# CS 6240 Parallel Processing with Map Reduce
## Section 01 HW4

Shabbir Saifee

Design Discussion

Data transformation:

Note: For Each line the transformation is documented in the source code

| **Step** | **Input** | **Output** |
|---|---|---|
| Preprocessing | Text file | PairRDD: (PageName: String, AdjacencyList: List[String]) |
| Initial page Rank | PairRDD: (PageName: String, AdjacencyList: List[String]) | PairRDD: (PageName: String, PageRank: Double) |
| Page Rank Calculation(10 iterations) | PairRDD: (PageName: String, AdjacencyList: List[String]) | PairRDD: (PageName: String, PageRank: Double)<br>with updated page ranks |
| Top K Pages | PairRDD: (PageName: String, PageRank: Double) | Array[(String, Double)] |

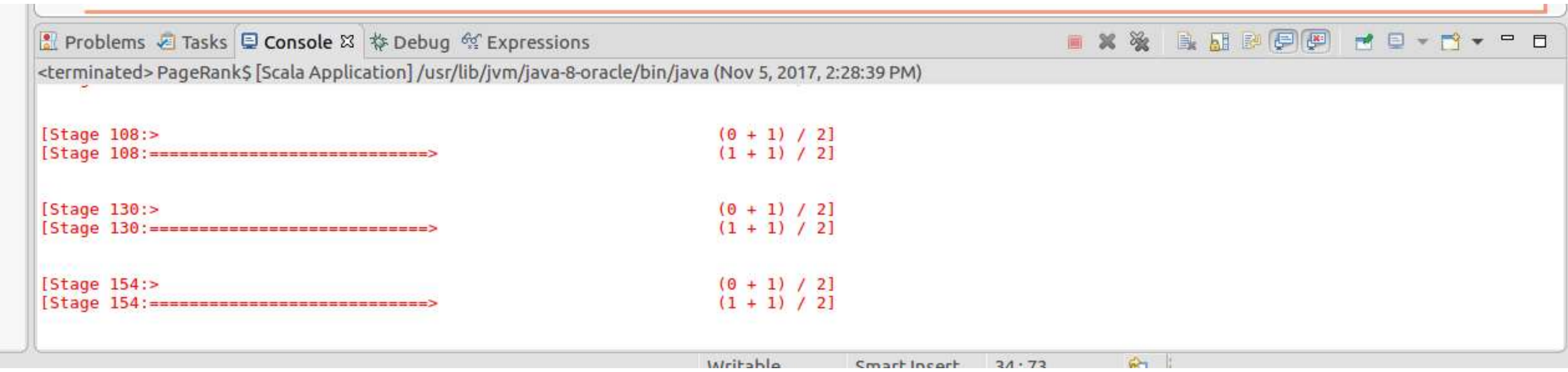
Transformations with narrow dependencies:

- map
- flatmap
- filter

Transformations with wide dependencies

- join
- reduceByKey
- reduce
- union

Total Stages in spark : 154



<terminated> PageRank$ [Scala Application] /usr/lib/jvm/java-8-oracle/bin/java (Nov 5, 2017, 2:28:39 PM)

```
[Stage 108:>                                                         (0 + 1) / 2]
[Stage 108:============================>                             (1 + 1) / 2]


[Stage 130:>                                                         (0 + 1) / 2]
[Stage 130:============================>                             (1 + 1) / 2]


[Stage 154:>                                                         (0 + 1) / 2]
[Stage 154:============================>                             (1 + 1) / 2]
```

Performance Comparison

| Run# | Hadoop Run Time | Spark Run Time |
|---|---|---|
| Run1 | 6072 seconds | 6150 seconds |
| Run2 | 4368 seconds | 4736 seconds |

Theoretically spark should be faster than Hadoop, but from above results it seems that Spark is taking more time. The plausible reason behind this could be that the parsing process is not executed parallelly on the spark whereas, on Hadoop it is parallel thus reducing the overall execution time.

