CNUTCon 2017 全球运维技术大会

苏宁大数据平台运维实践

王志强

苏宁云商.大数据平台技术总监

主办方 Geekbang 极客邦科技 InfoQ

Geekbang.
极客邦科技

极客搜索

全站干货，一键触达，只为技术

s.geekbang.org

扫描二维码立即体验

有没有一种搜索方式，能整合 InfoQ 中文站、极客邦科技旗下12大微信公众号矩阵的全部资源?
极客搜索，这款针对极客邦科技全站内容资源的轻量级搜索引擎，做到了!

扫描上方二维码，极客搜索!

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

这里只有
技术领导者

EGO会员第二季招募季正式开启

报名时间：9月1日-9月15日

扫描添加E小欧，

邀您进入EGO会员预报名群

立即报名

E小欧

你认为以下哪些关键字对运维最重要？

智能

易用

稳定

高效

性能

流程

MTTR

自动化

安全

规范

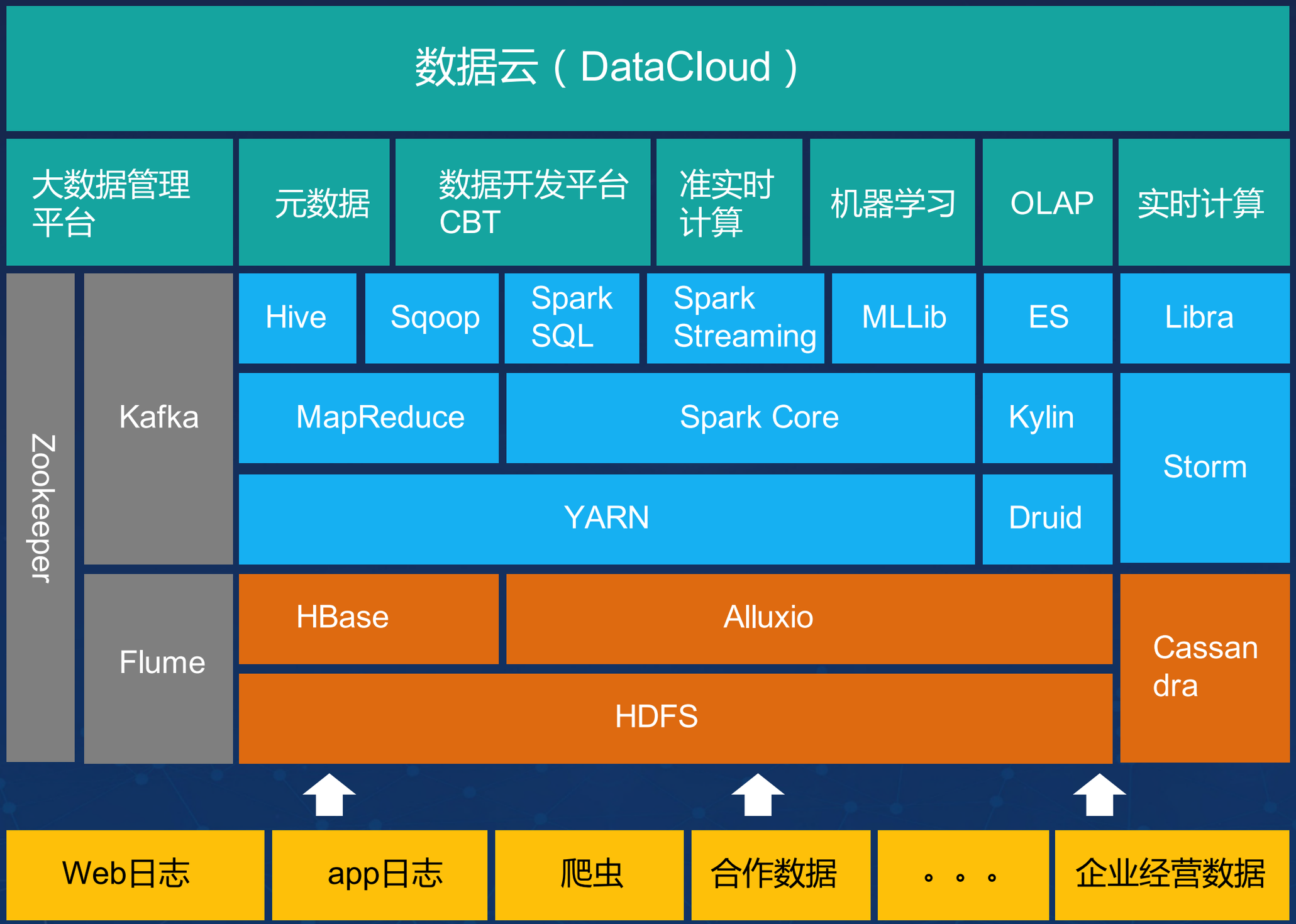
审计

SLA

故障预测

TABLE OF **CONTENTS**

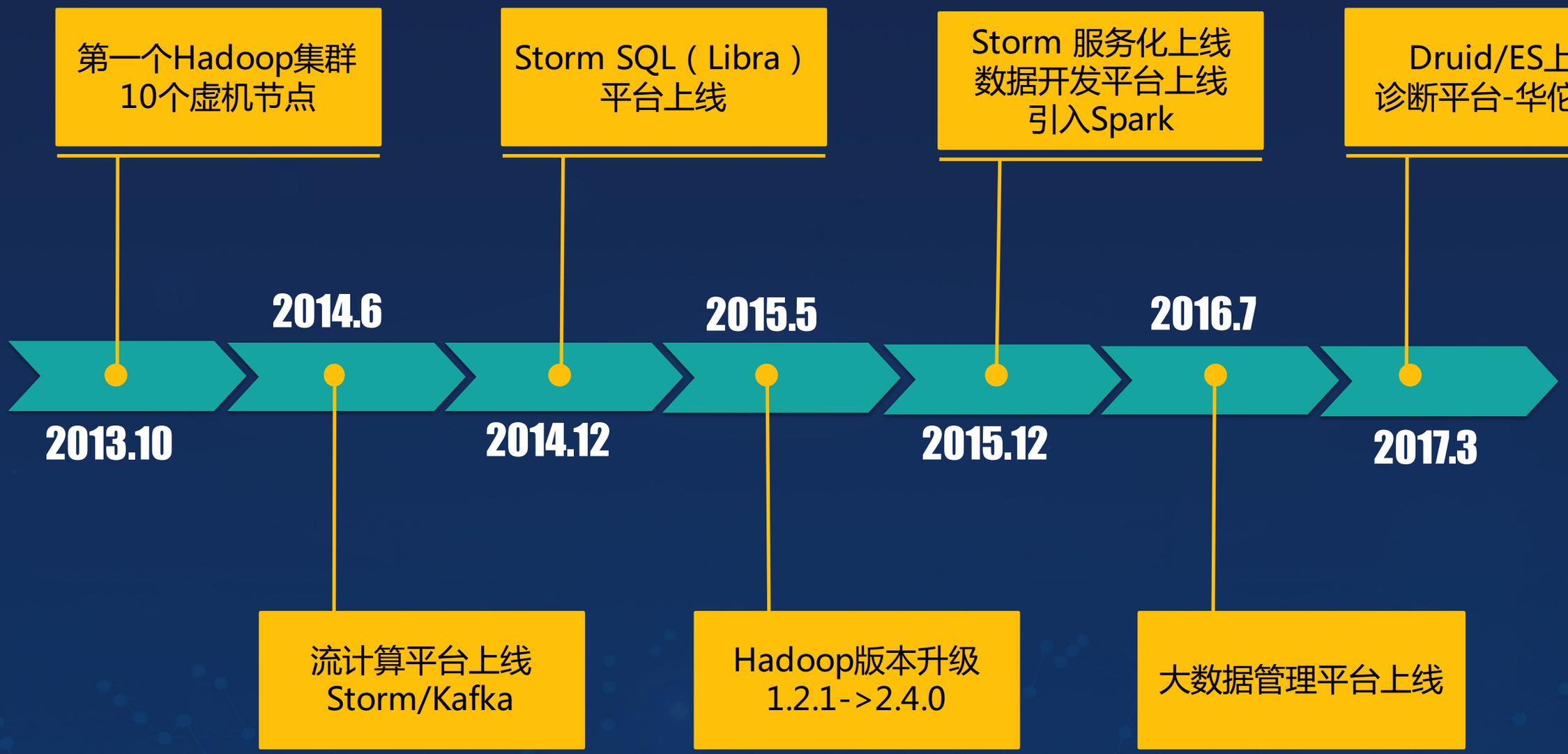
苏宁大数据平台软件栈介绍



典型数据流向


WebServer
Http请求
LogServer
Flume
Kafka
计算
Storm/Spark Streaming
输出
Kafka
poll
业务系统
流式
离线
Flume/DataFlow
数据交换
HDFS/Hive
计算
MR/Spark
输出
HDFS/Hive
数据交换
FTP
FTP
Push
数据交换
数据交换
业务系统
DB
DB

平台发展历程


第一个Hadoop集群
10个虚机节点
2013.10
流计算平台上线
Storm/Kafka
2014.6
2014.12
Storm SQL（Libra）
平台上线
Hadoop版本升级
1.2.1->2.4.0
2015.5
2015.12
Storm 服务化上线
数据开发平台上线
引入Spark
大数据管理平台上线
2016.7
Druid/ES上线
诊断平台-华佗上线
2017.3

平台规模

离线计算

1500节点
5万任务流，30万任务/天
处理500TB/天
新增30TB/天
150账户，500+开发人员
Hive:MR:Spark=90:3:7

KV存储(HBase)

160节点，6个集群
50个NameSpace
600+张表
高峰期200万QPS

流式计算

Storm：
1400虚拟机，41个集群
单集群最大174节点
1500流计算任务

Spark Streaming：
93节点，80个任务

OLAP引擎

Druid：
39节点
30个DataSource

Elastic Search：
15节点
4个Index

TABLE OF **CONTENTS**

大数据平台运维的痛点

- 痛点1. 部署及运维复杂
- 痛点2. 无资源使用视图
- 痛点3. 任务相互影响，资源隔离性差
- 痛点4. 排查问题耗时长，应用优化门槛高

痛点1. 部署及运维复杂

平台管理员

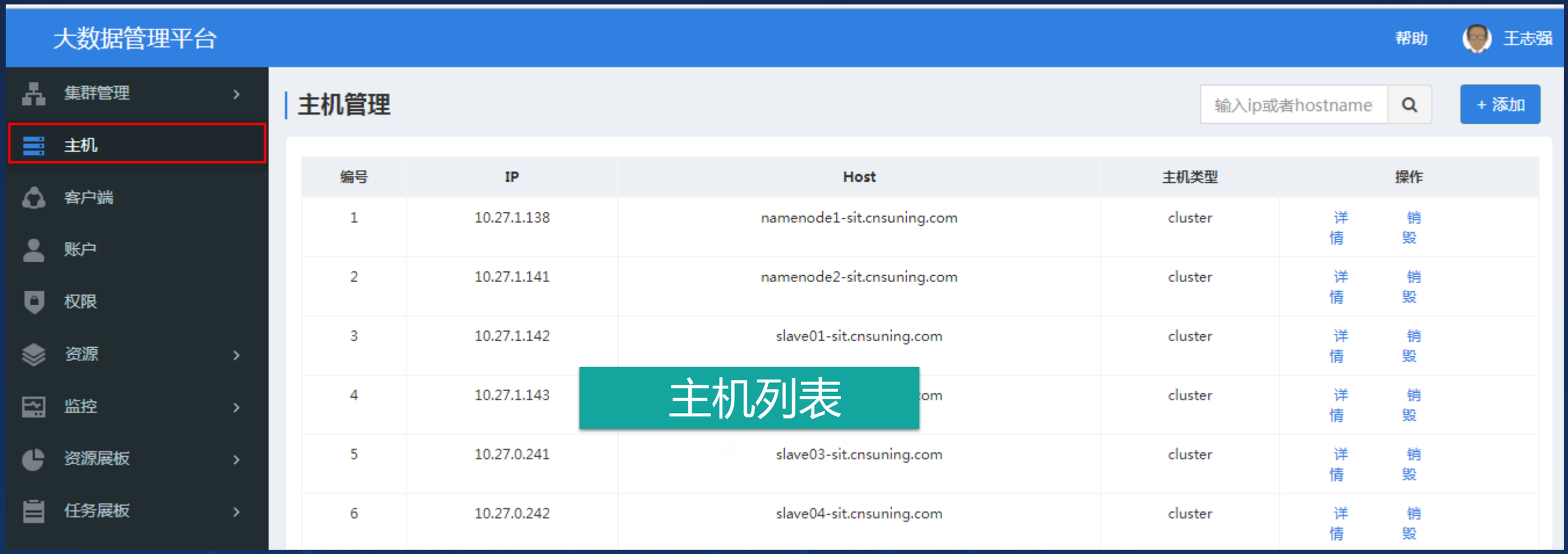
- 集群组件辣么多，部署环境辣么多，累成狗。。。
- 堡垒机->跳板机->运维机，就为了开个权限。。。
- 谁懂Flume运维的苦？

