CS6240 - Homework 4 - Ronn George Jacob

* PageRank in Spark

1. Pseudocode

```
// Code to generate graph

var graph: List[(Int, Int)] = List.empty
for (vertex <- 1 to args * args) {
  if (vertex % args != 0) {
   // Setting the neighbor of vertex to be vertex +1
    graph = graph :+ (vertex, vertex + 1)
  }
  else {
   // This would be for dangling vertices.
    graph = graph :+ (vertex, 0)
  }
}


// Code to calculate page rank
// Setting value of number of vertices and iterations.
 val k = 100
 val iters = 10

// Setting value of alpha
 val alpha = 0.15D

 val G = (k*k).toDouble
 val single_start = spark.sparkContext.parallelize(List((0, 0.0))).partitionBy(new
HashPartitioner(1))
val graph = spark.sparkContext.parallelize(createGraph(k)).distinct()
                    .groupByKey().partitionBy(new HashPartitioner(1)).cache()

    // Initializing PR value to 1.0/(k^2)
  var ranks = graph.mapValues(v => 1.0 / (k * k).toDouble).cache()

  // Adding single dummy vertex
  ranks = ranks.union(single_start)

    var pr_init: Double = 0.0

    for (i <- 1 to iters) {
//      logger.info("+++++++++++++++++< Iteration "+ iters.toString+ ">+++++++++++++++++
+")
      var contribs = (graph.join(ranks).flatMap { case (_, (urls, rank)) =>
        val size = urls.size
        urls.map(p => (p, rank / size))
      }).reduceByKey(_ + _)

      if (contribs.lookup(0).isEmpty){
        // If no mass has been transferred, we would not add any page rank mass.
        pr_init = 0.0
      } else{
        pr_init = contribs.lookup(0).head / G
      }

    val mass_transfer = ranks.leftOuterJoin(contribs).filter(_._1 != 0)
        .mapValues { case (v, new_v) => new_v.getOrElse(0.0) }

      //  alpha * (1/|G|) + (1-alpha) * (mass_transferred to single dummy/|G|+v)
      ranks = mass_transfer.mapValues(v => ((alpha/G) + (1-alpha) * (pr_init + v)))
      println("start of lineage after iteration " + i)
      println(ranks.toDebugString)
      println("end of lineage after iteration " + i)

    }
//    logger.info(ranks.toDebugString)
ranks.sortBy(_._1, true, numPartitions = 1).saveAsTextFile(“ranks")
```

2. Lineage

Iteration 1

```
start of lineage after iteration 1
(1) MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
 |  MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
 |  MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
 |  MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
 |  MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
 |  CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
 |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
 |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
 |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
 |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
       |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
       +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
          |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
 |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
 +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
 |  ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
 +-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
    |  MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
    |  MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
```

```
   | CoGroupedRDD[10] at join at PageRankMod.scala:61 []
   | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
   |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
   +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
     +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
       | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
       +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
         | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
   | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
   | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
   |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
   | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
   |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
   +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
     +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
       | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
       +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
         | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
   | ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
   +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
end of lineage after iteration 1
```

Iteration 2

```
start of lineage after iteration 2
(1) MapPartitionsRDD[31] at mapValues at PageRankMod.scala:77 []
 | MapPartitionsRDD[30] at mapValues at PageRankMod.scala:74 []
 | MapPartitionsRDD[29] at filter at PageRankMod.scala:73 []
 | MapPartitionsRDD[28] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[27] at leftOuterJoin at PageRankMod.scala:73 []
 | CoGroupedRDD[26] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
 | MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
 | MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
 | MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
 | CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
 | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
 | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
 |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
 |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
   +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
     | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
     +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
       | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
 | ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
 +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
 | ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
 +-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
   | MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
   | MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
```

```
  |  CoGroupedRDD[10] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
  |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
  +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
|  ShuffledRDD[25] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[24] at flatMap at PageRankMod.scala:61 []
  |  MapPartitionsRDD[23] at join at PageRankMod.scala:61 []
  |  MapPartitionsRDD[22] at join at PageRankMod.scala:61 []
  |  CoGroupedRDD[21] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
  |  MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
  |  MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
  |  MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
  |  MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
  |  CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
  |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
  +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
  |  ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
  +-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
    |  MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
    |  MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
```

```
     | CoGroupedRDD[10] at join at PageRankMod.scala:61 []
     | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
     |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
     +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
       +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
         | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
         +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
           | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
     | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
     | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
     |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
     | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
     |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
     +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
       +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
         | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
         +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
           | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
     | ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
     +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
end of lineage after iteration 2
```

