Cassandra Persistence for Online Systems

What Actually Works

John Sumsion
FamilySearch

Introduction

- Cassandra for online systems
- Introduction to Family Tree
- Event-sourced persistence model
- Surprises & Solutions

Cassandra for Online Systems

- KillrVideo from Datastax Academy
- Classic use cases (from 2014)
  - Product Catalog / Playlist
  - Recommendation Engine
  - Sensor Data/IOT
  - Messaging
  - Fraud Detection

https://www.datastax.com/2014/06/what-are-people-using-cassandra-for

Cassandra for Online Systems

- CQL-based schemas (record & fields)
- Blob-based schemas (JSON inside blob)
- Time-series schemas (sensor data)
- Event-sourced schemas (events & views)
- Restrictions:
  - No joins
  - No transactions
  - General-purpose Indexes & Materialized Views newly available if using Cassandra 3

Cassandra for Online Systems

Keys for schema design:

1. Denormalize at write time for queries
2. Keep denormalized copies in sync at edit time
3. Avoid schemas that cause many, frequent edits on the same record
4. Avoid schemas that cause edit contention
5. Avoid inconsistency from read-before-write

Cassandra for Online Systems

What we did that worked:

1. Event sourced schema with multiple views
2. Event denormalization, with consistency checks
3. Flexible schema (JSON in blob)
4. Limits and throttling to deal with hotspots

- Details follow for Family Tree

Family Tree - Introduction

- Family Tree for the entire human family
  - 1.2B persons
  - 800M relationships
  - 7.8M registered users
  - 3.8M Family Tree contributors
- Free registration, Open Edit
- Supported by growing record collection
- World-wide user base
- Backed by Apache Cassandra (DSE)

James Mead Sumsion (1873–

Secure https://familysearch.org/tree/pedigree/KWZ3-PW9/landscape

Messages Volunteer Get Help

FamilySearch

Family Tree Search Memories Indexing Temple

John Sumsion

TREE PERSON FIND LISTS WHAT'S NEW

Help Others

Landscape

Print | Show

+
–

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 9 March 1926
Salt Lake City, Salt Lake, Utah...
Lela Coplan
1895–1983 • KWZ3-PW3
Children

George Sumsion
1842–1907 • KWJ8-4W5
Marriage: 23 Jan 1862
Salt Lake City, Salt Lake, Utah
Ann Elizabeth Bird
1844–1922 • KWVP-QXP
Children

Albert Coplan
1866–1933 • KW8Q-TD8
Marriage: 27 Jan 1893
Sanpete, Utah, United States
Emma Charlotte Jensen
1870–1954 • KW8Q-TDD
Children

Daniel Sumsion
1810–1882 • KWV3-BX7
Marriage: 1837
Colerne, Wilts, Eng.
Eleanor Spender
1816–1880 • KWV3-BXW
Children

John Bird
1806–1855 • LR6L-RRX
Marriage: 10 May 1830
Saint Martin Church, Birmingh...
Ann Russen
1808–1871 • KWJF-QL9
Children

Willis Coplan
1816–1899 • KWJZ-1G2
Marriage: 14 Jul 1859
Salt Lake City, Salt Lake, Utah
Amelia Andrew Angier
1837–1915 • KWJZ-1GK
Children

Peter Jens Jensen
1822–1893 • KWNV-777
Marriage: 12 September 1863
Salt Lake City, St-Lk, Ut
Christina Charlotte Oman
1844–1921 • KWNV-77Q
Children

John Bird
1771–1833 • L691-697
Marriage: 15 October 1799
Birmingham St Martin, Warwic...
Elizabeth Pittaway
1780–1860 • LHZZ-XMB
Children

Thomas Russen
1766–1820 • LKKM-BT4
Marriage: 26 Oct 1796
Dodderhill, Worcestershire, En...
Sarah Tongue
1770–1862 • LKKM-BTS
Children

John Bird
1735–1807
Marriage:
Bickenhill,
Mary Bate
1736–1809
Children

Edward Pi
1721–1831
Marriage:
Solihull, W
Jane Smit
1759–1811
Children

Adam Rus
1742–1802
Marriage:
Upton Wan
Sarah Dar
1739–1811
Children

William T
1737–1784
Marriage:
Alvechurch
Mary Grif
1740–1818
Children

Family Tree - Introduction

- Multiple views of person
  - Pedigree page
  - Person page
  - Person card popup
  - Person change history
  - Descendancy page

Pedigree Page

James Mead Sumsion (1873–

Secure https://familysearch.org/tree/pedigree/KWZ3-PW9/lar

FamilySearch

Messages Volunteer Get Help

Family Tree Search Memories Indexing Temple

John Sumsion

TREE PERSON FIND LISTS WHAT'S NEW

Help Others

Landscape

Print | Show

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 9 March 1926
Salt Lake City, Salt Lake, Utah...
Lela Coplan
1895–1983 • KWZ3-PW3
Children

George Sumsion
1842–1907 • KWJ8-4W5
Marriage: 23 Jan 1862
Salt Lake City, Salt Lake, Utah
Ann Elizabeth Bird
1844–1922 • KWVP-QXP
Children

Albert Coplan
1866–1933 • KW8Q-TD8
Marriage: 27 Jan 1893
Sanpete, Utah, United States
Emma Charlotte Jensen
1870–1954 • KW8Q-TDD
Children

Daniel Sumsion
1810–1882 • KWV3-BX7
Marriage: 1837
Colerne, Wilts, Eng.
Eleanor Spender
1816–1880 • KWV3-BXW
Children

John Bird
1806–1855 • LR6L-RRX
Marriage: 10 May 1830
Saint Martin Church, Birmingh...
Ann Russen
1808–1871 • KWJF-QL9
Children

Willis Coplan
1816–1899 • KWJZ-1G2
Marriage: 14 Jul 1859
Salt Lake City, Salt Lake, Utah
Amelia Andrew Angier
1837–1915 • KWJZ-1GK
Children

Peter Jens Jensen
1822–1893 • KWNV-777
Marriage: 12 September 1863
Salt Lake City, St-Lk, Ut
Christina Charlotte Oman
1844–1921 • KWNV-77Q
Children

John Bird
1771–1833 • L691-697
Children

Thomas Russen
1766–1820 • LKKM-BT4
Marriage: 26 Oct 1796
Dodderhill, Worcestershire, En...
Sarah Tongue
1770–1862 • LKKM-BTS
Children

John Bird
1735–1807
Marriage:
Bickenhill,
Mary Bate
1736–1809
Children

Sarah Dar
1739–1811
Children

William T
1737–1784
Marriage:
Alvechurch
Mary Grif
1740–1818
Children

33 persons (plus children)

33 relationships (w/ details)

1 page view

Person Page (top)

