日志服务

常见问题

常见问题

问题列表

1. 日志服务是什么？

日志服务（Log Service，简称LS）是对日志收集、存储、订阅平台化服务。服务提供各种类型日志的实时收集，中心化管理、消费功能。

2. 日志服务可以用来做什么事情？

- 多种方式（API、SDK及Logtail接入服务）的日志写入途径。
- 通过Logtail自由定义日志的收集以及解析方式。
- 利用机器组管理数以千计机器上的日志收集。
- 提供实时日志消费与订阅功能。
- 简单易用的控制台配置方式，所有操作都可以在Web端完成。
- 后台与阿里云多个云产品无缝对接。

3. 日志服务的基本概念有哪些？

- 核心概念为：Project（项目、管理日志基础单元）、Logstore（日志库）、Shard（分区）、Topic（主题、对于Logstore二级分类）、Log（日志条数）、LogGroup（日志组）。
- 日志收集概念：Logtail配置（定义如何收集日志配置）、机器分组（分组）。

4. 日志服务有哪几部分组成？

主要有日志收集客户端、服务端以及其他系统。客户端目前有Windows、Linux版本日志收集Agent（Logtail），服务端处理日志服务API读写、以及配置操作，其它系统包括OSS等阿里云产品，即支持向OSS等云产品同步日志数据。

5. 日志服务如何定义一条日志？

日志包含三部分：时间（必填），日志内容（Key：Value对组成），以及元数据（Source，日志来源IP）。

问题列表

1. 日志服务如何存储、管理用户的日志？
2. 删除日志库，日志数据是否丢失？
3. 日志服务日志保存多长时间？可否修改这个保存时限？

1. 日志服务如何存储、管理用户的日志？

日志库（Logstore）是日志服务中的日志存储和查询的基本单元，通常用于存储一类日志数据。目前，支持在控制台或者通过API完成对日志库的增删改查操作。日志库创建完成后，用户通过API或SDK向指定日志库写入日志数据。如果用户希望收集阿里云ECS服务器的数据，日志服务提供的Logtail日志收集服务同样非常方便地收集到日志数据。

2. 删除日志库，日志数据是否丢失？

删除日志库会导致日志数据丢失，请谨慎操作。

3. 删除日志库，日志数据是否丢失？

日志服务有三项功能都与日志保存时间有关，分别如下：

- LogHub（日志中枢）/LogSearch（日志索引与查询）：根据需求自行设置。
- LogShipper（日志投递）：日志投递至OSS、ODPS后，生命周期在以上产品中设置。

日志采集

问题列表

1. 日志采集失败，应如何解决？
2. 采集到的日志数据是乱码，应如何解决？
3. 日志服务可以采集哪些日志？

4. 日志服务有哪些渠道采集日志？应该如何选择这些渠道？
5. 日志服务如何采集ECS上的日志？
6. 日志服务可以采集历史日志吗？
7. 日志服务采集数据的能力如何？有何限制？
8. Logtail采集NAS上的日志需要注意什么？

1. 日志采集失败，应如何解决？

1. 请检查匹配规则是否已通过（比较常见的是设置时候的日志和实际日志存在不一致）。
2. log文件是否实时更新，如果以前的老日志会不被采集。
3. 时间要含年份等日期信息。
4. 有延迟（日志服务读取大约1-2分钟）请耐心等待。
5. 机器组里查看对应的机器心跳是否fail。
6. 不支持非UTF-8编码的数据。
7. 请核实一下日志内的时间，比较常见因为时区问题导致时间过久被丢弃。

如果问题仍未解决，请联系售后技术支持。

2. 采集到的日志数据是乱码，应如何解决？

日志服务插入的数据要求是UTF-8编码的，如果是其他的字符集可能出现乱码的情况。

如果用户的数据是通过SDK插入的，可以在代码写入的时候做字符集转码；如果用户的数据是通过Logtail写入的，可以检查一下Logtail监控的日志文件的编码。

如问题还未解决,请联系售后技术支持

3. 日志服务可以采集哪些日志？

日志服务支持带有时间戳的文本日志和syslog，日志的时间必须是最近7天以内的，并且不能超过当前时间15分钟。

4. 日志服务有哪些渠道采集日志？应该如何选择这些渠道？

日志服务支持用户直接使用API写入；同时提供Linux和Windows版本的Logtail，用于采集磁盘文件上实时更新的日志。

1. 如果应用程序生成的日志不落磁盘，则可直接使用API写入到日志服务。
2. 实时写入磁盘的日志，可以通过Logtail来采集。

5. 日志服务如何采集ECS上的日志？

可以使用Logtail来采集ECS上落在磁盘上的日志，过程如下：

1. 在日志服务控制台上，首先创建一个Logstore
2. 配置Logtail采集的配置
3. 创建机器分组
4. 通过安装脚本自助安装Logtail客户端
5. 将Logtail的配置应用到需要的机器分组即可

