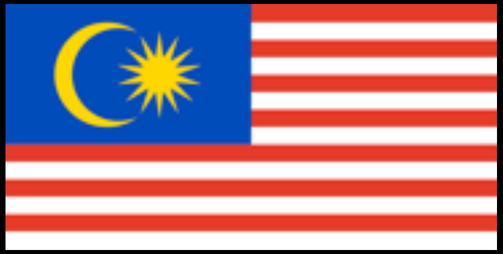
NewSQL trends and how MariaDB makes it happen

Colin Charles, MariaDB (SkySQL Ab)
colin@mariadb.org | byte@bytebot.net
http://skysql.com/ | http://mariadb.org/
http://bytebot.net/blog/ | @bytebot on Twitter

Open World Forum, Paris, France
4 October 2013

whoami



- Chief Evangelist, MariaDB at SkySQL Ab
  - Formerly Monty Program Ab (merged with SkySQL Ab)
- Formerly MySQL AB/Sun Microsystems
- Using/developing/hacking on MySQL since 2000
- Previously on FESCO for The Fedora Project, and hacked on OpenOffice.org code (2000-2005)

Agenda

- What is SQL? NoSQL? NewSQL?
- What is MariaDB?
- How does MariaDB help with being more "NewSQL compatible"?

SQL

- 1970 E.F. Codd paper "A Relational Model of Data for Large Shared Data Banks"
- ANSI standard in 1986
- Manage your data in relational systems, with transaction controls, SQL operators, and better cross-vendor compatibility

NoSQL/big data

- “not only SQL”
- 2004 Google paper “MapReduce: Simplified Data Processing on Large Clusters” -> Hadoop ~2008 “big data movement”
- 2006 Google paper “BigTable: A Distributed Storage System for Structured Data” -> NoSQL DBs like MongoDB ~2009
- Built without having a standardised structured query language
- Non-relational, distributed, horizontally scalable, schema-free, easy replication, eventually consistent

Eventual Consistency (BASE) & ACID

- ACID: Atomicity, Consistency, Isolation, Durability
    - the base of many transactional engines, like InnoDB/XtraDB
- BASE: Basically available, Soft state, Eventual Consistency
    - this is based on existence of CAP theorem which states that its impossible for a distributed computer system to simultaneously provide the following 3 guarantees:

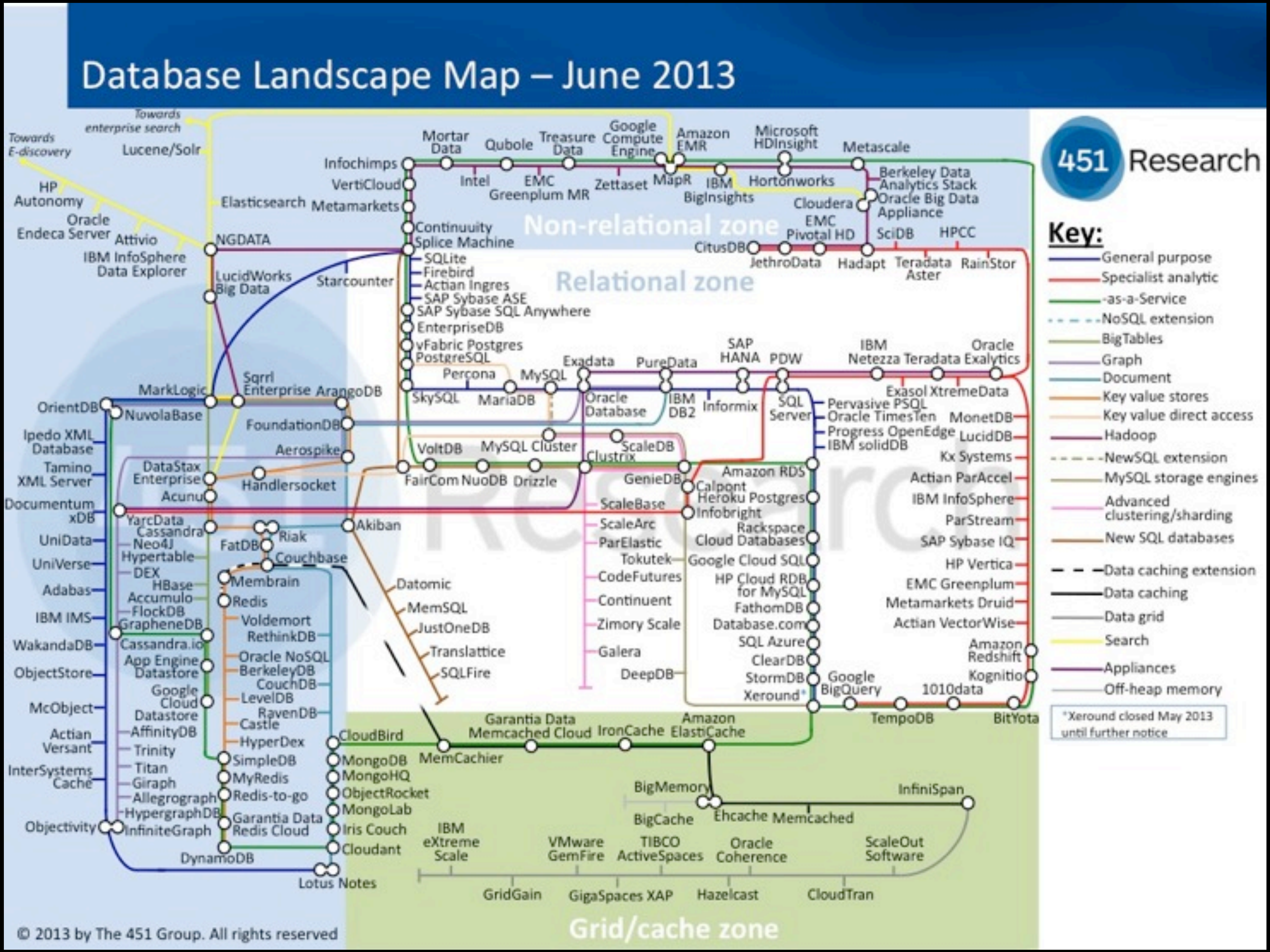
1. Consistency (all nodes have the same data)
2. Availability (guarantee that every request receives response)
3. Partition tolerance (continued operation with failure of parts of system)

- Satisfy 2 at the same time, but not 3

NewSQL

- 2012 Google paper "Spanner: Google's Globally-Distributed Database"
- goals: scalable, multi-version, globally distributed, synchronously replicated
- comes with a unique clock API allowing non-blocking reads in past, lock-free read-only transactions, atomic schema changes, etc.
- "We believe it is better to have application programmers deal with performance problems due to overuse of transactions as bottlenecks arise, rather than always coding around the lack of transactions."

Is this the landscape?



