** be able to propose data resources with the purpose of making them accessible to the public.
Req_016	As an **editor**, I **must** be able of adding, modifying or deleting catalog entries to make them accessible or inaccessible to DMP users.
Req_017	As an **editor**, I **must** be able to update or remove data assets, ensuring data accuracy and completeness.
Req_018	As an **editor**, I **must**be able to designate catalogue entries as private, ensuring that they are not visible to the public.
Req_019	As an **editor**, I **must** be able to define data access controls for specific data assets and user groups.
Req_020	As an **editor**, I **must** have the ability to generate new data resources so they can be accessible to other DMP users.
Req_021	As an **editor**, I **must** be able to review user-proposed data resources, enabling me to make informed decisions regarding their inclusion or exclusion.
Req_022	As an **editor**, I **must** be able to answer to user-proposed data resources, enabling me to inform the user regarding their inclusion or exclusion.
Req_023	As an **editor**, I **must** be able to track and audit changes made to data assets for compliance and accountability.
Req_024	As an **editor**, I **may** be able to generate data visualizations with the purpose of making them available to users through the data catalog or data resource

	pages.
Req_025	As an **editor**, I **may** be able to modify the status of user-proposed data resources, facilitating user notifications and maintaining records.
Req_026	As an **editor**, I **may** collaborate with other editors and readers by sharing data assets and insights within the platform.
Req_027	As an **editor**, I **may** be able to respond to user reviews and comments in order to help other users.
Req_028	As an **editor**, I **may** be able to assess and authorize the display of user comments and feedback, with the aim of retaining only constructive comments.
Req_029	As an **editor**, I **may** be able to determine the accessibility of my data resources, allowing me to grant access to specific registered users.
Req_030	As an **administrator**, I **must** have full control over user access permissions and roles within the DMP, to enhance overall security.
Req_031	As an **administrator**, I **must** be able to create and manage user accounts, including adding, modifying, and deactivating them.
Req_032	As an **administrator**, I **must** be able to generate, modify, and archive metadata categories, enhancing the categorization of data for users.
Req_033	As an **administrator**, I **must** be able to manage the visibility of consumer-generated data presentations for other consumers, ensuring the privacy and security of the DMP.
Req_034	As an **administrator**, I **must** be able to monitor and manage system performance and data storage capacity to ensure optimal performance.
Req_035	As an **administrator**, I **must** be able to generate and export usage and compliance reports for auditing purposes.
Req_036	As an **administrator**, I **may** be able to to customize global settings for the data catalogue, encompassing aspects like colour theme, branding/logo, and titling, with the aim of enhancing the user experience.
Req_037	As an **administrator**, I **may** be able to customize further global settings for the data catalog, such as custom Cascading Style Sheets (CSS), homepage display elements and layout, displayed catalog elements, and more, all geared towards enhancing the user experience.
Req_038	As an **administrator**, I **may** be able to set default sorting preferences for data catalogues, aiming to enhance the user experience.
Req_039	As an **administrator**, I **may** set up automated backups and disaster recovery procedures to protect data integrity.
Req_040	As an **administrator**, I **may** be able to monitor and analyze user feedback to gain a deeper understanding of the DMP.

5.1.3 Non-Functional Requirements

Non-functional requirements are a set of specifications that present the operational capabilities and constraints of the system and aim to enhance its functionality, and if they are not met, the system may prove useless. The following non-functional requirements have been defined for the ULTIMO DMP:

- **Performance**: Determines the speed at which the system or a specific component thereof, reacts to specific user actions within a designated workload.

- **Scalability**: Evaluates the highest workloads at which the system will continue to satisfy performance requirements.
- **Openness**: The Platform should support the open data, so it can be fully accessible for the public.
- **Reliability**: The system should run and be accessible at all possible situations.
- **Maintainability**: The system's capability to either maintain or restore itself to a predefined state in the event of a failure or during maintenance activities.
- **Availability**: Ensure that the platform is accessible to the users at a given point in time.
- **Extensibility**: The addition of new features to the system should be possible.
- **Robustness**: The ability of the system to react to varying levels of disturbance and remain functioning under all conditions.
- **User-friendly**: Users and potential future Passengers without technical expertise should be able to easily use and navigate throughout the platform.
- **Security** : Ensure that the system and all data are can be protected at all times against malicious attacks or unauthorized access.
- **Usability**: Ensure that users, existing and new and passengers can easily learn and use the system from any device.
- **External**: The portal should be reachable via a city-provided internet domain name.

5.2 Abstract Architecture

The architectural depiction of the ULTIMO Big Data Collection System (BDCS) is presented in Figure 5. Within the ULTIMO ecosystem, three operational sites have been designated for the deployment of autonomous fleets during specific time intervals. These operational sites encompass **Herford in Germany, Grorud-Valley in Oslo, Norway, and Geneva**.

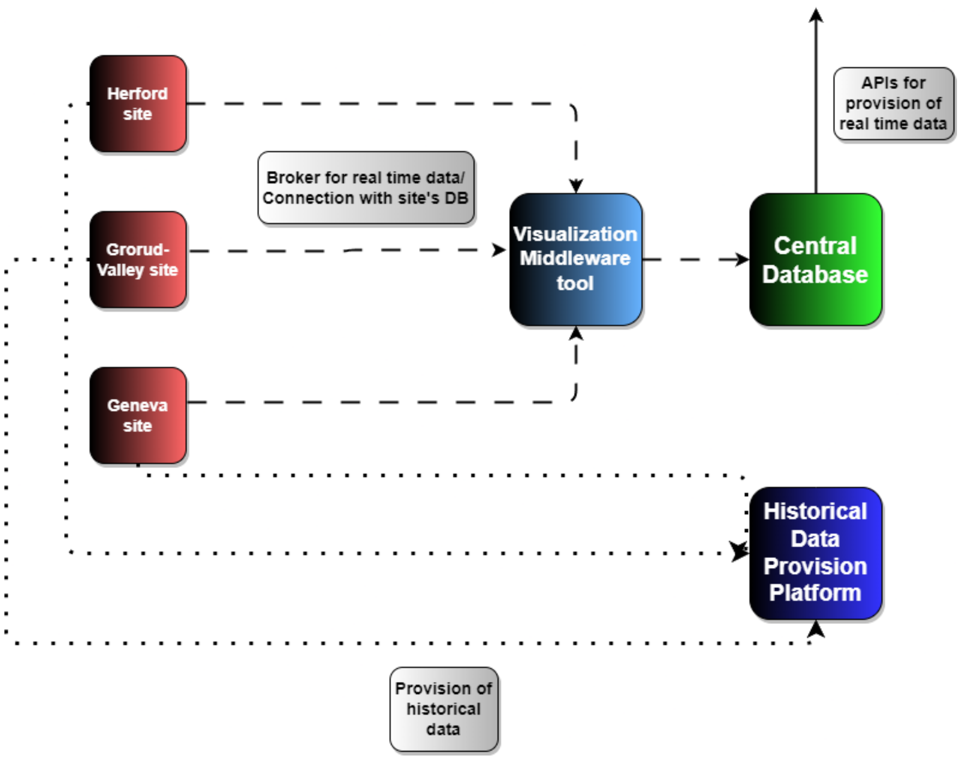

Figure 5: Abstract overview of ULTIMO BDCS Architecture

Each of these sites is required to establish dual modes of connectivity with the ULTIMO BDCS. The first mode pertains to real-time connectivity during the fleet's operational phase. The establishment of real-time connectivity can be achieved through two distinct approaches. The initial method involves the provisioning of data via a dedicated broker, responsible for transmitting data from the vehicles to ULTIMO's Central Database. Implementing this approach necessitates a degree of effort on the part of the respective sites, encompassing the configuration of the broker, alignment with message format standards, and determination of an appropriate data transmission frequency. It is crucial to strike a balance, ensuring data sufficiency without overburdening ULTIMO BDCS with excessive computational resource demands.

The second approach for the real-time data provision entails a direct connection between the technical team of ULTIMO BDCS and the local database at each operational site. In this scenario, the operational site is relieved of any direct-action requirements. Data will be seamlessly redirected to ULTIMO BDCS in accordance with the specific data message formats and security measures employed by each system. The sole responsibility of the operational sites in this context is the provision of comprehensive system documentation and detailed descriptions of their system specifications.

The assessment of real-time connections' quality and the data provisioning process will be subject to monitor through the utilization of a middleware visualization tool. This tool will facilitate the evaluation of the stability of brokers and streamline the pre-operational inspection procedure, rendering it more efficient and manageable.

All real-time data will be archived within ULTIMO's Central Database. Subsequently, this repository will serve as the source for forwarding data to the Historical Data Platform. This strategic arrangement ensures that entities necessitating data access for purposes such as data analysis, impact assessment, and KPI computation (once these KPIs have been comprehensively defined), as well as the execution of mutually agreed-upon services, can readily obtain the requisite data.

