Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Exercice 2 - Inclusion dependencies with apache spark

Distributed Data Management

Generelles Konzept

Für das Lösen der Aufgabe haben wir uns an das in der Vorlesung vorgestelltes Paper (*Scaling Out the Discovery of Inclusion Dependencies S. Kruse, T. Papenbrock, F. Naumann)* orientiert.

Unser Algorithmus besteht aus folgenden Schritten:

1. CSV Daten lesen
2. Daten in einzelne Spalten konvertieren und mappen
   a. Die geladenen Tabellen/Dataframes splitten
   b. Duplikate aus den einzelnen Spalten entfernen
   c. Mapping von Werten zu Attributen `{(v1 -> a1),(v1 -> a2)...}`
3. Attribute anhand der Werte gruppieren `(v1 -> {a1, a2})`
4. Listen für Inclusion Dependencies erzeugen `(a1, {a2})`
5. Inclusion Dependencies check:
   a. Gruppiere anhand des ersten Attributs `(a1 -> {{a2},{a3}})`
   b. Schnittmenge der Attribute bilden `{a2 ∩ a3}`
6. Filtern von leeren Listen
7. Daten sammeln (collect) und ausgeben