Statistical ML for big data analysis with Spark

Group Members: Karthik Garimella, Sanjeev Raichur, Sandeep Alfred

Introduction:

Implementation of machine learning algorithms to predict the average temperature for each state in the United States over time. Considering this is a regression problem, four powerful regression methods were implemented namely, Random Forest Regression, Gradient Boosted Regression, Decision Tree Regression, and Elastic Net Linear Regression. The preprocessed data of 10,000 records was put through a pipeline of ML algorithms. Using metrics like Rsquared and RMSE, an appropriate model was determined. The entire task was performed in parallel using PySpark and different combinations of virtual machines. The computation speeds for each combination of VMs were recorded.

Link to the code: https://github.com/karthikgarimella37/SAT5165_PySpark_Project
Data Downloaded from: https://asmith.ucdavis.edu/

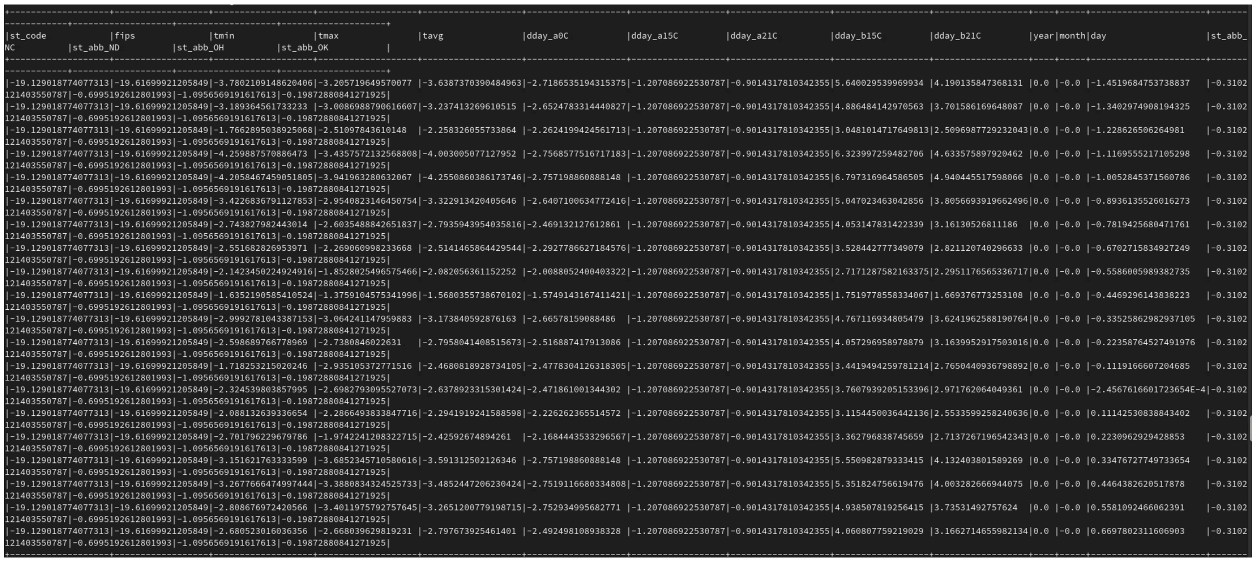
ML Methods implemented by each group member:

1. Karthik Garimella: Gradient Boosted Regression and Decision Tree Regression.
2. Sanjeev Raichur: Elastic Net Linear Regression.
3. Sandeep Alfred: Random Forest Regression.

Code Explanation and Results:

- **Data Preprocessing:**

1. Dropped highly sparse columns like **stability,** separated the day, month and year columns from date.
2. Created dummy variables for all the categorical variables.
3. Centered and scaled the data using **`StandardScaler()`** from **`pyspark.ml.feature.`** The result is shown in figure 1.

