UNIVERSITÄT LEIPZIG

PARALLEL OBJECT MATCHING IN THE CLOUD

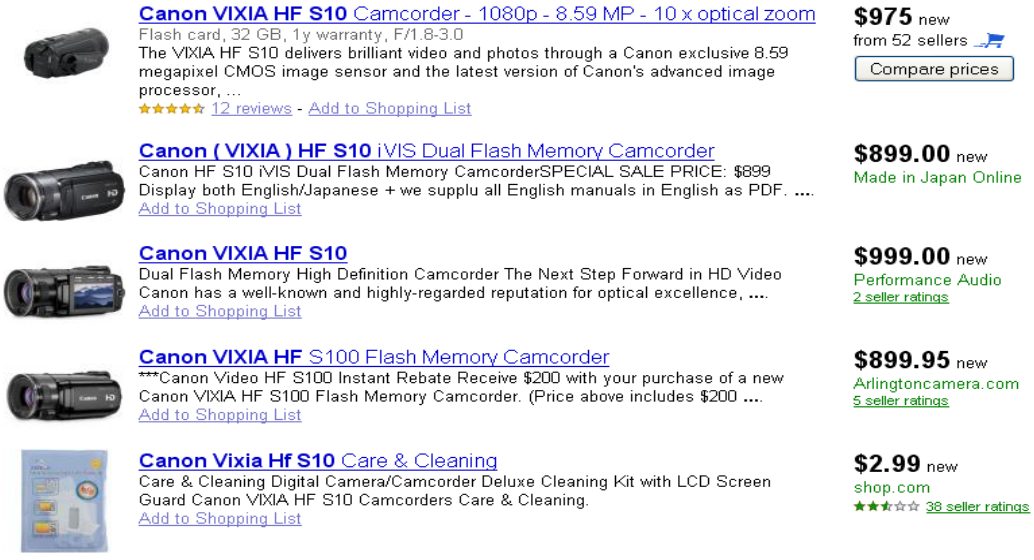
Erhard Rahm

Lars Kolb

Andreas Thor

Object Matching
(entity resolution, deduplication …)

- Identification of semantically equivalent objects
  - within one data source or between different sources
  - to merge them, compare them, improve data quality, etc.



Canon VIXIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x optical zoom
Flash card, 32 GB, 1y warranty, F/1.8-3.0
The VIXIA HF S10 delivers brilliant video and photos through a Canon exclusive 8.59 megapixel CMOS image sensor and the latest version of Canon's advanced image processor, ...
12 reviews - Add to Shopping List

$975 new
from 52 sellers
Compare prices

Canon (VIXIA) HF S10 iVIS Dual Flash Memory Camcorder
Canon HF S10 iVIS Dual Flash Memory CamcorderSPECIAL SALE PRICE: $899 Display both English/Japanese + we supplu all English manuals in English as PDF. ...
Add to Shopping List

$899.00 new
Made in Japan Online

Canon VIXIA HF S10
Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video Canon has a well-known and highly-regarded reputation for optical excellence, ...
Add to Shopping List

$999.00 new
Performance Audio
2 seller ratings

Canon VIXIA HF S100 Flash Memory Camcorder
***Canon Video HF S100 Instant Rebate Receive $200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. (Price above includes $200 ...
Add to Shopping List

$899.95 new
Arlingtoncamera.com
5 seller ratings

Canon Vixia Hf S10 Care & Cleaning
Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning.
Add to Shopping List

$2.99 new
shop.com
38 seller ratings

Duplicate web entities: Example 2

A survey of approaches to automatic schema matching
E Rahm, PA Bernstein - the VLDB Journal, 2001 - Springer
The VLDB Journal 10: 334–350 (2001) / Digital Object Identifier (DOI) 10.1007/s007780100057
... A survey of approaches to automatic schema matching ... Erhard Rahm 1 , Philip A. Bernstein 2 ... 1 Universitat Leipzig, Institut fur Informatik, 04109 Leipzig, Germany; (e-mail: rahm@ ...
Cited by 2436 - Related articles - All 72 versions

[CITATION] A survey of approaches to automatic schema matching
PA Bernstein, E Rahm - VLDB Journal, 2001
Cited by 19 - Related articles

