

讯飞语音云实时分析架构与实践

科大讯飞云平台

刘君

CONTEXT



I. 讯飞语音云概况

II. 语音云数据平台架构

III. 语音云实时分析实践

IV. 集群监控、运维及优化

V. 后续发展

Q&A

讯飞语音云概况



讯飞语音云简介

讯飞语音云的主要能力

业务带来的实时需求

语音识别：音频转文字（分钟级到5小时）

语音合成：文字转音频

人脸、声纹识别

语义理解

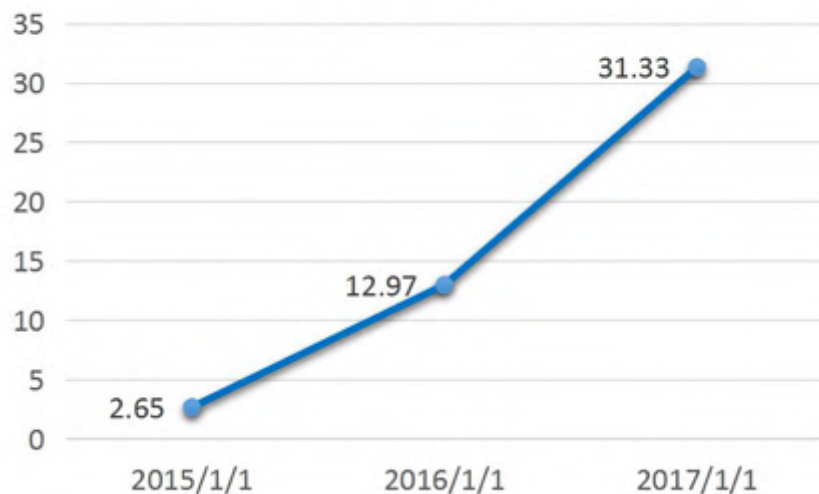
评测

全双工带上下文的AIUI语音交互
重要指标的实时统计（BI&大屏）

实时监控集群、服务的健康状况（实时监控平台）

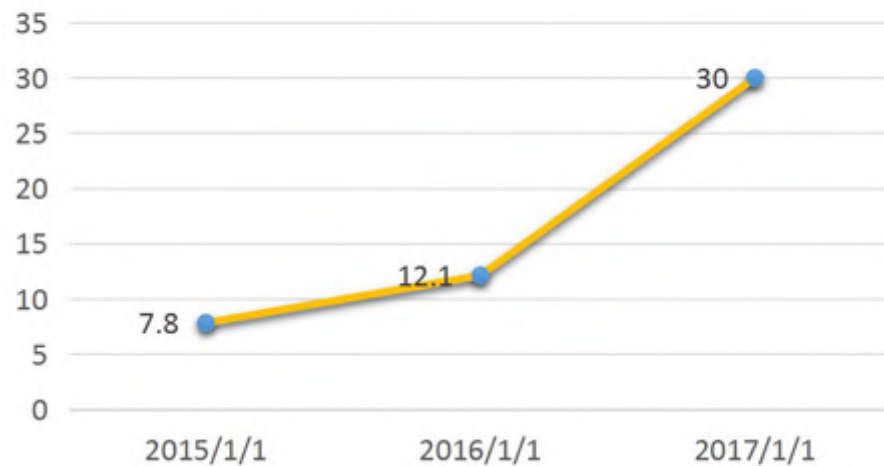
及时定位问题（云知道实时检索）

业务量 (亿)



单日数据量：**20TB/500亿条**

第三方创业团队 (万)



