DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D5.4 Analytic modelling relationships between metrics, data and project methodologies

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. This document provides a framework and methodology for the usage of the DataBench Toolbox with an associated validation and assessment approach as a basis for further sustainable evolution of the Toolbox. The document presents a methodology with for how projects can relate to criteria and metrics for the description of system, tool and benchmark features as they are related to collected data about this for the identification of project relevant tools and benchmarks. The document starts with a description of the Big Data and AI Pipeline Framework and Methodology. The document introduces 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 and tools. The framework also serves as a basis for demonstrating the similarities among Big Data Projects. 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 Visualisation and User Interaction as well as API Access. This is illustrated with Big Data and AI Pipeline examples and is further supported with more detailed technical Blueprints which present categorisations of architectural blueprints for realisations of the various steps of the Big Data and AI pipeline. Technical Benchmarks are being related to the Big Data and AI Pipeline Framework, and also related to Components, Tools and standards The DataBench Toolbox 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 supports observing the popularity, importance and the visibility of technologies with selected terms, here applied for the popularity of Big Data and AI tools and relevant benchmarks.

Deliverable D5.4	**Analytic modelling relationships between metrics, data and project methodologies**
Work package	WP5
Task	5.3
Due date	31/12/2020
Submission date	22/12/2020
Deliverable lead	SINTEF
Version	2.0
Authors	SINTEF (Arne J. Berre, Aphrodite Tsalgatidou, Brian Elvesæter) Lead Consult (Todor Ivanov) POLIMI (Chiara Francalanci, Giulio Costa, Gianmarco Ruggiero) JSI (Inna Novalija, Marko Grobelnik, Besher M.Massri) ATOS (Tomás Pariente, Iván Martínez and Ricardo Ruiz)
Reviewers	Ricardo Ruiz, Federica Acerbi

Keywords
Benchmarking, Big Data, AI, Pipeline, Framework, Blueprint, Methodology, Toolbox, Observatory

Disclaimer

Copyright Notice

Table of Contents

List of Figures

List of Tables

Executive Summary

The objectives of the work in workpackage 5 "Technical Evaluation using the DataBench Toolbox", is to provide a framework and methodology for the usage of the DataBench Toolbox with an associated validation and assessment approach as a basis for further sustainable evolution of the DataBench Toolbox. The objective of this document, D5.4: "Analytic modelling relationships between metrics, data and project methodologies" is to describe the usage of relevant feature criteria and metrics associated the identification of relevant Big data and AI benchmarks and tools. 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. A high-level usage patterns for this has been identified through a set of pipeline steps. The document starts with a description of the Big Data and AI Pipeline Framework and Methodology. In this 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 and tools. The framework also serves as a basis for demonstrating the similarities among Big Data Projects. 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 Visualisation and User Interaction as well as API Access. Examples Big Data and AI Pipeline steps are illustrated in particular for the Big Data types of IoT, Graph and SpatioTemporal data. The DataBench Pipeline Framework is further associated with more detailed technical Blueprints which present categorisations of architectural blueprints for realisations of the various steps of the Big Data and AI pipeline with variations depending on the processing types and the main data types involved. Some examples of relevant blueprints are described further in Annex A. Technical Benchmarks are being related to the Big Data and AI Pipeline Framework, and also related to Components, Tools and standards with feature criteria for various technical areas. Existing Big Data and AI Technical Benchmarks have been classified according to the Big Data and AI Pipeline Framework that is presented. These are benchmarks that are suitable for benchmarking of technologies related to the different parts of the pipeline and associated technical areas. The DataBench Toolbox supports the identification and use of existing benchmarks according to these steps in addition to all the different technical areas and different data types in the BDV Reference Model. The DataBench Observatory is accessed via the toolbox, as a tool for observing the popularity, importance and the visibility of Big Data and AI technologies and benchmarks. The conclusions also present future evolution and sustainability plans, followed by an Annex on Use case independent Blueprints.

1. Introduction

1.1. Objective

The objectives of WP5 " Technical Evaluation using the DataBench Toolbox " is to provide a framework and methodology for the usage of the DataBench Toolbox with an associated validation and assessment approach as a basis for further sustainable evolution of the DataBench Toolbox.

The WP5 workpackage is using the results from WP1, 2 and 3 and the result of this workpackage will be the technical validation of the DataBench framework from WP1 and 2 and the Toolbox from WP3 – both with possible extensions based on the usage and feedback from actual projects. This includes how to validate and assess the correspondence of the technical metrics and the resulting benchmarks collected and refined in WP1 and 2 and integrated in the Toolbox developed in WP3, to make sure that they effectively correspond to the intentions of the original tools and needs of the project communities.

The initial WP5 deliverables D5.1 "Initial Evaluation of DataBench Metrics", D5.2 "Final evaluation of DataBench metrics" and D5.3 "Assessment of technical usability, relevance, scale and complexity" provide the input for this deliverable D5.4 “Analytic modelling relationships between metrics, data and project methodologies". D5.4 provides the methodology and setup which is being supported by the DataBench Toolbox and which is further validated in the deliverable D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios".

In this work, we address sustainability of the DataBench Toolbox. The developed methodology has been refined, adjusted and put in the form to be reusable by external users and suitable for continuous monitoring of the Big Data projects in the future, as part of the DataBench Toolbox.
The objective of this document, D5.4: "Analytic modelling relationships between metrics, data and project methodologies” is to describe the usage of the feature metrics from D5.2. This is for Benchmarks and tools in conjunction with steps of a project methodology that provides support for the selection of benchmarks suitable for the kind of tools and components that are relevant for use in the different steps of a Big Data and AI Pipeline, and also to support the identification of related tools and components and possible standards. This is based on data collected from various existing benchmark analysis, as well as for data collected about available tools in various categories and on continuous input on pipelines and results from relevant projects.

The complementary deliverable D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios" will describe how current projects, and in particular Big Data PPP projects, has related to the methodology and use of the DataBench toolbox.

Organisations rely on evidence from the benchmarking domain to provide answers to how their processes are performing. There is extensive information on how and why 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 organisation 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.

In this deliverable, we focus on the DataBench Methodology approach for Technical Benchmarks which are using a Big Data and AI pipeline model as an overall framework. Technical areas are further classified depending on the various areas of the BDVA 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 [2].

The DataBench framework is accompanied by a Handbook (D4.4) and the DataBench 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 (i) to identify the use cases where they can achieve the highest possible business benefit and return on investment, so they can prioritize their investments; (ii) to select the best technical benchmark to measure the performance of the technical solution of their choice; and, (iii) 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 users' journey (before, during and in the ex-post evaluation of their Big Data and AI technology investment) and from both the technical and business viewpoint.

1.2. Structure of the Report

The report is structured as follows:

- Chapter 1 Introduction outlines the main objectives and the structure of the report.
- Chapter 2 describes the Big Data and AI Pipeline Framework and Methodology. In this 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 and tools. 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 [3]. We also discuss the relation to the BDVA reference model from the BDVA Strategic Research and Innovation Agenda (SRIA) [1] and the relation to the areas in the Strategic Research, Innovation and Deployment Agenda (SRIDA) for a European AI, Data and Robotics Partnership (AI PPP SRIDA) [4].

- Chapter 3 shows Big Data and AI Pipeline Examples in particular for the Big Data types of IoT, Graph and SpatioTemporal data based on Big Data and AI Pipeline Examples from the DataBio project [5].
- Chapter 4 describes the DataBench Pipeline Framework associated with more detailed technical Blueprints – which present categorisations 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). Specialisations can also be more complex aggregations/compositions of multiple specialisations/patterns. These blueprints are a basis for selecting specialisations of the pipeline that will fit with the needs of various projects and instantiations.
- Chapter 5 describes how the DataBench Toolbox is supporting the Methodology. This chapter describes the architecture of the DataBench Toolbox and the functionalities that it supports.
- Chapter 6 presents Technical Benchmarks – related to the Big Data and AI Pipeline Framework, and also related to Components, Tools and standards. The chapter presents how existing Big Data and AI Technical Benchmarks have been classified according to the Big Data and AI Pipeline Framework that has been presented in chapter 3. These are benchmarks that are suitable for benchmarking of technologies related to the different parts of the pipeline and associated technical areas.
- Chapter 7 explains how the DataBench Toolbox Observatory – for Benchmarks and Tools 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.
- Chapter 8 presents conclusions and future evolution and sustainability plans, followed by an Annex A on Use case independent Blueprints - Finally, the conclusions in section 7 present a summary of the contributions and the plans for further evolution and usage of the DataBench Toolbox.

2. The Big Data and AI Pipeline Framework Methodology

The DataBench Framework for Big Data and AI Benchmarks is based on Big Data Value Association (BDVA) reference architecture. In order to have an overall perspective on Big Data and AI systems, the usage of a top-level generic pipeline has recently been introduced by the DataBench project as a complementary data and control flow perspective for the description and analysis of technologies used in the context of a Big Data and AI Application.

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

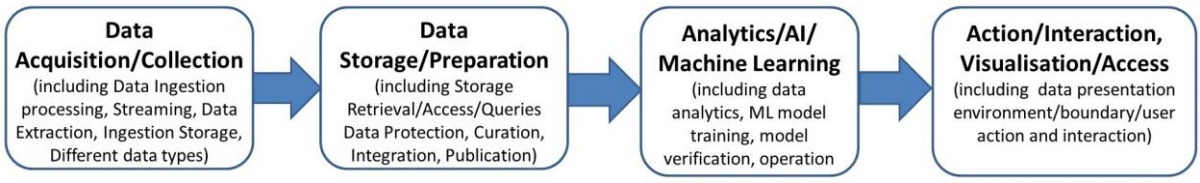

Figure 1: Top level Generic Big Data and AI Pipeline pattern

As it can be seen in Figure 1, this pipeline is quite high level. Therefore, it can be easily specialised 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 Figure 2 depicts the steps of this pipeline in relationship with the type of data processing and the type of data being processed.

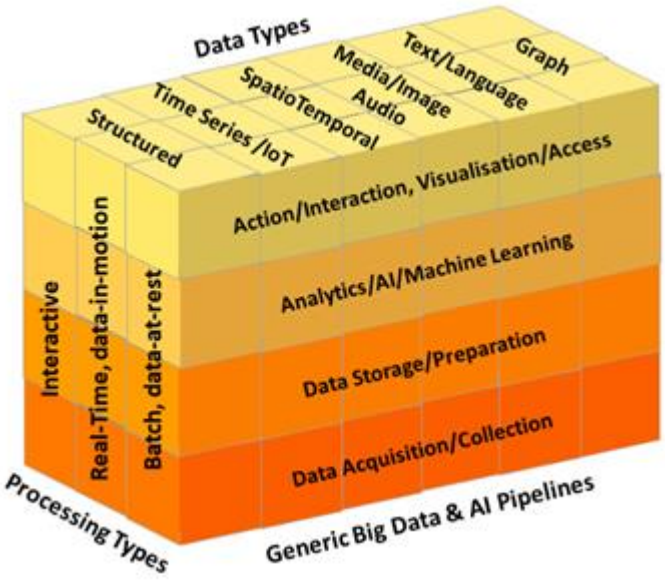

Figure 2: Top level Generic Big Data and AI Pipeline cube

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 also be considered for the area of Data Protection, with Privacy and anonymisation mechanisms to facilitate data protection. It also has links to trust mechanisms like Blockchain technologies, smart contracts and various forms for encryption. This area is also associated with the area of CyberSecurity, Risk and Trust.

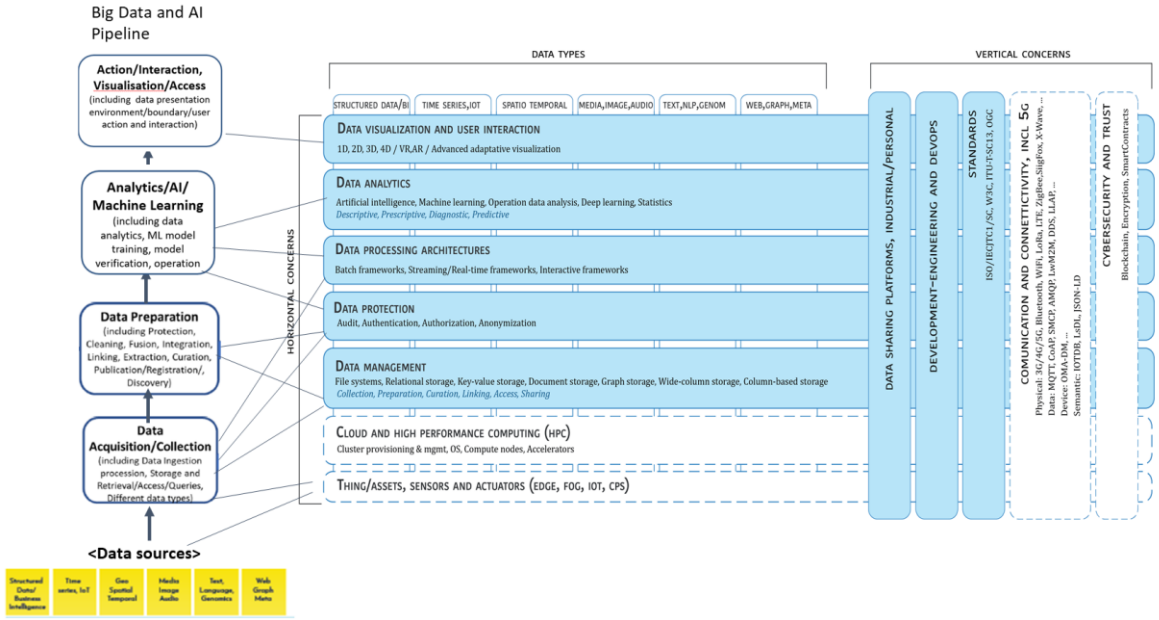

Figure 3: Big Data and AI Pipeline using technologies from the BDV reference model

The BDV Reference Model shown in Figure 3 has been developed by the BDVA, 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 showed to the left of the BDV Reference model in Figure 3, with lines related 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 and also 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, Visualisation and Access

This step (including data presentation environment/boundary/user action and interaction) identifies the boundary towards the environment for action/interaction, typically through a visual interface with various data visualisation 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 specialised 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 will be a cycle starting from training data and later using operational data. The steps of the Big Data and AI Pipeline Framework are also harmonised with the ISO SC42 AI Committee standards [8]. It is in particular harmonised with the steps of 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 [2]. The following figure shows how the Big Data and AI Pipeline can also be related to the recent AI PPP Ecosystem and Enablers (from SRIDA AI).

