

大数据专家认证

京东数据仓库业务篇 V3.0

初级版

宫敬财

Monday, July 25, 2016

www.jd.com

- 京东数据仓库的背景
- 京东数据仓库总体架构
- 京东数据仓库规范
- 京东数据集市

京东数据仓库的背景，没有数据仓库的时候

大数据专家认证

JDW
京东
数据
仓库
业务



京东数据仓库的背景，DW上线之后

大数据专家认证



JDW
京东
数据
仓库
业务

京东数据仓库的背景，JDW 上线之后



获取
数据

随心
所欲

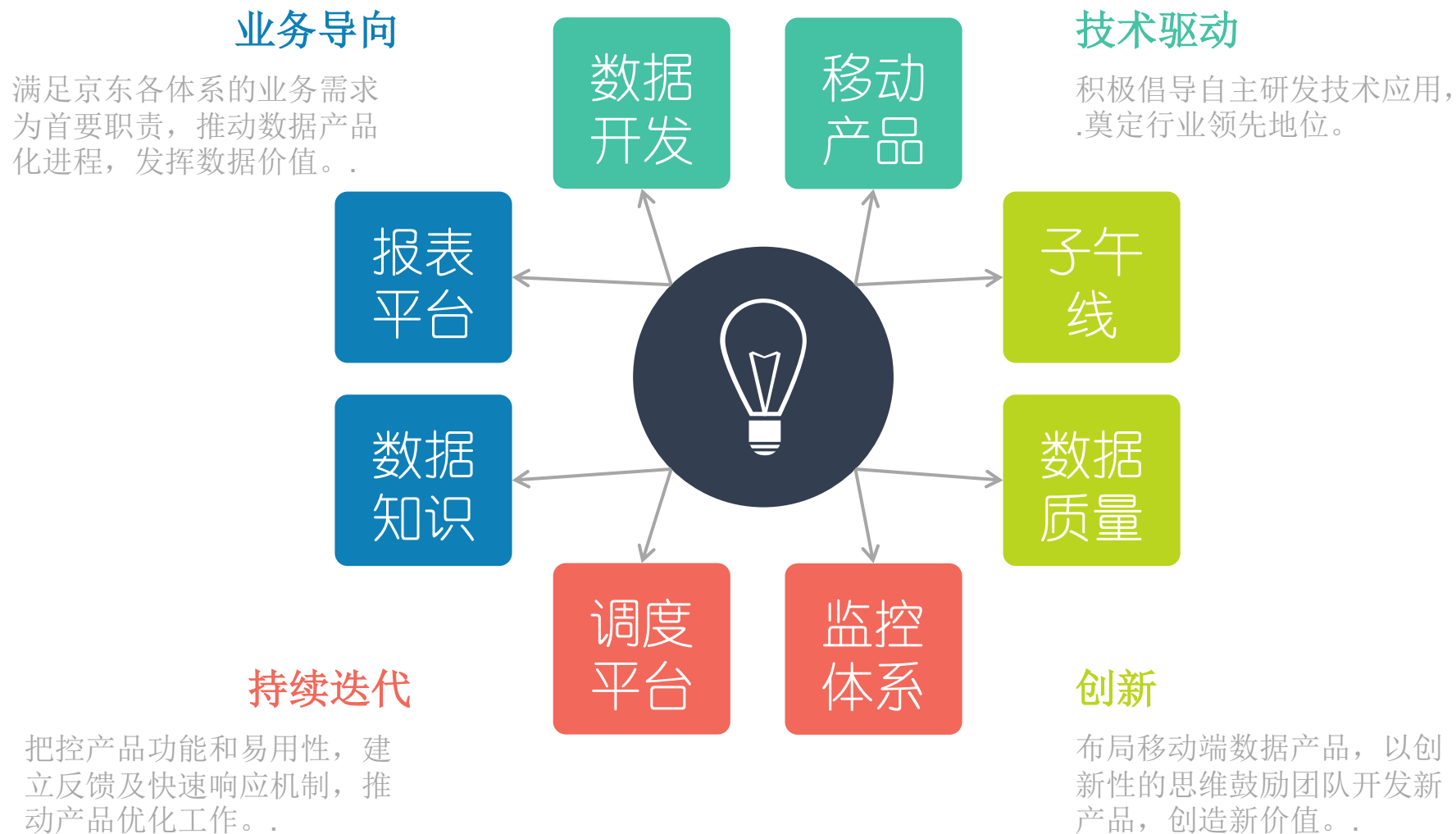
大数据专家认证

JDW
京东
数据
仓库
业务

京东数据仓库架构-概述

大数据专家认证

JDW
京东
数据
仓库
业务



京东数据仓库架构-概述

调度中心


[平台首页](#)
[产品分类](#)
[调度中心](#)



 宫敬财

调度中心

任务管理

脚本管理

变量维护

任务管理

任务类型: 调度任务

任务子类型: 请选择任务子类型

关键字: 支持ID、名称、描述

查询

高级搜索

负责人: 请选择负责人

任务状态:
 ☐ 禁用
 ☒ 空闲
 ☒ 队列中
 ☒ 执行中
 ☒ 失败

新建

拷贝

修改

删除

启用

禁用

重跑

终止

重置状态

管理任务关系

绑定变量

上线申请

下线申请

导出任务及父任务

导出任务及子任务

<input type="checkbox"/>	任务ID	任务名称	任务子类型	运行规则	任务状态	审批状态	执行方式	上次运行时间	结果
<input type="checkbox"/>	223366	107000_dataraw_test	数据计算(py/sh/zi...	0 0 14 * * ?	空闲	测试任务(14天)	正常执行	2016-05-11 14:00:04	2
<input type="checkbox"/>	227189	2016618hall	数据计算(py/sh/zi...	0 0 9 * * ?	执行中	测试任务(24天)	手动重跑	2016-05-11 12:04:52	2
<input type="checkbox"/>	212650	205000_pprelation_xx	普通任务	0 0 10 1/3 * ?	空闲	线上任务	正常执行	2016-05-10 10:00:13	2
<input type="checkbox"/>	163758	44172_app_user_profile_user_kuanbiao	数据计算(py/sh/zi...	00 00 17 23 ...	空闲	测试任务(14天)	正常执行	2016-04-23 22:57:29	2
<input type="checkbox"/>	201071	630017_xx	普通任务	0 0 10 ? * M...	空闲	线上任务	正常执行	2016-05-09 10:00:08	2
<input type="checkbox"/>	201057	635001_xx	普通任务	0 30 8 * * ?	空闲	线上任务	正常执行	2016-05-10 06:39:40	2

