Dedoop Efficient Deduplication with Hadoop

Lars Kolb, Andreas Thor, Erhard Rahm Database Group, University of Leipzig http://dbs.uni-leipzig.de

Motivation

What is deduplication?

- Task of identifying duplicates, i.e., entities referring to the same real-world object
- Broad range of applications, e.g.,
 - Duplicate customers in enterprise databases
 - Product offers for price comparison portals

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

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 repu Duplicate Add to Shopping List

Canon VIXIA HF S100 Flash Memory Camcorder



***Canon Video HF S100 Instant Rebate Receive \$200 with your purchase of a ne Canon VIXIA HF S100 Flash Memory Camcorder Add to Shopping List

Canon Vixia Hf S10 Care & Cleaning

Similar but different

\$899.95 new Arlingtoncamera.com 5 seller ratings

\$899.00 new

\$999.00 new

2 seller ratings

Performance Audio

Made in Japan Online

\$2.99 new

Dedoop

Deduplication framework based on MapReduce

- Browser-based specification of deduplication workflows
- Automatic transformation into executable MapReduce workflows
- Automatic load-balancing
- Automatic elimination of redundant pair-comparisons







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

shop.com

★★★☆☆ 38 seller ratings

Deduplication is expensive!

- Pair-wise comparison of input entities
- Application / combination of multiple (domain-specific) similarity measures
- \rightarrow Execution on cloud infrastructure

Workflow configuration

Browser-based specification of advanced deduplication workflows

- Rich toolset of common blocking techniques and similarity functions
- Support of machine learning-based match classifiers

Experiment 1 X + Multiple v	vorkflows
✓ Hadoop Cluster	Vivorktiow Definition
Running Cluster Launch EC2 Cluster Namenode : hdfs://gkpc3.informatik.uni- Jobtracker : gkpc3.informatik.uni-leipzig WebUI port : 50030	Input Data Mode : Self-Join R-S Join Domain Source : hdfs://gkpc3.informatik.uni-leipzig.de/input_data/GoogleS Id Attribute : Attribute 0 Normalize attribute values Output Directory : hdfs://gkpc3.informatik.uni-leipzig.de/output
Connect to cluster (e.g., Amazon EC2)	Blocking Strategy : Standard Blocking (BlockSplit) Key Generator : PrefixBlockingKeyGenera V Attributes : gs_title, gs_authors V Blocking Cenerator :
 ✓ Hadoop Distributed File System Name ▲ Size ☑ input_data ☑ DBLP.txt ☑ GoogleScholar.txt ⑧ guality perfect.csv 238.41KE 	Matching Classification : O Weighted Average / Threshold O Machine Learning Training data file : hdfs://gkpc3.informatik.uni-leipzig.de/input_data/train_500_1 Classifier type : weka.classifiers.functions.LibSVM Classifier Options : -K 0 -C 10
 train_500_1.txt 15.01KB map_reduce output 	Metric : TFIDFSimilarity v Attribute : gs_authors v Metric : NGramSimilarity v Attribute : gs_title v n : 3 •
HDFS file manager	Data Source definition & File Viewer HDFS fileset browser

Workflow management

Automatic mapping of specified workflow(s) to a sequence of MR jobs

Efficient cluster utilization

Basic blocking approach with MapReduce

- Map determine blocking key for every input entity and output (blockkey, entity) pair
- Partitioning by blocking key and block-wise redistribution to r reduce tasks
- Reduce matching of entities of the same block
- 1. Load imbalances
- Susceptible to severe load imbalances due to skewed block sizes
- Execution time dominated by a few tasks that process the largest blocks
- \rightarrow Dedoop: Automatic techniques for balancing workload across all reduce tasks Example: 2 reduce tasks, blocking 9 products by type: 3 MP3 players vs. 6 cell phones

Basic approach

- Simple block distribution (1 block per task)
- BlockSplit approach
- Data Analysis Job: Split large blocks into sub-blocks

4

2

9

VS.

9

pairs

Block distribution based on size of sub-blocks

Load balancing





- Automatic workflow submission to Hadoop cluster incl. progress monitoring
- Simultaneous handling of multiple workflows and multiple clusters
- Convenient cluster management (e.g., automatic launching of Amazon EC2 VMs)
- **Screenshots**



- Entities can be assigned to multiple blocks (e.g., multi-phase blocking)
- Parallel execution may lead to unnecessary comparisons (same pair in multiple blocks)
- \rightarrow Dedoop: Automatic techniques for eliminating redundant pair comparisons

Hadoop Distributed File System