# 《短视频信息数据挖掘》

# Linux系统与开源软件设计课程报告

姓名：易晓博

学号：2018282110414

2018年12月01日

1. **项目背景**

近年来国内短视频风潮兴起，成为国民进行消遣娱乐、消磨时间的重要途经，同时蕴含于其中商机和问题也逐渐暴露出来。例如：短视频中人气较高的主播成为网络明星，带来大量收入和流量，并且能引领话题风向；短视频中火爆的小吃美食，则为街边小吃美食的商铺带来机遇；在短视频快速发展的同时，其中也出现一些有违法律道德的信息，则需要依法封禁。

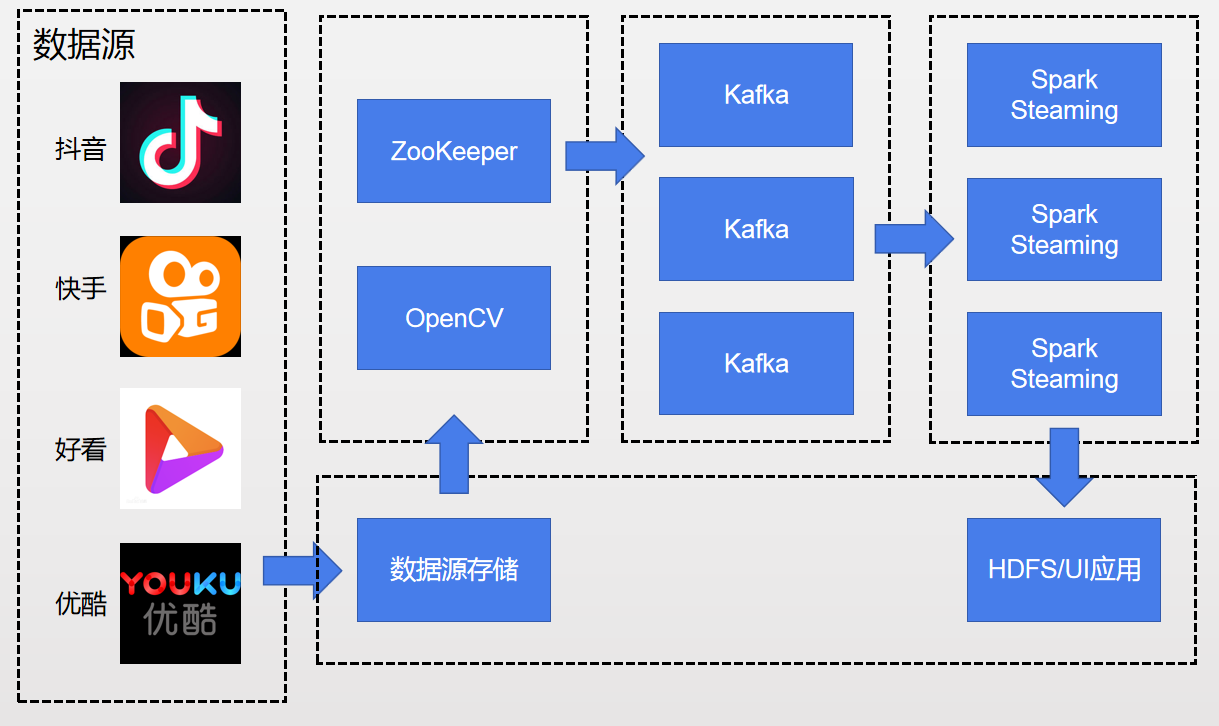
短视频信息数据挖掘主要针对于国内部分主流短视频平台，爬取短视频信息内容，对其数据包括但不限于标题、视频内容、作者、播放量、关注度、评论内容等进行数据挖掘；发现其中有价值的信息，例如可以引导网络流量的话题，民众喜爱的综艺、美食，将价值信息作为项目产品转售给第三方；同时如果发现有不良的违法信息，则及时向有关部门进行反映，协助监管部门进行监管。

1. **需求分析**

项目并无直接的用户，主要致力于有效挖掘出某个时间段内短视频信息，获取用户的话题风向、爱好兴趣等有价值信息。

1. **系统架构**

本项目通过自动爬取国内抖音、快手、百度好看、优酷等短视频内容，进行分布式存储，并进行视频流数据分析。主要使用的框架为OpenCV、Kafka、Spark、HDFS，具体架构图如下：



OpenCV（Open Source Computer Vision Library）是一个基于BSD许可证开源的库。这个库使用C++编写，但是该库也有大量的Python、Java and MATLAB/OCTAVE的接口。OpenCV包含了数百个CV算法，能够用来处理和分析图片及视频文件。

