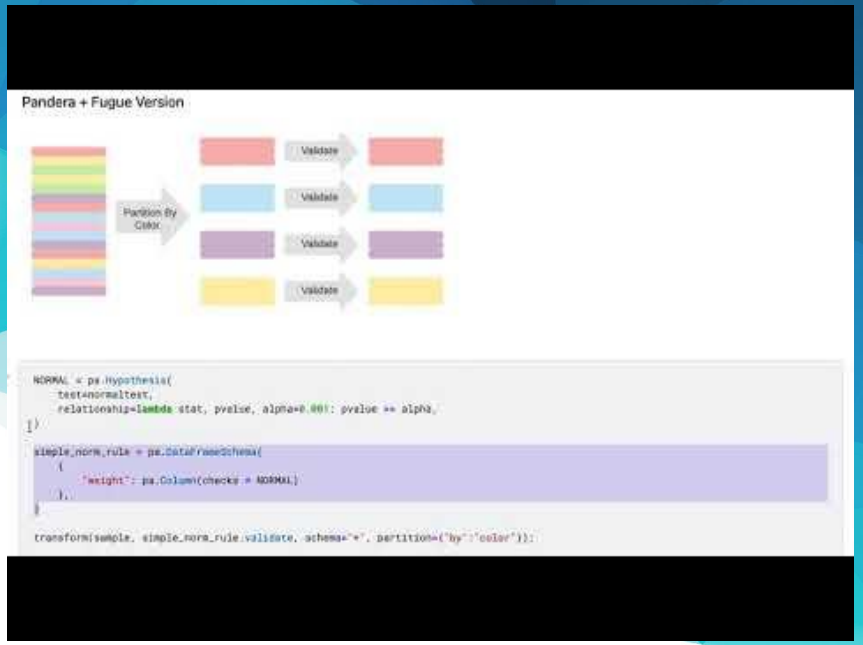
Scalable Data Validation With Fugue

Presented by: Han Wang

Demo


Pandera + Fugue Version
Partition By color
Validate
Validate
Validate
Validate
NORMAL = pa.Hypothesis(
test=normaltest,
relationship=lambda stat, pvalue, alpha=0.001: pvalue >= alpha,
)
simple_norm_rule = pa.DataFrameSchema(
{
"weight": pa.Column(checks = NORMAL)
},
)
transform(sample, simple_norm_rule.validate, schema="*", partition={"by":"color"})

Summary

- Great Expectations is feature rich, Pandera is straightforward
- Both have great support on pandas dataframes
- Both lack intuitive and scalable interface to do per-group validations
- Fugue can let both focus on single logical partition, pandas based validations while Fugue can handle the rest:
  - Various data sources, various computing engines
  - Partitioning
  - Computing orchestration
- Fugue makes validations simpler and faster

Fugue Is Open Sourced

```
pip install fugue[sql]
```

https://github.com/fugue-project/fugue

Kaggle Notebook

https://www.kaggle.com/goodwanghan/pydata-2021-scalable-data-validation-with-fugue

slack.fugue.ai

Thank You!

The Fugue Project
https://github.com/fugue-project/fugue