金砖国家职业技能大赛（金砖国家未来技能和技术挑战赛）

机器学习与大数据

BRICS-FS-02-RU

样题（全国选拔赛暨国内决赛）

2024 年 02 月

目录

1 参赛形式

本次赛项为个人赛。

2 竞赛内容

本次竞赛由 4 个模块组成，要求参赛选手按顺序完成所有竞赛内容。竞赛时会向参赛选手提供统一的赛题文件、竞赛数据集、基础操作说明文件，以及为保障每个任务模块的独立性与公平性所需的技术基础条件。

竞赛内容包含基于机器学习与大数据的以下任务模块：

模块 A：大数据

模块 B：数据仓库

模块 C：数据分析

模块 D：机器学习

如果参赛选手不遵守职业健康安全环境要求，或使自己和其他选手面临危险，他们可能会被取消比赛资格。

参赛选手完成模块后，将对结果进行评分。

3 项目模块和时间要求

3.1 项目模块和时间要求

机器学习与大数据赛项共 4 个模块，要求参赛者总共用时 480min。具体项目模块名称和时间要求参照表 1.

表 1 项目模块和时间要求清单

序号	模块名称	竞赛内容完成时间
1	模块 A：大数据	120min
2	模块 B：数据仓库	120min
3	模块 C：数据分析	120min
4	模块 D：机器学习	120min

3.2 任务内容

模块 A：大数据（120min）

模块描述：

模块 A 专注于 Hadoop 生态系统的搭建与应用。参赛选手将通过此模块学习如何配置 Hadoop 平台，并利用它进行基础的数据处理。模块还涵盖了使用 Sqoop 工具进行数据迁移以及使用 Hive 进行数据查询的基本操作。通过这些操作，参赛选手能够掌握数据从输入到处理的全流程，为处理更复杂的数据分析任务打下基础。同时，介绍了 Spark 技术，帮助选手理解如何通过内存计算来优化数据处理速度和效率。这一模块旨在为选手提供实际操作 Hadoop 及其相关工具的经验，以应对各种大数据场景的需求。

任务一：Hadoop 基础环境搭建

本环节需要使用`ec2-user`用户完成相关配置,安装 Hadoop 需要配置 Linux 前置基础环境，具体要求如下：

1、根据每节点的不同 ip 地址，请完成所有节点`/etc/hosts`文件的调整，并实现节点间相互免密登录；

2、在 master 节点将 Hadoop 解压到`/usr/local/src/`目录下，按以下需求修改集群配置文件，并将修改完成后的解压包分发至 slave1、slave2 节点相应路径，同时在`/etc/profile`文件配置好 Hadoop 相关环境变量，初始化 Hadoop 环境 namenode；

