DTCC 2014中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2014
大数据技术探索和价值发现

Hadoop生态技术在阿里全网商品搜索实战

阿里巴巴 - 王峰

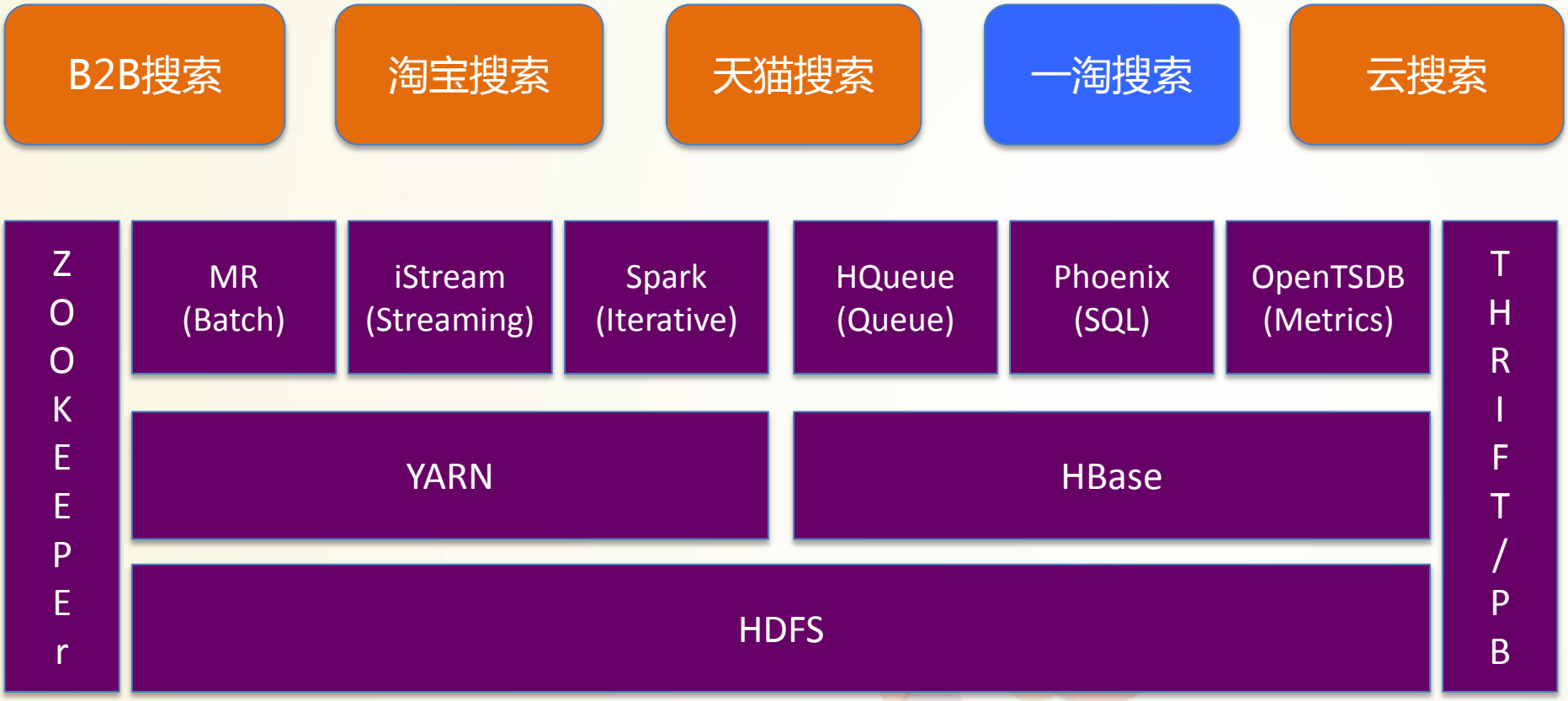
自我介绍

- 真名：王峰　　淘宝花名：莫问
- 微博：淘莫问
- 2006年硕士毕业后加入阿里巴巴集团
- 淘及搜索事业部（高级技术专家）
- 目前负责搜索离线系统团队
- 技术方向：分布式计算与存储

大纲

- 阿里搜索离线技术平台
- 阿里全网商品搜索系统架构
- 阿里电商网页库存储方案
- 阿里全网商品实时处理流程

阿里搜索离线技术平台


B2B搜索
淘宝搜索
天猫搜索
一淘搜索
云搜索
ZOOKEEPEr
MR (Batch)
iStream (Streaming)
Spark (Iterative)
HQueue (Queue)
Phoenix (SQL)
OpenTSDB (Metrics)
THRIFT/PB
YARN
HBase
HDFS

