MU5IN852

Bases de Données
Large Echelle

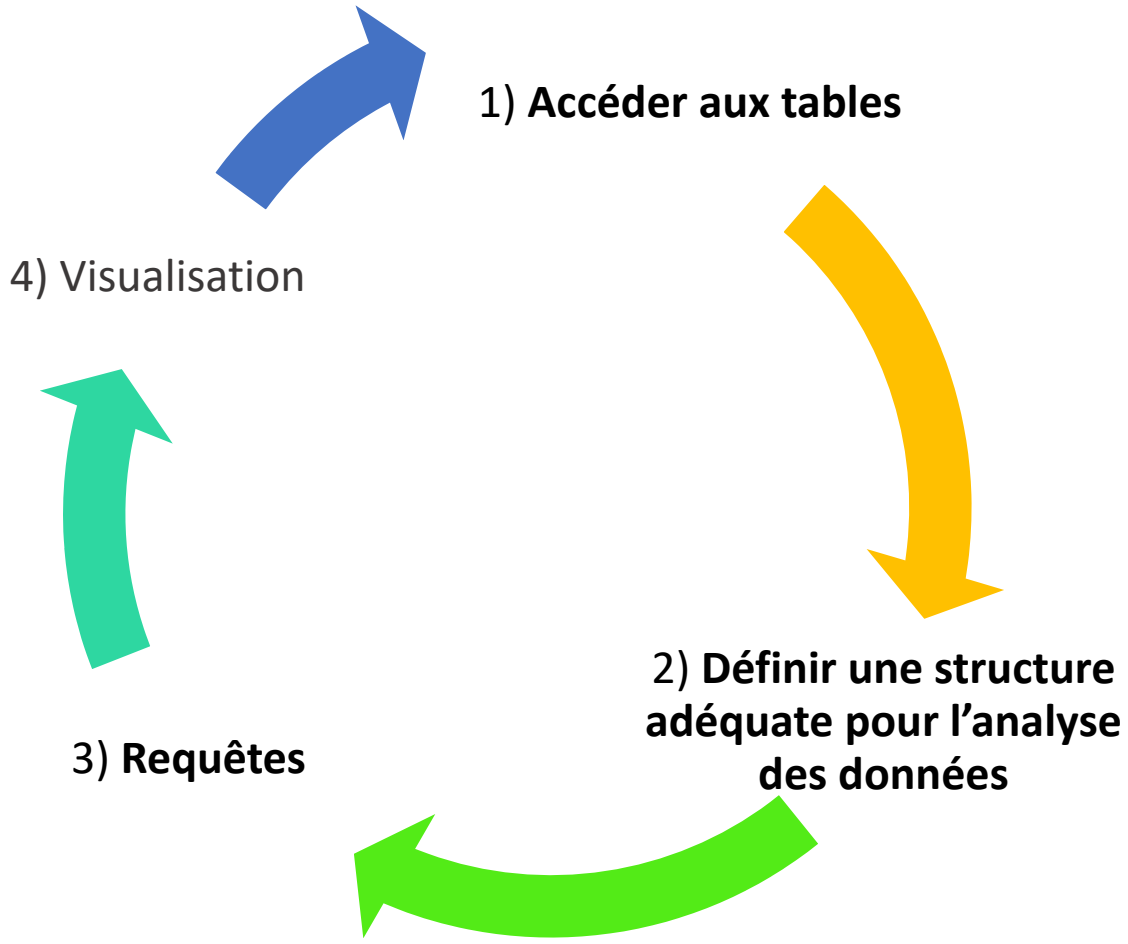
TME avec databricks

septembre 2021

hubert.naacke@lip6.fr

But des TME BDLE

Objectifs des séances



Savoir faire : outils

shell	Cloud	notebook	pyspark
• grands fichiers	• import/export • partage	• Jupyter • Colab • Databricks	• SQL avancé • viz mplotlib, seaborn, …

