DATASTAX

Apache Cassandra 2.0 – #Cassandra

```
USE aarhus;
SELECT * FROM presenters WHERE name = 'Hayato Shimizu';

 name             | title                    | company   | area
------------------+--------------------------+-----------+------
 Hayato Shimizu   | Solutions Architect      | DataStax  | EMEA
```

DataStax and Cassandra

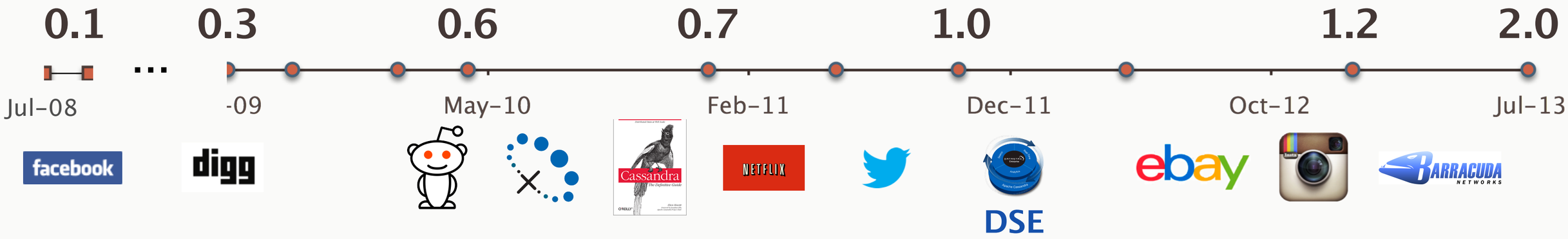
DATASTAX

- Commercial company behind Apache Cassandra
- Cassandra is a highly distributed database

