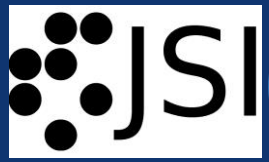
SCIKIT LEARN VS DASK VS APACHE SPARK BENCHMARKING ON THE EMNIST DATASET

Filip Zevnik, Din Music, Carolina Fortuna, Gregor Cerar

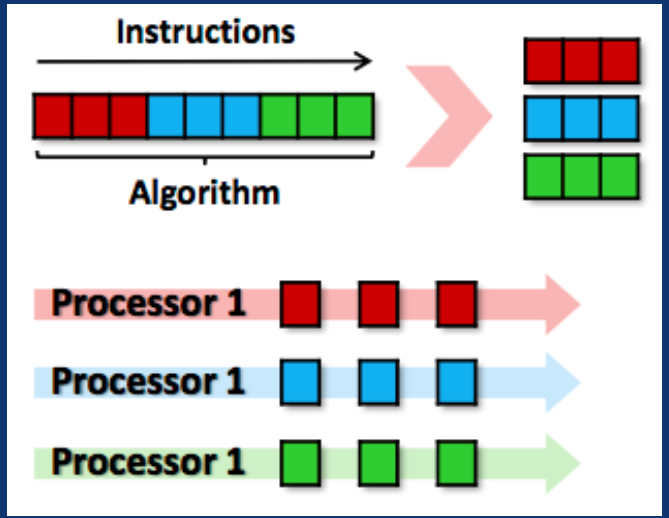


OUTLINE

- Introduction
- Technologies
- Related work
- Dataset
- Benchmarks
- Results
- Future research
- Conclusions

INTRODUCTION

- Large datasets
- Parallel processing
- Linux or Windows?
- Data manipulation
- Machine learning



