Fintech场景下大数据处理的挑战与实践

徐佳晶 @ 人人贷互联网信贷事业群

GITC

Nov. 2017

AGENDA

01 我看互金这6年

- 业务 / 获客方式的转变
- 用户数、交易数的激增
- 风控思维的转变

02 风控：传统金融 VS Fintech

- 人 VS 机器
- 评分卡 VS 模型
- 从业人员skillset

03 技术团队面临的挑战

- 数据量
- 计算复杂度
- 服务可靠性

04 经验 & 实践

- 由一起线上事故说起
- Kafka
- HBase
- 其它

05 再过三五年……

- 行业
- 政策
- 团队
- 技术

业务 / 获客方式的转变

线下网点，业务人员地推

插卡、陌拜、线下活动.....

- 开设线下门店，配置业务人员
- 增加门店、提高人均产能
- 核心业务系统

电销

电话外呼

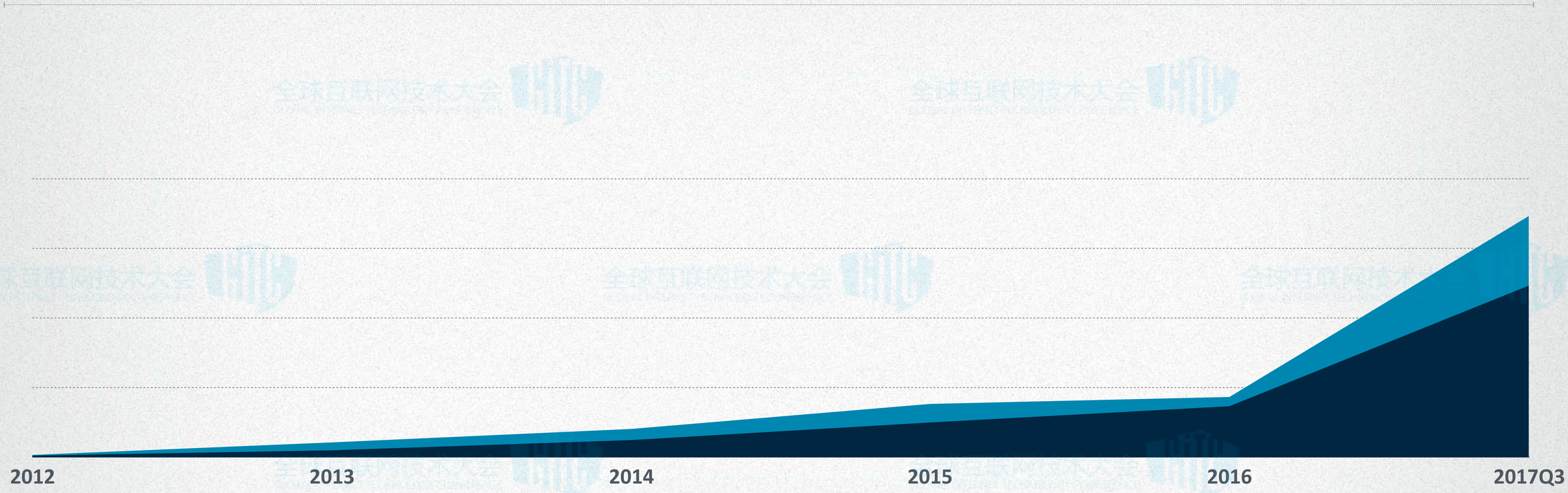
- 客户名单获取
- 扩大规模、提高名单质量、提升电销人员效率、优化外呼策略
- CRM

互联网方式

渠道、合作、流量交换

- 更偏向互联网获客模式，导流、引流、精准客户营销、投放
- 提高转化率、合作渠道数量与质量
- 中间件、系统群、云、大数据环境.....

