轻松玩转 SPL

—— Splunk 搜索入门和进阶

王国栋

2018 年 3 月 30 日

splunk>

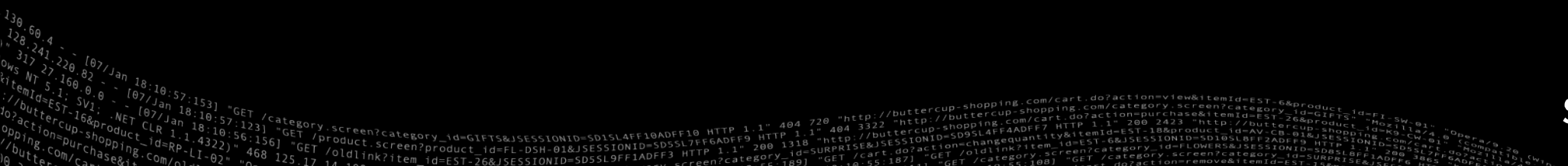
我

码农一枚

- Splunk 上海研发中心工程师
- 热爱编程
- 3 年 splunk 开发经验
- 8 年大数据平台开发经验
- 系统设计
- 开源爱好者

大纲

- 搜索基础
  - SPL
  - 命令
- 搜索进阶
  - 宏
  - 子查询
- 性能优化实践
  - 任务查看器
  - 高效 SPL 的技巧

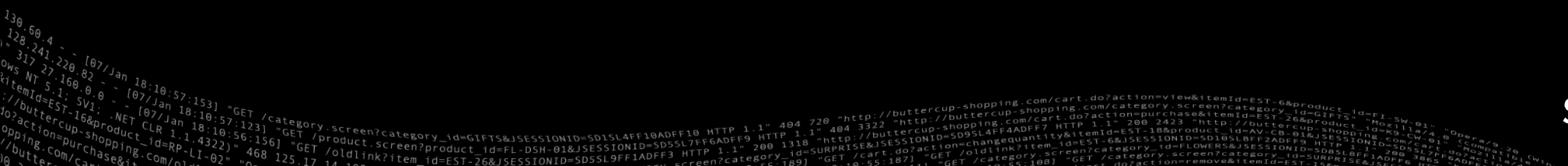


搜索基础

- SPL 大数据忍者必备技能

SPL

- Splunk Search Processing Language
- 面向流程
- 类似流式处理
- 一个 SPL 语句可以包含命令，关键词，参数名称，参数值
  - eval/stats/table/fields …
  - OR/AND/WHERE/AS/BY…



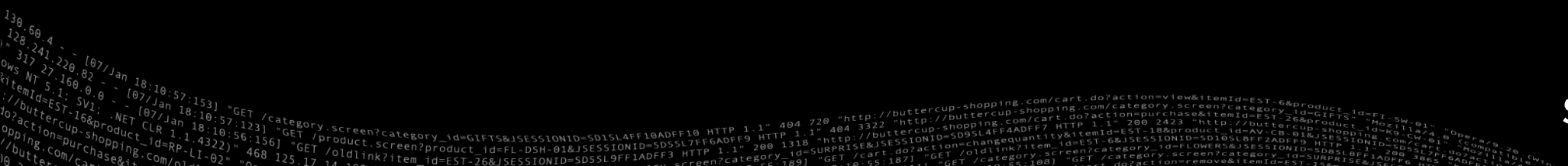
简单搜索例子

- 命令
  - 流式命令（streaming command）
  - 生成器命令（generating command）
  - 转换命令（transform command）
- 隐含的 Search 命令

```
index=_internal source="*metrics.log”
| eval mb = kb / 1024
| stats sum(mb) by series
| rename sum(mb) as "总大小(MB)"
```

常用的 SPL 命令

- Streaming
  - spath
  - eval
  - rex
  - fields
  - lookup
  - head
- Generating
  - search
  - metadata
- Transforming
  - chart
  - timechart
  - stats
  - table



简单查询的例子

将复杂的字符串用双引号引用是一个好主意

新搜索

```
index=* sourcetype="aws:description" name="*calvin*"
```

✓ 175 个事件 (17/11/13 15:28:41.000 之前的部分结果)　无事件采样

新搜索

```
index=* sourcetype="aws:description" name="*calvin*" | stats count
```

✓ 941 个事件 (17/11/13 15:29:51.000 之前的部分结果)　无事件采样

新搜索

```
index=* sourcetype="aws:description" name="*calvin*"
| append [search index=* sourcetype="aws:description" name="*frank*" | head 1]
```