TECHNOLOGIES

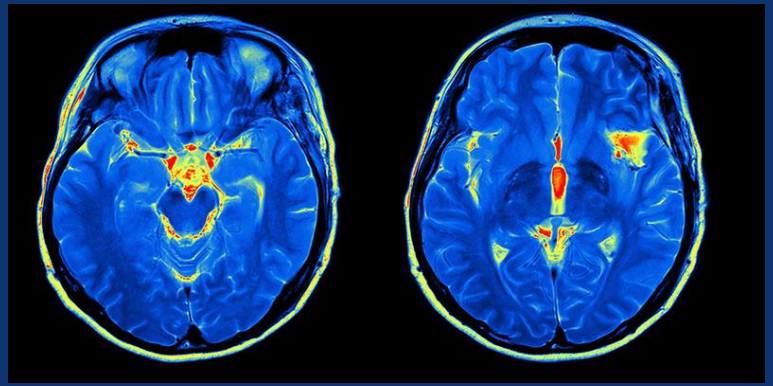
- SciKit Learn → smaller datasets

- Dask → SciKit parallelisation
- Walmart, Blue Yonder, Grubhub

- Apache Spark → analytics engine
- Yelp, Urban Institute, CrowdStrike

RELATED WORK

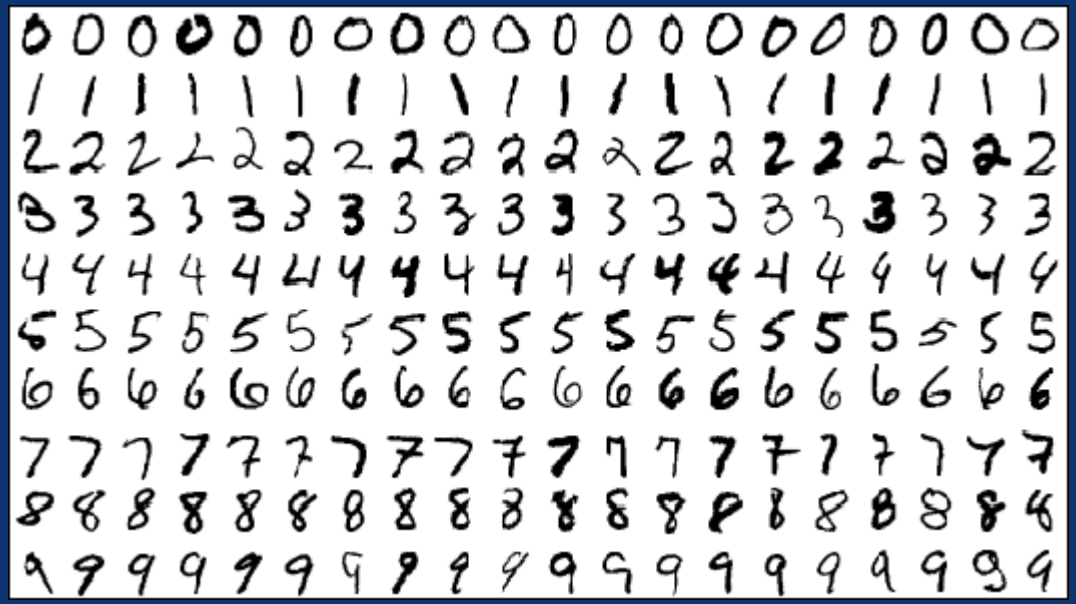
- AD campaigns → advertisement pipeline example
- Neuroimaging pipelines → adding a value of one to each voxel, histogram and BIDS app
- Satellite data processing pipeline → analysis of satellite data



- https://ieeexplore.ieee.org/document/8943502
- https://ieeexplore.ieee.org/document/9006205
- https://ieeexplore.ieee.org/document/7530084

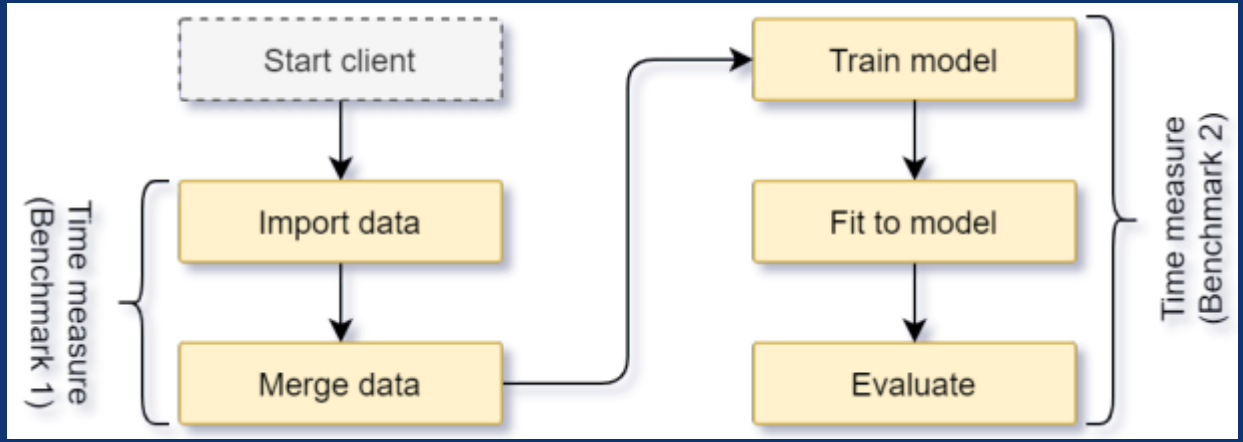
DATASET

- Extended Modified National Institute of Standards and Technology (MNIST) dataset
- Larger MNIST dataset
- Minimal efforts on formatting and pre-processing
- Classic example
- 20% - 80% split