To facilitate the operation of services reliant on real-time data, dedicated APIs will be deployed. This measure is taken to circumvent the need for multiple entities to directly connect to the Central Database. In the Central Database, the respective teams for the implementation of state-of-the-art services, for the impact assessment and for the KPIs calculation should have partial or full access. This approach is adopted for reasons pertaining to both security and computational resource management.

Within the framework of the Historical Data Platform, operational sites will, when feasible, contribute historical data from previous operational periods. Additionally, they will supply data message formats or simulation data, if accessible. This collaborative effort is designed to facilitate the exchange of valuable information, and assure fast learning cycles, while upholding the security of this information and consolidating it within a centralized repository.

5.3 Candidate tools for ULTIMO BDCS

Revisiting Figure 5, we hereby provide a detailed exposition of the prospective tools individually associated with each of the illustrated entities. With regard to the technological aspect of the broker:

- **MQTT (Message Queuing Telemetry Transport) [55]**: MQTT is a lightweight and efficient messaging protocol commonly used in IoT (Internet of Things) applications and industrial automation. It operates on a publish-subscribe model, allowing devices to communicate by publishing messages to specific topics and subscribing to receive messages on those topics. MQTT's low overhead makes it ideal for constrained environments where bandwidth and power are limited, enabling real-time data exchange between devices and applications.
- **Interactive Brokers (IBKR) [56]**: Interactive Brokers is a popular brokerage platform that provides a wide range of tools and services for traders and investors. It offers a powerful trading platform with advanced charting, research, and analysis tools. IBKR also provides access to various markets, including stocks, options, futures, forex, and more. Its API (Application Programming Interface) allows developers to create custom trading algorithms and integrate third-party applications, making it a versatile choice for both individual and institutional traders.
- **MetaTrader4 (MT4) [57]**: MetaTrader 4 is a widely used trading platform in the forex and CFD (Contract for Difference) markets. It offers a user-friendly interface, customizable charts, technical indicators, and automated trading capabilities through Expert Advisors (EAs). Traders can analyze market data, execute trades, and implement automated strategies efficiently. MT4's popularity is attributed to its robust trading features, extensive user community, and support for algorithmic trading, making it a favored choice among forex traders worldwide.

As concerns the Middleware Visualization tool, the options are the following:

- **Lenses [58]**: Lenses is a powerful middleware visualization tool designed for Apache Kafka, a widely used event streaming platform. Lenses provides a user-friendly interface for managing, monitoring, and troubleshooting Kafka clusters and the data flowing through them. With Lenses, users can easily visualize Kafka topics, consumer groups, and data pipelines, making it simpler to understand the real-time data streaming within their organization. It offers features like data lineage tracking, schema management, and SQL query capabilities, empowering users to gain valuable insights and ensure the reliability of their Kafka-based data infrastructure.
- **Kibana [59]**: Kibana, part of the Elastic Stack, is a popular open-source data visualization tool primarily used for log and event data analysis. It provides a versatile platform for visualizing and exploring data stored in Elasticsearch. Users can create interactive dashboards, charts, and maps to gain insights into their data. Kibana is commonly used for log analysis, security monitoring, and operational data visualization, making it an essential tool for organizations leveraging Elasticsearch for log and event management.
- **Grafana [60]**: Grafana is a widely adopted open-source platform for monitoring and observability, offering rich visualization capabilities for various data sources, including time-series databases like Prometheus, InfluxDB, and Graphite. It enables users to

build customizable dashboards and panels, helping to visualize and analyze metrics, logs, and traces. Grafana's flexibility and support for plugins make it a versatile choice for monitoring infrastructure, applications, and services, fostering better decision-making and system performance optimization.

As concerns the Database technologies, the main components are:

- **MongoDB [6]**: MongoDB is a popular NoSQL database that is designed for flexibility and scalability. It uses a document-oriented data model, which allows data to be stored in a semi-structured format, making it well-suited for handling a wide range of data types and evolving schemas. MongoDB is known for its high performance and horizontal scalability, making it an excellent choice for applications dealing with large volumes of data and high read/write workloads. It is often used in modern web applications, IoT, and real-time analytics due to its ability to handle unstructured or semi-structured data efficiently.
- **MySQL [61]**: MySQL is an opensource relational database management system (RDBMS) known for its speed, reliability, and ease of use. It follows the traditional relational database model with structured tables and SQL-based querying. MySQL is widely used in web applications, content management systems (CMS), and e-commerce platforms. It offers features like ACID compliance, replication, and clustering for high availability and data redundancy. MySQL's extensive community and robust support make it a dependable choice for businesses of all sizes.
- **PostgreSQL [62]**: PostgreSQL is an advanced opensource relational database known for its powerful features, extensibility, and support for complex data types. It is highly extensible, allowing users to define custom data types, operators, and functions. PostgreSQL supports ACID transactions and offers various indexing techniques for efficient querying. It is often chosen for applications that require complex data processing, geographic information systems (GIS), and data warehousing. Additionally, PostgreSQL's active and dedicated community ensures ongoing development and support for the database system.

Finally, the main options for the Historical Data Platform are:

- **CKAN (Comprehensive Knowledge Archive Network) [63]:** CKAN is an open-source data management platform designed for efficiently storing, sharing, and managing large sets of data and files. It's particularly useful for organizations, governments, and institutions looking to publish and manage open data initiatives. CKAN provides a user-friendly web interface for data cataloguing, metadata management, and data publishing. It supports versioning, access controls, and data previews, making it easier to find and reuse datasets. CKAN is a valuable tool for promoting transparency and data sharing within communities and organizations.
- **Dataverse [64]:** Dataverse is an opensource web application designed for research data management and sharing. It allows researchers to create data repositories to store, share, publish, and cite their research data. Dataverse provides data citation support, role-based access control, and dataset versioning. It is particularly popular in the academic and research communities for ensuring data preservation, accessibility, and reproducibility.

- **Amazon S3 (Simple Storage Service) [65]:** Amazon S3 is a cloud-based object storage service provided by Amazon Web Services (AWS). While it's primarily known for scalable and durable storage of various data types, it can also serve as a historical data platform. Users can store and archive large volumes of historical data in S3 buckets, taking advantage of features like data lifecycle policies for automatic data retention and archival. S3 is often used in conjunction with AWS data analytics and processing services to gain insights from historical data stored in the cloud.

5.4 CKAN platform

The final choice of the tool for the Historical Data Platform is CKAN. In this section, we present the main characteristics and features of CKAN.

5.4.1 Registration

In order to gain access to the platform, users who are not currently registered must undergo a registration process. This procedure entails the submission of specific information, including:

- **Username**: The registrant is required to select a username comprised solely of alphanumeric characters, hyphens (-), and underscores (_).
- **Full Name**: This name will be publicly displayed on the user's profile.
- **E-Mail Address**: This information is to remain confidential and will not be disclosed to other platform users.
- **Password**: The registrant must input the same secured password in both designated fields.

Upon successful registration, the reader will subsequently be able to log in utilizing their chosen username and password. It is worth noting that the option to retain this pertinent information for registered readers is available but not obligatory.

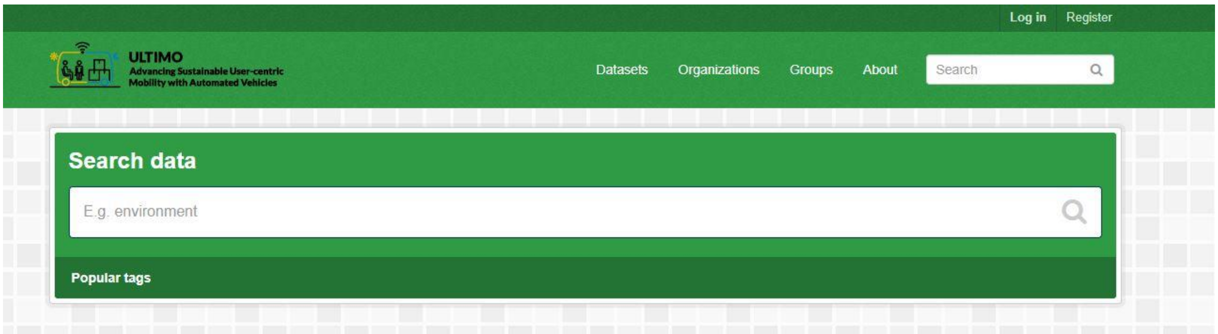

Figure 6: The introduction page of ULTIMO CKAN

5.4.2 Organization

Organizations serve as the primary mechanism for governing access to, as well as the creation and modification of datasets within CKAN. Within this system, each dataset is associated with a single organization, and each organization is vested with the authority to regulate access to the datasets under its purview.

