EXTRACT: A distributed data-mining software platform for EXTReme dAta aCross The compute continuum

Alba Cañete Garrucho
alba.canete@bsc.es
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

Xavier Palomo Teruel
xavier.palomo@bsc.es
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

Oriol Feliu Juárez
oriol.feliu@bsc.es
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

Elli Kartsakli
elli.kartsakli@bsc.es
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

Eduardo Quiñones Moreno
eduardo.quinones@bsc.es
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

KEYWORDS
Extreme data, Compute Continuum, Reinforcement Learning, Decision-Making, Societal impact, Fairness, Ethics

ACM Reference Format:

1 PROJECT INTRODUCTION

Data has become the fuel of today’s digital economy, driving innovation and enabling decision-making across various domains[4]. Extreme data, characterized by its massive scale, diverse composition, and dispersed sources, presents a great challenge to conventional computing methods.

High-performance computing (HPC) effectively addresses handling large volumes of data and complex computations at unparalleled speeds. On the other hand, in the context of extreme data, edge computing enables real-time processing at the source of data generation allowing to handle data streams with high velocity in scenarios where immediate insights are required. Lastly, cloud computing complements HPC and edge computing by providing virtually unlimited scalability and storage capacity.

The Compute Continuum is a conceptual all-encompassing framework connecting this diverse range of computing environments, spanning from large-scale, centralized cloud data centers to smaller, decentralized edge devices, as well as high-speed HPC systems. This framework provides a cohesive infrastructure for addressing the challenges posed by extreme data.

By harnessing the strengths of the Compute Continuum, the Horizon Europe EXTRACT project (2023–2025)[5] seeks to enable the development, deployment, and execution of data mining workflows that address the unique challenges posed by extreme data characteristics. This comprehensive approach not only ensures the reliability and scalability of data processing but also contributes to creating positive impacts across various societal, scientific, environmental, and economic domains.

As seen in Figure 1, the EXTRACT software platform is structured in four main software layers. The lower layer, termed the Compute continuum, implements an Interoperability Abstraction Layer using Kubernetes multiclusters[7][2]. This layer handles the deployment and scheduling of data-mining workflows, abstracting heterogeneous resources for users. Above it lies the Data-driven orchestrator, which has explicit awareness of extreme data characteristics and workflow description. Its role is to select the most appropriate computing resource, as well as to enable data-driven mechanisms to minimize data movements across the compute continuum, utilizing technologies such as COMPSs[8] and Lithops[14]. The next layer on the figure is the Data mining framework. The technologies used in this layer, like TensorFlow[1] and PyTorch[11],
will be integrated with the Data Infrastructure layer, enabling workflows to consume data directly from the object storage content layer (OVHCloud S3[10]) or leveraging the metadata (Nuvla Datacatalog[13]), semantic and staging engine layers for data preparation and ETLs tasks.

2 PER USE CASE

Urban areas face numerous threats, one of them being natural disasters like earthquakes and floods. Responding effectively to these emergencies requires advanced planning and real-time decision-making to ensure the safety and well-being of residents and visitors. In this context, EXTRACT presents a novel approach to emergency preparedness with its Personalized Evacuation Routing (PER) use case.

The PER system developed by EXTRACT is going to be tested in the city of Venice, which is characterized by its intricate network of canals and bridges, as well as its susceptibility to flooding. In such a complex urban landscape, traditional evacuation strategies may fall short. The PER system addresses the critical need for efficient and equitable evacuation routes in urban environments, with a focus on fairness and effectiveness.

In conclusion, the EXTRACT’s PER system represents a significant advancement in urban emergency preparedness, providing a holistic solution merging state-of-the-art technology with ethical and equitable principles. By empowering individuals with personalized evacuation routes, the use case aims to enhance the safety and resilience of urban communities during crises.

3 ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union’s Horizon Europe Programme under the EXTRACT Project, grant agreement n°101093110.

REFERENCES