INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

IA Frameworks

Introduction & Spark

Philippe Besse & Brendan Guillouet

November 4th, 2019

Institut National des Sciences Appliquées de Toulouse

Table of contents

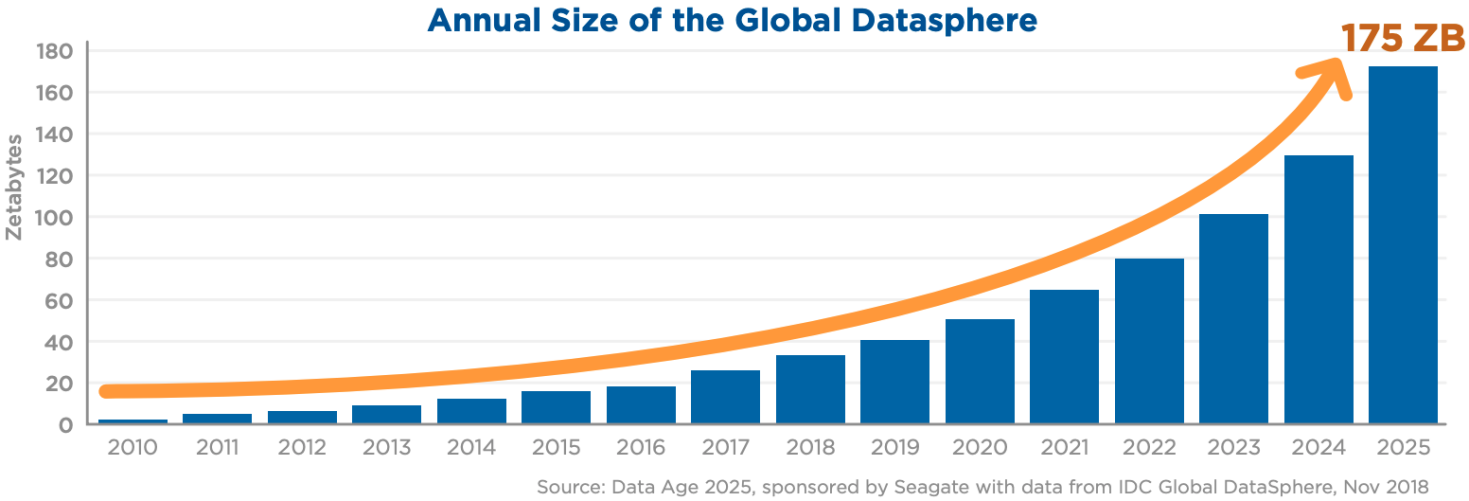
Introduction

Introduction

Definition & Context

Big Data and its 3 Vs :

- **Volume** : Exponential growth.
  - According to IDC : from 33 zeta bytes in 2018 to 175 in 2025. IDC [2018]

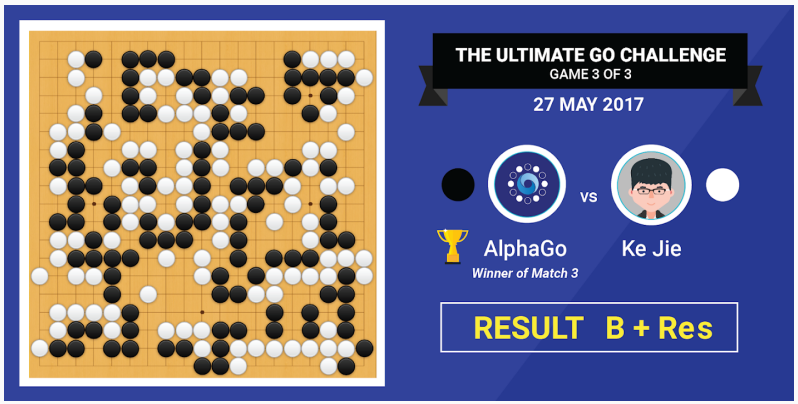


- **Variety** : numeric, signal, images, time series, graph, text etc...
- **Velocity** : real time decision.
- (**Valorisation**).

IA & Big Data

Now Artificial Intelligence (again)

- Big data,
- computation power availability : GPU, cloud computing.
- 'new' algorithms : Deep Learning, Reinforcement learning,
- that produce impressive results, (Alpha Go, Atari Game defeated, write poetry etc..).

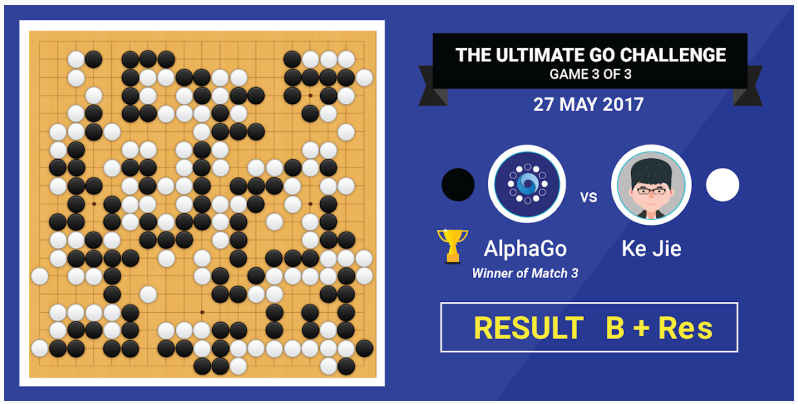


- .. and benefits some good marketing.

IA & Big Data

Now Artificial Intelligence (again)

- Big data,
- computation power availability : GPU, cloud computing.
- 'new' algorithms : Deep Learning, Reinforcement learning,
- that produce impressive results, (Alpha Go, Atari Game defeated, write poetry etc..).



- .. and benefits some good marketing.

