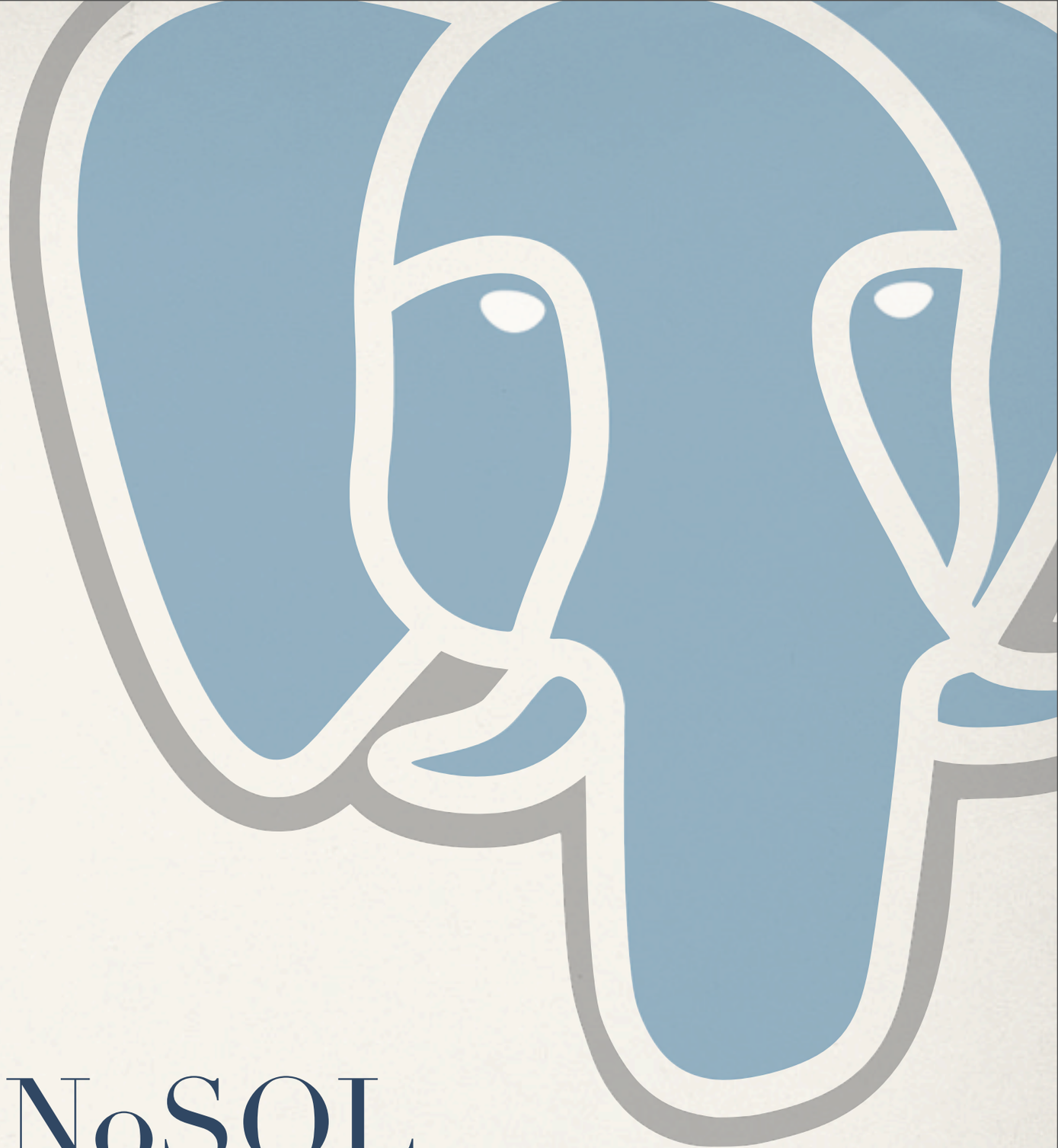


Perspectives on NoSQL

PGCon 2010

Gavin M. Roy <gmr@myyearbook.com>

What is NoSQL?

