MobiSpaces: An Architecture for Energy-Efficient Data Spaces for Mobility Data

Conference Paper · December 2023

DOI: 10.1109/BigData59044.2023.10386539

28 authors, including:

Georgios Santipantakis
University of Piraeus
35 PUBLICATIONS **282** CITATIONS
SEE PROFILE

Nikolaos Koutroumanis
University of Piraeus
11 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Yannis Theodoridis
University of Piraeus
383 PUBLICATIONS **11,431** CITATIONS
SEE PROFILE

Dimosthenis Kyriazis
University of Piraeus
342 PUBLICATIONS **4,008** CITATIONS
SEE PROFILE

MobiSpaces: An Architecture for Energy-Efficient Data Spaces for Mobility Data

Christos Doulkeridis[1], Georgios M. Santipantakis[1], Nikolaos Koutroumanis[1], George Makridis[1], Vasilis Koukos[1], George S. Theodoropoulos[1], Yannis Theodoridis[1], Dimosthenis Kyriazis[1], Pavlos Kranas[2], Diego Burgos[2], Ricardo Jimenez-Peris[2], Mariana M G Duarte[3], Mahmoud Sakr[3], Esteban Zimányi[3], Anita Graser[4], Clemens Heistracher[4], Kristian Torp[5], Ioannis Chrysakis[6,7,8], Theofanis Orphanoudakis[6], Evgenia Kapassa[9], Marios Touloupou[9], Jürgen Neises[10], Petros Petrou[11], Sophia Karagiorgou[11], Rosario Catelli[12], Domenico Messina[13], Marcelo Corrales Compagnucci[14], Matteo Falsetta[15]

[1]*University of Piraeus*, Piraeus, Greece, [2]*LeanXcale*, Madrid, Spain

[3]*Université Libre de Bruxelles,*, Brussels, Belgium, [4]*Austrian Institute of Technology*, Vienna, Austria

[5]*Aalborg University*, Denmark, [6]*Netcompany-Intrasoft*, Luxembourg [7]*Ghent University*, Belgium

[8]*KU Leuven University*, Belgium, [9]*University of Nicosia*, Cyprus, [10]*Fujitsu*, Germany, [11]*UBITECH*, Cyprus

[12]*Engineering Ingegneria Informatica S.p.A.*, Italy [14]*White Label Consultancy*, Copenhagen, Denmark [15]*GFT*, Italy

[1]{cdoulk,gsant,koutroumanis,gmakridis,vkoukos,gstheo,ytheod,dimos}@unipi.gr,

[2]{pavlos,diego.burgos,rjimenez}@leanxcale.com, [3]{mariana.machado.garcez.duarte,mahmoud.sakr,esteban.zimanyi}@ulb.be,

[4]{Anita.Graser,Clemens.Heistracher}@ait.ac.at, [5]torp@cs.aau.dk,

[6]{Ioannis.Chrysakis,Theofanis.Orphanoudakis}@netcompany.com,

[9]{kapassa.e,touloupos.m}@unic.ac.cy, [10]juergen.neises@fujitsu.com, [11]{ppetrou,skaragiorgou}@ubitech.eu,

[12]Rosario.Catelli@eng.it, [13]domenico.messina@linux.com, [14]mc@whitelabelconsultancy.com, [15]Matteo.Falsetta@gft.com

***Abstract*—In this paper, we present an architecture for mobility data spaces enabling trustworthy and reliable data operations along with its main constituent parts. The architecture makes use of a data lake for scalable storage of diverse mobility data sets, on top of which separate computing and storage layers are implemented to allow independent scaling with a data operations toolbox providing all data operations. Furthermore, to cater for mobility analytics, machine learning and artificial intelligence support, an edge analytics suite is provided that encompasses distributed algorithms for mobility analytics and federated learning, thereby exploiting edge computing technologies. In turn, this is supported by a resource allocator that monitors the energy consumption of data-intensive operations and provides this information to the platform for intelligent task placement in edge devices, aiming at energy-efficient operations. As a result, an end-to-end platform is proposed that combines data services and infrastructure services towards supporting mobility application domains, such as urban and maritime.**

***Index Terms*—data spaces, mobility data, data governance, edge analytics**

I. Introduction

Data spaces [4], [17] comprise trusted platforms for managing the complete lifecycle of data, covering various data models, metadata descriptors, ontologies for semantic meaning, as well as data services for accessing, processing and analysing data. Franklin et al. [7] originally introduced the concept of data spaces aiming to cover all data sources within an organization, irrespective of data model, data format or data location.

Recently, the concept of data spaces has been revived due to the observed situation worldwide that clearly indicates that companies and organizations mostly operate as "data silos" or "data islands", without well-established procedures to facilitate trusted and standardized data sharing, exchange and interoperability, leading to a waste of resources due to unnecessary and repetitive data-related operations.

This is intensified in the domain of mobility and transportation [26], which constitutes one of the main pillars of the global economy, because mobility-related data: (a) is (by nature) produced and collected in a completely decentralized manner, (b) contains sensitive information about the individual person or object tracked, (c) can be analysed to discover hidden mobility patterns and extract invaluable insights and knowledge, and (d) constitutes a complex data type, the management and exploration of which requires advanced processing algorithms.

