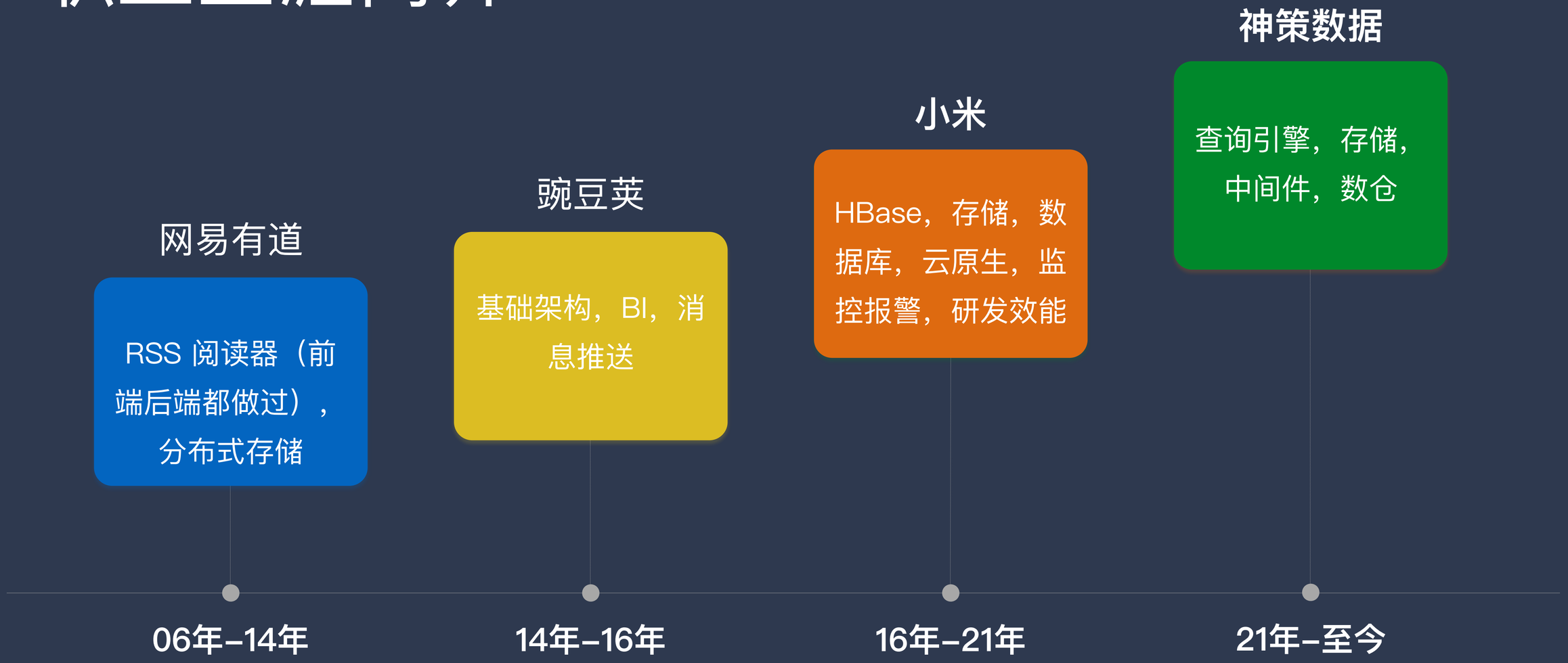
从互联网到 To B 服务

私有化部署对架构师的挑战

张铎 神策数据首席架构师

职业生涯简介


网易有道
RSS 阅读器（前端后端都做过），分布式存储
06年–14年
豌豆荚
基础架构，BI，消息推送
14年–16年
小米
HBase，存储，数据库，云原生，监控报警，研发效能
16年–21年
神策数据
查询引擎，存储，中间件，数仓
21年–至今