✓ 142 个事件 (17/11/13 15:30:59.000 之前的部分结果)　无事件采样

如果你会写 SQL

那么，你离熟练掌握 SPL 不远了

SQL

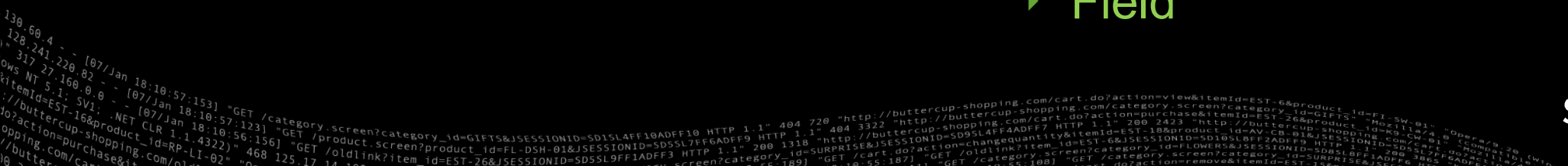
- 面向结果集
- Schema 在查询的时候已经固定
- 可以对数据 CRUD

- 统计聚合函数
- Database
- Table
- Column

SPL

- 面向过程
- Schema on read. 搜索执行的时候动态加载 schema
- 数据不可（immutable）

- 统计聚合函数
- Index
- Sourcetype
- Field

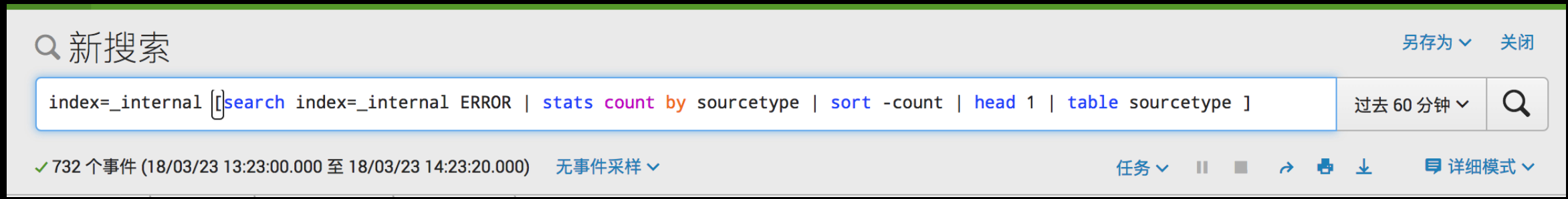


搜索进阶

高级的搜索功能让你成为 SPL 老司机

搜索进阶
子查询

- 执行顺序，总是在主查询之前运行
- 慎用子查询，尤其是嵌套子查询
- Limits.conf： 默认返回一万个结果，默认子查询要在 60 秒之内结束



搜索进阶

多值字段的处理

```
sourcetype="aws:config"
| eval rela=mvzip('relationships{}.resourceType', 'relationships{}.resourceId')
| table rela
| mvexpand rela
| eval rela=split(rela,",")
| eval "资源类型"=mvindex(rela,0) | eval "资源ID"=mvindex(rela,1)
| fields - rela
```


```
relationships: [ [-]
  { [-]
    name: Contains NetworkInterface
    resourceId: eni-00000000
    resourceType: AWS::EC2::NetworkInterface
  }
  { [-]
    name: Is associated with SecurityGroup
    resourceId: sg-00000000
    resourceType: AWS::EC2::SecurityGroup
  }
  { [-]
    name: Is contained in Subnet
    resourceId: subnet-00000000
    resourceType: AWS::EC2::Subnet
  }
  { [-]
    name: Is attached to Volume
    resourceId: vol-00000000
    resourceType: AWS::EC2::Volume
  }
  { [-]
    name: Is contained in Vpc
    resourceId: vpc-00000000
    resourceType: AWS::EC2::VPC
  }
]
```