FamilySearch

Messages Volunteer Get Help ▾

John Sumsion ▾

TREE ▾ PERSON ▾ FIND LIS

Help Others

James Mead Sumsion

27 April 1873 – 29 June 1958 • KWZ3-PW9

View Tree Unwatch View My Relationship

Details Memories 7 Ordinances

Vitals | Other | Family | Sources 22 | Discussions 0 | Notes 0

Life Sketch

Add

Vital Information

Open Details

Name
James Mead Sumsion

Sex
Male

Birth
27 April 1873
Springville, Utah, Utah, United States

Christening
5 June 1873
Springville, Utah, Utah, United States

Death
29 June 1958
Springville, Utah, Utah, United States

Burial
2 July 1958
Springville, Utah, Utah, United States

Other Information

Open Details | Add

Alternate Name
Birth Name
James Mead Sumsion

Residence
1880
Springville, Utah, Utah, United States

1910
Springville Ward 5, Utah, Utah, United States

1920
Springville Ward 5, Utah, Utah, United States

1930
Springville, Utah, Utah, United States

1940
Springville Election Precinct, Springville, Ward 5, Utah, Utah

Research Help

Record Hints

James Mead Sumsion
United States, GenealogyBank Obituaries, 1980-2014

James Mead Sumsion
United States, GenealogyBank Obituaries, 1980-2014

Show Details

Latest Changes

Parent-Child Source Attached
8 May 2017
by Cherise Armstrong

Child Added
8 May 2017
by Cherise Armstrong

Source Attached
8 May 2017
by Cherise Armstrong

Show all

Search Records

FamilySearch
ancestry
findmypast
MyHeritage

Print

Pedigree
Portrait Pedigree
Fan Chart
Family

Person Page (bottom)

Family
Family with Sources

e Duplicates
Report Abuse
Delete Person Unavailable

Family Members

Spouses and Children

Show All | Add Spouse

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 2 Mar 1908
Utah Co.
Ina Maycock
1882–1908 • KWJX-XCP
Preferred
Children

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 5 November 1919
Salt Lake City, Salt Lake, Utah, United States
Rose Wischer
1878–Deceased • LZGM-F87
Preferred
Children

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 9 March 1926
Salt Lake City, Salt Lake, Utah, United States
Lela Coplan
1895–1983 • KWZ3-PW3
Preferred
Children (1)
James Coplan Sumsion
1928–2002 • KWZ3-P71
Add Child

Add Child with an Unknown Mother

Marriage: 23 Jan 1862
Salt Lake City, Salt Lake, Utah
Ann Elizabeth Bird
1844–1922 • KWVP-QXP
Children (13)
Annabell Sumsion
1863–1863 • LLS4-CR5
Julia Ann Sumsion
1864–1922 • KWNK-W3J
George Albert Sumsion
1866–1932 • KWZB-3TV
Henry La Cell Sumsion
1868–1956 • KWC5-MSC
William Ernest Sumsion
1870–1934 • KWJ8-8H8
James Mead Sumsion
1873–1958 • KWZ3-PW9
Ellen Grace Sumsion
1874–1960 • KWZ4-ZLY
Christie Ann Sumsion
1876–1964 • KWVP-1MB
Minetta Sumsion
1879–1969 • KW88-9MY
Walter Sumsion
1881–1882 • K2MC-3LD
Daniel Bird Sumsion
1883–1885 • K2MC-3GT
Hilda Sumsion
1886–1888 • K2MC-3TS
Clara Rebecca Sumsion
1888–1888 • K2MC-3P9
Add Sibling

Sources

Open Details | Add Source | Attach from Source Box

James Mead Sumsion in entry for James Coplan Sumsion, "United States, GenealogyBank Obituaries, 1980-2014"

James M. Sumsion, "Utah Marriages, 1887-1935"

James Mead Sumsion, "Utah Marriages, 1887-1935"

18 persons (w/ details)

18 relationships (w/ details)

1 page view

Person Page (bottom)

James Mead Sumsion (1873–1

Secure https://familysearch.org/tree/person/KWZ3-PW9/det

1940
Springville Election Precinct, Springvil

Family Members

Spouses and Children

Show All Add Spouse

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 2 Mar 1908
Utah Co.
Ina Maycock
1882–1908 • KWJX-XCP
Preferred
Children

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 5 November 1919
Salt Lake City, Salt Lake, Utah, United States
Rose Wischer
1878–Deceased • LZGM-F87
Preferred
Children

James Mead Sumsion
1873–1958 • KWZ3-PW9
Marriage: 9 March 1926
Salt Lake City, Salt Lake, Utah, United States
Lela Coplan
1895–1983 • KWZ3-PW3
Preferred
Children (1)
James Coplan Sumsion
1928–2002 • KWZ3-P71
Add Child

Add Child with an Unknown Mother

Parents and Siblings

Hide All Add Parent

George Sumsion
1842–1907 • KWJ8-4W5
Marriage: 23 Jan 1862
Salt Lake City, Salt Lake, Utah
Ann Elizabeth Bird
1844–1922 • KWVP-QXP

Children (13)

Annabell Sumsion
1863–1863 • LLS4-CR5
Julia Ann Sumsion
1864–1922 • KWNK-W3J
George Albert Sumsion
1866–1932 • KWZB-3TV
Henry La Cell Sumsion
1868–1956 • KWC5-MSC
William Ernest Sumsion
1870–1934 • KWJ8-8H8
James Mead Sumsion
1873–1958 • KWZ3-PW9
Ellen Grace Sumsion
1874–1960 • KWZ4-ZLY
Christie Ann Sumsion
1876–1964 • KWVP-1MB
Minetta Sumsion
1879–1969 • KW88-9MY
Walter Sumsion
1881–1882 • K2MC-3LD
Daniel Bird Sumsion
1883–1885 • K2MC-3GT
Hilda Sumsion
1886–1888 • K2MC-3TS
Clara Rebecca Sumsion
1888–1888 • K2MC-3P9

Add Sibling

ait Pedigree
Fan Chart
Family
Family with Sources

Tools
Possible Duplicates
Report Abuse
Delete Person Unavailable

Person Card Popup

Parents and Siblings

Hide All Add Pare

George Su
1842–1907

Marriage: 23 Jan 1862
Salt Lake City, Salt Lake, Utah

Ann Elizabeth Bird
1844–1922 • KWVP-QXP

Report Abuse

Delete Person Unavailable

Children (13)

