



**Déployez un
modèle dans le
cloud**

Plan

- 01 Problématique
- 02 Présentation des données
- 03 Big data
- 04 Spark
- 05 Architecture big data
- 06 Application de la solution sur le Cloud
- 07 Conclusion

Problématique

«**Fruits**» start-up de l'AgriTech souhaite proposer une solution innovante de récolte des fruits avec des robots cueilleurs intelligents.

mettre en place une application mobile de reconnaissance des fruits.



Fruits!

Mission :

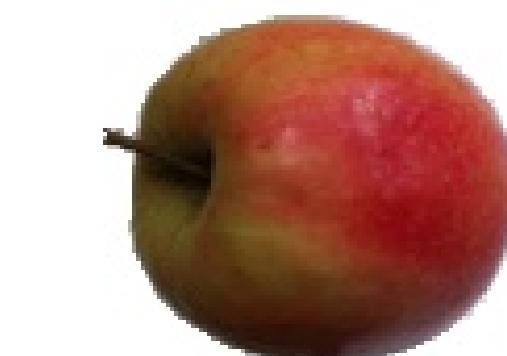
Développer une première architecture Big Data :

- Préprocessing des images et réduction de dimension
- Anticipation du passage à l'échelle

Présentation des données

Jeu de données

- Nombre total d'images : 90483.
- Taille de l'ensemble d'entraînement : 67692 images (un fruit ou un légume par image).
- Taille de l'ensemble de test : 22688 images (un fruit ou un légume par image).
- Le nombre de classes : 131 (fruits et légumes).
- Taille de l'image : 100x100 pixels
- Photos sous tous les angles (rotation 3 axes)



Le Big Data

Données massives



Explosion de la quantité de données ,Partage des données, Recherche Analyse/visualisation des données ,Stockage des données

Traitement des flux de données



Données relationnelles structurées,
non structurées: sms, images, textes..



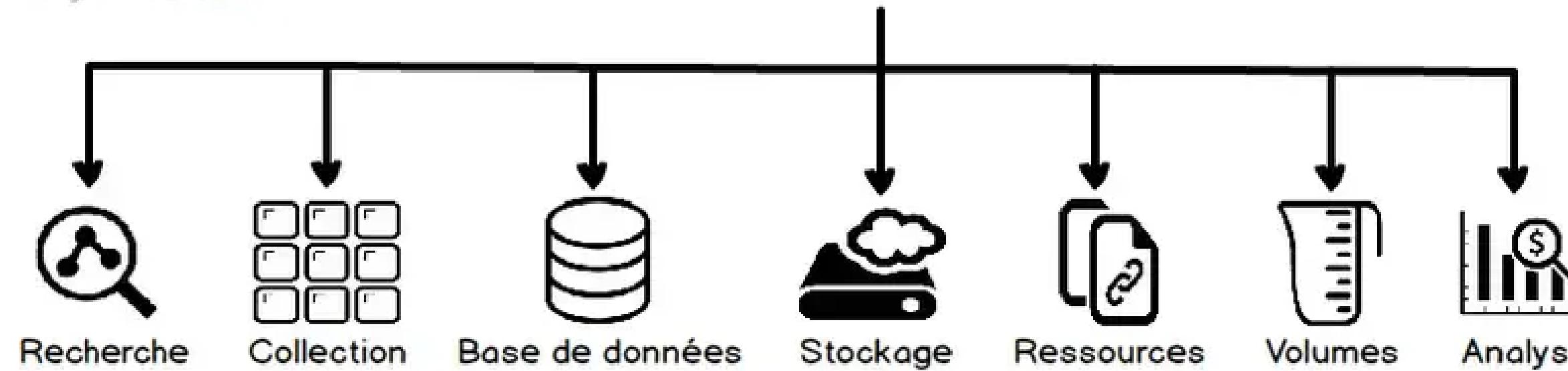
La fréquence de création,
collecte et partage de ces
données.



Volume de données extrêmement important à traiter
augmenter capacités de calcul



Big Data



Le Big Data n'est pas seulement un terme utilisé pour désigner un grand volume de données, mais également un moyen de rassembler, stocker, organiser et analyser de nombreux types de données.

Solutions de stockage big data



Google
Cloud Storage



Microsoft Azure
Blob Storage



Amazon Web Services
S3



Apache
Hadoop

SOLUTION : une Infrastructure distribuée

LE STOCKAGE DISTRIBUÉ

Volume : passage à l'échelle possible

Variété : capacité d'évolution

Vélocité : partitionnement

Résilience :

- redondance
- tolérance aux pannes

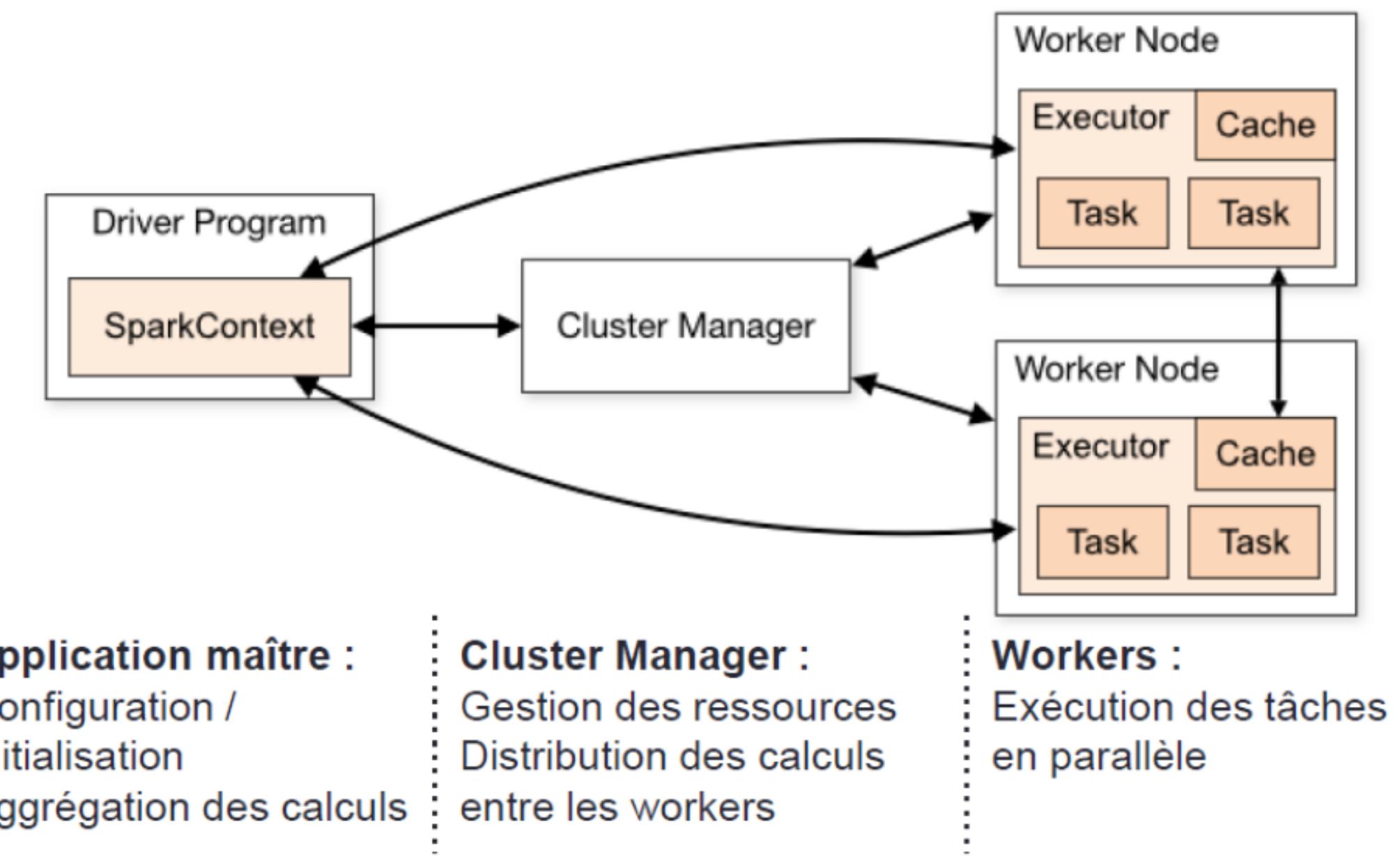
Spark



SOLUTION : une Infrastructure distribuée

CALCULS DISTRIBUÉS

- Diviser les opérations en micro opérations distribuables entre différentes machines, réalisables en parallèle
- Agréger les résultats sur une même machine



