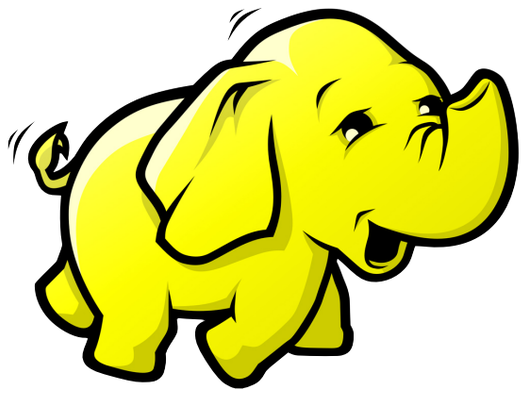
7 Things To Know

When Buying  for an  !

Alekh Jindal, Jorge Quiané, Jens Dittrich

1

What Shoes? Why Shoes?

Analyzing MR Jobs

(HadoopToSQL, Manimal)

Generating MR Jobs

(PigLatin, Hive)

Executing MR Jobs

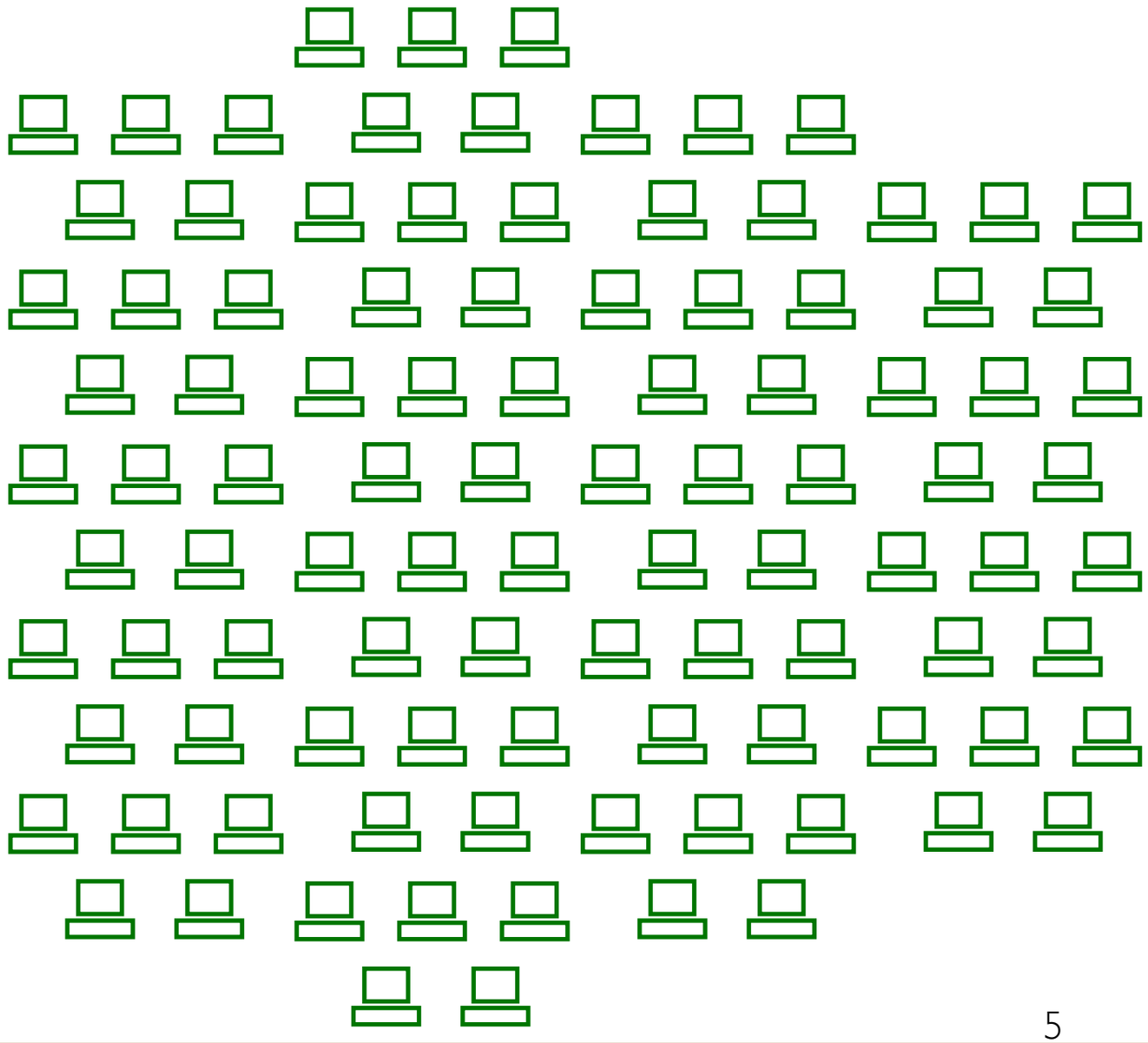
(Hadoop++, epiC)

Data Layouts & Access Paths !!

