Parallel Programming –PageRank

106062530 張原嘉

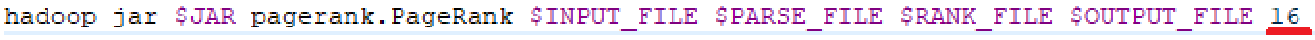
1. Instruction

在編譯程式上，因為有附 makefile 檔案，只要將 makefile 與 src 資料夾放一起並鍵入 make 即可編譯。執行方面，有一個 execute.sh 批次檔可供調整輸入參數，以下對各項參數稍作解釋

✔輸入檔大小

```
INPUT_FILE=/user/ta/PageRank/Input/input-50G
```

✔Iteration 次數



```
hadoop jar $JAR pagerank.PageRank $INPUT_FILE $PARSE_FILE $RANK_FILE $OUTPUT_FILE 16
```

✔最終 output 檔名

```
hdfs dfs -getmerge $OUTPUT_FILE pagerank.txt
```

2. Implementation

程式主要分為三個部分：Parsing、Ranking、Sorting，每個部份有各自的 Mapper/Reducer。

✔Parsing

首先 Mapper 是用來解析 Input File 及去除 out-link 的情況。根據需要，mapper 會產生<k,value>（k = 0, 1, 2…NumOfReducer, value = PageTitle）、<k,value>（k = title T1, T2, T3…, value = LinkToTitle）兩種 Key-Value Pair，且他們以前面是否有多一個空白為區隔。在建第二種 K-V Pair 時，若遇到 Page 後面沒有 Link（也就是遇到 Dangling node）則會加入一些字符以識別，最後用一個變數 N 紀錄總 page 數量。要送給

Reducer 之前，Partitioner 得用來將剛剛 Mapper 的 K-V Pair 分堆，一堆是前面有空白的，這堆會直接送給 K 值對應到的 Reducer（若<0, PageTitle>會送給 Reducer 0 號 、<1, PageTitle>會送給 Reducer 1 號，以此類推。）另一堆則以 mod 方式計算這個 K-V pair 要被送給哪個 Reducer。最後 Reducer 除了為先前 Mapper 傳過來的 K-V Pair 設定 PageRank 初值，也會建立一個 HashSet，不論是哪種 K-V Pair 都會被加到 HashSet 裡，並以垂直分隔線 " | " 分隔 Link。例如，＜title1, rank|L1|L2|L3＞。若第二種 K-V Pair 含有 Dangling Node，則結果會是＜title, rank|＞，排除 out-link 的方法則是在最後面判斷其 PageRank 值，若不為 0 則寫入 output，為 0 則不寫入。

✔Ranking

接下來是實際計算 PageRank 的 Ranking。計算 PageRank 迴圈終止條件有兩個，一個是如果沒有在 Input 輸入第四個參數（也就是 Iteration 次數），那麼終止條件即為偏差值（Error 值）小於 0.001，如果有的話，則依據輸入的 Iteration 次數或 Error 值小於 0.001 作為終止條件。而在 mapper 方面，會傳送兩種 K-V Pair 給 Reducer，一種是原本傳入 RankMapper 的 K-V Pair，目的是為了 Iteration 計算；另外一種是記錄著 Link、PageRank 的 K-V Pair，目的是 PageRank 加總後，給下一輪計算用。另外，Mapper 拿到的資料是用 split（ " \\| " ）切割的，字串陣列 value_arr 存放切割完的資料，value_arr［ 0 ］放 PageRank 值，value_arr［ 1]（如果有的話）則是 Link 資訊。如果此陣列長度為 1 ，代表它是 Dangling Node（因為只有

PageTitle，沒有 link 之後的資訊），此時即更新 DanglingSum 和 Dangling 值。若大於 1，則為普通之 Page Node，更新其 PageRank 值成 PR（t） / C（t），C（t）為 value_arr.length-1。接著是產生 K-V Pair，假設 A 連到 BC， A 的 PageRank 為 10，則送出 < B,5 > 、 < C,5 > 的 K-V pair 給 Reducer（ A 連到 B、 C ，且 B、 C 皆存在 ）。