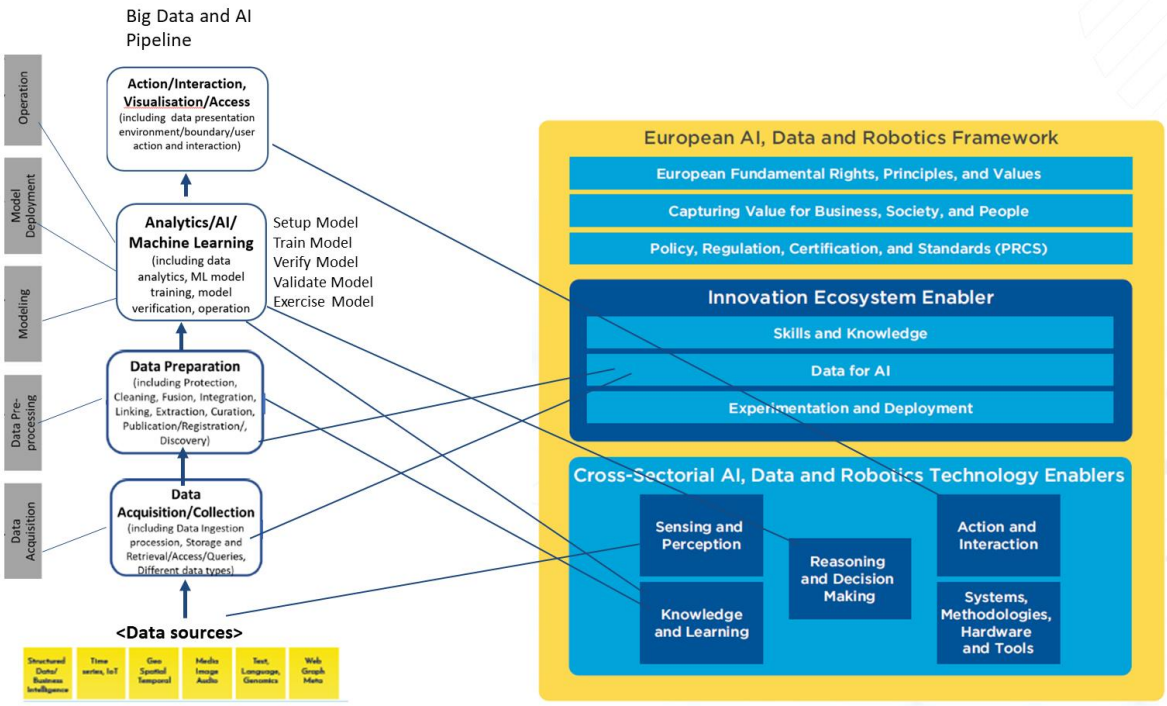

Figure 4: Big Data and AI Pipeline and the European AI and Robotics Framework

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, synthesizing 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 optimisation, search, planning, diagnosis and relies on methods to ensure robustness and trustworthiness.

Action/Interaction, Visualisation 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.

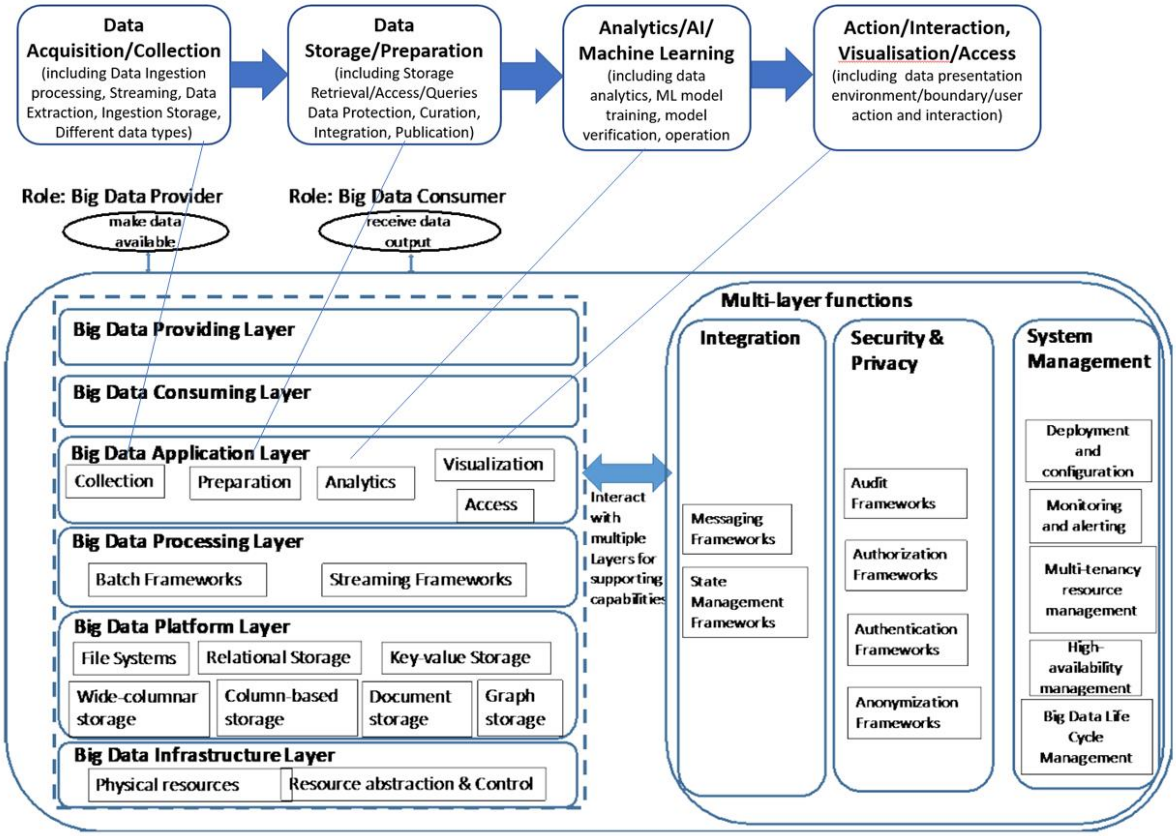

Figure 5: Big Data and AI Pipeline and the ISO 20547-3 Big Data Reference Architecture.

The steps of the Big Data and AI Pipeline Framework are also harmonized with the ISO SC42 AI Committee standards [8]. It is in particular harmonized with the steps of 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 [2].

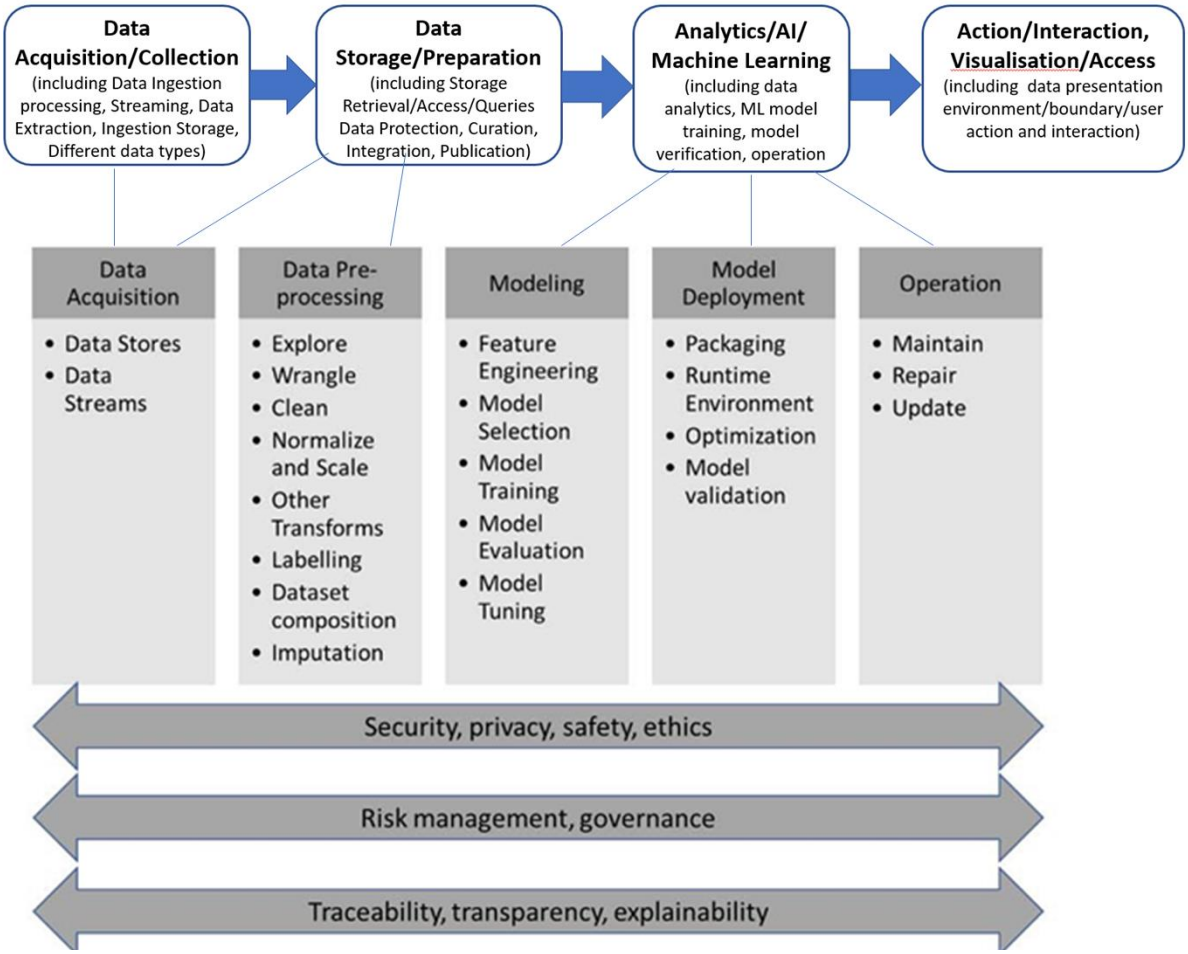

Figure 6: Big Data and AI Pipeline and the steps in ISO/IEC 23053 standard

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

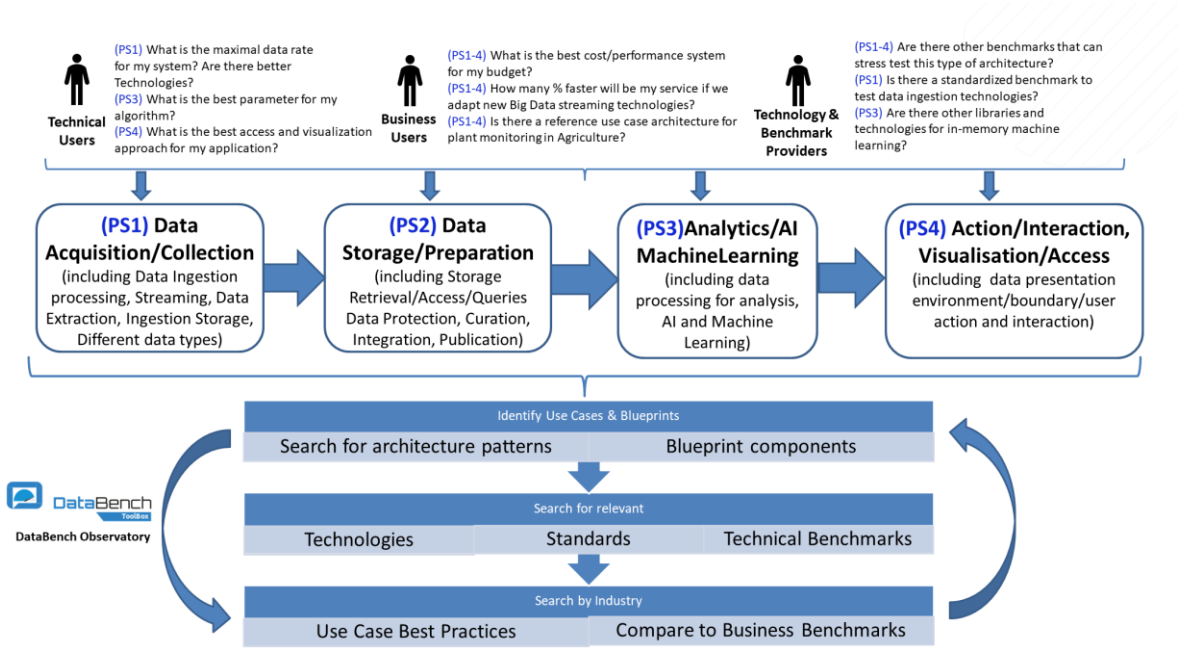

Figure 7: Big Data and AI steps related to blueprints, benchmarks, technologies, standards

Figure 7: Big Data and AI steps related to blueprints, benchmarks, technologies, standards – shows how the DataBench Toolbox can be used with the perspective of each of the pipeline steps described earlier, with additional selection criteria that can be chosen by different kinds of users. Technical Users can search based platform and architecture features. Business Users can search based on business and application oriented features. Benchmark and Technology Providers can search for benchmark-specific features. They can all also combine features including use of toolbox-specific features

The selection criteria for the identification of relevant benchmarks, technologies, blueprints is supported by searches and selection using the following different kinds of Features with their associated qualitative selection criteria and metrics as described further in D5.2 "Final Evaluation of DataBench Metrics" [6]:

Business Features such as Application Area, Business Goals, Business KPIs, Company Size, Industry, Level of BDA solutions maturity and Level of business process integration.

