Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. This paper was not previously presented to another examination board and has not been published.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted.

Meschede, 20th September 2023.

Casimir Giesler
MatNr: 123454678
Email: curie.marie@fh-swf.de
Corresponding Author

This Declaration must be signed by all listed Authors to be valid.

Math stuff for pyspark

1 Performance measurement

To evaluate the scalability of the different implementation, both in rows and columns as well as in cluster size, the DUS Airports Hadoop cluster is used. The Spark native implementation `pyspark.ml.regression.LinearRegression` is used as baseline. Each permutation of number of rows, number of columns, number of nodes and applied algorithm is measured at least five times.

1.1 Data scalability

The test shows, that the custom implementations perform significantly worse than the PySpark implementation, except the map-reduce LU implementation which is the only method which performs better than the PySpark implementation. QR consistently performs worst, as visualized in figures 1 and 2. Besides the worse performance, the QR and SVD implementations still show a linear algorithmic complexity and are somewhat scaleable.

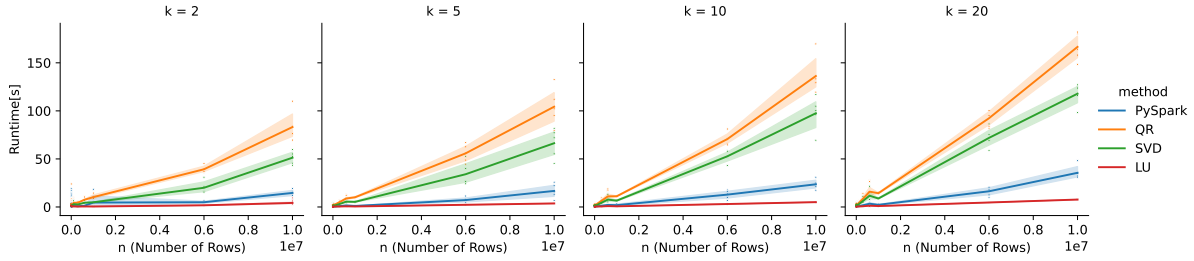

Figure 1: *Runtime comparision for the different implementations with linear scale*

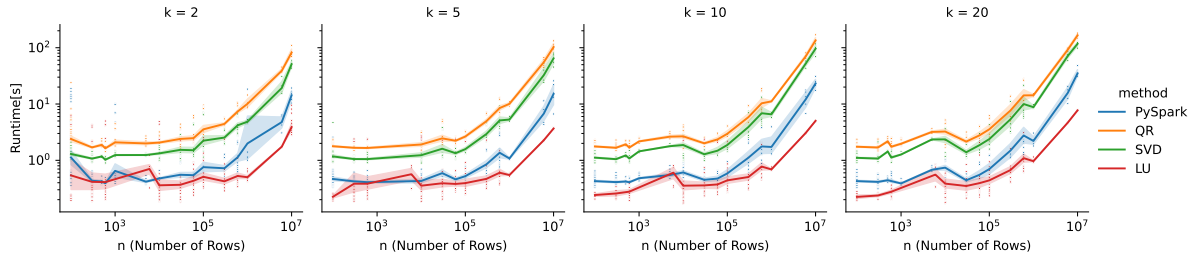

Figure 2: *Runtime comparision for the different implementations with logarithmic scale*