Caching in the Memory Hierarchy:
5 Minutes Ought to Be Enough for Everybody

Anastasia Ailamaki

with Raja Appuswamy, Renata Borovica, Manos Karpathiotakis, Tahir Azim, Matt Olma, Manos Athanassoulis, Yannis Alagiannis, and Goetz Graefe

DATA-INTENSIVE APPLICATIONS AND SYSTEMS

The five-minute rule

Jim Gray and Gianfranco Putzolu, circa 1987:

“Should I keep data item X in memory or on disk?”

Five-minute rule formulation

Break-even Reference Interval (seconds) =

$$\frac{PagesPerMBofRAM}{AccessPerSecondPerDisk}$$

Technology ratio

x

$$\frac{PricePerDiskDrive}{PricePerMBofDRAM}$$

Economic ratio

Five-minute rule formulation

Break-even Reference Interval (seconds) = *(400 secs)*

$$\frac{\text{PagesPerMBofRAM (1024)}}{\text{AccessPerSecondPerDisk (15)}}$$

Technology ratio

x

$$\frac{\text{PricePerDiskDrive (\$30k)}}{\text{PricePerMBofDRAM (\$5k)}}$$

Economic ratio

Popular rule of thumb for engineering data management systems

The five-minute rule

Jim Gray and Gianfranco Putzolu, circa 1987:

"Should I keep data item X in memory or on disk?"

Answer, circa 1987:

"Pages referenced every 5 minutes should be memory resident"

Answer, circa 2018: ???

The five-minute rule, 30 years later

[ADMS2017]

What has changed?

- Disk, RAM price ratio
- (Way) deeper storage hierarchy
- Different data formats -> Different access costs

Update I: RAM became CHEAP

New Disk, DRAM price ratio

Parameter	Disk (then)	Disk (now)	DRAM (then)	DRAM (now)
Unit cost ($)	$30,000	$49	$5,000	$80
Unit capacity	180MB	2TB	1MB	16GB
Random IO/s	15	200	-	-

- Capacity:↑10,000×, Cost: ↓1,000×, HDD Performance:↑10×

New Disk, DRAM price ratio

Parameter	Disk (then)	Disk (now)	DRAM (then)	DRAM (now)
Unit cost ($)	$30,000	$49	$5,000	$80
Unit capacity	180MB	2TB	1MB	16GB
Random IO/s	15	200	-	-

- Capacity:↑10,000×, Cost: ↓1,000×, HDD Performance:↑10×

Page size (4KB)	Then	Now
RAM-HDD	5 mins	**5 hours**

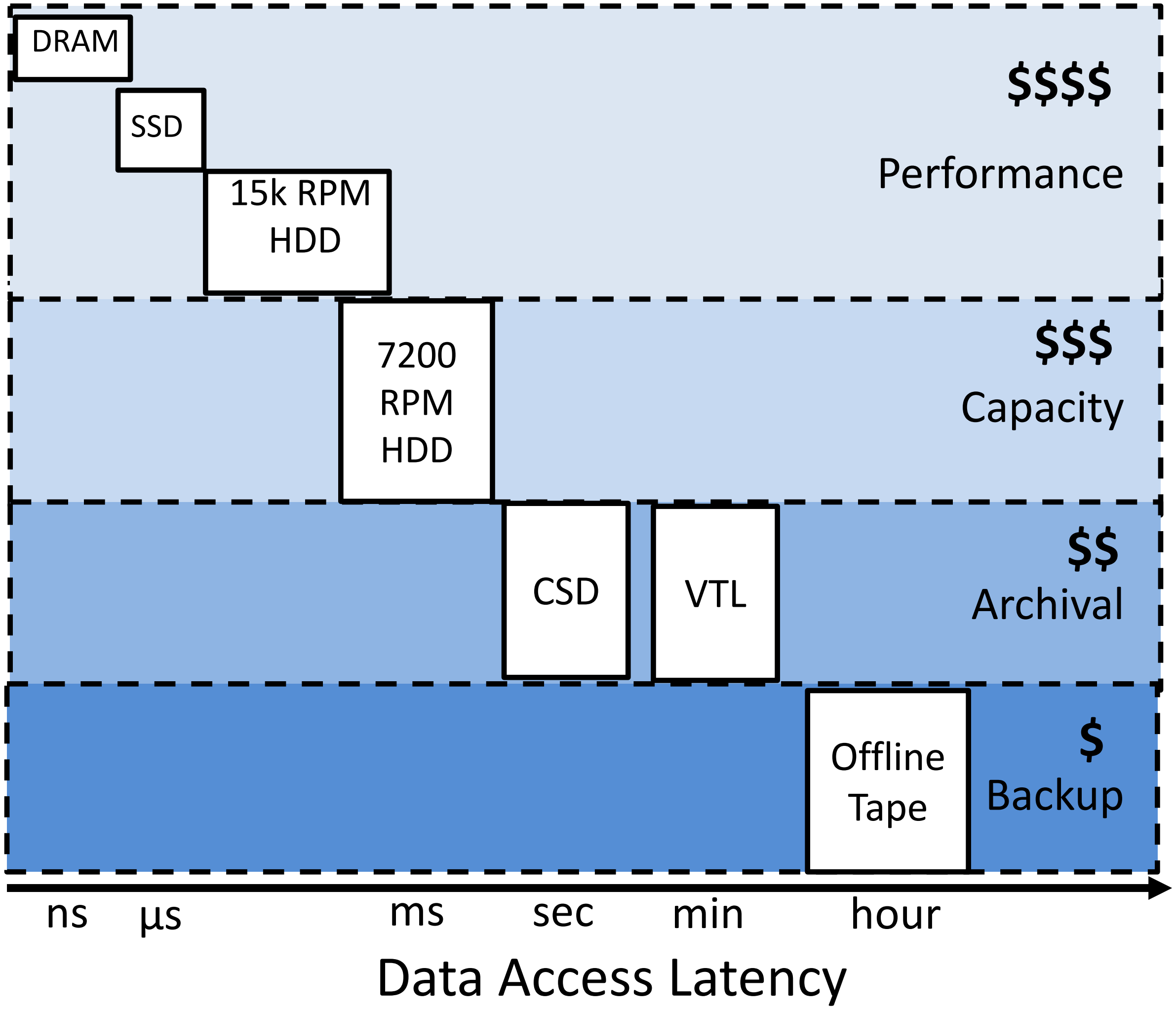
- RAM-HDD break-even 60× higher due to fall in DRAM price

