Hadoop Filter:

Test data: Test.csv

Simple regex: .\*TT.\* (match each row where Intersecti = .\*TT.\*)

Complex regex: [A-Za-z]\*L/[A-Za-z]\* (match letters followed by L followed by a slash mark followed by letters)

Cluster size: 2,3,5,7,10

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| regex | 2 machine | 3 machine | 5 machine | 7 machine | 10 machine |
| .\*TT.\* |  |  |  |  |  |
| [A-Za-z]\*L/[A-Za-z]\* |  |  |  |  |  |

Hadoop join:

Test data: D1-D5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5\*25 | 5\*360 | 25\*825 | 360\*825 | 825\*9493 |
| D1.OBJECTID=D2.OBJECTID | D1.OBJECTID=D3.OBJECTID | D2.LOT=D4.OBJECTID | D3.OBJECTID=D4.Stories | D4.OBJECTID=D5.AddNum |
| D1.OBJECTID=D2.LOT | D1.OBJECTID=D3.FACILITYID | D2.HSENO=D4.UniqueID | D3.OBJECTID=D4.Units | D4.Units=D5.CompAddNum |

Task1: