

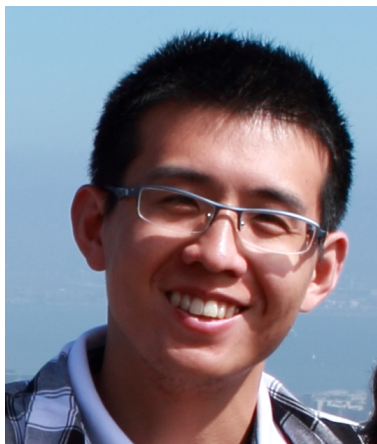
构建数据驱动的动画工作室

--ES及ELK实践

2016.09.10 @ 深圳Vivo



About Me



- 2012.04 硕士毕业于上海华东理工大学
- 2012 – 2014 思科中国研发中心
 - 开发测试（自动化测试和系统测试）
 - 思科认证网络专家 – 路由交换, 网络安全
- 2014年底 某动画制作公司IT部
 - 企业网络规划和维护
 - Linux系统维护
 - DevOps (Python)
- 2015年底开始接触elasticsearch与大数据技术

IT与数字运营

- 基础设施：
 - 200台+Intel Xeon工作站
 - 300台+堆叠式及刀片服务器
 - 物理+虚拟化混合环境
 - 独立运营的数据中心
- 应用与服务
 - Kerberos统一认证登录
 - 渲染农场
 - 分布式存储系统
 - Puppet自动化配置管理

运营挑战

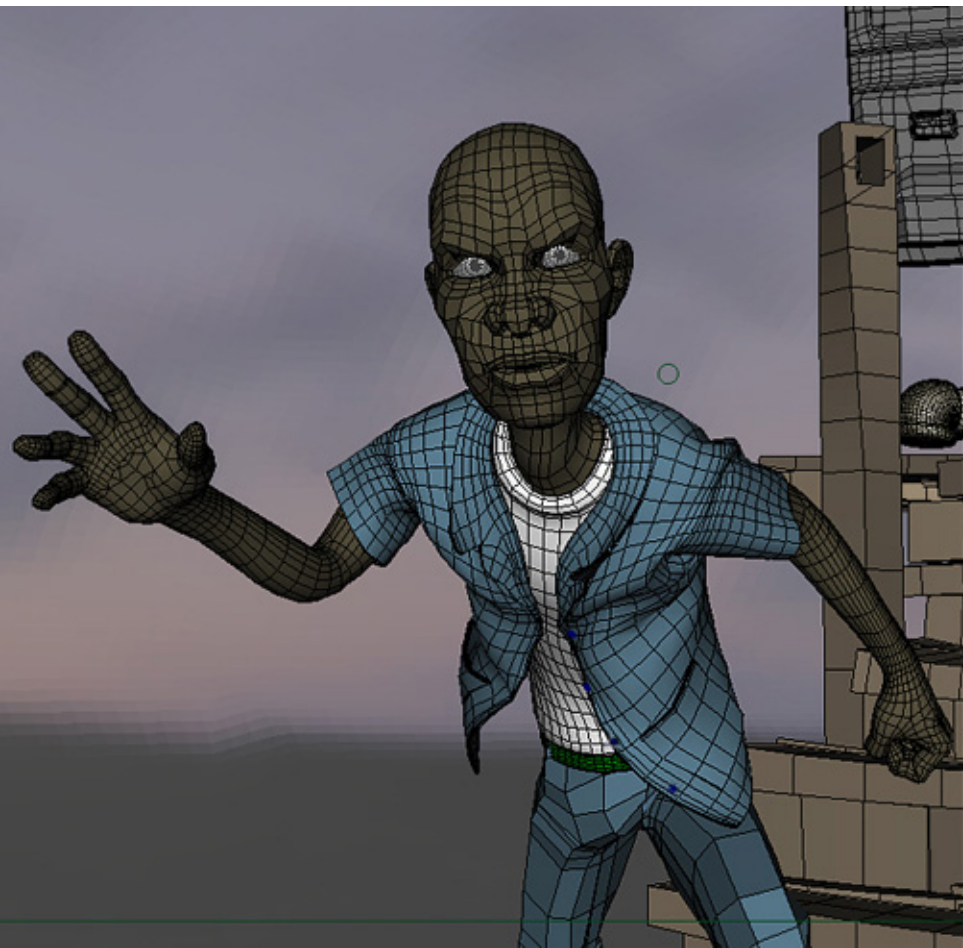
系统服务

峰/谷能耗管理

渲染农场
状态与效率

使用Elasticsearch增加渲染农场可见度





为什么需要渲染农场



电影成片

$100 \text{ 分钟} * 60 * 24 = 144000 \text{ 帧} <- 1920 \times 1080 \text{ 像素}$
 $= 2986 \text{ 亿像素}$

