# Big data astronomy with Apache Spark

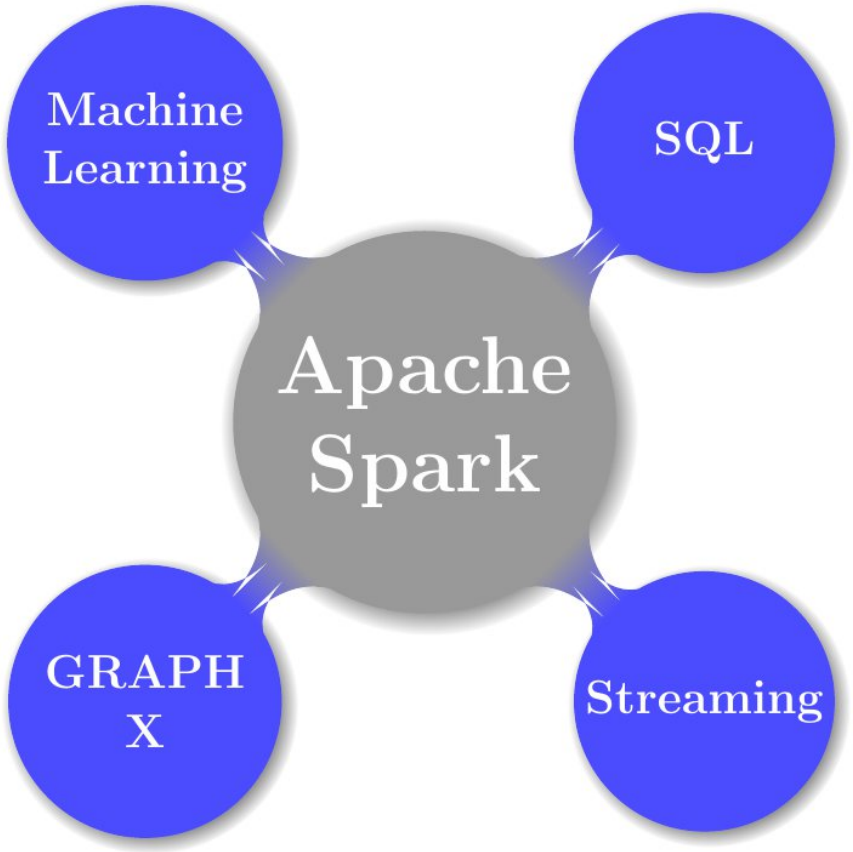
C. Arnault, J-E. Campagne, J. Peloton, **S .Plaszczynski**...

IJCLab
Irène Joliot-Curie
Laboratoire de Physique des 2 Infinis

université
PARIS-SACLAY

- Open source cluster-computing framework to deal with large datasets
- 2004: MapReduce (Google)
- 2006: Hadoop
- 2009: Apache Spark (~Hadoop++ with cache at that time)

Used by 1000+ companies in the world - occasionally used in science (Particle Physics, Genomics, Astronomy).




https://astrolabsoftware.github.io/

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

## spark-fits

Distribute FITS data with Apache Spark: Binary tables, images and more! API for Scala, Java, Python and R.

Learn More

## spark3D

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

Learn More

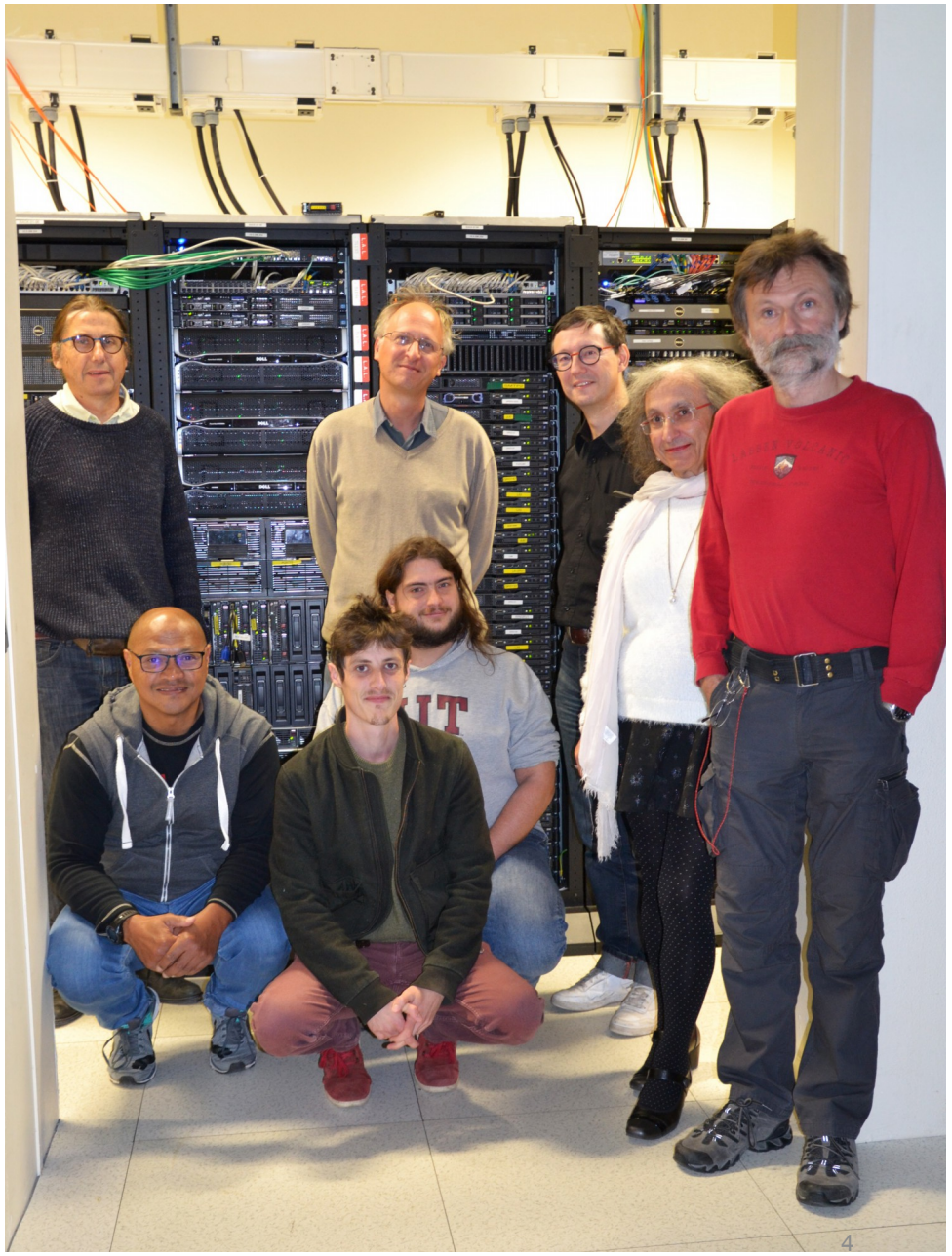
## Interfaces

Interface Scala and Spark with your favourite languages: C/C++/Fortran and more!

Learn More

# The team @ IJCLab

- Researchers/IT mix (~10 people)
- We maintain a small *Spark cluster* @"VirtualData" (Paris-Saclay):
  - 10 machines
  - 18 cores/36 GB RAM each
  - HDFS (multi-TB storage)
- Main focus on astronomy
  - Involved into LSST
- use NERSC supercomputers for brute force
  - Scaling up to 1000s of cores



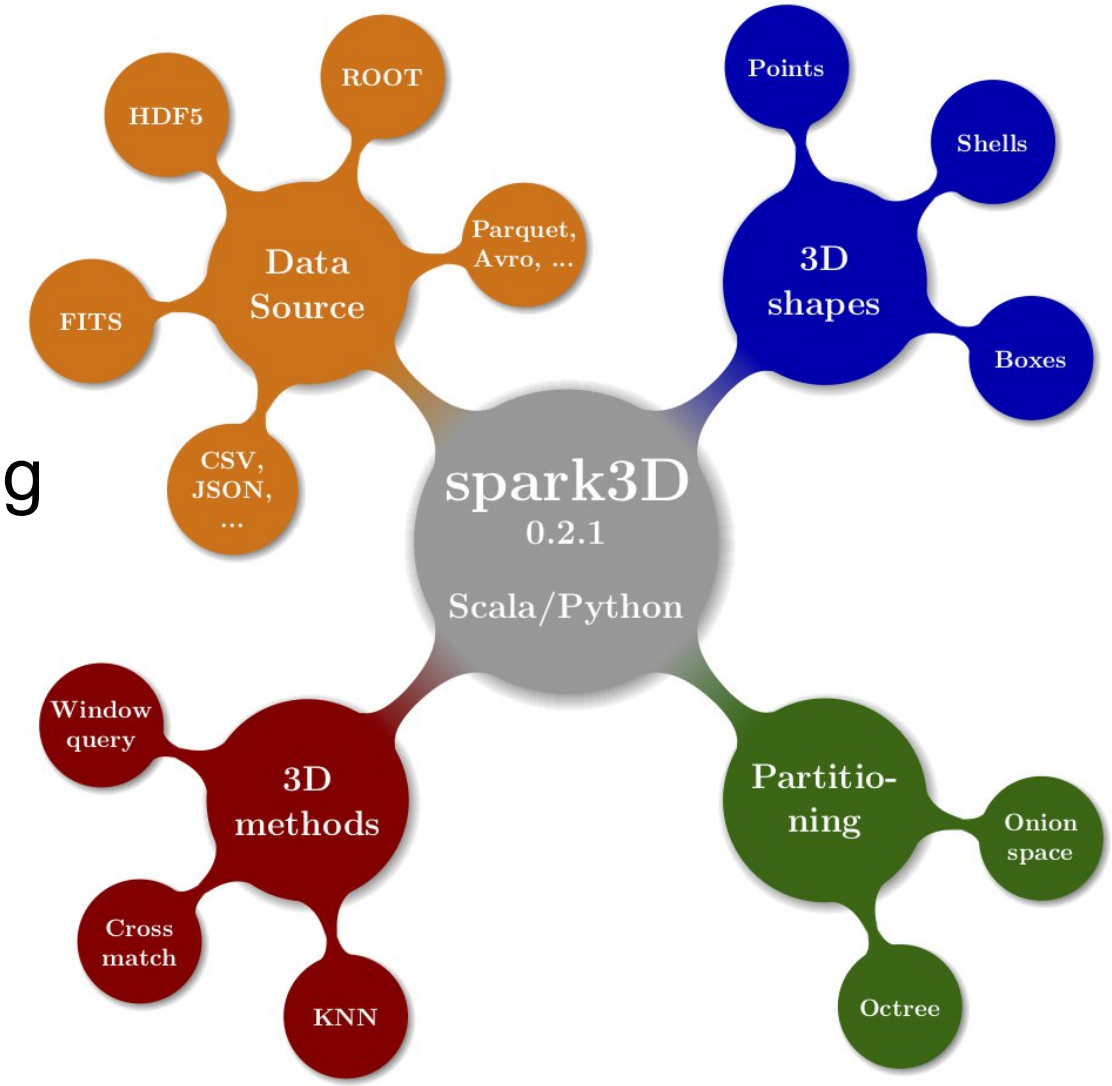
# Spark-Fits


- Astronomy requires *structured* data, FITS is a standard
- wrote a high quality FITS reader
  - Manipulation of binTables and images in a distributed environment
- Different API: Scala, Python, Java and R
- Used by different communities: LSST, HST, SKA, CosmoHub