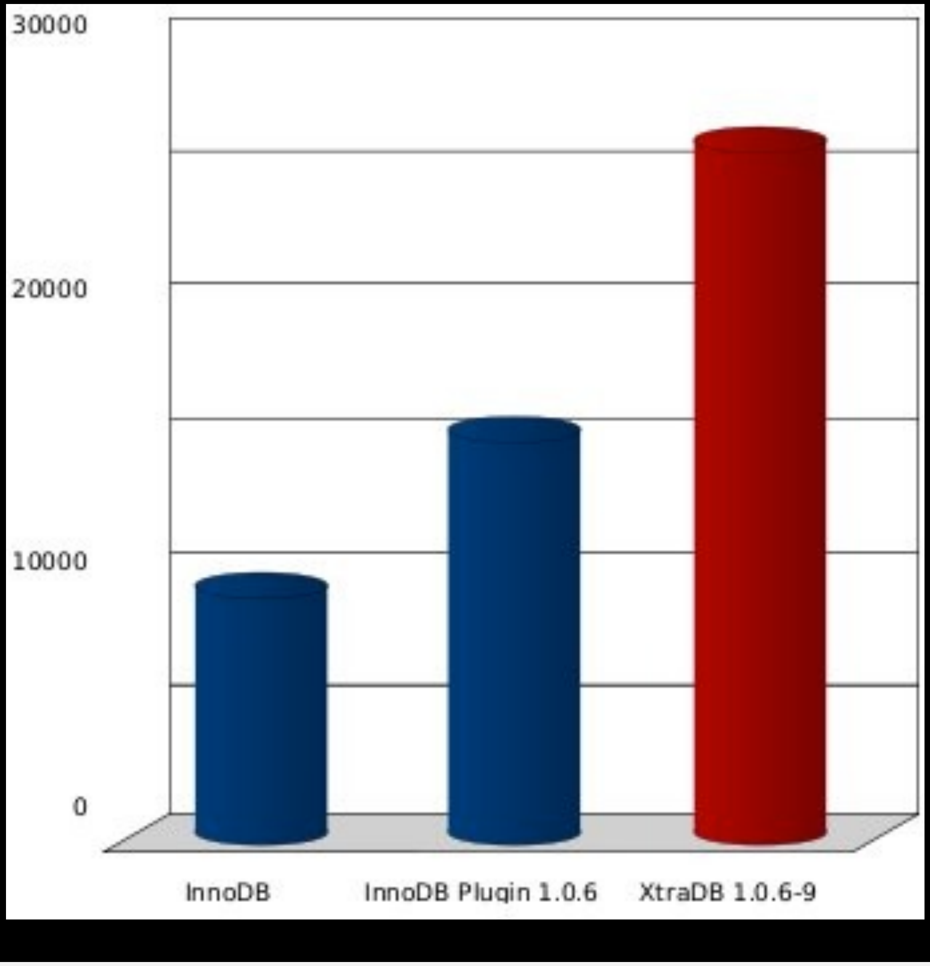
Source: https://blogs.the451group.com/information_management/files/2013/06/451db_map_06.13.jpg

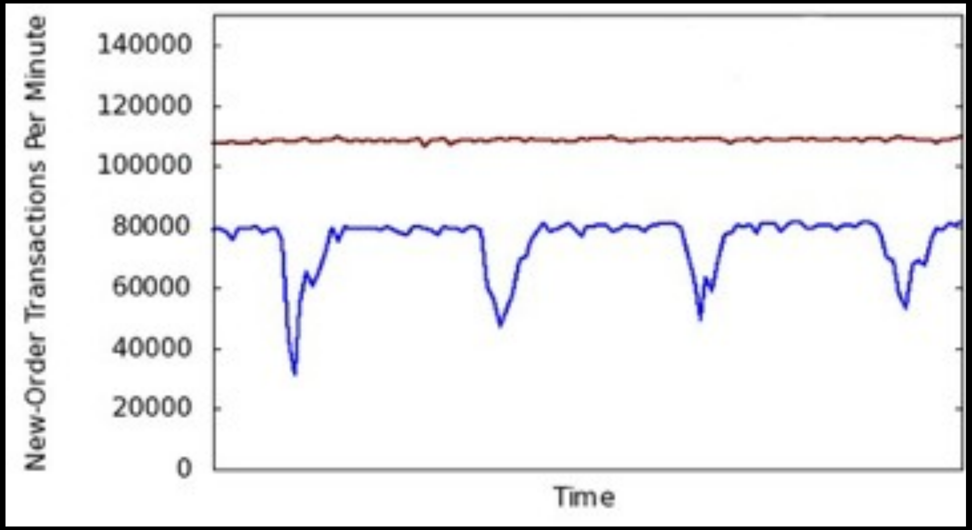
What is MariaDB?

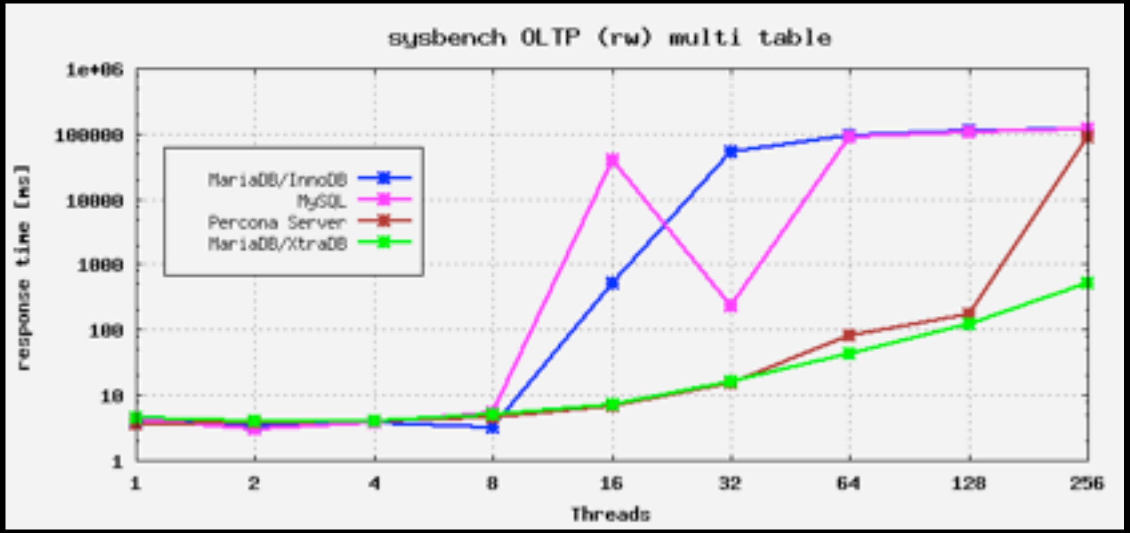
- Community developed, feature enhanced, backward compatible with MySQL database
- Correct ownership: community of contributors are 50% of committers (LinkedIn, Twitter, SkySQL, Taobao, Facebook, Codership, etc.)
- 100% compatible drop-in replacement, GPLv2
- Same client libraries, client-server protocol, master-slave replication, SQL dialect
- Governed by The MariaDB Foundation

XtraDB

- ENGINE=InnoDB uses XtraDB by default
- Less checkpointing (smoother), less flushing to disk, stable performance





