**大数据实践**

**一、参考资料**

0. 启动/关闭环境：

1. hdfs:

start-dfs.sh

stop-df.sh

1. yarn:  
   start-yarn.sh

stop-yarn.sh

1. zookeeper:

/opt/shared-area/zk.sh start

/opt/shared-area/zk.sh stop

1. kafka:

/opt/shared-area/kf.sh start

/opt/shared-area/kf.sh stop

顺序开 逆序关，建议每天重新开之前先关闭

hdfs和yarn 是一个账号启动，所有账号都可以用，但其他账号看不见进程，所以可以每天我打开

1. 每个账号都要更新环境变量

source /etc/profile

需要添加环境变量，

sudo /etc/profile.d/my\_env.sh

向以上文件中写入

1. Linux(Ubuntu 20以上， 我是22.04) mysql8.0 https://blog.csdn.net/LogosTR\_/article/details/125602116
2. 离线数仓：https://www.bilibili.com/video/BV1UN411j79o?spm\_id\_from=333.788.player.switch&vd\_source=b372db3f69edca674c69713403538e45&p=7
3. 实时数仓：https://www.bilibili.com/video/BV1wZ4y1a7Y6/?spm\_id\_from=333.1007.top\_right\_bar\_window\_custom\_collection.content.click&vd\_source=b372db3f69edca674c69713403538e45
4. Navicat: https://blog.csdn.net/weixin\_50670076/article/details/136350060
5. Hadoop: https://www.bilibili.com/video/BV1Qp4y1n7EN?spm\_id\_from=333.788.videopod.episodes&vd\_source=b372db3f69edca674c69713403538e45&p=22
6. 服务器：
7. 公网ip:

hadoop102 ip:47.96.250.106

hadoop103 ip:121.196.201.14

hadoop104 ip:116.62.199.244

私网ip:

hadoop102 ip:172.30.198.254

hadoop103 ip:172.23.124.143

hadoop104 ip:172.26.233.63

账号：姓名首字母

密码：123456

端口：自己去下面四个选一个，选完后面记录一下（每个人在三台服务器上用同一个端口吧，好记）

1. 服务器ssh： /etc/ssh/sshd\_config

添加:  
Port 20010 :zys

Port 20020 :hw

Port 20030 :xcxy

Port 20040 ：gzq

Port 20050 : xyx

对所有相关端口执行以下命令

ufw allow 200X0

创建用户：

groupadd group (执行一次即可)

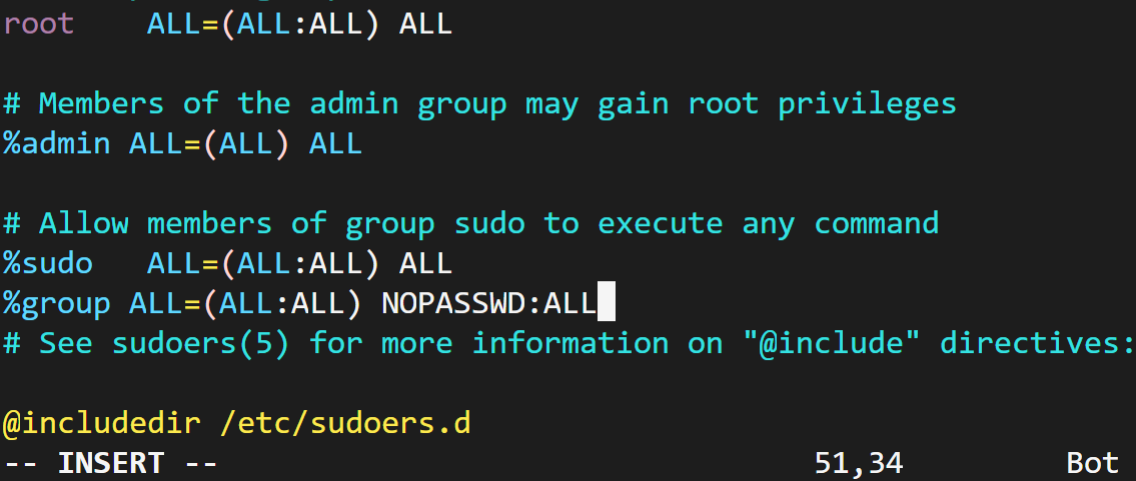
useradd -g group -s /bin/bash -d /home/xyx xyx

