电商数据仓库构建

一、数据仓库

定义：数据仓库（ Data Warehouse ）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，主要用于支持管理决策。

特点：

面向主题——数据仓库是针对一个大的领域主题（如销售情况）而设计，不是根据具体某个功能。

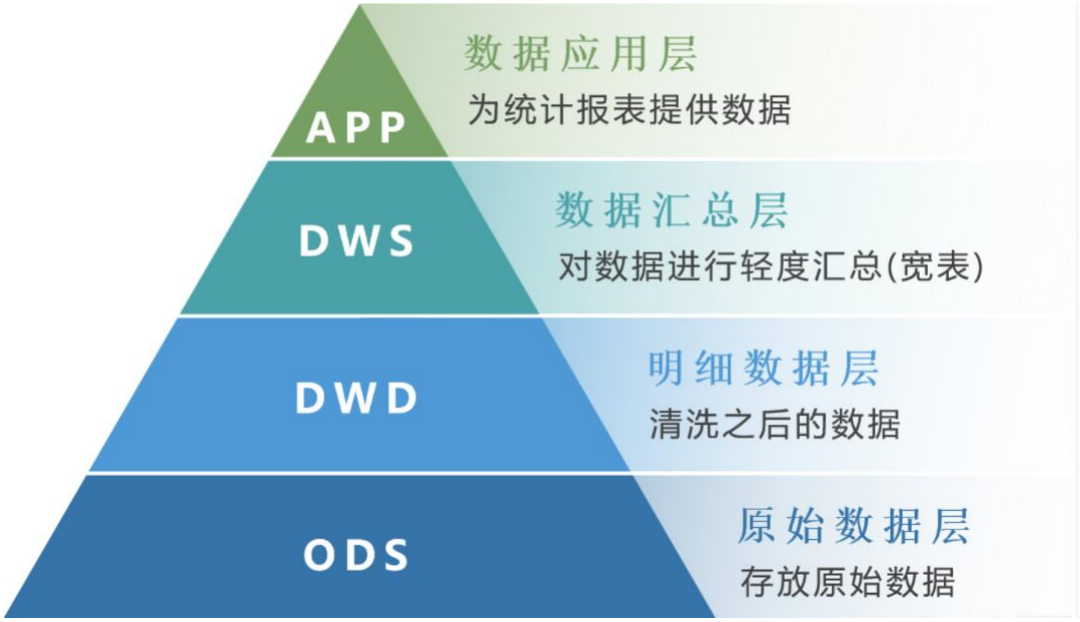
集成——数据仓库中的数据由多个不同来源的数据集加工集成而来。

相对稳定——数据仓库构建后，不会频繁地更新数据。

反映历史变化——数据仓库保存的是历史数据，不是实时数据。

用于支持管理决策——将现有数据分析和整理，从中得出某种趋势、结论，为决策的拟定提供数据支持。

二、数据仓库分层结构



ODS层：从数据源中抽取数据后放入本层，未经过清洗

DWD层：对ODS数据层做一些数据的清洗和规范化的操作，比如去除空数据、脏数据、离群值等，形成维度表、事实表

DWS层： 整合汇总数据，分析某一个主题域的数据服务层 ，形成主题大表

ADS（APP）层： 提供用于数据分析、数据可视化等任务的数据

三、电商数据仓库构建

1、数据采集（ODS层）

本项目采用天猫天池双十一公开数据集Tianchi User Behavior Dataset，包含以下两张表：

tianchi_fresh_comp_train_item.csv：

列名	说明
item_id	商品编号
item_geohash	地理位置（多数为空）
item_category	商品类别

tianchi_fresh_comp_train_user.csv

列名	说明
user_id	用户编号
item_id	商品编号
behavior_type	行为类型（1点击 2收藏 3加购物车 4购买）
user_geohash	地理位置（多数为空）
item_category	商品类别
time	操作时间