Datasets are categorized into either public or private datasets. Public datasets are readily viewable by any registered user of CKAN. In contrast, private datasets can only be accessed by logged-in readers who are concurrently recognized as members of the specific organization to which the dataset is affiliated. Importantly, private datasets remain concealed from dataset searches unless the logged-in reader, or the reader identified through an API key, possesses the requisite permissions to access them.

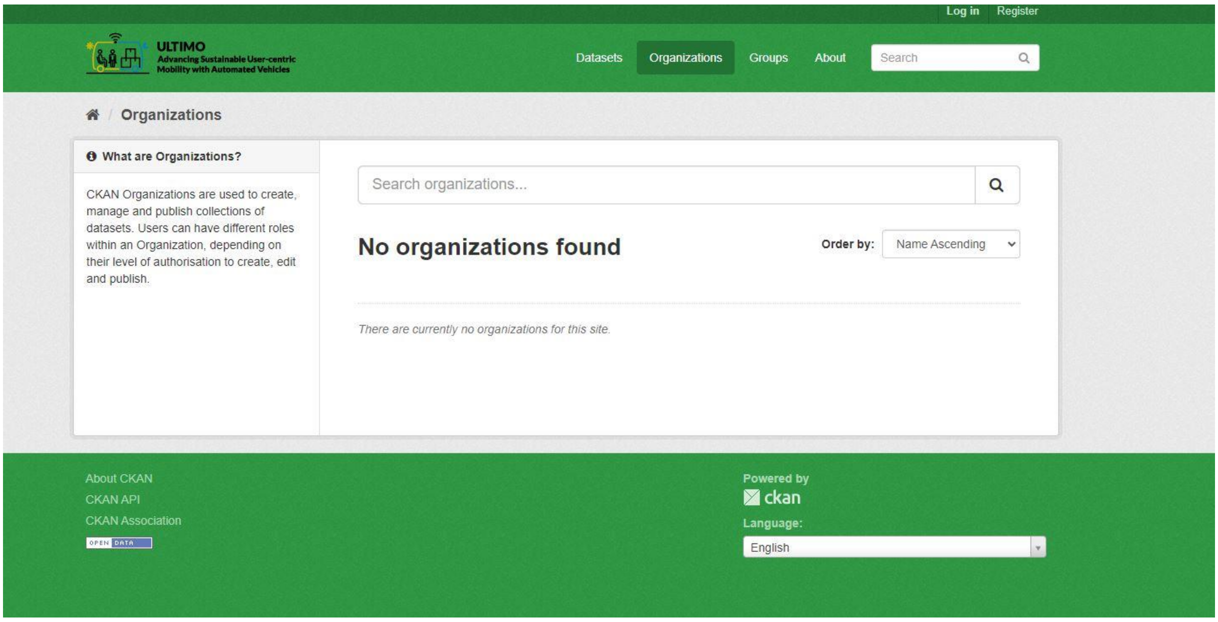

Figure 7: The Organization menu of ULTIMO CKAN

Upon joining an organization, a reader is assigned one of three distinct roles by an organization administrator: member, editor, or admin.

A member has the capability to:

- **Access** the **organization's private datasets.**

An editor enjoys the privileges of a member and, in addition, possesses the authority to:

- **Introduce new datasets** to the organization.
- **Modify or delete** any of the organization's datasets.
- **Alter the visibility status** of datasets, switching them between public and private.

An organization administrator is granted the full range of privileges held by an editor, and, in addition, has the power to:

- **Include users** within the organization, with the option to **designate their role** as a member, editor, or admin.
- **Modify the roles** of any user within the organization, including other administrators.
- **Remove members, editors, or fellow administrators** from the organization.
- **Modify the organization's attributes**, such as its title, description, or image.
- **Delete** the organization itself.

When a user initiates the creation of a new organization, they automatically assume the initial role of administrator for that organization.

5.4.3 Group and Organization

CKAN Groups serve as a valuable tool for the creation and management of dataset collections. These collections may encompass catalogued datasets related to specific projects, teams, themes, or provide a straightforward means for users to locate and search their own published datasets.

It is major to underline the distinction between Organizations and Groups, primarily rooted in the realm of authorization. Within an Organization, the editor or administrator possesses the capability to generate new datasets and make edits to datasets associated with that particular Organization. Conversely, in a Group, the purview of the editor or administrator is limited to removing existing datasets from the web and adding them to the group or removing datasets from the group. Creating new datasets or editing datasets belonging to the broader community is beyond the scope of a Group's editor or administrator. Additionally, it is noteworthy that only Organizations have the capacity to house private datasets, which are exclusively accessible to members of the respective Organization.

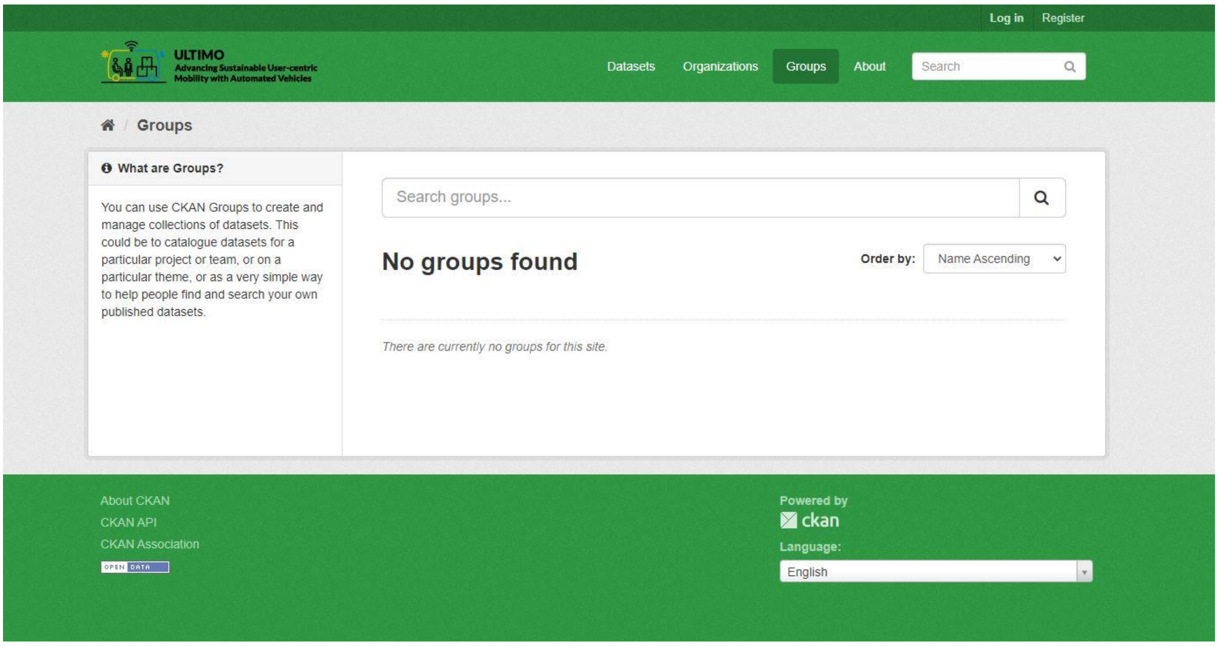

Figure 8: The Group menu of ULTIMO CKAN

Organizations exercise control over which databases can be introduced, modified, and published by users. In the CKAN system, each dataset is required to be affiliated with precisely one Organization. If a user holds the position of an editor or administrator within an Organization, they are empowered to create new datasets within that Organization and exercise authority over the editing and publication of datasets belonging to that Organization.

Conversely, Groups are designed with the purpose of curating datasets, bringing together related datasets into a cohesive unit. When a user assumes the role of a group editor or administrator, they may add existing datasets to the group from the web and remove datasets from the group. However, they lack the ability to introduce entirely new datasets to the platform or modify datasets contained within the group. Groups are intended to be utilized by a community of users on the platform, specifically those who consume information rather than those responsible for publishing datasets, with the goal of facilitating the aggregation of similar datasets for easy access and reference.

5.4.4 Dataset

To successfully complete the addition of a new dataset, the following essential information must be provided:

- **Title**: The dataset's title should be concise yet distinctive, ensuring uniqueness across CKAN.
- **Description**: A comprehensive description of the dataset should be included, encompassing details such as the data's source and any pertinent information that users will require when utilizing the dataset.
- **Tags**: Users have the option to add tags that facilitate data discovery and establish associations with related datasets. Tags should be separated by hitting the <return> key. Any incorrectly entered tags can be deleted using the provided delete button prior to saving the dataset.
- **License**: It is imperative to specify the dataset's licensing information to inform users about how they can legally utilize the data. This field should provide a dropdown menu of predefined licenses. If a user needs to use a license not present in the list, they should contact the site administrator for assistance.
- **Organization**: If the user is a member of an organization, they can select the appropriate organization from a dropdown menu to designate ownership of the dataset. It is crucial to ensure that the default selection is the correct organization before proceeding.

