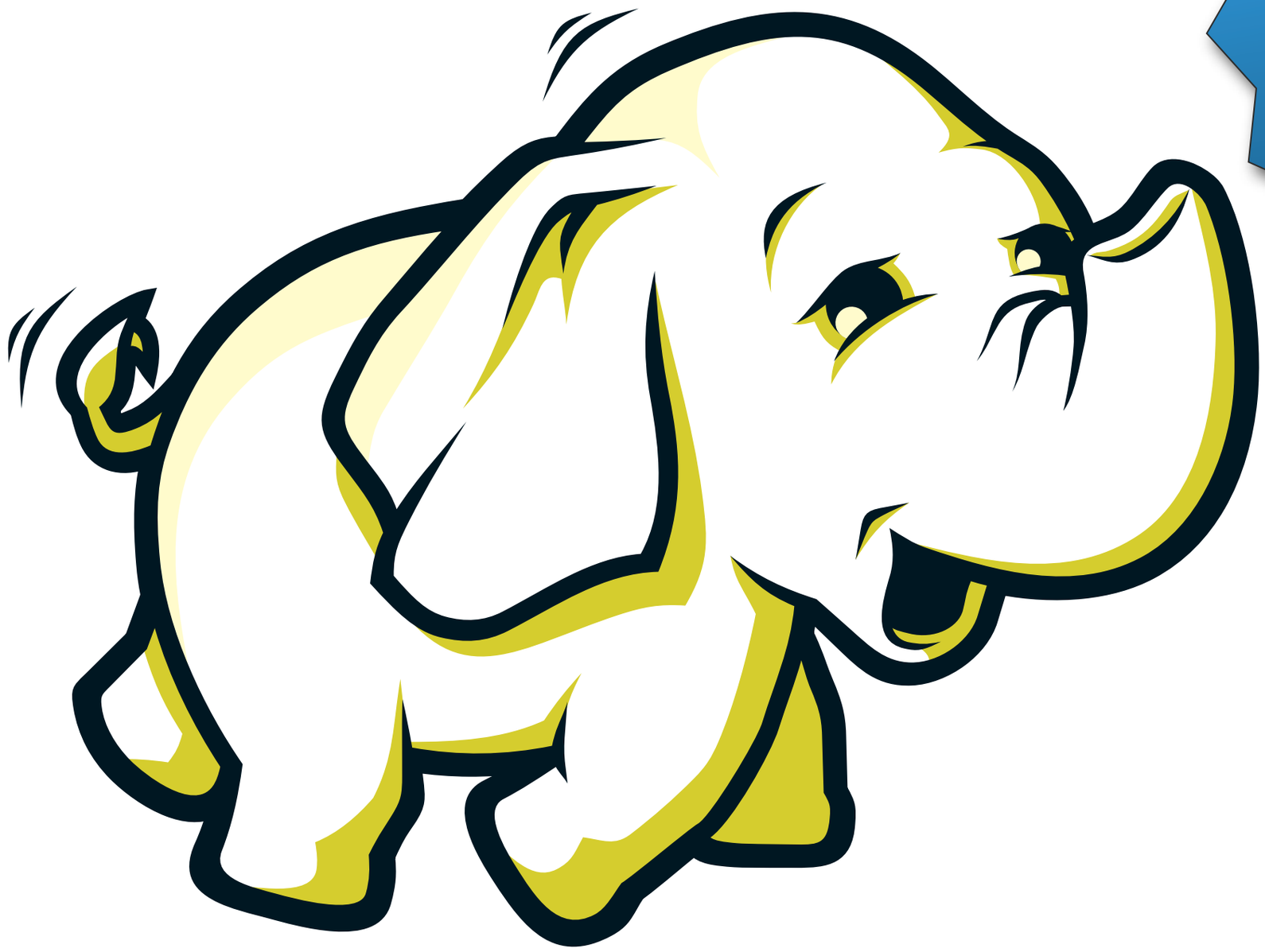
Datameer

Buzzword Berlin



Hadoop: A Reality Check

#1

(intro)

My **Path**

nutch

hadoop

hp

NOKIA
Connecting People

katta

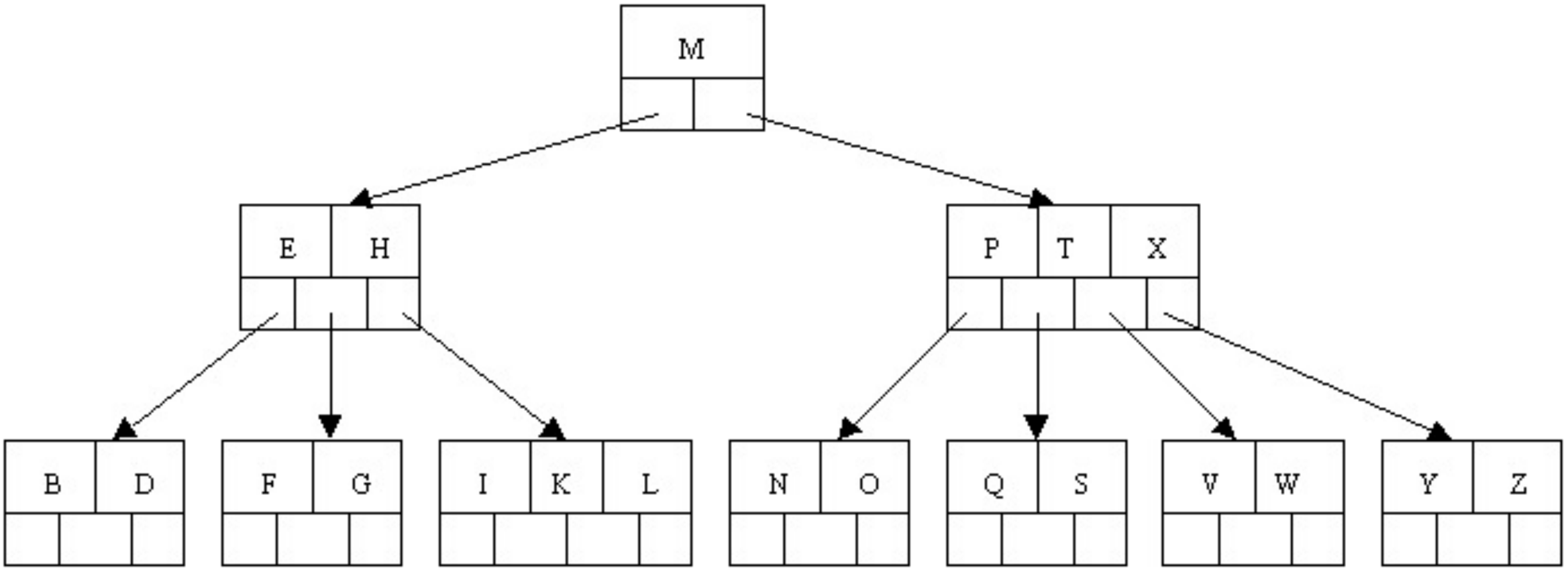
EMI

Datameer

Database Harddrive

time to find one record	=	$\log_b N * 10ms$
$\log_{100}(100{,}000{,}000) * 10ms$	=	40ms
time to read record	=	10ms
10,000,000 * 50ms	=	**5.8 days**