用户数、交易数的激增



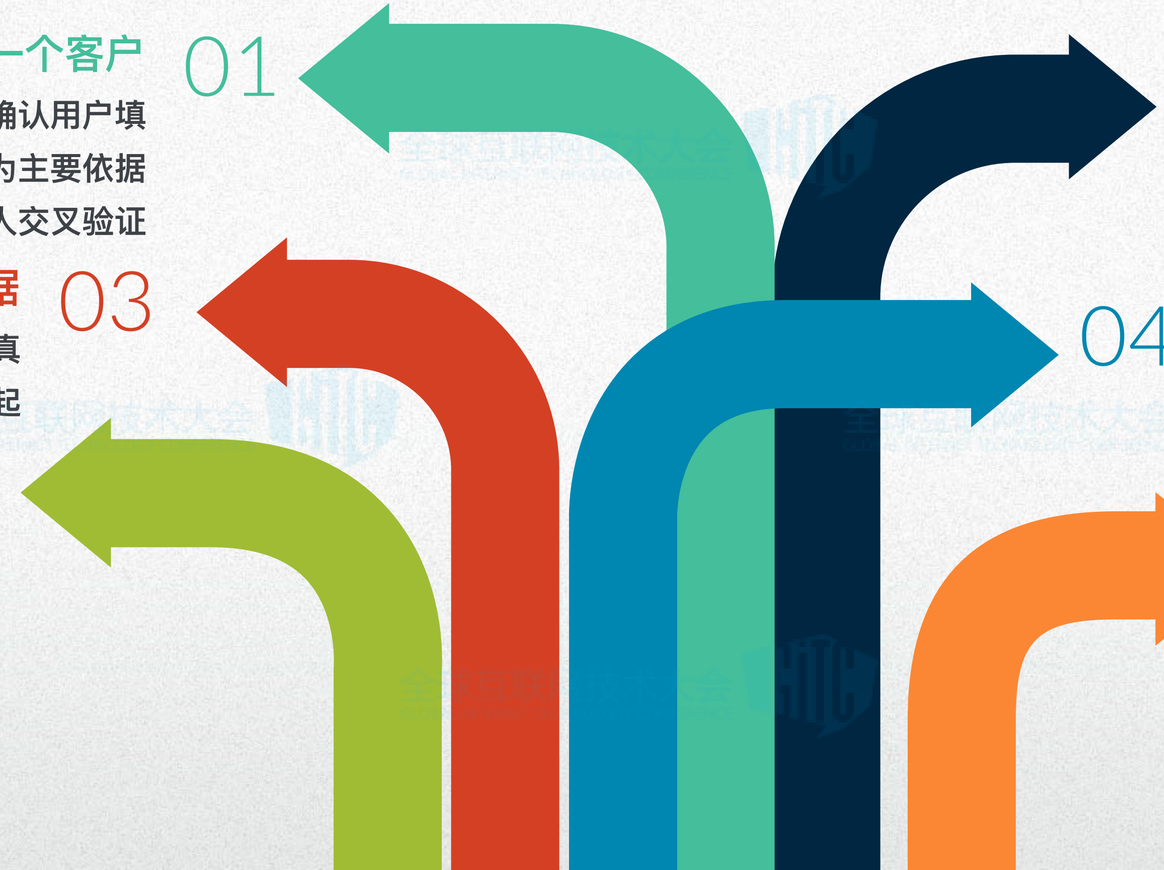
第一单！第一千单！第一万单！

10亿！50亿！100亿！……

新增50万用户 / 月，10亿 / 月

风控思维的转变

“本人”、“真实意愿”、“借款用途”、“还款意愿”、还款能力”



01 人工审核每一个客户

电核、面审、实地，以确认用户填写的信息的真实性为主要依据

结合联系人交叉验证

02 部分应用外部数据

人工搜索开放数据

一些行业内部黑名单，精准命中

03 对接专业三方数据

主要用于信息验真

三方数据公司的崛起

04 自动化数据验真

面部识别、身份证比对、活体检测

大量外围数据交叉验证

将三方数据引入模型

05 自动化审核

直拒、直批 + 人工审核

全自动化审核

06 “团伙识别”

