流式计算在苏宁的发展历程

苏宁云商.大数据平台研发中心

张毅

关于我和我的小伙伴们

团队

大数据平台研发中心

职责：

提供集团各个业务所需要的存储和计算能力。
保证平台的稳定、高效运行。
提高平台易用性。

目标：

打造稳定、易用、高效的平台，提高数据分析效率，实现人人都是数据分析师。

我

大数据攻城狮

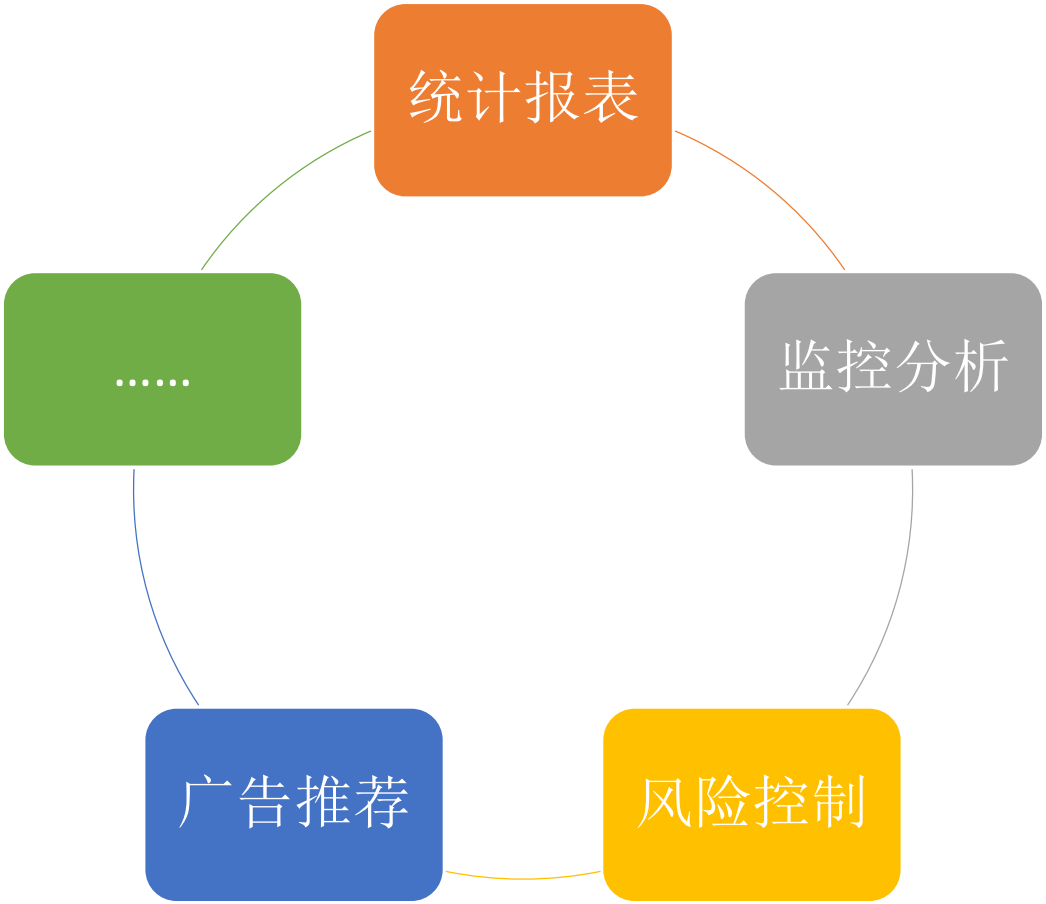
6年工作经验，3年半的流式计算领域相关经验。作为核心人员参与了苏宁流式计算平台的整个发展历程，并主导了Libra(sql on Storm)项目的研发。目前主要关注Storm,Spark Streaming 及sql on Storm等技术

目录

实时计算过程


实时计算过程
数据的产生与收集
数据实时计算
实时查询服务
开源技术
Flume，Kafka, ……
Storm, Spark……
Cassandra Redis,……

