“大数据技术应用”
样题

任
务
书

模块一： 大数据集群搭建

题目一： 基础配置

题目说明：

1．比赛框架

本次比赛为分布式集群搭建，共三台节点，其中 master 作为主节点，slave1、slave2 为从节点；

2．比赛内容

基础配置：修改主机名、主机映射、时区修改、时间同步、定时任务、免密访问；

JDK 安装：环境变量；

Zookeeper 部署：环境变量、配置文件 zoo.cfg、myid；

Hadoop 部署：环境变量、配置文件修改、设置节点文件、格式化、开启集群；

Hive 部署：Mysql 数据库配置、服务器端配置、客户端配置。

任务一： 基础环境配置

任务说明：

相关安装包已经存放至环境/usr/package277/中

对应 ntp 和 mysql 已安装，可直接对其进行操作和配置

1.修改主机名，便于识别节点；

2.工具包已保存在环境中；

3.修改 hosts 文件，添加集群节点映射，按照给出的节点 IP 和对应的主机名进行设置；

4.要求各节点时区修改为中国时区（ 中国标准时间 CST+8）

5.安装 ntp 服务，要求主节点 master 为本地时钟源，从节点设置定时任务同步本地时间；

6.集群中数据传输需要节点之间免密访问，要求设置主节点之间到从节点的

免密访问；

7.Hadoop 技术基于 Java 语言，要求本地源下载对应安装包进行安装配置，注意安装路径要求，无需更改文件名，注意添加环境变量。

本环境用于为基础设置部分，用于后续的集群搭建。

任务要求：

1．按照左侧虚拟机名称修改对应主机名（分别为 master、slave1、slave2，使用 hostnamectl 命令）。

2．修改 host 文件添加左侧 master、slave1、slave2 节点 IP 与主机名映射（使用内网 IP）。

3．时区更改为上海时间（CST+0800 时区）。

4．环境已经安装 NTP，修改 master 节点 NTP 配置，设置 master 为本地时间服务器，屏蔽默认 server，服务器层级设为 10。

5．开启 NTP 服务。

6．添加定时任务--在早十-晚五时间段内每隔半个小时同步一次本地服务器时间（24 小时制、使用用户 root 任务调度 crontab，服务器地址使用主机名）。

7．主节点生成公钥文件 id_rsa.pub(数字签名 RSA，用户 root，主机名 master)。

8．建立master自身使用root用户 ssh 访问 localhost 免密登录。

9．建立master 使用root用户到 slave1 的 ssh 免密登录访问。

10．建立master 使用root用户到 slave2 的 ssh 免密登录访问。

11．将 jdk 安装包解压到/usr/java 目录(安装包存放于/usr/package277/，路径自行创建，解压后文件夹为默认名称，其他安装同理）。

12．配置系统环境变量 JAVA_HOME，同时将 JDK 安装路径中 bin 目录加入 PATH 系统变量，注意生效变量，查看 JDK 版本。

题目二： Zookeeper 搭建

题目说明：

Zookeeper 是一个分布式服务框架，是 Apache Hadoop 的一个子项目，它主要是用来解决分布式应用中经常遇到的一些数据管理问题，如：统一命名服务、状态同步服务、集群管理、分布式应用配置项的管理等。

预装的配置文件 zoo_sample.cfg 下面默认有五个属性，分别是：

1．tickTime

心跳间隔，单位是毫秒，系统默认是 2000 毫秒，也就是间隔两秒心跳一次。

tickTime 的意义：客户端与服务器或者服务器与服务器之间维持心跳，也就是每个 tickTime 时间就会发送一次心跳。通过心跳不仅能够用来监听机器的工作状态，还可以通过心跳来控制 Flower 跟 Leader 的通信时间，默认情况下 FL 的会话时常是心跳间隔的两倍。

2．initLimit

集群中的 follower 服务器(F)与 leader 服务器(L)之间初始连接时能容忍的最多心跳数（tickTime 的数量）。

3．syncLimit

集群中 flower 服务器（F）跟 leader（L）服务器之间的请求和答应最多能容忍的心跳数。

4．clientPort

客户端连接的接口，客户端连接 zookeeper 服务器的端口，zookeeper 会监听这个端口，接收客户端的请求访问，端口默认是 2181。

5．dataDir

该属性对应的目录是用来存放 myid 信息跟一些版本，日志，跟服务器唯一的 ID 信息等。

在集群 Zookeeper 服务在启动的时候，会回去读取 zoo.cfg 这个文件，从这个文件中找到这个属性然后获取它的值也就是 dataDir 的路径，它会从这个路径下面读取 myid 这个文件，从这个文件中获取要启动的当前服务器的地址。

