Lambda已死，去ETL化IOTA才是未来

易观CTO 郭炜

郭炜
易观 CTO

郭炜先生2015年加入易观，担任易观CTO，构建易观技术团队完成易观大数据采集、平台、数据挖掘等技术架构与体系，从无到有完成易观混合云搭建、易观SDK升级并发布易观秒算实时计算平台，目前易观大数据平台日处理数据量30T，272亿条，月活用户5.5亿。

郭炜先生毕业于北京大学，加入易观之前，曾任联想研究院大数据总监，万达电商数据部总经理，并曾在中金、IBM、Teradata公司担任大数据方向重要岗位，对大数据前沿领域研究，包括视频、智能WIFI等大数据软硬数据一体技术有独特的见解。

目前易观大数据混合云的数据规模

Analysys 易观
你要的数据分析

终端覆盖

累计终端覆盖：22.5 亿

- 设备类型
- 品牌
- 机型
- 价格区间
- 屏幕尺寸
- 网络制式
- 运营商
- 摄像头画质

产品覆盖

监测APP数量：266万+

行业覆盖：309 个

- 现状分析
- 趋势数据
- 增速分析
- APP人群画像

4大类指标体系与数百个创新指标

用户覆盖

MAU：5.5 亿

DAU：7900万

标签类型：8365个

人群画像

- 年龄
- 性别
- 职业
- 婚育状况
- 资产状况
- 消费水平
- 常住地
- 人群特征

行为特征

- 领域偏好度TGI
  - 45个一级领域
  - 309个二级领域
- 应用偏好度TGI
  - 4万+应用
  - 1万+游戏
- 兴趣偏好度TGI
  - 细分功能
  - 内容分类
  - 商品分类
  - 品牌分类