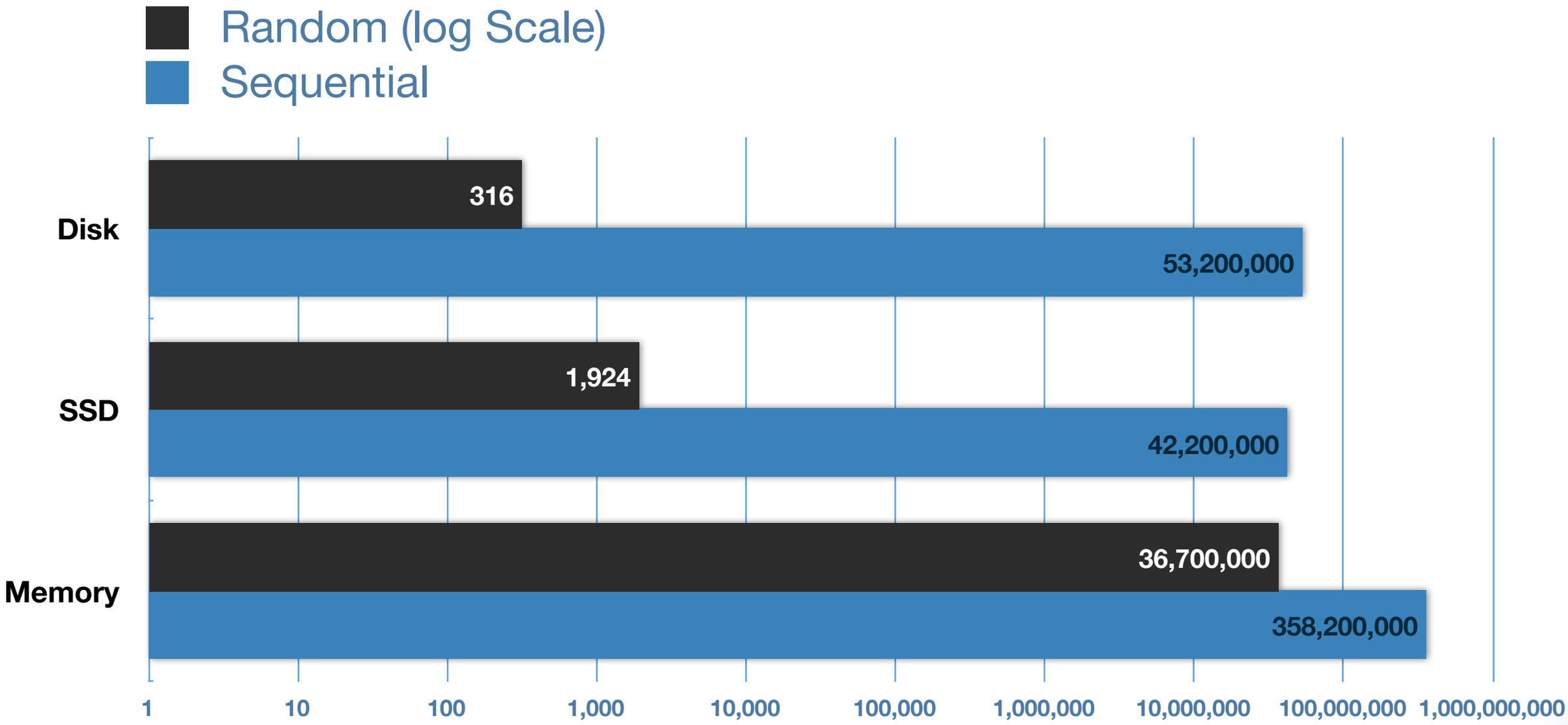
B-Tree



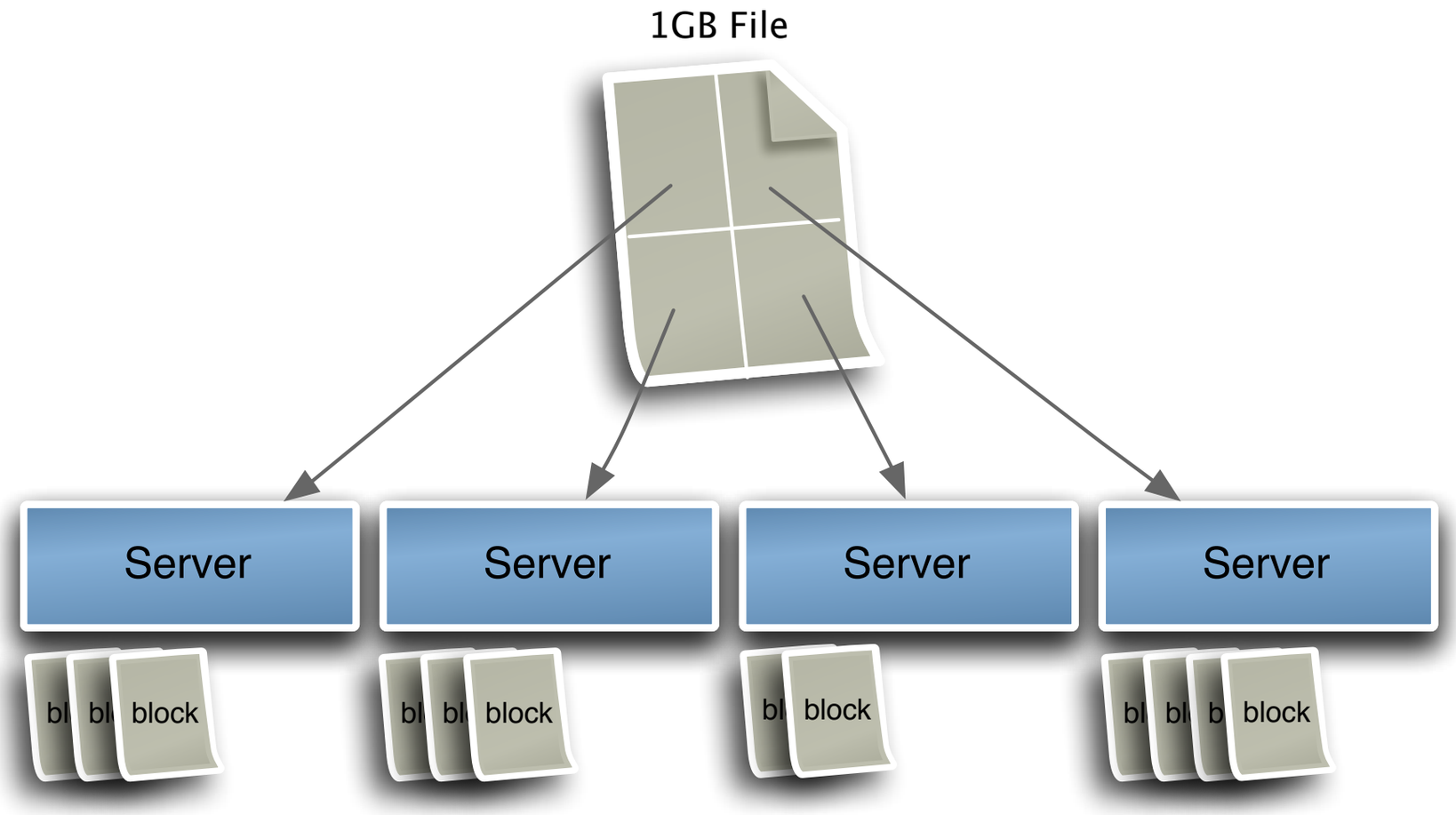
Hadoop Harddrive

throughput	=	10MB/s
time to transfer record	=	10ms
10,000,000 * 10ms	=	**1.5 days**
random reads	=	**(5.8 days)**

