西门子大数据平台分析

作者：　谭璧栋　日期：2019-10-21

目录

1 、系统总体设计

1.1 需求规定：对天猫西门子交易数据进行处理，获取最终报表信息。

1）Rf_Info_User（用户信息）

买家会员名，联系手机，用户总积分，会员等级、交易总额，退款总额

2）Rf_Info_Order（订单信息）

买家会员名，订单编号，交易明细、宝贝标题，商家编码，价格，购买数量，订单状态

1.2 运行环境

操作系统：Linux

软件环境：Hadoop 平台　HBase　Hive　Sqoop　Zookeeper　Mysql　Oozie

1.3 整体设计逻辑介绍：

第一步：因为存在中文，hive 识别的 utf-8 编码，使用工具将文件转换成所需的编码格式。然后每天定时将交易数据放到 LInux 指定存放数据的文件夹。

第二步：每天利用 Oozie 将工作流整合，并且定时扫描存放数据的文件夹，判断是否有新数据。

有则将数据导入到 HDFS 指定位置，然后备份一份元数据。

如果没有数据，则发出警告，然后过一段时间再次扫描文件夹。

第三步：处理原始数据，数据导入 Hive 表时自动除去脏数据。

1）数据存在双引号

2）删除数据第一行的字段。

第四步：将处理完的数据导入 HDFS 做数据备份。

第五步：定义数据整合规则，判断会员的信息，并且对会员信息进行更新、追加。

第六步：执行 Hive 语句，通过筛选出元数据中的信息，并且将处理好的数据导入相应的三张表：TM_ItemList、TM_OrderList、TM_Return。

第七步：执行 Hive 语句，筛选出所需要的信息，并且创建临时表。

第八步：定义会员积分、会员等级规则。

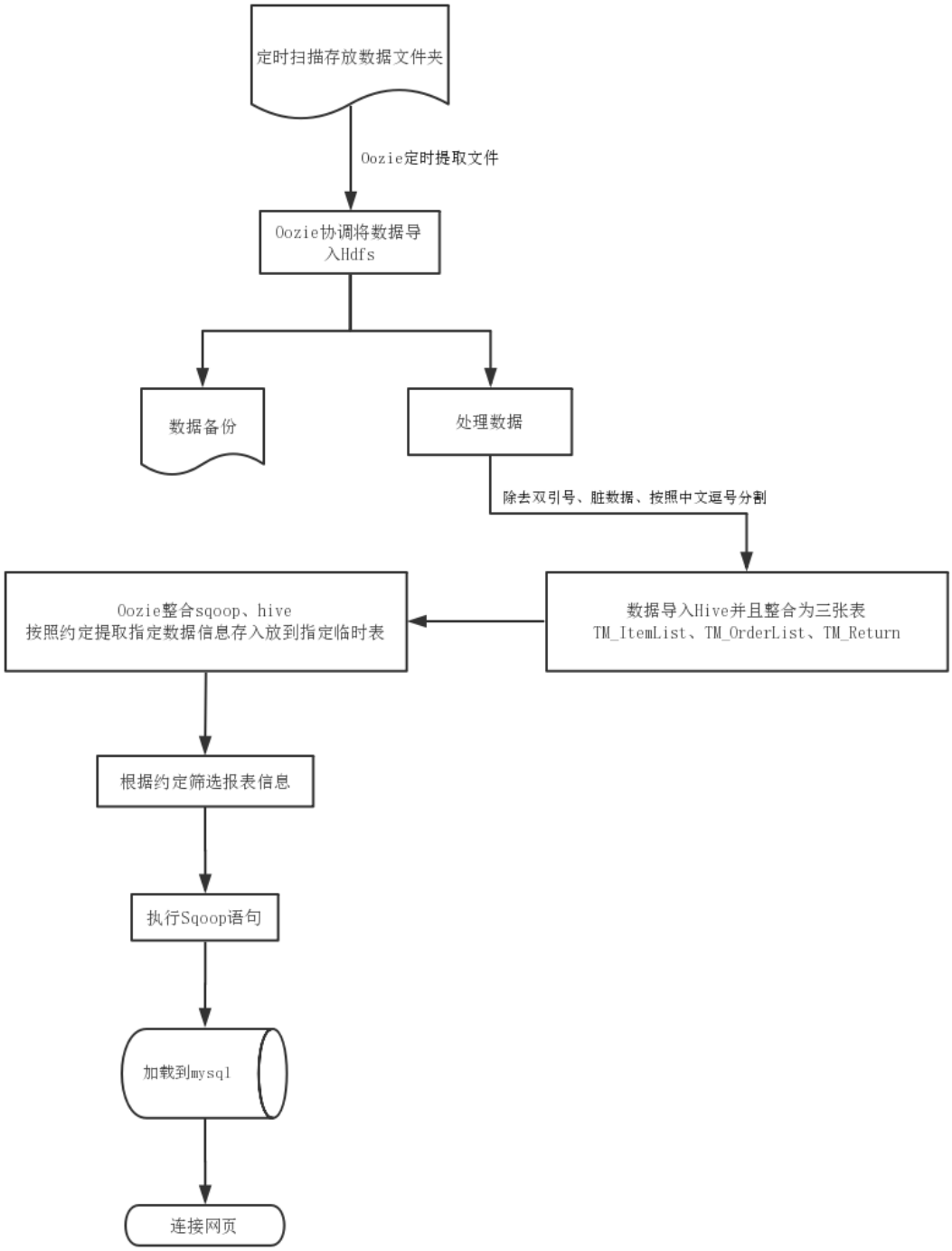
第九步：根据对会员积分、会员等级规则，执行 hive 语句筛选出临时表中对应的数据。

