ZZ052–大数据应用与服务赛项试题 06

一、背景描述

当今时代，数据正在迅速膨胀并变大，一天之中，互联网产生的全部内容可以达到 EB 级别，能够轻松刻满 1.68 亿张光盘。在商业、经济及其它领域中，决策将日益基于数据和分析而作出，而并非基于经验和直觉。那么，要怎样基于大数据做出正确的决策呢？大数据首先需要解决的问题就是数据存储的问题，由于数据量非常之大，想通过传统单一的节点的存储显得力不从心，搭建分布式的文件存储系统成为了一个完美的解决方案。解决了数据存储的问题，我们需要从数据中提取有用信息，通过数据分析手段让数据发挥出真正的价值。但往往采集的原始数据中包含了一些无用数据以及噪声数据，如果直接基于这些脏数据进行分析，往往会让分析结果产生偏差甚至错误，从而造成决策上的失准。因此，我们有必要对这些原始数据进行清洗，以保证其数据准确性、完整性和可用性，提高数据的质量。在解决脏数据的困扰后，我们需要采取各种数据分析手段，提取数据中的价值，得到可靠的结果，并以图表等直观的方式将分析结果进行展现。然后从业务层面对分析结果进行分析和解释，从而指引我们做出正确的决策，真正获取“数据财富”。

气候变化正在迅速地改变地球。随着全球气温不断升高、

1

海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，指引相关部门尽早做出举措以应对气候变化，保护人类赖以生存的家园，你的团队将运用大数据技术对天气数据进行分析及决策。搭建大数据平台集群环境以应对海量天气数据的存储，结合数据库的毫秒级的响应，为天气决策系统提供数据存储及查询保障。通过数据清洗技术，去除数据中的噪音，提高数据质量。通过数据标注技术，结合业务认知，对数据进行分类标注，为后续通过人工智能算法模型决策奠定基础。通过各种数据分析技术，让看似杂乱无章的数据，变得灵动，找出天气变化的内在规律。通过数据可视化技术，让数据分析结果及天气变化规律以一种最为直观的方式呈现。最后从业务层面对天气数据分析结果进行分析及解释，使气象学家更好的了解气候变化，并做出精准决策应对气候问题。你们作为该大数据小组的技术人员，请按照下面任务完成本次工作。

二、模块一：平台搭建与运维

（一）任务一：大数据平台搭建

1.子任务一：Zookeeper 集群安装配置

本任务需要使用 root 用户完成相关配置，具体要求如下:

（1）在 master 节点将/usr/local/src 目录下的 apache-zookeeper-3.5.7-bin.tar.gz 包解压到/opt 路径下，

将完整命令截图粘贴到对应答题报告中;

（2）把解压后的 apache-zookeeper-3.5.7-bin 文件夹更名为 zookeeper-3.5.7，将完整命令及结果截图粘贴到对应答题报告中;

（3）在 master 节点修改/root/.bash_profile 文件，设置 Zookeeper 环境变量，将环境变量配置内容截图粘贴到对应答题报告中;

（4）将/opt/zookeeper-3.5.7/conf 目录下的 zoo_sample.cfg 文件更名为 zoo.cfg ，将完整命令截图粘贴到对应答题报告中;

（5）修改/opt/zookeeper-3.5.7/conf/zoo.cfg 配置文件，配置 zookeeper 服务器存储快照文件（zookeeper 节点数据）的目录为 /opt/zookeeper-3.5.7/data 目录，将修改的内容截图粘贴到对应答题报告中;

（6）修改/opt/zookeeper-3.5.7/conf/zoo.cfg 配置文件，配置 master 节点为 zookeeper 集群的第一号服务器、slave1 节点为 zookeeper 集群的第二号服务器、 slave2 节点为 zookeeper 集群的第三号服务器，并且将 master、slave1、slave2 节点与集群中的 Leader 节点交换信息的端口号设置为 2888、选举 Leader 端口号设置为 3888，将修改的内容截图粘贴到对应答题报告中;