[CITATION] A survey of approaches to automatic schema matching
R Erhard, AB Philip - VLDB Journal, 2001
Cited by 10 - Related articles

[CITATION] A survey of approaches to automatic schema matching, in 'The VL
E Rahm, PA Bernstein - Vol, 2001
Cited by 3 - Related articles

[CITATION] A survey of approaches to semantic schema matching
E Rahm, PA Bernstein - The VLDB Journal 10: 334, 2001
Cited by 5 - Related articles

[CITATION] A survey of approaches to automatic schema mapping" the VLDB ...
E Rahm, PA Bernstein - Vol
Cited by 3 - Related articles

Duplicates due to
- Order of authors
- Extraction errors
- Different titles
- Typos
- ...

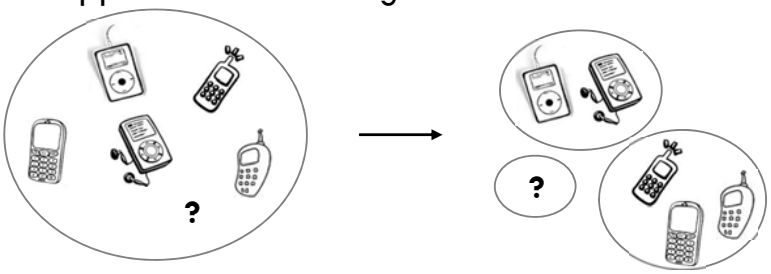
Object Matching Problem

- Lots of research work
  - String similarities, usage of structural information
  - Combined use of several matchers
  - Application of machine learning, …
- Study of real-world match systems/problems [VLDB'10]
  - Effective matching is difficult: F-Measure <75% for product data
  - Matching is expensive: scalability issues for $O(n^2)$

[VLDB'10] Koepcke, Thor, Rahm: Evaluation of entity resolution approaches on real-world match problems. VLDB 2010

How to speed up object matching?

- **Blocking** to reduce search space
  - Group similar objects within blocks based on *blocking key*
  - Restrict object matching to objects from the same block
  - Alternative approach: Sorted Neighborhood



- **Parallelization**
  - Split match computation in sub-tasks to be executed in parallel
  - Exploitation of cloud infrastructures and frameworks like Map/Reduce

Outline

- Motivation
- Blocking-based Object Matching with MapReduce
- Load Balancing
  - Problem
  - Block-Split Approach
- Experimental Results
- Conclusions & Future Work

MapReduce

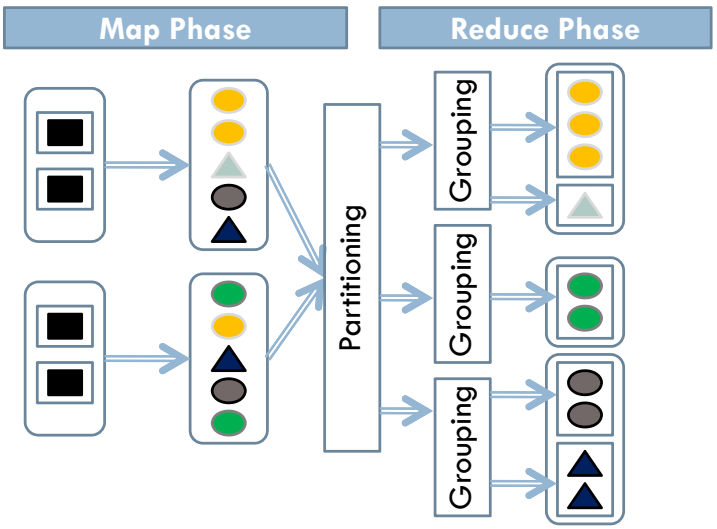
7

- Programming model for distributed computation
- Dataflow defined by map and reduce functions
  - map: $(key_{in}, value_{in}) \rightarrow list(key_{tmp}, value_{tmp})$
  - reduce: $(key_{tmp}, list(value_{tmp})) \rightarrow list(key_{out}, value_{out})$
- MapReduce framework hides all messy details
  - Automatic parallelization
  - Robustness, e.g., handles node failures
  - Scalability
  - ...