Laws of Physics

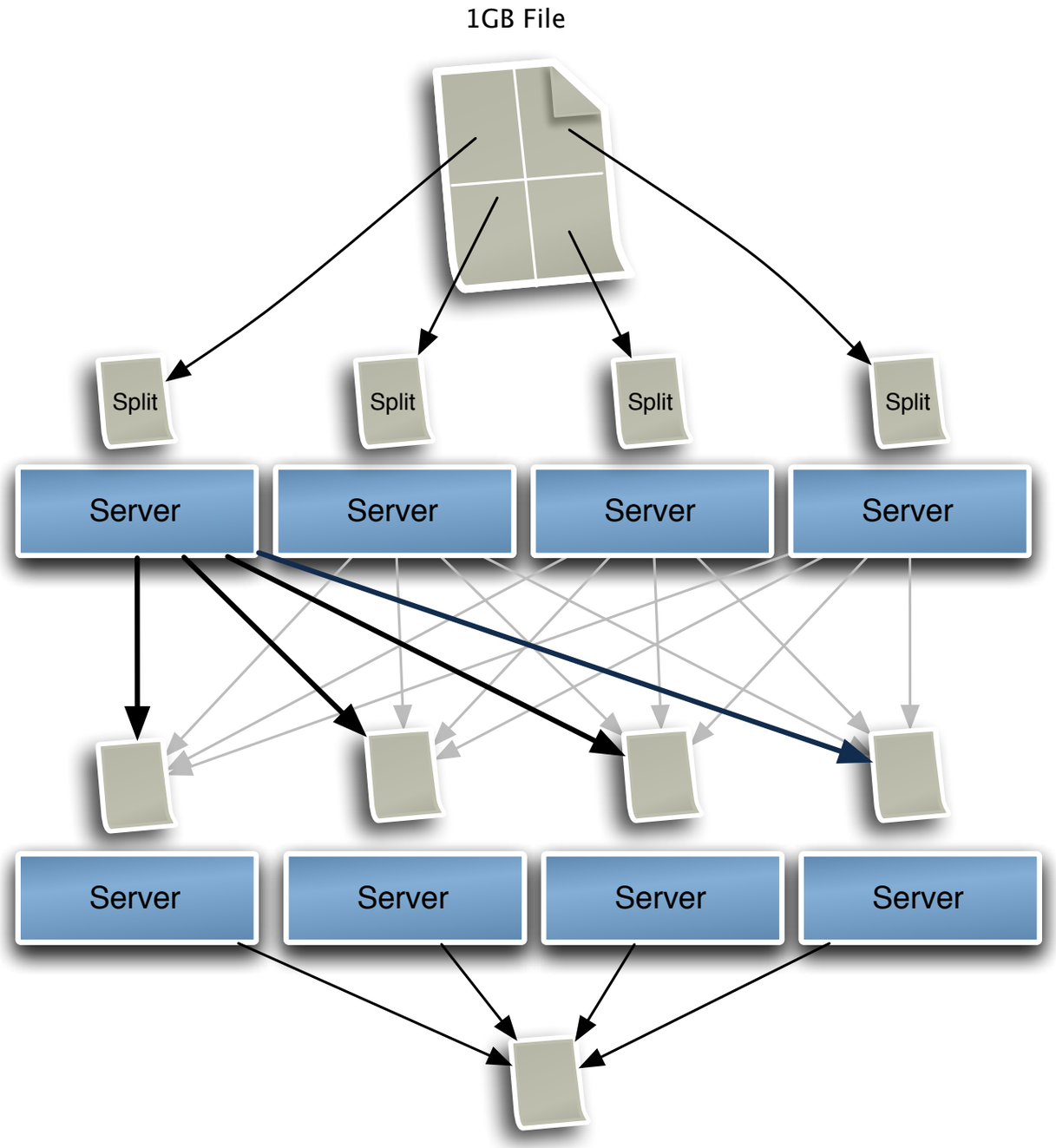


Hadoop File System



- **Files are split into blocks**
- **Blocks are stored on different servers**
- **Blocks are replicated**

Hadoop Computation



- Map Reduce
- Files are divided into splits
- Each server compute a split
- Data is grouped
- Aggregation from groups

Hadoop a brief history

Nutch: The Free Search Alternative to Google

Stefan Krempl *10.06.2004*

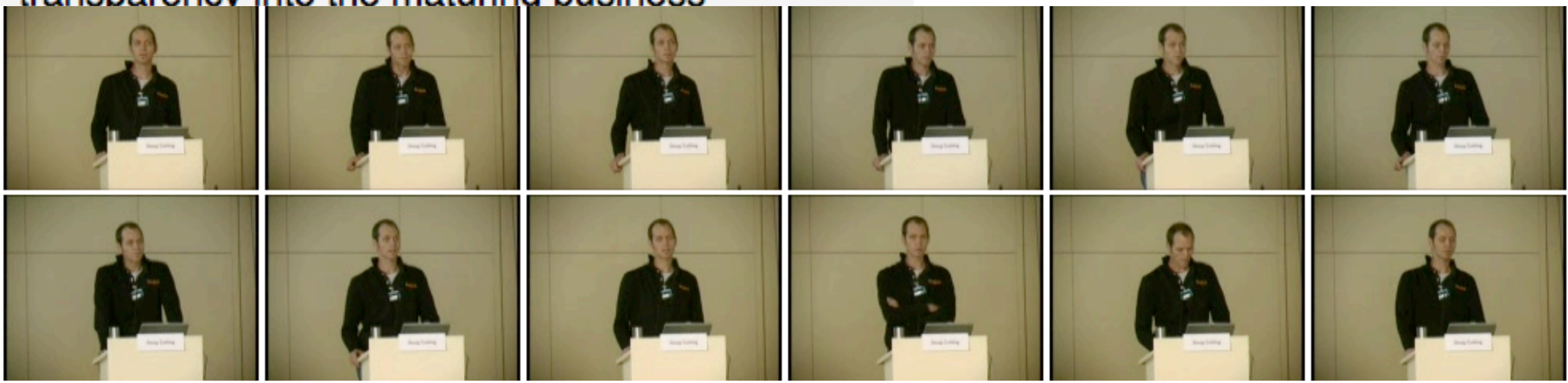
An open deployable search algorithm is set to allow webmasters to launch their own search engines and transparency into the maturing business

