



大数据导论作业汇报

# 基于Spark的 数据处理分析

# 小组简介



**王天睿**

CS2208班

Hadoop集群搭建, 数据分析, PPT制作汇报



**输入标题**

Supporting text here.

You can use the icon library in iSlide ([www.islide.cc](http://www.islide.cc)) to filter and replace existing icon elements with one click.



**王雯琪**

CS2203班

爬虫, 数据预处理, 数据集下载, PPT制作





# 目

**PART01**  
**测试环境**

**PART02**  
**数据集特征**

# 录

**PART03**  
**选题目的**

**PART04**  
**课程及作业感悟**



# PART01

## 测试环境



# 测试环境

设备名称	Magicwang
处理器	12th Gen Intel(R) Core(TM) i5-12500H 2.50 GHz
机带 RAM	16.0 GB (15.7 GB 可用)
设备 ID	46E6BF7C-7000-41FF-92EC-73436BE166A4
产品 ID	00342-30683-44360-AAOEM
系统类型	64 位操作系统, 基于 x64 的处理器
笔和触控	为 10 触摸点提供触控支持

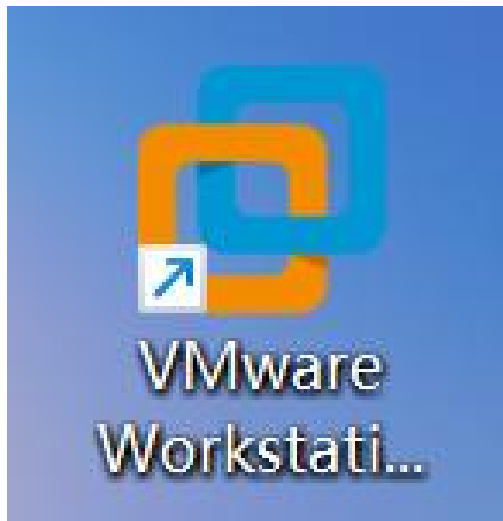
机带 RAM
设备 ID
产品 ID
系统类型

## <电脑配置>

Superluckyqi-Max
Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz 2.30 GHz
16.0 GB (15.8 GB 可用)
3A0DF7FF-A7C5-42E8-B5D6-63C47785729E
00342-35895-21184-AAOEM
64 位操作系统, 基于 x64 的处理器

# 测试环境

使用Vmware创建两台虚拟机进行Spark集群搭建，两台虚拟机命名master1和Slave，系统为Ubuntu 64位（ubuntu-22.04.3-desktop-amd64.iso），两台虚拟机都分配30G硬盘，4G内存。



master1

开启此虚拟机

编辑虚拟机设置

## 设备

内存	4 GB
处理器	2
硬盘 (SCSI)	30 GB
CD/DVD (SATA)	正在使用文件 a...
CD/DVD 2 (SATA)	正在使用文件 D:...
软盘	正在使用文件 a...
网络适配器	自定义 (VMnet...
USB 控制器	存在
声卡	自动检测
打印机	存在
显示器	自动检测

slave

开启此虚拟机

编辑虚拟机设置

## 设备

内存	4 GB
处理器	2
硬盘 (SCSI)	30 GB
CD/DVD (SATA)	正在使用文件 a...
CD/DVD 2 (SATA)	正在使用文件 D:...
软盘	正在使用文件 a...
网络适配器	桥接模式 (自动)
USB 控制器	存在
声卡	自动检测
打印机	存在
显示器	自动检测

# 测试环境

## 在虚拟机上使用Hadoop 3.3.5 集群搭建



```
hadoop@Master: /usr/local/hadoop$ jps
4357 Jps
3974 ResourceManager
4296 JobHistoryServer
3801 SecondaryNameNode
3598 NameNode
```