Application Features such as BDV Reference Model Horizontal and Vertical Layer elements, Analytics type, Data size, Data type(s), Machine Learning approach and Workload type.

Platform and Architecture Features such as Architecture patterns, Platform type, Platform-level performance, Processing types and Storage/Database types.

Benchmark-specific Features such as Benchmark synthetic/real data type, Benchmark type, Benchmarking aspect, Benchmarking performance metrics, Execution environment, Input data format, Output data format.

Toolbox-specific Features such as Knowledge nugget classification, Pipeline steps, Technology category, Type of user and User journey.

2.1. DataBench Big Data and AI Toolbox project methodology

In the following we outline the steps of the suggested DataBench Big Data and AI Toolbox project methodology for the identification of relevant Big Data and AI pipelines – relating to feature metrics and collected data about benchmarks, tools and standards.

1. Identify Application area and domain (D4.4, D4.3)
2. Identify Business Objectives/KPIs - do Self-assessment Survey (D4.4, D4.3)
3. Identify Technical user, Business user, Benchmark/Tech provider perspective (D5.4)
4. Identify key technical objectives (BDV Reference model area scope)
5. Identify and map to the 4 pipeline steps
6. Identify relevant Big data types and processing types + architectural patterns
7. Identify any relevant Blueprints - use case independent and domain / use-case specific
8. Identify relevant standards for big data and AI technologies (ISO SC42) (optimal) - for future
9. Identify relevant technologies - consider appropriateness - open sources
10. Identify relevant benchmarks - consider analysis of technologies
11. Consider use of relevant benchmarks - data input / execution
12. Analyse and report results - for the project exploitation - and possible for DataBench knowledge nuggets (marketing)

The following checkpoint list has been suggested for projects that want to take advantage of the DataBench Toolbox for selection and reporting of use of Big Data and AI technologies:

A. Identify Application area and domain (D4.4, D4.3), Business Objectives/KPIs – Consider to do the DataBench Self-assessment Survey from the perspective of any applications (D4.4, D4.3)

B. Identify key technical objectives (BDV Reference model area scope), Identify and map to the 4 pipeline steps, Identify relevant Big data types and processing types + architectural patterns

C. Identify any relevant Blueprints - use case independent and domain / use-case specific

D. Identify relevant standards for big data and AI technologies (ISO SC42), Identify relevant technologies - consider appropriateness - open source

E. Identify relevant benchmarks - consider analysis of technologies. Consider use of relevant benchmarks for the analysis of relevant/provided tools and components.

F. Analyse and report results - for the project exploitation - and possible for DataBench knowledge nuggets (marketing)

In the D5.5 "Final report on methodology for evaluation of industrial analytic projects scenarios" [6] the usage of this methodology is reported for a representative set of Big Data PPP projects, further project experiences and results are continuously being reported as "Knowledge Nuggets" through the DataBench Toolbox.

3. Big Data and AI Pipeline Examples

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 Section 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 focused sectors (agriculture, forestry, and fishery) and intelligently process, analyse and visualize them.

3.1. 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 (a.k.a. events) 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) 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") 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. 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).

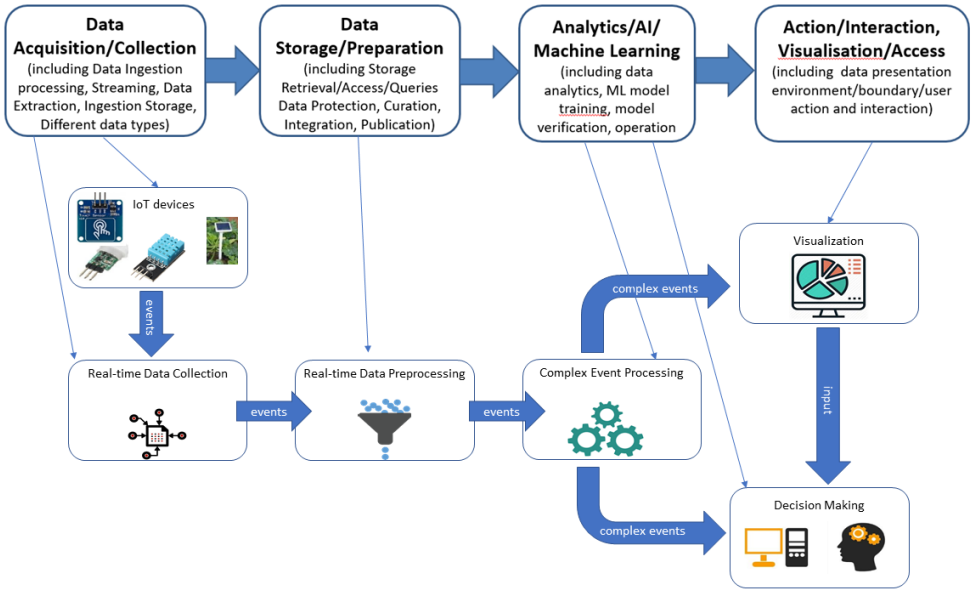

Figure 8: Mapping of "Pipeline for IoT data real-time Processing and decision making"

Figure 8 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 described in chapter 2.

3.2. Pipeline for Linked Data Integration and Publication

In DataBio project and some other agrifood 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 few other related projects) resulted in creating a repository of over 1 billion triples, being 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 realised in 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 uses cases related to the bioeconomy sectors. The main goal of these pipelines 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, and including also 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 Figure 9. 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-process 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, vi) identify links with other datasets, and publishes the generated datasets as Linked Data and applying 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 [14] were normally used in case of data coming from relational databases, tools like GeoTriples [15] was chosen mainly for geospatial data in the form of shapefiles, tools like RML Processor [16] for CSV, JSON, XML data formats, services like Ephedra [17] (within Metaphactory platform) for Restful APIs.

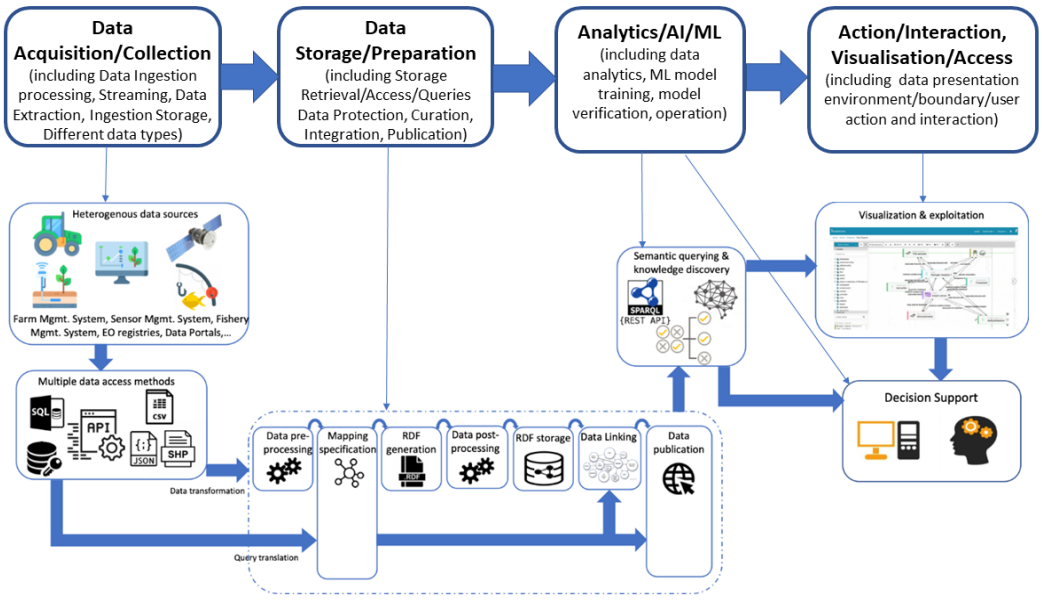

Figure 9: Mapping of "Pipeline for Linked Data Integration and Publication"

Figure 9 shows the steps of the pipeline for Linked Data Integration and Publication" that we have described above and their mapping to the steps of top level Generic Big Data and AI Pipeline pattern that we have described in chapter 2.

3.3. Pipeline for Earth Observation and Geospatial Data Processing

The pipeline for Earth Observation and Geospatial data processing, 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. 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, 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. Vector data usually describes some spatial features in the form of points, lines or polygones. Raster data, on the other hand, is usually generated from imaging-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 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 visualisation 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 Section 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 Figure 10.

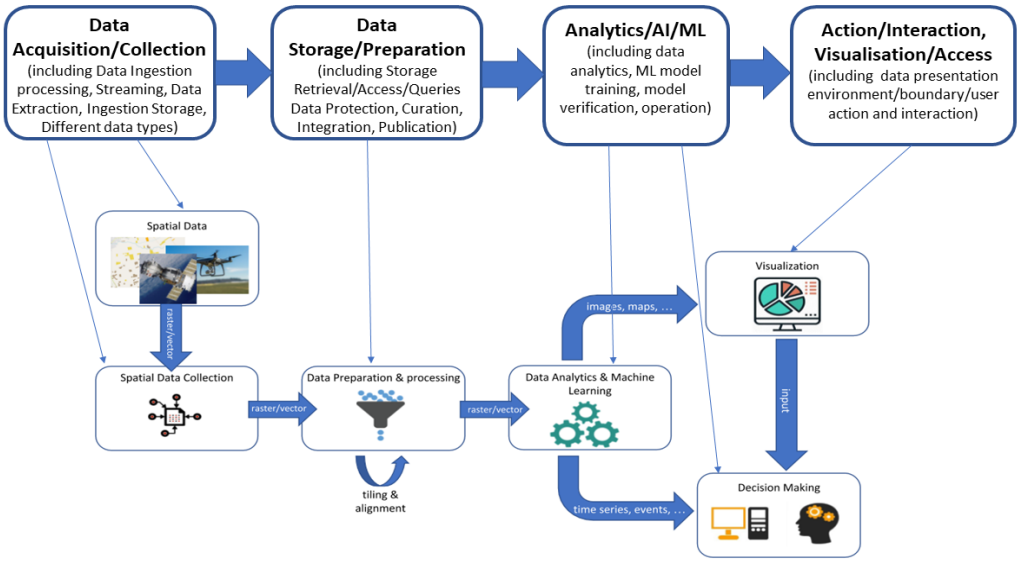

Figure 10: Mapping of "Pipeline for Earth Observation and Geospatial Data Processing"

4. Data Bench 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 identify architectural blueprints specifying the technical systems/components needed at different stages in a pipeline.

An important contribution towards the standardization of Big Data reference architectures, is given by Pääkkönen and Pakkala (2015) [18]. Pääkkönen and Pakkala proposed a reference architecture (shown in Figure 11), which is based on the six different pipelined stages, applied to data coming produced by several data sources.

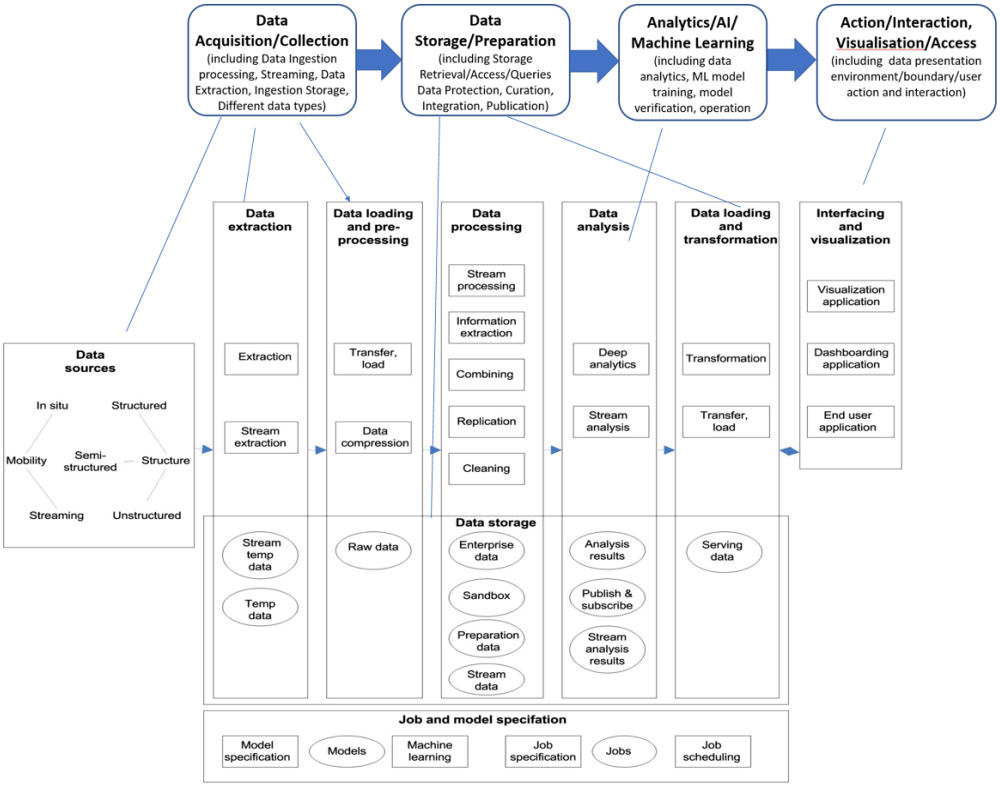

Figure 11: High-Level Design of Big Data Reference Architecture]

Data is distinguished between *in situ* data, not moved from where it is produced, and *streaming* data flow to be processed in real time (data-in-motion). The proposed phases are described below.

