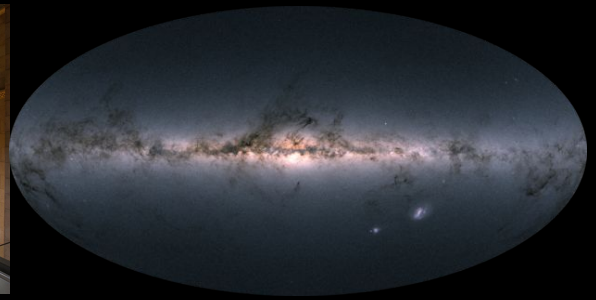
Widening Big data mining for astronomy

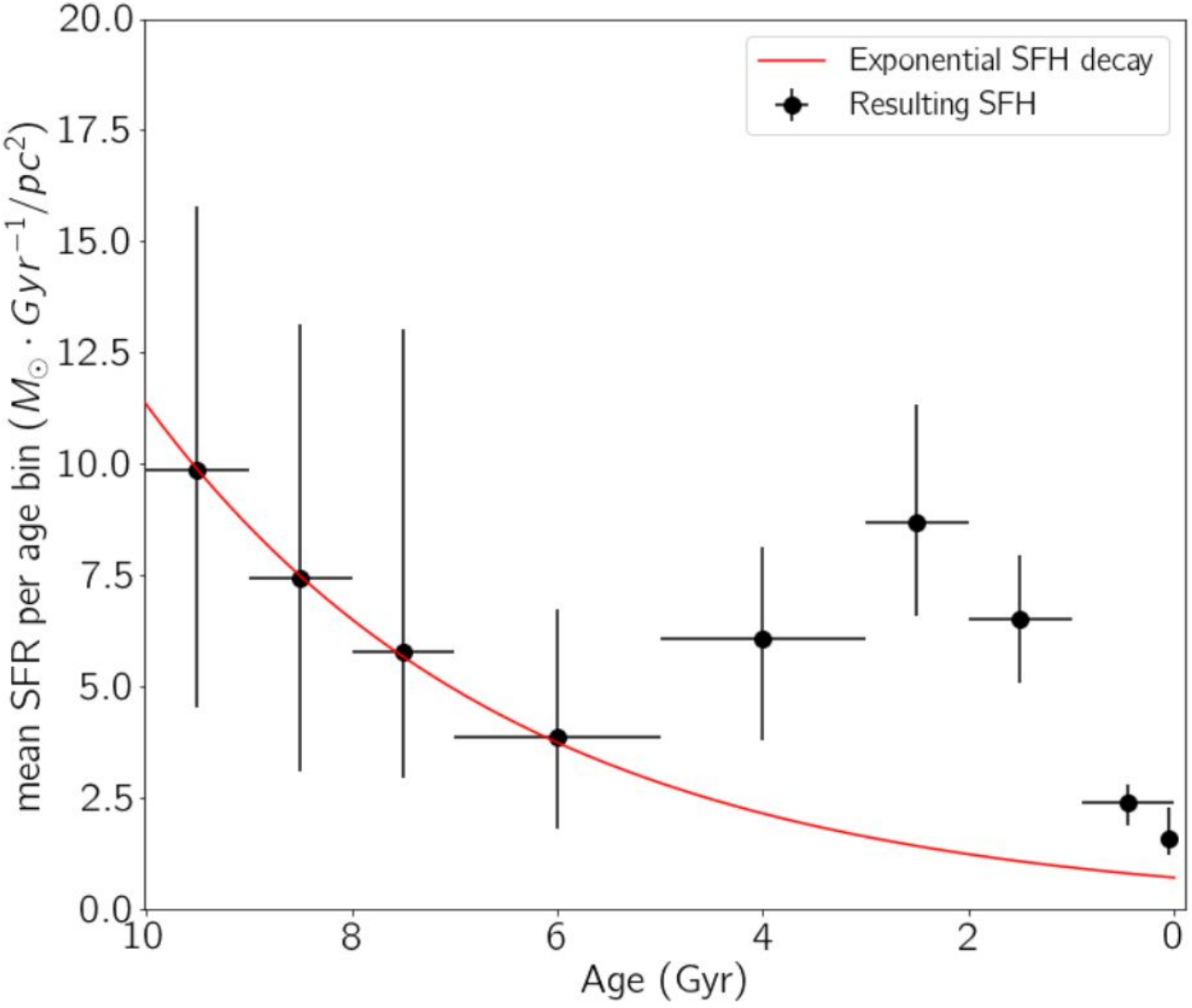
Roger Mor and GaiaUB team



Why do we care?

