BD4NRG
Big Data for Next Generation Energy

D3.1:
BD4NRG Governance – 1st Technology Release

WP3 – Data Governance and Management

August 2021

www.bd4nrg.eu

Disclaimer

Copyright Message

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613

Full Title	Big Data for Next Generation Energy		
Topic	DT-ICT-11-2019 | Big data solutions for energy		
Funding scheme	H2020- IA: Innovation Action		
Start Date	January 2021	Duration	36
Project URL	www.bd4nrg.eu		
Project Coordinator	ENG		
Deliverable	D3.1: BD4NRG Governance – 1st Technology Release		
Work Package	WP3: Data Governance and Management		
Delivery Month (DoA)	M9	Version	0.1
Actual Delivery Date	30/11/2021		
Nature	Other	Dissemination Level	Public
Lead Beneficiary	ED		
Authors	ED, ENG, NTUA, TNO, UNINOVA, IDSA, FIWARE, ATOS		
Quality Reviewer(s):	Harris Doukas (NTUA)		
Keywords	Governance, Data, Middleware, HTAP, Homogenization, Interoperability, Blockchain, Pipeline		

Grant Agreement Number: 872613 Acronym: BD4NRG

Preface

BD4NRG focuses on addressing emerging challenges in big data management for energy sector with an innovative open holistic solution for smart grid-tailored, near real time, energy-specific and AI-based open Big Data Analytics modular framework. The vision is to deliver holistic services for techno-economic optimal management of Electric Power and Energy Systems (EPES) value chain. Services range from optimal risk assessment for energy efficiency investments planning, to optimized management of grid and non-grid owned assets, improved efficiency, and reliability of electricity networks operation, while at the same time contributing to achieve fair energy prices to the consumers and laying the foundations for an EU-level energy-tailored data sharing economy. The BD4NRG Toolbox will be implemented and validated in real-life pilots in 12 large-scale demo sites across 10 countries for:

- Increasing the efficiency and reliability of the electricity network BD-4-NET
- Optimizing the management of assets connected to the grid BD-4-DER
- De-risking investments in energy efficiency and increasing the efficiency and comfort of buildings BD-4-ENEF

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Who We Are

	Participant Name	Short Name	Country Code	Logo
1	ENGINEERING – INGEGNERIA INFORMATICA SPA	**ENG**	IT	
2	NATIONAL TECHNICAL UNIVERSITY OF ATHENS	**NTUA**	GR	
3	RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN	**RWTH**	DE	
4	EUROPEAN DYNAMICS LUXEMBOURG SA	**ED**	LU	
5	INTERNATIONAL DATA SPACES EV	**IDSA**	DE	
6	EUROPEAN NETWORK OF TRANSMISSION SYSTEM OPERATORS FOR ELECTRICITY AISBL	**ENTSO-E**	BE	
7	PANEPISTIMIO DYTIKIS ATTIKIS	**UNIWA**	GR	
8	ATOS SPAIN SA	**ATOS**	ES	
9	FUNDACION CARTIF	**CARTIF**	ES	
10	UNIVERZA V LJUBLJANI	**UNILJ**	SL	
11	ENEL X SRL	**ENELX**	IT	
12	REN - REDE ELECTRICA NACIONAL SA	**REN**	PT	
13	CENTRO DE INVESTIGACAO EM ENERGIA REN - STATE GRID SA	**RDN**	PT	
14	UNINOVA-INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIASASSOCIACAO	**UNINOVA**	PT	
15	ENERCOUTIM - ASSOCIACAO EMPRESARIALDE ENERGIA SOLAR DE ALCOUTIM	**ENERC**	PT	
16	FIWARE FOUNDATION EV	**FIWARE**	DE	
17	CENTRICA BUSINESS SOLUTIONS BELGIUM	**CENTRICA**	BE	
18	NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO	**TNO**	NL	
19	ASM TERNI SPA	**ASM**	IT	

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	Participant Name	Short Name	Country Code	Logo
20	VIDES INVESTICIJU FONDS SIA	**LEIF**	LV	
21	COMSENSUS, KOMUNIKACIJE IN SENZORIKA, DOO	**COMSENSUS**	SL	
22	HOLISTIC IKE	**HOLISTIC**	GR	
23	INTERUNIVERSITAIR MICRO-ELECTRONICA CENTRUM	**IMEC**	BE	
24	TERRASIGNA SRL	**TS**	RO	
25	UBIMET GMBH	**UBIMET**	AT	
26	ELEKTRO LJUBLJANA PODJETJE ZADISTRIBUCIJO ELEKTRICNE ENERGIJE D.D.	**EKL**	SL	
27	BORZEN, OPERATER TRGA Z ELEKTRIKO, D.O.O.	**BORZEN**	SL	
28	AJUNTAMIENTO DE SANT CUGAT DEL VALLES	**AJSCV**	ES	
29	ELES DOO SISTEMSKI OPERATER PRENOSNEGA ELEKTROENERGETSKEGA OMREZJA	**ELES**	SL	
30	E-LEX - STUDIO LEGALE	**ELEX**	IT	
31	OSMANGAZI ELEKTRIK DAGITIM ANONIM SIRKETI	**OEDAS**	TR	
32	VEOLIA SERVICIOS LECAM SOCIEDAD ANONIMA UNIPERSONAL	**VEOLIA**	ES	
33	STICHTING EG	**EGI**	NL	
34	CINTECH SOLUTIONS LTD	**CN**	CY	
35	EMOTION SRL	**EMOT**	IT	

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Contents

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Figures

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Tables

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Executive Summary

The BD4NRG project encompasses a large number of stakeholders in the electricity sector as well as cross-sectorial areas such as mobility and buildings. The Big Data analytic services, which are to be implemented over the course of BD4NRG, will be demonstrated in 12 Large Scale Pilots. This results in a significant landscape of requirements and goals, which along with Security and Compliance Guidelines constitute a set of source material for the BD4NRG Governance Layer and the subsequent 1st Technology release. The following report provides an overall view of WP3 components but beyond that attempts to establish an early system architecture of the governance layer which will form the basis for the next Technology releases.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1 Introduction

1.1 Scope of this deliverable

Scope of this deliverable is to provide an overall view of WP3 components especially those present in the 1st Technology release of BD4NRG Governance Layer but beyond that to also establish a unified system architecture of the governance layer which will form the basis for all Technology releases based on the BD4NRG Conceptual and Reference Architecture, the pilot functional requirements and finally security and compliance guidelines. In short, the BD4NRG Governance Layer will be comprised of a FIWARE compliant- DLT/blockchain-based implementation for a decentralized data sovereignty and governance architecture for cross-entity data sharing, which will integrate IDSA and GAIA-X conceptual architectures, and truly extend the FIWARE Context broker to seamlessly integrate with hybrid IoT/blockchain off-chain data sharing solutions. Scalability of the planned infrastructure will be achieved through utilizing the blockchain for hash storing, which uniquely will refer to the information content which is managed off-chain in a decentralized way (e.g., IPFS). Hence, we will be gaining the immutability, traceability, accountability, and notarization/time stamping benefits offered by Distributed Ledgers and blockchain technologies while at the same time effectively managing DLT intrinsic difficulty of scaling up. This effort aims to provide the necessary middleware to act as a mediator between data users (Applications and Tools) and data providers who may want to decide case by case whether to disclose their data or not. State of the art solutions are used to guarantee traceability, provenance tracking and accountability.

1.2 Overall Work Package 3 Data Architecture

A Data Architecture for the whole Work Package 3 has been initially designed to serve as a masterplan of the BD4NRG data location, data description, data flows and data availability. It is intended to deliver a conceptual infrastructure to support data quality, data stewardship, data integration, data provisioning as well as system collaboration among applications and data repositories (see Figure 1).

The architecture describes the BD4NRG data system. Specifically, it shows – in left-to-right order – how the data flows from the Large Scale Pilot raw datasets, through the various layers of transformation (Interoperability & Homogenization Functions) and quality checks (Data Quality & Compliance) to obtain the Golden Data Source, through data integration tasks (Dynamic Data Provider Functions), through operational data store (HTAP), all the way to the business intelligence, analytics and decision support applications that use the Integrated Query Engine for querying optimized data structures for data processing.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

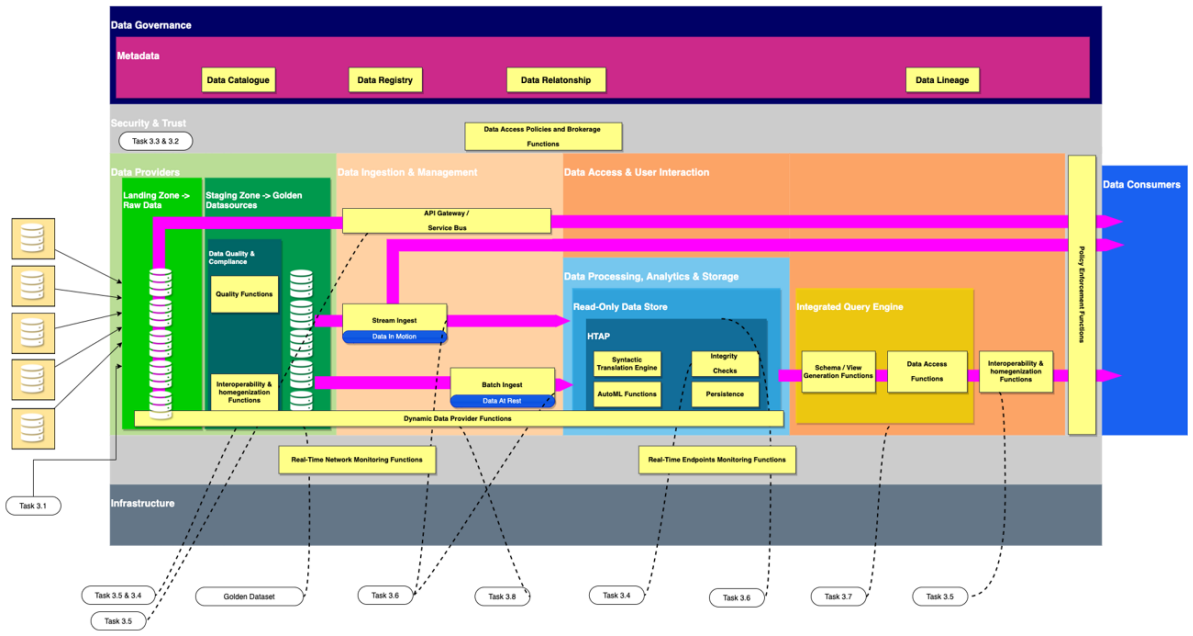


Figure 1 Work Package 3 Overall Data Architecture: BD4NRG Data System

The architecture is organized around three vertical layers that cover specific aspects along the data processing chain and two horizontal layers that are responsible for cross-cutting issues that have impact on the vertical layers. Everything is built upon a common infrastructure. Similarly, to the Big Data Value Reference Model the vertical layers do not imply the adoption of a layered architecture model since data can be served to consumers/client applications without being processed.

The purpose of each layer is here briefly described:

- **Data Providers:** introduces new data into the BD4NRG Data system for discovery, transformation, analysis, and access. The data can come from different sources such as static files (e.g. Comma Separated Values or JavaScript Object Notation), social media, third-party services as well as sensory data from intelligent devices. Raw data is – within this layer – properly processed to create the so-called golden datasets. These datasets are unique data, including all the necessary meta-data, the necessary quality and following the created interoperability standards to be ingested by the BD4NRG data system.
- **Data Ingestion & Management:** where the data starts their journey i.e. how data can be delivered and distributed within the BD4NRG data system for immediate use or storage. Two main mechanisms are considered, namely: real-time streaming ingestion (for data in motion) and batch ingestion (for data at rest). The specific data ingestion mechanism strictly depends on the application use case.
- **Data Access & User Interaction:** facilitates access and navigation of the data, allowing to the generation of different business views to be represented by the data. At the center of this

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

layer lies a high-performance real time analytics database to support interactive dashboards and analytics platforms while being used as data warehouse. A query engine is also part of the layer and serves as Rapid Application Development (RAD) by presenting fast logical views of data to consumer applications.

- **Data Governance:** all considered data will be channelized through data governance processes to ensure data stewardship, data classification, data transparency, data discovery, data ownership and data quality by adding business and technical metadata.
- **Security & Trust:** is responsible to support and maintain security and trust beyond anonymization and privacy. Data access policies are here defined and enforced as well as dedicated monitoring functions will be also deployed at infrastructure level.
- **Infrastructure:** is the core infrastructure of the BD4NRG data system. The layer deals with networking, computing and storage needs to ensure that large and diverse formats of data can be stored and transferred in a cost-efficient, secure and scalable way. The key requirement of any Big Data storage is that it can handle very massive quantities of data and that it keeps scaling with the growth of the organization, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to applications.

Within each layer a set of high-level functional components and features have been envisioned and connected to the work to be carried out of each Wp3's task. These components are deeply presented and documented in this document.

1.3 Structure of the document

This document is organized per task efforts in the following manner:

Chapter 2 analyses the efforts behind Task 3.1 regarding the DLT & Smart contracts for B2B cross-stakeholder trusted off-chain data sharing and re-use. Chapter 3 presents the functional and early technical approach for the Data Access Policy Brokerage Components (T3.2). Chapter 4 discusses the functional requirements for the Legal, Regulatory, Privacy and Cybersecurity Management & Compliance Tools in relation to T2.5 and the relevant deliverable D2.4 BD4NRG Security, Privacy, Standards & Regulatory Compliance Specs. Chapter 5 presents WP3 Governance layer Data Quality Compliance & Certification Tools approach.

Chapter 5, 6, 7 & 8 build upon the previous chapters and present a unified approach for

- Interoperability & Homogenization.
- Seamless Elastic Distributed Data Management and HTAP (Heterogeneous Data Capture, Cleansing, Integration, Polyglot Persistence).
- Integrated Querying of Streaming Data and Data at Rest.
- Automated Parallelization of Data Streams and Intelligent Data Pipelining.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

2 DLT & Smart contracts for B2B cross-stakeholder trusted off-chain data sharing and re-use

2.1 Blockchain platform/distributed ledgers: a short intro

Blockchain has been considered a disruptive technology, just like it was Internet, promising innovation in the commercial and financial area which is comparable to the impact the web has had on communication (Could Blockchain Have as Great an Impact as the Internet?). It is revolutionising the way we interact based on these main key advantages:

- **Traceability and data storage**: decentralised and distributed system that becomes a secure way to track changes in information over time.
- **Trust**: the creation of trust among untrusted participants.
- **Peer-to-peer transactions**: the absence of intermediaries promotes democracy.

In the energy sector, decentralization, decarbonization, and digitalization are the three key pillars of the global energy transition. Blockchain technology could help to address the challenges faced by future energy systems. Many use cases can be applied in the field of IoT and P2P trading, and thanks to the advanced cryptography that characterises this technology, it can solve the problems of cybersecurity and privacy preserving management, already discussed in the previous section. Coupled with smart contracts, blockchain can enable novel business solutions. The concept of smart contracts, first described by Nick Szabo long before the emergence of blockchain technology, denotes "a set of promises, specified in digital form, including protocols within which the parties perform on these promises". In their simplest form, for example, they are used to transfer assets to a buyer once the payment is made — comparable to a vending machine. When combined with blockchain, the rules set out in smart contracts must be fulfilled for a change of state to be triggered within the blockchain. This is the idea behind Ethereum, a platform that provides a blockchain with an integrated 'Turing-complete' programming language, meaning that it is suitable for any smart contract or transaction that can be defined mathematically.

The disruptive power of this technology, which upsets existing paradigms, goes much faster than legislative adaptation for use. However, one must be aware that the possibility that a premature regulatory framework outlined during these stages of development might limit the innovative potential of such technologies or even incur in the case of creating inefficient or inadequate legislative instruments. European Commission is taking the first steps towards blockchain regulation. The Commission on 24 September 2020 adopted a package of legislative proposals (EC communication - Digital finance package, s.d.) for the regulation of crypto-assets, updating certain financial market rules for crypto-assets, and creating a legal framework for regulatory sandboxes of financial supervisors in the EU for using blockchains in the trading and post trading of securities. Both energy generation and blockchain technology are enabling more democratic and decentralized markets. Future power networks will host more decentralized and new components with increased autonomy. The increased use of electric vehicles is expected to increase the complexity of future power systems and markets. Blockchain technology is a perfect match for future power markets and grids which develop decentralization at the core.

However, in the BD4NRG project the blockchain development will be used mostly to validate and secure data for the trusted data sharing service. This will be possible with an intensive usage of smart contracts.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The term smart contract was introduced in 1994 by Nick Szabo, where he first described how the computer-based execution of contracts between two parties could be secured without requiring any third party: “A set of promises, including protocols within which the parties perform on the other promises. The protocols are usually implemented with programs on a computer network, or in other forms of digital electronics, thus these contracts are ‘smarter’ than their paper-based ancestors”. In the blockchain context, it is generally related to computer code that is stored on a blockchain and that can be accessed by one or more parties. These programs are often self-executing and make use of blockchain properties like tamper-resistance, decentralised processing, and so on. In this interpretation, used for example by the Ethereum Foundation, a smart contract is not necessarily related to the classical concept of a contract, but can be any kind of computer program. It is called a ‘contract’ because the code that runs on Ethereum can control valuable things like ETH, currency notes or other digital assets (Learn about Ethereum, s.d.).

The possibilities are endless for smart contracts. They can be used to code and automate business processes that can be shared and executed among multiple parties offering increased trust and reliability in the process, often with significant gains in efficiency and cost reduction. Smart contracts can also be used to hard-code agreements between parties involving value and other types of asset transfer and allow them to be very transparent and run automatically based on predetermined rules, making it impossible for a party to back out. Smart contracts are used for tokenisation. A token is the digital representation of an asset on the blockchain. This asset can be both digital or physical, as well as tangible or intangible. The most important platform for the generation of tokens today is the Ethereum blockchain.

Smart contracts can directly manage payments between two (or more) actors on a blockchain, being completely self-enforcing. It is possible to transact cryptocurrency or tokens. A cryptocurrency is a digital currency that uses encryption techniques to secure and verify its transactions, while the term token usually indicates a digital representation of a specific asset. Unlike a currency, a token is not native to a blockchain, but it is created on top of a blockchain, and it is governed by a smart contract.

The value of the tokens depends on the value of the underlying assets and services that the tokens represent. Two main categories of tokens are defined on Ethereum: ERC-20 (ERC-20 Token Standard, s.d.) standard for Fungible Tokens (FTs) that are inter-changeable and not-unique and so can be used to represent a value like a currency note, and ERC-721 (ERC-721 Non-Fungible Token Standard, s.d.) standard for Non-Fungible Tokens (NFTs), representing a unique asset like a collectable good. One of the first NFT projects to gain significant traction was CryptoKitties (CryptoKitties, s.d.), a game developed on Ethereum that allows players to collect, breed and trade virtual cats.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

2.2 Smart Energy cross-domain notarization

2.2.1 Hybrid data sharing description

The blockchain data sharing structure is a collection of blocks storing a list of valid transactions and the hash of the previous block (Figure 2). The linked structure makes tamper attempts evident since any changes in one of the already registered blocks will break the hash contained in the following blocks. Each transaction must have a valid cryptographic signature and, since the blocks are timestamped and ordered chronologically, this means that it is possible to assure the provenance property by tracking entries at the moment of their registration in the blockchain.

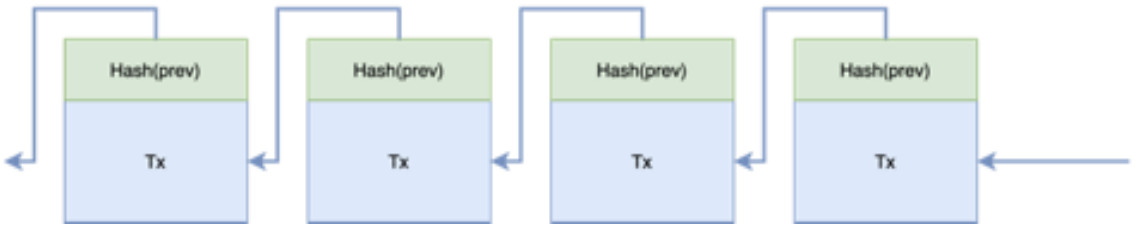


Figure 2 - Blockchain data structure

The blockchain network for data sharing energy transactions can be modelled as a private permissioned network in which each prosumer is associated with a node of the blockchain network. Depending on the resources available and the degree of trust between peers within the network, different network configurations can occur. Optionally it will be possible to rely one public blockchain according to the specific context needs.

Nodes can be light or full, depending on their computational power and data storage capacity (e.g., nodes 2 and 5 in Figure 3). Moreover, some prosumers may decide to delegate the control of transactions to another node they trust, and not use their own node.

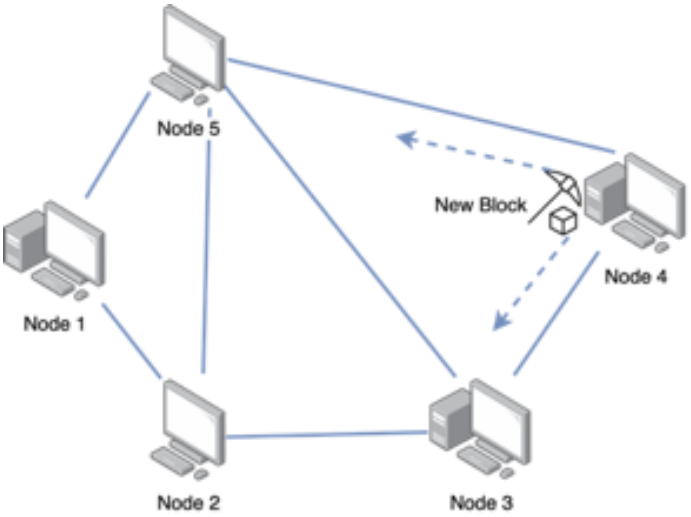


Figure 3 - Blockchain Network

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Our solution will combine the on-chain data storage together with distributed databases for off-chain data storage according to a hybrid approach. In order to minimize the amount of data to be stored on the blockchain, raw data from the data sources will be stored using a distributed database: the blockchain will be then used periodically to store a fixed-length hash of the data received in the last time interval. The hybrid approach combines the scalability of a distributed NoSQL database with the blockchain, making tampering attempts evident, enabling provenance tracking, non-reputability, and the use of self-enforcing smart contracts.

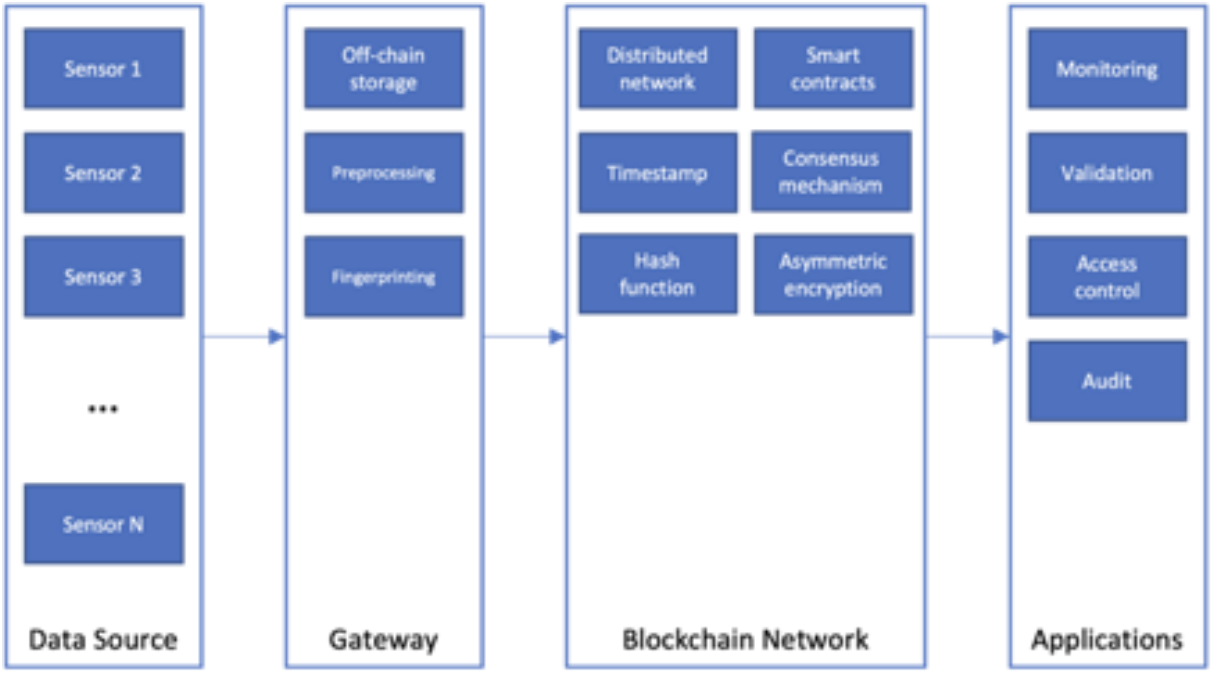


Figure 4 - Data Flow

Each data source (Figure 4) will be connected to a gateway, an embedded device which takes care of pre-processing, calculating digital fingerprints (hash codes) to assure data immutability via notarization, and storing the data off-chain. The possibility of handling data from the Data Storage is also an option. A private network based on Ethereum will be used as blockchain network. The Ethereum nodes provide the networking capabilities and the rules needed to determine consensus among them. The private network will be used as a platform for running smart contracts and register data with a timestamp.

Thanks to dedicated APIs, every node in the network can be queried and it is possible to build dedicated applications on top of this stack for monitoring the process, validating the off-chain data against the data stored on-chain, provide limited access for auditing purposes, or provide even finer-grained access to specific users for specific purposes, interacting with smart contracts deployed in the network.

2.2.2 Technological components

The proposed solution will be based on Ethereum in term of blockchain consensus mechanism, in addition, connection with IoT devices will be implemented via MQTT, a lightweight protocol suitable for unstable connections and low-power devices. Additionally, connection with NoSQL Data Storage will be supported for off-chain data handling.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Ethereum

Blockchains can be distinguished according to the consensus mechanism and programming capabilities. Considering the consensus mechanism, blockchains differ in the definition of the participation of nodes in the distributed network and the roles that they can perform. In particular, open blockchains and permissioned (or private) blockchains can be distinguished. Considering the programming capabilities, it is possible to differentiate between blockchains programmable via simple scripting and blockchains providing Turing-complete computational capabilities, enabling the creation of "smart contracts". Ethereum was the first blockchain supporting smart contracts and it is still the most notable example of Turing-complete programmable blockchain.

Ethereum was proposed by Vitalik Buterin in 2013 (Ethereum Whitepaper, s.d.) and further detailed by Gavin Wood in the 'yellow paper' (Ethereum: A Secure Decentralised Generalised Transaction Ledger (Wood, s.d.)). As the Ethereum website (Ethereum, s.d.) reports, "Ethereum is a decentralized platform that runs smart contracts." These contracts run on the "Ethereum Virtual Machine" (EVM); a distributed computing network made up of all the devices running Ethereum nodes. Like other blockchains, Ethereum has a native cryptocurrency called Ether (ETH) and it has a double use: it is used as an incentive for the network "validators", but also to regulate the use of the blockchain computational resources.

Each operation on the Ethereum network has its own cost, determined by the computation, storage and bandwidth required, and this cost is measured in a unit called Gas. gasLimit (the maximum amount of gas that can be spent) and gasPrice (the price-per-gas) are standard transaction parameters in Ethereum and also invoking a smart contract function both must be specified. Since smart contracts are run on a virtual machine, the presence of the gas price and limit have the purpose of preventing denial-of-service attacks. More specifically, is impossible to run infinite loops (due to the gas limit) and the non-optimised usage of system resources would be expensive due to the gas price. Gas and Ether are different concepts. Ether is a currency, while Gas is an internal transaction pricing mechanism that measures the amount of computational effort that it will take to execute certain operations (each line of code of an application requires a certain amount of gas to be executed in the Ethereum network). Gas is always paid in Ether. Having a separate unit allows maintaining a distinction between the actual valuation of the cryptocurrency (Ether), and the computational cost (Gas).

MQTT

MQTT is an OASIS standard messaging protocol for the Internet of Things (IoT). It is designed as an extremely lightweight publish/subscribe messaging transport that is ideal for connecting remote devices with a small code footprint and minimal network bandwidth. MQTT today is used in a wide variety of industries, such as automotive, manufacturing, telecommunications, oil and gas, etc.

2.3 Solution Deployment approach

The section 3.5.1 describes the role of the gateways in enabling IoT devices for blockchains. These are embedded devices (or similar, low-power devices) operating at the edge level connecting and managing the different IoT devices (Figure 5).

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Connection with IoT devices will be implemented via MQTT, a lightweight protocol suitable for unstable connections and low-power devices. The gateway periodically will receive information about the status of the devices connected, pre-processes input data as needed, and store them to a distributed database in the cloud, using an asynchronous communication queue.

The gateway periodically calculates a fingerprint of the data received from the device and stores the result on the blockchain.

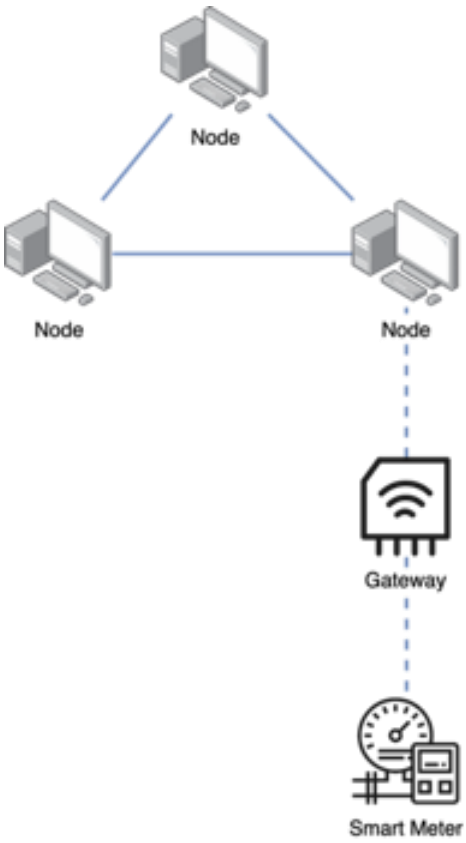


Figure 5 - IoT Gateway

MQTT communication will be authenticated and each IoT device (Figure 6) assigned a gateway and a specific topic for its messages, so it will be possible to keep track of the origin of the information received.

Blockchain transactions will be signed with the private cryptographic key of the gateway that sent them, so that each reading will be associated in the smart contract with the corresponding gateway.

Off-chain data consistency verification will be performed by calculating on-the-fly the fingerprint of the data to be verified and comparing it with the fingerprint stored on the blockchain for the same time interval. If the two fingerprints match, the information is unaltered; if not, there may be a possible manipulation or malfunction.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

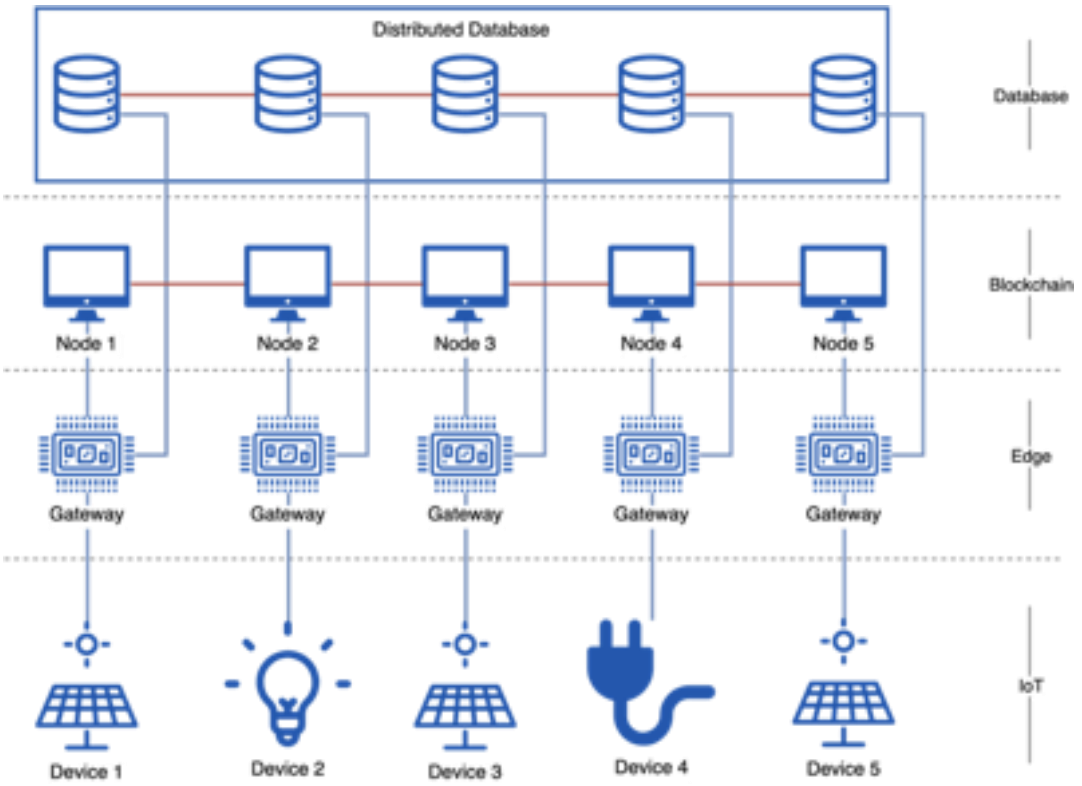


Figure 6 - Platform Architecture

2.4 Interaction with other BD4NRG ecosystem components

Within the Data Service layer, the Hybrid data sharing module will interact mainly with:

- The Data Storage for the extraction of harmonized data in order to calculate related fingerprint.
- The Blockchain network (Ethereum) for the physical fingerprints (hash codes) storage in the chain blocks to assure data immutability.
- This will be the main mechanism to assure immutability of data stored into the Pilots Data Storage.

