SACC 2016 第八届中国系统架构师大会
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 | 美好生活 触手可得

饿了么平台

日均交易额
2亿元+

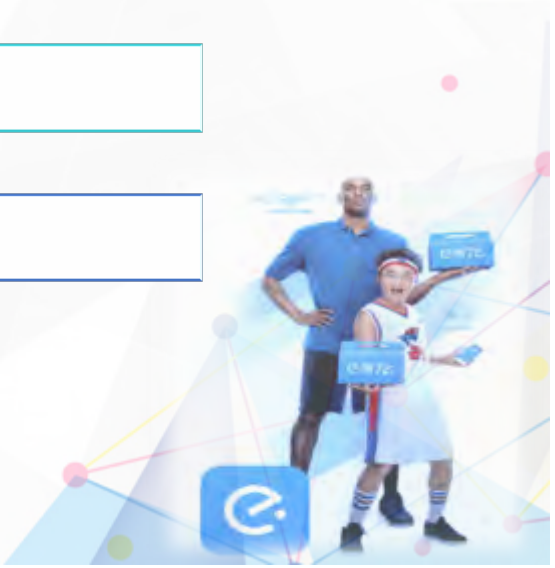
日均订单量
500w+

覆盖城市
1000+

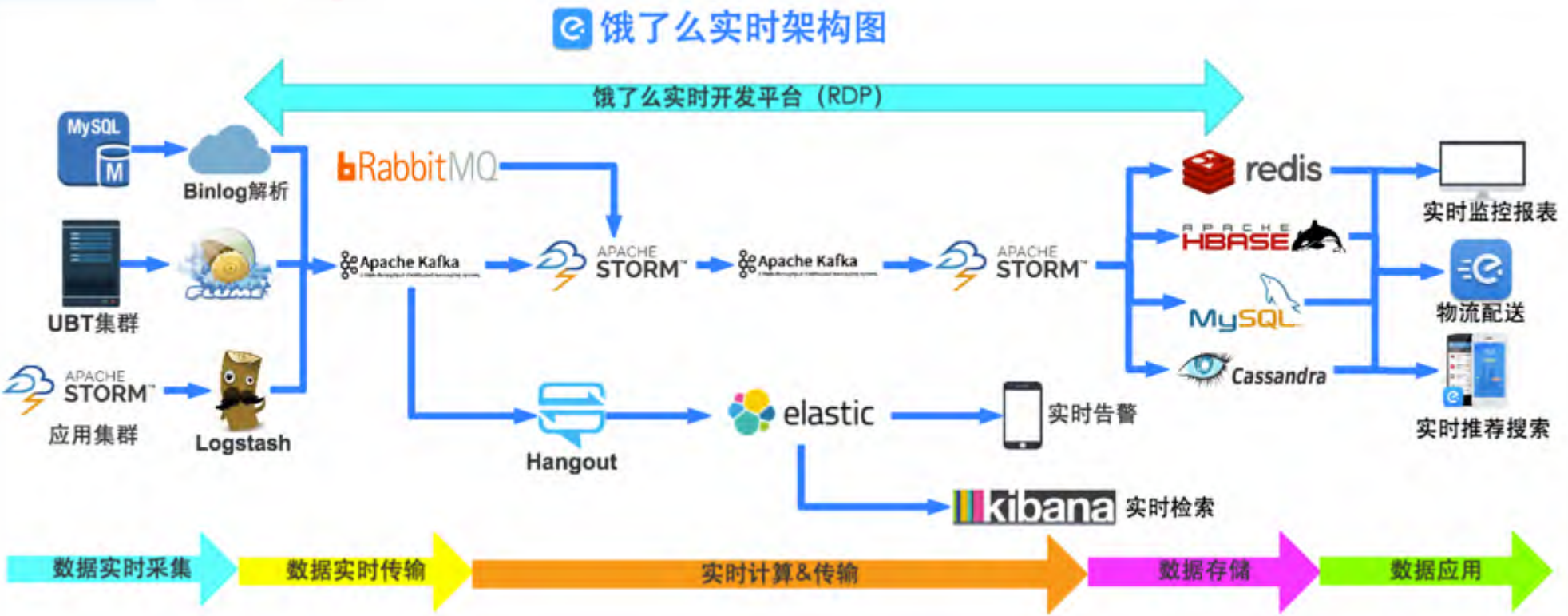
员工数
15000+

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



每一个组件也许都踩过N次的坑。。。

从0到1的实践

数据源：

+

数据收集：

python PK FLUME

消息队列：

Apache Kafka
A distributed streaming platform.

实时引擎：

APACHE STORM™ PK Spark Streaming

数据存储：

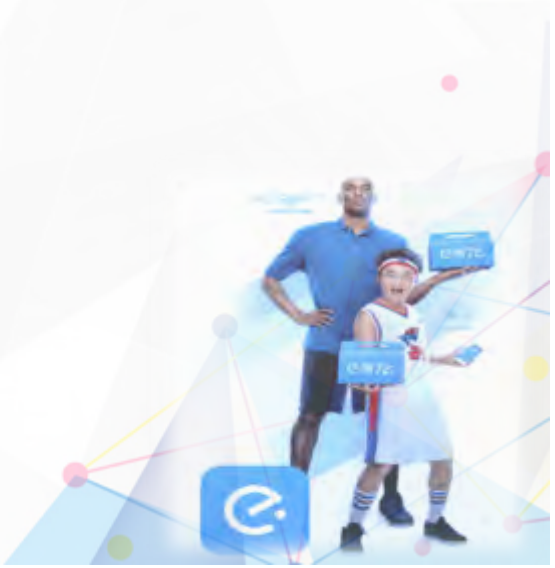
redis

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）

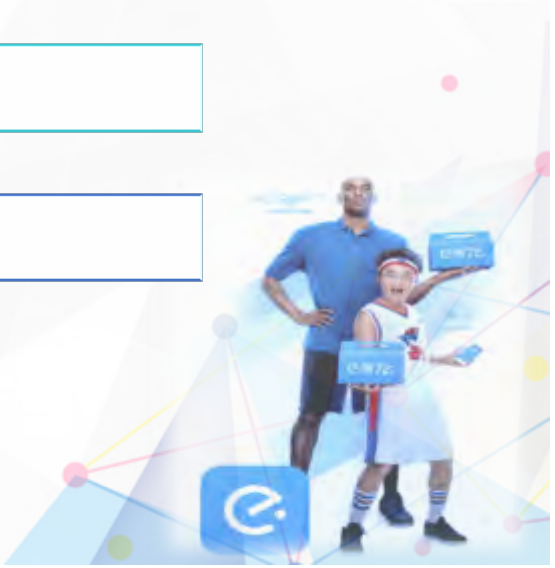


从0到1的总结

- 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



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....