2

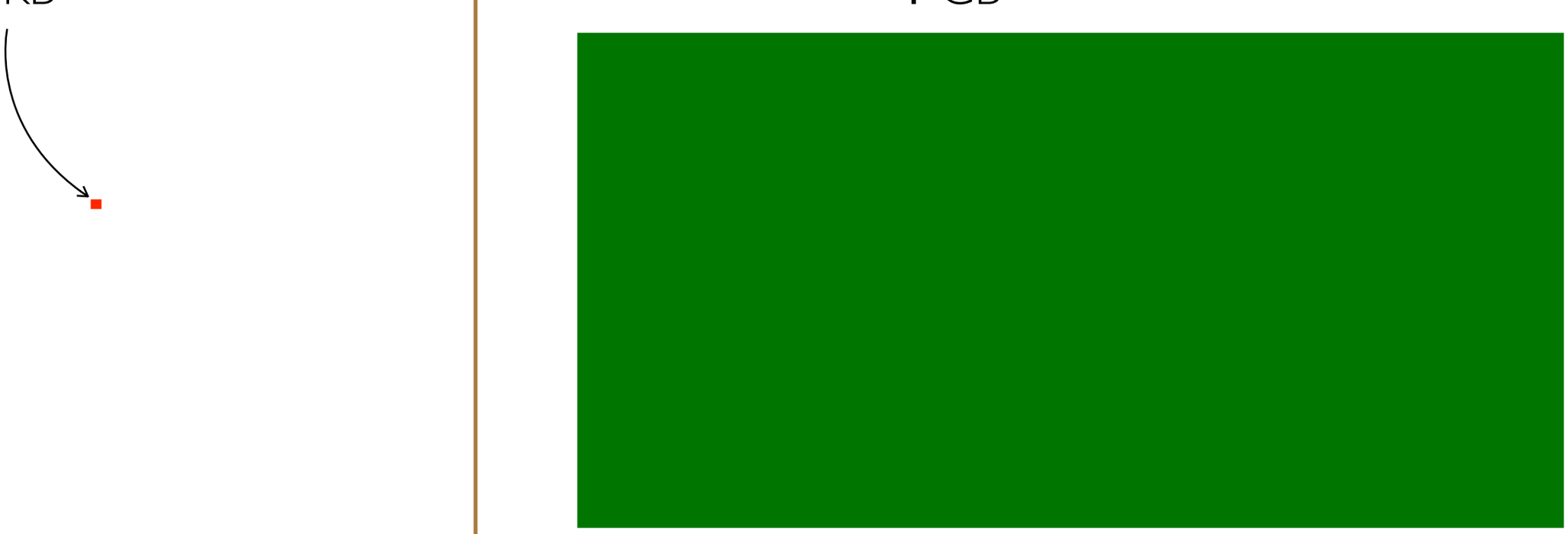
Why Elephant Needs Different Shoes?

Very Large Scale Storage & Execution

DBMS	MapReduce
	

Large Data Block Sizes

DBMS	MapReduce
8 KB	1 GB



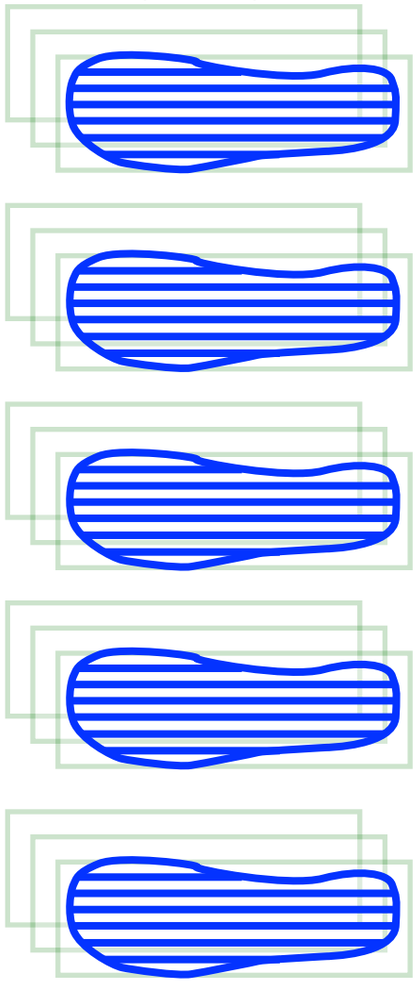
Block Level Data Replication

DBMS	MapReduce
	

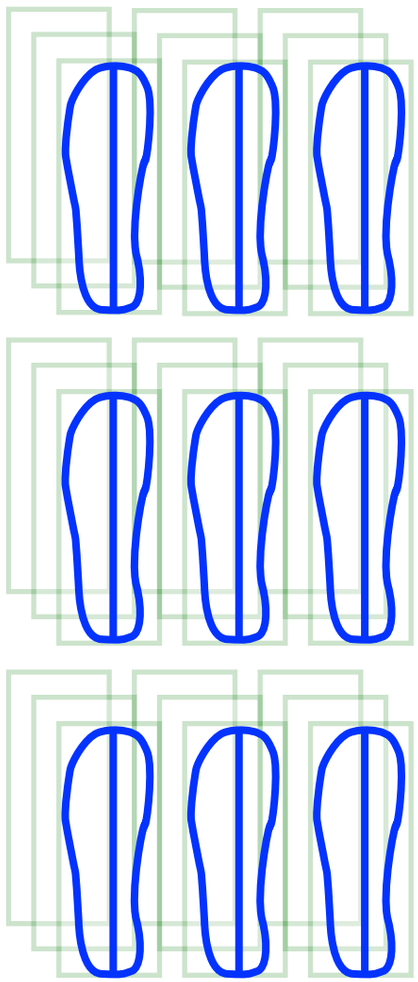
What’s Wrong with Old Shoes?

Current Data Layouts in Hadoop

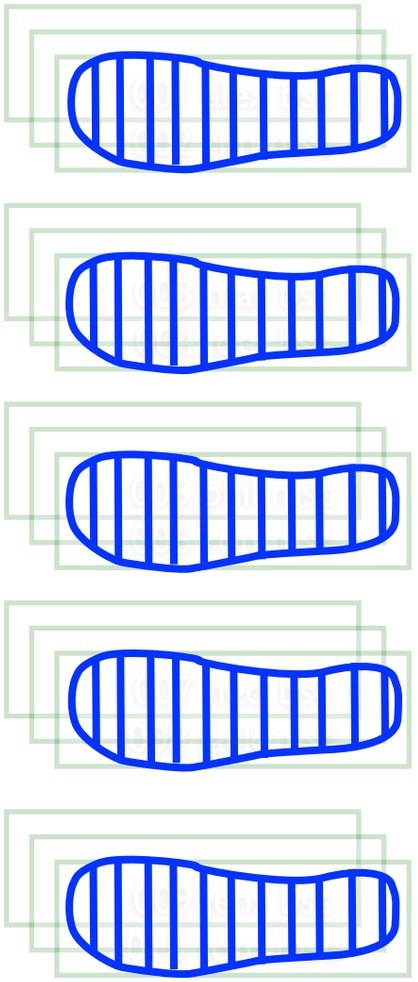
Row
(default)