The definition is still very unclear, but IA is still NOT here.

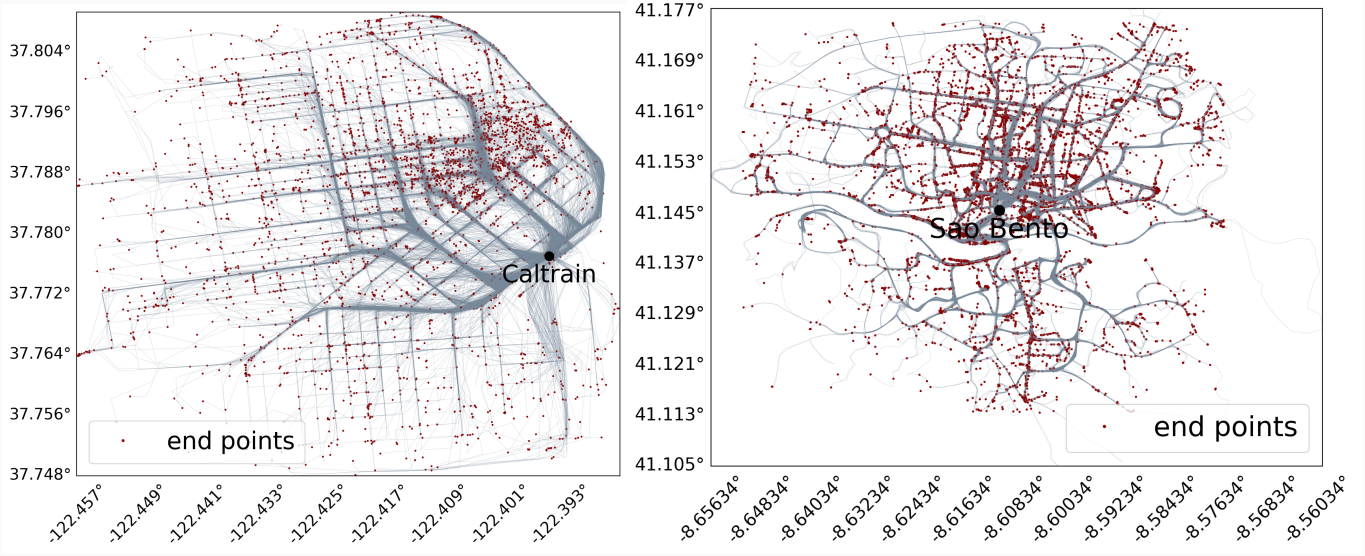
A wide range of application fields...

..at different levels of advancement.

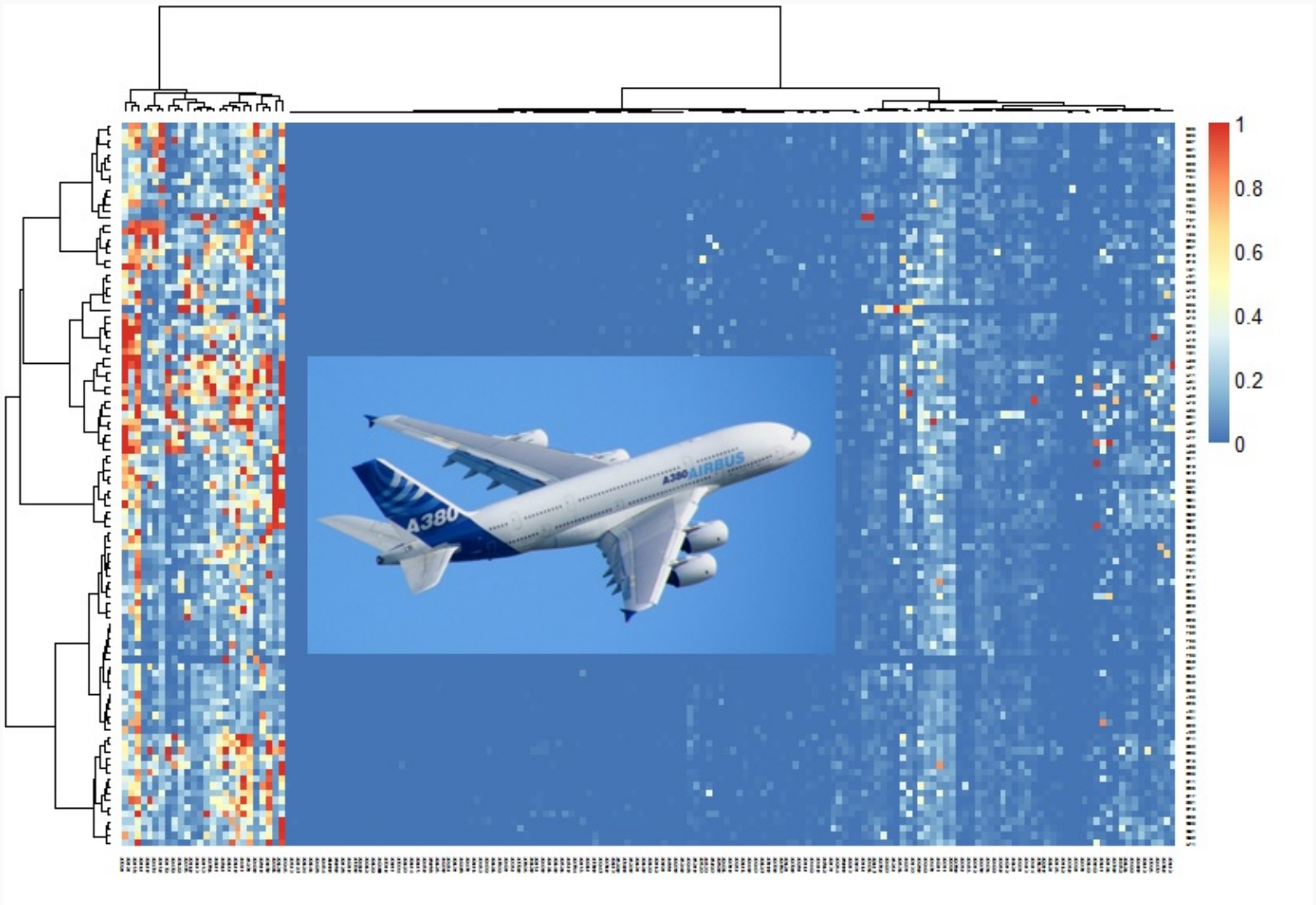
- **E-commerce** : Recommendation system, advertising, targeting.
- **Text translation** : DeepL , Google Translate.
- **Image & Video** : Classification, videosurveillance etc..
- **Game** : Chess, Go, Atari Game, StarCraft.

