Big Data and AI Pipeline Framework: Technology Analysis from a Benchmarking Perspective

Arne J. Berre, Aphrodite Tsalgatidou, Chiara Francalanci, Todor Ivanov, Tomas Pariente-Lobo, Ricardo Ruiz-Saiz, Inna Novalija, and Marko Grobelnik

Abstract Big Data and AI Pipeline patterns provide a good foundation for the analysis and selection of technical architectures for Big Data and AI systems. Experiences from many projects in the Big Data PPP program has shown that a number of projects use similar architectural patterns with variations only in the choice of various technology components in the same pattern. The project DataBench has developed a Big Data and AI Pipeline Framework, which is used for the description of pipeline steps in Big Data and AI projects, and supports the classification of benchmarks. This includes the four pipeline steps of Data Acquisition/Collection and Storage, Data Preparation and Curation, Data Analytics with AI/Machine Learning, and Action and Interaction, including Data Visualization and User Interaction as well as API Access. It has also created a toolbox which supports the identification and use of existing benchmarks according to these steps in addition to all of the different technical areas and different data types in the BDV Reference Model. An observatory, which is a tool, accessed via the toolbox, for observing the popularity,

A. J. Berre
SINTEF Digital, Oslo, Norway
e-mail: Arne.J.Berre@sintef.no

A. Tsalgatidou (✉)
SINTEF, Oslo, Norway

Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

C. Francalanci
Politecnico di Milano, Milan, Italy

T. Ivanov
Lead Consult, Sofia, Bulgaria

T. Pariente-Lobo · R. Ruiz-Saiz
ATOS Research and Innovation, Madrid, Spain

I. Novalija · M. Grobelnik
Jožef Stefan Institute, Ljubljana, Slovenia

E. Curry et al. (eds.), *Technologies and Applications for Big Data Value*,
https://doi.org/10.1007/978-3-030-78307-5_4

importance and the visibility of topic terms related to Artificial Intelligence and Big Data technologies has also been developed and is described in this chapter.

Keywords Benchmarking · Big Data and AI Pipeline · Blueprint · Toolbox · Observatory

1 Introduction

Organizations rely on evidence from the benchmarking domain to provide answers to how their processes are performing. There is extensive information on why and how to perform technical benchmarks for the specific management and analytics processes, but there is a lack of objective, evidence-based methods to measure the correlation between Big Data Technology (BDT) benchmarks and business benchmarks of an organization and demonstrate return on investment. When more than one benchmarking tool exist for a given need, there is even less evidence as to how these tools compare to each other, and how the results can affect their business objectives. The DataBench project has addressed this gap by designing a framework to help European organizations developing BDT to reach for excellence and constantly improve their performance, by measuring their technology development activity against parameters of high business relevance. It thus bridges the gap between technical and business benchmarking of Big Data and Analytics applications.

The Business Benchmarks of DataBench focus on Quantitative benchmarks for areas such as Revenue increase, Profit increase and Cost reduction and on Qualitative benchmarks for areas such as Product and Service quality, Customer satisfaction, New Products and Services launched and Business model innovation. The business benchmarks are calculated on the basis of certain business Key Performance Indicators (KPIs). The business KPIs selected by the project are valid metrics and can be used as benchmarks for comparative purposes by researchers or business users for each of the industry and company-size segments measured. These indicators have been classified in the following four features which group relevant indicators from different points of views: Business features, Big Data Application features, Platform and Architecture features, Benchmark-specific features. For each feature, specific indicators have been defined. Actually, none of the existing Big Data benchmarks make any attempt to relate the technical measurements parameters, metrics and KPIs (like Latency, Fault tolerance, CPU utilization, Memory utilization, Price performance, Energy, bandwidth, data access patterns, data processed per second, data processed per joule, query time, execution time, number of completed jobs) with the business metrics and KPIs (like operational efficiency, increased level of transparency, optimized resource consumption, improved process quality and performance), customer experience (increased customer loyalty and retention, precise customer segmentation and targeting, optimized customer interaction and service), or new business models (expanded revenue streams from existing products, creation of new revenue streams from entirely (new) data products).

This chapter presents a Big Data and AI Pipeline Framework that supports technology analysis and benchmarking for both the horizontal and vertical technical priorities of the European Big Data Value Strategic Research and Innovation Agenda [1], and also for the cross-sectorial technology enablers of the AI, Data and Robotics Strategic Research, Innovation and Deployment Agenda [2]. In the following sections, we focus on the DataBench approach for Technical Benchmarks which are using a Big Data and AI Pipeline model as an overall framework, and they are further classified depending on the various areas of the Big Data Value (BDV) Reference Model. Technical benchmarks are also related to the areas of the AI Strategic Research, Innovation and Deployment Agenda (SRIDA) [1] and the ISO SC42 Big Data and AI Reference models [3].

The DataBench Framework is accompanied by a Handbook and a Toolbox, which aim to support industrial users and European technology developers who need to make informed decisions on Big Data Technologies investments by optimizing technical and business performance. The Handbook presents and explains the main reference models used for technical benchmarking analysis. The Toolbox is a software tool that provides access to benchmarking services; it helps stakeholders (1) to identify the use cases where they can achieve the highest possible business benefit and return on investment, so they can prioritize their investments; (2) to select the best technical benchmark to measure the performance of the technical solution of their choice; and (3) to assess their business performance by comparing their business impacts with those of their peers, so they can revise their choices or their organization if they find they are achieving less results than median benchmarks for their industry and company size. Therefore, the services provided by the Toolbox and the Handbook support users in all phases of their journey (before, during and in the ex-post evaluation of their BDT investment) and from both the technical and business viewpoints.

In the following section, we present the Big Data and AI Pipeline Framework, which is used for the description of pipeline steps in Big Data and AI projects, and which supports the classification of benchmarks; the framework also serves as a basis for demonstrating the similarities among Big Data projects such as those in the Big Data Value Public-Private Partnership (BDV PPP) program [4]. We also discuss its relationship with the BDV Reference Model and the Strategic Research and Innovation Agenda (SRIA) [1] and the Strategic Research, Innovation and Deployment Agenda (SRIDA) for a European AI, Data and Robotics Partnership (AI PPP SRIDA) [2]. In Sect. 3, we present Big Data and AI Pipeline examples from the DataBio project [5] for IoT, Graph and SpatioTemporal data. In Sect. 4, we present categorizations of architectural blueprints for realisations of the various steps of the Big Data and AI Pipeline with variations depending on the processing types (batch, real-time, interactive), the main data types involved and on the type of access/interaction (which can be API access action/interaction or a Human interaction). Specializations can also be more complex aggregations/-compositions of multiple specializations/patterns. These blueprints are a basis for selecting specializations of the pipeline that will fit the needs of various projects and instantiations. Section 5 presents how existing Big Data and AI Technical Benchmarks have been classified according to the Big Data and AI Pipeline

Framework that is presented in Sect. 2. These are benchmarks that are suitable for benchmarking of technologies related to the different parts of the pipeline and associated technical areas. Section 6 describes the DataBench Toolbox as well as the DataBench Observatory, which is a tool (accessed via the toolbox) for observing the popularity, importance and the visibility of topic terms related to Artificial Intelligence and Big Data, with particular attention dedicated to the concepts, methods, tools and technologies in the area of Benchmarking. Finally, the conclusions in Sect. 7 present a summary of the contributions and plans for further evolution and usage of the DataBench Toolbox.

2 The Big Data and AI Pipeline Framework

The Big Data and AI Pipeline Framework is based on the elements of the Big Data Value (BDV) Big Data Value Reference Model, developed by the Big Data Value Association (BDVA) [1]. In order to have an overall usage perspective on Big Data and AI systems, a top-level generic pipeline has been introduced to understand the connections between the different parts of a Big Data and AI system in the context of an application flow. Figure 1 depicts this pipeline, following the Big Data and AI Value chain.

The steps of the Big Data and AI Pipeline Framework are also harmonized with the ISO SC42 AI Committee standards [3], in particular the *Collection, Preparation, Analytics* and *Visualization/Access* steps within the Big Data Application Layer of the recent international standard ISO 20547-3 Big Data reference architecture within the functional components of the Big Data Reference Architecture [3, 6]. The following figure shows how the Big Data and AI Pipeline can also be related to the BDV Reference Model and the AI PPP Ecosystem and Enablers (from SRIDA AI). Benchmarks often focus on specialized areas within a total system typically identified by the BDV Reference Model. This is in particular useful for the benchmarking of particular technical components. Benchmarks can also be directly or indirectly linked to the steps of a Big Data and AI Pipeline, which is useful when benchmarks are being considered from a Big Data and AI application perspective, where practical project experiences has shown that these steps can easily be recognized in most application contexts.

