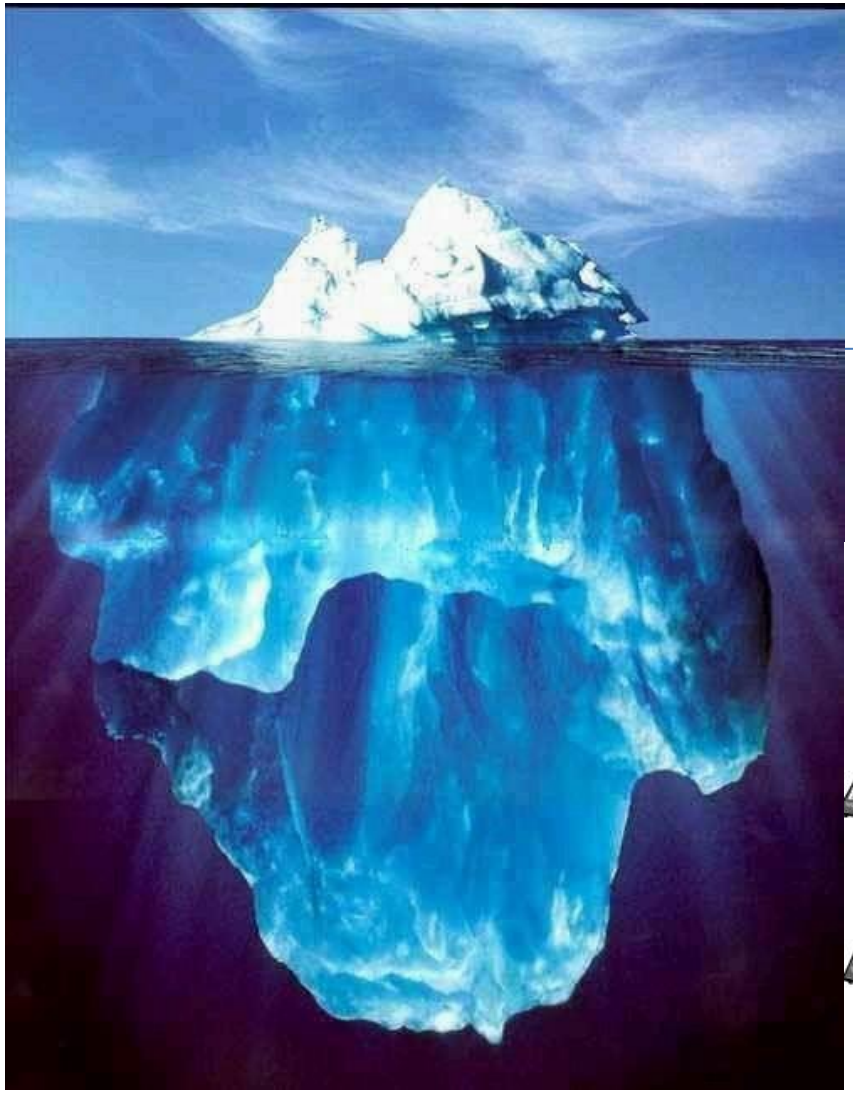
DTCC 2014中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2014
大数据技术探索和价值发现