Updated rule: Store only extremely "cold" data in HDD

Update II: Hierarchy became CHEAP

Modern (deep) storage hierarchy

[VLDB2016]



Multitier hierarchy with price and performance matching workload requirements

The performance tier

Five-minute rule with SATA SSD

Parameter	Disk (now)	DRAM (now)	SATA SSD (now)
Unit cost ($)	$49	$80	560
Unit capacity	2TB	16GB	800GB
Cost/MB	0.00002	0.005	0.0007
Random IO/s	200	-	67k/20k

- Two properties of SSDs
  - Middleground between DRAM and HDD w.r.t cost/MB
  - 100-1000× higher random IOPS than HDD
- Two new rules with SSDs
  - DRAM-SSD rule: SSD as a primary store
  - SSD-HDD rule: SSD as a cache

Break-even interval for SATA SSD

Parameter	Disk (now)	DRAM (now)	SATA SSD (now)
Unit cost ($)	$49	$80	560
Unit capacity	2TB	16GB	800GB
Cost/MB	0.00002	0.005	0.0007
Random IO/s	200	-	67k (r)/20k (w)

Page size (4KB)	2007	Now
RAM-HDD	1.5h	5 hours
RAM-SSD	15m	7 m (r)/24m (w)

5-minute rule now ~applicable to SATA SSD

Break-even interval for SATA SSD

Parameter	Disk (now)	DRAM (now)	SATA SSD (now)
Unit cost ($)	$49	$80	560
Unit capacity	2TB	16GB	800GB
Cost/MB	0.00002	0.005	0.0007
Random IO/s	200	-	67k (r)/20k (w)

Page size (4KB)	2007	Now
RAM-HDD	1.5h	5 hours
RAM-SSD	15m	7 m (r)/24m (w)
SSD-HDD	2.25h	1 day

5-minute rule now ~applicable to SATA SSD

With 1 day interval, all active data will be in RAM/SSD

Trends in performance tier

- SSDs inching closer to the CPU
  - SATA -> SAS/FiberChannel -> PCIe -> NVMe -> DIMM
  - NVMe PCIe SSDs are server accelerators of choice

Device	Capacity	Price ($)	IOPS (k) r/w	B/W (GBps)
SATA SSD	800GB	560	67/20	0.5/0.46
Intel 750	1TB	630	460/290	2.5/1.2

Trends in performance tier

- SSDs inching closer to the CPU
  - SATA -> SAS/FiberChannel -> PCIe -> NVMe -> DIMM
  - NVMe PCIe SSDs are server accelerators of choice
- Storage Class Memory devices (ex: 3D Xpoint)
  - Faster than Flash, Denser than DRAM, and non-volatile
  - Standardized, byte-addressable, NVDIMM-P soon

Device	Capacity	Price ($)	IOPS (k) r/w	B/W (GBps)
SATA SSD	800GB	560	67/20	0.5/0.46
Intel 750	1TB	630	460/290	2.5/1.2
Intel P4800X	384GB	1520	550/500	2.5/2

Break even interval for PCIe SSD/NVM

Device	Capacity	Price ($)	IOPS (k) r/w	B/W (GBps)
SATA SSD	800GB	560	67/20	0.5/0.46
Intel 750	1TB	630	460/290	2.5/1.2
Intel P4800X	384GB	1520	550/500	2.5/2

Page size (4KB)	Now
RAM-SATA SSD	7 m (r) / 24m (w)
RAM-Intel 750	41 s (r) / 1m (w)
RAM-P4800X	47 s (r) / 52s (w)

DRAM-NVM break-even interval is shrinking

Interval disparity between reads and writes is shrinking

Break even interval for PCIe SSD/NVM

Device	Capacity	Price ($)	IOPS (k) r/w	B/W (GBps)
SATA SSD	800GB	560	67/20	0.5/0.46
Intel 750	1TB	630	460/290	2.5/1.2
Intel P4800X	384GB	1520	550/500	2.5/2

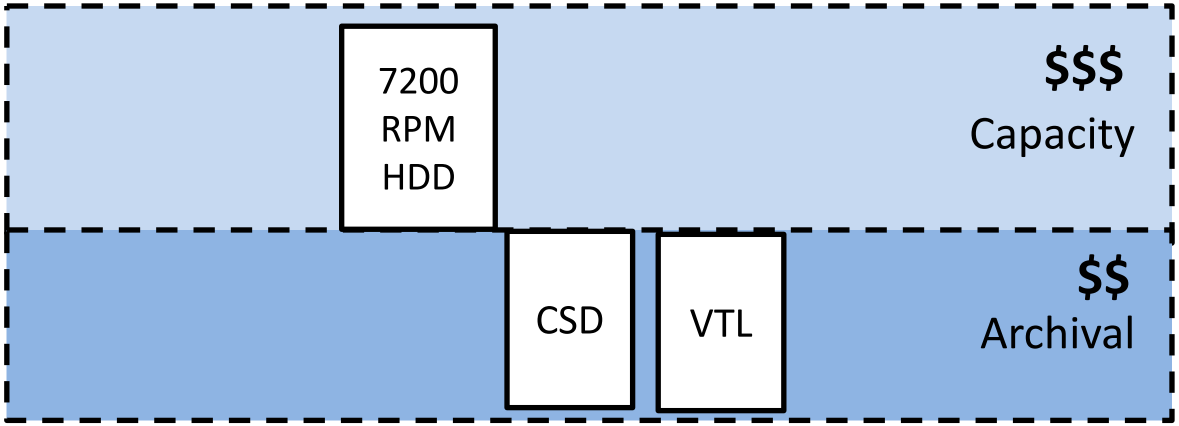
Page size (4KB)	Now
RAM-SATA SSD	7 m (r) / 24m (w)
RAM-Intel 750	41 s (r) / 1m (w)
RAM-P4800X	47 s (r) / 52s (w)

DRAM-NVM break-even interval is shrinking

Interval disparity between reads and writes is shrinking

Impending shift from DRAM to NVM-based data management engines

(Extending) the capacity tier



Trends in high-density storage

- HDD scaling falls behind Kryder's rate
  - PMR provides 16% improvement in areal density, not 40%

Trends in high-density storage

- HDD scaling falls behind Kryder's rate
  - PMR provides 16% improvement in areal density, not 40%
- Tape density continues 33% growth rate
  - IBM's new record: 201 Billion bits/sq. inch
  - But high access latency

Trends in high-density storage

- HDD scaling falls behind Kryder's rate
  - PMR provides 16% improvement in areal density, not 40%
- Tape density continues 33% growth rate
  - IBM's new record: 201 Billion bits/sq. inch
  - But high access latency
- Flash density outpacing rest
  - 40% density growth due to volumetric + areal techniques
  - But high cost/GB

Trends in high-density storage

- HDD scaling falls behind Kryder’s rate
  - PMR provides 16% improvement in areal density, not 40%
- Tape density continues 33% growth rate
  - IBM’s new record: 201 Billion bits/sq. inch
  - But high access latency
- Flash density outpacing rest
  - 40% density growth due to volumetric + areal techniques
  - But high cost/GB
- Cold storage devices (CSD) filling the gap
  - 1,000 high-density SMR disks in MAID setup
  - PB density, 10s latency, 2-10GB/s bandwidth

Break-even interval for tape

Metric	DRAM	HDD	SpectraLogic T50e tape library
Unit capacity	16GB	2TB	10 * 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100GB/s	200MB/s	4 * 750 MB/s

- DRAM-tape break-even interval: 300 years!

“Tape: The motel where data checks in and never checks out”

- Jim Gray

- Kaps is not the right metric for tape
  - Maps, TB-scan better

Alternate comparison metrics

Metric	DRAM	HDD	SpectraLogic T50e tape library
Unit capacity	16GB	2TB	10 * 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100GB/s	200MB/s	4 * 750 MB/s
$/Kaps (amortized)	9e-14	5e-9	8e-3
$/TBScan (amortized)	8e-6	3e-3	3e-2

HDD 1,000,000× cheaper w.r.t Kaps, only 10× w.r.t TBScan

HDD—tape gap shrinking for sequential workloads

Implications for the capacity tier

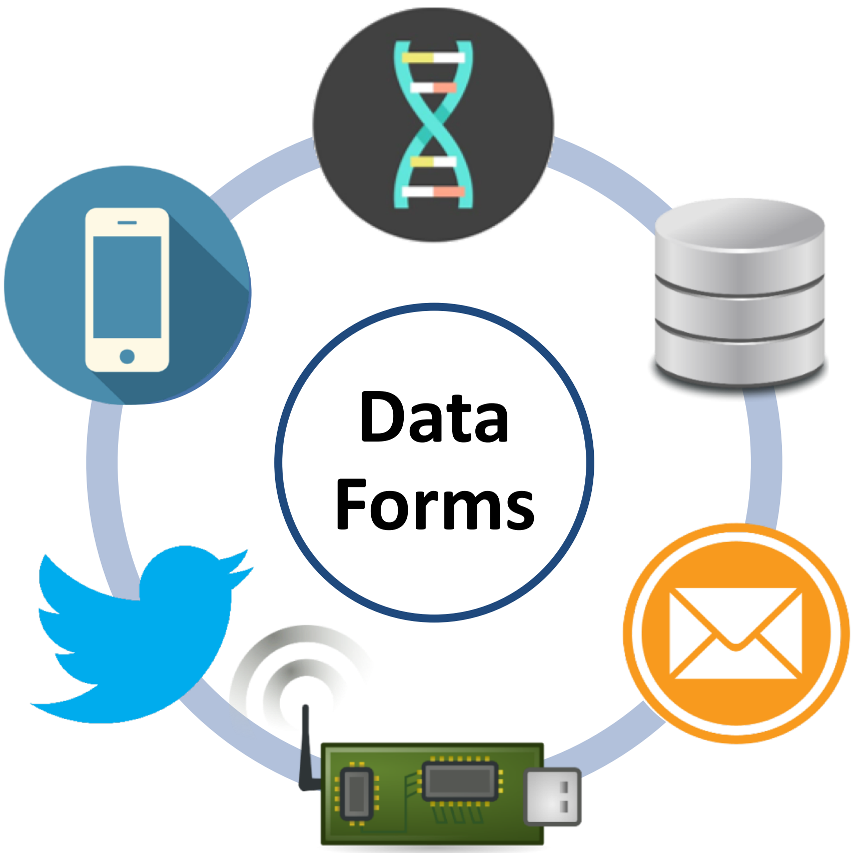
- Traditional tiering hierarchy
  - HDD based capacity tier. Tape, CSD only used in archival.
- Clear division in workloads
  - Only non-latency sensitive, batch analytics in capacity tier
- Is it economical to merge the two tiers?
  - "40% cost savings by using a cold storage tier" [Skipper, VLDB'16]
- Can batch analytics be done on tape/CSD?
  - Query Execution in Tertiary Memory Databases [VLDB'96]
  - Skipper: Cheap data analytics over cold storage devices [VLDB'16]
  - Nakshatra: Running batch analytics on an archive [MASCOTS'14]

Time to revisit traditional capacity—archival division of labor

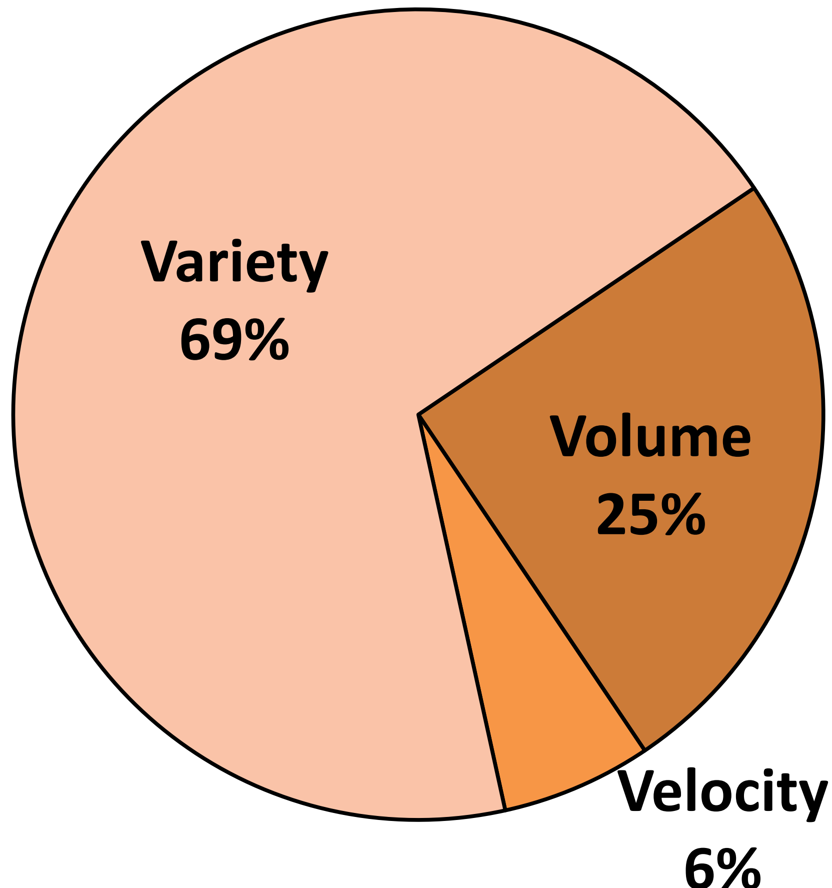
Update III:
Data became HETEROGENEOUS

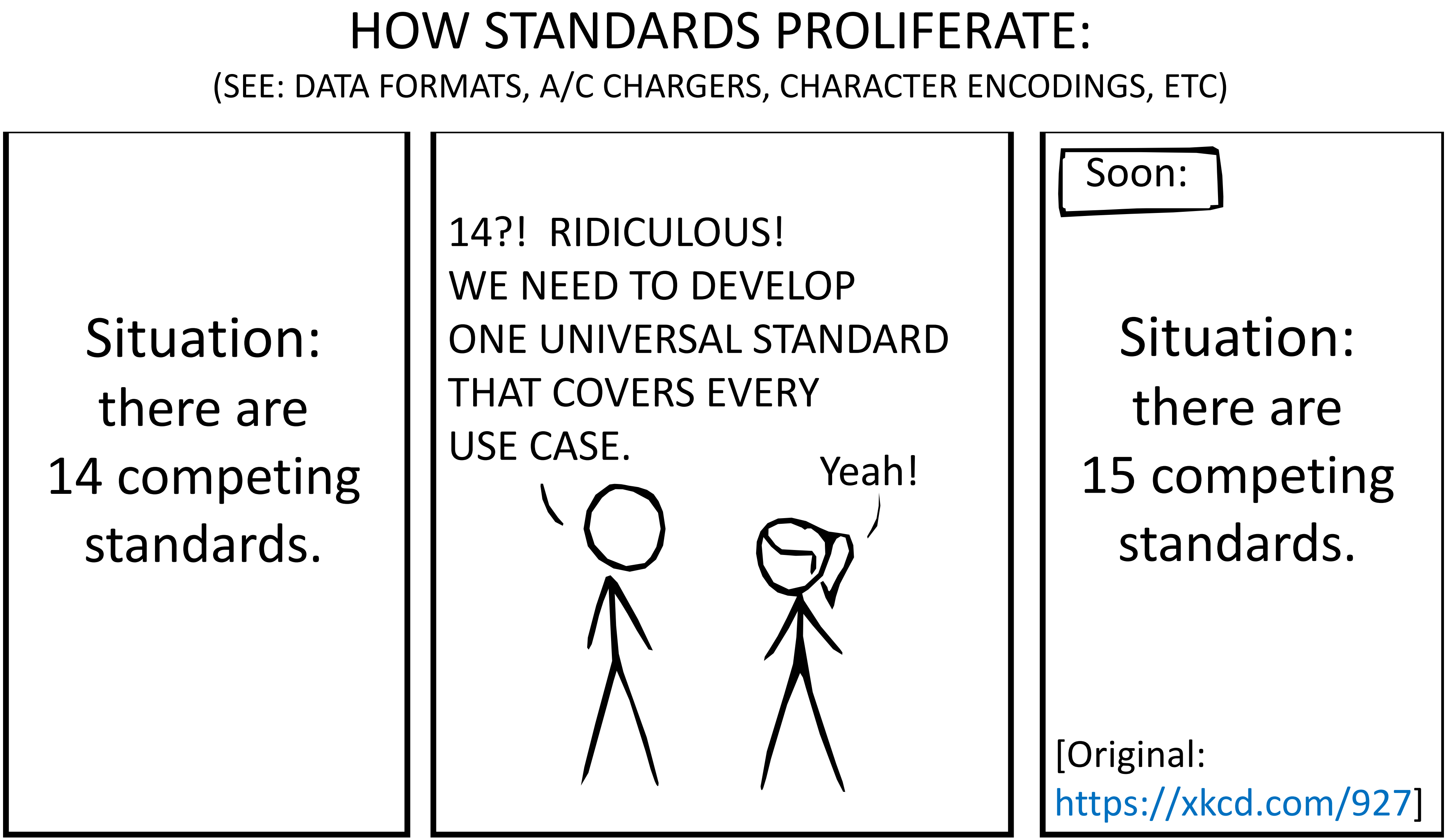
Data heterogeneity introduces challenges

