Design Discussion

The pre-processing algorithm follows the steps as specified in the problem statement. We have chosen the parser that has been provided along with the problem. In the pre-processing algorithm, we use the parser to find the node-name and adjacency list. While finding node-name and adjacency list, we ignore all the links that have the symbol '~' in the link. We also remove the path name and .html suffix thereby retaining only the page name.

Pseudo-Code for Pre-Processing

The pseudo-code to pre-process the Wikipedia dump and obtain the graph is as follows

```
method map(Key k, value v):
    name ← Parse v to obtain page-name
    adjacencyList ← Parse v to obtain adjacencyList

    emit(name, adjacencyList)

    for each record r in adjacencyList:
        emit(r, null);
end

method reduce(pageName p, adjacencyList [l1, l2, l3, …]):
    pageCountCounter ← new Global Counter
    if ( isInlinksPresent() or isOutLinksPresent()):
        pageCountCounter ← pageCountCounter + 1
    if(adjacencyList is null):
        emit(p, new adjacencyList[])
        return
    emit(p, adjacencyList)
```

Pseudo Code for Page Rank

The Page rank algorithm is modified to handle the dangling node adjustment. The dangling node adjustment is done in map function and is updated through global counter.

```
method map(Key k , Value v):
      alpha ← Initialize alpha component of page rank
      pageCount ← Extract pageCount from counters
     delta ← get delta from counters
      emit(k, v)

      newPageRank ← v.pageRank + (1-alpha) * delta/pageCount

      for each entry e in v.adjacencyList :
            emit(e, newPageRank/sizeOf(adjacencyList)

     if(v.adjacencyList is empty):
       emit(dummy, newPageRank)
END

method reduce(Node n , List[dummy, c1,c2,c3]):
     if(n is dummy):
          for each contribution 'c' in List :
                totalContributions ← totalContributions + c
          update delta gobal counter with totalContribution
          return
   Node n1 ← Initialize new node
   for each entry e in List:
            if(e is adjacencyList):
                 n1.adjacencylist ← e
            else:
              totalContributions ← totalContributions + e

  n1.pageRank = totalContributions
  emit(n,n1)
END
```

Pseudo-Code for Top-K records

The pseudo code to find top-k records is same as the pseudo-code provided in the module. We use Tree-Map data structure to find top-K records.

```
Class Mapper {
 localTopK

 setup() {
  initialize localTopK
 }

 map(…, x) {
  if (x is in localTopK)
   // Adding x also evicts the now
   // (k+1)-st record from localTopK
   localTopK.add( x )
 }

 cleanup() {
  for each x in localTopK
   emit( dummy, x )
 }
```

```
reduce(dummy, [x1, x2,…]) {

 initialize globalTopK

 for each record x in input list
  if (x is in globalTopK)
   // Adding x also evicts the now
   // (k+1)-st record from globalTopK
   globalTopK.add(x)

 for each record x in globalTopK
  emit(NULL, x)
}
```

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? (5 points)

The amount of data transferred between mapper and reducer is as shown below :

```
Map-Reduce Framework
        Map input records=7012253
        Map output records=68931961
        Map output bytes=3613989271
        Map output materialized bytes=1247135949
        Input split bytes=9646
        Combine input records=0
        Combine output records=0
        Reduce input groups=3150469
        Reduce shuffle bytes=1247135949
        Reduce input records=68931961
        Reduce output records=2292317
        Spilled Records=143046037
        Shuffled Maps =954
        Failed Shuffles=0
        Merged Map outputs=954
        GC time elapsed (ms)=113152
        CPU time spent (ms)=10986920
        Physical memory (bytes) snapshot=92688470016
        Virtual memory (bytes) snapshot=392017997824
        Total committed heap usage (bytes)=80684253184
PageRankDriver$globalCounter
        pageCount=2292317
```

```
HDFS: Number of bytes read=9646
HDFS: Number of bytes written=1433487552
```

From the above screen-shots of syslog, we can get the following information

Map Input Records: 7012253
Map Output Records: 68931961

Reduce Input Records: 68931961
Reduce Output Records: 2292317

HDFS: Number of bytes read = 9646
HDFS: Number of bytes written = 1433487552

Performance Comparison

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.

