

# Parquet files for large (and big) data dissemination

Romain Lesur

INSEE (France)

2023-11-22

# We tried **Parquet** for bulk data dissemination...



*Detailed results of the Census requested with DuckDB*

**...and our users loved it!**

# What is **Parquet**?

- A **big data file format** created in 2013
- **Open source** (Apache Foundation)
- Optimised for analytics workloads (OLAP)
  - Columnar-oriented
  - Efficient compression



# Parquet



# Parquet files are **easy to create**

A file format supported by a large ecosystem



**DuckDB**



Create a Parquet file with a simple command!  
available in R, Python, Rust, Java, JavaScript,  
C#, C++...

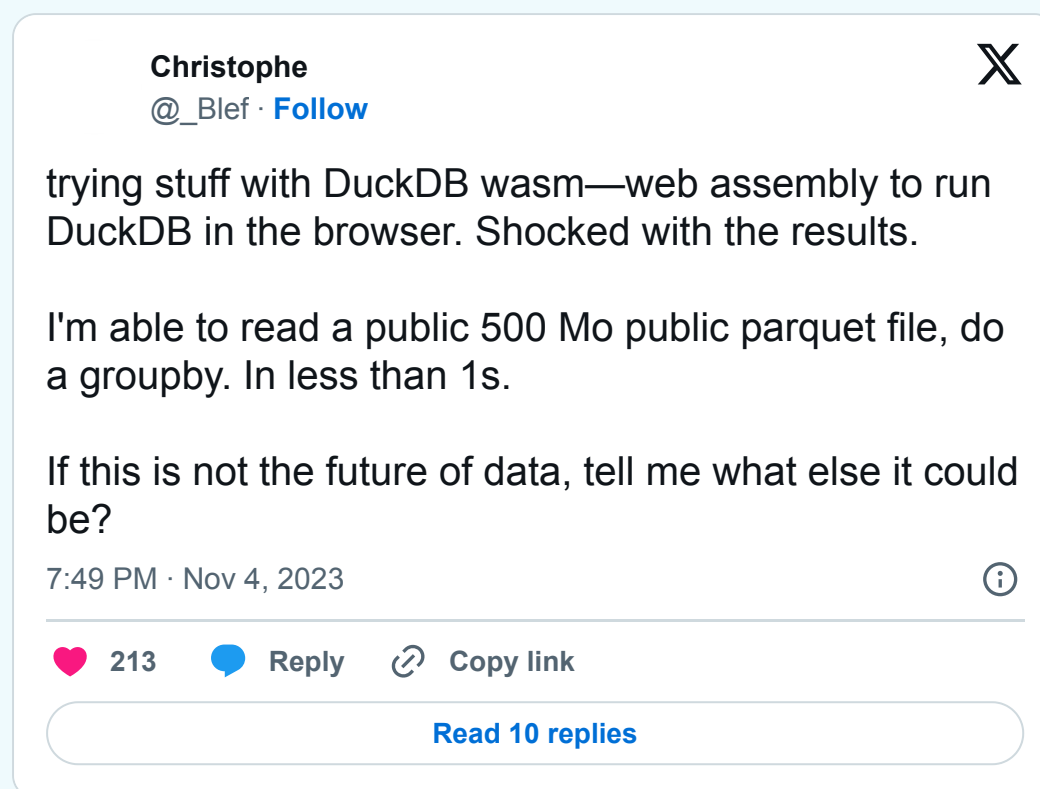
# Parquet files are **easy to use**

## Requests are **efficient**

*Only the required chunks of data are read*

## With **DuckDB**, Parquet files can be requested **over the web**

*No need to download the whole file!*



# Parquet as a **default** file format for dissemination?

- Parquet can handle any size of data (small, large and big)
- Adopted as INSEE's internal default file format, replacing SAS format (SAS7BDAT)
- Our experiment for bulk data dissemination was a great success!

**Is it the future of open data?**

*FMI, read Robin Linacre's post “**Why parquet file are my preferred API for bulk open data?**”*