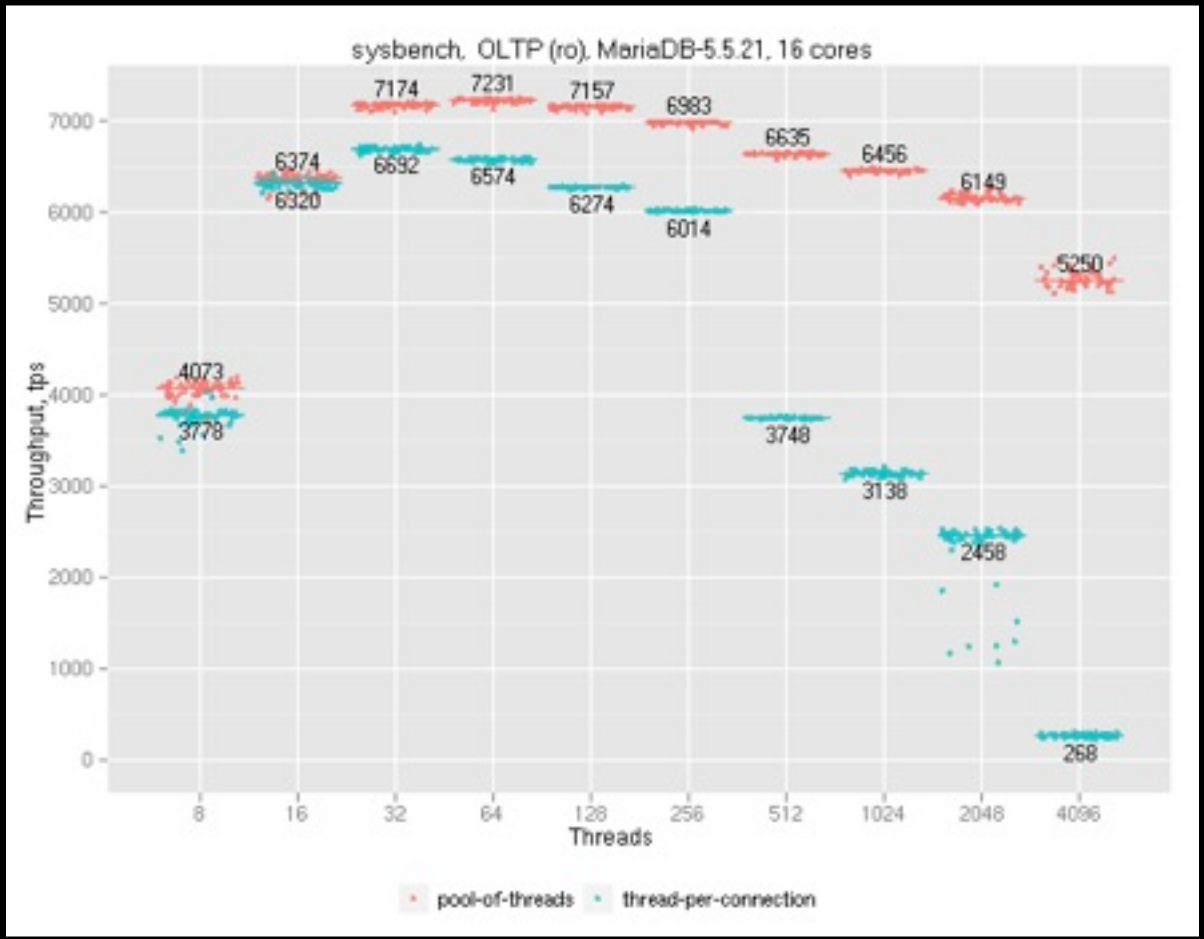


Opensource threadpool

- Modified from 5.1 (libevent based), great for CPU bound loads and short running queries

- Windows (threadpool), Linux (epoll), Solaris (event ports), FreeBSD/OSX (kevents)

- No minimization of concurrent transactions with dynamic pool size

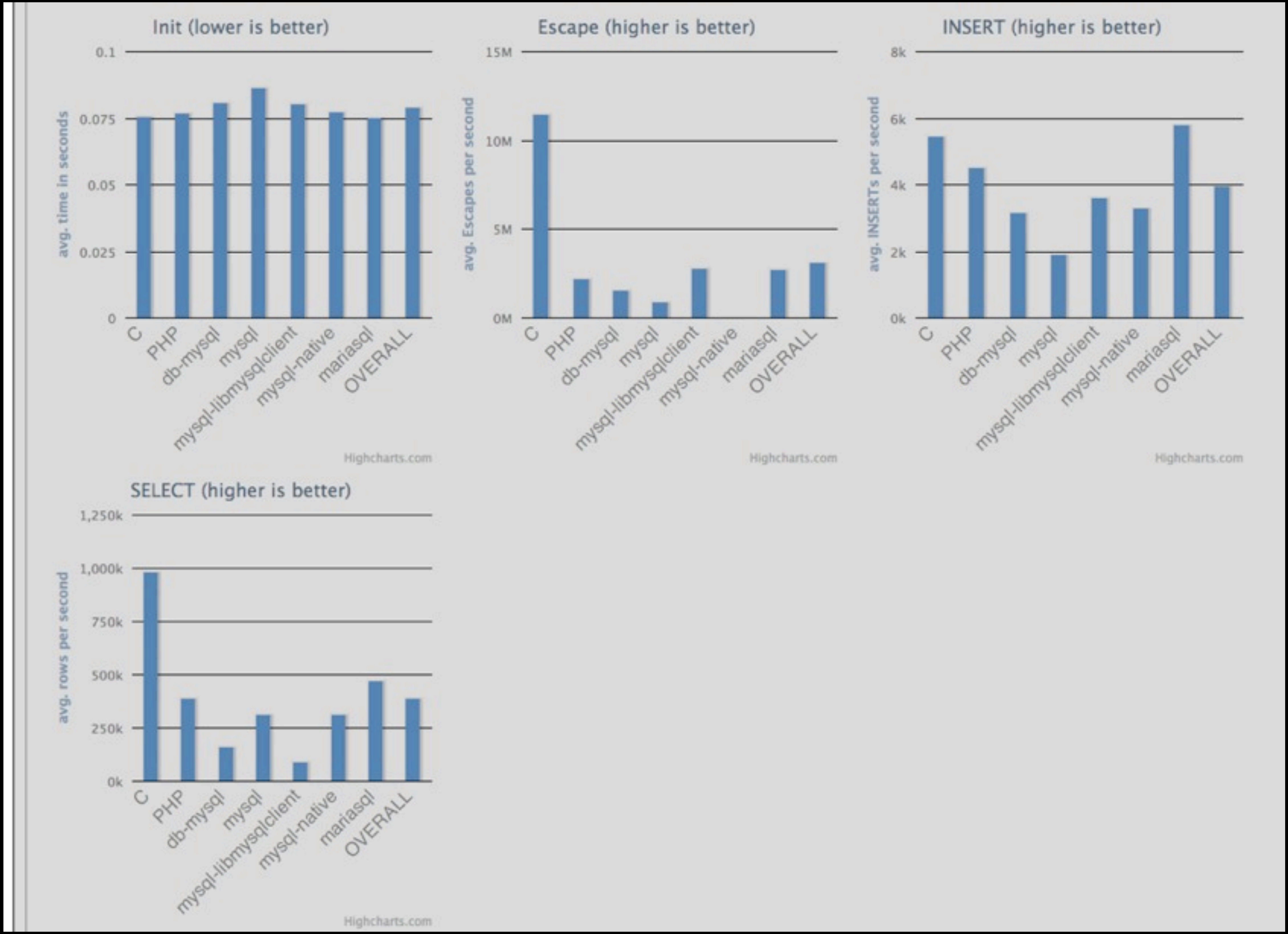


Better for DBAs

- non-blocking client library
  - start operation, do work in thread, operation processed, result travels back
  - use cases: multiple queries against single server (utilize more CPUs); queries against multiple servers (SHOW STATUS on many machines)