Hadoop集群搭建成功

Hadoop下载

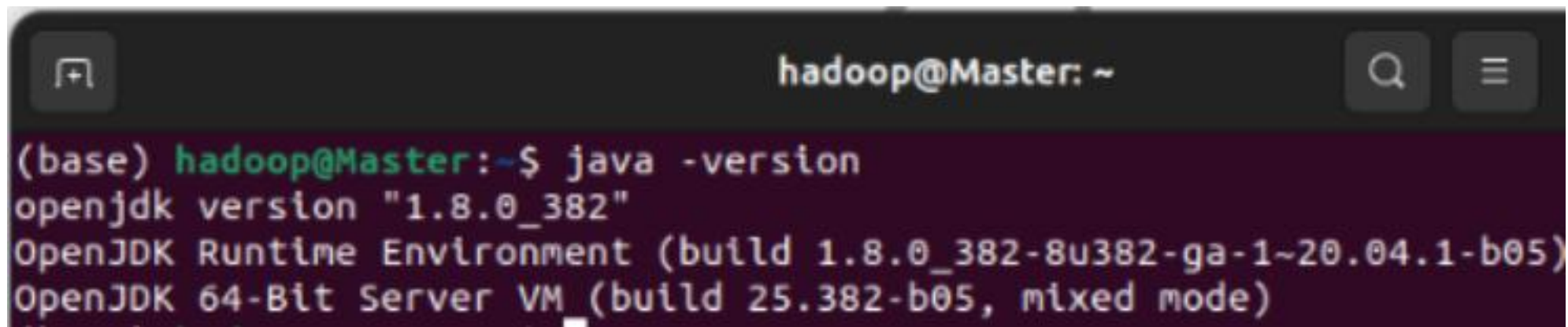
Hadoop安装成功

Index of /hadoop/common/hadoop-3.3.5			
Name	Last modified	Size	Description
Parent Directory	-	-	-
CHANGELOG.md	2023-03-15 19:35	53K	
CHANGELOG.md.asc	2023-03-15 19:35	833	
CHANGELOG.md.sha512	2023-03-15 19:35	153	
RELEASENOTES.md	2023-03-15 19:35	4.4K	
RELEASENOTES.md.asc	2023-03-15 19:35	833	
RELEASENOTES.md.sha512	2023-03-15 19:35	156	
hadoop-3.3.5-march64.tar.gz	2023-03-24 10:56	692M	
hadoop-3.3.5-march64.tar.gz.asc	2023-03-24 10:56	833	
hadoop-3.3.5-march64.tar.gz.sha512	2023-03-24 10:56	168	
hadoop-3.3.5-rat.txt	2023-03-15 19:35	2.0M	
hadoop-3.3.5-rat.txt.asc	2023-03-15 19:35	833	
hadoop-3.3.5-rat.txt.sha512	2023-03-15 19:35	161	
hadoop-3.3.5-site.tar.gz	2023-03-15 19:35	38M	
hadoop-3.3.5-site.tar.gz.asc	2023-03-15 19:35	833	
hadoop-3.3.5-site.tar.gz.sha512	2023-03-15 19:35	165	
hadoop-3.3.5-src.tar.gz	2023-03-15 19:35	35M	
hadoop-3.3.5-src.tar.gz.asc	2023-03-15 19:35	833	
hadoop-3.3.5-src.tar.gz.sha512	2023-03-15 19:35	164	
hadoop-3.3.5.tar.gz	2023-03-15 19:35	674M	
hadoop-3.3.5.tar.gz.asc	2023-03-15 19:35	833	
hadoop-3.3.5.tar.gz.sha512	2023-03-15 19:35	160	

```
hadoop@windupbird-virtual-machine: /usr/local/hadoop
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/icon_success_sml.gif
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/expanded.gif
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/external.png
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/icon_info_sml.gif
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/logo_apache.jpg
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/bg.jpg
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/newwindow.png
hadoop-3.3.5/share/doc/hadoop/hadoop-hdfs-nfs/images/h3.jpg
hadoop@windupbird-virtual-machine:~$ cd /usr/local/
hadoop@windupbird-virtual-machine: /usr/local$ sudo mv ./hadoop-3.3.5/ ./hadoop
hadoop@windupbird-virtual-machine: /usr/local$ ls
bin  etc  games  hadoop  include  lib  man  sbin  share  src
hadoop@windupbird-virtual-machine: /usr/local$ sudo chown -R hadoop ./hadoop
hadoop@windupbird-virtual-machine: /usr/local$ cd ./hadoop
hadoop@windupbird-virtual-machine: /usr/local/hadoop$ ./bin/hadoop version
Hadoop 3.3.5
Source code repository https://github.com/apache/hadoop.git -r 706d88266abcee09e
d78fbaa0ad5f74d818ab0e9
Compiled by stevel on 2023-03-15T15:56Z
Compiled with protoc 3.7.1
From source with checksum 6bbd9afcf4838a0eb12a5f189e9bd7
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3
.3.5.jar
hadoop@windupbird-virtual-machine: /usr/local/hadoop$
```

