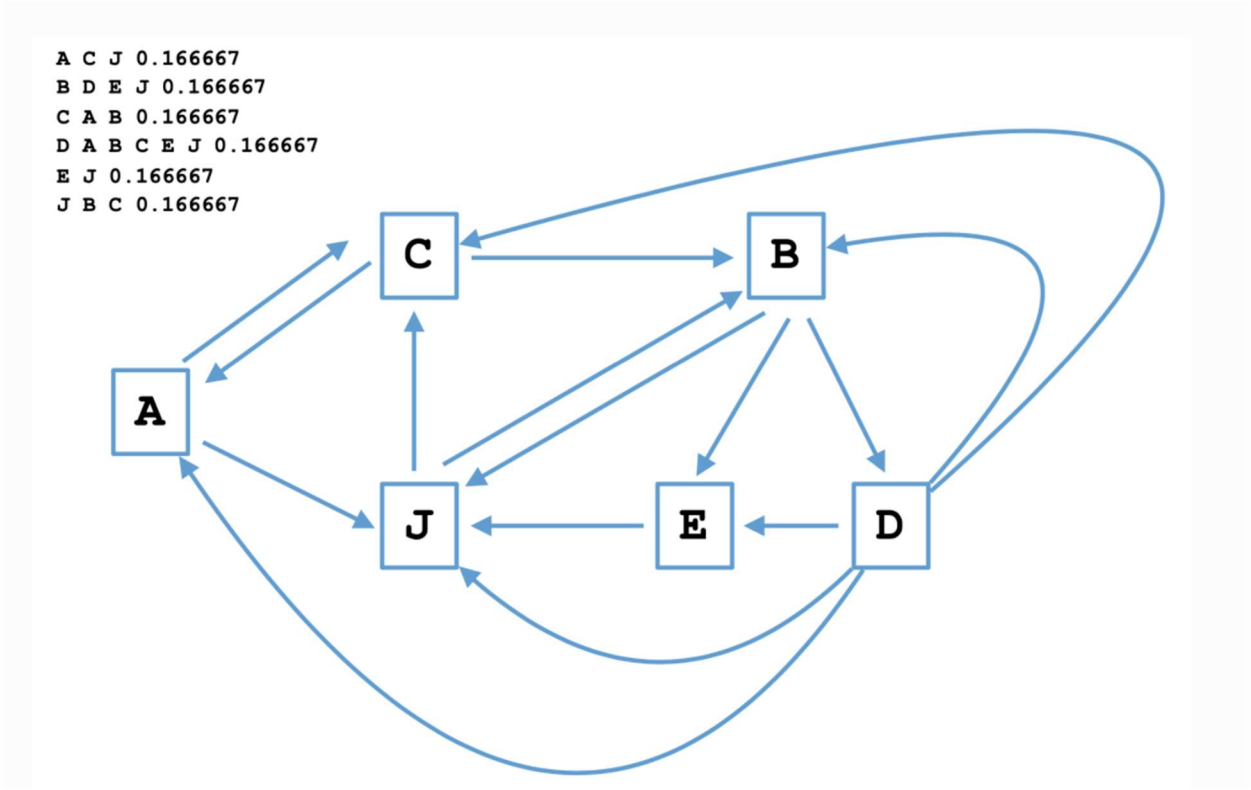
# Lab 3 Report:

Name: Rajan Shantanu Chaturvedi

NetId: rsc9044

## Objective:

The objective of the MapReduce Program is to calculate the Page rank of the following nodes:




## Procedure to achieve the Objective:

Below procedure is used to complete the objective of finding the pagerank of the given nodes:

- Writing the Mapper
- Writing the Reducer
- Writing the Driver
- Compiling the Java files
- Creation of the Jar
- Transfer the input file to HDFS.
- Run the MapReduce job

Writing the Mapper:

Following mapper is written which is splitting the input lines. Using indexes of splited inputs Key and Intermediate value is being output in the form Key(C) Value (A, PR/2).

```
import java.io.IOException;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce. Mapper;
import javax.naming.Context;

public class PageRankMapper
        extends Mapper<LongWritable, Text, Text, Text> {

    @Override
    public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        String line = value.toString();
        String[] splits = line.split(" ");
        String source = splits[0];
        int connections = splits.length - 2;
        double start_pr = Double.parseDouble(splits[splits.length - 1]);
        double end_pr = start_pr / connections;
        StringBuilder Nodes = new StringBuilder();
        for(int i = 1; i < splits.length - 1; i++) {
            String end_val = source + "," + String.valueOf(end_pr);
            System.out.println(splits[i] + " " + end_val);
            Nodes.append(splits[i] + " ");
            context.write(new Text(splits[i]), new Text(end_val));
        }
        System.out.println(Nodes.toString());
        context.write(new Text(source), new Text(Nodes.toString().trim()));
    }
}
```

