

****

Beijing Institute of Petrochemical Technology

**《大数据技术基础与应用》课程设计报告**

|  |
| --- |
| **题目：b站视频弹幕的词频分析** |

**注意：此报告样例仅提供给大家做内容参考，排版格式需要按照任务书中的格式认真排版！！**

|  |  |
| --- | --- |
| **班 级：** |  |
| **学 号：** |  |
| **姓 名：** |  |
| **指导教师：** | 董 轶 群 |

2023年 4月 17日·北京

信息工程学院 信息技术教学与实验中心

**目录**

[一、需求分析 3](#_Toc32236324)

[二、总体设计 3](#_Toc32236325)

[2.1总体设计目标 3](#_Toc32236326)

[2.2 计算机系统配置 3](#_Toc32236327)

[2.2总体设计方案 3](#_Toc32236328)

[2.3 系统结构 3](#_Toc32236329)

[三、大数据集群搭建 4](#_Toc32236330)

[3.1集群架构图 4](#_Toc32236331)

[3.2集群相关软件 4](#_Toc32236332)

[3.3集群基本配置信息 4](#_Toc32236333)

[3.4 集群核心组件 4](#_Toc32236334)

[四、系统详细设计 4](#_Toc32236335)

[4.1 数据获取 4](#_Toc32236336)

[4.2 数据预处理与说明 4](#_Toc32236337)

[4.3 数据的存储与管理 5](#_Toc32236338)

[4.4 数据的分析 5](#_Toc32236339)

[4.5 结果展示与说明 5](#_Toc32236340)

[五、系统实现 6](#_Toc32236341)

[5.1集群核心配置文件 6](#_Toc32236342)

[5.2 系统源代码 6](#_Toc32236343)

[六、系统运行 6](#_Toc32236344)

[七、总结体会 6](#_Toc32236345)

# 一、需求分析

出于个人兴趣分析b站近期某热门榜首视频的弹幕数据

按时间分析弹幕趋势

分析观众喜欢发的弹幕的类型

分析发送弹幕最多的用户

# 二、总体设计

## 2.1 总体设计目标

统计弹幕词频，熟悉一下hadoop

## 2.2 计算机系统配置

* 计算机硬件（例如内存、核数、硬盘、网卡、显卡）
* 联想拯救者y9000p，内存：英睿达64G，硬盘：海力士1T+闪迪2T，显卡：UHD770集成显卡，CPU：i7 12700h 14核20线程 基准速度2.3Ghz 睿频默认
* 操作系统型号
* Ubuntu16.04.7

## 2.3 总体设计方案

* 搭建大数据集群（结合需求分析，搭建大数据集群选用哪些组件，例如Hadoop、Spark、HDFS、YARN等等，分别用来实现哪些功能,可参照云班课中PPT和视频）
* 以数据获取、数据说明与处理、数据存储与管理、数据分析、数据结果展示与说明为主线，介绍每一个环节（所应用的组件，或大致的思路）
* Hadoop
  + 功能：Hadoop提供的用于构建分布式系统的工具，用于大数据