Reducer 一開始有個 setup Function，是為了能取得從 Mapper 送過來的一些數值，如 Dangling、DanglingSum，如此才能重複計算 PageRank。對於每一個接收到的 K-V Pair，若 key 值相同，則後面的 value 會一直往後增長下去，此時一樣用 split（ " \\| " ）方式切割資料，一個 Value 可能被切割成 v1, v2, v3…vn，每個 value 前面如果有!（這個!是前面 Mapper 做的標記），代表是一個新 rank 值，此時就要累加 rank 值。直到沒有!，就把圖建回去，以便之後跟新的 Rank 值做比較。當全部累加完後，即使用公式計算出新 Rank 值並與舊值比較，此值也會給主程式當作迴圈終止的條件。最後， Reducer 會建立新的 K-V Pair，如 < P1, NewRank|L1|L2| > ，如此下輪的 Mapper 即可拿到符合格式的 input pair。

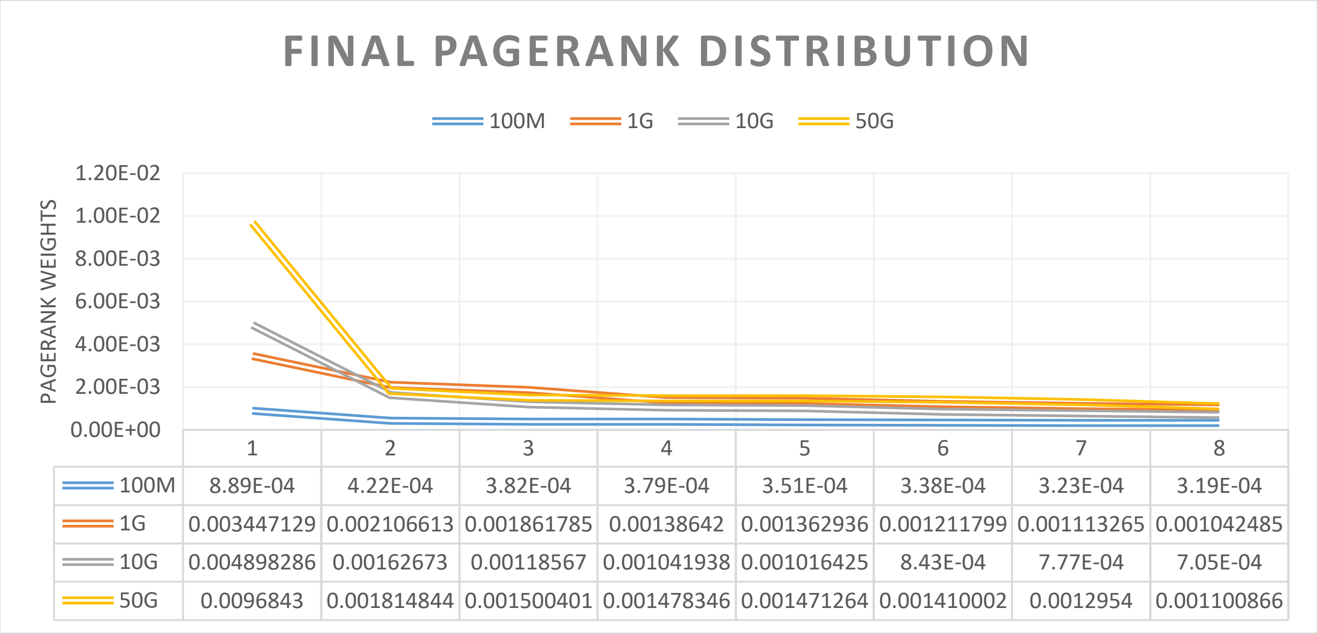
✔Sorting

Sorting 的 Mapper 部份，K-V Pair 是 < Pagetitle 和 PageRank,PageRank > Partitioner 部份則是根據 PageRank 與 Average 值（1/N）做分類，PageRank > Average 回傳 0 ，反之回傳 1。Comparator 與日前 Lab5 類似，若現在要排序的 Rank 值小於要比較的 Rank 值，則回傳 1，反之回傳-1，即根據 Rank 值由大到小的排列。最後 Reducer 輸出的格式即為

