AI-Based Hybrid Data Platforms

Vassil Vassilev, Sylvia Ilieva, Iva Krasteva, Irena Pavlova, Dessislava Petrova-Antonova, and Wiktor Sowinski-Mydlarz

Abstract The current digital transformation of many businesses and the exponential growth of digital data are two of the key factors of digital revolution. For the successful meeting of high expectations, the data platforms need to employ the recent theoretical, technological, and methodological advances in contemporary computing and data science and engineering. This chapter presents an approach to address these challenges by combining logical methods for knowledge processing and machine learning methods for data analysis into a hybrid AI-based framework. It is applicable to a wide range of problems that involve both synchronous operations and asynchronous events in different domains. The framework is a foundation for building the GATE Data Platform, which aims at the application of Big Data technologies in civil and government services, industry, and healthcare. The platform implementation will utilize several recent distributed technologies such as Internet of Things, cloud, and edge computing and will integrate them into a multilevel service-oriented architecture that supports services along the entire data value chain, while the service orchestration guarantees a high degree of interoperability, reusability, and automation. The platform is designed to be compliant with the open-source software, but its open architecture supports also mixing with commercial components and tools.

Keywords Vertical layering · Semantic technologies · Horizontal integration · Service-oriented architectures · Cloud · Application containerization · Service orchestration

V. Vassilev (✉) · W. Sowinski-Mydlarz
Cyber Security Research Centre, London Metropolitan University, London, UK

GATE Institute, Sofia University, Sofia, Bulgaria
e-mail: v.vassilev@londonmet.ac.uk; w.sowinsky-mydlarz@londonmet.ac.uk

S. Ilieva · I. Krasteva · I. Pavlova · D. Petrova-Antonova
GATE Institute, Sofia University, Sofia, Bulgaria
e-mail: sylvia@gate-ai.eu; iva.krasteva@gate-ai.eu; irena.pavlova@gate-ai.eu; dessislava.petrova@gate-ai.eu

E. Curry et al. (eds.), *Data Spaces*, https://doi.org/10.1007/978-3-030-98636-0_8

1 Introduction

Europe is home to more than 50 Big Data Centers of Excellence (CoE), participating in the European Network of National Big Data Centers of Excellence [1]. Big Data for Smart Society Institute at Sofia University (GATE) is building the first Big Data CoE in Eastern Europe. Its advanced infrastructure and unique research ecosystem aim to create data services and analytical and experimentation facilities to deal with the challenge of contemporary digital revolution. The GATE Data Platform will enable high-quality research with wide scope and big impact along the entire data value chain. The platform will also support data-driven innovations and will serve the needs of multiple projects within different application domains—Future City, Smart Industry, Intelligent Government, and Digital Health. As a by-product of these activities, the GATE Data Platform will create an advanced and sustainable ecosystem for both application developers and nontechnical businesses to exploit the full potential of the available services and acquired knowledge and data. For this purpose, the GATE Data Platform will also enable creating Data Spaces with high-quality pre-processed and curated data sets, aggregating and semantically enriching data from heterogeneous sources. The acquired knowledge for management and usage of data will be made available through reusable intelligent cross-domain data models and data processing services. The GATE Data Platform will enable start-ups, SMEs, and large enterprises, as well as other organizations in a wide range of societal sectors, to build advanced data-driven services and vertical applications. This way, the GATE Data Platform will become a focal point for sharing data, services, technology, and knowledge that eases the creation of an ecosystem of diverse stakeholders, adds value to the businesses, and facilitates creation of new business and commercial models for digital transformation of the industry and the society.

These ambitious goals can be achieved effectively only with wider employment of the achievements of contemporary computing and data science and technologies. To utilize their potential, the data platforms must adopt a hybrid approach in order to address the data processing from theoretical, technological, engineering, and organizational standpoint. Artificial Intelligence allows to utilize many powerful concepts, to build complex models, to automate difficult tasks, and to manage the complexity of technical projects through knowledge representation and problem solving, decision making, and action planning, execution monitoring, and explanation. This article presents a framework for developing a hybrid data platform, which embodies many AI techniques adding intelligence along the entire data value chain.

The chapter relates to the *data management*, *data processing architectures*, *data analytics*, and *data visualization* technical priorities of the European Big Data Value Strategic Research and Innovation Agenda [2]. It addresses the respective horizontal concerns of the BDV Technical Reference Model as well as the vertical concerns of the development perspective— *engineering and DevOps*, *cybersecurity*, and *data sharing*. The chapter also relates to the *data, knowledge, and learning*, *reasoning and decision making*, *action, interaction, and explainable AI*, and *systems, hard-*

ware, methods, and tools enablers of the recent AI, Data and Robotics Strategic Research, Innovation and Deployment Agenda [3].

The rest of the chapter is organized as follows. Firstly, it reviews some of the relevant reference architectures, component models, and data platforms, existing within the European Big Data space. Next, it presents the requirements for the GATE Data Platform considering the complex role of GATE CoE as an academic and innovation hub as well as business accelerator in several domains. After these preliminaries, the next section presents the multi-layer approach to hybridization, adopted for building the GATE Data Platform. In a subsequent section, the implementation of this concept is discussed and the final chapter presents one of the flagship projects of GATE, which will leverage from the GATE Data Platform. Finally, conclusions and directions for future work are presented.

2 Brief Overview of Architectures, Frameworks, and Platforms

This section provides a brief overview of some of the most prominent reference models and component frameworks for data processing across Europe. Several examples of platforms operating at other European Big Data centers are also presented.

2.1 Reference Architectures for Big Data Processing

The European Big Data Value Association (BDVA) has proposed a reference architecture for Big Data systems [2]. It has a two-dimensional structure with components structured into horizontal and vertical concerns (Fig. 1). The horizontal concerns cover the entire data processing value chain together with the supporting technologies and infrastructure for data management, analysis, and visualization. The main sources of Big Data, such as sensors and actuators, are presented along the horizontal dimension. The vertical concerns include cross-cutting issues, relevant to all horizontal concerns—data types and formats, standards, security, and trust. Communications, development, and use are other important vertical concerns which add engineering to the vertical concerns.

This reference model provides a very high-level view of data processing without imposing any restrictions on the implementation, regardless of the area of applicability. There are more specific reference architectures developed with particular application areas in focus, such as the hierarchical model Industrie 3.0 and the three-dimensional model Industrie 4.0, which account for more detailed relationship between the business processes, but they are focused entirely on the needs of industry.

