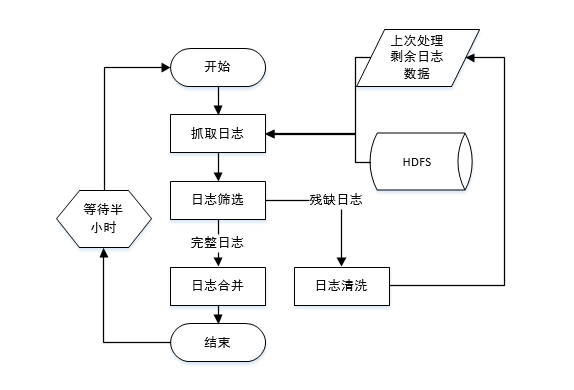
统一能力平台日志合并规则描述

本文档针对AI+能力日志，包括（识别、合成、语义）能力日志的合并以及清除规则进行了描述和说明。

能力日志分为3种能力类型，已日志中的 ability字段进行区分、以下依次进行描述，合并后的日志格式（ES格式）见附件[《ES日志描述.xlsx》](ES日志描述.xlsx)

日志合并清洗大致流程如下：



1. 日志每隔半小时（15分钟）处理一次
2. 抓取半小时到当前时间数据，与上次处理的剩余日志数据合并做筛选
3. 日志筛选，根据规则将下述规则的音频作为完整日志，
   1. 语音识别：同一sid包含cmd=ssb、auw、grs、sse四种状态的日志作为完整日志；残缺日志，再次处理仍未被选中合并，同一sid下如存在ssb和sse则也进行合并。
   2. 语音合成：同一sid包含cmd=ssb、txtw、grs、sse四种状态的日志作为完整日志；残缺日志，再次处理仍未被选中合并，同一sid下如存在ssb和sse则也进行合并。
4. 日志合并、清洗规则详见下文。

# 语音识别日志（ability=ab\_asr）

## 日志结构介绍



识别能力日志分为4个类型以cmd字段进行区分，每个类型的日志字段一致但是需要从不同的cmd类型的日志中获取字段值进行日志的合并，cmd的值为ssb（sessionbegin）、auw（audiowrite）、grs（getresult）、sse（sessionend）中的一个。

日志格式和示例如下。

{

"ability": "ab\_asr",

"svc": "iat",

"cmd": "auw",

"sid": "2-4dc8-a79b-fe87c33b331c",

"csid": "a-40e6-88b3-0d7ae7ad81ee",

"busid": null,

"appid": "pc20onli",

"uid": "userid",

"ret": "0",

"call\_time": "17/07/27-14:29:27 108",

"type": "",

"input\_text": "",

"result": "这是测试文本内容",

"data": "http://xxx.xxx.xxx.xxx:8080/1234567.pcm",

"data\_content": "1-100",

"where": "",

"aue": "audio/L16;rate=8000",

"auf": "raw",

"extend\_params": "",

"area": "",

"auth\_id": "72-d1-12-13-33-11\_12313323-0"

