Sogou 搜狗

GOPS2018
Shenzhen

# 用户视角下的
# 搜狗大数据平台建设

搜狗 | 技术工程部- 冯晋

# 个人介绍



- 在搜狗工作超过 10年
- 负责过 搜索 、运维 、云平台 、大数据 等
  多个方向的产品和技术研发，参与过搜狗早期多个技术项目的落地工作
- 目前团队主要负责公司 大数据基础平台建设 、数据的管理和应用 等方向

# 目录

GOPS2018
Shenzhen


1. 搜狗大数据业务概况

2. 搜狗基础运维平台简介

3. 搜狗大数据产品化实践

# 搜狗大数据概况



**搜狗是一家典型的大数据公司**

**大规模** 搜索引擎是典型的大数据系统，搜索数据量**2000亿+**

**高并发** 输入法行业第一，DAU用户规模**4亿+**

# 行业大数据方向演化



## 大数据系统演化时间线

# 搜狗大数据历史

GOPS2018
Shenzhen

## 1.0时代

专用的搜索大数据时代

2004

2009

# 2.0时代

## 行业接轨/大规模应用时代

2009

2016

# 3.0时代

## AI驱动/产品商业化
## 大数据时代

2016

未来

# 搜狗大数据整体架构


数据应用
数据报表
离线计算
数据工程师
交互式查询
BI分析
SQL类用户
模型训练
广告投放
研发工程师
数据计算
离线分析
Hive Spark MR
即席查询 / 多维分析
Presto SparkSQL Kylin
实时计算引擎
Flink Spark Streaming
数据存储
Hadoop
Kafka
Hbase
数据采集Agent
Sogou-Storage（批处理）
tail2kafka （实时）
数据源
广东IDC
北京IDC
互联网数据
服务监控报警平台
资源调度管理平台

