On the Performance of HuggingFace Storage layer and the latest release of it

By Quentin Lhoest, Aug 12, 2025

Let me explain why Hugging Face Datasets storage is faster than AWS S3 and why today's release changes everything.

Hugging Face uses Xet: a dedupe-based storage which enables fast deduped uploads.

Unlike traditional remote storage, uploads are faster on Xet because duplicate data is only uploaded once.

For example: if some or all of the data already exists in other files on Xet, it is not uploaded again, saving bandwidth and speeding up uploads. Deduplication for Parquet is enabled through Content Defined Chunking (CDC).

Thanks to Parquet CDC and Xet deduplication, saving a dataset on Hugging Face is faster than on any traditional remote storage.

This is especially efficient for appending rows or columns, and also for insert/delete operations (provided they don't change every single block of data).

Today I released V2 of the pyspark\_huggingface package, a Spark Data Source to read/write HF datasets.

This release is huge: it includes Parquet CDC and Xet support. Accelerating dataset saving on HF dramatically

This is a new paradigm for Spark, for Parquet, for data engineers. You don't want to miss this great evolution.

Write the parquet data locally (disk or ram) and only save the new data to remote storage after hashing and asking the server which data is new.

# References

[1] HuggingFace storage backends official document: <https://huggingface.co/docs/hub/en/storage-backends>

[2] Parquet Content-Defined-Chunking: <https://huggingface.co/blog/parquet-cdc>

[3] <https://github.com/huggingface/pyspark_huggingface>

[4] From Files to Chunks: Improving HF Storage Efficiency: <https://huggingface.co/blog/from-files-to-chunks>

[5] From Chunks to Blocks: Accelerating Uploads and Downloads on the Hub : <https://huggingface.co/blog/from-chunks-to-blocks>

[6] <https://github.com/huggingface/xet-core>

[7] [Git is for data , xethub blog, CIDR 2023](https://xethub.com/blog/git-is-for-data-published-in-cidr-2023)

[8] [Git is for data, Y. Low et al, XetData, CIDR 2023](https://github.com/dimitarpg13/rag_architectures_and_concepts/blob/main/docs/Git_is_for_data_Low_CIDR_2023.pdf)