interTwin

Data management for the interTwin project

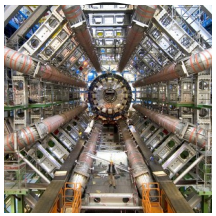
Paul Millar[1], Tim Wetzel[1], Dijana Vrbanec[1], Daniele Spiga[2], Andrea Manzi[3]

International Symposium on Grids and Clouds 2023, Academia Sinica

[1] DESY, [2] INFN, [3] EGI Foundation

interTwin is funded by Horizon Europe under grant agreement n° 101058386

Why interTwin



interTwin overall objective

Co-design and implement the prototype of an interdisciplinary Digital Twin Engine.

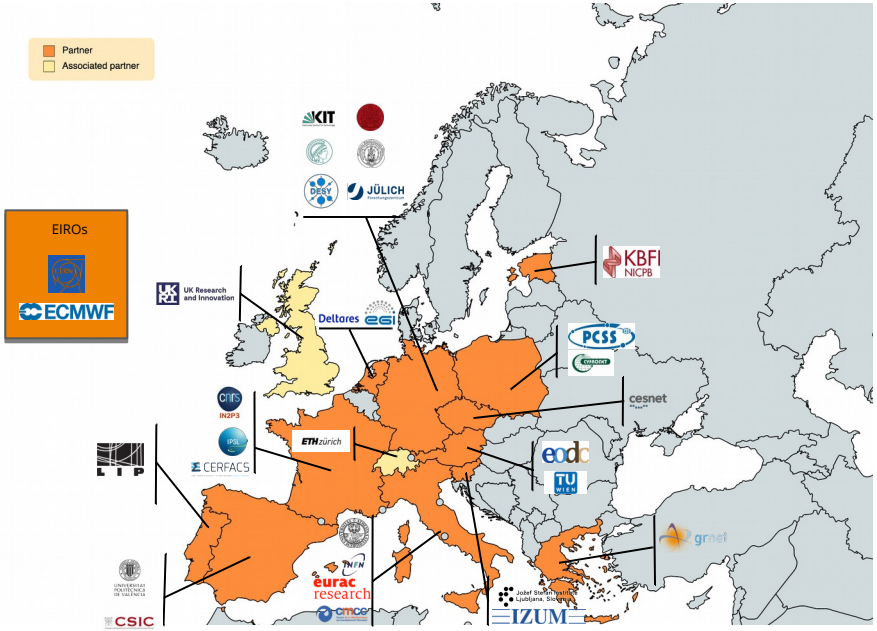
Digital Twin Engine

- It is an **open-source platform** based on open standards.
- It offers the capability to integrate with **application-specific Digital Twins**.
- Its functional specifications and implementation are based on
  - a **co-designed interoperability framework**
  - conceptual model of a DT for research - **the DTE blueprint architecture**.

What is a Digital Twin?

> *A digital twin is a virtual representation of an object or system helping in decision-making and maintenance. It takes in real-time data and keeps track of the history of the object or system.*

Consortium Overview



EGI Foundation as coordinator

29 **Participants**, including 1 affiliated entity and 2 associated partners

Consortium at a glance

10 Providers	11 Technology providers	14 Community representants
cloud, HTC , HPC resources and access to Quantum systems	delivering the DTE infrastructure and horizontal capabilities	from 5 scientific areas; requirements and developing DT applications and thematic modules

DT Applications: High Energy Physics

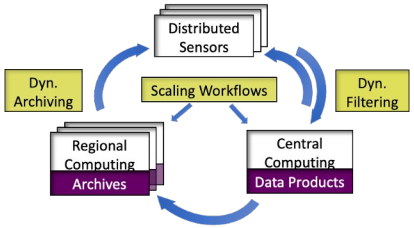
DT of Large Hadron Collider (LHC) detector components

seeking for strategies to face the increased need for simulated data expected during the future LHC runs. The primary goal is to provide a fast simulation solution to complement the Monte Carlo approach. ***Faster and deeper cycles of optimisation of the experiment parameters*** in turn will enable breakthroughs in experimental design.

DT of the Standard Model in particle physics

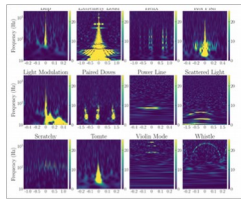
competitive results in Lattice QCD require the ***efficient handling of Petabytes of data***, therefore the implementation of advanced data management tools is mandatory. On the side of algorithmic advancement, ML algorithms have recently started to be applied in Lattice QCD. The goal is to ***systematize the inclusion of ML for large scale parallel simulations***.

