CS6240-SHUBHAM DEB

	Runtime on 6 m/c	Runtime on 11 m/c
Pre-Processing	58	94.124626
Pagerank	652650.621013	290913.454154
Top-100	78803.199821	34215.73116

	Runtime on 6 m/c	Runtime on 11 m/c
Spark Execution Time	731511	325222
Hadoop Execution Time	3996077	2758318

As expected the spark execution time is faster than Hadoop:
1) It runs in-memory on the cluster, and that it isn't tied to Hadoop's MapReduce two-stage paradigm. This makes repeated access to the same data much faster.
2) Also, Spark uses "lazy evaluation" to form a directed acyclic graph (DAG) of consecutive computation stages. In this way, the execution plan can be optimized, e.g. to minimize shuffling data around. In contrast, this should be done manually in MapReduce by tuning each MR step.
3) Spark can launch tasks much faster because MapReduce starts a new JVM for each task, which can take seconds with loading JARs, parsing configuration XML, etc while Spark keeps an executor JVM running on each node.

PAGERANK EXECUTION STEPS IN SCALA

The spark application performs the following steps to compute pagerank:
1) We Read the data using sc.textFile() function
   Input: each line from the input
   Output: RDD<String> containing all the pages with their links
   This is a narrow dependency.
2) We pass the data read from the above step and pass it to a parser(a Java file) and then use filter to eliminate any bad links.
This is a **narrow** dependency.

3) The output from the above step is then used so that we can split on a delimiter **“:”** which is used to separate the pagename from the outgoing links and we map it to a RDD of pagename as the key and adjacency list as the value.

Input: RDD<String> containing all the pages

Output: RDD<Array<[String]>in which String is split on **“:”**Hence first element would be the pagename and the the second element would be the adjacency list of outlinks.

This is a **narrow** dependency.

4) We add the dangling nodes to the above RDD by going through each dangling node and mapping each outgoing link as a RDD containing empty adjacency list. For example:

A → B,C will have in the RDD:

(A,(B,C)),(B.()),(C,())

Input: RDD<Array<[String]>in which String is split on **“:”**Hence first element would be the pagename and the the second element would be the adjacency list of outlinks.

Output: RDD<String,Array<[String]>in which pagename is the key and the arraylist would be the adjacency list containing it’s outlinks.

This is using **wide** dependency as here I am trying to aggregate all the dangling nodes by reducing them by key.

5) In each pagerank iteration, I am calculating delta using by joining the adjacency list RDD and the pagerank RDD, which would help me to determine whether or not to add the pagerank to delta if the adjacency list is null or not.

We filter out all the dangling nodes and do aggregation on each pagename as the key and get the sum of all the dangling nodes by using reduce(). This is a wide dependency.

6) Then we calculate all the contributions from the incoming nodes and then aggregate them based on key.

Input: RDD<String,Array<[String]>in which pagename is the key and the arraylist would be the adjacency list containing it’s outlinks.

Output: RDD<String,Double> in which each String pagename has a corresponding sum of contributions from other nodes.

This is a wide dependency.

7) In order to assign contribution to links that have no incoming links, we filter them out and assign them a pagerank of 0 as they get no contributions from other nodes by using subtractByKey()

This is wide dependency.
8) Then I append all the noInlinks to the contributors RDD and we map each of these values to calculate new pagerank value:

0.15*(1/numnodes)+(0.85*delta)/numnodes+(0.85*contributions)

This is a narrow dependency.
9) Then we repeat the iterations
10) We sort the output by giving pagerank as the key and pagename as the value and by applying function top(100) we get the top 100 links.
This is a narrow dependency.

My Program has 202 stages.

