DEEP-Hybrid DataCloud project: Hybrid services for distributed e-infrastructures

A. Costantini[1], D.C. Duma[1], G. Donvito[2], J. Marco de Lucas[8],

[1] INFN-CNAF, Bologna, Italy
[2] INFN Bari, Bari, Italy
[1]1 Univ. de Cantabria, Spain

E-mail: `alessandro.costantini@cnaf.infn.it`

Abstract. DEEP Hybrid DataCloud is an Horizon 2020 project that addresses the need to support intensive computing techniques that require specialized HPC hardware, like GPUs or low-latency interconnects, to explore very large datasets. Launched in November 2017 the H2020 DEEP Hybrid-DataCloud - DEEP-HDC is lasting for 30 months and is combining the expertise of 10 large European research organisations. The project proposes to deploy under the common label of "DEEP as a Service" a set of building blocks that enable the easy development of applications requiring these techniques: deep learning using neural networks, parallel post-processing of very large data, and analysis of massive online data streams.

1. Introduction

Machine learning 'as-a-service' can clearly be seen as one of the main user requirements being asked for when thinking about using large-scale computing infrastructure. With ever more data being available, the wish to exploit this data is omnipresent. While machine learning pipelines for small-scale data sets are available in the form of several software libraries, large-scale learning tasks provide another level of challenge. With the need for a proper design of the learning task at hand, additional tasks result in the need to organize large-scale data, the provision of necessary computing power and storage capacity and, since large-scale data is commonly distributed, the orchestration of various infrastructure components at different places. It is obvious that such learning tasks cannot be managed by a user with domain knowledge in the field of application, only. Therefore, support by the infrastructure layer must break down the complexity of the task and allow the user to focus on what she/he is skilled on, i.e., modelling of the problem, evaluating and interpreting the results of the machine learning algorithms. As a consequence, infrastructure providers have to understand the needs of their user communities and help them to combine their services in a way that encapsulates technical details the end user does not have to deal with. The goal of the DEEP project is to develop such applications along the lines of selected Use Cases. Indeed, this is the goal of the DEEP-HDC project [1] aimed at deploying under the common label of "DEEP as a Service" a set of building blocks that enable the easy development of applications requiring these techniques: deep learning using neural networks, parallel post-processing of very large data, and analysis of massive online data streams.

The targeted platforms for the released products are the already existing and the next generation e-Infrastructures deployed in Europe, such as the European Open Science Cloud

(EOSC) [2], the European Grid Infrastructure (EGI) [3] and the computing infrastructures that will be funded by the upcoming H2020 EINFRA-12 call. DEEP-HDC is funded by the H2020 EINFRA-21-2017 Research and Innovation action under the topic Platform-driven e-Infrastructure innovation [4]. It is carried on by a Consortium that brings together technology providers with a proven long-standing experience in software development and large research communities belonging to diverse disciplines: Biological and Medical Science, Computing Security, Physical Sciences, Citizen Science and Earth Observation.

DEEP-HDC started on 1st November 2017 and will run for 30 months until April 2020. The EU contribution for the project is 2.98 million euros.

2. Project Objectives

The DEEP-Hybrid-DataCloud project started with the global objective of promoting the usage of intensive computing services by different research communities and areas, and their support by the corresponding e-Infrastructure providers and open source projects. Other objectives followed by the project are:

- Focus on intensive computing techniques for the analysis of very large datasets considering highly demanding use cases.
- Evolve up to production level, intensive computing services exploiting specialized hardware.
- Integrate intensive computing services under a hybrid cloud approach.
- Define a "DEEP as a Service" solution to offer an adequate integration path to developers of final applications.
- Analyse the complementarity with other ongoing projects targeting added value services for the cloud.

The DEEP-Hybrid-DataCloud project aims to provide a bridge towards a more flexible exploitation of intensive computing resources by the research community, enabling access to the latest technologies that require also last generation hardware and the scalability to be able to explore large datasets. It is structured into six different work packages, covering Networking Activities (NA) devoted to the coordination, communication and community liaison; Service Activities (SA) focused on the provisioning of services and resources for the execution of the data analysis challenges; and Joint Research Activities (JRAs), dealing with the development of new components and technologies to support data analysis.

