

# 芒果TV ELK日志系统实践



# 关于我

刘波涛

芒果TV研发工程师

# 日志文件重要性

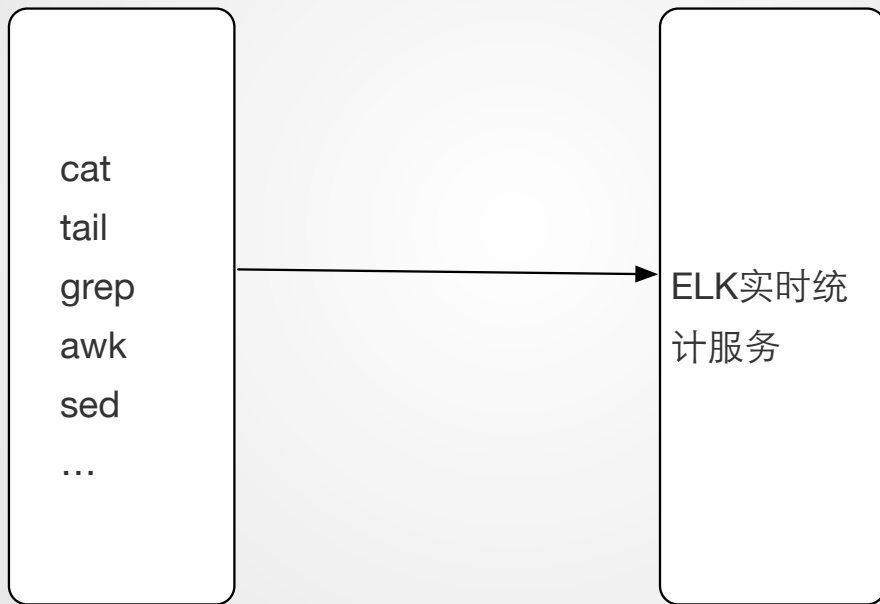
Linux哲学: 万物皆文件

运维哲学: 日志管理是保障高质量服务的基础