业务开发

- 我开发用到一个用户行为的数据，在哪个表里？
- 配置crontab的方式串联任务，伤不起。。。
- 不同账户间的任务相互依赖严重，跨账户依赖难实现

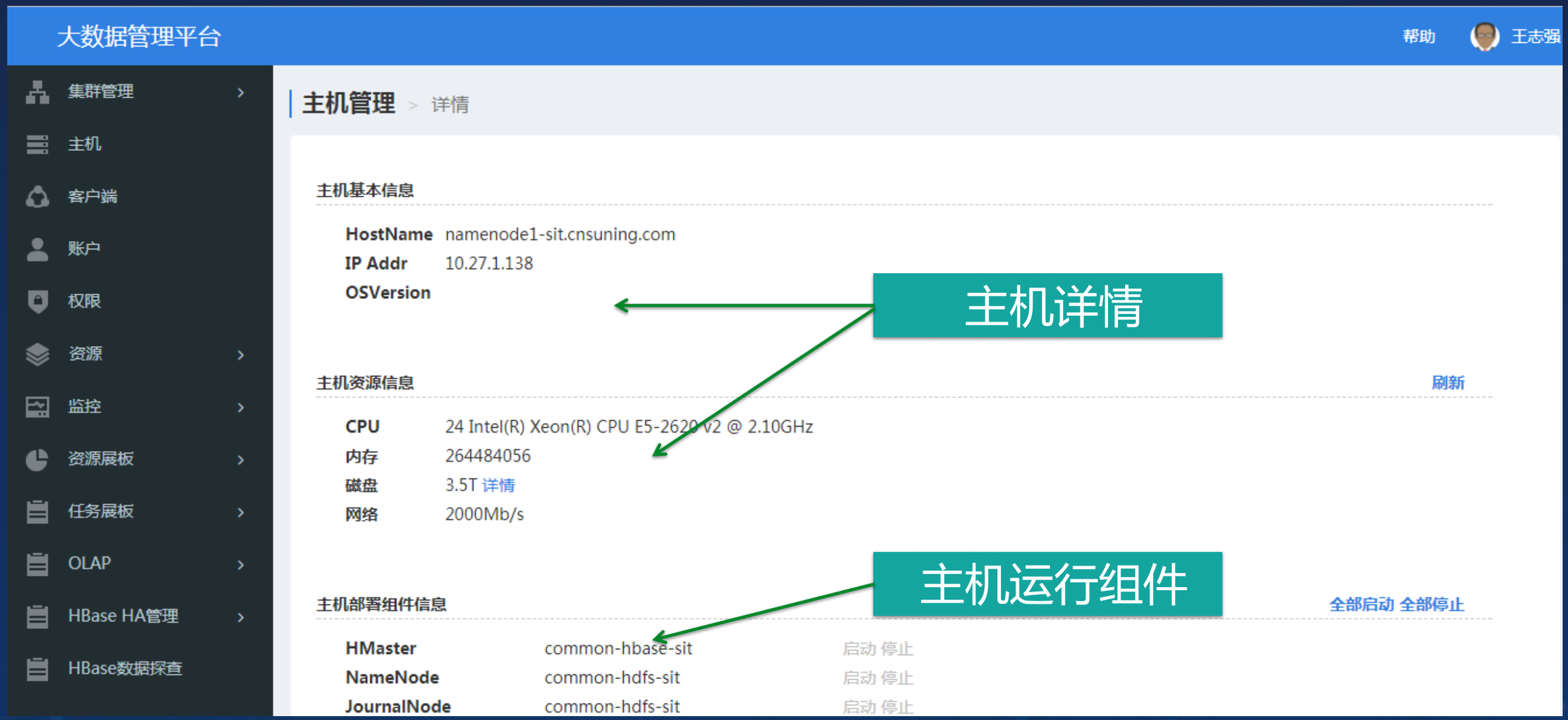
解决1. 平台化、自动化

大数据管理平台：主机管理，集群管理自动化



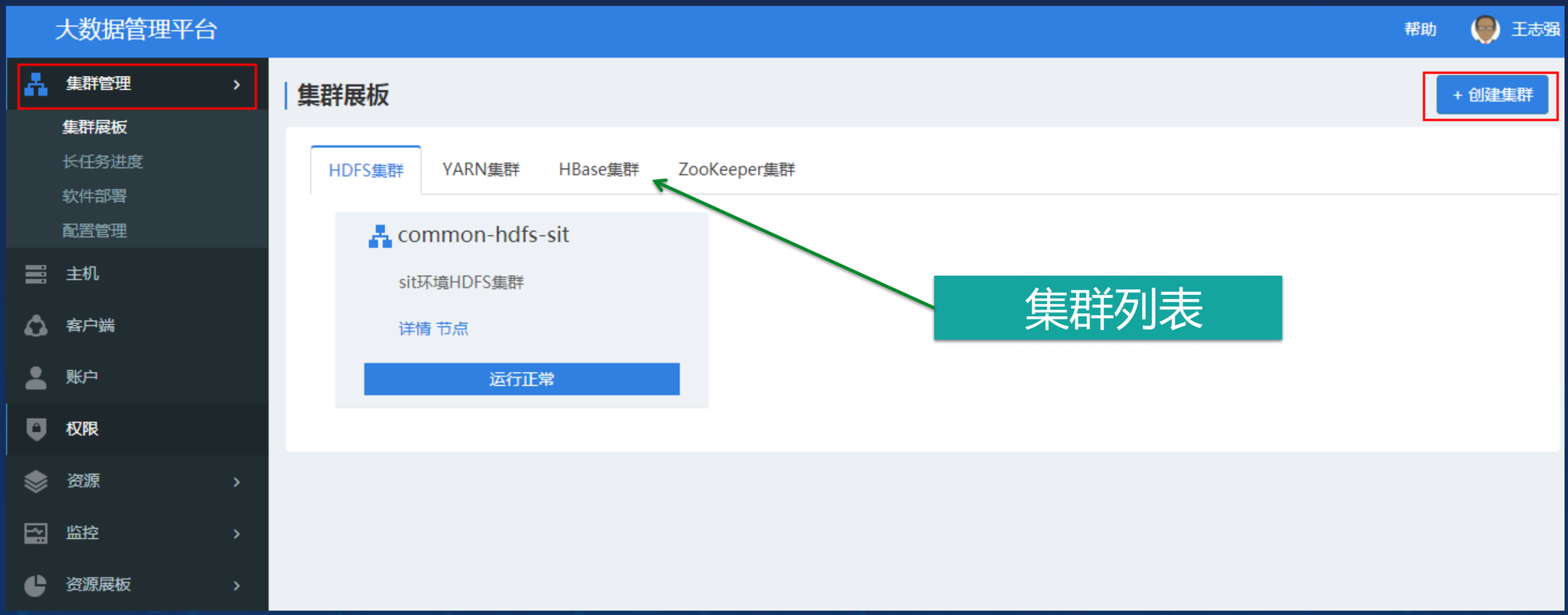
解决1. 平台化、自动化

大数据管理平台：主机管理，集群管理自动化



解决1. 平台化、自动化

大数据管理平台：主机管理，集群管理自动化

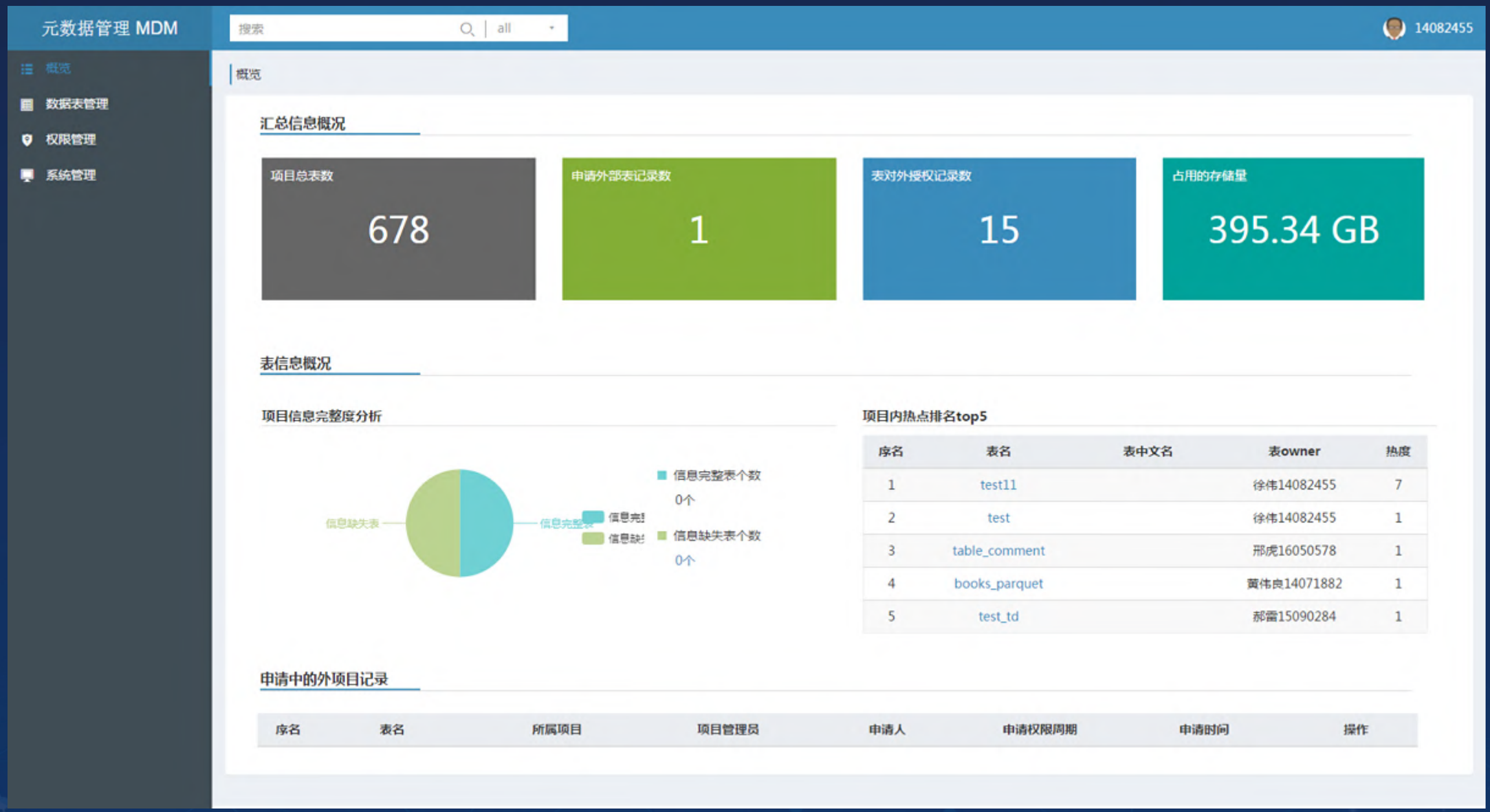


解决1. 平台化、自动化

大数据管理平台：主机管理，集群管理自动化

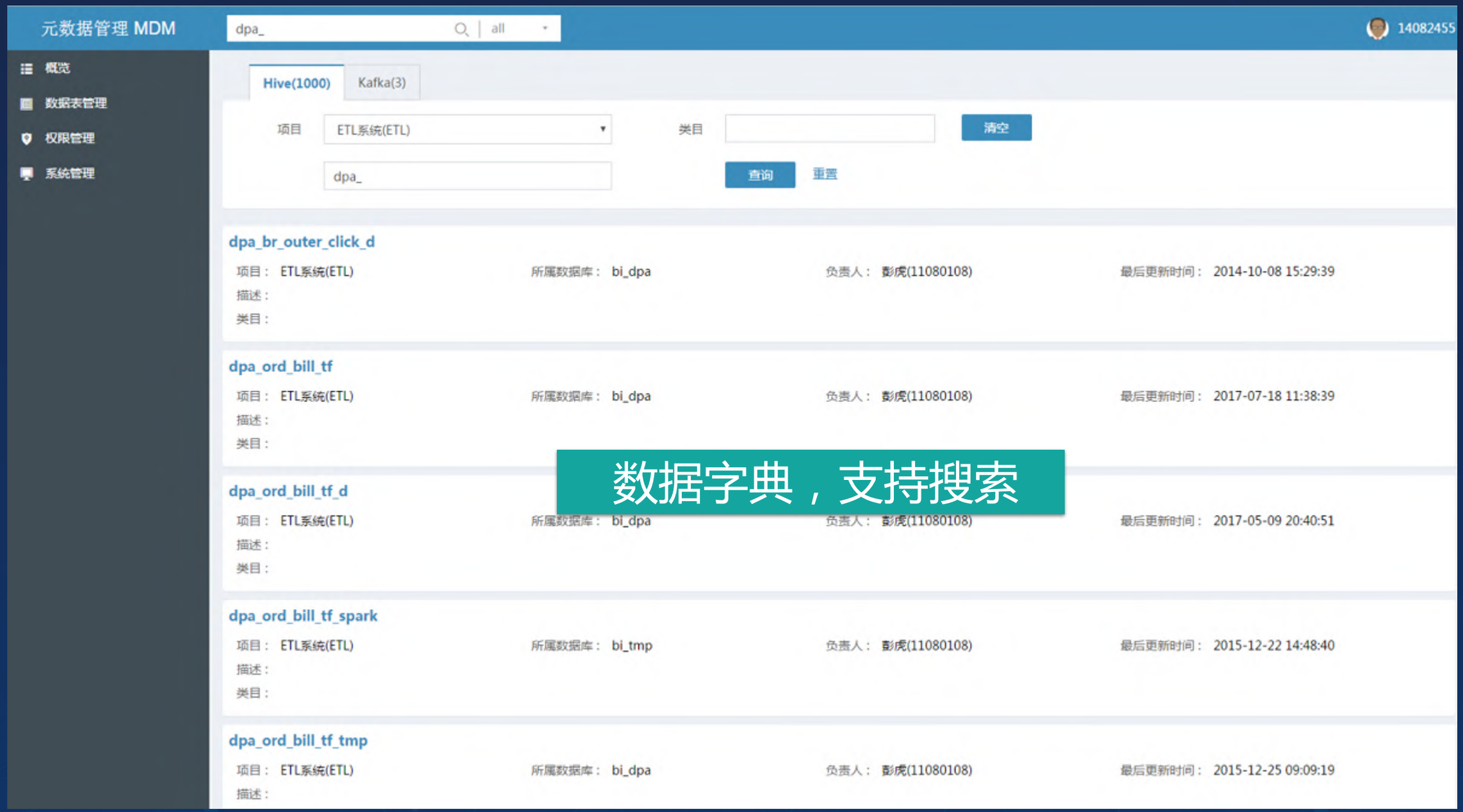
解决1. 平台化、自动化

元数据管理：数据字典，权限申请审批实施自动化



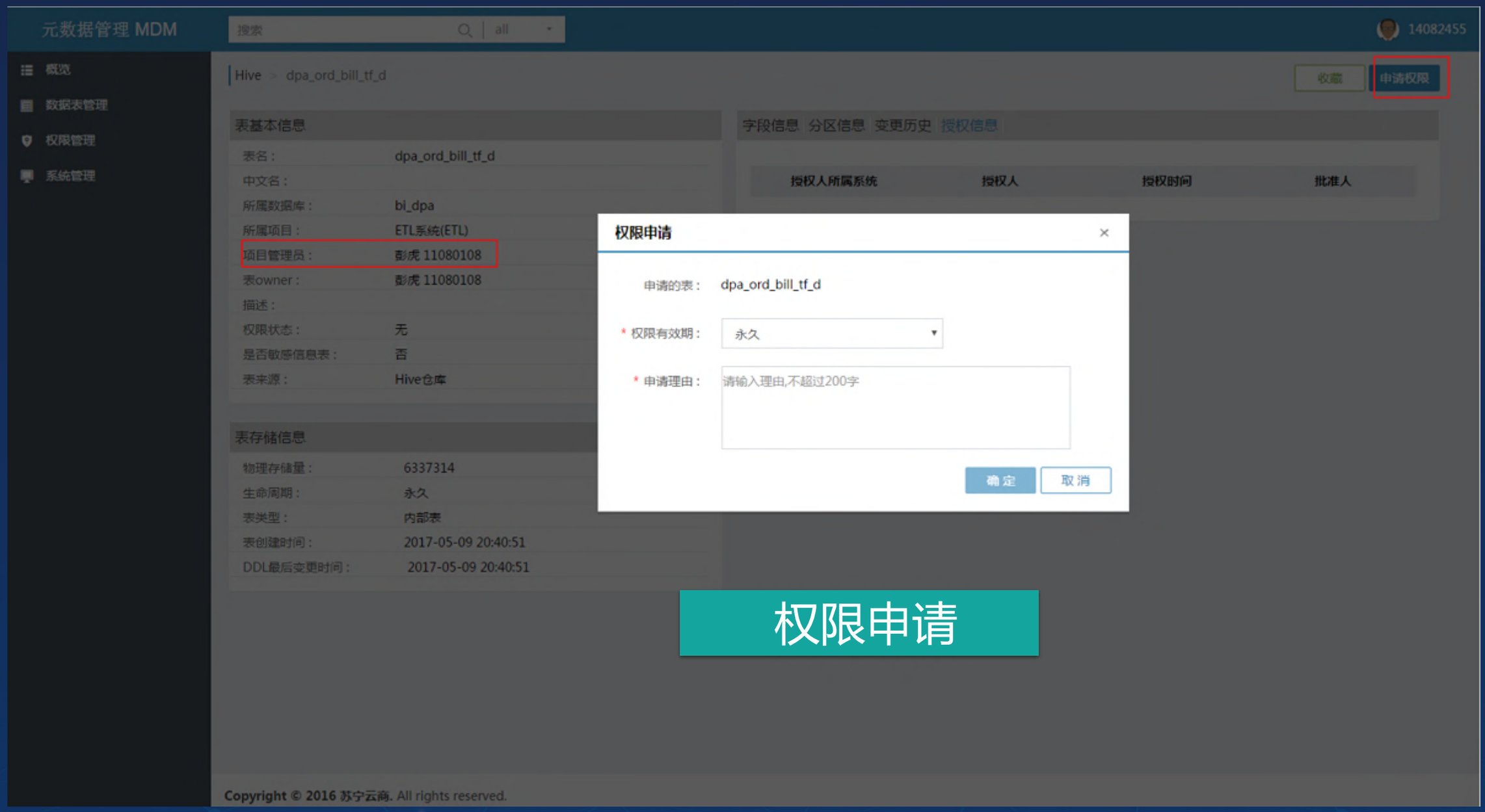
解决1. 平台化、自动化