# “大云”Hadoop平台及应用


王宝晗
中国移动研究院 云计算系统部
wangbaohan@chinamobile.com

# 电信运营商具有更多的数据



**移动互联网服务商**

专业SNS　博客　消息

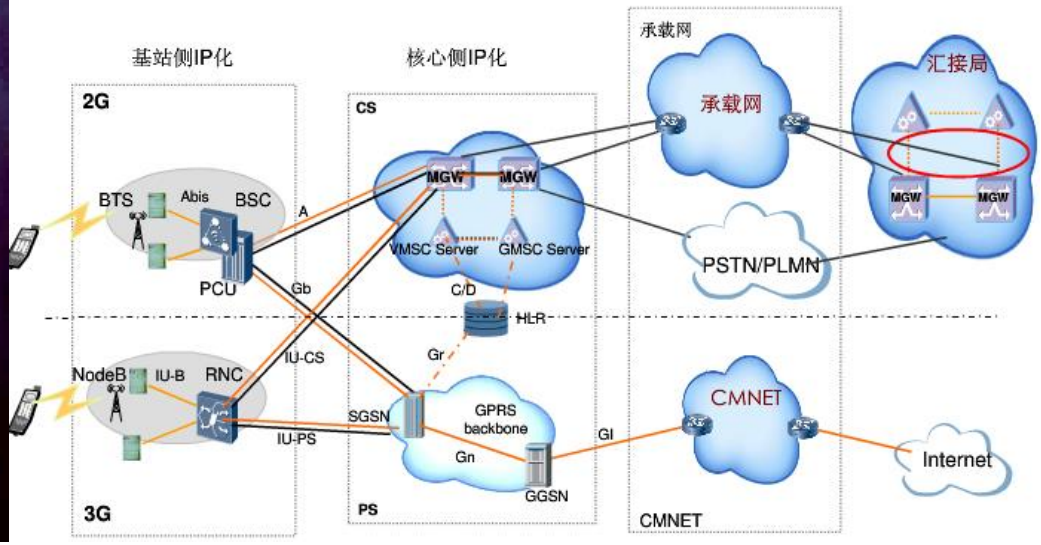
电商　图片　视频　优惠券

新闻　点评　音乐　微博

地图　问答　SNS　签到　论坛

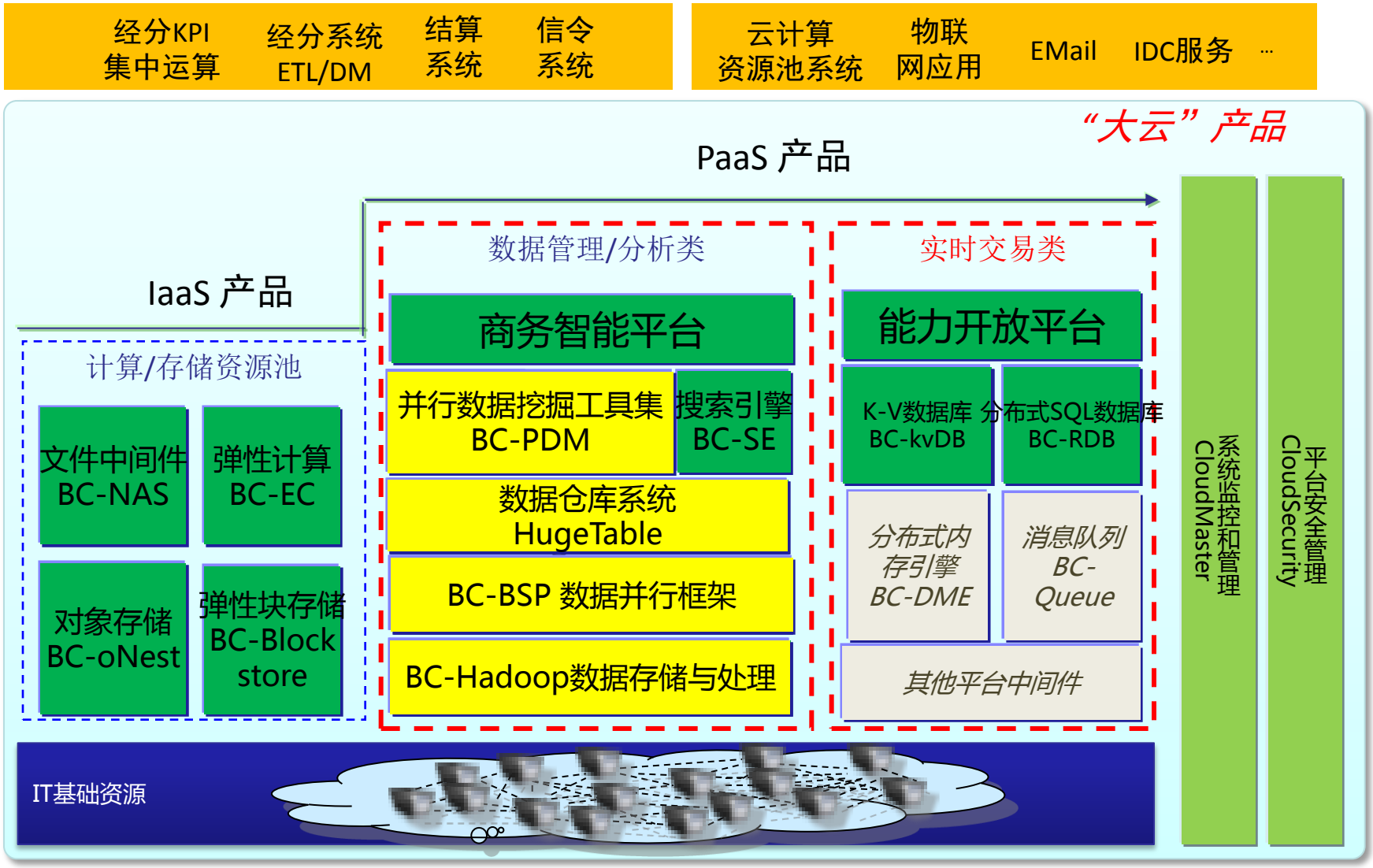
**电信运营商**

2G、3G、4G、WIFI



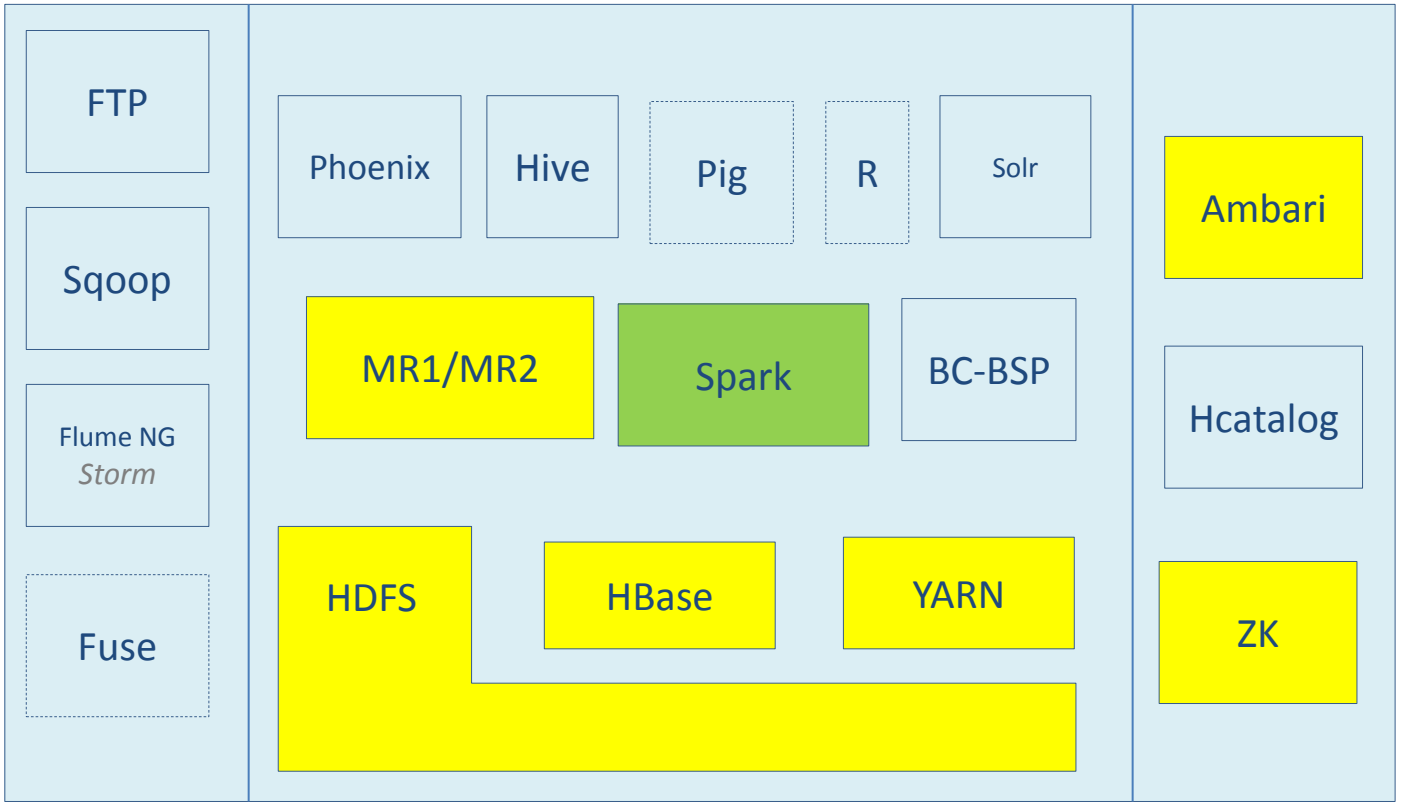

除了像移动互联网服务商那样关注“结果”，电信运营商还需要关注“过程”！

# 中国移动“大云” 云计算平台



# BC-Hadoop项目介绍

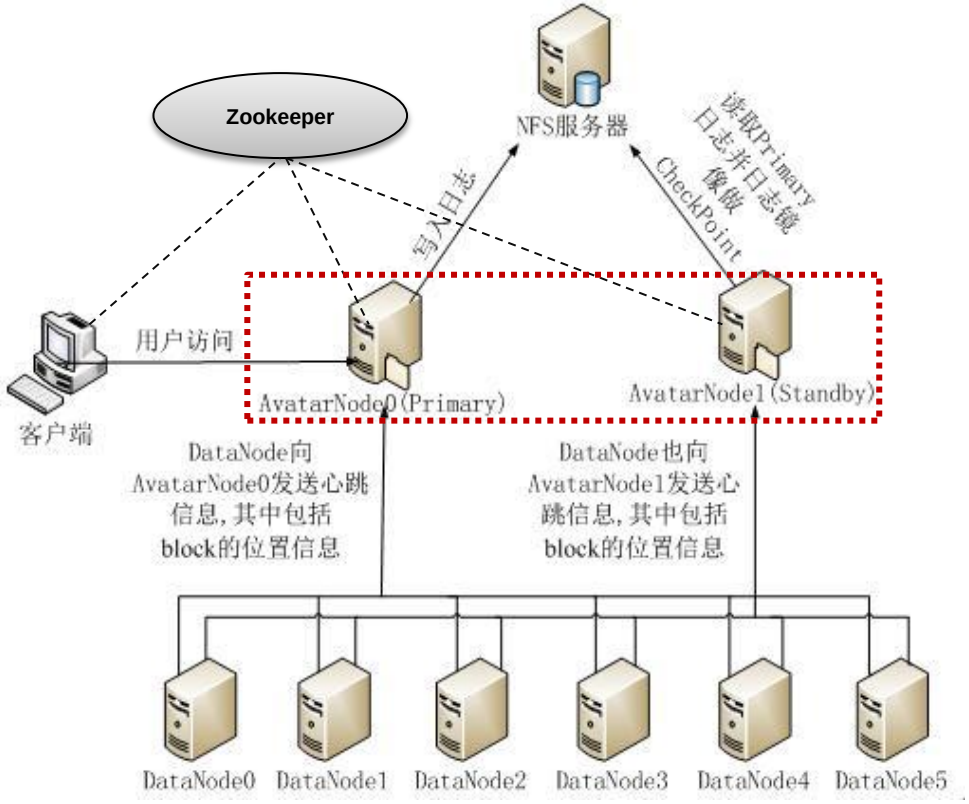
**BC-Hadoop：对开源Hadoop/HBase进行扩展和增强，为大云其他组件提供基本的存储计算能力。分别基于Hadoop 1.0 和 2.0 提供1.0和2.0两个版本。**