*Peloton, Arnault, Plaszczynski (2018) Comput. Softw. Big Sci, 2,7*

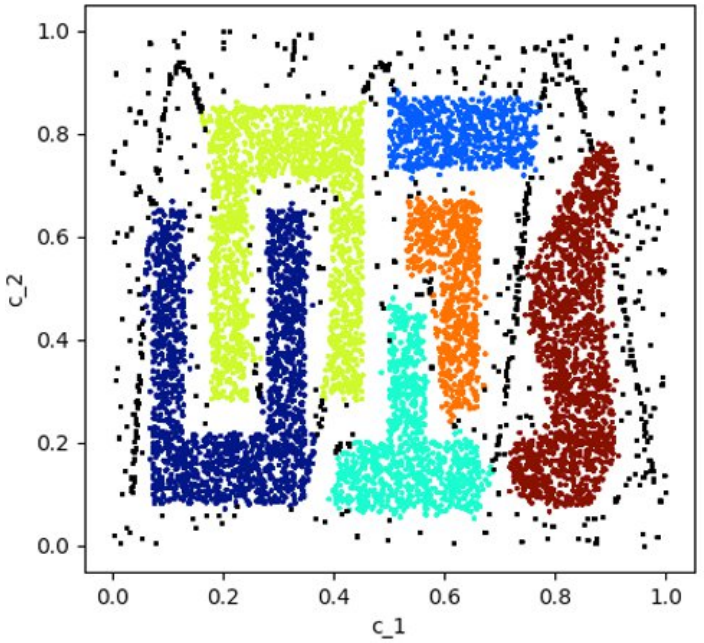
# Spark3D

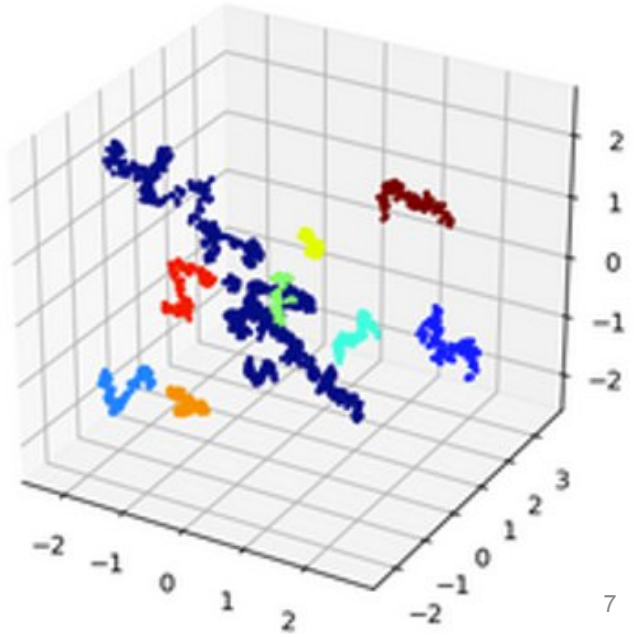
- Extension of Spark SQL for manipulation of 3D data
- 3D distributed partitioning
  - KDTree, Octree, shells, ...
- Distributed spatial queries & data mining
  - KNN, join, dbscan, …
  - Typical usage on million/billion rows
- Visualisation
  - Client/server architecture



# RP-DBSCAN

- Density Based Clustering (of Appl. with Noise) designed to be run on Spark cluster
- first really efficient algorithm: Hwanjun Song, Jae-Gil Lee SIGMOD'18, June 10-15, 2018, Houston, TX, USA)
- cell based+random partitioning
- Re-implemented in Scala (J.E Campagne 2019). Tested on our Spark Cluster.



# LSST

**In a nutshell:**

- new generation telescope in Chili (US+FR)
- wide+deep extra-galactic survey
- 8.4 m primary mirror
- World's largest CCD camera: 3.2 Gpixels

**In numbers:**

- 10-year survey, starting 2022 (+1?)
- ~half of the sky  every few days
- 1,000 images/night = 15TB/night
- 10 million transient candidates per night
- billions of galaxies for cosmology (DESC)

# User analysis

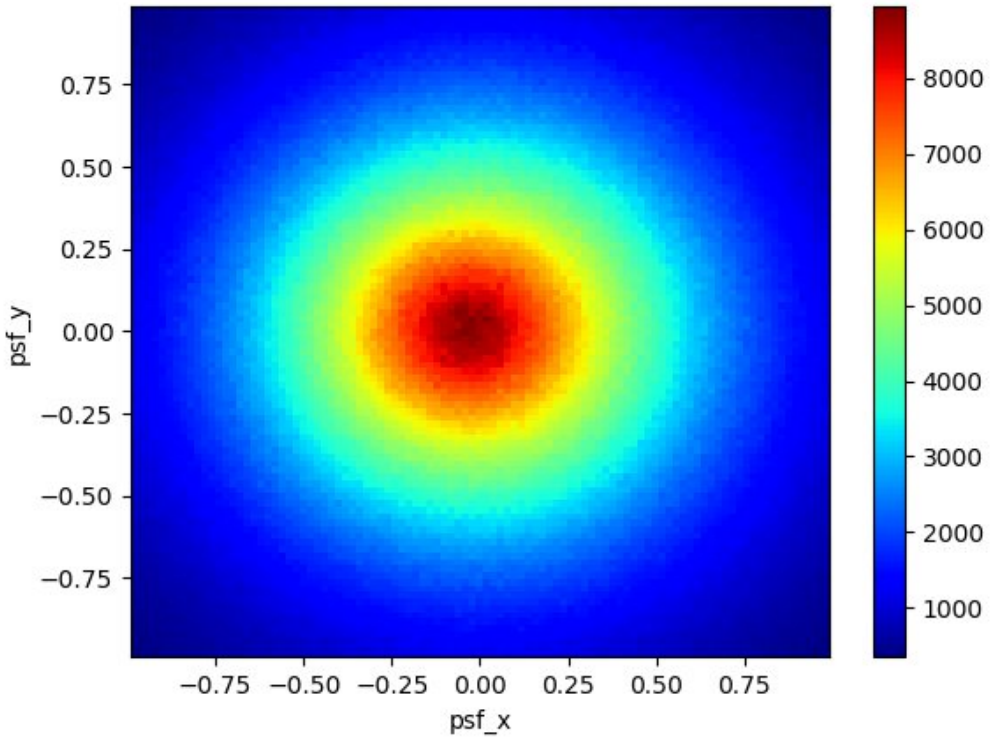
- spark-sql dataframes well suited (and simple) for *interactive* analysis of galactic data
- Fast sim : 6 $10^9$ galaxies

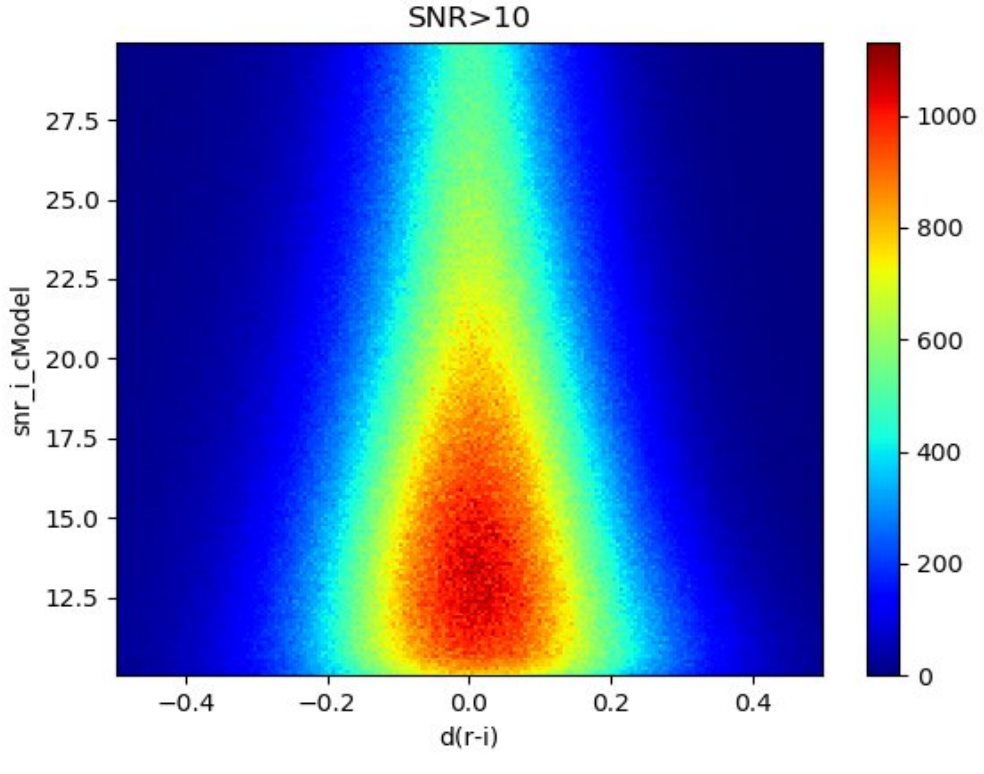
Plaszczynski, Peloton, Arnault,Campagne (2019), Astr. & Comp., 28.

