**电子商务实时数仓**

**利用数据仓库分层：**

**ODS层（原始数据层）{加载原始日志，数据，数据保持不做加工处理}**

**DIM层（维度层）{保存维度数据，主要对业务事实的描述信息，例如何人，何事等}**

**DWD层（明显数据层）{对ODS层数据进行清洗，主要去除空值，脏数据，超过极限范围的数据、脱敏等保存业务事实明显表，例如一行信息代表一次业务行为，例如一次下单 }**

**DWS层（服务数据层）{以DWD为基础，按天进行轻度汇总，一行信息代表一个主题对象一天的汇总行为，例如一个用户一天下单次数}**

**DWT层（数据服务层）{以DWS层位基础，对数据进行累计汇总，例如一行信息代表一个主题对象的累计行为，一个用户从注册那天开始至今天一共下了多少次单}**

**ADS层（数据应用层）{为各种统计做报表提供数据}**

（**背景**）

电子商务数据仓库，是在电子商务迅速发展的背景下产生的一种数据管理与应用系统。它主要用于存储、处理和分析电商平台在运营过程中产生的大量数据，为企业的决策提供数据支持。

(需求)

1电子商务的快速发展

互联网技术的广泛应用，电子商务市场迅速扩张，交易额逐年增长，电商平台积累了海量的用户行为数据、交易数据、商品数据等，为电商数仓的建设提供了丰富的数据资源。

2数据驱动的决策需求

企业需要通过对数据的深入分析，了解市场趋势、用户需求和自身运营状况，从而制定合理的营销策略、优化商品供应链和提高用户满意度。电商数仓为数据驱动的决策提供了基础

3大数据技术的进步

大数据技术得到了迅猛发展，为电商数仓的建设提供了技术支持。通过大数据技术，企业可以实现对海量数据的快速存储、处理和分析，挖掘数据的价值

4数据隐私和安全法规

数据隐私和安全法规不断完善，企业需要合规地收集、存储和使用用户数据。电商数仓在确保数据安全的前提下，为企业提供了合规的数据应用能力。

5业务场景多样化

电商平台涉及多个业务场景，如商品推荐、精准营销、用户画像、库存管理等。电商数仓可以根据不同业务场景，提供定制化的数据解决方案。

6企业数字化转型

在数字化时代（海量数据（消费的购物习惯）），企业纷纷进行数字化转型，以提升竞争力。电商数仓作为企业数字化基础设施的重要组成部分，有助于企业实现数据资产的积累和变现。