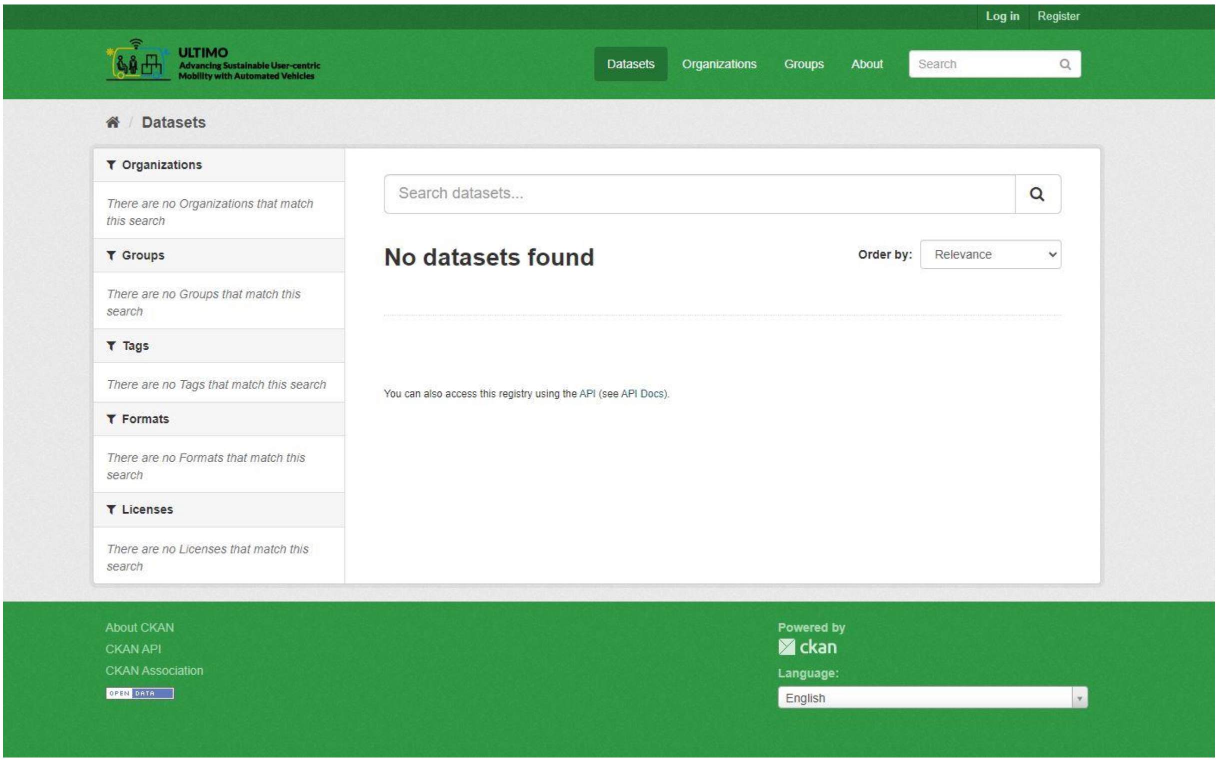

Figure 9: The Dataset menu of ULTIMO CKAN

After this initial input, CKAN proceeds to display the "Additional Data" screen, which includes the following elements:

- **Visibility**: Users must specify the visibility of the dataset. A "Public" dataset is accessible to any user of the site, while a "Private" dataset can only be viewed by members of the organization that owns the dataset. Private datasets will not appear in searches conducted by other users.
- **Author**: Provide the name of the individual or organization responsible for producing the data.
- **Author email**: Enter an email address for the author, which users can use to send queries related to the dataset.
- **Maintainer / Maintainer email:** If necessary, provide details for a second person responsible for maintaining the dataset, including their name and email address.
- **Custom Fields**: Users can include additional fields for the dataset by specifying the field name and its corresponding value in this section.

For existing datasets, users have the ability to explore them by either previewing or downloading the dataset.

5.4.5 Visualization

CKAN's resource views encompass various available visualizations of resource data or file contents. These views can take the form of tables, bar charts, maps, and more. The primary features of resource views within CKAN are as follows:

- **Multiple Views for One Resource**: It is possible for a single resource to have multiple views of the same data. For instance, a dataset may include both a tabular grid view and various graphical representations.
- **Editorial Control**: Dataset editors have the authority to select which views to display, adjust their order, and configure them individually to meet specific visualization requirements.
- **External Embedding**: Individual views can be embedded on external websites or platforms, making it convenient to share and showcase data visualizations beyond the CKAN environment.

Different types of resource views are implemented through custom plugins, which can be activated on a particular CKAN site. Once these plugins are integrated, site administrators have the flexibility to determine default views and assess the suitability of a resource for specific types of visualizations. For instance, tabular data may be suitable for a table view if it is uploaded to the DataStore, while spatial data may be more apt for a map view. In the current version of CKAN, there are three core visualization options:

- **Grid**: This view presents data in a tabular format, similar to a spreadsheet or database table.
- **Graph**: The graph view allows for the creation of various types of graphical representations, such as bar charts, line graphs, or pie charts.
- **Map**: The map view can be configured in two ways: with or without cluster markers. It provides a visual representation of spatial data.

CKAN also offers numerous additional view plugins, some of which include:

- **Dashboard**: This view combines multiple views into a single dashboard, offering a comprehensive overview of data.
- **PDF Viewer**: It enables the viewing of PDF files directly on the resource page, facilitating easy access to document-based data.
- **GeoJSON Map**: This view renders GeoJSON files on an interactive map, making it suitable for displaying geographical data.
- **Choropleth Map**: This visualization tool is designed for displaying data from the DataStore on a choropleth map, allowing for the depiction of regional variations.
- **Basic Charts**: It provides a range of alternative graph types and rendering options for diverse visualization needs.

These resource views and plugins enhance CKAN's capability to effectively visualize and present data in various formats, catering to the specific requirements of users and datasets.

5.5 Next steps

The principal forthcoming action required of all operational sites involves the completion of the "Questionnaire for ULTIMO Operations Sites: ULTIMO Data Management Platform," as detailed in Appendix III. This questionnaire serves as a critical step in achieving a comprehensive understanding of the ULTIMO ecosystem. It will enable the determination of the mode of connection (interfaces, plugins etc.) with each site and the identification of specific challenges unique to each.

The questionnaire comprises three primary sections. The initial section pertains to the fundamental characteristics of each site. This section encompasses the provision of information regarding the site's fleet, including the number of vehicles, their OEMs, the geographical area covered by the fleet, and the scheduled operational hours per day.

The second section of the questionnaire focuses on the specifications of the local Data Collection System utilized by the sites. It encompasses details about the sources of generated data, the primary attributes of the data and their respective schemas, as well as the frequency of data generation. Additionally, knowledge of the main technology tools employed at the site is essential to facilitate the seamless integration of ULTIMO DMP.

Finally, the last section delves into technical aspects concerning the method of connection with ULTIMO DMP, encompassing protocols, data formats, and the selection of common tools. The process of tool selection and the refinement of the ULTIMO DMP structure is currently in progress. The primary factors guiding the tool selection process encompass their open-source availability, their ability to facilitate the exchange of extensive data among various entities, and the extent of familiarity that the technical team at CERTH/ITI possesses with these tools. This process will reach its conclusion upon the receipt and evaluation of feedback from the operational sites via the Questionnaire by the CERTH/ITI team.

The initial tool slated for implementation is the Historical Data Platform, CKAN. As previously outlined, CKAN will serve as the central repository through which all sites can share data from past operational periods, thereby enabling the dissemination of information regarding data formats and data message structures. The availability of historical data holds the potential to expedite the development of services within the ULTIMO consortium. The technical aspects of the CKAN installation have been successfully concluded, and its distribution and population will occur during the forthcoming period among ULTIMO consortium members.

6 Conclusions

This deliverable describes the activities within "T5.4 Deployment Data Collection". The task objective is to analyse the data project requirements taking into account the envisioned services of WP4, along with the evaluation requirements and related KPIs from the pilot sites. Moreover, this task provides a foundation of services (e.g. Kafka, Apache Spark, machine learning libraries, etc.). Specifically, this deliverable provides the design and implementation of the ULTIMO storage platform and the big data execution engines that will be used.

The SHOW data management platform has been utilised as a design baseline, taking into account the data collected and stored there, along with the data models derived. A state-of-the-art analysis for data collection has been conducted, whereas the adopted data models have been described, based on the Transmodel and NeTEx standards. Then, the ULTIMO Data Registry has been presented, along with the ULTIMO Big Data Collection System and Open Data Portal. A detailed ULTIMO Data List has been also presented that will form the base for data collection activities within the development of services (WP4) and pilot site activities (WP5). Specifically, a thorough description of the on-going work of data types has been provided, while the first implemented version of the data collection mechanisms, the Big Data Management Platform, the Open Data portal based on CKAN have been described.

