- Why Hadoop and HBase?
- Social Media Monitoring
  - Prospective Search and Coprocessors
- Challenges & Lessons Learned
- Resources to get started

Software Architect
@ sentric

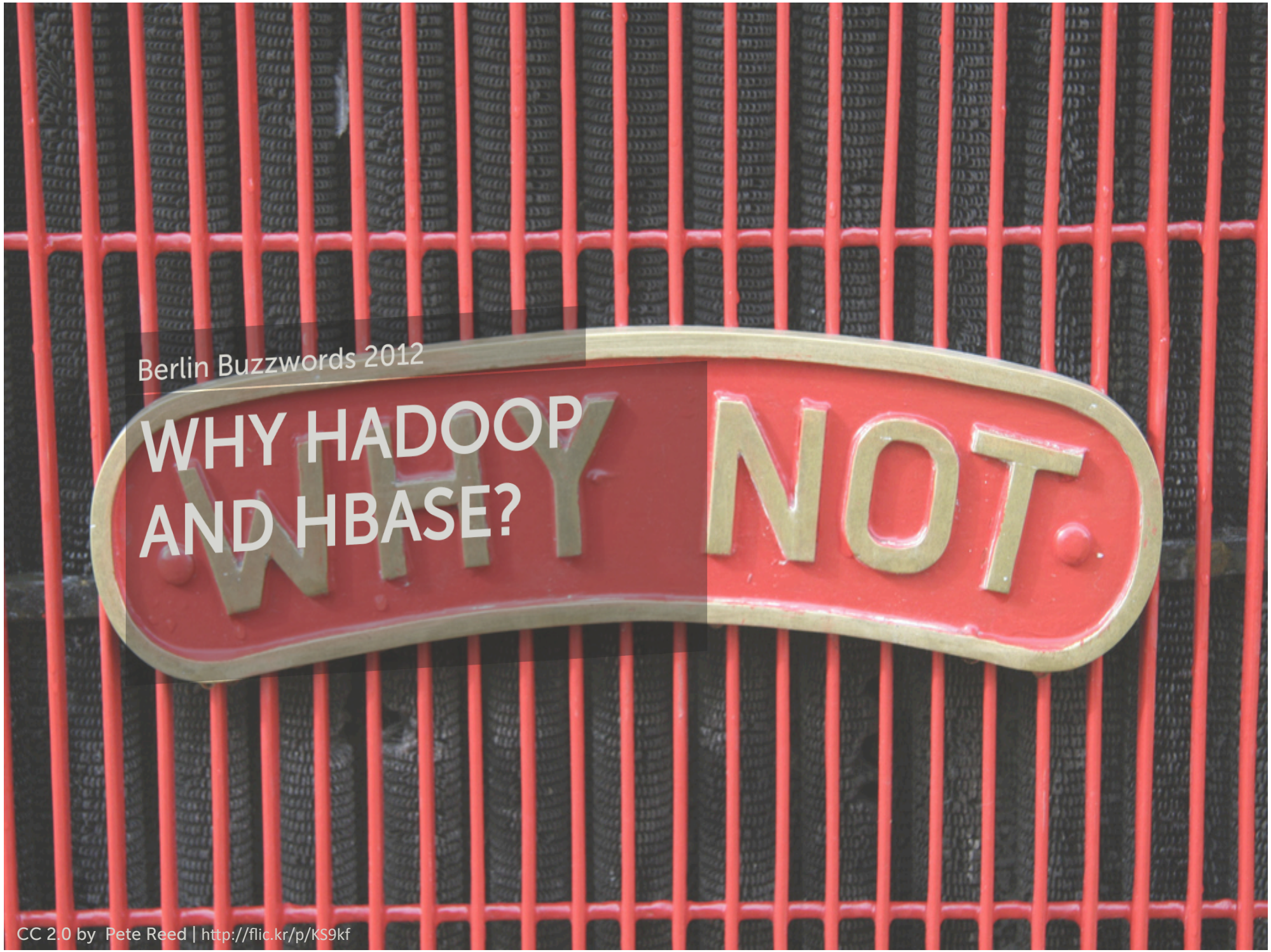
Co-founder and organizer of the Swiss HUG

Contact:

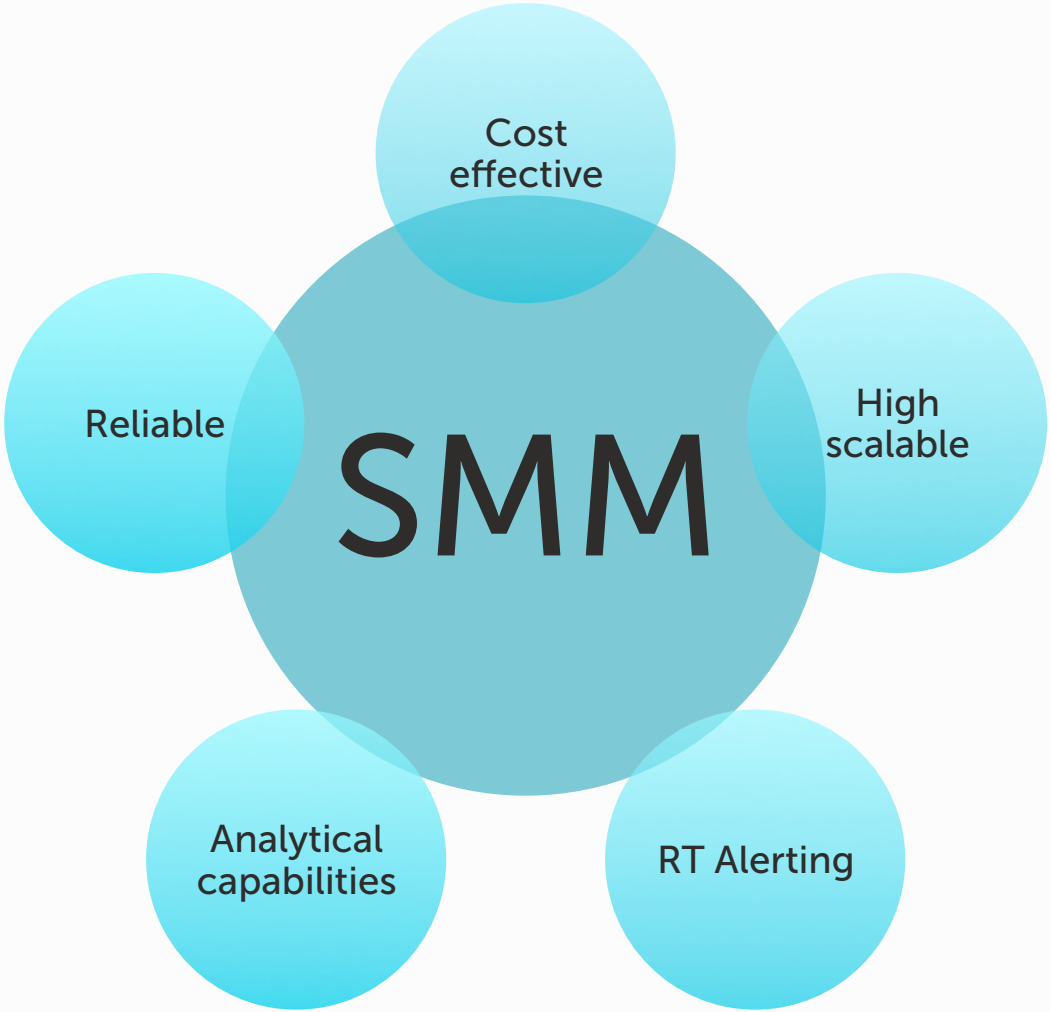
christian.guegi@sentric.ch
http://www.sentric.ch
@chrisgugi