6. 日志服务可以采集历史日志吗？

用户可以通过API写入7天以内的数据，7天之前的数据写入会失败。但是，Logtail暂不支持采集历史数据。

7. 日志服务采集数据的能力如何？有何限制？

用户可根据需求调整日志库（Logstore）的分区（Shard）数量。在ECS环境，Logtail采集的速率被限制在1MB/秒。

8. Logtail采集NAS上的日志需要注意什么？

例如Nginx accesslog采集场景，Web Server的nginx配置一般来说都是相同的，传统的方式会写在不同机器上相同名称的文件（Logtail可以正常采集）。使用NAS后，如果多台机器的Nginx日志写入了NAS的相同文件（并发写相同文件场景），Logtail采集日志会缺失或出错。因此，请注意在使用NAS时，保证不同Web Server的日志写入NAS中的不同文件。

问题列表

1. Logtail是什么？

Logtail是日志服务提供的一种便于日志接入的日志采集客户端。通过在您的机器上安装Logtail来监听指定的日志文件并自动把新写入到文件的日志上传到您所指定的日志库。

2. Logtail是否可以采集静态不变的日志文件？

Logtail基于文件系统的修改事件来监听文件的变化，并将实时产生的日志发送到日志服务。如果日志文件没有发生任何修改行为，日志文件内容将不会被Logtail采集。

3. Logtail支持哪些平台？

目前支持Linux 64位和Windows Server2003（含）以后 32/64 位系统。

Linux：

- Aliyun Linux
- Ubuntu
- Debian
- CentOS
- OpenSUSE

Windows：

- Windows 7 (Client) 32bit
- Windows 7 (Client) 64bit
- Windows Server 2003 32bit
- Windows Server 2003 64bit
- Windows Server 2008 32bit
- Windows Server 2008 64bit
- Windows Server 2012 64bit

4. 如何安装、升级Logtail客户端？

安装：目前需要用户通过安装脚本自助安装Logtail客户端。升级：Logtail客户端的升级由日志服务定期完成，升级过程数据采集不中断。

5. 如何配置使用Logtail客户端？

请参考：控制台配置Logtail采集日志说明。

6. Logtail如何工作？

1. 用户在控制台配置需要监控的目录、日志文件名以及相应的解析规则（正则表达式）等。

2. 用户机器上，日志文件发生修改，Logtail收到来自文件系统的事件并读取新产生的日志。
3. Logtail根据正则表达式解析日志格式并发往日志服务。

7. Logtail是否支持日志轮转？

对于日志文件a.LOG，当文件达到一定大小或创建超过一定时间后，a.LOG被mv为a.LOG.1（或其它名称），然后新建一个a.LOG继续写入日志。这个过程称为轮转。Logtail基于文件系统的事件通知，可以自动处理日志轮转的场景。

8. Logtail如何处理网络异常？

网络异常、写入Quota满时，Logtail会将采集到的日志内容写入本地磁盘缓存，并在稍后进行重试。磁盘缓存最大支持500MB，新缓存会覆盖旧缓存；超过24小时未发送成功的缓存文件将被自动删除。

9. Logtail日志采集延时如何？

Logtail基于事件进行日志采集，一般会在3秒内将日志发往日志服务。

10. Logtail如何处理历史日志？

Logtail只用于采集实时日志，如果日志内容的时间与Logtail处理该日志的系统时间相差5分钟以上，即被认为是历史日志。

11. 日志服务修改日志采集配置后多久可以生效？

用户在控制台应用配置到机器组后，Logtail最迟会在3分钟之内加载新配置并生效。

12. 如何调查Logtail采集日志问题？

完整步骤logtail日志采集异常排查。常见问题如下：

1. 查看Logtail心跳是否正常，如不正常，请尝试重新安装Logtail。
2. 确认日志采集配置中的日志文件是否为实时生成。
3. 查看日志采集配置的正则表达式是否与日志内容相匹配。如正则匹配错误，可以在Logtail运行日志查看到相关错误。错误日志路径Linux:/usr/local/ilogtail/ilogtail.LOG。