关系图谱

风控：传统金融 VS Fintech

人 VS 机器

- 50件 / 人 / 天 VS **5000件 / 小时，全年无休**
- 培训、初审、终定、质检..... VS **只要没bug、机器够**

评分卡 VS 模型

- feature有限，调整权重，谨慎 VS **大量数据维度 & 调整极快且“浪”**
- 半年一次迭代 VS **一周多次迭代 & AB Test**
- 套用规律、借鉴规律 VS **发现规律、验证规律、学习规律**
  模型稳定、固化，模型不可识别的都为异常 **VS 识别与模型的差异并进行非监督学习，发现新的模型**

从业人员skillset

- 行业经验 VS **数据分析、挖掘能力**
- 银行（信用卡、抵押贷）、小贷、保险相关从业经验 VS **机器学习、神经网络、AI**
- 金融、统计相关专业 VS **CS**
- SAS、SQL、Excel VS **Python、MR、Hive、Spark、R**

技术团队面临的挑战

数据量

几百张表*几十列；百万行；二维，范式建模

几十张表*几千列；千万行起；稀疏、维度建模

+5TB / 月（压缩后，40%）

计算复杂度

“在10000用户间建立单向关系网络”

“在100万用户间建立双向关系图谱”

“从短信中筛选特定关键字。样本不多，大概2000多万”

服务可靠性

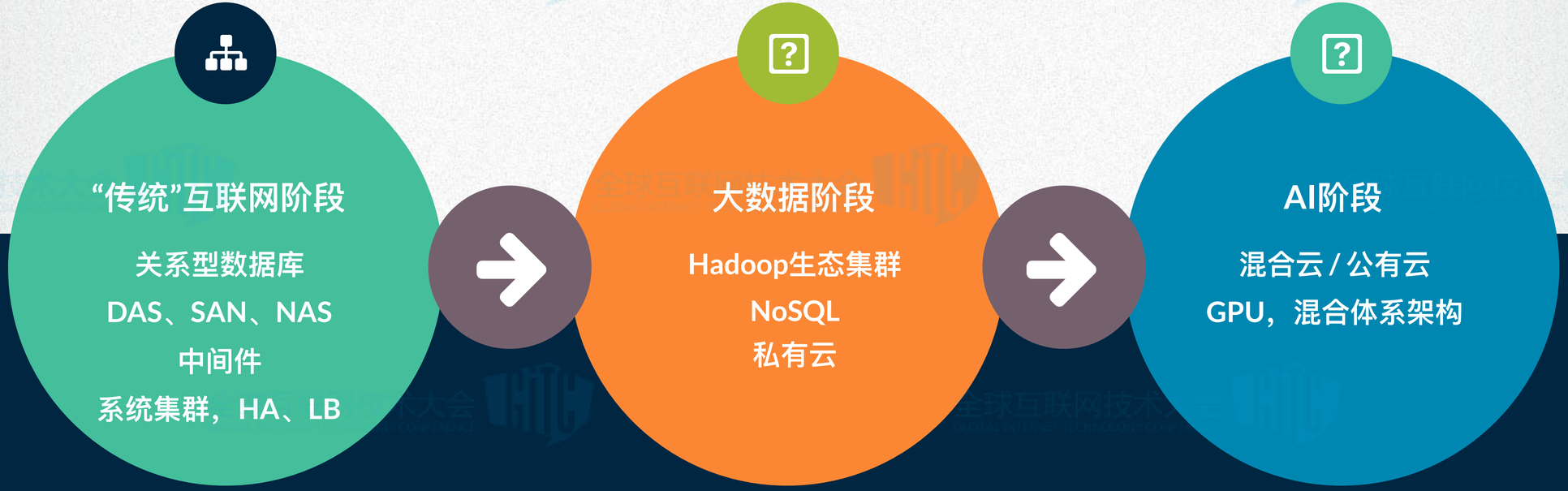
“目前系统压力大，通知前线，压一下进件量”

