

# 用户视角下的 搜狗大数据平台建设

搜狗 | 技术工程部 - 冯晋





GOPS2018  
Shenzhen

# 1 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



GOPS2018  
Shenzhen

# | 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



GOPS2018  
Shenzhen

# | 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+**



GOPS2018  
Shenzhen

# 行业大数据方向演化

## 大数据系统演化时间线







GOPS2018  
Shenzhen

# | 搜狗大数据历史

1.0时代

专用的搜索大数据时代



2004

2009



GOPS2018  
Shenzhen

# 2.0时代

行业接轨/大规模应用时代

2009



2016



GOPS2018  
Shenzhen

[ 3.0时代 ]

AI驱动/产品商业化  
大数据时代

2016



未来





GOPS2018  
Shenzhen

# 搜狗大数据整体架构





GOPS2018  
Shenzhen

# | 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



GOPS2018  
Shenzhen

# | 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





GOPS2018  
Shenzhen

# | 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

资源列表 概要 日志 白盒 属性

机器 空格支持批量 &支持且查询 Go!

机群	域名	套餐	容灾块	IP
<input type="checkbox"/> 线上环境	web01.corp.zw.ted	正常 NAT	U1 ZW-1136	10.142.39.196
<input type="checkbox"/> 线上环境	web02.corp.zw.ted	正常 NAT	U1 ZW-1038	10.142.75.203

重装 自助重启 备注 下线 延时下线 HA/公网IP/NAT 搬迁 Profile ^ 更多 ^





# 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

模块	nginx ×	wapsearchhub ×	resin ×	searchhub ×	balck_agent ×	data_agent ×	datamem ×
	finalpagemem ×	minoritymem ×	querymem ×	resinmem ×	xcb_blackagent ×	[+]	

