# 《Docker优化分布式平台搭建方案》

姓名：易晓博

学号：2018282110414

2019年01月07日

1. **Docker简介**

Docker是使用Google公司推出的Go语言进行开发实现， 基于Linux内核的CGroup、Namespace以及AUFS技术，对进程进行封装隔离，属于操作系统层面的虚拟化技术。由于隔离的进程独立于宿主和其他隔离的进程，因此也被称为容器。



与传统虚拟机和Docker容器技术比较，虚拟机不仅包括应用也包括操作系统。 操作系统一般要占用几个GB的磁盘空间，而Docker容器共用宿主的内核，只需按照应用和必要的类库，非常轻量。



如果在测试服务器上使用Docker，会极大节省应景资源，简单企业成本。 Docker支持镜像构建和镜像部署，有类似Github的Docker hub的镜像开源仓库hub.docker.com, 这上面有100000多免费镜像资源，目标国内也有很大Docker镜像网站。使用Docker容器，只需获取镜像，一键安装即可。同时，也可以将常用的系统软件集成构建成镜像后部署到其他容器。支持测试环境和线上环境使用Docker容器技术进行软件的构建发布。

1. **Docker的优势**

最近学习研究使用了Docker，根据Docker的特性考虑改进“短视频信息数据挖掘”项目中分布式平台搭建。因为在平常的大数据平台的开发运维工作中，发现现有的大数据集群部署方式复杂极其复杂，对于每一个大数据组件的安装都需要修改原有的配置文件信息，再修改过程中也很容易出错。当然市面上也有开源的和商用的集群管理组件如Ambari（HDP），CDH，但是开源的管理组件安装服务仍然容易出错并且错误原因不好查找，而商用的集群管理工具不能扩展，导致很多平台功能不能往下开发。但是我们平台的环境搭建特色就是一站式部署，减少中间手动的改动流程，让环境部署更轻松，更便捷。

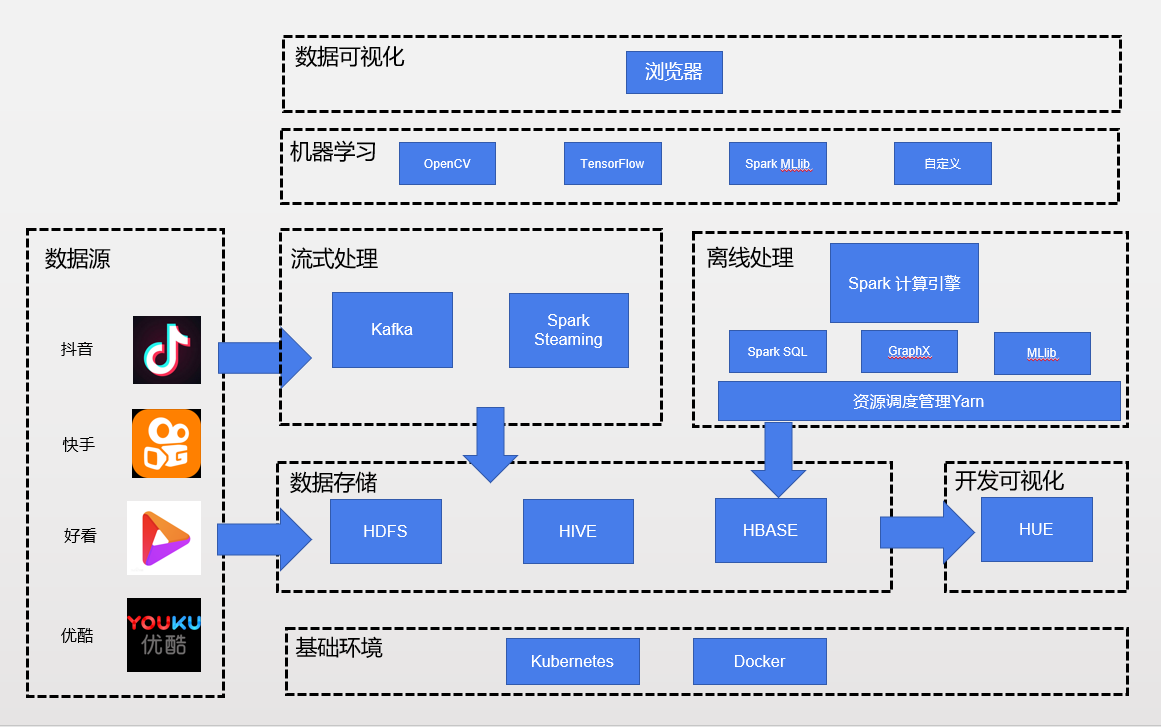
1. **使用Docker商业目的**

使用Docker来部署大数据分布式平台，其主要目的是为了打造统一的数据仓库与计算平台，解决后续大数据组件运维难的问题，降低集群搭建与服务运维的成本。考虑到本身做的是一个分布式的数据治理平台，本身平台可以不单只处理短视频信息，根据客户数据源的不同，进行相应的调整，可以在市场上获得不同的客户群体。

