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

Filip Zevnik, Din Music, Carolina Fortuna, Gregor Cerar
Outlines:

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

Machine Learning -> Samples vs time each technology (2 workers, Windows)

Machine Learning -> Samples vs time each technology (2 workers, Linux)
RESULTS

Import and merge times
(Linux, 1 worker, 100'000 samples)
FUTURE RESEARCH

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

<table>
<thead>
<tr>
<th>Number of samples (x1000)</th>
<th>Spark</th>
<th>Dask</th>
<th>Scikit</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>100</td>
<td>0.73</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>150</td>
<td>0.73</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>200</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>250</td>
<td>0.71</td>
<td>0.70</td>
<td>0.73</td>
</tr>
</tbody>
</table>
CONCLUSIONS

• Machine Learning → Apache Spark
• Import and merge → Dask
• Apache Arrow
• Smaller datasets → classical technologies (SciKit learn)
• Linux > Windows