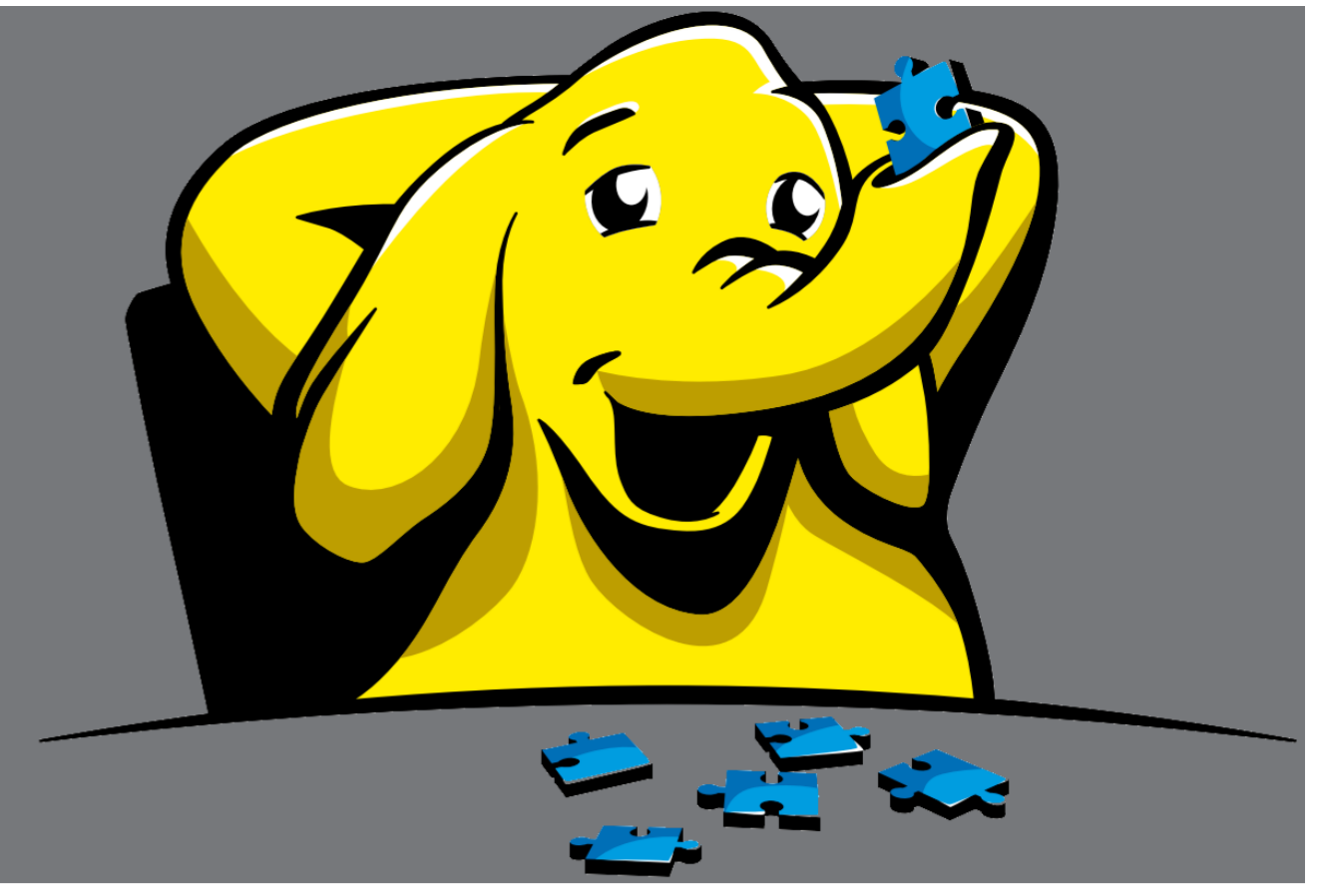


# Dedoop

## 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

	<b>Canon (VIXIA) HF S10 iVIS Dual Flash Memory Camcorder</b> Canon HF S10 iVIS Dual Flash Memory CamcorderSPECIAL SALE PRICE: \$899 Display both English/Japanese + we supplu all English manuals in English as PDF. ... <a href="#">Add to Shopping List</a>	<b>\$899.00</b> new Made in Japan Online
	<b>Canon VIXIA HF S10</b> Dual Flash Memory High Definition Camcorder The Next Step Forward in HD Video Canon has a well-known and highly-regarded reputation. <a href="#">Add to Shopping List</a>	<b>\$999.00</b> new Performance Audio 2 seller ratings
	<b>Canon VIXIA HF S100 Flash Memory Camcorder</b> ***Canon Video HF S100 Instant Rebate Receive \$200 with your purchase of a new Canon VIXIA HF S100 Flash Memory Camcorder. <a href="#">Add to Shopping List</a>	<b>\$899.95</b> new Arlingtoncamera.com 5 seller ratings
	<b>Canon Vixia Hf S10 Care &amp; Cleaning</b> Care & Cleaning Digital Camera/Camcorder Deluxe Cleaning Kit with LCD Screen Guard Canon VIXIA HF S10 Camcorders Care & Cleaning. <a href="#">Add to Shopping List</a>	<b>\$2.99</b> new shop.com 38 seller ratings

Labels: Duplicate, Similar but different, Accessory product

#### Deduplication is expensive!

- Pair-wise comparison of input entities
  - Application / combination of multiple (domain-specific) similarity measures
- 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

Multiple workflows

Connect to cluster (e.g., Amazon EC2)

Input file(s)

Blocking approach

Matching approach (e.g., SimMeasures)

