京东虚假交易识别系统

京东商城 广告部 寿如阳

虚假交易识别的需求与挑战

虚假交易的危害

商业分析

未经剔除虚假交易成分的商业数据，例如销售额、转化率等等，不能反映平台真实的业绩与成长，将误导商业决策

客户体验

刷单伪造交易历史骗取消费者对商品、品牌的认可；冲击平台的搜索排名体系以吸引流量，这些使得平台对商品质量的客观评价指标对消费者失去参考价值

电商生态

平台会因虚假交易损失在消费者当中的信誉；而刷单者的不正当竞争也会冲击卖方体系，造成商家流失；廉价的刷单成本更会侵蚀正规的营销渠道，例如广告业务

刷单行业现状

成为营销、赚钱手段

- 渗透各类综合、垂直电商以及O2O平台

规模化、市场化、产业化

- 涌现出各种刷单公司、集团、平台，“从业人员”众多，“产业”日趋成熟
- 提供一站式服务，并且敢于承诺“服务质量”

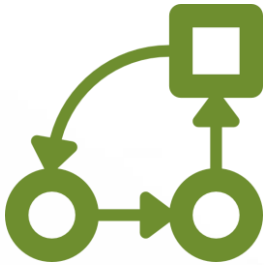
手法隐蔽、逼真、多变

- 多种刷单软件和工具协助提高效率以及隐藏刷单者真实身份
- “人肉”刷单，模拟真实客户行
- 刷单者通过社区分享反侦察心得

反刷单的挑战

多维度数据引证

基于单个行为特点的识别方法，面对逼真的刷手行为日渐困难，需要多种维度数据上深入挖掘实体信用指标作为依据



策略的敏捷迭代

适应刷单手法的变化，决策识别系统能够预警并演进

快速、准确与高召回

对层出不穷的刷单手段，既要抓得多，又要抓得准，还要抓得快

系统需求


Essentials
Adaptability
Reproducibility
Flexibility
Dimensionality

系统需求

分布式大数据系统基本需求

- 高可用性
- 可扩展性
- 低延迟

Essentials

系统需求

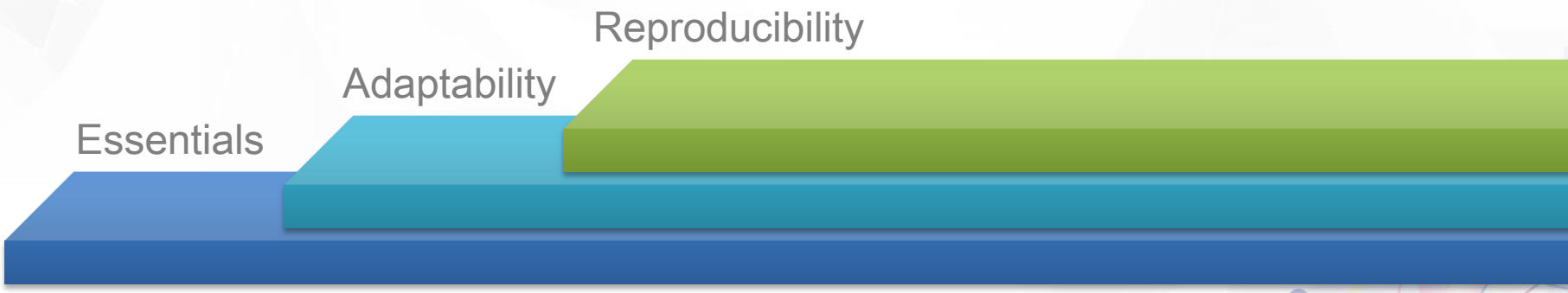
多样化数据源适应性

- 多种业务类型：订单、账户、支付、物流、评论
- 不同数据形式：批量数据（数据仓库）、流式数据（消息队列）、京东云
- 对于批量数据，解决到达时间不一致问题；对于流式数据，使用流式处理，同时落地为批量数据
- 数据的变化