"NoSQL is a movement promoting a loosely defined class of non-relational data stores that break with a long history of relational databases. These data stores may not require fixed table schemas, usually avoid join operations and typically scale horizontally. Academics and papers typically refer to these databases as structured storage." - Wikipedia

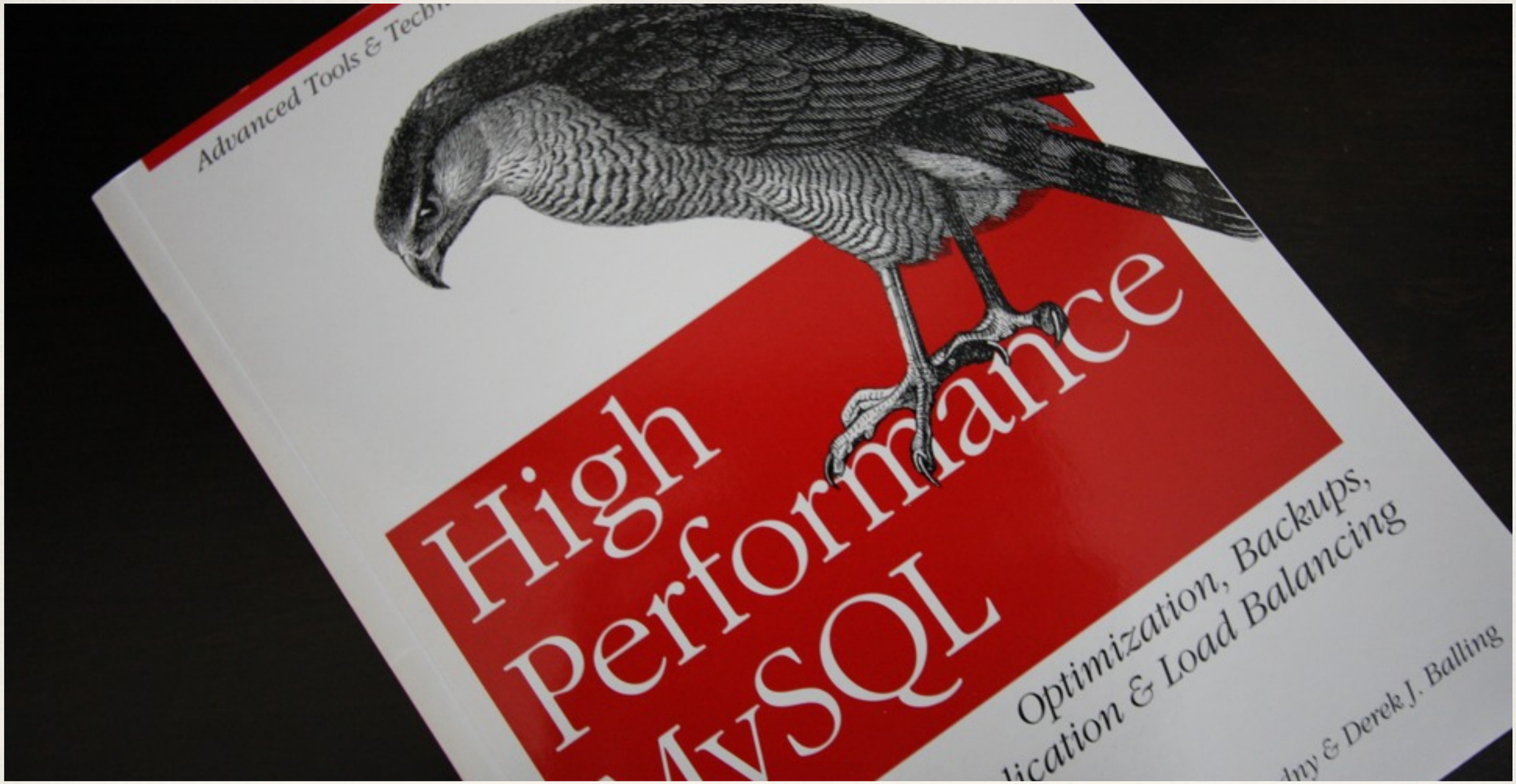
http://en.wikipedia.org/wiki/NoSQL

NoSQL: What does it mean?