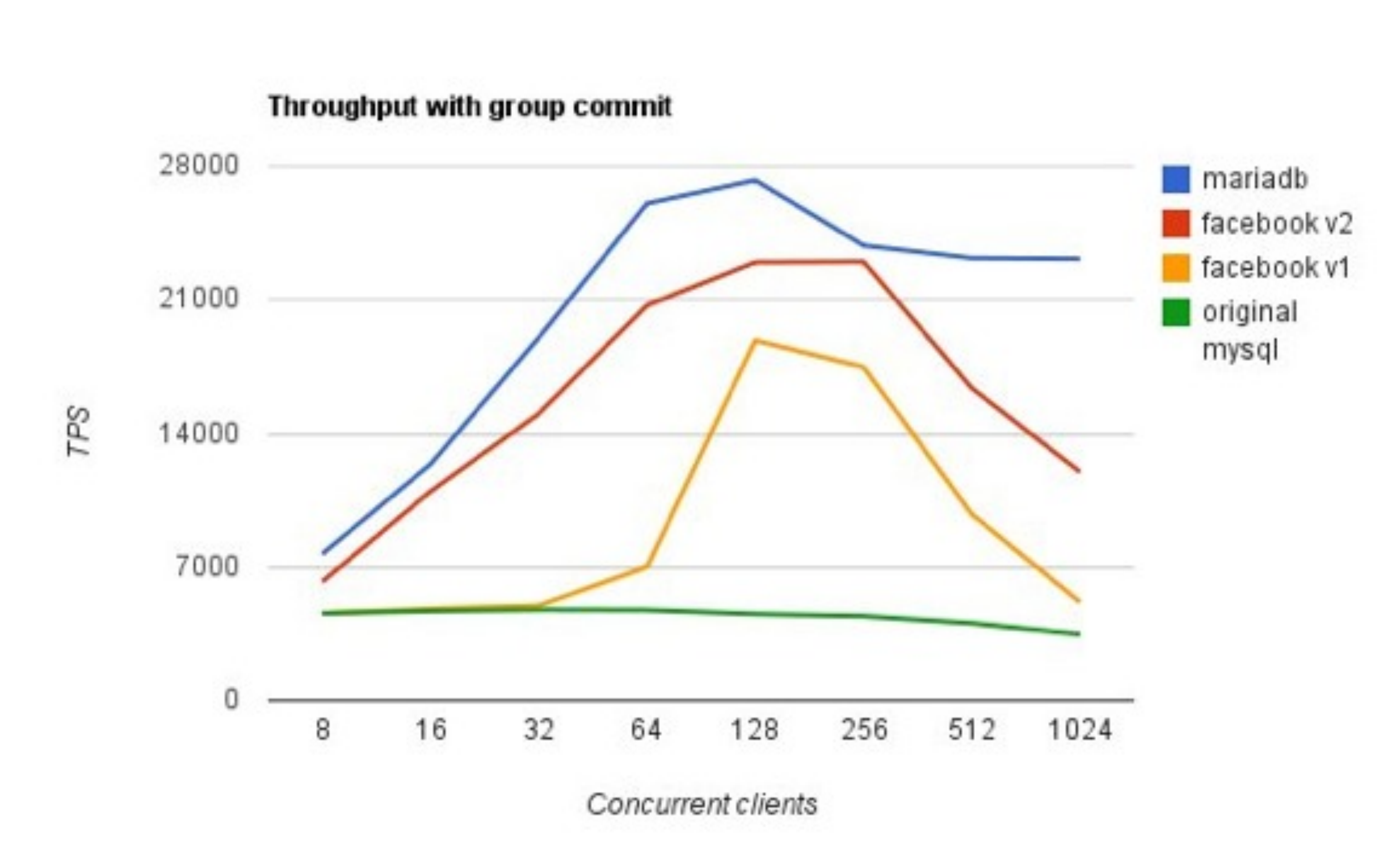
- fast node.js driver available: mariasql
- https://github.com/mscdex/node-mariasql
- SELECT now has LIMIT ROWS EXAMINED to consume less resources
  - `SELECT * from t1, t2 LIMIT 10 ROWS EXAMINED 1000;`

mariasql is fast!



Group commit in the binary log

- sync_binlog=1, innodb_flush_log_at_trx_commit=1
- https://www.facebook.com/note.php?note_id=10150261692455933
- http://kb.askmonty.org/en/group-commit-for-the-binary-log