Optionally the Hybrid data sharing module could interact directly with the BD4NRG assets, and external data providers in case of direct raw data handling; in this case the interaction will require specific data connectors to integrate the datasets with different standard and access interfaces.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

3 Data Access Policy Brokerage

3.1 BD4NRG Data Access Policy Brokerage Functionality

The central idea behind the Data Access Policy Brokerage is the development and enforcement of mechanisms for the application of Data Access Policies in the BD4NRG datasets. These mechanisms will be used by the upper layer's BD4NRG components specifying user access rights and rules based on:

- Volume of data access
- Frequency of data access
- Request Device (e.g., laptop, smartphone etc.)
- Person or Group Role (e.g., location, IP, role in the company etc.)
- Historical data regarding access patterns (e.g., frequency, usual dates or hours of access, usual duration of access, previously accessed data etc.)
- Access Control Type (MAC, DAC or RBAC/ABAC)

In more detail the Data Access Policy Brokerage component will be responsible for:

- The development of appropriate context model for semantically describing associations between types of access depending on the data objects and circumstances under which this access should be allowed
- Based on the aforementioned context model the creation, management and maintenance of BD4NRG Data Access Policies
- Enforcement of DAPs through specific User Access Rights
- Support of the User Access Rights administration through specific software components which will provide the following functionalities:
    - Identity and Access Management (IAM) for users, groups and devices management.
    - Policy Enforcement Point (PEP) Proxy which will carry out or enforce policy decisions.
    - Policy Decision Point (PDP) which will evaluate and issue authorization decisions.
    - Policy Administration Point (PAP) which will manage access policies.
    - Data Usage Control.
    - DAP Annotation Management.
    - Validity checking of DAP annotations (e.g., spotting inconsistent annotations).
    - Tracking dependencies between DAP annotations.
    - Ensuring compliance with pertinent organizational policies, especially with respect to the available encryption and data distribution strategies.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

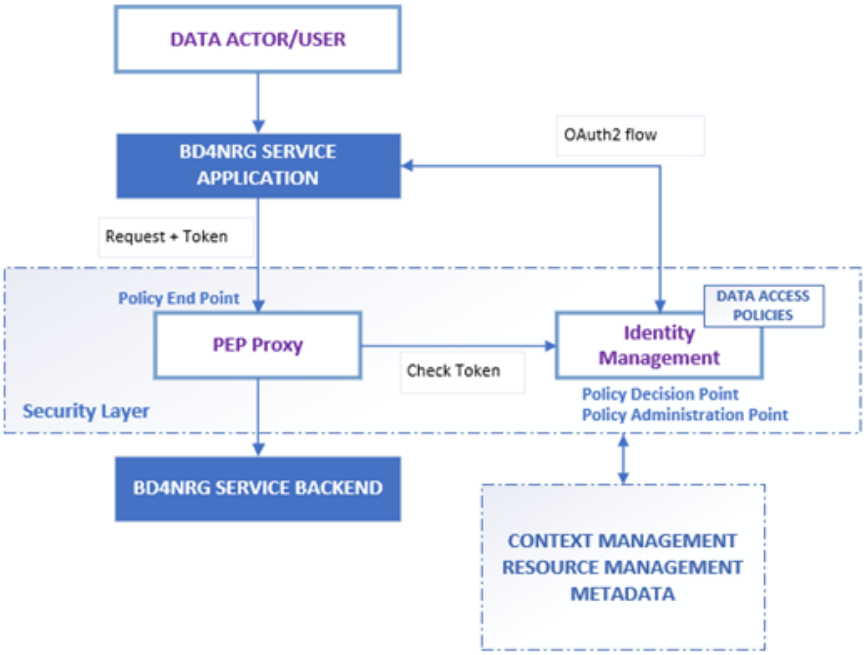


Figure 7 - BD4NRG Governance Layer Data Access Concept Architecture

In regards to BD4NRG context the Data Access Policy Brokerage is responsible for:

- Users, devices and groups management
- OAuth 2.0 - based Single Sign On
- Application - scoped roles and permissions management
- Support for local and remote PAP/PDP
- JSON Web Tokens (JWT) and Permanent Tokens support
- Securing service backends
- Basic and complex AC policies support
- OAuth 2.0 Access Tokens support
- PDP configuration

3.2 FIWARE Approach

FIWARE [1] is an open-source initiative defining a universal set of standards for context data management which facilitates the development of Smart Solutions for different domains such as Smart Cities, Smart Industry, Smart Agrifood, and Smart Energy.

In any smart solution there is a need to gather and manage context information, processing that information and informing external actors, enabling them to actuate and therefore alter or enrich the current context. The FIWARE Context Broker component is the core component of any “Powered

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

by FIWARE" platform. It enables the system to perform updates and access to the current state of context.

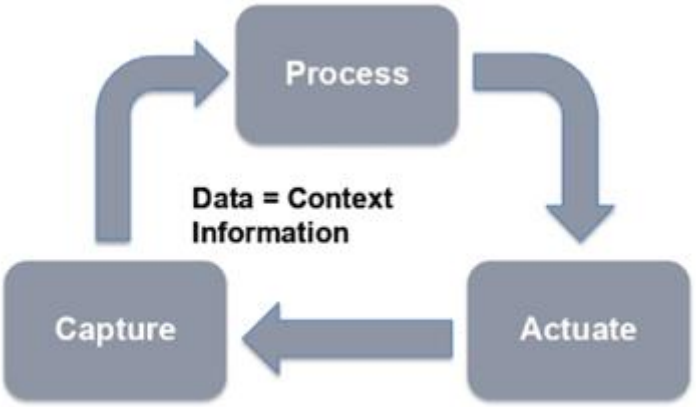


Figure 8 - The FIWARE Approach

The Context Broker in turn is surrounded by a suite of additional platform components, which may be supplying context data (from diverse sources such as a CRM system, social networks, mobile apps or IoT sensors for example), supporting processing, analysis and visualization of data or bringing support to data access control, publication or monetization.

This section presents an implementation of the proposed above architecture using the generic enablers provided by FIWARE. Specifically, the GEs used in this implementation of the data usage architecture are the following:

The **Keyrock Identity Management GE** [2] will be responsible for Identity Management. Using Keyrock enables OAuth 2.0-based authentication and authorization security to services and applications. In the context of this implementation, Keyrock will play the role of IdP, which will manage authorization policies (PAP) and will decide which Data Consumers will be able to access which resources in the data infrastructure (aPDP). Therefore, Data Providers and Data Consumers will perform the authentication process relying on Keyrock. Guaranteeing the unequivocal identification of all the agents involved in the data usage architecture is mandatory to ensure a secure way of providing or consuming data. By using Keyrock, Data Providers will implement authorization policies to constrain Data Consumers access to the data infrastructure.

The main identity management concepts within Keyrock are:

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- Users
  - Have a registered account in Keyrock.
  - Can manage organizations and register applications.
- Organizations
  - Are group of users that share resources of an application (roles and permissions).
  - Users can be members or owners (manage the organization).
- Applications
  - has the client role in the OAuth 2.0 architecture and will request protected user data.
  - Are able to authenticate users using their Oauth credentials (ID and secret) which unequivocally identify the application
  - Define roles and permissions to manage authorization of users and organizations
  - Can register Pep Proxy to protect backends.
  - Can register IoT Agents.

Keyrock plays the role of PAP (Policy Administration Point) and PIP (Policy Information Point) in architecture. It sets rules for data usage. Keyrock will provide both a GUI and an API interface.

Software Requirements	https://fiware-idm.readthedocs.io/en/7.4.0/#software-requirements
How to build and install	https://fiware-idm.readthedocs.io/en/7.4.0/#how-to-build-install
API Overview	https://fiware-idm.readthedocs.io/en/7.4.0/#api-overview

The **Wilma proxy GE** [3] brings support of proxy functions within OAuth 2.0-based and XACML-based access control authentication schemas. Wilma can be combined with other security components such as Keyrock and Authzforce to enforce access control to backend applications. This means that only permitted users will be able to access Generic Enablers or REST services. Identity Management allows to manage specific permissions and policies to resources allowing different access levels for users.

In the scope of this implementation a Wilma instance will be in charge of enforcing access policies over requests sent to the data infrastructure. When a Data Consumer will be authenticated through Keyrock, an OAuth 2.0 token will be generated, which will be included in every request sent to the Data Provider's data infrastructure. Wilma will intercept requests and asks Keyrock to validate the token, verifying the Data Consumer's identity. Since Keyrock will also act as the aPDP, it will check the Data Consumer's access authorization policies. In case that the Data Consumer's request complies with the established policies, Wilma will grant access to the requested resource.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Admin Guide	https://fiware-pep-proxy.readthedocs.io/en/7.4.0/admin_guide/
How to build and install	https://fiware-pep-proxy.readthedocs.io/en/7.4.0/#how-to-build-install
API Overview	https://fiware-pep-proxy.readthedocs.io/en/7.4.0/#api-overview

3.3 IDSA/FIWARE Integration

Data Spaces must provide means for guaranteeing organizations joining Data Spaces that they can trust the other participants and that they will be able to exercise sovereignty on their data. That requires the definition of common Building Blocks, based on mature security standards that will be used by all participants in the Data Space.

A first fundamental building block to support within Data Spaces has to do with Identity Management (IM). This building block allows identification, authentication, and authorization of organizations, individuals, machines and other actors participating in a Data Space. While this building block can be implemented on the basis of third-party Open-Source technologies like KeyCloak, just to mention one popular example, FIWARE brings the Keyrock component which supports OpenIdConnect, SAML 2.0 and OAuth2 standards. Quite relevant for Data Spaces deployed in Europe, Keyrock also resolves integration with eIDAS, a building block provided by the European Commission that enables the mutual recognition of national electronic identification schemes (eID) across borders, allowing European citizens to use their national eIDs when accessing online services from other European countries.

A second fundamental building block will be the one which facilitates trusted data exchange among participants, providing certainty that participants involved in the data exchange are who they claim to be, and that they comply with defined rules/agreements. Trust refers to the fact that data providers and data consumers can rely on the identity of the members of the data ecosystem and beyond that, between different security domains. Here, IDS Connector technology, as described in the IDS Reference Architecture Model (RAM) emerges as a solid basis and the FIWARE Community has incubated an Open-Source implementation of this technology that has already been tested on how it can integrate with the rest of core FIWARE components.

Figure 9 illustrates how an AI service provider application and an AI service consumer application, each hosted in different organizations and participating in a common Data Space “powered by FIWARE”, may interact with each other. It shows the role of components that are part of the incubated FIWARE IDS Connector implementation and their integration with other FIWARE components.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

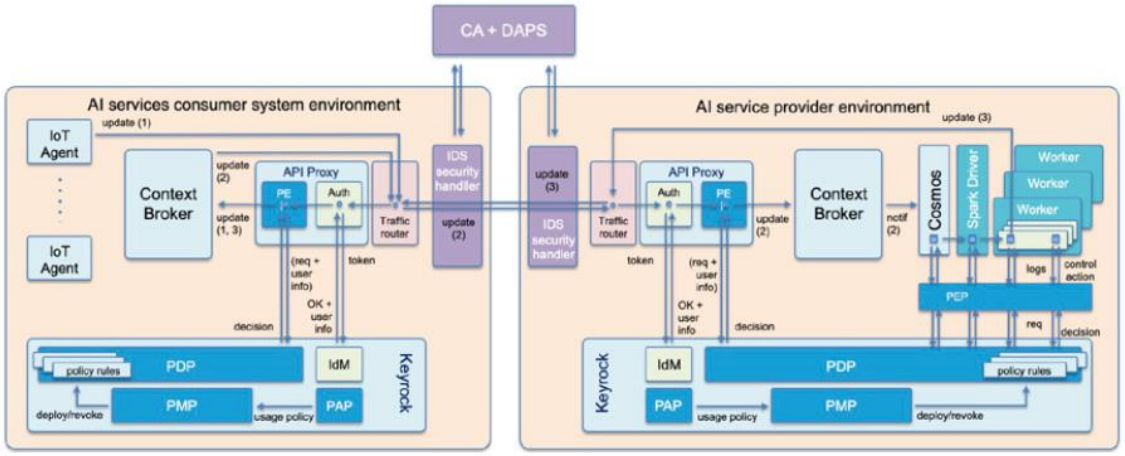


Figure 9 - Integration of FIWARE IAM and IDS Connector functionalities

In FIWARE, the implementation of IDS Connectors [4] comes closely linked to the deployment of Kubernetes clusters. In this respect, the implementation of the concept of IDS connectors would not be limited to the implementation of the procedures used to exchange data among IDS connectors but also the monitoring and control of what is deployed in the associated clusters (e.g., guaranteeing that only certified components/applications are deployed) as well as the control of the flow of data within the cluster, ensuring enforcement of data access/usage control policies. For this last purpose, the usage of products supporting service mesh concepts like Istio are essential. In figure 9, such components are identified as “traffic routers”.

In the IDS RAM, identity and access/usage control is currently performed at organization level. The Certification Authority (CA) and Dynamic Attribute Provisioning Service (DAPS) are in charge of verifying that organizations exchanging data are known and trustworthy. The IDS Connectors ensure that data exchanges are authorized at organization level.

On top of this authentication and data access/usage control procedures, performed at organization level via IDS Connectors, FIWARE security components enable an additional and finer grained authentication and access control at the end user level.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

4 Legal, Regulatory, Privacy and Cyber-security Management and Compliance Tools

The evolution of the energy systems is characterised by two main factors: first, the wider deployment of distributed renewable energy sources; and second, the massive integration of connected devices. Indeed, the increasing installations of renewable energy source is responding to the need of ecological sustainability of the power generation with consequent reduction of greenhouse gas emissions, in line with the United Nations Sustainable Development Goals and with a wider awareness from citizens and governments on the climate change matter. However, the renovation of the energy generation/transmission/distribution chain is requiring at the same time the transformation of the grid management based on an updated information to promptly and accurately ensure quality of service and operational constraints (including inter-alia reliability and robustness). For this reason, the energy infrastructures are evolving towards the integration and potential convergence of physical operational technology and cyber information technology through the adoption of industrial internet and internet of things (IOT) [5], enabling continuous infrastructures monitoring and maintenance to guarantee that their day-to-day operations can continue uninterrupted by failures, as well as to mitigate potential physical and cyber threats.

In this context, data sharing – and in general data governance and management - is a key aspect to be considered in the novel smart grids in compliance with current (and constantly evolving) legal and regulatory frameworks to appropriately ensure privacy and security dimensions of the involved stakeholders.

4.1 Compliant data governance

Specifically, as shown in the above sections, BD4NRG will use a security layer to securely share data according to “BD4NRG Data Access Policies”. The set of tools described in this section, i.e. “Privacy, Legal, Regulatory and Cyber-security Management and Compliance Tools” (from now on called with the acronym “PRECYSE Tools”), is responsible to assure that all legal, regulatory and cybersecurity guidelines and principles are incorporated in the design and development of the Data Governance and Management layers.

To achieve this goal (i.e. compliant data governance), the PRECYSE Tools are built on top of the BD4NRG Legal and Security Conceptual Framework, defined in D2.4 “BD4NRG Security, Privacy, Standards & Regulatory Compliance Specs”, and interact with the security layer as shown in Figure 10.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

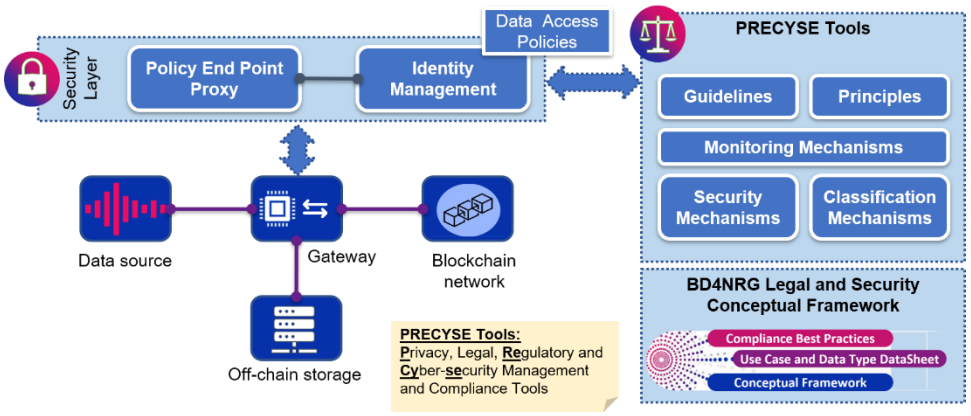


Figure 10 - The PRECYSE Tools and their interactions with other BD4NRG components

The PRECYSE Tools make available **guidelines and principles** from the Legal and Security conceptual framework addressing and mitigating potential privacy and security concerns. These guidelines and principles will be used to define and manage the data access policies.

As reported in the ethics requirements (D11.1), the BD4NRG partners and components will **not** share any data referring – or referable – to individuals (i.e. personal data as defined in GDPR [6]).

Based on the ethics requirements, BD4NRG has performed a preliminary analysis of the BD4NRG use cases, and from this analysis emerged that only 33% is in a medium-high risk zone, i.e. there are data sources potentially containing personal data but this will not be shared by applying opportune anonymisation/pseudonymisation techniques.