上线公告: 测试任务有效期为30日, 超过有效期的任务将自动禁用

[新版说明](#)
[操作手册](#)

JDW
京东
数据
仓库
业务

大数据专家认证

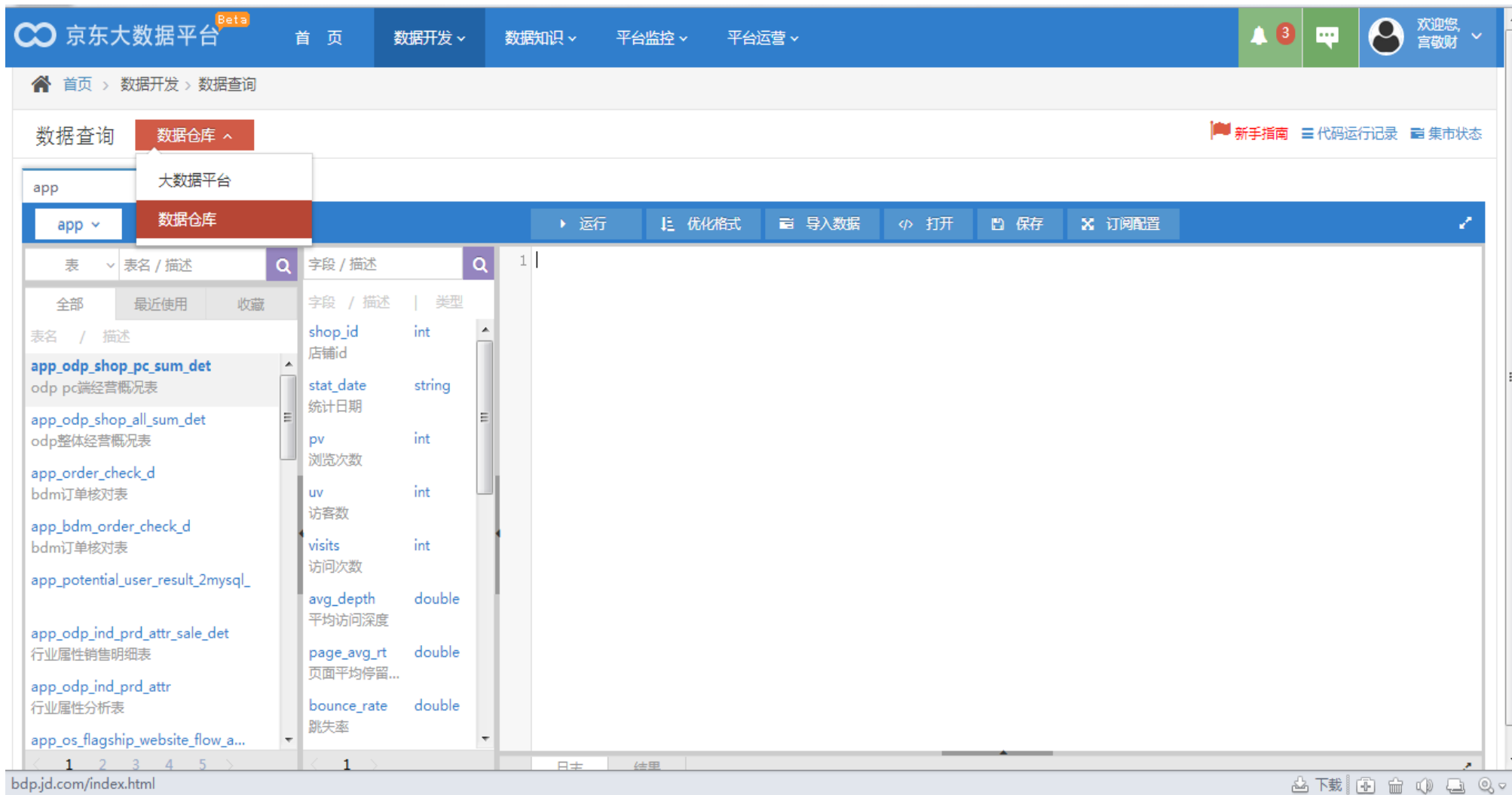
京东数据仓库架构-概述

知识管理平台

The screenshot shows the '数据知识' (Data Knowledge) section of the JD Data Warehouse platform. The interface includes a top navigation bar with links like '首页', '数据直通车', '数据开发', '数据知识', '数据应用', '平台监控', and '平台运营'. A left sidebar lists various management functions such as '元数据库', '集群目录', '业务目录', '维表管理', '运营管理', and '系统维护'. The main content area displays statistics for '今日更新表' (29), '今日新增表' (5314), and '今日下线表' (84). Below these, there are sections for '我收藏的表' (Tables I收藏), '我负责的表' (Tables I负责), and '数据资料库' (Data Library). The '我收藏的表' section lists several tables with their status (e.g., '将下线', '使用中') and actions (e.g., '取消收藏'). The '数据资料库' section lists documents like 'FDM模型的使用说明' and '最近访问的表' (Recently accessed tables).

京东数据仓库架构-概述

数据集成开发平台



The screenshot displays the JD Big Data Platform (京东大数据平台) interface. The top navigation bar includes links for 首页 (Home), 数据开发 (Data Development), 数据知识 (Data Knowledge), 平台监控 (Platform Monitoring), and 平台运营 (Platform Operation). The user is logged in as 欢迎您, 宫敬财.

The main content area is titled 数据查询 (Data Query) and 数据仓库 (Data Warehouse). A dropdown menu for 数据仓库 is open, showing options for 大数据平台 and 数据仓库. The 数据仓库 option is selected, leading to a table view of data warehouse tables.

表名 / 描述	字段 / 描述	类型
shop_id	店铺id	int
stat_date	统计日期	string
pv	浏览次数	int
uv	访客数	int
visits	访问次数	int
avg_depth	平均访问深度	double
page_avg_rt	页面平均停留...	double
bounce_rate	跳失率	double

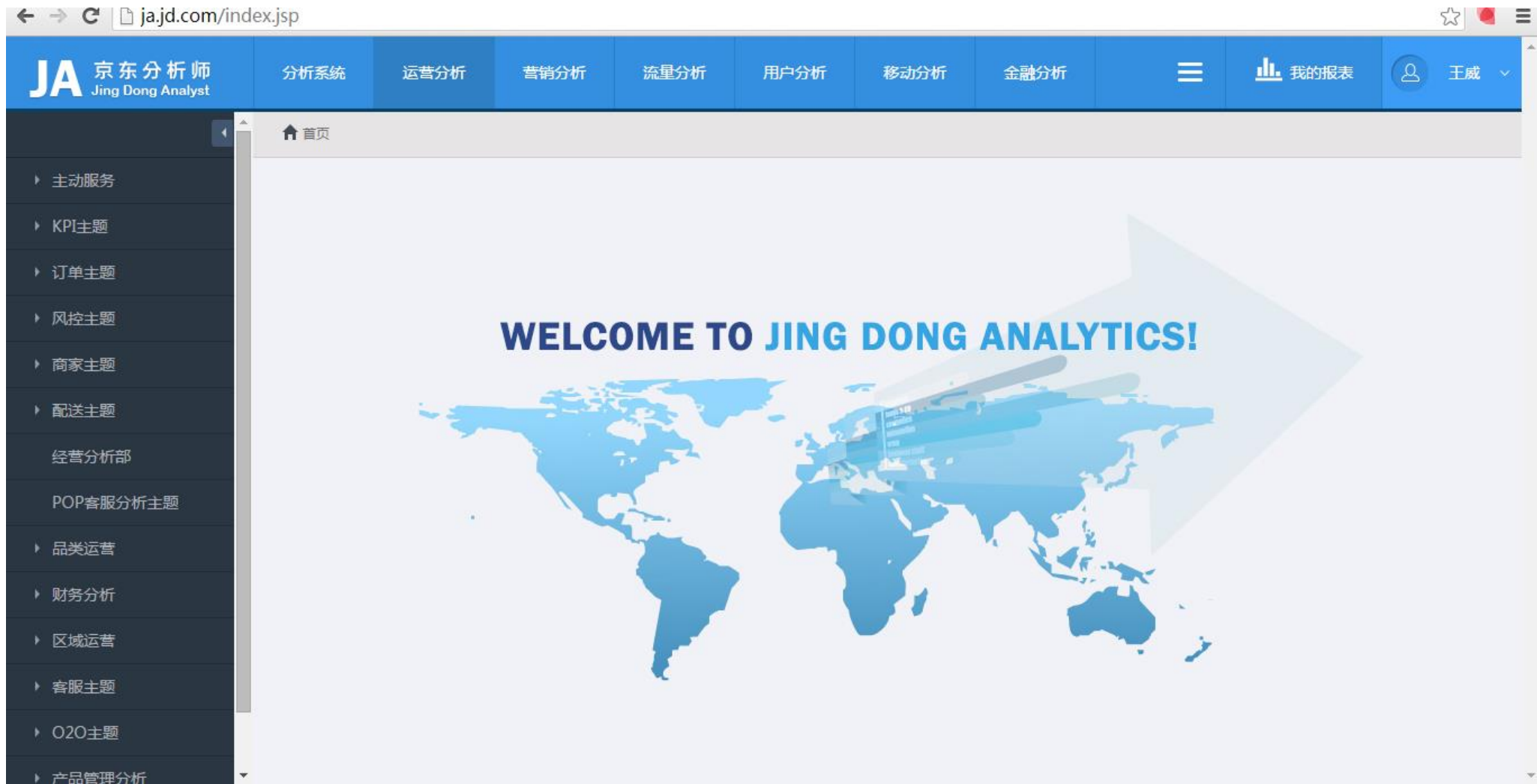
The interface also includes a sidebar with 数据查询 and 数据仓库, and a top right section with 新手指南, 代码运行记录, and 集市状态. The bottom of the page shows the URL bdp.jd.com/index.html and various utility icons.

大数据专家认证

JDW
京东
数据
仓库
业务

京东数据仓库架构-概述

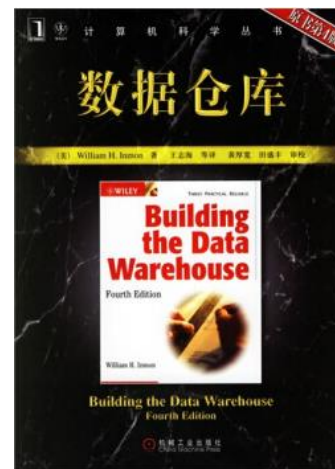
京东分析师



京东数据仓库- 定义

数据仓库始于20世纪80年代中期。由数据仓库之父 W.H Inmon在1991年出版的 “Building the Data Warehouse” （《数据仓库》）一书中提出了准确而又广泛被大家接受的定义。

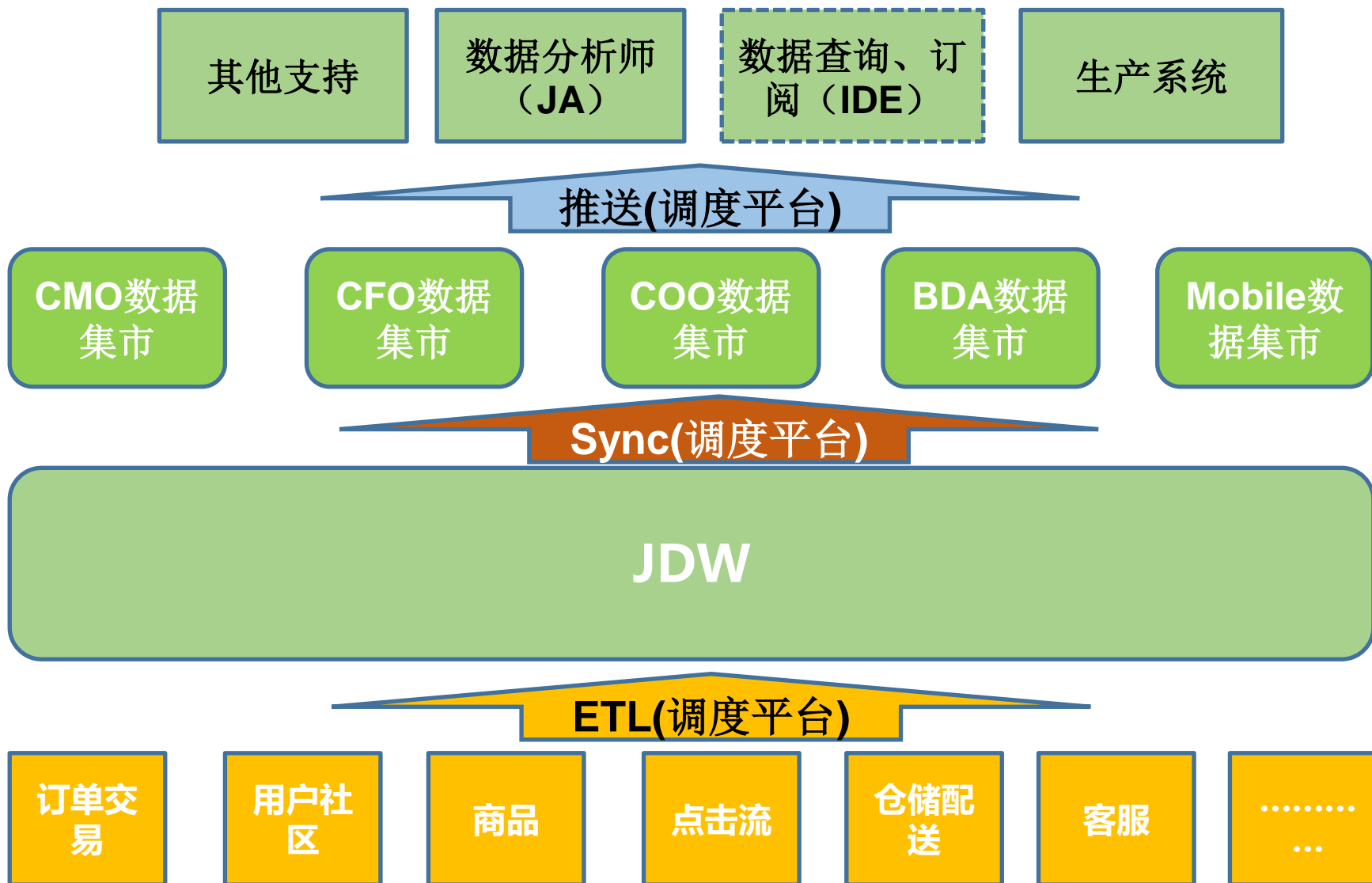
- 面向主题的（ Subject Oriented ）
- 集成的（ Integrated ）
- 非易失的（ Non-Volatile ）
- 随时间变化的（ Time Variant ）



