PROJET 8 :

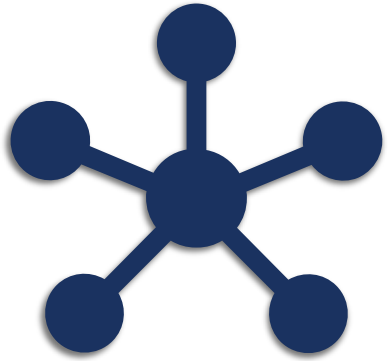
DÉPLOYER UN MODÈLE DANS LE CLOUD

SHERALI ASSEFY

E

16/01/2022

PLAN DE PRESENTATION



Problématique & présentation du jeu de données

Présentation de la chaine de traitement et de l'environnement Big Data choisi dans le cloud

Conclusion

CONTEXTE / MISSIONS

- ✓ **« Fruits» start-up de l'AgriTech** souhaite proposer une solution innovante de récolte des fruits avec des robots cueilleurs intelligents.
- ✓ **Première étape :** mettre en place une application mobile de reconnaissance des fruits.
- ✓ **Mission :** développer une première architecture Big Data :
  - ✓ Préprocessing des images et réduction de dimension
  - ✓ Anticipation du passage à l'échelle

JEU DE DONNÉES

- **Source :**
  - Jeu de données Kaggle (https://www.kaggle.com/moltean/fruits)
  - Données sous licence du MIT / Auteur : Dr Milhai Oltean

kaggle

- **Caractéristiques**
  - Nombre total d'images : 90 483
  - Données entrainement : 67 692 images (un fruit ou légume par image)
  - Données test : 22 688 images (un fruit ou légume par image)
  - Nombre total de classes : 131(fruit & légumes)
    - Plusieurs variétés du même fruit ou légume (exemple : Pommes Granny Smith, Dame rose, Rouge,...)
    - Le label est inscrit dans le nom de dossier de chaque variété
  - Taille de l'image : 100 x 100 pixels

CHAINE DE TRAITEMENT BIG DATA

- Big Data et enjeux
- Architecture mise en place
- Pré-traitement des données

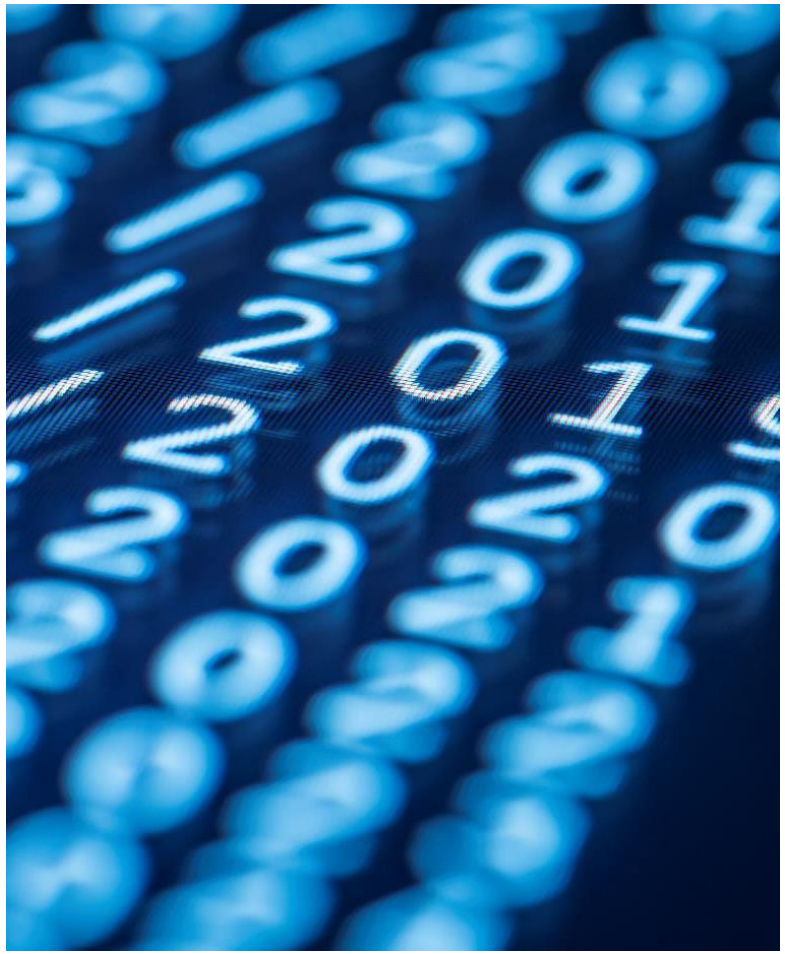
COMMENT DÉFINIR LE BIG DATA ?