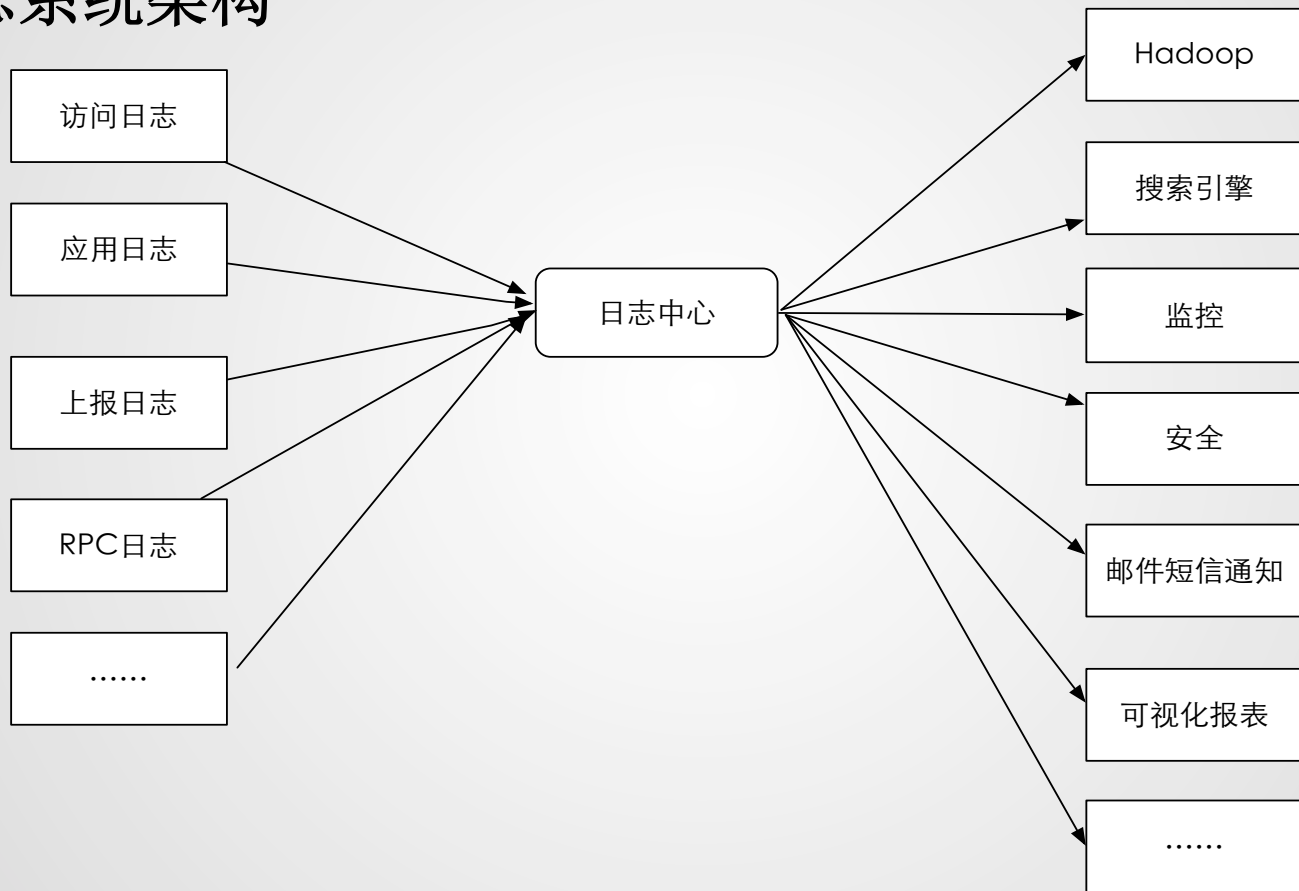
# 日志架构演变

专业运维人员(强大正则功底)

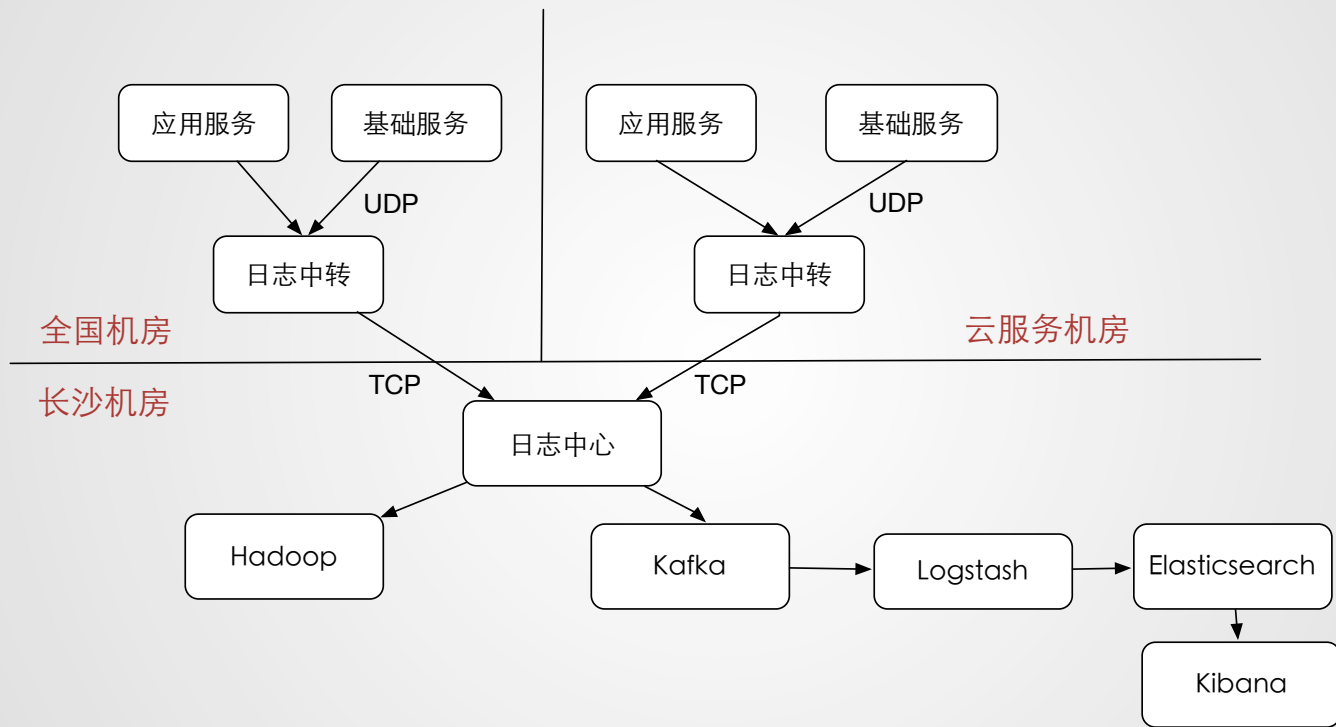
运维人员、开发人员(简单操作)



