DDM Exercise: Spark Hands-On

Team: ChewbAKKA
Felix Gohla
Timofei Kornev

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Algorithm

- Re-implementation of the Sindy algorithm proposed in the paper “Scaling Out the Discovery of Inclusion Dependencies” by Sebastian Kruse, Thorsten Papenbrock, and Felix Naumann (link)
- 4 steps:
  0. Read in input files
  1. Obtain attribute sets:
     1.1. Create reversed index (in the paper, corresponds to the step of creating tuples with a value and a singleton set with a column name)
     1.2. Aggregate and group column names by value
  2. Check for inclusions by creating pairs (<dependent column>, <referenced columns>) and intersecting referenced columns grouped for each dependent column
  3. Print out the results to the console

DDM Exercise: Spark Hands-On

Team: ChewbAKKA
Felix Gohla
Timofei Kornev