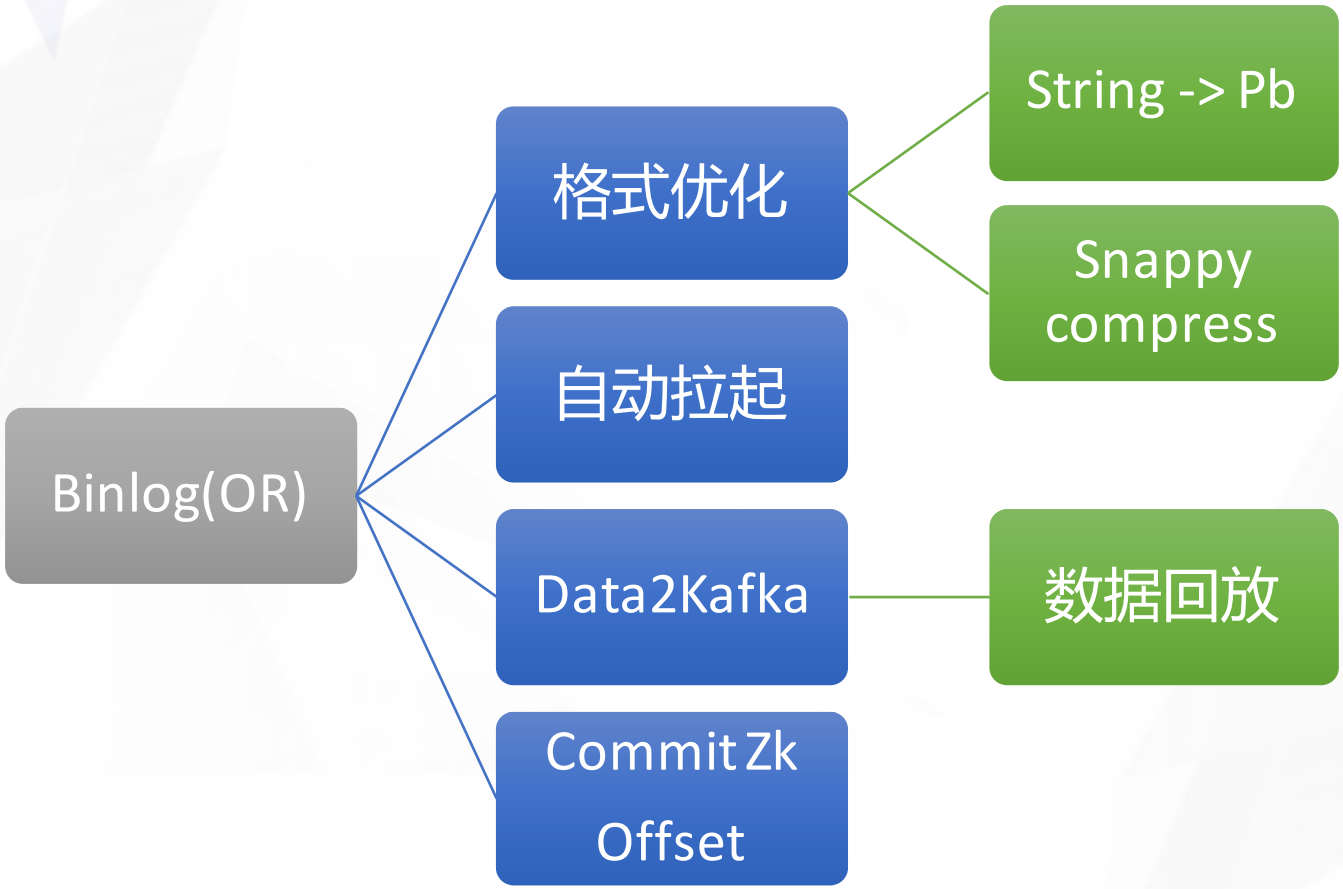
5、资源：机器&人力、经验匮乏...