# 日志系统架构



# ELK系统架构



# Rsyslog

## Nginx: 通过syslog模块转发

应用程序：通过调用syslog函数写入

## 规范统一格式

## v8版本以上(原生支持Kafka)

```
module(load="omkafka")
local5.*      action(type="omkafka" topic="mpp-elk" broker="10.200.0.127:9091,10.200.0.127:9092,10.200.0.127:9093,10.200.0.127:9094,10.200.0.126:9091,10.200.0.126:9092,10.200.0.126:9093,10.200.0.126:9094,10.200.0.126:9095,10.200.0.126:9096,10.200.0.126:9097,10.200.0.126:9098" partitions.number="14" confParam=["queue.buffering.max.messages=2000000"] queue.size="360000000" queue.maxdiskspace="99G" queue.highwatermark="216000000" queue.discardmark="288000000" queue.type="LinkedList" queue.dequeuebatchsize="4096" queue.timeoutenqueue="0" queue.maxfilesize="4G" queue.saveonshutdown="on" queue.workerThreads="14" template="access")
```

# Rsyslog-Avoid-Block

关闭HUPIsRestart配置选项(低版本)

监控rsyslog服务,一旦crashes能够马上重启

传输方式由TCP改为UDP(恶性循环)



# Kafka

Kafka vs Redis

强大消息堆积能力

日志领域高度成熟

支持Hadoop数据并行加载

高性能(顺序写单机写入TPS约在百万秒/s)