Passwd xyx

123456

1. 管理员权限

sudo vim /etc/sudoers



更改主机名：

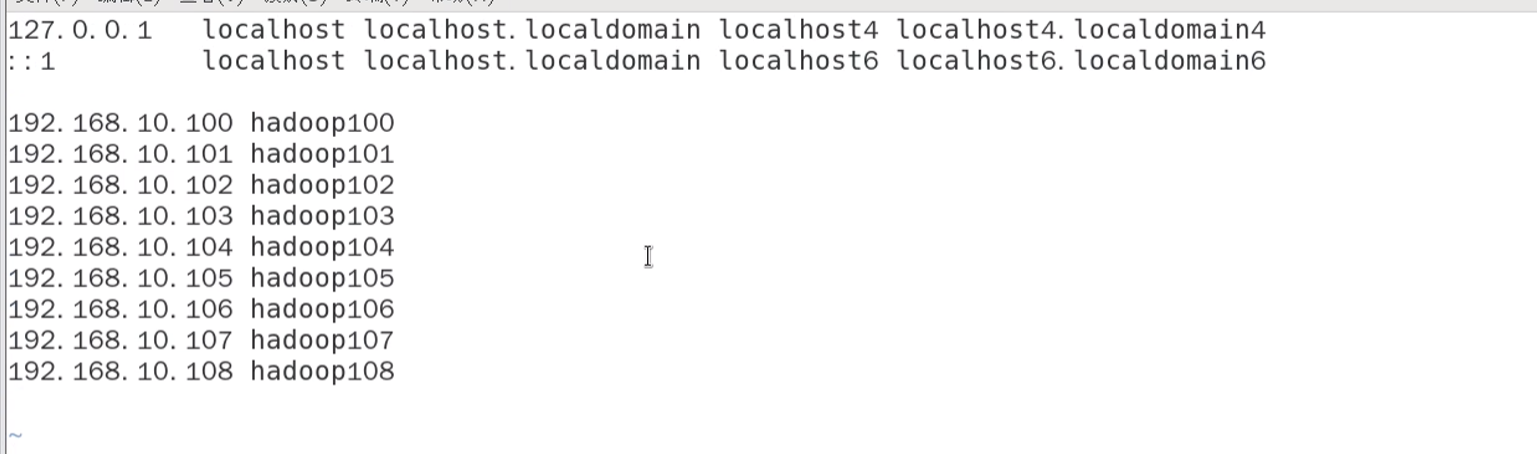
sudo vim /etc/hostname

Hadoop:  
1. 主机名称映射： /etc/hosts

1. 权限管理

sudo chmod -R 770 module/ software/

（ip改为自己服务器的ip）



**windows配置：**

服务器地址映射： C:/Windows/System32/drivers/etc/hosts

内容同上

1. 家目录：

sudo mkdir -p /home/gzq

sudo chown gzq:group /home/gzq

sudo chmod 755 /home/gzq

1. ssh免密：

在102上执行

cd ~/.ssh

ssh-keygen -t rsa

ssh-copy-id hadoop102

ssh-copy-id hadoop103

ssh-copy-id hadoop104

三台都要给copy,同一台也是，其他两台主机一样

！！！！要用hadoop前配免密

1. xsync脚本使用方式：

功能，将/路径复制到三台服务器当中（先配ssh免密）

（1） 非sudo: /opt/shared\_area/xsync.sh /路径

（2） sudo :

1. 启动hadoop:

./sbin/start-dfs.sh

./sbin/start-yarn.sh

1. hadoop使用

向hdfs发送文件：

hadoop fs -put /本地文件路径 /hdfs路径

从hdfs删除文件

hadoop fs -rm -r /hdfs路径

修改权限：

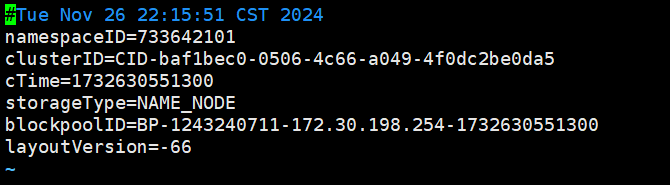
hadoop fs -chmod 777 /

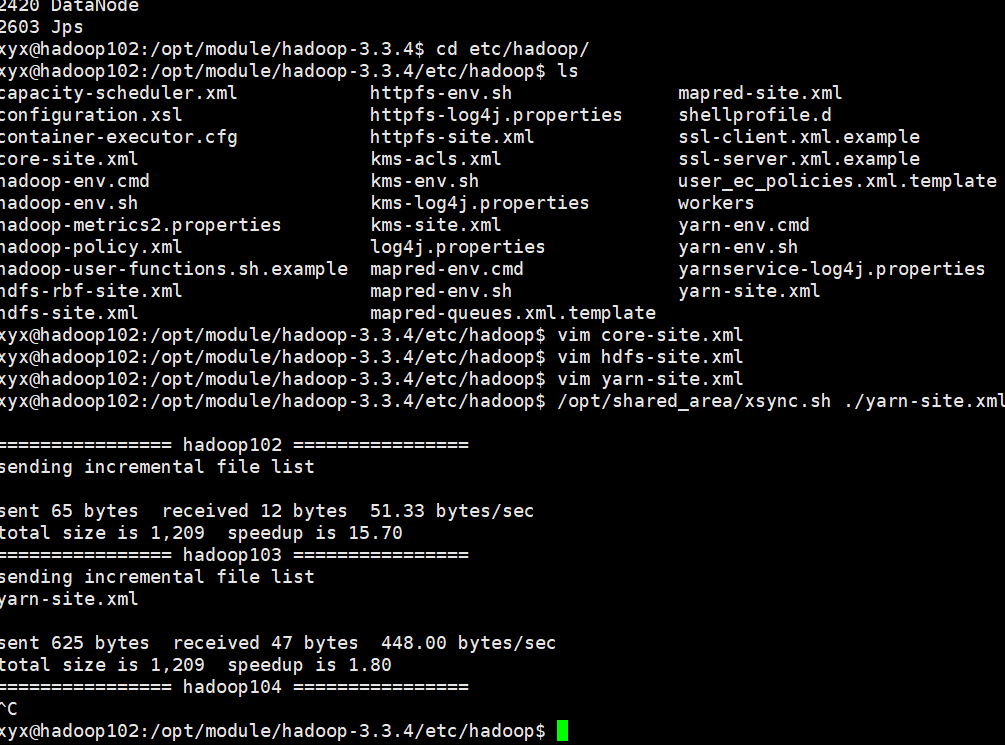
* hadoop集群配置
* hadoop102:

1. http访问hdfs NameNode (2nn)

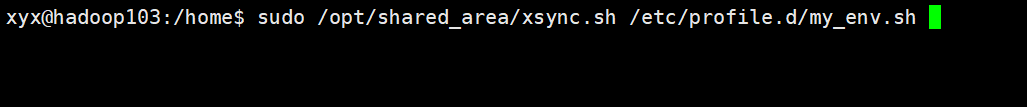
NameNode: 47.96.250.106:9870

2nn: 116.62.199.244:9868



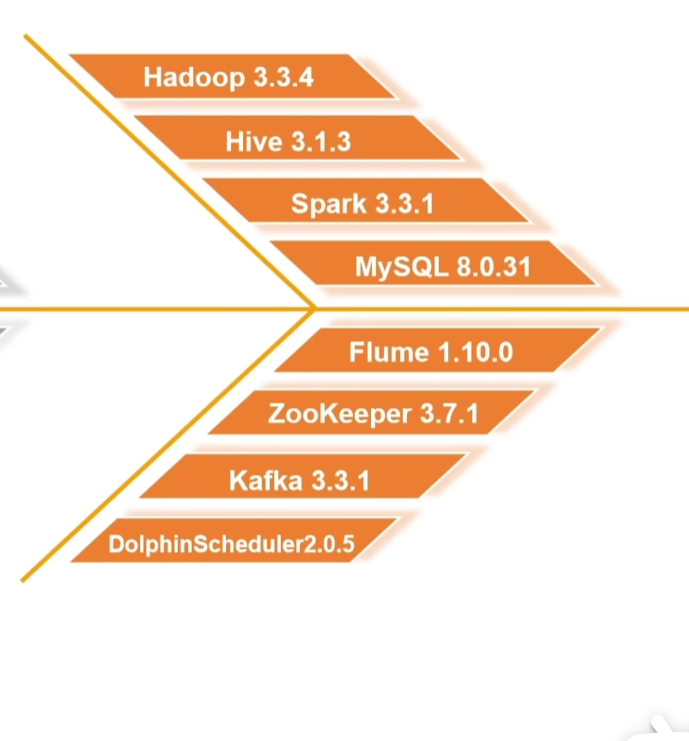


!!!!!!!!!!!!!!!!!!!!!!!104