About me

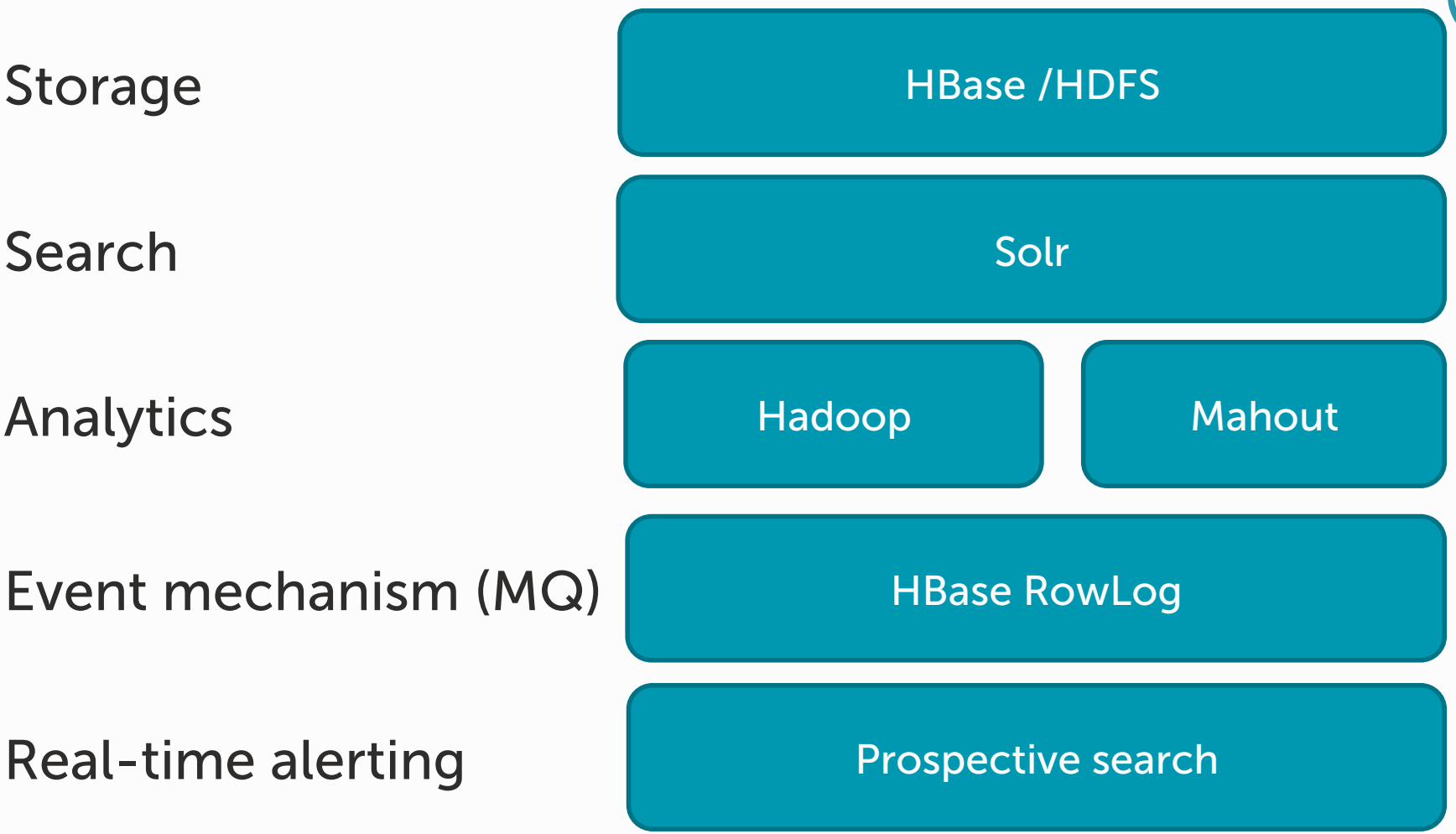
- Spin-off of MeMo News AG, the leading provider for Social Media Monitoring & Analytics in Switzerland
- Big Data expert, focused on Hadoop, HBase and Solr
- Objective: Transforming data into insights