消费场景

- 家庭生活
- 娱乐社交
- 购物消费
- 运动健康
- 工作/商务
- 旅游出行
- 学习教育

地理位置

- 国家级别定位
- 省份级别定位
- 城市级别定位
- 商圈级别定位
- 地理围栏定位
- POI定位
- LBS轨迹

数据基础资源

数据存储容量5.8PB

每日处理数据条数271亿

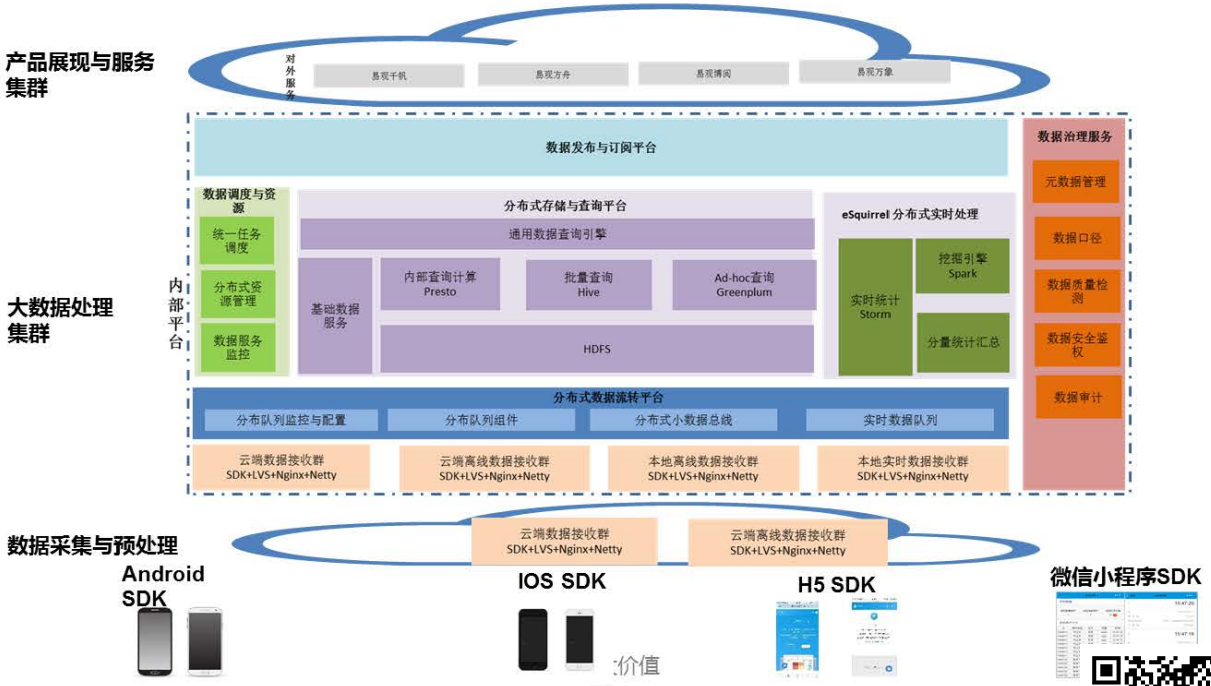
