

PB级企业电商离线数仓项目实战【下】

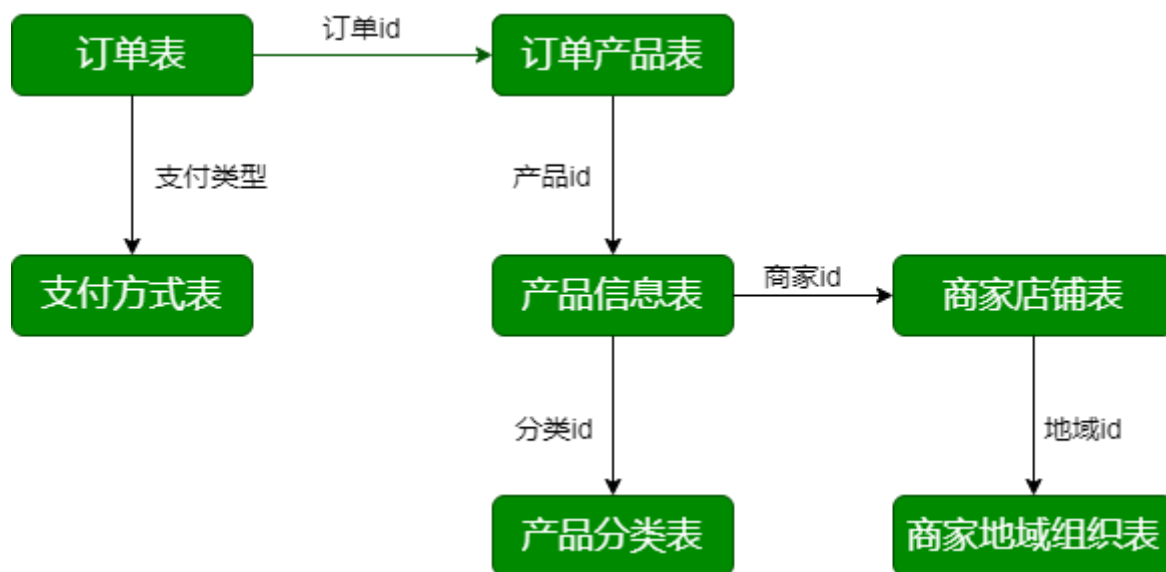
第一部分 电商分析之--核心交易

第1节 业务需求

本主题是电商系统业务中最关键的业务，电商的运营活动都是围绕这个主题展开。

选取的指标包括：订单数、商品数、支付金额。对这些指标按销售区域、商品类型进行分析。

第2节 业务数据库表结构



业务数据库：数据源

- 交易订单表 (trade_orders)
- 订单产品表 (order_product)
- 产品信息表 (product_info)
- 产品分类表 (product_category)
- 商家店铺表 (shops)
- 商家地域组织表 (shop_admin_org)
- 支付方式表 (payments)

交易订单表

```

1 CREATE TABLE `lagou_trade_orders` (
2   `orderId` bigint(11) NOT NULL AUTO_INCREMENT COMMENT '订单id',
3   `orderNo` varchar(20) NOT NULL COMMENT '订单编号',
4   `userId` bigint(11) NOT NULL COMMENT '用户id',
5   `status` tinyint(4) NOT NULL DEFAULT '-2' COMMENT '订单状态
-3:用户拒收 -2:未付款的订单 -1: 用户取消 0:待发货 1:配送中 2:用户确认收货',
6   `productMoney` decimal(11,2) NOT NULL COMMENT '商品金额',
7   `totalMoney` decimal(11,2) NOT NULL COMMENT '订单金额（包括运费）',
8   `payMethod` tinyint(4) NOT NULL DEFAULT '0' COMMENT '支付方式,0:未知;1:支付宝, 2: 微信;3、现金; 4、其他',
9   `isPay` tinyint(4) NOT NULL DEFAULT '0' COMMENT '是否支付 0:未支付 1:已支付',
10  `areaId` int(11) NOT NULL COMMENT '区域最低一级',
11  `tradeSrc` tinyint(4) NOT NULL DEFAULT '0' COMMENT '订单来源
0:商城 1:微信 2:手机版 3:安卓App 4:苹果App',
12  `tradeType` int(11) DEFAULT '0' COMMENT '订单类型',
13  `isRefund` tinyint(4) NOT NULL DEFAULT '0' COMMENT '是否退款
0:否 1: 是',
14  `dataFlag` tinyint(4) NOT NULL DEFAULT '1' COMMENT '订单有效标志
-1: 删除 1:有效',
15  `createTime` varchar(25) NOT NULL COMMENT '下单时间',
16  `payTime` varchar(25) DEFAULT NULL COMMENT '支付时间',
17  `modifiedTime` timestamp NOT NULL DEFAULT '0000-00-00
00:00:00' COMMENT '订单更新时间',
18  PRIMARY KEY (`orderId`)
19 ) ENGINE=InnoDB AUTO_INCREMENT=355 DEFAULT CHARSET=utf8;

```

备注：

- 记录订单的信息
- status。订单状态
- createTime、payTime、modifiedTime。创建时间、支付时间、修改时间

订单产品表

```

1 CREATE TABLE `lagou_order_product` (
2   `id` bigint(11) NOT NULL AUTO_INCREMENT,
3   `orderId` bigint(11) NOT NULL COMMENT '订单id',
4   `productId` bigint(11) NOT NULL COMMENT '商品id',
5   `productNum` bigint(11) NOT NULL DEFAULT '0' COMMENT '商品数量',
6   `productPrice` decimal(11,2) NOT NULL DEFAULT '0.00' COMMENT '商品价格',
7   `money` decimal(11,2) DEFAULT '0.00' COMMENT '付款金额',
8   `extra` text COMMENT '额外信息',
9   `createTime` varchar(25) DEFAULT NULL COMMENT '创建时间',
10  PRIMARY KEY (`id`),
11  KEY `orderId` (`orderId`),
12  KEY `goodsId` (`productId`)
13 ) ENGINE=InnoDB AUTO_INCREMENT=1260 DEFAULT CHARSET=utf8;

```

备注:

- 记录订单中购买产品的信息，包括产品的数量、单价等

产品信息表

```

1 CREATE TABLE `lagou_product_info` (
2   `productId` bigint(11) NOT NULL AUTO_INCREMENT COMMENT '商品id',
3   `productName` varchar(200) NOT NULL COMMENT '商品名称',
4   `shopId` bigint(11) NOT NULL COMMENT '门店ID',
5   `price` decimal(11,2) NOT NULL DEFAULT '0.00' COMMENT '门店价',
6   `issale` tinyint(4) NOT NULL DEFAULT '1' COMMENT '是否上架 0:不上架 1:上架',
7   `status` tinyint(4) NOT NULL DEFAULT '0' COMMENT '是否新品 0:否 1:是',
8   `categoryId` int(11) NOT NULL COMMENT 'goodsCatId 最后一级商品分类ID',
9   `createTime` varchar(25) NOT NULL,
10  `modifyTime` datetime DEFAULT NULL ON UPDATE CURRENT_TIMESTAMP COMMENT '修改时间',
11  PRIMARY KEY (`productId`),
12  KEY `shopId` (`shopId`) USING BTREE,
13  KEY `goodsStatus` (`issale`)
14 ) ENGINE=InnoDB AUTO_INCREMENT=115909 DEFAULT CHARSET=utf8;

```

备注:

- 记录产品的详细信息，对应商家ID、商品属性（是否新品、是否上架）
- createTime、modifyTime。创建时间和修改时间

产品分类表

```
1 CREATE TABLE `lagou_product_category` (  
2   `catId` int(11) NOT NULL AUTO_INCREMENT COMMENT '品类ID',  
3   `parentId` int(11) NOT NULL COMMENT '父ID',  
4   `catName` varchar(20) NOT NULL COMMENT '分类名称',  
5   `isShow` tinyint(4) NOT NULL DEFAULT '1' COMMENT '是否显示  
0:隐藏 1:显示',  
6   `sortNum` int(11) NOT NULL DEFAULT '0' COMMENT '排序号',  
7   `isDel` tinyint(4) NOT NULL DEFAULT '1' COMMENT '删除标志 1:有  
效 -1: 删除',  
8   `createTime` varchar(25) NOT NULL COMMENT '建立时间',  
9   `level` tinyint(4) DEFAULT '0' COMMENT '分类级别，共3级',  
10  PRIMARY KEY (`catId`),  
11  KEY `parentId` (`parentId`, `isShow`, `isDel`)  
12 ) ENGINE=InnoDB AUTO_INCREMENT=10442 DEFAULT CHARSET=utf8;
```

备注：产品分类表，共分3个级别

```
1 -- 第一级产品目录  
2 select catName, catid from lagou_product_category where level =  
3 1;  
4 -- 查看电脑、办公的子类（查看二级目录）  
5 select catName, catid from lagou_product_category where level =  
6 2 and parentId = 32;  
7 -- 查看电脑整机的子类（查看三级目录）  
8 select catName, catid from lagou_product_category where level =  
9 3 and parentId = 10250;
```

商家店铺表

```

1 CREATE TABLE `lagou_shops` (
2   `shopId` int(11) NOT NULL AUTO_INCREMENT COMMENT '商铺ID, 自增',
3   `userId` int(11) NOT NULL COMMENT '商铺联系人ID',
4   `areaId` int(11) DEFAULT '0',
5   `shopName` varchar(100) DEFAULT '' COMMENT '商铺名称',
6   `shopLevel` tinyint(4) NOT NULL DEFAULT '1' COMMENT '店铺等级',
7   `status` tinyint(4) NOT NULL DEFAULT '1' COMMENT '商铺状态',
8   `createTime` date DEFAULT NULL,
9   `modifyTime` datetime DEFAULT NULL COMMENT '修改时间',
10  PRIMARY KEY (`shopId`),
11  KEY `shopStatus` (`status`)
12 ) ENGINE=InnoDB AUTO_INCREMENT=105317 DEFAULT CHARSET=utf8;

```

备注：记录店铺的详细信息

商家地域组织表

```

1 CREATE TABLE `lagou_shop_admin_org` (
2   `id` int(11) NOT NULL AUTO_INCREMENT COMMENT '组织ID',
3   `parentId` int(11) NOT NULL COMMENT '父ID',
4   `orgName` varchar(100) NOT NULL COMMENT '组织名称',
5   `orgLevel` tinyint(4) NOT NULL DEFAULT '1' COMMENT '组织级别
6   1:总部及大区级部门;2: 总部下属的各个部门及基部门;3:具体工作部门',
7   `isDelete` tinyint(4) NOT NULL DEFAULT '0' COMMENT '删除标志,1:删除;0:有效',
8   `createTime` varchar(25) DEFAULT NULL COMMENT '创建时间',
9   `updateTime` varchar(25) DEFAULT NULL COMMENT '最后修改时间',
10  `isShow` tinyint(4) NOT NULL DEFAULT '1' COMMENT '是否显示,0:否 1:是',
11  `orgType` tinyint(4) NOT NULL DEFAULT '1' COMMENT '组织类型,0:总裁办;1:研发;2:销售;3:运营;4:产品',
12  PRIMARY KEY (`id`),
13  KEY `parentId` (`parentId`)
14 ) ENGINE=InnoDB AUTO_INCREMENT=100332 DEFAULT CHARSET=utf8;

```

备注：记录店铺所属区域

支付方式表

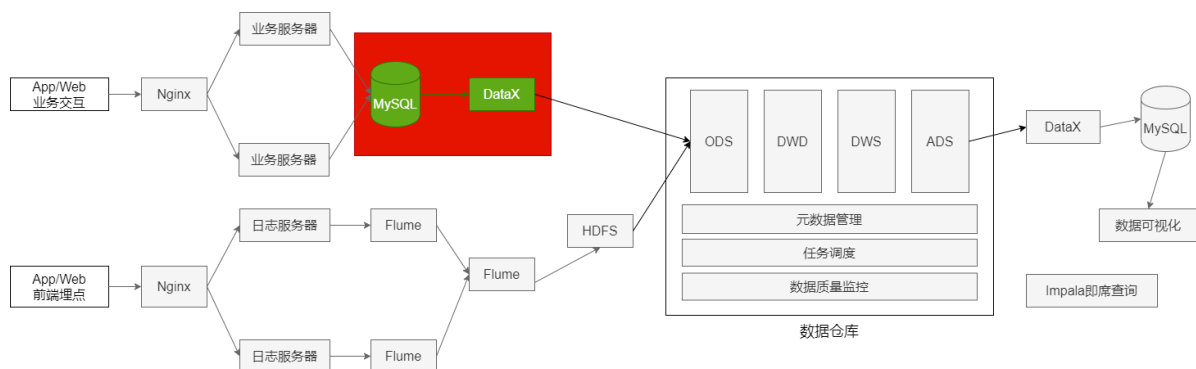
```

1 CREATE TABLE `lagou_payments` (
2   `id` int(11) NOT NULL,
3   `payMethod` varchar(20) DEFAULT NULL,
4   `payName` varchar(255) DEFAULT NULL,
5   `description` varchar(255) DEFAULT NULL,
6   `payOrder` int(11) DEFAULT '0',
7   `online` tinyint(4) DEFAULT NULL,
8   PRIMARY KEY (`id`),
9   KEY `payCode` (`payMethod`)
10 ) ENGINE=InnoDB DEFAULT CHARSET=utf8;

```

备注：记录支付方式

第3节 数据导入

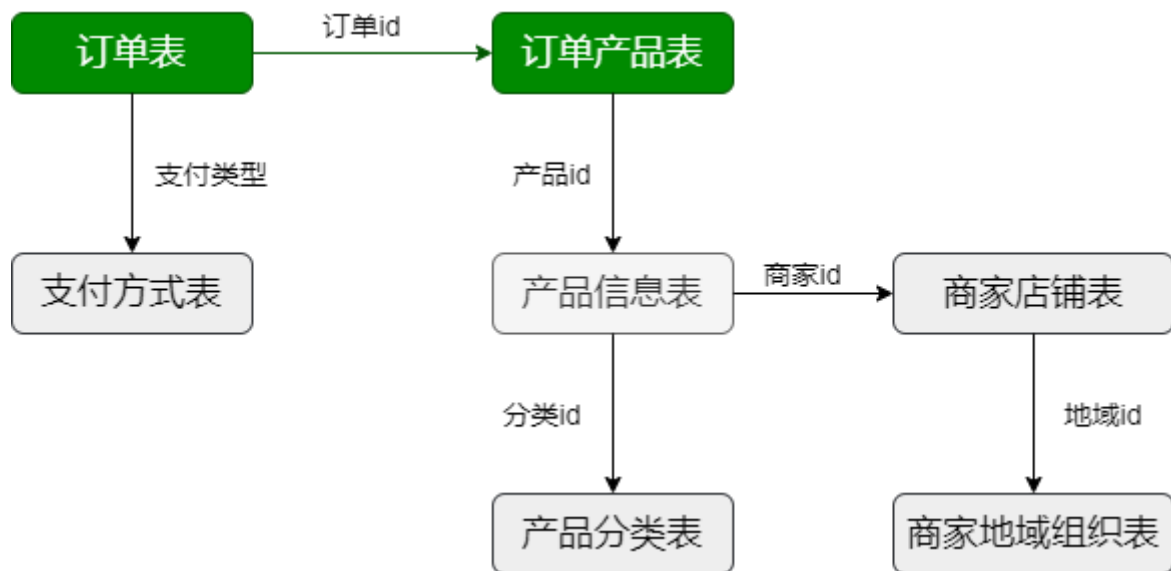


已经确定的事情：DataX、导出7张表的数据。

MySQL 导出：全量导出、增量导出（导出前一天的数据）。

业务数据保存在MySQL中，每日凌晨导入上一天的表数据。

- 表数据量少，采用全量方式导出MySQL
- 表数据量大，而且根据字段能区分出每天新增数据，采用增量方式导出MySQL



3张增量表:

- 订单表 lagou_trade_orders
- 订单产品表 lagou_order_produce
- 产品信息表 lagou_product_info

4张全量表:

- 产品分类表 lagou_product_category
- 商家店铺表 lagou_shops
- 商家地域组织表 lagou_shop_admin_org
- 支付方式表 lagou_payment

3.1、全量数据导入

MySQL => HDFS => Hive

每日加载全量数据，形成新的分区；(ODS如何建表有指导左右)

MySQLReader ==> HdfsWriter

ebiz.lagou_product_category ==> ods.ods_trade_product_category

1、产品分类表

/data/lagoudw/json/product_category.json

```

1 {
2   "job": {

```

```
3      "setting": {
4          "speed": {
5              "channel": 1
6          }
7      },
8      "content": [{
9          "reader": {
10             "name": "mysqlreader",
11             "parameter": {
12                 "username": "teacher",
13                 "password": "teacher123",
14                 "column": [
15                     "catId", "parentId", "catName",
16                     "isShow",
17                     "sortNum",
18                     "isDel",
19                     "createTime",
20                     "level"
21                 ],
22                 "connection": [{
23                     "table": [
24                         "lagou_product_category"
25                     ],
26                     "jdbcUrl": [
27                         "jdbc:mysql://hadoop1:3306/ebiz"
28                     ]
29                 }]
30             }
31         },
32         "writer": {
33             "name": "hdfswriter",
34             "parameter": {
35                 "defaultFS": "hdfs://hadoop1:9000",
36                 "fileType": "text",
37                 "path":
38                     "/user/data/trade.db/product_category/dt=$do_date",
39                 "fileName": "product_category_$do_date",
40                 "column": [
41                     {
42                         "name": "catId",
43                         "type": "INT"
44                     },
45                     {
46                         "name": "parentId",
47                         "type": "INT"
```



```

47         },
48         {
49             "name": "catName",
50             "type": "STRING"
51         },
52         {
53             "name": "isShow",
54             "type": "TINYINT"
55         },
56         {
57             "name": "sortNum",
58             "type": "INT"
59         },
60         {
61             "name": "isDel",
62             "type": "TINYINT"
63         },
64         {
65             "name": "createTime",
66             "type": "STRING"
67         },
68         {
69             "name": "level",
70             "type": "TINYINT"
71         }
72     ],
73     "writeMode": "append",
74     "fieldDelimiter": ",",
75 }
76 }
77 }]
78 }
79 }

```

备注:

- 数据量小的表没有必要使用多个channel; 使用多个channel会生成多个小文件
- 执行命令之前要在HDFS上创建对应的目录: /user/data/trade.db/product_category/dt=yyyy-mm-dd

```

1 do_date='2020-07-01'
2
3 # 创建目录
4 hdfs dfs -mkdir -p
   /user/data/trade.db/product_category/dt=$do_date
5
6 # 数据迁移
7 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/product_category.json
8
9 # 加载数据
10 hive -e "alter table ods.ods_trade_product_category add
   partition(dt='$do_date')"

```

2、商家店铺表

lagou_shops ==> ods.ods_trade_shops

/data/lagoudw/json/shops.json

```

1 {
2     "job": {
3         "setting": {
4             "speed": {
5                 "channel": 1
6             },
7             "errorLimit": {
8                 "record": 0
9             }
10        },
11        "content": [{
12            "reader": {
13                "name": "mysqlreader",
14                "parameter": {
15                    "username": "teacher",
16                    "password": "teacher123",
17                    "column": [
18                        "shopId", "userId", "areaId",
19                        "shopName", "shopLevel", "status", "createTime", "modifyTime"
20                    ],
21                    "connection": [{
22                        "table": [
23                            "lagou_shops"
24                        ]
25                    }
26                ]
27            }
28        }]
29    }
30 }

```

```

23         ],
24         "jdbcurl": [
25             "jdbc:mysql://hadoop1:3306/ebiz"
26         ]
27     }
28 }
29 },
30
31 "writer": {
32     "name": "hdfswriter",
33     "parameter": {
34         "defaultFS": "hdfs://hadoop1:9000",
35         "fileType": "text",
36         "path":
37             "/user/data/trade.db/shops/dt=$do_date",
38         "fileName": "shops_$do_date",
39         "column": [{
40             "name": "shopId",
41             "type": "INT"
42         },
43         {
44             "name": "userId",
45             "type": "INT"
46         },
47         {
48             "name": "areaId",
49             "type": "INT"
50         },
51         {
52             "name": "shopName",
53             "type": "STRING"
54         },
55         {
56             "name": "shopLevel",
57             "type": "TINYINT"
58         },
59         {
60             "name": "status",
61             "type": "TINYINT"
62         },
63         {
64             "name": "createTime",
65             "type": "STRING"
66         }

```

```

67         "name": "modifyTime",
68         "type": "STRING"
69     }
70 ],
71     "writeMode": "append",
72     "fieldDelimiter": ",",
73 }
74 }
75 }]
76 }
77 }

```

```

1  do_date='2020-07-02'
2
3  # 创建目录
4  hdfs dfs -mkdir -p /user/data/trade.db/shops/dt=$do_date
5
6  # 数据迁移
7  python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/shops.json
8
9  # 加载数据
10 hive -e "alter table ods.ods_trade_shops add
    partition(dt='$do_date')"

```

3、商家地域组织表

lagou_shop_admin_org ==> ods.ods_trade_shop_admin_org

/data/lagoudw/json/shop_org.json

```

1  {
2      "job": {
3          "setting": {
4              "speed": {
5                  "channel": 1
6              },
7              "errorLimit": {
8                  "record": 0
9              }
10         },
11         "content": [{

```

```
12         "reader": {
13             "name": "mysqlreader",
14             "parameter": {
15                 "username": "teacher",
16                 "password": "teacher123",
17                 "column": [
18                     "id", "parentId", "orgName",
19                     "orgLevel", "isDelete", "createTime", "updateTime", "isShow",
20                     "orgType"
21                 ],
22                 "connection": [{
23                     "table": [
24                         "lagou_shop_admin_org"
25                     ],
26                     "jdbcurl": [
27                         "jdbc:mysql://hadoop1:3306/ebiz"
28                     ]
29                 }]
30             },
31             "writer": {
32                 "name": "hdfswriter",
33                 "parameter": {
34                     "defaultFS": "hdfs://hadoop1:9000",
35                     "fileType": "text",
36                     "path":
37                     "/user/data/trade.db/shop_org/dt=$do_date",
38                     "fileName": "shop_admin_org_$do_date.dat",
39                     "column": [{
40                         "name": "id",
41                         "type": "INT"
42                     },
43                     {
44                         "name": "parentId",
45                         "type": "INT"
46                     },
47                     {
48                         "name": "orgName",
49                         "type": "STRING"
50                     },
51                     {
52                         "name": "orgLevel",
53                         "type": "TINYINT"
```

```

54         {
55             "name": "isDelete",
56             "type": "TINYINT"
57         },
58         {
59             "name": "createTime",
60             "type": "STRING"
61         },
62         {
63             "name": "updateTime",
64             "type": "STRING"
65         },
66         {
67             "name": "isShow",
68             "type": "TINYINT"
69         },
70         {
71             "name": "orgType",
72             "type": "TINYINT"
73         }
74     ],
75     "writeMode": "append",
76     "fieldDelimiter": ",",
77 }
78 }
79 }
80 }
81 }

```

```

1  do_date='2020-07-01'
2
3  # 创建目录
4  hdfs dfs -mkdir -p /user/data/trade.db/shop_org/dt=$do_date
5
6  # 数据迁移
7  python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/shop_org.json
8
9  # 加载数据
10 hive -e "alter table ods.ods_trade_shop_admin_org add
    partition(dt='$do_date')"

```

4、支付方式表

lagou_payments ==> ods.ods_trade_payments

/data/lagoudw/json/payments.json

```
1  {
2      "job": {
3          "setting": {
4              "speed": {
5                  "channel": 1
6              },
7              "errorLimit": {
8                  "record": 0
9              }
10         },
11         "content": [{
12             "reader": {
13                 "name": "mysqlreader",
14                 "parameter": {
15                     "username": "teacher",
16                     "password": "teacher123",
17                     "column": [
18                         "id", "payMethod", "payName",
19                         "description", "payOrder", "online"
20                     ],
21                     "connection": [{
22                         "table": [
23                             "lagou_payments"
24                         ],
25                         "jdbcUrl": [
26                             "jdbc:mysql://hadoop1:3306/ebiz"
27                         ]
28                     }]
29                 }
30             },
31             "writer": {
32                 "name": "hdfswriter",
33                 "parameter": {
34                     "defaultFS": "hdfs://hadoop1:9000",
35                     "fileType": "text",
36                     "path":
37                         "/user/data/trade.db/payments/dt=$do_date",
38                     "fileName": "payments_${do_date}.dat",
39                     "column": [{
```

```
39         "name": "id",
40         "type": "INT"
41     },
42     {
43         "name": "payMethod",
44         "type": "STRING"
45     },
46     {
47         "name": "payName",
48         "type": "STRING"
49     },
50     {
51         "name": "description",
52         "type": "STRING"
53     },
54     {
55         "name": "payOrder",
56         "type": "INT"
57     },
58     {
59         "name": "online",
60         "type": "TINYINT"
61     }
62 ],
63 "writeMode": "append",
64 "fieldDelimiter": ",",
65 }
66 }
67 }
68 }
69 }
```



```

1 do_date='2020-07-01'
2
3 # 创建目录
4 hdfs dfs -mkdir -p /user/data/trade.db/payments/dt=$do_date
5
6 # 数据迁移
7 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/payments.json
8
9 # 加载数据
10 hive -e "alter table ods.ods_trade_payments add
   partition(dt='$do_date')"

```

3.2、增量数据导入

3张增量表：

- 订单表 lagou_trade_orders
- 订单产品表 lagou_order_produce
- 产品信息表 lagou_product_info

初始数据装载（执行一次）；可以将前面的全量加载作为初次装载

每日加载增量数据（每日数据形成分区）；

1、订单表

lagou_trade_orders ==> ods.ods_trade_orders

/data/lagoudw/json/orders.json

备注：条件的选择，选择时间字段 modifiedTime

```

1 {
2   "job": {
3     "setting": {
4       "speed": {
5         "channel": 1
6       },
7       "errorLimit": {
8         "record": 0
9       }

```

```

10     },
11     "content": [{
12         "reader": {
13             "name": "mysqlreader",
14             "parameter": {
15                 "username": "teacher",
16                 "password": "teacher123",
17                 "connection": [{
18                     "querySql": [
19                         "select orderId, orderNo, userId,
status, productMoney, totalMoney, payMethod, isPay, areaId,
tradeSrc, tradeType, isRefund, dataFlag, createTime, payTime,
modifiedTime from lagou_trade_orders where
date_format(modifiedTime, '%Y-%m-%d')='$do_date'"
20                     ],
21                     "jdbcUrl": [
22                         "jdbc:mysql://hadoop1:3306/ebiz"
23                     ]
24                 }]
25             }
26         },
27
28         "writer": {
29             "name": "hdfswriter",
30             "parameter": {
31                 "defaultFS": "hdfs://hadoop1:9000",
32                 "fileType": "text",
33                 "path":
"/user/data/trade.db/orders/dt=$do_date",
34                 "fileName": "orders_$do_date",
35                 "column": [{
36                     "name": "orderId",
37                     "type": "INT"
38                 },
39                 {
40                     "name": "orderNo",
41                     "type": "STRING"
42                 },
43                 {
44                     "name": "userId",
45                     "type": "BIGINT"
46                 },
47                 {
48                     "name": "status",
49                     "type": "TINYINT"

