https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-large.csv.bz2.

Simple Linear Regression

- Calculate the m slope and the b intercept for the small dataset

Slope (m): 2.698633750152394 *Intercept (b): 3.970225230889111*

- Calculate the m slope and the b intercept for the large dataset

Slope (m): 2.6656349732213727 *Intercept (b): 4.532229802078271*

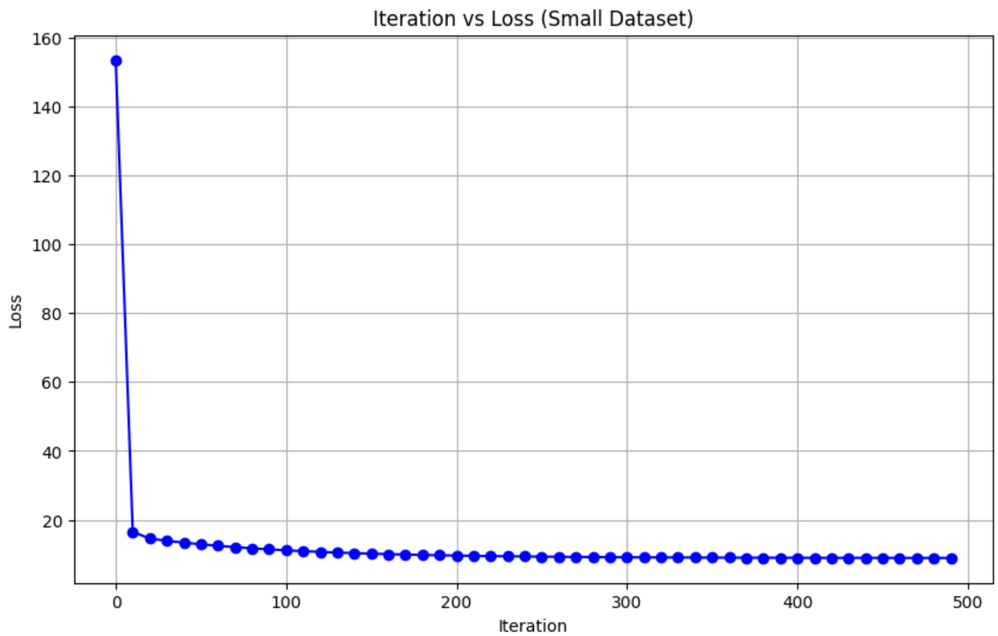
https://storage.googleapis.com/met-cs-777-data/taxi-data-sorted-large.csv.bz2.

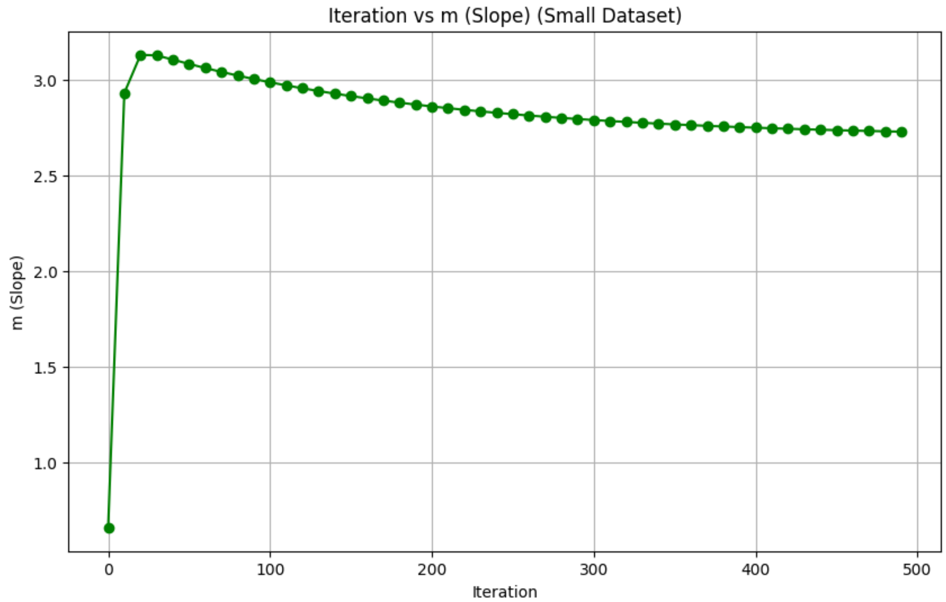
Additional Task using BatchGradientDescent and Desired Iterations and LR for Testing Aim

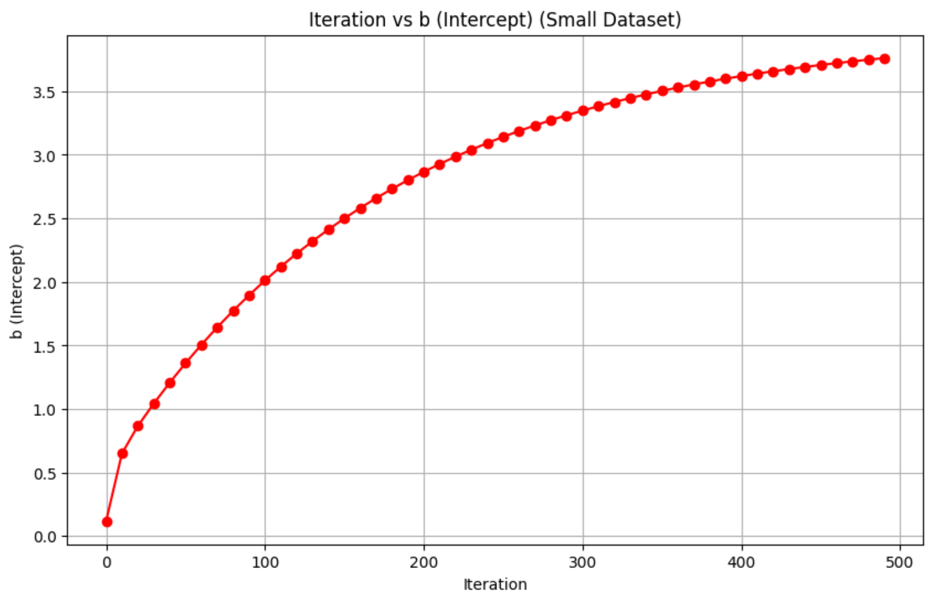
- Calculate the m slope and the b intercept for the small dataset

Final parameters for small dataset (iteration=2000, learning_rate = 0.01): *Slope (m): 2.6986391293785004* *Intercept (b): 3.9701887413601833* *Approximately constant loss after 2000 iterations.*

Final parameters for small dataset (iteration=1000, learning_rate = 0.0001): *Slope (m): 2.780701661507673* *Intercept (b): 0.608503802919607* *This needs to more time and bigger Alpha like 0.001 to decrease the Loss to get stable value*





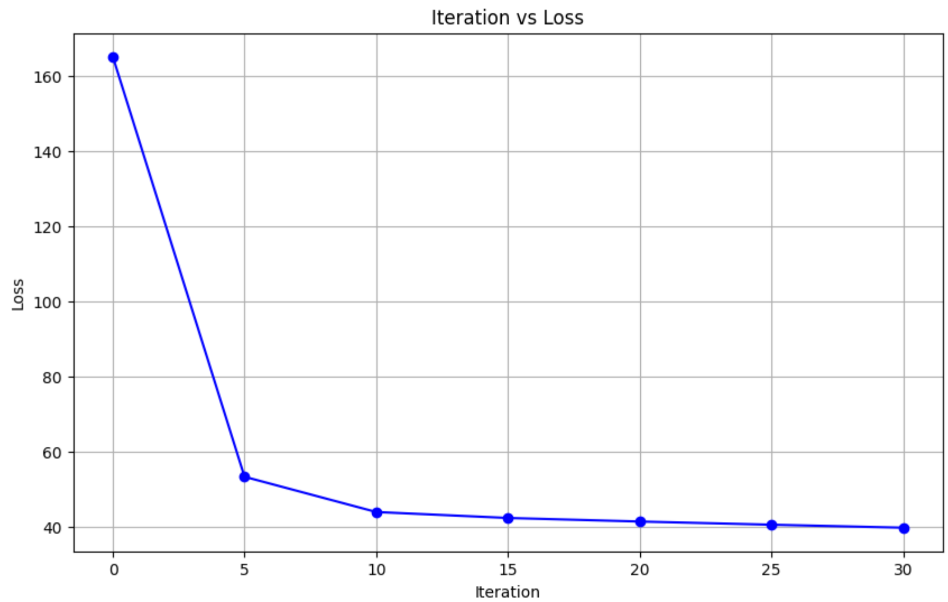


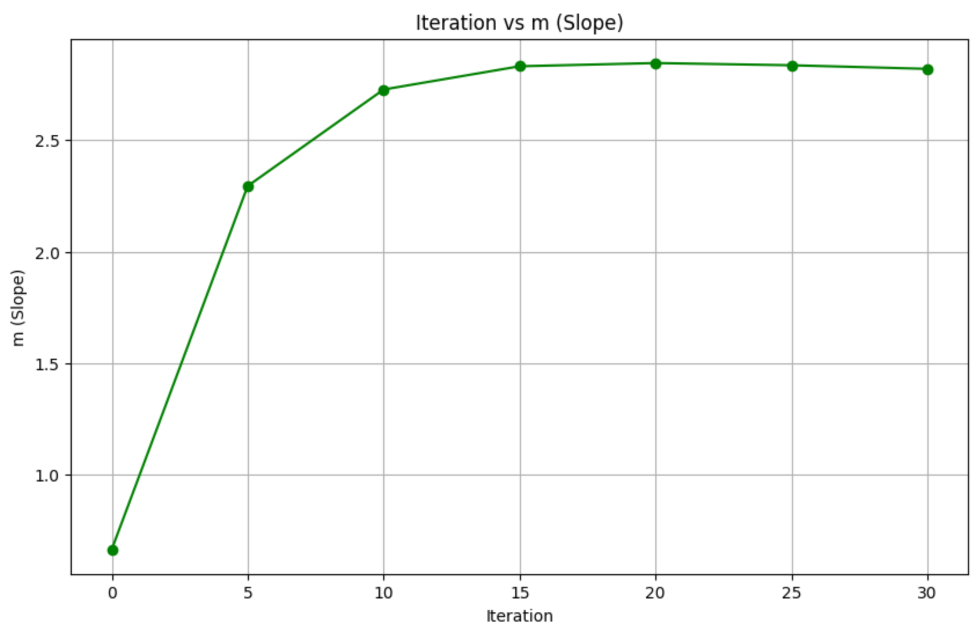
Above diagrams show the limitation of the range for answer to optimize the iteration and learning rate well approximately.

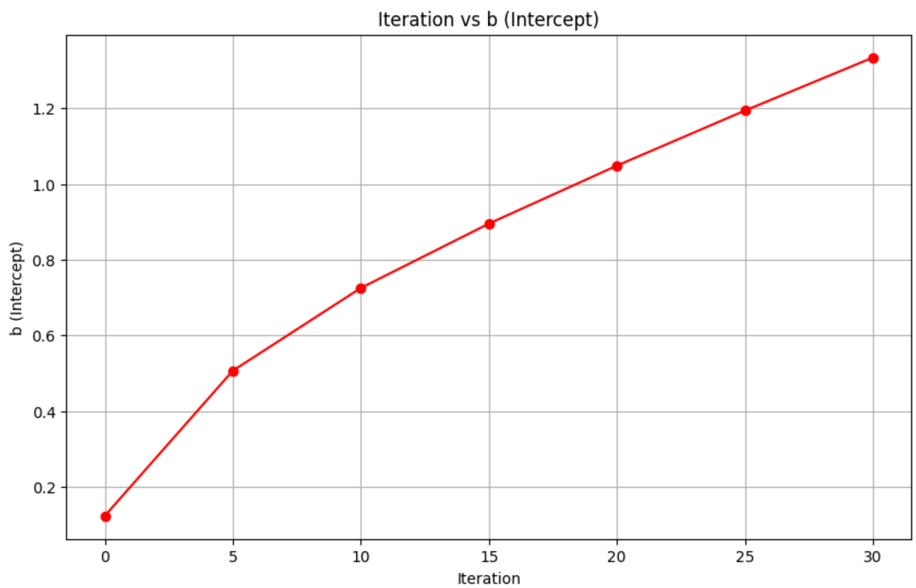
- Calculate the m slope and the b intercept for the large dataset

Final parameters for large dataset (iteration=35, learning_rate = 0.01): *Slope (m): 2.804932076974661* *Intercept (b): 1.4434139285985357 → ~ 3.9* *This model with mentioned hyperparameters needs more iterations (more time) to have stable and less Loss value*

Final parameters for large dataset (iteration=100, learning_rate = 0.0001): *Slope (m): 0.5958643526175813* *Intercept (b): 0.11339460453658075* *This model with mentioned hyperparameters needs more iterations (more time) and a little more learning rate like 0.001*


Iteration vs Loss
Loss
Iteration
160
140
120
100
80
60
40
0
5
10
15
20
25
30


Iteration vs m (Slope)
m (Slope)
Iteration
2.5
2.0
1.5
1.0
0
5
10
15
20
25
30



Above shows we need more iteration for large dataset, but there was some credit limitation for me to stop earlier. Diagrams show there is a limitation for m and loss however b can still increase well.
Note: All above models could estimate the slope m (parameter) with good accuracy however for getting intercept b (parameter) exists some issues and need to change the hyperparameters like LR and iterations.

Find the Parameters using Gradient Descent

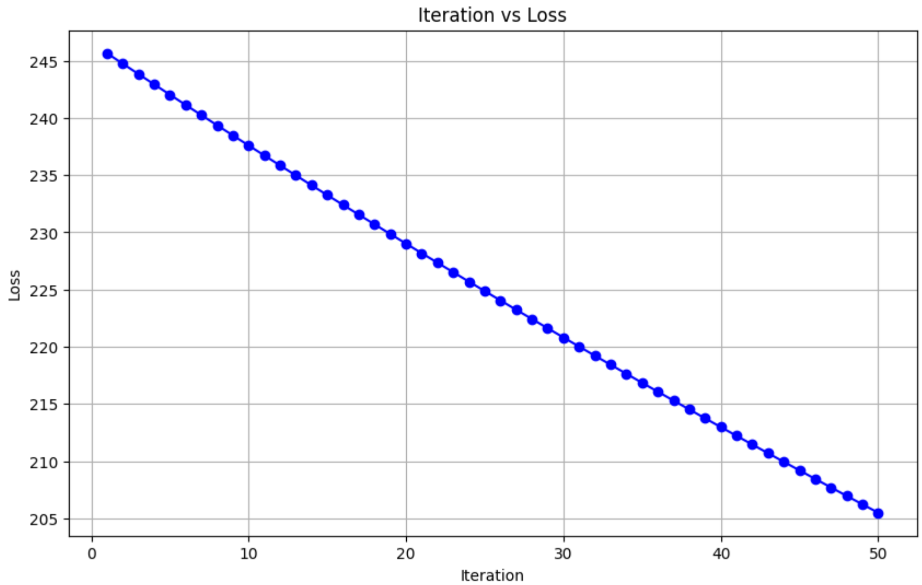
- Print out the costs and model parameters in each iteration. The maximum number of iterations is 50.

```
Iteration 1: Loss = 245.6573, m = 0.006661, b = 0.001229
Iteration 2: Loss = 244.7450, m = 0.013306, b = 0.002456
Iteration 3: Loss = 243.8368, m = 0.019937, b = 0.003680
Iteration 4: Loss = 242.9327, m = 0.026552, b = 0.004903
Iteration 5: Loss = 242.0327, m = 0.033152, b = 0.006124
Iteration 6: Loss = 241.1368, m = 0.039737, b = 0.007342
Iteration 7: Loss = 240.2450, m = 0.046307, b = 0.008559
Iteration 8: Loss = 239.3572, m = 0.052861, b = 0.009774
Iteration 9: Loss = 238.4735, m = 0.059401, b = 0.010986
Iteration 10: Loss = 237.5937, m = 0.065926, b = 0.012197
Iteration 11: Loss = 236.7180, m = 0.072436, b = 0.013405
Iteration 12: Loss = 235.8462, m = 0.078930, b = 0.014611
Iteration 13: Loss = 234.9783, m = 0.085410, b = 0.015816
Iteration 14: Loss = 234.1144, m = 0.091876, b = 0.017018
Iteration 15: Loss = 233.2545, m = 0.098326, b = 0.018219
Iteration 16: Loss = 232.3984, m = 0.104761, b = 0.019417
Iteration 17: Loss = 231.5462, m = 0.111182, b = 0.020614
Iteration 18: Loss = 230.6979, m = 0.117588, b = 0.021808
Iteration 19: Loss = 229.8534, m = 0.123980, b = 0.023001
Iteration 20: Loss = 229.0127, m = 0.130356, b = 0.024191
```

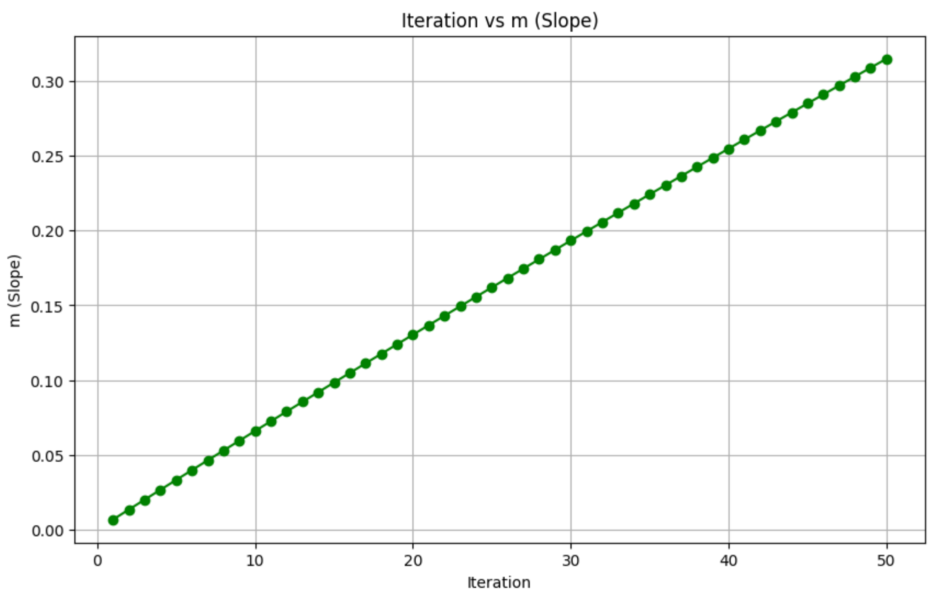
```
Iteration 21: Loss = 228.1759, m = 0.136719, b = 0.025380
Iteration 22: Loss = 227.3429, m = 0.143066, b = 0.026566
Iteration 23: Loss = 226.5136, m = 0.149399, b = 0.027751
Iteration 24: Loss = 225.6881, m = 0.155718, b = 0.028933
Iteration 25: Loss = 224.8664, m = 0.162022, b = 0.030114
Iteration 26: Loss = 224.0483, m = 0.168311, b = 0.031293
Iteration 27: Loss = 223.2340, m = 0.174586, b = 0.032469
Iteration 28: Loss = 222.4234, m = 0.180847, b = 0.033644
Iteration 29: Loss = 221.6164, m = 0.187094, b = 0.034817
Iteration 30: Loss = 220.8131, m = 0.193326, b = 0.035988
Iteration 31: Loss = 220.0135, m = 0.199543, b = 0.037157
Iteration 32: Loss = 219.2175, m = 0.205747, b = 0.038324
Iteration 33: Loss = 218.4251, m = 0.211936, b = 0.039489
Iteration 34: Loss = 217.6362, m = 0.218111, b = 0.040653
Iteration 35: Loss = 216.8510, m = 0.224272, b = 0.041814
Iteration 36: Loss = 216.0693, m = 0.230419, b = 0.042973
Iteration 37: Loss = 215.2912, m = 0.236552, b = 0.044131
Iteration 38: Loss = 214.5166, m = 0.242671, b = 0.045286
Iteration 39: Loss = 213.7455, m = 0.248776, b = 0.046440
Iteration 40: Loss = 212.9779, m = 0.254866, b = 0.047592
Iteration 41: Loss = 212.2138, m = 0.260943, b = 0.048742
Iteration 42: Loss = 211.4531, m = 0.267006, b = 0.049890
Iteration 43: Loss = 210.6959, m = 0.273055, b = 0.051036
Iteration 44: Loss = 209.9422, m = 0.279090, b = 0.052180
Iteration 45: Loss = 209.1918, m = 0.285111, b = 0.053323
Iteration 46: Loss = 208.4449, m = 0.291118, b = 0.054463
Iteration 47: Loss = 207.7013, m = 0.297112, b = 0.055602
Iteration 48: Loss = 206.9611, m = 0.303092, b = 0.056739
Iteration 49: Loss = 206.2243, m = 0.309058, b = 0.057874
Iteration 50: Loss = 205.4908, m = 0.315011, b = 0.059007
Final parameters:
Slope (m): 0.3150106118230866
Intercept (b): 0.059007033958606216
```

- Calculate the m slope and the b intercept based on your calculations

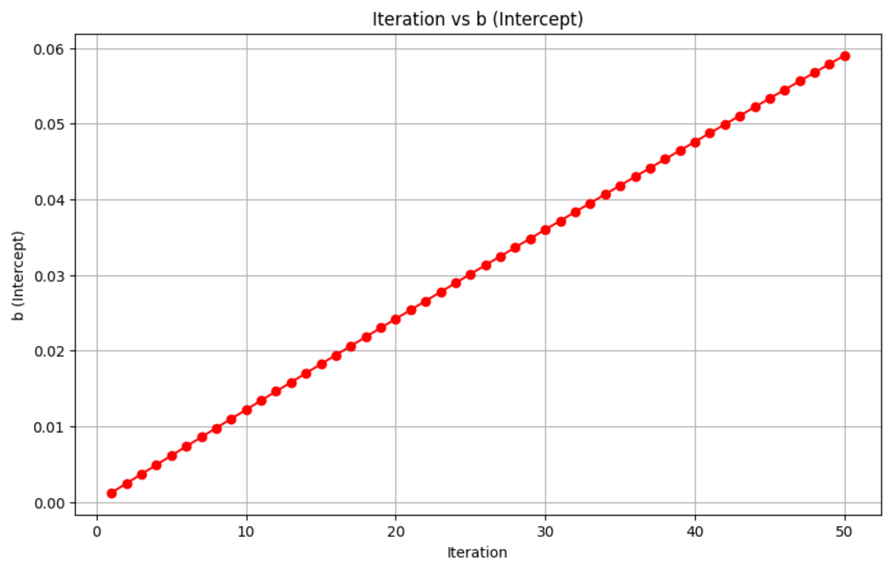
Final parameters:
Slope (m): 0.3150106118230866
Intercept (b): 0.059007033958606216



Iteration vs Loss
Loss
Iteration
245
240
235
230
225
220
215
210
205
0
10
20
30
40
50



Iteration vs m (Slope)
m (Slope)
Iteration
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50



● Comment on how you can interpret the parameters of the model. What is the meaning of m and b in this case?

The model parameters consist of the slope m and the intercept b. As the model iterates, it minimizes the loss function (in this case, the Mean Squared Error) to find the best fit line for the data points representing the relationship between trip time and fare amount. A decrease in loss with each iteration indicates that the model is learning to better predict fare amounts based on trip times. Slope (m) quantifies the change in fare amount per unit increase in trip time. In our results, a slope of approximately 0.315 suggests that for each additional second of trip time, the fare amount increases by about 0.315 currency units. Intercept (b) represents the estimated fare amount when the trip time is zero. An intercept of approximately 0.059 indicates that if there were a trip with a duration of zero seconds (hypothetically), the starting fare would be around 0.059 currency units. So, these parameters help create a linear regression model that estimates fare amounts based on the trip time, enabling fare prediction for taxi services. Simple Linear Regression gives an exact result but can be computationally expensive for large datasets. Gradient Descent provides an approximate result, but it's faster and more scalable for large datasets, especially in distributed environments like PySpark on GCP. If tuned correctly, Gradient Descent should give results very close to Simple Linear Regression, but for very large datasets, the trade-off between accuracy and efficiency comes into play.

Fit Multiple Linear Regression using Gradient Descent

● Print out the costs and model parameters in each iteration.

