BIG DATA

Hands On Session - 3

SPARK

K V Subramaniam

Usha Devi B G

Dept of Computer Science and Engineering

BIG DATA

SPARK Hands-on

SPECIFICATIONS

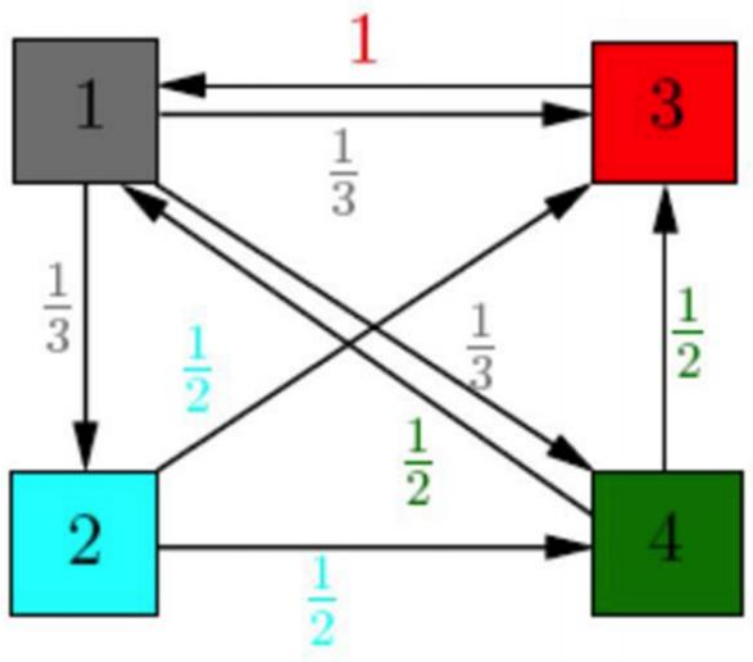
1. Hadoop: 3.2
2. Java: 1.8
3. Apache Spark 3.0
4. Dataset: Please download the dataset from the forum.

BIG DATA

Problem Statement

- Find the ranks of the 4 pages whose links have been given in the input file after 5 iterations of PageRank

Node/ Page	Edges/H yperlinks
1	3
1	2
1	4
2	3
2	4
3	1
4	3
4	1



BIG DATA

Code snippet

- lines = textfile (“urls.txt”))
- links = lines.map (lambda urls: urls.split()).groupByKey().cache()
- ranks = links.map(lambda url_neighbors: (url_neighbors[0], 1.0))
- for iteration in range(MAXITER)):
- contribs = links.join(ranks).flatMap(lambda url_neighbors_rank: computeContribs

  (url_neighbors_rank)
- ranks = contribs.reduceByKey(add).mapValues (lambda rank: rank * 0.85 + 0.15)

Node/ Page	Edges/H yperlinks
1	3
1	2
1	4
2	3
2	4
3	1
4	3
4	1

1. Start each page with a rank of 1
2. On each iteration, update each page’s rank to

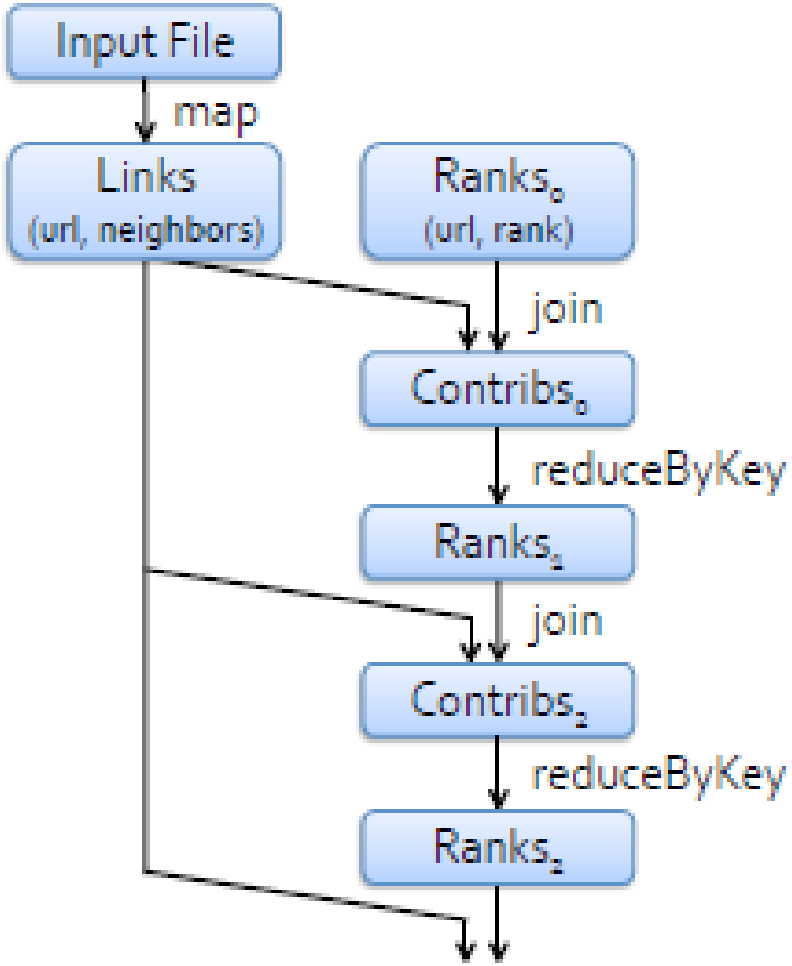
$$\Sigma_{i \in neighbors} rank_i / |neighbors_i|$$

DAG representation

- The Spark Driver will first convert this program into a DAG representation
- What does the DAG representation contain?
  - Each RDD is a node in the graph and
  - all transformations/actions on the RDD as edges

BIG DATA

DAG representation of Page Rank




```
lines = textfile ("urls.txt"))

links = lines.map (lambda urls:
urls.split()).groupByKey().cache()

ranks = links.map(lambda url_neighbors:
(url_neighbors[0], 1.0))

 for iteration in range(MAXITER)):

        contribs = links.join(ranks).flatMap(
lambda url_neighbors_rank:
computeContribs
(url_neighbors_rank)

        ranks =
contribs.reduceByKey(add).mapValues(la
mbda rank: rank * 0.85 + 0.15)
```

Steps to run PySpark Program

$ cd spark_dir

Load input data to HDFS

$ bin/spark-submit <path_to_file.py> <program_parameters>

$ bin/spark-submit pagerank.py <path/on/HDFS> 5

Problem Statement

- Update the given code to not accept number of iterations as a parameter.
- Your code should run till convergence with precision of 5 decimal digits (0.0001). Also print out the number of iterations it runs for.

THANK YOU

K V Subramaniam
Usha Devi B G

Department of Computer Science and Engineering