单机渲染速度：6小时/帧

服务器指标：Xeon E5v3 2.5GHz (Haswell) + 128G

单机渲染完成： $144000 * 6 = 864000 \text{ 小时} \approx 100 \text{ 年}$

幕后制作

工作量 x 3

So, 100台机器需要连续3年满负荷运转

关于渲染农场



- 多机多进程任务调度与处理系统
- 计算资源使用极度密集
- 整合生产流水和创作软件

几个问题

- 前一天都做了哪些场景的画面渲染？
- 每个场景有多少帧？
- 各个制作部门都提交了多少渲染任务？
- 他们的平均内存占用量有多大？
- 在这些任务里极清帧(4K)的平均每帧渲染时间和效率？

毛估估一下？ 可是...

- 每天处理的任务量：8千 - 9万
- 每个任务所需时间：10分钟 - 8小时
- 涉及到的场景及镜头数：20 - 40个

Elasticsearch

- 定义数据结构模型
- 搭建ES基础设施
- Index和mapping设定
- Go!

数据模型，数据模型！

- Nested field? ✓ - Kibana兼容
- Parent/children?
- Mapping属性：
 - 数字/字符串
 - 是否需要解析

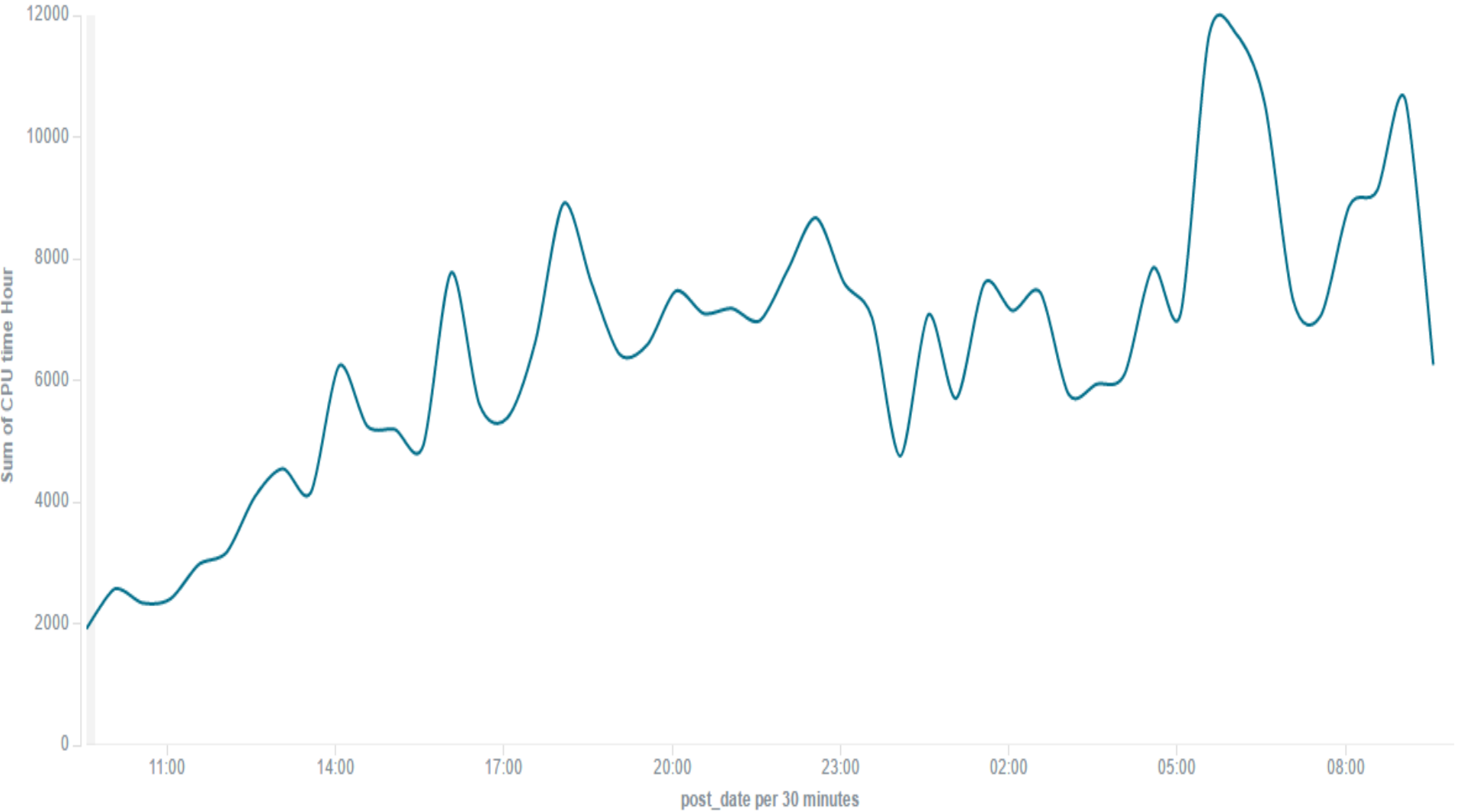
示例：
灯光组
农场CPU使用

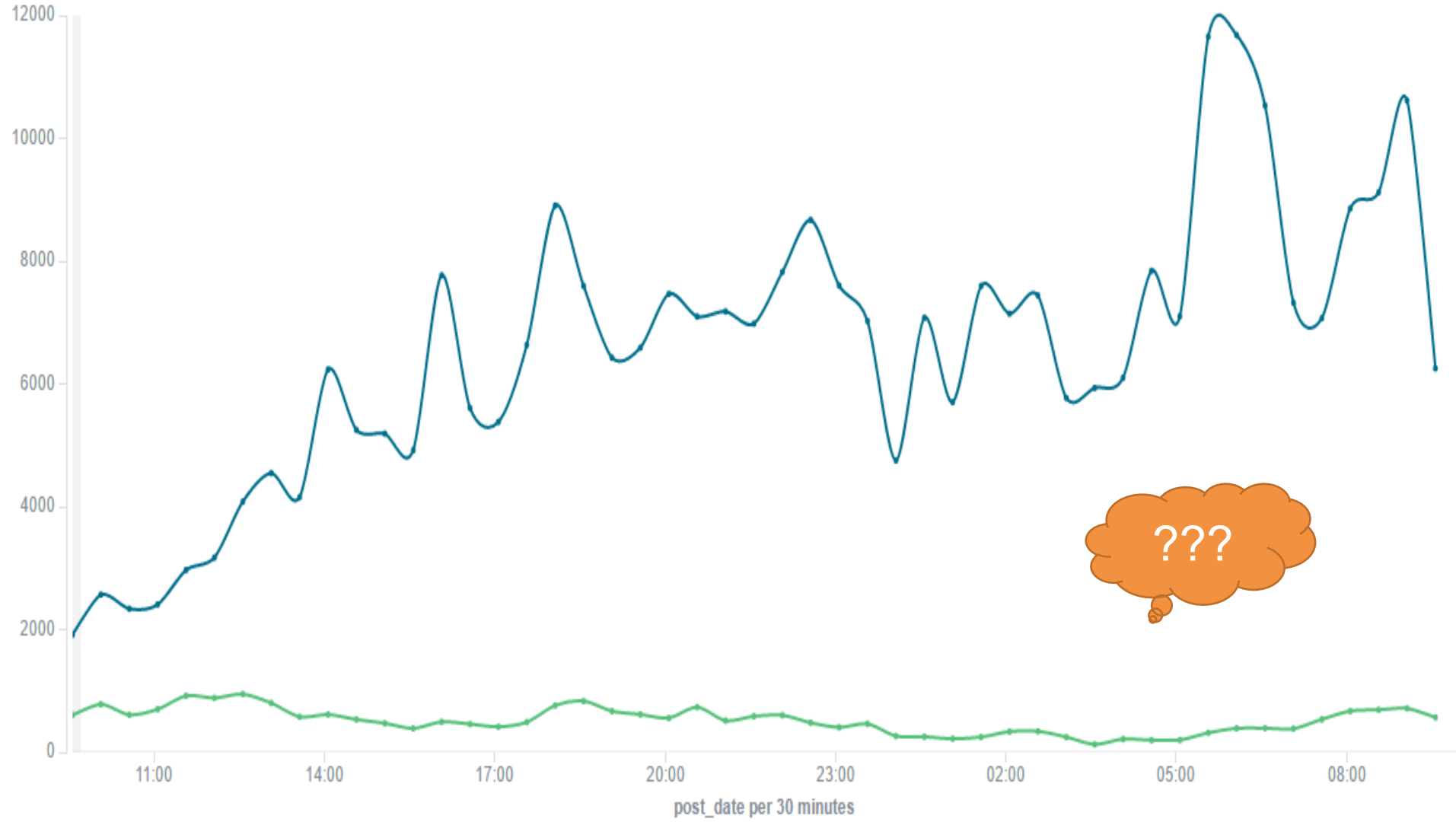
