New Journey of HBase in Alibaba and Cloud

八年磨一剑，HBase在阿里巴巴和云上的新征程

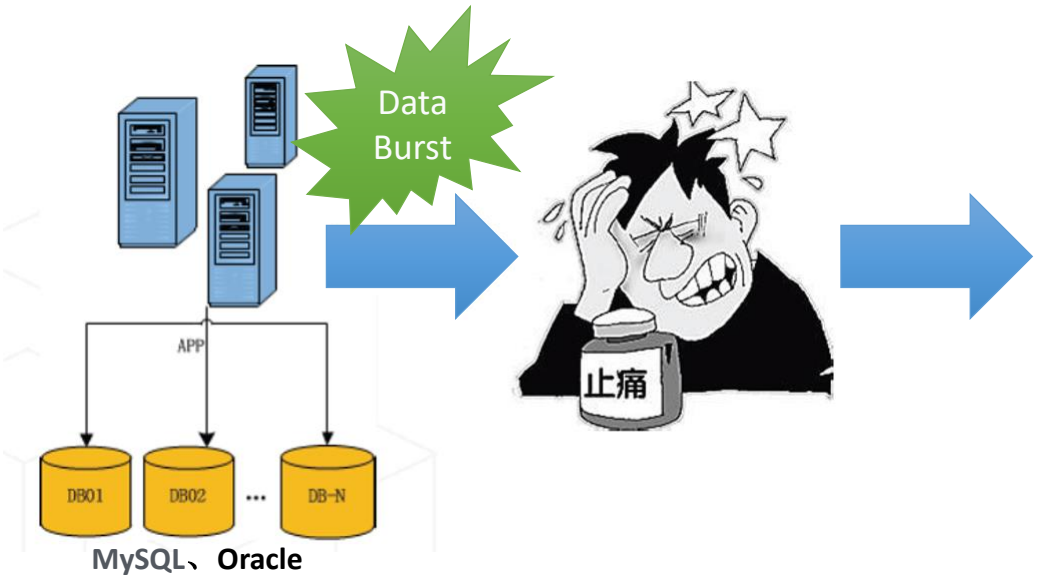
Chunhui Shen and Long Cao

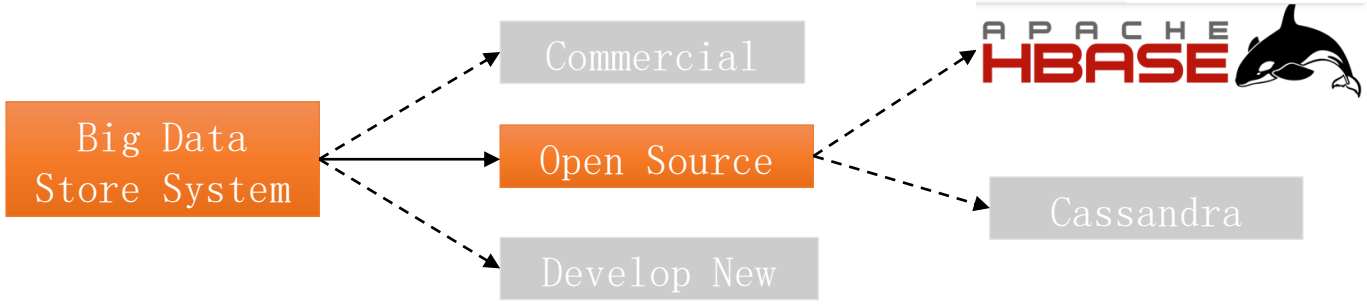
August 17,2018

Content

01 AliHB-Introduction of Alibaba HBase

HBase History in Alibaba

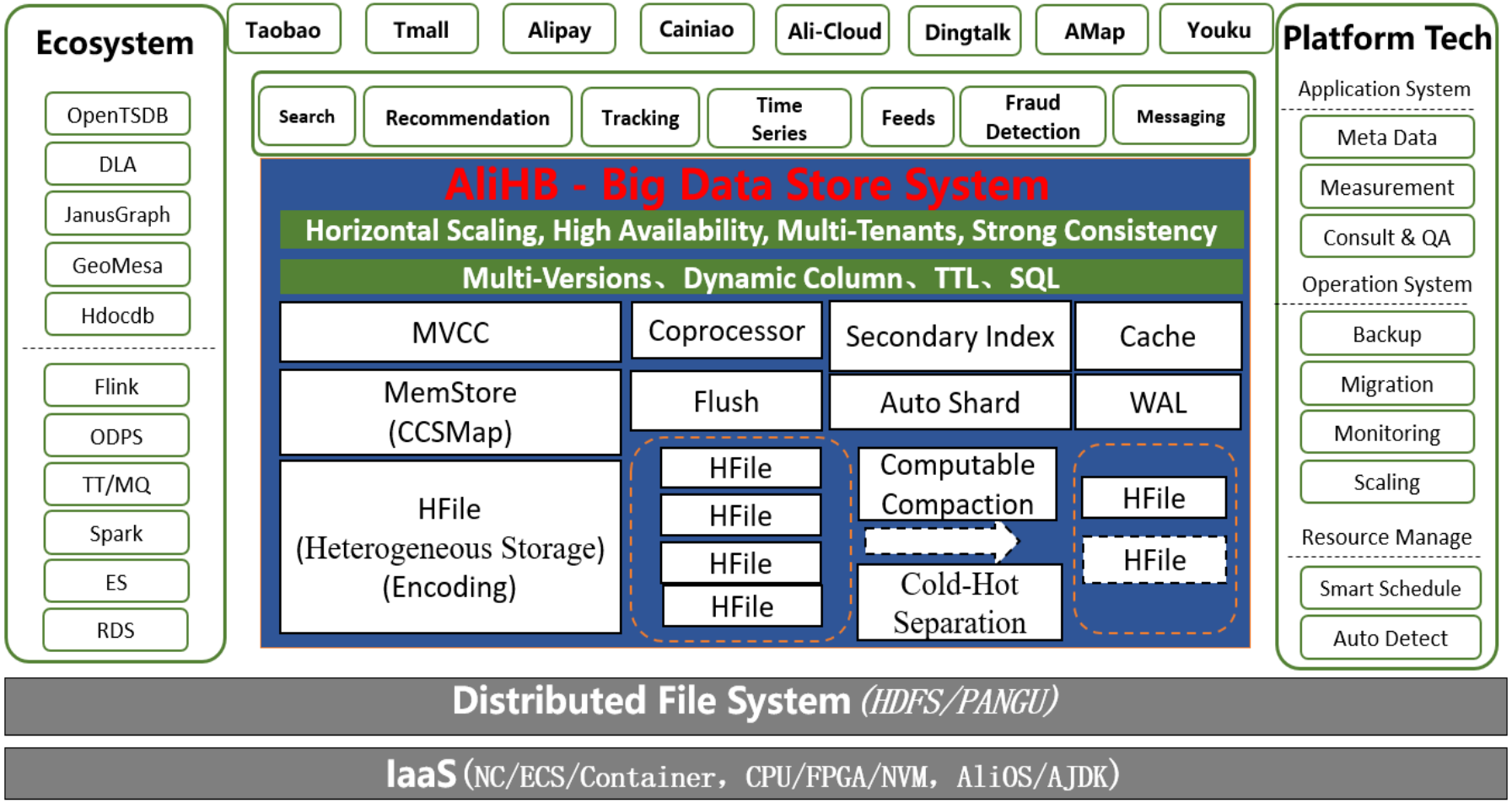




- Why HBase
  - Began using since 2010
  - Active community
  - Hadoop ecosystem
  - Facebook successful case
  - Google famous paper: Big Table

- Used Version
  - 0.20->0.90->0.92->0.94->0.98->1.1->2.0
- The earliest case in 2010-2011
  - Search Store
  - Taobao History Order
  - Alipay Risk Management
- Internal branch AliHB

Overview of AliHB



- **Performance**
  - High-Performance Data Structure、Lock-Free、Group IO
- **Feature**
  - SQL、Secondary Index
  - Multi-Tenants、Cold-Hot Separation、Async API
- **Stability**
  - High Availability Architecture
  - Faster MTTR
  - Verification in Double 11 Shopping Day
- **Efficient Maintenance**
  - Effective Monitoring
  - Full Path Trace
  - No-pause migration

