1.1 数据仓库概述

由 王成明 创建, 最后修改于2018-05-03

本页内容：

相关页面：

- 代码清单

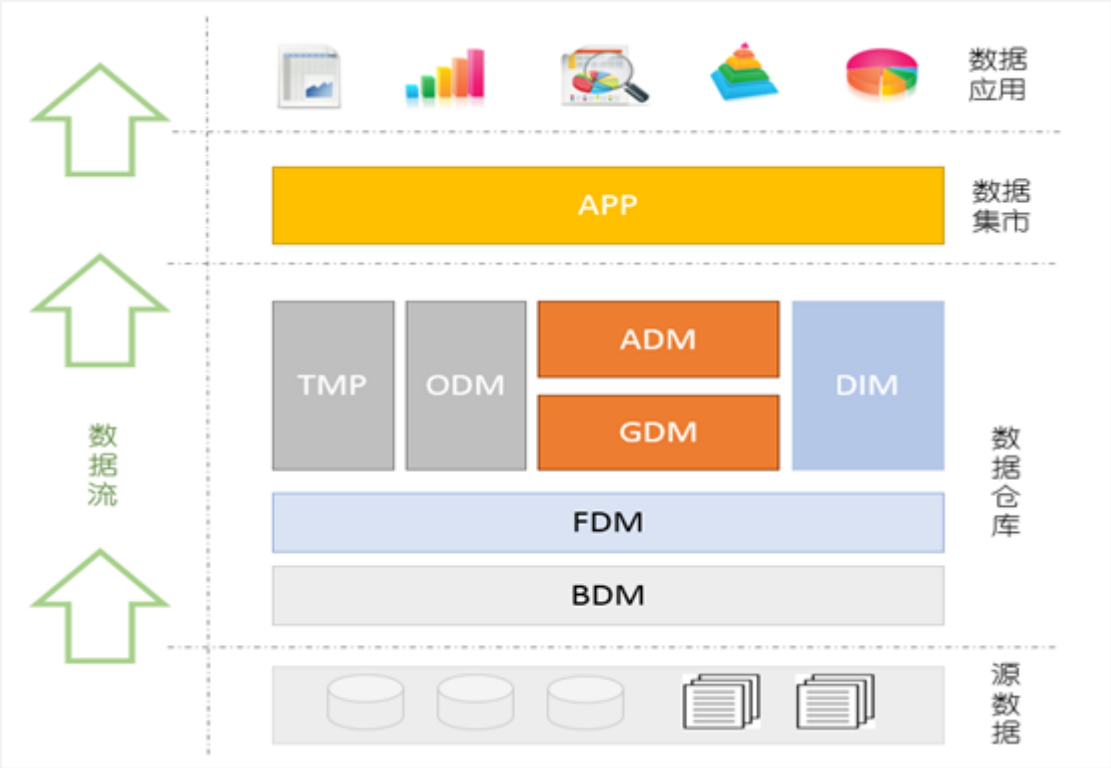
一、概述

《数据仓库参考手册》是保证数据仓库的 标准性、稳定性、一致性， 用于指导数据仓库模型管理，任务管理，命名规范，为维护、管理一个企业级数据仓库做好坚实的基础。

二、数据仓库概况

1、数据架构

1.1 京东数据仓库(JDW)上接数据源，下接数据集市，整体的架构示意图如下：



图表 -1京东数据仓库整体数据架构

源数据是指源业务系统的产生的业务数据，一般情况下存储在数据库中和文本文件中。数据经抽取加载进入数据仓库，建模人员在数据仓库中进行转换和清洗，然后同步到数据集市。数据集市根据集中使用数据的部门和特定的需求建立，如营销集市、运营集市、销量预测集市，各集市间相互独立。集市用户进行数据应用设计，以支持各种数据应用，如报表展现、营销数据支持等。

