Cassandra
A Decentralized, Structured Storage System

Avinash Lakshman
and
Prashant Malik
Facebook

Published: April 2010, Volume 44, Issue 2
Communications of the ACM
http://dl.acm.org/citation.cfm?id=1773922

Presented by James Owens
Old Dominion University
For CS795 on 11//2014

About the Authors

Avinash Lakshman

- Currently:
    - Hedvig / Quexascale ? – 2013
- Notable Works:
    - **Dynamo**: amazon's highly available key-value store - 2007
    - **Cassandra**: a decentralized structured storage system - 2010
    - **Cassandra**: structured storage system *on a p2p network* - 2009
    - System and method for providing high availability data - 2010

Prashant Malik

- Currently:
    - LimeRoad ? – 2013
- Notable works:
    - **Cassandra**: a decentralized structured storage system - 2010
    - **Cassandra**: structured storage system on a p2p network - 2009
    - Asynchronous communication within a server arrangement - 2007
    - Publishing digital content within a defined universe such as an organization in accordance with a digital rights management (DRM) system - 2009

Rank 4 search results on author name - Google Scholar 11/14/2014
http://www.bizjournals.com/sanjose/news/2013/06/25/ex-facebook-amazon-data-engineer.html
http://techcircle.vccircle.com/2013/05/30/ex-facebook-tech-lead-prashant-malik-joins-social-commerce-firm-limeroad-as-cto/

Significance

Provides a platform for data storage and retrieval which supports very high write throughput and tolerates continuous component failure.

Integrates strategies from many other technologies, cited over 916 times. [Google Scholar, Nov 2014]

Difficulties

- Dense reading.
- No Diagrams.
- Simultaneously defines a general purpose tool(Cassandra) and specific implementation (Inbox Search)
- No clear separation of the above.

Approach

- Handling Density:
    - Inbox Search
        - Big-Picture View of Cassandra
            - Data Model
            - API
            - Read/Write Model
        - How Cassandra solves Inbox Search
    - Cassandra Internals…

Why was Cassandra Created?

- Solution to the **Inbox Search** problem
  - Consider the Facebook context:
    - Many simultaneous users
    - Billions of writes per day
    - Need for Scalability
- Cassandra is used for multiple services within Facebook.

Inbox Search Problem

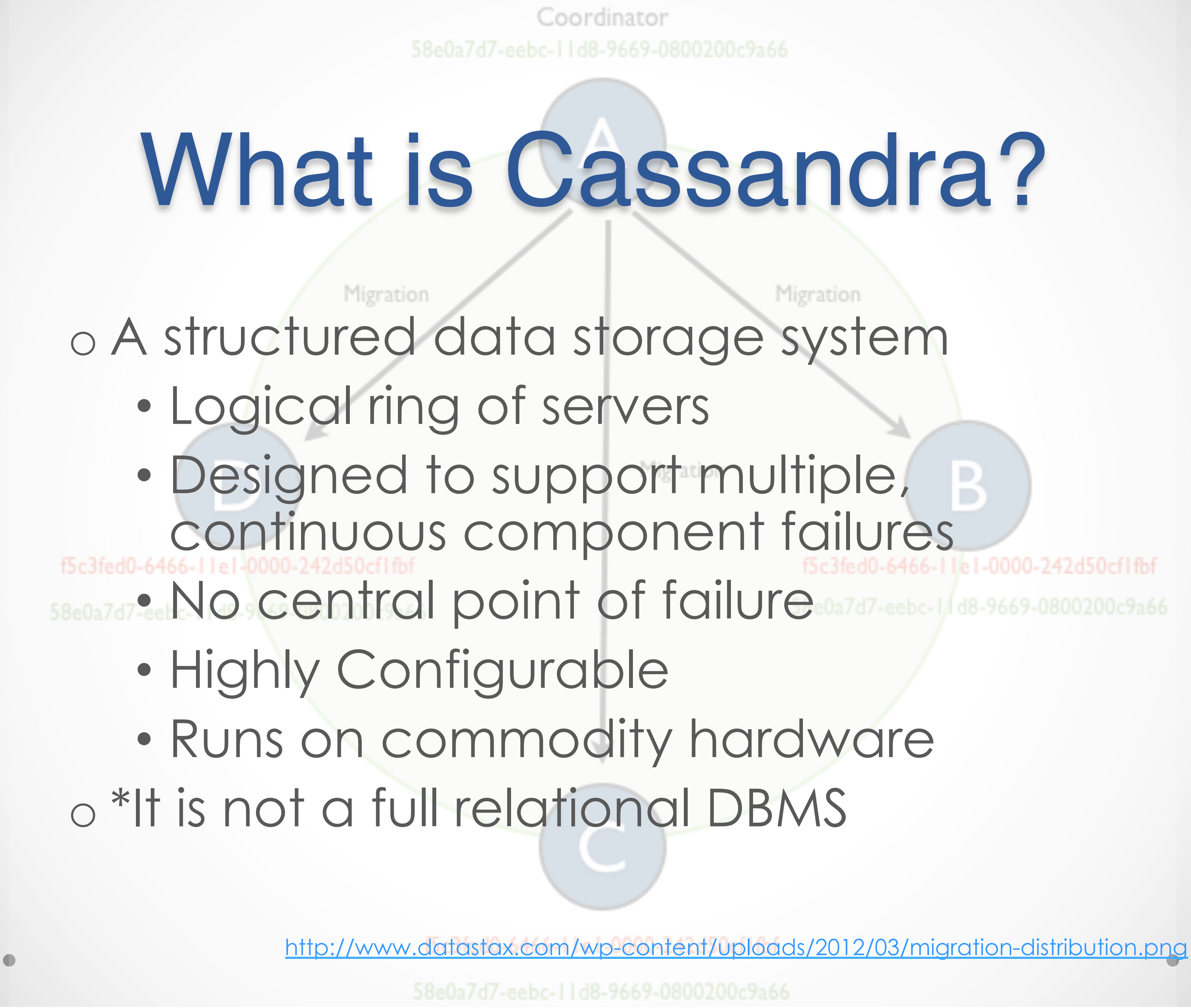
- A user wants to search his or her inbox for messages using one of two strategies
    - Term Search - keyword
    - Interactions - name

What is Cassandra?

