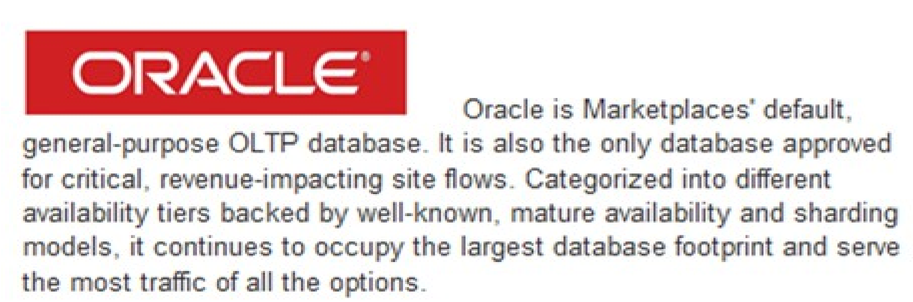
Scale

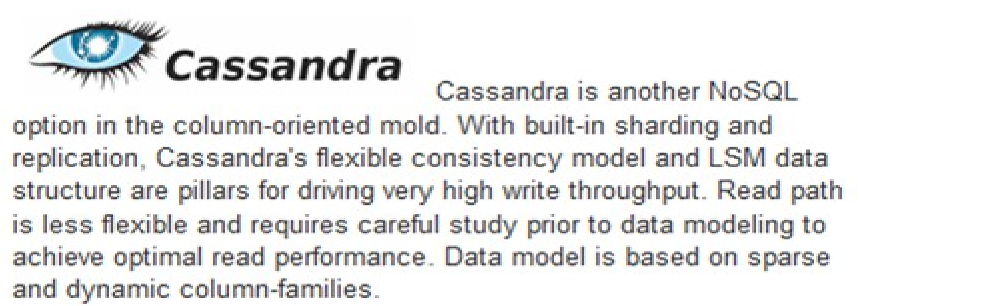
Anurag Jambhekar, Sr Manager Database Engineering

Feng Qu, Principal Database Engineer

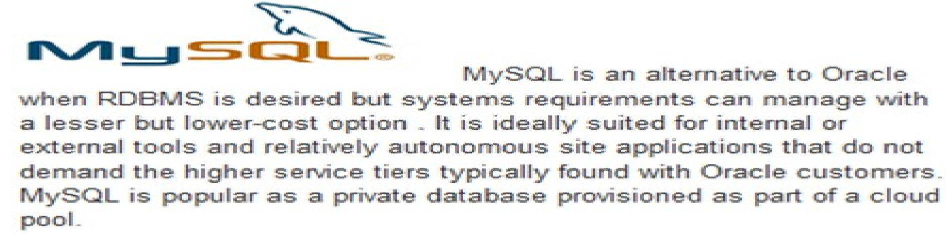
Databases in eBay marketplace's transactional platform



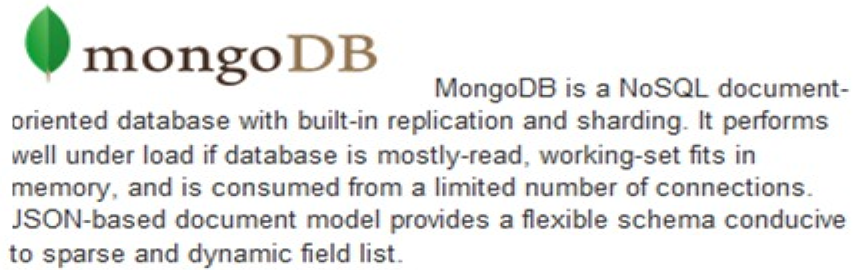
Oracle is Marketplaces' default, general-purpose OLTP database. It is also the only database approved for critical, revenue-impacting site flows. Categorized into different availability tiers backed by well-known, mature availability and sharding models, it continues to occupy the largest database footprint and serve the most traffic of all the options.



Cassandra is another NoSQL option in the column-oriented mold. With built-in sharding and replication, Cassandra's flexible consistency model and LSM data structure are pillars for driving very high write throughput. Read path is less flexible and requires careful study prior to data modeling to achieve optimal read performance. Data model is based on sparse and dynamic column-families.



MySQL is an alternative to Oracle when RDBMS is desired but systems requirements can manage with a lesser but lower-cost option . It is ideally suited for internal or external tools and relatively autonomous site applications that do not demand the higher service tiers typically found with Oracle customers. MySQL is popular as a private database provisioned as part of a cloud pool.



MongoDB is a NoSQL document-oriented database with built-in replication and sharding. It performs well under load if database is mostly-read, working-set fits in memory, and is consumed from a limited number of connections. JSON-based document model provides a flexible schema conducive to sparse and dynamic field list.

