Big Data cosmology with Spark

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

LAL, Univ. Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France

November 19, 2019

Computing since 2 decades

- processors freq was frozen ($P \propto f^3$)
- →multi-core architecture, GPU, (FPGA)

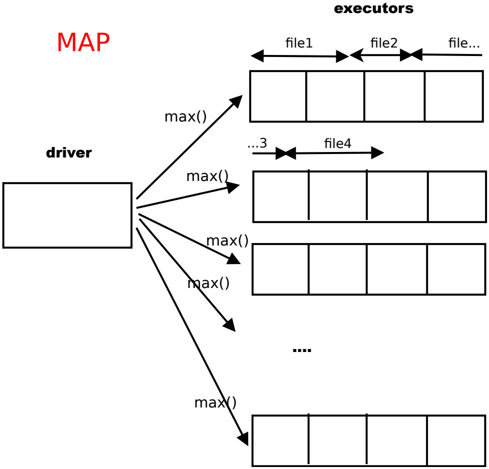
HPC (High *Performance* Computing): optimize ”arithmetic efficiency” (#ops/#data moves)

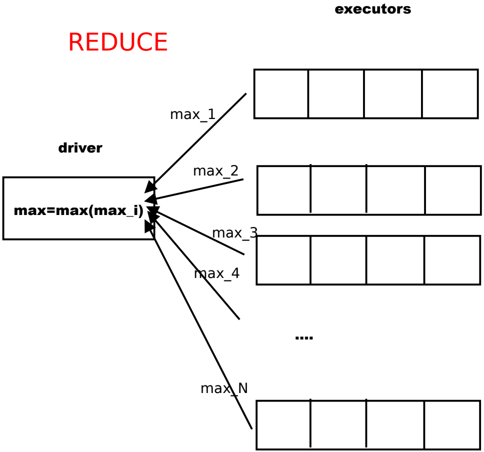
- complicated (`OpenMP, MPI, C++11/14/17,vectorization, CUDA/OpenCL...`)
- work on (very expensive) supercomputers

HTC (High *Throughput* Computing), aka ”Big Data”

- 2004 `Google`: `mapReduce` programming model: foundation of *distributed computing* on data centers
- 2006 `Hadoop` ecosystem develops..
- 2004 `scala` (from `java` ecosystem).
- 2009 `Spark`: research project at UC. Berkeley
- 2015 `Spark` SQL (dataframes)
- today: adopted by 1000++ companies, very active community, open-source
- 2018 `https://astrolabsoftware.github.io`

mapReduce


executors
MAP
file1
file2
file...
max()
...3
file4
driver
max()
max()
max()
....
max()



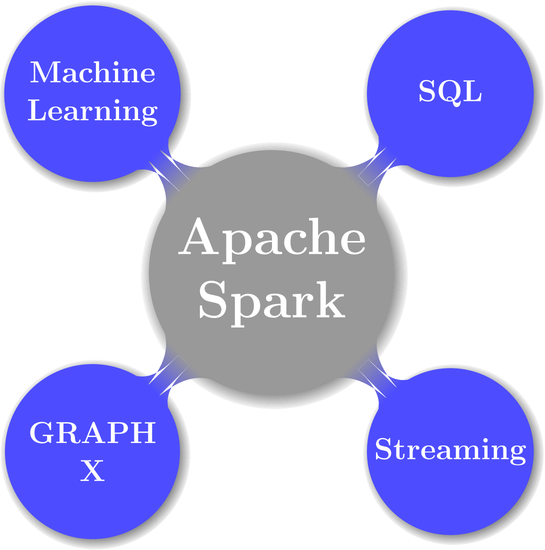
```
dataframe.select(max("variable"))
```

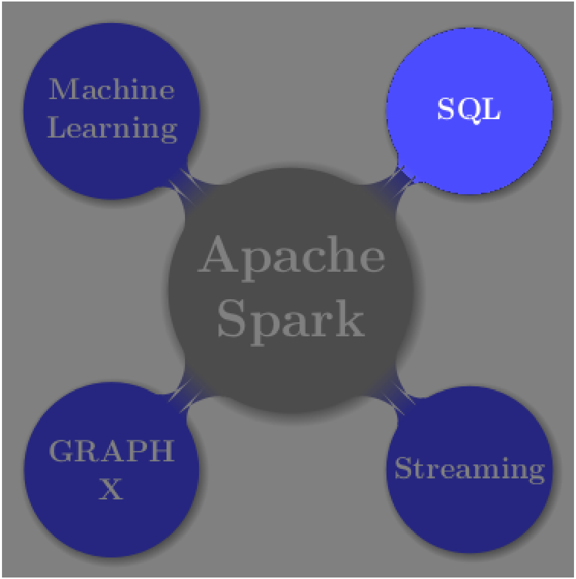
Spark in practice

- a set of (highly optimized) objects/functions to perform distributed computing hiding its complexity
- scala (java), python, R
- this is *Functional Programming* but you don’t need to know it!

Advantages

1. coarse grain parallelization over huge datasets
2. automatic pipeline optimization (lazy evaluation)
3. put data in cache →interactive work
4. scaling

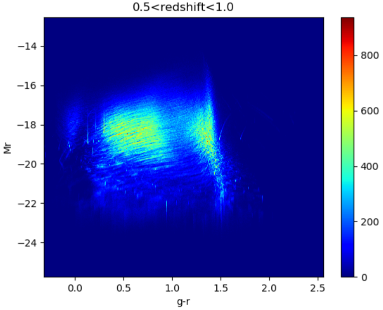

Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

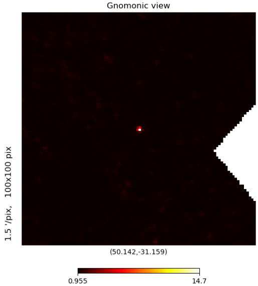

Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

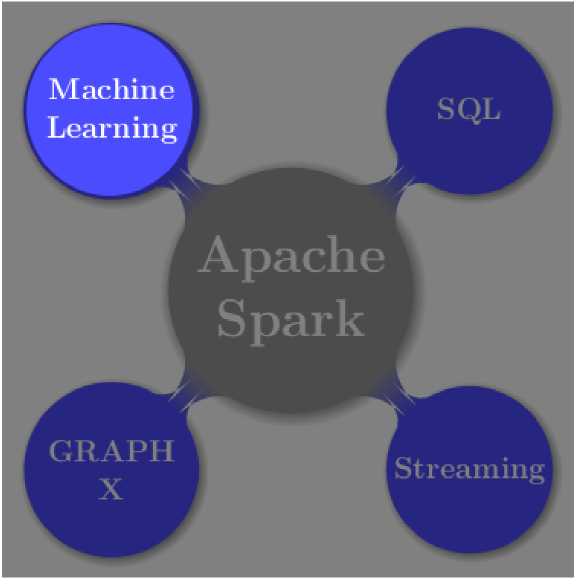
(interactive) data analysis

based on *dataframe* (named columns of known type... =n-tuple)

- histogram of redshifts on 6.10^9 simulated galaxies: $\simeq$10s
- cross-matching catalogs: (DC2) $80M \times 500M$: 3 mins
- data mining (outliers study)