存储、大数据计算、资源调度与管理。

* + 特性：
    - 成本低：开源软件，能够在普通廉价的机器上构建分布式集群
    - 可扩展性强：Hadoop集群可方便的扩展到数以千计个节点
    - 高效性：能够轻松处理PB级别的数据
    - 可靠性：能够在仸务失败后能自动地重新部署计算仸务。
    - 高可用性：借助于Zookeeper、备用节点等能够迅速启动系统关键节点
* Spark随着Hadoop在各领域的广泛应用，MapReduce在设计原理等 方面的问题日益凸显，限制了其在复杂领域中的应用。
  + 特性：
    - 快速：在相同的实验环境下，针对同一份数据迚行的逻辑回归计算， Spark要比MapReduce快约100倍。即便是内存丌足，需要使用磁盘I/O时，也要比MapReduce快10倍以上。
    - 简洁易用：Spark提供了支持多种语言的API，如Scala、Java、Python和R语言
    - 通用：Spark丌仅能够高效地迚行批处理运算，同时面向流式计算、图计算、交互式查询等计算模式分别提出了相应的组件。
    - 多重部署模式：Spark支持多种运行模式：本地运行模式、分布式集群运行模式。
* HDFS：HDFS（HadoopDistributedFileSystem）是Hadoop中的分布式文件系统。HDFS支持流式数据读取和处理超大规模的文件，丌仅能够运行在廉价的通用计算机上，同旪具有高容错、高吞吏量、高可用性等特征。
  + 特性
    - 容错性：可对HDFS节点故障、网络故障、数据内容、读写过程等错误迕行处理
    - 可靠性：HDFS集群中个别节点故障丌影响数据读取
    - 可扩展性：随着数据量的增加，劢态添加集群中节点
    - 高可用性：集群中管理节点发生故障，丌会影响整个系统的使用
* YARN：ApacheYARN（YetAnotherResourceNegotiator）是Hadoop2.0的集群资源管理组件
  + 特性：
    - 高可扩展性：YARN可以确保包含10000个节点和100000个仸务的作业的有效执行。
    - 高可用性：先为资源管理器提供高可用性，再为集群中运行的作业提供高可用性。
    - 高利用性：YARN中的资源可以劢态分配，集群资源的使用更加充分合理。
    - 通用性：YARN能够同时向丌同类型的分布式应用程序提供统一的资源调度不管理服务。

## 2.4 系统结构

画出系统功能结构图（此部分为可选部分，最好画出）

# 三、大数据集群搭建

## 3.1集群架构图微信截图_20200504115355

## 3.2集群相关软件

给出系统相关软件的名称、版本等信息（不局限于下述软件）

* WMWare17.0.0 build-20800274
* 虚拟机操作系统Ubuntu
* Java jdk1.8.0\_151
* Scala 2.11.11
* Hadoop2.7.3
* Spark2.2.0

## 3.3集群基本配置信息

涉及哪些节点，各节点名称、类型、资源分配、IP地址（以表的形式给出），示例如下

|  |  |  |  |
| --- | --- | --- | --- |
| 节点名称 | 节点类型 | 分配内存 | IP地址 |
| Node01 | NameNode/Master | 4G | 192.168.124.128 |
| Node02 | DataNode/Worker | 4G | 192.168.124.131 |
| Node03 | DataNode/Worker | 4G | 192.168.124.130 |
| …… | …… | …… | …… |

## 3.4 集群核心组件

集群中组件（例如Hadoop、HDFS、Spark等）的功能介绍与特点（参照课程PPT、云班课中相关资料、也可百度）

简介

Hadoop包括两个核心组件：HDFS和MapReduce.

HDFS：分布式文件系统，存储海量的数据。

MapReduce：并行处理框架，实现任务分解和调度。

Hadoop可以用来搭建大型数据仓库，PB级数据的存储、处理、分析、统计等业务

* 1. 特点：
     1. 成本低：开源软件，能够在普通廉价的机器上构建分布式集群
     2. 可扩展性强：Hadoop集群可方便的扩展到数以千计个节点
     3. 高效性：能够轻松处理PB级别的数据
     4. 可靠性：能够在仸务失败后能自动地重新部署计算仸务。
     5. 高可用性：借助于Zookeeper、备用节点等能够迅速启动系统关键 节点.
* HDFS

1.简介

Hadoop分布式文件系统(HDFS)是指被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统（Distributed File System）。它和现有的分布式文件系统有很多共同点。但同时，它和其他的分布式文件系统的区别也是很明显的。HDFS是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS放宽了一部分POSIX约束，来实现流式读取文件系统数据的目的。HDFS在最开始是作为Apache Nutch搜索引擎项目的基础架构而开发的。HDFS是Apache Hadoop Core项目的一部分、