Column*



PAX**



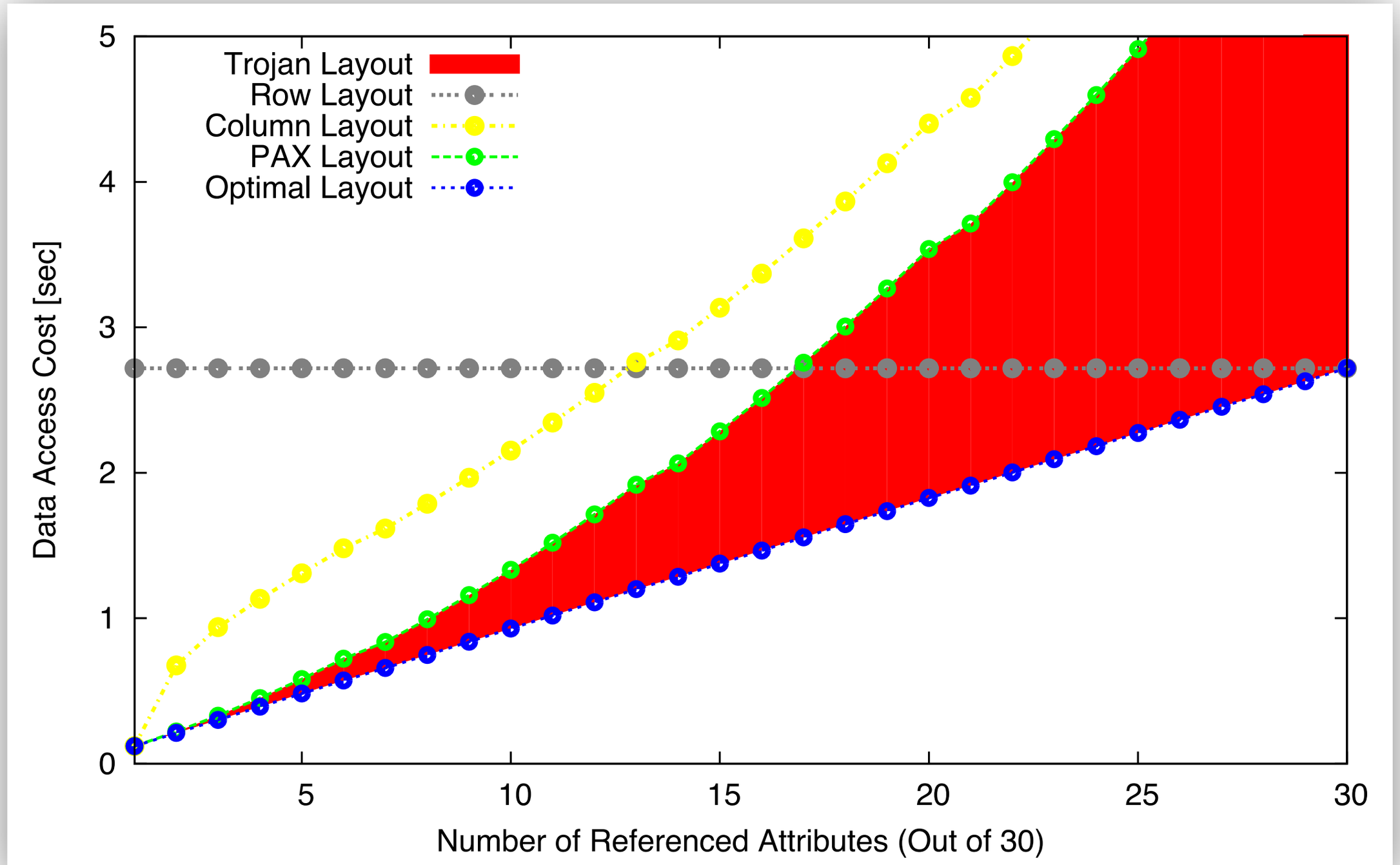
* A. Floratou et al. Column-Oriented Storage Techniques for MapReduce. PVLDB, April, 2011

** Y. He et al. RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems. ICDE, 2011

Current Data Layouts in Hadoop

	Row	Column	PAX
Non-required Reads			
Network Costs			
Data Block Placement			
Tuple Reconstruction			