配置Logtail采集日志数据时，如果Logtail机器组心跳状态不正常，可使用Logtail自动诊断工具或人工诊断的方式排查问题。

自动诊断

日志服务提供Logtail自动诊断工具，排查步骤请参考 Logtail自动诊断工具。

人工诊断

Logtail心跳失败一般由以下原因造成，请逐个排查。

1. 网络未联通

请执行以下命令查看网络连通性，确保网络正常。

经典网络

telnet logtail.cn-<region>-intranet.log.aliyuncs.com 80

VPC网络

telnet logtail.cn-<region>-vpc.log.aliyuncs.com 80

公网

telnet logtail.cn-<region>.log.aliyuncs.com 80

2. 未安装Logtail

请执行以下命令查看客户端状态，如未安装Logtail客户端，请参考Logtail安装，务必按照您日志服务Project所属Region以及网络类型进行安装。

Linux查看客户端状态:

```
sudo /etc/init.d/ilogtaild status
```

Windows查看客户端状态：

```
控制面板 -> 管理工具 -> 服务
查看LogtailDaemon、LogtailWorker两个Windows Service运行状态。
```

3. 安装时所选参数错误

日志服务是地域化的，需要在安装时为客户端指定正确的服务端访问入口，请查看您已安装的客户端使用的配置：

- Linux：/usr/local/ilogtail/ilogtail_config.json
- Windows x64 : C:\Program Files (x86)\Alibaba\Logtail\ilogtail_config.json

Windows x32 : C:\Program Files\Alibaba\Logtail\ilogtail_config.json

确认以下两点：

客户端连接的网络入口所属Region是否与您Project所在Region一致。网络入口列表参考服务入口。

- 是否根据您的机器所属网络环境选择了正确的域名。如VPC环境如果选择了内部域名，是无法联通的。可以Telnet测试ilogtail_config.json中配置的域名，如：telnet logtail.cn-hangzhou-intranet.log.aliyuncs.com 80。

4. 服务端配置了错误的IP或用户标识

一般来说，Logtail在机器上获取IP的方式为：

- 如果本机在文件/etc/hosts中设置了主机名绑定，需要确认绑定的IP。执行命令hostname可以查看主机名。

  如果没有设置主机名绑定，会取本机的第一块网卡IP。

  在服务器上查看IP地址：

- Linux：/usr/local/ilogtail/app_info.json
- Windows x64 : C:\Program Files (x86)\Alibaba\Logtail\app_info.json

  Windows x32 : C:\Program Files\Alibaba\Logtail\app_info.json

  如果服务端机器组内填写的IP与客户端获取的IP不一致，则根据情况进行修改：

  若服务端机器组填写了错误IP，请修改机器组内IP并保存，等待1分钟再查看。

  若修改了机器上的网络配置（如/etc/hosts修改），请重新启动Logtail以获取新的IP。

  如有需要，可以执行以下命令重启Logtail。

  - Linux：sudo /etc/init.d/ilogtaild stop; sudo /etc/init.d/ilogtaild start
  - Windows : **控制面板** -> **管理工具** -> **服务** -> **重启LogtailWorker**

5. 未配置AccessKey

请检查文件/usr/local/ilogtail/ilogtail.LOG，是否有以下错误：Unauthorized ErrorMessage:no authority, denied by ACL

如果出现以上错误，说明您的主账号没有配置AccessKey，因此Logtail不能正常运行。请参考[5分钟快速入门]中配置AccessKey步骤，正确配置AccessKey。

如果您的问题仍未解决，请提工单联系我们。

Kafka是分布式消息系统，由于其高吞吐和水平扩展，被广泛使用于消息的发布和订阅。以开源软件的方式提供，各用户可以根据需要搭建Kafka集群。

日志服务(Log Service)是基于飞天Pangu构建的针对日志平台化服务。服务提供各种类型日志的实时采集，存储，分发及实时查询能力。通过标准话的Restful API对外提供服务。

