Towards the Big Data Strategies for EISCAT-3D

Yin Chen[(1)], Ingemar Häggström [(2)], Alex Hardisty[(1)], Gergely Sipos[(3)], Małgorzata Krakowian [(3)], Nuno L. Ferreira[(3)], Ville Savolainen[(4)]

(1) Cardiff University, 5 The Parade, Roath, Cardiff, UK, {Y. Chen, Alex.Hardisty}@cs.cardiff.ac. uk

(2) EISCAT, Box 812, SE-981 28 Kiruna, Sweden, Ingemar.Haggstrom@eiscat.se

(3) EGI.eu, Science Park 140, Amsterdam, the Netherlands,{gergely.sipos, malgorzata.krakowian, nuno.ferreira}@egi.eu

(4) CSC-IT, P. O. Box 405, FIN-02101 Espoo, Finland, ville.savolainen@csc.fi

The design of the next generation incoherent scatter radar system, EISCAT-3D, opens up opportunities for physicists to explore many new research fields in the studies of the atmosphere and near-Earth space. On the other hand, it also introduces significant challenges in handling the large-scale experimental data which will be massively generated at great speeds and volumes. During its first operation stage in 2018, EISCAT-3D will produce 5PB data per year, and the total data volume will rise up to 40PB per year in its full operations stage in 2023. This refers to the so-called big data problem, whose size is beyond the capabilities of the current database technology [1]. To unlock the value from these data, new forms of processing and platforms of tools are needed.

Advanced e-Science infrastructures such as, EGI, EUDAT, and PRACE, and their enabling technologies are making large-scale computational capacities more accessible to researchers of all scientific disciplines. The European Grid Infrastructure (EGI), a not-for-profit foundation created to manage the infrastructures on behalf of the National Grid Initiatives and European Intergovernmental Research Organisations, operates more than 370,000 logical CPUs, 248 PB disk and 176 PB of disk capacity (June 2013 statistics). EUDAT is a European project aiming to take the first steps towards building a Collaborative Data Infrastructure for European scientific data products. It will offer services for data storage and replication, data staging to computational resources (and vice versa) and services for data cataloguing and discovery. PRACE is the pan-European supercomputing infrastructure that forms the top-tier of HPC provision across Europe, with the aim of enabling high impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness.

We propose e-Science approaches to tackle the challenges of processing and searching EISCAT-3D data, and will provide solutions for:

1. Staging EISCAT-3D lower-level data (voltage data) into the large-scale e-Science storage, such as EGI or EUDAT;
2. Providing various processing and mining facilities such as, auto-correlation and spatial/temporal integration, to allow individual scientists to analyse data as their will;
3. Providing advanced searching facilities to enable individual scientists to search through all level of data and identify specific signatures, e.g., plasma features, meteors, space debris, astronomical features.

The new data processing and searching strategy will offer more flexible way for EISCAT users to analyse and discover interesting data patterns which are not yet available. Space physicists will be able to make better use of the observation data and exploit the growing wealth of them. This will eventually lead to a new data-centric way of conceptualising, organising and carrying out research activities which could lead to an introduction of new approaches to solve problems that were previously considered extremely hard or, in some cases, impossible to solve and also lead to serendipitous discoveries and significant breakthrough [1].

[1] C. Thanos, S. Manegold and M. Kersten, “Big Data”, *ERCIM Special Theme: Big Data*, No. 89, Apr. 2012.

Relevant conference session: Data Management