Kibana
Lucene
search:
prod_dept:lgt

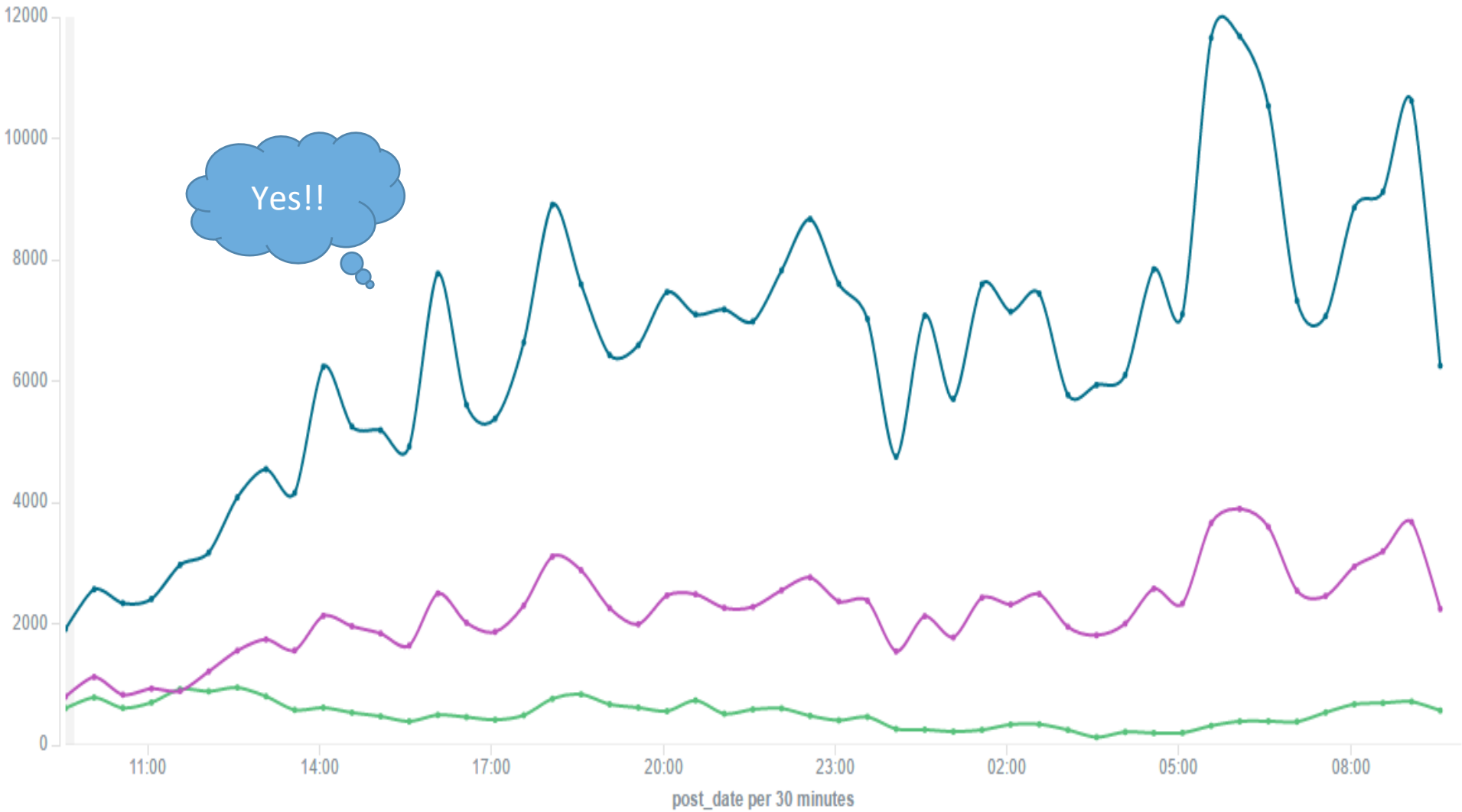
Save Search
As:
dept_lighting_all

In Visualize:

From saved search: dept_lighting_all
Bucket -> Date Histogram <date field>
Metrics -> Sum <CPU Time Hour>







构造Dashboard: 24小时summary

创建图表:

部门任务数
量柱状图
提交任务总
数时间线图
各场景任务
数量的时间
线图

... ..

