Cloud storage

Hugging Face Datasets

The Hugging Face Dataset Hub is home to a growing collection of datasets that span a variety of domains and tasks.

It's more than a cloud storage: the Dataset Hub is a platform that provides data versioning thanks to git, as well as a Dataset Viewer to explore the data, making it a great place to store AI-ready datasets.

This guide shows how to import data from other cloud storage using the filesystems implementations from `fsspec`.

Import data from a cloud storage

Most cloud storage providers have a `fsspec` FileSystem implementation, which is useful to import data from any cloud provider with the same code.
This is especially useful to publish datasets on Hugging Face.

Take a look at the following table for some example of supported cloud storage providers:

Storage provider	Filesystem implementation
Amazon S3	s3fs
Google Cloud Storage	gcsfs
Azure Blob/DataLake	adlfs
Oracle Cloud Storage	ocifs

This guide will show you how to import data files from any cloud storage and save a dataset on Hugging Face.

Let's say we want to publish a dataset on Hugging Face from Parquet files from a cloud storage.

First, instantiate your cloud storage filesystem and list the files you'd like to import:

```
>>> import fsspec
>>> fs = fsspec.filesystem("...")  # s3 / gcs / abfs / adl / oci / ...
>>> data_dir = "path/to/my/data/"
>>> pattern = "*.parquet"
>>> data_files = fs.glob(data_dir + pattern)
["path/to/my/data/0001.parquet", "path/to/my/data/0001.parquet", ...]
```

Then you can create a dataset on Hugging Face and import the data files, using for example:

```
>>> from huggingface_hub import create_repo, upload_file
>>> from tqdm.auto import tqdm
>>> destination_dataset = "username/my-dataset"
>>> create_repo(destination_dataset, repo_type="dataset")
>>> for data_file in tqdm(fs.glob(data_dir + pattern)):
...     with fs.open(data_file) as fileobj:
...         path_in_repo = data_file[len(data_dir):]
...         upload_file(
...             path_or_fileobj=fileobj,
...             path_in_repo=path_in_repo,
...             repo_id=destination_dataset,
...             repo_type="dataset",
...         )
```

Check out the huggingface_hub documentation on files uploads here if you're looking for more upload options.

Finally you can now load the dataset using 🤗 Datasets:

```
>>> from datasets import load_dataset
>>> ds = load_dataset("username/my-dataset")
```