关于神策数据

成立 8 年+

- 总部：北京
- 分公司：上海、深圳、合肥、武汉、成都、西安等。业务辐射全国/全球企业客户

成员 1200+

- 成员规模行业前列
- 创始团队均来自百度，是国内第一批互联网大数据践行者，从 0 到 1 构建了百度大数据分析平台

总融资 ¥ 19 亿+

- 完成 2 亿美元的 D 轮融资，由 Tiger Global、凯雷投资集团领投，明势资本、DCM、线性资本、红杉中国、华平投资、Bessemer Ventures、M31 资本、襄禾资本、五源资本、GGV 纪源资本跟投，凡卓资本担任本轮融资独家财务顾问

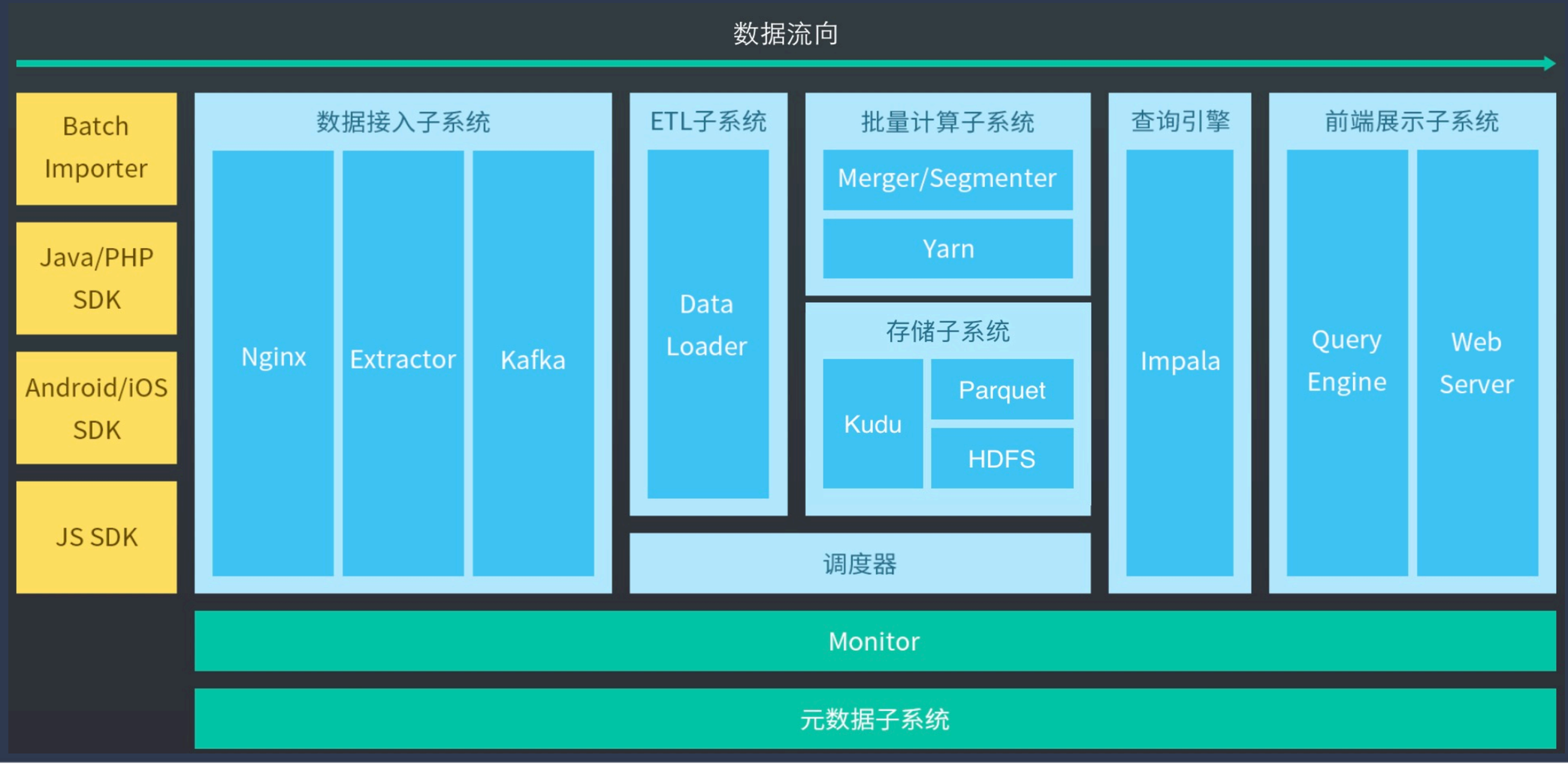
付费客户 2000+

- 私有化案例占超70%
- 覆盖行业30+。金融、互联网、品牌零售、企业服务、高科技、汽车、融合媒体、互联网+等。