In order to achieve these objectives, we propose to evolve existing cloud services, taking into account the following design principles:

- Evolve the required services from TRL6 to TRL8 under an open framework and considering existing standards for interoperability.
- Re-use if possible existing cloud services in production, and in particular those being adopted for proposed e-infrastructure of the European Open Science Cloud.
- Consider the integration of existing specialized resources into cloud services having in mind the point of view of the current daily management of those resources, like for example current HPC data centres.
- Ensure that the resulting framework will have a low learning curve for the developers of the solutions, by delivering a DEEP catalogue that can be directly exploited by users to build their applications.
- Assure the scalability and performance of the solution developed, which is key to guarantee the interest both of resource providers and users.

3. Research Communities and requirements

The Research Communities participating in the DEEP-HDC project enable to cover differet scientific areas ranging from Biological and Medical Science, Computing Security, Physical Sciences, Citizen Science and Earth Observation.

3.1. Medical science

Deep learning approaches to biomedical image analysis have opened new opportunities in how diseases are diagnosed and treated. However, one drawback of automated analysis of medical images with deep learning is the requirement for sophisticated IT infrastructures (hardware, software, network). In this project, we will explore various ways to apply deep learning to analyse images in the context of retinopathy [5]. Requirements injected into the project:

- develop and evaluate a deep learning tool facilitating the classification of retinopathy stage and progression based on digital color fundus retinal photography images.
- improve automated classification retinopathy stage (Healthy, Mild, Medium, Severe) and reconstruct disease progression by means of deep learning.
- explore construction (training) of deep learning models using inherently distributed training data.
- address the need for a comprehensive and automated method for large-scale screening programs based on medical images.

3.2. Computing Security

The usage of specialised hardware in order to speed up packet capturing, pre-processing and classification for network analysis is becoming very popular in recent times. Intrusion detection, or deep packet inspection systems are highly demanded application frameworks by network security analysts. The common feature to these network applications is that it is needed to process a continuous flow of information and react promptly to the generated events. Requirements injected into the project:

- Definition of an architecture for data intake, analysis and storage.
- Research and development of different tools e.g. monitoring tools, decision support modules, prediction module with cloud supports that allow data analysis.
- Research and development of intelligent modules using ML/DL to analyse and to get meaningful insights of massive online data. Applying and testing different approaches toward cyber-security aspects focused on event detection.
- Accelerate Use Case development using cloud e-infrastructure modern technology advantages such as isolated independent environment, that provide built-in security, portability, and flexibility features.

3.3. Physical Sciences

Quantum Chromodynamics (QCD) is the theory that describes the interaction responsible for the confinement of quarks inside hadrons, the so-called strong interaction. Investigating the properties of QCD requires different techniques depending on the scale of energy we are interested in. In this respect, data analysis presents technical challenges to the researchers involved. Those challenges can be traced down to the lack of a flexible environment to move data across the network and analyse them in a flexible way to improve the efficiency of the process of configuration analysis. Requirements injected into the project:

- designing a data configuration tool to have general applicability for similar usage scenarios in other scientific areas.

3.4. Citizen Science

The potential of applying deep learning techniques for plant classification and its usage for citizen science in large-scale biodiversity monitoring has been discussed in recent publications [6]. The predictions can be confidently used as a baseline classification in citizen science communities which in turn can share their data with biodiversity portals. Requirements injected into the project:

- Produce a tool that is able to classify plants species from images.
- Have the results produced by the developed tools validated by biodiversity experts.
- Deploy this tool to automated monitoring of biodiversity.