阿里搜索Hadoop/HBase集群

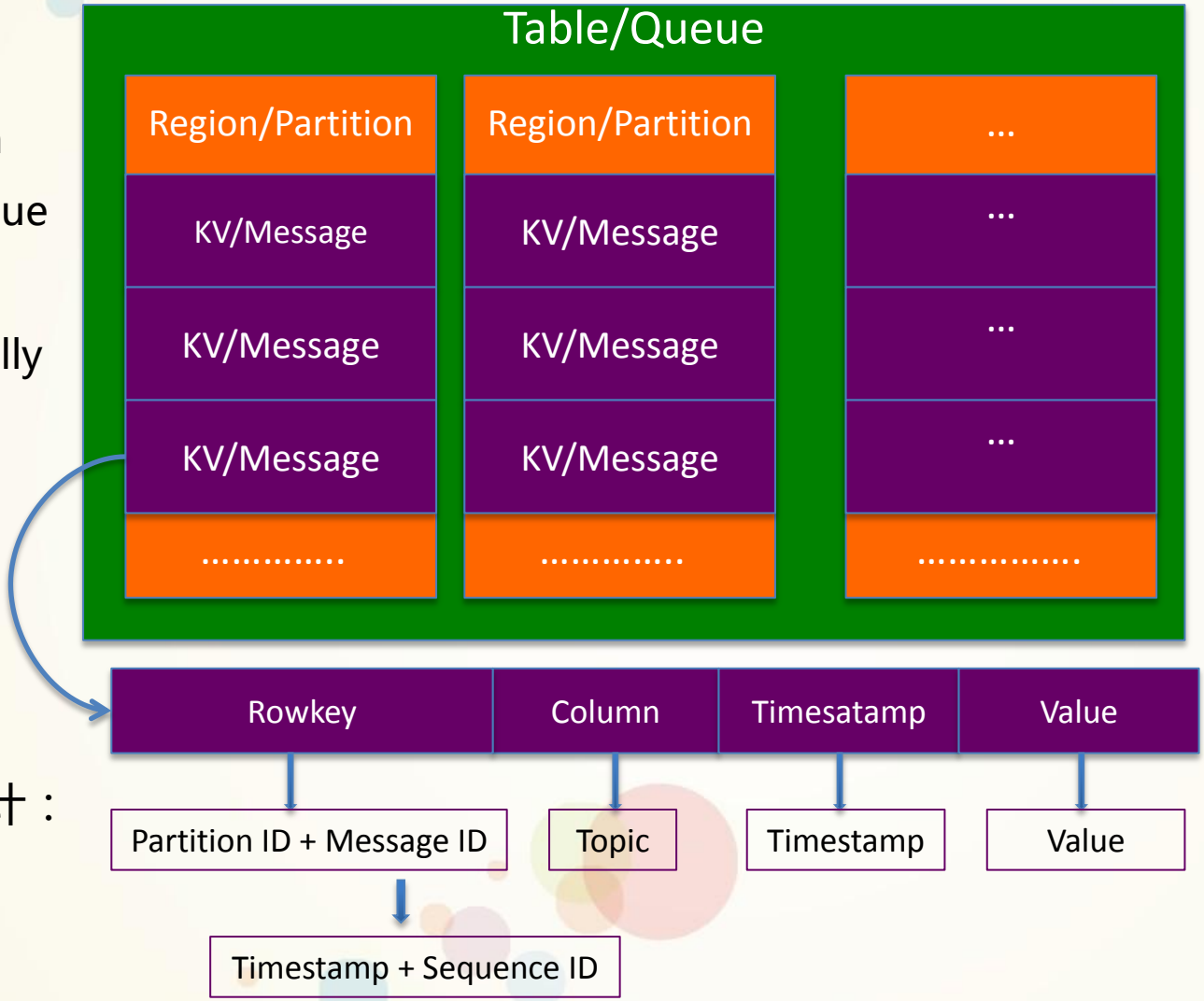
- Hadoop：基于2.2的阿里搜索定制版
- HBase：基于0.94的阿里搜索定制版
- 部署方式：Hadoop/HBase共同部署
- 集群规模：机器总数已近千台，分2个集群
- 硬件配置：
  - 24/32Core CPU
  - 48/96GB Memory
  - 12 * 1T/2T Sata Disk

阿里搜索分布式存储技术体系

- HDFS（分布式文件系统）
  - HBase（NoSQL数据库）
    - Phoenix（SQL On HBase）
    - OpenTSDB（Metrics On HBase）
    - HQueue（Queue On HBase）