○ A structured data storage system

- Logical ring of servers
- Designed to support multiple, continuous component failures
- No central point of failure
- Highly Configurable
- Runs on commodity hardware

○ *It is not a full relational DBMS


Coordinator
58e0a7d7-eebc-11d8-9669-0800200c9a66
A
Migration
Migration
Migration
D
B
f5c3fed0-6466-11e1-0000-242d50cf1fbf
f5c3fed0-6466-11e1-0000-242d50cf1fbf
58e0a7d7-eebc-11d8-9669-0800200c9a66
58e0a7d7-eebc-11d8-9669-0800200c9a66
C
58e0a7d7-eebc-11d8-9669-0800200c9a66

http://www.datastax.com/wp-content/uploads/2012/03/migration-distribution.png

Data Model

- Distributed Multidimensional Map
    - < Key : Value > Pairs
        - String Key
            - Typically 16 – 36 B
        - Object Value

Data Model

- Column Types:
  - Column Families
  - Super Column Families
    - A superset of column families
    - (The Value Model supports recursion)

- A Visual…

Data Model

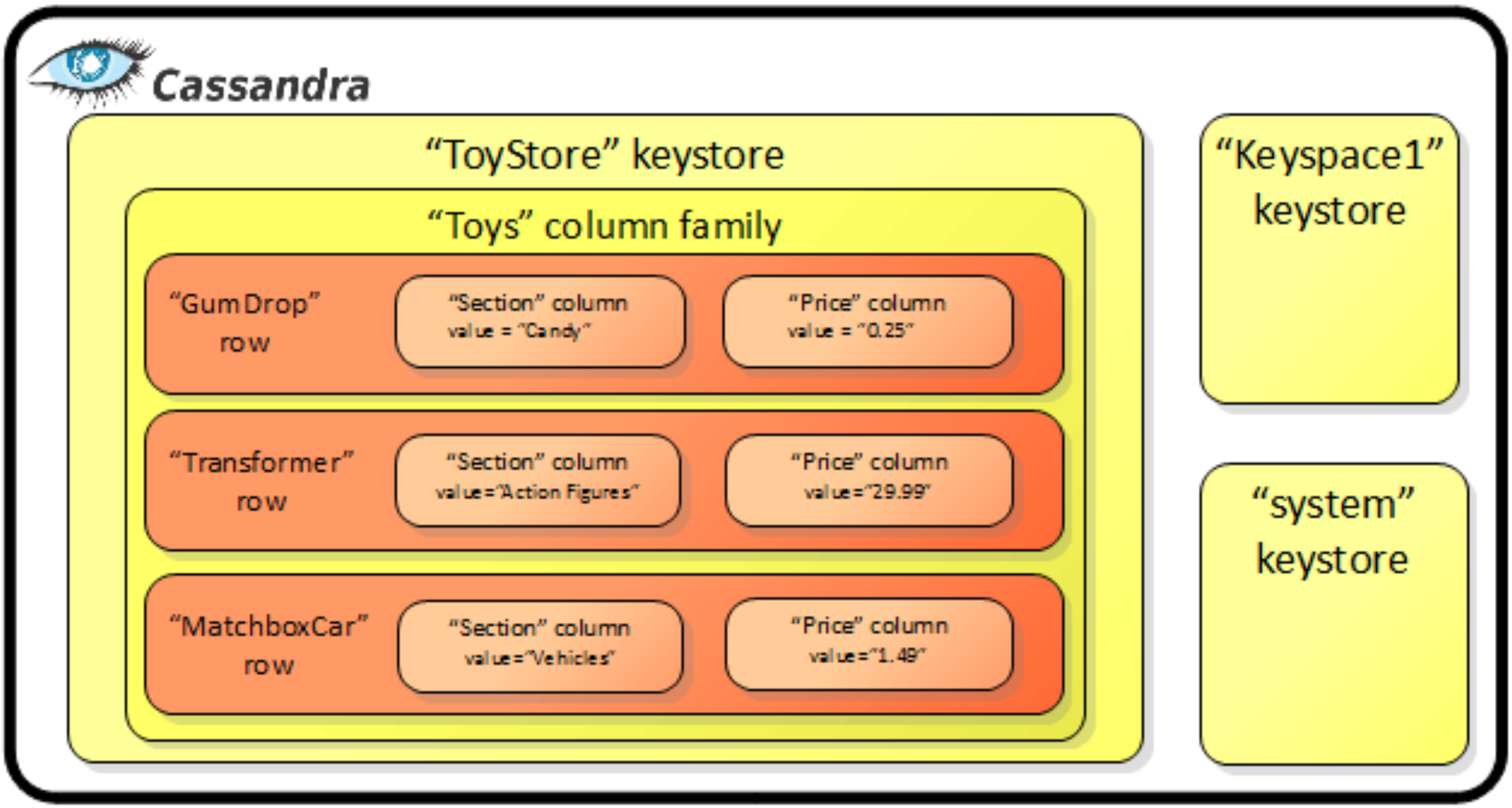

Cassandra
"ToyStore" keystore
"Toys" column family
"GumDrop" row
"Section" column value = "Candy"
"Price" column value = "0.25"
"Transformer" row
"Section" column value="Action Figures"
"Price" column value="29.99"
"MatchboxCar" row
"Section" column value="Vehicles"
"Price" column value="1.49"
"Keyspace1" keystore
"system" keystore