DT Applications: astronomy and astrophysics



DT for noise simulation of next-generation radio telescopes

Providing DTs to simulate the noise background of radio telescopes (**MeerKat**) will support the identification of rare astrophysical signals in (near-)real time. The result will contribute to a realisation of "**dynamic filtering**" (i.e. steering the control system of telescopes/sensors in real-time).



DT of the Virgo Interferometer

meant to **realistically simulate** the noise in the detector, in order to study how it reacts to external disturbances and, in the perspective of the **Einstein Telescope**, to be able to detect noise "glitches" in **quasi-real time**, which is currently not possible. This will allow sending out **more reliable triggers** to observatories for multi-messenger astronomy.

DT Application: Climate change and impact decision support tools

DT of the Earth

addressing complementary topics such as:

- Climate change, long-term predictions of extreme natural events (storms & fires)
- Early warning for extreme events (floods & droughts)
- Climate change impacts of extreme events (storms, fires, floods & droughts)

Interoperability & link with DestinE

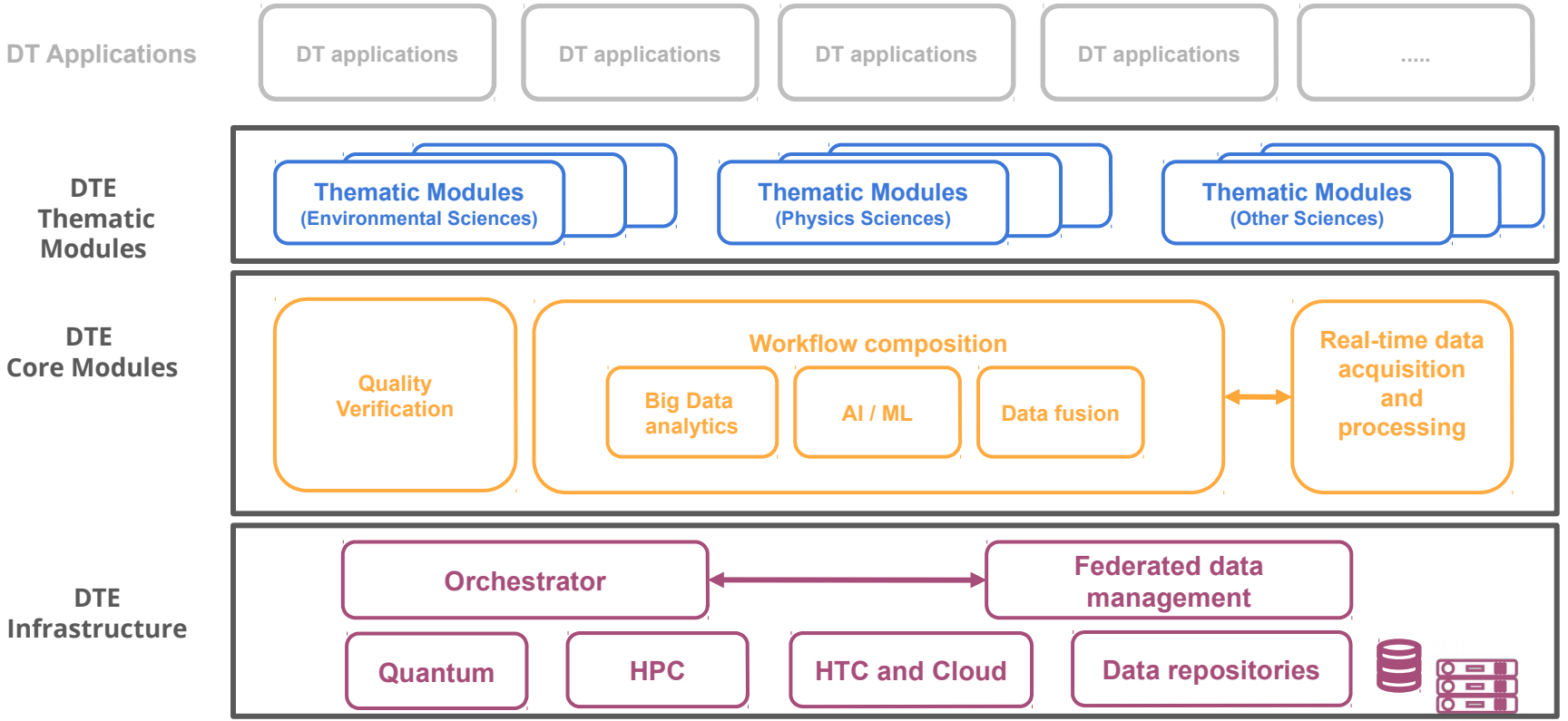
- **Interoperability:** A joint pilot activities with **DestinE** to **design a compatible architecture** that addresses the requirements of the largest set of user communities.