（7）在 /opt/zookeeper-3.5.7/data 目录下创建名字为 myid 的文件（如果文件夹不存在则需自行创建），根据

3

/opt/zookeeper-3.5.7/conf/zoo.cfg 配置文件中的信息在 myid 文件中添加合适的编号，将 myid 文件中添加的内容截图粘贴到对应答题报告中;

（8）在 master 节点上面将配置的 zookeeper 环境变量文件及 zookeeper 解压包拷贝到 slave1、slave2 节点的 /opt 路径下，将命令和结果截图粘贴到对应答题报告中;

（9）在 slave1 节点中根据 /opt/zookeeper-3.5.7/conf/zoo.cfg 配置文件中的信息修改/opt/zookeeper-3.5.7/data/myid 配置文件中的编号，将 myid 文件中修改的内容截图粘贴到对应答题报告中;

（10）在 slave2 节点中根据 /opt/zookeeper-3.5.7/conf/zoo.cfg 配置文件中的信息修改/opt/zookeeper-3.5.7/data/myid 配置文件中的编号，将 myid 文件中修改的内容截图粘贴到对应答题报告中;

（11）启动 Zookeeper 集群，使用 jps 查看 master 节点、slave1 节点、slave2 节点的进程，将查看结果截图粘贴到对应答题报告中;

2.子任务二: Kafka 完全分布式集群搭建

本任务需要使用 root 用户完成相关配置，已安装 Zookeeper 及需要配置的前置环境，具体要求如下:

（1）在 master 节点将/usr/local/src 目录下的 kafka_2.12-2.4.1.tgz 包解压到/opt 路径下，将完整命令

截图粘贴到对应答题报告中;

（2）把解压后的 kafka_2.12-2.4.1 文件夹更名为 kafka-2.4.1，将完整命令及结果截图粘贴到对应答题报告中;

（3）在 master 节点修改/root/.bash_profile 文件，设置 Kafka 环境变量，将环境变量配置内容截图粘贴到对应答题报告中;

（4）在 master 节点上面修改 Kafka 的配置文件 server.properties，需要在该文件中指定 broker 的全局唯一编号为 1，启用 topic 的删除功能，指定连接 zookeeper 集群地址（master、slave1、slave2 全部需要指定），将修改的内容截图粘贴到对应答题报告中;

（5）在 master 节点上面将配置的 Kafka 环境变量文件及 Kafka 解压包拷贝到 slave1、slave2 节点，将命令和结果截图粘贴到对应答题报告中;

（6）启动 Kafka 集群，使用 jps 查看 master 节点、slave1 节点、slave2 节点的进程，将查看结果截图粘贴到对应答题报告中。

3.子任务三：Spark standalone 安装配置

本任务需要使用 root 用户完成相关配置，具体要求如下:

（1）在 master 节点将/usr/local/src 目录下的

5

spark-3.1.1-bin-hadoop3.2.tgz 安装包解压到/opt 路径，将完整命令截图粘贴到对应答题报告中;

（2）把解压后的 spark-3.1.1-bin-hadoop3.2 文件夹更名为 spark-3.1.1，将完整命令及结果截图粘贴到对应答题报告中;

（3）在 master 节点修改/root/.bash_profile 文件，设置 Spark 环境变量，将环境变量配置内容截图粘贴到对应答题报告中;

（4）在 master 节点上将/opt/spark-3.1.1/conf 目录下的 spark-env.sh.template 文件更名为 spark-env.sh ，将完整命令截图粘贴到对应答题报告中;

（5）在更名后的 spark-env.sh 文件中进行修改，需要指定 standalone 模式运行时的 Master 进程运行在 master 节点上，指定 Master 进程内部通信端口号为 7077，将修改的内容截图粘贴到对应答题报告中;

（6）在 master 节点上面修改/opt/spark-3.1.1/conf/workers.template 文件名为 workers 后在/opt/spark-3.1.1/conf/workers 文件中指定 standalone 模式运行时 Worker 进程需要分别在 master、slave1、slave2 节点上运行，将修改的内容截图粘贴到对应答题报告中;

