Part1:CSV

Query1:

1. people_small:

Times= [2.8261451721191406, 2.5780556201934814, 2.9060657024383545, 2.1227662563323975, 1.6604135036468506, 1.9731497764587402, 2.1690430641174316, 2.408829689025879, 1.5697500705718994, 2.004345655441284, 1.9895951747894287, 2.4438366889953613, 1.99562668800354, 2.364877700805664, 2.792221784591675, 2.1269426345825195, 1.9278030395507812, 2.504514455795288, 2.5396158695220947, 2.656501531600952, 7.676696538925171, 2.2416865825653076, 1.949371099472046, 2.5807223320007324, 1.7884860038757324]

Maximum-7.676696538925171

Median-2.2416865825653076

Minimum-1.5697500705718994

2. people_medium:

Times= [8.795993566513062, 2.7709062099456787, 3.459489345550537, 2.73333477973938, 2.774613380432129, 3.3857572078704834, 1.0863640308380127, 1.3073811531066895, 1.2147548198699951, 1.4942514896392822, 1.4146785736083984, 1.17466139793396, 1.512695074081421, 1.3782105445861816, 1.2148373126983643, 1.0674138069152832, 1.3006441593170166, 1.1299054622650146, 1.063173532485962, 1.5997936725616455, 1.5673189163208008, 2.293553590774536, 2.5886282920837402, 2.1696722507476807, 1.4427571296691895]

Maximum-8.795993566513062

Median-1.4942514896392822

Minimum-1.0674138069152832

3. people_large:

Times= [19.078083515167236, 10.30107045173645, 9.355169534683228, 9.338501930236816, 8.984233140945435, 9.237159013748169, 10.175524711608887, 9.495015859603882, 9.524571418762207, 9.553520917892456, 9.483301877975464, 10.385676383972168, 9.104756116867065, 9.145282983779907, 9.667745590209961, 9.484043836593628, 9.55661392211914, 9.553150177001953, 9.76085090637207, 9.67215609550476, 9.33333706855774, 10.090212106704712, 11.05758547782898, 9.267585039138794, 8.930577278137207]

Maximum-19.078083515167236

Median-9.524571418762207

Minimum-8.930577278137207

Query2:

1. people_small

Times= [1.0506083965301514, 0.8749825954437256, 0.8942551612854004, 0.8913588523864746, 0.8842654228210449, 0.8827142715454102, 0.748884916305542, 0.889495849609375, 0.770775556564331, 0.7891192436218262, 0.8161303997039795, 0.679218053817749, 0.6575963497161865, 0.8129754066467285, 0.7395892143249512, 0.7507965564727783, 0.7010273933410645, 0.6588602066040039, 0.7727901935577393, 0.7998287677764893, 0.6778042316436768, 0.7664315700531006, 0.7043015956878662, 0.7833249568939209, 0.9410476684570312]

Minimum: 0.6575963497161865

Maximum: 1.0506083965301514

Median:0.7833249568939209

2.Medium_people

Times= [9.511051654815674, 1.6432874202728271, 1.4081990718841553, 1.2930700778961182, 1.4013268947601318, 1.505082368850708, 1.493194818496704, 1.4803965091705322, 1.4171562194824219, 1.3319523334503174, 1.4879350662231445, 1.5465431213378906, 1.3161051273345947, 1.265610933303833, 1.3680527210235596, 1.3460164070129395, 1.329399585723877, 1.4311659336090088, 1.780893087387085, 1.999767780303955, 4.957162380218506, 2.5755615234375, 4.0908942222595215, 2.7740116119384766, 1.25608491897583]

Minimum: 1.25608491897583

Maximum: 9.511051654815674

Median: 1.4803965091705322

3.Large_people:

Times= [25.840360164642334, 10.262852907180786, 9.765518426895142, 10.374146223068237, 10.30000638961792, 13.18604588508606, 16.960474252700806, 9.894081354141235, 10.886994361877441, 10.131798505783081, 10.247770547866821, 14.541703224182129, 11.092097520828247, 10.214712858200073, 13.824081182479858, 10.28542709350586, 10.13494086265564, 12.030067920684814, 10.500078201293945, 9.775466918945312, 12.287995338439941, 10.458235740661621, 10.528601169586182, 10.343077421188354, 10.178638935089111]

Minimum: 9.765518426895142

Maximum: 25.840360164642334

Median: 10.374146223068237

Query3:

1. people_small

Times= [7.05475378036499, 0.22829842567443848, 0.17508840560913086, 0.13164997100830078, 0.17127227783203125, 0.11963224411010742, 0.11418318748474121, 0.11794686317443848, 0.11147952079772949, 0.11405348777770996, 0.10207629203796387, 0.09774065017700195, 0.09884023666381836, 0.1120758056640625, 0.0943448543548584, 0.10058474540710449, 0.10610675811767578, 0.08982324600219727, 0.09073257446289062, 0.09618234634399414, 0.08459186553955078, 0.24522948265075684, 0.08231139183044434, 0.08487749099731445, 0.0812673568725586]

Minimum: 0.0812673568725586

Maximum: 7.05475378036499

Median: 0.10610675811767578

2. people_medium

Times= [1.133746862411499, 0.8042430877685547, 0.7776443958282471, 0.770451545715332, 0.7428460121154785, 0.7740330696105957, 0.7508988380432129, 0.7643604278564453, 0.7319583892822266, 0.7494640350341797, 0.7258727550506592, 0.762136697769165, 0.7237794399261475, 0.7385966777801514, 0.722480058670044, 0.7232637405395508, 0.722722053527832, 0.7319579124450684, 0.7223720550537109, 0.7295448780059814, 0.7440962791442871, 0.7904965877532959, 0.7307436466217041, 0.7179923057556152, 0.8539345264434814]

Minimum: 0.7179923057556152

Maximum: 1.133746862411499

Median: 0.7428460121154785

4. people_large:

Times= [69.66668367385864, 70.02246284484863, 66.16232252120972, 13.248416423797607, 9.847792387008667, 10.10477089881897, 9.019025087356567, 9.065463781356812, 9.033227443695068, 8.893892765045166, 9.15992021560669, 10.397667407989502, 8.900520324707031, 9.332217931747437, 8.611953973770142, 9.114089965820312, 9.380403757095337, 8.304866552352905, 9.284451246261597, 11.314353466033936, 8.782903909683228, 8.92414927482605, 8.673853158950806, 9.610411882400513, 9.219095468521118]

Minimum: 8.304866552352905

Maximum: 70.02246284484863

Median: 9.219095468521118

Part2:

Parquet

Query1:

1.Small_people:

Times=[3.592094659805298, 1.9111311435699463, 4.660661220550537, 2.0677990913391113, 3.5173211097717285, 1.8251233100891113, 2.011431932449341, 4.5094568729400635, 2.1988840103149414, 4.888187885284424, 3.3914294242858887, 4.6223742961883545, 2.054983139038086, 1.9899849891662598, 1.8527131080627441, 2.130173683166504, 1.8972604274749756, 1.9640090465545654, 1.7503914833068848, 2.2390127182006836, 2.1655287742614746, 1.9177300930023193, 1.556185007095337, 1.816366195678711, 2.085245132446289]

Minimum: 1.556185007095337

Maximum: 4.888187885284424

Median: 2.0677990913391113

2. Medium_people

Times= [9.346247673034668, 1.8039329051971436, 1.824580192565918, 1.4068779945373535, 1.1421396732330322, 1.2987914085388184, 1.4536137580871582, 1.5323312282562256, 1.2668023109436035, 1.370760440826416, 1.693920612335205, 2.5058066844940186, 2.7750449180603027, 1.7230358123779297, 1.567136526107788, 1.3037102222442627, 1.4229991436004639, 2.3873798847198486, 2.145026445388794, 1.9222254753112793, 1.8563637733459473, 1.5956859588623047, 3.241365432739258, 1.135662317276001, 1.0719661712646484]

Minimum: 1.0719661712646484

Maximum: 9.346247673034668

Median: 1.5956859588623047

3.large_people

Times=[14.528090476989746, 12.749746084213257, 13.446943044662476, 13.492405652999878, 9.622485160827637, 14.4599928855896, 11.202822208404541, 9.336459636688232, 7.29622483253479, 6.984270334243774, 10.606008291244507, 7.66433572769165, 12.285788297653198, 13.638594150543213, 11.67634916305542, 7.658676385879517, 7.5294189453125, 6.667465448379517, 6.978668212890625, 7.403420448303223, 6.3388991355896, 7.1837921142578125, 6.984291076660156, 7.636161804199219, 6.574344158172607]