语音云数据平台架构



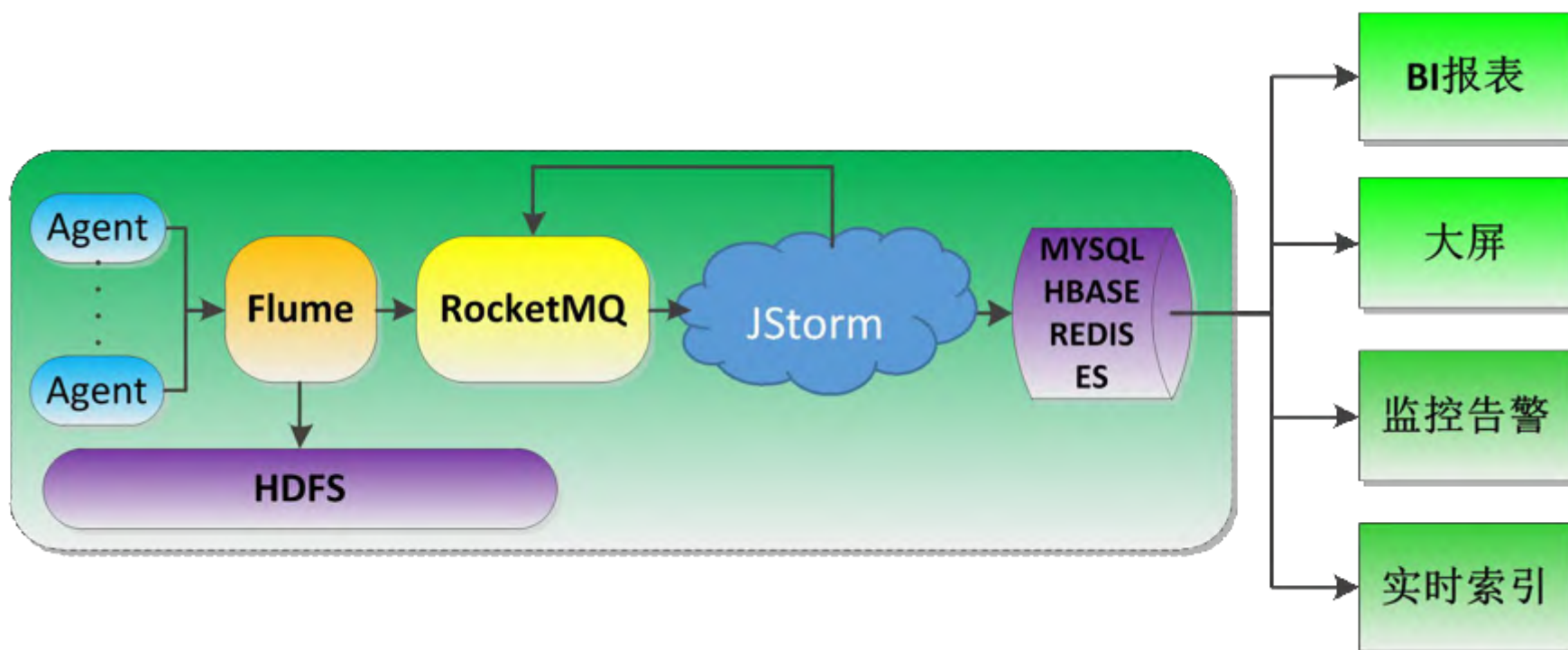
数据产品（BI&大屏，监控，云知道，DMP，etc）

数据服务（Hbase,Mysql,Redis,Mongo,ES）



数据通道（Flume,RocketMQ,DataX）

语音云实时数据流水线



语音云实时分析实践



BI&大屏

实时监控平台

云知道

BI&大屏

语音云数据特性

异地多活

- 三地（合肥，北京，广州）云服务

富文本

