Déployer un modèle dans le cloud

Pré-traitement pour l’application ‘Fruits !’

Introduction des données

- Dimensions des images : 100x100px
- Images prêtes à l'emploi
- 90k images (67k pour l'entraînement)
- 131 fruits/légumes

Besoins

- **Traitements à réaliser :**
  - Featurisation
  - Réduction de dimension
- **Augmentation rapide du volume de données**

Réponse

- **Traitements à réaliser :**
  - Featurisation : ResNet50
  - Réduction de dimension : SVD
- **Données volumineuses :**
  - Serveur EC2
  - Serveur S3

Déroulement

- **Éléments de l’architecture Big Data**
- **Chaîne de traitement des données :**
  - Featurisation
  - Réduction de dimension
- **Démonstration**
- **Conclusion**

I- Éléments de l’architecture Big Data

- **Un serveur pour stocker les données**
  - Qui soit robuste aux pannes
  - Qui permette un transfert rapide des données
- **Un serveur pour faire les calculs**
  - Qui permette de s’adapter au volume de données

I- Serveur de stockage : S3 Standard

- **Avantages :**
  - Durable
  - Scalable
  - Pas de frais minimaux + faibles coûts
  - Faible latence
  - Proximité avec les serveurs de calcul
- **Inconvénients**
  - Lecture aléatoire

I- S3 : Connexion

- **Connexion avec un compte IAM**
  - **Configuration sur terminal** : aws configure
- **En ligne de commande :**
  - aws s3 cp fichier_local s3://bucket/fichier_local
- **Sur python :**
  - Avec boto

I- Serveur de calcul : EC2

- **Avantages :**
  - Instance redimensionnable
  - Coût d'exécution bas
  - Choix du système d'exploitation (AMI)
- **Inconvénients :**
  - Capacité de calcul non adaptable en temps réel
  - Coût du stockage élevé
  - Éphémère

Amazon EC2

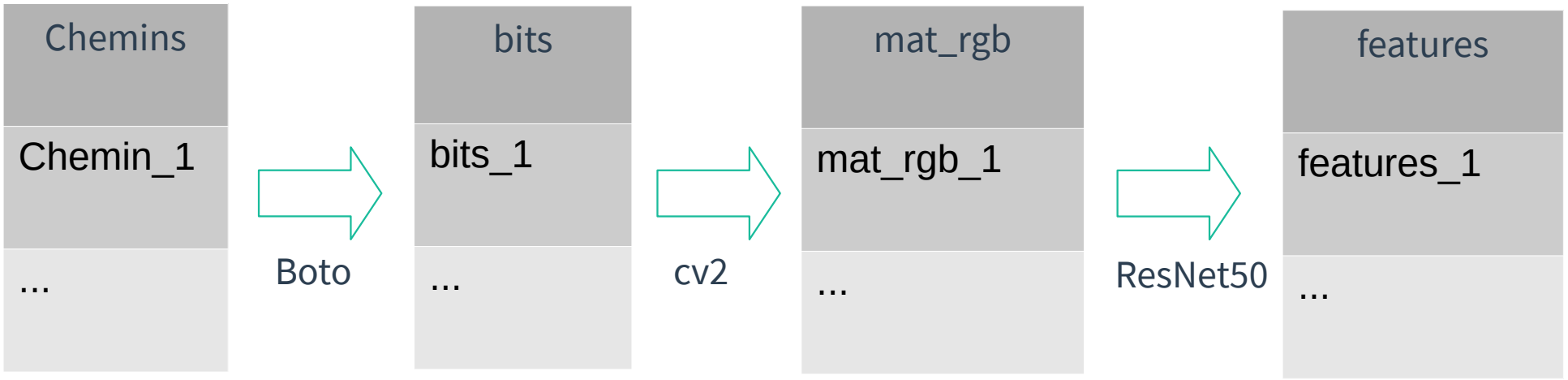
I- EC2 : Connexion

- **Système : Ubuntu**
- **Groupe de sécurité : Autorise toutes les connections ssh**
- **Avec Jupyter Notebook :**
  - **Sur EC2** : Jupyter notebook –no-browser –port=8080
  - **En local** : ssh -i clé_ec2 -L 8080:localhost:8080 ubuntu@dns_publique
  - **Ouvrir la page** : http://127.0.0.1:8080/?token=...

I- EC2 : Instance choisie

- **T2-medium**
  - 4 Go de RAM
  - 8 Go de mémoire système
  - 2 vCPU

II- Featurisation : ResNet50



Nombre d'images : 5
Nombre de variables : 32768

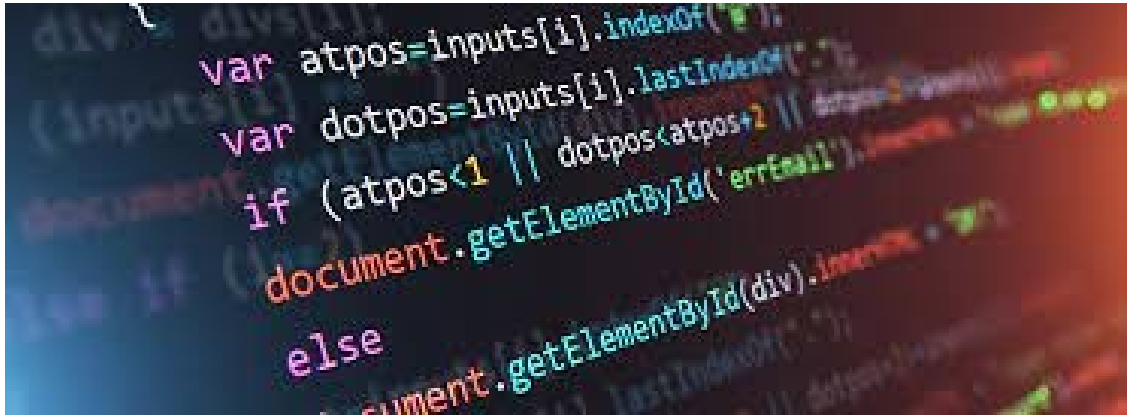
III- Réduction de dimension : SVD

$$A = U \Sigma V^T$$

U : m×k,
Σ: k×k,
V: n×k.

- **Outils pyspark**
  - RowMatrix
  - Compute SVD
- **Nombre de composantes**
  - 2

IV- Démonstration



Conclusion

- Utilisation des serveurs S3 et EC2
- Featurisation avec ResNet50
- Réduction de dimension avec SVD
- Parallélisation des calculs avec Pyspark
- Temps d’exécution pour 5 images et 2 composantes : 45 s

Améliorations

- **Utiliser un EMR pour traiter encore plus de données**
- **Calculer le nombre maximal de composantes avec la SVD**
  - Permet de connaître le % de la variance conservée
- **Utiliser pyspark pour la lecture de fichier**
  - Nécessite une bonne maîtrise des versions des logiciels