EVALUATE THE DASK DISTRIBUTED COMPUTING FRAMEWORK IN RESPECT TO VARIOUS SCIENTIFIC COMPUTING TASKS

EERO VAINIKKO

Course Coordinator

ARTJOM LIND

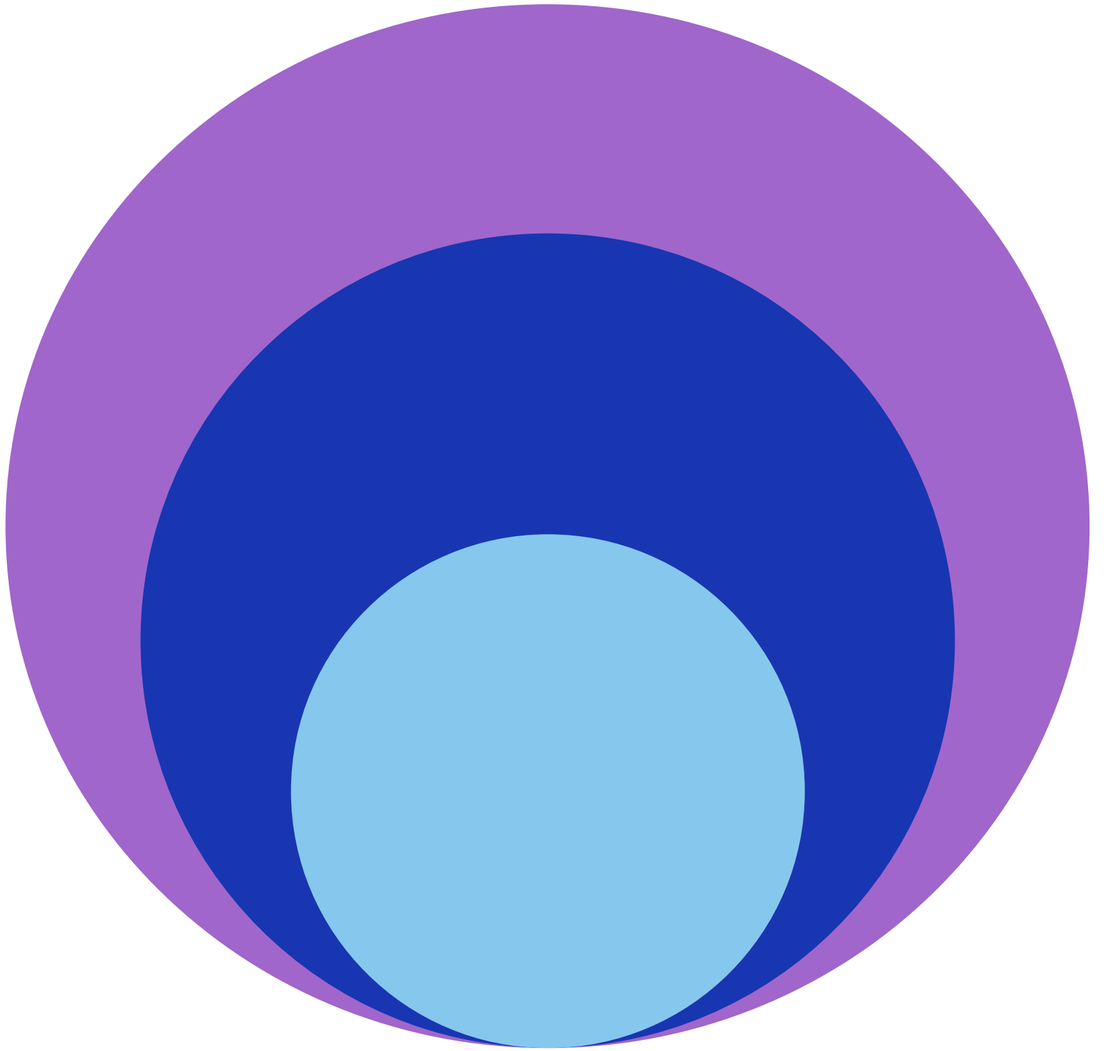
Topic Supervisor

JEYHUN ABBASOV

Student

I am going to talk...