```

```
50     },
51     {
52         "name": "productMoney",
53         "type": "Float"
54     },
55     {
56         "name": "totalMoney",
57         "type": "Float"
58     },
59     {
60         "name": "payMethod",
61         "type": "TINYINT"
62     },
63     {
64         "name": "isPay",
65         "type": "TINYINT"
66     },
67     {
68         "name": "areaId",
69         "type": "INT"
70     },
71     {
72         "name": "tradeSrc",
73         "type": "TINYINT"
74     },
75     {
76         "name": "tradeType",
77         "type": "INT"
78     },
79     {
80         "name": "isRefund",
81         "type": "TINYINT"
82     },
83     {
84         "name": "dataFlag",
85         "type": "TINYINT"
86     },
87     {
88         "name": "createTime",
89         "type": "STRING"
90     },
91     {
92         "name": "payTime",
93         "type": "STRING"
94     },
95 }
```

```

95         {
96             "name": "modifiedTime",
97             "type": "STRING"
98         }
99     ],
100     "writeMode": "append",
101     "fieldDelimiter": ",",
102 }
103 }
104 }
105 }
106 }

```

```

1  -- MySQL 中的时间日期转换
2  select date_format(createTime, '%Y-%m-%d'), count(*)
3  from lagou_trade_orders
4  group by date_format(createTime, '%Y-%m-%d');

```

```

1  do_date='2020-07-12'
2
3  # 创建目录
4  hdfs dfs -mkdir -p /user/data/trade.db/orders/dt=$do_date
5
6  # 数据迁移
7  python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/orders.json
8
9  # 加载数据
10 hive -e "alter table ods.ods_trade_orders add
    partition(dt='$do_date')"

```

2、订单明细表

lagou_order_product ==> ods.ods_trade_order_product

/data/lagoudw/json/order_product.json

```

1  {
2      "job": {
3          "setting": {

```

```

4         "speed": {
5             "channel": 1
6         },
7         "errorLimit": {
8             "record": 0
9         }
10    },
11    "content": [{
12        "reader": {
13            "name": "mysqlreader",
14            "parameter": {
15                "username": "teacher",
16                "password": "teacher123",
17                "connection": [{
18                    "querySql": [
19                        "select id, orderId, productId,
productNum, productPrice, money, extra, createTime from
lagou_order_product where date_format(createTime, '%Y-%m-%d')
= '$do_date' "
20                ],
21                "jdbcUrl": [
22                    "jdbc:mysql://hadoop1:3306/ebiz"
23                ]
24            }]
25        },
26    },
27
28    "writer": {
29        "name": "hdfswriter",
30        "parameter": {
31            "defaultFS": "hdfs://hadoop1:9000",
32            "fileType": "text",
33            "path":
"/user/data/trade.db/order_product/dt=$do_date",
34            "fileName": "order_product_$do_date.dat",
35            "column": [{
36                "name": "id",
37                "type": "INT"
38            },
39            {
40                "name": "orderId",
41                "type": "INT"
42            },
43            {
44                "name": "productId",

```

```
45         "type": "INT"
46     },
47     {
48         "name": "productNum",
49         "type": "INT"
50     },
51     {
52         "name": "productPrice",
53         "type": "Float"
54     },
55     {
56         "name": "money",
57         "type": "Float"
58     },
59     {
60         "name": "extra",
61         "type": "STRING"
62     },
63     {
64         "name": "createTime",
65         "type": "STRING"
66     }
67 ],
68 "writeMode": "append",
69 "fieldDelimiter": ",",
70 }
71 }
72 }
73 }
74 }
```

```

1 do_date='2020-07-12'
2
3 # 创建目录
4 hdfs dfs -mkdir -p
   /user/data/trade.db/order_product/dt=$do_date
5
6 # 数据迁移
7 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/order_product.json
8
9 # 加载数据
10 hive -e "alter table ods.ods_trade_order_product add
   partition(dt='$do_date')"

```

3、产品明细表

lagou_product_info ==> ods.ods_trade_product_info

/data/lagoudw/json/product_info.json

```

1 {
2     "job": {
3         "setting": {
4             "speed": {
5                 "channel": 1
6             },
7             "errorLimit": {
8                 "record": 0
9             }
10        },
11        "content": [{
12            "reader": {
13                "name": "mysqlreader",
14                "parameter": {
15                    "username": "teacher",
16                    "password": "teacher123",
17                    "connection": [{
18                        "querySql": [
19                            "select productid, productname,
20                                shopid, price, issale, status, categoryid, createtime,
                                modifytime from lagou_product_info where
                                date_format(modifyTime, '%Y-%m-%d') = '$do_date' "

```

```
21         "jdbcurl": [  
22             "jdbc:mysql://hadoop1:3306/ebiz"  
23         ]  
24     }]  
25 }  
26 },  
27  
28 "writer": {  
29     "name": "hdfswriter",  
30     "parameter": {  
31         "defaultFS": "hdfs://hadoop1:9000",  
32         "fileType": "text",  
33         "path":  
34         "/user/data/trade.db/product_info/dt=$do_date",  
35         "fileName": "product_info_$do_date.dat",  
36         "column": [{  
37             "name": "productid",  
38             "type": "BIGINT"  
39         },  
40         {  
41             "name": "productname",  
42             "type": "STRING"  
43         },  
44         {  
45             "name": "shopid",  
46             "type": "STRING"  
47         },  
48         {  
49             "name": "price",  
50             "type": "FLOAT"  
51         },  
52         {  
53             "name": "issale",  
54             "type": "TINYINT"  
55         },  
56         {  
57             "name": "status",  
58             "type": "TINYINT"  
59         },  
60         {  
61             "name": "categoryid",  
62             "type": "STRING"  
63         },  
64         {  
            "name": "createTime",
```



```

65         "type": "STRING"
66     },
67     {
68         "name": "modifytime",
69         "type": "STRING"
70     }
71 ],
72 "writeMode": "append",
73 "fieldDelimiter": ",",
74 }
75 }
76 }
77 }
78 }

```

```

1  do_date='2020-07-12'
2
3  # 创建目录
4  hdfs dfs -mkdir -p
   /user/data/trade.db/product_info/dt=$do_date
5
6  # 数据迁移
7  python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
   /data/lagoudw/json/product_info.json
8
9  # 加载数据
10 hive -e "alter table ods.ods_trade_product_info add
    partition(dt='$do_date')"

```

第4节 ODS层建表与数据加载

ODS建表：

- ODS层的表结构与源数据基本类似（列名及数据类型）；
- ODS层的表名遵循统一的规范；

4.1 ODS层建表

所有的表都是分区表；字段之间的分隔符为，；为表的数据数据文件指定了位置；

```
1 DROP TABLE IF EXISTS `ods.ods_trade_orders`;
2 CREATE EXTERNAL TABLE `ods.ods_trade_orders` (
3     `orderid` int,
4     `orderno` string,
5     `userid` bigint,
6     `status` tinyint,
7     `productmoney` decimal(10, 0),
8     `totalmoney` decimal(10, 0),
9     `paymethod` tinyint,
10    `ispay` tinyint,
11    `areaaid` int,
12    `tradesrc` tinyint,
13    `tradetype` int,
14    `isrefund` tinyint,
15    `dataflag` tinyint,
16    `createtime` string,
17    `paytime` string,
18    `modifiedtime` string)
19 COMMENT '订单表'
20 PARTITIONED BY (`dt` string)
21 row format delimited fields terminated by ','
22 location '/user/data/trade.db/orders/';
23
24 DROP TABLE IF EXISTS `ods.ods_trade_order_product`;
25 CREATE EXTERNAL TABLE `ods.ods_trade_order_product` (
26     `id` string,
27     `orderid` decimal(10,2),
28     `productid` string,
29     `productnum` string,
30     `productprice` string,
31     `money` string,
32     `extra` string,
33     `createtime` string)
34 COMMENT '订单明细表'
35 PARTITIONED BY (`dt` string)
36 row format delimited fields terminated by ','
37 location '/user/data/trade.db/order_product/';
38
39 DROP TABLE IF EXISTS `ods.ods_trade_product_info`;
40 CREATE EXTERNAL TABLE `ods.ods_trade_product_info` (
41     `productid` bigint,
```

```
42     `productname` string,
43     `shopid` string,
44     `price` decimal(10,0),
45     `issale` tinyint,
46     `status` tinyint,
47     `categoryid` string,
48     `createtime` string,
49     `modifytime` string)
50 COMMENT '产品信息表'
51 PARTITIONED BY (`dt` string)
52 row format delimited fields terminated by ','
53 location '/user/data/trade.db/product_info/';
54
55 DROP TABLE IF EXISTS `ods.ods_trade_product_category`;
56 CREATE EXTERNAL TABLE `ods.ods_trade_product_category`(
57     `catid` int,
58     `parentid` int,
59     `catname` string,
60     `isshow` tinyint,
61     `sortnum` int,
62     `isdel` tinyint,
63     `createtime` string,
64     `level` tinyint)
65 COMMENT '产品分类表'
66 PARTITIONED BY (`dt` string)
67 row format delimited fields terminated by ','
68 location '/user/data/trade.db/product_category';
69
70 DROP TABLE IF EXISTS `ods.ods_trade_shops`;
71 CREATE EXTERNAL TABLE `ods.ods_trade_shops`(
72     `shopid` int,
73     `userid` int,
74     `areaid` int,
75     `shopname` string,
76     `shoplevel` tinyint,
77     `status` tinyint,
78     `createtime` string,
79     `modifytime` string)
80 COMMENT '商家店铺表'
81 PARTITIONED BY (`dt` string)
82 row format delimited fields terminated by ','
83 location '/user/data/trade.db/shops';
84
85 DROP TABLE IF EXISTS `ods.ods_trade_shop_admin_org`;
86 CREATE EXTERNAL TABLE `ods.ods_trade_shop_admin_org`(`
```

```

87     `id` int,
88     `parentid` int,
89     `orgname` string,
90     `orglevel` tinyint,
91     `isdelete` tinyint,
92     `createtime` string,
93     `updatetime` string,
94     `isshow` tinyint,
95     `orgType` tinyint)
96 COMMENT '商家地域组织表'
97 PARTITIONED BY (`dt` string)
98 row format delimited fields terminated by ','
99 location '/user/data/trade.db/shop_org/';
100
101 DROP TABLE IF EXISTS `ods.ods_trade_payments`;
102 CREATE EXTERNAL TABLE `ods.ods_trade_payments`(
103     `id` string,
104     `paymethod` string,
105     `payname` string,
106     `description` string,
107     `payorder` int,
108     `online` tinyint)
109 COMMENT '支付方式表'
110 PARTITIONED BY (`dt` string)
111 row format delimited fields terminated by ','
112 location '/user/data/trade.db/payments/';

```

4.2 ODS层数据加载

DataX仅仅是将数据导入到了 HDFS，数据并没有与Hive表建立关联。

脚本的任务：数据迁移、数据加载到ODS层；

对于增量加载数据而言：初始数据加载；该任务仅执行一次，不在脚本中。

/data/lagoudw/script/trade/ods_load_trade.sh

```

1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ] ;then
6      do_date=$1

```

```
7 else
8     do_date=`date -d "-1 day" +%F`
9 fi
10
11 # 创建目录
12 hdfs dfs -mkdir -p
13 /user/data/trade.db/product_category/dt=$do_date
14 hdfs dfs -mkdir -p /user/data/trade.db/shops/dt=$do_date
15 hdfs dfs -mkdir -p /user/data/trade.db/shop_org/dt=$do_date
16 hdfs dfs -mkdir -p /user/data/trade.db/payments/dt=$do_date
17 hdfs dfs -mkdir -p /user/data/trade.db/orders/dt=$do_date
18 hdfs dfs -mkdir -p
19 /user/data/trade.db/order_product/dt=$do_date
20 hdfs dfs -mkdir -p
21 /user/data/trade.db/product_info/dt=$do_date
22
23 # 数据迁移
24 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
25 /data/lagoudw/json/product_category.json
26 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
27 /data/lagoudw/json/shops.json
28 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
29 /data/lagoudw/json/shop_org.json
30 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
31 /data/lagoudw/json/payments.json
32 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
33 /data/lagoudw/json/orders.json
34 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
35 /data/lagoudw/json/order_product.json
36 python $DATAX_HOME/bin/datax.py -p "-Ddo_date=$do_date"
37 /data/lagoudw/json/product_info.json
38
39 # 加载 ODS 层数据
40 sql="
41 alter table ods.ods_trade_orders add partition(dt='$do_date');
42 alter table ods.ods_trade_order_product add
43 partition(dt='$do_date');
44 alter table ods.ods_trade_product_info add
45 partition(dt='$do_date');
46 alter table ods.ods_trade_product_category add
47 partition(dt='$do_date');
48 alter table ods.ods_trade_shops add partition(dt='$do_date');
49 alter table ods.ods_trade_shop_admin_org add
50 partition(dt='$do_date');
```

```
37 alter table ods.ods_trade_payments add
    partition(dt='${do_date}');
38 "
39
40 hive -e "$sql"
```

特点：工作量大，繁琐，容易出错；与数据采集工作在一起；

第5节 缓慢变化维与周期性事实表

5.1、缓慢变化维

缓慢变化维（SCD；Slowly Changing Dimensions）。在现实世界中，维度的属性随着时间的流失发生缓慢的变化（缓慢是相对事实表而言，事实表数据变化的速度比维度表快）。

处理维度表的历史变化信息的问题称为处理缓慢变化维的问题，简称SCD问题。处理缓慢变化维的方法有以下几种常见方式：

- 保留原值
- 直接覆盖
- 增加新属性列
- 快照表
- 拉链表

1、保留原始值

维度属性值不做更改，保留原始值。

如商品上架售卖时间：一个商品上架售卖后由于其他原因下架，后来又再次上架，此种情况产生了多个商品上架售卖时间。如果业务重点关注的是商品首次上架售卖时间，则采用该方式。

2、直接覆盖

修改维度属性为最新值，直接覆盖，不保留历史信息。

如商品属于哪个品类：当商品品类发生变化时，直接重写为新品类。

3、增加新属性列

在维度表中增加新的一列，原先属性列存放上一版本的属性值，当前属性列存放当前版本的属性值，还可以增加一列记录变化的时间。

缺点：只能记录最后一次变化的信息。

维度属性值变化前：

商品ID	品类
SPU001	科技

维度属性值变化后：

商品ID	当前品类	原品类	变更时间
SPU001	教育	科技	2020-01-01

4、快照表

每天保留一份全量数据。

简单、高效。缺点是信息重复，浪费磁盘空间。

适用范围：维表不能太大

使用场景多，范围广；一般而言维表都不大。

5、拉链表

拉链表适合于：**表的数据量大**，而且数据会发生新增和变化，但是大部分是不变的（数据发生变化的百分比不大），且是缓慢变化的（如电商中用户信息表中的某些用户基本属性不可能每天都变化）。主要目的是节省存储空间。

适用场景：

- 表的数据量大
- 表中部分字段会被更新
- 表中记录变量的比例不高
- 需要保留历史信息

5.2、维表拉链表应用案例

user id	mobile	reg_date	start_date
001	13551111111	2020-03-01	2020-06-20
002	13561111111	2020-04-01	2020-06-20
003	13571111111	2020-05-01	2020-06-20
004	13581111111	2020-06-01	2020-06-20
002	13562222222	2020-04-01	2020-06-21
004	13582222222	2020-06-01	2020-06-21
005	13552222222	2020-06-21	2020-06-21
004	13333333333	2020-06-01	2020-06-22
005	13533333333	2020-06-21	2020-06-22
006	13733333333	2020-06-22	2020-06-22
001	13554444444	2020-03-01	2020-06-23
003	13574444444	2020-05-01	2020-06-23
005	13555554444	2020-06-21	2020-06-23
007	18600744444	2020-06-23	2020-06-23
008	18600844444	2020-06-23	2020-06-23

1、创建表加载数据（准备工作）

```
1  -- 用户信息
2  DROP TABLE IF EXISTS test.userinfo;
3  CREATE TABLE test.userinfo(
4      userid STRING COMMENT '用户编号',
5      mobile STRING COMMENT '手机号码',
6      regdate STRING COMMENT '注册日期')
7  COMMENT '用户信息'
8  PARTITIONED BY (dt string)
9  row format delimited fields terminated by ',';
10
11 -- 拉链表（存放用户历史信息）
12 -- 拉链表不是分区表；多了两个字段start_date、end_date
13 DROP TABLE IF EXISTS test.userhis;
14 CREATE TABLE test.userhis(
15     userid STRING COMMENT '用户编号',
16     mobile STRING COMMENT '手机号码',
17     regdate STRING COMMENT '注册日期',
18     start_date STRING,
19     end_date STRING)
20 COMMENT '用户信息拉链表'
21 row format delimited fields terminated by ',';
22
```



```

23  -- 数据(/data/lagoudw/data/userinfo.dat)
24  001,1355111111,2020-03-01,2020-06-20
25  002,1356111111,2020-04-01,2020-06-20
26  003,1357111111,2020-05-01,2020-06-20
27  004,1358111111,2020-06-01,2020-06-20
28
29  002,1356222222,2020-04-01,2020-06-21
30  004,1358222222,2020-06-01,2020-06-21
31  005,1355222222,2020-06-21,2020-06-21
32
33  004,1333333333,2020-06-01,2020-06-22
34  005,1353333333,2020-06-21,2020-06-22
35  006,1373333333,2020-06-22,2020-06-22
36
37  001,1355444444,2020-03-01,2020-06-23
38  003,1357444444,2020-05-01,2020-06-23
39  005,1355554444,2020-06-21,2020-06-23
40  007,1860074444,2020-06-23,2020-06-23
41  008,1860084444,2020-06-23,2020-06-23
42
43  -- 静态分区数据加载（略）
44  /data/lagoudw/data/userinfo0620.dat
45  001,1355111111,2020-03-01
46  002,1356111111,2020-04-01
47  003,1357111111,2020-05-01
48  004,1358111111,2020-06-01
49  load data local inpath '/data/lagoudw/data/userinfo0620.dat'
    into table test.userinfo
50  partition(dt='2020-06-20');
51
52  -- 动态分区数据加载：分区的值是不固定的，由输入数据确定
53  -- 创建中间表(非分区表)
54  drop table if exists test.tmp1;
55  create table test.tmp1 as
56  select * from test.userinfo;
57  -- tmp1 非分区表，使用系统默认的字段分割符'\001'
58  alter table test.tmp1 set serdeproperties('field.delim','=',');
59  -- 向中间表加载数据
60  load data local inpath '/data/lagoudw/data/userinfo.dat' into
    table test.tmp1;
61
62  -- 从中间表向分区表加载数据
63  set hive.exec.dynamic.partition.mode=nonstrict;
64  insert into table test.userinfo
65  partition(dt)

```

```
66 select * from test.tmp1;
```

与动态分区相关的参数

hive.exec.dynamic.partition

- Default Value: `false` prior to Hive 0.9.0; `true` in Hive 0.9.0 and later
- Added In: Hive 0.6.0

Whether or not to allow dynamic partitions in DML/DDL.

表示开启动态分区功能

hive.exec.dynamic.partition.mode

- Default Value: `strict`
- Added In: Hive 0.6.0

In `strict` mode, the user must specify at least one static partition in case the user accidentally overwrites all partitions. In `nonstrict` mode all partitions are allowed to be dynamic.

Set to nonstrict to support INSERT ... VALUES, UPDATE, and DELETE transactions (Hive 0.14.0 and later).

strict: 最少需要有一个是静态分区

nonstrict: 可以全部是动态分区

hive.exec.max.dynamic.partitions

- Default Value: `1000`
- Added In: Hive 0.6.0

Maximum number of dynamic partitions allowed to be created in total.

表示一个动态分区语句可以创建的最大动态分区个数，超出报错

hive.exec.max.dynamic.partitions.pernode

- Default Value: `100`
- Added In: Hive 0.6.0

Maximum number of dynamic partitions allowed to be created in each mapper/reducer node.

表示每个mapper / reducer可以允许创建的最大动态分区个数，默认是100，超出则会报错。

hive.exec.max.created.files

- Default Value: 100000
- Added In: Hive 0.7.0

Maximum number of HDFS files created by all mappers/reducers in a MapReduce job.

表示一个MR job可以创建的最大文件个数，超出报错。

2、拉链表的实现

userinfo(分区表) => userid、mobile、regdate => 每日变更的数据（修改的+新增的） / 历史数据（第一天）

userhis（拉链表） => 多了两个字段 start_date / end_date

```
1  -- 步骤:
2  -- 1、userinfo初始化（2020-06-20）。获取历史数据
3  001,1355111111,2020-03-01,2020-06-20
4  002,1356111111,2020-04-01,2020-06-20
5  003,1357111111,2020-05-01,2020-06-20
6  004,1358111111,2020-06-01,2020-06-20
7
8  -- 2、初始化拉链表（2020-06-20）。userinfo => userhis
9  insert overwrite table test.userhis
10 select  userid, mobile, regdate, dt as start_date, '9999-12-
    31' as end_date
11       from test.userinfo
12       where dt='2020-06-20';
13
14 -- 3、次日新增数据（2020-06-21）；获取新增数据
15 002,1356222222,2020-04-01,2020-06-21
16 004,1358222222,2020-06-01,2020-06-21
17 005,1355222222,2020-06-21,2020-06-21
18
19 -- 4、构建拉链表(userhis)（2020-06-21）【核心】 userinfo(2020-06-
    21) + userhis => userhis
```

```

20 -- userinfo: 新增数据
21 -- userhis: 历史数据
22
23 -- 第一步: 处理新增数据【userinfo】（处理逻辑与加载历史数据类似）
24 select  userid, mobile, regdate, dt as start_date, '9999-12-
31' as end_date
25     from test.userinfo
26    where dt='2020-06-21';
27
28 -- 第二步: 处理历史数据【userhis】（历史包括两部分: 变化的、未变化的）
29 -- 变化的: start_date:不变; end_date: 传入日期-1
30 -- 未变化的: 不做处理
31
32 -- 观察数据
33 select A.userid, B.userid, B.mobile, B.regdate, B.start_Date,
34        B.end_date
35    from (select * from test.userinfo where dt='2020-06-21') A
36        right join test.userhis B
37         on A.userid=B.userid;
38
39 -- 编写SQL, 处理历史数据
40 select B.userid,
41        B.mobile,
42        B.regdate,
43        B.start_Date,
44        case when B.end_date='9999-12-31' and A.userid is not
45 null
46            then date_add('2020-06-21', -1)
47            else B.end_date
48        end as end_date
49    from (select * from test.userinfo where dt='2020-06-21') A
50        right join test.userhis B
51         on A.userid=B.userid;
52
53 -- 最终的处理（新增+历史数据）
54 insert overwrite table test.userhis
55 select  userid, mobile, regdate, dt as start_date, '9999-12-
56 31' as end_date
57     from test.userinfo
58    where dt='2020-06-21'
59
60 union all
61
62 select B.userid,
63        B.mobile,

```

```

61         B.regdate,
62         B.start_Date,
63         case when B.end_date='9999-12-31' and A.userid is not
null
64             then date_add('2020-06-21', -1)
65             else B.end_date
66         end as end_date
67     from (select * from test.userinfo where dt='2020-06-21') A
68     right join test.userhis B
69     on A.userid=B.userid;
70
71
72 -- 5、第三日新增数据（2020-06-22）：获取新增数据
73 004,13333333333,2020-06-01,2020-06-22
74 005,13533333333,2020-06-21,2020-06-22
75 006,13733333333,2020-06-22,2020-06-22
76
77 -- 6、构建拉链表（2020-06-22） userinfo(2020-06-22) + userhis =>
userhis
78
79 -- 7、第四日新增数据（2020-06-23）
80 001,13554444444,2020-03-01,2020-06-23
81 003,13574444444,2020-05-01,2020-06-23
82 005,13555544444,2020-06-21,2020-06-23
83 007,18600744444,2020-06-23,2020-06-23
84 008,18600844444,2020-06-23,2020-06-23
85
86 -- 8、构建拉链表(2020-06-23)

```

处理拉链表的脚本(测试脚本):

/data/lagoudw/data/userzipper.sh

```

1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ] ;then
6      do_date=$1
7  else
8      do_date=`date -d "-1 day" +%F`
9  fi
10
11  sql="

```

```

12 insert overwrite table test.userhis
13 select  userid, mobile, regdate, dt as start_date, '9999-12-
14         31' as end_date
15     from test.userinfo
16     where dt='$do_date'
17
18 union all
19
20 select B.userid,
21        B.mobile,
22        B.regdate,
23        B.start_Date,
24        case when B.end_date='9999-12-31' and A.userid is not
25        null
26            then date_add('$do_date', -1)
27            else B.end_date
28        end as end_date
29     from (select * from test.userinfo where dt='$do_date') A
30     right join test.userhis B
31         on A.userid=B.userid;
32
33 "
34
35 hive -e "$sql"

```

拉链表的使用：

```

1  -- 查看拉链表中最新数据(2020-06-23以后的数据)
2  select * from userhis where end_date='9999-12-31';
3
4  -- 查看拉链表中给定日期数据("2020-06-22")
5  select * from userhis where start_date <= '2020-06-22' and
6  end_date >= '2020-06-22';
7
8  -- 查看拉链表中给定日期数据("2020-06-21")
9  select * from userhis where start_date <= '2020-06-21' and
10 end_date >= '2020-06-21';
11
12 -- 查看拉链表中给定日期数据("2020-06-20")
13 select * from userhis where start_date <= '2020-06-20' and
14 end_date >= '2020-06-20';

```

3、拉链表的回滚

1	06-20拉链表数据(sh xxx.sh 2020-06-20; 在2020-06-21日凌晨发出命令):				
2	001	13551111111	2020-03-01	2020-06-20	9999-12-31
3	002	13561111111	2020-04-01	2020-06-20	9999-12-31
4	003	13571111111	2020-05-01	2020-06-20	9999-12-31
5	004	13581111111	2020-06-01	2020-06-20	9999-12-31
6					
7	001	13551111111	2020-03-01	2020-06-20	9999-12-31
8	002	13561111111	2020-04-01	2020-06-20	9999-12-31
9	003	13571111111	2020-05-01	2020-06-20	9999-12-31
10	004	13581111111	2020-06-01	2020-06-20	9999-12-31
11					
12	06-21拉链表数据(sh xxx.sh 2020-06-21):				
13	001	13551111111	2020-03-01	2020-06-20	9999-12-31
14	002	13561111111	2020-04-01	2020-06-20	2020-06-20
15	002	13562222222	2020-04-01	2020-06-21	9999-12-31
16	003	13571111111	2020-05-01	2020-06-20	9999-12-31
17	004	13581111111	2020-06-01	2020-06-20	2020-06-20
18	004	13582222222	2020-06-01	2020-06-21	9999-12-31
19	005	13552222222	2020-06-21	2020-06-21	9999-12-31
20					
21	001	13551111111	2020-03-01	2020-06-20	9999-12-31
22	002	13561111111	2020-04-01	2020-06-20	2020-06-20
23	002	13562222222	2020-04-01	2020-06-21	9999-12-31
24	003	13571111111	2020-05-01	2020-06-20	9999-12-31

25	004	13581111111	2020-06-01	2020-06-20	2020-
	06-20	1			
26	004	13582222222	2020-06-01	2020-06-21	9999-
	12-31	2			
27	005	13552222222	2020-06-21	2020-06-21	9999-
	12-31	2			
28					
29	06-22拉链表数据:				
30	001	13551111111	2020-03-01	2020-06-20	9999-
	12-31				
31	002	13561111111	2020-04-01	2020-06-20	2020-
	06-20				
32	002	13562222222	2020-04-01	2020-06-21	9999-
	12-31				
33	003	13571111111	2020-05-01	2020-06-20	9999-
	12-31				
34	004	13581111111	2020-06-01	2020-06-20	2020-
	06-20				
35	004	13582222222	2020-06-01	2020-06-21	2020-
	06-21				
36	004	13333333333	2020-06-01	2020-06-22	9999-
	12-31				
37	005	13552222222	2020-06-21	2020-06-21	2020-
	06-21				
38	005	13533333333	2020-06-21	2020-06-22	9999-
	12-31				
39	006	13733333333	2020-06-22	2020-06-22	9999-
	12-31				
40					
41	001	13551111111	2020-03-01	2020-06-20	9999-
	12-31	2			
42	002	13561111111	2020-04-01	2020-06-20	2020-
	06-20	1			
43	002	13562222222	2020-04-01	2020-06-21	9999-
	12-31	2			
44	003	13571111111	2020-05-01	2020-06-20	9999-
	12-31	2			
45	004	13581111111	2020-06-01	2020-06-20	2020-
	06-20	1			
46	004	13582222222	2020-06-01	2020-06-21	2020-
	06-21	1			
47	004	13333333333	2020-06-01	2020-06-22	9999-
	12-31	2			
48	005	13552222222	2020-06-21	2020-06-21	2020-
	06-21	1			