# 目录

1. 搜狗大数据业务概况

2. 搜狗基础运维平台简介

3. 搜狗大数据产品化实践

# 搜狗基础运维平台简介



## 资源管理中心（ROC）

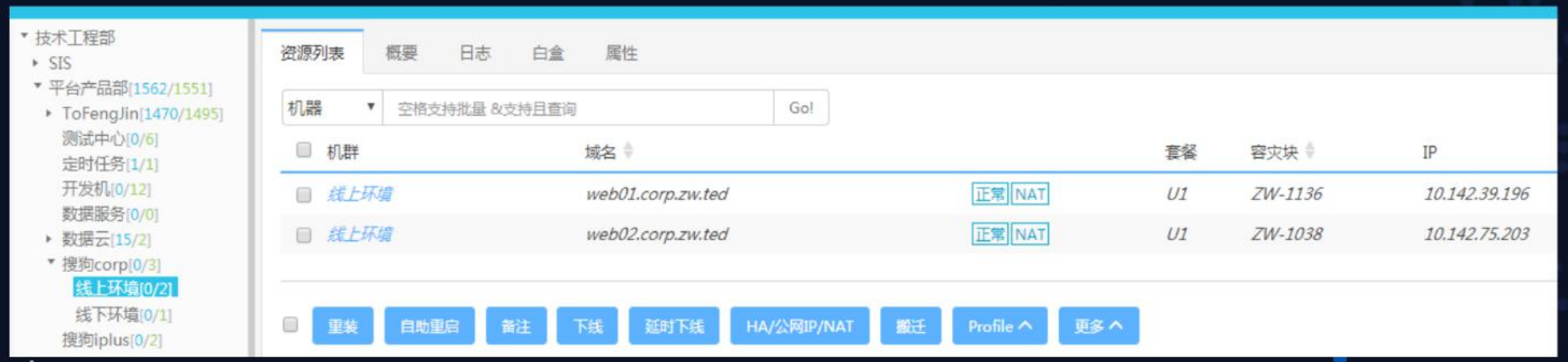
- 开源->自研
- 分散平台->集中展示
- 更广义的运维平台

# 搜狗基础运维平台简介



## 关于机器的管理

- 从机器到机群
- 机群叶子为最小模块单位
- 实现了所有日常操作组件

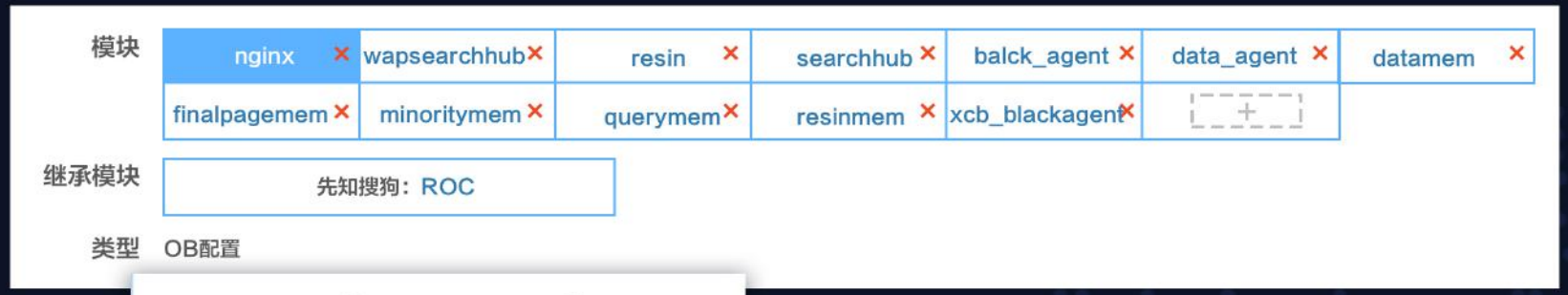


# 运维平台简介



## 进程与日志管理

- 模块管理/Sogou-Observer
- 日志管理/Sogou-Storage




```
OB_PREFIX="djt-pb-log"
