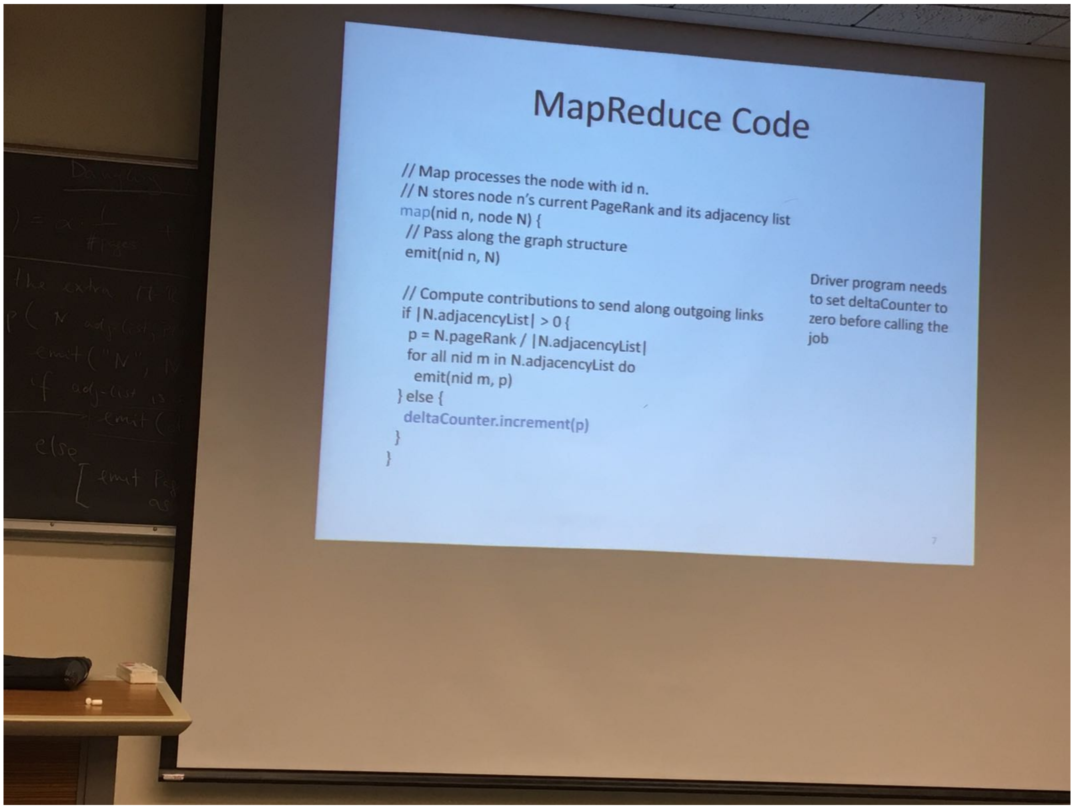
Design Discussion

I used the same Bz2parser.java file and converted the code in it to a MapReduce program (Parsing code in the Mapper and Identity reducer). The output of the file is in the following format

Pagename1:Adjacencylist1
Pagename2:Adjacencylist2
.
.
PagenameN:AdjacencylistN

I have used fairly the same algorithm given in the Prof. Mirek's slides. After pre-processing is done, the file is read in the PageRankmapper and are assigned the initial pagerank as 1/numOfNodesInGraph.
To identify if the Mapper emits an adjanceny list or pagerank, I emit a node with null adjacency list when pagerank has to be emitted and similary I emit 0.0 pagerank when adjacency list has to be emitted for a particular nodeID.