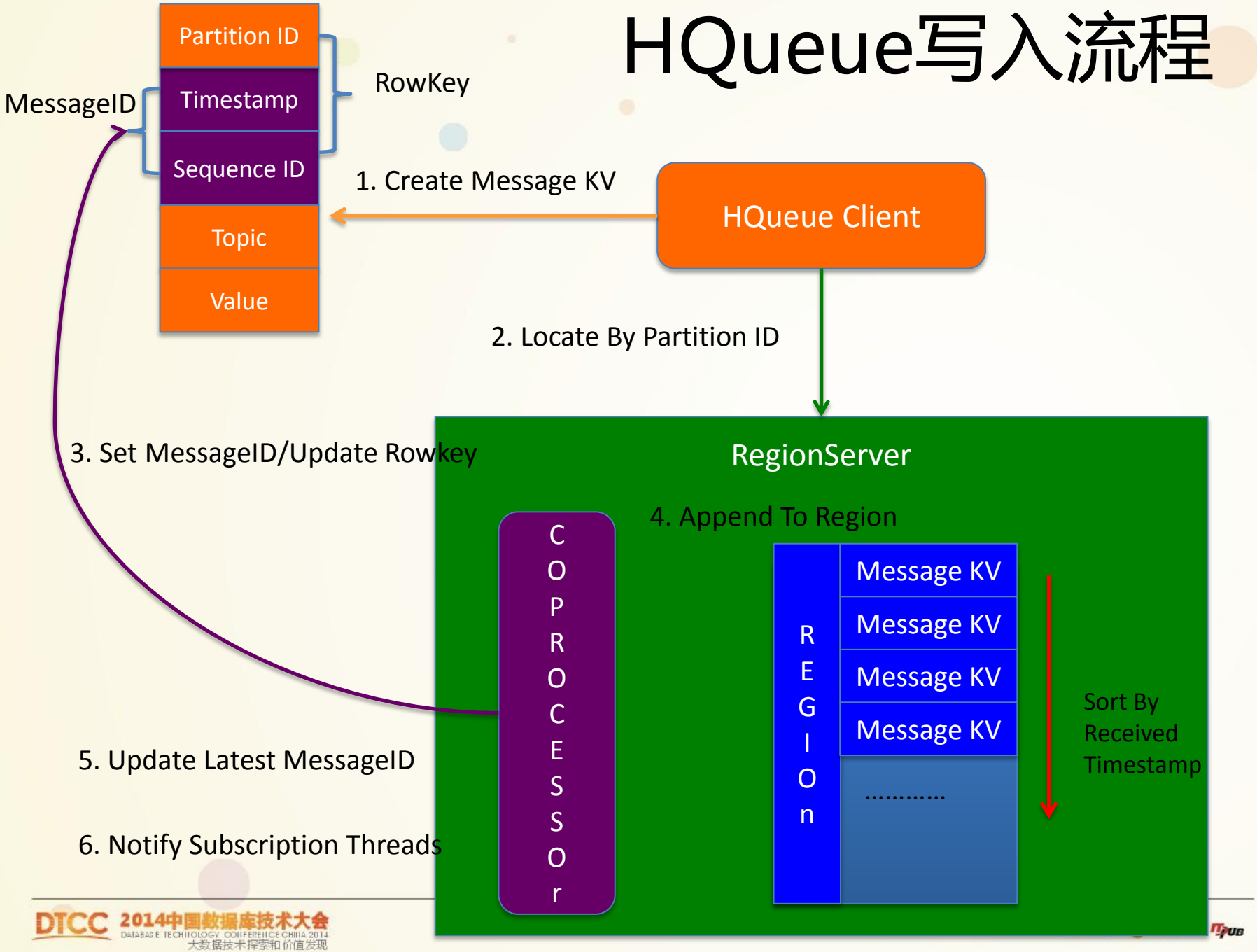
HQueue存储结构

1. Queue is a HBase Table
2. Partition is a HBase Region
3. Message is a HBase KeyValue
4. Message is stored in Partition/Region sequentially