- Coined as the name of a database project in 1998 by Carlo Strozzi
- Re-introduced as a general term in 2009 by Eric Evans at Rackspace
- Better phrased at No Relationships or No ACID?
- Departing from traditional RDBMS technology to scale large traffic websites

What's new about NoSQL?

- Key/Value data stores are not new
  - DBM (1979), BerkelyDB (1986)
- Document Databases are not new
  - Lotus Notes (1989)
- The use case is new...



Oxymoron

Scaling large database driven websites is hard

Impetus for NoSQL Databases

memcached + RDBMS

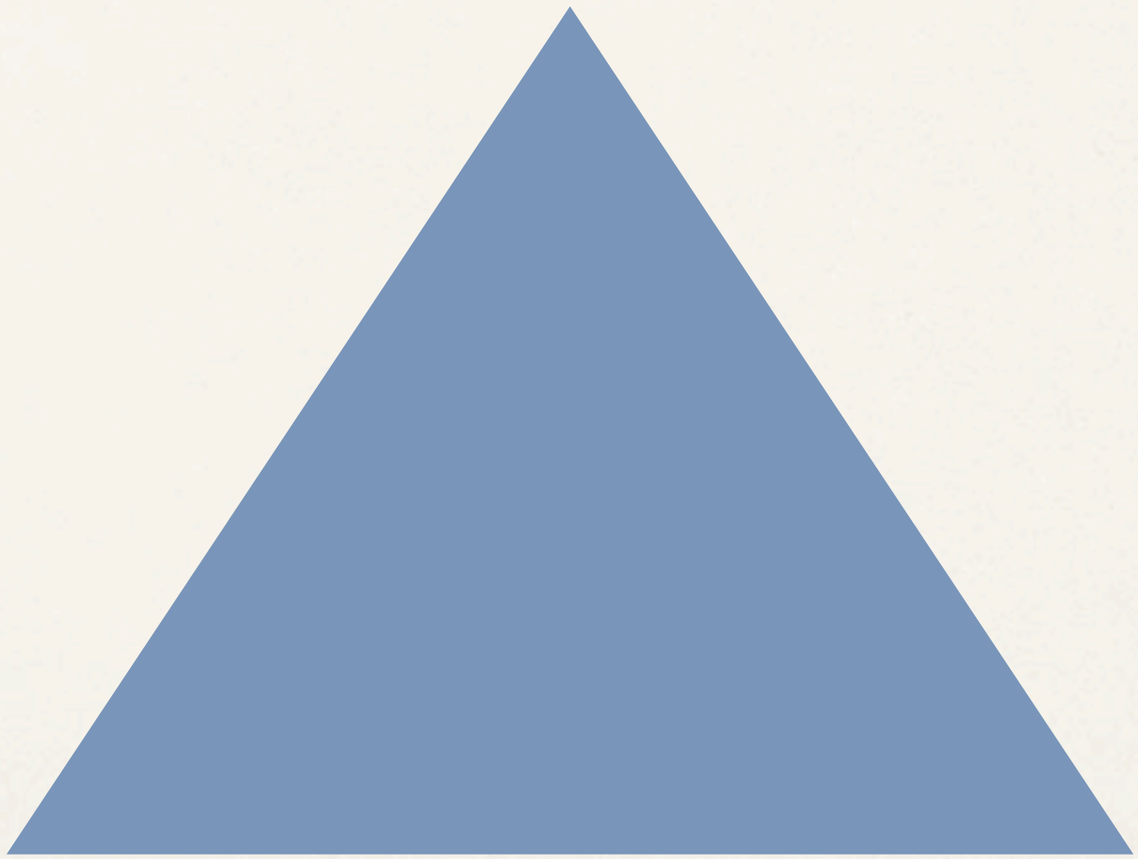
Commodity Hardware

"I know of one company that's managing to scale portions of their PostgreSQL servers by purchasing $250,000 servers. This would cover my 50 node EC2 cluster for two years."

