



GOPS 2016
Shanghai



全球运维大会

2016

重新定义运维

上海站

会议时间： 9月23日-9月24日

会议地点： 上海·雅悦新天地大酒店

主办单位：



指导单位：



阿里大数据计算平台运维实践

范伦挺（萧一）
阿里巴巴-大数据计算



目录



1

大规模计算平台运维挑战

2

自动化平台建设

3

数据驱动精细化运维

4

运维转型思考



阿里大数据架构演进

ODPS平台上线

单集群超5K

异地多活/离在线混布



异地多集群/异地灾备

单集群超10K



大规模计算平台运维挑战

1. 规模大、小概率事件常态化

- 各类硬件故障
- 网络链路不稳定
- 工具容易出问题

2. 多机房多地域

- 延时增加
- 资源不均衡



目录

1 大规模计算平台运维挑战



2 自动化平台建设

3 数据驱动精细化运维

4 运维转型思考



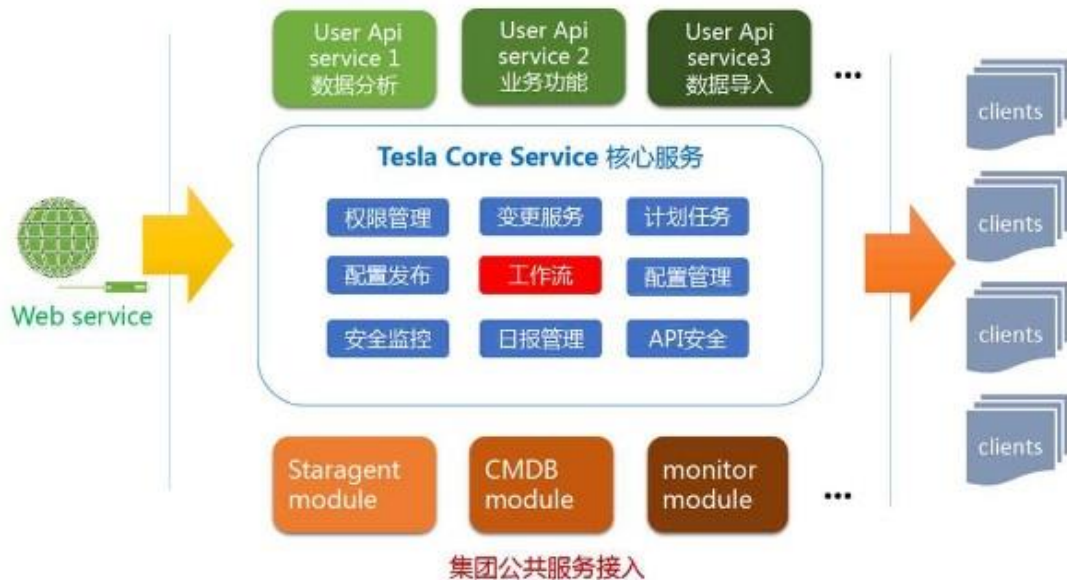
自动化平台建设

1. 自动化意义

- 稳定
- 提高效率


2. 自动化方向


- 变更
- 问题排查
- 硬件维修
- 交付检查



自动化平台建设

1. 变更自动化


 Tesla
Bigdata Design

首页 产品分站 ChangeLine^新 配置中心[▼] 工作流[▼] 工具[▼] 任务管理 


ODPS 工作流 / ODPS [▼]

常用功能

- 配置文件管理
- 拓扑图管理
- 插件脚本管理
- 中控机管理
- BPM工作流管理
- 集群树视图管理
- 全链路





工作流 

- UpdateTunnel2.0
- UpdateNginxConf
- UpdateNginx
- UpdateMetaBlack
- UpdateJreVersion
- UpdateDHFrontend
- Tunnel_smoke
- TunnelUpdate0.2
- TunnelUpdate0.1
- TunnelOnline2.0
- TaskGrayRollback
- TaskGrayRegister

组合工作流 

- TrunkUpdate
- SprintUpdate
- SmallVersionUpdate
- serviceModeRestart
- ReleaseUpdate
- ReleaseTrunkUpdate
- ReleaseSwitch
- PackageServiceUpdate0.1
- OneUpdateAmp
- OdpsSprint
- OdpsSmallVersion
- OdpsHotFix

我的任务

- 在集群()上执行工作流(PackageUpd...
- 在集群()上执行工作流(ServiceUpdat...
- 在集群()上执行工作流(ServiceUpdat...
- 在集群()上执行工作流(ServiceUpdat...