49	005	13533333333	2020-06-21	2020-06-22	9999-
	12-31	2			
50	006	13733333333	2020-06-22	2020-06-22	9999-
	12-31	2			
51					
52	001	13551111111	2020-03-01	2020-06-20	9999-
	12-31	2			
53	002	13561111111	2020-04-01	2020-06-20	2020-
	06-20	1			
54	002	13562222222	2020-04-01	2020-06-21	9999-
	12-31	2			
55	003	13571111111	2020-05-01	2020-06-20	9999-
	12-31	2			
56	004	13581111111	2020-06-01	2020-06-20	2020-
	06-20	1			
57	004	13582222222	2020-06-01	2020-06-21	2020-
	06-21	1			
58	004	13333333333	2020-06-01	2020-06-22	9999-
	12-31	2			
59	005	13552222222	2020-06-21	2020-06-21	2020-
	06-21	1			
60	005	13533333333	2020-06-21	2020-06-22	9999-
	12-31	2			
61	006	13733333333	2020-06-22	2020-06-22	9999-
	12-31	2			
62					
63	06-23	拉链表数据:			
64	001	13551111111	2020-03-01	2020-06-20	2020-
	06-22				
65	001	13554444444	2020-03-01	2020-06-23	9999-
	12-31				
66	002	13561111111	2020-04-01	2020-06-20	2020-
	06-20				
67	002	13562222222	2020-04-01	2020-06-21	9999-
	12-31				
68	003	13571111111	2020-05-01	2020-06-20	2020-
	06-22				
69	003	13574444444	2020-05-01	2020-06-23	9999-
	12-31				
70	004	13581111111	2020-06-01	2020-06-20	2020-
	06-20				
71	004	13582222222	2020-06-01	2020-06-21	2020-
	06-21				
72	004	13333333333	2020-06-01	2020-06-22	9999-
	12-31				

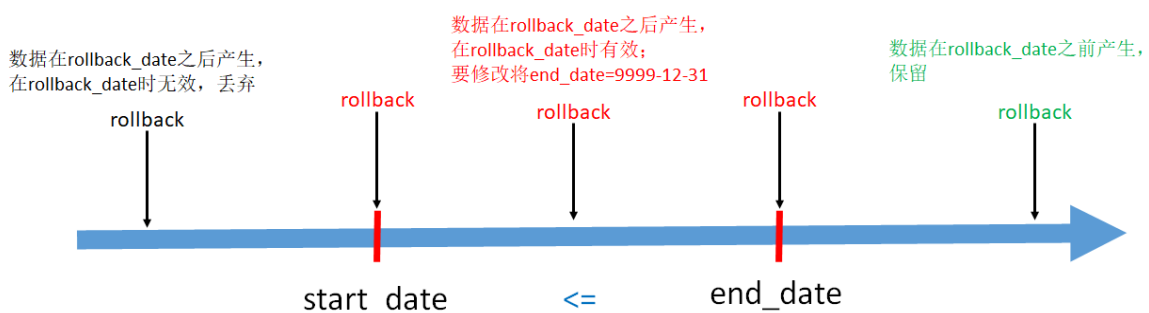
73	005	1355222222	2020-06-21	2020-06-21	2020-06-21
74	005	1353333333	2020-06-21	2020-06-22	2020-06-22
75	005	1355554444	2020-06-21	2020-06-23	9999-12-31
76	006	1373333333	2020-06-22	2020-06-22	9999-12-31
77	007	1860074444	2020-06-23	2020-06-23	9999-12-31
78	008	1860084444	2020-06-23	2020-06-23	9999-12-31

06-23拉链表的状态。假设回滚到2020-06-22, rollback_date

001	1355111111	2020-03-01	2020-06-20	2020-06-22	end_date:9999-12-31; s <= r <= e
001	1355444444	2020-03-01	2020-06-23	9999-12-31	
002	1356111111	2020-04-01	2020-06-20	2020-06-20	end_date < rollback_date 保留
002	1356222222	2020-04-01	2020-06-21	9999-12-31	end_date:9999-12-31
003	1357111111	2020-05-01	2020-06-20	2020-06-22	end_date:9999-12-31
003	1357444444	2020-05-01	2020-06-23	9999-12-31	
004	1358111111	2020-06-01	2020-06-20	2020-06-20	end_date < rollback_date 保留
004	1358222222	2020-06-01	2020-06-21	2020-06-21	end_date < rollback_date 保留
004	1333333333	2020-06-01	2020-06-22	9999-12-31	end_date:9999-12-31
005	1355222222	2020-06-21	2020-06-21	2020-06-21	end_date < rollback_date 保留
005	1353333333	2020-06-21	2020-06-22	2020-06-22	end_date:9999-12-31
005	1355554444	2020-06-21	2020-06-23	9999-12-31	
006	1373333333	2020-06-22	2020-06-22	9999-12-31	end_date:9999-12-31
007	1860074444	2020-06-23	2020-06-23	9999-12-31	
008	1860084444	2020-06-23	2020-06-23	9999-12-31	

由于种种原因需要将拉链表恢复到 rollback_date 那一天的数据。此时有：

- end_date < rollback_date, 即结束日期 < 回滚日期。表示该行数据在 rollback_date 之前产生, 这些数据需要原样保留
- start_date <= rollback_date <= end_date, 即开始日期 <= 回滚日期 <= 结束日期。这些数据是回滚日期之后产生的, 但是需要修改。将end_date 改为 9999-12-31
- 其他数据不用管



按以上方案进行编码：

1、处理 end_date < rollback_date 的数据，保留

```
1 select userid, mobile, regdate, start_date, end_date, '1' as
   tag
2   from test.userhis
3  where end_date < '2020-06-22';
```

2、处理 start_date <= rollback_date <= end_date 的数据，设置 end_date=9999-12-31

```
1 select userid, mobile, regdate, start_date, '9999-12-31' as
   end_date, '2' as tag
2   from test.userhis
3  where start_date <= '2020-06-22' and end_date >= '2020-06-
   22';
```

3、将前面两步的数据写入临时表tmp（拉链表）

```
1 drop table test.tmp;
2 create table test.tmp as
3 select userid, mobile, regdate, start_date, end_date, '1' as
   tag
4   from test.userhis
5  where end_date < '2020-06-22'
6
7 union all
8
9 select userid, mobile, regdate, start_date, '9999-12-31' as
   end_date, '2' as tag
10  from test.userhis
11  where start_date <= '2020-06-22' and end_date >= '2020-06-
   22';
12
13 select * from test.tmp cluster by userid, start_date;
```

4、模拟脚本

/data/lagoudw/data/zippertmp.sh

```
1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ] ;then
6      do_date=$1
7  else
8      do_date=`date -d "-1 day" +%F`
9  fi
10
11  sql="
12  drop table test.tmp;
13
14  create table test.tmp as
15  select userid, mobile, regdate, start_date, end_date, '1' as
    tag
16      from test.userhis
17      where end_date < '$do_date'
18
19  union all
20
21  select userid, mobile, regdate, start_date, '9999-12-31' as
    end_date, '2' as tag
22      from test.userhis
23      where start_date <= '$do_date' and end_date >= '$do_date';
24  "
25
26  hive -e "$sql"
```

逐天回滚，检查数据；

方案二：保存一段时间的增量数据(userinfo)，定期对拉链表做备份（如一个月做一次备份）；如需回滚，直接在备份的拉链表上重跑增量数据。处理简单

5.3、周期性事实表

有如下订单表，6月20号有3条记录(001/002/003)：

订单创建日期	订单编号	订单状态
2020-06-20	001	创建订单
2020-06-20	002	创建订单
2020-06-20	003	支付完成

6月21日，表中有5条记录。其中新增2条记录（004/005），修改1条记录（001）：

订单创建日期	订单编号	订单状态
2020-06-20	001	支付完成（从创建到支付）
2020-06-20	002	创建订单
2020-06-20	003	支付完成
2020-06-21	004	创建订单
2020-06-21	005	创建订单

6月22日，表中有6条记录。其中新增1条记录（006），修改2条记录（003/005）：

订单创建日期	订单编号	订单状态
2020-06-20	001	支付完成
2020-06-20	002	创建订单
2020-06-20	003	已发货（从支付到发货）
2020-06-21	004	创建订单
2020-06-21	005	支付完成（从创建到支付）
2020-06-22	006	创建订单

订单事实表的处理方法：

- 只保留一份全量。数据和6月22日的记录一样，如果需要查看6月21日订单001的状态，则无法满足；

- 每天都保留一份全量。在数据仓库中可以在找到所有的历史信息，但数据量大了，而且很多信息都是重复的，会造成较大的存储浪费；

使用拉链表保存历史信息，会有下面这张表。历史拉链表，既能满足保存历史数据的需求，也能节省存储资源。

订单创建日期	订单编号	订单状态	begin_date	end_date
2020-06-20	001	创建订单	2020-06-20	2020-06-20
2020-06-20	001	支付完成	2020-06-21	9999-12-31
2020-06-20	002	创建订单	2020-06-20	9999-12-31
2020-06-20	003	支付完成	2020-06-20	2020-06-21
2020-06-20	003	已发货	2020-06-22	9999-12-31
2020-06-21	004	创建订单	2020-06-21	9999-12-31
2020-06-21	005	创建订单	2020-06-21	2020-06-21
2020-06-21	005	支付完成	2020-06-22	9999-12-31
2020-06-22	006	创建订单	2020-06-22	9999-12-31

1、前提条件

- 订单表的刷新频率为一天，当天获取前一天的增量数据；
- 如果一个订单在一天内有多次状态变化，只记录最后一个状态的信息；
- 订单状态包括三个：创建、支付、完成；
- 创建时间和修改时间只取到天，如果源订单表中没有状态修改时间，那么抽取增量就比较麻烦，需要有个机制来确保能抽取到每天的增量数据；

数仓ODS层有订单表，数据按日分区，存放每天的增量数据：

```

1 DROP TABLE test.ods_orders;
2 CREATE TABLE test.ods_orders(
3   orderid INT,
4   createtime STRING,
5   modifiedtime STRING,
6   status STRING
7 ) PARTITIONED BY (dt STRING)
8 row format delimited fields terminated by ',';

```

数仓DWD层有订单拉链表，存放订单的历史状态数据：

```

1 DROP TABLE test.dwd_orders;
2 CREATE TABLE test.dwd_orders(
3   orderid INT,
4   createtime STRING,
5   modifiedtime STRING,
6   status STRING,
7   start_date STRING,
8   end_date STRING
9 )
10 row format delimited fields terminated by ',';

```

2、周期性事实表拉链表的实现

1、全量初始化

```

1 -- 数据文件order1.dat
2 001,2020-06-20,2020-06-20,创建
3 002,2020-06-20,2020-06-20,创建
4 003,2020-06-20,2020-06-20,支付
5
6 load data local inpath '/data/lagoudw/data/order1.dat' into
7 table test.ods_orders partition(dt='2020-06-20');
8
9 INSERT overwrite TABLE test.dwd_orders
10 SELECT orderid, createtime, modifiedtime, status,
11         createtime AS start_date,
12         '9999-12-31' AS end_date
13 FROM test.ods_orders
14 WHERE dt='2020-06-20';

```

增量抽取

```
1  -- 数据文件order2.dat
2  001,2020-06-20,2020-06-21,支付
3  004,2020-06-21,2020-06-21,创建
4  005,2020-06-21,2020-06-21,创建
5
6  load data local inpath '/data/lagoudw/data/order2.dat' into
   table test.ods_orders partition(dt='2020-06-21');
```

增量刷新历史数据

```
1  -- 拉链表中的数据分两部实现：新增数据(ods_orders)、历史数据
   (dwd_orders)
2
3  -- 处理新增数据
4  SELECT orderid,
5         createtime,
6         modifiedtime,
7         status,
8         modifiedtime AS start_date,
9         '9999-12-31' AS end_date
10 FROM test.ods_orders
11 where dt='2020-06-21';
12
13 -- 处理历史数据。历史数据包括：有修改、无修改的数据
14 -- ods_orders 与 dwd_orders 进行表连接
15 -- 连接上，说明数据被修改
16 -- 未连接上，说明数据未被修改
17 select A.orderid,
18        A.createtime,
19        A.modifiedtime,
20        A.status,
21        A.start_date,
22        case when B.orderid is not null and A.end_date>'2020-
06-21'
23            then '2020-06-20'
24            else A.end_date
25        end end_date
26 from dwd_orders A
```

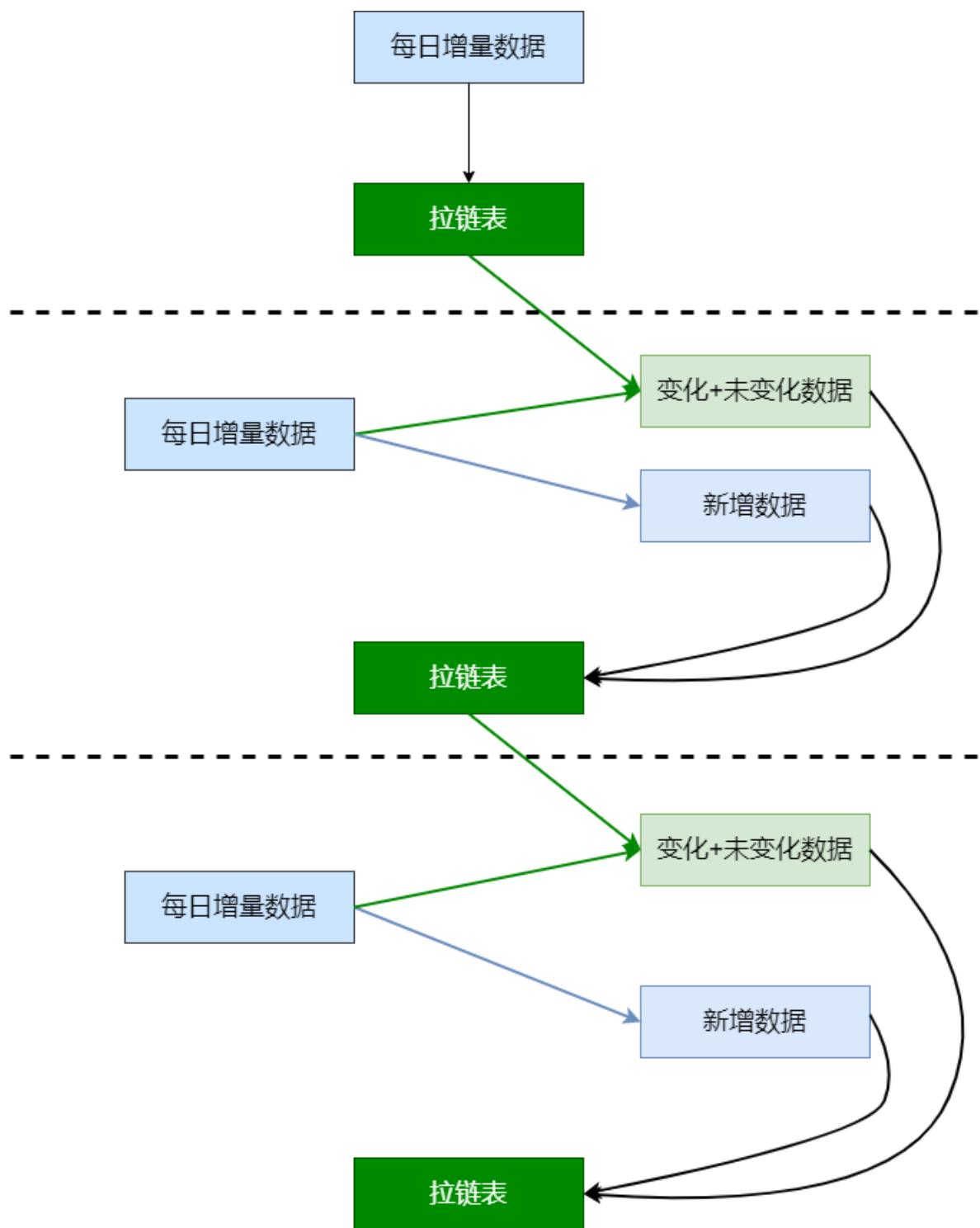


```

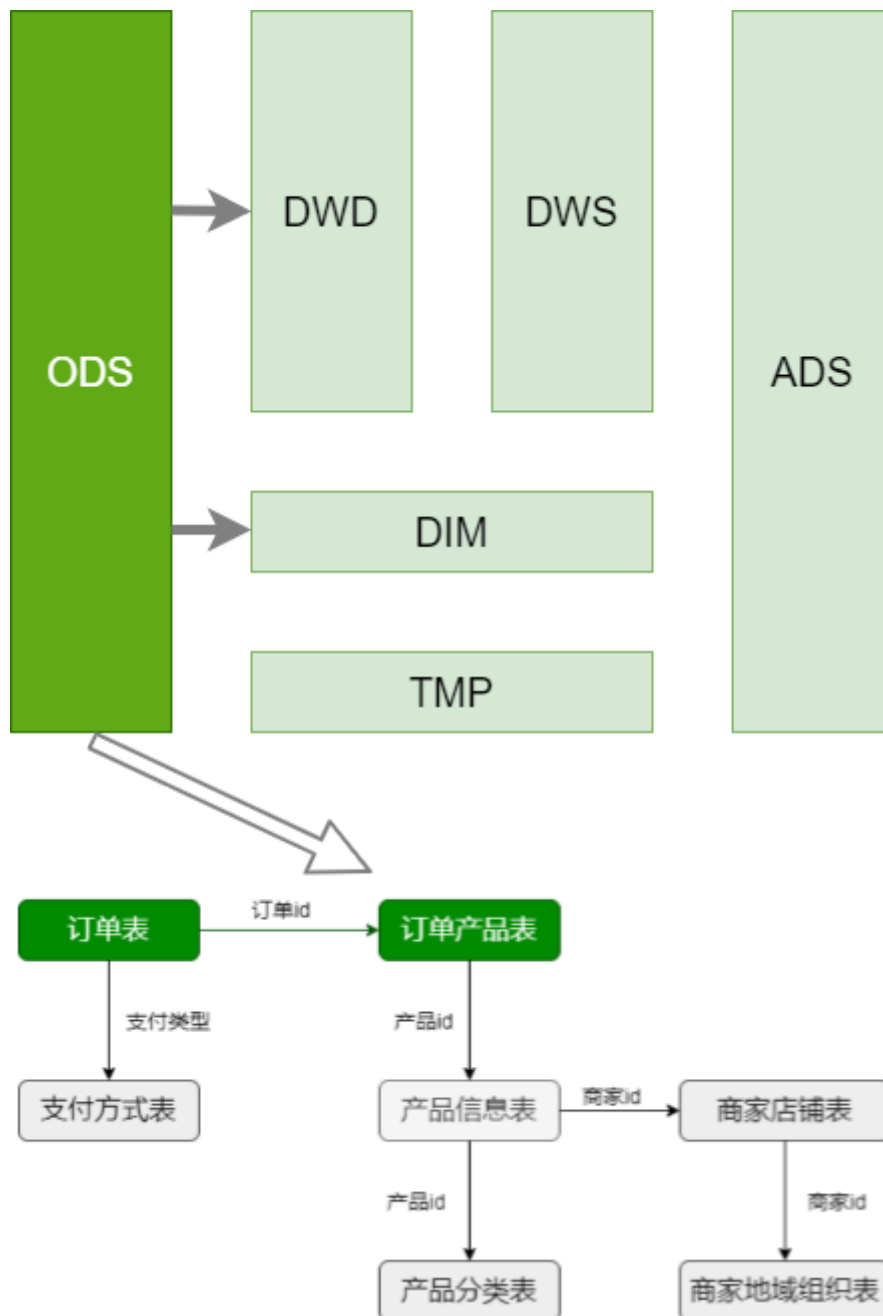
27         left join (select * from ods_orders where dt='2020-06-
28 21') B
29         on A.orderid=B.orderid;
30 -- 用以上信息覆写拉链表
31 insert overwrite table test.dwd_orders
32 SELECT orderid,
33         createtime,
34         modifiedtime,
35         status,
36         modifiedtime AS start_date,
37         '9999-12-31' AS end_date
38 FROM test.ods_orders
39 where dt='2020-06-21'
40
41 union all
42
43 select A.orderid,
44        A.createtime,
45        A.modifiedtime,
46        A.status,
47        A.start_date,
48        case when B.orderid is not null and A.end_date>'2020-
49 06-21'
50             then '2020-06-20'
51             else A.end_date
52        end end_date
53 from dwd_orders A
54        left join (select * from ods_orders where dt='2020-06-
55 21') B
56        on A.orderid=B.orderid;

```

5.4、拉链表小结



第6节 DIM层建表加载数据



首先要确定哪些是事实表、哪些是维表。绿色的是事实表，灰色的维表

用什么方式处理维表，每日快照、拉链表？

小表使用每日快照：产品分类表、商家店铺表、商家地域组织表、支付方式表

大表使用拉链表：产品信息表

6.1 商品分类表

数据库中的数据是规范的（满足三范式），但是规范化的数据给查询带来不便。

备注：这里对商品分类维度表做了逆规范化

省略了无关信息，做成了宽表

```
1 DROP TABLE IF EXISTS dim.dim_trade_product_cat;
2 create table if not exists dim.dim_trade_product_cat(
3     firstId int,                -- 一级商品分类id
4     firstName string,          -- 一级商品分类名称
5     secondId int,              -- 二级商品分类Id
6     secondName string,         -- 二级商品分类名称
7     thirdId int,               -- 三级商品分类id
8     thirdName string           -- 三级商品分类名称
9 )
10 partitioned by (dt string)
11 STORED AS PARQUET;
```

实现：

```
1 select T1.catid, T1.catname, T2.catid, T2.catname, T3.catid,
   T3.catname
2   from (select catid, catname, parentid
3         from ods.ods_trade_product_category
4         where level=3 and dt='2020-07-01') T3
5
6   left join
7
8   (select catid, catname, parentid
9     from ods.ods_trade_product_category
10    where level=2 and dt='2020-07-01') T2
11  on T3.parentid=T2.catid
12
13  left join
14
15  (select catid, catname, parentid
16    from ods.ods_trade_product_category
17   where level=1 and dt='2020-07-01') T1
18  on T2.parentid=T1.catid;
```

数据加载：

/data/lagoudw/script/trade/dim_load_product_cat.sh

```
1 #! /bin/bash
2
```

```
3 source /etc/profile
4
5 if [ -n "$1" ]
6 then
7     do_date=$1
8 else
9     do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 insert overwrite table dim.dim_trade_product_cat
14 partition(dt='$do_date')
15 select
16     t1.catid,          -- 一级分类id
17     t1.catname,        -- 一级分类名称
18     t2.catid,          -- 二级分类id
19     t2.catname,        -- 二级分类名称
20     t3.catid,          -- 三级分类id
21     t3.catname         -- 三级分类名称
22 from
23     -- 商品三级分类数据
24     (select catid, catname, parentid
25      from ods.ods_trade_product_category
26      where level=3 and dt='$do_date') t3
27
28 left join
29     -- 商品二级分类数据
30     (select catid, catname, parentid
31      from ods.ods_trade_product_category
32      where level=2 and dt='$do_date') t2
33 on t3.parentid = t2.catid
34
35 left join
36     -- 商品一级分类数据
37     (select catid, catname, parentid
38      from ods.ods_trade_product_category
39      where level=1 and dt='$do_date') t1
40 on t2.parentid = t1.catid;
41 "
42
43 hive -e "$sql"
```

6.2 商品地域组织表

商家店铺表、商家地域组织表 => 一张维表

这里也是逆规范化的设计，将商家店铺表、商家地域组织表组织成一张表，并拉宽。

在一行数据中体现：商家信息、城市信息、地域信息。信息中包括 id 和 name；

```
1 drop table if exists dim.dim_trade_shops_org;
2 create table dim.dim_trade_shops_org(
3   shopid int,
4   shopName string,
5   cityId int,
6   cityName string ,
7   regionId int ,
8   regionName string
9 )
10 partitioned by (dt string)
11 STORED AS PARQUET;
```

实现

```
1 select T1.shopid, T1.shopname, T2.id cityid, T2.orgname
   cityname, T3.id regionid, T3.orgname regionname
2   from
3   (select shopid, shopname, areaid
4     from ods.ods_trade_shops
5     where dt='2020-07-01') T1
6
7 left join
8
9   (select id, parentid, orgname, orglevel
10     from ods.ods_trade_shop_admin_org
11     where orglevel=2 and dt='2020-07-01') T2
12 on T1.areaId=T2.id
13
14 left join
15
16   (select id, orgname, orglevel
17     from ods.ods_trade_shop_admin_org
18     where orglevel=1 and dt='2020-07-01') T3
19 on T2.parentid=T3.id
20 limit 10;
```

/data/lagoudw/script/trade/dim_load_shop_org.sh

```
1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 insert overwrite table dim.dim_trade_shops_org
14 partition(dt='$do_date')
15 select t1.shopid,
16         t1.shopname,
17         t2.id as cityid,
18         t2.orgname as cityName,
19         t3.id as region_id,
20         t3.orgname as region_name
21 from (select shopId, shopName, areaId
22       from ods.ods_trade_shops
23       where dt='$do_date') t1
24
25 left join
26 (select id, parentId, orgname, orglevel
27  from ods.ods_trade_shop_admin_org
28  where orglevel=2 and dt='$do_date') t2
29 on t1.areaId = t2.id
30
31 left join
32 (select id, parentId, orgname, orglevel
33  from ods.ods_trade_shop_admin_org
34  where orglevel=1 and dt='$do_date') t3
35 on t2.parentid = t3.id;
36 "
37
38 hive -e "$sql"
```

6.3 支付方式表

对ODS中表的信息做了裁剪，只保留了必要的信息。

```
1 drop table if exists dim.dim_trade_payment;
2 create table if not exists dim.dim_trade_payment(
3     paymentId string,          -- 支付方式id
4     paymentName string        -- 支付方式名称
5 )
6 partitioned by (dt string)
7 STORED AS PARQUET;
```

/data/lagoudw/script/trade/dim_load_payment.sh

```
1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 insert overwrite table dim.dim_trade_payment
14 partition(dt='$do_date')
15 select id, payName
16     from ods.ods_trade_payments
17     where dt='$do_date';
18 "
19
20 hive -e "$sql"
```

6.4 商品信息表

使用拉链表对商品信息进行处理。

1、历史数据 => 初始化拉链表(开始日期：当日；结束日期：9999-12-31)【只执行一次】

2、拉链表的每日处理【每次加载数据时处理】

- 新增数据。每日新增数据(ODS) => 开始日期：当日；结束日期：9999-12-31
- 历史数据。拉链表(DIM) 与 每日新增数据(ODS) 做左连接
 - 连接上数据。数据有变化，结束日期：当日；
 - 未连接上数据。数据无变化，结束日期保持不变；

1、创建维表

拉链表要增加两列，分别记录生效日期和失效日期

```
1 drop table if exists dim.dim_trade_product_info;
2 create table dim.dim_trade_product_info(
3     `productId` bigint,
4     `productName` string,
5     `shopId` string,
6     `price` decimal,
7     `issale` tinyint,
8     `status` tinyint,
9     `categoryId` string,
10    `createTime` string,
11    `modifyTime` string,
12    `start_dt` string,
13    `end_dt` string
14 ) COMMENT '产品表'
15 STORED AS PARQUET;
```