流式计算在苏宁的应用场景



苏宁集群现状

Storm

- 开源Storm 0.9.3
- 400+节点
- 23个集群
- 70+topology
- 复杂算法
- 数据清洗、实时推荐、商户实时统计等

Libra

- 自研
- 300+节点
- 8个集群
- 700+sql
- 70+topology
- 简单、sql可描述
- 流量分析、性能监控等

Spark Streaming

- 开源Spark1.5.2
- 500+节点(hadoop共用)
- 30+ streaming任务
- 统计分析。

目录

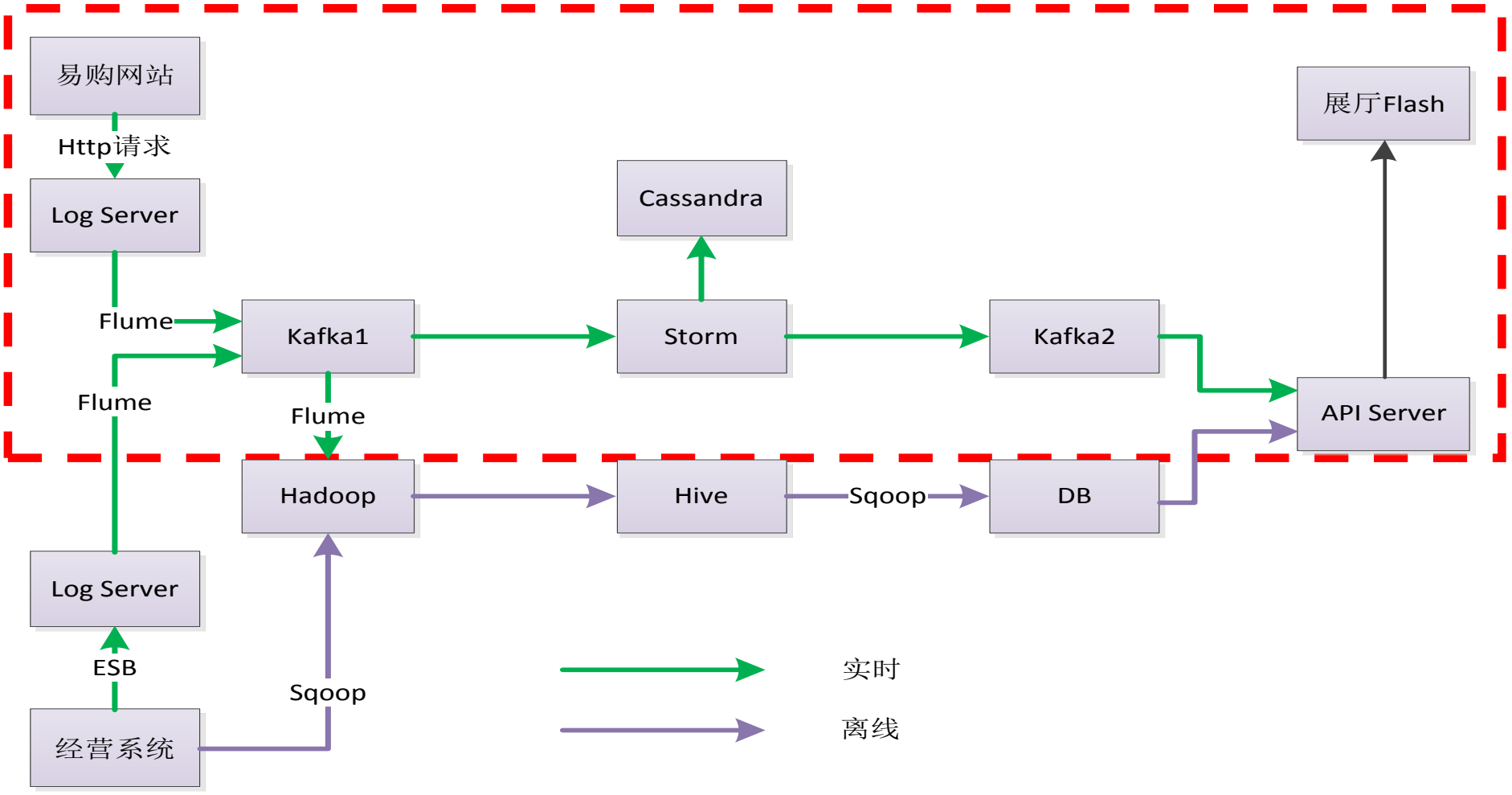
发展历程

2014.06

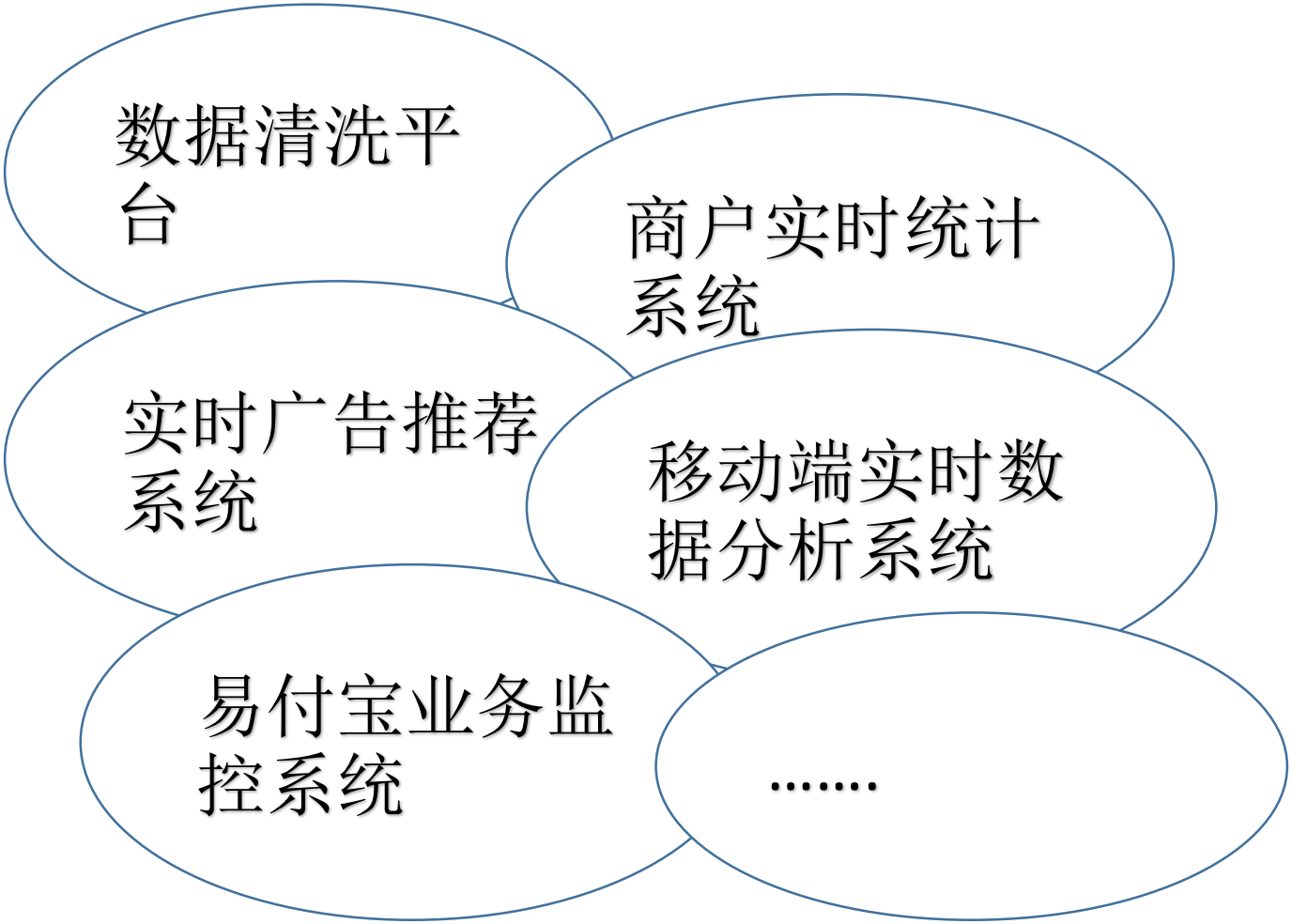
苏宁第一个storm项目

--展厅系统

展厅数据流向


易购网站
Http请求
Log Server
Flume
Flume
Kafka1
Flume
Cassandra
Storm
Kafka2
展厅Flash
API Server
Hadoop
Hive
Sqoop
DB
Log Server
ESB
经营系统
Sqoop
实时
离线