The development of the Big Data Management Platform and the Open Data portal of the ULTIMO infrastructure is an iterative procedure, which is going to be further enhanced during the next period, taking into account the envisioned architecture, the services development within WP4, as well as the related KPIs. The final version of this deliverable will be provided in "D5.6 Big Data Collection Platform and Data Management Portal - Final version", where all components will be described in more detail in terms of data formats and functionalities, as well as data transfer functions, taking into account the developed services and the related data.

References

[1] D. C. Luckham, The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Reading, MA, USA: Addison-Wesley, 2001.

[2] G. Cugola and A. Margara, “Tesla: A formally defined event specification language,” in Proc. 4th ACM Int. Conf. Distrib. Event-Based Syst. (DEBS’10), 2010, pp. 50–61.

[3] G. Cugola and A. Margara, “Processing flows of information: From data stream to complex event processing,” ACM Comput. Surv., vol. 44, no. 3, pp. 15:1–15:62, Jun. 2012.

[4] M. Roudjane, D. Rebaïne, R. Khoury and S. Hallé, "Predictive Analytics for Event Stream Processing," 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC), Paris, France, 2019, pp. 171-182, doi: 10.1109/EDOC.2019.00029.

[5] M. Li, M. Liu, L. Ding, E. A. Rundensteiner and M. Mani, "Event Stream Processing with Out-of-Order Data Arrival," 27th International Conference on Distributed Computing Systems Workshops (ICDCSW'07), Toronto, ON, Canada, 2007, pp. 67-67, doi: 10.1109/ICDCSW.2007.35.

[6] MongoDB. (2020, December 2020). The most popular database for modern apps. https://www.mongodb.com/

[7] Apache Hadoop. (2021, 4 January). Apache Hadoop. https://ha-doop.apache.org/

[8] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

[9] Apache Spark. (2021, 4 January). Apache Spark™ - Unified Analytics Engine for Big Data. https://spark.apache.org/

[10] Zhao, H., Yao, L., Zeng, Z., Li, D., Xie, J., Zhu, W., & Tang, J. (2020). An edge streaming data processing framework for autonomous driving. Connection Science, 1-28.

[11] Zaharia, M., Das, T., & Li, H. (2013). Discretized streams: Fault-tolerant streaming computation at scale. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles.

[12] Apache Storm. (2021, 4 January). Apache Storm. https://storm.apache.org/

[13] BULUT, I. S., & ILHAN, H. (2019, September). Cloud Based Vehicle and Traf-fic Information Sharing Application Architecture for Industry 4.0 (IoT). In 2019 International Conference on Information and Telecommunication Technolo-gies and Radio Electronics (UkrMiCo) (pp. 1-7). IEEE.

[14] Apache Samza. (2021, 4 January). Apache Samza. https://samza.apache.org/

[15] Noghabi, S. A., Paramasivam, K., Pan, Y., Ramesh, N., Bringhurst, J., Gupta, I., & Campbell, R. H. (2017). Samza: stateful scalable stream processing at LinkedIn. Proceedings of the VLDB Endowment, 10(12), 1634-1645.

[16] Debauche, O., El Moulat, M., Mahmoudi, S., Boukraa, S., Manneback, P., & Lebeau, F. (2018). Web monitoring of bee health for researchers and bee-keepers based on the internet of things. Procedia computer science, 130, 991-998.

[17] Apacha Flume. (2021, 4 January). Welcome to Apache Flume - Apache Flume. https://flume.apache.org/

[18] Kumar, S., & Goel, E. (2018, April). Changing the world of Autonomous Vehi-cles using Cloud and Big Data. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 368-376). IEEE.

[19] Samant, S. S., Chhetri, M. B., Vo, Q. B., Kowalczyk, R., & Nepal, S. (2018, July). Towards end-to-end qos and cost-aware resource scaling in cloud-based iot data

processing pipelines. In 2018 IEEE International Conference on Services Computing (SCC) (pp. 287-290). IEEE.

[20] Apache Flink Documentation. (2021, 4 January). Apache Flink 1.12 Docu-mentation: Apache Flink Documentation. https://ci.apache.org/pro-jects/flink/flink-docs-release-1.12/

[21] Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012, August). Fog computing and its role in the internet of things. In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13-16).

[22] Battulga, D., Miorandi, D., & Tedeschi, C. (2020, February). FogGuru: a Fog Computing Platform Based on Apache Flink. In 2020 23rd Conference on In-novation in Clouds, Internet and Networks and Workshops (ICIN) (pp. 156-158). IEEE.

[23] Apache Kafka. (2020, December 23). Apache Kafka. https://kafka.apache.org/

[24] Amazon Kinesis. (2021, 4 January). Amazon Kinesis. https://aws.ama-zon.com/kinesis/

[25] Li, C., Luo, Q., Mao, G., Sheng, M., & Li, J. (2019). Vehicle-mounted basestation for connected and autonomous vehicles: Opportunities andchallenges. IEEE Wireless Communications, 26(4), 30-36.

[26] Cohn, M. (2004). User stories applied: For agile software development. Addison-Wesley Professional.

[27] SHOW (2020). D4.1: Open modular system architecture and tools. Deliverable of the Horizon-2020 SHOW project, Grant Agreement No. 875530.

[28] Transmodel. (2023, September 30). Transmodel at a glance. http://www.trans-model-cen.eu/transmodel-at-a-glance/

[29] NeTEx. (2023, September 30). NeTEx. Network Timetable Exchange. http://netex-cen.eu/

[30] INSPIRE. (2023, September 30).INSPIRE- Transport Network Data. http://www.pti.org.uk/system/files/meeting_files/papers/090209_-_PTIC_-__INSPIRE_-_v1_%28Issued%29.pdf

[31] TAP TSI. (2023, September 30). TELEMATIC APPLICATION FOR PASSENGER (TAP TSI). http://taf-jsg.info/?page_id=51

[32] DATEX II. (2023, September 30).DATEX II, webinar series. https://www.datex2.eu/

[33] DVC (2023, September 30). Data Communication on Vehicles. http://www.transmodel-cen.eu/standards/dcv

[34] IFORT (2023, September 30). Identification of Fixed Objects in Public Transport. http://www.transmodel-cen.eu/standards/ifopt

[35] SIRI (2023, September 30). Standard Interface for Real-time Information. http://www.transmodel-cen.eu/standards/siri/

[36] OJP Standard (2023, September 30).Open API for distributed journey planning. http://www.transmodel-cen.eu/standards/ojp/

[37] OPRA Standard (2023, September 30).Operating Raw Data and statistics exchange. http://www.transmodel-cen.eu/standards/opra/

[38] Eurobus. (2023, September 30).Eurobus/Transmodel: 1992-1994 (DGXIII). http://www.transmodel-cen.eu/overview/history/eurobus/

[39] Cassiope. (2023, September 30). CASSIOPE: 1989 – 1992 (DGXIII). http://www.transmodel-cen.eu/overview/history/cassiope/

[40] AVRIS and SIJPRIS (2023, September 30).Transmodel implementation in Slovenia. http://www.transmodel-cen.eu/implementations/slovenia

[41] SPIDER STIB (2023, September 30). Transmodel implementation in Belgium. http://www.transmodel-cen.eu/implementations/belgium/
[42] NOPTIS (2023, September 30).Transmodel implementation in Sweden. http://www.transmodel-cen.eu/implementations/sweden/
[43] MobilitX (2023, September 30). Transmodel implementation in Marseille. http://www.transmodel-cen.eu/implementations/france/marseille
[44] TITAN (2023, September 30).Transmodel implementation in Lyon. http://www.transmodel-cen.eu/implementations/france/lyon
[45] CHOUETTE. (2023, September 30). CHOUETTE.http://www.transmodel-cen.eu/implementations/tools/chouett
[46] Kizoom, N., & Miller, P. (2008). A Transmodel based XML schema for the Google Transit Feed Specification with a GTFS/Transmodel comparison. Kizoom Ltd., London
[47] Benvenuti, F., Diamantini, C., Potena, D., & Storti, E. (2017). An ontology-based framework tosupport performance monitoring in public transport systems. Transportation Research Part C: Emerging Technologies, 81, 188-20
[48] Transmodel. (2023, September30). Common Concepts, Public Transport, Timing Information and Vehicle Scheduling. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2015/01/TUTORIAL-Part1-3-v0.2.pdf
[49] Transmodel. (2023, September 30). Operations Monitoring and Control.http://www.transmodel-cen.eu/wpcontent/uploads/sites/2/2015/01/TUTORIAL_Part4_v2.1.pdf
[50] Transmodel. (2023, September 30). Fare Management. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2019/10/TUTORIAL_Part5_v2.3.pdf
[51] Transmodel. (2023, September 30). Passenger Information. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2019/10/TUTORIAL_Part6_v2.2.pdf
[52] Transmodel. (2023, September 30). Driver Management. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2015/01/TUTORIAL_Part7_v2.1.pdf
[53] Transmodel. (2023, September 30). Management Information and Statistics. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2015/01/TUTORIAL_Part8_v2.1.pdf
[54] Transmodel team, (last access 30September 2023). Definitions of concepts. http://www.transmodel-cen.eu/wp-content/uploads/sites/2/2015/01/TRM6_DataDefinitions.pdf
[55] Egli, Peter. (2017). MQTT - Message Queueing Telemetry Transport Introduction to MQTT, a protocol for M2M and IoT applications. doi: 10.13140/RG.2.2.13210.54721.
[56] Interactive Brokers (IBKR). (2023, September 30).IBKR Trading Platforms.https://www.interactivebrokers.hu/en/trading/trading-platforms.php
[57] Metatrader 4 (MT4). (2023, September 30). Trading Platform. https://www.metatrader4.com/en/trading-platform
[58] Lenses. (2023, September 30). Use Cases. https://lenses.io/usecases/
[59] Kibana. (2023, September 30). https://www.elastic.co/kibana
[60] Grafana. (2023, September 30). https://grafana.com/
[61] MySQL. (2023, September 30). https://www.mysql.com/
[62] PostgreSQL. (2023, September 30). https://www.postgresql.org/about/
[63] Comprehensive Knowledge Archive Network (CKAN). (2023, September 30). Features. https://ckan.org/features
[64] Dataverse.(2023, September 30). Services https://www.dataverse.gr/index.php/services-2/

