Data Engineering – Interview Exercise

HENI wants to have a data pipeline that can ingest blockchain transactions and produce datasets for NFT analytics dashboard. As part of the data engineering team, you are responsible for developing this pipeline using PySpark for cleansing and transforming the data to produce a dataset that will be used by the application front end.

The data are in file data.parquet inside the ZIP file, with columns:

* load\_timestamp – timestamp when the record is arrived from the source
* transaction\_timestamp – timestamp when the transaction is completed
* token\_id – NFT token ID
* from\_address – seller address
* to\_address – buyer address
* amount – transaction value

Your tasks are:

* Remove duplicates of the blockchain transaction based on the key columns: transaction\_timestamp, from\_address, and to\_address
* Aggregate the de-duplicated data across all files by NFT\_TOKEN to get
  1. Daily top-5 NFT tokens with the highest amount
  2. Daily top-5 NFT tokens with the accumulated transaction amounts
* Write these two aggregated data separately in 2 parquet files with columns:
  1. date, token\_id, amount
  2. date, token\_id, accumulated\_amount
* The process should be idempotent
* Code reusability is important when writing production pipelines, consider how you would make your code modular and reusable
* Testing is critical for code-quality, consider the best strategy to test your code

# Output requirements

* Please provide README.md to explain on how to install the dependencies, run the tests, and run the entire pipeline from start to finish
* Consider on how you would use linting and type-checking in your solution
* Please provide all required files and supporting documents in a ZIP file. Please save this to either Dropbox, Google Drive, or Microsoft OneDrive. And then share via email the details of how to access the file.

You should spend 2-3 hours on this exercise. If you run short of time, please detail what you would have done if you had more time in the supporting document.

Good luck!