Log Service Loghub提供公共的日志采集、分发通道，用户如果不想自己搭建、运维kafka集群，可以使用Log Service LogHub功能。

Log Service Loghub & Kafka 概念映射

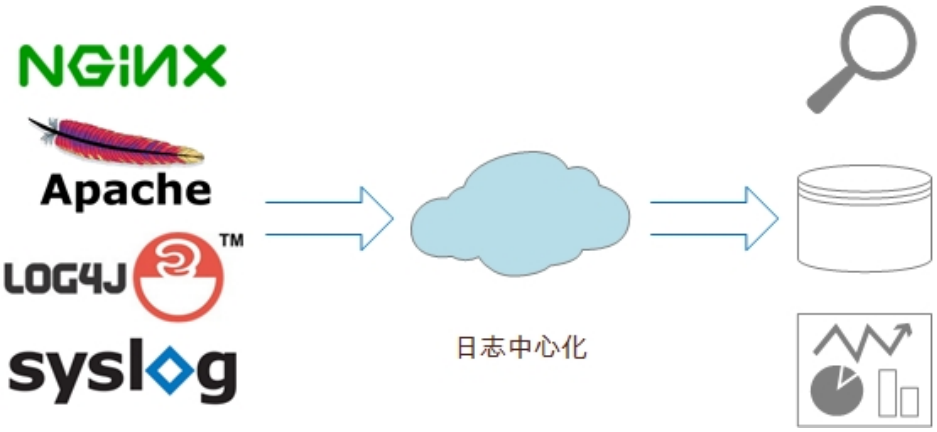
概念	Kafka	Loghub
存储对象	topic	logstore
水平分区	partition	shard
数据消费位置	offset	cursor

Loghub & Kafka 功能比较

功能	Kafka	LogHub
使用依赖	自建或共享Kafka集群	Log Service服务
通信协议	TCP 打通网络	Http (restful API)，80端口
访问控制	无	基于云账号的签名认证+访问控制
动态扩容	暂无	支持动态shard个数弹性伸缩(Merge/Split)，对用户无影响
多租户Qos	暂无	基于shard的标准化流控
数据拷贝数	用户自定义	暂不开放，默认3份拷贝
failover/replication	调用工具完成	自动，用户无感知
扩容/升级	调用工具完成，影响服务	用户无感知
写入模式	round robin/key hash	暂只支持round robin/key hash
当前消费位置	保存在kafka集群的zookeeper	服务端维护、无需关心
保存时间	配置确定	根据需求动态调整

日志采集场景下客户端测评

DT时代，数以亿万计的服务器、移动终端、网络设备每天产生海量的日志。中心化的日志处理方案有效地解决了在完整生命周期内对日志的消费需求，而日志从设备采集上云是始于足下的第一步。



三款日志采集工具

Logstash

- 开源界鼎鼎大名ELK stack中的”L”，社区活跃，生态圈提供大量插件支持
- Logstash基于JRuby实现，可以跨平台运行在JVM上
- 模块化设计，有很强的扩展性和互操作性。

Fluentd

- 开源社区中流行的日志采集工具，td-agent是其商业化版本，由Treasure Data公司维护，是本文选用的评测版本。
- Fluentd基于CRuby实现，并对性能表现关键的一些组件用C语言重新实现，整体性能不错。
- Fluentd设计简洁，pipeline内数据传递可靠性高
- 相较于Logstash，其插件支持相对少一些。

- Logtail

- 阿里云日志服务的生产者，经过3年多阿里集团大数据场景考验
- 采用C++语言实现，对稳定性、资源控制、管理等下过很大的功夫，性能良好
- 相比于Logstash、Fluentd的社区支持，Logtail功能较为单一，专注日志采集功能。

功能对比

功能项	Logstash	Fluentd	Logtail
日志读取	轮询	轮询	事件触发
文件轮转	支持	支持	支持

Failover处理 (本地checkpoint)	支持	支持	支持
通用日志解析	支持grok（基于正则表达式）解析	支持正则表达式解析	支持正则表达式解析
特定日志类型	支持delimiter、key-value、json等主流格式	支持delimiter、key-value、json等主流格式	支持delimiter、key-value、json等主流格式
数据发送压缩	插件支持	插件支持	LZ4
数据过滤	支持	支持	支持
数据buffer发送	插件支持	插件支持	支持
发送异常处理	插件支持	插件支持	支持
运行环境	JRuby实现，依赖JVM环境	CRuby、C实现，依赖Ruby环境	C++实现，无特殊要求
线程支持	支持多线程	多线程受GIL限制	支持多线程
热升级	不支持	不支持	支持
中心化配置管理	不支持	不支持	支持
运行状态自检	不支持	不支持	支持cpu/内存阈值保护

