**青岛清算**

**ETL设计开发规范**

（V 1.0版）

| 郑重申明 |
| --- |
| * 本文档及其附件含有 铂涛集团 专有信息，未经许可，不得将本文档及其附件中涉及的资料泄露除 有关人员以外的其它人员。 |

文档属性

| 属性 | 内容 |
| --- | --- |
| 客户名称 | 青岛清算 |
| 项目名称 | 风控项目ETL开发规范 |
| 项目编号 |  |
| 文件主题 |  |
| 文档副标题 |  |
| 文档编号 |  |
| 文档版本 | V1.0 |
| 版本日期 | 2017年11月7日星期二 |
| 文档状态 | 初稿 |
| 内部参照 |  |
| 撰稿人 | 陈力杰 |
| 审批人 |  |
| 相关人员 |  |

文档变更

| 版本 | 修订日期 | 修订人 | 描述 |
| --- | --- | --- | --- |
| V1.0 | 2017/11/7 | 陈力杰 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

文档送呈审批人

| 审批人姓名 | 审批人签名 | 目的 | 描述 |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

目 录

[1 ETL平台架构图 3](#_Toc424135584)

[2 资料库规划 3](#_Toc424135585)

[3 命名规划 4](#_Toc424135586)

[3.1 存储表空间命名 4](#_Toc424135587)

[3.2 Table 命名 5](#_Toc424135588)

[3.3 维度规范 5](#_Toc424135589)

[3.4 索引创建规范 5](#_Toc424135590)

[3.5 补录数据命名规范 5](#_Toc424135591)

[3.6 EDW表规定添加字段 6](#_Toc424135592)

[3.7 Transformation、jobs命名 6](#_Toc424135593)

[3.8 控件命名 6](#_Toc424135594)

[4 开发规范 7](#_Toc424135595)

[4.1 Transformation开发 7](#_Toc424135596)

[4.2 connection 8](#_Toc424135597)

# 命名规划

## 存储库设计命名

| 库 | 用户 | 描述 |
| --- | --- | --- |
| STG | STG | 存储STG层所有的表 |
| LAKE | LAKE | 存储LAKE层所有的表 |
| HUB | HUB | 存储LAKE,DIM层所有的表 |
| INSIGHT | INSIGHT | 存储HUB层所有的表 |

## Table 命名

| 模块 | 模块名称 | 目标表 | 例如 |
| --- | --- | --- | --- |
| STG | STG||系统名称 | 模块名称\_表名称 | STG||系统名称\_表名称 |
| 例如: | | |
| STGOMS | STGOMS\_表名称 | STGOMS\_ORDER |
| STGDCMS | STGDCMS\_表名称 | STGDCMS\_PRODUCT |
| LAKE | STG||系统名称 | 模块名称\_表名称 | STG||系统名称\_表名称 |
| 例如: | | |
| ODSOMS | ODSOMS\_表名称 | ODSOMS\_ORDER |
| ODSDCMS | ODSDCMS\_表名称 | ODSDCMS\_PRODUCT |
| HUB | DWDIM | 表名称\_DIM | CUSTOMER\_DIM |
| DW||主题 | 模块名称\_表名称 | DW||主题\_表名称 |
| 例如: | | |
| DWCF | DWCF\_表名称 | DWCF\_CASHFLOW\_ANALYZE |
| DWSF | DWSF\_表名称 | DWSF\_PROFIT\_ANALYZE |
| INSIGHT | DM||主题 | 模块名称\_表名称 | DM||主题\_表名称 |
| 例如: | | |
| DMCF | DMCF\_表名称 | DMCF\_CASHFLOW\_ANALYZE |
| DMSF | DMSF\_表名称 | DMSF\_PROFIT\_ANALYZE |

## 维度规范

* + 维表默认都有两条记录,维度ID分别为:
* 0: 不限 (表示该行记录的维度ID可以为空)
* -1: 找不到（表示该行记录的维度代码在维表里面找不到维度成员）

## 索引创建规范

* + 唯一索引的命名方式： 表名\_U1 (1…N)
  + 普通索引的命名方式: 表名\_N1 (1…N)
  + 在原则上数据仓库模型都应该有唯一键

## 补录数据命名规范

* + 补录数据区分面向公共和面向主题两部分
* APD\_表名
* APD\_主题\_表名
  + 补录数据直接从补录系统抽取到ODS层，所以在ODS层存储命名如下:
* ODSAPD\_表名
* ODSAPD\_主题\_表名
  + 补录系统的表以LAST\_MODIFIED\_DATE作为增量时间戳

所以补录的表需要以下几个字段:Created\_Date,Created\_BY,Last\_Modified\_Date,Last\_modified\_By

* + 所有的补录表都放在ODS层，而不放入DW,DM层
  + 如果补录表是关系对照表，则需要在结尾添加rela表示(relationship)
  + 其它类型的补录表则需要在结尾添加def.

