CS. 6240-02 HW3 Vikas Janardhanan

Design Discussion (20 points total)

Preprocessing step:

- All page names and adjacency list items containing the "~" character is removed at the preprocessing stage
- Dangling nodes are handled: There might be some pages which don't have any outgoing links. Such pages will only appear as a outgoing link from another page. Preprocessing step adds such pages as an output with empty adjacency list.
- Preprocessing steps also keeps track of total number of pages and stores the updated count in a global counter.

Pseudocode:

Preprocess:

// Reads from the .bz file and creates a corresponding pagename, adjacency list representation

```
Map(key k,inputLine)
        pagename=extract_pagename(inputLine)
        html=extract_html(inputline)
        if pagename doesn't contain "~":
                adjList=get_linkedpages(html) //using WikiParser
                emit(pagename,adjList);
                for each linkedpage in adjList:
                        emit(linkedpage,null);

Reduce(Key pagename,adjLists)
        if incoming link present or outgoing links present:
                Increment global counter to keep track of number of pages
        if adjLists is null:
                Emit(pagename,null)
        else
                Emit(pagename,adjList)
```

PageRank computation:

A flag is used to understand if the run is the first one. If it is the initial pagerank is set for all nodes at the mapper level.
Contribution from dangling nodes is accumulated at reducer. The dangling node contribution from run i is updated to global counter-delta. This delta is added to pagerank of all nodes during i+1 Map phase.
TotalPageRank after each iteration is accumulated and printed at the reducer to check for convergence.

```
Mapper(key pagename,PageNode n)
        If(FirstRun)
                newPageRank=1.0/TOTALPAGES; // TOTALPAGES is a global counter
        else
                newPageRank=n.pageRank+alpha*(DELTA/TOTALPAGES) //DELTA is
also global counter updated by the reducer during previous run
        n.pageRank=newPageRank
        emit(pagename,n)
        if(n is danglingnode)
                emit("~~",n)
        else
                for outlink in n.adjList:
                        n.pageRank=newPageRank/adjList.size()
                        emit(outlink,n)

Combiner(key pagename,pageNodes)
        PageNode n
        //Accumulate all dangling node contribution used to calculate delta in
reducer
        If(key == "~~")
                For each pageNode in pageNodes
                        accDelta+=pageNode.pageRank
                n.pageRank=accDelta
                emit(pagename,n);
                return;
        for each pageNode in pageNodes:
                if pageNode.hasadjList:
                        emit(pagename,pageNode)
                else:
```

```
                Contrib+=pageNode.pageRank
            n.pageRank=Contrib
            emit(pagename,n)

Reducer(key pagename,pageNodes)
    PageNode n
    If(key == "~~")
        For each pageNode in pageNodes
            accDelta+=pageNode.pageRank
        Update accDelta to global counter DELTA
        return;
    for each pageNode in pageNodes:
        if pageNode doesn't have adjList:
            contrib+=pageNode.pageRank
        n.pageRank=(alpha/TOTALPAGES)+((1.0-alpha)*contrib))
        emit(pagename,n)
```

Top K algorithm:

```
TopKMapper
    PriorityQueue q //Also create custom comparator to sort based on
pagerank.
Map(key pagename,PageNode n)
    q.add(n)
    If q.size() >100
        q.poll()
cleanup()
    for each page entry in q:
        emit(null,page)

TopKReducer
    PriorityQueue<PageRankNode> q //Also create custom comparator to sort
based on pageRank value. PageRankNode has pagename,pageRank value.
Reduce(NullWritable key,pageRankNode n)
    q.add(n)
    If q.size() >100
        q.poll()
    Print all pageRankNode details in q.
```

CS. 6240-02 HW3 Vikas Janardhanan

Report the amount of data transferred from Mappers to Reducers, and from Reducers to HDFS, in each
iteration of the PageRank computation. Does it change over time? If so, briefly discuss why or why not?

Map input records=3178227
Map output records=71167233
Map output bytes=3597197322
Reduce input groups=3178228
Reduce shuffle bytes=1134720463
Reduce input records=17002949
Reduce output records=3178227
HDFS: Number of bytes read=1477660084
HDFS: Number of bytes written=1477620154

No it doesn't change the same because each iteration of PageRank algorithm works on the same set of records(same pages with same adjacency list). Only the pagerank values are changed in each iteration.