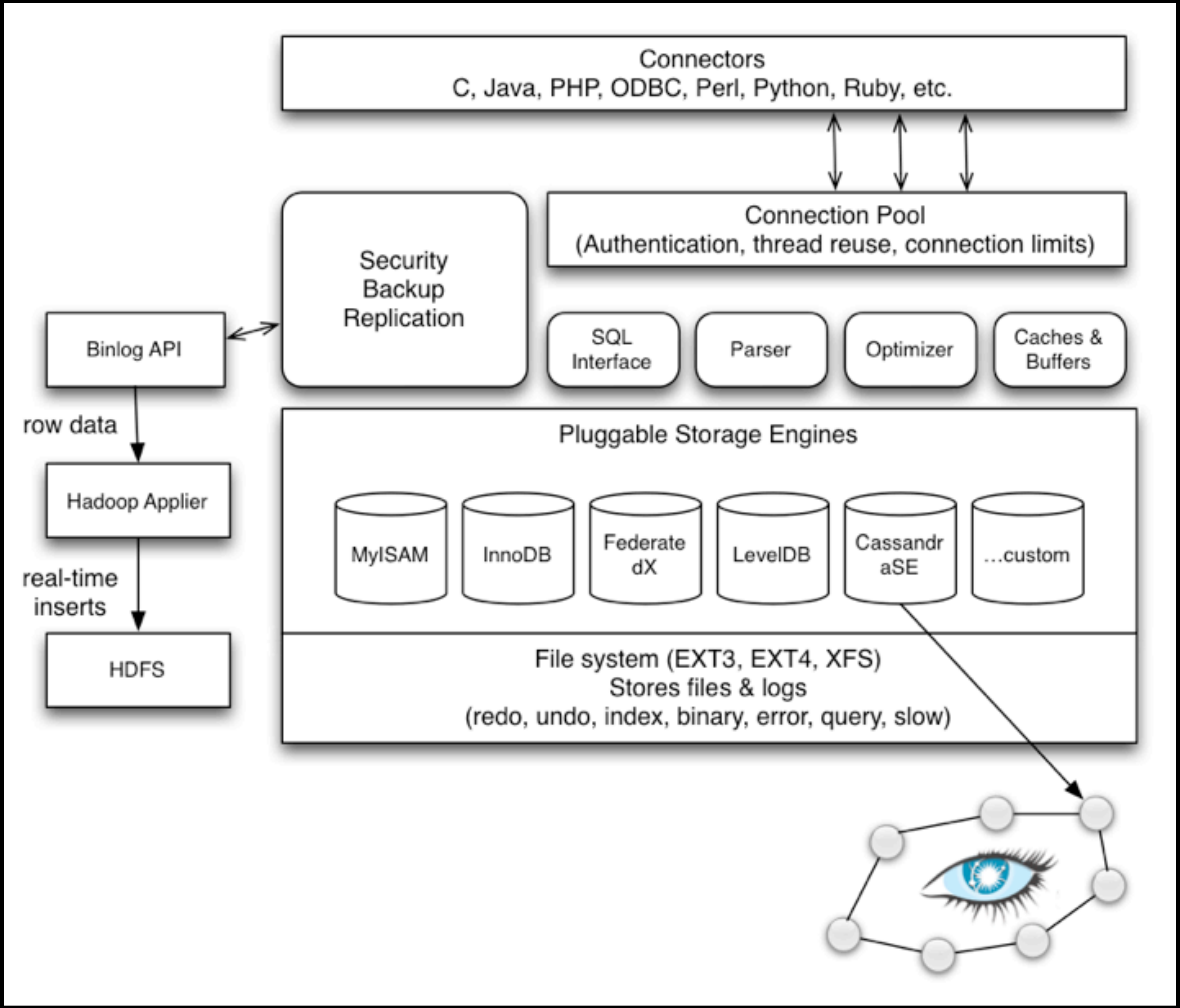


Progress reporting

- For ALTER TABLE or LOAD DATA INFILE

```
MariaDB [mail]> alter table mail engine = maria;
Stage: 1 of 2 'copy to tmp table' 17.55% of stage done
MariaDB [mail]> select id, user, db, command, state,
-> time_ms, progress from
information_schema.processlist;
+---------+-------------------+-----------+----------+
| command | state             | time_ms   | progress |
+---------+-------------------+-----------+----------+
| Query   | copy to tmp table | 23407.131 |   17.551 |
+---------+-------------------+-----------+----------+
1 row in set (0.47 sec)
```

Pluggable Engines



Connectors
C, Java, PHP, ODBC, Perl, Python, Ruby, etc.

Connection Pool
(Authentication, thread reuse, connection limits)

Security
Backup
Replication

Binlog API

SQL Interface

Parser

Optimizer

Caches & Buffers

row data

Hadoop Applier

real-time inserts

HDFS

Pluggable Storage Engines

MyISAM

InnoDB

FederatedX

LevelDB

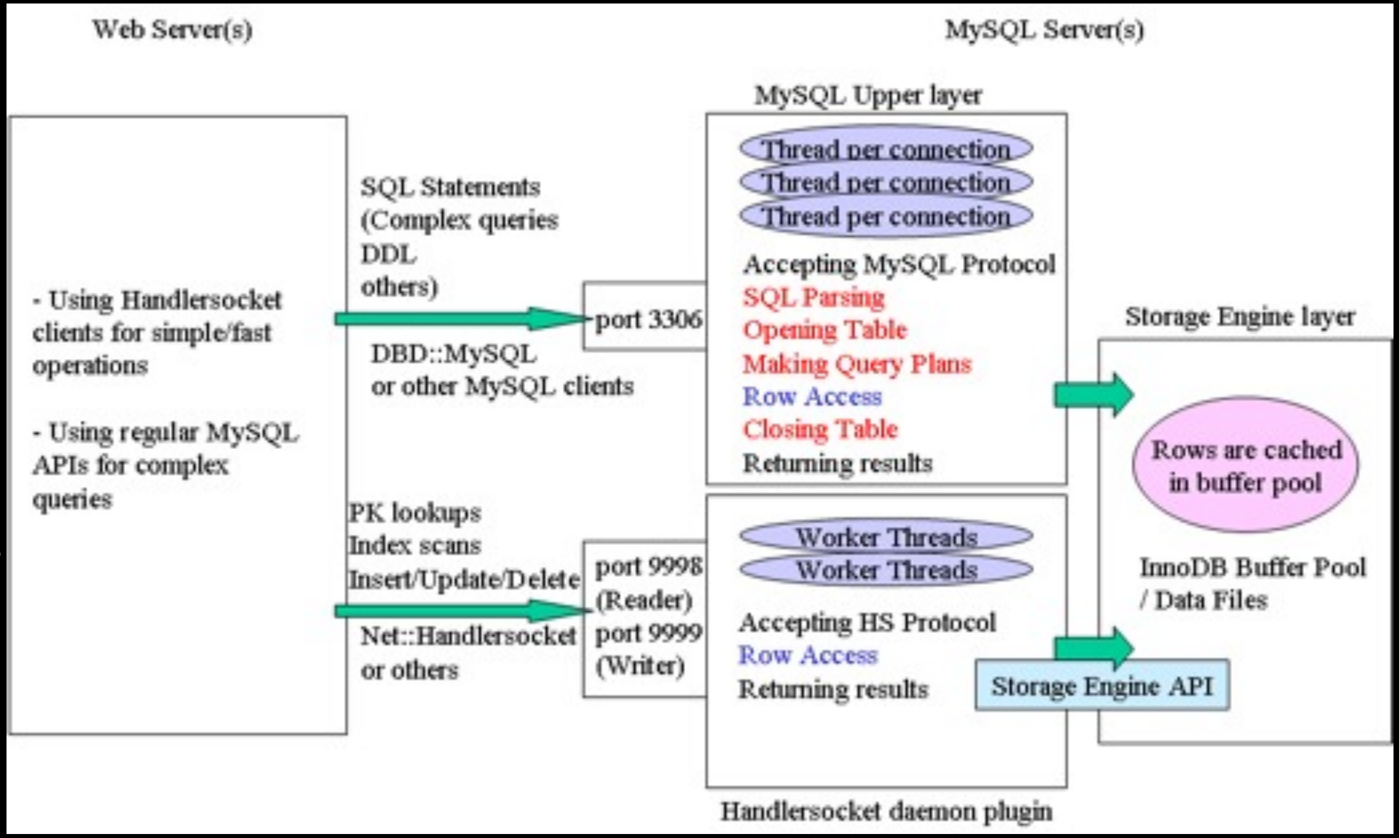
CassandraSE

...custom

File system (EXT3, EXT4, XFS)
Stores files & logs
(redo, undo, index, binary, error, query, slow)

HandlerSocket

- Direct access to InnoDB/XtraDB for CRUD operations
- SQL: 105,000 qps (60% usr, 28% sys)
- memcached: 420,000 qps (8% usr, 88% sys)
- HandlerSocket: 750,000 qps (45% usr, 53% sys)



MariaDB

Dynamic columns

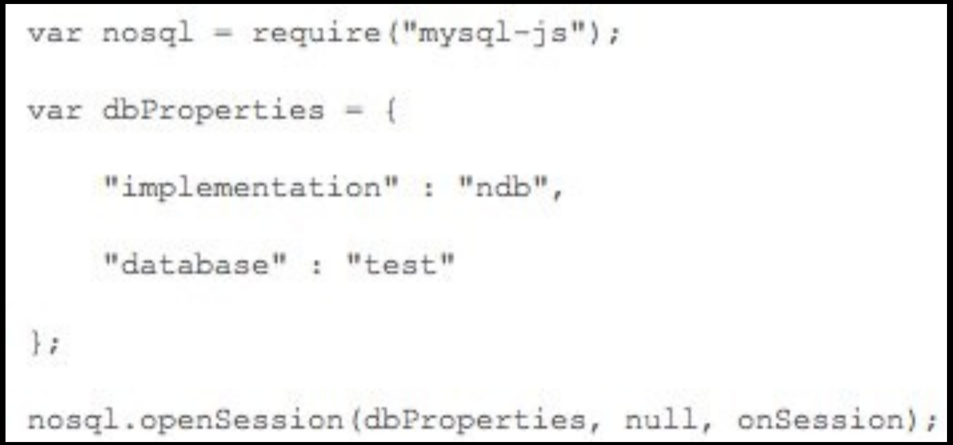
- Allows you to create virtual columns with dynamic content for each row in table
- Basically a blob with handling functions (GET, CREATE, ADD, DELETE, EXISTS, LIST, JSON)
- Store different attributes for each item (like a web store). Hard to do relationally
- In MariaDB 10: name support (instead of referring to columns by numbers, name it), convert all dynamic column content to JSON array, interface with Cassandra
- https://kb.askmonty.org/en/dynamic-columns/
- INSERT INTO tbl SET dyncol_blob=COLUMN_CREATE("column_name", "value");