而我们所使用Docker部署的大数据基础平台具有高度的独立性和可移植性，降低了开发、搭建平台的难度和时间，能够达到分钟级快速搭建集群。

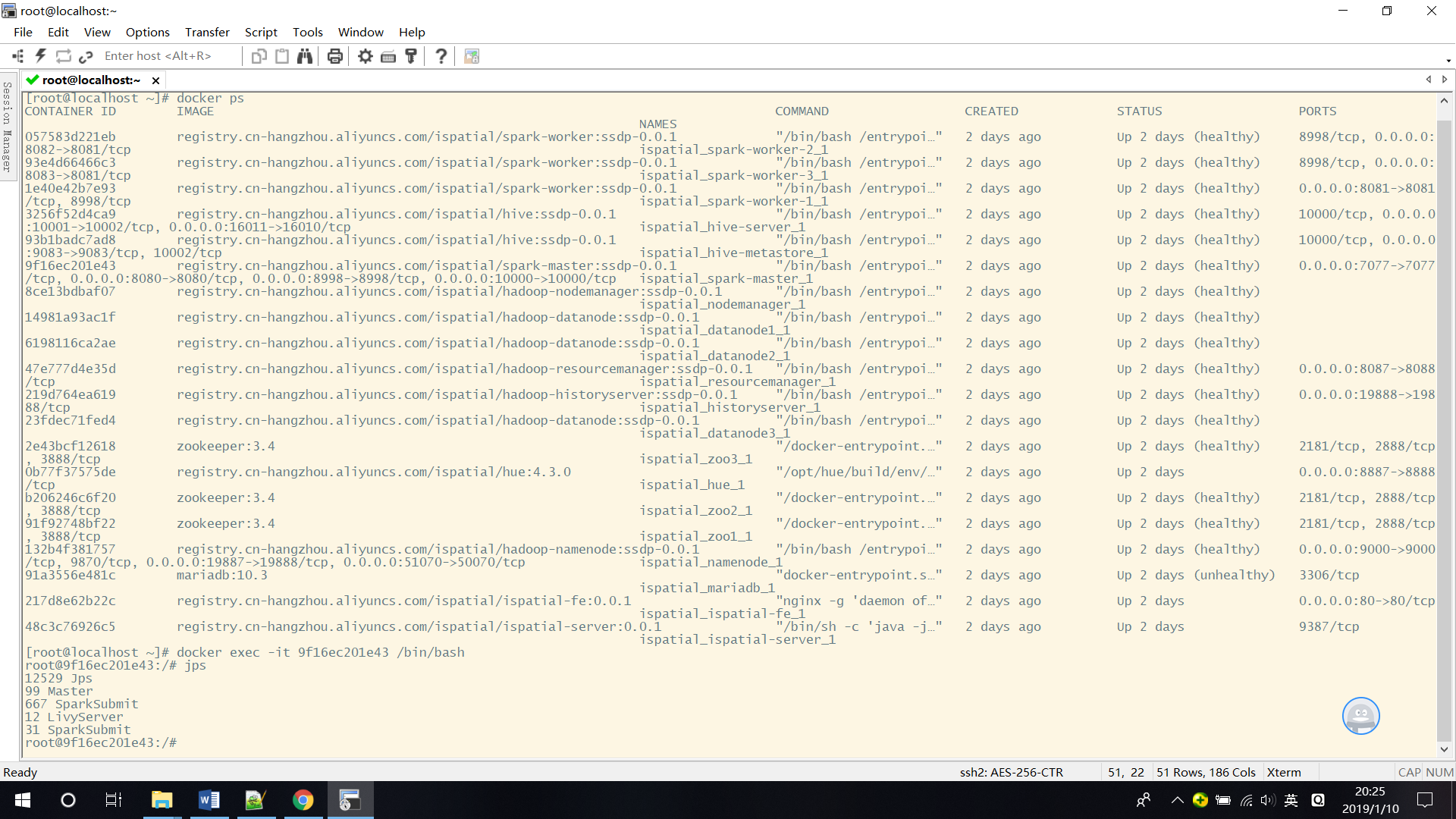
1. **改进软件架构**

使用Docker之后，我们之前的软件架构也有所调整，并且加入了Hive、Hbase、Hue等组件，这是为了后续平台更好地应对非结构化数据而新加入的，同时利用Hue这一可视化的开源图形界面，可以极大的提高开发人员的运维效率。