Benchmarks are useful both for the evaluation of alternative technical solutions within a Big Data and AI project and for comparing new technology develop-

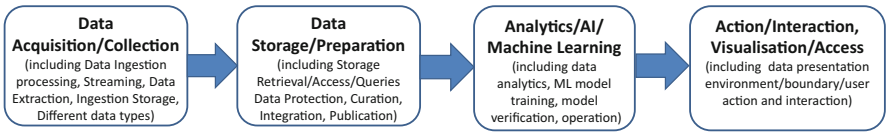

Fig. 1 Top-level, generic Big Data and AI Pipeline pattern

ments/products with alternative offerings. This can be done both from a technical area perspective for selected components and from a pipeline step perspective when seen from the steps of a Big Data and AI application.

As it can be seen in Fig. 1, this pipeline is at a high level of abstraction. Therefore, it can be easily specialized in order to describe more specific pipelines, depending on the type of data and the type of processing (e.g. IoT data and real-time processing). The 3D cube in Fig. 2 depicts the steps of this pipeline in relation to the type of data processing and the type of data being processed. As we can see in this figure, the type of data processing, which has been identified as a separate topic area in the BDV Reference Model, is orthogonal to the pipeline steps and the data types. This is due to the fact that different processing types, like batch/data-at-rest and real-time/data-in-motion and interactive, can span across different pipeline steps and can handle different data types, as the ones identified in the BDV Reference Model, within each of the pipeline steps. Thus, there can be different data types like structured data, times series data, geospatial data, media, image, video and audio data, text data, including natural language data, and graph data, network/web data and metadata, which can all imply differences in terms of storage and analytics techniques.

Other dimensions can similarly be added for a multi-dimensional cube, e.g. for Application domains, and for the different horizontal and vertical technology areas of the BDV Reference Model, and for the technology locations of the Computing Continuum/Trans Continuum—from Edge, through Fog to Cloud and HPC—for the actual location of execution of the four steps, which can happen on all these levels. The same orthogonality can be considered for the area of Data Protection, with Privacy and anonymization mechanisms to facilitate data protection. It also has links to trust mechanisms like Blockchain technologies, smart contracts and various

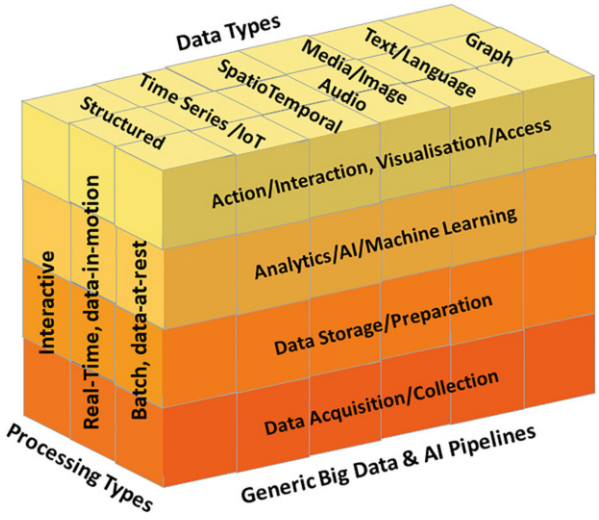

Fig. 2 Top-level, generic Big Data and AI Pipeline cube

forms for encryption. This area is also associated with the area of CyberSecurity, Risk and Trust.

The BDV Reference Model shown in Fig. 3 has been developed by the BDVA [1], taking into account input from technical experts and stakeholders along the whole Big Data Value chain as well as interactions with other related Public–Private Partnerships (PPPs). An explicit aim of the BDV Reference Model in the SRIA 4.0 document is to also include logical relationships to other areas of a digital platform such as Cloud, High Performance Computing (HPC), IoT, Networks/5G, CyberSecurity, etc.

The following describes the steps of the Big Data and AI Pipeline shown on the left of the BDV Reference Model in Fig. 3, with lines connecting them to the typical usage of some of the main technical areas.

Data Acquisition/Collection This step includes acquisition and collection from various sources, including both streaming data and data extraction from relevant external data sources and data spaces. It includes support for handling all relevant data types and also relevant data protection handling for this step. This step is often associated with the use of both real-time and batch data collection, and associated streaming and messaging systems. It uses enabling technologies in the area using data from things/assets, sensors and actuators to collect streaming data-in-motion as well as connecting to existing data sources with data-at-rest. Often, this step also includes the use of relevant communication and messaging technologies.

Data Storage/Preparation This step includes the use of appropriate storage systems and data preparation and curation for further data use and processing. Data storage includes the use of data storage and retrieval in different databases systems—both SQL and NoSQL, like key-value, column-based storage, document storage and graph storage, as well as storage structures such as file systems. This is an area where there historically exist many benchmarks to test and compare various data storage alternatives. Tasks performed in this step also include further data preparation and curation as well as data annotation, publication and presentation of the data in order to be available for discovery, reuse and preservation. Further in this step, there is also interaction with various data platforms and data spaces for broader data management and governance. This step is also linked to handling associated aspects of data protection.

Analytics/AI/Machine Learning This step handles data analytics with relevant methods, including descriptive, predictive and prescriptive analytics and use of AI/Machine Learning methods and algorithms to support decision making and transfer of knowledge. For Machine learning, this step also includes the subtasks for necessary model training and model verification/validation and testing, before actual operation with input data. In this context, the previous step of data storage and preparation will provide data input both for training and validation and test data, as well as operational input data.

Action/Interaction, Visualization and Access This step (including data presentation environment/boundary/user action and interaction) identifies the boundary between the system and the environment for action/interaction, typically through

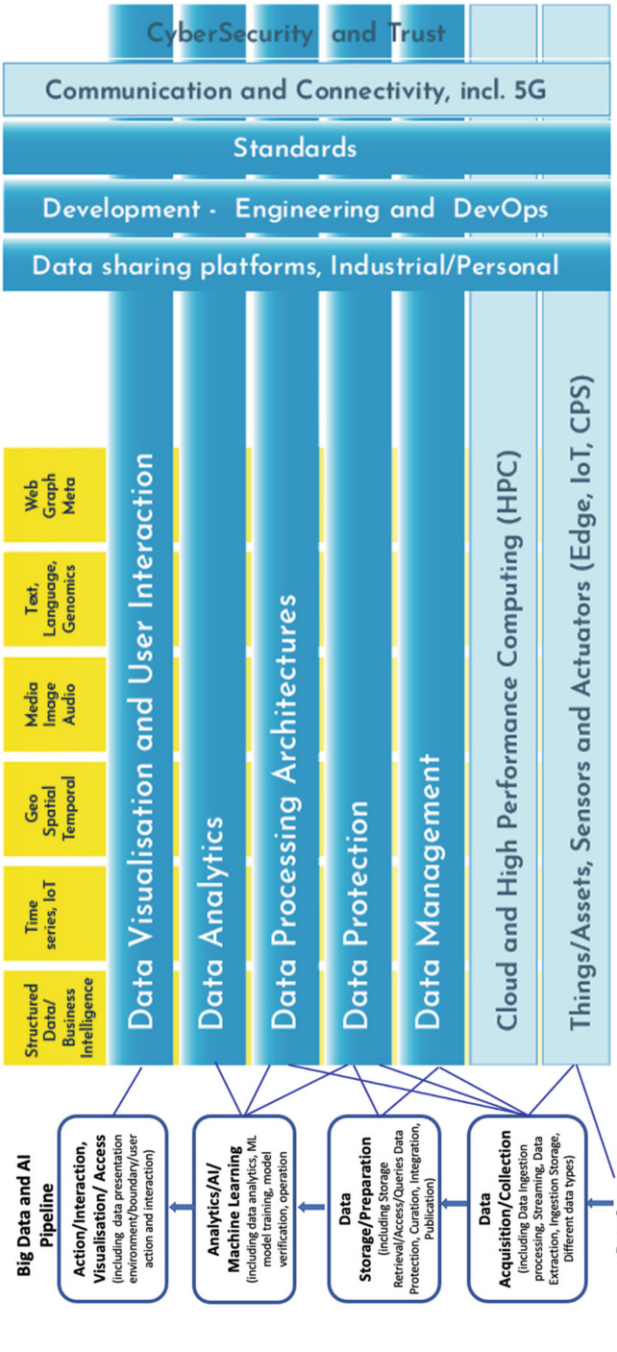

Fig. 3 Big Data and AI Pipeline using technologies from the BDV Reference Model

a visual interface with various data visualization techniques for human users and through an API or an interaction interface for system boundaries. This is a boundary where interactions occur between machines and objects, between machines, between people and machines and between environments and machines. The action/interaction with the system boundaries can typically also impact the environment to be connected back to the data acquisition/collection step, collecting input from the system boundaries.

The above steps can be specialized based on the different data types used in the various applications, and are set up differently based on different processing architectures, such as batch, real-time/streaming or interactive. Also, with Machine Learning there is a cycle starting from training data and later using operational data (Fig. 4).

The steps of the Big Data and AI Pipeline can relate to the AI enablers as follows:

Data Acquisition/Collection Using enablers from *Sensing and Perception* technologies, which includes methods to access, assess, convert and aggregate signals that represent real-world parameters into processable and communicable data assets that embody perception.

