



Apache Arrow

Robin Delbos INFO5



Sommaire

- Introduction 3
- Qu'est ce que Apache Arrow 4-8
- Autres alternatives 9
- Démonstration 10
- Questions 11



Introduction

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

APACHE

ARROW





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



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.

	session_id	timestamp	source_ip
Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138

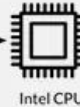
Traditional Memory Buffer

Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138

Arrow Memory Buffer

	session_id	timestamp	source_ip
Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138

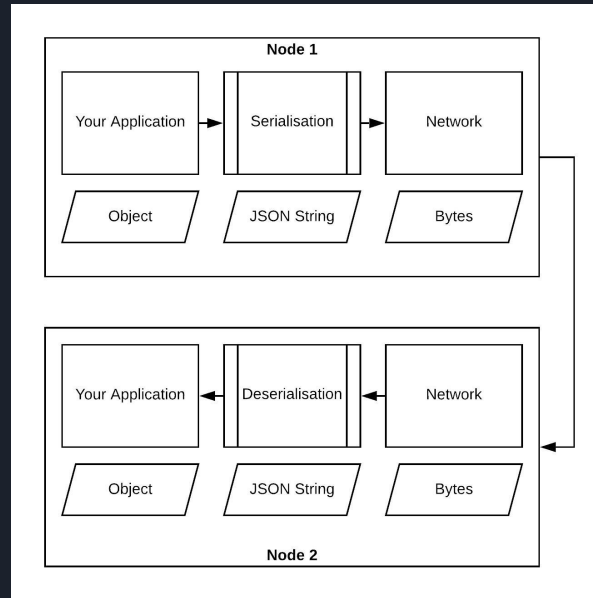
```
SELECT * FROM clickstream  
WHERE session_id = 1331246351
```



Intel CPU

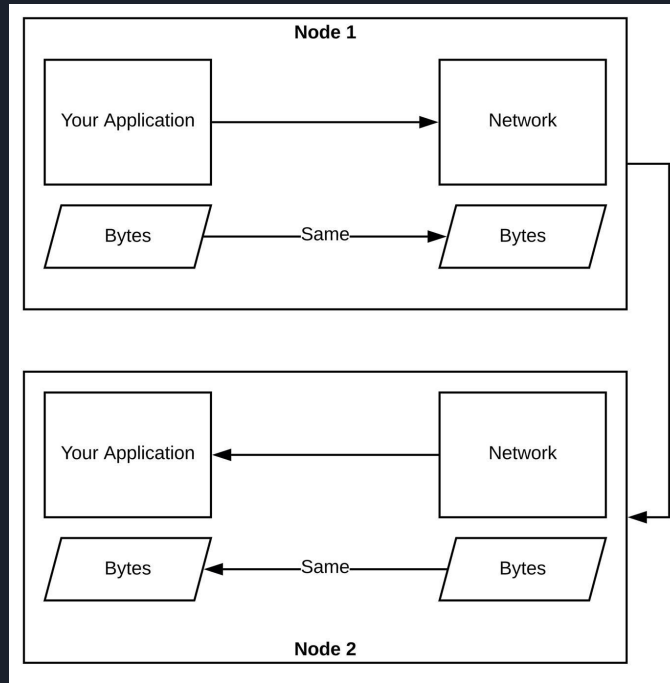
Sans utiliser Arrow

- 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 ?