中国用户行为分析行业技术与应用标准定义者

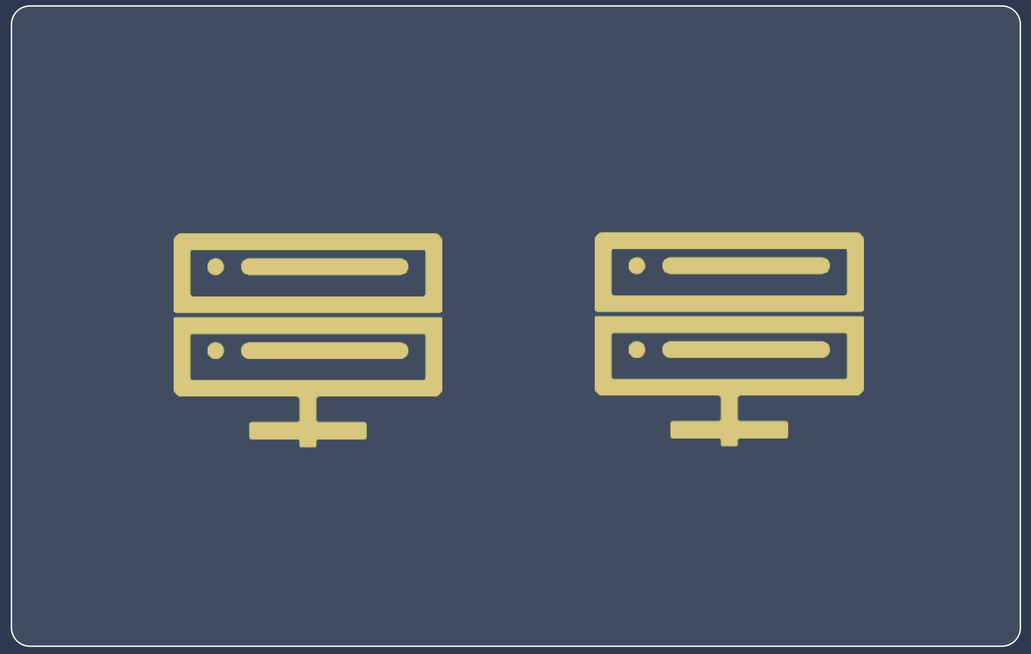
- 2018、19年连续两年荣获中国信通院评选的“最佳大数据产品奖”
- 与国家信息通信研究院联合发布中国用户行为分析行业技术与应用标准

用户行为分析平台


数据流向
Batch Importer
Java/PHP SDK
Android/iOS SDK
JS SDK
数据接入子系统
Nginx
Extractor
Kafka
ETL子系统
Data Loader
批量计算子系统
Merger/Segmenter
Yarn
存储子系统
Kudu
Parquet
HDFS
调度器
查询引擎
Impala
前端展示子系统
Query Engine
Web Server
Monitor
元数据子系统

