|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| XX数据仓库 | | | | | |
| 整体设计方案 | | | | | |
|  | | | | | |
| 大数据架构师股份有限公司 | | | | | |
|  |
|  |
|  |
|  |
| 档案编号 | XX-SJCK-ZTSJFA01 | | | 版本号 | V1.0 |  |
| 编制 | 彭文华 | 审批 | 彭文华 | 生效日期 | 2020年9月12日 |  |
|  | | | | | |  |
|  | | | | | |  |
|  | | | | | |  |
|  | | | | | |  |

变更记录

|  |  |  |  |
| --- | --- | --- | --- |
| 版本内容 | 更改人 | 审批人 | 更改日期 |
| 创建文档 | 彭文华 | 彭文华 | 2020-9-12 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**目 录**

1. 编写指南 4

2. 业务需求分析 5

2.1. 业务访谈 5

2.1.1. 统计需求 5

2.1.2. 分析需求 5

2.2. 业务需求分析 6

2.2.1. 业务流程分析 6

2.2.2. 固定报表需求 7

2.2.3. 多维分析需求 8

2.3. 业务数据分析 8

2.3.1. XX业务系统及关键表 8

2.4. 业务分析小结 9

3. 数据仓库设计 10

3.1. 主题域划分 10

3.2. 梳理指标体系 10

3.3. 确定维度 11

3.4. 数据仓库分层设计 12

3.4.1. ODS(操作数据层)设计 12

3.4.2. DWD（明细数据层）设计 14

3.4.3. DWB（基础数据层）设计 14

3.4.4. DWS（数据服务层）设计 14

4. 调度设计 14

4.1. 数据更新类型 14

4.2. 数据更新频率 15

4.3. 表处理方式 15

4.4. 任务调度 15

# 编写指南

数据仓库整体设计方案应包含系统设计阶段可能包括的基本信息，重点在于设计方案和需求分析文档相结合，描述系统整体情况。

报告编制要求：

这里列出本系统设计报告编制的经验性要求，须由系统设计人员参照其进行裁剪以确定本次报告编制的相关规定。

|  |  |
| --- | --- |
| 序号 | 要求 |
| 1 | 对各种结构图要求采用Visio绘制。 |
| 2 | 软件体系结构图通常包括：子系统实体、子系统间数据流、相关说明。 |
| 3 | 软件结构图通常包括：模块实体、模块间数据流、数据存储、相关说明。 |
| 4 | 运行体系图包括：进程实体、启动顺序、并发性、数据交换。 |
| 5 | 物理结构图包括：物理节点实体、物理节点分布、物理节点描述、物理节点间相互关系。 |
| 6 | 对数据库结构的描述要求采用结构图方式描述，而非采用过多表格。 |
| 7 | 数据库结构图主要包括：表实体、表间关系、主要字段、相关说明四部分。 |
| 8 | 对于平台性产品，要求有系统层次的划分图。 |
| 9 | 通常要隔离应用对数据库的直接访问。 |

# 业务需求分析

以各业务部门的需求访谈为基准，梳理所需实现的内容

## 业务访谈

在X年X月X日，对财务部门进行业务访谈，内容如下：

### 统计需求

1. 每月分别统计公司发展/关闭客户数量，包括本月新增和月末总数。
2. 每月统计各家客户交易笔数、交易金额、比上月增加的百分比、比去年同期增加的百分比、自年初累计金额（按客户类型、地区维度、商品类目分别统计）。
3. 每月统计各类客户分收益情况。
4. 每月统计平台总成交额、总费用、总收益（按业务线分别统计）。
5. 。。。
6. 。。。
7. 。。。