- **On parle de « mégadonnées » ou « données massives »**

- **Les 3 « V » du Big Data :**
  - **Volume** : trop important pour être stocké et/ou traité sur une seule machine avec des performances acceptables.
    - Dépassement de la capacité de RAM
    - Dépassement des capacités de stockage
  - **Vitesse** : à laquelle les données sont reçues et éventuellement traitées.
  - **Variété:** Nombreux types de données (structures / non structurées & semi-structurées tels que le texte, l'audio et la vidéo)

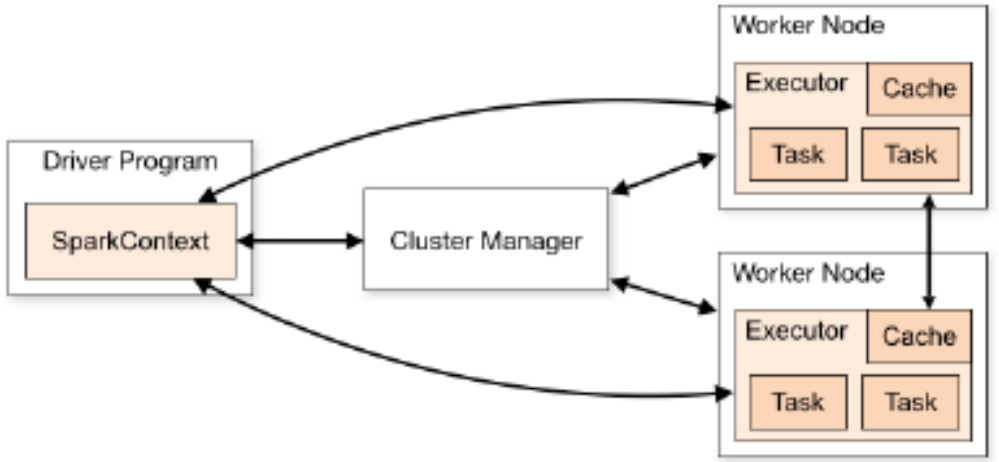
Ces dernières années on parle également de :

- ***Valeur:*** *les données possèdent une valeur intrinsèque*
- ***Véracité:*** *fiabilité des données*



BIG DATA / COMMENT TRAITER CES ENJEUX ?

- **Traitement par calculs distribués (MapReduce):**
  - Diviser les opérations en plus petites opérations distribuables entre différentes machines et traitement en parallèle
  - Agrégation des résultats sur une même machine



Application maître :
Configuration / Initialisation
Agrégation des calculs

Cluster Manager :
Gestion des ressources
Distribution des calculs entre les workers

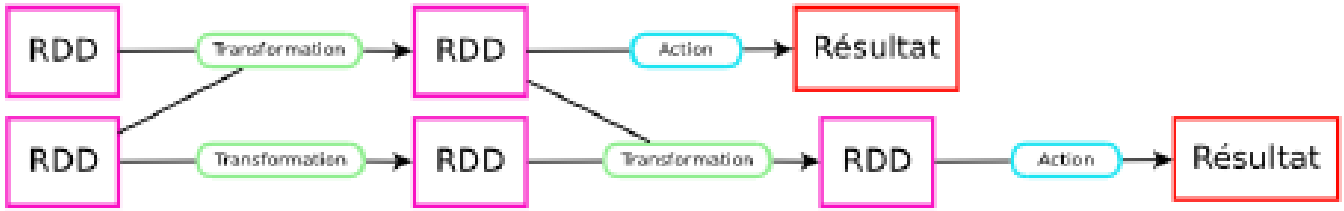
Workers :
Exécution des tâches en parallèle

BIG DATA / COMMENT TRAITER CES ENJEUX ?