1) Pre-processing:
    a) 6 Machines: 1694674 ms
    b) 11 Machines: 817642 ms

2) Page-Rank:
    a) 6 Machines: 1249807 ms
    b) 11 Machines: 924072 ms

3) Top-100 records:
    a) 6 Machines: 56117 ms
    b) 11 Machines: 54897 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

The speedup for the different phases is as follows:

Pre-Processing: time on 6 machines / time on 11 machines
1694674 / 817642 = 2.072

Page Rank: time on 6 machines / time on 11 machines
1249807 / 924072 = 1.352

Top-100 records: time on 6 machines / time on 11 machines
56117/54897 = 1.022

Pre-processing: In preprocessing, we are getting speedup of 2.072 which indicates good parallelism.

Page-Rank: In page-rank, we are getting a speedup of 1.35 which indicates good degree of parallelism. Since we are using pageName as the key, we can expect good load balancing, since page names are mostly unique.

Top-100 records: In top-100 records job, we are getting a scale-up of 1.07 which indicates that parallelism achieved by increasing the number of machines is not significant. this is due to the fact while calculating top-100 records, all the work is done by a single reducer. Hence increasing the number of machines does not significantly increase parallelism

Report the top-100 Wikipedia pages with the highest Page Ranks, 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?

Simple:

0.004546006683085723,United_States_09d4
0.003421367184609151,Wikimedia_Commons_7b57
0.002837854108871443,Country
0.0019128958770782945,England
0.001893623868425183,Europe
0.00189063406222648,United_Kingdom_5ad7
0.0018863222244596136,Water
0.0018104723850219677,Germany
0.0018078957577896406,France
0.001782692850423973,Earth
0.0017761070835650714,Animal

0.0016941543889209573,City
0.001509080873469262,Week
0.001405517538039679,Asia
0.0013858594240463266,Sunday
0.0013650997188877235,Monday
0.0013513117627065167,Wednesday
0.0013393075653381711,Wiktionary
0.0013347131212078185,Money
0.0013180487964604157,Friday
0.0013129832114232316,Plant
0.0013033386613353231,Saturday
0.0012863850538281122,Thursday
0.0012772265399668993,Tuesday
0.001269229690069004,Computer
0.0012675352719571284,English_language
0.0012409054831182553,Government
0.001238183685821667,Italy
0.001233980088923705,India
0.0011661170373948593,Number
0.0011197351835199247,Spain
0.0011042272681450372,Day
0.0010892818070885526,Japan
0.0010846976936956952,Canada
0.0010500379943580577,People
0.0010277930280467103,Human
9.99615400010956E-4,Wikimedia_Foundation_83d9
9.88229643421576E-4,China
9.865821321559352E-4,Australia
9.841244432521332E-4,Energy
9.579487834890896E-4,Sun
9.561043558319615E-4,index
9.529843537532963E-4,Food
9.426820142851911E-4,Science
9.254493149480996E-4,Mathematics
8.77314920381879E-4,Television
8.619921237236076E-4,Capital_(city)
8.562723041227841E-4,Russia

8.490433902998857E-4,State
8.304663346933987E-4,Year
8.24436214386361E-4,Music
8.028678351902631E-4,Greece
7.989810328259008E-4,Language
7.973423846701247E-4,Scotland
7.858037185653014E-4,Metal
7.791319865380842E-4,Wikipedia
7.742408413577361E-4,Greek_language
7.666509675199593E-4,Planet
7.572501148718657E-4,2004
7.425917694556622E-4,Sound
7.399459549363584E-4,Religion
7.297523562026461E-4,London
7.22161986810549E-4,Africa
6.915433441532776E-4,Geography
6.897878216281243E-4,Law
6.885288873252533E-4,20th_century
6.867768695327796E-4,Liquid
6.748111972622312E-4,19th_century
6.743361866947158E-4,World
6.664292548423801E-4,Society
6.656713852587338E-4,Scientist
6.485885665918626E-4,Atom
6.385336481006825E-4,Latin
6.385322348334243E-4,History
6.334749826187613E-4,Light
6.298045040975084E-4,Sweden
6.294929190212851E-4,Poland
6.28990051279928E-4,War
6.200517157442243E-4,Culture
6.18163753559403E-4,Netherlands
6.100708588939789E-4,Building
5.990230705211959E-4,Turkey
5.973182211532703E-4,Plural
5.959241778441696E-4,God
5.908338330469127E-4,Information

5.791455474047494E-4,Centuries
5.786729406795148E-4,Chemical_element
5.743756115712242E-4,Portugal
5.655358989376201E-4,Denmark
5.588416922845075E-4,Austria
5.576199710949825E-4,Cyprus
5.553568962748591E-4,Capital_city
5.537459902187112E-4,Ocean
5.469907345039025E-4,North_America_e7c4
5.461051008241 56E-4,Inhabitant
5.454983643637147E-4,Moon
5.444232908047917E-4,Species
5.433209501325888E-4,Disease
5.424871932837135E-4,Biology
5.409444816228805E-4,Book

