数据仓库

一、数据仓库概念

数据仓库的输入数据包括：业务数据、用户数据和爬虫数据等。

业务数据：处理事务产生的数据。比如用户在电商网站中登陆、下单、支付的过程中，和网站后台数据库进行增删改查交互所产生的数据，保存在MySQL、Oracle等数据库中。

用户行为数据：通过埋点收集与客户端交互过程产生的数据，保存在日志服务器。比如页面浏览、点击、停留、评论、点赞、收藏等。

爬虫数据：通过技术手段获取其他公司的数据。

二、需求及架构设计

需求

1、采集平台

用户行为数据采集平台搭建

业务数据采集平台搭建

2、离线需求

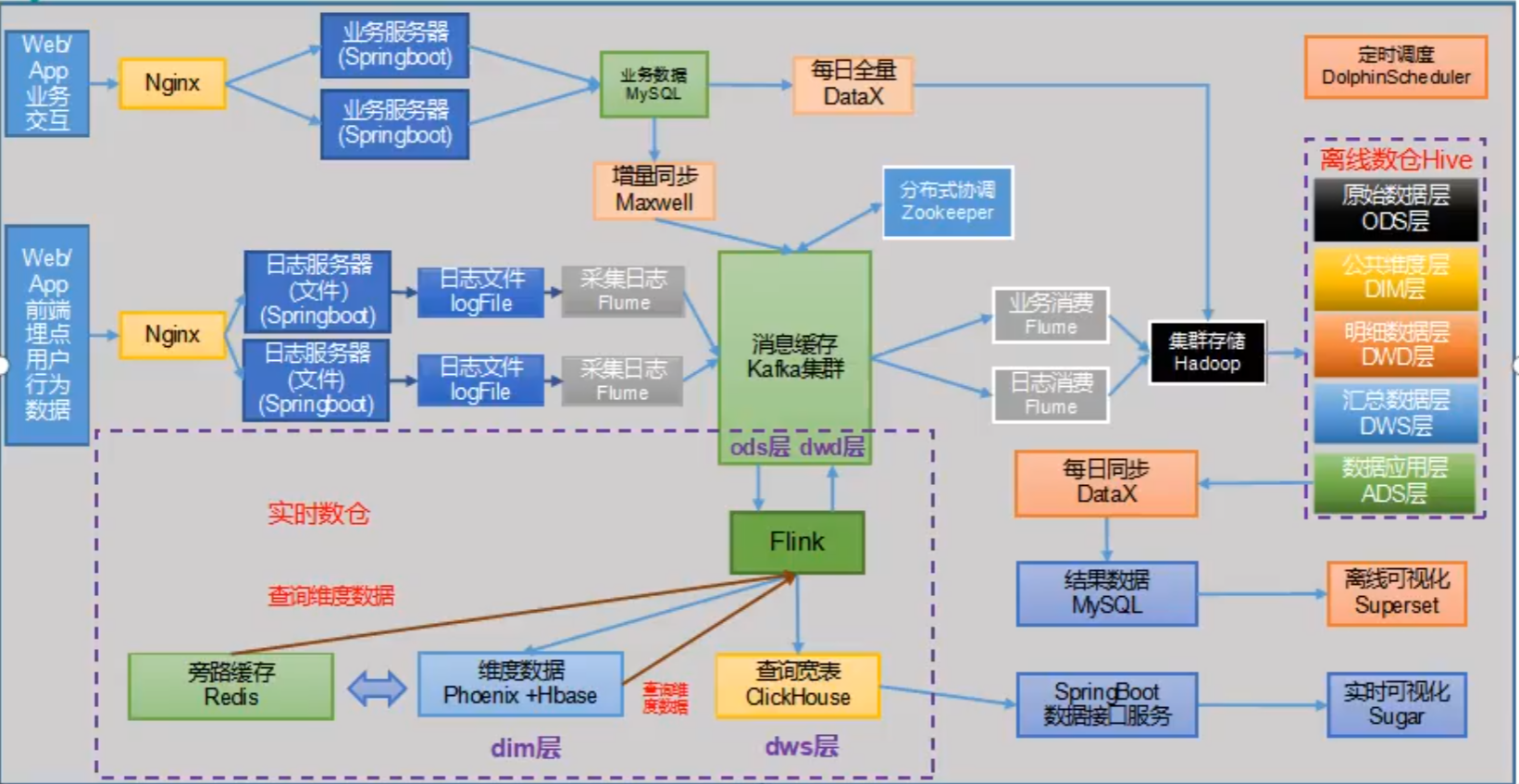
3、实时需求

架构设计

1、技术选型



2、系统数据流程图



数据同步的方式：

全量数据同步：表的全部数据。

增量数据同步：表的新增及变化的数据。

采集数据的目的，为了后期的统计分析做准备

需求1：当前电商网站所以得注册用户的数量

Select count(\*) from t\_user

需求2：当前电商网站最近1周的新增注册用户的数量

Select count(\*) from t\_user where regdate >= 10-05 and regdate <= 10-12

增量生成的一个表 t\_user\_7

Select count(\*) from t\_user\_7