任务一： Zookeeper 搭建

任务要求：

1．将 zookeeper 安装包解压到指定路径/usr/zookeeper（安装包存放于/us

r/package277/）。

2．配置系统变量 ZOOKEEPER_HOME，同时将 Zookeeper 安装路径中 bin 目录加入 PATH 系统变量，注意生效变量。

3．Zookeeper 的默认配置文件为 Zookeeper 安装路径下 conf/zoo_sample.cfg，将其修改为 zoo.cfg。

4．设置数据存储路径(dataDir)为/usr/zookeeper/zookeeper-3.4.14/zkdata。

5．设置日志文件路径(dataLogDir)为/usr/zookeeper/zookeeper-3.4.14/zkdatalog。

6．设置集群列表（要求 master 为 1 号服务器，slave1 为 2 号服务器，slave2 为 3 号服务器）。

7．创建所需数据存储文件夹、日志存储文件夹。

8．数据存储路径下创建 myid，写入对应的标识主机服务器序号。

9．启动服务，查看进程 QuorumPeerMain 是否存在。

10．查看各节点服务器角色是否正常(leader/follower)。

题目三： Hadoop 集群搭建

题目说明：

Hadoop 是由 Java 语言编写的，在分布式服务器集群上存储海量数据并运行分布式分析应用的开源框架，其核心部件是 HDFS 与 MapReduce。

HDFS 是一个分布式文件系统：引入存放文件元数据信息的服务器 Namenode 和实际存放数据的服务器 Datanode，对数据进行分布式储存和读取。

MapReduce 是一个计算框架：MapReduce 的核心思想是把计算任务分配给集群内的服务器里执行。通过对计算任务的拆分（Map 计算/Reduce 计算）再根据任务调度器（JobTracker）对任务进行分布式计算。

任务一： Hadoop 完全分布式集群搭建

任务要求：

1．将 Hadoop 安装包解压到指定路径/usr/hadoop（安装包存放于/usr/package277/）。

2．配置环境变量 HADOOP_HOME，将 Hadoop 安装路径中 bin 目录和 sbin 目录加入 PATH 系统变量，注意生效变量。

3．配置 Hadoop 运行环境 JAVA_HOME。

4．设置全局参数，指定 HDFS 上 NameNode 地址为 master,端口默认为 9000。

5．指定临时存储目录为本地/root/hadoopData/tmp(要求为绝对路径，下同)。

6．设置 HDFS 参数，指定备份文本数量为 2。

7．设置 HDFS 参数，指定 NN 存放元数据信息路径为本地/root/hadoopData/name；指定 DN 存放元数据信息路径为本地/root/hadoopData/data(要求为绝对路径)。

8．设置 HDFS 参数，关闭 hadoop 集群权限校验（安全配置），允许其他用户连接集群；指定 datanode 之间通过域名方式进行通信。

9．设置 YARN 运行环境$JAVA_HOME 参数。

10．设置 YARN 核心参数，指定 ResourceManager 进程所在主机为 master，端口为 18141;指定 mapreduce 获取数据的方式为 mapreduce_shuffle。

11．设置计算框架参数，指定 MR 运行在 yarn 上。

12．设置节点文件，要求 master 为主节点； slave1、slave2 为子节点。

13．对文件系统进行格式化。

14．启动 Hadoop 集群查看各节点服务。

15．查看集群运行状态是否正常。

题目四： Hive 集群搭建

题目说明：

1．比赛框架

本次比赛为分布式集群搭建，共三台节点，其中 master 作为主节点，slave1、slave2 为从节点；

2．比赛内容

基础配置：修改主机名、主机映射、时区修改、时间同步、定时任务、免密访问；

JDK 安装：环境变量；

Zookeeper 部署：环境变量、配置文件 zoo.cfg、myid；

Hadoop 部署：环境变量、配置文件修改、设置节点文件、格式化、开启集群；

Hive 部署：Mysql 数据库配置、服务器端配置、客户端配置安装数据库

任务一： 安装数据库

任务说明：

相关安装包已经存放至环境/usr/package277/中，对应 ntp 和 mysql 已安装，可直接对其进行操作和配置。

任务要求：

1．环境中已经安装 mysql-community-server，关闭 mysql 开机自启服务。

2．开启 MySQL 服务。

3．判断 mysqld.log 日志下是否生成初临时密码。

4．设置 mysql 数据库本地 root 用户密码为 123456。

任务二： Hive 基础环境配置

任务说明：

Hive 是基于 Hadoop 的一个数据仓库工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供 SQL 查询功能，能将 SQL 语句转变成 MapReduce 任务来执行。