Spark

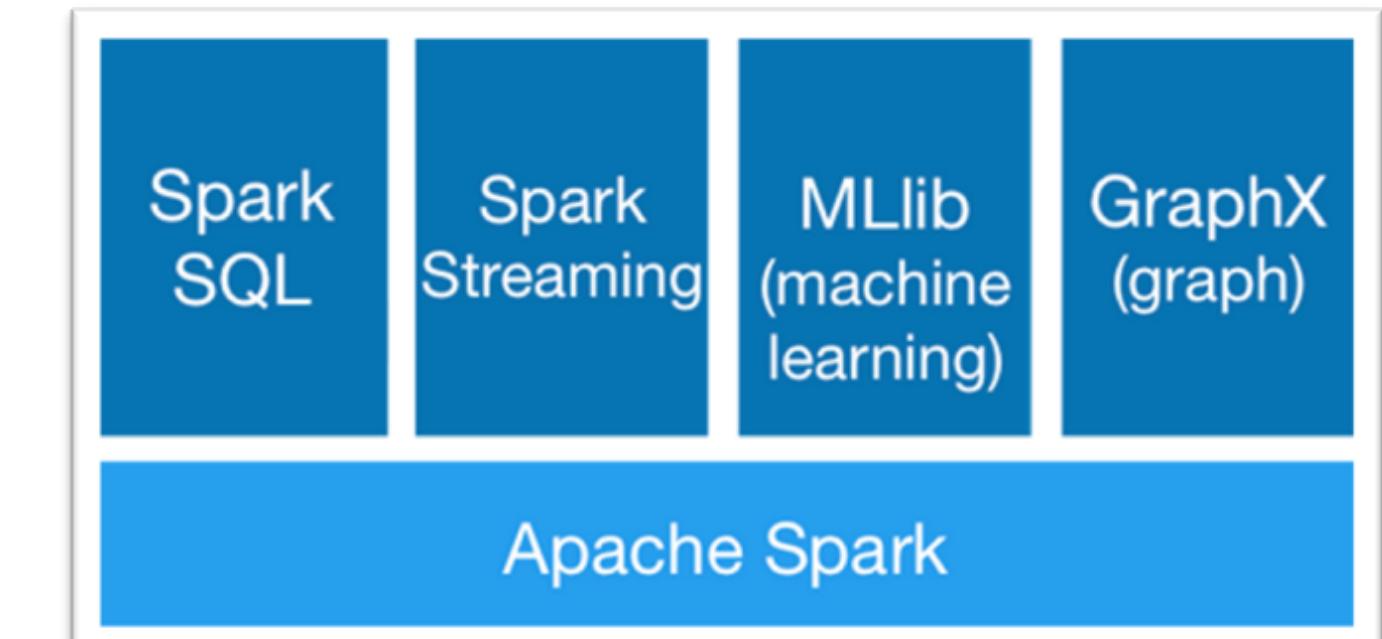
Principe de fonctionnement

Framework open source
de
calcul distribué
in-memory

Plate-forme unificatrice
riche en fonctionnalités

Gère et coordonne
l'exécution de tâches
sur des données à
travers un groupe
d'ordinateurs