Image is Public Domain: Released July 11, 2010 by divconq.com

http://www.divconq.com/category/cassandra/page/2/

Data Model

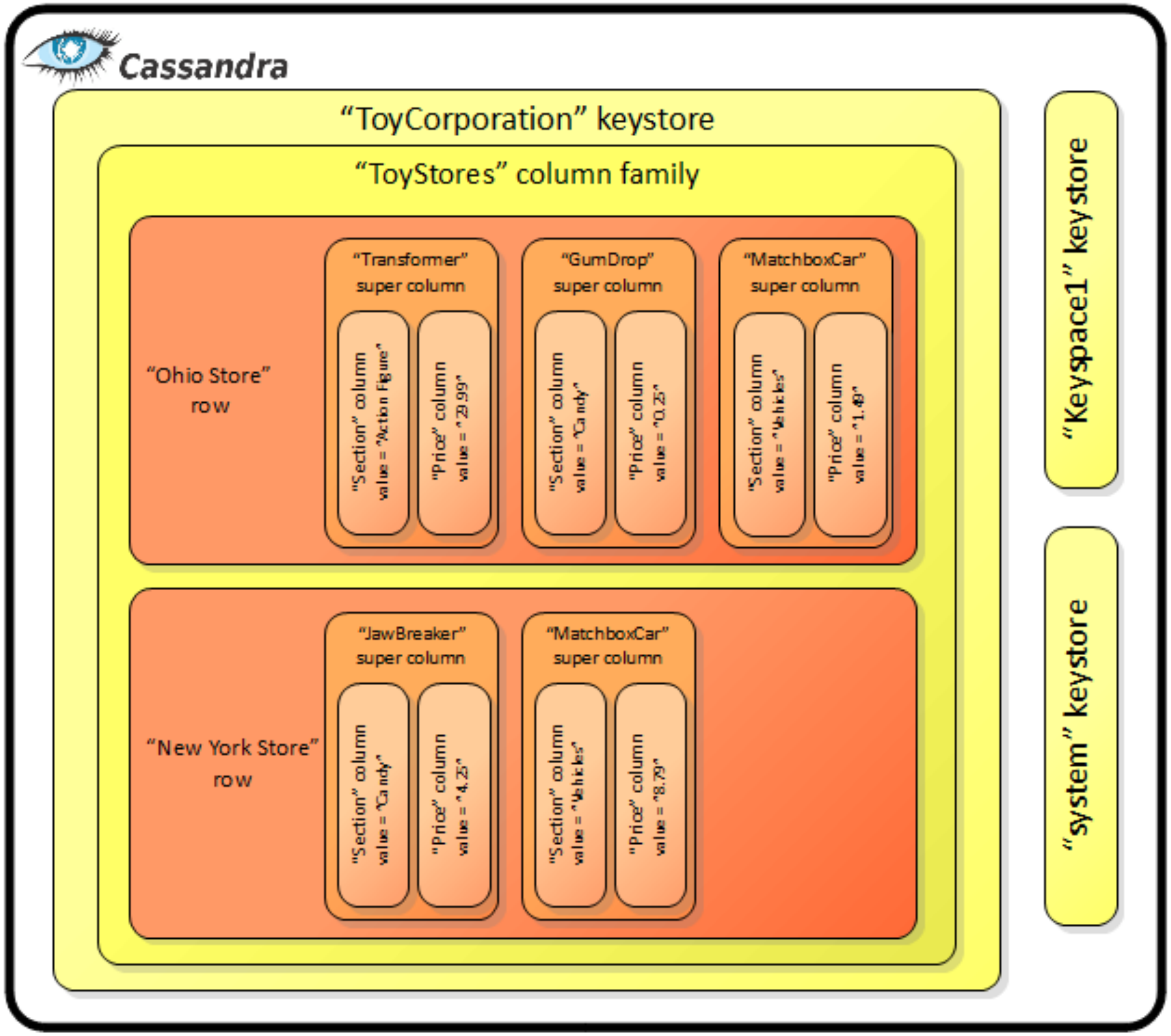

Cassandra
“ToyCorporation” keystore
“ToyStores” column family
“Ohio Store” row
“Transformer” super column
“Section” column value = “Action Figure”
“Price” column value = “29.99”
“GumDrop” super column
“Section” column value = “Candy”
“Price” column value = “0.25”
“MatchboxCar” super column
“Section” column value = “Vehicles”
“Price” column value = “1.49”
“New York Store” row
“JawBreaker” super column
“Section” column value = “Candy”
“Price” column value = “4.25”
“MatchboxCar” super column
“Section” column value = “Vehicles”
“Price” column value = “8.79”
“Keyspace1” keystore
“system” keystore

Image is Public Domain: Released July 11, 2010 by divconq.com

http://www.divconq.com/category/cassandra/page/2/

API

- insert (table, key, rowMutation)
- get (table, key, columnName)
- delete (table, key, columnName)

- columnName – (Path)
  - May refer to any column, column family, super column family, or a column within a super column

Read/Write Overview

- *Read and Write requests are processed by any node.
- The searching node determines which particular nodes contain the data.
- Writes
  - Issues write to all nodes, waits for a quorum of commits
- Reads (Variable)
  - Closest Node – Low integrity
  - All nodes + quorum – High integrity

Inbox Search Problem

- A user wants to search his or her inbox for messages using one of two strategies
  - Term Search - keyword
  - Interactions - name

Term Search

- Key – user ID
  - Super Column – (Inverted index)
    - CF - words that make up a message
      - C - Individual Message Identifiers

Search by Interactions

- Key – user ID
    - Super Column – (Inverted index)
        - CF – Recipient ID
            - C -Individual Message Identifiers

Inbox Search Problem

- Cassandra provides ‘hooks’ for intelligent caching:
  - e.g. user clicks on ‘inbox’, primes index
  - Production Performance numbers

Latency Stat	Search Interactions	Term Search
Min	7.69 ms	7.78 ms
Median	15.69 ms	18.27 ms
Max	26.13 ms	44.41 ms