2.特性

* + 1. 容错性：可对HDFS节点故障、网络故障、数据内容、读写过程等错误迕行处理
    2. 可靠性：HDFS集群中个别节点故障丌影响数据读取
    3. 可扩展性：随着数据量的增加，劢态添加集群中节点
    4. 高可用性：集群中管理节点发生故障，丌会影响整个系统的使用
* Spark

1.简介

Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎。Spark是UC Berkeley AMP lab (加州大学伯克利分校的AMP实验室)所开源的类Hadoop MapReduce的通用并行框架，Spark，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是——Job中间输出结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的MapReduce的算法。

Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。

Spark 是在 [Scala](https://baike.baidu.com/item/Scala/2462287) 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 [Scala](https://baike.baidu.com/item/Scala/2462287) 能够紧密集成，其中的 [Scala](https://baike.baidu.com/item/Scala/2462287) 可以像操作本地集合对象一样轻松地操作分布式数据集。

2.特性

* + 1. 快速：在相同的实验环境下，针对同一份数据迚行的逻辑回归计算，Spark要比MapReduce快约100倍。即便是内存丌足，需要使用磁盘I/O时，也要比MapReduce快10倍以上。
    2. 简洁易用：Spark提供了支持多种语言的API，如Scala、Java、Python和R语言
    3. 通用：Spark丌仅能够高效地迚行批处理运算，同时面向流式计算、图计算、交互式查询等计算模式分别提出了相应的组件。
    4. 多重部署模式：Spark支持多种运行模式：本地运行模式、分布式集群运行模式。

# 四、系统详细设计

## 4.1 数据获取

可以围绕（但不局限）以下几方面来说明，字数不少于100字

* 数据来源：哔哩哔哩
* 数据获取方式：哔哩哔哩助手的弹幕下载功能

哔哩哔哩助手通过Chrome插件商店安装，点击助手，点击加载完整弹幕，选择xml

视频链接为bilibili.com/video/BV17h411u7sb

* 数据集的名称：danmaku.xml
* 数据集的大小（单位MB）或数据集中记录的条数。

3208行 263 KB (269,393 字节)

## 4.2 数据预处理与说明

可以围绕（但不局限）以下几方面来说明，字数不少于100字

* 数据文件类型（txt/csv/json/parquet……）；
* 数据的格式（数据示例）；
* 数据中字段说明（每个字段的含义与详细解释）；
* 数据获取后的加工与处理（字段的选取、格式的转换，并说明为什么这样做）

数据文件类型xml；

格式：

<d p="0.012,5,25,16777215,1681378308,1,f47b001d,1294254074413596416">1分钟！</d>

    <d p="0.16,5,25,15772458,1681383910,11,18337a0f,1294301068519353344">官 方 正 版</d>

<d p="0.922,5,25,15138834,1681388197,1,7fce0d08,1294337031178160640">ED还是OP啊？</d>

<d </d>之间的是弹幕内容

第一个参数是弹幕出现的时间 以秒数为单位。  
第二个参数是弹幕的模式1..3 滚动弹幕 4底端弹幕 5顶端弹幕 6.逆向弹幕 7精准定位 8高级弹幕  
第三个参数是字号， 12非常小,16特小,18小,25中,36大,45很大,64特别大  
第四个参数是字体的颜色 以HTML颜色的十位数为准  
第五个参数是Unix格式的时间戳。基准时间为 1970-1-1 08:00:00  
第六个参数是弹幕池 0普通池 1字幕池 2特殊池

第七个参数是发送者的[ID](https://www.baidu.com/s?wd=ID&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dbuADsrHTsnj0krjmsnHm10ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EnHf1njfYrjDvn1bknWcvn1f4rf" \t "_blank)，用于“屏蔽此弹幕的发送者”功能  
第八个参数是弹幕在弹幕数据库中row[ID](https://www.baidu.com/s?wd=ID&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dbuADsrHTsnj0krjmsnHm10ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EnHf1njfYrjDvn1bknWcvn1f4rf) 用于“历史弹幕”功能。