# Kafka-Options

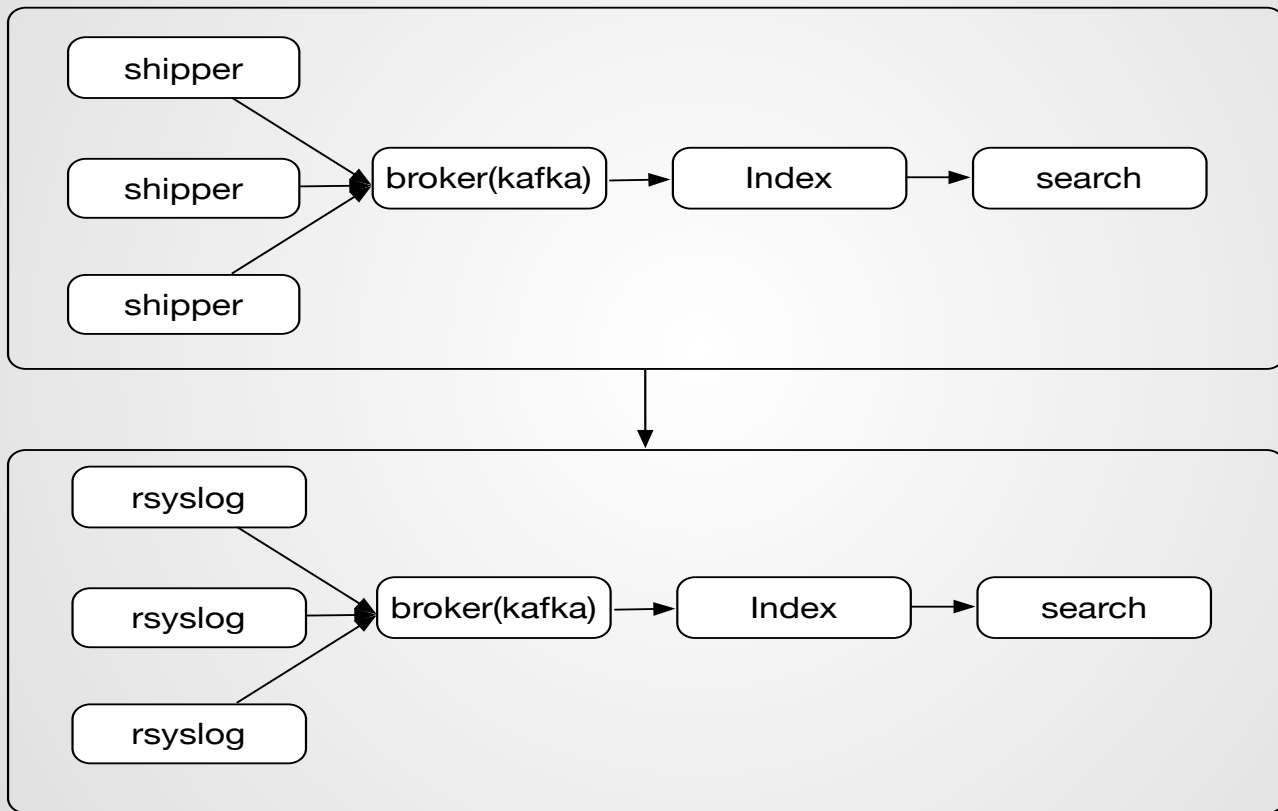
Partition Number(数量必须大于消费者数量)

Broker Number(配置和内核数相同)

num.network.threads

num.io.threads

# Logstash

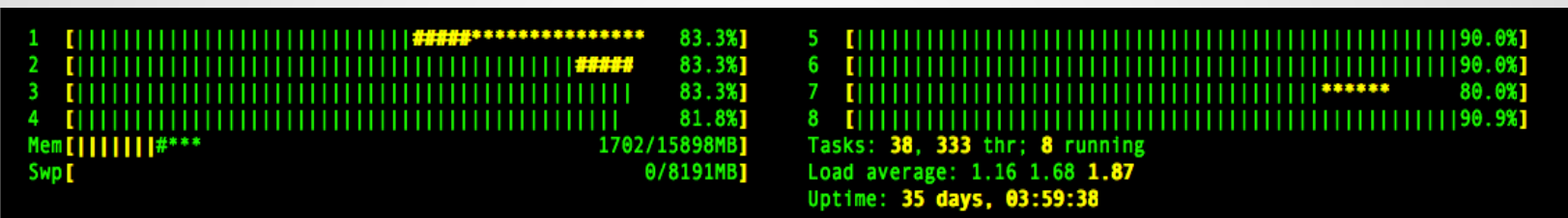


# Logstash

配置

```
[redacted@redacted conf.d]# ls  
redacted-input-kafka.conf redacted-filter-nginx-mpp.conf redacted-output-elastic.conf
```

性能问题,大量消耗CPU和内存



容易僵死

自定义JAVA程序替代Logstash(支持kafka,syslog输入,ES输出)

启动多个进程进行消费

# Elasticsearch

减少副本数量(副本数为0)

以写为主,读为辅助(随机写 磁盘瓶颈 使用SSD替代传统硬盘)