Gère les machines du
cluster



Spark

Principe de fonctionnement

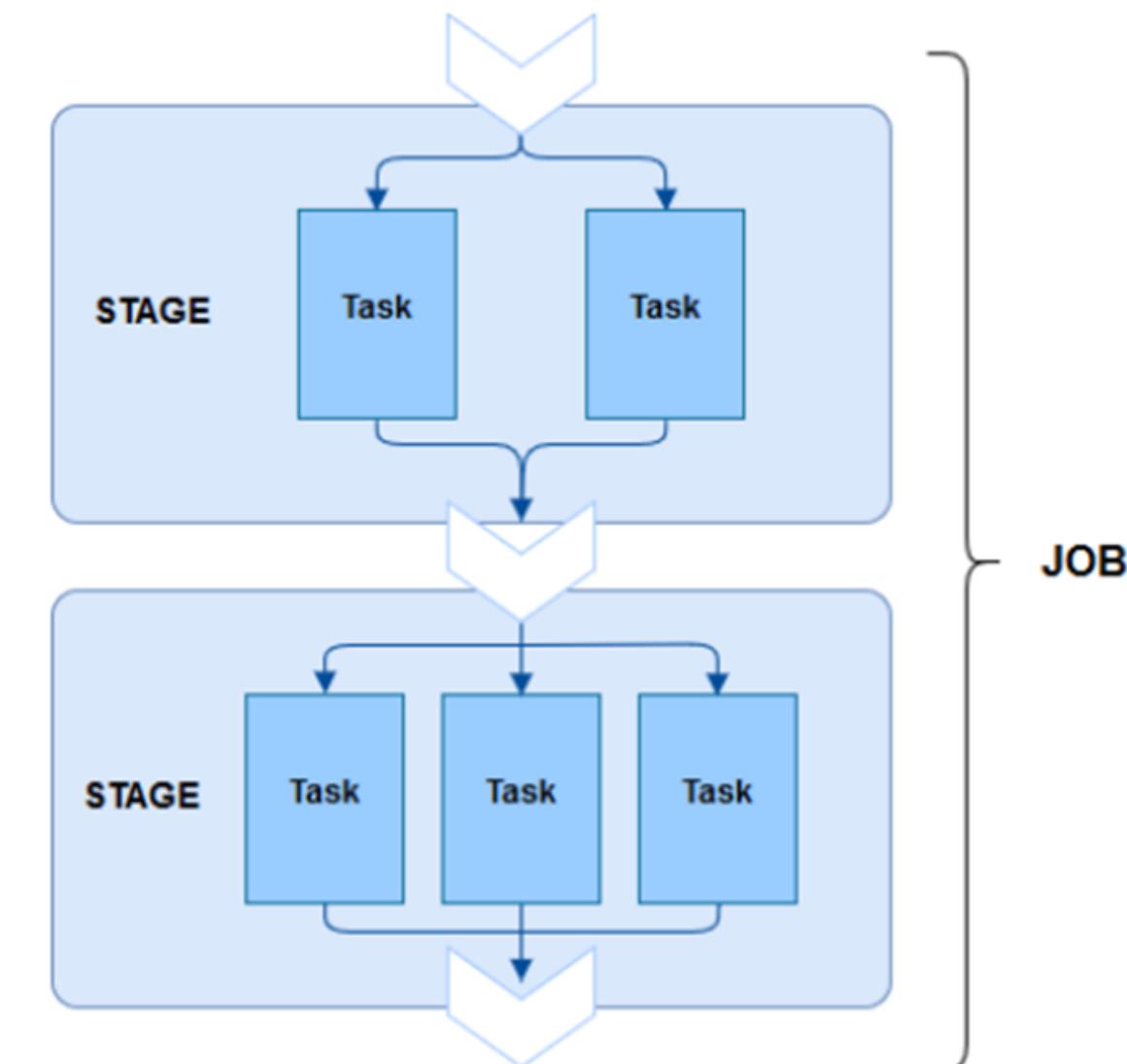
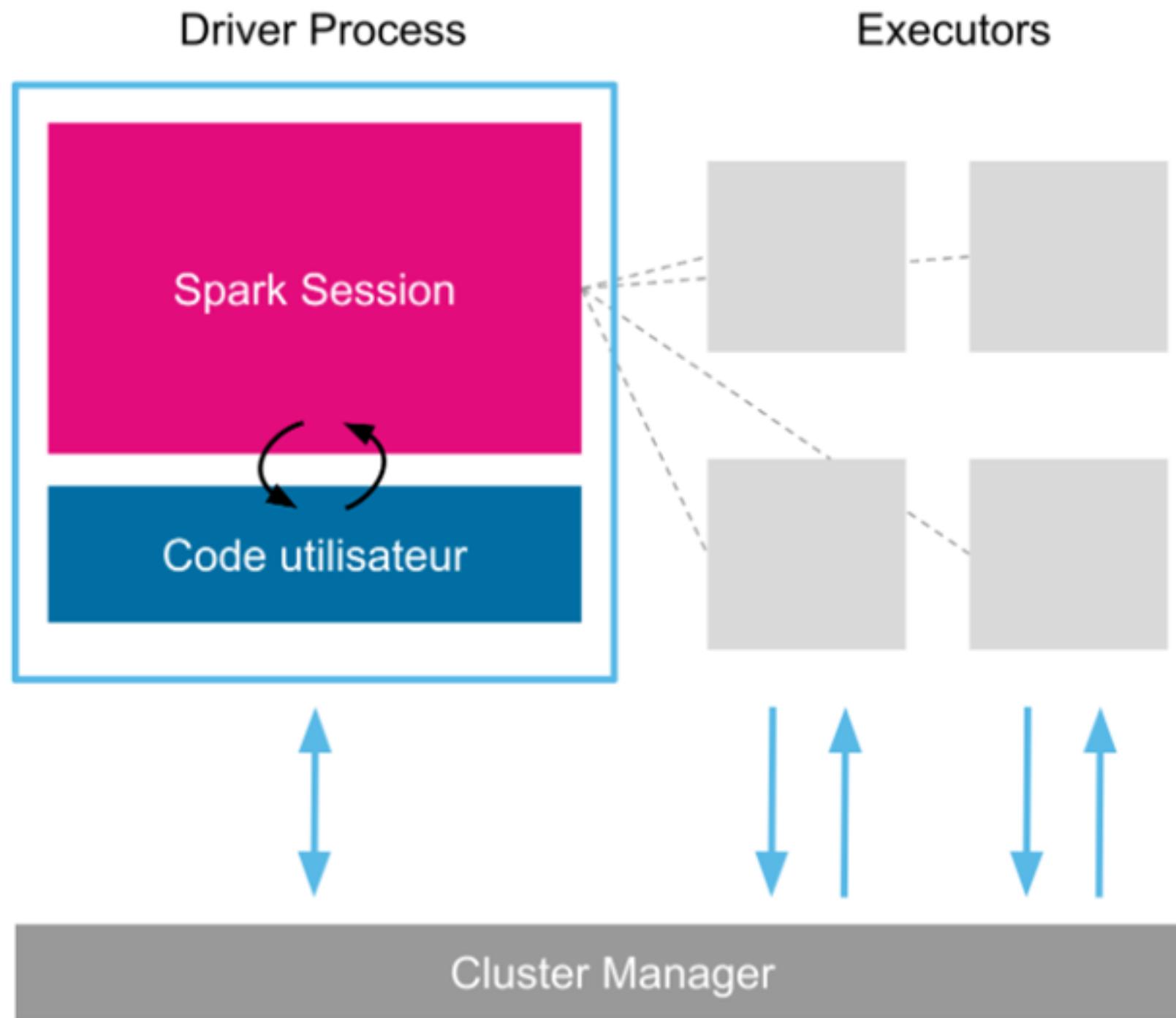
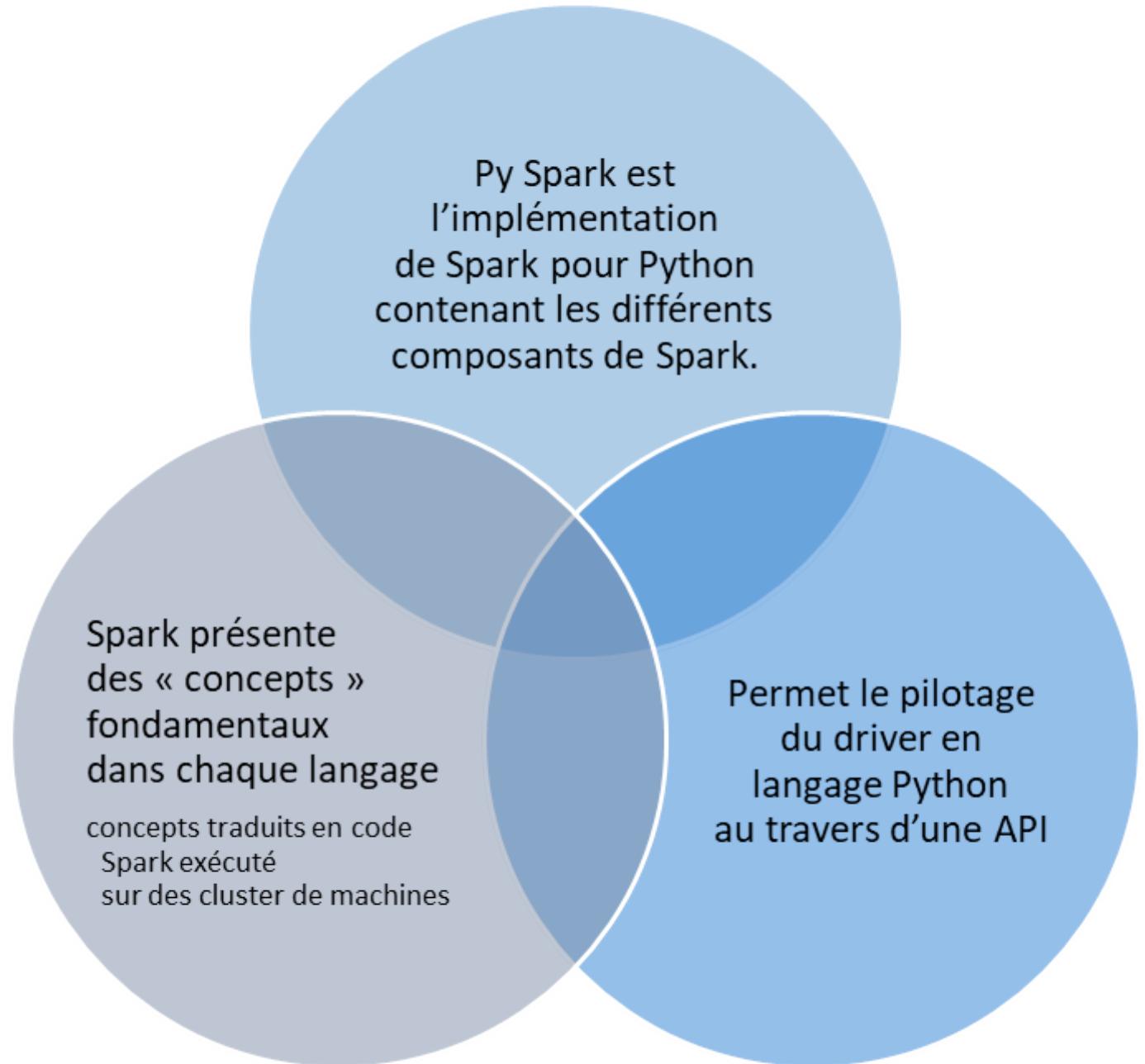


