项目架构

项目和技术背景

本项目是一个基于 Apache Spark 的电影推荐系统。系统为用户推荐出最合适的 5 部电影。

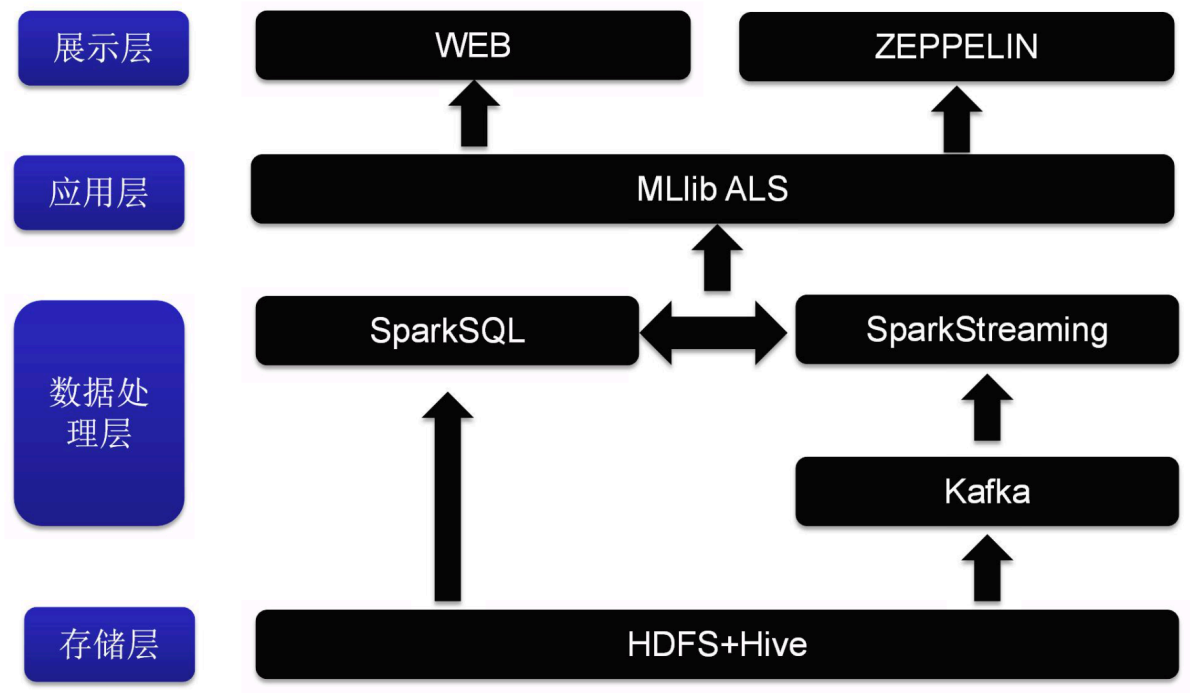
整个系统运用了 Spark, Hadoop, Hive, Kafka 等常用的大数据组件，分为离线推荐和实时推荐 2 个主要的技术路线。

项目架构

在项目中，运用了诸如 Hadoop，Spark，Kafka 等常用的大数据的组件

在项目中，主要有以下几层,以及:

1. 存储层（最底层）: HDFS 作为底层存储（文件系统），Hive 做为数据仓库
    - HDFS 作为文件存储，不管是存储性能还是稳定性还是吞吐量都占用很好的优势，HDFS 是企业必备组件。
    - Hive 是类 SQL 的查询引擎，Hive 作为数据仓库来存在。
2. 离线数据处理: SparkSQL
    - 数据清洗和预处理，后续做推荐的数据准备，对数据进行预处理，处理成模型需要的数据。
3. 实时数据处理: kafka, SparkStreaming
    - 运用到离线处理时的一些技术。
    - 将离线数据以一定的方式通过 kafka 产生一个消息队列，传到消息队列中，然后通过实时的消息处理框架 SaprkStreaming 接收这个消息队列。
    - 对于这个消息队列中的数据实时产生一些推荐结果。
    - 然后把实时数据对接到应用层。
4. 数据应用层: MLlib
    - 接收来自离线处理中的离线数据，将数据切分层训练集和测试集。通过测试集来测试训练的结果来验证模型是否具有可用性。
    - 为各个用户产生推荐结果。
5. 数据展示和对接: Zeppelin
    - 不仅能做数据流，也能做一些数据展示。Zepplin 包含一些图形是展示。