- *Data extraction*. It is the very first stage, in which the real data are requested and extracted, which either can mean stored temporarily, or transferred from the sources.
- *Data loading and pre-processing*. Data (or stream) is transformed and eventually compressed.
- *Data processing.* Data (or stream) is processed, combined with other datasets (or streams), integrated, cleaned and, usually, stored in a structured format. This phase comprehends all the activities aiming at improving data quality.
- *Data analysis*. The real data analysis happens in this stage, aiming at discovering knowledge from data.
- *Data loading and transformation*. The objective of this phase is to store the new information generated and transfer it to the visualization applications (or other end-user applications).
- *Interfacing and visualization*. Applications of this stage include dashboards, which usually show the behaviour of some predefined KPIs, visualization applications, which provide users more freedom in their monitoring, or different end-user applications.

Furthermore, Pääkkönen and Pakkala specify also the jobs intervening in the data analysis phase. The jobs can be saved and scheduled with a scheduling tool, and they can adopt machine learning models and algorithms to be trained or run on the extracted data (their structure is shown in the lower part of Figure 11). Finally, they provided the example of mapping the reference architecture, with some real-world industrial IT architectures.

The overall goal of introducing a reference architecture is to simplify the realisation of a big data system, by defining the important phases/stages that are commonly present in industrial applications. Moreover, it is favorable that the reference architecture is technology independent, but it offers a mapping between its components and the real big data technologies. The first concept helps the conceptual design of the infrastructure, while the mapping provides guidelines in the choice of the proper solution.

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. This classification work has been performed with an empirical bottom-up approach, starting from Big Data Analytics (BDA) use cases and then recognizing the commonalities among the technical requirements of different use cases and designing a general architectural blueprint depicted in Figure 12.

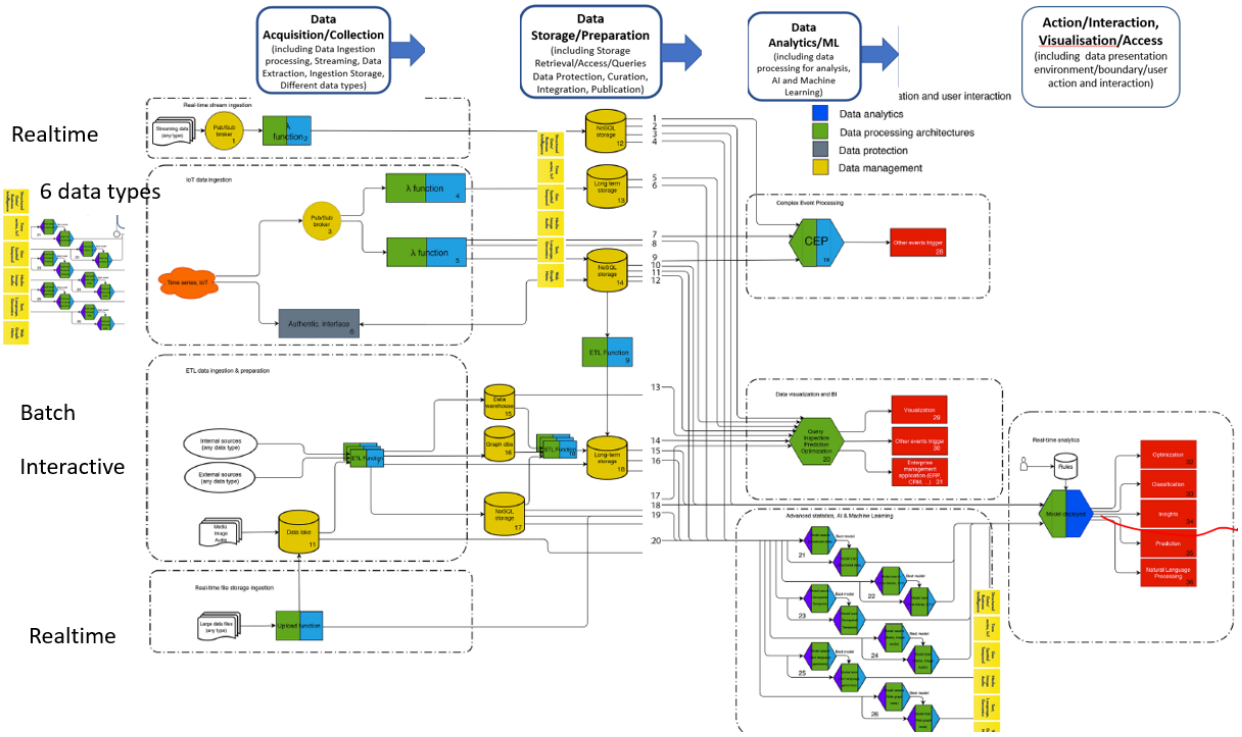

Figure 12: General architectural blueprint for BDA pipelines

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 in [6] the DataBench deliverables D2.2, D2.3 and D2.4 and

desk analysis D4.3, D4.3 and D4.4) that have shown the most frequent BDA use cases per industry. The most frequent BDA use cases are shown in Figure 12. We have then 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 these use cases and then inferred the general blueprint which is depicted on the above figure.

This general blueprint can be instantiated to account for the different requirements of different use cases and projects. In Databench, we have derived these 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 to the selection of a use case, to the specification of technical requirements, down to the selection and benchmarking of specific technologies for the different components of the use-case-specific blueprint.

The objective of this contribution is twofold: (1) to provide a framework for the evaluation of Big Data technologies designing an architecture that includes the most common components in Big Data infrastructures through which it is possible to associate technologies and compatible benchmarking tools and (2) to populate this model with hundreds of popular Big Data technologies and tens of benchmarks to make it usable. DataBench designs a benchmarking process to support the development and adoptions of Big Data technologies focusing on the measurement of parameters that are relevant for the business by studying available benchmarking tools and providing a set of metrics connecting technical and business performance.

In DataBench, 27 use-case specific Big Data architectures have been merged into a general blueprint, more than 1400 Big Data technologies available on the market have been considered, mapped to the general blueprint and evaluated in terms of popularity and functionality, ending with selecting 285 solutions for which a technical schematization has been done. Fifty-one (51) benchmarking tools have been considered, mapped to the general blueprint and schematized in their turn. Finally, selected technologies have been associated with compatible benchmarks. This mapping work is described in D4.3 [6] and has been done in WP4 in constant cooperation with WP5 to guarantee the soundness and completeness of the mapping work.

In the end, the relevance of a conscious Big Data technology selection is confirmed, thus the importance of having a tool facilitating the identification of adequate benchmarks is demonstrated by taking into account three use-case specific scenarios evaluating the differences in terms of costs and performances depending on the chosen technology.

4.1. General Blueprint Building Blocks

In the analysis taken as an input to construct the general blueprint, the starting point for the creation of use-case specific architectures has been the finding of common IT systems for Big Data and analytics pipelines. By data pipeline we refer to an ensemble of in-series architectural components in which the output of a component is the input of the next one. Some of the identified pipelines have been then used as building blocks for the construction of the twenty-seven use-case specific blueprints. The first step for the construction of the general blueprint has been to identify those building blocks that are part of the use-case specific blueprints and overall, four data storage management systems and four data processing exploitation systems have been identified. Although they have been largely and deeply explained in the previous research work by Gianmarco Ruggiero [13], since they are

the core of the general blueprint, it is important to include a brief explanation of the most relevant aspects of these building blocks.

The building blocks that had been used to design use-case specific architectures and the conceptual areas of the BDV reference model have been the starting point of the design of the general blueprint. During the construction of the architecture, some requirements to be satisfied have been taken into account:

- **Consistent abstraction level between use-case specific and general blueprint.** The abstraction level of the building blocks is higher than the one of the use-case specific architecture thus a simple merging of the identified architectural building blocks was not enough.
- **Bijection.** As the general blueprint is derived from the use-case specific architecture, it must be possible to derive every use-case specific architecture starting from the general blueprint. In this regard, an example is provided in the next section.
- **Readability.** The final general blueprint should be as clear and readable as possible.
- **BDVA data types consistent.** The general blueprint should include all the data types in the classification proposed by the Big Data Value Association. Since the core building blocks have been explained in the previous sections and a more detailed explanation is available in the research work by Gianmarco Ruggiero, this report only focuses on two details of this general blueprint that are worth to be specified:
    - **BDVA data types.** As mentioned above, one requirement of the general blueprint is to be consistent with the Big Data Value Association data types classification. BDVA divides data types into six different categories: structured data / business intelligence, time series/IoT, Geo-Spatial-Temporal, Media Image Audio, Text, Language Genomics, Web Graph Meta. The consistency of the general blueprint with the BDVA classification is particularly evident in the “Advanced statistics, AI & Machine Learning area” where ”Model Search” and ”Model Train” architectural components are replicated for every data type. Although the mapping of technologies and benchmarks consistently with BDVA data types requires a level of abstraction that goes beyond the purpose of this research work, further research on the general blueprint is being done by companies collaborating in DataBench project and in this context a lower abstraction level, thus a division consistent with BDVA data types is crucial.
    - **Missing components.** Some of the use-case specific blueprints involve architectural components that are not used by any other. For the sake of clarity those components have been discarded in the general blueprint.

The general blueprint is represented in Figure 12. It is also interesting to notice that it is possible to see the components of the general blueprint in the perspective of the horizontal concerns of the BDV reference model. In Figure 13 and Figure 14 every horizontal concern of the BDV reference model has been assigned a specific color and each component of the general blueprint has been associated to one or more conceptual area by graphically associating the respective color. While the BDV reference model explicates the pre-mentioned conceptual areas, it does not explain how the different areas interact. By mapping the different components of the general blueprint to the horizontal concerns of the BDV reference model, it has been possible to highlight the interaction among the different Big Data conceptual areas.

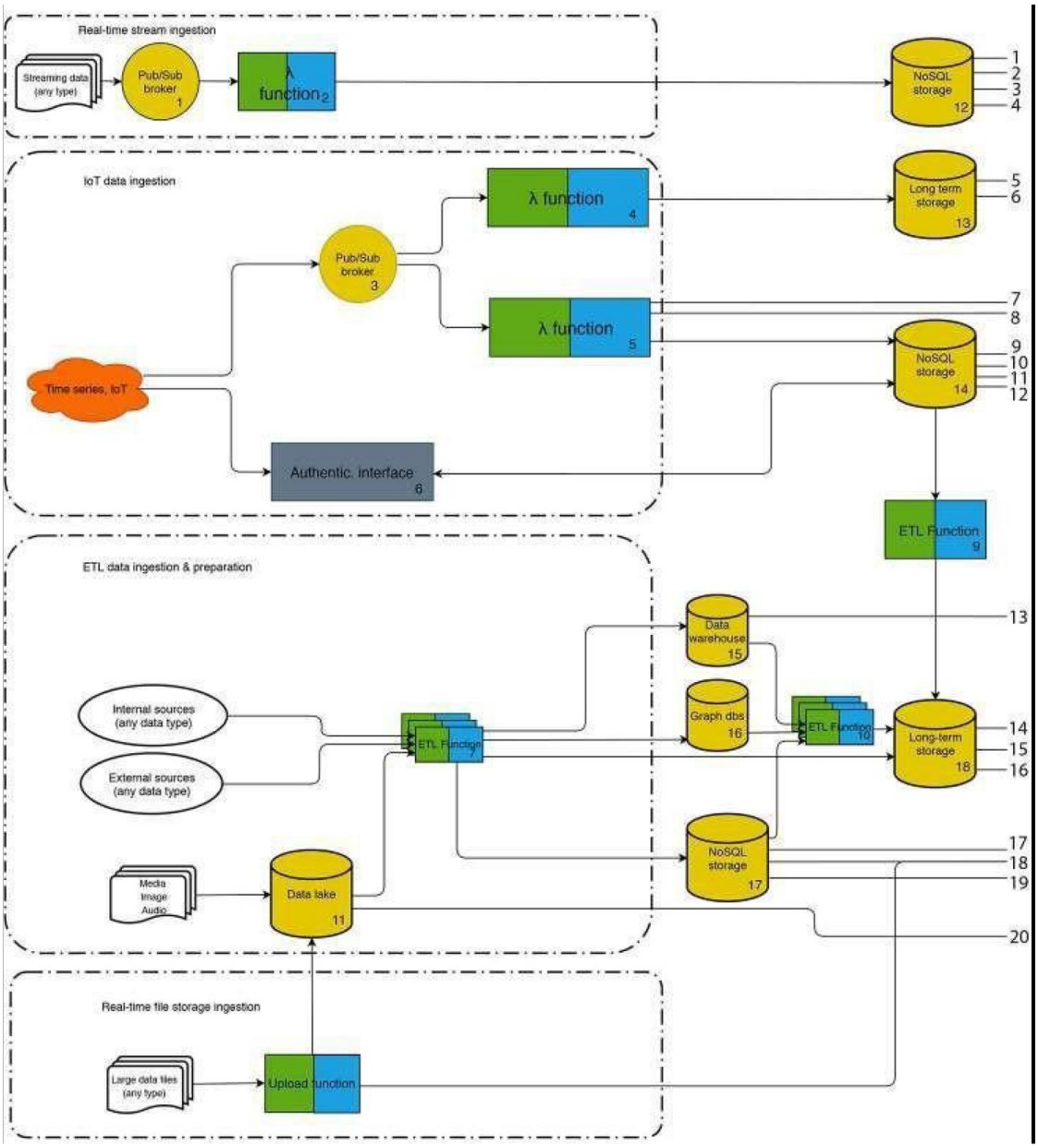

Figure 13: Mapping between BDV reference model and general blueprint, left side

The figure identifies the following blueprints related to the pipeline steps of Data Acquisition and Data Storage and Preparation:

- Real -time stream ingestion
- IoT data ingestion
- ETL data ingestion & preparation
- Real-time file storage ingestion
- Data storage and ETL preparation

The following related blueprints are described in more detail in Annex A - Data management systems, Extract-Transform-Load storage architecture, Real-time stream storage architecture, Real-time file storage architecture, IoT backend architecture, Data processing and exploitation systems.