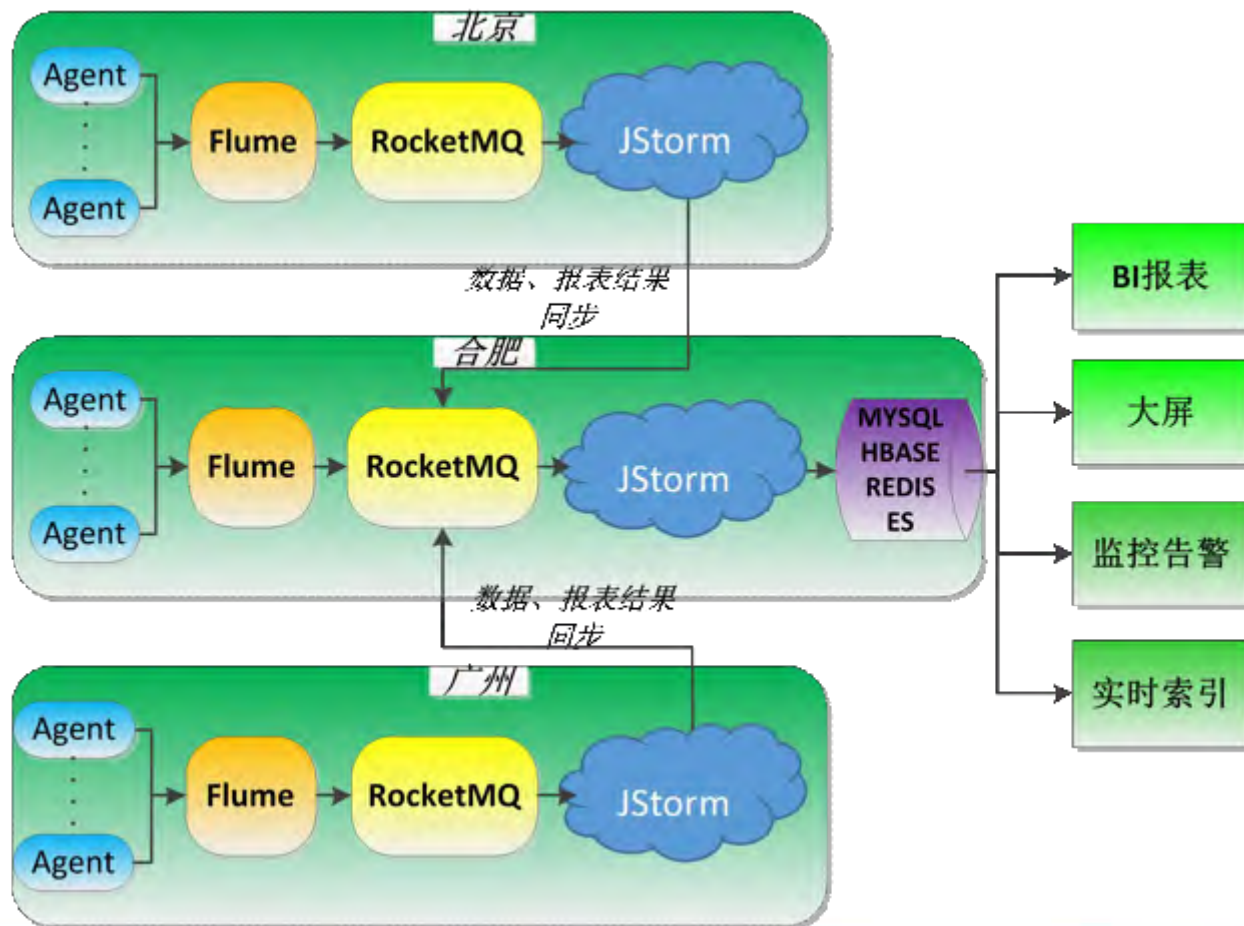
- 日志数据(2/3)+语音数据(1/3)
- 日志数据复杂

乱序数据

- 单个语音会话会产生很多条日志
- 乱序到达

BI&大屏

异地多活



BI&大屏

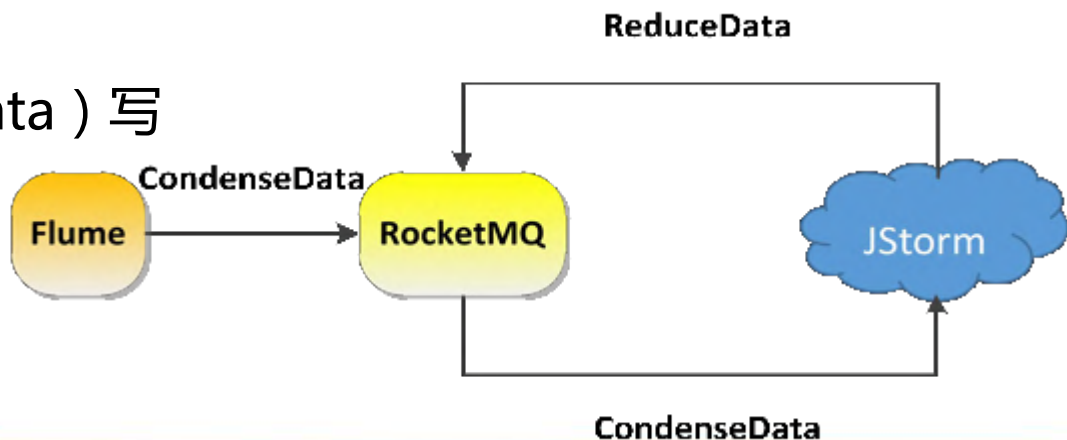
异地多活

- PV
- UV
 - 合肥Redis集群保存全量用户信息，用于实时比对
 - 三地将uid及其统计指标汇总到合肥，统一计算新增、活跃
 - 离线任务每天定时增量更新Redis中的全量用户信息

