Towards Extreme and Sustainable Graph Processing for Urgent Societal Challenges in Europe

Radu Prodan, Dragi Kimovski
University of Klagenfurt, Austria

Andrea Bartolini
University of Bologna, Italy

Michael Cochez, Alexandru Iosup
VU Amsterdam, The Netherlands

Evgeny Kharlamov
Bosch Center for Artificial Intelligence, Renningen, Germany

Jože Rožanec
Jožef Stefan Institute, Ljubljana, Slovenia

Laurenţiu Vasiliu
Peracton Ltd., Galway, Ireland

Ana Lucia Vărbănescu
University of Twente, The Netherlands

***Abstract*—The Graph-Massivizer project, funded by the Horizon Europe research and innovation program, researches and develops a high-performance, scalable, and sustainable platform for information processing and reasoning based on the massive graph (MG) representation of extreme data. It delivers a toolkit of five open-source software tools and FAIR graph datasets covering the sustainable lifecycle of processing extreme data as MGs. The tools focus on holistic usability (from extreme data ingestion and MG creation), automated intelligence (through analytics and reasoning), performance modelling, and environmental sustainability tradeoffs, supported by credible data-driven evidence across the computing continuum. The automated operation uses the emerging serverless computing paradigm for efficiency and event responsiveness. Thus, it supports experienced and novice stakeholders from a broad group of large and small organisations to capitalise on extreme data through MG programming and processing.**

Graph-Massivizer validates its innovation on four complementary use cases considering their extreme data properties and coverage of the three sustainability pillars (economy, society, and environment): sustainable green finance, global environment protection foresight, green AI for the sustainable automotive industry, and data centre digital twin for exascale computing. Graph-Massivizer promises 70% more efficient analytics than AliGraph, and 30% improved energy awareness for extract, transform and load storage operations than Amazon Redshift. Furthermore, it aims to demonstrate a possible two-fold improvement in data centre energy efficiency and over 25% lower greenhouse gas emissions for basic graph operations.

***Index Terms*—Extreme data, graph processing, serverless computing, sustainability**

I. Introduction

In 1736, mathematician Leonhard Euler solved the problem of traversing the seven bridges of Königsberg through a nascent form of graph theory [1]. It took nearly one hundred years until physicists like Gustav Kirchhoff combined graph theory with computational techniques to formulate the fundamental laws of modern electrical circuits and engineering. Pioneering solutions and discovering new problems succeeded rapidly, and today thousands of computational methods (algorithms) and findable, accessible, interoperable, and reusable (FAIR) graph datasets exist. However, current computational capabilities fail when faced with the extreme scale of existing graph datasets and their complex workflow. Even for known solutions, an average graph application can exceed one hundred years to complete on an average computer, threatening to offset the benefits of this technology for most use cases.

The use, interoperability, and analytical exploitation of graph data are essential for the European data strategy. Graphs or linked data are crucial to innovation, competition, and prosperity and establish a strategic investment in technical processing and ecosystem enablers. Graphs are universal abstractions [2] that capture, combine, model, analyse, and process knowledge about real and digital worlds into actionable insights through item representation and interconnectedness. For societally relevant problems, graphs are extreme data that require further technological innovations to meet the needs of the European data economy. Digital graphs help pursue the *United Nations Sustainable Development Goals (UN SDG)* by enabling better value chains, products, and services for more profitable or green investments in the financial sector and deriving trustworthy insight for creating sustainable communities. All science, engineering, industry, economy, and society-at-large domains can leverage graph data for unique analysis and insight, but only if graph processing becomes easy-to-use, fast, scalable, and sustainable.

The *Graph-Massivizer project* funded by the Horizon Europe research and innovation program of the European Union investigates a high-performance, scalable, gender-neutral, secure, and sustainable platform for multilingual information processing and reasoning based on the *massive graph (MG)* representation of extreme data in the form of general graphs, knowledge graphs and property graphs. They integrate patterns and store interlinked descriptions of objects, events, situations, concepts, and semantics. Graph-Massivizer supports the any-*volume* graph challenge by supporting up to billions of vertices and trillions of edges. It tackles the *velocity* graph challenge of dynamically changing topologies and proposes a novel *viridescence* graph challenge for sustainable processing at scale. The support for extreme data extends existing graph

Graph-Massivizer project receives funding from the Horizon Europe research and innovation program of the European Union.

978-1-6654-5113-0/22/$31.00 ©2022 IEEE
DOI 10.1109/CloudSummit54781.2022.00010

processing technological capabilities by orders of magnitude for at least one "V"-characteristic in four use cases.

The project delivers the *Graph-Massivizer toolkit* of five open-source software tools and FAIR graph datasets covering the sustainable lifecycle of processing extreme data as MG. The tools focus on holistic 1) usability (starting from extreme multilingual data ingestion and MG creation), 2) automated intelligence (through analytics and reasoning), 3) performance modelling, and 4) environmental sustainability tradeoffs supported by credible data-driven evidence 5) across high-performance computing (HPC) systems and computing continuum. The automated operation based on the serverless computing paradigm [3] protected by state-of-the-art cybersecurity measures supports experienced and novice stakeholders from a broad group of large and small organisations to capitalise on extreme data through MG programming and processing.

