

SACC 第八届中国系统架构师大会  
2016 SYSTEM ARCHITECT CONFERENCE CHINA 2016

架构创新之路

# 饿了么实时架构演进

常盛

2016.10.27

The way to do really BIG things is to do really SMALL things, and grow them bigger.

饿了么 | 美好生活 触手可得



饿了么平台



日均交易额  
**2亿元+**



日均订单量  
**500w+**



覆盖城市  
**1000+**



员工数  
**15000+**



e 分享提纲

从0到1

高速发展

成熟与完善

新方向&新挑战

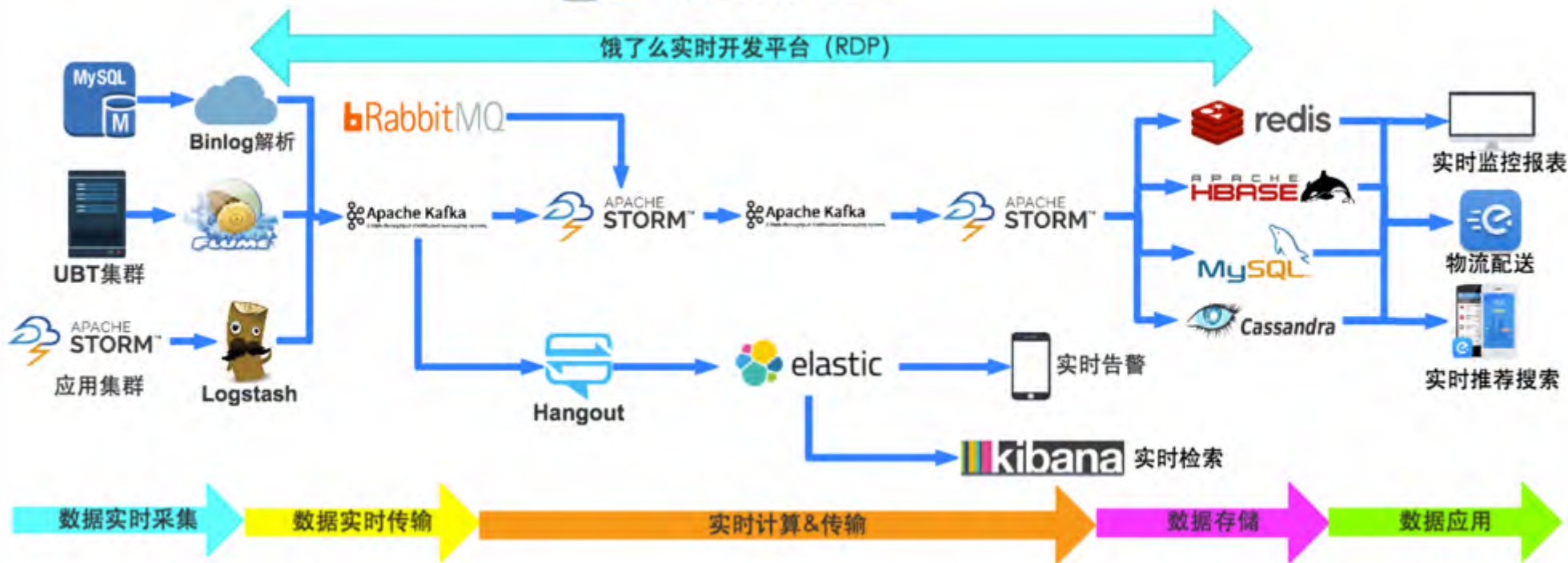
未来构想

Q&A



# 实时平台架构图

## 饿了么实时架构图



每一个组件也许都踩过N次的坑。。。





## e 从0到1的实践

数据源：



+

饿了么

首页

我的订单

数据收集：



python

PK



消息队列：



Apache Kafka  
A distributed streaming platform.

实时引擎：



APACHE  
STORM™

PK

Spark  
Streaming

数据存储：



redis



## e 从0到1的应用

- 页面访问性能
- 主站&各页面PV/UV
- 脚本JS错误



页面性能（分城市）



JS脚本错误（分运营商）



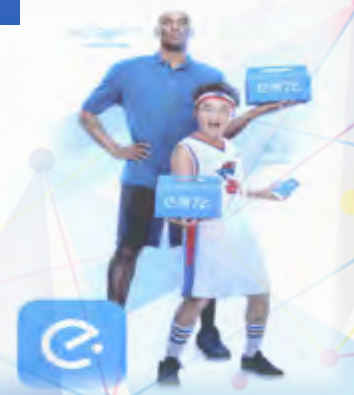
## e 从0到1的总结

巧妇难为无米之炊（业务场景）

欲善其事，必利其器（技术工具）

兵马未动，粮草（监控）先行

他山之石，可以攻玉（高效、快速）



 分享提纲

从0到1

高速发展

成熟与完善

新方向&新挑战

未来构想

Q&A





## e 高速发展之烦恼

- 1、业务场景：交易、推荐、测试...
- 2、平台问题：雪崩、规范...
- 3、监控：粒度粗、告警多...
- 4、压力：大促销、多维指标....
- 5、资源：机器&人力、经验匮乏...