Savoir faire : autonomie en TP

Détecter

- Comportement imprévu
- Résultat incorrect

Isoler

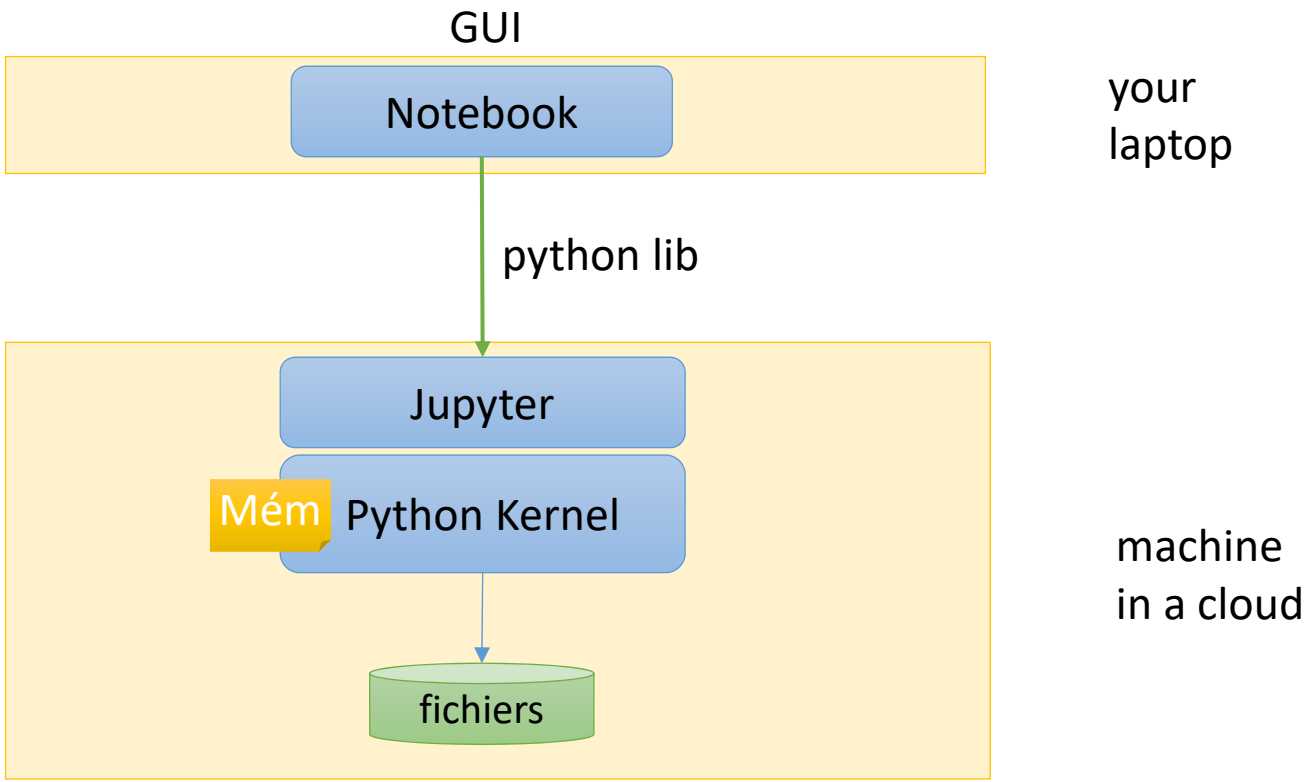
- Etat **avant** le problème
- Etat simplifié