Writing the Reducer:

This reducer is taking the input (Output of the Mapper job) and summing all the intermediate Pageranks of the Key and output the final result which is depicting the page ranks of all the nodes.

```
import java.io.IOException;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import javax.naming.Context;

public class PageRankReducer
        extends Reducer<Text, Text, NullWritable, Text> {

    @Override
    public void reduce(Text key, Iterable<Text> values, Context context)
            throws IOException, InterruptedException {
        double pr_val = 0.0;
        String t = "";
        for(Text value: values) {
            String line = value.toString();

            if(!line.contains(",")) {
                t = key + " " + line;
                continue;
            }
            String[] splits = line.split(",");
            double pr = Double.parseDouble(splits[1]);
            pr_val += pr;
        }
        context.write(NullWritable.get(), new Text(t + " " + pr_val + ""));
    }
}
```

## Writing the Driver:

Below mentioned driver code is mainly calling the Mapper and Reducer functions 3 times to give us the final output.

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;


public class PageRankDriver {

    public static void main(String[] args) throws Exception {
        if (args.length != 2) {
            System.err.println("Usage: PageRank <input path> <output path>");
            System.exit(-1);
        }
        String input = args[0];
        String output = args[1];
        for(int i = 0; i < 3; i++) {
            Job job = Job.getInstance();
            job.setJarByClass(PageRankDriver.class);
            job.setJobName("Page Rank" + i + "");
            job.setNumReduceTasks(1);

            FileInputFormat.addInputPath(job, new Path(input));
            FileOutputFormat.setOutputPath(job, new Path(output));
            input = output + "part-r-00000";
            output = output + i + "";

            job.setMapperClass(PageRankMapper.class);
            job.setReducerClass(PageRankReducer.class);

            job.setOutputKeyClass(Text.class);
            job.setOutputValueClass(Text.class);
            System.exit(job.waitForCompletion(true) ? 0 : 1);

        }
    }
}
```

## Compiling the Java files:

After Mapper, Reducer and Driver are written, we compiled the java files in their respective classes.

```
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ javac -classpath `hadoop classpath` PageRankMapper.java
[rsc9044@hlog-2 pagerank]$ javac -classpath `hadoop classpath` PageRankReducer.java
[rsc9044@hlog-2 pagerank]$ javac -classpath `hadoop classpath`:. PageRankDriver.java
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ ls -ltrh
total 3.0K
-rwxrwxrwx 1 rsc9044 rsc9044 1.3K Oct 10 21:35 PageRankDriver.java
-rwxrwxrwx 1 rsc9044 rsc9044 1.2K Oct 10 21:52 PageRankMapper.java
-rwxrwxrwx 1 rsc9044 rsc9044  877 Oct 10 21:52 PageRankReducer.java
-rw-rw-r-- 1 rsc9044 rsc9044 2.2K Oct 10 21:54 PageRankMapper.class
-rw-rw-r-- 1 rsc9044 rsc9044 2.2K Oct 10 21:54 PageRankReducer.class
-rw-rw-r-- 1 rsc9044 rsc9044 1.8K Oct 10 21:54 PageRankDriver.class
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
```

## Creation of the Jar:

Jar file is created after all the Java files are compiled.

```
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ jar cvf PageRank.jar *.class
added manifest
adding: PageRankDriver.class(in = 1806) (out= 1007)(deflated 44%)
adding: PageRankMapper.class(in = 2174) (out= 955)(deflated 56%)
adding: PageRankReducer.class(in = 2246) (out= 962)(deflated 57%)
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ ls -ltrh
total 3.5K
-rwxrwxrwx 1 rsc9044 rsc9044 1.3K Oct 10 21:35 PageRankDriver.java
-rwxrwxrwx 1 rsc9044 rsc9044 1.2K Oct 10 21:52 PageRankMapper.java
-rwxrwxrwx 1 rsc9044 rsc9044  877 Oct 10 21:52 PageRankReducer.java
-rw-rw-r-- 1 rsc9044 rsc9044 2.2K Oct 10 21:54 PageRankMapper.class
-rw-rw-r-- 1 rsc9044 rsc9044 2.2K Oct 10 21:54 PageRankReducer.class
-rw-rw-r-- 1 rsc9044 rsc9044 1.8K Oct 10 21:54 PageRankDriver.class
-rw-rw-r-- 1 rsc9044 rsc9044 3.6K Oct 10 21:55 PageRank.jar
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
```

## Transfer the input file to HDFS:

Then Input file is transferred to HDFS for processing.