71% of data scientists:
Analysis more difficult due to variety, not volume [Paradigm4]



Variety, Volume, Velocity Importance [NVP Survey]





No “one data format to rule them all”

Looking under the carpet: Loading and tuning are expensive

Instant access to data → **Avoid data loading (In situ querying)**

Interactive response time → **Building indexes is expensive!**

Five-minute rule assumes ready-to-go data

Reducing amount of (raw) data accessed

- –**Partition data to a favorable state**
- –**Build appropriate indexes and caches**

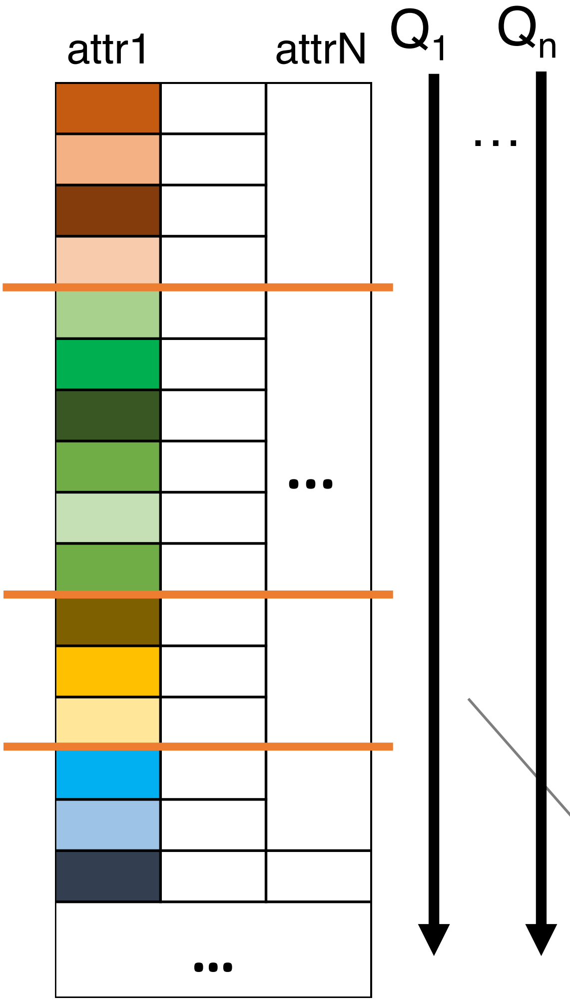
What to invest in?

- –**Evict based on cost of re-caching**

What to evict?

Logical partitioning

[VLDB2017]



Enable data skipping

Fine-grained access path selection

Capture implicit clustering

Iteratively partition dataset

Homogeneous

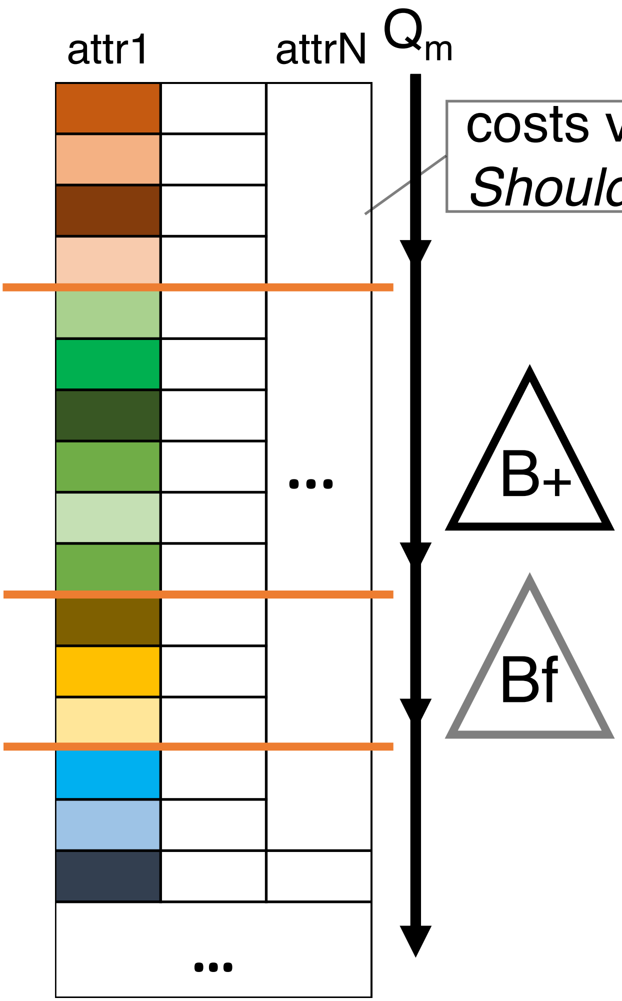
Query-based

1) Collect data statistics at runtime
2) Calculate number of sub-partitions

Increase disjointness: Reduce distinct values
Remove tails: Reduce excess kurtosis

Set the "ground" for reducing data access

Online index tuning



Index tuning on partition level

Choose what & when to build

What

- Value-Existence (i.e., Bloom filters)
- Value-Position (i.e., B+ Trees)

When

- Based on randomized algorithm
- Cost of scan vs. cost of build + gain

Build and drop based on budget

Maximize gain: build cost vs performance

Evicting heterogeneous data

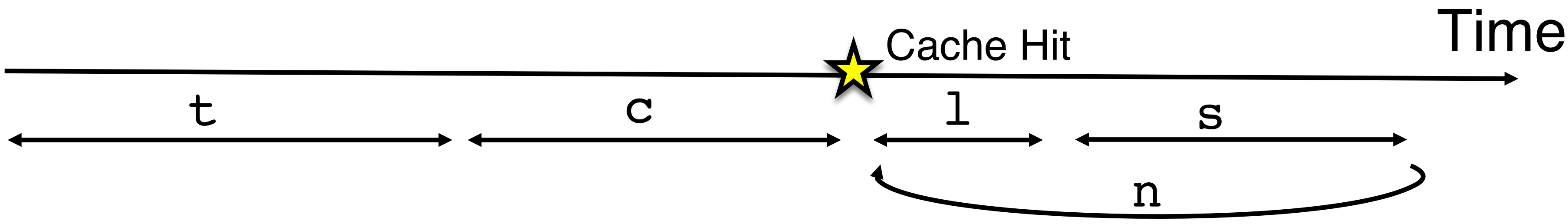
Extreme 1:
(LRU assumes) all cached items have equal weight

Extreme 2:
weight(XML) >> weight(JSON) >> weight(CSV) >> …

cached representation != raw representation

must account for widely varying weights

Benefit metric for het. datasets



- Cost of operator execution: **t**
- Cost of "materialization": **c**
- Cost of finding a match: **l**
- Cost of scanning the cache: **s**
- Number of times operator invoked: **n**
- Cache size: **B**

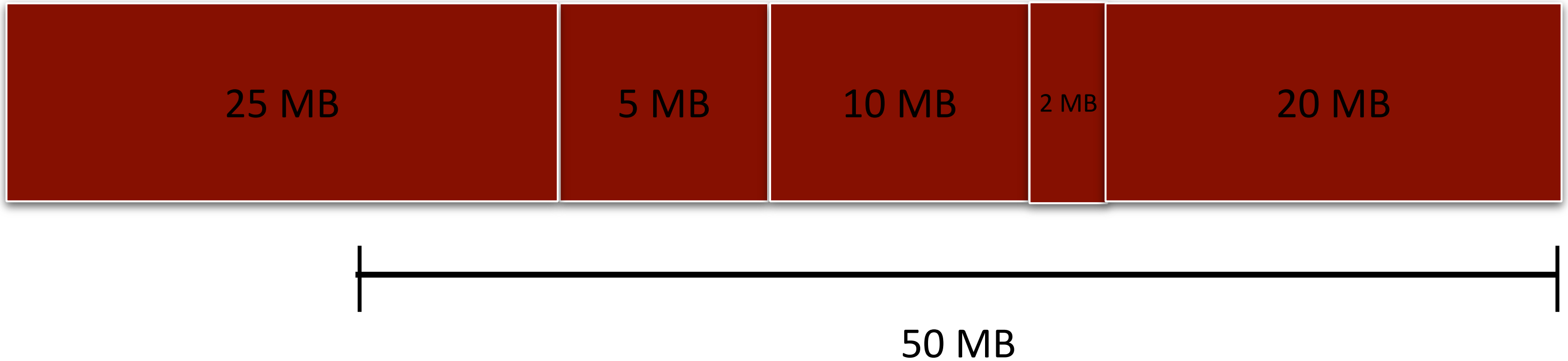
Materialization cost depends on data type & format

Metric: `(n*(t+c-s-l))/log(B)`

(ReCache) eviction policy: 1st try

[VLDB2018]