LOCAL OUTPUT ON THE SIMPLE DATASET

(0.0056317761828779825,United_States_09d4)
(0.004452925815711511,Wikimedia_Commons_7b57)
(0.0038936167925799777,England)
(0.0036235380070835004,Germany)
(0.002502809113545449,France)
(0.002286138935432594,Inhabitant)
(0.0019897961357322426,City)
(0.001825965155487697,Wiktionary)
(0.0016859620084768742,Japan)
(0.001672375718734812,Computer)
(0.0015980022724908394,Animal)
(0.0015670612492369128,United_Kingdom_5ad7)
(0.0014855343285828833,Country)
(0.0014832207971928778,India)
(0.0014193312256378136,Europe)
(0.001392009453407515,Australia)
(0.0013590294669440668,Italy)
(0.0013535999878973707,Water)
(0.0013535930247973317,Canada)
(0.0013273625430613575,English_language)
(0.0013095659832212524,Spain)
(0.0013038802962975933,Television)
(0.0012237050763191648,Plant)
(0.0011847206456424815,Earth)

(0.0011708715610589135,London)
(0.0011301703438957623,Football_(soccer))
(0.0011077969975849234,Scotland)
(0.0011020457124940365,China)
(0.0010944118909705682,Greece)
(0.001077281675249741,Music)
(0.0010531599821375737,Money)
(0.0010248024051610172,Food)
(0.0010153567658880197,Metal)
(0.00101459036875912,Capital_(city))
(0.0010089137415324613,Capital_city)
(9.943825827883526E-4,Netherlands)
(9.89050981165042E-4,Movie)
(9.705426520233404E-4,Brazil)
(9.688727120344198E-4,2005)
(9.645355290931428E-4,U.S._state_5a68)
(9.617486392316953E-4,Human)
(9.328485428473629E-4,Greek_mythology)
(9.224918509929682E-4,Book)
(9.19275663873436E-4,Poland)
(9.156871942643847E-4,Mathematics)
(8.986023120334558E-4,Russia)
(8.940553946464043E-4,Number)
(8.915101430774436E-4,2006)
(8.507721257824496E-4,Actor)
(8.491937572292197E-4,Language)
(8.43412744647354E-4,Government)
(8.352054469396938E-4,2004)
(8.332988364146013E-4,People)
(8.260298404399345E-4,California)
(8.209176407783963E-4,Year)
(8.081879626228026E-4,Sweden)
(8.077524748873112E-4,God)
(8.005969745351859E-4,Religion)
(7.748492086392958E-4,University)
(7.59281322041569E-4,Fruit)
(7.46089360338014E-4,Asia)
(7.356857887588011E-4,Science)

(7.224594838804932E-4,Film)
(7.190202490862731E-4,Internet_Movie_Database_7ea7)
(7.189072439391822E-4,Car)
(7.124823612692201E-4,19th_century)
(7.089423000292747E-4,Internet)
(7.047140020793375E-4,Chemical_element)
(6.975177246064461E-4,Africa)
(6.959247028249421E-4,World_War_II_d045)
(6.934862204272168E-4,Disease)
(6.840455233327224E-4,Company)
(6.656471178596653E-4,Species)
(6.609006536752668E-4,Latin)
(6.587157406917282E-4,North_America_e7c4)
(6.551583625117247E-4,River)
(6.492060438283372E-4,Video_game)
(6.441711033987011E-4,Fish)
(6.348819437762124E-4,Prefecture)
(6.318657664598953E-4,1970s)
(6.295816801265288E-4,Island)
(6.215007552289683E-4,Singer)
(6.072968162510523E-4,Liquid)
(6.01990275418283E-4,Sport)
(6.015846530812625E-4,Chad)
(5.860762483547293E-4,German_language)
(5.84478056779573E-4,1960s)
(5.812168330294553E-4,County)
(5.811338060346504E-4,Band)
(5.806128128053537E-4,Christianity)
(5.804848299854361E-4,New_York_City_1428)
(5.785307842246899E-4,Greek_language)
(5.776675762957429E-4,Tool)
(5.764501299864103E-4,War)
(5.730857637694089E-4,20th_century)
(5.578549975354585E-4,2001)
(5.57567212284005E-4,Mammal)
(5.562540803872637E-4,Austria)
(5.54014678332751E-4,2003)
(5.530394490389887E-4,Km²)