However, as per best practices, it is important that the PRECYSE Tools ensure continuous **monitoring** with appropriate mechanisms in order to guarantee that conditions are not changing and infringing any of the guidelines and principles. For this purpose, the BD4NRG Legal and Security Conceptual Framework suggests the adoption of the assessment checklist during the development process.

A special focus will be devoted to the interactions among data source, gateway and blockchain network. Indeed, for any reason it will **not** be allowed to store any data potentially referring to individuals, due to the fact that data erasure cannot be performed, infringing the GDPR art.17, as well as the data cannot be rectified (infringing the GDPR art.16). According to the current specifications, gateway will store references to original data stored in off-chain storage. These references are the results of hashed bunch of data.

4.2 Component Assessment

This BD4NRG Component Assessment Checklist is a means of monitoring and evaluating potential privacy concerns in a specific component of the BD4NRG architecture.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Table 1 - BD4NRG Component Assessment Checklist

Component Name	*<Identifier – the name of the component>*		
Component Owner	*<Identifier - the partner who develops the component and is accountable for its behaviour>*		
Assessment Date	*<Date – the assessment date>*	**Assessment Operator**	*<Identifier – the development partner who performs the assessment >*
Data	**Data Source(s)**	*<List of identifiers – the source(s) of data used by the component >*	
	Data Owner(s)	*<List of Identifiers - the partner(s) who own the data source(s) >*	
	Personal Data	*<List of Values - Options: {Yes / No / NA / Not Sure} >*	
Protection	**Purpose**	*<List of Values - If data contain personal data, specifies the purpose of treatment>*	
	Protection Mechanisms	**<**List of Values - If data contain personal data, specifies the designed mechanisms to be applied options: {Anonymisation, Pseudonymisation, Aggregation, Minimisation, Other-Specify} > <In case of uncertainty, please, contact immediately ELEX and Coordinator>	

The adoption of this component assessment checklist allows the BD4NRG team to promptly identify any potential concern during the whole development lifecycle and a contact with the ELEX and the BD4NRG Coordinator to establish appropriate countermeasures. This allows to mitigate and address any potential concerns.

At the same time, **security mechanisms and classification mechanisms** will be identified during the next development phases to identify potential security concerns

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The monitoring mechanisms, both for privacy and security concerns, and the assessments will allow to evaluate and trigger updates on the Data Access Policies, taking into account the foundations laid by the BD4NRG Legal and Security Conceptual Framework (see Figure 11) and ensure their compliance.

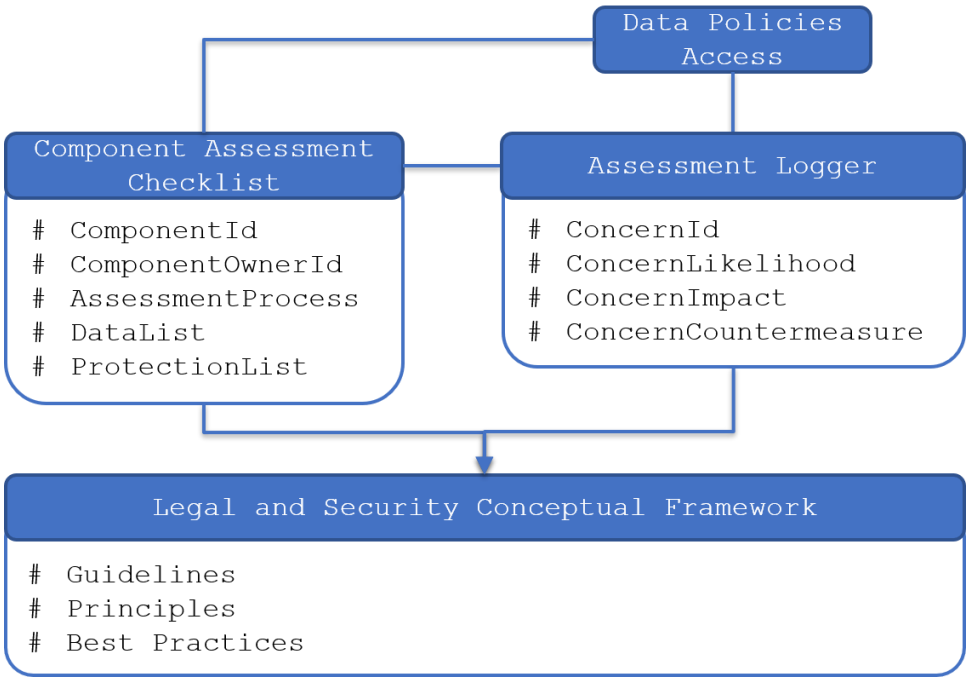


Figure 11 - Compliant Data Policies Access

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

5 Data Quality Compliance and Certification Tools

The present chapter summarizes some of the most outstanding options and mechanisms available as of today to pursue the compliance with standards of data quality that assure the shared information is trustable and valid, as well as the tools to conduct a proper certification process over the components that provide such data.

5.1 Data quality requirements

Data quality is important regarding the compliance of both qualitative or quantitative pieces of information. There are many definitions of data quality, but we can generally accept that high quality refers to the ability of data to fit its intended uses in operations, decision making and planning.

Data quality deals with many variables, like data structure, format and correct representation the real-world construct to which it refers. Furthermore, apart from these definitions, as the number of data sources increases, the question of internal data consistency becomes significant, regardless of fitness for use for any particular external purpose.

Going into the Data Quality field, even though there is no general agreement on all the properties defining the quality of data, it is indeed possible to provide certain metrics that would help to assess data. Such list could include, and not be restricted to, the following metrics (as suggested in sources such as [7]):

- **Completeness** helps to provide a measure on whether all the necessary data is present in a specific dataset or not. An example metric for completeness is the percent of 1data fields that have values entered into them.
- **Accuracy** focuses on how data reflects the real-world object it refers to. An example metric for accuracy could be to find the percentage of values that are correct compared to the actual value.
- **Consistency** is basic to ensure data remains consistent on a daily basis. An example metric for consistency is the percent of values that match across different records/reports.
- **Validity** contributes to check how well data conforms to required value attributes. An example metric for validity consists in finding the percentage of data that present values within the domain of acceptable ones.
- **Timeliness** reflects the accuracy of data at a specific point in time. An example metric for timeliness is the percent of data you can obtain within a certain time frame (e.g. weeks or days).
- **Integrity** relates to how all the data quality metrics already mentioned are maintained as data moves between different systems. Hence, a potential useful metric for integrity would be the percent of data that is the same across multiple systems.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Figure 12 - Metrics to check Data Quality

Another important issue to consider regarding quality of data is security. Before proceeding with any data exploitation, is of the essence to understand that quality and security represent two basis aspects that add value to data and their implementation becomes a real need. Thus, Data Quality and Data Security may be considered as key topics that oppose different challenges in Big Data such as high volumes and heterogeneity of data, credibility of data and their sources, the speed at which data is collected and processed, etc.

More specifically, when dealing with the infrastructure of distributed energy resources (DERs) it is crucial to consider security at the device, network, or application levels of DERs.

The lack of in-place security controls for critical infrastructures led to many cyber-physical attacks during the past 5 years [8]. These recent cyber-physical attacks speak to the urgency of securing all aspects of critical infrastructure.

Several industry standards and guidelines have been developed to take care of this issues. For example, the North American Electric Reliability Corporation developed cybersecurity requirements for critical infrastructure protection [9], but these requirements apply to security issues of the bulk power system and are not applicable to cybersecurity issues of DERs.

Similarly, the National Institute of Standards and Technology (NIST) has developed a cybersecurity framework [10] that instructs organizations how to develop processes to manage cyber risk to a system. The cybersecurity framework, however, does not address the cybersecurity risks of DERs and their communications with the grid.

NIST also developed many other standards, including:

- NIST SP 800-53: Security and Privacy Controls for Federal Information Systems and Organizations NISTIR 7628: Guidelines for Smart Grid Cybersecurity
- NIST SP 800-82: Guide to Industrial Control Systems Security.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

In addition to the North American Electric Reliability Corporation and NIST, other organizations have developed security standards and guidelines for power systems, including, but not limited to:

- The International Electrotechnical Commission (IEC) 62351 standards provide security for information exchange in power systems.
- The U.S. Department of Energy developed the Cybersecurity Capability Maturity Model to provide cybersecurity benchmarks and guidance for utilities on effective risk-management processes with consideration of specific organizational requirements and constraints.
- The Institute of Electrical and Electronics Engineers (IEEE) C37.240 standards provide cybersecurity requirements for substation automation, protection, and control systems.

Existing cybersecurity frameworks address cybersecurity issues related to the distribution grid as whole, but there are not sufficient guidelines and procedures that can help vendors, utilities, aggregators, government institutions, and other industry partners adopt and implement the procedures to secure the data and communications of DERs that are connected to the distribution grid. Therefore, to create an effective and efficient way of securing next-generation DERs that will be connected to the distribution grid, there should be a standard that industry can follow.

5.2 Data sources/data assets

The main concern regarding data sources is to provide a framework for the right data type/structure that will be fed into data models. The goal is to ensure that the right data will be selected as input for operational models, creating a smooth process which decreases the amount of both: bugs and incorrect results in the system.

With that goal in mind smart data models are born, which are main trend currently regarding standards of high quality for data structures nowadays.

Smart Data Models have been designed to foster the use of a reference architecture and the development of compatible common data models that can ensure data coherence and reliability.

A smart data model includes three elements:

- The **schema**, or technical representation of the model defining the technical data types and structure
- The **specification** of a written document for human readers
- The **examples of payloads** for validation of input data based on NGSI v2-v3, NGSI

In particular and linking with the information already exposed in D2.3 and D2.5, the FIWARE smart data model initiative has mapped pilots to relevant models according to their individual data needs across different FIWARE sectors with strong focus on energy but also Agrifood, environment, aeronautics etc.)

As of today, up to 604 models in total are currently available in the Smart Data Models repository; however, FIWARE provides an agile process allowing for fast incorporation of new standards in days thanks to use cases that can be documented (there are now 97 models in preparation as “incubated”).

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

5.3 IDS Certification Scheme

In search of a certification process that could satisfy the requirements posed by a scenario such as the one envisioned by BD4NRG, the initial step consists in checking the background and knowledge within the consortium. Hence, a closer look into IDS' certification scheme is mandatory.

The goal of the International Data Spaces is to establish data spaces that guarantee data sovereignty and safe collaboration between partners. The foundation of IDS is trust, which must be established through a rigorous, transparent certification process.

The International Data Spaces certification scheme encompasses all processes, rules and standards governing the certification of participants and core components within the International Data Spaces.

The participants in the data space collaborate by sharing their valuable data. The success of this collaboration depends on trust among all parties involved, and this, in turn, is guaranteed by a transparent certification of the environment and components. This two-pronged approach provides a high level of trust and security.

Operational environment certification [11]

The operational environment includes the physical environment, defined processes and organizational rules. This evaluation is based on international standards and provides an assessment of the trustworthiness of the operational environment along two dimensions, the evaluation depth and the extent of the security requirement.

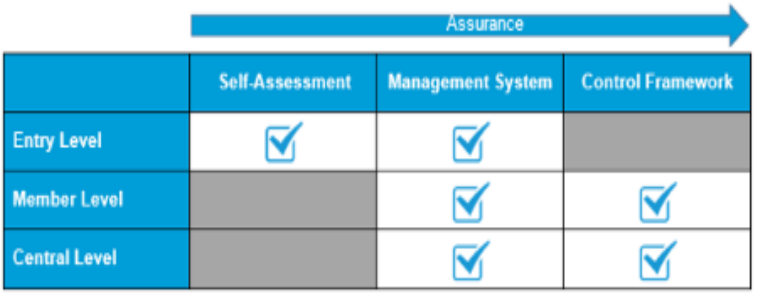


Figure 13 - Operational environment certification assurance levels

Core Component Certification [12]

We evaluate and certify core components based on whether they provide the required functionality and security regarding the security profiles. As core components are developed by different companies for a wide variety of industries and usage scenarios, core component certification differs, providing assurance levels and security profiles that may be applied based on the needs of the data provider and data consumer. The core component certification includes interoperability testing to ensure the interoperability and compatibility of all components in the IDS.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

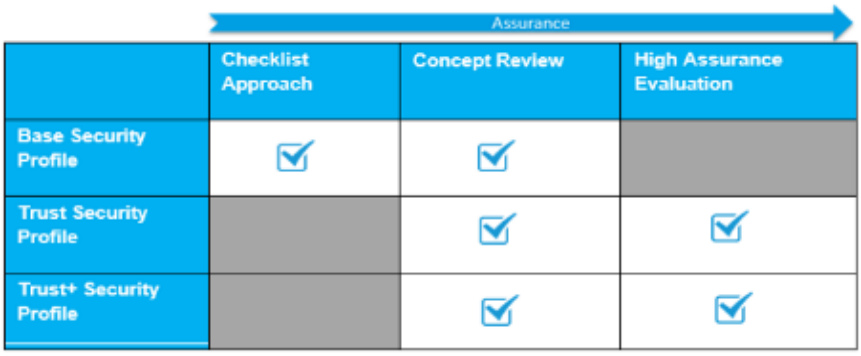


Figure 14 - Core Component certification assurance levels

The exercise that BD4NRG partners will perform in the next execution steps will imply the evaluation of how this kind of certification process may apply over the data providers implied in the LSPs and over the components that will shape BD4NRG architecture. Hence, the elaboration of a solid proposal on to what extent something similar could be put in place in this particular case.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

6 Interoperability & Homogenization

6.1 Data Integration & Homogenization sub-layer

References to the Agents and software packages to transform the direct payloads from sensor and legacy application to the reference architecture to enable the advanced analysis. This part of the architecture will be based on the FIWARE Reference Architecture and will deeply use both FIWARE and IDS ecosystems, leveraging the true connector component.

6.2 FIWARE Reference Architecture

The reference FIWARE architecture is depicted in Figure 15.

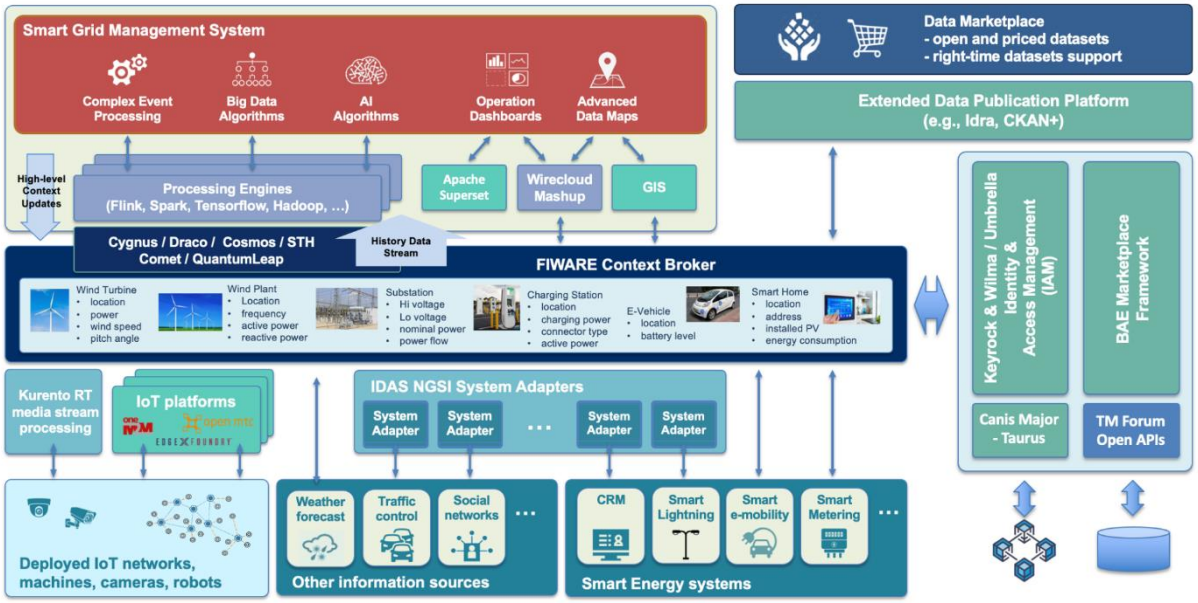


Figure 15 - FIWARE Reference Architecture

This reference architecture will have to be adopted and merged with the LSP requirements and needs to create on one hand a specific instance of the architecture integrating the legacy systems and sensor for each LSP and on the other hand to identify the core elements (e.g., context broker), that could be shared across different pilots.

6.3 Connectors (Interfaces)

As explained in the Par 3.3 specific connectors will be used also for the homogenisation aspects, one under analysis is the Engineering True Connector. This will use a Base Security Profile, build and developed using Apache Camel/Spring Boot and Info Model alignment 4.1.X.

It has been already tested the integration with other HTTP/HTTPS connectors (GEC, Boost4.0) during the different Integration Camp events and it has been also already integrated with DSC WIP.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The Connector main focus is on the Execution Core Container, as foreground of the Market4.0 project and it is a basic data APP, inside FIWARE Data APP WIP. The true connector can work in IDS Trusted Environment and it is composed by a **Sender Router**, it gets message from Data APP, adds Token received from DAPS and sends message to Receiver ECC. In addition, it registers transaction to CH and enforces response to UC.