# 测试环境

Java环境：下载自openjdk-8-jdk，为1.8.0\_382

A terminal window with a dark background. The title bar shows 'hadoop@Master: ~'. The prompt is '(base) hadoop@Master:~\$'. The command 'java -version' has been executed, resulting in three lines of output: 'openjdk version "1.8.0\_382"', 'OpenJDK Runtime Environment (build 1.8.0\_382-8u382-ga-1~20.04.1-b05)', and 'OpenJDK 64-Bit Server VM (build 25.382-b05, mixed mode)'.

```
(base) hadoop@Master:~$ java -version
openjdk version "1.8.0_382"
OpenJDK Runtime Environment (build 1.8.0_382-8u382-ga-1~20.04.1-b05)
OpenJDK 64-Bit Server VM (build 25.382-b05, mixed mode)
```



# 测试环境

## Spark: 3.5.0, 用到其中的pyspark



Download Libraries Documentation Examples Community Developers

### Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.5.0-bin-without-hadoop.tgz](#)
4. Verify this release using the 3.5.0 [signatures](#), [checksums](#) and [project release KEYS](#) by following these [procedures](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

### Link with Spark

Spark artifacts are [hosted in Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark  
artifactId: spark-core_2.12  
version: 3.5.0
```

### Installing with PyPi

PySpark is now available in pypi. To install just run `pip install pyspark`.

### Convenience Docker Container Images

Spark Docker Container images are available from [DockerHub](#), these images contain non-ASF software and may be subject to different license terms.



```
windupbird@windupbird-virtual-machine: ~/Downloads/spark-3.2.4-bin-hadoop3.2  
windupbird@windupbird-virtual-machine:~/Downloads/spark-3.2.4-bin-hadoop3.2$ bin/pyspark  
Python 3.10.12 (main, Jun 11 2023, 05:26:28) [GCC 11.4.0] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
23/10/31 14:47:41 WARN Utils: Your hostname, windupbird-virtual-machine resolves to a loopback address: 127.0.1.1; using 192.168.198  
.129 instead (on interface ens33)  
23/10/31 14:47:41 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
23/10/31 14:47:43 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where  
applicable  
Welcome to  
  
██████████ version 3.2.4  
  
Using Python version 3.10.12 (main, Jun 11 2023 05:26:28)  
Spark context Web UI available at http://192.168.198.129:4040  
Spark context available as 'sc' (master = local[*], app id = local-1698734867559).  
SparkSession available as 'spark'.  
>>> S
```

# 测试环境

Python: Anaconda3 (2023.09-0)

其中Python版本为3.11.5



```
hadoop@Master: ~  
(base) hadoop@Master:~$ python3  
Python 3.11.5 (main, Sep 11 2023, 13:54:46) [GCC 11.2.0] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> 
```

```
hadoop@Master: ~/Downloads  
hadoop@Master:~$ cd ~/Downloads  
hadoop@Master:~/Downloads$ ls  
Anaconda3-2023.09-0-Linux-x86_64.sh  hadoop-3.3.5.tar.gz  
baidunetdisk_4.17.7_amd64.deb        spark-3.4.1-bin-without-hadoop.tgz  
hadoop@Master:~/Downloads$ sh ./Anaconda3-2023.09-0-Linux-x86_64.sh  
  
