



Real Time Data Analysis

November 30, 2012 // HBDC, Beijing, China

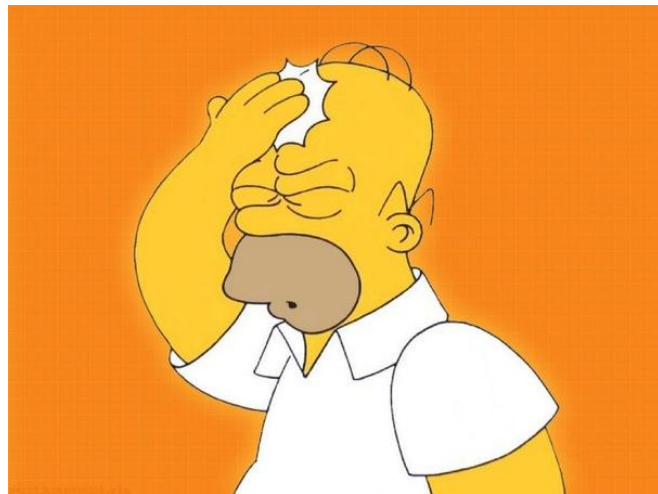


Nikita Shamgunov, CTO

- BS, MS, PhD in CS
- 8 years as a Senior Database Engineer at Microsoft SQL Server, Facebook, MemSQL



- Moor's law is over
- But not for data growth
- All kinds of data
 - Log
 - Image
 - JS
 - Structured



- Every mega successful company is a data driven company
- Google, Facebook, Amazon are obsessed with it
- What they are doing now, everyone will be doing in 5 years

- The data you've collected recently is usually more important than the data you collected a year ago
- And the value drops exponentially
- **Half Life of Data**



- Large web destination
 - How does the website perform in every country
 - What is the 99% page load time.
 - How does it correlate with revenue?

- We ship code every week
 - Which commits are regressing the key metrics
 - How can we pinpoint what the problem is?
- I want to track the performance of every little function and act upon my insights

 facebook zynga

- I want to perform A/B testing and serve ads out of a data store
 - I want to record every impression and every click and make decisions about it in real time.
 - How does it correlate with revenue?

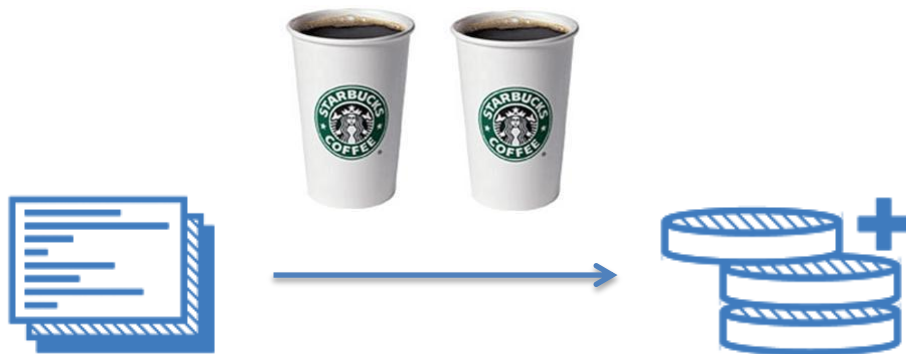
- How to store the state of multi-threaded applications and cope with faster-moving data streams?

- I want to train my models as fast as possible and test them immediately
 - I need to collect data and push it through a model using convenient tools
 - Once the model is ready I want to use it to make real time decisions when serving web pages.

- I wish I had a faster machine
- I wish I had a faster machine
- I wish I had a faster machine



■ Loading data for analysis is painful.



- Queries take too long to run
- The system cannot handle query volume
- Cannot sustain predictable performance levels



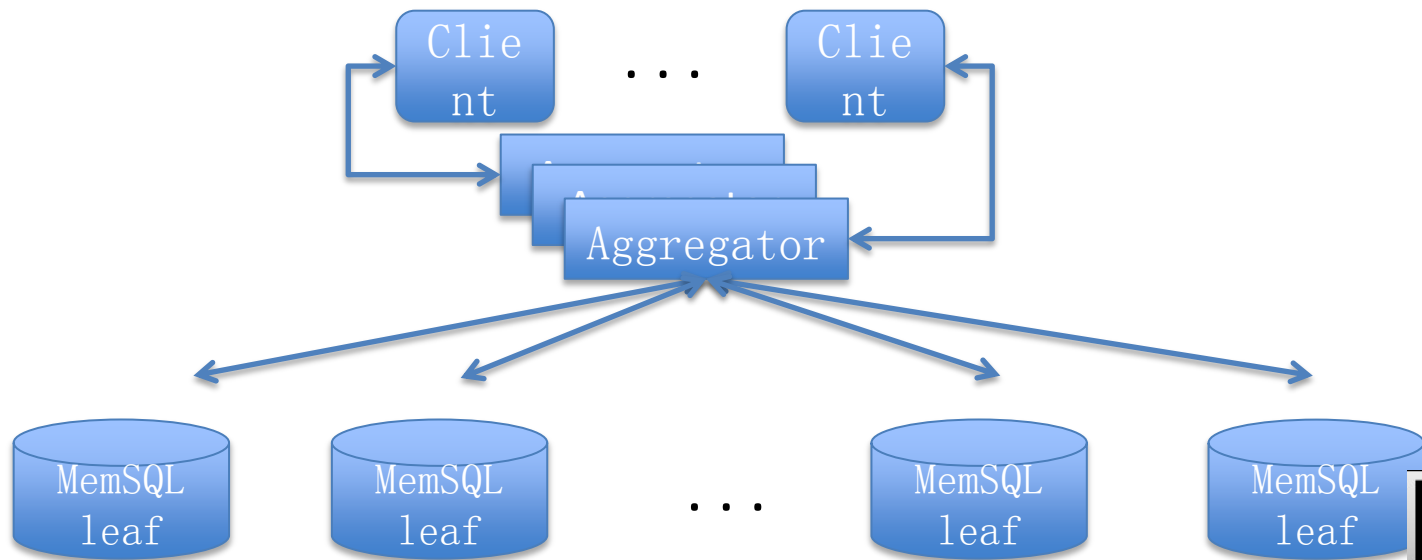
- Storm by Twitter (Nathan Marz)
- Cloudera Impala
- MemSQL

- MemSQL is a distributed, in-memory SQL database
- Capable of processing and analyzing the most demanding of workloads
- Two things we fix:
 - Data latency (the batched load)
 - Query latency



- For **data latency**, MemSQL provides
 - Ultra-fast data load
 - Real-time stream capture
- For **query latency**, MemSQL provides
 - Distributed query execution
 - Efficient SQL-to-C++ conversion
 - Lock-free data structures

- Shared-nothing architecture
- Distributed query optimizer
- Highly available through leaf-node replication
- Uses hash-partitioning



- Logging and snapshotting to disk
- No buffer pool, hence sequential IO only
 - Random read/write in RAM
 - Sequential IO on disk
- Native MemSQL replication
 - Ships snapshot to provision, then reads from transaction log
 - Skinny log – no indexes, which are reconstituted on recovery



- Consume live application data
- Issue complex, ad-hoc queries
- 48-server cluster on EC2
 - 384 cores
 - 2.7 TB of capacity in RAM



DEMO TIME.



CONTACT ME
nikita@memsql.com

WEB
www.memsql.com

380 10th ST
San Francisco, CA 94103

200 Park S Ave
New York, NY 10003

