Using Arrow and DuckDB for data engineering

Anónimo

Abstract This talk will provide an overview and code examples of how to use Apache Arrow and DuckDB to analyze very large datasets on regular computers. It will highlight how R users can leverage their knowledge of the tidyverse to work with datasets that don't fit in memory quickly and easily. This talk will also feature how to leverage the Arrow standards to develop interoperable data pipelines that use other languages such as Python.

Palabras clave: Data Engineering - Apache Arrow - DuckDB - Big Data - Apache Parquet

The data analysis landscape is changing as datasets are growing increasingly larger. However, recent advances in technologies such as Apache Arrow and DuckDB are making accessible to anyone the analysis of datasets that used to require complex infrastructure. Using the {arrow} and {duckdb} packages opens up the door to analyzing gigabytes of data in second using the same interface as with the {tidyverse}. By learning just a few concepts, R users can enjoy working easily and efficiently with very large datasets directly from their everyday computer. Additionally, because the underlying data structures used by Arrow and DuckDB are based on standards implemented in multiple languages, these technologies make it straightforward to collaborate with others, or to integrate them in more complex data pipelines.

In this talk, I will present an overview of the Apache Arrow standards, and how R users can benefit from adopting it to analyze easily large datasets. We will use the example of a real dataset analysis (with code made available on GitHub) to explore data formats used to store these large datasets on disks (e.g., Parquet), how the {arrow} and {duckdb} packages can be leveraged to analyze data, and how these tools integrate with the already familiar {tidyverse} interface. We will explore how to take advantage of the compatibility with other tools and languages to build data pipelines. Finally, I will demonstrate some of the geospatial features these tools support to illustrate the versatility of these tools.