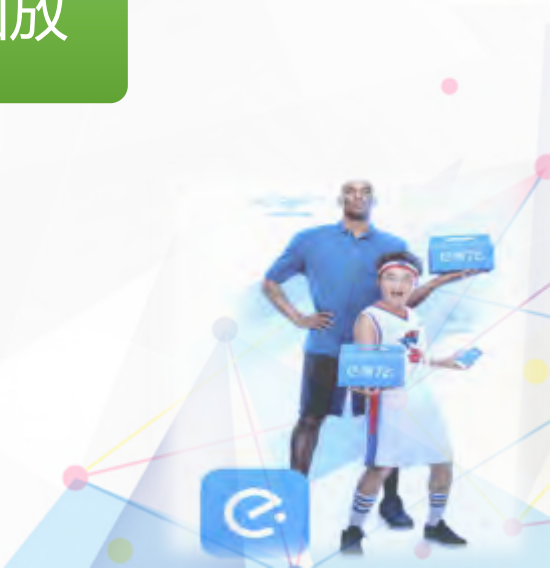
高速发展之行动

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

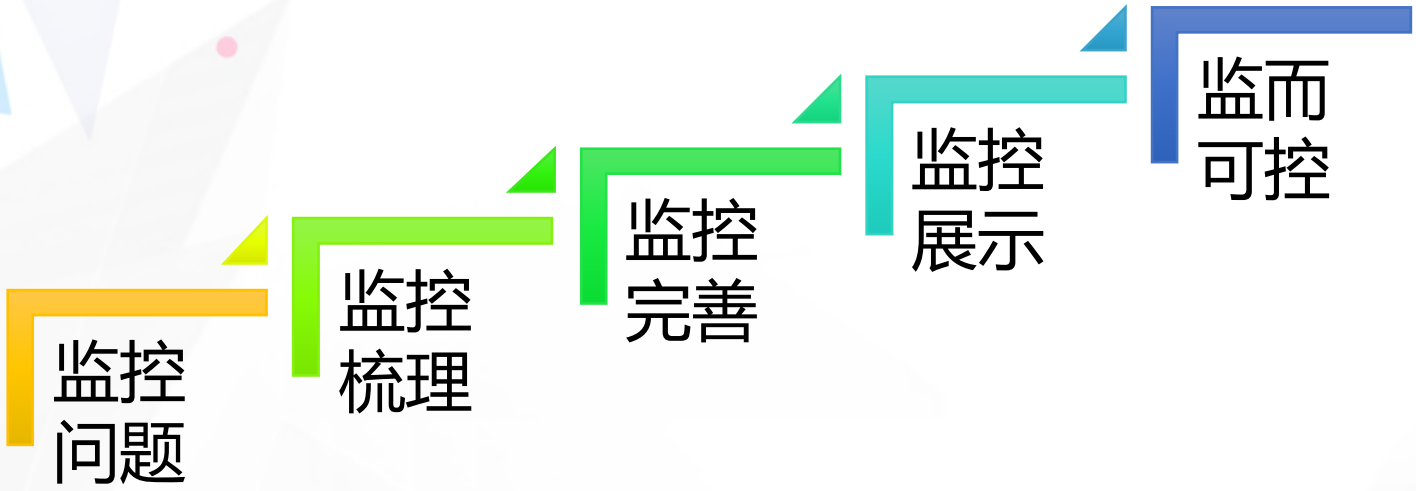

Binlog(OR)
格式优化
String -> Pb
Snappy compress
自动拉起
Data2Kafka
数据回放
Commit Zk Offset



高速发展之应用优化

JVM HLL方案	双链路高可用方案	实时链路梳理压测
• 存储方式优化 • qps降低4倍	• 链路双写 • 应用自动降级 • 数据自动恢复	• 延迟毫秒级 • Storm Heatbeat调整 • 灾备演练