In parallel, the **Receiver Route** gets the message from Sender ECC, validates the relevant token with DAPS and sends the message to Data APP. Then it sends response to Sender ECC and finally, registers the transaction to CH.

This process has been drafted in the following Figures 16, 17

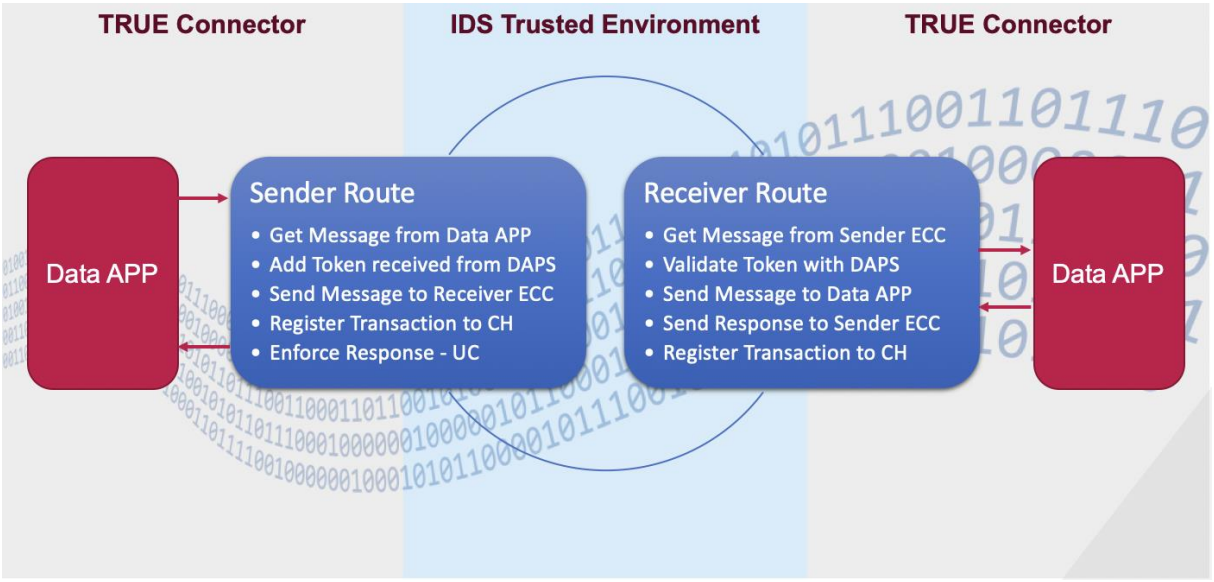


Figure 16 – True Connector

Next figure demonstrates the relationship and the positing of the True Connector inside the FIWARE and IDS ecosystem and in relation with the Context Consumer and the Context Provider.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

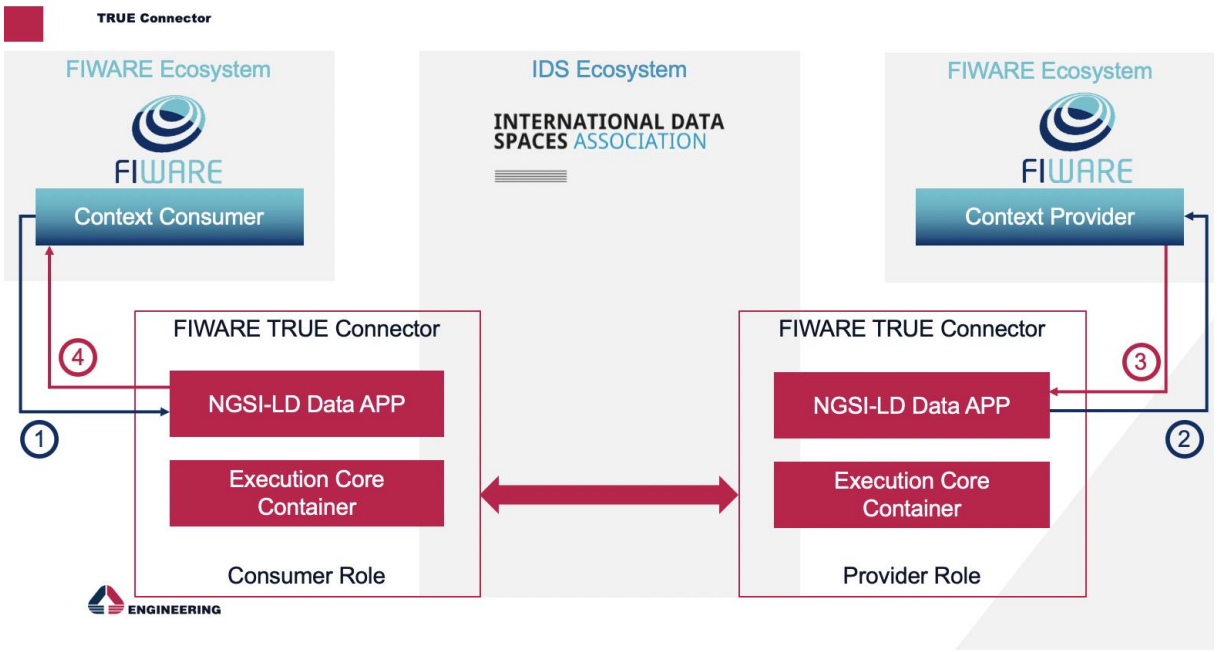


Figure 17 – True Connector

6.3.1 IDS Reference Testbed

The interoperability of the TRUE Connector with other IDS components has been tested on the IDS Reference Testbed.

The IDS Reference Testbed is a set up with Open-Source IDS components implementing the IDS specifications for establishing connections and communication. Its main uses are:

- Component behavior testing
- Interoperability testing against IDS components (Connector, DAPS, CA, Metadata Broker)
- Preparation for IDS certification
- Starting point for creation of data spaces

The testbed in Version 1.0 consists of a CA, DAPS, Meta Data Broker and two connectors as depicted in Figure 18. Other IDS and non-IDS components will continue to be integrated into the testbed, constituting different versions of it which can be applicable to different use cases.

Following the successful integration of the TRUE connector with the current version of the IDS Reference Testbed, other FIWARE components will be integrated in its subsequent versions. Following components have been planned:

- FIWARE context broker
- Keyrock

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The complete description and documentation of the IDS Reference Testbed can be found on Github in the repository https://github.com/International-Data-Spaces-Association/IDS-testbed. The available documentation includes:

- Description of IDS Reference Testbed, uses and current version
- Installation and interconnectivity guide
- User guide for the integration of external connectors
- Link to automated Test Suite to conduct automated testing

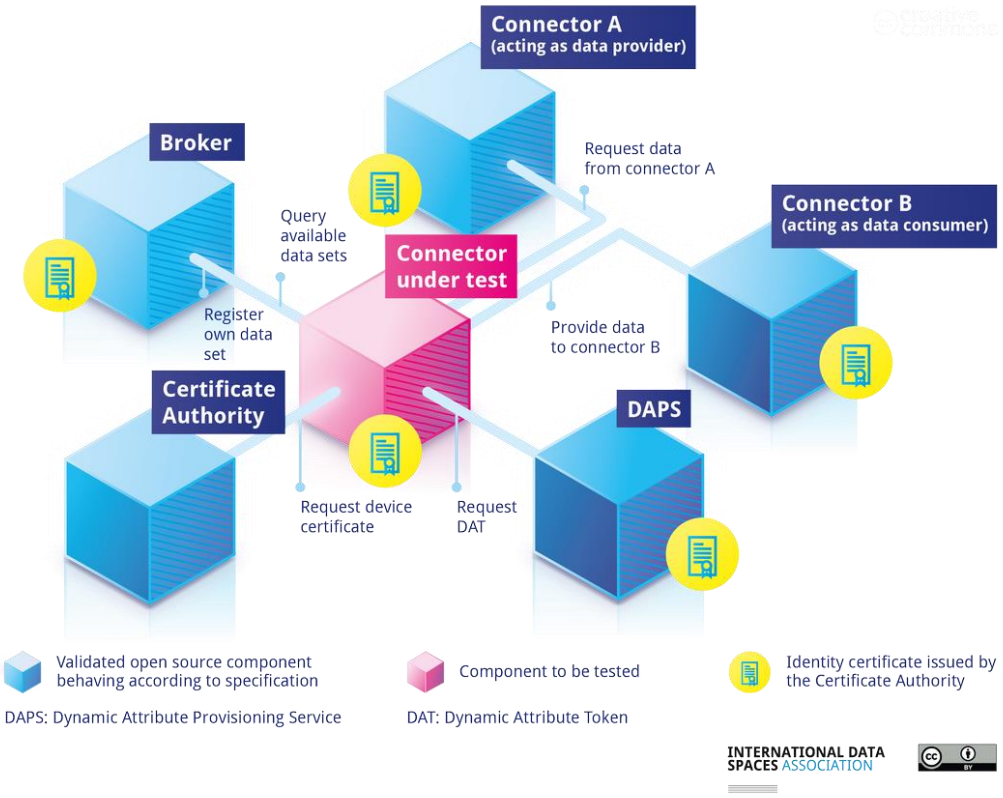


Figure 18 – IDS Reference Testbed V1.0

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

7 Heterogenous Data Capture and Adaptive Polyglot Persistence (HTAP)

7.1 Component Context in relation to the Project

BD4NRG's objectives for the component
To develop a middleware layer, on the top of the common IP infrastructure, to enable the exchange of the information between assets, systems, data hubs and actors in a common manner.
To develop of edge/IoT big data management enablers including elastic streaming capturing management

7.2 Component Overview & Features

The need to store, process and analyze data is pushing more and more the emergence of data-driven initiatives consolidated around three main pillars, namely: Business Intelligence, Artificial Intelligence and Data exchange [13]. The HTAP component is thought to work within modern data platform efficiently and effectively. So far, the overall objectives of the HTAP component are presented in Figure 1.

Overview		
Highly-heterogenous and fast amounts of data require the definition of a seamless elastic distributed data management process to ensure the storage of quality data with value to be used by all the other components envisioned within the BD4NRG solution. The overall process is materialized in the Heterogenous Data Capture and Adaptive Polyglot Persistence (HATP) component.		
Objectives & Features		
O1	Serve as a traditional data warehouse	
O2	Support for data analytics	
O3	Support for solid data plumbing from heterogenous data sources leveraging the most disparate technologies (Apache Kafka, HDFS, Filesystem, etc.)	
O4	Support for innovative data management techniques to ensure data quality, automated Machine Learning (AutoML)	
Business Cases Requirements (from deliverable d2.1 – Data Value Chain and Requirements)		
Rq_13	Components have edge computing capabilities	Support different deployment strategies
Rq_18	Real-time grid meter data should be integrated in the system	Streaming Ingestion
Rq_24	Data set should have historical	Data Persistence

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	operation data	
Rq_26	Incident report data should integrate information from different data sources	Data heterogeneity
Rq_30	Asset monitoring data that are needed for calculation of AHI and AUI need to be available	Batch Ingestion
Rq_35	Charging station data shall be stored for result evaluation	Data Persistence
Rq_36	EV data shall be stored for result evaluation	Data Persistence
Rq_37	Historic data of maintenance must be collected and evaluated for being implemented in predictive maintenance	Batch Ingestion
Rq_38	Historic data of faults must be collected and evaluated for being implemented in predictive maintenance	Batch Ingestion
Rq_42	Historic data of faults occurred	Batch Ingestion
Rq_45	Systems need availability of building information	Support for different data categories
Rq_46	Data must be pre-processed before integration	Data Pre-processing
Rq_49	Real-time data must be available	Streaming Ingestion
Rq_52	Historical RES data must be provided	Batch ingestion
Rq_54	Systems must receive and analyzes smart meters data for IoT	Streaming Ingestion
Rq_55	The system must collect data provided by the storage	Batch Ingestion

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Expected Input	
In_1	Real-time process/operational data from smart devices
In_2	Data in batching from file collectors
In_3	Data Ingestion configurations
Component Interactions	
Rel_1	Message Broker
Rel_2	Integrated Query Engine
Rel_3	Dynamic Data Provider
Rel_4	Golden Data Sources (for batching)
Software Requirements	
Sw_Rq_1	Support Batch ingestion
Sw_Rq_2	Support Streaming Ingestion
Sw_Rq_3	Support for data persistence of both structured and un-structured data
Sw_Rq_4	Support for single server and clustered deployment
Sw_Rq_5	Portability across different computing platform and paradigms
Sw_Rq_6	Support for AutoML functions

Software Quality Model (ISO/IEC 25010)	
Runtime	
Performance	The HTAP component delivers the storage functionality to the BD4NRG platform. Therefore, the performance of this component needs to be in line with the volume and speed of the other interacting components. In particular, the HTAP needs to perform adequately to guarantee the collected data organization as well as to facilitate their future usage. It supports batch and streaming ingestion mechanisms and the possibility to persist both structured and un-structured data. According to the use case, the performance of the component can be tuned by using single-server and/or clustered deployment.
Compatibility	Aligned with the Interoperability and Homogenization data models
Usability	The component is a core element of the BD4NRG platform; thus, it needs to be usable while ensuring the proper treatment of possible failures
Reliability	The component is a core element of the BD4NRG platform; thus, it needs to be reliable while ensuring the proper treatment of possible failures
Security	The component will offer authentication and authorization mechanisms, however advanced security mechanisms will be properly handled by the Security & Trust layer
Non-runtime	
Portability	The component will be provided as docker image to be easily installed and deployed within distinct hardware and/or software environments with

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	minimum effort
Reusability	The component will be generic enough to be able to be integrated in other projects
Scalability	The component will use configuration properties for cluster tuning and horizontal scalability

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

7.3 Component Logical View

DIAGRAM

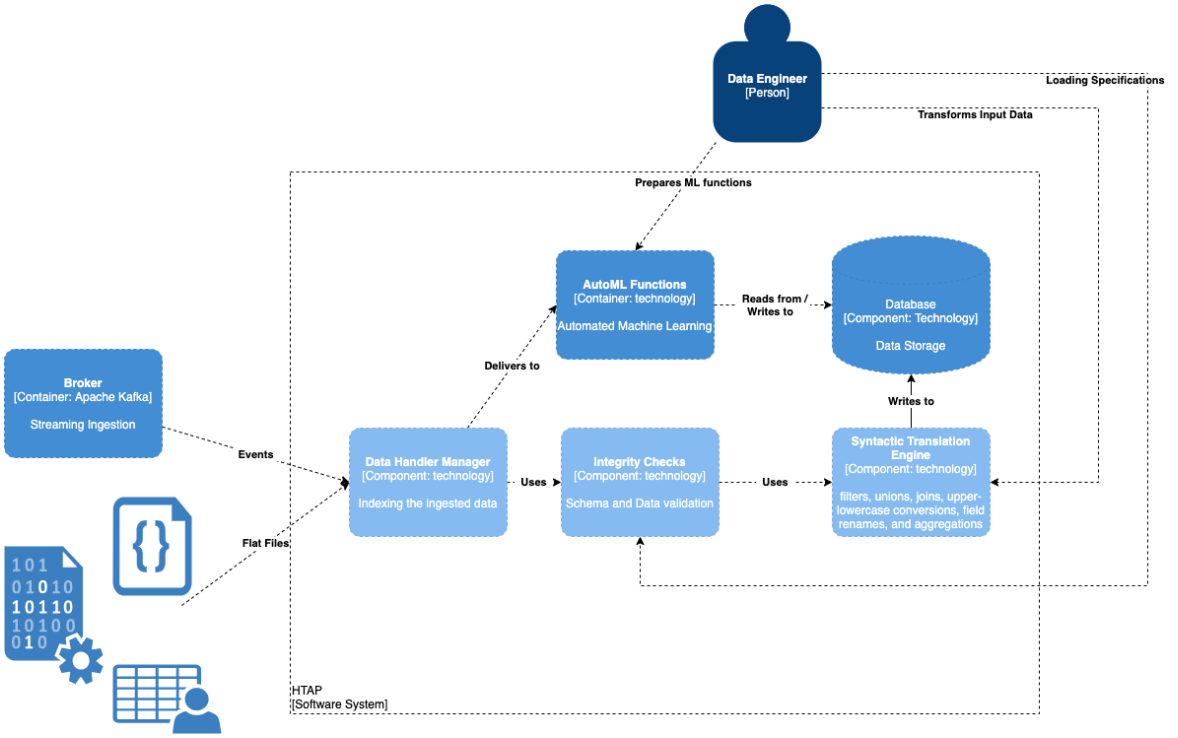


Figure 19 - Logical View of the HTAP component

COMPONENTS DESCRIPTION	
Data Handler Manager	the component responsible to index the data ingested using both batch and stream processes. As soon as the data is ingested, it is handled by the component to be indexed in order to respond to any query to those data.
AutoML Functions	The component responsible to provide methods to improve the efficiency and effectiveness of the machine learning tasks by delivering clean and effective training data. Information about missing data, correlation, cardinality, and distribution of each feature helps in the process of parametrizing and configuring the machine learning tasks.
Integrity Checks	The component is responsible for defining the ingestion specifications, i.e., defining the schema of the data to be ingested as well as to specify a set of expressions to evaluate on the top of ingested data.
Syntactic Translation Engine	The component is responsible for transforming and filtering data during the ingestion time. Some examples are unions, joins, string conversion mechanisms, field rename and/or data aggregations.
Data Storage / Persistence	The component responsible for storing the data. Depending on the application, requirements and dataset, the component can have different shapes and forms: bucket, folders, high-performing databases, virtualized instances, etc..