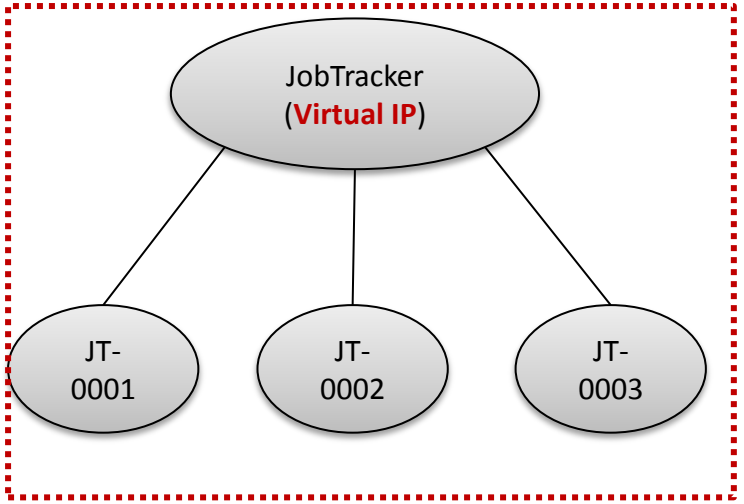



**黄色框** 是 BC-Hadoop的组件，正在整合Spark

# BC-Hadoop 1.0 主节点HA

参考Facebook **AvatarNode** 的实现，采用双主NameNode的自动故障检测与切换，大大缩短了NameNode切换时间和对应用系统的影响




实现了多个JobTracker的自动故障检测和切换

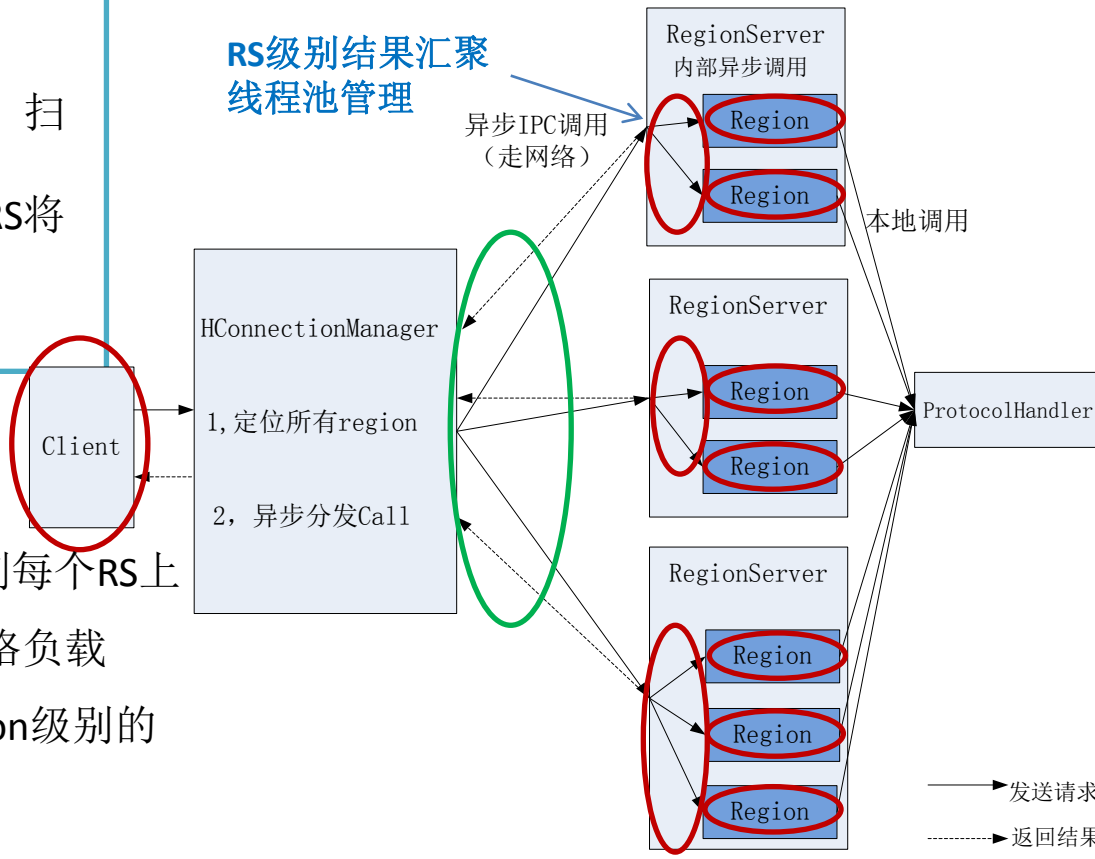
- 多个JobTracker启动并注册到Zookeeper
- 选举其中一个JobTracker作为Active
- 作业状态数据保存在HDFS
- Failover时，从HDFS读取作业数据，并继续执行作业

# HBase Coprocessor优化 – CP本地汇聚

目前Apache Hbase社区的实现机制是以Region为单位执行请求，每个请求直接发送到Region上，每个Region执行处理后将结果直接返回给Client

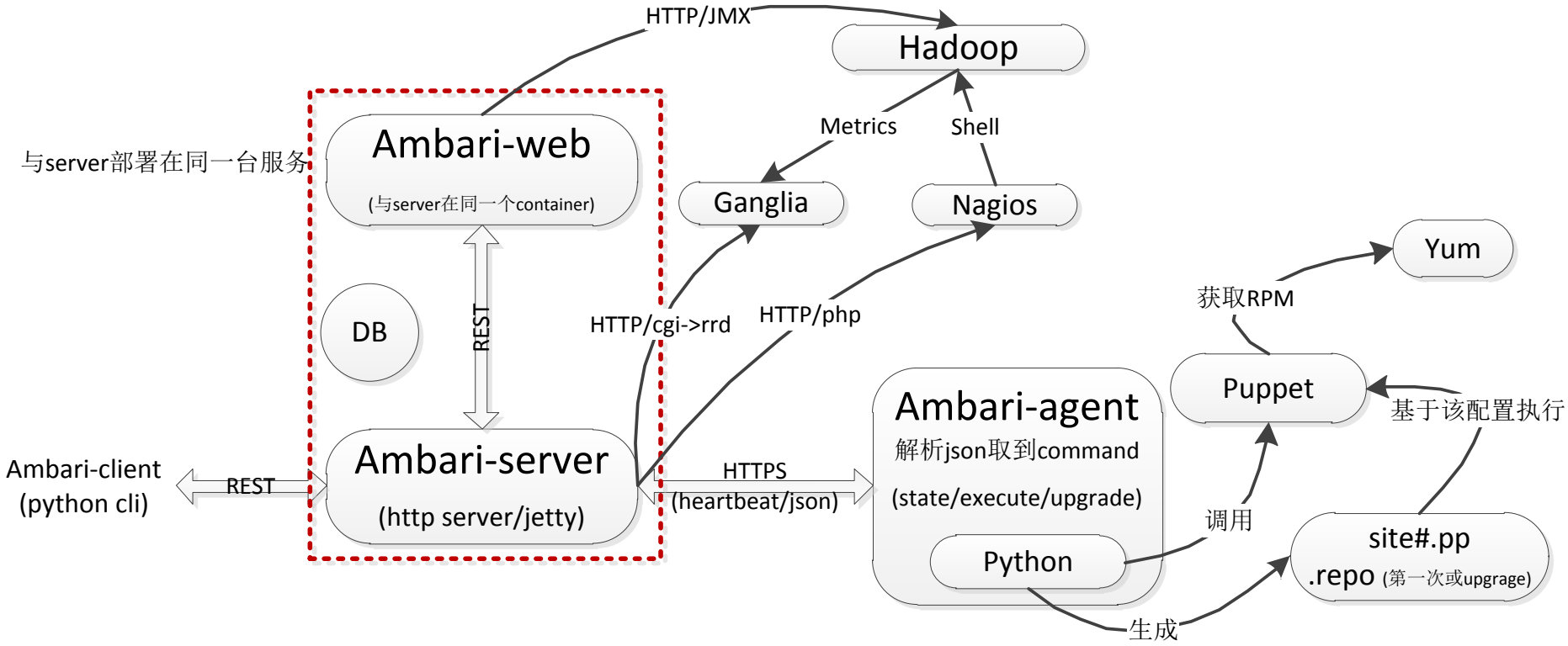
- Coprocessor本地汇聚
  - 以RS为单位发出CP计算请求
  - 每个RS对其管理的Region并行扫描，扫描结果在RS节点先做一次汇总
  - 当RS上所有Region均计算完毕，则RS将其本地汇聚结果返回给Client
  - Client将各RS返回结果进行汇总

- 优点
  - 计算分摊：Client端的计算被分布到每个RS上
  - 减轻网络负载： 减轻Client端的网络负载
  - 编程灵活：可以分别定义RS和Region级别的处理函数