系统需求

结果可复现性

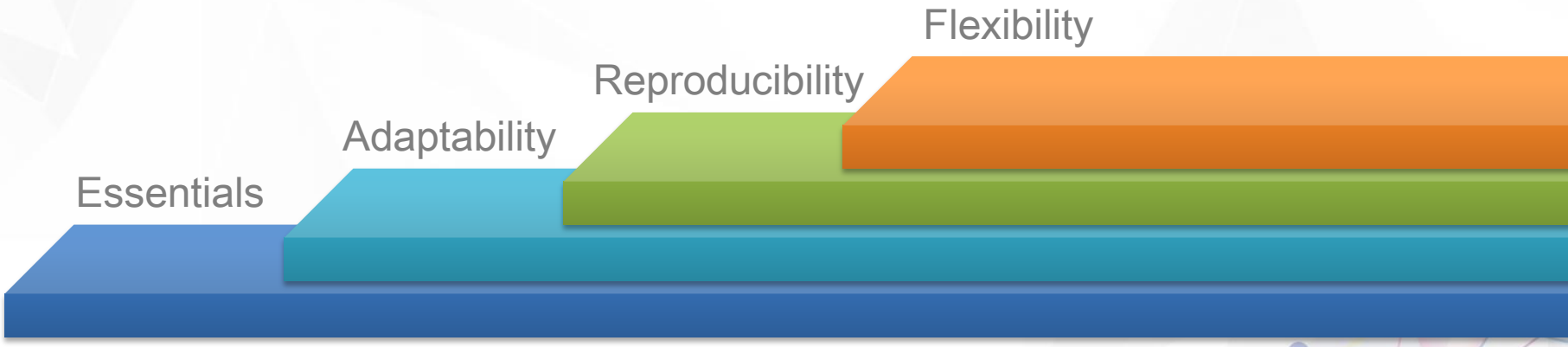
- 判定需要保留现场历史，以便回溯判定的过程
  - 当时点的用于生成特征的业务数据
  - 当时点的用于识别刷单的特征数据
  - 当时点的策略及系统（模型规则、参数、代码、配置）
- 有助于解决分歧、复议



系统需求

决策系统灵活性

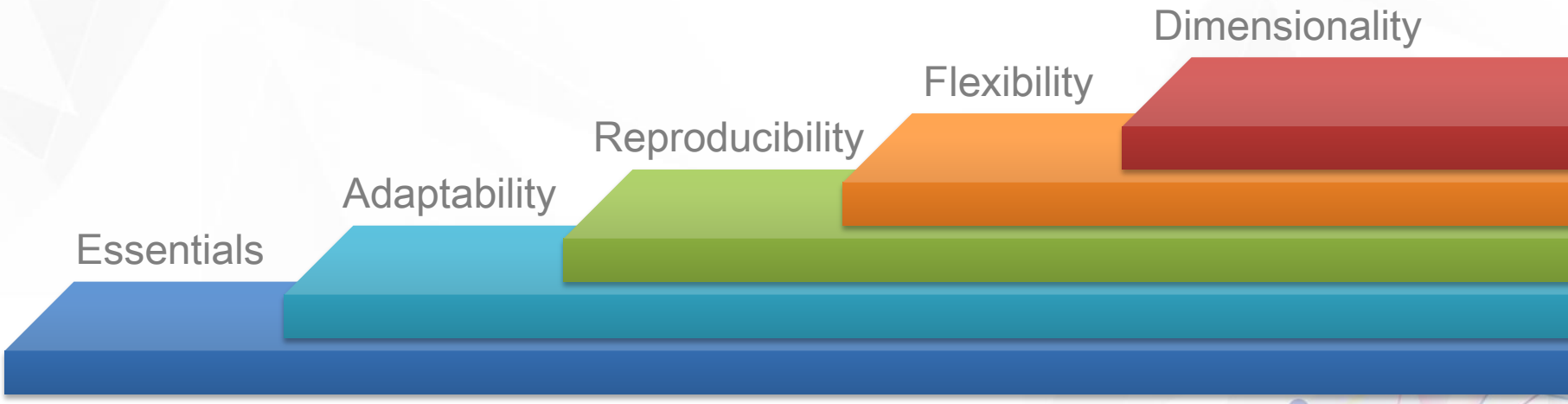
- 可扩展：支持多模型规则协作
- 热插拔：随时上线、下线模型规则，支持突发业务变更
- 应对业务变化：机器学习算法与业务规则结合；通用、稳定的模型与专用、易变的逻辑隔离