主要模块

在项目中，主要有以下的模块

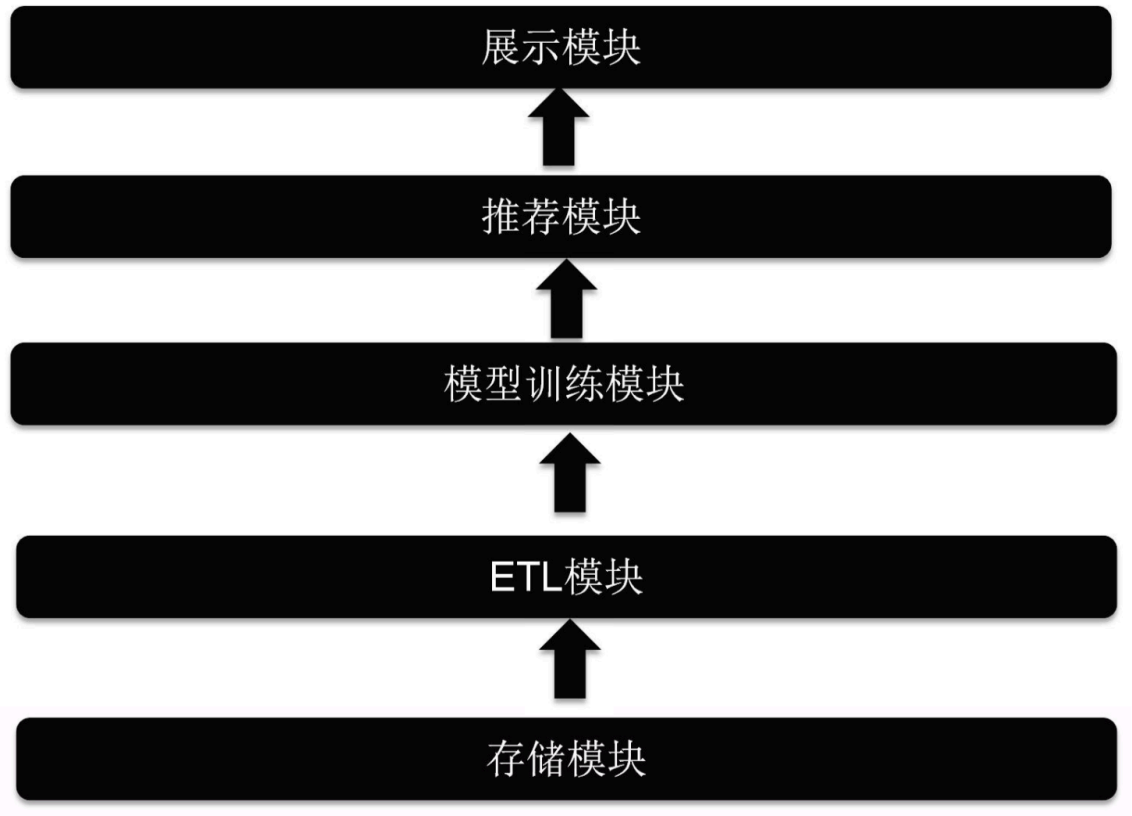
存储模块: 搭建和配置 HDFS 分布式文件存储系统，并且把 Hbase 和 MYSQL 做为备用的存储数据库。

ETL模块: 加载原始数据，清洗，加工，为模型训练模块和推荐模块准备所需要的各种数据。

模型训练模块: 负责产生模型，以及寻找最佳的模型。

推荐模块: 包含离线推荐和实时推荐，离线推荐负责把推荐结果存储到存储模块中。实时推荐则负责产生实时消息队列，并且消费实时消息产生推荐结果，最后存储到存储模块中。

数据展示模块: 负责展示项目中所用的数据。



项目重难点

1. 数据仓库的准备
    - 主要是集群搭建，机器必须要没问题，或者数据丢失或报警。
    - Hive 要对接到 SparkSql 上，即 Hive 要成功实现 SparkSql 的管理。Hive 默认是安装的 Derby，Derby 有一个问题，只支持单用户访问，所以这里我们用 MySql，这样可以多用户访问 Hive 数据。
    - Zepplin 要访问到 Hive 里的数据。
2. 数据的处理
    - 在数据的处理或者加工会遇到一些问题，一些解决思路。
3. 实时数据流
    - 实时数据流一直在各个企业中都是重难点。
    - 数据实时性
    - 数据的一致性，完整性
    - 保证应用不会崩掉，重启后还能有效处理