Apache Arrow

Robin Delbos INFO5

Sommaire

Introduction

- Apache Arrow a été Crée en octobre 2016.
- Communauté très active (dernière version stable datant du 19 octobre).

Plusieurs bibliothèques disponibles

- C
- C++
- C#
- Go
- Java
- JavaScript
- Rust
- MATLAB (via la librairie C++)
- Python (via la librairie C++)
- R (via la librairie C++)
- Ruby (via la librairie C++)

Qu’est ce que Apache Arrow

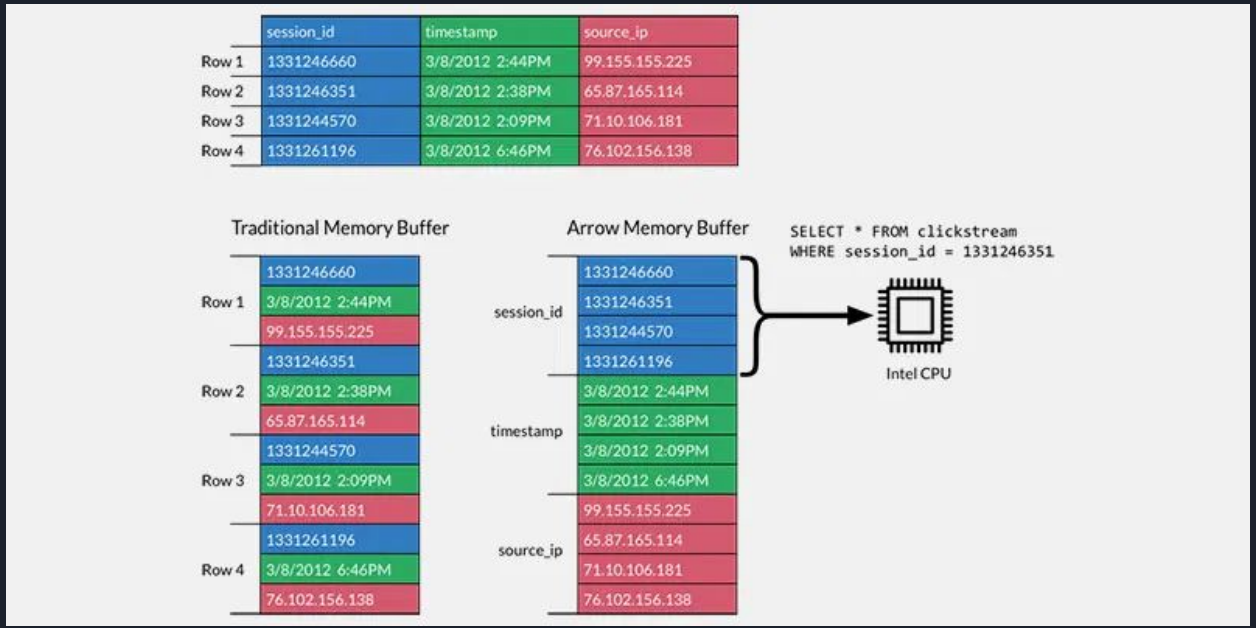
- Apache Arrow est une plateforme de développement pour l'analyse en mémoire.
- Elle spécifie un format de mémoire en colonnes standardisé, indépendant du système ou langage.

Avec Arrow


Pandas
Drill
Spark
Impala
Arrow Memory
Parquet
HBase
Cassandra
Kudu

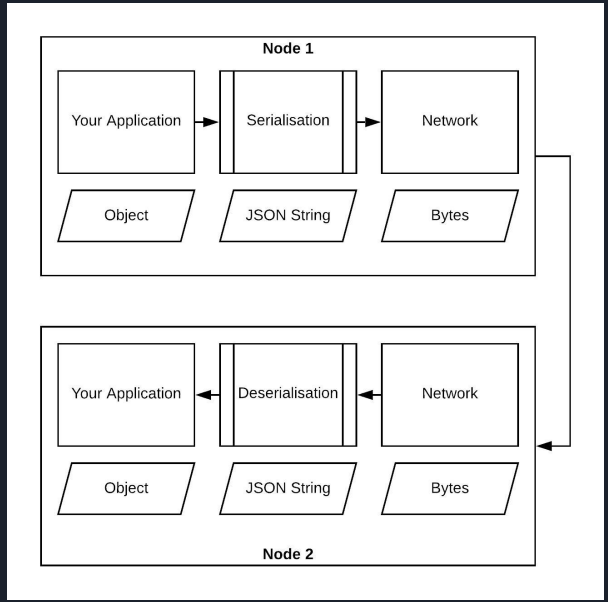
Le format de Apache Arrow

- Format standardisé en colonne.
- Sans ce format, chaque base de données et chaque langage doit mettre en œuvre son propre format de données interne.



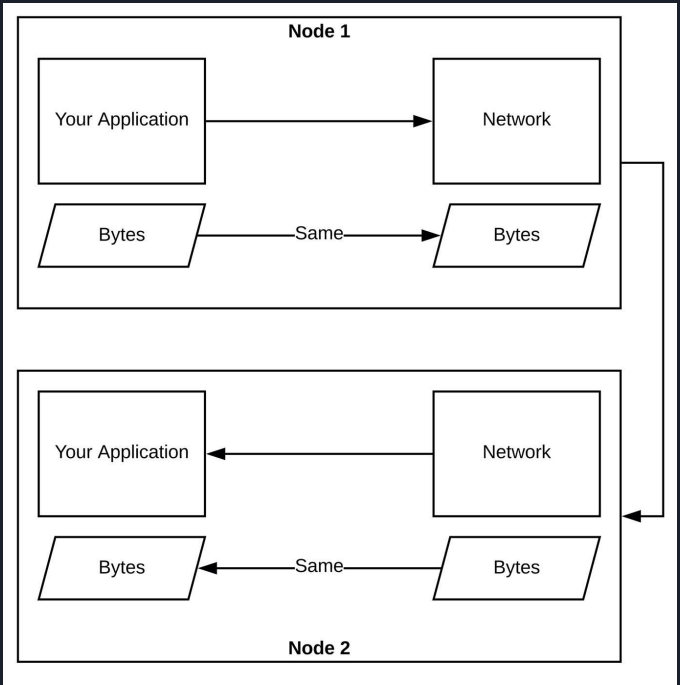
Sans utiliser Arrow

- Le déplacement des données d'un système à l'autre implique une sérialisation et une désérialisation coûteuses.



En utilisant Arrow

- Transfert de données sans copie.
- Accès et un échange de données rapides.



Autres alternatives

- Apache Parquet et Apache ORC sont des exemples populaires de formats de données en colonnes sur disque.
    - Les fichiers Parquet et ORC sont conçus pour le stockage sur disque, Arrow est conçu pour le stockage en mémoire.

- Pandas

Démonstration

Questions ?