Iteration 1, Cost: 3945948.4832964763, m: [1.76455744e-01 5.58341313e-03 5.05884956e-03 1.30863370e-04], b: 0.00014766810137397475
Iteration 2, Cost: 30620522546237.824, m: [-4.91568090e+02 -3.33534778e+00 -1.82126249e-02 -7.16602758e-04], b: -0.0012832041487742204
Iteration 3, Cost: 5.9407654746179265e+19, m: [6.84703545e+05 1.68969899e+03 4.00796484e+01 1.37903008e+00], b: 2.2165874171162856

Iteration 4, Cost: 2.877334532430775e+25, m: [-4.76515564e+08 -8.18270020e+05 -2.78798311e+04 -9.59435614e+02], b: -1542.2917867199478
Iteration 5, Cost: 3.4739967184554596e+30, m: [1.65575739e+11 2.62777662e+08 9.68721520e+06 3.33373838e+05], b: 535899.5592234605
Iteration 6, Cost: 1.0425697423701348e+35, m: [-2.86836605e+13 -4.48788097e+10 -1.67816599e+09 -5.77522193e+07], b: -92836930.88757718
Iteration 7, Cost: 7.732009906073271e+38, m: [2.47017783e+15 3.85541679e+12 1.44520094e+11 4.97350091e+09], b: 7994923915.752392
Iteration 8, Cost: 1.400472674113528e+42, m: [-1.05128226e+17 -1.64015511e+14 -6.15062574e+12 -2.11667071e+11], b: -340255724232.15234
Iteration 9, Cost: 6.047059187804543e+44, m: [2.18451046e+18 3.40793687e+15 1.27806836e+14 4.39833281e+12], b: 7070338849884.747
Iteration 10, Cost: 5.914447880207313e+46, m: [-2.16042465e+19 -3.37033098e+16 -1.26397672e+15 -4.34983800e+13], b: -69923832302923.44
Iteration 11, Cost: 1.1685104604038749e+48, m: [9.60280966e+19 1.49806738e+17 5.61821396e+15 1.93344704e+14], b: 310802440471880.06
Iteration 12, Cost: 3.4667189488073156e+48, m: [-1.65402220e+20 -2.58032475e+17 -9.67701222e+15 -3.33023818e+14], b: -535337214920475.7
Iteration 13, Cost: 4.52331699996137e+47, m: [5.97462333e+19 9.32059332e+16 3.49550948e+15 1.20294145e+14], b: 193373354179930.3
Iteration 14, Cost: 1.1188275563055752e+47, m: [-2.97141694e+19 -4.63550049e+16 -1.73845538e+15 -5.98270454e+13], b: -96172231955605.62
Iteration 15, Cost: 4.684578319225702e+46, m: [1.92272543e+19 2.99950988e+16 1.12490857e+15 3.87125010e+13], b: 62230511415595.56
Iteration 16, Cost: 3.0870823484209363e+46, m: [-1.56083188e+19 -2.43494503e+16 -9.13179345e+14 -3.14260708e+13], b: -50517543642953.7
Iteration 17, Cost: 3.0437633705785483e+46, m: [1.54984213e+19 2.41780065e+16 9.06749689e+14 3.12048013e+13], b: 50161851847885.0
Iteration 18, Cost: 4.326623988196347e+46, m: [-1.84780695e+19 -2.88263482e+16 -1.08107680e+15 -3.72040787e+13], b: -59805716111811.95
Iteration 19, Cost: 3.998013143739844e+44, m: [1.77625015e+18 2.77100396e+15 1.03921183e+14 3.57633412e+12], b: 5748972437973.125
Iteration 20, Cost: 1.6923205374885324e+43, m: [-3.65446099e+17 -5.70107043e+14 -2.13807673e+13 -7.35795778e+11], b: -1182794869048.3535
Iteration 21, Cost: 1.8019995916379536e+42, m: [1.19250268e+17 1.86033985e+14 6.97684881e+12 2.40100640e+11], b: 385962808486.2698
Iteration 22, Cost: 3.795572039794101e+41, m: [-5.47293938e+16 -8.53795457e+13 -3.20199465e+12 -1.10193152e+11], b: -177135964598.563
Iteration 23, Cost: 1.3885406102596688e+41, m: [3.31025394e+16 5.16409400e+13 1.93669508e+12 6.66492482e+10], b: 107138952457.28644
Iteration 24, Cost: 8.132977016596443e+40, m: [-2.53341764e+16 -3.95220994e+13 -1.48219985e+12 -5.10082907e+10], b: -81996046008.28568
Iteration 25, Cost: 7.214715230309111e+40, m: [2.38611634e+16 3.72241267e+13 1.39601970e+12 4.80424949e+10], b: 77228516745.63748
Iteration 26, Cost: 9.311255403326586e+40, m: [-2.71072923e+16 -4.22882040e+13 -1.58593766e+12 -5.45783141e+10], b: -87734873915.50308
Iteration 27, Cost: 4.308206565767179e+38, m: [1.84386991e+15 2.87648727e+12 1.07877268e+11 3.71247858e+09], b: 5967825334.056229

Iteration 28, Cost: 1.3167251080243498e+37, m: [-3.22351412e+14 -5.02882460e+11 -1.88595447e+10 -6.49030554e+08], b: -1043319297.9417276
Iteration 29, Cost: 1.1250438721083714e+36, m: [9.42251305e+13 1.46990577e+11 5.51265712e+09 1.89712180e+08], b: 304963242.9203994
Iteration 30, Cost: 1.999121463458819e+35, m: [-3.97193021e+13 -6.19659326e+10 -2.32388767e+09 -7.99739595e+07], b: -128558382.68057543
Iteration 31, Cost: 6.352132398897435e+34, m: [2.23893657e+13 3.49261842e+10 1.30983827e+09 4.50767539e+07], b: 72461175.35443851
Iteration 32, Cost: 3.2934831683466013e+34, m: [-1.61216594e+13 -2.51516105e+10 -9.43285410e+08 -3.24620394e+07], b: -52182767.394184604
Iteration 33, Cost: 2.6208859407853697e+34, m: [1.43815627e+13 2.24347798e+10 8.41334248e+08 2.89536990e+07], b: 46543295.334203824
Iteration 34, Cost: 3.064201606768887e+34, m: [-1.55503685e+13 -2.42596154e+10 -9.09861491e+08 -3.13117919e+07], b: -50333739.65345676
Iteration 35, Cost: 5.05976026176973e+31, m: [6.31898012e+11 9.85313936e+08 3.68972101e+07 1.26987585e+06], b: 2041430.9389074743
Iteration 36, Cost: 1.05940429264371e+30, m: [-9.14350864e+10 -1.43014598e+08 -5.42207078e+06 -1.86496054e+05], b: -299693.1517055435
Iteration 37, Cost: 7.115869042255265e+28, m: [2.36971436e+10 3.66770631e+07 1.31384954e+06 4.53132725e+04], b: 72941.32783062349
Iteration 38, Cost: 1.055220608733595e+28, m: [-9.12543622e+09 -1.44568450e+07 -6.06465324e+05 -2.07722882e+04], b: -33291.500770278915
Iteration 39, Cost: 2.892900304765641e+27, m: [4.77802893e+09 7.29164920e+06 2.06970830e+05 7.22121264e+03], b: 11708.165992209615
Iteration 40, Cost: 1.3218060367855156e+27, m: [-3.22972733e+09 -5.15329906e+06 -2.61529708e+05 -8.90173666e+03], b: -14209.54934607969
Iteration 41, Cost: 9.405648707176155e+26, m: [2.72443478e+09 4.17227655e+06 8.68248000e+04 3.08649159e+03], b: 5061.574326429305
Iteration 42, Cost: 9.937888892560422e+26, m: [-2.80045807e+09 -4.41917640e+06 -2.36413663e+05 -8.03741753e+03], b: -12820.161499050333
Iteration 43, Cost: 1.934537383845732e+23, m: [3.90725095e+07 1.94818324e+04 -7.02840257e+04 -2.32025116e+03], b: -3629.7937925412825
Iteration 44, Cost: 2.573929465591671e+21, m: [-4.50684712e+06 -4.04396222e+04 -7.28334385e+04 -2.40799021e+03], b: -3770.8365156652167
Iteration 45, Cost: 1.3250825911008849e+20, m: [1022581.02627742 -24668.12062785 -72509.6916343   -2396.85256706], b: -3752.93457564294
Iteration 46, Cost: 1.6199481751689216e+19, m: [-357412.57419683 -20642.02083468 -72590.18188679   -2399.62626118], b: -3757.395244946362
[CONTEXT ratelimit_period="5 MINUTES [skipped: 11]"]
Iteration 47, Cost: 3.805904552407058e+18, m: [173220.90326966 -14616.41681331 -72558.88261306 -2398.5528171], b: -3755.671705226879
Iteration 48, Cost: 1.525432535898835e+18, m: [-109607.11023038 -10819.69071639 -72575.16715977   -2399.11692177], b: -3756.580630865344
Iteration 49, Cost: 9.668078649616465e+17, m: [87310.03685172 -7177.7890024 -72563.3726641 -2398.71475833], b: -3755.9363956638376
Iteration 50, Cost: 9.201627450789531e+17, m: [-85179.34391764 -4927.3247565 -72573.17639541 -2399.05596398], b: -3756.487273210688
[CONTEXT ratelimit_period="5 MINUTES [skipped: 11]"]

Iteration 51, Cost: 1.2666466520842488e+18, m: [99980.34951036 -2821.59408051 -72562.03764286 -2398.67660339], b: -3755.880009207661
Iteration 52, Cost: 9660310018079194.0, m: [-8664.32029398 -2426.84273255 -72568.24502728 -2398.89212252], b: -3756.227712844304
Iteration 53, Cost: 445849806920755.7, m: [1702.01961324 -1907.58818955 -72567.47600656 -2398.86771714], b: -3756.1898541065125
Iteration 54, Cost: 85243361221846.03, m: [-527.23676057 -1472.99562214 -72567.42896757 -2398.86833547], b: -3756.1923502964314
Iteration 55, Cost: 35210825368322.54, m: [244.29948244 -1100.37455424 -72567.18987996 -2398.8625411], b: -3756.1846805126147
Iteration 56, Cost: 17720952096710.367, m: [-135.36945611 -795.42626237 -72566.99992693 -2398.85865558], b: -3756.180236428806
[CONTEXT ratelimit_period="5 MINUTES [skipped: 11]"]
Iteration 57, Cost: 9383145192232.69, m: [110.80963556 -552.16882835 -72566.75320676 -2398.85305861], b: -3756.173214834645
Iteration 58, Cost: 5403794061119.298, m: [-92.94971855 -366.83743261 -72566.51051062 -2398.84786937], b: -3756.1670404798792
Iteration 59, Cost: 3909671220976.619, m: [113.9229461 -230.83744453 -72566.21911202 -2398.84130345], b: -3756.158865024209
Iteration 60, Cost: 4303293259074.533, m: [-140.8427138 -137.08730976 -72565.92742287 -2398.83506014], b: -3756.151442222664
Final parameters: m = [-140.8427138 -137.08730976 -72565.92742287 -2398.83506014], b = -3756.151442222664

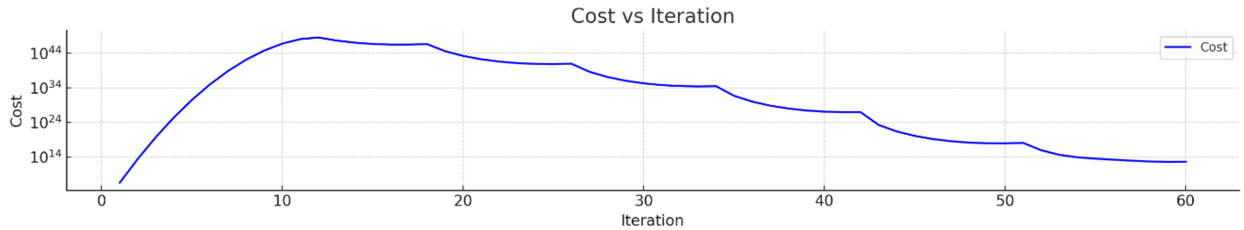
- What are m and b values based on your calculations?

