Hadoop and Cassandra

sitting in a tree...

Jake Luciani (@tjake)
Strange Loop 2011

K.I.S.S.I.N.G!?

Keeping

Integration

Simple

So

I

Never

Get-Woken-At-4am

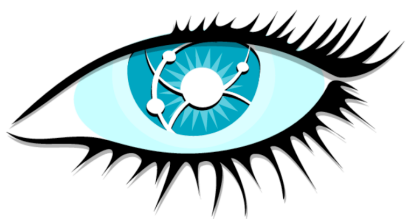
What makes Cassandra great?

Dynamo based

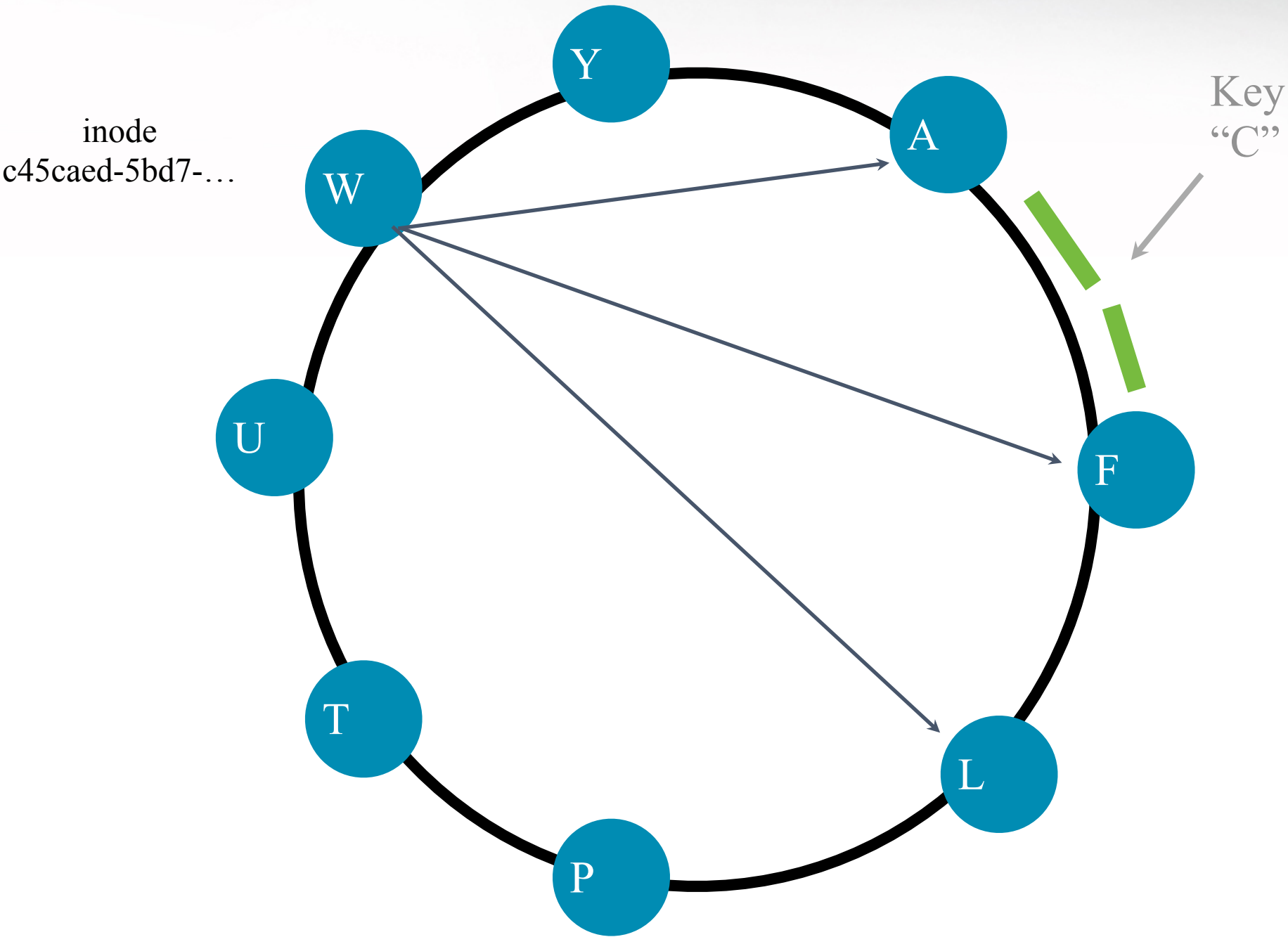
- Masterless
- Homogeneous
- Failure and Recovery at it's core
- Tunable consistency guarantees per read/write