展厅项目之后，……



存在的问题

需求多变、
流程复杂？
……

产品

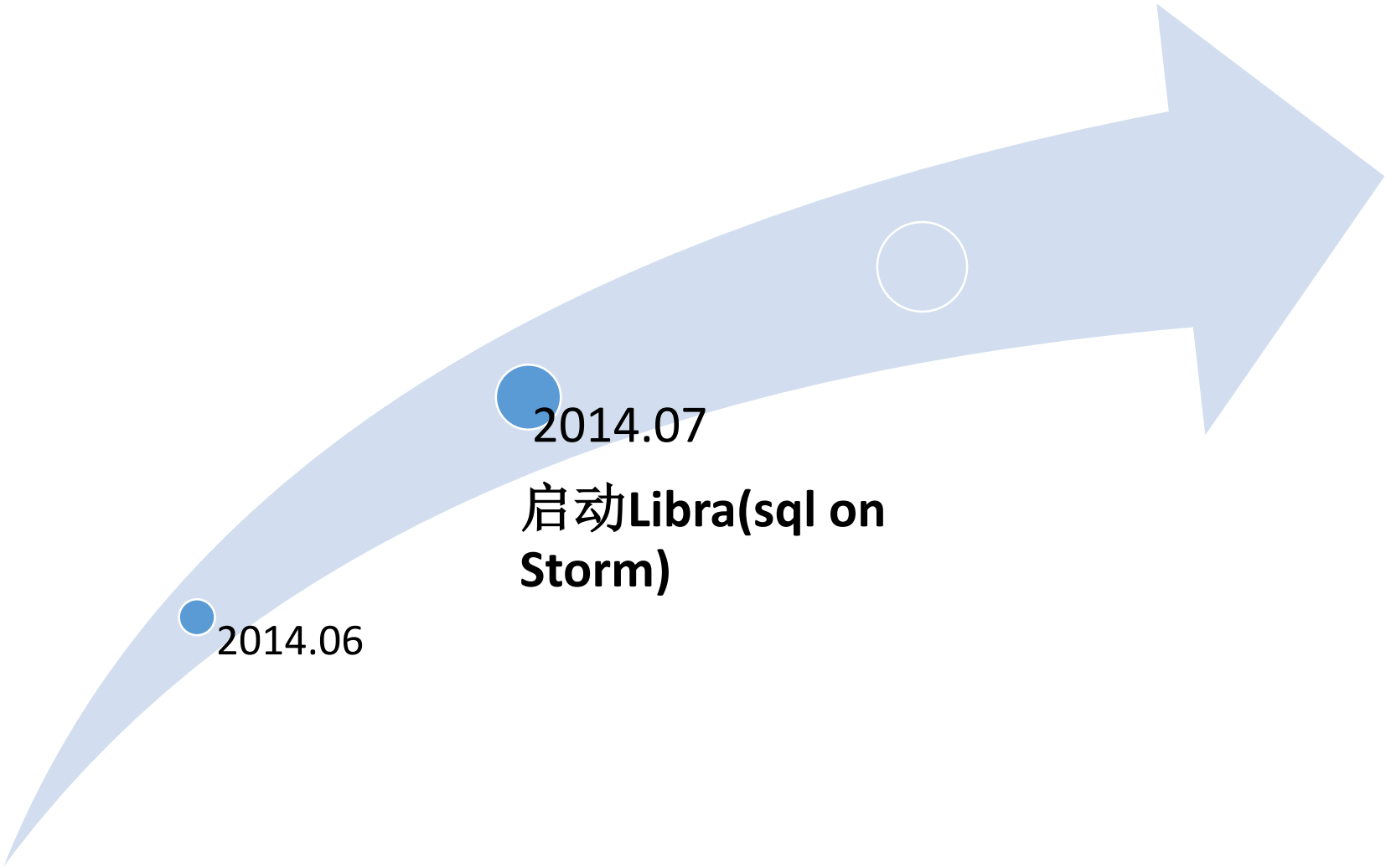
Kafka、
Storm、
Cassandra
……

开发

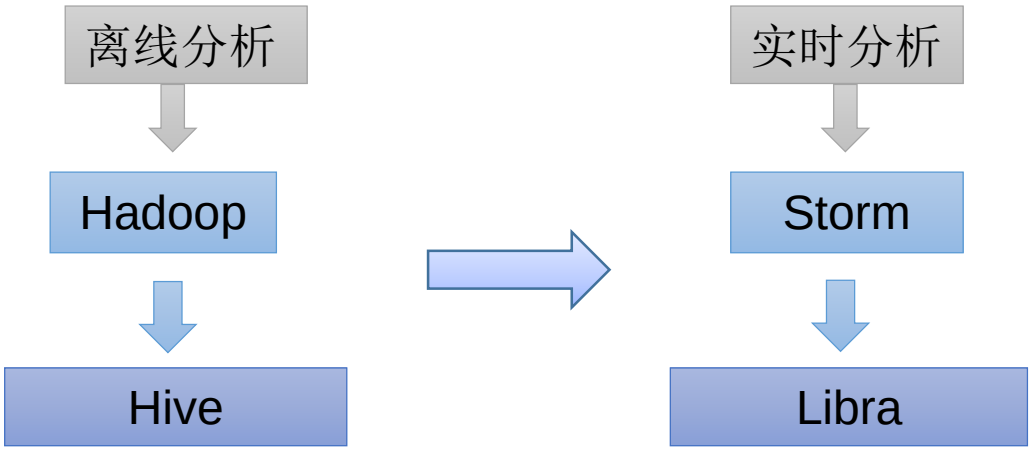
产品迭代快
开发周期长
……

ALL

发展历程



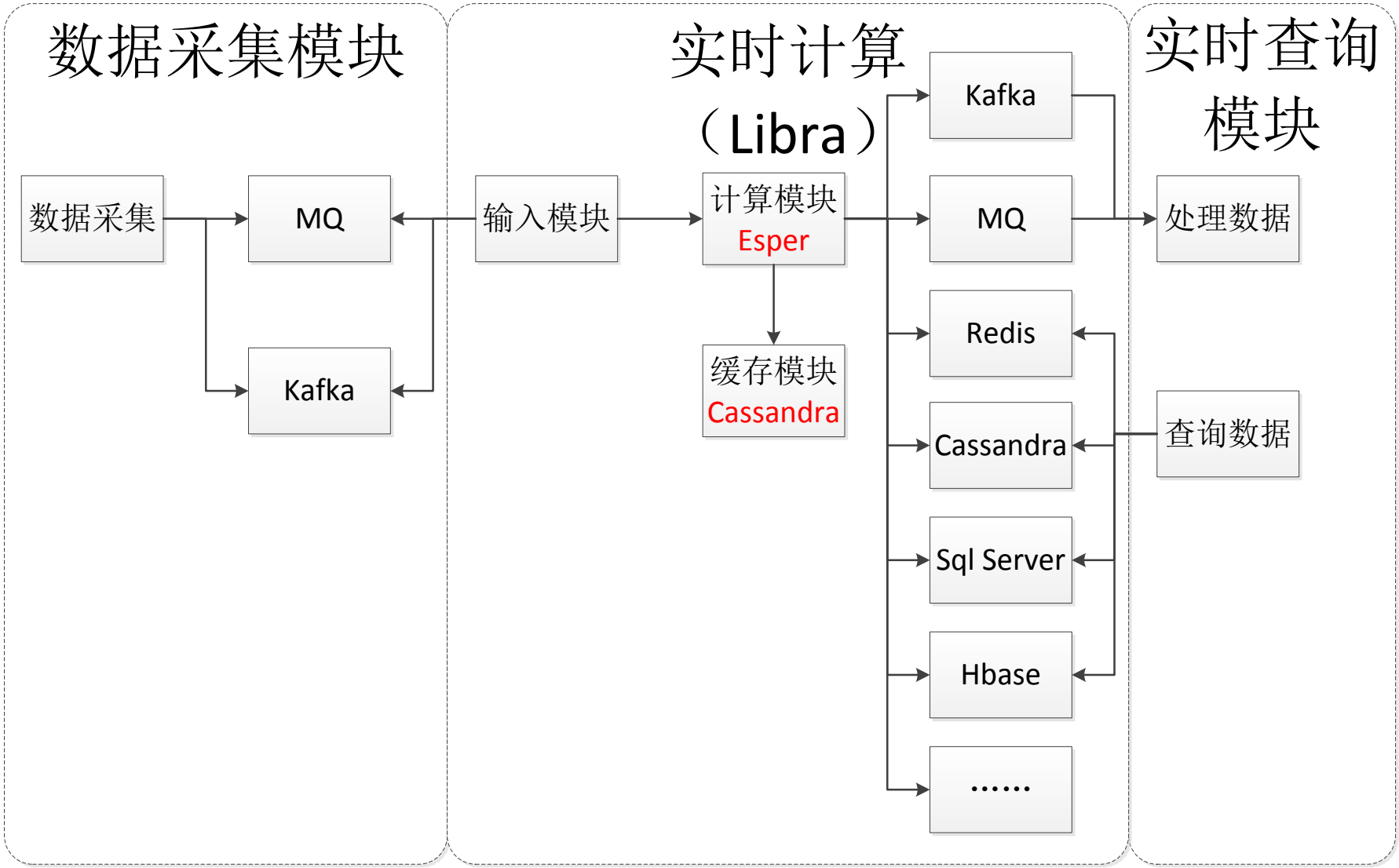
Libra是什么？



通过类SQL语句实现统计分析需求，不必额外开发代码。

以类SQL语句的形式提供实时计算规则，不用专门编码实现

Libra 数据流向


数据采集模块
实时计算（Libra）
实时查询模块
数据采集
MQ
Kafka
输入模块
计算模块
Esper
缓存模块
Cassandra
Kafka
MQ
Redis
Cassandra
Sql Server
Hbase
……
处理数据
查询数据

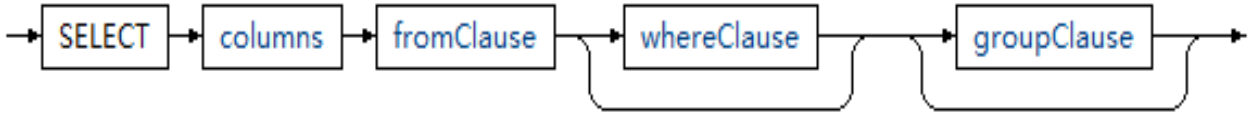
Libra主要功能点

- 支持类SQL语法--ANTLR
- 保证数据不丢--Cassandra
- 动态更改计算规则--Dubbo

SQL语法解析--ANTLR

Select count(distinct userid) **from** login **group by** ip

ANTLR



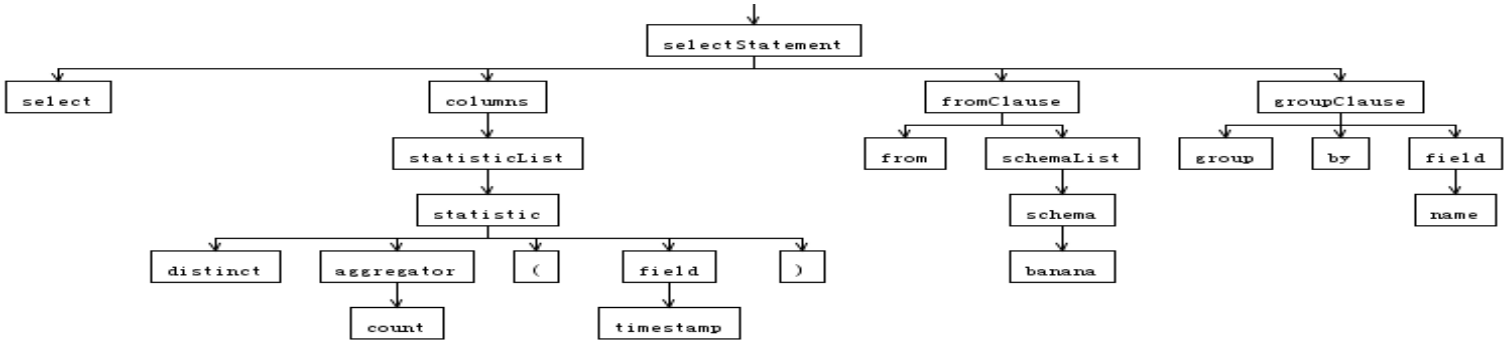
词法分析

select distinct count from group by

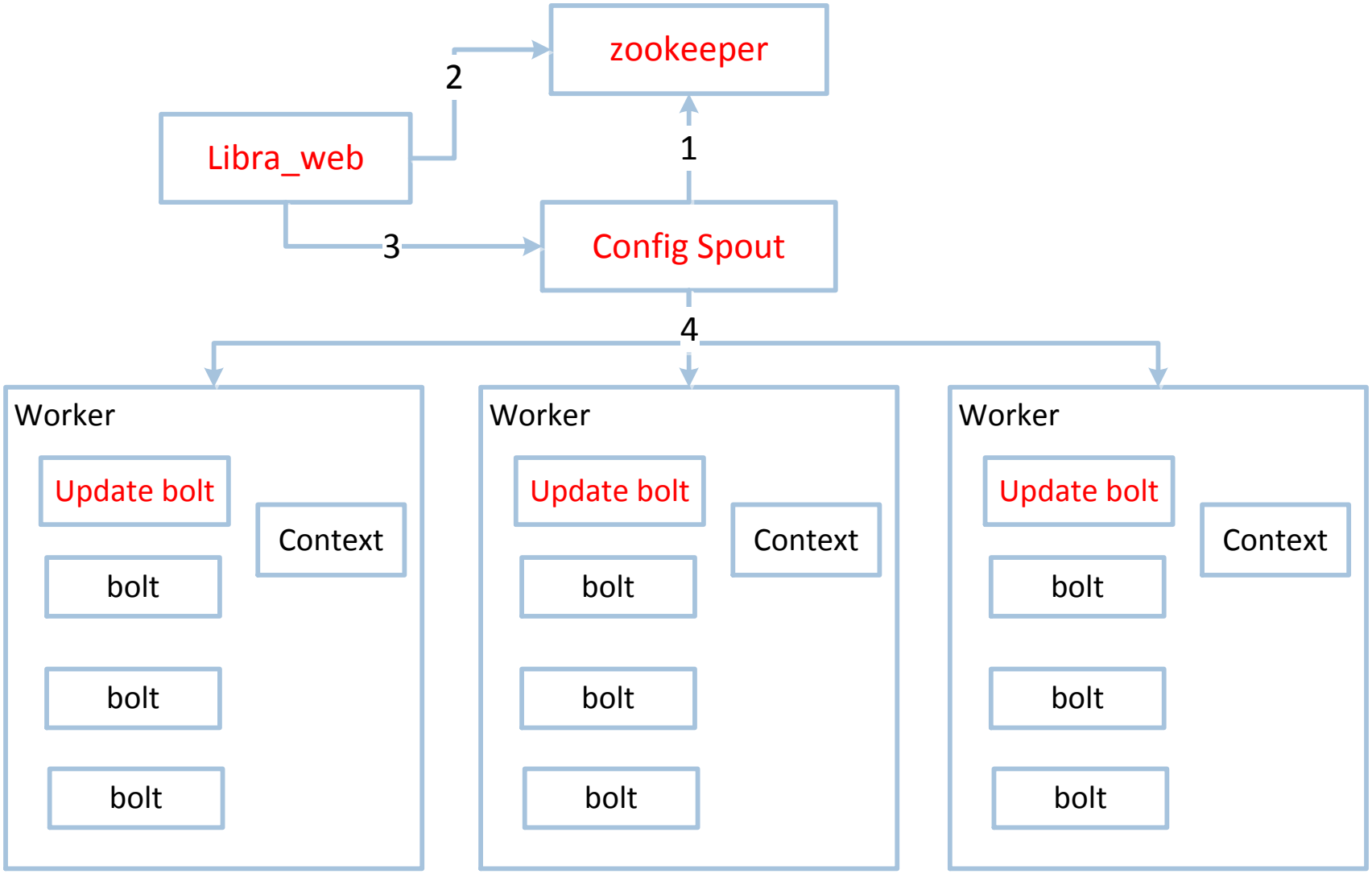
语法分析

按 ip 分组(**group by**), 对userid去重(**distinct**), 输出去重后 userid 数量（ **count** ）

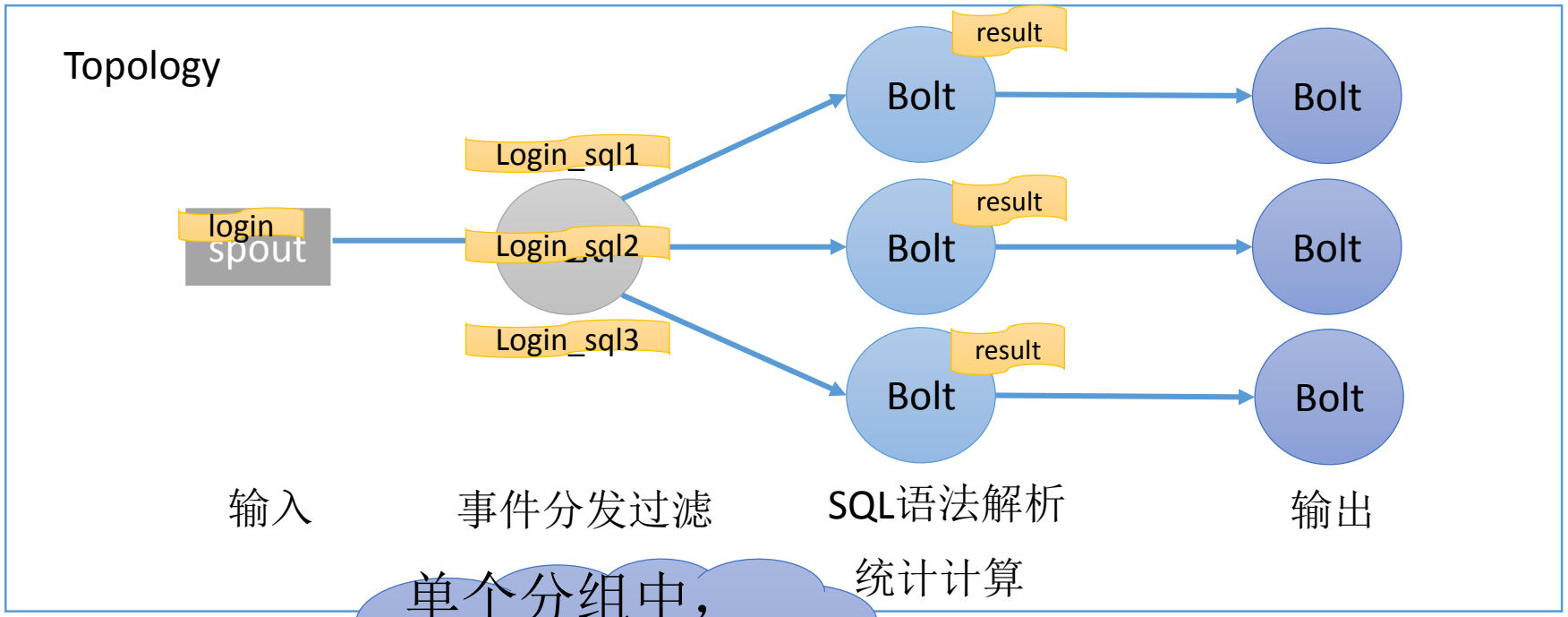
抽象语法树 – 逻辑运算



动态更改计算规则


zookeeper
2
Libra_web
1
3
Config Spout
4
Worker
Update bolt
Context
bolt
bolt
bolt
Worker
Update bolt
Context
bolt
bolt
bolt
Worker
Update bolt
Context
bolt
bolt
bolt

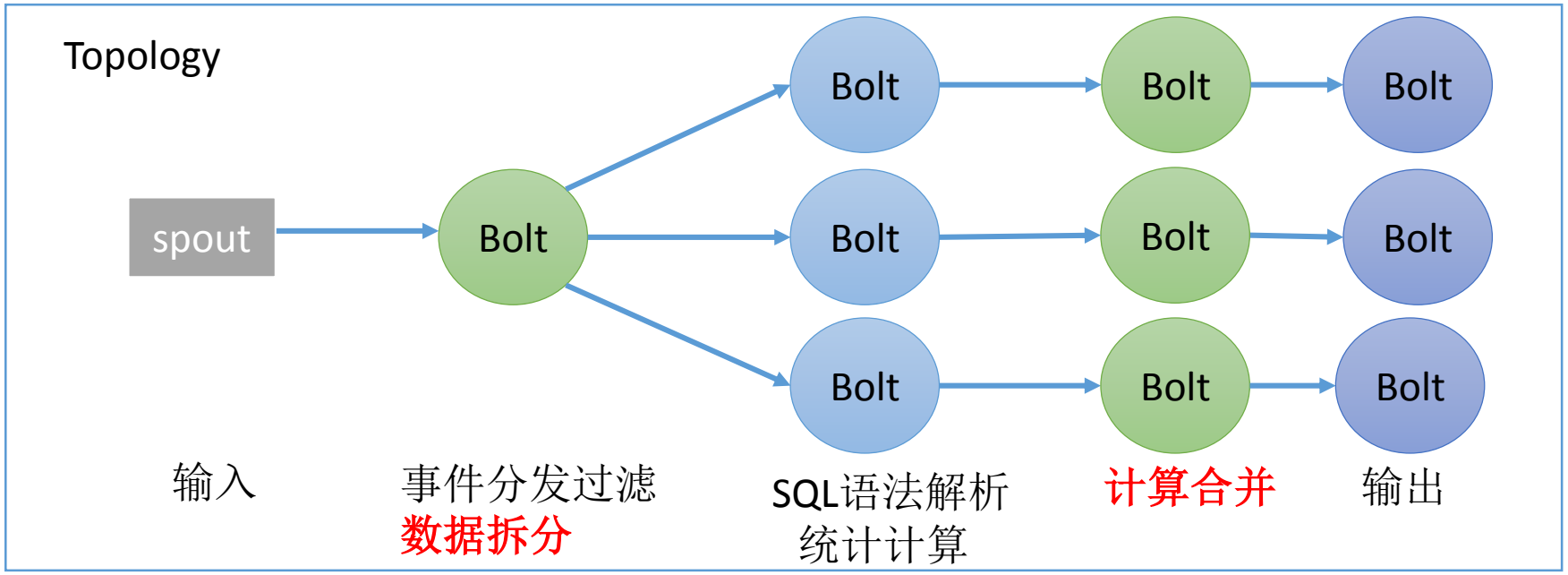
Libra 1.0 DAG


Topology
login
spout
Login_sql1
Login_sql2
Login_sql3
result
Bolt
result
Bolt
result
Bolt
Bolt
Bolt
Bolt
输入
事件分发过滤
SQL语法解析
统计计算
输出


单个分组中，
distinct计算过
于庞大。


OOM

Libra 2.0 DAG



Select count(distinct userid) **from** login **group by** ip

事件名	sql
Login_sql1	sql1
Login_sql2	sql2
Login_sql3	sql3

分发：ID = event+sql+group by+ distinct

合并：ID = event+sql+group by

集群数据均匀分布

Storm SQL: STORM-1040

采用ANTLR语法解析

- 支持根据关键字排序
- 支持分布式部署
- 支持将Kafka作为外部表
- 支持输出到Kafka
- ……

发展历程

2014.06

2014.08

2014.11

Libra集群
30+节点数。

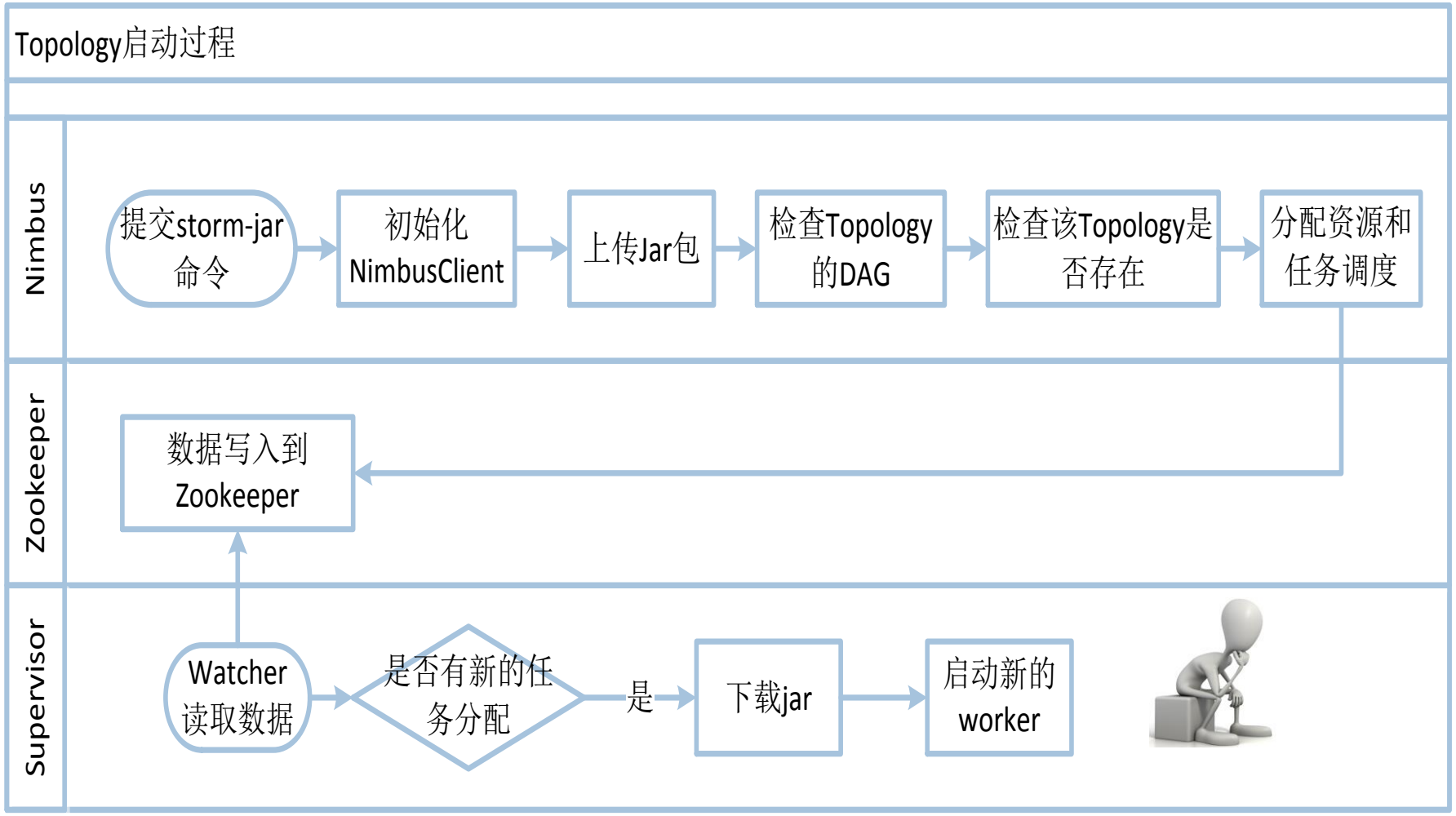
最大的
Topology用
60+worker

2015.4

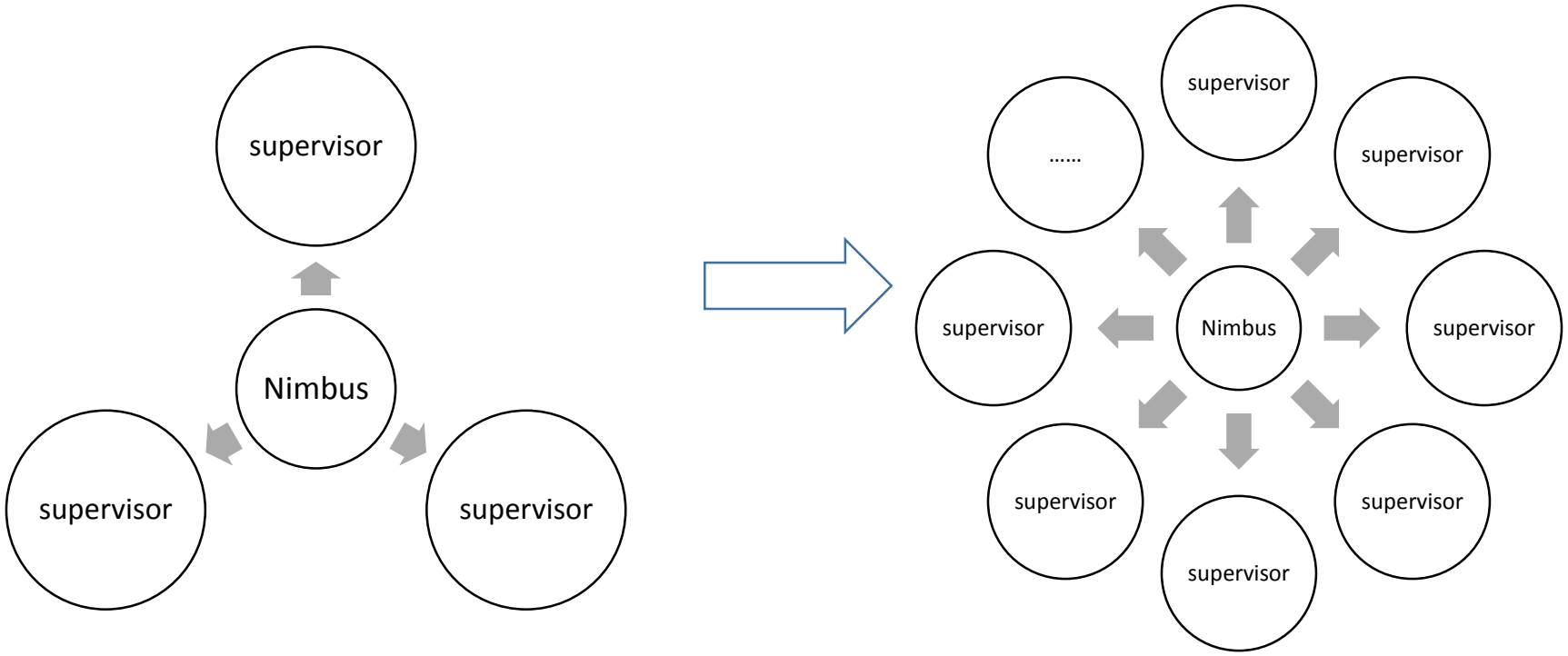
Libra集群
80+节点数。

最大的
Topology用
150+worker

Topology启动过程



Jar包下载的规模效应



jar 包大小80MB+，Nimbus带宽200mbps：

30个supervisor， 下载jar时间 = (30*80*8)/200 = 96s

80个supervisor， 下载jar时间 = (80*80*8)/200 =**256s**

Jar包分离方案

- 将Topology代码包和依赖包分离
- 使用Rsync，提前将依赖包分发到各个节点
- 修改Nimbus源码，在Topology提交时，加载依赖包
- 修改Supervisor源码，在启动Worker时，加载依赖包

优化效果：

代码包6MB

80个Supervisor， 下载Jar时间 = (80*6*8)/200 = 19.2s

Dist Cache: STORM-876

- contributed by Yahoo
- Jar包存储在Dist Cache中
- 支持HDFS作为后端存储，提升副本数分散压力

发展历程

2014.8

2014.11

2015.4

节点数
150+

2015.6

节点数
250+

监控困难

Storm监控

比较挫的方法：

通过crontab + shell脚本，定时的抓取Storm-UI页面信息，监控集群和Topology的健康状况。

问题：

1、新增监控比较麻烦。
2、无法快速定位问题。
3、无历史数据展示。

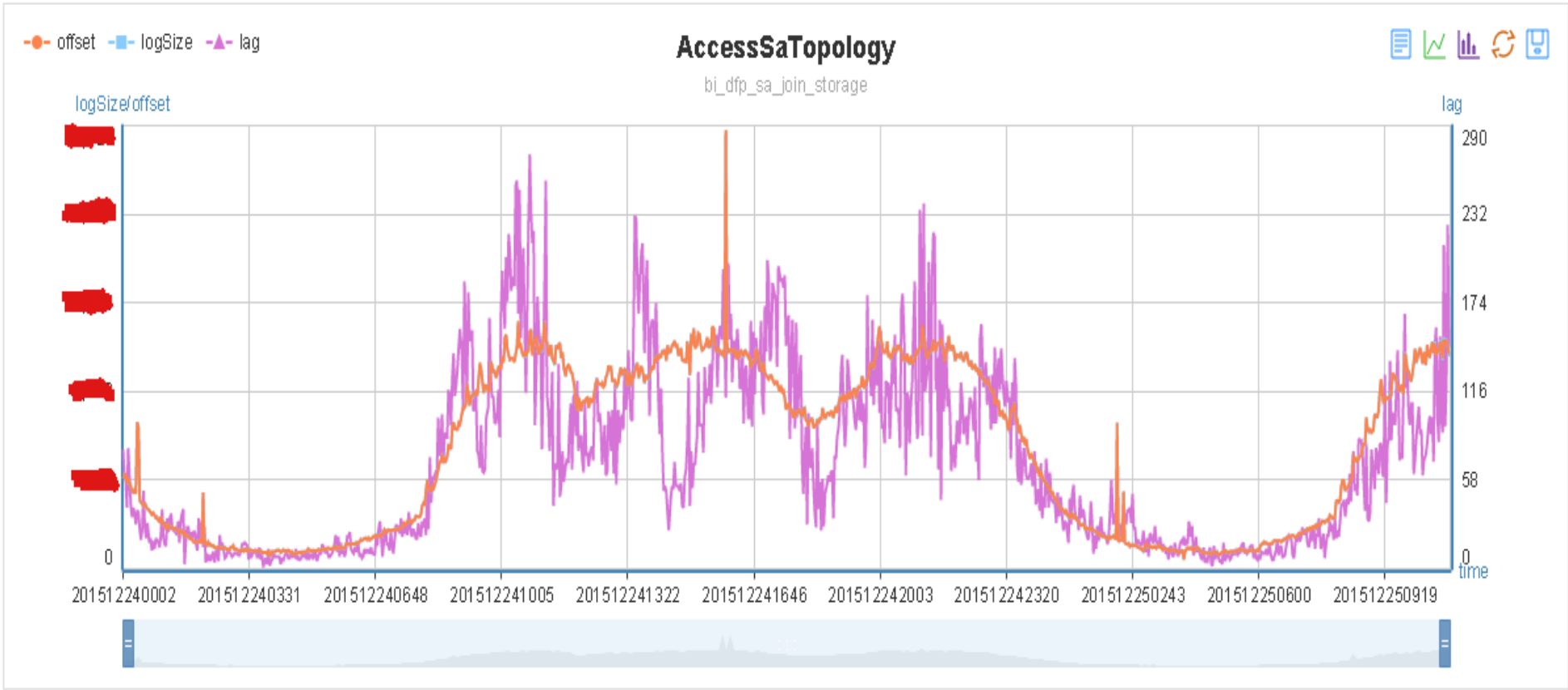
Storm-monitor

Storm-monitor

Kafka

topologyName: AccessSaTopology ▾ topic: 全部 ▾ 查询类型: 一分钟 ▾

开始时间: 2015-12-24 00:01:00 结束时间: 2015-12-25 11:01:00 查询 刷新



发展历程

2014.11

2015.4

2015.6

最大Storm
集群节点
数100+

2015.8

最大Storm
集群节点
数200+

Storm Zookeeper IO高

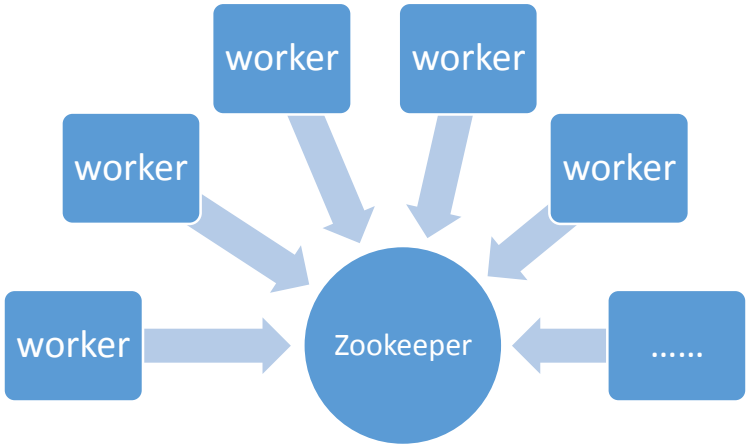
- 网络IO

  400worker

  每次心跳23.4KB/worker

  心跳间隔1s

  ZK Leader 带宽：~9.14MB。



方案：

worker.heartbeat.frequency.secs 调整为3

Storm Zookeeper IO高

- 磁盘IO

  每次心跳20KB-50KB/worker； 440个worker； 3313个executor

  ZK Leader IOUtil: **50%+**

可选方案：

●增强磁盘性能，比如换为SSD盘。

● 继续降低worke心跳频率。

●将zookeeper参数forceSync设置为no。

结果：

ZK Leader的IOUtil降低至10%。

发展历程

2014.11

2015.4

2015.6

250+节点

2015.8

350+节点

2015.11

500+节点

应用管理麻烦
资源隔离困难
组件繁杂多样
…

Storm 服务化

Storm 服务化--CPU绑定

CPU超分1:3

- 大促期间，业务之间相互影响

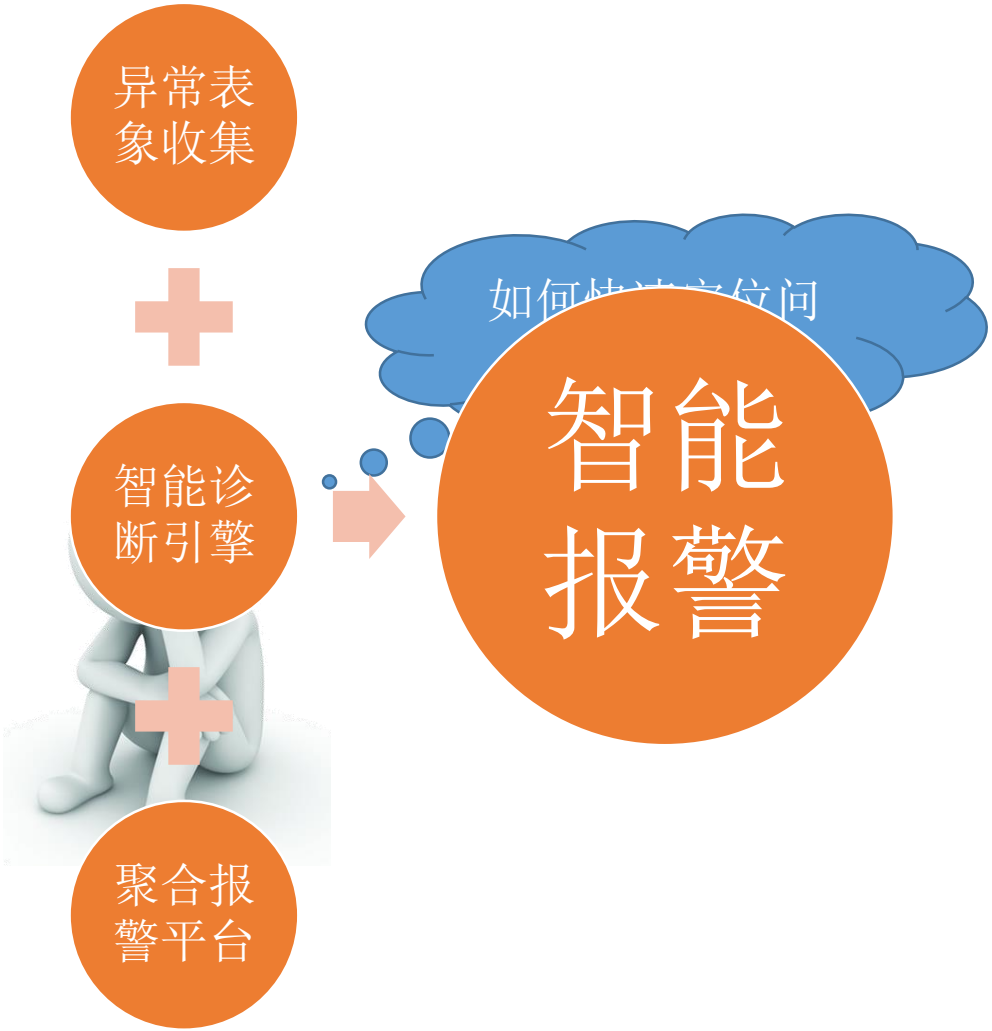
CPU1:1未绑定

- CPU Steal Time 很高，达到40%

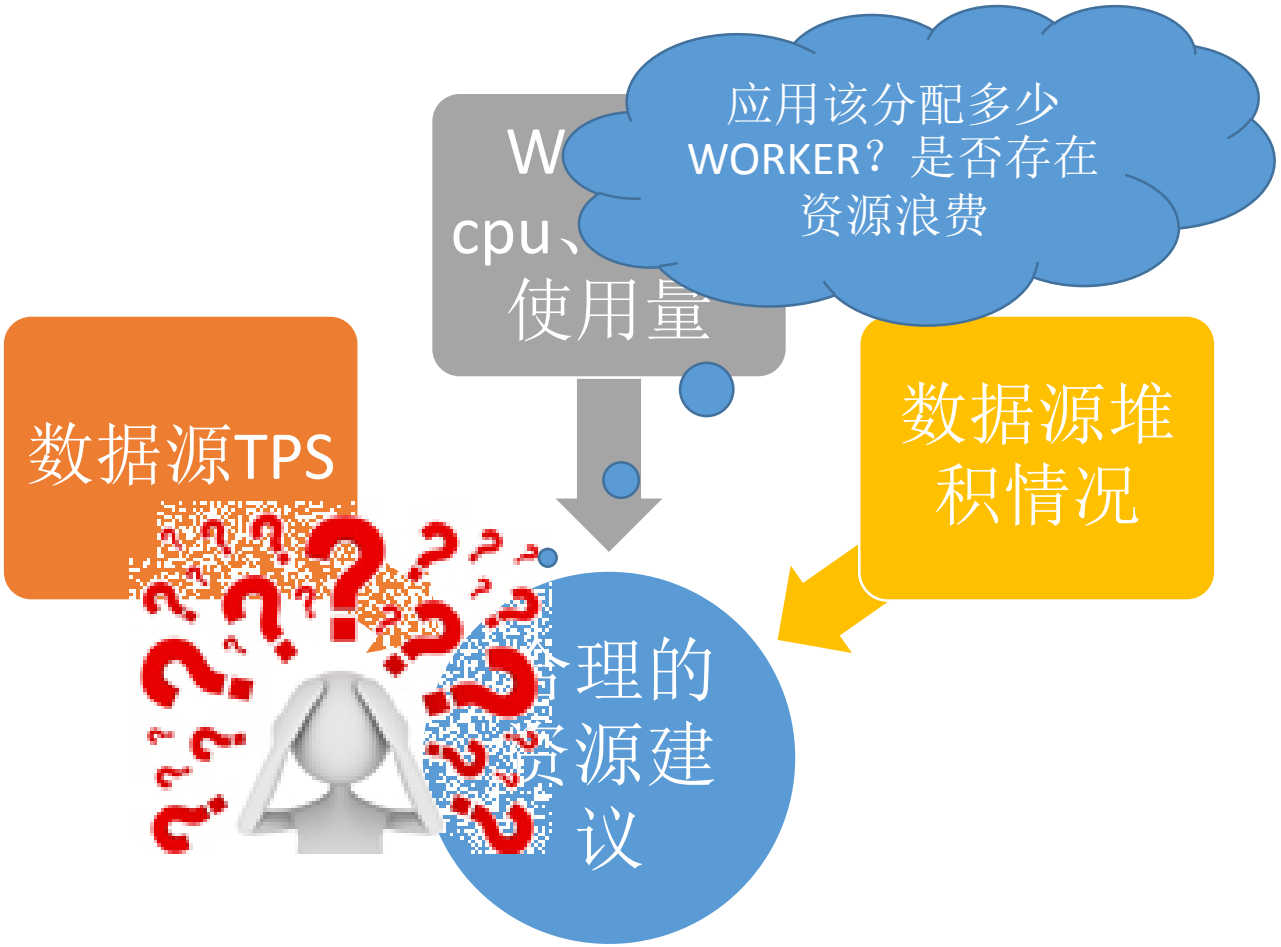
CPU绑定

- CPU Steal Time 非常低，接近0%

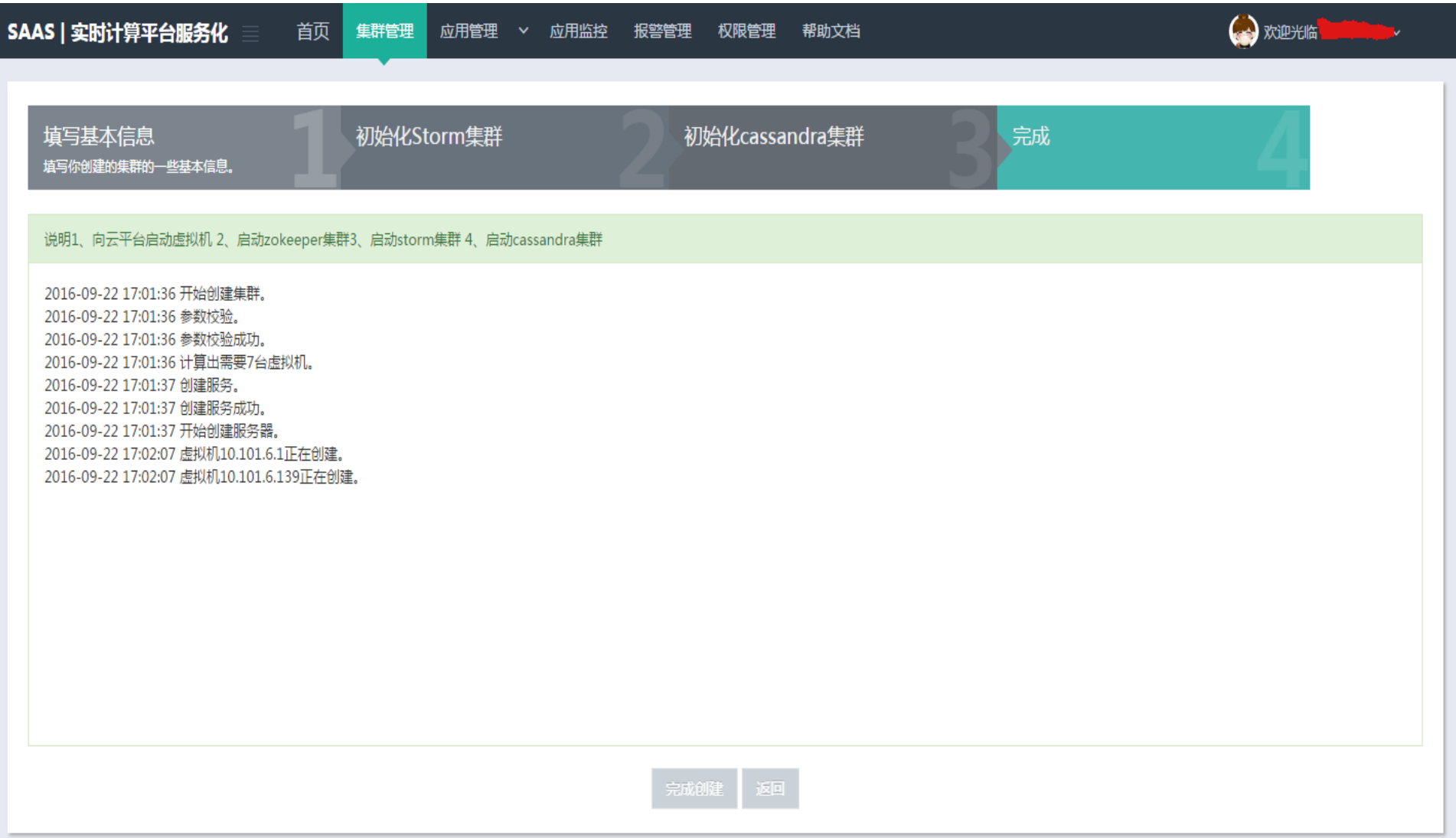
Storm 服务化



Storm 服务化--扩缩容建议



Storm 服务化


SAAS | 实时计算平台服务化
首页
集群管理
应用管理
应用监控
报警管理
权限管理
帮助文档
欢迎光临
填写基本信息
填写你创建的集群的一些基本信息。
1
初始化Storm集群
2
初始化cassandra集群
3
完成
4
说明1、向云平台启动虚拟机 2、启动zokeeper集群3、启动storm集群 4、启动cassandra集群
2016-09-22 17:01:36 开始创建集群。
2016-09-22 17:01:36 参数校验。
2016-09-22 17:01:36 参数校验成功。
2016-09-22 17:01:36 计算出需要7台虚拟机。
2016-09-22 17:01:37 创建服务。
2016-09-22 17:01:37 创建服务成功。
2016-09-22 17:01:37 开始创建服务器。
2016-09-22 17:02:07 虚拟机10.101.6.1正在创建。
2016-09-22 17:02:07 虚拟机10.101.6.139正在创建。
完成创建
返回
2016 大数据平台组 © 苏宁

发展历程

2014.7

libra项目

2015.6

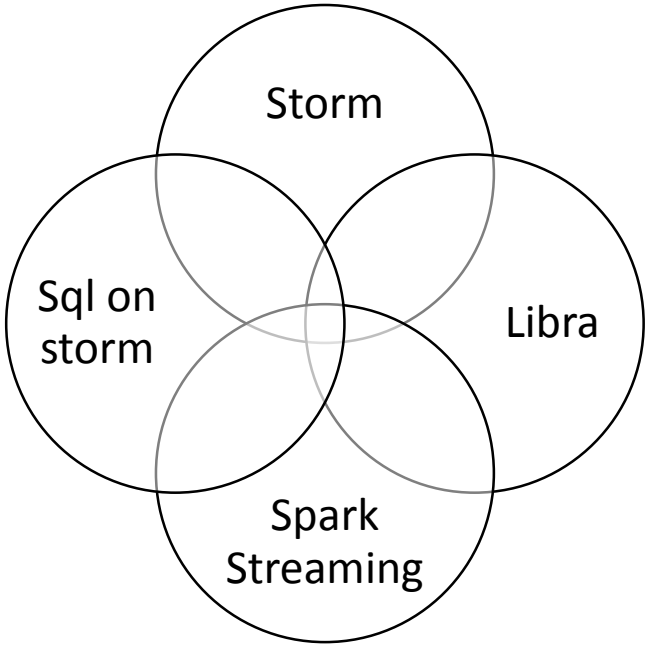
监控平台

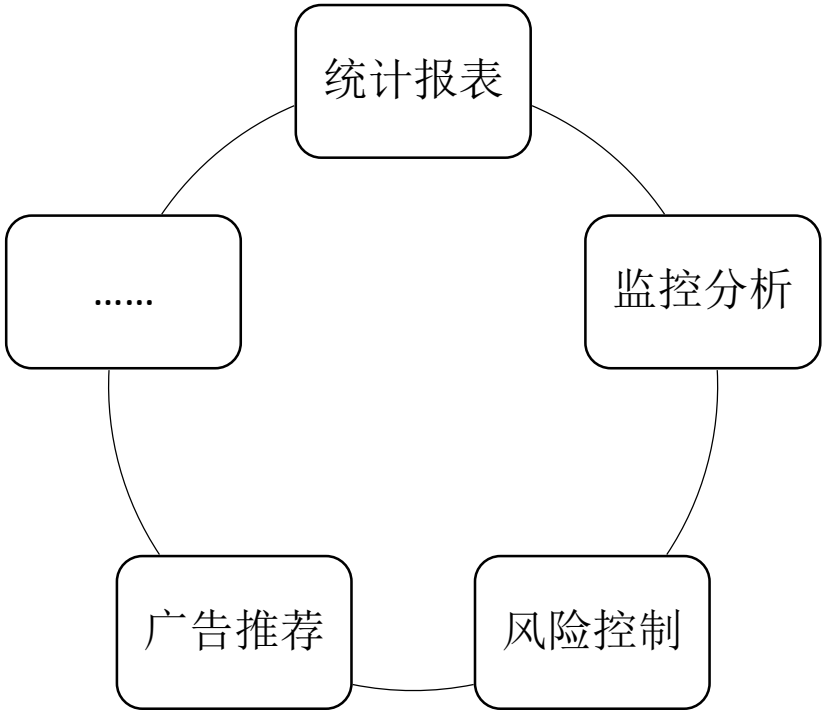
2015.12

Storm服务平台

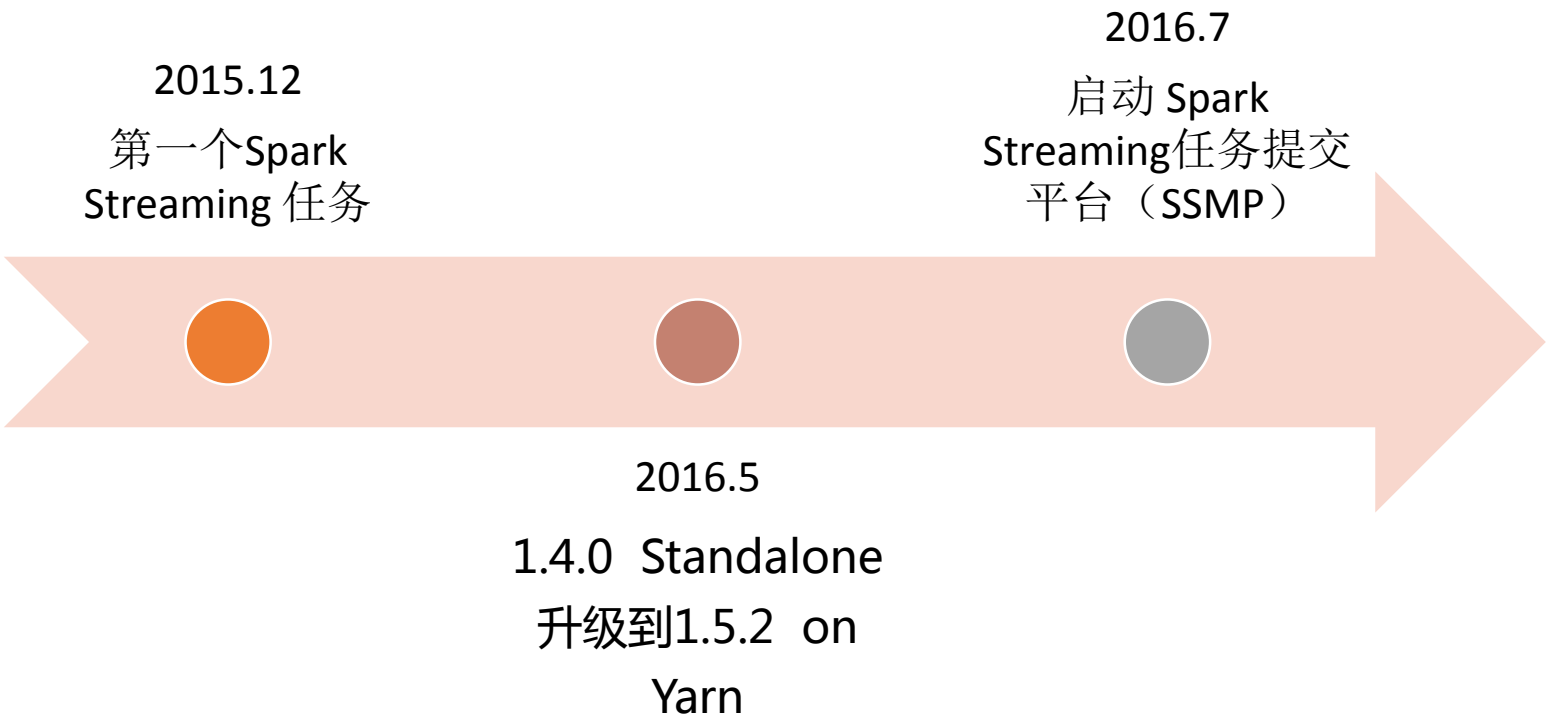
Storm是否适用于所有的场景？

Storm & Libra & Spark Streaming?

