Data Engineer

Thank you for taking the time to solve Jua.ai's Geospatial Data Engineering assessment. This challenge is intended to give us deeper insight into your skill level and approach to problem-solving.

At Jua, we're extensively working with multi-dimensional data about the weather. Most of this data is stored as multidimensional arrays in NetCDF or HDF5 files, which are quite good for storing and consuming huge amounts of data but aren't as good for data analysis.

The assessment aims to check your skills in geospatial data transformation and analysis.

Source data for the task can be downloaded from the ARCO data set, which is hosted here: gs://gcp-public-data-arco-era5/raw/date-variable-single_level.

Using the provided data, please create a pipeline to transform the **year 2022 total precipitation** data into the Apache Parquet dataset, supporting the following features:

- Regular queries with filtering by timestamp.
- Filtering by the hierarchical geospatial index of your choice, i.e. H3 (you will need to add the index to ERA5 data).

Please provide the example queries for the created dataset as a part of your submission.

You can use any open-source tools and frameworks. We can, however, ask you to explain your choice.

Requirements and submission guidelines

- The solution can be implemented in any modern programming language.
- Code and commit messages should be treated the same way when writing production code.
- Please consider code quality and readability.
- Provide a README with instructions on how to set up, run, and test your app.
- Submission can be a ZIP file or repository hosted in a public location (e.g., Github).

You're welcome to approach this task creatively and have some fun.

Thank you, and enjoy!