BigTable based

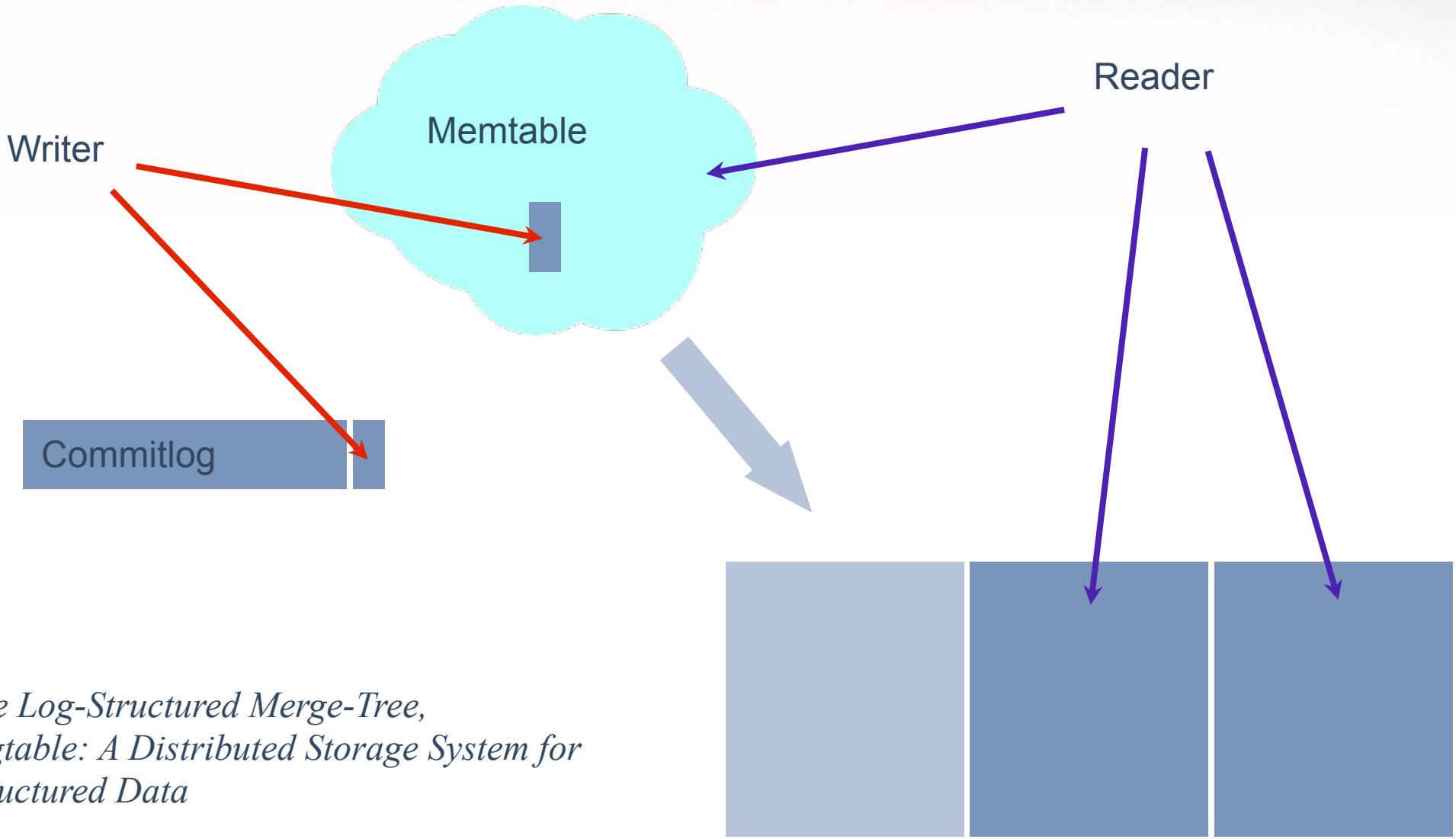
- Powerful data model
- Low Latency
- OLTP



P2P replication (Dynamo)



Locally-managed data (BigTable)



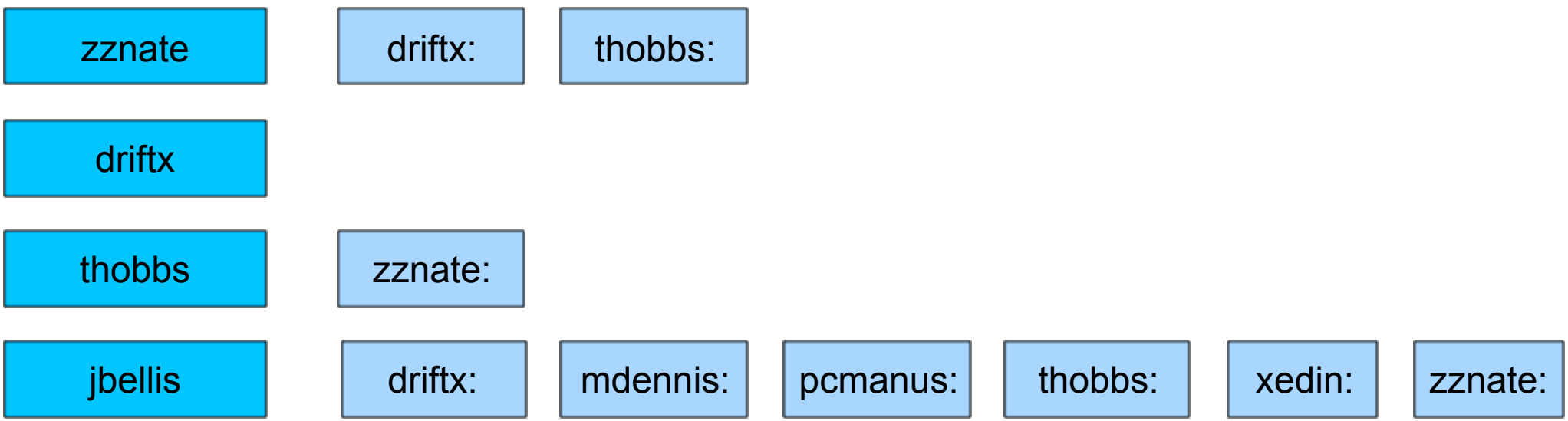
The Log-Structured Merge-Tree,
Bigtable: A Distributed Storage System for Structured Data