2、初始数据加载（历史数据加载，只做一次）

```
1 insert overwrite table dim.dim_trade_product_info
2 select productId,
3     productName,
4     shopId,
5     price,
6     issale,
7     status,
8     categoryId,
9     createTime,
10    modifyTime,
11    -- modifyTime非空取modifyTime，否则取createTime; substr取
    日期
12    case when modifyTime is not null
```

```

13         then substr(modifyTime, 0, 10)
14         else substr(createTime, 0, 10)
15     end as start_dt,
16     '9999-12-31' as end_dt
17 from ods.ods_trade_product_info
18 where dt = '2020-07-12';

```

3、增量数据导入（重复执行，每次加载数据执行）

/data/lagoudw/script/trade/dim_load_product_info.sh

```

1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 insert overwrite table dim.dim_trade_product_info
14 select productId,
15         productName,
16         shopId,
17         price,
18         issale,
19         status,
20         categoryId,
21         createTime,
22         modifyTime,
23         case when modifyTime is not null
24             then substr(modifyTime,0,10)
25             else substr(createTime,0,10)
26         end as start_dt,
27         '9999-12-31' as end_dt
28     from ods.ods_trade_product_info
29     where dt='$do_date'
30
31 union all
32
33 select dim.productId,

```

```

34         dim.productName,
35         dim.shopId,
36         dim.price,
37         dim.isSale,
38         dim.status,
39         dim.categoryId,
40         dim.createTime,
41         dim.modifyTime,
42         dim.start_dt,
43         case when dim.end_dt >= '9999-12-31' and ods.productId
is not null
44             then '$do_date'
45             else dim.end_dt
46         end as end_dt
47     from dim.dim_trade_product_info dim left join
48         (select *
49          from ods.ods_trade_product_info
50          where dt='$do_date' ) ods
51     on dim.productId = ods.productId
52 "
53
54 hive -e "$sql"

```

第7节 DWD层建表加载数据

要处理的表有两张：订单表、订单产品表。其中：

- 订单表是周期性事实表；为保留订单状态，可以使用拉链表进行处理；
- 订单产品表普通的事实表，用常规的方法进行处理；
 - 如果有数据清洗、数据转换的业务需求，ODS => DWD
 - 如果没有数据清洗、数据转换的业务需求，保留在ODS，不做任何变化。这个是本项目的处理方式

订单状态：

- -3：用户拒收
- -2：未付款的订单
- -1：用户取消
- 0：待发货
- 1：配送中
- 2：用户确认收货

订单从创建到最终完成，是有时间限制的；业务上也不允许订单在一个月之后，状态仍然在发生变化；

7.1、DWD层建表

备注：

- 与维表不同，订单事实表的记录数非常多
- 订单有生命周期；订单的状态不可能永远处于变化之中（订单的生命周期一般在15天左右）
- 订单是一个拉链表，而且是分区表
- 分区的目的：订单一旦终止，不会重复计算
- 分区的条件：订单创建日期；保证相同的订单在用同一个分区

```
1  -- 订单事实表(拉链表)
2  DROP TABLE IF EXISTS dwd.dwd_trade_orders;
3  create table dwd.dwd_trade_orders(
4      `orderId`      int,
5      `orderNo`      string,
6      `userId`       bigint,
7      `status`       tinyint,
8      `productMoney` decimal,
9      `totalMoney`   decimal,
10     `payMethod`     tinyint,
11     `isPay`         tinyint,
12     `areaId`        int,
13     `tradeSrc`      tinyint,
14     `tradeType`     int,
15     `isRefund`      tinyint,
16     `dataFlag`      tinyint,
17     `createTime`    string,
18     `payTime`       string,
19     `modifiedTime`  string,
20     `start_date`    string,
21     `end_date`      string
22 ) COMMENT '订单事实拉链表'
23 partitioned by (dt string)
24 STORED AS PARQUET;
```

7.2、DWD层数据加载

```
1  -- 备注：时间日期格式转换
2  -- 'yyyy-MM-dd HH:mm:ss' => timestamp => 'yyyy-MM-dd'
3  select unix_timestamp(modifiedtime, 'yyyy-MM-dd HH:mm:ss')
4      from ods.ods_trade_orders limit 10;
5
6  select from_unixtime(unix_timestamp(modifiedtime, 'yyyy-MM-dd
   HH:mm:ss'), 'yyyy-MM-dd')
7      from ods.ods_trade_orders limit 10;
```

/data/lagoudw/script/trade/dwd_load_trade_orders.sh

```
1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 set hive.exec.dynamic.partition.mode=nonstrict;
14 set hive.exec.dynamic.partition=true;
15 INSERT OVERWRITE TABLE dwd.dwd_trade_orders
16 partition(dt)
17 SELECT orderId,
18         orderNo,
19         userId,
20         status,
21         productMoney,
22         totalMoney,
23         payMethod,
24         isPay,
25         areaId,
26         tradeSrc,
27         tradeType,
28         isRefund,
29         dataFlag,
30         createTime,
```

```

31         payTime,
32         modifiedTime,
33         case when modifiedTime is not null
34             then from_unixtime(unix_timestamp(modifiedTime,
35 'yyyy-MM-dd HH:mm:ss'), 'yyyy-MM-dd')
36             else from_unixtime(unix_timestamp(createTime,
37 'yyyy-MM-dd HH:mm:ss'), 'yyyy-MM-dd')
38         end as start_date,
39         '9999-12-31' as end_date,
40         from_unixtime(unix_timestamp(createTime, 'yyyy-MM-dd
41 HH:mm:ss'), 'yyyy-MM-dd') as dt
42 FROM ods.ods_trade_orders
43 WHERE dt='$do_date'
44
45 union all
46
47 SELECT A.orderId,
48         A.orderNo,
49         A.userId,
50         A.status,
51         A.productMoney,
52         A.totalMoney,
53         A.payMethod,
54         A.isPay,
55         A.areaId,
56         A.tradeSrc,
57         A.tradeType,
58         A.isRefund,
59         A.dataFlag,
60         A.createTime,
61         A.payTime,
62         A.modifiedTime,
63         A.start_date,
64         CASE WHEN B.orderid IS NOT NULL AND A.end_date >
65 '$do_date'
66             THEN date_add('$do_date', -1)
67             ELSE A.end_date END AS end_date,
68         from_unixtime(unix_timestamp(A.createTime, 'yyyy-MM-dd
69 HH:mm:ss'), 'yyyy-MM-dd') as dt
70 FROM (SELECT * FROM dwd.dwd_trade_orders WHERE
71 dt>date_add('$do_date', -15)) A
72 left outer join (SELECT * FROM ods.ods_trade_orders
73 WHERE dt='$do_date') B
74 ON A.orderId = B.orderId;
75
76 "

```

第8节 DWS层建表及数据加载

DIM、DWD => 数据仓库分层、数据仓库理论

需求：计算当天

- 全国所有订单信息
- 全国、一级商品分类订单信息
- 全国、二级商品分类订单信息
- 大区所有订单信息
- 大区、一级商品分类订单信息
- 大区、二级商品分类订单信息
- 城市所有订单信息
- 城市、一级商品分类订单信息
- 城市、二级商品分类订单信息

需要的信息：订单表、订单商品表、商品信息维表、商品分类维表、商家地域维表

订单表 => 订单id、订单状态

订单商品表 => 订单id、商品id、商家id、单价、数量

商品信息维表 => 商品id、三级分类id

商品分类维表 => 一级名称、一级分类id、二级名称、二级分类id、三级名称、三级分类id

商家地域维表 => 商家id、区域名称、区域id、城市名称、城市id

订单表、订单商品表、商品信息维表 => 订单id、商品id、商家id、三级分类id、单价、数量（订单明细表）

订单明细表、商品分类维表、商家地域维表 => 订单id、商品id、商家id、三级分类名称、三级分类名称、三级分类名称、单价、数量、区域、城市 => 订单明细宽表

8.1、DWS层建表

dws_trade_orders（订单明细）由以下表轻微聚合而成：

- dwd.dwd_trade_orders (拉链表、分区表)
- ods.ods_trade_order_product（分区表）
- dim.dim_trade_product_info（维表、拉链表）

dws_trade_orders_w（订单明细宽表）由以下表组成：

- ads.dws_trade_orders (分区表)
- dim.dim_trade_product_cat（分区表）
- dim.dim_trade_shops_org（分区表）

```
1  -- 订单明细表(轻度汇总事实表)。每笔订单的明细
2  DROP TABLE IF EXISTS dws.dws_trade_orders;
3  create table if not exists dws.dws_trade_orders(
4     orderid      string,          -- 订单id
5      cat_3rd_id  string,          -- 商品三级分类id
6      shopid      string,          -- 店铺id
7      paymethod   tinyint,         -- 支付方式
8      productsnum bigint,          -- 商品数量
9      paymoney    double,          -- 订单商品明细金额
10     paytime     string            -- 订单时间
11 )
12 partitioned by (dt string)
13 STORED AS PARQUET;
14
15 -- 订单明细表宽表
16 DROP TABLE IF EXISTS dws.dws_trade_orders_w;
17 create table if not exists dws.dws_trade_orders_w(
18    orderid string,                -- 订单id
19     cat_3rd_id string,            -- 商品三级分类id
20     thirdname string,             -- 商品三级分类名称
21     secondname string,           -- 商品二级分类名称
22     firstname  string,           -- 商品一级分类名称
23     shopid     string,           -- 店铺id
24     shopname   string,           -- 店铺名
25     regionname string,           -- 店铺所在大区
26     cityname   string,           -- 店铺所在城市
27     paymethod  tinyint,          -- 支付方式
28     productsnum bigint,          -- 商品数量
```



```

29     paymoney double,          -- 订单明细金额
30     paytime string           -- 订单时间
31 )
32 partitioned by (dt string)
33 STORED AS PARQUET;

```

8.2、DWS层加载数据

/data/lagoudw/script/trade/dws_load_trade_orders.sh

备注：dws_trade_orders/dws_trade_orders_w 中一笔订单可能出现多条记录！

```

1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 insert overwrite table dws.dws_trade_orders
14 partition(dt='${do_date}')
15 select t1.orderid    as orderid,
16         t3.categoryid as cat_3rd_id,
17         t3.shopid     as shopid,
18         t1.paymethod  as paymethod,
19         t2.productnum as productsnum,
20         t2.productnum*t2.productprice as pay_money,
21         t1.paytime    as paytime
22 from (select orderid, paymethod, paytime
23       from dwd.dwd_trade_orders
24       where dt='${do_date}') T1
25
26 left join
27
28 (select orderid, productid, productnum, productprice
29   from ods.ods_trade_order_product
30   where dt='${do_date}') T2
31 on t1.orderid = t2.orderid
32

```

```

33     left join
34
35     (select productid, shopid, categoryid
36         from dim.dim_trade_product_info
37         where start_dt <= '$do_date'
38             and end_dt >= '$do_date' ) T3
39     on t2.productid=t3.productid;
40
41 insert overwrite table dws.dws_trade_orders_w
42 partition(dt='$do_date')
43 select t1.orderid,
44         t1.cat_3rd_id,
45         t2.thirdname,
46         t2.secondname,
47         t2.firstname,
48         t1.shopid,
49         t3.shopname,
50         t3.regionname,
51         t3.cityname,
52         t1.paymethod,
53         t1.productsnum,
54         t1.paymoney,
55         t1.paytime
56     from (select orderid,
57                 cat_3rd_id,
58                 shopid,
59                 paymethod,
60                 productsnum,
61                 paymoney,
62                 paytime
63             from dws.dws_trade_orders
64             where dt='$do_date') T1
65
66     join
67
68     (select thirdid, thirdname, secondid, secondname,
69         firstid, firstname
70         from dim.dim_trade_product_cat
71         where dt='$do_date') T2
72     on T1.cat_3rd_id = T2.thirdid
73
74     join
75
76     (select shopid, shopname, regionname, cityname
77         from dim.dim_trade_shops_org

```

```
77         where dt='$do_date') T3
78         on T1.shopid = T3.shopid
79     "
80
81     hive -e "$sql"
```

备注：要自己准备测试数据！

- dwd.dwd_trade_orders (拉链表、分区表)
- ods.ods_trade_order_product (分区表)
- dim.dim_trade_product_info (维表、拉链表)
- dim.dim_trade_product_cat (分区表)
- dim.dim_trade_shops_org (分区表)

保证测试的日期有数据。

构造测试数据（拉链表分区表）：

```
1  insert overwrite table dwd.dwd_trade_orders
2  partition(dt='2020-07-12')
3  select
4  orderid,
5  orderno,
6  userid,
7  status,
8  productmoney,
9  totalmoney,
10 paymethod,
11 ispay,
12 areaid,
13 tradesrc,
14 tradetype,
15 isrefund,
16 dataflag,
17 '2020-07-12',
18 paytime,
19 modifiedtime,
20 start_date,
21 end_date
22 from dwd.dwd_trade_orders
23 where end_date='9999-12-31';
```

第9节 ADS层开发

需求：计算当天

- 全国所有订单信息
- 全国、一级商品分类订单信息
- 全国、二级商品分类订单信息
- 大区所有订单信息
- 大区、一级商品分类订单信息
- 大区、二级商品分类订单信息
- 城市所有订单信息
- 城市、一级商品分类订单信息
- 城市、二级商品分类订单信息

用到的表：

- dws.dws_trade_orders_w

9.1、ADS层建表

```
1  -- ADS层订单分析表
2  DROP TABLE IF EXISTS ads.ads_trade_order_analysis;
3  create table if not exists ads.ads_trade_order_analysis(
4      areatype string,                -- 区域范围：区域类型（全国、大
    区、城市）
5      regionname string,             -- 区域名称
6      cityname string,               -- 城市名称
7      categorytype string,           -- 商品分类类型（一级、二级）
8      category1 string,              -- 商品一级分类名称
9      category2 string,              -- 商品二级分类名称
10     totalcount bigint,              -- 订单数量
11     total_productnum bigint,        -- 商品数量
12     totalmoney double               -- 支付金额
13 )
14 partitioned by (dt string)
15 row format delimited fields terminated by ',';
```

9.2、ADS层加载数据

/data/lagoudw/script/trade/ads_load_trade_order_analysis.sh

备注：1笔订单，有多个商品；多个商品有不同的分类；这会导致一笔订单有多个分类，它们是分别统计的；

```
1  #!/bin/bash
2
3  source /etc/profile
4
5  if [ -n "$1" ]
6  then
7      do_date=$1
8  else
9      do_date=`date -d "-1 day" +%F`
10 fi
11
12 sql="
13 with mid_orders as (
14 select regionname,
15         cityname,
16         firstname category1,
17         secondname category2,
18         count(distinct orderid) as totalcount,
19         sum(productsnum) as total_productnum,
20         sum(paymoney) as totalmoney
21 from dws.dws_trade_orders_w
22 where dt='$do_date'
23 group by regionname, cityname, firstname, secondname
24 )
25 insert overwrite table ads.ads_trade_order_analysis
26 partition(dt='$do_date')
27 select '全国' as areatype,
28        '' as regionname,
29        '' as cityname,
30        '' as categorytype,
31        '' as category1,
32        '' as category2,
33        sum(totalcount),
34        sum(total_productnum),
35        sum(totalmoney)
36 from mid_orders
37
38 union all
```

```
39
40 select '全国' as areatype,
41        '' as regionname,
42        '' as cityname,
43        '一级' as categorytype,
44        category1,
45        '' as category2,
46        sum(totalcount),
47        sum(total_productnum),
48        sum(totalmoney)
49    from mid_orders
50 group by category1
51
52 union all
53
54 select '全国' as areatype,
55        '' as regionname,
56        '' as cityname,
57        '二级' as categorytype,
58        '' as category1,
59        category2,
60        sum(totalcount),
61        sum(total_productnum),
62        sum(totalmoney)
63    from mid_orders
64 group by category2
65
66 union all
67 select '大区' as areatype,
68        regionname,
69        '' as cityname,
70        '' as categorytype,
71        '' as category1,
72        '' as category2,
73        sum(totalcount),
74        sum(total_productnum),
75        sum(totalmoney)
76    from mid_orders
77 group by regionname
78
79 union all
80
81 select '大区' as areatype,
82        regionname,
83        '' as cityname,
```

```
84         '一级' as categorytype,
85         category1,
86         '' as category2,
87         sum(totalcount),
88         sum(total_productnum),
89         sum(totalmoney)
90     from mid_orders
91 group by regionname, category1
92
93 union all
94
95 select '大区' as areatype,
96        regionname,
97        '' as cityname,
98        '二级' as categorytype,
99        '' as category1,
100       category2,
101       sum(totalcount),
102       sum(total_productnum),
103       sum(totalmoney)
104     from mid_orders
105 group by regionname, category2
106
107 union all
108
109 select '城市' as areatype,
110        '' as regionname,
111        cityname,
112        '' as categorytype,
113        '' as category1,
114        '' as category2,
115        sum(totalcount),
116        sum(total_productnum),
117        sum(totalmoney)
118     from mid_orders
119 group by cityname
120
121 union all
122
123 select '城市' as areatype,
124        '' as regionname,
125        cityname,
126        '一级' as categorytype,
127        category1,
128        '' as category2,
```

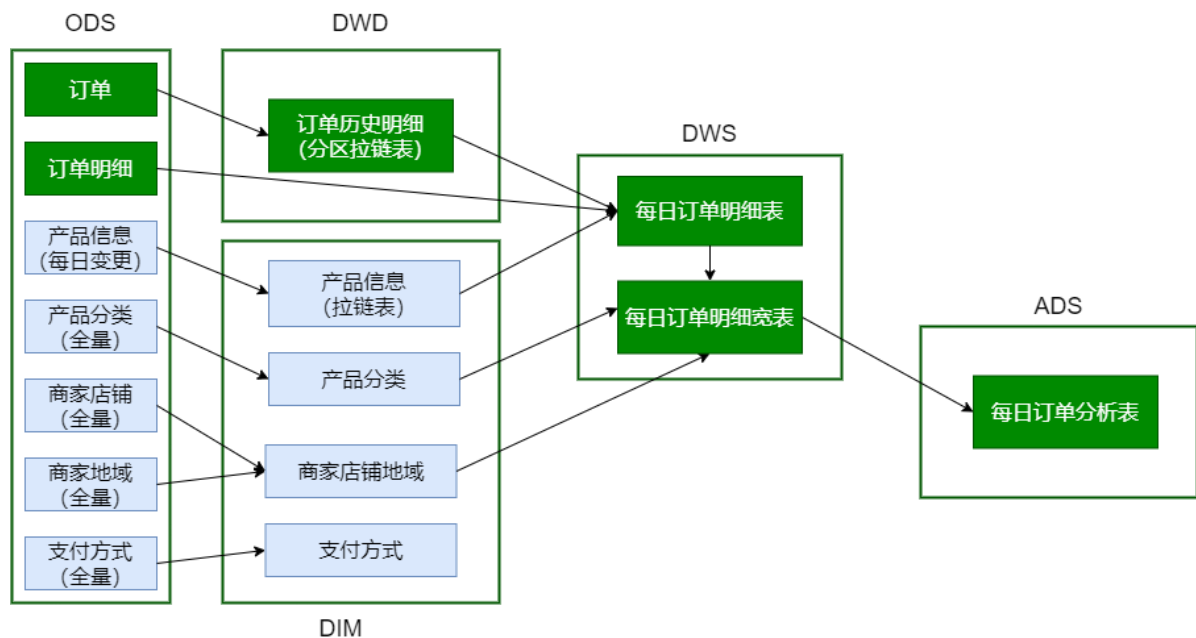
```
129         sum(totalcount),
130         sum(total_productnum),
131         sum(totalmoney)
132     from mid_orders
133 group by cityname, category1
134
135 union all
136
137 select '城市' as areatype,
138        '' as regionname,
139        cityname,
140        '二级' as categorytype,
141        '' as category1,
142        category2,
143        sum(totalcount),
144        sum(total_productnum),
145        sum(totalmoney)
146     from mid_orders
147 group by cityname, category2;
148 "
149
150 hive -e "$sql"
```

备注：由于在dws.dws_trade_orders_w中，一笔订单可能有多条记录，所以在统计订单数量的时候要用count(distinct orderid)

第10节 数据导出

ads.ads_trade_order_analysis 分区表，使用DataX导出到MySQL

第11节 小结



脚本调用次序：

```

1  # 加载ODS数据（含DataX迁移数据）
2  /data/lagoudw/script/trade/ods_load_trade.sh
3
4  # 加载DIM层数据
5  /data/lagoudw/script/trade/dim_load_product_cat.sh
6  /data/lagoudw/script/trade/dim_load_shop_org.sh
7  /data/lagoudw/script/trade/dim_load_payment.sh
8  /data/lagoudw/script/trade/dim_load_product_info.sh
9
10 # 加载DWD层数据
11 /data/lagoudw/script/trade/dwd_load_trade_orders.sh
12
13 # 加载DWS层数据
14 /data/lagoudw/script/trade/dws_load_trade_orders.sh
15
16 # 加载ADS层数据
17 /data/lagoudw/script/trade/ads_load_trade_order_analysis.sh
  
```

主要技术点：

- 拉链表。创建、使用与回滚；商品信息表、订单表（周期性事实表；分区表+拉链表）
- 宽表（逆规范化）：商品分类表、商品地域组织表、订单明细及订单明细宽表（轻度汇总的事实表）

第二部分 任务调度系统Airflow

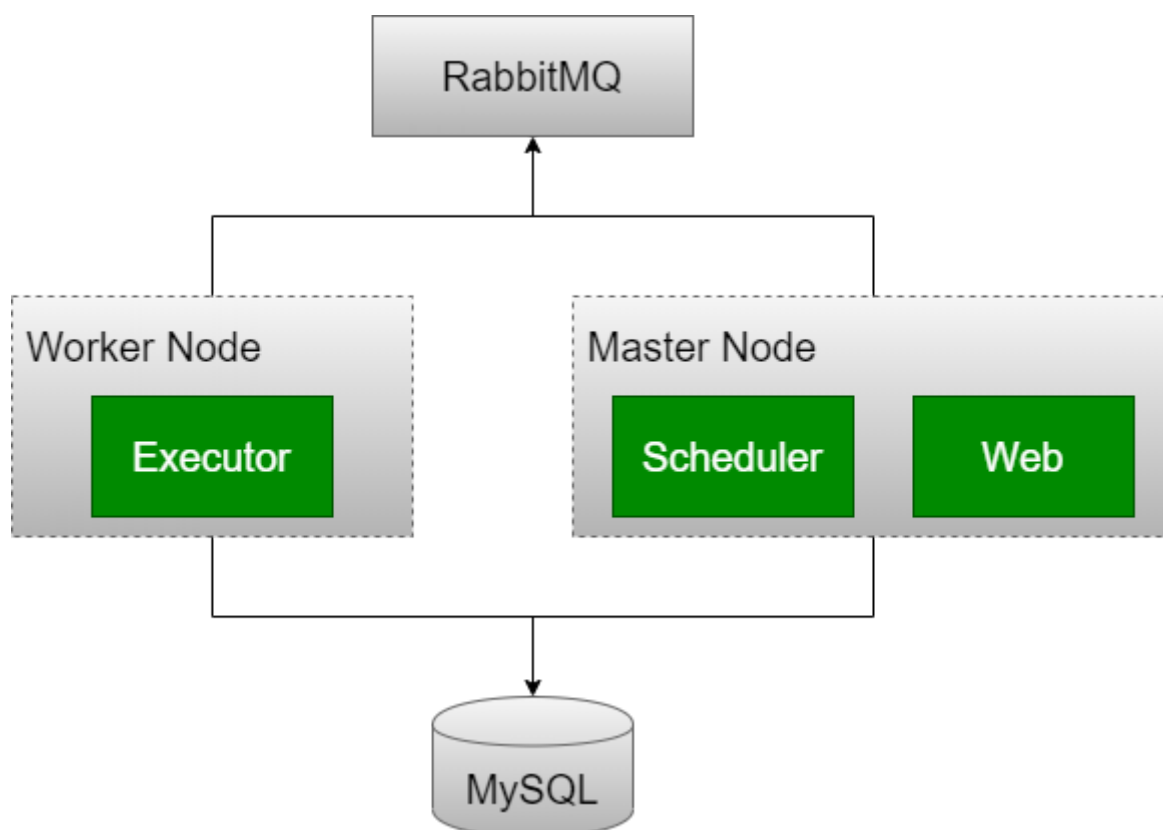
第1节 Airflow简介

Airflow 是 Airbnb 开源的一个用 Python 编写的调度工具。于 2014 年启动，2015 年春季开源，2016 年加入 Apache 软件基金会的孵化计划。

Airflow 将一个工作流制定为一组任务的有向无环图（DAG），并指派到一组计算节点上，根据相互之间的依赖关系，有序执行。Airflow 有以下优势：

- 灵活易用。Airflow 是 Python 编写的，工作流的定义也使用 Python 编写；
- 功能强大。支持多种不同类型的作业，可自定义不同类型的作业。如 Shell、Python、Mysql、Oracle、Hive 等；
- 简洁优雅。作业的定义简单明了；
- 易扩展。提供各种基类供扩展，有多种执行器可供选择；

1.1、体系架构



Webserver 守护进程。接受 HTTP 请求，通过 Python Flask Web 应用程序与 airflow 进行交互。Webserver 提供功能的功能包括：中止、恢复、触发任务；监控正在运行的任务，断点续跑任务；查询任务的状态，日志等详细信息。

Scheduler 守护进程。周期性地轮询任务的调度计划，以确定是否触发任务执行。

Worker 守护进程。Worker负责启动机器上的executor来执行任务。使用celeryExecutor后可以在多个机器上部署worker服务。

1.2、重要概念

DAG (Directed Acyclic Graph) 有向无环图

- 在Airflow中，一个DAG定义了一个完整的作业。同一个DAG中的所有Task拥有相同的调度时间。
- 参数：
 - dag_id: 唯一识别DAG
 - default_args: 默认参数，如果当前DAG实例的作业没有配置相应参数，则采用DAG实例的default_args中的相应参数
 - schedule_interval: 配置DAG的执行周期，可采用crontab语法

Task

- Task为DAG中具体的作业任务，依赖于DAG，必须存在于某个DAG中。Task在DAG中可以配置依赖关系
- 参数：
 - dag: 当前作业属于相应DAG
 - task_id: 任务标识符
 - owner: 任务的拥有者
 - start_date: 任务的开始时间

第2节 Airflow安装部署

2.1、安装依赖

- CentOS 7.X
- Python 3.5或以上版本（推荐）
- MySQL 5.7.x
- Apache-Airflow 1.10.11
- 虚拟机可上网，需在线安装包

备注：后面要安装三个软件Airflow、Atlas、Griffin，相对Hadoop的安装都较为复杂

- 正式安装软件之前给虚拟机做一个快照
- 按照讲义中指定的软件安装
- 按照讲义的步骤执行对应的命令，命令的遗漏会对后面的安装造成影响

Airflow 视频录制的时候选择的是最新版本，现在Airflow版本升级了，讲义中的安装步骤适合Airflow 1.10.11。

请注意以下两条命令：

```
1  -- 下载的时候指定 airflow 的版本
2  pip install apache-airflow==1.10.11 -i
   https://pypi.douban.com/simple
3
4  -- 下载的时候指定 mysqlclient 的版本
5  pip install mysqlclient==1.4.6
```

