GENIUS
Data Mining Work Package
WP-430

Leiden, November 20th 2015

Overview:

- GAIA will implement an advanced data analytics framework that will allow performing complex queries to the archive.

- The complexity of the queries and the size of the archive are the main drivers to approach those advanced queries using Big Data technologies.

- This framework has been called **GDAF** (Gaia Data Analytics Framework)

GENIUS, year 2 (I) - STATUS

Building on the learning process undertaken during the first GENIUS year:

- Work Package team configured (16-17 members)
- Requirements and features defined (*although given its early stages, they may change*)
- Framework has been completely designed at following levels:
  - Hardware infrastructure
  - Big Data framework (Spark & Hadoop)
  - Services configured for development
- First scientific use cases identified
- Grand challenge progress

GENIUS, year 2 (II) - STATUS

System description:

- Hardware infrastructure – Provided by CSUC (6 nodes, 16 cores each), 384GB RAM and about 70TB of storage capacity

- Big Data framework, 3 main components already configured:

  HDFS: Distributed file system where the Gaia will be stored for Data analytics

  YARN: Job scheduling and resource manager

  Spark: Data Analytics distributed engine

GENIUS, year 2 (III) - STATUS

Grand challenges:

Astronomical problems that require high level of computation, data processing and that will provide great scientific value thanks to the quality of Gaia data and the features provided by GDAF.

Global comparison of existing galactic models with real Gaia Data.

Tasks identification and status:

- Scientific problem identification: what is the shape for the initial mass function (IMF) and star formation rate (SFR) taking into account millions of observations in the GAIA catalog.
- Establish a joint probability distribution for IMF and SFR parameters.
- Due to the complexity of the models it'll be necessary to use Markov chain Monte Carlo (MCMC) methods in order to obtain an i.d.d. sample from the joint posterior probability distribution of the parameters to be estimated.
- Status:
  - The estimation of IMF parameters is done
  - The estimation of the SFR is in progress
  - Joint estimation of IMF and SFR is in progress

GENIUS, year 2 (IV) - GOALS

Goals for next year:

- Consolidate current platform configuration

Interaction with other WPs:

- Architecture, providing a frame to the platform, identifying interactions with other components and subsystems and merging it to the global archive architecture.
- Visualization features and capabilities. Which features can offer current visualization infrastructure → WP980 integration

GAIA Archive for GDAF:

- A prototype to convert GAIA data into Hadoop compatible format has been tested both with GACS data and TGAS simulated data. To be evaluated with ESDC and define a ‘formal’ procedure.

GENIUS, year 2 (V) - GOALS

Use cases: Definition of ‘simple’ use cases and recipes to build a ‘library’ set to build more elaborated use cases and pipelines

- Clustering on TGAS: Clustering in the space of spatial positions and proper motions. We should rediscover proper motion groups and maybe even find new groups.
- Cross match between TGAS and other photometric surveys and do a search for exotic objects. For example, sub dwarfs or white dwarfs. This might be through a classifier, or again with a mixture of Gaussians

Grand Challenge:

- Continue current work on Grand Challenge
- Identify more grand challenges, Gaia community?

Milestones

- Define a GDAF platform as a deliverable of GENIUS project despite being 1 year ahead of its official release date (Gaia-DR3)

GENIUS, year 2 (VI) – FINAL CONSIDERATIONS & COMMENTS

- GDAF can be deployed in other data centers next to their ‘copy’ of the Gaia archive.

- Documentation where all relevant GDAF information will be available (how to use it –submit jobs, monitor…-, use cases and recipes, etc.)

- Current GDAF Information source:

  http://wiki.cosmos.esa.int/gaia-dpac/index.php/CU9:970:973

  https://gaia.am.ub.es/Twiki/bin/view/GENIUS/DataMiningGENIUSPage (update pending)

  http://gaia.esac.esa.int/dpacsvn/DPAC/CU9/docs/WP970_Science_Enabling_Apps/nonECSS/TechNotes/GAIA-C9-SP-UB-FJL-003/GAIA-C9-SP-UB-FJL-001.pdf

- H2020 project call for 2017 (COMPET-4-2017: Scientific data exploitation), challenge: *Support the data exploitation of European missions and instruments, in conjunction, when relevant, with international missions.*