1．讲指定版本的 Hive 安装包解压到指定路径，添加系统并生效；

2．修改 Hive 运行环境；

3．由于客户端需要和 Hadoop 通信，为避免 jline 版本冲突问题，将 Hive

中 lib/jline-2.12.jar 拷贝到 Hadoop 中，保留高版本。

任务要求：

1．将 Hive 安装包解压到指定路径/usr/hive（安装包存放于/usr/package 277/）。

2．配置环境变量 HIVE_HOME,将 Hive 安装路径中的 bin 目录加入 PAT H 系统变量，注意生效变量。

3．修改 HIVE 运行环境，配置 Hadoop 安装路径 HADOOP_HOME。

4．修改 HIVE 运行环境,配置 Hive 配置文件存放路径 HIVE_CONF_DIR。

5．修改 HIVE 运行环境，配置 Hive 运行资源库路径 HIVE_AUX_JARS_ PATH。

6．解决 jline 的版本冲突，将$HIVE_HOME/lib/jline-2.12.jar 同步至$H ADOOP_HOME/share/hadoop/yarn/lib/下。

任务三： 配置 HIVE 元数据至 MySQL

任务说明：

slave1 作为服务器端需要和 Mysql 通信，所以服务端需要将 Mysql 的依赖包放在 Hive 的 lib 目录下。mysql-connector-java 是 MySQL 的 JDBC 驱动包，用 JDBC 连接 MySQL 数据库时必须使用该 jar 包。

任务要求：

1．驱动 JDBC 拷贝至 hive 安装目录对应 lib 下（依赖包存放于/usr/packa ge277/）。

2．配置元数据数据存储位置为/user/hive_remote/warehouse。

3．配置数据库连接为 MySQL。

4．配置连接 JDBC 的 URL 地址主机名及默认端口号 3306，数据库为 hive，如不存在自行创建，ssl 连接方式为 false。

5．配置数据库连接用户。

6．配置数据库连接密码。

任务四： 配置 HIVE 客户端

任务说明：

1．master 作为客户端,可进入终端进行操作。

2．关闭本地模式。

3．将 hive.metastore.uris 指向 metastore 服务器 URL。

任务要求：

1．配置元数据存储位置为/user/hive_remote/warehouse。

2．关闭本地 metastore 模式。

3．配置指向 metastore 服务的主机为 slave1，端口为 9083。

任务五： 启动 Hive

任务说明：

1．服务器端初始化数据库，并启动 metastore 服务。

2．客户端开启 Hive client，即可根据创建相关数据操作。

任务要求：

1．服务器端初始化数据库，启动 metastore 服务。

2．客户端开启进入 hive，创建 hive 数据库。

题目五： Spark 搭建

题目说明：

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 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 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。

任务一： Spark 集群搭建

任务说明：

Spark 是 Hadoop 的子项目。在相关环境中将 Spark 安装到基于 Linux 的系统中。

任务要求：

1．将 Spark 安装包解压到指定路径。/usr/spark/spark-2.4.3-bin-hadoop2.7（安装包存放于/usr/package277/）。

2．文件/etc/profile 中配置环境变量 SPARK_HOME，将 Spark 安装路径中的 bin 目录加入 PATH 系统变量，注意生效变量。

3．修改配置文件 spark-env.sh，设置主机节点为 master。

4．修改配置文件 spark-env.sh，设置 java 安装路径。

5．修改配置文件 spark-env.sh，设置节点内存为 8g。

6．修改配置文件 spark-env.sh，设置 hadoop 安装目录、hadoop 集群的配置文件的目录。

7．修改 slaves 文件，添加 spark 从节点 slave1、slave2。

8．开启集群，查看各节点进程(主节点进程为 Master，子节点进程为 Worker)。

模块二： 大数据集群运维

题目一： 平台运维

题目说明：

随着科技的进步,如今社会已经进入到大数据时代,不同行业开始将各种各样的高新技术引进到企业中,以此来提升企业生产效率.而机房是信息系统的载

体,所以在大数据时代要想提升信息技术的效率,就需要加强机房管理和运维工作,以此来保证信息系统的安全稳定.