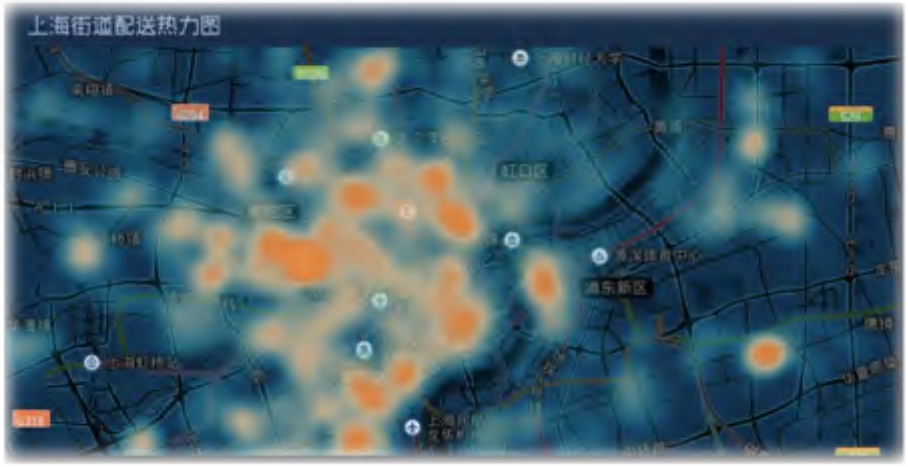
高速发展监控优化



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

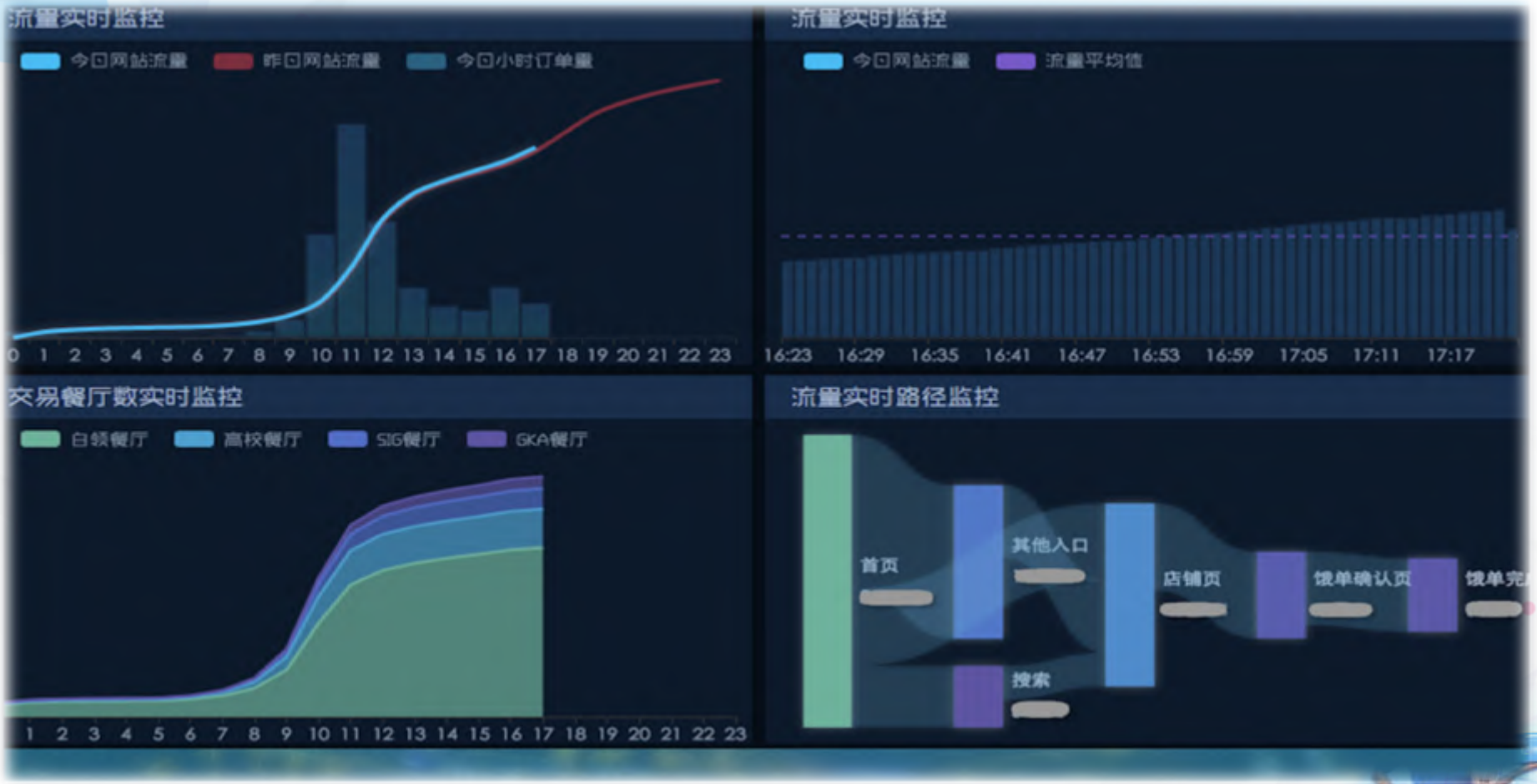
用户行为实时测试系统



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



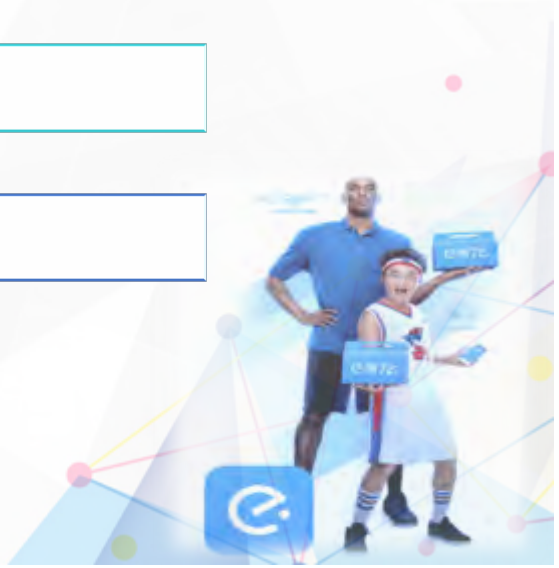
订单流量实时指标

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

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

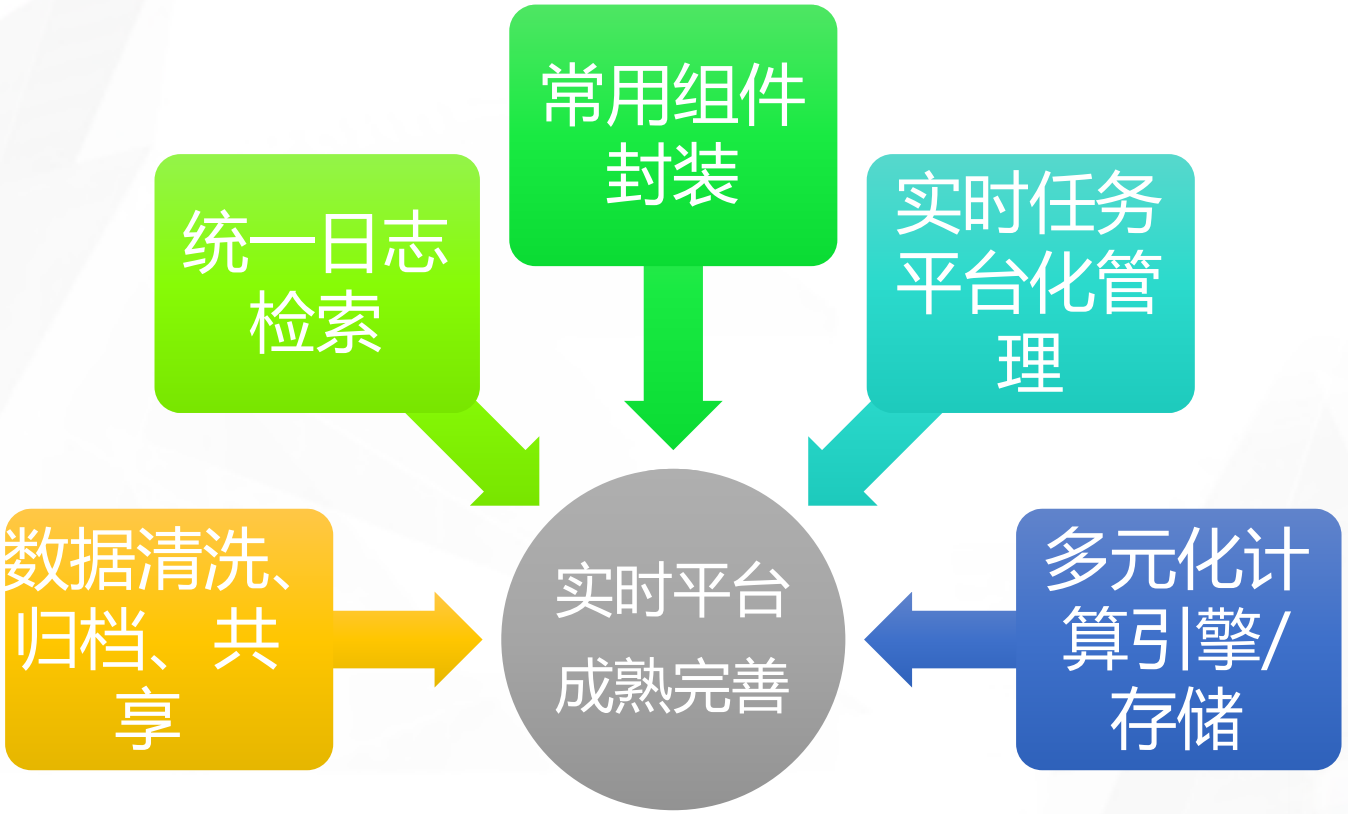


成熟与完善之窘境初现

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？