- **Collaboration with ECMWF:** Demonstrators of **data handling across interTwin and DestinE DTs** for the Extremes and Climate in production-type configurations.

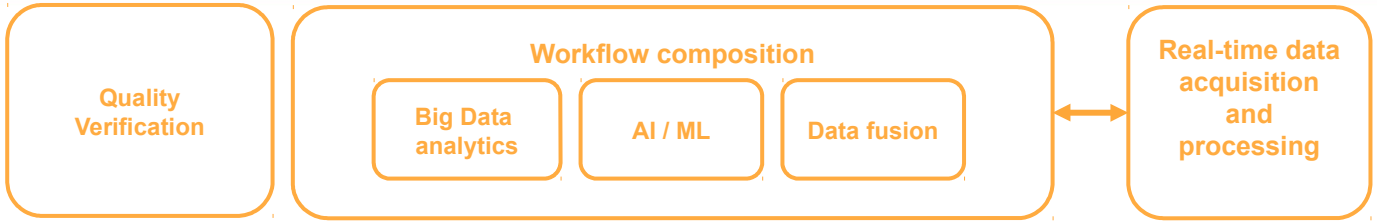
- **Collaboration with DestinE:** Development of **common software architecture concepts** that are also **applicable to other major DTs initiatives**.

ECMWF

interTwin components



DTE Core Capabilities [1 / 2]

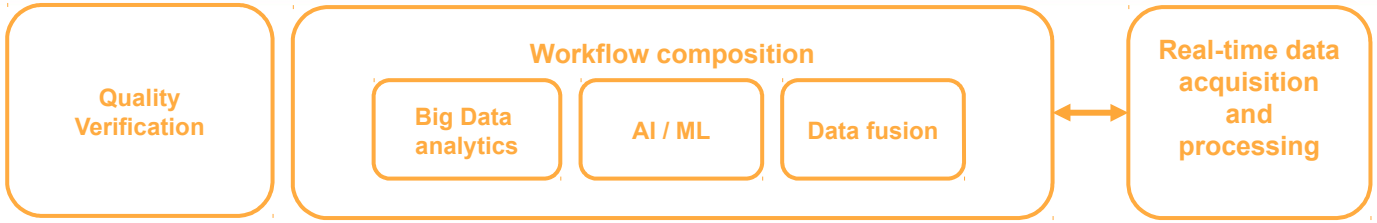


The DTE Core Modules offer horizontal capabilities to facilitate the creation and the operations of data-intensive and compute-intensive DT applications:

- **Advanced workflow composition**: executes DT workflows that can invoke other Core module capabilities. An **interdisciplinary processing graph** will serve as a link and API for the supported workflow engines guaranteeing a **common user experience** and facilitating the integration of discipline specific tools.

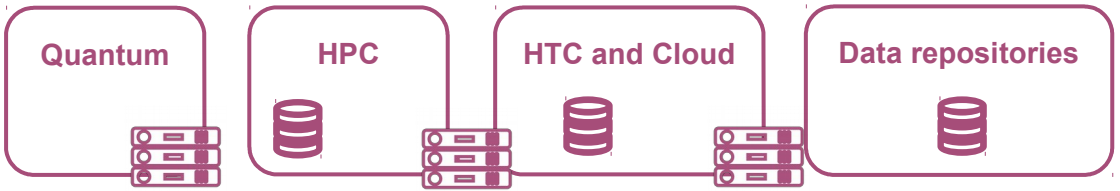
- **Data Fusion:** implements and integrates processes for merging datasets from different sources. This includes **linking of observational and modelled data**, and the **harmonization of different types of observational data** like gridded datasets with vector based datasets like point streams of data from ground stations.

DTE Core Capabilities [2 / 2]



- **AI workflow and method lifecycle management**: toolbox for realizing complex AI setup
- **Real time acquisition and data analytics**: delivers high performance data ingestion by applying the paradigm of "**serverless**" computing to DTs. This module expands the advanced workflow composition module to trigger on the fly processing upon data acquisition.
- **Validation, verification, and uncertainty tracing for model quality**: toolkit that provides developers with the possibility to **design and validate DT Models guaranteeing quality and reliability of the DT applications outputs**. Offered as a "Model Validation as a Service" enabling customisations of best practices and standard quality measures for scientific disciplines and applications.

