# 分布式与并行计算项目设计报告书

### 班级

* 大数据二班

### 小组成员（学号）

* 陈刚刚（2018080910007）
* 杨峰（2018080910162）
* 杨化翰（2018080910166）
* 徐彬迅（2018080910151）
* 杨春春（2018080910160）

## 前言

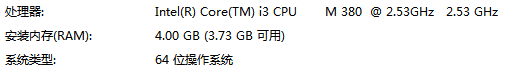
这个地方对项目进行一些描述，例如这个项目要涉及到什么方面的技术

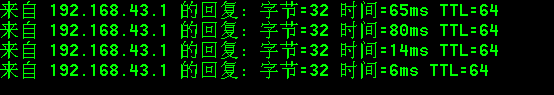
1. 我们在着手这个项目中首先了解这个项目的整体结构和需要准备的系统和数据，按以下步骤
2. 安装lunix系统，搭建Hadoop的环境，安装Java，将Hadoop导入lunix系统里，再利用Java技术来检测数据
3. 中间件技术。中间件作为构建，是分布式系统中介于应用层和网络层的一个功能层次,它能够屏蔽操作系统（或网络协议）的差异，实现分布式异构系统之间的互通或互操作。分布式 应用软件借助这种软件在不同的域之间共享资源。中间件位于客户机服务器的操作系统之上，管理计算资源和网络通信。它扩展了C/S结构,形成了一个包括客 户端、中间件和服务器的多层结构。  
     
   基于中间件的分布式计算技术以中间件为桥梁,通过把数据转移到计算之处的计算方式,把网络系统的所有组件集成为一个连贯的可操作的异构系统,从而达到网络“透明”的目的。  
     
   早期的中间件比较简单，仅具有消息通讯和事务管理功能。但随着需求的多样化，也产生了许多不同类别的中间件产品，如：消息中间件(MOM ：Message-Oriented Middleware)、数据库中间件(Database Middleware)、远程过程调用中间件（RPC ：Remote Process Call）、对象请求代理中间件(ORB ：Object Request Broker)和事务处理中间件(TP Monitor：Transaction Process Monitor)。）、网络、分布式、数据等进行描述。
4. 搭建互通局域网络。以保证A局域网路由器能让B局域网主机访问，A局域网路由器操作虚拟服务器映射
5. 在通过分布式系统，分布式数据库应用数据

## 项目框架介绍

这个地方首先写你们这个项目所有节点的架构，最好能够用图表的方式来表示，要能体现两个方面的信息：

1. 每个节点的硬件性能





1. 节点的网络框架和网络传输性能

节点的网络框架: 传输层-应用层：定义应用程序的功能， 剩下三层主要通过网络的端到端的数据流。OSI七层模型是一个理论模型，更多的则是把它作为分析、评判各种网络技术的依据。

物理层 为数据链路层提供物理链接，在其串行传送比特流(传送数据的单位比特)。  
数据链路层 负责在网络节点间德尔线路上通过检查、流量控制和重发手段，无差错的传送数据(帧单位)。为做到这一点，在每一帧中必须同时带有同步、地址、差错控制及流量控制等信息。  
网络层 网络层的任务就是选择合适的路由和交换节点，使数据源在传送层传下来的分组信息能够无误的按照地址找到目的地，并交付给相应的传输层，也就是完成网络的寻址功能。  
传输层 传送层是高低层之间的接口层。数据传送的单位是报文，当报文较长时将它分割若干组，然后交给网络层进行传送。传输层是计算机网络协议最关键的一层，该层以上的个各层不在管理信息传输问题。  
会话层 该层对传输的报文提供管理服务，在两个不同的系统互相通信的应用进程之间建立、组织、协调交互。列如：确定是双工或者是半双工工作  
表示层 该层的任务是把传送的数据的抽象语法变为传送语法、即把不同计算机内部不同表示形式转换为通信中的标准鄙视形式。此外对传送的数据加密或解密、正文压缩或解压。  
应用层 该层直接面向用户，是OSI中最高层。为用户提供应用的接口，既提供不同计算机间的文件传送、访问与管理、电子邮件的处理、不同计算机通过网络交互访问的需您终端功能。