。。。。。。

成熟与完善之思考



成熟与完善之数据共享

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

成熟与完善之数据检索

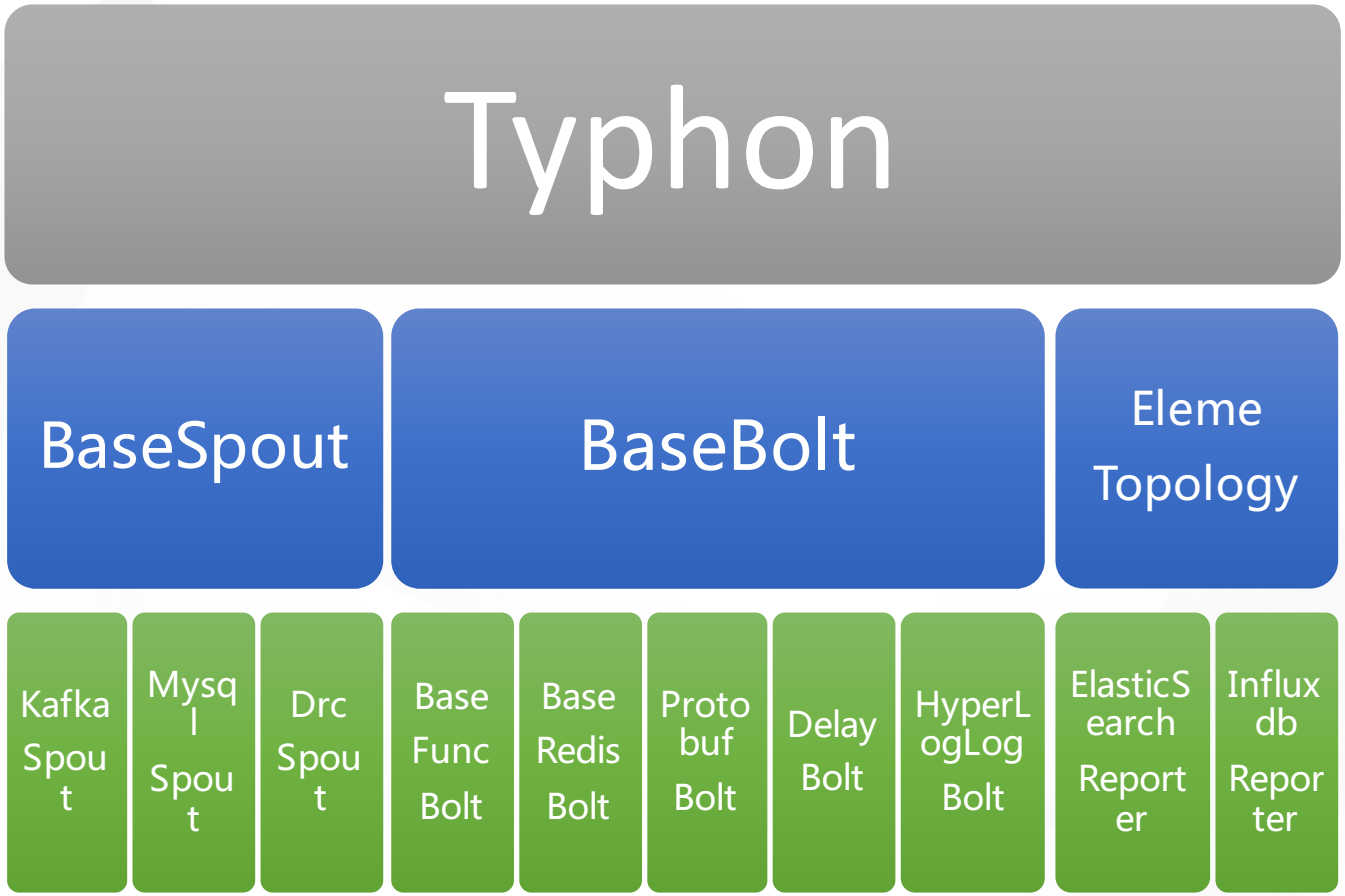
- 应用日志/监控实时收集
  - 日志按应用合理限流

- 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

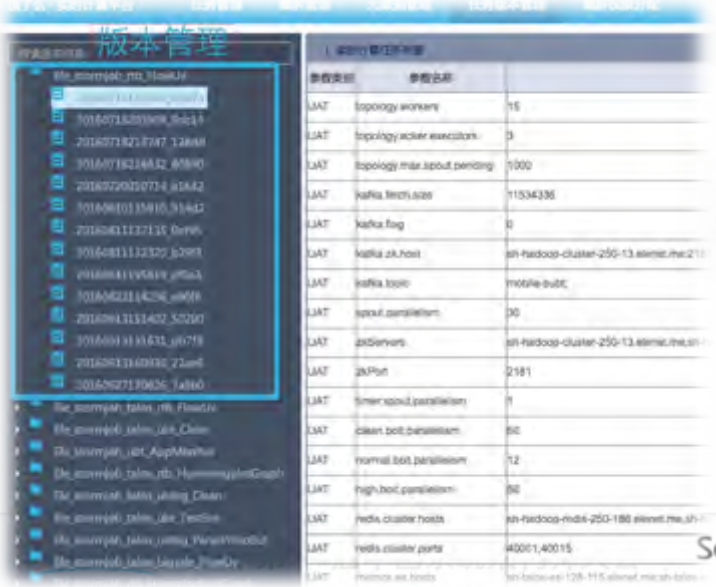
- ElasticSearch检索性能
  - 2个集群，上海20台ssd机器
  - qps : 13w/s 25亿/天
  - Cpu 5% Heap Usage：20%

- 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
Protobuf Bolt
Delay Bolt
HyperLogLog Bolt
ElasticSearch Reporter
Influxdb Reporter

