Aix*Marseille université
Initiative d'excellence

La Data Science ambitieuse peut être indolore

Saïda Guezoui et Thomas Lo Coco

Inspirée de l'article « **Ambitious data science can be painless** »

Hatef Monajemi, Riccardo Murri, Eric Jonas, Percy Liang, Victoria Stodden, David Donoho, publié le 22 Juin 2019

Introduction

- **Contexte** : Les programmes développés et utilisés en Data Science demandent une puissance de calcul de plus en plus importante
- **Problématique** : Le matériel traditionnel comme les ordinateurs personnels ou les infrastructures internes d'une université ou entreprise ne sont pas capables d'effectuer ces lourds calculs dans un temps raisonnable.
- **Objectif** : Présenter des services ou des infrastructures du cloud* permettant de mettre en place des architectures capables de traiter des calculs lourds.

Les étapes (et les obstacles) d'une étude ambitieuse en Data Science

Chronologie d'une expérience

Les obstacles principaux d'une étude ambitieuse

Calcul intensif

Comment entraîner un réseau de neurones avec plusieurs milliards de paramètres ?

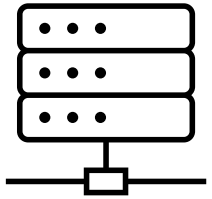
Analyse des données

Comment récupérer efficacement les informations d'exécution (instructions, nombre de cœurs utilisés, erreurs,…) ?

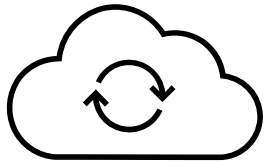
Reproductibilité

Comment faire pour que mon programme soit facilement ré-exécutable dans un environnement différent ?

Avantages et inconvénients



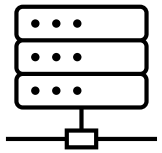
High-Performance Computer



Server Cloud

Avantages	+ Le coût est faible + Pas de transfert de données	+ Accès immédiat aux ressources + Evolutif + Fiable / Robuste + Configurable + Accès au root et à la gestion des logiciels + Le coût dépend seulement de l'utilisation
Inconvénients	- Pas d'accès root - Les ressources ne sont pas extensibles - Les ressources ne sont pas forcément accessibles. - Politique de fonctionnement fixée	- Transfert des données dans un serveur tiers - Les APIs peuvent être difficiles à utiliser

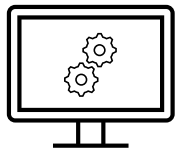
Présentation des services du cloud



IaaS

Infrastructure as a Service

- Application
- Environnement d’exécution
- Middleware
- Système d’exploitation
- Serveur



PaaS

Platform as a Service

- Application
- Environnement d’exécution
- Middleware
- Système d’exploitation
- Serveur

SaaS

Software as a Service

- Application
- Environnement d’exécution
- Middleware
- Système d’exploitation
- Serveur

Gérer par l’utilisateur

Gérer par le fournisseur

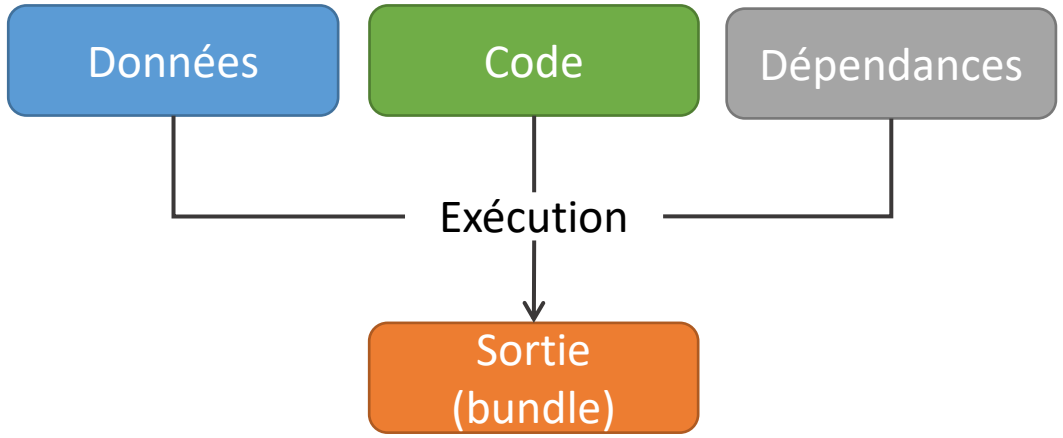
CodaLab

Objectif : Faciliter et sécuriser la reproduction des résultats dans le domaine de la science des données.

Fonctionnement :

CodaLab facilite la reproduction d'un code en **conteneurisant** l'environnement d'exécution.

De plus, le travail est organisé et présenté grâce à un « worksheet » comme un Jupyter Notebook.

