<http://www.codelast.com/%E5%8E%9F%E5%88%9Bpig%E4%B8%AD%E7%9A%84%E4%B8%80%E4%BA%9B%E5%9F%BA%E7%A1%80%E6%A6%82%E5%BF%B5%E6%80%BB%E7%BB%93/>

（1）关系（relation）、包（bag）、元组（tuple）、字段（field）、数据（data）的关系

1 关系就是包，是一个外部包outer bag

2 {}里面的东西是包的内容，包里面是一堆元组的集合

3 一个元组（tuple）是若干字段（field）的一个有序集（ordered set）。在pig中表示数据时，用小括号()括起来的东西表示一个元组。

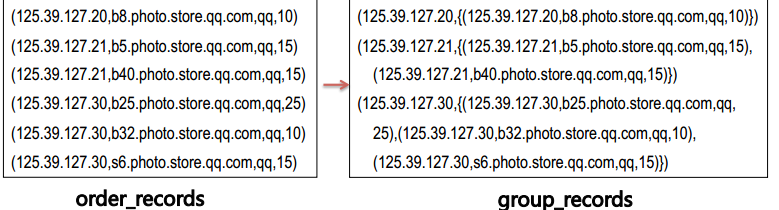
Outer bag：

records:{ip: chararray,host: chararray,sp: chararray, traffic: int}

inner bag：

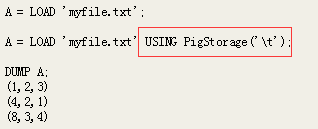
group\_records: {group: chararray, order\_records: {(ip: chararray, host: chararray,

sp: chararray, traffic: int)}}



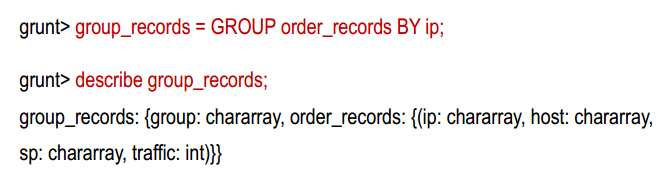
几个比较重要的用法总结：

Load载入：



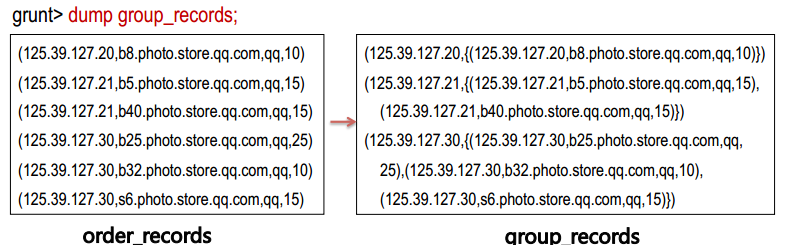
如果不使用Using关键字，默认的PigStorage方法使用的分隔符就是tab

Group分组：

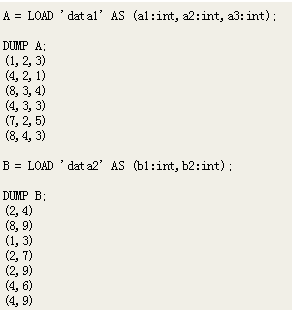


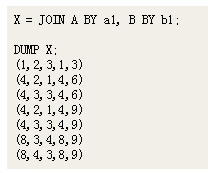
如上图所示：

按照ip分组之后，新产生的关系(relation)或者是包(bag)会自动命名一个新的字段：group，这个字段就是上面分组的依据。



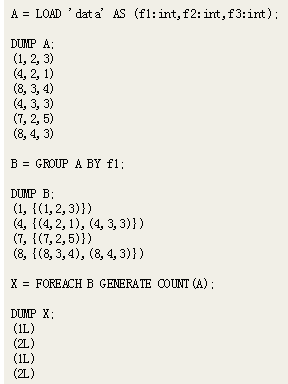
Join联结：





下面是一些用到的函数，这些函数在使用的时候，一般都需要先进行group分组，然后使用foreach…generate…来产生新的字段关系。

Count：



Sum：