自动化平台建设

1. 变更自动化

28210 (CF编号: 219974)

紧急变更 ODPS frontend 紧急发布申请

当前进度 100%

当前步骤 无

当前状态 详情 >

克隆 变更文档

批量设置自动 批量设置手动

1 准备

100%

提交changefree 通知用户 审批

执行人: [头像] CF审批人: [头像]

2 工作流: ODPS 前端frontend_server升级

状态 执行成功

影响集群 [集群名称]

只可查看

查看执行详情

执行成功

© 2016-08-17 14:34:49



自动化平台建设

1. 变更自



批量设置手动 批量设置自动 ? 批量设置报警方式: 旺旺 短信 电话

第 1 步

自手

执行成功

报警方式: ☒ 旺旺 ☐ 短信 ☐ 电话 超时时间(秒): 900, 剩余自动重试次数: 0, 重试间隔: 0 秒 第 1 次执行, 1秒

命令

发送旺旺(集群 [redacted] 开始执行工作流 FrontendUpdate . 执行人: [redacted] 给用户组: [redacted] 用户列表: 59337

> 查看输出

第 2 步

自手

执行成功

报警方式: ☒ 旺旺 ☐ 短信 ☐ 电话 超时时间(秒): 900, 剩余自动重试次数: 0, 重试间隔: 0 秒 第 1 次执行, 1秒

命令

发送短信(集群 [redacted] 开始执行工作流 FrontendUpdate . 执行人: [redacted] 给用户组: [redacted] 用户列表: 59337

