1004 hw4 Benchmarking Results

Rui Jiang, NetID: rj1407

11 april 2019

1 Introduction

A brief report of your benchmarking results:

1. A summary of all numerical results for each query/size/storage combination.
2. How do the results in parts 1, 2, and 3 compare?
3. What did you try in part 3 to improve performance for each query?
4. What worked, and what didn't work?

1.1 summary

(a) part1

queries	data files	minimum-time	median-time	maximum-time
csv_avg_income	small	0.7026236057281494	0.8502302169799805	8.965166807174683
csv_avg_income	medium	1.9216511249542236	2.551579713821411	8.645233154296875
csv_avg_income	large	8.115004539489746	8.895373344421387	19.3701651096344
csv_max_income	small	1.0988335609436035	1.580608606338501	7.058758020401001
csv_max_income	medium	1.166165828704834	1.3885726928710938	7.82256007194519
csv_max_income	large	7.743648529052734	8.693727016448975	19.286871671676636
csv_sue	small	0.06338071823120117	0.07914185523986816	6.200247764587402
csv_sue	medium	0.4113173484802246	0.44638872146606445	5.762178659439087
csv_sue	large	8.293389320373535	8.670412302017212	18.197821617126465

(b) part2

queries	data files	minimum-time	median-time	maximum-time
pq_avg_income	small	1.4893548488616943	2.4563770294189453	9.794914245605469
pq_avg_income	medium	2.6616287231445312	4.787017107009888	8.227185487747192
pq_avg_income	large	6.899017333984375	8.339593172073364	11.140564203262329 4
pq_max_income	small	1.446739673614502	2.0630996227264404	7.276565790176392
pq_max_income	medium	1.0488393306732178	1.7234728336334229	7.348182916641235
pq_max_income	large	4.913762331008911	7.8775954246521	11.524076223373413
pq_sue	small	0.06737303733825684	0.08722949028015137	1.1302649974822998
pq_sue	medium	0.14275407791137695	0.17364239692687988	1.1947340965270996
pq_sue	large	3.988520622253418	5.020178318023682	10.895919799804688

Compared to part1, minimum, median, and maximum time for each query on large data file significantly decrease, while all the time for each query on small and medium data file are not much different from results of part1 or even slightly increase.

(c) part3

queries	data files	minimum-time	median-time	maximum-time
pq_avg_income	small	0.5433940887451172	1.0724928379058838	4.228980302810669
pq_avg_income	medium	0.2626044750213623	0.4732506275177002	3.0550758838653564
pq_avg_income	large	4.364762544631958	5.477892875671387	9.059809684753418
pq_max_income	small	0.7000463008880615	1.21645188331604	4.37148380279541
pq_max_income	medium	0.5486409664154053	0.9072511196136475	5.967069149017334
pq_max_income	large	0.49623775482177734	1.2071934509277344	9.842505931854248
pq_sue	small	0.08102941513061523	0.09958171844482422	0.699988603591919
pq_sue	medium	0.06943058967590332	0.08196449279785156	0.6407155990600586
pq_sue	large	0.506645679473877	0.6634480953216553	1.951178789138794

The above result is the decreased running time after optimization.
(1)For 'avg_income', I first sort the dataframe by 'zipcode', for all three data files, I set num_partitions as 100 and partition column 'zipcode' and improved the performance, however when I set num_partitions as 50 or 250, it didn't work;
(2)for 'max_income', I first sort the dataframe by both 'last_name' and 'income', and then I set num_partitions as 50 and partition column 'last_name' and improved the performance significantly especially for large data file, however when I set num_partitions to 100 it didn't work;
(3)for 'sue', I tried: sort the dataframe by both 'first_name' and 'income', set num_partitions to 5 and improved the performance significantly, I also tried setting num_partitions to 500, and it became very slow.

If I only change the HDFS replication factor to 1, then the result would be as below:

queries	data files	minimum-time	median-time	maximum-time
pq_avg_income	small	1.3780009746551514	1.942859411239624	3.86772084236145
pq_avg_income	medium	0.2626044750213623	0.4732506275177002	3.0550758838653564
pq_avg_income	large	9.797628164291382	11.318755865097046	19.90740132331848
pq_max_income	small	1.445481300354004	2.003634214401245	5.822409152984619
pq_max_income	medium	1.4722049236297607	6.447839975357056	14.472556829452515
pq_max_income	large	9.175496578216553	12.592344522476196	21.682058095932007
pq_sue	small	0.06980419158935547	0.09876894950866699	1.1285512447357178
pq_sue	medium	0.13917946815490723	0.16200971603393555	0.6746749877929688
pq_sue	large	4.855283498764038	5.851213455200195	8.306337356567383

As we can see from the table, all the running time for small data file seems to decrease a little(not very significant),but the running time for medium and large data files(especially large) increases significantly.