成熟与完善之平台化

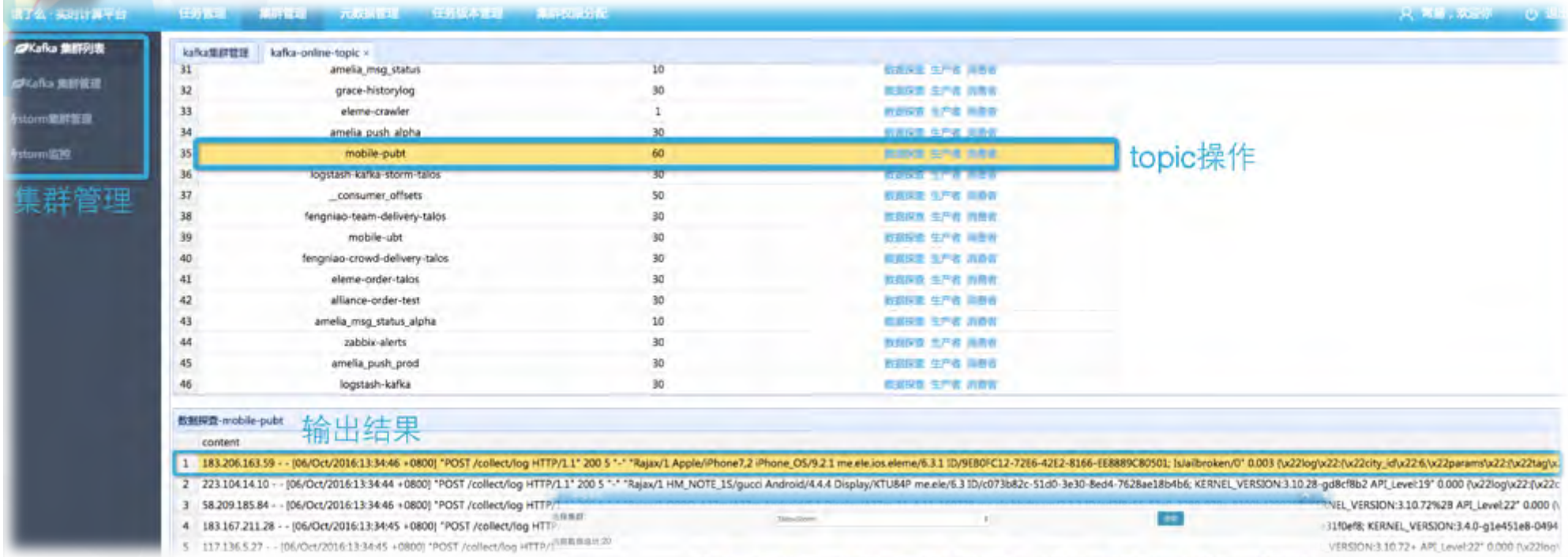


- 项目任务管理、权限控制
- 多版本控制，方便回滚
- 发布前置检测

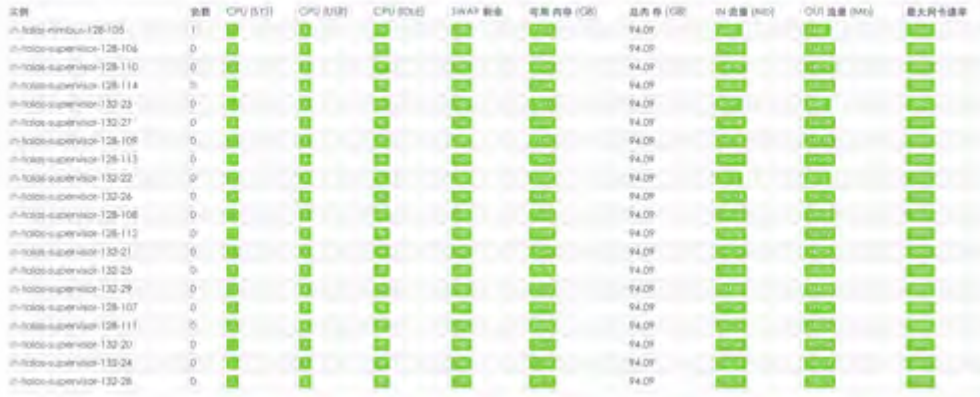
成熟与完善之平台化

- 实时任务Metric监控指标
- 自定义监控

成熟与完善之平台化

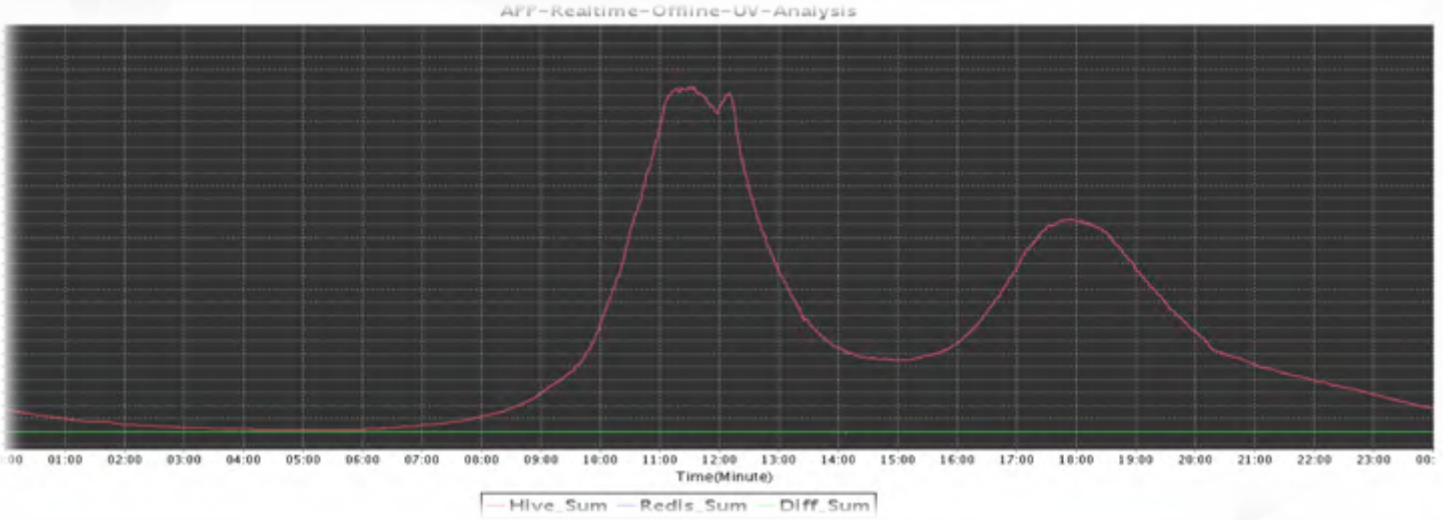


- 数据源探查
- 实时集群管理
- 容量监控



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



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：

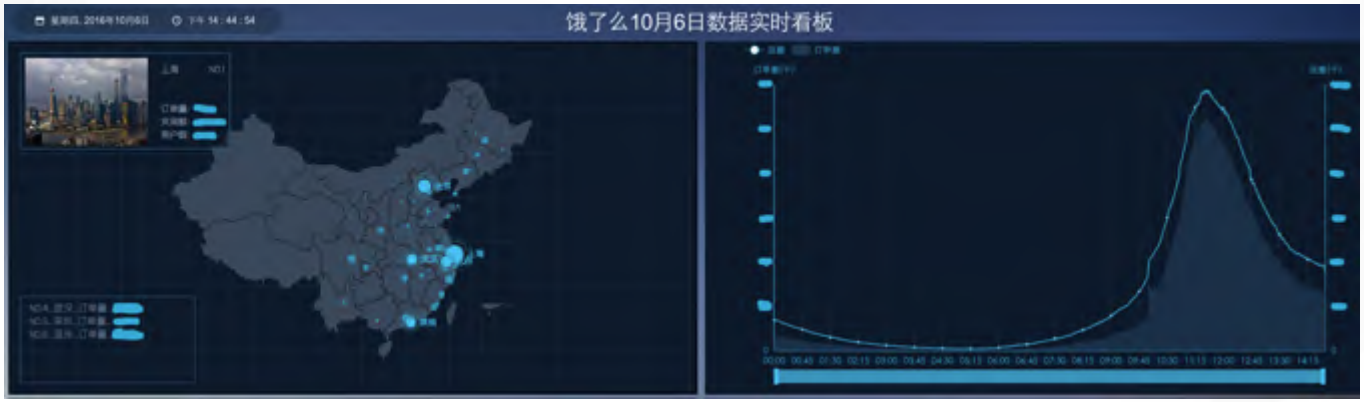
消息队列：

实时平台：

数据存储：

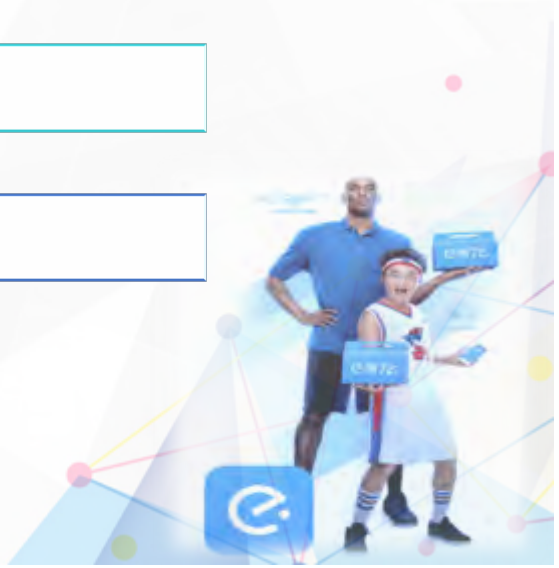