Iteration 60,
Cost: 4303293259074.533, m: [-140.8427138, -137.08730976, -72565.92742287, -2398.83506014], b: -3756.151442222664

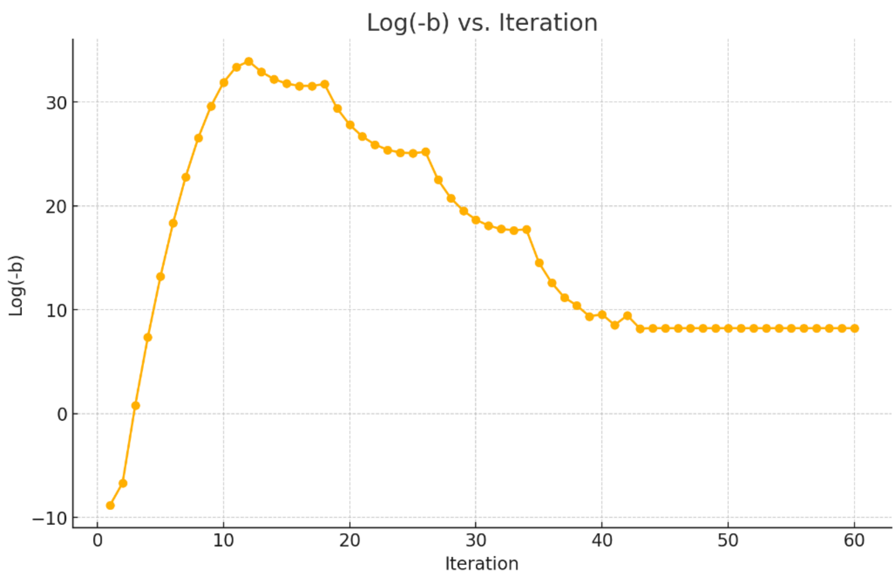
Final parameters: m = [-140.8427138 -137.08730976 -72565.92742287 -2398.83506014], b = -3756.151442222664

- Comment on how you can interpret the parameters of the model. What is the meaning of m_i and *b* in this case

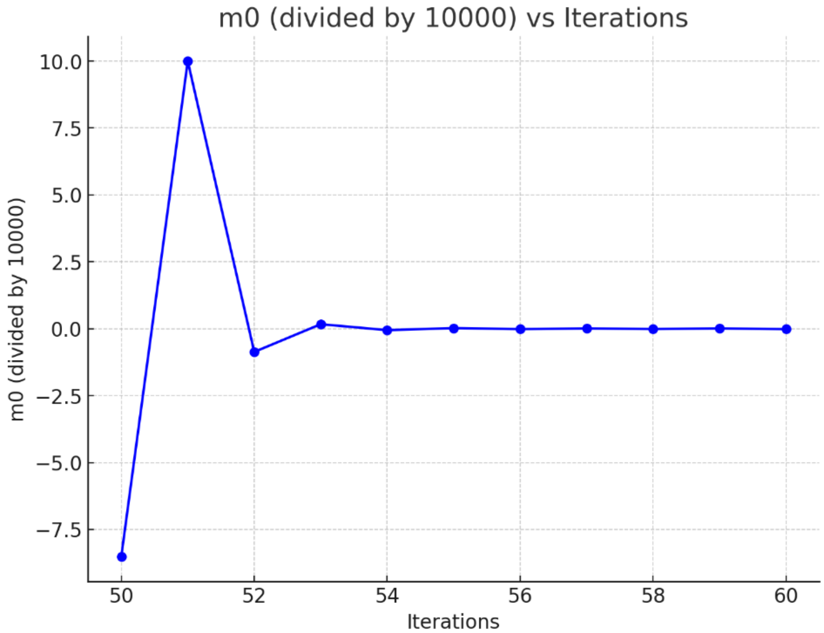
About m values, these represent the coefficients for the independent variables (or features) in the regression model. Each element in the m vector corresponds to the weight (slope) assigned to each feature, which quantifies how much that feature contributes to the prediction of the dependent variable. A positive m means that the feature has a positive relationship with the dependent variable, while a negative m indicates a negative relationship. Regarding b value, this represents the intercept of the regression model. It is the value of the dependent variable when all independent variables are equal to zero. Essentially, it acts as the baseline value of the prediction when no other features contribute to the outcome. Based on the final results of the iterations: m=[−140.84,−137.08,−72565.92,−2398.83]: Each value corresponds to the slope for a specific feature. The larger negative values, like -72565.92, indicate a significant negative impact of the associated feature on the output. (m0: trip_time_in_secs, m1= trip_distance, m2= fare_amount, m3= tolls_amount). b=−3756.15: This is the intercept, indicating the baseline prediction of the dependent *variable when all features are zero. The iterations indicate that the model adjusts these parameters over time to minimize the cost, eventually reaching the final set of values. But still we need more iteration for it, however due to credit I had to limit by 60 iterations.*

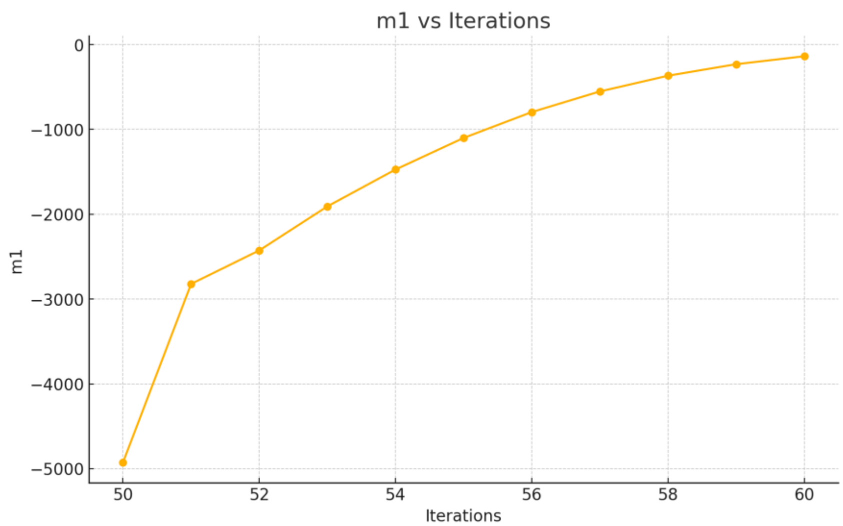


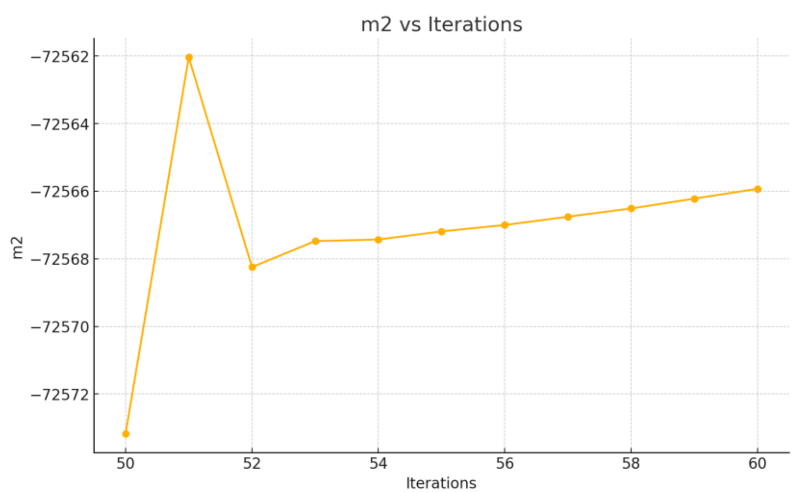
Above trend shows that if we have some more iterations like iteration=500, cost might decrease to near zero well (credit limitation)

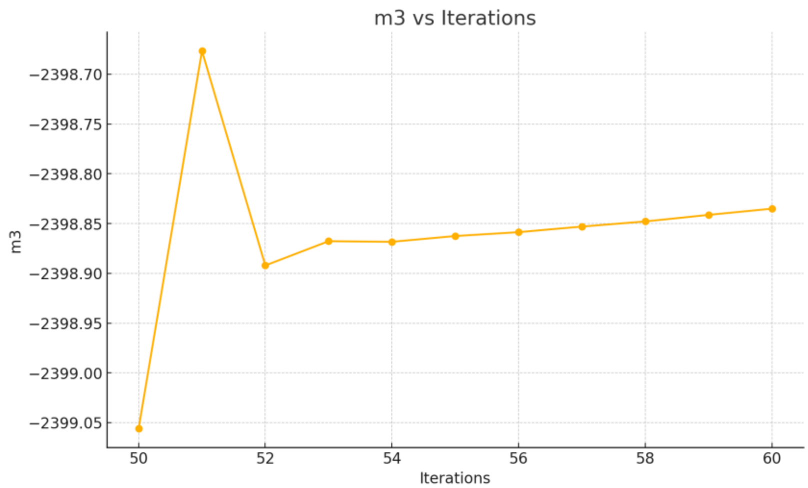


I want to show the convergence of method here to specific value. It needs more iterations.


m0 (divided by 10000) vs Iterations
m0 (divided by 10000)
Iterations
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
50
52
54
56
58
60