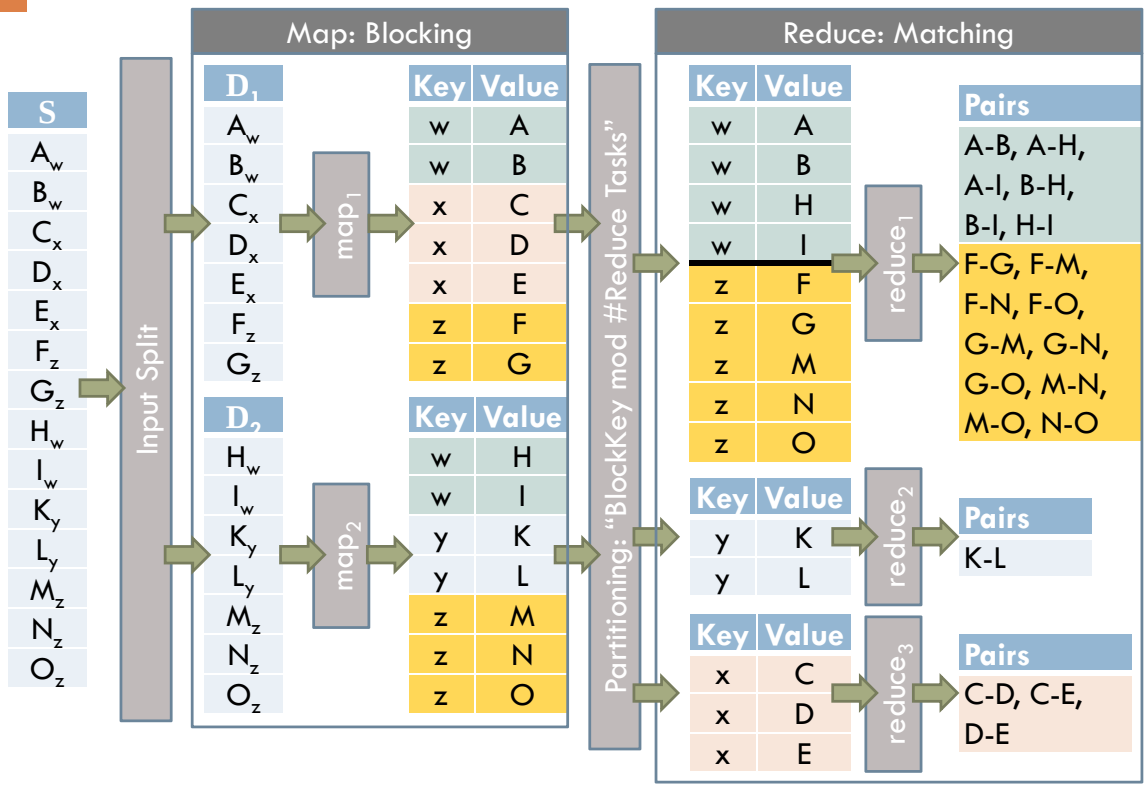
MapReduce

8

- **Map** function applied on each input object to generate **key-value pairs**
- Each key-value pair is assigned to a **reduce task**
- **Reduce** function is invoked for each object group with same key



Blocking + MapReduce: Basic scheme



Load Balancing

- Data skew leads to unbalanced workload
  - Large blocks prevent utilization of more than a few nodes
  - Deteriorates scalability and efficiency
  - Unnecessary costs (you also pay for underutilized machines!)
- Key ideas for load balancing
  - Additional MR job to determine blocking key distribution, i.e., number and size of blocks (per input partition)
  - Global load balancing that assigns (nearly) the same number of pairs to reduce tasks

Load Balancing Approaches

- Two load balancing strategies for parallel object matching with general blocking [ICDE'12]
  - **BlockSplit:** Split large blocks into sub-blocks
  - **PairRange:** Global enumeration and tailored distribution of all pairs
- Variation for Sorted Neighborhood [CSRD'11]

[ICDE'12] Kolb, Thor, Rahm: *Load Balancing for MapReduce-based Entity Matching.* **Proc. Int. Conf. on Data Engineering, 2012 (to appear)**

