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

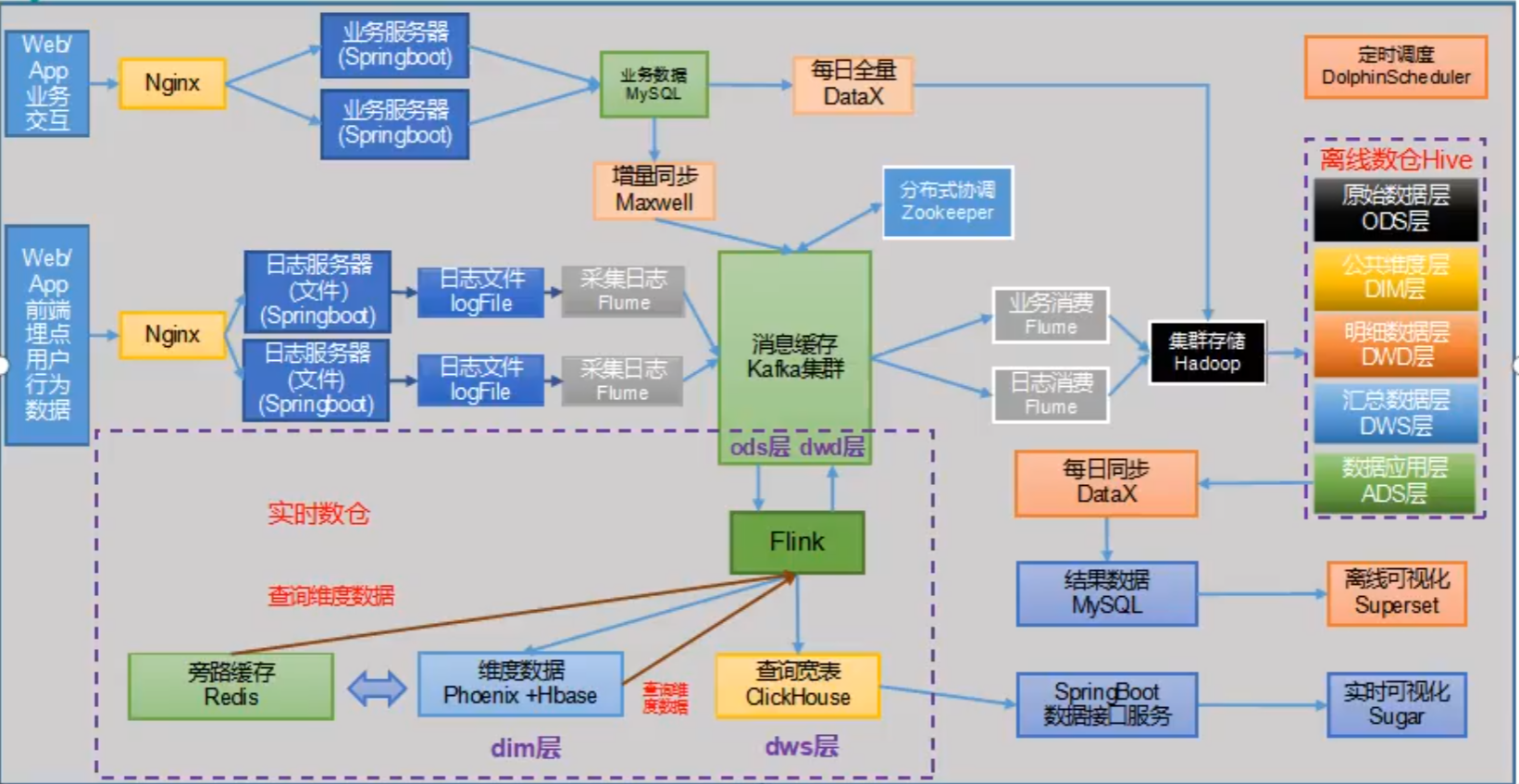
3、实时需求

架构设计

1、技术选型



2、系统数据流程图



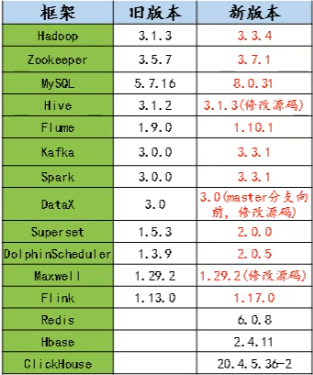
3、框架版本选型

（1）Apache

（2）CDH

（3）HDP

2和3合并成CDP



4、服务器选型

（1）物理机

（2）云主机

（3）企业选择

5、集群规模

（1）每天活跃用户数量，每人平均数据，相乘。

（2）每条1k左右，每天条数，除以两个1024，得到大概g。

（3）按一年算，再乘以365。

（4）如果副本有3个，再乘以3。

（5）预留20%-30%Buf，除以0.7。

（6）得出服务器大小和数量。

三、用户行为日志

1、概述

埋点：用户的各项行为信息以及行为所处的环境信息，包括代码埋点（前端/后端），可视化埋点，全埋点。

2、内容

页面浏览记录、动作记录、曝光记录、启动记录和错误记录。

环境信息：用户信息，时间信息，地理位置信息，设备信息，应用信息，渠道信息。

页面浏览记录：页面id，页面对象。

动作记录：对象类型，对象id。

曝光记录：对象类型，对象id。

启动记录：广告id等信息。

错误记录：报错相关的页面，动作，曝光，启动等信息。

四、服务器的安装

（1） 准备三台模拟机

（2）修改模拟机ip

设置静态ip，sudo vim /etc/sysconfig/network-scripts/ifcfg-ens33

BOOTPROTO=static

IPADDR=192.168.10.101

GATEWAY=192.168.10.2

DNS1=192.168.10.2

（3）修改主机名

sudo vim /etc/hostname

Sudo vim /etc/hosts

192.168.10.101 master

192.168.10.102 slave1

192.168.10.103 slave2

设置Linux虚拟机的虚拟网络

五、编写集群分发脚本

1、使用rsync命令原始拷贝

Rsync -av /opt/module root@slave2:/opt/

2、在/home/root/bin中创建xsync文件

#1. 判断参数个数

If [$# -lt l]

then

Echo Not Enough Arguement!

exit;

fi

#2.遍历集群所有机器

for host in master slave1 slave2

do

echo ============ $host =============

