The NoMaD Laboratory and Big-Data Analytics: Extracting hidden information from repositories of computational materials science

Fawzi Mohamed, Luca M. Ghiringhelli, Chrisitian Carbogno, Claudia Draxl, Alessandro De Vita, Daan Frenkel, Francesc Illas, Risto Nieminen, Angel Rubio, Kristian Sommer Thygesen, and *Matthias Scheffler*(*)

Initiatives like the NoMaD Repository [1] give access to the raw data of computational material-science studies performed with a variety of codes. To take advantage of the wealth of information hidden in this very heterogeneous, open access data, the base layer of the NoMaD Lab [2] − an infrastructure to perform advanced big-data analytics and complex queries over this kind of data − is presented.

First a *translation layer* transforms the data into a standardized format that uses an hdf5 or a json file. This file uses a flexible classification system that can be easily extended, to describe the data stored. As a consequence, all data are stored in a uniform, robust and extensible representation.

Then the reactive (responsive, resilient, message-driven and dynamically resizable) application that started the standardization might complete the data calculating derived quantities. Finally Apache Flink [3] (a fast large-scale data-processing engine) is used to efficiently perform complex queries on the extracted data.

Illustrative examples of the queries and data analytics enabled by the NoMaD Lab, and of its scalability, are demonstrated.

(*) Work done in collaboration with the NoMaD team [2]

[1] http://nomad-repository.eu/ [2] http://nomad-lab.eu/

[3] http://flink.apache.org/