Data Storage/Preparation Using enablers from *Knowledge and learning technologies*, including data processing technologies, which cover the transformation, cleaning, storage, sharing, modelling, simulation, synthesising and extracting of insights of all types of data, both that gathered through sensing and perception as well as data acquired by other means. This will handle both training data and operational data. It will further use enablers for *Data for AI*, which handles the availability of the data through data storage through data spaces, platforms and data marketplaces in order to support data-driven AI.

Analytics/AI/Machine Learning Using enablers from *Reasoning and Decision making* which is at the heart of Artificial Intelligence. This technology area also provides enablers to address optimization, search, planning, diagnosis and relies on methods to ensure robustness and trustworthiness.

Action/Interaction, Visualization and Access Using enablers from *Action and Interaction*—where Interactions occur between machines and objects, between machines, between people and machines and between environments and machines. This interaction can take place both through human user interfaces as well as through various APIs and system access and interaction mechanisms. The action/interaction with the system boundaries can typically also be connected back to the data acquisition/collection step, collecting input from the system boundaries.

These steps are also harmonized with the emerging pipeline steps in ISO SC42 AI standard of "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML), ISO/IEC 23053, with Machine Learning Pipeline with the related steps of *Data Acquisition, Data Pre-processing, Modeling, Model Deployment and Operation*.

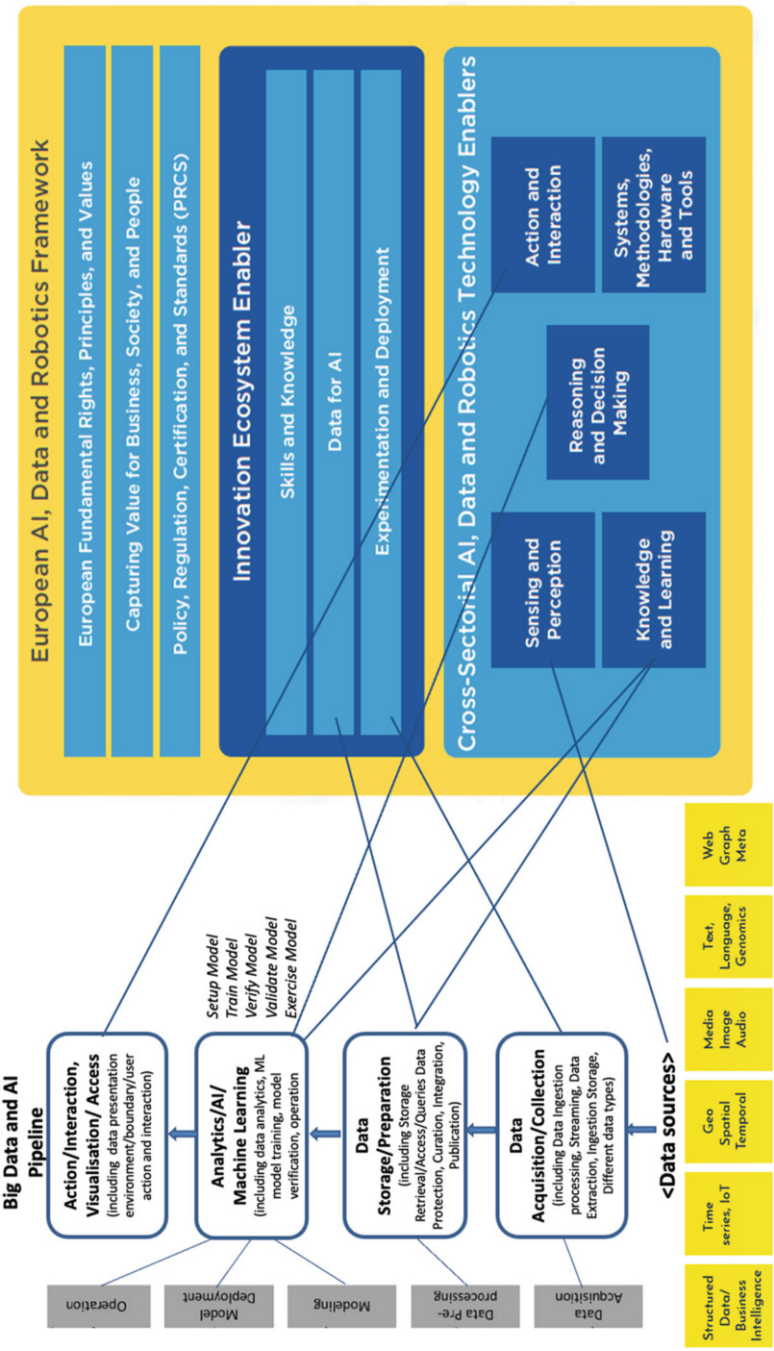

Fig. 4 Mappings between the top-level generic Big Data and AI Pipeline and the European AI and Robotics Framework

Benchmarks can be identified related both to technical areas within the BDV Reference Model and the AI Frameworks and to the various steps in the DataBench Toolbox that supports both perspectives.

3 Big Data and AI Pipeline Examples for IoT, Graph and SpatioTemporal Data: From the DataBio Project

In the following, we present example pipelines which handle different data types. Specifically, they handle IoT data, Graph data and Earth Observation/Geospatial data. Each pipeline is mapped to the four phases of the top-level Generic Big Data and AI Pipeline pattern, presented in Sect. 2. All these pipelines have been developed in the DataBio project [5], which was funded by the European Union's Horizon 2020 research and innovation programme. DataBio focused on utilizing Big Data to contribute to the production of the best possible raw materials from agriculture, forestry, and fishery/aquaculture for the bioeconomy industry in order to produce food, energy and biomaterials, also taking into account responsibility and sustainability issues. The pipelines that are presented below are the result of aggregating Big Data from the three aforementioned sectors (agriculture, forestry and fishery) and intelligently processing, analysing and visualising them.

Pipeline for IoT Data Real-Time Processing and Decision Making
The "Pipeline for IoT data real-time processing and decision making" has been applied to three pilots in the DataBio project from the agriculture and fishery domain, and, since it is quite generic, it can also be applied to other domains. The main characteristic of this pipeline is the collection of real-time data coming from IoT devices to generate insights for operational decision making by applying real-time data analytics on the collected data. Streaming data (a.k.a. events) from IoT sensors (e.g. are collected in real-time, for example: agricultural sensors, machinery sensors, fishing vessels monitoring equipment. These streaming data can then be pre-processed in order to lower the amount of data to be further analysed. Pre-processing can include filtering of the data (filtering out irrelevant data and filtering in only relevant events), performing simple aggregation of the data, and storing the data (e.g. on cloud or other storage model, or even simply as a computer's file system) such that conditional notification on data updates to subscribers can be done. After being pre-processed, data enters the complex event processing (CEP) [7] component for further analysis, which generally means finding patterns in time windows (temporal reasoning) over the incoming data to form new, more complex events (a.k.a. situations or alerts/warnings). These complex events are emitted to assist in decision-making processes either carried out by humans ("human in the loop" [8]) or automatically by actuators, e.g. sensors that start irrigation in a greenhouse as a result of a certain alert. The situations can also be displayed using visualization tools to assist humans in the decision-making process (as, e.g., in [8]).

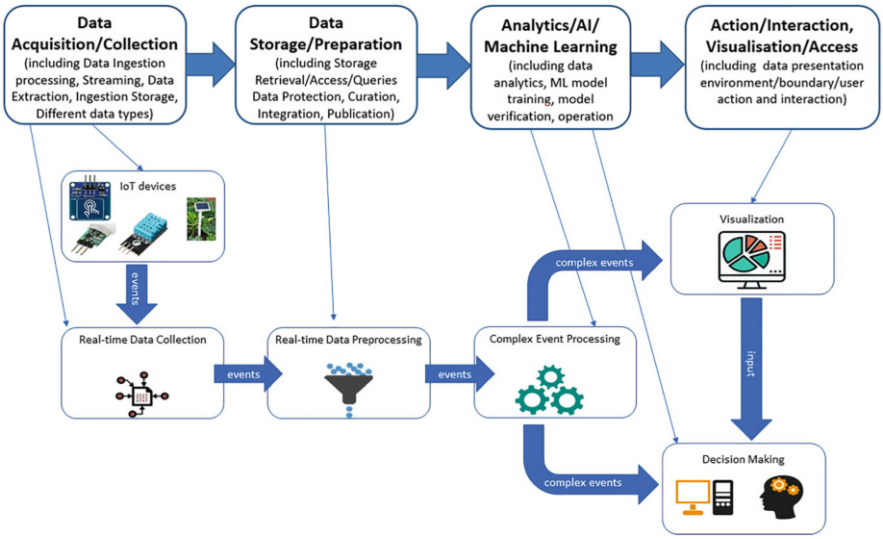

Fig. 5 Mapping of steps of the "Pipeline for IoT data real-time processing and decision making" to the "Generic Big Data and AI Pipeline" steps

The idea is that the detected situations can provide useful real-time insights for operational management (e.g. preventing a possible crop pest or machinery failure).

Figure 5 shows the steps of the pipeline for real-time IoT data processing and decision making that we have just described and their mapping to the steps of top-level Generic Big Data and AI Pipeline pattern that we have analysed in Sect. 2.