> 查看输出

第 3 步

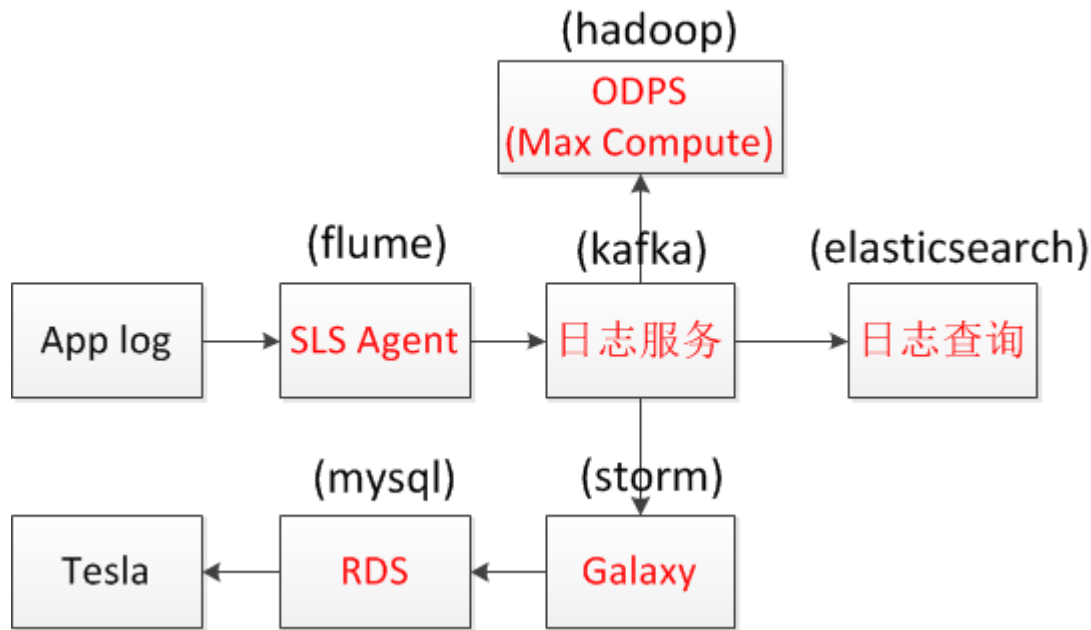
执行成功

报警方式: ☒ 旺旺 ☐ 短信 ☐ 电话 超时时间(秒): 120, 剩余自动重试次数: 0, 重试间隔: 0 秒 第 1 次执行, 1秒



自动化平台建设

2. 问题排查

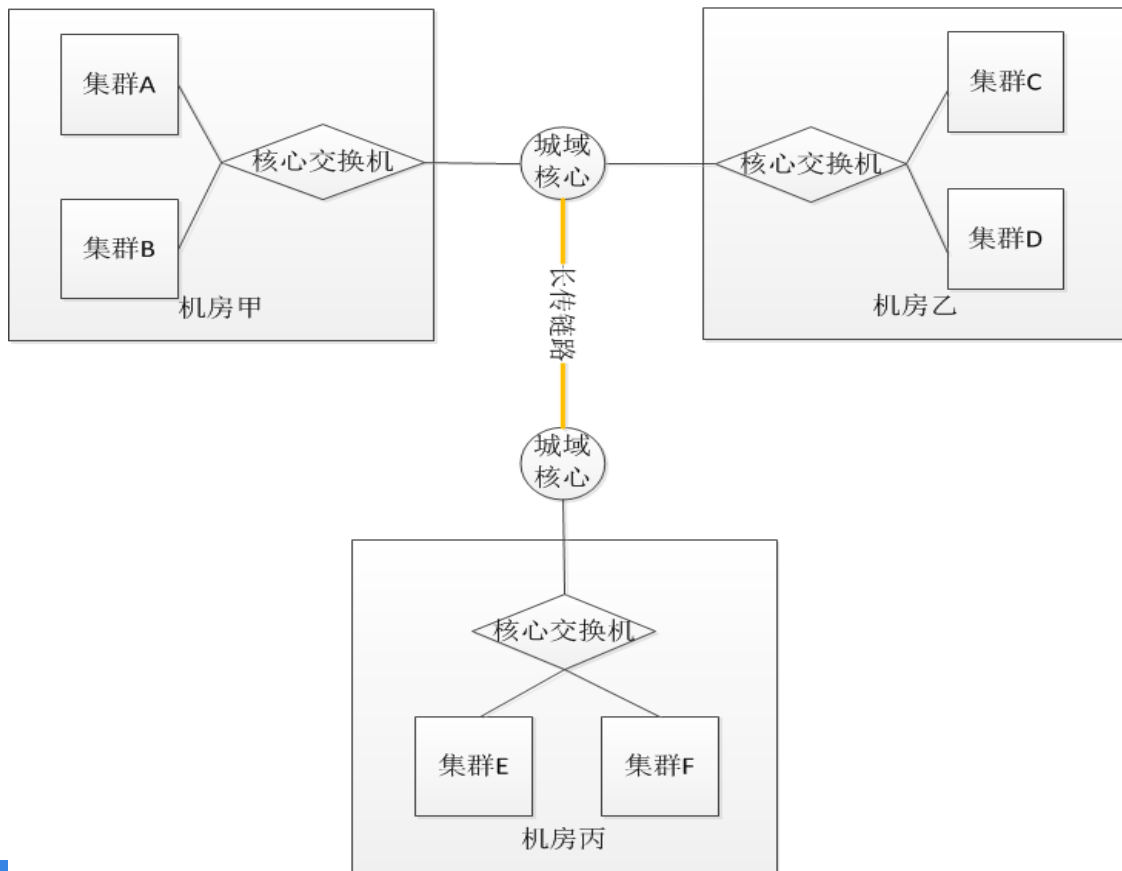