用于支持管理决策的数据集合

京东数据仓库架构-数据流

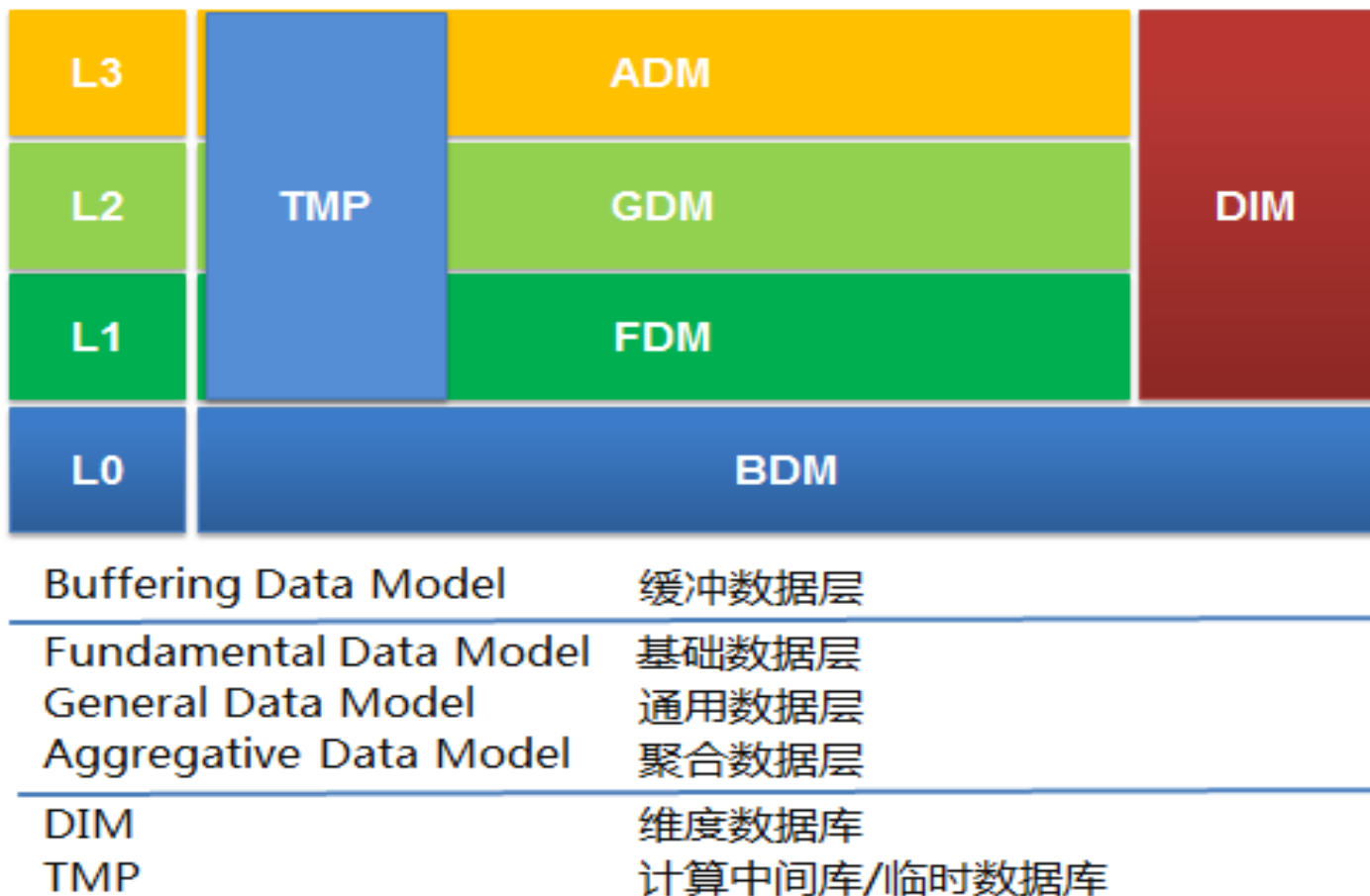
大数据专家认证



JDW
京东
数据
仓库
业务

京东数据仓库架构-架构

EDW的核心数据架构分为四层：缓冲数据层、基础数据层、通用数据层、聚合数据层，其次是临时层和维度层。其示意图如下：



京东数据仓库架构-说明

大数据专家认证

JDW
京东
数据
仓库
业务

序号	数据层次	简称	数据层次用途简述
1	缓冲数据层	BDM	源业务系统数据的快照，保存细节数据，按天保存
2	基础数据层	FDM	按业务概念组织细节数据。
3	通用数据层	GDM	根据京东核心业务价值链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	聚合数据层	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数据集市
5	维度层	DIM	维度是对具体分析对象的分析角度，维度要具备丰富的属性，历史信息的可追溯性，对通用的维表要保持一致性。
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层。

序号	可用数据库	数据库简称	数据仓库用途描述	环境描述
1	EDW_4468_app_test	app_test	数据仓库（多数据层次） APP层开发测试使用	开发测试环境
2	EDW_4468_tmp_test	tmp_test	数据仓库（多数据层次） GDM、ADM开发测试环境	开发测试环境
3	EDW_Learning		供初级人员练习使用	初级人员练习使用