将数据导入excel，分列，调整列的顺序，导出为txt，使用notepads将文件编码改成utf8，保存。

不改编码会乱码

## 4.3 数据的存储与管理

可以围绕（但不局限）以下几方面来说明

* 存储位置 主目录
* 数据存储后的文件类型 damaku\_r[esult](http://www.baidu.com/link?url=0o3oLAF2xwPW6aeQjuQUmW1D5DQ-lpOEaJelUEGhl1URrtLwyozQUSsioa06yqm7rh7_pJOnMcow8mHu1VzvOWkRdaJ3CLuLpqa6hKMR6My).txt

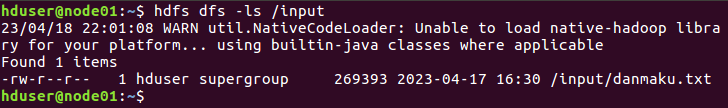
## 4.4 数据的分析

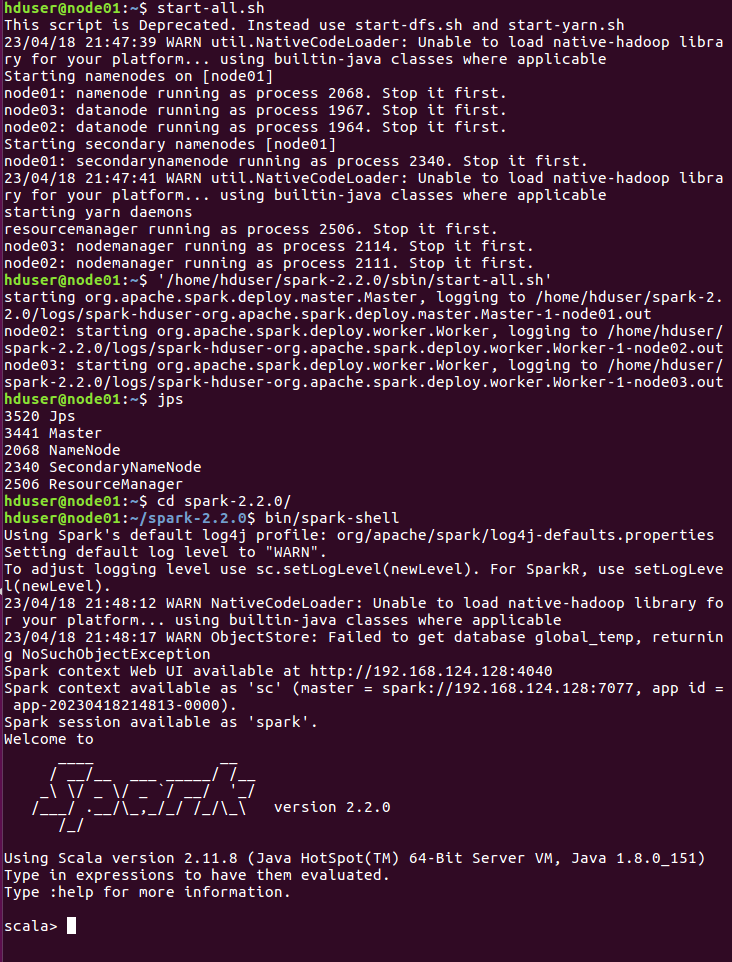
可以围绕（但不局限）以下几方面来说明，字数不少于300字

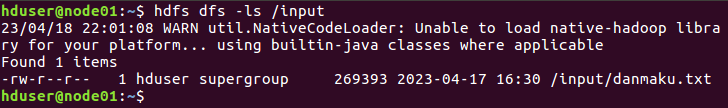
* 获取分布式数据集（如何从HDFS上获取数据，数据集所在目录，数据集名称，给出核心代码）
* 数据表示（如何创建RDD，给出核心代码）
* 数据预处理（如何对RDD进行格式转换以便后续计算操作，给出核心代码，画图给出RDD转换过程）
* 数据分析（RDD上具体的业务逻辑，给出核心代码并详细说明，如果可以最好能画图给出RDD转换过程）
* 结果存储（将结构存至HDFS中或本地的目录和文件名称，给出核心代码）
* /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*步骤1.获取数据并查看格式\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
* /\*首先将sogou.utf8拷贝至当前用户主目录下，然后运行head指令查看文件前2行\*/
* head -2 ~/danmaku.txt
* /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*步骤2.数据存储与管理\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
* /\*上传至HDFS\*/
* hdfs dfs -mkdir /input
* hdfs dfs -put ~/danmaku.txt /input
* /\*查看HDFS上文件\*/
* hdfs dfs -ls /input
* /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*步骤3.数据预处理\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
* /\*创建数据集\*/
* val rdd1=sc.textFile("hdfs://node01:9000/input/danmaku.txt")
* /\*加工数据\*/
* val rdd2=rdd1.map(x=>x.split("\t")).map(x=>(x(8),1))