The Graph-Massivizer integrated toolkit proposes ambitious technological breakthroughs in line with the *Horizon Europe Strategic Plan* and data strategy, providing a unique, fundamental building block in its target green and digital transformation. The project leverages the world-leading roles of European researchers in graph processing and serverless computing. It uses leadership-class European infrastructure in the computing continuum, from pre-exascale HPC facilities to local computing clusters with state-of-the-art networking and in-place cybersecurity capabilities. Graph-Massivizer validates its innovation on four complementary use cases considering their extreme data properties and coverage of the sustainability pillars: economy, society, and environment. It also leverages leadership-class graph datasets from Europe, such as the LOD Laundromat [4], the most extensive FAIR collection of knowledge graphs worldwide with 38 billion facts and rules.

II. Graph-Massivizer Concept

Graph-Massivizer researches and develops a novel integrated toolkit for sustainable development and operation of MG processing on extreme data (see Figure 1).

The *graph operational layer* facilitates generating, transforming, and manipulating extreme data through *basic graph operations (BGO)*, comprising graph creation, enrichment, query, and analytics.

a) Graph creation: implemented by the Graph-Inceptor tool translates extreme data from various static and event streams or follows heuristics to generate synthetic data, persist it, or publish it within a graph structure.

b) Graph enrichment, graph query, and graph analytics: are three BGO implemented by the Graph-Scrutinizer tool. They analyse and expand extreme datasets using probabilistic reasoning and machine learning (ML) algorithms for graph pattern discovery, low memory footprint graph generation, and low latency error-bounded query response. The output is a new graph, a query, or an enriched structured dataset.

The *graph processing layer* provides sustainable and energy-aware serverless graph analytics on the underlying heterogeneous HPC infrastructure, following three phases.

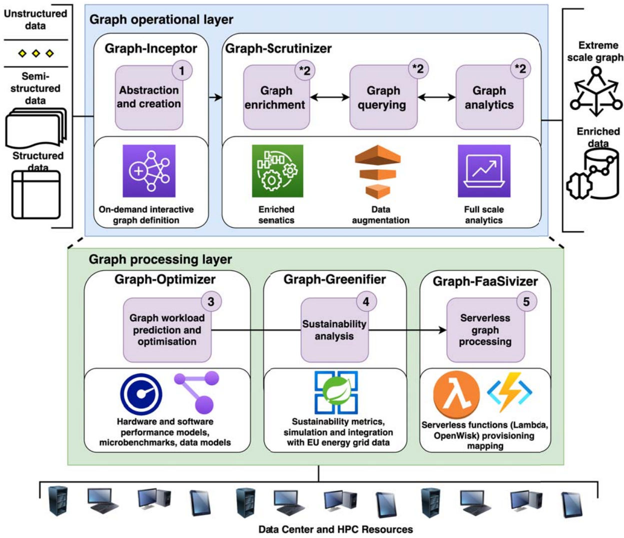

Fig. 1. Graph-Massivizer conceptual architecture.

c) Graph workload modelling and optimisation: represented by the Graph-Optimiser tool, analyses and expresses graph processing workloads into a workflow of BGO. It further combines parametric BGO performance and energy models with hardware models into accurate performance and energy consumption predictions for the workload running on a given multi-node, heterogeneous infrastructure of CPUs, GPUs, and FPGAs. The predictions indicate the most promising combinations of BGO optimisations and infrastructure, i.e., a codesigned solution for the given workload while guaranteeing its performance and energy consumption bounds.

d) Sustainability analysis: implemented by the Graph-Greenifier tool, collects, studies and archives performance and sustainability data from operational data centres and national energy suppliers on a large scale. This phase simulates multi-objective infrastructure sustainability profiles for operating graph analytics workloads, trading off performance and energy (e.g., consumption, CO_2, methane, greenhouse gas (GHG) emissions) metrics. Its ultimate goal is to model the impact of specific graph analytics workloads on the environment for evidence-based decision making.

e) Serverless BGO processing: implemented by the Graph-Serverlizer tool, uses performance and sustainability models and data from the previous phases to deploy serverless graph analytics on the computing continuum. It relies on novel scheduling heuristics, infrastructure partitioning and environment-aware processing for scalable orchestration of serverless graph analytics with an accountable performance and energy consumption tradeoff.

The *hardware infrastructure layer* considered by Graph-Massivizer consists of geographically distributed data centres distributed across the Cloud HPC, mid-range Fog, and low-end Edge computing continuum. A data centre is a collection of nodes connected through high-performance networks. A node represents a heterogeneous set of tightly coupled devices comprising commodity multiprocessors and specialised accelerators such as GPU or FPGA.

III. Graph-Massivizer Architecture

Graph-Massivizer researches and develops an modular architecture of five integrated tools, presented in this section.

A. Graph-Inceptor: Extreme Data Ingestion, Massive Graph Creation, and Storage

The Graph-Inceptor tool consists of two graph creation and storage components, detailed below.

a) Graph creation: consists of a series of data extract, transform and load conversions into graph data, parameterizable graph generators and techniques to stream the data to an endpoint or persist it to the graph storage. The ingested data undergoes a use case-specific preprocessing phase, comprising multilingual data reported in 100 languages. The ingestion supports batch and streaming, posing additional challenges related to data generation velocity. Graph-Massivizer envisages a solution compatible with Apache Spark, offering highly scalable processing capabilities with low latency. The graph creation phase consumes raw data or uses insights from the Graph-Scrutinizer tool to enrich the graph further.

b) Graph storage: generates complementary graph views covering two types of graphs: *RDF graphs* [5] modelling semantic aspects of the data and labelled *property graphs* addressing data modelled with edge properties. *RDF-star graphs* [6] logically unify the two views and cover most real-world use cases. Graph storage is a flexible component that stores data in multiple formats, depending on the graph and processing requirements. Moreover, it converts the data into a different form when repeating analytics requires switching from edge-centric to node-centric storage. Graph-Inceptor uses information from Graph-Scrutinizer to predict the type of ingested data, optimise the ingestion, and distribute the graph to the geographical locations where computational analytics occur. Data redundancy handled at the filesystem level ensures replication, fault tolerance and consistency across the nodes.