DTE Infrastructure: Computing, Storage, Federation services and policies



- Solutions to provision access to a wide range of compute providers, including **HTC, HPC, Cloud, Quantum systems** and a stable interface to the storage systems of interest for the use cases in the project.
- Integral part of the Infrastructure are resource providers who allow access to their resources within the project
- Development of homogeneous security and access policies and resource accounting

DTE Infrastructure providers

HPC (6)

TU Wien
GRNET
PSNC
UKRI
JSI (EURO HPC)
JUELICH

Cloud (6+)

TU Wien
EODC
GRNET
PSNC
UKRI
JUELICH
EGI Federation

HTC (2+)

UKRI
KBFI
EGI Federation

DTE Infrastructure: AI Orchestrator and Federated Data Management

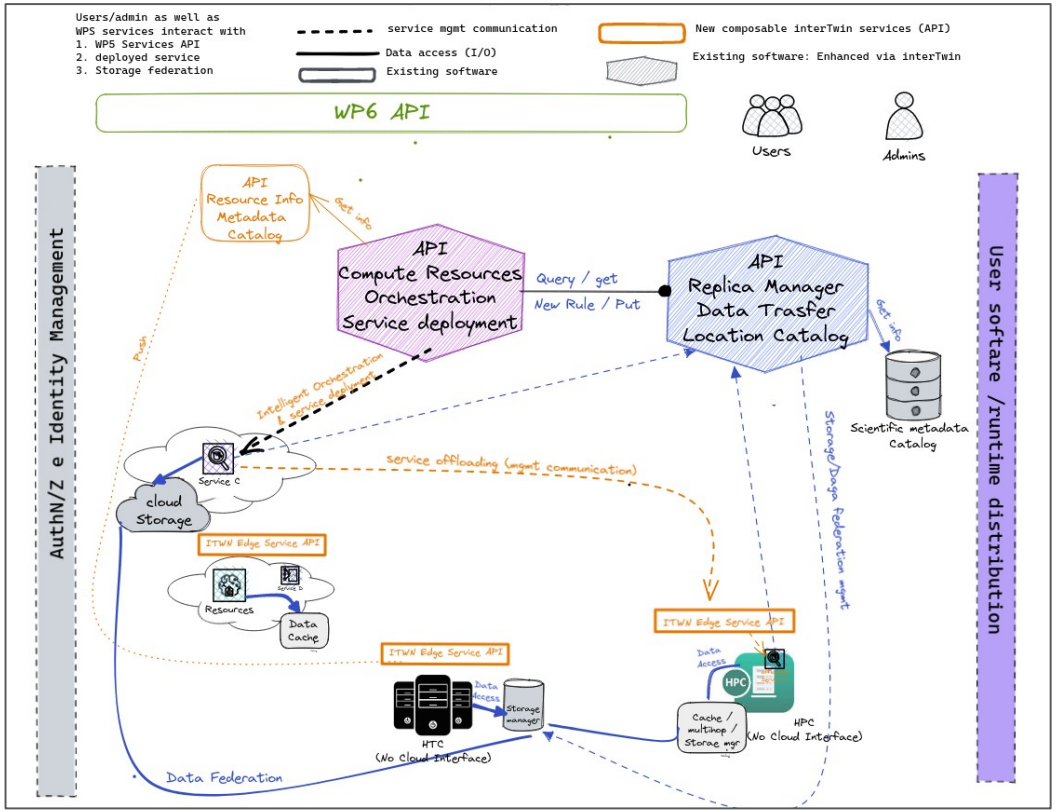
Challenge

To match storage and compute in complex workflows like those expected in the project, a series of non trivial decision must be taken (e.g. choose data source and compute sites, how to make data available to compute, etc).