Minimum: 6.3388991355896

Maximum: 14.528090476989746

Median: 7.66433572769165

Query2:

1.Small_people:

Times=[3.8122024536132812, 1.5559861660003662, 1.3371710777282715, 1.6613092422485352, 1.5331988334655762, 1.6793580055236816, 1.5916380882263184, 1.4866786003112793, 1.2759816646575928, 1.446500539779663, 1.8436288833618164, 1.9657549858093262, 2.0847690105438232, 4.7889440059661865, 2.1130661964416504, 1.8926260471343994, 1.3903231620788574, 1.3558573722839355, 1.4109783172607422, 1.3311429023742676, 1.337981939315796, 1.405228853225708, 1.33198881149292, 1.5545167922973633, 1.2371079921722412]

Minimum: 1.2371079921722412

Maximum: 4.7889440059661865

Median: 1.5331988334655762

2. Medium_people

Times=[6.5427374839782715, 4.958266973495483, 3.760910749435425, 1.332059383392334, 1.2370519638061523, 1.345959186553955, 1.4916555881500244, 1.4981815814971924, 1.4861936569213867, 1.5185487270355225, 1.1992161273956299, 1.172914743423462, 1.6910371780395508, 1.1840732097625732, 1.3319125175476074, 1.1599645614624023, 1.433272123336792, 1.503483772277832, 1.244389533996582, 1.1341590881347656, 1.2096319198608398, 1.2823021411895752, 1.4256184101104736, 1.3948116302490234, 1.3791909217834473]

Minimum: 1.1341590881347656

Maximum: 6.5427374839782715

Median: 1.3791909217834473

3.large_people

Times=[9.880967378616333, 7.867677927017212, 6.521807432174683, 6.847893953323364, 8.107550859451294, 5.791763067245483, 5.982555627822876, 6.208383321762085, 5.885224342346191, 5.3721888065338135, 8.457029104232788, 6.161182880401611, 6.15587043762207, 6.140912771224976, 5.631833553314209, 5.756205797195435, 5.93156361579895, 5.808679819107056, 5.836531639099121, 4.980981111526489, 5.80924654006958, 8.629687309265137, 7.028022050857544, 7.918824195861816, 6.044981479644775]

Minimum: 4.980981111526489

Maximum: 9.880967378616333

Median: 6.140912771224976

Query3:

1.Small_people:

Times= [1.0397896766662598, 0.11899733543395996, 0.09272456169128418, 0.09023284912109375, 0.09560036659240723, 0.08410000801086426, 0.0939018726348877, 0.0838627815246582, 0.08351516723632812, 0.08820652961730957, 0.07738733291625977, 0.08656167984008789, 0.0769190788269043, 0.07941126823425293, 0.07668304443359375, 0.07225775718688965, 0.11474990844726562, 0.07274651527404785, 0.07215237617492676, 0.07615900039672852, 0.06850790977478027, 0.07089805603027344, 0.08357453346252441, 0.06939029693603516, 0.06551122665405273]

Minimum: 0.06551122665405273

Maximum: 1.0397896766662598

Median: 0.08351516723632812

2. Medium_people

Times=[0.8759422302246094, 0.14467787742614746, 0.14624643325805664, 0.15556788444519043, 0.14722561836242676, 0.14063453674316406, 0.12961792945861816, 0.14082837104797363, 0.1407485008239746, 0.12273573875427246, 0.134857177734375, 0.12160205841064453, 0.11947488784790039, 0.11557674407958984, 0.13402009010314941, 0.1257646083831787, 0.13229036331176758, 0.12276697158813477, 0.1484978199005127, 0.13206720352172852, 0.11611127853393555, 0.12091755867004395, 0.12758851051330566, 0.11504530906677246, 0.22190570831298828]