Annabell Sumsion
1863–1863 • LLS4-CR5

Julia Ann Sumsion
1864–1922 • KWNK-W3J

George Albert Sumsion
1866–1932 • KWZB-3TV

Henry La Cell Sumsion
1868–1956 • KWC5-MSC

William Ernest Sumsion
1870–1934 • KWJ8-8H8

James Mead Sumsion
1873–1958 • KWZ3-PW9

Ellen Grace Sumsion
1874–1960 • KWZ4-ZLY

Christie Ann Sumsion
1876–1964 • KWVP-1MB

Minetta Sumsion
1879–1969 • KW88-9MY

Walter Sumsion
1881–1882 • K2MC-3LD

Daniel Bird Sumsion
1883–1885 • K2MC-3GT

Hilda Sumsion
1886–1888 • K2MC-3TS

Clara Rebecca Sumsion
1888–1888 • K2MC-3P9

Add Siblings

James Mead Sumsion

KWZ3-PW9 Unwatch

BIRTH 27 April 1873
Springville, Utah, Utah, United States

DEATH 29 June 1958
Springville, Utah, Utah, United States

22 SOURCES 0 DISCUSSIONS

7 MEMORIES

TEMPLE B C I E SP SS

TREE PERSON

Person Change History

FamilySearch

Messages Volunteer Get Help ▾

John Sumsion ▾

Go to: James Mead Sumsion

James Mead Sumsion

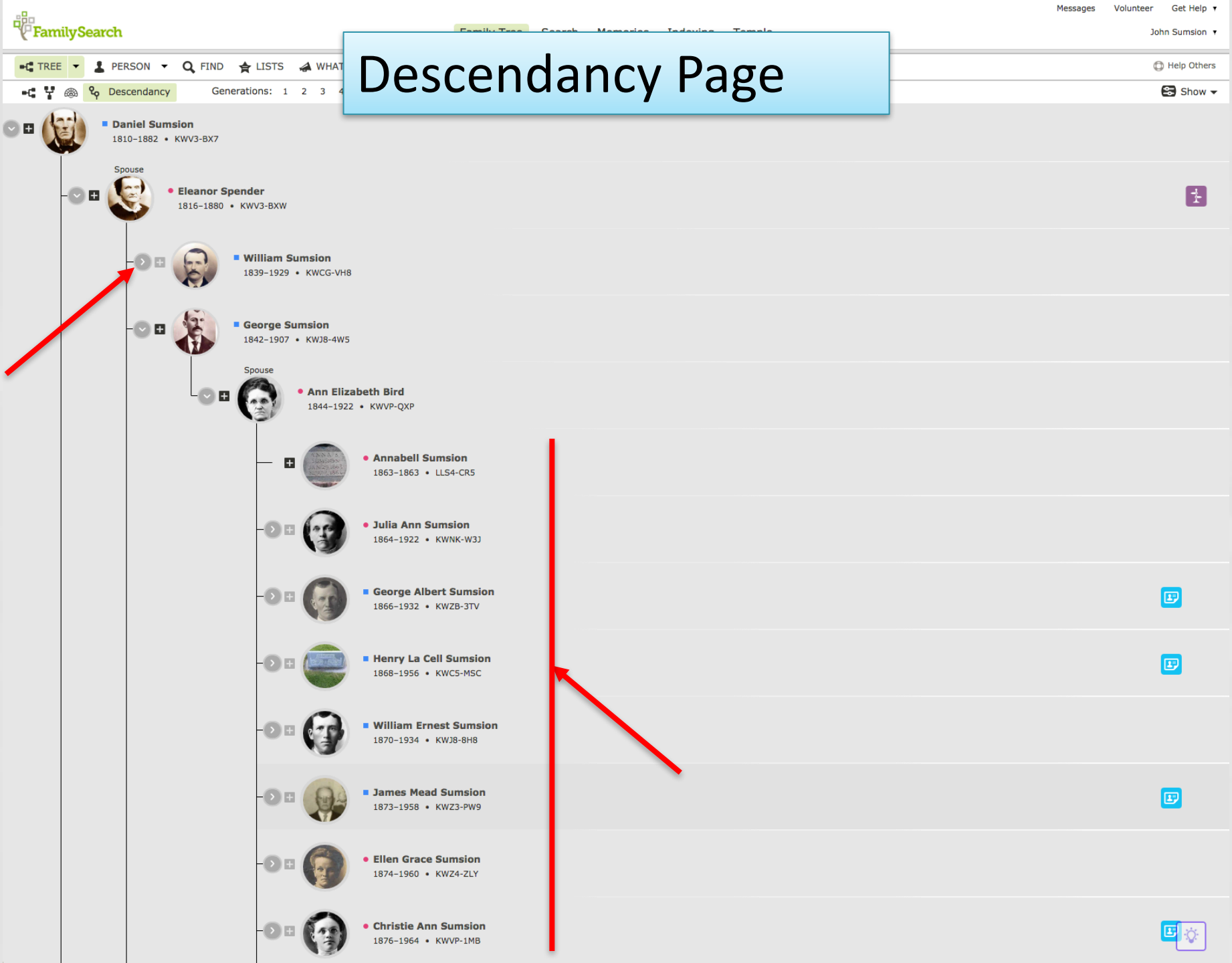
27 April 1873 – 29 June 1958 • KWZ3-PW9

View My Relationship

All Changes

Information	Details	
Parent-Child Source Attached 8 May 2017 by Cherise Armstrong	James Coplan Sumsion, "United States, GenealogyBank Obituaries, 1980-2014"	Reference
Child Added 8 May 2017 by Cherise Armstrong	Child James Coplan Sumsion 1928-2002 Show Relationship	Reference
Source Attached 8 May 2017 by Cherise Armstrong	James Mead Sumsion in entry for James Coplan Sumsion, "United States, GenealogyBank Obituaries, 1980-2014"	Current
Source Attached 24 March 2017 by terilynnferney1	James M. Sumsion, "Utah Marriages, 1887-1935"	Current
Photo Attached 2 December 2016 by John Sumsion	Show Photo	Current
Document Attached 1 December 2016 by John Sumsion	Show Document	Current
Source Attached 3 August 2016 by heathergcurtis	James Mead Sumsion, "Utah Marriages, 1887-1935"	Current
Source Attached 26 June 2016 by ChristensenClydeS1	James Mead Sumsion, "Utah Marriages, 1887-1935" Reason This Source Is Attached name and years match. The date listed in Utah Marriage record is 4 March 1908	Current
Couple Event Added 26 June 2016 by ChristensenClydeS1	Couple Event Marriage 04 Mar 1908 Salt Lake, Utah, United States Reason for Adding This Information name and years match. The date listed in Utah Marriage record is 4 March 1908	Reference

Descendancy Page



Family Tree

- Flexible schema
- 4th major iteration over 10 years
- Schema still adjusted relatively often (6 mo)

Family Tree

- API stats:
  - 300M API requests / peak day
  - 300K API requests / peak minute
  - 150M API requests / off-peak day
- DB stats:
  - 1.5B reads / peak day
  - 58K reads / sec (peak)
  - 10M writes / peak day

Family Tree

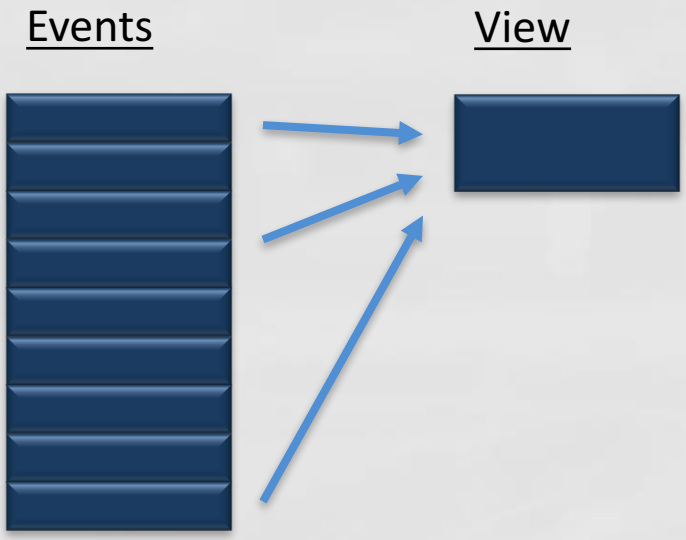
- DB stats:
  - 20TB of data (without 3x replication)
  - 7.5TB of that is canonical
  - 12.5TB is derivative, denormalized for queries
- DB size:
  - 60TB of disk used (replication factor = 3)
  - Able to drop most derivative data in emergency

Family Tree

- API performance
  - Peak day P90 is 22ms (instead of 2-5 sec on Oracle)
- DB performance
  - Peak day P90 is 2.3ms
  - Peak day P99 is 9.9ms
- Person page
  - Able to be served from 2 person reads
  - Still lots of room for optimization
  - Front-end client still over-reading