Schéma de l'architecture logique d'exécution

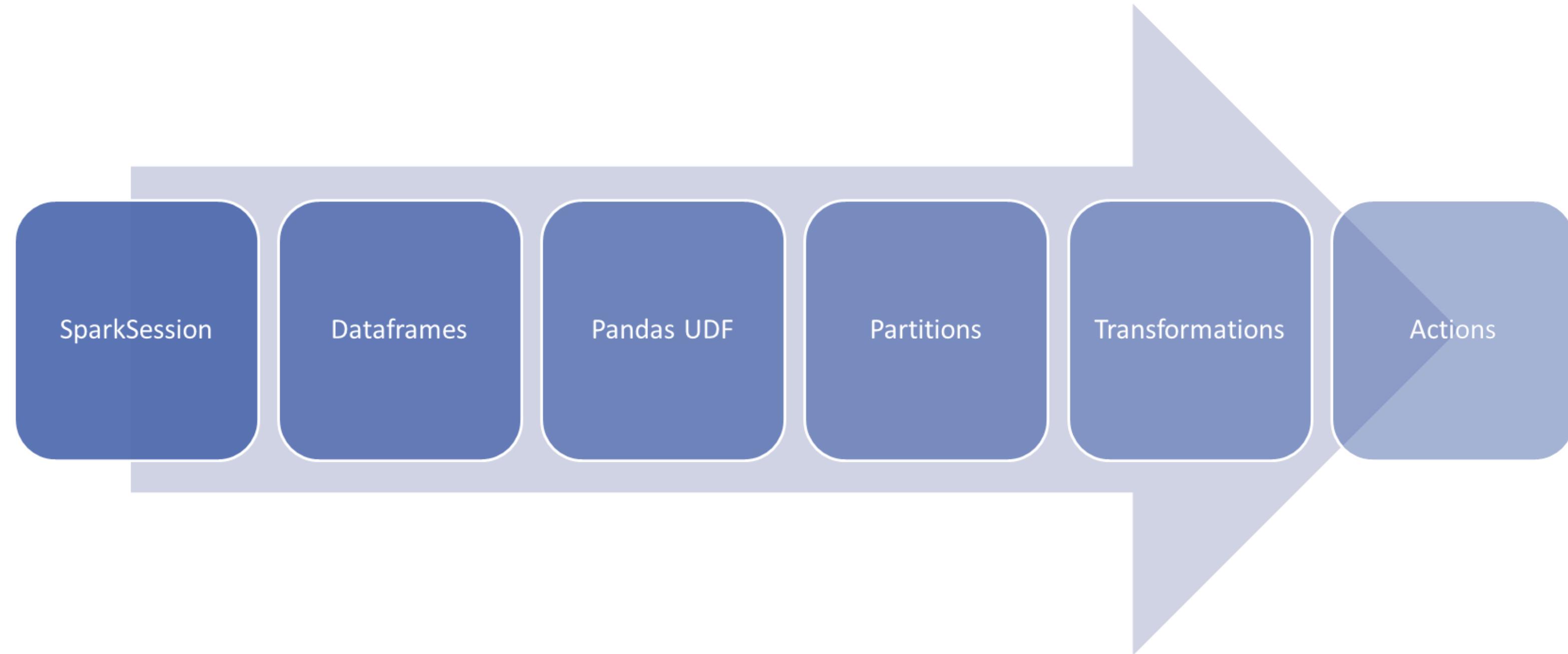
PySpark

Principe de fonctionnement



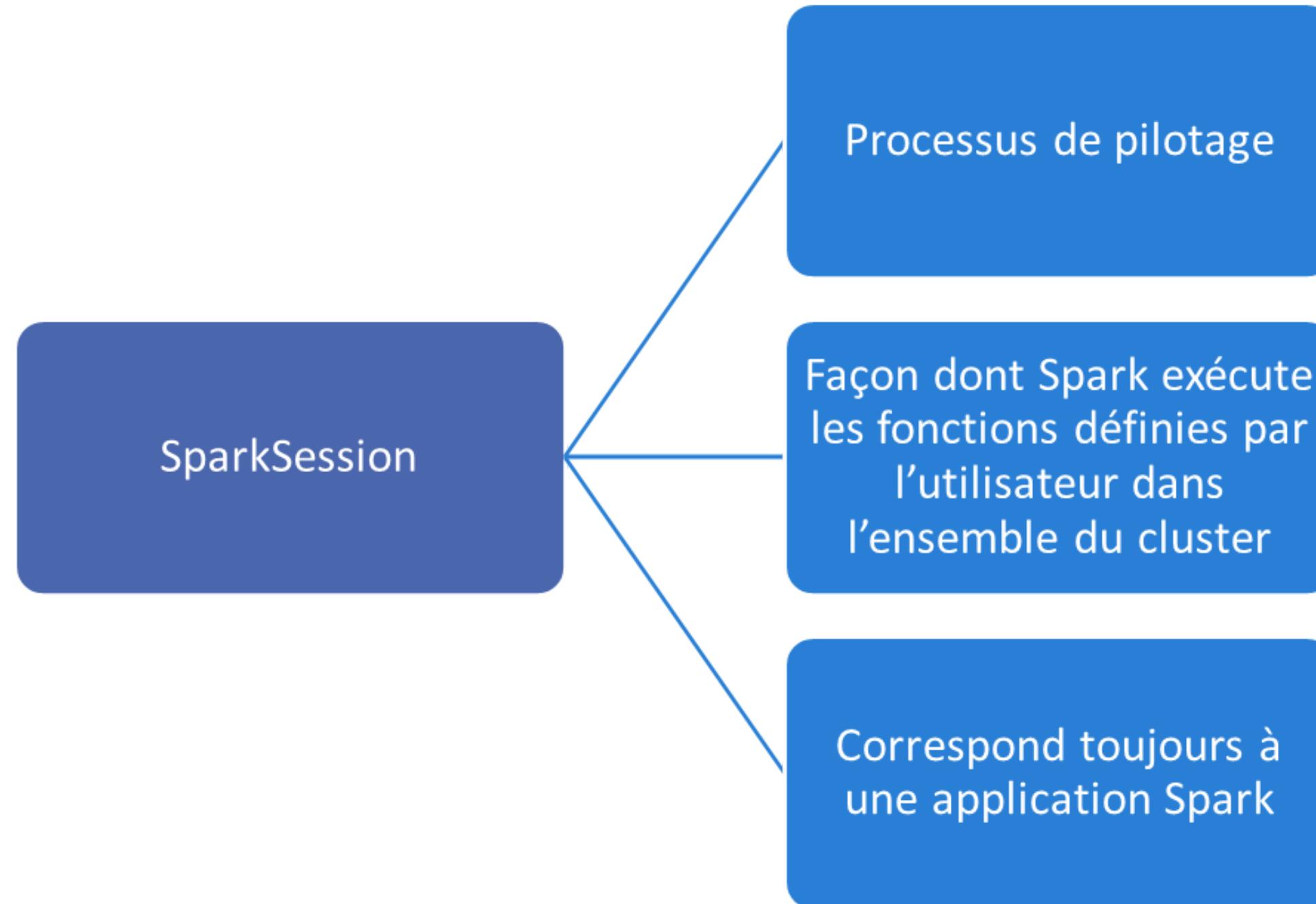
Spark

Concepts principaux



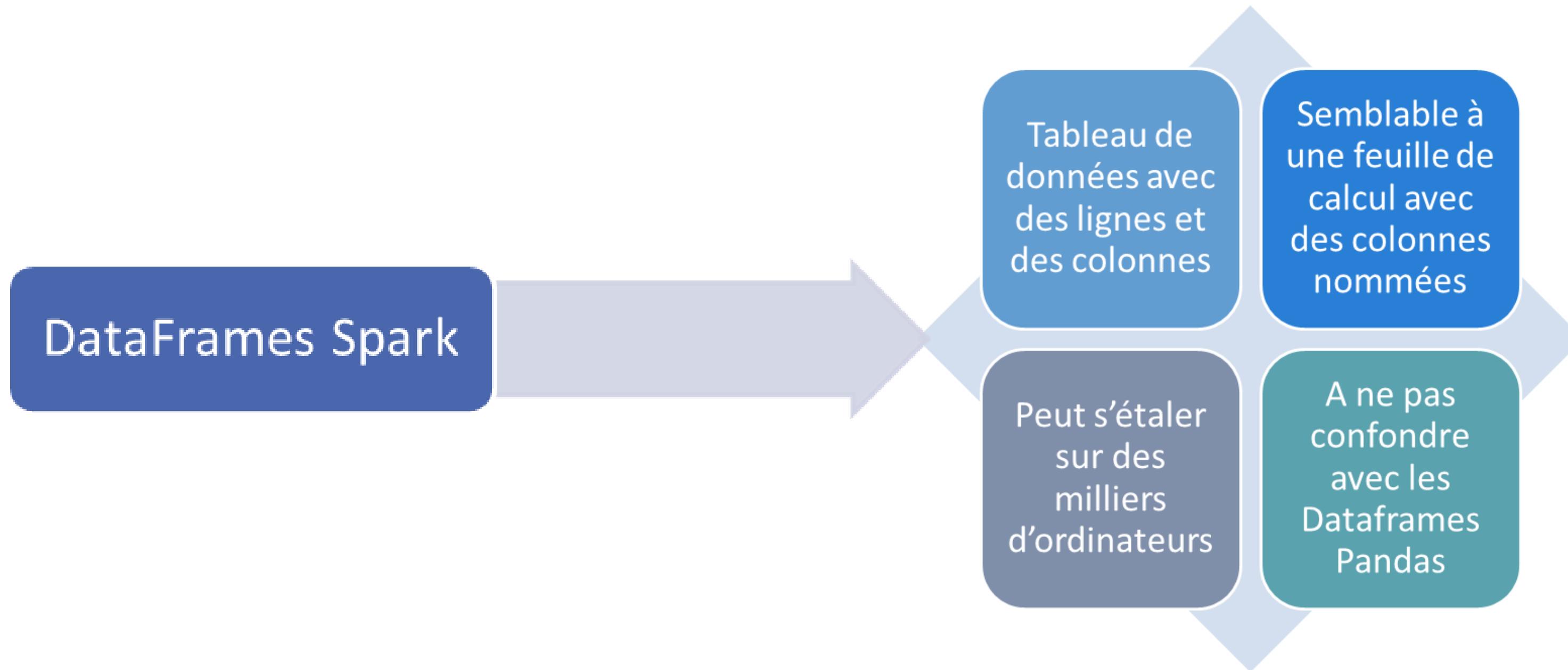
Spark

Concepts principaux



Spark

Concepts principaux



Spark

Concepts principaux

Pandas UDF (User Defined Function)

Fonction définie par l'utilisateur

Permet des opérations vectorisées

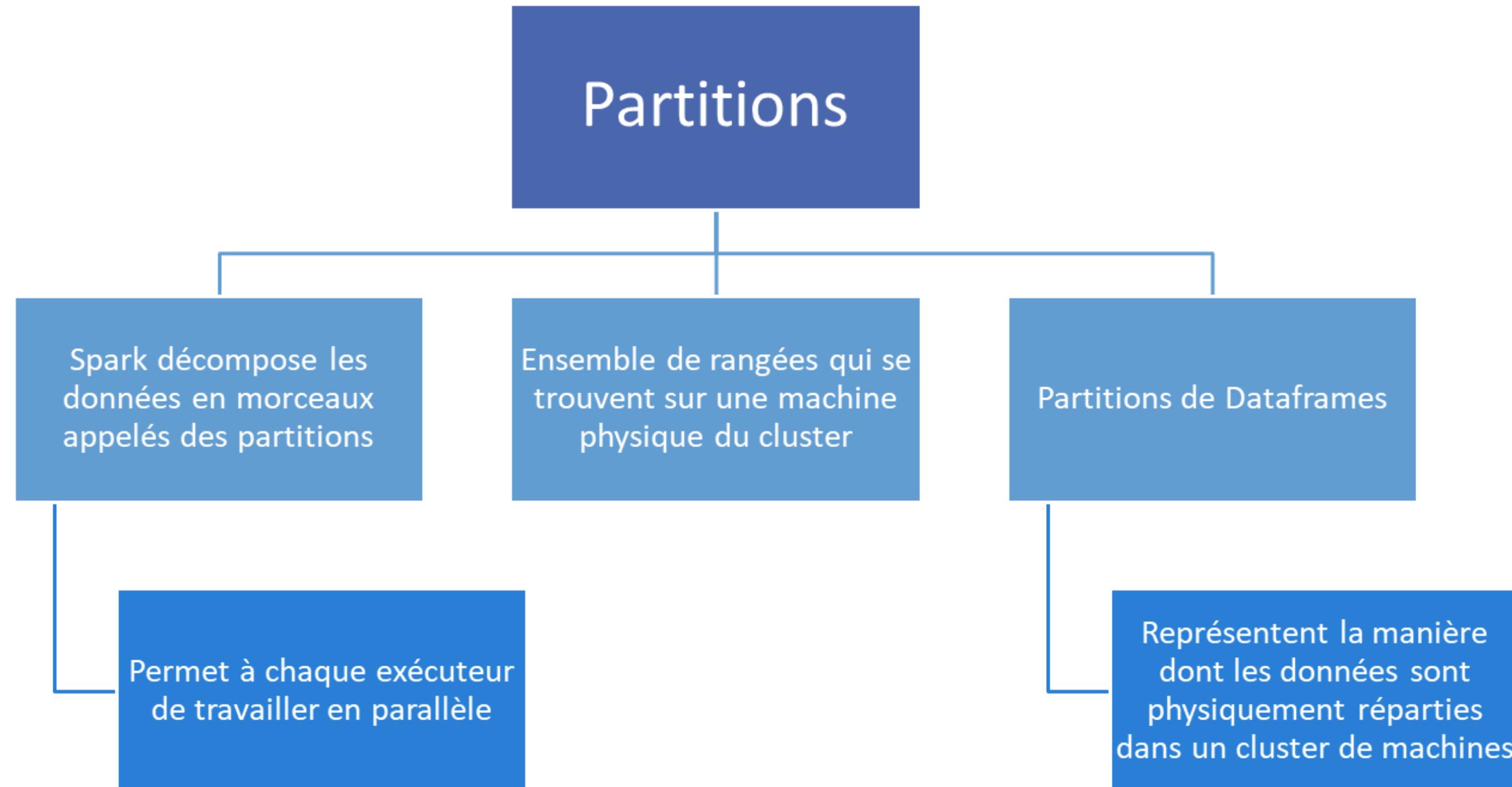
Apache Arrow pour transférer des données

Pandas pour travailler avec les données

Performances jusqu'à 100 fois supérieures aux UDF Python

Spark

Concepts principaux



Spark

Concepts principaux

Transformations



RDD: Résilient Distributed Dataset
Structures fondamentales de Spark
Objets Immuables



Passage d'un RDD en un autre RDD = Transformation



Ne renvoient aucun résultat
➤Lazy Evaluation
➤Spark n'agira pas sur les transformations tant que nous n'aurons pas appelé une action

Spark

Concepts principaux

The diagram consists of three light blue chevron-shaped arrows pointing from left to right. The first arrow is large and blue, containing the word 'Actions' in white. The second and third arrows are smaller and lighter blue, containing explanatory text. The second arrow contains the text 'Permet de déclencher un calcul' and the third arrow contains the text 'Demande à Spark de calculer un résultat à partir d'une série de transformations'.

Actions

Permet de déclencher un calcul

Demande à Spark de calculer un résultat à partir d'une série de transformations



Architecture big data



AWS –Qu'est que c'est ?

Service de cloud computing à la demande ,les plus populaires sont Elastic Compute Cloud (EC2) et Simple Storage Service(S3) et IAM (Identity and Access Management)