Résoudre

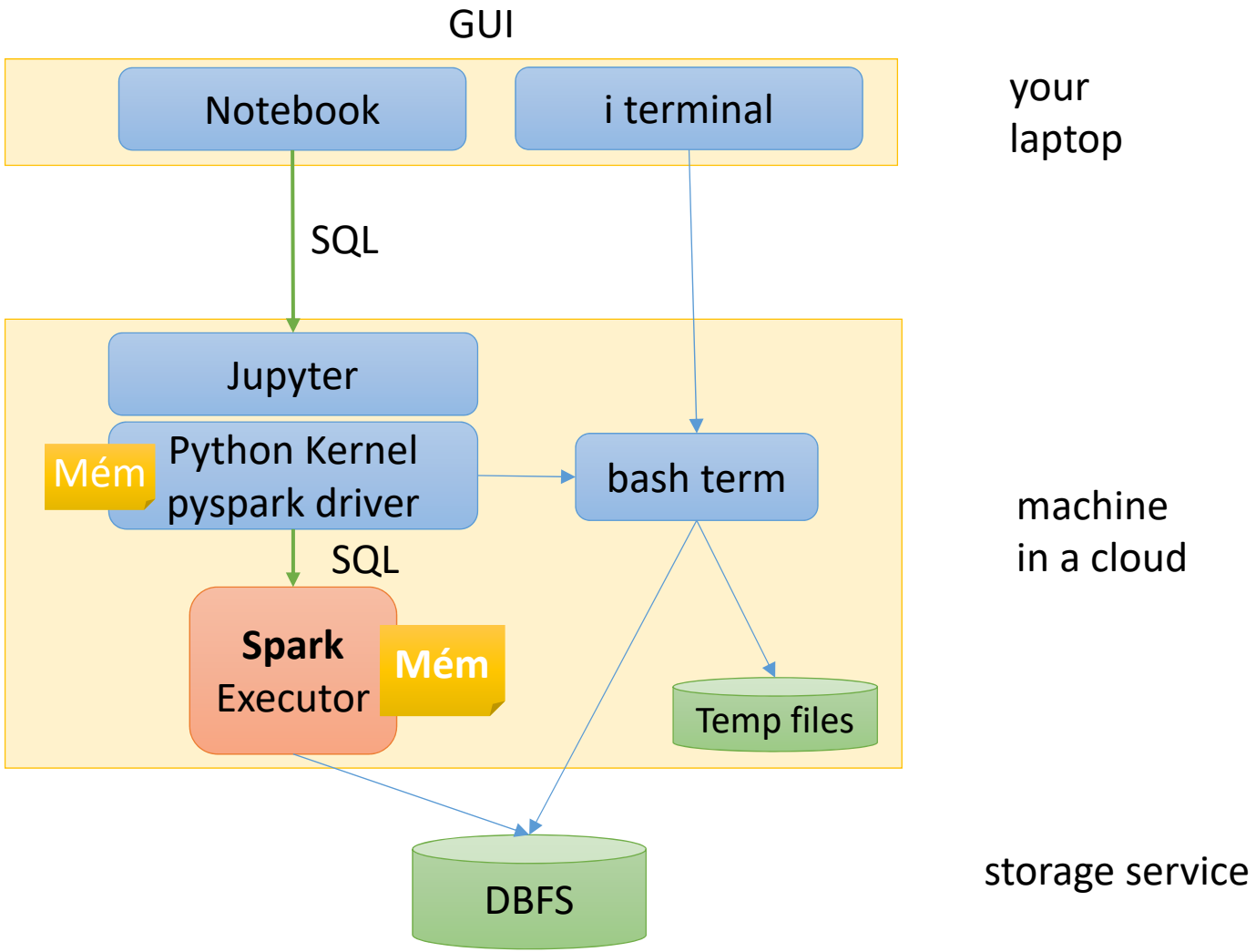
- Reproduire
- Comprendre

TP avec un notebook : Comparaison Colab / Databricks

- Architecture d'un notebook colab



Architecture databricks en TME



TP avec notebook databricks

Outil scalable

- Plateforme Spark
  - Moteur SQL
    - Databricks community : community.cloud.databricks.com
  - Interface : Notebook mixant SQL et python
    - Supporte la syntaxe SQL : tag %sql
    - Le résultat d'une requête SQL est manipulable en python

TP avec notebook databricks

- Lire les indications sur la page de l'UE BDLE
  http://www-bd.lip6.fr/wiki/site/enseignement/master/bdle/tmes/databricks
- Importer les datasets dans Databricks
- Importer un notebook avec des exemples
- Obtenir des ressources de calcul
  - Démarrer un cluster
  - Associer un notebook à un cluster
- Se familiariser avec l'interface utilisateur du notebook

Datasets

- Dossier partagé : PUBLIC_DATASET
  https://nuage.lip6.fr/s/H3bpyRGgnCq2NR4
- Fichiers IMDB dans le dossier
  - imbd/vldb2015/csvfiles_sample001

- Contient aussi d'autres datasets à disposition

Biblio

- **Spark SQL : Relational Data Processing in Spark**
  - **in SIGMOD 2015**
  - Databricks, MIT CSAIL, UC Berkeley AmpLab

    https://people.csail.mit.edu/matei/papers/2015/sigmod_spark_sql.pdf

    https://dl.acm.org/doi/pdf/10.1145/2723372.2742797
- Site de l'équipe BD
  - https://www-bd.lip6.fr/wiki/site/enseignement/start
  - Requêtes avancées sur données semi-structurées
    - UE Modèles et langages pour les BD avancées
      - M1 4IN801