DÉPLOIEMENT D'UN MODÈLE DANS LE CLOUD

1. INTRODUCTION
2. ENJEUX DES DONNÉES MASSIVES
3. TRAITEMENT DES DONNÉES
4. ARCHITECTURE ET INGÉNIERIE DES DONNÉES
5. PERSPECTIVES ET CONCLUSION

1 INTRODUCTION

Le présent projet vise à concevoir une application mobile qui permettra à ses utilisateurs de prendre en photo un fruit afin d'obtenir des informations détaillées sur ce dernier. Cette application sera capable d'identifier le fruit en question, de fournir des informations nutritionnelles ainsi que des conseils de conservation.

Pour ce faire, l'équipe de développement travaillera à la mise en place d'un environnement Big Data, incluant une phase de preprocessing afin d'optimiser la qualité des données collectées. Une étape de réduction de dimension sera également effectuée pour garantir une utilisation efficace des ressources informatiques et ainsi, permettre un traitement rapide des informations.

Le développement de cette application mobile permettra aux utilisateurs de mieux comprendre les fruits qu'ils consomment, en fournissant des informations utiles sur leur provenance, leurs caractéristiques et leurs bienfaits pour la santé. Ce projet a pour ambition d'améliorer l'expérience utilisateur tout en contribuant à la promotion d'une alimentation saine et équilibrée.

Le jeu de données contient plus de 90 000 images de 130 types différents de fruits.

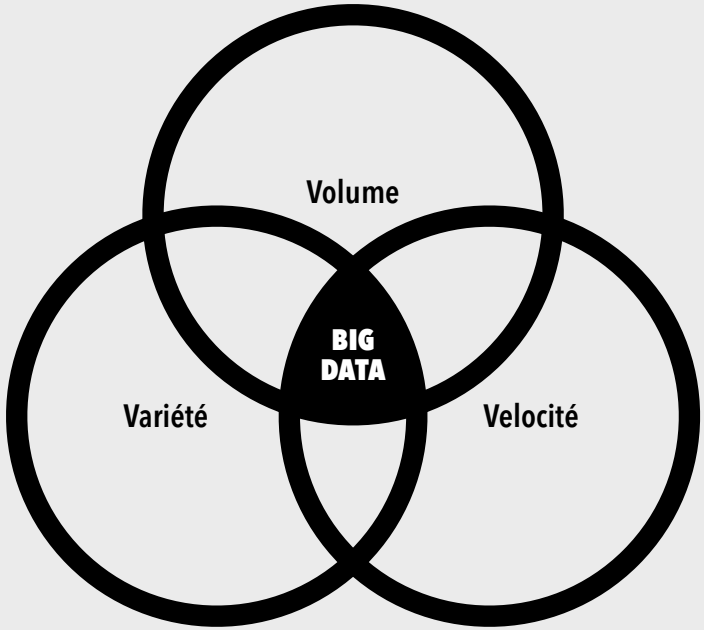
Notamment des pommes, des bananes, des oranges, des fraises, des citrons, des mangues, etc. **Les images ont été capturées à partir de différents angles et avec différentes tailles, couleurs et formes**. Le jeu de données est organisé en différents dossiers correspondant à chaque type de fruit, avec des sous-dossiers pour chaque variété de fruit.

Exemple

2 ENJEUX DES DONNÉES MASSIVES

La mise en place d'un écosystème cohérent, robuste et évolutif est un enjeu majeur pour la gestion et l'analyse des données massives

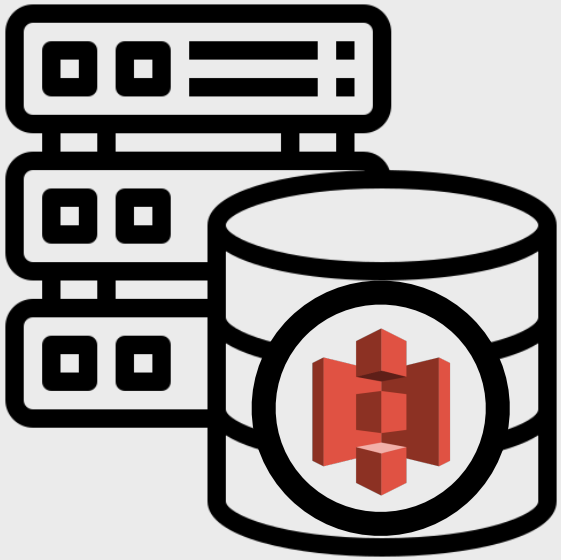
Les données massives se caractérisent par leur volume (très grande quantité de données), leur vélocité (rythme élevé de collecte et de génération des données) et leur variété (diversité des types et sources de données). Ces caractéristiques nécessitent des technologies adaptées pour leur gestion et leur analyse, telles que le stockage et le traitement distribué et l'analyse en temps réel.