调整图表布局

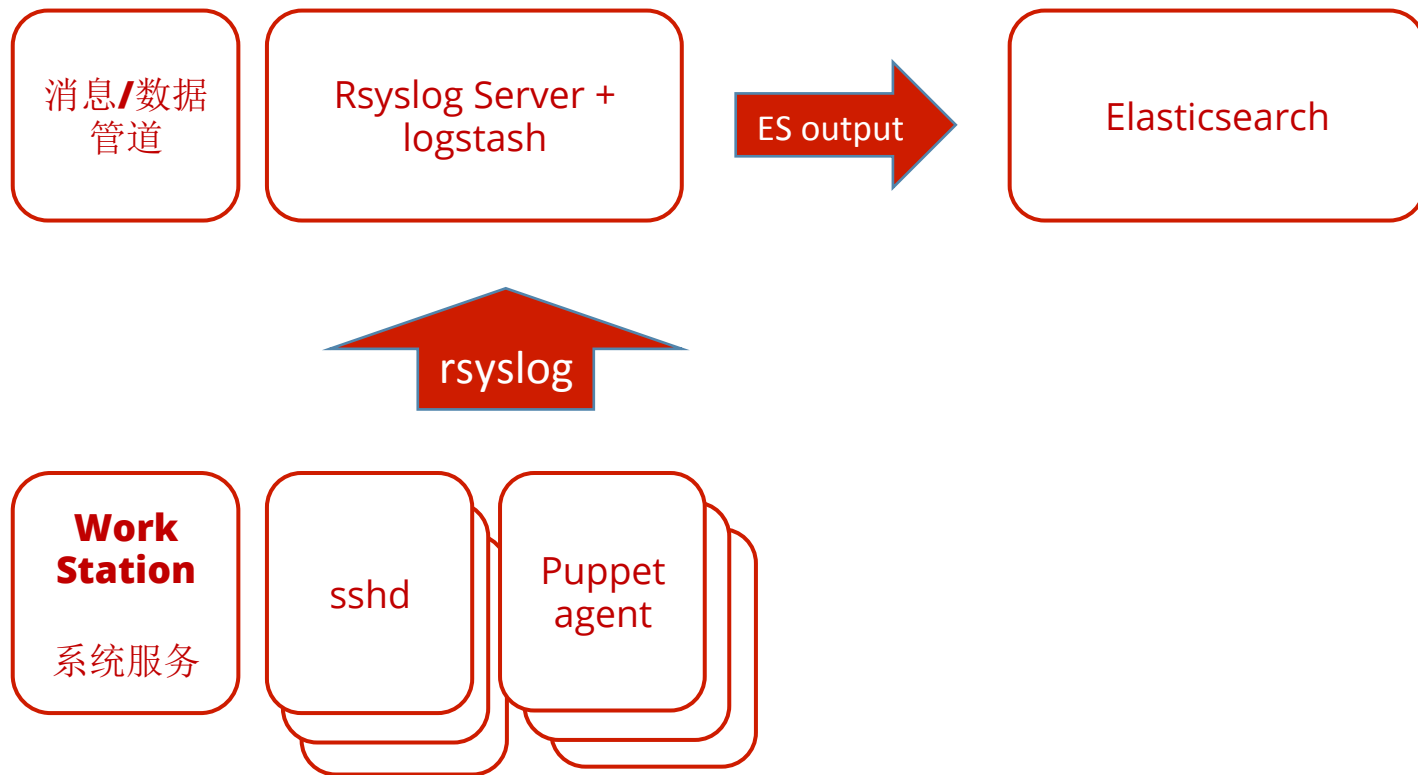
调整右上角time range为Last 24h

将Dashboard保存并勾选“Store time
with dashboard”

