Cassandra 1.0 and Beyond

Jake Luciani, DataStax
jake@datastax.com, 11/11/11

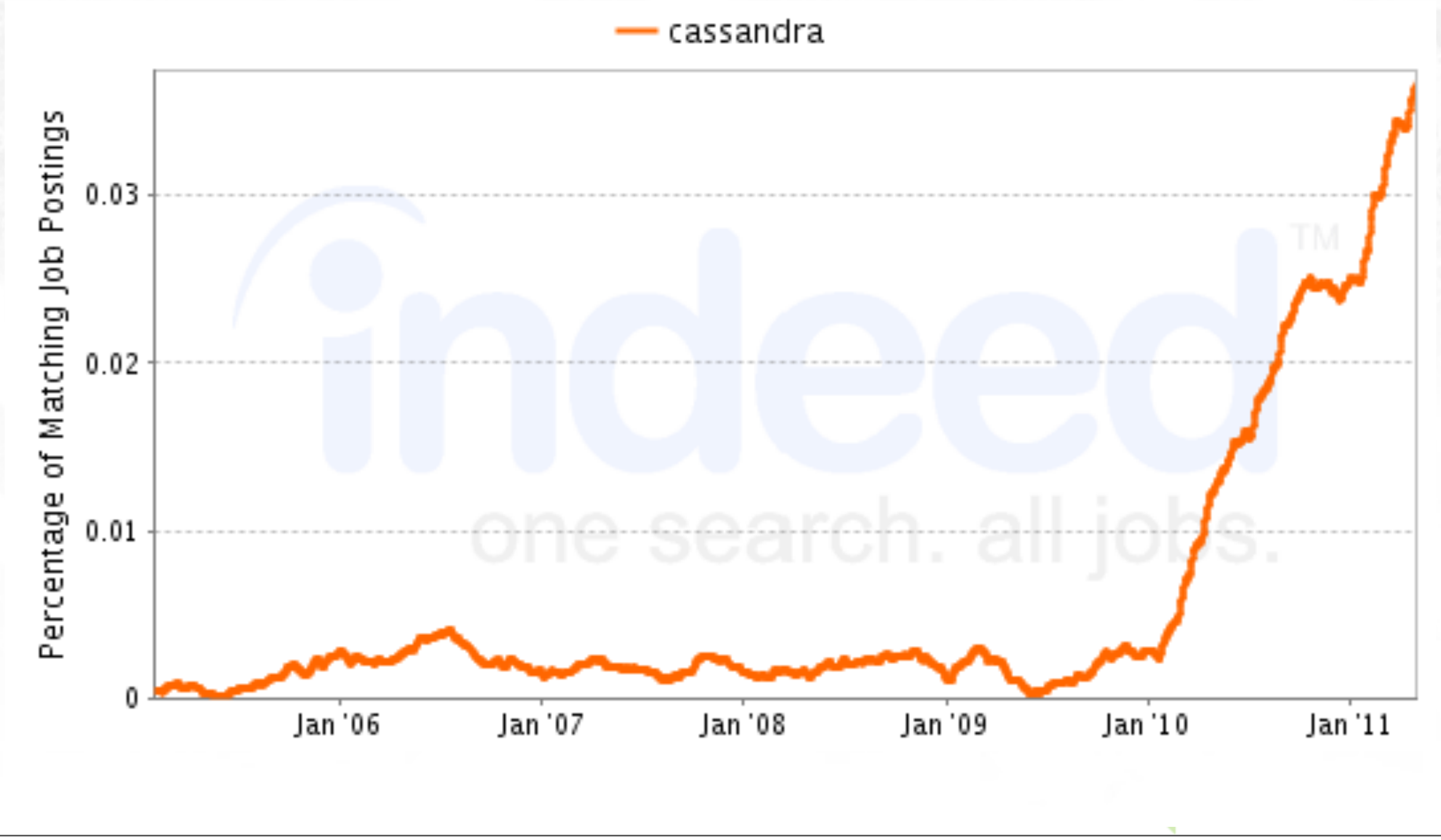
Presented by

Produced by

About me

- http://twitter.com/tjake
- Cassandra Committer
- Thrift PMC
- Early DataStax employee
- Ex-Wall St. (happily)