（7）在 master 节点上面将配置的 Spark 环境变量文件及 Spark 解压包拷贝到 slave1、slave2 节点的 /opt 路径

6

下，将命令和结果截图粘贴到对应答题报告中;

(8)启动Spark集群,使用jps查看master节点、slave1节点、slave2节点的进程，将查看结果截图粘贴到对应答题报告中。

（二）任务二: 数据库配置维护

1.子任务一: 数据库配置

MySQL 是一个多用户数据库，具有功能强大的访问控制系统，可以为不同用户指定不同权限。root 用户是超级管理员，拥有所有权限，包括创建用户、删除用户和修改用户密码等管理权限。

为了实际项目的需要，可以定义不同的用户角色，并为不同的角色赋予不同的操作权限。当用户访问数据库时，需要先验证该用户是否为合法用户，再约束该用户只能在被赋予的权限范围内操作。具体任务要求如下:

（1）为本地主机数据库创建一个名为 user01 的用户，密码为 123321，将完整命令及结果截图粘贴到对应答题报告中;

（2）将用户名 user01 修改为 user1001，将完整命令及结果截图粘贴到对应答题报告中;

（3）使用新用户 user1001 登录 MySQL 数据库，将完整命令及结果截图粘贴到对应答题报告中;

（4）授予用户 user1001 对 WeatherDB 数据库中所有

7

表的所有权限，将完整命令及结果截图粘贴到对应答题报告中（MySQL 数据库中已创建好 WeatherDB 数据库，如果不存在则需要自己建库并导入数据，提供的几个 sql 是数据源文件。);

2.子任务二：数据表与数据管理

本环节需要使用 MySQL 数据库系统完成关于天气信息的建库、建表、数据的插入、数据表的管理等操作。具体要求如下:

（1）在 MySQL 数据库中创建名字为 WeatherTestDB 的数据库，将完整命令及运行结果截图粘贴到对应答题报告中;

（2）在 MySQL 数据库中删除名字为 WeatherTestDB 数据库，将完整命令及运行结果截图粘贴到对应答题报告中;

（3）在 MySQL 数据库的 WeatherDB 库中，创建一个名为 weather_base_info_table 的数据表，包含的字段见下面的数据表说明，要求所有字段非空，数据库引擎为 InnoDB，默认字符集为 utf8。将完整命令及运行结果截图粘贴到对应答题报告中;

表 1　数据表字段

列名	数据类型	备注
city_id	int	城市 ID
city_name	varchar	城市名称
month	varchar	日期，形式为年-月，比如:2022-04
avg_high_temp	int	平均最高温度

8

avg_low_temp	int	平均最低温度
extreme_high_temp	int	极端最高温度
extreme_low_temp	int	极端最低温度

（4）使用 SQL 命令修改 weather_base_info_table 表中 city_id 字段的类型和长度为 varchar(255)，将完整命令及结果截图粘贴到对应答题报告中;

（5）使用 SQL 命令删除 weather_base_info_table 表中的 city_id 字段，将完整命令及结果截图粘贴到对应答题报告中;

（6)使用 SQL 命令删除 weather_base_info_table 表，将完整命令及结果截图粘贴到对应答题报告中;

3.子任务三：维护数据表

本环节需要使用 MySQL 数据库系统内置的数据分析语法按照不同维度对天气相关数据进行数据分析等操作，具体要求如下:

（1）在 weather_day 表中使用 SQL 语句查询鞍山城市 2022 年晴天有多少天(天气为晴则为晴天),将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

（2）在 weather_month 表中使用 SQL 语句查询重庆每年极端最高温度超过 40℃的天气分别有多少天，将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

（3）在 weather_month 表中使用 SQL 语句统计 2022

9

年极端最高温度与极端最低温度超过 40℃ 的城市分别有哪些，将完整 SQL 语句和运行结果的后 5 条数据以及总数据行数截图粘贴到对应答题报告中;