Iteration 3

```
start of lineage after iteration 3
(1) MapPartitionsRDD[42] at mapValues at PageRankMod.scala:77 []
 | MapPartitionsRDD[41] at mapValues at PageRankMod.scala:74 []
 | MapPartitionsRDD[40] at filter at PageRankMod.scala:73 []
 | MapPartitionsRDD[39] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[38] at leftOuterJoin at PageRankMod.scala:73 []
 | CoGroupedRDD[37] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[31] at mapValues at PageRankMod.scala:77 []
 | MapPartitionsRDD[30] at mapValues at PageRankMod.scala:74 []
 | MapPartitionsRDD[29] at filter at PageRankMod.scala:73 []
 | MapPartitionsRDD[28] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[27] at leftOuterJoin at PageRankMod.scala:73 []
 | CoGroupedRDD[26] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
 | MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
 | MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
 | MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
 | MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
 | CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
 | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
 | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
 |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
 |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
 +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
   +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
     | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
```

```
    +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
      |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
|  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
+-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
|  ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
  |  MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
  |  MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
  |  CoGroupedRDD[10] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
  |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
  +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
|  ShuffledRDD[25] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[24] at flatMap at PageRankMod.scala:61 []
  |  MapPartitionsRDD[23] at join at PageRankMod.scala:61 []
  |  MapPartitionsRDD[22] at join at PageRankMod.scala:61 []
  |  CoGroupedRDD[21] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
  |  MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
  |  MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
  |  MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
  |  MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
  |  CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
  |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  | ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
  +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
  | ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
  +-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
    | MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
    | MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
    | CoGroupedRDD[10] at join at PageRankMod.scala:61 []
    | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
    |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
    +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
      +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
        | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
        +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
          | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
    | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
    | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
    |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
    | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
    |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
    +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
      +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
        | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
        +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
          | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
    | ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
    +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
| ShuffledRDD[36] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[35] at flatMap at PageRankMod.scala:61 []
  | MapPartitionsRDD[34] at join at PageRankMod.scala:61 []
  | MapPartitionsRDD[33] at join at PageRankMod.scala:61 []
  | CoGroupedRDD[32] at join at PageRankMod.scala:61 []
  | ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      | ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        | ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  | MapPartitionsRDD[31] at mapValues at PageRankMod.scala:77 []
  | MapPartitionsRDD[30] at mapValues at PageRankMod.scala:74 []
  | MapPartitionsRDD[29] at filter at PageRankMod.scala:73 []
  | MapPartitionsRDD[28] at leftOuterJoin at PageRankMod.scala:73 []
  | MapPartitionsRDD[27] at leftOuterJoin at PageRankMod.scala:73 []
  | CoGroupedRDD[26] at leftOuterJoin at PageRankMod.scala:73 []
  | MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
  | MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
  | MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
  | MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
  | MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
  | CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
  | PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  | MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
|      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
|  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
|      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
+-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
  +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
    |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
    +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
      |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
|  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
+-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
|  ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
  |  MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
  |  MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
  |  CoGroupedRDD[10] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
  |      CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
  +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
|  ShuffledRDD[25] at reduceByKey at PageRankMod.scala:64 []
+-(1) MapPartitionsRDD[24] at flatMap at PageRankMod.scala:61 []
  |  MapPartitionsRDD[23] at join at PageRankMod.scala:61 []
  |  MapPartitionsRDD[22] at join at PageRankMod.scala:61 []
  |  CoGroupedRDD[21] at join at PageRankMod.scala:61 []
  |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
  |      CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
  +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
    +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
      |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
      +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
        |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
  |  MapPartitionsRDD[20] at mapValues at PageRankMod.scala:77 []
  |  MapPartitionsRDD[19] at mapValues at PageRankMod.scala:74 []
  |  MapPartitionsRDD[18] at filter at PageRankMod.scala:73 []
  |  MapPartitionsRDD[17] at leftOuterJoin at PageRankMod.scala:73 []
  |  MapPartitionsRDD[16] at leftOuterJoin at PageRankMod.scala:73 []
  |  CoGroupedRDD[15] at leftOuterJoin at PageRankMod.scala:73 []
  |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
  |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
```

```
    |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
    |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
    |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
    +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
      +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
        |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
        +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
          |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
    |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
    +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
    |  ShuffledRDD[14] at reduceByKey at PageRankMod.scala:64 []
    +-(1) MapPartitionsRDD[13] at flatMap at PageRankMod.scala:61 []
      |  MapPartitionsRDD[12] at join at PageRankMod.scala:61 []
      |  MapPartitionsRDD[11] at join at PageRankMod.scala:61 []
      |  CoGroupedRDD[10] at join at PageRankMod.scala:61 []
      |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
      |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
      +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
        +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
          |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
          +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
            |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
      |  PartitionerAwareUnionRDD[9] at union at PageRankMod.scala:55 []
      |  MapPartitionsRDD[8] at mapValues at PageRankMod.scala:53 []
      |     CachedPartitions: 1; MemorySize: 664.1 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
      |  ShuffledRDD[7] at partitionBy at PageRankMod.scala:50 []
      |     CachedPartitions: 1; MemorySize: 898.5 KB; ExternalBlockStoreSize: 0.0 B; DiskSize: 0.0 B
      +-(4) ShuffledRDD[6] at groupByKey at PageRankMod.scala:50 []
        +-(4) MapPartitionsRDD[5] at distinct at PageRankMod.scala:49 []
          |  ShuffledRDD[4] at distinct at PageRankMod.scala:49 []
          +-(4) MapPartitionsRDD[3] at distinct at PageRankMod.scala:49 []
            |  ParallelCollectionRDD[2] at parallelize at PageRankMod.scala:49 []
      |  ShuffledRDD[1] at partitionBy at PageRankMod.scala:48 []
      +-(4) ParallelCollectionRDD[0] at parallelize at PageRankMod.scala:48 []
end of lineage after iteration 3
```