“系统需要加硬盘，周末停机维护”

24 * 365，SLA

系统架构演进

"ABC"

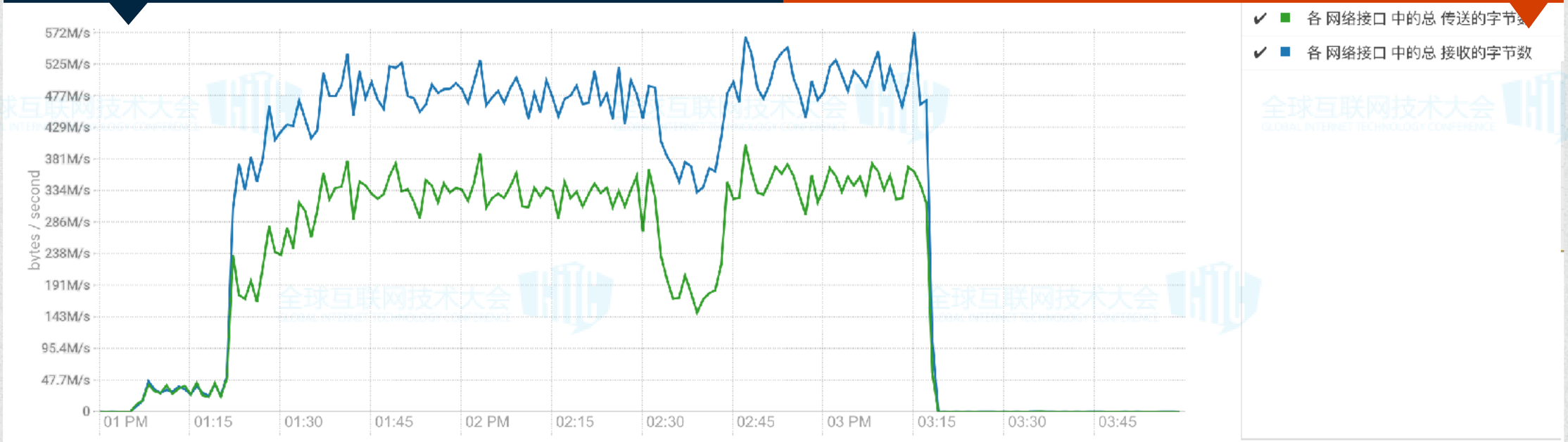

"传统"互联网阶段
关系型数据库
DAS、SAN、NAS
中间件
系统集群，HA、LB
大数据阶段
Hadoop生态集群
NoSQL
私有云
AI阶段
混合云 / 公有云
GPU，混合体系架构

“金融”互联网 VS “互联网”金融

一次线上事故

Kafka队列积压

随着业务量的增加，Kafka队列的积压问题日益频繁且严峻。除了Kafka本身的运维优化外，通过监控发现网络架构问题，最终调整解决



线上参数调优 → Consumer异步处理 → 过时消息丢弃 → 网络限流 → 问题解决

Kafka

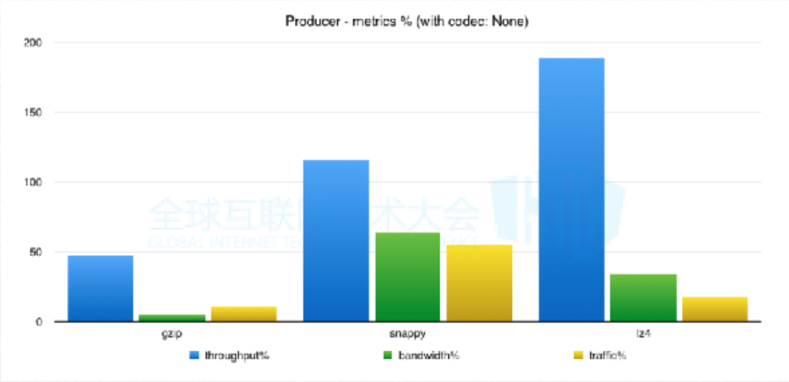
~100 MPS；95th消息大小：～500KB；95th消息处理时长：～0.7s；95th消息延迟：～1.2s