Several challenges [18] need to be addressed at technical level in order to harness the merits of mobility data and mobility analytics [6], [20]. First, a distributed and decentralized platform is required that is able to cope with big data for scalable and low-latency processing. In addition, data governance services are required to ensure trustworthiness and reliable data manipulation, while respecting privacy which is a key concern for mobility data. Second, specialized data operations must be supported, in particular providing mobility-aware operations that can handle spatio-temporal data and trajectories of moving objects at scale. Third, data analytics and machine learning methods tailored for this domain are essential, especially in a federated and decentralized context, thus pushing data analysis near to the sources where data acquisition takes place. Last but not least, infrastructure support in terms of deployment and orchestration needs to be

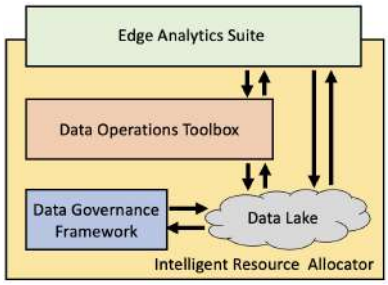

Fig. 1. A bird's-eye view of the MOBISPACES architecture.

mobility-aware and mobility-optimized, to enable intelligent placement of services on decentralized resources, leading to energy-efficiency.

To address these challenges, in this paper, we present MOBISPACES, an architecture for energy-efficient data spaces for mobility data (see Fig. 1 for a high-level view), following a decentralized approach that exploits Edge computing technologies. The architecture comprises both *data services* as well as *infrastructure services*, following the GAIA-X paradigm that provides both a data and an infrastructure ecosystem [17]. The data services affect the entire lifecycle of mobility data (the "data path") and include data governance, data management and data analysis services. Moreover, we exploit the data locality of mobility data in order to perform decentralized operations at the Edge, ranging from local, privacy-preserving aggregations, to federated learning. As a result, the architecture enhances privacy, drastically reduces communication overheads, and exploits processing at the Edge, thus also reducing the energy consumption of data-intensive operations that typically run in the Cloud. Another innovative aspect of MOBISPACES is that it is designed to support resource allocation at Edge devices, by intelligent task placement according to an optimization objective that (additionally) takes into account energy consumption, besides execution time.

The rest of this paper is structured as follows: Sect. II presents an overview of the MOBISPACES architecture. Then, its main components are presented: the Data Governance Framework (Sect. III, the Data Operations Toolbox (Sect. IV), the Edge Analytics Suite (Sect. V), and the Intelligent Resource Allocator (Sect. VI). Sect. VII provides evaluation results of MOBISPACES. Finally, Sect. VIII summarizes the key points of the paper.

II. OVERVIEW OF THE MOBISPACES ARCHITECTURE

Data spaces [4], [17] promise to deliver a principled approach to sovereign and trustworthy data sharing, by establishing a trusted and secure environment for data exchange. This is going to impede the current situation of organizations and companies operating as data silos, with limited or ad-hoc data communication mechanisms. Consequently, the proliferation of data spaces technologies may have a huge impact in the data economy. Also, mobility is among the top priorities of the EU for data spaces. For instance, the Mobility Data Space (https://mobility-dataspace.eu/) brings together data providers who wish to monetize their data and data consumers who need access to mobility data sets for developing innovative data services.

The MOBISPACES architecture is divided into two main parts: (a) the centralized or cloud-based part of the architecture (the Cloud), and (b) the part that is distributed to the edge of the overall solution (the Edge). As a result, some of its internal building blocks have been designed to be deployed either in a centralized infrastructure (i.e., in a private cloud) or at the edge, or at both sides.

Data ingestion. The first design decision regarding the MOBISPACES platform concerns data ingestion from diverse data sources of mobility data. These data sources may provide input data both as a stream in real-time as well as batches of historical data. To provide a uniform data ingestion mechanism with fault-tolerance and scalability, we rely on a message queue which acts as main data entry in the platform. We select Apache Kafka for this purpose, due to its salient features that combine scalability, high throughput, durability, real-time processing, support for flexible schemas, and easy integration, making it an ideal platform for data ingestion services.

Scalable storage. At a technical level, the data lake provides scalable storage of data in raw, or minimally-processed form, in a distributed, cloud-based environment. The most common deployment for data lakes relies on HDFS, offering fast data loading in a schema-free way, highly-efficient streaming data access patterns as well as fault-tolerance. Thus, high throughput data accesses on large mobility data sets can be supported via HDFS. Even though other storage solutions providing low latency do exist (e.g., NoSQL, NewSQL), MOBISPACES includes a variety of database engines (as will become clear in Sect. IV) and mainly uses the data lake as a repository for these engines to ingest data from.

III. THE DATA GOVERNANCE FRAMEWORK

Data governance operations are essential for modern data marketplaces, in order to promote trusted data exchange and ensure reliability. This is also the case for mobility data which are prevalent in everyday operations, from route discovery to fleet monitoring. MOBISPACES provides *a set of data governance services* to address this challenge. Fig. 2 illustrates the internal architecture of the Data Governance Framework, while its functionality is described below.

Data sharing and exchange. A very important goal of MOBISPACES concerns standardized data sharing and exchange. To ensure compatibility with initiatives, such as GAIA-X (https://gaia-x.eu/) and IDS (https://internationaldataspaces.org/), MOBISPACES is designed to use *data connectors* for data exchange. More specifically, we have selected IDS connectors that provide a virtual container that maintains control over the data: who can access the data, for how long, which data, etc. In this way, interoperability with other data spaces in the mobility or in other relevant domains is supported. The ultimate aim of MOBISPACES data space is to provide a means of collaboration upon mobility data among different stakeholders [17]. We also plan to make some of the offered