1.2 京东数据仓库数据架构各层的用途描述如下：

表格-1数据流向描述：

序号	数据层次	简称	数据层次用途简述
1	缓冲数据模型	BDM	源业务系统数据的快照，保存细节数据，按天分区，会保持最近一段时间数据。一般情况下，每个BDM表对应着源业务系统的一个表或者一个日志文件，数据结构与线上基本是对应的。绝大多数的数据快照是经过增量抽取策略抽过来了，对于不支持增量抽取策略或者数据量极少的表采用全量抽取的策略。
2	基础数据模型	FDM	基础数据模型，用来保存源业务系统数据的快照，数据永久保存。对于有更新操作的数据来说，采用拉链的方式优化存储。对于没有更新操作的数据来说，采用流水方式存储。
3	通用数据模型	GDM	根据京东核心业务主题按照星型模型或雪花模型设计方式建设的最细业务粒度汇总层。在本层需要进行指标与维度的标准化，保证指标数据的唯一性。
4	聚合数据模型	ADM	根据不同的业务需求采用星型或雪花型模型设计方法构建的按维度汇总数据。
5	维度模型	DIM	维度表可以看作是用户来分析数据的窗口，维度表中包含事实数据表中事实记录的特性，有些特性提供描述性信息，有些特性指定如何汇总事实数据表数据，以便为分析者提供有用的信息，维度表包含帮助汇总数据的特性的层次结构。例如，包含订单信息的维度表通常包含将订单分为区域、省份、城市等若干类的层次结构。在维度表中，每个表都包含独立于其他维度表的事实特性，例如，客户维度表包含有关客户级别的数据。维度表中的列字段可以将信息分为不同层次的结构级。
6	临时层	TMP	用来降低加工过程计算难度，提高运行效率的临时表，用完即舍，不保存历史数据。
7	中间层(操作数据模型)	ODM	在加工通用模型的时候，对于多个模型都使用到的公共数据需要清洗转换的时候，用来封装清洗转换逻辑保存清洗后的数据，供加工通用模型使用，中间层数据保存历史状态。
8	应用数据模型	APP	应用数据模型按照具体的需求进行设计，其数据直接供前端报表工具展现使用，或者推送到其他系统做相关的数据支撑。