系统需求

服务多维度应用

- 识别结果在高维度上聚合，生成个体风险指标
- 除用于反刷单外，帮助构建风险、信用账户体系、商家信用体系、商品质量监控



京东订单交易数据

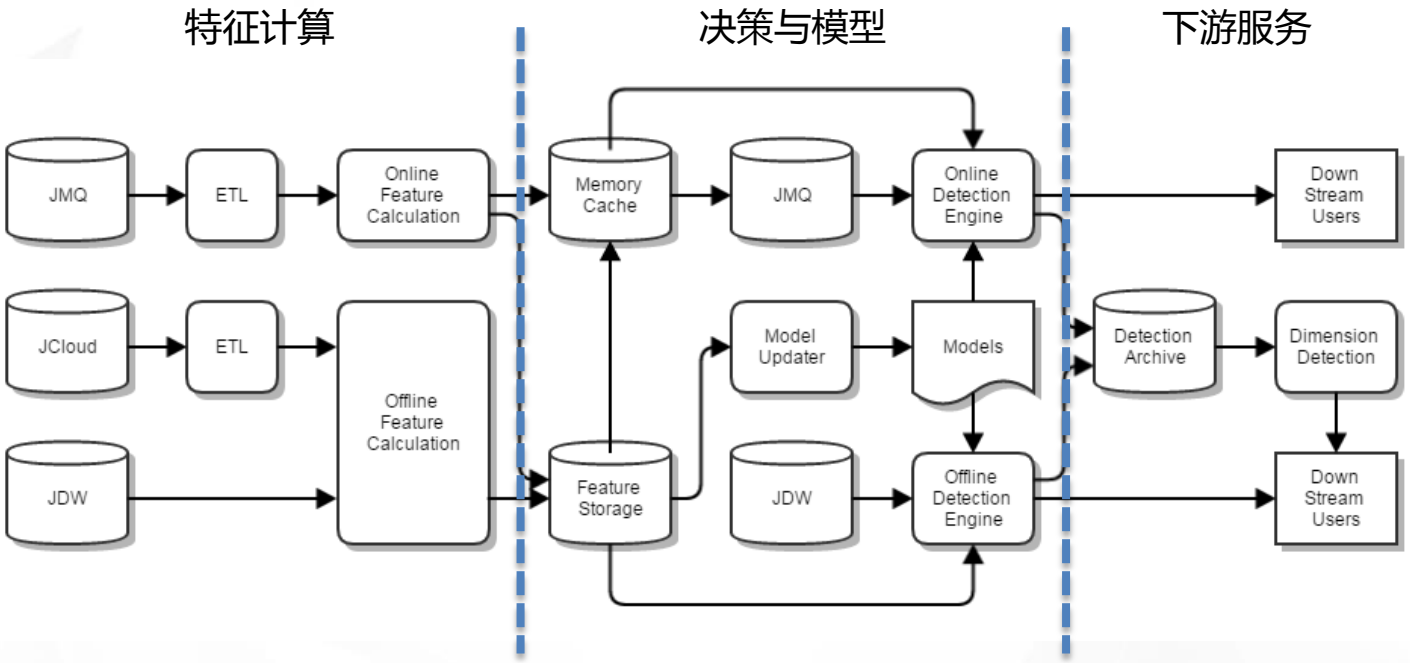
特点

- 生命周期长：从用户产生消费冲动到对商品发表评论，一个订单关联到的数据跨度可长达数周甚至数月
- 数据种类多：日志、买卖方属性、商品属性、交易属性、支付、物流等等数据
- 数据多变：在订单生命周期内交易数据的变动是十分常见的