- Since 2018 the GaiaUB team have participated in more than 20 publications using bayesian and/or data mining techniques and/or using the GDAF Big Data platform prototype
    - Including the Gaia cross-match explained by F. Torra yesterday and the work of Alfred Castro coming later today
- These works have more than 3000 citations in total.
- But...Will this be the only way of quantify the impact in the future? E.g. what about reproducibility?

Unexpected SFR enhancement
Mor et al. (2019)



Roger Mor
Annie Robin
Francesca Figueras
Xavi Luri
and
Santi Roca-Fabrega

- Share the simulations
- Share the code
- **Share the capability of analysis**

Open Science

- Most of the ESA space mission data are public to ensure a universal accessibility

- However, new times come with large amounts of data (e.g. RVS spectra or CU8 data in Gaia as seen yesterday)

- One of the challenges for the coming years is to provide Big data mining tools **to a wider community** boosting the use of large data sets to produce scientific results.

Ultimate Goal looking to the future

"The access to the data itself is not enough anymore"

- **To offer a Open Science platform** (based on Big Data and Data Mining) **through the European Open Science Cloud portal (EOSC)**
    - To **enable** the **analysis** of the data

Collaboration

We are in collaboration with:

- Barcelona Supercomputing Centre
- Universidade da Coruña (UDC)
- University of Edinburgh
- CNRS
- Port d'Informació Científica (PIC/CIEMAT)
- University of Lisbon (UL)
- and some others

Roadmap to the Open Science platform

- Prototype: Gaia Data Analytics Framework (GDAF) ✓

- Ambition: To provide a self-deployable template to deploy a *"GDAF-like"* cluster in main the commercial cloud services.

- Challenge:To provide a self-deployable environment to deploy a "*GDAF-like*" cluster in both the commercial cloud services and local physical environments

Roadmap to the Open Science platform

- Prototype: Gaia Data Analytics Framework (GDAF) 

- Ambition: To provide a self-deployable template to deploy a *“GDAF-like”* cluster in the commercial cloud services.

- Challenge:To provide a self-deployable environment to deploy a *“GDAF-like”* cluster in both the commercial cloud services and local physical environments

The prototype:
Gaia Data Analytics Framework (GDAF)

GDAF – Environment

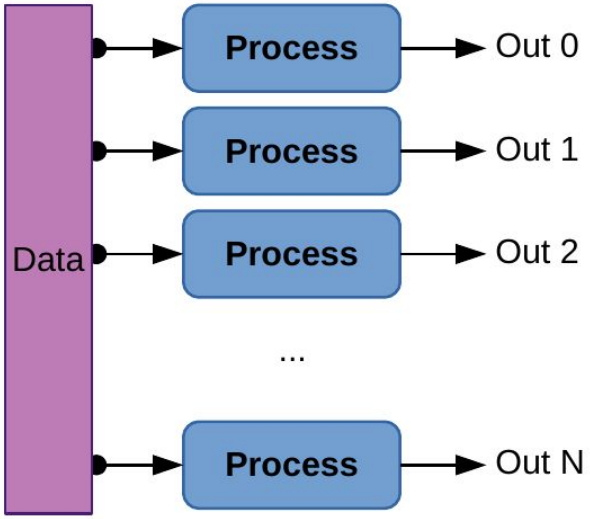
- 6 Nodes
  - 96 Cores (2 x 8 Intel Xeon 2,6 GHz each)
  - 4 TFLOPs
  - 384 GB RAM (8 x 8 GB DDR4 each)
  - 72 TB disk (12 x 1 TB HD, SATA 6 Gb/s each)



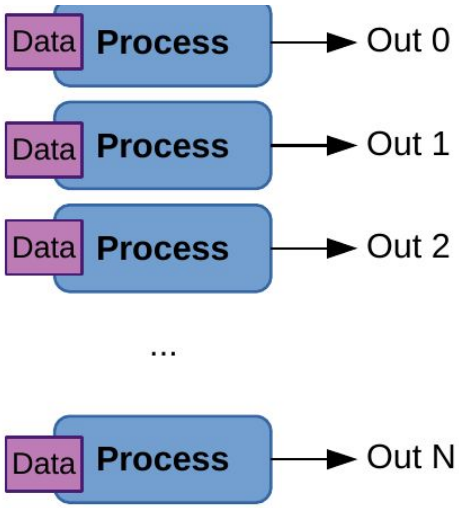
Apache Spark and Apache Hadoop

Big Data – How?