B. Graph-Scrutinizer: MG Analytics and Reasoning

The Graph-Scrutinizer tool exploits MG using analytics, querying, sampling, and probabilistic reasoning.

a) Graph analytics and querying: use statistical and ML models to provide critical insights for each use case. On top of the graph algorithms and ML models, Graph analytics supports arbitrary user queries to gain insights into facts and patterns observed in the graphs. This component builds on top of open-source software with proven distributed processing capabilities, such as Apache GraphX, to enable extreme data computation (i.e., billions of edges). Graph-Scrutinizer creates compatibility with PyTorch Geometric to scale ML models such as neural networks and other embedding approaches and integrates summarisation approaches to scale approximate reasoning techniques to these MG.

b) Graph sampling: optimises the analysis or query on a data subset depending on the operation type. The graph sampler supports a wide range of graph sampling strategies, such as random walks, meta path, node degree distribution, topK, edge weight, or pruning ones, such as the k-core graph. These techniques retain critical information from the original graph but reduce the size such that the analysis becomes feasible on small-scale infrastructure. Furthermore, they enable sophisticated analytics with a slight loss in precision [7] or lower energy consumption. While such results are probabilistic, scientific literature shows the insights are close to those obtained by processing the whole graph.

c) Probabilistic reasoning: provides capabilities to enrich MG with expanding extreme data sets, enable transparent graph queries, and provide insights into the underlying uncertain knowledge valuable to the stakeholders. For this purpose, it uses ML algorithms for graph pattern discovery, low memory-footprint graph generation (with no materialisation), real-time query latency, and error-bounded query response.

d) Graph enrichment: materialises persistent insights obtained through analytics or probabilistic reasoning into the graph and makes them available without repeated computation. Successive enrichments enlarge the information explicitly encoded in the graph, discover new patterns, and provide a rich ground for further analytics and reasoning.

C. Graph-Optimizer: Workload Modelling with Performance and Energy Guarantees

The Graph-Optimizer tool uses optimised BGO and composition rules to capture and model the workload. It further combines the workload model with hardware and infrastructure models, predicting performance and energy consumption. Combined with design space exploration, such predictions select codesigned workload implementations to fit a performance objective and guarantee their runtime bounds.

a) Hardware models: They capture the BGO-relevant *hardware operations (H-Ops)* per device (CPU, GPU, FPGA), including memory latency, bandwidth, throughput for basic arithmetic operations, or I/O communication bandwidth. Newly designed, specialised microbenchmarks determine runtime and energy costs for H-Ops.

b) Graph processing workload models: All BGO and composition rules model sequences of H-Ops. Graph-Optimizer proposes a formalism to express workloads combining BGO in sequences of H-Ops through composition rules, following ideas like GraphBLAS [8] and GBTL-CUDA [9]. Graph-Optimizer generates different workload designs and the corresponding models using automated design-space exploration and selects promising designs for implementation (via semi-automated code generation), upscaling, and deployment using workload prediction.

c) Workload execution prediction: Workload (i.e., H-Ops sequences) and hardware models (i.e., H-Ops costs) combined in the presence of data models for the actual inputs can predict the performance and energy consumption on heterogeneous (multi-node, multi-device, heterogeneous) infrastructures.