## e 高速发展之行动

### 数据源

- 流量日志
- 交易订单
- 物流配送

### 架构优化

- Binlog Parser
- JVM HLL方案
- 双链路高可用
- 链路压测

### 数据存储

- Hbase
- MySQL
- RedisCluster

### 1. 数据源：

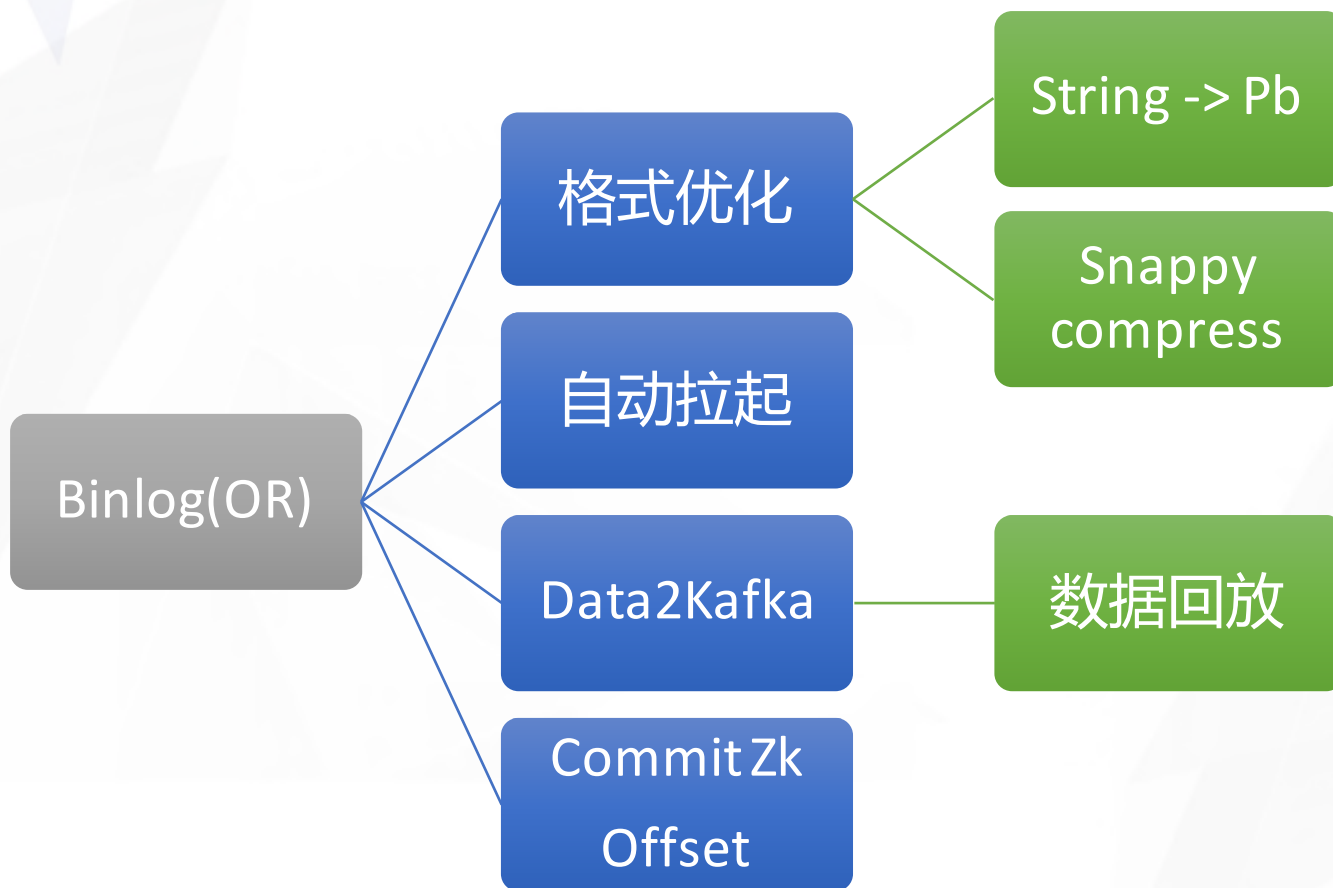
- 流量（UBT集群）
- 交易订单（MySQL）
- 物流配送（MySQL、Rabbitmq）

### 2. 数据存储：

- RedisCluster (双写)
- Hbase
- MySQL (DAL控制)



## 高速发展之架构优化



## e 高速发展之应用优化

### JVM HLL方案

- 存储方式优化
- qps降低4倍

### 双链路高可用方案

- 链路双写
- 应用自动降级
- 数据自动恢复

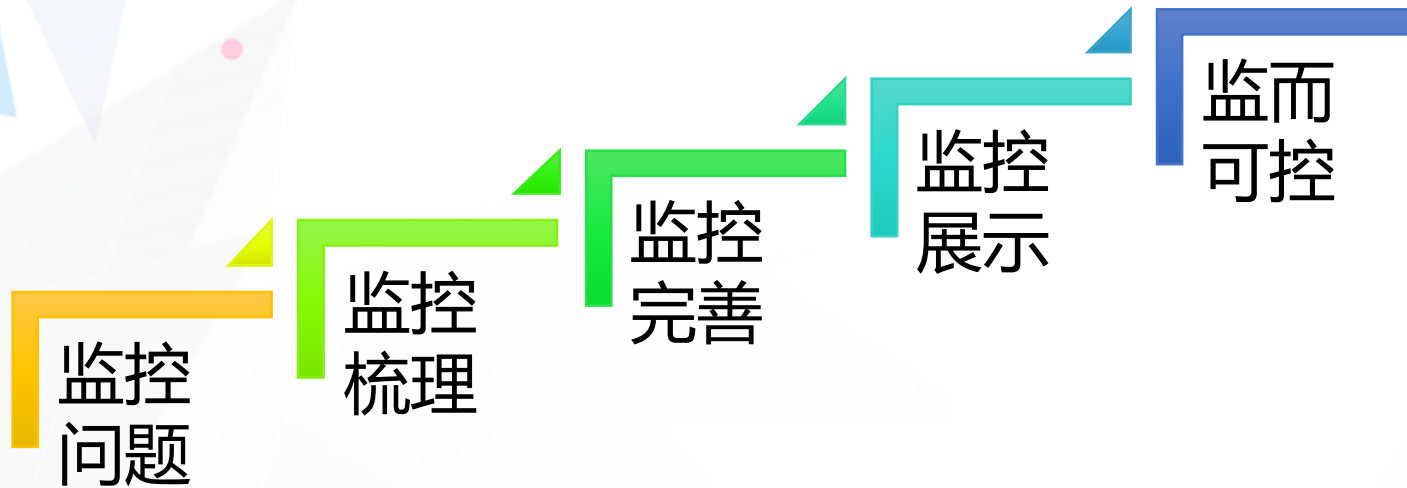
### 实时链路梳理压测

- 延迟毫秒级
- Storm  
Heatbeat调整
- 灾备演练





## e 高速发展监控优化



1. 监控问题：粒度粗、无效告警多 .....
2. 监控梳理：全链路、机器级别、服务级别、应用级别 .....
3. 监控完善：链路延迟、组件Metric、DB性能 .....
4. 监控展示：阈值实时告警; 数据入库展示 .....
5. 监而可控：应用自动拉起、监控驱动优化 .....