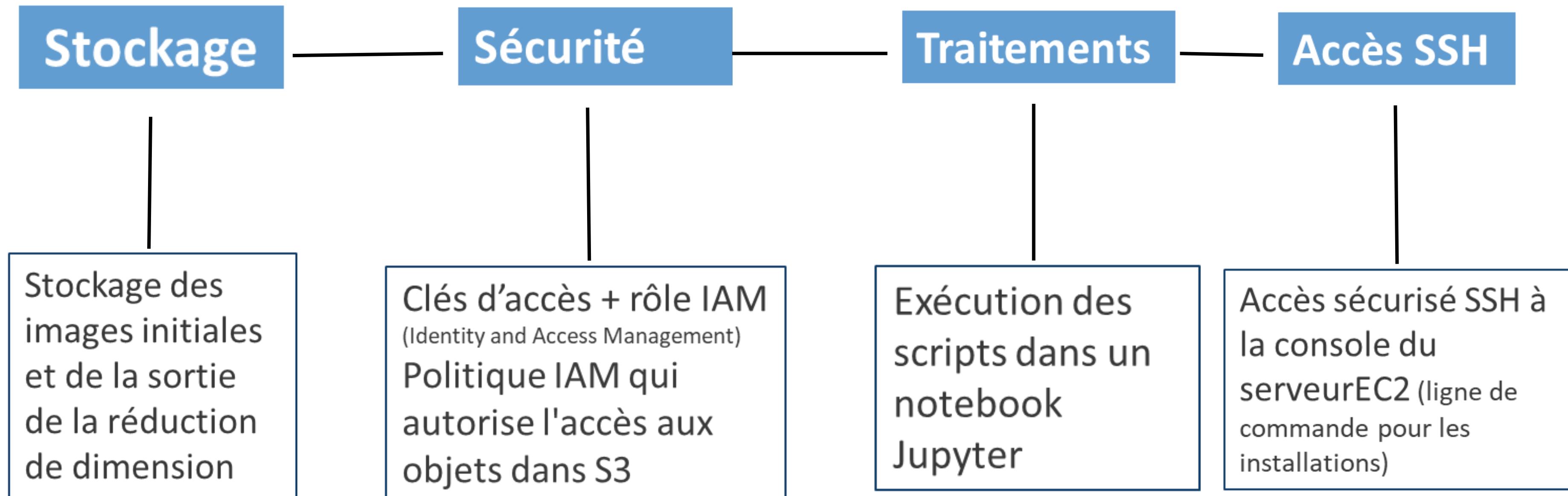
EC2 : est une Infrastructure as a Service (IaaS), car Amazon fournit un accès à une partie de ses serveurs mais c'est à l'utilisateur de gérer le système opératif, run time et data

EMR est une Solution PAAS (Plateforme As A Service) permet de louer des instances EC2 avec des applications préinstallées et configurées telles que Apache Spark.

S3: permet de stocker des données de manière infinie. La tarification s'adapte automatiquement à l'utilisation.

IAM : Permet de gérer les services AWS accessibles à un compte IAM (utilisateur).

Architecture big data



Stockage des images initiales et de la sortie de la réduction de dimension

Clés d'accès + rôle IAM (Identity and Access Management)
Politique IAM qui autorise l'accès aux objets dans S3

Exécution des scripts dans un notebook Jupyter

Accès sécurisé SSH à la console du serveurEC2 (ligne de commande pour les installations)



Amazon S3

► Instantané de compte

Afficher le tableau de bord de Storage Lens

Storage Lens offre une visibilité sur l'utilisation du stockage et les tendances d'activité. [En savoir plus](#)

Compartiments (2) [Info](#)



Copier l'ARN

Vider

Supprimer

Créer un compartiment

Les compartiments sont des conteneurs pour les données stockées dans S3. [En savoir plus](#)

Rechercher des compartiments par nom

< 1 > |

Nom	Région AWS	Accéder	Date de création
aws-logs-156809840539-eu-west-3	Europe (Paris) eu-west-3	Les objets peuvent être publics	03 Nov 2023 06:10:31 PM CET
testprojet8	Europe (Paris) eu-west-3	Les objets peuvent être publics	02 Nov 2023 10:06:05 AM CET

Amazon EMR

The first screenshot shows the 'Version Amazon EMR' selection dropdown set to 'emr-6.3.0'. Below it, the 'Offre d'applications' section lists various services: Spark, Core Hadoop, HBase, Presto, PrestoSQL, and Custom. A large list of optional applications is shown, with several checked (e.g., Hadoop 3.2.1, JupyterHub 1.2.0, Spark 3.1.1, TensorFlow 2.4.1) and some highlighted in blue.

The second screenshot shows the 'Configuration de cluster' section. It specifies 'Principale (m5.xlarge), Unité principale (m5.xlarge), Tâche (m5.xlarge)' as group sizes. The 'Dimensionnement et mise en service du cluster' section indicates 'Taille du noyau: 1 instance' and 'Taille de la tâche: 2 instances'.

The third screenshot shows the 'Configuration de mise en service' section where instance counts are set to 1 for the 'Unité principale' and 2 for 'Tâche - 1'. To the right, 'Groupes d'instances' are listed as 'Principale (m5.xlarge), Unité principale (m5.xlarge), Tâche (m5.xlarge)'. The 'Dimensionnement et mise en service du cluster' section is also visible.

Actions d'amorçage – facultatif Info

Utilisez les actions d'amorçage pour installer des logiciels ou personnaliser la configuration de votre instance.

Actions d'amorçage (1)		
Nom	Emplacement Amazon S3	Arguments
amorçage	s3://testprojet8/bootstrap-emr.sh	-

Supprimer **Modifier** **Ajouter**

Paire de clés Amazon EC2 pour SSH sur le cluster Info

Rôle Identity and Access Management (IAM) Info

Fonction du service Amazon EMR Info

Profil d'instance EC2 pour Amazon EMR Info

Récapitulatif Info

Actions d'amorçage – facultatif

Actions d'amorçage (1)

Emplacement Amazon S3

Arguments

Supprimer

Modifier

Ajouter

Services

Rechercher

Cloner un cluster

Amazon EMR

[Amazon EMR](#) > [EMR sur EC2: Clusters](#) > [Mon cluster](#)

Mon cluster

Mis à jour il y a 1 minute [C](#) Résilier [Cloner dans AWS CLI](#) [Cloner](#)

Récapitulatif	
Informations sur le cluster	Applications
ID de cluster j-2M32PEOIXXNIP	Version d'Amazon EMR emr-6.3.0
Configuration de cluster	Gestion des clusters
Groupes d'instances	Destination des journaux dans Amazon S3 aws-logs-156809840539-eu-west-3/elasticmapreduce
Capacité	Interfaces utilisateur d'application persistantes
1 primaire(s) 1 unité(s) principale(s) 2 tâche(s)	Serveur d'historique Spark Serveur de chronologie YARN
	DNS public du nœud primaire ec2-13-38-11-34.eu-west-3.compute.amazonaws.com
	Connexion au nœud primaire à l'aide de SSH

[Propriétés](#) [Actions d'amorçage](#) [Instances \(Matériel\)](#) [Étapes](#) [Applications](#) [Configurations](#) [Surveillance](#) [Événements](#) [identifications \(0\)](#)

Journaux de cluster Info	Résiliation du cluster Info
Archiver les fichiers journaux dans Amazon S3 Activé	Chiffrement pour les journaux Désactivé
Emplacement Amazon S3 s3://aws-logs-156809840539-eu-west-3/elasticmapreduce/	Option de résiliation Résilier manuellement le cluster
	Temps d'inactivité -
	Protection contre la résiliation Désactivé

[Modifier la résiliation du cluster](#)

Réseau et sécurité Info		
Réseau	Configuration de sécurité	Autorisations
Cloud privé virtuel (VPC) vpc-0ab657811cf54a09a	Configuration de sécurité Aucun	Fonction du service pour Amazon EMR EMR_DefaultRole
Sous-réseau(x) et zone(s) de disponibilité subnet-01d79b5a57ff9dc8c eu-west-3a	Paire de clés EC2 regis_ec2	Profil d'instance EC2 EMR_EC2_DefaultRole
▶ Groupes de sécurité EC2 (pare-feu)		Rôle d'autoscaling personnalisé EMR_AutoScaling_DefaultRole

[Amazon EMR](#) > [EMR sur EC2: Clusters](#)

Clusters (19) [Info](#)

Filtrer les clusters par statut ▾ [Rechercher des clusters](#) [Filtrer les clusters par date et heure de création](#)