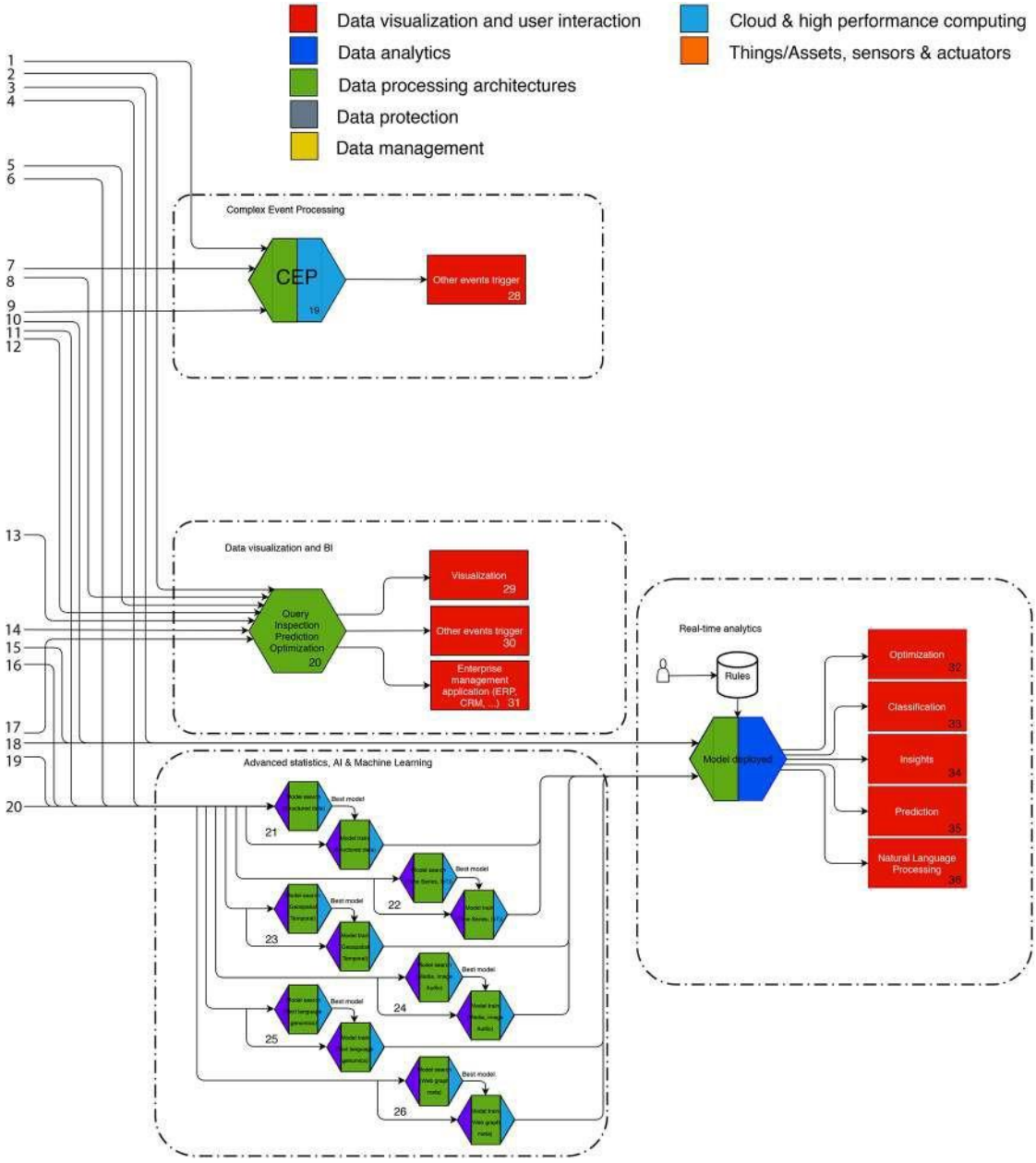

Figure 14: Mapping between BDV reference model and general blueprint, right side

Figure 14 identifies the following blueprints related to the pipeline steps of Data Analytics and Data Storage and Preparation:

- Data visualization and AI
- Advances statistics, AI and Machine Learning

- Real-time analytics
- Complex event processing

The following related blueprints are described in more detail in Annex A: Data visualization and business intelligence architecture, Data analytics and machine learning architecture, Real-time analytics architecture and Complex event processing.

In addition to these initial blueprints it is assumed that further blueprints will be identified during analysis of architectural patterns found in various big data and AI pipeline analysis being realized through various projects that are providing their results as input to the DataBench Toolbox.

5. DataBench Toolbox – supporting the Methodology

The DataBench Toolbox has been explained in detail in DataBench deliverable D3.4, as well as in the project Handbook in deliverable D4.4 [6]. The Toolbox allows access to resources related to big data benchmarking in the form of two main catalogues: 1) the benchmarking tools catalogue, which provides a list of existing benchmarking tools and solutions, some of them integrated for potential deployment; and 2) a Knowledge Nuggets catalogue that gives access to a knowledge base related to technical and business benchmarking, as well to information about the DataBench generic pipelines and blueprints presented in this document.

Figure 15 shows the main building blocks of the Toolbox. These are the 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.

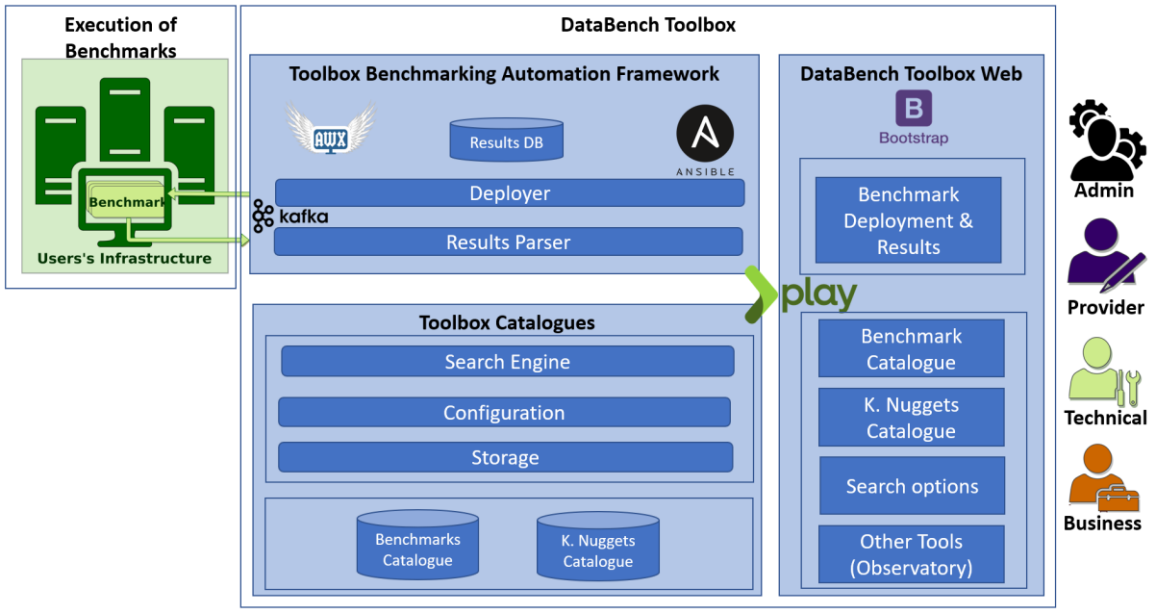

Figure 15: DataBench Toolbox Functional Architecture Overview

The access to the Toolbox is done via a dedicated web user interface accessible via https://databench.ijs.si/. The DataBench Toolbox Web building block shown in Figure 15 allows browsing and navigating through the information stored in the DataBench Toolbox Catalogues building block, giving the possibility to search and find technical benchmarking tools and knowledge related to big data and AI, such as use cases, lessons learned, business KPIs in different sectors of application, architectural blueprints of reference and other aspects related to benchmarking big data and AI from a business perspective, etc.

The Toolbox Benchmarking Automation Framework building block shown in Figure 15, 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 is not providing a playground to deploy and execute benchmarks. The automation of

This project has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No 780966.

the deployment and execution of the benchmarks is achieved via the generation of Ansible Playbooks [19] and enabled by an AWX project [20] 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 section 3.1 of DataBench Deliverable D3.4 which can be found in [6].

The Toolbox main page offers a set of so-called user journeys. User journeys are textual descriptions of tips and advice on how to use and navigate throughout the Toolbox for different types of users (technical, business, benchmarking providers and administrators). The front page also gives access to other tools such as the DataBench Observatory explained in section 8. The search options included in the menu provides three main type of search: 1) search by a clickable representation of the BDV Reference Model; 2) a guided search enabling the selection of categorized tags used for the annotation of the benchmarks and knowledge nuggets; and 3) search by a clickable depiction of the big data architectural blueprints and the generic pipeline presented in section 2. The latter type of searching, depicted in Figure 16, enables accessing technical benchmarks as well as nuggets related to the clicked elements.

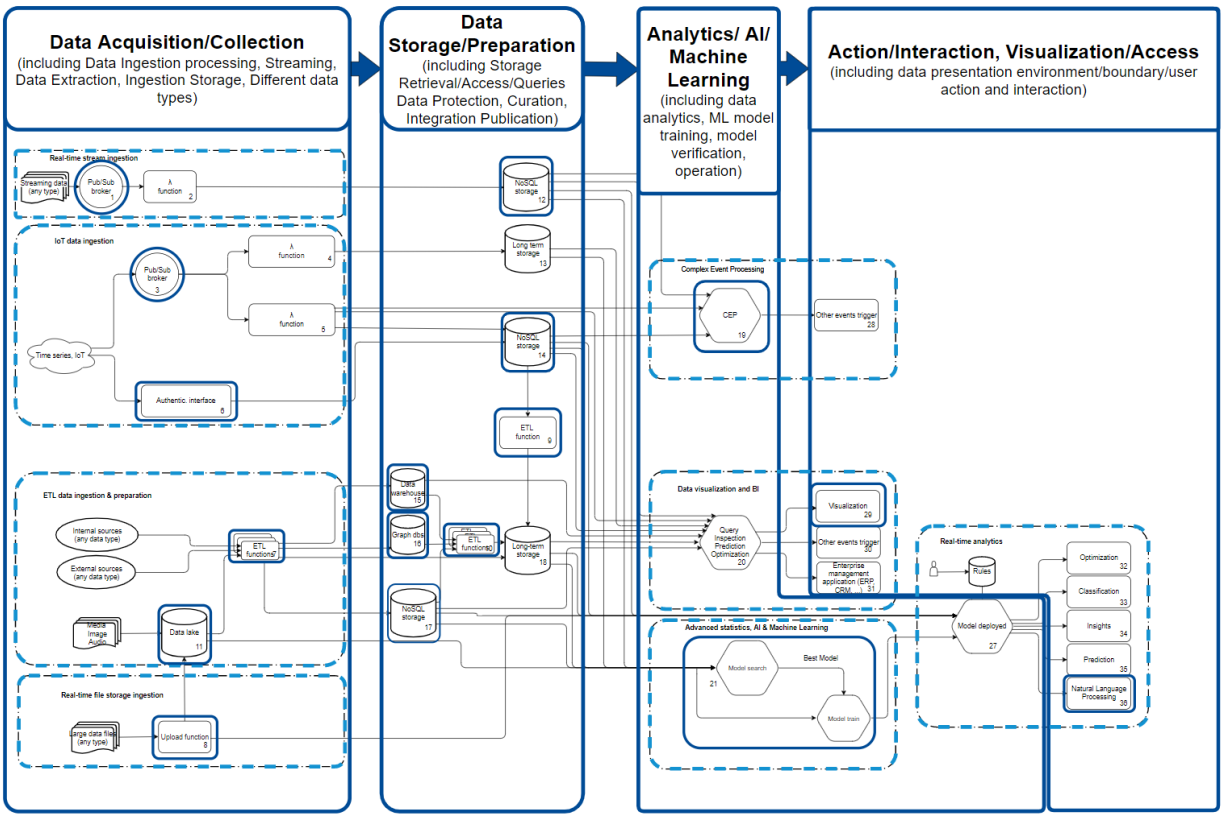

Figure 16: Search by Pipeline/Blueprint available from the DataBench Toolbox

Besides this search, the DataBench Toolbox offers several knowledge nuggets especially related to pipelines and blueprints. It is worth mentioning the ones explaining in detail the main 4 steps of a Generic Data Pipeline and the Generic Big Data Analytics Blueprint. There are also specific nuggets related to mapping either the generic pipeline or the blueprint to different architectures for a specific sector or use case. In order to find these pieces of knowledge, users might just use the full text search box located at the top right corner of the

Toolbox and type “blueprint” or the industry vertical of their choices (e.g. “agriculture). The Guided Search also offers the possibility of selecting tags related to industries or blueprints.

