vaex

The perfect DataFrame library for Python data apps

Jovan Veljanoski & Maarten Breddels

02-Dec-2022

https://vaex.io

@

PyData Global 2022

ABOUT US

Jovan Veljanoski

Senior data-scientist @ Tiqets.com

Previously ML-specialist @ CTS.co

Former astrophysicist (PhD)

Co-Founder of vaex.io

✉ jovan.veljanoski@gmail.com

in linkedin.com/in/jovanvel/

@jovanvaex

github.com/jovanveljanoski

Maarten Breddels

Freelancer / consultant / data scientist

Core Jupyter-Widgets developer

Former astrophysicist (PhD)

Founder of vaex.io

Principal author of Vaex

Author of Solara, ipyvolume

✉ maartenbreddels@gmail.com

G www.maartenbreddels.com

@maartenbreddels

github.com/maartenbreddels

WHAT IS VAEX?

- High-performance, out-of-core DataFrame library in Python
- Out-of-core algorithms: working with data larger than RAM is just as easy
- Work with billions (10^9) of samples on a single node (machine/laptop) interactively
- Familiar API: similar to Pandas but not built on Pandas
- Easy installation:
  - `pip install vaex`
  - `conda install -c conda-forge vaex`
- Free & Open Source, MIT License

VAEX: KEY CONCEPTS

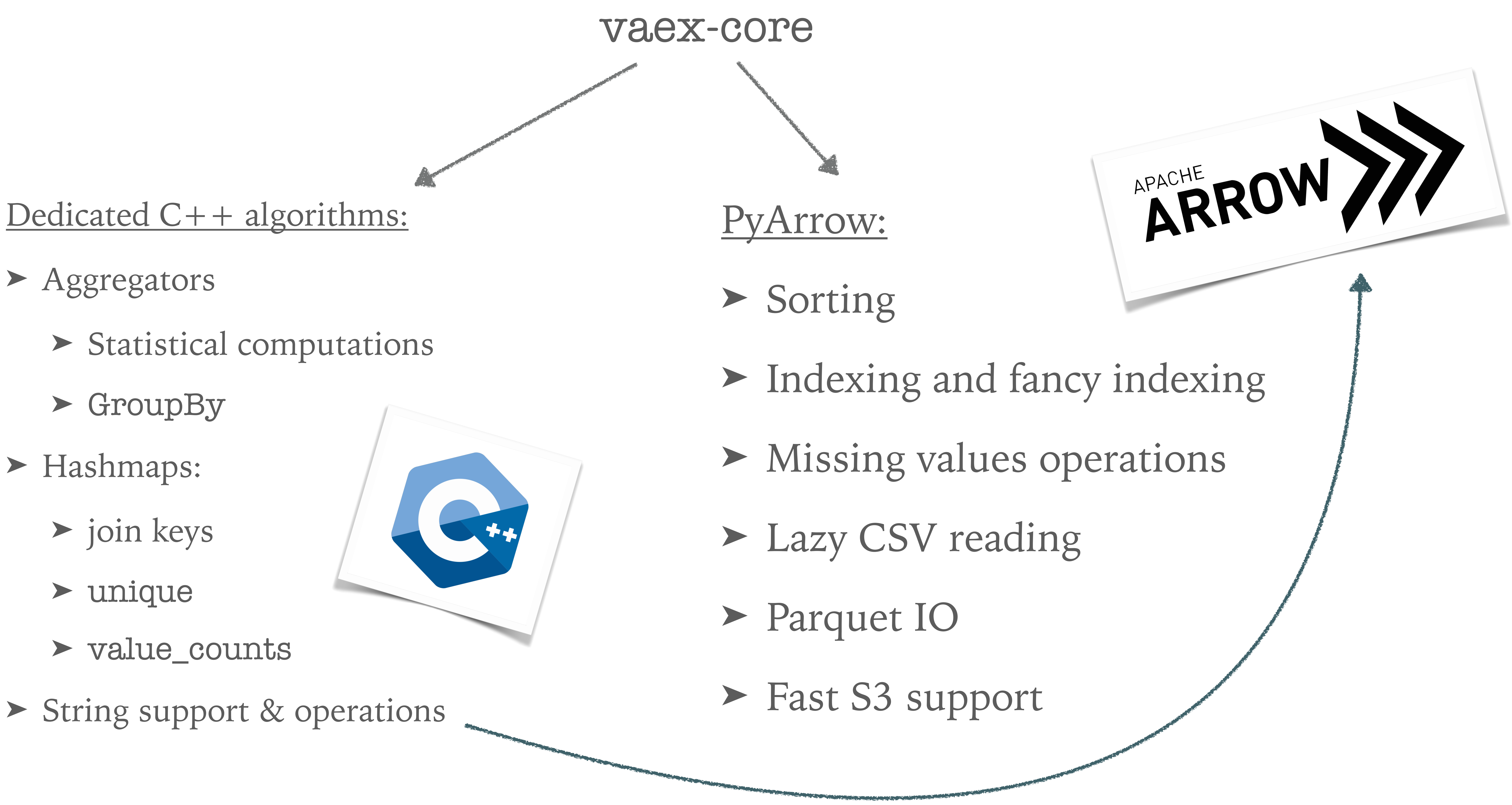
- Streaming data + memory mapping + columnar storage - work with datasets the size of your hard-drive (Arrow, HDF5, Parquet, CSV)

- Expression system - memory and computational efficiency

- Lazy evaluations - control flow, performance increase

- High performance - efficient C++ algorithms, Just-in-Time compilation via Numba, Pythran, Cuda or Metal

VAEX: UNDER THE HOOD



vaex-core

Dedicated C++ algorithms:

- Aggregators
  - Statistical computations
  - GroupBy
- Hashmaps:
  - join keys
  - unique
  - value_counts
- String support & operations

PyArrow:

- Sorting
- Indexing and fancy indexing
- Missing values operations
- Lazy CSV reading
- Parquet IO
- Fast S3 support

VAEX: HOW IT WORKS

Code

```
df = vaex.open('./data.hdf5')
```

```
df2 = df[df['x'] < 5]
```

```
df2['z'] = df2['x'] + df2['y'] * 10
```

What happens

```
df == {
  'data': {'x': Column(fd='data.hdf5', name='x'),
           'y': Column(fd='data.hdf5', name='y')},
  'state': {}
}
```

```
df2 == {
  'data': {'x': Column(fd='data.hdf5', name='x'),
           'y': Column(fd='data.hdf5', name='y')},
  'state': {
      'filters': 'y < 5'
  }
}
```

```
df2 == {
  'data': {'x': Column(fd='data.hdf5', name='x'),
           'y': Column(fd='data.hdf5', name='y')},
  'state': {
      'filters': 'y < 5',
      'virtual_columns': {
        'z': 'x + y * 10'
      }
  }
}
```

VAEX: APPLICATIONS AND COOL FEATURES

- Machine Learning via vaex-ml:
  - Out-of-core preprocessing of data
  - Binding to popular ML libraries (scikit-learn, xgboost, lightgbm, catboost, keras)
- Interactive exploration & visualisations of very large datasets (not only tabular?)
- DataFrame Server
- Cloud friendly - read and write directly from/to GCP/AWS
- Performance optimisations - ideal as a data app backend

VAEX AS A DATA APPS BACKEND

Needs of modern data applications:

- Fast, responsive, interactive
- Support for large & growing datasets
- Repetitive operations
- Support for custom logic
  - statistics
  - computations
  - statistical & ML modelling
  - visualisations

How can Vaex help:

- High performance (C++, PyArrow, JIT)
- Memory mapping - shared between processes
- Caching - shared between processes
- Delayed operations
- Async evaluations
- Fingerprinting of files
- Early stopping of operations as needed
- progress bars :)

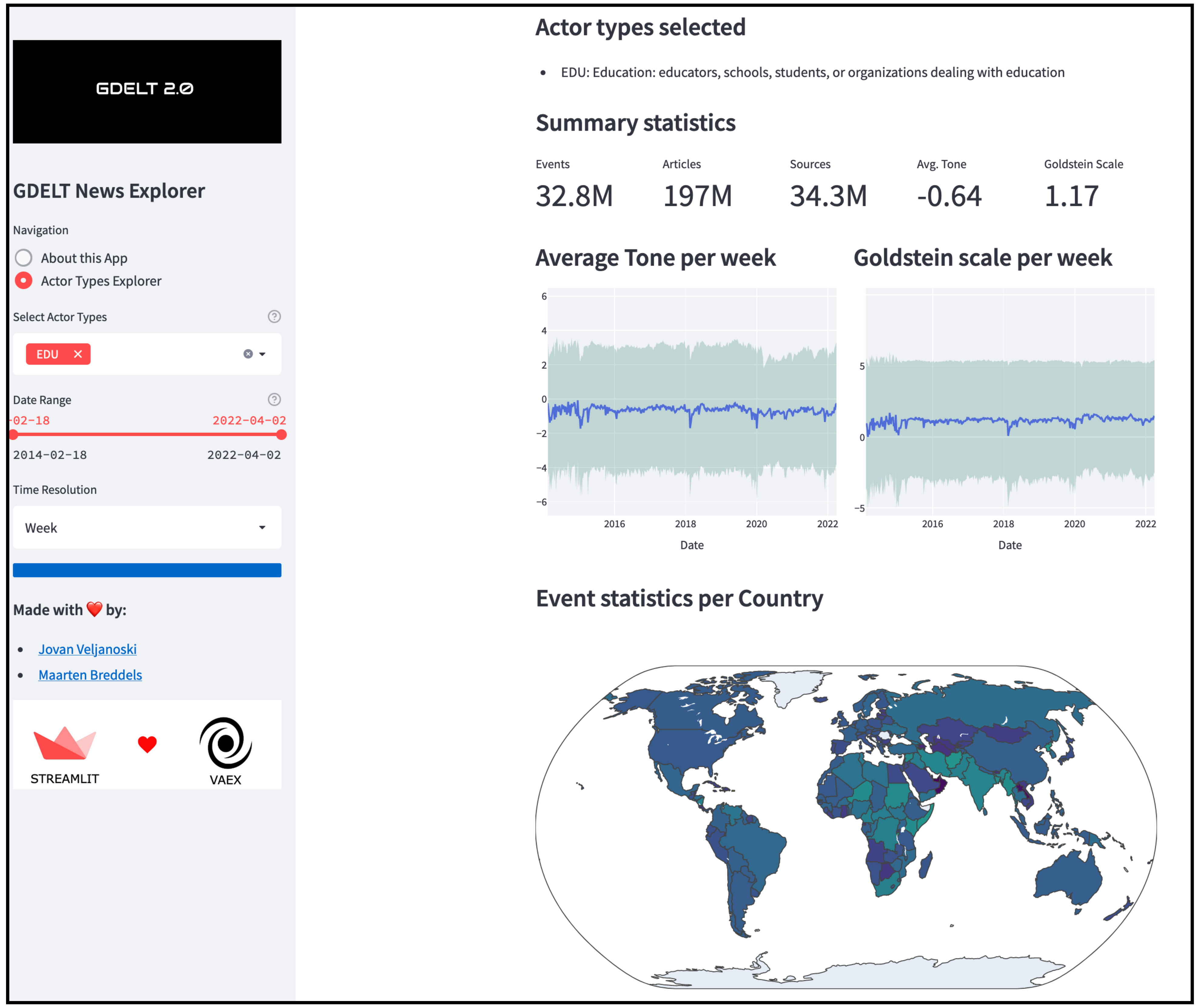
"Never do a live demo"

-Many people

VAEX IN PRODUCTION: STREAMLIT & DASH EXAMPLES

https://streamlit.vaex.io

https://dash.vaex.io/



GDELT News Explorer

Navigation

○ About this App
● Actor Types Explorer

Select Actor Types: EDU

Date Range: -02-18 – 2022-04-02

2014-02-18 2022-04-02

Time Resolution: Week

Made with ❤ by:

- Jovan Veljanoski
- Maarten Breddels

Actor types selected

- EDU: Education: educators, schools, students, or organizations dealing with education

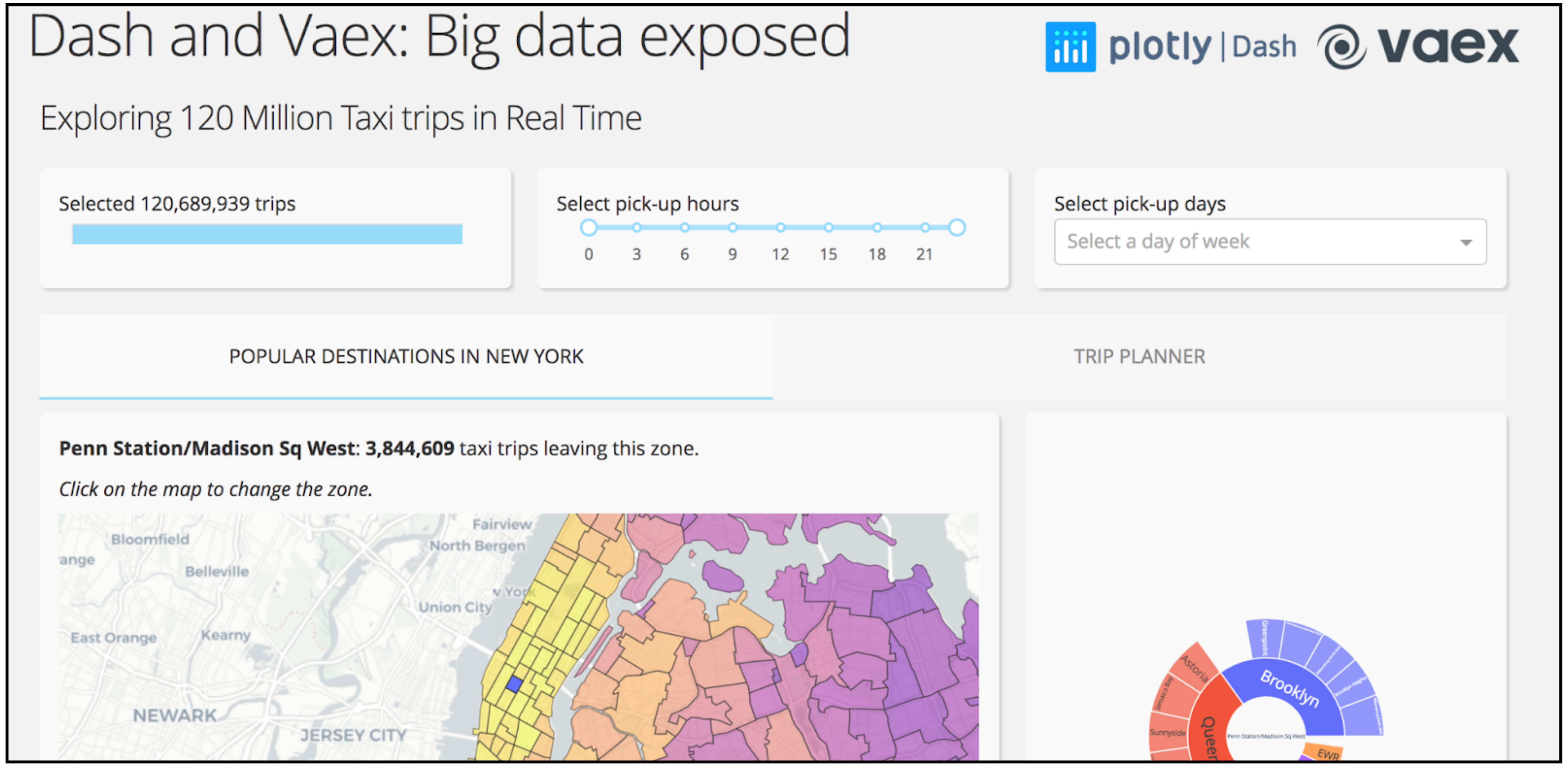
Summary statistics

Events	Articles	Sources	Avg. Tone	Goldstein Scale
32.8M	197M	34.3M	-0.64	1.17

Average Tone per week

Goldstein scale per week

Event statistics per Country



Dash and Vaex: Big data exposed

Exploring 120 Million Taxi trips in Real Time

Selected 120,689,939 trips

Select pick-up hours

Select pick-up days: Select a day of week

Penn Station/Madison Sq West: 3,844,609 taxi trips leaving this zone.

Click on the map to change the zone.

VAEX IN PRODUCTION

- Memory mapping shares data
- Integrated caching system can cash common queries
- Well tested with Plotly-Dash, Flask, FastAPI
- Cloud storage support (S3, GCP via `pyarrow`, `fsspec`)
- Sub-package structure: pick what you need
  - `vaex-core`
  - `vaex-viz`
  - `vaex-ml`
  - ….

VAEX IN THE WILD

Galileo

Field: Data processing & Machine learning

Use: Processing engine for all NLP insights and analytics on users' modelling data, scaling to hundreds of GB.

BioSkryb GENOMICS

Field: Genomics

Use: Interactive exploration of genomics data (x240 performance increase over previous solution)

plotly

Field: Data visualisations & Dashboards

Use: The primary backend of the Plotly Dashboard Engine; Plotly clients get access to big data with no setup.

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Field: Astronomy & Astrophysics

Use: Remote interactive exploration, visualisation & analysis of large datasets.

VAEX.IO: CONSULTANCY

- Feature development
- Support
- Retainers
- Performance
- Training
- Collaborations

Flow back to Open Source development and maintenance

vaex

CONTACT

- ✉ contact@vaex.io - support / consultancy / training
- https://github.com/vaexio/vaex
- @vaex_io
- Documentation: https://vaex.io/docs
- Blog: https://vaex.io/blog
- Examples: https://github.com/vaexio/vaex-examples