}

## 合成规则以及对应关系

1. 规则描述

spark对每个完整日志，对syncid字段进行排序，然后只保留cmd=ssb、cmd=sse之间的日志，具体字段提取规则如下：

|  |  |  |
| --- | --- | --- |
|  | 字段名称 | 赋值字段 |
| 日志字段及合成规则 | sid | cmd==ssb日志的sid字段 |
| svc | cmd=ssb日志的svc字段 |
| ability | cmd==ssb日志的ability字段（ab\_asr） |
| appid | cmd==ssb日志的appid字段 |
| syncid | 此字段仅用于日志排序 int |
| csid | cmd==ssb日志的csid字段 |
| busid | cmd==ssb日志的busid字段 |
| uid | cmd==ssb日志的uid字段 |
| ret | cmd==sse日志的ret字段,或者出现该字段为非0的数值 |
| call\_time | cmd==ssb日志的call\_time字段 |
| type | cmd==ssb日志的type字段 |
| input\_text | 为空 |
| result | cmd==auw或者cmd==grs日志的result字段拼接 |
| data\_list | cmd==auw的data字段和data\_content字段拼成一个Object，并放在data\_list下成为Array，并已data\_content首个字符串(","分割得到的第一个值)的数值大小排序 |
| data | cmd==auw的data字段 |
| data\_content | cmd==auw的data\_content字段 |
| where | cmd==ssb日志的where字段 |
| aue | cmd==ssb日志的aue字段 |
| auf | cmd==ssb日志的auf字段 |
| end\_time | cmd==sse的call\_time字段 |
| rest\_time | end\_time时间减去call\_time，毫秒级 |
| extend\_params | cmd==ssb日志的extend\_params字段 |
| area | cmd==ssb日志的area字段 |
| auth\_id | cmd==ssb日志的auth\_id字段 |
| output |  |
| online | cmd==ssb日志的online字段 |

说明：绝大部分字段都从ssb的日志中获取，例外情况有：

end\_time：取自session\_end的call\_time

ret：取自排序后第一个ret!=0的字段或sse的ret值

result：cmd==auw和cmd=grs日志的result字段拼接（注意忽略空值或null）。

rest\_time ：end\_time减去call\_time

data\_list：cmd==auw的data字段和data\_content字段拼成一个Object，并放在data\_list下成为Array，并已data\_content首个字符串(","分割得到的第一个值)的数值大小排序

data\_list.data：cmd==auw的data字段

data\_list.data \_content：cmd==auw的data\_content字段

## 清理方案

1、对筛选掉的日志，添加筛选标记，第二次仍未被作为完整日志处理，则清洗掉。

# 语音合成引擎日志（ability=ab\_tts）

## 合成规则及对应关系

|  |  |  |
| --- | --- | --- |
|  | 字段名称 | 赋值字段 |
| 日志字段及合成规则 | svc | cmd=ssb日志的svc字段 |
| csid | cmd==ssb日志的csid字段 |
| busid | cmd==ssb日志的busid字段 |
| appid | cmd==ssb日志的appid字段 |
| uid | cmd==ssb日志的uid字段 |
| sid | cmd==ssb日志的sid字段 |
| ret | cmd==sse日志的ret字段,或者出现该字段为非0的数值 |
| call\_time | cmd==ssb日志的call\_time字段 |
| type | cmd==ssb日志的type字段 |
| input\_text | cmd==txtw日志的input\_text字段 |
| result | 空 |
| data\_list | 空 |
| data |  |
| data\_content |  |
| where | cmd==ssb日志的where字段 |
| aue |  |
| auf |  |
| ability | cmd==ssb日志的ability字段（ab\_tts） |
| end\_time | cmd==sse的call\_time字段 |
| rest\_time | end\_time减去call\_time |
| extend\_params | cmd==ssb日志的extend\_params字段 |
| area | cmd==ssb日志的area字段 |
| auth\_id | cmd==ssb日志的auth\_id字段 |
| output |  |
| online |  |

说明：绝大部分字段都从ssb的日志中获取，例外情况有：

end\_time：取自session\_end的call\_time

ret：取自排序后第一个ret!=0的字段或sse的ret值

input\_text：cmd==txtw日志的input\_text字段

rest\_time ：end\_time减去call\_time

## 清洗方案

1、对筛选掉的日志，添加筛选标记，第二次仍未被作为完整日志处理，则清洗掉。

# 语义能力引擎日志（ability=ab\_nlp）

Nlp日志为单条日志格式无需合并，直接保存即可

# 指标计算：

服务量：count（sid）

交互时长：sum（resttime），rest\_time（会话耗时），某时间段resttime的和

成功率：count（ret=0）/count（\*），某时间段内引擎返回码成功的个数/引擎总计调用个数

并发量：由讯飞提供接口，从redis中获取。

最大并发量：某时间段内并发量的峰值

平均并发量：某时间段内并发量的平均值。