成熟与完善之应用

- 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



新方向之选型

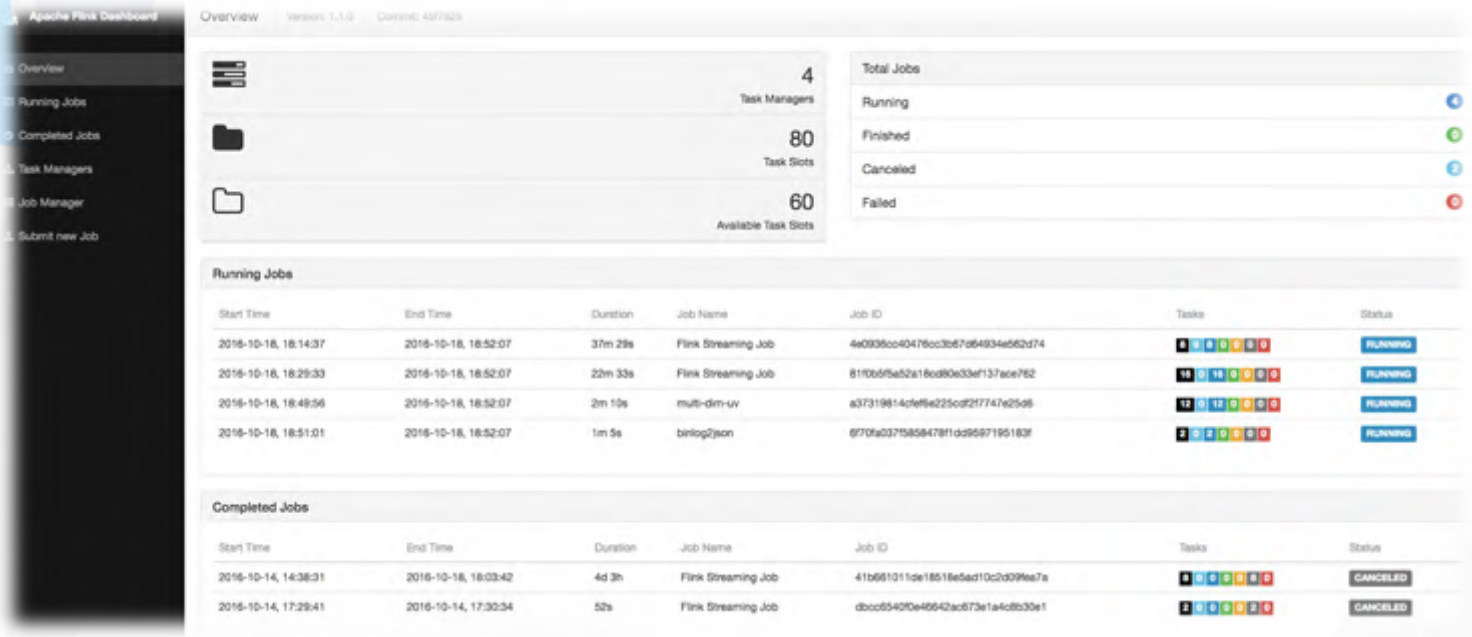
BI同学想说：
如何快速实时数据报表&&实时数据分析

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

Flink PK JStorm

新方向之Flink实践

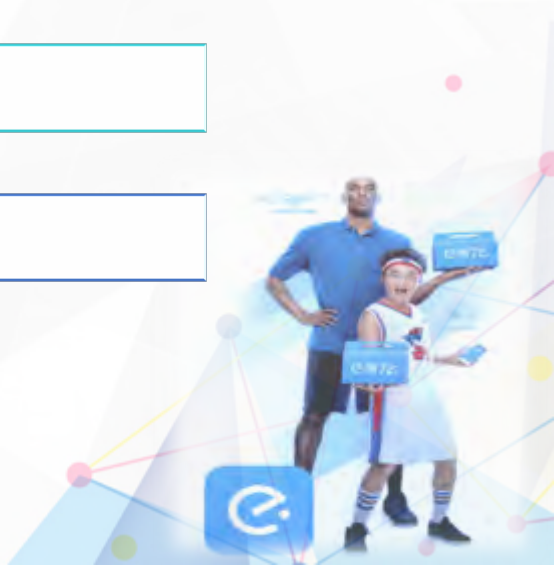


Flink Features：

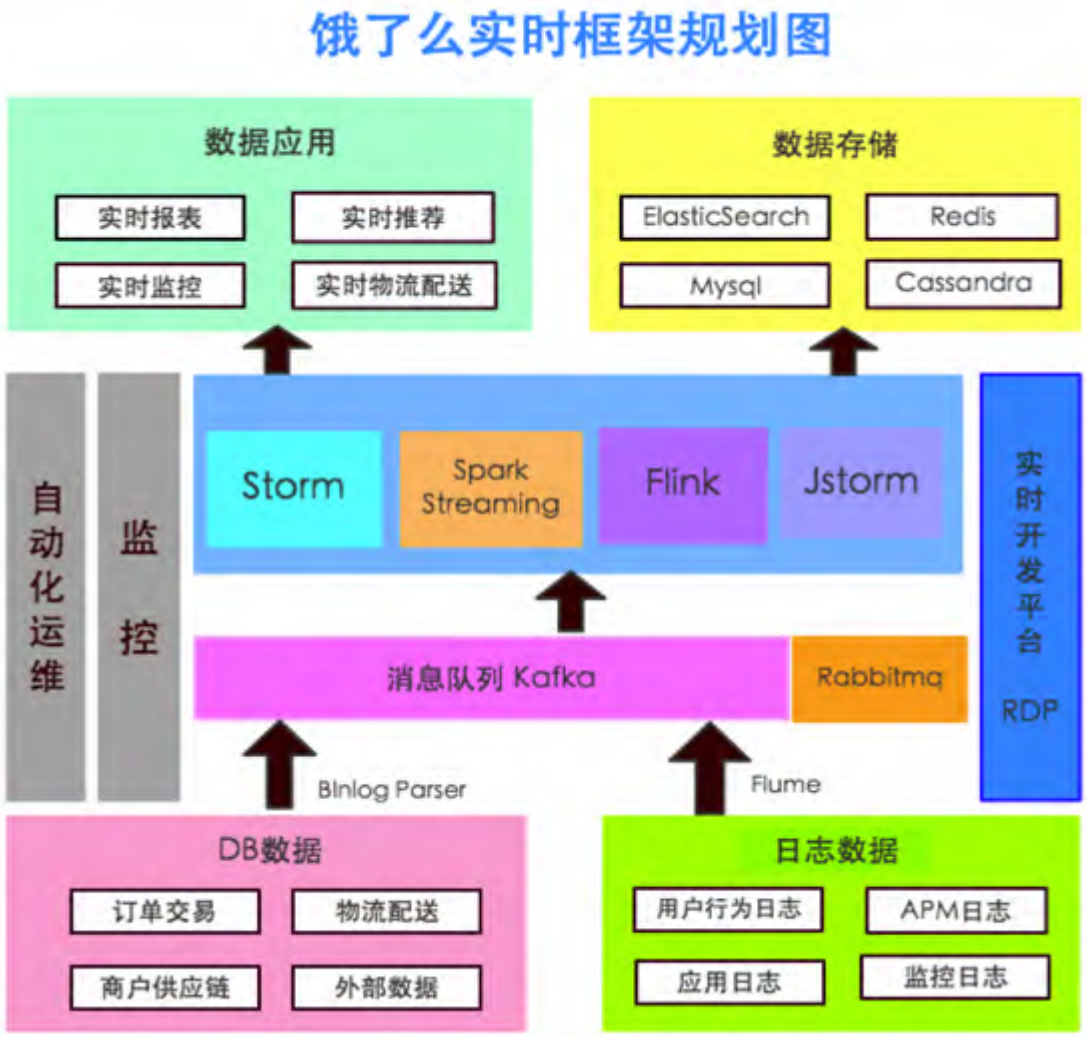
- 高吞吐、低延迟
- 支持 Event Time
- exactly-once
- 高度灵活的流式窗口
- 迭代和增量迭代

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