- **Public** : administrations, e-health (*open data*)
- **Industry** : 4.0 industry, Autonomous vehicle.
- **Meteorology** : Hybridation of learning and physical model
- **Certification**.
- **Internet of things**.

Variety : Trajectory clustering from GPS data.



Analysis of flight incident messages



2019 internship of INSA student

- Detection of field (agriculture) and batiment (military) from satellite images. *SSII & Startup*
- Anonymization of name on text documents. *SSII*
- Fall detection of elder people. *PME*
- Autonomous drone flight. *SSII*
- Rain prediction in Ivory Coast. *Startup*
- Detection of diabetic retinopathy from x-ray photo. *Startup-Norway*
- Estimation of fish results. *Research Center*

Job Specification

Task

- Managing data Wharehousing
- Ensure Data accessability & consumption via API.
- Continuously monitoring and testing the system to ensure optimized performance.
- Extraction Transformation & Load
- Cleaning and organizing raw data.
- Programming skills.
- Using descriptive statistics to get a big-picture view of their data.
- Analyzing interesting trends found in the data.
- Creating visualizations and dashboards to help the company interpret and make decisions with the data.
- Presenting the results of a technical analysis to business clients or internal teams.
- Evaluating statistical models to determine the validity of analyses.
- Using machine learning to build better predictive algorithms.
- Testing and continuously improving the accuracy of machine learning models.
- Continuously Perform scientific Watch.
- Building data visualizations to summarize the conclusion of an advanced analysis.

Job Specification

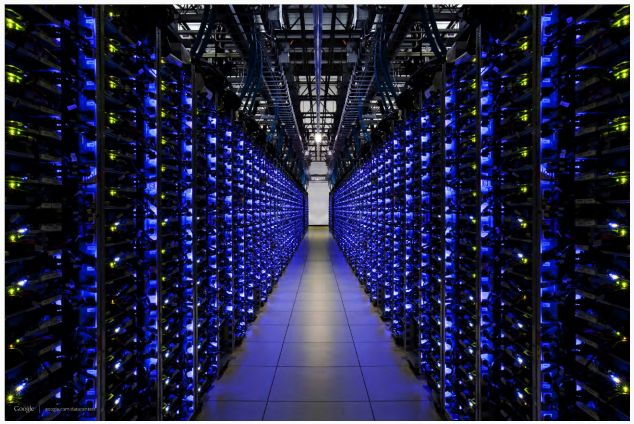
Task	Data Engineer	Data Analyst	Data Scientist
• Managing data Wharehousing	✔✔	✗	✗
• Ensure Data accessability & consumption via API.	✔✔	✗	✗
• Continuously monitoring and testing the system to ensure optimized performance.	✔✔	✗	✗
• Extraction Transformation & Load	✔✔	✔	✔
• Cleaning and organizing raw data.	✔	✔	✔
• Programming skills.	✔	✔	✔
• Using descriptive statistics to get a big-picture view of their data.	✗	✔✔	✔
• Analyzing interesting trends found in the data.	✗	✔✔	✔
• Creating visualizations and dashboards to help the company interpret and make decisions with the data.	✗	✔✔	✔
• Presenting the results of a technical analysis to business clients or internal teams.	✗	✔✔	✔
• Evaluating statistical models to determine the validity of analyses.	✗	✔	✔✔
• Using machine learning to build better predictive algorithms.	✗	✔	✔✔
• Testing and continuously improving the accuracy of machine learning models.	✗	✗	✔✔
• Continuously Perform scientific Watch.	✗	✗	✔✔
• Building data visualizations to summarize the conclusion of an advanced analysis.	✗	✗	✔✔

Job Specification

Task	Data Engineer	Data Analyst	You're Here ↓ Data Scientist
• Managing data Wharehousing	✔✔	✗	✗
• Ensure Data accessability & consumption via API.	✔✔	✗	✗
• Continuously monitoring and testing the system to ensure optimized performance.	✔✔	✗	✗
• Extraction Transformation & Load	✔✔	✔	✔
• Cleaning and organizing raw data.	✔	✔	✔
• Programming skills.	✔	✔	✔
• Using descriptive statistics to get a big-picture view of their data.	✗	✔✔	✔
• Analyzing interesting trends found in the data.	✗	✔✔	✔
• Creating visualizations and dashboards to help the company interpret and make decisions with the data.	✗	✔✔	✔
• Presenting the results of a technical analysis to business clients or internal teams.	✗	✔✔	✔
• Evaluating statistical models to determine the validity of analyses.	✗	✔	✔✔
• Using machine learning to build better predictive algorithms.	✗	✔	✔✔
• Testing and continuously improving the accuracy of machine learning models.	✗	✗	✔✔
• Continuously Perform scientific Watch.	✗	✗	✔✔
• Building data visualizations to summarize the conclusion of an advanced analysis.	✗	✗	✔✔

What do we need to start ?

