

# Building Scalable Data Collection

cron based db insertion sucks



# Steal good ideas from LJ

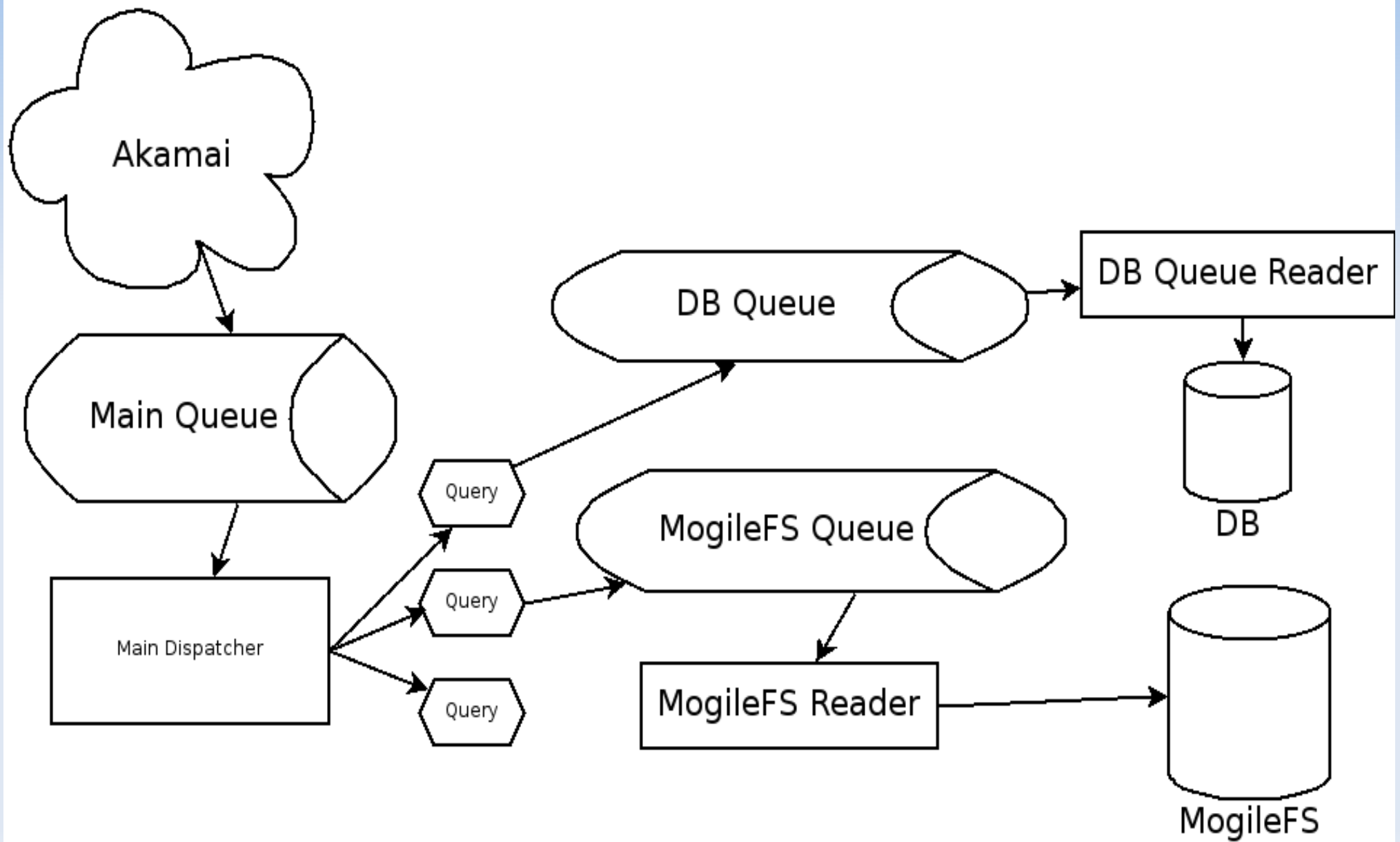
perlbal

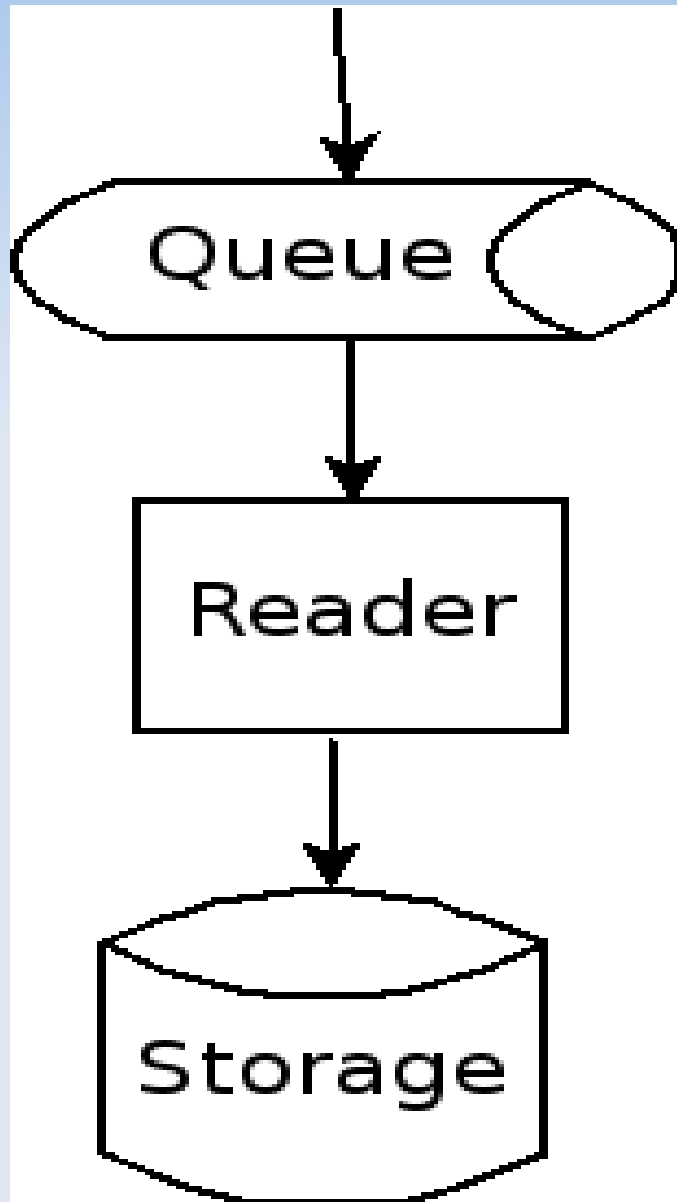
memcached

mogilefs