[65] Amazon S3 (Simple Storage Service). (2023, September 30). https://aws.amazon.com/s3/?nc=sn&loc=1

Appendix I

Privacy Policy

This platform has been created and is managed by ULTIMO Consortium ("ULTIMO"), its use being subject to these Terms and Conditions of Use.

Please read carefully the Terms and Conditions of Use presented here.

Scope of the Privacy Policy

The Privacy Policy consists of two parts: the part for passengers and their data and the part of pilot sites and the insurance of the data provided in order to design the ULTIMO's services.

ULTIMO is dedicated to safeguarding the privacy of individuals and the data from pilot sites, ensuring the security of personal data, whether it is stored electronically in our databases or physically on our premises. In line with both national and EU legal regulations, including the General Data Protection Regulation (EU) 679/2016, referred to as "the Regulation," ULTIMO has released this transparent and lawful Privacy Policy. The primary objective is to offer comprehensive information to individuals and pilot sites regarding the data collection and subsequent processing carried out while delivering our public services.

Acceptance of Terms and Conditions of Use

The ULTIMO User recognizes that through the act of registering and consistently utilizing ULTIMO and its applications; they acknowledge and consent to abide by these Terms and Conditions of Use. ULTIMO partners acknowledge that by granting access to their data lists, they too are accepting these Terms and Conditions of Use.

Accuracy and integrity of information

While all the means at our disposal have been used to ensure the accuracy and integrity of the information on the system, ULTIMO assumes no responsibility whatsoever if the information provided contains errors, inaccuracies or inaccuracies, especially information has been provided by partners.

The content contained in this system does not constitute and can never be construed as an advice, guarantee, commitment or suggestion given by ULTIMO, and there is no room for any liability on the part of ULTIMO.

Platform access

ULTIMO retains the sole and perpetual authority to suspend, limit, or terminate access to the system, whether in its entirety or specific functionalities, particularly for purposes of system maintenance, updates, or repairs. ULTIMO operators have the discretion to terminate the system either definitively or temporarily without prior notice, without infringing upon any contractual obligations established with third parties.

Intellectual Property Rights

All copyrights and other intellectual property rights inherent to the texts, graphics, questionnaires, images and other content made available, regardless of how they are represented graphically, are owned by ULTIMO or are used by ULTIMO with permission or as owner or as third party with the right to do so.

The use or transmission, as well as the availability of trademarks, logos or other content contained do not grant, nor can be interpreted as granting to public any intellectual property rights that may be registered, licensed, authorized or otherwise similar rights thereto.

Any attempts to alter, reproduce, transmit, copy, sell or otherwise misrepresent the contents of the system and protected by this provision, as well as any action that may cause damage to ULTIMO and jeopardize the integrity are strictly prohibited and shall be punished according to the applicable legislation in force.

Responsibility

The entirety of copyrights and other intellectual property rights associated with the texts, graphics, questionnaires, images, and all other content provided, regardless of their visual representation, either belong to ULTIMO or are utilized by ULTIMO with proper authorization, whether as the rightful owner or through a third party possessing the appropriate rights.

The utilization or dissemination, as well as the accessibility of trademarks, logos, or any other contained content, does not confer, nor should it be construed as conferring, any intellectual property rights to the public that may be registered, licensed, authorized, or possess similar rights.

Any attempts to modify, replicate, transmit, duplicate, trade, or otherwise misrepresent the contents protected by this provision within the system, as well as any actions that may result in harm to ULTIMO or compromise its integrity, are strictly forbidden and will be subject to legal penalties according to the prevailing legislation.

Change of Terms and Conditions of Use

ULTIMO retains the authority to amend, update, or entirely remove these Terms and Conditions of Use at its discretion, and such modifications will be properly announced on this page or in another designated section. These changes will also be made accessible to our partners and the public for their validation.

It is incumbent upon both the public and our partners to routinely check this page containing the Terms and Conditions of Use to stay informed about the contents herein and to remain apprised of any supplementary information.

Validity of Terms and Conditions of Use

If any provision of these Terms and Conditions of Use is inconsistent with applicable law, that provision shall be deemed to be unwritten, not affecting the validity of the remaining provisions of these Terms and Conditions of Use.

Jurisdiction and applicable law

These Terms and Conditions of Use are subject to Belgian Law. Any disputes, legal matters, or issues stemming from the use of this system will be exclusively resolved within the ULTIMO consortium's jurisdiction, with the clear waiver of any alternative venues.

Clarifications

If you have any questions about these Terms and Conditions of Use, you should send a request for clarification to ULTIMO through the ULTIMO helpdesk service with the following email address ultimodmp@gmail.com

Data Management and Responsibility

This Privacy Policy applies to all personal data and data from pilot sites contained in the Platform managed by ULTIMO.

The processing of personal data and data from pilot sites by ULTIMO is governed by the following principles in accordance with the GDPR: a) legality, fairness and transparency in the processing of data; b) limitation of purpose; c) minimization of data; d) accuracy; e) conservation limitation; f) integrity and confidentiality.

1. **Lawfulness, fairness and transparency**: ULTIMO ensures that personal data and data from pilot sites are collected and processed lawfully, fairly and in a transparent manner in relation to the data subject.

2. **Purpose limitation**: ULTIMO ensures that personal data and data from pilot sites are collected only for specified, explicit and legitimate purposes.

3. **Data minimization**: ULTIMO takes relevant technical and organizational measures so that personal data and data from pilot sites will be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

4. **Accuracy**: ULTIMO shall take all necessary steps to ensure that the personal data and data from pilot sites it collects and processes are always accurate and, where necessary, kept up-to-date.

5. **Storage limitation**: ULTIMO does not store the personal data and data from pilot sites it collects for longer than is necessary for the purposes for which the personal data and data from pilot sites are processed. However, it is possible to keep them for a longer period, when processing is necessary:

i) to comply with a legal obligation on the processing;

ii) to fulfil a duty carried out for the public interest;

iii) for achieving purposes in the public interest;

iv) for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to implementation of the appropriate technical and organizational measures, including pseudonymization, and only where these purposes cannot be achieved through data anonymization;

v) for the establishment, exercise or defence of legal claims.

6. **Integrity and confidentiality**: ULTIMO ensures that personal data and data from pilot sites are processed in a manner that ensures appropriate security of the personal data and data from pilot sites, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures.

Purposes of treatment

The data processed within the framework of ULTIMO's usage is gathered during the User's system registration and each subsequent login. ULTIMO will handle User and/or data subject data, both manually and through automated means, for the following specific purposes:

- Registration on the ULTIMO application and services.
- Enhancing access to User information regarding reported mobility conditions.
- Improving services within the autonomous mobility sector.
- Ensuring the continuity of provided Public Transport services.
- Advancing scientific knowledge in the area by comparing and designing common response patterns based on anonymized data, when appropriate. The data of the holders will not be treated for any other purpose that does not observe the purposes thus described.

Within the scope of this Platform, ULTIMO manages data related to User identification, public transport behaviour, data from social media, and, in certain cases, health data. ULTIMO acknowledges the sensitive nature of health data and implements appropriate technical and organizational measures to ensure its protection.