- 12000+ Nodes , 100+ Clusters , 200+ Million OPS , 100+ PB Data
- 20+ BU , 6000+ Users , 100+ Production Changes per Day

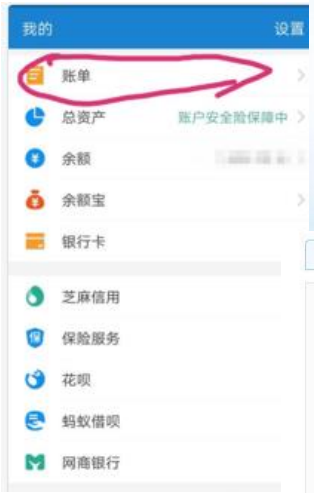
Open Source and Community

- Contributing to open source since 2011
- 3 PMC, 6 Committers in Alibaba
- Sponsor the Chinese HBase Technology Community
  - Already Organized 2 HBase Meetup
  - At least one HBase Related tech article one day
  - Tens of thousands of readers now, and more are coming
- Hosting HBase Con Asia 2018
- Promote the use of HBase through several conference talks
- Hope more people to join in HBase Community

中国HBase技术社区微信订阅号
hbasegroup

Core Scenarios in Alibaba

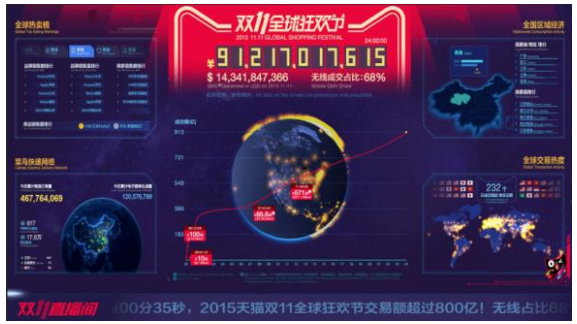
Message, Orders, Feeds …

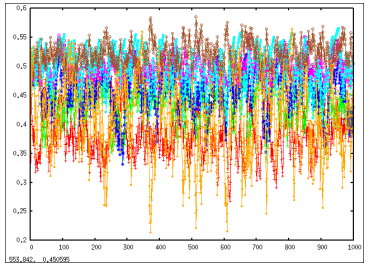


Alipay Bills

Cainiao Logistics

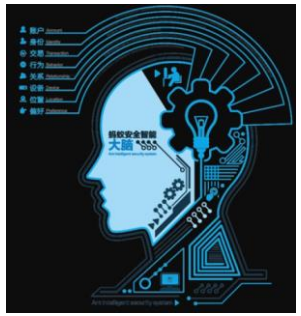
Monitor, Log, Tracking, IoT Data…





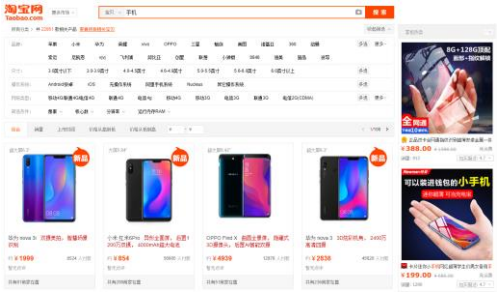
AI Storage

Ant Intelligent Security



Intelligent Customer Service

Recommendation Search, BI Report…

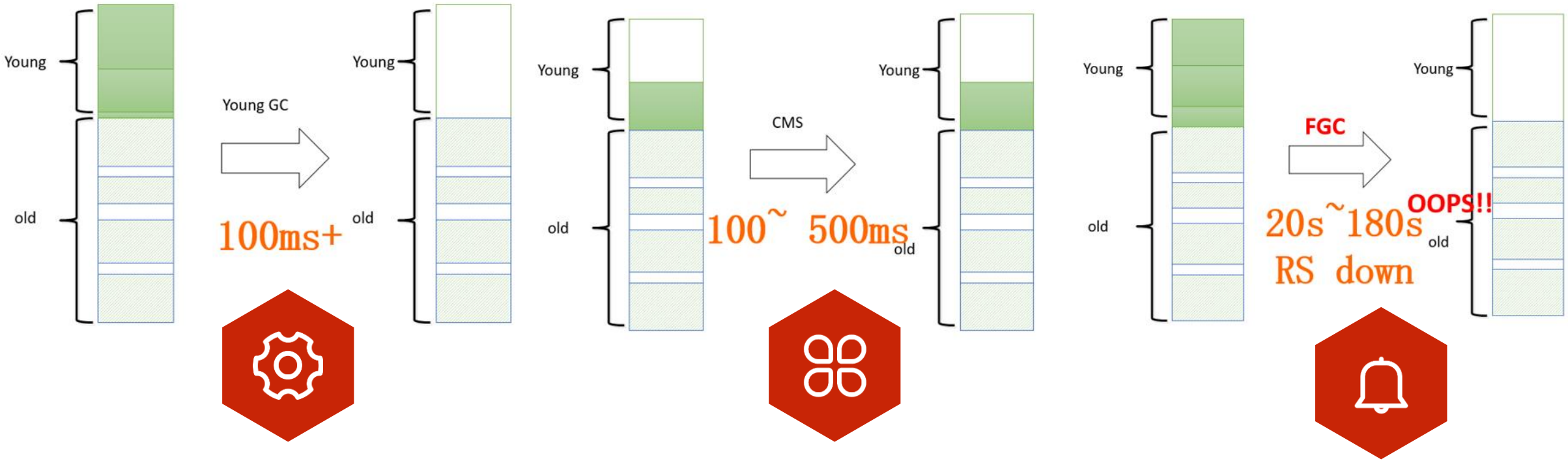


Ali-HBase

02 Recent Key Challenge & Improvements

GC Trouble

GC Problems Under100GB Memory



Frequent
Slow Request

Very Slow
Request

Service
Unavailable

GC Trouble

Only for offline application

Exploring a Thorough Solution

Rewriting with C++

GC Trouble

Type	Pause Time	Frequency
YGC	100ms+	Once per 5 Secs
CMS	100~500ms	Once per 5 Mins
FGC	20s-180s	Once per 7~60 Days

Type	Pause Time	Frequency
YGC	**5ms**	Once per 5 Secs
CMS	100ms	Once per 5 Hours
FGC	N/A	N/A

Allocation and reclaim the major memory by hbase itself, rather than JVM

CCSMap

BucketCacheV2

New GC algorithm in AJDK

ZenGC

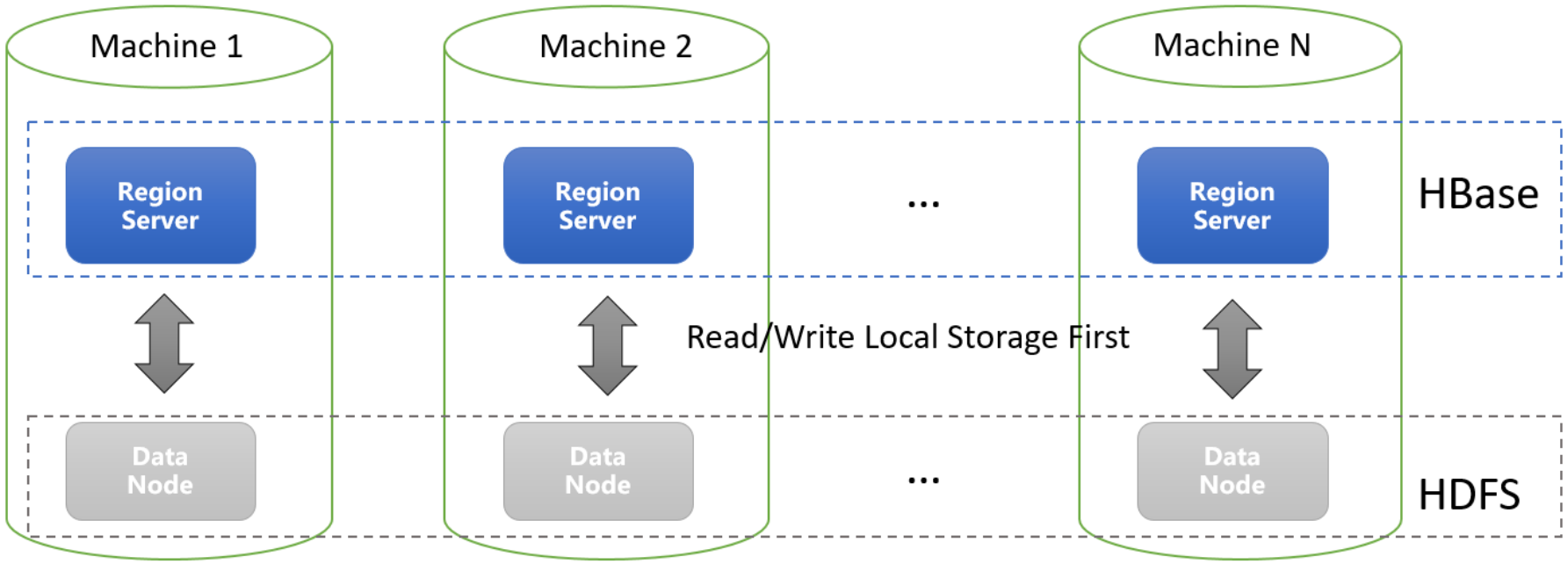
Try best to reuse object(In Core Path) when programming