互联网 -> To B 服务

几个大的 SaaS 集群



上千个私有化部署小集群

为什么私有化部署

1. 正常的架构师都会选 SaaS、大集群

2. 商业模式、业务需求优先于技术架构

   - 除非技术上做不出来

3. 不做私有化部署卖不出去

   - 在国内，没有公司放心把核心数据给一个创业公司
   - 政策限制，有些行业只能私有化部署

Part 1 技术挑战

混合部署

使用客户的
Hadoop 集群

- 各种版本，各种认证方式

使用客户的消息队列

- 不让随便建 Topic

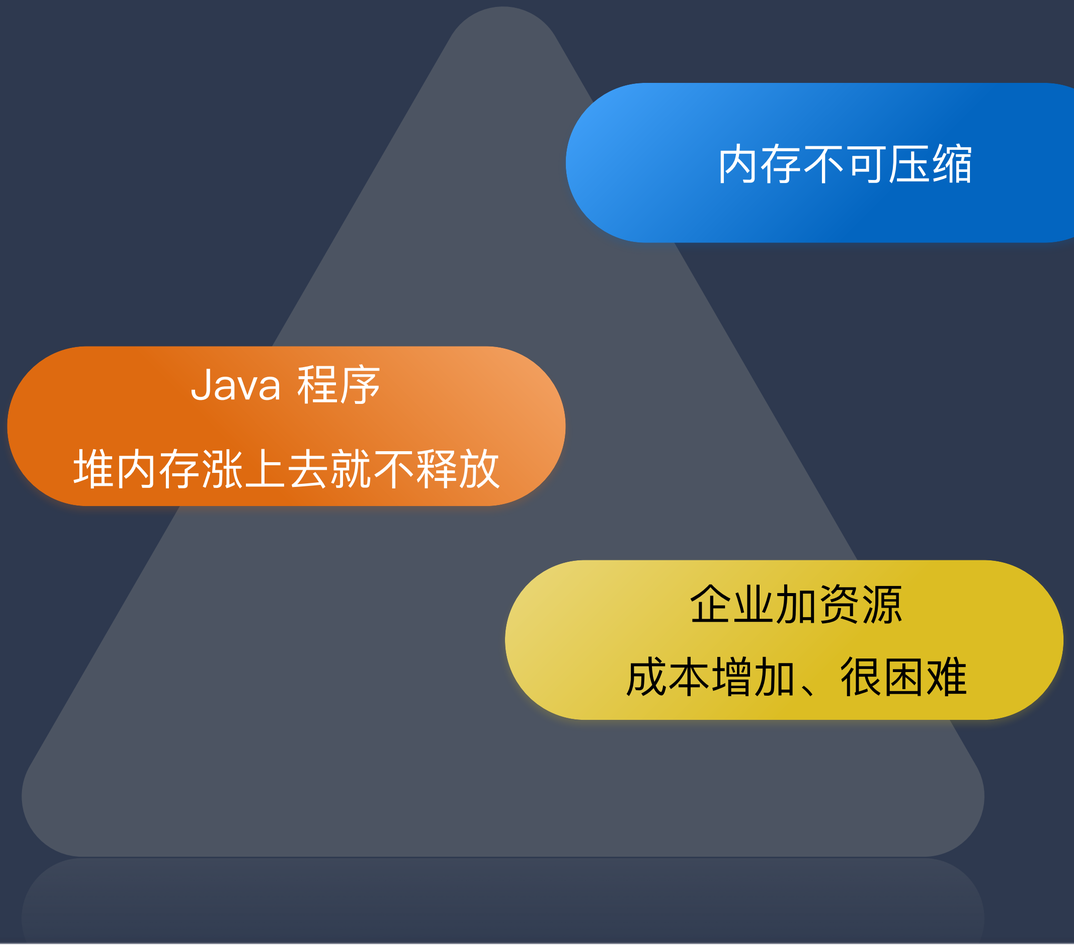
客户自己 SDK 打点

- 怎么兼容?

客户自己采集数据，
再导入神策

- 走 nginx 仿照 SDK 转一道，性能不理想
- 走批量直接入库，需要转换，不够实时
- 开发一个 Flink 任务跑在客户集群上?

