AWARE:
Workload-aware, Redundancy-exploiting Linear Algebra: Reproducibility Guide

This is the guide to reproduce the paper. To make the execution work there is a few setup requirements.

To produce a full reproduction of the paper it is expected that one have access to a Spark (v 3.2.0) cluster with Hadoop (v 3.3.1) running Java 11. But any single machine should be able to run small scale experiments covering most of the experiments shown. For baseline experiments

The specifics of the resources used in the paper is: 6 Cluster nodes and a main node with 32 virtual cores and 128 GB RAM each. Local storage requirements is at least 100GB, and distributed HDFS is 2.0 TB.

The machines have similar specifications as m5a.8xlarge in AWS. In there it should be simple to start a spark cluster, but setting such up is not covered in this guide. Note that if one wants to, it has to be able to switch hadoop and spark versions, to run all baseline experiments.

Further dependencies are:

- Java 11 and 8 available on main node
- Maven 3.6+
- Git
- rsync (installed per default on Ubuntu)
- ssh (also installed per default on Ubuntu)
- Python 3.6+
- pdflatex - If you want to make the paper.

1 Setup / Verification

verify install and setup.

change parameters to run what you want

2 Data Preparation

create datasets.

3 Micro benchmarks

4 local execution

5 distributed execution

6 plotting

7 compilation of paper