3.5. Earth Observation

The application of NN to pattern recognition in satellite images, and their combination with other in-situ measurements, opens new possibilities in areas like Ecosystems and Biodiversity. The EU Copernicus Programme relies on a family of dedicated Earth Observation missions called the Sentinels. The data acquired from these missions are systematically downlinked and processed to operational user products, and made available under an open and free license [7].

- Enable monitors for ecosystems and the design of better policies regarding the environment.
- Enable the capabilities of the European biodiversity platforms such as LifeWatch.
- Develop a detection and prediction system that combines the latest Deep Learning techniques with satellite data. An environment where you can process and analyse different satellite maps, choosing the place, the date, etc.

4. DEEP Overall architecture

The DEEP PaaS layer is based on the components developed and integrated in the INDIGO-DataCloud project [8]. The architecture is depicted in Figure 1 and the main components are briefly described hereafter:

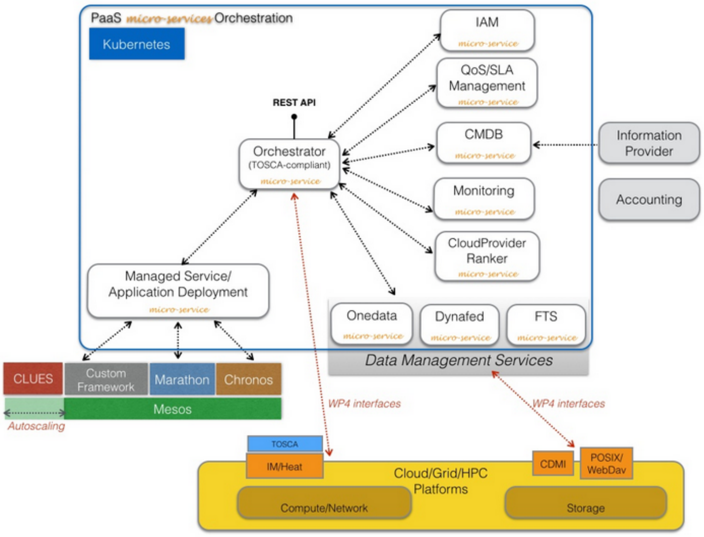

Figure 1. The architecture of the DEEP PaaS layer is based on the building blocks provided by the INDIGO-DataCloud project.

The PaaS Orchestrator is the core component of the PaaS layer. It receives high-level deployment requests and coordinates the deployment process over the IaaS platforms.

The Identity and Access Management (IAM) Service provides a layer where identities, enrolment, group membership, attributes and policies to access distributed resources and services can be managed in an homogeneous and interoperable way.

The Monitoring Service is in charge of collecting monitoring data from the targeted clouds, analysing and transforming them into information to be consumed by the Orchestrator.

The Cloud Provider Ranker (CPR) is a rule-based engine that allows to rank cloud providers in order to help the Orchestrator to select the best one for the requested deployment. The ranking algorithm can take into account preferences specified by the user and other information like SLAs and monitoring data.

The SLA Management (SLAM) Service allows the handshake between users and a site on a given SLA.

The Managed Service/Application (MSA) Deployment Service is in charge of scheduling, spawning, executing and monitoring applications and services on a distributed infrastructure; the core of this component consists of an elastic Mesos cluster with slave nodes dynamically provisioned and distributed on the IaaS sites.

The Infrastructure Manager (IM) deploys complex and customized virtual infrastructures on a IaaS site providing an abstraction layer to define and provision resources in different clouds and virtualization platforms.

The Data Management Services is a collection of services that provide an abstraction layer for accessing the data storage in a unified and federated way.

The Information Provider and Accounting System collects detailed information from an IaaS provider about the current status of the resources from the amount of resources of CPU, RAM or storage to the availability of a service.

