数据平台实时化实践

苏宁大数据中心 — 王富平

流式计算的成熟

- 流式技术完成普及：以storm为代表的流式计算框架在业界普及开来，得到了广泛实践
- 批流处理范式统一：谷歌dataflow提出了批流统一模型，spark－streaming与flink等平台践行这一理念

在线应用场景爆发

- 大数据实时分析：仪表盘、实时监控
- 在线学习：用户实时意图、在线排序模型

数聚平台做什么？

- AI ？
- OLAP ？

- 实时多维分析平台，提供面向业务的可视化解决方案

走向平台化

- 统一管理、资源共享
- 聚焦分析、提升能力

需求长什么样？

- 分析部门：指标实时计算
- 财务部门：支持复杂的业务指标计算
- 业务部门：可视化方案、多维分析

指标

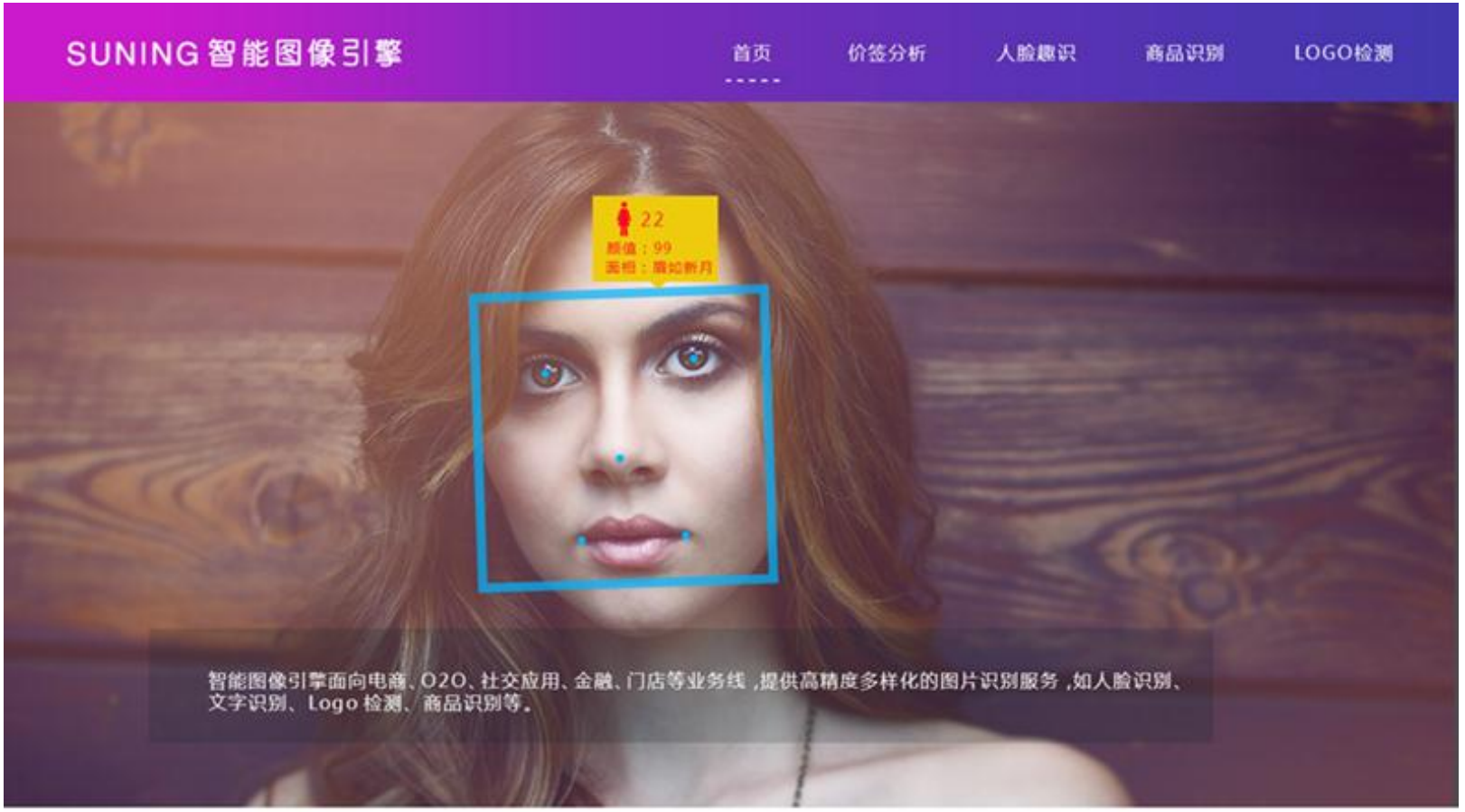
（XX产品结构实际进销差率 - 其他产品结构进销差额 / 其他产品结构付款金额）* 变动比例（销售占比平均值-销售占比实际值）* 产品总付款金额