Events & Views

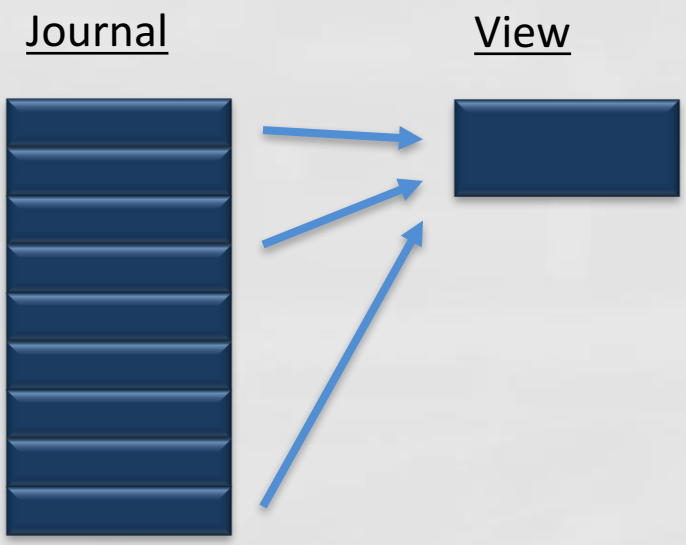
- Events are CANONICAL
- Multiple, derivative views
  - View computed from events
  - Views can be deleted (recomputed from events)
- Views stored in DB
  - For faster reads
- Event Sourcing



https://martinfowler.com/eaaDev/EventSourcing.html

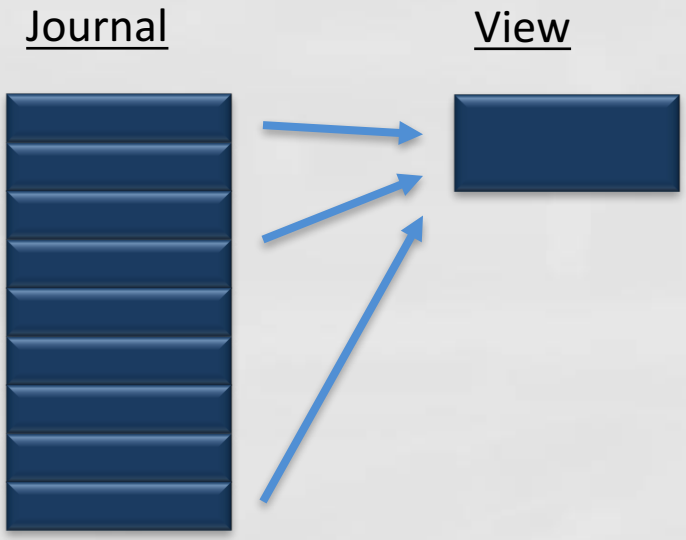
Events & Views

- Views optimized for Read
  - 100 reads : 1 write
- Different use case?
  - Might justify a new view
  - Might just change views
- Family Tree views
  - Person Card (summary)
  - Full Person View
  - Change History



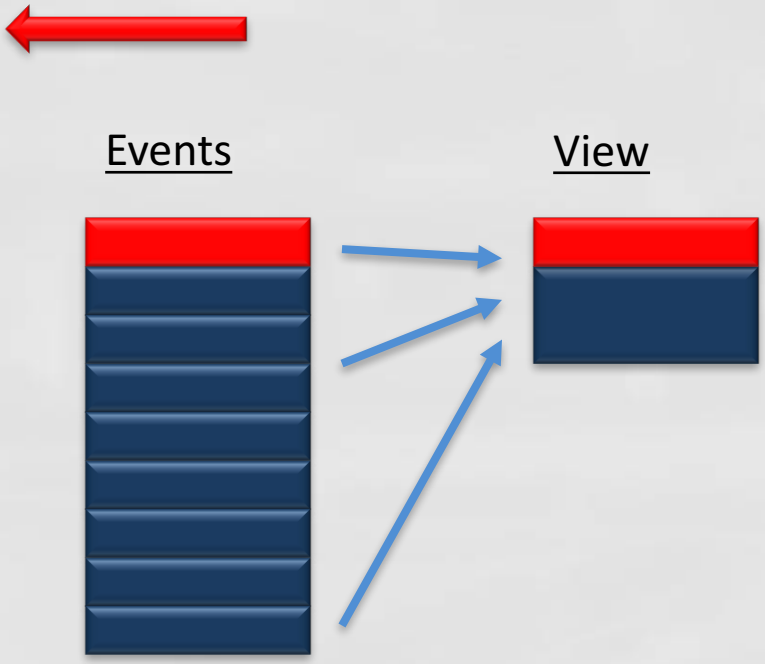
Events & Views

- Types of reads
  - Full View Refresh
  - Incremental View Refresh
  - Fast Path Read
    (no refresh needed)



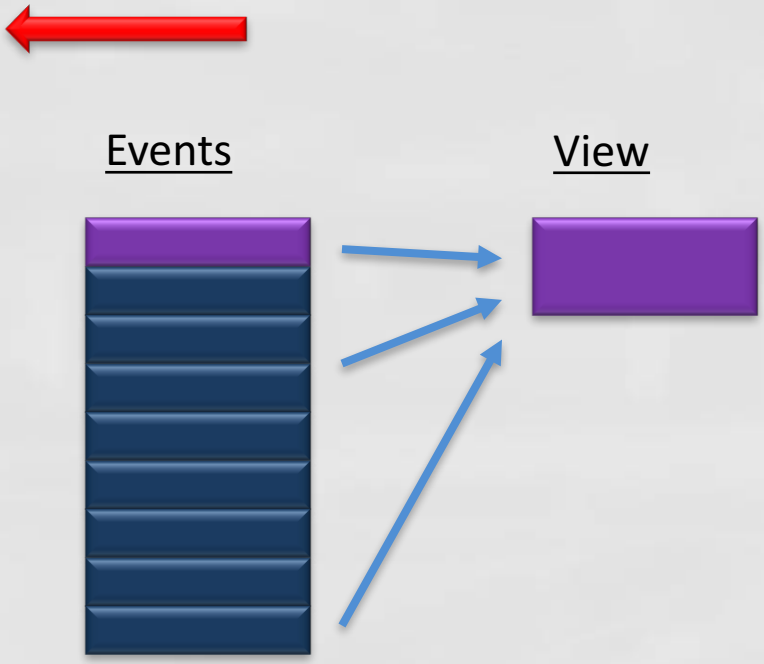
Events & Views

- Types of reads
  - Full View Refresh
  - Incremental View Refresh
  - Fast Path Read
    (no refresh needed)



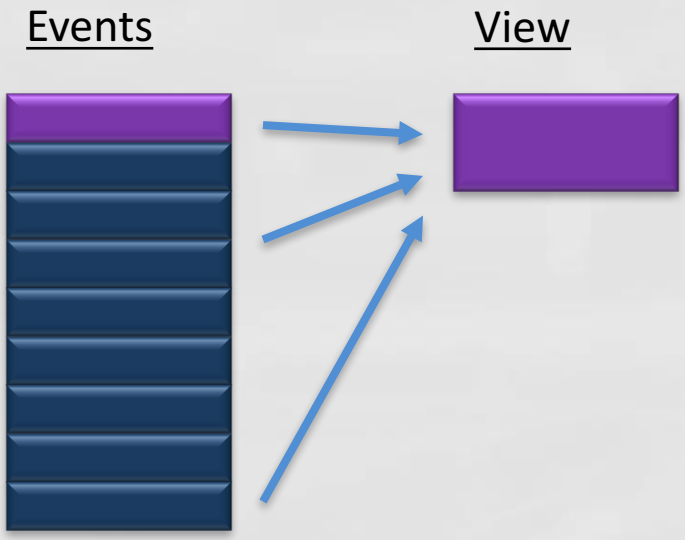
Events & Views

- Types of reads
  - Full View Refresh
  - Incremental View Refresh
  - Fast Path Read
    (no refresh needed)



