

L'Optimisation & Chez Spache & Chez Spache & Chez Spache & Chez Spache & Chez &



DONNEES DISTRIBUEES

Spark est un framework open source de traitement de données distribué. Il s'agit d'un ensemble d'outils et de composants logiciels développés pour fournir une alternative rapide et généralisée à MapReduce, le modèle de traitement de données associé à Apache Hadoop.

APACHE SPARK



GRANDE ECHELLE

Spark est conçu pour le traitement de données à grande échelle, permettant de traiter des volumes massifs de données en les distribuant sur un cluster de machines.



NOMBREUX SECTEURS

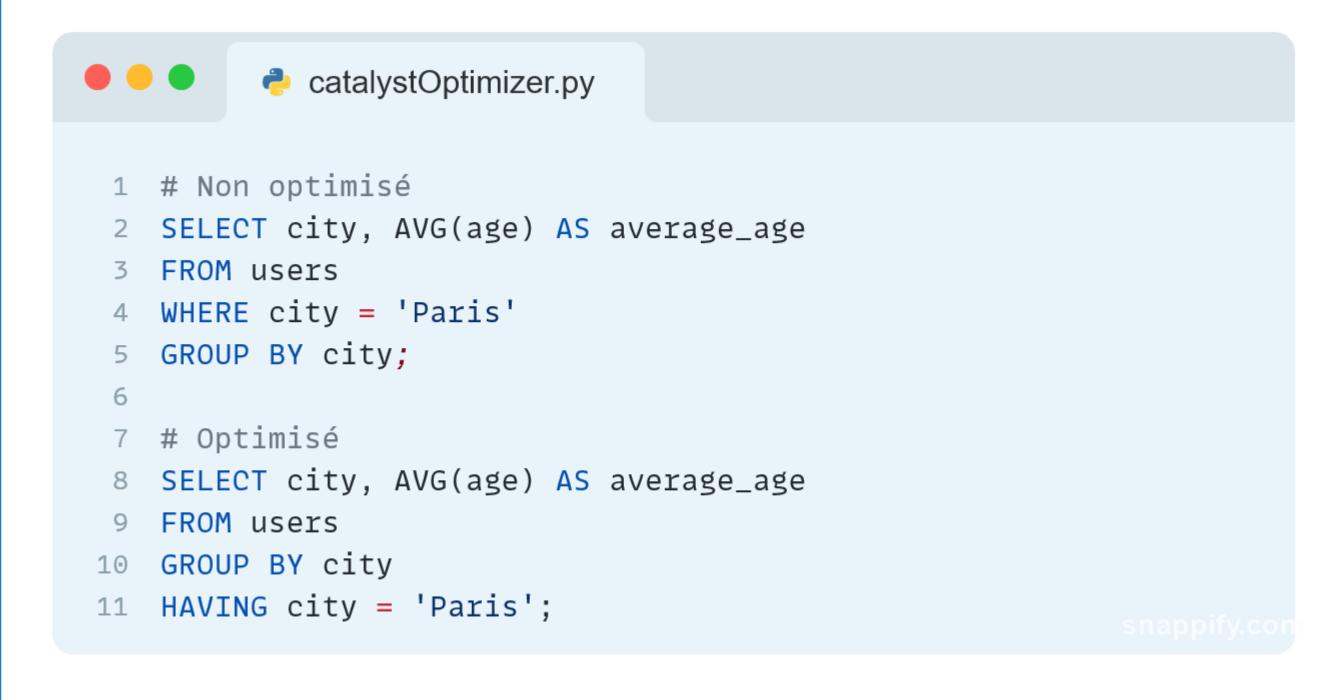
L'utilisation de Spark est répandue dans de nombreux secteurs pour des applications telles que l'analyse de données, le traitement de flux en temps réel, le machine learning et plus encore. Sa polyvalence et sa rapidité en font un choix très populaire chez les développeurs.

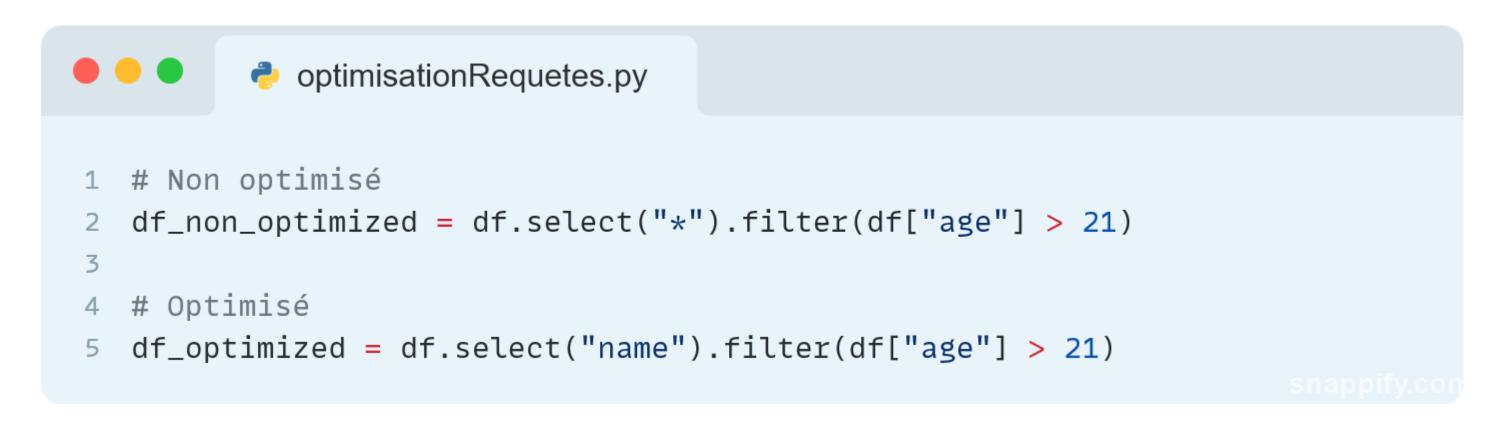
EXEMPLES

OPTMISATION DES REQUÊTES : Imaginons que l'on dispose d'une base donnée contenant les toutes les informations associées à des utilisateurs et que l'on souhaite récupérer le nom de tous ceux âgés de plus de 21 ans,

nous pourrions le faire de deux manières différentes.

Dans cet exemple, la version optimisée sélectionne uniquement les colonnes nécessaires avant d'appliquer le filtre. Cela réduit la quantité de données traitées, améliorant ainsi les performances.





CATALYST OPTMIZER: dans un premier temps nous filtrons les utilisateurs de la ville de Paris avant de calculer la moyenne d'âge. Il est possible de réécrire cette requête de manière plus efficace en effectuant la filtration après le calcul de la moyenne. Cela peut réduire le nombre d'enregistrements à considérer lors du calcul de la moyenne et donc être plus performant. Ce genre d'optimisation peut facilement être repéré par le Catalyst Optimizer.

METHODES D'OPTMISATION





Master II – Ingénierie Logiciel Année 2023 – 2024

FOUILLEN Mathis, GOMBERT Gwenaël RASEMONT Florian, RIBEIRO GOMES Lise

