CATIE
Solutions pour la société numérique

Hadoop - Spark

09 Janvier 2020

Romain QUÉRAUD - Cours INSEEC

r.queraud@catie.fr

RÉGION Nouvelle-Aquitaine

CATIE
Solutions pour la société numérique

Apache
Hadoop

RÉGION Nouvelle-Aquitaine

CATIE
Solutions pour la société numérique

Deux usages

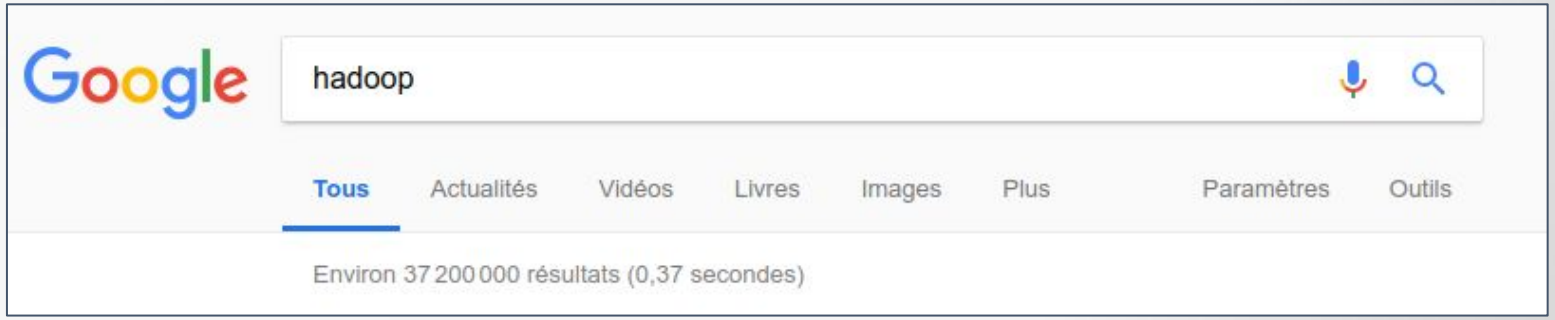
Stockage

- **Sécurisé**
- **Répliqué**