Pipeline for Linked Data Integration and Publication

In the DataBio project and some other agri-food projects, Linked Data has been extensively used as a federated layer to support large-scale harmonization and integration of a large variety of data collected from various heterogeneous sources and to provide an integrated view on them. The triplestore populated with Linked Data during the course of DataBio project (and a few other related projects) resulted in creating a repository of over one billion triples, making it one of the largest semantic repositories related to agriculture, as recognized by the EC innovation radar naming it the "Arable Farming Data Integrator for Smart Farming". Additionally, projects like DataBio have also helped in deploying different endpoints providing access to the dynamic data sources in their native format as Linked Data by providing a virtual semantic layer on top of them. This action has been realized in the DataBio project through the implementation of the instantiations of a "Pipeline for the Publication and Integration of Linked Data", which has been applied in different use cases related to the bioeconomy sectors. The main goal of these pipeline instances is to define and deploy *(semi-)automatic processes* to carry out the necessary steps to transform and publish different input datasets for various heterogeneous sources as Linked Data. Hence, they connect different data-processing components to carry out

the transformation of data into RDF format [9] or the translation of queries to/from SPARQL [10] and the native data access interface, plus their linking, as well as the mapping specifications to process the input datasets. Each pipeline instance used in DataBio is configured to support specific input dataset types (same format, model and delivery form).

A high-level view of the end-to-end flow of the generic pipeline and its mapping to the steps of the Generic Big Data and AI Pipeline is depicted in Fig. 6. In general, following the best practices and guidelines of Linked Data Publication [11, 12], the pipeline takes as input selected datasets that are collected from heterogeneous sources (shapefiles, GeoJSON, CSV, relational databases, RESTful APIs), curates and/or pre-processes the datasets when needed, selects and/or creates/extends the vocabularies (e.g., ontologies) for the representation of data in semantic format, processes and transforms the datasets into RDF triples according to underlying ontologies, performs any necessary post-processing operations on the RDF data, identifies links with other datasets and publishes the generated datasets as Linked Data, as well as applies required access control mechanisms.

The transformation process depends on different aspects of the data like the format of the available input data, the purpose (target use case) of the transformation and the volatility of the data (how dynamic is the data). Accordingly, the tools and the methods used to carry out the transformation were determined firstly by the format of the input data. Tools like D2RQ [13] were normally used in the case of data coming from relational databases; tools like GeoTriples [14] was chosen mainly for geospatial data in the form of shapefiles; tools like RML Processor [15] for CSV, JSON, XML data formats; and services like Ephedra [16] (within Metaphactory platform) for Restful APIs.

Pipeline for Earth Observation and Geospatial Data Processing

The pipeline for Earth Observation and Geospatial data processing [17], developed in the DataBio project, depicts the common data flow among six project pilots, four of which are from the agricultural domain and two from the fishery domain. To be more specific, from the agricultural domain there are two smart farming pilots, one agricultural insurance pilot and one pilot that provides support to the farmers related to their obligations introduced by the current Common Agriculture Policy [18]. The two pilots from the fishery domain were in the areas of oceanic tuna fisheries immediate operational choice and oceanic tuna fisheries planning.

Some of the characteristics of this pipeline include the following:

- Its initial data input is georeferenced data [19], which might come from a variety of sources such as satellites, drones or even from manual measurements. In general, this will be represented as either in the form of vector or raster data [20]. Vector data usually describes some spatial features in the form of points, lines or polygons. Raster data, on the other hand, is usually generated from image-producing sources such as Landsat or Copernicus satellites.
- Information exchanged among the different participants in the pipeline can be either in raster or vector form. Actually, it is possible and even common that the

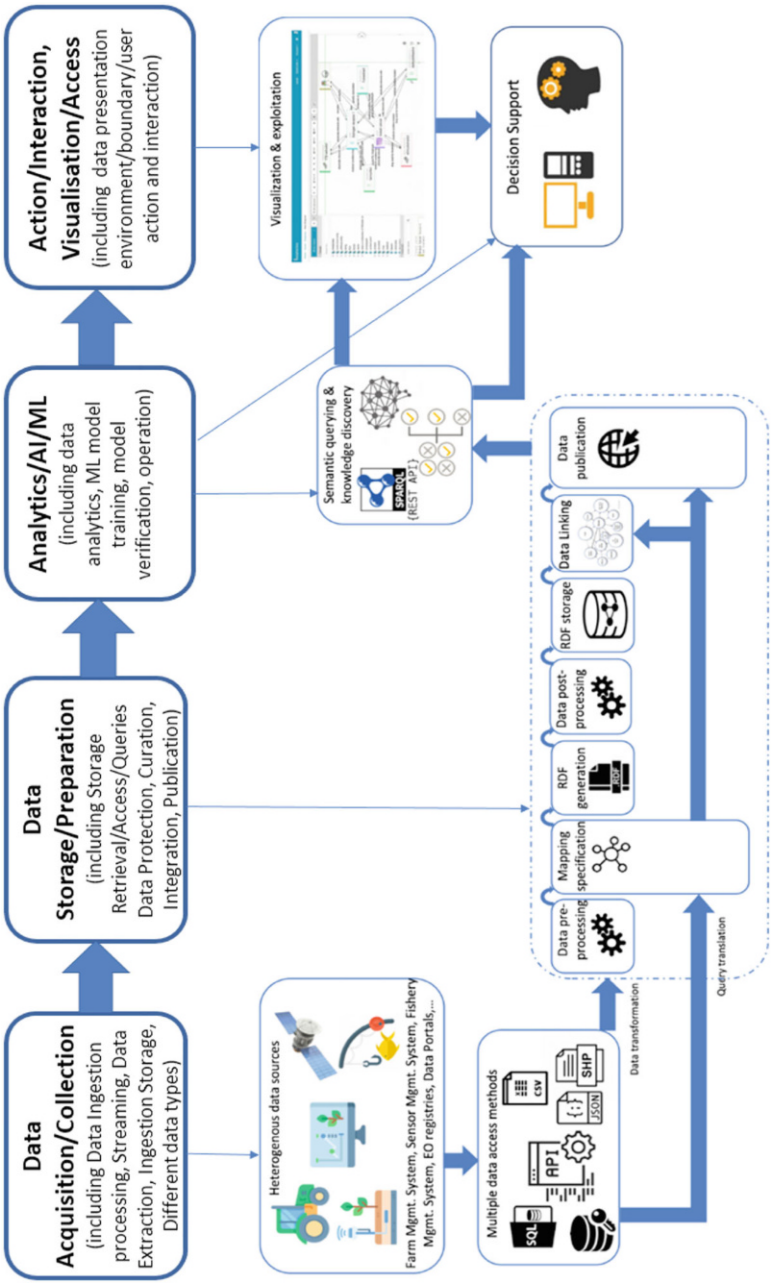

Fig. 6 Mapping of steps of the "Pipeline for linked data integration and publication" to the "Generic big data and AI Pipeline" steps

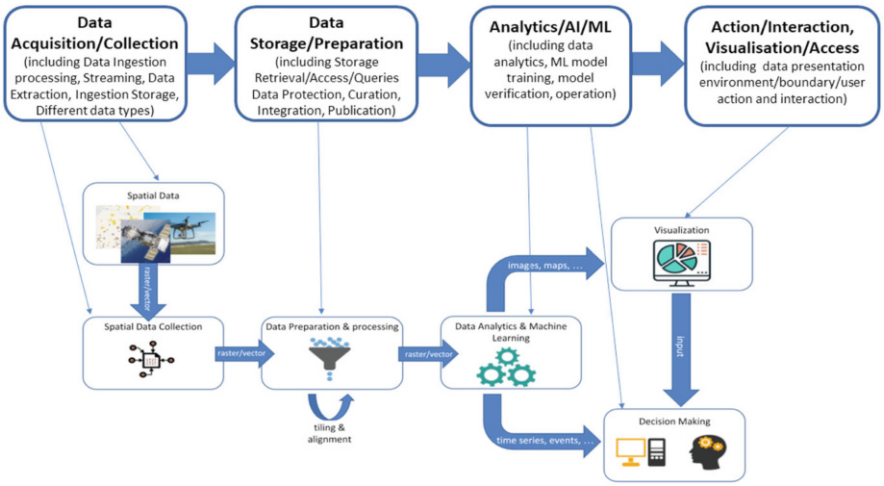

Fig. 7 Mapping of steps of the "Pipeline for earth observation and geospatial data processing" to the "Generic Big Data and AI Pipeline" steps

form of the data will change from one step to another. For example, this can result from feature extraction based on image data or pre-rendering of spatial features.

- For visualization or other types of user interaction options, information can be provided in other forms like: images, maps, spatial features, time series or events.

Therefore, this pipeline can be considered as a specialization of the top-level Generic Big Data and AI Pipeline pattern, presented in Sect. 2, as it concerns the data processing for Earth Observation and Geospatial data. The mapping between the steps of these two pipelines can be seen in Fig. 7.