- HPC



- Hardware dependent
- Data not distributed



- Commodity hardware
- Brings code to data

Spark

hadoop

Java

Apache Zeppelin

If you are interested in using the GDAF platform please contact us:

rmor@fqa.ub.edu

Roadmap to the Open Science platform

- Prototype: Gaia Data Analytics Framework (GDAF)

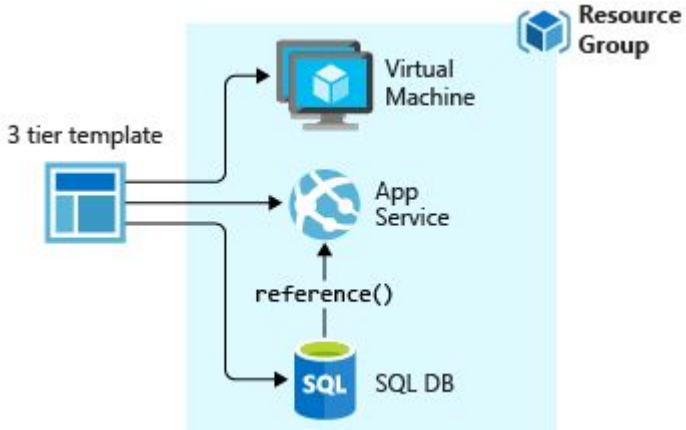
- Ambition: To provide a self-deployable template to deploy *"GDAF-like"* cluster in the commercial cloud services.

- Challenge:To provide a self-deployable environment to deploy *"GDAF-like"* cluster in both the commercial cloud services and local physical environments

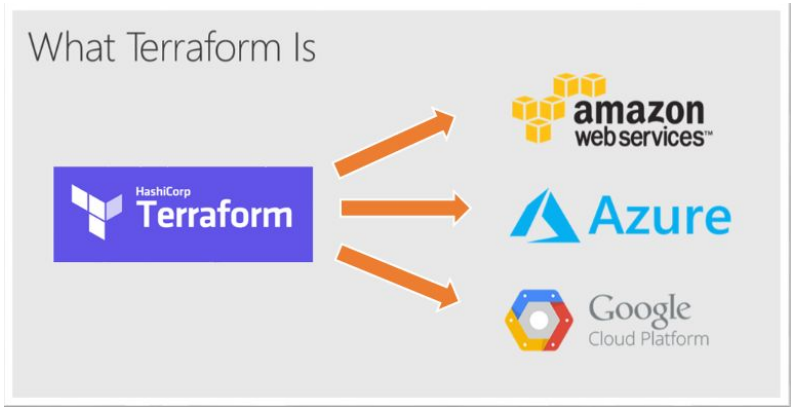
Ambition for the coming years

To provide a self-deployable template to deploy *"GDAF-like"* in the commercial cloud services.