反刷单：在更长的时间跨度上，从海量持续变动的数据中挖掘刷单行为的痕迹

系统架构设计实践

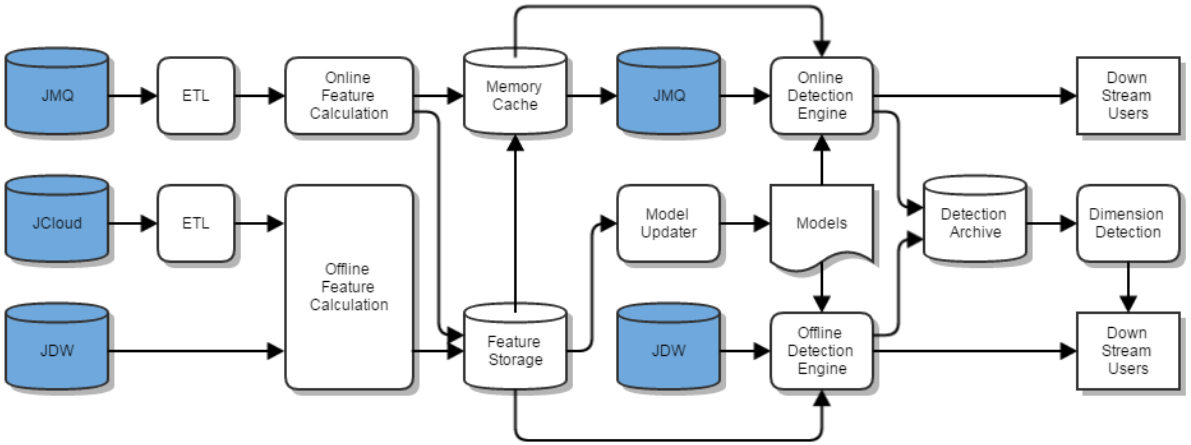
京东反刷单系统架构



Hadoop Stack + Spark

- 因地制宜，根据数据和作业的特点选择适合的数据处理技术
- 精简选择，用简洁一致的解决方案处理复杂多变需求

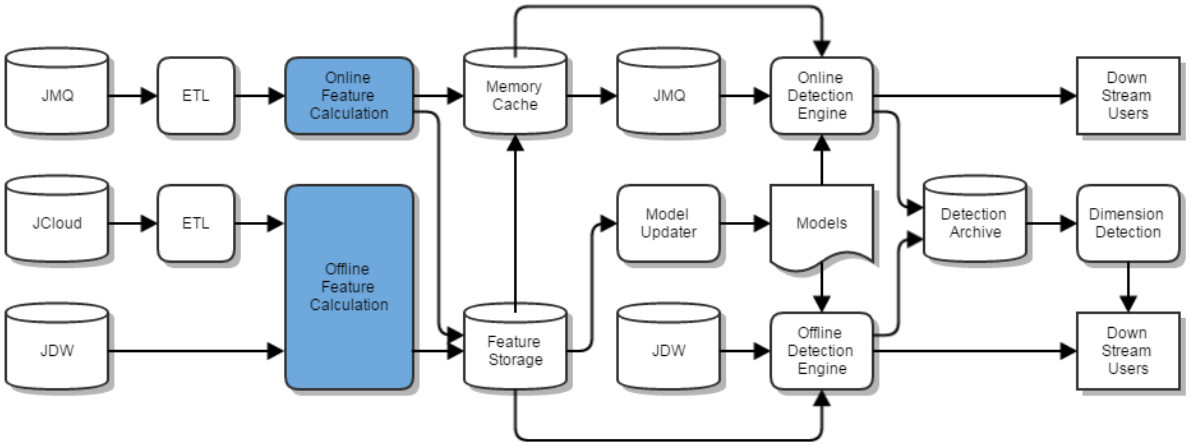
京东反刷单系统架构



数据预处理

- 批量数据、批量作业
  - 数据仓库：Hive + Pig
  - 云平台等数据源：定时任务
  - ETL：Pig
- 流式数据
  - 持久化：Camus
  - ETL：Spark Streaming
- 作业管理和调度：Oozie