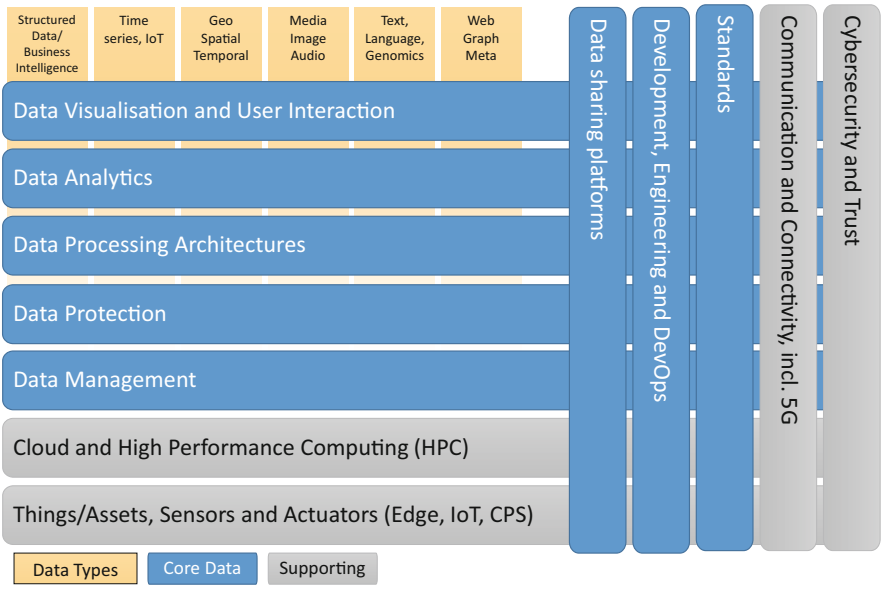

Fig. 1 Big Data Value Association reference architecture

One of the more generic enhancements of the BDVA reference architecture has been developed under the EU Horizon 2020 project OPEN DEI [4]. It aligns the reference architecture of BDVA with the requirements of open platforms and large-scale pilots for digital transformation. The OPEN DEI reference architecture framework (RAF) is built upon six fundamental principles which are generally applicable to digital platforms for data-driven services:

- Interoperability through data sharing
- Openness of data and software
- Reusability of IT solutions, information, and data
- Security and privacy
- Avoiding vendor lock-in
- Supporting a data economy

OPEN DEI reference architecture is three-dimensional (Fig. 2), with the third dimension providing directions for implementation according to the underlying philosophy of the framework. The horizontal layers include `Field Level Data Spaces`, `Edge Level Data Spaces`, and `Cloud Level Data Spaces` in which data is shared. The `Smart World Services` included in the `Field Level Data Spaces` enable interaction with IoT, automation systems, and humans. The `Edge Level Data Spaces` provide services for data acquisition, brokering, and processing. `Cloud Level Data Spaces` include different operations on the cloud such as data storage, data integration, and data intelligence. These Data Spaces offer the services to the main orthogonal dimension of the RAF—the `X-Industry Data Spaces`. The `X-Industry Data Spaces`

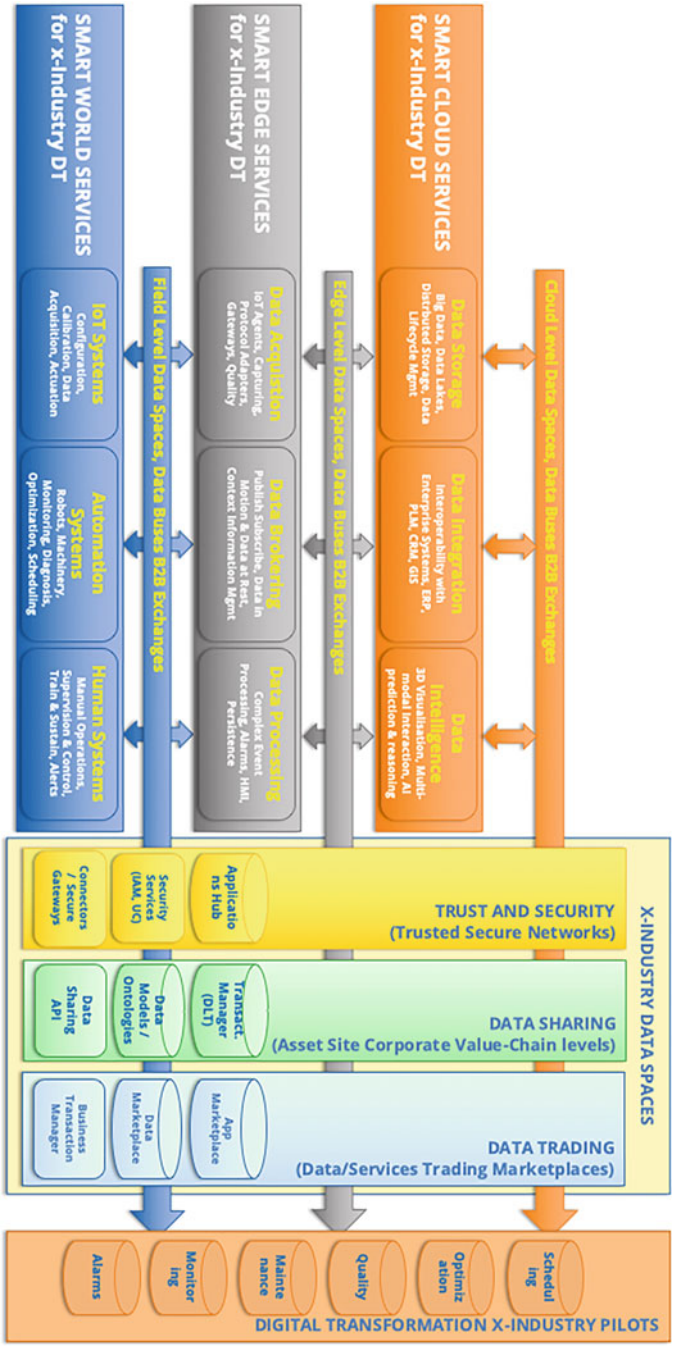

Fig. 2 OPEN DEI reference architecture

provide trustful and secure communication, data sharing, and data trading through appropriate technical infrastructure and development frameworks. All these Data Spaces support the implementation of Digital Transformation X-Industry Pilots for specific business scenarios. The main enhancement of the BDVA reference architecture by OPEN DEI RAF is in embedding the innovation and commercialization directly into the architecture through the concepts of *Data Spaces*, *smart services*, and *industry pilots*.

2.2 Component Frameworks

One of the most significant efforts to provide support for building data platforms has been undertaken by FIWARE Foundation. The FIWARE framework comprises open-source software components which can be assembled together and with other third-party components to accelerate the development of smart solutions [5]. The FIWARE component model is *broker*-based and provides an API for utilizing the functionality of the components. For this purpose FIWARE offers the so-called `Generic Enablers`, which provide support for common and specific operations for interfacing with data sources, processing, analysis, and visualization of context information, as well as usage control, publishing, and monetizing of data. The key enabler is the `Context Broker` which integrates all platform components and allows applications to update or consume the context information in a highly decentralized and large-scale manner. The `Context Broker` is the only mandatory component of any platform or solution which builds upon the FIWARE platform. A number of applications in the areas of smart agriculture, smart cities, smart energy, and Smart Industry are built upon FIWARE components.

More recent effort to provide technical support for assembling applications based on existing components is undertaken by the International Data Space Association (IDSA). Its component model elaborates further the broker architecture by standardization of two additional elements of the broker pattern—the *data provider* and the *data consumer* [6]. On the data provider side IDSA architecture is based on the concept of `Data Space` together with an associated `Connector`, while on the consumer side it operates through `DataApps` and `AppStore`. Similarly to FIWARE, the IDSA framework has an open architecture and supports large-scale system integration. IDSA has mainly in focus B2B industrial channels with extensive communications or distributed networks of institutions involved in collaborative work such as national and international government or public systems. As a member of IDSA Gate Institute considers building its own platform so that it can expose and integrate IDSA-compatible components.

2.3 Data Platforms

Platforms operating at other European Big Data Centers of Excellence include general-purpose as well as application-specific ones. The most relevant to the task of developing the GATE Data Platform, built specifically to provide support for Big Data and AI projects regardless of their application domain, are briefly presented here.