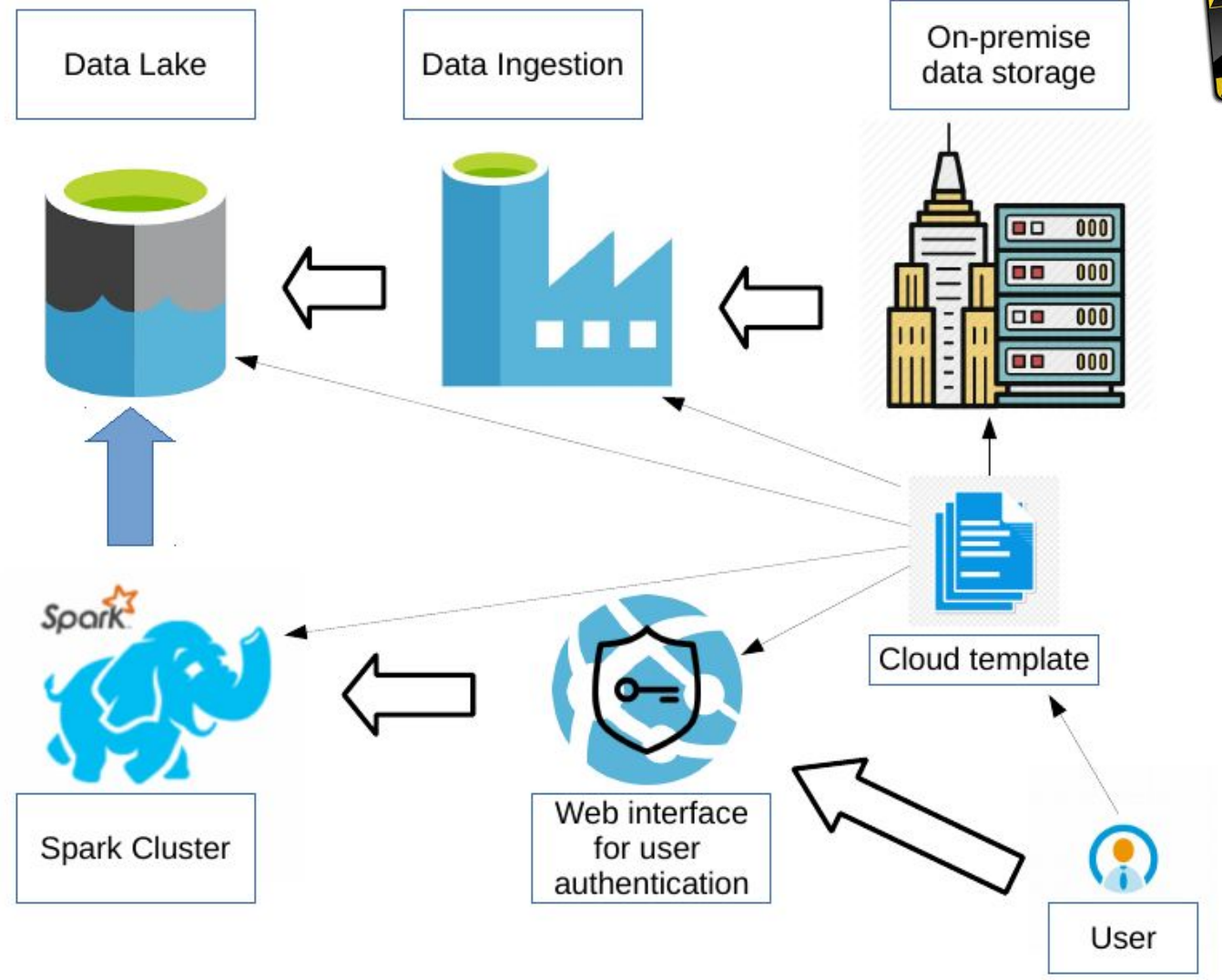
- 1 template for each cloud provider



- Standardizable template



Example in Microsoft Azure cloud



Roadmap to the Open Science platform

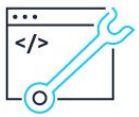
- Prototype: Gaia Data Analytics Framework (GDAF)

- Ambition: To provide a self-deployable template to deploy *"GDAF-like"* in the commercial cloud services.
  - Testing *"GDAF-like"* cluster in the commercial cloud services

- Challenge:To provide a self-deployable environment to deploy GDAF in both the commercial cloud services and local physical environments

Preliminary tests in the commercial cloud

- Microsoft Azure  
- Google Cloud  
- Amazon web services  Pending...

Next scheduled test in the roadmap

- Migrate BGM FASt (Mor et. al. 2018) to the Microsoft Azure Cloud (During 2020)
  - Bayesian inference environment to compare Milky Way simulations with observations
- Try to reproduce in the cloud the results obtained with BGM FASt in Mor et al. (2019) about IMF and the SFH (probably during 2020)

Roadmap to the Open Science platform

- Prototype: Gaia Data Analytics Framework (GDAF)

- Ambition: To provide a self-deployable template to deploy GDAF in the commercial cloud services.
  - Testing GDAF in the commercial cloud services

- Challenge:To provide a self-deployable environment to deploy a “GDAF-like” cluster in both the commercial cloud services and local physical environments

Challenge for the coming years

To provide a self-deployable environment to deploy GDAF in both the commercial cloud services and local physical environments

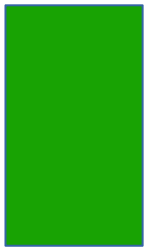
Virtualized environment with Kubernetes

First performance test scheduled

- BGM FASt (Mor et. al. 2018) is being used at University of Edinburgh for testing the preliminary stages of the virtualized platforms

To take away

- We are working in Big data and Data Mining platforms for astronomy
- If you are interested in using our GDAF prototype you can contact us through: rmor@fqa.ub.edu
- In our first steps towards an Open Science platform:
  - We are working to expand GDAF to the Commercial Cloud Services
  - We are in the preliminary stages for an efficient virtualization of a *“GDAF-like”* environment
- Our ultimate Goal for the future is to offer and Open Science platform through EOSC portal



Thanks for attending!