db shards

Glue together with POE the wrong way  
around

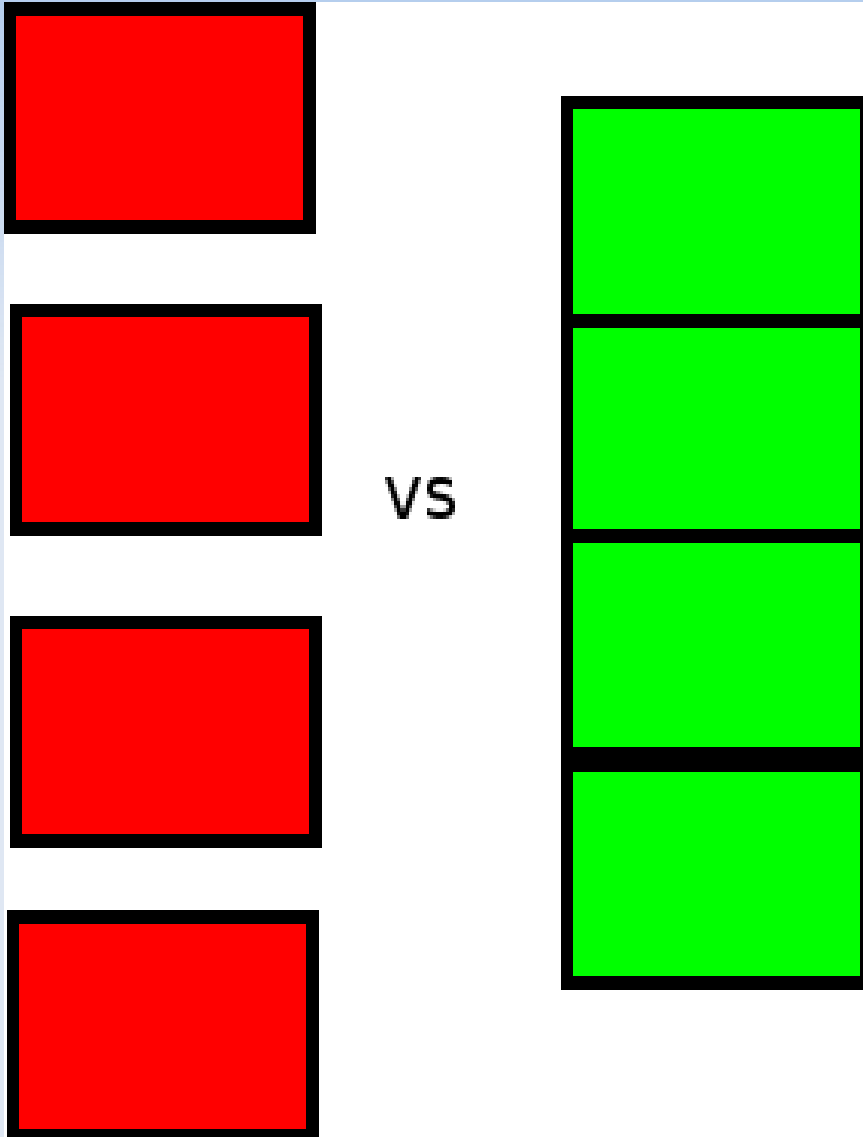




Smooths spikes

Aggregates into lumps

DB doesn't fall over



Larger lumps of data are faster to process and transport

# MogileFS

Distributed load balanced storage

Like memcached for disk

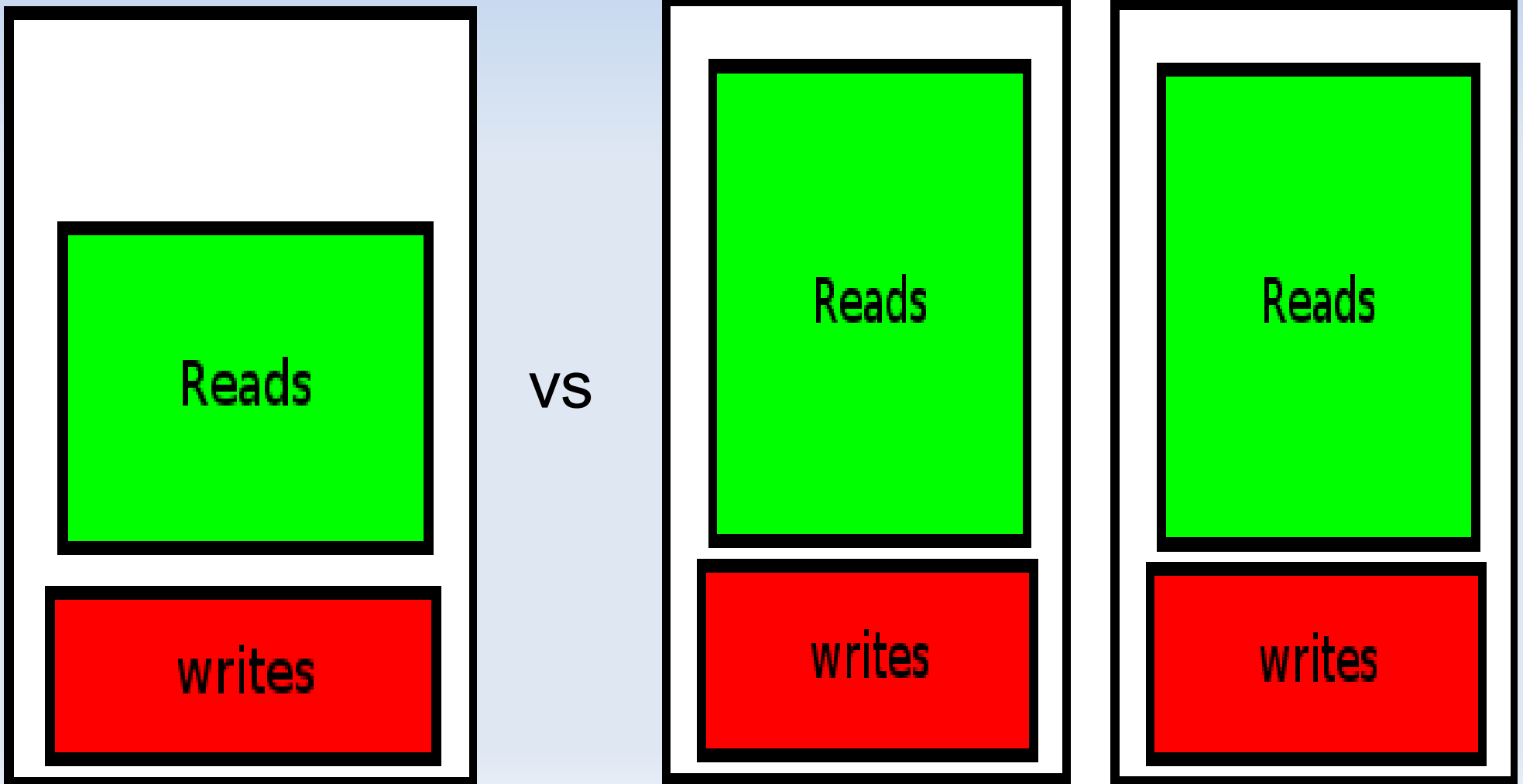
Uses mysql – too many inserts is bad

JSON as compromise record encoding

