National Research Data Infrastructure Conference

1. NFDI4Energy Conference

10.5281/zenodo.10607895

Published: 02.02.2024

Decentralized energy data storages through an Open Energy Database Server

An ETL workflow for research databases

Florian Maurer[1] [https://orcid.org/0000-0001-8345-3889],
Jonathan Sejdija[1] [https://orcid.org/0009-0009-3615-4160], and
Volker Sander[1] [https://orcid.org/0009-0006-8850-8520]

[1]University of Applied Sciences Aachen, Aachen, Germany

Abstract: In the research domain of energy informatics, the importance of open data is rising rapidly. This can be seen as various new public datasets are created and published. Unfortunately, in many cases, the data is not available under a permissive license corresponding to the FAIR principles, often lacking accessibility or reusability. Furthermore, the source format often differs from the desired data format or does not meet the demands to be queried in an efficient way. To solve this on a small scale a toolbox for ETL-processes is provided to create a local energy data server with open access data from different valuable sources in a structured format. So while the sources itself do not fully comply with the FAIR principles, the provided unique toolbox allows for an efficient processing of the data as if the FAIR principles would be met. The energy data server currently includes information of power systems, weather data, network frequency data, European energy and gas data for demand and generation and more. However, a solution to the core problem - missing alignment to the FAIR principles - is still needed for the National Research Data Infrastructure.

Keywords: Open Data, Database, Time-series

1 Motivation

In the domain of energy system simulation, the need for open datasets for better parameterization of simulations and input data, as well as verification and evaluation of the simulations output is crucial[1]. Often the curation and download of datasets takes a lot of time and is done variously, as there is no common collection of existing datasets and how they can be used. While the access to such datasets is mostly free of charge, API keys with registration are often required. Datasets exist in various different formats which often also need proprietary programs to visualize and edit them. Often an ETL (Extract, Transform, Load) workflow can be used to create a local data warehouse for research data. This is needed for datasets which do not allow public redistribution [2], which does not correspond to the FAIR principles (Findable, Accessible, Interoperable, Reusable)[3]. (Re-)Sharing such datasets as public records on existing Research Data Management (RDM) tools like the Open Energy Platform [4] or Zenodo[5] is therefore not possible.

To work around this redistribution limitation, an ETL workflow is given to provide the dataset in a structured database to query and filter datasets efficiently for research. More specifically, the use of time series databases keeps the efficiency, even when working on large datasets. To reduce the needed accesses and the load on open data sets, such a database can be reused in various institutions across research, and create tools which query the database mid simulation to retrieve the latest records. A centralized self-hosted database also aggregates data for research related to energy and can be extended by further datasets, while providing a unified database access and language (SQL) to access the data conveniently. The presented tool focuses on different use cases than existing Research Data Management tools like Zenodo or Coscine, as it makes it possible to conveniently update and append to datasets in real time and query data in a structured way, which is often not possible in the source datasets.

2 Data sources

The target of the project is to include different valuable data sources, in a way which allows easy reanalysis and usage in simulation projects. As the database is hosted only inside the institutes network, other research data which can not be published for everybody can be used too, like the data from EEX which is available to buy. On the other hand the conversion tools to easily process files with bought access - if an organization has licensed access - can be made freely available. The most important datasets available include:

1. power plant data and generation capacities from Markstammdatenregister [1]
2. weather Data from ECMWF[2]
3. energy usage/generation data from ENTSO-E[3]
4. gas usage/import data from ENTSO-G[4]
5. power plants from Open-Power-System-Data[5]
6. cross Border Trading results from JAO[6]
7. grid information from SciGrid[7]
8. frequency data from 50Hertz
9. building topology from TABULA IWU[8]
10. wind turbine information
11. data from EEX (if paid access is available) [9]

3 Technology Stack

From a technology perspective, Python is used as a scripting language to execute the extraction and transformation of various data sources. While some datasets present only the current state at the time of access (for example the MaStR) or have a fixed time horizon, others can be updated on a daily or weekly base. If the database is empty because the ETL-processes are being run for the first time, all data will be extracted