### 分析需求

1)．退款退货分析：

对退款退货行为进行分析

1. 单客户退款10000元以上
2. 单客户月退货3单以上
3. 3天内退款情况分析

要求显示退款/退货原因、处理情况等内容。

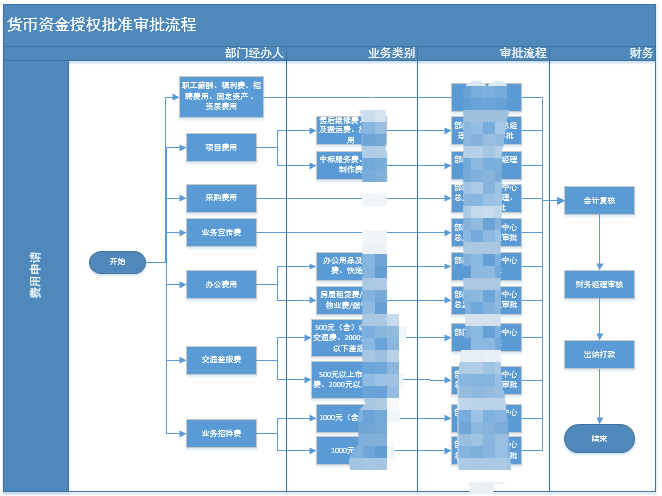
2)．交易分析：

1. 月超大额交易（按时间/笔数/金额排序）
2. 月大额提现（分客户、地区、天、时段等维度）
3. 。。。
4. 。。。

## 业务需求分析

以业务方提出的具体数据需求为基础，梳理业务流程，辨别干系人，了解业务规则。然后根据业务需求，设计对应的实现方式和内容。

### 业务流程分析



对业务流程进行详细的梳理和阐述，每个流程都要明确干系人、业务流转方向、流转规则等内容。

### 固定报表需求

表1． 平台交易额统计表（月报）

XXXX年XX月

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 客户类型 | 平台交易额 | | 自营交易额 | |
| 笔数 | 金额 | 笔数 | 金额 |
| 钻石客户 |  |  |  |  |
| 金牌客户 |  |  |  |  |
| 银牌客户 |  |  |  |  |
| 普通客户 |  |  |  |  |

注：每一列需增加同比、环比数据

数据不含未付款订单，含已退款退货订单。

筛选条件：年月。

每个报表需求都需要保护以下内容：

* 报表名称
* 筛选条件
* 表头
* 业务逻辑说明

### 多维分析需求

表2． 月度大额提现多维分析表

XXXX年XX月

分析指标：提现金额、提现次数、提现客户数

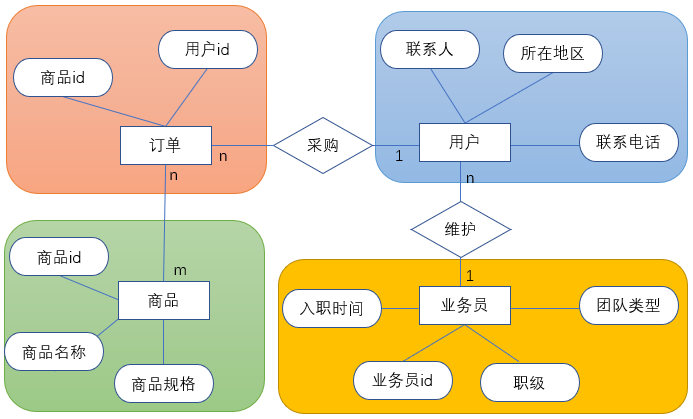
分析维度：分客户、地区、天、时段等维度

筛选条件：年月、地区等

## 业务数据分析

根据业务需求，反向找到相应的业务系统，提取相应数据表，分析表间关系。必须要有ER图和表结构，便于之后的建模。如果内容太多，可以引用

### XX业务系统及关键表



XX系统关键ER关系图

|  |  |  |  |
| --- | --- | --- | --- |
| XX系统用户表 User\_info | | | |
| 字段名 | 注释 | 类型 | PK |
| UserID | 用户id | Varchar2（10） | PK |
| UserName | 用户名称 | Varchar2（50） |  |

XX系统关键表结构

## 业务分析小结



根据业务流、关注重点和信息系统等梳理出一张业务全景图，作为划分业务域的依据。

# 数据仓库设计

## 主题域划分

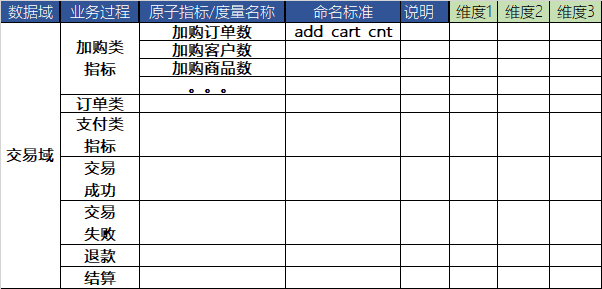
以业务目标为核心，划分主题域：

* 用户域
* 销售域
* 活动域
* 商品域
* 仓储域
* 配送域
* 结算域

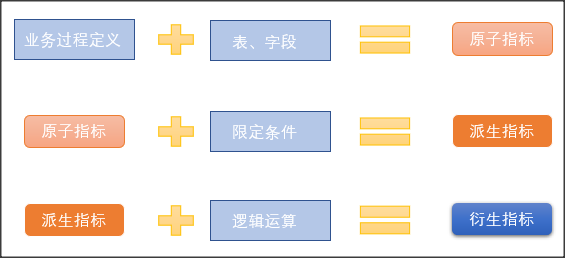
主题域的划分规则有：按业务流划分、按需求划分、按部门职责划分和按产品功能划分。一般都是按照业务流划分。