GC Trouble

New BucketCache in HBase-2.0

CCSMap in HBase-3.0

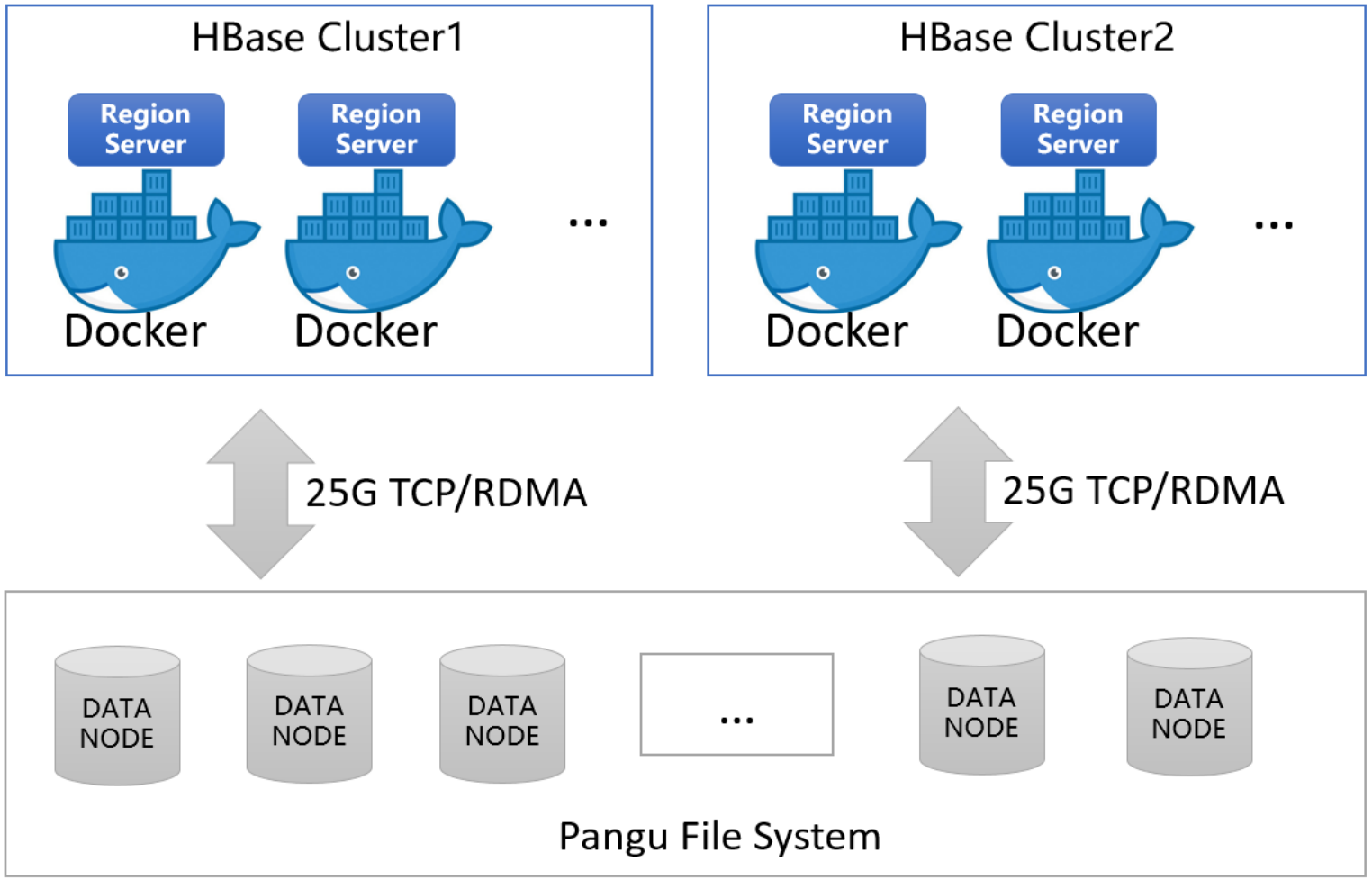
Separation of Computing & Storage



Localized Deployment

- Low IO latency with Short-Circuit Read
- Unbalanced storage space, especially between clusters
- Difficult to increase the usage ratio of CPU and Disk (both), especially when lots of scenarios
- Cluster scaling is slow because of datanode decommission

Separation of Computing & Storage

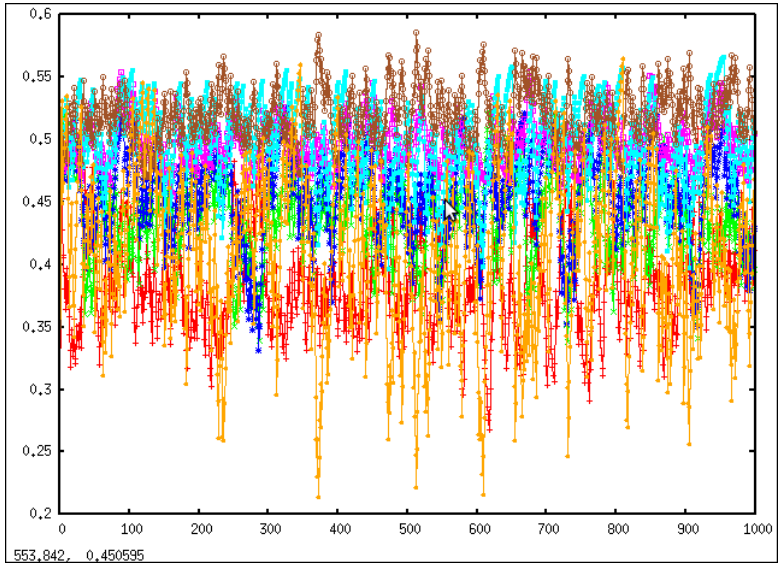


Shared-Storage Deployment

- Big shared storage, more balanced
- Compute node can scale independently
- Storage node can scale independently
- Auto-scaling become feasible
- Based on load statistics, smart schedule between clusters
- Share compute resources with other applications