Inbox Search Solution

- The full solution requires understanding of the underlying architecture.
- Key points are:
  - Any node can service a query
    - Global knowledge of data stores
  - Query all relevant data stores
  - Take most-recent, good response
    - Quorum, flexibility
    - Time Threshold

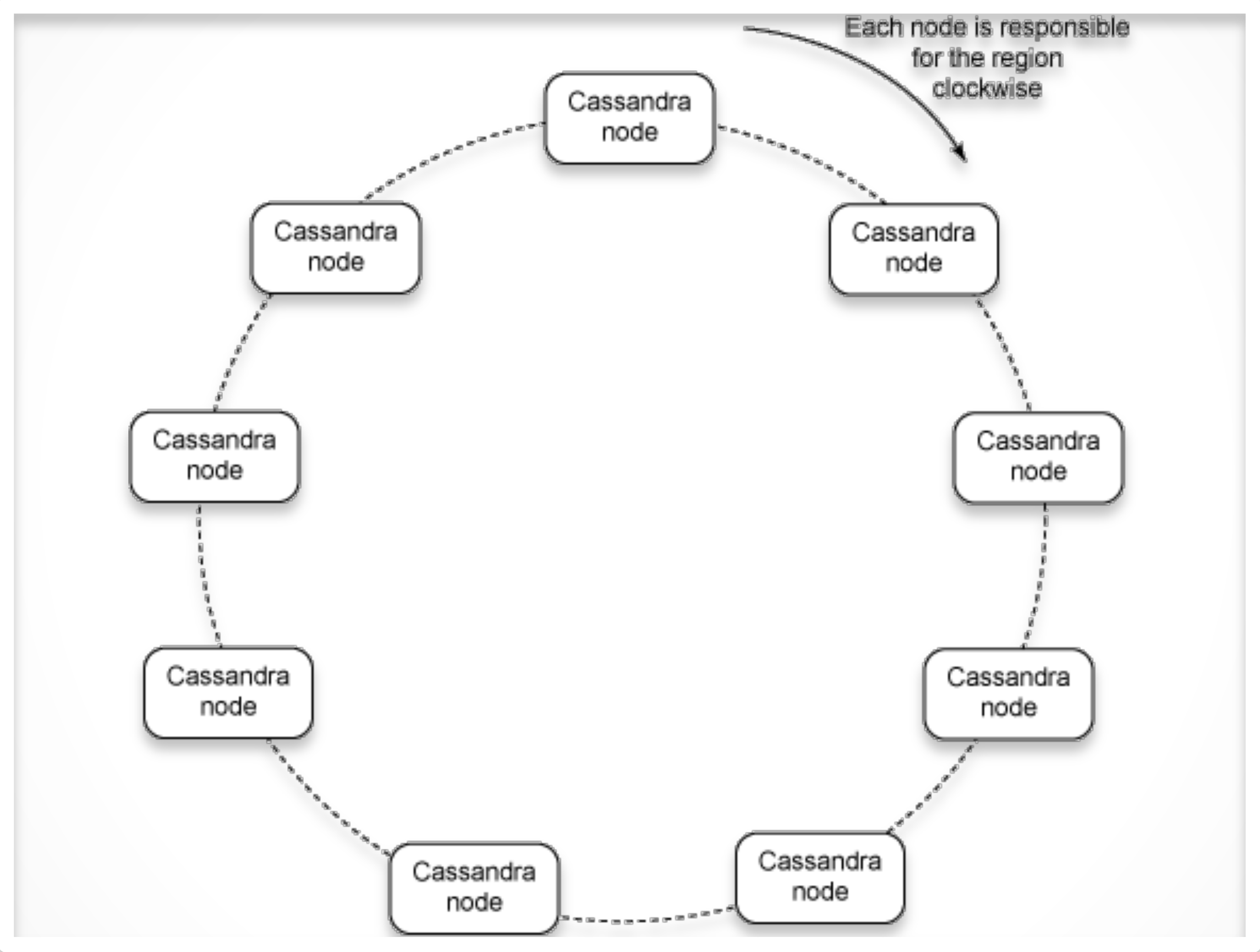
Questions

?