| Section | analysis | python | scala |
|---|---|---|---|
| 4.3.2 | load(HDU) | 2.8 ± 0.1 | 8.8 ± 0.2 |
| | PZ + show(5) | 12.4 ± 0.6 | 13.7 ± 1.2 |
| 4.3.4 | cache (count) | 97.7 ± 4.0 | 95.4 ± 5.0 |
| 4.3.5 | stat(z) | 3.9 ± 1.5 | 4.9 ± 2.5 |
| | stat(all) | 9.8 ± 1.0 | 11.0 ± 0.9 |
| | minmax(z) | 1.8 ± 0.3 | 3.2 ± 0.7 |
| 4.3.6 | histo (dataframe) | 11.5 ± 1.5 | 13.0 ± 0.8 |
| | histo (UDF) | 114.9 ± 5.6 | 13.9 ± 1.2 |
| | histo (pandas UDF) | 43.3 ± 4.5 | - |
| 4.3.7 | 1 shell | 30 ± 3 | 13 ± 2 |
| | all shells (10) | 307 ± 34 | 130 ± 18 |

# Validation

- DESC released a very detailed galactic simulation "DC2"
- ran the whole reconstruction pipeline -> measured catalog
- cross-match with Spark 50M x 80M : ~3 mins



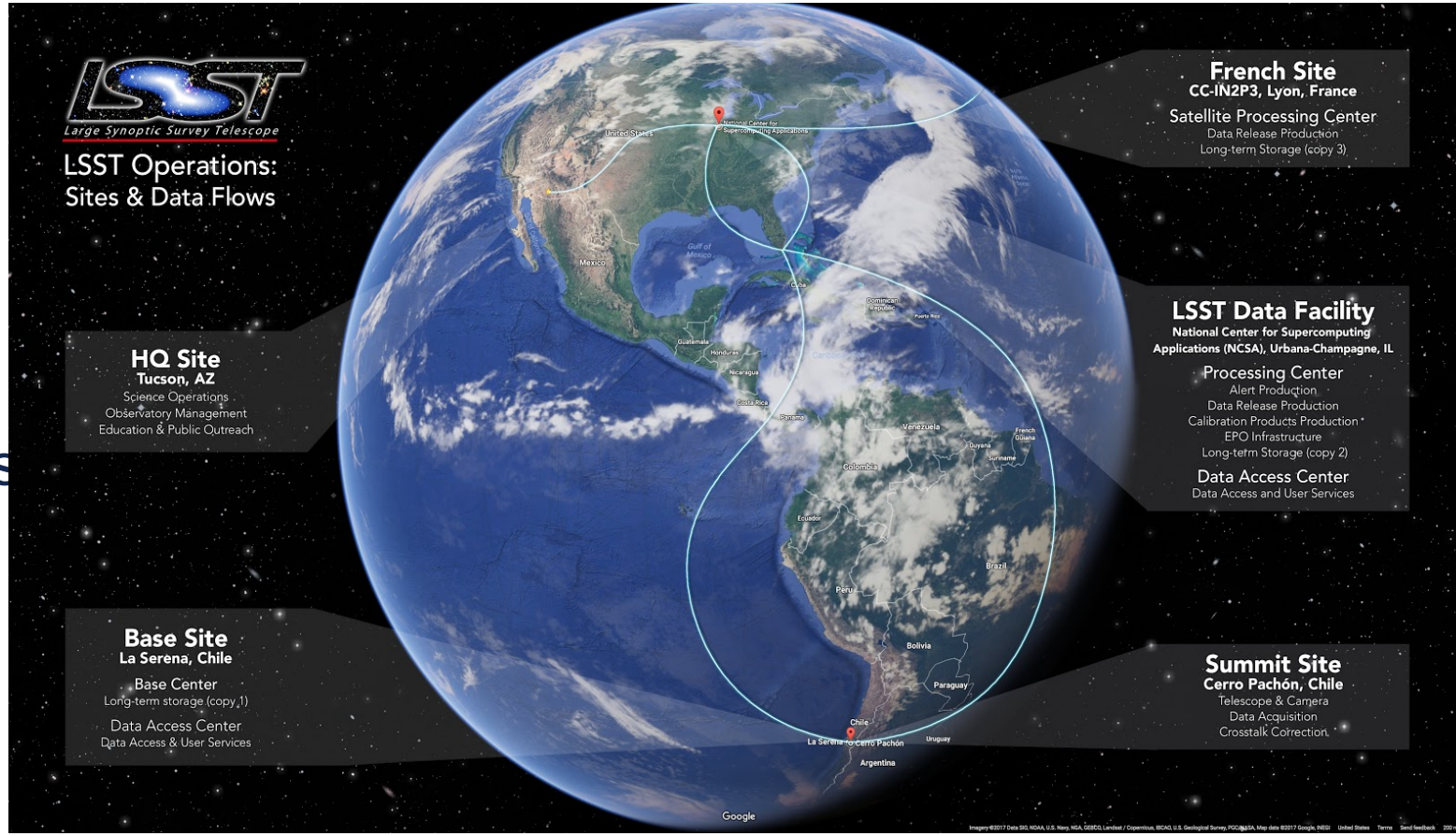



# SparkCorr

- **counting the number of pairs** in bins is a classical task (auto/cross correlations) : tomography
- combinatorics is huge
- existing (MPI) algorithms can make it up to ~10M in mins
- but not up to 100-200M necessary for LSST
- first results :
  - 1 min for 100M
  - 3 mins for 500M (32 nodes @NERSC)
- link to geometric graphs

# The transient sky

- Forecasted: 10 million alerts per night…
- ~82KB/alert, 800 GB/night (3PB in 2030)
- 98% of alerts must be transmitted with 60 seconds of readout...
- … and processed before the next night!
- Wires to send alerts worldwide are not infinitely big...