元数据管理：数据字典，权限申请审批实施自动化



解决1. 平台化、自动化

元数据管理：数据字典，权限申请审批实施自动化



权限申请

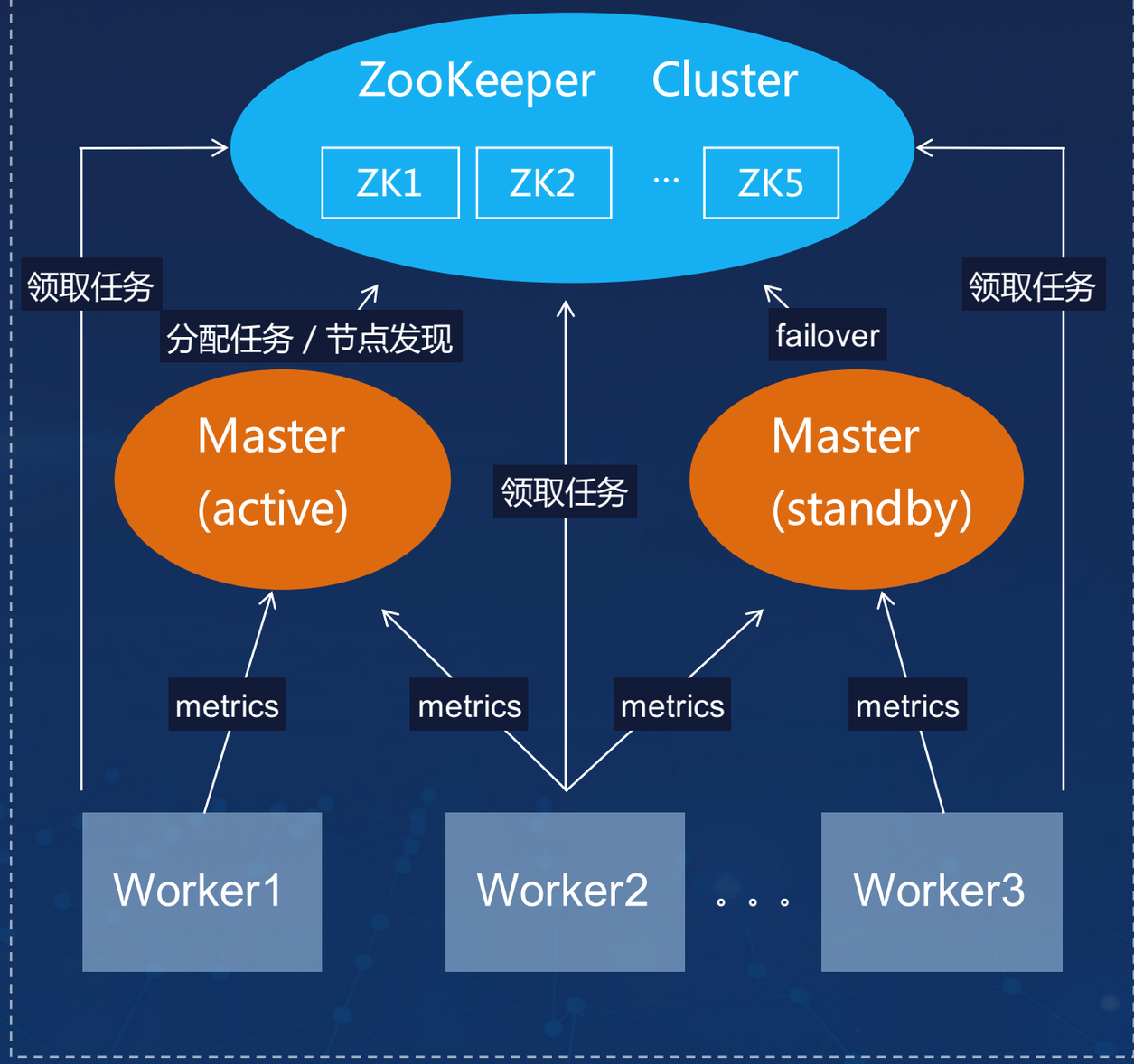
解决1. 平台化、自动化

数据流管理平台：集成Flume，智能扩缩容，插件式

2节点小集群
对接<1000qps业务

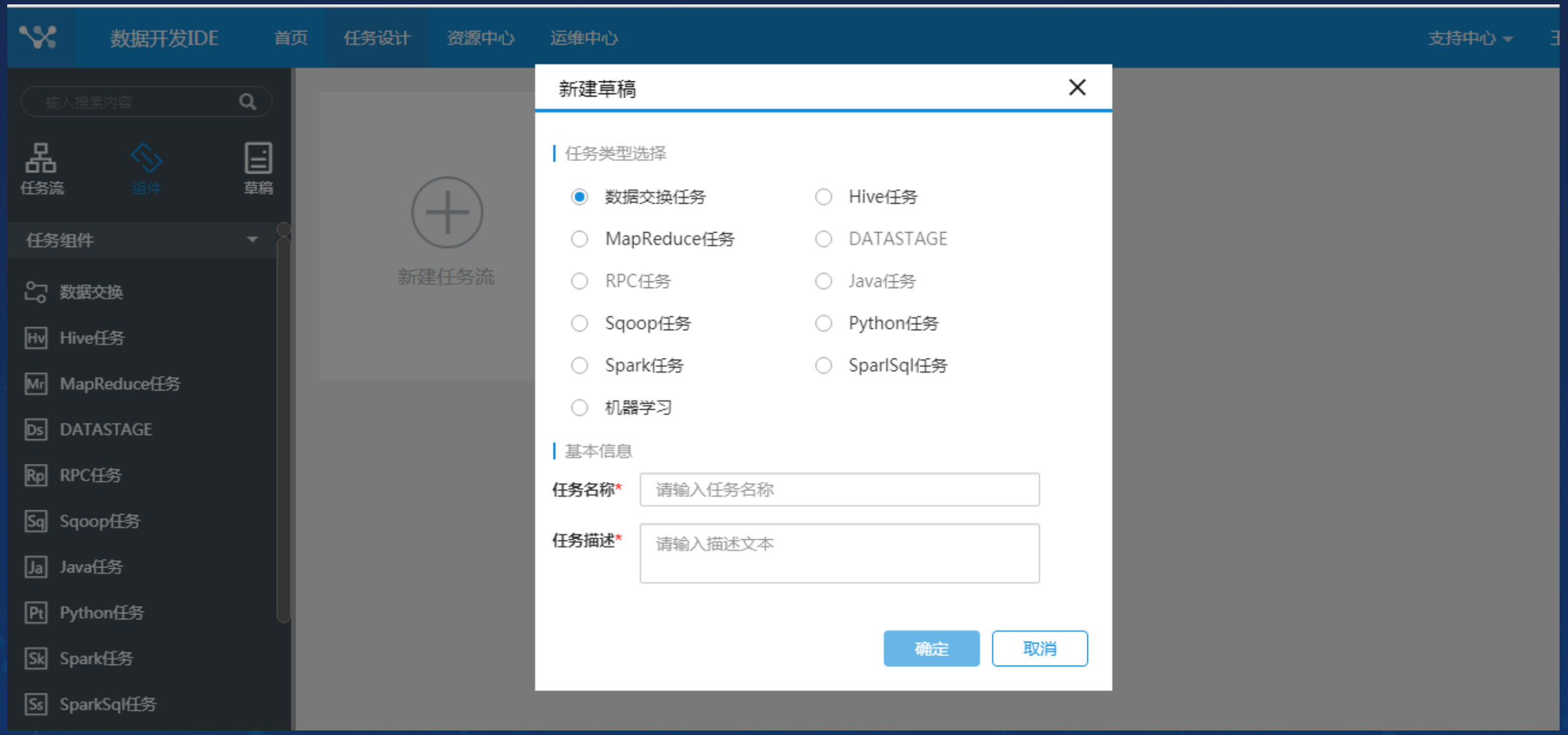
10节点中集群
对接1000-10000qps业务

30节点大集群
对接>10000qps业务



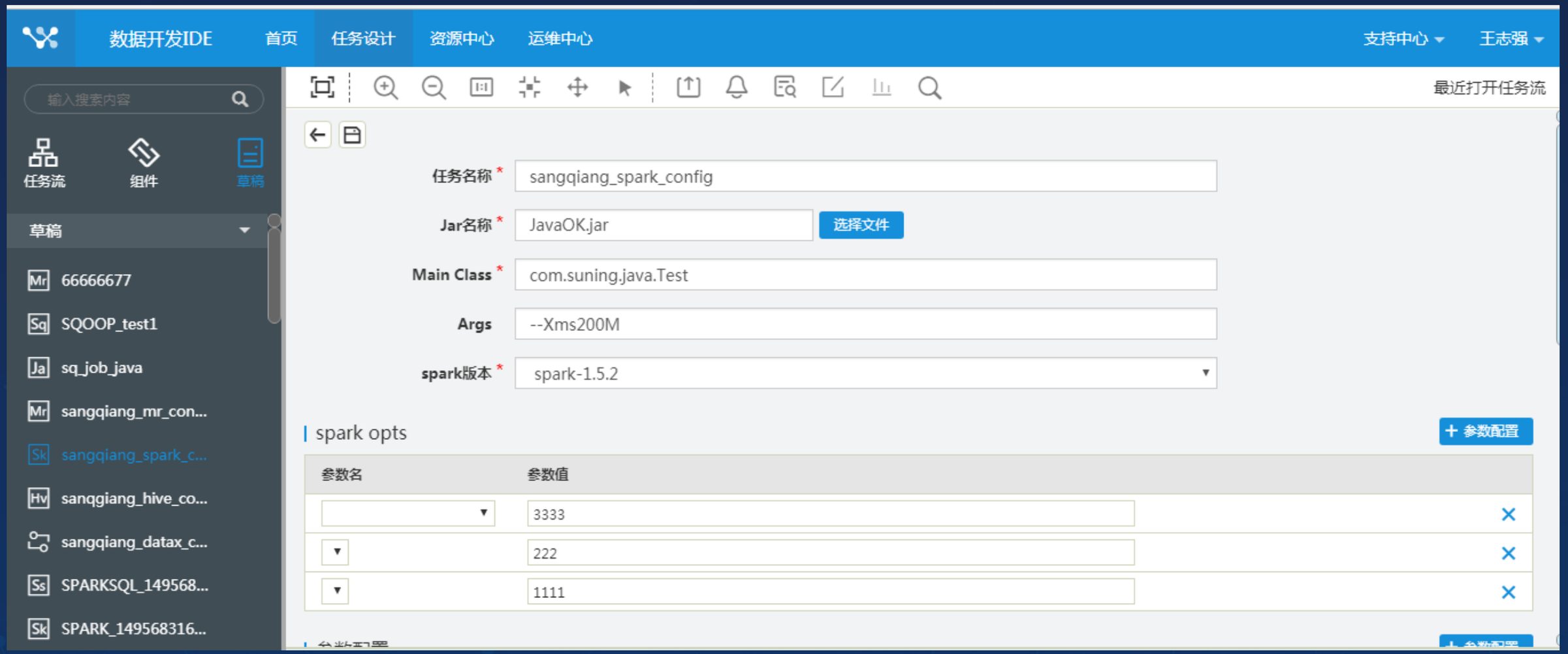
解决1. 平台化、自动化

数据开发平台：支持10种不同的任务类型，支持任务流/任务管理，解决复杂依赖问题，可扩展



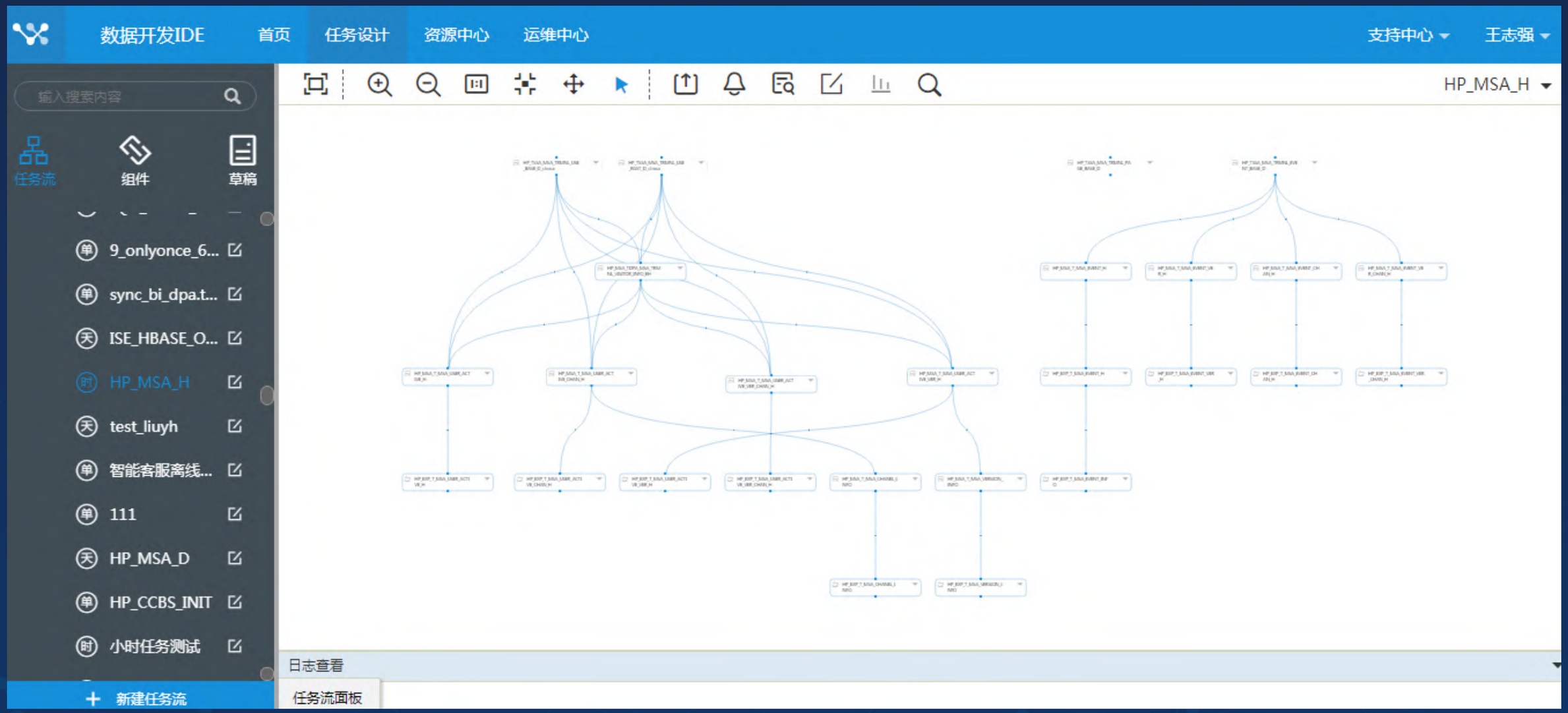
解决1. 平台化、自动化

数据开发平台：支持10种不同的任务类型，支持任务流/任务管理，解决复杂依赖问题，可扩展

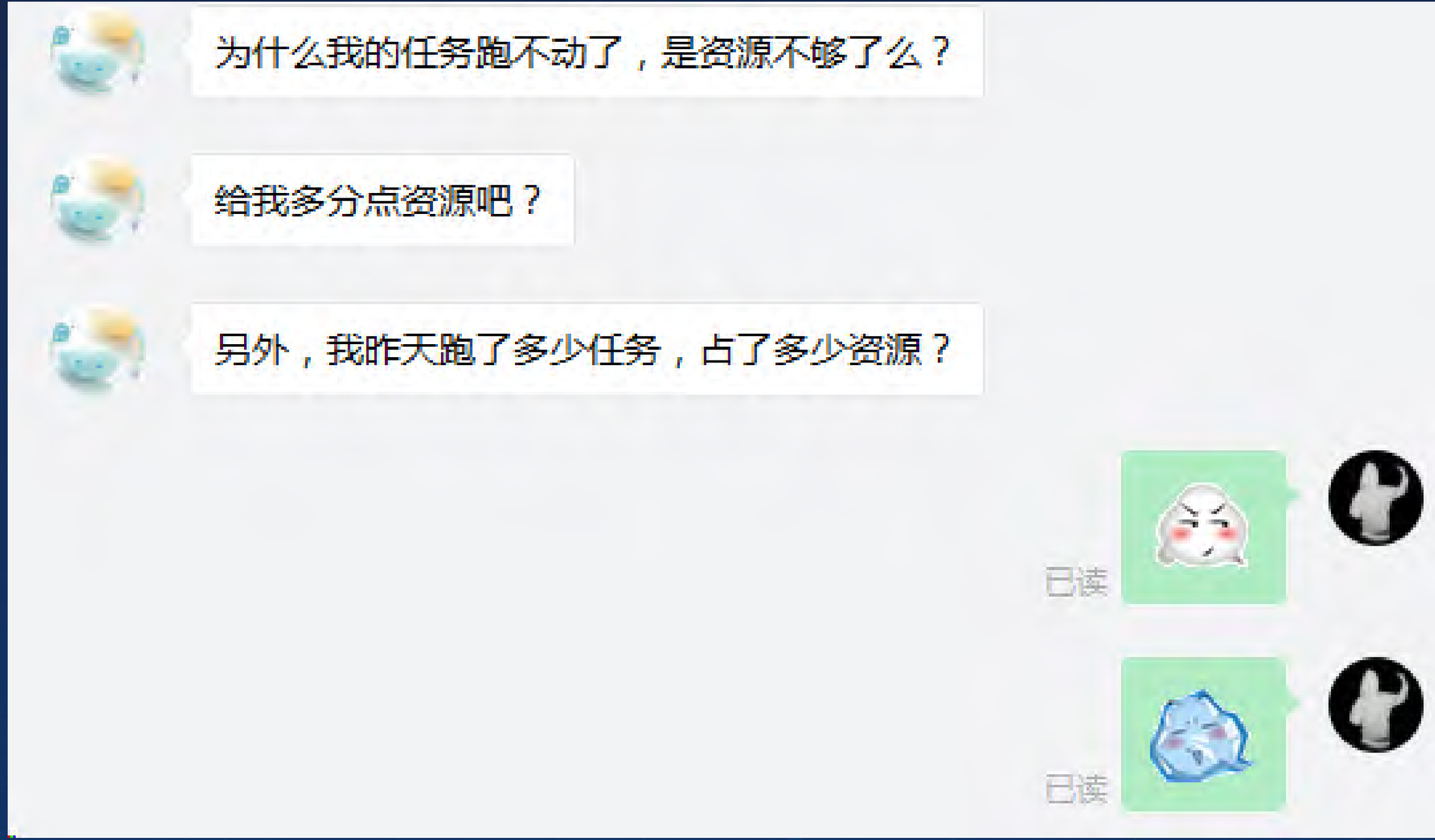


