SPARKING PANDAS: AN EXPERIMENT

PyConOtto - Florence '17

Francesco Bruni

brunifrancesco

WHO I AM

MSc in Telecommunication Engineering

Functional pythonista

Currently working with geo data

OUTLINE

Why *Sparking* Pandas

Functional data processing pipelines

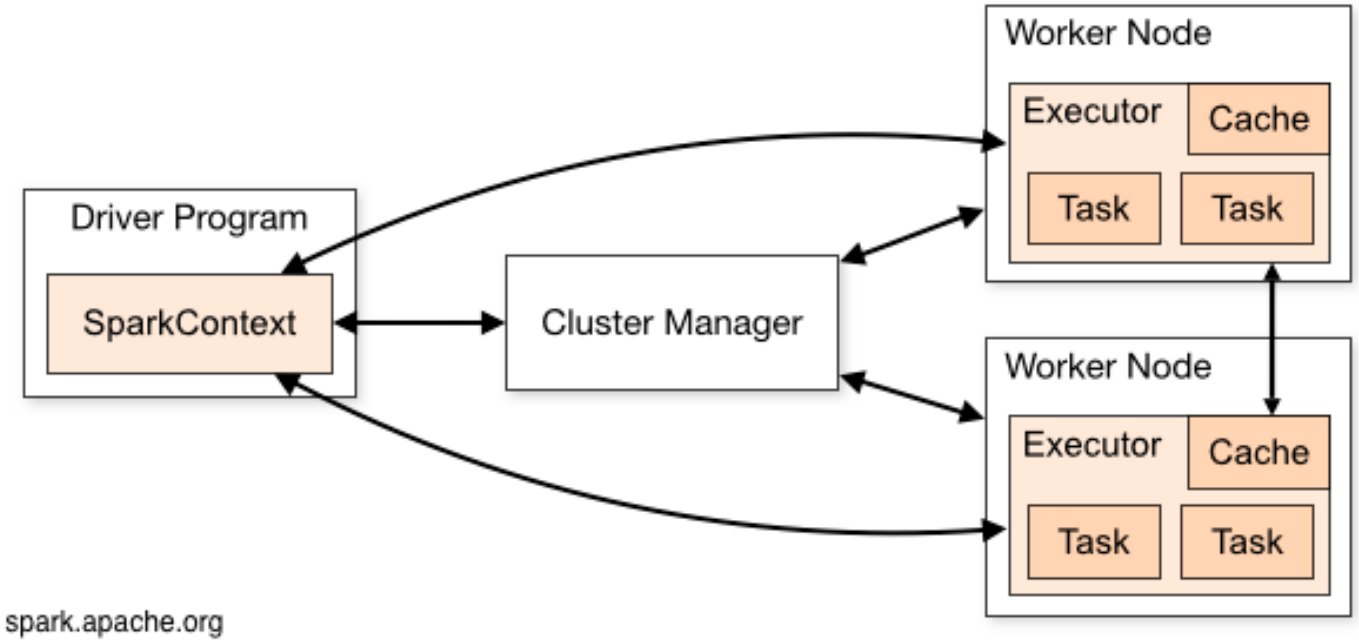
A real world application

Conclusions

WHY *SPARKING* PANDAS

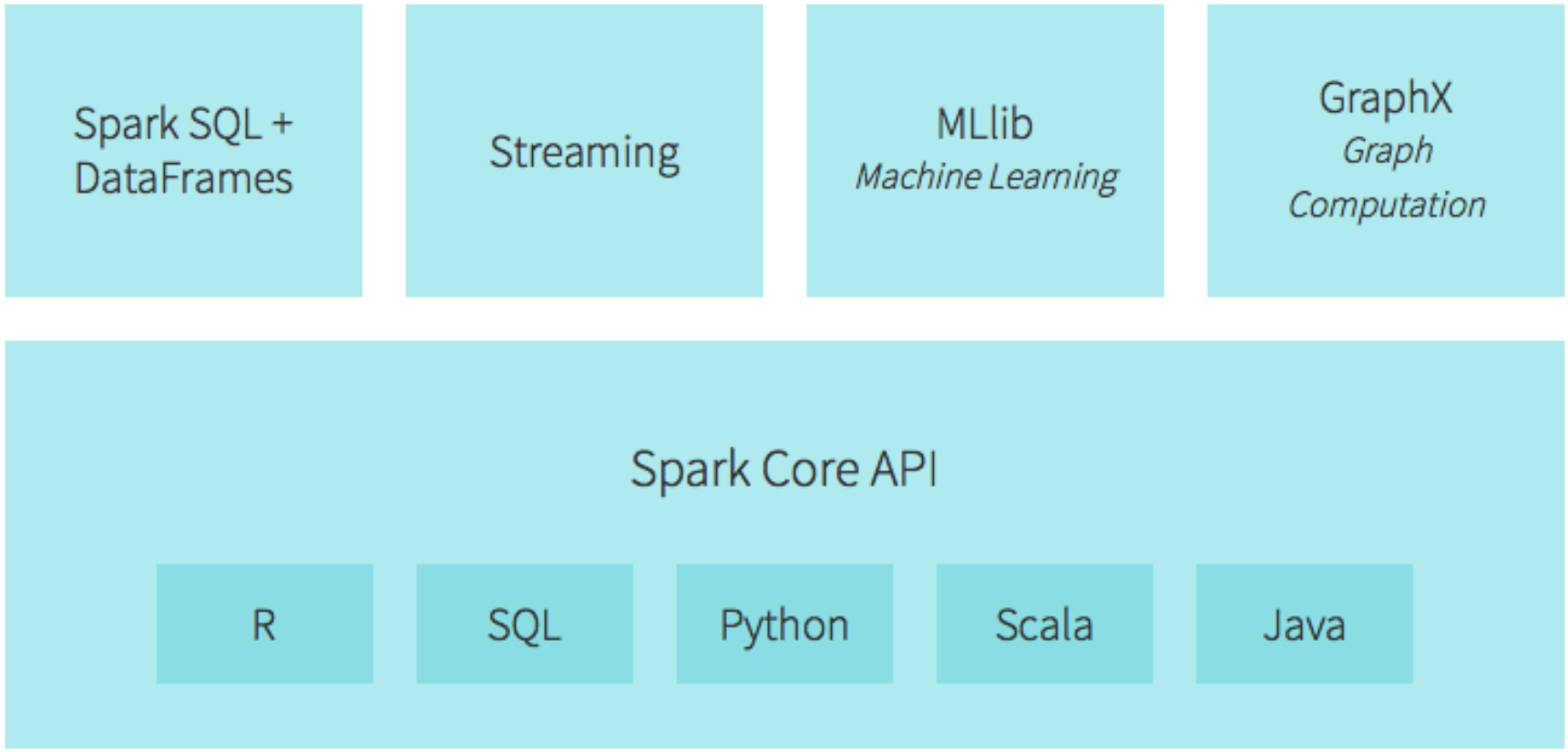
What if your data don't fit into memory?