OB_PROC_NAME="nginx"
OB_BASE_DIR="/search/nginx"
OB_APP_MODE="nginx"
OB_USER="root"
OB_PID_FILE="nginx.pid"
OB_LOG_LIMIT="128"
```

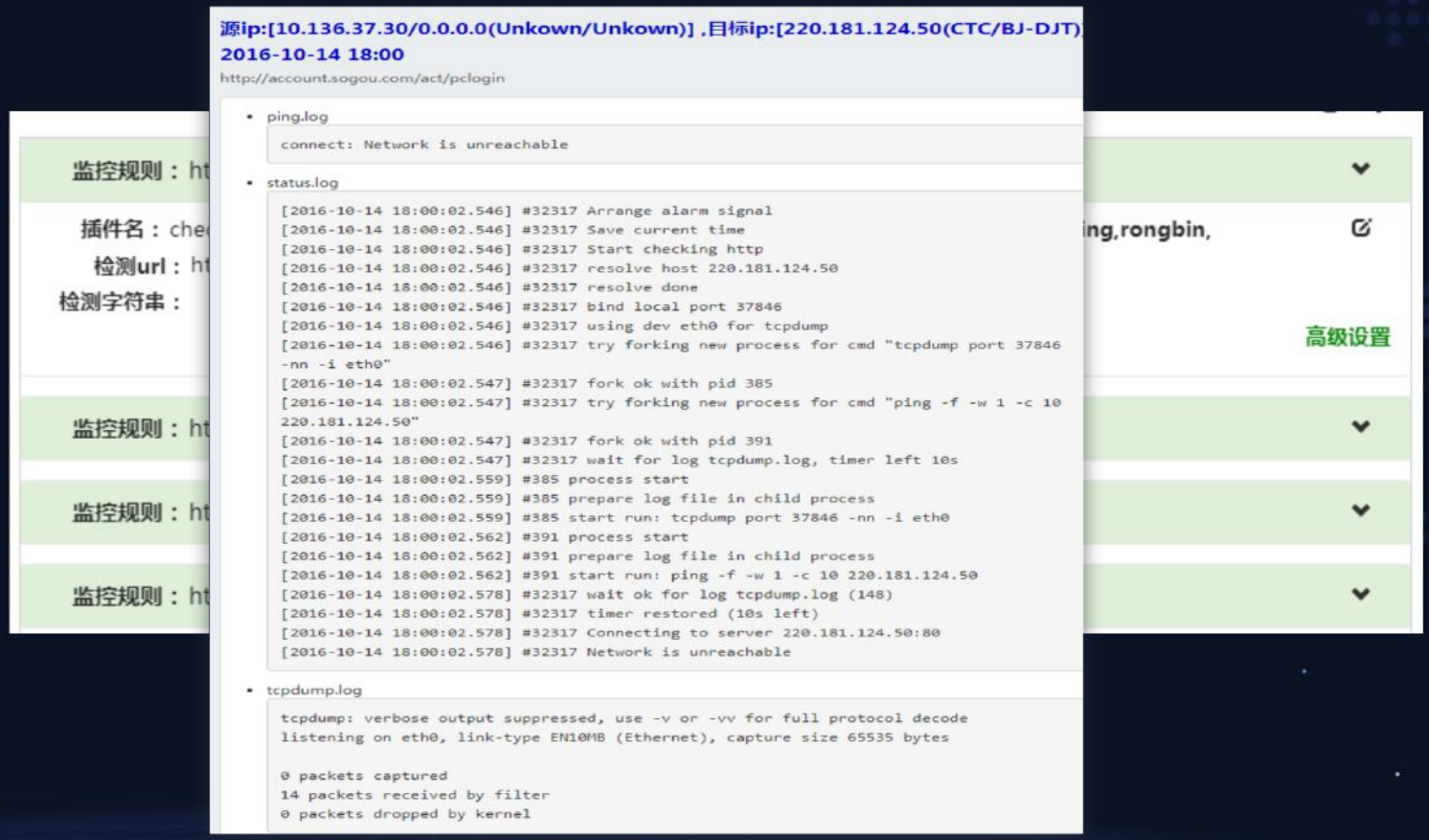
```
STORE_BASE_NAME="error_log access_log"
STORE_SAVE_DAYS="180"
STORE_APP_MODE="nginx"
STORE_LZO=0
STORE_RSYNC_BWLIMIT=10000
STORE_RSYNC_PROC_NUM=20
STORE_CPU_LIMIT="0-3"
```