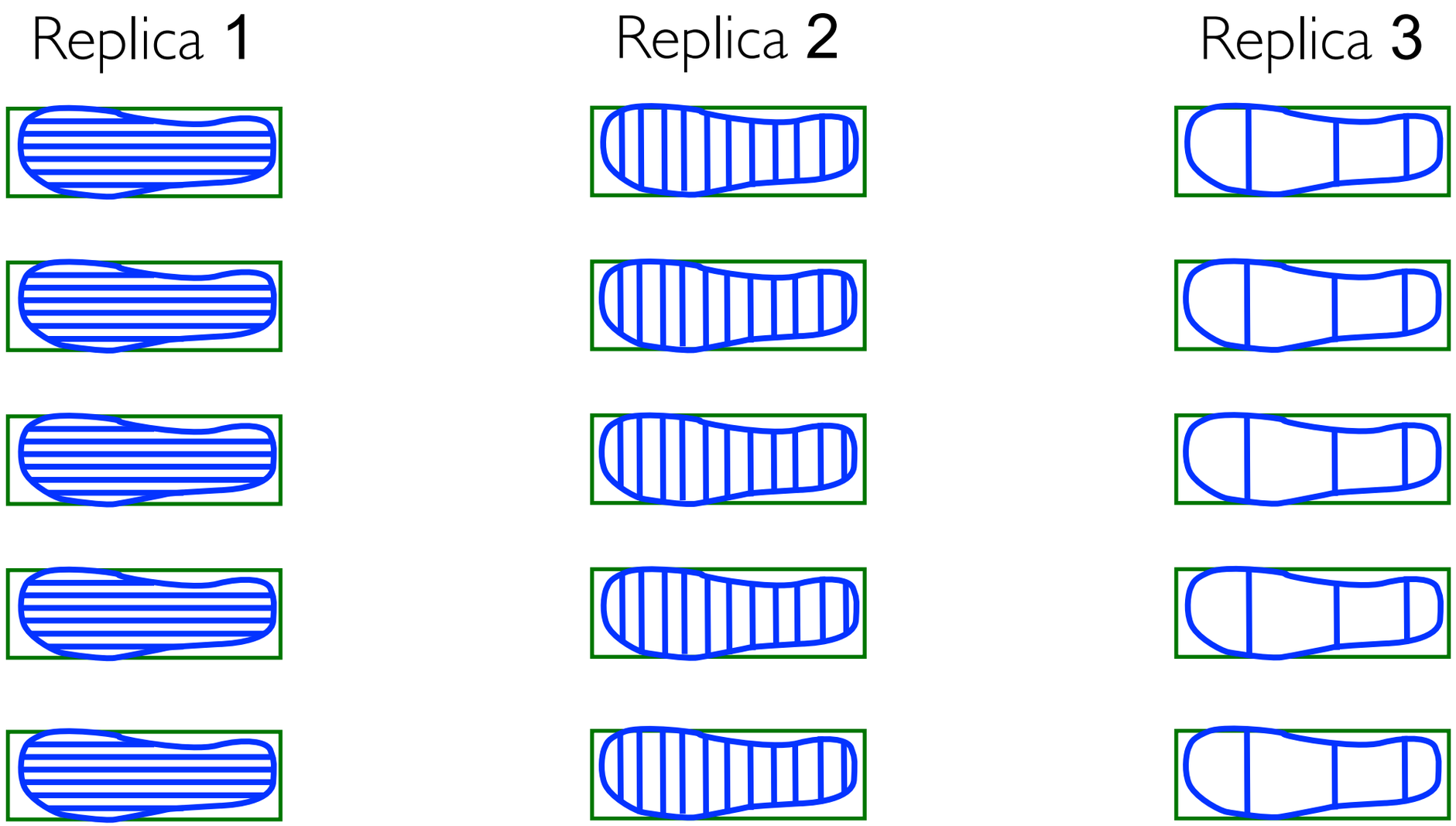
Current Data Layouts in Hadoop



4

What Shoes do We Propose?

Trojan Data Layouts



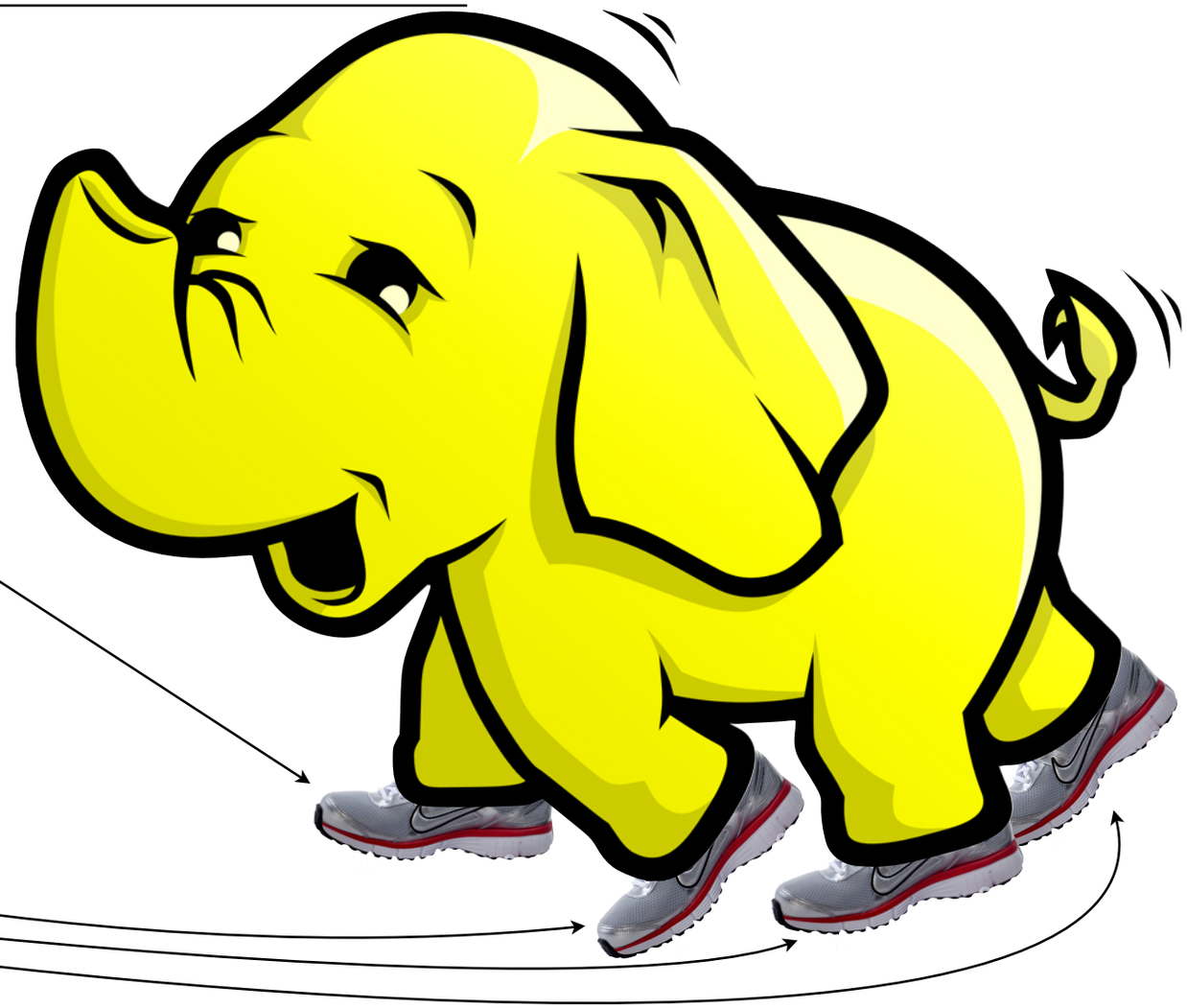
Trojan Data Layouts

	Row	Column	PAX	Trojan
Non-required Reads				
Network Costs				
Data Block Placement				
Tuple Reconstruction				

Challenges in Trojan Data Layouts

How do we design shoe for one leg?