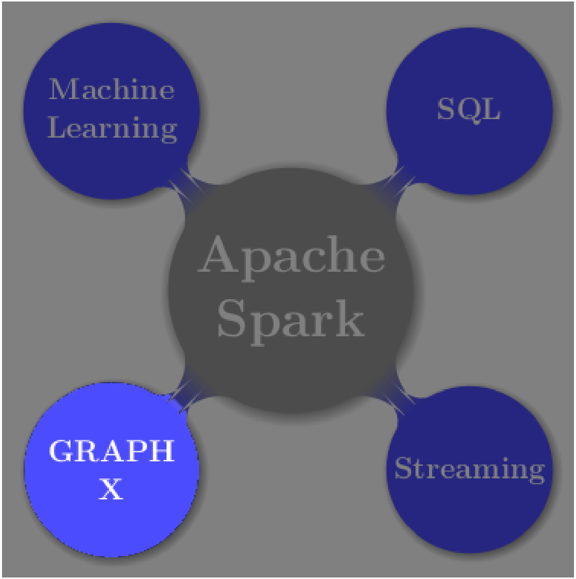





Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

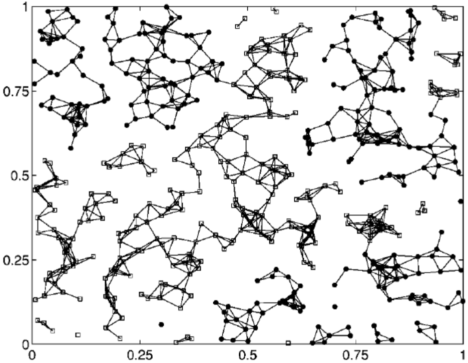
Spark ML

- ‘google theorem’: data size matters more than algorithm
- distributed classical algorithms: trees, SVM, regressions, MLP but no deep-learning (CNNs)
- external libs to Keras exists but can it beat GPUs?

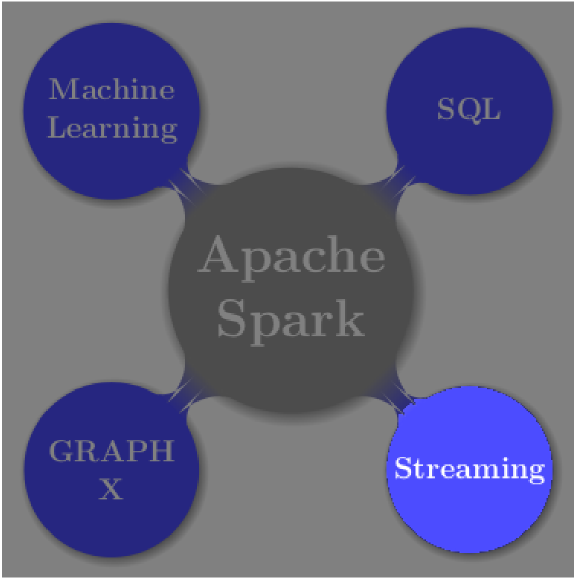

Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

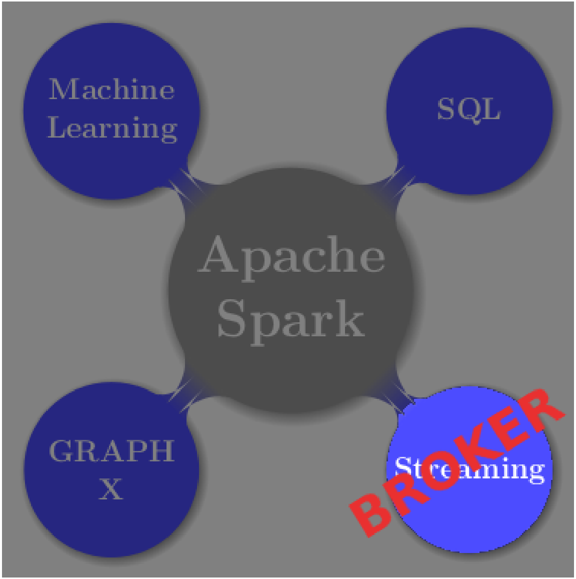
Graph-X

- Cosmic Web, Nbody-sims
- FoF, skeleton, MST, topology, Count-in-cell, 2pcf++



Dall,Christensen (2002)


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming
BROKER

FP

- we all code in a *imperative* way (nothing to do with procedural vs. OO)
- inherited from Turing machines (variable, states)
- but before (1933) was $\lambda-$ calculus
- rather theoretical language (used in math/logic, theorems proofs etc.)
- basic objects are ”functions” not variables (const). closer to math meaning.
- some paradigms to code
- was used in confidential languages (Lisp, Haskell)
- rediscovered today (scala, computing power)
- quite concise, clean, robust. allows (often) to scale.