商品销售数据分析实验

一. 案例信息

1. 实验内容

现有商品销售订单表数据如下：

```
5349043,2015-05-22 06:16:28,河北省,丰宁满族自治县,7,1,125
5339045,2015-05-01 23:56:20,江苏省,连云港市,2,3,30
5339046,2015-05-31 15:50:40,浙江省,安吉县,8,2,22
5339047,2015-05-04 22:05:00,西藏,昂仁县,4,3,24
5339048,2015-05-31 21:47:18,江苏省,盱眙县,4,2,16
5339049,2015-05-22 01:31:47,山东省,泗水县,6,2,20
5339050,2015-05-24 20:09:26,广东省,鹤山市,6,1,10
5339051,2015-05-25 05:15:38,西藏,工布江达县,9,3,435
5339052,2015-05-24 12:35:45,内蒙古,卓资县,9,3,435
5339053,2015-06-01 01:45:10,浙江省,宁波市,8,4,44
5339054,2015-05-13 01:57:56,广东省,广州市,2,1,9
5339055,2015-05-22 01:38:12,山东省,单县,6,1,10
5339056,2015-05-19 23:54:48,山东省,淄博市,10,3,15
```

字段从左致右分别为：订单编号，销售日期，省份，城市，商品编号，销量，销售额

商品详细表如下：

```
1,product_a,1,Category A,99.9
2,product_b,2,Category B,9.9
3,product_c,1,Category A,89.9
4,product_d,2,Category B,8.9
5,product_e,1,Category A,100
6,product_f,2,Category B,10
7,product_g,1,Category A,121
8,product_h,2,Category B,11
9,product_i,1,Category A,145
10,product_j,2,Category B,5
11,product_k,3,Category New,998.0
```

字段从左致右分别为：商品编号，商品名称，分类编号，分类名称，商品价格

2. 实验目的

- 掌握Hive表创建的方法
- 掌握Hive数据加载的方法
- 掌握Hive连接查询及子查询的方法
- 掌握Hive窗口函数、聚合函数的使用方法

3. 实验环境

- Hadoop == 3.1.0
- CentOS == 8
- Hive == 3.1.2

二. 实验指导

1. 关联技术

无

2. 实验步骤

- 非分区表数据加载
- 分区表数据加载（静态分区）
- 分区表数据加载（动态分区）
- 分桶表数据加载

3. 实验效果

三. 实验操作

01. 步骤一：表创建及数据加载

步骤操作说明

1. 创建表

销售订单表：

```
create table t_dml (
  detail_id bigint,
  sale_date date,
  province string,
  city string,
  product_id bigint,
  cnt bigint,
  amt double
)row format delimited
fields terminated by ',';
```

商品详细表：

```
create table t_product (
    product_id bigint,
    product_name string,
    category_id bigint,
    category_name string,
    price double
)row format delimited
fields terminated by ',';
```

2. 加载数据

```
load data local inpath '/opt/data/t_dml.csv' into table t_dml;
load data local inpath '/opt/data/t_product.csv' into table t_product;
```

02. 步骤二：销售数据分析

步骤操作说明

1. 查询t_dml中的销售记录的时间段：

```
select max(sale_date), min(sale_date) from t_dml;
```

2. 查询各商品类别的总销售额

```
select t.category_name, sum(t.amt) as total_money
 from
 ( select a.product_id, a.amt, b.category_name
 from t_dml a
 join t_product b
 on a.product_id=b.product_id
 ) t
group by t.category_name;
```

3. 查询售量排行榜

店主想知道哪个商品最畅销以及销量排行榜，请查询销量前10的商品，显示商品名称，销量，排名。

```
select a.product_name , t.cnt_total,
 rank() over (order by t.cnt_total desc) as rk
 from
 ( select product_id, sum(cnt) as cnt_total
 from t_dml
 group by product_id
 order by cnt_total desc
 limit 10
 ) t
 join t_product a
 on t.product_id=a.product_id;
```

03. 步骤三：创建中间表

店主想知道各个市县的购买力，同时也想知道自己的哪个商品在该地区最热卖，通过创建中间表，优化查询。

步骤操作说明

1. 创建结果存放表：

```
create table t_city_amt
( province string,
 city string,
 total_money double
);
create table t_city_prod
( province string,
 city string,
 product_id bigint,
 product_name string,
 cnt bigint
);
```

2. 插入数据

```
insert into t_city_amt
select province,city,sum(amt)
from t_dml group by province,city
```

```
insert into t_city_prod
select t.province,t.city,t.product_id,t.product_name,sum(t.cnt) from
(
select a.product_id,b.product_name,a.cnt,a.province,a.city
from t_dml a join t_product b
on a.product_id = b.product_id
) t
group by t.province,t.city,t.product_id,t.product_name
```

3. 优化

```
from
( select a.*, b.product_name
 from t_dml a
 join t_product b
 on a.product_id=b.product_id
) t
insert overwrite table t_city_amt
 select province, city, sum(amt)
 group by province, city
insert overwrite table t_city_prod
 select province, city, product_id, product_name, sum(cnt)
 group by province, city, product_id, product_name;
```

04. 步骤四：统计指标

步骤操作说明

1. 统计各省最强购买力地区：

```
select province, city, total_money
 from
 (
  select province, city, total_money,
  dense_rank() over (partition by province order by total_money desc) as rk
  from t_city_amt
 ) t
where t.rk=1
order by total_money desc;
```

2. 统计各地区的最畅销商品

```
select province, city, product_id, product_name
 from
 ( select province, city, product_id, product_name,
 dense_rank() over (partition by province order by cnt desc) as rk
 from t_city_prod
 ) t
where t.rk=1
order by province, city;
```

四. 实验扩展

无

1. 创新扩展点

无

2. 创新实现步骤

01. 步骤一: 无

02. 步骤二:无

五. 附录

技术资料

- 官方文档:https://hadoop.apache.org/docs/r1.0.4/cn/quickstart.html

---