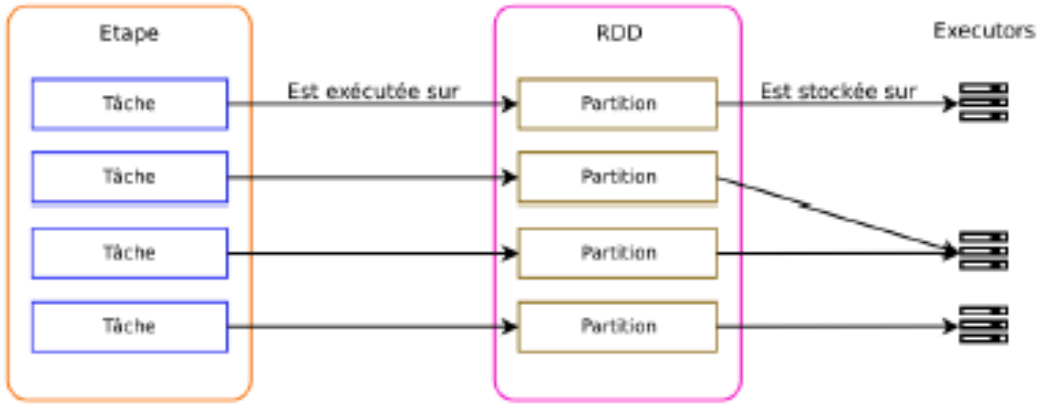
- **Stockage : système de fichier distribué (ex : HDFS)**

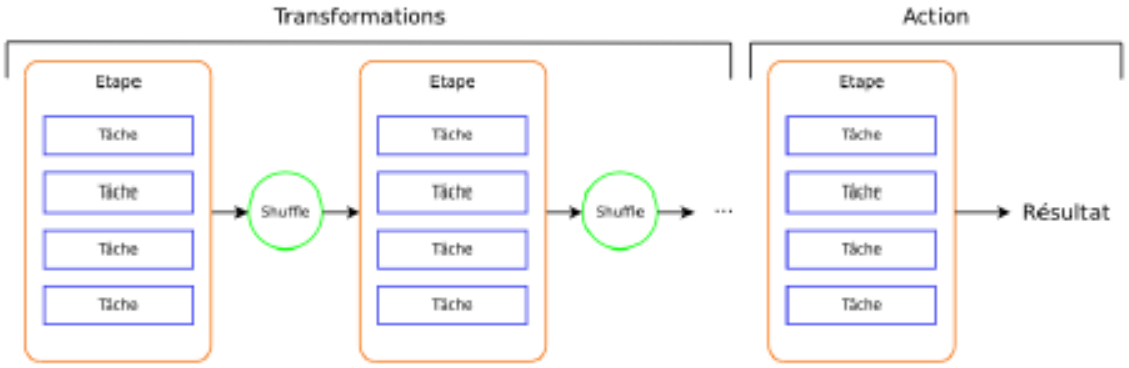
- **Tolérance aux pannes :**
  - **Utilisation de Resilient Distributed Datasets (RDD):**
    - **Division des données en partitions**
    - **Duplication des données (sur plusieurs machines)**

  - **Graphe Acyclique Orienté (DAG) :**
    - **Panne : Régénération à partir des nœuds parents**
    - **Nœuds (RDD ou Résultats) : liés par des actions et transformations**



