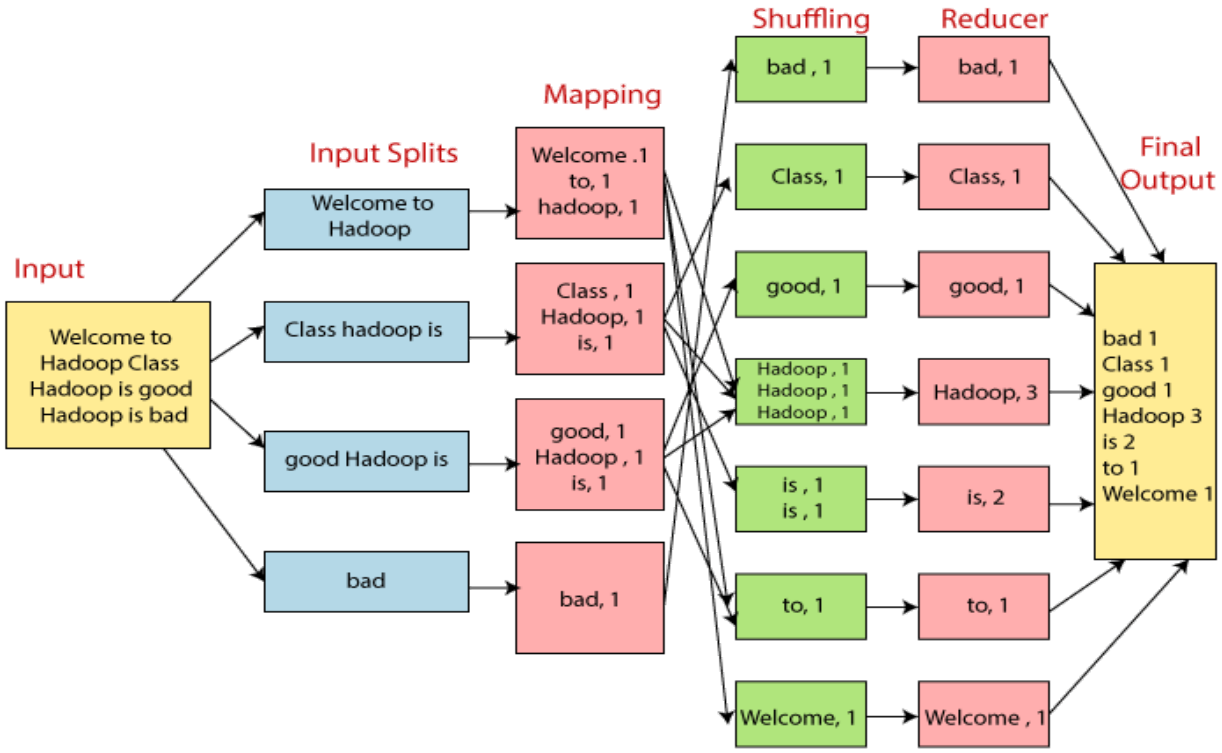
Big Data - 1η ομαδική εργασία

Ομάδα
Σωτηρίου Σωτήριος, dai19052
Κακαλές Δημήτριος, dai19062

Αλγόριθμοι

Το σχήμα που ακολουθεί η διαδικασία είναι το εξής:



Mapper

Ο mapper παίρνει το αρχείο εισόδου και το χωρίζει σε γραμμές. Μετά ξεχωρίζει τα πεδία, κρατάει ως κλειδί (key) το <ip-filename> (το filename σχηματίζεται όπως λέει η εκφώνηση) και ως τιμή (value) την ημερομηνια <date> για κάθε εγγραφή. Το output του mapper έχει τη μορφή <ip-filename, date>.

Reducer

Ο reducer παίρνει ως είσοδο απο τον mapper τα στοιχεία με τη μορφή <ip-filename,[date1, date2, …]>, προσθέτει όλες τις τιμές date σε ένα HashSet για να κρατήσει όλες τις διαφορετικές ημερομηνίες από μια φορά και δίνει ως output τα ζεύγη <ip, filename> για τα οποία το μέγεθος του HashSet είναι μεγαλύτερο του 1. Δηλαδή τα ζεύγη <ip, filename> για τα οποία το set περιέχει πάνω από μια ημερομηνία.

Παράδειγμα εκτέλεσης

Δεδομένα:

ip,date,time,zone,cik,accession,extention,code,size,idx,norefer,noagent,find,crawler,browser
101.81.133.jja,2017-06-26,00:00:00,0.0,1423227.0,0001104659-17-041312,-index.htm,200.0,12352.0,1.0,0.0,0.0,9.0,0.0,
101.81.133.jja,2017-06-26,00:00:00,0.0,1318349.0,0000913760-17-000077,-index.htm,200.0,7713.0,1.0,0.0,0.0,9.0,0.0,
101.81.133.jja,2017-06-27,00:00:00,0.0,1504389.0,0001662252-17-000130,-index.htm,200.0,6649.0,1.0,0.0,0.0,9.0,0.0,
104.247.35.caa,2017-06-26,00:00:00,0.0,1113148.0,0001193125-16-455456,.txt,200.0,50057.0,0.0,0.0,0.0,10.0,0.0,
104.247.35.caa,2017-06-26,00:00:00,0.0,1113148.0,0000769993-16-001076,.txt,200.0,22615.0,0.0,0.0,0.0,10.0,0.0,
104.247.35.caa,2017-06-26,00:00:00,0.0,1113148.0,0000215457-16-001570,.txt,200.0,14159.0,0.0,0.0,0.0,10.0,0.0,
107.22.225.dea,2017-06-26,00:00:00,0.0,1617406.0,0001193125-17-160961,d364947dex81.htm,200.0,32915.0,0.0,0.0,0.0,10.0,0.0,
107.23.85.jfd,2017-06-26,00:00:00,0.0,1171040.0,0001193125-14-194534,-index.htm,200.0,2828.0,1.0,0.0,0.0,10.0,0.0,
107.23.85.jfd,2017-06-26,00:00:00,0.0,1171040.0,0001193125-14-309551,-index.htm,200.0,2803.0,1.0,0.0,0.0,10.0,0.0,
107.23.85.jfd,2017-06-27,00:00:00,0.0,1171040.0,0001193125-14-309551,-index.htm,200.0,2803.0,1.0,0.0,0.0,10.0,0.0,

Εξοδος mapper:

101.81.133.jja -index.htm, 2017-06-26
101.81.133.jja -index.htm, 2017-06-26
101.81.133.jja -index.htm, 2017-06-27
104.247.35.caa 0001193125-16-455456.txt, 2017-06-26
104.247.35.caa 0000769993-16-001076.txt, 2017-06-26
104.247.35.caa 0000215457-16-001570.txt, 2017-06-26
107.22.225.dea d364947dex81.htm, 2017-06-26
107.23.85.jfd -index.htm, 2017-06-26
107.23.85.jfd -index.htm, 2017-06-26
107.23.85.jfd -index.htm, 2017-06-27

Είσοδος Reducer:

101.81.133.jja -index.htm, 2017-06-26, 2017-06-26, 2017-06-27
104.247.35.caa 0001193125-16-455456.txt, 2017-06-26
104.247.35.caa 0000769993-16-001076.txt, 2017-06-26
104.247.35.caa 0000215457-16-001570.txt, 2017-06-26
107.22.225.dea d364947dex81.htm, 2017-06-26
107.23.85.jfd -index.htm, 2017-06-26, 2017-06-26, 2017-06-27

Έξοδος reducer:

101.81.133.jja, -index.htm
107.23.85.jfd, -index.htm

Η εκτέλεση του προγράμματος γίνεται με την παρακάτω εντολή:

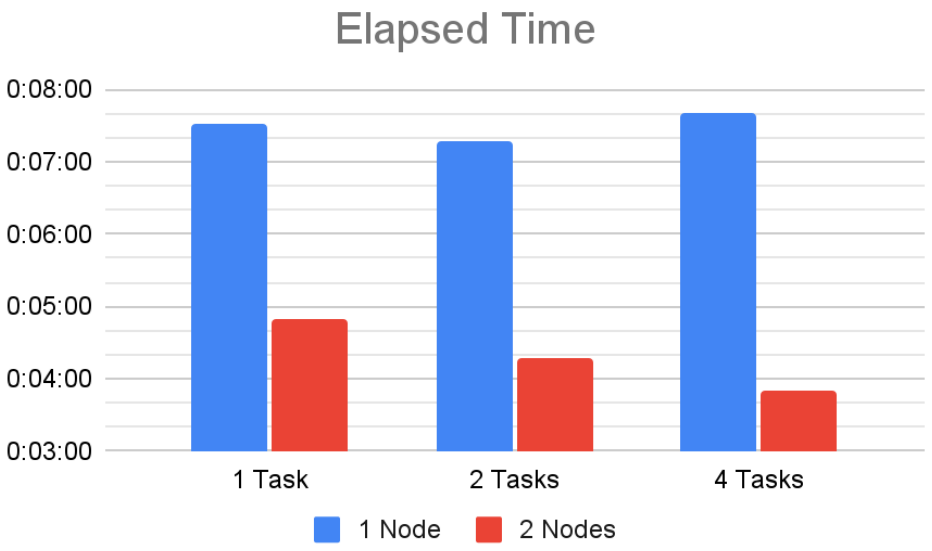
```
~/hadoop/bin/hadoop jar logs.jar LogsMR <input_folder> <output_folder> <numReduceTasks>
```

Μετρήσεις

	1 Node		
	Elapsed Time	Average Map Time	Average Reduce Time
1 Task	0:07:32	0:00:15	0:01:26
2 Tasks	0:07:18	0:00:14	0:00:46
4 Tasks	0:07:40	0:00:13	0:00:24

	2 Nodes		
	Elapsed Time	Average Map Time	Average Reduce Time
1 Task	0:04:50	0:00:19	0:01:23
2 Tasks	0:04:18	0:00:19	0:00:44
4 Tasks	0:03:51	0:00:17	0:00:23

Παρατηρείται στις μετρήσεις μας πως η αύξηση του αριθμού των κόμβων από 1 σε 2 προκαλεί μεγάλη μείωση στο συνολικό χρόνο που τρέχει η διαδικασία map-reduce.



Επίσης μια ακόμη παρατήρηση είναι πως όσο αυξάνουμε τα reduce tasks ο μέσος χρόνος του reduce μειώνεται.