Xmp is a new emerging database system which offer best of relational and NoSQL model. It provide high transaction throughput with ACID compliance and auto sharding.)

#cassandra13

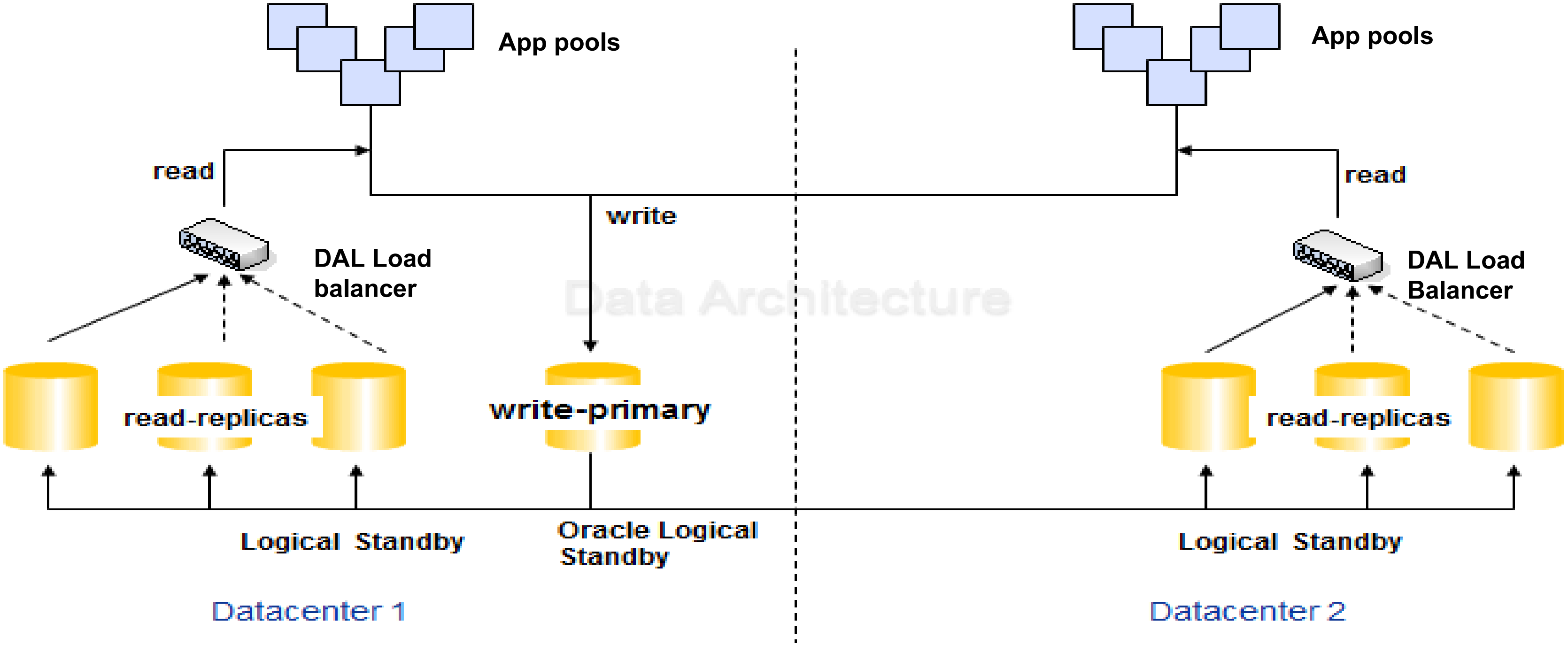
Database Scalability Patterns

- Logical database host and transparent mapping to physical database
- Multiple read copies
  - One write database and multiple read databases
- Horizontal Scaling
  - Shard data across multiple database hosts using Mod/Range/Lookup based Sharding
- Vertical scaling of database hardware
- Archiving of data
- Auto sharding in NoSQL makes it even easier to add capacity