设置`filedldata: format :doc_value` 避免Heap crash

增加`Index.refresh_interval` 时间(默认为一秒),降低压力

合理使用TCP,UDP索引模式(我们使用Http模式)

关闭系统swap

内核配置修改

对数据聚合进行处理`string2int`

定时删除旧索引(保存2个星期)

# 服务器参数调整

## TCP参数

```
net.ipv4.tcp_fin_timeout = 30
net.ipv4.tcp_keepalive_time = 1200
net.ipv4.tcp_syncookies = 1
net.ipv4.tcp_tw_reuse = 1
net.ipv4.tcp_tw_recycle = 1
net.ipv4.ip_local_port_range = 1024 65000
net.ipv4.tcp_max_syn_baklog = 8192
net.ipv4.tcp_max_tw_buckets = 5000
```

## 调整Linux的最大文件数

```
ulimit -SHn 65535
```

# 磁盘

## iostat

```
[root@localhost ~]# iostat -x 1 5
Linux 2.6.32-358.el6.x86_64 (localhost.localdomain)    2015年10月15日  _x86_64_    (24 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           9.45    0.00    0.54    0.05    0.00   89.95

Device:            rrqm/s   wrqm/s     r/s     w/s    rsec/s    wsec/s  avgrq-sz  avgqu-sz   await  svctm   %util
sda                 4.90     5.52     1.35     0.77     52.28     50.31     48.29     0.00     2.07   0.78   0.17
sdb                 0.02    2224.07    20.03    122.72    1703.81    18730.76    143.15     0.79     5.52   0.16   2.25
```

如果util接近100%则说明产生的I/O请求太多,I/O系统已经满负载

磁盘可能存在瓶颈如果idle小于70%,I/O的压力比较大,说明读取进程中有较多的wait

## vmstat

```
[root@localhost ~]# vmstat 2
procs -----memory----- ---swap-- -----io----- --system-- -----cpu-----
 r  b   swpd   free   buff   cache   si   so    bi   bo   in   cs  us  sy  id  wa  st
12  0  4267028 558908 147512 32044600    1    1   37   392    1    1  9  1 90  0  0
 9  0  4267028 531364 147512 32073924    0    0    0    0 6888 5396 13  1 86  0  0
 1  0  4267028 529716 147512 32081780    0    0    0  9634 7645 5906 13  1 86  0  0
 0  0  4267028 511768 147528 32100348    0    0    0  7428 6867 5571 11  1 88  0  0
 5  0  4267028 489936 147528 32120188    0    0    0 11544 8170 5873 13  1 86  0  0
```

