Fetching and Processing NYC Yellow Taxi Data Using EMR, MobaXterm, and S3.

**Amazon EMR Cluster:** A cloud tool to process big data.

**MobaXterm:** A program to connect to my EMR cluster.

**Amazon S3:** A cloud storage place to save my data.

**Website Data:** The TLC data at https://d37ci6vzurychx.cloudfront.net/trip-data/ for Parquet files.

In MobaXterm CLI:

* I created a folder and moved into it on the EMR master node: **mkdir nyc\_taxi\_data**

**cd nyc\_taxi\_data**

* I downloaded all 12 months of 2017 Parquet files at once using a loop**:**

**for month in {01..12}**

**do**

**wget https://d37ci6vzurychx.cloudfront.net/trip-data/yellow\_tripdata\_2021-$month.parquet**

**done**

This fetched files like yellow\_tripdata\_2021-01.parquet to yellow\_tripdata\_2021-12.parquet.

* I checked the files with: **ls**
* I uploaded all files from the nyc\_taxi\_data folder to S3:

**aws s3 cp . s3://grp-bucket101/nyc-taxi/2017/ --recursive**

I verified the upload in the AWS S3 console under **s3://proj-data-grp-3/nyc-taxi/2021/.**