Data Model (BigTable)

- ColumnFamilies contain rows + columns
- Sparse model: no rewriting for ALTER
- (Not really schemaless for a while now)

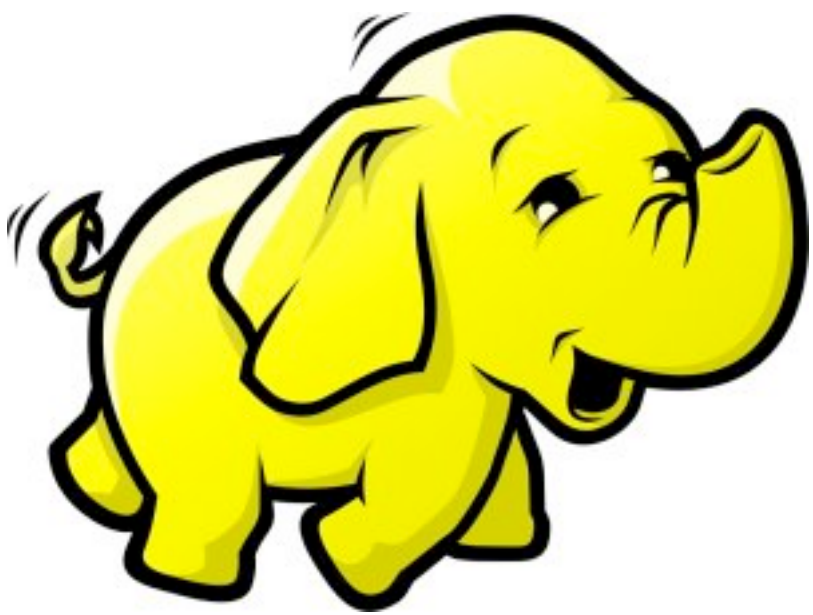
Wide rows (BigTable)

- Sparse model + sstable storage means we can have arbitrarily large rows
- Columns are sorted by comparator



What makes Hadoop great?

- MapReduce Framework
- Pig
- Hive
- OLAP



How can I use Cassandra with Hadoop?

- ColumnFamilyInputFormat
- ColumnFamilyOutputFormat
- Pig Driver
- Hive Driver

- Run TaskTrackers on Cassandra Nodes
  - Data Locality

How do I run Hadoop jobs on data created in my application?

- *Run Hadoop on all nodes and access the data live?*

How do I run Hadoop jobs on data created in my application?

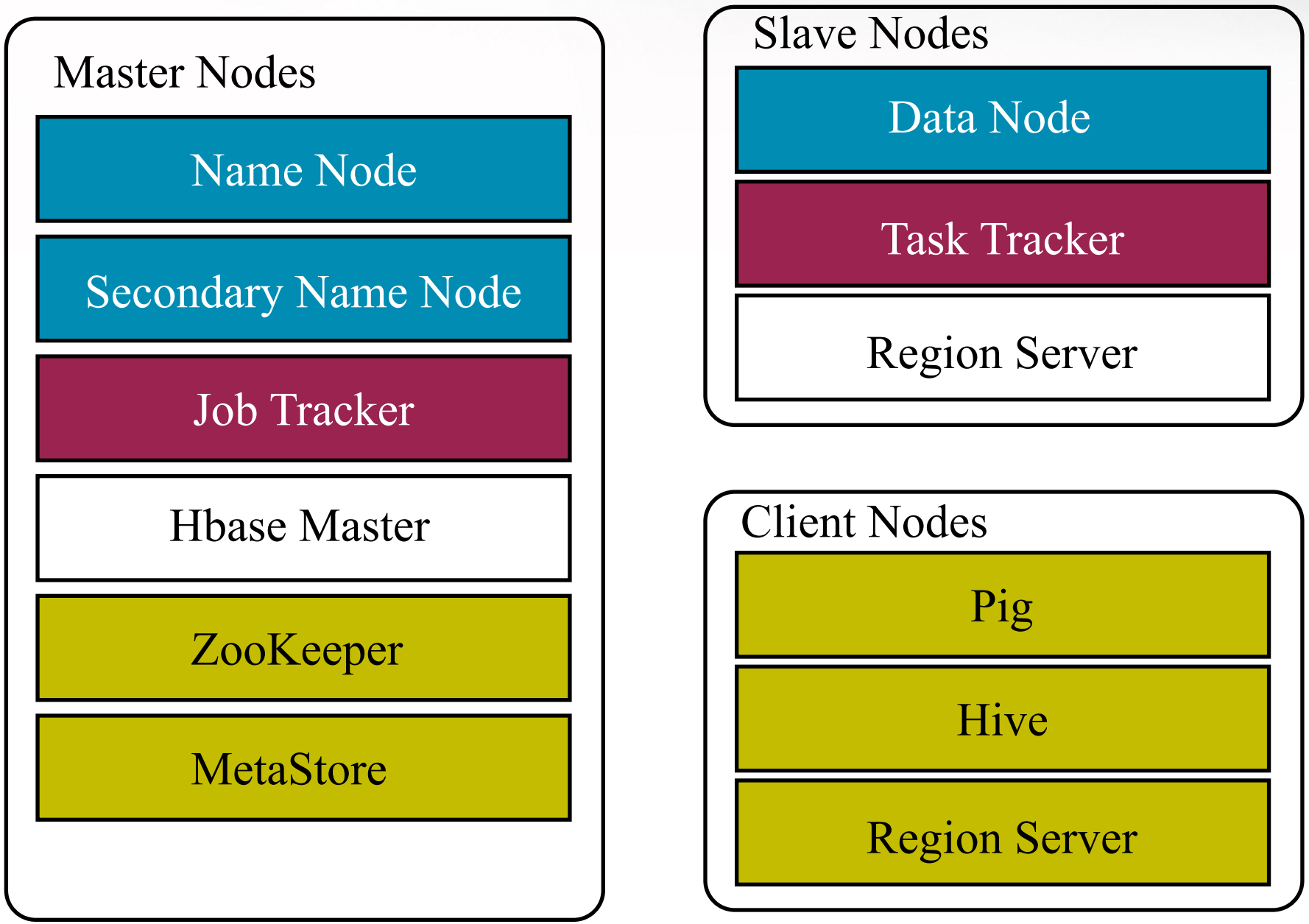
- *Run Hadoop on all nodes and access the data live?* **No. This will kill your OLTP performance.**