Welcome to Anaconda3 2023.09-0  
  
In order to continue the installation process, please review the license  
agreement.  
Please, press ENTER to continue  
>>>  
=====
```

End User License Agreement - Anaconda Distribution

```
=====
```

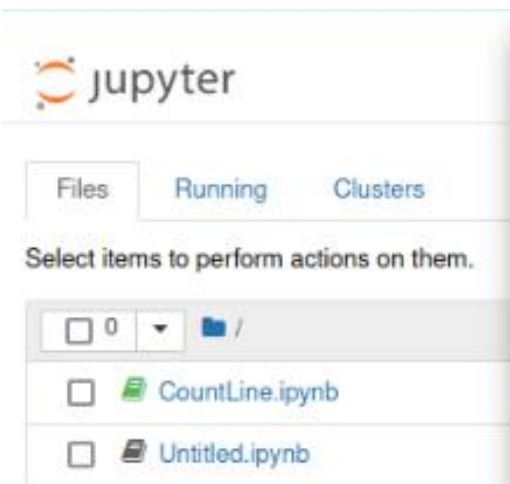
Copyright 2015-2023, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Distribution (which was formerly known as Anaconda Individual Edition).

# 测试环境

Jupyter notebook (和pyspark交互使用)



```
hadoop@Master: ~  
j(base) hadoop@Master:~$ jupyter notebook  
  
[W 23:50:09.791 NotebookApp] WARNING: The notebook server is listening on all IP  
addresses and not using encryption. This is not recommended.  
[W 23:50:10.418 NotebookApp] Loading JupyterLab as a classic notebook (v6) exten  
sion.  
[W 2023-11-02 23:50:10.422 LabApp] 'ip' has moved from NotebookApp to ServerApp.  
This config will be passed to ServerApp. Be sure to update your config before o  
ur next release.  
[W 2023-11-02 23:50:10.422 LabApp] 'password' has moved from NotebookApp to Serv  
erApp. This config will be passed to ServerApp. Be sure to update your config be
```



# 测试环境

Wordcloud和plotly





**PART02**

**数据集特征**

# 数据集特征

数据集1（100MB）：

豆瓣影评数据

爬虫获取+数据预处理

数据分析可视化

豆瓣影评

文件夹

D:\浏览器

99.7 MB (104,638,880 字节)

100 MB (104,857,600 字节)

豆瓣影评比较经典  
适合新手进行爬取  
我们用其来进行爬虫  
和数据处理初练习