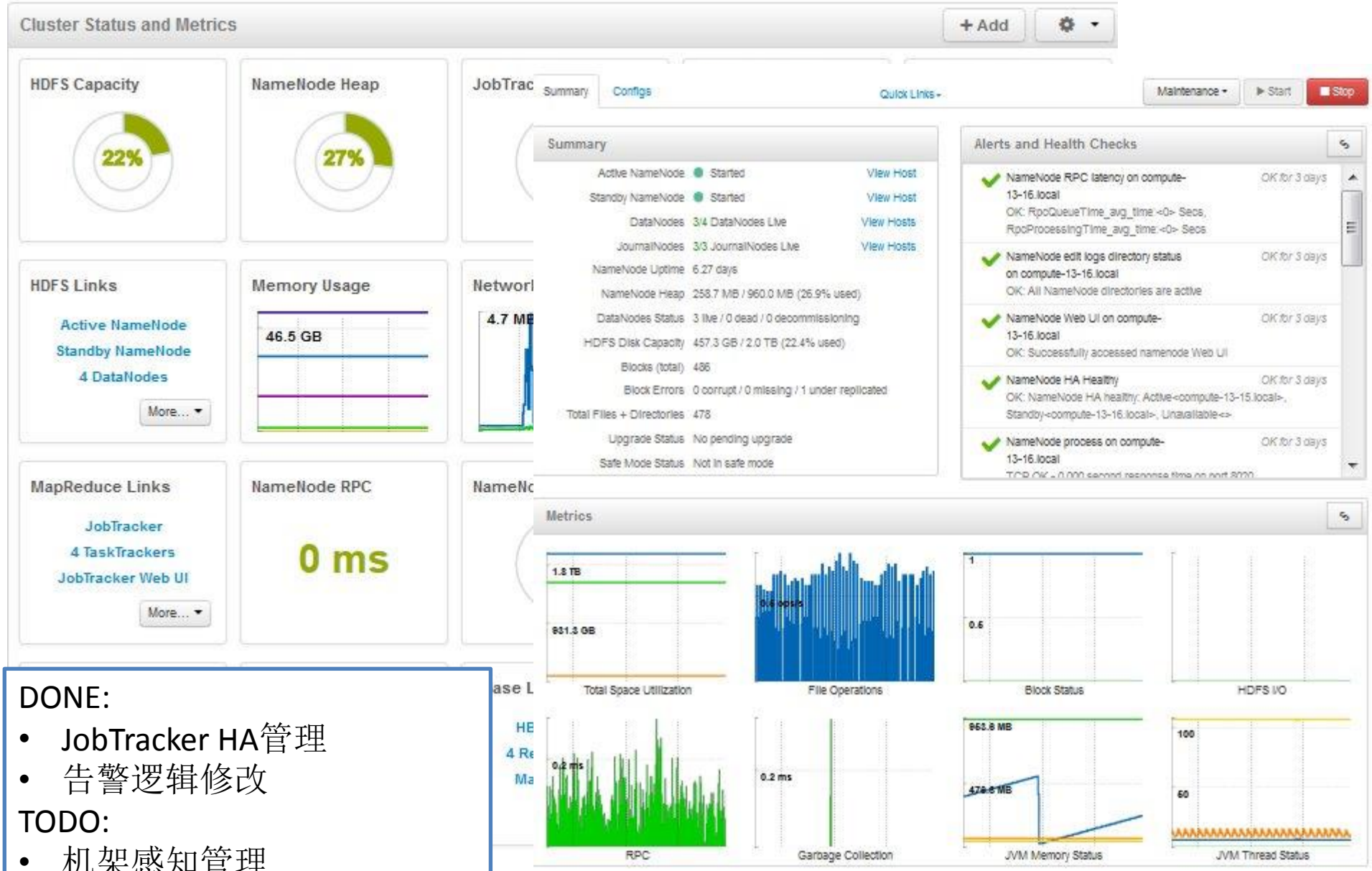


# 基于Ambari的Hadoop监控管理工具

- Apache Ambari是对Hadoop进行部署、监控和管理的开源项目
  - Puppet部署和管理hadoop服务
  - Ganglia 收集hadoop 服务数据与生成图表
  - Nagios监控集群服务状态并报警



# 基于Ambari的Hadoop监控管理工具




DONE:

- JobTracker HA管理
- 告警逻辑修改

TODO:

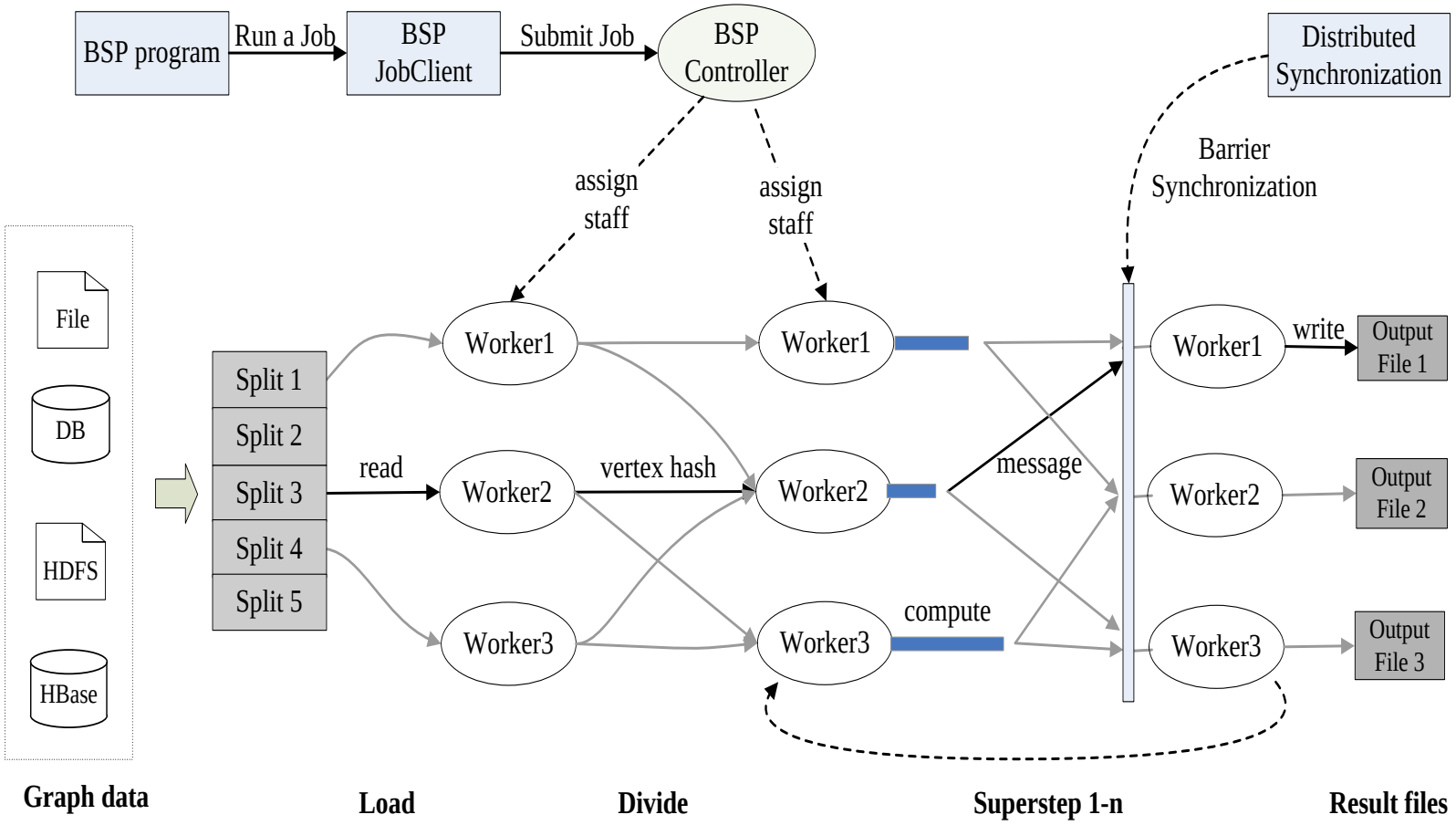
- 机架感知管理

# 走向YARN

- 功能测试及源码熟悉
  - 资源管理：cpu，内存
  - 资源调度：fair scheduler，DRF
  - Appmaster实现机制
- TODO
  - 集成计算框架如spark
  - 多租户，资源抽象由slot变为<vcore,mem>