- Joe Stump, SimpleGeo

Brewer's CAP Theorem: Pick Two

Consistency



Availability

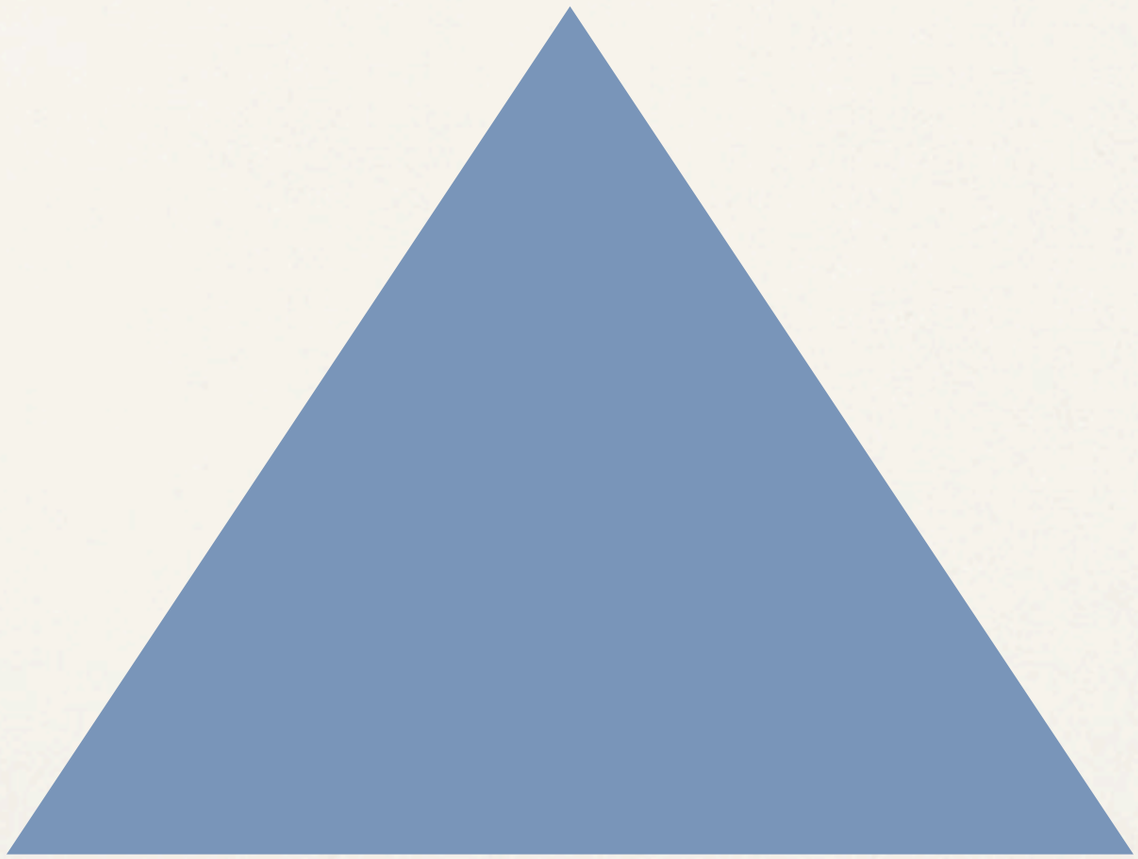
Partition Tolerance

"No set of failures less than total network failure is allowed to cause the system to respond incorrectly"