Job Trends from Indeed.com

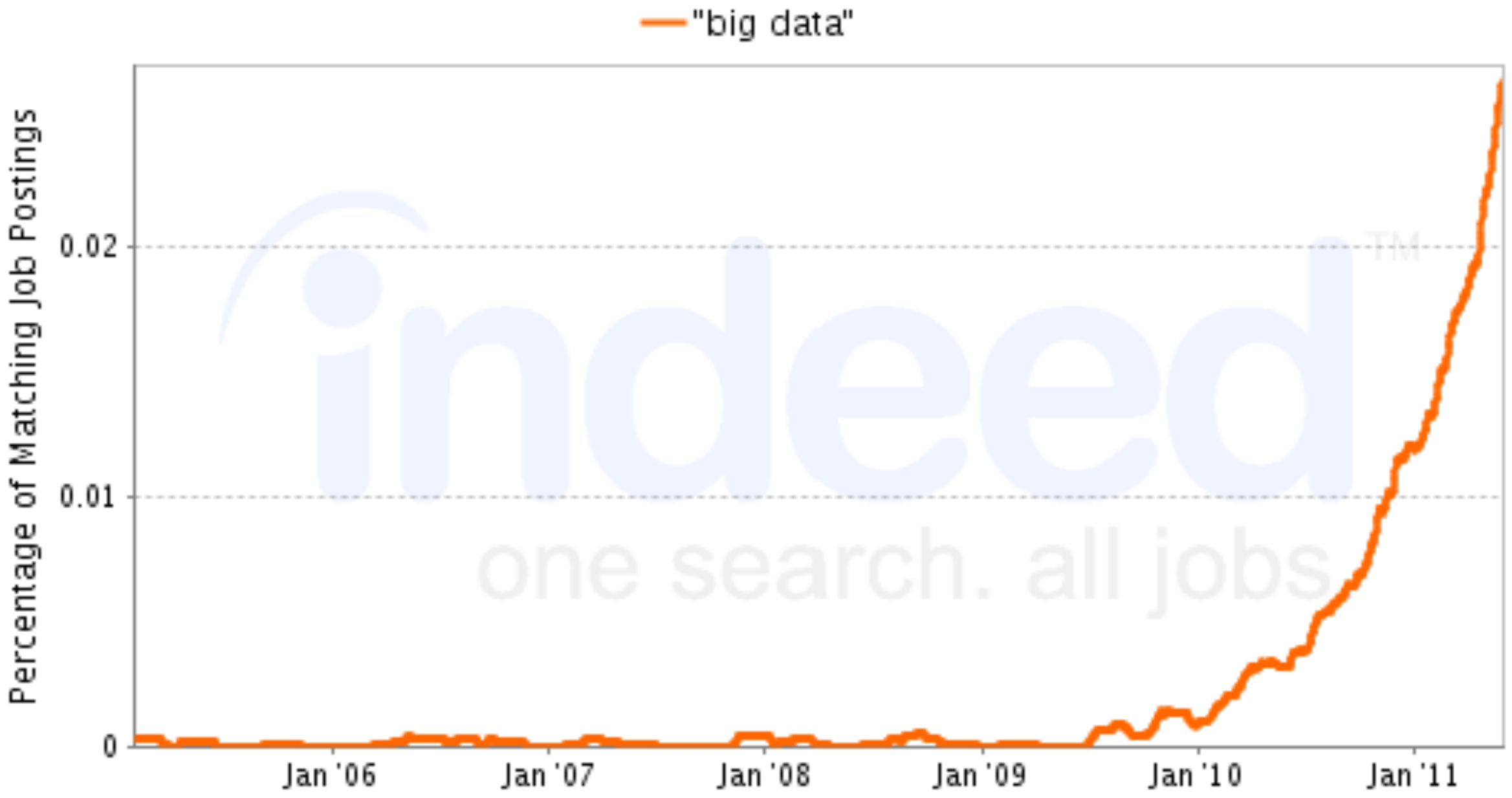


Who?

- Financial
- Social Media
- Advertising
- Entertainment
- Energy
- E-tail
- Health care
- Government

http://datastax.com/cassandrausers

Why?


"big data"
Percentage of Matching Job Postings
0.02
0.01
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11

Why Cassandra?

Better technology for OLTP

- Multi-master, multi-DC
- Linearly scalable
- Larger-than-memory datasets
- Best-in-class performance (*not just writes!*)
- Fully durable
- Integrated caching
- Tunable consistency
- Monolithic