# 图计算平台（BC-BSP）

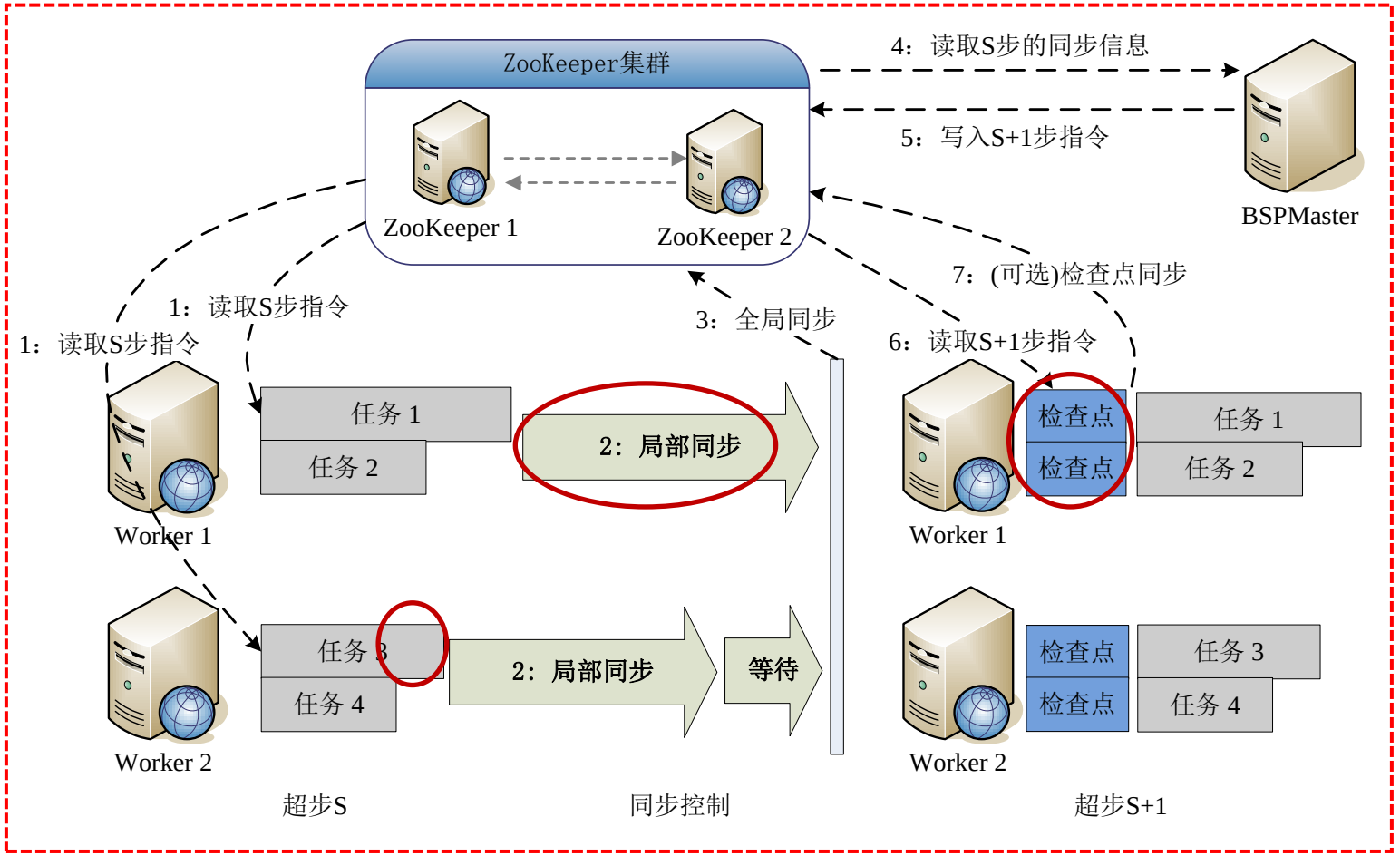
BC-BSP：针对社交网络分析、用户精准营销、搜索引擎PageRank计算等图计算领域的数据挖掘需求而研发的并行计算框架，针对迭代计算，计算效率优于MapReduce框架




http://www.github.com/cmri/

# 图计算平台（BC-BSP）

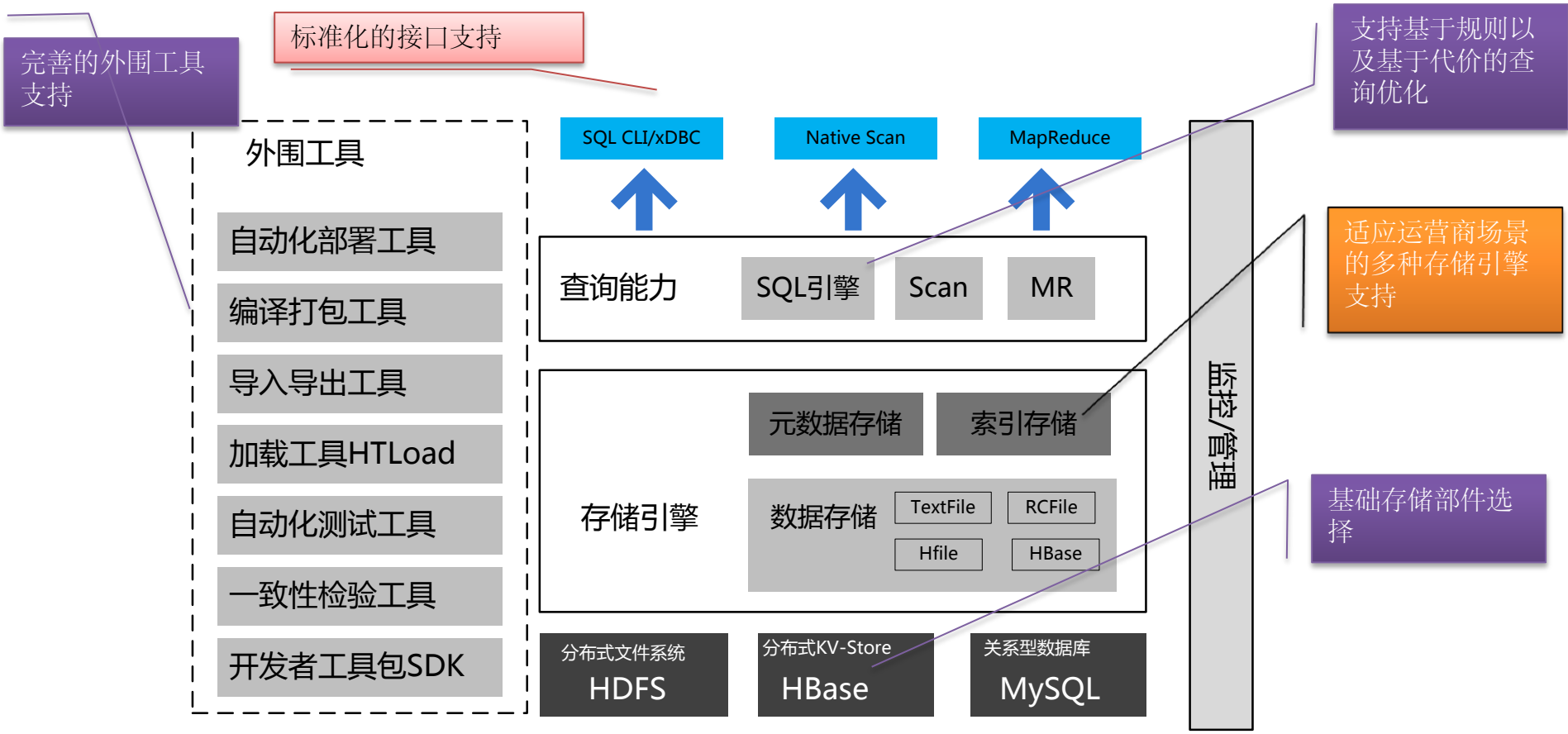
大同步并行:




- 先对同一个Worker上同一Job的多个Task进行局部同步，再以Worker为粒度进行全局同步
- 周期性进行数据检查点保存，与自动恢复
- 在计算任务的同时进行数据交换，降低数据交换时间