2.2、Python环境准备

备注：提前下载 Python-3.6.6.tgz

备注：使用linux122安装

```
1  # 卸载 mariadb
2  rpm -qa | grep mariadb
3  mariadb-libs-5.5.65-1.el7.x86_64
4  mariadb-5.5.65-1.el7.x86_64
5  mariadb-devel-5.5.65-1.el7.x86_64
6
7  yum remove mariadb
8  yum remove mariadb-libs
9
10 # 安装依赖
11 rpm -ivh mysql57-community-release-el7-11.noarch.rpm
12
13 yum install readline readline-devel -y
14 yum install gcc -y
15 yum install zlib* -y
16 yum install openssl openssl-devel -y
17 yum install sqlite-devel -y
18 yum install python-devel mysql-devel -y
```

```
19
20 # 提前到python官网下载好包
21 cd /opt/lagou/software
22 tar -zxvf Python-3.6.6.tgz
23
24 # 安装 python3 运行环境
25 cd Python-3.6.6/
26 # configure文件是一个可执行的脚本文件。如果配置了--prefix，安装后的所有
    资源文件都会放在目录中
27 ./configure --prefix=/usr/local/python3.6
28 make && make install
29 /usr/local/python3.6/bin/pip3 install virtualenv
30
31 # 启动 python3 环境
32 cd /usr/local/python3.6/bin/
33 ./virtualenv env
34 . env/bin/activate
35
36 # 检查 python 版本
37 python -V
```

2.3、安装Airflow

```
1 # 设置目录（配置文件）
2 # 添加到配置文件/etc/profile。未设置是缺省值为 ~/airflow
3 export AIRFLOW_HOME=/opt/lagou/servers/airflow
4
5 # 使用豆瓣源非常快。-i：指定库的安装源（可选选项）
6 pip install apache-airflow==1.10.11 -i
    https://pypi.douban.com/simple
```

备注：

- apache-airflow==1.10.11，需要指定安装的版本，重要！！
- 软件安装路径在\$AIRFLOW_HOME（缺省为~/airflow），此时目录不存在
- 安装的是版本是1.10.11，不指定下载源时下载过程非常慢

2.4、创建数据库用户并授权

```
1  -- 创建数据库
2  create database airflowlinux122;
3
4  -- 创建用户airflow, 设置所有ip均可以访问
5  create user 'airflow'@'%' identified by '12345678';
6  create user 'airflow'@'localhost' identified by '12345678';
7
8  -- 用户授权, 为新建的airflow用户授予Airflow库的所有权限
9  grant all on airflowlinux122.* to 'airflow'@'%';
10 SET GLOBAL explicit_defaults_for_timestamp = 1;
11 flush privileges;
```

2.5、修改Airflow DB配置

```
1  # python3 环境中执行
2  pip install mysqlclient==1.4.6
3  airflow initdb
```

备注:

- mysqlclient==1.4.6, 需要指定安装的版本, 重要!!!
- 有可能在安装完Airflow找不到 \$AIRFLOW_HOME/airflow.cfg 文件, 执行完 airflow initdb才会在对应的位置找到该文件。

勘误开始 (2021年5月)

在执行 `airflow initdb` 命令时, 如遇上如下报错:

```
1  ModuleNotFoundError: No module named
   'sqlalchemy.ext.declarative.clsregistry'
```

这是由于 SQLAlchemy 模块版本低导致的错误。执行以下命令后, 重新执行 `airflow initdb` 命令。

```
1  pip install SQLAlchemy==1.3.23
```

勘误结束 (2021年5月)

修改 \$AIRFLOW_HOME/airflow.cfg:

```
1 # 约 75 行
2 sql_alchemy_conn =
  mysql://airflow:12345678@linux123:3306/airflowlinux122
3
4 # 重新执行
5 airflow initdb
```

可能出现的错误: Exception: Global variable explicit_defaults_for_timestamp needs to be on (1) for mysql

解决方法:

```
1 SET GLOBAL explicit_defaults_for_timestamp = 1;
2 FLUSH PRIVILEGES;
```

2.6、安装密码模块

安装password组件:

```
1 pip install apache-airflow[password]
```

修改 airflow.cfg 配置文件 (第一行修改, 第二行增加) :

```
1 # 约 281 行
2 [webserver]
3 # 约 353行
4 authenticate = True
5 auth_backend = airflow.contrib.auth.backends.password_auth
```

- 添加密码文件

python命令, 执行一遍; 添加用户登录, 设置口令

```
1 import airflow
2 from airflow import models, settings
3 from airflow.contrib.auth.backends.password_auth import
  PasswordUser
4
```

```
5 user = PasswordUser(models.User())
6 user.username = 'airflow'
7 user.email = 'airflow@lagou.com'
8 user.password = 'airflow123'
9
10 session = settings.Session()
11 session.add(user)
12 session.commit()
13 session.close()
14 exit()
```

2.7、启动服务

```
1 # 备注：要先进入python3的运行环境
2 cd /usr/local/python3.6/bin/
3 ./virtualenv env
4 . env/bin/activate
5
6 # 退出虚拟环境命令
7 deactivate
8
9 # 启动scheduler调度器：
10 airflow scheduler -D
11
12 # 服务页面启动：
13 airflow webserver -D
```

备注：airflow命令所在位置：/usr/local/python3.6/bin/env/bin/airflow

安装完成，可以使用浏览器登录 linux122:8080；输入用户名、口令：airflow / airflow123

2.8、修改时区

Airflow默认使用UTC时间，在中国时区需要用+8小时。将UTC修改为中国时区，需要修改Airflow源码。

1、在修改 \$AIRFLOW_HOME/airflow.cfg 文件


```
1 # 约 65 行
2 default_timezone = Asia/Shanghai
```

2、修改 timezone.py

```
1 # 进入Airflow包的安装位置
2 cd /usr/local/python3.6/bin/env/lib/python3.6/site-packages/
3
4 # 修改airflow/utils/timezone.py
5 cd airflow/utils
6 vi timezone.py
```

第27行注释，增加29-37行：

```
1      27 utc = pendulum.timezone('UTC')
2      28
3      29 from airflow import configuration as conf
4      30 try:
5      31         tz = conf.get("core", "default_timezone")
6      32         if tz == "system":
7      33             utc = pendulum.local_timezone()
8      34         else:
9      35             utc = pendulum.timezone(tz)
10     36 except Exception:
11     37         pass
```

备注：以上的修改方式有警告，可以使用下面的方式（推荐）：

```

1      27 utc = pendulum.timezone('UTC')
2      28
3      29 from airflow import configuration
4      30 try:
5      31         tz = configuration.conf("core",
    "default_timezone")
6      32         if tz == "system":
7      33             utc = pendulum.local_timezone()
8      34         else:
9      35             utc = pendulum.timezone(tz)
10     36 except Exception:
11     37         pass

```

修改utcnow()函数 (注释掉72行, 增加73行内容)

```

1      62 def utcnow():
2      63         """
3      64         Get the current date and time in UTC
4      65
5      66         :return:
6      67         """
7      68
8      69         # pendulum utcnow() is not used as that sets a
    TimezoneInfo object
9      70         # instead of a Timezone. This is not pickable and
    also creates issues
10     71         # when using replace()
11     72         # d = dt.datetime.utcnow()
12     73         d = dt.datetime.now()
13     74         d = d.replace(tzinfo=utc)
14     75
15     76         return d

```

3、修改 airflow/utils/sqlalchemy.py

```

1  # 进入Airflow包的安装位置
2  cd /usr/local/python3.6/bin/env/lib/python3.6/site-packages/
3
4  # 修改 airflow/utils/sqlalchemy.py
5  cd airflow/utils
6  vi sqlalchemy.py

```

在38行之后增加 39 - 47 行的内容：

```
1      38 utc = pendulum.timezone('UTC')
2      39 from airflow import configuration as conf
3      40 try:
4      41         tz = conf.get("core", "default_timezone")
5      42         if tz == "system":
6      43             utc = pendulum.local_timezone()
7      44         else:
8      45             utc = pendulum.timezone(tz)
9      46 except Exception:
10     47     pass
```

备注：以上的修改方式有警告，可以使用下面的方式（推荐）：

```
1      38 utc = pendulum.timezone('UTC')
2      39 from airflow import configuration
3      40 try:
4      41         tz = configuration.conf("core",
    "default_timezone")
5      42         if tz == "system":
6      43             utc = pendulum.local_timezone()
7      44         else:
8      45             utc = pendulum.timezone(tz)
9      46 except Exception:
10     47     pass
```

4、修改airflow/www/templates/admin/master.html

```
1  # 进入Airflow包的安装位置
2  cd /usr/local/python3.6/bin/env/lib/python3.6/site-packages/
3
4  # 修改 airflow/www/templates/admin/master.html
5  cd airflow/www/templates/admin
6  vi master.html
```

```

1 # 将第40行修改为以下内容:
2     40         var UTCseconds = x.getTime();
3
4 # 将第43行修改为以下内容:
5     43         "timeFormat":"H:i:s",

```


重启airflow webserver

```

1 # 关闭 airflow webserver 对应的服务
2 ps -ef | grep 'airflow-webserver' | grep -v 'grep' | awk
  '{print $2}' | xargs -i kill -9 {}
3
4 # 关闭 airflow scheduler 对应的服务
5 ps -ef | grep 'airflow' | grep 'scheduler' | awk '{print $2}'
  | xargs -i kill -9 {}
6
7 # 删除对应的pid文件
8 cd $AIRFLOW_HOME
9 rm -rf *.pid
10
11 # 重启服务（在python3.6虚拟环境中执行）
12 airflow scheduler -D
13 airflow webserver -D

```

2.9、Airflow的web界面

<div>  <div> DAGs Data Profiling Browse Admin Docs </div> <div>2020-08-25 02:50:45 UTC</div> </div>							
DAGs							
<div>Search: <input type="text"/></div>							
	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	example_bash_operator	@ 0 * * *	Airflow				
	example_branch_dop_operator_v3	* * * * *	Airflow				
	example_branch_operator	@daily	Airflow				
	example_complex	None	airflow				
	example_external_task_marker_child	None	airflow				
	example_external_task_marker_parent	None	airflow				
	example_http_operator	1 day, 0:00:00	Airflow				

Trigger Dag: 人为执行触发

Tree View: 当dag执行的时候, 可以点入, 查看每个task的执行状态 (基于树状视图)。状态: success、running、failed、skipped、retry、queued、no status

Graph View: 基于图视图 (有向无环图), 查看每个task的执行状态

Tasks Duration: 每个task的执行时间统计, 可以选择最近多少次执行

Task Tries: 每个task的重试次数

Gantt View: 基于甘特图的视图, 每个task的执行状态

Code View: 查看任务执行代码

Logs: 查看执行日志, 比如失败原因

Refresh: 刷新dag任务

Delete Dag: 删除该dag任务

2.10、禁用自带的DAG任务

停止服务:

```
1 # 关闭 airflow webserver 对应的服务
2 ps -ef | grep 'airflow-webserver' | grep -v 'grep' | awk
  '{print $2}' | xargs -i kill -9 {}
3
4 # 关闭 airflow scheduler 对应的服务
5 ps -ef | grep 'airflow' | grep 'scheduler' | awk '{print $2}' |
  xargs -i kill -9 {}
6
7 # 删除对应的pid文件
8 cd $AIRFLOW_HOME
9 rm -rf *.pid
```

修改文件 \$AIRFLOW_HOME/airflow.cfg:

```
1 # 修改文件第 136 行
2     136 # load_examples = True
3     137 load_examples = False
4
5 # 重新设置db
6 airflow resetdb -y
```

重新设置账户、口令:

```
1 import airflow
```

```

2 from airflow import models, settings
3 from airflow.contrib.auth.backends.password_auth import
  PasswordUser
4
5 user = PasswordUser(models.User())
6 user.username = 'airflow'
7 user.email = 'airflow@lagou.com'
8 user.password = 'airflow123'
9
10 session = settings.Session()
11 session.add(user)
12 session.commit()
13 session.close()
14 exit()

```

重启服务

```

1 # 重启服务
2 airflow scheduler -D
3 airflow webserver -D

```

2.11、crontab简介

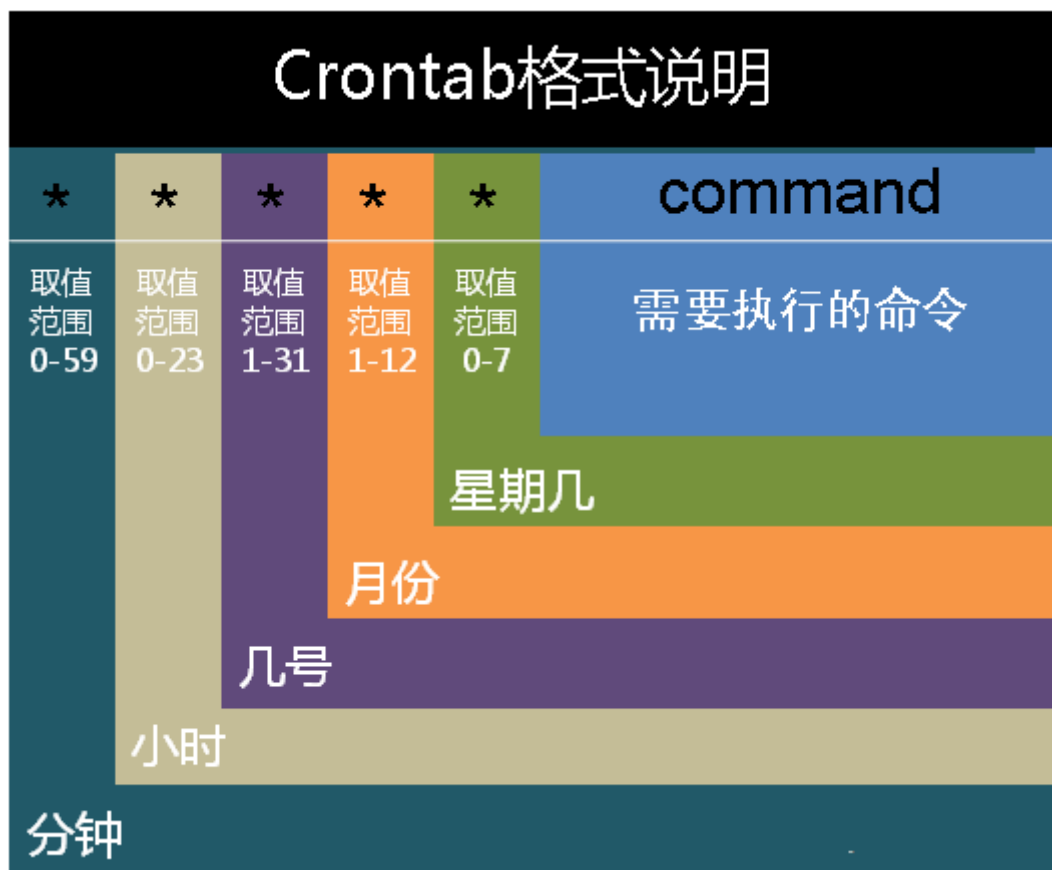
Linux 系统则是由 cron (crond) 这个系统服务来控制的。Linux 系统上面原本就有非常多的计划性工作，因此这个系统服务是默认启动的。

Linux 系统也提供了Linux用户控制计划任务的命令：crontab 命令。



- 日志文件：ll /var/log/cron*
- 编辑文件：vim /etc/crontab
- 进程：ps -ef | grep crond ==> /etc/init.d/crond restart

- 作用：任务（命令）定时调度（如：定时备份，实时备份）
- 简要说明：cat /etc/crontab



在以上各个字段中，还可以使用以下特殊字符：

- * 代表所有的取值范围内的数字。如月份字段为*，则表示1到12个月；
- / 代表每一定时间间隔的意思。如分钟字段为*/10，表示每10分钟执行1次；
- 代表从某个区间范围，是闭区间。如2-5表示2,3,4,5，小时字段中0-23/2表示在0~23点范围内每2个小时执行一次；
- , 分散的数字（不连续）。如1,2,3,4,7,9；

注：由于各个地方每周第一天不一样，因此Sunday=0（第1天）或Sunday=7（最后1天）。

crontab配置实例

```

1 # 每一分钟执行一次command（因cron默认每1分钟扫描一次，因此全为*即可）
2 * * * * * command
3
4 # 每小时的第3和第15分钟执行command
5 3,15 * * * * command
6
```

```

7  # 每天上午8-11点的第3和15分钟执行command
8  3,15  8-11  * * *  command
9
10 # 每隔2天的上午8-11点的第3和15分钟执行command
11 3,15  8-11  */2 * *  command
12
13 # 每个星期一的上午8点到11点的第3和第15分钟执行command
14 3,15  8-11  * * 1  command
15
16 # 每晚的21:30执行command
17 30 21 * * *  command
18
19 # 每月1、10、22日的4:45执行command
20 45 4 1,10,22 * *  command
21
22 # 每周六、周日的1 : 10执行command
23 10 1 * * 6,0  command
24
25 # 每小时执行command
26 0 */1 * * *  command
27
28 # 晚上11点到早上7点之间，每隔一小时执行command
29 * 23-7/1 * * *  command

```

第3节 任务集成部署

3.1、Airflow核心概念

- DAGs：有向无环图(Directed Acyclic Graph)，将所有需要运行的tasks按照依赖关系组织起来，描述的是所有tasks执行的顺序；
- Operators：Airflow内置了很多operators
 - BashOperator 执行一个bash 命令
 - PythonOperator 调用任意的 Python 函数
 - EmailOperator 用于发送邮件
 - HTTPOperator 用于发送HTTP请求
 - SqlOperator 用于执行SQL命令
 - 自定义Operator
- Tasks：Task 是 Operator的一个实例；

- Task Instance: 由于Task会被重复调度, 每次task的运行就是不同的 Task instance。Task instance 有自己的状态, 包括 `success`、`running`、`failed`、`skipped`、`up_for_reschedule`、`up_for_retry`、`queued`、`no_status` 等;
- Task Relationships: DAGs中的不同Tasks之间可以有依赖关系;
- 执行器 (Executor) 。Airflow支持的执行器就有四种:
 - SequentialExecutor: 单进程顺序执行任务, 默认执行器, 通常只用于测试
 - LocalExecutor: 多进程本地执行任务
 - CeleryExecutor: 分布式调度, 生产常用。Celery是一个分布式调度框架, 其本身无队列功能, 需要使用第三方组件, 如RabbitMQ
 - DaskExecutor: 动态任务调度, 主要用于数据分析
 - 执行器的修改。修改 `$AIRFLOW_HOME/airflow.cfg` 第 70行: `executor = LocalExecutor`。修改后启动服务

3.2、入门案例

放置在 `$AIRFLOW_HOME/dags` 目录下

```

1  from datetime import datetime, timedelta
2
3  from airflow import DAG
4  from airflow.utils import dates
5  from airflow.utils.helpers import chain
6  from airflow.operators.bash_operator import BashOperator
7  from airflow.operators.python_operator import PythonOperator
8
9  def default_options():
10     default_args = {
11         'owner': 'airflow',           # 拥有者名称
12         'start_date': dates.days_ago(1), # 第一次开始执行
                                           的时间
13         'retries': 1,                 # 失败重试次数
14         'retry_delay': timedelta(seconds=5) # 失败重试间隔
15     }
16     return default_args
17
18 # 定义DAG
19 def task1(dag):
20     t = "pwd"
21     # operator支持多种类型, 这里使用 BashOperator
22     task = BashOperator(

```

```

23         task_id='MyTask1',                                # task_id
24         bash_command=t,                                    # 指定要执行的命
    令
25         dag=dag                                            # 指定归属的dag
26     )
27     return task
28
29 def hello_world():
30     current_time = str(datetime.today())
31     print('hello world at {}'.format(current_time))
32
33 def task2(dag):
34     # Python Operator
35     task = PythonOperator(
36         task_id='MyTask2',
37         python_callable=hello_world,                        # 指定要执行的函
    数
38         dag=dag)
39     return task
40
41 def task3(dag):
42     t = "date"
43     task = BashOperator(
44         task_id='MyTask3',
45         bash_command=t,
46         dag=dag)
47     return task
48
49 with DAG(
50     'HelloWorldDag',                                       # dag_id
51     default_args=default_options(),                       # 指定默认参数
52     schedule_interval="*/2 * * * *"                      # 执行周期，每分
    钟2次
53 ) as d:
54     task1 = task1(d)
55     task2 = task2(d)
56     task3 = task3(d)
57     chain(task1, task2, task3)                            # 指定执行顺序

```

```
1 # 执行命令检查脚本是否有错误。如果命令行没有报错，就表示没问题
2 python $AIRFLOW_HOME/dags/helloworld.py
3
4 # 查看生效的 dags
5 airflow list_dags -sd $AIRFLOW_HOME/dags
6
7 # 查看指定dag中的task
8 airflow list_tasks HelloWorldDag
9
10 # 测试dag中的task
11 airflow test HelloWorldDag MyTask2 20200801
```

3.3、核心交易调度任务集成

核心交易分析

```
1 # 加载ODS数据（DataX迁移数据）
2 /data/lagoudw/script/trade/ods_load_trade.sh
3
4 # 加载DIM层数据
5 /data/lagoudw/script/trade/dim_load_product_cat.sh
6 /data/lagoudw/script/trade/dim_load_shop_org.sh
7 /data/lagoudw/script/trade/dim_load_payment.sh
8 /data/lagoudw/script/trade/dim_load_product_info.sh
9
10 # 加载DWD层数据
11 /data/lagoudw/script/trade/dwd_load_trade_orders.sh
12
13 # 加载DWS层数据
14 /data/lagoudw/script/trade/dws_load_trade_orders.sh
15
16 # 加载ADS层数据
17 /data/lagoudw/script/trade/ads_load_trade_order_analysis.sh
```

备注： `depends_on_past`，设置为True时，上一次调度成功了，才可以触发。

`$AIRFLOW_HOME/dags`

```
1 from datetime import timedelta
2 import datetime
3 from airflow import DAG
```

```
4 from airflow.operators.bash_operator import BashOperator
5 from airflow.utils.dates import days_ago
6
7 # 定义dag的缺省参数
8 default_args = {
9     'owner': 'airflow',
10    'depends_on_past': False,
11    'start_date': '2020-06-20',
12    'email': ['airflow@example.com'],
13    'email_on_failure': False,
14    'email_on_retry': False,
15    'retries': 1,
16    'retry_delay': timedelta(minutes=5),
17 }
18
19 # 定义DAG
20 coretradedag = DAG(
21     'coretrade',
22     default_args=default_args,
23     description='core trade analyze',
24     schedule_interval='30 0 * * *',
25 )
26
27 today=datetime.date.today()
28 oneday=timedelta(days=1)
29 yesterday=(today-oneday).strftime("%Y-%m-%d")
30
31 odstask = BashOperator(
32     task_id='ods_load_data',
33     depends_on_past=False,
34     bash_command='sh
35 /data/lagoudw/script/trade/ods_load_trade.sh ' + yesterday,
36     dag=coretradedag
37 )
38
39 dimtask1 = BashOperator(
40     task_id='dimtask_product_cat',
41     depends_on_past=False,
42     bash_command='sh
43 /data/lagoudw/script/trade/dim_load_product_cat.sh ' +
44 yesterday,
45     dag=coretradedag
46 )
47
48 dimtask2 = BashOperator(
```

```
46     task_id='dimtask_shop_org',
47     depends_on_past=False,
48     bash_command='sh
/data/lagoudw/script/trade/dim_load_shop_org.sh ' + yesterday,
49     dag=coretradedag
50 )
51
52 dimtask3 = BashOperator(
53     task_id='dimtask_payment',
54     depends_on_past=False,
55     bash_command='sh
/data/lagoudw/script/trade/dim_load_payment.sh ' + yesterday,
56     dag=coretradedag
57 )
58
59 dimtask4 = BashOperator(
60     task_id='dimtask_product_info',
61     depends_on_past=False,
62     bash_command='sh
/data/lagoudw/script/trade/dim_load_product_info.sh ' +
yesterday,
63     dag=coretradedag
64 )
65
66 dwdtask = BashOperator(
67     task_id='dwd_load_data',
68     depends_on_past=False,
69     bash_command='sh
/data/lagoudw/script/trade/dwd_load_trade_orders.sh '+
yesterday,
70     dag=coretradedag
71 )
72
73 dwstask = BashOperator(
74     task_id='dws_load_data',
75     depends_on_past=False,
76     bash_command='sh
/data/lagoudw/script/trade/dws_load_trade_orders.sh ' +
yesterday,
77     dag=coretradedag
78 )
79
80 adstask = BashOperator(
81     task_id='ads_load_data',
82     depends_on_past=False,
```

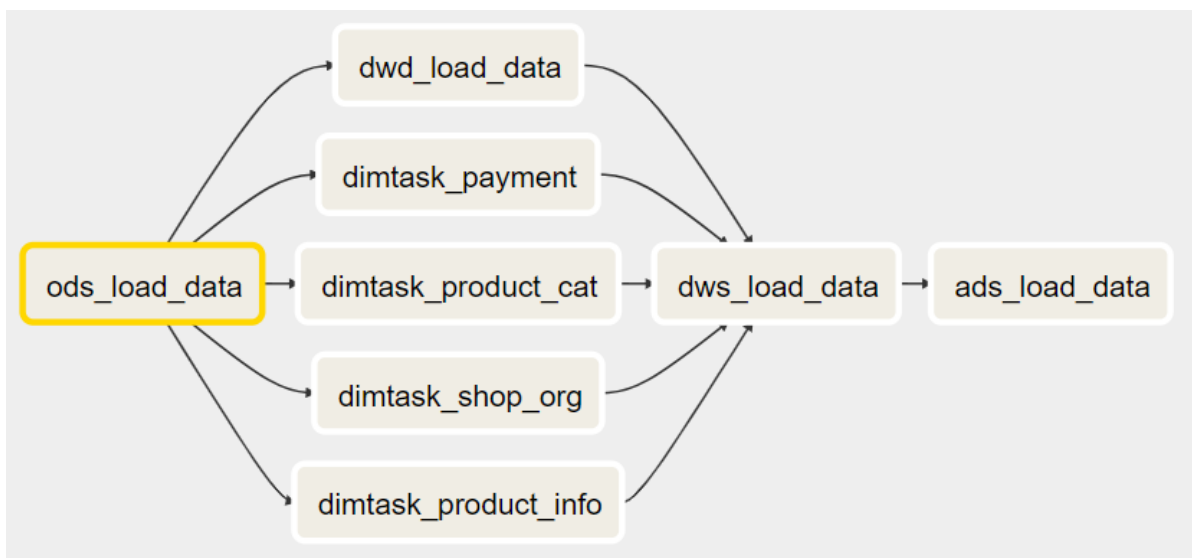
```

83     bash_command='sh
/data/lagoudw/script/trade/ads_load_trade_order_analysis.sh '
+ yesterday,
84     dag=coretradedag
85 )
86
87 odstask >> dimtask1
88 odstask >> dimtask2
89 odstask >> dimtask3
90 odstask >> dimtask4
91 odstask >> dwdtask
92
93 dimtask1 >> dwstask
94 dimtask2 >> dwstask
95 dimtask3 >> dwstask
96 dimtask4 >> dwstask
97 dwdtask >> dwstask
98
99 dwstask >> adstask

```

airflow list_dags

airflow list_tasks coretrade --tree



第三部分 元数据管理工具Atlas (扩展)

第1节 数据仓库元数据管理

元数据（MetaData）狭义的解释是用来描述数据的数据。广义的来看，除了业务逻辑直接读写处理的那些业务数据，所有其它用来维持整个系统运转所需的信息 / 数据都可以叫作元数据。如数据库中表的Schema信息，任务的血缘关系，用户和脚本 / 任务的权限映射关系信息等。

管理元数据的目的是为了让用户能够更高效的使用数据，也是为了让平台管理人员能更加有效的做好数据的维护管理工作。

但通常这些元数据信息是散落在平台的各个系统，各种流程之中的，它们的管理也可能或多或少可以通过各种子系统自身的工具，方案或流程逻辑来实现。

元数据管理平台很重要的一个功能就是信息的收集，至于收集哪些信息，取决于业务的需求和需要解决的目标问题。

元数据管理平台还需要考虑如何以恰当的形式对这些元数据信息进行展示；进一步的，如何将这些元数据信息通过服务的形式提供给周边上下游系统使用，真正帮助大数据平台完成质量管理的闭环工作。