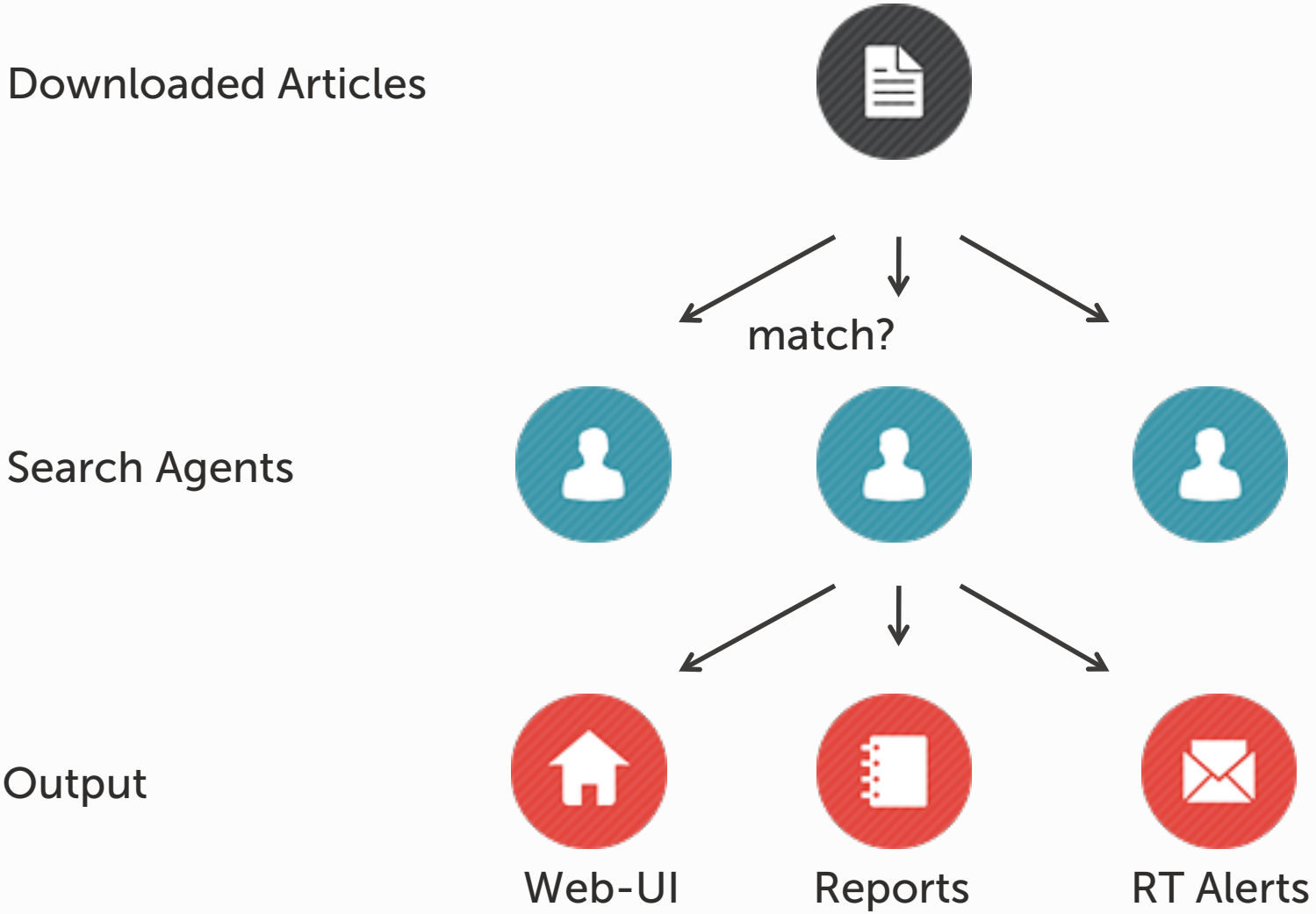



Information Gathering
Information Processing
Analysis & Interpretation
Insight Presentation