阿里实时日志分析架构



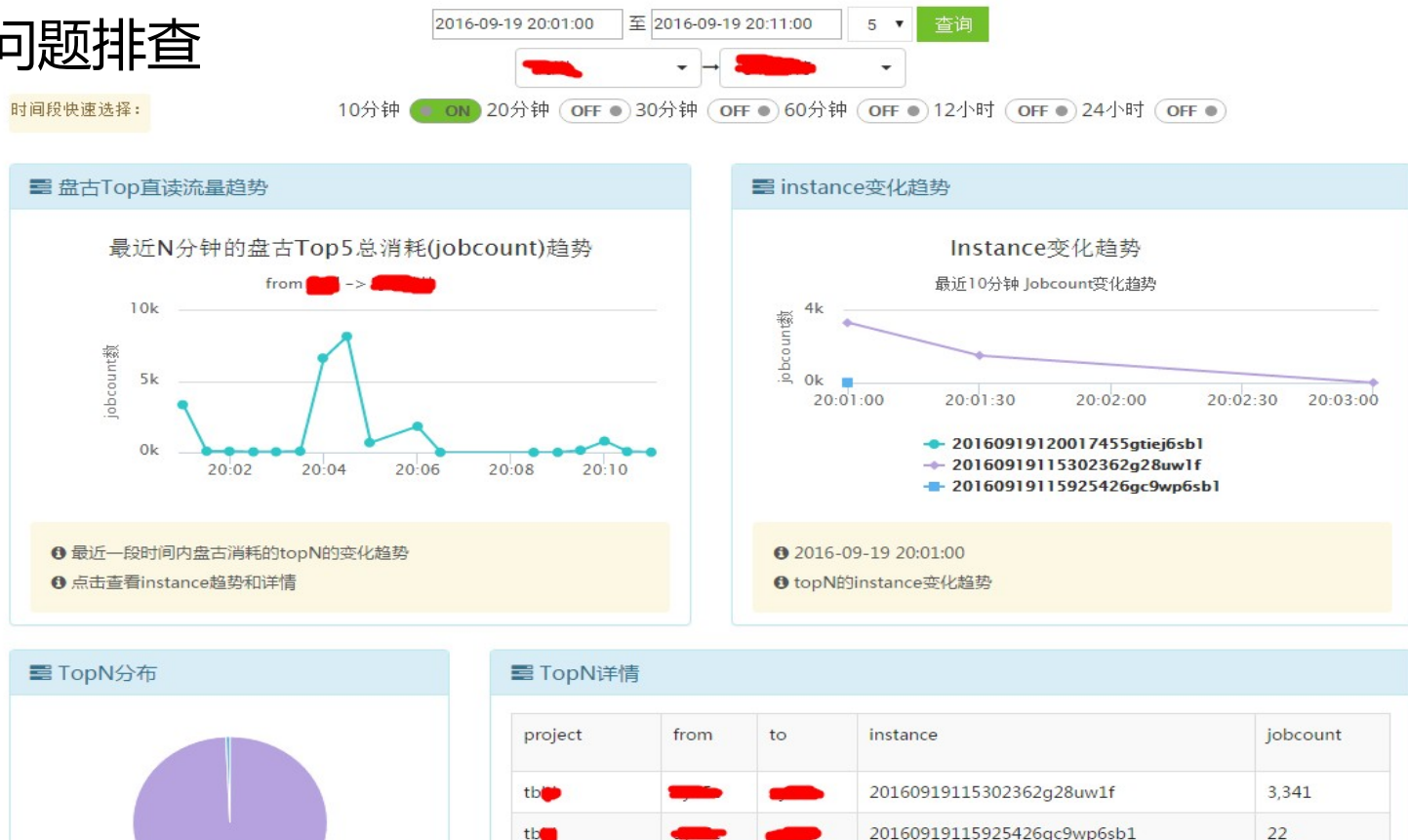
自动化平台建设

2. 问题排查



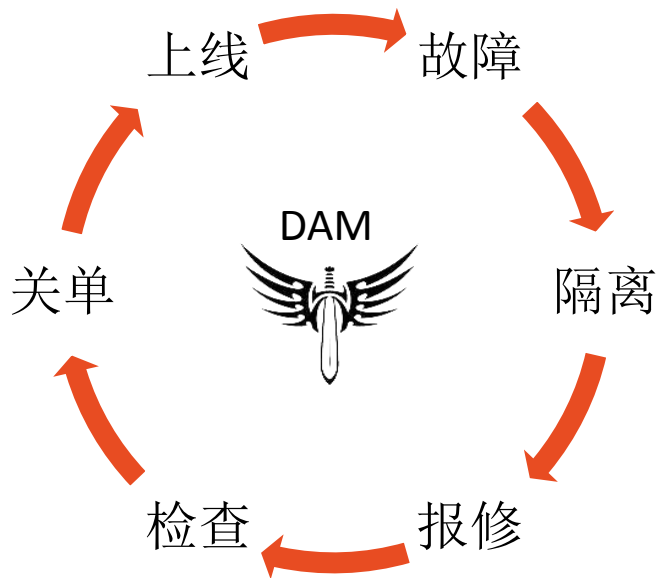
自动化平台建设

2. 问题排查



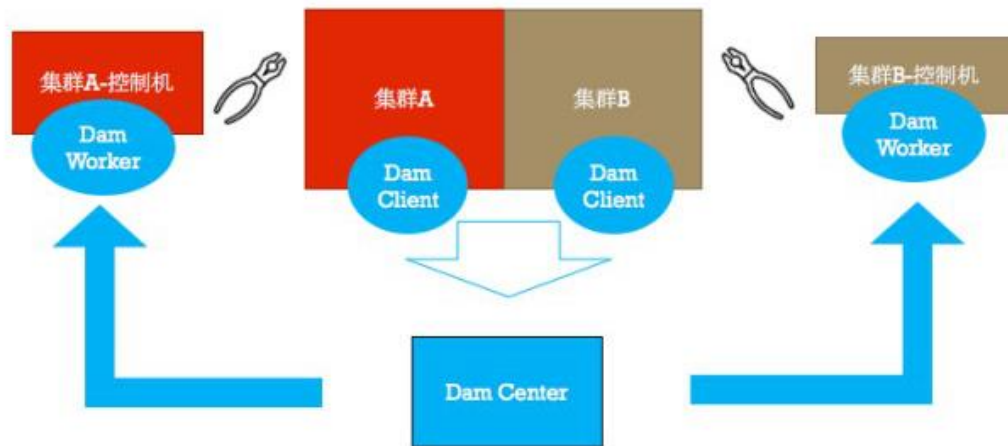
自动化平台建设

3. 硬件维修-DAM



自动化平台建设

3. 硬件维修-DAM



自动化平台建设

3. 硬件维修-DAM

Dam-checker信息来源：

- 硬盘/板卡：kernel log/smartctl/tsar
- 内存：ipmitool/mcelog/stream
- CPU/风扇：mcelog/cpu频率/ipmitool
- 网络/网卡/交换机端口：tsar/kernel log
- 主板：排除以上可能



