Efficient distributed discovery of bidirectional order dependencies

Bidirectional Order Dependencies (bODs)

BODs capture order relationships between lists of attributes in a relational table. They can express that, e.g., sorting books by *publication date* in ascending order also sorts them by *age* in descending order. The knowledge about order relationships is useful for many data management tasks, such as query optimization, data cleaning, or consistency checking. Because the bODs of a specific dataset are usually not explicitly given, they need to be discovered.

age↑ ↦ year-of-birth↓

age	yob
19	2001
25	1995
25	1995
31	1989
45	1975

https://www.co2.earth/annual-co2

year	CO2
2019	411.49
2018	408.59
2017	406.59
2016	404.28

year↑ ↦ CO2↑

salary↑ ↦ tax↑

salary	tax
5k	1k
6k	1.5k
8k	2k
10k	3k

Discovery

cases↑, r0↓ ↦ dt↑

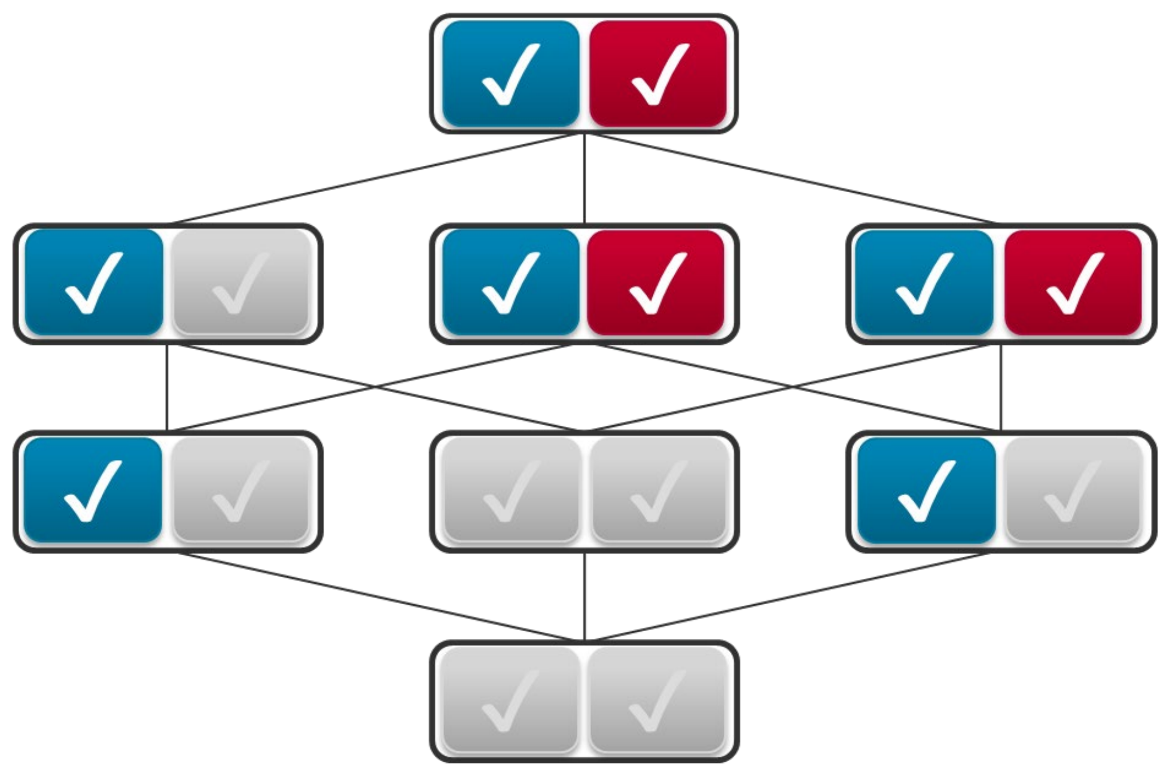
~~r0↑ ↦ dt↑~~

…

	cases	r0	doubling time
t_1	46	1.5	4 d
t_2	57	1.8	7 d
t_3	102	1.7	10 d
t_4	102	1.4	12 d
t_5	188	1.4	14 d