每页 50 个 ∨ 格式 预览 ∨

	资源ID	资源类型
1	eni-00000000	AWS::EC2::NetworkInterface
2	sg-00000000	AWS::EC2::SecurityGroup
3	subnet-00000000	AWS::EC2::Subnet
4	vol-00000000	AWS::EC2::Volume
5	vpc-00000000	AWS::EC2::VPC

foreach

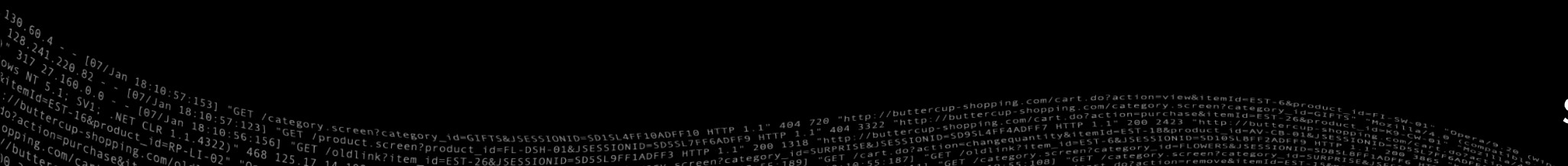
循环处理多个字段

```
sourcetype="aws:config"
| foreach tags{}.* [eval <<MATCHSTR>>='<<FIELD>>']
| table tags.*
```

```
tags: { [-]
  abc: def
  aws_config_test: aws_config_test
}
```

每页 50 个 ✓ 格式 预览 ✓

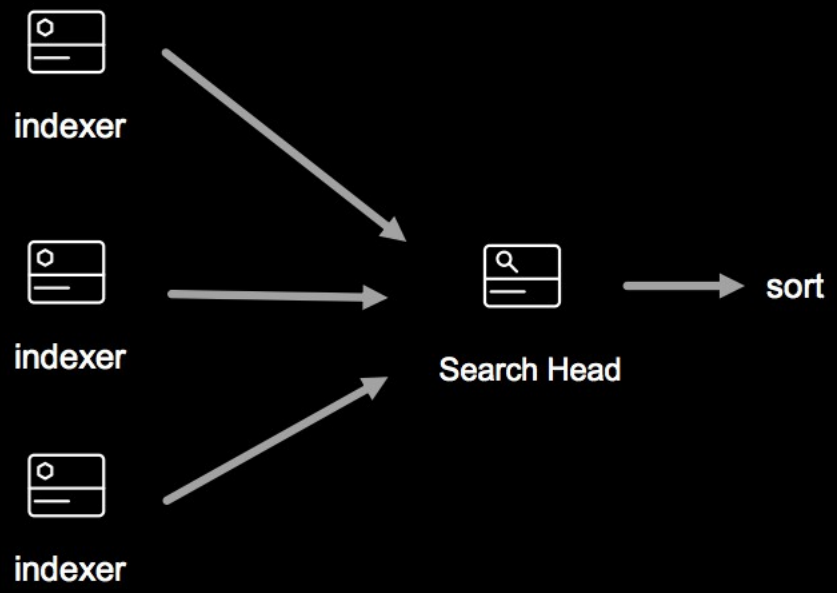
	tags.abc	tags.aws_config_test
1	def	aws_config_test



搜索进阶

分布式环境下的搜索

- 不能在索引集群上分布式执行的命令
  - head
  - sort
- 统计类的转换命令不一定可以分布式执行



高级搜索

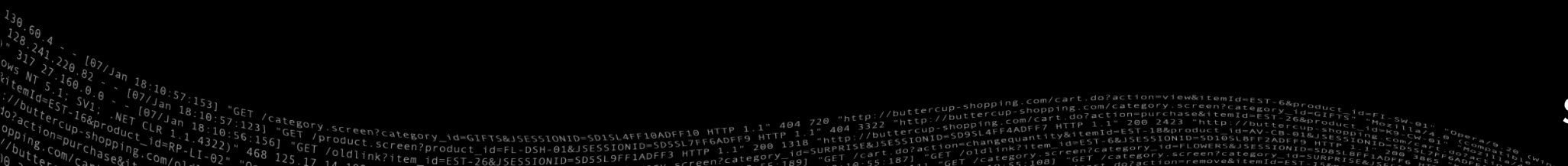
- 宏　macros.conf
  - 非常有利于组织 SPL 搜索代码，保持很好的可读性
  - SPL 中的宏跟编程语言里面的函数一般重要
  - 增加代码复用，每次只需要改宏即可

```
[jenkins_log_filter_level(3)]
args = master_name,level,log_txt
definition = index=jenkins_console host="$master_name$" $level$ source ="logger://*"\
| search "$log_txt$" |eval Message=log_source+",  ".message |rename level as LogLevel|rename log_thrown as Exception
iseval = 0
```