Items to Evict Chosen by Unmodified Greedy Dual

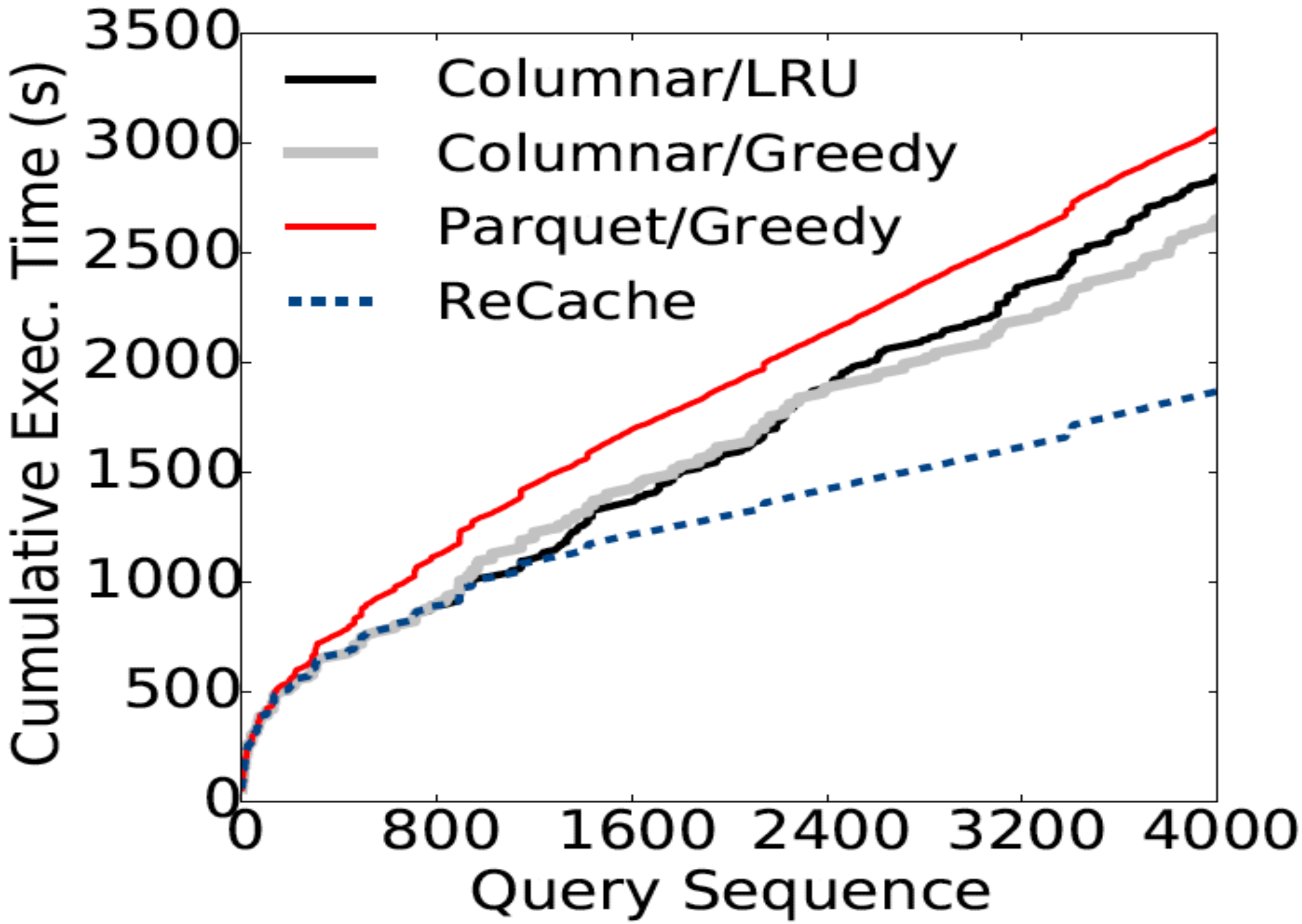


Unnecessary removals!

(ReCache) eviction policy

Sort candidates by size -> Minimize # removals

Queries on CSV+JSON Symantec Data



ReCache is 40% faster than Parquet, 34% than relational columnar, plus another 8% due to cache eviction policy

The five-minute rule, 30 years later

- Growing DRAM-HDD & shrinking DRAM-NVM intervals

Most performance critical data will sit in SSD/NVM

- Rapid improvements in SSD/NVM density

All randomly accessed data can sit in SSD/NVM

- Shrinking HDD—tape/CSD difference w.r.t $/TBscan

Can merge archival+capacity tier into cold storage tier

Sequential batch analytics can be hosted in new tier

- Growing data heterogeneity -> Non-uniform access costs

Need techniques to i) separate "hot–cold data", and ii) decide on eviction based on "re-cache cost"