GDAF – Gaia Data Analytics Platform

CU9 work Package: 970 Science enabling Applications → 973 Data Mining

Currently being partially funded and developed under the frame of the GENIUS project (FP7).

Goal: Design, develop and deploy a data mining platform for scientific exploitation of the Gaia archive.

- Requirements:
  - Gaia data access scenarios summary - GAIA-C9-TN-LEI-AB-026-1
  - WP970 SRS

- WP with members of several institutions. Spread effort.
- CSUC (Consorci de Serveis Universitaris de Catalunya)

GDAF – Gaia Data Analytics Platform (I)

HARDWARE

- Global features:
  - 6 nodes
  - 96 cores
  - 4 TFLOPs
  - 384 GB RAM
  - 72 TB disc

6 servers RSTORAGE 12D+ E5V3, each of one with:

2 x Intel Xeon™ E5-2640v3 8 Core 2,6GHz, 22nm, 20MB, 90W
8 x 8GB DDR4 2133MHz ECC REG
2 x SSD Toshiba 128GB 19nm PCIe 6Gb/s MLC 7mm 19nm NAND Flash Memory Multi-Level Cell. 510/460MB/s. R/W
12 x HD 1TB, SATA 6 Gb/s 7.200rpm, 3,5 64MB Nearline Enterprise Storage
1 x Ethernet PCI-E x4 Gigabit Dual Port RJ45
1 x Asus® 10GbE SFP+ Dual Port LC PCI-E 3.0 x8

SOFTWARE FRAMEWORK

CentOS 6 with:

- Cloudera 5.4.4 - http://www.cloudera.com/ (Hadoop Distribution)
- Spark 2.0.1 - http://spark.apache.org/

GDAF – Gaia Data Analytics Platform (II)

Scientific use cases:

- Based on recipes, using either Gaia DR1, GOG simulated data, TGAS, User specific data...
- User: Astronomers with scientific needs and requirements.

Functionality:

- Usage based on recipes, code reuse.

Validation use cases:

- Exploration of validation tests for the Validation WP using Spark & ecosystem on DR1.
- Proof of concept. Several minutes x validation test.

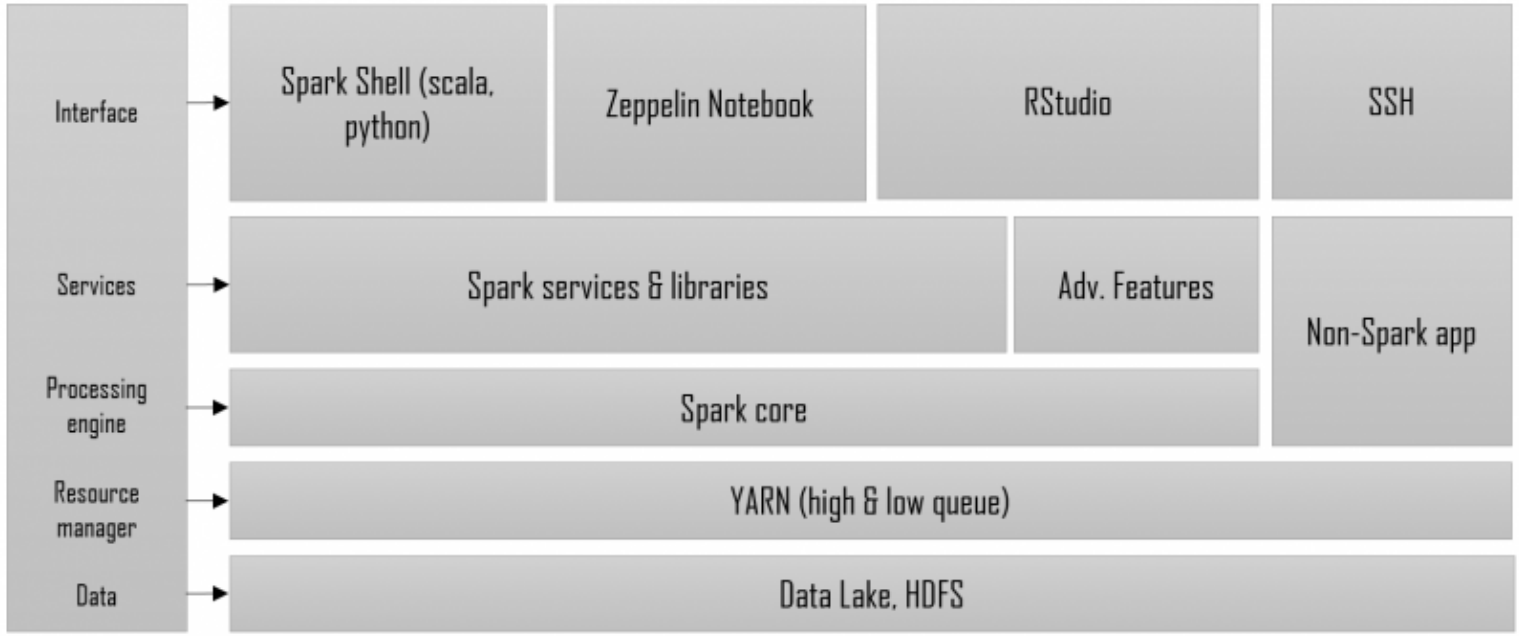
GDAF – Gaia Data Analytics Platform (III)