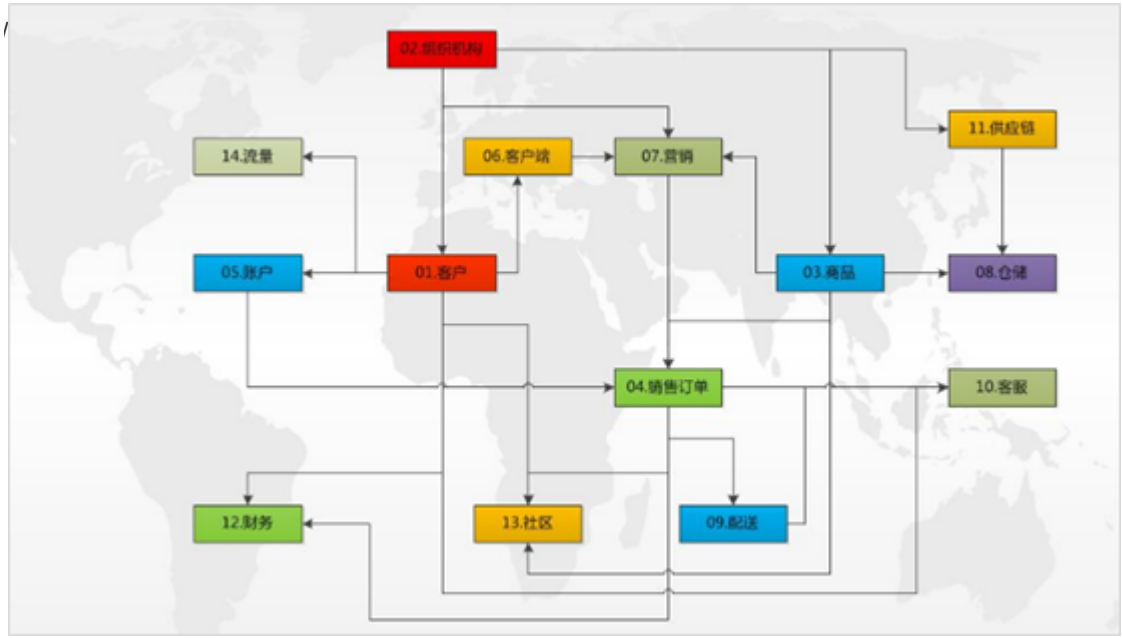
2、通用模型主题

主题列表如下：

表格 -2通用模型主题：

主题划分	主题简称	主题英文名	覆盖内容(包含但不仅限于)
01.客户	CUS	Customer	供应商、POP商家、团购商家、合作伙伴、用户
02.组织机构	ORG	Organization	员工、部门
03.商品	ITM	ITEM	商品相关
04.销售订单	ORD	Order	订单相关
05.账户	ACT	Account	与账户关联的礼品卡、余额、积分、优惠券等
06.客户端	CLI	Client	移动客户端、PC客户端、移动设备
07.营销	CAM	Campaign	促销、活动、优惠券、礼品卡
08.仓储	INV	Inventory	仓储相关
09.配送	DIS	Distribution	配送相关
10.客服	CSC	Customer Service Center	售后、备件库、呼叫中心、工单
11.供应链	SCM	Supply Chain Management	采购、采购退货(退供应商)
12.财务	FIN	Finance	财务单、结算、发票、应收应付
13.社区	COM	Community	用户的关注、订阅、俱乐部、社区
14.流量	TRA	Traffic	流量相关

主题示意图：



三、命名规范

1、京东数据仓库层次命名

命名规范包括对数据仓库数据架构中的各层次的命名规划。

在命名过程中需要建立一个统一的中英文词库，在翻译过程中，首先到词库中查找相应的单词缩写或者词根，然后根据表名和字段名的命名规则进行命名。表名和字段名在用词、词缩写方式上都应保持一致，避免出现同一词有多种缩写的情况。

参考物理模型，附件：JD_JDW_C2E_ROOT.csv

1.1、数据缓冲数据层

数据缓冲层规则：表名= BDM_源库名称_源表名。

注，BDM表名一般是直通车自动生成。

源数据为分库分表则需要负责人为源表名部位指定一个逻辑表名，库名会默认取分库排序的第一个库名。

如果库名表名存在相同的表，后接入的可以指定标识。

1.2、基础数据层

基础数据层规则：表名= FDM + 源库名称 +源表名+ 加载策略。

加载策略：拉链表 _chain结尾；流水表无；

注：FDM表名是由数据接入工具直通车自动生成，分库分表与重复接入同BDM。

1.3、通用数据层

通用数据层规则：表名= GDM + 主题前缀 + 主体 + 策略

命名格式举例：GDM_M03_ITEM_SKU_DA

按照 表格- 2通用模型主题 中对主题的分类、主题编号，我们对每一主题中的表在物理命名时会加上一个前缀，以方便区分来自哪一个主题。主体是对业务分析范围或者源表名称的简写。策略是表示增量还是全量，_DA 表示每日全量。增量默认没有策略后缀.

ODM命名规范与GDM一致。

1.4、聚合数据层

聚合数据层规则：表名= ADM+ 主题前缀 + 主体+后缀(日/周/月/季/年/其他)

ü 主题前缀：编号与通用模型主题一致，字母为S开头，如：流量S14

ü 主体：到词库中查找相应的单词缩写

ü 后缀：D 日汇总 W周汇总 M 月汇总 Q季汇总 Y年汇总

命名格式举例： ADM_S04_DNST_ORD_XXX_D

3常用主体缩写

序号	聚合层主题	主体英文简写
1	日百	DNST
2	IT数码	IT
3	通讯汽车	CC
4	家电	HEA
5	图书	BOOK
6	采销	SCM
7	仓储	STORE
8	配送	DIS
9	POP	POP
10	售后	AFS

11	市场	MKT
12	信息部	INFO
13	发展战略	DS
14	财务	FIN
15	人资	HR
16	海外	EPT
17	自有品牌	PB

1.5、临时层

临时层规则：表名= TMP + 层次名 + [业务含义的表名]

1.6、维度层

DIM层规则：

维表表名= DIM + 业务含义的表名

注：DIM表一般设计为非分区的单表

代码表,用来保存代码值和描述的代码表，其表名采用DIM_D99_开头，加主体，后加该代码字段的字段名。

命名格式举例：dim_d99_ord_pay_type_cd

表结构如下：

1.7、其他层次

在京东数据仓库中除上述分层之外，还存在APP层和集市层的特殊需求，这两层数据是在我们设计的数据仓库层次基础上，根据不同部门的需求，从数据仓库各层次中抽取需要的数据信息按照APP层和集市层的存储与命名规范生成的数据层次。