自动化平台建设

3. 硬件维修-效果

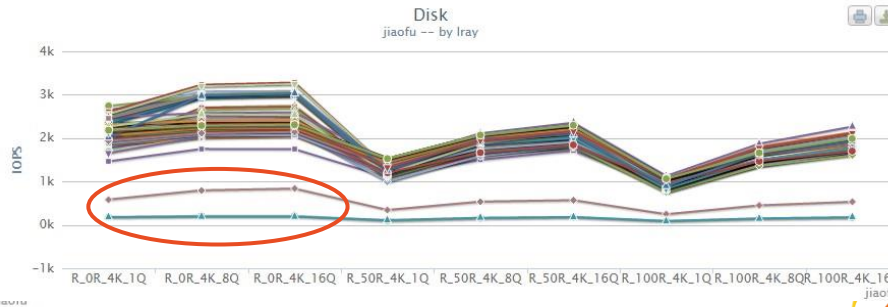
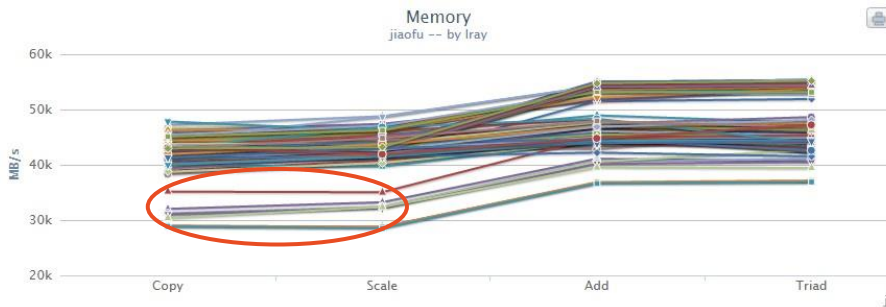
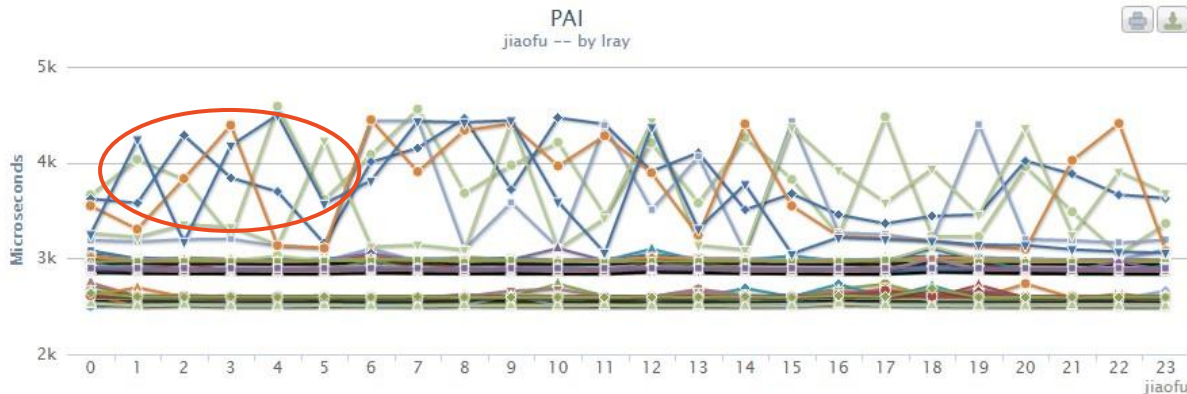
集群	机器数(dam覆盖)	异常工单	7天工单趋势	今天工单	24小时新增工单	1小时内新增	稳定分趋势	稳定分
		2		31		0		97.92
		8		37		0		95.76
		6		37		1		98.46
		8		82		0		97.81
		8		90		2		95.69
		5		115		1		93.43
		76		241		0		95.50



自动化平台建设

4. 交付检查

- 软件交付检查
- 硬件交付检查



目录

1 大规模计算平台运维挑战

2 自动化平台建设

➔ 3 数据驱动精细化运维

4 运维转型思考



数据驱动精细化运维



数据驱动精细化运维

1. Inode优化

以4T单盘为例，ext4默认格式后inode总数为2.4亿，每个inode的大小是256 Bytes，一台机器挂12块磁盘。

Inode占用= $2.4 \times 10^8 \times 256 \times 12 / 1024 / 1024 / 1024 = \mathbf{686.6GB}$

万台集群= $686.6 \times 10000 / 1024 = \mathbf{6.54PB}$

平均文件大小10M计算，单盘所需inode为420000，即使按照10倍计算inode仅需 $42 \times 10^5 / (2.4 \times 10^8) = 1.75\%$ ，至少可以**节省6PB**