- **Disponible**

Configuration

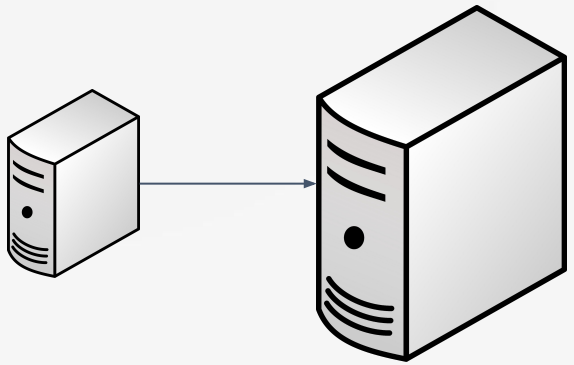
Calcul

- **Possible ?**
- **Rapidité**

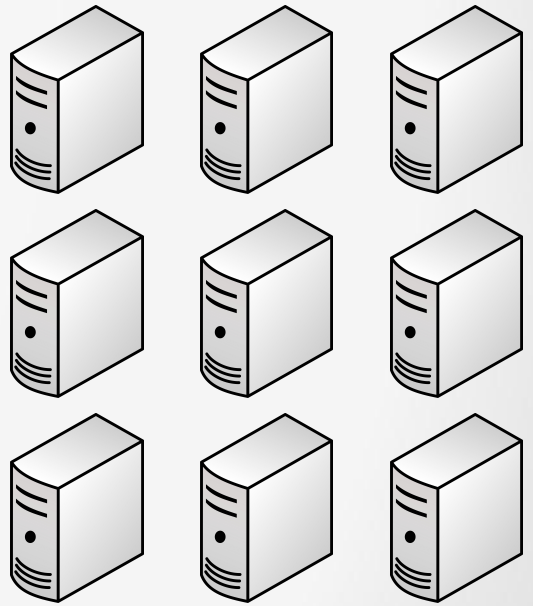


Un besoin de scalabilité

Scalabilité verticale

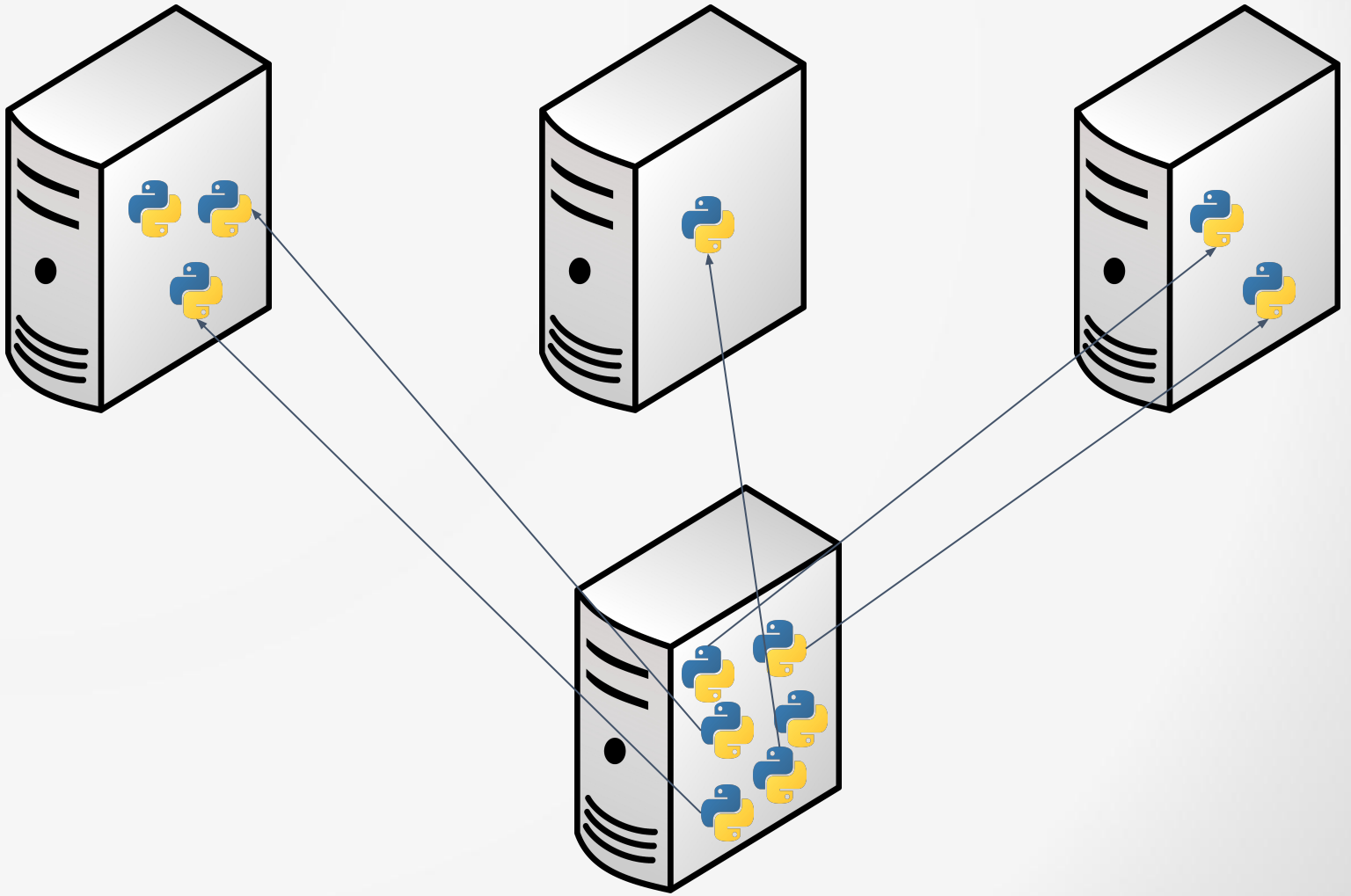


Scalabilité horizontale



CATIE
Solutions pour la société numérique

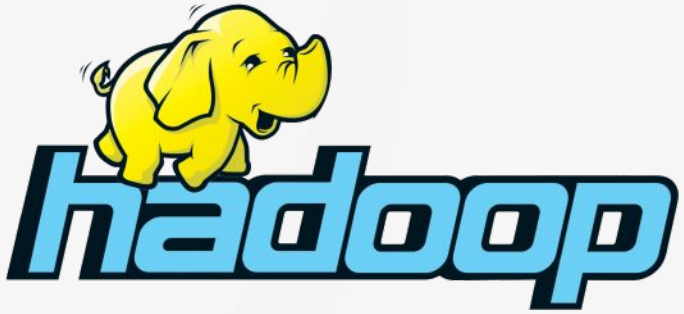
Uniquement distribué ?