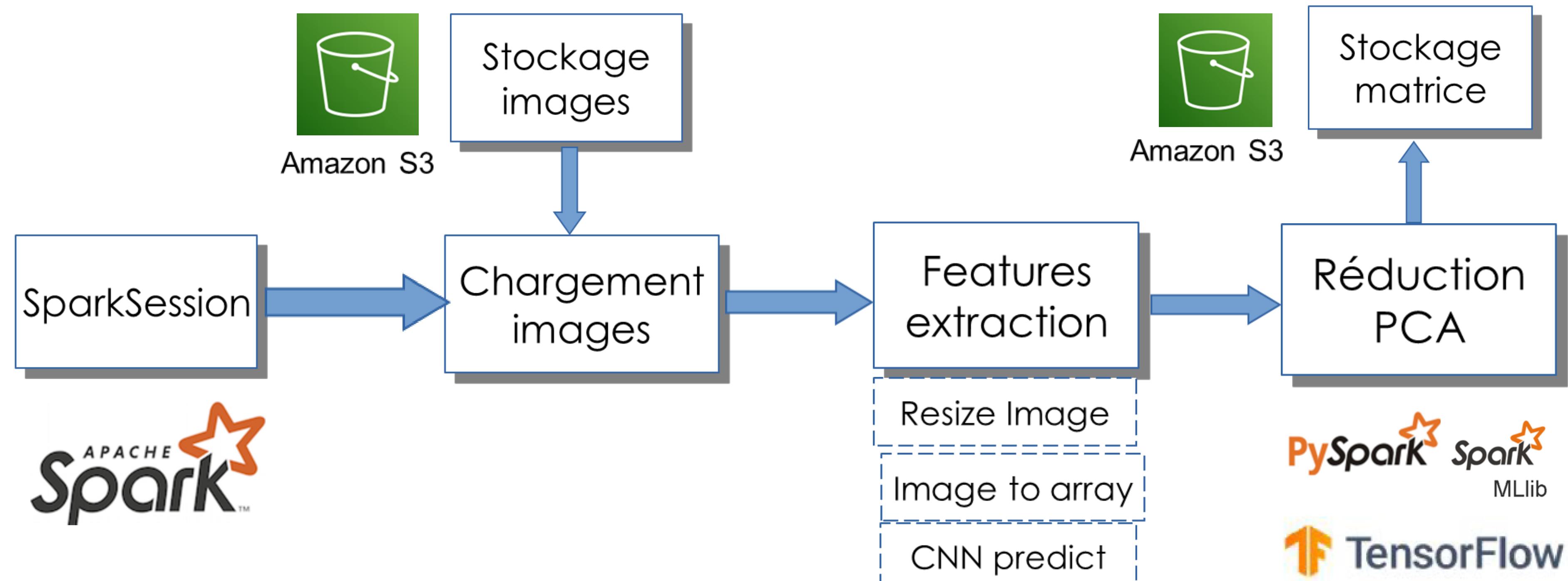
ID de cluster	Nom du cluster	Statut	Heure de création (UTC+01:00)	Temps é
j-2M32PEOIXXNIP	Mon cluster	✗ Résilié Demande utilisateur	9 novembre 2023 18:15	36 minutes
j-3MJSIUHU602WE	Mon cluster	✗ Résilié Demande utilisateur	9 novembre 2023 15:59	1 heure,
j-2M4FZ9HGM07P6	Mon cluster	✗ Résilié Demande utilisateur	7 novembre 2023 12:58	1 heure,
i-3c12a9rl7vni	Mon cluster	✗ Résilié	7 novembre 2023 10:43	1 heure