Full Data Set :

0.0016411787135201484,United_States_09d4
0.0014777878917566 54,2006
7.823689422762323E-4,United_Kingdom_5ad7
6.782337449031476E-4,2005
5.165559432788325E-4,Biography
5.085971411620261E-4,Canada
5.064652914716857E-4,England
5.055358041377 88E-4,France
4.714976593027057E-4,2004
4.315121311511 94E-4,Germany
4.164059596585 25E-4,Australia
4.0057633296339224E-4,Geographic_coordinate_system
3.7963233850248307E-4,2003
3.672903743039332E-4,India
3.602718154255603E-4,Japan
3.0922034741062297E-4,Italy
3.044861962657224E-4,2001
3.0086254091458536E-4,2002
2.9509712360619 54E-4,Internet_Movie_Database_7ea7

2.8864756669765175E-4,Europe
2.8494108523079625E-4,2000
2.758462312057176E-4,World_War_II_d045
2.668441558022778E-4,London
2.5420871074074237E-4,English_language
2.5244608304332225E-4,Population_density
2.521112667910454E-4,1999
2.5205251492564895E-4,Spain
2.469514984143491E-4,Record_label
2.3793952477492816E-4,Russia
2.359043969236522E-4,Race_(United_States_Census)_a07d
2.3363058268814036E-4,Wiktionary
2.2306905721327443E-4,Wikimedia_Commons_7b57
2.176987050515585E-4,1998
2.0822142653032522E-4,1997
2.0773867552061962E-4,Music_genre
2.059041876242581E-4,New_York_City_1428
2.0526639447688416E-4,Scotland
1.952897290718076E-4,1996
1.919128508011 94E-4,Sweden
1.9161046908470756E-4,Football_(soccer)
1.9044109490254027E-4,Television
1.8474098833800082E-4,Square_mile
1.8399765214677203E-4,1995
1.8307167262777074E-4,Census
1.8204814405544512E-4,California
1.8127305523787615E-4,China
1.7847678677382676E-4,Netherlands
1.7617521126518486E-4,New_Zealand_2311
1.758017533695162E-4,1994
1.678858753642945E-4,1991
1.6597866611499108E-4,1993
1.652497637163338E-4,1990
1.640456008699301E-4,Public_domain
1.640249865203445E-4,New_York_3da4
1.5914603505585005E-4,1992
1.5748519228304757E-4,United_States_Census_Bureau_2c85
1.5633080250120144E-4,Film

1.5448406827780074E-4,Ireland
1.5435451034009743E-4,Norway
1.5401776991551556E-4,Actor
1.535327406457997E-4,Scientific_classification
1.498268800003559E-4,Population
1.495311807929372E-4,1989
1.482283114311815E-4,January_1
1.4646696116109593E-4,Latin
1.4642425340389767E-4,1980
1.4426142646083615E-4,Brazil
1.4403872923623423E-4,Mexico
1.4341849369626998E-4,Marriage
1.424042832864494E-4,1986
1.4032876760378417E-4,French_language
1.389171341175259E-4,1979
1.3868364328856642E-4,1985
1.3829160870527747E-4,1982
1.3825237418156004E-4,1981
1.3737896253439728E-4,1974
1.3721915914058981E-4,Poland
1.371787929988921E-4,Politician
1.3567923823914422E-4,South_Africa_1287
1.3566247336178138E-4,Switzerland
1.3543259642785882E-4,1984
1.3532755377256096E-4,1983
1.3527134264757152E-4,1987
1.3507196668768135E-4,Per_capita_income
1.3395389211669537E-4,1970
1.322262428086799E-4,1988
1.3222361651328964E-4,1976
1.3209529833566695E-4,Album
1.305822219158411E-4,Record_producer
1.3058162169410218E-4,1975
1.2918460659115381E-4,1969
1.2885551578071386E-4,Paris
1.283366743173938E-4,Greece
1.2830672454033486E-4,Km²
1.2828424736917155E-4,1945

1.28131222100640 12E-4,1972
1.27335168599441 12E-4,Soviet_Union_ad1f
1.26831083820602 33E-4,1977
1.26125976338635 4E-4,1978
1.24864272086049 68E-4,1973