CATIE
Solutions pour la société numérique

Écosystème Hadoop

YAHOO!



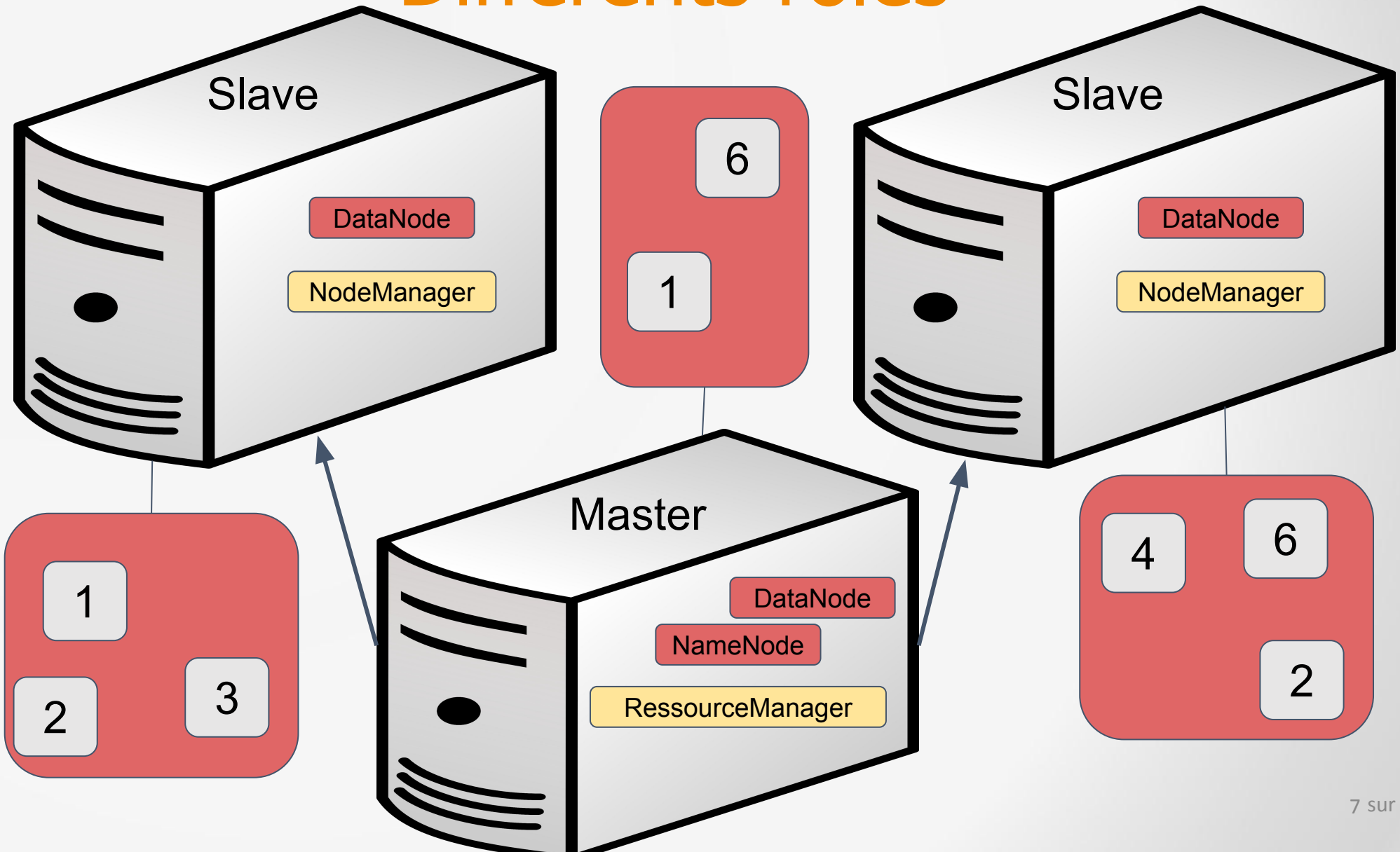
HDFS

YARN

Map - Reduce

CATIE
Solutions pour la société numérique