# Brokers : collect, enrich, redistribute





Covering LSST science pillars:

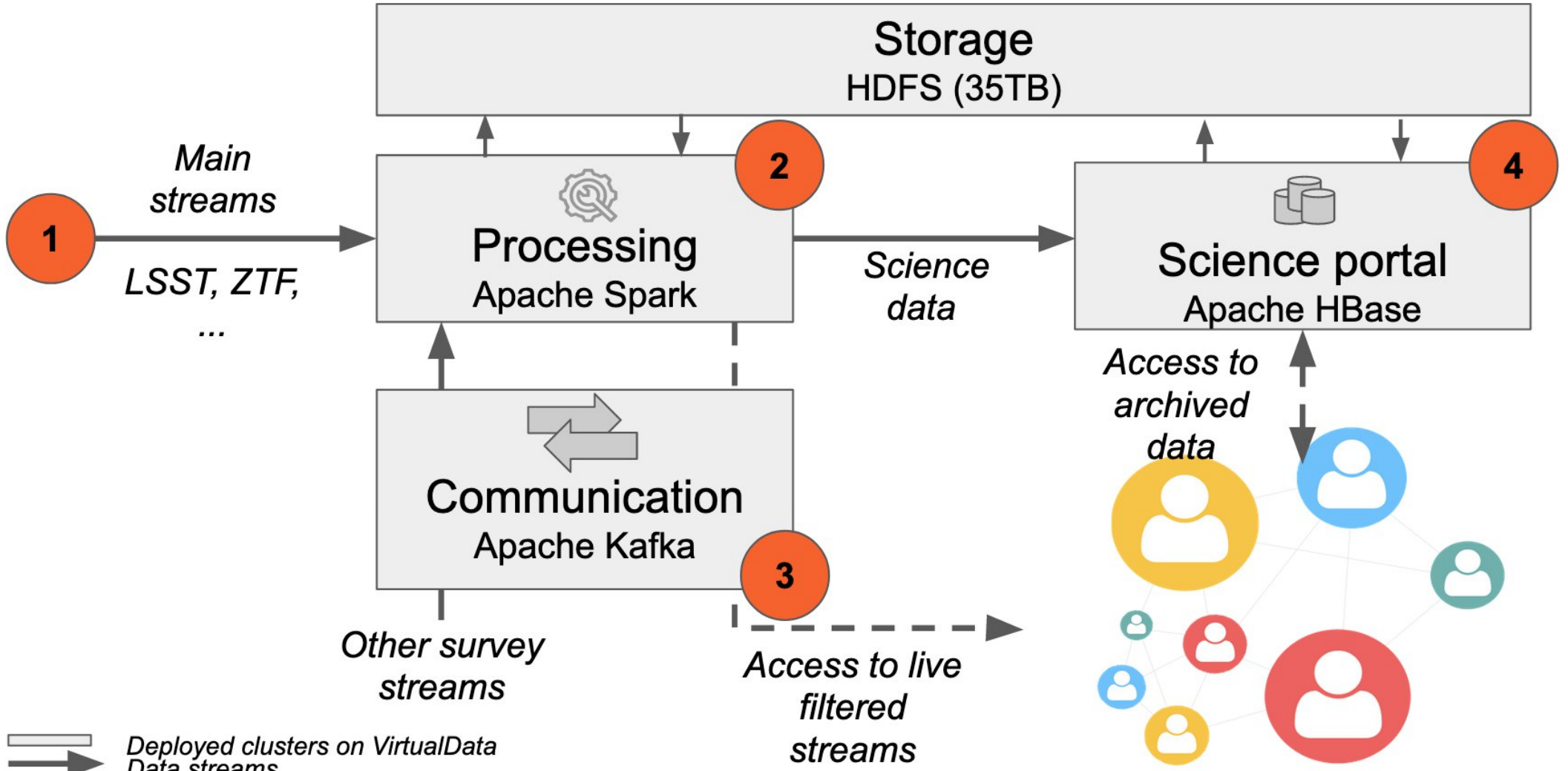
- Probing Dark Matter and Dark Energy
- Taking an Inventory of the Solar System
- Exploring the Transient Optical Sky
- Mapping the Milky Way

# Fink broker

- ~10 million alerts/night (~1TB/night).

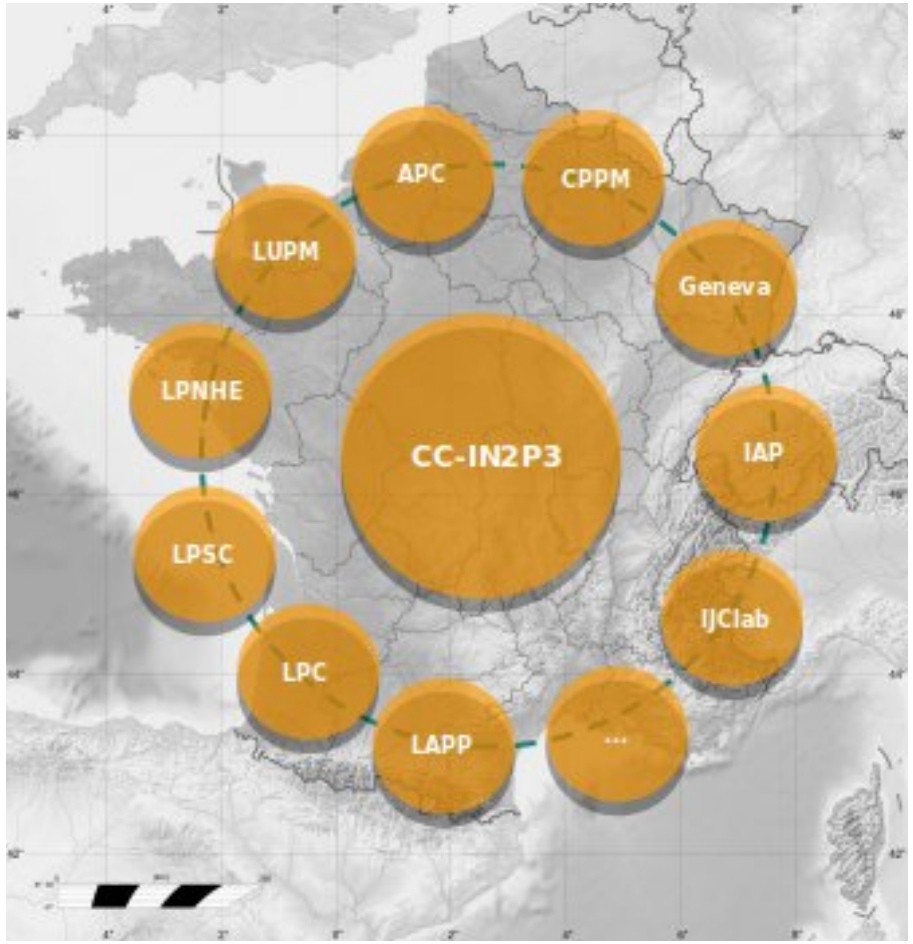
- Our solution: distributing the load
  - Distributed computation (Apache Spark Streaming)
  - Distributed streaming (Apache Kafka)
  - Distributed database (Apache HBase)
  - Distributed graph (JanusGraph)
- Fink: broker using the big data ecosystem, based on cloud infrastructures.
  - Focus on supernovae (ML) , microlensing, multi-messenger astronomy (GRB,GW ...)