日志文件采集场景 - 性能对比

日志样例:以Nginx的access log为样例，如下一条日志365字节，结构化成14个字段：

在接下来的测试中，将模拟不同的压力将该日志重复写入文件，每条日志的time字段取当前系统时间，其它13个字段相同。

相比于实际场景，模拟场景在日志解析上并无差异，有一点区别是：较高的数据压缩率会减少网络写出流量。

Logstash

logstash-2.0.0版本，通过grok解析日志并写出到kafka（内置插件，开启gzip压缩）。

日志解析配置：

```
grok {
patterns_dir=>"/home/admin/workspace/survey/logstash/patterns"
match=>{ "message"=>"%{IPORHOST:ip} %{USERNAME:rt} - \[%{HTTPDATE:time}\] \"%{WORD:method}
```

```
%{DATA:url}\" %{NUMBER:status} %{NUMBER:size} \"%{DATA:ref}\" \"%{DATA:agent}\" \"%{DATA:cookie_unb}\"
\"%{DATA:cookie_cookie2}\" \"%{DATA:monitor_traceid}\" %{WORD:cell} %{WORD:ups}
%{BASE10NUM:remote_port}" }
remove_field=>["message"]
}
```

测试结果：

写入TPS	写入流量 (KB/s)	CPU使用率 (%)	内存使用 (MB)
500	178.22	22.4	427
1000	356.45	46.6	431
5000	1782.23	221.1	440
10000	3564.45	483.7	450

Fluentd

td-agent-2.2.1版本，通过正则表达式解析日志并写入kafka（第三方插件fluent-plugin-kafka，开启gzip压缩）。

日志解析配置：

```
<source>
type tail
format /^(?<ip>\S+)\s(?<rt>\d+)\s-
\s\[(?<time>[^\]]*)\]\s"(?<url>[^\"]+)"\s(?<status>\d+)\s(?<size>\d+)\s"(?<ref>[^\"]+)"\s"(?<agent>[^\"]+)"\s"(?<
cookie_unb>\d+)"\s"(?<cookie_cookie2>\w+)"\s"(?
<monitor_traceid>\w+)"\s(?<cell>\w+)\s(?<ups>\w+)\s(?<remote_port>\d+).*$/
time_format %d/%b/%Y:%H:%M:%S %z
path /home/admin/workspace/temp/mock_log/access.log
pos_file /home/admin/workspace/temp/mock_log/nginx_access.pos
tag nginx.access
</source>
```

测试结果：

写入TPS	写入流量 (KB/s)	CPU使用率 (%)	内存使用 (MB)
500	178.22	13.5	61
1000	356.45	23.4	61
5000	1782.23	94.3	103