（4）统计黑龙江省每年下雪天有多少天（每天的天气中包含雪即为下雪天），将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

（5）统计天气晴朗天数最多的前三个省份信息（天气包含晴则为天气晴朗），将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

表 2 （5）中查询字段相关说明

字段名字	字段含义
province_name	省份名称
cloudless	天气晴朗天数

（6）在 weather_month 表中使用 SQL 语句统计 2022 年 4 月 01 日这天中温差超过 15℃的城市有哪些,将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

表 3 （6）中查询字段相关说明

字段名字	字段含义
city_id	城市 ID
city_name	城市名称

（7）统计每个省份中属于亚热带季风气候的城市数量，查询结果按照城市数量从高到低进行排序，将完整 SQL 语句及运行结果截图粘贴到对应答题报告中;

表 4 （7）中查询字段相关说明

字段名字	字段含义
province_name	省份名称
climate_num	省内亚热带季风气候对应的城市数

三、模块二：数据获取与处理

（一）任务一：数据获取与清洗

1.子任务一：数据获取

根据 distribution.csv 文件统计单条数据缺失字段计数的最大值，将结果输出到控制台，输出格式如下:

===单条数据缺失字段计数的最大值为******

将控制台输出截图并粘贴到结果文件中。

2.子任务二：HDFS 文件上传下载

本任务需要使用 Hadoop，HDFS 命令，已安装 Hadoop 及需要配置前置环境，具体要求如下:

（1）在 master 节点的 hadoop 环境中，使用 HDFS 命令列出 HDFS 的文件和目录，将完整命令及结果截图粘贴到对应答题报告中;

（2）使用 HDFS 命令创建一个名为 bigdata 目录，将完整命令及结果截图粘贴到对应答题报告中;

（3）使用 HDFS 命令将/opt/eurasia_mainland.csv 文件上传到 HDFS 文件系统的/bigdata 目录下，将完整命令及结果截图粘贴到对应答题报告中;

(4)使用 HDFS 命令将/bigdata/eurasia_mainland.csv

11

文件下载到/root 目录下，将完整命令及结果截图粘贴到对应答题报告中;

(5)使用 HDFS 令查看/bigdata/eurasia_mainland.csv 文件的数据内容，将完整命令及结果截图粘贴到对应答题报告中。

（二）任务二: 数据标注

本任务是使用 Python 对给定的天气数据进行标注，并进行持久化存储。使用 Python 完成此任务，天气数据集具有多个字段，各字段信息如下表所示。

表 5　天气数据集字段说明表

数据字段	字段说明
city	城市名
hightest_tem	最高温
lowest_tem	最低温
weather	天气
date	年月日
wind_direction	风向
wind_level	风力等级

具体要求如下:

（1）使用 Python 读取“长春天气信息.xlsx”。

（2）在末尾新增一列数据为“当日是否解冻”，若当日最高温大于 0，并且风力小于等于 2 级，打标签为‘是’；否则打标签为‘否’。标记完成后保存到当前目录，文件命名为“annotation.xlsx”,并将数据截图粘贴到答题报告对应

12

位置。

（三）任务三：数据统计

1.子任务一：处理异常值数据

HDFS 文件系统中/bigdata/eurasia_mainland.csv 文件存储了欧亚大陆各个国家的灾害数据，数据中有以下内容：

表 6　灾害数据集字段说明表

c_year	年份
c_country	国家
hazard_type	灾害类型
disaster_subtype	灾害子类型
area	区域
disaster_frequency	灾害频次
c_death_toll	总死亡人数
c_people_affected	总受灾人数
c_economic_loss	总经济损失

编写 MapReduce 程序，实现以下功能：清除年份、国家区域为空的数据，将清理后的数据保存到 HDFS 中/clean_data 目录下，若目录不存在，请自行创建，使用命令查看该文件的大小，将完整命令及结果截图粘贴到对应答题报告中。