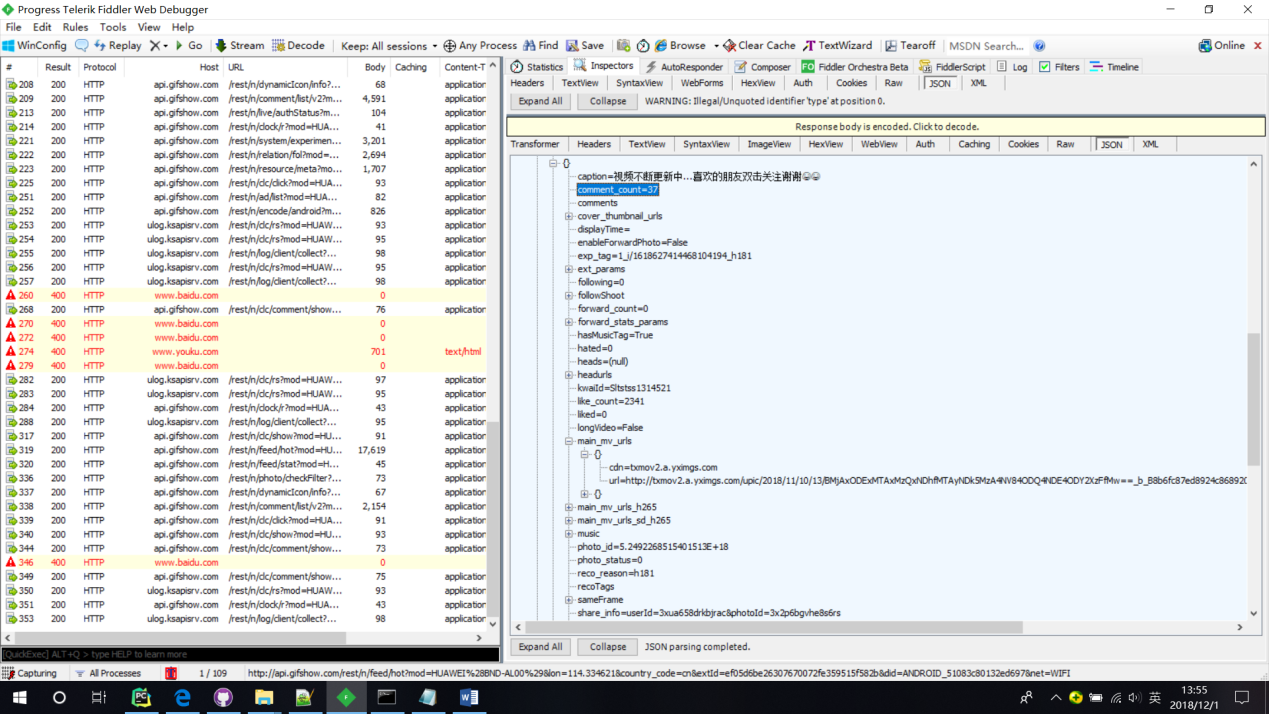
Apache Kafka是一个分布式的流平台，它提供了一个发布和订阅流记录（streams of records）的系统。这些记录能够按照可容错的方式进行存储，消费者可以处理这些数据。

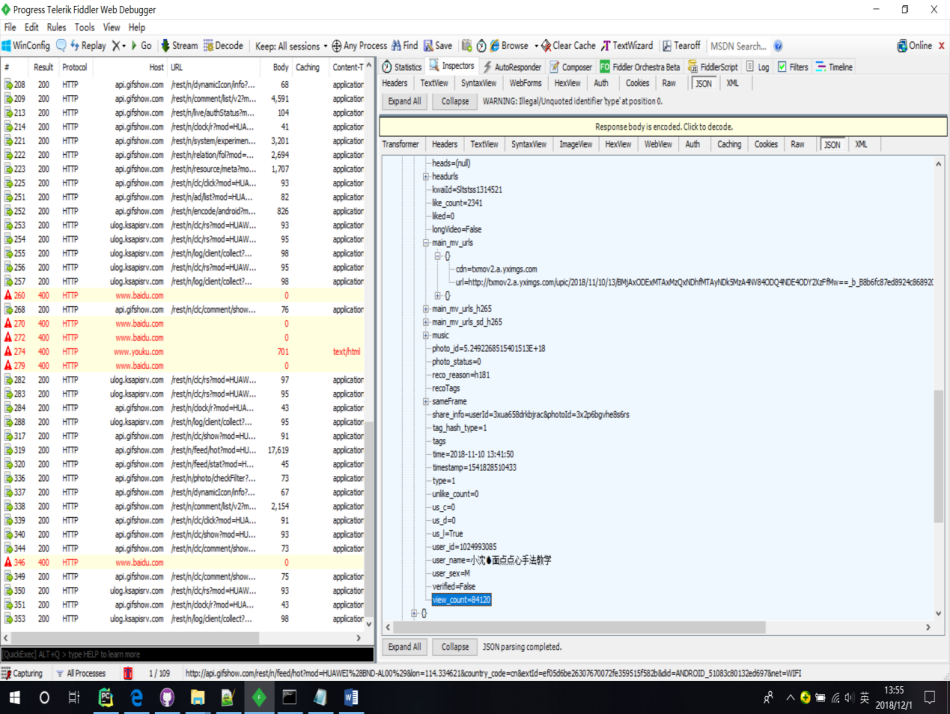
Apache Spark是一个快速、通用的集群计算系统。它提供了用于SQL和结构化数据处理的模块、用于机器学习的MLlib、用于图像处理的GraphX以及Spark Streaming。