LE STOCKAGE DISTRIBUÉ

S3 (Simple Storage Service) d'AWS garantit une haute disponibilité et une durabilité élevée des données stockées grâce à son architecture de stockage distribué

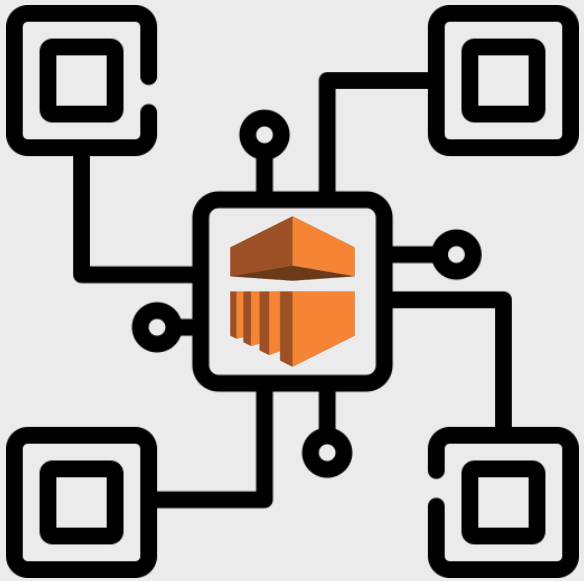
Le stockage distribué est une méthode de stockage de données qui utilise plusieurs serveurs pour stocker les données plutôt qu'un seul serveur. Les données sont réparties sur plusieurs nœuds ou serveurs, ce qui permet d'augmenter la capacité de stockage globale et de réduire les risques de perte de données en cas de panne d'un serveur.



LE CALCUL DISTRIBUÉ

EMR permet de traiter de grandes quantités de données en parallèle, ce qui réduit considérablement le temps de traitement global

Il peut être utilisé pour des analyses en temps réel, des requêtes interactives et des analyses de données en batch. EMR simplifie également la mise en place, la configuration et la gestion des clusters Hadoop et Spark, qui sont des outils de traitement de données à grande échelle. EMR offre également un modèle de tarification à la demande et gère automatiquement les ressources pour réduire les coûts.



LE CALCUL DISTRIBUÉ AVEC SPARK EN DÉTAIL

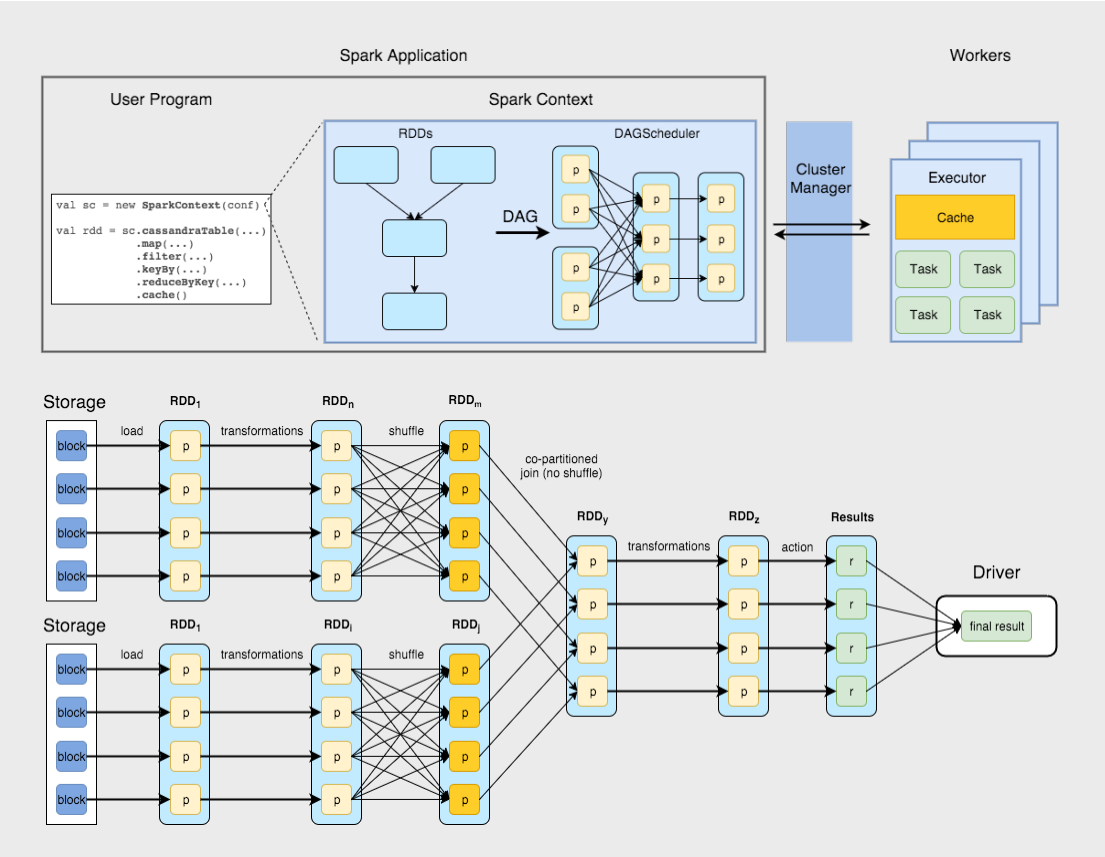
1. Spark divise les données en petits morceaux appelés partitions, qui sont traitées en parallèle sur plusieurs nœuds d'un cluster.