解决1. 平台化、自动化

数据开发平台：支持10种不同的任务类型，支持任务流/任务管理，解决复杂依赖问题，可扩展



痛点2. 无资源使用视图



解决2. 资源可视化、人民币化

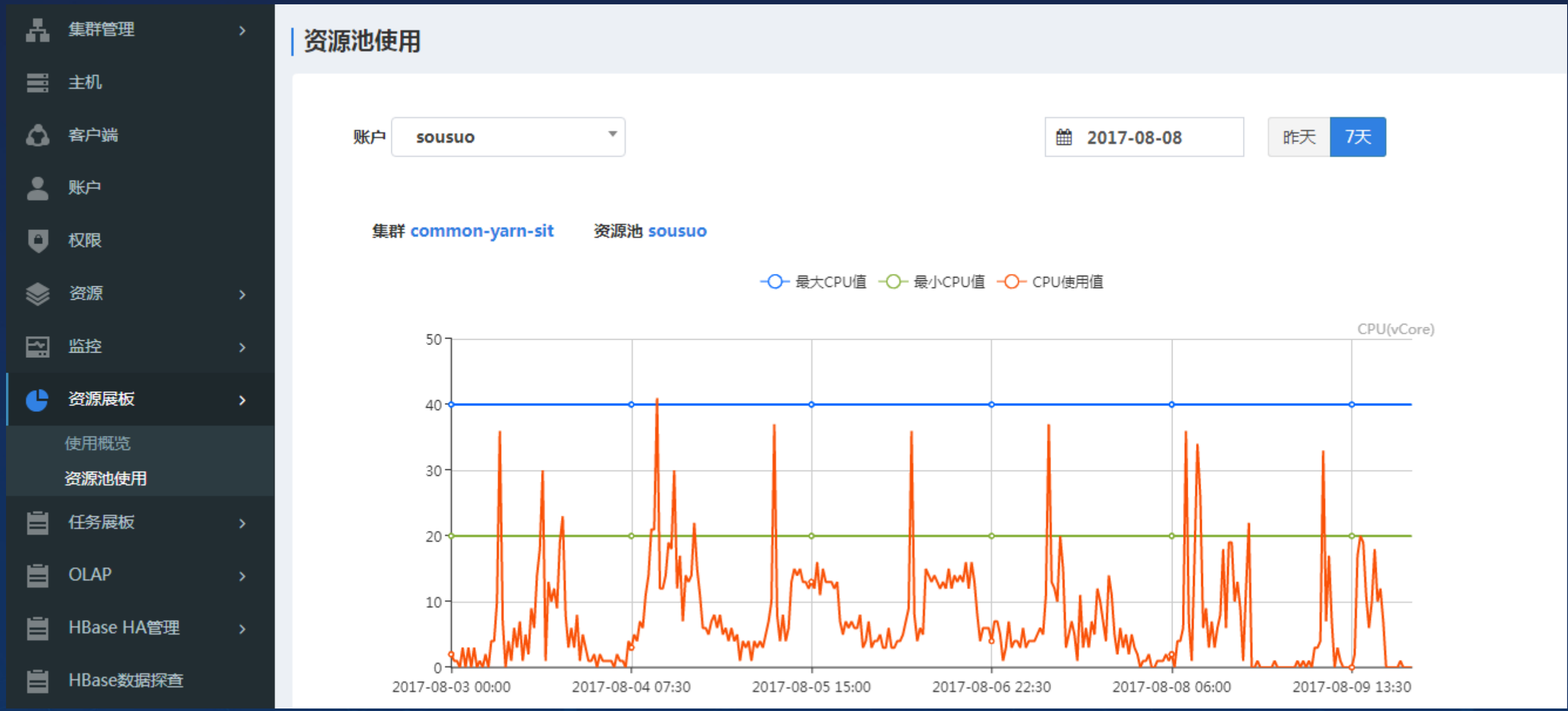
- ✓ 存储/计算资源计量计费
- ✓ 资源池使用可视化
- ✓ 任务展板

解决2. 资源可视化、人民币化

存储/计算资源计量计费

解决2. 资源可视化、人民币化

资源使用可视化

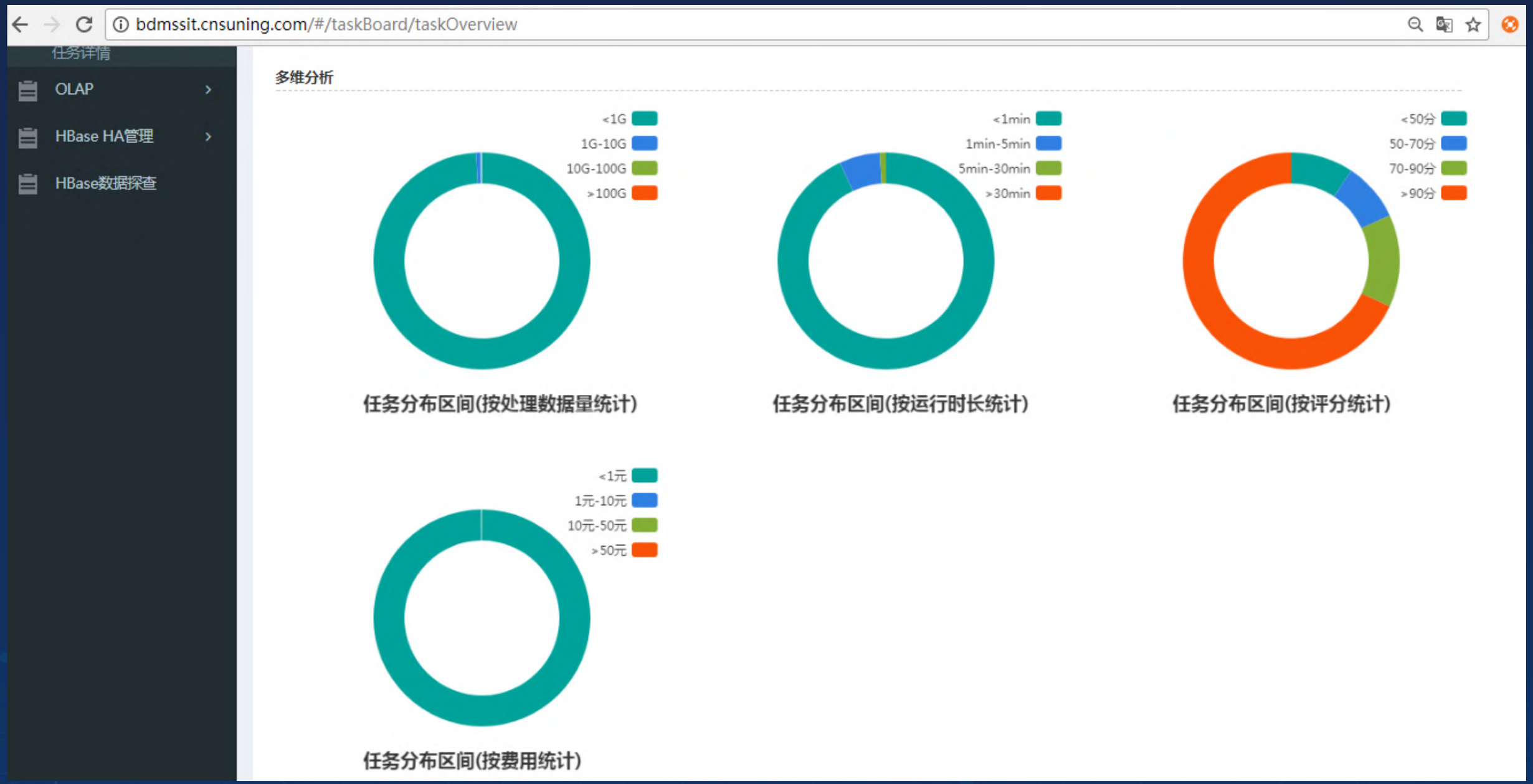


解决2. 资源可视化、人民币化

任务展板

解决2. 资源可视化、人民币化

任务展板



解决2. 资源可视化、人民币化

任务展板

大数据管理平台

帮助 王志强

- 集群管理
- 主机
- 客户端
- 账户
- 权限
- 资源
- 监控
- 资源展板
- 任务展板
  - 排行榜
  - 任务概览
  - 任务详情
- OLAP
- HBase HA管理
- HBase数据探查