memcached plugin

- innodb_memcache daemon plugin contains an embedded memcached with InnoDB as a storage backend
- Data accessed using memcache's key-value protocol
- Access different tables at same time; or same table using different columns as memcache key
- PHP? PECL mysqlnd_memache extension
- Works with NDBCLUSTER too

node.js with NDBCLUSTER & InnoDB

- end-to-end JavaScript development, from browser to server
- native access to storage layer, bypassing SQL layer
- Also asynchronous interface

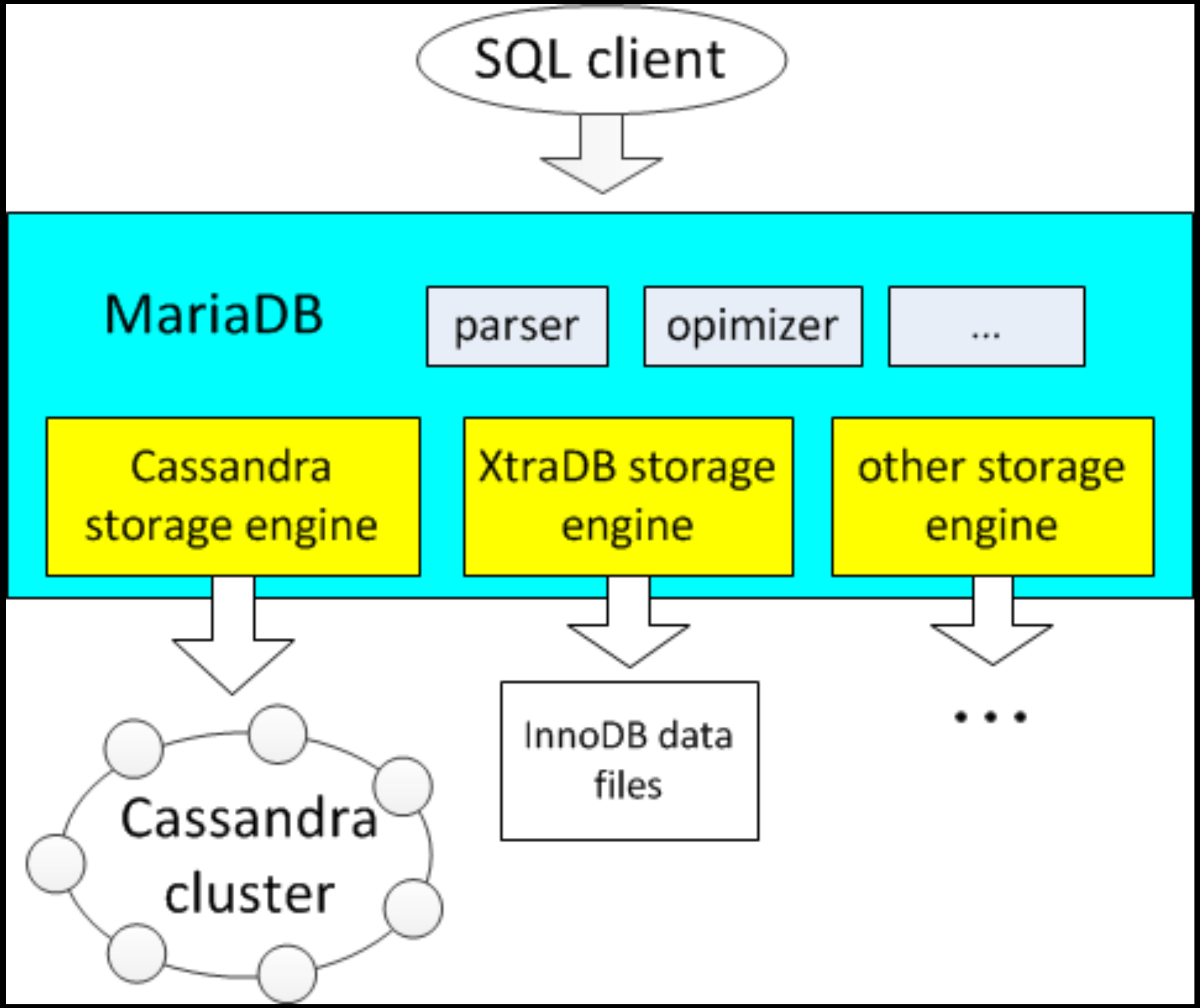


```
var nosql = require("mysql-js");

var dbProperties = {

    "implementation" : "ndb",

    "database" : "test"

};

nosql.openSession(dbProperties, null, onSession);
```

CassandraSE

- Access data in your Cassandra cluster from MariaDB
- CQL is great but now you can just work in SQL without switching (think ORMs too)

- Use cases: sensor data, log collection & analysis, form versioning, time series data, user activity tracking
- MariaDB users want a globally replicated table that's fault tolerant?



CONNECT

- CONNECT will speak XML or even grab data over an ODBC connection
- You can CONNECT to Oracle (via ODBC), join results from Cassandra (via CassandraSE) and have all your results sit in InnoDB
- Turn on engine condition pushdown

MySQL JSON import/ export

- EXPLAIN in MySQL 5.6 has JSON output now
- mysqljsonimport
- mysqljsonexport
- Export/import a database using the above commands in JSON format

LevelDB

- LevelDB: single statement transactions, secondary indexes, crash proof slave replication, non-blocking schema change, hot backup, LevelDB per mysqld/schema/ table, transaction support (read snapshots/ batch updates)
  - https://kb.askmonty.org/en/leveldb-storage-engine/

MariaDB

HBase

- HBase: Mapping cells, rows is ongoing work using the HBase API
  - https://kb.askmonty.org/en/hbase-storage-engine/
- Honeycomb
  - does automatic sharding, automatic replication & failover, map/reduce integration with Hadoop, offline map/reduce bulkoad

To recap

- node.js: mariasql, interface to NDBCLUSTER/InnoDB
- memcached: interface to NDBCLUSTER/InnoDB
- HandlerSocket: directly to InnoDB
- Dynamic columns: in-server feature
- CassandraSE: interface to Cassandra cluster
- JSON: import, export, and manipulation
- Future: HBase, LevelDB

What's wrong with MySQL?