http://planetcassandra.org

http://www.datastax.com

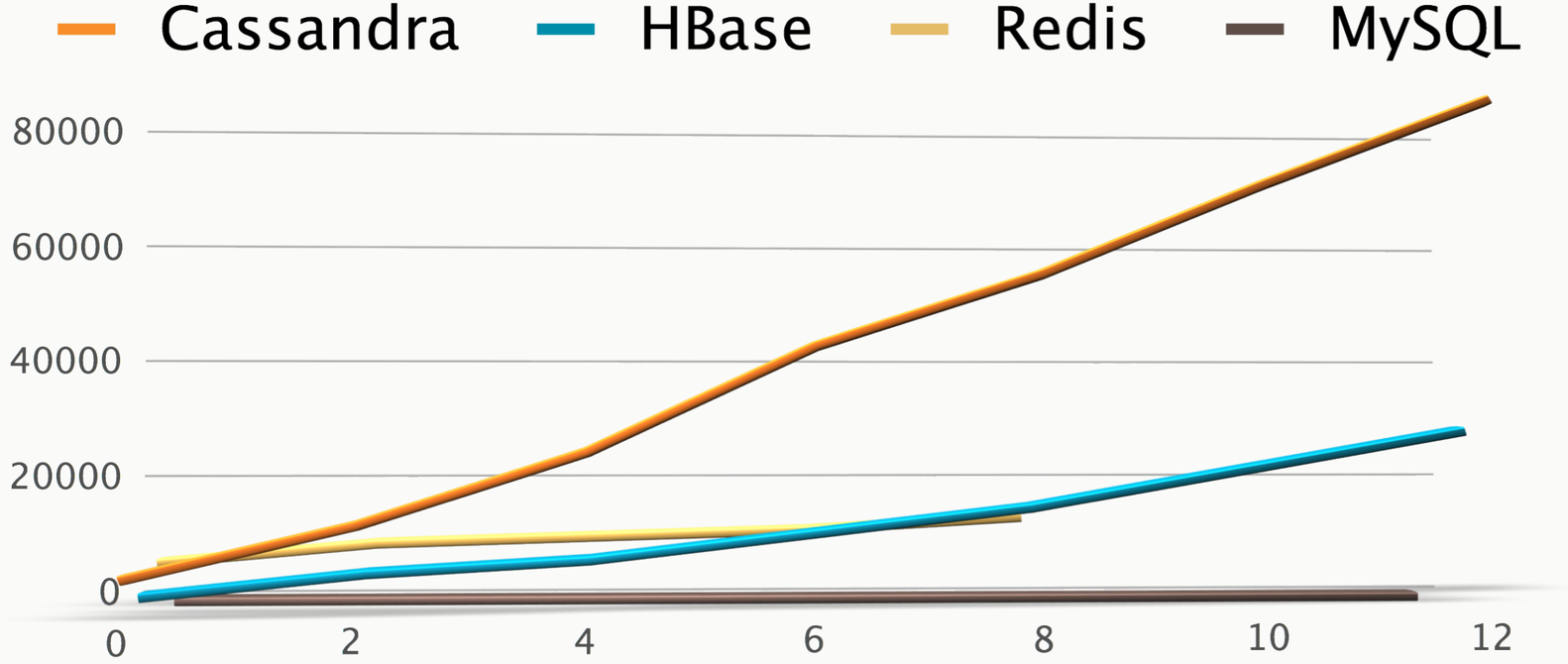
Five Years of Cassandra

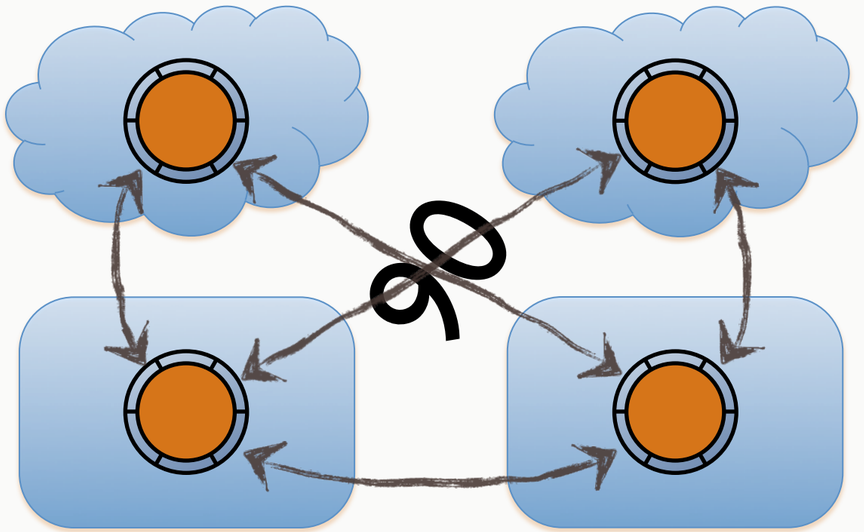


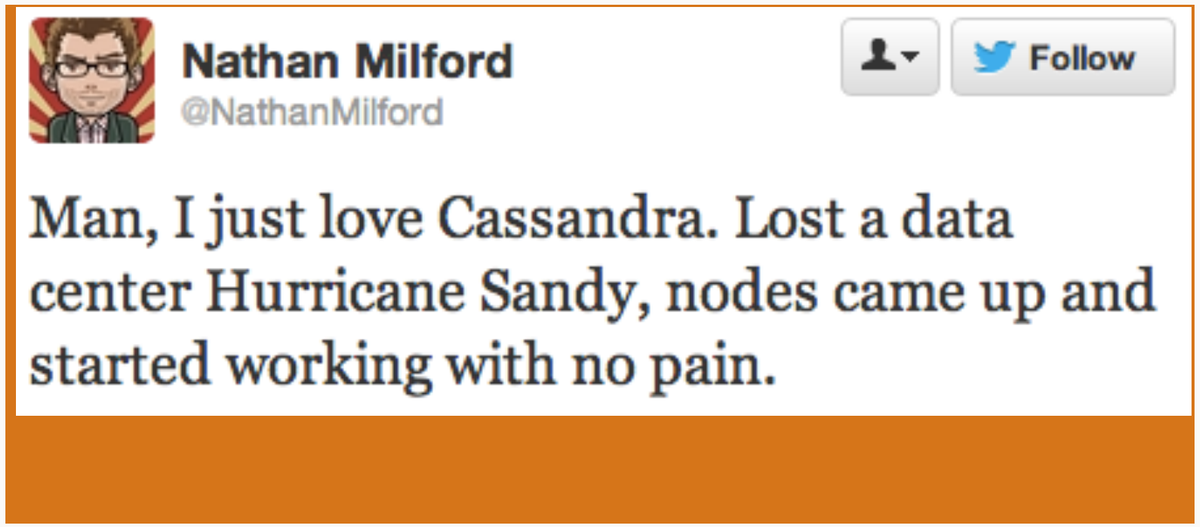
Core values

- Massive scalablility
- High performance
- Reliability/Availability

http://vldb.org/pvldb/vol5/p1724_tilmannrabl_vldb2012.pdf







New Core Value

DATASTAX

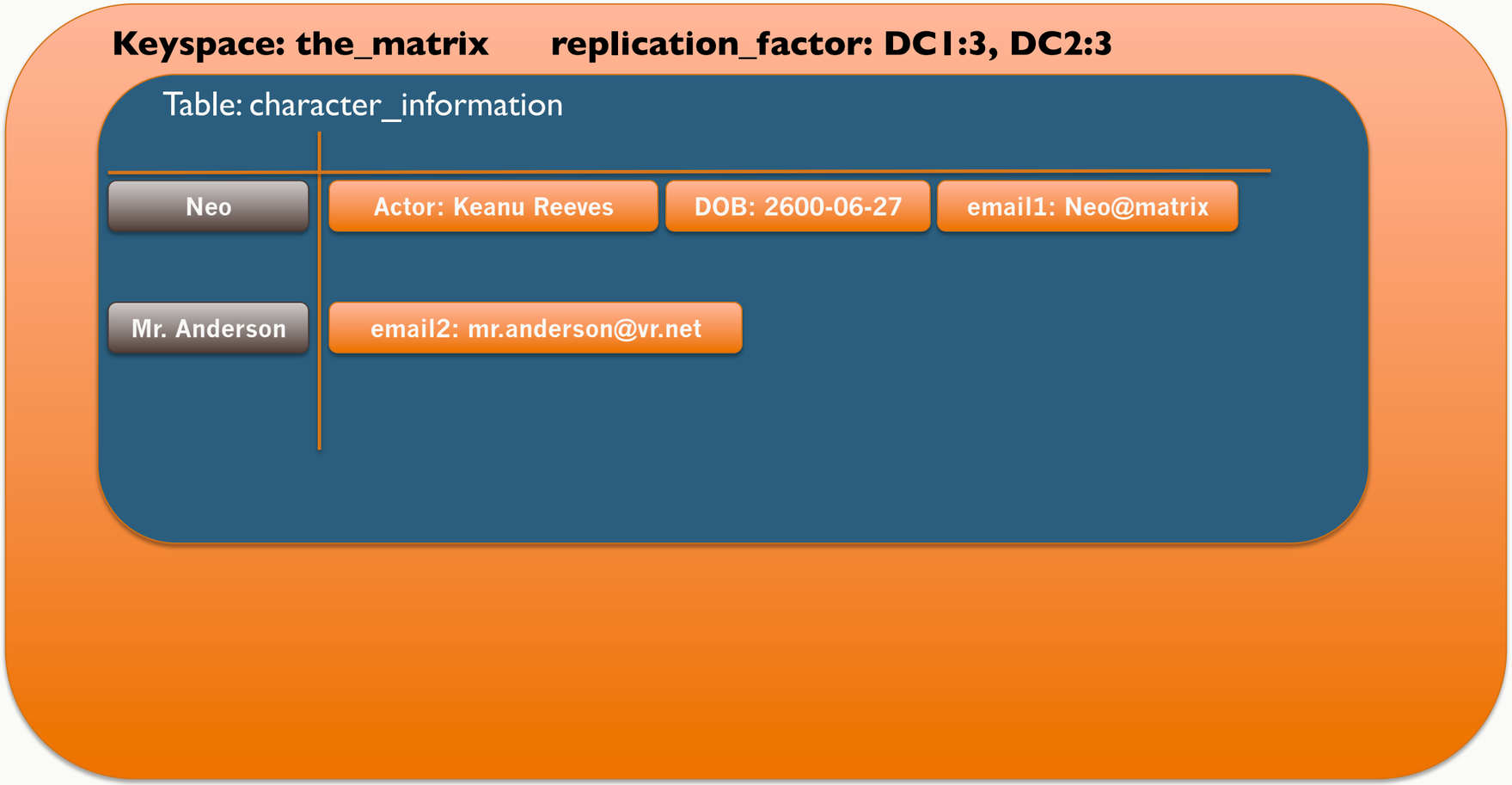
- Massive scalability
- High performance
- Reliability/Availability
- **Ease of use**

```
CREATE TABLE users (
  id text PRIMARY KEY,
  name text,
  state text,
  birth_date int,
  email text
);


INSERT INTO users
(id, name, state, birth, email)
VALUES
('hshimizu', 'Hayato Shimizu', 'Surrey',
'1-1-1995');


SELECT * FROM users
WHERE id = 'hshimizu';
```

DATASTAX

Cassandra Basics

Data Storage Structure



Cassandra Architecture – Data Replication

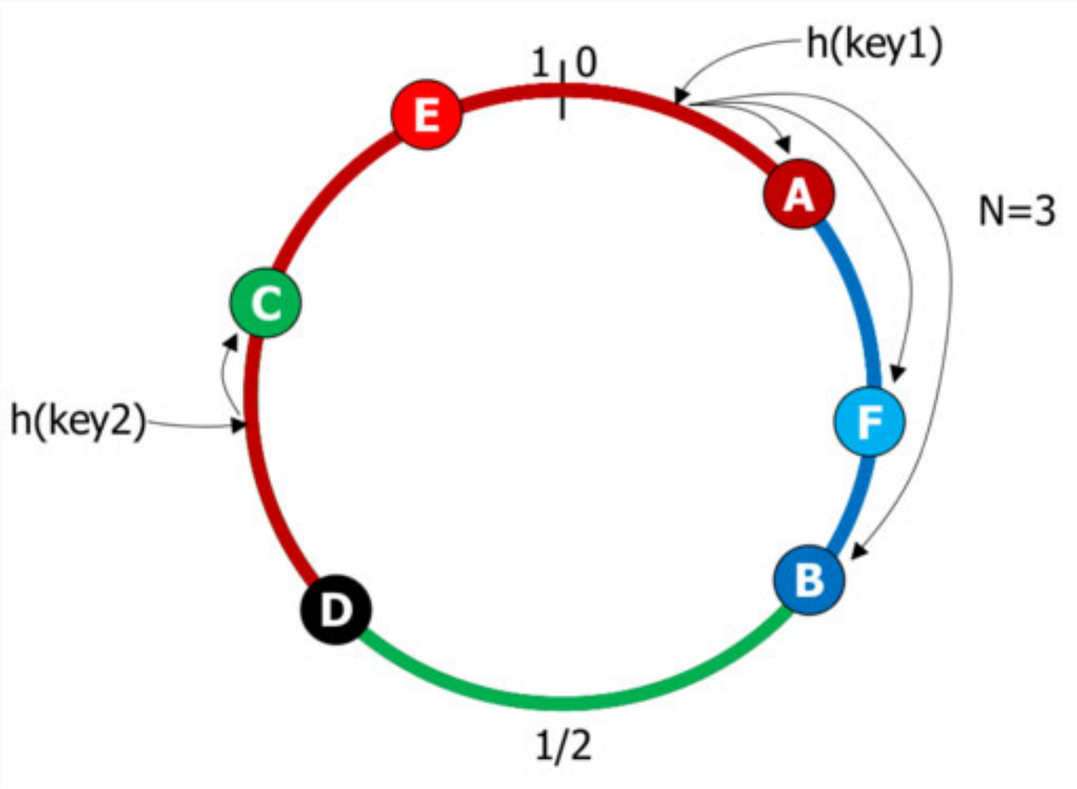
DATASTAX

Token Range: 0 -> 2^{127}

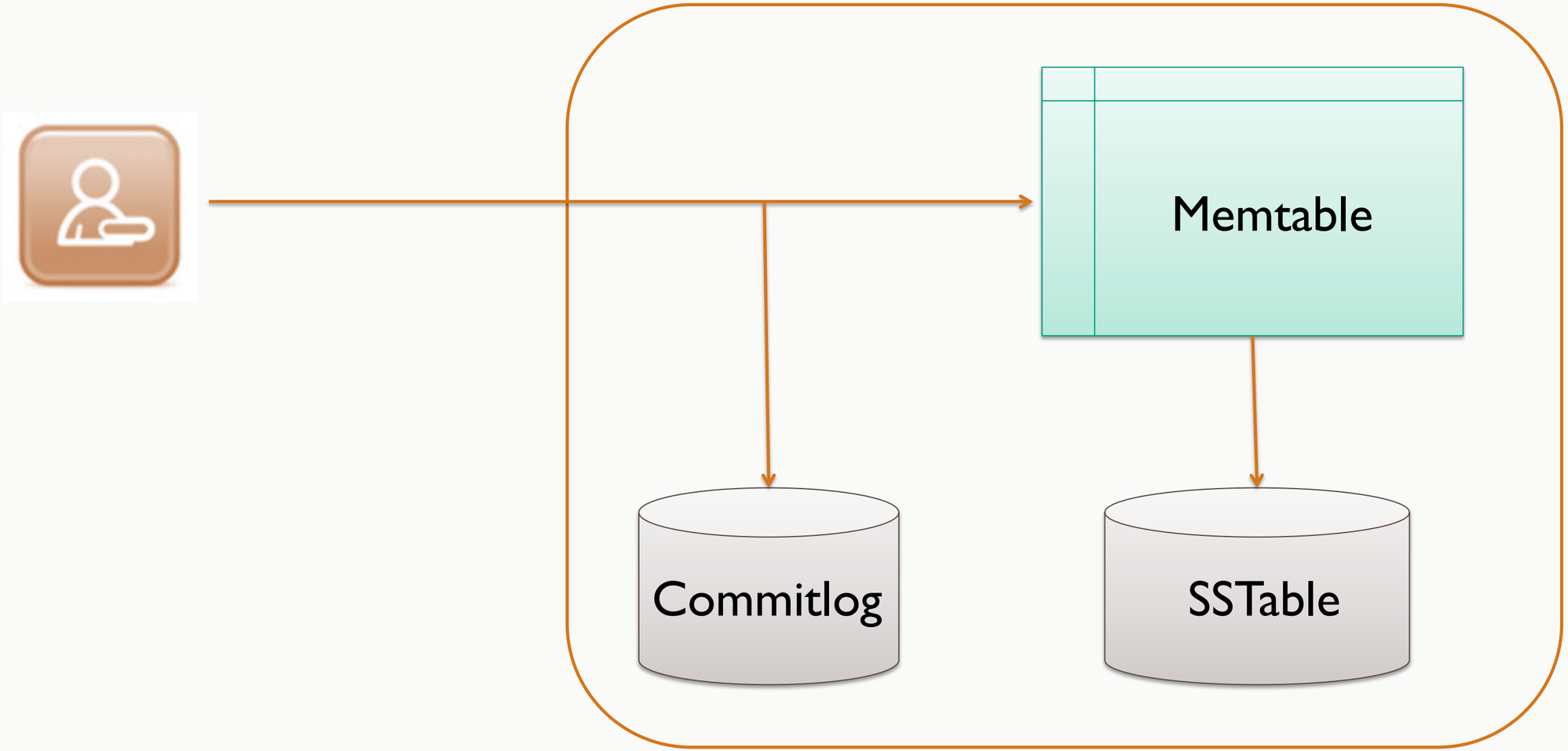
C* offers active everywhere strategy

C* offers flexible replication strategies with **TUNABLE CONSISTENCY**

One, Two, Three, Quorum,
Local Quorum, Each Quorum, All



Cassandra Architecture - Writes



Cassandra 1.2

1.2 for Developers

- **CQL3**
  - SQL Like
  - Collections – set, list, map
  - Data dictionary
  - Auth support
- **Tracing**
- **Atomic batches**

Authentication

- [cassandra.yaml]
- authenticator: PasswordAuthenticator
- # DSE offers KerberosAuthenticator as well

```
CREATE USER robin WITH PASSWORD 'manager' SUPERUSER;
ALTER USER cassandra WITH PASSWORD 'newpassword';
DROP USER cassandra;
```

Authorization

- [cassandra.yaml]
- authorizer: CassandraAuthorizer

```
GRANT select ON audit TO jonathan;
GRANT modify ON users TO robin;
GRANT all ON ALL KEYSPACES TO lara;
```

Native Protocol

DATASTAX

CQL native protocol: efficient, lightweight, asynchronous

Java (GA): https://github.com/datastax/java-driver

.NET (GA): https://github.com/datastax/csharp-driver

Python (Beta): https://github.com/datastax/python-driver

Coming soon: C++, PHP, Ruby, others

1.2 for Operators

- Virtual nodes
- JBOD improvements
- Off-heap bloom filters, compression metadata
- “Dense node” support (5-10TB/machine)
- Parallel leveled compaction

1.2.5+

- ~1/2 memory usage in partition summary
- Improved compaction throttle
- Removed cell-name bloom filters
- Thread-local allocation
- LZ4 compression (default in 2.0)
- (1.2.7) CQL Input/Output for Hadoop
- (1.2.7) Range tombstone performance
- (1.2.9) Larger default LCS filesize (160MB > 5MB)

DATASTAX

Cassandra 2.0

2.0

- Lightweight transactions
- Triggers (experimental)
- Improved compaction
- CQL cursors
- Streaming re-written