Minimum: 0.11504530906677246

Maximum: 0.8759422302246094

Median: 0.13229036331176758

3.large_people

Times= [8.650033473968506, 6.790310859680176, 6.998730421066284, 6.4841296672821045, 6.690676927566528, 5.913714408874512, 5.9656853675842285, 6.283844470977783, 6.314756155014038, 5.893944978713989, 5.962340831756592, 5.902162790298462, 5.854931592941284, 5.962604761123657, 5.8770976066589355, 5.940180540084839, 5.995205402374268, 5.959359407424927, 5.844036340713501, 5.835162878036499, 5.849179267883301, 5.966971397399902, 6.0410215854644775, 6.335839033126831, 6.023918628692627]

Minimum: 5.835162878036499

Maximum: 8.650033473968506

Median: 5.9656853675842285

Part 3

I tested 3 ways:

1) Sorting based on zipcode

2) Changing HDFS replication factor from 3 to 1

3) Dataframe repartition (10)

1)Sorting based on zipcode:

Query1:

For small dataset:

[11.137639999389648, 2.083547592163086, 3.2870595455169678, 1.7365531921386719, 1.2080597877502441, 1.2443194389343262, 1.213925838470459, 1.1487109661102295, 1.712317705154419, 1.142817497253418, 1.11812162399292, 1.178238868713379, 1.2395846843719482, 0.9973793029785156, 1.0685033798217773, 1.0679552555084229, 1.4928827285766602, 1.3688304424285889, 1.1982474327087402, 1.2228083610534668, 1.221083164215088, 1.6081087589263916, 1.2057652473449707, 1.6235270500183105, 1.2815673351287842]

Minimum: 0.9973793029785156
Maximum: 11.137639999389648
Median: 1.2228083610534668

For Medium dataset:

[6.385270833969116, 3.9873952865600586, 2.944810152053833, 2.257990837097168, 2.673874616622925, 2.8285632133483887, 2.425804376602173, 2.184701442718506, 2.715941905975342, 2.588367223739624, 2.410320520401001, 2.150815010070801, 2.570774793624878, 2.3815090656280518, 2.324833631515503, 1.8954219818115234, 2.8439786434173584, 2.3351352214813232, 2.391603946685791, 2.0073800086975098, 2.552232265472412, 2.32342791557312, 2.0846962928771973, 2.3841280937194824, 2.7838151454925537]

Minimum: 1.8954219818115234
Maximum: 6.385270833969116
Median: 2.410320520401001

For large dataset:

[13.032573938369751, 5.328117370605469, 5.496618032455444, 7.533630847930908, 7.729750633239746, 11.664348363876343, 8.334502220153809, 8.454636096954346, 9.556930541992188, 7.986032724380493, 4.440638065338135, 4.794765949249268, 7.914449214935303, 6.240906000137329, 6.6278486251831055, 7.906900644302368, 7.085848569869995, 11.048851013183594, 13.7359778881073, 7.382605314254761, 5.582719087600708, 6.405520439147949, 4.897575378417969, 6.362250089645386, 6.884901523590088]

Mean deviation: 1.81569903259

Minimum: 4.440638065338135

Maximum: 13.7359778881073

Median: 7.382605314254761

Query2:

For small dataset:

[10.419291973114014, 9.187857866287231, 4.991171598434448, 4.7162017822265625, 5.4911439418792725, 10.137105464935303, 4.540773391723633, 5.530250310897827, 5.5322792530059814, 5.155280828475952, 8.048673391342163, 6.259162902832031, 6.229900598526001, 5.924079656600952, 6.854030132293701, 5.208340883255005, 4.8773674964904785, 5.576457500457764, 5.242650032043457, 5.762898921966553, 5.239511966705322, 4.622494697570801, 4.603212833404541, 5.749209403991699, 6.62775182723999]

Minimum: 4.540773391723633
Maximum: 10.419291973114014
Median: 5.5322792530059814

For Medium dataset:

[12.18917465209961, 10.797689437866211, 6.626230239868164, 8.917567014694214, 9.261852264404297, 5.56609320640564, 5.749328851699829, 4.4820380210876465, 6.86072564125061, 5.352772235870361, 5.001183032989502, 6.074818134307861, 6.352311611175537, 5.436468124389648, 7.1578919887542725, 6.950881242752075, 4.739631175994873, 4.044867753982544, 6.427323341369629, 6.9934165477752686, 4.618765354156494, 5.290529727935791, 6.153432369232178, 8.098942518234253, 5.396251201629639]