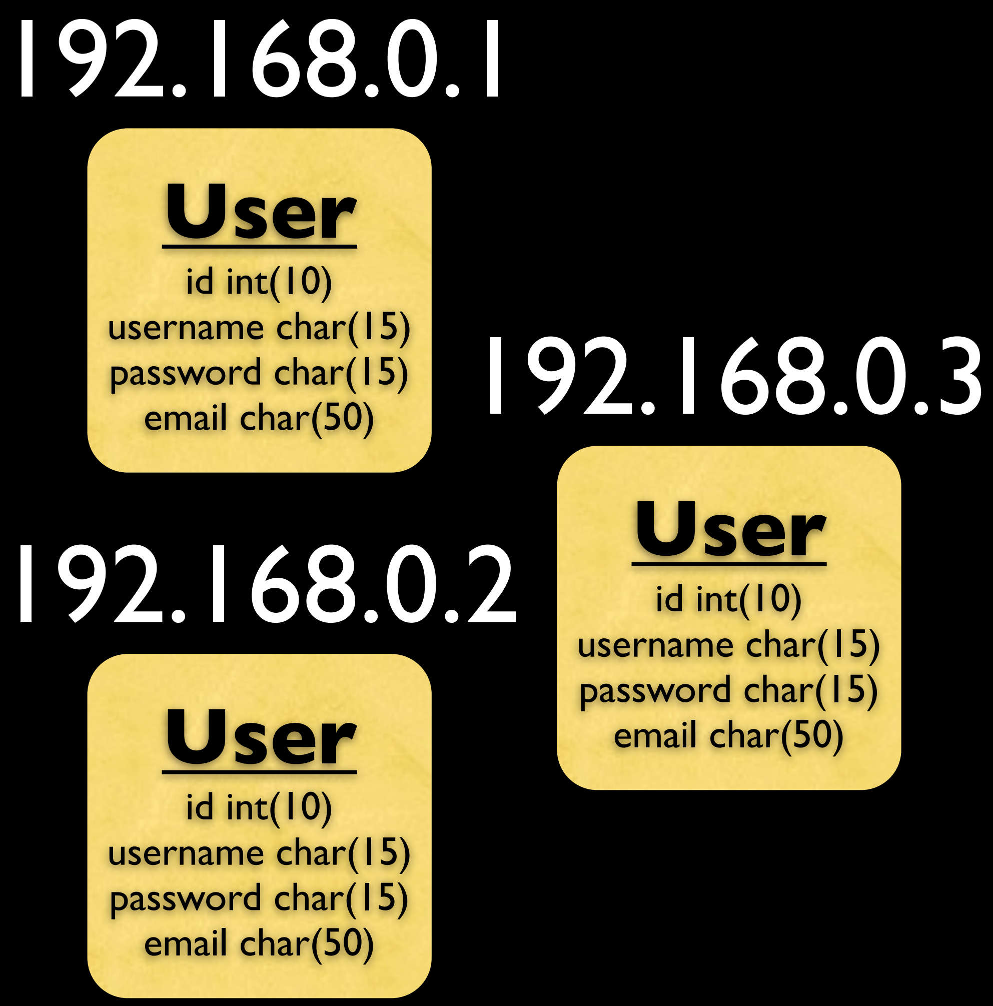
- Hard to have high availability built-in
  - need 3rd party tools like MHA/mysqlfailover
- Multi-master replication is missing (MMM)
- Out of box high scalability is missing
  - 3rd party tools like Scalr/RightScale
- Synchronous geographic replication
  - NDBCLUSTER has this; bandwidth/latency can be issues

Sharding

How do you Shard: alphabetically? n-users/shard? postal codes?

Sharding choices:

- Own framework
- SPIDER
- Tungsten Replicator
- Tumblr JetPants
- ShardQuery
- MySQL Fabric



192.168.0.1

User

id int(10)
username char(15)
password char(15)
email char(50)

192.168.0.2

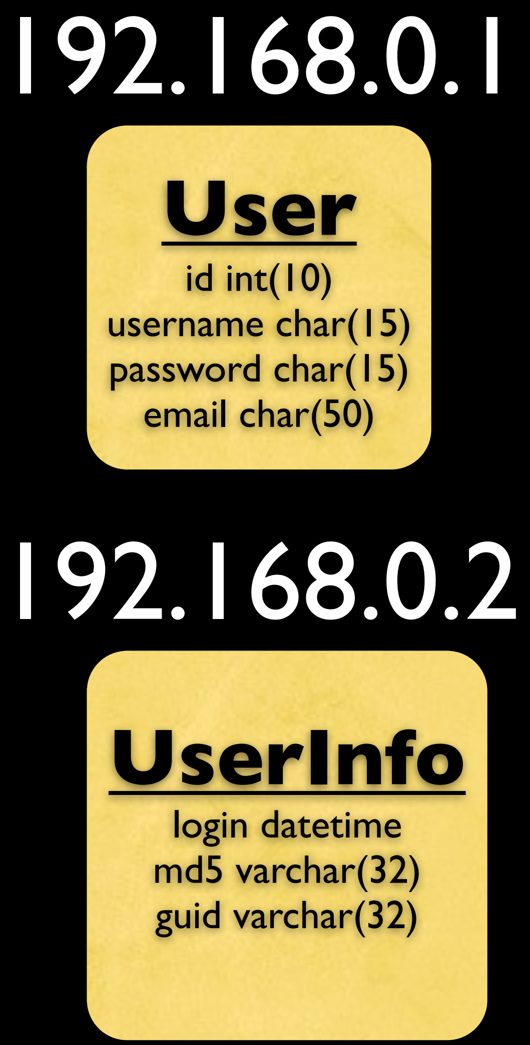
User

id int(10)
username char(15)
password char(15)
email char(50)

192.168.0.3

User

id int(10)
username char(15)
password char(15)
email char(50)



192.168.0.1

User

id int(10)
username char(15)
password char(15)
email char(50)