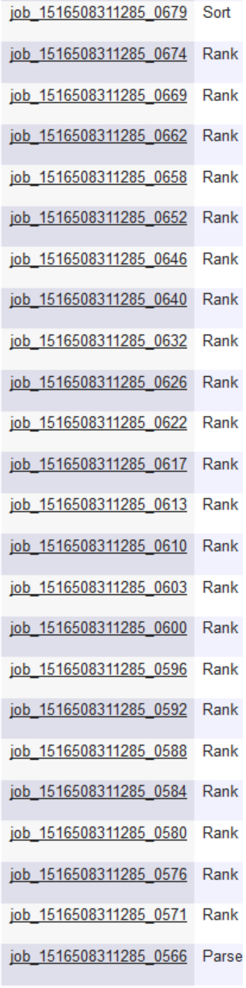
PageTitle PageRank。

3. Experiment & Analysis

✔Analyze the distribution of PageRank weights

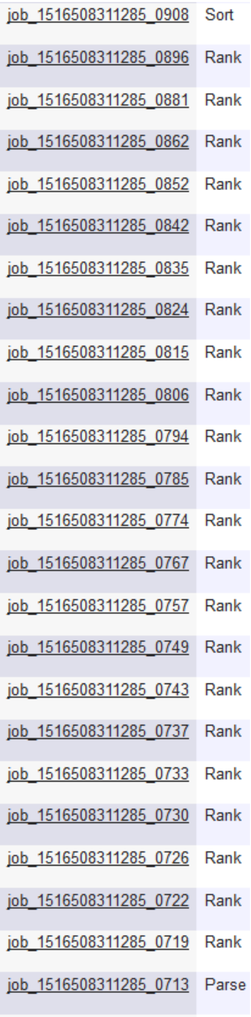


	1	2	3	4	5	6	7	8
100M	8.89E-04	4.22E-04	3.82E-04	3.79E-04	3.51E-04	3.38E-04	3.23E-04	3.19E-04
1G	0.003447129	0.002106613	0.001861785	0.00138642	0.001362936	0.001211799	0.001113265	0.001042485
10G	0.004898286	0.00162673	0.00118567	0.001041938	0.001016425	8.43E-04	7.77E-04	7.05E-04
50G	0.0096843	0.001814844	0.001500401	0.001478346	0.001471264	0.001410002	0.0012954	0.001100866



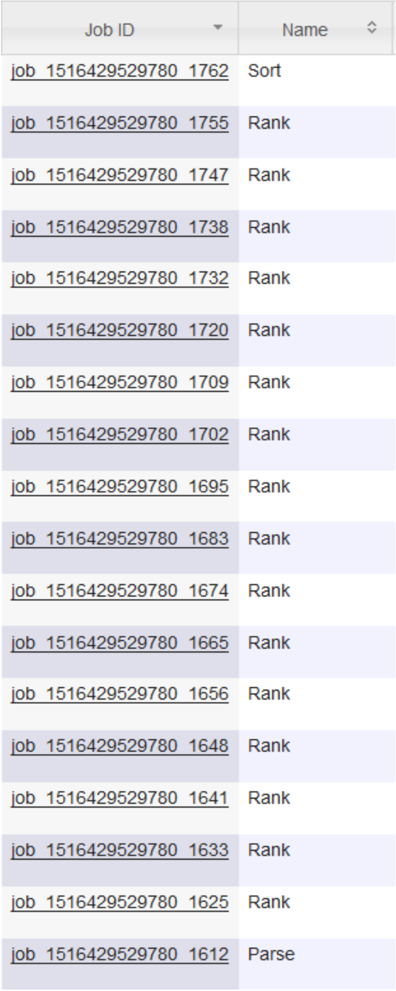
job_1516508311285_0679	Sort
job_1516508311285_0674	Rank
job_1516508311285_0669	Rank
job_1516508311285_0662	Rank
job_1516508311285_0658	Rank
job_1516508311285_0652	Rank
job_1516508311285_0646	Rank
job_1516508311285_0640	Rank
job_1516508311285_0632	Rank
job_1516508311285_0626	Rank
job_1516508311285_0622	Rank
job_1516508311285_0617	Rank
job_1516508311285_0613	Rank
job_1516508311285_0610	Rank
job_1516508311285_0603	Rank
job_1516508311285_0600	Rank
job_1516508311285_0596	Rank
job_1516508311285_0592	Rank
job_1516508311285_0588	Rank
job_1516508311285_0584	Rank
job_1516508311285_0580	Rank
job_1516508311285_0576	Rank
job_1516508311285_0571	Rank
job_1516508311285_0566	Parse