# Fink @ VirtualData



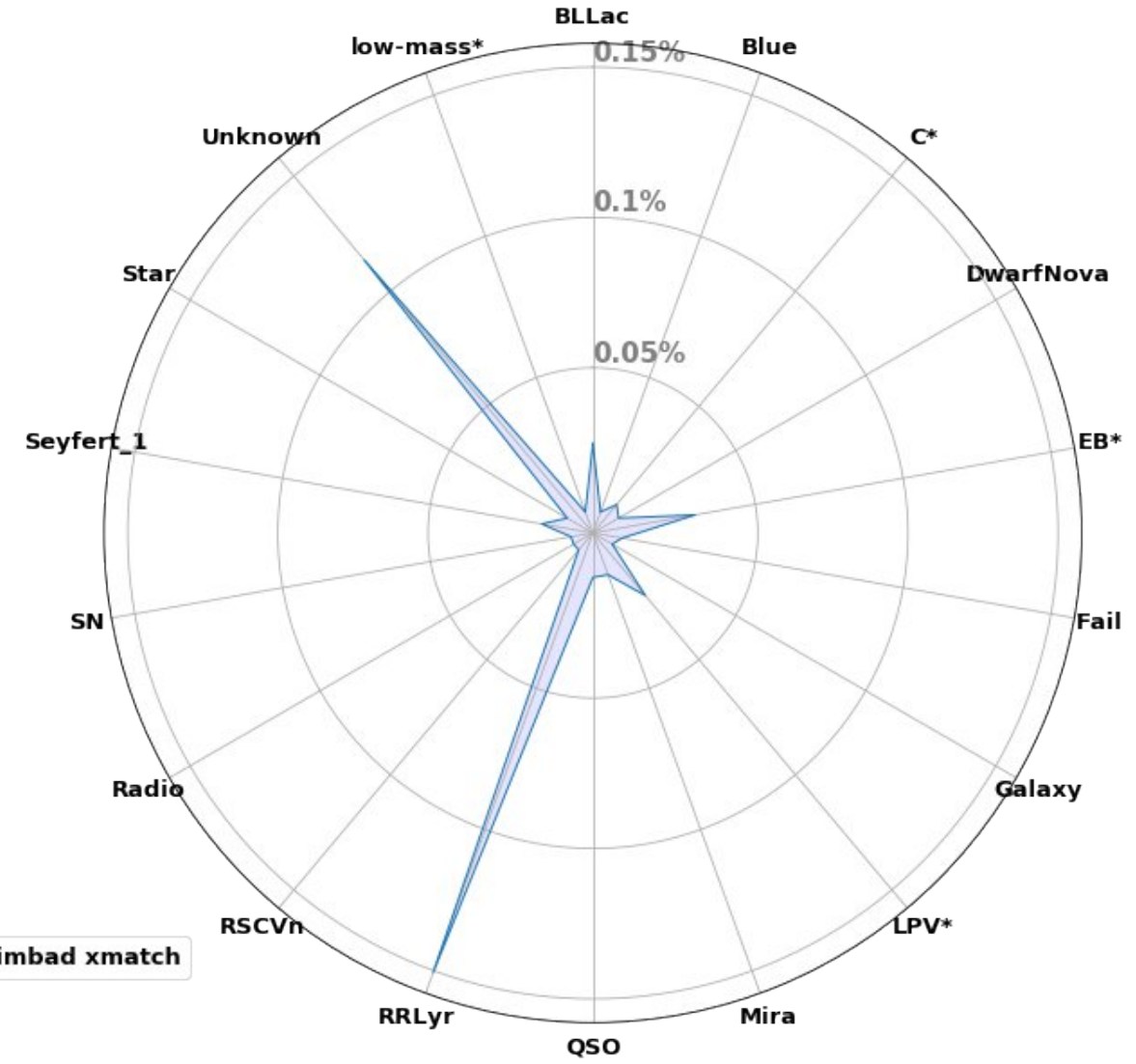
# Performances

- Processing ZTF data @Virtual Data
  - Pathfinder for LSST
  - ~100K alerts/night (~10 GB/night)
- Simulating LSST streams, and beyond
  - We deployed a Kafka cluster to generate streams on-demand.
  - Processing routinely 20K alerts/min, and peak processing at 100K alerts/min.
  - Bottleneck is the storage, not (yet) the CPU! At a rate of 1TB/night, disks are quickly full...
- Trying to have cc-in2p3 on board
- answer: dec 2020



# Cross-matching at scale

- Cross-match service
- Question: Are there already known objects in the stream?
- Problem: Standard cross-match is very slow! (we receive 10,000 alerts / 30 seconds)
- Our solution: Distributed cross-match using xmatch service @ CDS Strasbourg.
- Cross-match of thousands of objects per second.



# Conclusions

- We've been investigating the Spark solution for some years: huge potential for astronomy/cosmology
- simple/efficient interactive analysis with dataframes (teaching)
- efficient 3D tools (partitioning, DBSCAN,correlations...)
- Spark streaming: new technology that  allows to follow modern alerts rates
- we use some ML methods in the broker but not obvious to interface to existing code (pytorch)
- beginning to have some experience on graphs: promising.
- but community is not yet ready to switch

# Discussion

- user side: easy /efficient data analysis with dataframes (nothing new, in HEP ntuples known since the 80's...). But python/notebooks inertia (+publication pressure).
- little developers (language not understood by the former)

