厦門大學
XIAMEN UNIVERSITY

数据仓库大作业
中期汇报

——报告展示

小组成员：苏一涵、陈昊睿、刘灵菲、陆浩楠

指导老师：王鸿吉

by：苏一涵

目录

项目名称：

对电商用户行为数据的分析

小组成员及分工：

苏一涵：需求分析、概念设计（事实表，维度表）和逻辑设计明确各表字段、数据类型、分区规则，组织调配整体项目任务，汇报

陈昊睿：确认数据源，负责数据的ETL脚本开发与数据加载

刘灵菲：数据分析可视化，负责编写分析 SQL完成可视化图表与结论报告

陆浩楠：负责实验项目的整合测试，完成最终的汇报ppt

项目背景简要介绍：

随着数字经济的快速发展，电商行业已成为零售领域的核心支柱，用户行为数据成为平台优化运营、提升转化效率的核心资产。淘宝作为国内领先的电商平台，积累了海量用户浏览、收藏、加购、购买等行为数据，这些数据蕴含着用户消费偏好、商品热度分布、转化路径规律等关键业务信息，对精准营销、商品运营、用户留存具有重要指导意义。

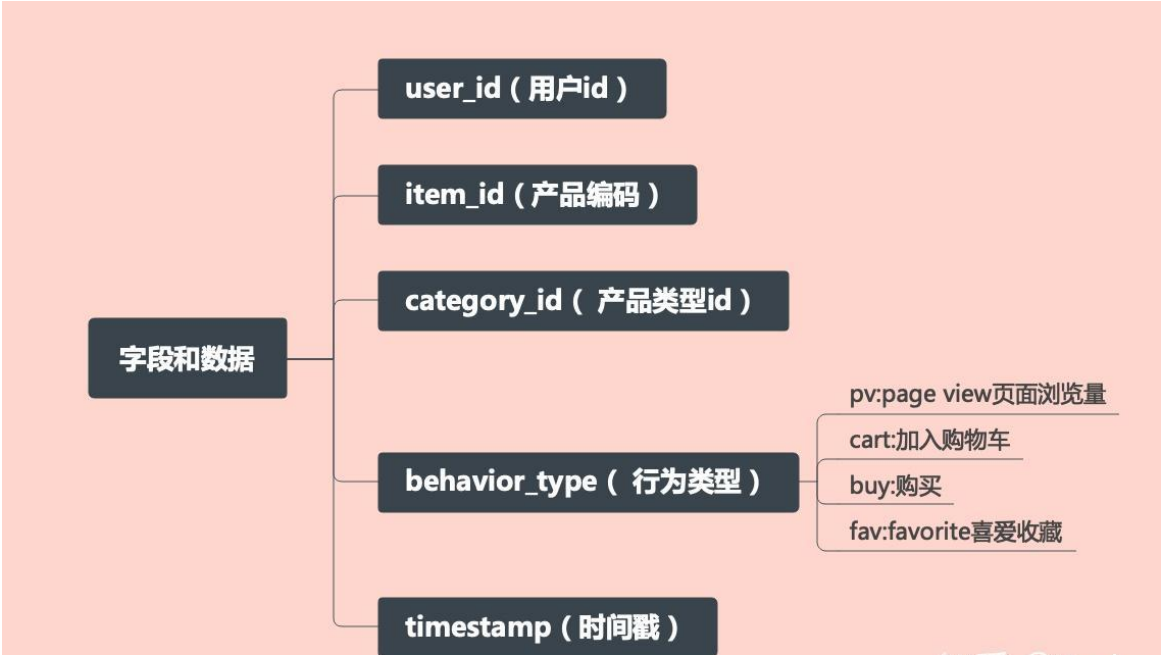
项目背景简要介绍：

本项目选取阿里天池公开数据集《淘宝用户行为数据集》来进行分析

天池实验室 > 数据集 > 公共数据集 > 正文

淘宝用户购物行为数据集

作为核心数据源，该数据集包含 1000 万 + 条真实用户行为记录，涵盖用户 ID（userid）、商品 ID（itemid）、商品类目 ID（categoryid）、行为类型（type：pv 浏览、buy 购买、cart 加购、fav 收藏）、时间戳（timestamp）五大核心字段，数据真实且覆盖电商核心业务场景。