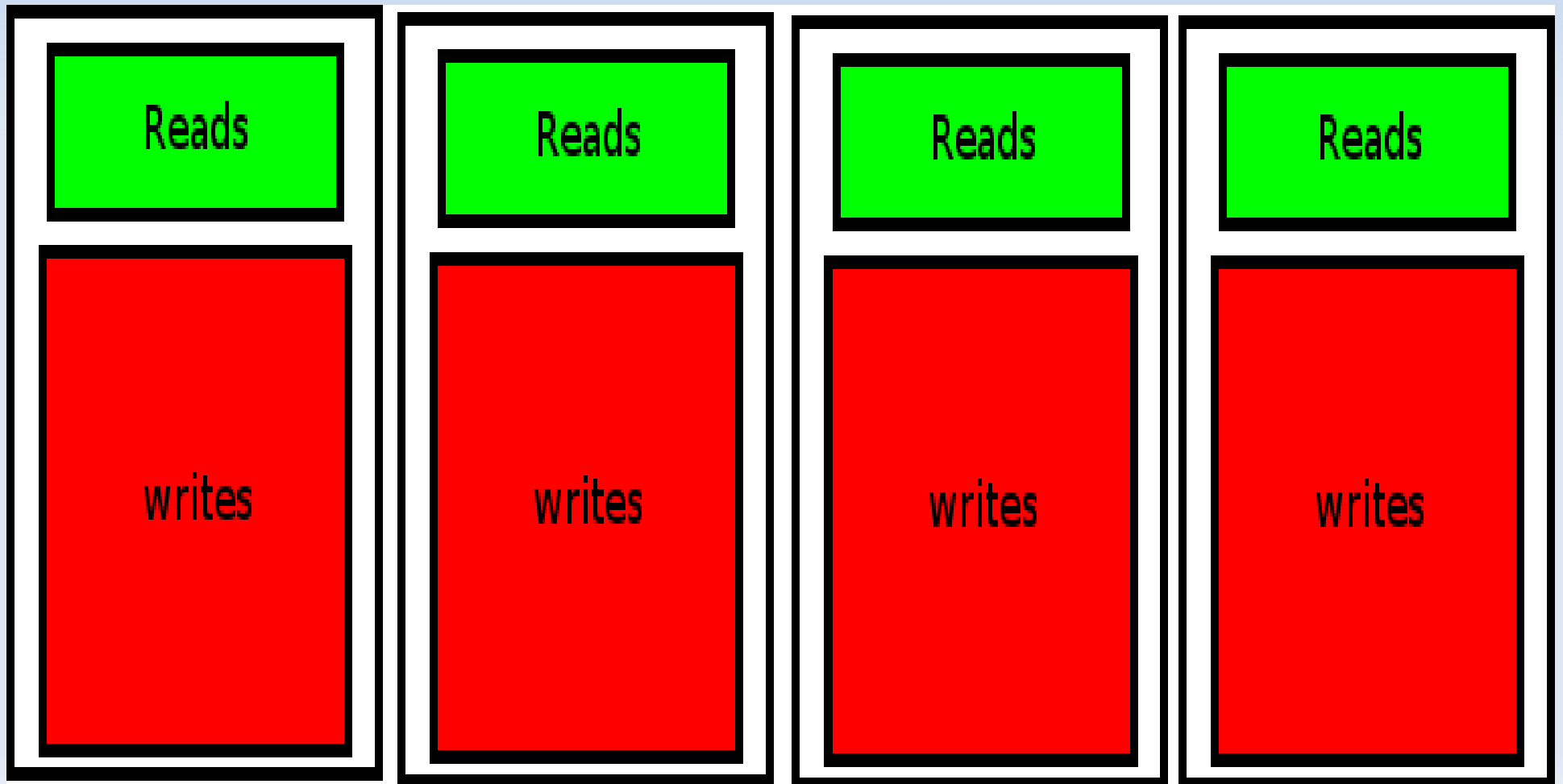
Aggregate data into large gzipped files

Index position of records in sql db

# DB reads scale with clusters

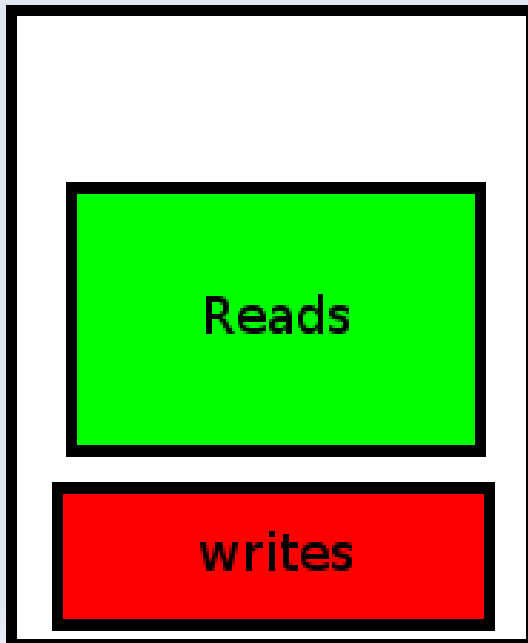