- *Use Scribe, Flume or Sqoop to put the data into HDFS, run my Hadoop jobs then load the data back into yet another system to consume the data?!*

How do I run Hadoop jobs on data created in my application?

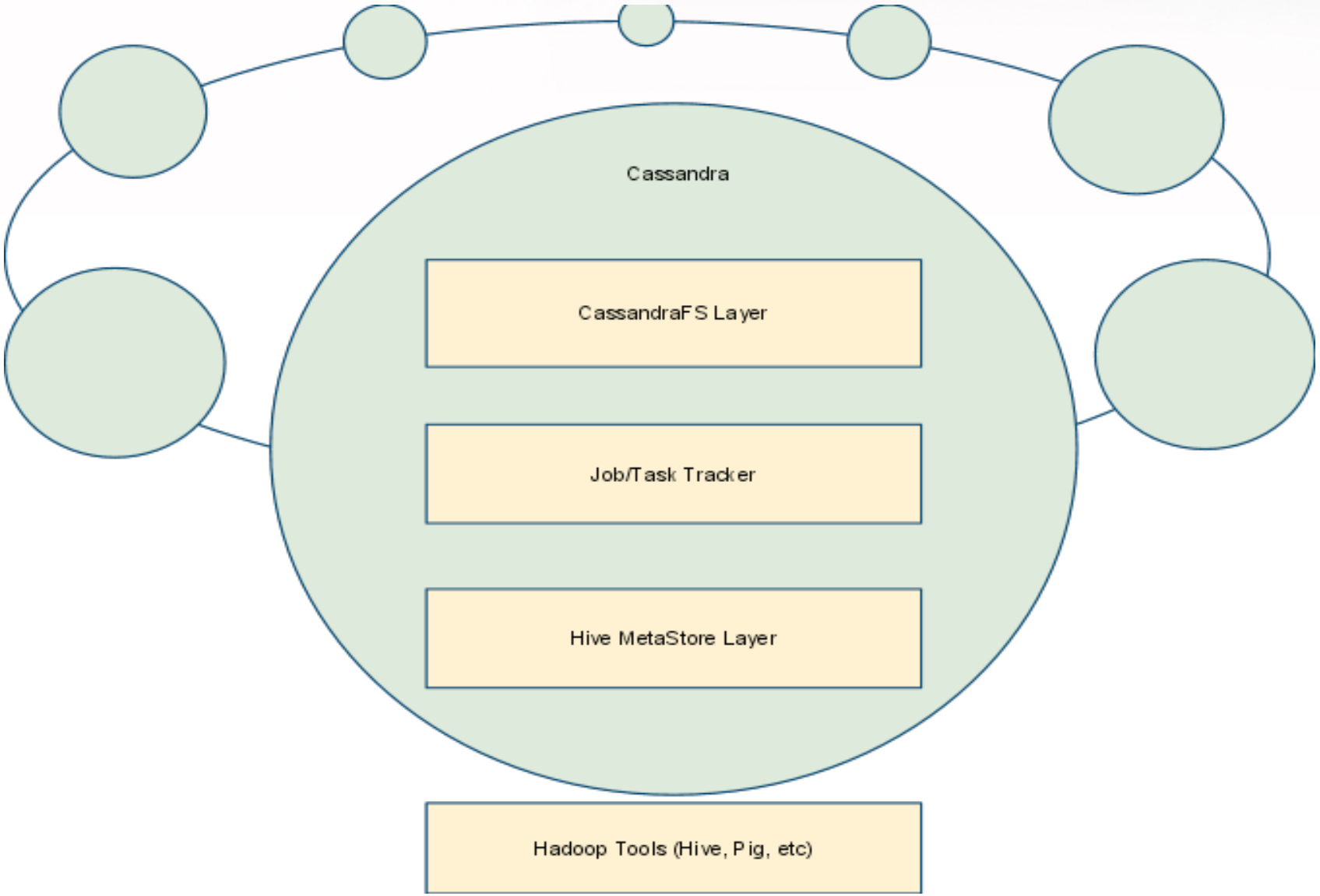
- *Run Hadoop on all nodes and access the data live?* **No. This will kill your OLTP performance.**
- *Use Scribe, Flume or Sqoop to put the data into HDFS, run my Hadoop jobs then load the data back into yet another system to consume the data?!* **Maybe. But there must be a better way.**
- Replicate the data from my live cluster to a secondary cluster, run Hadoop then replicate results back to the live cluster? **Yes!**

The Traditional Hadoop Stack




Stop
adding new
services to monitor!
I want my
name-node back!
Your
stacktraces
are novels!