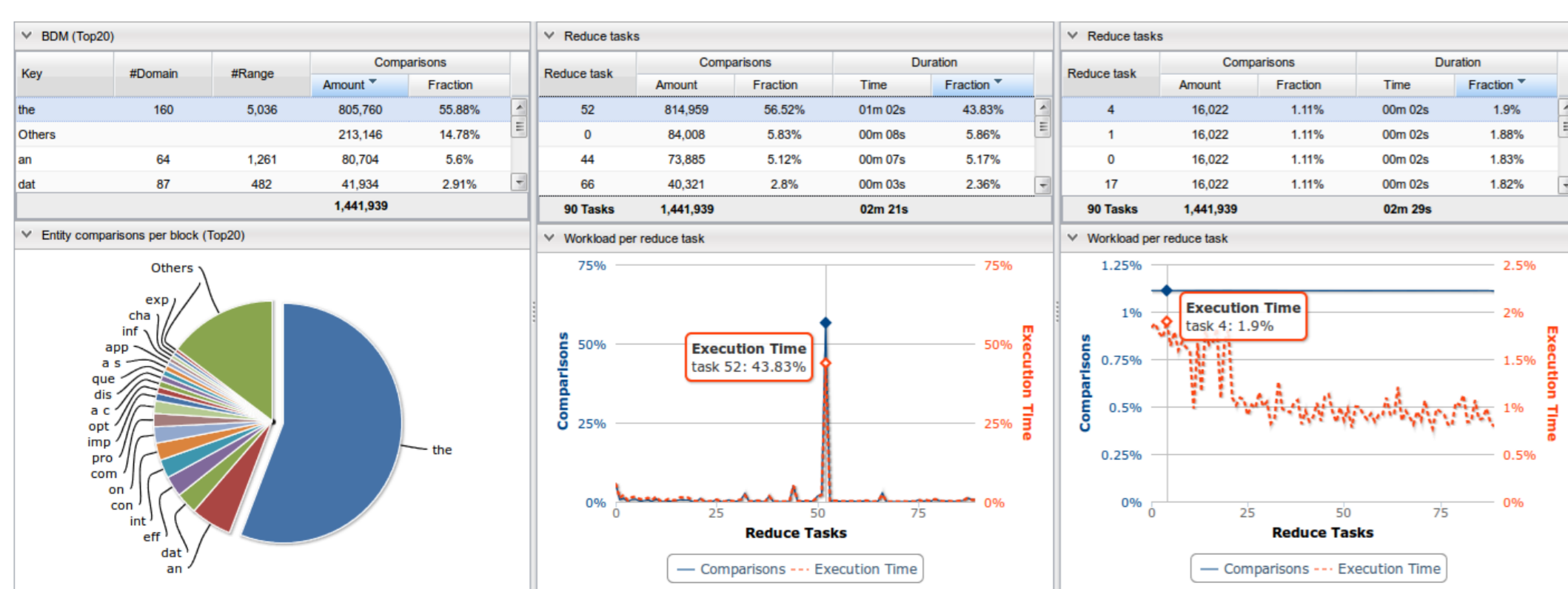
Data Exchange via HDFS file manager

HDFS fileset browser

#### Workflow management

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

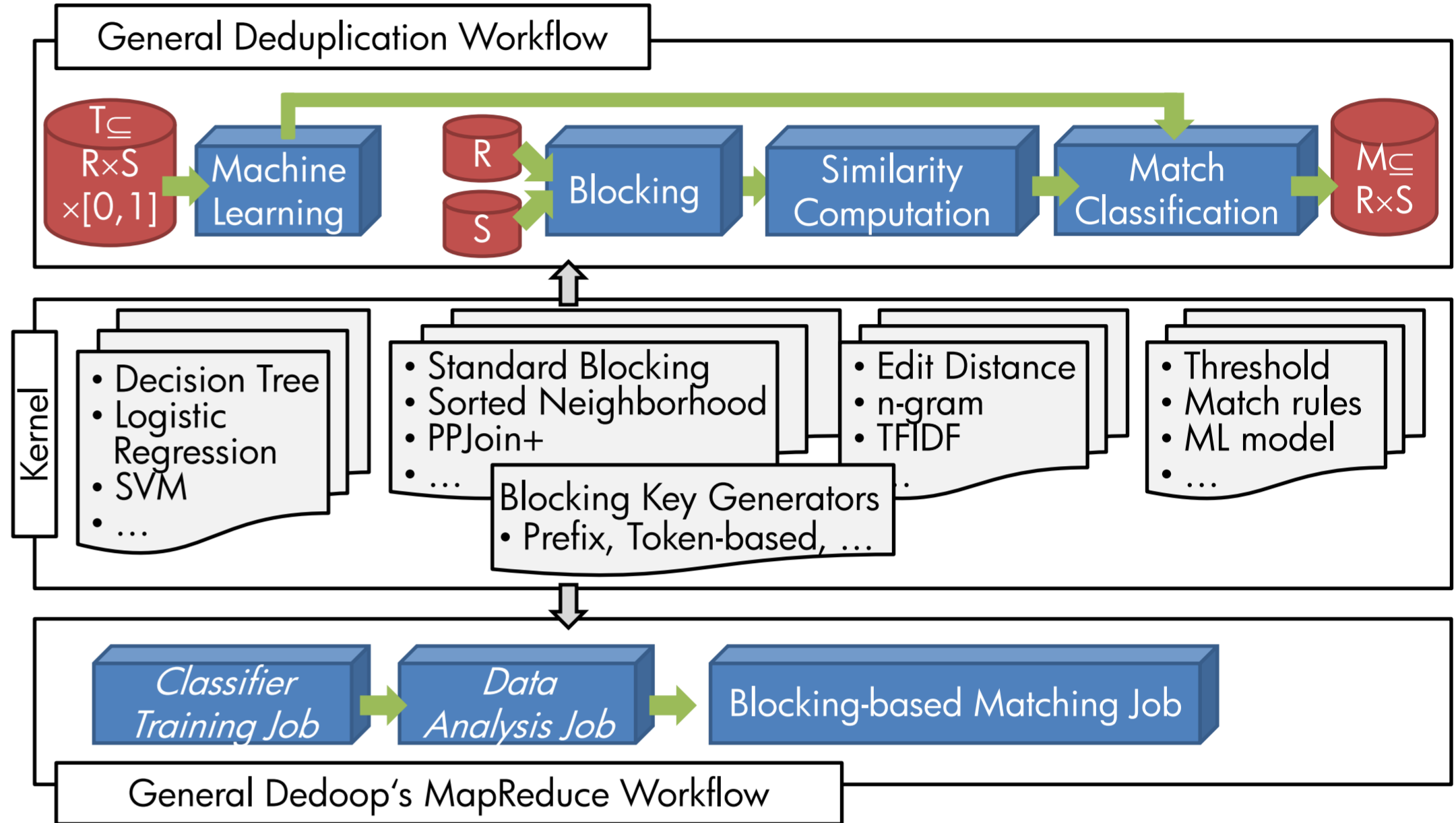