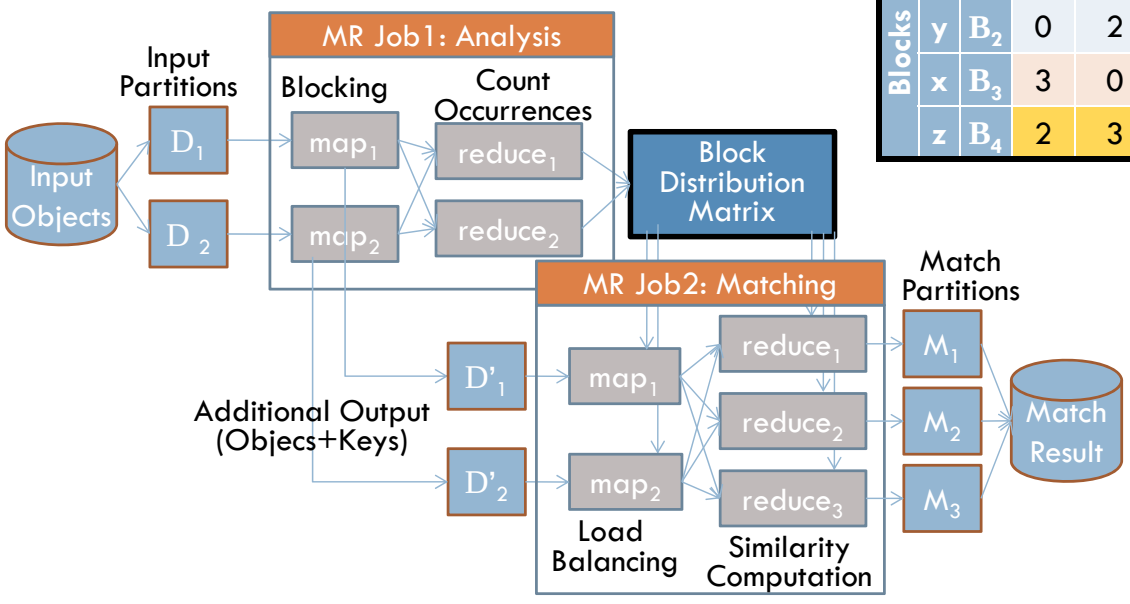
[CSRD'11] Kolb, Thor, Rahm: *Multi-pass Sorted Neighborhood Blocking with MapReduce.* Computer Science - Research and Development, 2011 ("Best of BTW2011")

[BTW'11] Kolb, Thor, Rahm: *Parallel Sorted Neighborhood Blocking with MapReduce.* Proc. BTW, 2011

Load Balancing for MR-based Object Matching

Partition	D_1							D_2						
Object	A	B	C	D	E	F	G	H	I	K	L	M	N	O
Blocking Key	w	w	x	x	x	z	z	w	w	y	y	z	z	z

		Partition		Overall	
		D_1	D_2	#O	#P
Blocks w	B_1	2	2	4	6
Blocks y	B_2	0	2	2	1
Blocks x	B_3	3	0	3	3
Blocks z	B_4	2	3	5	10



BlockSplit

- Large blocks split into *m* sub-blocks
  - according to *m* input partitions
  - large if $\#P_{Block} > \#P_{Overall} / \#Reducer$
- Two types of match tasks
  - Single (small blocks and sub-blocks)
  - Two sub-blocks
- Greedy load balancing
  - Sort match tasks by number of pairs in descending order
  - Assign match task to reducer with lowest number of pairs
- **Example**
  - r=3 reduce tasks, split B_4 in m=2 sub-blocks
  - B_4's match tasks: $B_{4.1}$, $B_{4.2}$, and $B_{4.1\times2}$