vmstat 2 查看b堵塞进程情况

# 磁盘

## 检查磁盘性能

```
# fio -filename=/dev/sda3 -  
direct=1 -iodepth 1 -thread  
-rw=randrw -ioengine=psync  
-bs=1k -size=1G -numjobs=10  
-runtime=120 -  
group_reporting -  
name=mytest
```

```
...-]# fio -filename=/dev/sda3 -direct=1 -iodepth 1 -thread -rw=randrw -ioengine=psync  
-bs=1k -size=1G -numjobs=10 -runtime=120 -group_reporting -name=mytest  
mytest: (g=0): rw=randrw, bs=1K-1K/1K-1K, ioengine=psync, iodepth=1  
...  
mytest: (g=0): rw=randrw, bs=1K-1K/1K-1K, ioengine=psync, iodepth=1  
fio 2.0.7  
Starting 10 threads  
Jobs: 10 (f=8): [#####] [0.0% done] [153K/142K /s] [150 /139 iops] [eta 1158033717d:17h:47  
Jobs: 10 (f=10): [#####] [3.3% done] [166K/161K /s] [163 /158 iops] [eta 01m:59s]  
Jobs: 1 (f=1): [#####] [10.2% done] [130K/150K /s] [127 /147 iops] [eta 18m:00s]  
mytest: (groupid=0, jobs=10): err= 0: pid=24227  
read: io=23387KB, bw=199396 B/s, iops=194, runt=120104msec  
clat (usec): min=89, max=340649, avg=16451.24, stdev=16074.47  
lat (usec): min=90, max=340649, avg=16451.52, stdev=16074.47  
clat percentiles (usec):  
| 1.00th=[ 1720], 5.00th=[ 2672], 10.00th=[ 3760], 20.00th=[ 5728],  
| 30.00th=[ 7648], 40.00th=[ 9536], 50.00th=[12480], 60.00th=[15680],  
| 70.00th=[19072], 80.00th=[23936], 90.00th=[32640], 95.00th=[42240],  
| 99.00th=[73216], 99.50th=[100864], 99.90th=[168960], 99.95th=[191488],  
| 99.99th=[259072]  
bw (KB/s) : min= 0, max= 49, per=9.84%, avg=19.10, stdev= 7.38  
write: io=23440KB, bw=199848 B/s, iops=195, runt=120104msec  
clat (msec): min=2, max=308, avg=34.49, stdev=27.43  
lat (msec): min=2, max=308, avg=34.49, stdev=27.43  
clat percentiles (msec):  
| 1.00th=[ 5], 5.00th=[ 8], 10.00th=[ 10], 20.00th=[ 15],  
| 30.00th=[ 19], 40.00th=[ 23], 50.00th=[ 28], 60.00th=[ 33],  
| 70.00th=[ 40], 80.00th=[ 50], 90.00th=[ 68], 95.00th=[ 86],  
| 99.00th=[ 141], 99.50th=[ 167], 99.90th=[ 215], 99.95th=[ 235],  
| 99.99th=[ 277]  
bw (KB/s) : min= 0, max= 51, per=9.80%, avg=19.10, stdev= 5.31  
lat (usec) : 100=0.01%, 250=0.02%, 500=0.01%, 750=0.01%  
lat (msec) : 2=1.00%, 4=4.86%, 10=20.61%, 20=26.47%, 50=35.77%  
lat (msec) : 100=9.38%, 250=1.86%, 500=0.03%  
cpu : usr=0.92%, sys=1.22%, ctx=588804, majf=0, minf=1295891  
IO depths : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%  
submit : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%  
complete : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%  
issued : total=r=23387/w=23440/d=0, short=r=0/w=0/d=0  
  
Run status group 0 (all jobs):  
READ: io=23387KB, aggrb=194KB/s, minb=194KB/s, maxb=194KB/s, mint=120104msec, maxt=120104msec
```



# Elasticsearch

Mapping:压缩\_source(compress:true)

Mapping:禁用all(include\_in\_all:false)

自定义分词:尽量不使用标准分词使用ik

ES\_HEAP\_SIZE: -Xms = -Xmx 不超过内存50%

index.cache.field.type = soft

index.cache.field.max\_size:50000

index.cache.field.expire:10m

index.fielddata.cache: soft

# Elasticsearch SSD优化参数

磁盘RAID0

mmap索引文件格式 (index.store.type: mmapfs)

indices.store.throttle.type:none

indices.memory.index\_buffer\_size: 30%

index.merge.scheduler.max\_merge\_count: 6

index.translog.flush\_threshold\_size:5gb

index.translog.flush\_threshold\_ops: 500000

index.gateway.local.sync:30s

index.merge.scheduler.max\_thread\_count: 3

关闭文件系统ATIME (atime off)

# Kibana

K3 VS K4

原生 VS 自定义可视化

# 监控报警

Nginx 5xx/s

Mysql 慢日志、错误日志

Redis 慢日志、错误日志

程序错误日志

DNS劫持

# 统计报表

Nginx 响应时间

Nginx 正常响应占比

Nginx QPS统计

CDN视频流加载时间

# 搜索引擎

生成热门搜索

统计搜索转化率

调整搜索权重

感谢关注和支持