My Question

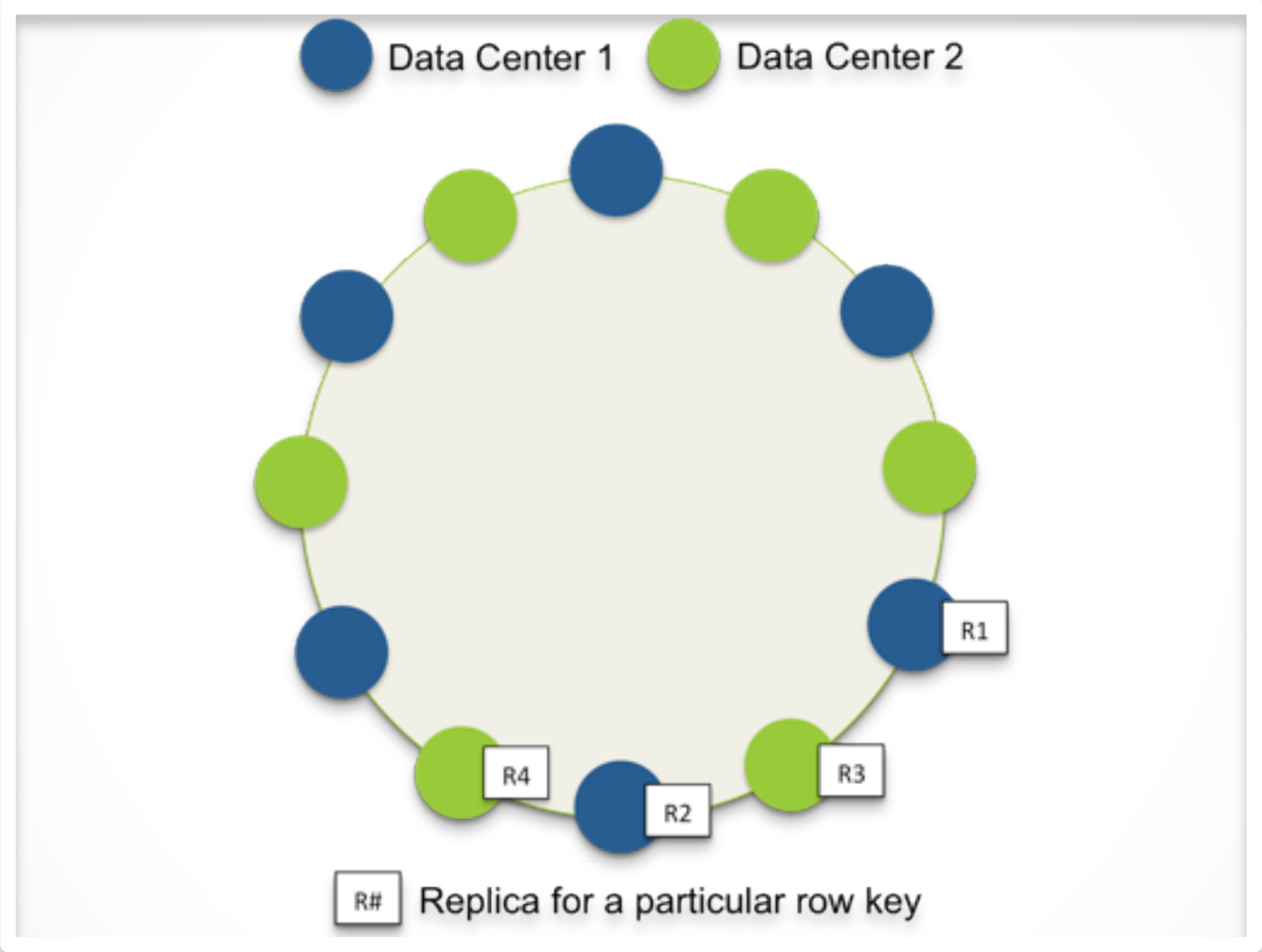
What should I emphasize in the architecture?

Cassandra Architecture
General Structure



http://www.ibm.com/developerworks/library/os-apache-cassandra/figure003.gif

Cassandra Architecture
General Structure

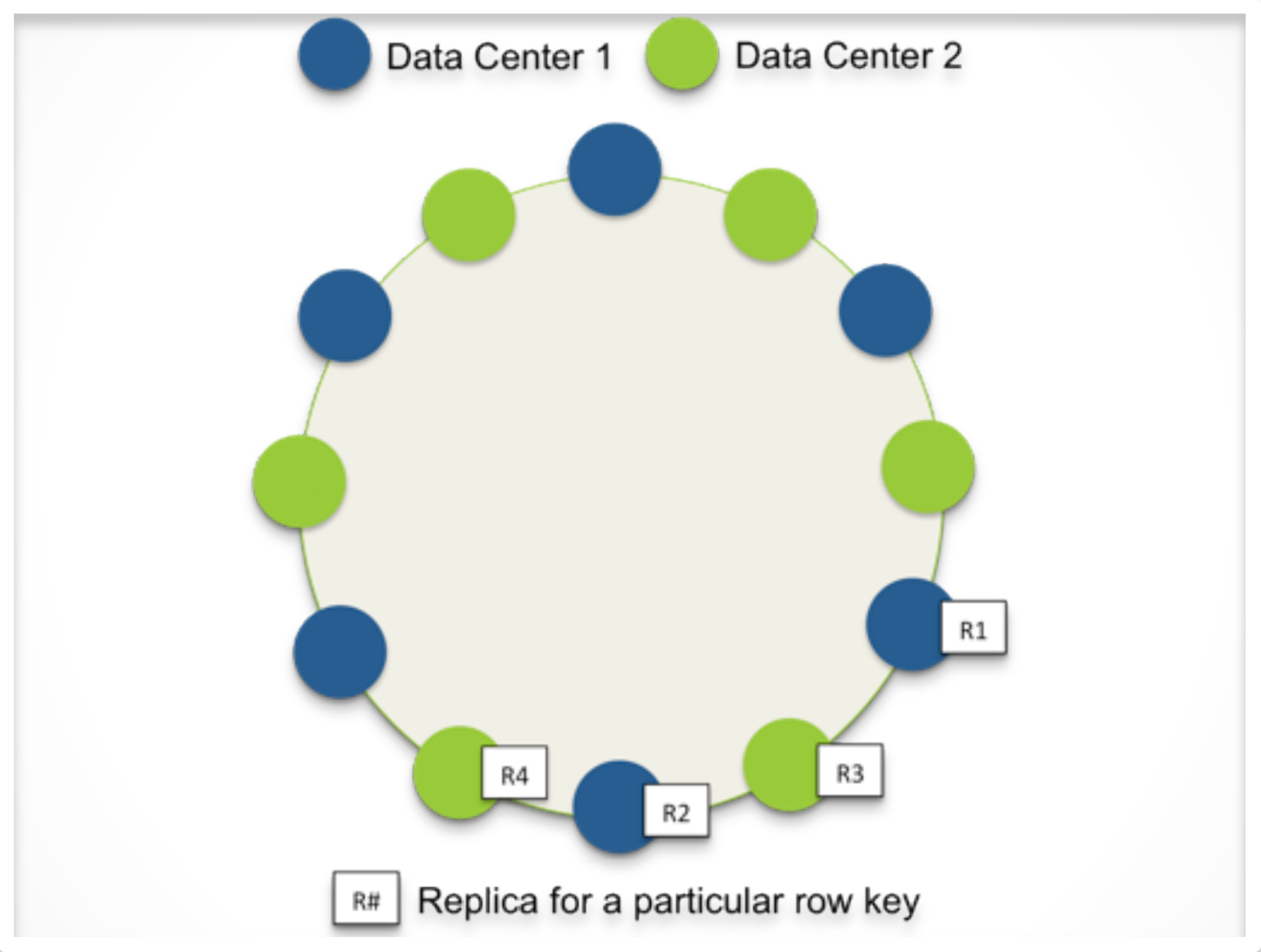


http://www.datastax.com/docs/1.0/cluster_architecture/replication

Cassandra Architecture Data Partitioning

- Consistent Hashing Algorithm
  - Logical Ring of hash values
  - Each node is given a position on this ring
  - Each node is responsible for a LEFT range of hashes.
  - Each data item's RIGHT neighbor is responsible for storage and replication
- Recall the nodes communicate about ranges so any one node knows the locations which **should** contain data for a particular key.