03.09.2003 15:09

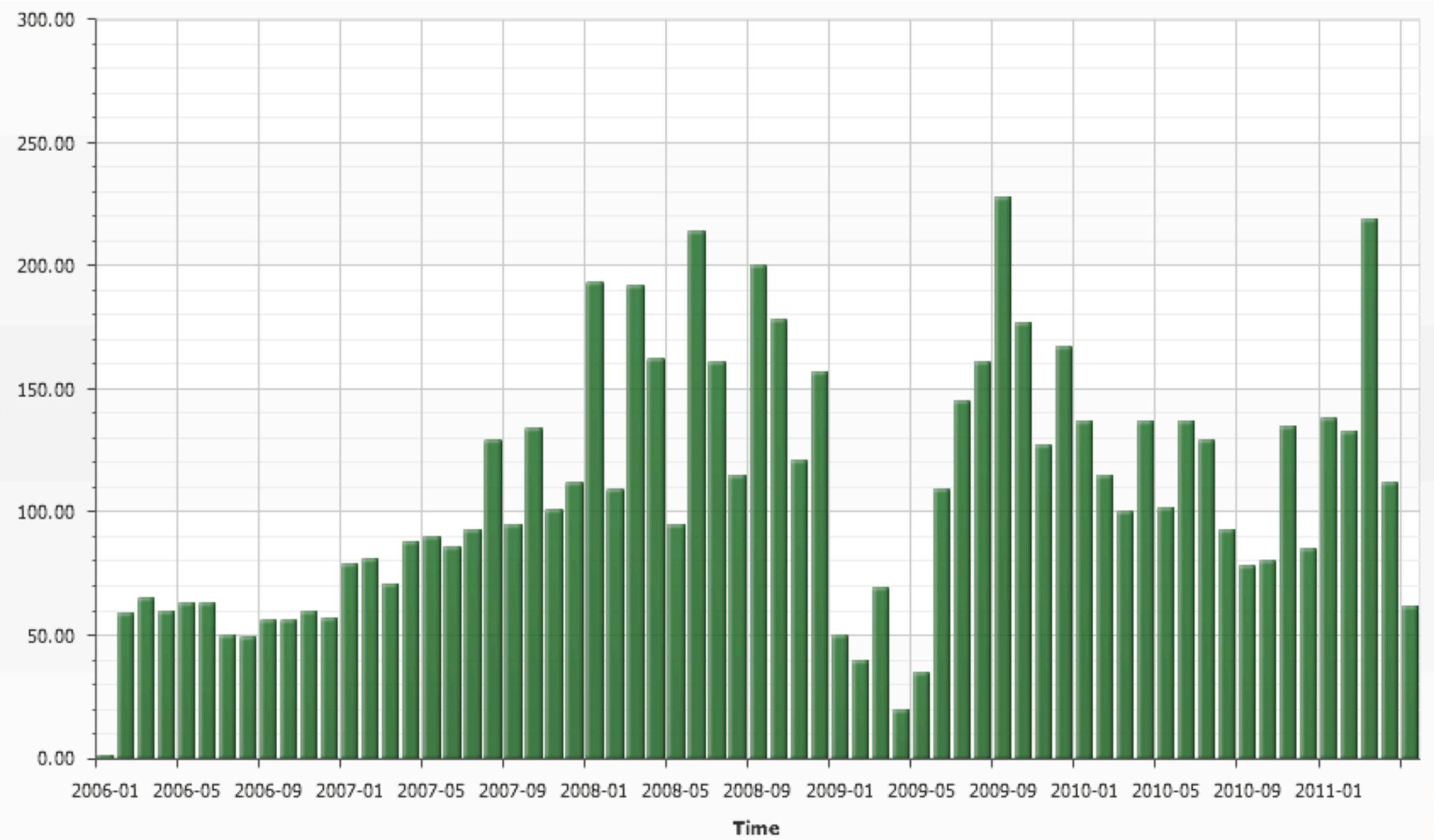
Open-Source-Suchmaschine will in die

vorlesen / MP3-Download

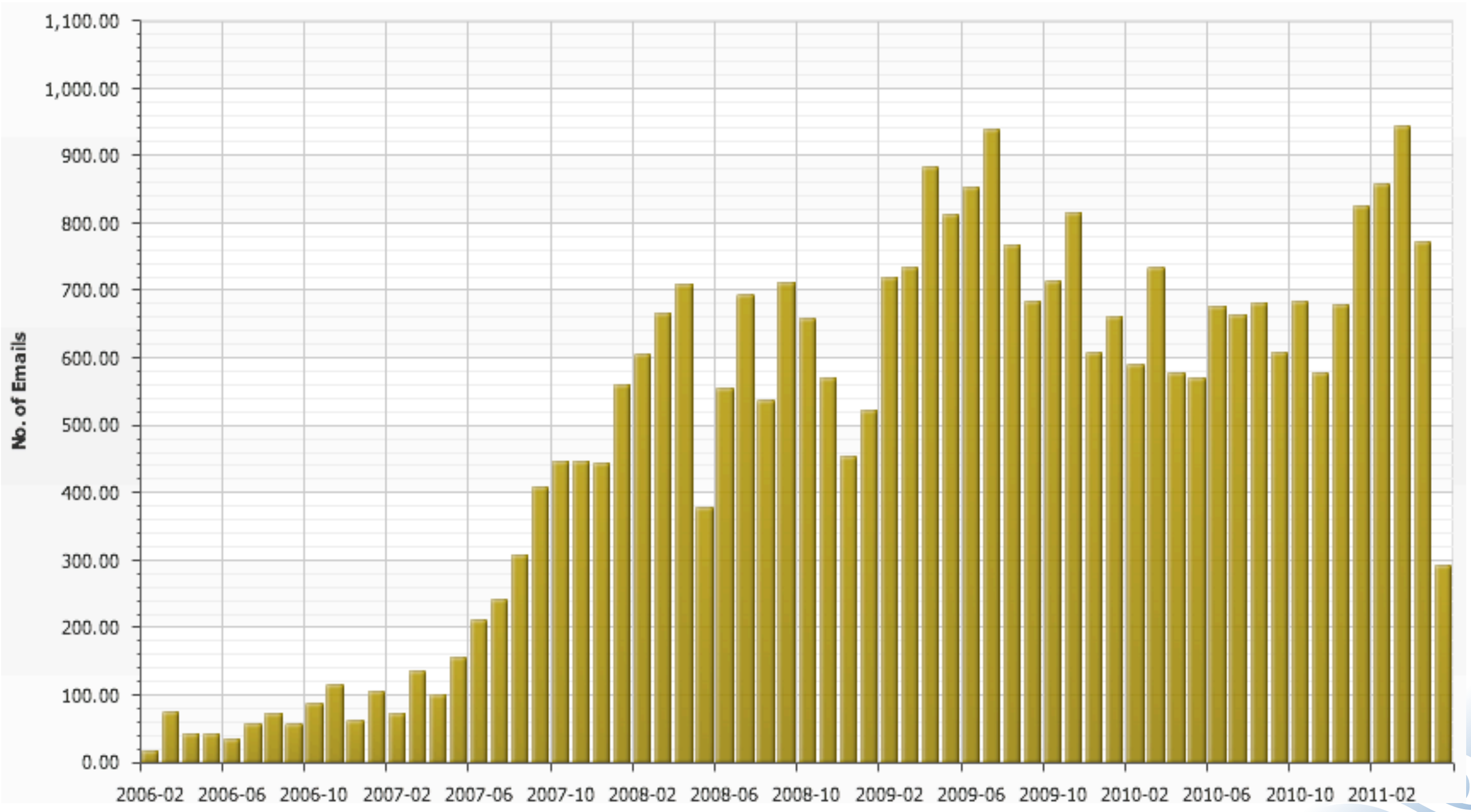
Das Projekt Nutch will eine Alternative zu den bekannten