线下门店客流分析

实现门店监控视频的智能分析，以获取准确的客流人数、

客流成分组成、商品热区、顾客滞留分析等丰富的

商业数据信息，进而为商业智能系统提供数据保障



应用场景

电商 O2O

识别图像中商品，实现自动化精准化的商品检索和推荐，增强用户购买体验；识别图像中的 Logo，实现品牌的精准化推荐。

金融

使用文字识别 OCR 对身份证、银行卡等开户证件进行识别，减少用户输入，增强用户体验；对关键交易使用人脸识别进行刷脸支付，控制风险。

社交

对社交应用中上传的图片进行审核，有效识别色情、涉政等违规图片；对图片中人脸特征进行分析，并实现换脸、美颜、PK 大咖等娱乐体验。

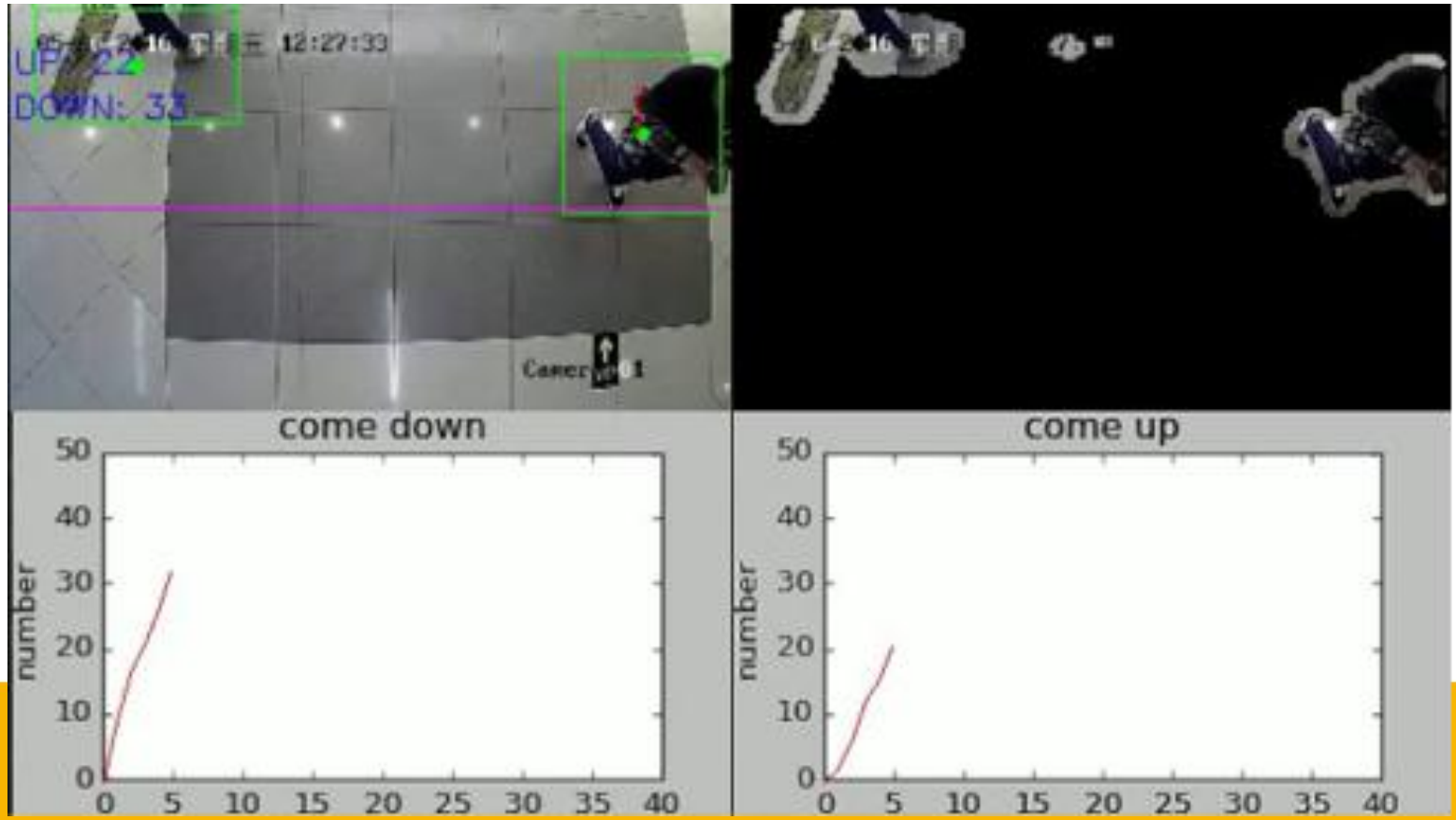
门店

使用文字识别 OCR 对门店价签进行自动识别和搜索；检测和统计顾客基本信息，用于客流统计、用户分析、商品推荐等商业应用。

智能门店监控：客流统计

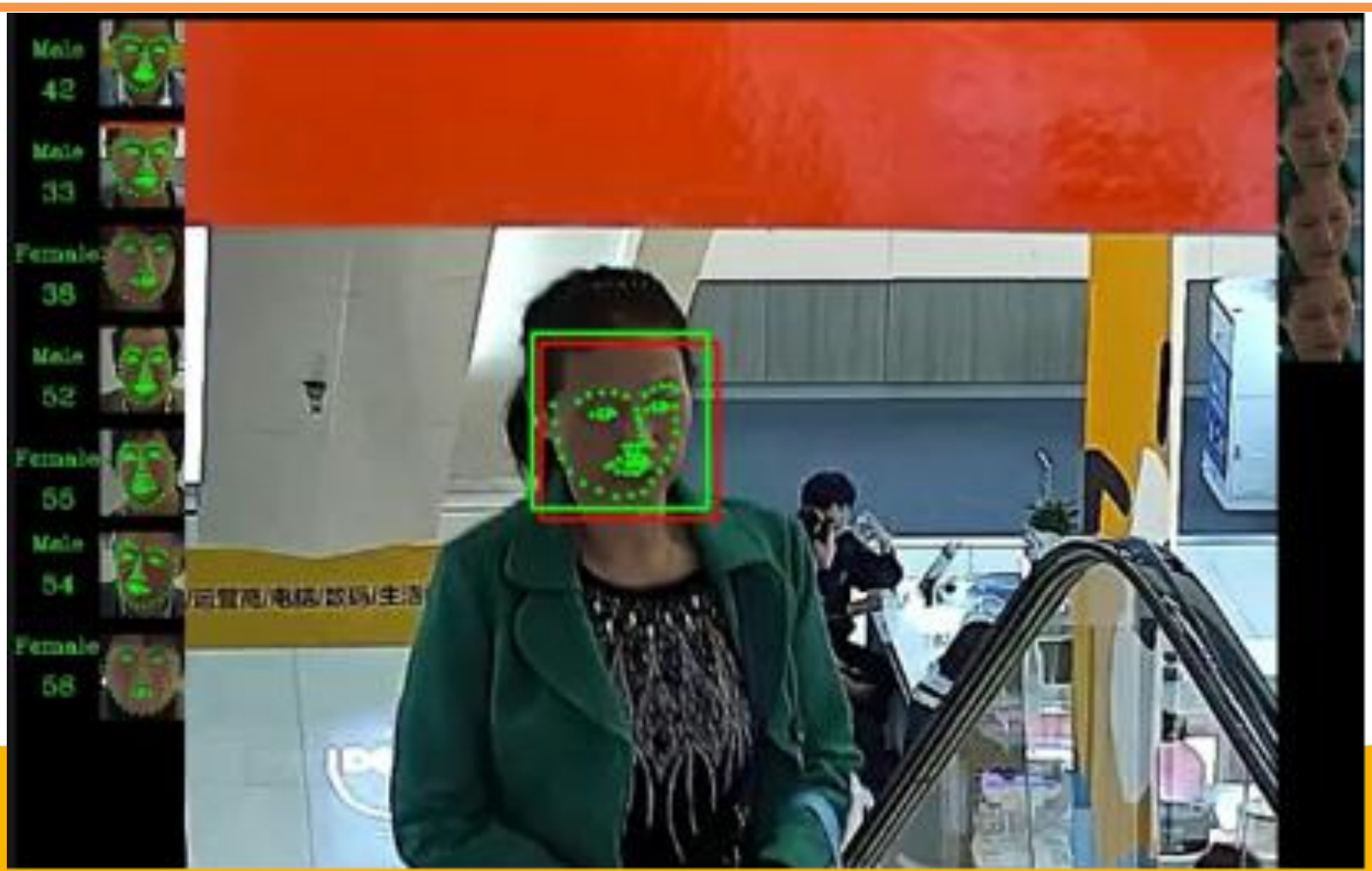
方案：

1. 采用置顶数字摄像头
2. 检测进入区域的人体
3. 跟踪并确认有效客流



智能门店监控：人脸分析

1. 利用特征点跟踪有效估计人脸姿态，利用正面人脸进行去重分析。
2. CNN网络分析年龄及性别。