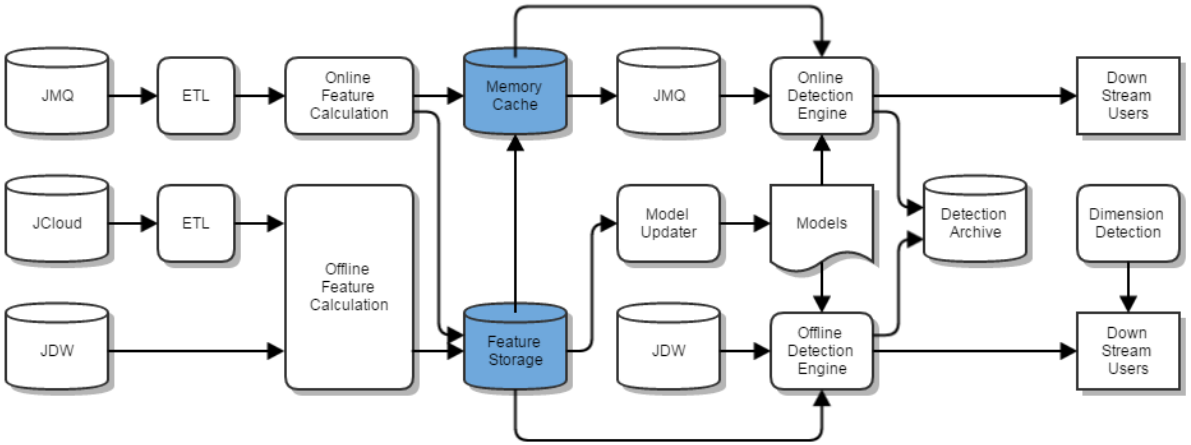
京东反刷单系统架构



特征计算

- 离线特征
  - 初级特征：特征工厂 (Feature Factory)
  - 高阶特征
    - 图模型算法：Spark GraphX
    - 传统机器学习方法：Spark MLlib
    - 聚类、序列分析等方法：自实现
- 在线特征
  - 时间窗口统计：Spark Streaming