2、数据预处理（DWD层）

为了使数据量尽量压缩、根据问题需要以及操作方便，使用pandas对数据进行预处理

通用预处理方案参考：

（1）去除关键列为空的数据，如行为类别、商品类别等

（2）将日期列拆成年、月、日三列，方便分时段统计

（3）从商品表中去除完全没有用户交互过的商品

（4）同一用户同一天对同一个商品进行重复行为的，将次数合并，缩减行数

（5）将地理位置列的空白填满（与上方最近一个值相同）

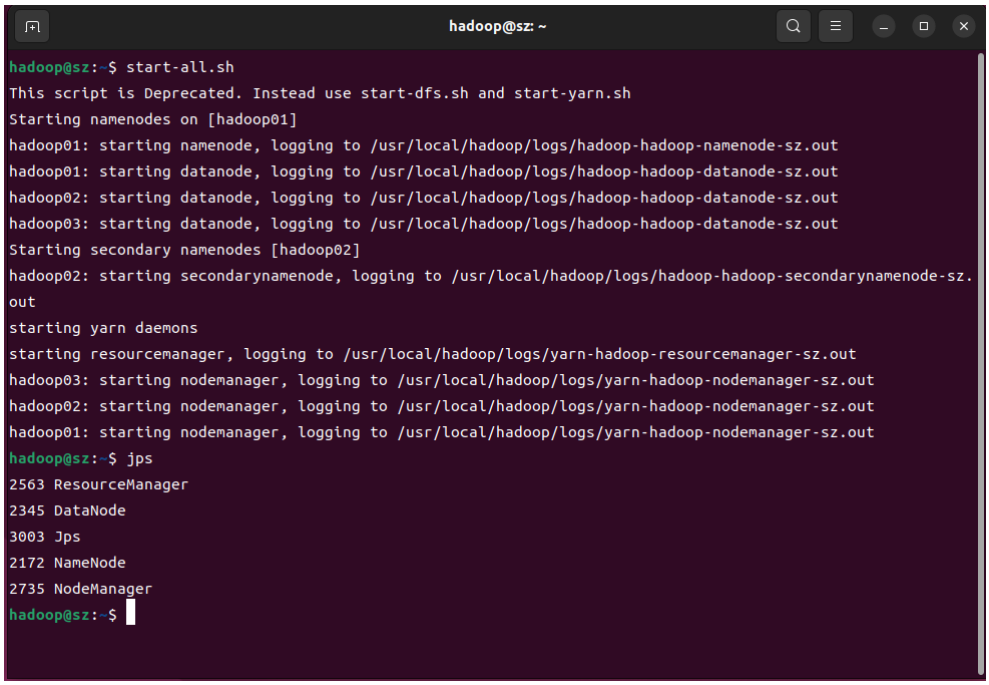
（6）如果要建分区表，需要将数据文件按分区字段拆成多个小文件

其他预处理方案需根据实际问题需要而定。

假设处理好后的文件为 clean_item.csv 和clean_user.csv，放在一个data文件夹中

3、数据导入

打开虚拟机并启动hadoop



数据文件较大时，可不需要拷贝到虚拟机内，直接通过共享文件夹上传

重新挂载共享文件夹，并将当前用户设为所有者：

```
sudo vmhgfs-fuse .host:/ /mnt/hgfs -o allow_other -o uid=$(id -u) -o gid=$(id -
g)
```

将处理好的数据文件复制到外部共享文件夹，查看确认：

```
sudo ls /mnt/hgfs/VMShare
```

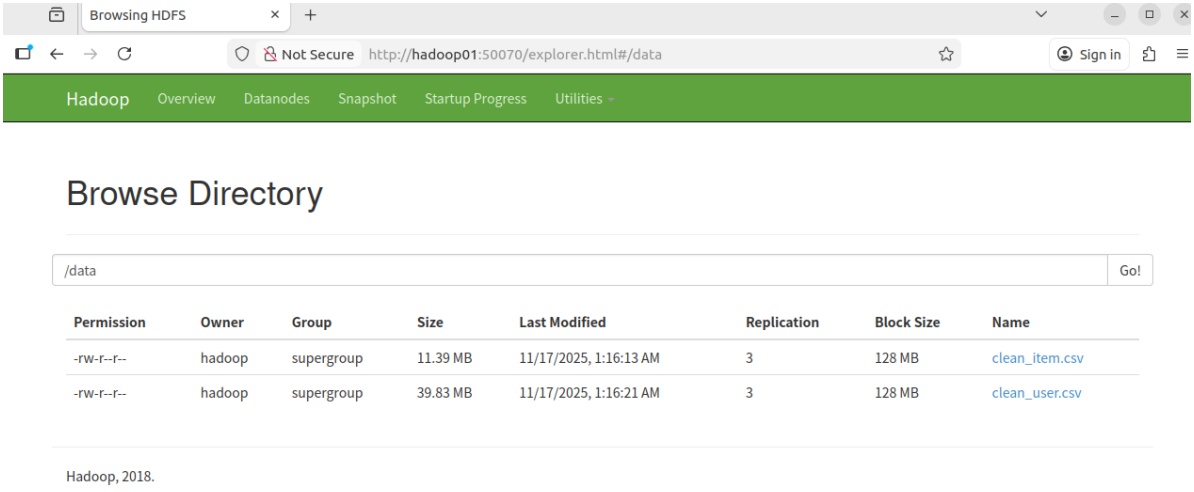
上传文件到HDFS，先创建一个文件夹

```
hdfs dfs -mkdir /data
```

上传

```
hdfs dfs -put /mnt/hgfs/VMShare/clean_item.csv /data
hdfs dfs -put /mnt/hgfs/VMShare/clean_user.csv /data
```

可到hdfs网页端查看是否上传成功，地址：hadoop01:50070



4、创建Hive表

（1）启动hive，创建个人数据库并使用；

（2）建商品表item、用户行为表behavior，要求建为**外部表**

Tips：在建表语句末尾加tblproperties参数，该表在后续导入数据时将跳过首行

```
create 建表语句... tblproperties ("skip.header.line.count"="1");
```

（3）导入数据到表（从hdfs导入，如果有分区小数据文件，需将不同类型商品放到不同分区，可用代码批量生成语句）

（4）简单查询测试

例：

查询某个用户所有行为

查询某天（如双十二）所有购买行为

统计各类商品购买数

……

自行设计问题进行查询（5个以上）