- Local or cloud computing ? *(GCP, AWS, AZURE)*
- Size of disk ? Size of RAM ? CPU or GPU ? Hom many of its ?
- From where and how did you get the data : local/bucket ?
- How are they stored ? Csv, Data base ? Which one ?
- Distributed architecture ? *(Hadoop, Spark)*
- In which format do you expect the results ?
- How it is automate ?



DataFarming

A new economic model

Situation :

- Algorithm are available and free (python library, github etc..)
- Most used Software (Tensorflow, Spark,...) are under free licences (GNU, MIT, Apache)
- **Margin** on hardware are not really valuable

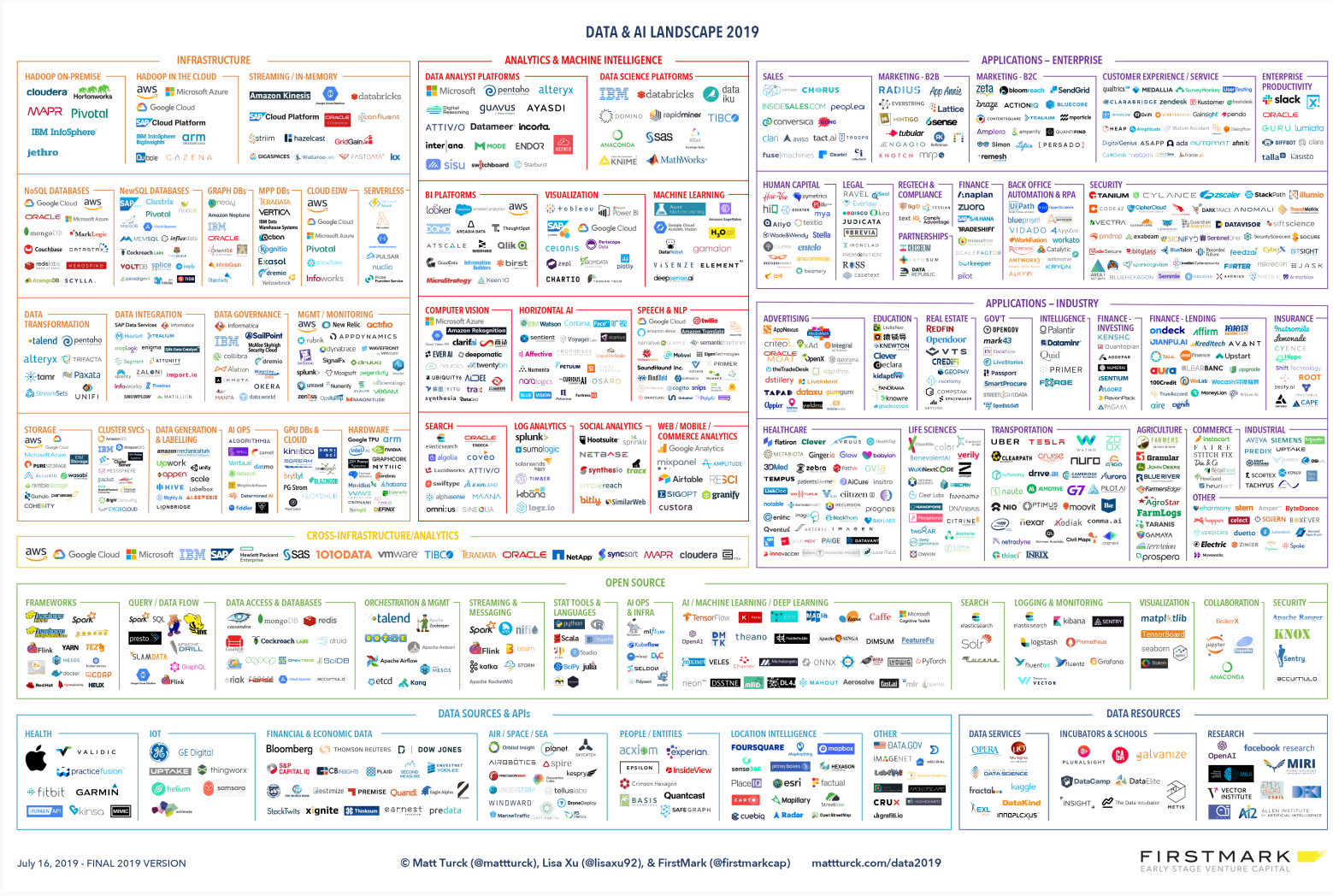
New Strategy : **Sell services**

- Enthought (**Canopy**), Continuum analytics (**Anaconda**), Horthon Works, Cloudera (**Hadoop, Spark...**), Databricks (**Spark**), Oxdata (**H20**), Revolution Analytics (**RHadoop**) – Microsoft

New market : **Cloud computing** (Pick-And-Shovel Play)

- Actors : Amazon Web Service Platform, Microsoft Azure, Google Cloud Computing **aaS**, Software **aaS**, Service **aaS**...
- A huge variety of services, Storage, Instances, ML Pipeline, DataFlow, etc....

Big Data & AI Landscape (Turck, 2019)



New (?) Data Science

As a **statistician**.

- Deep comprehension of algorithms
- Scientific watch

What's really new for a **Data Scientist**.

- Programming skills.
- Better understanding of all data-pipeline technology.
- Manage bigger data.
- New questions about ethics.

$\Longrightarrow$ Learn to "**self-learn**".