3、启动 Hadoop 集群，查看集群是否正常。

任务二：Sqoop 安装配置

本环节需要使用 ec2-user 用户完成相关配置,已安装 Hadoop 及需要配置前置环境，具体要求如下：

提示：如果遇到 jar 包缺失问题，可参考`/home/ec2-user/packages/`路径下 jar 包

1、将 master 节点 Sqoop 安装包解压到`/usr/local/src/`目录下；

2、修改`/etc/profile`文件，设置 Sqoop 环境变量；

3、修改并配置 Sqoop 相关配置文件，然后分发到每个节点；

4、测试 sqoop 组件是否运行正常。

任务三：Hive 安装配置

本环节需要使用`ec2-user`用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

1、将 master 节点 Hive 安装包解压到`/usr/local/src/`目录下；

2、修改`/etc/profile`文件，设置 Hive 环境变量；

3、按照以下要求，修改并配置 Hive 相关配置文件，然后分发到每个节点；

* `slave2`节点 mysql 数据库作为 Hive 元数据库；

* 配置 Hive 远程元数据启动模式，将 slave2 节点作为服务端，其它节点为客户端。

4、初始化 Hive 元数据；

5、确认 Hive 组件是否运行正常

任务四：Spark 安装配置

本环节需要使用`ec2-user`用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体要求如下：

1、将 master 节点 Spark 安装包解压到`/usr/local/src/`目录下；

2、修改`/etc/profile`文件，设置 Spark 环境变量；

3、按照以下要求，修改并配置 Hive 相关配置文件，然后分发到每个节点；

* master 节点启动 Master，slave1、slave2 节点启动 Worker

* 配置 DRIVER 内存大小为 4G

* 配置 EXECUTOR 内存大小为 4G

* 配置 EXECUTOR 核心数为 2

4、启动 Spark 集群，查看集群是否正常。

模块 B：数据仓库（120min）

模块描述：

模块 B 专注于数据仓库技术的应用，包括数据的抽取、存储和清洗。该模块主要任务是使用 Sqoop 工具从 MySQL 数据库中全量或增量地抽取数据到 Hive 的 ODS 层。这一过程中，参赛者将学习如何处理和转移大量数据，并保持数据的一致性和准确性。在数据成功抽取到 ODS 层后，参赛者需要执行数据清洗操作，以提高数据质量，然后将清洗后的数据转移到数据仓库的更高层次—DWD 层。在 DWD 层中，数据会被进一步聚合和处理，目的是为了支持更复杂的数据分析和业务决策。通过这些步骤，参赛者不仅能够实践数据抽取和清洗的技能，还能深入理解数据仓库在实际业务中的应用，特别是如何通过多层数据处理来支持和优化数据分析和决策过程。

环境说明：

在实现模块 A 的基础环境下，完成模块 B 所有任务

任务一：ods 层数据抽取

使用 Sqoop 工具，将 Mysql 的 brics 库中表 APP_LAUNCH_LOGS、USER_PLAYBACK_DATA 、 USER_PORTRAIT_DATA 、 VIDEO_RELATED_DATA 、USER_INTERACTION_DATA_PART1 的数据分别全量抽取到 Hive 的 ods 库中对应表 app_launch_logs 、 user_playback_data 、 user_portrait_data 、video_related_data 、 user_interaction_data 中 ， 将 表USER_INTERACTION_DATA_PART2 的数据增量抽取到 Hive 的 ods 库中对应表

user_interaction_data。

1. 抽取 brics 库中`APP_LAUNCH_LOGS`的全量数据进入 Hive 的 ods 库中表`app_launch_logs`。字段排序、类型不变，同时添加静态分区，分区字段的 key 为`date_load`,类型为 String，且值为`20220527`；

2. 抽取 brics 库中`USER_PLAYBACK_DATA`的全量数据进入 Hive 的 ods 库中表`user_playback_data`。字段排序、类型不变，同时添加静态分区，分区字段 key 为`date_load`,类型为 String，且值为`20220527`；

3. 抽取 brics 库中`USER_PORTRAIT_DATA`的全量数据进入 Hive 的 ods 库中表`user_portrait_data`。字段排序、类型不变，同时添加静态分区，分区字段 key 为`date_load`,类型为 String，且值为`20220527`；

4. 抽取 brics 库中`VIDEO_RELATED_DATA`的全量数据进入 Hive 的 ods 库中表`video_related_data`。字段排序、类型不变，同时添加静态分区，分区字段 key 为`date_load`,类型为 String，且值为`20220527`；

5. 首先抽取 brics 库中`USER_INTERACTION_DATA_PART1`的全量数据进入 Hive 的 ods 库中表`user_interaction_data`；再抽取 brics 库中`USER_INTERACTION_DATA_PART2`的增量数据进入 Hive 的 ods 库中表`user_interaction_data`，根据`USER_INTERACTION_DATA_PART1`表中`index`字段作为增量字段，只将新增的数据抽入，字段排序、类型不变。

任务二：dwd 层数据清洗

调用 spark 集群环境并编写代码，将 ods 库中相应表数据按要求抽取到 Hive 的 dwd 库中对应表中。

1.将 ods 库中`app_launch_logs`表数据按照以下要求抽取到 dwd 库中`dim_app_launch_logs`表。

* 以`user_id`字段进行聚合，`launch_date`和`launch_type`字段合并成 list 类型。

* 添加`dwd_insert_user`、`dwd_modify_user`列，均填写“user1”值。

* 按照`user_id`字段顺序排序。

2.将 ods 库中`user_playback_data`表数据按照以下要求抽取到 dwd 库中`dim_user_playback_data`表。

* 以`user_id`、`date`、`item_id`字段进行聚合，`playtime`字段合并成 list 类型。

* 添加`dwd_insert_user`、`dwd_modify_user`列，均填写“user1”值。

3. 将 ods 库中`video_related_data`表数据按照以下要求抽取到 dwd 库中`dim_video_related_data`表。

* 以`item_id`字段进行聚合，其它字段合并成 list 类型。

* 添加`dwd_insert_user`、`dwd_modify_user`列，均填写“user1”值。

4. 将 ods 库中`user_portrait_data`表数据按照以下要求抽取到 dwd 库中`dim_user_portrait_data`表。

* 删除多余的行重复值。

* 添加`dwd_insert_user`、`dwd_modify_user`列，均填写“user1”值。

* 添加动态分区，分区字段类型为 int，且值为`territory_code`字段内容。

5. 将 ods 库中`user_interaction_data`表数据按照以下要求抽取到 dwd 库中`dim_user_interaction_data`表。