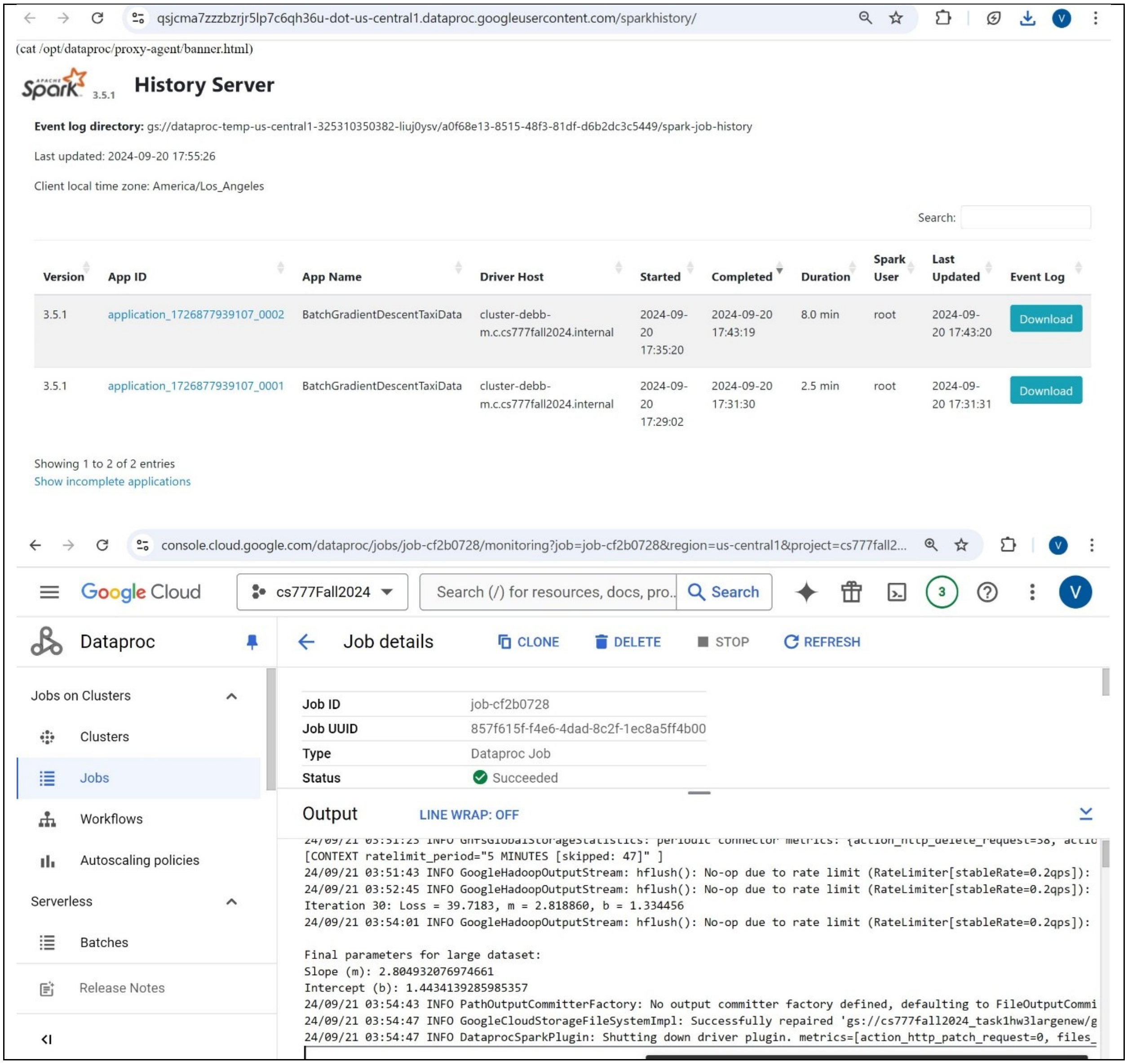

m1 vs Iterations
m1
Iterations
0
−1000
−2000
−3000
−4000
−5000
50
52
54
56
58
60

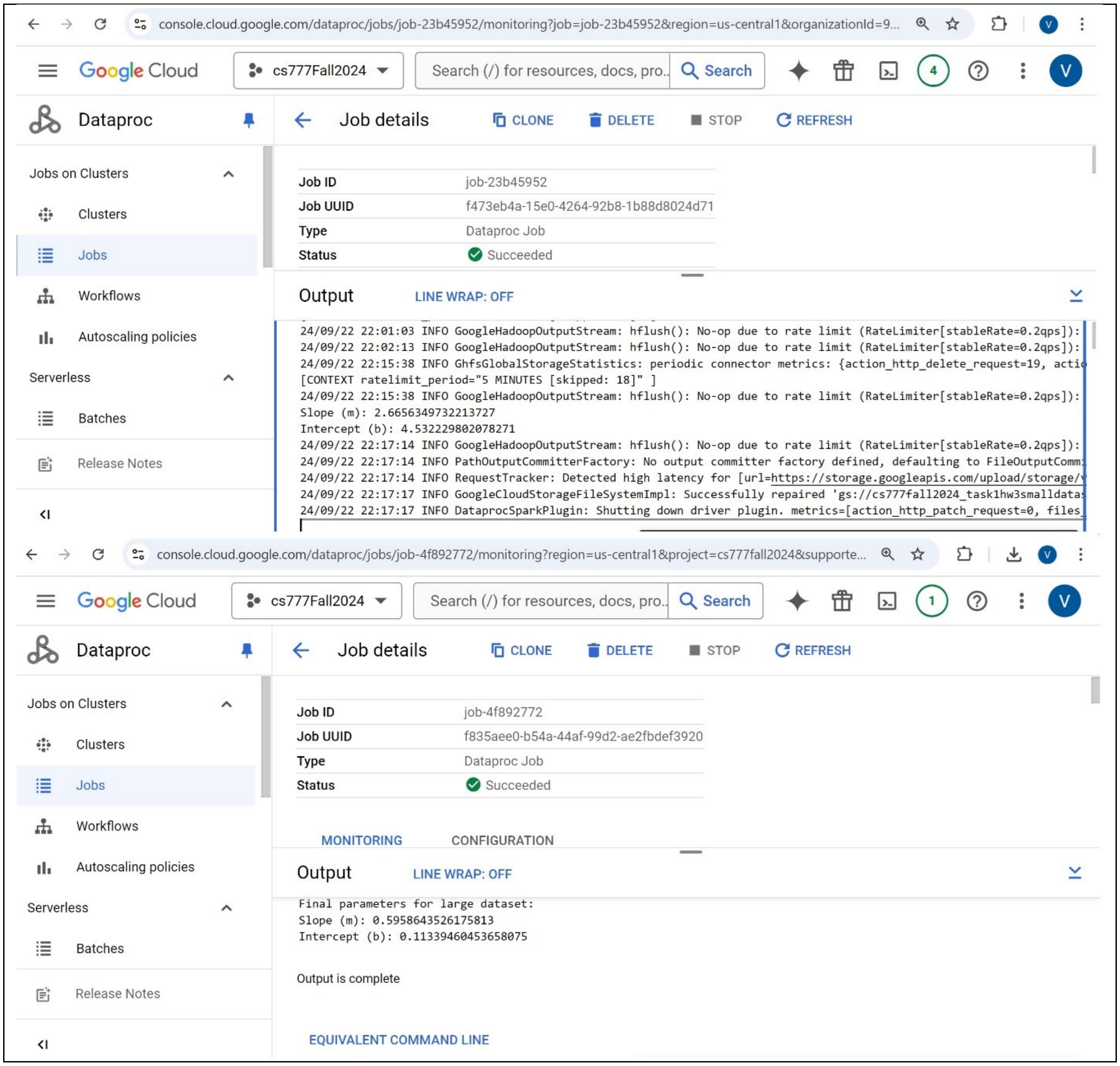

m2 vs Iterations
m2
Iterations
−72562
−72564
−72566
−72568
−72570
−72572
50
52
54
56
58
60

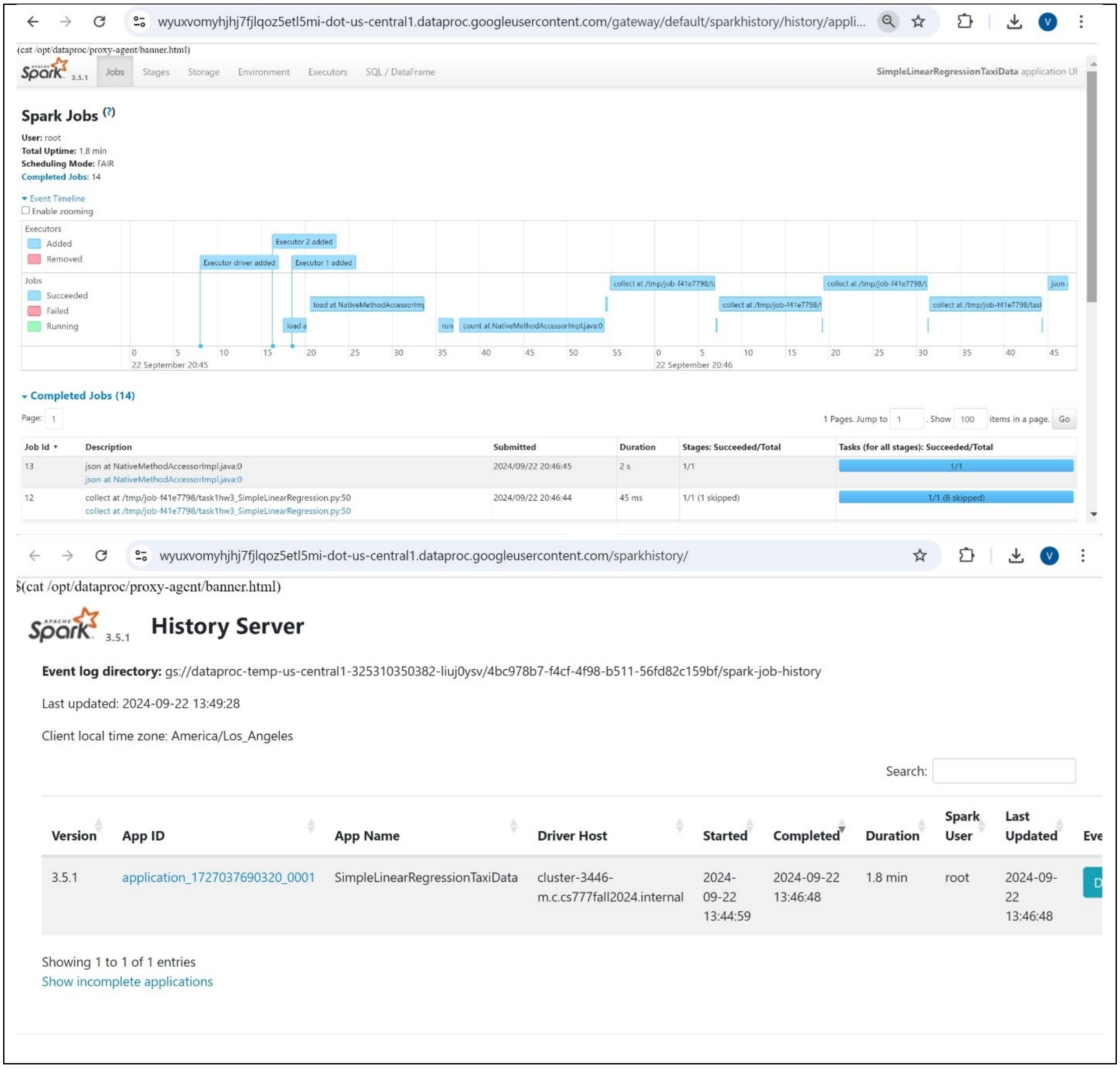


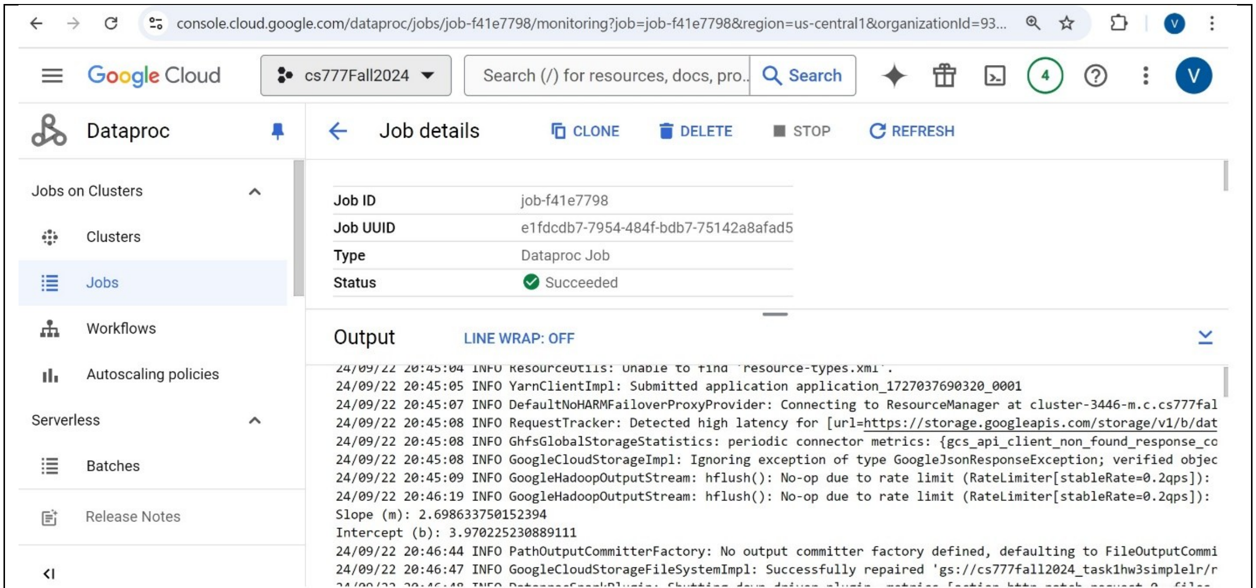
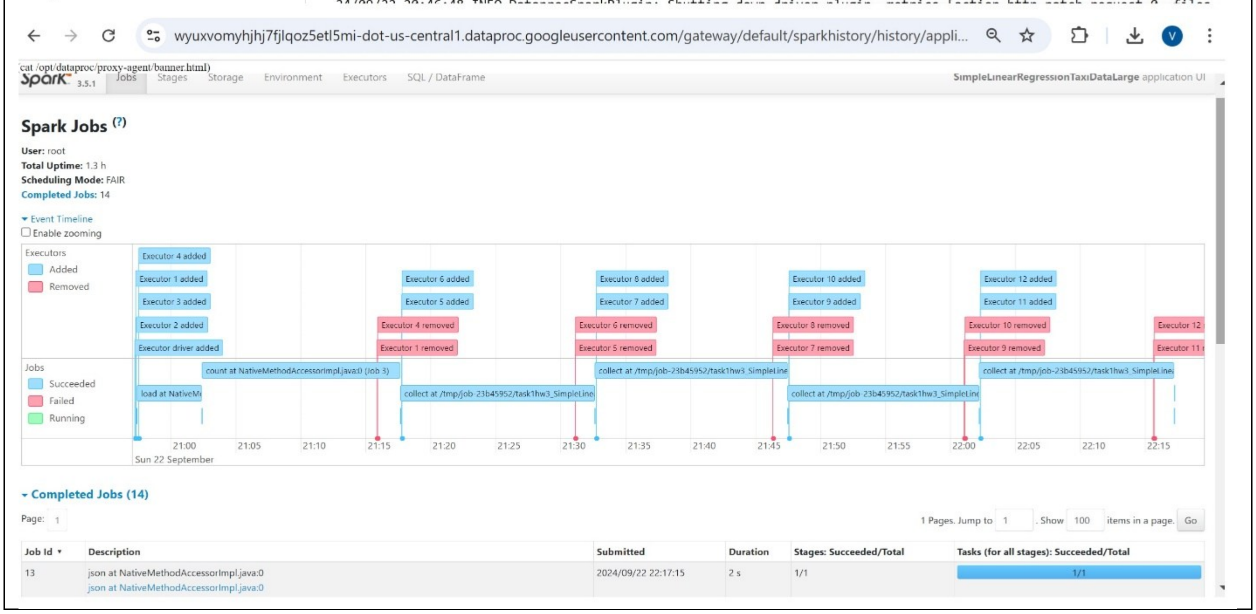
Spark History Output:

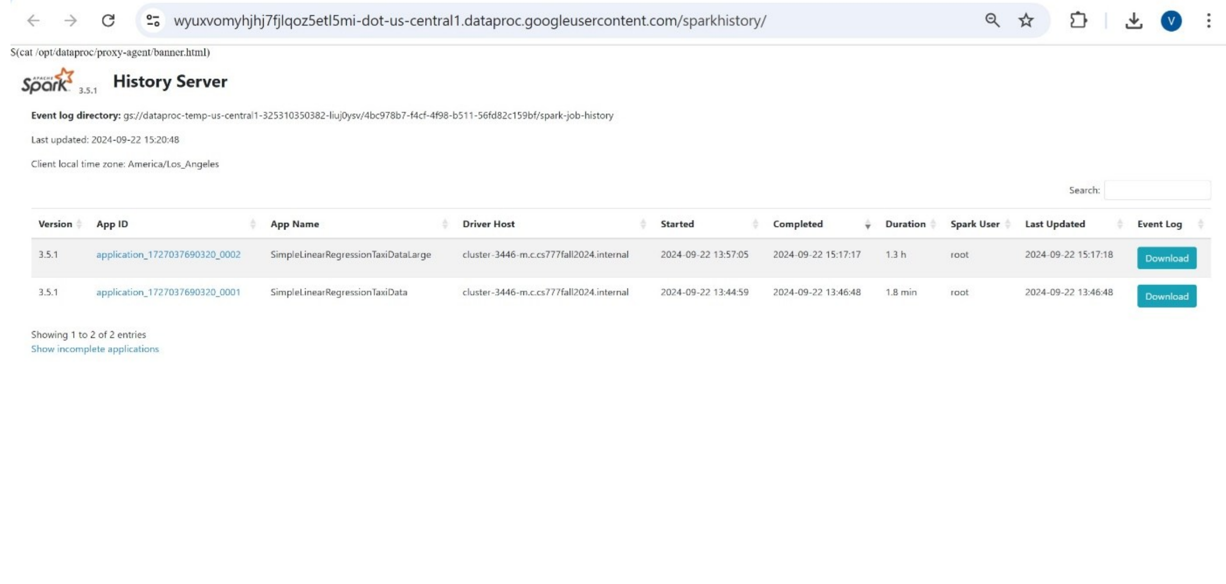
Simple Linear Regression


qsjcma7zzzbzrjr5lp7c6qh36u-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
History Server
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/a0f68e13-8515-48f3-81df-d6b2dc3c5449/spark-job-history
Last updated: 2024-09-20 17:55:26
Client local time zone: America/Los_Angeles
Search:
Version | App ID | App Name | Driver Host | Started | Completed | Duration | Spark User | Last Updated | Event Log
3.5.1 | application_1726877939107_0002 | BatchGradientDescentTaxiData | cluster-debb-m.c.cs777fall2024.internal | 2024-09-20 17:35:20 | 2024-09-20 17:43:19 | 8.0 min | root | 2024-09-20 17:43:20 | Download
3.5.1 | application_1726877939107_0001 | BatchGradientDescentTaxiData | cluster-debb-m.c.cs777fall2024.internal | 2024-09-20 17:29:02 | 2024-09-20 17:31:30 | 2.5 min | root | 2024-09-20 17:31:31 | Download
Showing 1 to 2 of 2 entries
Show incomplete applications
console.cloud.google.com/dataproc/jobs/job-cf2b0728/monitoring?job=job-cf2b0728&region=us-central1&project=cs777fall2...
Google Cloud
cs777Fall2024
Search (/) for resources, docs, pro...
Search
Dataproc
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Release Notes
Job ID job-cf2b0728
Job UUID 857f615f-f4e6-4dad-8c2f-1ec8a5ff4b00
Type Dataproc Job
Status Succeeded
Output
LINE WRAP: OFF
[CONTEXT ratelimit_period="5 MINUTES [skipped: 47]"]
24/09/21 03:51:43 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/21 03:52:45 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
Iteration 30: Loss = 39.7183, m = 2.818860, b = 1.334456
24/09/21 03:54:01 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
Final parameters for large dataset:
Slope (m): 2.804932076974661
Intercept (b): 1.4434139285985357
24/09/21 03:54:43 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommi
24/09/21 03:54:47 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task1hw3largenew/g
24/09/21 03:54:47 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_


console.cloud.google.com/dataproc/jobs/job-23b45952/monitoring?job=job-23b45952&region=us-central1&organizationId=9...
Google Cloud
cs777Fall2024
Search (/) for resources, docs, pro...
Search
Dataproc
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Release Notes
Job ID job-23b45952
Job UUID f473eb4a-15e0-4264-92b8-1b88d8024d71
Type Dataproc Job
Status Succeeded
Output
LINE WRAP: OFF
24/09/22 22:01:03 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/22 22:02:13 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/22 22:15:38 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {action_http_delete_request=19, actio
[CONTEXT ratelimit_period="5 MINUTES [skipped: 18]"]
24/09/22 22:15:38 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
Slope (m): 2.6656349732213727
Intercept (b): 4.532229802078271
24/09/22 22:17:14 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/22 22:17:14 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputComm
24/09/22 22:17:14 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/upload/storage/v
24/09/22 22:17:17 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task1hw3smalldata
24/09/22 22:17:17 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_
console.cloud.google.com/dataproc/jobs/job-4f892772/monitoring?region=us-central1&project=cs777fall2024&supporte...
Google Cloud
cs777Fall2024
Search (/) for resources, docs, pro...
Search
Dataproc
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Release Notes
Job ID job-4f892772
Job UUID f835aee0-b54a-44af-99d2-ae2fbdef3920
Type Dataproc Job
Status Succeeded
MONITORING
CONFIGURATION
Output
LINE WRAP: OFF
Final parameters for large dataset:
Slope (m): 0.5958643526175813
Intercept (b): 0.11339460453658075
Output is complete
EQUIVALENT COMMAND LINE


wyuxvomyhjhj7fjlqoz5etl5mi-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
SQL / DataFrame
SimpleLinearRegressionTaxiData application UI
Spark Jobs (?)
User: root
Total Uptime: 1.8 min
Scheduling Mode: FAIR
Completed Jobs: 14
Event Timeline
Enable zooming
Executors
Added
Removed
Executor 2 added
Executor driver added
Executor 1 added
Jobs
Succeeded
Failed
Running
22 September 20:45
22 September 20:46
Completed Jobs (14)
Page: 1
1 Pages. Jump to 1 . Show 100 items in a page. Go
Job Id
Description
Submitted
Duration
Stages: Succeeded/Total
Tasks (for all stages): Succeeded/Total
13
json at NativeMethodAccessorImpl.java:0
json at NativeMethodAccessorImpl.java:0
2024/09/22 20:46:45
2 s
1/1
1/1
12
collect at /tmp/job-f41e7798/task1hw3_SimpleLinearRegression.py:50
collect at /tmp/job-f41e7798/task1hw3_SimpleLinearRegression.py:50
2024/09/22 20:46:44
45 ms
1/1 (1 skipped)
1/1 (8 skipped)
wyuxvomyhjhj7fjlqoz5etl5mi-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
$(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
History Server
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/4bc978b7-f4cf-4f98-b511-56fd82c159bf/spark-job-history
Last updated: 2024-09-22 13:49:28
Client local time zone: America/Los_Angeles
Search:
Version
App ID
App Name
Driver Host
Started
Completed
Duration
Spark User
Last Updated
3.5.1
application_1727037690320_0001
SimpleLinearRegressionTaxiData
cluster-3446-m.c.cs777fall2024.internal
2024-09-22 13:44:59
2024-09-22 13:46:48
1.8 min
root
2024-09-22 13:46:48
Showing 1 to 1 of 1 entries
Show incomplete applications


console.cloud.google.com/dataproc/jobs/job-f41e7798/monitoring?job=job-f41e7798&region=us-central1&organizationId=93...
Google Cloud
cs777Fall2024
Search (/) for resources, docs, pro...
Search
Dataproc
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Release Notes
Job ID job-f41e7798
Job UUID e1fdcdb7-7954-484f-bdb7-75142a8afad5
Type Dataproc Job
Status Succeeded
Output
LINE WRAP: OFF
24/09/22 20:45:05 INFO YarnClientImpl: Submitted application application_1727037690320_0001
24/09/22 20:45:07 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-3446-m.c.cs777fal
24/09/22 20:45:08 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/storage/v1/b/dat
24/09/22 20:45:08 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_co
24/09/22 20:45:08 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified objec
24/09/22 20:45:09 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/22 20:46:19 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
Slope (m): 2.698633750152394
Intercept (b): 3.970225230889111
24/09/22 20:46:44 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommi
24/09/22 20:46:47 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task1hw3simplelr/r

wyuxvomyhjhj7fjlqoz5etl5mi-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appli...
Jobs Stages Storage Environment Executors SQL / DataFrame
SimpleLinearRegressionTaxiDataLarge application UI
Spark Jobs
User: root
Total Uptime: 1.3 h
Scheduling Mode: FAIR
Completed Jobs: 14
Event Timeline
Enable zooming
Executors
Added
Removed
Jobs
Succeeded
Failed
Running
Sun 22 September
Completed Jobs (14)
Job Id Description Submitted Duration Stages: Succeeded/Total Tasks (for all stages): Succeeded/Total
13 json at NativeMethodAccessorImpl.java:0 2024/09/22 22:17:15 2 s 1/1 1/1


wyuxvomyhjhj7fjlqoz5etl5mi-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

S(cat /opt/dataproc/proxy-agent/banner.html)

3.5.1 History Server

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/4bc978b7-f4cf-4f98-b511-56fd82c159bf/spark-job-history

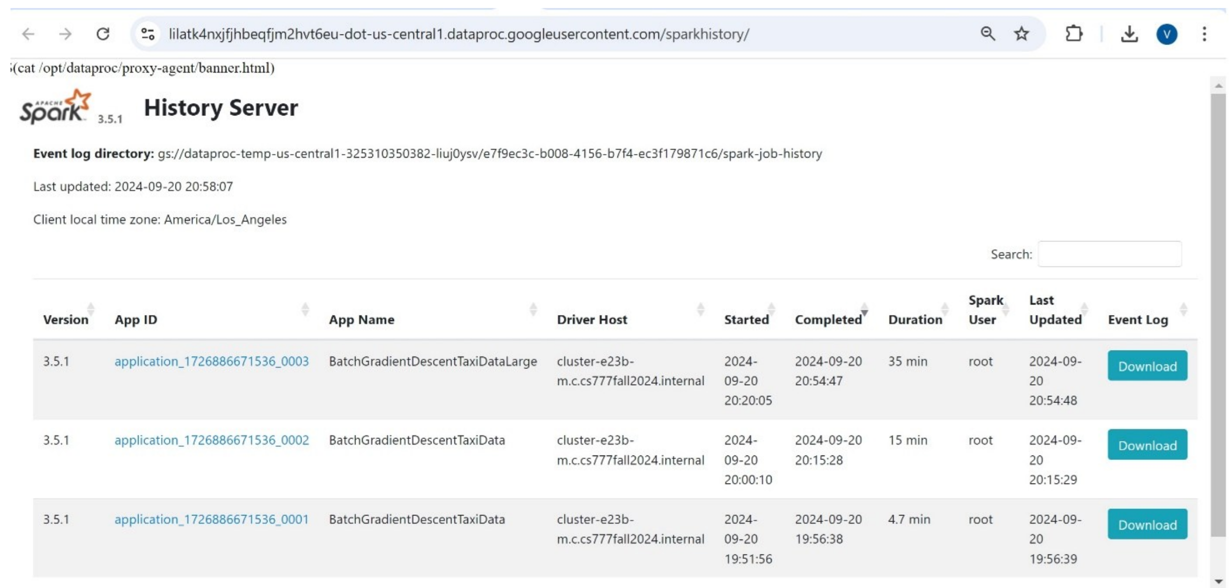
Last updated: 2024-09-22 15:20:48

Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1727037690320_0002	SimpleLinearRegressionTaxiDataLarge	cluster-3446-m.c.cs777fall2024.internal	2024-09-22 13:57:05	2024-09-22 15:17:17	1.3 h	root	2024-09-22 15:17:18	Download
3.5.1	application_1727037690320_0001	SimpleLinearRegressionTaxiData	cluster-3446-m.c.cs777fall2024.internal	2024-09-22 13:44:59	2024-09-22 13:46:48	1.8 min	root	2024-09-22 13:46:48	Download

Showing 1 to 2 of 2 entries
Show incomplete applications


lilatk4nxjfjhbeqfjm2hvt6eu-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

$(cat /opt/dataproc/proxy-agent/banner.html)

3.5.1 History Server

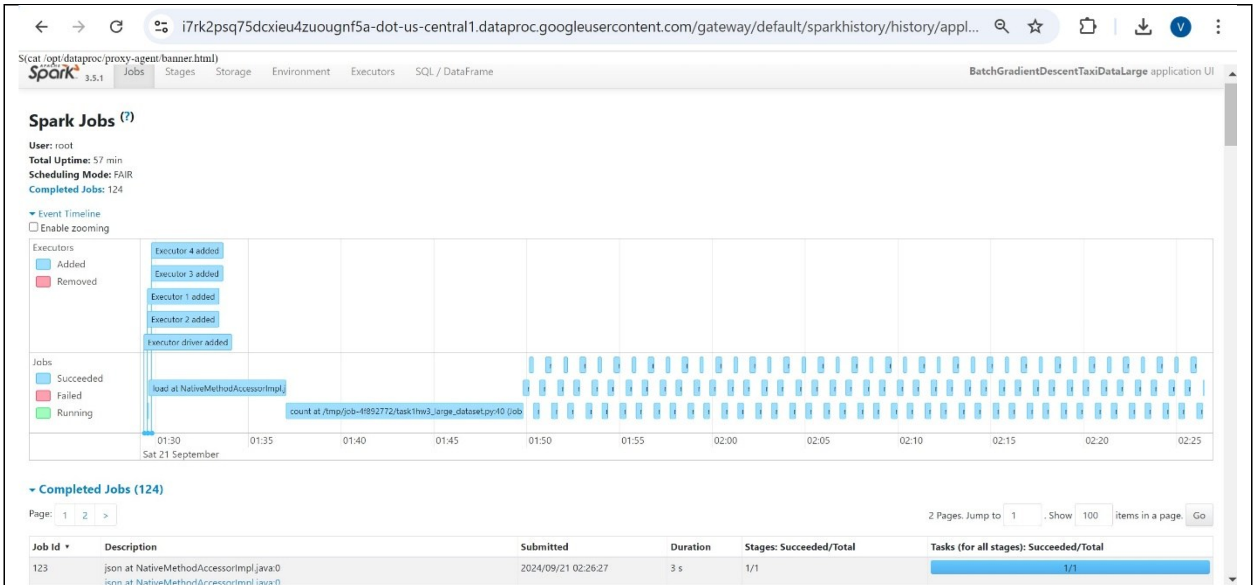
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/e7f9ec3c-b008-4156-b7f4-ec3f179871c6/spark-job-history