- Seth Gilbert & Nancy Lynch, 2002

CAP Theorem and PostgreSQL

Consistency

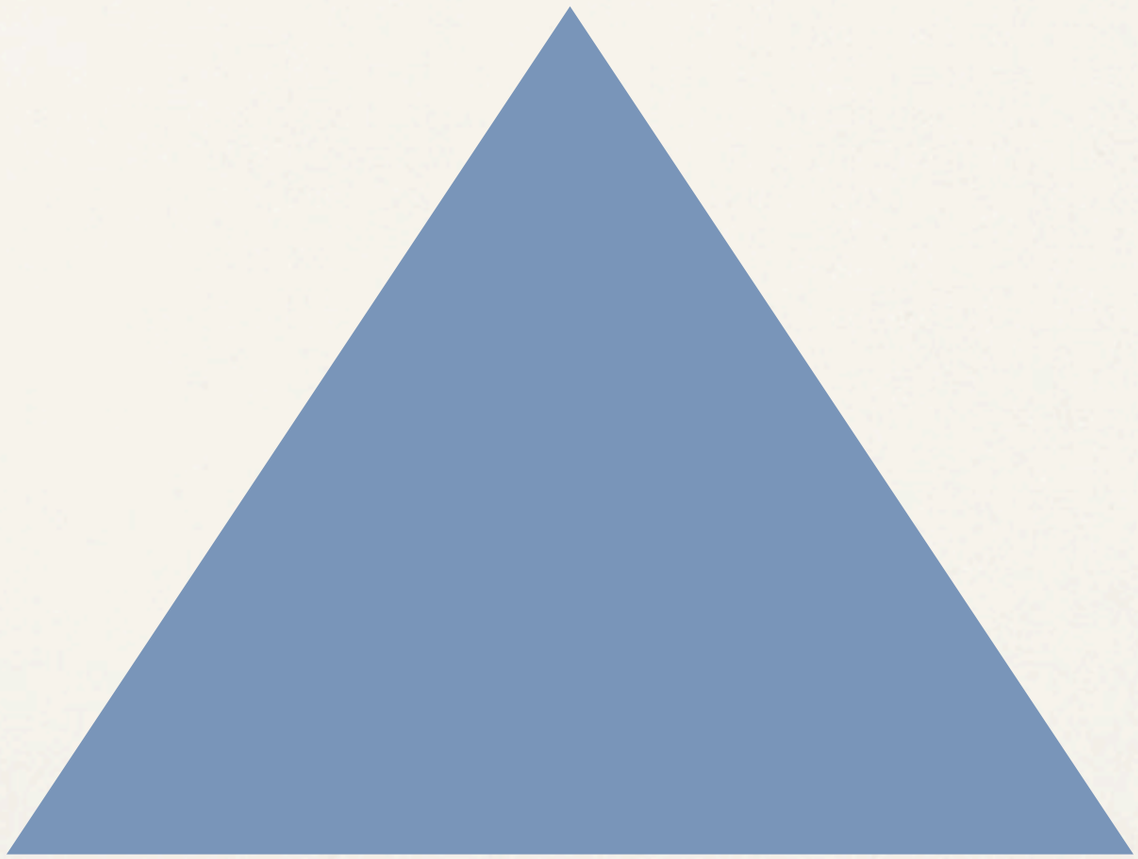


Availability

Partition Tolerance

CAP Theorem and NoSQL

Consistency



Availability

Partition Tolerance

fsync = off

a.k.a. Database Administrators running with scissors


Cassandra

Apache Cassandra

- Developed initially at Facebook, used at Digg and others
- Column-oriented key/value store
- Durability via Commit Log, similar to WAL
- Eventually Consistent across nodes
- Focuses on Availability and Partition Tolerance
- Talks via the Thrift protocol, query via Map-Reduce