1G



job_1516508311285_0908	Sort
job_1516508311285_0896	Rank
job_1516508311285_0881	Rank
job_1516508311285_0862	Rank
job_1516508311285_0852	Rank
job_1516508311285_0842	Rank
job_1516508311285_0835	Rank
job_1516508311285_0824	Rank
job_1516508311285_0815	Rank
job_1516508311285_0806	Rank
job_1516508311285_0794	Rank
job_1516508311285_0785	Rank
job_1516508311285_0774	Rank
job_1516508311285_0767	Rank
job_1516508311285_0757	Rank
job_1516508311285_0749	Rank
job_1516508311285_0743	Rank
job_1516508311285_0737	Rank
job_1516508311285_0733	Rank
job_1516508311285_0730	Rank
job_1516508311285_0726	Rank
job_1516508311285_0722	Rank
job_1516508311285_0719	Rank
job_1516508311285_0713	Parse

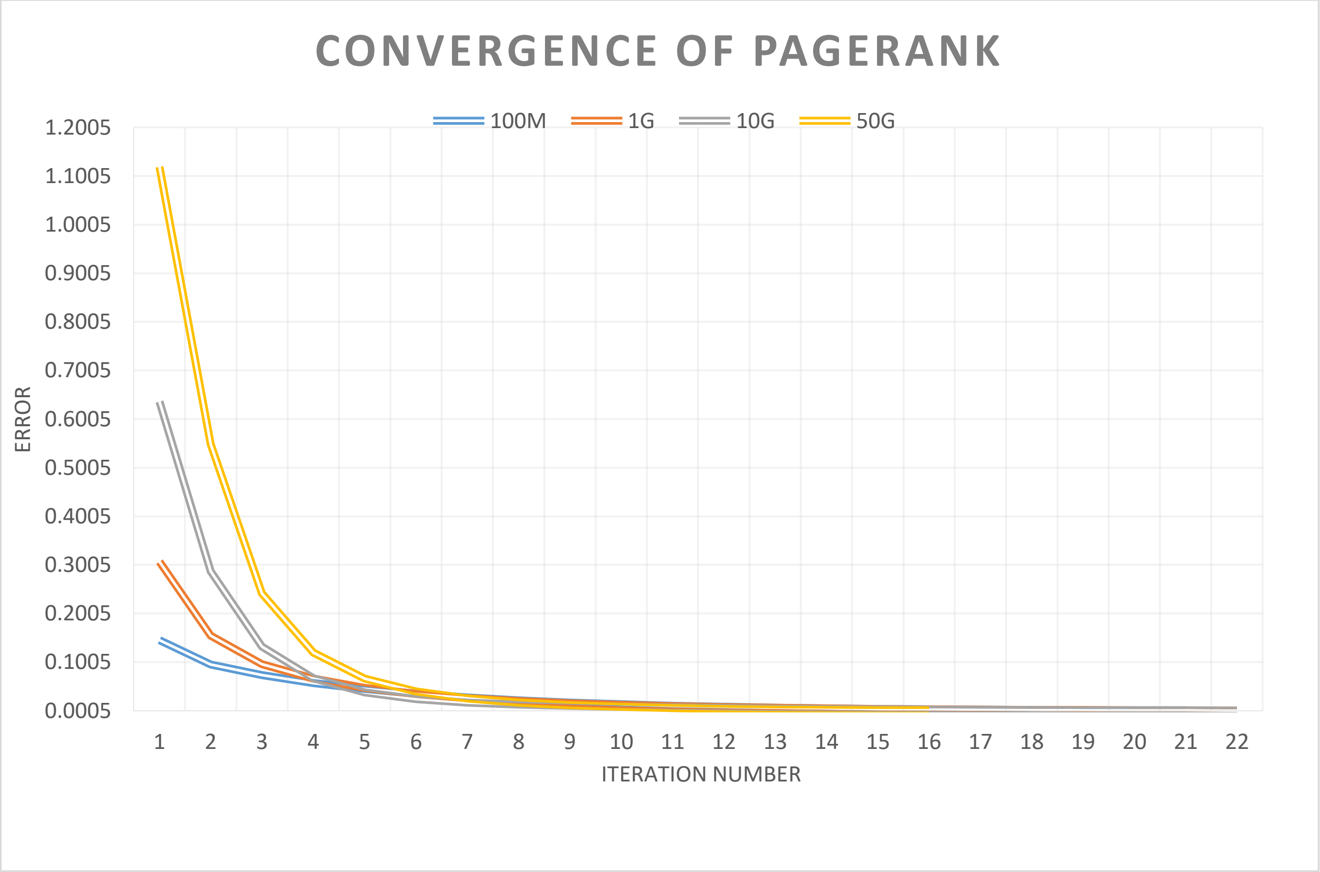
10G



Job ID	Name