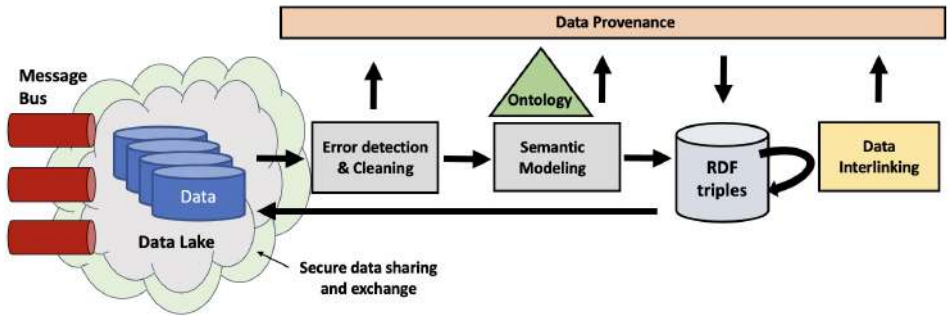

Fig. 2. The Data Governance Framework.

services available for external users to unleash the potential of the MOBISPACES data space through data sharing and data integration and to leave room for collaboration with other relevant projects or initiatives. Interestingly, collaborative data sharing [8] is promoted by EU through the establishment of several legal frameworks (e.g., European Strategy for Data, Data Act, and Data Governance Act). Data integration is also important because no central schema is required for data sharing and exchange and there are no restrictions in terms of data formats as any kind of data can be ingested to the data lake and thus can be shared via our data space. Thus, MOBISPACES data space offers flexibility and less complexity to organizations and users who want to share and exchange their data to finally create value propositions.

Error detection and cleaning. GPS data typically contains noise due to limitations of mobile sensor technology. For example, the accuracy of GPS devices is known to be a couple of meters, while various obstacles such as buildings or bridges may have a negative effect on this. In the sea, similar issues exist as large vessels are required to use the Automatic Identification System (AIS) to report their position, and AIS data is noisy. Oftentimes, errors are introduced during the data acquisition process (e.g., longitude and latitude columns are reversed), data transformation (e.g., date-time conversion), or data filtering (e.g., spatial area clipping may remove GPS positions and lead to infeasible trajectories). To cope with such cases, MOBISPACES includes an error detection and cleansing mechanism that goes beyond rule-based error detection.

Semantic modeling. Apart from using formal data models for data representation, semantic models are required in order to achieve semantic interoperability. To this end, ontologies and semantic web technologies (RDF, OWL) applied on mobility data are the facilitators. In MOBISPACES, we build upon previous work that has created conceptualizations for the aviation and maritime domain [27], which is extended also for the urban mobility domain. In this way, we provide a semantic model for diverse data sources: positioning data (GPS/Radar data), weather forecasts, contextual information, geographical data, as well as complex events. However, additional ontologies and vocabularies can be used to describe new data sources.

Having ontologies for the conceptualization of the domain is only the first step. Subsequently, a data transformation mechanism is required to generate a semantic data representation from raw data. We opt for RDF, the de-facto standard in the World Wide Web, which can be used to associate mobility data with external data sources (e.g., wikidata, etc.). RDF-Gen [21] is a state-of-the-art tool for RDF data generation with several salient features, including support for diverse data sources, both streaming and archival, various input formats, flexibility, near to the sources processing, as well as scalability. By applying RDF-Gen to the output records of the Error detection and cleaning, we obtain a common semantic representation, where cleansed data is stored as RDF triples in an RDF store.

Data interlinking. Another challenge for effective data governance concerns data interlinking. In MOBISPACES, we focus on spatial and spatio-temporal link discovery, aiming to associate (link) entities based on various spatial or/and temporal relations. Indicative examples include linking trajectories with protected regions or nearby ports. After link discovery, these associations are stored back to the RDF store as new RDF triples that explicitly record the discovered relations. In essence, these materialized join results are exploited at query time for improved performance. To further enable easy data access, we support a mechanism that extracts integrated data from the RDF store and puts them back in the data lake in tabular format. In this way, upstream query engines can avoid the hurdle of data ingestion from RDF stores. Essentially, RDF acts as an integration mechanism that aligns and enriches data sets under common data schema.

Data provenance. MOBISPACES offers a service that tracks the provenance of data within the platform. Typically, a data set that is used for successfully training an ML model has undergone a series of data manipulation operations. Examples include data imputation, exclusion of outliers, data normalization, enrichment with other data, as well as different kinds of transformations. The data provenance mechanism aims to track this history and represent it in a formal way, so that it can be queried. For the representation, we adopt PROV-O, a W3C recommendation for the representation and interchange of provenance information generated in different systems and under different contexts. As such, the information collected by the data provenance component is stored in the form of RDF triples and it can be used to query the provenance of a data set together with the actual data.

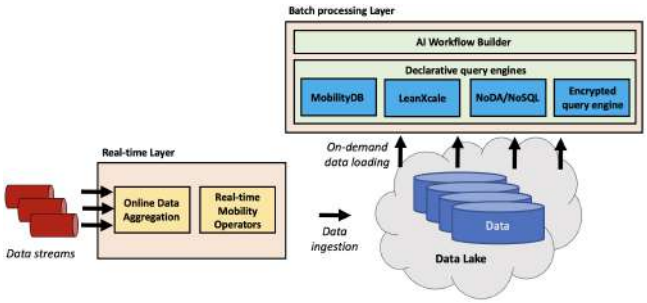

Fig. 3. The Data Operations Toolbox.

IV. Data Operations Toolbox

MobiSpaces offers efficient and scalable tools for querying mobility data, supporting a wide range of diverse scenarios. As depicted in Fig. 3, the data operations toolbox consists of a *batch layer* and a *real-time layer*. Regarding the batch layer, MobiSpaces adopts a declarative querying approach ("SQL on everything") over all data sources, also popularized lately by Presto [22]. Users can express complex data processing tasks either by means of workflows (via a workflow builder) or by using SQL. Interestingly, SQL-based queries are supported for different types of storage of mobility data: raw or minimally-processed files, relational data, non-relational data, mobility data, as well as for encrypted data. As regards the real-time layer, this involves online data aggregation (e.g., for data compression) and a library of efficient mobility operators, which can be deployed at the Edge and support both data ingestion and in-situ querying.