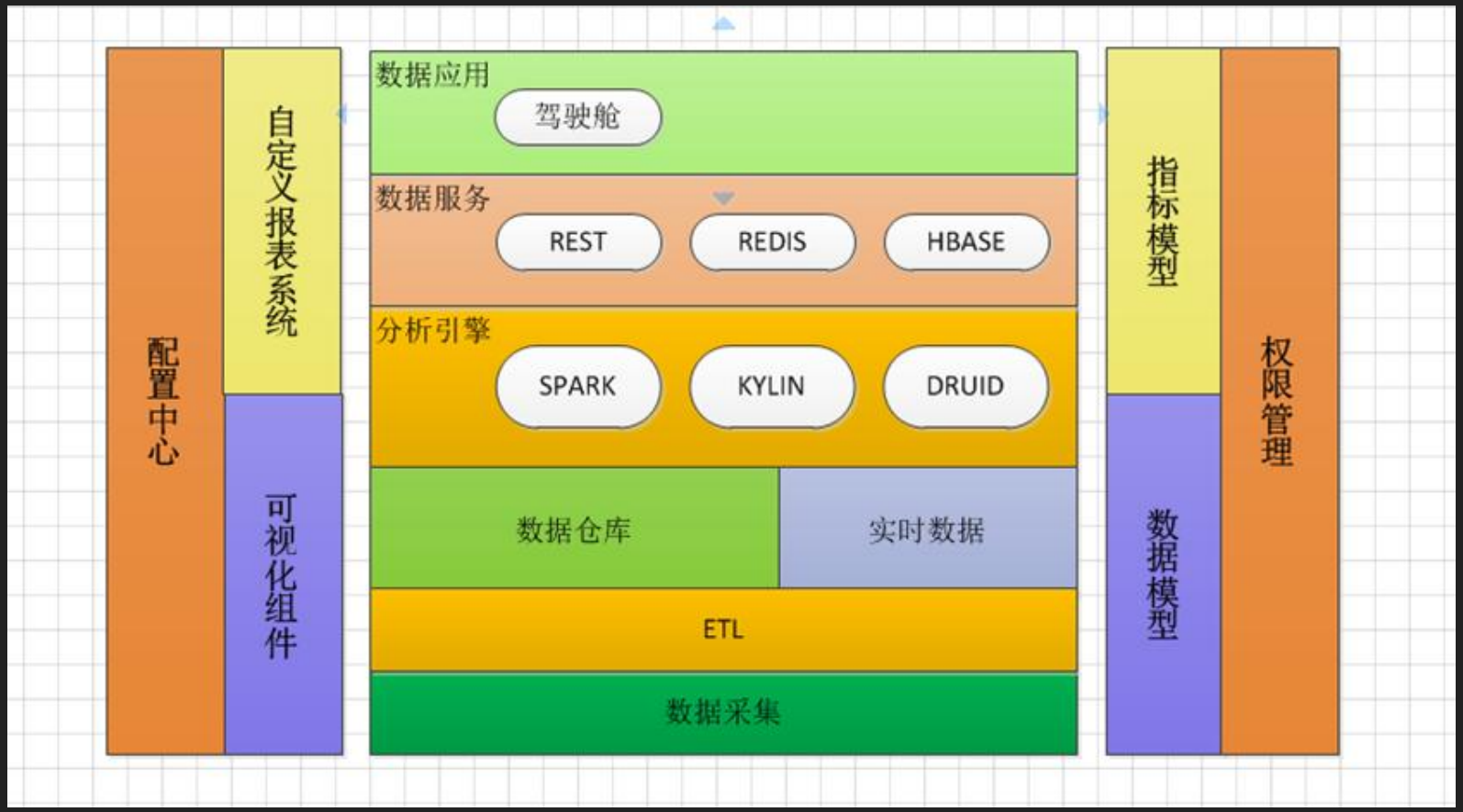
智能门店监控：顾客行为分析

1. 对于置顶或偏置摄像头，都能够有效获取商品热力图。

2. 对顾客进行去重和滞留分析，能够最大限度抑制销售员影响。

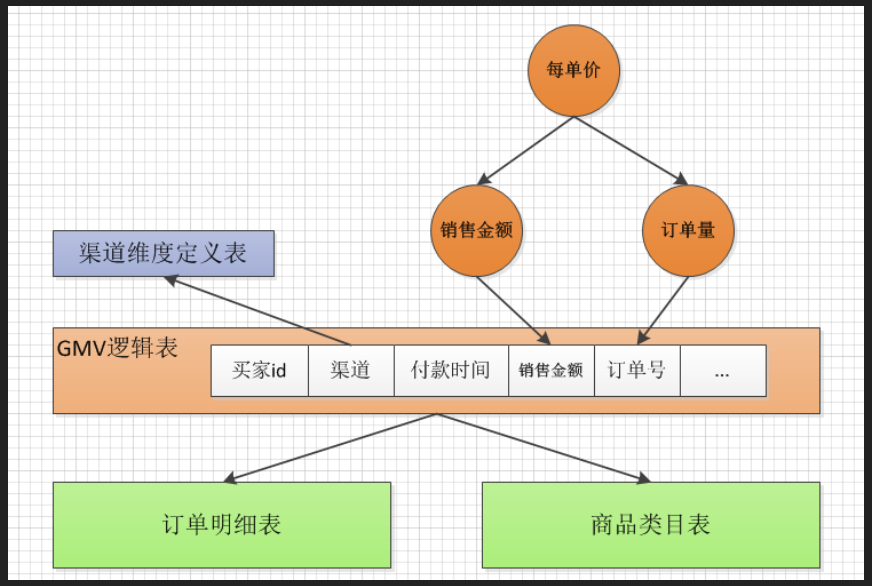
3. 进一步分析顾客与销售员行为。

数聚架构图


数据应用
驾驶舱
数据服务
REST
REDIS
HBASE
分析引擎
SPARK
KYLIN
DRUID
数据仓库
实时数据
ETL
数据采集
自定义报表系统
可视化组件
配置中心
指标模型
数据模型
权限管理

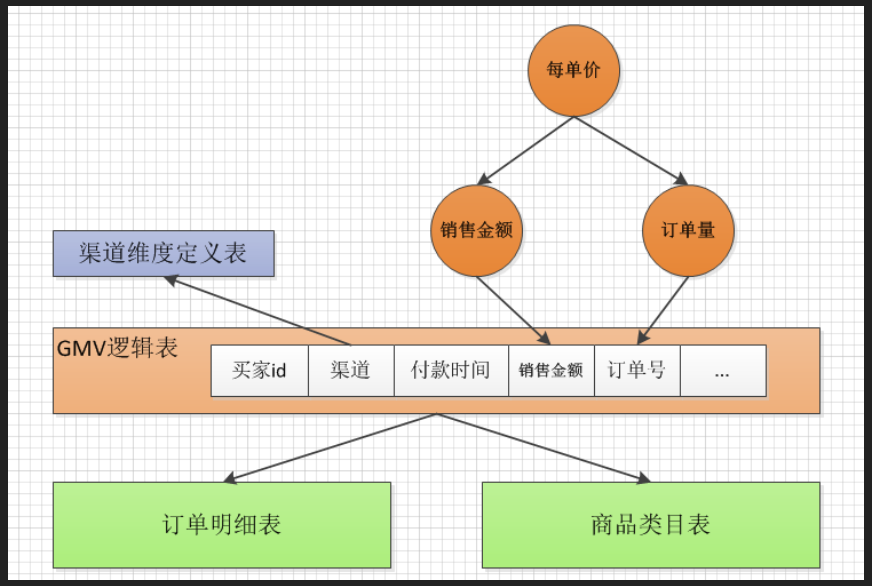
数据模型

- 物理表：数据仓库的事实表
- 逻辑表：基于物理表生成的视图表，由三部分组成：时间字段、维度字段、指标字段



指标模型

- 一级指标：基于逻辑表直接计算得出
- 二级指标(衍生指标)：依赖其他指标计算得出



保障指标实时产出？

- druid：利用druid进行实时分析

变态实时需求：数据量大，刷新频率高，并发量大

- 维度固定：大区－城市－门店
- spark－streaming硬算后直接入hbase、redis等kv数据库

DRUID性能测试

druid测试集群3台机器，数据接口部署在10.27.15.33一台机器上						
查询指标	视角、过滤条件	时间维度	时间范围	并发	是否有缓存	（所有线程查询任务完成）总耗时
ZB_XS_0001_SJ_01 GMV ZB_XS_0005_SJ_01 销售数量 ZB_XS_0003_SJ_01 订单数量 ZB_YY_0001_SJ_01 买家数 ZB_CW_0001_SJ_01 进销差 ZB_XS_0006_SJ_01 销售单价 ZB_XS_0004_SJ_01 订单单价 ZB_YY_0002_SJ_01 客单价 ZB_XS_0001_SJ_07 GMV环比	视角：大区-品类-品牌 过滤条件：大区：1060、1061 排序： GMV排序 top100	周	2016年5月份1-4周	1	No	5496ms
