Distributed Computing Tasks II - Join Strategies

Manual join by grouping the RDDs by the join key (geographical_location_oid)

```
// Manual join for both RDDs
        val joinedRdd = saltedDetectionsRdd.cartesian(locationsRdd)
       .filter { case (detection, location) => detection._2 == location.getInt(0) }
       .map { case (detection, location) => (detection._2, (detection, location)) }
```

This code performs a manual join on two RDDs based on a common key (the second element of detection and the first integer element of location). The resulting RDD contains pairs of joined elements, with the join key as the first element of the tuple.

1. val joinedRdd = saltedDetectionsRdd.cartesian(locationsRdd)

This line creates a new RDD joinedRdd by computing the Cartesian product of saltedDetectionsRdd and locationsRdd. The Cartesian product is a set of all possible pairs of elements from the two RDDs. This operation is equivalent to a full outer join in SQL.

2. filter { case (detection, location) => detection._2 == location.getInt(0) }

This line filters the resulting RDD from the previous step to only include pairs where the second element of detection (accessed using _2) is equal to the first integer element of location (accessed using getInt(0)). This is equivalent to a join condition in SQL.

3. .map { case (detection, location) => (detection._2, (detection, location)) }

This line transforms the filtered RDD by mapping each pair to a new tuple. The first element of the tuple is the second element of detection (the join key), and the second element is a tuple containing both detection and location.

Created Unit test to test the manual join: run sbt test

```
24/08/17 11:23:22 INFO DAGScheduler: Job 0 finished: collect at ManualJoinTest.scala:34, took 2.168519 s
[info] ManualJoinTest:
[info] Manual Join Test
[info] - should return joined common records
[info] Run completed in 12 seconds, 375 milliseconds.
[info] Total number of tests run: 1
[info] Suites: completed 1, aborted 0
[info] Tests: succeeded 1, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[success] Total time: 18 s, completed Aug 17, 2024 11:23:22 AM
24/08/17 11:23:22 INFO SparkContext: Invoking stop() from shutdown hook
```

Code is in \src\test\scala\ManualJoinTest.scala

Drafted by: Kader