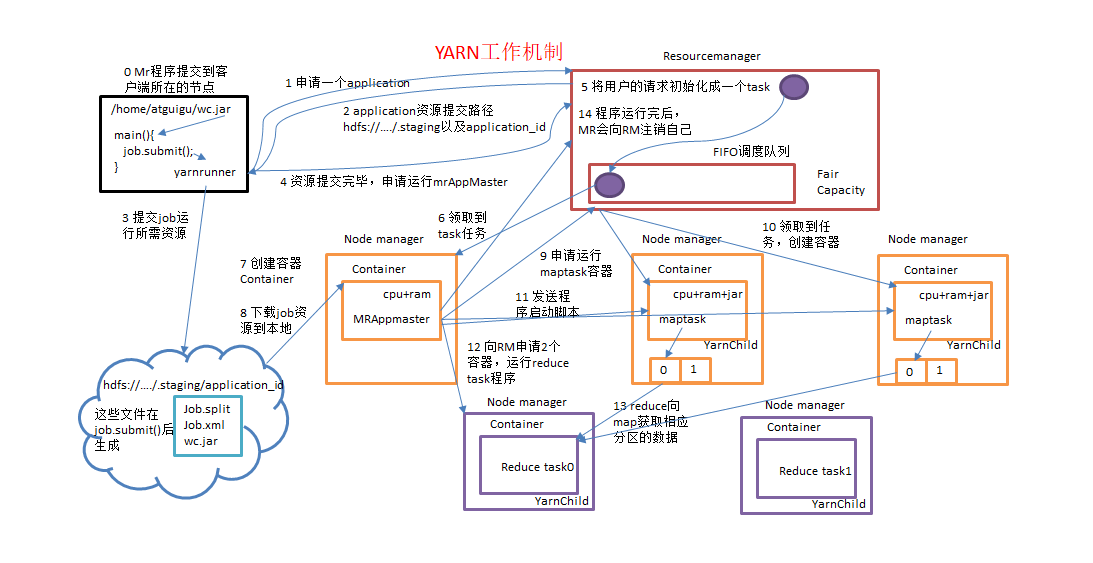
网络传输:是指用一系列的线路(光纤，双绞线等)经过电路的调整变化依据网络传输协议来进行通信的过程。其中网络传输需要介质，也就是网络中发送方与接收方之间的物理通路，它对网络的数据通信具有一定的影响。常用的传输介质有:双绞线、同轴电缆、光纤、无线传输媒介。网络协议即网络中(包括互联网)传递、管理信息的一些规范。

网络传输过程: 数据从上层进入到传输层，加上源端口和目标端口成为数据段(如果是UDP则成为数据报)，再进入网络层加上源IP和目标IP成为数据包，再进入链路层加上源MAC地址和目标MAC地址成为数据帧，这段过程是一种"加头"封装数据的过程。数据经过网络传输到达目标主机后，逐层"剃头"解包，最终得到data纯数据内容。

## 项目技术介绍

hadoop的架构核心包括：

1. 分布式文件系统（HDFS）：主要负责集群数据的存储和读取
2. 分布式计算系统（MapReduce）：实现任务分解和调度
3. 分布式资源管理系统（YARN）：负责海量数据的资源调度



利用集群的威力进行海量数据的高速运算和存储，使得Hadoop基本框架在用java编写数据时具有高可靠性，高扩展性，高效性和高容错性

## 项目实施过程

在这个项目开始之前，我们采纳老师的意见准备一个U盘,网线和检测器等等，在一切工具准备完毕之后检测网线是否接通。在网线接通之后我们在老师建议下在https://man.linuxde.net/download/下载Linux系统并在老师指导下进行安装，在Linux安装工程中最难的就是里面的选项，在一层一层的选项中我们最终安装好系统，把前期工作都准备好之后，接下来的最头痛的问题来里网络搭建过程，网络搭建步骤：

1. 准备3台已经安装好的Linux系统
2. 前期准备好的网线和路由器分别插入三台电脑中，实现两两互通
3. 添加VMnet2（一号机）和VMnet3（二号机），并取消其DHCP功能
4. 解决网卡不一致的问题，为其设置静态IP、子网掩码和网关

利用命令：gedit/ect/sysconfing/network-scripts/ifcfg配置网址

！q保持退出，将内网虚拟机加入VMnet2并重启网络服务，最后利用ifconfig查看是否获取到设置的IP地址

1. 运用以上步骤将VMnet3加入进来
2. 在内网虚拟机上Ping外网虚拟机，

网络搭建完成，配置好网络中离不开团队的合作，有的准备网线。有的准备数据，每一步都在老师辛苦指导下完成，每一步都离不开老师的帮助，在上一步的基础上我们开始着单机模式脚本的搭建，在大家一致的意见下我们搭建的步骤如下：