5. DEEP as a Service solution

The high level decomposition of the DEEP as a Service design is depicted in Figure 2 and consists on the following key components:

- The DEEP open Catalog where the users, communities, etc. can browse, store and download relevant modules for building up their applications (like ready to use machine learning frameworks, complex application topologies, etc.).
- An application modeler or composition tool, that will be used to build up complex application topologies in an easy way.
- A runtime engine, that will take the defined topology as input, provision the required computing resources and deploy the application.
- The DEEP PaaS layer, that will coordinate the overall workflow execution to choose the appropriate Cloud sites and manage the deployment of the applications to be executed.
- The DEEP as a Service solution, that will offer the application functionality to the user.
- The EINFRA/EOSC data services, to be integrated with the DEEP solutions in order to provide access to any of the data facilities existing in the European Open Science Cloud

The system is designed with extensibility in mind, taking great care in designing a framework which can be updated easily and where a component can be replaced with a new one in case it is needed. Many of the anticipated changes to our system in future phases will only require adding additional functionality on top of existing components, remaining backwards compatible with previous versions.

6. Conclusions

In the present contribution the DEEP-XDC project and its objectives have been presented and discussed. These objective, together with the related needs proper of the research communities

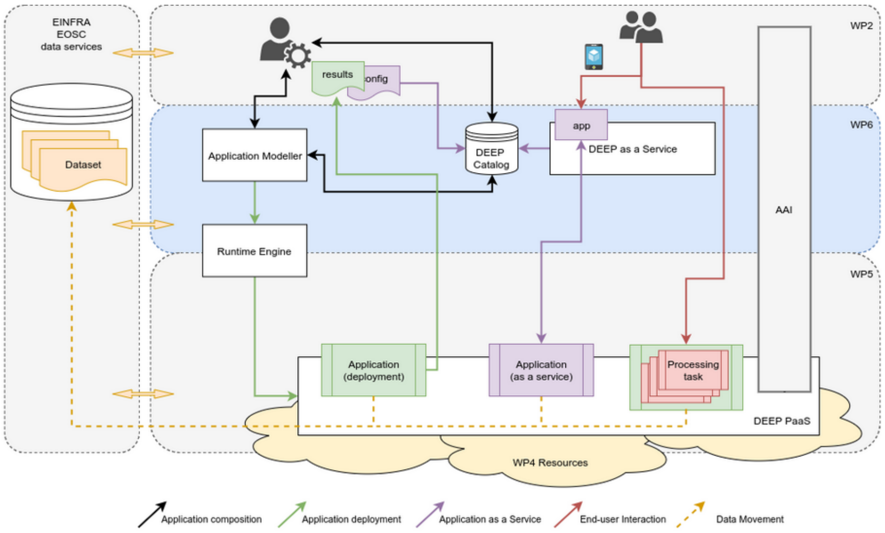

Figure 2. DEEPaaS high level architecture.

involved in the project, are the real driver to develop innovative and reliable open source solutions able to fill up the technology gaps that currently prevent effective exploitation of distributed computing and storage resources by many scientific communities.

Moreover, DEEP-HDC project can complement and integrate with other running projects and communities and with existing multi-national, multi-community infrastructures. As an example,DEEP-HDC is collaborating with the eXtreme-DataCloud (XDC) [9] project aimed at developing scalable technologies for federating storage resources and managing data in highly distributed computing environments.

As an added value both projects (DEEP-HDC and XDC) have the common objective to open new possibilities to scientific research communities in Europe by supporting the evolution of e-Infrastructure services for exascale computing. Those services are expected to become a reliable part of the final solutions for the research communities available in the European Open Science Cloud Service Catalogue.

7. References

[1] Web site: www.deep-hybrid-datacloud.eu
[2] Web site: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[3] Web site: https://www.egi.eu/
[4] Web site: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-21-2017.html
[5] Yau, J.W., Rogers, S.L., Kawasaki, R., Lamoureux, E.L., Kowalski, J.W., et al. (2012) Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 35: 556–564.
[6] Ignacio Heredia, 2017. Large-Scale Plant Classification with Deep Neural Networks. Proceedings of the Computing Frontiers Conference , 259-262.
[7] http://www.copernicus.e u/main/data-access
[8] Web site: https://www.indigo-datacloud.eu
[9] Web site: www.extreme-datacloud.eu