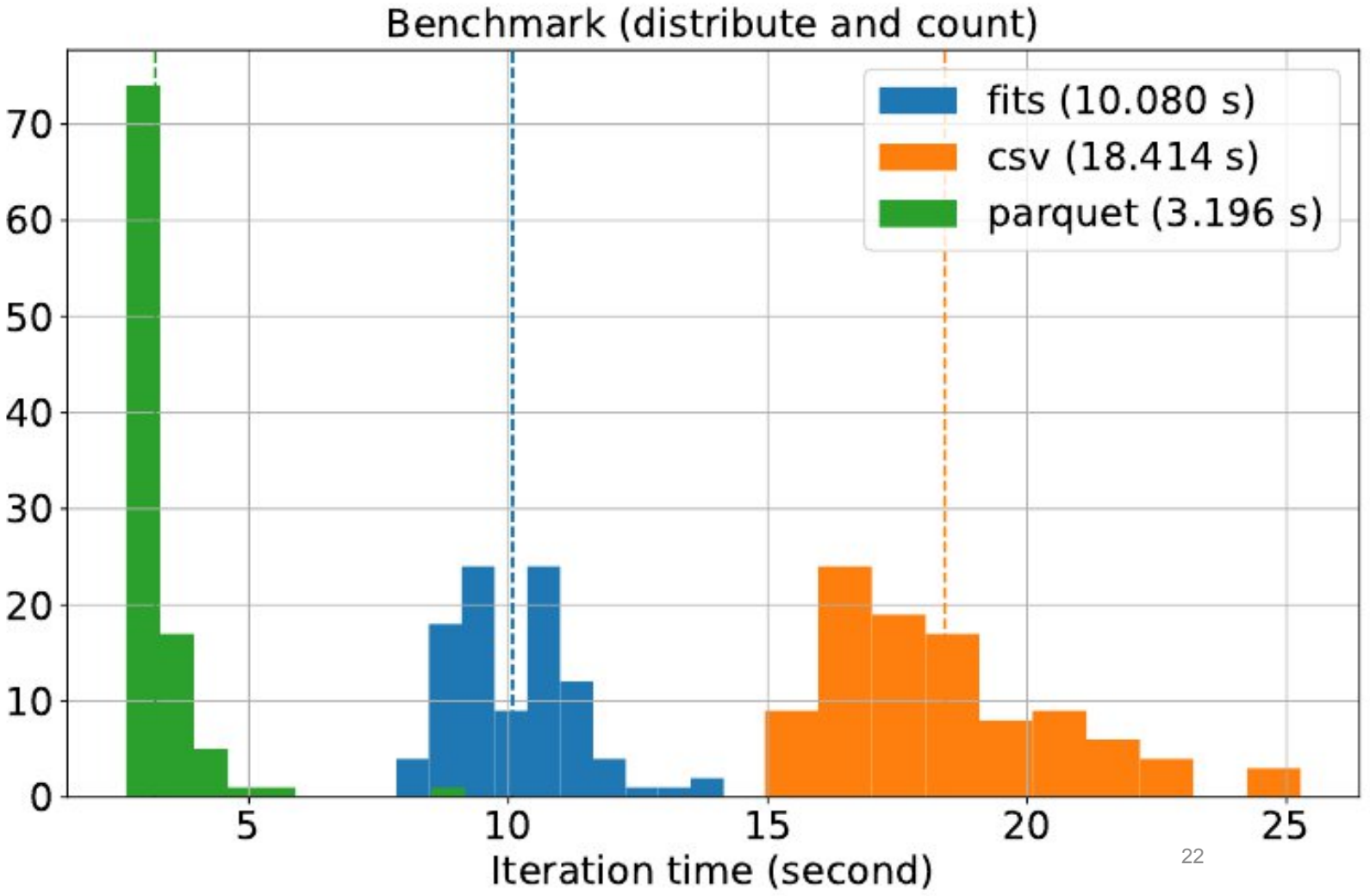
Gap between users/developers has grown (same in industry)

1. quick&dirty (but efficient) single-person analysis (physicists)
2. build a solid projet (engineers)

# Backup

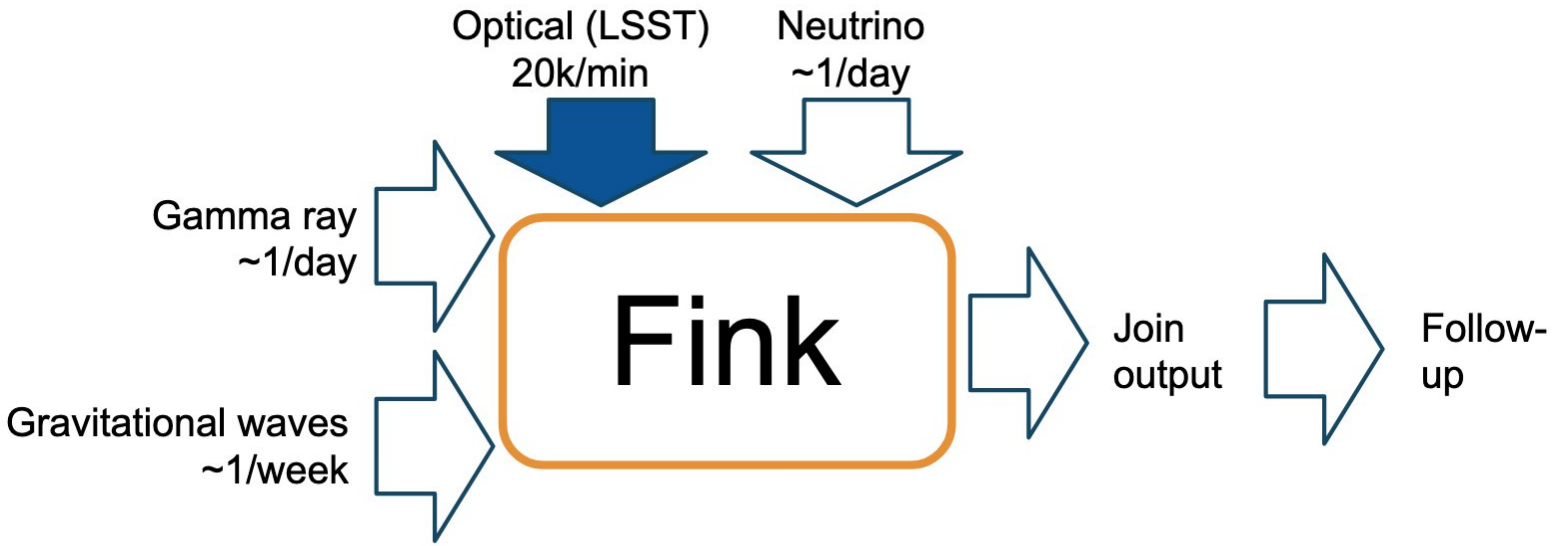
# Spark-fits performances

- Billion rows
- Comparison to other connectors:
  - CSV
  - Parquet
- Comparable performances



# Multi-messenger astronomy

- Identifying interesting LSST alerts is only part of the story
  - we need coordination with other surveys and facilities existing networks to allow follow-up of interesting sources.
  - Gamma ray bursts, neutrino telescopes, gravitational wave detectors, ...
- Performing stream-stream join with Apache Spark