CouchDB
relax

Apache CouchDB

- Major deployments include the BBC and Engine Yard
- Key/Value Document Store
- Durability by append-only use
- MVCC
- Focuses on Availability and Partition Tolerance
- Talks JSON via HTTP REST
- Query by document or JavaScript functions (Map-Reduce)


mongoDB

MongoDB

- Major deployments include Sourceforge, Foursquare, Bit.ly, & Github
- Key/Value Document Store
- No durability without replication
- In place updates
- Focuses on Consistency and Partition Tolerance
- Data stored in BSON (Binary JSON)
- Own communication protocol

MongoDB & Ad-Hoc Queries

SELECT * FROM things WHERE j != 3 AND k > 10;

db.things.find({j: {$ne: 3}, k: {$gt: 10}});

MongoDB & Ad-Hoc Queries

SELECT * FROM things OFFSET 10 LIMIT 10;

db.things.find().skip(10).limit(10);

Project Voldemort

- Developed at LinkedIn
- Key/Value Document Store
- Durability at pluggable data storage layer (BerkeleyDB)
- Focuses on Availability and Partition Tolerance
- Multiple serialization formats for data
- Own communication protocol


Redis

Redis

- Development sponsored by VMWare
- Key/Value Document Store
- In-memory database with durability via snapshots
- Focuses on Availability and Partition Tolerance
- No set data serialization format
- Own communication protocol similar to POP3

Tokyo Cabinet/Tyrant

- Key / Value DBM Implementation and Network Daemon
- Durability via WAL and Shadow Paging
- Focuses on Availability and Partition Tolerance
- No set data serialization format
- Protocols: Tokyo Tyrant Binary Protocol, memcached Compatible Text Protocol, HTTP REST

NoSQL Solves Scaling Problems...

- NoSQL users use memcached too!
  - Reddit uses memcached in front of Cassandra
    http://blog.reddit.com/2010/05/reddits-may-2010-state-of-servers.html
  - Twitter uses memcached in front of Cassandra
    http://nosql.mypopescu.com/post/407159447/cassandra-twitter-an-interview-with-ryan-king
  - Facebook uses memcached in front of Cassandra
    http://www.facebook.com/note.php?note_id=39391378919#!/notes.php?id=9445547199
  - BusinessInsider.com uses memcached in front of MongoDB
    http://journal.uggedal.com/tags/mongodb

YeSQL

How does PostgreSQL stack up?

PostgreSQL Scales

- Hot Standby + WAL Streaming adds native read only active standby
- Horizontal scaling via sharding using plProxy
- Scales with hardware
- Projects like pgPool-II and GridSQL
- New projects like PostgreSQL-XC
- Connection pooling (pgBouncer, pgPool-II)
- Replication: Londiste, Slony, Bucardo, Mammoth Replicator

KVPBench

- http://github.com/gmr/kvpbench
- Python app
- Benchmarked in OS X 10.6.3
  - 2.8 GHz Intel Core i7
  - 8GB DDR3 Ram
  - 1 Seagate 7200 RPM 1TB Drive