job_1516429529780_1762	Sort
job_1516429529780_1755	Rank
job_1516429529780_1747	Rank
job_1516429529780_1738	Rank
job_1516429529780_1732	Rank
job_1516429529780_1720	Rank
job_1516429529780_1709	Rank
job_1516429529780_1702	Rank
job_1516429529780_1695	Rank
job_1516429529780_1683	Rank
job_1516429529780_1674	Rank
job_1516429529780_1665	Rank
job_1516429529780_1656	Rank
job_1516429529780_1648	Rank
job_1516429529780_1641	Rank
job_1516429529780_1633	Rank
job_1516429529780_1625	Rank
job_1516429529780_1612	Parse

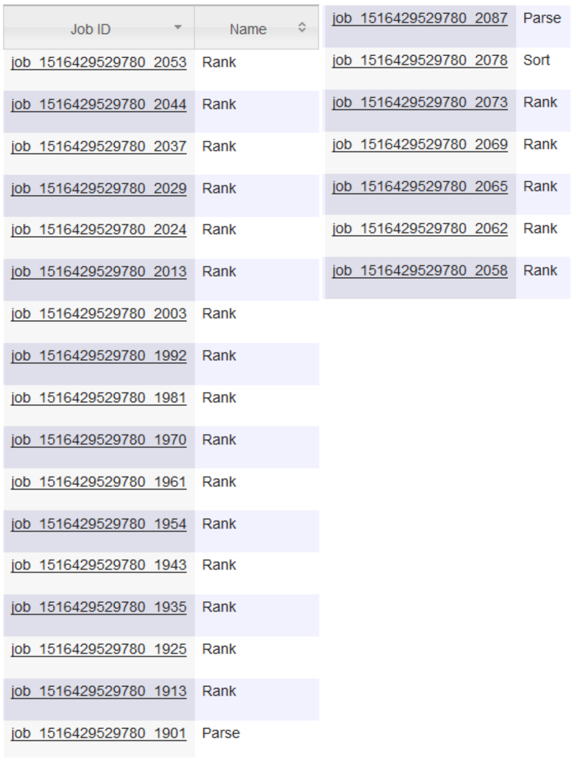
50G

✔Analyze the converge rate



✔Performance analysis with different settings

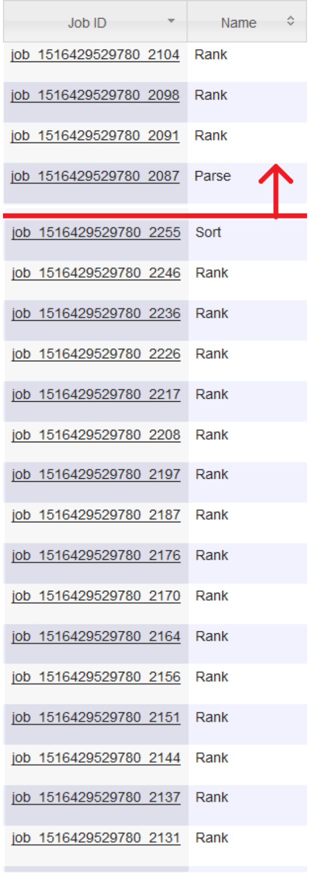
NumOfReducer	Time
1	120min
2	86min
4	84min
8	10min
16	9min
32	14min