Online data aggregation. Online aggregation refers to the idea of converting aggregation from a batch process to an online process. Further, an end-user can see how the result of the aggregation changes as more data is processed. The classic batch approach only returns the final result to the end-user. Online aggregation is an active area of research, particularly related to distributed architectures and well-suited in a streaming-data environment. In a distributed edge architecture, the online aggregation sends partial results to a coordinator that reports the complete result. The aggregation at the edge enhances privacy and reduces communication costs. These cost reductions are explored in MobiSpaces.

Mobility operators. In edge compute nodes the mobility data represent an individual moving object, i.e., the owner/host of the edge device, and possibly the objects in its close neighbourhood. The real-time requirement is easy to satisfy from the scalability point of view. The challenge remains to develop expressive operators in a streaming mode, and to package them in a library implementation that can run in a variety of edge platforms. MobilityDB [28] implements an extensive API for moving object data management, as an extension to PostgreSQL. Its type system and spatio-temporal processing functions have been isolated and packaged into the MEOS library [25]. In MobiSpaces, preliminary tests have been carried out to use MEOS on edge devices including Rassperry PI 4 and NVIDIA AGX Orin SBC. The ambition is to stabilize MEOS on these platforms, and to extend it with streaming APIs for trajectory simplification, cleaning, and compression.

Declarative querying. To harness the merits of declarative querying, the concept of "SQL on everything" was recently proposed [22]. In this approach, SQL is advocated to query data in a declarative way, irrespective of the storage model. Extending this approach for spatio-temporal data is proposed by NoDA [16], which provides a common SQL interface over different scalable NoSQL stores (MongoDB, HBase and Redis). In MobiSpaces, the ambition is to provide the same functionality over raw or minimally-processed data, such as those stored in data lakes. Towards this goal, we have proposed TrajParquet [15], an extension of Apache Parquet for trajectory data, following the paradigm of [19] for spatial data.

In the case of structured data, MobiSpaces supports LeanXcale [14], a novel relational database that combines the benefits of both SQL and NoSQL worlds. Firstly, it is SQL-compliant including a relational distributed query engine that allows both inter-/intra-query and inter-/intra-operation parallelism. This allows to push down query operations to the storage, to be executed in parallel, which results to efficient query processing in terms of both latency and energy consumption, as only minimum data needs to be transmitted over the network. Secondly, it ensures database transactions with its ultra-scalable transaction engine that allows for linear and limitless scaling. Thirdly, aggregation operations can be executed in real-time, as data is being ingested concurrently at very high rates, ensuring data consistency in terms of database transactions.

To support scenarios that require a richer repertoire of mobility operators, MobilityDB [24], [28] is used as a database server providing a powerful spatio-temporal extension to the PostgreSQL open-source relational database management system. MobilityDB allows for the storage and querying of spatio-temporal data. MobilityDB provides a range of temporal data types and functions, including support for mobility data with different types time domains: discrete instances, continuous time periods, and gaps of definition.

Moreover, in many real-world cases where privacy preservation over edge or centralised data is a major concern, encrypted query processing is required to secure users' sensitive information. MobiSpaces introduces an encrypted query engine which applies Fully Homomorphic Encryption (FHE) and querying processing over edge and centralised data sources of diverse schemas and granularities to efficiently align, aggregate and serve actionable and trainable Artificial Intelligence Operations (AIOps) and Data Operations (DataOps) without sacrificing efficiency and accuracy.

Workflow Builder. The Workflow Builder oversees the delicate task of providing a graphical interface for interaction with the end-user. It allows users to create complex, AI-based batch and ETL workflows through a declarative methodology and a highly customisable, easy-to-interact canvas-and-palette interface. Besides the graphical interface, it contains two other internal components. The *AI Catalogue* consists of a list of

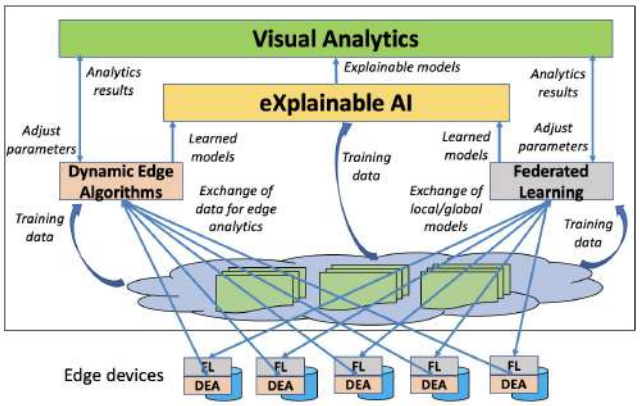

Fig. 4. The Edge Analytics Suite.

statistical and ML/DL tools and algorithms on which the technology underpinning MOBISPACES is based. One of its key features is extensibility: additional AI tools and algorithms can be incorporated, thus facilitating the exchange of results from the Edge Analytics Suite components (Sect. V). The *Workflow Status Manager* is responsible for tracking the execution status of each workflow, which may consist of several processing nodes distributed in different clusters, thus the interaction with the MOBISPACES Cloud/Edge Operations components is fundamental (see Sect. VI).