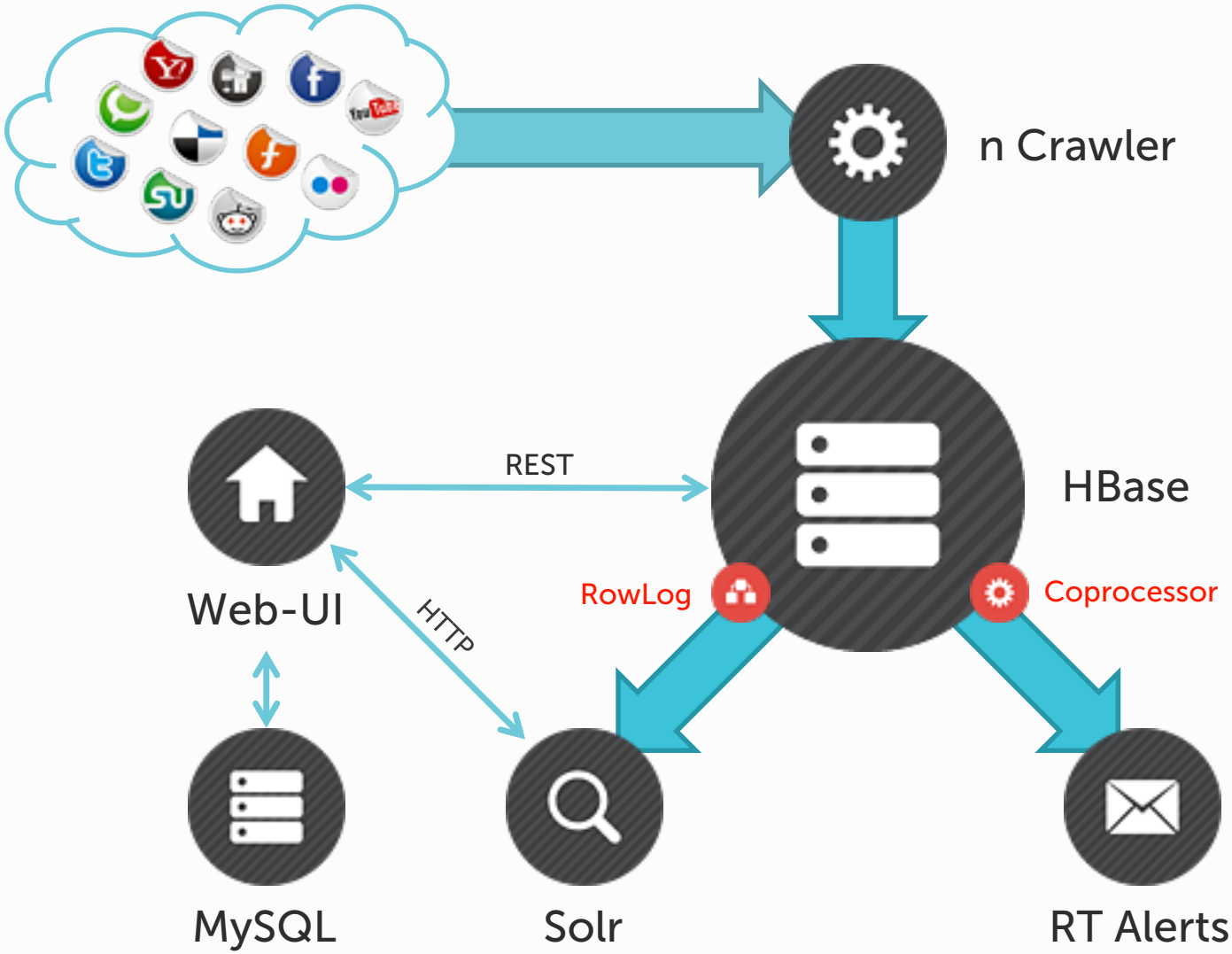


Requirements


Storage
HBase /HDFS
Search
Solr
Analytics
Hadoop
Mahout
Event mechanism (MQ)
HBase RowLog
Real-time alerting
Prospective search