项目背景简要介绍：

基于该数据集构建电商用户行为分析数据仓库，旨在通过数仓分层设计、ETL 数据处理、多维分析等技术手段，系统挖掘用户行为规律、商品热度特征及转化效率瓶颈，为电商平台的运营决策（如精准推荐、活动策划、库存优化）提供数据支撑，也对我们这学期ooad的项目设计有所帮助，同时践行数据仓库 “需求分析→设计→实现→分析” 的完整开发流程，达成课程实践目标。

项目设计和开发工具列表：

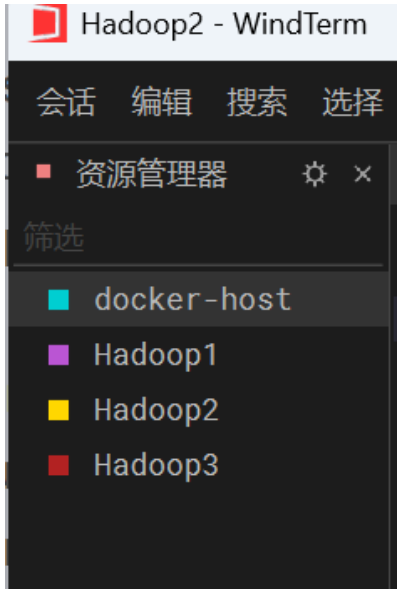
1

StarUML

用于绘制事实表和维度表

2

WindTerm



管理Hadoop集群

3

Hive-on-Spark环境

在 Hive 中集成 Spark，Hive 既作为元数据存储，又负责解析 HQL 语句，将 Hive 的运行引擎更换为 Spark，速度更快

4

MySQL：数据库管理

DataX：将异构数据同步到MySQL数据库

QuickBI：完成数据可视化处理

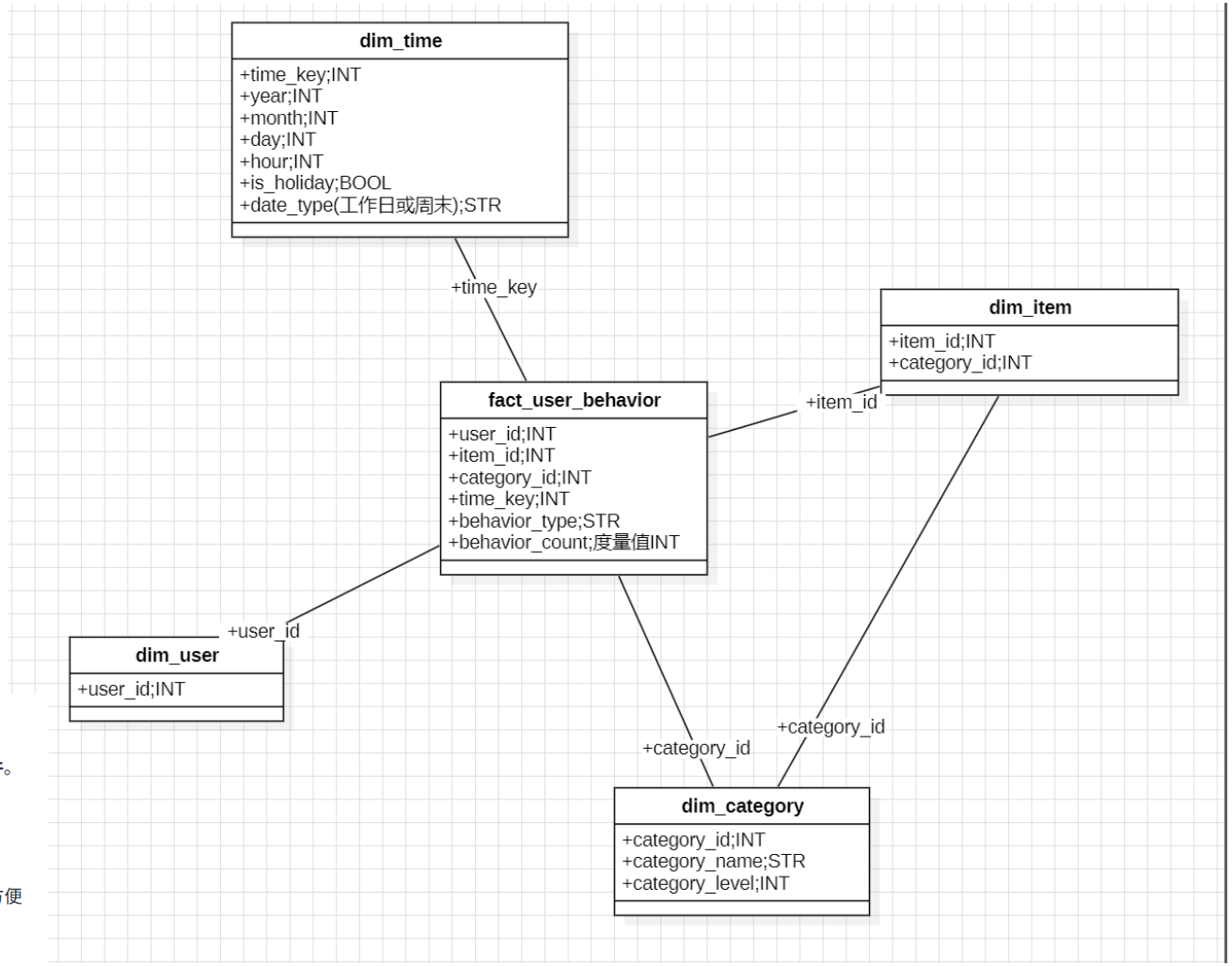
等

项目进度：

已经完成项目分工

根据数据源和需求分析初步构建好事实表和维度表

因为天池的数据集商品只有id，所以商品类目这边暂定引入《淘宝商品类目数据集》



淘宝商品类目数据集详细介绍

此仓库提供了一份宝贵的资源文件——**淘宝商品类目数据共831247条，2019年全部淘宝分类MYSQL导出文件**。以下是该数据集的详细介绍：

数据集概述

本数据集包含2019年8月整理的全部淘宝商品类目，总计831247条记录。数据以MYSQL导出文件形式提供，方便用户直接在数据库环境中使用。