Lightweight Transactions: the problem

DATASTAX

Session 1

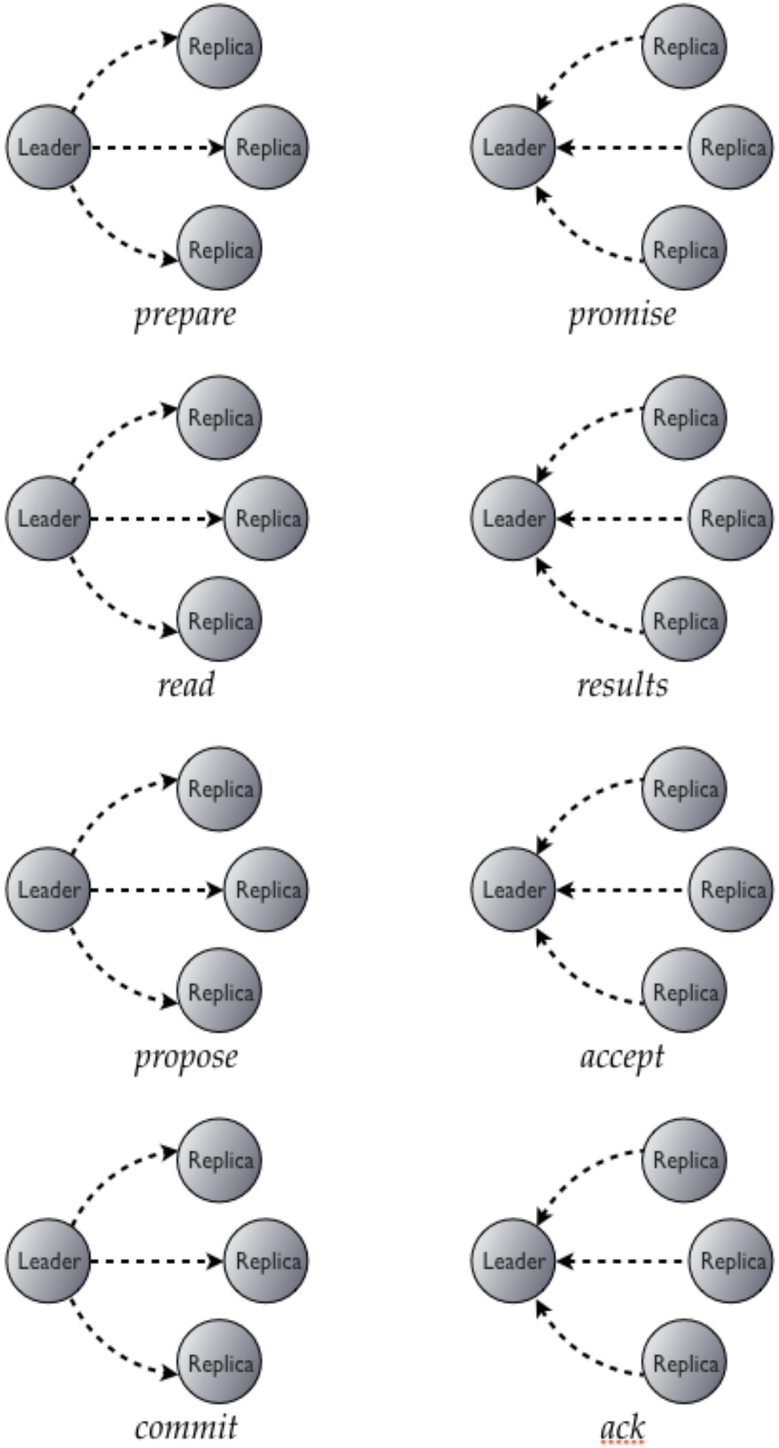
```
SELECT * FROM users
WHERE username = 'jbellis'

[empty resultset]

INSERT INTO users (...)
VALUES ('jbellis', ...)
```

Session 2

```
SELECT * FROM users
WHERE username = 'jbellis'

[empty resultset]

INSERT INTO users (...)
VALUES ('jbellis', ...)
```

Paxos

- All operations are quorum-based
- An elected leader prepares the participating replicas a ballot
- Replicas would reply with the promise
- Each replica sends information about unfinished operations to the leader during prepare
- Paxos Made Simple
- Paxos Made Live – An Engineering Perspective



LWT: details

- 4 round trips vs 1 for normal updates
- Paxos state is durable
- ConsistencyLevel.SERIAL
- http://www.datastax.com/dev/blog/lightweight-transactions-in-cassandra-2-0

Using LWT

```
CREATE TABLE USERS IF NOT EXISTS (
    username text,
    email text
    …
);

INSERT INTO USERS (username, email, ...)
VALUES (‘jbellis’, ‘jbellis@datastax.com’, ... )
IF NOT EXISTS;

UPDATE USERS
SET email = ’jonathan@datastax.com’, ...
WHERE username = ’jbellis’
IF email = ’jbellis@datastax.com’;
```

LWT: Use with caution

- Great for 1% of your application
- Eventual consistency is your friend
  - http://www.slideshare.net/planetcassandra/c-summit-2013-eventual-consistency-hopeful-consistency-by-christos-kalantzis