第十步：通过 Sqoop 将筛选出的信息导入 MySQL。

第十一步：通过 SpringBoot+Vue 等技术编辑创建前端、后端平台，并且通过 mybatis 连接 MySQL 数据库，实现数据报表的展现。

第十二步：运行测试。

1.4 业务逻辑图


定时扫描存放数据文件夹
Oozie定时提取文件
Oozie协调将数据导入Hdfs
数据备份
处理数据
除去双引号、脏数据、按照中文逗号分割
数据导入Hive并且整合为三张表
TM_ItemList、TM_OrderList、TM_Return
Oozie整合sqoop、hive
按照约定提取指定数据信息存入放到指定临时表
根据约定筛选报表信息
执行Sqoop语句
加载到mysql
连接网页

2 、数据表基本设计思路和处理流程

1、提取数据，并且放到指定目录下

2、通过 Oozie 在指定时间，自动对存放数据的目录进行扫描，提取实时更新的原始数据，并且导入 HBASE，然后做数据备份。

3、采用 hadoop 体系结构对大批量的数据进行运算，筛选数据，排除脏数据。采用 hadoop 集群的方式对大数据进行运算。

4、将相同的数据类型将元数据整合为三张表：

TM_ItemList：表存放交易明细

TM_OrderList：表里存放用户的交易方面的信息

TM_Return：表里面存放的用户退货方面的信息

5、筛选出三张表中数据、存放到五张临时表中

1）Member_Deal（积分临时表）

买家会员名、订单编号、买家实际支付金额、退款编号、退款状态、买家退款金额

2）All_Money（金额表）

买家会员名、交易总金额、退款总金额

3）Member_Points_Grade（会员积分等级表）

买家会员名、会员积分、会员等级

4）Info_Member（会员信息表）

买家会员名 、买家支付宝账号、收货人姓名、 收货地址 、联系电话、联系手机

5）Info_Order（订单信息表）