项目进度：

完成对Hadoop集群的配置和初步组建

共使用3台虚拟机作为服务端，用windows本机做为客户端

三台centOS虚拟机的角色分配如下

	hadoop01	hadoop02	hadoop03
角色	主节点	从节点	从节点
NameNode	√		
DataNode	√	√	√
ResourceManager	√		
NodeManager	√	√	√
SecondaryNameNode		√	
Historyserver	√		

具体ETL等之后的过程还在规划

项目进度：

节点名称	角色类型	承担组件及职责
hadoop01	主节点	- NameNode：HDFS 主节点，管理 HDFS 元数据（文件目录、权限等）； - DataNode：存储实际数据（主节点可同时作为数据节点）； - ResourceManager：YARN 主节点，负责集群资源（CPU、内存）的全局调度； - NodeManager：YARN 从节点组件，管理本节点的资源（主节点也可部署）； - HistoryServer：记录并展示 MapReduce 等作业的历史运行信息。
hadoop02	从节点	- DataNode：存储实际数据； - NodeManager：管理本节点资源； - SecondaryNameNode：辅助NameNode合并编辑日志，减轻主节点负担。
hadoop03	从节点	- DataNode：存储实际数据； - NodeManager：管理本节点资源。

感 谢 聆 听

汇报人：苏一涵、陈昊睿、刘灵菲、陆浩楠