## EDW表规定添加字段

* + Created\_date --创建日期
  + Load\_batch --装载批次

## Transformation、jobs命名

* + 名字长度不要超过30个字符

| 模块 | 名称 | | 例如 | |
| --- | --- | --- | --- | --- |
| Transformation | | | | |
| STG | tf \_ Folder Name \_ Business Name | | tf\_STGOMS\_GET\_ORDERS | |
| ODS | tf \_ Folder Name \_ Business Name | | tf\_ODSOMS\_GET\_PRODUCT | |
| DW | tf \_ Folder Name \_ Business Name | | tf\_DWAR\_GET\_ORDERS | |
| DM | tf \_ Folder Name \_ Business Name | | tf\_DMAR\_GET\_PRODUCT | |
| APP | tf \_ Folder Name \_ Business Name | | tf\_APPEPM\_GET\_PRODUCT | |
| Jobs | | | | |
| STG | | jb\_JOBS name | | jb\_STGOMS\_ORDERS |
| ODS | | jb\_mapping name | | jb\_ODSOMS\_ORDERS |
| DW | | jb\_mapping name | | jb\_DWAR\_NAME1 |
| DM | | jb\_mapping name | | jb\_DMAR\_NAME1 |
| APP | | jb\_mapping name | | jb\_APPEPM\_NAME1 |

## 控件命名

| 控件 | 控件名称 | 规范 | 例如 |
| --- | --- | --- | --- |
| Table Input | 源系统 | SRC\_SYSTEM \_Schema | SRC\_OMS |
| UNION | 合并 | UN\_SHORT NAME | UN\_ORDERS |
| Call DB Procedure | 存储过程 | SP\_PROCEDURE NAME | SP\_GET\_PRODUCT |
| Sequence Generator | 序列 | SEQ\_NAME | SEQ\_CUSTOMER\_DIM |
| Group by | 汇总 | AGG\_SHORT NAME | AGG\_PRODUCT |
| Filter rows | 过滤 | FLT\_SHORT NAME | FLT\_GREATER\_100 |
| Database lookup | 查找 | LKP\_SHORT NAME | LKP\_CUSTOMER\_NAME |
| Merge Join | 关联 | JNR\_JOIN COLUMN | JNR\_CUSTOMER\_CODE |
| Row Normaliser / Row denormaliser | 列转行 | NRM\_SHORT NAME | NRM\_MONTH |
| Switch/case | 分组 | RTR\_ROUTER COLUMN | RTR\_CUSTOMER\_CODE |
| Sort rows | 排序取几位 | RNK\_SHORT NAME | RNK\_CUSTOMER\_10 |

注释：合并的功能控件目前还没有找到

# 开发规范

## Transformation开发

* + 所有的Mapping都添加三个参数(Parameter):
* LOAD\_BATCH VARCHAR2(30) 🡪装载批次(表示模块数据是那一次装载的)
* INCR\_START\_DT VARCHAR2(20) 🡪增量起始时间(表示增量区间的起点)
* INCR\_END\_DT VARCHAR2(20) 🡪增量截至时间(表示增量区间的终点)
  + Mapping增量要求:
* Source 🡪 STG : Truncate & Insert
* Source 🡪 ODS: Truncate & Insert
* STG 🡪 ODS : Update&Insert/Delete &Insert/ Truncate & Insert
* ODS 🡪 DW : Delete & Insert
* DW 🡪 DM: Delete&Insert
* 对于所有的 转换要求尽量使用增量（根据增量区间）
* 对于数据源表确实无法提供增量时间戳的数据源，则进行全量抽取（这种方式只适用于元数据表,对于大数据量的表则不给与抽取，直到该表提供增量时间戳）
* 对于STG🡪ODS，该表又没有增量时间戳的，则使用Update & Insert,此时需要判断本次全量抽取的数据，对于ODS表原有的数据那些记录发生了变化，对于发生变化的记录进行更新，同时标记最后更新日期，对于没有发生变化的记录，则不做更新，对于新增的记录则直接插入。
* 对于Source 🡪 ODS: 要求是每次全量装载的表
  + Reload方案:
* 经常会进行数据的重复装载，这就要求在转换的设计，开始时要考虑数据的重载机制，对于目前需要考虑Reload的有ODS🡪DW,DW🡪DM,重载需要在 Job里面添加一个sql控件 并内置添加以下脚本
* Delete from (Target Table)

Where (increment column)>=to\_date($$incr\_start\_dt,’YYYY-MM-DD HH24MI’)

And (increment column)<to\_date($$incr\_end\_dt,’YYYY-MM-DD HH24MI’)

And (source\_code = $$SOURCE\_CODE);

Commit;

* + Source Qualifier 控件使用:
* 如果转换中存储存在两个及以上的Source表，同时他们来自相同的源系统，尽量使用一个源表 并添加关联条件，而不是用Joiner控件
* 在where中添加增量条件:

时间戳字段>= TO\_DATE($$INCR\_START\_DT,’YYYY-MM-DD HH24MI’) And

时间戳字段 < TO\_DATE($$INCR\_END\_DT,’YYYY-MM-DD HH24MI’) (注意这里是小于，而不是小于等于)

(source\_code= $$SOURCE\_CODE )

## connection

* + 管理规则

Connection的维护需要由系统管理员进行统一管理，每个项目组如果需要可以进行申请，

对于已经定义好了连接，可以共享

* + 命名规则
* C\_ System Name \_ Schema Name

例如: C\_EDW\_DWAR，C\_OMS\_EAI

## Job规范

* + 一个job尽量只有一个转换，因为本身Job没有状态，调度平台需要判断转换的返回状态，如果多个转换，只能判断其中一个。