Events & Views

- Read Optimizations
  - Row Cache for view tables 14G (out of 60G)
  - CL=ONE for Fast Path Read
  - Upgrade to LOCAL_QUORUM
    - if read fails
    - if view refresh is required
- Write Optimization
  - Group events into tx record
  - Split txs to avoid over-copy



Events & Views

- Sample Cassandra Schema (event table):

```
CREATE TABLE person_journal (
    entity_id text,
    record_id timeuuid,
    sub_id int,
    type text,
    subtype text,
    content blob,
    primary key ((entity_id), record_id, sub_id, type, subtype))
    with compaction = { 'class': 'SizeTieredCompactionStrategy' };
```

Events & Views

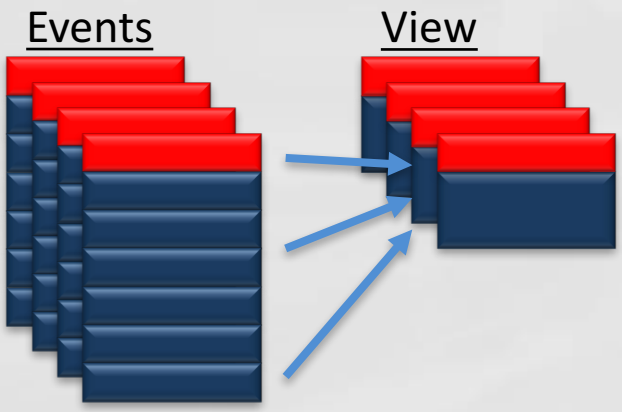
- Sample Cassandra Schema (view table):

```
CREATE TABLE person_view (
    entity_id text,
    record_id timeuuid,
    sub_id int,
    type text,
    subtype text,
    content blob,
    primary key ((entity_id), record_id, sub_id, type, subtype))
    with caching = 'ALL’
    and compaction = { 'class': 'LeveledCompactionStrategy' }
    and gc_grace_seconds = 86400;
```

Consistency Model

Classes of Writes:

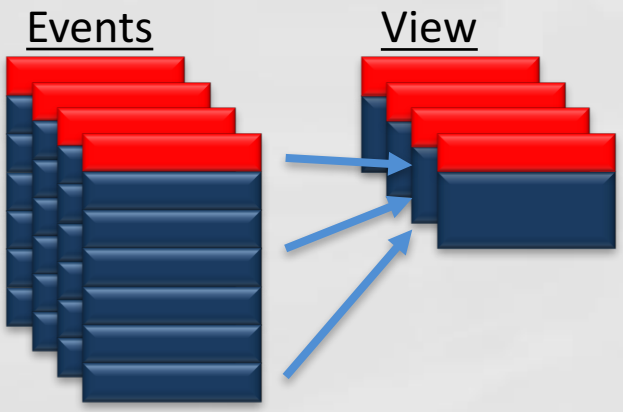
1. Single record edits
2. Multiple record edits
   - 2-4 records
   - Simple changes
3. Composite multi-record edits
   - Many records
   - Complex changes



Consistency Model

Write Process:

1. Create & write single “command” record
2. Pre-read affected records (views)
3. Pre-apply events (non-durable)
4. Check for rule violations
5. Write events
6. Post-read new affected records
7. Check for rule violations
   - Revert if problems



Consistency Model

Failure Modes:

1. Rule violation
   - Bad request response
   - NO write
2. Race condition
   - Conflict response
   - Revert

Consistency Model

Failure Modes:

3. Read Timeout at CL=ONE
   - Retry with LOCAL_QUORUM
   - Down node often is ignored
4. Write Timeout
   - Internal error response
   - Janitor follow-up later (from queue)
   - Idempotent writes

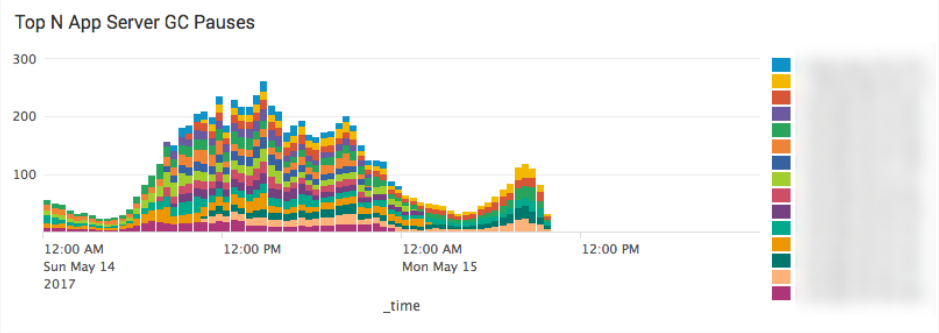
Surprises & Solutions

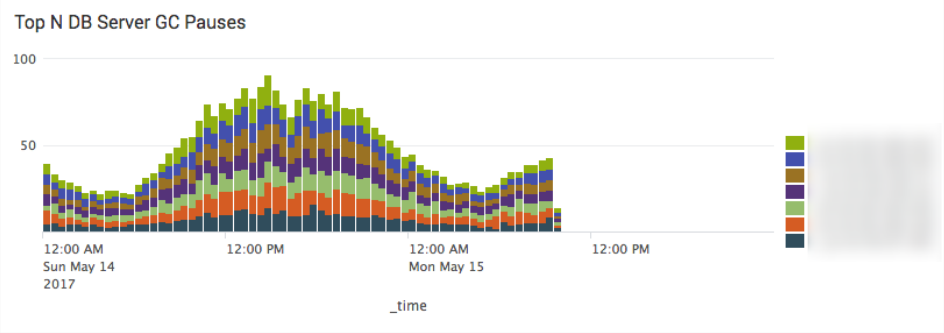
Surprises:

- Disproportionate Rate issues
- NTP Time issues
- Consistency issues

Disproportionate Rate

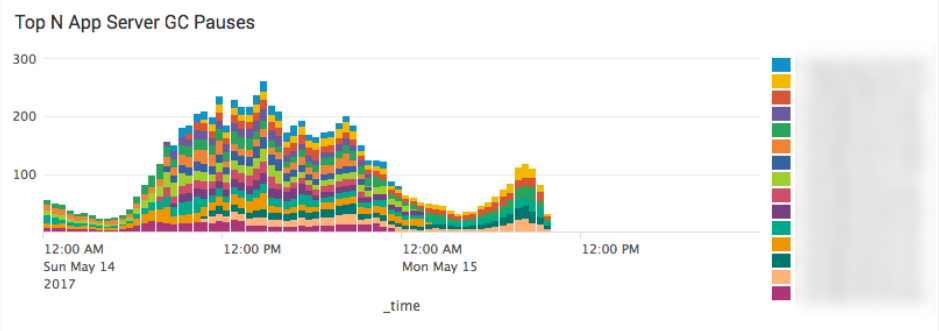
- Surprise: Bytes matter, not queries
  - Number of queries has less to do with latency
  - Large number of bytes cause CPU from Java GC
  - Multiple copies of large edited blobs add up