订单编号、总金额、买家实际支付金额、订单状态 、订单创建时间、订单付款时间、宝贝标题、宝贝种类、店铺 Id、店铺名称、确认收货时间、买家会员名

6、创建存放报表数据的代理表，用于将报表数据通过 sqoop 传递到 MySQL。

7、创建会员积分等级表（Member_Points_Grade）

1）积分规则如下：

通过产品消费金额来计算。

1）每笔交易明细一笔积分。

2）实际支付金额 10 元以下的不算积分。

3）消费 1 元，发放 100 积分。

4）积分类型交易积分、退货积分。

5）积分=总金额*100-退款金额*100

2）会员需要有等级。

通过产品金额来计算，

1）会员消费在 0-2000 之间 普通会员。

2）会员消费在 2000-5000 之间 银卡会员。

3）会员消费在 5000 以上 金卡会员。

8、从五张临时表中数据进行计算以及筛选，最后生成客户所需的报表信息

1）My_Info_User（用户信息）

买家会员名，联系手机，用户总积分，会员等级、交易总额，退款总额

2）My_Info_Order（订单信息）

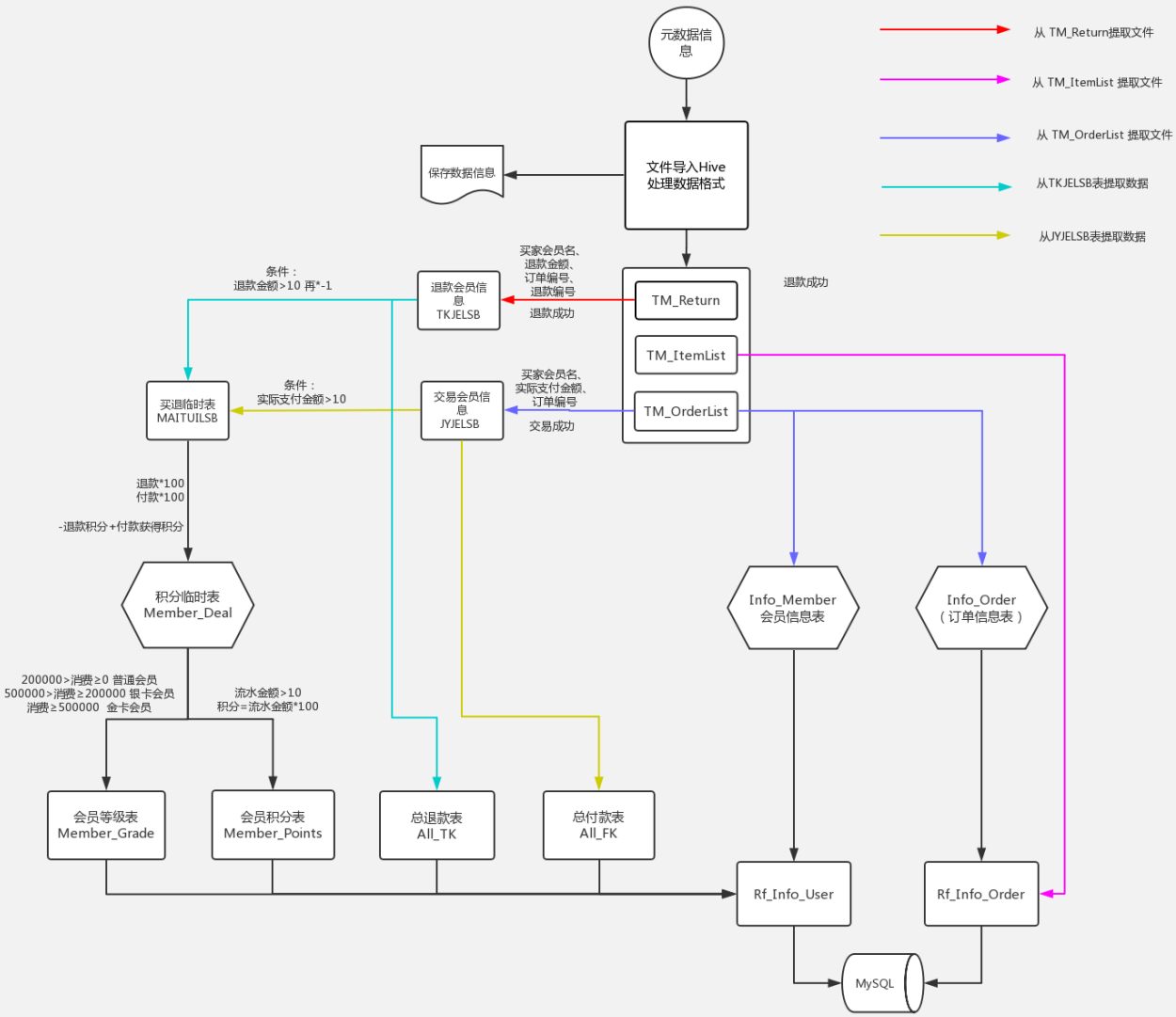
买家会员名，订单编号，交易明细、宝贝标题,商家编码,价格,购买数量,订单状态

9、将所得报表导入 mysql 数据库

10、前端网页，连接数据

3 、数据流程

3.1 项目数据流向图



3.2 Hive 中数据表结构设计

Info_Member（会员信息表）		
字段名	类型	中文描述
Member_Name	string	买家会员名
Alipay_Account_Number	string	买家支付宝账号
Name_Of_Oonsignee	string	收货人姓名
Address	string	收货地址
Contact_Phone_Number	string	联系电话
Contact_Phone	string	联系手机