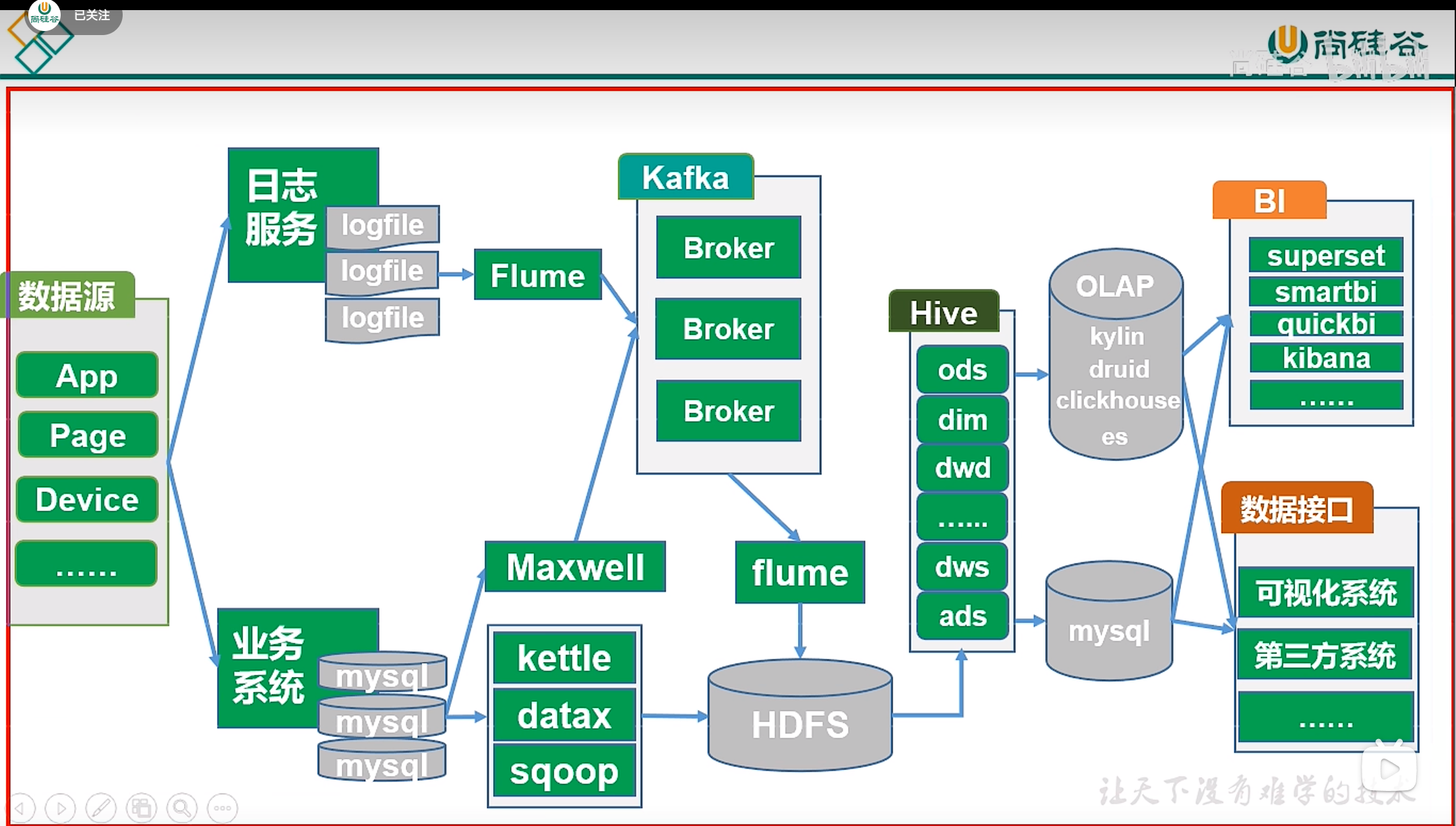


需要三台主机的root密码

离线架构：



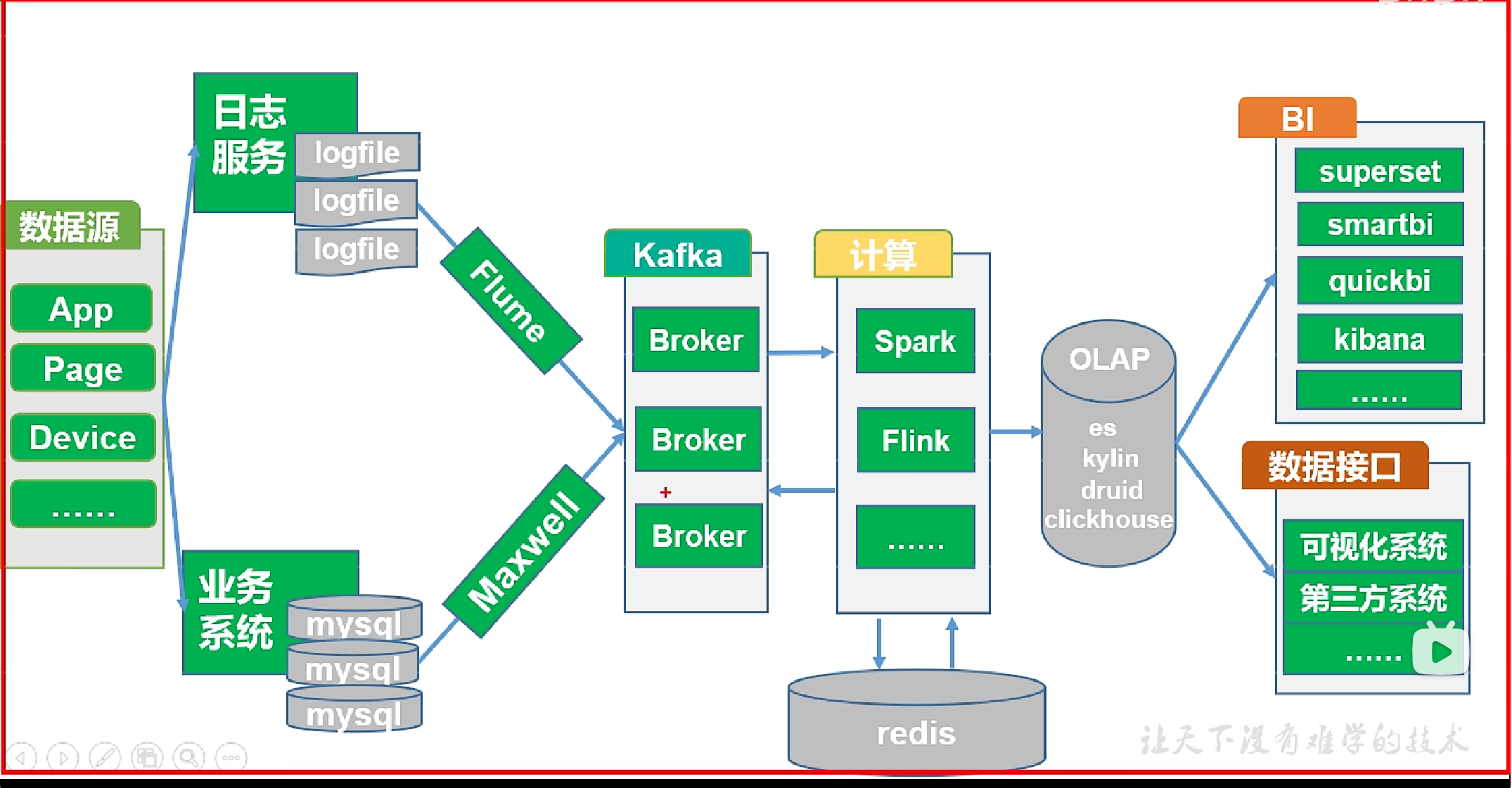
离线架构



全量备份：dataX

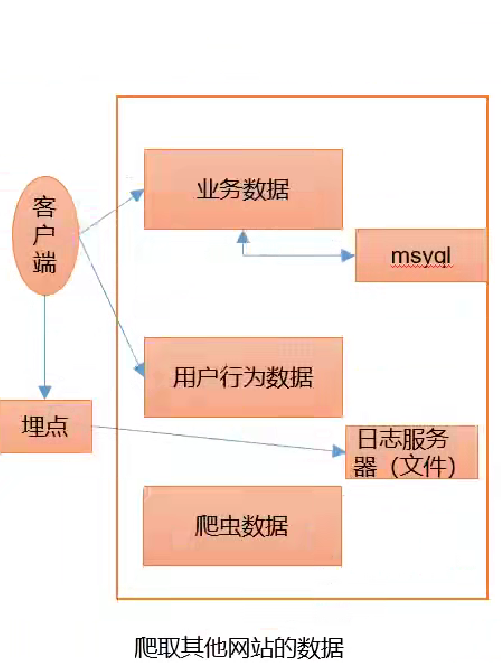
增量：Maxwell -> Kafka -> Flume -> HDFS

实时架构：