4 DataBench Pipeline Framework and Blueprints

The top-level Generic Big Data and AI Pipeline pattern discussed in the previous sections has been used as a reference to build architectural blueprints specifying the technical systems/components needed at different stages in a pipeline. For example, in the data acquisition phase of a pipeline, a software broker synchronizing data source and destination is needed. A data acquisition broker will then send data to a lambda function that transforms data in a format that can be stored in a database. In DataBench, we have identified and classified all these technical components with an empirical bottom-up approach as follows: we started from Big Data Analytics (BDA) use cases and then we recognized the commonalities among the technical requirements of the different use cases. Finally, we designed a general architectural blueprint, an overview of which is depicted in Fig. 8. This figure has been detailed in Figs. 9 and 10 for better readability.

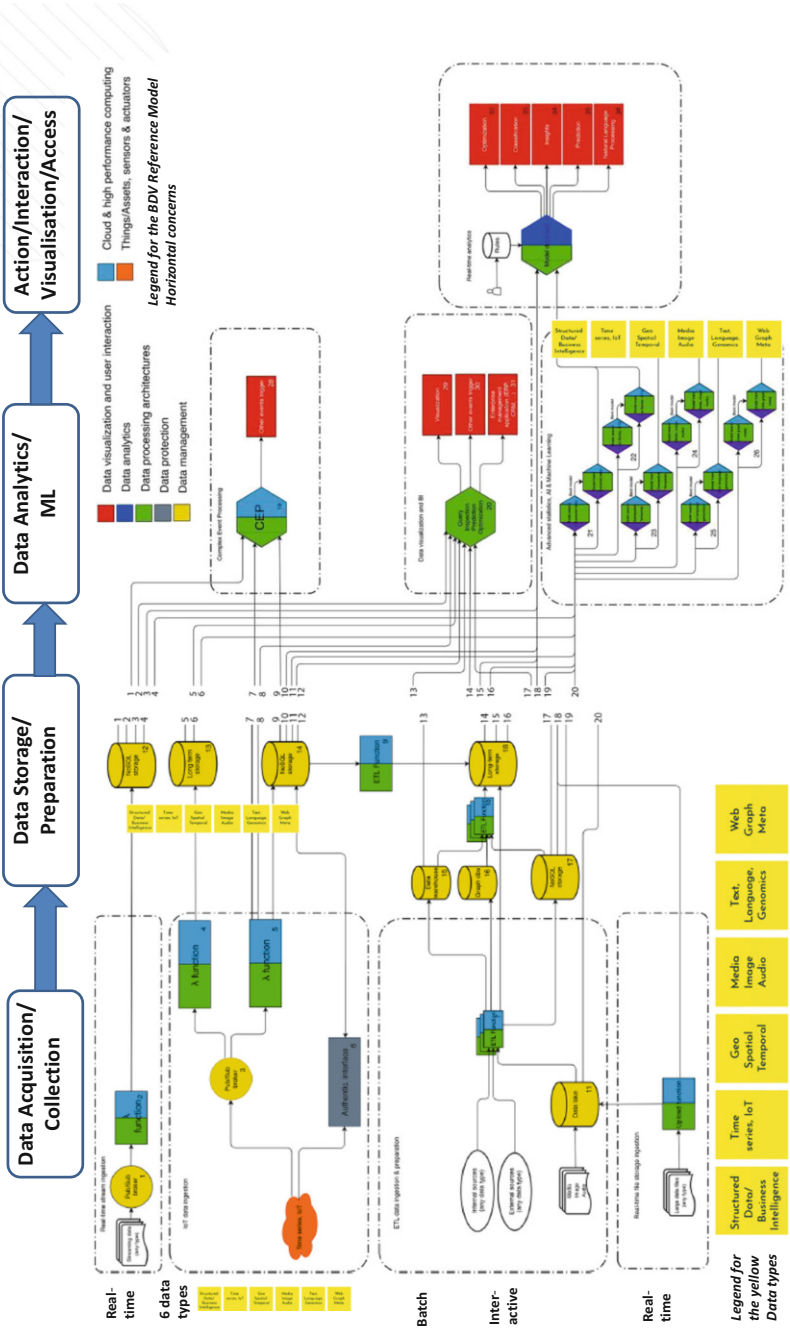

Fig. 8 General architectural blueprint for BDA Pipelines

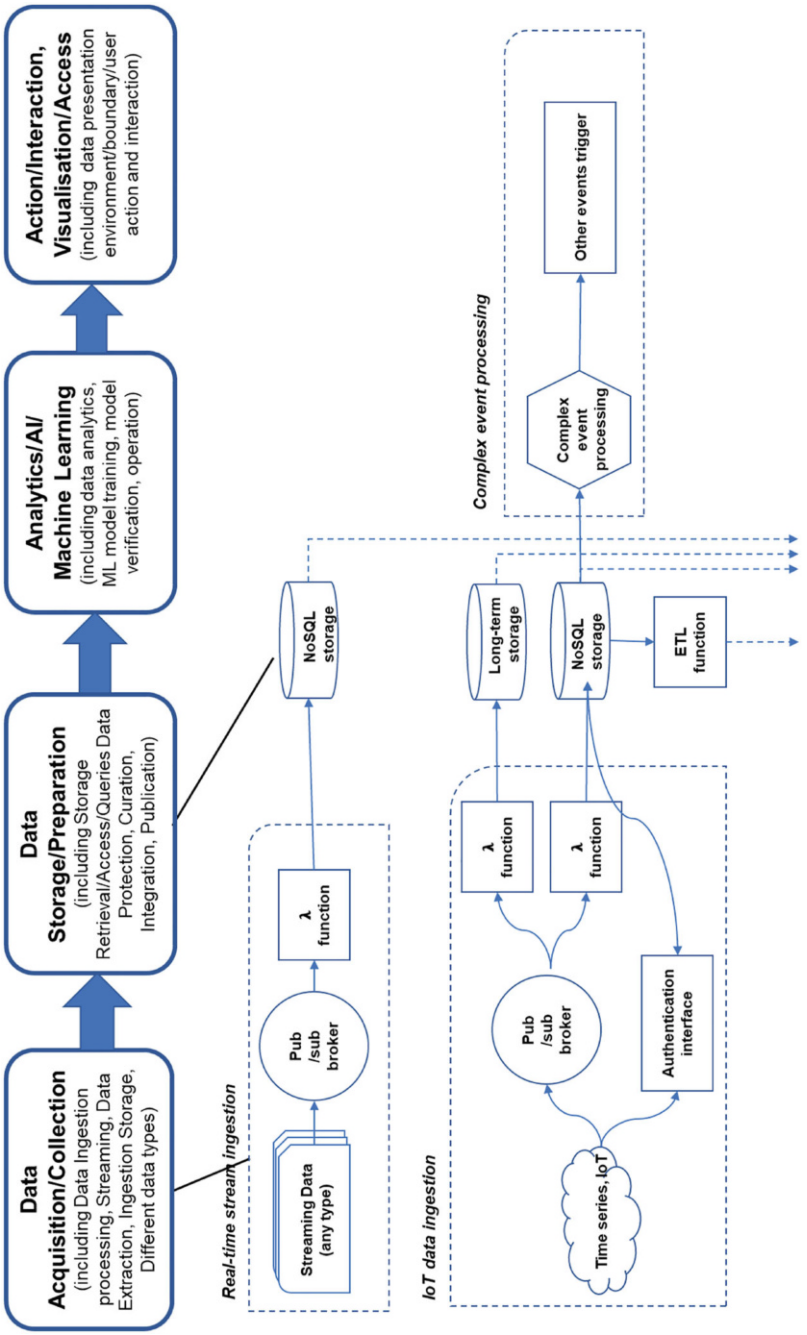

Fig. 9 General architectural blueprint for BDA Pipelines (detail 1)