How do we design shoes for all legs?



How do we make the shoes from the design?

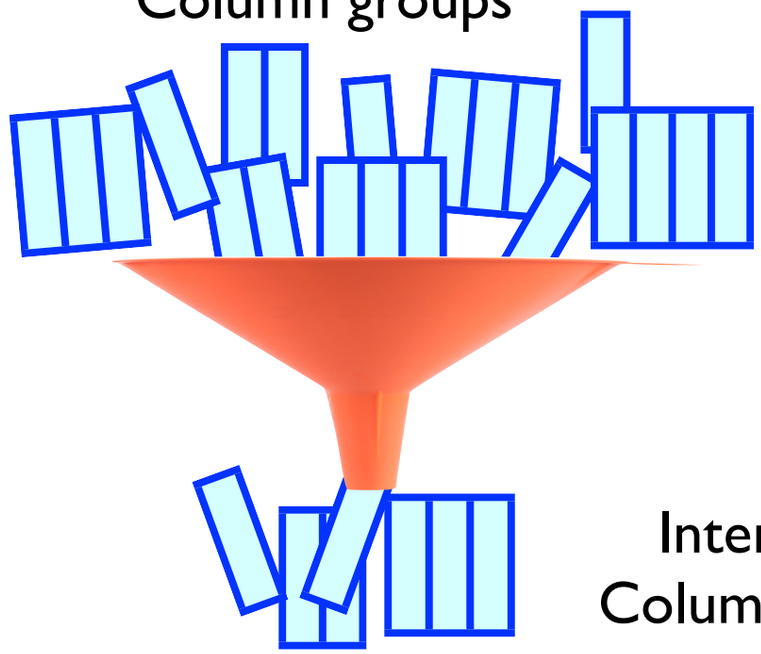
How Do We Design the Shoes?