Différents rôles



CATIE
Solutions pour la société numérique

Outils autour d'hadoop

HDFS

YARN

Map - Reduce

CATIE
Solutions pour la société numérique

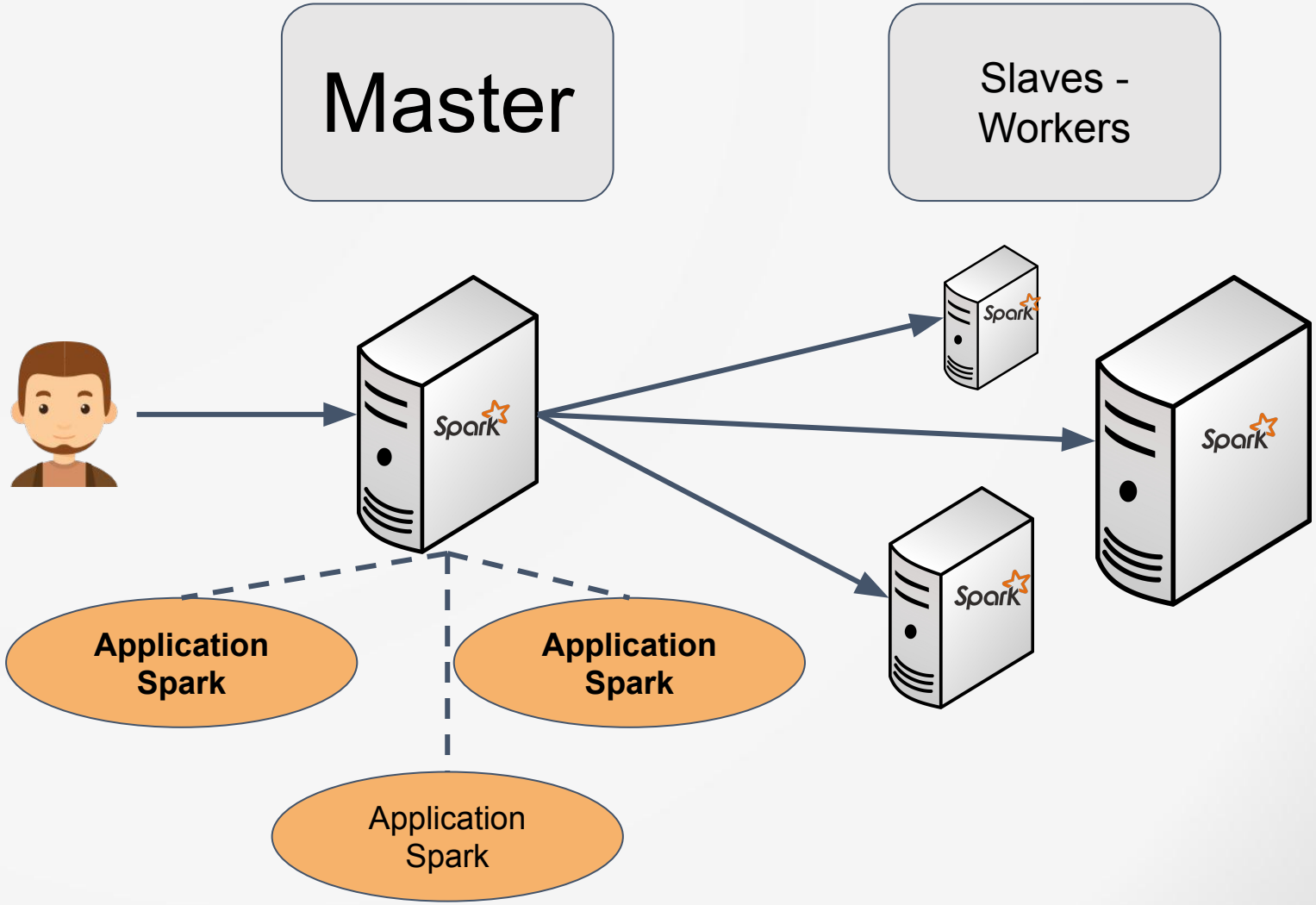
Pour la performance

Spark

RÉGION Nouvelle-Aquitaine

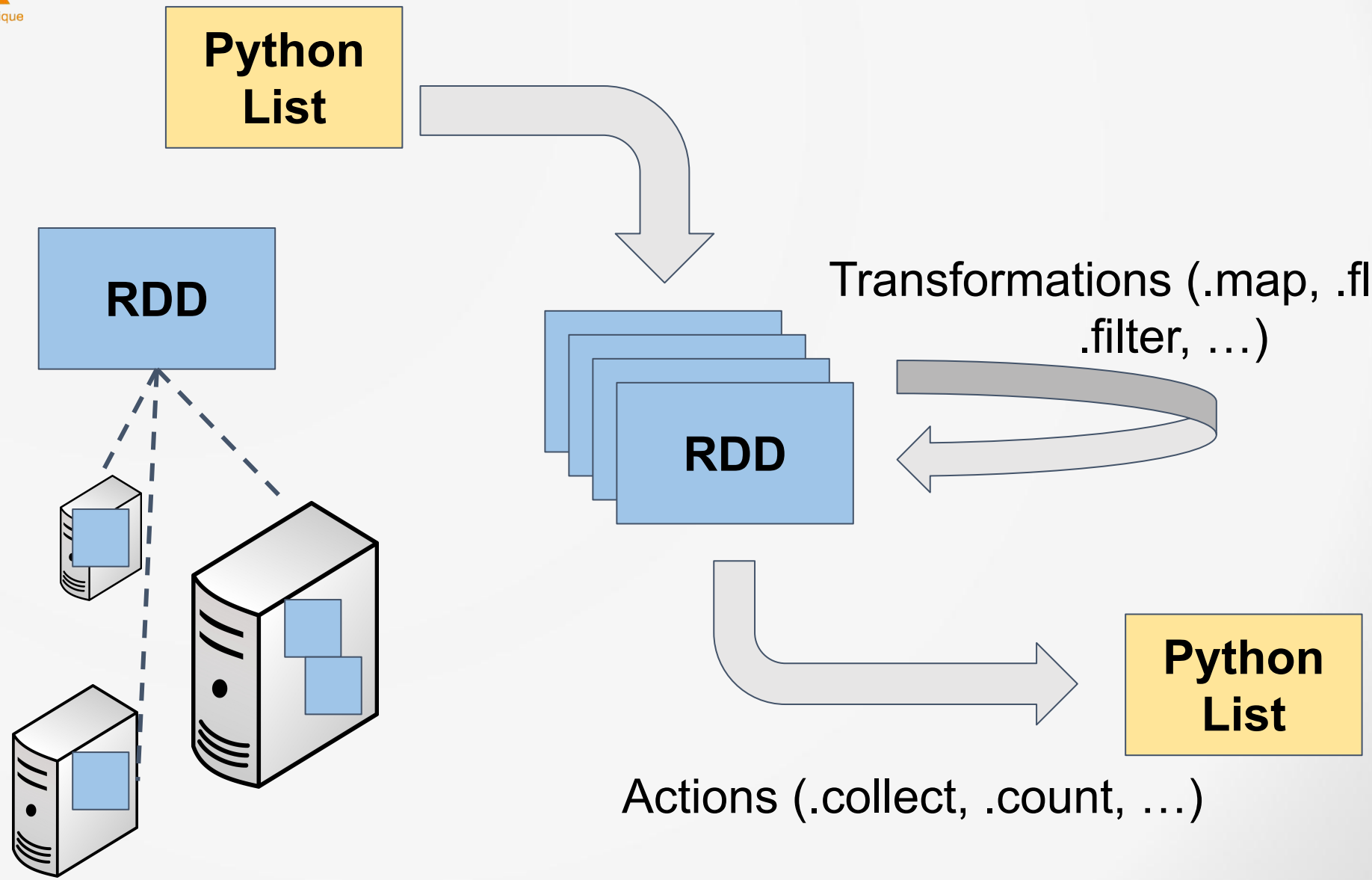
CATIE
Solutions pour la société numérique