高级搜索

现在所有的搜索命令都无法满足我的要求

- 自定义搜索命令
  - 子进程
  - 需要符合跟 splunk 的交互协议
  - Splunk SDK
- 例子
  - 根据 git 代码提交的历史日志，向 git 服务器拿到两次提交之间的代码改动



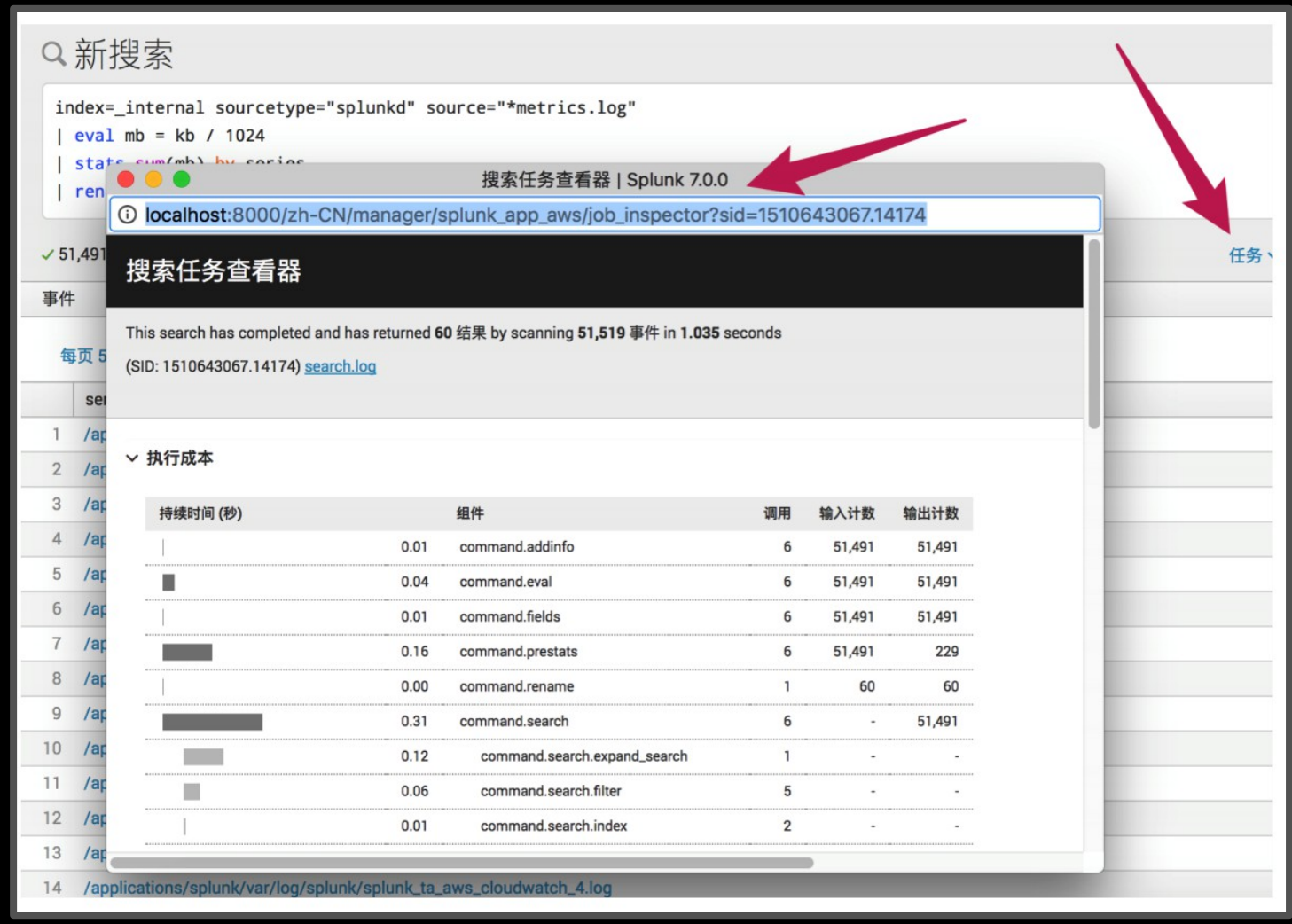
SPL 最佳实践

那些年，我们填过的坑

SPL 最佳实践

用好任务查看器

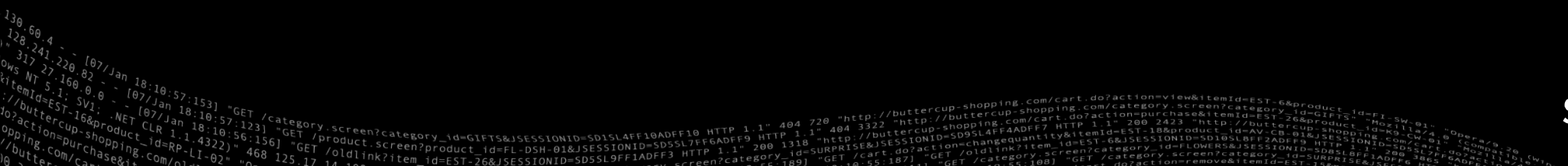
- 查看搜索各阶段耗费的时间
  - Index 定位的时间
  - 原始数据读取时间
  - 字段提取时间
  - 字段过滤时间
  - 任务分发时间
- 查看任务的各项属性
  - 搜索时间范围
  - 搜索的结果集数量和扫描的事件数量



SPL 最佳实践

如何写高效的查询

- 尽可能少的扫描数据
  - 合理的时间范围
  - 过滤掉无关字段，越早越好
    - 尽量避免 wildcard 通配符
    - 尽量将字段过滤放到搜索的前面
  - 尽量避免 pipe
    - index=_internal | search sourcetype=splunkd ␣ index=_internal sourcetype=splunkd
- 慎用排序 sort， dedup 和 join
  - 多考虑用 stats 来实现类似的功能
- 利用好 sourcetype， host 和 source 字段



SPL 最佳实践

血的教训

`aws-config-notification-sourcetype`

region="ap-southeast-1”

```
southeast-1" OR region="ap-southeast-1"))
11-15-2017 11:04:46.671 INFO  UnifiedSearch - Expanded filtering search =
((index="main" OR index="main" OR index="default")
sourcetype="aws:config:notification" region="ap-southeast-1")
11-15-2017 11:04:46.672 INFO  UnifiedSearch - base lispy: [ AND 1 ap
sourcetype::aws:config:notification southeast [ OR index::default index::main ]
]
11-15-2017 11:04:46.672 INFO  UnifiedSearch - Processed search targeting
arguments
11-15-2017 11:04:46.673 INFO  SearchParser - PARSING: prehead limit=10000
null=false keeplast=false
```

1.52	command.search	36
0.41	command.search.index	18
0.12	command.search.expand_search	1

`aws-config-notification-sourcetype`

region="ap-southeast-1*"

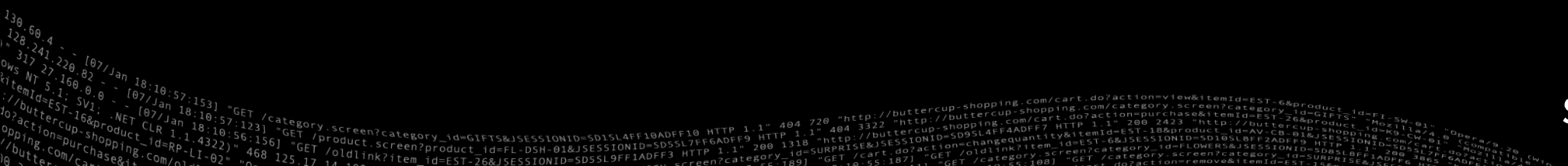
```
11-15-2017 11:06:01.105 INFO  UnifiedSearch - Expanded filtering search =
((index="main" OR index="main" OR index="default")
sourcetype="aws:config:notification" region="ap-southeast-1*")
11-15-2017 11:06:01.106 INFO  UnifiedSearch - base lispy: [ AND 1* ap
sourcetype::aws:config:notification southeast [ OR index::default index::main ]
]
11-15-2017 11:06:01.106 INFO  UnifiedSearch - Processed search targeting
arguments
11-15-2017 11:06:01.106 INFO  SearchParser - PARSING: prehead limit=10000
null=false keeplast=false
```

8.54	command.search	26
6.81	command.search.index	13
0.19	command.search.expand_search	1

SPL 最佳实践

如果我的 SPL 运行起来很慢

- 使用任务查看器检查各项性能指标，找出瓶颈
- 使用快速模式，而不是智能，或者详细模式
- 将搜索中的最常用的字段放到索引时做字段提取
- 升级 splunk 的版本



谢谢

Q&A