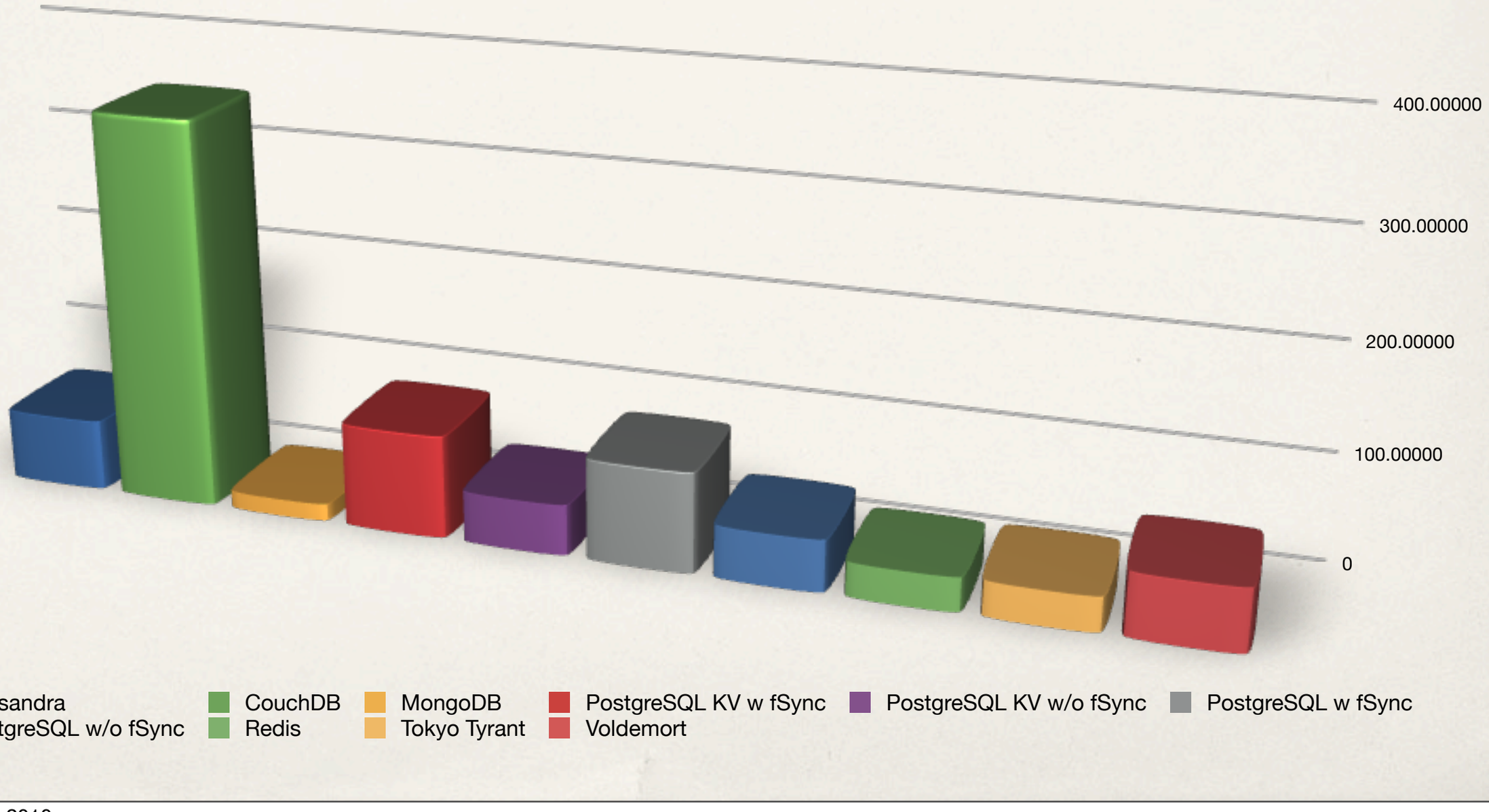
KVPBench Details

- Supreme Court Decision Data
- Default Configurations
- Load Test: 100% write, individual inserts, 98k rows
- Random Work: 75% Read, 25% Read+Write (Update)
  - 10 Backends
  - 10k Queries per Backend