V. EDGE ANALYTICS SUITE

Decentralized analytics and learning algorithms are at the core of the MOBISPACES platform, due to the nature of mobility data generation and acquisition. Typically, moving objects emit their spatio-temporal positions (using GPS), which are accumulated in a cloud-based infrastructure. Consequently, our vision is to move analytics closer to the location where data is captured, offloading computation to edge devices, while at the same time minimizing the data that is necessary to be collected centrally (thereby also enhancing privacy).

Fig. 4 depicts the internal architecture of the Edge Analytics Suite. It consists of two main decentralized components that provide dynamic edge analytics algorithms and methods for federated learning over trajectory data. In addition, an eXplainable AI component is included, which offers interpretability and explainability for the models learnt by the decentralized components. Last, but not least, a visual analytics component closes the loop and enables human interaction with the results of data analytics, the learned models and their interpretation, by means of interactive visualizations that can assist a domain expert in discovering meaningful patterns of movement.

Dynamic edge algorithms. Dynamic algorithms for the Compute Continuum (CC) can be deployed in more widely distributed networks with many heterogeneous devices, i.e., devices with different CPU architectures and compute capabilities. As such, the CC as a whole may include many different devices, from small and low-power ones like Raspberry PIs to large centralized servers that are able to communicate with the aforementioned Raspberry PIs. Algorithms for the CC can benefit from being executed at the edge, near the source of data generation, resulting in reduced data transfer.

Federated learning. Federated learning (FL) enables learning in settings where centralized learning is not possible due to communication bandwidth limits or organizations that are not willing or able to share training data with each other. These are common issues in the mobility domain where we deal with data from moving objects and where privacy is a an important concern. While existing FL frameworks (such as TensorFlow Federated, PySyft, and Flower) lack dedicated tools for movement data, MOBISPACES provides a framework for FL from movement data with mobility-aware training pipelines (including simulation of FL environments and training appropriate data processing) as well as mobility-aware ML models [11] for local training, model aggregation approaches, and appropriate metrics [12]. In MOBISPACES, vessels act as edge devices that perform model training on local AIS data collected by surrounding vessels.

eXplainable AI. Explainable AI (XAI) research is focused on creating techniques that allow for more transparent and understandable artificial intelligence (AI) models. XAI aims to enable human users to understand, trust, and effectively manage AI systems. In MOBISPACES, the XAI component is focused on exploring and demonstrating the potential of machine learning and artificial intelligence technologies in the Mobility Data Space environment and more specifically for time series and spatio-temporal underlying data. The XAI framework is designed to accommodate pre-built machine learning models and associated data sets. The outputs of the framework would be the predictions or insights generated by the models, and the visualisations of explanations of the models provided by the XAI component of the framework using techniques such as heatmaps, scatter plots, or line charts.

VA for mobility data. The field of mobility visual analytics has made significant progress, thanks to the growing availability of large data sets and mobile and edge devices [2]. Visual analytics tools enable users to interactively explore and understand complex data sets, in different mobility-related fields [5], [6], [9], [13]. In MOBISPACES, a privacy-aware Visual Analytics (VA) component enables users to interact and understand highly distributed mobility data, bridging the gap between users' knowledge and the predictions provided by other components of MOBISPACES platform. To achieve its objectives, the VA component consists of several internal services, including a data connector, a data converter, and a visualization dashboard.

VI. RESOURCE ALLOCATION AND ORCHESTRATION

Infrastructure support in terms of deployment and orchestration is necessary to enable intelligent placement of services on decentralized resources, leading to energy-efficiency. Providing the end-user services declared in the respective AI workflows appropriate to security and privacy constraints as well as mobility-aware and energy efficient asks for intelligent resource allocation and execution deployment mechanisms.

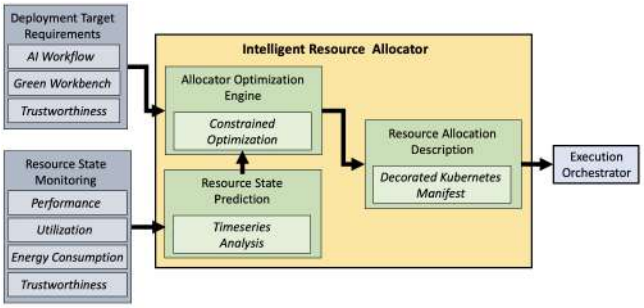

Fig. 5. Intelligent resource allocation and orchestration.

The *Intelligent Resource Allocator*, depicted in Fig. 5, is the core component, which will optimize the allocation based on various constraints provided as input. Resource allocation is modeled as a constrained optimization problem using this input, and assigning resources for the orchestration by the *Execution Orchestrator*. The target of deployment will be based on Kubernetes clusters. The Workflow Builder will provide a workflow description in .yml format. Then, the Intelligent Resource Allocator will produce the optimized target description in the form of a decorated .yml-manifest.

The Workflow Status Manager of the Workflow Builder (Sect. IV) monitors the continuous status update information of the different processing nodes. Based on this, it can return an answer regarding the correct execution of each workflow, i.e., the distribution specified by the Intelligent Resource Allocator and the start-up initiated by the Execution Orchestrator. Thus, the interaction of the Workflow Builder with both these components allows the workload to be meticulously planned and distributed to the right execution environment, considering (i) the availability of both cloud- and edge-side resources, (ii) data access policies, and (iii) their location and availability based on mobility.

Other inputs to the Intelligent Resource Allocator include energy benchmarking and trustworthiness conditions that provide resource needs. Energy benchmarking of the green workbench enables the estimation of energy consumption of a resource. In this way, the platform has available information about the energy consumption for different processing tasks for each type of device. On the other hand, trustworthiness scores and resource state will be monitored. Trustworthiness evaluation is based on a Trustworthiness model first introduced by the Industrial Internet consortium [1] and assessed within the Horizon 2020 project SecureIoT [23]. This way system characteristics like security, privacy, resilience, reliability, safety will be evaluated continuously and shall be used as a trustworthiness profile for resource allocation.