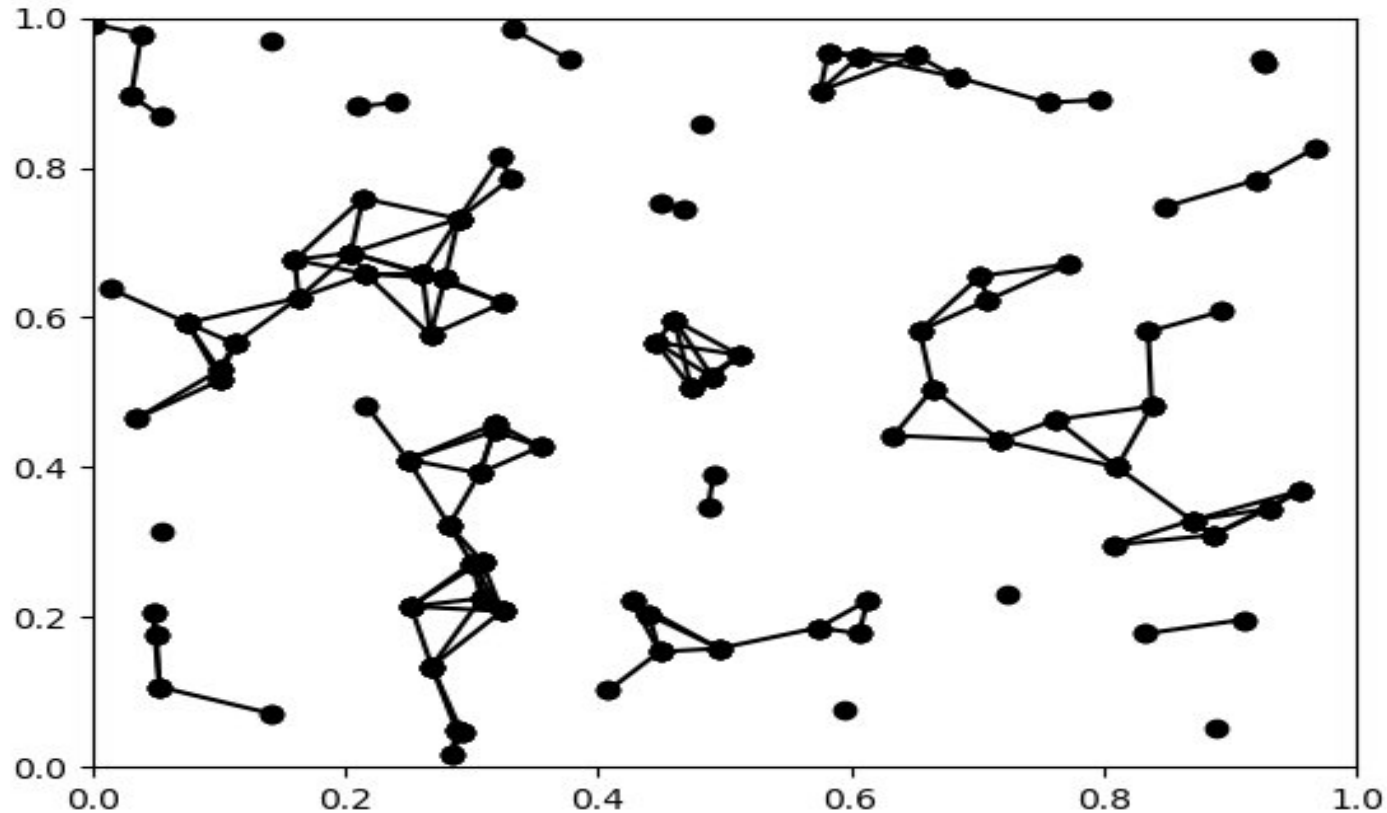


# Apache Spark cluster @ VirtualData

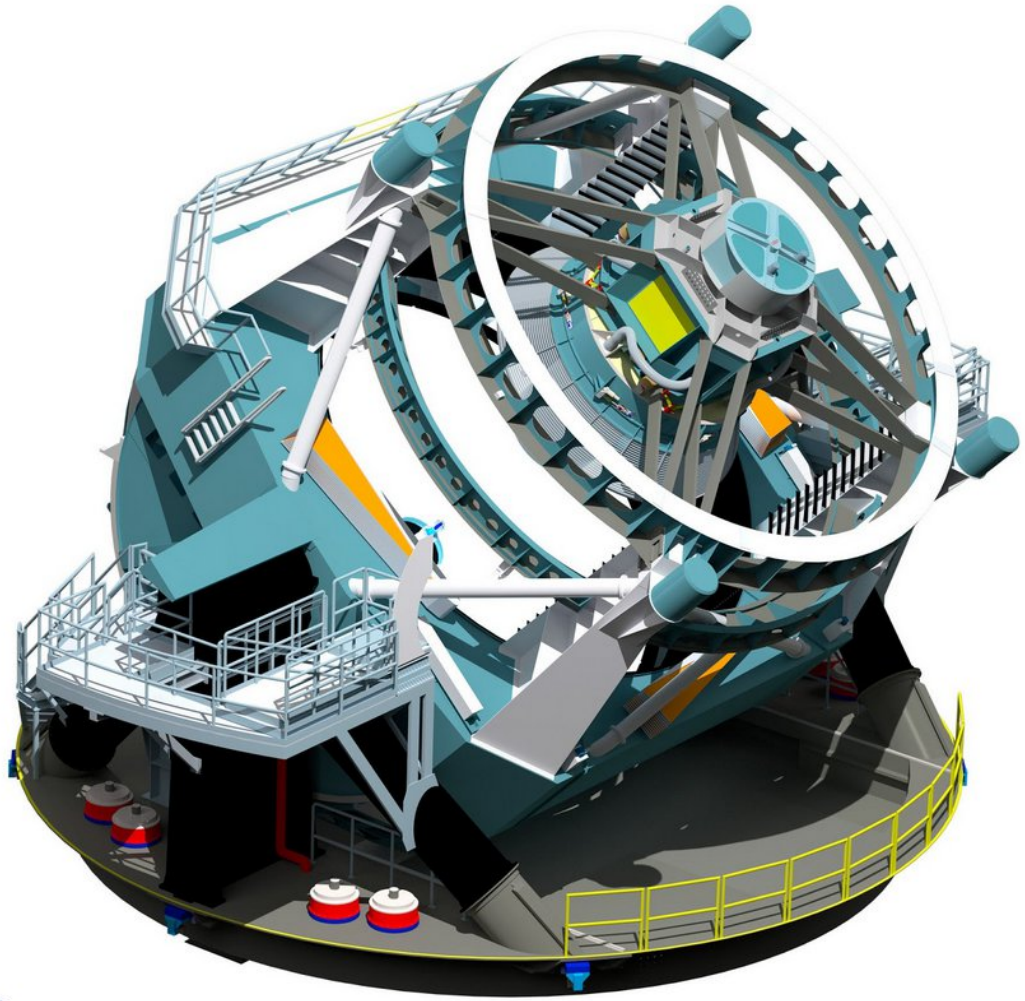
- Multi-tenant cluster, mainly used for genomics and astronomy, and more recently geosciences.
  - Computing: 1+9 machines (18 CPU each, 2GB RAM/core)
  - Cluster manager: Apache Mesos / transitioning to Kubernetes
  - Storage: 35 TB on HDFS (replication factor x3) / transitioning to Ceph/S3
  - Software distribution: CernVM FS
  - Monitoring: Ganglia, Grafana

Scaling jobs from MB to TB in one place!

# Geometric Graphs



# LSST data products



**Now**

**Raw Data**
Sequential 30s image, 15TB/night

**60s**

**Prompt Data Product**
Difference Image Analysis
Alerts: up to 10 million per night

Public data!

**24h**

**Prompt Products DataBase**
Images, Object and Source catalogs from DIA
Orbit catalog for ~6 million Solar System bodies

**Year**

**Annual Data Release**
Accessible via the LSST Science Platform &
LSST Data Access Centers.

**End**

**Final 10yr Data Release**
Images: 5.5 million x 3.2 Gpx
Catalog: 15PB, 37 billion objects