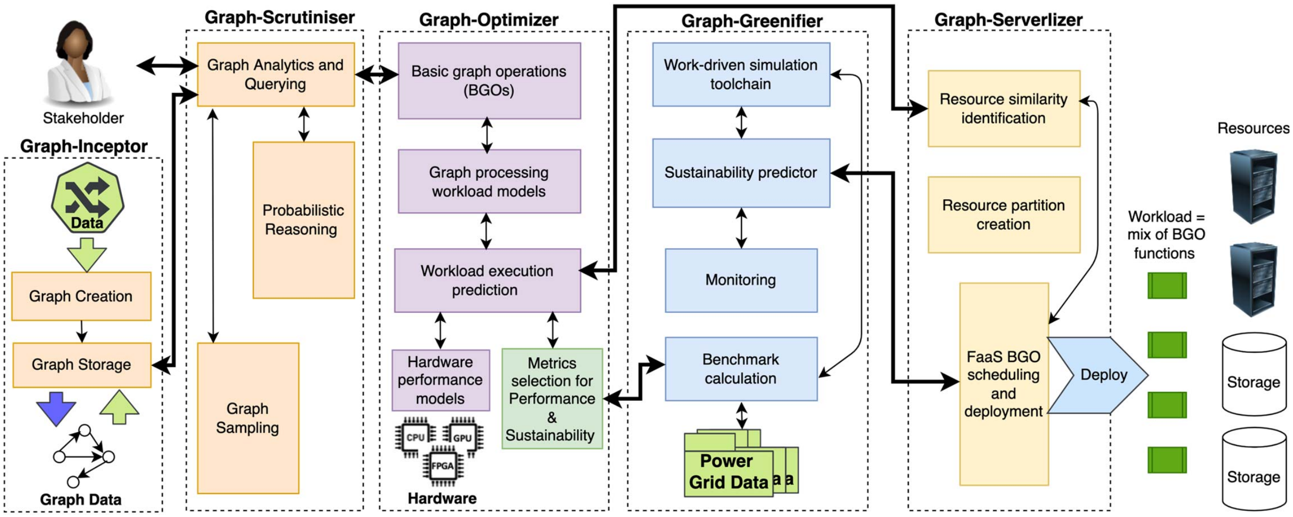
Stakeholder
Graph-Inceptor
Data
Graph Creation
Graph Storage
Graph Data
Graph-Scrutiniser
Graph Analytics and Querying
Probabilistic Reasoning
Graph Sampling
Graph-Optimizer
Basic graph operations (BGOs)
Graph processing workload models
Workload execution prediction
Hardware performance models
Metrics selection for Performance & Sustainability
CPU
GPU
FPGA
Hardware
Graph-Greenifier
Work-driven simulation toolchain
Sustainability predictor
Monitoring
Benchmark calculation
Power Grid Data
Graph-Serverlizer
Resource similarity identification
Resource partition creation
FaaS BGO scheduling and deployment
Deploy
Workload = mix of BGO functions
Resources
Storage
Storage

Fig. 2. Graph-Massivizer technical architecture.

D. Graph-Greenifier: Sustainable and Energy-aware Massive Graph Processing

The Graph-Greenifier tool provides sustainability analysis and decision-making capabilities for extreme graph processing workloads. It aims to inform providers and consumers on operational sustainability aspects, requiring mutual information sharing, reducing the energy consumption for graph analytics, and increasing the use of electricity from renewable sources. Graph-Greenifier has four design components.

a) Monitoring: Graph-Massivizer identifies relevant sustainability metrics based on extensive state-of-the-art analysis and establishes effective means to monitor and calculate them depending on installed measurement capabilities. For example, it uses software energy counters where hardware counters and power meters are unavailable and deploys the infrastructure to collect and archive these metric calculations.

b) Power grid data interface: Graph-Greenifier automates data gathering based on the offer and price of electrical energy on the open market and its energy source and greenness. It uses monitoring data from local data centre operators such as the SURF operational dataset [10], availability insights from cloud operators and third-party aggregators of user reports such as Outage Report[1], and energy- and sustainability-related data published by national infrastructures such as the EU-wide ENTSO-E (https://www.entsoe.eu/) transparency platform or the Dutch national data source (https://energieopwek.nl/) coupled with credible energy-consumption analyses such as those provided by national statistics bureaus.

c) Work-driven simulation toolchain: Graph-Greenifier utilises and extends the OpenDC [11] simulator, encompassing public information from national energy suppliers to model the impact of graph processing on the climate and, therefore, society. The simulator uses the selected sustainability indicators, such as carbon footprint, CO_2 and methane emissions calculated by the sustainability predictor, to estimate the impact of different scenarios.

d) Sustainability predictor: Graph-Greenifier extends and upscales Graph-Optimizer's predictions to rank graph processing scenarios based on performance, energy efficiency and sustainability at scale. Data centre operators use this ranking to choose the most sustainable operational procedures at runtime, with the help of Graph-Serverlizer that steers the workload to more appropriate infrastructure resources. The process is transparent and evidence-based.

e) Sustainability benchmark: Graph-Greenifier proposes a sustainability benchmark providing runtime energy labels, including information about energy sources derived from data centres and energy operation models. Graph-Greenifier creates a closed monitoring loop, encompassing simulation and estimations of the emitted GHG pollutants for performing the BGO processing. The GHG estimates engage the data centre operators and other stakeholders in meaningful dialogues for reaching informed graph processing decisions with reduced impact on the environment.

E. Graph-Serverlizer: Scalable Serverless Graph Analytics over a Codesigned Continuum Infrastructure

The Graph-Serverlizer tool encapsulates BGO as serverless functions and automates their deployment on the computing continuum according to the performance and sustainability metrics and labels communicated by Graph-Optimizer and Graph-Greenifier. The serverless technology allows deploying BGO with minimal operational delay and reduced *"pay-as-you-go"* financial cost. Furthermore, it lessens the burden on developers by providing transparent runtime management. Graph-Serverlizer has three components.

a) Similarity resource partitioner: Graph-Serverlizer employs a three-step approach that partitions the underlying

[1] https://github.com/atlarge-research/outage_report_characterization

infrastructure nodes by considering the resources, I/O, and network BGO requirements. It first applies resource extraction to identify HPC, Cloud and Edge codesigned infrastructure nodes' characteristics based on resource types, such as processing cores, memory, storage, and energy consumption. Next, the multilayer infrastructure facilitation clusters the infrastructure nodes comprising specialised hardware (GPUs, FPGAs) requested by the Graph-Optimizer design(s) in resource layers based on their topological (betweenness centrality) and similarity relationships between the related resources. Finally, the layer partitioning clusters each resource layer of the multilayer infrastructure in disjoint resource partitions of nodes with similar resource types, including operational delay, sustainability, and energy consumption metric.