数据合作伙伴1200+个

每秒处理数据61万条

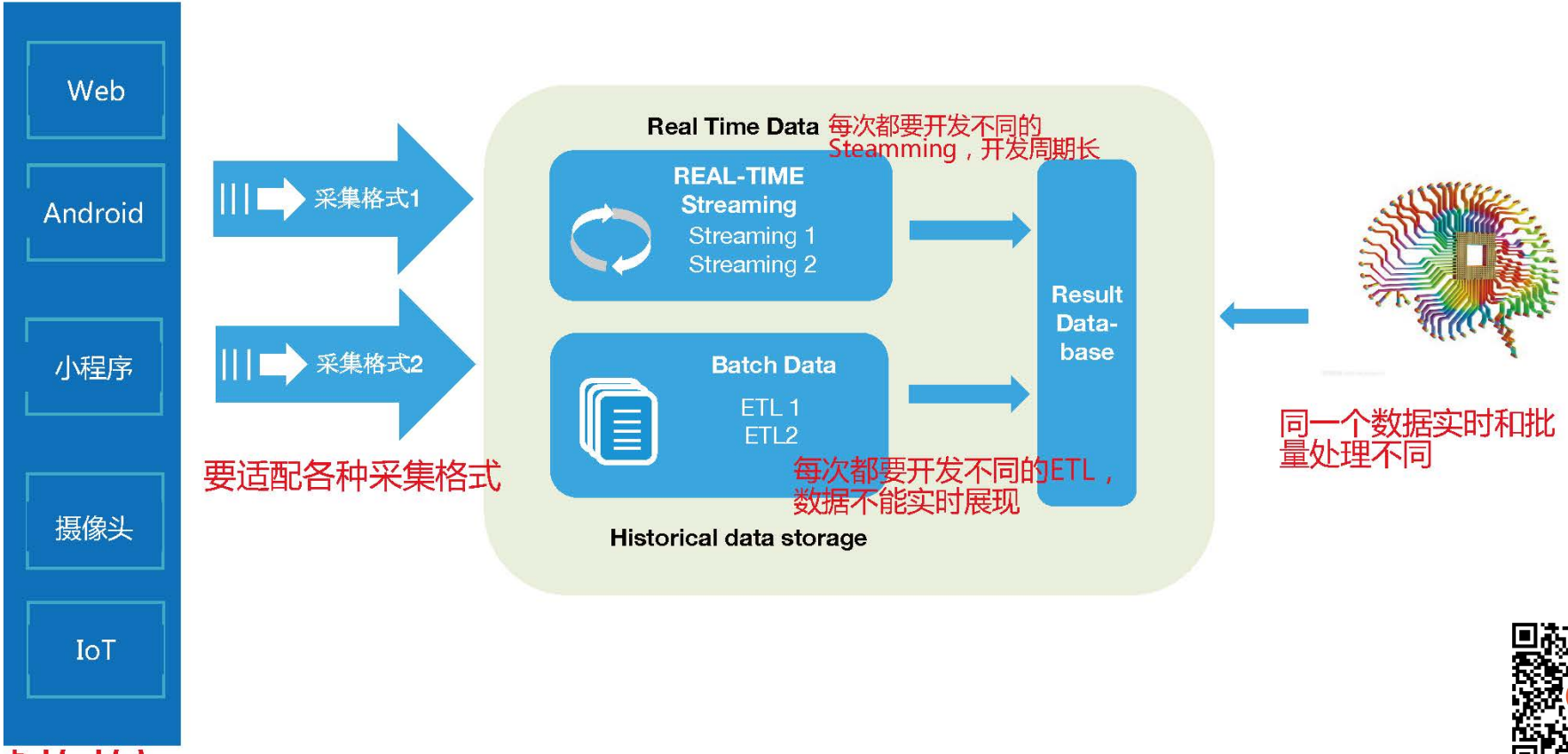
实时数据处理

现在大数据混合云的困境

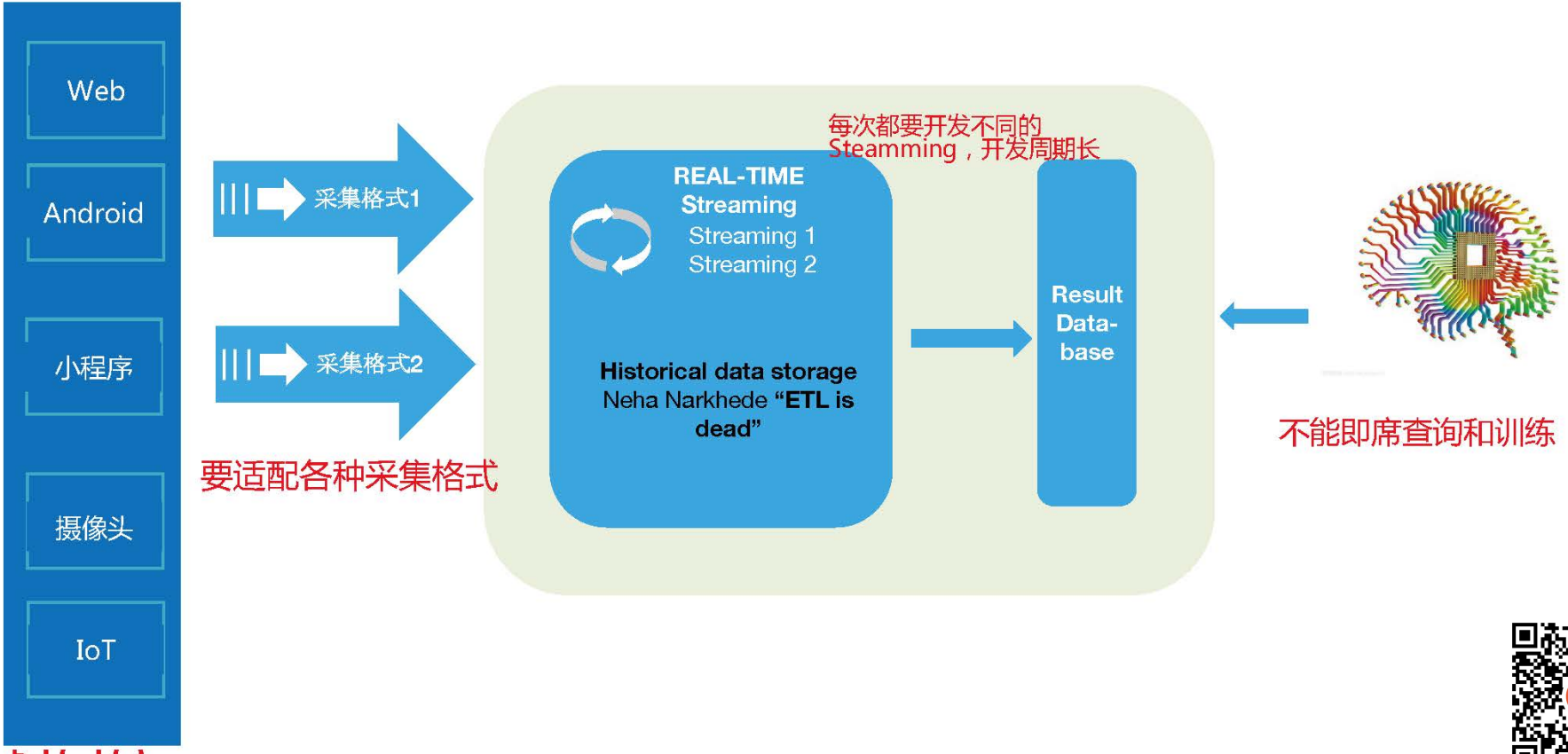
- IOT大潮来临
- 数据量级巨大
- 数据格式不相统一
- 数据业务多变
- 数据需要实时查询



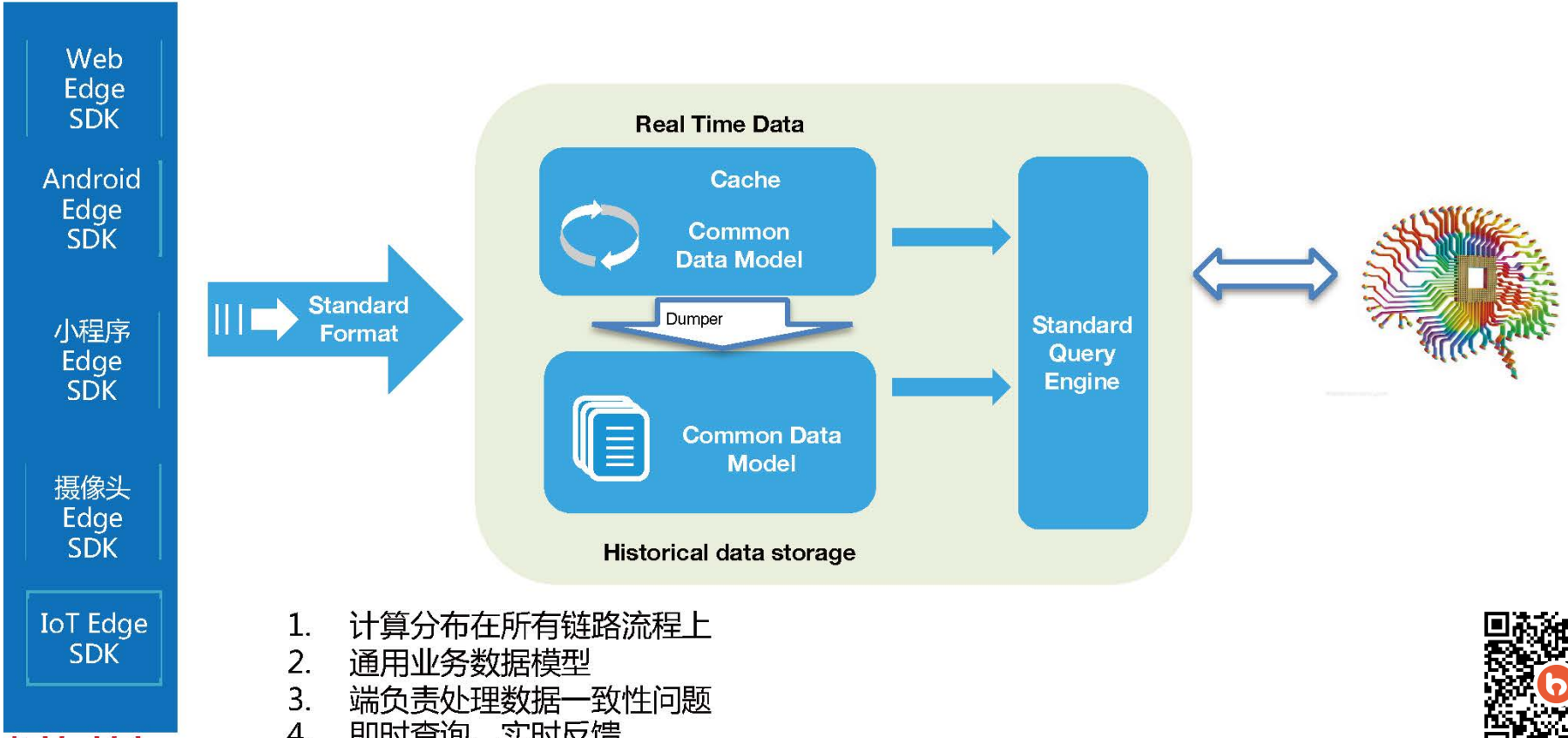
传统大数据架构下的挑战——LAMBDA架构



传统大数据架构的挑战——KAPPA架构



边缘计算下的大数据和AI架构——IOTA架构



1. 计算分布在所有链路流程上
2. 通用业务数据模型
3. 端负责处理数据一致性问题
4. 即时查询、实时反馈

数据模型与存储

- Common Data Model：贯穿整体业务始终的数据模型，这个模型是整个业务的核心，要保持SDK、cache、历史数据、查询引擎保持一致。对于用户数据分析来讲可以定义为“主-谓-宾”或者“对象-事件”这样的抽象模型来满足各种各样的查询。以大家熟悉的APP用户模型为例，用“主-谓-宾”模型描述就是“X用户 – 事件1 – A页面（2018/4/11 20:00）”。当然，根据业务需求的不同，也可以使用“产品-事件”、“地点-时间”模型等等。模型本身也可以根据协议（例如 protobuf）来实现SDK端定义，中央存储的方式。此处核心是，从SDK到存储到处理是统一的一个Common Data Model。
- 实时数据缓存区，这部分是为了达到实时计算的目的，海量数据接收不可能海量实时入历史数据库，那样会出现建立索引延迟、历史数据碎片文件等问题。因此，有一个实时数据缓存区来存储最近几分钟或者几秒钟的数据。这块可以使用Kudu或者Hbase等组件来实现。这部分数据会通过Dumper来合并到历史数据当中。此处的数据模型和SDK端数据模型是保持一致的，都是Common Data Model，例如“主-谓-宾”模型。
- 历史数据沉浸区，这部分是保存了大量的历史数据，为了实现Ad-hoc查询，将自动建立相关索引提高整体历史数据查询效率，从而实现秒级复杂查询百亿条数据的反馈。例如可以使用HDFS存储历史数据，此处的数据模型依然SDK端数据模型是保持一致的Common Data Model。

新一代边缘计算的大数据混合云

Analysys 易观
你要的数据分析

应用层

渠道分析 | 转化、留存分析 | 用户画像 | 营销 | CRM | ERP

自助查询 (Superset)

服务层

Query Engine

数据治理

- 元数据管理
- 数据追踪
- 数据质量稽核
- 数据安全鉴权
- 数据审计

数据层

数据管理

- 任务调度 Dispatch