[1] https://www.marktstammdatenregister.de/MaStR/Datendownload
[2] https://climate.copernicus.eu/climate-reanalysis
[3] https://transparency.entsoe.eu/
[4] https://transparency.entsog.eu/
[5] https://data.open-power-system-data.org/
[6] https://jao.eu
[7] https://www.power.scigrid.de/pages/downloads.html
[8] https://webtool.building-typology.eu/
[9] https://www.eex.com/en/market-data

starting from a pre-defined start date. Otherwise, the scripts will download the recent data from the data source by first selecting the latest date available in the database for the current data source and updating it until the most recent timestamp.

As a database, a PostgreSQL database is used with the open source extensions TimescaleDB and PostGIS. Without the consideration of the time axis, most databases would become slow due to the huge size of the primary index on the datetime column[6], particularly because of its rather complex maintenance needed for updates and ongoing inserts. Knowing about the time column as an ordered index allows to store data in smaller chunks across the time axis, which gives a huge performance boost[7], [8]. Unlike other time series databases, the fact that a common SQL based database is used, makes it easy to store additional structured datasets. Therefore, TimescaleDB provides needed features to join and link datasets, while also providing higher performance than other time series databases[9].

For an easy setup, a dockerized container is available to start the database which also contains a management interface through the open-source pgAdmin[10]. The open-source tool can the be used to automatically retreive and update open datasets from various sources through an ETL workflow. The datasets are stored in a SQL-compatible format, allowing them to be accessed through standard query methodology. Various other tools can then access the PostgreSQL database for relational OLAP (Online analytical processing) and create visualizations, for example with Grafana or other Python tools.

This allows to efficiently reuse datasets which are not REUSE compliant inside an organization.

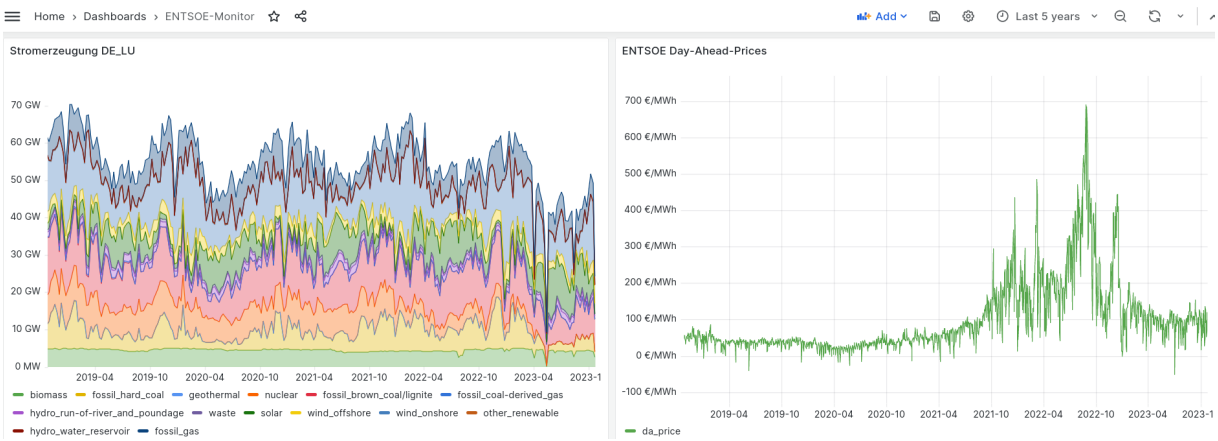

Figure 1. Sample visualization of ENTSO-E data from the database done with Grafana

4 Summary

The provided tool can be used to set up a modern time series database which contains various datasets needed for research of energy systems. It is open-source available and welcomes contributions for further datasets which are not covered yet. By utilizing modern and widely used technology, accessing open datasets becomes more efficient and convenient, which also helps to compare and analyze similar datasets to assure the quality of open datasets [11], [12]. In the future, the focus will be on adding new datasets to cover more of the available open datasets. The aim is to include datasets that are not easily accessible, such as those that are split across multiple CSV files and cannot be efficiently queried. However, a broader approach would be to address

the issue of restrictive data licenses on a larger scale. To achieve this, national research data infrastructure projects are urgently needed for the research community to systematically manage scientific and research data.

Data availability statement

The tool and scripts are open-source available under the MIT License at https://github.com/NOWUM/open-energy-data-server/.

Author contributions

Conceptualization, F.M.; software, F.M. and J.S.; validation, F.M. and J.S.; investigation, F.M., J.S. and V.S.; data curation, F.M. and J.S.; writing—original draft preparation, F.M., J.S.; writing—review and editing, F.M., J.S. and V.S.; supervision, V.S.; All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding.

References

[1] S. Pfenninger, J. DeCarolis, L. Hirth, S. Quoilin, and I. Staffell, "The importance of open data and software: Is energy research lagging behind?" *Energy Policy*, vol. 101, pp. 211–215, Feb. 1, 2017, ISSN: 0301-4215. DOI: 10.1016/j.enpol.2016.11.046. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421516306516 (visited on 07/13/2022).

[2] L. Hirth, "Open data for electricity modeling: Legal aspects," *Energy Strategy Reviews*, vol. 27, p. 100 433, Jan. 1, 2020, ISSN: 2211-467X. DOI: 10.1016/j.esr.2019.100433. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211467X19301269 (visited on 07/13/2022).

[3] M. D. Wilkinson, M. Dumontier, IJ. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160 018, 1 Mar. 15, 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://www.nature.com/articles/sdata201618 (visited on 10/27/2023).

[4] E. Kötter, B. Müller, L. Hülk, and M. Glauer (OvGU), "The OpenEnergy Platform (OEP) - A web-platform to improve transparency and reproducibility of energy system analyses," May 11, 2017. [Online]. Available: https://policycommons.net/artifacts/1554568/the-openenergy-platform-oep/2244377/ (visited on 01/25/2024).

[5] European Organization For Nuclear Research and OpenAIRE, "Zenodo," 2013. DOI: 10.25495/7GXK-RD71. [Online]. Available: https://zenodo.org/ (visited on 01/25/2024).

[6] InfluxDB. "Compare InfluxDB vs TimescaleDB." (2023), [Online]. Available: https://www.influxdata.com/comparison/influxdb-vs-timescaledb/ (visited on 10/24/2023).

[7] S. N. Z. Naqvi, S. Yfantidou, and E. Zimányi, "Time series databases and influxdb," *Studienarbeit, Université Libre de Bruxelles*, vol. 12, 2017. [Online]. Available: https://www.devopsschool.com/blog/wp-content/uploads/2022/09/influxdb_2017.pdf (visited on 10/24/2023).

[8] "Timescale Documentation — About hypertables." (), [Online]. Available: https://docs.timescale.com/use-timescale/latest/hypertables/about-hypertables/ (visited on 10/24/2023).

[9] I. González Liaño and M. Vázquez Fernández, "Use of TimescaleDB as a database for ocean-meteorological data storage," *Instrumentation Viewpoint*, no. 21, pp. 40–41, 2021, ISSN: 1886-4864. [Online]. Available: https://upcommons.upc.edu/handle/2117/360213 (visited on 10/24/2023).

[10] pgAdmin. "pgAdmin - PostgreSQL Tools." (2023), [Online]. Available: https://www.pgadmin.org/ (visited on 10/24/2023).

[11] L. Hirth, J. Mühlenpfordt, and M. Bulkeley, "The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform," *Applied Energy*, vol. 225, pp. 1054–1067, Sep. 1, 2018, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2018.04.048. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918306068 (visited on 09/15/2021).

[12] A. S. Kumar, I. Kouveliotis-Lysikatos, and L. Soder, "Comparison of Openly Available Power System Data for the Nordic Region," p. 6,