# 高速发展之成果



用户行为实时测试系统



上海街道配送热力图

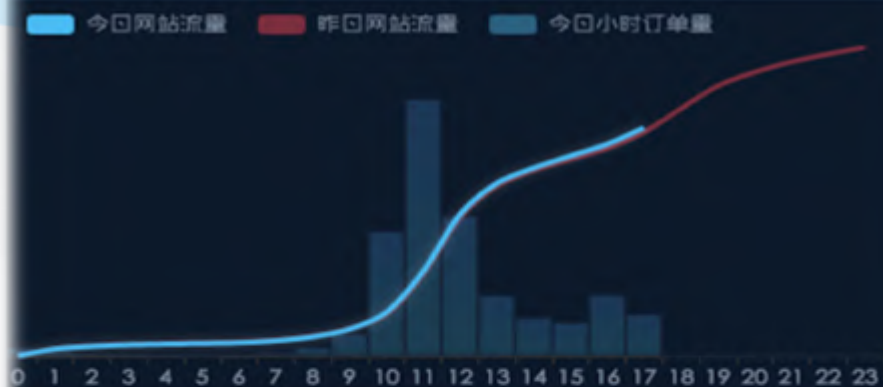


全国订单实时热力图



## 高速发展之成果

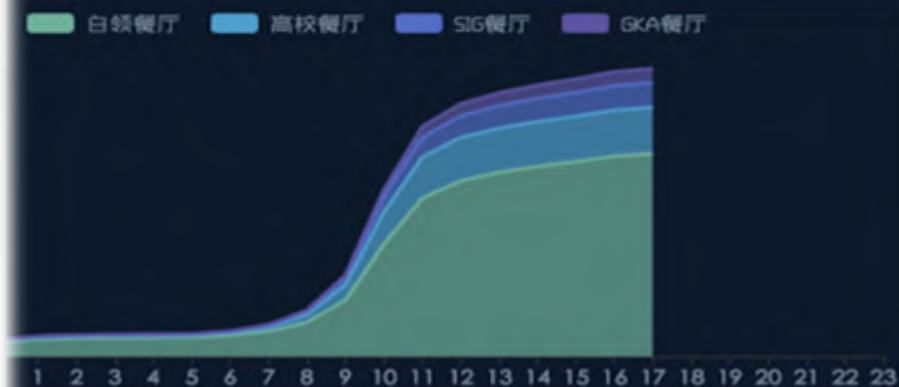
流量实时监控



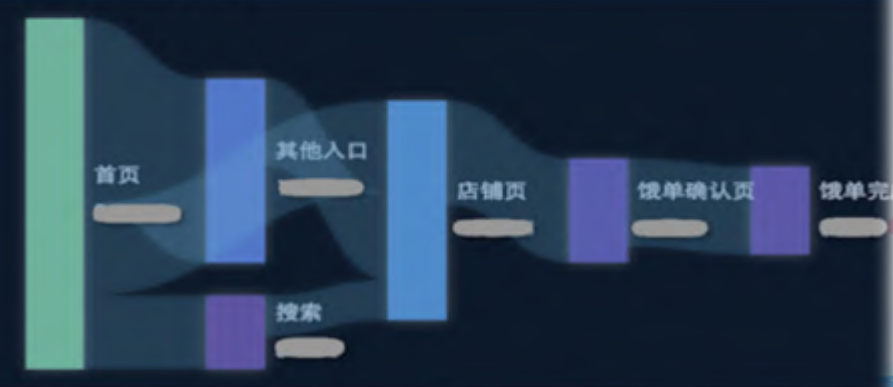
流量实时监控



交易餐厅数实时监控



流量实时路径监控



## 订单流量实时指标



## e 高速发展之总结

- 实时集群容量  
200台+ ( 7 storm集群 )
- 数据量 qps : 10w/s
- 计算量 400w/s
- 链路压力  
各组件20%以内

### 总结经验

实时链路压测

监控梳理完善

实时计算经验

### 疑难杂症

集群公用  
权限控制

平台问题

自定义监控

开发周期  
代码规范





 分享提纲

从0到1

高速发展

成熟与完善

新方向&新挑战

未来构想

Q&A



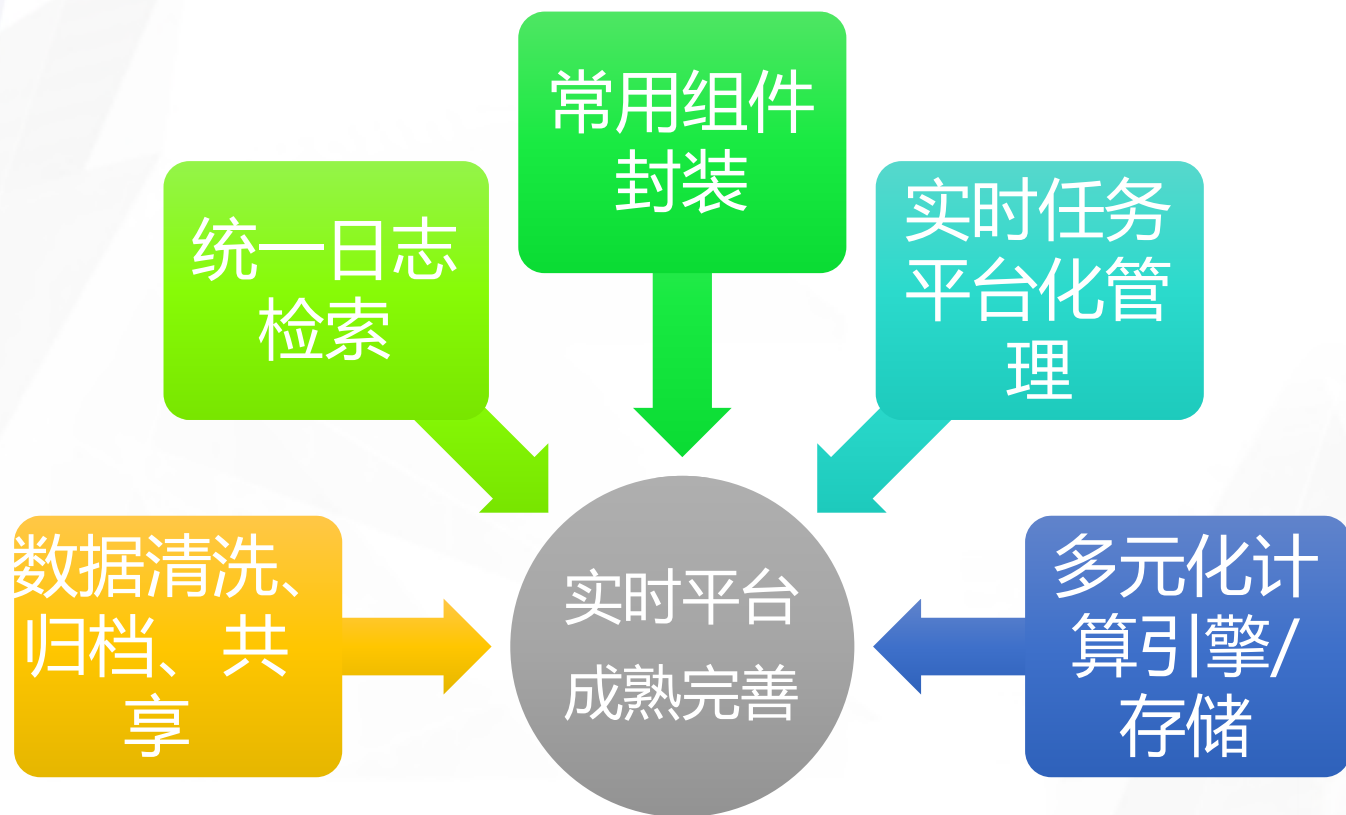
## e 成熟与完善之窘境初现

1. 实时计算平台？不就Storm吗？