The Brisk Hadoop Ring

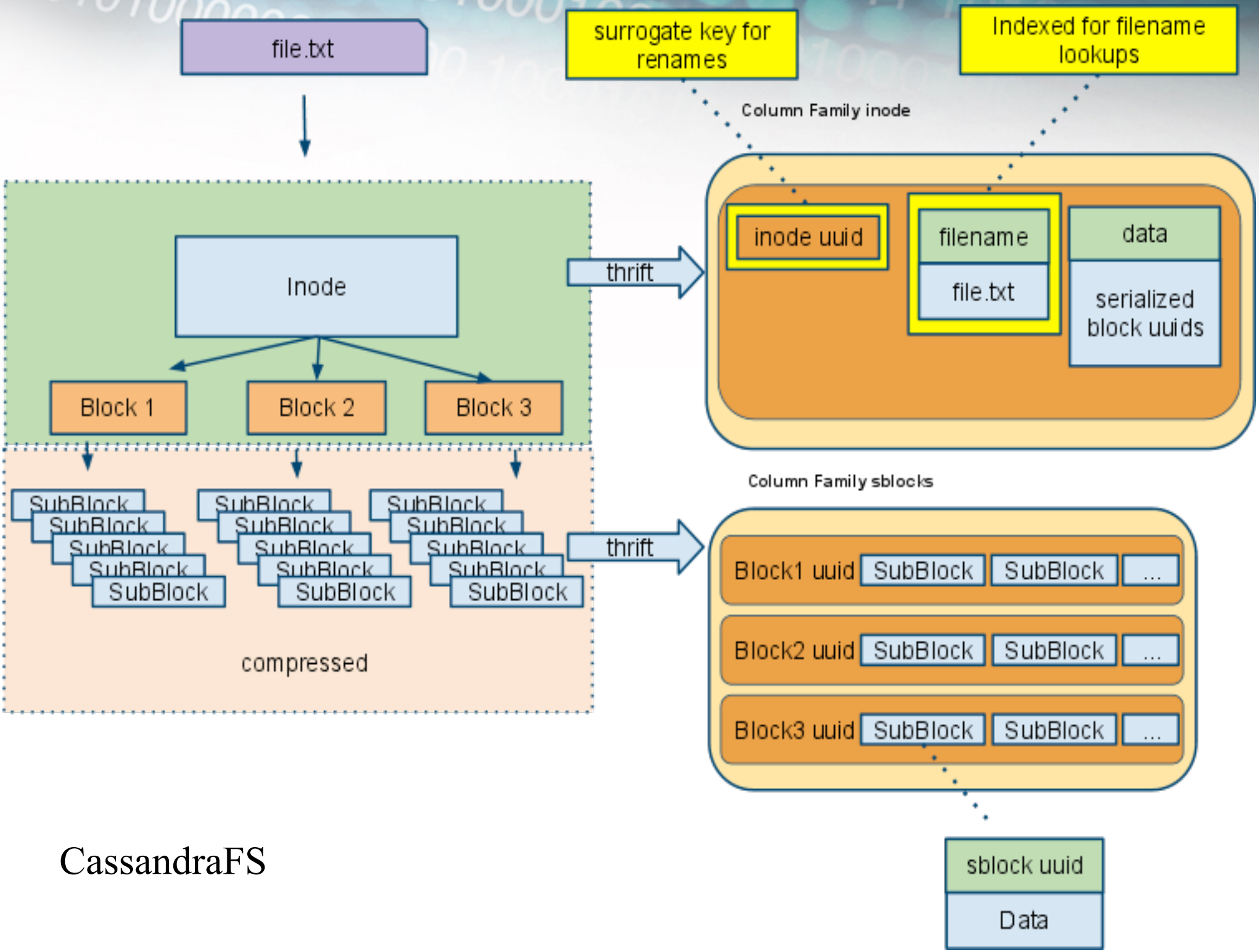


Brisk

- Easy to deploy and operate
- No single points of failure
- Scale and change nodes with no downtime
- Cross-DC, multi-master clusters
- Allocate resources for OLAP vs OLTP
  - With no manual ETL
- Based on Cassandra 0.8 and Hadoop 0.20.203

