

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

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

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.

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

userid	age	gender
111	25	female
222	32	male
2. 输出：userid 为key，value 是表id 和 pageid
1. 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

2.

key	value
111	<1,1>
111	<1,2>
222	<1,1>

key	value
111	<2,25>
222	<2,32>

reduce 函数

1. 输入：map 的输出根据 key (userid) 排序

1.

key	value
111	<1,1>
111	<1,2>
111	<2,25>

Red

key	value
222	<1,1>
222	<2,32>

2. 输出：根据 key (userid) join value (表id、age) ， 输出 pv.pageid, u.age

1.

pv_users	
Pageid	age
1	25
2	25
pageid	age
1	32