Commits per Month



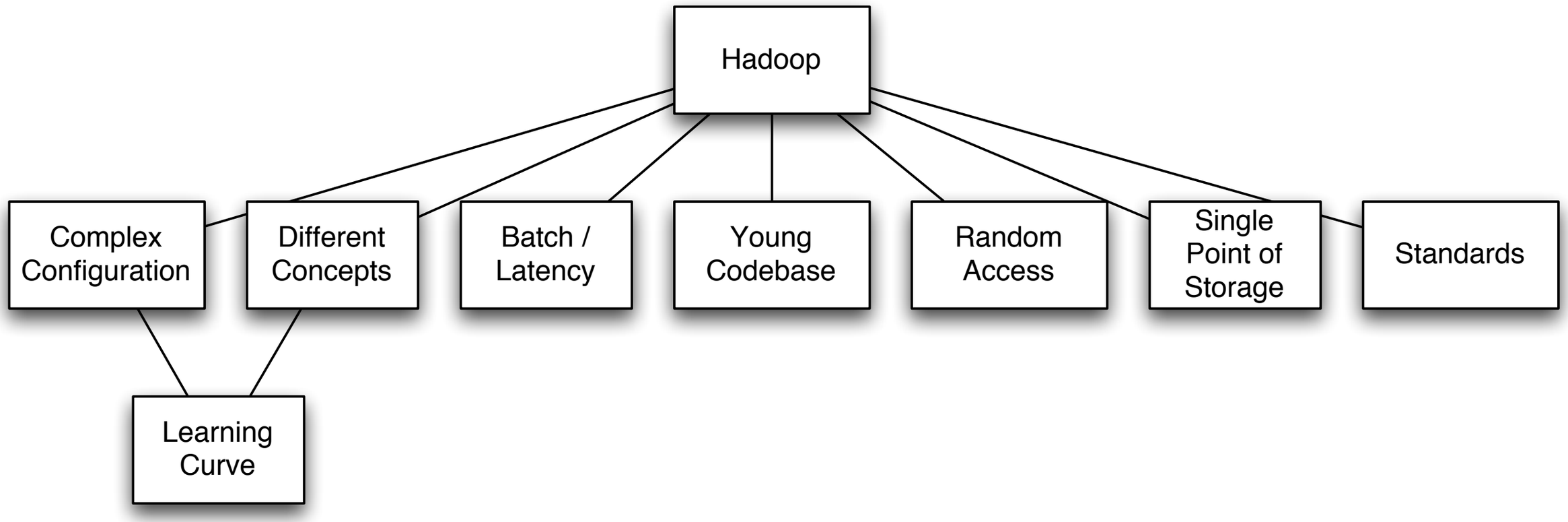
EMails per Month



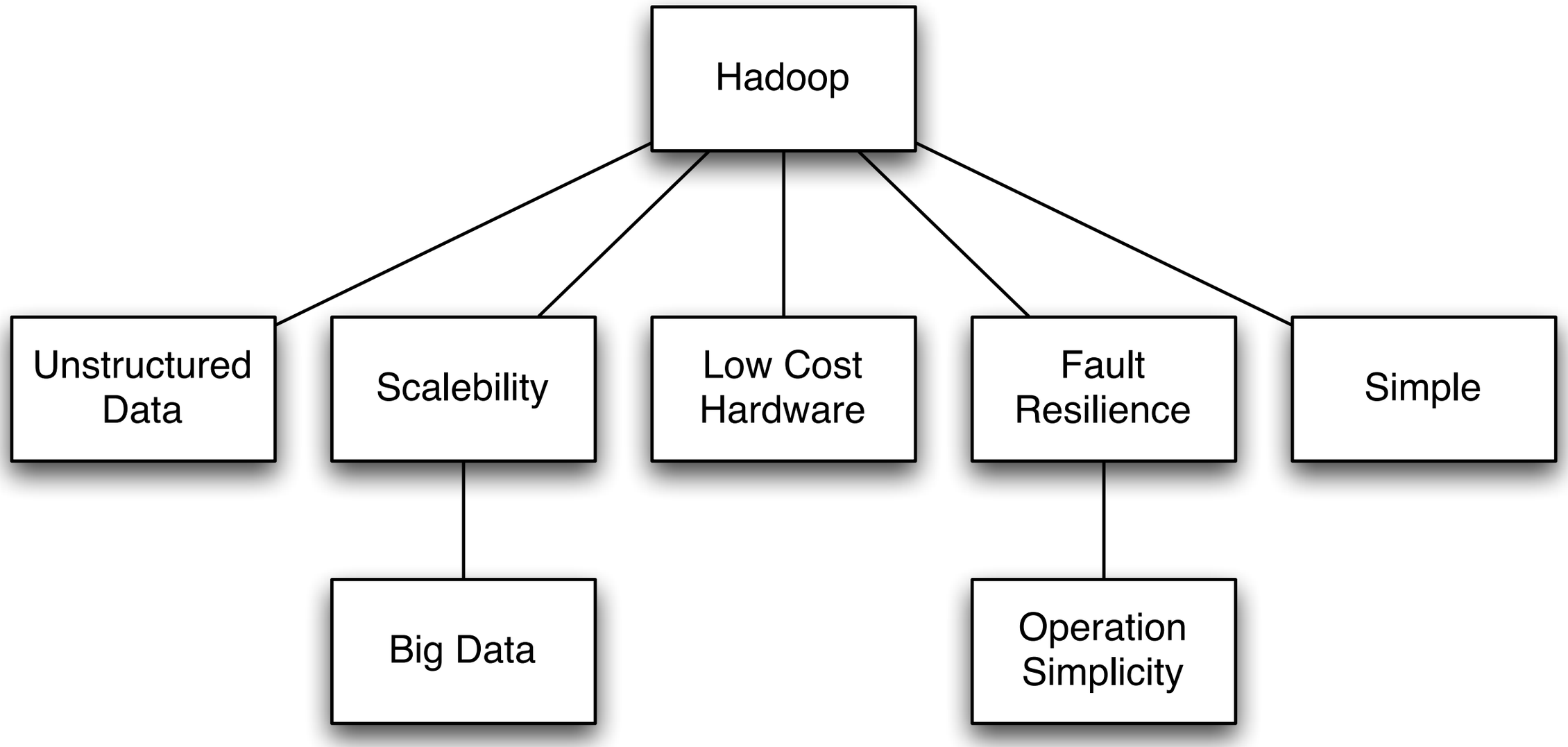
#2

(SWOT)

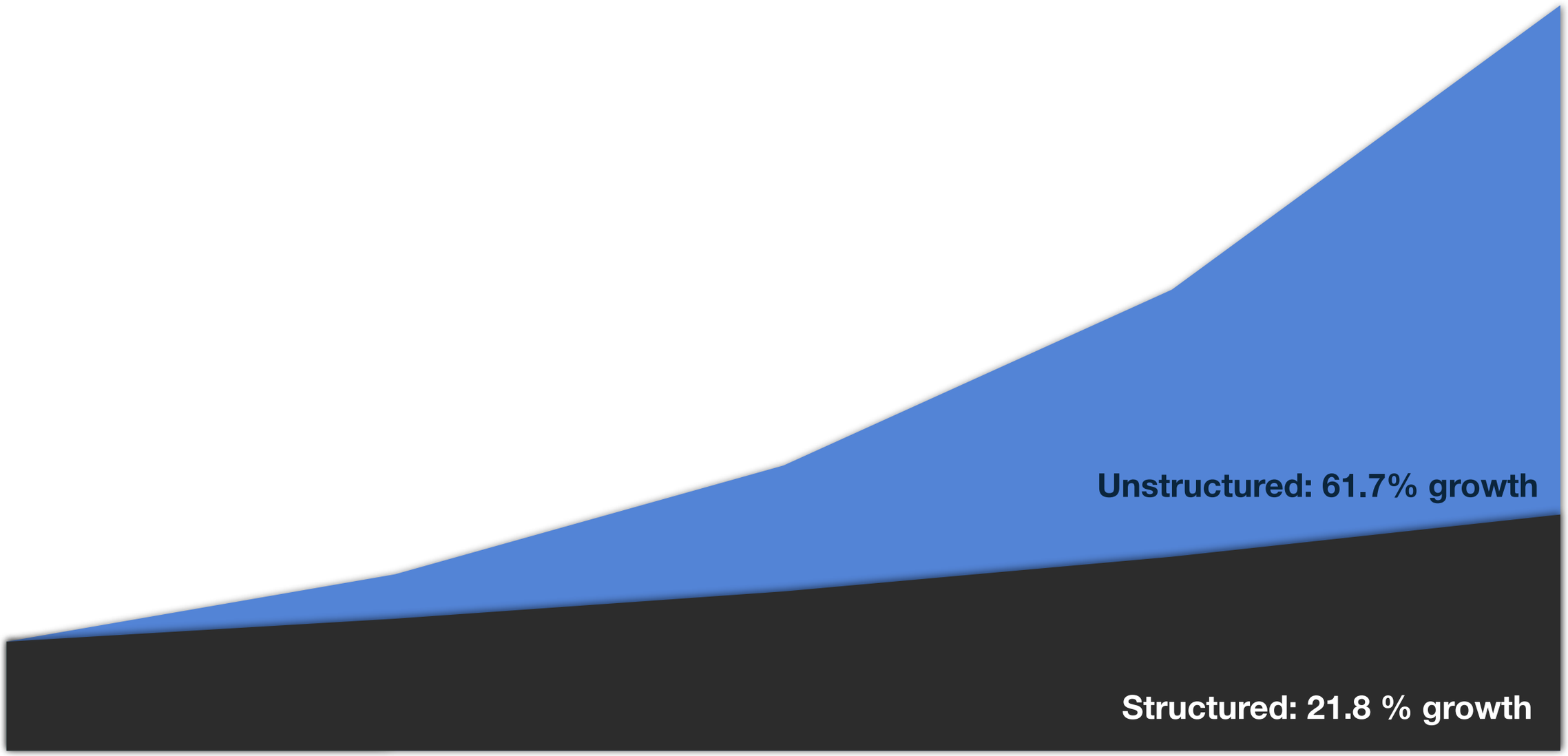
Weaknesses


Hadoop
Complex Configuration
Different Concepts
Batch / Latency
Young Codebase
Random Access
Single Point of Storage
Standards
Learning Curve

Strength


Hadoop
Unstructured Data
Scalebility
Low Cost Hardware
Fault Resilience
Simple
Big Data
Operation Simplicity

Unstructured Data Growth



Enterprise data doubles every three years (Forrester)

What is Hadoop really about?