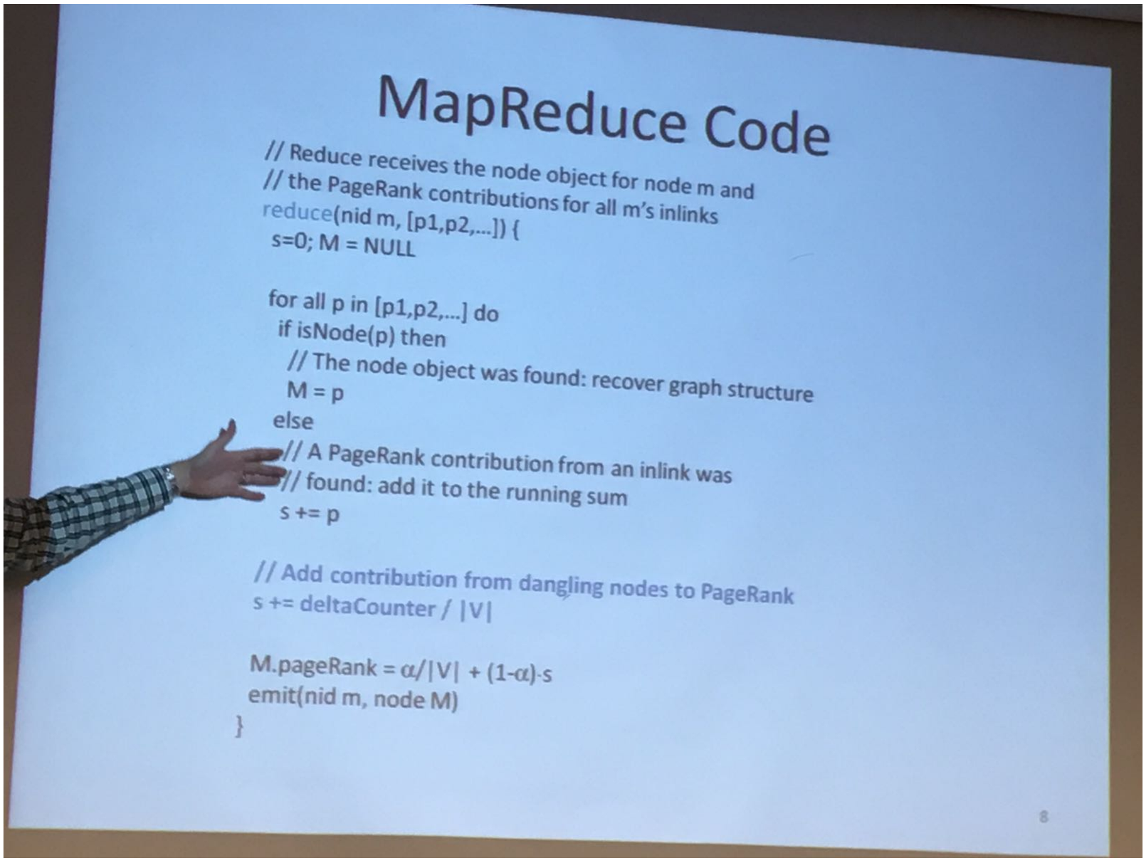
MapReduce Code

```
// Map processes the node with id n.
// N stores node n's current PageRank and its adjacency list
map(nid n, node N) {
  // Pass along the graph structure
  emit(nid n, N)

  // Compute contributions to send along outgoing links
  if |N.adjacencyList| > 0 {
    p = N.pageRank / |N.adjacencyList|
    for all nid m in N.adjacencyList do
      emit(nid m, p)
  } else {
    deltaCounter.increment(p)
  }
}
```

Driver program needs to set deltaCounter to zero before calling the job

7



MapReduce Code

```
// Reduce receives the node object for node m and
// the PageRank contributions for all m's inlinks
reduce(nid m, [p1,p2,...]) {
 s=0; M = NULL

 for all p in [p1,p2,...] do
  if isNode(p) then
   // The node object was found: recover graph structure
   M = p
  else
   // A PageRank contribution from an inlink was
   // found: add it to the running sum
   s += p

 // Add contribution from dangling nodes to PageRank
 s += deltaCounter / |V|

 M.pageRank = α/|V| + (1-α)·s
 emit(nid m, node M)
}
```

8

I have handled the delta counter same way by making a global counter.

For Top-100 calculation my pseudo code is as follows:

```
Map(…,Text value){
        //Read pagerank and NodeID from the value
        Emit(pagerank, NodeID)
}
getPartition(pagerank, NodeID){
        //send everything to one reducer to calculate top 100 globally
        return 0
}
KeyComparator(pagerank, NodeID){
        //Sort in the descending order of key pagerank
}

Redude(pagerank, NodeID){
        //Use a global counter to keep track of 100 records
        While (count < 100)
                Emit (NodeID, pagerank)
                GlobalCounter.increment(1)
}
```

AMOUNT OF DATA TRANSFERRED:

Iteration 1:

Mappers to Reducers: 3491142191
Reducers to HDFS: 1488422905

Iteration 2:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488794201

Iteration 3:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488782321

Iteration 4:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488721200

Iteration 5:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488715040

Iteration 6:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488702069

Iteration 7:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488684713

Iteration 8:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488651387

Iteration 9:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488617567

Iteration 10:

Mappers to Reducers: 3500132524
Reducers to HDFS: 1488587310

1st iteration has the least Mappers to Reducers bytes (3491142191) among all iteration as there was no initial pagerank and I assigned the same 1/nodes value to all the nodes which required less bytes.
After this all the iterations have same Mapper to Reducers bytes because we're emitting the same node data every time.