b) Sustainable BGO function operation: Graph-Serverlizer targets three BGO function operation phases, considering heterogeneous resources and sustainability profiles. Firstly, feature partitioning identifies nodes with similar features, cost and sustainability characteristics to the resource requirements of the BGO encompassed as functions. Afterwards, function scheduling allocates appropriate virtual instances within the nodes of the same feature partition and highly connected network layer partition based on the performance and sustainability metrics provided by Graph-Optimizer. Graph-Serverlizer envisions a scheduling algorithm inspired by matching theory opposing two conflicting players (i.e., BGO and hardware nodes), bounded by the sustainability metrics provided by Graph-Greenifier.

c) BGO serverlization: implements function wrapping techniques for BGO and the required execution libraries for a set of serverless platforms, such as OpenWisk or Lambda.

d) Cybersecure deployment: addresses runtime aspects of the serverless functions and provides elastic and scalable deployment of the BGO while minimising the operational costs. It further features real-time sustainability analysis and automated decision-making with a reduced negative impact on the environment. Graph-Serverlizer employs state-of-the-art security and privacy mechanisms [12] installed at the use case providers to protect against malicious attacks.

IV. Graph-Massivizer Use Cases

Graph-Massivizer selects four *real-world use cases (UC)* with complementary economic, social, and environmental sustainability profiles. They need to tackle complementary extreme data processing and massive graph analytics challenges, going order of magnitude beyond big data in at least three "V"-characteristics each. Graph-Massivizer proposes a novel "V"-characteristic called "Viridescence", representing the sustainability of processing extreme data from an environmental perspective. Table I summarises seven V-characteristics enhanced from "big-to-extreme" dimensions and their balanced distribution across the four use cases.

TABLE I
"Big-to-extreme" V-characteristic enhancements in Graph-Massivizer use cases.

	Volume	*Velocity*	*Value*	*Veracity*	*Variety*	*Viscosity*	*Viridescence*
UC-1	✓		✓	✓			✓
UC-2	✓				✓	✓	✓
UC-3		✓		✓	✓		✓
UC-4	✓	✓			✓		✓

A. UC-1: Green and Sustainable Finance

Green finance focuses on financial products, investments, and services that channel investment into green-focused companies. Green finance comprises green bonds, green lending, and green equity investment. Green finance aims to achieve economic growth while reducing pollution and GHG emissions, reducing waste, and improving efficiencies. Sustainable finance considers environmental, social, and governance (ESG) factors for investment decisions leading to long-term sustainable economic activities and projects. UC-1 focuses on improving and optimising green investments and trading, facing significant barriers. Existing historical securities' data, particularly on ESG data (starting from the early 2010s), is not enough for in-depth, massive volume testing, de-risking financial algorithms and training ML models due to erroneous, scattered, unavailable, proprietary, incomplete, or expensive data. The ever-growing complexity of novel ML-driven financial algorithms that require much data for advanced ML training and simulations further amplifies the data scarcity. Therefore, a common practice is using one historical record per security to optimise a financial strategy. However, a financial model that overfits the training data shows good returns and low risk but can deliver disastrous financial losses in live trading. Few ML insurance policies are commercially available and considered risky as they do not cover ML losses.

a) Synthetic Financial Data Multiverse: Graph-Massivizer aims to remove the limitations of financial data providers (low volume, accessibility, high costs) by enabling semi-automated creation of realistic and affordable synthetic extreme financial data sets, unlimited in size and accessibility. Financial companies use the financial multiverse for improved ML-based green investment and trading simulations, free of critical biases such as prior knowledge, overfitting, and indirect contaminations due to present data scarcity.

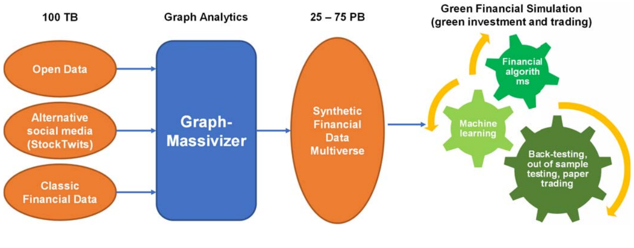
100 TB
Graph Analytics
25 – 75 PB
Green Financial Simulation (green investment and trading)
Open Data
Alternative social media (StockTwits)
Classic Financial Data
Graph-Massivizer
Synthetic Financial Data Multiverse
Financial algorithms
Machine learning
Back-testing, out of sample testing, paper trading

Fig. 3. UC-1: green and sustainable finance.

b) Methodology: UC-1 (Figure 3) defines a *financial MG (F-MG)* as a hybrid metadata structure representing time-series, text, values, boolean and monetary taxonomies. On top of it, it researches novel financial algorithms and testing methodology operating in five steps: 1) map the historical financial data structure into an F-MG; 2) generate synthetic data using the graph that preserves the original historic data statistical features; 3) interpolate missing data (e.g. gaps, errors) using ML-based inference and reasoning; 4) simulate various green financial investment and trading, and 5) recommend the "greenest" investments and trading opportunities.

c) Extreme financial data (Table II): UC-1 uses five types of significant financial data sources of 10 TB in size: purchased historical data, data acquired from financial data providers (e.g., S&PCapitalIQ, Thomson Reuters, Factset), open social media data (e.g., StockTwits, Twitter), open gov. data (USA, UK, Ireland), and various NGO repositories fulfilling the GDPR requirements. Based on this data, Graph-Massivizer aims to generate an extreme synthetic financial data multiverse of 1 PB to 5 PB for highly advanced and accurate financial simulations (backtesting). UC-1 will provide a *Financial Data Multiverse* service to stream this extreme volume of data to financial customers' facilities and offer selective samples for internal testing and open data.

TABLE II
EXTREME FINANCIAL DATA CHARACTERISTICS.

Data Characteristic	*Big financial data state-of-the-art*	*Extreme financial data dimension*
Volume	Real financial data: 10 TB	Synthetic streaming data: 1 PB to 5 PB
Value	> €200000 per year	€5000 – €30 000 per year
Veracity	20% missing data	100% complete data
Viridescence	Unsustainable resource intensive analytics	Sustainable and energy-accountable graph analytics

B. UC-2: Global Foresight for Environment Protection

Global foresight for environment protection focuses on geopolitical and business aspects related to sustainable environmental goals, including climate action, responsible production and consumption patterns, clean water and sanitation, and clean and affordable energy. The foresight comprehends insights on future trends and scenarios to guide decision-making in developing better policies. A contextual graph built through data available from the Common Crawl, Linked Open Data Cloud, and global media news provides unique insights into the convergence of the three societal systems (economy, politics, science) in mass media. Foresight requires complex data and operations to extract causal templates, understand future events, and predict consequences. Geopolitical and business foresight concerning global warming and the environment are essential for taking timely decisions, maximising impact, and mitigating negative scenarios.

a) Environment Protection Foresighter: Graph-Massivizer removes the limitations of classic strategic geopolitical and business foresight (e.g., expert surveys) by resorting to MG encoding crawling Web data and multilingual live media news. The ML-based analysis enables frequent horizon scanning, megatrend analysis, and higher scalability for processing and combining different data sources and events while avoiding human-prone biases. UC-2 provides an innovative subscription-based foresight service, targeting the four environmental SDGs of the UN: climate action, responsible production and consumption, clean water and sanitation, and clean and affordable energy.

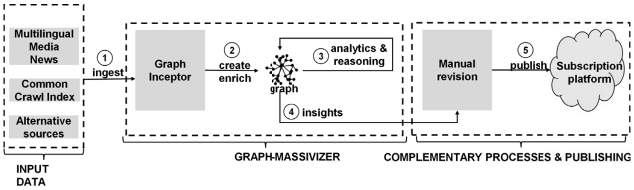

Fig. 4. UC-2: global foresight for environment protection.

b) Methodology: UC-2 (Figure 4) introduces the *foresight MG (FOR-MG)* as a hybrid structure consisting of Web data (Common Crawl Index), multilingual news media data, and expert domain knowledge correlating past scenarios with current environmental geopolitical and business events. ER fulfils its goals by mapping historical news media events and web data structure into a FOR-MG, sampling the graph data while preserving essential structure information and performing analytics using statistical, ML models, and reasoning methods. Horizon scanning systematically analyses new data to identify potential threats, risks, and emerging issues and opportunities (e.g., predicting new FOR-MG nodes and their relationships). Megatrend analysis explores the relation of subgraph structures to shifts in behaviours and attitudes, geographies, and industries (e.g., through embeddings clustering and data drift analysis). It predicts their impact by projecting their expected structure on forecasting future nodes and relations (link prediction). Human experts analyse the analytics insights to provide a reasonable interpretation.

c) Extreme news media data (Table III): UC-2 uses Web data available at the Common Crawl, Linked Open Data Cloud, and historical and streaming news media data provided by a media event monitoring system. Both datasets encompass over seven years of multilingual data and enable the creation of an ever-evolving dynamic graph of billions of nodes and trillions of edges describing events and meaningful context.

TABLE III
EXTREME FORESIGHT DATA CHARACTERISTICS.

Data Characteristic	*Big foresight data state-of-the-art*	*Extreme foresight data dimension*
Volume	Millions of vertices and edges	Billions of vertices and edges
Viscosity	Manual, weekly insight creation	Automatic daily insight creation
Variety	Media data in nearly 100 languages	Common Crawl Index, Linked Open Data Cloud
Viridescence	Unsustainable resource intensive analytics	Sustainable and energy-accountable graph analytics

C. UC-3: Green AI for Sustainable Automotive Industry

Following Industry 4.0 in manufacturing, the automotive industry undergoes a massive transformation towards digitalisation across the whole value chain, from multi-tier suppliers and original equipment manufacturers to logistics, customers, and recycling companies. The automotive value chain involves extreme data flows of heterogeneous, distributed, fast-growing and often disconnected or hardly compatible information. ML methods face new challenges and opportunities to holistically analyse the massive and unprecedented data integrated across these chains, to support decisions that fundamentally change automotive manufacturing processes towards a sustainable, circular, and climate-neutral automotive industry.

a) Green Manufacturing Line Diagnoser: Graph-Massivizer enables new graph-based encoding that captures several value-chain stages to predict their outcome better and detect anomalies. Better and quicker analysis prevents defect propagation and unnecessary waste, contributing to a sustainable, circular, and climate-neutral automotive industry. By combining graph-based ML methods with digital twins, Graph-Massivizer provides new insights and boosts the efficiency and scalability of the diagnosis beyond that of more expensive alternatives (e.g., excessive sensor deployment for continuous monitoring). The insights gained will help optimise manufacturing operations and improve the operational quality of the resulting products.

b) Methodology: UC-3 (Figure 5) defines manufacturing *MG (Man-MG)* that integrates a collection of digital twins expressed as OWL 2 ontologies and SHACL constraints. Man-MG captures several domains of discrete manufacturing, such as welding, specifications of production robots and the embedding manufacturing environment and products such as car bodies-in-white (BiW). The digital twins align with the ISO 18 278, 14 327, and 14 732 standards that guarantee scalability to various discrete production machines and processes for broad adoption. UC-3 plans to enrich and integrate its extreme welding data as a Man-MG enhanced with digital twins in the form of ontologies that determine the shape of the data and allow for ML reasoning. UC-3 evaluates the ML methods for Man-MG alignment, summarization, search, pattern mining, quality assurance and trustworthiness analyses. UC-3 demonstrates their benefits for automated welding of BiW and surrounding manufacturing environments by predicting the quality of welding spots produced by welding machines, their optimal configuration, remote diagnostics, schedules of their repairs and accessory replacements.

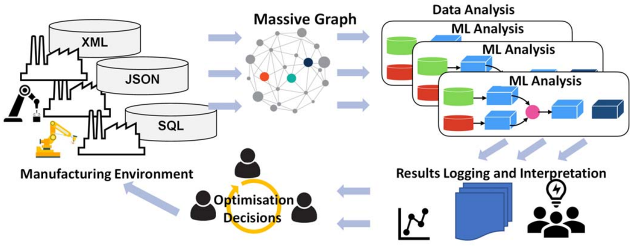

Fig. 5. UC-3: green AI for the sustainable automotive industry.

c) Extreme manufacturing data (Table IV): One BiW assembly line with 1500 car bodies per day comprises approximately 14 billion daily data items, corresponding to half a billion records in large integrated tables with more than 250 fields. Sensor data from one welding spot consists of 1000 samples. Each BiW contains roughly 6000 spots resulting in tens of billions of samples per day for one factory with several BiW productions lines. One robot's welding controlling system has up to 10 000 programs yielding millions of robotic program combinations across a production line.

TABLE IV
EXTREME MANUFACTURING DATA CHARACTERISTICS.

Data Characteristic	*Big manufacturing data state-of-the-art*	*Extreme manufacturing data dimension*
Variety	100s of measurement types	1000s of IoT-driven measurement types
Velocity	Gigabytes per day	Terabytes per day
Veracity	Sparse, incomplete, and low-quality data	Complete high-quality synthetic data
Viridescence	Unsustainable resource intensive analytics	Sustainable and energy-accountable graph analytics

D. UC-4: Data Centre Digital Twin for Sustainable Exascale Computing

Supercomputers are the backbone of HPC, supporting computational science discoveries and massive engineering analyses. They maximise their societal and economic impact through high computation and scientific throughput per investment. As the community focuses on peak performance in its race towards exascale machines, two critical factors limit HPC sustainability. Firstly, energy consumption is a vital factor in the *total cost of ownership (TCO)* of data centres and a de-facto barrier to their peak performance. While data-driven heat dissipation models exist in digital devices, they do not capture complex spatiotemporal dependencies between the cooling equipment, computing nodes, and computational workloads. Secondly, system utilisation is critical and directly impacts a supercomputer's productivity, quantified by the science, research and innovation throughput. However, system utilisation is hard to maximise while preserving fairness and fulfilling the requirements of jobs and workloads.

a) Data Centre Digital Twin: Graph-Massivizer targets "sustainable science throughput" through scalable energy-aware, exascale operation and traceable TCO understanding, including sustainability indicators and their environmental effects (e.g., GHG emissions). The Graph-Massivizer tools will enable the creation of a novel, graph-based digital twin of a data centre; this digital twin will further support the construction of sustainable exascale computing operational models to support scientific discovery in the next decade.

b) Methodology: UC-4 (Figure 6) leverages the holistic monitoring data to produce a *data centre MG (DC-MG)*, representing a digital twin describing spatial, semantic, and temporal relationships between the monitoring metrics, hardware nodes, cooling equipment and jobs. The DC-MG supports the deployment of performance prediction and what-if analysis using ML methods to change its configuration for maximising utilisation or sustainability requirements and observing the effects in simulation. For this purpose, UC-4 relies on a mathematical and visually connected DC-MG model to locate undesired effects like highly demanded racks with prohibitive energy consumption. UC-4 investigates different clustered partitions of the HPC facility tuned for incoming workloads with accurate processing and queuing time estimates to improve its operational model based on these DC-MG analyses. UC-4 targets a sustainable computing operation at exascale by optimising two parameters. Firstly, it improves the power usage effectiveness and GHG emissions of the data centre by creating and training DC-MG to capture the spatiotemporal-ontological dependencies among computation, computing nodes, and cooling equipment and predict the impact of the spatial power distribution on cooling efficiency and cost. Secondly, it improves the global resource utilisation based on predictive workload, resource consumption and job queuing models, maximising the science throughput.

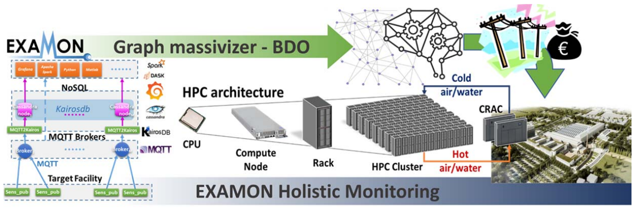

Fig. 6. UC-4: data centre digital twin for exascale computing.

c) Extreme HPC monitoring data (Table V): UC-4 uses the Marconi100 data centre scaled to the EuroHPC Leonardo (https://leonardo-supercomputer.cineca.eu/) pre-exascale supercomputer. The holistic monitored data includes approximately one million sensors producing 21 000 metrics per second on the M100 system. The Leonardo supercomputer overpasses the complexity of the graph by two orders of magnitude compared to the current deployment.

TABLE V
EXTREME HPC MONITORING DATA CHARACTERISTICS.

Data Characteristic	*Big HPC data state-of-the-art*	*Extreme HPC data dimension*
Volume	10 TB of monitoring data	1 PB
Variety	21 000 metrics per second	2 million metrics per second
Veracity	1 million metrics	10 million metrics
Viridescence	Unsustainable resource intensive analytics	Sustainable and energy-accountable graph analytics

V. CONCLUSIONS AND AMBITION

Graph-Massivizer brings the opportunity for European green financial investments, automotive, and media industries to accelerate at supercomputing speed and get a competitive advantage on graph-based powerful analytics, with evidence of improved performance and sustainability. While other forms of analysis rely on present assumptions about *"what happened"* or *"what happens"*, correctly building and employing graphs can further reveal the predictive patterns suggesting what *"might happen"* with clear evidence for each connection or inference step. Graph processing facilitates solving problems in many use cases driven by metrics related to costs, fraud, equipment failure, and inefficiencies and enables extra revenues from better intelligence. The large-scale graph analytics market still traverses a developing phase, hampered by the lack of technology research and use case adoption. Graph-Massivizer provides for Europe these missing links.

Graph-Massivizer's ambition: is to brings a new dimension to the scale and complexity of extreme data analytics of European ICT security products and services. Graph-Massivizer aims to lower the extreme data streaming ingestion latency to 500 ms and targets an analytics throughput of over three million triples per second for MG with ten billion nodes and 100 billion edges. Graph-Massivizer will achieve this scalability through 80% accurate BGO performance and energy consumption models of codesigned commodity and specialised hardware accelerators while reducing energy use two-fold and GHG emissions by 25% for BGO. A serverless operational engine validates the promise of targeting 40% faster deployment and 70% faster graph analytics than business enterprise solutions like Aligraph. Graph-Massivizer give the EU a winning position in this area with very few European vendors and no uncertain paths.

The Graph-Massivizer project will start on January 1, 2023 and will last three years.

REFERENCES

[1] L. Euler, "The solution of a problem relating to the geometry of position," *Commentarii academiae scientiarum Petropolitanae*, vol. 8, pp. 128–140, 1741.

[2] S. Sakr *et al.*, "The future is big graphs: a community view on graph processing systems," *Communications of the ACM*, vol. 64, no. 9, pp. 62–71, Sep. 2021.

[3] C. Abad, I. Foster, N. Herbst, and A. Iosup, "Serverless computing (Dagstuhl seminar 21201)," *Dagstuhl Reports*, vol. 11, no. 4, pp. 34–93, 2021.

[4] W. Beek, L. Rietveld, H. R. Bazoobandi, J. Wielemaker, and S. Schlobach, "LOD laundromat: A uniform way of publishing other people's dirty data," in *The Semantic Web – ISWC 2014*, ser. LNISA, vol. 8796. Springer, 2014, pp. 213–228.

[5] V. Bonstrom, A. Hinze, and H. Schweppe, "Storing RDF as a graph," in *IEEE/LEOS 3rd International Conference on Numerical Simulation of Semiconductor Optoelectronic Devices*. IEEE, 2003, pp. 27–36.

[6] B. Kasenchak, A. Lehnert, and G. Loh, "Use case: Ontologies and rdf-star for knowledge management," in *The Semantic Web: ESWC 2021 Satellite Events*. Springer, 2021, pp. 254–260.

[7] G. Salha, R. Hennequin, V. A. Tran, and M. Vazirgiannis, "A degeneracy framework for scalable graph autoencoders," in *IJCAI'19: Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3353–3359.

[8] A. Buluç and J. R. Gilbert, "The combinatorial BLAS: design, implementation, and applications," *The International Journal of High Performance Computing Applications*, vol. 25, no. 4, pp. 496–509, 2011.

[9] P. Zhang, Bloomington, A. Lumsdaine, S. Misurda, and S. McMillan, "GBTL-CUDA: Graph algorithms and primitives for gpus," in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2016, pp. 912–920.

[10] O. Laursen, V. Kristian, A. Uta, A. Iosup, P. Melis, D. Podareanu, and V. Codreanu, "Beneath the SURFace: An MRI-like view into the life of a 21st century datacenter," https://zenodo.org/record/3878143, 2020.

[11] F. Mastenbroek *et al.*, "OpenDC 2.0: Convenient modeling and simulation of emerging technologies in cloud datacenters," in *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2021, pp. 455–464.

[12] M. Golec, R. Ozturac, Z. Pooranian, S. S. Gill, and R. Buyya, "iFaaSBus: A security- and privacy-based lightweight framework for serverless computing using iot and machine learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3522–3529, May 2022.