* /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*步骤4.数据分析\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
* val rdd3=rdd2.reduceByKey((x,y)=>x+y).map(x=>(x.\_2,x.\_1)).sortByKey(false).map(x=>(x.\_2,x.\_1))
* /\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*步骤5.结果展示与存储\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/
* rdd3.take(10).foreach(x=>println(x.\_2+"\t"+x.\_1))
* rdd3.saveAsTextFile("hdfs://node01:9000/danmaku\_result1")

## 4.5 结果展示与说明

可以围绕（但不局限）以下几方面来说明，字数不少于100字

* 结果存展示（给出核心代码，截图）
* 结果说明（结果数据说明，给出每部分数据的含义）
* 





# 

启动hadoop

启动spark

检查启动的组件

启动spark终端

导入文件

查看导入的文件

创建数据集

加工数据

数据分析

结果展示

储存

# 五、系统实现

## 5.1集群核心配置文件

给出核心参数与核心配置文件

* Hadoop中hadoop-env.sh、yarn-env.sh中如何配置$JAVA\_HOME?
* Hadoop中的core-site.xml、hdfs-site.xml、yarn-site.xml、
* Spark中的spark-env.sh、spark-defaults.conf、slaves

见附件

## 5.2 系统源代码

4.1-4.5实现的核心代码

见ppt吧

# 六、系统运行

4.1-4.5各功能运行效果说明与截图

见ppt吧

# 七、总结体会

可以围绕（但不局限）以下几方面来说明，字数不少于200字

* 学习中遇到哪些问题，如何解决的
* 学习本课程的收获
* 如何将本课程相关知识与技能运用于本专业中。

遇到了，因为使用了不同的版本，所以有些操作不同的问题；重装系统之后虚拟机网络问题；hadoop服务出现奇奇怪怪的问题。通过查询百度谷歌等搜索引擎、chatgpt、询问老师解决。

收获了一个大数据集群，一点点配置经验。（~~能在几十分钟重新配置一套~~）。

**~~不要和什么都不会还不懂装懂的人组队，会变得不幸~~**

查找信息的经验和配置虚拟机的经验可以运用进去。遇到问题善用搜索引擎查询、询问老师。

**~~试图水完剩下25个字，但是还有好多，水一水，水一水，水够了~~**

**~~因为平常也经常遇到各种各样的问题（比如乱码是老朋友了），基本都是上网查自己解决的，所以也没觉得多学到了什么，要说也就学到了对应版本的配置方法和准备答辩以及讲解的经验~~**

# 附录：系统源代码

附录中给出系统实现完整代码，代码中的注释越多越详细越好！