Déployez un modèle dans le cloud

Fruits!

Meltem YILDIRIM

30/11/2021

Sommaire

I. Problématique

- **Fruits! :**
  - Startup AgriTech
  - Solutions innovantes pour la récolte des fruits
    - → Robots cueilleurs intelligents
  - Application mobile de reconnaissance de fruits
- **Missions :**
  - Développer un environnement Big Data
  - Preprocessing
  - Réduction de dimension

Fruits!

II. Jeu de données

- **Datas :** Fruits 360
- **Propriétés :**
  - Training set : **67692** images
  - Test set : **22688** images
  - Nombre de catégories de fruits : **131**
  - Taille des images : **100x100** pixels
  - Test multiple fruits : **103** images

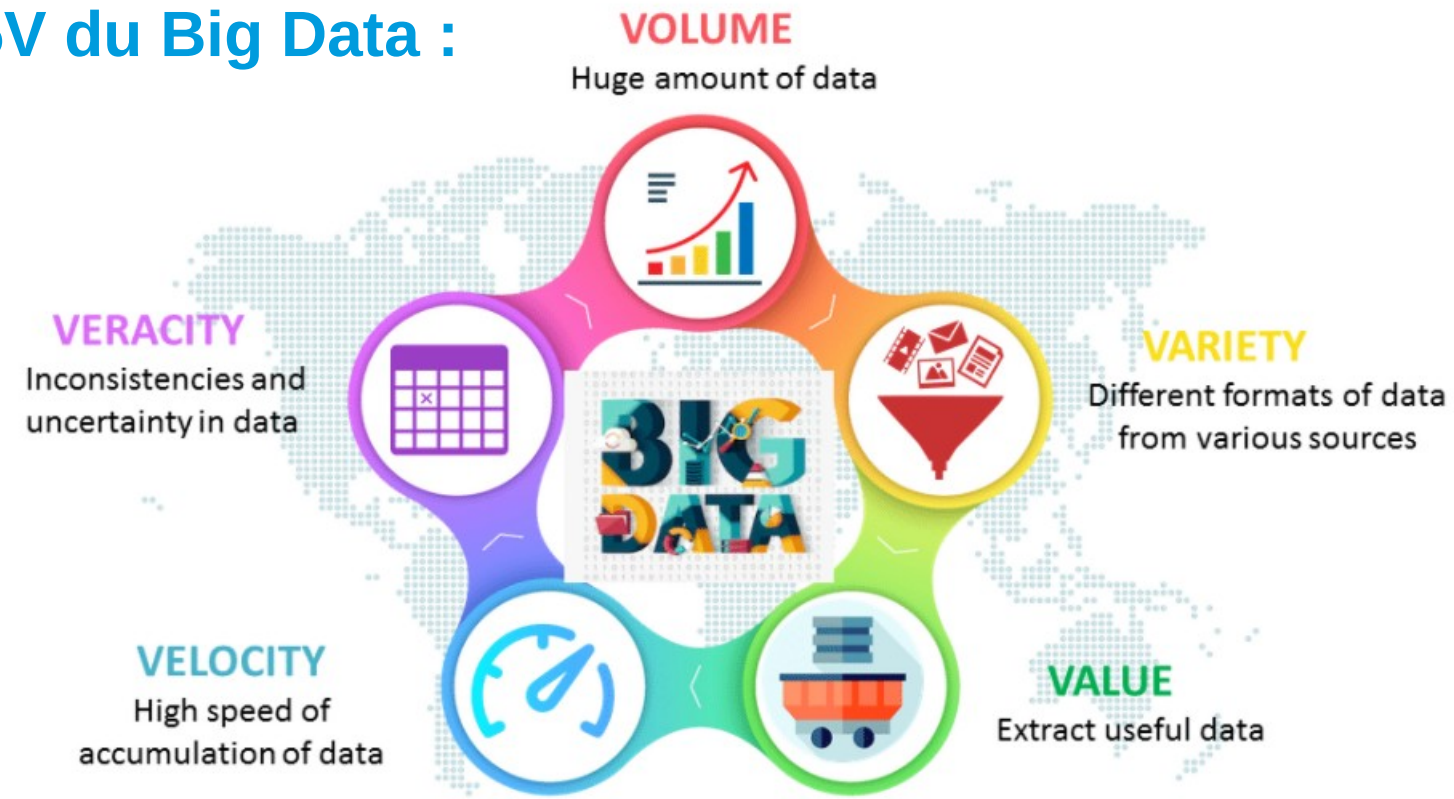
Plum

Watermelon

Apple Red 3

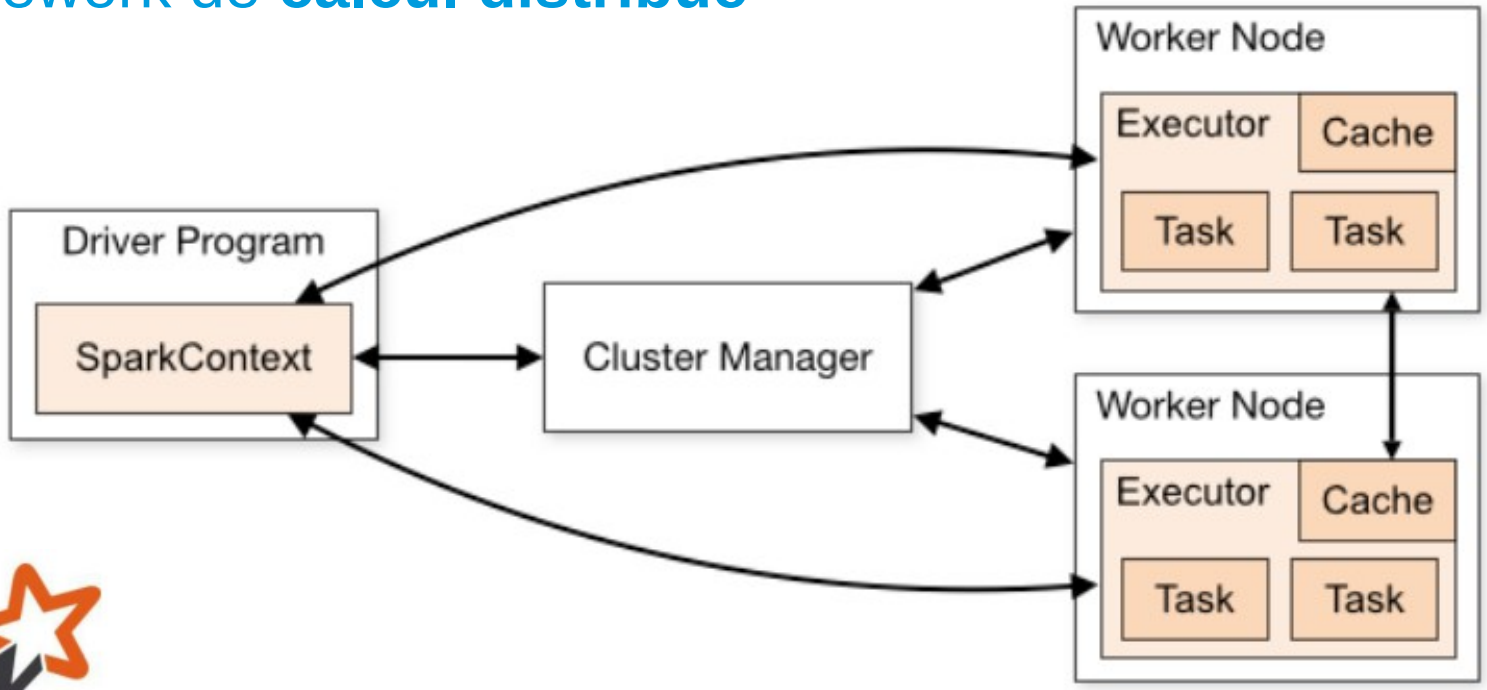
III. Big Data

- **Les 5V du Big Data :**