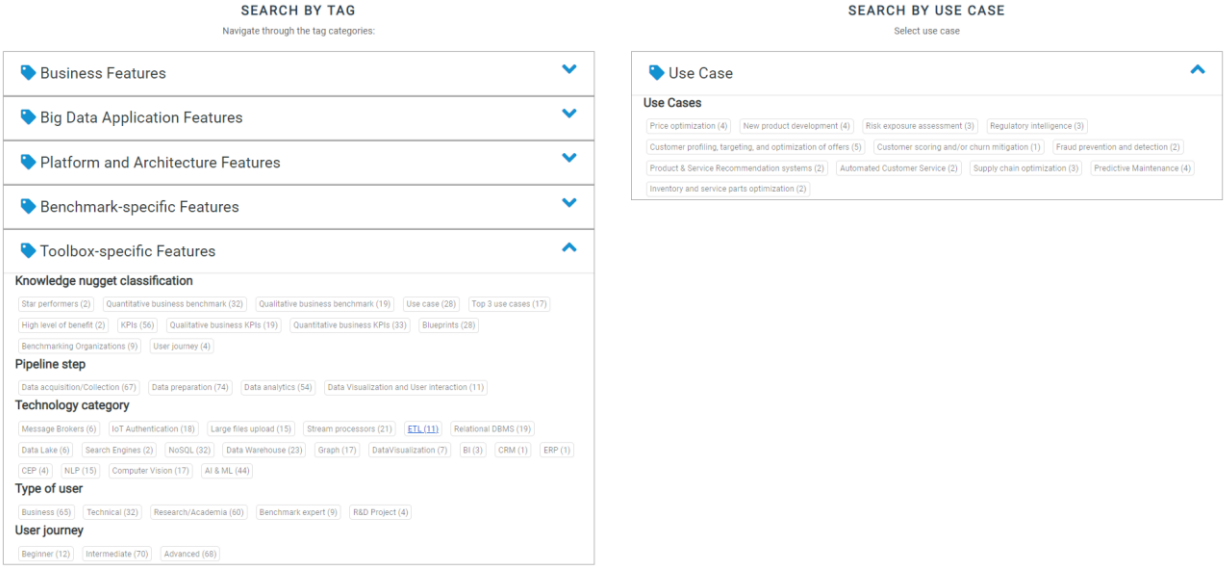

Figure 17: Guided Search by tag and by use case, available from the DataBench Toolbox

6. Technical Benchmarks and Components, Tools and Standards

6.1. Technical Benchmarks– related to the Big Data and AI Pipeline Framework

The goal of the DataBench framework with the supporting DataBench Toolbox 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 BDVA 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 [21]. 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 the Benchmark matrix. Figure 18 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, Spatio-Temporal 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 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.

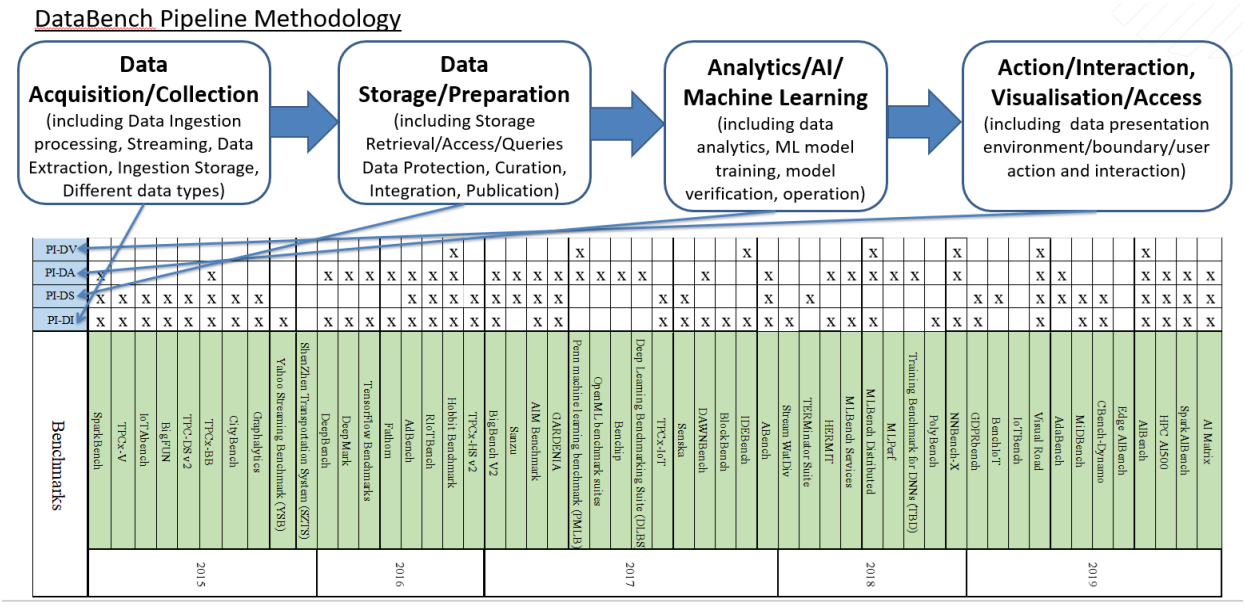

Figure 18: DataBench Pipeline mapping to Benchmarks

As we can see in Figure 18, 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 technologies like batch processing. In this case, focusing on a specific processing category reduces the number of retrieved benchmarks as the example in Figure 19, 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.

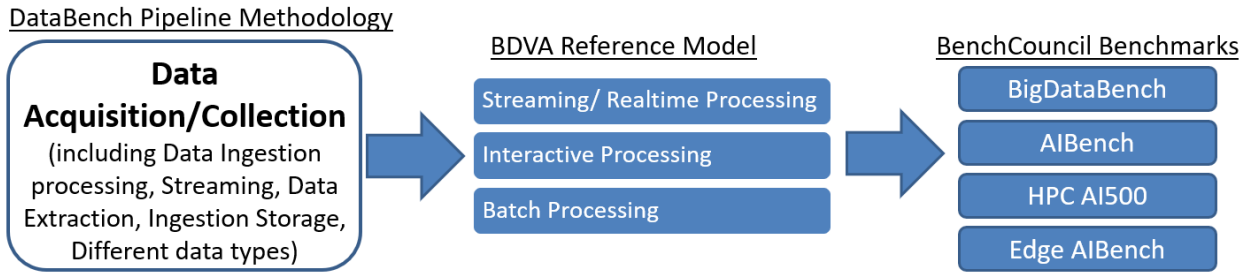

Figure 19: DataBench Pipeline step mapping to specific category of Benchmarks

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 web interface as shown on Figure 20. The main Guided Benchmark Search implements closely both the DataBench Indicators Ecosystem and the Benchmark Matrix, described in WP1. More details for the user interface of the Guided Benchmark Search are provided in D5.2 and in the DataBench Benchmarking

Handbook D4.4 [6]. The objective here is to demonstrate how the different searches can be used and the results that can be obtained in the current version of the DataBench Toolbox.

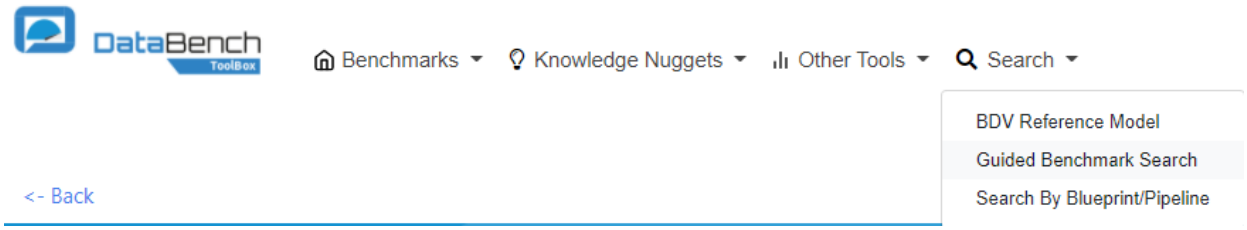

Figure 20: DataBench Benchmark Search Menu Options

The first search option is based on the BDV Reference Model that was already presented and shown in section 2, Figure 4. In the search page of the ***BDV Reference Model*** [29] all the different horizontal and vertical layers are depicted. Each layer can be selected and links to the search results of both benchmarks and knowledge nuggets. Table 1 below illustrates the search results for each of the 6 data types marked with yellow. The top 10 benchmarks are listed in blue, whereas the top 5 knowledge nuggets for each data type are listed in orange.

Go to Search --> BDV Reference Model and select one of the Data Types:					
Structured Data/ Business Intelligence	Time series, IoT	Geo Spatial Temporal	Media Image Audio	Text, Language, Genomics	Web Graph Meta
The search returns the following list of benchmarks (only the top 10):					
BigBench V2	owperf (CLASS)	IDEBench	AIBench	BigBench V2	HiBench
HiBench	Yahoo Streaming Benchmark (YSB)	MLPerf	AIMatrix	HiBench	Berlin SPARQL Benchmark (BSBM)
Yahoo Streaming Benchmark (YSB)	AIoTBench	Training Benchmark for DNNs (TBD)	BigDataBench	AIBench	BigFrame
Yahoo! Cloud Serving Benchmark (YCSB)	BenchIoT		CloudSuite	AIMatrix	CloudRank-D

AMP Lab Big Data Benchmark	CityBench		DAWNBench	BigDataBench	gMark
BigDataBench	CloudRank-D		Deep Learning Benchmarking Suite (DLBS)	CloudRank-D	Graphalytics
BigFrame	CloudSuite		DeepMark (Convnet)	CloudSuite	Hobbit Benchmark
BigFUN	DeepMark (Convnet)		Edge AI Bench	DAWNBench	LinkBench
CALDA	Edge AI Bench		Fathom	DeepMark (Convnet)	LIQUID
CloudRank-D	HERMIT		HPC AI500	Edge AI Bench	MiDBench
The search returns the following list of knowledge nuggets (only the top 5):					
NoSQL - Key-Value DB	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Linked Data Integration and Publication Pipeline pattern
Benchmarks features matrix	Data features Star Diagram	Data features Star Diagram	Computer Vision	Data features Star Diagram	Benchmarks features matrix
Data features Star Diagram	IoT Ingestion and Authentication	Earth Observation and Geospatial Pipeline pattern	Data features Star Diagram	DataBench Generic Data Pipeline (4 steps data value chain)	Data features Star Diagram
Data Lake	IoT Pipeline pattern	Technical benchmarks Star Diagram	Technical benchmarks Star Diagram	Natural Language Processing (NLP)	Linked Data Benchmark Council (LDBC)
Data Warehouse	Technical benchmarks Star Diagram	Transaction Processing Performance Council		Technical benchmarks Star Diagram	NoSQL - Graph DB

Table 1: Results from BDV Reference Model Search

It is clear from above search results, that for most data types there are at least 10 benchmarks that can be selected from, except the Geo Spatial Temporal data type to which are related only to three benchmarks. With respect to knowledge nuggets, there is almost equal number for the different data types and will probably increase with the evolution of the DataBench Toolbox.

Table 2 provides a snippet of the results that users obtain when searching for benchmarks based on one of the four pipeline steps shown on Figure 19 and accessible via the ***Search by Blueprint/Pipeline*** [30]. The top 10 benchmarks returned by the searching for each step are marked in blue. In the case of step four "Actions/Interaction, Visualisation/Access", only six benchmarks are returned and classified under this step.

Go to Search --> "Search By Blueprint/Pipeline" and select one of the four categories (in yellow):			
Data Acquistion/ Collection	Data Storage/ Preparation	Analytics/AI/Machine Learning	Action/Interaction, Visualisation/Access
The search returns the following list of benchmarks (only the top 10 in blue):			
BigBench V2	BigBench V2	BigBench V2	AIBench
HiBench	HiBench	HiBench	Hobbit Benchmark
owperf (CLASS)	owperf (CLASS)	AdBench	IDEBench
Yahoo Streaming Benchmark (YSB)	Yahoo Streaming Benchmark (YSB)	AIBench	NNBench-X
AdBench	Yahoo! Cloud Serving Benchmark (YCSB)	AIM Benchmark	Penn Machine Learning Benchmark (PMLB)
AIBench	AdBench	AIMatrix	VisualRoad
AIM Benchmark	AIBench	AIoTBench	
AIMatrix	AIM Benchmark	ALOJA	
AIoTBench	AIMatrix	Benchip	
ALOJA	AIoTBench	BigBench	

The search returns the following list of knowledge nuggets (only the top 5 in orange):			
ETL	NoSQL - Key-Value DB	AI&ML Developmemt Platforms (IDEs)	Benchmarks features matrix
IoT Ingestion and Authentication	Benchmarks features matrix	AI&ML Frameworks	Data Visualization tools
Large files ingestion	Control Event Processing (CEP)	AI&ML Libraries	Search engines
Message brokers and Pub-Sub	Data Lake	AI&ML Platforms	Use case independent blueprints - Data Processing and Exploitation Systems
	Data Warehouse	Benchmarks.AI. Directory of AI Benchmarks	Use case independent blueprints - Data processing and exploitation systems - Data visualization and business intelligence architecture

Table 2: Results from Search by Blueprint/Pipeline

In the second part of the table with orange are marked the top 5 knowledge nuggets returned by the search for each pipeline step. Some knowledge nuggets like the last two in the "Actions/Interaction, Visualisation/Access" step describe common blueprints and architectural patterns that can be followed as best practices when implementing similar types of systems. All the current independent blueprints are described in Annex A. The rest of the knowledge nuggets represent subcategories of technologies, which are common implementation choices for the pipeline steps where they appear. For example, in the Data Storage step users can search in the NoSQL-Key Value DBs list to identify NoSQL technology that can be optimal for their data requirements. Then, after selecting one or more promising NoSQL engines they can search for the best NoSQL benchmark (e.g. YCSB) that can provide them with enough results and metrics to pick the best of the selected NoSQL engines.

6.2. Components, Tools and Standards

The DataBench Observatory is another tool available as part of the Toolbox which can be also used in the selection of the most appropriate technologies when implementing Big Data and AI architectures. The metrics provided in the Observatory can be very useful to identify the popularity and scientific relevance of emerging technologies.

For the more advanced technical users that work daily with benchmarks or develop their own benchmarks it can be very helpful to follow and be familiar with the most active benchmark communities and organizations. The Toolbox provides an extensive list of these benchmark communities and organizations. Among them we can highlight the following ones:

- Transaction Processing Performance Council (TPC)
- Standard Performance Evaluation Corporation (SPEC)
- Securities Technology Analysis Center (STAC)
- Linked Data Benchmark Council (LDBC)
- International Open Benchmarking Council (BenchCouncil)
- Hobbit platform and community (Hobbit)
- MLPerf Community (MLPerf)
- BDVA Big Data Benchmarking Sub Group (BDVA TF6 SG7)

Another important argument for following the above communities is that they also regularly publish benchmark results usually in combination with detailed reports explaining what hardware and software components were tested and how exactly were done the performance tests. This type of results (for example validated and published by TPC, SPEC and BenchCouncil) can have at least two important applications:

1) can be used as reference results by practitioners to validate and position their performance results;
2) can decide based on the published results which combination of hardware and software setup will be the best pick for their application requirements.