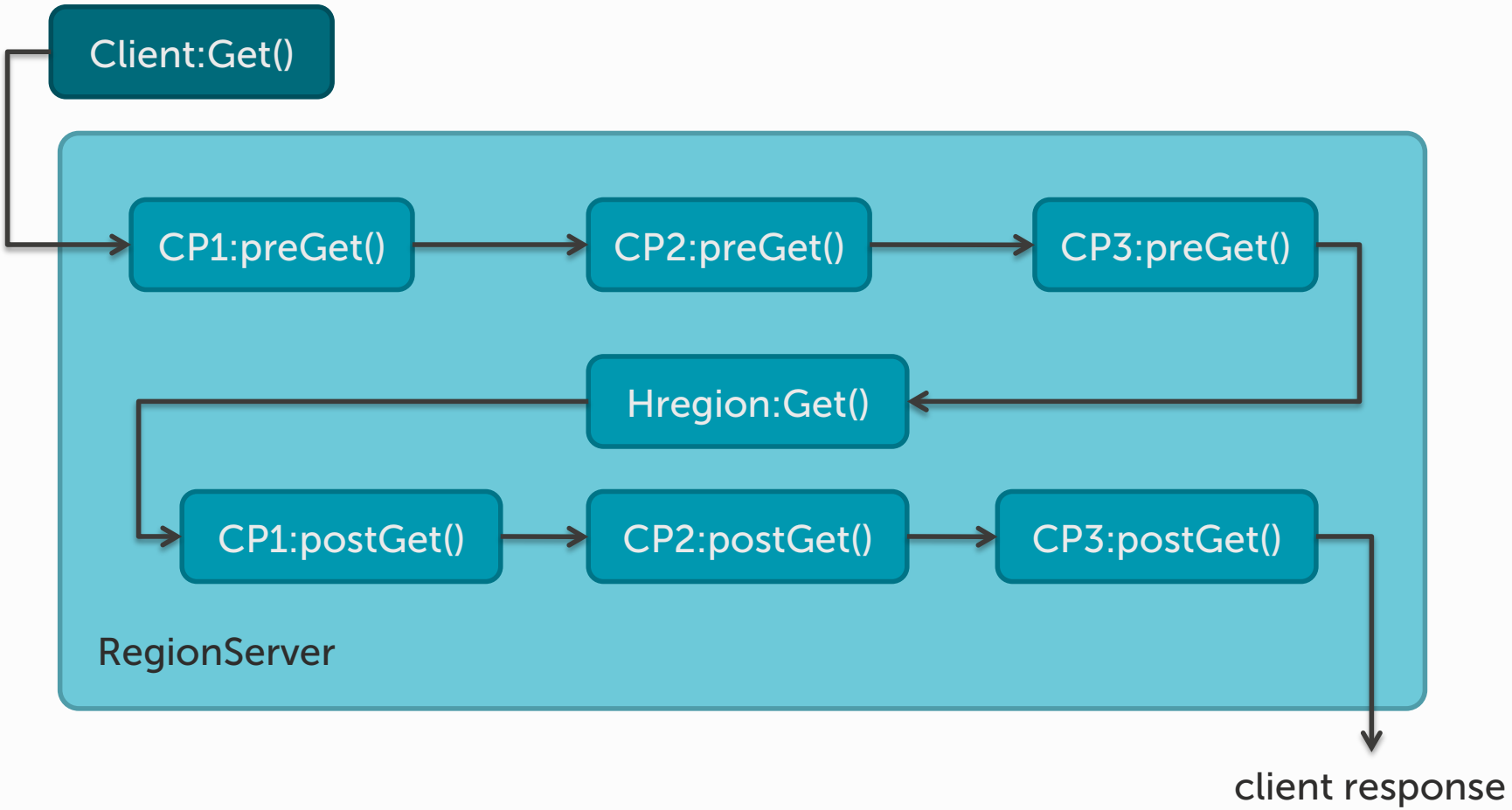
Icons by http://dryicons.com



Icons by http://dryicons.com

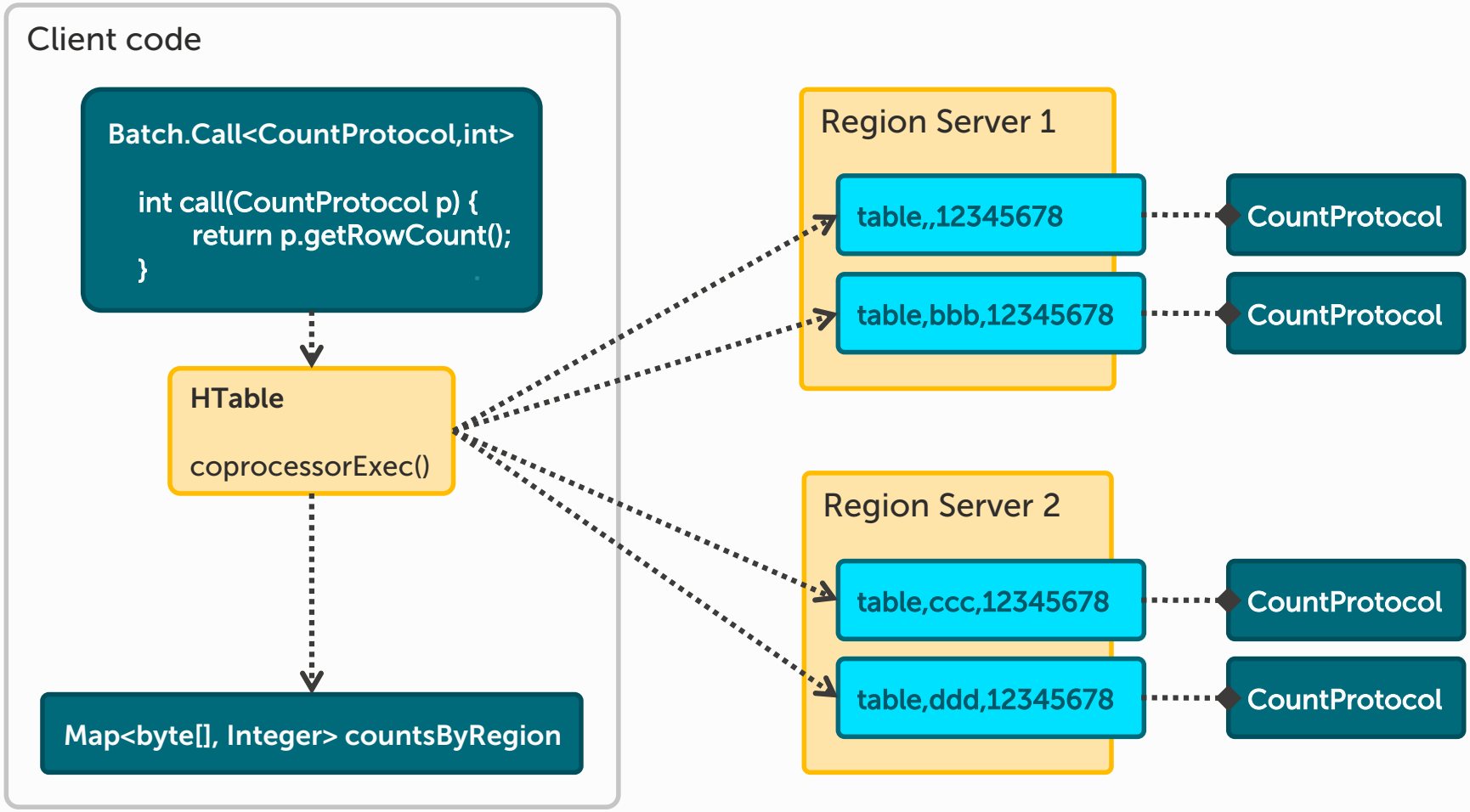
- Inspired by Google Bigtable coprocessors
- HBase version 0.92
- Embed code directly into server processes
- High-level call interface for clients
- Automatic scaling, load balancing, request routing

- Like a database trigger
  - Provides event based hooks
- Concrete Implementations
  - RegionObserver
    - CRUD or DML type operations
  - MasterObserver
    - DDL or metadata operations and cluster administration
  - WALObserver
    - Write-ahead-log appending and restoration