2. 业务实现实时应用，如何开发、上线、监控？
  1. 消息队列里同一份数据，总是被多次重复处理？
  2. 各种乱七八糟的报警，如何屏蔽或自定义监控？
  3. 应用又报错，哪里可以快速、高效检索应用日志？

。 。 。 。 。

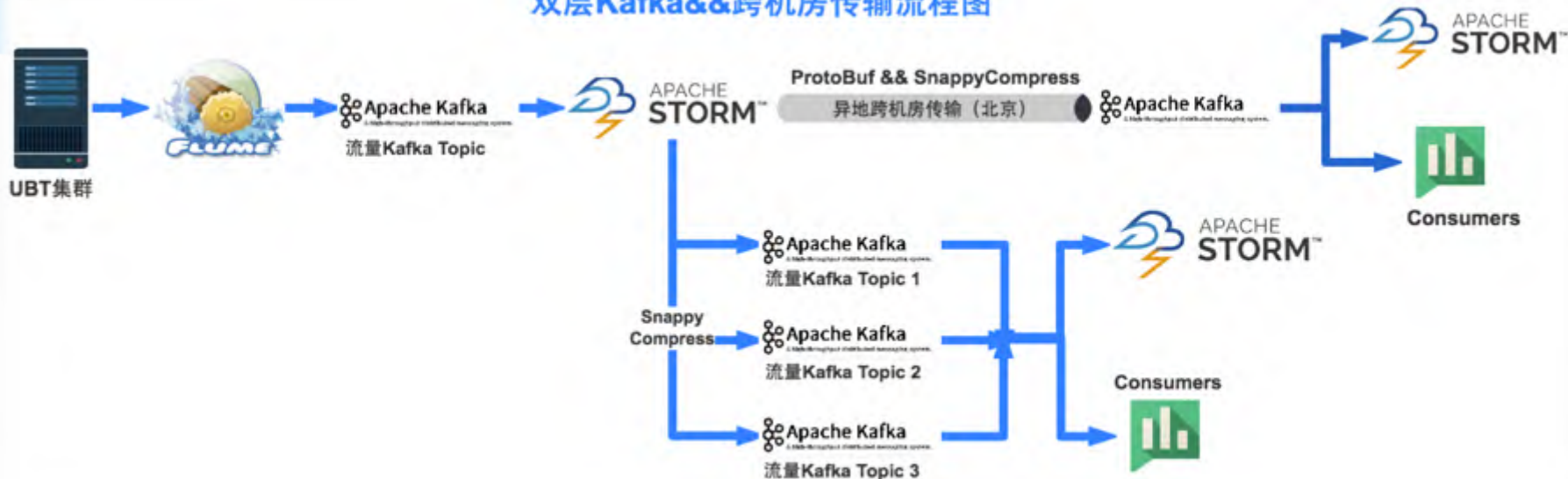


## e 成熟与完善之思考



## 成熟与完善之数据共享

双层Kafka&&跨机房传输流程图



- 清洗、分离、归档
- 数据复用
- 跨IDC传输
  - 99% Delay 2s以内
  - Protobuf + Snappy Compress





## e 成熟与完善之数据检索

### 实时数据检索



- 应用日志/监控实时收集
- 日志按应用合理限流

- Hangout VS Flume 性能对比
- 30 Flume Agent 高峰延迟10min
- 1 Hangout 10线程 数据无延迟

- Elasticsearch检索性能
- 2个集群，上海20台ssd机器
- qps : 13w/s 25亿/天
- Cpu 5% Heap Usage : 20%

- Elasticsearch监控
- Metric监控
- KOPF、HQ ...



 成熟与完善之组件封装

# Typhon

BaseSpout

BaseBolt

Eleme  
Topology

Kafka  
Spout

Mysql  
Spout

Drc  
Spout

Base  
Func  
Bolt

Base  
Redis  
Bolt

Proto  
buf  
Bolt

Delay  
Bolt

HyperL  
ogLog  
Bolt

ElasticS  
earch  
Report  
er

Influx  
db  
Repor  
ter



# 成熟与完善之平台化

饿了么 | 实际计算平台 | 任务管理 | 集群管理 | 元数据管理 | 任务成本管理 | 数据权限分配

我的任务 | BackupMan任务 | 所有任务 | 所有项目 | 项目管理

名称: 请输入任务名称

实际计算任务列表

名称	项目名称	状态	集群UI	集群	日志查询	任务类型	任务操作	系统操作
Ele_stormjob_talos_ubt_Clean	bigDataRealtime	已上线	Ele_stormjob_talos_ubt_Clean	sh-talos	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_ubtbg_Clean	bigDataRealtime	已上线	Ele_stormjob_talos_ubtbg_Clean	sh-talos	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_ubt_TestEnv	bigDataRealtime	已上线	Ele_stormjob_talos_ubt_TestEnv	sh-talos	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_rtb_FlowUV	bigDataRealtime	已上线	Ele_stormjob_talos_rtb_FlowUV	sh-talos	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_rtb_HummingbirdGraph	bigDataRealtime	已下线	Ele_stormjob_rtb_HummingbirdGraph	sh-dt	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_bigsafe_FlowUV	bigDataRealtime	已上线	Ele_stormjob_talos_bigsafe_FlowUV	sg-talos-LD1	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_ubtbg_ParseProtoBuf	bigDataRealtime	已上线	Ele_stormjob_talos_ubtbg_ParseProtoBuf	sg-talos-LD1	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_talos_rtb_HummingbirdGraph	bigDataRealtime	已上线	Ele_stormjob_talos_rtb_HummingbirdGraph	sh-talos	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_ubt_AppMonitor	bigDataRealtime	已上线	Ele_stormjob_ubt_AppMonitor	sh-dt	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警
Ele_stormjob_rtb_FlowUV	bigDataRealtime	已上线	Ele_stormjob_rtb_FlowUV	sh-dt	Log Link	storm	kill   发布   更新   管理   详情   删除   设置   告警	kill   发布   更新   管理   详情   删除   设置   告警

ES日志链接

版本管理

任务列表

- Ele\_stormjob\_rtb\_FlowUV
- Ele\_stormjob\_talos\_rtb\_HummingbirdGraph
- Ele\_stormjob\_talos\_ubt\_TestEnv
- Ele\_stormjob\_talos\_ubtbg\_ParseProtoBuf
- Ele\_stormjob\_talos\_ubtbg\_Clean
- Ele\_stormjob\_talos\_ubt\_Clean
- Ele\_stormjob\_rtb\_HummingbirdGraph
- Ele\_stormjob\_rtb\_FlowUV
- Ele\_stormjob\_talos\_bigsafe\_FlowUV
- Ele\_stormjob\_talos\_rtb\_HummingbirdGraph
- Ele\_stormjob\_talos\_ubt\_AppMonitor
- Ele\_stormjob\_talos\_ubt\_TestEnv
- Ele\_stormjob\_talos\_ubtbg\_ParseProtoBuf
- Ele\_stormjob\_talos\_ubtbg\_Clean
- Ele\_stormjob\_talos\_ubt\_Clean
- Ele\_stormjob\_rtb\_HummingbirdGraph
- Ele\_stormjob\_rtb\_FlowUV

参数名称	参数值
topology workers	15
topology actor executors	3
topology max.zoo.out.pending	1000
kafka fetch.size	11534336
kafka flag	0
kafka zk.host	sh-hadoop-cluster-250-13.elastic.mech...
kafka topic	mobile-bufc
spark.parallelism	30
zooServers	sh-hadoop-cluster-250-13.elastic.mech...
zkPort	2181
zooKeeper.parallelism	1
client.rpc.parallelism	50
normal.zoo.parallelism	12
high.zoo.parallelism	30
redis.cluster.hosts	sh-hadoop-mid6-250-180.elastic.mech...
redis.cluster.port	4000/1,40015
redis.conf.paths	sh-hadoop-128-115.elastic.mech...

- 项目任务管理、权限控制
- 多版本控制，方便回滚
- 发布前置检测





## 成熟与完善之平台化



- 实时任务Metric监控指标
- 自定义监控

### 告警设置

· 邮件地址

· 电话号码

· 监控项

· snooze

· 是否启动

关闭 更新告警



# 成熟与完善之平台化

饿了么·实时计算平台 | 任务管理 | 集群管理 | 元数据管理 | 任务版本管理 | 实时资源分配

### Kafka 集群列表

ID	Topic Name	Partitions	State	Producer	Consumer
31	amelia_msg_status	10	健康状态	生产者	消费者
32	grace-historylog	30	健康状态	生产者	消费者
33	eleme-crawler	1	健康状态	生产者	消费者
34	amelia_push_alpha	30	健康状态	生产者	消费者
35	mobile-pubt	60	健康状态	生产者	消费者
36	logstash-kafka-storm-talos	30	健康状态	生产者	消费者
37	_consumer_offsets	50	健康状态	生产者	消费者
38	fengniao-team-delivery-talos	30	健康状态	生产者	消费者
39	mobile-ubt	30	健康状态	生产者	消费者
40	fengniao-crowd-delivery-talos	30	健康状态	生产者	消费者
41	eleme-order-talos	30	健康状态	生产者	消费者
42	alliance-order-test	30	健康状态	生产者	消费者
43	amelia_msg_status_alpha	10	健康状态	生产者	消费者
44	zabbix-alerts	30	健康状态	生产者	消费者
45	amelia_push_prod	30	健康状态	生产者	消费者
46	logstash-kafka	30	健康状态	生产者	消费者

**topic操作**

### 数据探查 - mobile-pubt

输出结果

ID	Content
1	183.206.163.59 -- [06/Oct/2016:13:34:46 +0800] "POST /collect/log HTTP/1.1" 200 5 "-" "Rajax/1 Apple/Phone7.2 iPhone_OS/9.2.1 me.ele.ios/6.3.1 ID/9E80FC12-7266-42E2-8166-E8889C80501; isjaibroken/0" 0.003 {vx22log/vx22:/vx22city_id/vx226/vx22params/vx22:/vx22tag/vx...
2	223.104.14.10 -- [06/Oct/2016:13:34:44 +0800] "POST /collect/log HTTP/1.1" 200 5 "-" "Rajax/1 HM_NOTE_15/gucci Android/4.4.4 Display/KTU84P me.ele/6.3 ID/c073b82c-51d0-3e30-8ed4-7628ae18b4b6; KERNEL_VERSION:3.10.28-gd8c9fb2 API_Level:19" 0.000 {vx22log/vx22:/vx22c...
3	58.209.185.84 -- [06/Oct/2016:13:34:46 +0800] "POST /collect/log HTTP/1.1" 200 5 "-" "Rajax/1 HM_NOTE_15/gucci Android/4.4.4 Display/KTU84P me.ele/6.3 ID/c073b82c-51d0-3e30-8ed4-7628ae18b4b6; KERNEL_VERSION:3.10.28-gd8c9fb2 API_Level:19" 0.000 {vx22log/vx22:/vx22c...
4	183.167.211.28 -- [06/Oct/2016:13:34:45 +0800] "POST /collect/log HTTP/1.1" 200 5 "-" "Rajax/1 HM_NOTE_15/gucci Android/4.4.4 Display/KTU84P me.ele/6.3 ID/c073b82c-51d0-3e30-8ed4-7628ae18b4b6; KERNEL_VERSION:3.4.0-g1e451e8-0494" 0.000 {vx22log/vx22:/vx22c...
5	117.136.5.27 -- [06/Oct/2016:13:34:45 +0800] "POST /collect/log HTTP/1.1" 200 5 "-" "Rajax/1 HM_NOTE_15/gucci Android/4.4.4 Display/KTU84P me.ele/6.3 ID/c073b82c-51d0-3e30-8ed4-7628ae18b4b6; KERNEL_VERSION:3.10.72+ API_Level:22" 0.000 {vx22log/vx22:/vx22c...

- 数据源探查
- 实时集群管理
- 容量监控

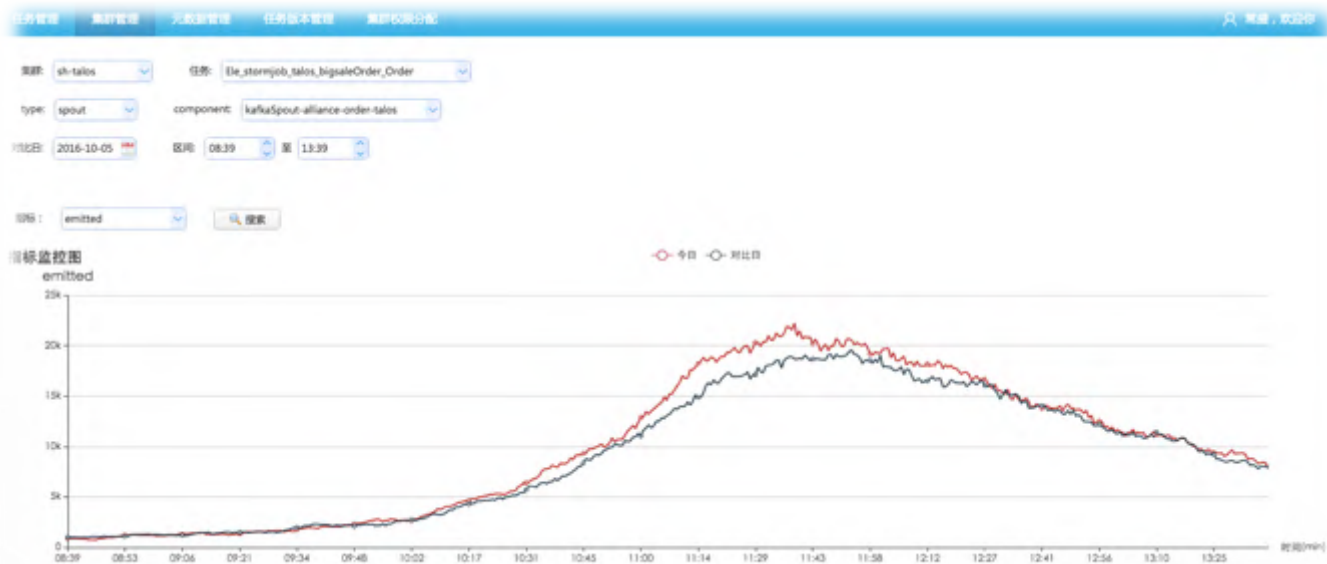
实例	负载	CPU (B)	CPU (R)	CPU (D)	JVM 负载	可用内存 (GB)	总内存 (GB)	内存量 (MB)	CPU 温度 (MB)	最大内存速率
sh-node-supervisor-128-105	1	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-104	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-110	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-114	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-23	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-27	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-108	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-113	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-22	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-26	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-106	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-112	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-21	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-25	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-29	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-107	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-128-111	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-20	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-24	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00
sh-node-supervisor-132-28	0	2	2	2	2	94.00	94.00	2.00	2.00	2.00



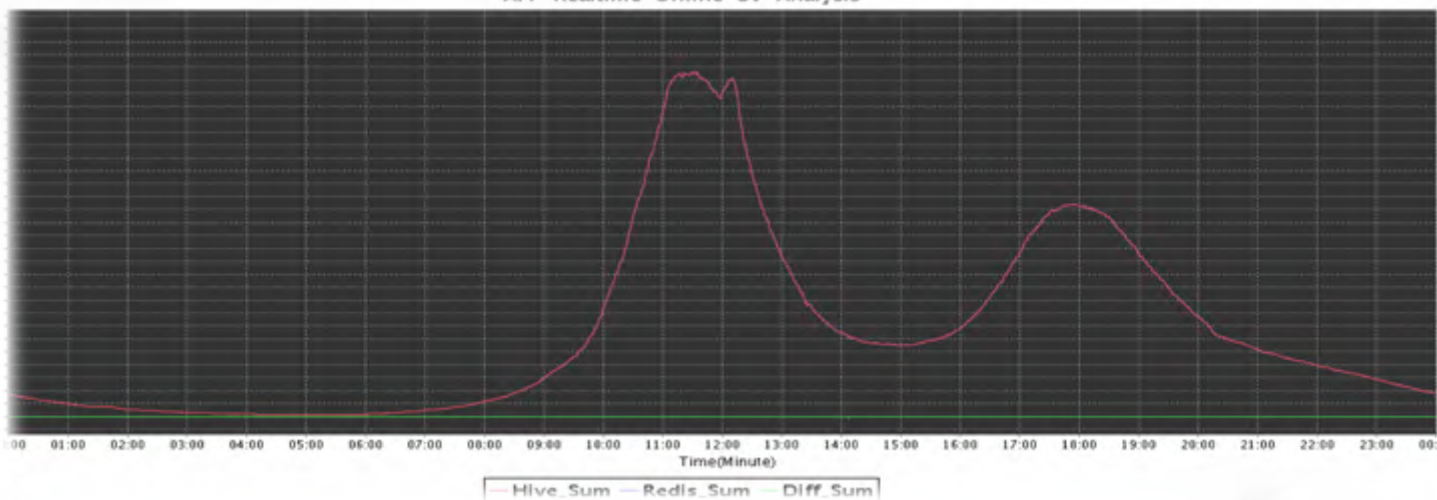


## e 成熟与完善之平台化

- 实时任务指标管理
- Binlog Parser 管理
- 数据质量管理



AFF-Realtime-Offline-UV-Analysis



## e 成熟与完善之多元化

### Grace ( 监控离线性能 )

### Redis时间响应系统

数据源 :