CS. 6240-02 HW3 Vikas Janardhanan

Performance Comparison (15 points total)

Run your program in Elastic MapReduce (EMR) on the four provided bz2 files, which comprise the full
English Wikipedia data set from 2006, using the following two configurations:

- 6 m4.large machines (1 master and 5 workers)
- 11 m4.large machines (1 master and 10 workers)

Report for both configurations (i) pre-processing time, (ii) time to run ten iterations of PageRank, and
(iii) time to find the top-100 pages. There should be 2*3=6 time values. (6 points)

From logs:

6 Machines:

Preprocessing completed in 1600273 ms
Page Rank computation completed in 113923 ms
Page Rank computation completed in 132959 ms
Page Rank computation completed in 127923 ms
Page Rank computation completed in 133218 ms
Page Rank computation completed in 127823 ms
Page Rank computation completed in 121873 ms
Page Rank computation completed in 113808 ms
Page Rank computation completed in 122732 ms
Page Rank computation completed in 131760 ms
Page Rank computation completed in 135824 ms
Top 100 computation completed in 55863 ms

11 Machines :

Preprocessing completed in 1014500 ms
Page Rank computation completed in 89939 ms
Page Rank computation completed in 91847 ms
Page Rank computation completed in 93865 ms
Page Rank computation completed in 95878 ms
Page Rank computation completed in 91866 ms
Page Rank computation completed in 93888 ms
Page Rank computation completed in 95847 ms
Page Rank computation completed in 92703 ms
Page Rank computation completed in 94725 ms
Page Rank computation completed in 91708 ms

CS. 6240-02 HW3 Vikas Janardhanan

Top 100 computation completed in 52722 ms

6 machines:

(i) pre-processing time: **1600273 ms**
(ii) time to run ten iterations of PageRank: **1261843 ms**
(iii) time to find the top-100 pages: **55863 ms**

11 machines:

(i) pre-processing time: **1014500 ms**
(ii) time to run ten iterations of PageRank: **932266** ms
(iii) time to find the top-100 pages: **52722 ms**

Critically evaluate the runtime results by comparing them against what you had expected to see and
discuss your findings. Make sure you address the following question: Which of the computation phases
showed a good speedup? If a phase seems to show fairly poor speedup, briefly discuss possible
reasons—make sure you provide concrete evidence, e.g., numbers from the log file or analytical
arguments based on the algorithm's properties. (4 points)

Here, speedup is being analyzed using the results we got from EMR. So lets take 6 machine time as the serial time taken. Optimal speedup hence would be 2 since we double the number of worked machines from 5 to 10.

We have,
Speedup = time taken in 6 machine setup/time taken in 11 machine setup

Preprocess speedup: 1600273/1014500 = 1.58
PageRank speedup: 1261843/932266 = 1.35
Top 100 Speedup: 1.06

CS. 6240-02 HW3 Vikas Janardhanan

Preprocess and PageRank:
We achieve good amount of parallelism as expected in both these jobs. There are many keys for the mapper and reducer phase in both these jobs and hence we can expect good distribution of work among the worker nodes. So as expected good speedup is achieved after increasing the number of worker nodes.

Top 100:
In case of Top 100, each mapper emits the local top 100 and the global top 100 is computed at a single reducer. Hence in the reduce phase no matter the number of machines available only one of them will be busy where the reducer computes the global top 100 from the local ones. Therefore, in this case, speedup isn't achieved on doubling the number of worker nodes. Parallelism exists in this case only in the mapper phase.

Report the top-100 Wikipedia pages with the highest PageRanks, along with their rank values and sorted
from highest to lowest, for both the simple and full datasets. Do they seem reasonable based on your
intuition about important information on Wikipedia? (5 points)

Yes they seem to be reasonable. Country name and year related pages seems to be amongst the most highly ranked ones. As expected majority of Wikipedia articles would contain references to countries and years. So, our PageRank values are consistent with this intuition.

Machine6:

United_States_09d4:0.002905049211098
2006:0.002601052663040
United_Kingdom_5ad7:0.001381933693276
2005:0.001198785776077
Biography:0.000950231857881
Canada:0.000902948043433
England:0.000897252898679
France:0.000888489576305