Triggers - Experimental

- CREATE TRIGGER <name> ON <table> USING <classname>;
- Expect Changes in 2.1

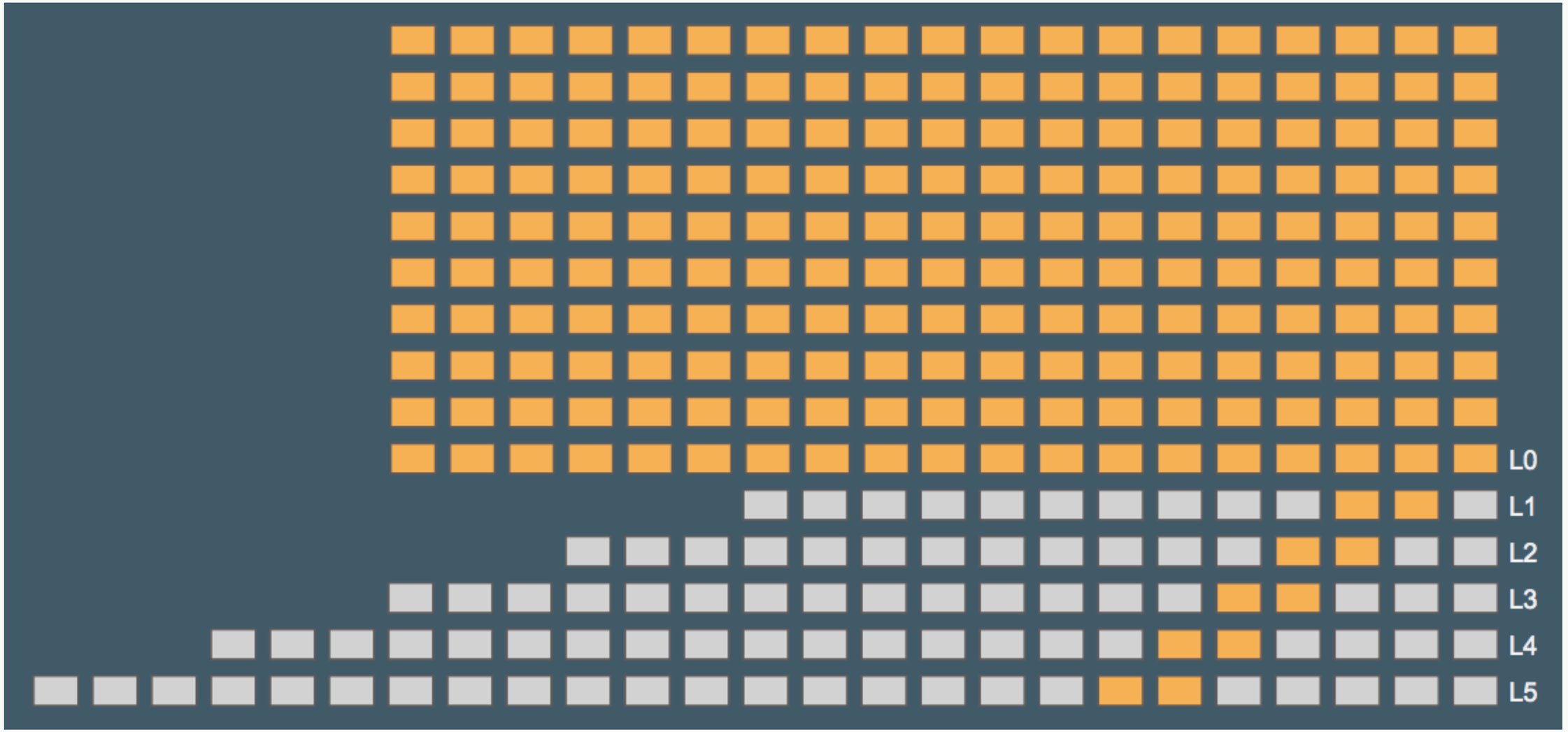
```
class MyTrigger implements ITrigger
{
    public Collection<RowMutation> augment(ByteBuffer key, ColumnFamily update)
    {
        ...
    }
}
```