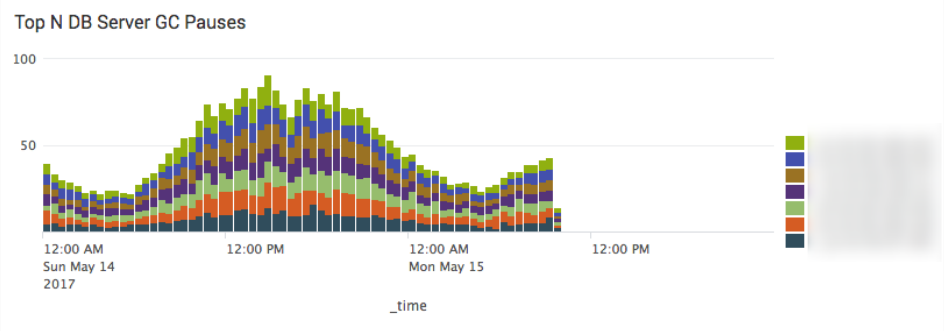




Disproportionate Rate

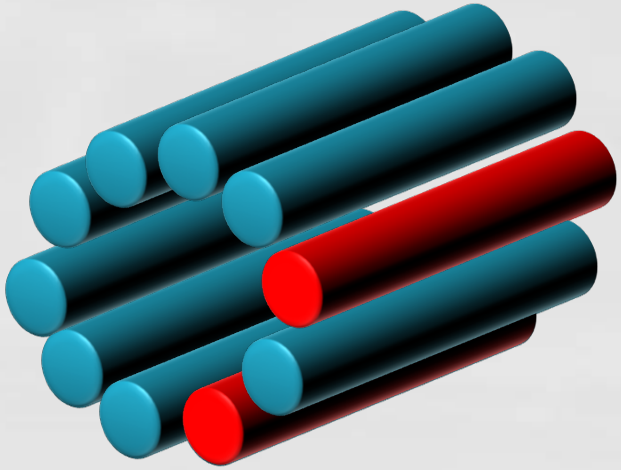
- Surprise: VERY Large Views
  - Well-known historical persons
  - Vanity genealogy (connecting to royalty)
  - 50+ names, 100+ spouses, 500+ parents
  - Many more bytes / request than normal (skews GC)





Disproportionate Rate

- Surprise: Single nodes matter, not total cluster
  - Slow node affects all traffic on that node
  - Events & Views on same node, worse hotspots



- Surprise: Replica set surprisingly resilient

Disproportionate Rate

- Solution #1:
  - Reduce size of views
  - Family Tree data limits (control) & data cleanup (fix)
  - Emergency blacklist for certain records until they can be manually trimmed
- Solution #2:
  - Throttle duplicate requests
  - Throttle problem clients
  - Reduce rate of requests to specific replica set

Disproportionate Rate

- Solution #3:
  - Spread views by prepending key prefix
  - Events on different set of nodes than views
  - Put each type of view on different set of nodes
  - Spread traffic out
- Solution #4:
  - Prevent merge / edit wars (limits)
  - Emergency lock records / suspend accounts

Disproportionate Rate

- Solution #5:
  - Split command up into contiguous events
  - Avoid over-copying large transactions
  - Split batches when writing
  - Retry writes if writes time out (janitor & queue)
- Solution #6:
  - Change view tables to LCS (leveled compaction)
  - Lower gc_grace_seconds for view tables to 2d
  - Emergency: Truncate view tables

Disproportionate Rate

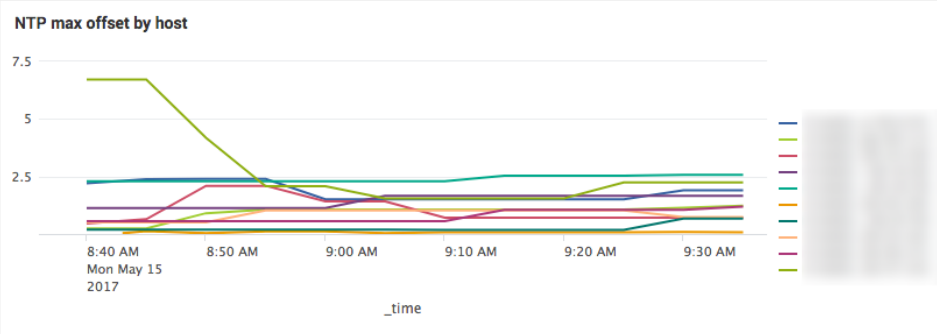
- Solution #7:
  - Pre-compute common views
  - Spread out pub-sub consumers with queue delays
  - Prevents incremental view refresh races from pub-sub consumers

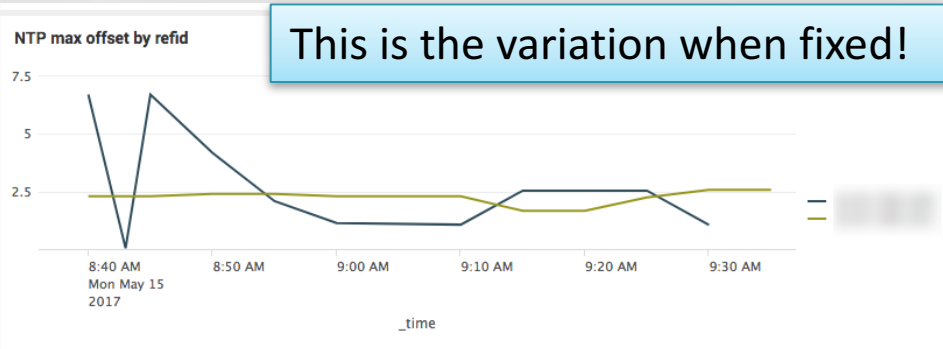
NTP Time

- NTP Time Issues:
  - Event transaction id is V1 time-based UUID
  - UUID generated on app server
  - Sequence of writes across app servers
  - App server time out of sync (broken NTP)
  - Arbitrary event reordering

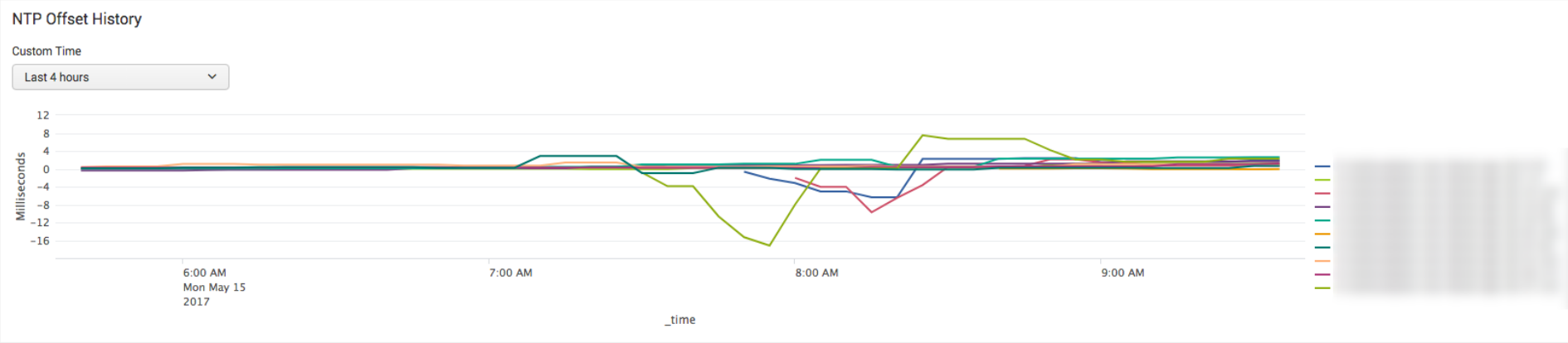
NTP Time

- Solution #1:
  - Fix NTP config, of course
  - Monitor / alert on NTP sync issues