- An **orchestrator with predictive AI capabilities** will be developed, basing its decisions on static configurations (site description, network connections, cost, ...) as well as on the data collected in previous runs

- The data management actions are implemented by a **Federated Data Management** layer interoperable with various repository types and including data transfer, caching tools and cataloguing systems. Through this layer the DTE can manage for instance HPC data ingress/egress

DTE Infrastructure Architecture Overview



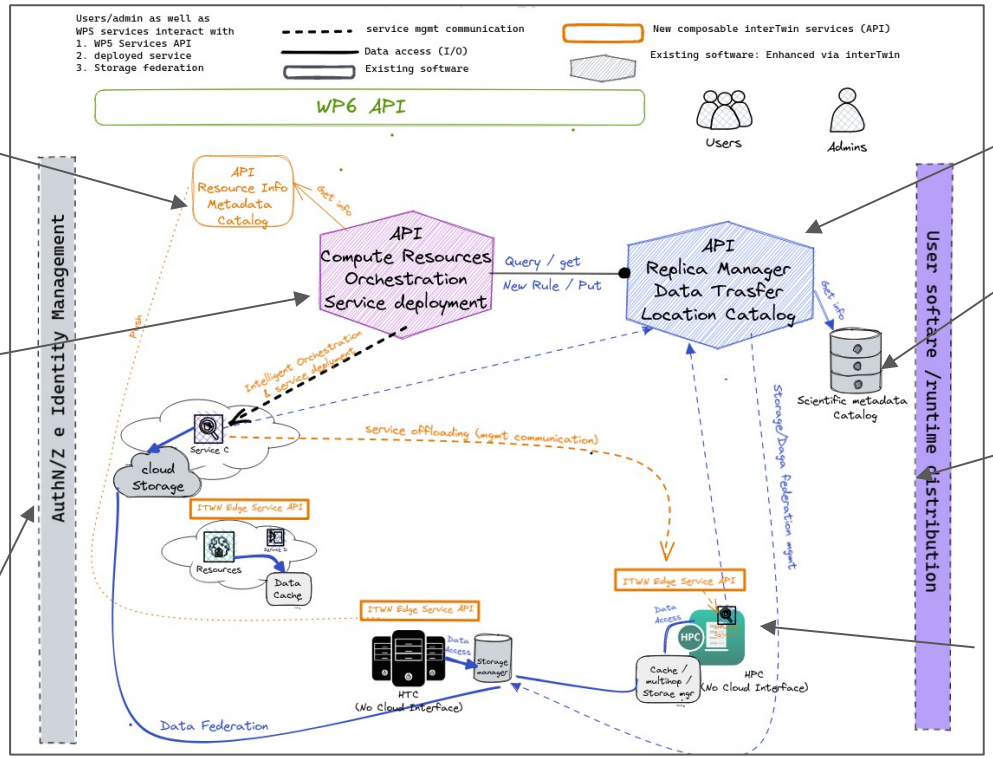
- Is based on and extends the **ESCAPE data lake blueprint** architecture and the **C-SCALE blueprint** for distributed cloud-based access of Copernicus data
- Architecture components as a combination of **existing and future developments** in software
- **Modular architecture** to make integration of loosely coupled systems feasible
  - i.e. distributed storage and HPC & HTC resources hosted by **EGI, PRACE & EuroHPC**

DTE Infrastructure Architecture
foundations & developments

WIP

Indigo PaaS Orchestrator

CheckIn & other OIDC providers



Rucio & FTS

TBD

CVMFS & other filesystems

TBD

Data Management Requirements

Implement new or update existing software to support data access and management in a federated environment, use-cases arising from DTE core and thematic modules.

- Consistent with the "building on **ESCAPE DataLake**" goal
- Event based (storage may announce the arrival of new data)

Transferring data into or out of storage systems – including HPC & Edge – with multi-hop as necessary

Abstraction layer to isolate application code from underlying storage technology

Tracking & querying **dataset locality**, managing **replicas** and **policy-based transfers** (to reconcile discrepancies)

Challenges for federated data

The DTE architecture is not yet completed and will get input from these areas:

- Interfacing with HPC resources: AuthN/Z, alternative protocols and mainly user-focused storage
- Interoperability with other infrastructures (such as DestinE, C-SCALE, OneData): different approaches, technologies and “openness”
- Support for event-driven processing
- Coping with unreliable network connections
- Potential multi-master interactions with data set definitions outside of RUCIO

Summary

interTwin is designing and testing the next-generation blueprint for AI/ML-driven federated computing infrastructure

- Requirements for the infrastructure emerge from several scientific use cases in a codesign approach
- The ESCAPE datalake (with RUCIO and FTS) will be the base for the federated data management in interTwin
- With new communities involved, there will be new requirements for RUCIO compared with the status quo
- The project will hopefully kickstart more widespread use of distributed workflows and make cooperation between different scientific fields and communities easier in the future

Thank you!

www.intertwin.eu

info@intertwin.eu

intertwin_eu

intertwin