Résumé des notions clés des tutoriels PySpark d’Apache

Ce document présente une synthèse fidèle mais non verbatim des tutoriels officiels Apache Spark:
- Quickstart DataFrame
- Quickstart Spark Connect
- Quickstart Pandas API on Spark
- Testing PySpark

1. Création de DataFrames
Les DataFrames sont créés depuis des listes Python, des schémas, ou des sources de données externes. Ils sont dis

2. Visualisation des données
La méthode `show()` permet d'afficher un aperçu. D'autres méthodes comme `printSchema()` renseignent la structure

3. Sélection et accès aux données
Les opérations courantes incluent `select`, `filter`, `where`, l'accès par colonne et l'utilisation d'expressions Spark (`F.c

4. Application de fonctions
Spark permet l'application de fonctions via `withColumn`, les UDF, ou les fonctions SQL intégrées.

5. GroupBy et agrégations
Les groupes permettent des opérations comme `count()`, `sum()`, `mean()`, etc.

6. Entrée / Sortie des données
Spark lit et écrit dans divers formats: CSV, JSON, Parquet, ORC, tables SQL, etc.

7. SQL avec Spark
Le moteur SQL intégré permet d'enregistrer des DataFrames en tables temporaires et d’exécuter des requêtes SQL s

8. Spark Connect
Spark Connect sépare le client du serveur Spark.
- On initialise un serveur Spark.
- On se connecte via un client Python.
- Les DataFrames sont manipulés de façon similaire.

9. Pandas API on Spark
Interface compatible Pandas mais distribuée:
- Création de DataFrames en style Pandas
- Opérations similaires (sélection, groupby, plots)
- Gestion des données manquantes

10. Construction d’une application PySpark
Les applications suivent une structure standard: création de session, définition de transformations, écriture des résulta

11. Tests PySpark
Apache recommande `pytest` avec une session Spark locale.