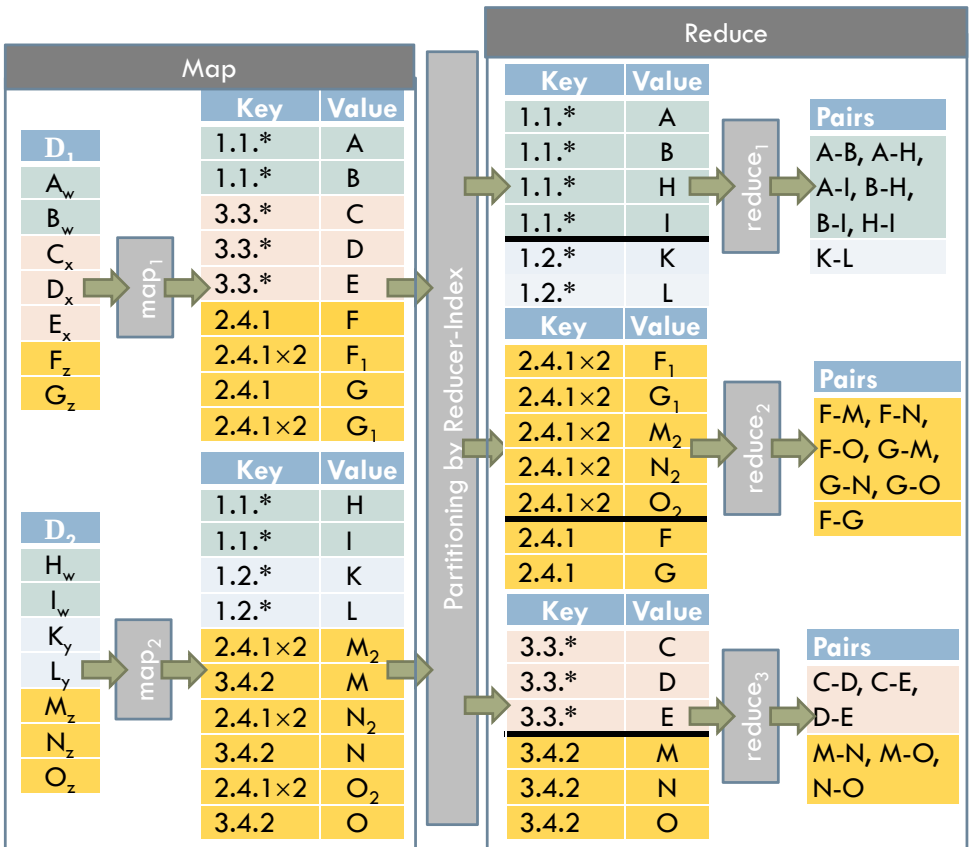
Blocks		Partition D_1	Partition D_2	Overall #O	Overall #P
w	B_1	2	2	4	6
y	B_2	0	2	2	1
x	B_3	3	0	3	3
z	B_4	2	3	5	10

Block Tasks	#P	Reducer
B_1	6	
$B_{4.1\times2}$	6	
B_3	3	
$B_{4.2}$	3	
B_2	1	
$B_{4.1}$	1	