Minimum: 4.044867753982544
Maximum: 12.18917465209961
Median: 6.153432369232178

For large dataset:

[15.636760950088501, 9.895964860916138, 9.94747281074524, 7.364735841751099, 7.825048208236694, 7.1292946338653564, 10.661897659301758, 8.416216850280762, 8.227037906646729, 7.118448495864868, 7.941537857055664, 7.91771936416626, 9.56234335899353, 9.183664083480835, 8.756763219833374, 8.489778757095337, 7.7220752239227295, 8.049550294876099, 6.209645748138428, 8.588666677474976, 6.4813315868377686, 10.861623048782349, 8.187833309173584, 15.55777883529663, 10.432417154312134]

Minimum: 6.209645748138428

Maximum: 15.636760950088501

Median: 8.416216850280762

Query3:

For small dataset:

[4.634995698928833, 4.0617711544036865, 3.764634370803833, 4.697351694107056, 4.24166464805603, 3.5246922969818115, 3.8021059036254883, 4.461017370223999, 3.264220952987671, 3.8329036235809326, 3.894723892211914, 3.8013978004455566, 3.8695924282073975, 3.8524444103240967, 3.8444716930389404, 3.827620267868042, 3.839717149734497, 3.8351240158081055, 3.919407844543457, 3.796341896057129, 3.875028133392334, 3.946566343307495, 3.764228582382202, 3.869180917739868, 3.8593735694885254]

Minimum: 3.264220952987671
Maximum: 4.697351694107056
Median: 3.8524444103240967

For Medium dataset:

[0.6800203323364258, 1.8442630767822266, 0.7122209072113037, 1.8035426139831543, 0.5848767757415771, 0.5645334720611572, 0.448683500289917, 1.228987455368042, 0.6256527900695801, 0.3667316436767578, 0.5789785385131836, 0.39589738845825195, 0.4582052230834961, 3.7809863090515137, 0.65932297706604, 0.49038219451904297, 0.48612451553344727, 0.440044641494751, 3.9084770679473877, 1.3314487934112549, 0.8817832469940186, 1.1736531257629395, 0.7032301425933838, 0.5613477230072021, 0.3868255615234375]

Minimum: 0.3667316436767578
Maximum: 3.9084770679473877
Median: 0.6256527900695801

For large dataset:

[3.9520273208618164, 5.635193824768066, 4.882921457290649, 5.138877868652344, 4.223238468170166, 3.8494672775268555, 4.637498617172241, 4.238116979598999, 5.059812068939209, 5.211129426956177, 4.684208393096924, 5.413644313812256, 3.9788386821746826, 3.7359426021575928, 3.940619468688965, 4.621798515319824, 5.867504119873047, 4.835549831390381, 4.670418739318848, 4.780267715454102, 4.559459447860718, 4.744747638702393, 4.033041000366211, 5.080227851867676, 6.595410108566284]

Minimum: 3.7359426021575928

Maximum: 6.595410108566284

Median: 4.684208393096924

2. Changing HDFS replication factor from 3 to 1:

Query1:

For small dataset:

[4.507534980773926, 2.118612766265869, 4.901369333267212, 1.8856725692749023, 2.0644423961639404, 1.8840899467468262, 1.6036872863769531, 5.014885425567627, 1.8475189208984375, 2.050931453704834, 5.452807903289795, 1.5166635513305664, 4.8572328090667725, 2.0510308742523193, 5.252667427062988, 5.88262152671814, 1.6510372161865234, 2.1132912635803223, 1.740708827972412, 1.5527582168579102, 1.8898730278015137, 2.000495672225952, 3.805272340774536, 2.148963451385498, 3.601222038269043]

Minimum: 1.5166635513305664

Maximum: 5.88262152671814

Median: 2.0644423961639404

For Medium dataset:

[5.67313027381897, 3.8859028816223145, 2.4685263633728027, 3.3557355403900146, 1.3599953651428223, 1.2017176151275635, 1.130974292755127, 1.5731120109558105, 1.2411563396453857, 0.8820958137512207, 1.7394514083862305, 1.005722999572754, 1.3826274871826172, 0.9217913150787354, 1.4889700412750244, 0.996476411819458, 1.3759784698486328, 1.4848556518554688, 1.1532464027404785, 0.9977939128875732, 0.8630383014678955, 1.1587820053100586, 1.3336424827575684, 1.3340001106262207, 0.992485761642456]