京东反刷单系统架构

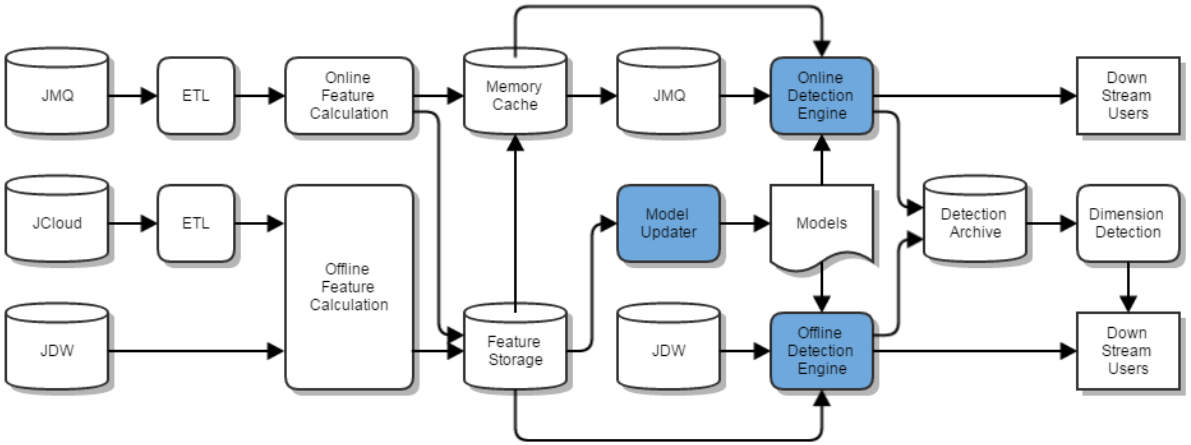


特征管理

- 离线特征：特征仓库 (Feature Warehouse)
  - 模型训练更新
  - 特征共享
- 在线特征：JimDB
  - 实时特征检索

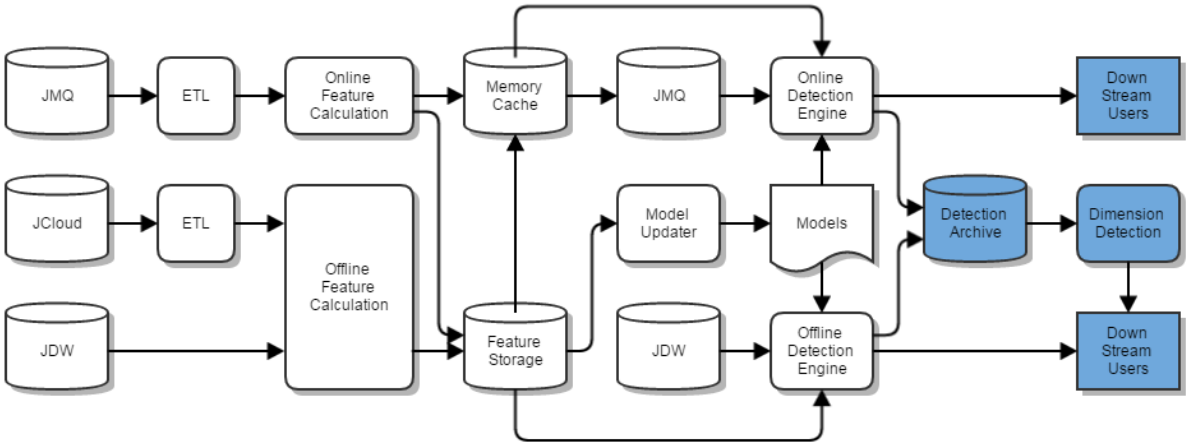
京东反刷单系统架构

模型与决策引擎系统



- 模型训练与更新
  - 浅层模型方法：Spark MLlib
  - 其他方法：自实现
  - 深度学习方法：评估调研中
- 决策系统
  - 基于模型方法：Spark
  - 基于规则方法：Drools

京东反刷单系统架构



结果归档与推送

- 归档
  - 数据压缩：Avro
- 推送
  - 实时请求：JSF RPC框架
  - 消息推送：JMQ

系统需求与架构实践

如何满足分布式系统基本需求？

需求	方案
高可用性	监控 + 主从、旁路系统
可扩展性	一切皆分布式
低延迟	
数据时效性	数据降级
计算容量不足	优化特征计算
订单交易数据属性	订单生命周期内多次识别刷单可疑度

系统需求与架构实践

- 监控无处不在
  - 任务监控
    - 开源框架：Oozie、Spark -> 集成原生监控
    - 京东框架：JMS、JimDB、JSF -> 京东统一监控平台
  - 数据质量监控
    - 上游数据/下游推送：量级监控
    - 离线/在线识别结果：识别统计报表与Dashboard
- 计算的权衡
  - 历史数据 <-> 最新数据
  - 手动优化 <-> 自动生成
  - 单次识别 <-> 多次识别
- 分段旁路系统
  - 候场环境与灾备：特征计算、模型与决策系统、下游推数三阶段