数据应用
实时报表
实时推荐
实时监控
实时物流配送
数据存储
ElasticSearch
Redis
Mysql
Cassandra
自动化运维
监控
Storm
Spark Streaming
Flink
Jstorm
实时开发平台
RDP
消息队列 Kafka
Rabbitmq
Binlog Parser
Flume
DB数据
订单交易
物流配送
商户供应链
外部数据
日志数据
用户行为日志
APM日志
应用日志
监控日志

不断完善平台之Next

平台性能提升

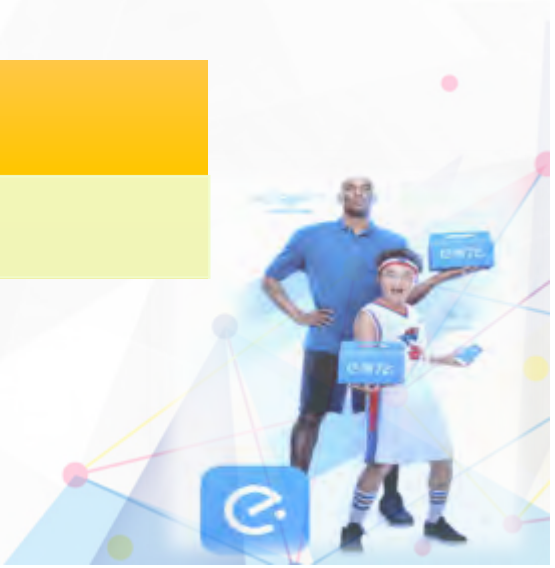
- Storm调优/资源隔离
- Flink、Jstrorm适配

平台自动化运维支撑

- 运维一键化/容量分析
- 资源审计、治理

多元化平台

- StreamCQL
- RDP优化



Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 | ChinaUnix | ITPUB