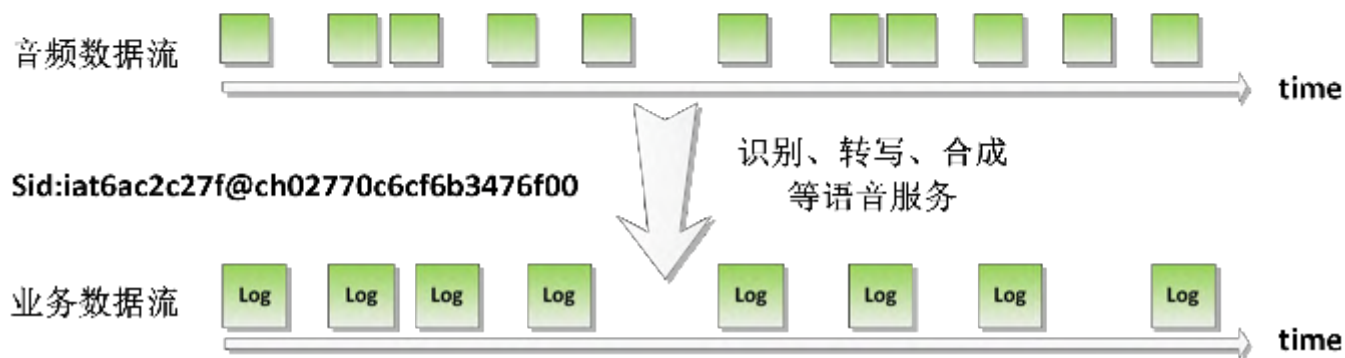
BI&大屏

富文本

- 目标：减少数据体量
- Flume端：
 - memory-channel共享数据
 - 在sink端做日志精简，提取有效字段（CondenseData）
 - 采用disruptor无锁队列来增大吞吐（5倍提升）
- 实时数据清洗：
 - 按会话聚合（ReduceData）写回rmq，供其他作业消费