Single Replica

Columns

Column groups

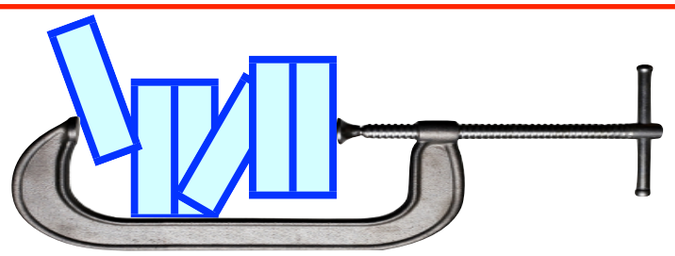
Filter

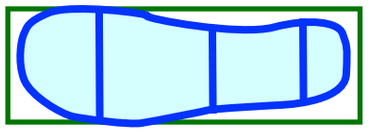


Interesting
Column groups

Novel Column Group
Interestingness

Pack



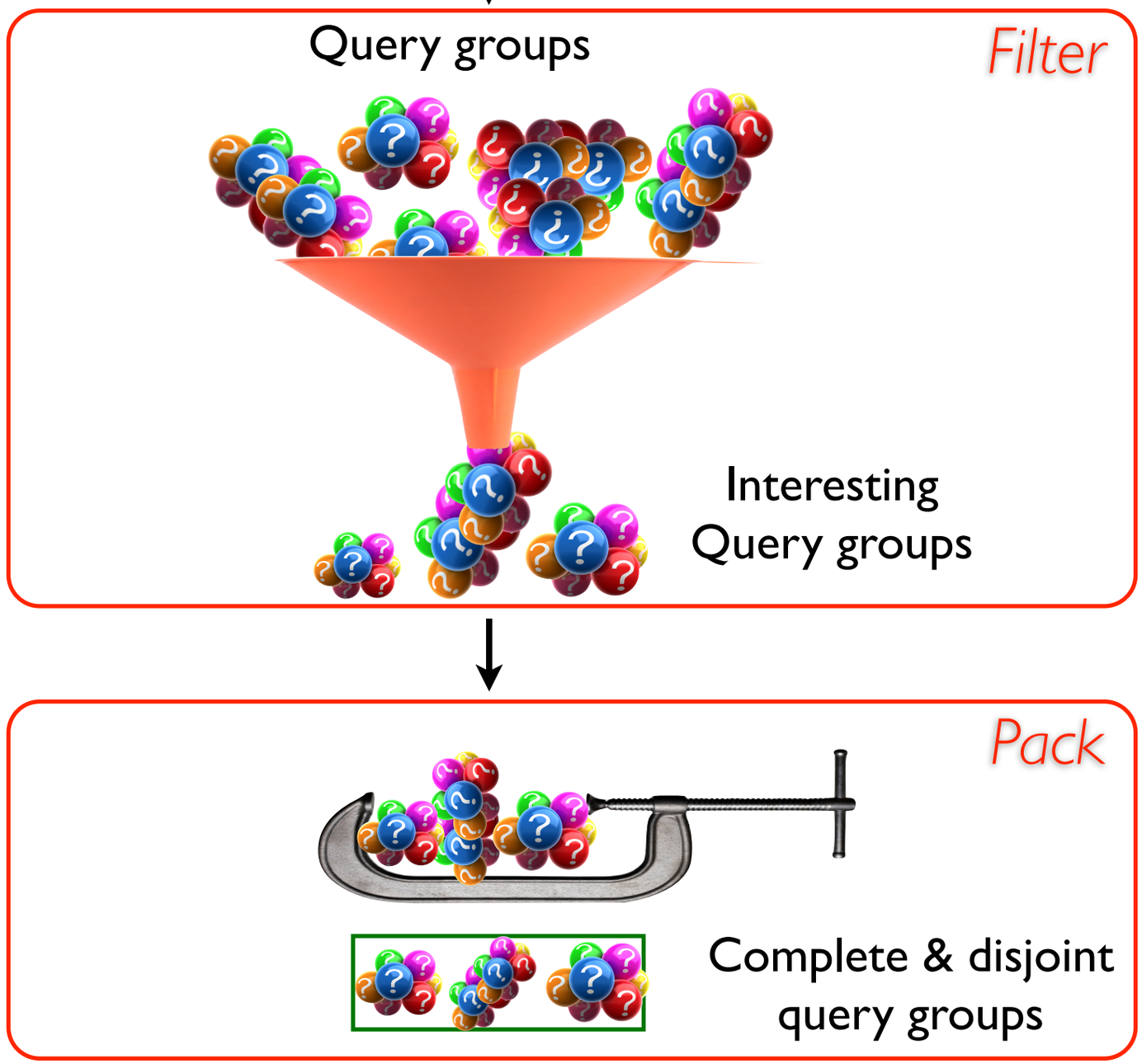


Complete & disjoint
column groups

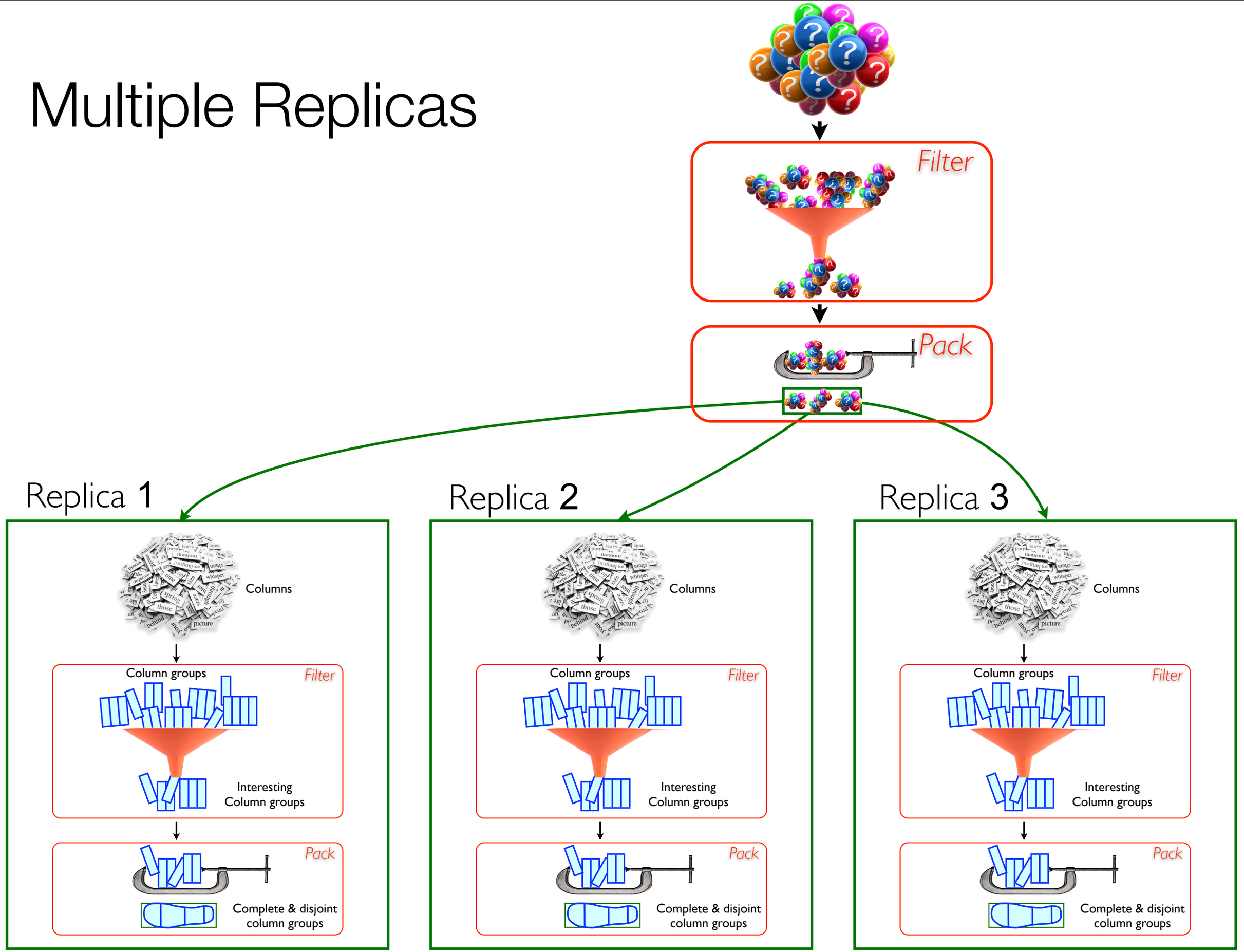
Column Group Packing
as 0-1 Knapsack

Multiple Replicas

Queries

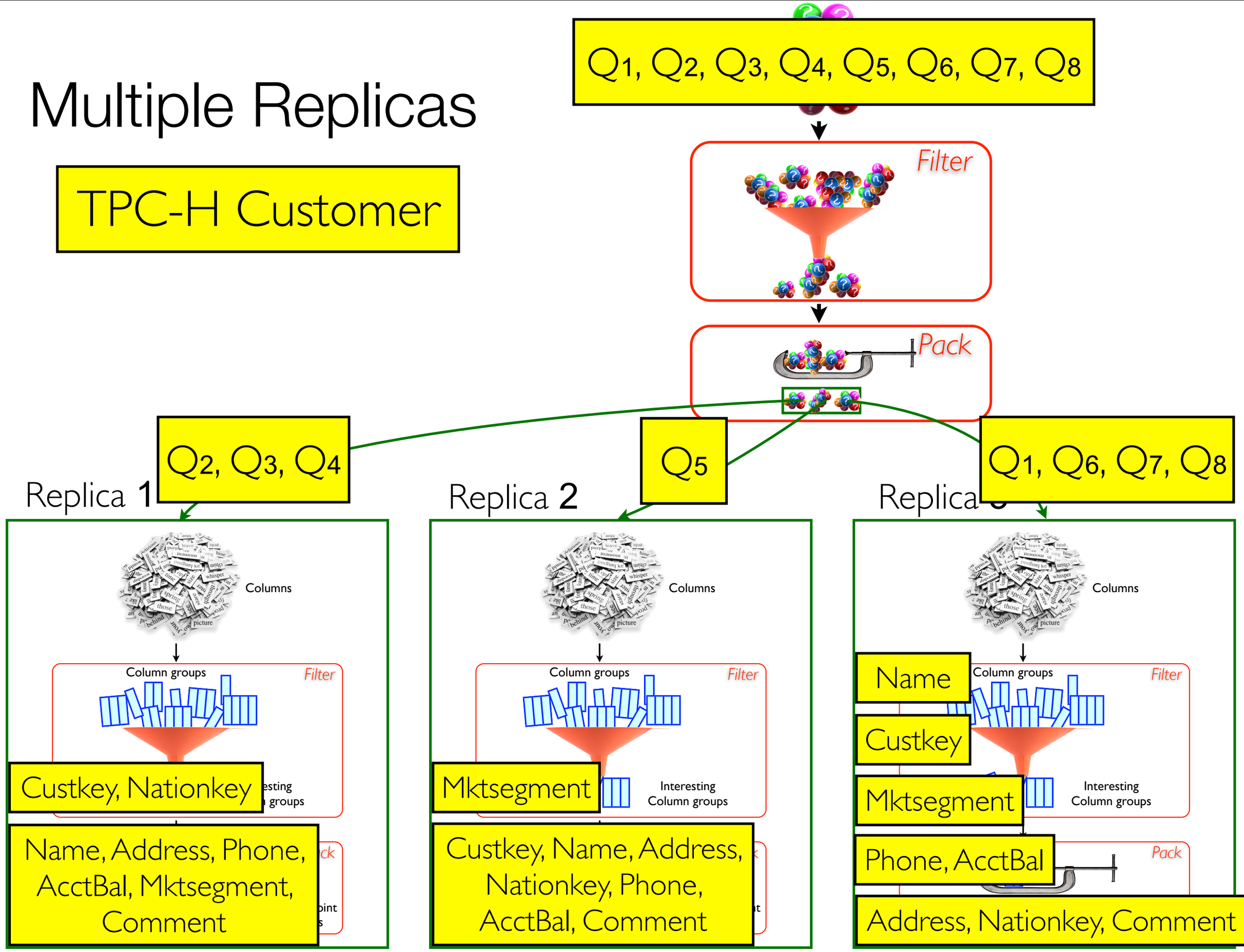


Multiple Replicas



Multiple Replicas

TPC-H Customer

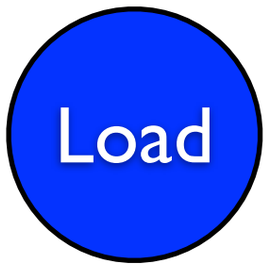


Trojan Layout Advantages