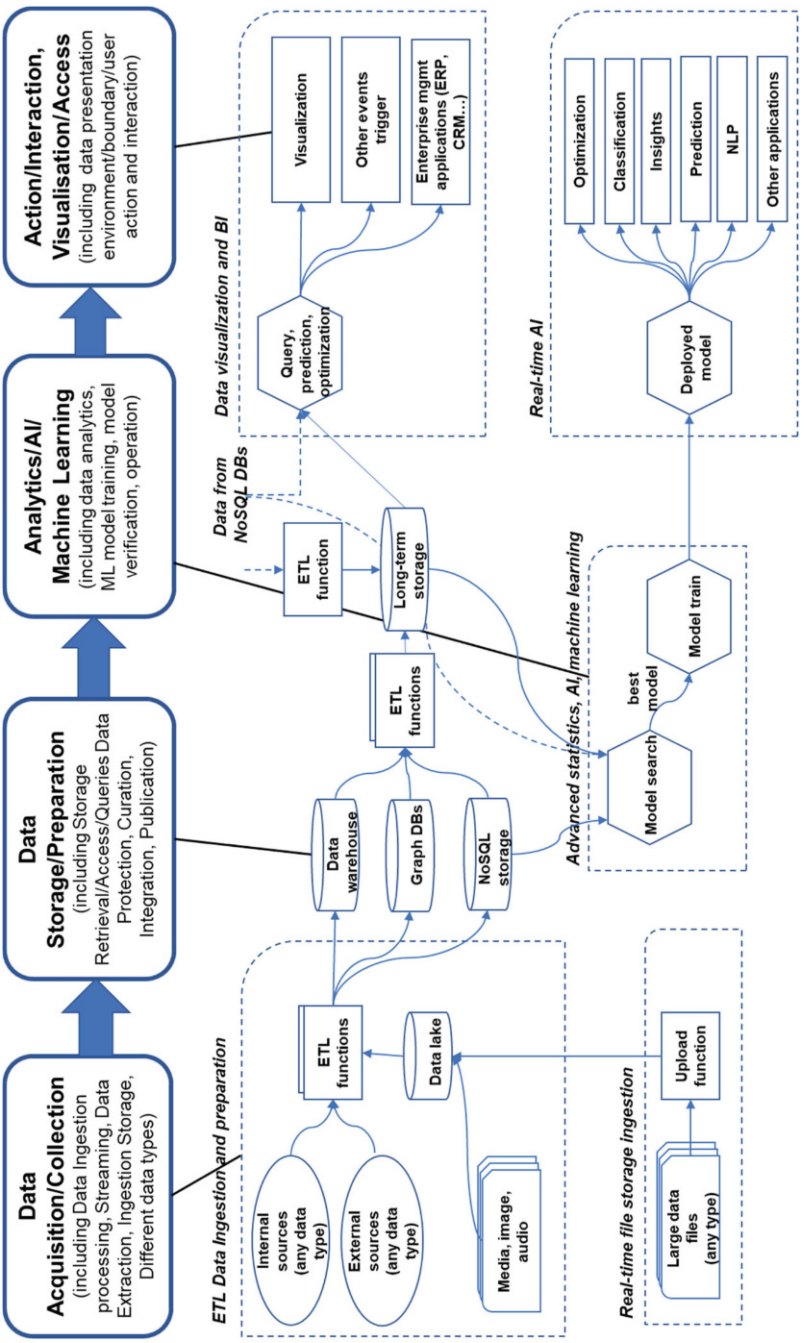

Fig. 10 General architectural blueprint for BDA Pipelines (detail 2)

The general blueprint is consistent with the Big Data Value Association data types classification. This can be seen, e.g., in the "Advanced statistics, AI & Machine Learning" area (see bottom right of Figs. 8 and 10). Specifically, the "Model Search" and "Model Train" architectural components (depicted clearly in Fig. 10) have been replicated for every data type in Fig. 8. The components of the general blueprint can be also seen in the perspective of the horizontal concerns of the BDV Reference Model. Thus, in Fig. 8, we have assigned a specific colour to every horizontal concern of the BDV Reference Model (see *Legend for the BDV Reference Model horizontal concerns*) and each component of the general blueprint has been associated with one or more horizontal concerns by using the respective colours. By mapping the different components of the general blueprint to the horizontal concerns of the BDV Reference Model, we can highlight the interaction among the different Big Data conceptual areas.

We have ensured the generality of this blueprint by addressing the needs of a cross-industry selection of BDA use cases. This selection has been performed based on a European-level large-scale questionnaire (see DataBench deliverables D2.2, D2.3 and D2.4 and desk analyses D4.3, D4.3 and D4.4 in [21]) that have shown the most frequent BDA use cases per industry. We have also conducted an in-depth case study analysis with a restricted sample of companies to understand the functional and technical requirements of each use case. Based on this body of knowledge, we have designed an architectural blueprint for each of the use cases and then inferred the general blueprint which is depicted in the above figure.

We would like to note that the general blueprint can be instantiated to account for the different requirements of different use cases and projects. In DataBench, we have derived use-case-specific blueprints from the general blueprint. The entire collection of use-case-specific blueprints is available from the DataBench Toolbox, as it is discussed in the following section. The Toolbox guides the user from the end-to-end process of the pipelines, the selection of a use case, the specification of technical requirements, down to the selection and benchmarking of specific technologies for the different components of the use-case-specific blueprint.

5 Technical Benchmarks Related to the Big Data and AI Pipeline Framework

As mentioned before, the goal of the DataBench framework is to help practitioners discover and identify the most suitable Big Data and AI technologies and benchmarks for their application architectures and use cases. Based on the BDV Reference Model layers and categories, we initially developed a classification with more than 80 Big Data and AI benchmarks (currently between 1999 and 2020) that we called Benchmark matrix [22]. Then, with the introduction of the DataBench Pipeline Framework, we further extended the benchmark classification to include the pipeline steps and make it easier for practitioners to navigate and search through

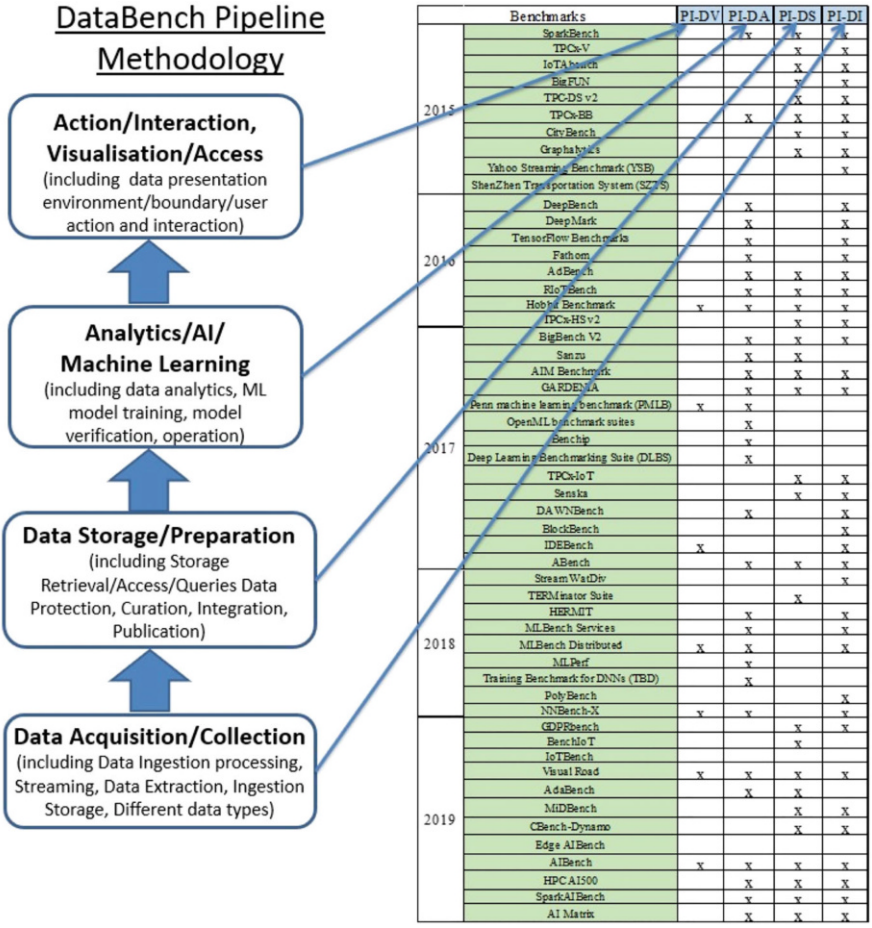

Fig. 11 DataBench Pipeline mapping to benchmarks

the Benchmark matrix. Figure 11 depicts the mapping between the four pipeline steps and the classified benchmarks.

In addition to the mapping of existing technical benchmarks into the four main pipeline steps, there also have been mappings for relevant benchmarks for all of the horizontal and vertical areas of the BDV Reference model. This includes vertical benchmarks following the different data types, such as Structured Data Benchmarks, IoT/Time Series and Stream processing Benchmarks, SpatioTemporal Benchmarks, Media/Image Benchmarks, Text/NLP Benchmarks and Graph/Metadata/Ontology-Based Data Access Benchmarks. It also includes horizontal benchmarks such as benchmarks for Data Visualization (visual analytics), Data Analytics, AI and Machine Learning; Data Protection: Privacy/Security Management Benchmarks

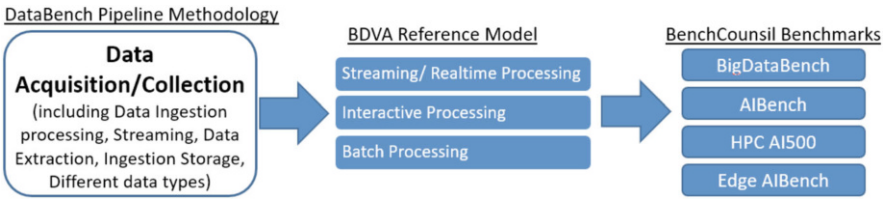

Fig. 12 DataBench Pipeline step mapping to specific category of Benchmarks

related to data management; Data Management: Data Storage and Data Management Benchmarks, Cloud/HPC, Edge and IoT Data Management Benchmarks.