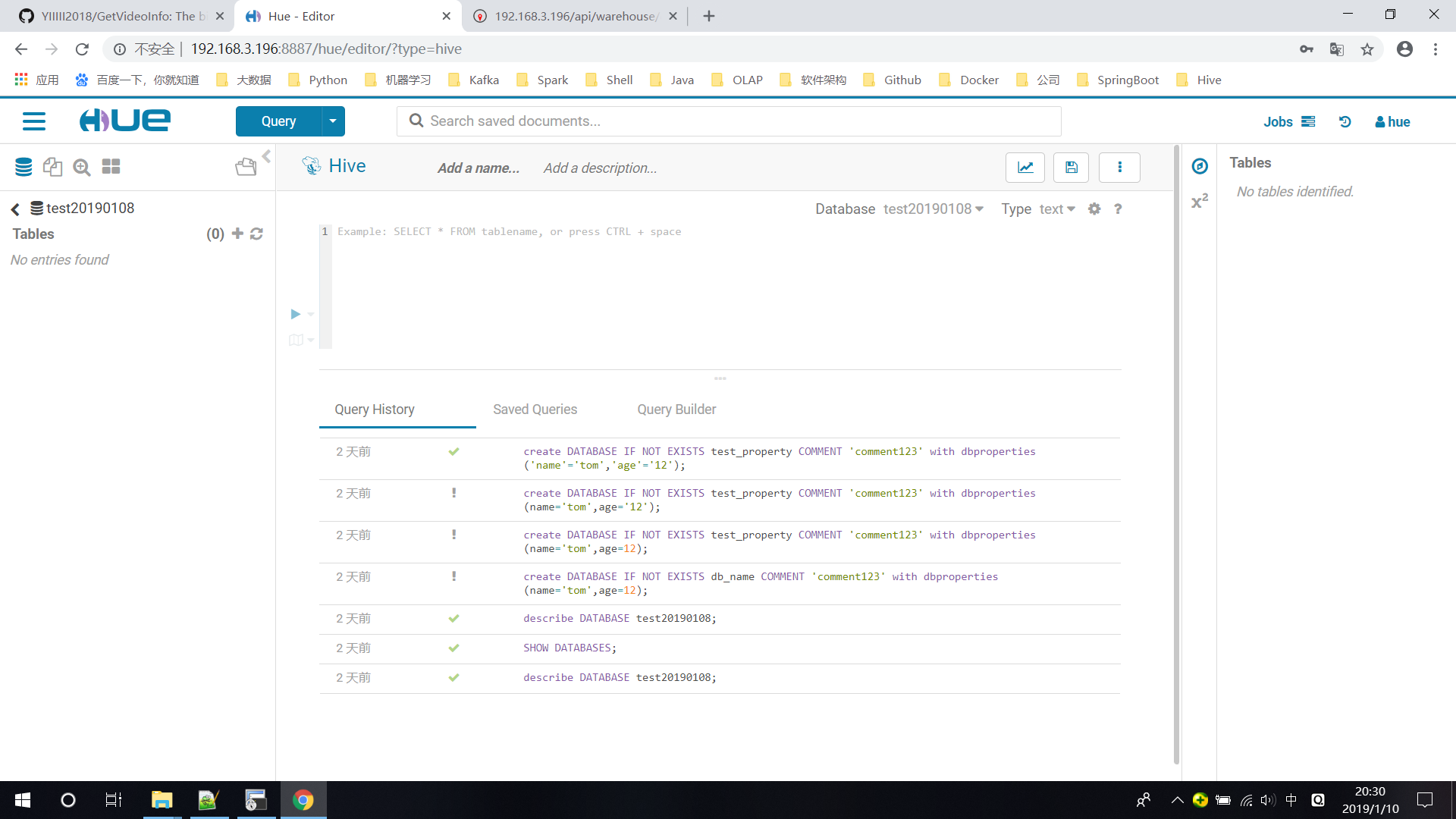


1. **Docker应用**

下面是Docker在虚拟机上运行的实例，后续平台实际搭建可以参考。从下图我们可以看到Spark的Master节点在Docker镜像中正常运行。



从下图我们可以看到平台正常使用Hue来间接管理Hive的数据存储过程。



从下图我们可以看到平台正常使用SpringBoot来调用Hive提供的服务，并根据所调用的API返回了正确Json对象。