Databases Tested

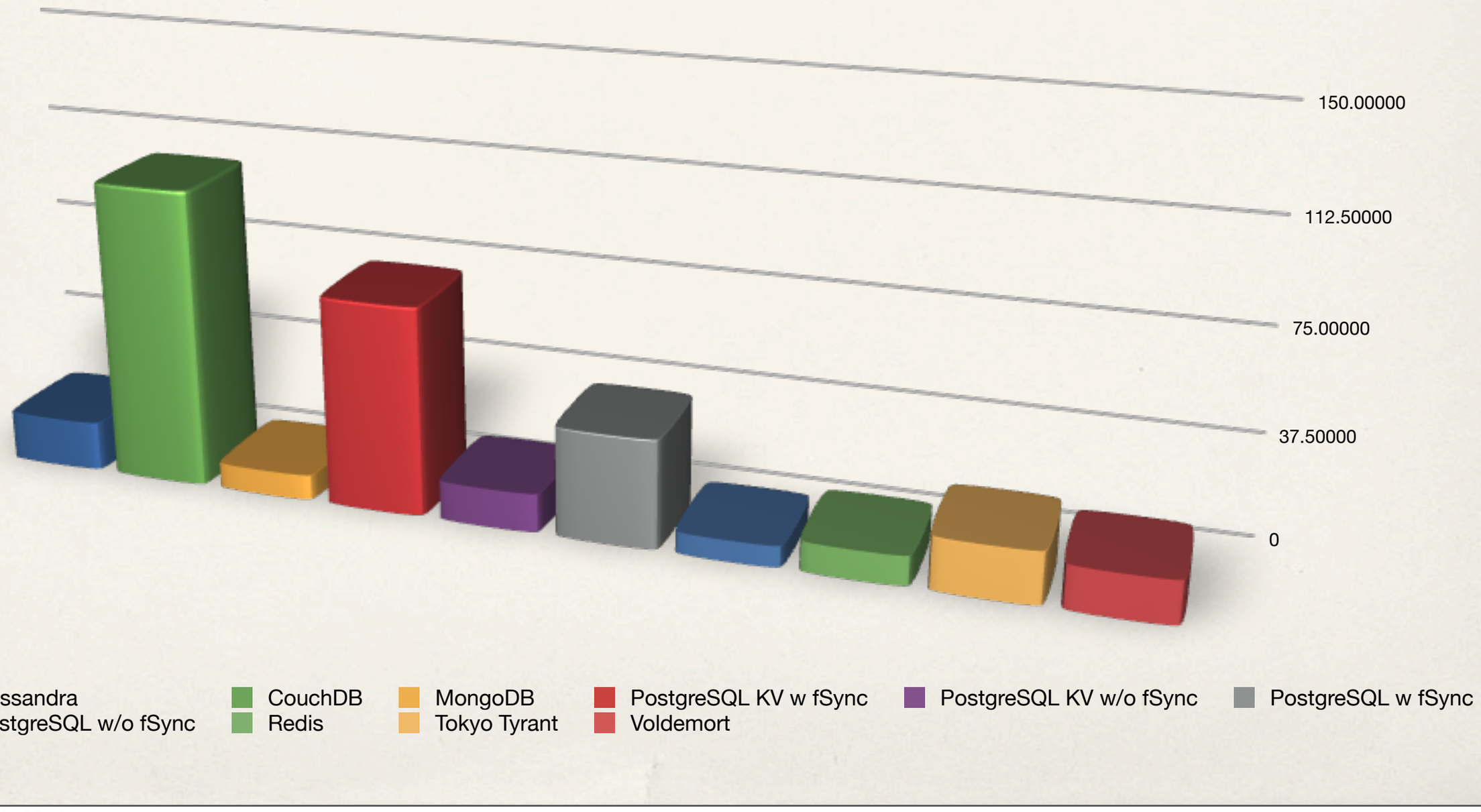
- Cassandra 0.6.1
- CouchDB 0.11.0
- MongoDB 1.4.2
- PostgreSQL 9.0b1
- Project Voldemort 0.80.2
- Redis 1.2.6
- Tokyo Tyrant 1.4.44

KVPBench Load Times



KVPBench Random Workload

10k Rows, 10 Workers, Random Workload



YeSQL?

- Amazing community of talented developers who know the subject matter deeply

- Mature code-base that is flexible enough to be a stable platform for innovation

- Fast, stable and well documented code base

- Performs as well as (and outperforms some) NoSQL databases while retaining deeper flexibility

Questions?

Follow me on Twitter:

http://twitter.com/crad

Rate this talk:

http://bit.ly/cF04fX

Email me:

gmr@myyearbook.com