注：可优化量根据系统平均文件大小定。



数据驱动精细化运维

2. 网络丢包分析



在生产时段，从整个集群角度看，平均每分钟丢包超过1000个的节点会有200多台，丢包比率达到2%。



数据驱动精细化运维

2. 网络丢包分析



万兆网络环境下单核cpu处理网卡中断能力不足，导致丢包。



数据驱动精细化运维

3. 用户资源分析



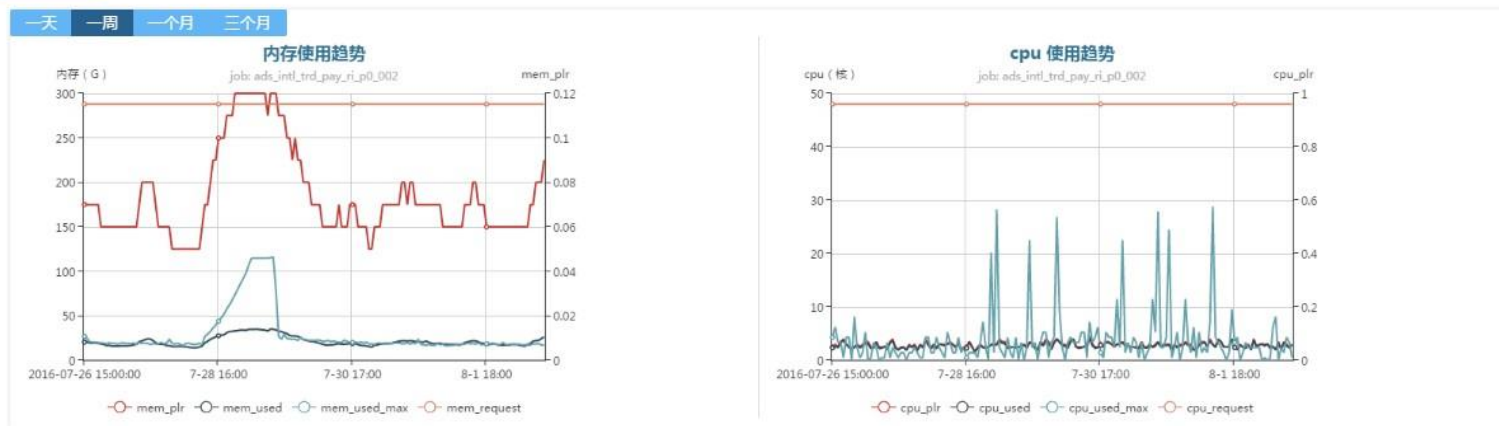
数据驱动精细化运维

3. 用户资源分析

job总体信息

集群		用户		jobid	logview	delay
				ads_intl_trd_pay_ri_p0_002-73-1469189062	logview	0
Worker总数	内存申请(LGV) (M)	内存使用(PUV) (M)	内存使用率(PLR)	cpu申请(LGV) (core)	cpu使用(PUV) (core)	cpu使用率(PLR)
48	294912	26987	0.09	48	2.95	0.06

job使用趋势



目录

1 大规模计算平台运维挑战

2 自动化平台建设

3 数据驱动精细化运维

➔ 4 运维转型思考



运维转型思考

1. 运维向运营转型

运维：稳定、安全 --活着（ “眼前的苟且” ）

运营：服务、效益 --发展（ “诗和远方” ）



运维转型思考

2. 自动化向产品化转型



运维转型思考

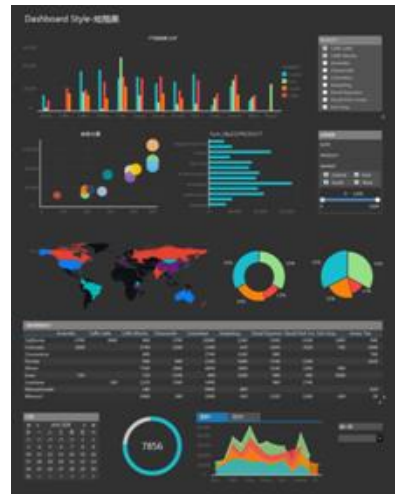
3. 效率向价值转型



喝咖啡？



数据分析



数据化/可视化

运维转型思考





Thanks

范伦挺（萧一），阿里巴巴集团大数据计算与服务
团队负责人，团队主要负责ODPS/ADS/GALAXY/HPC等
大数据计算平台运营服务。

Join us! lunting.fan@alibaba-inc.com