数据收集 :



消息队列 :



实时平台 :

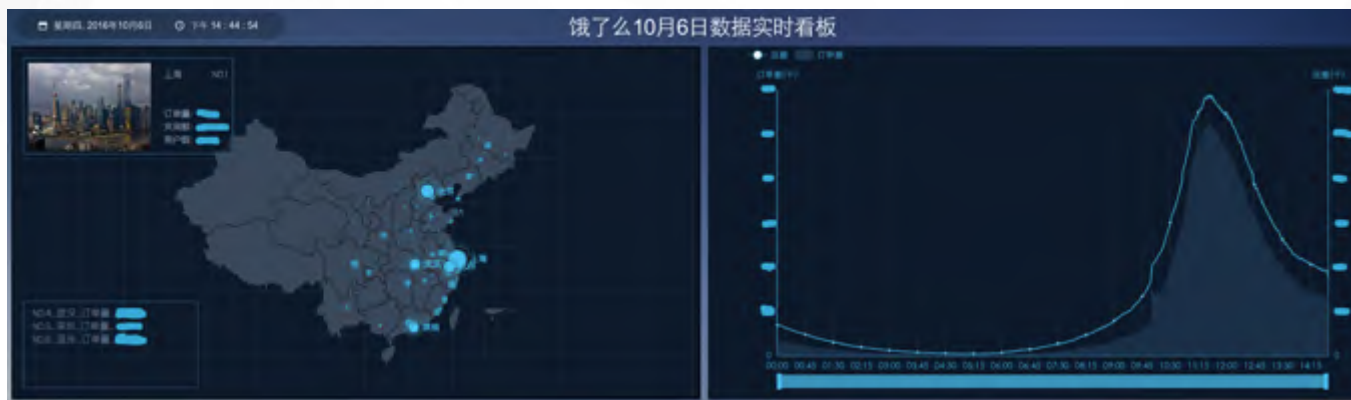


数据存储 :



## e 成熟与完善之应用

- 热卖美食实时推荐
- 猜你喜欢实时特征提取
- 准时达实时定位
- 饿了么实时看板指标
- 实时风控





 分享提纲

从0到1

高速发展

成熟与完善

新方向&新挑战

未来构想

Q&A





## e 新方向之选型

BI同学想说：  
如何快速实时数据报表&&实时数据分析

StreamCQL

PK

STORMSQL

实时计算同学想说：  
计算框架能否更快、更准确、更完美



PK



# 新方向之Flink实践

Overview Version: 1.1.0 Commit: 407565

Task Managers	4
Task Slots	80
Available Task Slots	60

Total Jobs	
Running	4
Finished	0
Canceled	2
Failed	0

### Running Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
2016-10-16, 18:14:37	2016-10-16, 18:52:07	37m 29s	Flink Streaming Job	4e0936cc40476cc3b67d64934e562d74	1 0 0 0 0 0 0	RUNNING
2016-10-16, 18:29:33	2016-10-16, 18:52:07	22m 33s	Flink Streaming Job	8110b6f5a52a16cd80e33e137ace762	18 0 18 0 0 0 0	RUNNING
2016-10-16, 18:48:56	2016-10-16, 18:52:07	2m 10s	multi-dim-uv	a37319814cfe9e225cd2f747e2508	17 0 17 0 0 0 0	RUNNING
2016-10-16, 18:51:01	2016-10-16, 18:52:07	1m 5s	binlog2json	670fa0375858478f1cd9567195183f	2 0 2 0 0 0 0	RUNNING

### Completed Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
2016-10-14, 14:38:31	2016-10-16, 18:03:42	4d 3h	Flink Streaming Job	41b661011de18518e8ad10c2d099ea7a	0 0 0 0 0 0 0	CANCELED
2016-10-14, 17:29:41	2016-10-14, 17:30:34	52s	Flink Streaming Job	dbcc0540f0e48642acd73e1a4c0b30e1	0 0 0 0 0 0 0	CANCELED

