

题目2

2 分析如下HiveQL，生成的MapReduce执行程序，map函数输入是什么？输出是什么，reduce函数输入是什么？输出是什么？

```
INSERT OVERWRITE TABLE pv_users
SELECT pv.pageid, u.age
FROM page_view pv
JOIN user u
ON (pv.userid = u.userid);
```

Page_view表和user表结构与数据示例如右

page_view		
pageid	userid	time
1	111	9:08:01
2	111	9:08:13
1	222	9:08:14

user		
userid	age	gender
111	25	female
222	32	male

解答

map 函数

1. 输入：page_view、user 的每一个行
 1. 示例：page_view 表 3 行，user 表 2 行
2. 输出：根据 userid 关联的 pv.pageid, u.age 数据
 1. 示例：输出 3 行
 2. 在map的输出value中为不同表的数据打上tag标记，在reduce阶段根据tag判断数据来源

reduce 函数

1. 输入：map 函数输出的 pv.pageid, u.age
2. 输出：把 pv.pageid, u.age 以覆盖方式写入 pv_users 表中