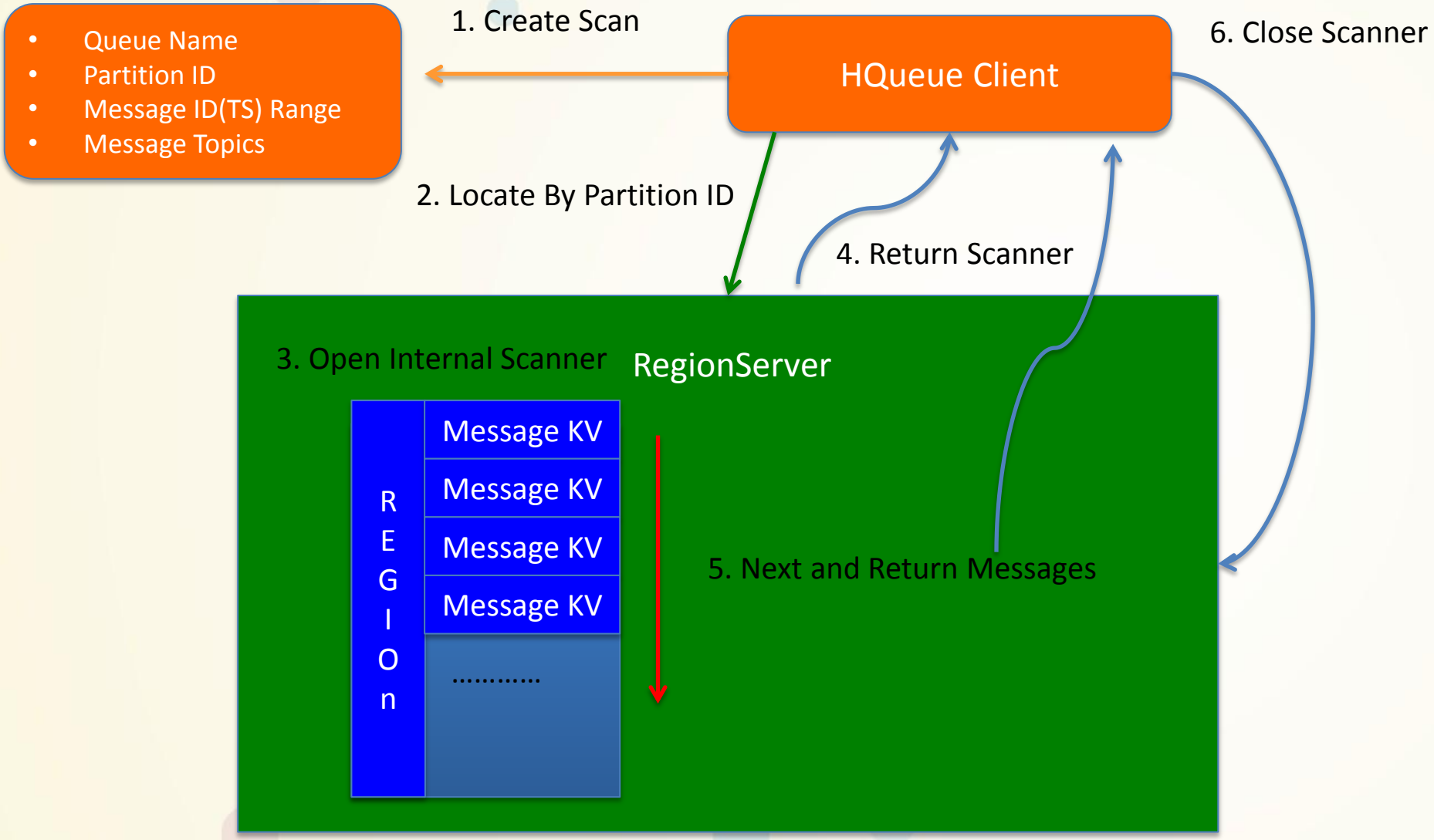


Message的KeyValue设计：

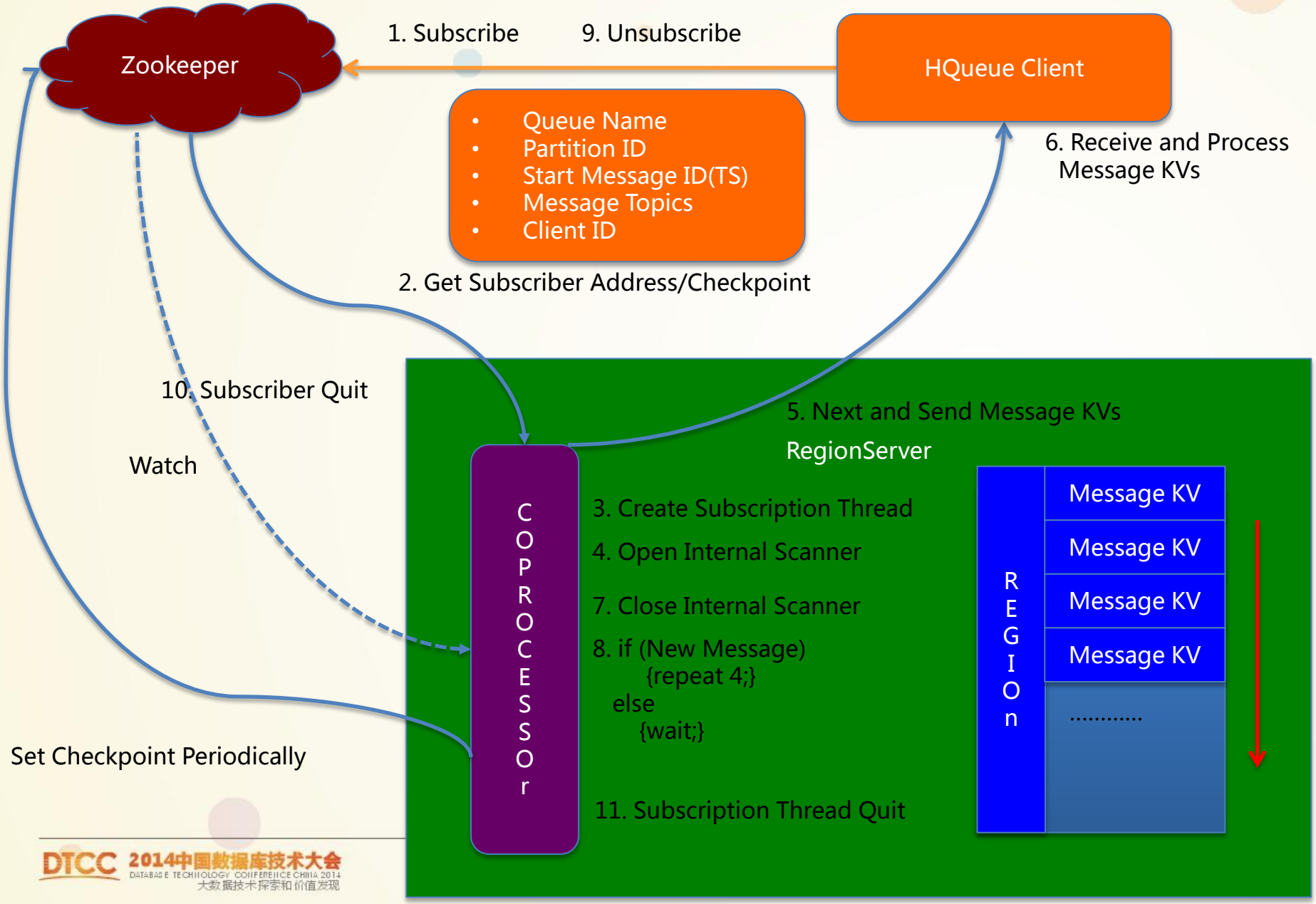
HQueue写入流程


Partition ID
Timestamp
Sequence ID
Topic
Value
RowKey
MessageID
1. Create Message KV
HQueue Client
2. Locate By Partition ID
3. Set MessageID/Update Rowkey
RegionServer
4. Append To Region
COPROCESSOr
REGIOn
Message KV
Message KV
Message KV
Message KV
.....
Sort By Received Timestamp
5. Update Latest MessageID
6. Notify Subscription Threads

HQueue读取流程


Queue Name
Partition ID
Message ID(TS) Range
Message Topics
1. Create Scan
HQueue Client
6. Close Scanner
2. Locate By Partition ID
4. Return Scanner
3. Open Internal Scanner
RegionServer
REGIOn
Message KV
Message KV
Message KV
Message KV
.....
5. Next and Return Messages

HQueue订阅流程


Zookeeper
1. Subscribe
9. Unsubscribe
HQueue Client
Queue Name
Partition ID
Start Message ID(TS)
Message Topics
Client ID
2. Get Subscriber Address/Checkpoint
6. Receive and Process Message KVs
10. Subscriber Quit
Watch
5. Next and Send Message KVs
RegionServer
COPROCESSOr
3. Create Subscription Thread
4. Open Internal Scanner
7. Close Internal Scanner
8. if (New Message)
{repeat 4;}
else
{wait;}
11. Subscription Thread Quit
REGIOn
Message KV
Message KV
Message KV
Message KV
.....
Set Checkpoint Periodically