减小组件内存占用



- 更精细的模块控制，客户用不到的组件就不启动，节省资源
- 引入新的 GC 算法，让堆内存可以收缩(ZGC、Shenandoah GC)

- CPU 不够是慢，内存不够直接挂

资源受限情况下的查询优化

01 针对用户行为分析场景定向优化

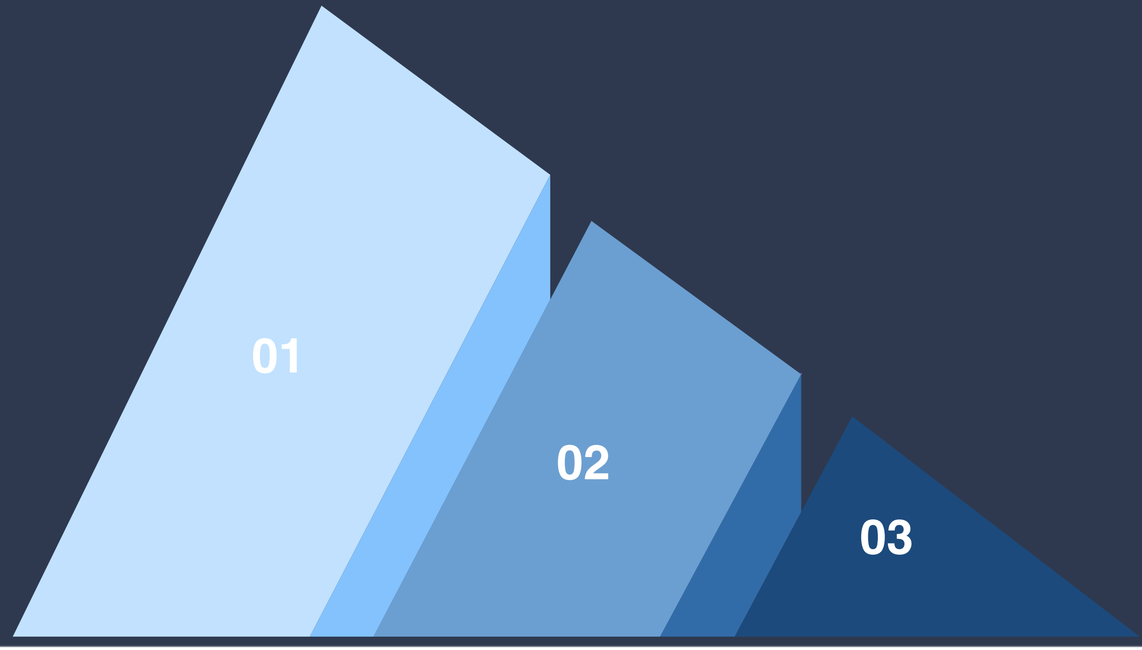
- 用户行为数据本身有序，重写 SQL，将 join 全排序变为归并排序
- 记录用户最后活跃时间，过滤不活跃用户
- 外连接消除
- 高基数分组优化

02 查询资源预估

- 资源不够先等待，避免谁都查不出来
- 基于历史资源消耗预估，用的越多越准确

03 神策数据数仓负载管理平台

- 让客户清楚自己的资源是怎么消耗的



Part 2 非（纯）技术挑战

企业部署环境各种奇怪的限制和要求

- 物理机，无法按照要求挂载磁盘，扩容时候配置还不一样
- 不给 sudo 权限，甚至有把 sudo 这个命令的 binary 直接删掉的
- 多网卡，不同机器组之间通信用不同的 IP
- 不通外网，不让采集监控数据，服务挂了也不知道
- 开权限，但是不开认证；要加密，但是没有 KMS
- ……

如何「兼容各种配置的机器」

- 前置检查，优先沟通，尽量推动客户改配置
- 在部署系统中抽象各种概念，例如随机盘，顺序盘，机器组，尽量确保在输入机器配置之后，可以自动生成程序配置，无需人工干预
- 机器性能不达标？如果客户坚持就走付费压测

如何解决「不让用 root，不给 sudo」的挑战

安装时必须给 root，这个不能妥协，通常客户会接受

运行期可以不给 root 或者 sudo

- CDH 会用不同用户启动服务？自研大数据组件部署工具，可以单一用户启动
- 为了兼容老的 CDH 环境，部署系统需要支持多用户和单用户两种模式
- 相对应的，各个服务不能假定自己的账号是什么，需要由部署系统传入
- 内部测试环境也不给 root，确保不会反复