Cassandra Architecture
Data Partitioning

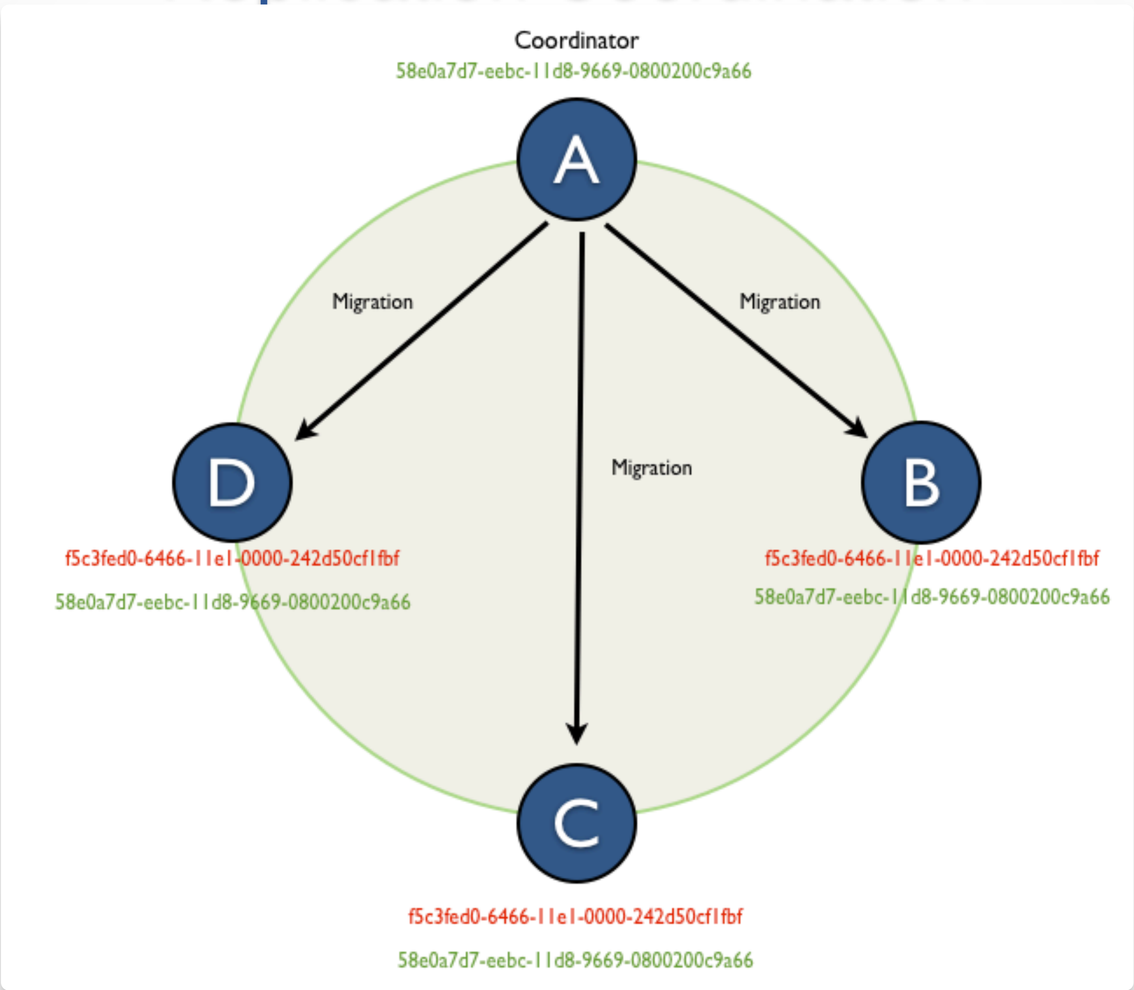


http://www.datastax.com/docs/1.0/cluster_architecture/replication

Cassandra Architecture
Replication Schemes

- Configurable:
  - Number of replicas
  - Replication Policies
    - Non-coordinator replicas are chosen by *picking* N-1 other nodes on ring.
      - Rack Unaware
    - Zookeeper (Elected leader) Abstraction
      - Rack Aware
      - Datacenter Aware

Replication Coordination


Coordinator
58e0a7d7-eebc-11d8-9669-0800200c9a66
A
Migration
Migration
Migration
D
f5c3fed0-6466-11e1-0000-242d50cf1fbf
58e0a7d7-eebc-11d8-9669-0800200c9a66
B
f5c3fed0-6466-11e1-0000-242d50cf1fbf
58e0a7d7-eebc-11d8-9669-0800200c9a66
C
f5c3fed0-6466-11e1-0000-242d50cf1fbf
58e0a7d7-eebc-11d8-9669-0800200c9a66

http://www.datastax.com/wp-content/uploads/2012/03/migration-distribution.png

Cassandra Architecture
Domain Awareness

- Each node has full awareness - handwaving
  - Cassandra nodes communicate via a Gossip Protocol
    - Based on “Scuttlebutt”