In technical terms, it is foreseen to access Resources states from local *Prometheus* timeseries databases using *Thanos* consolidating access to the distributed resources. The timeseries of those values as well as performance benchmarks shall be used for resource state prediction. Based on the resource requirements and the constraints resulting from the resource state prediction, the Intelligent Resource Allocator performs a constrained optimization resulting in values for the decoration of Kubernetes manifests, which origin from the Workflow Builder. These mainfests shall be used for the execution orchestration.

Data set size	Proc. time (sec)	#Records	#Records/sec
20%	379.61	3,699,135	9,745
40%	755.96	7,398,271	9,787
80%	1,512.78	14,796,543	9,781
100%	1,906.48	18,495,678	9,701

TABLE I
PERFORMANCE OF SEMANTIC REPRESENTATION OF TRAJECTORIES FROM RAW AIS DATA VIA THE DATA GOVERNANCE FRAMEWORK.

VII. EVALUATION SCENARIOS

To demonstrate indicative functionality offered by the MOBISPACES platform, we present three scenarios of processing and analyzing mobility data. These scenarios are manifested as *pipelines* of components, and serve different real-life use-cases that are of particular interest in the maritime domain.

A. Pipeline 1: Semantic representation of trajectories

This pipeline demonstrates how the Data Governance Framework (cf. Sect. III) ingests raw AIS data as a stream, performs basic data cleaning operations, transforms this data in RDF and enriches it with other data, and eventually outputs the semantically enriched data as RDF triples. For the demonstration, we use a Kafka topic to simulate the real stream of AIS data. Also, the MOBISPACES components run on a single virtual machine (VM). We use a publicly available data set (https://zenodo.org/record/2563256) of AIS messages from the area of Brest, and replay the stream (via the Kafka topic) for cleaning and semantic enrichment. To study the efficiency and scalability, we also used smaller portions of the data set, i.e., 20%, 40% and 80%, and repeated each experiment 10 times and report average values.

Table I reports the obtained results: for each portion of the data set (col.1), the average processing time (col.2), the number of records (col.3), and the throughput (col.4). We observe a linear increase in the processing time when we increase the size of the data. This is also reflected in the stable throughput (number of records per second), which is almost 10,000 records per second. Given that these results are obtained using a single VM and the task can be parallelized, since each record can be processed independently of others, this demonstrates scalability with data set size.

B. Pipeline 2: Trajectory compression and analytics

This pipeline aims to assess the impact of trajectory compression on data quality for the analysis, particularly in minimizing data stream size between vessels (Edge) and shore (Cloud). It demonstrates part of the functionality offered by the Data Operations Toolbox (cf. Sect. IV).

Initially, AIS data is gathered at the edge, i.e., the vessel. This includes the AIS data of the vessel itself, and possibly of its neighbourhood. Allowing vessels to repeat the broadcast signals of each other can solve several issues including, for

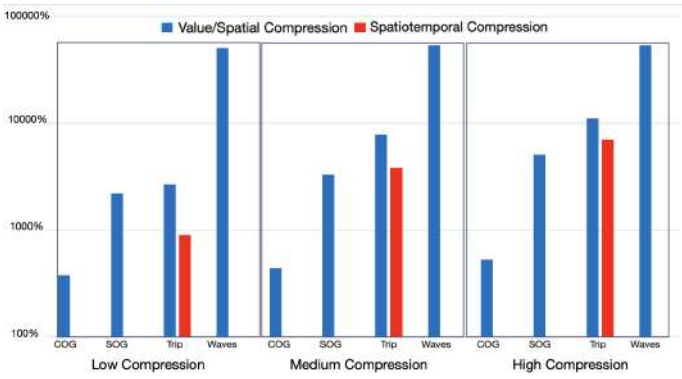

Fig. 6. The compression rate in log scale of (original size / compressed size) of the temporal and spatiotemporal signals in AIS data, using different simplification parameters.

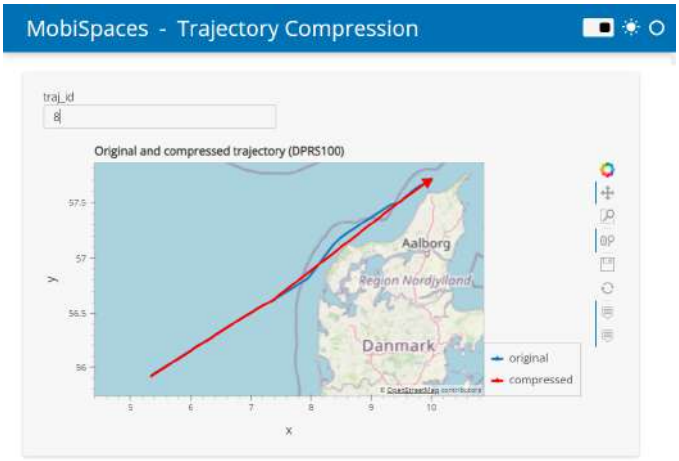

Fig. 7. Pipeline 2 trajectory compression VA app example: spatial trajectory visualization for comparing original and compressed trajectory showing results for DouglasPeucker, Single pass algorithm, Spatial-only distance set to 100m.