2.子任务二：数据统计

HDFS 文件系统中/bigdata/eurasia_mainland.csv 文件存储了欧亚大陆各个国家的灾害数据，数据中有以下内容：

编写 MapReduce 程序，实现以下功能：统计每个国家不同年份基于灾害类型为气候灾害受损经济最高的国家，并在控制台输出打印出气候灾害受损经济最高的 10 个国家，将输出结果截图粘贴到对应答题报告中。

四、模块三：业务分析与可视化

（一）任务一：数据可视化

1.子任务一：基于 Echarts 的数据可视化分析

气候变化正在迅速地改变地球。随着全球气温不断升高、海平面上升、极端天气事件频繁发生，人们对于地球的未来更加担忧。为了更好地了解气候变化的趋势、预测未来天气趋势，我们创建了“天气数据库”，用于收集、记录、可视化展示来自全球各地的气象数据和天气预报信息。请在“index.html”文件中编写代码实现功能，数据文件名为“chengdu.js”：

（1）文件内记录了四川省成都市 2021 年的天气数据，统计每个月的平均最高气温和每个月的平均最低气温，将统计得到的数据格式转换为 Echarts 所需的数据格式；

（2）将第 1 步中构建的数据作为输入，通过 Echarts 绘制柱状图；

（3）使用浏览器打开“index.html”文件，然后将渲染结果截图粘贴到答题报告对应位置。

14

2.子任务二：基于 Excel 进行数据分析与可视化

本环节需要使用办公软件 Excel 工具，对近几年天气数据进行分析与可视化。

近几年天气的数据在“E_weather.csv”中，数据表中记录 2011 年至 2022 年各城市的天气信息,包含城市、月份、平均高温、平均低温、平均空气质量指数、最佳空气质量指数、最佳空气质量日期、最差空气质量指数、最差空气质量日期 9 列，其中温度相关数据的单位：摄氏度（℃）。使用 Excel 打开“E_weather.csv”文件，对数据进行分析与可视化，具体要求如下：

（1）将 csv 数据表读取为 Excel 数据表，并分析每个数据字段类型，使字段能进行统计、计算等（参与计算的单元格中的值，如果存在字符，需要把字符替换为空，例如：单元格的值为:1a23a<--,替换之后的单元格的值为 123）。

（2）数据中每个城市每月的数据应该只有一份，数据中包含重复数据，请过滤掉重复日期的数据，并对数据根据日期升序进行排序。

（3)对数据进行统计分析,绘制出北京市 2018 年到 2021 年 12 个月份平均空气质量最差和最佳【带数据标记的折线图】。设置要求如下：

①设置图表标题为【空气质量波动】，标题加粗、居中显示。

②横坐标标签显示为月份，合理设置标签位置，使其

15

显示在轴的下方。

③纵坐标显示每月平均最佳和最差的空气质量指数，标题显示为“空气质量指数”。

④图例显示“最佳空气”和“最差空气”，并置于图像底部。

⑤显示涨/跌柱线，并将其修改为浅绿色，如下图所示：

图 1 结果示意图

（4）将绘制完成后的图表进行截图，粘贴到答题报告上对应位置。

3. 子任务三：基于 Python 实现历史最高温城市排名分析

现有一份关于 2011-2022 年全国各城市的每日天气数据

16

集，字段说明如下表:

表 7 每日天气数据集说明表

列名	字段说明
city	城市
hightest_tem	最高气温
lowest_tem	最低气温
weather	天气
date	日期
wind_direction	风向
wind_level	风力等级
month_day	月-日
weekday	星期

请编写代码实现功能，数据集为“clean_day.csv”。

绘图过程中如有使用中文字体的地方，请统一使用“SimSun”字体。

从数据集中取出 2011-2021 年的数据，计算出每个城市在这 10 年中出现的最高温度，然后将该指标排名前 10 的城市数据取出并绘制出这些城市最高温统计图进行分析，具体绘图要求如下:

（1）使用 Seaborn 绘制出以上指标的柱状图，主题设置为“whitegrid”，字体为“SimSun”，字体缩放因子设置为 3;

（2）柱状图颜色设置为 Seaborn 中调色板“hls”的默认颜色;