Introduction

Organization & Evaluation

Objectives - Tools

Tools you know

ML Python Libraries

Viz' Python Libraries

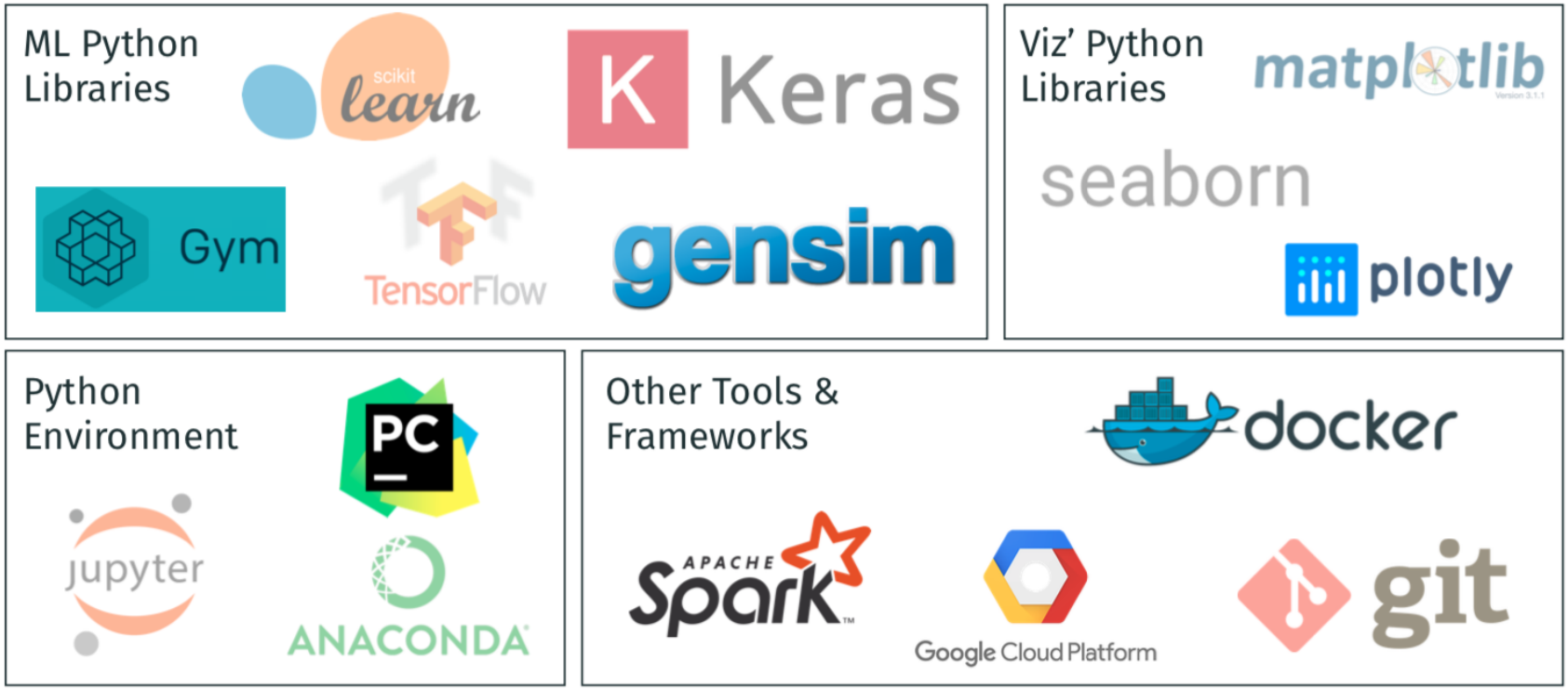
seaborn

Python Environment

Other Tools & Frameworks

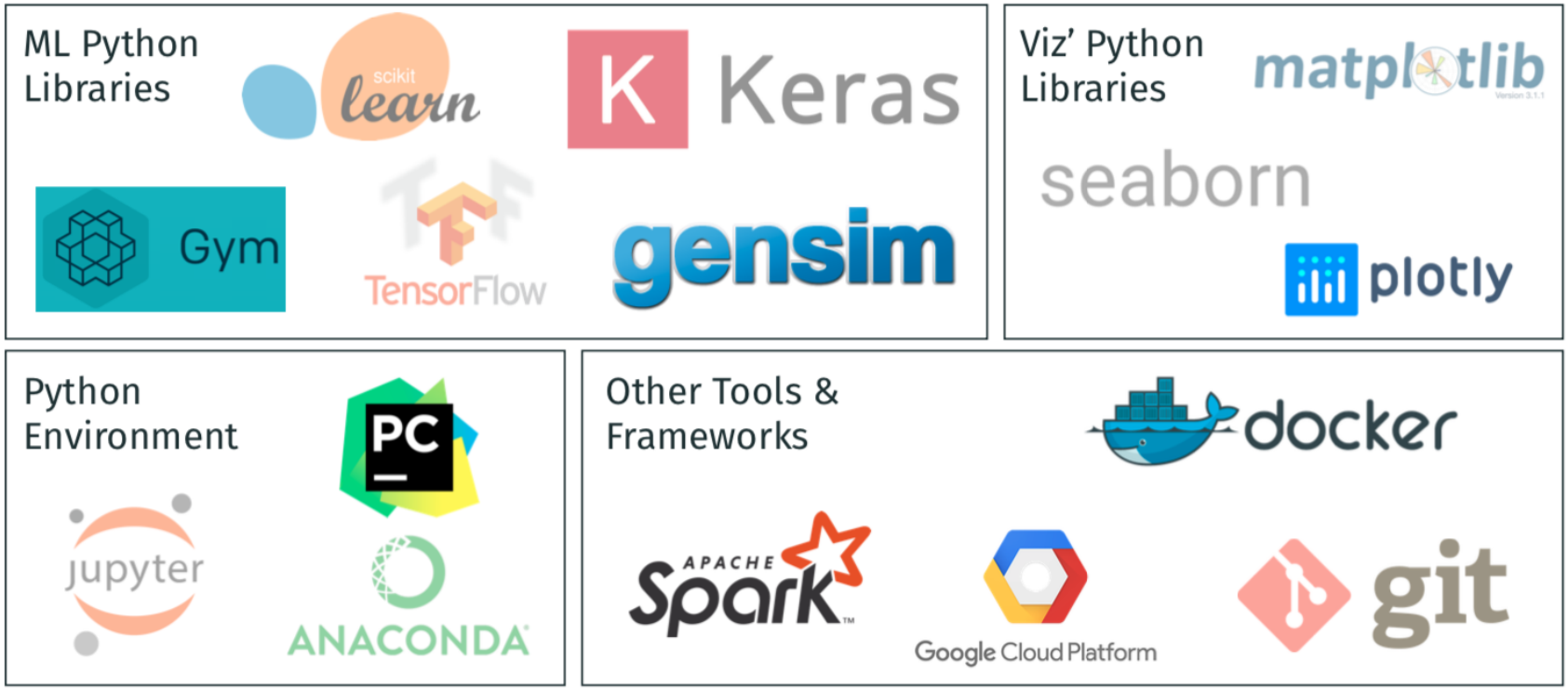
Objectives - Tools

Tools you'll discover



Objectives - Tools

Tools you'll discover



Question :

- What are the best tools to used ? In which circumstances ?
- Is it valuable to use CPU/GPU ? Own laptop/cloud resources ? Distributed architecture ?

Objectives - Data & Algorithm

Data

- Structured Data
- Time Series & Signal
- Image

Algorithm

- Machine Learning : *SVM*, *RF*, *Regression*.
- Deep Learning : *MLP*,*CNN*.

Objectives - Data & Algorithm

Data

- Structured Data
- Time Series & Signal
- Image

- Text (*Cdiscount*)
- Video Game (*PacMan-like*)
- Note Data (*MovieLense*)

Algorithm

- Machine Learning : *SVM, RF, Regression*.
- Deep Learning : *MLP,CNN*.

- Text cleaning : *TFIDF, steming*.
- More Deep Learning : *RNN, Words Embedding*.
- Reinforcement learning : *Policy Gradient / Q-learning*.
- Recommendation System : *NMF*.

Five Labs

1. Python complements. Introduction to Spark.
2. Cloud computing (GCP) and 'containerization' (Docker).
3. **Reinforcement Learning**. (AI Gym).
4. **NLP** on *Cdiscount*. (Gensim)