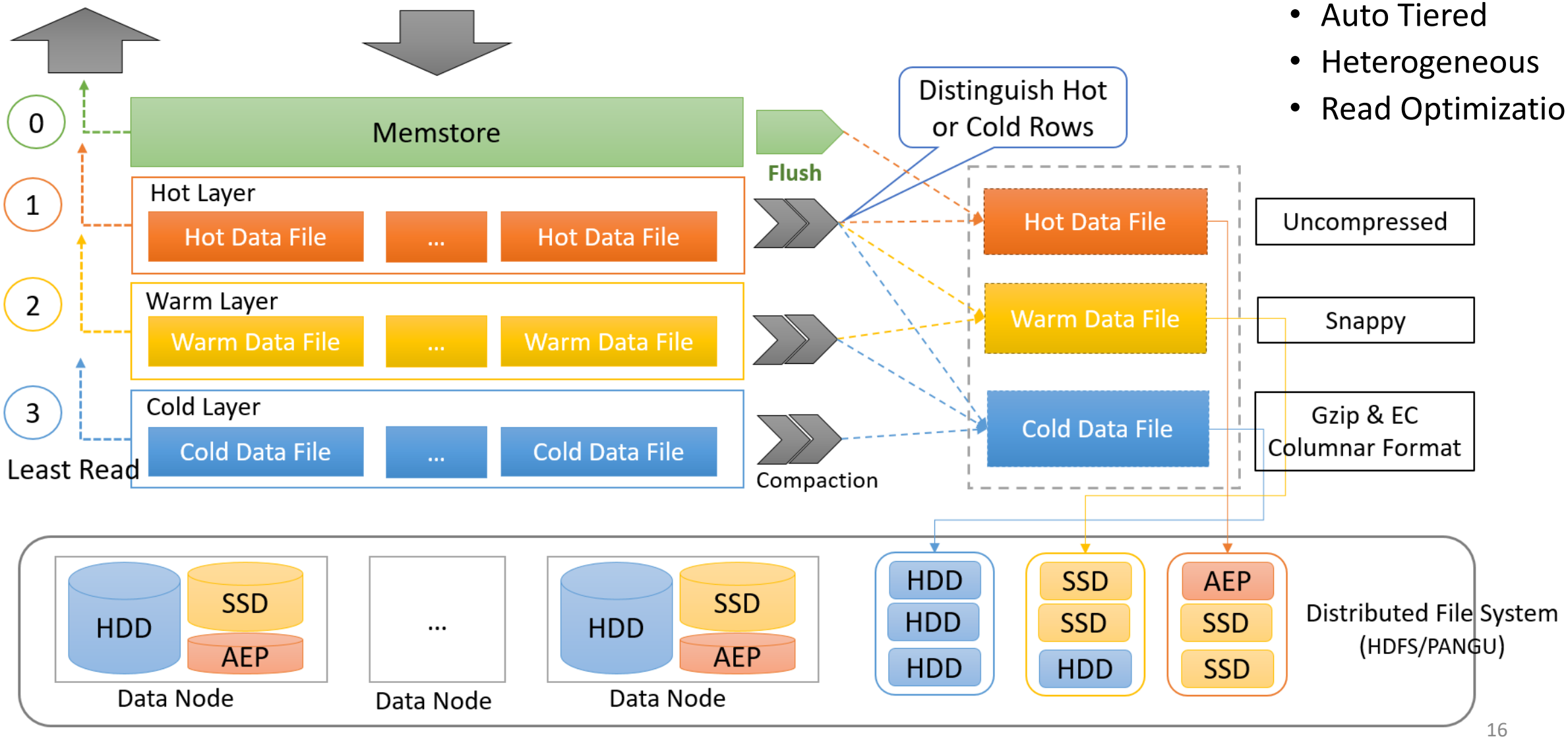
Heterogeneous Cold-Hot Storage

- **HBase has the capability to hold all the data of whole life cycle**
- **But in most cases, like monitor, trace, order, logistics**
  - **The recently generated data is often accessed, but occupy very little storage space**
  - **The history data is rarely visited, but occupy a lot of storage space**
- **Common solution**
  - **Cold storage system for history data**
  - **Hot storage system for recent data**
  - **Move the data from hot storage system to cold storage system periodically**



Heterogeneous Cold-Hot Storage

- Easy To Use
- Auto Tiered
- Heterogeneous
- Read Optimization



Diagnostic System

12000+ Nodes , 100+ Clusters , 6000+ Users

ERROR! TIMEOUT !

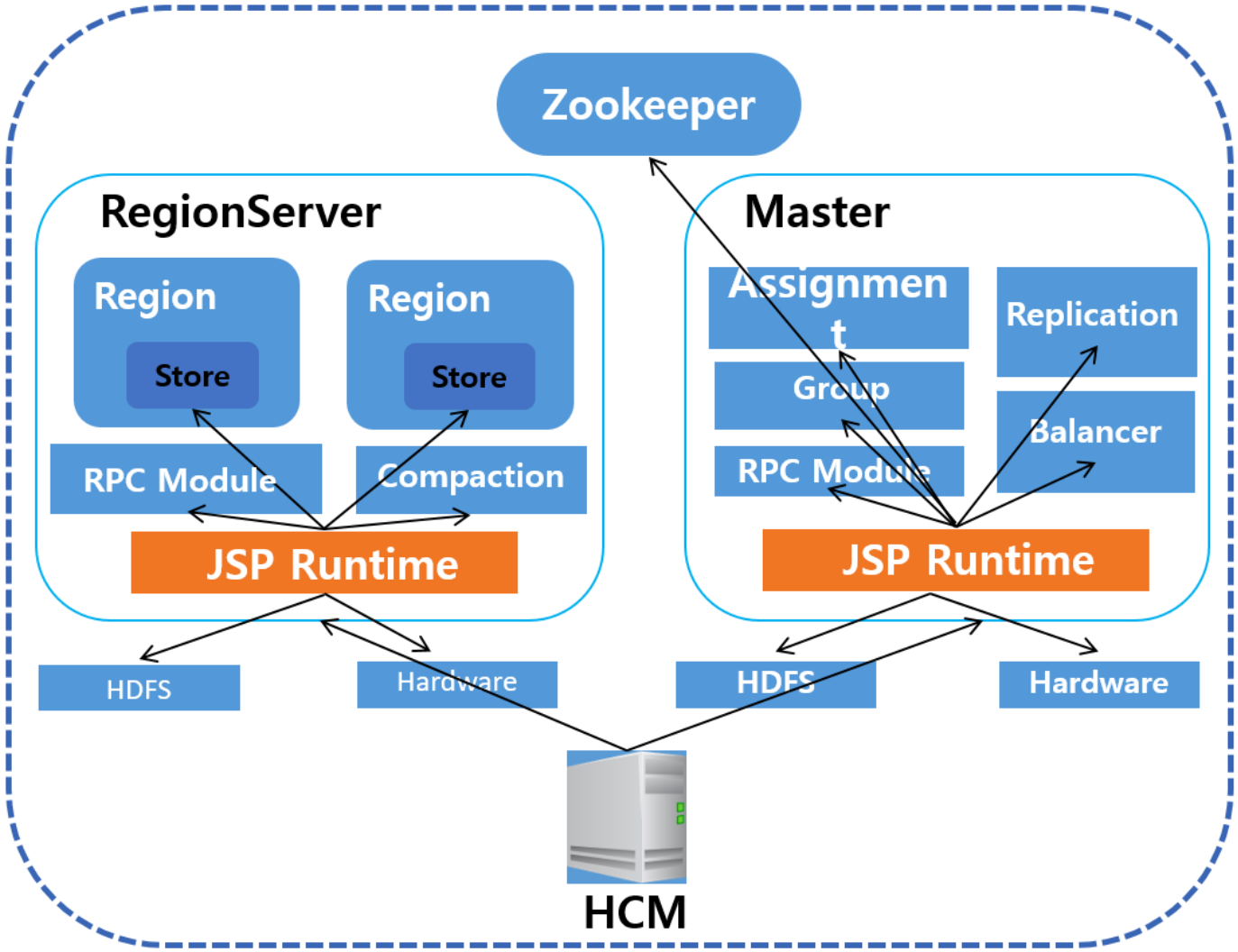
"Request Rush?" — Monitor
"Big Region?" — Web UI
"Full Disk?" — df
"Bad Disk? " — tsar,demsg
……

HBase Diagnostic Center

1. The unified entrance of trouble shooting
2. Experience/Solution => Function of Diagnostic System

Diagnostic System

1. One extra server for all
2. No Agent
3. Adding rule dynamically
4. Runtime information
5. Check all components
6. Only 10 seconds for a diagnosis