Database Logical Host

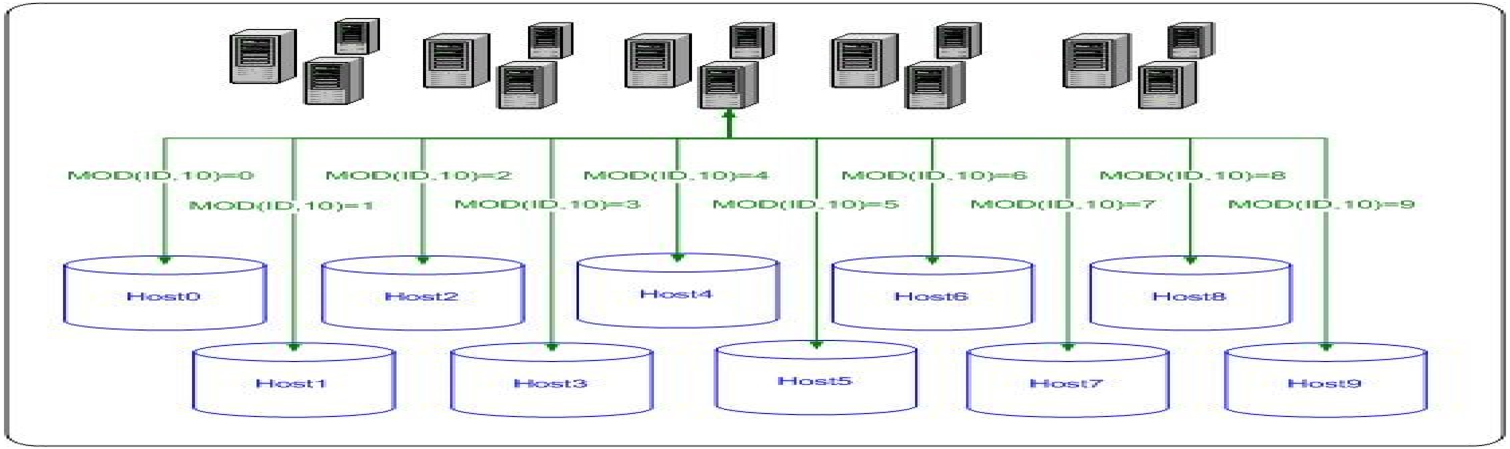
- eBay applications interact with the databases through a level of indirection that is referred to as Logical Databases.
- Logical databases to physical database relation is many-to-one
  - Smaller logicals from different DB families are mapped to same physical
  - DAL(Data Access Layer) does the translation from logical host to physical host
- Why Logical Databases?
  - Gives DBAs the independence to move objects based on load and traffic
  - Code modification is not required when the physical topology changes
  - Helps failover scenarios
  - Allows to quickly turn off less critical feature to minimize site impact