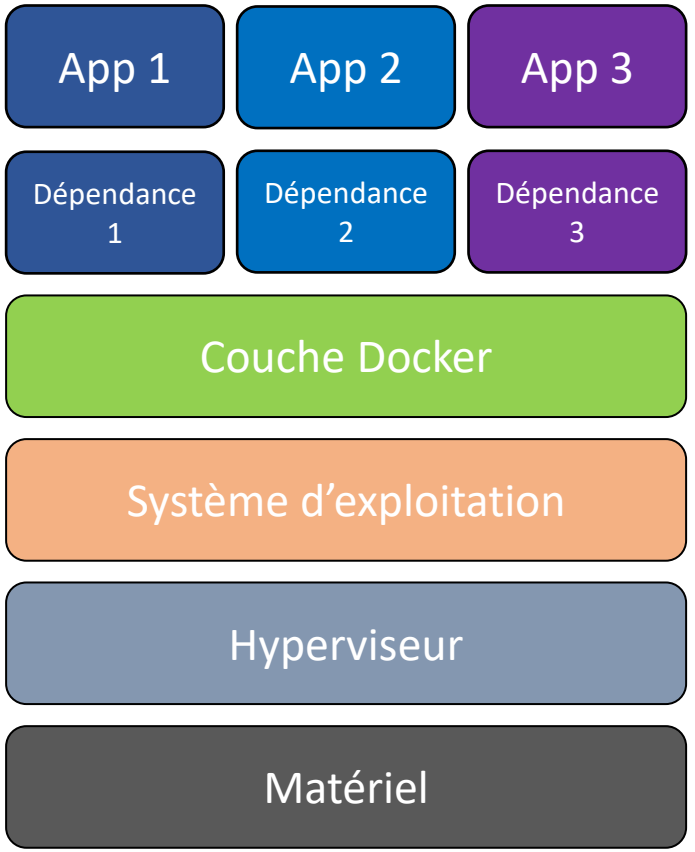


Qu’est-ce que Docker et les conteneurs?

Docker permet de mettre en place et manipuler des **conteneurs** via une API.

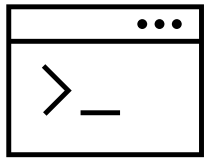
Conteneur

L’objectif d’un conteneur est d’isoler les ressources (nombre de processeurs utilisés, pourcentage de la RAM allouée, etc.) et l’ensemble des dépendances des applications.

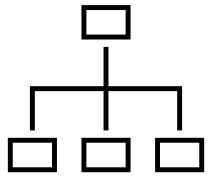


Les différentes couches dans un ordinateur : la couche Docker permet d’isoler les applications et leurs dépendances.

La stack ElastiCluster + ClusterJob

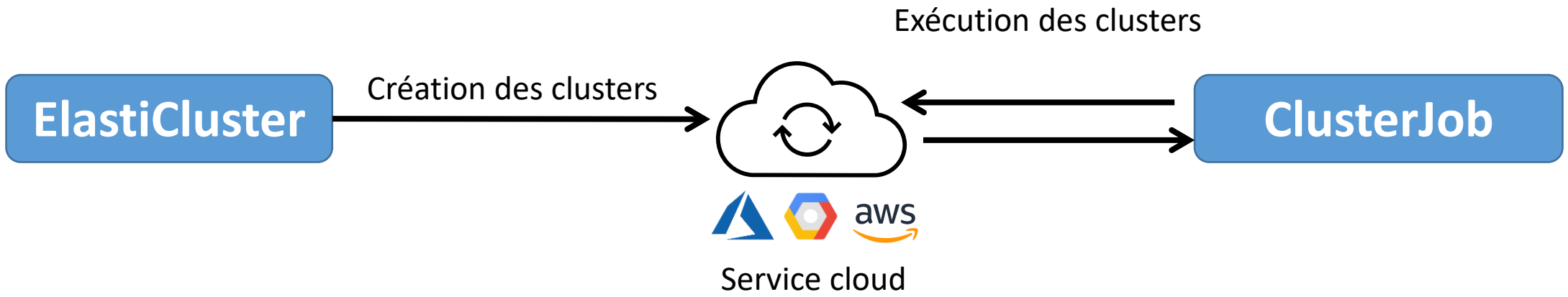


ElastiCluster est un programme en ligne de commande pour **créer**, **gérer** et **configurer** des **clusters de calcul hébergés** sur des **infrastructures.**



Clusterjob (**CJ**) est un **système de gestion des expériences** (**EMS**). Il permet de soumettre des tâches de calcul aux clusters de manière simple et reproductible.
CJ **produit** des **paquets de calcul "reproductibles"** pour les publications académiques.

Fonctionnement ElastiCluster-ClusterJob



1. **ElastiCluster permet de mettre en place des clusters** dans un service cloud comme Microsoft Azure, Google Cloud Platform et Amazon Web Service.

2. Une fois les clusters démarrés, **on utilise ClusterJob pour exécuter le programme** dans les clusters.

Conclusion

- Des infrastructures et des services du cloud comme CodaLab permettent de rendre la réalisation de ces expériences bien moins douloureuses.

- De plus, ces stacks technologiques aident aussi les scientifiques à maintenir leurs programmes et à les diffuser plus facilement.

Questions

Discussion

Que pensez-vous d'envoyer vos données et votre code à une entreprise extérieure comme Google ou AWS ?

Durant vos alternances ou stages avez-vous eu l'occasion de faire des calculs lourds en data science ? Des retours ?

Est-ce que vous préférez utiliser des serveurs locaux ou des services du cloud ?

Est-ce que le domaine de la big data/calcul intensif vous intéresse à l'avenir ?

Pensez-vous que c'est le travail d'un Data Scientist de chercher et de mettre en œuvre des solutions pour traiter ces problèmes ?

Sondage

Machine Learning	Deep Learning / Big Data	Aucun avis
19	10	2