京东数据仓库规范-FDM

一、FDM层表命名规范：

表名 = FDM + 源库名称 + 源表名 + 加载策略

加载策略：拉链表-chain，增量表-无后缀

FDM层表使用方法：

拉链表：分区：dp，dt，end_date

使用start_date和end_date进行日期范围控制

start_date <= #date# and end_date > #date#

可以取某日全量数据的快照

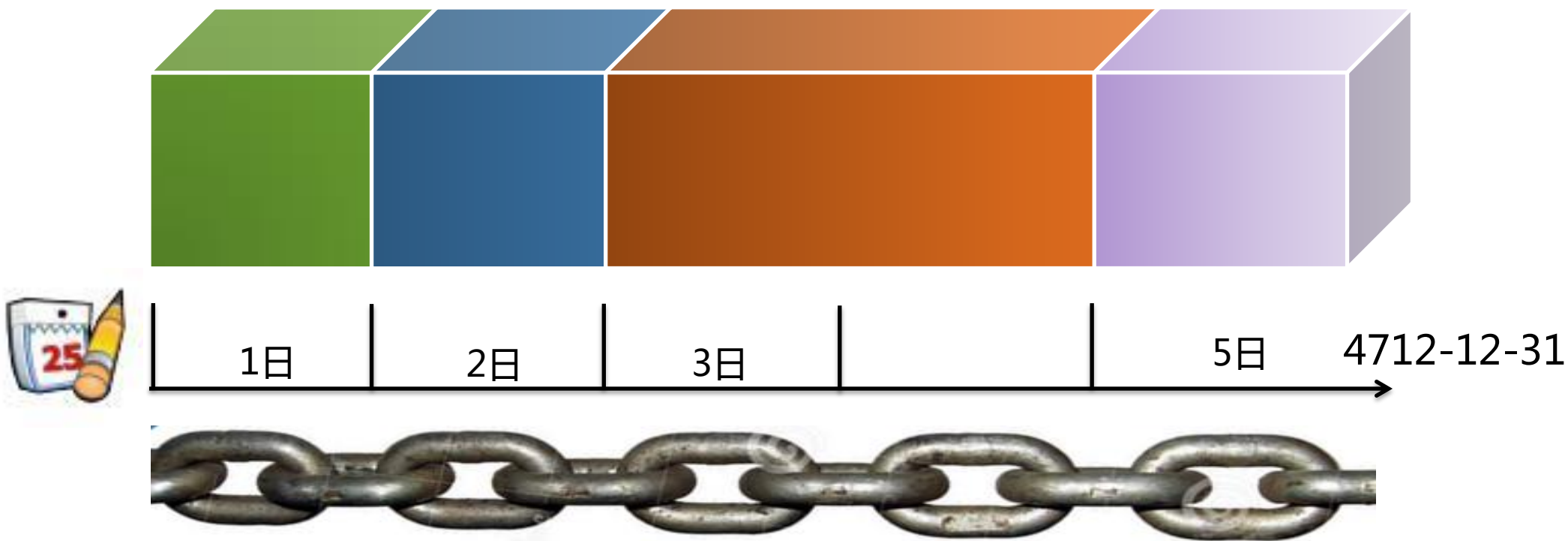
增量表：分区：dt

使用dt进行日期范围控制，可以查询某日的增量数据

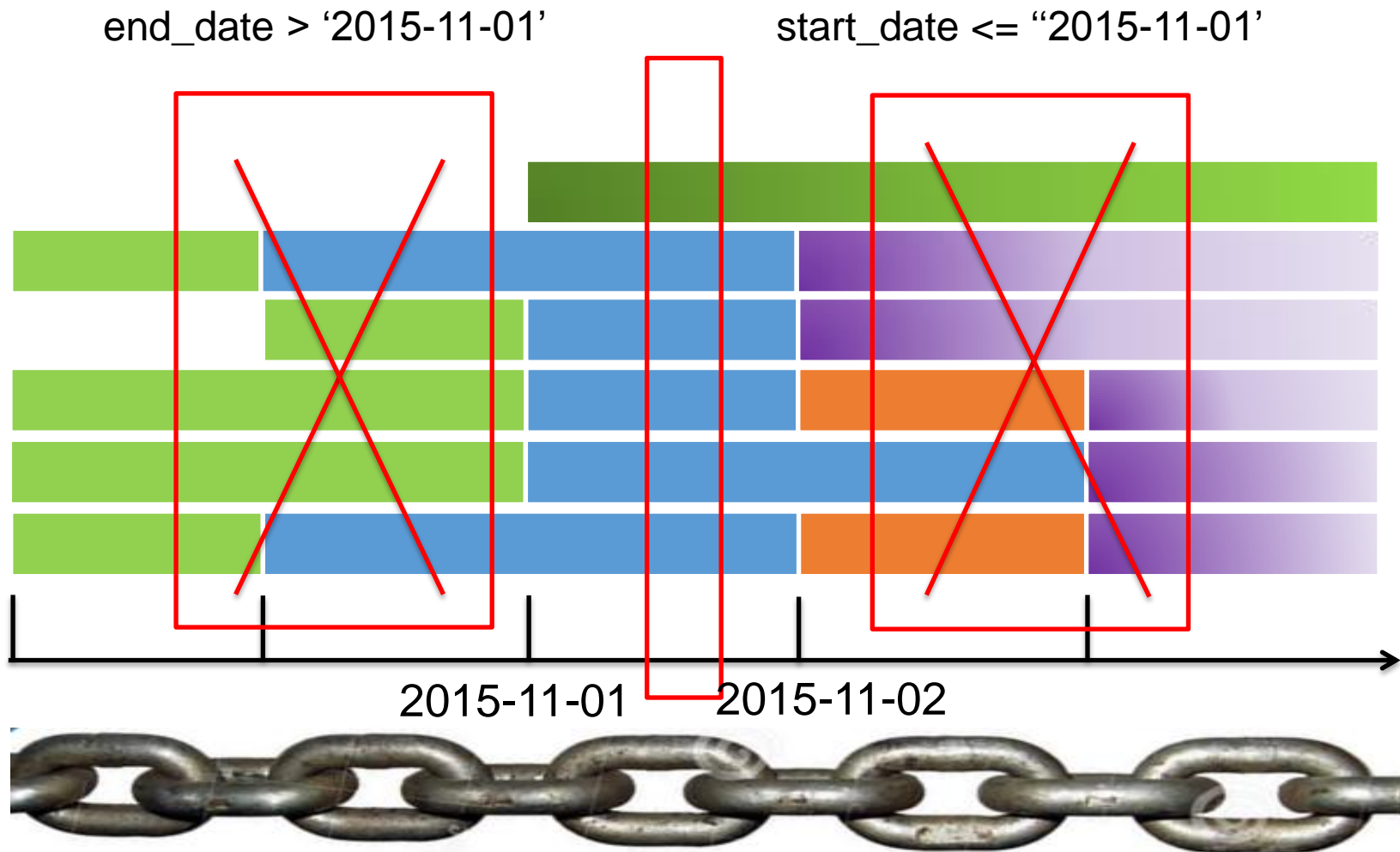
拉链表一条数据的形象展示

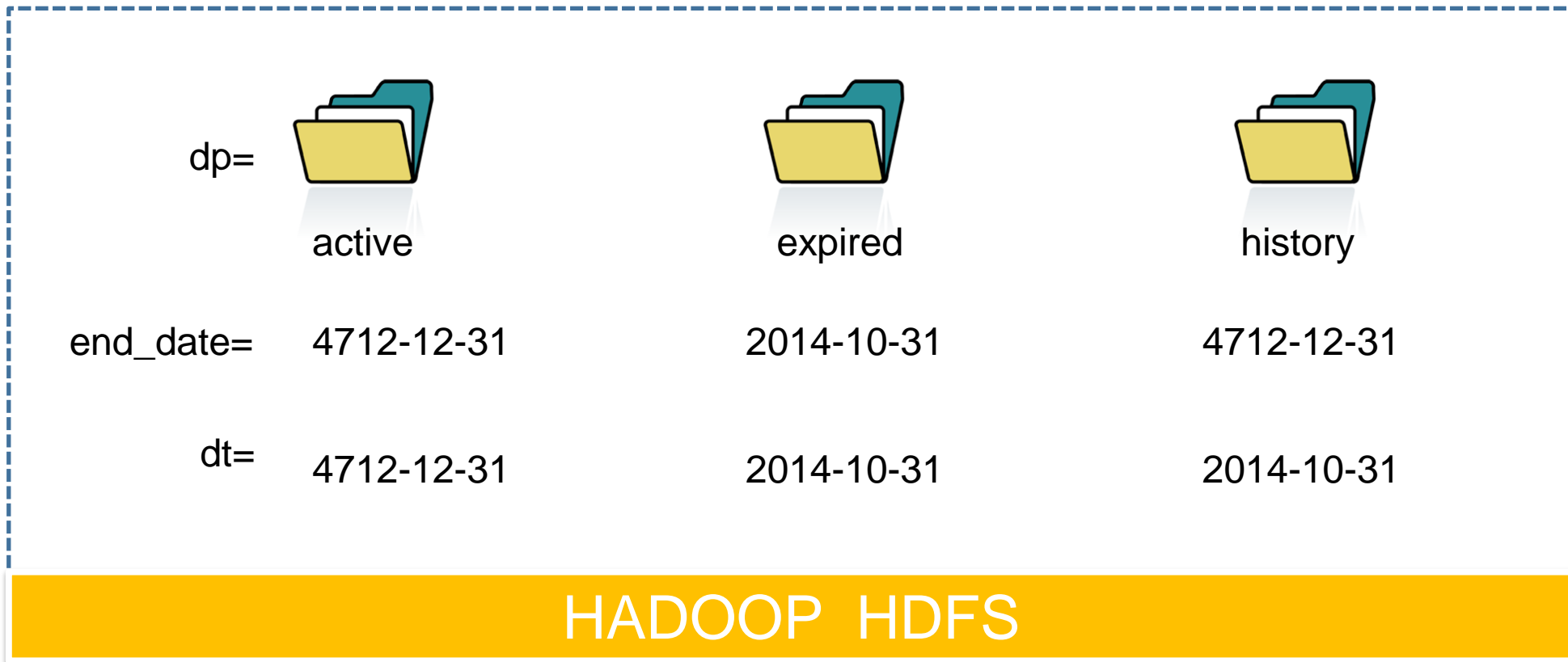
大数据专家认证

JDW
京东
数据
仓库
业务



拉链表多条数据的形象展示





分区使用案例R

一、全量表：

```
select * from gdm.gdm_self_m03_item_sku_da where dt = sysdate(-1)
```

二、增量表：

```
select * from fdm.fdm_pek_orderDetail where dt >= sysdate(-2)
```

三、普通拉链表：

```
select * from fdm.fdm_product_self_sku_1_chain where dp = 'ACTIVE' ;
```

```
select * from fdm.fdm_product_self_sku_1_chain  
where start_date <= '2014-10-01' and end_date > '2014-10-01'
```

四、有结转拉链表

```