（3）柱状图上需要显示数据标签，数据标签中需要带有单位（℃），颜色为黑色;

（4）设置图像标题为“2011-2021 中高温城市排名”;

17

（5）横轴标签为“城市”，纵轴标签为“最高温（℃）”；

（6）横轴的刻度标签为各城市的名字；

（7）纵轴的刻度范围是（0,55）。

（8）绘制完成后将图片粘贴到答题报告对应位置。

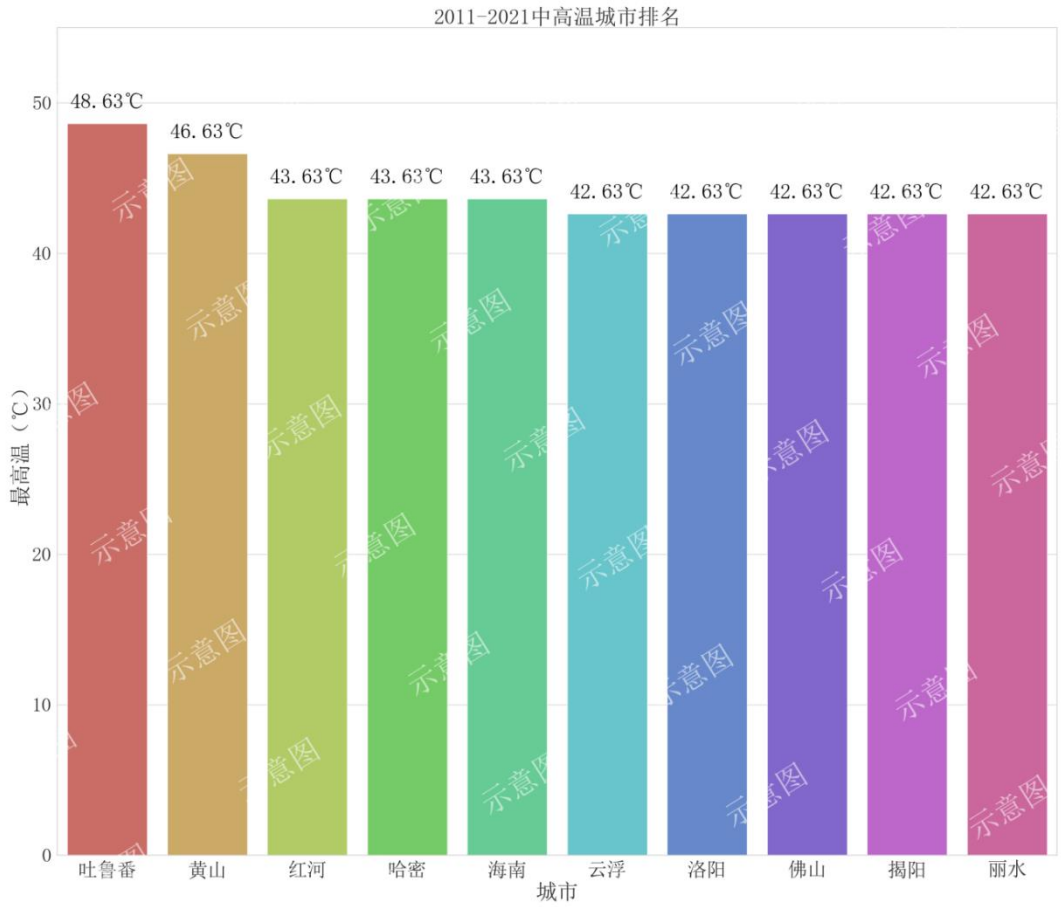

图 2 结果示意图（图中数据请以实际计算结果为准）

（二）任务二：业务分析

现有一份关于天气的可视化分析结果，请对结果进行业务分析，并给出可解释型结论。

下图是 3 个城市 10 年来每月平均风力等级走势图，请对该图以及图表中的数据进行分析并给出合理解释。

18

图 3 风力等级走势图

19