Teralab is an open Big Data and AI platform hosted by Institut Mines-Telecom (IMT)—a leading French institute of technology [7]. The platform aims to support and accelerate projects in Big Data and AI by providing technical, legal, and infrastructure tools and services as well as an ecosystem of specialists in those fields. The main asset toward providing various services is the diverse expertise hold by the Teralab teams that elaborate the best solution for each specific project.

ICE, the infrastructure and cloud research and test environment, is hosted by the RISE research institutes of Sweden and provides technical infrastructure, research data, and expert support [8]. As part of its services, ICE offers a tailor-made data platform that supports Big Data analytics and ML. The platform provides both Big Data services and customized development environment.

The Swiss Data Center has implemented RENKU platform as an open-source standalone solution with the aim of making the collaboration in data science teams more effective, trustful, and easy [9]. The RENKU platform can be deployed on a Kubernetes cluster within an organization. It supports versioning of data and code and allows customization of the environment. It enables traceability and reusability of all the artifacts developed in a data science project.

The discussed background provides a steppingstone for designing the GATE Data Platform and in particular for specifying the requirements for the platform, which are outlined in the next section. Presented reference architectures and component frameworks for data processing are designed to be general enough to support various usage scenarios and application domains and to provide common understanding of the architectural components and connections between them. By adhering to these reference architectures, the GATE platform will ensure high level of reusability of artifacts and processes, as well as of standardization and interoperability. On the other hand, the presented data platforms are a good example of different styles of utilization to be followed—from standalone instances, through service-oriented mode, to customized solutions. In addition, they demonstrate how various technologies provide support for the vast landscape of Big Data and AI projects.

3 Requirements for the GATE Data Platform

The complex role of GATE CoE as a scientific and innovation hub and business accelerator in several domains leads to multidimensional requirements:

- To empower the research on AI and Big Data technologies conducted within GATE CoE
- To enable development and exploitation of data-driven innovations
- To support the education and training activities on MSc, PhD, and professional level
- To facilitate creation of a Big Data ecosystem within the country, in the region, and in Europe

To reflect these objectives, the platform requirements were considered to be *holistic, symbiotic, open, evolving*, and *data-driven* [10], which fully aligns with the fundamental principles of BDVA and OPEN DEI reference architectures. Here we are briefly specifying them.

3.1 Requirements from Research Perspective

To support simultaneous work on research projects across the entire data value chain in different application areas, the following is required:

RR1 Vertical hybridization: Combining symbolic, statistical, and numerical AI methods with semantic technologies and Knowledge Graphs to derive value from domain knowledge

RR2 Horizontal integration: Combining multiple technologies to provide flexibility in the implementation of data services

RR3 Modularization and reusability: Integration of generic domain-independent components and data with domain-specific and problem-specific components and data for enabling the use and reuse of third-party components and APIs, such as the Fireware and Geospatial components

RR4 Synchronization and orchestration: Control over the execution of data services to support simultaneous use of the resources when working on different projects while executing the individual data services in an isolated and safe environment

RR5 Robustness and security: Coping with a wide range of problems, caused by human errors, technical faults, or external interventions

RR6 Multilevel explainability: Transparency of both the data and the operations in mapping the solutions to the problems by uncovering the circumstances and dependencies behind decisions, plans, processes, and events and thus explaining the specific results during data processing

3.2 Data-Driven Requirements

The specific requirements toward the data are:

DR1 Integration of diverse data sources: Mixing data coming from multiple data sources over different transport protocols

DR2 Support for data variability: To ensure possibility for processing data in structured, unstructured, and semi- structured formats

DR3 Multi-mode data processing: Support for different modes of data processing—batch, messaging, and streaming in both discrete and continuous flows

DR6 Scalability: Scaling out for processing large amounts of data without compromising the performance

DR5 End-to-end velocity: Capability to handle data through processing in parallel and mitigating bottlenecks and latency within the existing infrastructure

The produced datasets and metadata will be integrated into Data Spaces. A key step in realizing GATE data space is data acquisition, including public and private data. Data sharing will be realized by adhering to FAIR (findability, accessibility, interoperability, and reuse) principles:

FR1 Findability: Support of rich machine-readable metadata for automatic discovery of datasets and data services

FR2 Accessibility: Strict mechanisms for control, based on consumer profiling for both data and metadata

FR3 Interoperability: Well-defined data models, common vocabularies, and standardized ontologies for data processing

FR4 Reusability: Clear usage of licenses, detailed provenance information, and domain-relevant community standards

3.3 Service Provisioning Requirements

Currently, there is a wide variety of open-source and commercial products which can be used to implement the platform [11]. They need to be chosen in accordance with the service provisioning objectives and integrated to achieve the following:

SR1 Openness: Building on open standards, providing APIs and public data

SR2 Integration of open-source and commercial technologies: Exploiting open-source solutions as a cheaper alternative, providing for better customization and extendability, but also leveraging mature concepts, established methodologies, and stable commercial technologies to minimize migration and to foster quick innovation and commercialization

SR3 Technological agnosticism: Through relying on proven industrial and open-source solutions which support modularization, isolation, and interoperability without dependence on the underlying technology of implementation

SR4 Explainability: Through dedicated services to be able to provide for rational explanation at different level of operation, abstraction, and granularity

3.4 Data Governance Requirements

The Big Data is seen as an asset that needs to be effectively managed. This requires governance for the decentralized and heterogeneous data sharing and data processing. It should also facilitate building trust in AI as a key element for Data Spaces as defined by the recent European AI Data and Robotics Partnership [3]:

GR1 Data sovereignty and privacy: By implementing data connectors with guaranteed level of control following various compliance requirements such as GDPR, RAF, IDS, etc.

GR2 Non-repudiation and auditability: Enabling responsible development through maintenance of versioning, tracing, and auditing at all levels of operation

GR3 Trustfulness: Building trust between organizations in partnership and collaboration through enhanced data sovereignty and privacy, transparency, explainability, auditability, security, and control of the access and operation

As a conclusion, we can say that the GATE Data Platform must be open, extendible, and very flexible to improve the comprehensiveness of the different processes and enhance the transparency of its operation at theoretical, methodological, technological, and engineering levels. The architecture which can support all these requirements will need to strike a fine balance which cannot be achieved by simply endorsing the reference architectures or repeating the experience of other Big Data CoE.

4 Hybridization of Data Platforms

This section presents the theoretical, technological, and methodological choices behind the hybrid approach adopted for the GATE Data Platform.

4.1 Multilevel and Service-Oriented Architectures

Traditionally, AI as an area of advanced research and development has been divided into several sub-domains: knowledge representation, automated inference, problem solving, decision making, machine learning, etc. Although most of them are relevant to data processing, only a few are directly present in data platforms. There is an urgent need to bridge the different AI sub-domains on theoretical, methodological, technological, and engineering levels in order to add intelligence to data processing

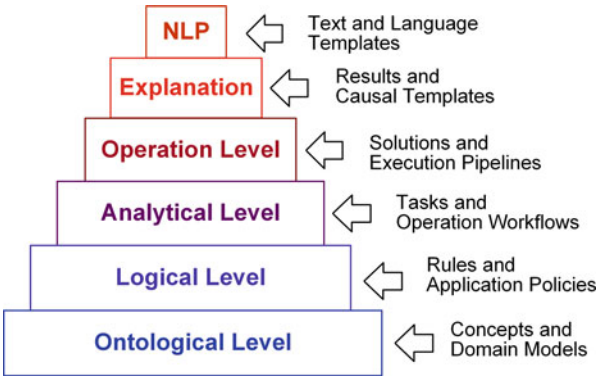

Fig. 3 Vertical layering of the GATE Data Platform

along the entire value chain and on all levels. Our approach, multilevel conceptualization, allows for a seamless integration of several AI technologies as shown in Fig. 3. The backbone of the architecture is the mapping between the levels. The *ontological*, *logical*, *analytical*, *operation*, *explanation*, and *linguistic* levels are based on a common foundation—the theory of situations and actions, which allows to model both the statics and the dynamics in a single framework [12]. Technological support for the mapping between the layers comes from the "layered cake" of the Semantic Web serialized languages. The software implementation is based on the service-oriented architectures (SOA), which utilize the containerization and orchestration capabilities of contemporary cloud technology.

4.2 Levels of Intelligence

The multilevel architecture of the platform can enhance Big Data projects through adding intelligence on several levels:

Ontological level: Models the metadata, the domain knowledge, the methods, and algorithms for data processing as well as the parameters of the processes they generate during execution. Fundamental concepts on this level in our approach are the *situations*, which describe the static state of affairs; *events*, which formalize the asynchronous results of completing the data processing; and *actions*, which model the synchronous data operations. On this level the model is represented as an OWL ontology.

Logical level: Specifies the heuristics which control the execution of data management operations, referring to the concepts and individuals on ontological level. They are modeled using SWRL rules.

Analytical level: Integrates the two previous levels into operational workflows, modeled as RDF graphs. These workflows are much simpler than BPEL work-

flows as there is no need to use the web services API for remote orchestration of the services.

Operational level: Controls the workflows for executing data processing operations such as collection, pre-processing, transportation, aggregation, storing, analyzing, and interpretation of the data, together with the events associated with the pathways during executing the data processing workflows. Each operation will be implemented as a software component, configured and deployed to the cloud containers using the metadata from the ontology, and controlled by the workflow monitoring tools available on the platform.

Explanation level: Generates rational explanation based on the representations of the causal dependencies and their logic on ontological, logical, and analytical levels, plus the results of data processing on operational level. It is based on a separate ontology, which can be extended to different domains and problems with domain-specific and problem-specific concepts, roles, and heuristics especially for explanation [13].

Linguistic level: The attributes of the various ontological concepts form a case-based grammar. It can be used for template-based generation of the text narrative of the explanation [14].

The top concepts of the ontological hierarchies are the OWL classes `Problem`, `Data`, and `Solution`, which represent the domain-specific and problem-specific information. The taxonomies of `Infrastructure` and `Resource` classes describe the available software and hardware components and tools for data processing on the platform. On logical level this model can be expanded further with domain-specific, problem-specific, and application-specific heuristics. From the OWL and SWRL representations, we can extract information to build a pure RDF graph, which forms the basis for the models on the next levels (Fig. 4).

Figure 5 illustrates the analytical model of a possible scenario for data analysis in the form of such an AND-OR graph. It can be expanded further on analytical level based on additional logical analysis and heuristic decisions. Later on, this graph can be used to control the execution of the workflow operations. The above multi-layer model of the data platform has been designed to allow seamless integration of knowledge and data representation, problem solving, data analytics, and action planning in a single conceptual framework. On the other hand, it splits the domain-specific from problem-specific and application-specific logics, which supports high modularization, interoperability, and reusability on all levels of operation. Our recent research also shows the possibility to integrate decision-making components with it based on stochastic process control, thus adding further capability to the framework [16]. The use of explicit representation of knowledge in OWL and SWRL also allows to address the problem of explainability, which is important for presenting the work of the platform in a rational way to both professionals and non-specialists [3].

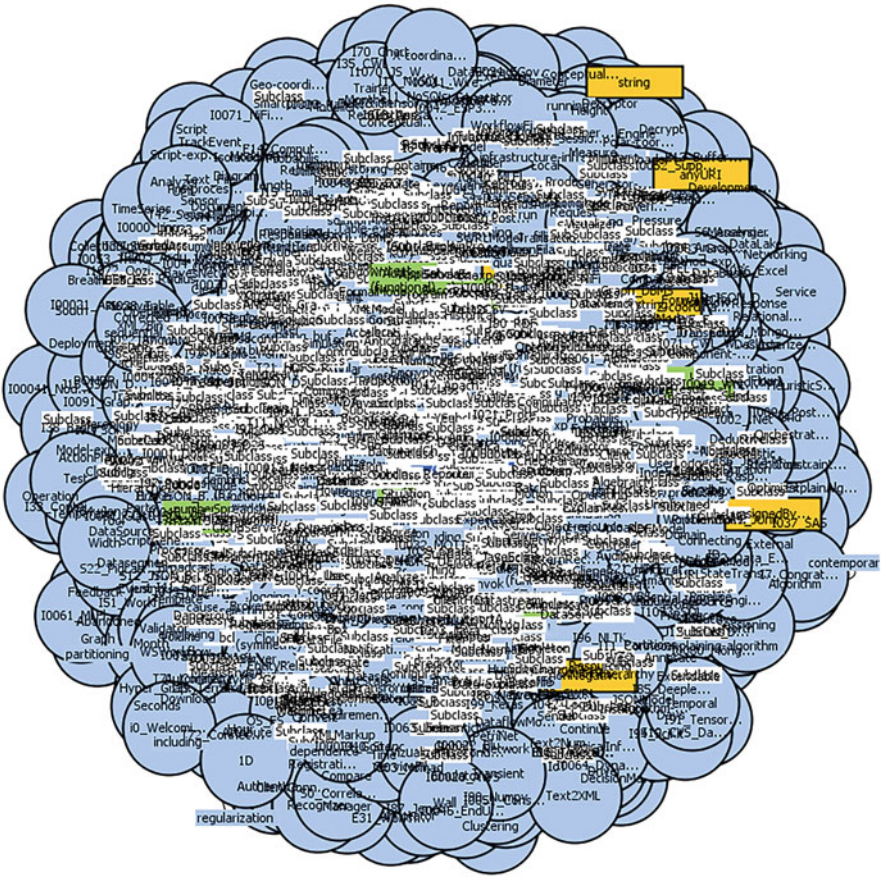

Fig. 4 Ontology of data processing on the platform

4.3 System Architecture

Many of the requirements for data platform can be met by contemporary SOA (Fig. 6). The cloud-based infrastructure for such an architecture is shown in Fig. 7. Its distinctive feature is the horizontal integration of application components through containerization and their orchestration using the service control languages of the container management system. This perfectly meets the requirements for the GATE Data Platform, supporting multiple different projects and activities in several application areas on different level and using a variety of technologies.

The system architecture shown in Fig. 7 is based on public domain software. However, this is not a limitation of the platform. Although it has been designed from the ground up using public domain software in mind, it does not exclude the use of commercial software. Nowadays the Big Data software vendors (IBM, Oracle,

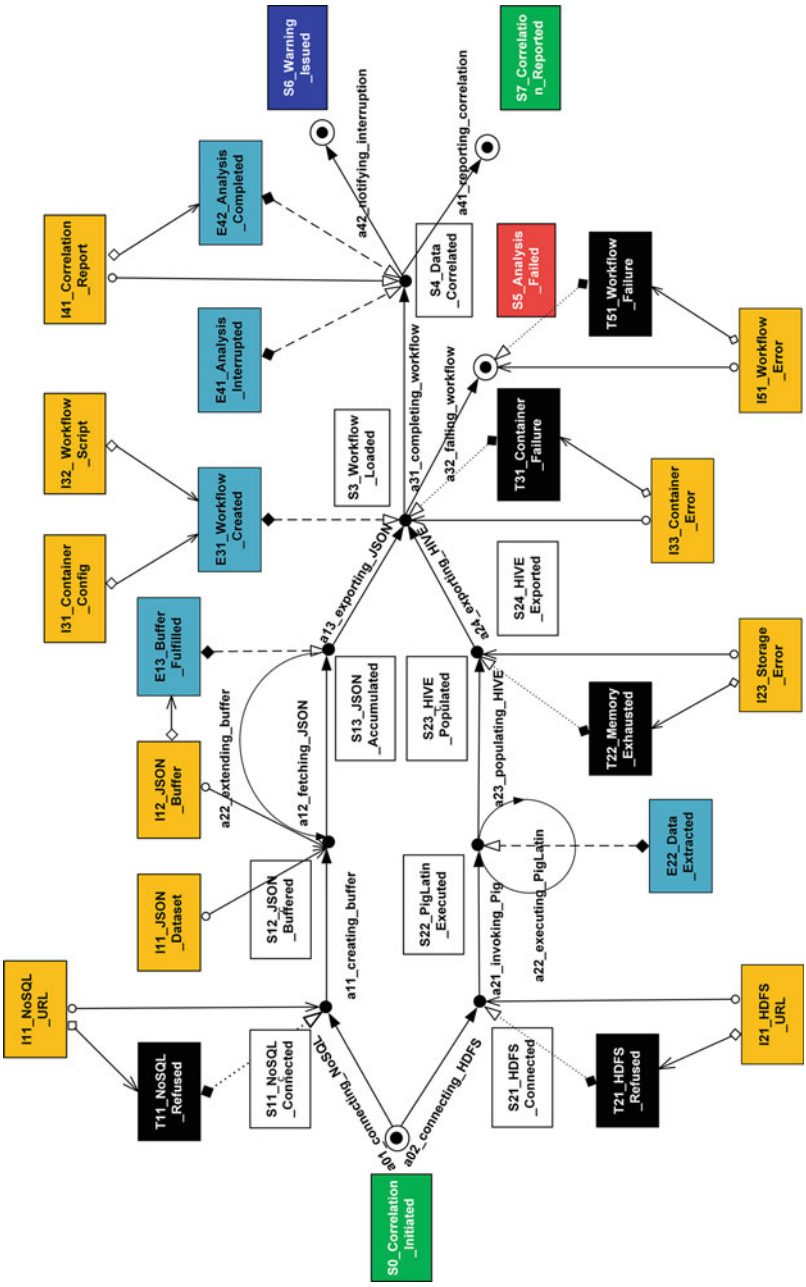

Fig. 5 Graph of the platform operations on analytical level

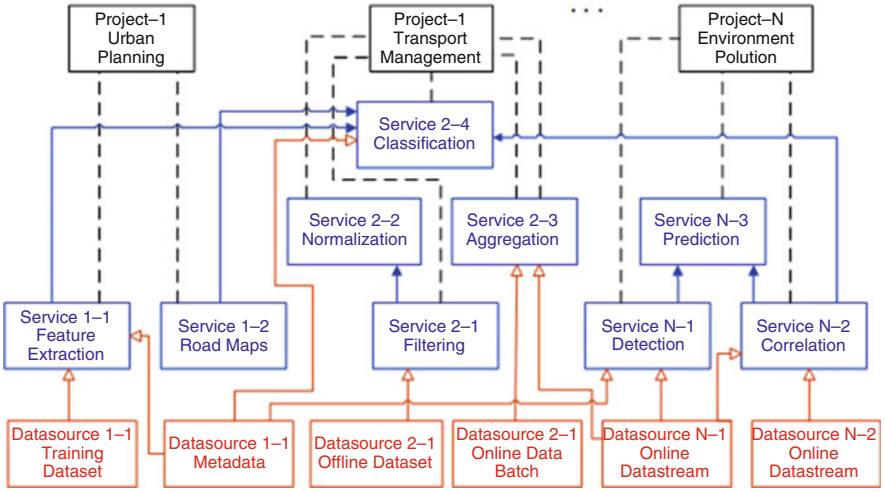

Fig. 6 Horizontal integration of the platform services

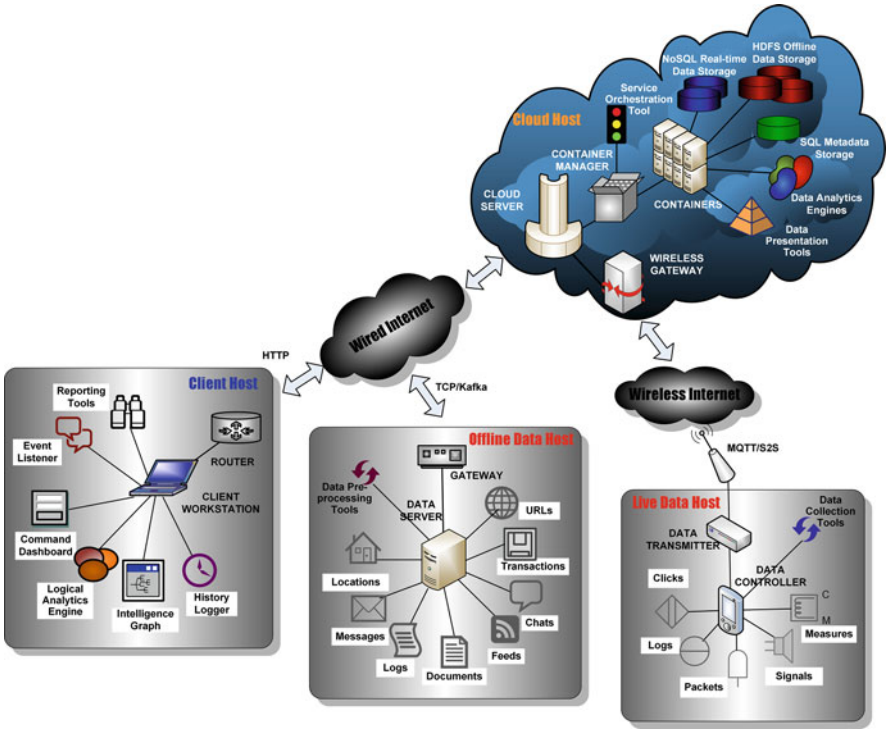

Fig. 7 Cloud-based system architecture of the GATE Data Platform

Microsoft, Hortonworks, Cloudera, MapR, etc.) also develop their own software based on open platforms and this allows compatibility between open-source and commercial technology stacks.

5 Implementation

The platform as presented can support a wide range of problems which involve both synchronous operations and asynchronous events. Such problems are typical in most of the application themes of GATE and especially in future cities and Smart Industry where the potential projects will need to deal with production line fault recovery, critical infrastructure protection, urban planning, public safety management, etc.