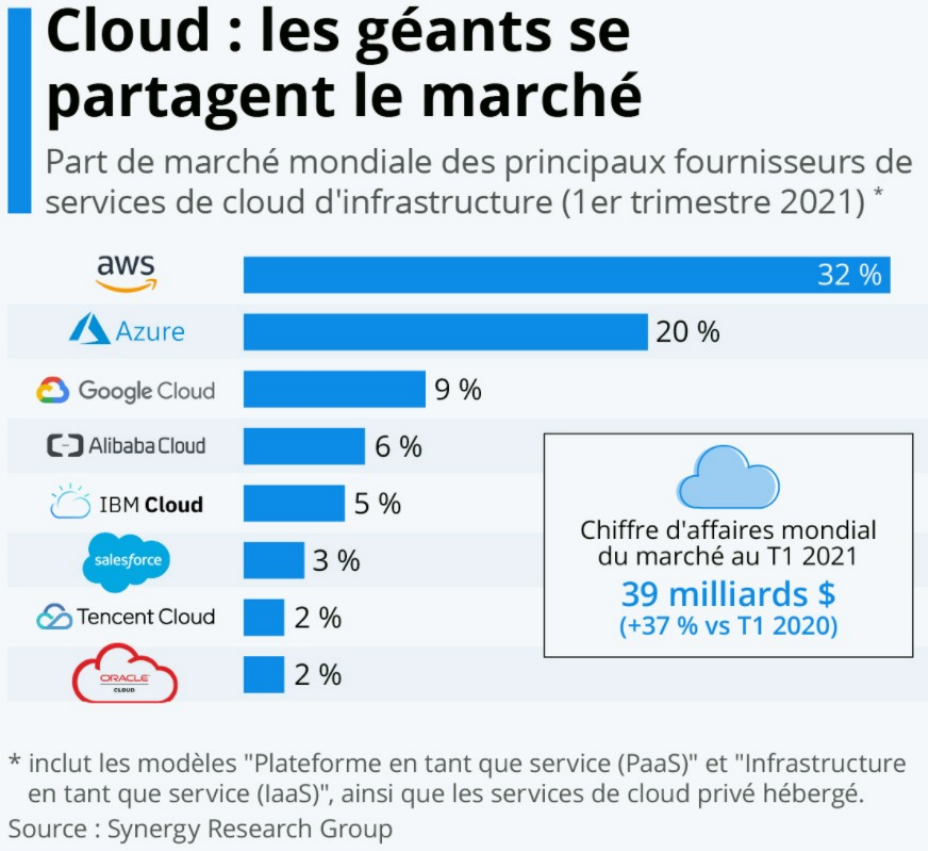


IV. Spark

- Framework de **calcul distribué**



V. Amazon Web Services

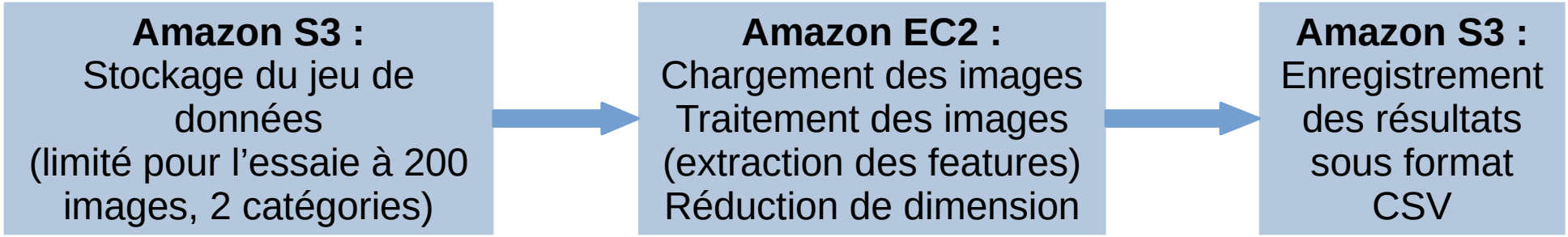


V. Amazon Web Services

- **Amazon S3 :**
  - Service de stockage illimité
- **Amazon EC2 :**
  - Service Web de calcul sécurisé et redimensionnable dans le cloud
  - Choix de processeur, de stockage, de système d'exploitation etc.