如何解决「网络环境复杂」的挑战

只有笨办法：使用域名互相通信，配置 /etc/hosts 来映射不同的 IP

困难点：不同的组件 hack 方法不一致，操作成本很高

- Hadoop：增加配置强制使用 hostname 来访问 datanode
- Kudu：配置 advertised_addresses
- Pegasus(skv)：只能用 IP，没有办法 hack。正在推动社区支持 FQDN

更优解：和客户沟通，降低复杂度

- 有些特殊行业没有办法，比如金融行业，外部可访问的服务和内部服务必须放在不同的网络分区里，中间要有防火墙
- 但具体哪些服务放哪边儿，还是可以谈的，谈的好能降低很多复杂度

如何解决「不通外网」的挑战

不通外网，最大的挑战就是监控和报警出不来

金融客户常见情况，政策要求，没有讨论的余地

*没有政策限制的行业，还是优先和客户沟通解决

监控可以本地看，报警必须想办法转出来

- 给报警机器增加 IP 白名单，让客户可以请求神策的服务进行报警
- 报警对接客户报警系统，让客户把报警邮件自动转给神策
- 安排驻场专门收报警（客户更倾向于这类属于免费增值服务）
- 提前约定好，客户自己收到报警再通知我们，但处理时效就无法保证了

如何解决「非常规的认证加密」

- 首先要搞清楚客户的真实需求

  是真的要安全，还是“为了安全而安全”

- 提供各种兼容回退方案

  是不是开云硬盘加密就可以？提供模拟的 KMS 保存根密钥

- 如果是真的要安全，那么要坚持底线

  安全不绝对就是绝对不安全

如何解决「版本收敛」

上千家客户，数十个组件，每个组件若干版本跑在线上，乘起来是个天文数字……

测试覆盖度足够是保证复杂产品最终质量可控的必要条件

不同组件版本绑定，升级一起升

- 极大降低 QA 工作量
- 需要定位为软件公司，有统一的开发和发布节奏

设置中继版本，跨越版本较多时需要先升级到中继版本

- 任意两个版本都可以直接升级，版本一多测试工作量仍然较大

Part 3 变与不变

架构师的职责

业务可以正常运行

在可控的成本下运行

让技术架构

可以“支撑”公司的业务

互联网 vs. 私有化部署（业务场景挑战不同）

大规模，高并发

vs.

资源受限，场景复杂

案例：设计对象存储服务

- 素材管理，需要一个内部可以上传，外部可以访问并裁剪的存储服务
- 标准的对象存储服务，最好直接用云，但私有化部署如何确定用哪个云?
- 自己做一个适配层，兼容各种主流云厂商的对象存储服务
- 客户不在云上怎么办？底层用 HDFS，适配层自己需要支持裁剪缩放
- 客户自己买了商用对象存储要对接？就当 HDFS 用，不用额外功能
- 库 VS 服务？还是需要服务，让使用方自己做各种配置不现实
- 但增加服务就要多耗资源，客户不愿意怎么办？做成标准的 HTTP 服务，提供嵌入其他服务中的姿势
- ……

互联网 vs. 私有化部署（商业模式不同）

运维成本相对不敏感

vs.

运维成本直接决定生死

案例：要不要加配置

- 一个后台合并 parquet 文件的任务，同时合并太多容易 OOM，在一个客户那里跑不过去
- 最快的改法：加一个配置，限制一下单次合并的文件数量，给这个客户配置
- 影响？配小了影响合并速度，配大了影响稳定性，不同客户配的还不一样，需要培训运维和交付人员，成本明显上升
- 结论：不要加配置。自适应，能选的文件全选上，代码里自动改成多轮合并，找一个性能和稳定性的平衡点，减少运维成本

写在最后

- 不存在一招鲜

  业务需求变了，关注点自然要变

- 要能搞清楚技术的极限

  私有化部署到底能不能赚钱?

  最终归宿是不是仍然是 SaaS?

谢谢