5.1 Enabling Technologies

Contemporary cloud technology relies on several cornerstones—application *containerization*, container *isolation*, process *synchronization*, and service *orchestration*. Cloud deployment is especially attractive for data platforms due to the support for SOA. Initially cloud hosting was pushed by big software vendors like Amazon, Google, Microsoft, and Oracle, which introduces dependence on the service providers. Hosting data platform on public cloud might not be feasible for project-oriented organizations such as GATE due to the large running costs. Fortunately, nowadays this computing paradigm can be implemented on the premises using open-source software [15, 17], which allows to untangle the dependence from the service providers. Of course, this introduces additional requirements for the maintenance. At the same time it gives more opportunities for system integration, software reuse, and optimization of the costs. Since the cloud service provision does not differ in the case of public from private cloud hosting, it can be easily combined, which would combine the benefits of both solutions.

The GATE Data Platform implementation relies on cloud software which exists in both vendor-proprietary and open-source versions. The two scripting languages for managing cloud resources supported by most container management systems—YAML [18] and CWL [19]—are sufficient for specification, deployment, and controlling the execution of the data services on the cloud and their orchestration in data processing workflows. The control can be implemented using cloud dashboards such as Apache Airflow [20], which monitors the deployment and execution of containerized software components. Such an approach has the advantage of reducing the requirements for the client and significantly simplifies the maintenance.

The layering of the GATE Data Platform allows additional automation of the component deployment, service configuration, and workflow orchestration on the cloud. The scripts needed can be generated directly from the OWL ontology, the SWRL heuristics, and the Knowledge Graphs created on the previous levels. When

the general task requires a local workflow of data processing operations, which has to be executed directly within the data management system, it can also be generated in the command language supported by it, like Oozie in Hadoop [21], JavaScript in NoSQL, or stored procedures in SQL databases.

5.2 Data Services

The data services must support the data processing operations along the entire Big Data value chain and will be the main focus of the research and innovation projects of GATE. The process of developing containerized components for data analysis on the cloud based on two popular methods for ML—SVM and NN—is described in [22]. The GATE Data Platform does not impose any restrictions on the programming languages, libraries, and tools, but will require parameterization to support the interoperability and reusability.

The data services will run within separate containers under the control of the container management system of the cloud. Each containerized component will be developed according to a common methodology, which will be based on the use of templates for configuration, deployment, and controlling the execution. This will increase the productivity of the development and will support additionally the automation of the deployment.

5.3 Engineering Roadmap

The major pathways supported by GATE Data Platform are the following:

Data warehousing and analysis of data at rest. Big Data requires powerful infrastructure capable of running HDFS-compatible data management system such as Hadoop [23], installed on cloud-enabled container management systems and executing services within containers. The enterprise tools for transporting the data are Kafka [24] and NiFi [25]. From platform perspective the analytical engines, including machine learning tools, are parametrized black boxes and their specifications will become part of the top-level ontology of the platform. The analysis will be performed by containerized applications organized in data processing workflows under the control of powerful tools such as Spark [26] or ad hoc programs which include ML software libraries such as TensorFlow in Python [27] or Deeplearning4j in Java [28].

Data streaming, integration, and analysis of data in motion. Using simple IoT controllers such as ESP32 [29] or Arduino [30], over wireless protocols such as MQTT [31], the sensor data can be transported and pre-processing in real-time on the fly, can be stored to NoSQL databases [32] for later analysis using suitable data analytics and machine learning tools.

Conceptual modeling of the explanation. Using ontological editors such as Protege the analysts and domain experts can develop problem-solving, machine learning, decision-making, and operation models for explanation. They can be then stored in graph databases such as GraphDB [33] for later use during explanation of the entire data processing from conceptual, theoretical, technological, and computational viewpoint.

Data services. As a by-product the cloud-based GATE Data Platform can also support various data services offered to third parties—downloading of datasets, broadcasting of live data streams, online and offline data pre- and post-processing, data analytics on demand, etc.

To employ the GATE Data Platform, the project teams must complete several steps:

1. Develop domain- and problem-specific ontologies to extend the ontology of data processing.
2. Specify the problem-solving heuristics for solving particular problems on logical level and index them against the ontologies.
3. Generate the working scenarios and extend them with decision-making heuristics to control the execution on analytical level.
4. Develop domain- and problem-specific ontologies for explanation of the methods, algorithms, tasks, solutions, and results.
5. Develop software components implementing specific methods for data management, data analysis, and data insight for various tasks on operational level.
6. Generate the containerization and orchestration scripts needed to deploy the software components using the ontological representation and metadata.

The operations required for setting up a project may look demanding, but the SOA of the platform allows to use design patterns. To leverage on this GATE considers adopting a suitable model-centric methodology of working and dedicated team training. On the other hand, the multi-layered architecture allows focusing on a specific vertical level and/or horizontal component which will lower the staff requirements.

After completing some project tasks the more universal components can be incorporated in the library of platform services for further use. Such an incremental development will allow the platform to grow and expand over time without the need for changing its core.

6 City Digital Twin Pilot

This section presents one of the flagship projects of GATE which will benefit from the use of data platform supporting hybrid AI. At the same time, it is a testbed of the philosophy behind it. City Digital Twin is a large interdisciplinary pilot project that aims at developing a Digital Twin platform for designing, testing, applying,

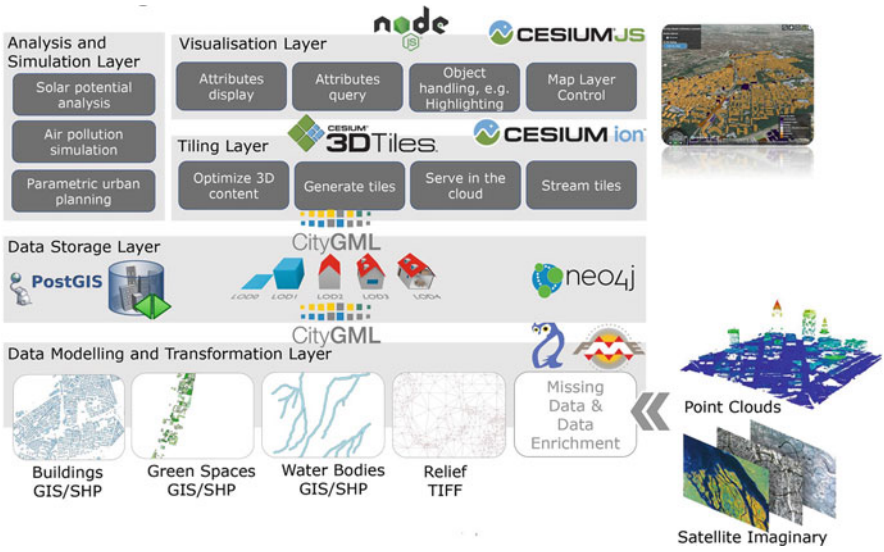

Fig. 8 Multi-layer framework of city Digital Twin pilot

and servicing the entire lifecycle of the urban environment. The project is focused on a spatially resolved exploration of a broad range of city-specific scenarios (urban climate, mobility of goods and people, infrastructure development, buildings' energy performance, air and noise pollution, etc.). The core of the platform is a semantically enriched 3D model of the city. Simulation, analytical, and visualization tools will be developed on top of it enabling the basic idea of the Digital Twin "design, test, and build first digitally." The technological framework used for initial implementation of the platform is shown in Fig. 8.