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

7.4 Component Deployment View

Technology	Technology description	Available connectors	HTAP module
Apache Druid https://druid.apache.org	Apache Druid is a real-time analytics database designed for fast slice-and-dice analytics ("OLAP" queries) on large data sets. Most often, Druid powers use cases where real-time ingestion, fast query performance, and high uptime are important. Druid is commonly used as the database backend for GUIs of analytical applications, or for highly-concurrent APIs that need fast aggregations. Druid works best with event-oriented data. Common application areas for Druid include: clickstream analytics including web and mobile analytics, network telemetry analytics including network performance monitoring, server metrics storage, supply chain analytics including manufacturing metrics, application performance metrics, digital marketing/advertising analytics, business intelligence/OLAP	Streaming Ingestion: • Apache Kafka • Apache Hadoop • Amazon Kinesis Deep Storage: • Microsoft Azure • Google Cloud Storage • AWS S3 • Apache Hadoop (HDFS) Metadata store: • MySQL • PostgreSQL Authentication: • Basic HTTP authentication • Apache Ranger • OpenID connect • Simple SSLContext File format • CSV • TFS • JSON • Apache Parquet • Apache Avro • Apache ORC • Protobuf	Persistence, Integrity Checks and Syntactic Translation Engine
MinIO https://min.io	MinIO offers high-performance, S3 compatible object storage. Native to Kubernetes, MinIO is the only object storage suite available on every public cloud, every Kubernetes distribution, the private cloud and the edge. MinIO is software-defined and is 100% open source under GNU AGPL v3	Connector and/or documentation on how to connect and use MinIO as a deep storage for druid is available on-line.	Persistence (for deep storage)

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

PyCaret https://pycaret.org	PyCaret is an open source, low-code machine learning library in Python that allows you to go from preparing your data to deploying your model within minutes in your choice of notebook environment.	Python client library for reading/writing data from/to druid	AutoML functions
Docker https://www.docker.com	Docker is a set of platforms as a service (PaaS) products that use OS-level virtualization to deliver software in packages called containers. Containers are isolated from one another and bundle their own software, libraries and configuration files; they can communicate with each other through well-defined channels. Because all of the containers share the services of a single operating system kernel, they use fewer resources than virtual machines.	Designed to develop cloud-native applications	Containerization platform
k8s	Generic container orchestration platform	Designed, but not limited to cloud-native.	Container orchestration

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

8 Integrated Querying of Streaming Data and Data at rest

8.1 Integrated querying component

The role of the Integrated Querying Component is twofold. On the one hand, it is the component responsible for allowing its users to explore, combine and express complex queries on BD4NRG datasets and on the other hand it is responsible for executing these queries in an efficient manner, facilitating the execution of complex queries that combine different datasets from different data sources alongside acceptable response time.

More specifically, it ensures that end users unfamiliar with formal query languages or the structure of the underlying data will still be able to describe complex queries on top of multiple datasets. The produced queries will then be used in order to retrieve raw data from BD4NRG. Moreover, these queries will be used by other components of the platform, such as the Visual Analytics Engine, in order to retrieve the required data for a specific analysis or visualization and they can also be used as templates that can be parameterized by the final users of the services so as to get customized results. The Integrated Querying component in collaboration with the Identity and Access Control Management module will also ensure that the generated queries will be used to provide data without any violation of data access policies defined within the platform.

Moreover, except static data, the Integrated querying component provides querying capabilities to streaming (real-time) data. Such capabilities can enable near real-time data analysis and unlock services like real-time anomaly detection and alerts.

In order to facilitate efficient querying to different and heterogeneous data sources, Presto[1] distributed SQL Query Engine for Big Data has been employed as the backbone of the Integrated Querying component. On top of the latter, an access control layer will be developed in order to allow access to data sources only for authenticated users that are authorized to access the requested data.

Regarding the user interface of the component, Apache Superset[2] enables querying to Presto through a user-friendly interface, along with visualization and query storing capabilities. Therefore, it is used as the main user interface for the Integrated Querying component in the initial phases of BD4NRG project. However, Apache Superset poses some restrictions in terms of both identity and access control management and the combination of different data sources. Hence, several alternative approaches will be examined, in order to abate these restrictions.

Concerning the communication with the Integrated Querying backbone, both the front end and other services, that intend to use it, can communicate with the latter through APIs or connectors.

A conceptual architecture for the Integrated Querying component is illustrated in the figure below:

[1] https://prestodb.io/
[2] https://superset.apache.org/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

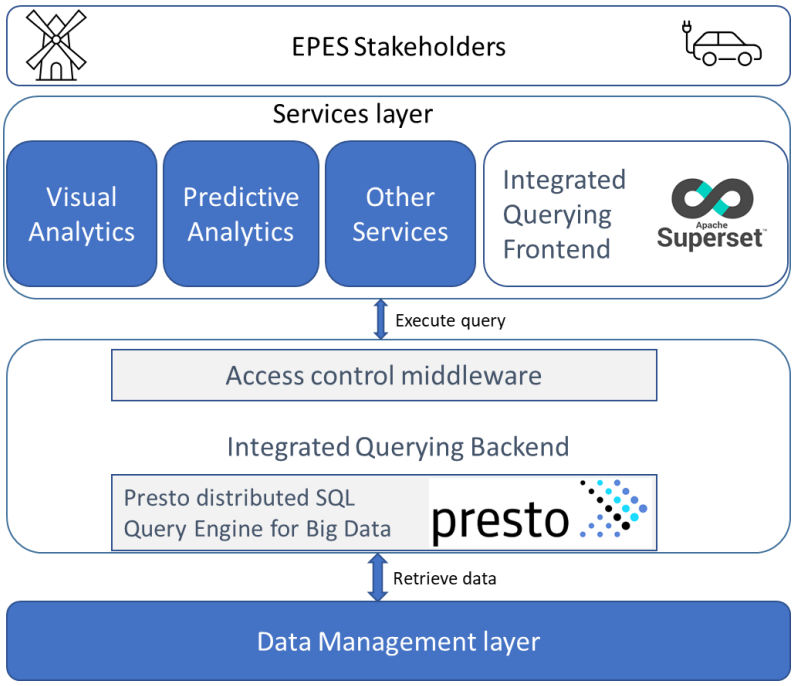


Figure 20 - Integrated Querying Conceptual Architecture

8.2 Querying data at rest and streaming data

As stated in the previous section, Presto enables complex queries that combine different datasets from different data sources. Moreover, it is compatible with most of the synchronous database technologies (relational and noSQL). Specifically, it is compatible (among others) with the technologies:

- MongoDB[3]
- PostgreSQL[4]
- Apache Hive[5] for HDFS or Object Stores (S3)
- Redis[6] Cache
- MySQL[7]
- ElasticSearch[8]
- Apache Cassandra[9]
- Apache Kafka[10]
- Apache Druid[11]

[3] https://www.mongodb.com/
[4] https://www.postgresql.org/
[5] https://hive.apache.org/
[6] https://redis.io/
[7] https://www.mysql.com/
[8] https://www.elastic.co/elasticsearch/
[9] https://cassandra.apache.org/
[10] https://kafka.apache.org/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Querying historical data (data at rest) is rather straightforward. To do that, the user has to create the query (e.g., through the Integrated Querying frontend) and the latter is executed through Presto, against the requested databases, if the user is authorized to access the requested resources.

On the other hand, it is much more complex to execute queries against real-time data (streaming data). This is because there is a continuous flow of data and the response time of a query is of crucial importance. In this context, there are several approaches for querying streaming data. The first (and the simpler) one is to query the latest values of data that entered the system. This is rather straightforward to be implemented using a key-value caching mechanism (e.g., Redis[12]). With this approach the user can retrieve the latest value that came from a sensor, through an API call, on condition that he is aware of the sensor id (key). Although this approach is very efficient in terms of response time, it enables only simple queries. If more complex queries to streams are required, Presto also enables SQL queries to Kafka streams, that go beyond querying by key and allow more complex operations like data aggregations and filtering. Of course, the response time of queries of the aforementioned approach depends heavily on the complexity of the query that is being executed and the volume of the ingested data, and as a result the response times of several queries may not be acceptable. Another approach is to perform stream queries, and instead of executing a query after an API call provide a connector to the requested data that filters the incoming data and provides results at a continuous fashion.

The first version of the Integrated querying module supports only the first two approaches through Apache Presto. Concerning the implementation of the last approach (continuous stream querying), more expensive and sophisticated tools should be utilized (e.g., Apache Spark[13]). Moreover, processing the output stream is much more complex than processing the results of an API call. Therefore, this approach will be followed only if there are strong pilot requirements for connecting to and consuming data streams.

[11] https://druid.apache.org/
[12] https://redis.io/
[13] https://spark.apache.org/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

8.3 Integrated Querying Interfaces

The following tables document the interface offered by the Integrated Querying Component:

Table 2 - List available datasets interface

List the available datasets	
Description	Retrieve a list of all the available datasets that are available on the BD4NRG platform. For each dataset, some metadata are provided, such as its title, a small description and more
HTTP Method	GET
Endpoint URL	/datasets/
Parameters	-
Request Body	-

Table 3 - List all variables of a dataset interface

List the variables of a dataset	
Description	Retrieve a list with all the variables that are included in a selected dataset. For each variable, some metadata are provided, such as its name, its dataset and its description.
HTTP Method	GET
Endpoint URL	/datasets/{dataset_id}/
Parameters	**dataset_id**: The identifier of the dataset from which the variables are obtained.
Request Body	-

Table 4 - Add a new dataset interface

Add a new dataset	
Description	Add a new dataset to the list of available datasets.
HTTP Method	POST
Endpoint URL	/datasets/
Parameters	**title:** the title of the dataset, **description:** a small description of the dataset,

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	catalog: the database catalog of the dataset, **schema**: the database schema of the dataset
Request Body	JSON

Table 5 - Execute a query interface

Execute a query	
Description	Execute a given query
HTTP Method	GET
Endpoint URL	/execute_query/
Parameters	**query:** The query to be executed
Request Body	JSON

Regarding the access of the Integrated Querying component through Apache Superset, the following Figure illustrates the provided user interface for creating a query for a dataset provided by Pilot 8 (Tilos_input):

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

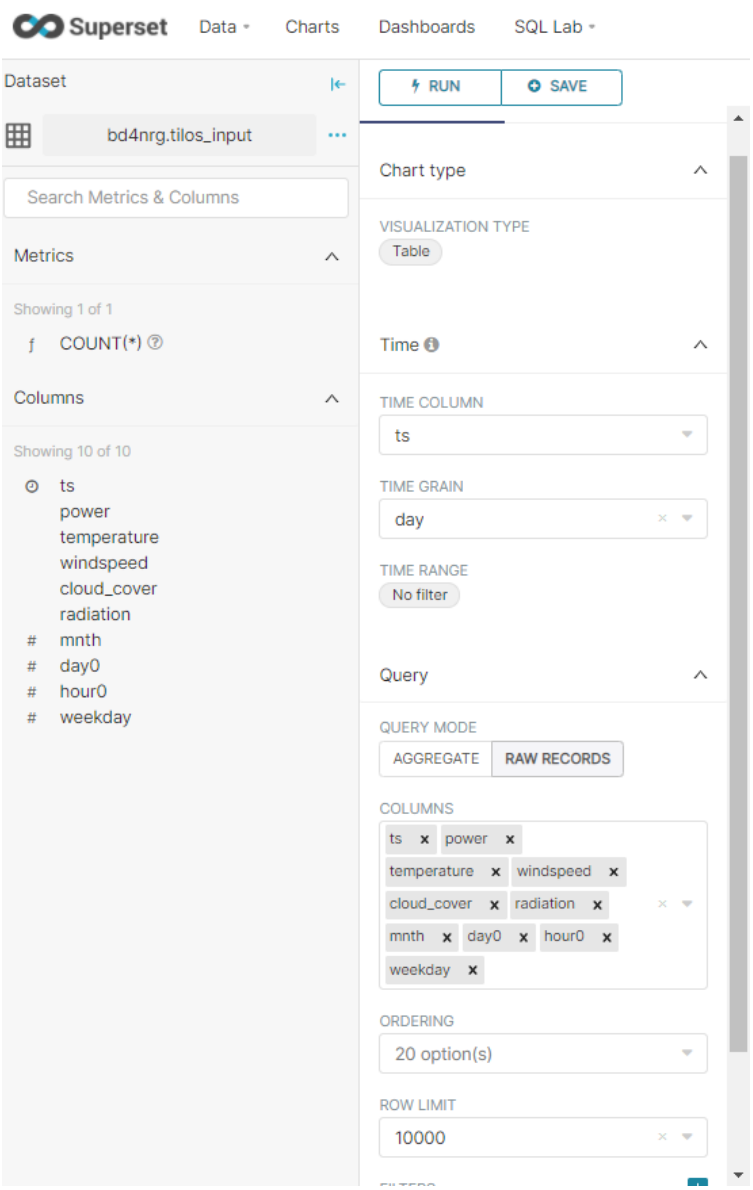
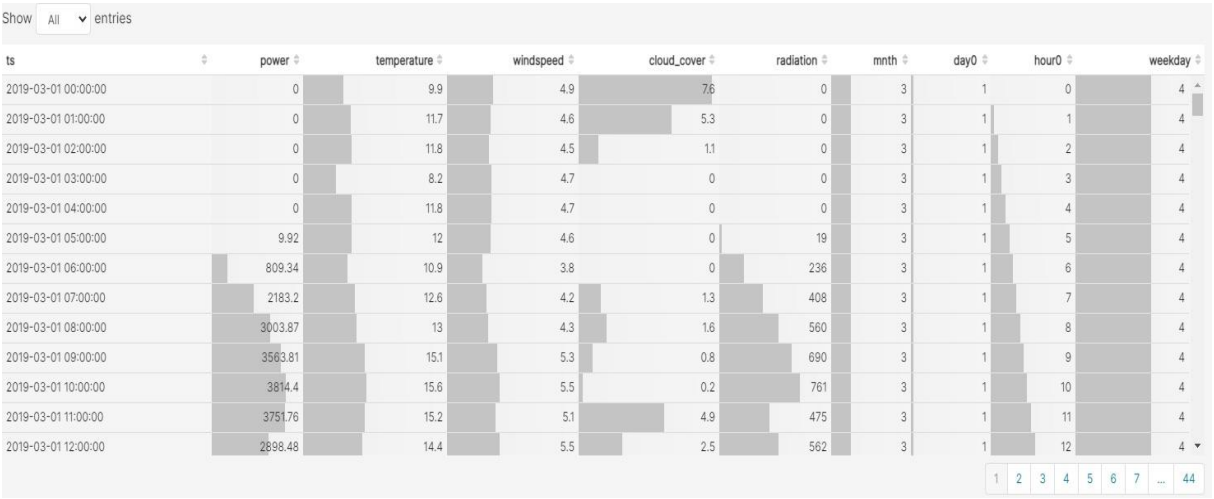


Figure 21 - Build Query user interface

In the following figure the results of the query are illustrated:

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.



Show All entries

ts	power	temperature	windspeed	cloud_cover	radiation	mnth	day0	hour0	weekday
2019-03-01 00:00:00	0	9.9	4.9	7.6	0	3	1	0	4
2019-03-01 01:00:00	0	11.7	4.6	5.3	0	3	1	1	4
2019-03-01 02:00:00	0	11.8	4.5	1.1	0	3	1	2	4
2019-03-01 03:00:00	0	8.2	4.7	0	0	3	1	3	4
2019-03-01 04:00:00	0	11.8	4.7	0	0	3	1	4	4
2019-03-01 05:00:00	9.92	12	4.6	0	19	3	1	5	4
2019-03-01 06:00:00	809.34	10.9	3.8	0	236	3	1	6	4
2019-03-01 07:00:00	2183.2	12.6	4.2	1.3	408	3	1	7	4
2019-03-01 08:00:00	3003.87	13	4.3	1.6	560	3	1	8	4
2019-03-01 09:00:00	3563.81	15.1	5.3	0.8	690	3	1	9	4
2019-03-01 10:00:00	3814.4	15.6	5.5	0.2	761	3	1	10	4
2019-03-01 11:00:00	3751.76	15.2	5.1	4.9	475	3	1	11	4
2019-03-01 12:00:00	2898.48	14.4	5.5	2.5	562	3	1	12	4

1 2 3 4 5 6 7 ... 44

Figure 22 - Query results user interface

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

9 Automated Parallelization of Data Streams and Intelligent Data Pipelining

9.1 Adaptive data pipelines

This section introduces the concepts of self-adaptation, self-healing, self-awareness, self-protection, self-configuration and self-optimization, which are at the basis of adaptive data pipelines. The next sub-section presents an overview of the changes in software divided into seven categories and describes each of them with examples. Adaptation triggers are listed and discussed both for long-term and short-term adaptation. The section is concluded with a discussion on adaptive data pipeline lifecycle in four phases: design and development, operation, maintenance and end-of-life.

