¿Que entendemos por un "Gaia Grand Challenge?

Plataforma de minería de datos en GAIA

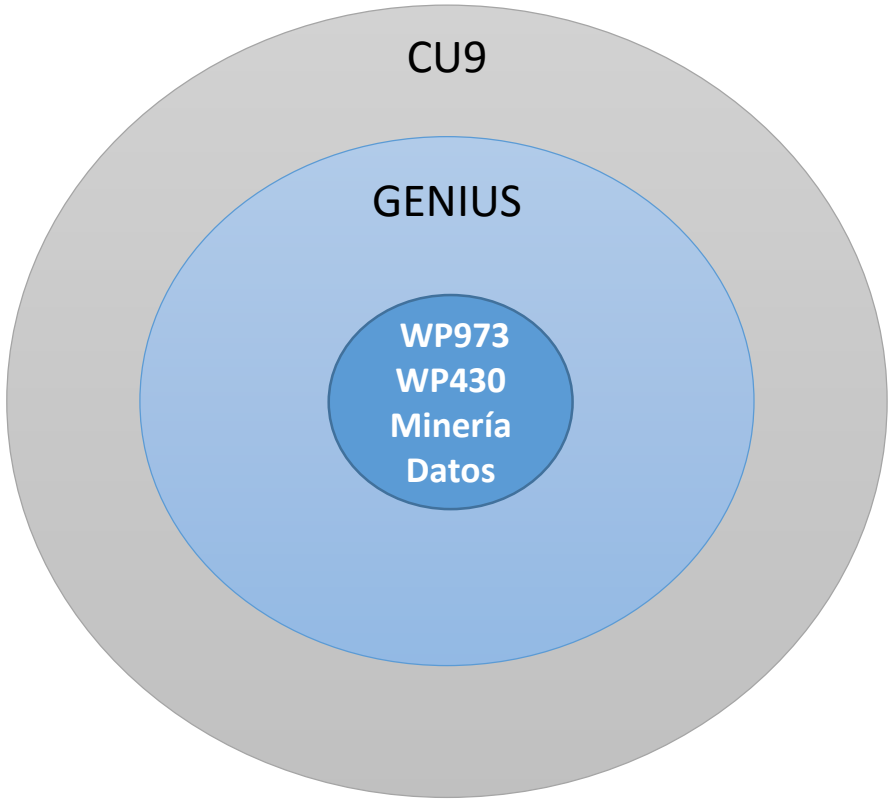
Francesc Julbe Lopez ICCUB
Barcelona 24-Mayo-2016

En el marco de trabajo de la CU9, se desarrolla una infraestructura para realizar Minería de datos:

Objetivo:
Proveer una infraestructura que ofrezca al astrónomo herramientas para realizar tareas de minería de datos:

- Análisis estadístico
- Machine Learning, etc.

… extracción de conocimiento



GENIUS, proyecto FP7 (FP7-SPACE-2013-1)

Minería de datos es un 'Sub Work-package' de:

- "Science Enabling Applications" en CU9
  ¡
- "Tools for data exploration" en GENIUS

¿Cuando? DR3 ~**finales 2018**

Proyecto GENIUS: Componente tecnológica y científica
3 años:

Componente Tecnológica:

Tamaño previsto del catalogo Gaia en DR3?

- DR1 ~ Formato Hadoop: **3.3 T**
- DR2 ~ ?
- DR3 ~ **20TB**

Tendencias y tecnologías Big Data?
Fase de aprendizaje:

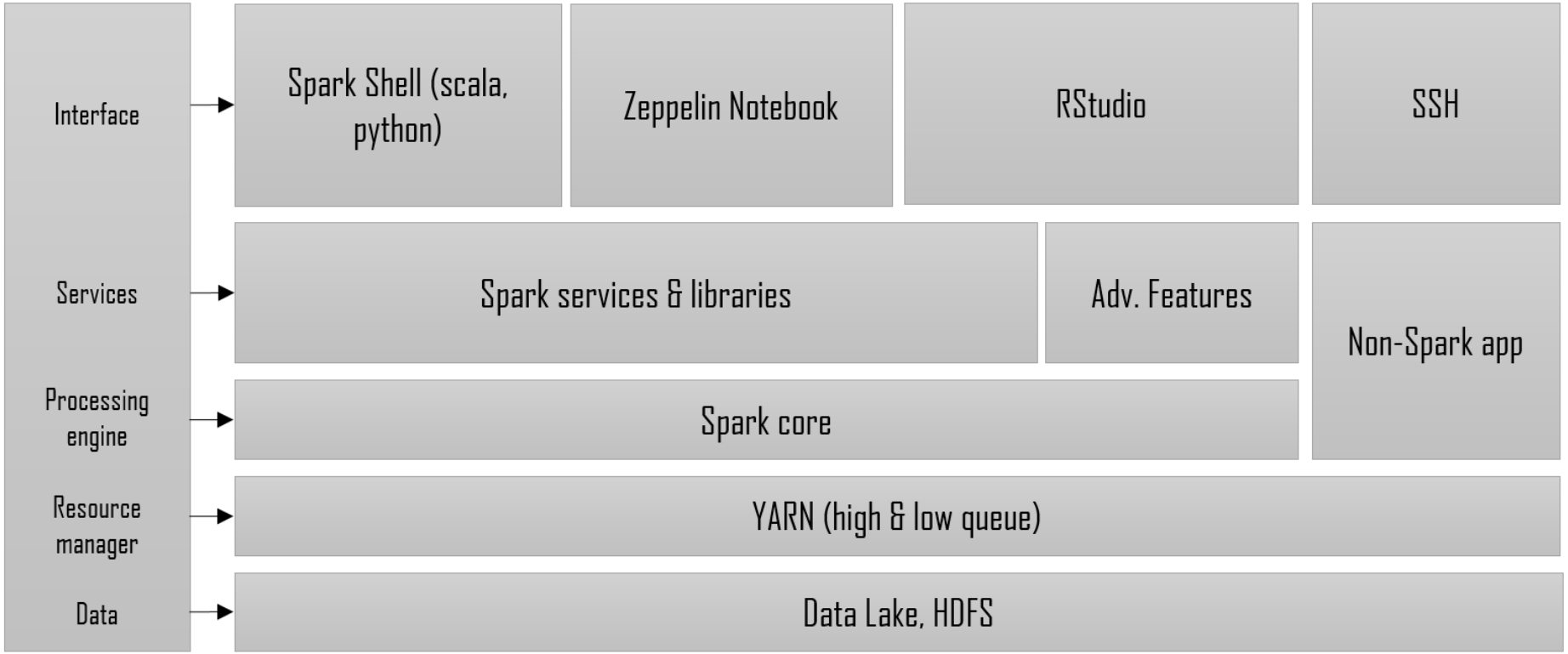
Componente científica:

Primera definición de los llamados “Grand Challenges”
Breves use cases de análisis de tecnologías:

- PCA con diferentes herramientas sobre espectros simulados con GOG (200k y 2M)

Características: Altamente escalable, facilidad de desplegar y de utilizar

Componente Tecnológica: Definición de la plataforma, componentes, etc.



Cluster financiado en parte por GENIUS y por CSUC.

- Plataforma: GDAF (Gaia Data Analytics Framework)
  - Definición de la estructura del archivo en entorno distribuido (tipo de ficheros, etc) y método de conversión a formato ‘Hadoop’.
  - Configuración de colas (YARN)
  - Herramientas de acceso: Notebook Zeppelin, Spark Shell, RStudio.

Objetivo hasta finalizar GENIUS (Primavera 2017): Desarrollo de más casos de uso sobre esta infraestructura con el fin de validar su ‘base’, usabilidad, etc.

Casos de uso ‘simples’ y casos de uso ambiciosos → Grand Challenges

Casos de uso científicos:

- Casos de uso ‘simples’

- *‘Librería’ de casos de uso*
- *Documentación*

Herramientas ‘out-of-the-box’ (Spark Mllib) y otras más especificas como SOM (Self Organizing Maps) y HMAC (Hierarchical Modal Association Clustering)

- Grand Challenges

- *Caso de uso ambicioso que mediante el análisis del catálogo Gaia y herramientas de HPC y Big Data van a suponer un avance en el desarrollo de alguna teoría científica.*

Objetivo:
Máxima reusabilidad de librerías, pipelines…’learning by example’

Grand Challenges

Global comparison of a galaxy model (e.g. the one used for the catalogue validation tasks described in WP940) with the actual Gaia data.

(1) 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? Uso Modelos Jerárquicos Bayesianos

Código 'reutilizable', todavía un prototipo muy optimizable (2 apps, 'IMF, 'SFR'), implementación de MCMC.
Conclusiones técnicas: Buena escalabilidad. En un cluster con 300 cores (unos 20 nodos), tardaría 50 días.

(2) Segunda aproximación → Presentación de Roger

Otras propuestas, no elaboradas

- Selection and exploration in the Fourier domain for variability analysis.
- Cross-correlate spatial position and variability information, to explore the structure of our galaxy using variable stars.
- Query per-CCD photometry to explore short-term variability