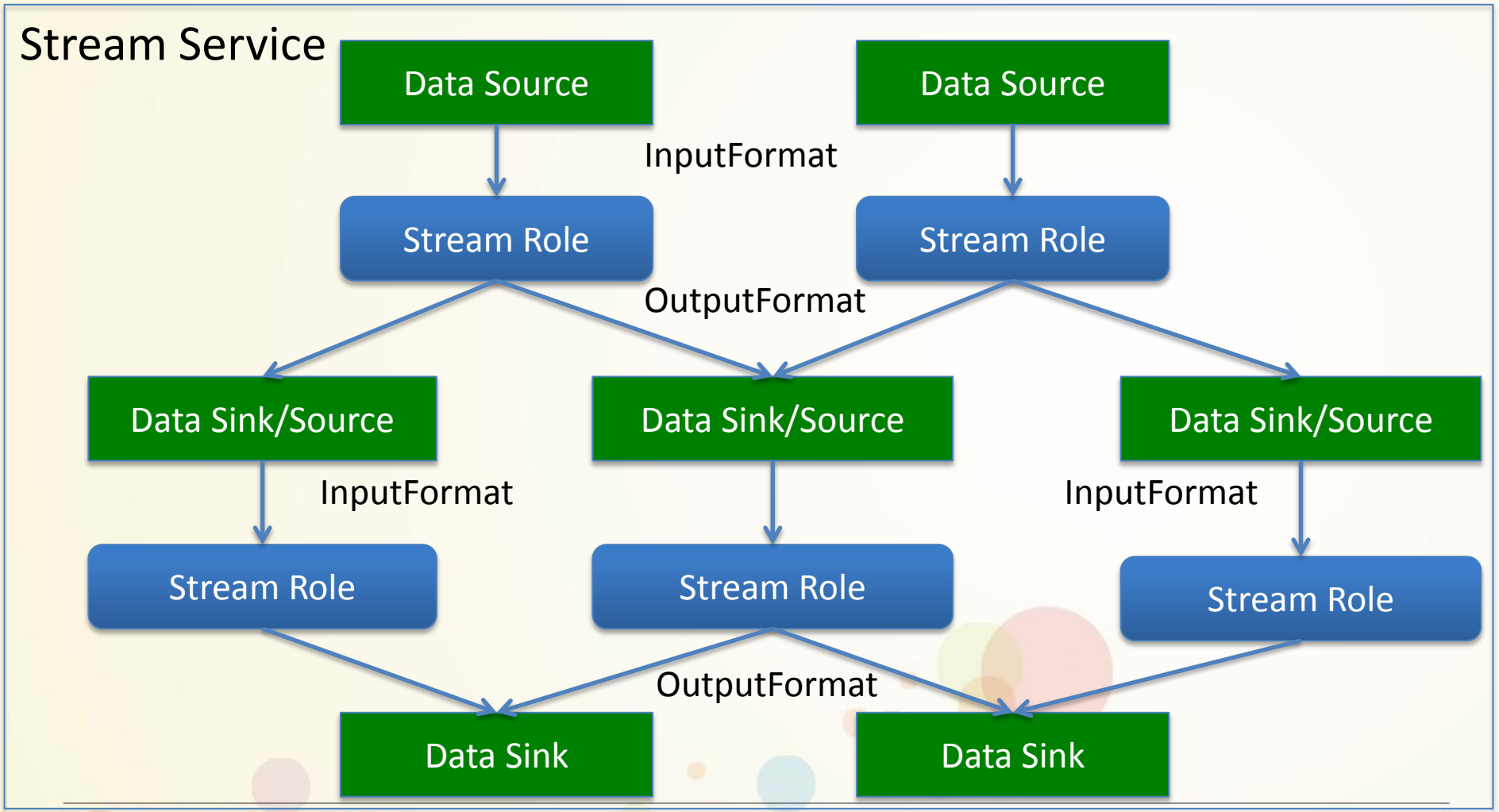
HQueue主要特性

- 读写高性能（新消息都在MemStore + 顺序存储）
- 消息持久化存储，不丢失（HFile + HLog）
- 消息支持TTL设定，自动清理过期消息（HBase TTL）
- 消息支持主动拉和订阅两种模式（HBase Client Wrapper + Coprocessor）
- 服务支持动态负载均衡（HBase Load Balance）
- 服务支持快速Failover (HBase MTTR)
- 支持多语言客户端（扩展HBase Thrift Server）
- 可与Hadoop计算平台无缝对接（HQueueInputFormat/OutputFormat）
- 可复用HBase集群直接部署管理，无需独立硬件（HQueue Shell）

阿里搜索分布式计算技术体系

- Hadoop YARN（统一管理所有计算模型）
  - MapReduce(批处理计算模型)
  - iStream（流计算模型）
  - Spark（迭代计算模型，规划中）