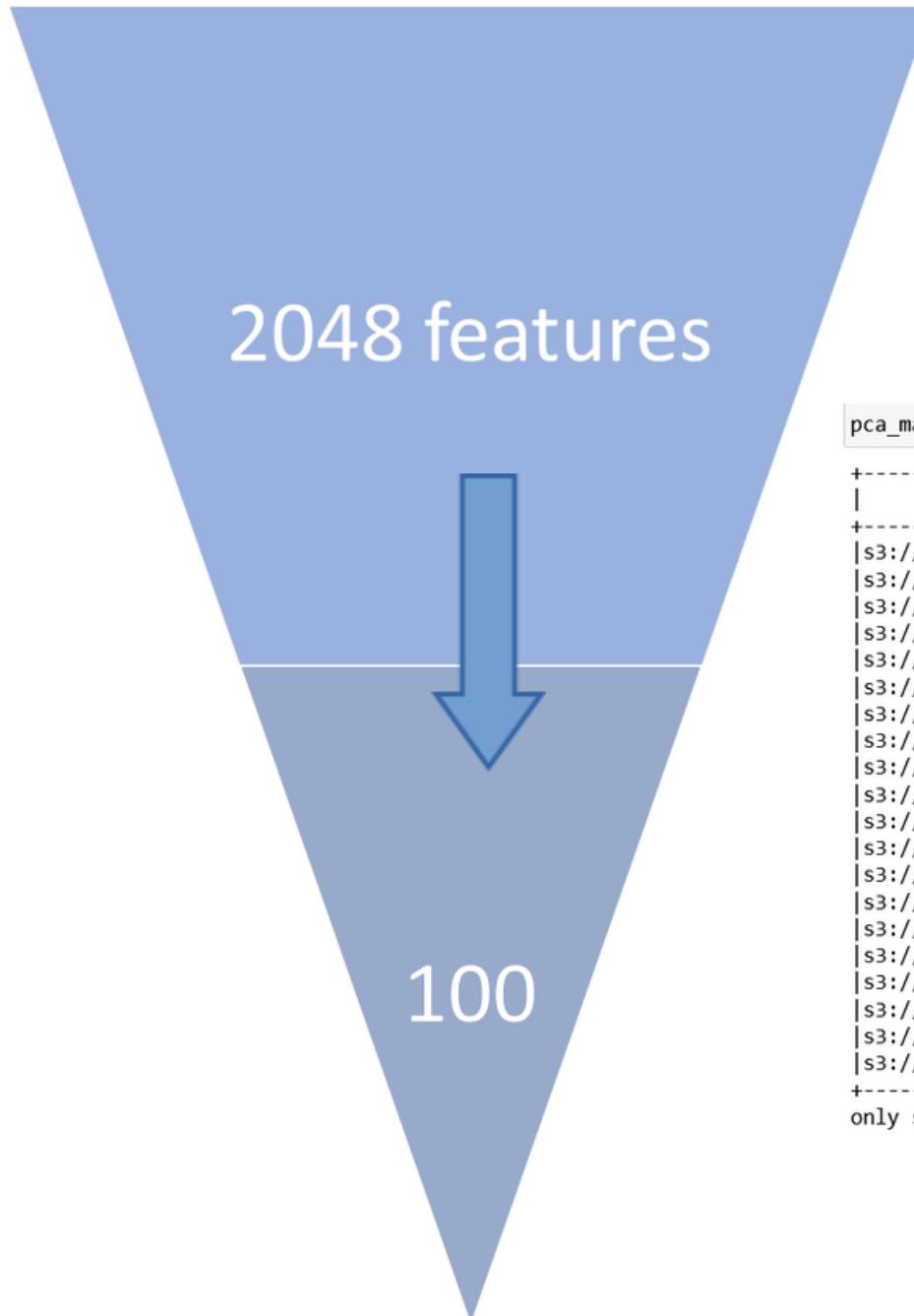
Application de la solution sur le Cloud

Etapes de la chaîne de traitement

Schéma du traitement des images dans notebook Jupyter



Réduction de dimension PCA



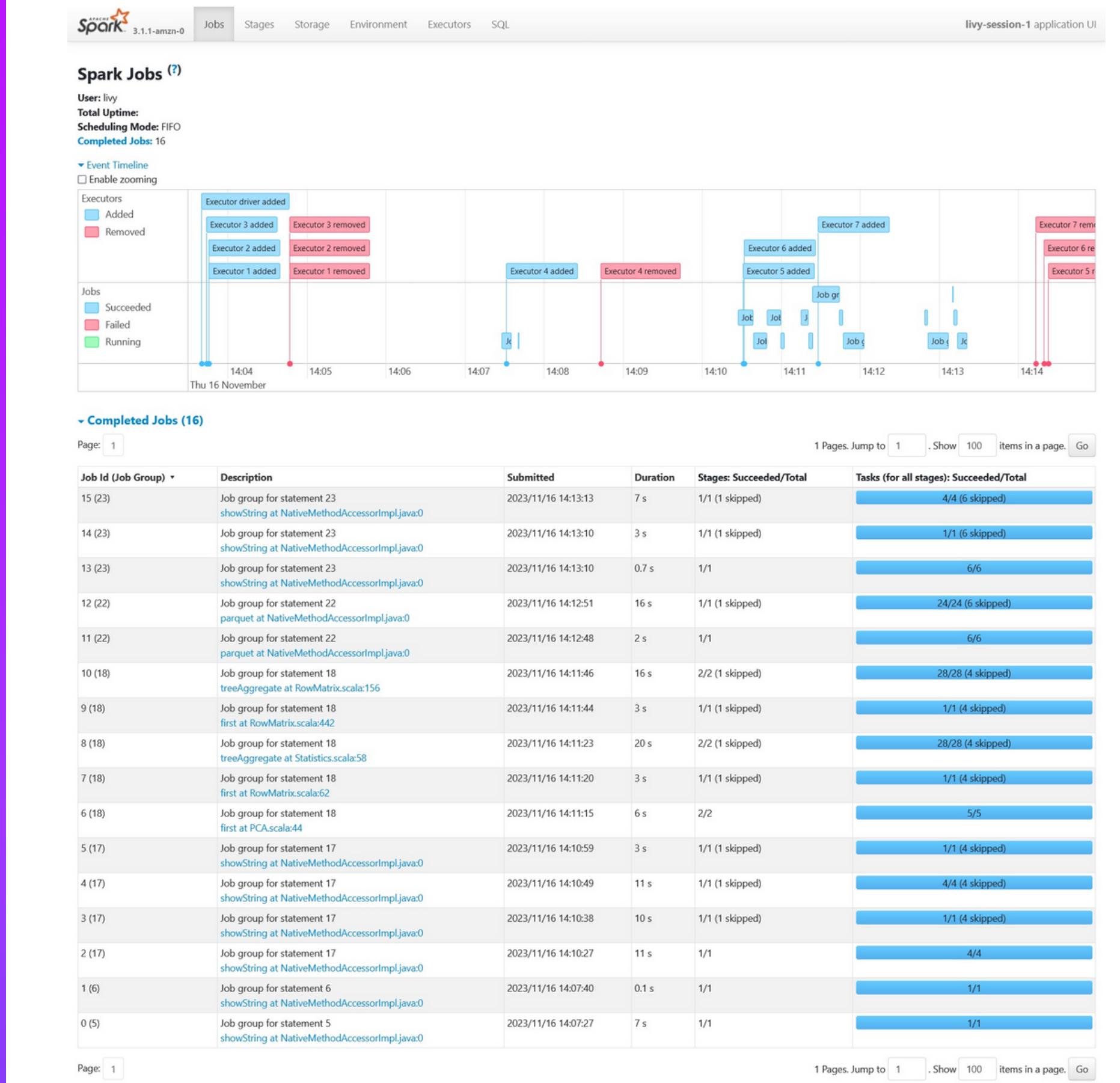
Composantes principales → Enregistrées au format parquet (réduit) dans S3

The screenshot shows the AWS S3 'Compartiments' view for a folder named 'testprojet8'. The 'Results/' folder contains 25 objects, all of which are parquet files. The files are organized into partitions based on their creation date and time. The columns in the table include 'Nom' (Name), 'Type' (Type), 'Dernière modification' (Last modified), 'Taille' (Size), and 'Classe de stockage' (Storage class). The sizes of the files range from 37.9 Ko to 63.7 Ko, and they were all modified on November 16, 2023, at various times between 03:12:55 PM and 05:15:07 PM CET.

Nom	Type	Dernière modification	Taille	Classe de stockage
_SUCCESS	-	16 Nov 2023 03:13:07 PM CET	0 o	Standard
part-00000-4904571-7eb4-41b5-94ef-0b9222070902-00000.parquet	parquet	16 Nov 2023 03:12:55 PM CET	51.9 Ko	Standard
part-00001-4904571-7eb4-41b5-94ef-0b9222070902-00001.parquet	parquet	16 Nov 2023 03:12:56 PM CET	60.3 Ko	Standard
part-00002-4904571-7eb4-41b5-94ef-0b9222070902-00002.parquet	parquet	16 Nov 2023 03:13:00 PM CET	63.7 Ko	Standard
part-00003-4904571-7eb4-41b5-94ef-0b9222070902-00003.parquet	parquet	16 Nov 2023 03:12:56 PM CET	52.7 Ko	Standard
part-00004-4904571-7eb4-41b5-94ef-0b9222070902-00004.parquet	parquet	16 Nov 2023 03:12:58 PM CET	51.6 Ko	Standard
part-00005-4904571-7eb4-41b5-94ef-0b9222070902-00005.parquet	parquet	16 Nov 2023 03:12:56 PM CET	62.7 Ko	Standard
part-00006-4904571-7eb4-41b5-94ef-0b9222070902-00006.parquet	parquet	16 Nov 2023 03:12:59 PM CET	63.3 Ko	Standard
part-00007-4904571-7eb4-41b5-94ef-0b9222070902-00007.parquet	parquet	16 Nov 2023 03:12:59 PM CET	58.7 Ko	Standard
part-00008-4904571-7eb4-41b5-94ef-0b9222070902-00008.parquet	parquet	16 Nov 2023 03:13:01 PM CET	38.3 Ko	Standard
part-00009-4904571-7eb4-41b5-94ef-0b9222070902-00009.parquet	parquet	16 Nov 2023 03:12:59 PM CET	51.8 Ko	Standard
part-00010-4904571-7eb4-41b5-94ef-0b9222070902-00010.parquet	parquet	16 Nov 2023 03:12:59 PM CET	51.6 Ko	Standard
part-00011-4904571-7eb4-41b5-94ef-0b9222070902-00011.parquet	parquet	16 Nov 2023 03:13:02 PM CET	48.9 Ko	Standard
part-00012-4904571-7eb4-41b5-94ef-0b9222070902-00012.parquet	parquet	16 Nov 2023 03:13:02 PM CET	49.2 Ko	Standard
part-00013-4904571-7eb4-41b5-94ef-0b9222070902-00013.parquet	parquet	16 Nov 2023 03:13:02 PM CET	50.0 Ko	Standard
part-00014-4904571-7eb4-41b5-94ef-0b9222070902-00014.parquet	parquet	16 Nov 2023 03:13:02 PM CET	60.0 Ko	Standard
part-00015-4904571-7eb4-41b5-94ef-0b9222070902-00015.parquet	parquet	16 Nov 2023 03:13:02 PM CET	64.4 Ko	Standard
part-00016-4904571-7eb4-41b5-94ef-0b9222070902-00016.parquet	parquet	16 Nov 2023 03:13:03 PM CET	50.5 Ko	Standard
part-00017-4904571-7eb4-41b5-94ef-0b9222070902-00017.parquet	parquet	16 Nov 2023 03:13:03 PM CET	51.9 Ko	Standard
part-00018-4904571-7eb4-41b5-94ef-0b9222070902-00018.parquet	parquet	16 Nov 2023 03:13:05 PM CET	50.7 Ko	Standard
part-00019-4904571-7eb4-41b5-94ef-0b9222070902-00019.parquet	parquet	16 Nov 2023 03:13:05 PM CET	52.7 Ko	Standard
part-00020-4904571-7eb4-41b5-94ef-0b9222070902-00020.parquet	parquet	16 Nov 2023 03:13:05 PM CET	39.5 Ko	Standard
part-00021-4904571-7eb4-41b5-94ef-0b9222070902-00021.parquet	parquet	16 Nov 2023 03:13:07 PM CET	37.9 Ko	Standard

Le serveur d'historique Spark

nous permet une vision beaucoup plus précise de l'exécution des différentes tâche sur les différentes machines du cluster



Conclusions

Enseignements

- Utiliser les outils du cloud pour manipuler des données dans un environnement Big Data
- Découverte de l'écosystème AWS
- Paralléliser des opérations de calcul avec Pyspark

Difficultés rencontrés

- Nombreuses possibilités techniques : choix complexes
- Débug complexe dû à des erreurs peu explicites (superposition Spark/Java)

Amélioration

- Passer à l'échelle, augmentation du nombre d'instances esclaves (nœuds)
- Déployer le modèle en production



Thank you