# 数据仓库系统（HugeTable）

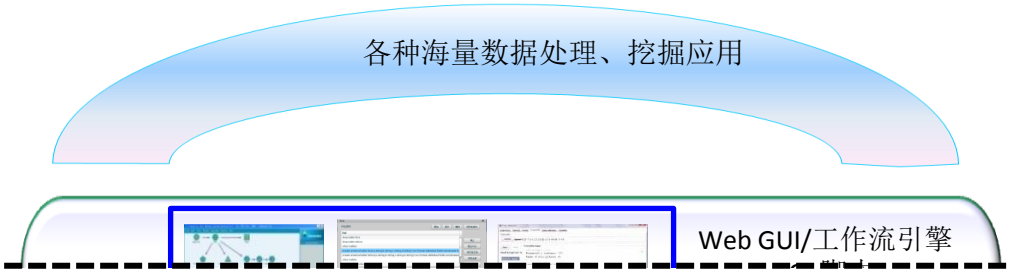
设计目标：具备海量数据管理能力；满足网管、经分、增值业务系统需求；方便的整合现有应用




- 支持数据的IUD操作
- HBase存储引擎：支持同一份数据进行实时查询和统计分析：Hive直接读Hfile进行统计，通过HBase实时查询
- Join优化：按照join key将两个表的数据存储在同一个HBase Table的不同column.

# 并行数据挖掘工具集（BC-PDM）

BC-PDM：支持SaaS模式的海量数据并行处理、分析与挖掘系统。适用于经营决策、用户行为分析、精准营销、网络优化、移动互联网等领域的智能数据分析与挖掘应用

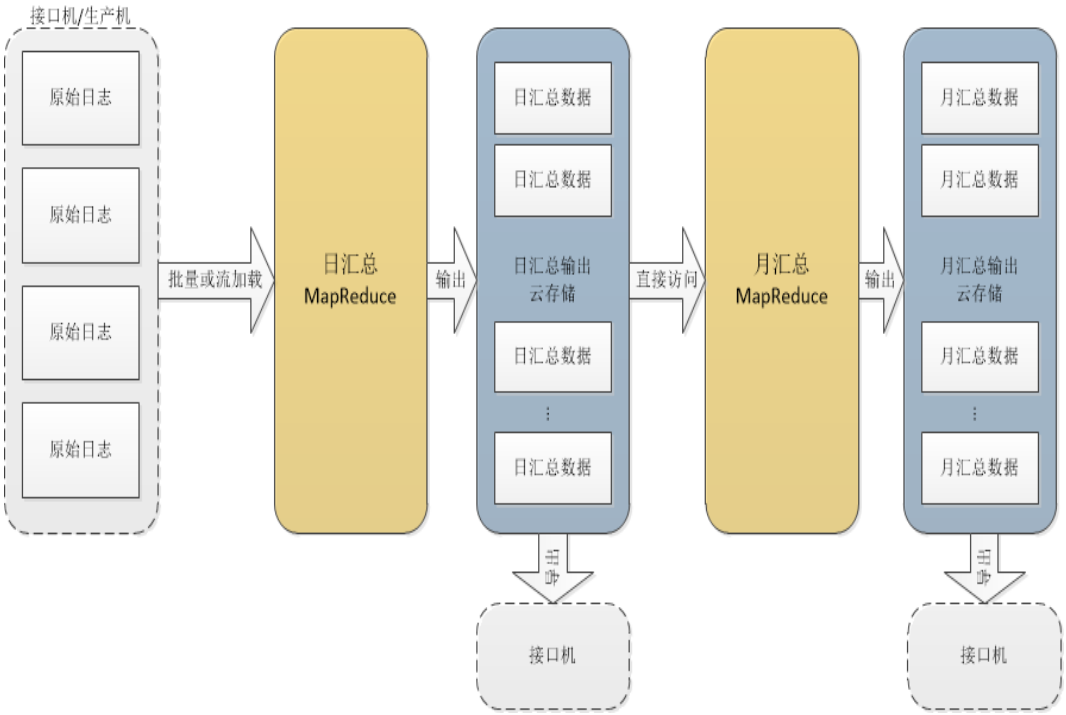



## 主要特点

- 数据交换：支持与RDB直接交换数据、支持CSV格式数据
- 数据ETL：支持数据清洗、转换、集成等7大类45种ETL

# 典型的应用场景之一：大数据批处理系统

**目标：针对海量结构化、非结构化数据的ETL操作。从各种数据源获取数据，并进行清洗、转换、去重、缺值补充等操作。通常采用MapReduce等并行计算技术。**




例图：分时段汇总的业务场景

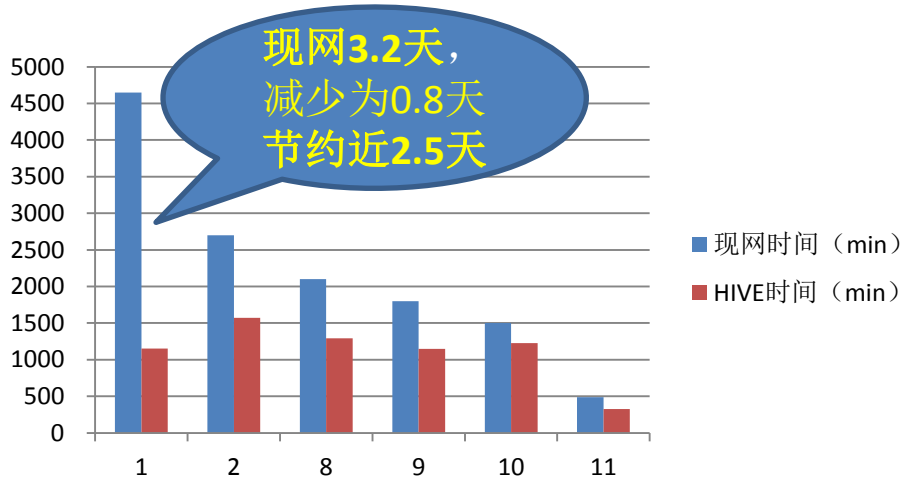
技术要求举例：

- ✓针对海量数据实时离线批处理运算（ETL），通常时间要求较为宽松，如几个小时级别。
- ✓数据ETL运算种类多，灵活性强，通常具有很强的定制化特征
- ✓数据通常需要导出到数据库、数据仓库，提供报表能力
- ✓需要灵活的调度的系统，便于系统需要和其他业务系统混合部署，提高资源利用水平

# “大云”应用案例之一：大数据ETL业务

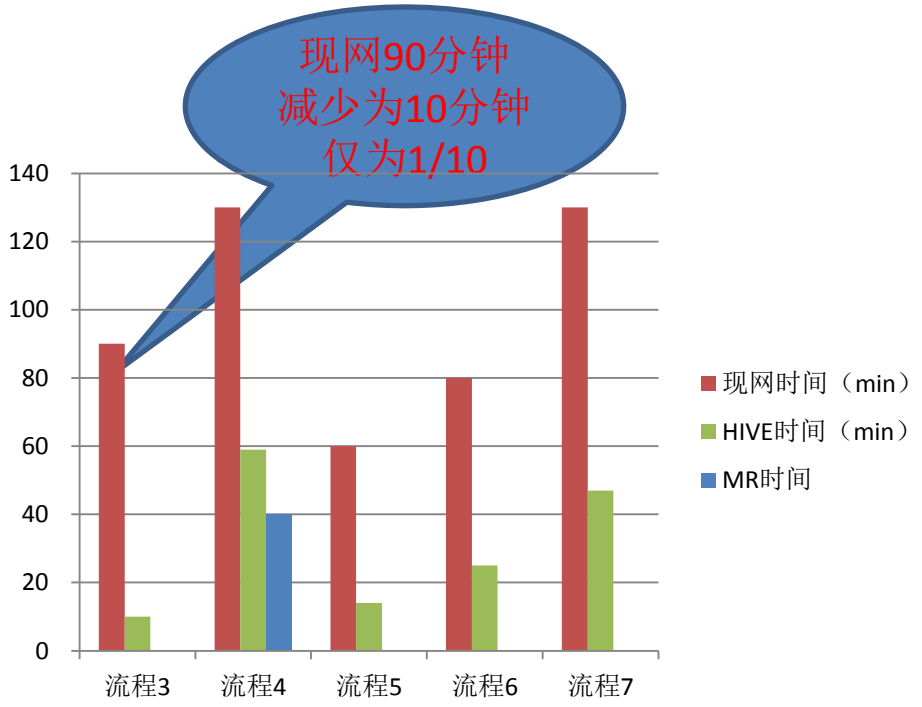
**滚详单类**

| 流程 | 现网时间（min） | 云ETL时间（min） | 加速比例 | 时间减少绝对值（小时） |
|---|---|---|---|---|
| 1 | 4650 | 1153 | 4.03 | 58.3 |
| 2 | 2700 | 1571 | 1.72 | 18.8 |
| 8 | 2100 | 1293 | 1.62 | 13.4 |
| 9 | 1800 | 1150 | 1.56 | 10.8 |
| 10 | 1500 | 1225 | 1.22 | 4.6 |
| 11 | 490 | 325 | 1.51 | 2.8 |