任务一： 集群动态添加 DataNode 节点

任务说明：

1．配置主节点 master 到新增节点 slave3 的免密登录（ssh 信任）。

2．安装 JDK 和 Hadoop，生效环境变量。

3．将 slave3 主机名加入到所有 Hadoop 配置文件中 slaves 文件中。

4．新节点启动 datanode 和 nodemanager 进程。

5．主节点刷新集群状态。

任务要求：

1．按照左侧虚拟机名称修改主机名并生效。

2．时间更改为为上海时间（CST+0800 时区）。

3．修改所有节点的 host 文件，添加 slave3 主机信息。

4．添加定时任务，每隔十分钟同步一次主节点时间。

5．建立master 使用root用户到 slave3 的 ssh 免密登录访问。

6．节点中安装 JDK 至/usr/java 下，并更新系统变量 JAVA_HOME，查看 JDK 版本。

7．节点中安装 hadoop 安装包至/usr/hadoop 下，并更新系统变量$HADOOP_HOME。

8．修改集群 slaves 文件，添加新节点主机名信息。

9．启动新节点中的 DataNode 和 NodeManager 进程。

10．刷新集群并查看节点状态。

任务二： 集群动态删除 DataNode 节点

任务说明：

1．添加 dfs.hosts.exclude 设置，配置下线节点信息。

2．关闭节点进程，进行数据迁移。

3．刷新集群，查看状态。

任务要求：

1．配置 HDFS 参数 dfs.hosts.exclude，指定拒绝加入集群的节点列表文件为参数文件同目录下的 excludes。

2．创建 excludes 文件，并写入需要删除的节点为 slave3。

3．集群强制重新加载配置，查看集群状态，查看节点退役状态。

4．关闭 slave3 上 Hadoop 相关进程，主机点重新查看集群节点状及死节点（此过程中，数据迁移较久，同时注意数据均衡）。

模块三： 数据采集与处理

题目一： 论坛数据采集

题目说明：

根据给出的网站地址（discuz 论坛）进行相关爬取操作，爬取形式不限，爬取相关的论坛的所有发帖数据。将爬去的数据存放在指定目录。

使用 hive 对数据进行统计，将结果写入指定文件。

任务一： 爬取指定网站数据

任务要求：

1．编写代码爬取给定网站的帖子 ID，用户名，积分，等级，标题，内容，并将数据写入/root/discuz/data.txt。

2．创建表并将数据导入 data 表中，并统计所有数据数目至/root/discuz01/目录下。

3．统计总用户数，并将最后数目写入/root/discuz02/目录下。

4．统计活跃用户 top10，将结果用户名及对应发帖数目写入。/root/discuz03/目录下，要求如下： 复合排列：先按照第二列发帖数倒叙排列，再按照第一

列用户名升序排列。

题目二： 商场数据采集

题目说明：

根据给出的网站地址（shopxo 商城）进行相关爬取操作，爬取形式不限，根据要求将爬去的数据存放在指定目录。

使用 hive 对数据进行统计，将结果写入指定文件。

任务一： 商城数据获取及分析

任务要求：

1．编写代码爬取给定网站的商品 ID、名称、价格、浏览量、销量、库存，并将数据写入/root/college020/goods.txt。

2．hive 中创建 shopxo(库).goods 表，要求字段包括 id,title,price,views,sales,stock。

3．查找缺失值，将表中价格为空（null）的数据，写入至/root/college023/

4．缺失值处理，title 中去除“连衣裙”、“女士”及空值 null 数据，创建中间表 goods1，存放过滤后的数据。

5．对中间表数据所有行进行统计，结果写入/root/college024/。

6．查询中间表 goods1，按照价格降序查找前三条商品信息(去重)，格式为 tile　price。结果写入/root/college025/。

7．第一个元素 title[0]作为对应商品品牌,对各品牌进行计数统计,将 TOP10 写入/root/college026/。

8．对上题排名第一的品牌进行分析，根据其商品特征前 6 名进行特征统计，结果写入/root/college027/。

模块四： 数据分析

题目一： Hadoop 数据分析

任务一： 简单数据统计 WordCount

任务说明：

单词计数是最简单也是最能体现 MapReduce 思想的程序之一，可以称为 MapReduce 版"Hello World"，该程序的完整代码可以在 Hadoop 安装包的"src/examples"目录下找到。单词计数主要完成功能是：统计一系列文本文件中每个单词出现的次数。

任务要求：

1．将数据 sonnet.txt 上传至 HDFS 文件系统/input/下，注意自行创建目录。

2．编写程序或使用 Hadoop 自带的开源 jar 包对数据进行 wordcount 操作，结果保存至 hdfs:/output/part--r-00000。

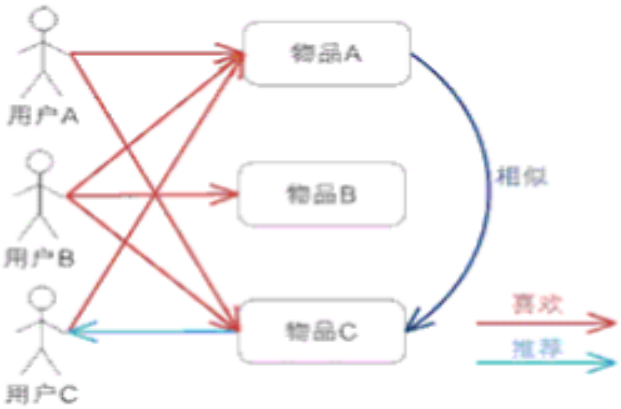
任务二： 用户电影推荐

任务说明：

随着互联网的急速发展，网络中的信息量以指数规律迅速扩展和增加，网络上的信息过载和信息迷航问题日益严重，使得人们被动接受不喜欢的事物，为解决这一问题，推荐技术应运而生，推荐系统通过预测用户对信息资源的喜好程度来进 行信息过滤，根据用户具体需求通过协同过滤等技术进行个性化推荐。