Slow	Expensive	Expertise
ETL	Data Warehouse	Business Intelligence

Fast	Economic	Self Service
Raw Load	Hadoop	Drag and Drop Spreadsheet

*AD

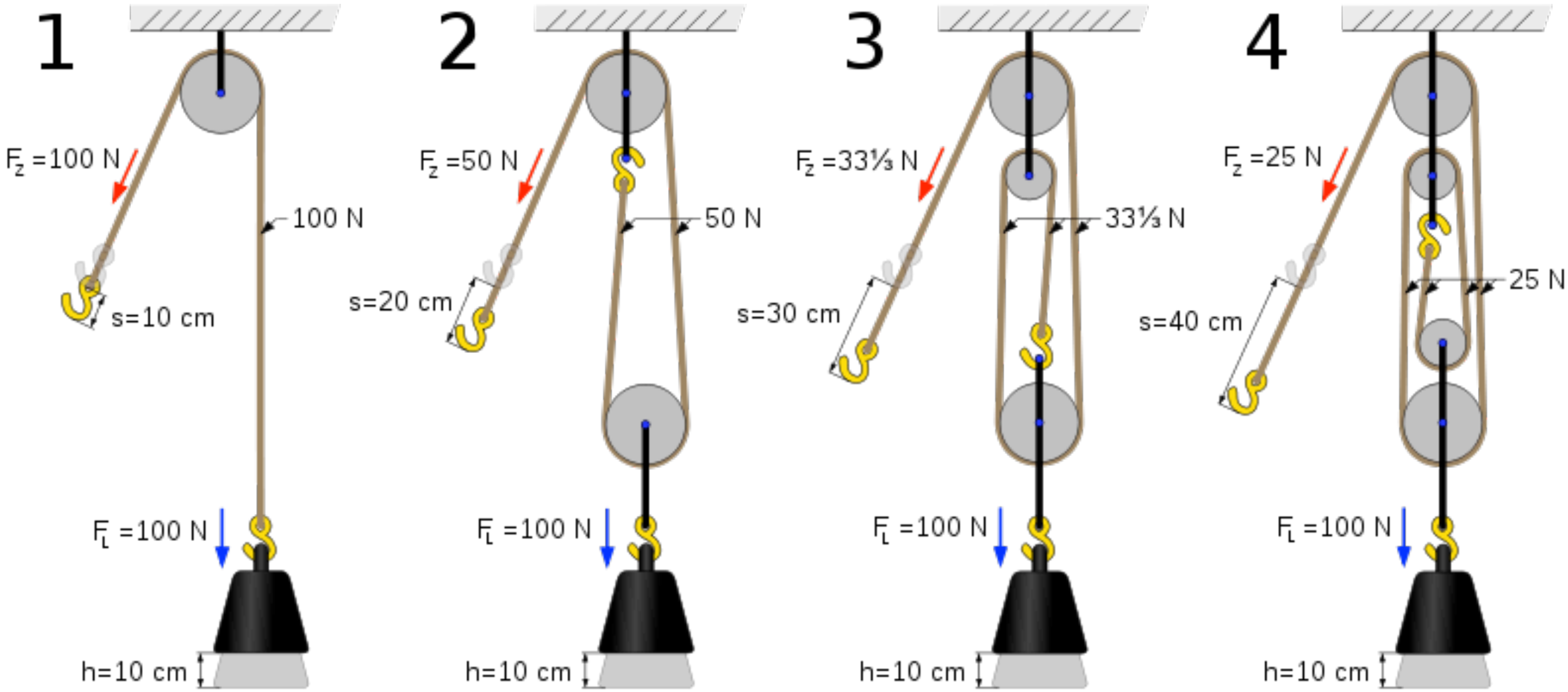
What is Hadoop really about?

We’re hiring!

#3

(Lessions Learned)

Hadoop vs. DB



Hadoop(mergesort)

DB(b-tree/index)