应该收集那些信息，没有绝对的标准，但是对大数据开发平台来说，常见的元数据信息包括：

- 表结构信息
- 数据的空间存储，读写记录，权限归属和其它各类统计信息
- 数据的血缘关系信息
- 数据的业务属性信息

数据血缘关系。血缘信息或者叫做Lineage的血统信息是什么，简单的说就是数据之间的上下游来源去向关系，数据从哪里来到哪里去。如果一个数据有问题，可以根据血缘关系往上游排查，看看到底在哪个环节出了问题。此外也可以通过数据的血缘关系，建立起生产这些数据的任务之间的依赖关系，进而辅助调度系统的工作调度，或者用来判断一个失败或错误的任务可能对哪些下游数据造成影响等等。

分析数据的血缘关系看起来简单，但真的要做起来，并不容易，因为数据的来源多种多样，加工数据的手段，所使用的计算框架可能也各不相同，此外也不是所有的系统天生都具备获取相关信息的能力。而针对不同的系统，血缘关系具体能够分析到的粒度可能也不一样，有些能做到表级别，有些甚至可以做到字段级别。

以Hive表为例，通过分析Hive脚本的执行计划，是可以做到相对精确的定位出字段级别的数据血缘关系的。而如果是一个MapReduce任务生成的数据，从外部来看，可能就只能通过分析MR任务输出的Log日志信息来粗略判断目录级别的读写关系，从而间接推导数据的血缘依赖关系了。

数据的业务属性信息。业务属性信息都有哪些呢？如一张数据表的统计口径信息，这张表干什么用的，各个字段的具体统计方式，业务描述，业务标签，脚本逻辑的历史变迁记录，变迁原因等，此外还包括对应的数据表格是由谁负责开发的，具体数据的业务部门归属等。数据的业务属性信息，首先是为业务服务的，它的采集和展示也就需要尽可能的和业务环境相融合，只有这样才能真正发挥这部分元数据信息的作用。

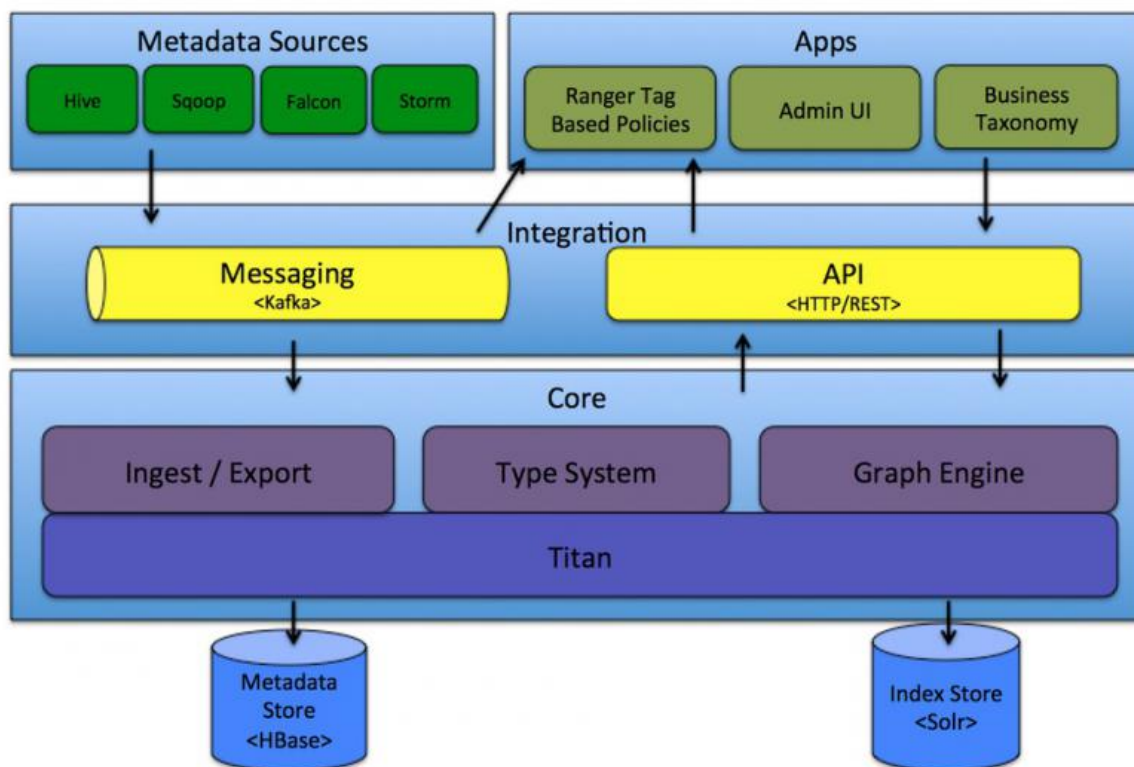
很长一段时间内，市面都没有成熟的大数据元数据管理解决方案。直到2015年，Hortonworks终于坐不住了，约了一众小伙伴公司倡议：咱们开始整个数据治理方案吧。然后，包含数据分类、集中策略引擎、数据血缘、安全和生命周期管理功能的Atlas应运而生。（类似的产品还有Linkedin 在2016年新开源的项目 whereHows）

第2节 Atlas简介

Atlas是Hadoop平台元数据框架；

Atlas是一组可扩展的核心基础治理服务，使企业能够有效，高效地满足Hadoop中的合规性要求，并能与整个企业数据生态系统集成；

Apache Atlas为组织提供了开放的元数据管理和治理功能，以建立数据资产的目录，对这些资产进行分类和治理，并为IT团队、数据分析团队提供围绕这些数据资产的协作功能。



Atlas由元数据的收集，存储和查询展示三部分核心组件组成。此外，还会有一个管理后台对整体元数据的采集流程以及元数据格式定义和服务的部署等各项内容进行配置管理。

Atlas包括以下组件：

- Core。Atlas功能核心组件，提供元数据的获取与导出(Ingets/Export)、类型系统(Type System)、元数据存储索引查询等核心功能
- Integration。Atlas对外集成模块。外部组件的元数据通过该模块将元数据交给Atlas管理
- Metadata source。Atlas支持的元数据数据源，以插件形式提供。当前支持从以下来源提取和管理元数据：
 - Hive
 - HBase
 - Sqoop
 - Kafka
 - Storm
- Applications。Atlas的上层应用，可以用来查询由Atlas管理的元数据类型和对象
- Graph Engine（图计算引擎）。Atlas使用图模型管理元数据对象。图数据库提供了极大的灵活性，并能有效处理元数据对象之间的关系。除了管理图对象之外，图计算引擎还为元数据对象创建适当的索引，以便进行高效的访问。在Atlas 1.0 之前采用Titan作为图存储引擎，从1.0开始采用 JanusGraph 作为图存储引擎。JanusGraph 底层又分为两块：
 - Metadata Store。采用 HBase 存储 Atlas 管理的元数据；
 - Index Store。采用Solr存储元数据的索引，便于高效搜索；

第3节 安装配置

重点讲解Atlas，不对Atlas的依赖组件做讲解，组件均采用单机模式安装。

编译才能安装。

3.1 安装依赖

- Maven 3.6.3（完成）
- HBase 1.1.2（不需要安装，需要软件包）
- Solr 5.5.1（不需要安装，需要软件包）
- atlas 1.2.0（需要编译）

官方只提供了源码，没有提供二进制的安装版本，因此Atlas需要编译。

3.2 安装步骤

1、准备软件包

apache-atlas-1.2.0-sources.tar.gz

solr-5.5.1.tgz

hbase-1.1.2.tar.gz

2、解压缩源码，修改配置

```
1 # 解压缩
2 cd /opt/lagou/software
3 tar zxvf apache-atlas-1.2.0-sources.tar.gz
4 cd apache-atlas-sources-1.2.0/
5
6 # 修改配置
7 vi pom.xml
8
9 # 修改
10 645 <npm-for-v2.version>3.10.8</npm-for-v2.version>
11 652 <hadoop.version>2.9.2</hadoop.version>
```

3、将HBase、Solr的包拷贝到对应的目录中

如果不拷贝这些包，就需要下载，下载 HBase 和 Solr 时速度很慢。这里提前下载完所需的这两个组件，拷贝到对应目录中。

```
1 cd /opt/lagou/software/apache-atlas-sources-1.2.0
2
3 # 创建目录
4 cd distro/
5 mkdir solr
6 mkdir hbase
7
8 # 拷贝软件包
9 cp /opt/lagou/software/solr-5.5.1.tgz ./solr/
10 cp /opt/lagou/software/hbase-1.1.2.tar.gz ./hbase/
```

4、maven设置阿里镜像

备注：重要，否则非常慢

```
1 cd $MAVEN_HOME/conf
2
3 # 在配置文件中添加
4 vi settings.xml
5
6 # 加在 158 行后
7     <mirror>
8         <id>alimaven</id>
9         <name>aliyun maven</name>
10
11         <url>http://maven.aliyun.com/nexus/content/groups/public/</url>
12     </mirror>
13     <mirrorOf>central</mirrorOf>
```

5、Atlas编译

```
1 cd /opt/lagou/software/apache-atlas-sources-1.2.0
2 export MAVEN_OPTS="-Xms2g -Xmx2g"
3 mvn clean -DskipTests package -Pdist,embedded-hbase-solr
```

编译过程中大概要下载600M左右的jar，持续的时间比较长。

```

[INFO] atlas-client-common 1.2.0 ..... SUCCESS [ 1.617 s]
[INFO] atlas-client-v1 1.2.0 ..... SUCCESS [ 1.342 s]
[INFO] Apache Atlas Server API 1.2.0 ..... SUCCESS [ 1.076 s]
[INFO] Apache Atlas Notification 1.2.0 ..... SUCCESS [ 30.059 s]
[INFO] atlas-client-v2 1.2.0 ..... SUCCESS [ 0.958 s]
[INFO] Apache Atlas Graph Database Projects 1.2.0 ..... SUCCESS [ 0.100 s]
[INFO] Apache Atlas Graph Database API 1.2.0 ..... SUCCESS [ 6.426 s]
[INFO] Graph Database Common Code 1.2.0 ..... SUCCESS [ 1.789 s]
[INFO] Apache Atlas JanusGraph DB Impl 1.2.0 ..... SUCCESS [02:18 min]
[INFO] Apache Atlas Graph Database Implementation Dependencies 1.2.0 ..... SUCCESS [ 20.978 s]
[INFO] Shaded version of Apache hbase client 1.2.0 ..... SUCCESS [01:02 min]
[INFO] Shaded version of Apache hbase server 1.2.0 ..... SUCCESS [ 1.783 s]
[INFO] Apache Atlas Authorization 1.2.0 ..... SUCCESS [01:17 min]
[INFO] Apache Atlas Repository 1.2.0 ..... SUCCESS [04:44 min]
[INFO] Apache Atlas UI 1.2.0 ..... SUCCESS [02:21 min]
[INFO] Apache Atlas Web Application 1.2.0 ..... SUCCESS [ 47.682 s]
[INFO] Apache Atlas Documentation 1.2.0 ..... SUCCESS [ 2.515 s]
[INFO] Apache Atlas FileSystem Model 1.2.0 ..... SUCCESS [ 0.733 s]
[INFO] Apache Atlas Plugin Classloader 1.2.0 ..... SUCCESS [01:45 min]
[INFO] Apache Atlas Hive Bridge Shim 1.2.0 ..... SUCCESS [ 17.256 s]
[INFO] Apache Atlas Hive Bridge 1.2.0 ..... SUCCESS [ 52.770 s]
[INFO] Apache Atlas Falcon Bridge Shim 1.2.0 ..... SUCCESS [ 2.693 s]
[INFO] Apache Atlas Falcon Bridge 1.2.0 ..... SUCCESS [ 15.507 s]
[INFO] Apache Atlas Sqoop Bridge Shim 1.2.0 ..... SUCCESS [ 2.748 s]
[INFO] Apache Atlas Storm Bridge Shim 1.2.0 ..... SUCCESS [ 28.811 s]
[INFO] Apache Atlas Storm Bridge 1.2.0 ..... SUCCESS [ 6.475 s]
[INFO] Apache Atlas Hbase Bridge Shim 1.2.0 ..... SUCCESS [ 1.359 s]
[INFO] Apache Atlas Hbase Bridge 1.2.0 ..... SUCCESS [ 41.697 s]
[INFO] Apache Atlas Kafka Bridge 1.2.0 ..... SUCCESS [ 1.641 s]
[INFO] Apache Atlas Distribution 1.2.0 ..... SUCCESS [04:00 min]
[INFO] -----
[INFO] BUILD SUCCESS
[INFO] -----
[INFO] Total time: 35:19 min

```

编译完的软件位置：/opt/lagou/software/apache-atlas-sources-1.2.0/distro/target

编译完的软件：apache-atlas-1.2.0-bin.tar.gz

6、Atlas安装

```

1 cd /opt/lagou/software/apache-atlas-sources-
  1.2.0/distro/target
2
3 # 解压缩
4 tar zxvf apache-atlas-1.2.0-bin.tar.gz
5 mv apache-atlas-1.2.0/ /opt/lagou/servers/atlas-1.2.0
6
7 # 修改 /etc/profile, 设置环境变量 ATLAS_HOME
8
9 # 启动服务(第一次启动服务的时间比较长)
10 cd $ATLAS_HOME/bin
11 ./atlas_start.py
12
13 # 检查后台进程 (1个atlas、2个HBase、1个solr后台进程)
14 ps -ef | grep atlas
15

```

```
16 /opt/lagou/servers/atlas-1.2.0/server/webapp/atlas
17 hbase-daemon.sh
18 org.apache.hadoop.hbase.master.HMaster
19 /opt/lagou/servers/atlas-1.2.0/solr/server
20
21 # 停止服务
22 ./atlas_stop.py
```

检查 solr 的状态:


```
1 cd /opt/lagou/servers/atlas-1.2.0/solr/bin
2 ./solr status
3
4 solr process 25038 running on port 9838
5 {
6   "solr_home":"/opt/lagou/servers/atlas-
7 1.2.0/solr/server/solr",
8   "version":"5.5.1 c08f17bca0d9cbf516874d13d221ab100e5b7d58 -
9 anshum - 2016-04-30 13:28:18",
10  "startTime":"2020-07-31T11:58:45.638Z",
11  "uptime":"0 days, 14 hours, 55 minutes, 11 seconds",
12  "memory":"54.8 MB (%11.2) of 490.7 MB",
13  "cloud":{
14    "zookeeper":"localhost:2181",
15    "liveNodes":"1",
16    "collections":"3"}}
```

检查 zk 状态:

```
1 echo stat|nc localhost 2181
```

Web服务: <http://linux122:21000/login.jsp>

Apache Atlas

 Username

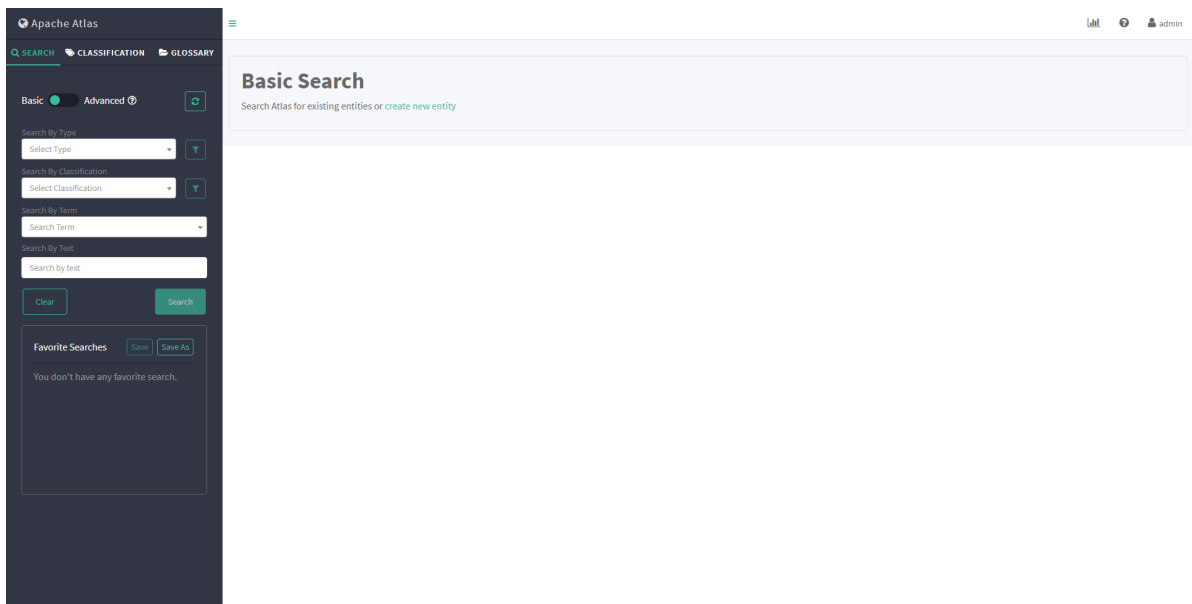
Password

Login

用户名 / 口令：admin / admin

账号的信息存储在文件 `conf/users-credentials.properties` 中。其中 Password 通过如下方式产生sha256sum 摘要信息：

```
1 | echo -n "admin" | sha256sum
```



第4节 Hive血缘关系导入

1、配置HIVE_HOME环境变量；将 `$ATLAS_HOME/conf/atlas-application.properties` 拷贝到 `$HIVE_HOME/conf` 目录下

```
1 | ln -s $ATLAS_HOME/conf/atlas-application.properties  
   | $HIVE_HOME/conf/atlas-application.properties
```

2、拷贝jar包

```

1 # $ATLAS_HOME/server/webapp/atlas/WEB-INF/lib/ 目录下的3个jar,
   拷贝到 $ATLAS_HOME/hook/hive/atlas-hive-plugin-impl/ 目录下
2 jackson-jaxrs-base-2.9.9.jar
3 jackson-jaxrs-json-provider-2.9.9.jar
4 jackson-module-jaxb-annotations-2.9.9.jar
5
6 ln -s $ATLAS_HOME/server/webapp/atlas/WEB-INF/lib/jackson-
   jaxrs-base-2.9.9.jar $ATLAS_HOME/hook/hive/atlas-hive-plugin-
   impl/jackson-jaxrs-base-2.9.9.jar
7
8 ln -s $ATLAS_HOME/server/webapp/atlas/WEB-INF/lib/jackson-
   jaxrs-json-provider-2.9.9.jar $ATLAS_HOME/hook/hive/atlas-
   hive-plugin-impl/jackson-jaxrs-json-provider-2.9.9.jar
9
10 ln -s $ATLAS_HOME/server/webapp/atlas/WEB-INF/lib/jackson-
   module-jaxb-annotations-2.9.9.jar $ATLAS_HOME/hook/hive/atlas-
   hive-plugin-impl/jackson-module-jaxb-annotations-2.9.9.jar

```

3、修改Hive的配置

hive-site.xml增加 hook

```

1 <property>
2   <name>hive.exec.post.hooks</name>
3   <value>org.apache.atlas.hive.hook.HiveHook</value>
4 </property>

```

\$HIVE_HOME/conf/hive-env.sh中添加HIVE_AUX_JARS_PATH变量

```

1 export HIVE_AUX_JARS_PATH=/opt/lagou/servers/atlas-
   1.2.0/hook/hive

```

4、批量导入hive数据

备注：Hive能正常启动；在执行的过程中需要用户名/口令：admin/admin

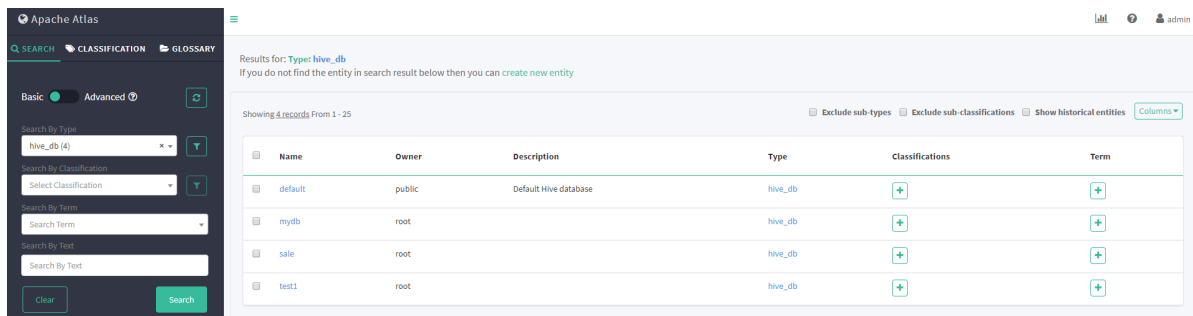
```

1 import-hive.sh

```

成功导出可以看见最后的提示信息：Hive Meta Data imported successfully!!!

在浏览器中可以看见：Search 中的选项有变化



Hive hook 可捕获以下操作：

- create database
- create table/view, create table as select
- load, import, export
- DMLs (insert)
- alter database
- alter table
- alter view

第5节 与电商业务集成

开发（建库、建表） => 导入数据 => 执行Hive脚本

导入Hive的血缘关系

电商业务建表语句（可省略）：

```

1  -- 创建DataBases;
2  CREATE DATABASE ODS;
3  CREATE DATABASE DIM;
4  CREATE DATABASE DWD;
5  CREATE DATABASE DWS;
6  CREATE DATABASE ADS;
7
8  -- 创建ODS表
9  DROP TABLE IF EXISTS `ods.ods_trade_orders`;
10 CREATE EXTERNAL TABLE `ods.ods_trade_orders`(
11     `orderid` int,
12     `orderno` string,
13     `userid` bigint,
14     `status` tinyint,
15     `productmoney` decimal(10,0),

```



```
16     `totalmoney` decimal(10,0),
17     `paymethod` tinyint,
18     `ispay` tinyint,
19     `areaid` int,
20     `tradesrc` tinyint,
21     `tradetype` int,
22     `isrefund` tinyint,
23     `dataflag` tinyint,
24     `createtime` string,
25     `paytime` string,
26     `modifiedtime` string)
27 COMMENT '订单表'
28 PARTITIONED BY (`dt` string)
29 row format delimited fields terminated by ','
30 location '/user/data/trade.db/orders/';
31
32 DROP TABLE IF EXISTS `ods.ods_trade_order_product`;
33 CREATE EXTERNAL TABLE `ods.ods_trade_order_product` (
34     `id` string,
35     `orderid` decimal(10,2),
36     `productid` string,
37     `productnum` string,
38     `productprice` string,
39     `money` string,
40     `extra` string,
41     `createtime` string)
42 COMMENT '订单明细表'
43 PARTITIONED BY (`dt` string)
44 row format delimited fields terminated by ','
45 location '/user/data/trade.db/order_product/';
46
47 DROP TABLE IF EXISTS `ods.ods_trade_product_info`;
48 CREATE EXTERNAL TABLE `ods.ods_trade_product_info` (
49     `productid` bigint,
50     `productname` string,
51     `shopid` string,
52     `price` decimal(10,0),
53     `issale` tinyint,
54     `status` tinyint,
55     `categoryid` string,
56     `createtime` string,
57     `modifytime` string)
58 COMMENT '产品信息表'
59 PARTITIONED BY (`dt` string)
60 row format delimited fields terminated by ','
```

```
61 location '/user/data/trade.db/product_info/';
62
63 DROP TABLE IF EXISTS `ods.ods_trade_product_category`;
64 CREATE EXTERNAL TABLE `ods.ods_trade_product_category` (
65     `catid` int,
66     `parentid` int,
67     `catname` string,
68     `isshow` tinyint,
69     `sortnum` int,
70     `isdel` tinyint,
71     `createtime` string,
72     `level` tinyint)
73 COMMENT '产品分类表'
74 PARTITIONED BY (`dt` string)
75 row format delimited fields terminated by ','
76 location '/user/data/trade.db/product_category';
77
78 DROP TABLE IF EXISTS `ods.ods_trade_shops`;
79 CREATE EXTERNAL TABLE `ods.ods_trade_shops` (
80     `shopid` int,
81     `userid` int,
82     `areaaid` int,
83     `shopname` string,
84     `shoplevel` tinyint,
85     `status` tinyint,
86     `createtime` string,
87     `modifytime` string)
88 COMMENT '商家店铺表'
89 PARTITIONED BY (`dt` string)
90 row format delimited fields terminated by ','
91 location '/user/data/trade.db/shops';
92
93 DROP TABLE IF EXISTS `ods.ods_trade_shop_admin_org`;
94 CREATE EXTERNAL TABLE `ods.ods_trade_shop_admin_org` (
95     `id` int,
96     `parentid` int,
97     `orgname` string,
98     `orglevel` tinyint,
99     `isdelete` tinyint,
100     `createtime` string,
101     `updatetime` string,
102     `isshow` tinyint,
103     `orgType` tinyint)
104 COMMENT '商家地域组织表'
105 PARTITIONED BY (`dt` string)
```

```

106 row format delimited fields terminated by ','
107 location '/user/data/trade.db/shop_org/';
108
109 DROP TABLE IF EXISTS `ods.ods_trade_payments`;
110 CREATE EXTERNAL TABLE `ods.ods_trade_payments`(
111     `id` string,
112     `paymethod` string,
113     `payname` string,
114     `description` string,
115     `payorder` int,
116     `online` tinyint)
117 COMMENT '支付方式表'
118 PARTITIONED BY (`dt` string)
119 row format delimited fields terminated by ','
120 location '/user/data/trade.db/payments/';
121
122 -- 创建DIM表
123 DROP TABLE IF EXISTS dim.dim_trade_product_cat;
124 create table if not exists dim.dim_trade_product_cat(
125     firstId int,                -- 一级商品分类id
126     firstName string,          -- 一级商品分类名称
127     secondId int,              -- 二级商品分类Id
128     secondName string,         -- 二级商品分类名称
129     thirdId int,               -- 三级商品分类id
130     thirdName string           -- 三级商品分类名称
131 )
132 partitioned by (dt string)
133 STORED AS PARQUET;
134
135 drop table if exists dim.dim_trade_shops_org;
136 create table dim.dim_trade_shops_org(
137     shopid int,
138     shopName string,
139     cityId int,
140     cityName string ,
141     regionId int ,
142     regionName string
143 )
144 partitioned by (dt string)
145 STORED AS PARQUET;
146
147 drop table if exists dim.dim_trade_payment;
148 create table if not exists dim.dim_trade_payment(
149     paymentId string,          -- 支付方式id
150     paymentName string         -- 支付方式名称