- Dask Alternatives
- Quick Comparison
- Experiment 1
- What is next? - Experiment 2

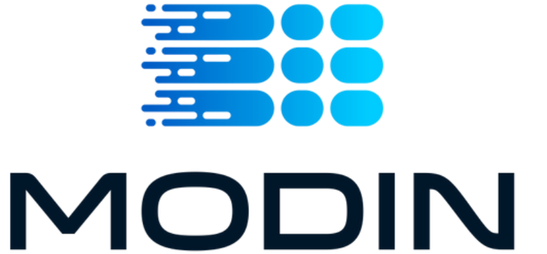


Dask Alternatives

1. **Dask**
   - scales Python code from multi-core local machines to large distributed clusters
2. **Ray**
   - scales data loading, writing, conversions, and transformations
3. **Modin**
   - uses Ray or Dask to provide an effortless way to speed up process
4. **Vaex**
   - high performance Python library for lazy Out-of-Core DataFrames

Source: https://www.dask.org/

Source: https://www.ray.io/



Source: https://modin.readthedocs.io/en/stable/

Source: https://vaex.io/

Quick comparison

	Maturity	Popularity	Scaling ability	Use cases	Scaling strategy
Dask	A	B	1TB+	Data Science	Clusters
Ray	A	A	1TB+	General	Clusters
Modin	C	C	10GB+	Data Frame	Clusters
Vaex	C	C	100GB+	Data Science	Lazy Loading

- **Maturity** - time since the first commit.
- **Popularity** - number of GitHub stars.
- **Scaling ability** - broad dataset size limits for each tool
- **Use cases** - use cases of the software libraries
- **Scaling strategy** - scaling strategy of the software libaries

Experiment 1 - Setup

Baseline

- Pandas

Dataset

- The r/place Parquet dataset - 12GB (22GB uncompressed)

