



Cloud Datastore: A NoSQL Database at Google Scale

Randy Shoup
Engineering Director, Google Cloud Platform

NoSQL-Search RoadShow SF
June 6, 2013

- Google Cloud Platform
- Google Scale
- Google Cloud Datastore
- Google Storage Infrastructure
- Parting Thoughts

The image features the Google Cloud Platform logo, which consists of a solid blue rectangular area at the bottom. Above this area, there are stylized, light blue and white clouds. The text "Google Cloud Platform" is centered in a blue, sans-serif font. A thin vertical line is positioned to the left of the text.

Google Cloud Platform

Google builds and operates one of the largest computing infrastructures in the world ...

- Dozens of data centers located around the world
- Designed from the ground up to run massive Internet-scale services
- Integrated design of facility and computing machinery
- Homogeneous hardware and system software
- Cluster-level networking fabric

All Google Computing is Cloud Computing ...

- Custom-built machines and network
- Cluster is typically thousands of machines
- Common pool of resources with central cluster management
 - Fungible units of compute, memory, storage, network
 - Sophisticated bin-packing to maximize utilization
- Hundreds to thousands of active jobs, from one task to thousands of tasks
- Mix of low-latency, user-facing jobs and batch workloads
- Massively multitenant



Google Compute Engine

...

Full Linux virtual machines running on Google's infrastructure.



Google Cloud Storage

...

Store, access, and manage application data.



Google BigQuery

...

Analyze terabytes of data in seconds.



Google App Engine

...

Platform as a Service: Powerful, scalable application development and execution environment.



Google Scale

Layering and Composition

- Compose complex systems from simple primitives
- As much as possible, make it possible to reason independently and intuitively about behavior of primitives
- All Google services rely on (often many!) lower layers of infrastructure

At Scale, Everything Breaks

- Service-level outages
 - Networking
 - Power
 - Oops
- Node-level outages (industry average)
 - >1% uncorrectable DRAM errors per server per year
 - 2-10% disk drive failure rate per year
 - ~2 crashes per server per year
 - >1 utility event per year

=> 2000 node service sees 10 server crashes per day (!)

Predictable Performance

- Systems at scale highly exposed to performance variability
 - Imagine an operation ... 1ms latency median, but 1 second latency at 99.99%ile (1 in 10,000)
 - Service using 1 machine -> 0.01% slow
 - Service using 5K machines -> 50% slow
- Consistent performance trumps low average latency
 - Low latency + inconsistent performance != low latency (!)
 - Far easier to program for consistent performance
 - Tail latencies are *much* more important than average latencies

Opinionated Platform

- Encourage scalable development practices
 - Small discrete units of processing
 - No single points of failure
 - Automated testing
 - Staged deployments
- Make it easy to do the right thing, and hard to do the wrong thing
- ==> "It Just Works" (TM)



Google Cloud Datastore

Based on High Replication Datastore in Google App Engine

- Multiple generations of evolution
 - Originally introduced with Google App Engine in 2008
- 3M+ applications, 300K unique developers
- Petabytes of storage
- 4.5T+ operations / month
- Layered on top of
 - MegaStore
 - BigTable
 - Colossus

Accessible

- RESTful interface
- HTTP with JSON or Protocol Buffer API
- Accessible from
 - Google Compute Engine
 - Google App Engine
 - Anywhere else
- Web-based interface for configuration and management
- Development server for local development

Fully Managed

- No planned downtime
 - Completely automated failover
- Replicated across multiple data-centers
 - All data replicated across multiple disks and multiple data centers
- Managed and operated as a service by Google
- 99.95% SLA

Scalable

- Arbitrary horizontal scaling
- Autoscales as traffic increases
- Autosharding as data increases
- More distributed as more data is stored

Resilient

- Cross-data center active-active replication
 - All data replicated across multiple disks and multiple data centers
- Synchronous replication via Paxos
- Application can seamlessly migrate between data centers with no data loss
- Applications can read locally in separate data centers with no inconsistency or replication lag
- Resilient to catastrophic failure ("meteorite durability")

Schemaless

- No configuration needed; just start writing data
- Arbitrary attributes on any entity
 - Different entities can have different attributes
 - Attributes can be multi-valued
- Arbitrary-depth parent-child relationships
- "Entity groups" can associate many related entities
 - E.g., all emails for a user

Consistency

- Strongly consistent, with atomic transactions
- Strong serial consistency within entity group
 - Will always Get an entity once Put
 - Never see partial transactions
- Strong consistency on reads and ancestor queries
- Multi-entity group transactions
 - Transactions can read / write entities within (small number of) entity groups
- Eventual consistency only when querying across many entity groups

Rich Query Features

- GQL is an ever-growing subset of SQL
- Filters
 - Equality (=, IN)
 - Inequality (!=, <, <=, >=, >)
 - AND, OR, NOT, sub-expressions
- Sort
- DISTINCT
- Projections, index-only queries
- Geo radius, Date range
- Cursors for paged iteration

Predictable Performance

- Fixed cost queries
 - Query latency scales in the size of the result set, not in the size of the overall data
 - Constant latency for queries over 1M or 1B or 1T entities
- All queries are index queries
- *"It's not a limitation, it's a discipline"*

The image features a decorative background with a white top half and a blue bottom half. The boundary between them is a scalloped edge with light blue, semi-transparent clouds. The text 'Google Storage Infrastructure' is centered in the white area.

Google Storage Infrastructure

Next-generation clustered file system, successor to GFS

- Exabyte scale global storage system
- Automatically sharded metadata layer
- Data blocks for a given stripe replicated to multiple different fault domains
 - Different disks, servers, racks
- Blocks distributed across entire cluster
 - Easy to load-balance reads
 - Efficient to recover

"You know you have a large storage system when you get paged at 1 AM because you only have a few petabytes of storage left." -- Google Engineer

Cluster-level structured storage

- Distributed multi-dimensional sparse map
 - (row, column, timestamp) -> cell contents
- Layered on Colossus for file storage
- Automatically splits and rebalances tablets based on size and load
- Fault-tolerant within data center
- Asynchronous, eventually-consistent replication

"If you look at every NoSQL solution out there, everyone goes back to the Amazon Dynamo paper or the Google BigTable paper" -- Jason Hoffman, Joyent

Geo-scale structured database

- Layered on BigTable for structured storage
- Multi-row transactions across machines
 - Strong ACID consistency within fine-grained partitions ("entity groups")
- Eventual consistency across partitions
- Synchronous cross-datacenter replication via Paxos
- Transparent failover



Parting Thoughts

"One Size Does Not Fit All"

- Everything is a tradeoff
 - Data structures are fundamental to performance and features of any storage system
 - No data structure can optimize for every possible use-case
- Polyglot persistence is expected
 - Column stores for analytics
 - Inverted indexes for search
 - Simple key-value stores
 - Scalable, powerful NearSQL systems
- We use everything at Google (!)

Scale Depends On ...

- Discipline, not permissiveness
- Sharing, not coupling
- Architecture, not language or programming environment
- Simplicity and elegance, not complexity

Questions?

and ... We are hiring!

<https://cloud.google.com/>

rshoup@google.com