基于物品的协同过滤(ItemCF)：通过用户对于不同 item 的评分来预测 item 之间的相似性，基于 item 之间的相似性最初推荐；简单来说，就是给用户推荐和他之前喜欢的物品相似的物品。

用户/物品	物品A	物品B	物品C
用户A	√		√
用户B	√	√	√
用户C	√		推荐



要求通过协同过滤算法 itemCF 针对用户评价电影的信息，对用户信息进行训练，计算出为每位用户推荐电影的分数，进而进行电影推荐。

（1）计算物品之间的相似度。

（2）根据物品的相似度和用户的历史行为给用户生成推荐列表。

任务要求：

1．将本地数据/root/movie/data.csv 上传至 HDFS 文件系统/input/下，注意自行创建目录。

2．编写程序实现评分矩阵，计算所有物品出现的组合列表，结果保存至本地/root/movie/output1/目录下 part-r-00000 文件中。

3．编写程序实现同现矩阵，对电影 ID 循环匹配并进行计数，结果保存至本地/root/movie/output2/目录下 part-r-00000 文件中。

4．编写程序实现对评分矩阵的转换，结果保存至本地/root/movie/output3_1/目录下 part-r-00000 文件中。

5．编写程序实现对同现矩阵的转换(即读入 output2 结果)，结果保存至本地/root/movie/output3_2/目录下 part-r-00000 文件中。

6．编写程序实现矩阵相乘，得到推荐结果，结果保存至本地/root/movie/output4/目录下 part-r-00000 文件中。

任务三： 互联网日志分析

任务说明：

随着计算机互联网技术的不断发展，社会经济水平的不断提高，各种通讯设备和移动工具的普及，网络成为人们生活的必需品，以某大型网站为例，每小时就产生 10G 的数据量，网络数据在飞速增长，在大数据时代背景下，需要对数据隐藏价值进行充分挖掘，加强对数据的分析。

对这些数据进行有效利用需要通过大数据相关方面的技术和工具，收集用户的习惯、风格等方面的数据，对用户进行有效的分析，对每个用户制定不同的平台营销与个性化服务。

任务要求：

1．将本地数据/root/internetlogs/journal.log 上传至 HDFS 文件系统/input/下，注意自行创建目录。

2．编写程序进行页面访问量统计，结果保存至本地/root/internetlogs/pv/目录下 part-00000 文件中。

3．编写程序进行页面独立 IP 的访问量统计，结果保存至本地/root/internetlogs/ip/目录下 part-00000 文件中，例如 1.80.249.223　1 表示此 IP 访问量为 1。

4．编写程序进行每小时访问网站的次数统计，结果保存至本地/root/internetlogs/time/目录下 part-00000 文件中。

5．编写程序进行访问网站的浏览器标识统计，结果保存至本地/root/internetlogs/browser/目录下 part-00000 文件中，具体查看步骤说明。

题目二： Hive 数据分析

任务一： 共享单车数据分析

任务说明：

现有数据为某年某段时间某地区的共享单车数据集，可以适用于大数据分析和挖掘。通过对共享单车的骑行规律，用户群体，单日活月用户等数据的分析，给出运营思路和方法上的建议，对共享单车的发展有一个整体的把握。基于对数据的分析，可以进行活动推广、会员特定优惠，也可进行专线共享大巴等活动策划。

任务要求：

1．数据/root/college/bike.csv 上传至 hdfs://college/目录下。

2．统计本次数据所有单车数量（以单车车号进行计算，注意去重），结果写入本地/root/bike01/000000_0 文件中。

3．计算单车平均用时，结果写入本地/root/bike02/000000_0 文件中，以分钟为单位，对数据结果取整数值（四舍五入）。

4．统计常年用车紧张的地区站点 top10，结果写入本地/root/bike03/000000_0 文件中。(以 stratstation 为准)。

5．给出共享单车单日租赁排行榜，结果写入本地/root/bike04/000000_0 文件中。（以 startdate 为准,结果格式为 2021-09-14）。