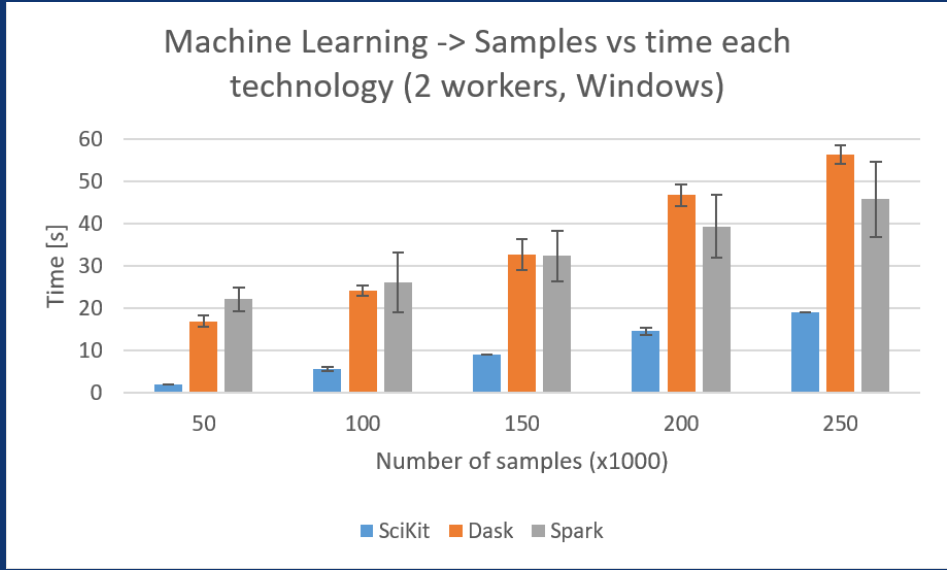


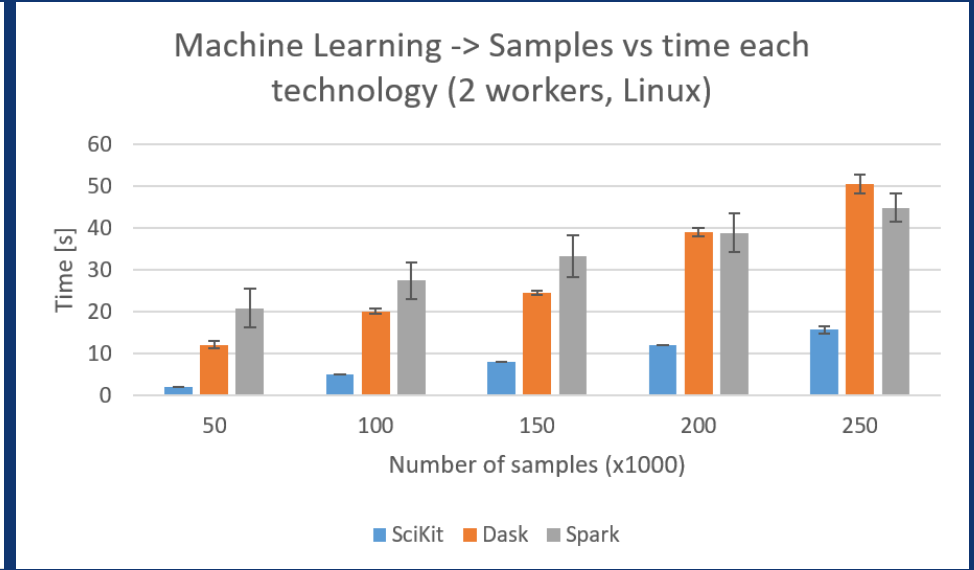
BENCHMARKS

- Machine Learning
- Import and Merge (pandas)
- 6 CPU, 10G RAM

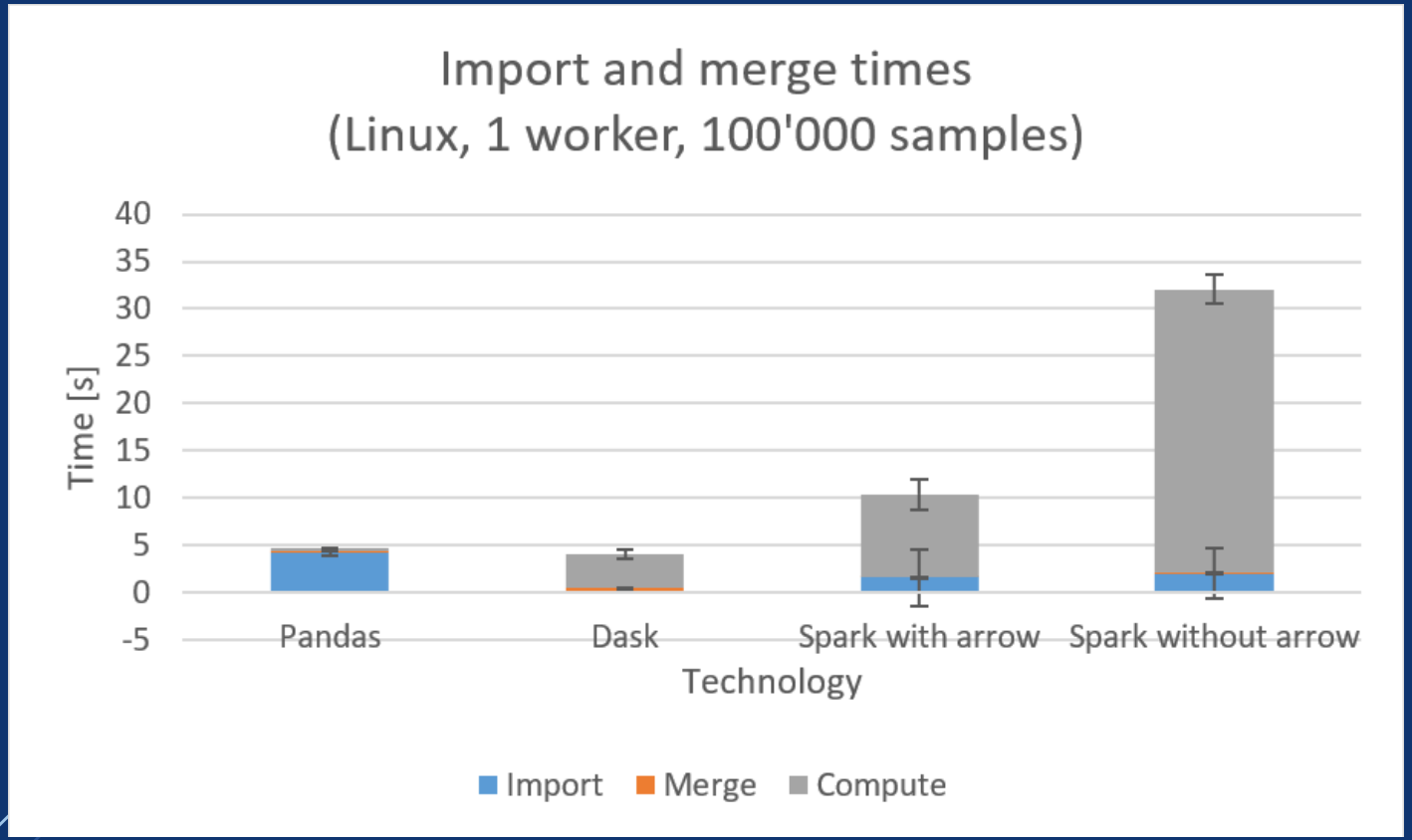


RESULTS


Machine Learning -> Samples vs time each technology (2 workers, Windows)
Time [s]
60
50
40
30
20
10
0
50
100
150
200
250
Number of samples (x1000)
SciKit
Dask
Spark


Machine Learning -> Samples vs time each technology (2 workers, Linux)
Time [s]
60
50
40
30
20
10
0
50
100
150
200
250
Number of samples (x1000)
SciKit
Dask
Spark

RESULTS


Import and merge times
(Linux, 1 worker, 100'000 samples)
Time [s]
40
35
30
25
20
15
10
5
0
-5
Pandas
Dask
Spark with arrow
Spark without arrow
Technology
Import
Merge
Compute

FUTURE RESEARCH

- Bigger dataset
- Different dataset
- Different algorithm (Nearest Neighbour, Neural Network model)

	Number of samples (x1000)				
	50	100	150	200	250
Spark	0.71	0.73	0.73	0.71	0.71
Dask	0.71	0.72	0.73	0.71	0.70
Scikit	0.70	0.71	0.70	0.71	0.73

CONCLUSIONS

- Machine Learning → Apache Spark

- Import and merge → Dask
- Apache Arrow

- Smaller datasets → classical technologies (SciKit learn)

- Linux > Windows