```

```
151 )
152 partitioned by (dt string)
153 STORED AS PARQUET;
154
155 drop table if exists dim.dim_trade_product_info;
156 create table dim.dim_trade_product_info(
157     `productId` bigint,
158     `productName` string,
159     `shopId` string,
160     `price` decimal,
161     `issale` tinyint,
162     `status` tinyint,
163     `categoryId` string,
164     `createTime` string,
165     `modifyTime` string,
166     `start_dt` string,
167     `end_dt` string
168 ) COMMENT '产品表'
169 STORED AS PARQUET;
170
171 -- 创建DWD表
172 -- 订单事实表(拉链表)
173 DROP TABLE IF EXISTS dwd.dwd_trade_orders;
174 create table dwd.dwd_trade_orders(
175     `orderId` int,
176     `orderNo` string,
177     `userId` bigint,
178     `status` tinyint,
179     `productMoney` decimal,
180     `totalMoney` decimal,
181     `payMethod` tinyint,
182     `isPay` tinyint,
183     `areaId` int,
184     `tradeSrc` tinyint,
185     `tradeType` int,
186     `isRefund` tinyint,
187     `dataFlag` tinyint,
188     `createTime` string,
189     `payTime` string,
190     `modifiedTime` string,
191     `start_date` string,
192     `end_date` string
193 ) COMMENT '订单事实拉链表'
194 partitioned by (dt string)
195 STORED AS PARQUET;
```

```

196
197 -- 创建DWS表
198 DROP TABLE IF EXISTS dws.dws_trade_orders;
199 create table if not exists dws.dws_trade_orders(
200     orderid      string,          -- 订单id
201     cat_3rd_id   string,          -- 商品三级分类id
202     shopid       string,          -- 店铺id
203     paymethod    tinyint,         -- 支付方式
204     productsnum  bigint,          -- 商品数量
205     paymoney     double,          -- 订单商品明细金额
206     paytime      string           -- 订单时间
207 )
208 partitioned by (dt string)
209 STORED AS PARQUET;
210
211 -- 订单明细表宽表
212 DROP TABLE IF EXISTS dws.dws_trade_orders_w;
213 create table if not exists dws.dws_trade_orders_w(
214     orderid string,              -- 订单id
215     cat_3rd_id string,           -- 商品三级分类id
216     thirdname string,           -- 商品三级分类名称
217     secondname string,          -- 商品二级分类名称
218     firstname string,           -- 商品一级分类名称
219     shopid string,              -- 店铺id
220     shopname string,            -- 店铺名
221     regionname string,          -- 店铺所在大区
222     cityname string,            -- 店铺所在城市
223     paymethod tinyint,          -- 支付方式
224     productsnum bigint,         -- 商品数量
225     paymoney double,            -- 订单明细金额
226     paytime string              -- 订单时间
227 )
228 partitioned by (dt string)
229 STORED AS PARQUET;
230
231 -- 创建ADS表
232 -- ADS层订单分析表
233 DROP TABLE IF EXISTS ads.ads_trade_order_analysis;
234 create table if not exists ads.ads_trade_order_analysis(
235     areatype string,            -- 区域范围：区域类型（全国、大
236                                -- 区、城市）
237     regionname string,          -- 区域名称
238     cityname string,            -- 城市名称
239     categorytype string,        -- 商品分类类型（一级、二级）
240     category1 string,           -- 商品一级分类名称

```

```

240     category2 string,           -- 商品二级分类名称
241     totalcount bigint,         -- 订单数量
242     total_productnum bigint,   -- 商品数量
243     totalmoney double         -- 支付金额
244 )
245 partitioned by (dt string)
246 row format delimited fields terminated by ',';

```

使用Sqoop加载数据（可省略）：

```

1  sqoop import \
2  --connect jdbc:mysql://linux123:3306/ebiz \
3  --username hive \
4  --password 12345678 \
5  --target-dir /user/data/trade.db/orders/dt=2020-07-21/ \
6  --table lagou_trade_orders \
7  --delete-target-dir \
8  --num-mappers 1 \
9  --fields-terminated-by ','
10
11 sqoop import \
12 --connect jdbc:mysql://linux123:3306/ebiz \
13 --username hive \
14 --password 12345678 \
15 --target-dir /user/data/trade.db/payments/dt=2020-07-21/ \
16 --table lagou_payments \
17 --delete-target-dir \
18 --num-mappers 1 \
19 --fields-terminated-by ','
20
21 sqoop import \
22 --connect jdbc:mysql://linux123:3306/ebiz \
23 --username hive \
24 --password 12345678 \
25 --target-dir /user/data/trade.db/product_category/dt=2020-07-
26 21/ \
27 --table lagou_product_category \
28 --delete-target-dir \
29 --num-mappers 1 \
30 --fields-terminated-by ','
31
32 sqoop import \
33 --connect jdbc:mysql://linux123:3306/ebiz \

```

```
33 --username hive \  
34 --password 12345678 \  
35 --target-dir /user/data/trade.db/product_info/dt=2020-07-21/ \  
36 --table lagou_product_info \  
37 --delete-target-dir \  
38 --num-mappers 1 \  
39 --fields-terminated-by ','  
40  
41 sqoop import \  
42 --connect jdbc:mysql://linux123:3306/ebiz \  
43 --username hive \  
44 --password 12345678 \  
45 --target-dir /user/data/trade.db/order_product/dt=2020-07-21/  
46 --table lagou_order_product \  
47 --delete-target-dir \  
48 --num-mappers 1 \  
49 --fields-terminated-by ','  
50  
51 sqoop import \  
52 --connect jdbc:mysql://linux123:3306/ebiz \  
53 --username hive \  
54 --password 12345678 \  
55 --target-dir /user/data/trade.db/shop_org/dt=2020-07-21/ \  
56 --table lagou_shop_admin_org \  
57 --delete-target-dir \  
58 --num-mappers 1 \  
59 --fields-terminated-by ','  
60  
61 sqoop import \  
62 --connect jdbc:mysql://linux123:3306/ebiz \  
63 --username hive \  
64 --password 12345678 \  
65 --target-dir /user/data/trade.db/shops/dt=2020-07-21/ \  
66 --table lagou_shops \  
67 --delete-target-dir \  
68 --num-mappers 1 \  
69 --fields-terminated-by ','  
70  
71 alter table ods.ods_trade_orders add partition(dt='2020-07-  
21');  
72 alter table ods.ods_trade_payments add partition(dt='2020-07-  
21');  
73 alter table ods.ods_trade_product_category add  
partition(dt='2020-07-21');
```

```
74 alter table ods.ods_trade_product_info add partition(dt='2020-07-21');
75 alter table ods.ods_trade_order_product add
partition(dt='2020-07-21');
76 alter table ods.ods_trade_shop_admin_org add
partition(dt='2020-07-21');
77 alter table ods.ods_trade_shops add partition(dt='2020-07-21');
```

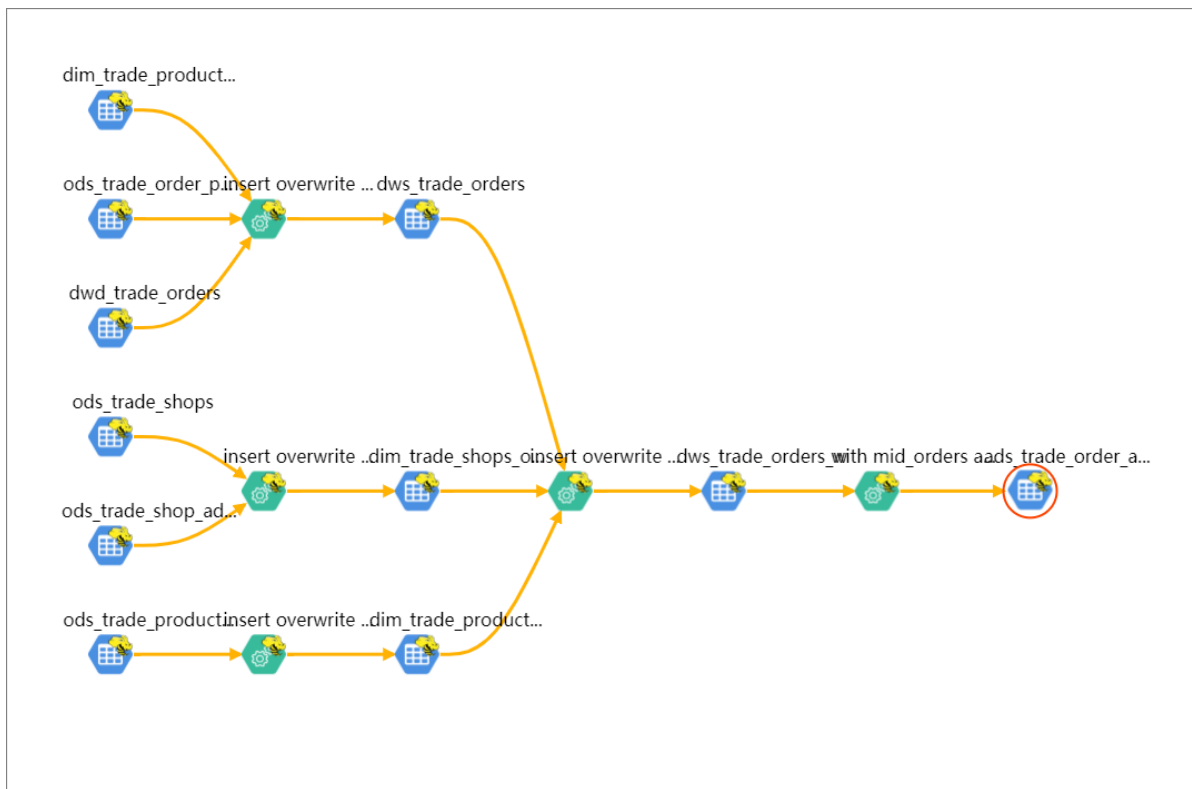
电商业务脚本(省略了ODS层数据加载):

```
1  # 加载DIM层数据
2  sh /data/lagoudw/script/trade/dim_load_product_cat.sh 2020-07-21
3  sh /data/lagoudw/script/trade/dim_load_shop_org.sh 2020-07-21
4  sh /data/lagoudw/script/trade/dim_load_payment.sh 2020-07-21
5  sh /data/lagoudw/script/trade/dim_load_product_info.sh 2020-07-21
6
7  # 加载DWD层数据
8  sh /data/lagoudw/script/trade/dwd_load_trade_orders.sh 2020-07-21
9
10 # 加载DWS层数据
11 sh /data/lagoudw/script/trade/dws_load_trade_orders.sh 2020-07-21
12
13 # 加载ADS层数据
14 sh /data/lagoudw/script/trade/ads_load_trade_order_analysis.sh 2020-07-21
```

创建 Classification: order_analysis

创建Glossary: ODS层 => 电商业务

查看血缘关系 ads_trade_order_analysis:



第四部分 数据质量监控工具 Griffin（扩展）

第1节 为什么要做数据质量监控

garbage in garbage out

1、数据不一致

企业早期没有进行统一规划设计，大部分信息系统是逐步迭代建设的，系统建设时间长短各异，各系统数据标准也不同。企业业务系统更关注业务层面，各个业务系统均有不同的侧重点，各类数据的属性信息设置和要求不统一。另外，由于各系统的相互独立使用，无法及时同步更新相关信息等各种原因造成各系统间的数据不一致，严重影响了各系统间的数据交互和统一识别，基础数据难以共享利用，数据的深层价值也难以体现。

2、数据不完整

由于企业信息系统的孤立使用，各个业务系统或模块按照各自的需要录入数据，没有统一的录入工具和数据出口，业务系统不需要的信息就不录，造成同样的数据在不同的系统有不同的属性信息，数据完整性无法得到保障。

3、数据不合规

没有统一的数据管理平台和数据源头，数据全生命周期管理不完整，同时企业各信息系统的录入环节过于简单且手工参与较多，就数据本身而言，缺少是否重复、合法、对错等校验环节，导致各个信息系统的数据不够准确，格式混乱，各类数据难以集成和统一，没有质量控制导致海量数据因质量过低而难以被利用，且没有相应的数据管理流程。

4、数据不可控

海量数据多头管理，缺少专门对数据管理进行监督和控制的组织。企业各单位和部门关注数据的角度不一样，缺少一个组织从全局的视角对数据进行管理，导致无法建立统一的数据管理标准、流程等，相应的数据管理制度、办法等无法得到落实。同时，企业基础数据质量考核体系也尚未建立，无法保障一系列数据标准、规范、制度、流程得到长效执行。

5、数据冗余

各个信息系统针对数据的标准规范不一、编码规则不一、校验标准不一，且部分业务系统针对数据的验证标准严重缺失，造成了企业顶层视角的数据出现“一物多码”、“一码多物”等现象。

第2节 数据质量监控方法

1、设计思路

数据质量监控的设计要分为4个模块：数据，规则，告警和反馈

- ①数据：需要被监控的数据，可能存放在不同的存储引擎中
- ②规则：值如何设计发现异常的规则，一般而言主要是数值的异常和环比等异常监控方式。也会有一些通过算法来发掘异常数据的方法
- ③告警：告警是指发告警的动作，这里可以通过微信消息，电话或者短信，邮件
- ④反馈：反馈是指对告警内容的反馈，比如说收到的告警内容，要有人员回应该告警消息是否是真的异常，是否需要忽略该异常，是否已经处理了该异常。有了反馈机制，整个数据监控才能形成闭环

2、技术方案

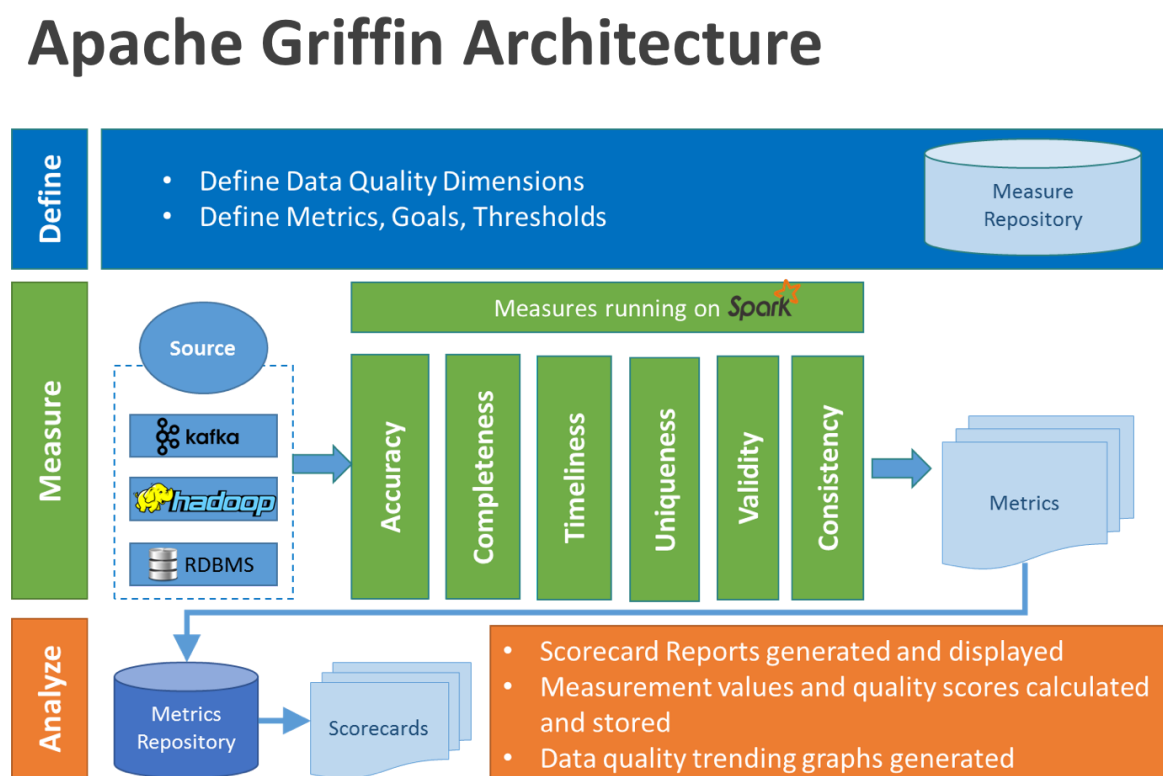
- 最开始可以先关注核心要监控的内容，比如说准确性，那么就对核心的一些指标做监控即可
- 监控平台尽量不要做太复杂的规则逻辑，尽量只对结果数据进行监控。比如要监控日质量是否波动过大，那么把该计算流程提前，先计算好结果表，最后监控平台只监控结果表是否异常即可

- 多数据源。多数据源的监控有两种方式：针对每个数据源定制实现一部分计算逻辑，也可以通过额外的任务将多数据源中的数据结果通过任务写入一个数据源中，再对该数据源进行监控
- 实时数据监控：区别在于扫描周期的不同，因此在设计的时候可以先以离线为主，但是尽量预留好实时监控的设计

第3节 Griffin架构

Apache Griffin是一个开源的大数据数据质量解决方案，它支持批处理和流模式两种数据质量检测方式，可以从不同维度（如离线任务执行完毕后检查源端和目标端的数据数量是否一致、源表的数据空值数量等）度量数据资产，从而提升数据的准确度、可信度。

Griffin主要分为Define、Measure和Analyze三个部分，如下图所示：



各部分的职责如下：

- Define：主要负责定义数据质量统计的维度，比如数据质量统计的时间跨度、统计的目标（源端和目标端的数据数量是否一致，数据源里某一字段的非空的数量、不重复值的数量、最大值、最小值、top5的值数量等）
- Measure：主要负责执行统计任务，生成统计结果
- Analyze：主要负责保存与展示统计结果

第4节 编译安装

4.1 相关依赖

重点讲解 Griffin，不对依赖组件做过多讲解，所有组件均采用单机模式安装。

- JDK (1.8 or later versions)
- MySQL(version 5.6及以上)
- Hadoop (2.6.0 or later)
- Hive (version 2.x)
- Maven
- Spark (version 2.2.1)
- Livy (livy-0.5.0-incubating)
- Elasticsearch (5.0 or later versions)

备注：

- Spark：计算批量、实时指标
- Livy：为服务提供 RESTful API 调用 Apache Spark
- Elasticsearch：存储指标数据
- MySQL：服务元数据

4.2 Spark安装

1、解压缩，设置环境变量 \$SPARK_HOME

```
1 tar zxvf spark-2.2.1-bin-hadoop2.7.tgz
2 mv spark-2.2.1-bin-hadoop2.7/ /opt/lagou/servers/spark-2.2.1/
3
4 # 设置环境变量
5 vi /etc/profile
6
7 export SPARK_HOME=/opt/lagou/servers/spark-2.2.1/
8 export PATH=$PATH:$SPARK_HOME/bin
9
10 source /etc/profile
```

2、修改配置文件 \$SPARK_HOME/conf/spark-defaults.conf

```
1 spark.master yarn
2 spark.eventLog.enabled true
3 spark.eventLog.dir
  hdfs://linux121:9000/spark/logs
4 spark.serializer
  org.apache.spark.serializer.KryoSerializer
5 spark.yarn.jars
  hdfs://linux121:9000/spark/spark_2.2.1_jars/*
```

备注：上面的路径要创建

拷贝 MySQL 驱动

```
1 cp $HIVE_HOME/lib/mysql-connector-java-5.1.46.jar
   $SPARK_HOME/jars/
```

将 Spark 的 jar 包上传到 hdfs://hadoop1:9000/spark/spark_2.2.1_jars/

```
1 hdfs dfs -mkdir -p /spark/logs
2 hdfs dfs -mkdir -p /spark/spark_2.2.1_jars/
3 hdfs dfs -put /opt/lagou/servers/spark-2.2.1/jars/*.jar
  /spark/spark_2.2.1_jars/
```

3、修改配置文件spark-env.sh

```
1 export JAVA_HOME=/opt/lagou/servers/jdk1.8.0_231/
2 export HADOOP_HOME=/opt/lagou/servers/hadoop-2.9.2/
3 export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
4 export SPARK_DIST_CLASSPATH=$(hadoop classpath)
5 export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop
```

4、yarn-site.xml 添加配置

```
1 <property>
2   <name>yarn.nodemanager.vmem-check-enabled</name>
3   <value>>false</value>
4 </property>
```

yarn.nodemanager.vmem-check-enabled: 是否检查虚拟内存。

修改所有节点，并重启yarn服务。

不添加该配置启动spark-shell，有如下错误：Yarn application has already ended! It might have been killed or unable to launch application master.

5、测试spark

spark-shell

```
1 // /wcinput/wc.txt : HDFS上的文件
2 val lines = sc.textFile("/wcinput/wc.txt")
3 lines.flatMap(_.split("
  ")).map(_._1).reduceByKey(_+_).collect()
```

4.3 Livy安装

1、解压缩，设置环境变量 \$LIVY_HOME

```
1 unzip livy-0.5.0-incubating-bin.zip
2 mv livy-0.5.0-incubating-bin/ ../servers/livy-0.5.0
3
4 # 设置环境变量
5 vi /etc/profile
6
7 export LIVY_HOME=/opt/lagou/servers/livy-0.5.0
8 export PATH=$PATH:$LIVY_HOME/bin
9
10 source /etc/profile
```

2、修改配置文件 conf/livy.conf

```
1 livy.server.host = 127.0.0.1
2 livy.spark.master = yarn
3 livy.spark.deployMode = cluster
4 livy.repl.enable-hive-context = true
```

3、修改配置文件 conf/livy-env.sh

```
1 export SPARK_HOME=/opt/lagou/servers/spark-2.2.1
2 export HADOOP_HOME=/opt/lagou/servers/hadoop-2.9.2/
3 export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
```

4、启动服务

```
1 cd /opt/lagou/servers/livy-0.5.0
2 mkdir logs
3
4 nohup bin/livy-server &
```

4.4 ES安装

1、解压缩

```
1 tar zxvf elasticsearch-5.6.0.tar.gz
2 mv elasticsearch-5.6.0/ ../software/
```

2、创建 elasticsearch用户组 及 elasticsearch 用户。不能使用root用户启动ES程序，需要创建单独的用户去启动ES 服务；

```
1 # 创建用户组
2 groupadd elasticsearch
3
4 # 创建用户
5 useradd elasticsearch -g elasticsearch
6
7 # 修改安装目录的宿主
8 chown -R elasticsearch:elasticsearch elasticsearch-5.6.0/
```

3、修改linux系统文件 /etc/security/limits.conf

```
1 elasticsearch hard nofile 1000000
2 elasticsearch soft nofile 1000000
3 * soft nproc 4096
4 * hard nproc 4096
```

4、修改系统文件 /etc/sysctl.conf

```
1 # 文件末尾增加:
2 vm.max_map_count=262144
3
4 # 执行以下命令，修改才能生效
5 sysctl -p
```

5、修改es配置文件

/opt/lagou/servers/elasticsearch-5.6.0/config/elasticsearch.yml

```
1 network.host: 0.0.0.0
```

/opt/lagou/servers/elasticsearch-5.6.0/config/jvm.options

jvm内存的分配，原来都是2g，修改为1g

```
1 -Xms1g
2 -Xmx1g
```

6、启动ES服务

```
1 # 到ES安装目录下，执行命令(-d表示后台启动)
2 su elasticsearch
3 cd /opt/lagou/servers/elasticsearch-5.6.0/
4 bin/elasticsearch -d
```

在浏览器中检查：<http://linux122:9200/>


```
{
  "name" : "14TSMYw",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "ZSy0cL5hTtiuEL_g0nDzUQ",
  "version" : {
    "number" : "5.6.0",
    "build_hash" : "781a835",
    "build_date" : "2017-09-07T03:09:58.087Z",
    "build_snapshot" : false,
    "lucene_version" : "6.6.0"
  },
  "tagline" : "You Know, for Search"
}
```

7、在ES里创建griffin索引

```
1 # linux122 为 ES 服务所在节点
2 curl -XPUT http://linux122:9200/griffin -d '
3 {
4   "aliases": {},
5   "mappings": {
6     "accuracy": {
7       "properties": {
8         "name": {
9           "fields": {
10             "keyword": {
11               "ignore_above": 256,
12               "type": "keyword"
13             }
14           },
15           "type": "text"
16         },
17         "tmst": {
18           "type": "date"
19         }
20       }
21     }
22   },
23   "settings": {
24     "index": {
25       "number_of_replicas": "2",
26       "number_of_shards": "5"
27     }
28   }
29 }
30 '
```

4.5 Griffin编译准备

1、软件解压缩

```
1 cd /opt/lagou/software
2 unzip griffin-griffin-0.5.0.zip
3 mv griffin-griffin-0.5.0/ ../servers/griffin-0.5.0/
4 cd griffin-0.5.0
```

2、在MySQL中创建数据库quartz，并初始化

/opt/lagou/servers/griffin-0.5.0/service/src/main/resources/Init_quartz_mysql_innodb.sql

备注：要做简单的修改，主要是增加 use quartz;

```
1 # mysql中执行创建数据库
2 create database quartz;
3
4 # 命令行执行，创建表
5 mysql -uhive -p12345678 < Init_quartz_mysql_innodb.sql
```

3、Hadoop和Hive

在HDFS上创建/spark/spark_conf目录，并将Hive的配置文件hive-site.xml上传到该目录下

```
1 hdfs dfs -mkdir -p /spark/spark_conf
2 hdfs dfs -put $HIVE_HOME/conf/hive-site.xml /spark/spark_conf/
```

备注：将安装 griffin 所在节点上的 hive-site.xml 文件，上传到 HDFS 对应目录中；

4、确保设置以下环境变量(/etc/profile)

```
1 export JAVA_HOME=/opt/lagou/servers/hadoop-2.9.2
2 export SPARK_HOME=/opt/lagou/servers/spark-2.2.1/
3 export LIVY_HOME=/opt/lagou/servers/livy-0.5.0
4 export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
```

4.6 Griffin编译

1、service/pom.xml文件配置

编辑 service/pom.xml (约113-117行) , 增加MySQLJDBC 依赖 (即删除注释) :

```
1 <dependency>
2     <groupId>mysql</groupId>
3     <artifactId>mysql-connector-java</artifactId>
4     <version>${mysql.java.version}</version>
5 </dependency>
```

2、修改配置文件 service/src/main/resources/application.properties

```
1 server.port = 9876
2
3 spring.application.name=griffin_service
4 spring.datasource.url=jdbc:mysql://linux123:3306/quartz?
  autoReconnect=true&useSSL=false
5 spring.datasource.username=hive
6 spring.datasource.password=12345678
7 spring.jpa.generate-ddl=true
8 spring.datasource.driver-class-name=com.mysql.jdbc.Driver
9 spring.jpa.show-sql=true
10
11 # Hive metastore
12 hive.metastore.uris=thrift://linux123:9083
13 hive.metastore.dbname=hivemetadata
14 hive.hmsHandler.retry.attempts=15
15 hive.hmsHandler.retry.interval=2000ms
16
17 # Hive cache time
18 cache.evict.hive.fixedRate.in.milliseconds=900000
19
20 # Kafka schema registry
21 kafka.schema.registry.url=http://localhost:8081
22
23 # Update job instance state at regular intervals
24 jobInstance.fixedDelay.in.milliseconds=60000
25
26 # Expired time of job instance which is 7 days that is
  604800000 milliseconds.Time unit only supports milliseconds
```

```
27 jobInstance.expired.milliseconds=604800000
28
29 # schedule predicate job every 5 minutes and repeat 12 times
   at most
30 #interval time unit s:second m:minute h:hour d:day,only
   support these four units
31 predicate.job.interval=5m
32 predicate.job.repeat.count=12
33
34 # external properties directory location
35 external.config.location=
36
37 # external BATCH or STREAMING env
38 external.env.location=
39
40 # login strategy ("default" or "ldap")
41 login.strategy=default
42
43 # ldap
44 ldap.url=ldap://hostname:port
45 ldap.email=@example.com
46 ldap.searchBase=DC=org,DC=example
47 ldap.searchPattern=(sAMAccountName={0})
48
49 # hdfs default name
50 fs.defaultFS=
51
52 # elasticsearch
53 elasticsearch.host=linux122
54 elasticsearch.port=9200
55 elasticsearch.scheme=http
56
57 # elasticsearch.user = user
58 # elasticsearch.password = password
59
60 # livy
61 livy.uri=http://localhost:8998/batches
62 livy.need.queue=false
63 livy.task.max.concurrent.count=20
64 livy.task.submit.interval.second=3
65 livy.task.appId.retry.count=3
66
67 # yarn url
68 yarn.uri=http://linux123:8088
69
```

```
70 # griffin event listener
71 internal.event.listeners=GriffinJobEventHook
```

备注:

- 默认端口是8080, 为避免和spark端口冲突, 这里端口修改为9876
- 需要启动Hive的 metastore 服务
- 如果Griffin、MySQL没有安装在同一节点, 请确认用户有权限能够远程登录

3、修改配置文件 service/src/main/resources/quartz.properties

```
1 # 将第26行修改为以下内容:
2 org.quartz.jobStore.driverDelegateClass=org.quartz.impl.jdbcjob
  store.StdJDBCDelegate
```

4、修改配置文件 service/src/main/resources/sparkProperties.json

sparkProperties.json 在测试环境下使用:

```
1 {
2   "file": "hdfs:///griffin/griffin-measure.jar",
3   "className": "org.apache.griffin.measure.Application",
4   "name": "griffin",
5   "queue": "default",
6   "numExecutors": 2,
7   "executorCores": 1,
8   "driverMemory": "1g",
9   "executorMemory": "1g",
10  "conf": {
11    "spark.yarn.dist.files": "hdfs:///spark/spark_conf/hive-
    site.xml"
12  },
13  "files": [
14  ]
15 }
```

备注: 修改第11行: "spark.yarn.dist.files": "hdfs:///spark/spark_conf/hive-site.xml"

sparkProperties.json 在生产环境中根据实际资源配置进行修改【生产环境】

```

1  {
2    "file": "hdfs:///griffin/griffin-measure.jar",
3    "className": "org.apache.griffin.measure.Application",
4    "name": "griffin",
5    "queue": "default",
6    "numExecutors": 8,
7    "executorCores": 2,
8    "driverMemory": "4g",
9    "executorMemory": "5g",
10   "conf": {
11     "spark.yarn.dist.files": "hdfs:///spark/spark_conf/hive-
site.xml"
12   },
13   "files": [
14   ]
15 }

```

5、修改配置文件 service/src/main/resources/env/env_batch.json

```

1  {
2    "spark": {
3      "log.level": "WARN"
4    },
5    "sinks": [
6      {
7        "type": "CONSOLE",
8        "config": {
9          "max.log.lines": 10
10       }
11     },
12     {
13       "type": "HDFS",
14       "config": {
15         "path": "hdfs:///griffin/persist",
16         "max.persist.lines": 10000,
17         "max.lines.per.file": 10000
18       }
19     },
20     {
21       "type": "ELASTICSEARCH",
22       "config": {
23         "method": "post",
24         "api": "http://liunx122:9200/griffin/accuracy",

```

```

25         "connection.timeout": "1m",
26         "retry": 10
27     }
28 }
29 ],
30 "griffin.checkpoint": []
31 }

```

备注：仅修改第24行

6、编译

```

1 cd /opt/lagou/servers/griffin-0.5.0
2 mvn -Dmaven.test.skip=true clean install

```

备注：

- 编译过程中需要下载500M+左右的jar，要将Maven的源设置到阿里
- 如果修改了前面的配置文件，需要重新编译

7、修改文件

编译报错：

```

[ERROR] ERROR in /opt/lagou/servers/griffin-
0.5.0/ui/angular/node_modules/@types/jquery/JQuery.d.ts (4137,26): Cannot
find name 'SVGElementTagNameMap'. [ERROR] ERROR in
/opt/lagou/servers/griffin-
0.5.0/ui/angular/node_modules/@types/jquery/JQuery.d.ts (4137,89): Cannot
find name 'SVGElementTagNameMap'.

