ID 213

Tipo de Comunicación: Poster

Sesión Científica: Instrumentacion y sipercomputacion

Titulo: Data mining and data analytics with the Gaia archive

Nombre (Autor que presenta): Francesc

Apellidos (Autor que presenta): Julbe Lopez

Apellidos y nombre de los autores: Sarro, Luis; Luri, Xavier; Tapiador, Daniel on behalf of Gaia-DPAC CU9-WP973

Resumen:

The Gaia Data Processing and Analysis Consortium (DPAC) Coordination Unit 9 is the last DPAC coordination unit, in charge of the design and implementation of the Gaia archive. We report on recent advances in the Data Mining Work Package of this Coordination Unit 9. The main goal of this work package is to make available to the community of Gaia archive users a platform that provides and enables the Knowledge Discovery and data analytics tools needed to fully exploit datasets of sizes intractable with traditional approaches. Initial developments have been funded by the GENIUS FP7 project and during its first two years of activity, we have undertaken a broad study and evaluation of Big Data technologies applicable Gaia scientific use cases. The simplest use cases have been implemented and tested as part of the platform validation. This so-called Gaia Data Analytics Framework (GDAF) deploys a set of Big Data services based on the Hadoop and Apache Spark framework for state-of-the-art distributed processing. GDAF should be in production for the third Gaia Data Release. As a more ambitious test case, we simulate large numbers of Gaia archive stellar entries and estimate the IMF (Initial Mass Function) and Star Formation Rate (SFR) using the GDAF prototype. The inference system is based on hierarchical Bayesian models and MCMC posterior probability samplers of both parametric and non-parametric (Gaussian Processes based) models. The simulations lack important details of the Gaia selection biases, but they serve to gauge the processing capabilities of the infrastructure.