## 数据集1

12313992142	力荐	2016-12-20 14:02:59	疯狂的赛车电影就退，那时候我们才开始
1521875200	很差	2017-02-14 20:09:11	没工作，却有房车周游世界谈恋爱，
0797495276	力荐	2016-08-31 17:38:03	开场大场面一镜到底歌舞段落惊艳，i
1536008174	力荐	2017-02-17 21:04:09	有多喜欢《爱乐之城》呢，走出电影
0797785152	推荐	2016-08-31 19:31:25	奥斯卡要横扫
153647446	很差	2017-02-17 22:30:51	沙拉盘群舞，民族大联欢，咱们老百姓
0797424114	力荐	2016-08-31 17:15:39	好听好看好甜好美的爱的四季爵士乐！
12644611418	较差	2016-12-26 12:20:41	#用光了九叔的25cents买票#演员肢
1278733103	还行	2017-01-15 23:55:24	We drifted apart while busy pursuing o
1507925133	推荐	2017-02-11 21:30:38	五年的时空在电光石火的瞬时凝视中
2987939151	力荐	2022-04-02 16:17:55	超级大延迟观影。前半段还并不很喜
1508121121	力荐	2017-02-11 21:59:50	不五星对不起现在的冲动，是虚高也
1522390194	力荐	2017-02-14 21:59:24	举重若轻，这哪里是爱情片，这是一
1404760182	力荐	2017-01-22 13:50:31	总有那么一个平行时空里，所有的好
0953251206	力荐	2016-10-09 00:01:42	最后10分钟太美，跟《妈咪》里那一
8964680172	还行	2021-05-30 15:38:31	意外的不喜欢，结尾丝毫不能感动我。
1494515181	力荐	2017-02-09 02:19:34	相信若干年后，这还会是一部对我意
1560828179	还行	2017-02-22 23:06:43	看豆瓣好友评分完全变成了窥私：这
0963284199	还行	2017-01-26 17:20:03	呵呵一个。内核如此单薄俗套，形式
1656901192	力荐	2017-03-17 09:09:48	和喜欢Before Sunrise系列一样喜欢。
1971502129	推荐	2017-06-01 13:58:18	精致的老套，有佳句无佳章
1525947162	推荐	2017-02-15 17:06:11	有情人终成前任，此去经年，在你当
1304802118	力荐	2017-01-02 06:21:29	太好看了，后半段基本是从头哭到尾。
0842307113	力荐	2016-09-13 14:07:26	又美又甜又让人心碎，一个“男孩遇上



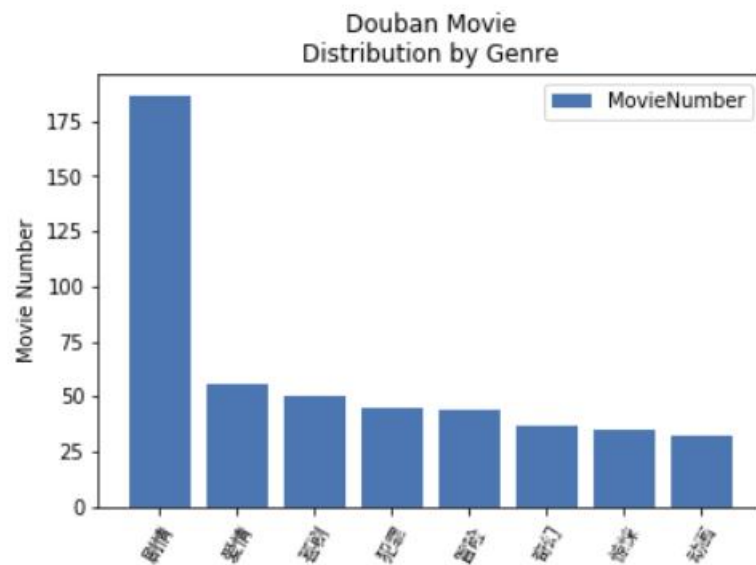
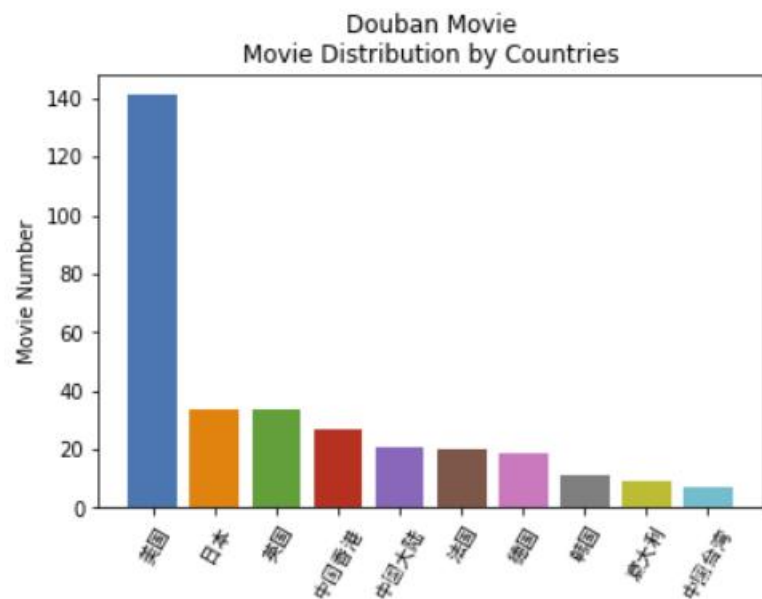
# 数据集特征

## 部分代码图