# DB writes don't scale with clusters

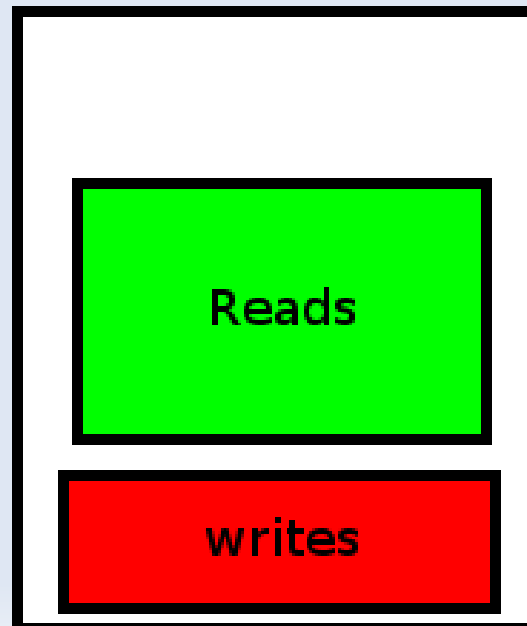


# DB Shards

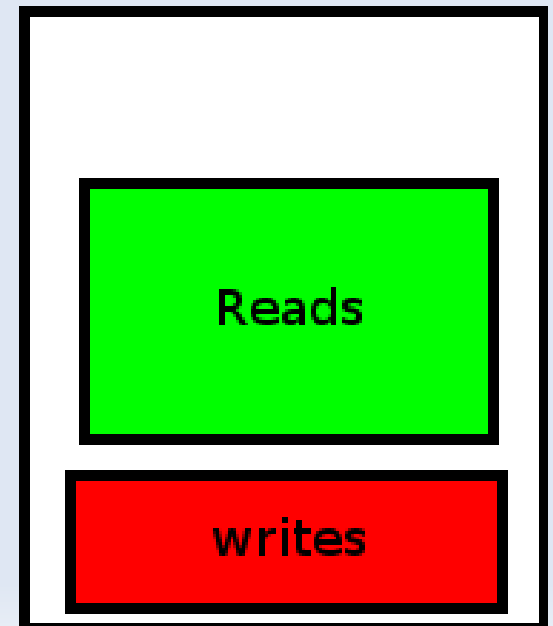
Customer A



Customer B



Customer C

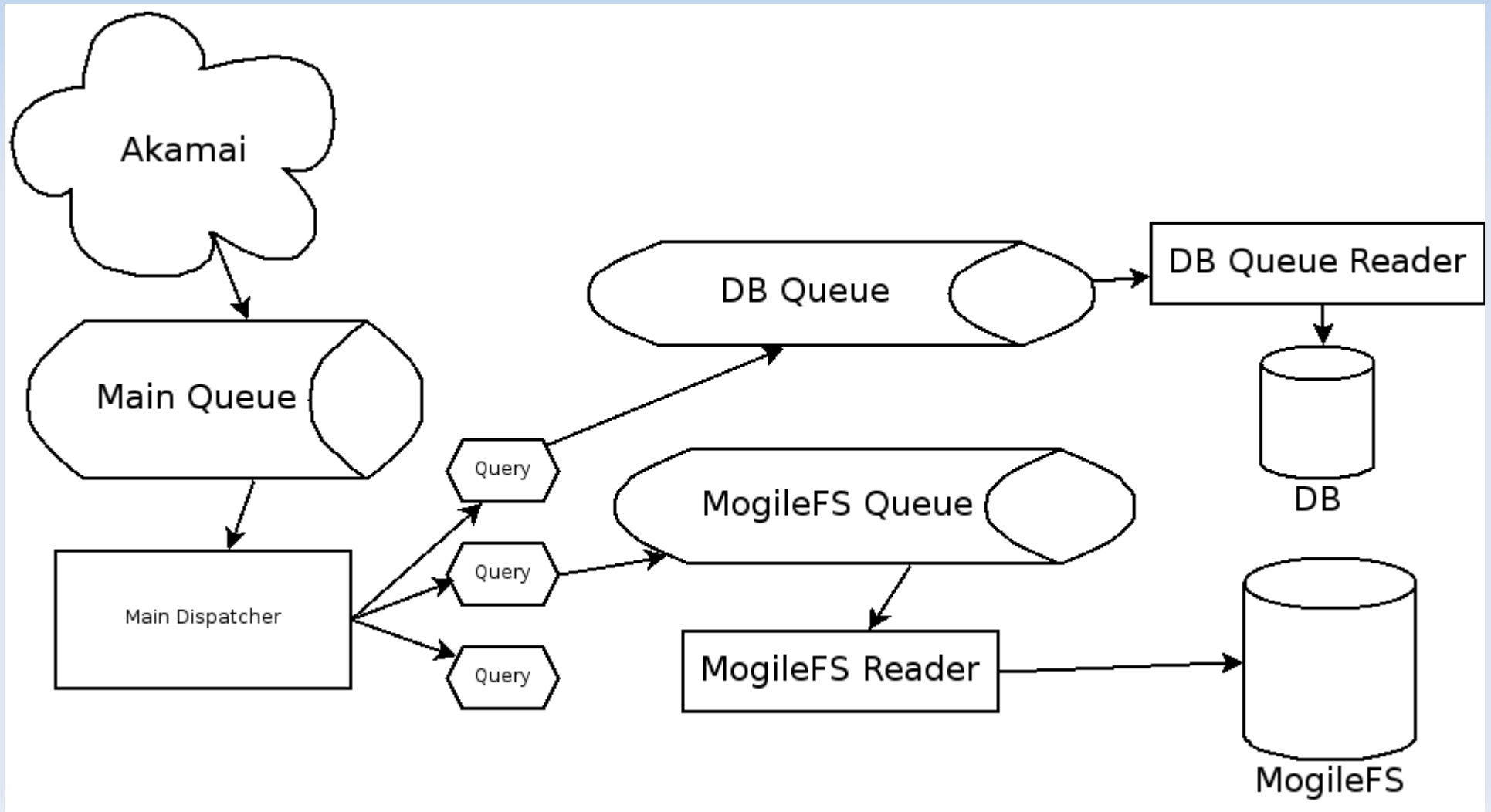


## Other implementations:

Apache/mod\_perl  
Event::Lib

Gearman  
TheSchwartz

# Problems and the future:



Slides are at <http://sketchfactory.com>

mail me at [mock@obscurity.org](mailto:mock@obscurity.org)