Partition / Consumer规划

- Partition越多越好?
  Kafka借助partition提升并发能力
  Partition内消息有序，而partition间顺序无保障
- Producer发送消息时注意partition倾斜（murmur2）
- Consumer数量略多于partition数量

消息压缩

- 压缩协议：gzip、snappy、lz4？无压缩吞吐最高?
- 考虑客户端的是否支持，Java、PHP、Python……



参数调优

- 吞吐 VS 延迟
- Producer
  max.request.size
  batch.size
- Consumer
  session.timeout.ms
  fetch.min.bytes
  fetch.max.wait.ms

HBase

数据量：～20TB；读取：～2,000 RPS，L1：400 RPS，L2：～200 RPS

集群 / Region规划

- 预置region
- Region倾斜
- Resign越多越好?
  Scan时阻塞遍历

Rowkey设计

- HBase无外键，选择合适的字段 / 属性作为rowkey
- 若数据按时间正序 / 逆序，考虑将时间戳置入rowkey
- 使rowkey尽量均匀地分散于region中，考虑使用MD5或其它哈希算法算法处理

Compaction优化

- 合适的StoreFile大小
- Compaction线程数

其它

Hadoop生态部署、管理工具

使用成熟的免费商业工具管理集群：

Cloudera CDH、IBM Biginsights

部署、扩 / 缩容、配置调整、监控

OLAP / OLTP边界

HBase、Mongodb、MR、Spark

热点数据识别、优化

缓存机制，合理的超时机制、缓存性价比

Mongodb索引优化

类似MySQL索引

前缀匹配

复合索引、超时索引

对写性能影响较大

运维、监控

ELK

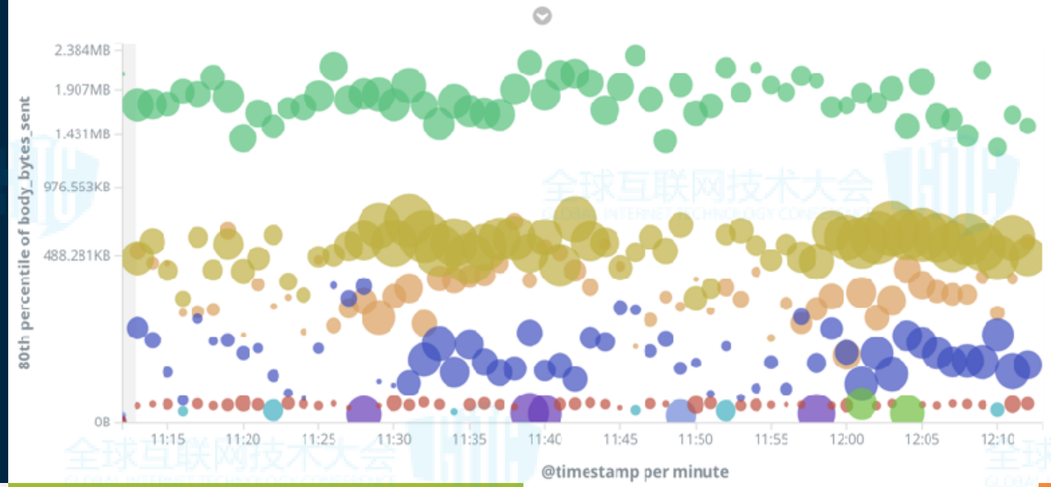
日志规范、script fields

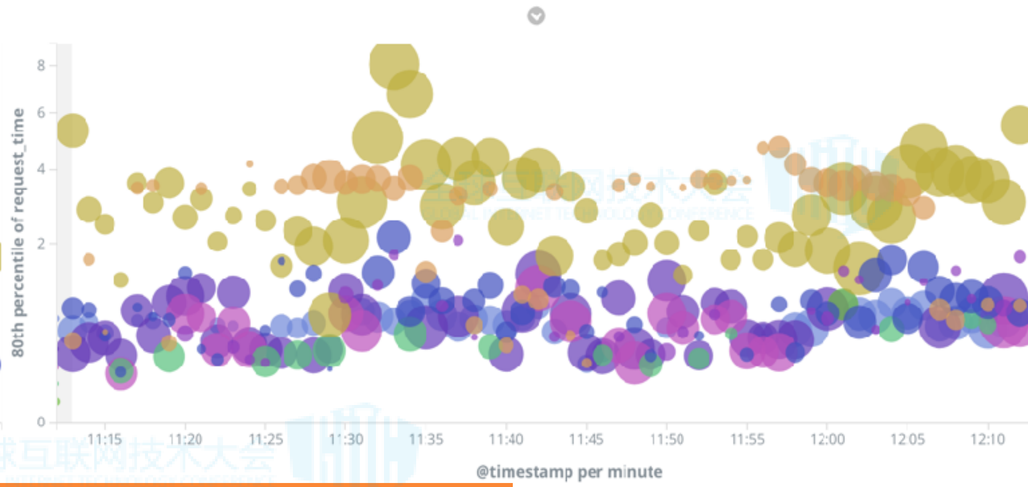
ELK Alert

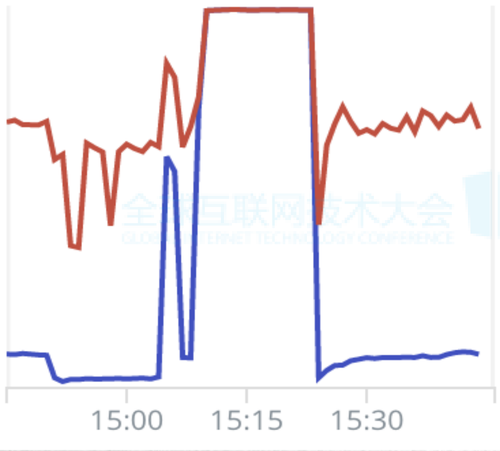
Nginx监控

响应体大小

请求响应时间



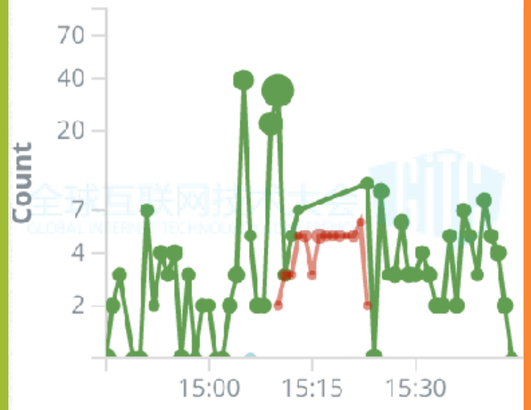




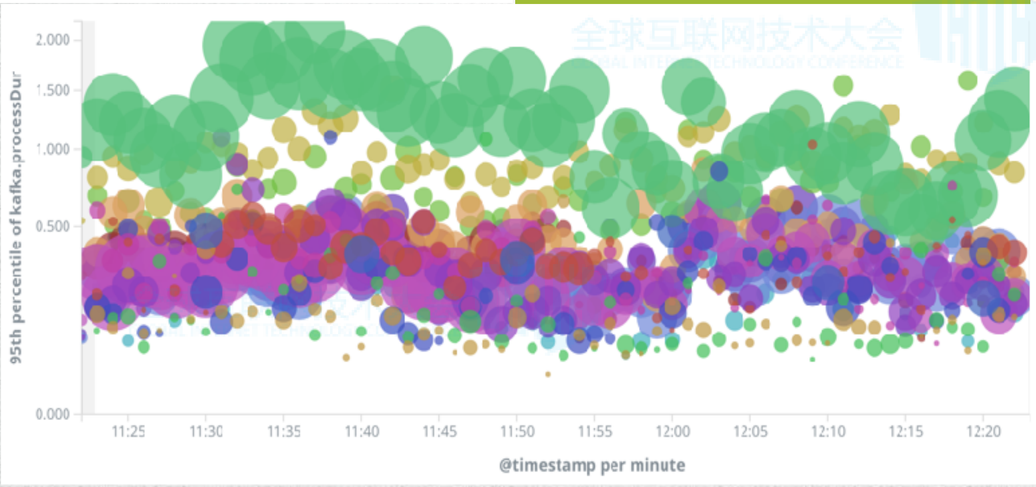
GC监控