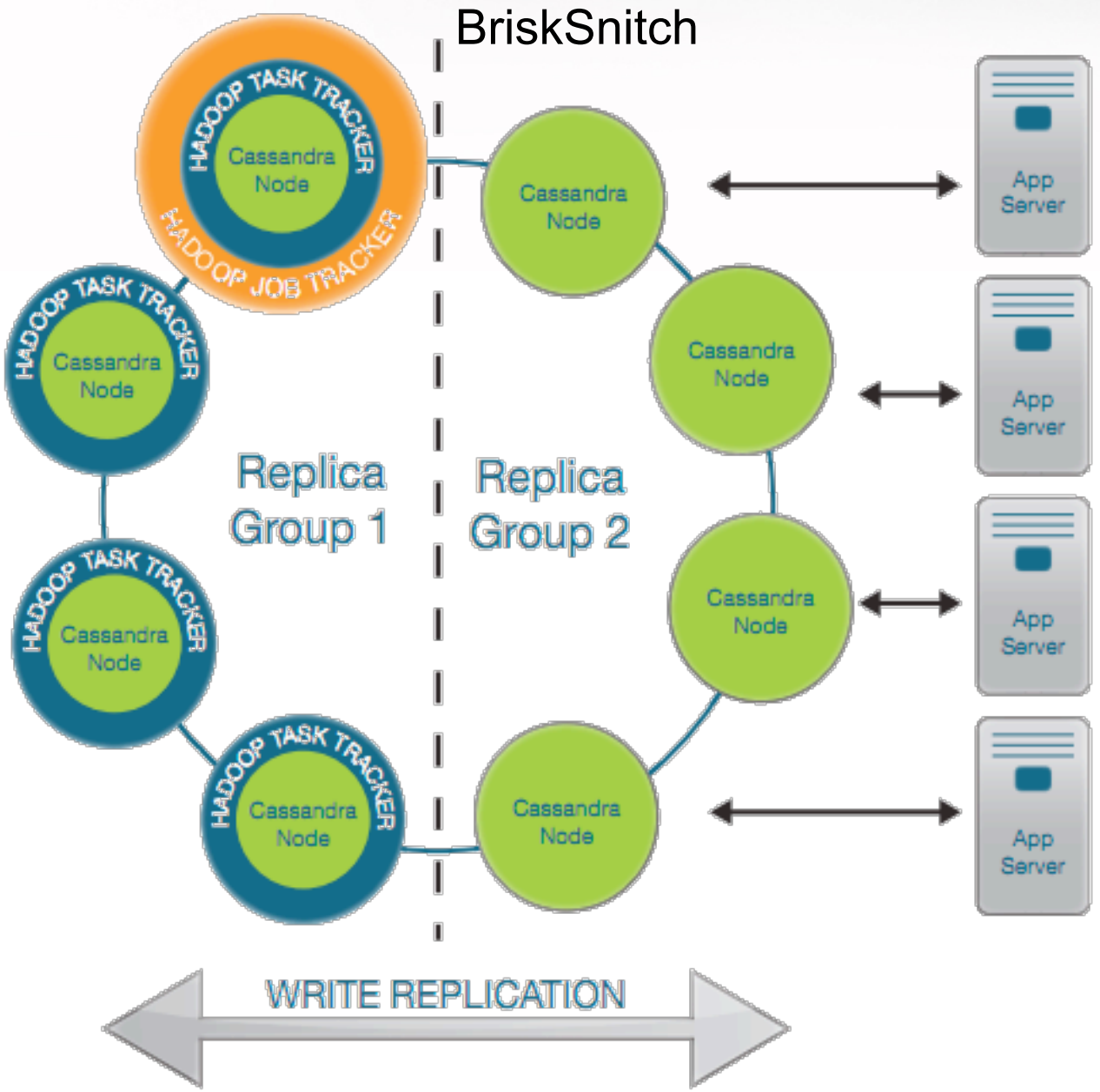
CassandraFS (HDFS Replacement)

- Built on ColumnFamilies
  - inode and sblocks
- Data stored as ByteBuffer internally — excellent fit for HDFS blocks
- Local reads mmap data directly (**no rpc**)
- Blocks are compressed with google snappy
- hadoop distcp hdfs:///mydata cfs:///mydata



CassandraFS

Hybrid Workloads



Hive: CFS and ColumnFamilies

CFS:

```
CREATE TABLE users (name STRING, zip INT);
LOAD DATA LOCAL INPATH 'kv2.txt' OVERWRITE INTO TABLE users;
select count(*), zip from invites group by bar;
```

Column Family (fixed):

```
CREATE EXTERNAL TABLE Keyspace1.Users(name STRING, zip INT)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler';
```

Column Family (dynamic):

```
CREATE EXTERNAL TABLE Keyspace1.Users
(row_key STRING, column_name STRING, value string)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler';
```

Pig: CFS and ColumnFamilies

CFS:

```
grunt> data = LOAD 'cfs:///example.txt' using PigStorage() as
(name:chararray, value:long);
```

ColumnFamily:

```
data = LOAD 'cassandra://Demo1/Scores' using
CassandraStorage() AS (key, columns: {T: tuple(name, value)});
```

ColumnFamily Slices:

```
data = LOAD 'cassandra://Demo1/
Scores&slice_start=M&slice_end=S' using CassandraStorage() AS
(key, columns: {T: tuple(name, value)});
```

Portfolio Manager Demo

- **Low-Latency Side** (OLTP)
  - Live tick prices for NASDAQ stocks
  - Thousands of Portfolios
- **Batch Analytics Side** (OLAP)
  - Historical Stock prices
  - Historical VAR (Value At Risk) calculation for each portfolio using Hive
  - VAR number is written back to Cassandra from Hive