		周	2016年5月份1-4周	1	YES	460ms
		周	2016年5月份1-4周	4	YES	500ms
		日	2016年5月1日-5月30日	1	NO	6374ms
		日	2016年5月1日-5月30日	1	YES	800ms
		日	2016年5月1日-5月30日	4	YES	1121ms
		日	2016年5月1日-5月30日	30	YES	4210ms

物流的需求难题：实时分析未完成配送订单数据

- 时间跨度大：一年前的订单也存在未配送
- 状态更新：配送中、配送失败、配送完成
- 数据更新集中：高峰期5分钟更新100万条数据

具体问题具体分析

- 实时要求不高，5分钟延迟
- 动态更新：insert、update、delete
- 总数据量2000w左右

两种方案

- MPP: Greenplum
- Redis + Spark

如果延迟要求是10S内 ？

如果数据是2亿 ？

保证EXACT－ONCE

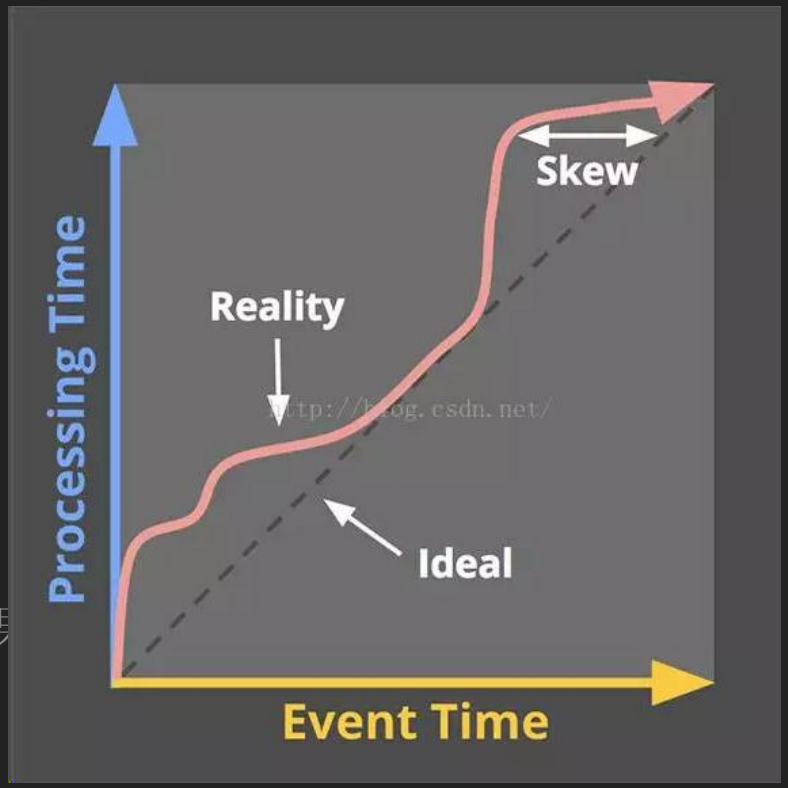
- 可靠的数据源
- 幂等性：处理逻辑保证幂等性

外部系统不稳定

- 采集系统宕机、数据延迟严重
- 外部系统重复发送数据

DATA-FLOW的愿景：批流统一

- 计算逻辑是什么
- 计算什么时候的数据（事件时间）的数据
- 在什么时候（处理时间）进行计算
- 后续数据的处理结果如何影响之前的处理结果




谢谢大家！