BI&大屏 乱序数据



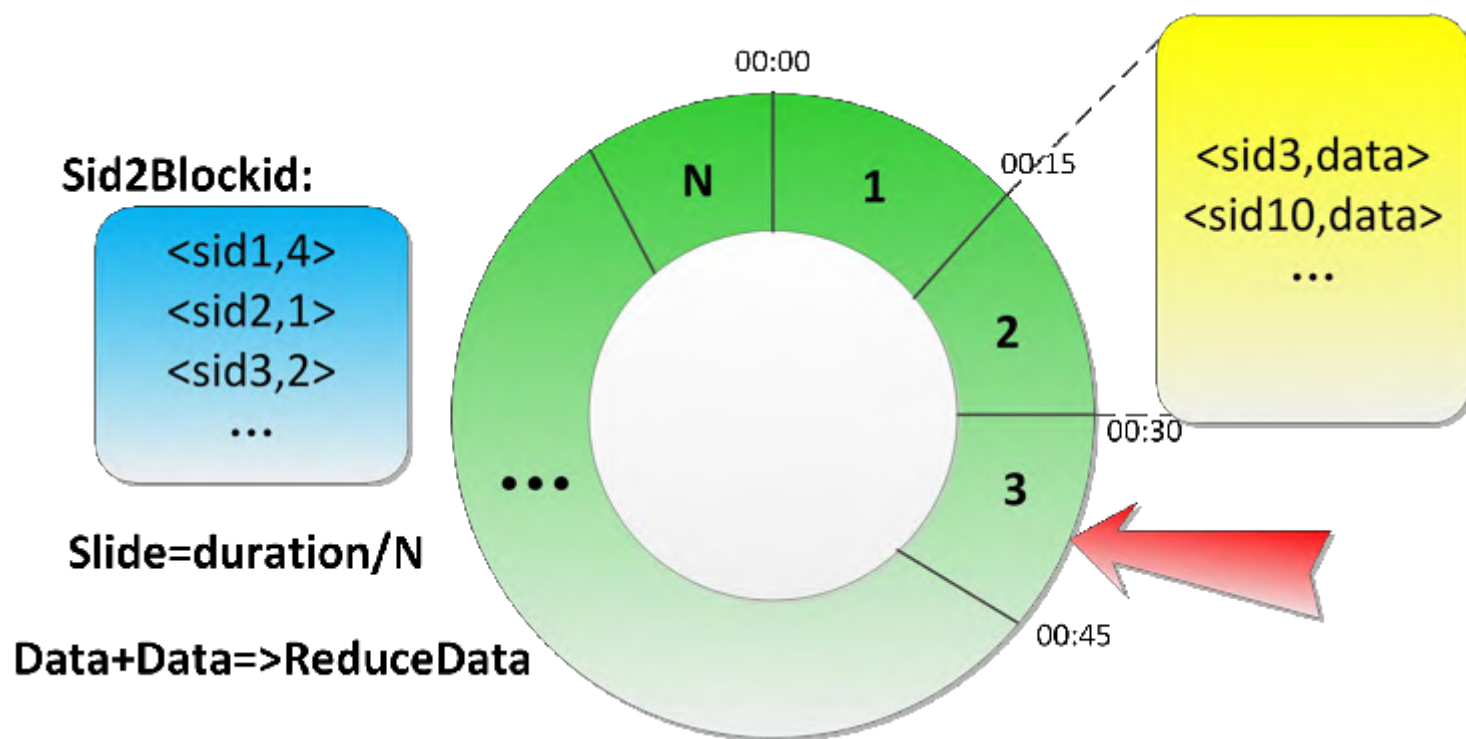
BI&大屏

乱序数据

- 实时清洗（会话聚合）——99%的会话长度在30秒以内
 - 按sid分组
 - 以块为单位Cache数据，对相同sid的数据做聚合
 - 对每块数据设定内存最大存活时间
 - 以块为单位flush超时了的聚合数据
- 聚合后
 - TPS：80W --> 1W
- 安全稳定但统计不够精确

BI&大屏

乱序数据



- 应用概况
- 应用趋势
- 渠道分析
- 终端属性
- 行为分析
- 自定义分析
- 错误分析
- SDK下载

关键实时指标 2017-05-12 13:38:43

新增用户	活跃用户	启动次数	累计用户
8,578 ⁺¹	129,185 ⁺²	293,812 ⁺¹⁸	14,563,256 ⁺¹

近日概况

日期	新增用户	活跃用户	新增用户占比	启动次数	次均使用时长	人均使用时长
前日	21,647	208,900	10.36%	631,816	00:00:37	00:01:47
昨日	21,553 [▼]	209,701 [▲]	10.28% [▼]	638,671 [▲]	00:00:38 [▲]	00:01:49 [▲]

应用规模

累计用户	14,554,678
一次性用户 (%)	4,156,645 (28.56%)
累计启动次数 (人均)	165,880,900 (11.40)