iStream - 基于YARN的流计算引擎



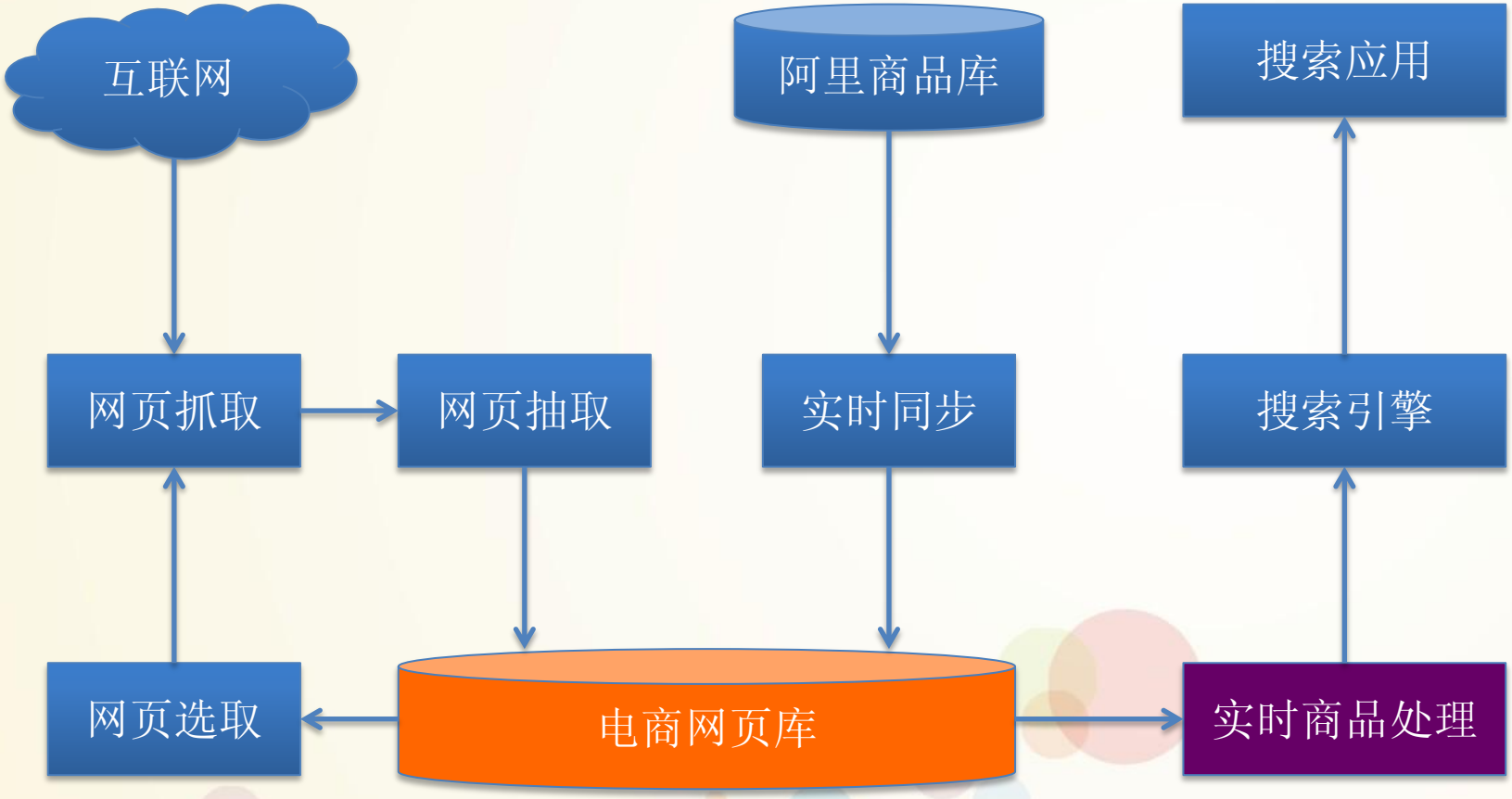
iStream技术概念

- Stream Service：运行在YARN上的一套流计算服务，例如：实时商品处理服务
- Stream Role：计算角色，即一组具备特定功能逻辑的Worker组合
- Data Source：计算服务的流式数据来源，可被切分为多个Data Split，一个Worker可以读取多个Splits。例如：数据库集群binlog服务，分布式消息队列
- Data Sink：计算服务的数据终端，可被切分为多个Data Partition，一个Worker可以向多个Partition分发写入。例如：数据库集群，分布式消息队列
- In/OutputFormat：控制Stream Role如何访问Data Source/Sink，不同存储介质的Source/Sink可以通过配套的In/OutputFormat来接入

iStream主要特性

- 计算和存储层分离，可灵活搭配消息队列
- 计算拓扑开放，可以根据业务变化动态调整
- 具备流处理进度管理能力，进度可视化以及监控报警
- 具备弹性调度能力，可根据进度动态调整计算资源数量
- 服务Metrics自动记录到OpenTSDB中，可WebUI查看
- 类似MR Streaming方式，支持多语言编程
- 可与MR等模型共享Hadoop集群，无需单独集群部署

阿里全网商品搜索系统架构



阿里电商网页库存储方案

- 发展历程
  - 2010年上线，学习Google网页库Bigtable存储方案，决定采用开源的HBase作为存储引擎
  - HBase经历了0.25，0.26，0.90，0.92，0.94（当前），5月将升级到0.98
  - 集群规模从30多台持续升级到300多台
  - Region数从1000多个增长到20000多个
  - 网页数从十亿增长到百亿

阿里电商网页库-Rowkey设计

- URL翻转
  - 例如：http://www.taobao.com 翻转为http://com.taobao.www
- 特点：同一网站内的网页/商品连续存储，各站点连续分布
- 优点：可以方便的在海量网页库中快速读取到某个站点的所有数据
- 缺点：不同网站的网页/商品数量以及变化频率差异较大，不同Region/RegionServer之间的I/O访问容易不均衡
- 解决方案：通过统计分析和抓取调度反馈，针对热点网站进行手动split，分解压力；升级到0.98后将会采取Stochastic Load Balancer根据I/O压力等综合指标进行负载均衡