Always Available Read Databases



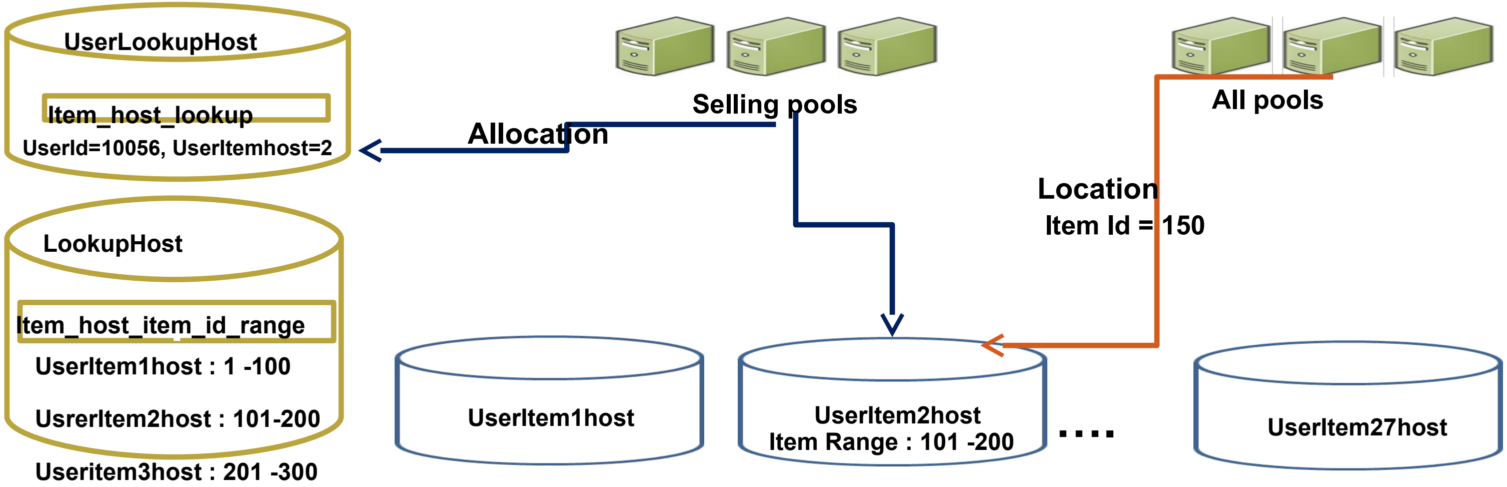
#cassandra13

MOD Based Model



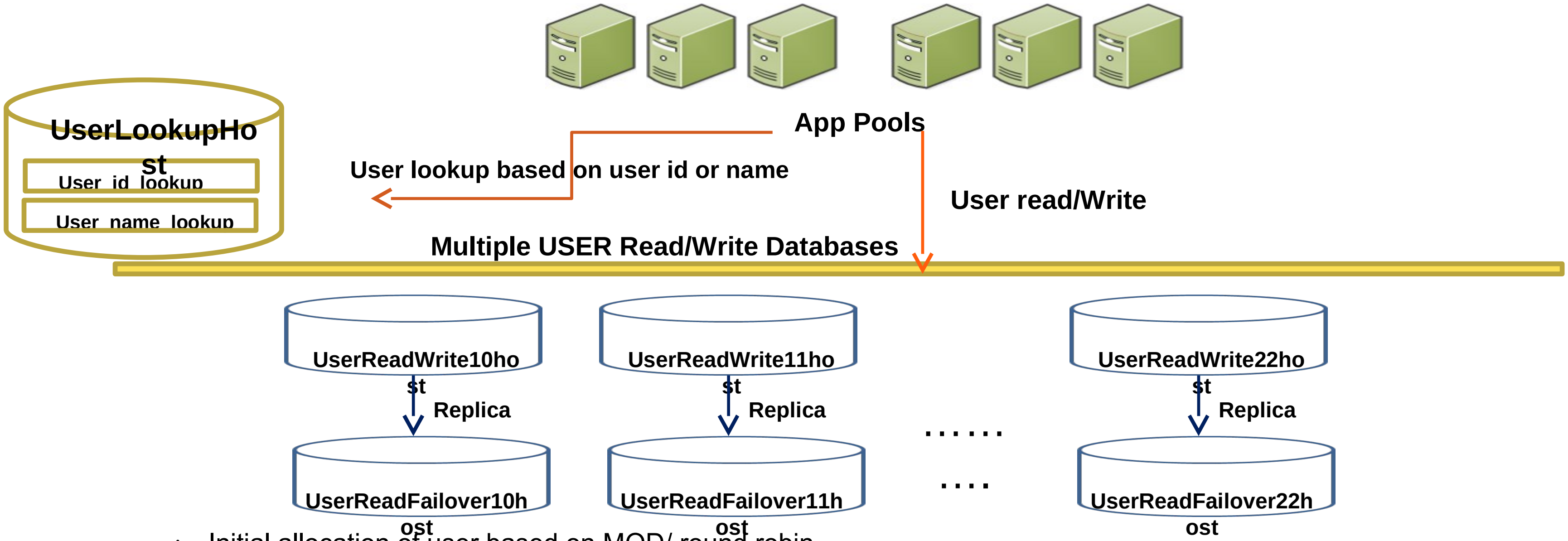
- Implicit allocation/routing based on a mathematical formula
- A modulus function of primary key provides host ID to allocate/use
- Applies mostly in case of user generated data
- Allows for simple segmentation with almost zero overhead
- Limits to scaling only to “N” based on initial configuration

Range Based Model – Item Split



- Allocation based on seller segment, and seller segment based on user host, seller level
  - All of a sellers items are listed in certain set of Item hosts
- Each User Item host has a range of IDs associated with it
  - All items listed on a item host get an ID within a specified Range
- ID being in a fixed range allows for location of items
- Indefinitely scalable

Lookup Based Model : User Split



- Initial allocation of user based on MOD/ round robin
- Location based on persistent mapping stored in UserLookup database family
  - Lookup based on user numeric id or name id.
- Each host has a subset of users and support both read and write
- Infinitely scalable

#cassandra13

Challenges of Traditional RDBMS

- Performance penalty to maintain ACID features
- Lack of native sharding and replication features
- Cost of software/hardware
- Higher cost of commit

Few words about myself

- Am a RDBMS veteran.
  - Started with Oracle 5 in early 1990s.
- For more than a decade, I worked on Oracle, MS SQL Server, MySQL in DoubleClick, Yahoo and Intuit
- In recent years, I am working at eBay marketplace focusing on NoSQL technology, like Cassandra, MongoDB.

Why Cassandra?