排行榜

组件 All

2017-08-09

任务总量Top10

排名	账户	任务数
1	csi	3347
2	bigdata	2232
3	erp	1274
4	lbi	1086
5	bi	779
6	srdss	361
7	scpdm	240
8	smmb	205
9	sopdm	191
10	pos	155

处理数据量Top10

排名	账户	处理数据量
1	spider	706.78GB
2	srdss	271.88GB
3	ztbd	210.14GB
4	bigdata	165.41GB
5	pcms	148.15GB
6	mobdss	118.66GB
7	sousuo	29.57GB
8	pos	27.91GB
9	scpdm	15.15GB
10	lbi	12.13GB

任务评分Top10（由高到低）

排名	账户	评分
1	[illegible]	100

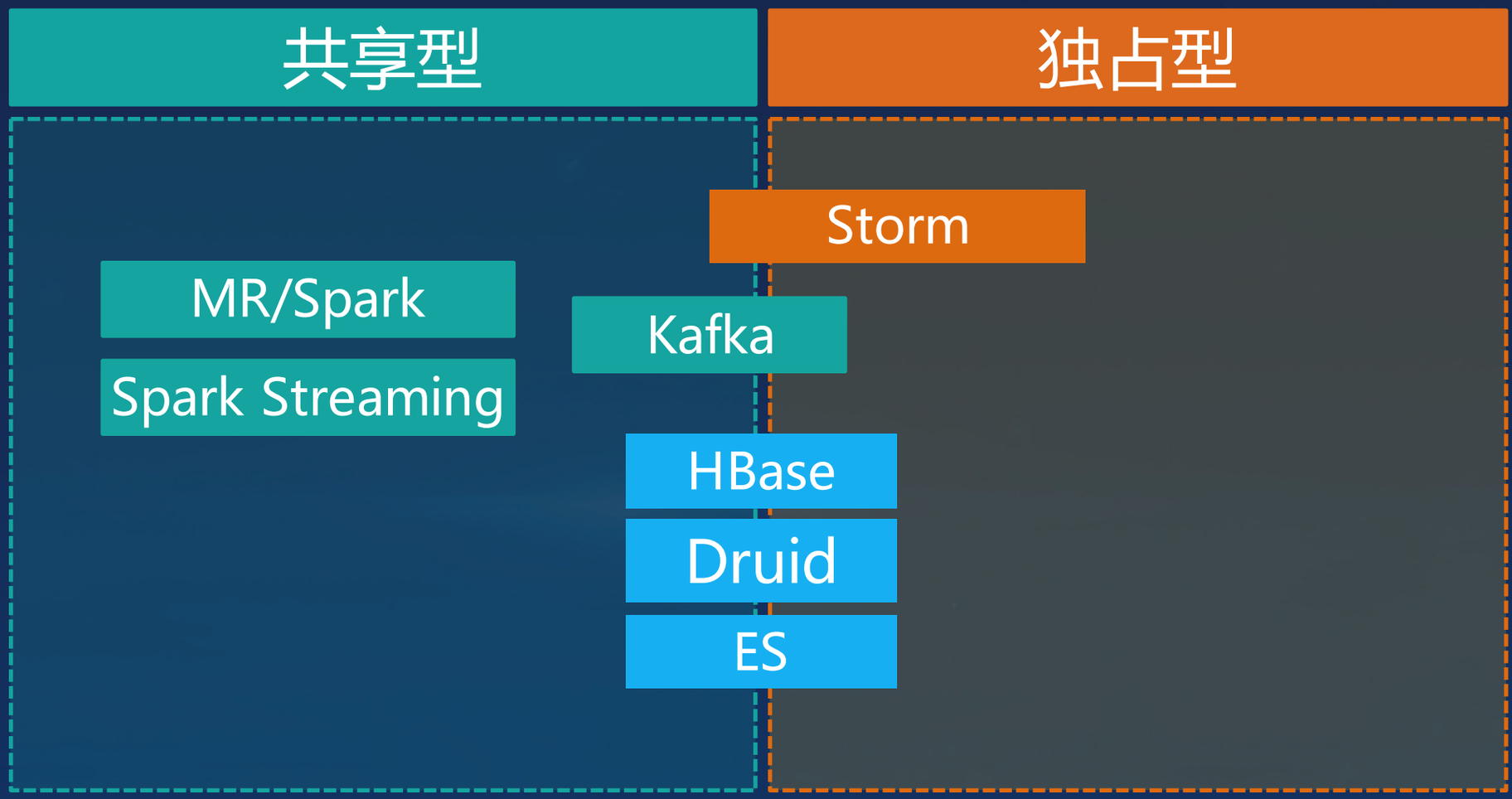
任务评分Top10（由低到高）

排名	账户	评分
1	[illegible]	48

痛点3. 任务互相影响，资源隔离性差

我的任务为什么突然变慢了？
是不是别人影响我了？

解决3. 差异化服务、物理隔离



资源浪费：离线、流本来就是错峰的

解决3. 差异化服务、物理隔离

共享型-YARN

FairScheduler：每个账户一个Pool，配置最大最小资源；最小资源是能保证的资源，最大资源是上限；

优先级: backport from 2.8；改进为7个优先级，每提升1个优先级，获得分配权重增加1倍；

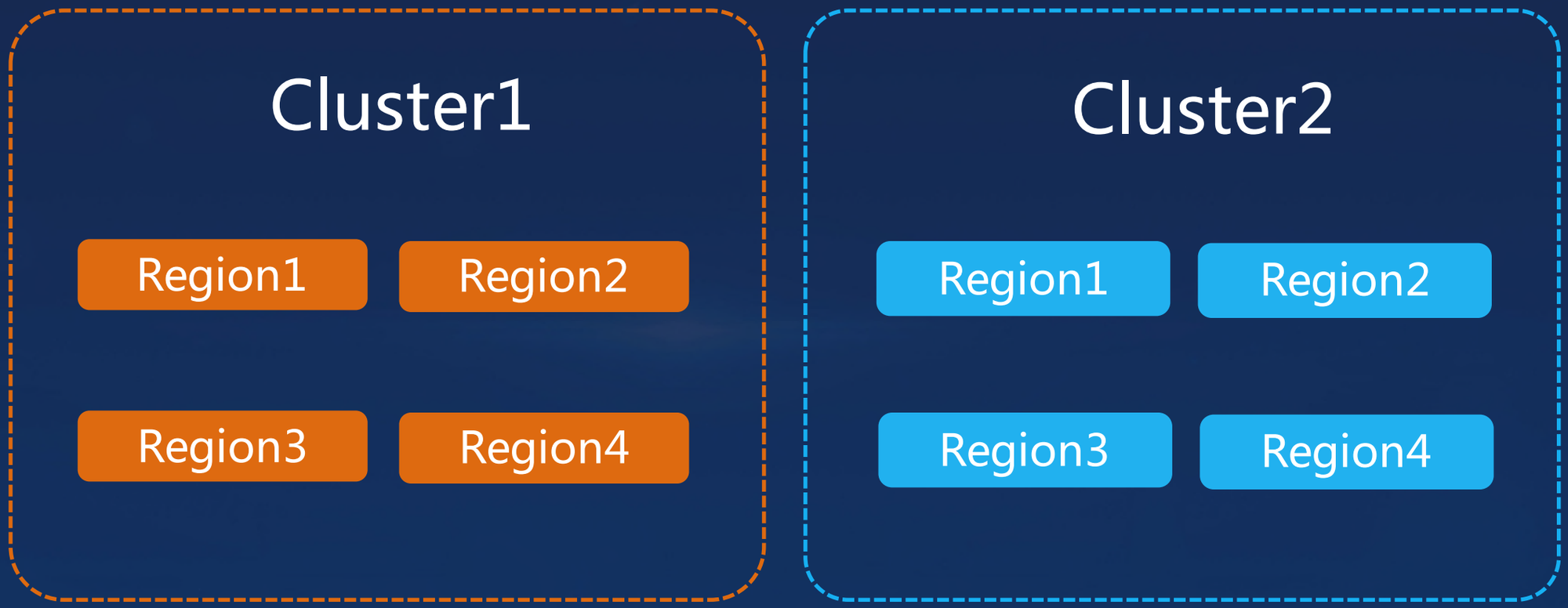
降低长GC的影响：尽快失败，合理设置 GCTimeLimit=90 GCHeapFree Limit=10 ParallelGCThreads=8；

Cgroup

开启抢占

解决3. 差异化服务、物理隔离

共享型-Kafka



避免大集群：对ZK的压力，对controller的压力

逻辑分区：按照不同的业务场景分配，避免不同保障级别业务的相互影响

解决3. 差异化服务、物理隔离

独占型-Storm

独立的物理区域

物理CPU与虚机CPU 1:1 亲和性绑定

CPU ST从40%+降低至1%

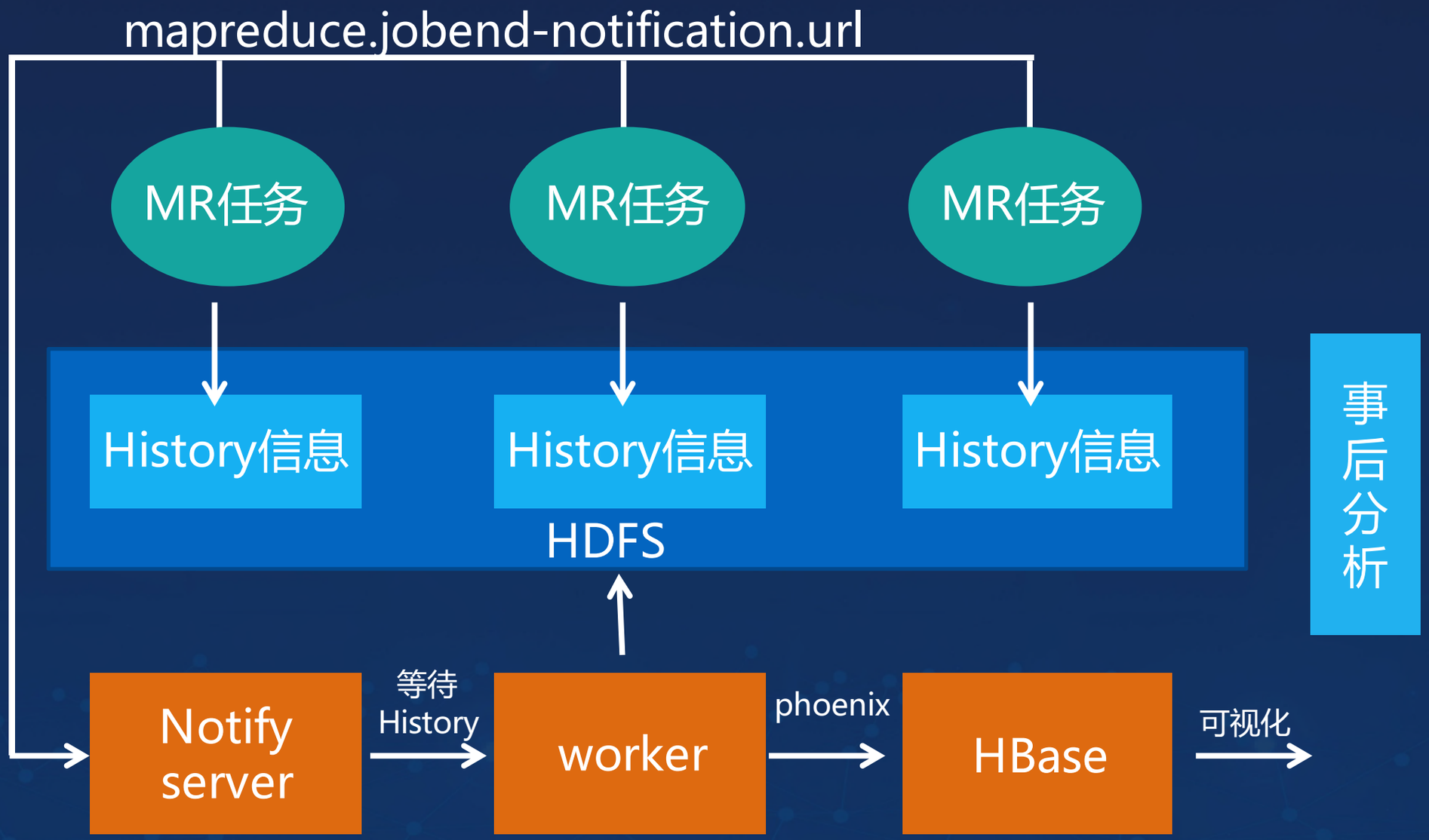
痛点4. 排查问题耗时长，应用优化门槛高

用户不清楚任务为什么失败

用户不清楚该如何优化任务

解决4. 智能诊断、优化建议

离线计算平台-MR任务分析平台



解决4. 智能诊断、优化建议

离线计算平台-MR任务分析平台

任务排行榜

评分最低Top20

排名	账户	jobID	评分
1	bi	job_1482378778761_19995963	44
2	aps	job_1482378778761_19986972	44
3	cloudytrace	job_1482378778761_19882779	44
4	aps	job_1482378778761_20003396	44
5	smmb	job_1482378778761_20040120	44
6	smmb	job_1482378778761_19987516	44
7	smmb	job_1482378	
8	lbi	job_1482378	
9	lbi	job_1482378778761_19888904	44
10	birec	job_1482378778761_20032814	44