- 资源管理 Yarn
- 服务监控 Monitor

即时查询引擎(Presto、Spark)

数据处理模型

- Event
- Profile
- DumpMR
- MergerMR

存储引擎

- HDFS
- HBase

第三方存储

- Mysql
- Redis
- Mongo
- ...

接收层

分布式数据传输接收平台

Netty | Extractor | KafKa | 策略配置

边缘计算

Import Tools | Java/C/PHP Edge SDK | Android/iOS Edge SDK | IOT Edge SDK

目前易观大数据混合云的数据规模

Analysys 易观
你要的数据分析

终端覆盖

累计终端覆盖：22.5 亿

- 设备类型
- 品牌
- 机型
- 价格区间
- 屏幕尺寸
- 网络制式
- 运营商
- 摄像头画质

产品覆盖

监测APP数量：266万+

行业覆盖：309 个

- 现状分析
- 趋势数据
- 增速分析
- APP人群画像

4大类指标体系与数百个创新指标

用户覆盖

MAU：5.5 亿

DAU：7900万

标签类型：8365个

人群画像

- 年龄
- 性别
- 职业
- 婚育状况
- 资产状况
- 消费水平
- 常住地
- 人群特征

行为特征

- 领域偏好度TGI
  - 45个一级领域
  - 309个二级领域
- 应用偏好度TGI
  - 4万+应用
  - 1万+游戏
- 兴趣偏好度TGI
  - 细分功能
  - 内容分类
  - 商品分类
  - 品牌分类

消费场景

- 家庭生活
- 娱乐社交
- 购物消费
- 运动健康
- 工作/商务
- 旅游出行
- 学习教育

地理位置

- 国家级别定位
- 省份级别定位
- 城市级别定位
- 商圈级别定位
- 地理围栏定位
- POI定位
- LBS轨迹

数据基础资源

数据存储容量5.8PB

每日处理数据条数271亿

数据合作伙伴1200+个

每秒处理数据61万条

实时数据处理

数据分析驱动业务升级

■ 易观千帆 ■ 易观方舟 ■ 易观标签云 ■ 易观行业解决方案

网址：www.analysys.cn

客户热线：4006-515-715

微博：@Analysys易观

Analysys 易观

你要的数据分析

微信扫码收听演讲音频