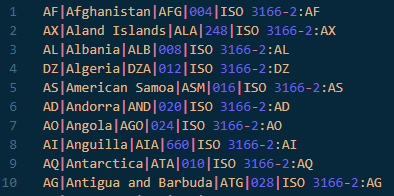
# 实验3　Hive表的Join操作

　　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。

## 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。实验数据在hdfs://master001:9000/2022s/hive\_join目录下。单机测试可以使用本科实验教学平台上的测试数据集（dataset\_example.zip）。所有输入文件使用UTF-8字符编码，字段之间使用竖线（|）分隔。如下面几张图所示：

country.tbl：



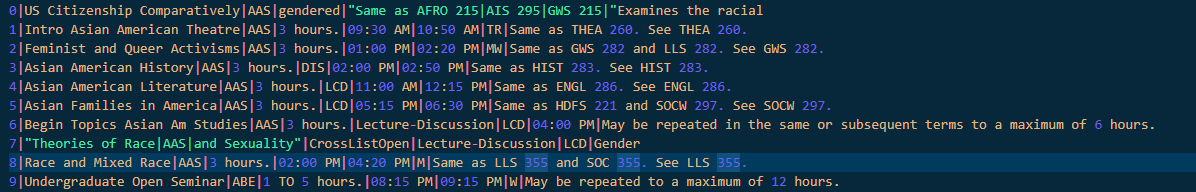
　　country.tbl为国家信息数据表，包含国家2字母编码（n\_alpha-2-code）、国家名（n\_name）、国家3字母编码（n\_alpha-3-code）、国家数字码（n\_numeric-code）和ISO国家代码（n\_ISO-3166-2）五个字段。

university.tbl：



university.tbl为世界大学信息表，包含学校序号（u\_key）、学校名称（u\_name）、学校所属国家2字母编码（u\_alpha-2-code）、学校所在省份（u\_province）、学校主页网址（u\_webpage）、学校域名（u\_domain）六个字段。

course.tbl：



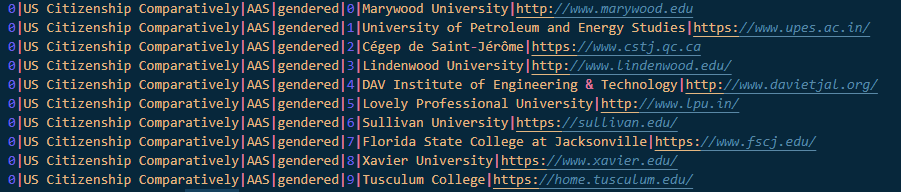
course.tbl为课程信息表，包含课程序号（c\_key）、课程名称（c\_name）、课程类别（c\_subject）、课程学时（c\_hours）、课程开始时间（c\_start\_time）、课程结束时间（c\_end\_time）、上课时间（c\_day）、课程描述（c\_description）八个字段。

## 实验任务

　　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。



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。



## 实验结果输出

　　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：

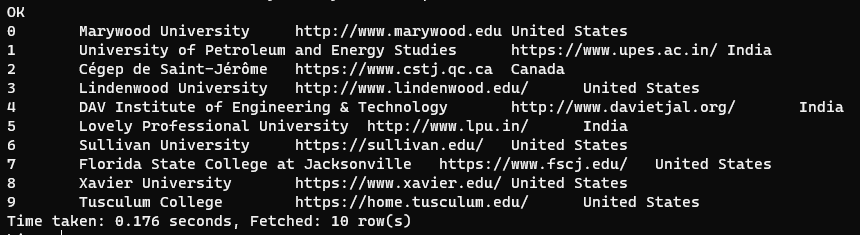
create table UniversityCountry(u\_key int, u\_name string, u\_webpage string, n\_name string) row format delimited fields terminated by '|' location '/user/nju\_stxy/output/';

create table CourseUniversity(c\_key int, c\_name string, c\_subject string, c\_hours string, u\_key int, u\_name string, u\_webpage string) row format delimited fields terminated by '|' location '/user/nju\_stxy/output/';

其中，row format delimited fields terminated by '|'表示数据集分隔符为竖线；location '/user/nju\_stxy/output/' 表示所管理的数据集在HDFS上的路径。

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，如下图所示：

UniversityCountry表：



CourseUniversity表：



## 实验报告

在最后提交的压缩包中，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。实验报告中请包含：

1. Map和Reduce的设计思路（含Key、Value类型）。

2. MapReduce程序中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。

3. 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。

4. 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。