处理数据量最多Top20

排名	账户	jobID	处理数据量
1	sousuo	job_1482378778761_20032979	2.57TB
2	sousuo	job_1482378778761_20039489	2.57TB
3	srdss	job_1482378778761_20002279	2.57TB
4	spider	job_1482378778761_19879587	2.44TB
5	spider	job_1482378778761_19875138	2.39TB
6	safe	job_1482378778761_19953908	1.99TB
7		job_1482378778761_19950502	1.99TB
8		job_1482378778761_19961139	1.99TB
9	srdss	job_1482378778761_19996728	1.46TB
10	osmos	job_1482378778761_19879911	1.41TB

不同维度的排行榜

总CPU时间最长Top20

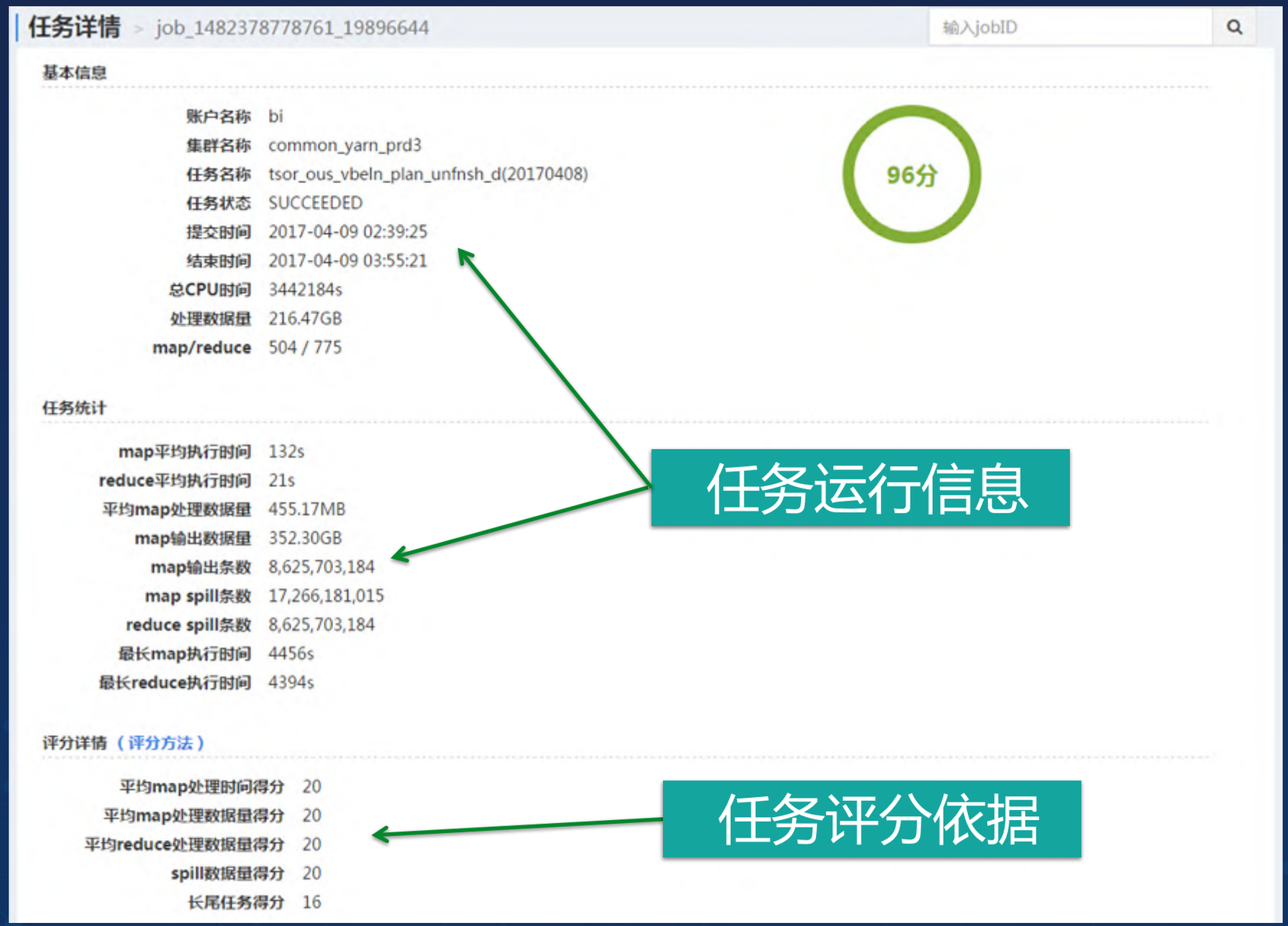
排名	账户	jobID	耗时
1	bi	job_1482378778761_19896644	3442184
2	bi	job_1482378778761_19903213	2259819
3	erp	job_1482378778761_20004500	2040751
4	erp	job_1482378778761_20010608	1533820

Task数排名Top20

排名	账户	jobID	Task数
1	sopdm	job_1482378778761_19987327	22908
2	aps	job_1482378778761_19979424	20925
3	aps	job_1482378778761_19977850	20391
4	ztbd	job_1482378778761_19916949	17344

解决4. 智能诊断、优化建议

离线计算平台-MR任务分析平台



解决4. 智能诊断、优化建议

离线计算平台-MR任务分析平台

任务详情 > job_1499408566046_8795319

输入任务ID

基本信息

账户名称	bjredbaby
集群名称	common_yarn_prd3
任务名称	create table tmp_zq_zs_order_06 stored ...)t(Stage-1)
任务状态	FAILED
提交时间	2017-08-10 17:14:18
结束时间	2017-08-10 17:15:37
总CPU时间	1.25min
处理数据量	0MB
map/reduce	1 / 4

失败

失败原因

失败原因	数字格式化异常
建议解决方案	["检查数字格式化"]

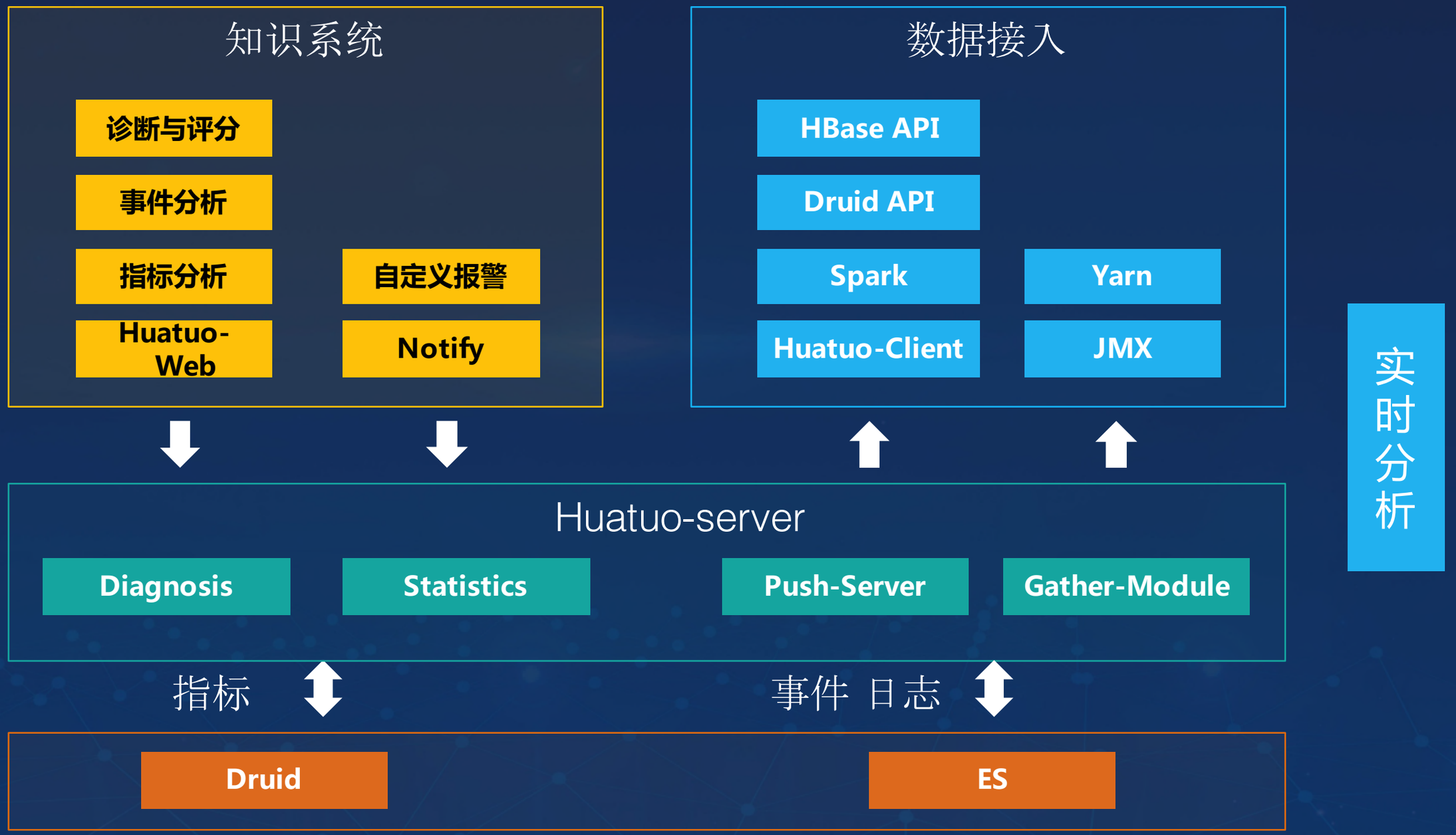
失败诊断信息

异常日志信息

Log Type: stderr Log Length: 243 log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.impl.MetricsSystemImpl). log4j:WARN

解决4. 智能诊断、优化建议

离线计算平台-华佗智能诊断平台(spark)



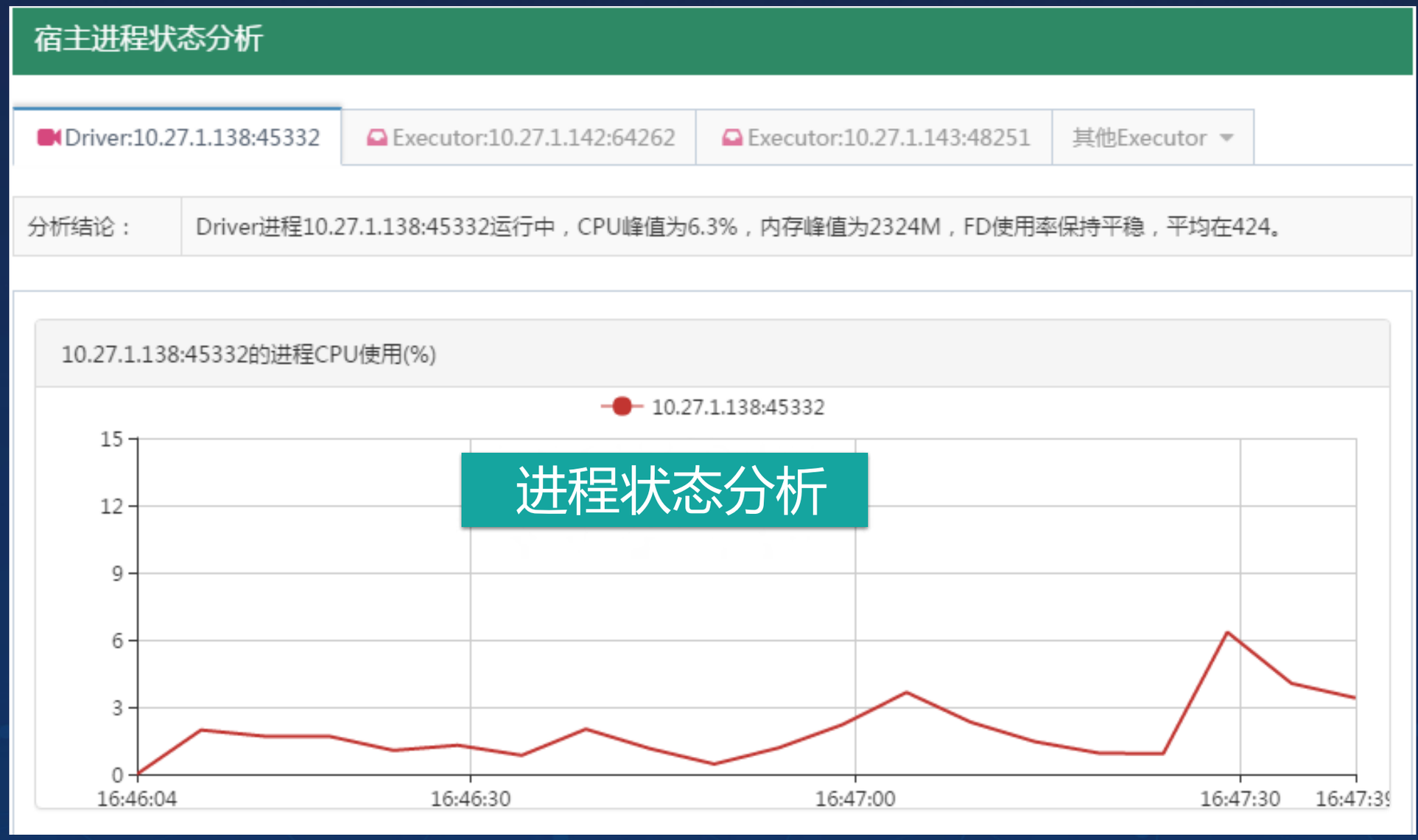
解决4. 智能诊断、优化建议

离线计算平台-华佗智能诊断平台(spark)

资源角度	性能角度	故障角度
• 宿主机器状态分析 • HDFS资源使用分析	• Task耗时链分析 • 长尾Task分析	• Shuffle数据倾斜 • HDFS Commit阻塞 • Executor丢失 • Spill事件分析 • 高维Parquet写性能诊断 • RDD Size Estimator耗时诊断 • 任务事件流
• Driver/Executor进程状态分析 • 资源利用率分析	• 任务调度Overhead分析 • Reduce并发度分析 • JDBC并发度分析Kafka读并发度分析	
• Cache 利用率分析 • Shuffle内存利用率分析		

解决4. 智能诊断、优化建议

离线计算平台-华佗智能诊断平台(spark)



