Big Data Analytics on Large-Scale Scientific Datasets in the INDIGO-DataCloud Project

Sandro Fiore
Euro-Mediterranean Center on
Climate Change Foundation
Lecce, Italy

Cosimo Palazzo
Euro-Mediterranean Center on
Climate Change Foundation
Lecce, Italy

Alessandro D'Anca
Euro-Mediterranean Center on
Climate Change Foundation
Lecce, Italy

Donatello Elia
Euro-Mediterranean Center on
Climate Change Foundation
Lecce, Italy

Elisa Londero
INAF - Trieste Astronomical
Observatory (OATs)
Trieste, Italy

Cristina Knapic
INAF - Trieste Astronomical
Observatory (OATs)
Trieste, Italy

Stephen Monna
Istituto Nazionale di Geofisica e
Vulcanologia (INGV)
Rome, Italy

Nicola M. Marcucci
Istituto Nazionale di Geofisica e
Vulcanologia (INGV)
Rome, Italy

Fernando Aguilar
Instituto de Física de Cantabria
(UC-CSIC)
Santander, Spain

Marcin Płóciennik
PSNC - Poznan Supercomputing and
Networking Center, IBCh PAS
Poznan, Poland

Jesús E. Marco De Lucas
Instituto de Física de Cantabria
(UC-CSIC)
Santander, Spain

Giovanni Aloisio
Euro-Mediterranean Center on
Climate Change Foundation
Università del Salento
Lecce, Italy

ABSTRACT

In the context of the EU H2020 INDIGO-DataCloud project several use case on large scale scientific data analysis regarding different research communities have been implemented. All of them require the availability of large amount of data related to either output of simulations or observed data from sensors and need scientific (big) data solutions to run data analysis experiments. More specifically, the paper presents the case studies related to the following research communities: (i) the European Multidisciplinary Seafloor and water column Observatory (INGV-EMSO), (ii) the Large Binocular Telescope, (iii) LifeWatch, and (iv) the European Network for Earth System Modelling (ENES).

CCS CONCEPTS

• **Applied computing** → *Environmental sciences*; *Astronomy*; *Physics*; *Computational biology*; • **Computer systems organization** → **Distributed architectures**; • **Information systems** → *Data management systems*; • **Computing methodologies** → *Distributed computing methodologies*;

KEYWORDS

Workflow, big data, scientific use case, ensemble analysis

ACM Reference format:
Sandro Fiore, Cosimo Palazzo, Alessandro D'Anca, Donatello Elia, Elisa Londero, Cristina Knapic, Stephen Monna, Nicola M. Marcucci, Fernando Aguilar, Marcin Płóciennik, Jesús E. Marco De Lucas, and Giovanni Aloisio. 2017. Big Data Analytics on Large-Scale Scientific Datasets in the INDIGO-DataCloud Project. In *Proceedings of CF'17, Siena, Italy, May 15-17, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3075564.3078884

1 INTRODUCTION

Data-driven scientific discovery is a key paradigm driving research innovation in multiple scientific domains such as astronomy, geophysics, biology and climate change.

Volume, complexity, and variety of data produced in these contexts need specific solutions able to manage large datasets and to take advantage of parallel processing so as to deliver results in (near) real-time.

In the (scientific) big data landscape, the Ophidia framework (an array-database solution for eScience), exploits parallel computing techniques, in-memory storage and smart data distribution methods to enhance scalability of traditional OLAP systems. In particular, an array-based storage model and a hierarchical data organization are adopted to distribute scientific datasets over multiple nodes and, hence, to enable efficient parallel data processing. In addition, an embedded analytics workflow manager supports the execution of more tasks concurrently, so that compute resource utilization is further improved. The workflow manager makes more flexible the

CF'17, Siena, Italy
© 2017 ACM. 978-1-4503-4487-6/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3075564.3078884