在准备好的hadoop和Java进行安装。解压Hadoop利用移动xftp6进行移动解压命令： tar -xvf hadoop-1.1.2.tar.gz

mv hadoop-1.1.2 /opt/hadoop

最后进行Java的安装和环境配置

1.新建目录cg\_single，将下载好数据tar.gz拷进该目录。

2.解压数据

3.在cg目录下新建data，logs两个文件夹。

4.进入cg/conf目录，把zoo\_sample.cfg文件复制一份名字改成zoo.cfg。

cp zoo\_sample.cfg zoo.cfg

5.修改zoo.cfg文件，需要修改以下几个地方。

dataDir=/data/cg/cg-3.4.9/data

dataLogDir=/datacg//zookeeper-3.4.9/logs

clientPort=2181

6.配置完以后，就可以启动zookeeper服务了，进入cg/bin目录，启动zookeeper服务器。

7.启动完成后，查看服务状态。

8.写入数据保持并退出

单机模式建立之后在三台计算机上调试并检测完毕，接下来对完全分布式模式调试过程，步骤如下：

1. 关闭三台计算机的防火墙，以防止打开的端口的太多，干扰试验。
2. 三台计算机之间打开免密登录
3. 配置Java\_HOME在${JAVA\_HOME}/ect/hadoop目录下的三个脚本：

Hadoop-env.sh

Yarn-env.sh

Mapred-env.sh

都需要配置JAVA-HOME

1. 三个节点职能分配
2. 集群配置
3. 复制配置到其他节点
4. 群起集群
5. 用yarn的主机节点查看yarn资源分配情况，用Namnode的主机节点查看hdfs情况
6. DataNode也可以配置成多个目录，每个目录存储的数据不一样。
7. 退役节点
8. 添加访问NameNode的节点
9. 数据调试

在团队的共同努力下完成最后的单机模式的调试过程，最终收获满满，数据的读取和运行过程取得圆满成功，虽然在数据读取过程中多次失败，有的是命令错误。有的是数据下载错误。甚至导致虚拟机崩溃，系统重装和环境变量的配置都多次失败，最后上网查资料和请教老师问题才得以解决。

## 项目结果分析和总结

**一,课程设计**

在这次项目中缺少技术积累和经验，导致不同的环节成功性很低这个问题很明显，而且在以前学习中也很普遍。对于技术的积累，涉及到知识库的建立，包含Linux系统、单机模式的搭建，特定问题的处理方法，技术的积累。由于没有知识库的积累，所以在不同项目中，成功的模块特别少，没有充分利用已有的资源。同时，一旦项目组人员有懒惰，对项目的影响特别大。知识库的建立方法：对于知识库的建立，因为工作量大，如果单靠一个人来做，肯定是低效的，而且还不一定成功。知识库内容的产生，必须由全体成员分享。前期做重点Linux系统的安装、技术的积累，记录,数据的收集和网线，路由器的准备，在这个项目中遇到很多棘手的问题，比如Linux系统的安装，提到系统的安装，里面涉及到的选项最让人头痛，比如分区问题,格式化高级问题,由于虚拟机没有学好导致很多的命令忘记,比如：文件的增删改查，接下来就是各种的环境变量的配置问题，首先你得先弄明白这个文件的根目录在什么地方没在利用根目录进行配置，多次的失败，在老师指导下利用% 名字 %进行配置，接下来的网络搭建过程中，由于配置的IP问题，和shell命令多次失败告终，缺少网络连接经验的积累和改进目前对于项目的认识，还比较混乱，缺少一套正规的、合适的方法，对于搭建和测试，完全凭已有的知识进行，导致项目进度、懒惰行为等不可控。其次学习JAVA技术很重要,因为在这次项目中我们对Java技术的认知及其少导致在运行代码过程中多次失败,配置Hadoop在Java上的过程中需要运用Java技术进行调试,因此在将来的学习中,Java技术是我们学习的一部分.网关和节点多次让我们陷于放弃状态,多次的连接失败和节点多次出现问题让团队灰心,在多次上网查阅质料和翻阅书籍都难以解决,最后在学长和老师的帮助下才得以解决,,虽然现在对项目有一点认识，但是对于新来的项目，又是按照之前的方式，得不到有效的改善。改进方法：还得从多向老师学习和长辈们学习，然后进行项目的总结，形成文档。并且要善于发现、多听多学并在以后的生活中养成做笔记的好习惯，让问题尽量都暴露出来，然后解决并整理总结，然后在后续同样的问题上进行改进。每一次项目必须要总结至少一次，并记录怎么做的,怎么解决的。并且在后续的学习生涯中对学科的挑战并吸取经验。Linux系统越来越受到电脑用户的欢迎，于是很多人开始学习Linux。Linux系统之所以会成为目前最受关注的系统之一，主要原因是它的免费，以及系统的开放性，可以随时取得程序的原代码，这对于程序开发人员是很重要的，同学和老师选择Linux不仅仅它可以实现多用户多任务最主要的它的可靠性和良好的稳定性，因此我们相信在今后的科学发展中离不开大数据和各种科学技术。最后在学习中一个好的老师和他正确引导很重要，我相信在老师的带领下学到很多东西和知识。