## Flink Features :

- 高吞吐、低延迟
- 支持 Event Time
- exactly-once
- 高度灵活的流式窗口
- 迭代和增量迭代



 分享提纲

从0到1

高速发展

成熟与完善

新方向&新挑战

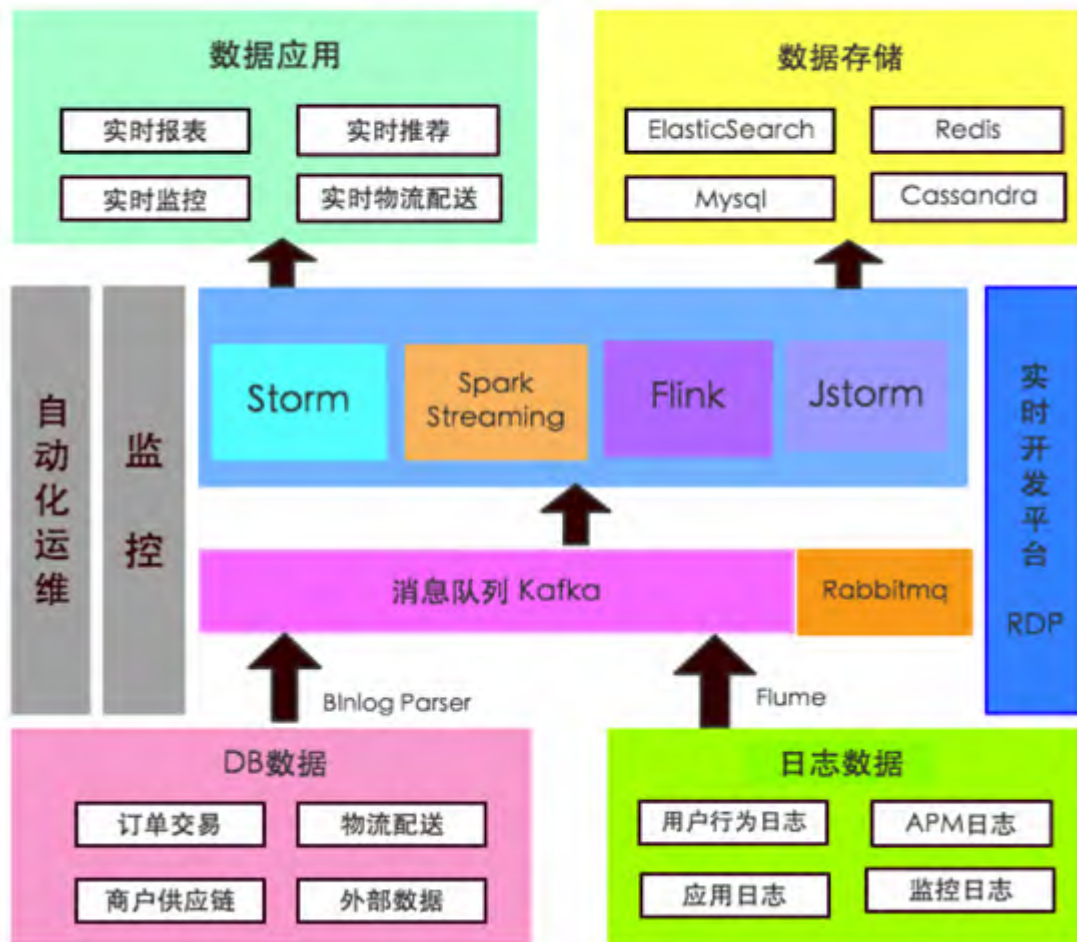
未来构想

Q&A



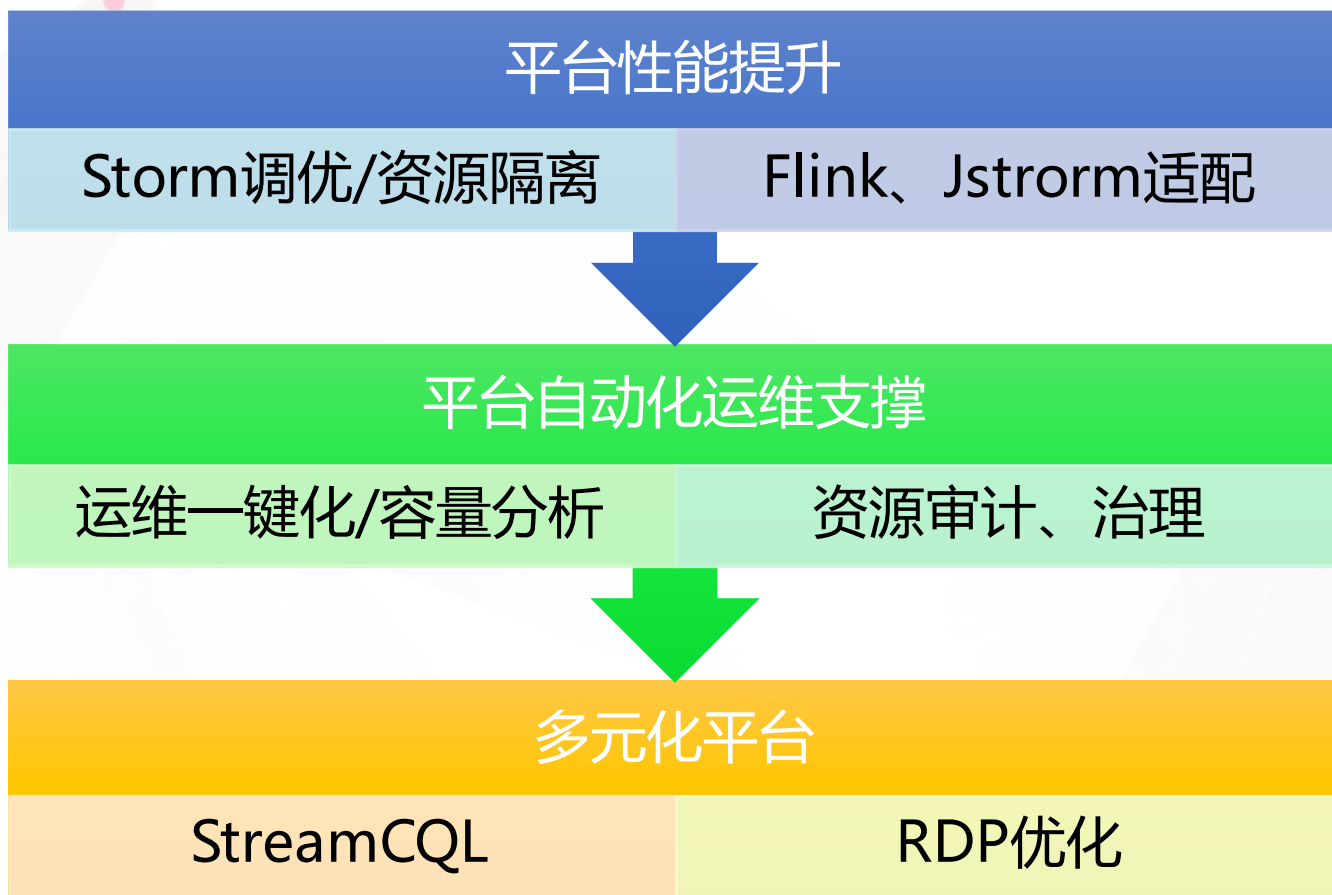
# 不断完善的平台框架图

## 饿了么实时框架规划图





## e 不断完善平台之Next



# Q&A

# THANKS

SequeMedia  
盛拓传媒

IT168.com  
中国网络 16 年

ChinaUnix

ITPUB  
www.itpub.net