The overall number of technical benchmarks that have been identified and described for these areas are close to 100. All identified benchmarks have been made available through the DataBench Toolbox.

As we can see in Fig. 11, the steps are quite general and map to multiple benchmarks, which is very helpful for beginners that are not familiar with the specific technology types. Similarly, advanced users can go quickly in the pipeline steps and focus on a specific type of technology like batch processing. In this case, focusing on a specific processing category reduces the number of retrieved benchmarks, like in the example in Fig. 12, where only four benchmarks from the BenchCouncil are selected. Then, if further criteria like data type or technology implementation are important, the selection can be quickly reduced to a single benchmark that best suits the practitioner requirements.

The above-described approach for mapping between the DataBench Pipeline Framework and the Benchmark matrix is available in the DataBench Toolbox. The toolbox enables multiple benchmark searches and guidelines via a user-friendly interface that is described in the following sections.

6 DataBench Toolbox

The DataBench Toolbox aims to be a one-stop shop for big data/AI benchmarking; as its name implies, it is not a single tool, but rather a “box of tools”. It serves as an entry point of access to tools and resources on big data and AI benchmarking. The Toolbox is based on existing efforts in the community of big data benchmarking and insights gained about technical and business benchmarks in the scope of the DataBench project. From the technical perspective, the Toolbox provides a web-based interface to search, browse and, in specific cases, deploy big data benchmarking tools, or direct to the appropriate documentation and source code to do so. Moreover, it allows to browse information related to big data and AI use cases, lessons learned, business KPIs in different sectors of application, architectural blueprints of reference and many other aspects related to benchmarking big data and

AI from a business perspective. The Toolbox provides access via a web interface to this knowledge base encapsulated in what is called "knowledge nuggets".

The main building blocks of the Toolbox are depicted in Fig. 13 and comprise a front-end DataBench Toolbox Web user interface, Toolbox Catalogues and the Toolbox Benchmarking Automation Framework, which serves as a bridge to the Execution of Benchmarks building block located outside the Toolbox.

The intended users of the Toolbox are technical users, business users, benchmarking providers, and administrators. The support and benefits for each type of user is highlighted in their dedicated user journeys accessible from the Toolbox front-page, except for administrators, who are needed to support all kinds of users and facilitate the aggregation and curation of content to the tool. The Toolbox Catalogues building block shown in Fig. 13 comprises the backend functionality and repositories associated with the management, search and browsing of knowledge nuggets and benchmarking tools. The Toolbox Benchmarking Automation Framework building block serves as a bridge to the Execution of Benchmarks building block located in the infrastructure provided by the user outside the Toolbox (in-house or in the cloud), as the Toolbox does not provide a playground to deploy and execute benchmarks. The automation of the deployment and execution of the benchmarks is achieved via the generation of Ansible Playbooks [23] and enabled by an AWX project [24] for process automation. The steps to be followed by a Benchmark Provider with the help of the Administrator to design and prepare the benchmark with the necessary playbooks for the automation from the Toolbox are described in detail in Sect. 3.1 "Support for adding and configuring benchmarks" of DataBench Deliverable D3.4, which can be found in [21]. Last but not least, the DataBench Toolbox Web building block is the main entry point for the users.

The DataBench Toolbox Web user interface is publicly available for searching and browsing, although some of the options are only available to registered users. Registering to the Toolbox is free and can be done from the front-page. This web user interface is accessible via https://toolbox.databench.eu/. Via this interface, a user can access (1) the Big Data Benchmarking Tools Catalogue (2) the Knowledge Nuggets Catalogue; (3) User journeys which provide a set of tips and advice to different categories of users on how to use and navigate throughout the Toolbox; (4) links to other tools such as the DataBench Observatory explained below in this chapter; and (5) search features. The Toolbox provides several search options, including searching via a clickable representation of the BDV Reference Model and via clickable depiction of the big data architectural blueprints and the generic pipeline presented in Sect. 2. The latter type of searching, depicted in Fig. 14, enables accessing technical benchmarks as well as nuggets related to the clicked elements.

As we mentioned before, one of the tools accessible from the Toolbox is the DataBench Observatory. This is a tool for observing the popularity, importance and the visibility of topic terms related to Artificial Intelligence and Big Data, with particular attention dedicated to the concepts, methods, tools and technologies in the area of Benchmarking.

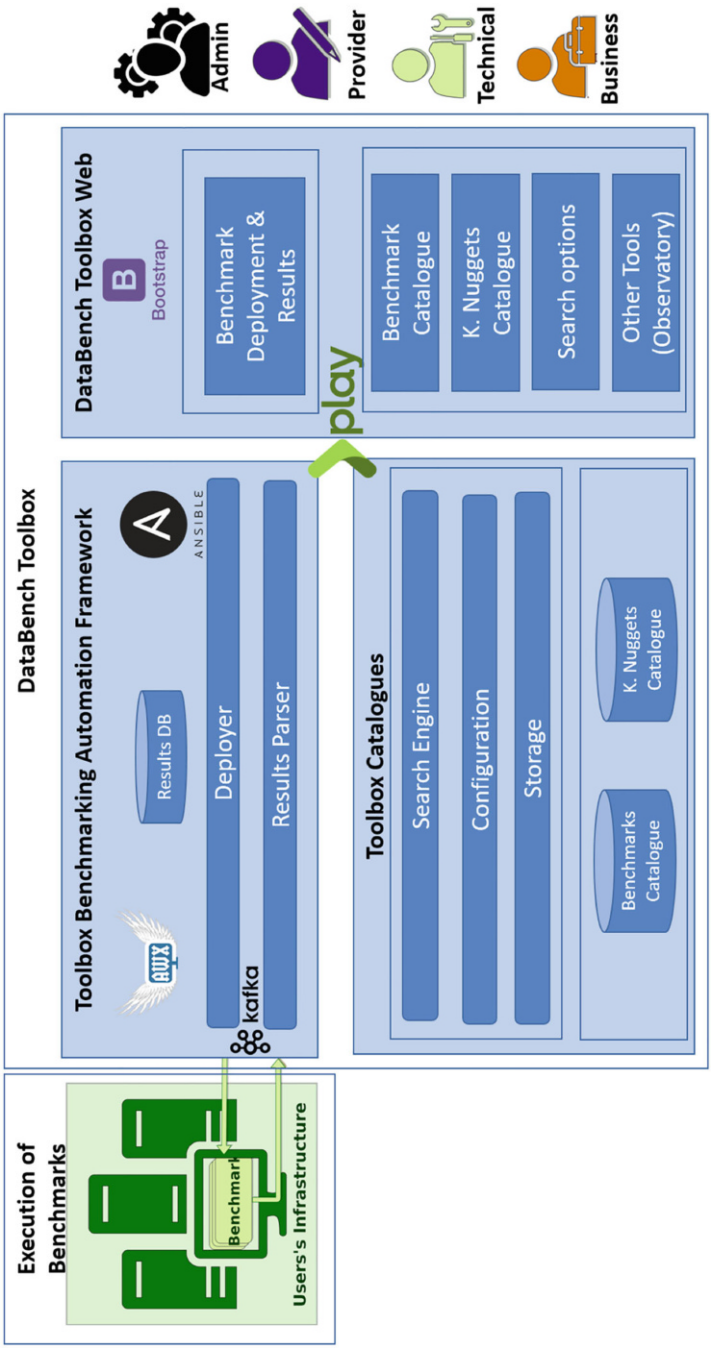

Fig. 13 DataBench Toolbox functional architecture

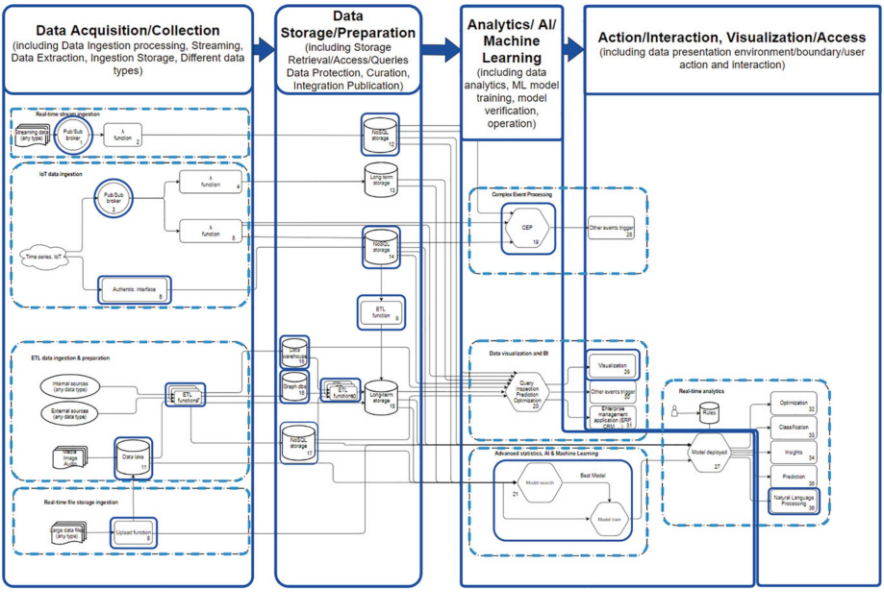

Fig. 14 Search by Pipeline/Blueprint available from the DataBench Toolbox

The DataBench Observatory introduces the popularity index, calculated for ranking the topic terms in time, which is based on the following components: (1) Research component, such as articles from the Microsoft Academic Graph (MAG) [25]; (2) Industry component, such as job advertisements from Adzuna service [26]; (3) Research and Development component, such as EU research projects, e.g. in CORDIS [27] dataset (4) Media component, such as cross-lingual news data from the Event Registry system [28]; (5) Technical Development component, such as projects on GitHub [29]; and (6) General Interest, such as Google Trends. The DataBench observatory provides ranking and trending functionalities, including overall and monthly ranking of topics, tools and technologies, as well as customized trending options. See, e.g., Fig. 15 that demonstrates that Microsoft tools are highly requested in job postings, Python is one of the most popular languages at GitHub (as it is also mentioned in [30, 31]) and users on the web search a lot for Node.js solutions. It is possible here to search for the popularity of various Big Data and AI tools and also for Benchmarks.