APACHE SPARK: THE COMPONENTS



spark.apache.org

APACHE SPARK: THE ARCHITECTURE



databricks.com

FUNCTIONAL DATA PROCESSING PIPELINES

High order functions

Immutable data

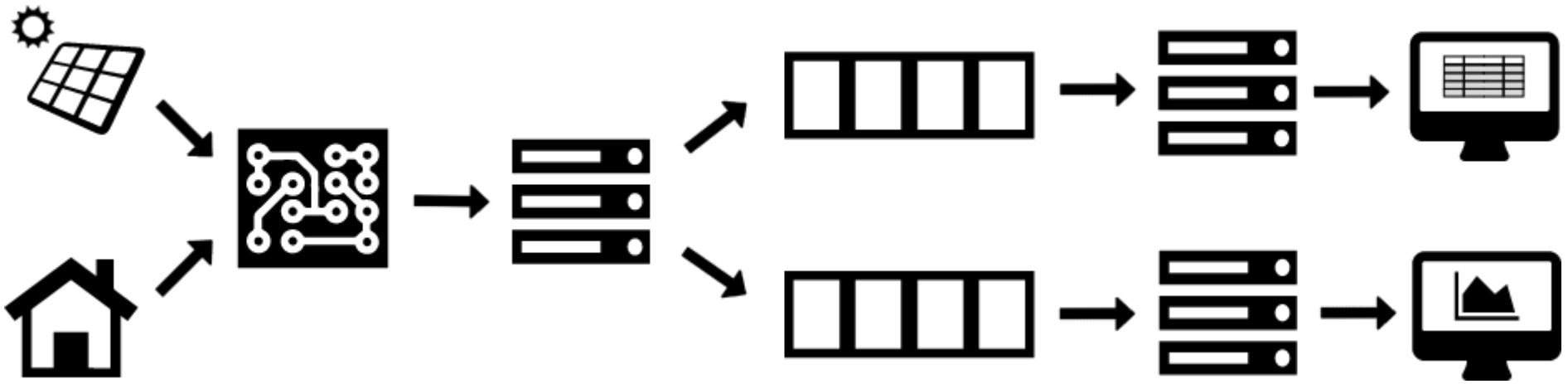
Lazy evaluation

THE EXPERIMENT

The scenario

Containerized application

THE SCENARIO

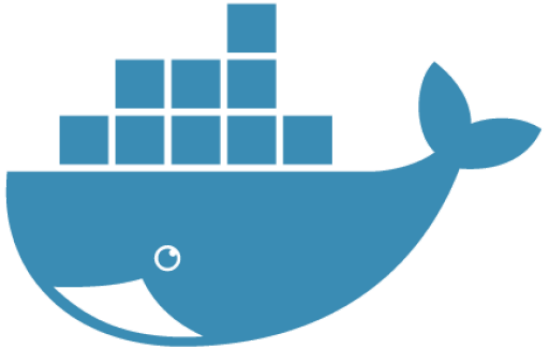


CONTAINERIZED APPLICATION

Containerized components

Constrained memory nodes

docker-composed ecosystem



HANDS ON CODE

Apache Spark basics

Linear regression

Near real time processing with Apache Kafka

CONCLUSIONS

Complex structure

Worth the effort with a lot of data

Worker nodes should be distribueted

Keep exploring :)

QUESTIONS?

brunifrancesco

https://github.com/brunifrancesco/docker-spark