AWS OUTPUT ON THE FULL DATASET

(0.0025875738790207057,United_States_09d4)
(0.0012121056986652548,2006)
(0.0011886074301786114,United_Kingdom_5ad7)
(9.813720091963068E-4,Biography)
(9.05332614942987E-4,2005)
(8.699140229635562E-4,England)
(8.451601055534671E-4,Canada)
(7.768478669111108E-4,Geographic_coordinate_system)
(7.156957671772966E-4,France)
(7.109486874046102E-4,2004)
(6.729246787356331E-4,Australia)
(6.459678697707187E-4,Germany)
(5.80326756208131E-4,2003)
(5.76801384038091E-4,India)
(5.753517188509887E-4,Japan)
(5.314390982957382E-4,Internet_Movie_Database_7ea7)
(5.03521183334416E-4,Europe)
(4.919495027986931E-4,Record_label)
(4.812265314253436E-4,2001)
(4.769405138358850 5E-4,2002)
(4.721562521826126E-4,Population_density)
(4.716412298141728E-4,World_War_II_d045)
(4.6287364816733034E-4,Music_genre)
(4.5896625055166904E-4,2000)
(4.4031846174467877E-4,Italy)
(4.374592860272672E-4,Wikimedia_Commons_7b57)
(4.3612645286178773E-4,Wiktionary)
(4.295857538539904E-4,London)
(4.1243067112942876E-4,English_language)
(4.0086609685105375E-4,1999)
(3.586719654052174E-4,Spain)
(3.519866622204175E-4,1998)
(3.39356817190978E-4,Russia)
(3.3364669754869263E-4,Television)
(3.3326907080584895E-4,1997)

(3.3067598592203054E-4,New_York_City_1428)
(3.241073293444151E-4,Football_(soccer))
(3.207926137704338E-4,Census)
(3.197823005176379E-4,1996)
(3.184817145567736E-4,Scotland)
(3.1120437108255074E-4,Square_mile)
(3.063556132480581E-4,1995)
(3.04389204291456E-4,China)
(3.0346263301045916E-4,Population)
(3.025322913242835E-4,Scientific_classification)
(2.981925122128793E-4,California)
(2.872214822972601E-4,1994)
(2.8475627813825597E-4,Public_domain)
(2.8431449705184774E-4,Sweden)
(2.8400317775039823E-4,Record_producer)
(2.8381166479992764E-4,Film)
(2.798896010099988E-4,New_Zealand_2311)
(2.7558963447707805E-4,New_York_3da4)
(2.7527888164922265E-4,United_States_Census_Bureau_2c85)
(2.734028557593524E-4,Netherlands)
(2.728070179443469E-4,Marriage)
(2.715767629087643E-4,1993)
(2.6867823779444023E-4,1991)
(2.6585027243587674E-4,Politician)
(2.6515832125038516E-4,1990)
(2.631174957469717E-4,1992)
(2.619885428973026E-4,Album)
(2.583031665436375E-4,Per_capita_income)
(2.5718403782976215E-4,Latin)
(2.5696898032223034E-4,Actor)
(2.5514176805721756E-4,Ireland)
(2.489060468269134E-4,Studio_album)
(2.4759837317354493E-4,Poverty_line)
(2.4599591922284627E-4,Km²)
(2.439309899334116E-4,1989)
(2.3851865820809982E-4,Norway)
(2.3668417613464737E-4,Website)
(2.3255569466904534E-4,1980)

(2.31339527337428E-4,Area)
(2.2792507774442981E-4,Animal)
(2.2616863236359078E-4,Personal_name)
(2.2439403869535274E-4,1986)
(2.2371376643876315E-4,Poland)
(2.2337147397289407E-4,Brazil)
(2.2141653270397313E-4,1985)
(2.2073847571755487E-4,1987)
(2.190826539884543E-4,1983)
(2.1856109366308812E-4,1982)
(2.167735427422231E-4,1981)
(2.1675144231500575E-4,1979)
(2.1640134086344412E-4,French_language)
(2.1624106833540628E-4,1984)
(2.1599701654168867E-4,1988)
(2.1570002920061077E-4,World_War_I_9429)
(2.154412826170668E-4,1974)
(2.152557401352954E-4,Paris)
(2.129990570281566E-4,Mexico)
(2.0897065377531982E-4,19th_century)
(2.087875732917075E-4,1970)
(2.0802790845061457E-4,January_1)
(2.0789003357067843E-4,USA_f75d)
(2.0611904973606018E-4,1975)
(2.0597554282342402E-4,1976)
(2.0541873037329303E-4,Africa)
(2.049813351372208E-4,South_Africa_1287)