**1 CDH3台**

**2配置虚拟网络编辑器子网**

**3镜像环境的搭建**

**4 maxwell的安装**

**5 HBase的安装**

**6 kafka的安装**

**7 Doris的安装**

**商品主题（技术）**

1订单事实表

2取消订单事实表

3交易退款支付成功事实表

4加购独立用户

5首页、详情页聚合统计

6按照版本、地区、渠道、新老访客对PV、UV、SV、dur进行聚合统计

**商品主题（技术）**

1订单事实表

****目的：****

****记录用户实际完成的购买行为，反映商品销售情况和用户消费趋势。****

****分析角度：****

****时间维度：日 / 周 / 月销量趋势、季节性波动（如节假日峰值）。****

****商品维度：热销 / 滞销商品排行、品类销售占比。****

****用户维度：新老用户订单量对比、高价值用户订单分布。****

****地区维度：地域销量差异、区域偏好分析。****

****渠道维度：不同流量来源（如 APP、小程序）的订单转化率。****

1. 取消订单事实表

**目的：**

**分析订单流失原因，优化购物流程，提升转化率。**

**分析角度：**

**时间维度：订单取消时间分布（如支付前 / 支付后取消）。**

**商品维度：高取消率商品特征（如价格区间、库存状态）。**

**用户维度：新用户取消率是否高于老用户、高频取消用户画像。**

**原因维度：取消原因分类（如价格敏感、配送问题、临时改变主意）。**

**流程节点：购物车放弃率、支付环节流失率。**

**3. 交易退款支付成功事实表**

**目的：**

**监控售后情况，评估商品质量和用户满意度。**

**分析角度：**

**时间维度：退款时间周期（如 7 天无理由退款占比）、退款高峰期。**

**商品维度：高退款率商品清单、质量问题集中的品类。**

**用户维度：退款用户复购率、恶意退款行为识别。**

**原因维度：退款原因分类（如商品质量、尺寸不符、描述不符）。**

**财务影响：退款金额占 GMV 比例、退款对利润率的影响。**

**4. 加购独立用户**

**目的：**

**衡量商品吸引力和用户购买意愿，优化商品推荐策略。**

**分析角度：**

**时间维度：加购用户数变化趋势、与促销活动的关联性。**

**商品维度：高加购率商品与实际转化率对比、加购但未购买的商品特征。**

**用户维度：不同用户群体（如新老用户、高价值用户）的加购偏好。**

**行为路径：加购后流失的用户后续行为（如是否转购竞品）。**

**营销效果：加购提醒功能对转化率的提升效果。**

**5. 首页、详情页聚合统计**

**目的：**

**优化页面布局和内容，提升用户体验和商品曝光率。**

**分析角度：**

**流量指标：PV（页面浏览量）、UV（独立访客数）、SV（会话数）。**

**停留指标：平均停留时间（dur）、页面跳出率。**

**位置分析：首页 / 详情页不同位置商品的点击 / 加购转化率。**

**版本差异：A/B 测试中不同页面版本的用户行为对比。**

**漏斗分析：从首页浏览到详情页再到下单的转化路径。**

**6. 按版本、地区、渠道、新老访客对 PV、UV、SV、dur 进行聚合统计**

**目的：**

**精细化分析用户行为，针对性优化运营策略。**

**分析角度：**

**版本维度：不同 APP / 小程序版本的用户活跃度差异，新版本功能对用户停留的影响。**

**地区维度：各地区用户访问习惯（如一线城市与下沉市场的浏览时长差异）。**

**渠道维度：不同推广渠道（如社交媒体、搜索引擎、广告投放）带来的流量质量对比。**

**新老访客维度：新用户浏览深度是否低于老用户、老用户回流周期分析。**

**交叉分析：例如某地区新用户通过特定渠道访问的停留时间异常低，可能提示本地化内容不足。**

统计过后能解决了以下问题：

商品策略优化

用户运营提升

渠道效果分析

活动效果评估

页面优化

异常订单监控

业务波动预警

**商品策略优化**

通过订单事实表和加购独立用户等指标分析，能明确不同商品的销售状况和用户购买意愿，助力企业精准选品，合理规划库存，减少滞销商品积压，提高资金周转率。

依据交易退款支付成功事实表中商品的退款率及原因，可识别出存在质量或描述问题的商品，以便及时改进产品或优化商品信息，提升用户满意度和商品竞争力。

**用户运营提升**

利用新老访客的相关统计指标，如按新老访客对 PV、UV、SV、dur 进行聚合统计，能深入了解不同用户群体的行为特征和需求差异，从而制定个性化的运营策略。对新用户，可设计针对性的引导流程和优惠活动，提高其留存和转化；对老用户，通过分析其购买频率、浏览偏好等，提供精准的推荐和专属服务，增强用户忠诚度。

加购独立用户

指标能帮助企业发现潜在购买意向较高的用户，通过加购提醒、个性化推荐等方式，促进这部分用户完成购买转化，提高整体转化率。

**营销效果评估**

**渠道效果分析**

按渠道对 PV、UV 等指标进行聚合统计，可清晰评估不同推广渠道带来的流量规模和质量，帮助企业明确哪些渠道的获客效果好，哪些渠道需要优化或调整投入。例如，若发现某个渠道的 UV 较高但转化率较低，可深入分析原因，是渠道受众与目标用户不匹配，还是落地页设计存在问题等，进而有针对性地进行优化。

**活动效果评估**

结合时间维度分析订单事实表、加购独立用户等指标在促销活动期间的变化情况，能准确衡量活动的效果。如活动期间订单量、加购用户数是否显著增长，活动前后用户的浏览行为和购买偏好有无变化等，为后续活动策划和优化提供数据支持，提高营销活动的投资回报率。