项目使用OpenCV将视频流转换为帧数据，将这些数据交给ZooKeeper进行动态分配，由Kafka生成消息队列，Spark Streaming作为消费者从消息队列中不断取走视频流数据进行分析，最后将分析的结果展示到UI界面，或者在HDFS中存储起来。

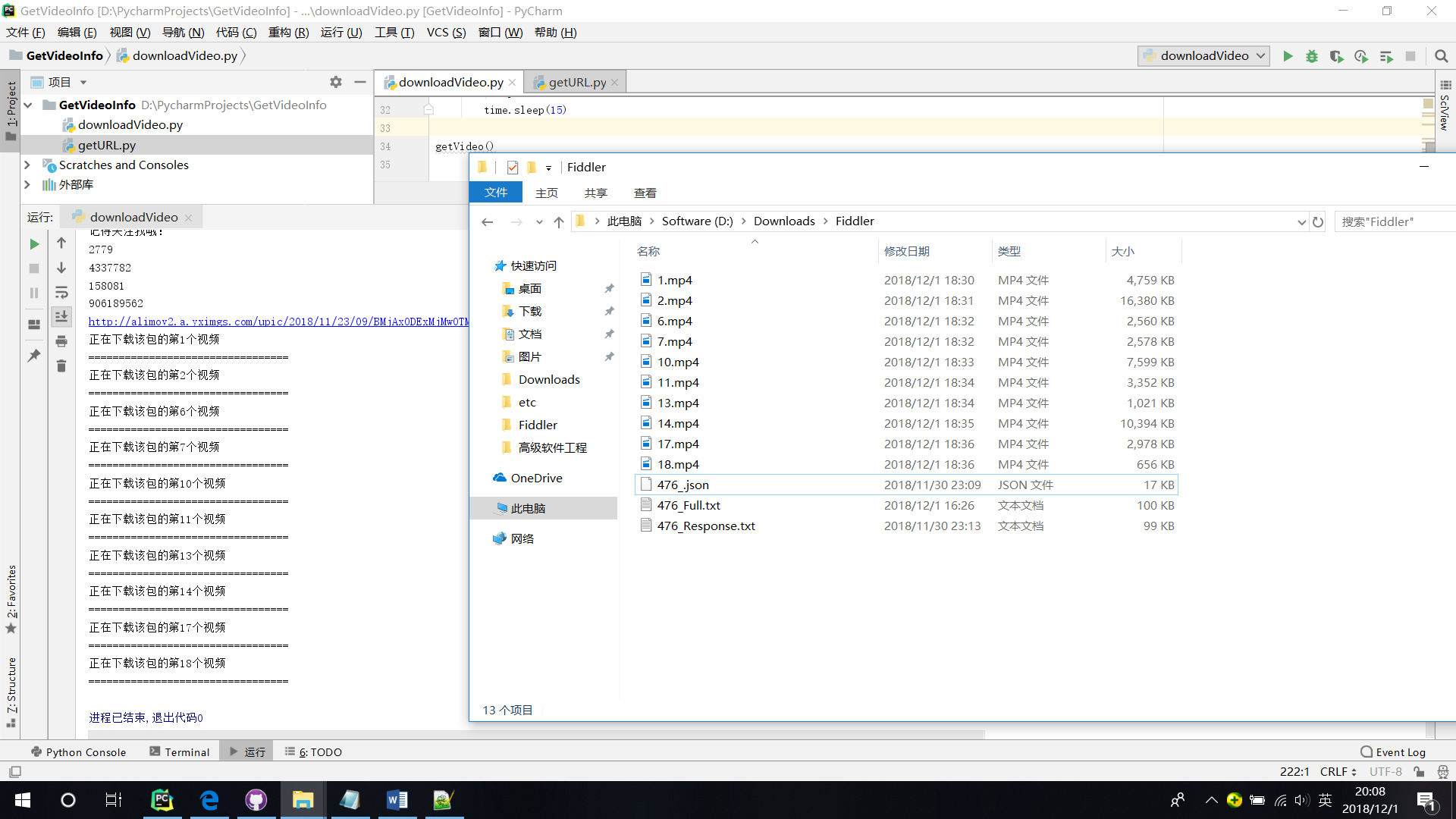
1. **阶段成果**
2. 视频爬虫处理

目前项目已经完成快手的爬虫处理，使用脚本从Fiddler批量下载数据包数据，从中解析出视频数据信息。包括视频发布者ID、视频播放量、视频关注度、评论、视频下载地址等等；爬虫Demo可以参考GetVideoInfo源码Demo，其中包含一个测试的数据包URL集以及下载完成待解析的视频。