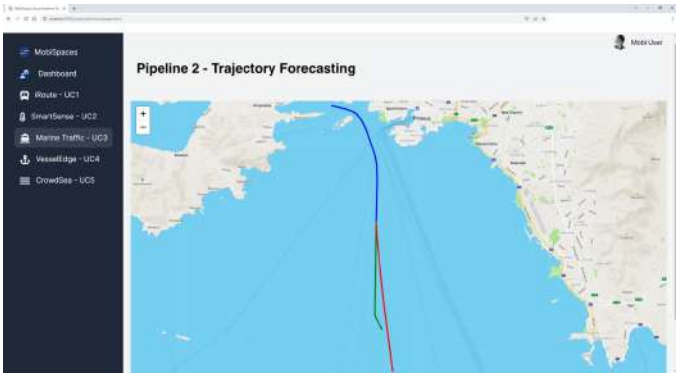

Fig. 8. An illustration of the output of the forecasting component visualised using the Visual Analytics module.

instance, extending the range of connectivity to the shore. The cost however is the increased size of data transmission at individual vessels, which may exceed the bandwidth capacity. Therefore we apply MOBISPACES mobility operators, in Sect. IV, to compresses trajectories at the edge, before they are transmitted to the shore. The MEOS library [25] is used for achieving this compression. We experimented the different variants of compression in MEOS including: *temporal-only* for signals speed over ground (SOG), course over ground (COG), and waves' height, *spatial-only* compression for the vessel trajectory ignoring the time dimension, and *spatio-temporal compression* which preserves speed information in the vessel trajectory. Fig. 6 depicts the achieved compression rate of multiple variants in log-scale. The compression rate reaches 1–2 orders of magnitude, depending on the level of detail in the original signal. In turn, this results in reduced communication cost and energy-efficient use of the available bandwidth.

Subsequently, in the cloud, trajectories are reconstructed by restoring a regular sampling interval enabling assessment of spatio-temporal differences between original and reconstructed trajectories. This enables us to compare the original and reconstructed trajectories so that analysts can tune the compression parameters to find a good fit (balancing the loss of accuracy due to compression and the bandwidth saved) for different applications, such as, for example, visualizations to establish a situation picture in operating rooms or automated forecasting, classification, and anomaly detection. Fig. 7 shows a MOBISPACES visual analytics (VA) app based on the MovingPandas [10] and Panel open source libraries. The compression rate may also be altered temporarily on-demand, for example, during specific events the compression could be reduced to get more accurate data (at the cost of higher bandwidth requirements) than during ordinary situations where higher compression can be applied.

C. Pipeline 3: Trajectory forecasting

This pipeline demonstrates the capabilities of the Edge Analytics Suite (cf. Sect. V). The primary goal of this pipeline is trajectory forecasting, but it also offers other mobility data processing and analytics. The participating components are responsible for trajectory cleaning, augmentation and prediction, alongside a powerful visual analytics tool that elegantly presents the output. The Cleaning component includes modules responsible for tasks like deduplication, speed and bearing calculation, outlier and stationarity detection, etc. The Augmentation component includes resampling, Point of Interest (POI) tagging, segmentation and flock detection modules. Lastly, the forecasting module includes a model that predicts the future locations of a moving object, creating a predicted trajectory by using a set of prior coordinates that have been received for the specific object [3]. Fig. 8) presents the results of the forecasting module as part of the Visual Analytics GUI.

Two different deployment scenarios are evaluated, one for cloud based deployment and one for edge. The data set used is provided by MarineTraffic and it includes 10K records of moving vessels in the Saronic Gulf in Greece, spanning approximately 45 minutes. For each scenario, a specific CPU has been chosen based on its Thermal Design Power (TDP). For Cloud deployment, we have picked a desktop grade Intel i9 processor with a TDP of 125 Watts, whereas for the Edge

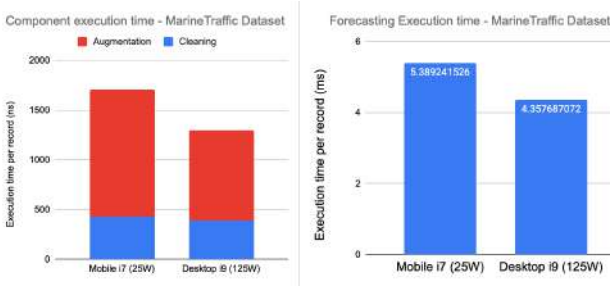

Fig. 9. Runtimes of the components of Pipeline 3 in two deployment scenarios with different CPUs (left: in nanoseconds, right: in msec).

scenario a mobile Intel i7 processor with a TDP of 25 Watts is used. Fig. 9 illustrates the difference in performance between these processors. As expected, the desktop-grade i9 processor outperforms the mobile i7, with the cleaning and augmentation phase needing 300 nanoseconds per record extra and the forecasting phase needing an extra 1 msec per record when deployed at the Edge. However, the difference between the two is low enough that in some scenarios (especially when lengthy data transfers are needed) an edge-only deployment can be substantially faster. Essentially, an Edge-only approach can be very beneficial if suitable nodes are deployed in places that allow for data to transfer over high-bandwidth connections between devices that are not very spatial distant, allowing for quick response time and time-savings compared to a Cloud-only approach that relies on large data transfers over connections with possible worse bandwidth and latency.

VIII. Conclusions

In this paper, we presented the architecture of MobiSpaces for mobility data spaces, supporting mobility data management and analytics. MobiSpaces provides a data governance framework for semantic modeling and generating enriched data descriptions, a data operations toolbox for efficient and scalable querying of mobility data in various contexts, and an edge analytics suite for distributed learning algorithms, explainability and visual analytics. A distinguishing feature of the architecture is that the combination of the aforedescribed data services with infrastructure services, namely intelligent resource allocation and task placement that takes into account the distributed nature of generation of mobility data.

