**港中文数据匹配最后一步的问题说明**

**表a：tmp\_dm\_bas.company\_info\_match3**

每一行代表一家公司。

需要字段：a.symbol, a.fullname, a.sfname, a.customer\_code, a.cust\_source

分别表示：上市公司代码，公司全称，匹配到的公司在顺丰登记的名字，月结账号，匹配精度（1，2均为精准，3为模糊）

**表b: gdl.tt\_waybill\_info**

每一行代表一条运单。

需要字段：consignor\_comp\_name, consignee\_comp\_name, freight\_monthly\_acct\_code，consignor\_addr\_native consignee\_addr\_native, freight\_payment\_type\_code

分别表示：寄件公司名称，收件公司名称，付费月结账号，寄件方月结账号，收件方月结账号，付费方式（1寄付，2到付，3第三方付）

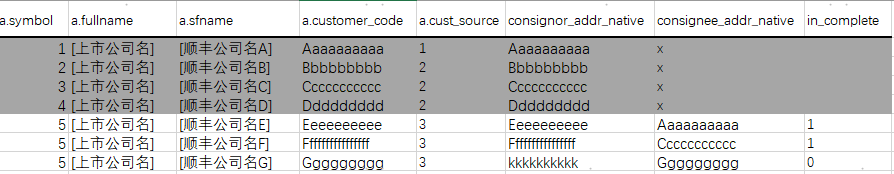
**背景**：表a中的a.fullname是研究的所有上市公司，且在顺丰是月结客户。但是由于名称有长有短，比如业务给的研究对象叫做“美邦服饰”，但是在顺丰是“美特斯邦威”，实际是一家。因此，对于a.cust\_source=3的需要用如下方法做筛选：

**逻辑**：

第一步： a中a.cust\_source=3的部分 left join b生成新表表1（输出a.symbol, a.fullname, a.sfname, a.customer\_code, a.cust\_source，consignor\_comp\_name, consignee\_comp\_name, consignor\_addr\_native consignee\_addr\_native，freight\_payment\_type\_code）, 即a.cust\_source=3的公司在顺丰的全部运单.

第一步的连接为a.customer\_code = b.freight\_monthly\_acct\_code，月结账号连接

第二步：判断表1中的a.customer\_code这个月结号是收方还是寄方，若为寄方，则判断收方月结账号是否在a.cust\_source为1,2对应的a.customer\_code中，是则记为1，否则为0。若为收方，则判断寄方月结账号是否在a.cust\_source为1,2对应的a.customer\_code中，是则记为1，否则为0。1,0记为新列in\_complete。此时该表每一行仍为一条运单。



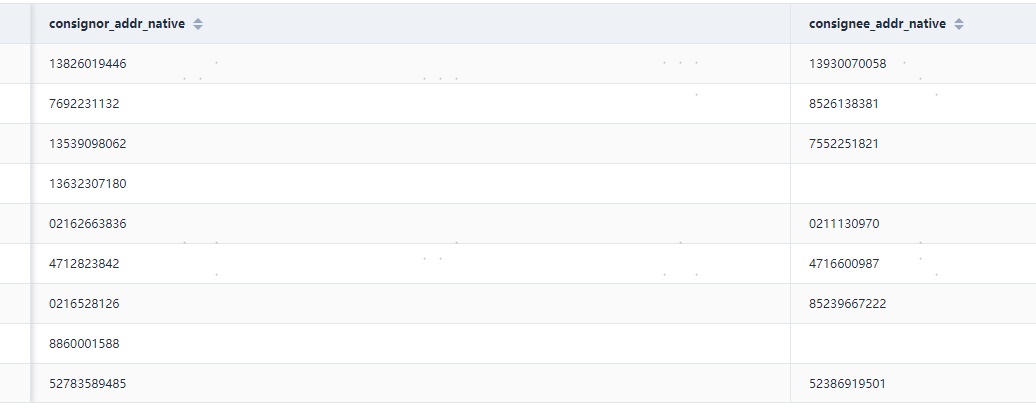
假如symbol为1,2,3,4,5就是要研究的全部公司。前四行并不会在建的表里出现，因cust\_source是1,2，只是为了解释。后三行实际表示的是同一个公司，因为模糊匹配导致它匹配到了三个。前两个是寄付，最后一个是到付，最后一个in\_complete为0是因为kkkkkk没出现在前四行，也就是第三个从没有和前四行的公司有快递往来。

第三步：统计a.cust\_source=3中的每家公司记为1的次数生成新的一列，记为n\_order。此时该表已经汇总为公司层面，每一行为一家公司。最终得到：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| a.symbol | a.fullname | a.sfname | a.customer\_code | a.cust\_source | n\_order |
| 000001 | [上市公司名] | [顺丰公司名E] | Eeeeeeeeee | 3 | 11 |
| 000001 | [上市公司名] | [顺丰公司名F] | Ffffffffffffffff | 3 | 20 |
| 000001 | [上市公司名] | [顺丰公司名G] | Ggggggggg | 3 | 3 |
| … | … | … | … | … |  |

注： 表b中的consignor\_comp\_name, consignee\_comp\_name, freight\_monthly\_acct\_code，freight\_payment\_type\_code做辅助判断用，可以输出到最后的结果。

**问题：** 运单表中的consignor\_addr\_native consignee\_addr\_native只有极少部分是月结账号，原因在于A寄给B的时候很少会写成B的公司全称或月结账号这种形式，因此无法判断B的公司类型，字段多为私人手机号码或公司缩写+编号。同样的，尝试consignor\_comp\_name, consignee\_comp\_name,也是这个问题，即A和B均为私人名或公司名简写，匹配不到结果。如下：



最后一张是匹配结果示例，其中最后一列是n\_order，几乎全部为0.