DOCUMENTATION

Ce code Python utilise PySpark pour effectuer le traitement de données de transport public. Voici un résumé des étapes clés :

1. Il configure la session Spark et l'accès au stockage Azure.

2. Il lit un fichier CSV depuis Azure Data Lake Storage Gen2.

3. Il effectue diverses transformations sur les données, notamment la conversion des heures, la catégorisation des retards, et le calcul de la durée du voyage.

4. Il groupe les données par heure de départ et calcule la moyenne des passagers pour chaque heure.

5. Il catégorise les heures de départ en "Heure de Pointe" ou "Heure Hors Pointe" en fonction d'un seuil de passagers.

6. Les données transformées sont écrites dans Azure Data Lake Storage Gen2.

7. Une boucle vérifie les fichiers bruts non traités, les traite, et limite le traitement à un certain nombre de fichiers ou jusqu'à la fin de la liste.

Ce code permet de traiter efficacement des données de transport public stockées dans Azure en utilisant PySpark et Azure Databricks.