Below follows a list with some of the available benchmark results published by different organizations and projects:

- Published SPEC Benchmark Results [31]
- ALOJA Project benchmarking results for HiBench and BigBench (TPCx-BB) [34]
- TPCx-IoT All Results - Sorted by Performance [35]
- TPCx-BB All Results - Sorted by Performance [36]
- TPCx-HS All Results - Sorted by Performance [37]
- HPC AI500 Ranking, Image Classification, Free Level [38]
- AIBench: A Datacenter AI Benchmark Suite Ranking [39]
- OpenML Dataset execution results [32]
- Visual Data Challenges from Maryland University [33]

Standards are now emerging in the areas of Big Data and AI technologies:

There are number of recently published Big Data and AI standards from ISO SC42 Artificial Intelligence (and Big Data):

- ISO/IEC TR 24028:2020 - Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 20547-1:2020 - Information technology — Big data reference architecture — Part 1: Framework and application process
- ISO/IEC TR 20547-2:2018 - Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
- ISO/IEC 20547-3:2020 - Information technology — Big data reference architecture — Part 3: Reference architecture
- ISO/IEC TR 20547-5:2018 - Information technology — Big data reference architecture — Part 5: Standards roadmap
- ISO/IEC 20546:2019 - Information technology — Big data — Overview and vocabulary

In addition to these there is a larger number of Big Data and AI standards in progress as follows:

ISO SC42 WG 1 Foundational standards and WG5 Computational AI

- ISO/IEC CD 23053.2 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
- ISO/IEC CD 22989.2 Artificial intelligence — Concepts and terminology
- ISO/IEC CD 38507 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations
- ISO/IEC AWI TR 24372 Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems
- ISO/IEC WD 42001 Information Technology — Artificial intelligence — Management system
- ISO/IEC WD 5392 Information technology — Artificial intelligence — Reference architecture of knowledge engineering
- ISO/IEC WD TS 4213 Information technology — Artificial Intelligence — Assessment of machine learning classification performance

ISO SC42 WG2 Big Data and WG4 Use cases and applications

- ISO/IEC CD TR 24030 Information technology — Artificial Intelligence (AI) — Use cases
- ISO/IEC CD 24668 Information technology — Artificial intelligence —Process management framework for Big data analytics
- ISO/IEC WD 5339 Information Technology — Artificial Intelligence — Guidelines for AI applications
- ISO/IEC WD 5338 Information technology — Artificial intelligence — AI system life cycle processes
- ISO/IEC WD 5259-1 Data quality for analytics and ML — Part 1: Overview, terminology, and examples
- ISO/IEC NP 5259-2 Data quality for analytics and ML — Part 2: Part 2: Data quality measures
- ISO/IEC WD 5259-3 Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines
- ISO/IEC WD 5259-4 Data quality for analytics and ML — Part 4: Data quality process framework

WG3 Trustworthy AI

- ISO/IEC AWI TR 24027 Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
- ISO/IEC CD 23894 Information Technology — Artificial Intelligence — Risk Management
- ISO/IEC AWI 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI-based systems
- ISO/IEC AWI TR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns
- ISO/IEC AWI 24029-2 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods

- ISO/IEC NP TS 6254 Information technology — Artificial intelligence — Objectives and methods for explainability of ML models and AI systems
- ISO/IEC AWI TR 5469 Artificial intelligence — Functional safety and AI systems
- ISO/IEC DTR 24029-1 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC AWI 24029-2 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods

These unpublished standards and technical reports are currently in the form of working drafts or committee drafts and are thus only available to those that are working with these documents. Comments and input for these documents is currently being provided by BDVA and Big Data PPP members through liaisons with ISO SC42 and national representatives.

Related to the pipeline steps and the BDV Reference model there are also a number of other relevant standards for instance in the context of domain information models or ontologies (for various domains) and supporting information modeling standards and linked data standards for instance from W3C and other ISO groups, there are also other relevant standards for IoT, CyberSecurity, Cloud, HPC, Robotics and others. These standards are being followed up by the BDVA TF6 SG6 Standards group also in collaboration with the EU Multi Stakeholder ICT Standards Group and other organisations like AIOTI, NESSI and ECSO.

In addition to these there are also other relevant Big Data and AI standardisation activities like the OECD AI Policy Observatory (oecd.ai), ETSI, CEN and others. The BDVA TF6 SG6 Standardisation group is keeping an overview of these ongoing activities and is also coordinating input to and involvement in these on behalf of the BDVA Community.

7. DataBench Toolbox Observatory – for Benchmarks and Tools

The DataBench Observatory is a tool for observing the popularity, importance and the visibility of topic terms related to the Artificial Intelligence and Big Data, with particular attention dedicated to the concepts, methods, tools and technologies in the area of Benchmarking. DataBench Observatory introduces the popularity index, calculated for ranking the topic terms in time, which is based on several components, such as research articles from the Microsoft Academic Graph (MAG) [24], job advertisements from Adzuna service [25], EU research projects [26], cross-lingual news data from Event Registry system [27], projects on Github [28], general interest from Google Trends.

The methodology behind the DataBench observatory and its implementation are in details discussed in DataBench deliverable D5.2 Final evaluation of DataBench metrics. The DataBench observatory provides ranking and trending functionalities, including overall and monthly ranking of topics, tools and technologies, as well as customized trending options. In particular, the users can sort the ranked list of tools and technologies based on data source. Figure 21 demonstrates that TPC and TensorFlow benchmarks are highly popular within Github project data (accessed in November 2020).

month: All Topics: benchmark

Search:

Topic	Papers	EU Projects	News	Github	Jobs	Search Volume	Total
TPC-DS	1.35	1	1.03	10	5.5	4	3.81
TensorFlow Benchmarks	1.01	1	1	10	5.5	3.92	3.74
SWIM	10	1	10	4.6	5.5	8.48	6.6
Fathom	1.92	10	1	4.6	5.5	10	5.5
GARDENIA	1.54	1	1.15	1	5.5	8.75	3.16
ALOJA	1.2	1	1.42	1	5.5	8.68	3.13
CALDA	1.05	1	1.46	1	5.5	7.74	2.96
MRBS	1.11	1	1.01	1	5.5	6.81	2.74
Linear Road	1.28	1	1	1	5.5	6.57	2.73
BigFUN	1	1	1	1	5.5	6.77	2.71
Cloudsuite	1.16	1	1.14	1	5.5	6.28	2.68

Figure 21: DataBench Popularity Index (Tools and Technologies, category: Benchmark,)

Figure 22 shows time series for selected benchmarks (accessed in November 2020). Users interested in particular benchmark can observe its popularity (score 10 is the maximum normalized popularity) within academic Papers (and other selected data sources).

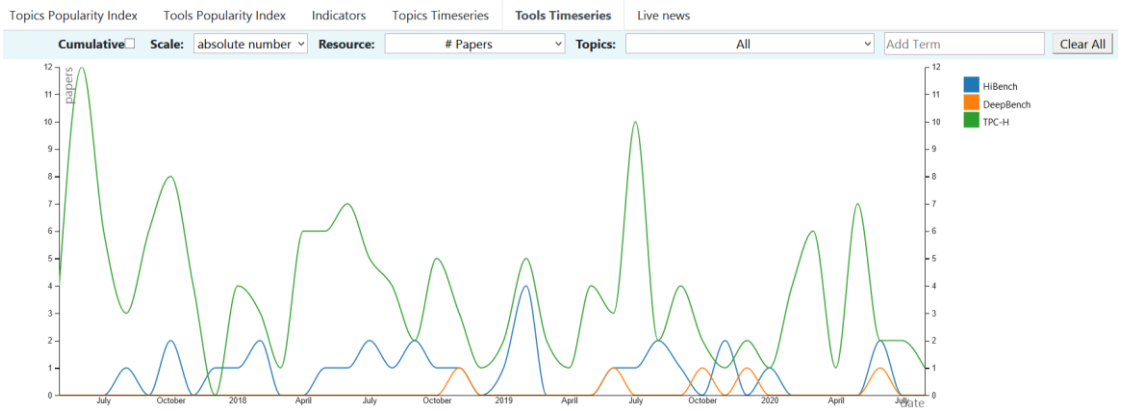

Figure 22: Time Series - Tools – selected benchmarks

The DataBench observatory (in the context of DataBench toolbox) is targeted at different user groups, such as business users, academic users etc. With DataBench observatory the academic users/researchers obtain an opportunity to find relevant topics in the area of Artificial Intelligence, Big Data and Benchmarking. Researchers who participate in the EU projects can search for popular topics within EU research and development domain, as well as industry related components, such as jobs and Github projects. Business users can use DataBench observatory for aligning the promising technologies, for observing the demand of tools and technologies on labour market, as well as for observing trends in time. The methodology behind DataBench index and implementation of the DataBench observatory allows for automatic system maintenance and updates based on the data availability.

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 [24] has been expanded with DataBench terms – over 1,700 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.

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

Figure 23: DataBench Popularity Index - Tools and Technologies: Graph Databases

Figure 23 illustrates the popular tools and technologies in the Graph databases category, sorted by popularity for Github data source (as accessed in November 2020).

8. Conclusions

This report has presented a Big Data and AI Pipeline Framework developed in the DataBench project, supported by the DataBench Toolbox. The Framework includes a number of dimensions including Pipelines steps, data processing types and types of different data. The relationship of the Framework is to existing and emerging Big Data and AI reference models such as the BDVA Reference Model and the AI PPP, and also the ISO SC42 Big Data Reference Architecture (ISO 20547) [2] and the emerging AI Machine learning framework (ISO 23053) [6] which the pipeline steps also have been harmonised with.

Further work is now related to populating the DataBench Toolbox with additional examples of actual Big Data and AI pipelines realised 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.

References

1. Big Data Value Strategic Research and Innovation Agenda (BDV SRIA). http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf
2. ISO/IEC 20547-3:2020, Information technology — Big data reference architecture — Part 3: Reference architecture
3. https://www.bdva.eu/PPP
4. https://ai-data-robotics-partnership.eu/wp-content/uploads/2020/09/AI-Data-Robotics-Partnership-SRIDA-V3.0.pdf
5. https://www.databio.eu/en/
6. https://www.databench.eu/public-deliverables/
7. ISO/IEC CD 23053.2: 2020 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
8. https://www.iso.org/committee/6794475/x/catalogue/
9. Wood, D., Lanthaler, M. & Cyganiak, R. (2014). RDF 1.1 Concepts and Abstract Syntax [W3C Recommendation]. (Technical report, W3C).
10. Harris, S. and 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. and Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool.

13. Gianmarco Ruggiero, "A methodology for the sizing and cost assessment of Big Data and Analytics (BDA) IT infrastructures," Politecnico di Milano, Graduation Thesis April 2020.
14. http://d2rq.org/
15. http://geotriples.di.uoa.gr/
16. https://rml.io/specs/rml/
17. https://www.metaphacts.com/ephedra
18. Pääkkönen, P. and Pakkala, D. (2015), 'Reference architecture and classification of technologies, products and services for big data systems', Big data research 2(4), 166–186.
19. https://docs.ansible.com/ansible/2.3/playbooks.html
20. https://www.ansible.com/products/awx-project
21. http://databench.ijs.si/knowledgeNugget/nugget/53
22. https://docs.ansible.com/ansible/2.3/playbooks.html
23. https://www.ansible.com/products/awx-project
24. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph
25. https://www.adzuna.co.uk
26. https://data.europa.eu/euodp/sl/data/dataset/cordisH2020projects
27. https://eventregistry.org
28. https://github.com
29. https://databench.ijs.si/bdva
30. https://databench.ijs.si/unicum
31. https://www.spec.org/results.html
32. https://www.openml.org/search?type=run
33. https://www.cs.umd.edu/hcil/varepository/benchmarks.php
34. http://aloja.bsc.es/
35. http://tpc.org/tpcx-iot/results/tpcxiot_results5.asp
36. http://tpc.org/tpcx-bb/results/tpcxbb_results5.asp
37. http://tpc.org/tpcx-hs/results/tpcxhs_results5.asp?version=2
38. https://benchcouncil.org/HPCAI500/ranking.html
39. https://www.benchcouncil.org/AIBench/ranking.html

Annex – Use Case Independent Blueprints

Data management systems

Data management systems building blocks can be divided into three logically distinct areas: ingestion, processing and storage.

- **Ingestion**: Data is transported from assorted resources (internal or external) to the processing function.
- **Processing**: Data is processed to make it consistent with the format required by the storage system.
- **Storage**: Data is persistently saved into a storage system.

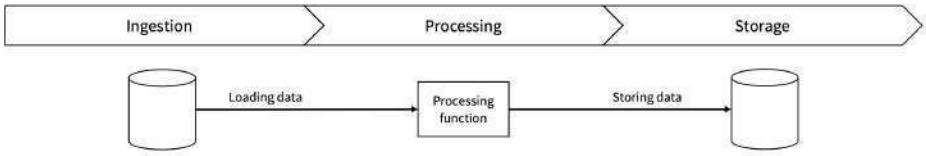

Figure 24: Data management systems phases

The use-case specific architectures rely on four different data management systems: *Extract-Transform-Load storage architecture*, *Real-time stream storage architecture*, *Real- time file storage architecture* and *IoT backend architecture*, that are briefly described in the following sections.

Extract-Transform-Load storage architecture

Data coming from different sources is extracted, transformed and loaded through an ETL system that is run periodically. Data is then stored or into a *SQL storage*, or into a *NoSQL* one. Both the frequency of ETL running and the type of database in which to store transformed data depend on the specific use case.