- Flexible schema
- Good performance for both reads/writes
- Horizontal linear scalability
- Built-in high availability
  - No SPOF
  - Automatic Geo load balancing and DR with multiple data centers deployment
- Automatic sharding for load balancing across multiple nodes
- Real-time data analytics, search with DSE, no more ETL
- No manual data purging required

Cassandra @ eBay

- Started in 2011
- Uses Apache Cassandra and DataStax Enterprise
- Java Hector client used most, and evaluating other options
- 10+ production clusters on 100+ bare metal nodes with 250TB storage provisioned across multiple data centers
- 100TB+ user data on local HDD and shared SSD array
- Cluster size varies from 6-node cluster to 32-node cluster

Cassandra Environment

- Software
  - Cassandra 1.0/1.1/1.2, DSE 3.0, RHEL 6.3, Hector 1.1
- Standard hardware:
  - HP DL380, 12 cores, 96GB RAM, 5.5TB raid10 HDD
  - Dell R620, 12 cores, 128GB RAM, 1TB raid10 HDD , 10 Gbps NIC
  - Violin 6200/6600 flash memory array
  - VM – coming soon
- Two types of cluster:
  - Dedicated cluster for large use case
  - Multi-tenant cluster for small use cases

eBay Scale

Data Centers:	3
Production Clusters:	10+
Physical nodes:	100+
Total data size:	100+ TB
Largest cluster:	32 nodes
Total reads:	5 billion/day
Total writes:	9 billion/day
Largest CF by size:	40 TB
Largest CF by rows:	32 billion
Busiest CF by reads:	1.6 billion/day
Busiest CF by writes:	6 billion/day

Packaging

- Build own RPMs on top of binary distribution
  - Different RPM for Apache Cassandra and DSE
- Easy to install/upgrade, easy to maintain
- Ensure deployment consistency across board and easy to identify deployment difference
- Built in pre-defined tuning parameters
- Using virtual nodes as default for new 1.2 clusters

Tuning Examples

- Kernel
  - Increase vm.max_map_count to fix *“attempt to allocate stack guard pages failed”* and *“java.lang.OutOfMemoryError: Map failed”*
- JVM heap
  - Keep heap size under control, and disable row cache(almost always)
- Compaction
  - LCS vs. STCS. Used LCS in 1.0, but had severe performance issue when # of SST grows to ~200K
  - Adjust compaction throughput as needed
- Hector client
  - Enable discovery mode and adjust timeout for different use cases

Scalability

- Benchmarking
  - Get performance baseline for new type of hardware
  - Enforce full scale testing in dedicated LnP env before going to production
- In general, **scale out** by adding more nodes to increase throughput
- Sometimes, it's cost-efficient to **scale up** at component level by Identifying scaling bottleneck, then resolve it accordingly
  - Network bandwidth: upgrade to 10 Gbps network
  - I/O latency: upgrade to SSD
  - Storage: add/expand data volume
  - CPU/Memory: haven't seen it yet