# 运维平台：关于监控

GOPS2018 Shenzhen

## 用户视角（黑盒监控）

- 支持多种监控插件
- 灵活的语义定制
- 完整的现场快照

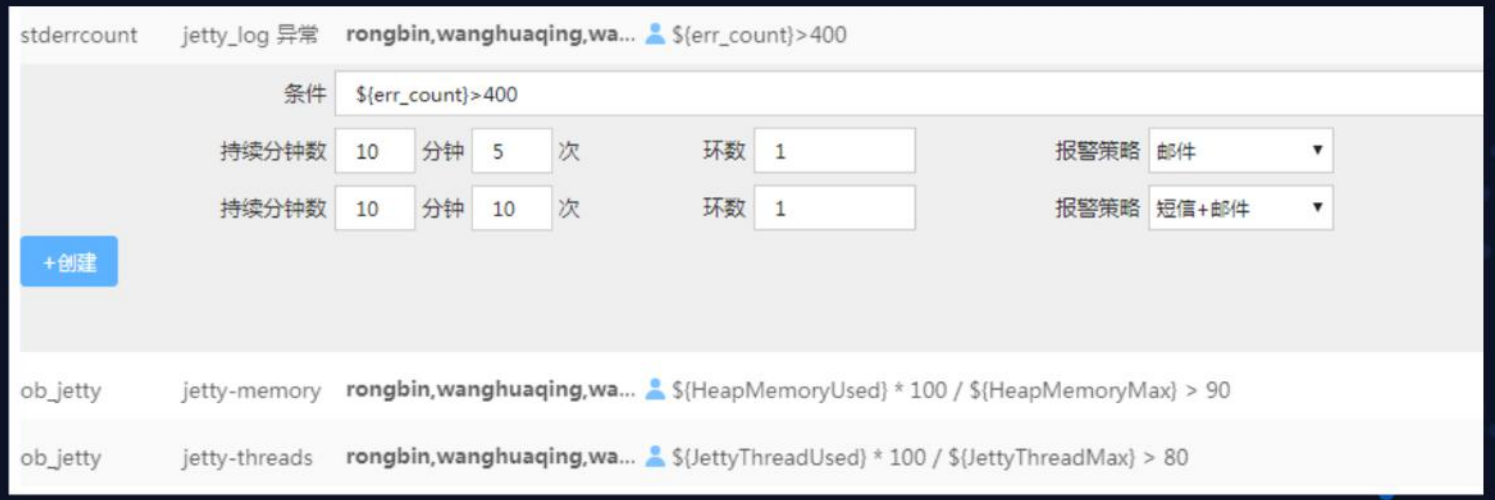


# 运维平台：关于监控



## 系统视角（白盒监控）

- 灵活语义定制
- 灵活报警策略



# 运维平台：资源管理

GOPS2018 Shenzhen

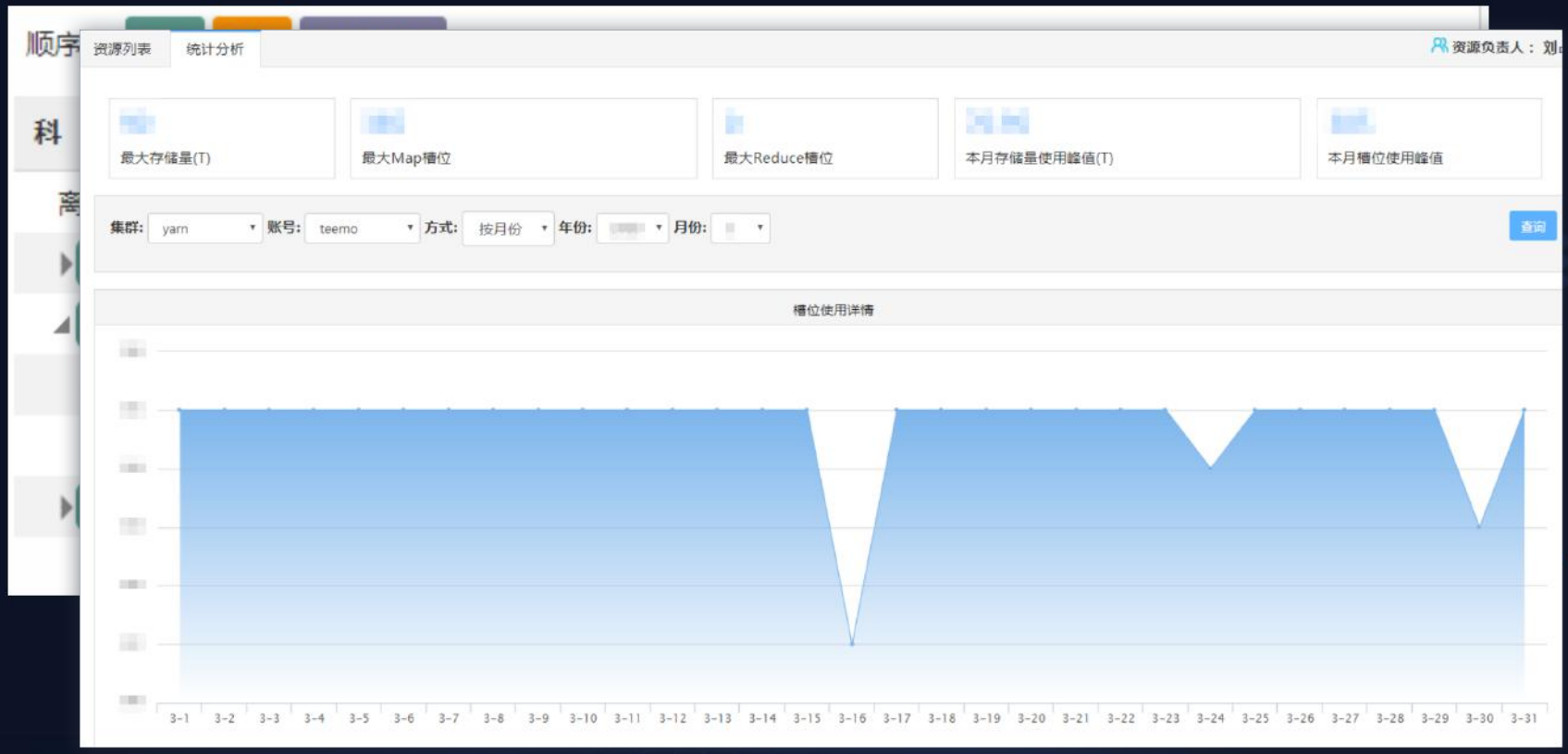
## 资源全流程管理

- 资源申请 → 业务目标 → 逐级审批
- 逐级审批 → 生成上线单
- 生成上线单 → 机群扩容 → 开放申请
- 开放申请 → 使用/结算

# 运维平台：资源管理

GOPS2018 Shenzhen

## 结算与优化



# 目录

1. 搜狗大数据业务概况

2. 搜狗基础运维平台简介

3. 搜狗大数据产品化实践

# 3.0时代

AI驱动/产品商业化
大数据时代

GOPS2018
Shenzhen

## 新问题 [1] ?

产品之间的
数据依赖更高

业务对数据的
安全更加敏感

跨产品的数据
共享的门槛很高

**如何让公司的数据 安全 高效的 共享**