系统需求与架构实践

如何满足大数据系统基本需求？

如何适应多样化数据源？

需求	方案
多种业务类型	特征工厂，对特征计算中数据依赖和过程的高度抽象
不同数据形式	通过ETL将数据统一到两类：流式数据和批量数据

特征工厂：初级特征要素

空间
Space

对象的时间跨度以及筛选条件，如限于过往半年的订单记录，限于移动端日志

维度
Dimension

按空间筛选后，聚合的字段，如账户名、商品标识符

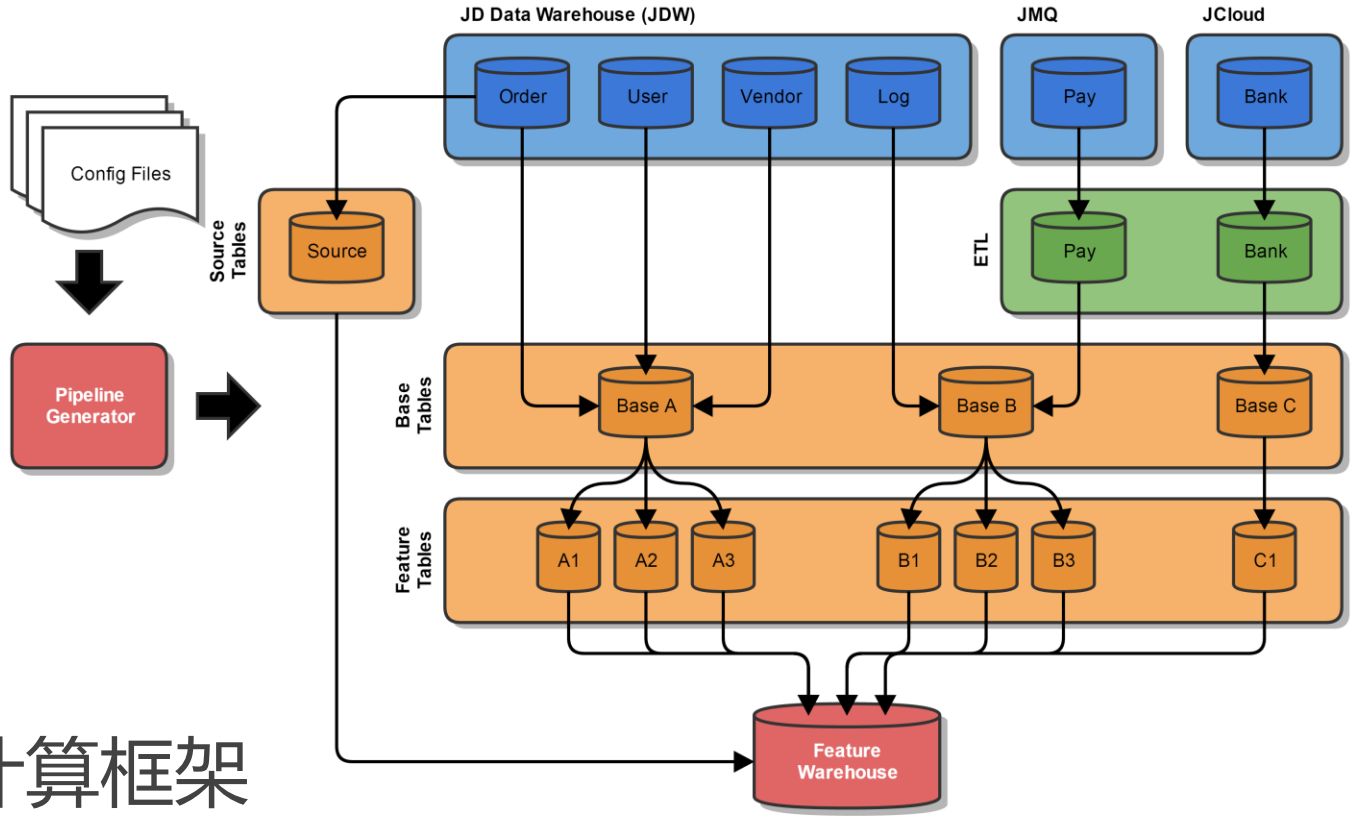
测度
Metrics

按维度聚合后，群组上的统计方式，如计数、均值、方差、信息增益

目标
Target

最后将测度按照维度关联到订单记录上时，目标的范围，如仅适用于当天的订单记录，仅适用于自营业务的订单记录

特征工厂：通用特征计算框架



初级特征计算框架

- 提供上述要素的配置语言表达形式
- 由配置自动构建计算特征的数据查询语言与作业调度

系统需求与架构实践

如何满足大数据系统基本需求？
如何适应多样化数据源？

如何让结果可复现？

需求	方案
保留现场历史	识别结果中包含 • 所有特征数据 • 所有模型识别结果 • 决策引擎配置 • 决策引擎代码版本（CI自动生成提交） 使用Avro压缩并存档

系统需求与架构实践

如何满足大数据系统基本需求？
如何适应多样化数据源？
如何让结果可复现？

如何提高决策系统灵活性？

需求	方案
可扩展	支持多模型协作的决策引擎 • 使用元分类器 + 模型的决策结构 • 元分类器下辖各类机器学习模型或者规则模型 • 元分类器与模型构成决策引擎拓扑
热插拔	决策引擎拓扑由配置动态生成 • 离线系统：配置更改即生效 • 在线系统：配置更改后由定时任务侦测并更新

系统需求与架构实践

如何满足大数据系统基本需求？
如何适应多样化数据源？
如何让结果可复现？
如何提高决策系统灵活性？

服务多维度应用

服务下游应用

THANKS

SequeMedia 盛拓传媒 | IT168.com 专注导购16年 | ChinaUnix.net | ITPUB www.itpub.net