解决4. 智能诊断、优化建议

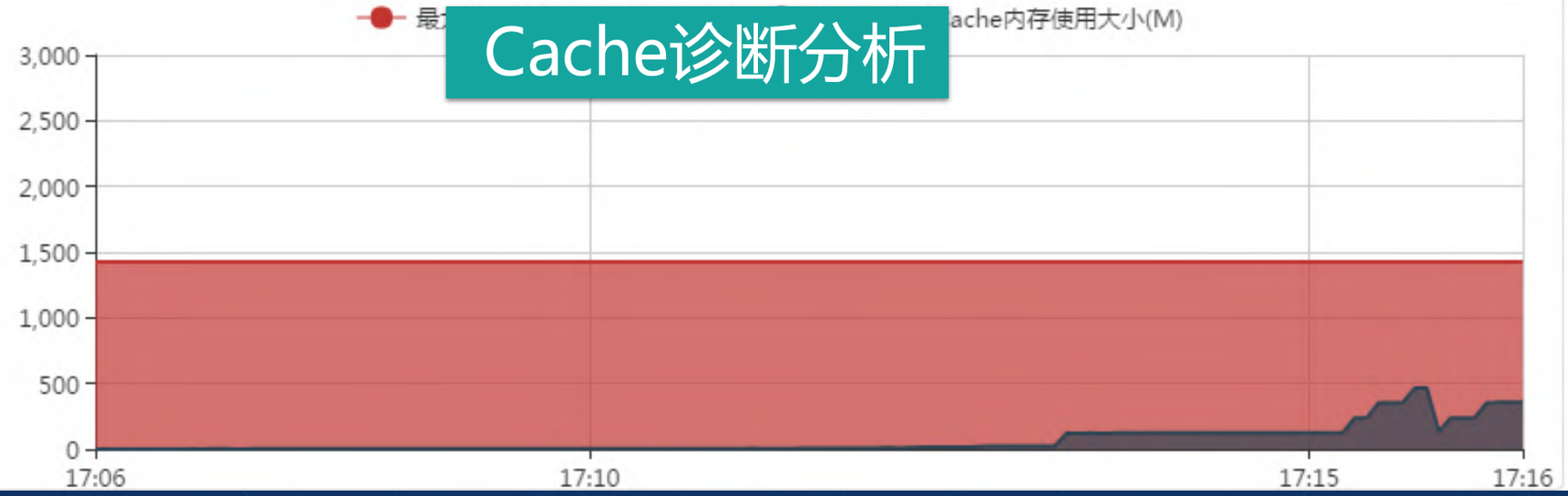
离线计算平台-华佗智能诊断平台(spark)

App Cache诊断分析

Cache建议：	RDD[id=3,name='MapPartitionsRDD']被重复调用了2次; RDD[id=9,name='MapPartitionsRDD']被重复调用了2次; RDD[id=12,name='MapPartitionsRDD']被重复调用了2次; RDD[id=17,name='MapPartitionsRDD']被重复调用了2次; 请将需要cache的rdd进行cache。

Executor:10.37.2.141:55693 | Executor:10.27.1.143:52320 | Executor:10.27.0.242:45151 | 其他Executor

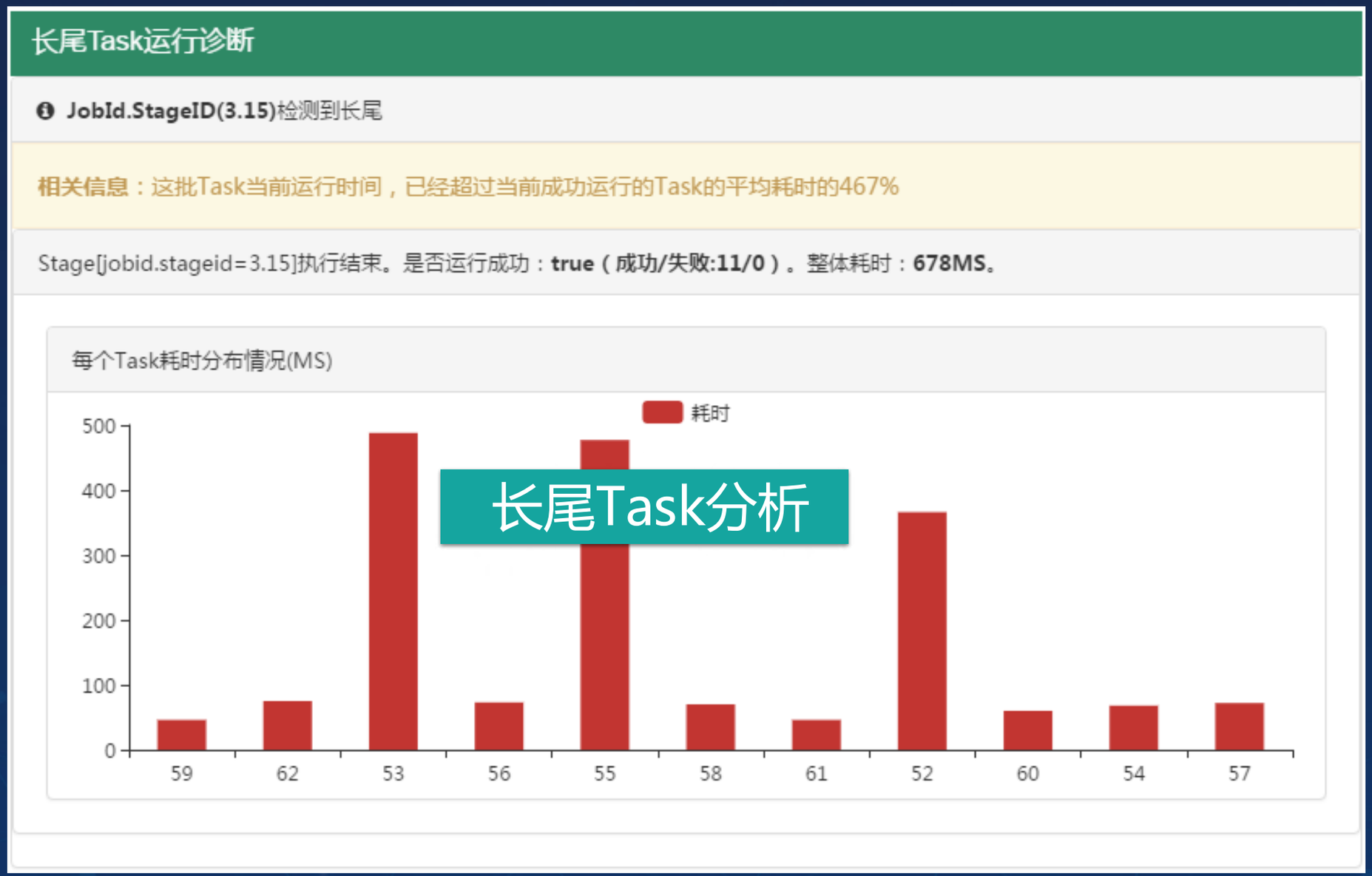
10.27.0.242:45151的Cache内存段使用



Cache诊断分析

解决4. 智能诊断、优化建议

离线计算平台-华佗智能诊断平台(spark)



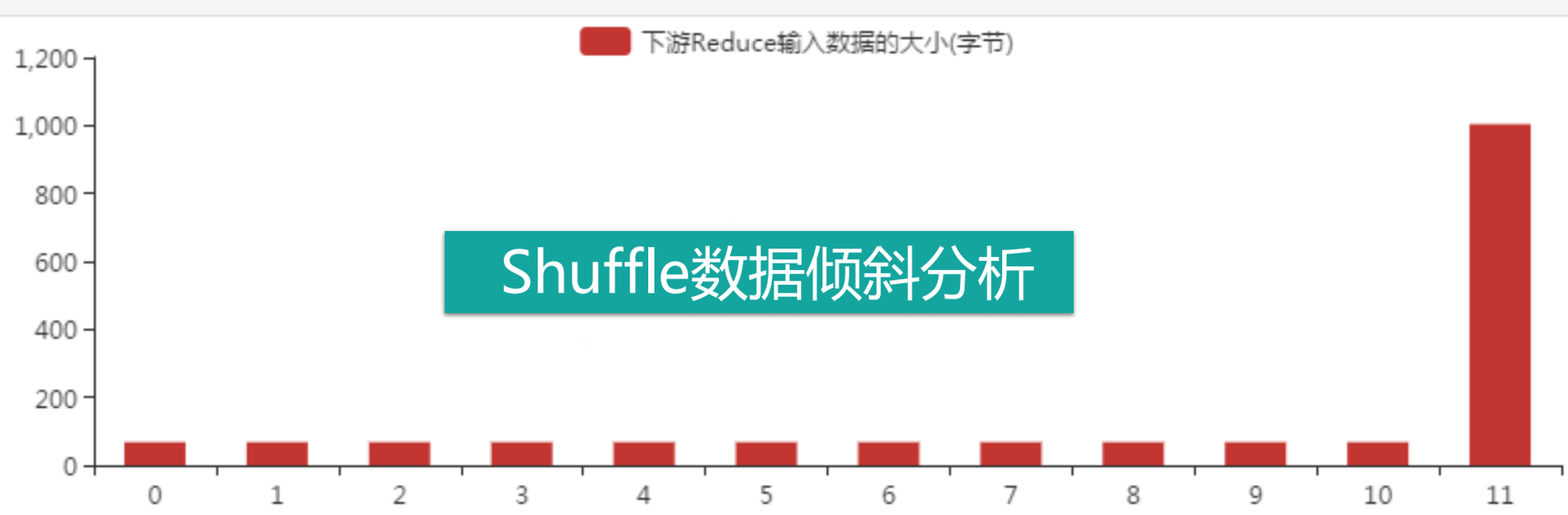
解决4. 智能诊断、优化建议

离线计算平台-华佗智能诊断平台(spark)

Shuffle数据倾斜诊断

事件： [**INFO**] ShuffleWriteStage[**9.17**]执行结束。任务耗时：**4918MS**。ShuffleWrite的Meta大小为：**228字节**。

下游Reduce输入数据的大小(字节)



Shuffle数据倾斜分析

解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统

进程重启

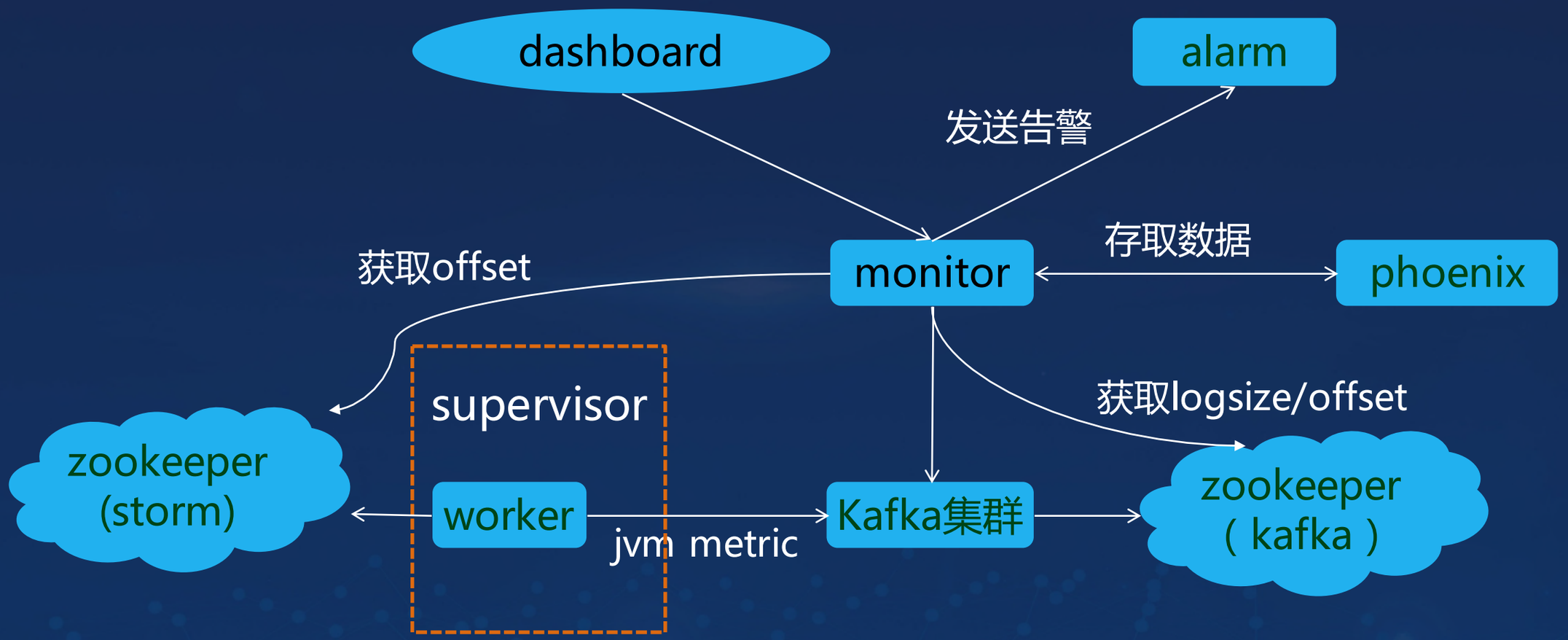
- 代码异常
- OutOfMemory
- 宿主机/虚机宕机

Kafka消息堆积

- 数据量暴增
- FGC/Capacity
- 并发不够
- 任务异常停止

解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统



解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统

ppcsHomePageB健康状态

2017年08月10日15时04分~2017年08月10日22时59分

1、Topology基本信息

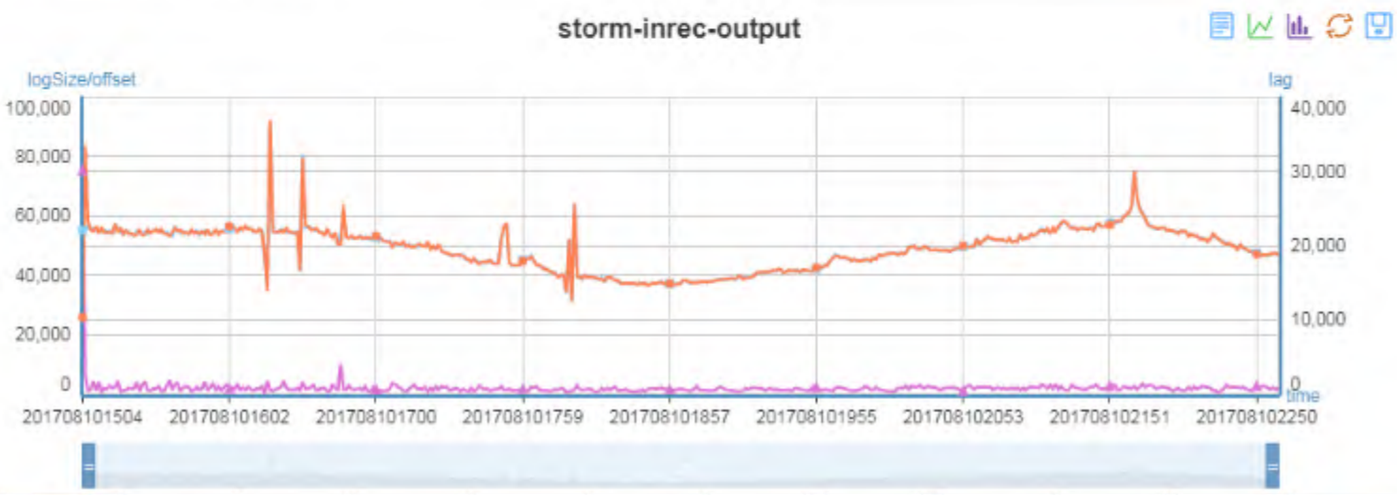
启动时长	Num workers	Num executors	Num tasks	spout.pending	message.timeout(s)	集群ID
7h54m52s	140	449	449	1000	360	ppcs-v-storm