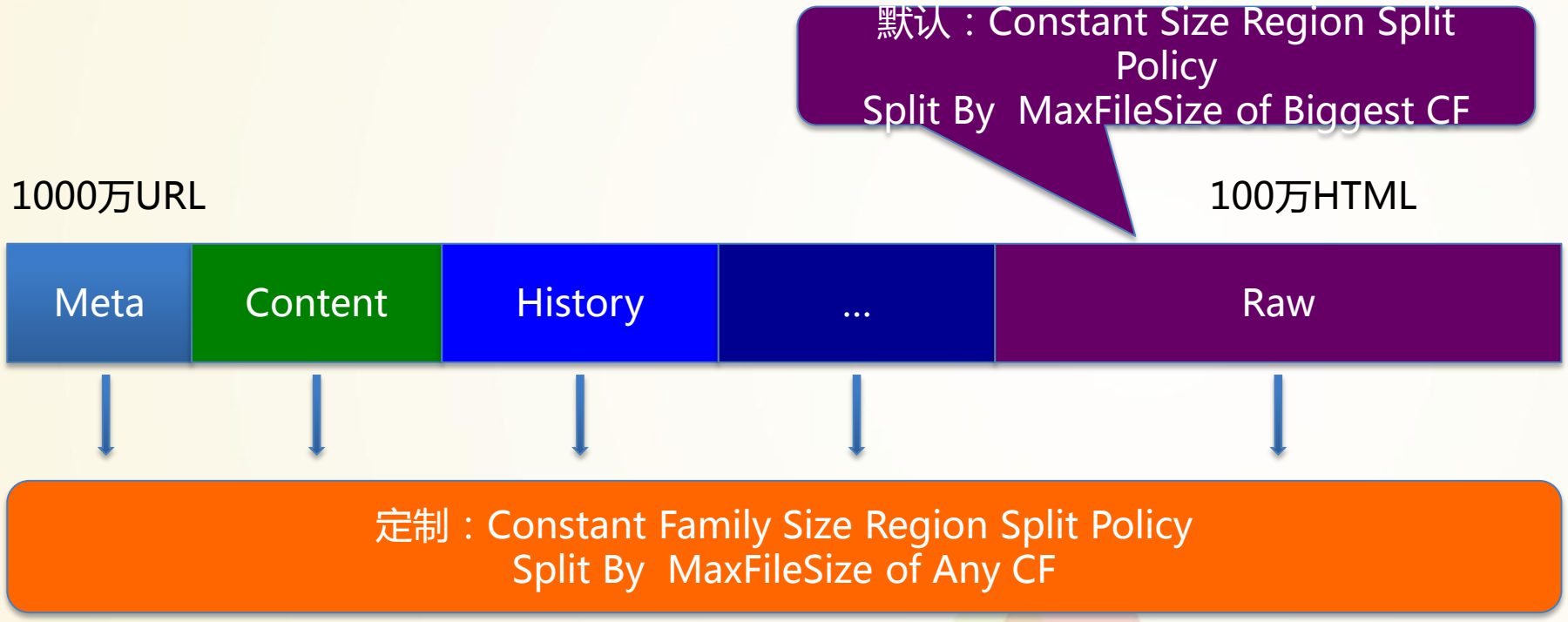
阿里电商网页库-CF设计

Column Family	内容	描述
Meta	网页元信息	url，host，type等
Content	网页抽取出的结构化信息	标题，价格等
Outllinks	网页的外链信息	页面的url链接
Algorithm	算法结果	分类，权重等
History	历史信息	多版本历史价格，销量等
Trace	网页处理的trace信息	时间点，错误信息等
Image	网页中的图片信息	图片url等
Raw	网页原始HTML	HTML

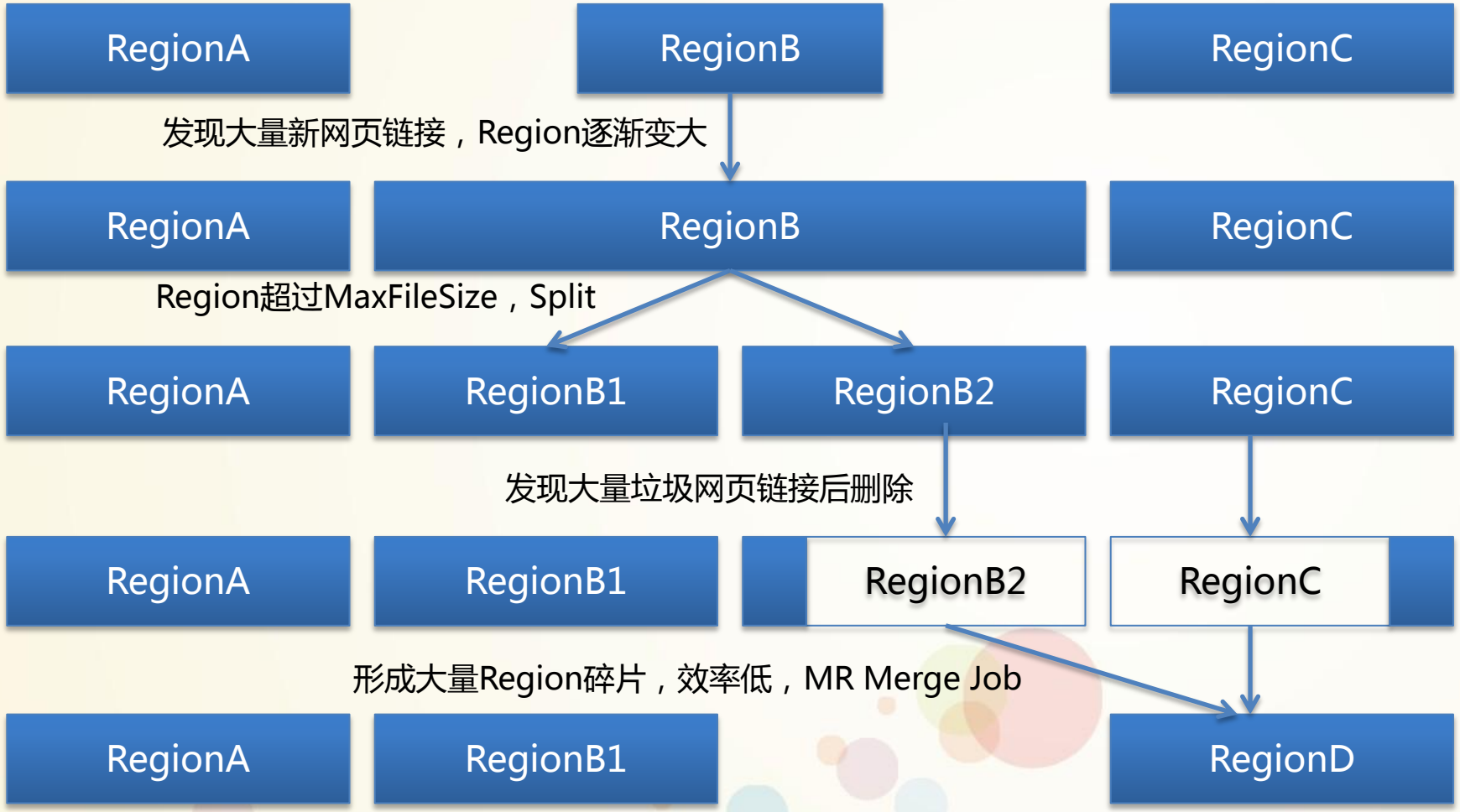
阿里电商网页库-I/O设计

参数名	参数值	描述
Compression	Snappy/Gzip	Meta，Content等CF访问频繁，用Snappy，速度快 Raw CF访问较少，用Gzip，压缩比高
Block Encoding	Diff	0.98后采用PrefixTree
Block Size	64KB-1MB	Meta，Content等CF都有Get需求，Block Size设置较小 Raw CF 只有Scan的需求，Block Size设置较大
Block Cache	InMemory	Meta CF内容较少，访问频率较高，InMerory=True，减少Cache换出概率
Bloom Filter	ROW	所有CF基本都是Row级别访问，无需ROWCOL