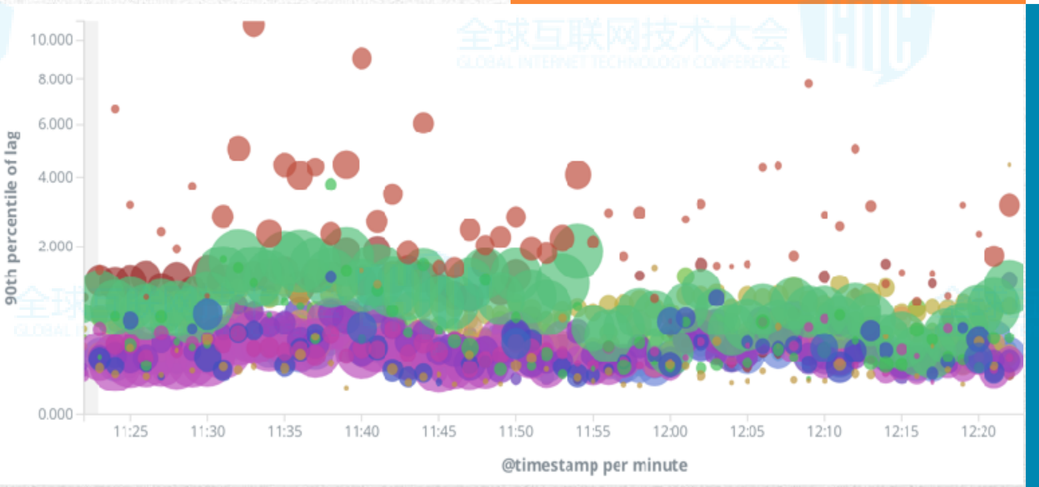
GC前后内存变化情况

Young、full GC频率



爬虫归属地监控





Kafka队列消费监控

单条消息处理时长

单条消息处理时效

再过三五年.....

Fintech将成为串接上下游高新、前沿技术的完整产业链，促进并推动其它技术领域更快地进化（产业化）

01 行业内部更开放透明的信息共享、上报机制

02 持照企业可更便捷地查询公民个人 / 征信信息

个人征信报告
公共事业缴费 / 欠费信息
公积金、社保缴费信息
法院执行、失信信息
学历、学位信息
.....

03 “跨界”专业人才的涌现与储（zheng）备（duo）

懂互联网的不（一定）懂金融，懂金融的不（一定）懂互联网
PMP、Codecademy、Coursera、Github、SOF?
注会、CFA? 零壹、起点、一本?

04 机器学习、神经网络会更深度地与风控手段结合

更“实用”的算法
性价比更高的xPU计算架构 / 集群 / 云服务
更“傻瓜”的库 / 语言

05 更成熟的数字合同技术，且受司法实践支持

数字签名签发、验证
电子合同的法律效力，司法鉴定、法院证据采纳
区块链技术

06 更可靠的身份识别技术

“又快又准”
人脸识别、比对，官方数据库
活体检测

THANKS