1st iteration has the least Reducer to HDFS bytes (1488422905) among all iterations because the initial page ranks(which consumed less bytes) got converged for the first time and also in the first iteration dangling nodes contribution was zero so there was missing mass in the output file. In 2nd iteration Reducer to HDFS bytes (1488794201) got increased and then till the 10th interation it kept on decreasing from the 2nd iteration because the pagerank values kept converging to a more stable value and did not require large bytes to store data.
Use of ***long*** data type for store the missing mass as global counter also effected some precision and hence some bytes were lost in each iteration after 2nd iteration.

Performance Comparison

- 6 m4.large machines (1 master and 5 workers)
  (i) pre-processing time – **47 minutes**
  (ii) time to run ten iterations of PageRank – **25 minutes**
  (iii) time to find the top-100 pages – **1 minute**

- 11 m4.large machines (1 master and 10 workers)
  (i) pre-processing time – **22 minutes 30 secs**
  (ii) time to run ten iterations of PageRank – **15 minutes**
  (iii) time to find the top-100 pages- **50 seconds**

The pre- processing phase shows good speed up (almost 2 times) because the number of worker machines also got doubled from 5 workers to 10 workers. Input files were split to more number of mapper workers. For both the runs input split was 106 = Number of Map task which was done by 5 machines in the 1st run and 10 machines in the 2nd run.

Iterations of Pagerank also showed good speed up but comparatively less than the pre-processing phase. In the pre-processing phase data read and write from/to HDFS happens only once so it was not a very intensive job in terms of data transferred through network. Iterations of pagerank was very intensive in terms of data read/write from HDFS (data transfer through network) as every job wrote ~1488422905 bytes to HDFS and the next job read the same amount of files. So the data transferred via network in the entire duration is ~1488422905*10.

Top100 pagerank calculation hardly showed any speedup as the processing happened in one reducer in both the runs. Network data transfer was also same in both the jobs.

Top 100 values from local run:

```
United_States_09d4:0.0454852507732353
Week:0.03772705069781671
Sunday:0.029203165906032547
Monday:0.028553038474976013
Wednesday:0.02825025222869564
Friday:0.02741482898377306
Saturday:0.027087909272236307
Day:0.02680241129328086
Thursday:0.02663301041406329
Tuesday:0.02647985715913479
Country:0.025347654552460554
Wikimedia_Commons_7b57:0.025030209342995477
Europe:0.01831436627025558
United_Kingdom_5ad7:0.01726088026809777
Earth:0.016532332974908595
France:0.014063151455722319
Water:0.013990694074367606
Germany:0.012923231029418638
Asia:0.012715042483263852
England:0.012657010570046909
City:0.012343067541579591
Animal:0.011625011036475933
Sun:0.011349370183746108
Year:0.01115665332330853
English_language:0.010747789338969041
Money:0.010391974369820988
Government:0.010244875782399979
Italy:0.010184014406384702
Number:0.010162091446642676
index:0.01001340626273539
India:0.009831644364015686
Canada:0.008722623886963899
Wiktionary:0.008586675916110575
Spain:0.008551767401417122
Plant:0.008538664323993882
Planet:0.008314112717676667
People:0.008269425955768829
Computer:0.007942147936913066
Japan:0.007920412477290431
Wikimedia_Foundation_83d9:0.007767705377697868
China:0.007633657798553193
Moon:0.007525919978151609
Australia:0.007486019926958532
Energy:0.00748433290420455
```

```
Russia:0.007260311288695717
Human:0.007199562564065275
Thor:0.007184595619813111
State:0.007095707402616285
Science:0.006838893673541622
20th_century:0.006781715769856163
Capital_(city):0.006651064831413421
19th_century:0.006442635008879791
Geography:0.006399262623998236
God:0.0063775301228415696
Greece:0.006312461093133384
Africa:0.006276024113201375
Greek_language:0.006231107226943432
Religion:0.006200854234929282
Mathematics:0.006191403199456686
Scotland:0.006099852514683112
Food:0.006053776661276938
2004:0.005984866932524986
February:0.005840806964818388
Language:0.005811990915690682
Poland:0.005751616796439403
Wikipedia:0.005734922742142722
Society:0.0057320667091768805
Sweden:0.005632091057134353
January:0.005608437803081466
World:0.0055824994171650285
Turkey:0.00555244722320288
History:0.0055453986994576454
Centuries:0.005530586076582796
Cyprus:0.0054937006824995775
Television:0.005463504916069171
Culture:0.005435489451664863
Law:0.005388481727693353
Odin:0.005381232839626957
Sound:0.005340248109379373
March:0.005314159609413709
Latin:0.005286269388816302
Month:0.005225667256954458
London:0.005204979562437963
Music:0.005189592129546086
War:0.005168338166881445
List_of_decades:0.005078875341472842
Denmark:0.005063828554266997
Portugal:0.00500231667778429
Greek_mythology:0.004976419222982018
Metal:0.004966963218853904
Plural:0.004916565541951172
Austria:0.004860750120543997
```

```
Scientist:0.004813227495006541
Liquid:0.004775088070909461
April:0.00476278120169472
Netherlands:0.004757545211132822
Light:0.0047563468275417554
Norse_mythology:0.004715423222849549
Information:0.0047076801314134735
Atom:0.004584355253553932
```

The order of page ranks seems fine to me. It consists of mostly country names and important nouns which can be important searched about pages so it can have better page rank than the others. Also this data is the Wikipedia data dump of 2006, it is the page with highest page rank. This seems to be the obvious important/searched upon page on internet. Also there are countries of world importance or social issues of that time which take the higher page ranks.