```
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ hadoop fs -ls /user/rsc9044/pagerank/
Found 2 items
-rw-rw----+  3 rsc9044 rsc9044         96 2021-10-10 22:01 /user/rsc9044/pagerank/input.txt
drwxrwx---+  - rsc9044 rsc9044          0 2021-10-10 21:59 /user/rsc9044/pagerank/output
[rsc9044@hlog-2 pagerank]$
```

## Run the MapReduce job:

Finally, we run the Map reduce job and as the status we can see in the log that Job is successful.

```
[rsc9044@hlog-2 pagerank]$ hadoop jar PageRank.jar PageRankDriver /user/rsc9044/pagerank/input.txt /user/rsc9044/pagerank/output
WARNING: Use "yarn jar" to launch YARN applications.
21/10/10 22:05:58 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
21/10/10 22:05:58 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunne
r to remedy this.
21/10/10 22:05:58 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/rsc9044/.staging/job_1622566668497_9996
21/10/10 22:05:59 INFO input.FileInputFormat: Total input files to process : 1
21/10/10 22:05:59 INFO mapreduce.JobSubmitter: number of splits:1
21/10/10 22:05:59 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
21/10/10 22:05:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622566668497_9996
21/10/10 22:05:59 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/10/10 22:05:59 INFO conf.Configuration: resource-types.xml not found
21/10/10 22:05:59 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/10/10 22:05:59 INFO impl.YarnClientImpl: Submitted application application_1622566668497_9996
21/10/10 22:05:59 INFO mapreduce.Job: The url to track the job: http://horton.hpc.nyu.edu:8088/proxy/application_1622566668497_9996/
21/10/10 22:05:59 INFO mapreduce.Job: Running job: job_1622566668497_9996
21/10/10 22:06:04 INFO mapreduce.Job: Job job_1622566668497_9996 running in uber mode : false
21/10/10 22:06:04 INFO mapreduce.Job:  map 0% reduce 0%
21/10/10 22:06:08 INFO mapreduce.Job:  map 100% reduce 0%
21/10/10 22:06:13 INFO mapreduce.Job:  map 100% reduce 100%
21/10/10 22:06:13 INFO mapreduce.Job: Job job_1622566668497_9996 completed successfully
21/10/10 22:06:13 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=203
                FILE: Number of bytes written=442585
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=223
                HDFS: Number of bytes written=151
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7788
                Total time spent by all reduces in occupied slots (ms)=13338
                Total time spent by all map tasks (ms)=1947
                Total time spent by all reduce tasks (ms)=2223
                Total vcore-milliseconds taken by all map tasks=1947
                Total vcore-milliseconds taken by all reduce tasks=2223
                Total megabyte-milliseconds taken by all map tasks=7974912
                Total megabyte-milliseconds taken by all reduce tasks=13658112
```

```
        File System Counters
                FILE: Number of bytes read=203
                FILE: Number of bytes written=442585
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=223
                HDFS: Number of bytes written=151
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7788
                Total time spent by all reduces in occupied slots (ms)=13338
                Total time spent by all map tasks (ms)=1947
                Total time spent by all reduce tasks (ms)=2223
                Total vcore-milliseconds taken by all map tasks=1947
                Total vcore-milliseconds taken by all reduce tasks=2223
                Total megabyte-milliseconds taken by all map tasks=7974912
                Total megabyte-milliseconds taken by all reduce tasks=13658112
        Map-Reduce Framework
                Map input records=6
                Map output records=21
                Map output bytes=281
                Map output materialized bytes=199
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=199
                Reduce input records=21
                Reduce output records=6
                Spilled Records=42
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=74
                CPU time spent (ms)=1300
                Physical memory (bytes) snapshot=1004912640
                Virtual memory (bytes) snapshot=7438946304
                Total committed heap usage (bytes)=2362441728
                Peak Map Physical memory (bytes)=636297216
                Peak Map Virtual memory (bytes)=3712348160
                Peak Reduce Physical memory (bytes)=368615424
                Peak Reduce Virtual memory (bytes)=3726598144
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=96
        File Output Format Counters
                Bytes Written=151
```

## Output:

The below final output displays the page ranks of all nodes after calling MapReduce job 3 times.

```
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$ hadoop fs -cat /user/rsc9044/pagerank/output/*
A C J 0.1166669
B D E J 0.20000040000000002
C A B 0.20000040000000002
D A B C E J 0.05555566666666667
E J 0.08888906666666667
J B C 0.3388895666666667
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
[rsc9044@hlog-2 pagerank]$
```