See also Fig. 16, which shows time series for topics from the areas of Artificial Intelligence, Big Data and Benchmarking. Users interested in the "Artificial Intelligence" topic can observe its high popularity (score 10 is the maximum normalized popularity) within academic papers.

In the context of the DataBench Toolbox, the DataBench Observatory is targeted at different user groups, such as industrial and business users, academic users, general public, etc. Each type of user can use the Observatory to explore popular topics as well as tools and technologies in areas of their interest.

month: All Topics: All

Search:

Topic	Categories	Papers	EU Projects	News	Github	Jobs	Search Volume	Total
Microsoft	infrastructure,computer_science	1.72	1.13	5.24	1.3	10	6.34	4.29
Google	infrastructure,computer_science	3.74	2.23	8.21	2.76	1.27	7.38	4.26
Amplitude	customer,computer_science	5	10	1.03	1.04	1	7.11	4.2
Vector	log,computer_science	10	1	1.19	2.32	1	9.35	4.14
Python	stat_tool,computer_science	1.81	1.06	1.2	10	1	8.96	4.01
Node	b2b_marketing,computer_science	8.71	1	1.17	1.62	1	10	3.92
ARM	hardware,computer_science	10	1	1.03	1.13	1	9.37	3.92

Fig. 15 DataBench popularity index (Tools and Technologies, accessed in November 2020)

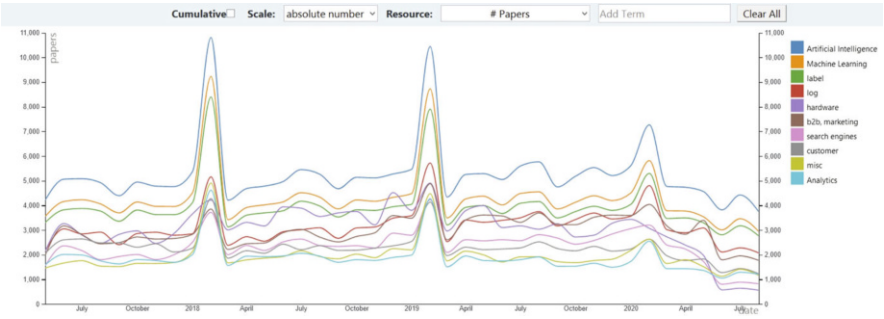

Fig. 16 Time series (Topics, accessed in November 2020)

Furthermore, in order to develop the DataBench observatory tool, we have composed a DataBench ontology based on Artificial Intelligence, Big Data, Benchmarking-related topics from Microsoft Academic Graph and extended/populated the ontology with tools and technologies from the relevant areas, by categories. Microsoft Academic Graph (MAG) taxonomy has been expanded with DataBench terms—over 1700 tools and technologies related to Benchmarking, Big Data, and Artificial Intelligence. New concepts have been aligned with MAG topic, MAG keyword, Wikipedia (for analysis in wikification) and Event Registry concepts. The DataBench ontology is used in the semantic annotation of the unstructured textual information from the available data sources. Figure 17 illustrates the popular tools and technologies in the Graph databases category, sorted by popularity for GitHub data source.

7 Conclusions

This chapter has presented a Big Data and AI Pipeline Framework developed in the DataBench project, supported by the DataBench Toolbox. The Framework contains a number of dimensions, including pipelines steps, data processing types and types

month: All Topics: graph_database

Search:

Topic	Papers	EU Projects	News	Github	Jobs	Search Volume	Total
Neo4j	5.95	8.5	10	10	10	10	9.07
Grakn	1	1	1	3.1	1	5	2.02
ArangoDB	1.12	1	1.73	2.5	1	7.43	2.46
OrientDB	1.18	1	1.96	1.6	1	6.41	2.19
GraphDB	1.18	1	1	1.6	1	6.11	1.98
Virtuoso	10	10	1	1	1	8.43	5.24
Microsoft Azure Cosmos DB	1.02	1	8.58	1	1	3.92	2.75
Graph Engine	2.13	1	2.59	1	1	7.51	2.54
TigerGraph	1.07	1	4.01	1	1	4.24	2.05
Dgraph	1	1	1.48	1	1	5.52	1.83
JanusGraph	1.02	1	1	1	1	5.89	1.82

Fig. 17 DataBench popularity index (Tools and Technologies, Category: Graph Database, sorted by GitHub data source, accessed in November 2020)

of different data. The relationship of the Framework is with existing and emerging Big Data and AI reference models such as the BDV Reference Model and the AI PPP, and also the ISO SC42 Big Data Reference Architecture (ISO 20547) [3] and the emerging AI Machine Learning Framework (ISO 23053) [6], with which the pipeline steps also have been harmonized.

Further work is now related to populating the DataBench Toolbox with additional examples of actual Big Data and AI Pipelines realized by different projects, and further updates from existing and emerging technical benchmarks.

The DataBench Toolbox observatory will continuously collect and update popularity indexes for benchmarks and tools. The aim for the DataBench Toolbox is to be helpful for the planning and execution of future Big Data and AI-oriented projects, and to serve as a source for the identification and use of relevant technical benchmarks, also including links to a business perspective for applications through identified business KPIs and business benchmarks.

Acknowledgements The research presented in this chapter was undertaken in the framework of the DataBench project.

"Evidence Based Big Data Benchmarking to Improve Business Performance" [32] funded by the Horizon 2020 Programme under Grant Agreement 780966.

References

1. Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Big Data Value Association.
2. Zillner, S., Bisset, D., Milano, M., Curry, E., García Robles, A., Hahn, T., Irgens, M., Lafrenz, R., Liepert, B., O'Sullivan, B., & Smeulders, A., (Eds.). (2020). Strategic research, innovation and deployment agenda – AI, data and robotics partnership. Third Release. "September 2020, Brussels. BDVA, euRobotics, ELLIS, EurAI and CLAIRE".
3. ISO/IEC 20547–3:2020, Information technology—Big data reference architecture—Part 3: Reference architecture, https://www.iso.org/standard/71277.html
4. https://www.bdva.eu/PPP

5. Södergård, C., Mildorf, T., Habyarimana, E., Berre, A. J., Fernandes, J. A., & Zinke-Wehlmann, C. (Eds.). *Big data in bioeconomy results from the European DataBio project*.
6. ISO/IEC CD 23053.2: 2020 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).
7. https://hazelcast.com/glossary/complex-event-processing/
8. https://www.sciencedirect.com/science/article/pii/S1364815214002618
9. Wood, D., Lanthaler, M., & Cyganiak, R. (2014). RDF 1.1 concepts and abstract syntax [W3C Recommendation]. (Technical report, W3C).
10. Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. W3C recommendation. W3C.
11. Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). Best practices for publishing linked data. W3C working group note 09 January 2014. https://www.w3.org/TR/ld-bp/
12. Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space (1st ed.). *Synthesis lectures on the semantic web: Theory and technology*, 1, 1–136. Morgan & Claypool.
13. http://d2rq.org/
14. http://geotriples.di.uoa.gr/
15. https://github.com/IDLabResearch/RMLProcessor
16. https://www.metaphacts.com/ephedra
17. https://www.europeandataportal.eu/en/highlights/geospatial-and-earth-observation-data
18. https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy_en
19. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/geo-referenced-data
20. https://gisgeography.com/spatial-data-types-vector-raster/
21. https://www.databench.eu/public-deliverables/
22. http://databench.ijs.si/knowledgeNugget/nugget/53
23. https://docs.ansible.com/ansible/2.3/playbooks.html
24. https://www.ansible.com/products/awx-project
25. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph
26. https://www.adzuna.co.uk
27. https://data.europa.eu/euodp/sl/data/dataset/cordisH2020projects
28. https://eventregistry.org
29. https://github.com
30. https://octoverse.github.com/
31. 2020 StackOverflow Developers Survey. https://insights.stackoverflow.com/survey/2020#technology-programming-scripting-and-markup-languages
32. https://www.databio.eu/en/