**产品优化与体验提升**

**页面优化**

首页、详情页聚合统计中的 PV、UV、平均停留时间、跳出率等指标，能直观反映用户对页面的关注度和使用体验。若某一页面的跳出率过高，可能意味着页面布局不合理、内容不吸引人或加载速度过慢，企业可据此对页面进行优化，如调整商品展示顺序、优化图片和文案、改善页面性能等，以提高用户的停留时间和购买转化率。

通过分析不同版本的相关指标差异，可评估产品迭代的效果。如果新版本上线后，某些关键指标如用户停留时间、转化率等有明显提升，说明产品优化方向正确；反之，则需要进一步分析原因，及时调整优化策略。

**风险监控与防范**

**异常订单监控**

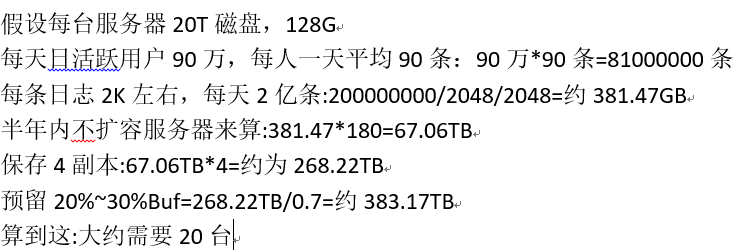
订单事实表、取消订单事实表和交易退款支付成功事实表等数据结合分析，有助于及时发现异常订单行为，如大量同一用户或同一地区的取消订单、退款订单，可能存在恶意刷单、欺诈等风险。通过设置预警机制，企业可及时采取措施进行调查和处理，降低损失。

**业务波动预警**

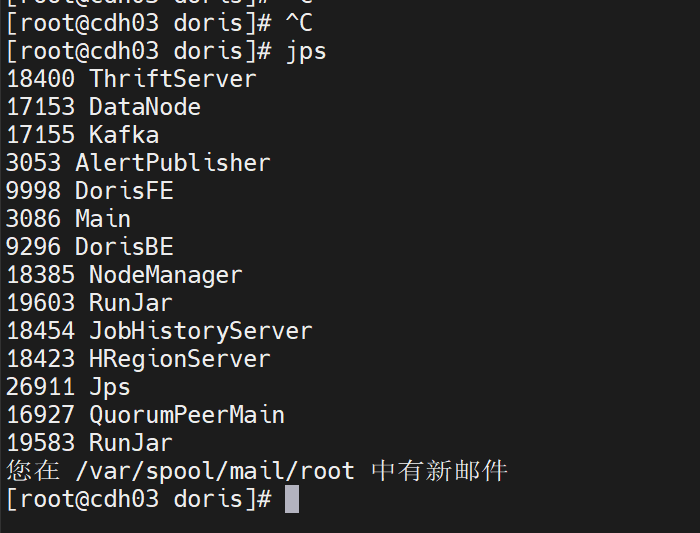
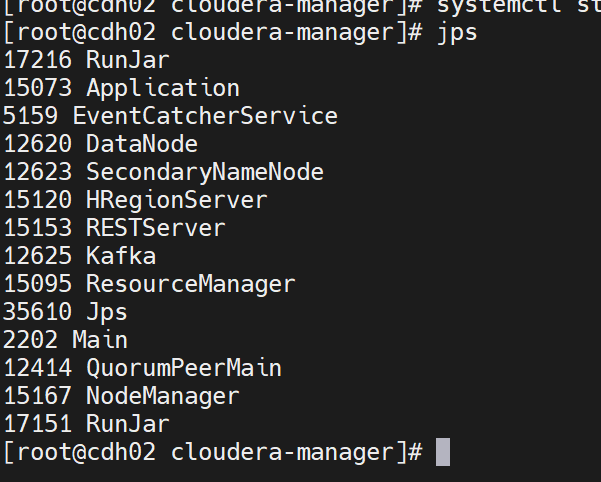
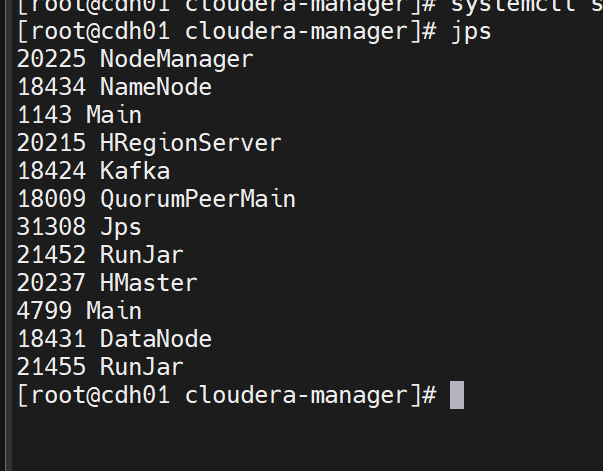
持续监测各指标的变化趋势，当出现异常波动时，如订单量突然大幅下降、首页 PV 骤减等，能及时发出预警信号，提醒企业关注业务状况，提前排查可能的原因，如市场竞争加剧、产品出现重大问题或外部环境变化等，以便企业及时调整策略，应对潜在风险