Diagnostic System

Shared on Apsara HBase

50+ Rules

80%+ Accuracy

HBase

- Compaction
- Stuck
- Balance Abnormal
- Table Abnormal
- Region Offline
- Replication Delay
- Too many files
- High Meta Load
- Multi Assign
- ……

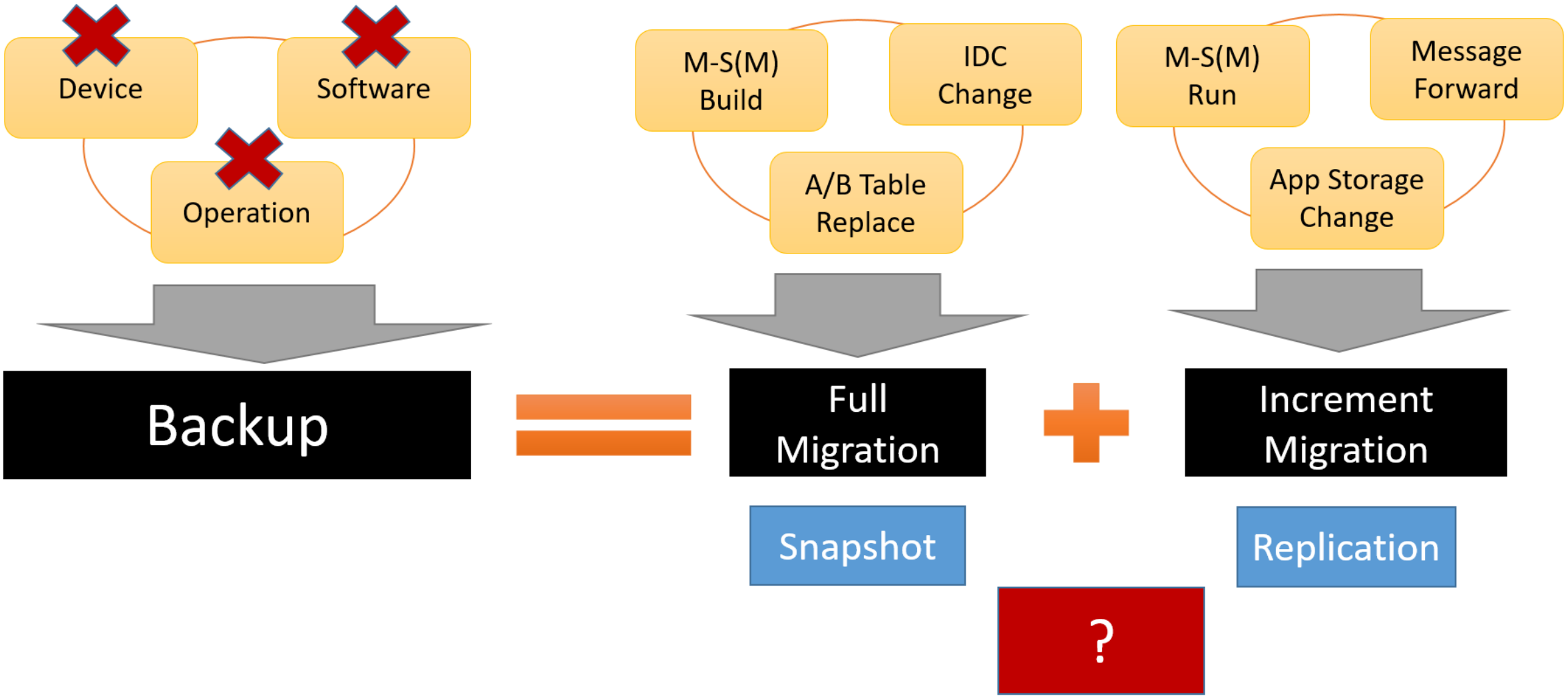
ZK/HDFS

- ZK Unavailable
- Block Miss
- NameNode Abnormal
- Full capacity of datanode
- Inconsistent state between two namenodes
- Too much Xceivers
- Disk not mounted
- ……

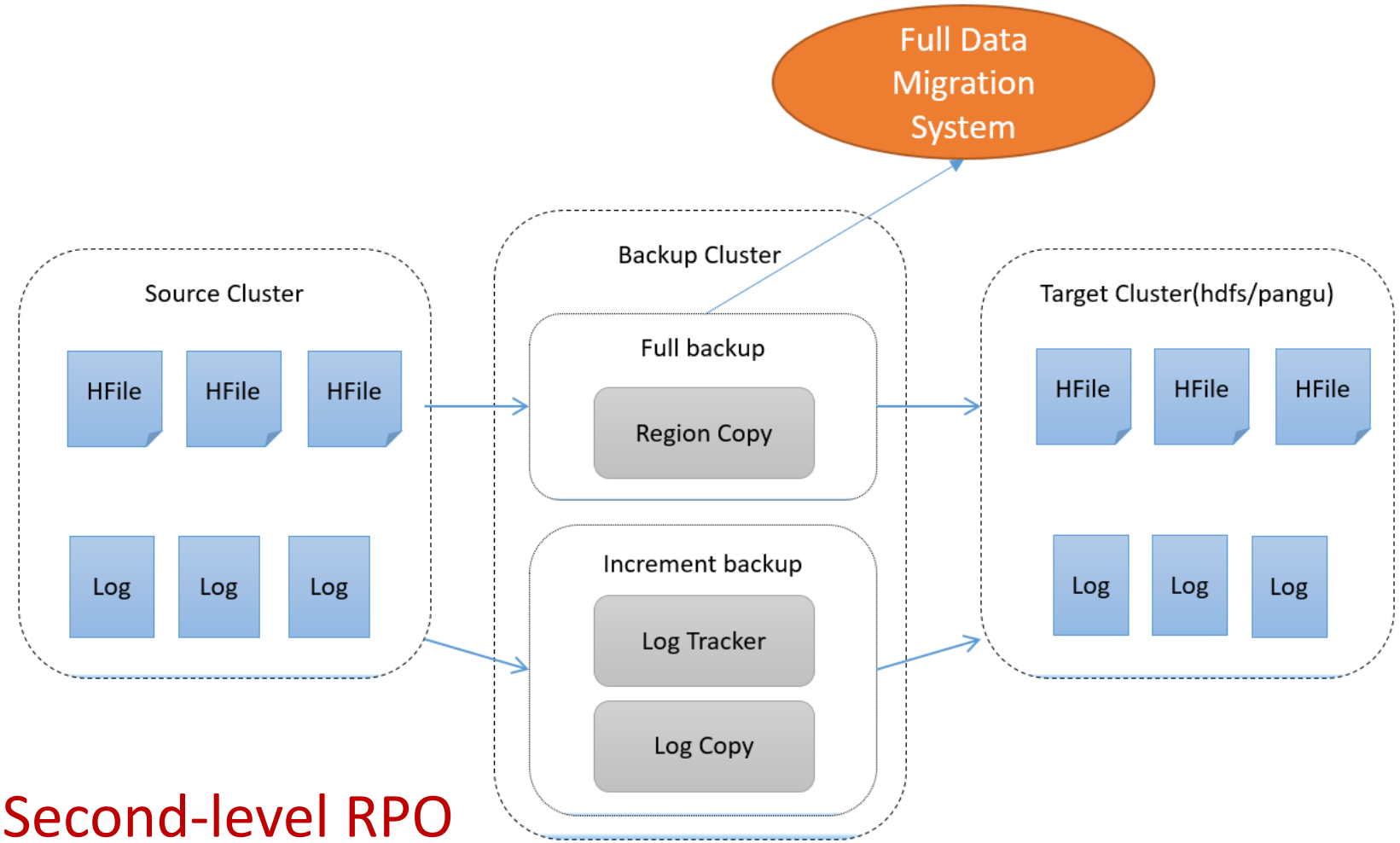
Hardware

- Insufficient disk space
- Slow Disk
- Bad Disk
- Too much TCP error
- Slow ping
- CPU hang
- Load too high
- Port is unreachable
- ……