集市层的命名规范与数据仓库各层的命名规范相对应。APP层命名规范如下：

表名= APP + 需求主体英文简称+ 业务主体 + 后缀(日/周/月/季/年)

2、字段命名

在数据仓库模型中，具有相同业务含义的字段在模型的不同业务范围、不同主题范围、不同的事实表中、不同的维度表中必须使用同一个中文和英文名，不允许出现同一业务含义的属性字段在数据仓库使用不同的中英文名称；命名翻译应该参考物理模型字典。

构成字段的英文名称中每个单词字母采用小写，单词间用下划线分割；

对于编号类字段，一般源系统有业务含义的编号用“××号”命名，无业务含义的编号用“××编号”命名，由数据仓库中生成的编号字段也用“××编号”命名。当中文用“号”时，英文名后缀一般用num或no，中文用“编号”时，英文名后缀一般用id。对于订单主题、组织机构主题和财务主题的标识符，我们习惯上用num或no，如：凭证号vou_num、卡号card_no、机构号org_num、会计科目号gl_num等；其他主题的标识符则习惯上用id，例如：客户编号cus_id、商品编号 item_id、地址编号addr_id、营销计划编号plan_id、营销活动编号activity_id、渠道编号chan_id、申请号apply_id、 订单编号ord_id 等；

字段属性的中文名称尽量保留实体的名称作为前缀，例如：“客户编号”、“渠道编号”；

代码表的列属性中文名与实体的中文名保持一致，代码列属性英文名为实体名去掉前缀，每单词字母小写，相应的描述字段属性后缀是desc，如“婚姻状况描述marg_status_desc”；

有一些按习惯称为“xx名称”，描述属性中文名可以用“xx名称”，英文名用“xx_name”，比如“货币代码”实体，其两个字段分别为“货币代码curr_cd”和“货币中文名称curr_cn_name”；

日期类型的字段，后缀应是dt，如“开户日期open_dt”等；时间类型后缀应是tm，如“出库时间out_wh_tm”等；

时间拉链用“开始日期start_date”和“结束日期end_date”，此处的开始日期和结束日期不体现业务含义，只体现数据有效性的周期(数据快照)；

中文名称中尽量不加“和”、“的”等字眼以及标点符号、空格、斜线、减号和其他非规范文字；

英文名中不能出现标点符号、空格、斜线、减号等特殊字符，否则DDL语句在执行时会报错；

维表及字段命名，统一使用“dim_”+名称(表名、字段名)的方式，在JDW模型各层维表命名规则原则上保持一致，衍生出的分支维表也应遵守维表命名原则

3、数据类型定义

数据仓库系统会以多个源系统作为其数据来源，一般情况下源系统的数据类型定义会多种多样，差异较大，所以数据仓库在数据类型的定义上，一方面要参考源系统相关字段的定义，还要考虑到仓库自身的特点和要求，综合起来进行设计。对于业务系统的数值型单据编号，一律采用bigint而非int 。

3.1、数据类型定义遵循原则：

数据仓库中的字段长度尽量满足相应源系统字段中最大长度的要求，当然也会考虑字段的业务含义，对于一些源系统定义过长，而从实际业务含义又不可能有那么长的字段，由仓库自行选择一个合适的长度定义；

为了提高通用性，代码字段尽量不用数字型，建议多采用String类型；

数据仓库在处理代码转换时，如果遇到一个例外取值，处理的的方式是在前面拼上一个特殊字符将其保留下来，以便后面知道这个例外取值的含义时可以进行相应恢复处理，这样就要求字段长度定义必须比源系统字段中最大长度还多2位；