9.1.1 Self-adaptation

Self-adaptation is a property of the dynamic system which is capable to adapt to the changes. First, awareness on the system's capabilities and on the current situation and context, based on system and environmental information, is a key. That is why in Figure 23 Self-Awareness is put around the concept of adaptation as a necessary surrounding block. The learning (or self-adaptation) of the studied systems relies on high quality and rich information. Typically, this is based on data in combination with the use of physical models or digital twins based on Big Data. This especially holds for inner system state awareness. This data is collected from previous operation, from special operating conditions (which are introduced intentionally), or from other similar systems. This is enabled by the observed trend of increased connectivity, which gives access to more environment data. Besides the observation of the current state, also the forecast of future situations is essential.

Using the awareness information, the overall performance of the complex system is maximized while dealing with constraints. Contrary to conventional optimization methods, there is a need for optimization approaches that can deal with a wide range of system deviations:

- ***System uncertainty***; such as production tolerances, component wear and ageing. This leads to changing system behaviour over time;
- ***Changing environmental conditions***: e.g., ambient conditions, data availability, service availability or traffic flow;
- ***Changing requirements***: e.g., legal limits, energy or production demands.

In this project four classes of self-adaptation are distinguished:

1. ***Self-optimization***: optimizing control and awareness settings based on changing system and environmental state (for example an attack);
2. ***Self-configuration***: system reconfiguration based on information on system capabilities (Anomaly) and on system state;
3. ***Self-protection***: in order to avoid damage or failure, including damage from external hacker attacks, system behaviour is adapted (incl. switching of (sub-) systems) for changing environmental state. This typically result in suboptimal performance to extend the lifetime;

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

4. ***Self-healing***: continue to realize maximum performance by system reconfiguration to overcome failures and anomalies.

Each class has its own goals and will require more specific optimisation criteria and key performance indicators. Based on the chosen self-adaptation strategy, the adaptation function can determine:

- how the performance can be optimised;
- how it can guarantee robustness;
- how it can guarantee explainability of the selected improvement actions.

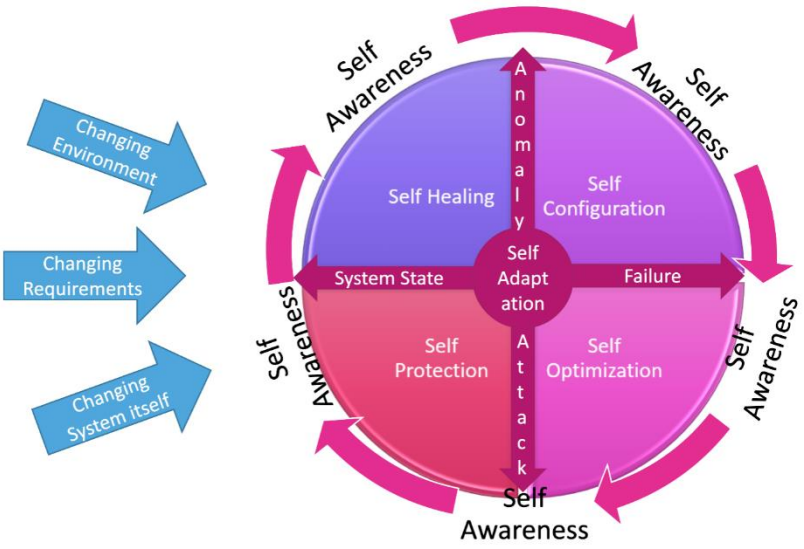


Figure 23 - The self-adaptation concept

9.1.2 Overview of changes in software

The types of changes in software can be divided into seven categories: failures, changes in the systems itself, environment, users, regulations and legislations, IT infrastructure and security incidents, as depicted in Figure 24.

Figure 24 – Types of changes in Software

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The changes in the IT infrastructure are connected both to hardware and platform software. Change of a data and application host is one of the modifications that can influence an application. Moving services and data between 3rd party providers requires often adaptation from the application. Change of the infrastructure type is also an important change for the application, for example moving services or data from cloud to edge. Changes in data processing or any underlying data storage solution often require integration effort and can be made seamless by applying adaptation mechanisms.

External changes in the environment can influence applications and their operation. We can divide external changes into foreseen and unforeseen. Foreseen adjustments include short horizon changes, for example people behaviour patterns and long horizon changes for example season change or planned constructions and extensions. Probable changes to the environment include power outage, storms, flood, and improbable changes are for example natural disasters, terrorism attack, large synchronized failures (due to faulty update) and external temperatures. Unforeseen changes include tornado, large power failure, IT infrastructure outage and earthquake.

Changes in the system itself include adjustments in data and in components (applications, services, software). The changes in data include adding and removing data sources, replacing data sources, modifications in APIs used for data access. The data structure (connected to modifications of underlying database structure), data format, underlying data model have influence on how data is received by applications. Modifications in frequency and the way the data is delivered in a timely manner to the application, that incudes swapping between non-real time to real-time data streams and switching between batch processing and streaming. Changes in data access type for example (change from an SQL query to a REST call) can influence the operation of an application, it usually requires manual integration efforts. Adjustments in data features affect results that data driven application provide, possibly as a service to other applications. The changes in data features include: statistical properties of data, change of data trends, features appearing and disappearing from data, new or unknown biases. Additionally, to adjustments in data features, the sizes of data sets can drastically change over time. This can lead to imbalances between training, testing and validation datasets that influence the generality of trained models. Changes to any data policy or governance influences the way data is accessed and how pipelines and applications can process the data. The changes in data governance can affect the whole data sets or parts of a set, preventing applications from accessing data for some point in time, and restricting how results of historical processing can be accessed by other applications. Data governance can also specify when data should be deleted and data retention policies can delete data and all its traces without informing any of the dependent applications. Changes to components include new software versions, new interfaces, APIs, services and any parameters for example configuration, hyper-parameters or model weights.

Software can be designed in a way that its goals, KPIs, constrains or business requirements are a moving target. In the modern software development, change of business requirements require manual steps that include extensions, additional components or building blocks, re-implantation or, in the worst case, redesign. In many cases asking an additional business question can result in months of development work.

9.1.3 Adaptation triggers

A short-term adaptation is usually a reactive process, designed to be triggered on a specific event or

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

deviation from a KPI. A short-term adaptation requires self-awareness components that observe the system itself, its environment, all surrounding systems and KPIs to assess how well in the system doing under the current circumstances. Events that can trigger a short-term adaptation include: new versions of components, deviation from the required performance and availability, expected external conditions and expected functionality. Other short-term adaptation triggers include: new values of parameters or configuration changes, failures of components or subsystems, addition or switching between data sources, security breaches or cyber-attacks. Any new information that leads to possible improvements can be an adaptation trigger, all system improvements need to be evaluated in order to justify the cost for adaptation. Infrastructure changes including unavailability or change of capacity of resources for example computational, communication, bandwidth, data loss, any changes in the 3rd party services can be a trigger for adaptation on any level of the system and can for example trigger resource allocation, migration processes or communication adaptation.

Usually a new goal, requirements or constraints is a reason for software rewrite or decommission, as the whole purpose of the software changes. It is foreseeable that these fundamental changes in the software purpose can also be tackled with systems adaptation. In order for this adaptation to be successful the software must be aware of its requirements, goals and constraints and formulate adaptation strategy that can influence the fundamental behaviour of the system.

Adaptation is a disruptive process and should always be evaluated on its present and future effects. A flexible system should be able to evaluate its adaptation process and be able to stop its progress. Planning, evaluation and execution of an adaptation strategy can take up systems resources and decrease the stability of the system. Other tasks can take priority over adaptation for example a spike in system use, infrequent large calculations or security upgrades. Adaptation planning, evaluation and execution often takes time, it is possible that in the meantime the adaptation strategy is no longer needed or is not optimal, this means one adaptation process can be stopped and another can start to keep up with the system's transition.

A long-term adaptation is not reactive, it requires planning and predicting future user needs and state of the whole ecosystem that the application exists in. Long term adaptation can include learning skills that the be use in the future, moving away from phased-out technology, retraining models to become more generic toward future situations and transfer learning.

9.1.4 Adaptive data pipeline lifecycle

The lifecycle of adaptive data pipelines can be divided into four parts: design and development, operation, maintenance and end-of-life as depicted in Figure 25.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

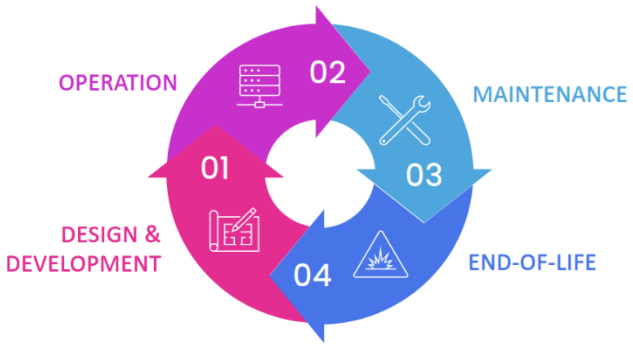


Figure 25 - The lifecycle of adaptive data pipelines

In the design and development stage of the lifecycle, system designers can use change-driven modularity technique: identify all axis of change of the system and design software modules to be ready for foreseeable changes. Design-for-change includes as requirement that every part of the software can change, designing modular software with a clear separation of concerns and clear interfaces, following single-responsibility design principle, for example with use of micro-services, minimizes the impact of change of each module on the whole system.

In operation a running software or its parts can be shamelessly updated with use of continuous integration and continuous delivery (CI/CD) methods and tools. A single change to the software, usually done manually by a developer, can be automatically built, tested, released, deployed, operated and monitored. To close the CI/CD loop based on measurements of the software performance, the new software changes need to be planned and implemented, this is nowadays mostly done by developers however, in the short-term and long-term adaptation in operation, techniques can aid the full automation of the CI/CD software development loop. With a clear separation of application from the infrastructure in Infrastructure-as-a-service (IaaS) it is possible to continuously optimize resource usage: this concerns CPU/GPU usage, storage, power optimization. Data locality techniques allow bring computation to the data to optimize I/O speed, comply with security rules (for example in secure multi-party computation) or avoiding costly data migration (federated learning). Observability and monitoring are important to assess application's health and fit-for-purpose, used also for context awareness and self-awareness. Data monitoring including lineage, provenance, data statistics, health of data source, and traceability of any data modification or processing, is especially needed for assessing data-driven and data-centric systems.

In modern software deployment seamless integration of new or improved software in operation is done with rolling updates for example in a form of blue-green deployment, A/B testing or canary releases. Similar techniques can be used to release adaptation strategy in predictable manner and reducing any downtime associated with a release. If a software is built in a modular and exchangeable way, it is possible to activate and deactivate parts of the system that are currently required, functions can scale up and down depending on rate of requests, storage and CPU/GPUs and added and removed on demand, new findable functionalities can appear in the pool of services available to the application, new self-describing data can be found and seamlessly integrated in existing applications.

In the maintenance phase of the adaptive data pipeline lifecycle the application is constantly observed and its performance analysed. Maintenance is usually a manual process performed by engineers aided by observability and monitoring tools providing traceability, data provenance, and

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

can help with root-cause analysis and data quality assessment. Other maintenance tasks include anomaly detection, reverse engineering, data maintenance and migration strategies, self-healing and predictive maintenance.

The end-of-life phase of the adaptive data pipeline lifecycle is reached when the application cannot adapt to the current circumstances and needs to be replaced with a new application. This phase focuses on graceful degradation, transfer of knowledge and repurposing of legacy ecosystems.

9.2 Dynamic data provider

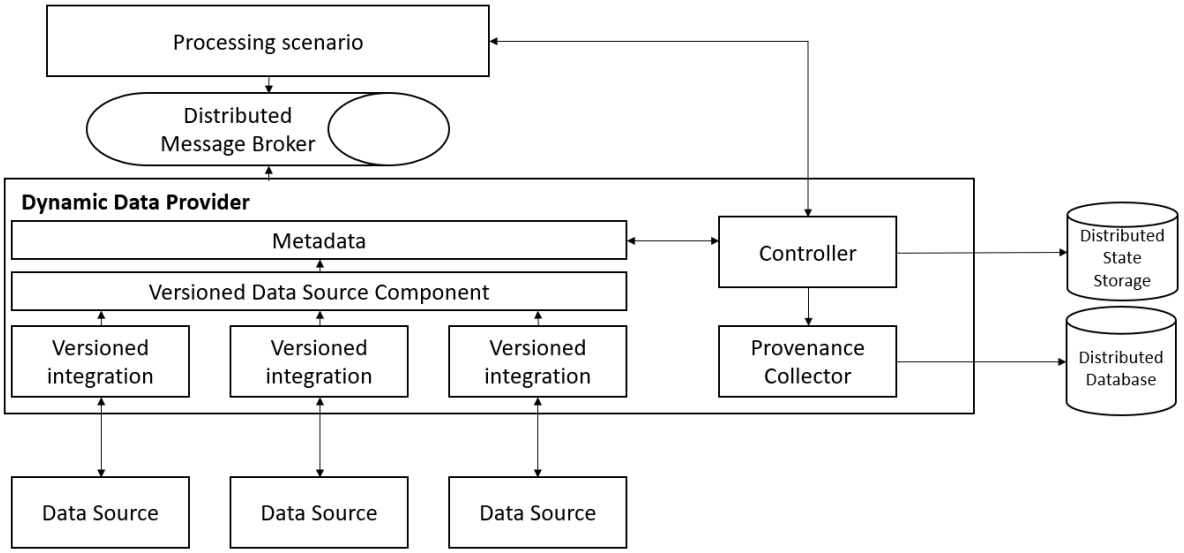


Figure 26 - Dynamic Data Provider Concept

Allowing a processing scenario to dynamically switch from one data source version to another, especially when considering the previously stated requirements, introduces several problems that have to be solved. For the usability requirement, it is extremely important that the processing scenario can be built by the user without the user having to introduce additional functionality in the processing scenario that allows the dynamic switching.

All that is required for the dynamic switching must be hidden from the user and it must be easy to use the functionality in the processing scenario. This requires some additional layer on which the processing scenario can be built that offers this dynamicity to the processing scenario through a simple interface. Constructing such a Dynamic Data Provider is, however, not a trivial task. The processing scenario is no longer using a specific data source and the Dynamic Data provider can choose from any of the available data sources versions.

Before it is possible to process any data, the data has to be retrieved from a data source. This requires the Dynamic Data provider to interact with the data source and specify what data is required. How the interaction with a data source is done, can greatly differ between different data

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

sources. Usually, this interaction is taken care of by a component called a driver, and such drivers are in many cases available and can be used by the Dynamic Data provider to interact with data sources.

However, there is an enormous number of different types of data sources available and including support for all of these data sources in the Dynamic Data Provider is very difficult to realize. Furthermore, a driver might not be available for every one of these data sources. In both of these cases, it could be beneficial to provide the user with means to indicate their own driver or to provide the Dynamic data provider with code that it can use to interact with the data source. With such a mechanism in place, using data sources that are not supported by the dynamic Data provider becomes much easier and much more flexible.

To hide the complexity of the heterogeneity of the data sources from user, the introduced Metadata layer can leverage the choice of the needed data source type and version. In such Metadata all data sources are described, taking into account the changes in time in versioning and their properties.

Controller is responsible for organizing the process of the propagation of the changes in data and monitoring its execution. Version Data source component is responsible for providing the required version from different data sources which are needed for analysis as an input on request from Controller.

The overall state of the system is kept in Distributed State database. With any change happening in the system the database receives an update. Data and Metadata are kept in distributed database.

9.2.1 Functionality of Adaptation Engine for Data Changes

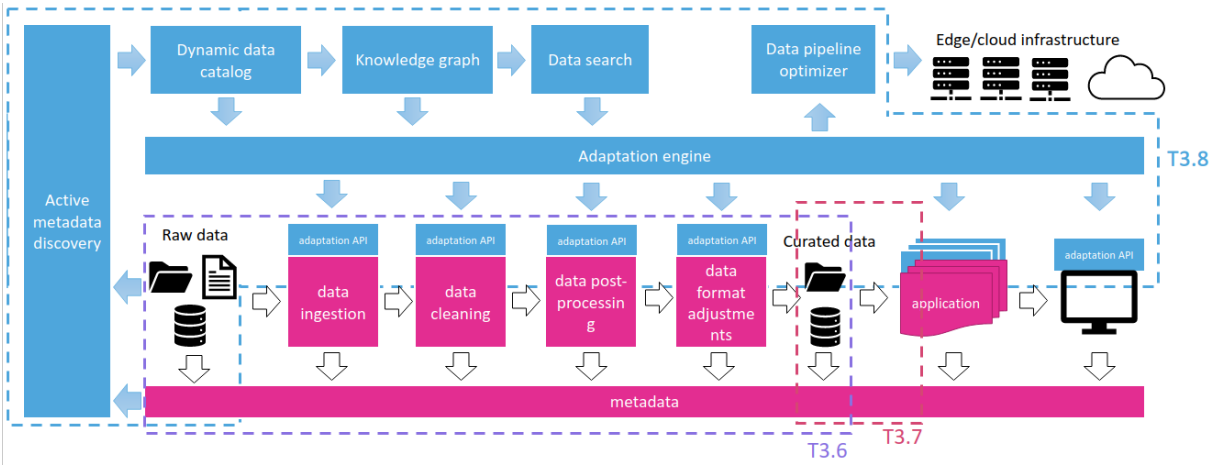


Figure 27 - Functional Design of Dynamic Data provider

Figure 27 depicts the functional design of the dynamic data provider and the information flow and interactions with the full data pipeline. The envisioned adaptation engine is focused on tackling changes in data, it is possible to extend the adaptation engine to deal with other changes as described in section 9.1.2.