Migration & Backup



Migration & Backup



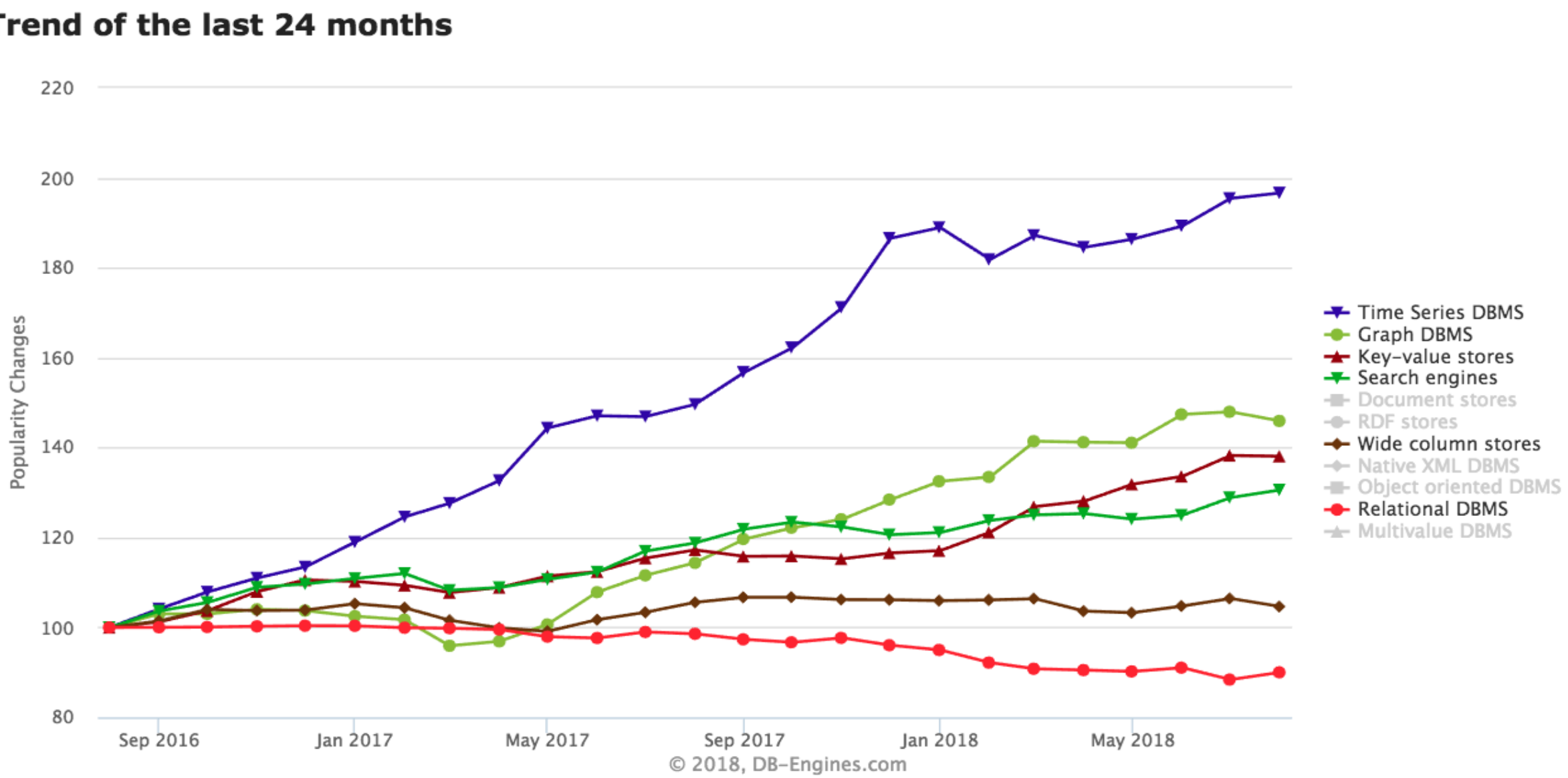
Second-level RPO
Minute-level RTO

Independent with HBase

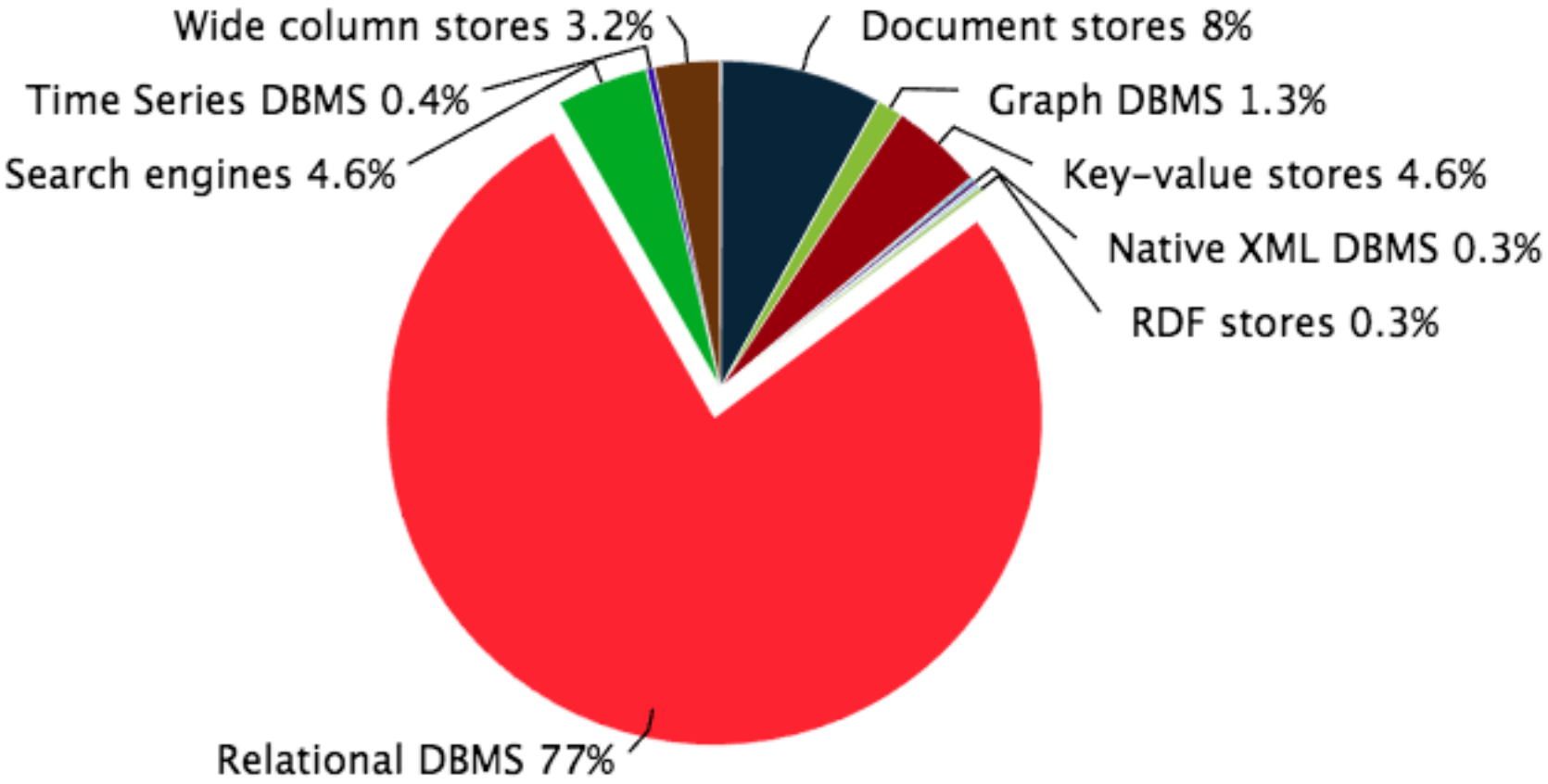
- almost no impact to service
- easy to upgrade
- support multi versions
- support the non-hbase target