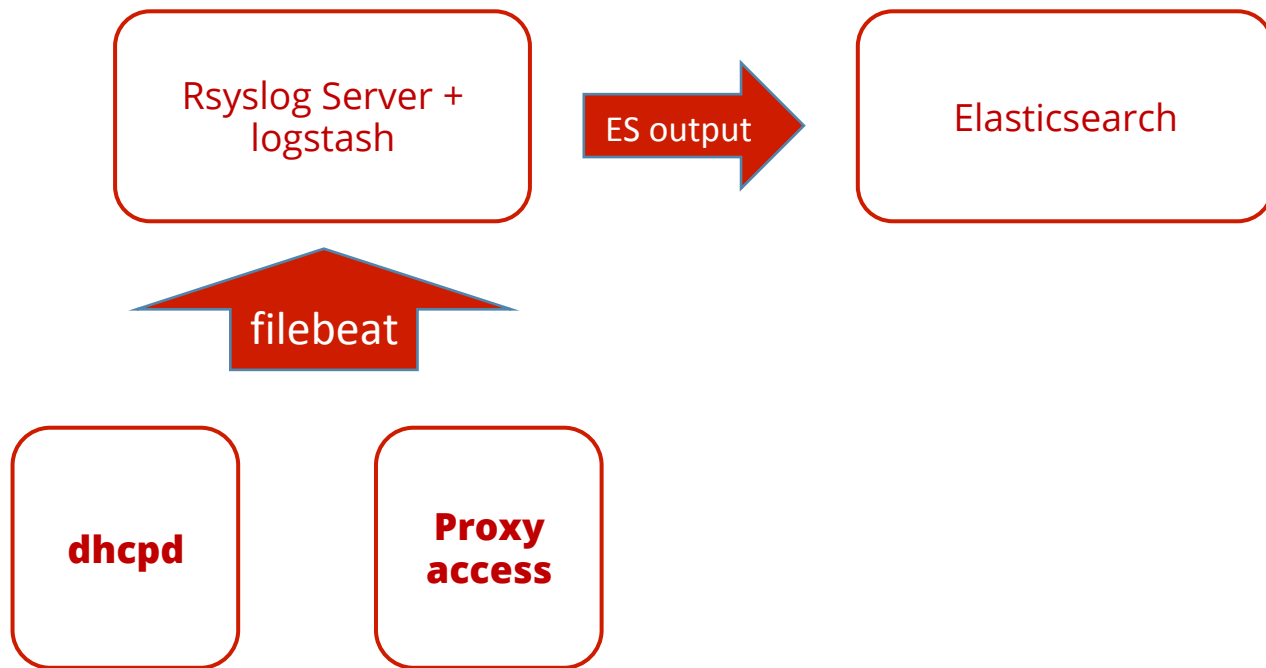
创建Dashboard: 24h_farm_summary

使用ELK提高IT运维效率

logstash + rsyslog remote logging



零星主机：filebeat/lumberjack



Let's Grok

- 为每个应用服务编写各自的grok filter
- 使用grokconstructor.appspot.com测试grok匹配
- Grok filter config版本控制

Sep 2 14:35:27 soma dhcpd: DHCPACK on ~~10.0.2.15~~ to c0:ee:fb:d9:35:18 (android-1b4bb90ad2557583) via ~~10.0.2.15~~

MATCHED	
device_name	android-1b4bb90ad2557583
gateway	10.0.2.1
dhcp_type	DHCPACK
client_mac	c0:ee:fb:d9:35:18
client_ip	10.0.2.15
before match:	Sep 2 14:35:27 10.0.2.15 -dhcpd:

使用机器人 处理机器数据

- 哪些设备在过去半小时内发送了最多的dhcp request? ✓
- 过去半小时内出现puppet配置应用错误的主机有哪些? ✓
- 哪些设备昨天收到了最多的ssh登陆尝试? ✓

Kibana + python

- 在ES中配置grok输出项的mapping属性并使用Kibana可视化
- 通过Python调用ES api实现基于ES的分析功能
- api的返回结果可供已有的IT监控系统使用

Elasticsearch

Pros vs Cons

- 平坦的学习曲线
- 开箱即用
- 无master的集群模式
- REST兼容
- 原生json
- 全文索引/Schemaless
- 无法删改field mapping
- 不支持列变换
- 无ACID
- BI工具支持不足

场景总结

- 数据结构要有良好的文档记录
- Data节点与client节点分离，data节点不处理REST请求
- 收集、协商公司内数据需求
- 定期备份，滚动升级
- 测试发掘更多的特性和插件

Thanks!



Elastic中文社区技术沙龙【深圳站】

Elastic中文社区技术沙龙【深圳站】



<http://elastic.co>

<http://elasticsearch.cn>

<http://elasticsearch.cn/article/99>

<http://www.vivo.com.cn>

<http://dev.vivo.com.cn>

<http://hr.vivo.com.cn>

<http://www.vivo.com>