Job ID	Name
job_1516429529780_2053	Rank
job_1516429529780_2044	Rank
job_1516429529780_2037	Rank
job_1516429529780_2029	Rank
job_1516429529780_2024	Rank
job_1516429529780_2013	Rank
job_1516429529780_2003	Rank
job_1516429529780_1992	Rank
job_1516429529780_1981	Rank
job_1516429529780_1970	Rank
job_1516429529780_1961	Rank
job_1516429529780_1954	Rank
job_1516429529780_1943	Rank
job_1516429529780_1935	Rank
job_1516429529780_1925	Rank
job_1516429529780_1913	Rank
job_1516429529780_1901	Parse

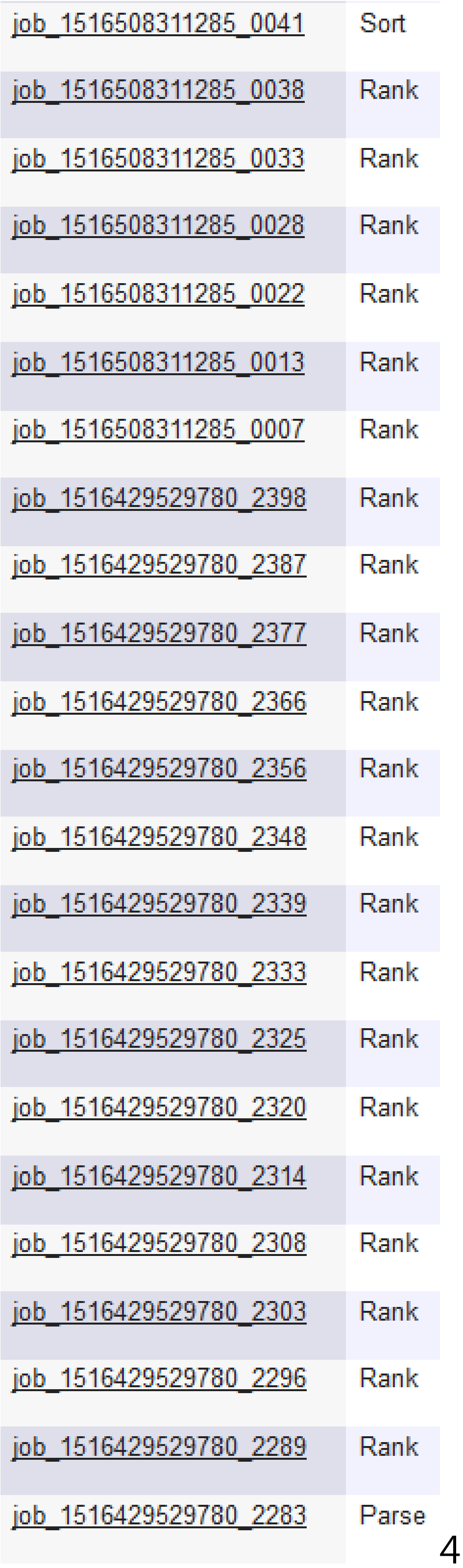
job_1516429529780_2087	Parse
job_1516429529780_2078	Sort
job_1516429529780_2073	Rank
job_1516429529780_2069	Rank
job_1516429529780_2065	Rank
job_1516429529780_2062	Rank
job_1516429529780_2058	Rank

1R



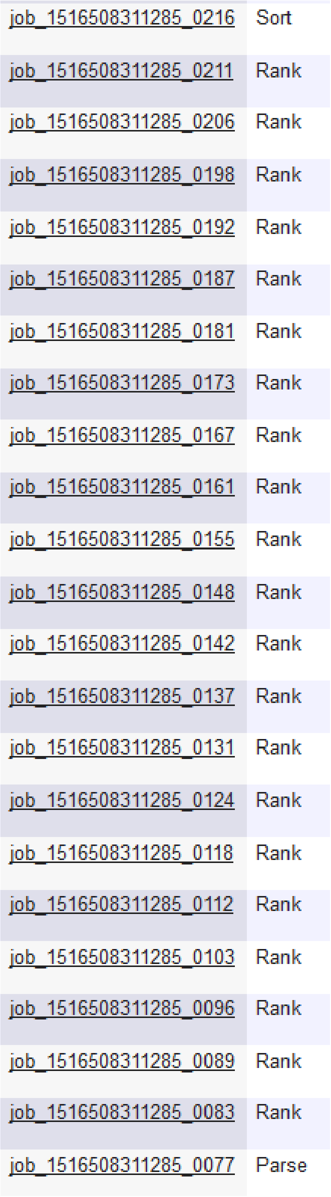
Job ID	Name
job_1516429529780_2104	Rank
job_1516429529780_2098	Rank
job_1516429529780_2091	Rank
job_1516429529780_2087	Parse
job_1516429529780_2255	Sort
job_1516429529780_2246	Rank
job_1516429529780_2236	Rank
job_1516429529780_2226	Rank
job_1516429529780_2217	Rank
job_1516429529780_2208	Rank
job_1516429529780_2197	Rank
job_1516429529780_2187	Rank
job_1516429529780_2176	Rank
job_1516429529780_2170	Rank
job_1516429529780_2164	Rank
job_1516429529780_2156	Rank
job_1516429529780_2151	Rank
job_1516429529780_2144	Rank
job_1516429529780_2137	Rank
job_1516429529780_2131	Rank