5. **Recommendation system** on *Movielens* with Spark/MLlib.

Evaluation

$\Longrightarrow$ Act like a Data scientist :

1. Select a new algorithm you want to understand.
2. Select the data and the infrastructure on which you want to make it run.
3. Make it run.
4. Share it in order to maker it easily reproducible for everyone.

$\Longrightarrow$ Two parts evaluation :

- A Github repository where you share your code.
  - How clear is your code ?
  - How easy it is to reproduce the results, following your instruction.
  - **Deadline** : January 11th, 2019.
- Oral Presentation.
  - In-Deep explanation of the chosen algorithm.
  - Choice of the tools /infrastructure used.
  - Difficulty you've met.
  - **Date** : January 14th, 2019.

Evaluation

$\Longrightarrow$ Some propositions of subject

- Generative Adversarial Network.
- Object segmentation & Localization
- Siamese Network.
- One-Shot Learning.
- Recurrent Neural Network (Text generation)
- Words Embedding Comparison. (Glove/Fasttestx/Bert/Elmo)
- Words Embedding to handle small learning dataset.
- RL : Actor Critic.
- RL : Application on real data (Pacman).
- Google CLoud : Comparizon of AutoML with other services.

$\Longrightarrow$ You're free to choose the infrastructure you want.

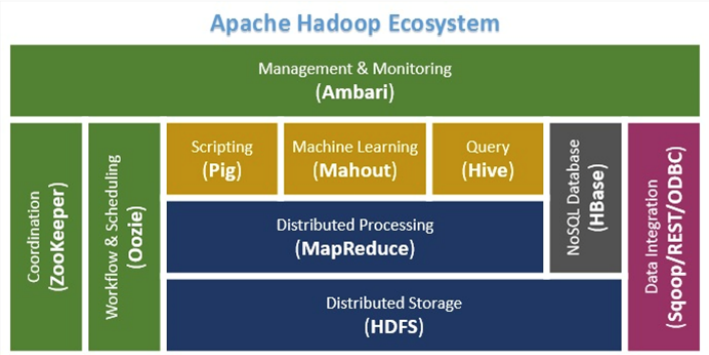
$\Longrightarrow$ Group of 4 to 5.