192.168.0.2

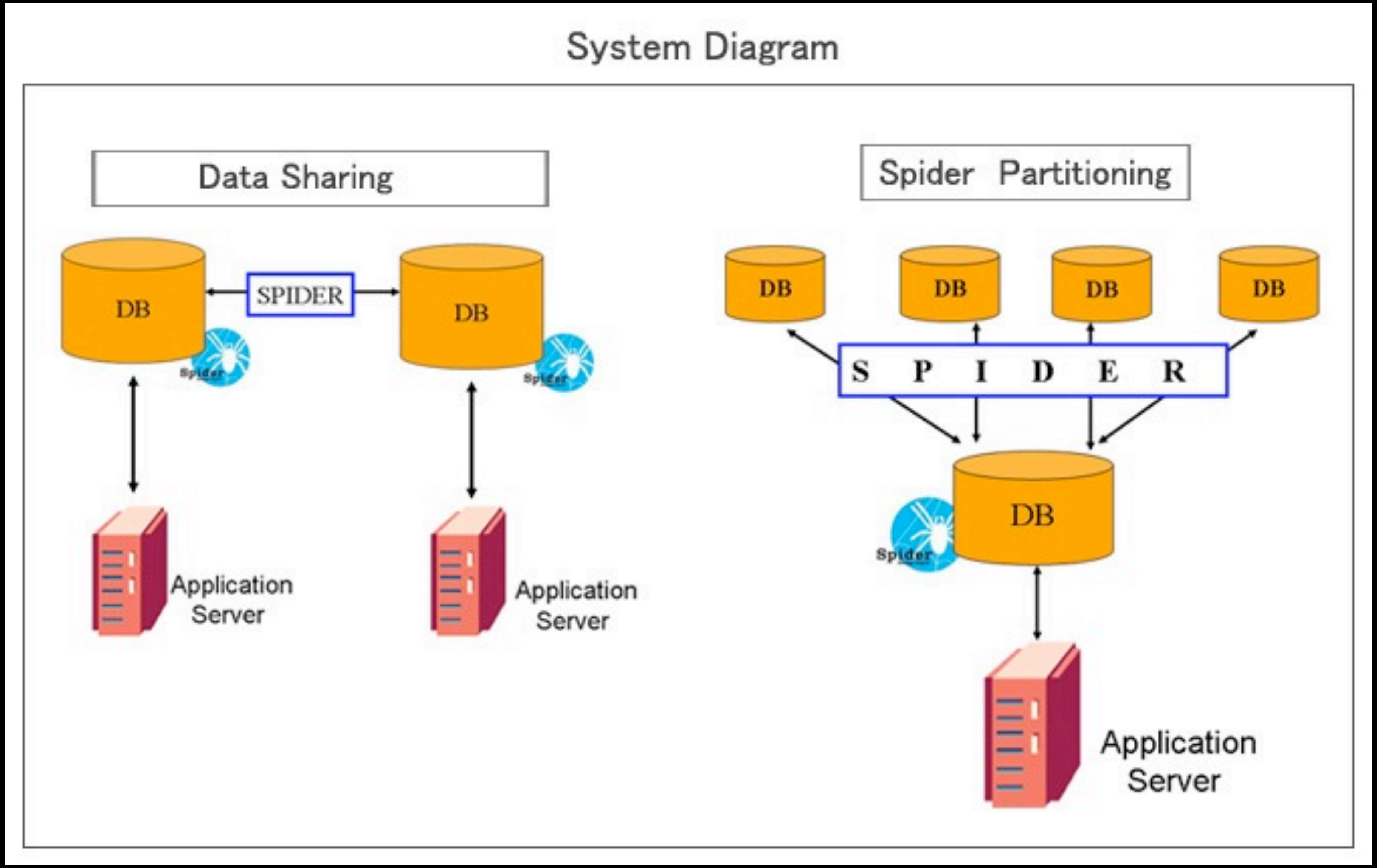
UserInfo

login datetime
md5 varchar(32)
guid varchar(32)

Better if INSERT heavy and there's less frequently changed data

SPIDER

- Storage engine to vertically partition tables



Online schema change

- MySQL doesn’t do online schema change in-server
- MySQL 5.6 & MariaDB 10.0 has this ability
- Today? External tools:
    - https://www.facebook.com/notes/mysql-at-facebook/online-schema-change-for-mysql/430801045932
    - http://www.percona.com/doc/percona-toolkit/2.1/pt-online-schema-change.html
    - http://code.openark.org/forge/openark-kit

MariaDB Galera Cluster

- High availability & scalability, with synchronous replication (transaction committed to all nodes or none), multi-master replication support, parallel replication (apply events on slaves in parallel), automatic node provisioning, data consistency (all slaves synced)
- Sounds rather NewSQL capable, no? :-)
- Load balancing with HAProxy
- Great for cloud environments, with work from Codership, Severalnines, Percona, ~~Monty Program~~ SkySQL

MariaDB deployed

happy users: **pap.fr**, **Paybox Services**, OLX, Jelastic, **Web of Trust**, Wikipedia, Craigslist, etc.

"We made the switch on Saturday -- and we're seeing benefits already -- our daily optimization time is down from 24 minutes to just 4 minutes" -- Ali Watters, CEO, travelblog.org

"MariaDB had these same bugs that we ran into with MySQL. However the big difference was that when we reported these bugs, they were quickly resolved within 48 hours!" -- Dreas van Donselaar, Chief Technology Officer, SpamExperts B.V. after migrating over 300 servers from MySQL 5.0 to MariaDB 5.1.

We upgraded the support.mozilla.org databases from Percona 5.1 to MariaDB 5.5. One of the engineers and I had a conversation where he mentioned that "one of our worst performing views on SUMO is doing waaaayyy better with the upgraded databases", that it "seems more stable" and that "I stopped receiving 'MySQL went away or disconnected emails' which came in once in a while." - Sheeri Cabral, Mozilla IT

"@nginxorg & @mariadb have helped me save $12000/year in infrastructure cost. I love it! Do more with less!" - Ewdison Then, CEO, Slashgear

"Migrating from MySQL 5.1 to MariaDB 5.2 was as simple as removing MySQL RPMs and installing the MariaDB packages, then running mysql_upgrade." - Panayot Belchev, proprietor, Host Bulgaria on providing MariaDB to over 7,000 of their web hosting customers.

Get MariaDB

- http://mariadb.org/ - comes with Multi-source replication, new engines and more
- Comes in all popular Linux distributions, plus we provide apt/yum repositories
  - default in Fedora, openSUSE, etc.
  - coming in RHEL7 (6 via Software Collections)
- http://kb.askmonty.org/v/distributions-which-include-mariadb
- Many software packages talk about us (Drupal, MediaWiki, WordPress, phpMyAdmin, Plone, etc.)

Other branches?

MySQL	Percona	MariaDB
5.5.20	7.7M	61M
5.5.22	16M	60M

MySQL	Percona	MariaDB
5.5.20	222299	1587843
5.5.22	438567	1540932

Track record

- We found a large MySQL security bug and MariaDB was first to be patched (sql/password.c & memcmp())
- We don't like regressions
  - http://www.skysql.com/blogs/hartmut/nasty-innodb-regression-mysql-5525
  - http://www.skysql.com/blogs/kolbe/heads-no-more-query-cache-partitioned-tables-mysql-5523
- We care about backward compatibility & introduce features carefully
  - XtraDB innodb_adaptive_checkpoint=none|**reflex**|estimate|keep_average (no more reflex...)

We really care about quality

- Automated test suite run upon every push
- Better QA & code coverage
- MySQL test cases: 1,765
- Percona Server test cases: 1,837
- MariaDB test cases: 2,180

In conclusion

- Use the right tool for the job
  - log analysis? Impala/Hive. Transaction management? RDBMS
- Don’t just use the “hot new tool” that “big site uses”
- Study all available solutions that provide maximum flexibility & capabilities for the job today... and in the foreseeable future

Future

- YOU define it!
- MongoDB?
- Solr? (we have SPHINXSE)
- Redis?
- Neo4J? (we have OQGRAPH)

References

- MapReduce: http://research.google.com/archive/mapreduce.html
- BigTable: http://research.google.com/archive/bigtable.html
- Spanner: http://research.google.com/archive/spanner.html
- NoSQL-Database.org
- mysql ha_memcache: http://dev.mysql.com/doc/refman/5.6/en/ha-memcached.html
- nosql to mysql with memcache: http://dev.mysql.com/tech-resources/articles/nosql-to-mysql-with-memcached.html
- node.js innodb/ndbcluster: https://blogs.oracle.com/MySQL/entry/tutorial_getting_started_with_the
- CassandraSE: https://kb.askmonty.org/en/cassandrase/
- HandlerSocket: http://yoshinorimatsunobu.blogspot.com/2010/10/using-mysql-as-nosql-story-for.html

MariaDB

Incredibly social

- Facebook: fb.com/mariadb.dbms
- Twitter: @mariadb
- Google Plus: gplus.to/mariadb
- LinkedIn groups

Au revoir

Colin Charles, colin@mariadb.org | byte@bytebot.net
http://bytebot.net/blog/ | @bytebot

SkySQL
skysql.com