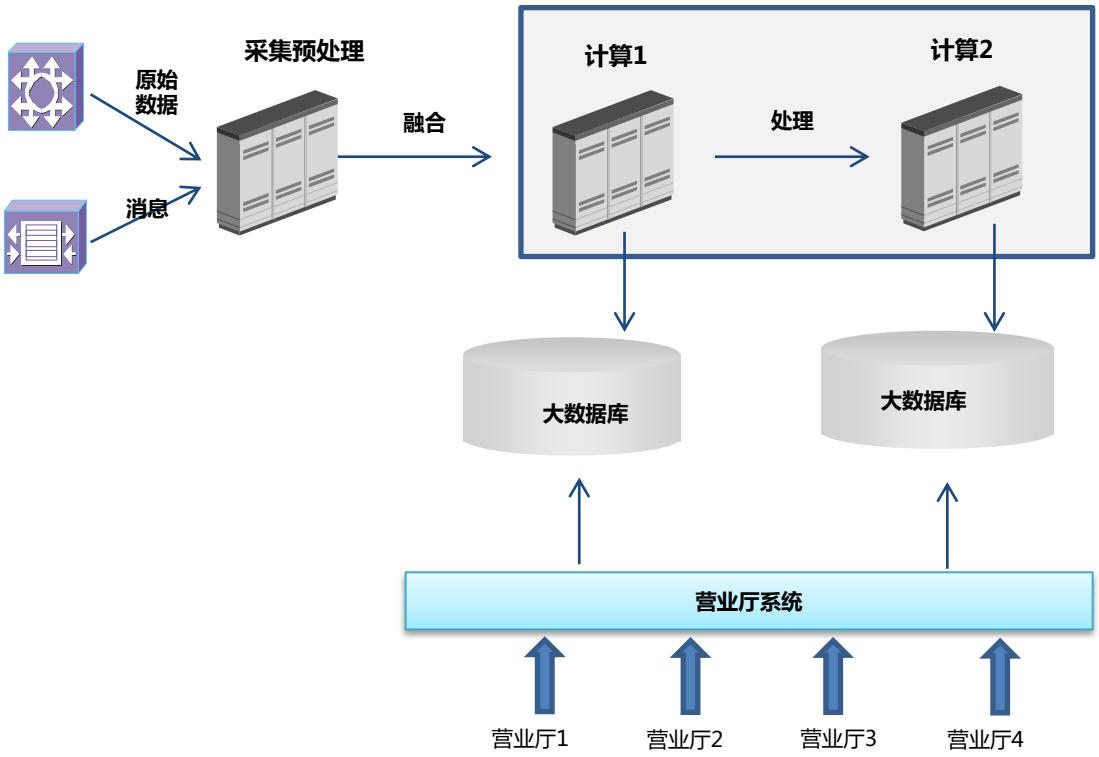

**出月表类**

| | 现网时间（min） | 云ETL时间（min） | MR时间 | 云ETL脚本加速比例 | 时间减少绝对值（小时） |
|---|---|---|---|---|---|
| 流程3 | 90 | 10 | 无 | 9.00 | 1.3 |
| 流程4 | 130 | 59 | 40 | 3.25 | 1.5 |
| 流程5 | 60 | 14 | 无 | 4.28 | 0.8 |
| 流程6 | 80 | 25 | 无 | 2.50 | 0.9 |
| 流程7 | 130 | 47 | 无 | 2.76 | 1.9 |



# 典型的应用场景之二：大数据查询系统

**目标：针对海量结构化、半结构化数据的精确定位、区段扫描等条件查询操作，用于网络优化、帐详单查询、故障定位、搜索引擎等业务场景。**




例图：帐详单查询系统

技术要求举例：

- ✓针对海量数据实施交互式查询，返回时间在1秒钟左右。
- ✓针对海量大数据规模实施查询，数据规模可以达到100TB-10PB规模。
- ✓数据插入通常采用批处理方式，而查询通常带有条件，通常返回结果数较少
- ✓系统具备较高的并发性，支持大量用户同时查询，依然可以在给定时间出口返回结果
- ✓数据具有很高的可靠性和可用性要求

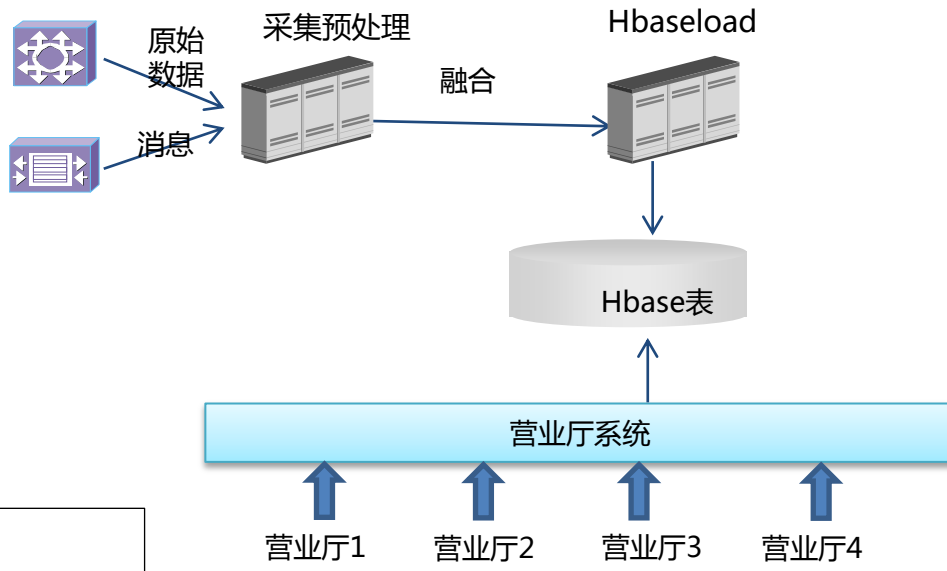
# “大云”应用案例之二：帐详单存储查询

**帐详单系统存储数量急剧膨胀，传统架构难以满足当前业务运营要求，系统面临扩容难题**

**方案介绍：**

- ✓ 某地市应用，每个月帐详单总体数据量100TB。
- ✓ 话单通过Bulk Load工具批量加载。
- ✓ 根据业务规则，key设计为：

  **Presplit+电话号码+业务类型+业务时间**




**运维经验：**

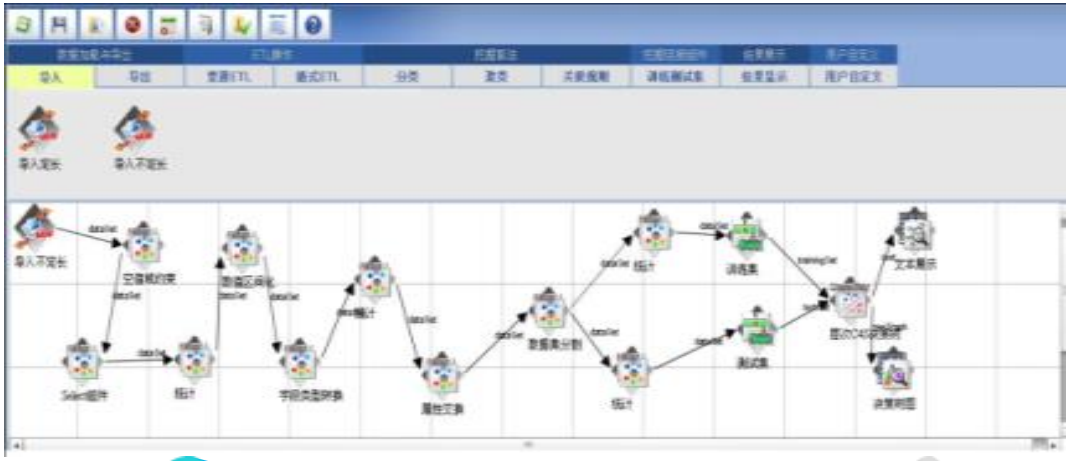
- ✓ 根据业务特点，预建分区，尽量避免split。
- ✓ 建议打开bloomfilter，提供查询效率。
- ✓ 查询请求量小时执行major_compact，合并小文件。
- ✓ 为了提高查询效率，单机cpu使用率最好低于60%。