Acknowledgment

The research leading to the results presented in this paper has received funding from the European Union's funded Project MobiSpaces under grant agreement no 101070279.

References

[1] The industrial internet of things: Managing and assessing trustworthiness for iiot in practice, Jun 2019.

[2] G. L. Andrienko, N. V. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao. Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Trans. Intell. Transp. Syst.*, 18(8):2232–2249, 2017.

[3] E. Chondrodima, N. Pelekis, A. Pikrakis, and Y. Theodoridis. An efficient LSTM neural network-based framework for vessel location forecasting. *IEEE Trans. Intell. Transp. Syst.*, 24(5):4872–4888, 2023.

[4] E. Curry, editor. *Real-time Linked Dataspaces - Enabling Data Ecosystems for Intelligent Systems*. Springer, 2020.

[5] Z. Deng, D. Weng, S. Liu, Y. Tian, M. Xu, and Y. Wu. A survey of urban visual analytics: Advances and future directions. *Computational Visual Media*, 9(1):3–39, 2023.

[6] C. Doulkeridis, A. Vlachou, N. Pelekis, and Y. Theodoridis. A survey on big data processing frameworks for mobility analytics. *SIGMOD Rec.*, 50(2):18–29, 2021.

[7] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Rec.*, 34(4):27–33, 2005.

[8] S. Geisler, M. Vidal, C. Cappiello, B. F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, and J. Rehof. Knowledge-driven data ecosystems toward data transparency. *ACM J. Data Inf. Qual.*, 14(1):3:1–3:12, 2022.

[9] H. Georgiou, C. Theodoridis, and Y. Theodoridis. Mobility analytics and covid-19 in greece. In *The Science behind the COVID Pandemic and Healthcare Technology Solutions*, pages 305–327. Springer, 2022.

[10] A. Graser. MovingPandas: Efficient Structures for Movement Data in Python. *GI_Forum*, 1:54–68, 2019.

[11] A. Graser, A. Jalali, J. Lampert, A. Weißenfeld, and K. Janowicz. Deep Learning From Trajectory Data: a Review of Neural Networks and the Trajectory Data Representations to Train Them. In *Proceedings of the BMDA*, 2023. (in press).

[12] A. Graser, V. Pruckovskaja, and C. Heistracher. On the Role of Spatial Data Science for Federated Learning. In *Spatial Data Science Symposium 2022 Short Paper Proceedings*, 2022. Publisher: eScholarship - University of California.

[13] S. E. Huang, Y. Feng, and H. X. Liu. A data-driven method for falsified vehicle trajectory identification by anomaly detection. *Transportation research part C: emerging technologies*, 128:103196, 2021.

[14] R. Jiménez-Peris, D. Burgos-Sancho, F. J. Ballesteros, M. Patiño-Martínez, and P. Valduriez. Elastic scalable transaction processing in LeanXcale. *Inf. Syst.*, 108:102043, 2022.

[15] N. Koutroumanis, C. Doulkeridis, C. Renso, M. Nanni, and R. Perego. TrajParquet: A trajectory-oriented column file format for mobility data lakes. In *Proceedings of SIGSPATIAL*. ACM, 2023.

[16] N. Koutroumanis, N. Kousathanas, C. Doulkeridis, and A. Vlachou. A demonstration of NoDA: Unified access to NoSQL stores. *Proc. VLDB Endow.*, 14(12):2851–2854, 2021.

[17] B. Otto, M. ten Hompel, and S. Wrobel. *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer Nature, 2022.

[18] N. Pelekis and Y. Theodoridis. *Mobility Data Management and Exploration*. Springer, 2014.

[19] M. Saeedan and A. Eldawy. Spatial Parquet: a column file format for geospatial data lakes. In *Proceedings of SIGSPATIAL*, pages 102:1–102:4. ACM, 2022.

[20] M. A. Sakr, C. Ray, and C. Renso. Big mobility data analytics: recent advances and open problems. *GeoInformatica*, 26(4):541–549, 2022.

[21] G. M. Santipantakis, K. I. Kotis, A. Glenis, G. A. Vouros, C. Doulkeridis, and A. Vlachou. RDF-Gen: generating RDF triples from big data sources. *Knowl. Inf. Syst.*, 64(11):2985–3015, 2022.

[22] R. Sethi, M. Traverso, D. Sundstrom, D. Phillips, W. Xie, Y. Sun, N. Yegitbasi, H. Jin, E. Hwang, N. Shingte, and C. Berner. Presto: SQL on everything. In *Proceedings of ICDE*, pages 1802–1813, 2019.

[23] T. Sibalija and J. P. Davim, editors. *Soft Computing in Smart Manufacturing*. De Gruyter, Berlin, Boston, 2022.

[24] ULB DSL. MobilityDB an open source geospatial trajectory data management & analysis platform, 2016.

[25] ULB DSL. MEOS mobility engine, open source., 2022.

[26] G. A. Vouros et al. Big data analytics for time critical mobility forecasting: Recent progress and research challenges. In *Proceedings of EDBT*, pages 612–623. OpenProceedings.org, 2018.

[27] G. A. Vouros, G. M. Santipantakis, C. Doulkeridis, A. Vlachou, G. L. Andrienko, N. V. Andrienko, G. Fuchs, J. M. C. Garcia, and M. G. Martinez. The datAcron ontology for the specification of semantic trajectories - specification of semantic trajectories for data transformations supporting visual analytics. *J. Data Semant.*, 8(4):235–262, 2019.

[28] E. Zimányi, M. A. Sakr, and A. Lesuisse. MobilityDB: A mobility database based on PostgreSQL and PostGIS. *ACM Trans. Database Syst.*, 45(4):19:1–19:42, 2020.