Minimum: 0.8630383014678955

Maximum: 5.67313027381897

Median: 1.3336424827575684

For large dataset:

[14.145169973373413, 12.802163124084473, 11.112159013748169, 15.328082799911499, 10.346418142318726, 9.46509051322937, 10.958763837814331, 9.351040840148926, 13.353733777999878, 10.711897611618042, 8.082474708557129, 8.555066108703613, 9.005520820617676, 11.191061735153198, 9.49665117263794, 14.22213625907898, 10.613188028335571, 8.529667854309082, 8.99878978729248, 10.311579465866089, 9.666844367980957, 9.544612407684326, 10.727580308914185, 9.723974704742432, 9.45245099067688]

Minimum: 8.082474708557129

Maximum: 15.328082799911499

Median: 10.311579465866089

Query2:

For small dataset:

[5.7764732837677, 2.347487211227417, 5.167581796646118, 6.38420033454895, 4.847326993942261, 6.326218605041504, 2.479163646697998, 6.087578535079956, 3.8443615436553955, 2.1496222019195557, 1.7486631870269775, 1.9082117080688477, 4.332461833953857, 1.498640537261963, 1.985051155090332, 1.9691226482391357, 2.371938705444336, 1.5243475437164307, 4.615681171417236, 2.3984668254852295, 1.587597370147705, 2.164006471633911, 1.3741686344146729, 1.7670557498931885, 8.577851057052612]

Minimum: 1.3741686344146729

Maximum: 8.577851057052612

Median: 2.371938705444336

For Medium dataset:

[0.9926819801330566, 1.0284347534179688, 0.9827566146850586, 0.7857859134674072, 0.7155485153198242, 0.8000729084014893, 1.3794009685516357, 0.8881971836090088, 1.1212084293365479, 0.7855956554412842, 0.9843478202819824, 0.9207019805908203, 1.1094245910644531, 1.196664810180664, 0.8765110969543457, 1.399653434753418, 0.7476305961608887, 0.9248461723327637, 0.6960132122039795, 1.335096836090088, 1.2284488677978516, 1.0159282684326172, 0.8310256004333496, 0.8747632503509521, 0.9960291385650635]

Minimum: 0.6960132122039795

Maximum: 1.399653434753418

Median: 0.9827566146850586

For large dataset:

[9.258645057678223, 10.016647577285767, 10.090998411178589, 6.114313125610352, 7.160630941390991, 6.849953889846802, 6.746578216552734, 6.794080495834351, 6.867067098617554, 6.135220766067505, 7.073768138885498, 6.767061710357666, 6.185437917709351, 7.153445720672607, 6.747191667556763, 7.532978773117065, 7.179020881652832, 8.122924566268921, 7.093442916870117, 6.601865530014038, 7.63100266456604, 8.874868392944336, 8.590531826019287, 8.338441371917725, 10.236039161682129]

Minimum: 6.114313125610352

Maximum: 10.236039161682129

Median: 7.153445720672607

Query3:

For small dataset:

[1.2610692977905273, 0.3719937801361084, 2.536013126373291, 0.8930716514587402, 0.14301538467407227, 0.08563613891601562, 0.15728116035461426, 0.9764659404754639, 0.12989068031311035, 0.10313582420349121, 2.4998650550842285, 1.3761630058288574, 2.1011788845062256, 2.0195326805114746, 0.1139523983001709, 0.11639904975891113, 0.1546945571899414, 0.18081355094909668, 1.358583688735962, 0.07890009880065918, 0.09893441200256348, 0.09642982482910156, 0.21034526824951172, 0.14525485038757324, 0.161786317821738]

Minimum: 0.07890009880065918

Maximum: 2.536013126373291

Median: 0.16178631782531738

For Medium dataset:

[0.197361230821973, 0.2031252384185791, 0.10891866683959961, 0.10550475120544434, 0.1756305694580078, 0.10602688789367676, 0.09562897682189941, 0.08991312980651855, 0.1028597354888916, 0.09114837646484375, 0.09412479400634766, 0.08963274955749512, 0.121609922375488, 0.09138655662536621, 0.09739494323730469, 0.09639573097229004, 0.09888863563537598, 0.09420728683468, 0.104617834091186, 0.09255409240722656, 0.08897995948791504, 0.093356609348242, 0.09199285507202148, 0.08977413177490234, 0.1121821403503418]