阿里电商网页库-Region Split



阿里电商网页库-Region Merge



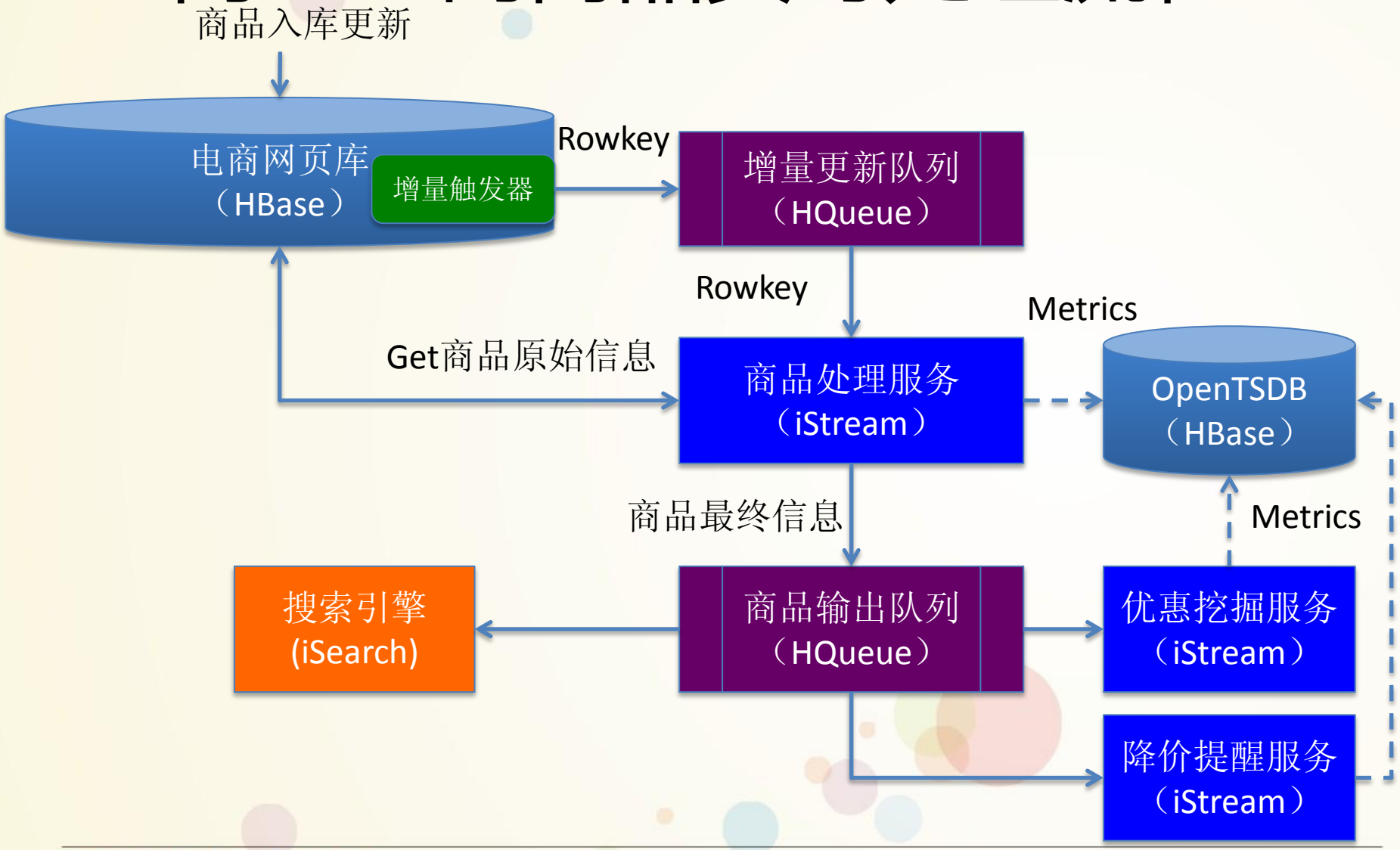
阿里电商网页库-Coprocessor

- Trace Coprocessor
  - 当字段值有变化时才进行写入，有效记录、跟踪字段变化趋势，例如：History CF，跟踪记录商品不断变化的历史价格
- Clone Coprocessor
  - 将put中某个cf中的column自动复制到其他cf中，例如：各个cf中需要追查的字段复制到log多版本cf中，方便追查问题
- Incremental Coprocessor
  - 如果put中的column符合某种条件，则将此put的rowkey推送到特定的hqueue中，实现增量更新队列，方便下游增量处理服务

阿里全网商品实时处理流程

- 全网商品搜索和全网网页搜索的区别：
  - 全网网页搜索：
    - 规模大，千亿级别
    - 整体时效性要求不高，索引分级构建
    - 覆盖率即使有短期遗漏，依然可以凭借庞大相关网页进行弥补
  - 全网商品搜索：
    - 规模中等，电商网页库百亿级别，有效商品页面几十亿
    - 时效性要求高，尤其是价格和库存状态，一天更新多次，且需及时生效
    - 覆盖率要求高，所有站点商品必须全部囊括，否则比价将出现缺陷

阿里全网商品实时处理流程



欢迎加盟阿里搜索！

微博：淘莫问