2. Il distribue ensuite les partitions sur les nœuds du cluster. Chaque nœud traite les partitions qui lui ont été attribuées.

3. Chaque nœud traite les partitions qui lui ont été attribuées en utilisant des transformations telles que les filtres, les maps, les réductions et les jointures.

4. Les résultats partiels de chaque partition sont ensuite combinés pour produire un résultat final.

5. Le résultat final est renvoyé à l'application qui a lancé le calcul distribué.



ENJEUX DE SÉCURITÉ ET CONFORMITÉ DES DONNÉES MASSIVES

La mitigation des risques de sécurité et la conformité permettent a **Fruits!** de mieux se protéger contre les menaces de sécurité et de gagner la confiance de ses clients

Sécuriser les données : Mise en place de politiques de sécurité strictes, la mise en place de pare-feu et de systèmes de détection d'intrusion, ainsi que l'utilisation de chiffrement et d'autres mesures de sécurité avancées.

Gérer les identités et les accès : Mise en place de politiques de gestion des identités et des accès, ainsi que l'utilisation de systèmes de gestion des identités et des accès.

Contrôler l'accès aux données : Mise en place de contrôles d'accès et d'authentification robustes, ainsi que la surveillance de l'accès aux données.

Respecter les réglementations : Mise en place de politiques et de procédures de conformité, ainsi que la formation des employés sur les exigences de conformité.

Surveiller les activités : Mise en place de systèmes de surveillance et d'analyse des journaux d'activité, ainsi que la mise en place de politiques de surveillance des utilisateurs.

3 TRAITEMENT DES DONNÉES

Le prétraitement d'images avec Pillow améliore la qualité, la cohérence et la taille des fichiers des images, accélère leur traitement et améliore la précision des modèles d'apprentissage automatique

Pillow est une bibliothèque open-source en Python pour le traitement d'images, utilisée pour effectuer des opérations telles que le redimensionnement, la rotation et la correction de couleurs.

La réduction de dimension accélère le temps de traitement des données et améliore la précision des modèles, ce qui peut améliorer la scalabilité et la performance globale de l'infrastructure