2004:0.000834786716495
Germany:0.000763110031137
Australia:0.000738825188152
Geographic_coordinate_system:0.000722554706546
2003:0.000672261798953
India:0.000651145680121
Japan:0.000645777402587
Italy:0.000542419130970
2001:0.000539561851684
2002:0.000533306832033
Internet_Movie_Database_7ea7:0.000527718734330
Europe:0.000514002813042
2000:0.000505009259219
World_War_II_d045:0.000487160678123
London:0.000470074453236
Population_density:0.000452275208234
1999:0.000446418899664
Record_label:0.000446183915032
English_language:0.000443664387416
Spain:0.000443156662935
Russia:0.000418420387229
Race_(United_States_Census)_a07d:0.000415283406358
Wiktionary:0.000408740523612
Wikimedia_Commons_7b57:0.000389606462805
1998:0.000385864293375
Music_genre:0.000375615703614
1997:0.000368118107863
Scotland:0.000362208528710
New_York_City_1428:0.000362021171877
Football_(soccer):0.000352446564407
1996:0.000345343839368
Sweden:0.00034021471471 4
Television:0.000339458759153
Square_mile:0.000327662198420
Census:0.000326717713307
1995:0.000325277961537
California:0.000322410941768

China:0.000318512347458
Netherlands:0.000313758387111
New_Zealand_2311:0.000312482578150
1994:0.000310555561147
1991:0.000296362568883
1993:0.000293635132755
1990:0.000291916513006
New_York_3da4:0.000289886879812
Public_domain:0.000289559817245
1992:0.000281442926449
United_States_Census_Bureau_2c85:0.000279107103114
Film:0.000278708552637
Actor:0.000276463548893
Scientific_classification:0.000275890010209
Norway:0.000273797725988
Ireland:0.000272592027011
Population:0.000270709495661
Poland:0.000270300703246
1989:0.000263930437892
1980:0.000257874272306
January_1:0.000257872304183
Marriage:0.000255737339646
Brazil:0.000255431810114
Latin:0.000254142613717
Mexico:0.000254010007714
Politician:0.000251315058529
1986:0.000250838427877
1985:0.000244661664647
1979:0.000244439542430
French_language:0.000243774554375
1982:0.000243751608508
1981:0.000243553693248
1974:0.000241356433487
Per_capita_income:0.000241205949684
Album:0.000239460974004
Switzerland:0.000239276855584
1984:0.000239006414395

1987:0.000238787242833
South_Africa_1287:0.000238736833875
1983:0.000238701883728
Record_producer:0.000235823714677
1970:0.000235026354548
1988:0.000233317564855
1976:0.000232281228604
1975:0.000229592645040
Km²:0.000229295667798
Paris:0.000226542759783
1969:0.000226441217981
Greece:0.000226040283926
1945:0.000225328429128
1972:0.000224947052252
1977:0.000223085419689
Personal_name:0.000222994194712
Soviet_Union_ad1f:0.000222497758946
1978:0.000221991918240

Machine 11:
United_States_09d4:0.002905032006329
2006:0.002601076649573
United_Kingdom_5ad7:0.001381924062984
2005:0.001198770556559
Biography:0.000950232035161
Canada:0.000902946609550
England:0.000897257267659
France:0.000888500227283
2004:0.000834763436315
Germany:0.000763160669759
Australia:0.000738818201905
Geographic_coordinate_system:0.000722725288700
2003:0.000672244999079
India:0.000651145556541
Japan:0.000645777074573
Italy:0.000542382834957
2001:0.000539560091809

2002:0.000533306438610
Internet_Movie_Database_7ea7:0.000527705682315
Europe:0.000514040822245
2000:0.000505010268098
World_War_II_d045:0.000487161249831
London:0.000470090427904
Population_density:0.000452296840027
1999:0.000446416487551
Record_label:0.000446173252360
English_language:0.000443684693668
Spain:0.000443189041423
Russia:0.000418420842199
Race_(United_States_Census)_a07d:0.000415281053465
Wiktionary:0.000408755453552
Wikimedia_Commons_7b57:0.000389619070516
1998:0.000385867406202
Music_genre:0.000375614670938
1997:0.000368117601535
Scotland:0.000362215204190
New_York_City_1428:0.000362018605227
Football_(soccer):0.000352449597580
1996:0.000345348468695
Sweden:0.000340213229927
Television:0.000339457888664
Square_mile:0.000327665338053
Census:0.000326718057923
1995:0.000325278095442
California:0.000322410990844
China:0.000318508037310
Netherlands:0.000313766166194
New_Zealand_2311:0.000312485381252
1994:0.000310560968765
1991:0.000296372652449
1993:0.000293615205418
1990:0.000291919248116
New_York_3da4:0.000289889335582
Public_domain:0.000289550296507