The processing of personal data contained in the Platform is contingent upon the prior consent of the data subject, and this consent can be revoked at any time by contacting ultimodmp@gmail.com. Withdrawal of consent does not affect the legality of processing based on the consent granted prior to revocation.

Rights of data subjects

As the owners of personal data and data from pilot sites, the User and pilot sites may, at any time and for free, exercise their rights of access, rectification or erasure, limitation, portability and opposition of their data, through the email address ultimodmp@gmail.com

If you exercise any of these rights, ULTIMO will analyse it and respond within a maximum of one month. The time limit may be extended by up to two months, subject to Article 12 (3) of the GDPR where necessary, taking into account the complexity of the application and the number of applications, in which case the data subject to be informed of the extension and the reasons for the delay within one month from the date of receipt of the request. If the request submitted by the data subject is not followed up, the data subject must be informed, without delay and within one month of the date of receipt of the request, of the reasons for the delay and the possibility of submitting complaints to a supervisory authority and bring legal action.

If the data subject submits the request by electronic means, and unless otherwise requested by the data subject, the information must be provided in an electronic format in common use.

However, you are hereby informed that if you believe that ULTIMO has violated or may have violated your data protection rights, you may submit a complaint to the National Data Protection Commission.

The User may also contact the Data Protection Officer ("Data Protection Officer" or "DPO") of ULTIMO on all matters related to the processing of their personal data and the exercise of their rights, through the email address ultimodmp@gmail.com or upon written request addressed to ULTIMO - under the care of WP3's forum.

With regard to the right to erase the data and in view of the identified purpose of clinical research, under Article 89 of the GDPR, the GDPR may be restricted, in accordance with Article 17 (3) (d) of the GDPR, only to the extent that such erasure of the data is likely to make it impossible or seriously detrimental to the attainment of the objectives of that treatment.

Shelf life

ULTIMO will retain personal data and data from pilot sites solely for the duration necessary to fulfill the specified purposes. This includes maintaining data as long as a user's login remains active and the pilot sites have granted consent, while the original purposes for data collection persist. In exceptional cases, data may be retained for extended periods, but exclusively for scientific research or statistical purposes. This extension is subject to the implementation of suitable technical and organizational measures aimed at safeguarding the data subject's rights and freedoms. Such measures may involve pseudonymization, provided that the intended objectives can still be achieved through this method. If these purposes are potentially impacted by new processing activities that no longer allow for the identification of data subjects, the same objectives can still be fulfilled.

Communication of data

ULTIMO may communicate personal data and data from pilot sites with respect to the rules established in the GDPR provided that:

- has obtained, unequivocally, the consent of the User for this purpose;
- the transmission is made in the context of the fulfilment of a legal obligation, a determination by the National Data Protection Commission or a court order; or
- the communication is intended to protect the vital interests of Users or any other legitimate purpose provided by law.

When ULTIMO communicates the data to third parties, the User and pilot sites will be informed and the identity of the recipients and the purpose of the treatment for which the data have been transferred will be transmitted to him.

Subcontractors

Users' personal data may be processed using appropriate service providers, carefully selected by ULTIMO in accordance with the GDPR, to adopt sufficient guarantees for the implementation of appropriate technical and organizational measures for the processing of data that meets the requirements of GDPR and ensure the protection of the rights of data subjects. Service providers will only treat personal data for the purposes defined by ULTIMO and in compliance with the instructions issued by ULTIMO, ensuring and observing the standards set forth in the GDPR and other applicable data protection laws.

Pilot sites' data may be processed using appropriate service providers, carefully selected by ULTIMO in accordance with the GDPR, to adopt sufficient guarantees for the implementation of appropriate technical and organizational measures for the processing of data that meets the requirements of GDPR and ensure the protection of the rights of data subjects. Service providers will only treat data from pilot sites for the purposes defined by ULTIMO and in compliance with the instructions issued by ULTIMO, ensuring and observing the standards set forth in the GDPR and other applicable data protection laws.

The identification of entities subcontracted by ULTIMO is made in the access to the Platform.

Security measures

In its commitment to protecting the data security, ULTIMO takes the most appropriate measures to protect the personal data and data from pilot sites contained in the system against its dissemination, loss, destruction, misuse, alteration, treatment or unauthorized access and / or illicit:

Confirm every point

- Access is restricted to personal data and data from pilot sites based on the "need to know" principle, strictly within the defined purposes.
- Utilization of Secure Sockets Layer (SSL) certificates to guarantee the security and confidentiality of data transmission between the system, users, operators, and pilot sites.
- Authentication mechanisms are in place for system entry and for specific actions such as loading location and health information.
- To ensure secure identification, a validation email and/or code is sent directly to the user's designated contact.
- Adherence to practices outlined in the ISO 27001 standard, which serves as the international benchmark for Information Security Management and stands as the premier structure for safeguarding information.
- Implementation of encryption practices and pseudonymization of data, incorporating enhanced security measures for location, health data, and information from social media, among others.
- Complete data access is limited to ULTIMO operators only, in accordance with their confidentiality obligations.

- Health data is logically and physically segregated from other data, with distinct access profiles due to its sensitive nature.
- ULTIMO professionals are bound by confidentiality obligations, which persist even after their engagement with ULTIMO concludes.
- Adoption of a code of conduct, internal regulations, or binding company policies pertaining to the safeguarding of personal data and data from pilot sites.
- Anonymization of data, where applicable, ensuring it aligns with lawful, legitimate, and clearly defined purposes.

However, it is the Users' responsibility to ensure that the personal equipment through which they access the ULTIMO is adequately protected against harmful software and computer viruses or other forms of improper access by third parties. The same condition holds for pilot sites as well.