select * from fdm.fdm_pek_orders_chain where dp = 'ACTIVE'
```

```
select * from fdm.fdm_pek_orders_chain where dp in ('ACTIVE','HISTORY')
```

```
select * from fdm.fdm_pek_orders_chain where dp in ('ACTIVE','HISTORY')  
and dt >= '2014-10-01'
```

分区使用案例W

一、全量表：

```
select * from gdm.gdm_self_m03_item_sku_da where dt >= sysdate(-2)
```

```
select * from gdm.gdm_self_m03_item_sku_da
```

二、普通拉链表：

```
select * from fdm.fdm_product_self_sku_1_chain
```

```
select * from fdm.fdm_product_self_sku_1_chain where dt > '2014-10-01'
```

四、有结转拉链表

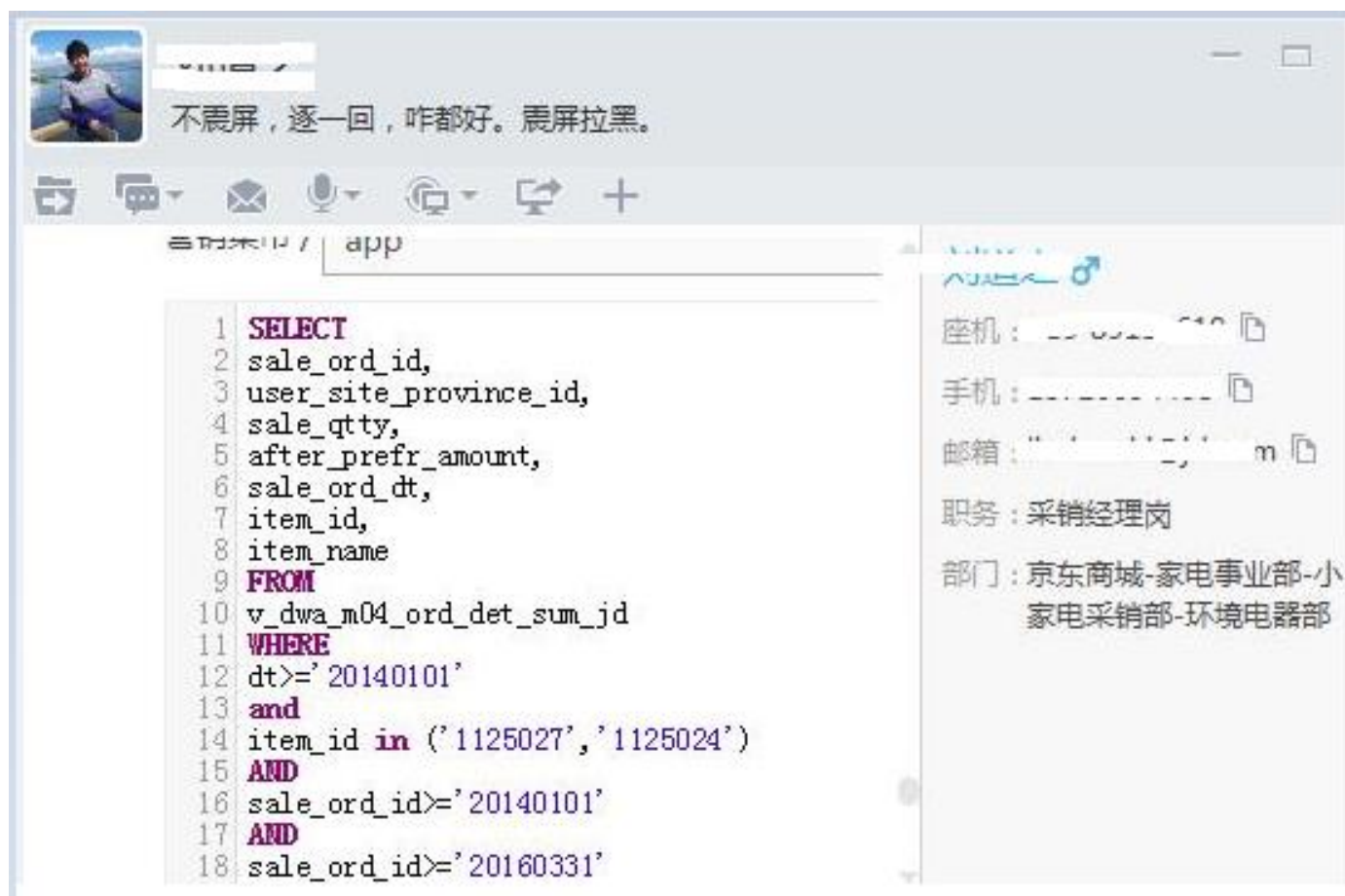
```
select * from fdm.fdm_pek_orders_chain where dt = '2014-10-01'
```

```
select * from fdm.fdm_pek_orders_chain where dp = 'ACTIVE' dt >= '2014-10-01'
```

```
select * from fdm.fdm_pek_orders_chain where and dt >= '2014-10-01'
```