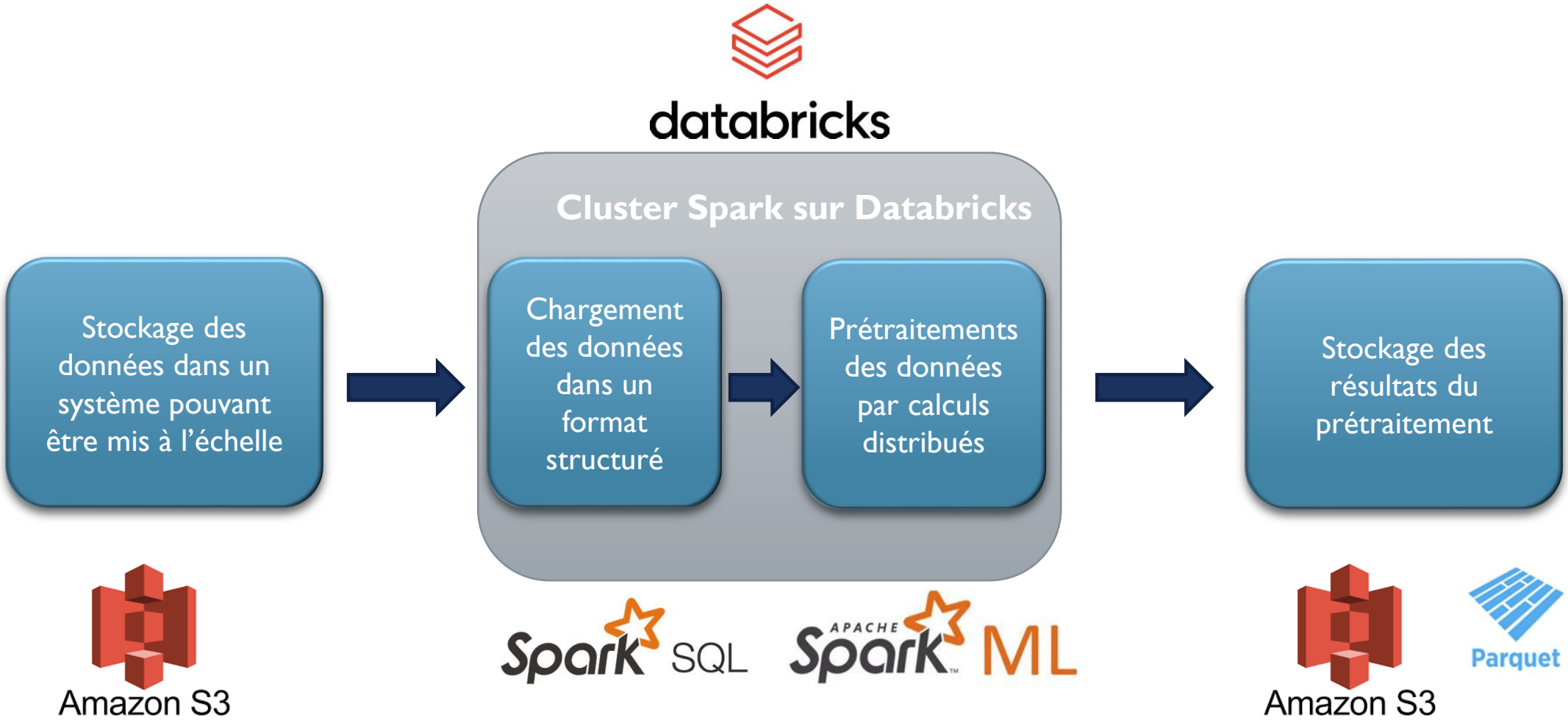
BIG DATA / COMMENT TRAITER CES ENJEUX ?



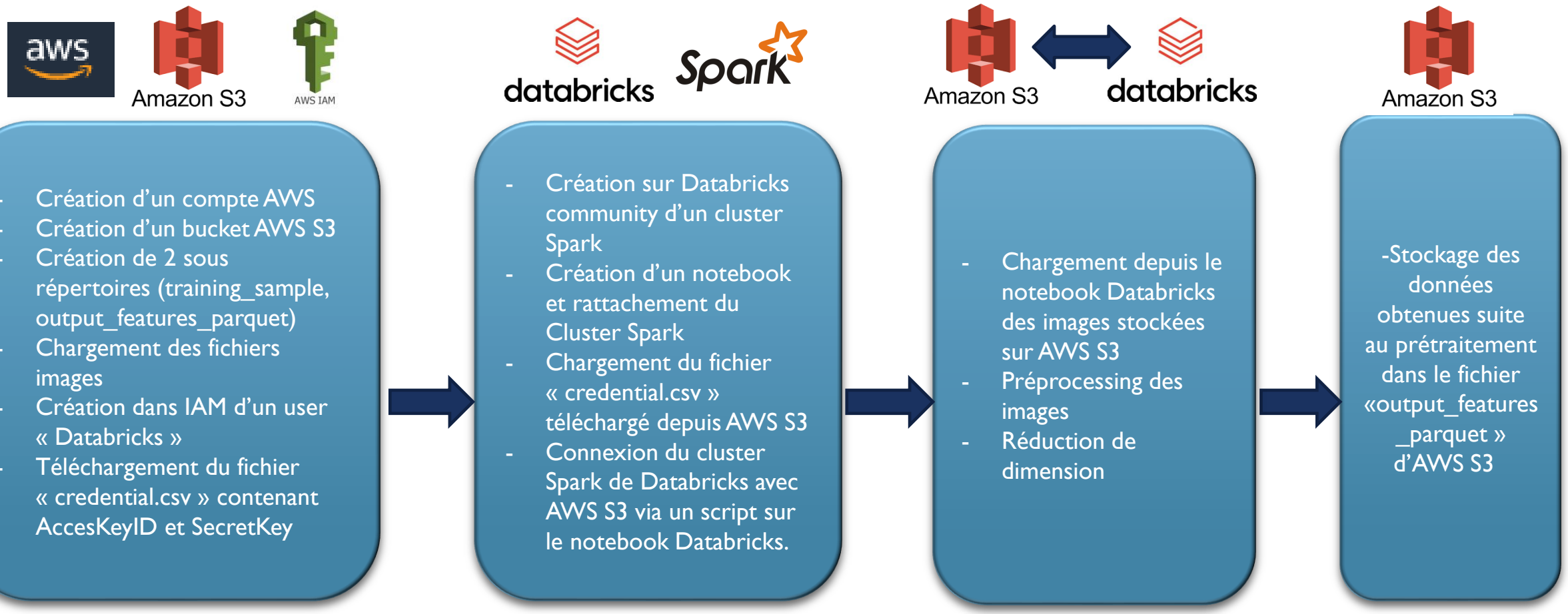


Shuffle = redistribution des données entre les nœuds

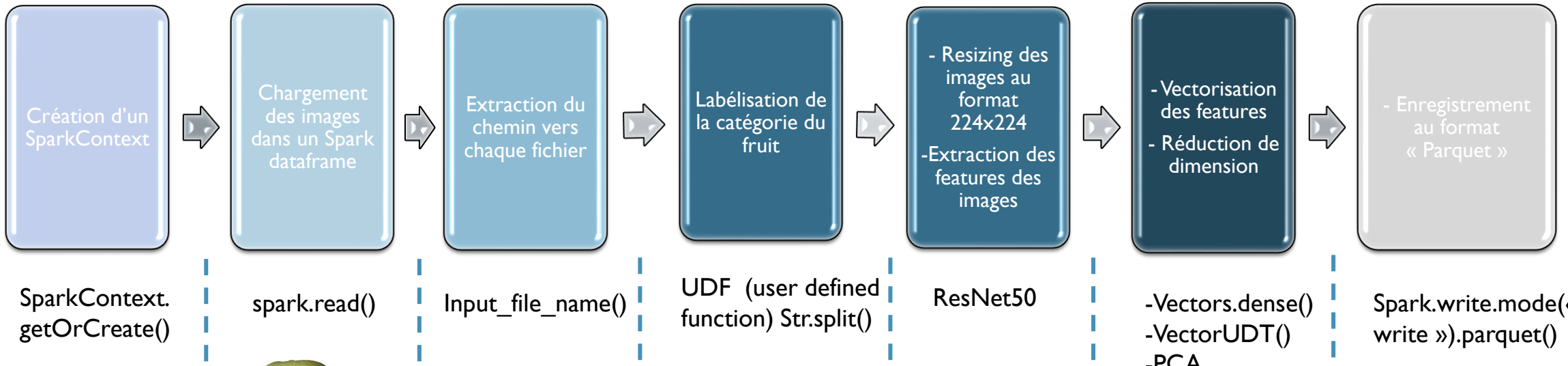
ARCHITECTURE MISE EN PLACE



PROCESS DE TRAVAIL
POUR LA MISE EN PLACE DE L'ARCHITECTURE CIBLE



PRÉTRAITEMENT DES DONNÉES



path	label	features_pca
dbfs:/mnt/dbp8v2/...	cabbage_white_1	[2.58817404745408...]
dbfs:/mnt/dbp8v2/...	cabbage_white_1	[3.21223116140212...]
dbfs:/mnt/dbp8v2/...	cabbage_white_1	[1.89440665733665...]
dbfs:/mnt/dbp8v2/...	cabbage_white_1	[2.68806227425664...]
dbfs:/mnt/dbp8v2/...	cabbage_white_1	[2.28526925334920...]

CONCLUSION

- Passage à l'échelle
- Conclusion

PASSAGE À L'ÉCHELLE

- **Code Spark/Python :** aucune modification à apporter

- **Stockage des fichiers:**
  - **AWS S3** : permet de stocker des données de manière infinie. La tarification s'adapte automatiquement à l'utilisation.