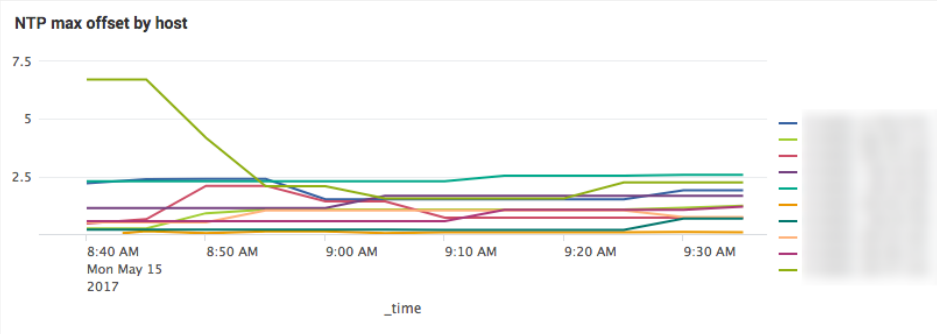


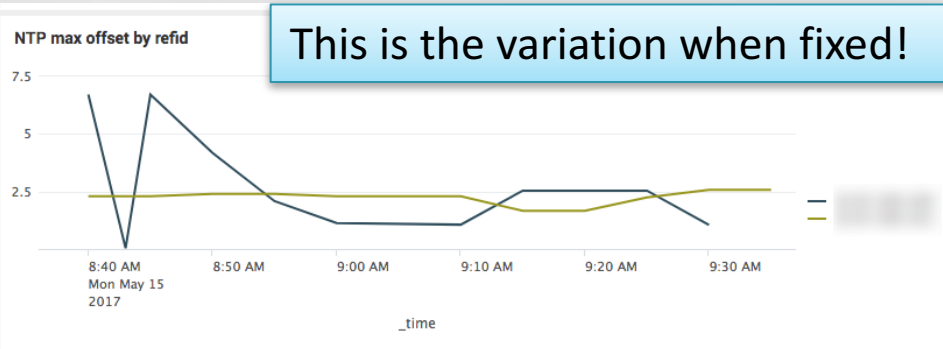
This is the variation when fixed!



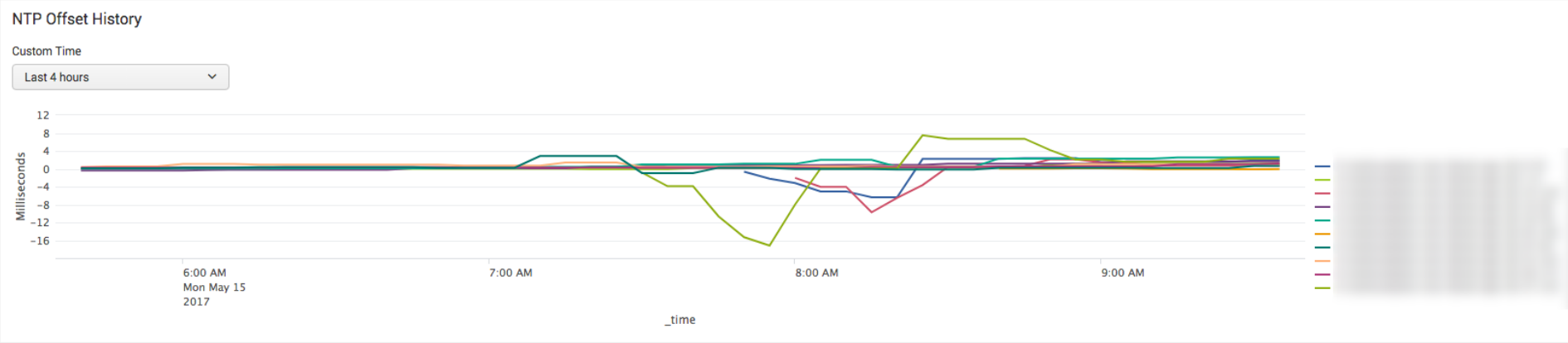
NTP Time

- Solution #2:
  - Keep V1 UUIDs in sequence at write time
  - Read prior UUID and wait up to 500ms until in past





This is the variation when fixed!



View Inconsistency

- Concurrent writes:
  - Concurrent incremental view refresh
  - Different view snapshots read (different nodes)
  - Overlapping view writes
  - Missing view data (as if write never happened)
- Partial writes:
  - Timeout on complex many-record write
  - Janitor not yet caught up replaying write
  - User refreshes and attempts again

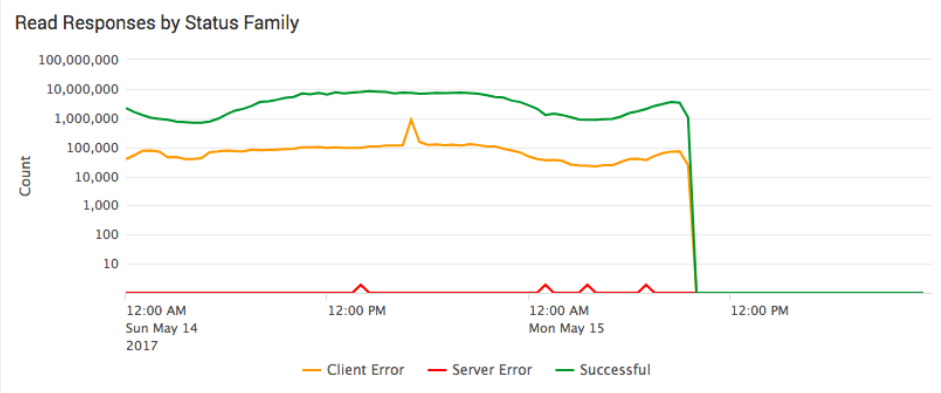
View Inconsistency

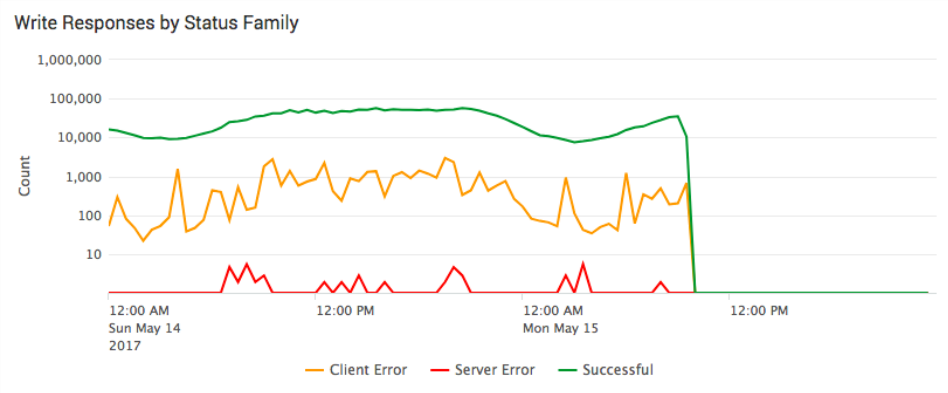
- Solution #1:
  - Observe view UUID during event preparation
  - Observe view UUID during write
  - Revert if different (concurrent write conflict)
- Solution #2:
  - Spark job to find inconsistencies
  - Semi-automated categorization & fixup
  - Address each source of inconsistency