Output of EMR:

Top 100 pages

```
(United_States_09d4,0.0029413056575926873)
(2006,0.0026270396443048645)
(United_Kingdom_5ad7,0.0013976958662404555)
(2005,0.0012110962940753648)
(Biography,9.531396066964096E-4)
(Canada,9.135398901488759E-4)
(England,9.070253721556724E-4)
(France,8.969192236018876E-4)
(2004,8.428114042729063E-4)
(Germany,7.709395485776397E-4)
(Australia,7.471082835414593E-4)
(Geographic_coordinate_system,7.198097255687983E-4)
(2003,6.789007801249108E-4)
(India,6.579907199093056E-4)
(Japan,6.531032183765519E-4)
(Italy,5.468696044097373E-4)
(2001,5.446597619539529E-4)
(2002,5.384745590260148E-4)
(Internet_Movie_Database_7ea7,5.325000677503746E-4)
(Europe,5.19029073541379E-4)
```

```
(2000,5.098431618320911E-4)
(World_War_II_d045,4.921595879864778E-4)
(London,4.749477290348249E-4)
(1999,4.507960320671722E-4)
(Population_density,4.503824037289746E-4)
(English_language,4.483127459425858E-4)
(Record_label,4.474978517851885E-4)
(Spain,4.469690338740799E-4)
(Russia,4.2250486846523964E-4)
(Race_(United_States_Census)_a07d,4.1687888776305066E-4)
(Wiktionary,4.093514438016729E-4)
(1998,3.895858302488107E-4)
(Wikimedia_Commons_7b57,3.886912567450177E-4)
(Music_genre,3.7948860503198305E-4)
(1997,3.7162470663100076E-4)
(Scotland,3.6590421518006513E-4)
(New_York_City_1428,3.657248147336632E-4)
(Football_(soccer),3.54598603696281E-4)
(1996,3.487282723347976E-4)
(Sweden,3.435385068323198E-4)
(Television,3.4284147981225084E-4)
(Census,3.289006470382112E-4)
(1995,3.284334016903185E-4)
(California,3.262394909522743E-4)
(China,3.220711964145969E-4)
(Square_mile,3.207221874696565E-4)
(Netherlands,3.171325493888409E-4)
(New_Zealand_2311,3.158999957373838E-4)
(1994,3.1347346119818954E-4)
(1991,2.9909125190972694E-4)
(1993,2.963510471962803E-4)
(1990,2.945810064511358E-4)
(New_York_3da4,2.9290005401207673E-4)
(Public_domain,2.9202509970855636E-4)
(1992,2.8415789786409524E-4)
(Film,2.8165238844106194E-4)
(Actor,2.785964235079639E-4)
(United_States_Census_Bureau_2c85,2.7808505651651403E-4)
(Scientific_classification,2.778206040110102E-4)
(Norway,2.76454464248674E-4)
(Ireland,2.7544680366643173E-4)
(Poland,2.7251041444603506E-4)
(Population,2.7179645872378545E-4)
(1989,2.6639954684079466E-4)
(January_1,2.602915050755522E-4)
(1980,2.60187226808109E-4)
(Marriage,2.578860391738928E-4)
(Brazil,2.5761403487918643E-4)
(Mexico,2.566082480931703E-4)
```

```
(Latin,2.5645063953744993E-4)
(1986,2.530776843280789E-4)
(Politician,2.5092820126455936E-4)
(1985,2.4688879578528984E-4)
(1979,2.465824781197497E-4)
(French_language,2.460938372461549E-4)
(1982,2.459046337026834E-4)
(1981,2.4573431056935843E-4)
(1974,2.4348948997373888E-4)
(Switzerland,2.4130941672518037E-4)
(South_Africa_1287,2.4126785164289768E-4)
(1984,2.4121171682879887E-4)
(1987,2.4094331032462204E-4)
(1983,2.4089177875400445E-4)
(Album,2.4062534509688323E-4)
(Per_capita_income,2.3955808937212985E-4)
(1970,2.3710763269708701E-4)
(Record_producer,2.3606302686968004E-4)
(1988,2.3547053619971852E-4)
(1976,2.3434286812575738E-4)
(Km²,2.3170219408590606E-4)
(1975,2.3163187770156534E-4)
(Paris,2.286795071654059E-4)
(1969,2.2846407474683698E-4)
(Greece,2.279818697116187E-4)
(1945,2.2713953023062908E-4)
(1972,2.269949069964333E-4)
(1977,2.2507523113519877E-4)
(Soviet_Union_ad1f,2.2419829213391314E-4)
(1978,2.239673887509946E-4)
(Personal_name,2.2315158065712285E-4)
```