To be able to cope with the changes happening with data in time, the data provider cannot be static. It should actively monitor the changes which happen in time, and trigger the Adaptation functionality to start.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

One of the most important components is an Active metadata discovery. This component is responsible for the periodic browsing through the newly coming data, its comparison with the historical data, and by that finding new properties, features, tags and anomalies in data behavior. The periodicity of browsing through data depends on the dynamicity of data itself and frequency of the changes in time.

The findings from the Active metadata discovery are put into Dynamic Data catalog, which updates its records according to the new input. This catalog is used to represent the schema, relationships within data, and includes the description information about the datasets of the organization. The relationships between different entities within catalog are translated into the Knowledge graph which allows the graphic representation of all datasets and the relationships between the entities. Data search functionality allows to browse and reason through the representation of data in knowledge graph, and apply graph algorithms for the effective search for needed subsets of data.

Adaptation Engine is running on permanent basis in order to control the integrity and correctness of all updates/ upgrades and fixes in Dynamic Data Provider layer.

The process of making Curated data in required format, of necessary quality and guaranteeing data integrity includes the following stages:

- Data ingestion: the mechanism to load the raw data coming from the data sources;
- Data cleaning: removing the noise from data, bias;
- Data postprocessing: filling in the gaps within data, filtering it, compressing, calculating means, averages and deviations, etc.;
- Data format adjustments: putting data into the required format.

After going through all these stages data becomes Curated, ready for use by the applications which need this data. It is stored in the database, and can be at any time pulled by the application.

Adaptation actions could happen at any of these stages. For Data ingestion it could be, for example, mean the change of frequency of accepting the data, or changing the model of loading the data. For Data cleaning and Data postprocessing the new models of better quality can be introduced. For format adjustment the new requirement for new format can come. Application itself can change, and therefore, could request data in different form and format. Adaptation engine should control the actions and introduction of new models into all the phases of the transformation of data from Raw to Curated and its use in application.

9.2.2 Change propagation with sequence diagram

The following diagram provides the reflection of the process of the propagation of the change in data, going through all the components of Self-adaptation functionality.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

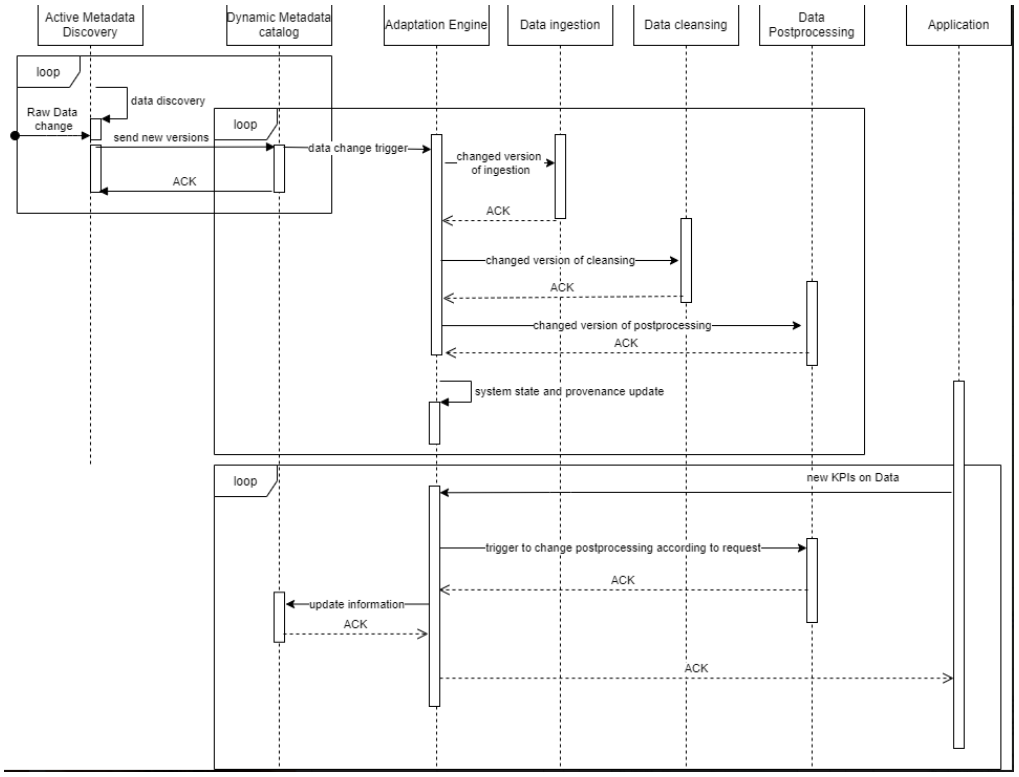


Figure 28 - Propagation of change in data

9.2.3 Controller

This module helps the Dynamic Data Provider to get data into its Distributed Database and Storage and deliver processed data to the Distributed Message broker. Controller is hence responsible to apply the proper tools for organizing the process of the propagation of the changes in data and monitoring its execution. Therefore, it is possible to propagate schema changes on the source to accommodate for schema drift in the distributed state storage.

9.2.4 Versioned data source type component (dynamic data catalog)

The versioned data source type component takes a role of a dynamic data catalog. A Dynamic Data Catalog is a collection of metadata that stays current with the data, combined with data management and search tools that helps data users to find the data that they need, serves as an inventory of available data, and provides information to evaluate fitness data for intended uses. Traditional data catalogs are intended to be aiding users in search for data in many separate data stores. The versioned data source component is used by the Controller to recognize changes in data providing adaptation triggers or constraining the search for new data sources and changing metadata. The versioned data source component provides the traces of the evolution of the data source, enabling the controller to evaluate where the change happened, which applications are using the data and what needs to be adapted in the pipeline to fit the current situation.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

9.2.5 Versioned Integration

Versioned Integration for Data sources is needed since the Data sources can also undergo through changes. For example, the sensors could be replaced, the databases which were SQL could become no-SQL, the frequency of coming data can change as well. For the traceability and reproducibility, it is very important to know when the change happens, and which version of the Data source was valid at the specific time period when the requested data came. It allows to avoid the situations when expected result from processing does not reflect the data sources versions which were valid at the requested period. To learn the changes in versioning in time, this component continuously analyzes available metadata for key metrics and statistics.

9.2.6 Provenance collector

Data received in the module must be compliant with the rules established in T3.4. This means data passed through a complete process of standardization and certification to comply with the internal rules posed by BD4NRG infrastructure.

The provenance collector module may handle the production of metadata that is paired with records that details the origin, changes to, and details supporting the confidence or validity of data, thus providing a sense of data lineage. This data provenance feature is important for tracking down errors within data and attributing them to sources. Additionally, the actions performed by this provenance collector can be useful in reporting and auditing for business and research processes.

9.2.7 Message Broker description

A message broker (also known as an integration broker or interface engine) is an intermediary computer program module that translates a message containing data/metadata from the Dynamic data provider layer to the formal messaging protocol of Data Processing scenario in a format which is required by that scenario. To be able to provide huge volumes of data to the processing scenario the message broker in Big Data system is typically made distributed. There the requested subset of data can reside persistently until it is read and used by the Processing Scenario.

9.2.8 Storage

From what can be inferred up to three separated types of information should be stored in the module, namely: 1) provenance data, 2) raw data and 3) metadata. Each data type must present the required fields to distinguish one from another and hence permit the database handle them all properly.

To achieve this successfully, there is the need to perform kind of a similar set of actions as the ones that would be carried out in a so-called data lake. This means there would not be a single repository but it would be flexible and could be divided into several separate layers instead.

On the one hand, a raw data layer could ingest data into raw as quickly and as efficiently as possible. To do so, data should remain in its native format, don’t allowing any transformations at this stage. In this particular case, it would be possible to get back to a point in time, since the archive is maintained. In addition, no overriding is allowed, which means handling duplicates and different

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

versions of the same data. Despite allowing the above, raw data would still need to be organized into folders.

As for the provenance data, once passed the respective data quality controls data is stored in a format that fits best for further cleansing and similar operations. The internal structure could be similar as the one employed with raw data, even introducing a higher granularity.

In the particular case of the metadata, is worth reminding in essence it consists in data about data, so implies mostly all the schemas, reload intervals, additional descriptions of the purpose of data, with descriptions on how it is meant to be used.

9.2.9 Data Pipeline Optimizer

Part of the Activity of the task T3.8 is the investigation of how data locality and the ability to move the computation closer to where the data resides, can improve overall throughput of the system and minimizes network congestion. Data Locality at the same time allows analysis of commercial and confidential data without sharing.

The functionality to be implemented is the ability to distribute Data Pipelines on the computation nodes, trying to exploit as much as possible data locality. For this purpose, the module needs to have access to the following information:

- available computation nodes on the infrastructure, possibly linked with data locality information (e.g., if a computational cluster is co-located in the same data center as the data we are trying to access)
- available datasets on the infrastructure (data catalogue), provided by one of the components to be implemented in the context of task T3.8 (Dynamic Data Catalogue)

Assuming that the computational nodes are running a container orchestrator (e.g., Kubernetes) the module will make sure to deploy and configure the needed data streaming technology according to the pipeline to be executed.

The component will provide API for the Adaptation Engine component to interact with, in order to describe the computation and the data to be used. Possible usage of TOSCA language to specify the topology the services, their components, relationships, and the data.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

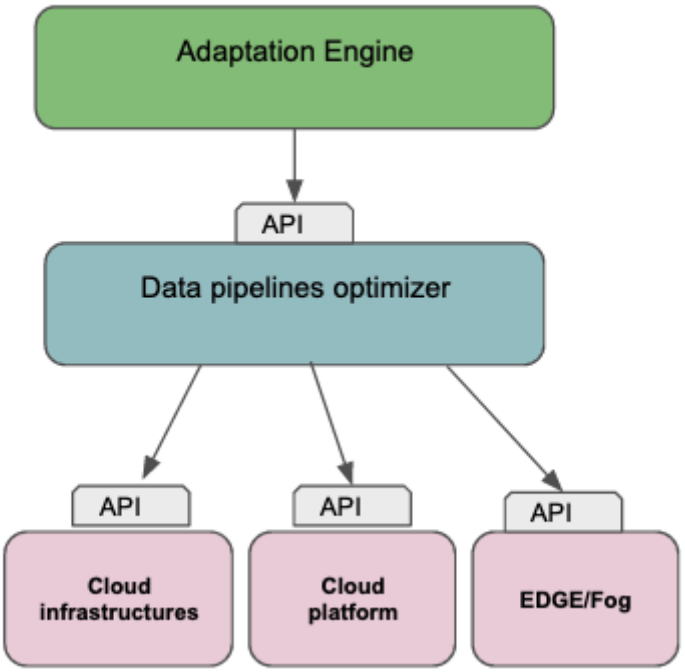


Figure 29 - Data Pipeline Optimizer

Infrastructure resource discovery needs to be implemented in order to understand the available computation nodes and if the given computation nodes is co-located with the data source, to explore data locality.

9.2.10 Technology stack

Technology	Technology description	Available connectors	Dynamic data provider module
Amundsen www.amundsen.io	Amundsen is a data discovery and metadata engine for improving the productivity of data analysts, data scientists and engineers when interacting with data.	Table Connectors: • Amazon Athena • Amazon Glue • Amazon Redshift • Apache Cassandra • Apache Druid • Apache Hive • CSV	Versioned integration

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Technology	Technology description	Available connectors	Dynamic data provider module
		• dbt • Delta Lake • Elasticsearch • Google BigQuery • IBM DB2 • Microsoft SQL Server • MySQL • Oracle (via dbapi or sql_alchemy) • PostgreSQL • PrestoDB • Trino (formerly Presto SQL) • Vertica • Snowflake Table Column Statistics: • Pandas Profiling Dashboard Connectors: • Apache Superset • Mode Analytics • Redash • Tableau ETL Orchestration: • Apache Airflow	
Marquez https://marquezproject.github.io/marquez/	Marquez is an open-source metadata service for the collection, aggregation, and visualization of a data ecosystem's metadata. It maintains the provenance of how datasets are consumed	RESTful API enabling integrations with: • Apache Airflow • Amundsen • Dagster	Versioned data source type component

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Technology	Technology description	Available connectors	Dynamic data provider module
	and produced, provides global visibility into job runtime and frequency of dataset access, centralization of dataset lifecycle management.		
Great Expectations https://greatexpectations.io/	Great Expectations helps data teams eliminate pipeline debt, through data testing, documentation, and profiling.	Connectors to: • Pandas • Apache Spark • SQL • Apache Airflow • S3 • Azure Blob Store • GCS • Athena • Big Query • MySQL • PostgreSQL • SQLite • Snowflake • Redshift	Data validation and profiling
Kubernetes https://kubernetes.io	Generic container orchestration platform	Designed, but not limited to cloud-native.	Container orchestration
Apache Storm https://storm.apache.org/	Apache Storm is a free and open source distributed real-time computation system. Apache Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing.	Integrates with streaming technologies: • Kestrel • RabbitMQ / AMQP • Kafka • JMS • Amazon Kinesis	Data Pipelines Optimizer
Apache Flink https://flink.apache.org	Apache Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Flink has been	DataStream Connectors: • Kafka • Cassandra • Elasticsearch	Data Pipelines Optimizer

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Technology	Technology description	Available connectors	Dynamic data provider module
	designed to run in all common cluster environments, perform computations at in-memory speed and at any scale.	• Kinesis • File Sink • RabbitMQ • Google Cloud PubSub • Hybrid Source • NiFi • Pulsar • Twitter • JDBC Table API Connectors: • Kafka • Upsert Kafka • Kinesis • JDBC • Elasticsearch • FileSystem • HBase • DataGen • Print • BlackHole • Hive	

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

10 Conclusions

The main purpose of this deliverable is to present in a clear and concise manner the 1st Technology release of BD4NRG Governance Layer based on a unified system architecture approach for all the relevant components per WP3 Task. This layer forms the middleware on top of the common IP infrastructure, in order to enable the exchange of information between assets, systems, data hubs and actors in a common manner. This involves (a) the definition of a data access policy and a legal/ethical/cybersecurity framework implementation, (b) the development of P2P DLT blockchain/smart contracts platforms and DAPPS for cross-stakeholders secure transparent data/models sharing and marketplace-based data and models monetization, (c) the development and configuration of modules that will assure the quality of retrieved datasets, (d) the development of edge/IoT big data management enablers including elastic streaming data capturing management, and (e) the building of the interoperability and homogenization layer to manage the data access.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

References

[1] https://www.fiware.org

[2] FIWARE Keyrock: https://fiware-idm.readthedocs.io

[3] FIWARE Wilma: https://fiware-pep-proxy.readthedocs.io

[4] FIWARE Foundation, "*Fiware for Data Spaces*", Position Paper, Version 1.0, June 2021 Available: https://www.fiware.org

[5] Bailey, et al., "The energy-sector threat: How to address cybersecurity vulnerabilities," 2020. [Online]. Available: https://www.mckinsey.com/business-functions/risk/our-insights/the-energy-sector-threat-how-to-address-cybersecurity-vulnerabilities.

[6] European Parliament, "Regulation (EU) 2016/679 General Data Protection Regulation (GDPR)," 2016. [Online]. Available: https://eurlex.europa.eu/eli/reg/2016/679/oj.

[7] CloverDX Blog, "*6 data quality metrics you can't afford to ignore*", 2021 [Online] Available: https://www.cloverdx.com/blog/6-data-quality-metrics-you-cant-ignore

[8] NEMA (National Electrical Manufacturers Association), "*Certification Procedures for Data and Communications Security of Distributed Energy Resources*", Danish Saleem and Cedric Carter, 2019 [Online] Available: https://www.nrel.gov/docs/fy19osti/73628.pdf

[9] North American Electric Reliability Corporation, CIP (Cybersecurity requirements for critical infrastructure protection) Standards [Online] Available: https://www.nerc.com/pa/Stand/Pages/CIPStandards.aspx

[10] National Institute of Standards and Technology (NIST), Cybersecurity Framework [Online] Available: https://www.nist.gov/cyberframework

[11] International Data Spaces Association (IDSA), Certification [Online] Available: https://internationaldataspaces.org/use/certification/

[12] International Data Spaces Association (IDSA), "*IDS Certification Explained*", Position Paper, Version 1.0, November 2019 Available: https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-IDS-Certification-Explained.pdf

[13] LUMIQ.AI, "Data platform modernization," https://medium.com/, Jun. 23, 2021. https://medium.com/lumiq-tech/data-platform-modernization-85128d475d2b

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.