CORE SERVICES

- HDFS HADOOP CORE
- YARN HUE

BIG DATA SERVICES & INTERFACES

- SPARK ZEPPELIN NOTEBOOK



GDAF – Gaia Data Analytics Platform (IV)

DATA – STORAGE

CSV files to an ASCII ‘formatted’ format and Hadoop ‘friendly’ format ‘Parquet’

```
|-ascii
|---Hipparcos-2
|-----Hipparcos2Header
|-----data
|---Tycho-2
|-----regular
|-----supplementary
|-----tycho2RegularHeader
|-----tycho2SupplHeader1
|-----tycho2SupplHeader2
|---external
|---gaia
|-----gdr1
|-------gaiaSource
|-------gaiaSourceHeader
|-------tgas
|-------tgasHeader
|-----tgas-sim
|-----tgasheader
|-parquet
|---Hipparcos-2
|---Tycho-2
|---external
|---gaia
|-----gdr1
|-------GaiaSourceHeaderSchema
|-------gaiaSource
|-------gaiaSourceFormatted
|-----tgas-sim
|-----um
|-tmp
|-validation
```

Data conversion procedure:

- ASCII to Parquet, standard procedure for any ASCII conversion *
- Parquet allows fast SQL querying to the archive
- Scala code
- ~ 60min for full sky conversion

http://gaia.esac.esa.int/dpacsvn/DPAC/CU9/software/SciEnablingApps/WP973/Tools/GacsAsciiToParquetConverter/

*What about Gbin?

GDAF – Gaia Data Analytics Platform (V)

Spark ‘features’

- *Transparent* parallel processing
- Multiple features: SQL Spark, Streaming, MLlib, Graph Processing
- Pipelines definition

Interfaces:

- SSH, scripts: spark-shell, spark-submit
- Zeppelin - Notebook

YARN – cluster manager

Virtual Machine

-> Goal is providing the community with a small test environment which replicates the GDAF platform.

TGAS

Subset or downsized GAIA

Other catalogues

‘Vbox image’

GDAF – Gaia Data Analytics Platform (VI)

Pending issues:

Will the platform be integrated at ESAC?

- Genius context, goal meet in terms of platform definition. Next phase needed (GENIUS ending)
- Efforts needed on concurrent scientific use cases.

Calendar. DR3 official date is ½ half of 2018? Early 2019?

- Cross-mission development, not only for Gaia.

Responsibilities?