6．给出建议维修的单车编号（使用次数），结果写入本地/root/bike05/000000_0 文件中。

7．给出可进行会员活动推广的地区，结果写入本地/root/bike06/000000_0 文件中。（以 stratstation 为准）。

8．给出可舍弃的单车站点，结果写入本地/root/bike07/000000_0 文件中。（以 endstation 为准）。

任务二： 贷款数据分析

任务说明：

首先创建对应数据库，根据数据类型创建表，最后将数据进行导入。本此数据不进行数据清洗。

任务要求：

1．将数据 loan.csv 上传到 hdfs 的/input/目录下。

2．创建数据库 hive。

3．在 hive 数据库下构建数据表 loan。

4．将提供的分析数据导入到表 loan 中，并统计数据至本地/root/college000/000000_0 文件中。

5．以信用得分 ProsperScore 为变量，对借款进行计数统计（降序），结果写入本地/root/college001/000000_0 文件中。

6．给出借款较多的行业 top5，结果写入本地/root/college002/000000_0 文件中。

7．分析贷款状态为违约（Defaulted）的贷款人就业信息，将结果 top3 写入/root/college003/000000_0 文件。

8．对数据中收入范围进行分组统计（降序），查看贷款人收入情况，结果写入/root/college004/000000_0 文件。

9．对信用得分上限及下限进行中间数求值作为职业信用分，对职业进行分组，计算职业信用分 top5（具体见步骤说明）。结果写入/root/college005/000000_0 文件。

10．支持度写到本地/root/college006/000000_0 文件中(保留五位小数)。

11．置信度写到本地/root/college007/000000_0 文件中(保留五位小数)。

题目三： 算法预测

题目说明：

1．使用随机森林算法完成基本建模任务

基本任务需要我们处理数据，观察特征，完成建模并进行可视化展示分析。

2．观察数据量与特征个数对结果影响

在保证算法一致的前提下，加大数据个数，观察结果变换。重新考虑特征工程，引入新特征后观察结果走势。

3．对随机森林算法进行调参，找到最合适的参数

掌握机器学习中两种经典调参方法，对当前模型进行调节。

任务一： 天气最高温度预测

任务说明：

1．使用随机森林算法完成基本建模任务

基本任务需要我们处理数据，观察特征，完成建模并进行可视化展示分析。

2．观察数据量与特征个数对结果影响

在保证算法一致的前提下，加大数据个数，观察结果变换。重新考虑特征工程，引入新特征后观察结果走势。

在数据分析和特征提取的过程中，我们的出发点都是尽可能多的选择有价值的特征，因为其实阶段我们能得到的信息越多，之后建模可以利用的信息也是越多的。

任务要求：

1．加载数据，并查看前 5 行数据。

2．补充代码，展示数据 features 的描述性统计信息。

3．设置布局为 2*2 画布，尺寸为 10*10。

4．补充参数，进行 one-hot 编码。

5．去掉标签。

6．设置 y 轴为 Importance（绘制柱状图）。

7．验证绘图结果。

8．研究“数据与特征对随机森林的影响”，设置 x 轴坐标，倾斜 45 度。

9．绘图展示 Max Temp、Prior Max Temp、Two Days Prior Max Temp、Friend Estimate（注意文件中说明）。

10．绘图展示 Historical Avg Max Temp、Prior Wind Speed、Prior Precipitation、Prior Snow Depth。

11．绘制 pairplot，设置 diag_kind 参数。

12．数据集切分，测试集比例 25%，随机种子为 42。

13．运行程序，查看平均温度误差。

任务二： 某运营商用户消费行为分析

任务说明：

当前，中国经济已由高速增长转向高质量发展阶段。党的十九届四中全会首次将数据增列为一种生产要素，2020 年 4 月 9 日，中央第一份关于要素市场化配置的文件《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》正式发布。《意见》指出了土地、劳动力、资本、技术、数据五个要素领域改革的方向，明确了完善要素市场化配置的具体措施。数据作为一种新型生产要素，成为了《意见》中备受关注的内容。

当前人类社会已经步入数据驱动的数字经济时代，数据要素空前提升了全要素生产率，成为与劳动力、资本、土地和技术并驾的生产要素，更成为数字经济时代的关键要素，是第三次工业革命的关键成果，也是第四次工业革命的重要基础。围绕海量的数据要素，以分布式计算、边缘计算、量子计算等展开的大数据分析挖掘提供了新的处理手段，数据资产管理提供了新的管理手段，共同推动数据要素向各行业全面渗透。

近年来，社会转型加速，国家不断加强培育数据要素市场，推动治理体系现代化、推进新型技术设施建设，致力于打造全新智慧城市。

在新型智慧城市建设中，5G 成为最重要的技术背景，其低时延、大带宽、广连接的技术优势，可以提升城市感知的灵活度、弹性和精细程度。在 5G 技术的加持下，我国新型智慧城市建设快速升级，深刻改变着城市面貌。

加快 5G 用户增长与城市发展深度融合，通过信息化手段解决城镇化过程中

带来的问题，即使城市可持续发展所需，也是产业新动能所在，通过模型精准识别 5G 需求潜在用户，促进 4G 向 5G 的转变，以实现基于 5G 深度应用的智慧城市建设至关重要。