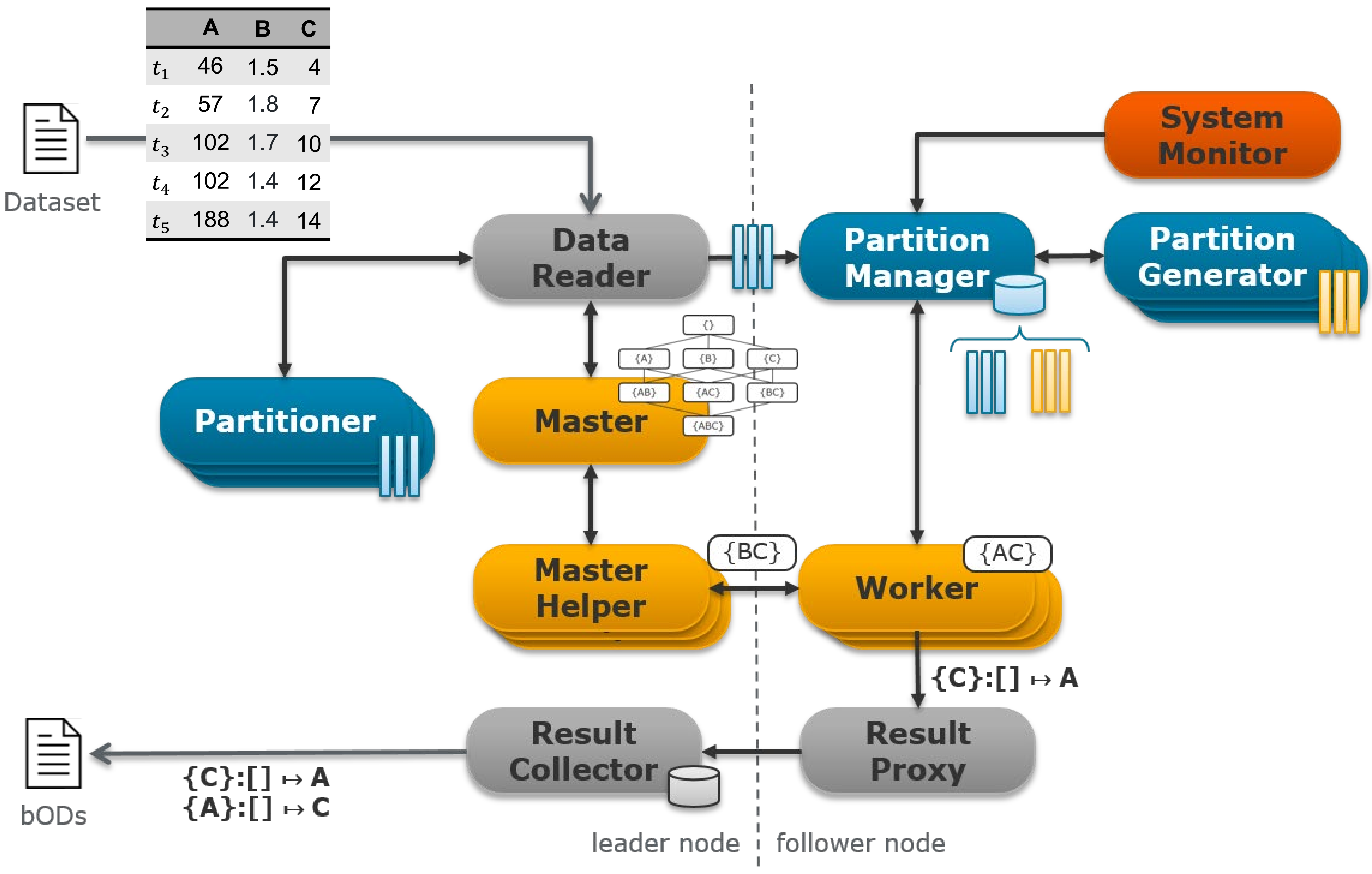
swap

split



Distributed algorithm

DISTOD is a distributed bOD discovery algorithm, whose execution time scales with the available hardware. DISTOD uses a scalable, robust, and elastic discovery approach based on **actor programming** that combines efficient pruning techniques for bOD candidates in a set-based canoncial form with a novel, reactive, and distributed search strategy.



