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).

# References

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