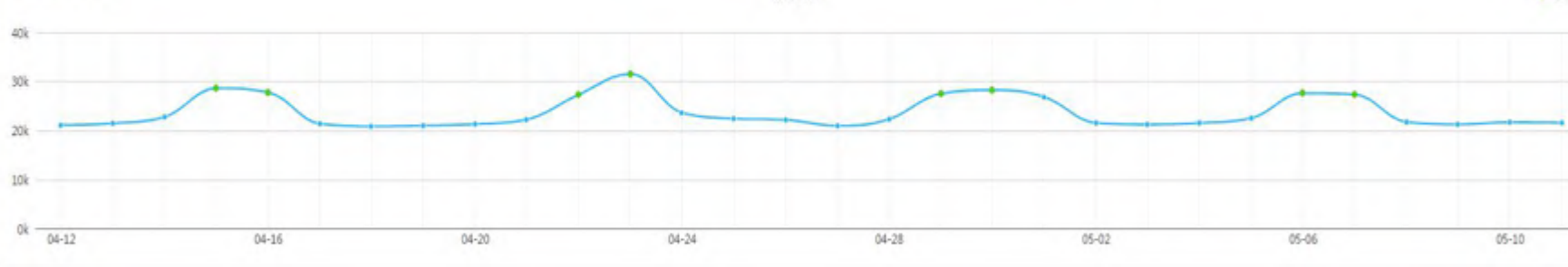
使用情况

首日留存率	17.16%
月活跃用户	1,776,774
月活跃率	11.8%

30日趋势分析

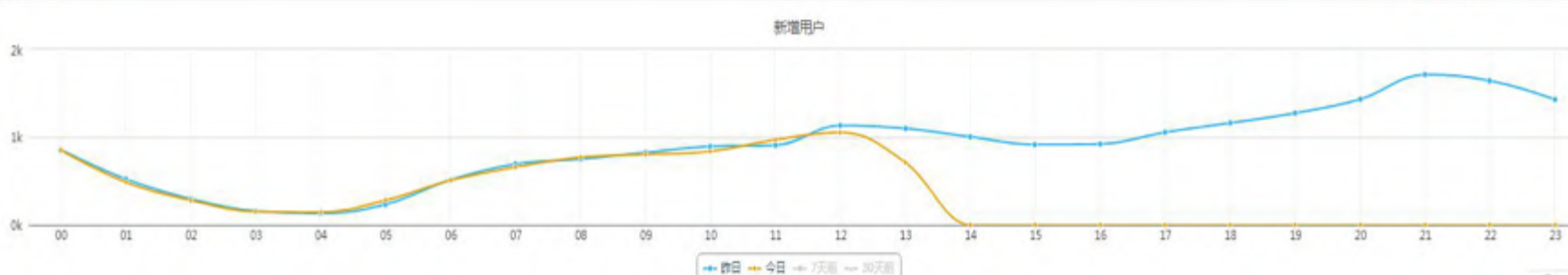
新增用户 活跃用户 新用户占比 启动次数 累计用户 次均使用时长 次均上传流量 次均下载流量