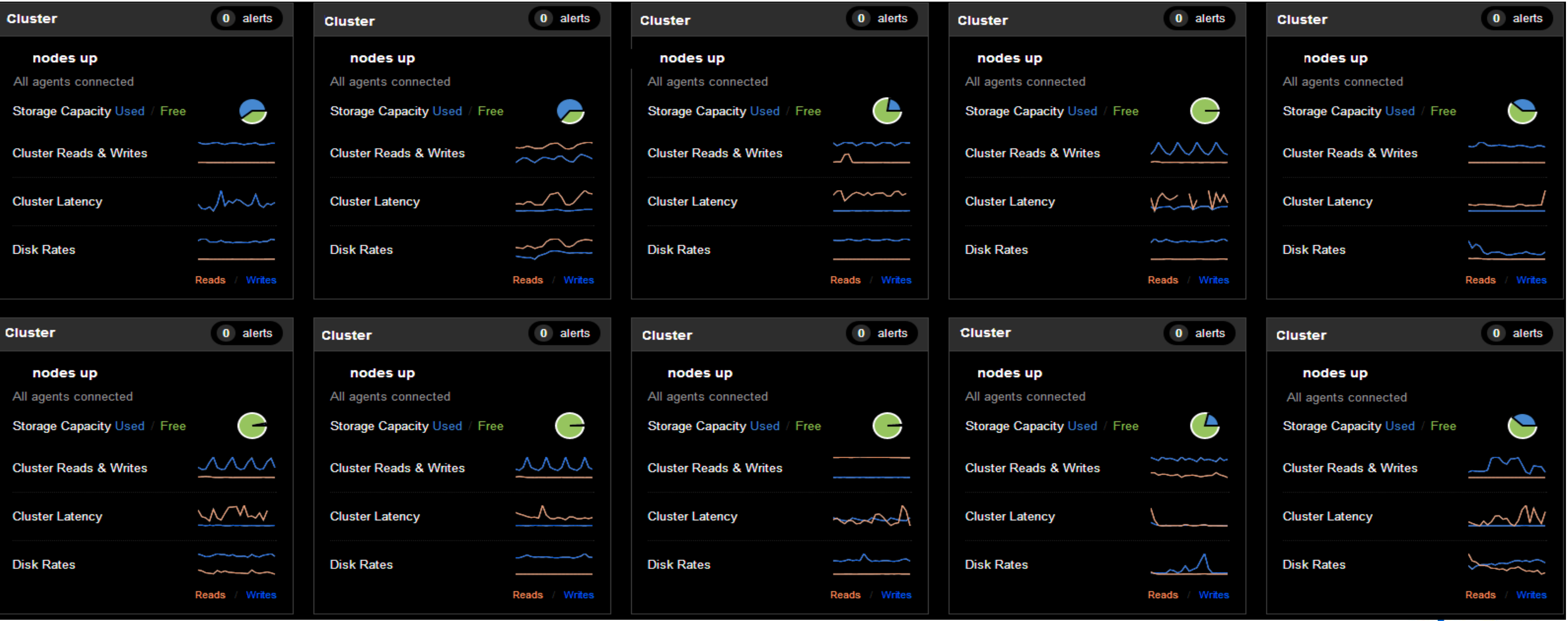
Operations Notes

- Routine repair is not really needed if there is no deletes. You still need run repair after bringing up a down node if it is dead for a while
- Use CNAME in client configuration to avoid client conf change in case of hardware replacement with new IP/name
- Reduce gc_grace to reduce overall data size
- Disable swap to avoid a slow node
- Bootstrapping didn’t work for 1st few times ☹, so you have keep trying
- Disable row cache, unless you have <100K rows
- Collect statistics, real-time or historical, to monitor system performance and provide dashboard report for management

Monitoring

- Integrate with NOC monitoring tool for 24*7 support
- Customized email alerts for non-critical events
  - Pending compactions
  - Garbage collections
  - Storage usage
- DataStax OpsCenter Enterprise
  - Monitor multiple clusters on one place
  - Rolling restart
  - Dashboard

OpsCenter



Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Cluster 0 alerts

nodes up

All agents connected

Storage Capacity Used / Free

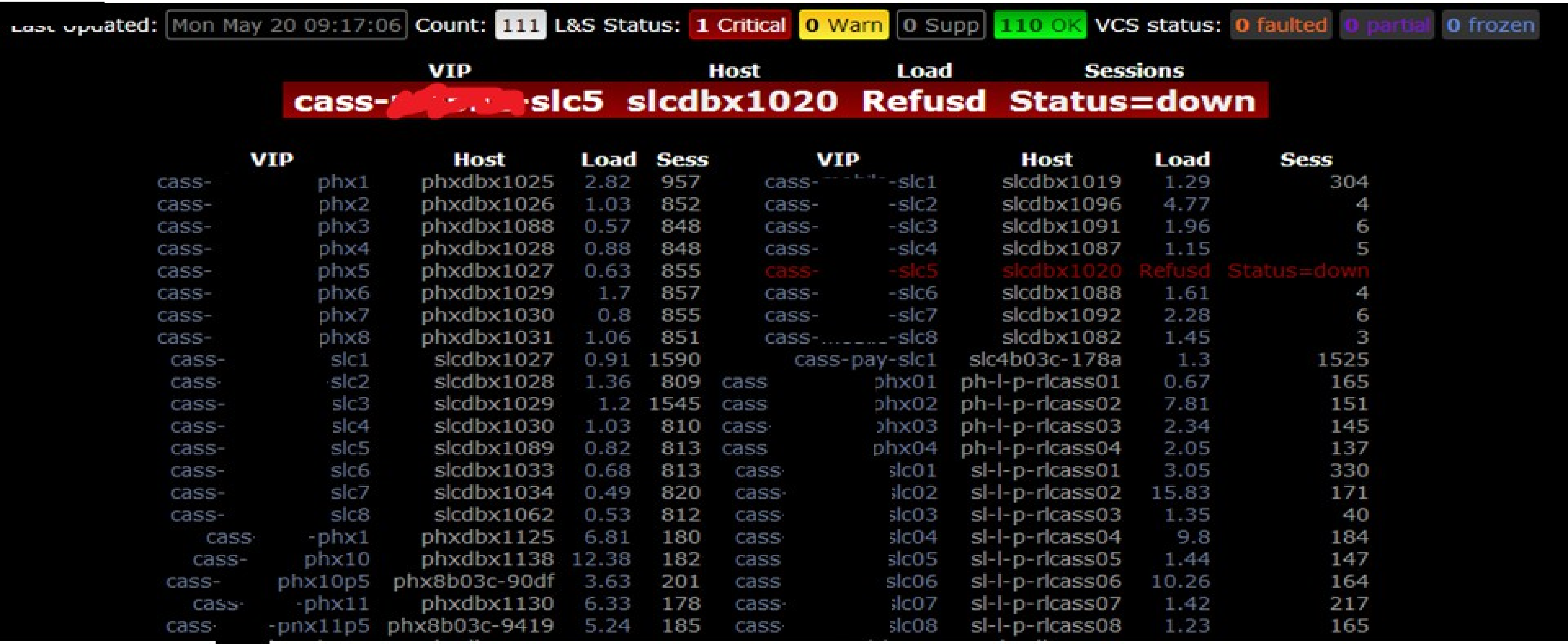
Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