3. Actions in the program - Lookup, saveAsTextFile, sortBy

This would inturn also call tasks like reduceByKey, join, union and mapValues for the dataset.

What was triggered can be determined by taking a look at the lineage
For eg:

```
2020-03-26 23:37:38 INFO  SparkContext:54 - Starting job: lookup at PageRankMod.scala:70
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Got job 11 (lookup at PageRankMod.scala:70) with 1 output partitions
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Final stage: ResultStage 101 (lookup at PageRankMod.scala:70)
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Parents of final stage: List(ShuffleMapStage 100)
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Missing parents: List()
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Submitting ResultStage 101 (ShuffledRDD[69] at reduceByKey at PageRankMod.scala:64), which has no missing parents
2020-03-26 23:37:38 INFO  MemoryStore:54 - Block broadcast_21 stored as values in memory (estimated size 3.3 KB, free 364.7 MB)
2020-03-26 23:37:38 INFO  MemoryStore:54 - Block broadcast_21_piece0 stored as bytes in memory (estimated size 2.0 KB, free 364.7 MB)
2020-03-26 23:37:38 INFO  BlockManagerInfo:54 - Added broadcast_21_piece0 in memory on 192.168.1.164:53182 (size: 2.0 KB, free: 364.8 MB)
```

```
2020-03-26 23:37:38 INFO  SparkContext:54 - Created broadcast 21 from broadcast at DAGScheduler.scala:1039
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Submitting 1 missing tasks from ResultStage 101 (ShuffledRDD[69] at
reduceByKey at PageRankMod.scala:64) (first 15 tasks are for partitions Vector(0))
2020-03-26 23:37:38 INFO  TaskSchedulerImpl:54 - Adding task set 101.0 with 1 tasks
2020-03-26 23:37:38 INFO  TaskSetManager:54 - Starting task 0.0 in stage 101.0 (TID 33, localhost, executor driver, partition 0,
ANY, 7649 bytes)
2020-03-26 23:37:38 INFO  Executor:54 - Running task 0.0 in stage 101.0 (TID 33)
2020-03-26 23:37:38 INFO  ShuffleBlockFetcherIterator:54 - Getting 1 non-empty blocks out of 1 blocks
2020-03-26 23:37:38 INFO  ShuffleBlockFetcherIterator:54 - Started 0 remote fetches in 1 ms
2020-03-26 23:37:38 INFO  Executor:54 - Finished task 0.0 in stage 101.0 (TID 33). 1349 bytes result sent to driver
2020-03-26 23:37:38 INFO  TaskSetManager:54 - Finished task 0.0 in stage 101.0 (TID 33) in 43 ms on localhost (executor driver)
(1/1)
2020-03-26 23:37:38 INFO  TaskSchedulerImpl:54 - Removed TaskSet 101.0, whose tasks have all completed, from pool
2020-03-26 23:37:38 INFO  DAGScheduler:54 - ResultStage 101 (lookup at PageRankMod.scala:70) finished in 0.051 s
2020-03-26 23:37:38 INFO  DAGScheduler:54 - Job 11 finished: lookup at PageRankMod.scala:70, took 0.057113 s
```