Portfolio Demo ColumnFamilies

Portfolios

Portfolio1	GOOG	LNKD	P	AMZN	AAPL
	5	7	50	100	4

LiveStocks

AAPL	LAST
	$950.00

StockHist

GOOG	2011-01-01	2011-01-02	2011-01-03	2011-01-04
	$2.99	$1.00	$9.29	$29.10

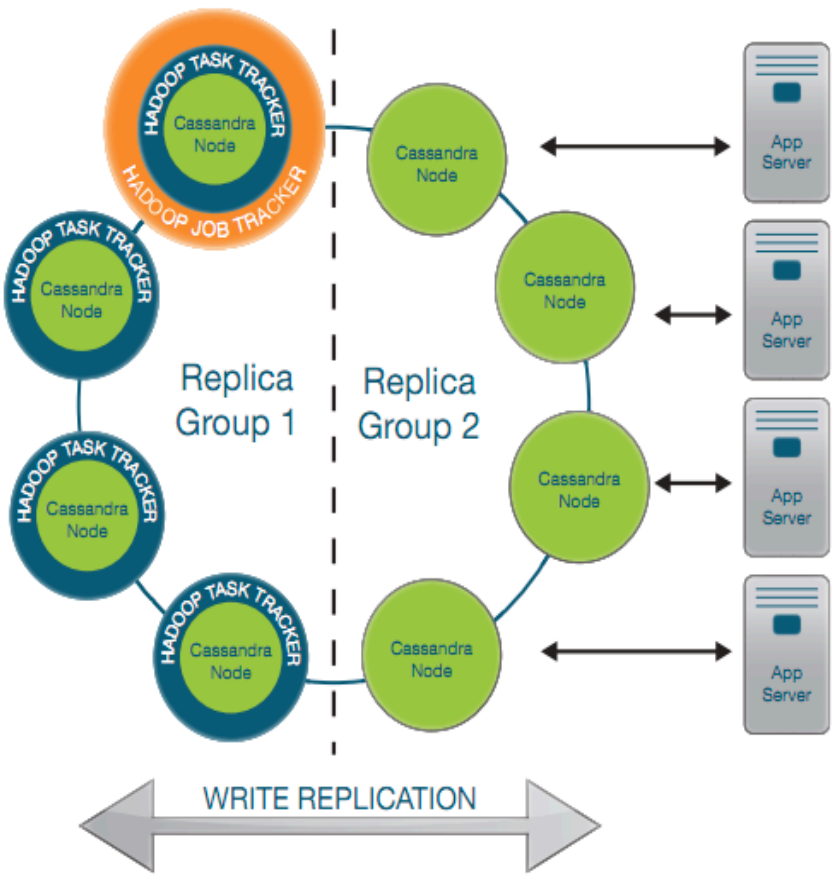
Portfolio Demo Workloads

OLAP:

- Hive <- Cassandra
  - Portfolios
  - Historical Prices
- Intermediate Results
- Hive -> Cassandra
- Every N minutes

OLTP:

- WebBased Portfolios
- Live Prices for today
- Historical Prices



Portfolio Manager Demo

Portfolio #5329 $13778.89 (+55%)

Largest Historical 10day Loss:

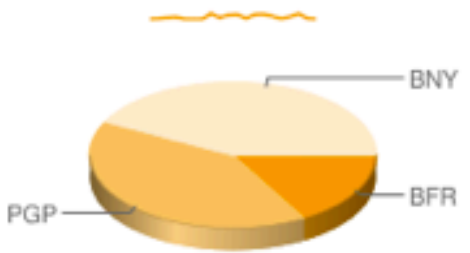
?

Portfolio #1970 $4325.45 (-59%)

Largest Historical 10day Loss:

?

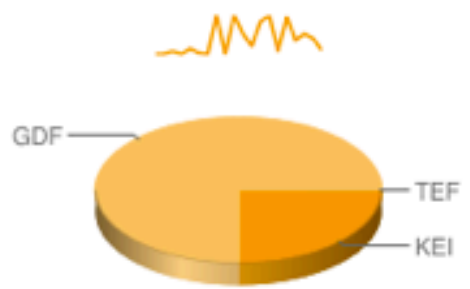
Portfolio #9700 $3526.94 (-85%)



Largest Historical 10day Loss:

?

Portfolio #4612 $796.03 (+98%)



Largest Historical 10day Loss:

?

Portfolio #6095 $721.58 (-33%)

