Inclusion Dependency Discovery with SINDY

Samik Real & Lisa Koeritz

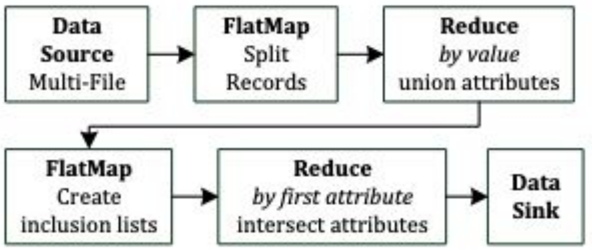
Inclusion Dependency (IND) Discovery - with SINDY

Implementation follows the steps of the SINDY algorithm from Kruse, Papenbrock and Naumann (2015) utilizing Spark Framework

1. read data into single columns, create one dataset and do cache-based pre-aggregation to create attribute sets

2. create inclusion lists using `explode()` and aggregate them for every attribute via `reduce()` and `intersect()`

3. disassemble lists into INDs, sort alphabetically and output into console


Data Source Multi-File
FlatMap Split Records
Reduce *by value* union attributes
FlatMap Create inclusion lists
Reduce *by first attribute* intersect attributes
Data Sink

implementation description from Kruse, Papenbrock and Naumann (2015)

```
C_CUSTKEY < P_PARTKEY
C_NATIONKEY < S_NATIONKEY , N_NATIONKEY
L_COMMIT < L_SHIP , L_RECEIPT
L_LINENUMBER < S_NATIONKEY , L_SUPPKEY , P_PARTKEY , P_SIZE , L_PARTKEY , C_NATIONKEY , O_ORDERKEY , N_NATIONKEY ,
S_SUPPKEY , C_CUSTKEY
L_LINESTATUS < O_ORDERSTATUS
L_ORDERKEY < O_ORDERKEY
L_PARTKEY < P_PARTKEY
L_SUPPKEY < P_PARTKEY , S_SUPPKEY , C_CUSTKEY
L_TAX < L_DISCOUNT
N_NATIONKEY < C_NATIONKEY , S_NATIONKEY
N_REGIONKEY < C_NATIONKEY , S_NATIONKEY , N_NATIONKEY , R_REGIONKEY
O_CUSTKEY < P_PARTKEY , C_CUSTKEY
O_SHIPPRIORITY < C_NATIONKEY , S_NATIONKEY , N_REGIONKEY , N_NATIONKEY , R_REGIONKEY
P_SIZE < L_SUPPKEY , P_PARTKEY , S_SUPPKEY , C_CUSTKEY , L_PARTKEY
R_REGIONKEY < C_NATIONKEY , S_NATIONKEY , N_REGIONKEY , N_NATIONKEY
S_NATIONKEY < C_NATIONKEY , N_NATIONKEY
S_SUPPKEY < L_SUPPKEY , P_PARTKEY , C_CUSTKEY
```