2R



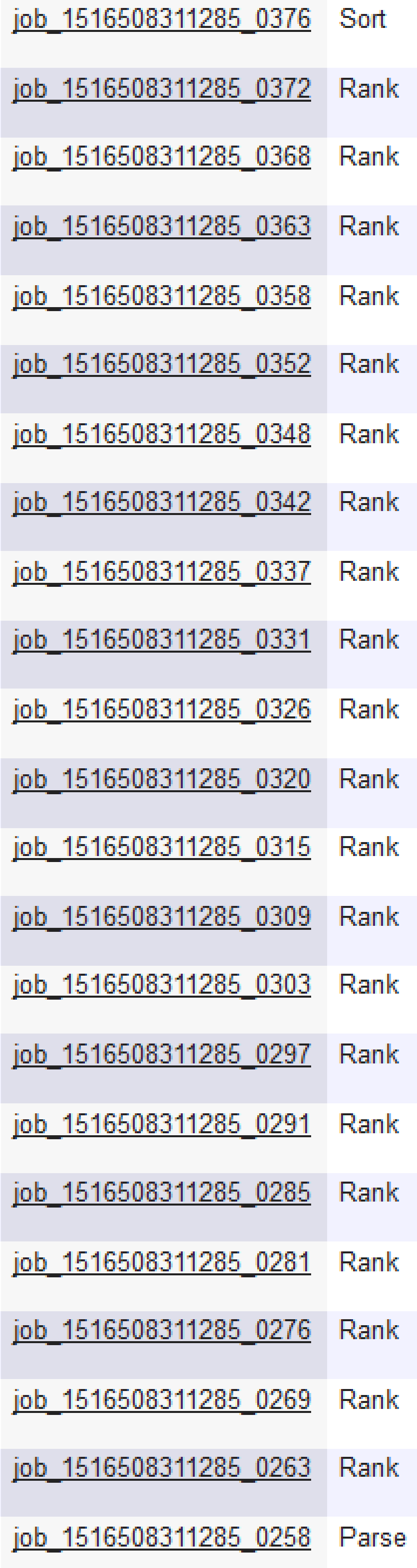
job_1516508311285_0041	Sort
job_1516508311285_0038	Rank
job_1516508311285_0033	Rank
job_1516508311285_0028	Rank
job_1516508311285_0022	Rank
job_1516508311285_0013	Rank
job_1516508311285_0007	Rank
job_1516429529780_2398	Rank
job_1516429529780_2387	Rank
job_1516429529780_2377	Rank
job_1516429529780_2366	Rank
job_1516429529780_2356	Rank
job_1516429529780_2348	Rank
job_1516429529780_2339	Rank
job_1516429529780_2333	Rank
job_1516429529780_2325	Rank
job_1516429529780_2320	Rank
job_1516429529780_2314	Rank
job_1516429529780_2308	Rank
job_1516429529780_2303	Rank
job_1516429529780_2296	Rank
job_1516429529780_2289	Rank
job_1516429529780_2283	Parse