日期类型字段由于格式多样，造成在信息加工处理过程中的格式转换复杂且易出错，因此对日期类型字段统一制定如下规范。

ü 以“2013-01-01 13:35:23.71”为例说明日期字段类型规范内容：

日期规范类型	格式规范	描述	举例	命名规范
时间戳字段	yyyy-MM-dd HH:mm:ss	带日期，精度到秒	2013-01-01 13:35:23	xxx_timestamp
时间字段	HH:mm:ss	不带日期，精度到秒	13:35:23	xxx_time
日期字段	YYYY-MM-DD	日期到天	2013-01-01	xxx_date

京东集团数据仓库 / YYYY数据仓库管理				
周（年的第几周）	YYYYWW	所在年份的第几周，由年份、月份和所在周数组成	20130101	xxx_yearweek
周（月的第几周）	YYYYMMW	由年份、月份和所在周组成	2013011	xxx_monthweek
月份字段	YYYYMM	由年份和月份组成，无分隔符	201301	xxx_month
季度字段	YYYYQ	由年份和所在季度组成	20131	xxx_quarterly
年字段	YYYY	年份	2013	xxx_year

ü 命名中的“xxx”代表具体的业务操作说明

ü 在默认情况下，JDW中的日期类型就以上述规范为标准，如果有特殊需求请向我们反馈，我们将完善到上述规则和数据仓库模型设计规范中

ü 需要精度到毫秒级的日期类型字段，此时我们会补充如下规范内容：

日期规范类型	格式规范	描述	举例	命名规范
时间戳字段（到毫秒）	yyyy-MM-dd HH:mm:ss.fff	带日期，精度到3位数的毫秒	2013-01-01 13:35:23.713	xxx_3timestamp

为了尽可能的保持仓库中数据类型的一致性以及规范性，建议仓库中的数据类型定义不宜过杂，使得仓库中的字段类型看起来比较整齐。

3.2、HIVE支持的数据类型

Hive支持两种数据类型，一类是原子数据类型，一类是复杂数据类型。

3.2.1、原子数据类型

包括数值型、布尔型和字符串类型，具体如下表所示：

表格 4-3 Hive原子数据类型：

类型	描述	示例
TINYINT	1个字节（8位）有符号整数	1
SMALLINT	2字节（16位）有符号整数	1
INT	4字节（32位）有符号整数	1
BIGINT	8字节（64位）有符号整数	1

FLOAT	4字节（32位）单精度浮点数	1.0
DOUBLE	8字节（64位）双精度浮点数	1.0
BOOLEAN	true/false	true
STRING	字符串	‘sss’,”sss”

3.2.2、复杂数据类型

包括数组（ARRAY）、映射（MAP）和结构体（STRUCT），具体如下表所示：

表格 4-4 Hive复杂数据类型：

类型	描述	示例
ARRAY	一组有序字段。字段的类型必须相同	Array(1,2)
MAP	一组无序的键/值对。键的类型必须是原子的，值可以是任何类型，同一个映射的键的类型必须相同，值得类型也必须相同	Map(‘a’,1,’b’,2)
STRUCT	一组命名的字段。字段类型可以不同	Struct(‘a’,1,1,0)

4、建表

建表语句示意：

```
use gdm;
CREATE EXTERNAL TABLE `gdm_m03_item_sku_area_limit_da`(
`item_sku_id` string COMMENT '商品sku编号',
`item_id` string COMMENT '商品编号',
`area_type_cd` string COMMENT '区域类型代码',
`area_id` string COMMENT '区域编号',
`limit_type_cd` string COMMENT '限制类型代码',
`data_type` string COMMENT '商品类型：1 自营 2 图书 3 POP 4 其他')
PARTITIONED BY (
```

`dt` string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'

TBLPROPERTIES (

'orc.compress'='SNAPPY')

;

ü 建表时需要加上 ‘表注释’，’字段注释’

5、任务命名

5.1、调度任务

执行及供下游依赖调用的调度任务： exe_gdm_m03_item_sku_da

用来加工目标表逻辑的任务，应该以exe_ 开头，后加表名，表示该任务是用于执行目标表的数据加工任务。

5.2、plumber推数方式

从hive推往orcl： hive2orcl_adm_s03_item_band_stat

从hive推往mysql： hive2mysql_adm_s03_item_band_stat

从hive推往sqlserver： hive2sqlserver_adm_s03_item_band_stat

从hive推往jss： hive2jss_adm_s03_item_band_stat

无标签