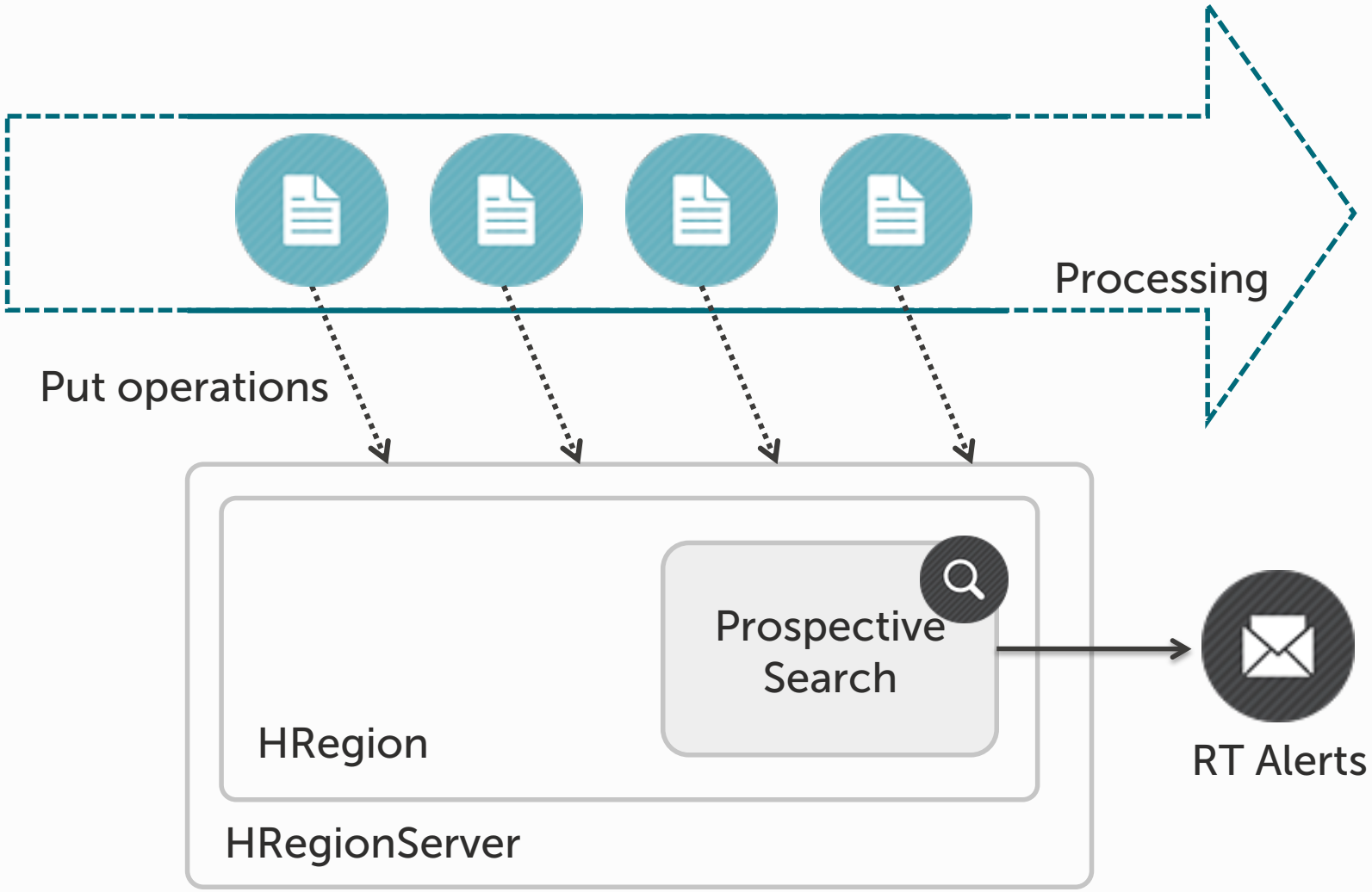
Observer Execution

- Comparable to stored procedures
  - Custom RPC protocol, used between client and region server
- Loaded in region server
- Client call APIs over single row or a row range
  - Framework translates row keys to region location
  - Parallel execution