4R



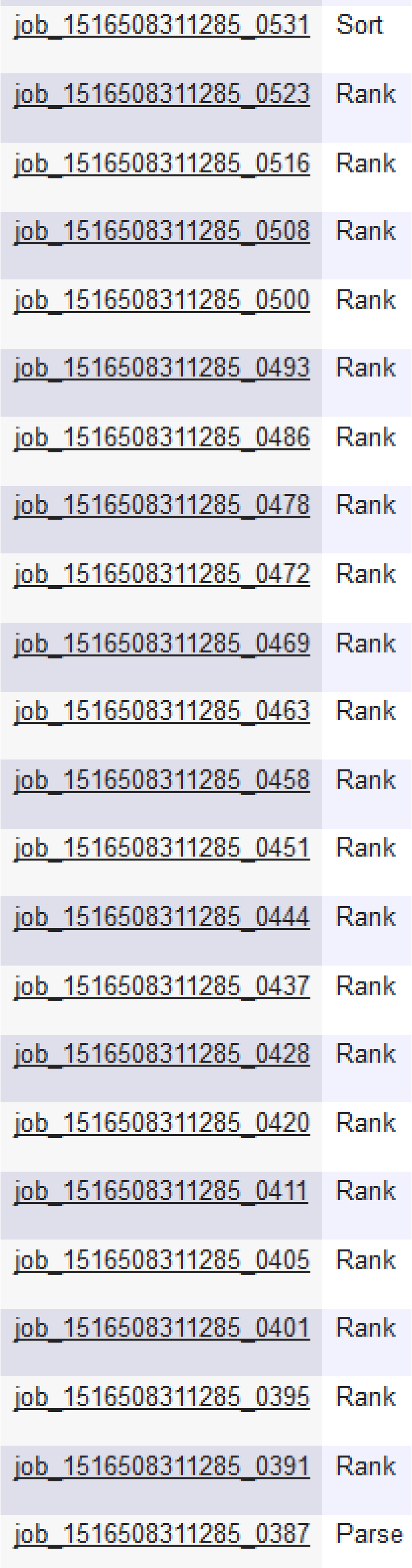
job_1516508311285_0216	Sort
job_1516508311285_0211	Rank
job_1516508311285_0206	Rank
job_1516508311285_0198	Rank
job_1516508311285_0192	Rank
job_1516508311285_0187	Rank
job_1516508311285_0181	Rank
job_1516508311285_0173	Rank
job_1516508311285_0167	Rank
job_1516508311285_0161	Rank
job_1516508311285_0155	Rank
job_1516508311285_0148	Rank
job_1516508311285_0142	Rank
job_1516508311285_0137	Rank
job_1516508311285_0131	Rank
job_1516508311285_0124	Rank
job_1516508311285_0118	Rank
job_1516508311285_0112	Rank
job_1516508311285_0103	Rank
job_1516508311285_0096	Rank
job_1516508311285_0089	Rank
job_1516508311285_0083	Rank
job_1516508311285_0077	Parse

8R



job_1516508311285_0376	Sort
job_1516508311285_0372	Rank
job_1516508311285_0368	Rank
job_1516508311285_0363	Rank
job_1516508311285_0358	Rank
job_1516508311285_0352	Rank
job_1516508311285_0348	Rank
job_1516508311285_0342	Rank
job_1516508311285_0337	Rank
job_1516508311285_0331	Rank
job_1516508311285_0326	Rank
job_1516508311285_0320	Rank
job_1516508311285_0315	Rank
job_1516508311285_0309	Rank
job_1516508311285_0303	Rank
job_1516508311285_0297	Rank
job_1516508311285_0291	Rank
job_1516508311285_0285	Rank
job_1516508311285_0281	Rank
job_1516508311285_0276	Rank
job_1516508311285_0269	Rank
job_1516508311285_0263	Rank
job_1516508311285_0258	Parse

16R



job_1516508311285_0531	Sort
job_1516508311285_0523	Rank
job_1516508311285_0516	Rank
job_1516508311285_0508	Rank
job_1516508311285_0500	Rank
job_1516508311285_0493	Rank
job_1516508311285_0486	Rank
job_1516508311285_0478	Rank
job_1516508311285_0472	Rank
job_1516508311285_0469	Rank
job_1516508311285_0463	Rank
job_1516508311285_0458	Rank
job_1516508311285_0451	Rank
job_1516508311285_0444	Rank
job_1516508311285_0437	Rank
job_1516508311285_0428	Rank
job_1516508311285_0420	Rank
job_1516508311285_0411	Rank
job_1516508311285_0405	Rank
job_1516508311285_0401	Rank
job_1516508311285_0395	Rank
job_1516508311285_0391	Rank
job_1516508311285_0387	Parse

32R