2、Topology配置诊断

- topology的bolt并发设置合理，无需调整。此次启动也无worker发生迁移

3、数据消费视图（2017年08月10日15时04分~2017年08月10日22时59分）

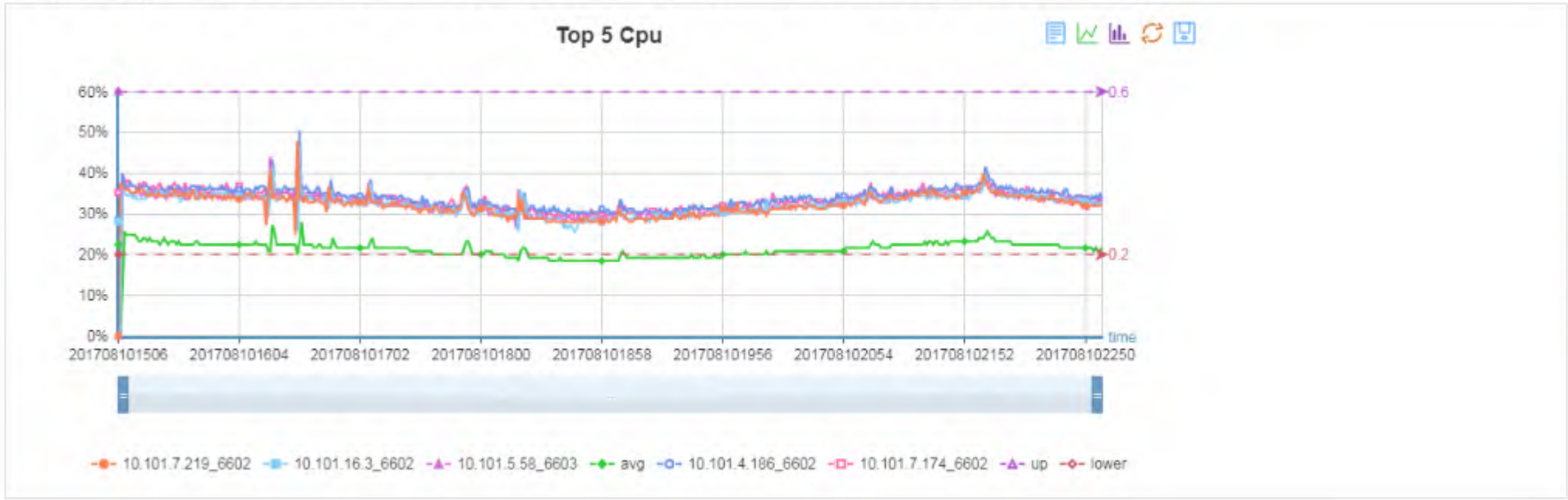
topic名： storm-inrec-output



解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统

4、Topology运行CPU（2017年08月10日15时04分～2017年08月10日22时59分）

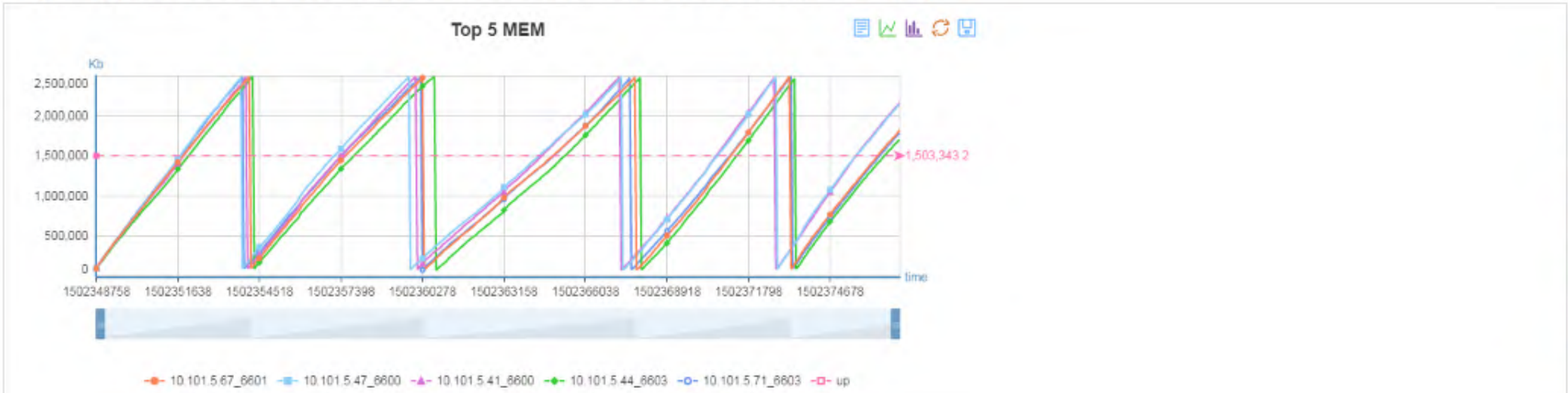


- worker:10.101.7.219_6602的cpu使用率远远高于平均使用率。
- worker:10.101.5.58_6603的cpu使用率远远高于平均使用率。
- worker:10.101.16.3_6602的cpu使用率远远高于平均使用率。
- worker:10.101.4.186_6602的cpu使用率远远高于平均使用率。
- worker:10.101.7.174_6602的cpu使用率远远高于平均使用率。

解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统

5、Topology运行MEM（2017年8月10日15时4分～2017年8月10日22时59分）



1 10.101.5.44 6603

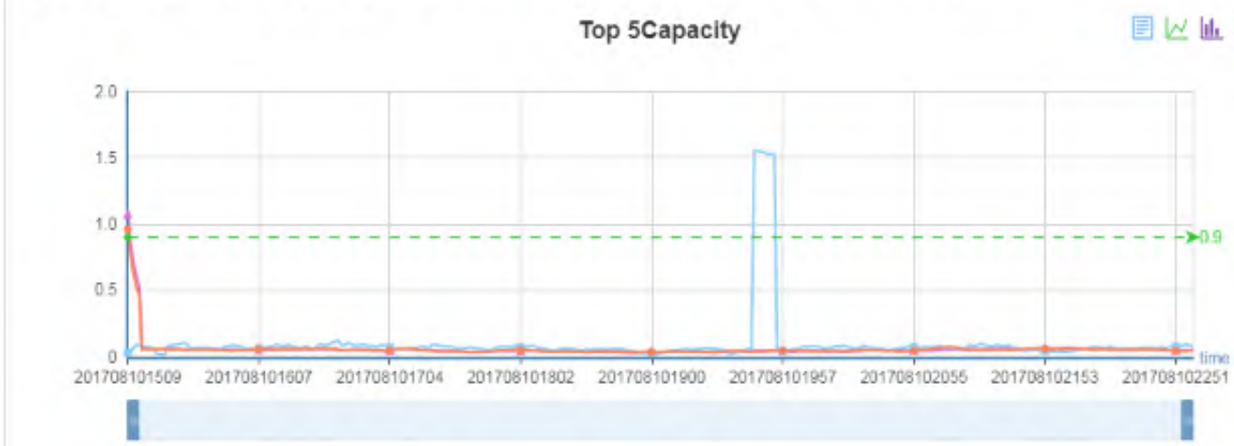
2 10.101.5.71 6603

3 10.101.5.47 6600

4 10.101.5.41 6600

5 10.101.5.67 6601

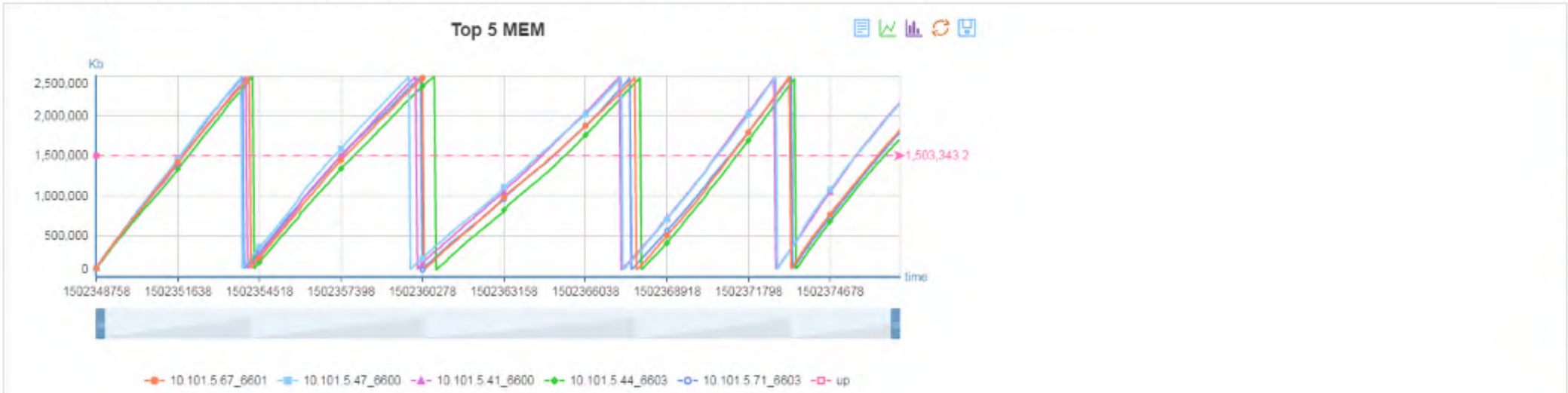
6、Topology运行Capacity（2017年08月10日15时09分～2017年08月10日22时59分）



解决4. 智能诊断、优化建议

流计算平台-健康状态监测系统

5、Topology运行MEM（2017年8月10日15时4分～2017年8月10日22时59分）



1 10.101.5.44 6603

2 10.101.5.71 6603

3 10.101.5.47 6600

4 10.101.5.41 6600

5 10.101.5.67 6601

6、Topology运行Capacity（2017年08月10日15时09分～2017年08月10日22时59分）

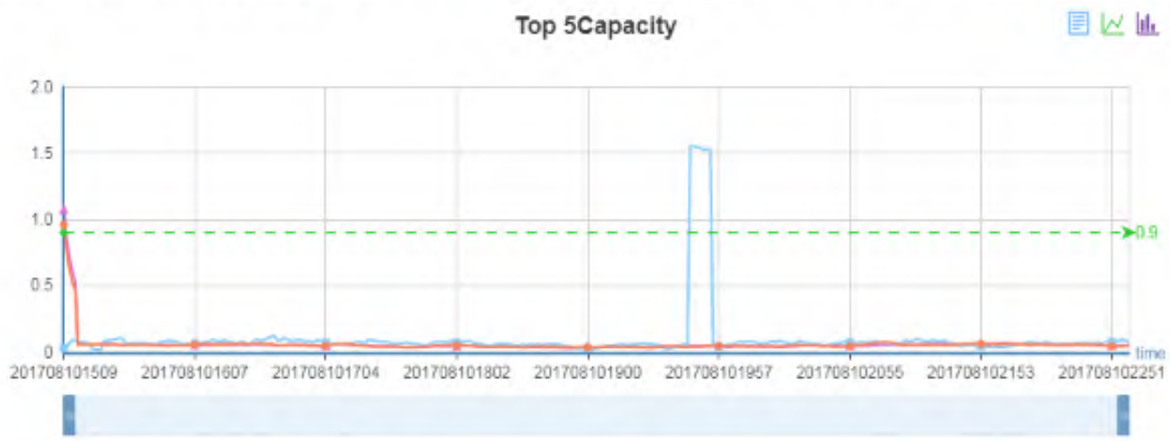


TABLE OF **CONTENTS**

苏宁大数据平台基本介绍

大数据平台运维的痛点及解决方案

平台优化及增强

DOING & TO DO

平台优化及增强-稳定

- Hive metaserver 连接数过高的问题

  修改bonecp的配置：maxConnectionsPerPartition=1

- Spark Streaming & Druid System CPU过高的问题

  设置vm.zone_reclaim_mode=0

- 透明大页导致System CPU过高的问题

  echo never >/sys/kernel/mm/transparent_hugepage/defrag

平台优化及增强-安全

- 账户/权限体系：每个系统一个账户，不允许跨账户写。

- Hive metaserver 密码加密

```
<property>
  <name>javax.jdo.option.ConnectionPassword</name>
  <value>OL2vkw+gDtQ=</value>
</property>
```

- 基于User/IP的访问控制策略：RPC层面控制，白名单

- skipTrash禁用：防止误删数据

平台优化及增强-扩展性

结合HDFS的瓶颈问题逐步优化

- 程序优化，扫全表：Hive慎用unix_timestamp方法
- 小文件合并
- YARN日志降低副本至1
- YARN日志单独放在另一个集群
- Federation + Alluxio 实现统一命名空间

TABLE OF **CONTENTS**

DOING & TO DO

- ✓ Flink推广
- ✓ OLAP平台建设
- ✓ 流计算消息回溯
- ✓ 多活&灾备
- ✓ 资源统一管理（明年）

你认为以下哪些关键字对运维最重要？

智能

易用

稳定

高效

性能

流程

MTTR

自动化

安全

规范

审计

SLA

故障预测

个人号

公众号

THANKS!

智 能 时 代 的 新 运 维

CNUTCon 2017