NOC Real-time Monitoring

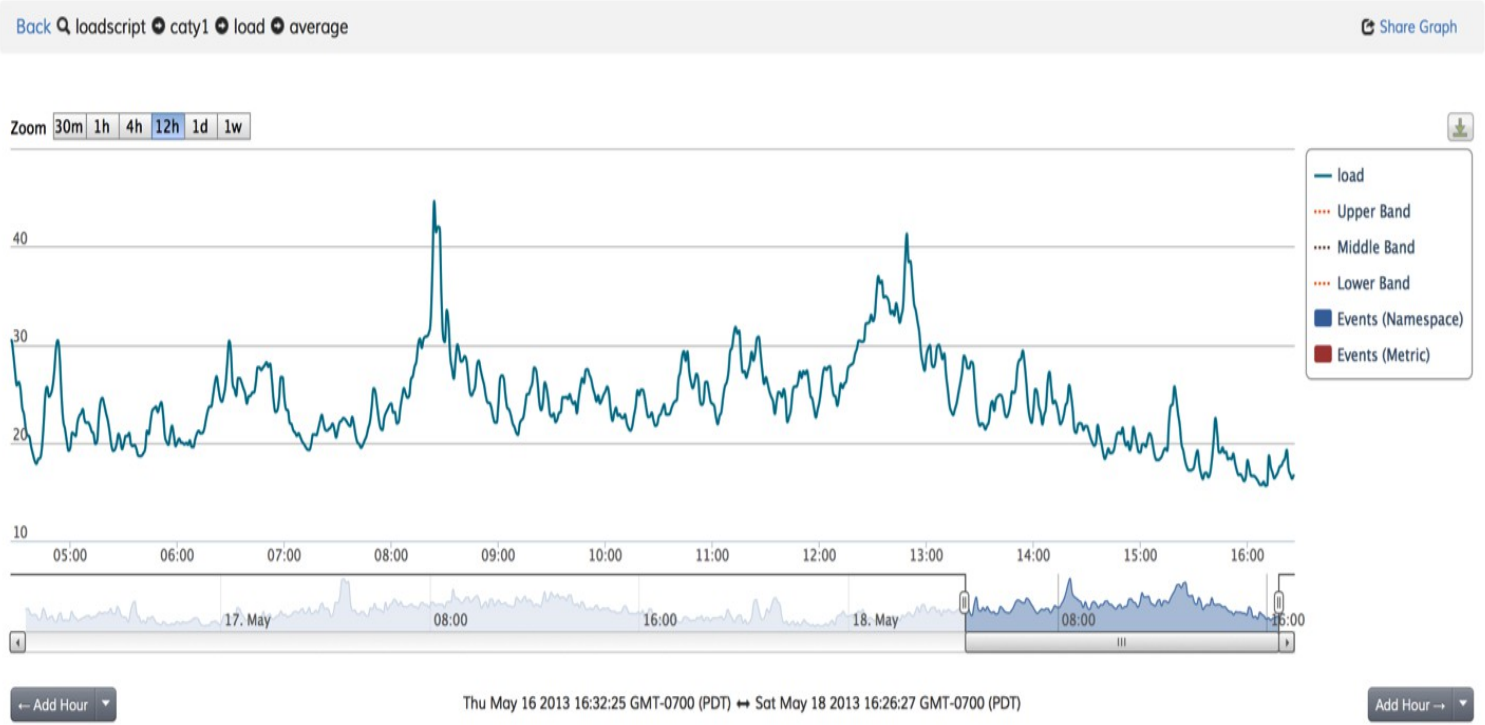


```
Last updated: Mon May 20 09:17:06  Count: 111  L&S Status: 1 Critical  0 Warn  0 Supp  110 OK  VCS status: 0 faulted  0 partial  0 frozen

                        VIP                       Host       Load      Sessions
           cass-[redacted]-slc5  slcdbx1020  Refusd  Status=down

          VIP                   Host        Load   Sess          VIP                  Host          Load        Sess
cass-[redacted]phx1       phxdbx1025   2.82    957   cass-[redacted]-slc1     slcdbx1019      1.29         304
cass-[redacted]phx2       phxdbx1026   1.03    852   cass-[redacted]-slc2     slcdbx1096      4.77           4
cass-[redacted]phx3       phxdbx1088   0.57    848   cass-[redacted]-slc3     slcdbx1091      1.96           6
cass-[redacted]phx4       phxdbx1028   0.88    848   cass-[redacted]-slc4     slcdbx1087      1.15           5
cass-[redacted]phx5       phxdbx1027   0.63    855   cass-[redacted]-slc5     slcdbx1020    Refusd Status=down
cass-[redacted]phx6       phxdbx1029    1.7    857   cass-[redacted]-slc6     slcdbx1088      1.61           4
cass-[redacted]phx7       phxdbx1030    0.8    855   cass-[redacted]-slc7     slcdbx1092      2.28           6
cass-[redacted]phx8       phxdbx1031   1.06    851   cass-[redacted]-slc8     slcdbx1082      1.45           3
  cass-[redacted]slc1       slcdbx1027   0.91   1590       cass-pay-slc1  slc4b03c-178a      1.3        1525
  cass·[redacted]·slc2      slcdbx1028   1.36    809  cass[redacted]phx01  ph-l-p-rlcass01     0.67         165
  cass-[redacted]slc3       slcdbx1029    1.2   1545  cass[redacted]phx02  ph-l-p-rlcass02     7.81         151
  cass-[redacted]slc4       slcdbx1030   1.03    810  cass·[redacted]phx03  ph-l-p-rlcass03    2.34         145
  cass-[redacted]slc5       slcdbx1089   0.82    813  cass[redacted]phx04  ph-l-p-rlcass04     2.05         137
  cass-[redacted]slc6       slcdbx1033   0.68    813    cass·[redacted]slc01  sl-l-p-rlcass01   3.05         330
  cass-[redacted]slc7       slcdbx1034   0.49    820    cass·[redacted]slc02  sl-l-p-rlcass02  15.83         171
  cass-[redacted]slc8       slcdbx1062   0.53    812    cass·[redacted]slc03  sl-l-p-rlcass03   1.35          40
      cass·[redacted]-phx1  phxdbx1125   6.81    180    cass·[redacted]slc04  sl-l-p-rlcass04    9.8         184
     cass-[redacted]phx10   phxdbx1138  12.38    182    cass[redacted]slc05   sl-l-p-rlcass05   1.44         147
 cass-[redacted]phx10p5  phx8b03c-90df   3.63    201    cass[redacted]slc06   sl-l-p-rlcass06  10.26         164
     cass·[redacted]-phx11  phxdbx1130   6.33    178    cass·[redacted]slc07  sl-l-p-rlcass07   1.42         217
 cass·[redacted]-pnx11p5 phx8b03c-9419   5.24    185    cass·[redacted]slc08  sl-l-p-rlcass08   1.23         165
```