VI. Chaîne de traitement des images

- **Principales étapes :**



VI. Chaîne de traitement des images

- **Étapes détaillées :**
  - Création d’un SparkSession
  - Création d’un SparkContext (coordonner l’exécution des tâches Spark sur le cluster)
  - Récupération des chemins d’accès des images avec boto3
  - Création d’un DataFrame Spark
  - Récupération des catégories de fruits à partir du chemin d’accès

```
+--------------------+-----------+
|            img_path|   category|
+--------------------+-----------+
|Lemon Meyer/0_100...|Lemon Meyer|
|Lemon Meyer/100_1...|Lemon Meyer|
```

VI. Chaîne de traitement des images

- **Étapes détaillées :**
  - Récupération du « Body » des chemins d’accès
  - Extraction des features avec ResNet50 pré-entrainé sur imagenet
  - 2048 features
  - Vectorisation


```
+--------------------+----------+--------------------+
|            img_path|  category|        features_udf|
+--------------------+----------+--------------------+
|Lemon Meyer/259_1...|Lemon Meyer|[1.67463469505310...|
|Lemon Meyer/225_1...|Lemon Meyer|[1.47705852985382...|
|Lemon Meyer/251_1...|Lemon Meyer|[1.62330150604248...|
|Lemon Meyer/r_133...|Lemon Meyer|[1.92805790901184...|
|Lemon Meyer/r_149...|Lemon Meyer|[2.15801882743835...|
|Lemon Meyer/r_65_...|Lemon Meyer|[0.89454418420791...|
|   Lemon/261_100.jpg|      Lemon|[1.16088235378265...|
|   Lemon/215_100.jpg|      Lemon|[0.17953847348690...|
```

VI. Chaîne de traitement des images

- **Étapes détaillées :**
  - Standardisation des features (StandardScaler)
  - Réduction de dimension avec ACP (50 features retenus)

img_path	category	pcaFeatures_udf
Lemon Meyer/259_1...	Lemon Meyer	[-2.5144509406579...
Lemon Meyer/225_1...	Lemon Meyer	[-5.8734443735693...
Lemon Meyer/251_1...	Lemon Meyer	[-5.7569685933784...
Lemon Meyer/r_133...	Lemon Meyer	[0.22440108064295...
Lemon Meyer/r_149...	Lemon Meyer	[-0.0686469784765...
Lemon Meyer/r_65_...	Lemon Meyer	[0.39107897707380...
Lemon/261_100.jpg	Lemon	[27.2603984841592...
Lemon/215_100.jpg	Lemon	[19.7795667110322...

  - Création d’un bucket et enregistrement des résultats sous forme CSV après avoir convertir en DataFrame Pandas

V. Chaîne de traitement des images

- **Configuration EC2 :**
  - Instance t2.xlarge (CPU = 4, Mémoire = 16 Gio, EBS = 30 Gio)
  - Ubuntu Server 18.04
  - Python 3.8.8, Java 11.0.11, Spark 3.2 avec Hadoop 3.3
  - AWS CLI, clé IAM
  - Anaconda, Jupyter Notebook
  - Packages

VII. Conclusion

- **Mise en place d’un environnement Big Data :**
  - AWS (EC2, S3)
  - Spark
- **Recommandations pour traiter tous les images :**
  - Choix d’un instance EC2 plus puissant
  - Elastic Map Reduce
  - Pas de changements niveau code

VII. Conclusion

- **Autres recommandations :**
  - Amélioration du modèle de transfer learning
  - Preprocessing d'images plus « réaliste » (arrière plan non blanc, multiples fruits, main tenant le fruit etc.)
  - Amélioration du nombre de composants à retenir
  - Utilisation d'autres techniques de prétraitement (couleur, bruitage, égalisation de l'histogramme, redimensionnement etc.)
  - Entraînement d'un modèle de classification

Merci de votre attention