Minimum: 0.08897995948791504

Maximum: 0.2031252384185791

Median: 0.09639573097229004

For large dataset:

[7.39433479309082, 2.4410183429718018, 1.48220419883723, 1.57194781303476, 1.5911743640899658, 1.6479272842407227, 5.022326469421387, 4.606472492218018, 7.766163349151611, 15.612988471984863, 8.956841230392456, 9.028645277023315, 8.778779029846191, 8.87227749824524, 8.977166891098022, 8.826448917388916, 8.811817407608032, 8.932800054550171, 9.935317516326904, 8.933541297912598, 8.977555513381958, 10.290614846853, 8.731968402862549, 9.49358868598938, 8.332775115966797]

Minimum: 1.4822041988372803

Maximum: 15.612988471984863

Median: 8.811817407608032

3.Datframe repartition(10):

Query1:

For small dataset:

[6.286993980407715, 5.295557975769043, 5.014646053314209, 6.4580771923065186, 4.669739484786987, 4.5722129344940186, 5.068404674530029, 4.9424707889556885, 5.0019402503967285, 4.751055717468262, 5.339811563491821, 4.466461420059204, 7.413817882537842, 4.696871042251587, 4.424874305725098, 5.345811367034912, 4.71056604385376, 5.0082478523254395, 5.0370190143585205, 4.547041654586792, 5.31323127821045, 5.3311619750596, 5.404607534408569, 5.406122207641602, 4.595912456512451]

Minimum: 4.424874305725098

Maximum: 7.413817882537842

Median: 5.014646053314209

For Medium dataset:

[6.775264024734497, 5.300170183181763, 7.560017347335815, 6.691128492355347, 5.300795793533325, 4.580033779144287, 5.049131631851196, 4.794148206710815, 4.968419313430786, 5.353922605514526, 5.567997217178345, 5.0336267948150635, 5.235072135925293, 5.757943391799927, 5.345193147659302, 5.697831869125366, 4.588623762130737, 4.898461818695068, 4.842574596405029, 5.1246654987335205, 5.302778005599976, 5.253861427307129, 4.822391510009766, 5.19081807136564, 4.622339487075806]

Minimum: 4.580033779144287

Maximum: 7.560017347335815

Median: 5.235072135925293

For large dataset:

[11.2881278991622, 8.30640435218811, 8.34809684753418, 8.923076391220093, 8.768110990524292, 6.64319109916687, 10.467523097991943, 7.944965124130249, 7.224200248718262, 6.61367130279541, 8.152810335159302, 6.487229824066162, 7.2294371128082275, 6.658957242965698, 7.2759690284729, 6.914740085601807, 6.647405624389648, 6.052124261856079, 6.702977418899536, 6.308215856552124, 5.772991418838501, 6.228223562240601, 7.514663219451904, 6.2736852169036865, 6.1807444561157]

Minimum: 5.772991418838501

Maximum: 11.288127899169922

Median: 6.914740085601807

Query2:

For small dataset:

[4.191436529159546, 4.800817012786865, 5.134759187698364, 4.984867572784424, 4.722151756286621, 4.919852256774902, 4.835162162780762, 4.077932596206665, 4.727035999298096, 5.1439878940582, 5.413954496383667, 5.7935521602630, 5.4985997676849365, 8.024978876113892, 8.199605941772461, 9.047823429107666, 5.495365142822266, 5.979604482650, 4.3860731124877, 5.956125020980835, 5.811970949172974, 5.5333120822906, 5.0809290409088135, 4.9295899868011475, 3.5712370872756]

Minimum: 3.571237087249756

Maximum: 9.047823429107666

Median: 5.134759187698364

For Medium dataset:

[4.7241251465845, 4.859148025512695, 5.0539128780399, 4.9965276718165, 4.472257137298584, 5.4653127145093, 4.532566785812378, 5.349400043487549, 4.828309774398804, 5.5812488641968, 4.33367800758545, 5.19016337394736, 5.289344549179077, 4.2094521552197, 5.752646446228027, 5.757079839706421, 5.376992702484131, 5.43952703595215, 4.499986410140991, 4.855646610201, 4.6160533819214, 5.166670799255371, 4.9297381575928, 4.86831068992675, 4.964639663696289]