```
print(response.status_code)
# 解析页面数据
soup = BeautifulSoup(response.text, 'html.parser')
# 所有评论数据
reviews = soup.find_all('div', {'class': 'comment'})
print('开始爬取第{}页, 共{}条评论'.format(page, len(reviews)))
sleep(random.uniform(1, 2))
# 定义空列表用于存放数据
user_name_list = [] # 评论者昵称
star_list = [] # 评论星级
time_list = [] # 评论时间
ip_list = [] # 评论者ip属地
vote_list = [] # 有用数
content_list = [] # 评论内容
```

# 数据集特征

## 数据处理展示图



# 数据集特征

数据集2 (xxxMB) :

地震数据

网上获取+数据预处理

数据分析可视化

## 数据集2

贴个数据集图

使用爬虫爬取的数据集较小

为了进行更大的数据处理

我们从网上下载了数据集2





# **PART03**

## **选题目的**

# 选题目的

2008年5月12日，新中国成立以来破坏性最强、波及范围最广、救灾难度最大的里氏8.0级地震突袭四川汶川。面对突如其来的灾难，在中国共产党的正确领导下，广大军民临危受命、迎难而上，开展了一场抢救速度最快、动员范围最广、投入力量最大的救援斗争，彰显出“万众一心、众志成城，不畏艰险、百折不挠，以人为本、尊重科学”的抗震救灾精神。

抗震救灾精神同中华文明的基因禀赋一脉相承，是中华民族精神的生动写照。新时代我们要继续弘扬抗震救灾精神，深刻理解其科学内涵，善用抗震救灾历史经验，为全面建设社会主义现代化国家提供强大精神动力。

当地震来临时，人们往往陷入恐慌和绝望之中。然而，大数据分析地震信息的研究正在改变这一现状。通过收集、分析和解读大量的地震数据，科学家们能够准确地预测地震的发生时间、地点等，为人们提供宝贵的时间来做好准备和采取适当的防护措施。

这样的预测意味着更多的生命可以被拯救，更多的伤害可以被避免。它给灾民和相关救援组织提供了宝贵的信息，使他们能够更好地理解和应对灾情。这意味着更快速、更高效的救援行动，为被困和受伤的人们带来更多的希望和机会。

选题目的和对社会的意义是挽救生命、保护安全、给人们带来希望和保障。它代表了人类对自然灾害的不懈探求和努力，体现了科技的力量和智慧，让我们对未来充满信心和期待。



# 选题目的





# 选题目的



# 选题目的



# 选题目的



# **PART04**

## **相关应用**



# 相关应用

以下为项目所用到的应用：



## vmware

采用虚拟机来模拟非当前操作系统，通过在虚拟的操作系统环境中进行操作



## Hadoop

实现分布式文件系统，利用集群访问超大数据集并进行高速计算



## spark

支持分布式数据集上的迭代作业，是数据处理计算的引擎

在vmware上建立虚拟机，采用ubuntu系统。

建立的虚拟机一台命名为master，一台命名为slave。

在配置好换源，java环境后先进行单节点的hadoop伪分布式配置，而后进行两个节点的hadoop集群配置。

而后下载spark，主要使用其中的pyspark功能作数据分析。

启动hadoop集群后，将数据集上传到hdfs系统后，就可以开始进行数据分析。

# 相关应用

下载anaconda既提供各种python的库，还配置了jupyter。

将用于数据分析的python代码在Jupyter notebook上运行，并且可以得到可视化的结果。

以下为项目所用到的应用：

Wordclouds和plotly都是用于在jupyter notebook上展示可视化的库。



## anaconda

Anaconda是一个开源的Python发行版本，其包含了conda、Python等180多个科学包及其依赖项，提供pandas，numpy等库。Anaconda中已经集成了Jupyter Notebook

## wordclouds

wordcloud是优秀的词云展示第三方库，以词语为基本单位，通过图形可视化的方式，更加直观和艺术化的展示文本

## Jupyter notebook

交互式笔记本，提供实时代码运行和数据可视化的平台

用于可视化的库，开源、可交互、支持40余种图表类型，涵盖统计、金融、地理、科学和3D图表。



## PART04

# 课程及作业感悟





# 问题挑战



# 问题挑战



## PART04

# 课程及作业感悟





大数据导论作业汇报

**谢谢观看!**