Cassandra Architecture
PHI Accrual Failure Detection

- Failure detection (prediction) via Gossip
  - A sliding window is used to calculate likelihood of failure, based on gossip:
  - (Phi, P(Failure))
  - { (1, 0.1) (2, 0.01) (3, 0.001) … }

Cassandra Architecture Failure Response

- System architecture does not reconfigure
  - The server will return eventually.
  - Recall: Scuttlebutt, sliding window and PHI
- Permanent modification of the ring is an administrator task.
- Phi represents a level of suspicion a particular node is down or unreachable
  - (*) This is used for timeout avoidance?
- (*) Protocols account for failure by design

Cassandra Architecture Request Handling

- Request arrives at any node:
  1. Servicing node looks up the data hosts.
     1. All nodes have knowledge of the data partition
  2. (*) Request is routed to all data hosts.
  3. If requests timeout, failure is returned
  4. **Identify the response containing data with the youngest timestamp, return it.**
  5. **Schedule update of nodes with older timestamps. – This ensures quorum integrity**

Cassandra Architecture
Local Data Persistence

Typical Write:

1. Write to commit log **– Dedicated Disk**
2. Update to in-memory structure

When in-memory structure crosses threshold
(data size, number of data items)

Data is sequentially written to commodity disks.

Over time these files are merged and reorganized
ala BigTable

Cassandra Architecture Local Data Persistence

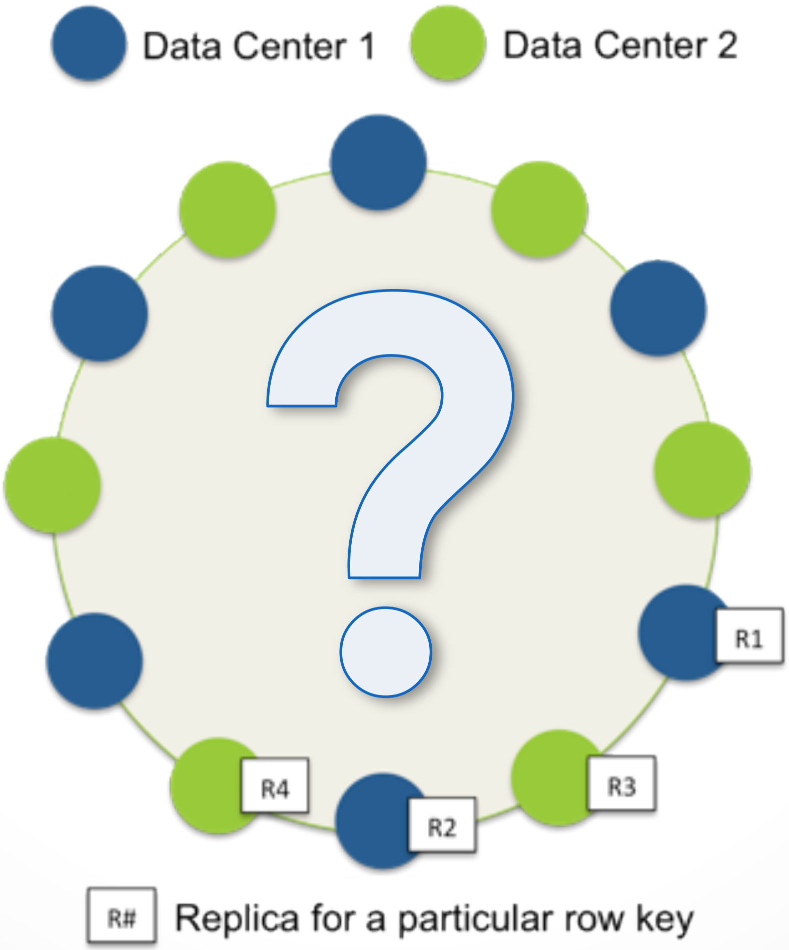
Typical Read: (A key can be in many files)

1. Bloom filter (index of keys in each file)
2. Get Values (reverse chronological order)
    1. Values (CF) have Column Indices allowing for direct access of columns.

I’ve left out much of the *chunking* details.

Optimizations, included at almost every conceivable point, are our of scope.