This would imply that reduceByKey operation is triggered by job which is of ID 11.

4.

A) It is clear from the lineage that Spark does reuse RDDs for an earlier action. For instance, the same shuffleRDD operation is being used many times throughout the lineage, and this means there is caching involved for the reuse.

B) Cache would be limited by the amount of in-memory data can be held. Persist would leverage external data sources/memory to store data.

* PageRank in MapReduce

1. Pseudocode

// The graph pseudocode stays the same as the scale program wherein
// a file is generated and used as input for the PageRank program.

```
var graph: = List.empty
for (vertex <- 1 to k*k) {
  if (vertex % args != 0) {
   // Setting the neighbor of vertex to be vertex +1
    graph = graph :+ (vertex, vertex + 1)
  }
  else {
   // This would be for dangling vertices.
    graph = graph :+ (vertex, 0)
  }
}
```

```
// This is the loop for the iteration wherein each output is taken as input for the subsequent iteration.

// Before running the iteration each of the vertex is assigned a page rank of 1/k*k

for (int i = 1; i <= iterations; i++) {
        if (i==1){
                prevIterationOutput = initialIterationInput;
                iterationOutput = iterations+i;
        }
        else{
                prevIterationOutput = iterations + (i - 1);
                iterationOutput = iterations + i;
        }

        run(new Iterations(), new String[]{prevIterationOutput, iterationOutput});
}
```

```
//The pseudocode for the iteration would consider dangling edges as the rank would be stored for each
// edge. The global counter is used and reset in each iteration
```

Mapper

```
// Retrieving the list of neighbours and page rank of a vertex which would have been the input
src, neighbours, rank = input

// Rank distributed among all the neighbours
updated_rank = (rank)/(neighbors.count);

// Updating the ranks of the neighbour split by the number of neighbours
for(String s:neighbours) {
        if (!s.isEmpty()) {
                neighbour.set(s);
                neighbour_rank.set(updated_rank.toString());
        }
}
```

Reduce

```
// The reducer would receive vertices


// Check if the vertex encountered is the single dummy vertex
if(vertex is dangling key){
        if(key is single dummy vertex){
                increase dangling_counter
        }
        else{
                update page_rank
        }
}


//Otherwise we would calculate the page rank.
if(vertex Is not dangling key)
        //Calculating page rank for the
        float PR = (alpha*sumPR) + ((1-alpha)/(k*k)) + (dangling_counter/k*k)
}
```

The pseudocode explains how the dangling page problem would be handled wherein we would have a check and do the above according to the appropriate conditions.

The program creates at least 20 Map tasks, by using NLineInputFormat.

```
job2.setInputFormatClass(NLineInputFormat.class);
NLineInputFormat.addInputPath(job2, new Path(intermeOut));
job2.getConfiguration().setInt("mapreduce.input.lineinputformat.linespermap", 50000);
```

MapReduce

Large Cluster : 14 min 43 seconds
Small Cluster: 15 min 44 seconds