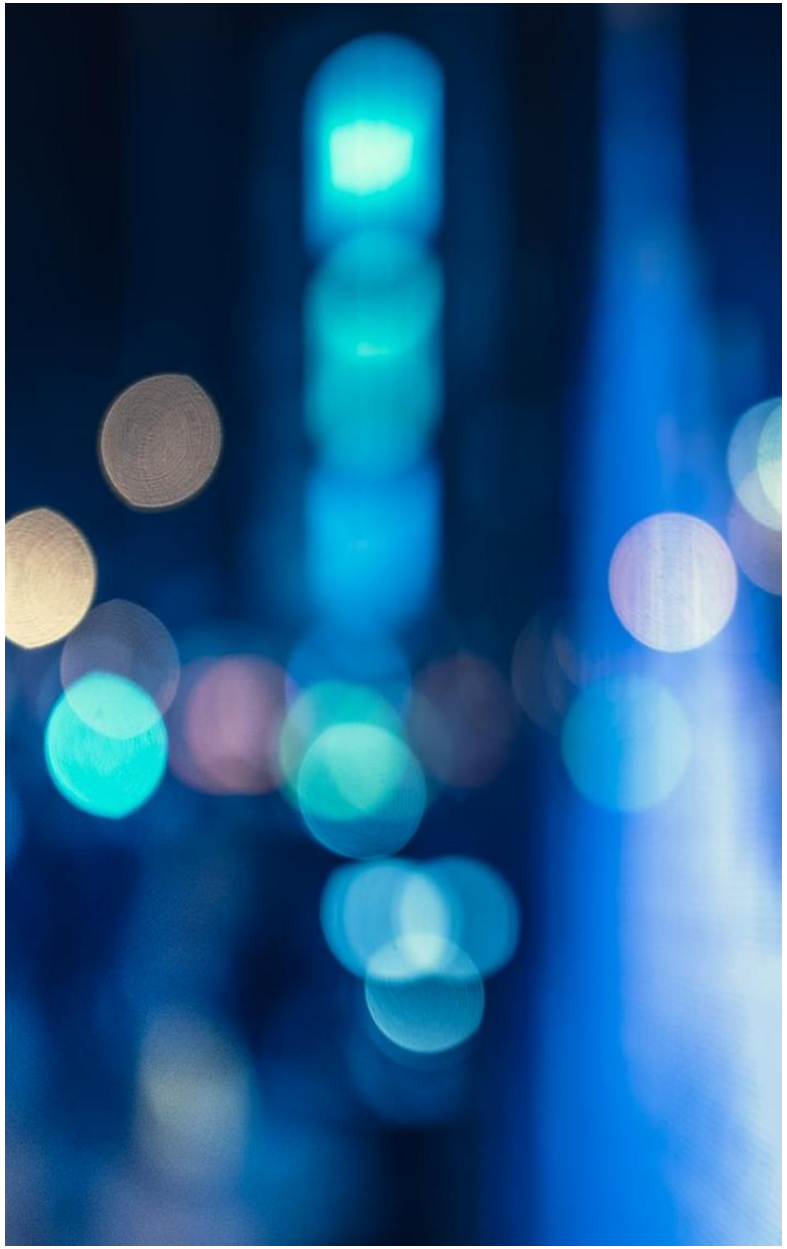
- **Infrastructure de calcul:**
  - **Databricks**: passage à l'échelle facilité (une simple case à cocher)

CONCLUSION

- **Réalisations :**
  - Création d'une architecture Big Data permettant de passer facilement à l'échelle
  - Réalisation des objectifs de la mission :
    - Stockage des données initiales dans le cloud (AWS S3)
    - Préprocessing des données (Databricks)
    - Réduction de dimension (Databricks)
    - Stockage des données préprocessées dans le cloud (AWS S3)

- **Difficultés:**
  - Choix complexes car il y a de nombreuses possibilités techniques.

- **Pistes d'amélioration:**
  - Optimisation du code pour réduire le temps de traitement des tâches
  - Prétraitement pour cas reels (plusieurs fruits, arrière plan,…)
  - Entrainer le modèle de transfer learning
  - Monitoring
  - Développer les cas d'usage (pathologies, qualité du fruit, maturité…)



MERCI DE VOTRE ATTENTION