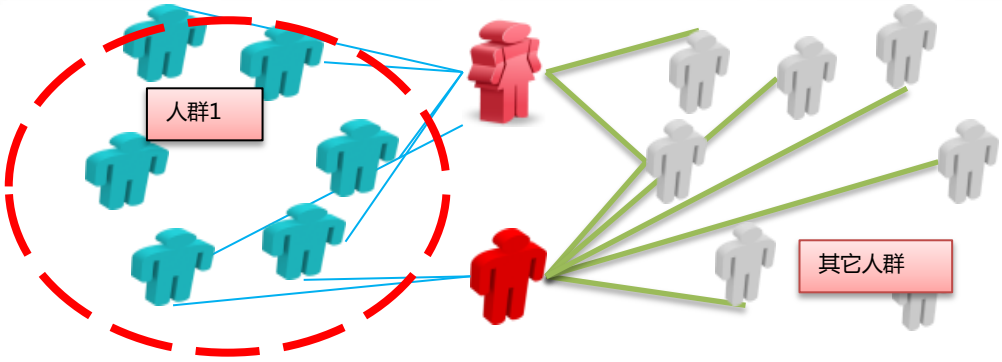
**业务流程：**

原始详单数据加载到预处理平台，通过预处理平台对数据进行清洗，清洗后通过bulk load进行数据加载。数据加载流程每10分钟启动一次，10分钟后营业厅可以对详单内容进行查询。

# 典型的应用场景之三：大数据挖掘系统

**目标：针对海量结构化、非结构化数据的进行深度挖掘。通常需要根据业务需求设计模型、训练集并选择算法（分类、聚类、关联、非结构化）。通常会使用各种分布式数据挖掘工具和算法**






例图：客户分类识别应用

技术要求举例：

- ✓针对海量数据实施全量数据挖掘，规模达到10TB-PB规模。
- ✓处理时间没有严格要求，通常达到几个小时，甚至更长时间
- ✓需要支持各种并行计算模式，如MapReduce、BSP等
- ✓数据挖掘系统需要较好的用户界面，用户通常具备业务知识，但是未必具备开发经验
- ✓系统可以和其他系统混合部署
- ✓数据具有一定的可靠性和可用性要求

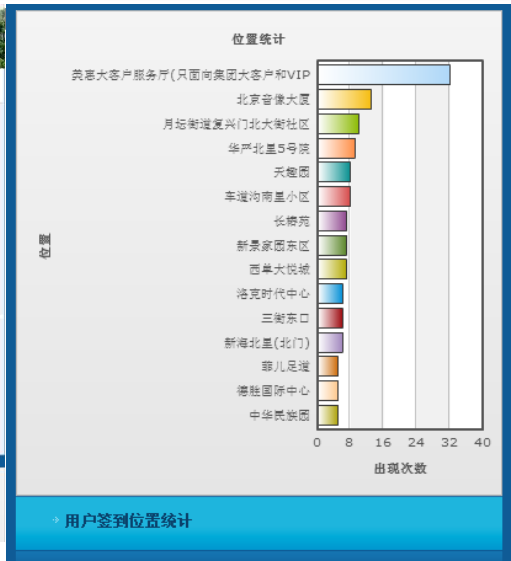
# “大云”应用案例之三： 微博爬取与挖掘

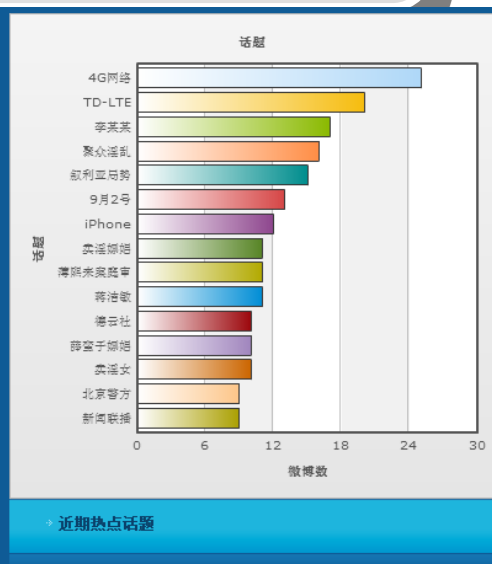
**目标**

针对微博的用户信息、交往关系、微博内容、位置等数据进行实时爬取与分析。可实现市场产品的定向营销和目标客户群体发现。支持用户对自定义条件的目标群体进行检索和整体社交关系分析；支持对目标群体中的各社团子群体发现和特征分析；支持目标群体中用户关注内容和位置聚集信息分析；针对个人用户发现交往行为变化和实时关注点

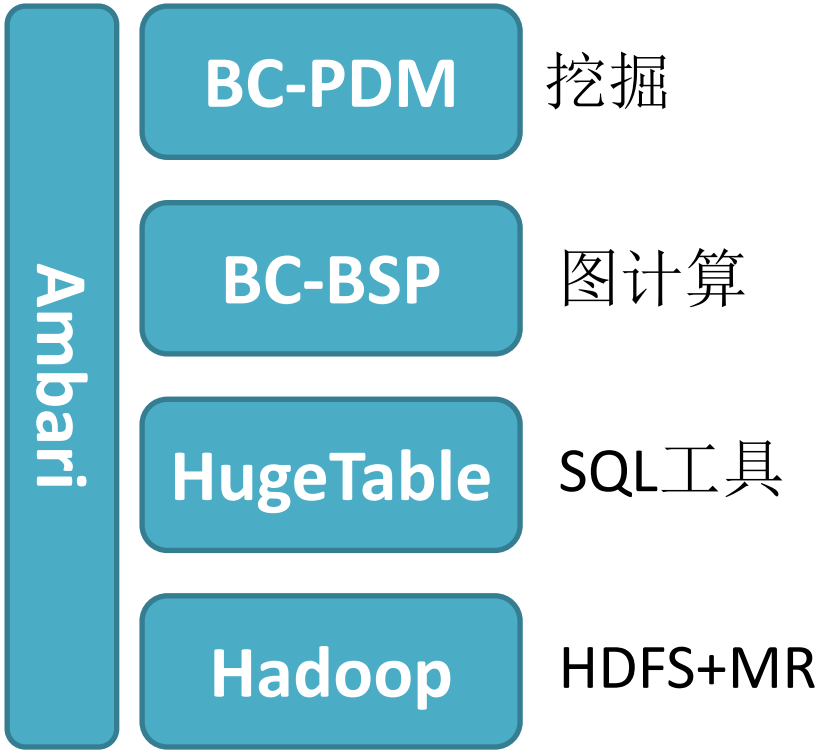
**主要功能**

1. **用户交往关系图生成**：利用粉丝关注关系和转发评论，构建用户交往关系图
2. **用户地点信息统计**：根据签到信息，统计用户常出现地点，发现活动规律
3. **热点话题发现**：从用户近期发布微博中发现用户关心的热点事件
4. **关键词提取**：从用户近期微博中提取出关键词，从中发现用户特征
5. **用户信息挖掘**：统计用户的性别、地域等基本信息
6. **个人分析**：对用户发微博的时段分布、用户近期密友等进行统计分析

# 总结



# 诚聘英才

## 中国移动苏州研发中心 欢迎您！

发邮件至 wangbaohan@chinamobile.com

# Q&A

# THANKS