**技术方面:**

**框架选型:**



**架构（技术层）**



**技术选型（最新版本号）:**

数据采集传输:Flume（1.9,0）,sqoop（1.4.6）

数据存储:MySQl（5.7.16）,HDFS（3.1.3）,HBase（2.0.5）

数据计算:Hive（3.1.2）,Spark（3.0.0）

数据查询:Presto（0.189）

数据可视化:Echarts（5.5.1）

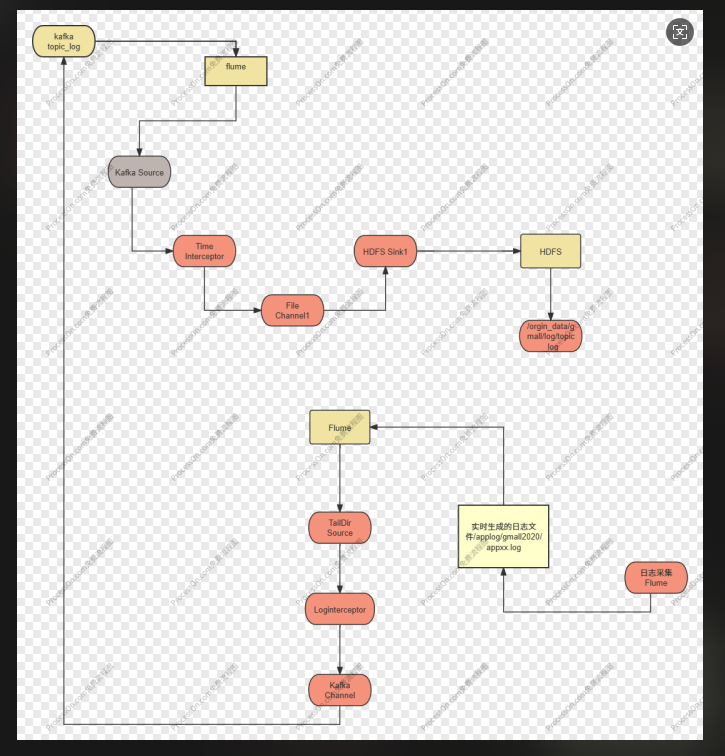
任务调度:Azkaban（3.85.4）

集群监控:Zabbix（7.2.0）

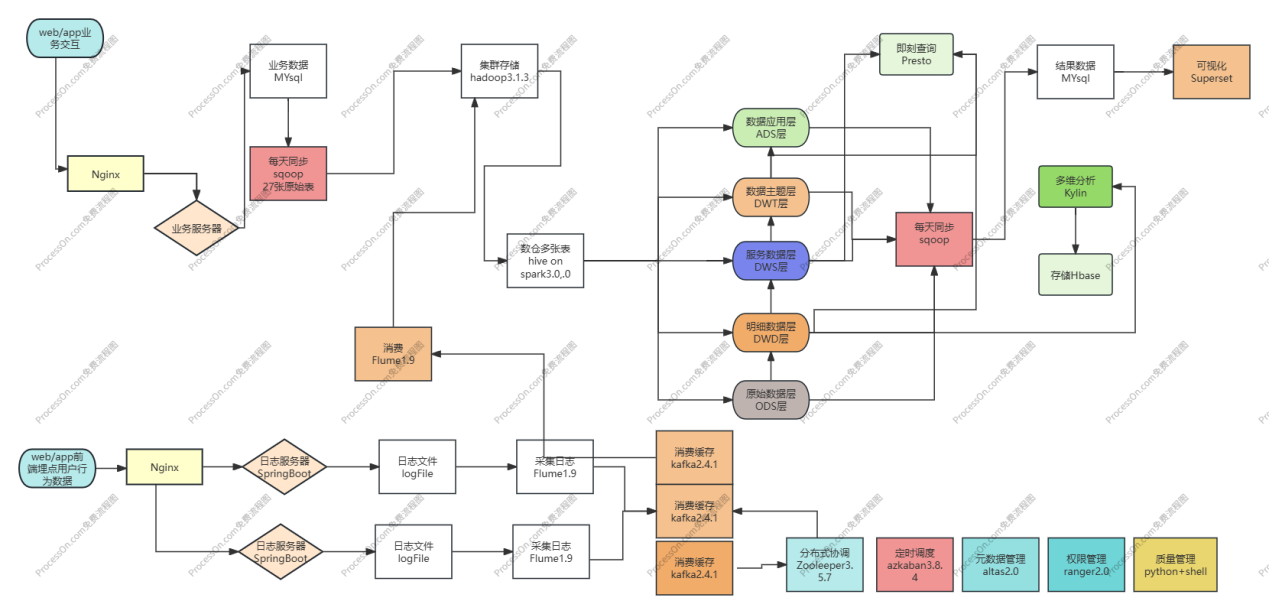
元数据管理:Atals（2.0.0）

权限管理:Ranger（2.0.0）

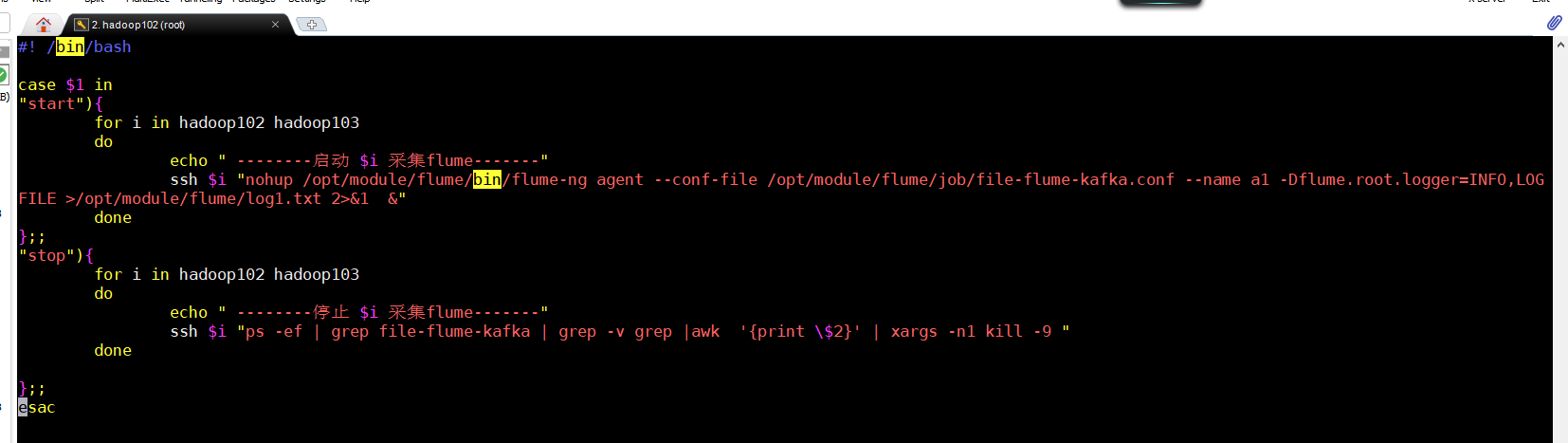
**Kafka结构流程**

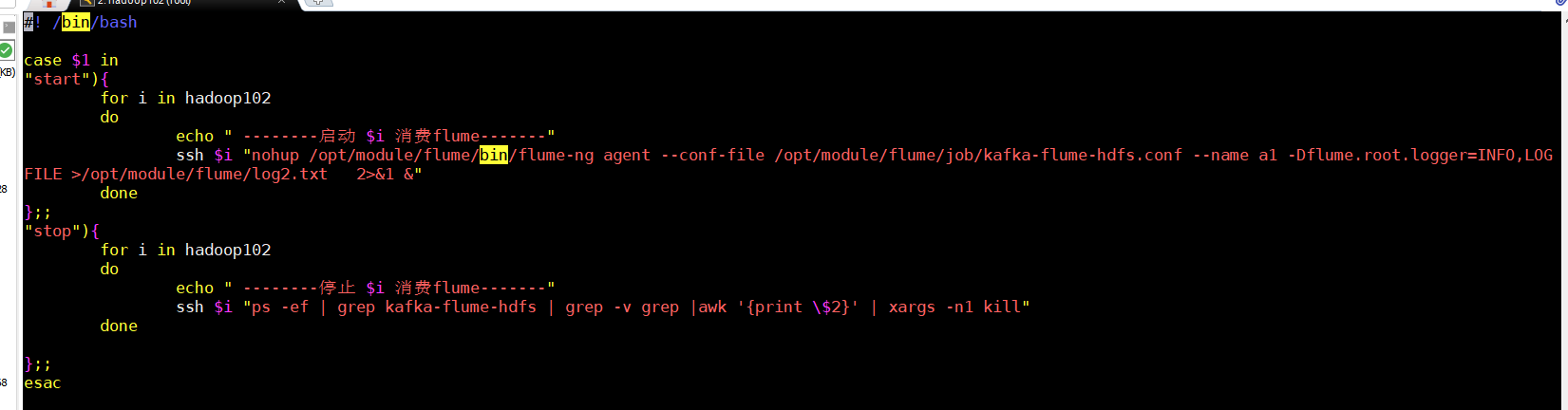