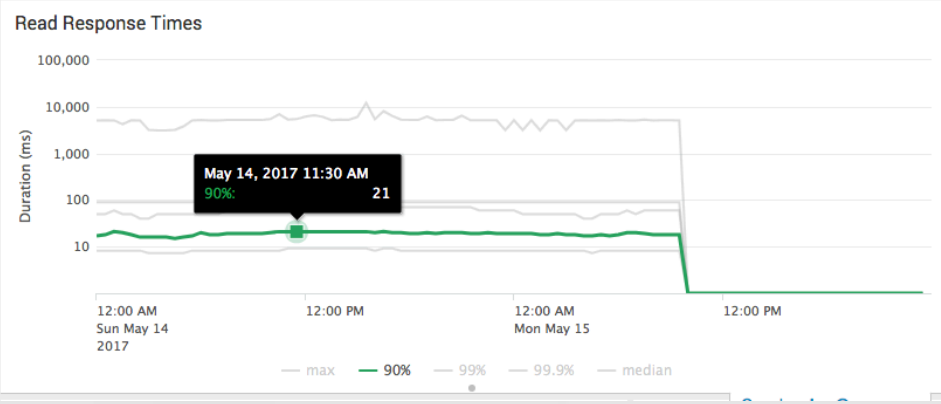
End Result

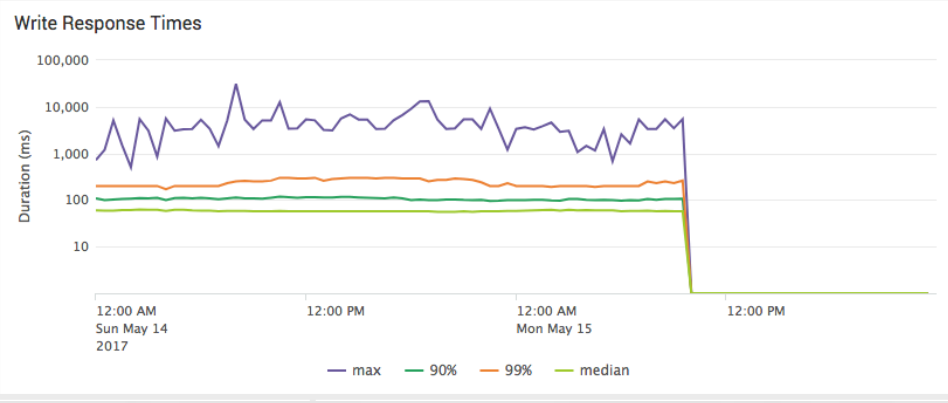
- Fantastic peak day performance
- Data consistency is good enough
- Consistency checks catching issues
- Quality of Family Tree improved with cleanups
- Splitting events / view – lots of flexibility
- Flexible schema – allows for agility
- Takes abuse from users and keeps running

End Result

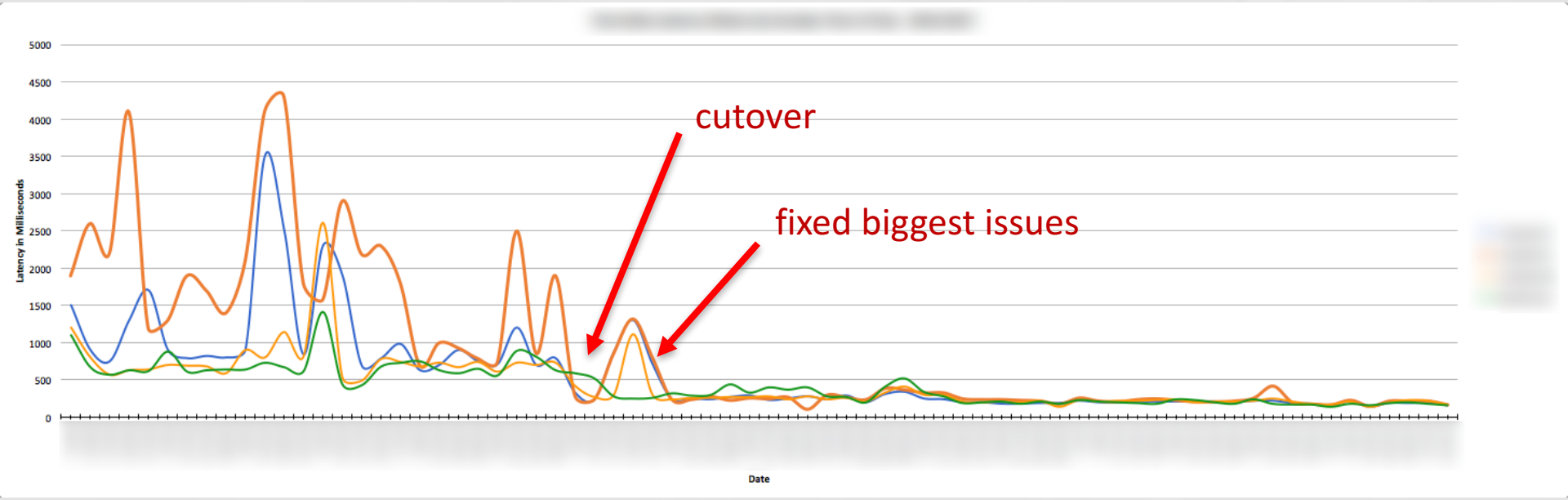

Read Responses by Status Family
Count
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
12:00 AM
Sun May 14
2017
12:00 PM
12:00 AM
Mon May 15
12:00 PM
Client Error
Server Error
Successful


Write Responses by Status Family
Count
1,000,000
100,000
10,000
1,000
100
10
12:00 AM
Sun May 14
2017
12:00 PM
12:00 AM
Mon May 15
12:00 PM
Client Error
Server Error
Successful


Read Response Times
Duration (ms)
100,000
10,000
1,000
100
10
May 14, 2017 11:30 AM
90%: 21
12:00 AM
Sun May 14
2017
12:00 PM
12:00 AM
Mon May 15
12:00 PM
max
90%
99%
99.9%
median


Write Response Times
Duration (ms)
100,000
10,000
1,000
100
10
12:00 AM
Sun May 14
2017
12:00 PM
12:00 AM
Mon May 15
12:00 PM
max
90%
99%
median

End Result



18 months, incl. 8 months before cutover

Learnings

- Event Sourced data model:
  - Very performant & scalable
  - Good enough consistency
- NTP time:
  - Must monitor / alert
  - Must deal with small offsets
- Consistency checks:
  - Long-term consistency must be measured
  - Fixes for measured issues must be applied

Questions

- Thanks:
  - To Apache for hosting the conference
  - To all Cassandra contributors
  - To Datastax for DSE
  - To FamilySearch for sending me