The development has started with implementation of a CityGML-compliant 3D model at ontological level, covering District Lozenets of Sofia. CityGML is an official international standard of the Open Geospatial Consortium (OGC). It is implemented as a GML application schema [34]. Because CityGML is based on GML, it can be used with the whole family of GML-compatible web services for data access, processing, and cataloguing, such as Web Feature Services, Web Processing Services, and Catalogue Services. It allows to model the significant objects in a city taking into account their 3D geometry, 3D topology, semantics, and visual appearance. Explicit relationships and component hierarchies between objects are supported and thus the 3D model is applicable to urban planning, environment analysis, 3D cadastres, and complex simulations [35]. Table 1 presents a mapping between the multi-layer framework of the GATE Data Platform and the technological layered framework of city Digital Twin platform. The city digital framework spans over several data pathways of the GATE Data Platform: (1) data at rest, (2) conceptual explanation, (3) data services, and in the future—(4) data in motion.

Table 1 Mapping of pilot layers to GATE Data Platform levels

	Modeling and transformation	Data storage	Analysis and simulation	Tiling	Visualization
Explanation	✓				✓
Operation		✓	✓	✓	✓
Analytics	✓		✓		
Logics	✓				
Ontology	✓				✓

The 3D model is based on three main data sources: cadastral data, covering most of thematic modules of CityGML standard, such as buildings, green spaces, relief, road network, etc.; high-resolution satellite image; and point cloud data. The satellite imaginary and point cloud data are used for semantic enrichment of the 3D model as well as for urban analysis, such as cadastre validation and urban change detection. Currently, the 3D model covers `Building` and `Relief` thematic modules in CityGML 2.0, including information about the buildings addresses as well as their intersection with the terrain. It is mainly developed using FME software, which allows to create and reuse data integration workflows. Additional transformations, related to the integration of the buildings and terrain, are performed, using MathLab. At operational level, the 3D model is stored in a 3D City Database [36], which can be implemented on either Oracle Spatial/Locator or PostgreSQL/PostGIS. PostGIS database is chosen for the current implementation of the platform.

Regarding the second data pathway, a domain-specific city data model and a corresponding ontology will be elaborated at ontological level for the purpose of urban planning. Thus, the full potential for mixing symbolic and graphic representation of information in Knowledge Graphs using graph databases, such as Neo4j [37] or GraphDB [33], will be exploited at the operational level. The domain model is needed to establish the basic concepts and semantics of the city domain and help to communicate these to GATE stakeholders. NGSI-LD [38] is chosen for its implementation, since it allows for specification of rules, which control the execution of data management operations at the logical level. NGSI-LD supports both the foundation classes which correspond to the core meta-model and the cross-domain ontology. The core meta-model provides a formal basis for representing a "property graph" model using RDF/RDFS/OWL. The cross-domain ontology is a set of generic, transversal classes which are aimed at avoiding conflicting or redundant definitions of the same classes in each of the domain-specific and problem-specific ontologies.

The third data pathway is realized through sharing the 3D model for user interaction through the web. The 3D model is currently uploaded to a Cesium ion platform, which optimizes and tiles it for the web. Cesium ion serves the 3D model in the cloud and streams it to any device. A web application is developed for visualization of the building model (Fig. 9) which will become part of the explanation level. It is currently hosted on a Node.js web server. Cesium.js is used

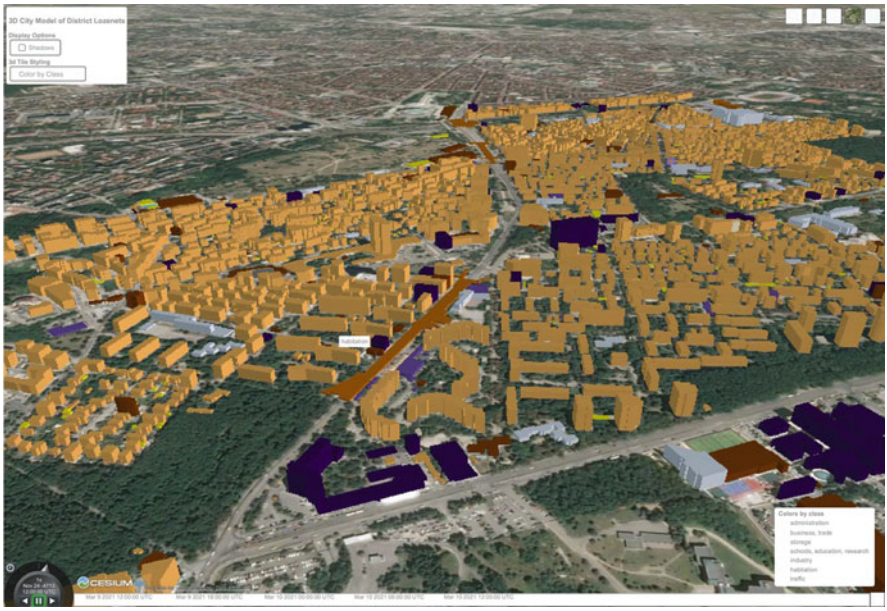

Fig. 9 3D model visualization of Lozenets District in Sofia

for its implementation due to its extensive support of functionality, such as attribute display and query, object handling, highlighting, map layer control, etc.

Several use cases in a process of implementation demonstrate the potential of the GATE Data Platform in urban setting. The first one is related to urban planning. The main idea behind it is to develop a tool for parametric urban design, supporting urban planning by taking into account neighborhood indicators related to population, green areas, transport connectivity, etc. The logic of the parametric urban design and its implementation using genetic algorithms fit within the logical, analytical, and operation level, respectively. The second use case deals with analysis and simulation of air quality, focusing on pollution dispersion independent of the wind direction and velocity, as well as the geometry of the buildings. In collaboration with researchers from Chalmers University, the wind flow in an urban environment is explored by applying computational fluid dynamics. The simulations themselves are implemented on the operational level, while their visualization corresponds to the explanation level of the GATE Data Platform.

The fourth, real-time data pathway of the GATE Data Platform will be fully operational after GATE City Living Lab is completed. The Living Lab will generate data for air, weather, and noise monitoring and will continuously deliver data to the cloud for real-time data analysis. A pilot implementation for processing of data about air pollution, generated by open air sensors across the city, is already on the way. When the GATE Data Platform is fully operational, we plan to transfer the

entire project to it, which would allow us to reuse various components in other projects related to the analysis of the urban environment.

7 Conclusion and Future Work

The data platform under development for the GATE CoE is unique in the way it combines theoretical, technological, and applied aspects in a simple but powerful multi-layered hybrid framework, based on AI and empowered by the cloud technologies. The separation of domain-specific from problem-specific knowledge at ontological, logical, and analytical levels allows detaching the tasks for specification and modeling from the technical tasks for processing the data, which is the cornerstone of the interoperability of both data and operations. At the same time, it facilitates explanation on different level and with different granularity. Furthermore, thanks to the containerization and the orchestration of data services, the platform adds high degree of automation, reusability, and extendibility. Still, this hybrid platform can be used in a uniform way, regardless of the mode of working—locally, remotely, over network, or on the cloud.

The possibility for vendor-independent implementation of the platform, based on open software, very well supports both academic teaching and professional training practices, which is an additional advantage for GATE. In order to leverage the full advantages of AI technologies for data processing presented in this chapter, the software development for research, innovation, and commercialization requires conceptual, methodological, and technological discipline which will gradually become a practice at the GATE CoE.

GATE has already completed the development of the ontological level of its data platform and currently proceeds with formulation of heuristics, which will guide its operations. The immediate follow-up plans include developing of an ontological model of explanation, which will complete its conceptual framework as an explainable AI framework. GATE is relying on both its academic partners and its industrial supporters to build the technical infrastructure needed to implement this solution, and it is expected that by the end of the year, the GATE Data Platform will be fully operational. Its belief is that this framework can be of interest to other organizations with similar goals within the European Big Data space.