Appendix II

ULTIMO Data List

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Static	**Name**	String	Name of the Vehicle	All Services	NeTEx	No
Static	**Manufacturer**	String	Name of OEM	All Services	NeTEx	No
Static	**Vehicle Type**	String	Classification according mode and capacity	All Services	NeTEx	No
Static	**Model**	String	Name of the model	All Services	NeTEx	No
Static	**Seating capacity**	Integer	Maximum number of seats available for passengers	Passenger Services, Operator Services	NeTEx	No
Static	**Standing Capacity**	Integer	Maximum number of standing spaces available for passengers	Passenger Services, Operator Services	NeTEx	No
Static	**Energy Type**	String	Gasoline or Electric	Predictive maintenance, optimizing battery health	NeTEx	No
Static	**Special needs capacity**	Integer	number of places for people with special needs	Passenger Services, Operator Services	NeTEx	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Static	**Wheelchair capacity**	Integer	number of places for wheelchairs	Passenger Services, Operator Services	NeTEx	No
Dynamic	**Connection Status**	Boolean	Online or Offline	All Services	SIRI	No
Dynamic	**Location**	Float	GPS location	All Services	SIRI	No
Dynamic	**Door Status**	Boolean	Open or Close	All Services	SIRI	No
Dynamic	**Energy level**	Float	Current energy level of the vehicle	Predictive maintenance, optimizing battery health	SIRI	No
Dynamic	**Odometer**	Integer	Current odometer reading of the vehicle	All Services	SIRI	No
Dynamic	**State of Charge**	Float	Rate %	All Services, especially for predictive maintenance and optimizing battery health	SIRI	No
Dynamic	**Speed**	Float	Current speed of the vehicle	All services	SIRI	No
Dynamic	**Occupancy**	Integer	Current occupancy of the vehicle	Passenger Services, Operator Services	SIRI	No
Dynamic	**Wheelchair on board**	Integer	Current number of wheelchairs in the vehicle	Passenger Services, Operator Services	SIRI	Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Dynamic	**Passengers with special needs**	Integer	Current number of passengers with special needs in the vehicle	Passenger Services, Operator Services	SIRI	Yes
Dynamic	**Orientation**	float	Direction of the vehicle	All Services	SIRI	No
Dynamic	**Heading**	Float	Angle between the direction in which the vehicle's front is pointing and the true north	All Services	SIRI	No
Dynamic	**Acceleration**	Float	Current acceleration of the vehicle	All Services	SIRI	No
Dynamic	**Navigation Mode**	String	Current navigation mode of the vehicle (autonomous, manual)	All Services	SIRI	No
Dynamic	**GNSS connection**	Boolean	Connected or Not Connected	All Services	SIRI	No
Dynamic	**Communication protocol**	String	The protocol for wireless communication	All Services	NeTEx	No
Dynamic	**Signal strength**	String	Level of the signal	All Services	SIRI	No
Dynamic	**Bandwidth**	Float	Description of network	All Services	SIRI	No
Dynamic	**Latency**	Float	Description of network	All Services	SIRI	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Dynamic	**CO_2 emissions**	Float	Current emissions of CO_2	Mitigation of disease spread, ensure public health, predictive maintenance	SIRI	No
Dynamic	**Energy consumption**	Float	Current Energy consumption	All Services, especially for predictive maintenance and optimizing battery health	SIRI	No
Dynamic	**Travelled kilometres**	Float	Kilometres from the beginning of ULTIMO	Passenger Services, Operator Services	SIRI	No
Dynamic	**Event**	Boolean	If an event exists is true.	Operator Services, Intervention Services	SIRI	No
Dynamic	**Type of event**	String	Emergency or incident	Operator Services, Intervention Services	SIRI	No
Dynamic	**Located event**	Time/location	Time and location of an existing event	Operator Services, Intervention Services	SIRI	No
Dynamic	**Situation**	String	Description about traffic situation	Operator Services, Intervention Services	SIRI	No
Dynamic	**Situation cause**	String	Description about traffic situation	Operator Services, Intervention Services	SIRI	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Dynamic	**Situation Reason**	String	Description about traffic situation	Operator Services, Intervention Services	SIRI	No
Dynamic	**Incident**	String	An unexpected event	Operator Services, Intervention Services	SIRI	No
Dynamic	**Alarm**	Boolean	A dysfunctionality of the system	Operator Services, Intervention Services	SIRI	No
Dynamic	**Emergency notification time**	Date/Time	The time of an emergency	Operator Services, Intervention Services	SIRI	No
Dynamic	**Emergency notification location**	float	The location of an emergency	Operator Services, Intervention Services	SIRI	No
Dynamic	**Vehicle is driving in reverse**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**Vehicle is braking**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**Break light**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**Strong braking**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Dynamic	**Severe braking**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**Shuttle switched to manual mode**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**DUI: klaxon triggered**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Dynamic	**DUI: buzzer triggered**	Boolean	A variable for emergency detection.	Operator Services, Intervention Services	SIRI	No
Mobility	**Stop places**	XML	A location where vehicles stop	All Services	NeTEx	No
Mobility	**Routes**	XML	A list of located points	All Services	NeTEx	No
Mobility	**Lines**	XML	A group of Routes	All Services	NeTEx	No
Mobility	**Service area**	GeoJSON	The geographic area which is covered by the service.	All Services		No
Mobility	**Passing time**	Time	Arrival time, departure time, waiting time.	Passenger Services, Operator Services	NeTEx	No
Mobility	**Delay**	Time	Deviation from the Timetable	Passenger Services, Operator Services	NeTEx	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Mobility	**Timetable planned**	XML	Planned timetable	All Services	SIRI	No
Mobility	**Timetable actual**	XML	Estimated timetable	All Services	SIRI	No
Mobility	**Operating Day**	String	Weekend/weekdays and so on	All Services	NeTEx	No
Mobility	**Day Type**	String	Properties of a day	All Services	NeTEx	No
Booking	**Load**	Integer	Number of travellers in service	Passenger Services, Operator Services	SIRI	No
Booking	**Vehicle availability**	Boolean	The vehicle could cover a need or not	Passenger Services, Operator Services	SIRI	No
Booking	**Desired pickup location**	Float	Initial location of a passenger	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Desired pickup time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Desired drop off location**	Float	Destination of a passenger	Passenger Services, Operator Services, Intervention Services	SIRI	Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Booking	**Desired drop off time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Planned pickup location**	Float	The location in which the passenger could board calculated by the system	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Planned pickup time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Planned drop off location**	Float	The location in which the passenger could debark calculated by the system	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Planned drop off time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual pickup location**	Float	The location in which the passenger actually boards.	Passenger Services, Operator Services, Intervention Services	SIRI	Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Booking	**Actual pickup time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual drop off location**	Float	The location in which the passenger actually debarks.	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual drop off time**	Time	Correspond time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Planned booking route**	GeoJSON	Planned vehicle route between pickup and drop off location.	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual booking route**	GeoJSON	Actual vehicle route between pickup and drop off location.	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Direct ride distance**	Float	Length of the fastest direct route between pickup and drop off location	Passenger Services, Operator Services, Intervention Services	SIRI	Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Booking	**Direct ride duration**	Float	Duration of the fastest direct route between pickup and drop off location.	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual ride distance**	Float	Length of the actual route between the actual pickup location and the actual drop off location	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Actual ride duration**	Float	Duration between the actual pickup time and the route actual drop off time	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
Booking	**Trip reason**	String	Work/shopping/school and so on	Passenger Services, Operator Services, Intervention Services	SIRI	Yes
External	**Temperature**	Float	Current temperature	All Services	SIRI	No
External	**Feels like**	Float	Current feels like temperature	All Services	SIRI	No
External	**Min Temperature**	Float	Prediction of minimum temperature	All Services	SIRI	No
External	**Max Temperature**	Float	Prediction of maximum temperature	All Services	SIRI	No

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
External	**Pressure**	Float	Current pressure	All Services	SIRI	No
External	**Humidity**	Float	Current humidity (%)	All Services	SIRI	No
External	**Wind deg**	Float	Wind degrees (azimuth degrees)	All Services	SIRI	No
External	**Wind speed**	Float	Current wind speed	All Services	SIRI	No
External	**Weather main**	String	Weather condition	All Services	SIRI	No
External	**Weather description**	String	Weather description	All Services	SIRI	No
External	**City traffic**		Current city traffic situation	All Services	SIRI	No
External	**Maps**		3D maps	All Services		No
External	**Noise levels**	Float	Noise level of the city	All Services	SIRI	No
Infrastructure	**Internal temperature**	Float	Current internal temperature of the vehicle	Passenger Services, Operator Services, Intervention Services	SIRI	No
Infrastructure	**Video-internal cameras**		Cameras installed in the vehicle	Passenger Services, Operator Services, Intervention Services		Yes
Infrastructure	**Video-external cameras**		Cameras installed out of the vehicle	Passenger Services, Operator Services, Intervention Services, V2X		Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Infrastructure	**Magnetic loops**		It calculates the inferences regarding the traffic condition	Passenger Services, Operator Services, Intervention Services, V2X		Yes
Infrastructure	**Lidar Sensor**		Cameras installed in points of congestion for measuring total traffic flow	Passenger Services, Operator Services, Intervention Services, V2X		No
Infrastructure	**Radar Sensor**		Radar Sensor is installed on the vehicle	Passenger Services, Operator Services, Intervention Services, V2X		No
Infrastructure	**Sensors for capturing wireless internet traffic**		Metadata showing the traffic between the vehicle's sensors and the cloud servers	All Services		Yes
Infrastructure	**Camera installed on traffic lights or bridge**		External camera on the vehicle for counting outside vehicles	Passenger Services, Operator Services, Intervention Services, V2X		Yes

Data Type	Attribute Name	Attribute Value Type	Attribute Description	Why/where need	Standard	Personal Data
Infrastructure	**Traffic in Vehicle's route**		Camera in the vehicle for traffic situation	Passenger Services, Operator Services, Intervention Services, V2X		Yes
Infrastructure	**Radio frequency sensor**		Bluetooth detections	All Services		Yes
Infrastructure	**Vehicle traffic camera**		Camera installed on the vehicle	Passenger Services, Operator Services, Intervention Services, V2X		Yes
Network	**Bluetooth Sensor**		Probe data, Social media data	Passenger Services, Operator Services, Intervention Services, V2X	SIRI	Yes
Network	**Network traffic metadata**			All Services		Yes
WiFi	**WiFi connection**			All Services		Yes
5G	**5G cellular network**			All Services		Yes

Appendix III

Questionnaire for ULTIMO operation sites: ULTIMO Data Management Platform

Introduction

This questionnaire is currently being circulated among the operational sites of ULTIMO. Its main purpose is to establish the data collection mechanisms employed by ULTIMO's sites and outline the primary workflow of the ULTIMO Data Collection System (DCS). The questionnaire is comprised of 3 parts:

- Main properties of the site
- Data Collection System of the site
- Data Provision in terms of ULTIMO

Part 1: Main properties of the site

1. How many vehicles are included in your fleet?
2. What is the original equipment manufacturer (OEM) of your vehicles?
3. What is the geographical operational area covered by your vehicles?
4. How many hours, approximately, does each vehicle spend on the road per day?

Part 2: Data Collection System of the site

5. What are the sources of the data, including vehicles, infrastructure, and external sources?
6. What are the specific data attributes captured by each of these sources?
7. How frequently is the data captured or updated?
8. What are the main technologies that are utilized in your site Data Collection System?

Part 3: Data Provision in terms of ULTIMO

9. Are you willing to provide historical (csv), real time-data (automation mechanisms like APIs -WEBSOCKET - MQTT - KAFKA) or both historical and real-time?
10. Do you agree on establishing a common data structure among all partners for ULTIMO?
11. Would it be realistic according to your requirements to change your implementation to fit the ULTIMO requirements?
12. Are you willing to use technologies and tools that are extra from the default strategy of your institution?
13. What is your preference regarding the provision of real-time data? Would you prefer a direct connection of ULTIMO's team to your existing system, or would you prefer the data to be provided through a separate broker?