Hadoop, MapReduce, Spark

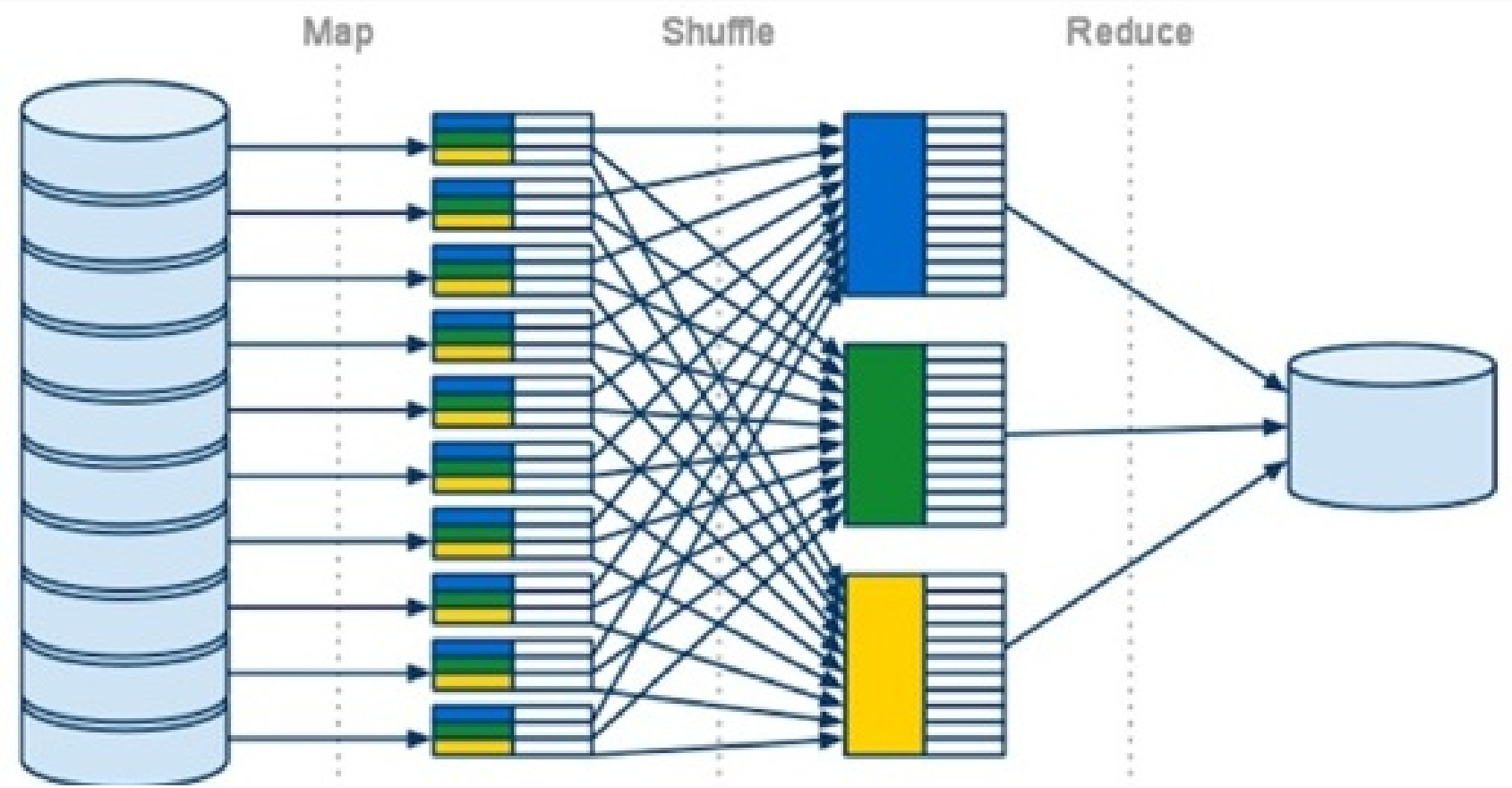
Hadoop, MapReduce

- Environment : *Google* and now **Apache** (2009)
- **Hadoop** *Distributed File System* (HDFS)
- Heterogeneous **distributed** data
- Hardware fault **tolerance**
- **Scalable** (node multiplication)
- **Parallelization** : *Map Reduce*
- strictly **Disk-Based**

- **(Key, Value)** Communication
- **Move** the algorithm, not the data
- *Immutable* **data**
- *Streaming*



Hadoop, MapReduce



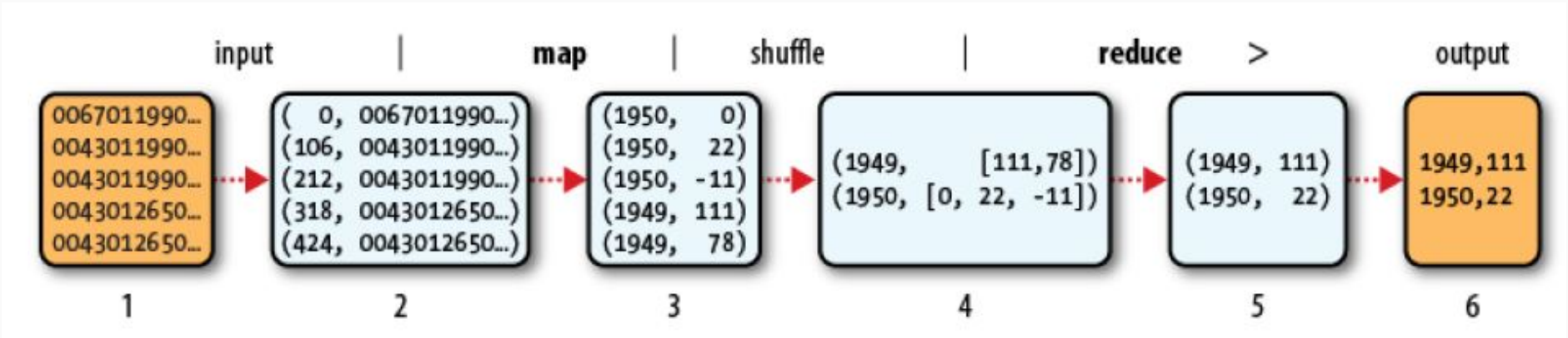
Hadoop Distributed File System (HDFS) & MapReduce

MapReduce

Algorithm :

1. `Input` ⇒ `list (k1, v1)`
2. **`Map()`** ⇒ `list (k2, v2)`
3. ***Shuffle***, *Combine* ⇒ `(k2, list(v2))`
   - Implicit
   - Identical key to same reduce node.
4. **`Reduce()`** ⇒ `list (k3, v3)`

Exemple : Max per year.



