Assignment 4 – Team Infinity

The source code can be find in the spark-tutorial directory.

Transformation pipeline:

Read all tables as datasets

→ use flatMap and map to get all colums

→ map each entry to the name of its column and use union in the reduce to get all cells

→ transform to dataframe

→ use a groupBy function in combination with collect_set to get the attribute sets proposed in the paper (Scaling Out the Discovery of Inclusion Dependencies S. Kruse, T. Papenbrock, F. Naumann, Proceedings of the conference on Database Systems for Business, Technology, and Web (BTW). pp. 445-454 (2015).)

→ explode each attribute set to get the keys of the inclusion lists and and transform to dataset

→ for each row reduce the attribute set by the key to get the inclusion list

→ transform to dataframe

→ use groupBy and collect set to get the aggregates

- Each column with an empty set in the inclusion list must have an empty set in the aggregates. To achieve that transform to a dataset, compute all intersects and filter the empty intersects.

→ sort the computed aggregates

→ collect and print in the given format