03 HBase Ecosystem & Multi-model DB & Cloud

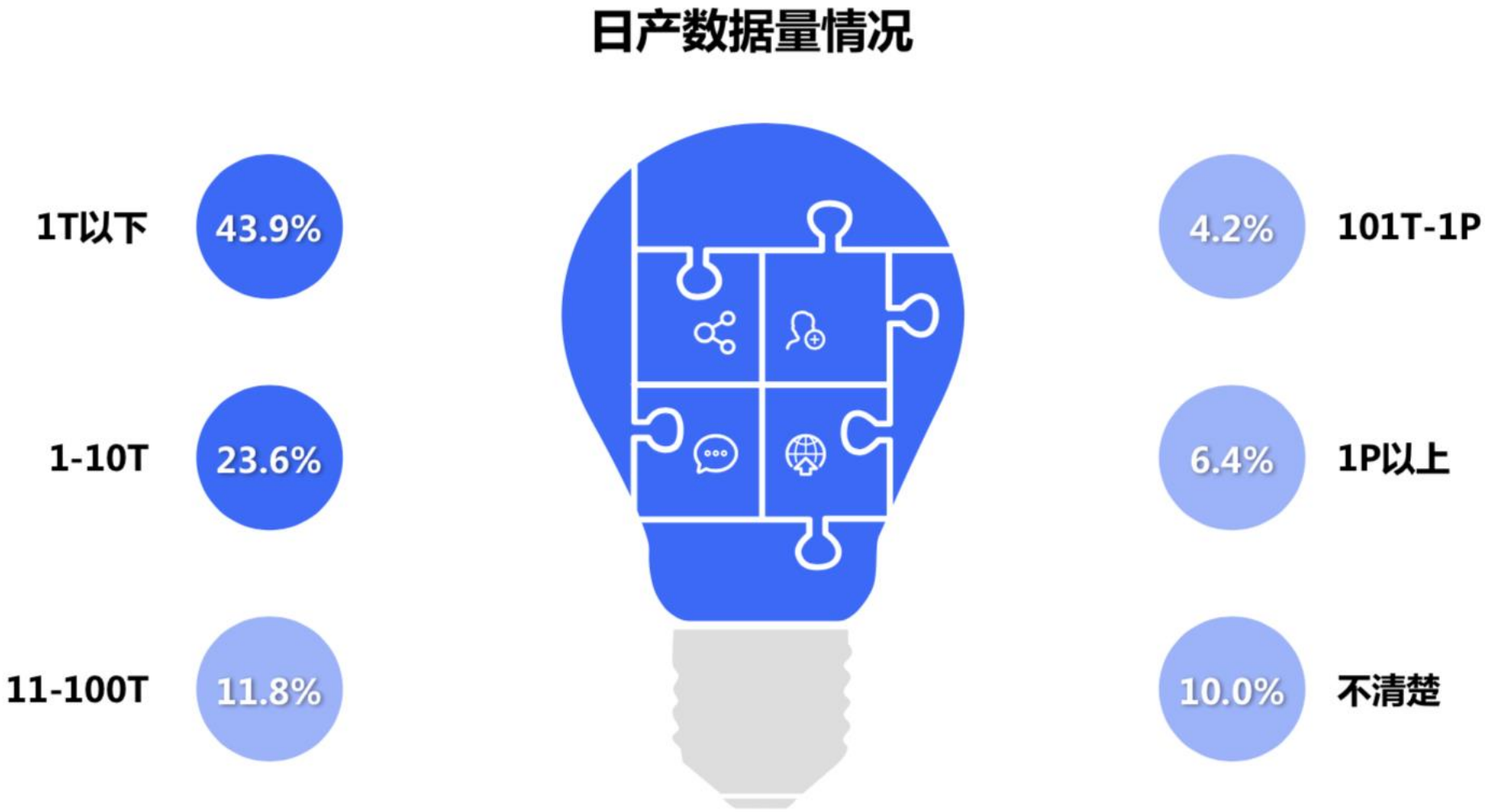
Popularity changes per DB category



Ranking scores per category in percent



Data size per day



数据来源：云栖社区《2017中国开发者调查报告》，7032人参与调查，2017年12月

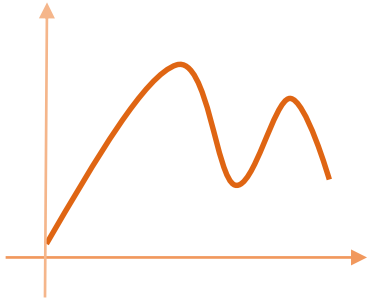
All in one

Key Value

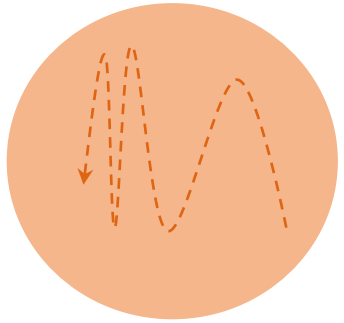
Relational

Doucument

Graph



Time Series



Geospatial

Tabular NoSQL

All in one

HBase

Key Value

Phoenix/AntsDB

Relational

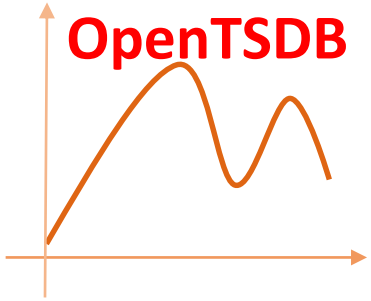
HBase

Doucument

JanusGraph

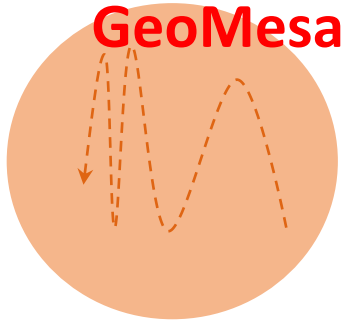
Graph

OpenTSDB



Time Series

GeoMesa



Geospatial

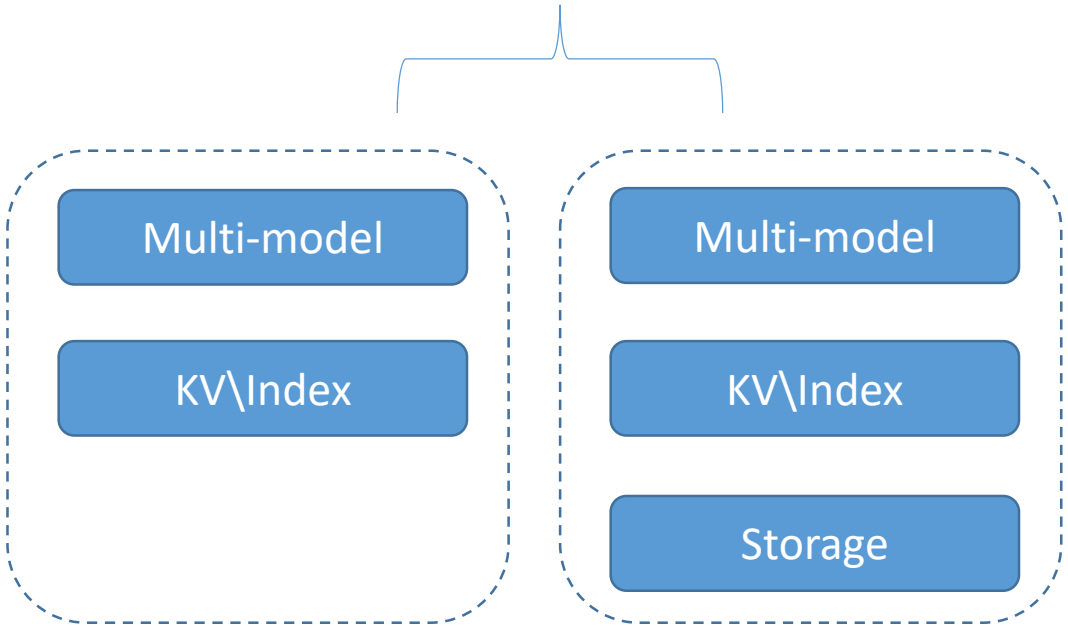
HBase

Tabular NoSQL

Multi-model - Native Or Layer

Neo4j
InfluxDB
CockroachDB
PG

HBase Ecosystem
DataStax
CosmosDB



HBase Meet Cloud – Benefits

Alibaba Group 阿里巴巴集团 APACHE HBASE

Cloud Native

New Hardware

RDMA
Flash
GPU
Non-volatile memory

Flexibility

Fast Add/Remove Resource
Insight
Fix bugs in time
Self-driven

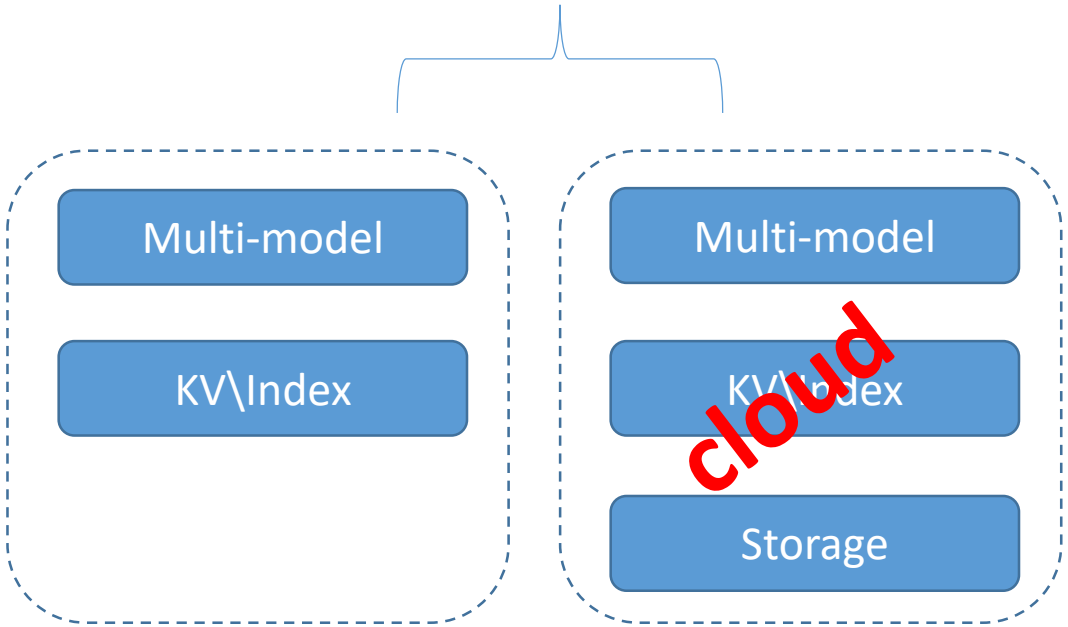
Cost Savings
(TCO)

End up paying for features
Flexibility
self-driven
Reduce human
……

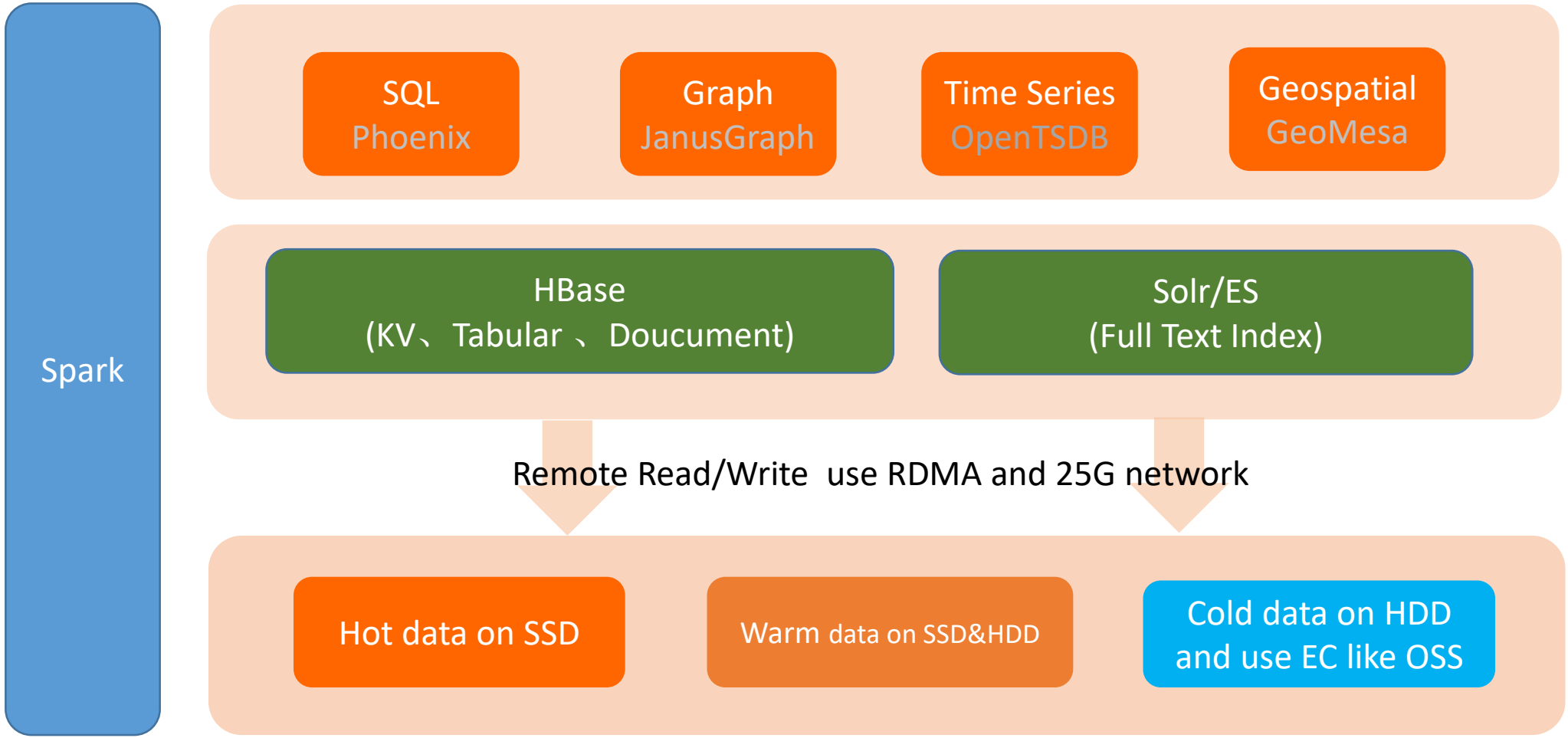
Multi-model - Native Or Layer

Neo4j
InfluxDB
CockroachDB
PG