**日志数据采集**



**flume怎么采集流程**





**测试清单**

**flume优化：**

调整内存设置：

增加Flume进程的JVM堆内存大小，以处理更大的数据量。

调整Flume的内存参数，如heap-size和direct-memory-size，以优化内存使用。

选择合适的Source：

根据数据源的特性选择合适的Source类型，例如，对于高吞吐量的场景，可以使用ExecSource或TailDirSource。

优化Channel选择：

使用内存效率更高的Channel，如MemoryChannel，但要注意数据的安全性。

对于需要数据持久性的场景，可以使用FileChannel，并适当调整其缓存目录和事务容量。

调整Channel参数：

增加Channel的容量（capacity）和事务大小（transactionCapacity），以提高吞吐量。

调整keep-alive参数，以防止Sink处理缓慢导致Source被阻塞。

优化Sink配置：

根据Sink的类型（如HDFS、Kafka等）调整相关参数，如批量大小、刷新间隔等。

使用异步Sink，如AsyncHBaseSink，以减少I/O等待时间。

多Agent级联：

在数据量大或网络延迟较高的场景下，使用多级Flume Agent进行数据聚合，减少单点压力。

网络优化：

优化网络配置，如增加带宽、减少网络延迟。

使用压缩数据传输，减少网络负载。

负载均衡：

使用Flume的负载均衡功能，如通过LoadBalancingSinkProcessor将数据均匀分发到多个Sink。

监控和日志：

启用Flume的监控功能，如JMX，以便实时监控Flume的性能指标。

分析Flume日志，识别性能瓶颈和错误。

系统优化：

优化操作系统参数，如网络缓冲区大小、文件系统缓存等。

确保磁盘I/O性能足够高，特别是对于使用FileChannel的场景。

自定义Interceptor：

使用自定义Interceptor来预处理数据，如过滤、转换和聚合，以减少下游系统的负担。