Info_Order（订单信息表）		
字段名	类型	中文描述
OrderID	string	订单编号
Amount	double	总金额
Actual_Amount	double	买家实际支付金额
OrderState	string	订单状态
Order_Creation_Time	string	订单创建时间
Order_Payment_Time	string	订单付款时间
Baby_Title	string	宝贝标题
Types_Of_Baby	string	宝贝种类
StoreId	string	店铺 Id
Shop_Name	string	店铺名称
Confirm_Receipt_Time	string	确认收货时间
Member_Name	string	买家会员名

Member_Deal（积分临时表）		
字段名	类型	中文描述
Member_Name	string	买家会员名
OrderID	string	订单编号
Actual_Amount	double	买家实际支付金额
Refund_Number	string	退款编号
Refund_Status	string	退款状态
Refund_Amount	double	买家退款金额

备注：通过对 TM_OrderList、TM_Return 两张表中的数据进行筛选，最终汇总成积分临时表，用于对会员的支付情况情况进行统计。根据要求算出会员的积分情况，然后按照要求出会员的积分等级。

All_Money（金额表）		
字段名	类型	中文描述
Amount_Trade	double	总交易金额
Refund_Amount	Double	退款总金额

Member_Points_Grade（会员积分等级表）		
字段名	数据类型	中文描述
Member_Name	string	买家会员名
Member_Points	double	会员积分
Member_Grade	string	会员等级

会员积分=（买家付款金额-买家退款金额）*100

会员等级：2000>消费≥0 普通会员

5000>消费≥2000 银卡会员

消费≥5000 金卡会员

创建存放最终报表数据的两个代理表

这两张表存放最终的报表信息，这样可以通过 sqoop 直接将报表数据传递到 MySQL 中对应的报表中。

1）Rf_Info_User（用户信息）

买家会员名，联系手机，用户总积分，会员等级、交易总额，退款总额

2）Rf_Info_Order（订单信息）

买家会员名，订单编号，交易明细、宝贝标题，商家编码，价格，购买数量，订单状态

Rf_Info_User（用户信息）			
字段名	类型	约束	中文描述
Member_Name	varchar(255)	主键	买家会员名
Contact_Phone	varchar(255)		联系手机
Member_Points	double		用户总积分
Member_Grade	varchar(255)		会员等级
Amount_Trade	double		交易总额
Refund_Amount	double		退款总额

Rf_Info_Order（订单信息）			
字段名	类型	约束	中文描述
Member_Name	varchar(255)	主键	买家会员名
OrderID	varchar(255)		订单编号
Deal_Detail	double		交易明细
Baby_Title	varchar(255)		宝贝标题
StoreId	varchar(255)		商家编码
Price	double		价格
Buy_Quantity	varchar(255)		购买数量
OrderState	varchar(255)		订单状态

3.3 数据库表结构设计

My_Info_User（用户信息）			
字段名	类型	约束	中文描述
Member_Name	varchar(255)	主键	买家会员名
Contact_Phone	varchar(255)		联系手机
Member_Points	double		用户总积分
Member_Grade	varchar(255)		会员等级
Amount_Trade	double		交易总额
Refund_Amount	double		退款总额

My_Info_Order（订单信息）			
字段名	类型	约束	中文描述
Member_Name	varchar(255)	主键	买家会员名
OrderID	varchar(255)		订单编号
Deal_Detail	double		交易明细
Baby_Title	varchar(255)		宝贝标题
StoreId	varchar(255)		商家编码
Price	double		价格
Buy_Quantity	varchar(255)		购买数量
OrderState	varchar(255)		订单状态

4 、数据整合

4.1 数据整合规则

记录会员交易记录，以及会员的电话号码、收货地址、收货人姓名的变动。

因为一个会员可能有多个地址，多个收货联系人，多个收货手机号。

整合规则：以第一次进入数据库的记录为准。因此当数据传入 hive 时，需要判断会员对应的信息，针对一个会员可能存在多个收货地址，多个收货联系人，多个收货联系电话的情况。对这些信息与用户的原有信息进行判断。从而达到对用户信息的追加以及更新。

