实验3：Hive表的Join操作

实验内容与要求

- 本实验要求学生使用MapReduce任务，实现指定表的Join操作，并将Join后的数据表存储在Hive中。
- 任务1：编写一个MapReduce程序，实现学校信息表（university.tbl）和国家信息表（country.tbl）的join操作，按国家2字母编码（n_alpha-2-code）和学校所属国家2字母编码（u_alpha-2-code）字段进行join。join后的表名为UniversityCountry，从左至右应包含学校信息表（university.tbl）的学校序号（u_key）、学校名称（u_name）、学校主页（u_webpage）字段和国家信息表（country.tbl）的国家名称（n_name）字段，如下图所示。

```
0|Marywood University|http://www.marywood.edu|United States
1|University of Petroleum and Energy Studies|https://www.upes.ac.in/|India
2|Cégep de Saint-Jérôme|https://www.cstj.qc.ca|Canada
3|Lindenwood University|http://www.lindenwood.edu/|United States
4|DAV Institute of Engineering & Technology|http://www.davietjal.org/|India
5|Lovely Professional University|http://www.lpu.in/|India
6|Sullivan University|https://sullivan.edu/|United States
7|Florida State College at Jacksonville|https://www.fscj.edu/|United States
8|Xavier University|https://www.xavier.edu/|United States
9|Tusculum College|https://home.tusculum.edu/|United States
```

实验3：Hive表的Join操作

实验内容与要求

- 本实验要求学生使用MapReduce任务，实现指定表的Join操作，并将Join后的数据表存储在Hive中。
- 任务2：编写一个MapReduce程序，将课程信息表（course.tbl）和学校信息表（university.tbl）join成课程-学校信息表。合并后的表名为CourseUniversity，从左至右应包含课程信息表的课程序号（c_key）、课程名称（c_name）、课程类型（c_subject）、课程学时（c_hours）信息和学校信息表（university.tbl）的学校序号（u_key）、学校名称（u_name）、学校主页（u_webpage）字段，如下图所示。

```
0|US Citizenship Comparatively|AAS|gendered|0|Marywood University|http://www.marywood.edu
0|US Citizenship Comparatively|AAS|gendered|1|University of Petroleum and Energy Studies|https://www.upes.ac.in/
0|US Citizenship Comparatively|AAS|gendered|2|Cégep de Saint-Jérôme|https://www.cstj.qc.ca
0|US Citizenship Comparatively|AAS|gendered|3|Lindenwood University|http://www.lindenwood.edu/
0|US Citizenship Comparatively|AAS|gendered|4|DAV Institute of Engineering & Technology|http://www.davietjal.org/
0|US Citizenship Comparatively|AAS|gendered|5|Lovely Professional University|http://www.lpu.in/
0|US Citizenship Comparatively|AAS|gendered|6|Sullivan University|https://sullivan.edu/
0|US Citizenship Comparatively|AAS|gendered|7|Florida State College at Jacksonville|https://www.fscj.edu/
0|US Citizenship Comparatively|AAS|gendered|8|Xavier University|https://www.xavier.edu/
0|US Citizenship Comparatively|AAS|gendered|9|Tusculum College|https://home.tusculum.edu/
```

实验3：Hive表的Join操作

实验内容与要求

- 通过使用MapReduce进行join操作后，需要将结果输出到HDFS的个人目录上，然后导入到Hive数据库中。
- 具体操作为：进入SQL On Hadoop页面，使用Hive建表管理上一步输出的结果，输入建表语句。执行上述指令后，在Hive上通过show tables能查看到对应的UniversityCountry和CourseUniversity表，并且能通过select语句查询表格的内容。
- 要求程序跑完后可通过Hive查看生成的表，详见“本科教学支撑平台”。

实验数据

- 本实验考虑在国家-学校-课程数据库中进行表的聚集操作。实验提供三张数据表：country.tbl, university.tbl, course.tbl。数据格式等具体信息详见“本科教学支撑平台”。
- 单机测试样例：提供部分数据作为单机测试样例，可在“本科教学支撑平台”中下载。该数据集主要供本地调试使用。
- 全部数据集：全部数据集位于集群的 HDFS 存储上，HDFS 存储位置为：hdfs://master001:9000/data/2022s/hive_join

实验报告提交要求

- 实验报告要求提交一个压缩包，除了包含源代码、JAR包、JAR包执行方式说明，还需要包含一个实验报告。实验报告中包含：
  - Map 和 Reduce 的设计思路（含 Key、Value 类型）。
  - MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
  - hive输出结果文件的部分截图（至少25条）。注意，由于实验的数据量较大，直接使用select * 命令可能需要很长的输出时间；应当使用limit关键字控制输出结果的数量。
  - 请在报告中包含在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容。每个MapReduce Job对应一个执行报告，报告提交要求与倒排索引试验相同。

实验报告提交要求

- 实验报告文件命名规则： MPLab3-小组编号-组长姓名.docx
- 实验报告提交至：本科教学支撑平台 http://cslabcms.nju.edu.cn/
- 实验提交截止日期：5月8日（包含当天）