## 梳理指标体系

在确定了主题以后，我们将梳理每个主题域的指标体系，诸如年销售额之类。



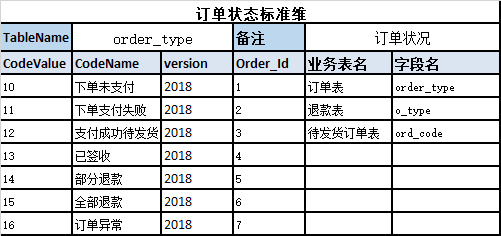
对每个主题域的业务过程，都梳理出可以量化的指标。



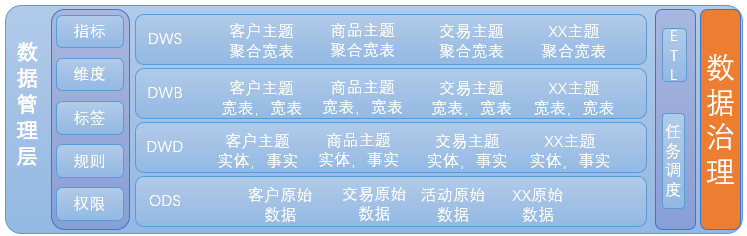
指标分为原子指标、派生指标、衍生指标。三者区分如上图所示。

## 确定维度

设计各个维度的主键、层次、层级，尽量减少冗余。将所需的所有维度全部梳理清楚，有部分维度是跨系统的，需要进行ID的Mapping。



## 数据仓库分层设计



数据仓库分层最少3层：ODS\DW\DM。DW和DM又可以各自拆分，这就有了4、5、6、7层的数仓。分层是数据仓库的总体架构设计核心内容。上图示例为4层架构。

### ODS(操作数据层)设计

需要对每一层的设计进行详细的阐述，每层的目标、任务、处理逻辑都要详细说明。

#### 功能描述

实现从外部系统导入数据到大数据平台,ODS层数据与数据源保持一致.

* 同步方式包括:
* 增量同步
* 全量同步
* 数据源类型包括:
* 数据库

#### 业务流程

* 流程图:



* 流程说明:

1. 调度系统定时启动数据同步作业.

2. 作业读取数据同步脚本,远程调用数据库接口.

3. 数据库接口返回远程调用结果.

4. 接收数据库接口返回的数据集,装载到ODS本地表.

* 异常处理:

调度系统捕获异常并处理

* 日志说明:

由调度系统负责记录作业(正常和异常)日志.

* 数据源说明:

1. 通过视图封装需要同步的物理表, 与业务进行隔离. 当物理表发生变动,尽量通过修改视图兼容.

2. 数据同步程序访问数据源视图.

* 数据粒度说明：

与数据源保持完全一致。

#### 作业设计

略

### DWD（明细数据层）设计

#### 功能描述

#### 业务流程

#### 作业设计

### DWB（基础数据层）设计

#### 功能描述

#### 业务流程

#### 作业设计

### DWS（数据服务层）设计

#### 功能描述

#### 业务流程

#### 作业设计

## 调度设计

## 数据更新类型

全量更新：每次获取数据时，都将原始表读取到目标表中；

增量更新：每次获取数据时，只获取上次同步之后的新数据；

## 数据更新频率

按日更新：每天更新一次，又分为月度数据日更、年度数据日更、每日数据日更。

按月更新：每月更新一次。

准实时更新：每N分钟更新一次。

实时更新：实时计算。

## 表处理方式

拉链表：每次获取数据时，获取所有变化的数据，同时保留所有历史数据；

代理键：缓慢变化维时使用代理键，避免维度代码重复。

CUBE：每个维度组合都计算一遍各指标数据。

## 任务调度

同时开始。

顺序进行。

前后依赖。

参数调用。

任务判断、流向支流。

报警邮件。