显示运营用户数量



时段分布

新增用户 活跃用户 启动次数



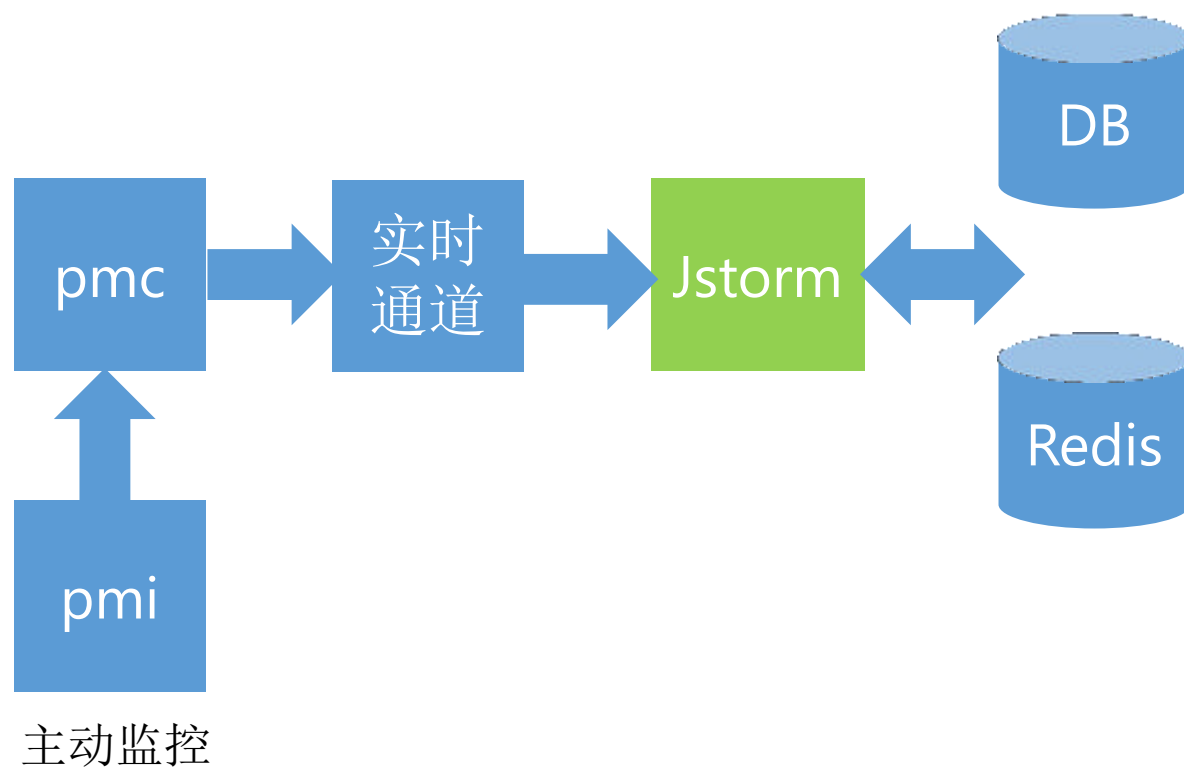
BI&大屏

实时监控平台

云知道

实时监控平台

序列号
磁盘使用率
Cpu占用率
物理内存
虚拟内存
线程数
句柄数
进程数
Tcp链接
流量



实时监控平台

告警类别

- 服务器基础指标监控告警
- 业务主动告警
- PMC客户端异常告警（进程异常关闭）
- 服务器无数据、无法ping通监控告警（宕机）

告警配置

- 告警触发条件可配置
- 报警时间可配置
- 告警通知人员可配置
- 邮件、短信告警方式



实时监控平台

规模

- 5000+线上应用
- 100+商业应用
- 100+业务
- 100+组件
- 1500+服务器

BI&大屏

实时监控平台

云知道

52

	飞来了一只小红鸟。	嗯，会做梦，真好，什么鸟飞走了？又
recResult		
nlpResult		
sub_ntt		

◎ 结构化日志 ◎ 备份日志

☐ AtsLog

```
[{"id": "isb7f4b17fz011b0d6e25c4cc000", "timestamp": "2017-05-10 17:13:09 0", "callLog": [{"timestamp": "2017-05-10 17:13:09 58", "loglevel": "INFO", "extras": {"ret": "0", "perf": "0.251", "ecps": "0.400", "syncid": "0", "_LOGLIB_LOG_RANDOM_ID_": "24834888", "pid": "001", "tid": "c984700", "param": "rule16k, account"}, {"timestamp": "2017-05-10 17:13:09 154", "loglevel": "INFO", "extras": {"_LOGLIB_LOG_RANDOM_ID_": "127605889", "resource id": "104", "resource type": "LB", "pid": "001", "tid": "c0ad1700", "fn": "on_task", "describes": []}, {"timestamp": "2017-05-10 17:13:09 60", "loglevel": "INFO", "extras": {"syncid": "0", "waitcost": "0.000", "stime": "2017-05-10 17:13:20 60", "workcost": "5.342", "etime": "1970-01-01 08:00:00 000", "syncid": "0", "LOGLIB_LOG_RANDOM_ID_": "127605889"}]}, {"id": "isb7f4b17fz011b0d6e25c4cc000", "timestamp": "2017-05-10 17:13:09 0", "callLog": [{"timestamp": "2017-05-10 17:13:20 60", "loglevel": "INFO", "extras": {"syncid": "0", "waitcost": "0.000", "stime": "2017-05-10 17:13:20 60", "workcost": "5.342", "etime": "1970-01-01 08:00:00 000", "syncid": "0", "LOGLIB_LOG_RANDOM_ID_": "127605889"}]}]
```

4

MssLog

```
[{"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 0, "callName": "pb", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 50", "logLevel": "INFO", "extra": {"LOGLIB_LOG_RANDOM_ID": "405368524", "pid": "36a1", "tid": "d143700"}, "describes": ["all params is void enable=true"]}, {"timestamp": "2017-05-10 17:13:09 46", "logLevel": "ERROR", "extra": {"LOGLIB_LOG_RANDOM_ID": "1689594078", "pid": "b7f4", "tid": "b7f4700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "b7f4", "tid": "b7f4700"}, "type": "pb"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 1, "callName": "sw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 150", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "1351510566", "pid": "891b", "tid": "90f6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "891b", "tid": "90f6700"}, "type": "sw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 2, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 353", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "1413172151", "pid": "2bcb", "tid": "39f2700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "2bcb", "tid": "39f2700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 3, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 553", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "8681222789", "pid": "2bcb", "tid": "30ef9700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "2bcb", "tid": "30ef9700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 4, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 753", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "322743480", "pid": "7660", "tid": "12ef6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "7660", "tid": "12ef6700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 5, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:09 953", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "1219439453", "pid": "0258", "tid": "76f6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "0258", "tid": "76f6700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 6, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:10 153", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "1602071086", "pid": "4116", "tid": "aaf6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "4116", "tid": "aaf6700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 7, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:10 353", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "512347770", "pid": "0058", "tid": "16ef5700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "0058", "tid": "16ef5700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 8, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:10 553", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "171010405", "pid": "7669", "tid": "45ef6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "7669", "tid": "45ef6700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 9, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:10 753", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "1251077869", "pid": "93de", "tid": "abef6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "93de", "tid": "abef6700"}, "type": "vsw"}, {"id": "iwb7f4b17f9gr01b0c425d3cc400", "ip": "", "syncid": 10, "callName": "vsw", "timestamp": "2017-05-10 17:13:09 0", "callLogs": [{"timestamp": "2017-05-10 17:13:10 953", "logLevel": "INFO", "extra": {"ret": "0", "ses_city": "gr", "ow_city": "gr", "LOGLIB_LOG_RANDOM_ID": "950615431", "pid": "0514", "tid": "96f6700"}, "describes": ["set common: session begin, login"]}], "session": {"begin": 1689594078, "end": 1689594078, "pid": "0514", "tid": "96f6700"}, "type": "vsw"}]
```