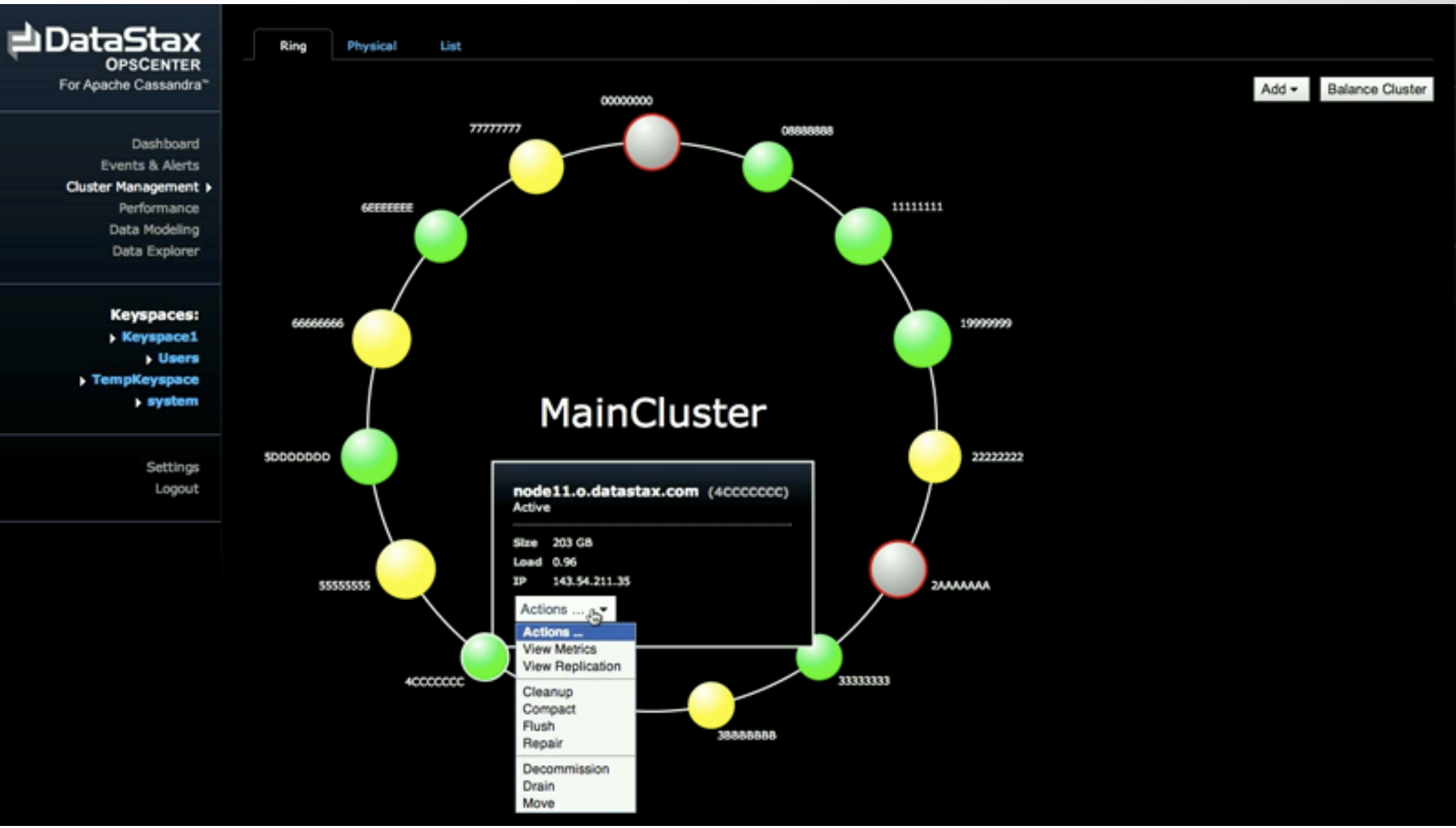
Largest Historical 10day Loss:

?

Portfolio #1126 $9193.65 (-82%)

Largest Historical 10day Loss:

?


DataStax
OPSCENTER
For Apache Cassandra™
Dashboard
Events & Alerts
Cluster Management
Performance
Data Modeling
Data Explorer
Keyspaces:
Keyspace1
Users
TempKeyspace
system
Settings
Logout
Ring
Physical
List
Add
Balance Cluster
00000000
08888888
11111111
19999999
22222222
2AAAAAAA
33333333
38888888
4CCCCCCC
55555555
5DDDDDDD
66666666
6EEEEEEE
77777777
MainCluster
node11.o.datastax.com (4CCCCCCC)
Active
Size 203 GB
Load 0.96
IP 143.54.211.35
Actions ...
Actions ...
View Metrics
View Replication
Cleanup
Compact
Flush
Repair
Decommission
Drain
Move

Job Tracker

View Full Details

	Job	Progress	Started ▼	Duration	User
	TeraSort	33% Maps: 1/2 Reduces: 0/1	5/12/11 12:27 PM	40s	root
	TeraSort	100% Maps: 2/2 Reduces: 1/1	5/12/11 12:26 PM	34s	root
	TeraGen	100% Maps: 2/2 Reduces: 0/0	5/12/11 12:25 PM	44s	root
	PiEstimator	100% Maps: 10/10 Reduces: 1/1	5/12/11 12:19 PM	49s	riptano
	select count (*) from server_request(Stage-1	100% Maps: 15/15 Reduces: 1/1	5/12/11 10:56 AM	1m 31s	riptano
	select count(*) size, method from myT...size(Stage-2)	100% Maps: 10/10 Reduces: 10/10	5/11/11 8:30 PM	57s	riptano
	select count(*) size, method from myT...size(Stage-1)	100% Maps: 2/2 Reduces: 10/10	5/11/11 8:29 PM	50s	riptano
	select count(*), method from myTest...method(Stage	100% Maps: 2/2 Reduces: 10/10	5/11/11 8:24 PM	46s	riptano
	select count(*), method from myTest...method(Stage	100% Maps: 2/2 Reduces: 10/10	5/11/11 8:12 PM	49s	riptano
	select count(*), method from myTest...method(Stage	100% Maps: 2/2 Reduces: 1/1	5/11/11 8:04 PM	21s	riptano
	select min(duration) from myTest(Stage-1)	100% Maps: 2/2 Reduces: 1/1	5/11/11 8:02 PM	24s	riptano
	select max(duration) from myTest(Stage-1)	100% Maps: 2/2 Reduces: 1/1	5/11/11 8:02 PM	24s	riptano
	select count(*) from myTest(Stage-1)	100% Maps: 2/2 Reduces: 1/1	5/11/11 8:00 PM	25s	riptano

Open Source

- https://github.com/riptano/brisk
- Apache License

Thank You

Jake Luciani

jake@datastax.com

@tjake