**Assignment 4\_Notes**

**Reddit\_etl.py**

Cluster – reddit2

w/o cache(): 34.355s

w cache(): 30.579s

Cluster – reddit3

w/o cache(): 39.711s

w cache(): 36.527s

Cluster – reddit4

w/o cache(): 55.309s

w cache(): 46.812s

**Broadcast Join**

We broadcast the *smaller* RDD/dataframe

Before broadcast, every RDD shuffles. After broadcast, we copy smaller RDD to each executor, and only do map() to larger RDD. (reflects in code)

Join() works with 2 RDDs, but here we have one RDD and one broadcast object. So we need to map to the larger one (that is RDD here) *manually* to join these two (create tuples manually).

Reddit\_etl.py: output-2 (reddit2)

45.781s

Text

Description automatically generated

Reddit\_etl\_no\_cache.py: output-3 (reddit2)

34.227s

Same as before.

Relative\_score.py: output-1 (reddit2) same cl

1m13.084s

Text, application

Description automatically generated

Relative\_score\_bcast.py: output (reddit2) same cl

Text

Description automatically generated

Weather\_etl.py: output-4 (weather-2)

Table

Description automatically generated with low confidence