Compaction

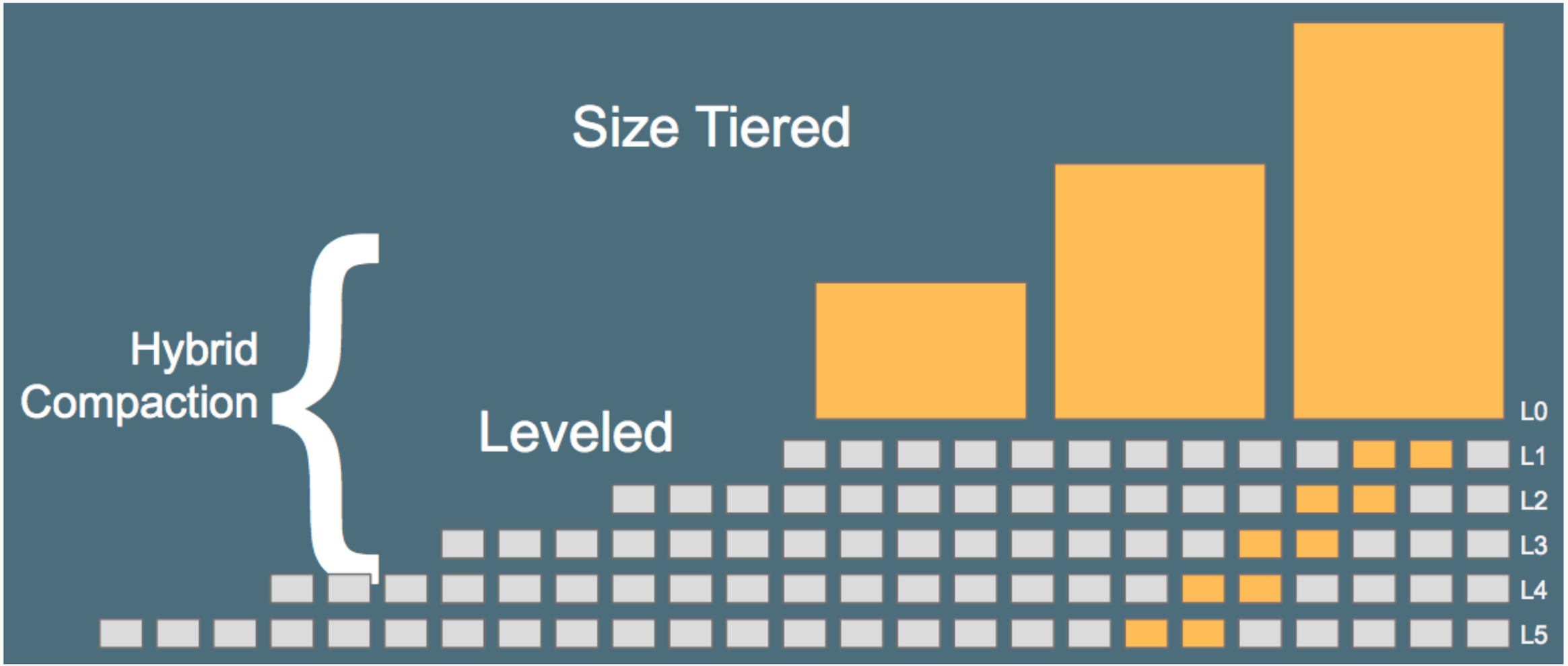
- Single-pass, always
- LeveledCompactionStrategy performs SizeTieredCompactionStrategy in Level 0

Healthy leveled compaction

Sad leveled compaction



STCS in L0



Cursors (before)

DATASTAX

```
CREATE TABLE timeline (
  user_id uuid,
  tweet_id timeuuid,
  tweet_author uuid,
  tweet_body text,
  PRIMARY KEY (user_id, tweet_id)
);
```

```
SELECT *
FROM timeline
WHERE (user_id = :last_key
       AND tweet_id > :last_tweet)
   OR token(user_id) > token(:last_key)
LIMIT 100
```

Cursors (after)

DATASTAX

```
SELECT * FROM timeline;
```

Misc. performance improvements

DATASTAX

Tracking statistics on clustered columns allows eliminating unnecessary sstables from the read path

- New half-synchronous, half-asynchronous Thrift server based on LMAX Disruptor
- Faster partition index lookups and cache reads by improving performance of off-heap memory
- Faster reads of compressed data by switching from CRC32 to Adler checksums
- JEMalloc support for off-heap allocation

Spring cleaning

- Removed compatibility with pre-1.2.5 sstables and pre-1.2.9 schema
- The potentially dangerous countPendingHints JMX call has been replaced by a Hints Created metric
- The on-heap partition cache (“row cache”) has been removed
- Vnodes are on by default
- the old token range bisection code for non-vnode clusters is gone
- Removed emergency memory pressure valve logic

Operational concerns

Java7 is now required!

- Leveled compaction level information has been moved into sstable metadata
- Kernel page cache skipping has been removed in favor of optional row preheating (preheat_kernel_page_cache)
- Streaming has been rewritten to be more transparent and robust.
- Streaming support for old-version sstables

DataStax Enterprise

- Analytics Integration
- Search Integration
- Security Enhancement
- Production Support

DATASTAX

http://planetcassandra.org

http://www.datastax.com