```

这个文件在编译之前是没有的

```

/opt/lagou/servers/griffin-
0.5.0/ui/angular/node_modules/@types/jquery/JQuery.d.ts

```

删除 4137 行

```

1 find<K extends keyof SVGElementTagNameMap>(selector_element: K
  | JQuery<K>): JQuery<SVGElementTagNameMap[K]>;

```

8、再次编译

```
1 | cd /opt/lagou/servers/griffin-0.5.0
2 | mvn -Dmaven.test.skip=true clean install
```

9、jar拷贝

编译完成后，会在service和measure模块的target目录下分别看到 service-0.5.0.jar 和 measure-0.5.0.jar 两个jar，将这两个jar分别拷贝到服务器目录下。

```
1 | # 将 service-0.5.0.jar 拷贝到 /opt/lagou/servers/griffin-0.5.0/
2 | cd /opt/lagou/servers/griffin-0.5.0/service/target
3 | cp service-0.5.0.jar /opt/lagou/servers/griffin-0.5.0/
4 |
5 | # 将 measure-0.5.0.jar 拷贝到 /opt/lagou/servers/griffin-
6 | 0.5.0/, 并改名
7 | cd /opt/lagou/servers/griffin-0.5.0/measure/target
8 | cp measure-0.5.0.jar /opt/lagou/servers/griffin-0.5.0/griffin-
9 | measure.jar
10 |
11 | # 将 griffin-measure.jar 上传到 hdfs:///griffin 中
12 | cd /opt/lagou/servers/griffin-0.5.0
13 | hdfs dfs -mkdir /griffin
14 | hdfs dfs -put griffin-measure.jar /griffin
```

备注：spark在yarn集群上执行任务时，需要到HDFS的/griffin目录下加载griffin-measure.jar，避免发生类org.apache.griffin.measure.Application找不到的错误。

4.7 启动Griffin服务

启动Griffin管理后台：

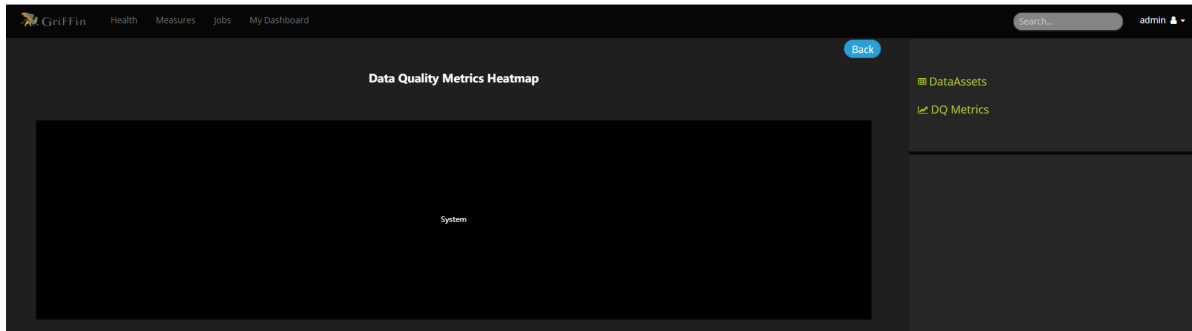
```
1 | cd /opt/lagou/servers/griffin-0.5.0
2 | nohup java -jar service-0.5.0.jar>service.out 2>&1 &
```

Apache Griffin的UI: <http://linux122:9876>

用户名口令：admin / admin



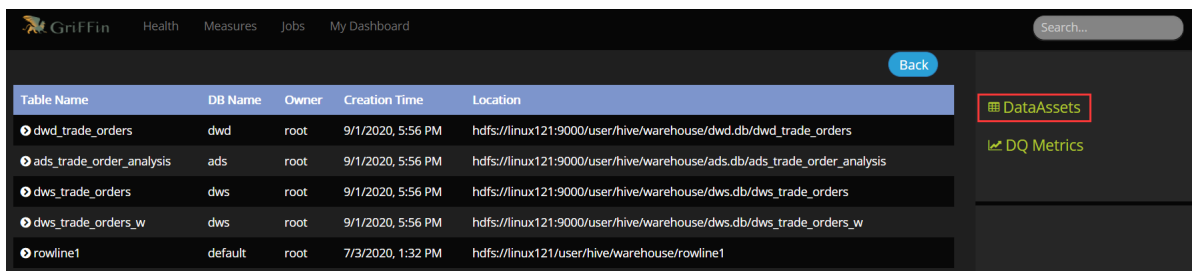
登录后的界面：



第5节 与电商业务集成

5.1 数据资产

单击右上角的 DataAssets 来检查数据资产

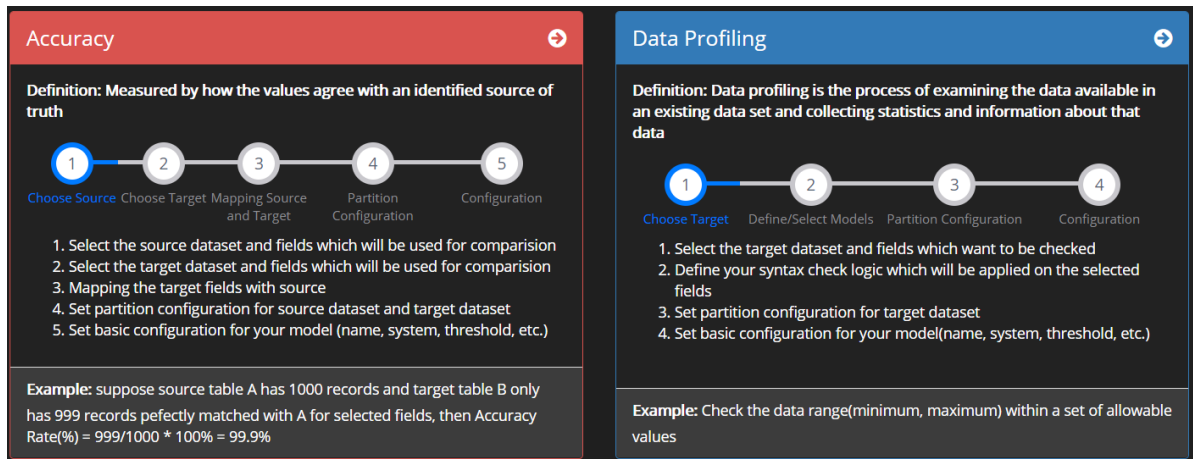


备注：这里的数据数据资产主要是保存在Hive上的表，要求 Hive Metastore 服务正常

5.2 创建 measure

- 如果要测量源和目标之间的匹配率，请选择 **Accuracy**（精确度验证）
- 如果要检查数据的特定值（例如：空列计数），请选择 **Data Profiling**（数据统计分析）
 - 统计表的特定列里面值为空、唯一或是重复的数量

- 统计最大值、最小值、平均数、中值等
- 用正则表达式来对数据的频率和模式进行分析



核心交易分析中有两张表：

- dws_trade_orders（订单明细）
- dws_trade_orders_w（订单明细宽表）

这两张表的数据量应该是相等的（Accuracy）

计算ODS层

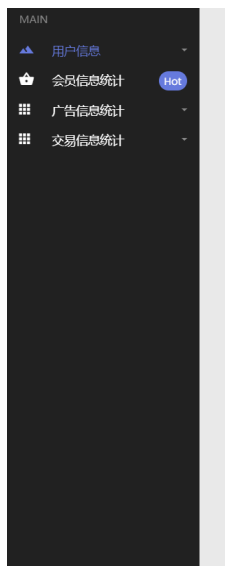
- ods_trade_orders(订单表)

订单表的数据量(Data Profiling)

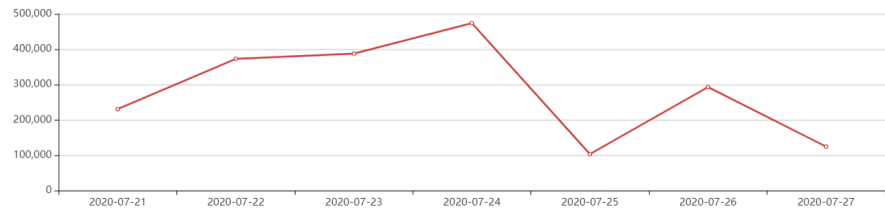
第五部分 数据可视化

ADS => DataX => MySQL => 浏览器呈现

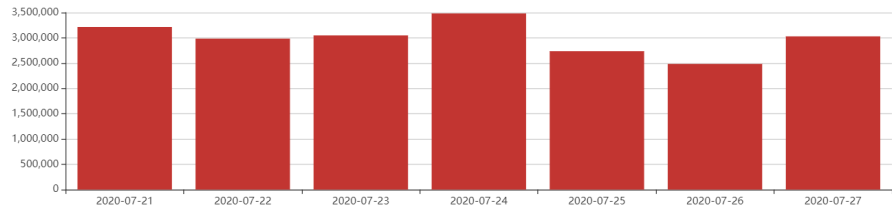
对统计数据展示，一般都是以图、表方式呈现；常见方式有 ECharts、HighCharts、G2、Chart.js、FineBI等。本项目使用SSM（Spring + SpringMVC + MyBatis）、ECharts。



统计7天会员增长

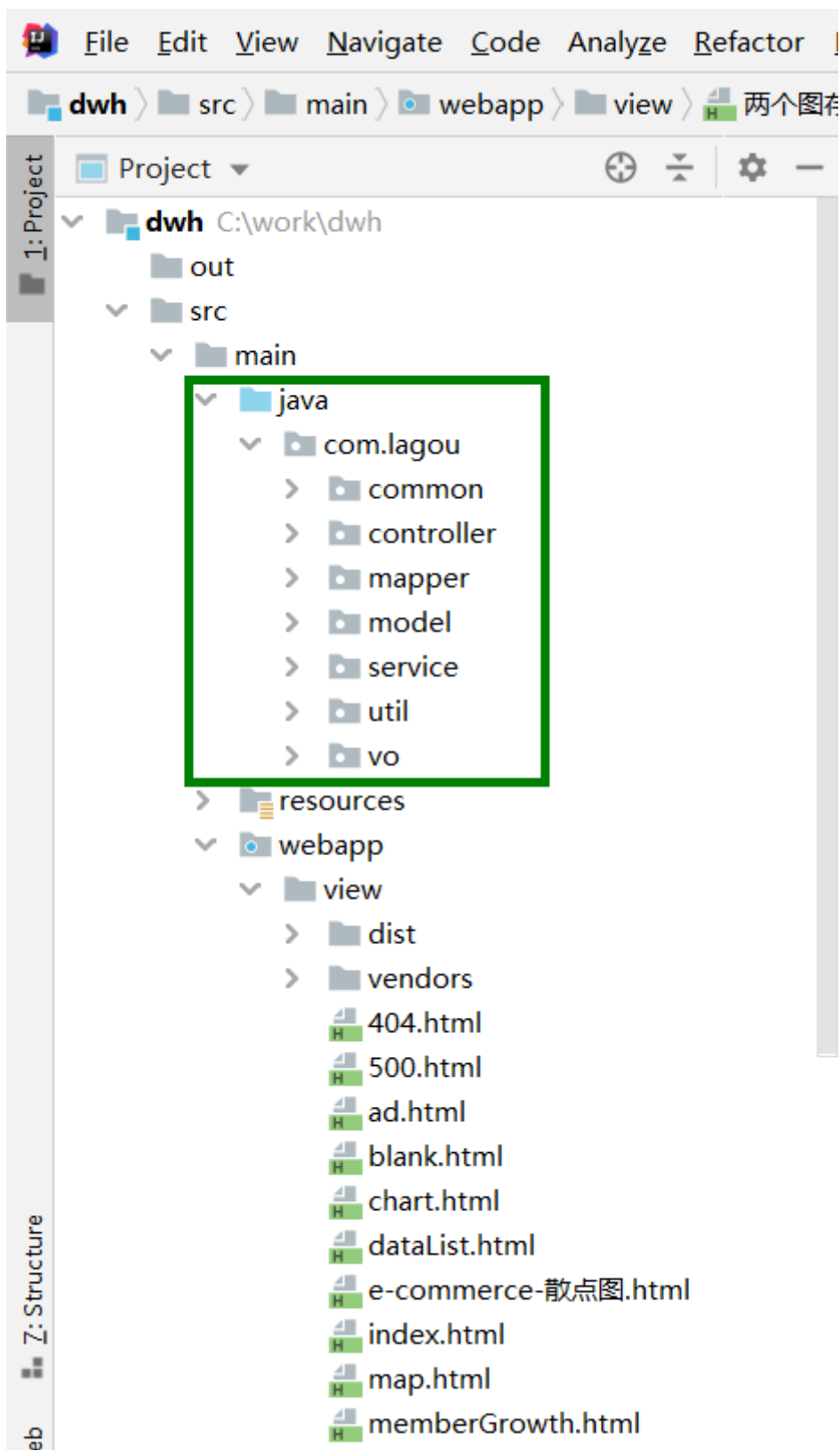


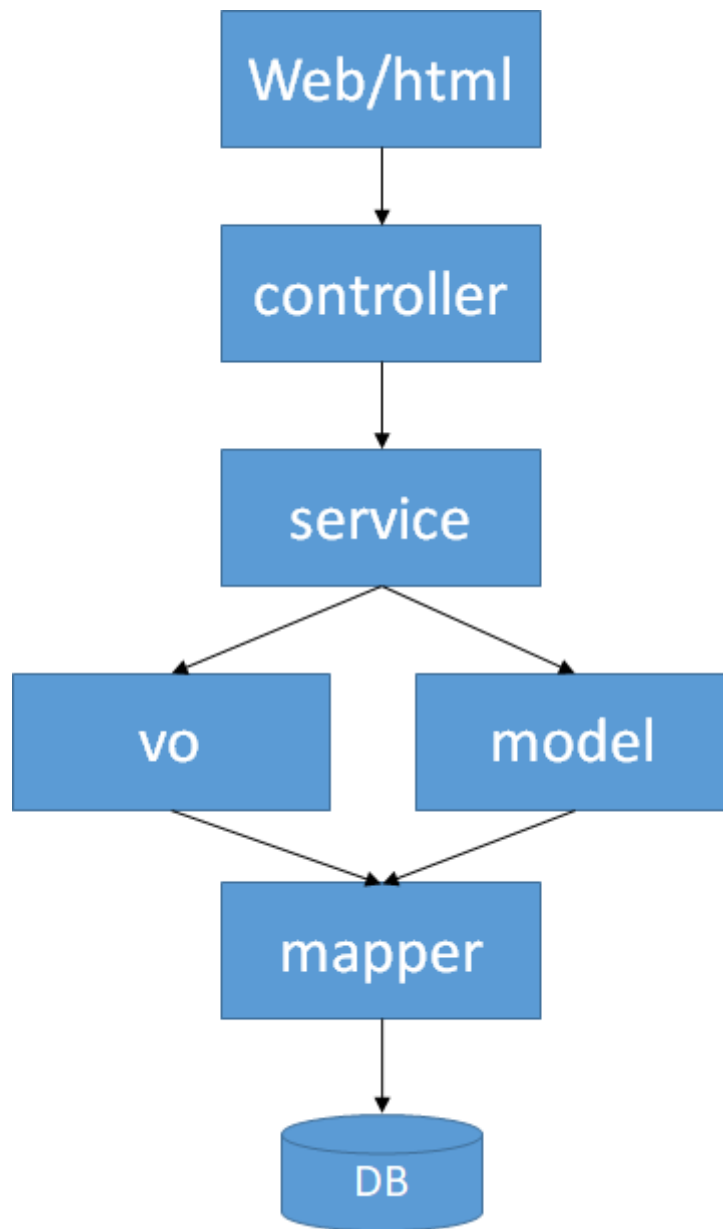
统计7天活跃会员



备注：

- src/main/resources/jdbc.properties: jdbc连接信息
- src/main/resources/log4j.properties: 日志存放位置
- MySQL中创建user表, 添加信息
- <http://localhost:8080/dwh/login.html>
- admin/admin





项目总结与回顾

1、数据仓库概念

数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。

OLAP（数据仓库）与OLTP（数据库）的区别；

数据仓库分层：ODS、DWD、DWS、ADS

为什么要分层：

- 清晰的数据结构
- 将复杂的问题简单化

- 减少重复开发
- 屏蔽原始数据的异常
- 数据血缘的追踪

数据仓库建模：维度建模、ER建模

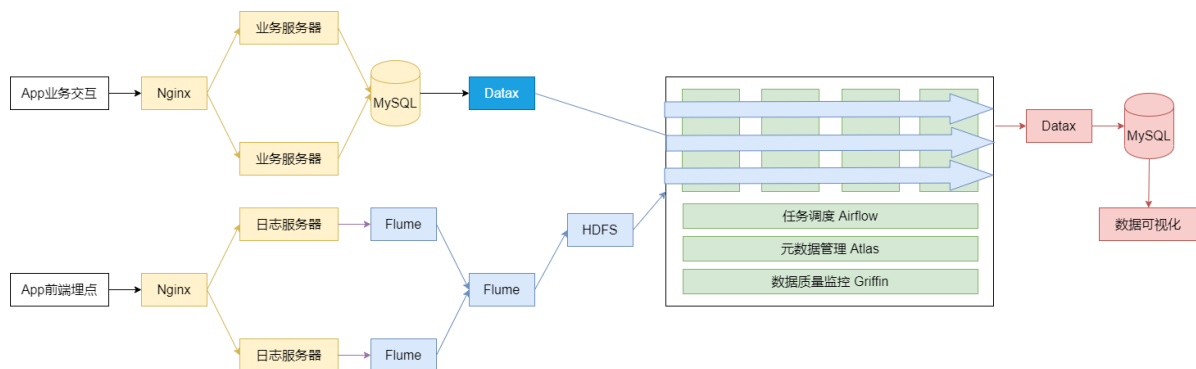
维度建模的4个步骤：

- 选择业务
- 定义粒度
- 选定维度
- 确定事实

集群的规划：

- 集群可以做水平扩展
- 初始时可依据数据量估算集群规模

框架版本的选型：CDH国内选用最多的版本



2、数据采集模块

Flume采集日志数据、DataX采集业务数据（数据的全量或增量）；

Flume组成、Put事务(Source到Channel是Put事务)、Take事务(Channel到Sink是Take事务)

Taildir Source：断点续传、监控多目录。Flume1.6以前需要自己自定义Source记录每次读取文件位置，实现断点续传。

File Channel：数据存储在磁盘，宕机数据可以保存。但是传输速率慢。适合对数据传输可靠性要求高的场景，比如，金融行业；

Memory Channel：数据存储在内存中，宕机数据丢失。传输速率快。适合对数据传输可靠性要求不高的场景，比如，普通的日志数据；

Kafka Channel：减少了Flume的Sink阶段，提高了传输效率；

HDFS Sink：如何避免小文件(HDFS文件的滚动方式)

Flume自定义拦截器：

- initialize 初始化
- intercept(Event event) 处理单个Event 【实现的重点】
- intercept(List events) 处理多个Event
- close 方法

设置Agent JVM heap为4G或更高，部署在单独的服务器上；

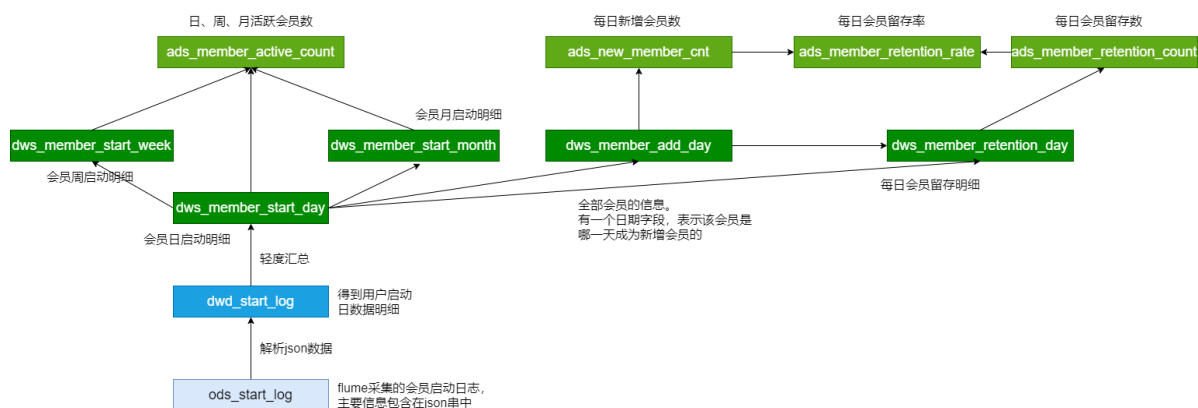
-Xmx与-Xms设置一致，减少内存抖动带来的性能影响，设置不一致容易导致频繁full gc；

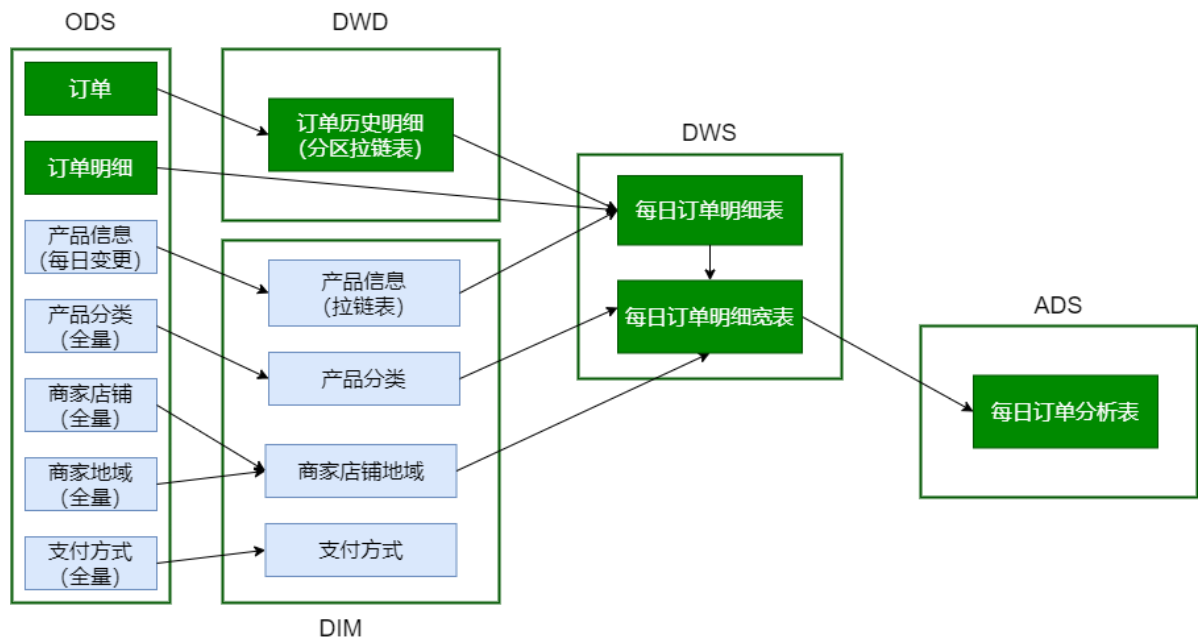
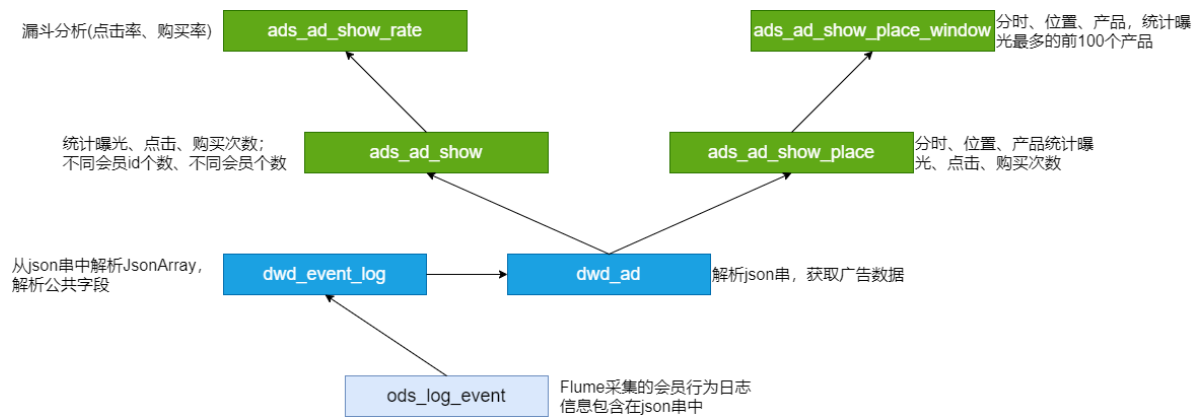
3、主题分析模块【重点】

会员活跃度分析、广告业务分析、核心交易分析；

Json数据的处理、动态分区、拉链表、宽表(逆规范化)、Tez引擎（缺点：对资源要求高）

ODS、DWD、DWS、ADS、DIM各层模型如何建立；





4、调度系统

5、元数据管理数据、数据质量监控 (扩展)

6、数据可视化