注：受GIL限制，Fluentd单进程最多使用1个cpu核心，可以使用插件multiprocess以多进程的形式支持更大的日志吞吐。

Logtail

logtail 0.9.4版本，设置正则表达式进行日志结构化，数据LZ4压缩后以HTTP协议写到阿里云日志服务，设置

batch_size为4000条。

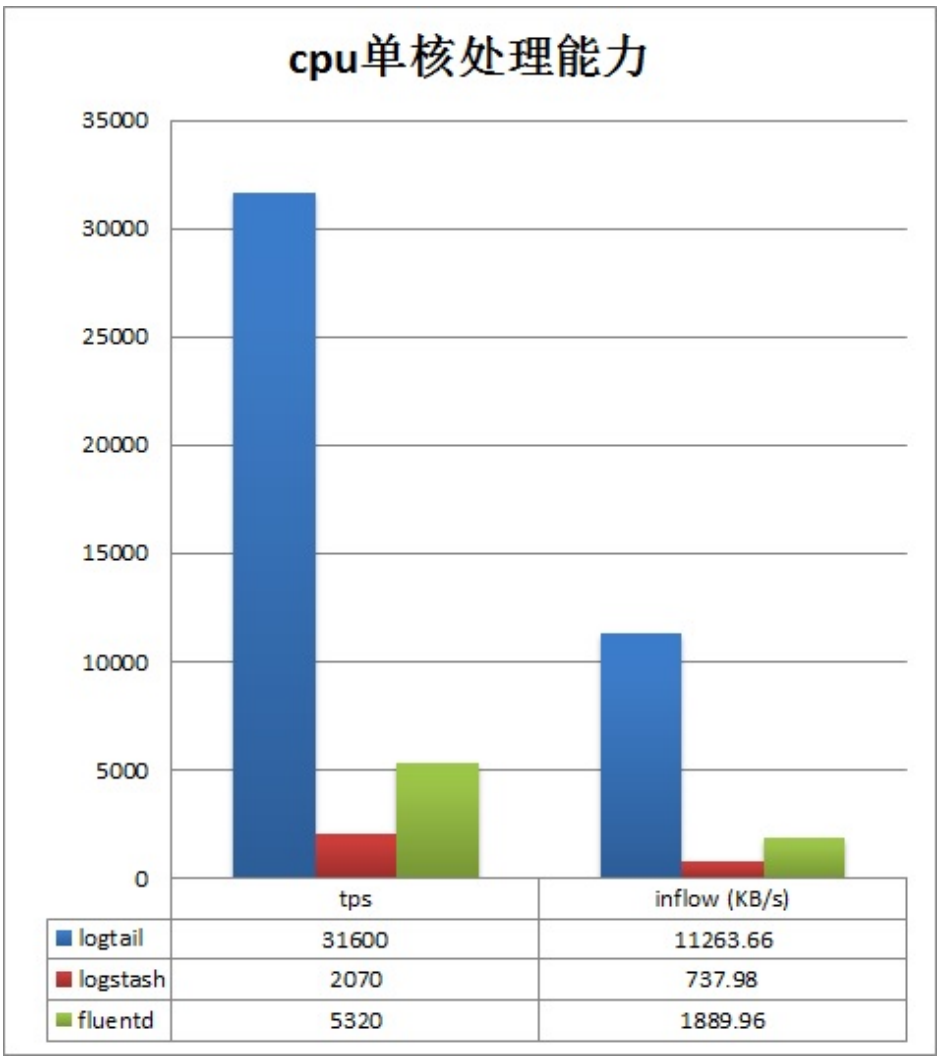
日志解析配置：

```
logRegex : (\S+)\s(\d+)\s-
\s\[([^]]+)]\s"([^"]+)"\s(\d+)\s(\d+)\s"([^"]+)"\s"([^"]+)"\s"(\d+)"\s"(\w+)"\s"(\w+)"\s(\w+)\s(\w+)\s(\d+).*
keys : ip,rt,time,url,status,size,ref,agent,cookie_unb,cookie_cookie2,monitor_traceid,cell,ups,remote_port
timeformat : %d/%b/%Y:%H:%M:%S
```

测试结果：

写入TPS	写入流量 (KB/s)	CPU使用率 (%)	内存使用 (MB)
500	178.22	1.7	13
1000	356.45	3	15
5000	1782.23	15.3	23
10000	3564.45	31.6	25

单核处理能力对比



总结

可以看到三款日志工具各有特点：

- Logstash支持所有主流日志类型，插件支持最丰富，可以灵活DIY，但性能较差，JVM容易导致内存使用量高。
- Fluentd支持所有主流日志类型，插件支持较多，性能表现较好。
- Logtail占用机器CPU/内存资源最少，性能吞吐量较好，针对常用日志场景支持全面，但缺少插件等机制，灵活性和可扩展性不如以上两个客户端。

日志查询

问题列表

1. 如何在日志数据中搜索IP地址？
2. 如何在日志中搜索包含空格的关键字？
3. 如何完成双重条件检索？
4. 日志服务提供哪些渠道查询采集的日志？
5. 日志服务提供什么样的查询能力？
6. 日志查询有什么限制？