继承模块

类型 OB配置

```
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

## 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

The screenshot displays a monitoring configuration and log view. At the top, it shows the source IP [10.136.37.30/0.0.0.0(Unkown/Unkown)] and target IP [220.181.124.50(CTC/BJ-DJT)] for the date 2016-10-14 18:00. The URL is http://account.sogou.com/act/plogin.

Configuration fields include:

- 监控规则: ht
- 插件名: che
- 检测url: ht
- 检测字符串:

Log sections include:

- ping.log: connect: Network is unreachable
- status.log: A detailed log of system events including alarm signals, time saving, http checking, host resolution, port binding, process forking, and ping attempts. The log ends with "Network is unreachable".
- tcpdump.log: tcpdump: verbose output suppressed, use -v or -vv for full protocol decode listening on eth0, link-type EN10MB (Ethernet), capture size 65535 bytes. It shows 0 packets captured, 14 packets received by filter, and 0 packets dropped by kernel.

On the right side, there is a sidebar with a dropdown menu, a search icon, and a button labeled "高级设置" (Advanced Settings).



GOPS2018  
Shenzhen

# | 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

stderrcount jetty\_log 异常 rongbin,wanghuaqing,wa...  $\$(err\_count)>400$

条件  $\$(err\_count)>400$

持续分钟数	10	分钟	5	次	环数	1	报警策略	邮件
持续分钟数	10	分钟	10	次	环数	1	报警策略	短信+邮件

+创建

ob\_jetty jetty-memory rongbin,wanghuaqing,wa...  $\$(HeapMemoryUsed) * 100 / \$(HeapMemoryMax) > 90$

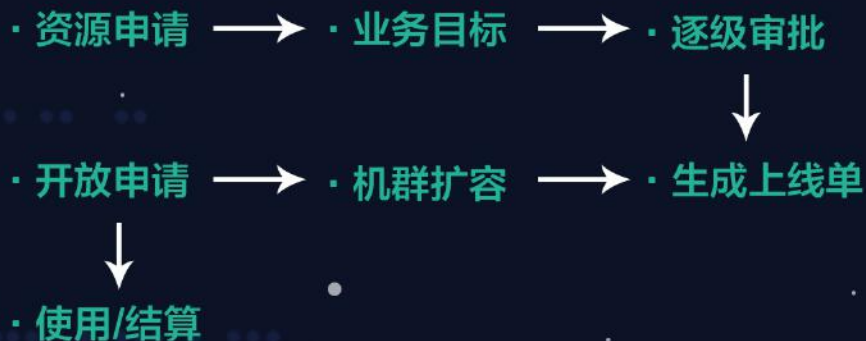
ob\_jetty jetty-threads rongbin,wanghuaqing,wa...  $\$(JettyThreadUsed) * 100 / \$(JettyThreadMax) > 80$



GOPS2018  
Shenzhen

# 运维平台：资源管理

## 资源全流程管理



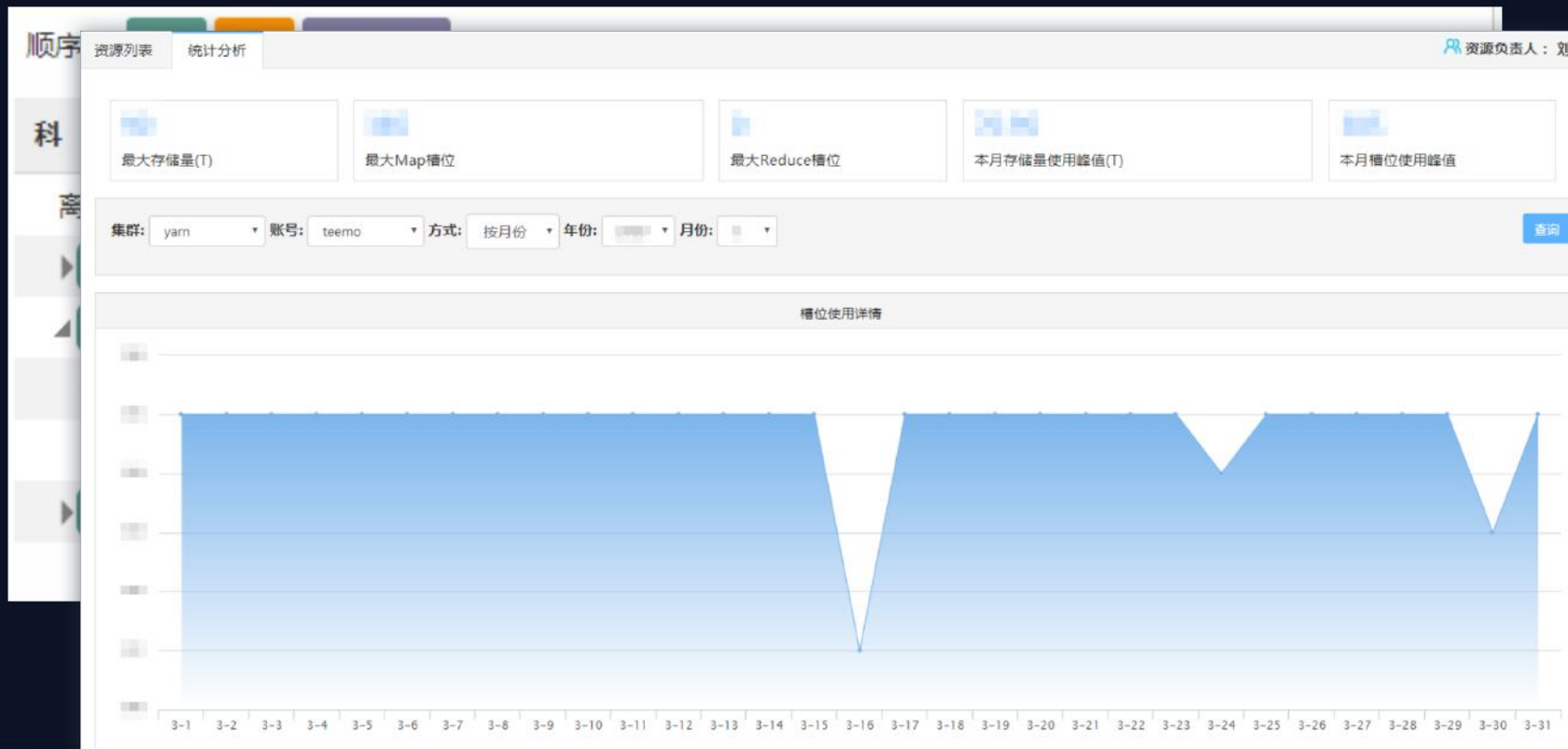




GOPS2018  
Shenzhen

# | 运维平台：资源管理

结算与优化





GOPS2018  
Shenzhen

# | 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



GOPS2018  
Shenzhen

## [3.0时代]

AI驱动/产品商业化  
大数据时代

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