DDM-SPARK

ELMO PRIMMER, CHRIS NECKEL

Step by step process

- Read data into lists of datasets

N_NATIONKEY,N_NAME,N_REGIONKEY,N_COMMENT
1,Germany,1,"Rich economy"
2,France,1,"Fine wine"
3,Brazil,2,"Coffee exporter"

- Reapartition into **n_cores * 3**
- Turn each column into RDD
- Map each column to set of **(value, Set(columnName))**
- Take union accross all DataFrames (files)

("Germany", Set("N_NAME")),
("France", Set("N_NAME"))

- Coalesce n_cores * 2
- Reduce by key (value of column)

("Germany", Set("N_NAME", "N_REGIONKEY"))

- Pair each column with the rest of the columns in its set for identifying potential INDs.

("N_NAME", Set("N_REGIONKEY")),
("N_REGIONKEY", Set("N_NAME"))

- Intersect column sets for each column across all values to identify actual INDs.

IND: ("N_NAME", Set("N_REGIONKEY"))

- Sort and print