黑马程序员《企业级360°全方位用户画像》项目课程介绍

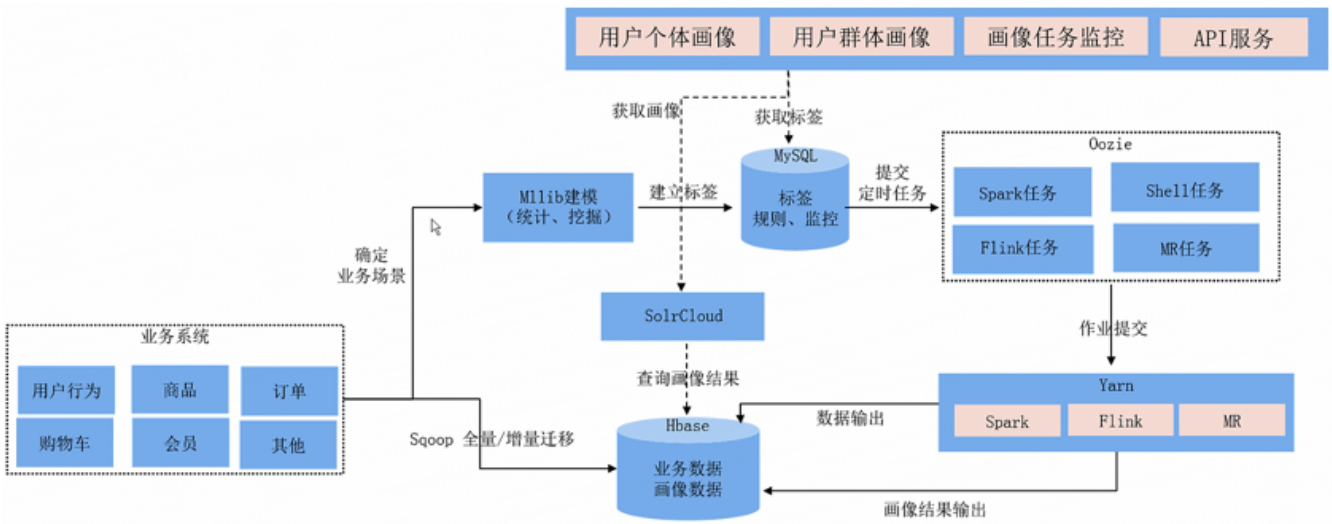
1、项目简介

企业级360°全方位用户画像系统是基于垂直电商平台构建的用户画像，采用了目前流行的大数据技术的实时+离线计算方案。

2、系统架构

2.1、架构图

企业级360°全方位用户画像架构图：



2.2、架构要点

要点：

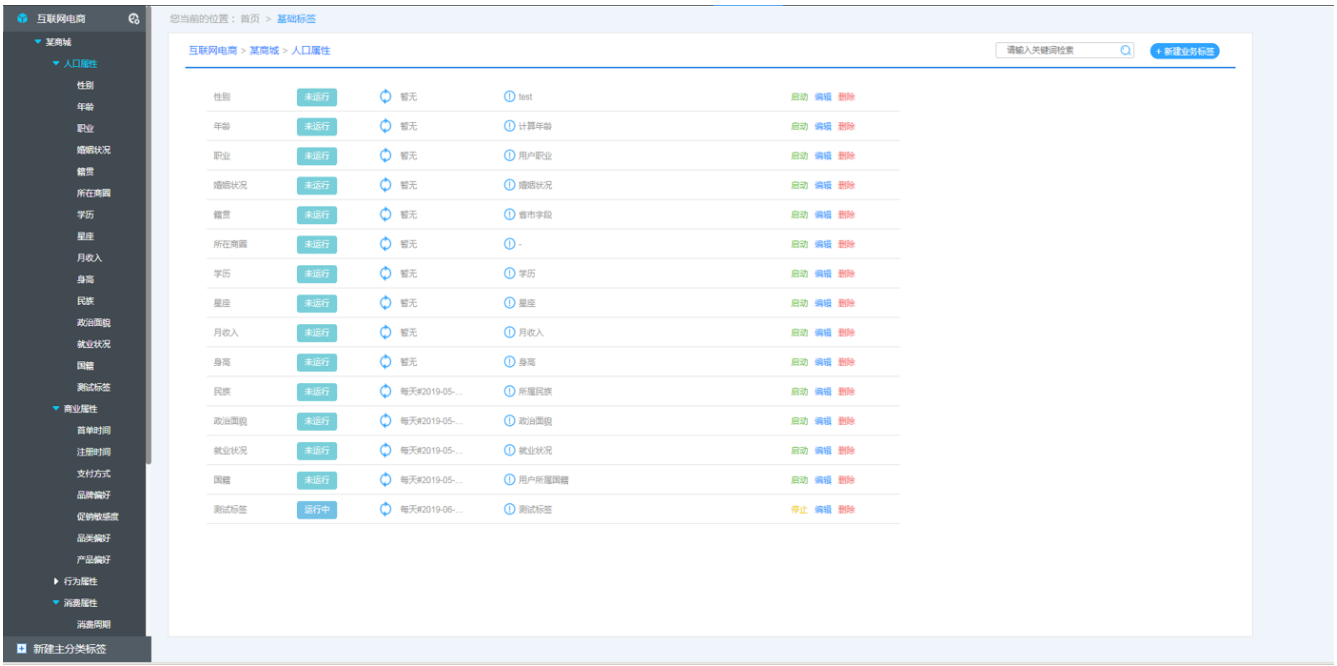
- 通过Sqoop迁移业务数据到HBase。

- 基于数据内容确定业务场景并使用SparkMLlib建模
- 建立标签及其规则关联算法模型
- 确定标签更新周期生成Oozie的定时工作流执行
- YARN执行作业完成后写入画像结果数据到HBase和Solr
- 通过RestAPI查询Solr并实时生成用户画像结果展示

