Spark Assignment

Distributed Data Management

Tobias Jordan & Nastassia Heumann - 7th July 2021

Our Approach

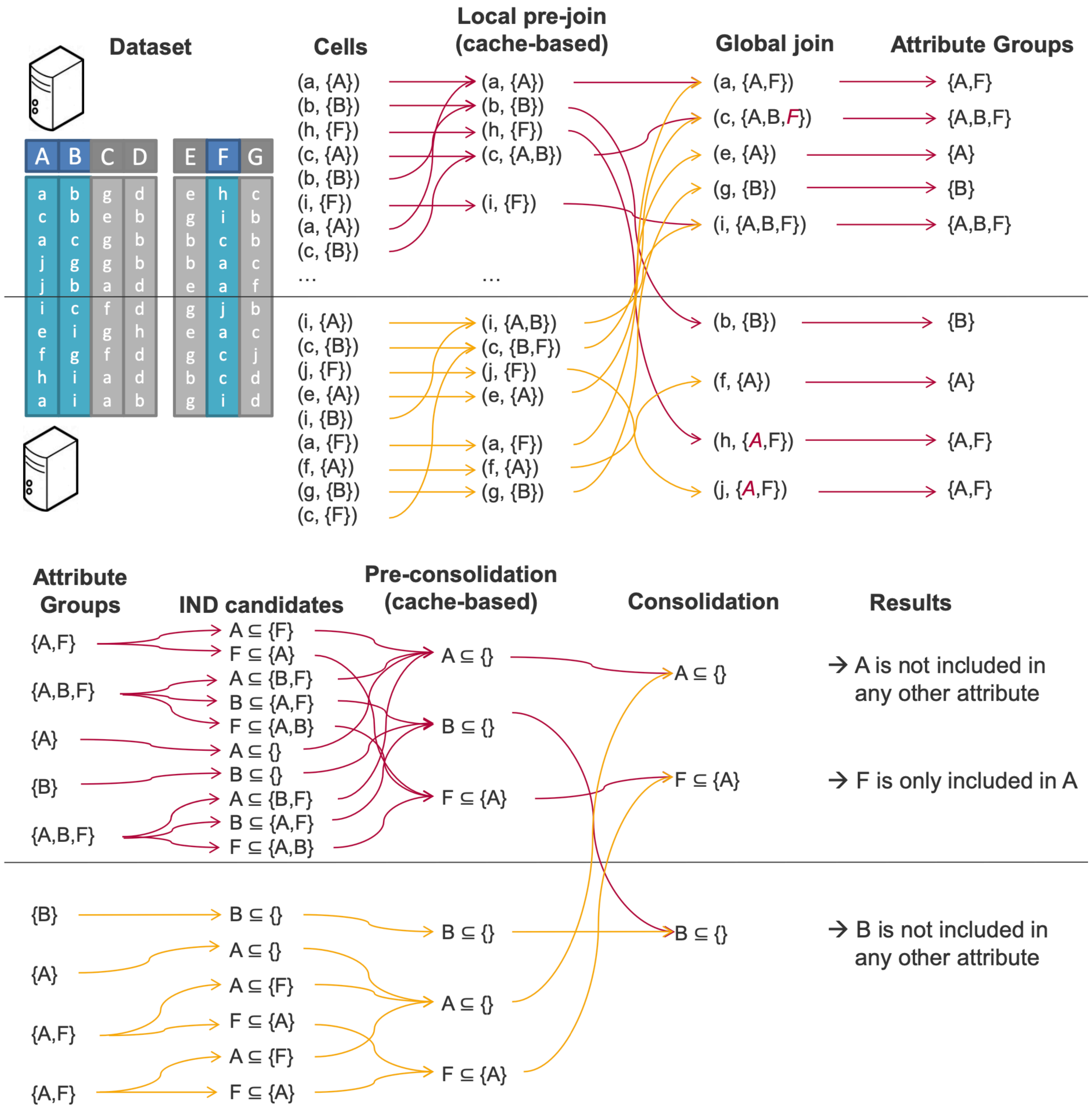
Sindy Algorithm

- Our implementation is based on the Sindy slides from Data Profiling (WS20/21)
- We strongly make use of the functional methods `map(), reduce(), filter(), sort(), foreach()`
- What we found challenging:
  - Figuring out how to handle various Scala data types (DataFrame, WrappedArray, Dataset, …)
  - Learning Scala Syntax
- What surprised us is how little code is needed to detect unary INDs

Our Approach

Sindy Algorithm

1. Read the datasets
2. Map all cell values to the column
3. Join all columns with the same cell value
4. Based on the joined columns, aka attribute groups, compute all IND candidates
5. Consolidate IND candidates
6. Print INDs



Dataset

A	B	C	D
a	b	g	d
c	b	e	b
a	c	g	b
j	g	g	b
j	b	a	d
i	c	f	d
e	i	g	h
f	g	f	d
h	i	a	d
a	i	a	b

E	F	G
e	h	c
g	i	b
b	c	b
b	a	c
e	a	f
g	j	b
e	a	c
g	c	j
b	c	d
g	i	d

→ A is not included in any other attribute

→ F is only included in A

→ B is not included in any other attribute