解决方案

1. 如何在日志数据中搜索IP地址？

在日志数据中搜索IP地址，支持全部匹配的方式检索。您可以直接在日志数据中直接搜索指定IP地址相关的日志信息，比如包含指定IP地址、过滤指定IP地址等。但是目前尚不支持部分匹配的方式检索，即不能直接搜索IP地址的一部分，因为小数点不是日志服务默认的分词项。如果需要的话，建议自行过滤，比如用SDK先下载数据，然后在代码里用正则或者用string.indexof等方法判断。

例如，在日志服务的Project中搜索条件如下：not ip:121.42.0 not status:200 not 360jk not DNSPod-Monitor not status:302 not jiankongbao not 301 and status:403。

搜索结果中仍会出现121.42.0网段地址。因为日志服务会认为121.42.0.x是一个词，所以只有搜121.42.0.x能搜到结果，而121.42.0的话不会搜到这个结果，同理加上not也就不会过滤该地址。

2. 如何在日志中搜索包含空格的关键字？

搜索包含空格的关键字时，如果直接搜索，则会得到包含空格左侧关键字或右侧关键字的所有日志。建议您在查询的包含空格的关键字时，把关键字用双引号包裹起来，将引号中的内容作为一个关键字进行搜索，搜索结果就是符合条件的日志内容。

例如，在以下日志中搜索包含关键字POS version的日志。

```
post():351]:&amp;nbsp;device_id:&amp;nbsp;BTAddr&amp;nbsp;:&amp;nbsp;B6:xF:xx:65:xx:A1&amp;nbsp;IMEI&amp;nbsp;:&amp;nbsp;35847xx22xx81x9&amp;nbsp;WifiAddr&amp;nbsp;:&amp;nbsp;4c:xx:0e:xx:4e:xx&amp;nbsp;|&amp;nbsp;user_id:&amp;nbsp;bb07263xxd2axx43xx9exxea26e39e5f&amp;nbsp;POS&amp;nbsp;version:903
```

如果直接搜索POS version，则会得到包含POS或者version的所有日志，不符合搜索要求。如果搜索"POS version"，则会得到包含关键字POS version的所有日志。

3. 如何完成双重条件检索？

双重条件检索时，只需同时输入两个语句即可。

例如，需要在Logstore中搜索数据状态不是OK或者Unknown的日志。直接搜索not OK not Unknown即可得到符合条件的日志。

4. 日志服务提供哪些渠道查询采集的日志？

日志服务提供了三种方式查询日志：

1. 通过日志服务控制台查询。
2. 通过SDK查询。详见SDK。
3. 通过Restful API查询，详见API。

5. 日志服务提供什么样的查询能力？

- 提供组合条件过滤查询，查询语法参见日志查询语法。
- 能够提供单次查询1000万日志的能力。用户可以根据一定的条件筛选出需要的日志，读取命中日志在时间维度上的分布，或者拿到原始日志。
- 查询提供了cache的功能，第二次查询相同的条件获得更加完整的查询结果。

6. 日志查询有什么限制？

- 最多能够查询10个词组成的组合条件。
- 单次查询结果最多获取100行原始数据。
- 单次查询最多处理1000万行数据。

日志服务提供了两项功能都和“读”有关：

日志收集与消费（LogHub)：提供公共的日志收集、分发通道。全量数据顺序（FIFO）读写，提供类似Kafka的功能

- 每个Logstore有一个或多个Shard，数据写入时，随机落到某一个Shard中
- 可以从指定Shard中，按照日志写入shard的顺序批量读取日志
- 可以根据Server端接收日志的时间，设置批量拉取Shard日志的起始位置（cursor）
- 日志在LogHub中，默认保留2天时间，在此期间，日志可消费

日志查询（Index）：在LogHub基础上提供海量日志查询功能，根据关键词的数据随机查询

- 通过关键词查找，只抓取符合要求的数据
- 支持关键词 AND、NOT、OR的布尔组合
- 数据查询不区分Shard

两者区别：

功能	日志查询(LogSearch)	日志收集与消费(LogHub)
关键词查找	支持	不支持

小量数据读取	快	快
全量数据读取	慢(100条日志100ms，不建议这样使用)	快（1MB日志10ms，推荐方式）
读取是否区分Topic	区分	不区分，只以Shard作为标识
读取是否区分Shard	不区分，查询所有Shard	区分，单次读取需要指定Shard
费用	较高	低
适用场景	监控、问题调查等需要过滤数据的场景	流式计算、批量处理等全量处理场景

日志投递