* 按照`index`字段顺序排序。

* 添加`dwd_insert_user`、`dwd_modify_user`列，均填写“user1”值。

* 添加动态分区，分区字段类型为 int，且值为`date`字段内容。

任务三：dws 层指标计算

1. 编写代码，根据 Hive 中数据统计不同地区、每种设备的用户数量和 APP 启动次数，存入 master 节点 MySQL 数据库`dws.regiontype`表中，最终结果排除存在缺失值的数据。

2. 编写代码，根据 Hive 中数据计算出每个地区的每位用户平均回放时长和所有地区每用户平均回放时长相比较结果（“high/low/same”），存入 master 节点 MySQL 数据库`dws.regionavgpt`表（表结构如下）中，最终结果排除存在缺失值的数据。

3. 结合 slave1 节点 mysql 数据库中的`brics.train`表分析，计算以`end_date`时间开始未来 7 天 APP 登录天数，补全到 train 表的`label`字段，并保存完成后的数据到 master 节点 MySQL 数据库`dws.train`表中。

4. 请结合 master 节点 MySQL 数据库 dws.train 表，以及 Hive 数据仓库中`ods.video_related_data`和`ods.user_playback_data`表对数据进行处理。

- 按照以下给出的自定义函数`target_encoding_spark`方法分别处理`father_id`、`cast`、`tag_list`字段，计算`father_id`、`cast`、`tag_list`各自对应的`score`新字段值；

- 分别保存处理完成的`score`字段数据到 master 节点 MySQL 数据库`dws.father_id_score`、`dws.tag_id_score`和`dws.cast_id_score`中。

5. 编写代码，根据 Hive 中数据统计连续两天启动过 APP 的用户，并计算 APP 启动次数，存入 master 节点 MySQL 数据库`dws.usercontinuelaunch`表中。

模块 C：数据分析（120min）

模块描述：

在这一模块中，选手需要处理和分析从 Hive 数据仓库和 MySQL 数据库中提取的数据。重点是通过各种图形（如饼图和柱状图）展示数据分析结果，帮助参赛选手理解和掌握如何将复杂数据转化为直观的图形信息。参赛选手将执行一系列分析，包括 APP 启动量分析、用户交互行为分析及视频内容偏好分析，每项分析后都需要生成相应的图表。此外，选手还需处理数据的预处理和清洗，确保分析的准确性。完成后，分析结果将保存回 MySQL 数据库，以便进行进一步的使用或展示。整个模块 C 的核心是让参赛选手通过实际的数据操作，学习如何提取有价值的信息，并有效地呈现出来，从而加深对数据分析在业务和决策中应用的理解。

任务一：APP 启动分析

编写 Python 工程代码，查询 Hive 数据仓库和 MySQL 中数据，完成任务要求的数据分析以及可视化图形展示。

1. 统计用户 APP 的总启动量。保存正确结果到变量`taskC_1_1`中，并运行已给出的答案保存代码，保存结果到指定位置。

2.分析 APP 启动类型，按照以下要求，以饼图进行展示。

- 计算不同启动类型的次数占比

- 显示百分比

3.请结合 Hive 数据仓库 dwd.dim_app_launch_logs 表数据,以及 master 节点 mysql 数据库中的 dws.train 表和 slave1 节点 mysql 数据库中的 brics.test 表对数据进行处理。

- 按照以下给出的自定义函数`gen_launch_seq`方法处理 dwd.dim_app_launch_logs 表的`date`字段，对应计算每一个`user_id`的`launch_seq`新字段值;

- 保存处理完成后的数据到 master 节点 MySQL 数据库`train.launch_seq`和`test.launch_seq`表。

任务二：交互行为分析

1.统计不同互动方式，按照以下要求，以柱状图进行展示。

- 计算不同互动方式数量，并显示在标签中。

- 从左到右，按照互动数量降序排序。

2.分析哪个视频互动次数最多？保存正确的`item_id`到变量`taskC_2_2`，并运行已给出的答案保存代码，保存结果到指定位置。

3.分析出连续三天都有过互动行为的不重复用户数量。保存正确结果到变量 taskC_2_3,并运行已给出的答案保存代码，保存结果到指定位置。

4.处理 Hive 数据仓库`dwd.dim_user_portrait_data`表中`device_ram`和`device_rom`字段，存在多个值时，只保留第一个的值（例如`51872;52472`处