STORE RAW DATA

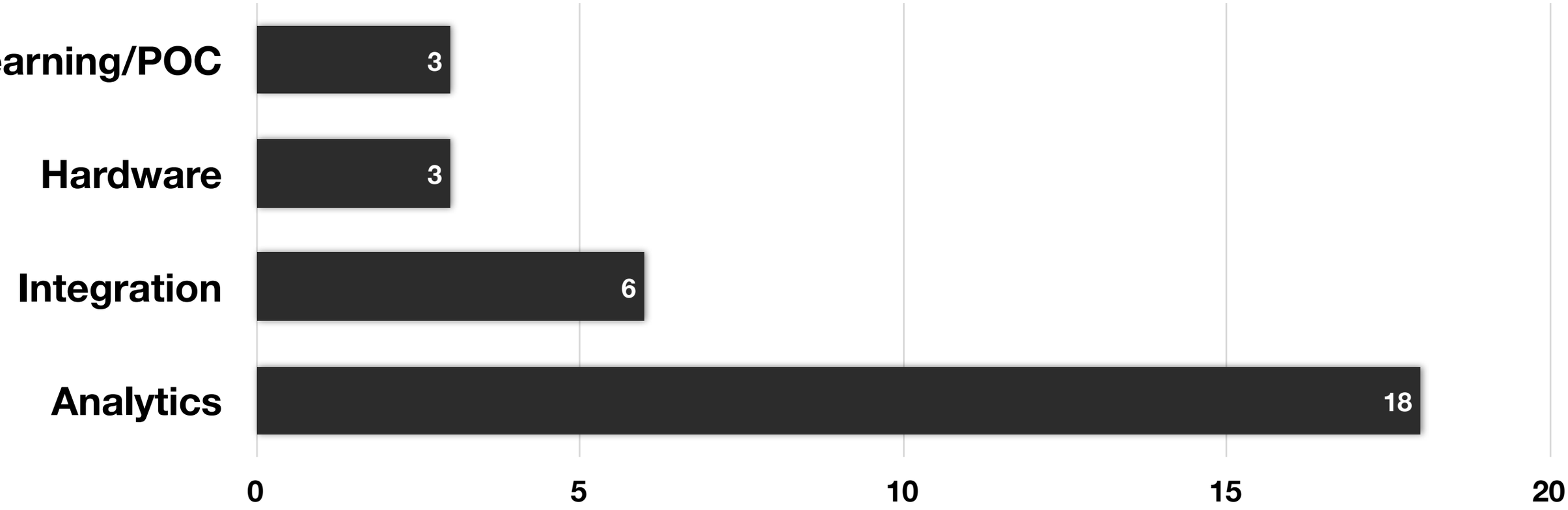
KEEP IT SIMPLE

For Startups

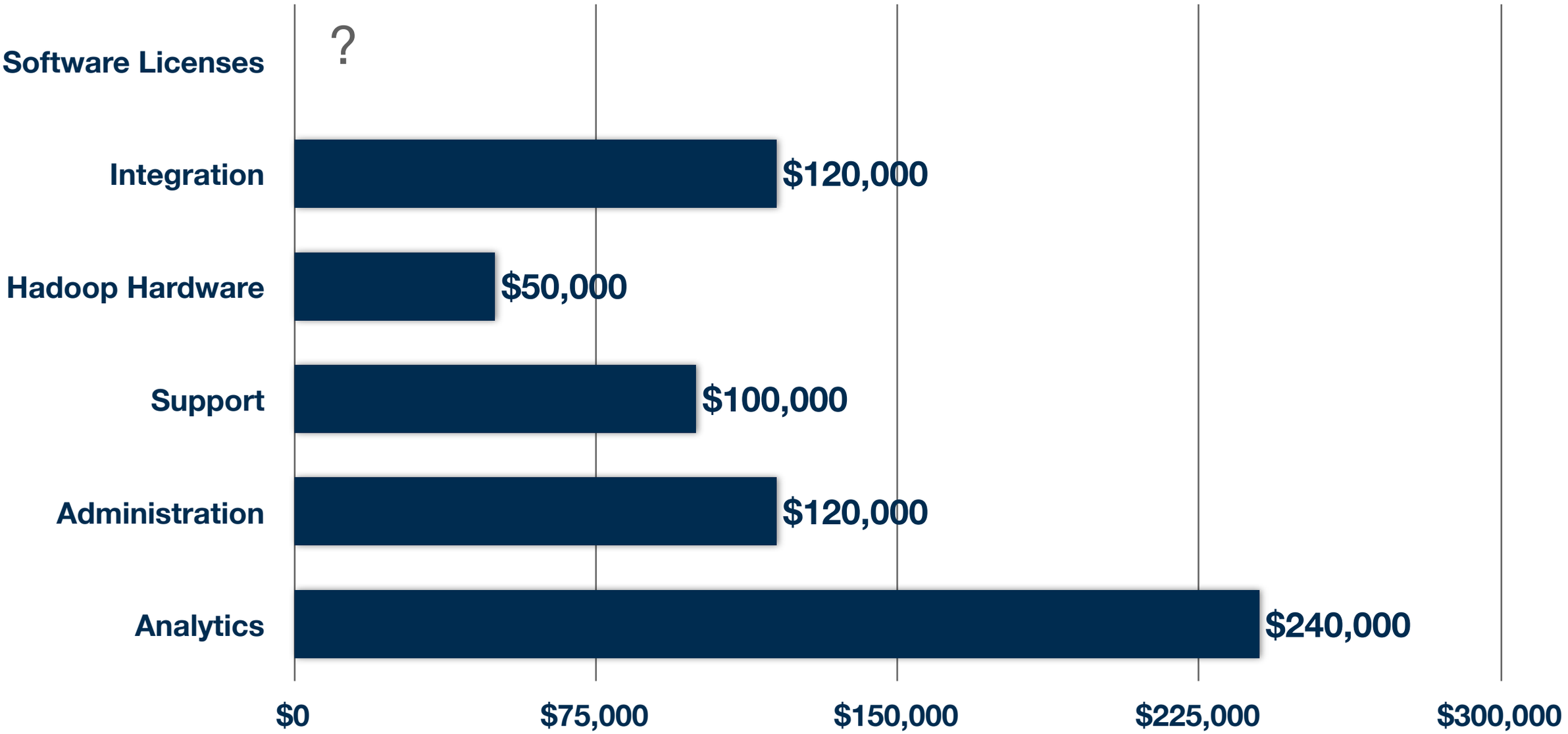
USE SQL

Iterate fast with known technologies

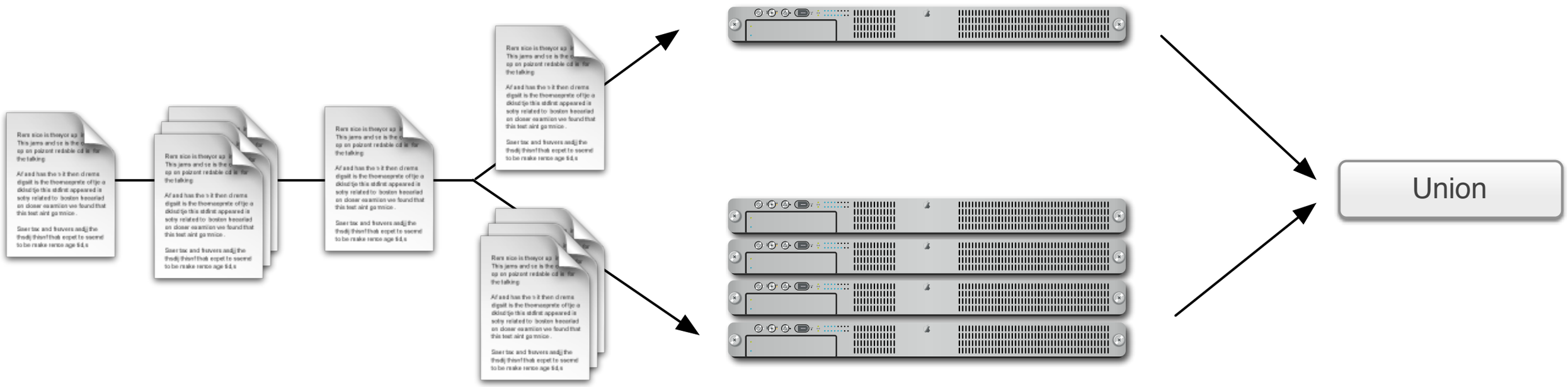
Implementation Effort (month)



Hadoop Cost


Software Licenses
?
Integration
$120,000
Hadoop Hardware
$50,000
Support
$100,000
Administration
$120,000
Analytics
$240,000
$0
$75,000
$150,000
$225,000
$300,000