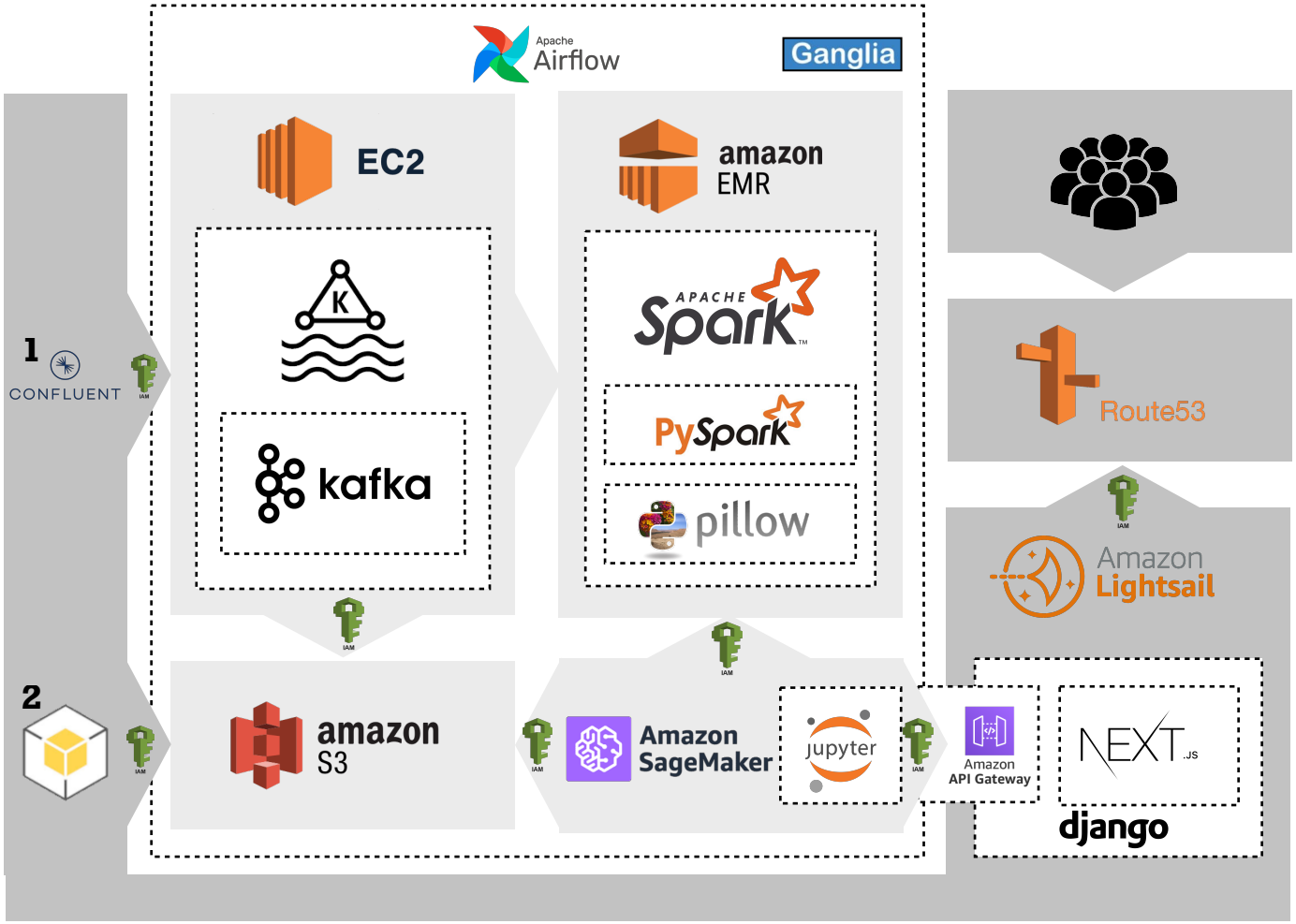
En appliquant une réduction de dimension à un fichier Parquet, on peut éliminer les colonnes non importantes ou corrélées, réduire la taille du fichier et améliorer l'efficacité de stockage et de traitement des données structurées de grande taille.

4 ARCHITECTURE ET INGÉNIERIE DES DONNÉES

Architecture simplifiée de la solution **Fruits!** complète

Les utilisateurs téléchargent des photos sur l'application Web Django. Les données de connexion sont envoyées à Kafka via Confluent pour une gestion centralisée des connexions. Les images sont stockées sur S3 via Boto3. Les données sont ensuite traitées par un cluster EMR via AWS SageMaker, qui renvoie les prédictions à Django pour affichage dans l'interface utilisateur ReactJS. Toutes les échanges sont sécurisés et unidirectionnels.

En utilisant les informations de connexion, notamment la géolocalisation, S3 ingère les données dans les buckets dédiés situés dans la région la plus proche de l'utilisateur. De même, EMR utilise ces informations pour lancer des clusters sur les serveurs les plus proches de la cible, réduisant ainsi le temps de réponse et améliorant la performance du système global. Cette optimisation de la distribution des ressources garantit une expérience utilisateur fluide et rapide.



PERSPECTIVES ET CONCLUSION