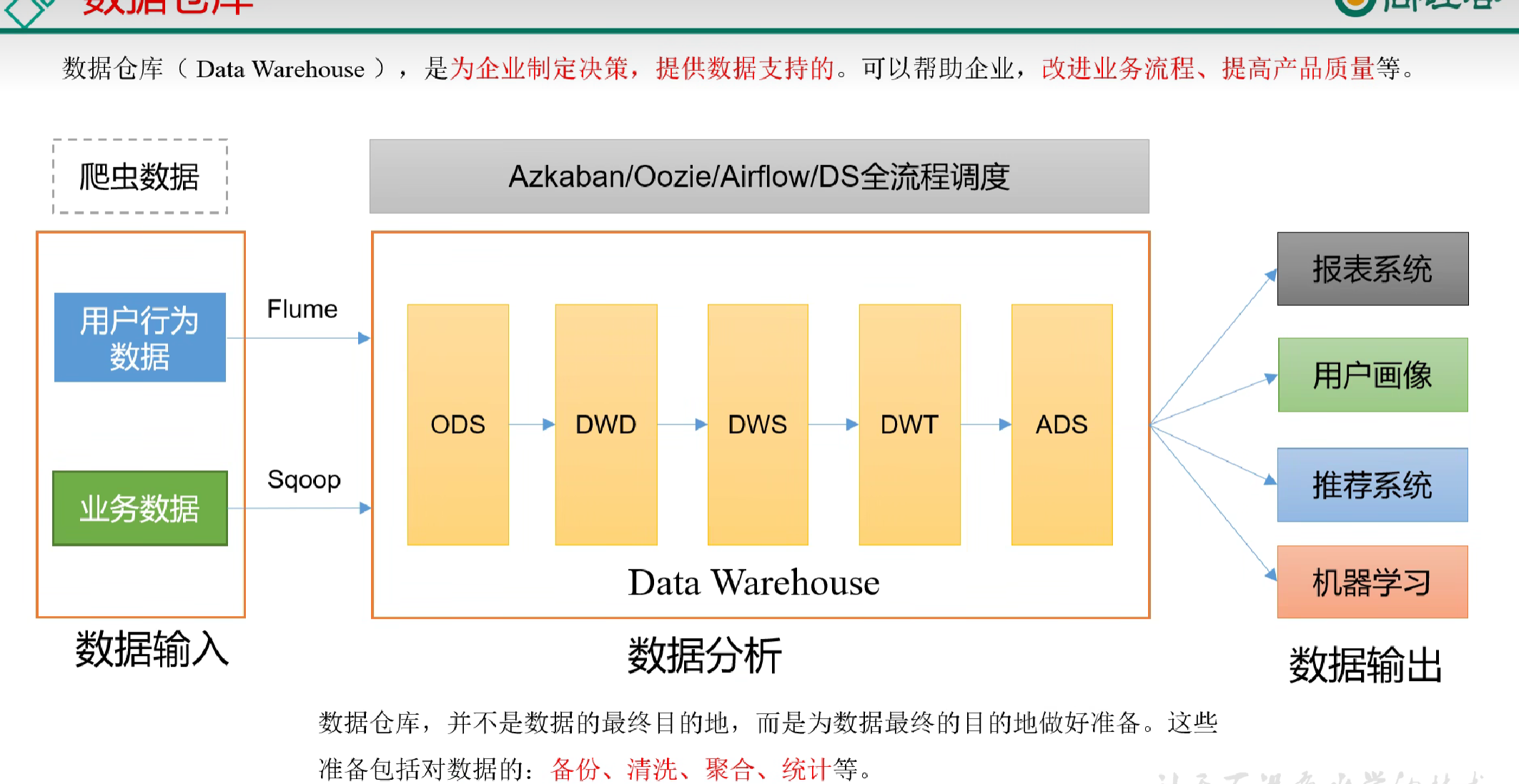


数据种类：

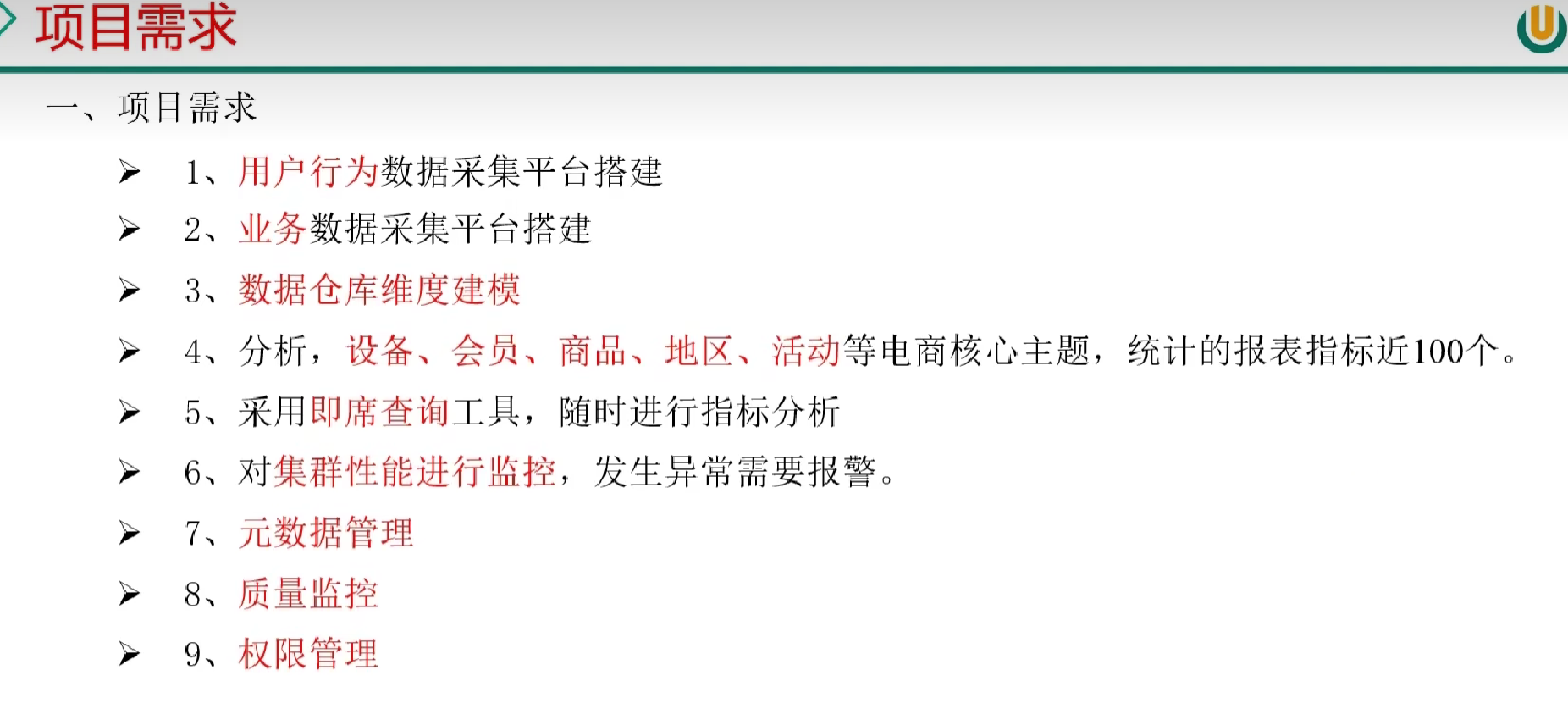


（咱们应该只有业务数据？）



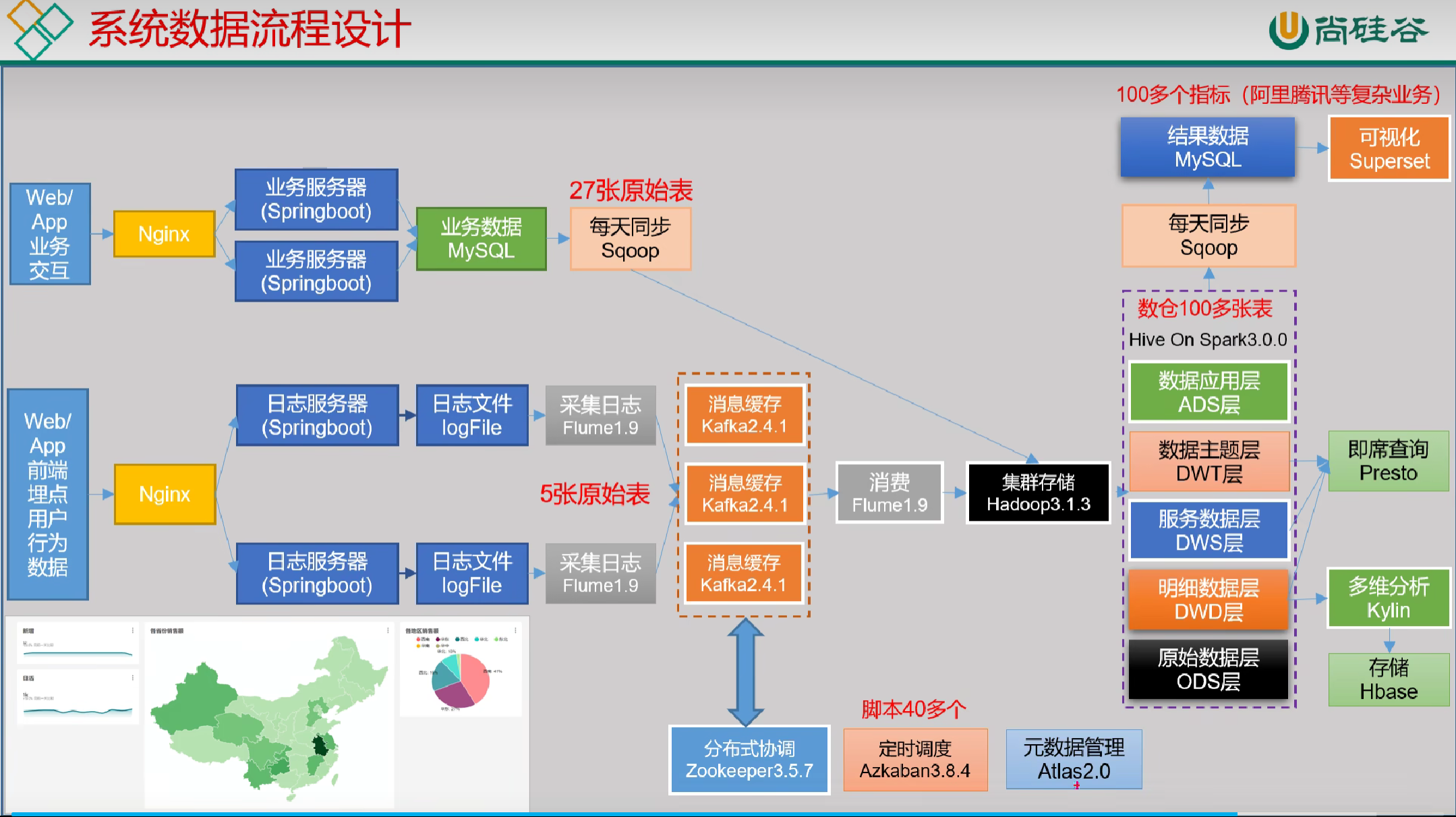


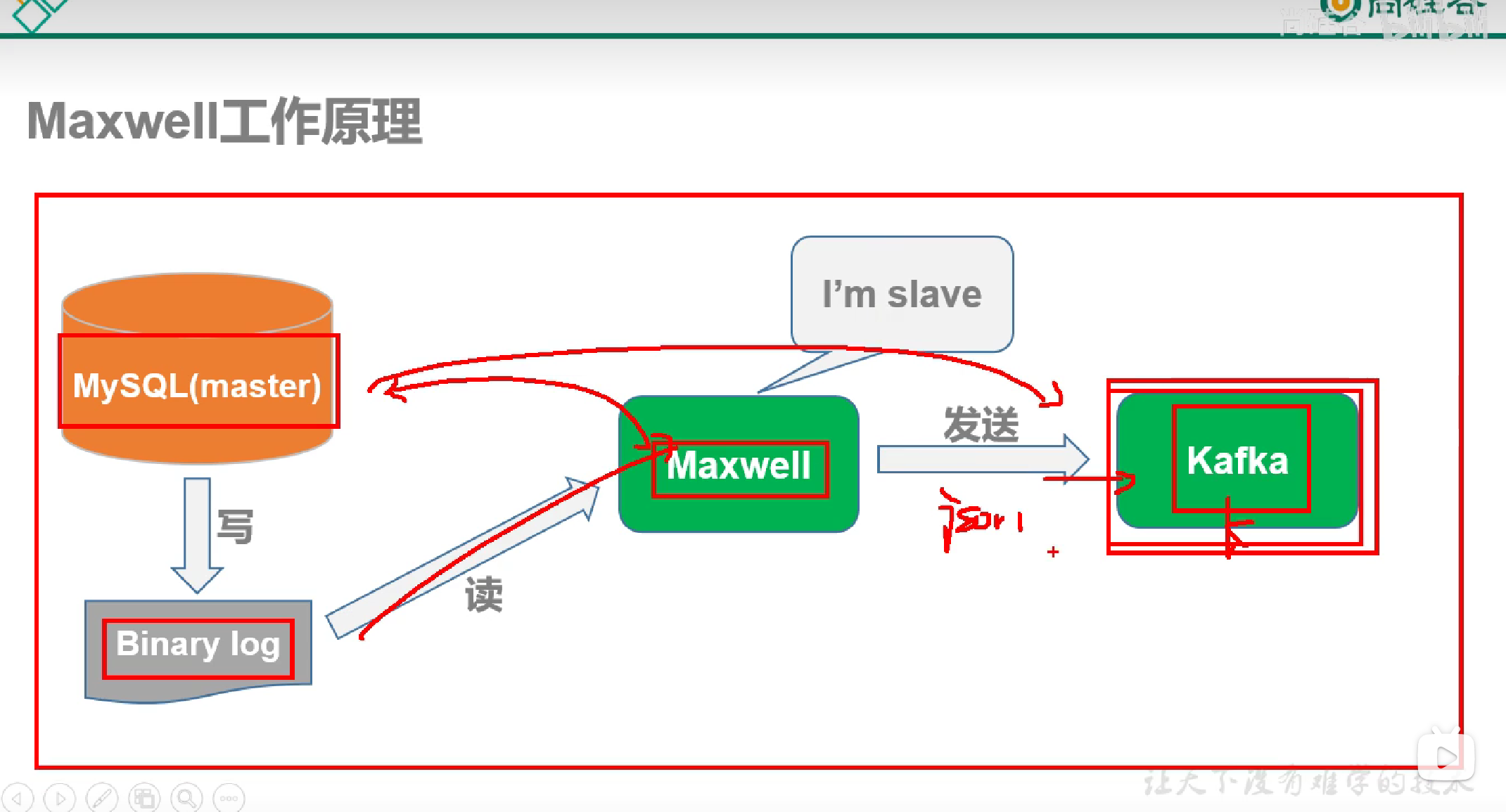
需求：



咱数据处理部分可能主要要做2&3