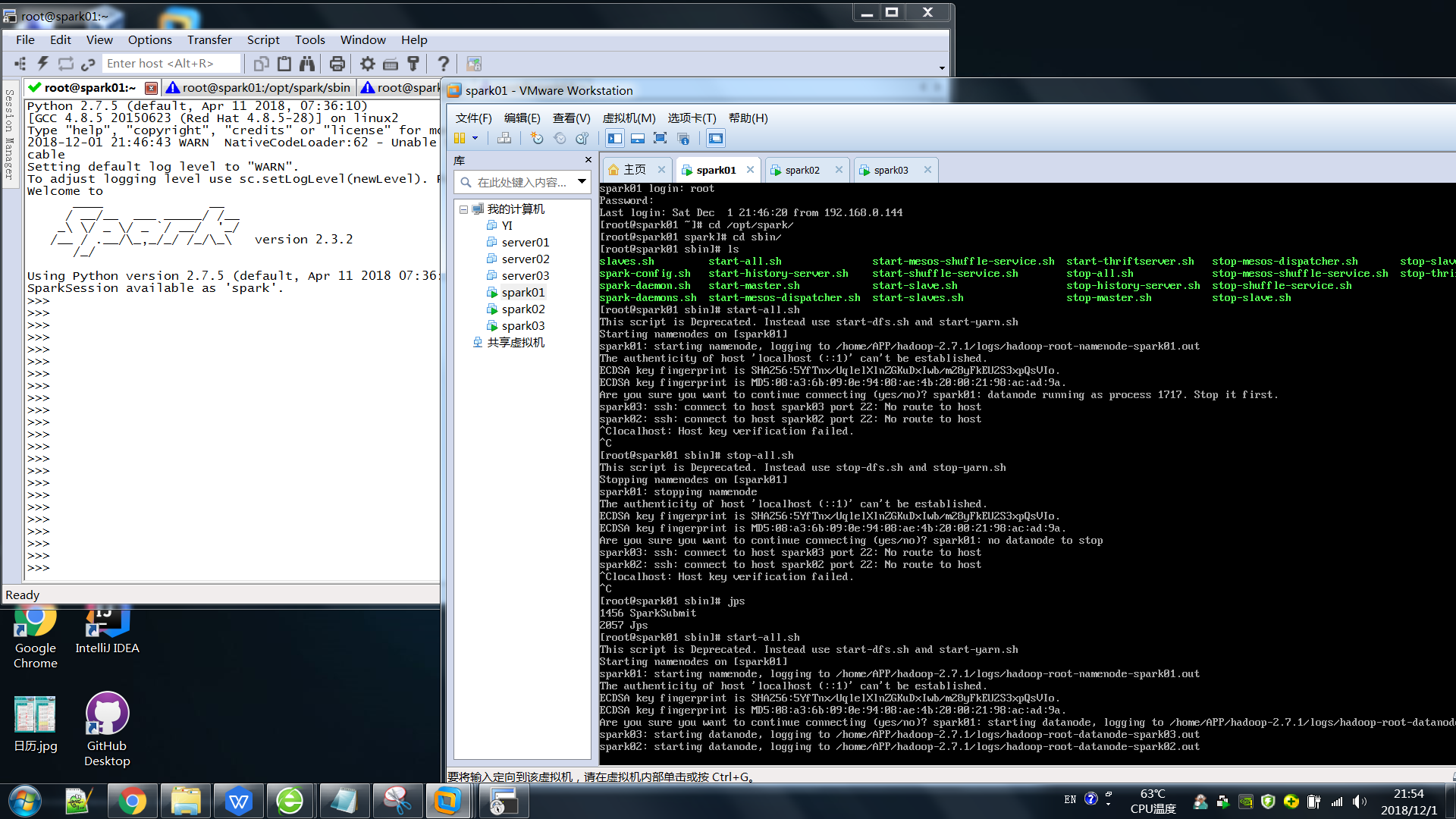


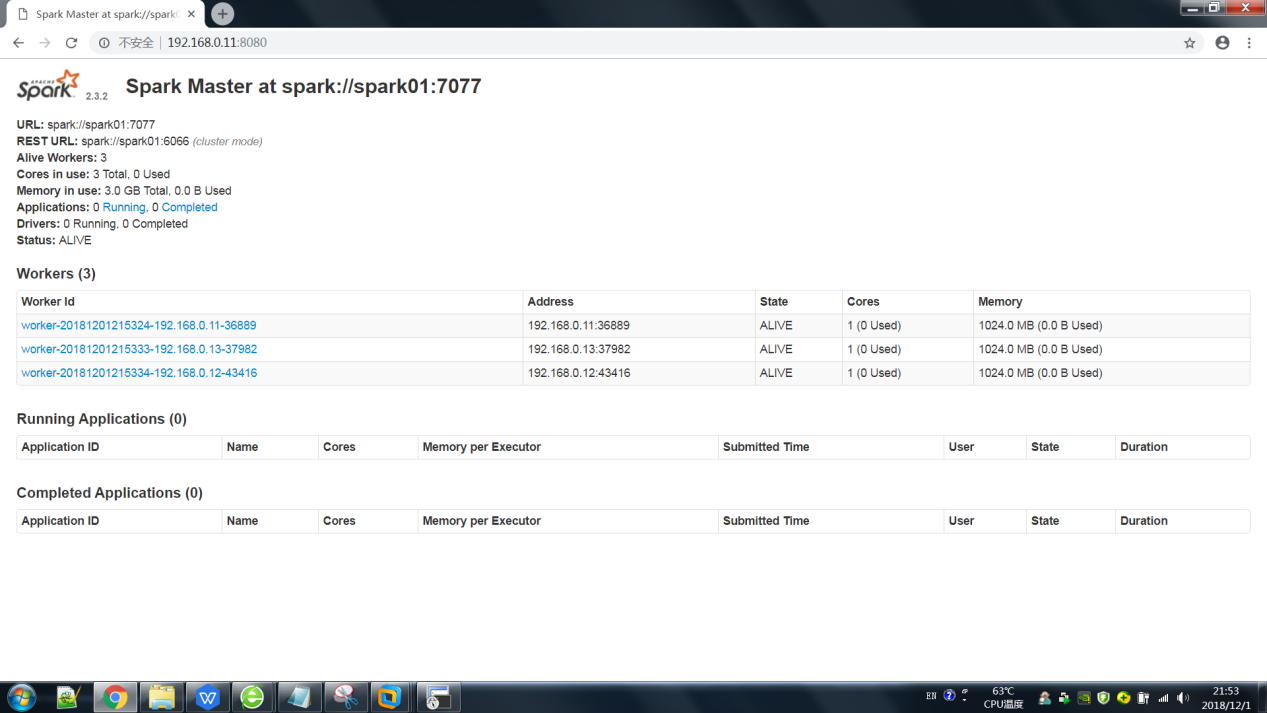
可以明显看到解包数据与在APP上看到的视频数据一致。并能正确从视频服务器端下载对应视频，留在处理解析。



1. ZooKeeper+Kafka+Spark+HDFS分布式平台框架搭建

在本地虚拟机上测试消息队列和SparkStreaming消费者获取数据正常，并能正常完成部分数据分析任务，本地PC性能有限，目前测试3个Spark节点。





Hadoop、ZooKeeper、Kafka、Spark分布式安装过程比较简单，项目后期会采用CDH进行集中管理，在此以上开源框架的配置与安装过程就不再过多赘述。

1. **待改进完善点**
2. 目前爬虫处理需要在手机端进行抓包，虽然抓取视频可以进行多线程处理，但是视频数据包的来源速度受限，一台手机仅一个数据包来源，且使用脚本从Fiddler解析数据包，属于半自动化的过程，希望能提高效率；
3. OpenCV目前工作缺少相应专业人员推进工作，下载的视频数据待处理解析；
4. 需要熟悉数理统计的人员，为清洗处理过的流数据进行建模分析，最好熟悉分类、回归、聚类算法，完成挖掘出有价值的视频信息的工作。