Top 100 values from EMR run

```
2006:0.005129149449041743
United_States_09d4:0.004492913048100462
United_Kingdom_5ad7:0.002518389237467937
2005:0.0022707837004375146
France:0.0018790856977073558
2004:0.001624343156990327
Germany:0.0015111180070295563
England:0.00147740430672144
Italy:0.00142744191736116
Canada:0.001333442820097365
2003:0.001242007682310821
Australia:0.0011528916758065868
Japan:0.001132273370032307
index:0.0011179708189963317
English_language:0.001087251938677357
India:0.0010807092676739548
Europe:0.0010184917985445016
World_War_II_d045:9.979047327229335E-4
2002:9.933718905106507E-4
Wikimedia_Commons_7b57:9.578856904930968E-4
2001:9.470229973092762E-4
Russia:9.4193670036599E-4
London:9.415947679397907E-4
Wiktionary:9.340206105601811E-4
Spain:9.322447750533802E-4
Biography:8.528673661294494E-4
2000:8.436482257689477E-4
1999:8.372676240584725E-4
Internet_Movie_Database_7ea7:7.385166283045563E-4
1998:7.147292341051591E-4
1997:6.969437116915292E-4
Latin:6.849769491685031E-4
```

```
Sexagenary_cycle:6.739044859623275E-4
January_1:6.720499510744478E-4
Netherlands:6.602002739536533E-4
China:6.565190874904540E-4
New_York_City_1428:6.494292875183218E-4
1996:6.451349938984680E-4
Scotland:6.354585022042599E-4
French_language:6.278248807458337E-4
1995:6.214962567718055E-4
Geographic_coordinate_system:6.137256908925661E-4
Sweden:6.114641967777905E-4
1991:6.047570171259578E-4
Gregorian_calendar:5.9885430177269E-4
1994:5.983139546354373E-4
Soviet_Union_ad1f:5.87199221659451E-4
1990:5.741665610645349E-4
1993:5.638555436874276E-4
1992:5.502804642970164E-4
Egypt:5.465092619963279E-4
1945:5.418940185793729E-4
International_Phonetic_Alphabet_96f8:5.398502298942869E-4
Greek_language:5.349913456227379E-4
1980:5.341064900117803E-4
1989:5.304447533131564E-4
Public_domain:5.297973987792992E-4
New_Zealand_2311:5.204552183295226E-4
1979:5.187426918130323E-4
Poland:5.165131290618411E-4
1974:5.148772970554488E-4
Television:5.148403204539094E-4
1986:5.14811560739523E-4
Paris:5.14122154711963E-4
1970:5.133354288759246E-4
1981:5.047591183641088E-4
1976:5.045023036319759E-4
European_Union_e368:5.033218268196827E-4
1969:5.005275839081883E-4
1975:5.004875507592081E-4
1982:4.986493708048685E-4
1985:4.940912247572052E-4
Greece:4.906916913891982E-4
1972:4.888869125985893E-4
Portugal:4.8683615611325436E-4
Austria:4.8605438723816876E-4
German_language:4.8470195003976147E-4
Switzerland:4.8448502853132617E-4
1984:4.8110635941921544E-4
Ireland:4.7840266995451895E-4
```

```
1971:4.779233883504469E-4
1973:4.7783779512257265E-4
1983:4.766190530592744E-4
1977:4.74645548461477E-4
1968:4.6936937990713453E-4
1987:4.684503010801367E-4
19th_century:4.680616296357845E-4
1967:4.660088277689139E-4
1978:4.649199223288549E-4
People's_Republic_of_China_82bf:4.642805359038263E-4
World_War_I_9429:4.626416308239336E-4
1988:4.6012133590514205E-4
Turkey:4.594435033814591E-4
Israel:4.580987216579595E-4
Belgium:4.574951344082336E-4
Mexico:4.5694478205556097E-4
Norway:4.5599934796782687E-4
Denmark:4.532754907724953E-4
South_Africa_1287:4.523927537854544E-4
Football_(soccer):4.51420881632047E-4
```