本项目基于每月用户更换 5G 套餐数据，分析其行为特征、洞察消费趋势，并结合当前 5G 资费产品库对客户进行识别,以满足客户实际需求,提供更优服务体验,助力 5G 套餐迁转发展战略实施。

本数据数据包括：基础信息、消费行为、超套信息、宽带信息、其他信息，共 46 个用户特征。

任务要求：

1．补充代码，加载训练特征数据和标签数据，注意使用数据目录。

2．补充代码，合并特征数据与标签数据作为训练数据 train，以‘user_id’为连接键。

3．将合并后的的数据，替换中文列名并保存文件为同目录下 train_data.csv，编码格式为 UTF-8。

4．补充代码，查看训练集各列中空值个数并打印。

5．要求对数值型变量缺失值进行中位数填充，使用数据填充函数 fillna()直接修改原对象。

6．补充代码，计算宽带带宽对应的 5G 用户占比。

7．读取预测数据样本，进行数据合并并重置索引。

8．划分数据集，要求对数据进行随机排列交叉验证，分割 10 组。

9．保存随机森林模型 RandomForestClassifier.pkl 至当前目录下。

10．保存逻辑回归模型 LogisticRegressionCV.pkl 至当前目录下。

11．保存贝叶斯模型 GaussianNB.pkl 至当前目录下。

12．保存 KNN 模型 KNeighborsClassifier.pkl 至当前目录下。

13．补充代码，调用随机森林模型预测样本概率。

14．补充代码，使用模型预测测试集中 5G 套餐转换意向信息。

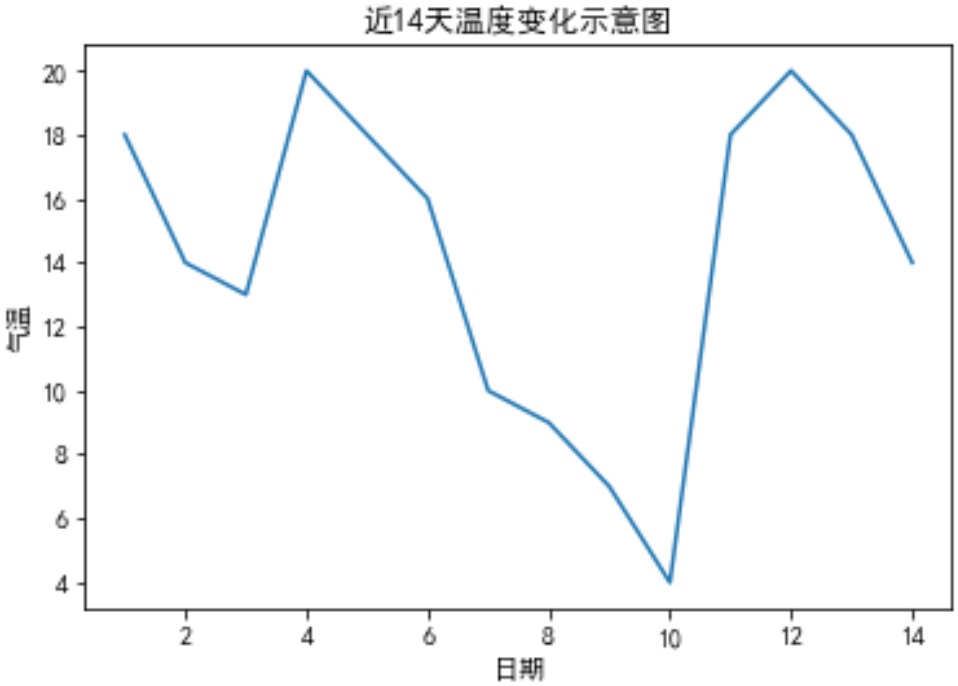
模块五： 数据可视化

题目一： 数据可视化

任务一： 天气温度趋势图

任务说明：

导入气温表的数据，实现近 14 天气温趋势示意图。



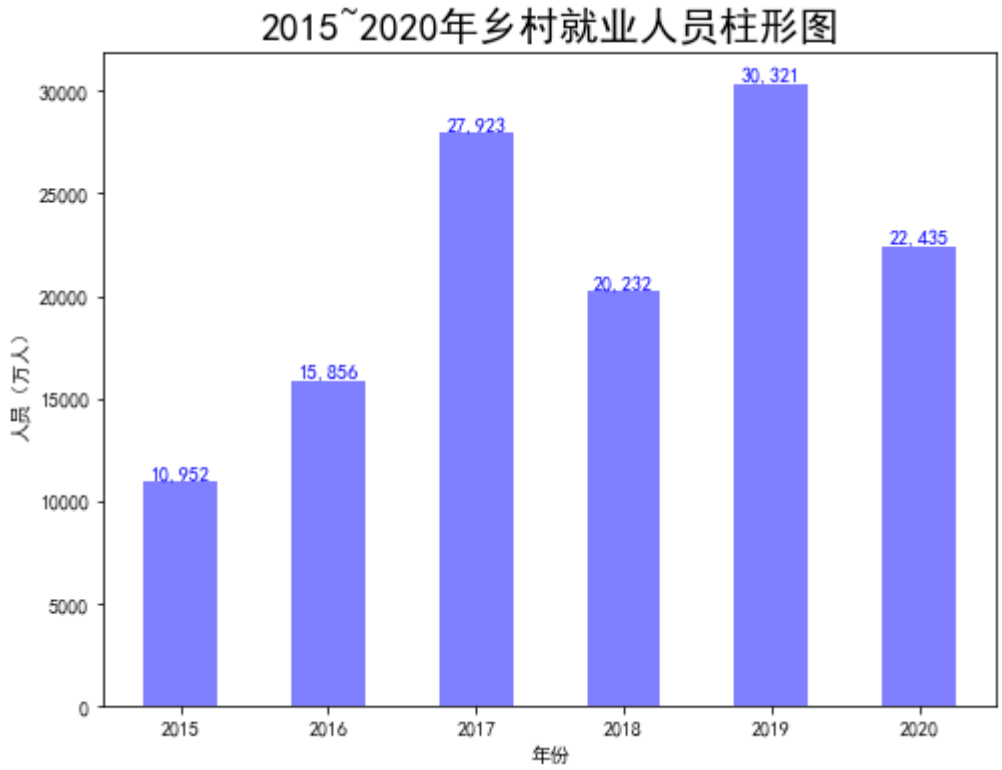
任务要求：

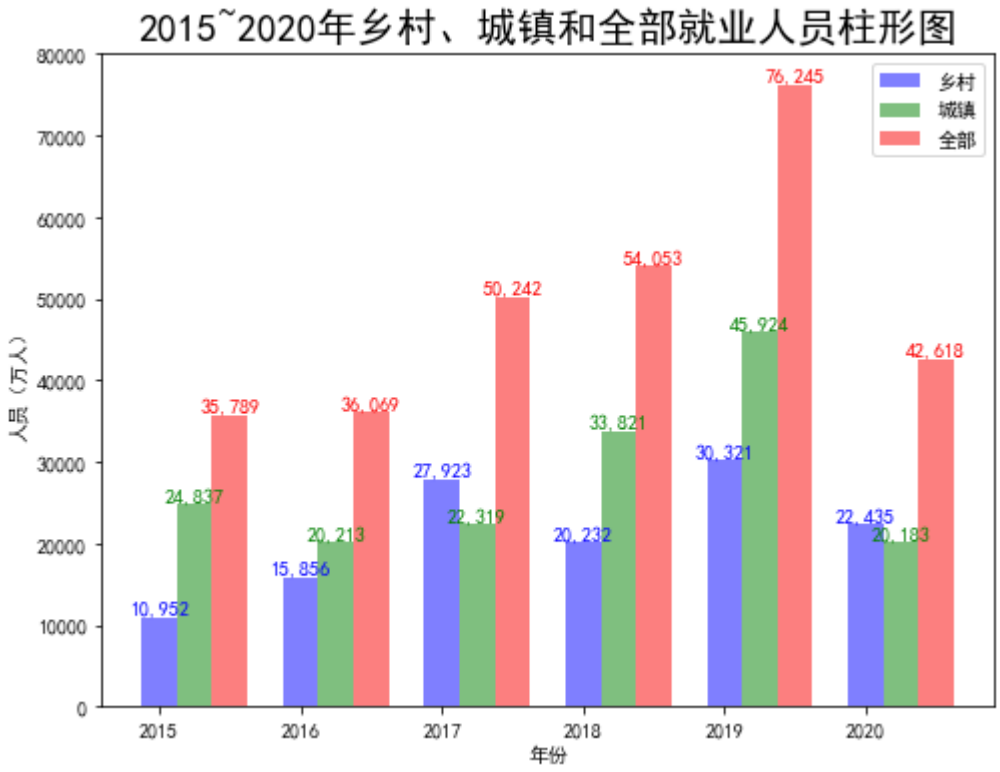
1. 读取目录下的气温数据文件。
2. 设置气温折线图横坐标轴名称。
3. 设置气温折线图纵坐标轴名称。
4. 设置气温折线图标题。

任务二： 就业形势分析图

任务说明：

根据要求绘制就业人员柱状图





任务要求：

1．读取就业形势数据并查看数据。

2．使用函数绘制柱状图并写入对应 x,y 坐标数据。

3．绘制多个并列柱状图，将乡村、城镇、全部就业人员数据赋值给 y 轴。

4．添加图例进行就业人员柱状图展示。