Typical Cassandra Use Cases at eBay

- **Write Intensive:** Metrics collection
  - Collecting metrics from tens of thousands devices periodically
  - 6+ billion writes per day
- **Read Intensive:** Recommendation backend
  - 4+ billion reads per day
- **Mixed workload:** Personalization
  - Data is bulked loaded from data warehouse periodically
  - Data is retrieved in real time when user visits ebay site

Metrics Collection



Thu May 16 2013 16:32:25 GMT-0700 (PDT) ↔ Sat May 18 2013 16:26:27 GMT-0700 (PDT)

#cassandra13

Metrics Collection

- Storing, serving real-time metrics data for tens of thousands devices for dashboards, triage Automation etc
- Statistics
  - 16-node 1.2 cluster
  - 2 copies of data in 2 data centers
  - 600M keys
  - 6B writes including rollups per day

Recommendation Backend

$1.99
You are here

#cassandra13

Recommendation Backend

- Recommendation based on user’s list/watch/sell activities
- Two 1.0 clusters, 40 nodes combined
- Statistics
  - 25B keys
  - 32TB data on SSD
  - 600M writes per day
  - 3B reads per day

Vendor Support

- Support from DataStax.
  - CASSANDRA-4142: OOM during repair with LCS
  - CASSANDRA-4287: Slow compaction
  - CASSANDRA-4427: Restarting a failed bootstrap node instajoins the ring
  - CASSANDRA-5107: Node fails to start because host id is missing
- Prompt responses from highly skilled support persons
- Support $$$ is not wasted, trust me.

Looking Forward

- Support more mission critical use cases, like personalization, anti-fraud etc
- Hardware virtualization
  - OpenStratus which is built on top of OpenStack
- Separate storage from compute node
  - To scale different components as needed
- More automations, such as self-serve deployment
- More best practice process
  - Troubleshooting/Tuning runbook

Before Q&A…

Thanks for your time.

We are looking for talented NoSQL experts to join database engineering team

email resume to

ajambhekar at ebay dot com