Mobile center classification (≈ *k-means*)

- Definition of an euclidean **distance**
- **Forgy**'s algorithm (1965)
  - **Initialization** of k centers
  - *MapReduce*'s steps **Itération**
    - **Map** : Affectation of each individual `(value)` to it's closest center `(key)`.
    - **Reduce** : Compute means of individuals with same keys.
    - **Update** of the k centers
- **Problem** : disk access to each iteration
- **Solution** **Spark**'s *Resilient Distributed Dataset*, (Zaharia *et al.*, 2012)

Resilient Distributed Dataset

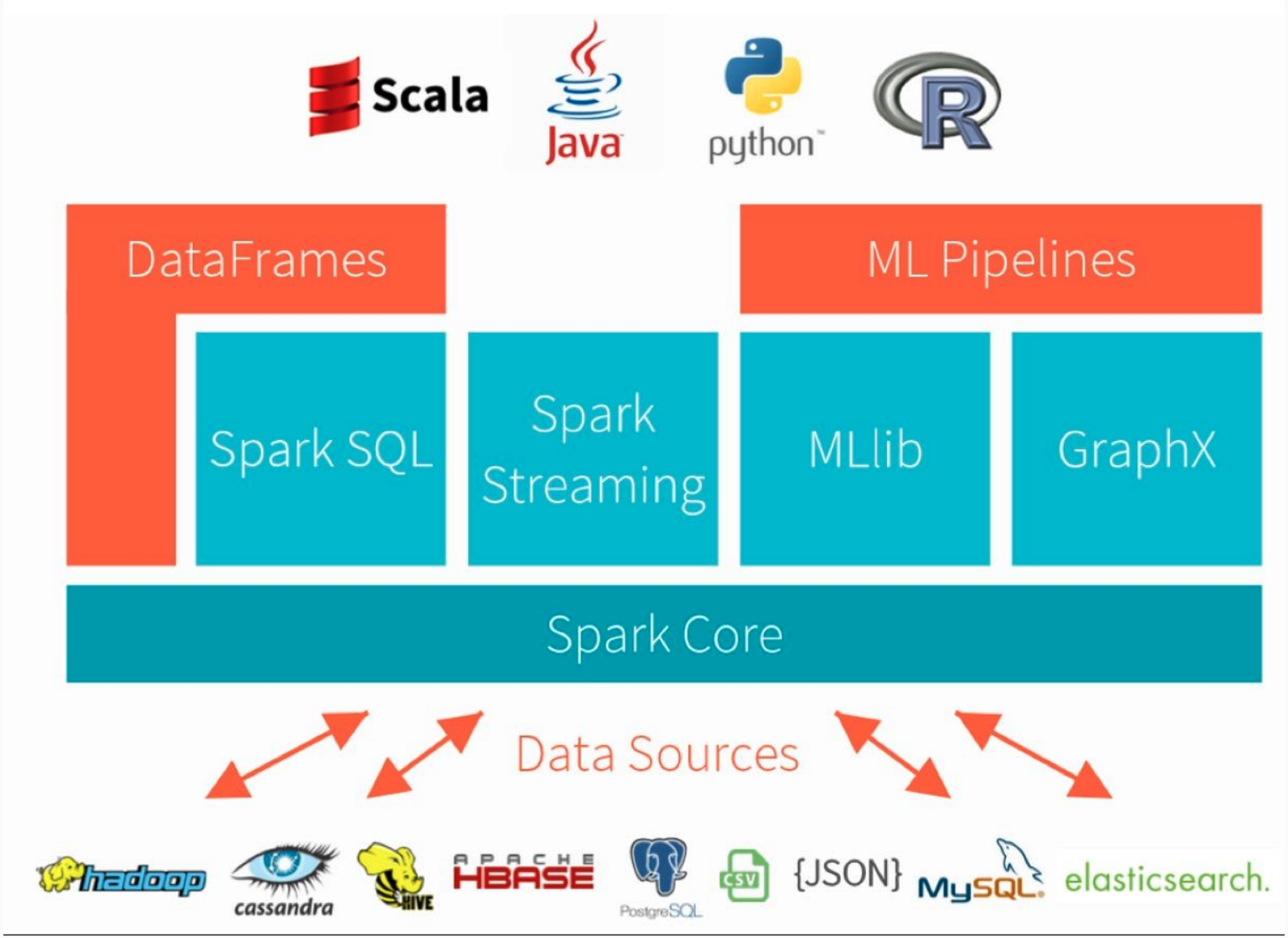
Spark's **RDD** supports in-memory processing computation.

- The state of the memory is stored as an object across the jobs.
- Data sharing in memory is 10 to 100 times faster than in disk.

Characteristics

- RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.
- Handle two types of operation :
  - **Transformation** allows to create a new RDD from an existing one.
  - **Action** applies computation on the RDD and return a value.
- **Transformation** are lazy, while **Action** aren't.

Spark's technology and its ecosystem



RDD vs DataFrame

Limit of **RDD** performances :

- Do not handle structure data.
  - Make not difference between and RDD of list and a RDD of dictionnaries.
- Do not handle easy access of part of the data (by columns name) etc..
- Lead to non-optimized performances.

DataFrame

- Based on Spark SQL API.
- Benefit all RDDs performances.
- Structured data (Columns, named etc..)
- Enable optimization and better computation performances.

MLlib VS SparkML Library

Spark 2.4

- MLlib
  - RDD Based library.
  - Available but deprecated.
- SparkML
  - DataFrame Based library.
  - Official library (Not as furnished as MLlib).
  - Pipeline object.

Lab

Launch the notebook with pyspark :

```
pyspark_notebook.sh
```

Four notebooks :

1. Python notebooks - *Cal4-PythonProg*.
2. Basic notions & RDD.
3. MLlib library.
4. Spark SQL & DataFrame.
5. SparkML & Pipeline.

NB : It is used here in local mode.

Références

IDC. The digitization of the world. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, 2018.