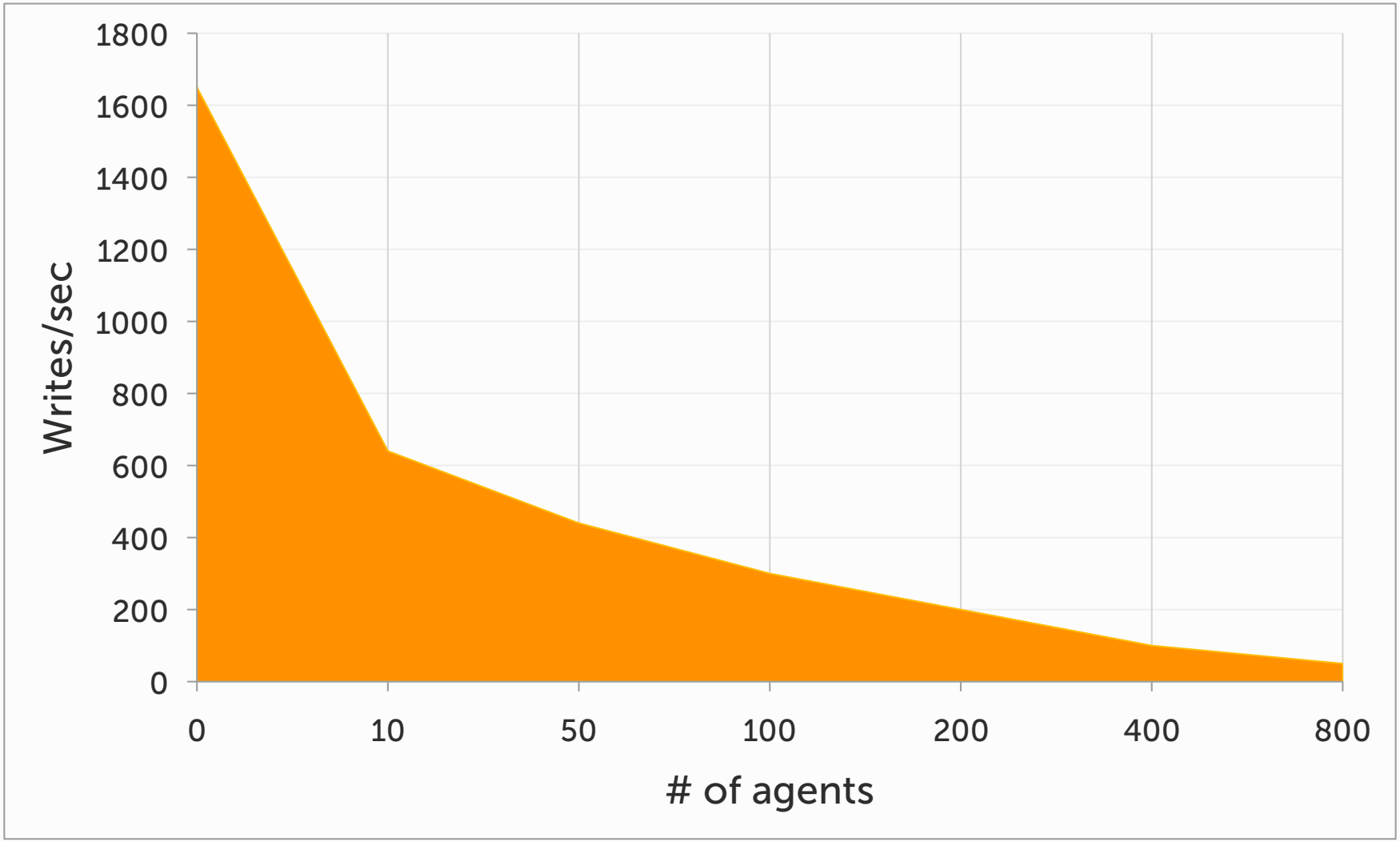
Endpoint Call Routine

- HBase Security (Version 0.94)
- Aggregate operations avg(), sum()
  - AggregatorProtocol
- HBASE-3529: Embedded search



Icons by http://dryicons.com

- Standard, virtualized test cluster: 4RS/DN, 1HM, 1NN, 3ZK
- Test dataset created from 2h of live index (1GB)
- Drive load on RS/DN


1800
1600
1400
1200
1000
800
600
400
200
0
Writes/sec
0
10
50
100
200
400
800
# of agents

- Everyone is still learning
- Some issues only appear at scale
- Production cluster configuration
  - Hardware issues
  - Tuning cluster configuration to our work loads
- HBase stability
- Monitoring health of HBase

- Be careful with expensive operations in coprocessors
- At scale, nothing works as advertised
- Monitoring/Operational tooling is most important
- Play with all the configurations and benchmark for tuning

- https://blogs.apache.org/hbase/entry/coprocessor_introduction
- http://hbase.apache.org/apidocs/index.html
- http://www.lilyproject.org/lily/about/playground/hbaserowlog.html
- http://www.github.com/sentric/HBasePS

sentric

Questions?

Christian Gügi

christian.guegi@sentric.ch

Thank you!

sentric