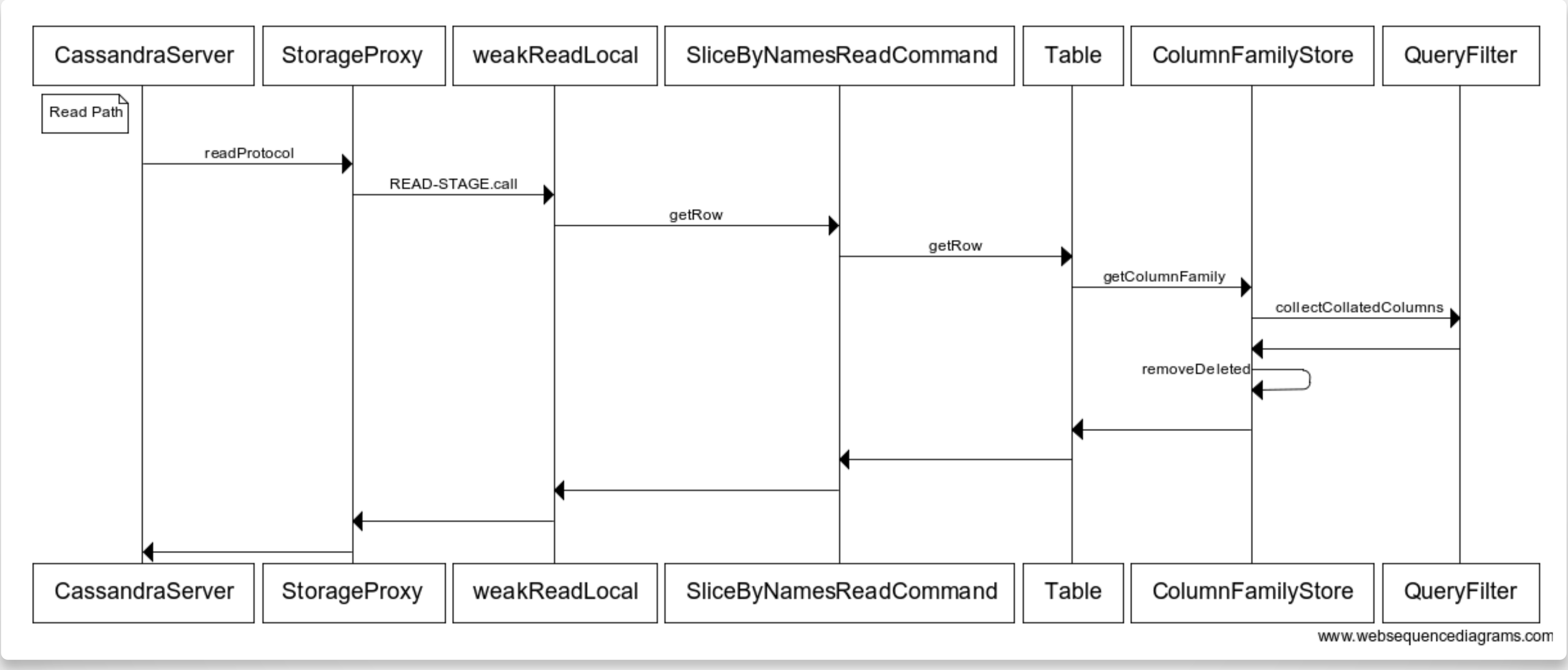
Questions



Data Center 1 Data Center 2

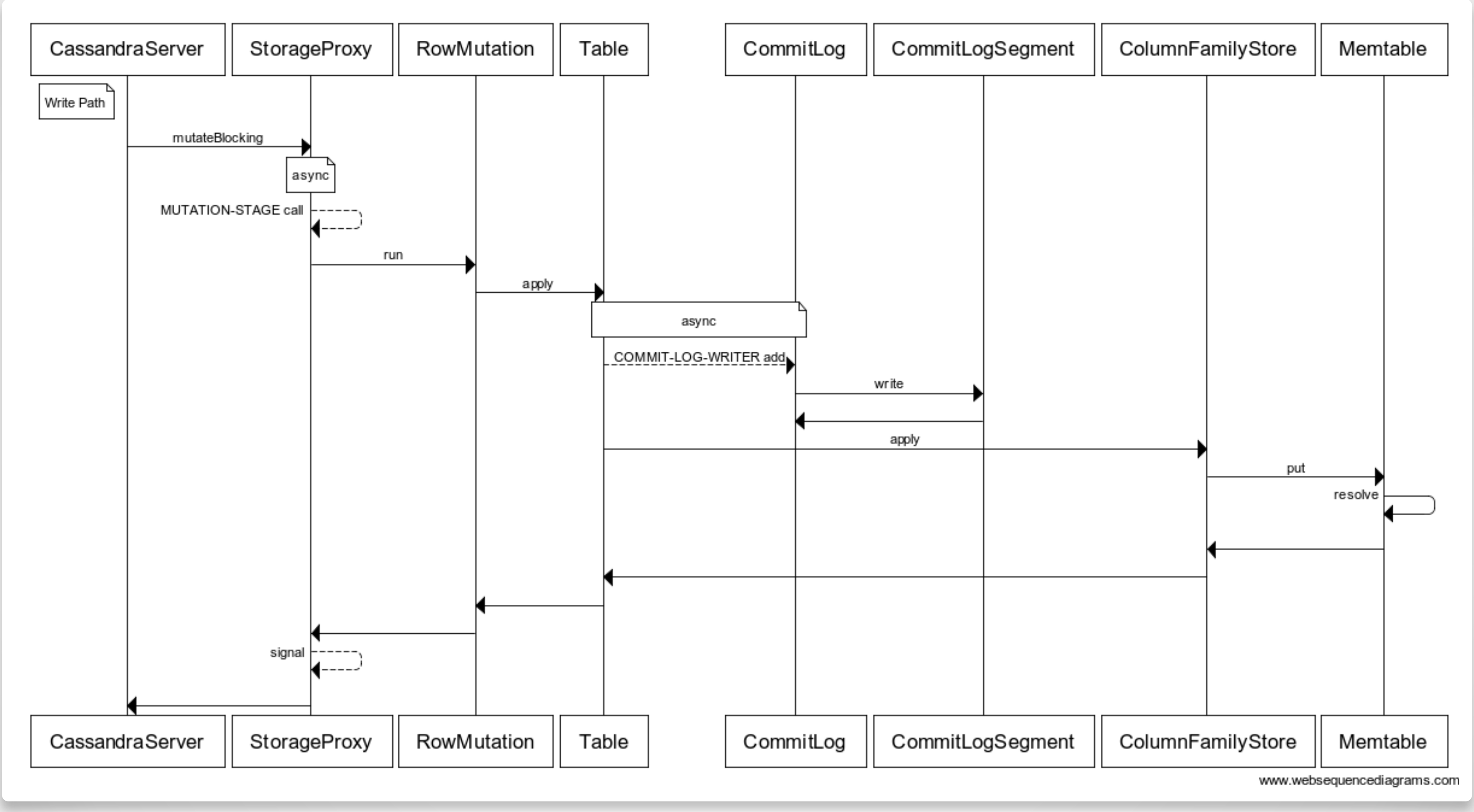
R# Replica for a particular row key

Read Path



http://prettyprint.me/prettyprint.me/2010/05/02/understanding-cassandra-code-base/index.html

Write Path



http://prettyprint.me/prettyprint.me/2010/05/02/understanding-cassandra-code-base/index.html

Questions

- Additional Information:
  - http://www.slideshare.net/DataStax/an-overview-of-apache-cassandra