集群监控、运维及优化



实时集群监控

时间	服务器	状态	开机时间	磁盘使用	cpu	物理内存	虚拟内存	句柄数	线程数	进程数	tcp连接数	流量
8:20:20	192.168.51.132	正常	07-26 18:49:30	/ 2.00% /disk2 1.00% /disk3 1.00% /disk4 1.00% /boot 9.00% /disk1 1.00%	1%	9%	5%	0	355	226	7	3Kbps
8:20:20	192.168.70.125	正常	12-01 14:23:15	C: 47.50% D: 16.90% E: 10.42% F: 19.75%	15%	84%	16%	107654	7038	181	1416	6798Kbps
8:20:20	192.168.71.202	正常	01-06 03:22:57	/ 12.00% /boot 18.00% /opt 1.00%	1%	30%	21%	0	137	90	4	5Kbps
8:20:19	192.168.71.21	正常	07-08 19:21:38	/boot 9.00%	0%	4%	2%	0	587	472	598	0Kbps
8:20:19	192.168.51.119	正常	09-25 18:18:36	/ 1.00% /disk2 63.00% /disk3 63.00% /disk4 64.00% /boot 9.00% /disk1 68.00% /disk5 58.00% /disk6 57.00% /disk7 58.00% /disk8 57.00%	1%	97%	58%	0	214	170	21	218Kbps
8:20:17	192.168.70.156	正常	12-04 13:14:55	C: 5.50% D: 19.96% E: 29.33% F: 20.61%	15%	86%	33%	108465	7028	171	1417	7195Kbps
8:20:17	192.168.70.126	正常	11-14 20:22:07	C: 48.20% D: 13.33% E: 73.62% F: 21.58%	15%	90%	15%	105983	7013	167	1406	7791Kbps
8:20:16	192.168.150.20	正常	06-18 15:47:46	/ 34.00% /boot 30.00% /opt 97.00%	10%	100%	69%	0	151	109	1113	80166Kbps
8:20:16	192.168.70.112	正常	10-31 06:44:55	/ 2.00% /boot 11.00%	0%	3%	1%	0	292	212	9859	154Kbps

监控时间	组件名	责任人	服务器IP	异常描述	路径	备注	处理建议
2015-11-04 16:05	JstormMonitor.class	刘君	42.84	错误码比率在 20%~100%之间，错误比率为 100%，主要错误码为：20002:3,备注：20002:Alarm time:2015-11-04 15:57:41 Topology spout sendtps alarm!tps=0 in 30s, topology:AdStatsTopology_lzwang2 spout:pullspout :,20002:Alarm time:2015-11-04 16:01:06 Topology spout sendtps alarm!tps=0 in 30s, topology:AdStatsTopology_lzwang2 spout:pullspout :,20002:Alarm time:2015-11-04 16:03:37 Topology spout sendtps alarm!tps=0 in 30s, topology:AdStatsTopology_lzwang2 spout:pullspout :	/service/jstorm-monitor/com/iflytek/jstorm/monitor/JstormMonitor.class		

运维

- 隔离JStorm的ZK 和应用的ZK
- 减少nimbus节点负载
- Nimbus HA、Daemon-Tools、Supervisor自检

作业优化

- 减少无意义的网络IO
 - 尽可能小的tuple
 - 对pipeline的数据可以将多个模块合并到一个task
 - 尽量让tuple走本地而非进程间通信
- 均匀hash避免数据热点
- 在吞吐量和时延间的选择
- 性能方面，trident < 使用ack机制普通接口 < 关掉ack机制的普通接口

后续发展



- 日志、音频数据从源端分离
- Beam的引入

Q&A

THANK YOU FOR YOUR
ATTENTION !