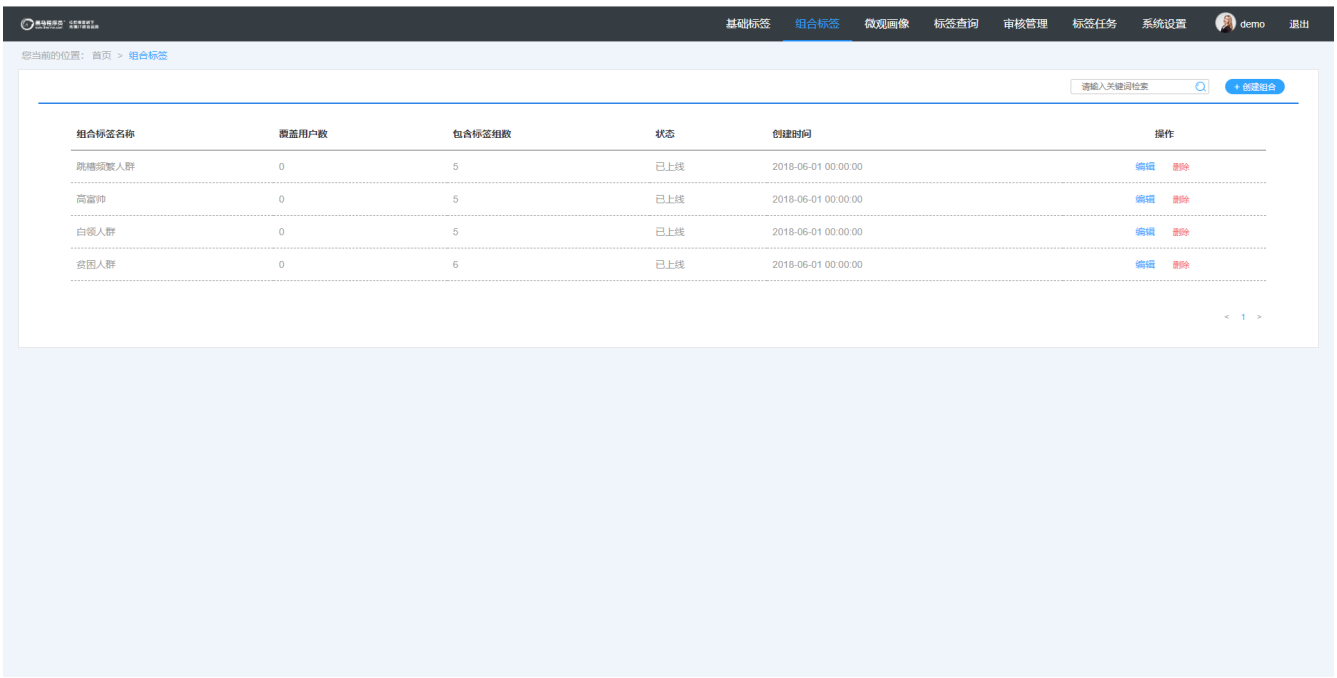
功能模块：

系统由7个模块组成，主要为基础标签、组合标签、微观画像、标签查询、标签任务、审核管理和系统设置。

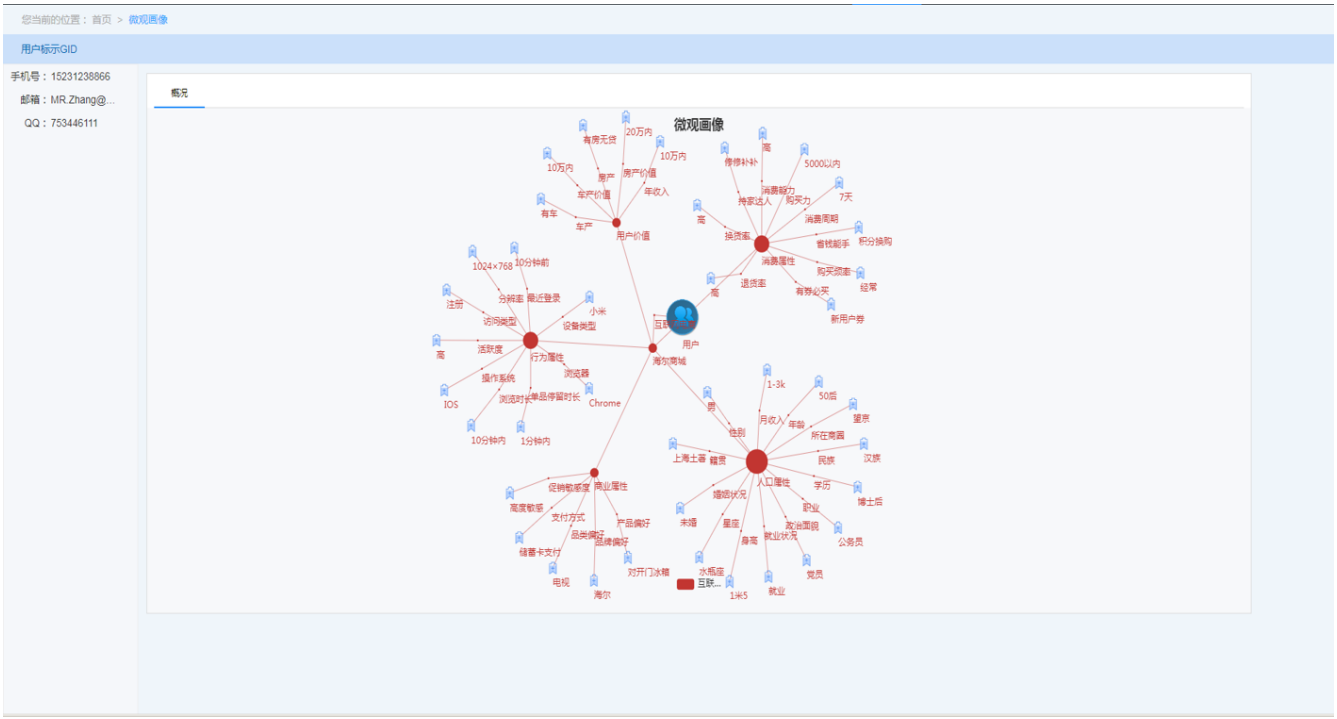
基础标签模块预览图：



组合标签模块预览图：



用户画像模块预览图：



3、技术选型

3.1、相关技术

- 数据迁移工具Sqoop
- 分布式存储和计算平台Hadoop
- 机器学习库Spark MLlib
- SQL on Hadoop方案Spark SQL
- 准实时计算Spark Streaming
- 分布式NoSQL数据库HBase
- 分布式索引和全文检索工具Solr Cloud
- 工作流调度引擎Oozie

3.2、技术解读

1. Sqoop

   Sqoop是一个在结构化数据和Hadoop之间进行批量数据迁移的工具，结构化数据可以是Mysql、Oracle等RDBMS。Sqoop底层用MapReduce程序实现抽取、转换、加载，MapReduce天生的特性保证了并行化和高容错率，而且相比Kettle等传统ETL工具，任务跑在Hadoop集群上，减少了ETL服务器资源的使用情况。在特定场景下，抽取过程会有很大的性能提升。

2. Hadoop

   Hadoop使用简单的编程模型跨计算机集群分布式处理大型数据集。它旨在从单个服务器扩展到数千台计算机，每台计算机都提供本地计算和存储。Hadoop自身不依靠硬件来提供高可用性，而是设计用于检测和处理应用层的故障，从而在计算机集群之上提供高可用性服务。