#3.变量所有目录，挨个发送

for file in $@

do

#4 判断文件是否存在

if [ -e $file]

then

#5.获取父目录

Pdir=$(cd -p $(dirname $file); pwd)

#6.获取当前文件的名称

fname =$host “mkdir -p $pdir”

Rsync -av $pdir/$fname $host:$pdir

else

Echo $file does not exists!

fi

done

done

3、修改脚本xsync具有执行权限

chmod 777 xsync

4、测试脚本

xsync xsync

六、设置ssh免密登陆

1、在master上生成公钥和私钥

ssh-keygen -t rsa

2、将公钥拷贝到其他机器上

ssh-copy-id slave1

ssh-copy-id slave2

3、重复以上步骤

六、安装JDK、Hadoop、Zookeeper、Kafka、Flume

七、Flume采集流程的配置

dfs.namenode.hadnler.count

Count = int(20\*math.log(x))

x为机器数量

配置TailDirSource和KafkaChannel，并配置日志校验拦截器。

TailDirSource→KafkaChannel→Kafka

八、日志采集Flume配置实操

（1）创建Flume配置文件

在flume的job目录中创建file\_to\_kafka.conf

#定义组件

a1.source = r1

a1.channels = c1

#配置source

a1.sources.r1.type=TAILDIR

a1.sources.r1.filegroups = f1

a1.sources.r1.filegroups.f1=/opt/module/applog/log/app.\*

a1,sources.r1.positionFile=/opt/module/flume/taildir\_position.json

#配置channel

a1.channels.c1.type=org.apache,flume.channel.kafka.Kafkachannel

a1.channels.c1.kafka.bootstrap.servers = master:9092

a1.channels.c1.kafka.topic = topic\_log

a1.channels.c1.parseAsFlumeEvent = false

#组装

a1.sources.r1.channels = c1

九、