- Multiple layouts for a given workload
- Default row layout still available
- Specialized replicas for different query sub-class
- Divide and conquer layout computation

How do We Ride the Elephant?

Putting It All Together



Create trojan layout configuration file in HDFS

dataset layout-**1** layout-**2** layout-**3**

Supply referenced attributes in JobConf

itemize UDF to transparently read the referenced attributes

Three Optimization Options:

- data locality (default)
- best layout
- best layout & locality

7

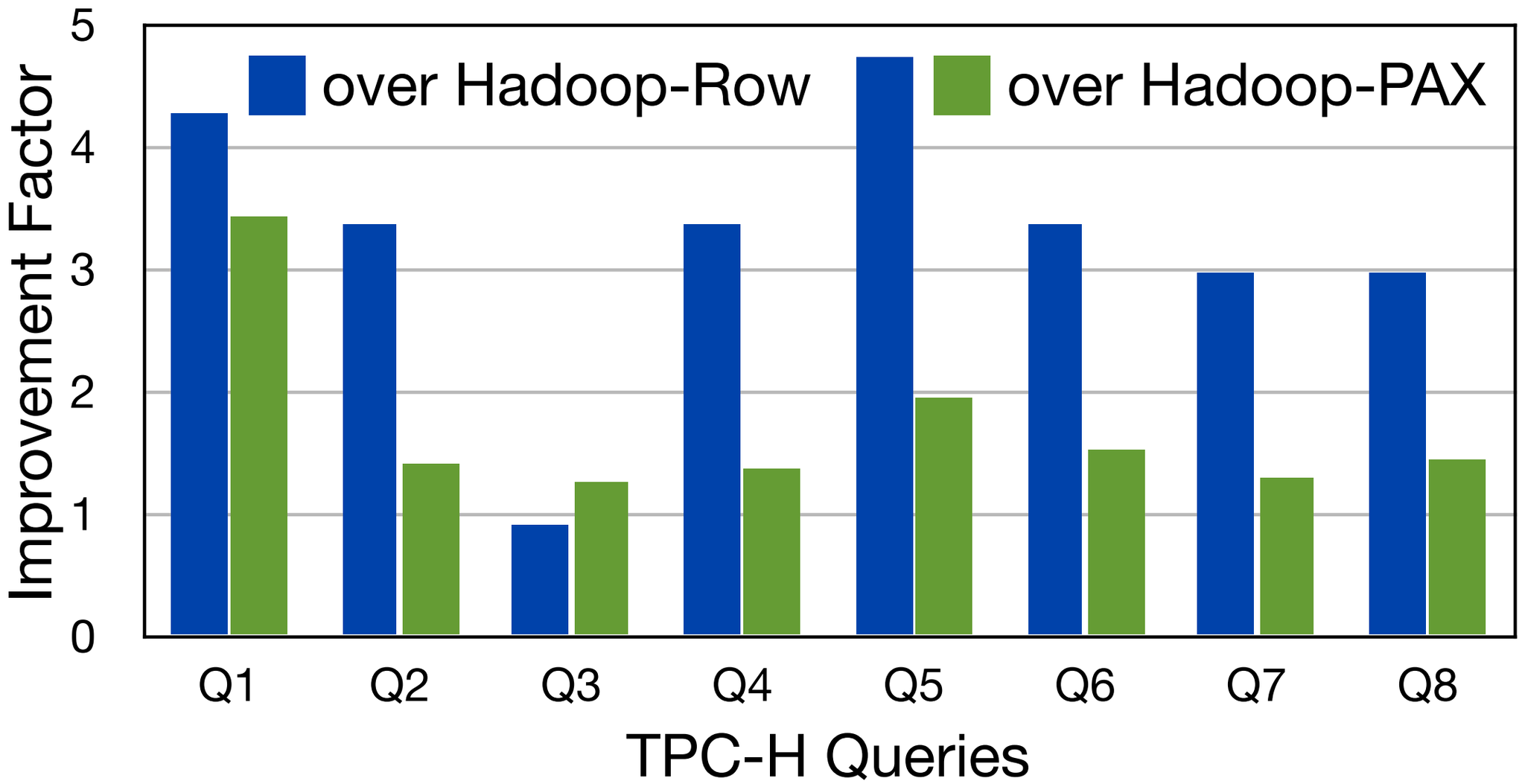
How were the Field Trials?