Minimum: 4.209452152252197

Maximum: 5.757079839706421

Median: 4.964639663696289

For large dataset:

[7.4364566802972, 6.39669132232666, 7.0680360706738, 6.9454419612882, 7.054111957550049, 6.977774143218994, 7.04552388192314, 6.46998834609535, 7.1612796951294, 6.1836667085205, 6.82288670539856, 6.2794699668428, 6.565807819366455, 6.311746597290039, 6.571455240249634, 6.329586505889893, 7.0022242069244385, 6.51337508690186, 5.9221043673096, 6.268640756607056, 7.100996494293213, 6.41338849067688, 6.0118508637085, 6.958583831787109, 6.7616223220825]

Minimum: 5.922104358673096

Maximum: 7.4364566802852

Median: 6.571455240249634

Query3:

For small dataset:

[4.0162632436255, 4.34418785095215, 3.3935136795043945, 3.8071630618286, 3.9825568199157715, 3.883819103240967, 3.921590805053711, 3.95148277281484, 3.92303870845947, 3.9374704360961914, 3.926884412765503, 3.93872666358975, 3.93407559483643, 3.9593920707702637, 4.5410296996167, 3.300466775894165, 3.9437897930908, 3.940541982650757, 3.9355006217956543, 3.9442718029022217, 3.93705272656055, 3.932429075241089, 3.9690380096435547, 3.91624450659375, 3.94045400619684]

Minimum: 3.300466775894165

Maximum: 4.541029691696167

Median: 3.9374704360961914

For Medium dataset:

[2.232896327972412, 0.481952969072, 0.40932750701904297, 0.33498191833496094, 0.2955892086029053, 0.30319166187168, 1.038245211098, 0.286045781862793, 0.28481261291504, 0.31280183711426, 0.78458571434021, 0.26830124855041504, 0.2554035186767578, 0.278782367729883, 0.246831655231934, 0.2445833683013916, 0.2704913618042, 0.3713676929473877, 0.241287321472168, 0.268621683172754, 0.2563195857666, 0.253469467308594, 0.234811541833496, 0.238836765280664, 0.2250993251800537]

Minimum: 0.2250993251800537

Maximum: 2.232896327972412

Median: 0.27878236770629883

For large dataset:

[11.128163337752, 9.465143442193, 8.1165971755145, 7.763831615447998, 7.45205020904541, 7.03240752201538, 7.9492356777116, 7.9270765140259, 7.7502214909985, 6.9397330284865, 7.814648151397705, 7.03085803985957, 6.941550254821777, 7.004456281661987, 6.9408798217344, 8.733825445175171, 7.125668048858643, 7.805485458374, 7.5434277057647705, 9.443923473358154, 7.0969834329775, 7.803034543991089, 8.7500257492043, 7.774768590927124, 8.244042696399]

Minimum: 6.939733028411865

Maximum: 11.12816333770752

Median: 7.774768590927124

Comparison Between result from part 1,2,3: (comparing based on median)

Query1:

Small_people:

Sorting based on zipcode(Part3) works the best here

Medium_people:

CSV file (Part1) and sorting based on zipcode (Part3) works best here (there is very little difference between csv file and sorting with zipcode)

Large_people:

Sorting based on zipcode(Part3) works the best here .

So overall Part3: sorting based on zipcode works best for query 1

Query2:

Small_people:

CSV file (Part1) works the best here

Medium_people:

Changing Hdfs replication factor (Part3) works best here.

Large_people:

Parquet (Part2) and dataframe repartition works (Part3) the best here.

So overall for different dataset different type of file optimization works best for query 2

Query3:

Small_people:

Parquet (Part2), CSV (part1) and Changing Hdfs replication factor (Part3) the best here.

Medium_people:

Changing Hdfs replication factor (Part3) works best here.

Large_people:

Sorting based on zipcode (Part3) works the best here.

So overall for different dataset different type of file optimization works best for query 3 but mostly changing

Hdfs replication factor (Part3) works best.

Soall three optimization technique used in part3 worked but for different dataset and different query.