1992:0.000281438708945
United_States_Census_Bureau_2c85:0.000279105730636
Film:0.000278702873905
Actor:0.000276459971619
Scientific_classification:0.000275889670506
Norway:0.000273898756834
Ireland:0.000272612850648
Population:0.000270725056028
Poland:0.000270336802547
1989:0.000263938335896
1980:0.000257873237556
January_1:0.000257860779852
Marriage:0.000255736025673
Brazil:0.000255457249352
Latin:0.000254143994131
Mexico:0.000254010212922
Politician:0.000251314347330
1986:0.000250838418147
1985:0.000244660491629
1979:0.000244437006268
French_language:0.000243775547458
1982:0.000243752808624
1981:0.000243552526072
1974:0.000241356353959
Per_capita_income:0.000241204517792
Album:0.000239468213543
Switzerland:0.000239262314441
1984:0.000239000443780
1987:0.000238793966442
South_Africa_1287:0.000238737715177
1983:0.000238712317022
Record_producer:0.000235816191851
1970:0.000235026298389
1988:0.000233317490668
1976:0.000232279330600
1975:0.000229593276736
Km²:0.000229303435496

Paris:0.000226542542286
1969:0.000226444508921
Greece:0.000226042703453
1945:0.000225328294457
1972:0.000224957020473
1977:0.000223085854211
Personal_name:0.000222993533765
Soviet_Union_ad1f:0.000222504696273
1978:0.000221989954110

Local:
United_States_09d4:0.006262544724786
Wikimedia_Commons_7b57:0.004747208297026
Country:0.003876365724522
England:0.002671922135423
United_Kingdom_5ad7:0.002601134687522
Europe:0.002596869321453
Water:0.002574354466479
Germany:0.002530449354759
France:0.002505287963670
Animal:0.002448377338789
Earth:0.002415278209120
City:0.002351471502280
Week:0.002001511083086
Asia:0.001914990086483
Sunday:0.001862134569055
Wiktionary:0.001852782958001
Monday:0.001835018059206
Money:0.001829512432229
Wednesday:0.001816810340586
Plant:0.001804491925776
Friday:0.001772335882083
Computer:0.001754541900025
Saturday:0.001752622942500
English_language:0.001740587000786
Thursday:0.001730001771798
Tuesday:0.001717577106620

Italy:0.001708058561767
Government:0.001697483228082
India:0.001695981244459
Number:0.001581564167344
Spain:0.001552759125444
Japan:0.001508667527512
Canada:0.001492041319103
Day:0.001464465924854
People:0.001438858098782
Human:0.001411405823849
Wikimedia_Foundation_83d9:0.001368796692070
Australia:0.001360222584530
China:0.001359788284978
Energy:0.001327318762273
Food:0.001310808958415
Sun:0.001287556346779
Science:0.001284969080235
Mathematics:0.001270144556436
index:0.001241487529925
Television:0.001219794283299
Capital_(city):0.001182880276552
Russia:0.001176045379887
State:0.001157764954977
Music:0.001151391016344
Year:0.001129201697040
Greece:0.001106092512366
Language:0.001102724635201
Scotland:0.001099696262596
Metal:0.001076010630010
Wikipedia:0.001066736475674
Greek_language:0.001055413679842
2004:0.001050808273882
Planet:0.001025016545382
Sound:0.001019759251166
Religion:0.001016537485765
London:0.001014542205818
Africa:0.000985072347359

20th_century:0.000949881084480
Law:0.000943346486874
Geography:0.000937807896327
Liquid:0.000931359006082
19th_century:0.000931027183352
World:0.000918976075697
Poland:0.000917086639133
Scientist:0.000907045227841
Society:0.000904654767803
Latin:0.000872131836493
Atom:0.000872042791711
History:0.000870224608641
Sweden:0.000863183526746
War:0.000862450991110
Light:0.000858963056627
Netherlands:0.000852474036754
Culture:0.000843762357625
Building:0.000834214377449
God:0.000817829464995
Turkey:0.000815952484510
Plural:0.000810584840039
Information:0.000807645701921
Centuries:0.000799830021777
Chemical_element:0.000787135149938
Portugal:0.000785096840683
Inhabitant:0.000781358780596
Denmark:0.000772022540671
Capital_city:0.000769642711050
Austria:0.000765172573502
Cyprus:0.000753189758101
Species:0.000751352208499
Ocean:0.000750281569105
Book:0.000749402361577
Disease:0.000747693955759
North_America_e7c4:0.000747152285160
University:0.000744790131240
Biology:0.000742444961229