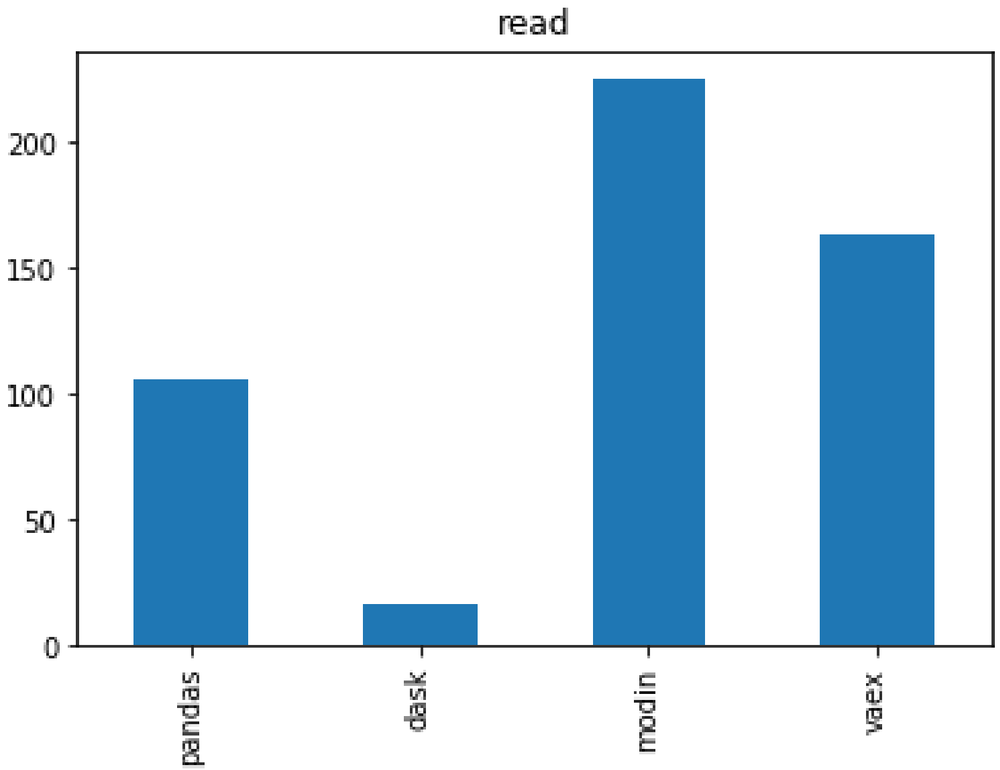
Setup

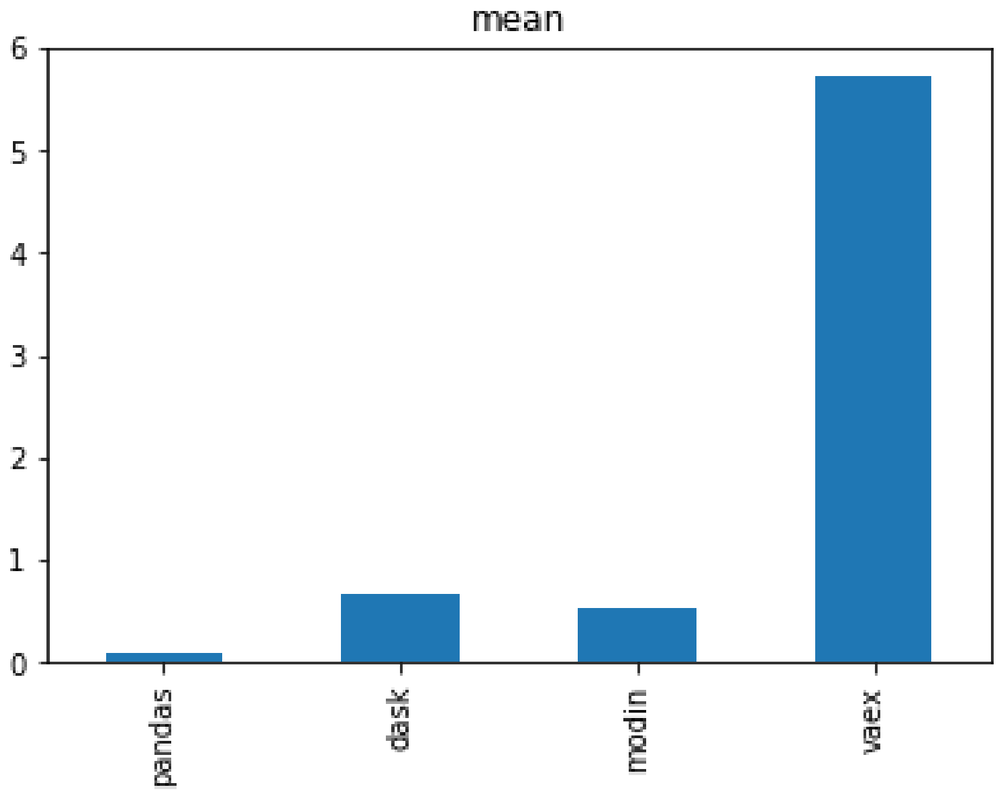
- Processor: Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
- Number of cores: 4
- Memory: 16.0 GB
- Hard Disk: 250 GB

Tests

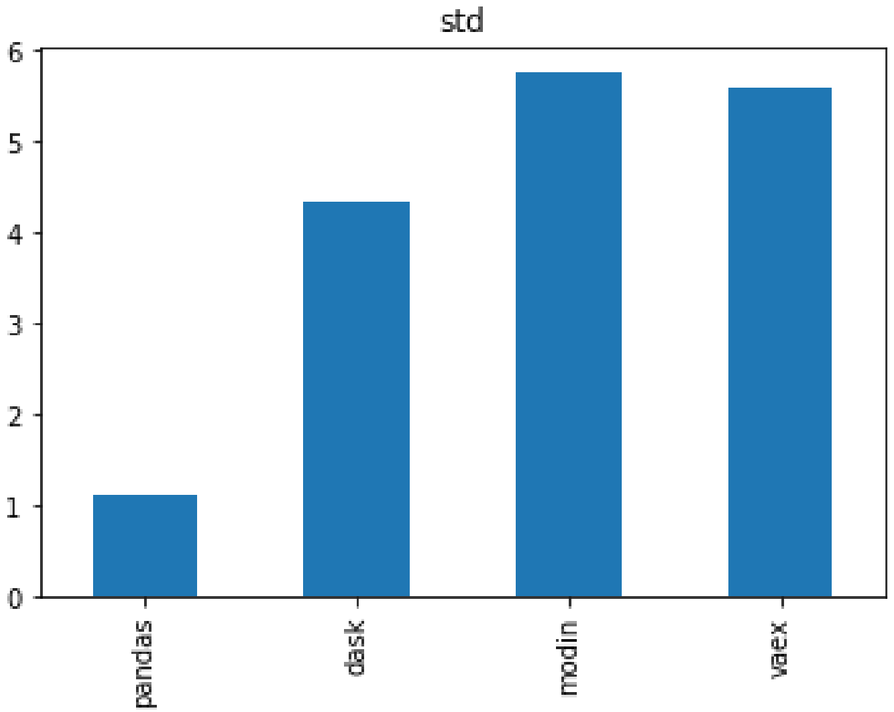
- Reading in the file. Speed comparison
- Compute metrics of a column: (mean, std)
- Finding the unique value of a column
- Cumulative sum of a column
- Groupby Aggregation