BlockSplit: MR-Dataflow

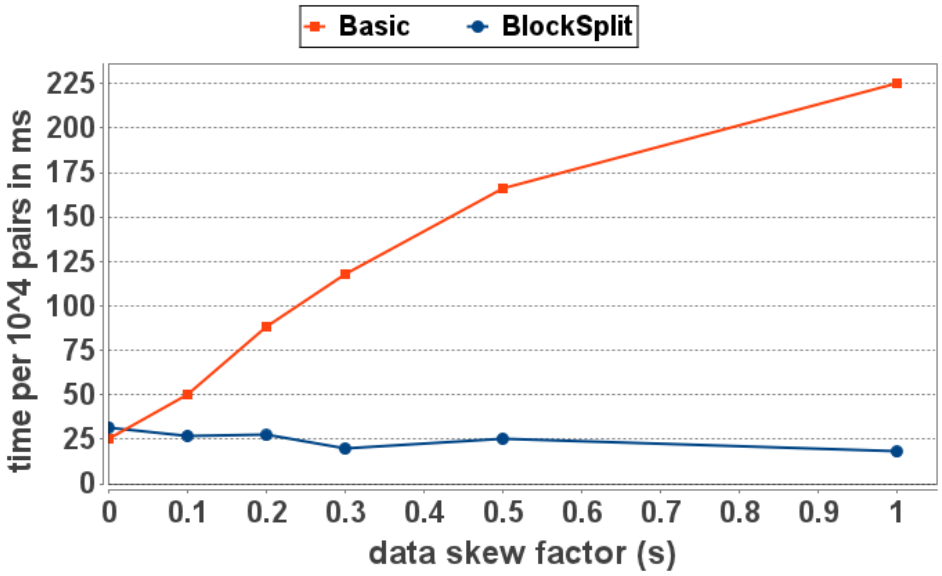
MapReduce Techniques

- MapKey = ReducerIndex + MatchTask
- Replicate objects of sub-blocks



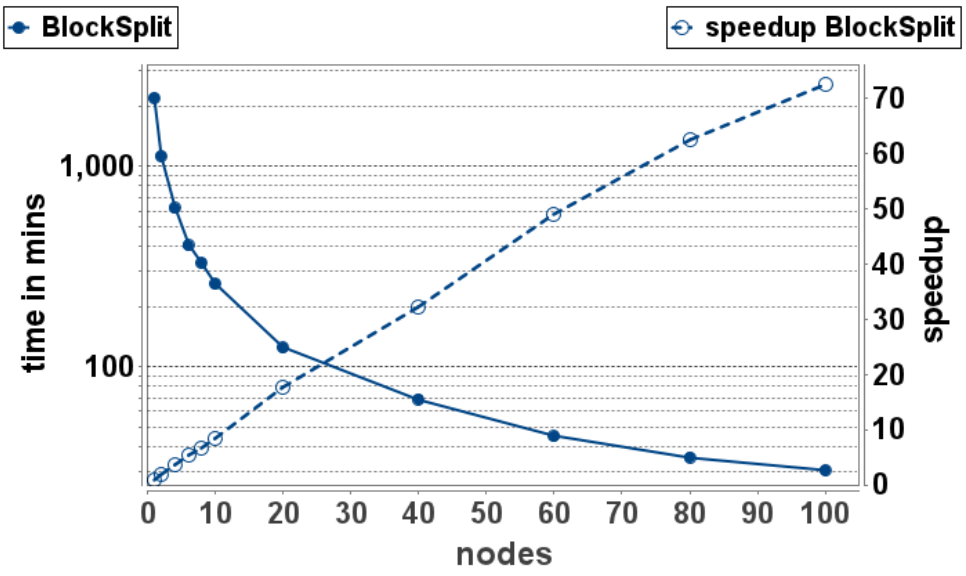
Evaluation: Data Skew

- Evaluation on Amazon EC infrastructure using Hadoop
- Matching of 114.000 product records
- BlockSplit robust against data skew



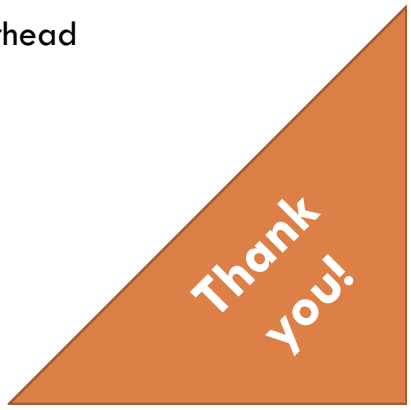
Evaluation: Scalability

- BlockSplit is scalable



Conclusions and Future Work

- Faster object matching by
  - Blocking
  - Parallel matching
- Straight-forward utilization of MapReduce possible
  - ... but doing it efficiently requires some work
- Effective load balancing approaches such as Block-Split
  - Additional MR job for analysis incurs minimal overhead
- Future Work
  - Load balancing for other data-intensive tasks
  - Analytic model for determining #reduce tasks
  - ...



References

- Kolb, L.; Thor, A.; Rahm, E.: Parallel Sorted Neighborhood Blocking with MapReduce. Proc. BTW conf., 2011
- Kolb, L.; Thor, A.; Rahm, E.: *Multi-pass Sorted Neighborhood Blocking with MapReduce*. CSRD 27(1), 2012
- Kolb, L.; Thor, A.; Rahm, E.: *Load Balancing for MapReduce-based Entity Resolution*. Proc. ICDE, 2012
- Koepcke, H.; Thor, A.; Rahm, E.: *Evaluation of entity resolution approaches on real-world match problems*. Proc. VLDB Endowment 3(1), 2010
- Koepcke, H.; Thor, A.; Rahm, E.: *Learning-based approaches for matching web data entities*. IEEE Internet Computing 14(4), 2010
- Koepcke, H.; Rahm, E.: *Frameworks for entity matching: A comparison*. Data & Knowledge Engineering, 2010