理后`51872`）。并保存清洗完成后的数据到 master 节点 MySQL 数据库`train.user_portrait_data`表。

任务三：视频偏好分析

1. 分析视频时长对用户偏好的影响，按照以下要求，以折线柱状图展示。

- 计算不同时间长度视频数量，以柱状图显示；计算不同长度视频观看用户数量，以折线图展示

- x 轴为视频时间长度；左边 y 轴为视频数量，右边 y 轴为用户数量

- 视频时长字段，从左到右升序排列。

2.通过视频回看数量分析，哪位演员的作品更受欢迎。保存正确的`cast`值到变量`taskC_3_2`，并运行已给出的答案保存代码，保存结果到指定位置。

3. 请结合 Hive 数据仓库 dwd.dim_user_playback_data 表数据，以及 master 节点 mysql 数据库中的 dws.train 表和 slave1 节点 mysql 数据库中的 brics.test 表对数据进行处理。

- 按照以下给出的自定义函数`get_playtime_seq`方法处理`dwd.dim_user_playback_data`表的`date`字段，对应计算每一个`user_id`的`playtime_seq`新字段值；

- 保存处理完成后的数据到 master 节点 MySQL 数据库`train.playback_seq`和`test.playback_seq`表。

4. 请结合 master 节点 MySQL 数据库`dws.item_scores`表，以及 slave1 节点`brics.train`和`brics.test`表对数据进行处理。

提示：Python 的 list 类型存入到 MySQL 的默认字段类型为 text

任务四：多维度分析

1.分析哪个年龄段的用户在看视频时更喜欢互动？按照以下要求，以玫瑰图进行展示。

- 统计不同年龄段有过交互行为的用户数量。

- 顺时针方向年龄段升序排列。

- 标签格式为`age:percent`，百分比保留两位小数。例如：`1:8.71%`

2. 计算 APP 登录量、视频回放次数、交互次数之间的转换率，并按照以下要求，以漏斗图展示。

- 从上到下依次为：APP 登录人数转化率、视频回看人数转化率、交互人数转化率。

- 每个环节用不同颜色标识。

- 标签格式为`阶段名称:percent`，百分比保留两位小数。例如：`lanuch:100.00%

3. APP 启动周期分析。计算 7 天内启动过 APP 两次及以上的客户，占启动过 APP 两次及以上的所有客户中的比例（保留两位小数）。保存正确答案到变量`taskC_4_3`，并运行已给出的答案保存代码，保存结果到指定位置。

模块 D：机器学习 (120min)

模块描述：

选手主要任务是构建一个预测模型来预测用户未来一周内 APP 的登录次数。通过处理和分析大量数据集，参赛选手将开发模型，使用用户 ID、登录开始时间和登录天数等信息作为输入。模块 D 要求选手掌握如何利用现有数据来训练回归模型，并通过模型预测用户的行为。模型的性能将通过均方误差（MSE）来评

估，任务核心在于使用机器学习工具和技术处理和分析数据，使参赛选手能够有效地从数据中学习并做出准确的预测。通过实际的机器学习项目，提高选手对数据科学和大数据技术的理解，以及如何在实际问题中应用这些技术。不仅使参赛选手的技术技能提升，还加深了他们对机器学习实际应用的认识。

评分标准

目标

本次数据为某 APP 启动数据，目标为训练一个回归模型预测用户未来 7 天登录 APP 的天数，并在测试数据集上尝试获得最优结果。

评价指标

MSE 值：`sklearn.metrics.mean_squared_error`

MSE 值（均方误差）：指参数估计值与参数真实值之差平方的期望值。是统计学中用来衡量回归模型预测误差的一种指标，可以评价数据的变化程度，MSE 的值越小，说明预测模型描述实验数据有更好的精确度，理论最小值为 0。

最终得分

$$score = 30/e^{mse^2-3.5}$$

Step1. 数据获取

- 训练集路径：slave1 节点 Mysql 数据库`brics.train`

- 测试集路径：slave1 节点 Mysql 数据库`brics.test`

Step2. 数据探查

Step3. 模型训练

Step4. 模型预测

Step5. 结果保存

任务结束

4 项目模块评分标准

项目模块评分标准参照表 2。

表 2 评分标准

模块	任务	配分
A	大数据	20
B	数据仓库	20
C	数据分析	30
D	机器学习	30
合计		100

注：样题最终解释权归组委会所有。