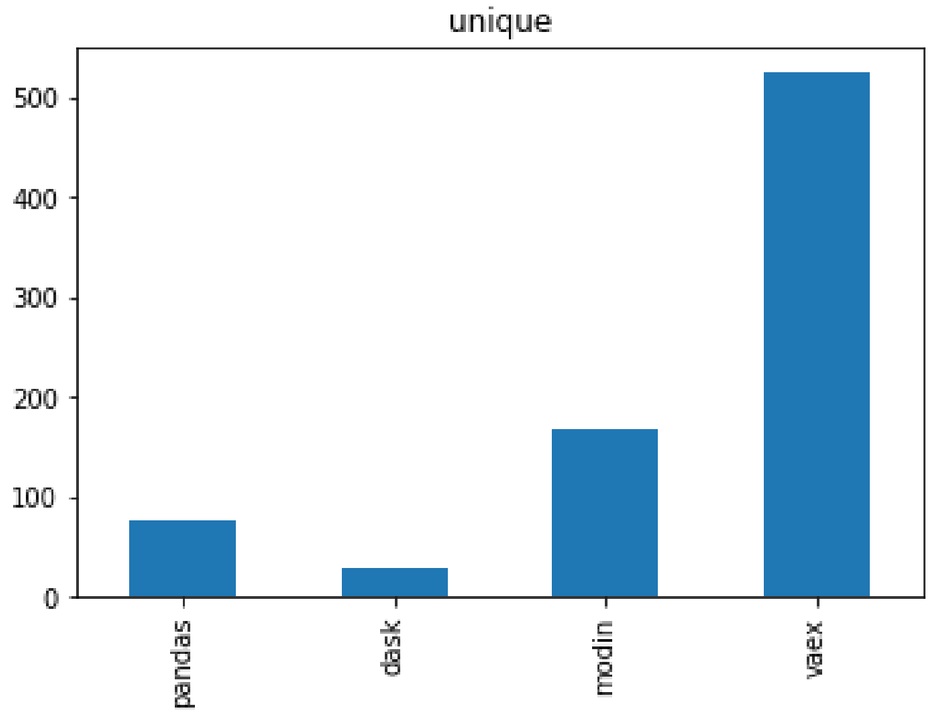
Experiment 1 - Results



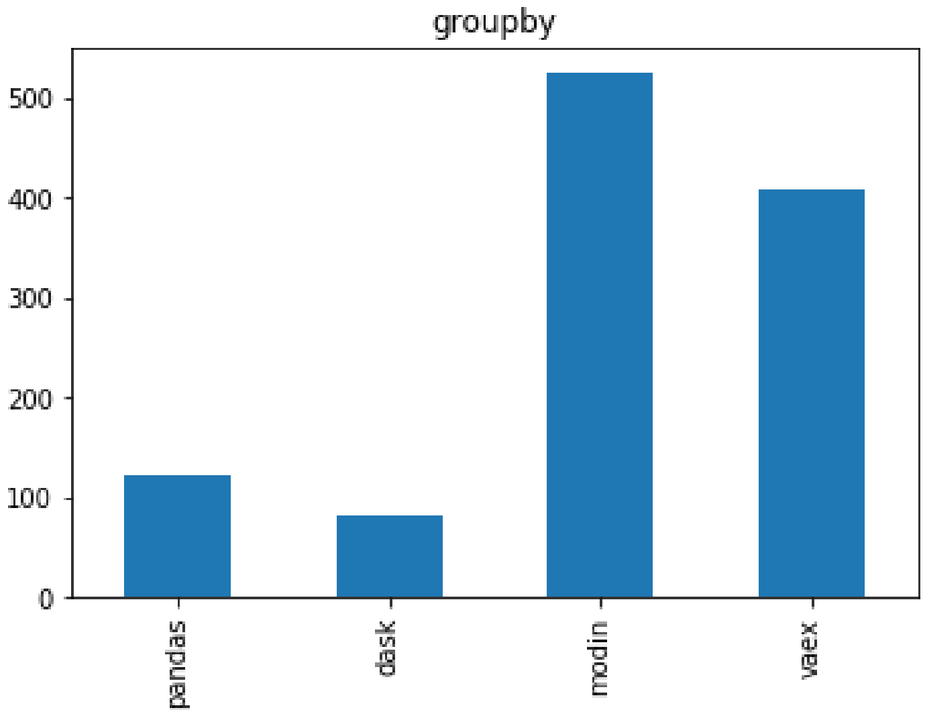


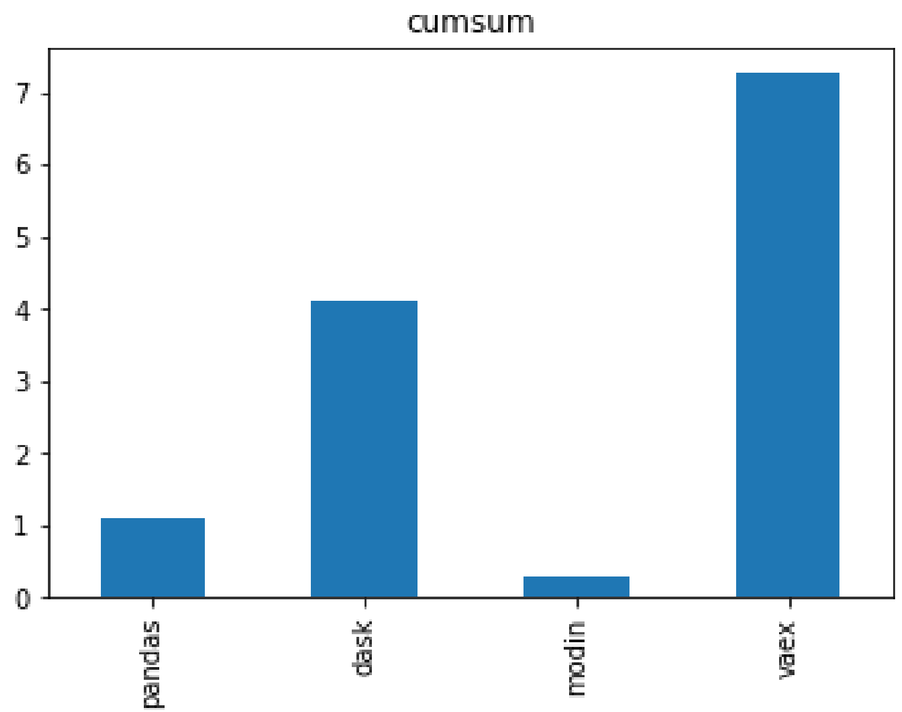
Experiment 1 - Results





Experiment 1 - Results





What is next? - Experiment 2

1 Application

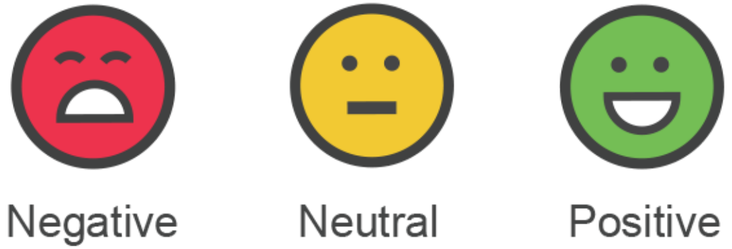
- Sentiment Analysis (data intensive app for testing distributed computing using Dask)

2 Data

- Amazon Product Reviews

3 Evaulate the Dask

- dataset sizes: 25GB, 50GB, 100GB



Source: https://thedatascientist.com/wp-content/uploads/2018/10/sentiment-analysis.png

References

[1]

M. Dugr´e, V. Hayot-Sasson and T. Glatard, ”A Performance Comparison of Dask and Apache Spark for Data-Intensive Neuroimaging Pipelines,” 2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS),

[2]

Shafi, Aamir and Hashmi, Jahanzeb and Subramoni, Hari and K., Dhabaleswar and Panda,. (2021). Efficient MPI-based Communication for GPU-Accelerated Dask Applications.

[3]

Rocklin, Matthew. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. 126-132. 10.25080/Majora-7b98e3ed-013.

[4]

D. Youssefi et al., ”CARS: A Photogrammetry Pipeline Using Dask Graphs to Construct A Global 3D Model,” IGARSS 2020 - 2020 IEEE International Geoscience andRemote Sensing Symposium, 2020, pp. 453-456, doi:10.1109/IGARSS39084.2020.9324020.