Cassandra: History

Google

Bigtable, 2006

amazon.com

Dynamo, 2007

facebook

OSS, 2008

Incubator, 2009

TLP, 2010

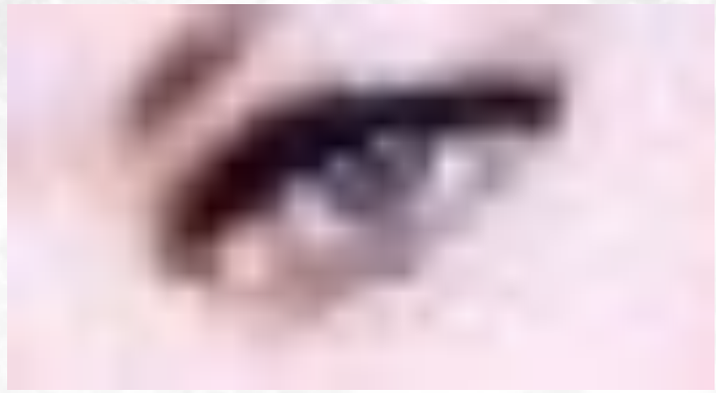
1.0, October 2011

What's with the eye: Theory #1

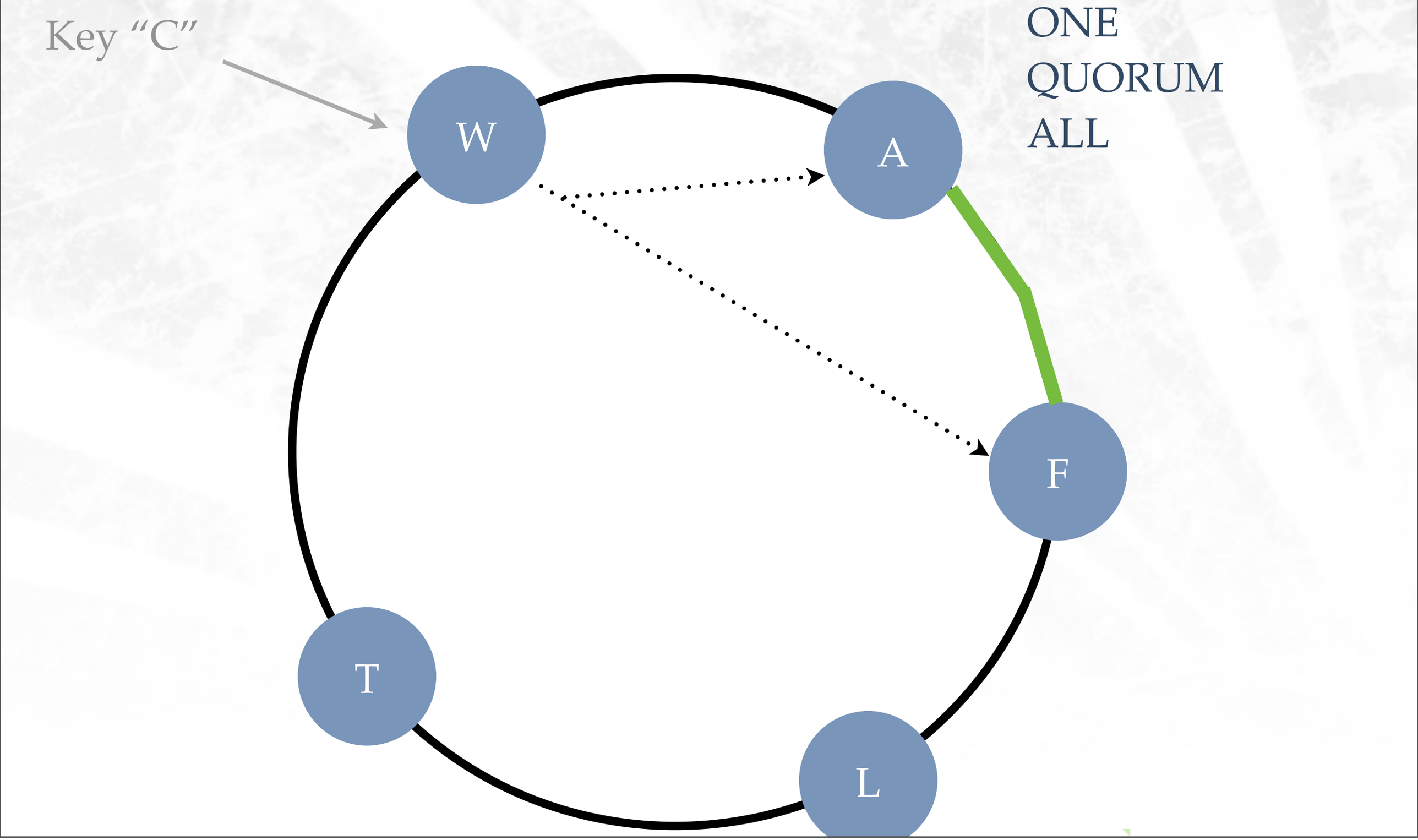
- PMC hosted a logo contest on 99designs.com
- http://99designs.com/logo-design/contests/open-source-project-logo-28940
- Called for a vote on cassandra-user
- Eye logo won out of 618 designs
-

What's with the eye: Theory #2

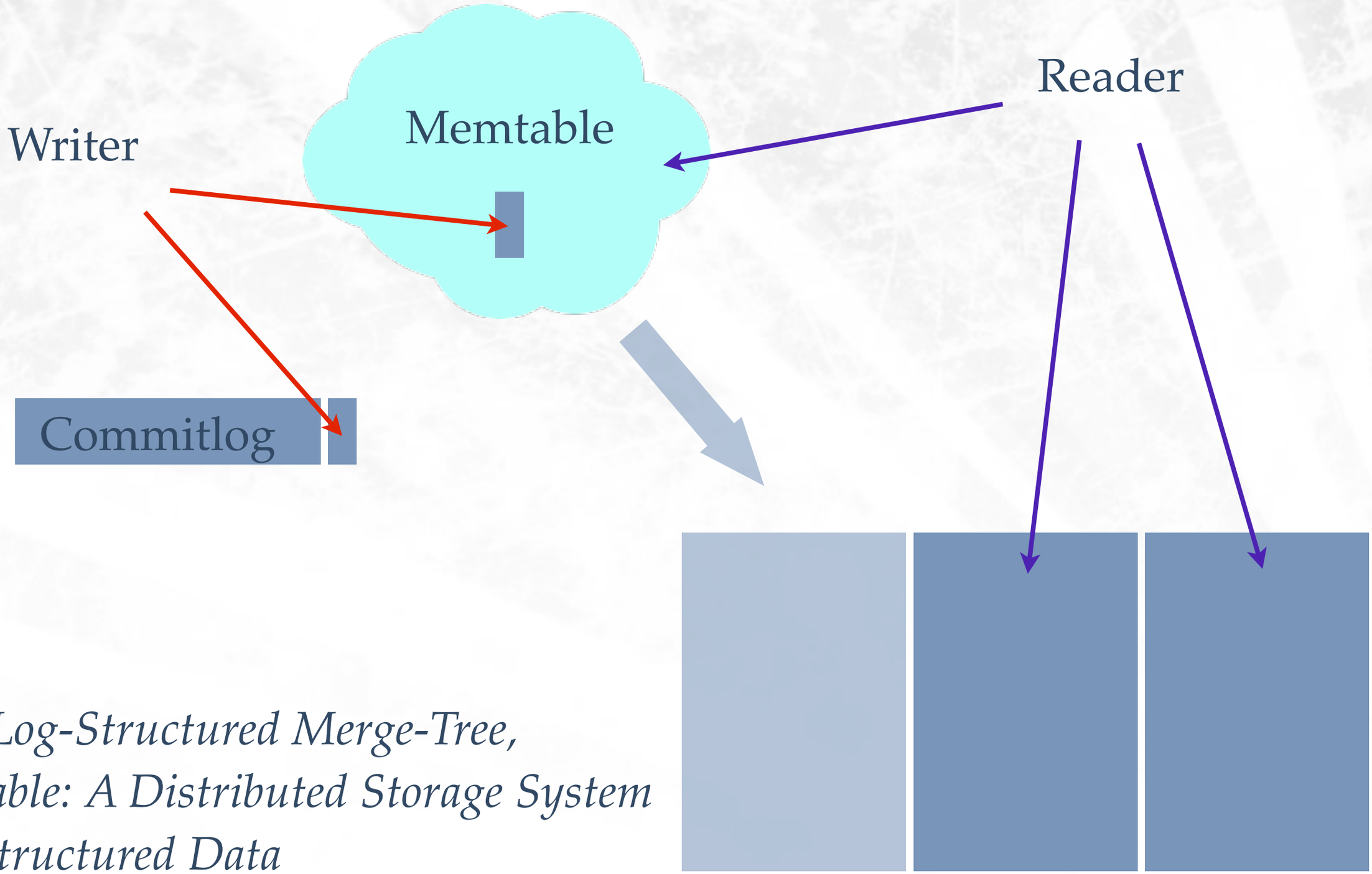
Dr. Cassandra and her husband, Cabala.



Cassandra: Dynamo

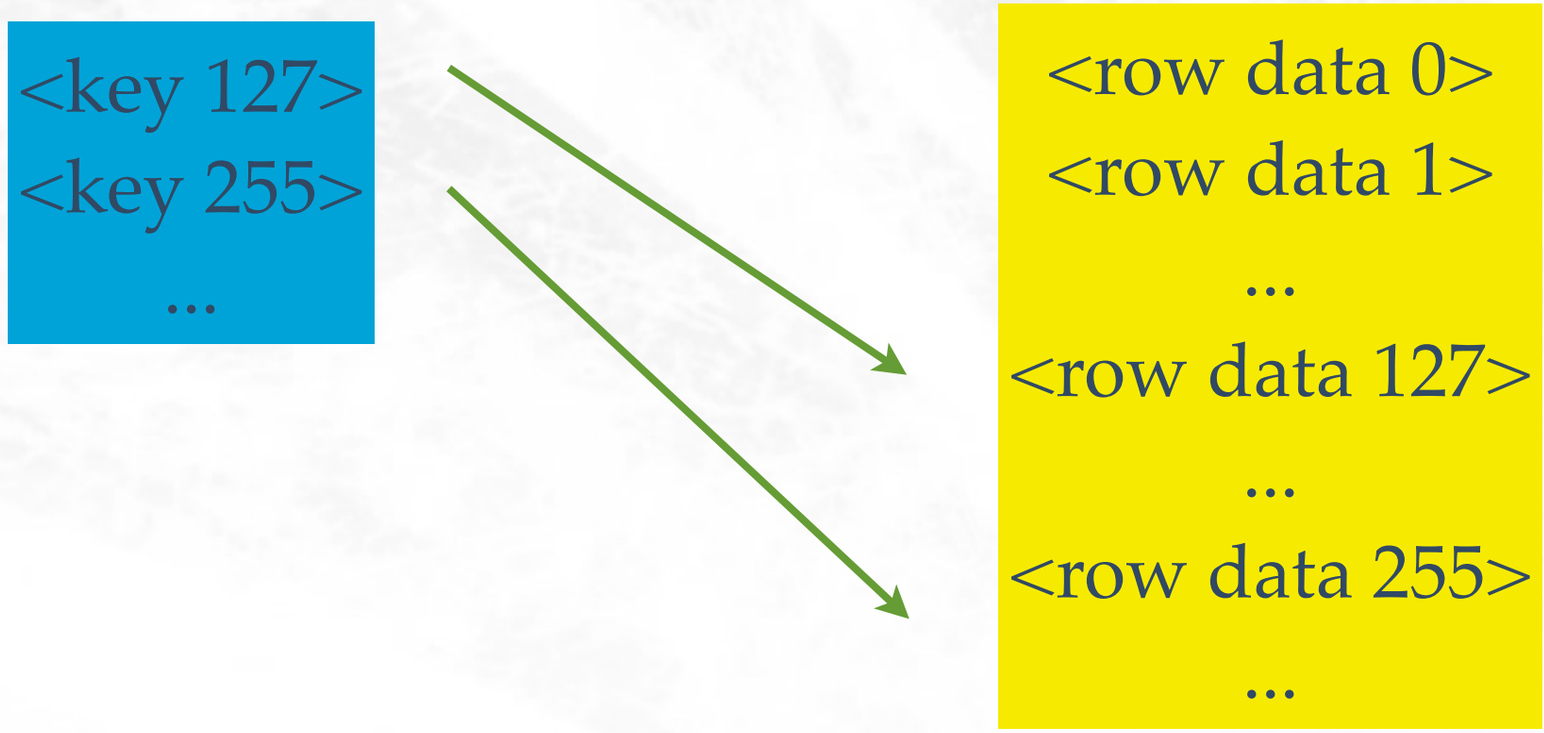


Cassandra: BigTable



The Log-Structured Merge-Tree, Bigtable: A Distributed Storage System for Structured Data

Cassandra: BigTable



Sorted [clustered] by row key

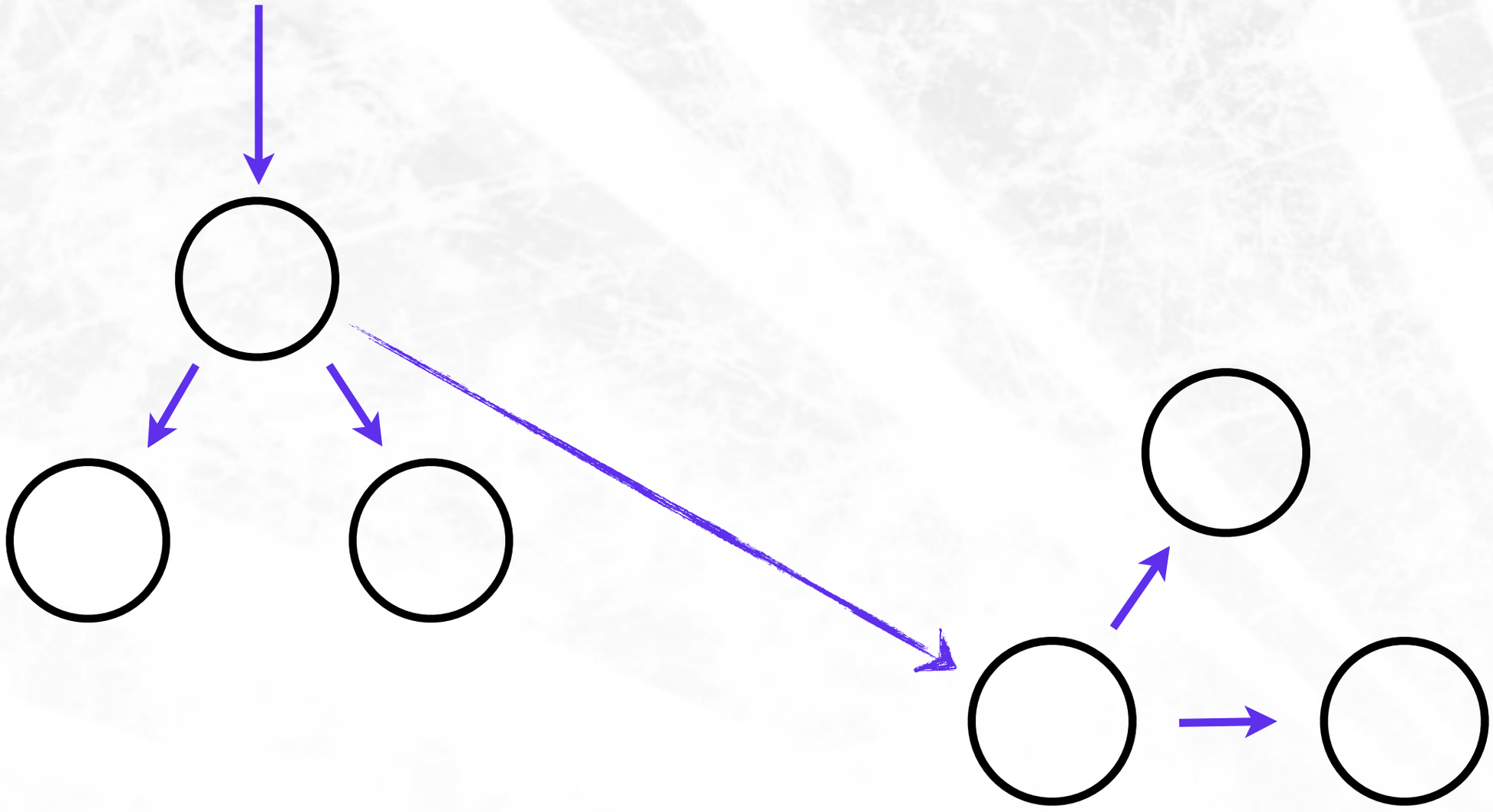
Apache Cassandra: The Early Years (0.4-0.6)

- Fixed ColumnFamilies
- Poor Read Performance (<50% of Write)
- Lots of Onions in the code
- Difficult to operate cluster (no rolling restarts)

0.7

- CREATE COLUMN FAMILY
- Expiring columns (TTL)
- Secondary (column) indexes
- Efficient streaming

0.7+ Multi-DC writes



0.8

- CQL
- Counters
- Automatic memtable tuning
- New bulk load interface

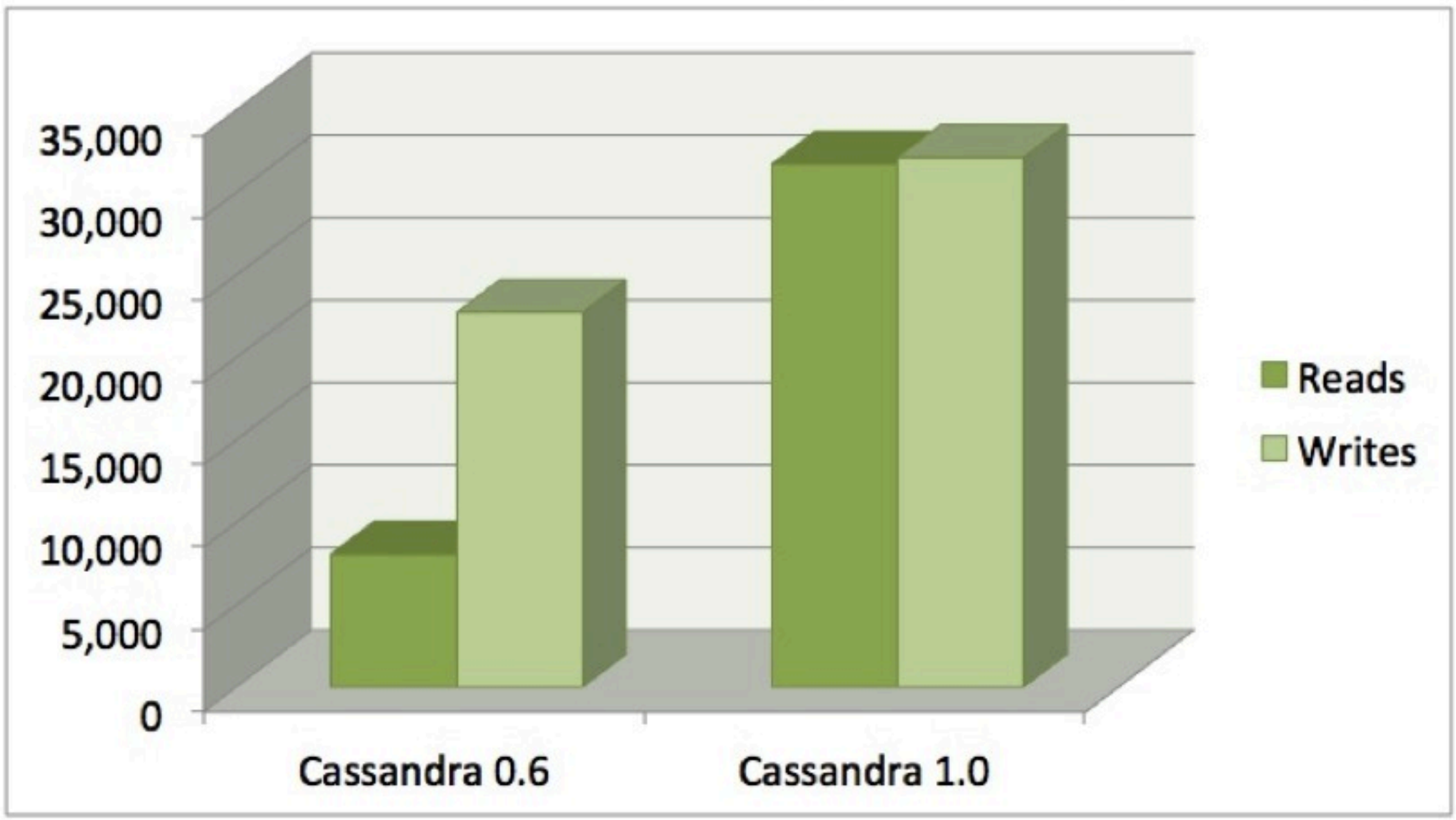
1.0

- Compression
- Read performance
- LeveledCompactionStrategy
- CQL 1.1
- Better Hinted Handoff Tracking

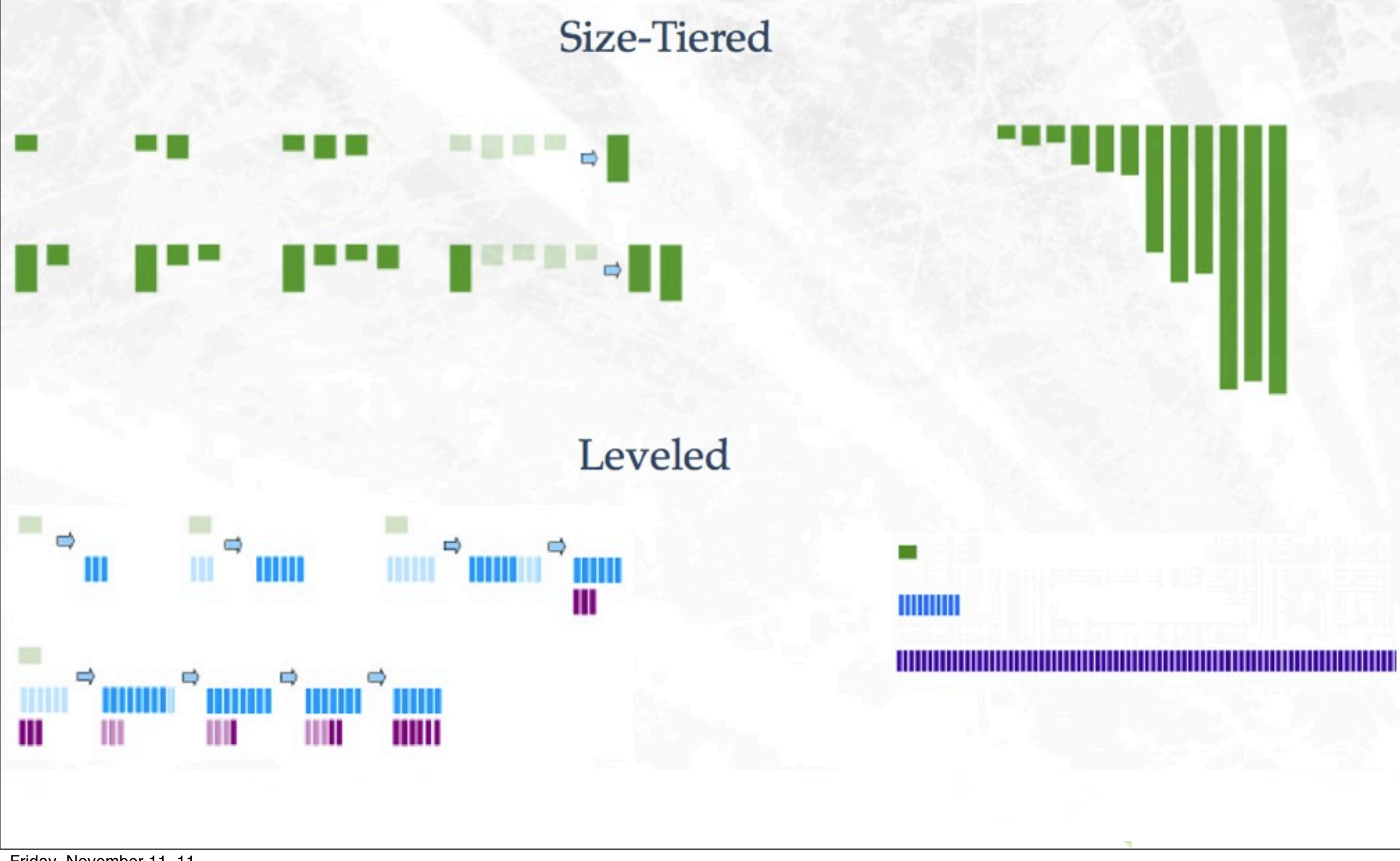
Read performance: maxtimestamp

- Sort sstables by maximum (client-provided) timestamp
- Only merge sstables until we have the columns request
- Allows pre-merging highly fragmented rows without waiting for compaction

A performance retrospective



Pluggable Compaction Strategies



Leveled Compaction Compared to LevelDB

- ColumnFamily instead of key/value
- Multithreading (experimental)
- Optional throttling (16MB/s by default)
- Per-sstable bloom filter for row keys
- Larger data files (5MB by default)
- Does not block writes if compaction falls behind

Dealing with the JVM

- JNA
  - mlockall()
  - posix_fadvise()
  - link()
- Memory
  - Move cache off-heap
  - In-heap arena allocation for memtables, bloom filters
  - Later: move compaction to another process?

CQL

```
cqlsh> CREATE INDEX state_idx ON users(state);
cqlsh> INSERT INTO users (uname, state, birth_date)
       VALUES ('bsanderson', 'UT', 1975)
```

users

	state	birth_date
bsanderson	UT	1975
prothfuss	WI	1973
htayler	UT	1968

state_idx

UT	bsanderson	htayler
WI	prothfuss	

```
cqlsh> SELECT * FROM users WHERE state='UT' AND birth_date > 1970;

        KEY | birth_date |          full_name | state |
 bsanderson |       1975 | Brandon Sanderson |    UT |
```

CQL 1.1

- ALTER
- Counter support
- TTL support

Post-1.0

- Ease of use

Post-1.0

- Ease of use

- Ease of use

Post-1.0

- Ease of use

- Ease of use
- Ease of use

Post-1.0 features

- CQL
  - “Native” transport
  - Composite columns (transpose)
  - Prepared statements
- Triggers
- Entity groups
- Lucene based secondary indexes
- Smarter range queries
  - Enables more-efficient (OLAP)

Cassandra + Hadoop

- Many clients use Cassandra with Hadoop
  - Cassandra -> Hadoop -> Cassandra
- ColumnFamilyInput/Output Format
- Pig Driver
- Hive Driver

- DataStax Enterprise (shameless plug)

The Cassandra EcoSystem

- Commercial entities starting around Cassandra
- Replication into Cassandra
  - GigaSpaces
  - Drizzle
  - EsperHA
- Solandra: Solr + Cassandra
- Virgil: REST Service
- Many Community Maintained API wrappers
- DataStax Enterprise: Cassandra + Hadoop + ?

Questions?

- @tjake
- jake@datastax.com