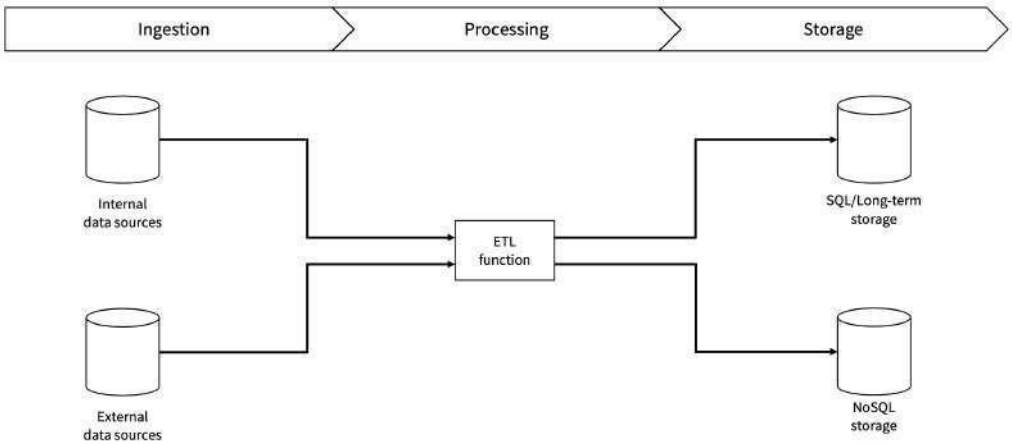

Figure 25: ETL storage architecture

Real-time stream storage architecture

Data coming from different streaming sources is acquired, processed and stored. Streaming data is ingested by a message broker and it is made available to a streaming processing platform (defined as lambda function in the figure) to make it storable. Processed data is then stored or into a data lake, or a NoSQL store. Data lake store is particularly suitable for use cases in which streaming data needs a batch processing or needs to be stored for a long time. NoSQL store is instead suitable for processes that need to receive data in a timely matter, nearly real-time.

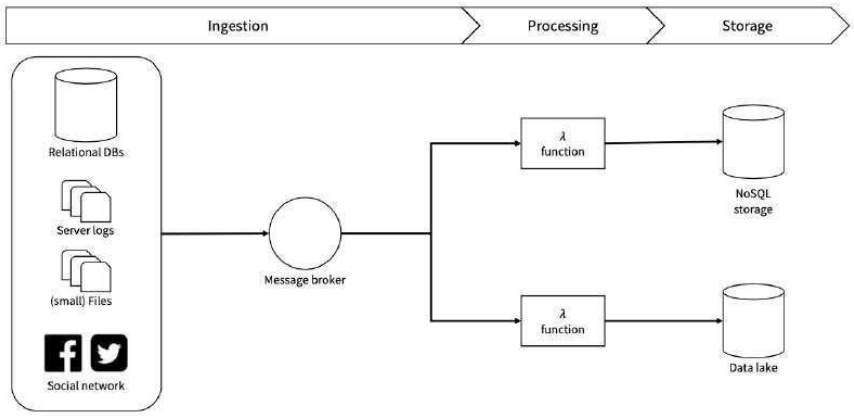

Figure 26: Real-time stream storage architecture

Real-time file storage architecture

The functionality of this architecture is very similar to the one of the previous building blocks but it is more suitable whenever the data to ingest, process and store consists of big files (e.g. images, videos). Data is firstly stored without any processing done into a data lake, here a message broker streams the files to a streaming processing platform or into a NoSQL storage, or into a DataWarehouse. NoSQL storage as in the previous building block is suitable for processes that need to receive data in a timely matter, nearly real-time. Data warehouse instead is particularly suitable whenever an organized and historical view of aggregated data is required.

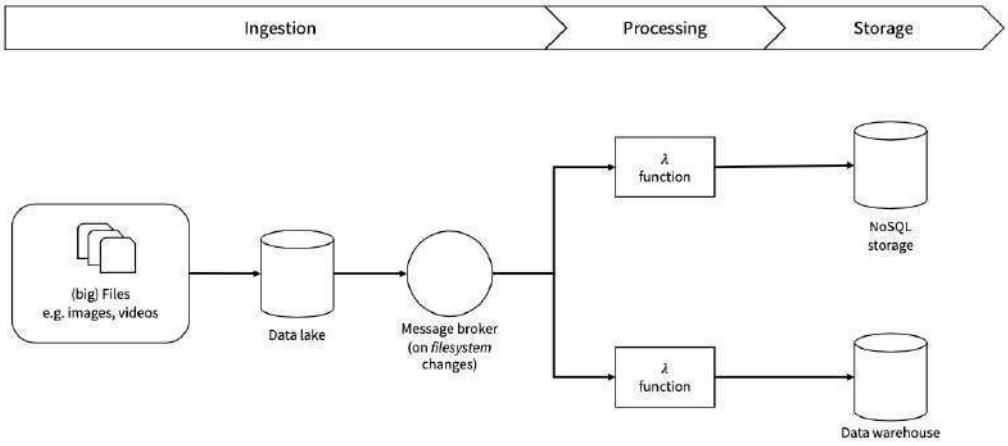

Figure 27: Real-time file storage architecture

IoT backend architecture

The most complex data management system architecture. IoT devices are provided with backend data access while transmitting new data streams. Authentication is managed by a synchronous interface that has access to a NoSQL database in which credentials are stored. Data streams are ingested by a message broker and the stream is, as in the previous data management building blocks, transformed by a stream processing platform and stored into a NoSQL storage or long term storage (e.g. data warehouse or data lake). The choice of storage systems depends on the specific needs of the use case.

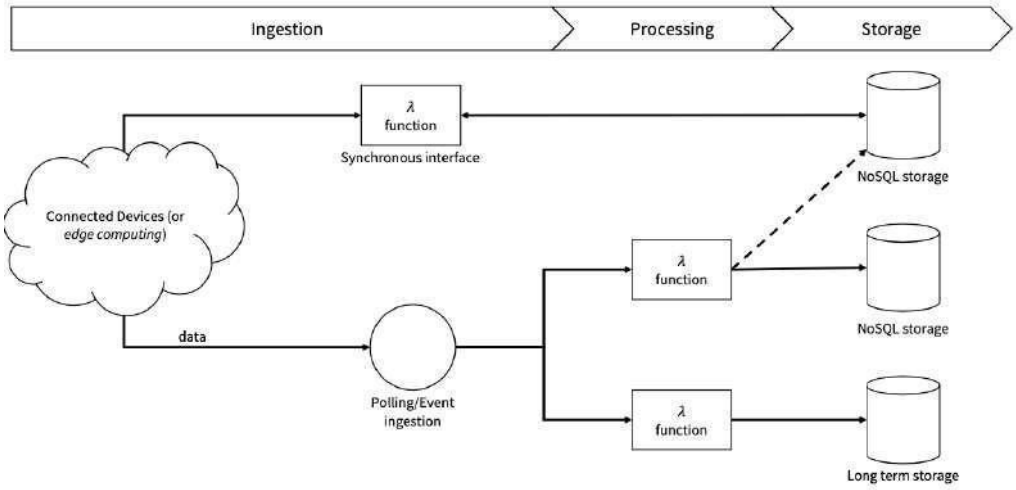

Figure 28: IoT backend architecture

Data processing and exploitation systems

Data processing and exploitation systems building blocks can be divided into three logically distinct areas: storage, exploitation and outcome.

- **Storage**. The last phase of data management building blocks is the first one of data processing and exploitation systems. Data is persistently saved into a storage system to be processed.

- **Exploitation**. Data is exploited by a generic processing subsystem (e.g. machine learning model or optimization algorithm).
- **Outcome**. It is the last phase of the entire pipeline and it strictly depends on the type of processing done in the exploitation phase. More details are given in the specific building blocks.

Figure 29: Data processing phases

The use-case specific architectures rely on four different data processing and exploitation systems: Data visualization and business intelligence architecture, Data analytics and machine learning, Real-time analytics and Complex event processing that are briefly described in the following sections.

Data visualization and business intelligence architecture

Data visualization & BI is by far the most adopted exploitation architecture. Data stored in a SQL or NoSQL storage is exploited by a processing sub-system that gives as an outcome interactive dashboards, email reports, visual workflows or predictive analytics. In most of data visualization & BI cases, data to be exploited is stored into a data warehouse, but there are also cases (e.g. in case of interactive dashboards) in which data is needed almost in real-time thus a NoSQL store is preferred.

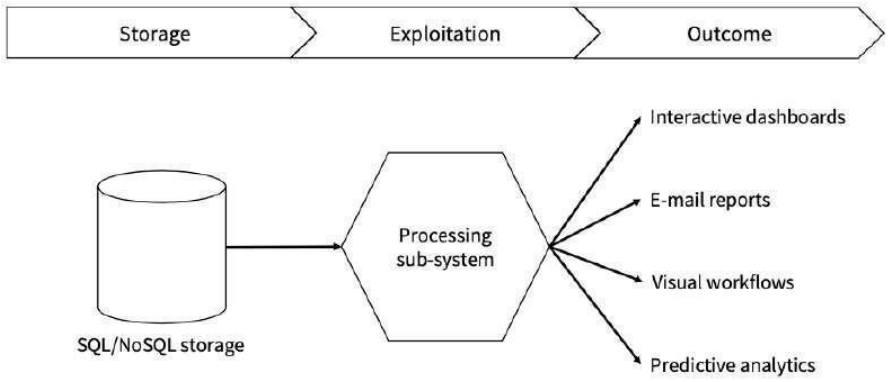

Figure 30: Data visualization and BI architecture

Data analytics and machine learning architecture

Advanced data analytics require the involvement of artificial intelligence and machine learning. This exploitation architecture is becoming more and more popular since it is the basis of data mining. The output of this architectural building block is or a machine learning model to be used by another processing function, or some insights or patterns extracted from the data. Involved storage systems are usually data lakes or data warehouses.

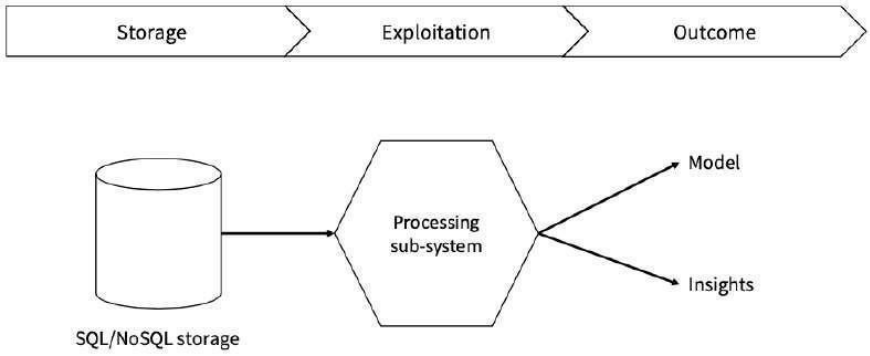

Figure 31: Data Analytics and machine learning architecture

Real-time analytics architecture

This architectural building block strictly relies on the usage of machine learning and artificial intelligence thus it is connected in pipeline with the data analytics and ma- chine learning architecture explained in the previous paragraph. A trained model, that is usually stored in a NoSQL database receives data coming from another storage com- ponent and gives real-time analytics on new incoming data to get insights and discover new knowledge about the data, which in turn is stored and possibly reused.

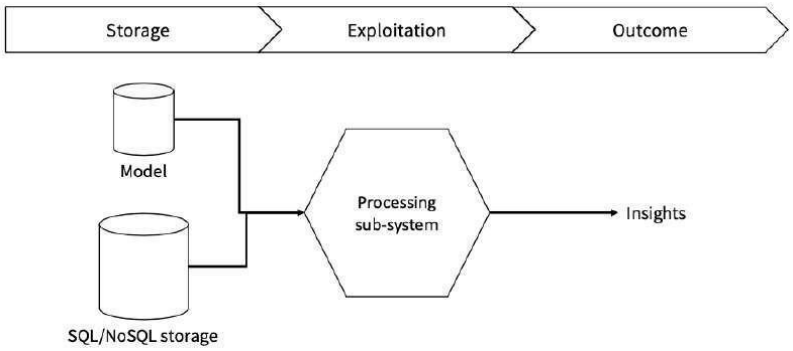

Figure 32: Real time analytics architecture

Complex event processing

Complex event processing (CEP) building block is all about query streaming of data. A stream processor receives data coming from a SQL or NoSQL storage giving as an output detection of some event patterns, event abstraction or event filtering. In this case, storage is not strictly needed but it is useful to permanently store the data stream for further processing.

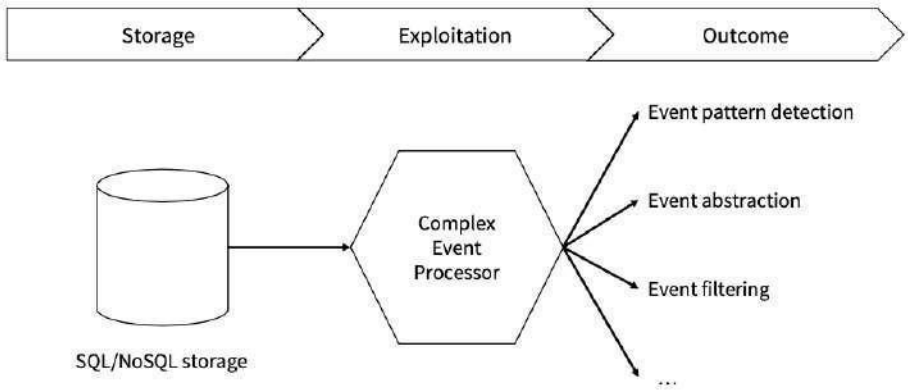

Figure 33: Complex event processing architecture