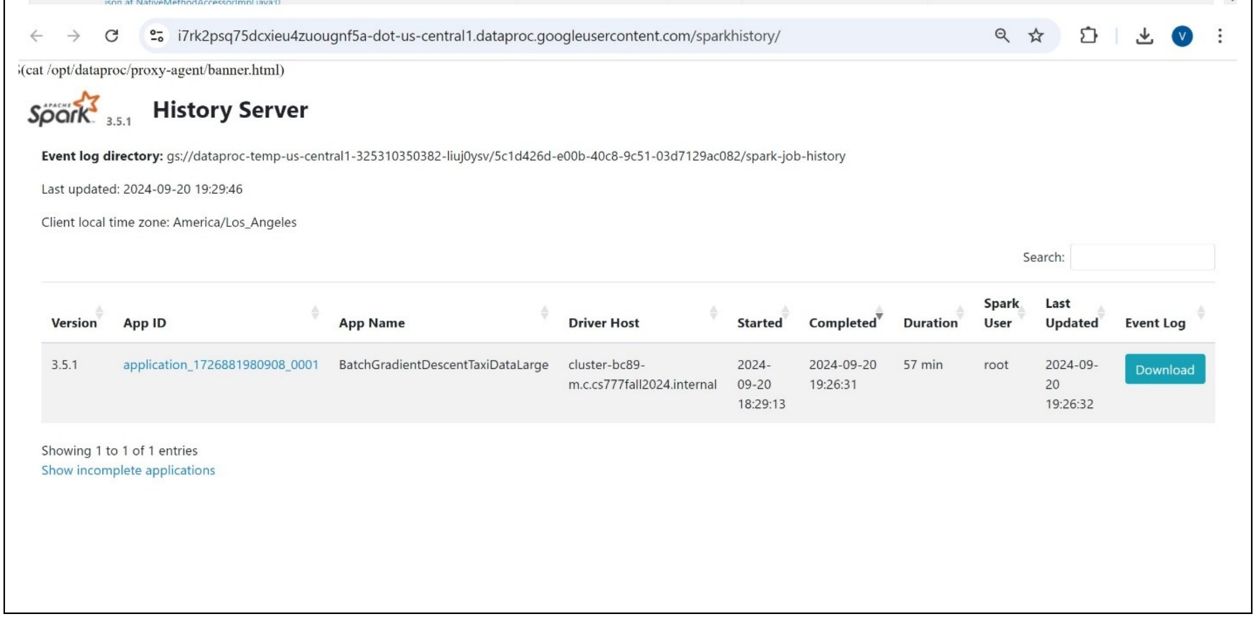
Last updated: 2024-09-20 20:58:07

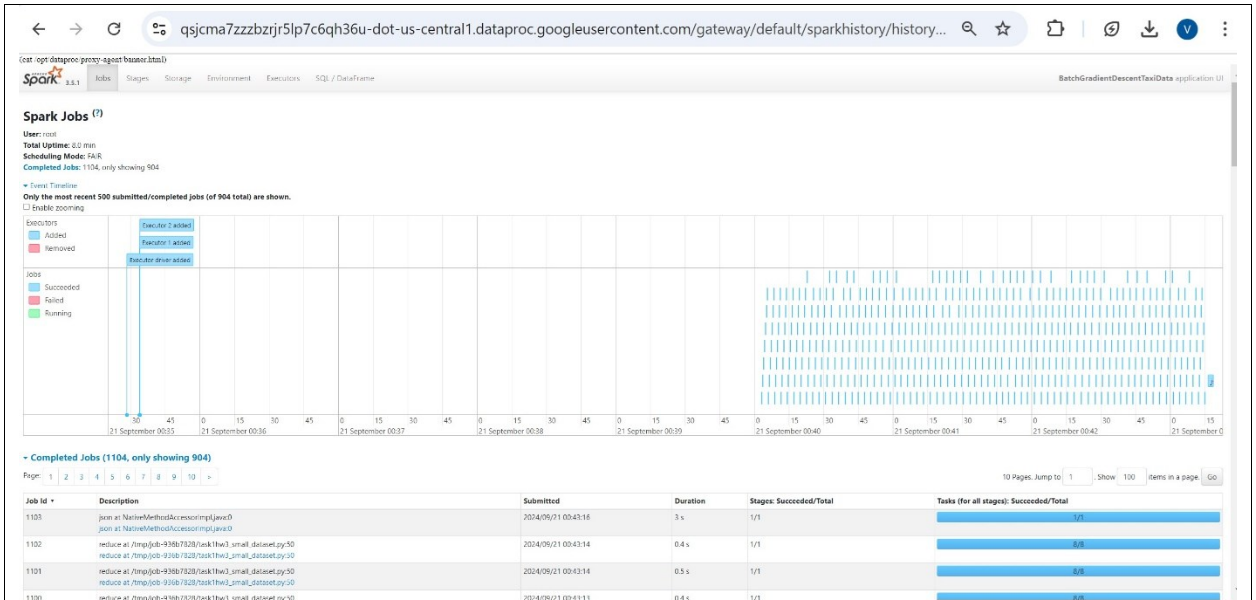
Client local time zone: America/Los_Angeles

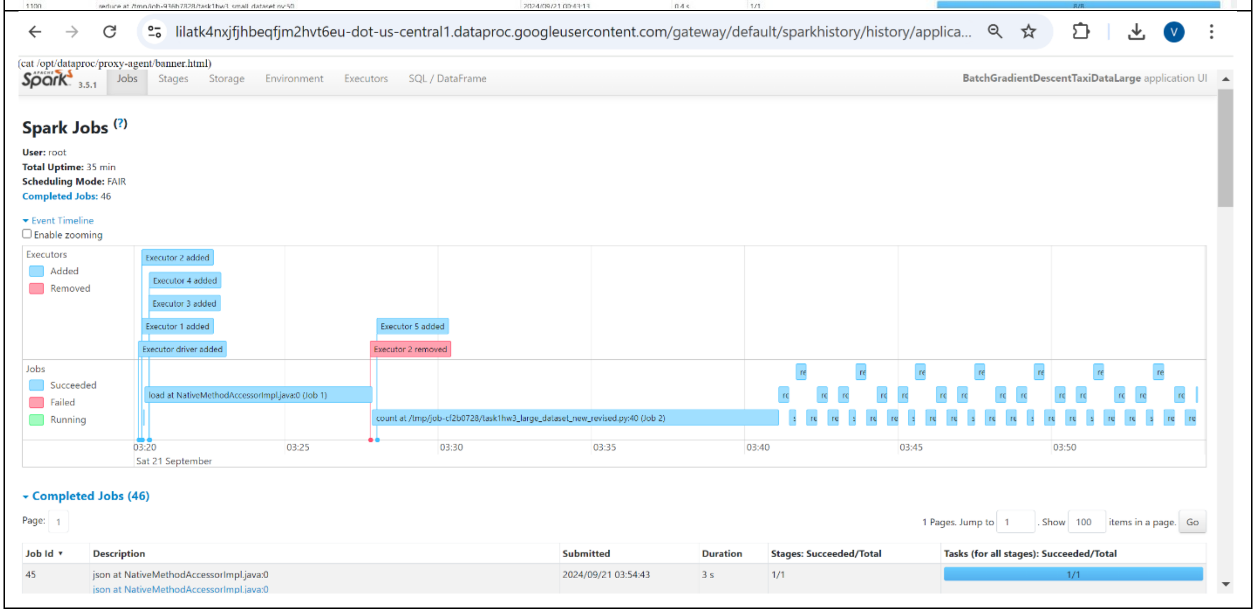
Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1726886671536_0003	BatchGradientDescentTaxiDataLarge	cluster-e23b-m.c.cs777fall2024.internal	2024-09-20 20:20:05	2024-09-20 20:54:47	35 min	root	2024-09-20 20:54:48	Download
3.5.1	application_1726886671536_0002	BatchGradientDescentTaxiData	cluster-e23b-m.c.cs777fall2024.internal	2024-09-20 20:00:10	2024-09-20 20:15:28	15 min	root	2024-09-20 20:15:29	Download
3.5.1	application_1726886671536_0001	BatchGradientDescentTaxiData	cluster-e23b-m.c.cs777fall2024.internal	2024-09-20 19:51:56	2024-09-20 19:56:38	4.7 min	root	2024-09-20 19:56:39	Download

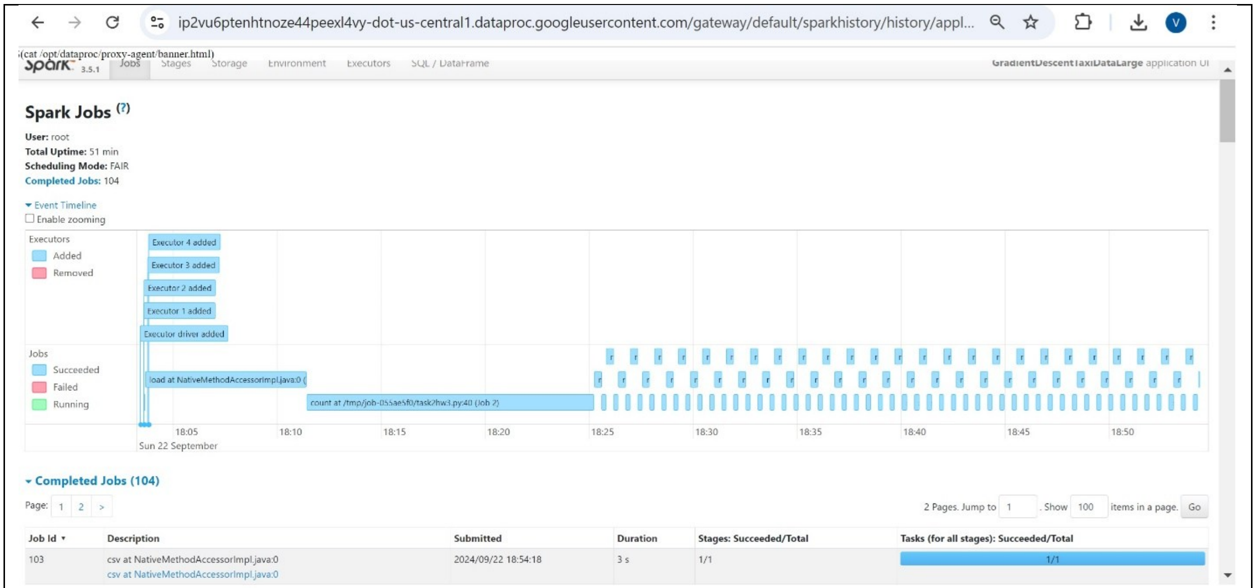

i7rk2psq75dcxieu4zuougnf5a-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appl...
S(cat /opt/dataproc/proxy-agent/banner.html)
3.5.1 Jobs Stages Storage Environment Executors SQL / DataFrame
BatchGradientDescentTaxiDataLarge application UI
Spark Jobs (?)
User: root
Total Uptime: 57 min
Scheduling Mode: FAIR
Completed Jobs: 124
Event Timeline
Enable zooming
Executors
Added
Removed
Executor 4 added
Executor 3 added
Executor 1 added
Executor 2 added
Executor driver added
Jobs
Succeeded
Failed
Running
load at NativeMethodAccessorImpl.j
count at /tmp/job-4f892772/task1hw3_large_dataset.py:40 (Job
01:30 01:35 01:40 01:45 01:50 01:55 02:00 02:05 02:10 02:15 02:20 02:25
Sat 21 September
Completed Jobs (124)
Page: 1 2 >
2 Pages. Jump to 1 . Show 100 items in a page. Go
Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total
123 | json at NativeMethodAccessorImpl.java:0 | 2024/09/21 02:26:27 | 3 s | 1/1 | 1/1

