Use parquet for storing the dataset #2

New issue

Open

rth opened on Aug 14, 2020 · edited by rth — Edits

Thanks for making this historical data available!

It might be worth switching to parquet (supported in pandas), the dataset would be much smaller and faster to load:

- historique_stations.csv: 1.5GB and 424MB zip compressed
- historique_stations.parquet (with snappy compression): 78 MB and probably less after converting dates and GPS coordinates to correct dtype

Stored with pandas.DataFrame.to_parquet,

```
df.to_parquet("historique_stations.parquet", compr
```

This would require adding pyarrow as a dependency

rth changed the title ~~Use parquet for for storing the dataset~~ Use parquet for storing the dataset on Aug 14, 2020

Assignees

No one assigned

Labels

No labels

Projects

No projects

Milestone

No milestone

Relationships

None yet

Development

Code with agent mode

No branches or pull requests

Participants