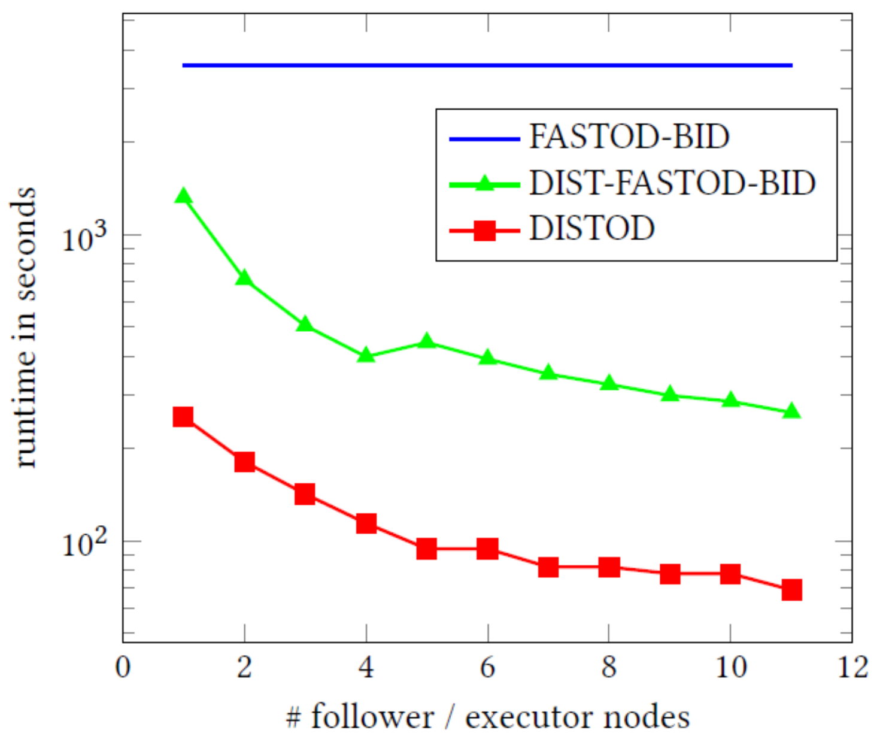
	A	B	C
t_1	46	1.5	4
t_2	57	1.8	7
t_3	102	1.7	10
t_4	102	1.4	12
t_5	188	1.4	14

Evaluation

Dataset	Columns	Rows	Results	FASTOD	FASTOD (Spark)	DISTOD
Adult	15	32 561	1 218	1h	6m	**1m**
TPC-H	16	6m	17 744	OOM	OOM	**17h**
Letter	17	20 000	2 263	4.5h	22m	**5m**
NCVoter	19	999 999	4 934	OOM	TL	**10h**
Flight	21	499 999	2 543	OOM	16m	**4m**
Horse	29	300	2.4m	OOM	TL	**7h**
FD-Reduced	30	250 000	90 313	44m	22m	**4m**

Reactive discovery is 4x – 12x faster

As scalable as batch-oriented discovery (Spark)



Sebastian Schmidl
Thorsten Papenbrock

Information Systems Group
Hasso Plattner Institute, University of Potsdam
Potsdam, Germany

E-Mail: firstname.lastname@hpi.de

https://hpi.de/naumann/s/distod

doi:10.1007/s00778-021-00683-4