大数据开发规范文档

> Author: koray
>
> Version: 1.0 (持续更新完善)
>
> Date: 2020-11-20

1. 数仓模型层次定义

数仓层级	命名	说明
数据源层	ODS	只映射COS系统上的数据，不做任何修改，起到备份数据的作用。
数据明细层	DWD	对ODS层的数据进行清洗和转换，并且提供一定的数据质量保证。
数据中间层	DWM	对DWD层的数据做轻度的聚合操作，生成一系列的中间表，提升公共指标的复用性，减少重复加工。
数据服务层	DWS	按照业务主题划分,生成字段比较多的宽表，用于提供后续的业务查询。
数据应用层	APP	以分析的主题对象为建模驱动，基于上层的应用和产品的指标需求，构建主题对象的全量宽表，供后续的业务使用。

2. 命名规范

1.数据库命名规范

- 规则： 项目名 + 公共功能
- 示例： **lazada_dw** （dw：data warehouse 数仓）

2.表命名规范

数仓层级	表命名规范	示例
ODS	ods_cos/mysql_主题(模块)_业务描述	ods_cos_item
DWD	dwd_表类型_主题(模块)_业务描述	dwd_snap_item_info (商品历史快照表)
DWM	dwm_主题(模块)_业务描述	dwm_item_day (商品的天维度表)
DWS	dws_主题(模块)_业务描述	dws_item_itemNum (商品的产品数维度宽表)
APP	app_rpt_主题(模块)_业务描述	app_rpt_item (商品统计展示表)
中间临时表	tmp_功能_主题(模块)_业务描述	tmp_etl_ods_item (商品数据清洗中间表)

> **注:**
>
> - 表名使用英文小写字母，单词之间用下划线分开，长度不超过30个字符，命名一般控制在小于等于6级。
> - 时间粒度：使用"c"代表当前数据，"h"代表小时数据，"d"代表天数据，"w"代表周数据，"m"代表月数据，"q"代表季度数据， "y"代表年数据。
> - 对象属性，用"t"表示表，用"v"表示视图。o

3. 数据集标准规范

1.数据类型规范

数据库	类型	源数据库数据类型	HIVE数据类型
MYSQL	字符	CHAR	STRING
MYSQL	字符	VARCHAR	STRING
MYSQL	字符	TEXT	STRING
MYSQL	字符	TINYTEXT	STRING
MYSQL	字符	BINARY	STRING
MYSQL	字符	VARBINARY	STRING
MYSQL	字符	MEDIUMTEXT	STRING
MYSQL	字符	LONGTEXT	STRING
MYSQL	数值	TINYINT	DECIMAL(15,2)
MYSQL	数值	SMALLINT	DECIMAL(15,2)
MYSQL	数值	MEDIUMINT	DECIMAL(15,2)
MYSQL	数值	INT	BIGINT
MYSQL	数值	BIGINT	BIGINT
MYSQL	数值	FLOAT	DECIMAL(15,2)
MYSQL	数值	DOUBLT	DECIMAL(15,2)
MYSQL	数值	DECIMAL	DECIMAL(15,2)
MYSQL	日期	DATE	STRING
MYSQL	日期	TIME	STRING
MYSQL	日期	DATETIME	STRING
MYSQL	日期	TIMESTAMP	STRING
MYSQL	日期	YEAR	STRING

2.数据字段规范

- **命名**

小写字母、数字、下划线组成，不同单词之间用下划线分开

- **partition分区列**

Hive partition列在Data中并不存储，这会导致当以文件形式对外提供数据时，数据会有缺失，所以我们要对所有的partition列进行冗余存储。

> 如：当以分区列为dt字段时，我们多添加一列hp_dt列即可。（hp：hive partition）

- **列操作**

可以修改列数据类型；新增列只能加到最后；严禁删除列。

- **注释**

注释本着简洁、详实、完整的原则，对于有业务含义的字段，在注释中需要枚举并解释其业务含义.

> 如: ”order_status” 订单状态字段：1待支付，2支付不成功，3支付成功.

- **类型**

日期时间等格式统一用string类型，字符串也是用string，数值的话会根据字段定义来确定，对于有小数点要求的，比如某些金额、利率，需要用到decimal类型，无小数点要求的用浮点类型double和整数类型(int / bigint)

- **时间字段格式**

爬虫数据传入的时间统一采用 **yyyy-MM-dd HH:ss:mm** 格式.

hive表的时间分区字段统一采用 **yyyy-MM-dd** 格式.