### Screenshots



### 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



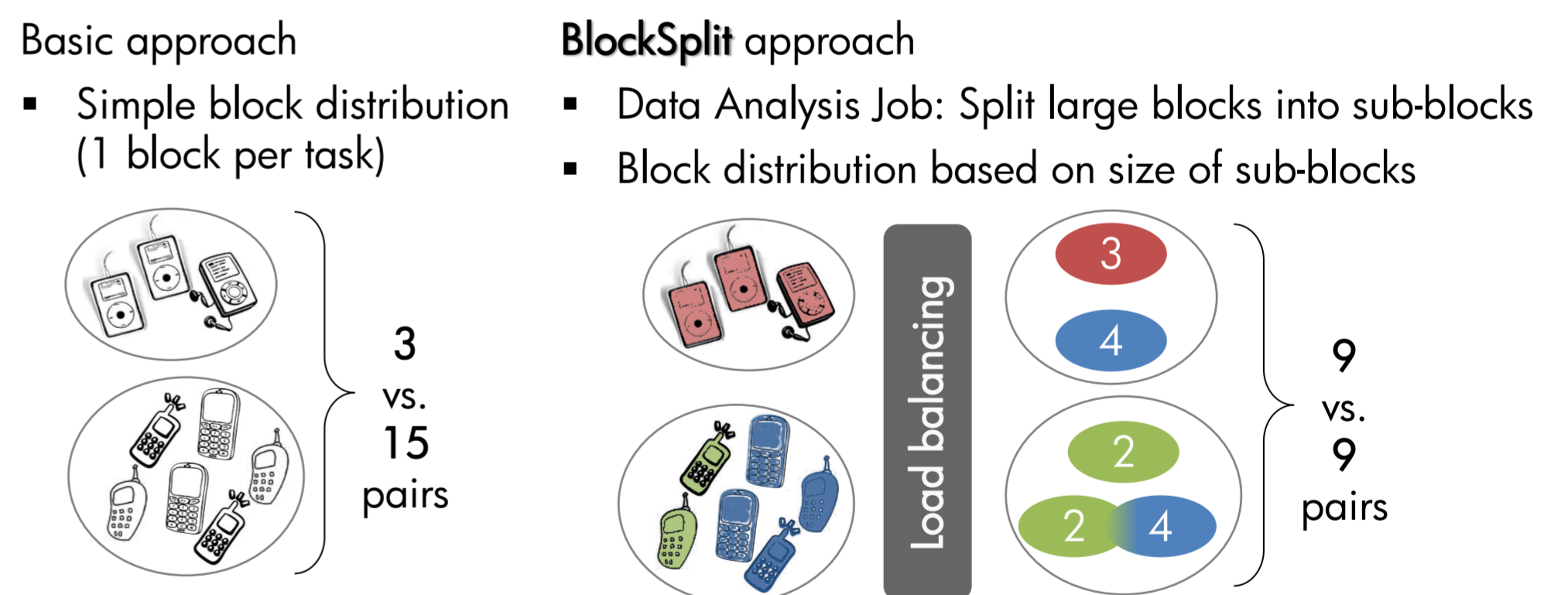
### Efficient cluster utilization

#### 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
- 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



#### 2. Redundant pair-comparisons

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

HDFS fileset browser

Progress monitoring

Initialization of EC2 VMs and Hadoop cluster setup