Small Data / Latency



Pull vs Push

Very slow

Local buffer = risk of lost data

Monitor many agents

Complicated

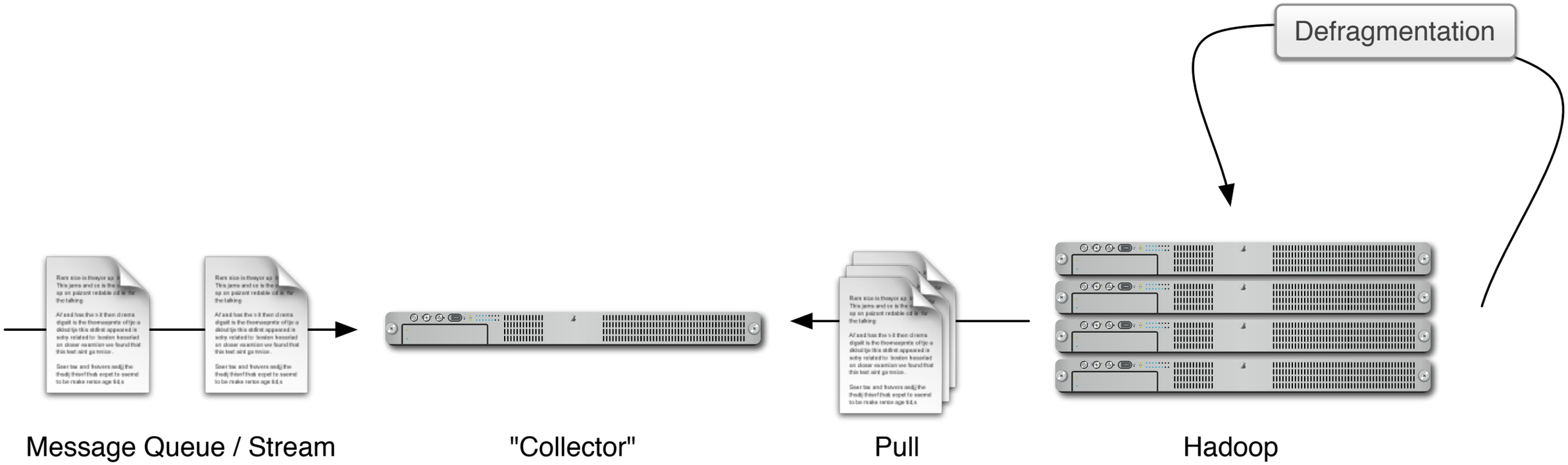
Simple

Pull as often as required

Just one system to monitor

Easier to secure

Working with Streams == push



Just Store Structured Data

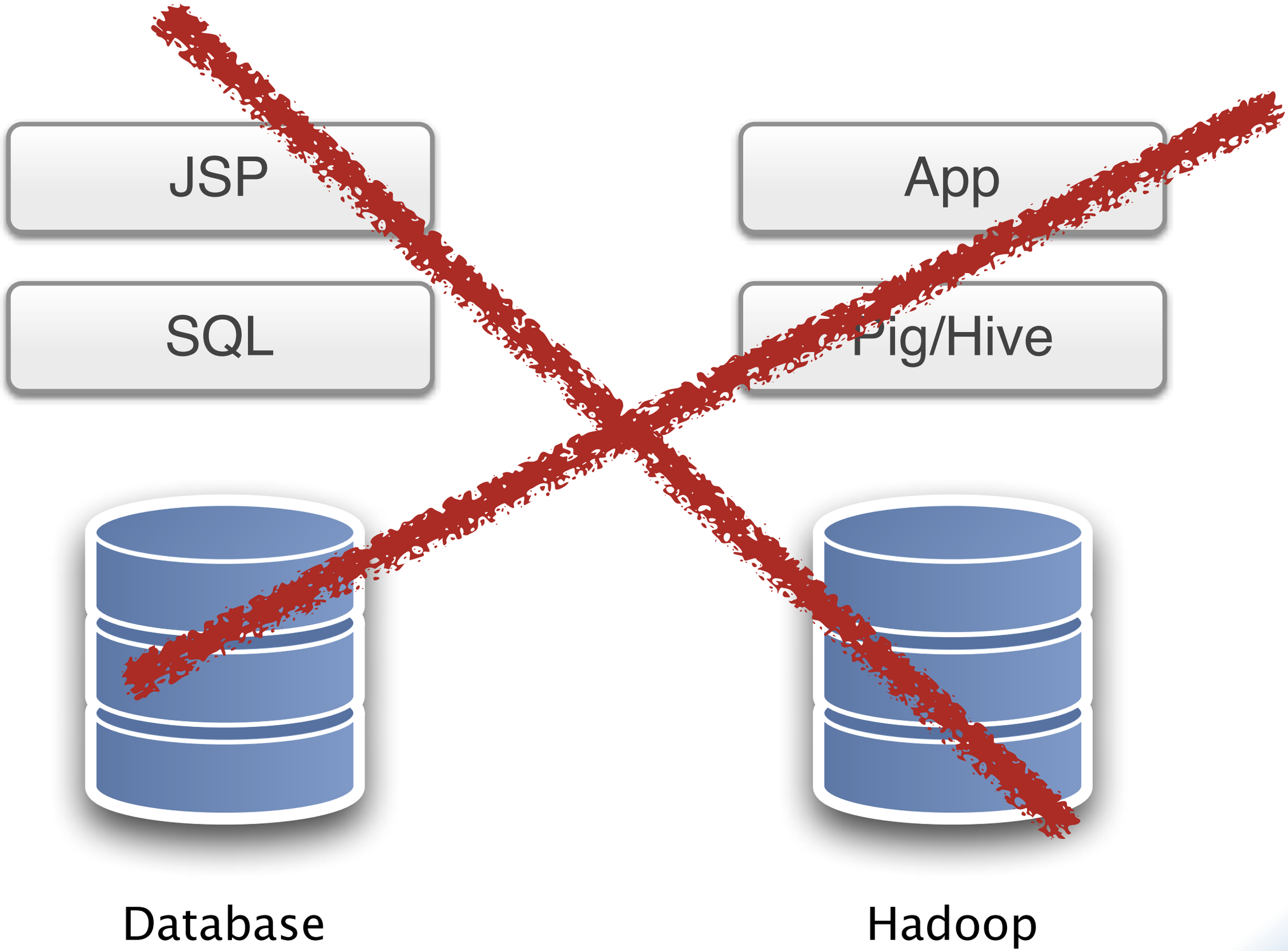
DBVisualizer

Database

Hive/Pig

Hadoop

Application Design



DON'T REMOVE

HOLDING UP THE

SERVERS ABOVE

Application Design

JSP (View)

Spring (Controller)

Hibernate (Model)

Database

Data Vis (View)

Flows (Controller)

Cascading Tuples (Model)

Hadoop

Application Design

Data Vis (View)

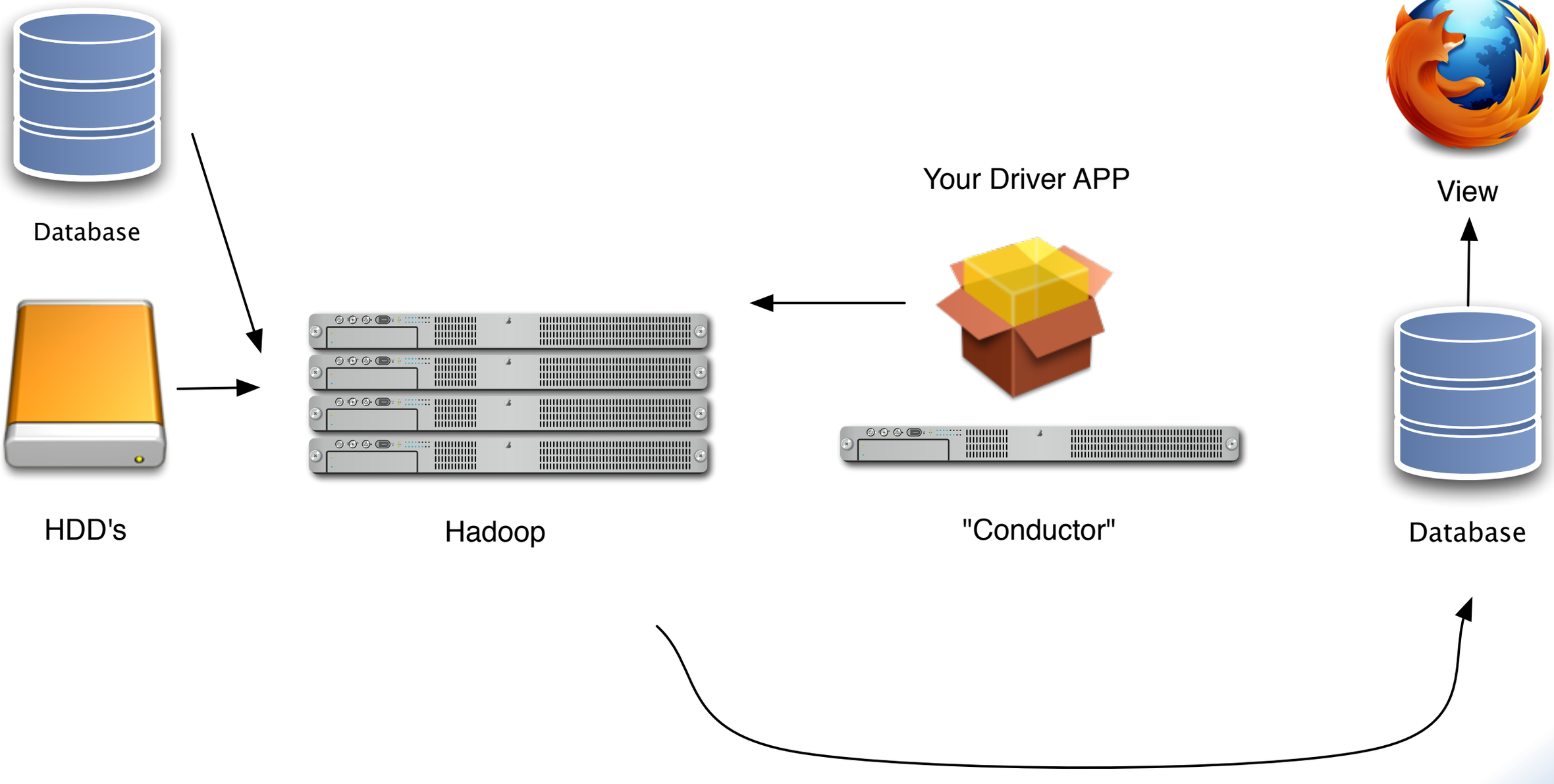
Flows (Controller)

Cascading Tuples (Model)

Hive/Pig

Hudson/App

Application Design



Buy vs Build

Pentenho

Database

Datameer

Hadoop

ORWELL

was an
OPTIMIST

We're hiring!

www.datameer.com

sg AT datameer.com

@datameer