Architecture



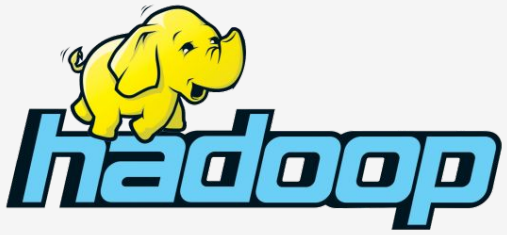
CATIE
Solutions pour la société numérique

Calculs



CATIE
Solutions pour la société numérique

Cluster Spark



+ YARN

Ou

Standalone

Apache Zeppelin
=~ Jupyter Notebook

Exercices RDD

TP disponible : https://github.com/rqueraud/cours_hadoop_3

CATIE
Solutions pour la société numérique

Pour la dernière séance

Cours Hadoop: L'environnement et les services (YARN, Zookeeper, Map-Reduce, …)

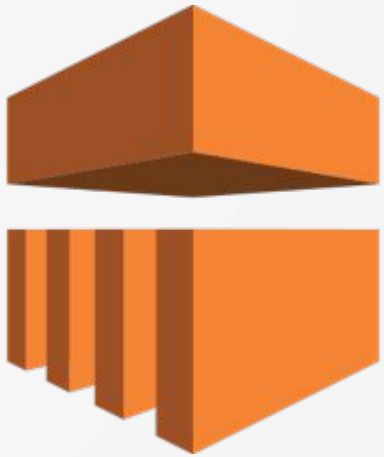
TP noté :

- Se créer un environnement sous AWS:
  - Lancer et paramétrer une EC2 (security-groups, …)
  - Utiliser S3 pour récupérer un fichier
  - …
- Paramétrer son environnement dans une instance
  - SSH et terminal
  - Configurer un cluster Hadoop ?
- Utiliser son environnement
  - Jupyter/Zeppelin Notebook
  - Python, PySpark (RDD, Dataframe ?, …)

CATIE
Solutions pour la société numérique

Intégration dans AWS

- Utiliser les machines EC2 IaaS
  - Ou les machines EC2 PaaS ⇒ Pré-configurées
- Hybrid Data / Calculs
- Remplacer HDFS



amazon
EMR

CATIE
Solutions pour la société numérique

DES QUESTIONS ?

Contact : r.queraud@catie.fr

RÉGION Nouvelle-Aquitaine

Spark Master at spark://ip-172-31-38-28.eu-west-1.compute.internal:7077

2.4.4

URL: spark://ip-172-31-38-28.eu-west-1.compute.internal:7077
Alive Workers: 1
Cores in use: 1 Total, 1 Used
Memory in use: 1024.0 MB Total, 1024.0 MB Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory
worker-20200102144854-172.31.43.214-40585	172.31.43.214:40585	DEAD	1 (0 Used)	1024.0 MB (0.0 B Used)
worker-20200102145120-172.31.43.214-37275	172.31.43.214:37275	ALIVE	1 (1 Used)	1024.0 MB (1024.0 MB Used)

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20200102145933-0000 (kill)	Spark shell	1	1024.0 MB	2020/01/02 14:59:33	ubuntu	RUNNING	30 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

CATIE

Solutions pour la société numérique

localhost:8158

Spark 2.4.4

Spark Worker at 172.31.43.214:37275

ID: worker-20200102145120-172.31.43.214-37275
Master URL: spark://ip-172-31-38-28.eu-west-1.compute.internal:7077
Cores: 1 (1 Used)
Memory: 1024.0 MB (1024.0 MB Used)

Back to Master

Running Executors (1)

ExecutorID	Cores	State	Memory	Job Details	Logs
0	1	RUNNING	1024.0 MB	**ID:** app-20200102145933-0000 **Name:** Spark shell **User:** ubuntu	stdout stderr

RÉGION Nouvelle-Aquitaine