Acknowledgments This research work has been supported by GATE project, funded by the H2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155, by Operational Programme Science and Education for Smart Growth under grant agreement no. BG05M2OP001-1.003-0002-C01, and by London Metropolitan University Transformation project no. 0117/2020, funded by UK HEIF. GATE is also grateful for the continuing support of the companies Rila Solutions and Ontotext. The understanding, the concepts, and the claims formulated in this material are of the authors only and should not be attributed to the official policy of any of these organizations.

References

1. KNOW-CENTER GmbH: European Network of National Big Data Centers of Excellence. Retrieved March 9, 2021 from https://www.big-data-network.eu/map/
2. Zillner, S., Curry, E., Metzger, A. et al. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Big Data Value Association.
3. Zillner, S., Bisset, D., Milano, M., Curry, E. et al. (Eds.). (2020). Strategic research, innovation and deployment agenda—AI, data and robotics partnership. Third Release. September 2020, Brussels. BDVA, euRobotics, ELLIS, EurAI and CLAIRE. Retrieved March 9, 2021 from https://ai-data-robotics-partnership.eu/wp-content/uploads/2020/09/AI-Data-Robotics-Partnership-SRIDA-V3.0.pdf
4. OpenDei Project: Reference Architecture for Cross-domain Digital Transformation. Retrieved March 9, 2021 from https://www.opendei.eu/wp-content/uploads/2020/10/
5. Fiware Foundation, e.V.: FIWARE-NGSI v2 Specification. Retrieved March 9, 2021 from http://fiware.github.io/specifications/ngsiv2/stable/
6. International Data Spaces Association: Reference Architecture Model Version 3.0 (2019). Retrieved March 9, 2021 from https://internationaldataspaces.org/publications/
7. Institut Mines-Telecom: Data Science for Europe. Artificial intelligence and big data platform. Retrieved March 9, 2021 from https://www.teralab-datascience.fr/?lang=en
8. RISE Research Institutes of Sweden: ICE Data center. Retrieved March 9, 2021 from https://www.ri.se/en/ice-datacenter
9. Swiss Data Science Center: Multidisciplinary Data Science Collaborations Made Trustful and Easy. Retrieved March 9, 2021 from https://datascience.ch/renku/
10. Petrova-Antonova, D., Krasteva, I., Ilieva, S., & Pavlova, I. (2019). Conceptual architecture of GATE big data platform. In *Proc. 20th Int. Conf. on Computer Systems and Technologies (CompSysTech)* (pp. 261–268). ACM.
11. Petrova-Antonova, D., Ilieva, S., & Pavlovam I. (2017). Big data research and application—a systematic literature review. *Serdica Journal of Computing, 11*(2), 73–114.
12. Bataityte, K., Vassilev, V., & Gill, O. (2020). Ontological foundations of modelling security policies for logical analytics. In *IFIP Advances in Information and Communication Technology (IFIPAICT)* (Vol. 583, pp. 368–380). Springer.
13. Chari, S., Seneviratne, O., Gruen, D., et al. (2020). Explanation ontology: A model of explanations for user-centered AI. In J. Pan, V. Tamma, C. d'Amato et al. (Eds.), *Proc. 19th Int. Semantic Web Conference (ISWC)*. LNCS (Vol. 12507, pp 228–243). Springer.
14. van Deemter, K., Theune, M., & Krahmer, E. (2005). Real versus template-based natural language generation: A false opposition? *Computational Linguistics, 31*(1), 15–24.
15. Docker, Inc.: Get Started with Docker. Retrieved February 19, 2021 from https://www.docker.com
16. Vassilev, V., Donchev, D., & Tonchev, D. (2021). Risk assessment in transactions under threat as a partially observable Markov decision process. In *50th Int. Conf. Optimization in Artificial Intelligence and Data Sciences (ODS2021)*, 14–17 Sep 2021, Rome, Springer.
17. Cloud Native Computing Foundation: Building sustainable ecosystems for cloud native software. Retrieved February 19, 2021 from https://www.cncf.io/
18. yaml.org: YAML Ain't Markup Language. Retrieved February 9, 2021 from https://yaml.org/
19. Amstutz, P., Crusoe, M., & Tanic, N. (Eds.). Common Workflow Language, v1.0.2. Retrieved February 19, 2021 from https://w3id.org/cwl/v1.0/
20. Apache Software Foundation: AirFlow. Retrieved February 19, 2021 from http://airflow.apache.org/
21. Apache Software Foundation: Oozie. Retrieved February 28, 2021 from https://oozie.apache.org/
22. Sowinsky-Mydlarz, W., Li, J., Ouazzane, K., & Vassilev, V. (2021). Threat intelligence using machine learning packet dissection. In *20th Int. Conf. on Security and Management (SAM21)*, Las Vegas, USA. Springer.

23. Apache Software Foundation: Hadoop. Retrieved February 19, 2021 from https://hadoop.apache.org/
24. Apache Software Foundation: Kafka. Retrieved February 19, 2021 from https://kafka.apache.org/
25. Apache Software Foundation: NiFi. Retrieved February 19, 2021 from https://nifi.apache.org/
26. Apache Software Foundation: Spark—Unified Analytics Engine for Big Data. Retrieved February 9, 2021 from https://spark.apache.org/
27. tensorflow.org: An end-to-end open source machine learning platform. Retrieved February 9, 2021 from https://www.tensorflow.org/
28. Eclipse Foundation: Deep Learning for Java. Retrieved February 9, 2021 from https://deeplearning4j.org/
29. Espressif Systems: SoCs. Retrieved February 9, 2021 from https://www.espressif.com/en/products/socs
30. Arduino: Arduino Pro. Retrieved February 9, 2021 from https://store.arduino.cc/new-home/iot-kits
31. MQTT.org: MQTT: The Standard for IoT Messaging. Retrieved February 19, 2021 from https://mqtt.org/
32. MongoDB, Inc.: The database for modern applications. Retrieved February 9, 2021 from https://www.mongodb.com/try/download/community
33. Ontotext: GraphDB—The Best RDF Database for Knowledge Graphs. Retrieved February 9, 2021 from https://www.ontotext.com/products/graphdb/
34. Open Geospatial Consortium Europe, CityGML. Retrieved March 8, 2021 from https://www.ogc.org/standards/citygml
35. Gröger, G., Kolbe, T., Nagel, C.. & Häfele, K. (2012). OGC City Geography Markup Language (CityGML) Encoding Standard, Wayland MA: Open Geospatial Consortium.
36. Kolbe, T., Nagel, C., Willenborg, B. et al. 3D City DB. Retrieved March 3, 2021 from https://www.3dcitydb.org/3dcitydb/d3ddatabase/
37. Neo4j, Inc.: Introducing Neo4J. Retrieved March 1, 2021 from https://neo4j.com/
38. Context Information Management (CIM) Industry Specification Group (ISG), NGSI-LD API. Retrieved March 3, 2021 from https://www.etsi.org/deliver/etsi_gs/CIM/
39. Vassilev, V., Sowinski-Mydlarz, W., Gasiorowski, P. et al. (2020) Intelligence graphs for threat intelligence and security policy validation of cyber systems. In *Advances in Intelligent Systems and Computing (AISC)* (Vol. 1164, pp. 125–140). Springer.