HBase Ecosystem
DataStax
CosmosDB



ApsaraDB HBase Platform – Cloud Native



ApsaraDB HBase Platform Advantage

	Item	ApsaraDB HBase (ALiyun Product) https://cn.aliyun.com/product/hbase	Apache HBase (Sofeware)
Basic	High availability	99.9% ~ 99.99%	N/A
	Data reliability	99.999999999%	N/A
Online Ability	Multi-master clustering	Multi-master clustering， Multi-AZ/Regon	NO
	GC	FGC NO， YGC 5ms	GC 20s~100s， YGC 100ms+
Reduce Cost	Storage Cost	Cut by 50%+ on share cloud disk， Total 3 Copy	Maybe on Cloud Disk， Total 9 Copy
	Support Cold Storage	Support OSS， Cut by 70% at less read	NO
Multi-model DB	Multi-model DB	KV， Tabular， SQL， Graph， Time Series， Geospatial Full Text index， Search	KV， Tabular
Enterprise Characteristics	Disaster recovery	Backup and Restore	NO， maybe3.0
	Security	user/password， ACL	Kerberos， ACL
	Analytics	Spark on HBase ， More optimization	Spark on HBase
	Version upgrade	Automatic upgrade	N/A
Self-driven	Database control system	15min Create a DB/Monitor Online add storage and node/Elastic Power in future	N/A
	Diagnostic System	Big request ， Big Table merge， Hot Region ……	NO

欢迎加入HBase中文社区

- HBase中文技术社区 http://www.hbase.group/

技术社区微信公众号

钉钉技术交流群

求贤若渴

沈春辉

浙江 杭州

欢迎加入

杭州、硅谷、深圳、北京

扫一扫上面的二维码图案，加我微信

谢谢观看

Thanks