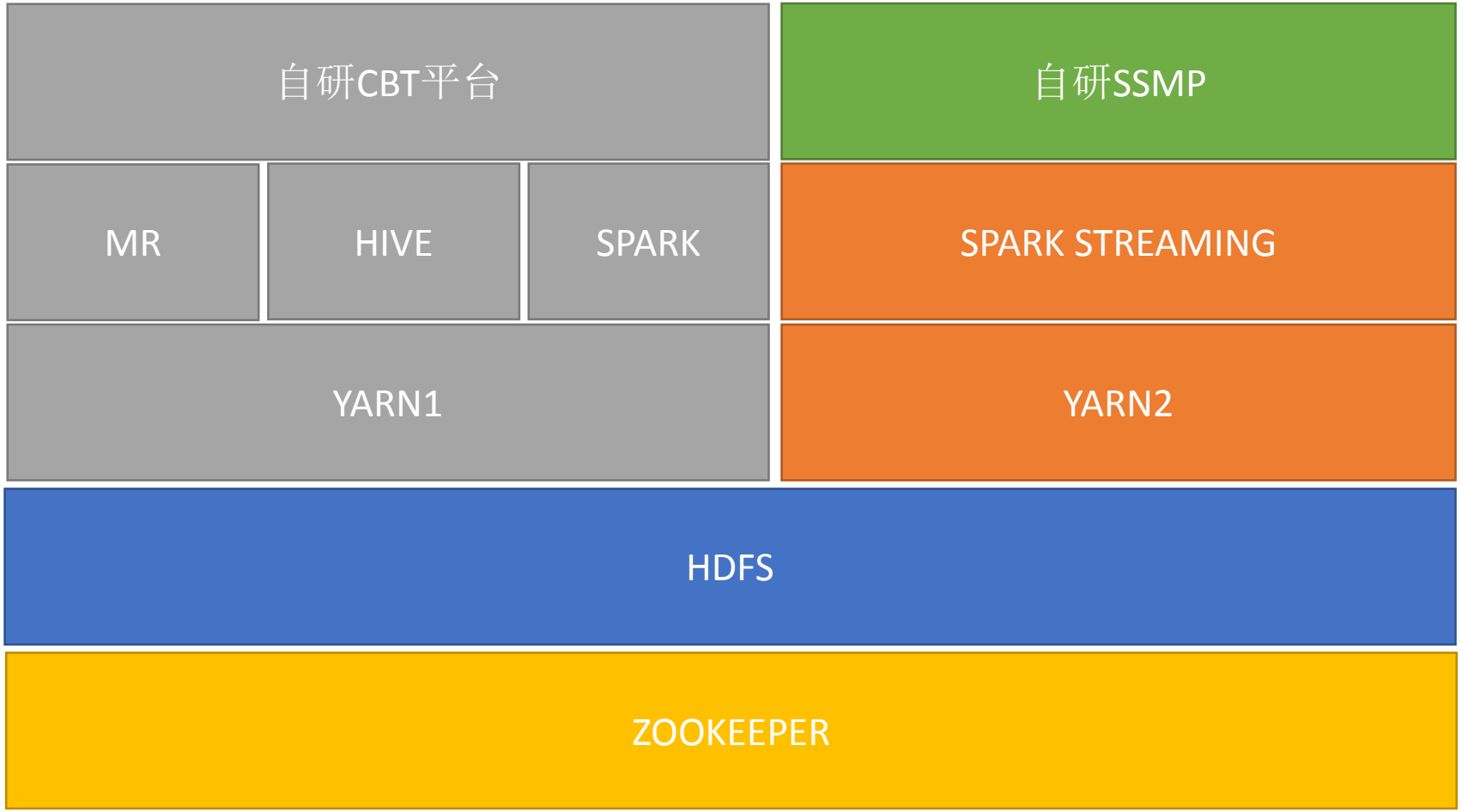




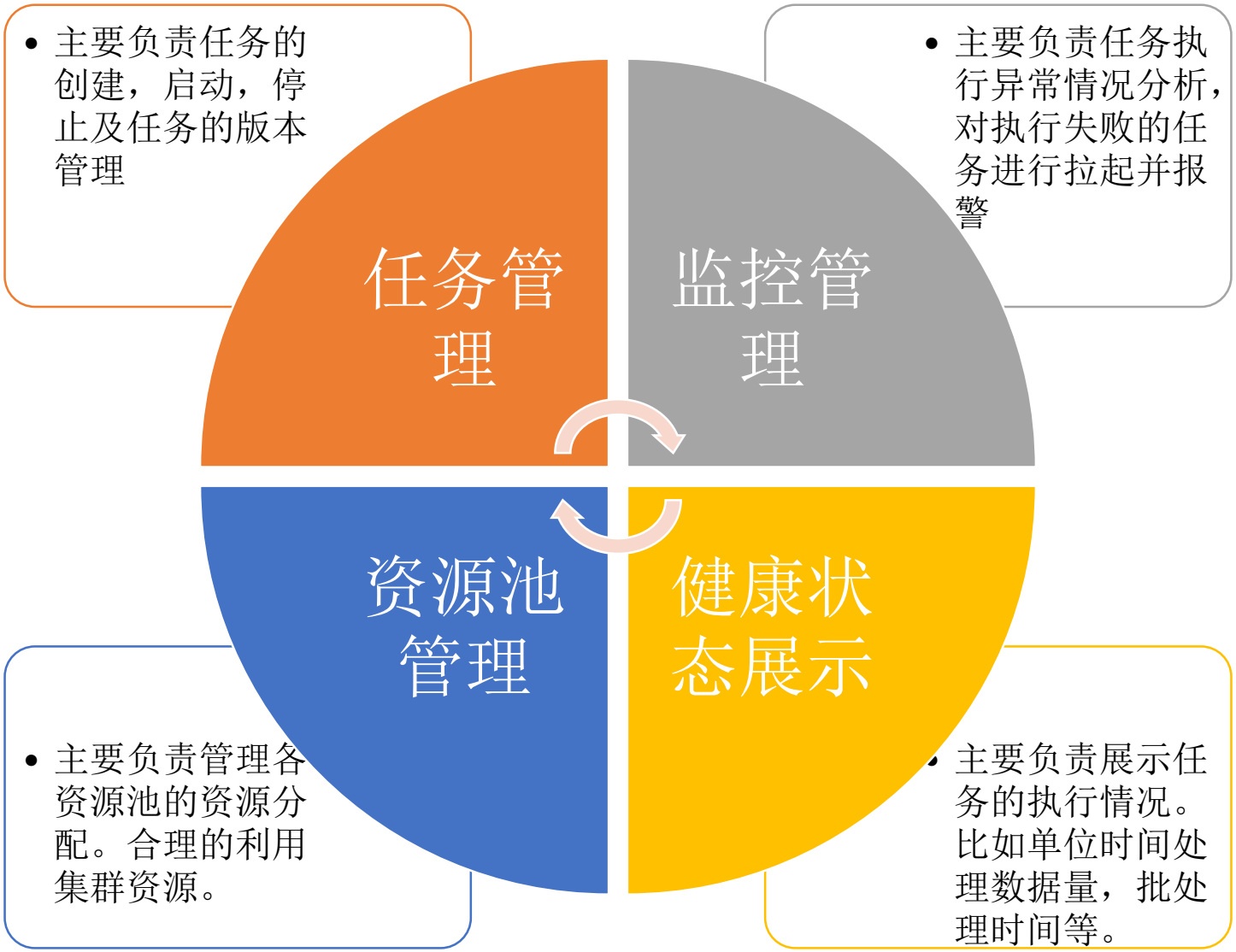
发展历程—Spark Streaming



Spark Streaming 部署结构

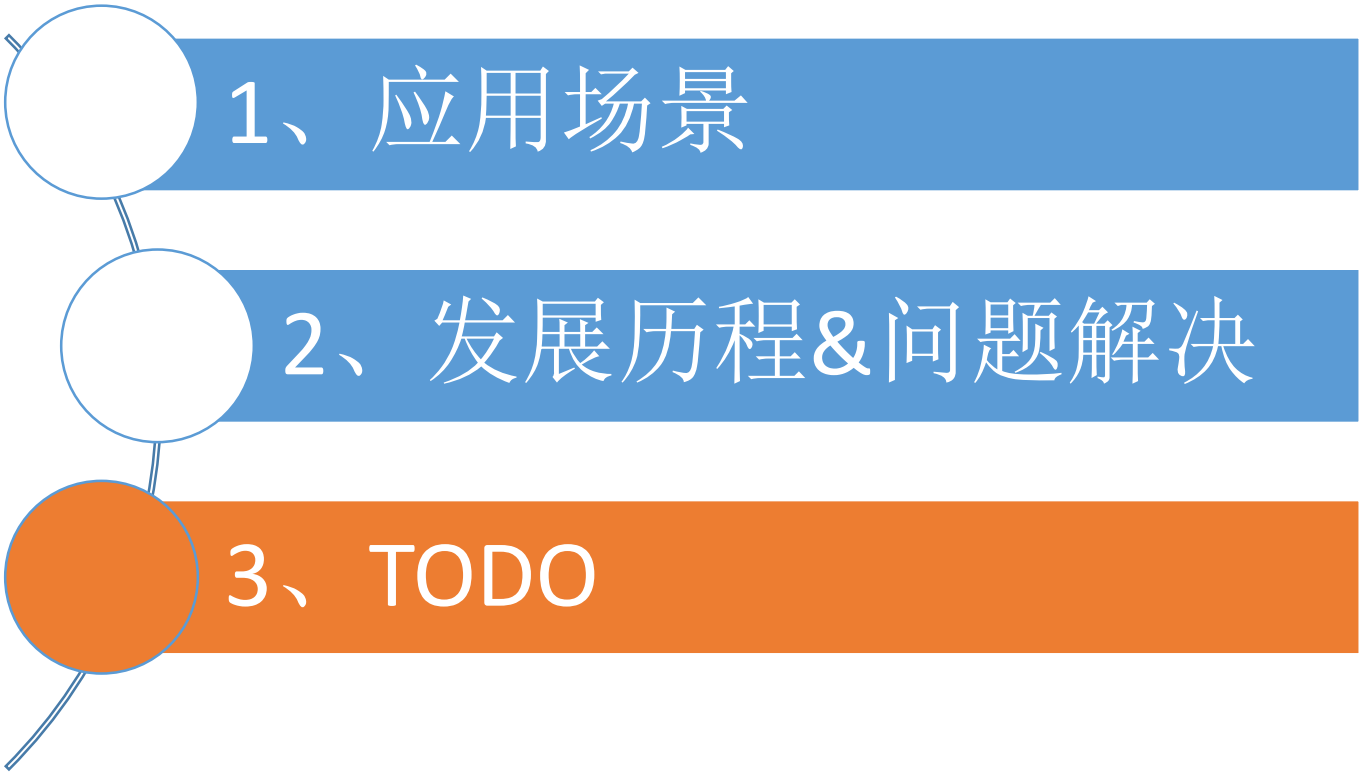


Spark Streaming 平台



Spark Streaming 平台

目录



TO DO

- Storm1.0 落地
- Spark 2.0 落地
- Storm自动扩缩容
- Storm exactly-once
- 统一的sql引擎
- 参与社区，回馈社区

The End

Thanks

如果你对苏宁大数据技术有兴趣，欢迎来聊聊~~~~
你懂的O(∩_∩)O~