Setup

- Datasets
  TPC-H Lineitem, TPC-H Customer, SSB LineOrder, SDSS PhotoObj
- Queries
  First 8 queries from the respective benchmark for each table
- Methodology
  focus on scan and projection operators i.e. map-phase-only jobs
  improvement: record reader time (I/O and tuple reconstruction)
- Hardware
  50 virtual nodes in a 10 node cluster

Per-replica Trojan Layout Performance

TPC-H Lineitem



Layout Quality

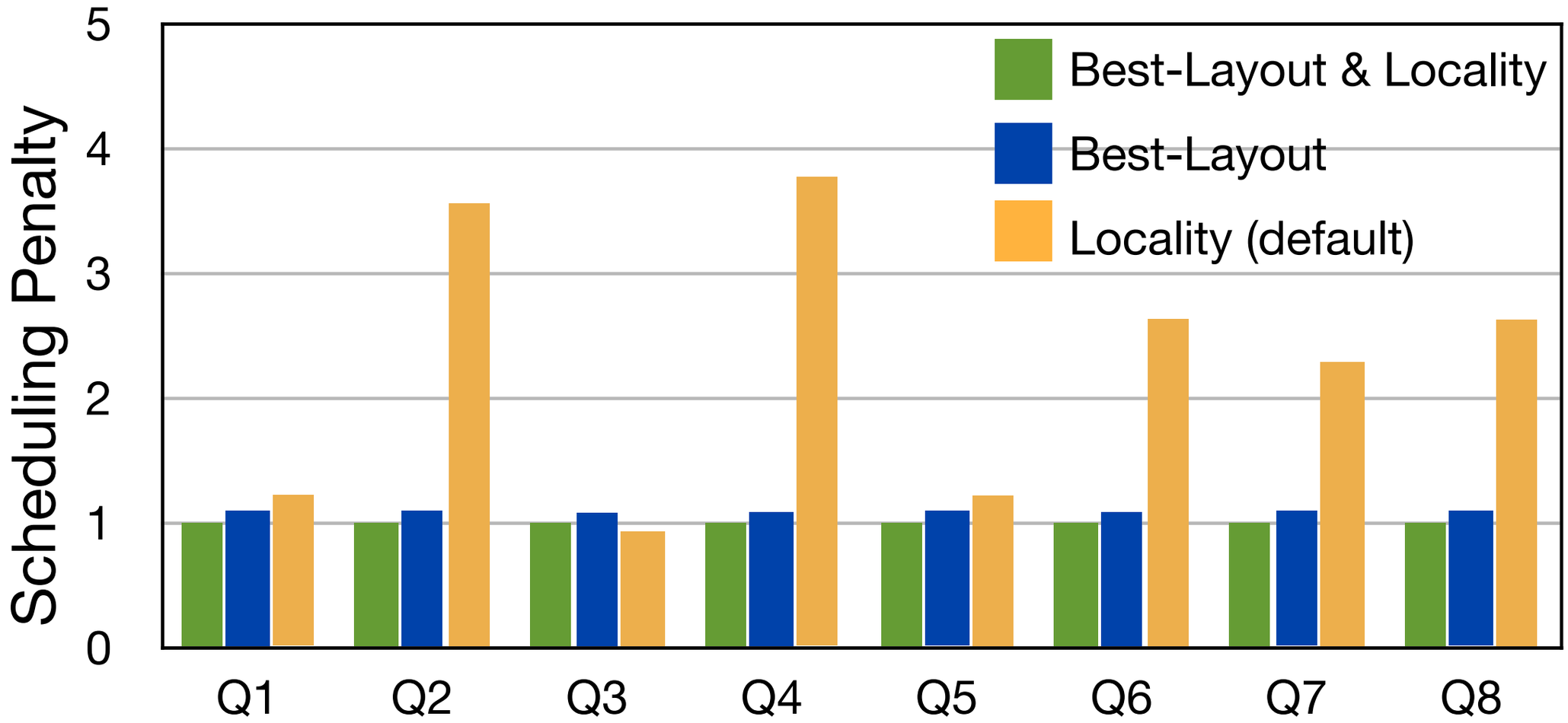
	#Non-required Attributes Read	#Joins in Tuple Reconstruction
HADOOP-ROW	525	0
HADOOP-PAX	0	139
HYRISE* Layout	2	64
Trojan Layout	14	20

>14% improvement over HYRISE

* M. Grund et al. HYRISE - A Main Memory Hybrid Storage Engine. PVLDB, November, 2010.

Scheduling Decisions

TPC-H Lineitem



Summary

- Data layouts crucial to MR job performance
- Exploit default data block replication in MR
- Novel algorithm to compute per-replica layouts
- Improvement: **4.8**x over Row, **3.5**x over PAX
- Better than HYRISE; **14**% improvement