**配置Maxwell的过程：**





发现没有my.sh，打开看看my\_env.sh好像是配环境变量的，在里边加了  
*#MAXWELL\_HOME*  
export MAXWELL\_HOME=/opt/module/maxwell-1.29.2  
export PATH=$PATH:$MAXWELL\_HOME/bin



配置mysql



添加了：

[mysqld]

#maxwell 需要指定 Binlog 日志以"行级别"的方式进行记录

binlog\_format=row

#MySQL服务器的唯一标识号

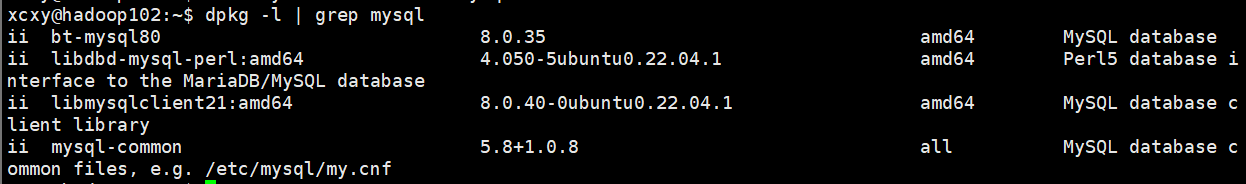
server\_id=1

#启用二进制日志 Binlog，指定 "master" 作为 Binlog 文件的前缀名

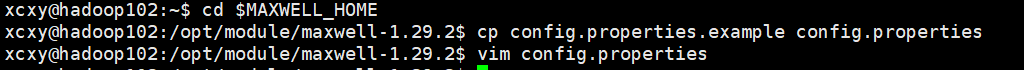
log-bin=master

重启mysql



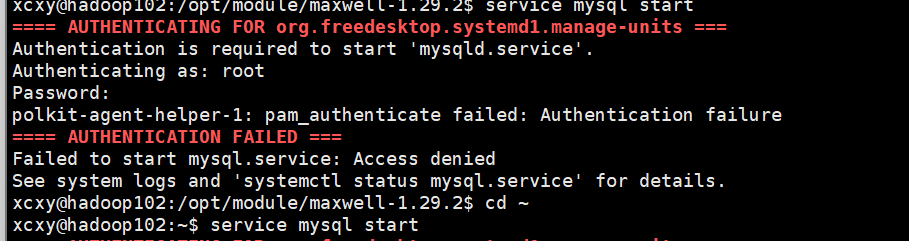


端口号没改↓↓↓修改配置文件！！！！！！！！！！！！！





启动mysql失败了，先暂停



**二、数据处理**

1.删除error列，

2.数据表解耦合——创建商家表

3.传入云服务器

https://chatgpt.com/share/674c9c6f-6d34-8009-b0b9-6e27434265c