具体规则：

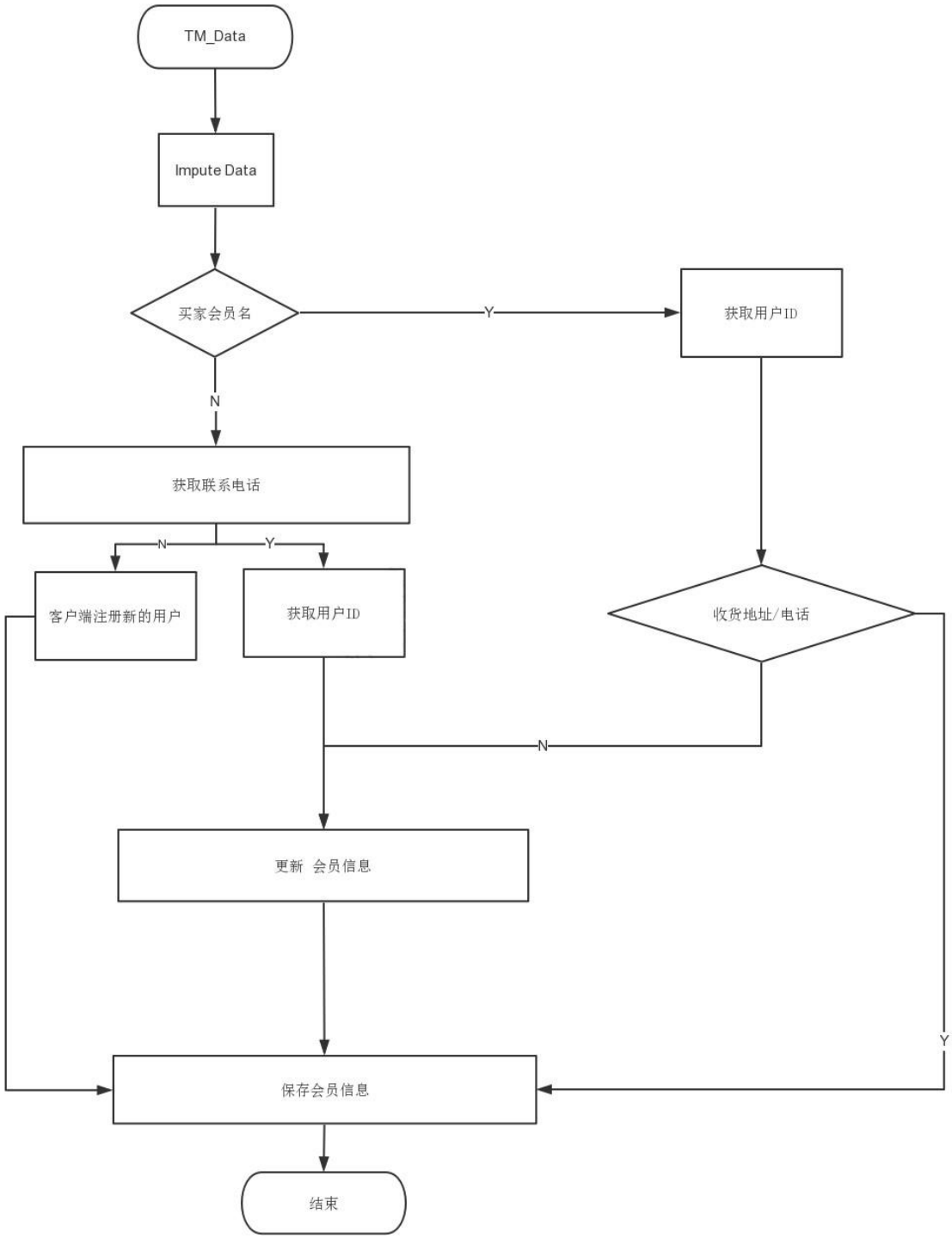
1）会员名如果不存在：则再次判断是否修改会员电话号码或者是收货地址

①若电话号码不存在，则注册新的用户。

②若存在，则判断会员名称是否相同，若相同，则更新电话号码。

2）会员名如果存在：通过会员名获取会员地址和联系方式，追加或更新会员交易信息

4.2 数据整合设计图：


TM_Data
Impute Data
买家会员名
Y
获取用户ID
N
获取联系电话
N
Y
客户端注册新的用户
获取用户ID
收货地址/电话
N
更新 会员信息
Y
保存会员信息
结束