反面教材

咨询问题收集：
这个SQL为什么会跑不出数据来？



在执行之前的SQL，
要检查检查再检查

在咨询之前的SQL，
要思考思考再思考

京东数据仓库规范-GDM

二、GDM数据层命名规则：

增量表名=GDM_主题前缀_主体

全量表名=GDM_主题前缀_主体_DA

主题前缀	主题名称	简称	业务覆盖范围
M01	客户	CUS	供应商、POP商家、团购商家、合作伙伴、用户
M02	组织机构	ORG	员工、部门
M03	商品	ITM	
M04	销售订单	ORD	订单相关
M05	账户	ACT	与账户关联的礼品卡、余额、积分、优惠券等
M06	客户端	CLI	移动客户端、PC客户端、移动设备
M07	营销	CAM	促销、活动、优惠券、礼品卡
M08	仓储	INV	
M09	配送	DIS	
M10	客服	CSC	售后、备件库、呼叫中心、工单
M11	地理区域	LOC	
M12	财务	FIN	
M13	社区	COM	用户的关注、订阅
M14	流量	TRA	Traffice
M15	供应链	SCM	采购、采购退货(退供应商)

京东数据仓库规范-ADM

三、ADM数据层命名规则：

表名 = ADM + 主题英文简称 + 主体 + 后缀(日/周/月/季/年/)

编号	聚合层主题	主题英文简写
1	日百采销	DNST
2	IT数码采销	IT
3	通讯汽车采销	CC
4	家电采销	HEA
5	图书采销	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

四、APP数据层命名规则：

表名 = APP_主体_后缀

后缀：全量加工为DA，非全量考虑采用数据周期为后缀

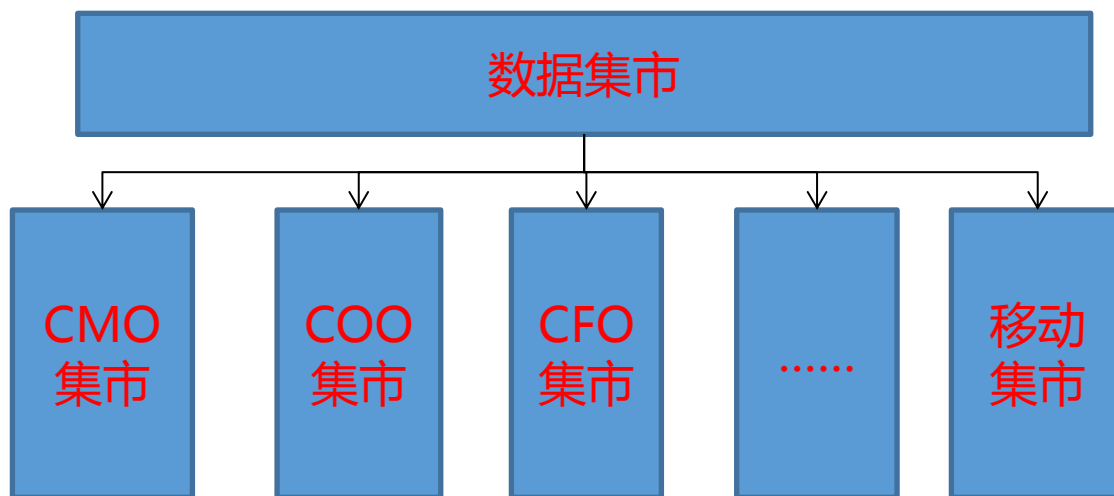
推送任务命名：

- 从hive推往orcle: 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
- 从hive推往hive: hive2hive_adm_s03_item_band_stat

京东数据集市-简介

一、集市简介：

京东数据集市是基于JDW（京东数据仓库）和BDP（大数据平台）构建的面向条线的数据环境，为各条线提供数据应用服务，包含CFO、CMO、COO、BDA、MOBILE等数据集市，目前已为公司23个一级部门提供服务和支持。数据集市包含基础数据层（dim、fdm、gdm、adm）和应用数据层（app）两大部分。



二、集市功能：

数据集市提供面向业务条线的基础数据，通过集成开发环境、调度系统、京东分析师等工具提供数据服务，主要包括以下几类：

- 资源独享，各个业务线资源独立
- 应用层APP表加工，用以支持报表服务
- 临时、周期性数据提取，包括明细和汇总级数据
- 数据推送服务，用以支持线上系统数据计算
- 大数据应用、数据挖掘类需求
- 特殊类型的ETL计算，如准实时库存查询

三、集市服务方式：

针对不同的数据服务需求类型，数据集市分别通过不同的数据出口进行服务，主要包括以下几个：

- **数据知识管理平台**

元数据查询:FDM,GDM,ADM,APP

- **IDE数据集成开发平台**

数据查询；临时、周期性数据提取、订阅；脚本开发、调试、发布

- **大数据平台调度系统**

应用层APP任务调度计算；数据推送服务；数据ETL

- **HIVE客户端**

数据查询；脚本开发、调试、发布

- **京东分析师**

京东数据服务知识库

biwiki.bdp.jd.com/w/index.php?title=首页

数据需求接口人




内控合规及技术研发，董月红 市场部需求，高伟

数据集市负责人

集市业务负责人	邮箱	集市业务接口人	邮箱	集市英文名称	集市简写	集市中文名称
胡浩	huhao@jd.com	谢蔚	xiaweiinfo@jd.com	IPC	IPC	销量预测

谢谢！
Thank you!

大数据专家认证

京东·大数据平台

北京市朝阳区北辰西路8号北辰世纪中心A座16层
16F Building A, North-Star Century Center, 8 Beichen West Street,
Chaoyang District, Beijing 100101