3. SparkMLlib

   MLlib是Spark的机器学习库，其目标是使机器学习可扩展且简单。它提供了常见的分类，回归，聚类和协同过滤等算法。

4. SparkSQL

   Spark SQL是用于结构化的数据处理，在执行时会在内部进行优化。支持标准SQL和DSL对数据进行查询、加载和写入。

5. SparkStreaming

   Spark Streaming是SparkCore API的扩展，可实现实时数据流的可扩展，高吞吐量，容错流处理。数据可以从许多来源（如Kafka，Flume，Kinesis或TCP）中提取，并且可以使用以高级函数表示的复杂算法进行处理多种算子，处理后的数据可以推送到文件系统，数据库和实时仪表板。

6. HBase

   HBase是建立在Hadoop文件系统之上的分布式面向列的NoSQL数据库。它是一个开源项目，支持横向扩展，可以提供快速随机读写数据。

7. SolrCloud

   Solr具有高可靠性，可扩展性和容错性，可提供分布式索引，复制和负载均衡查询，自动故障转移和恢复，集中配置等。Solr为很多大型联网站点的搜索和导航功能提供支持。SolrCloud是Solr提供的分布式搜索方案，提供超大规模、容错、分布式索引和检索能力。

8. Oozie

   Oozie是一个基于Hadoop平台上的可靠的、可扩展的工作流程调度系统，支持MapReduce、Shell、Hive、Sqoop、Spark等多种类型的作业。