

Analyzing LSST data with Apache Spark

Julien Peloton, Christian Arnault & Stéphane Plaszczyński

Laboratoire de l'Accélérateur Linéaire

LSST Calcul



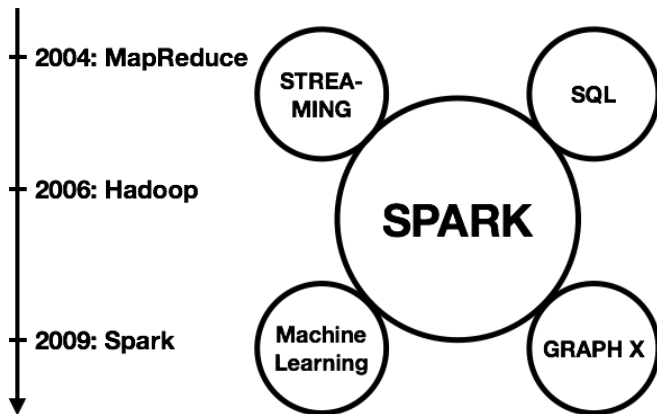
Un soupçon de démagogie pour commencer...

Hi, I would like to efficiently process TeraBytes of data.

No problem, we have a solution.

Apache Spark

- Apache Spark is a cluster-computing framework.
- Started as a research project at UC Berkeley in 2009.
- Open Source License (Apache 2.0).
- Used by +1000 companies over the world.



Current efforts: 1/3

- Develop tools to interface with the scientific world:
spark-fits (Python, Scala, Java)
 - Native Spark connector: no need to convert data format!
 - Peloton et al. (arXiv:1804.07501): distribute 1.2 TB of data in a few seconds (1280 cores).
 - Plaszczynski++ *in prep*: Spark for physicists: Simulation CoLoRe LSST 10 years, 6×10^9 galaxies (110GB)
 - Histogram redshift : 10s (136 cores)
 - Tomographical bin: 30s (136 cores)
- Multilanguage aspects: JNA (C/C++/Fortran \rightarrow Scala)

Current efforts: 2/3

- Develop methods specific to astronomical data sets: spark3D
 - Google Summer of Code 2018 project (1 internship).
 - Cross-match, neighbour search, DataBase queries,
 - Optimized for large data sets.

spark3D Quick-Start Installation About Fork me!

spark3D

Spark extension for processing large-scale 3D data sets:
Astrophysics, High Energy Physics, Meteorology, ...

Latest release v0.1.1

[Start](#) [Fork](#)

[Install Now](#)

⬇️ Load 3D object RDD
Distribute points, spheres, shells, boxes, and more using spark3D.
[Learn More](#)

⊠️ Partition your space
Partition the three-dimensional space to speed-up your search.
[Learn More](#)

🔍 Query, match, play!
Find objects based on conditions, cross-match data sets, and define your requests.
[Learn More](#)

Current efforts: 3/3

- Organisation for the development and the promotion of big data solution (<https://theastrolab.github.io>)


AstroLab Bring big data and science together

About Papers & Projects

AstroLab


Providing state-of-the-art cluster computing softwares to overcome modern science challenges

[Learn more](#)

 **spark-fits**

Distribute FITS data with Apache Spark: Binary tables, images and more!

[Learn More](#)

 **spark3D**

Apache Spark extension for processing large-scale 3D data sets: Astrophysics, High Energy Physics, Meteorology, ...

[Learn More](#)

>_ Multilanguage aspects

Interface Scala and Spark with your favourite languages.

[Learn More](#)

- Interfacing with LSST tools
 - Started (w/ J. Neveu): project using CTIO data (telescope auxilliaire)
 - Started: Image to catalogs
 - Planned: Connection to DC2 data
 - Planned: Connection with the Stack tools (data calibration, reduction, coaddition)

Infrastructures and community effort

Infrastructures

- Dedicated cluster at LAL (medium size)
 - 9 machines, 162 cores total, 307.8 GB RAM total.
 - Cannot handle yet the $O(100)$ TB data set
- **Currently, no specific Spark infrastructure at CC-IN2P3**
 - We can not work at CC currently, and we have no direct DC2 data access.

Contact within the community

- Contact with SLAC
- NERSC access (Spark installed at NERSC, run using Shifter).
- Meeting with CERN IT mid-July