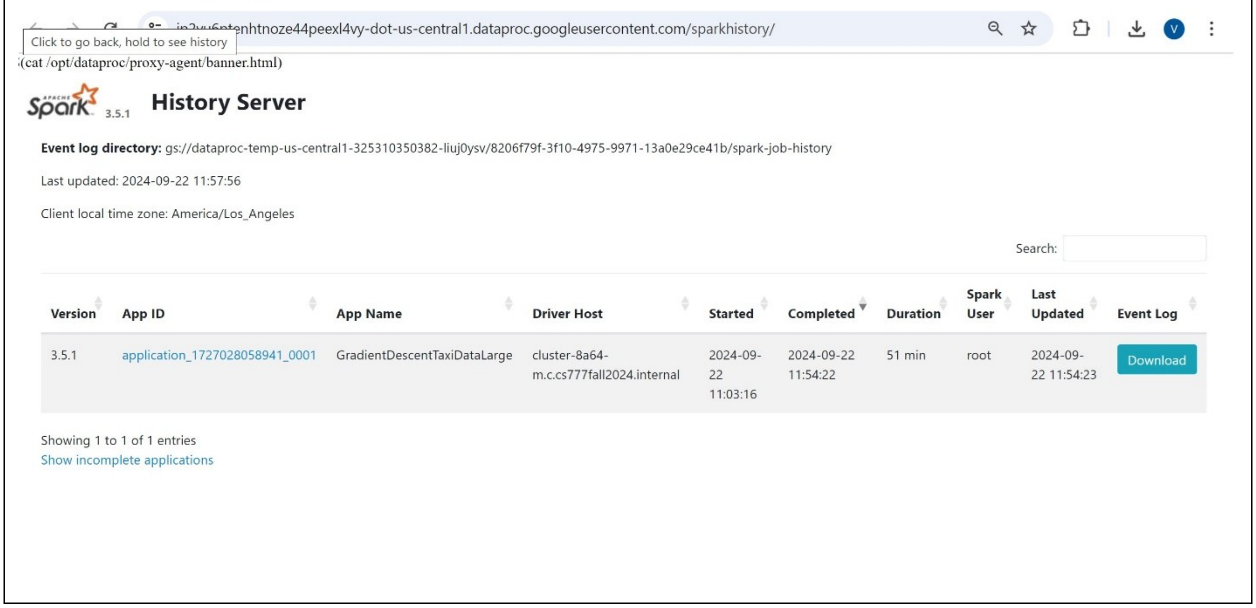

i7rk2psq75dcxieu4zuougnf5a-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
(cat /opt/dataproc/proxy-agent/banner.html)
3.5.1 History Server
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/5c1d426d-e00b-40c8-9c51-03d7129ac082/spark-job-history
Last updated: 2024-09-20 19:29:46
Client local time zone: America/Los_Angeles
Search:
Version | App ID | App Name | Driver Host | Started | Completed | Duration | Spark User | Last Updated | Event Log
3.5.1 | application_1726881980908_0001 | BatchGradientDescentTaxiDataLarge | cluster-bc89-m.c.cs777fall2024.internal | 2024-09-20 18:29:13 | 2024-09-20 19:26:31 | 57 min | root | 2024-09-20 19:26:32 | Download
Showing 1 to 1 of 1 entries
Show incomplete applications

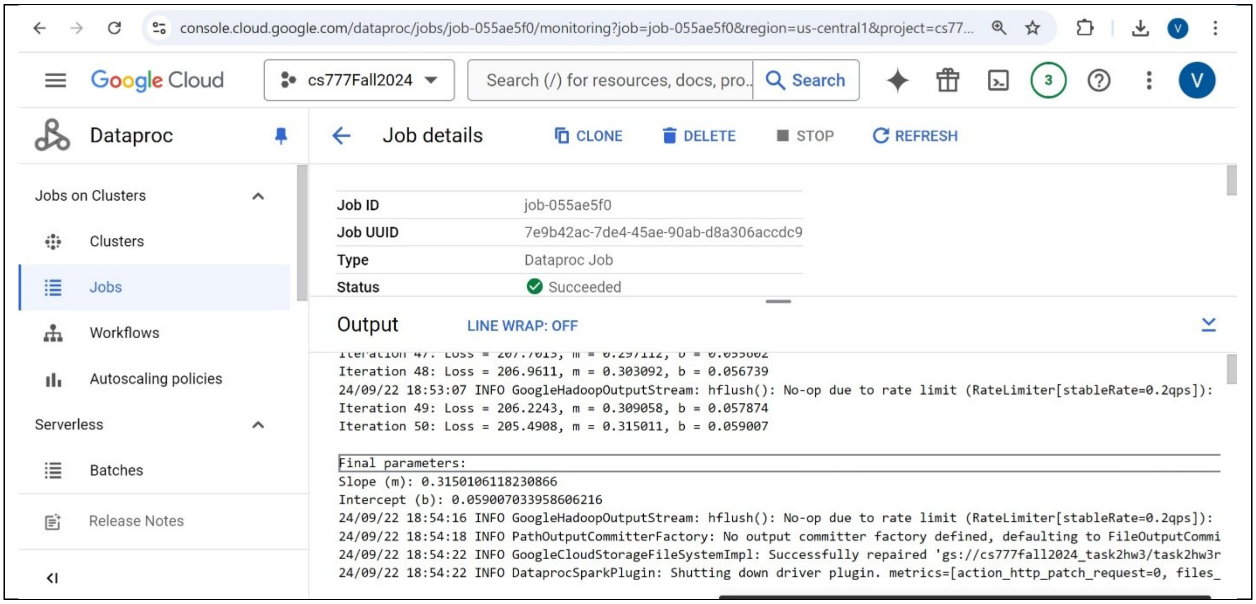

qsjcma7zzzbzrjr5lp7c6qh36u-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history...
Spark Jobs
BatchGradientDescentTaxiData application UI
Completed Jobs (1104, only showing 904)


lilatk4nxjfjhbeqfjm2hvt6eu-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/applica...
(cat /opt/dataproc/proxy-agent/banner.html)
Jobs
Stages
Storage
Environment
Executors
SQL / DataFrame
BatchGradientDescentTaxiDataLarge application UI
Spark Jobs
User: root
Total Uptime: 35 min
Scheduling Mode: FAIR
Completed Jobs: 46
Event Timeline
Enable zooming
Executors
Added
Removed
Executor 2 added
Executor 4 added
Executor 3 added
Executor 1 added
Executor driver added
Executor 5 added
Executor 2 removed
Jobs
Succeeded
Failed
Running
load at NativeMethodAccessorImpl.java:0 (Job 1)
count at /tmp/job-cf2b0728/task1hw3_large_dataset_new_revised.py:40 (Job 2)
03:20
Sat 21 September
03:25
03:30
03:35
03:40
03:45
03:50
Completed Jobs (46)
Page: 1
1 Pages. Jump to 1 . Show 100 items in a page. Go
Job Id
Description
Submitted
Duration
Stages: Succeeded/Total
Tasks (for all stages): Succeeded/Total
45
json at NativeMethodAccessorImpl.java:0
2024/09/21 03:54:43
3 s
1/1
1/1

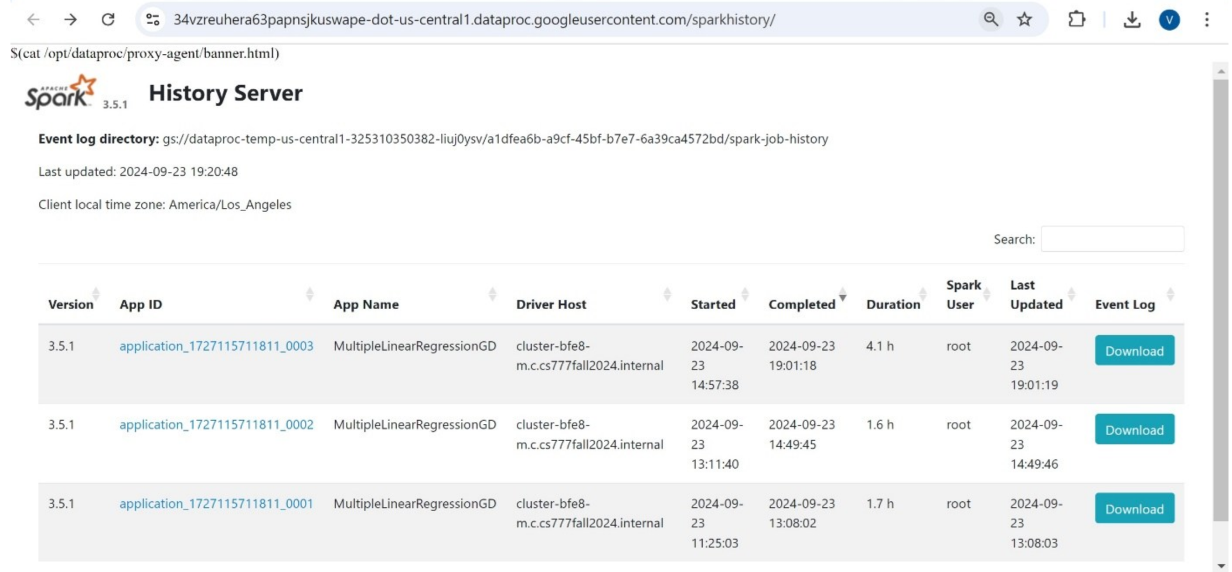
SGD


ip2vu6ptenhtnoze44peexl4vy-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/appl...
(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
Jobs
Stages
Storage
Environment
Executors
SQL / DataFrame
GradientDescentTaxiDataLarge application UI
Spark Jobs (?)
User: root
Total Uptime: 51 min
Scheduling Mode: FAIR
Completed Jobs: 104
Event Timeline
Enable zooming
Executors
Added
Removed
Executor 4 added
Executor 3 added
Executor 2 added
Executor 1 added
Executor driver added
Jobs
Succeeded
Failed
Running
load at NativeMethodAccessorImpl.java:0
count at /tmp/job-055ae5f0/task2hw3.py:40 (Job 2)
18:05
18:10
18:15
18:20
18:25
18:30
18:35
18:40
18:45
18:50
Sun 22 September
Completed Jobs (104)
Page: 1 2 >
2 Pages. Jump to 1 . Show 100 items in a page. Go
Job Id
Description
Submitted
Duration
Stages: Succeeded/Total
Tasks (for all stages): Succeeded/Total
103
csv at NativeMethodAccessorImpl.java:0
csv at NativeMethodAccessorImpl.java:0
2024/09/22 18:54:18
3 s
1/1
1/1


ip2vu6ptenhtnoze44peexl4vy-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/
Click to go back, hold to see history
(cat /opt/dataproc/proxy-agent/banner.html)
Spark 3.5.1
History Server
Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/8206f79f-3f10-4975-9971-13a0e29ce41b/spark-job-history
Last updated: 2024-09-22 11:57:56
Client local time zone: America/Los_Angeles
Search:
Version
App ID
App Name
Driver Host
Started
Completed
Duration
Spark User
Last Updated
Event Log
3.5.1
application_1727028058941_0001
GradientDescentTaxiDataLarge
cluster-8a64-m.c.cs777fall2024.internal
2024-09-22 11:03:16
2024-09-22 11:54:22
51 min
root
2024-09-22 11:54:23
Download
Showing 1 to 1 of 1 entries
Show incomplete applications


console.cloud.google.com/dataproc/jobs/job-055ae5f0/monitoring?job=job-055ae5f0&region=us-central1&project=cs77...
Google Cloud
cs777Fall2024
Search (/) for resources, docs, pro...
Search
Dataproc
Job details
CLONE
DELETE
STOP
REFRESH
Jobs on Clusters
Clusters
Jobs
Workflows
Autoscaling policies
Serverless
Batches
Release Notes
Job ID
job-055ae5f0
Job UUID
7e9b42ac-7de4-45ae-90ab-d8a306accdc9
Type
Dataproc Job
Status
Succeeded
Output
LINE WRAP: OFF
Iteration 48: Loss = 206.9611, m = 0.303092, b = 0.056739
24/09/22 18:53:07 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
Iteration 49: Loss = 206.2243, m = 0.309058, b = 0.057874
Iteration 50: Loss = 205.4908, m = 0.315011, b = 0.059007
Final parameters:
Slope (m): 0.3150106118230866
Intercept (b): 0.059007033958606216
24/09/22 18:54:16 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]):
24/09/22 18:54:18 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommi
24/09/22 18:54:22 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://cs777fall2024_task2hw3/task2hw3r
24/09/22 18:54:22 INFO DataprocSparkPlugin: Shutting down driver plugin. metrics=[action_http_patch_request=0, files_

SLR/SGD



34vzreuhera63papnsjkuswape-dot-us-central1.dataproc.googleusercontent.com/sparkhistory/

$(cat /opt/dataproc/proxy-agent/banner.html)

Spark 3.5.1 **History Server**

Event log directory: gs://dataproc-temp-us-central1-325310350382-liuj0ysv/a1dfea6b-a9cf-45bf-b7e7-6a39ca4572bd/spark-job-history

Last updated: 2024-09-23 19:20:48

Client local time zone: America/Los_Angeles

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.5.1	application_1727115711811_0003	MultipleLinearRegressionGD	cluster-bfe8-m.c.cs777fall2024.internal	2024-09-23 14:57:38	2024-09-23 19:01:18	4.1 h	root	2024-09-23 19:01:19	Download
3.5.1	application_1727115711811_0002	MultipleLinearRegressionGD	cluster-bfe8-m.c.cs777fall2024.internal	2024-09-23 13:11:40	2024-09-23 14:49:45	1.6 h	root	2024-09-23 14:49:46	Download
3.5.1	application_1727115711811_0001	MultipleLinearRegressionGD	cluster-bfe8-m.c.cs777fall2024.internal	2024-09-23 11:25:03	2024-09-23 13:08:02	1.7 h	root	2024-09-23 13:08:03	Download