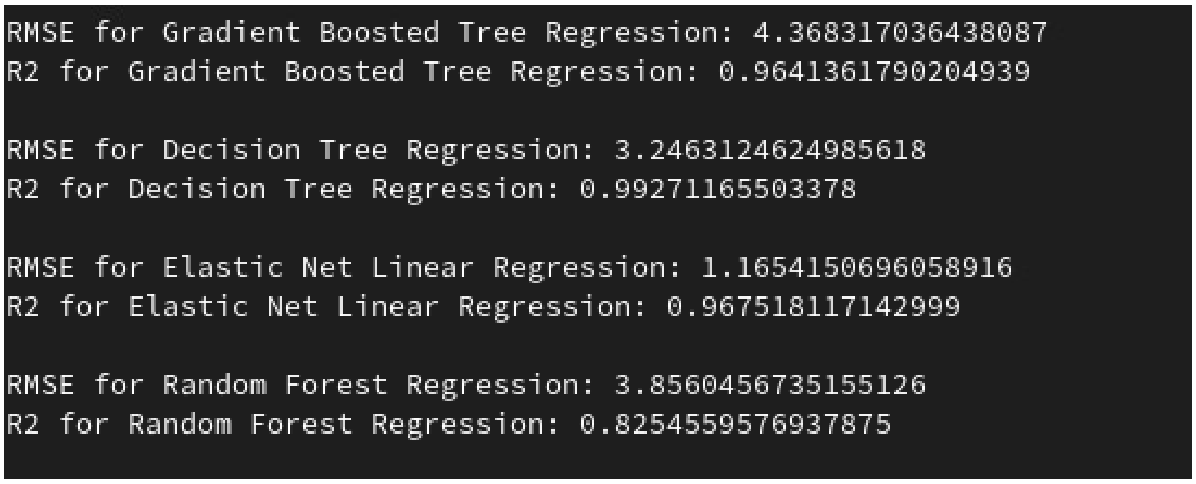


```
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+----+-----+----------------------+-------
------------+-------------------+-------------------+-------------------+
|st_code            |fips               |tmin               |tmax               |tavg               |dday_a0C           |dday_a15C          |dday_a21C          |dday_b15C          |dday_b21C          |year|month|day                   |st_abb_
NC          |st_abb_ND          |st_abb_OH          |st_abb_OK          |
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+----+-----+----------------------+-------
------------+-------------------+-------------------+-------------------+
|-19.129018774077313|-19.61699921205849|-3.7802109148620406|-3.205719649570077 |-3.6387370390484963|-2.7186535194315375|-1.207086922530787|-0.9014317810342355|5.640029539969934 |4.190135847368131 |0.0 |-0.0 |-1.4519684753738837   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-3.189364561733233 |-3.0086988790616607|-3.237413269610515 |-2.6524783314440827|-1.207086922530787|-0.9014317810342355|4.886484142970563 |3.701586169648087 |0.0 |-0.0 |-1.3402974908194325   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-1.7662895038925068|-2.51097843610148  |-2.258326055733864 |-2.2624199424561713|-1.207086922530787|-0.9014317810342355|3.0481014717649813|2.5096987729232043|0.0 |-0.0 |-1.22862650626498l    |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-4.259887570886473 |-3.4357572132568808|-4.003005077127952 |-2.7568577516717183|-1.207086922530787|-0.9014317810342355|6.323997259482706 |4.633575897920462 |0.0 |-0.0 |-1.1169555217105298   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-4.2058467459051805|-3.941963280632067 |-4.2550860386173746|-2.757198860888148 |-1.207086922530787|-0.9014317810342355|6.797316964586505 |4.940445517598066 |0.0 |-0.0 |-1.0052845371560786   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-3.4226836791127853|-2.9540823146450754|-3.322913420405646 |-2.6407100634772416|-1.207086922530787|-0.9014317810342355|5.047023463042856 |3.8056693919662496|0.0 |-0.0 |-0.8936135526016273   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.743827982443014 |-2.6035488842651837|-2.7935943954035816|-2.469132127612861 |-1.207086922530787|-0.9014317810342355|4.053147831422339 |3.16130526811186  |0.0 |-0.0 |-0.7819425680471761   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.551682826953971 |-2.269060998233668 |-2.5141465864429544|-2.2927786627184576|-1.207086922530787|-0.9014317810342355|3.528442777349079 |2.821120740296633 |0.0 |-0.0 |-0.6702715834927249   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.1423450224924916|-1.8528025496575466|-2.082056361152252 |-2.0088052400403322|-1.207086922530787|-0.9014317810342355|2.7171287582163375|2.2951176565336717|0.0 |-0.0 |-0.5586005989382735   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-1.6352190585410524|-1.3759104575341996|-1.5680355738670102|-1.5749143167411421|-1.207086922530787|-0.9014317810342355|1.7519778558334067|1.669376773253108 |0.0 |-0.0 |-0.4469296143838223   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.9992781043387153|-3.064241147959883 |-3.173840592876163 |-2.66578159088486  |-1.207086922530787|-0.9014317810342355|4.767116934805479 |3.6241962588190764|0.0 |-0.0 |-0.33525862982937105  |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.598689766778969 |-2.7380846022631   |-2.7958041408515673|-2.516887417913086 |-1.207086922530787|-0.9014317810342355|4.057296958978879 |3.1639952917503016|0.0 |-0.0 |-0.22358764527491976  |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-1.718253215020246 |-2.935105372771516 |-2.4680818928734105|-2.4778304126318305|-1.207086922530787|-0.9014317810342355|3.4419494259781214|2.7650440936798892|0.0 |-0.0 |-0.1119166607204685   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.324539803857995 |-2.6982793095527073|-2.6378923315301424|-2.471861001344302 |-1.207086922530787|-0.9014317810342355|3.7607939205153396|2.971762064049361 |0.0 |-0.0 |-2.4567616601723654E-4|-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.088132639336654 |-2.2866493833847716|-2.2941919241588598|-2.226262365514572 |-1.207086922530787|-0.9014317810342355|3.1154450036442136|2.5533599258240636|0.0 |-0.0 |0.11142530838843402   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.701796229679786 |-1.9742241208322715|-2.42592674894261  |-2.1684443533296567|-1.207086922530787|-0.9014317810342355|3.362796838745659 |2.7137267196542343|0.0 |-0.0 |0.2230962929428853    |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-3.151621763333599 |-3.6852345710580616|-3.591312502126346 |-2.757198860888148 |-1.207086922530787|-0.9014317810342355|5.550982879333415 |4.132403801589269 |0.0 |-0.0 |0.33476727749733654   |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-3.2677666474997444|-3.3880834324525733|-3.4852447206230424|-2.7519116680334808|-1.207086922530787|-0.9014317810342355|5.351824756619476 |4.003282666944075 |0.0 |-0.0 |0.4464382620517878    |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.808676972420566 |-3.4011975792757645|-3.2651200779198715|-2.752934995682771 |-1.207086922530787|-0.9014317810342355|4.938507819256415 |3.73531492757624  |0.0 |-0.0 |0.5581092466062391    |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
|-19.129018774077313|-19.61699921205849|-2.680523016036356 |-2.668039629819231 |-2.797673925461401 |-2.492498108938328 |-1.207086922530787|-0.9014317810342355|4.060807759219029 |3.1662714655982134|0.0 |-0.0 |0.6697802311606903    |-0.3102
121403550787|-0.6995192612801993|-1.0956569191617613|-0.19872880841271925|
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+----+-----+----------------------+-------
```

Figure 1. Centered and Scaled data

- **Implementation of ML algorithms:**

1. Implemented `Pipeline` from `pyspark.ml` to run several ML models.
2. Used `VectorAssembler` from `pyspark.ml.feature` to merge all the columns of the data into a vector column.
3. Used `randomSplit` method with seed 123 to split the data into training and testing sets. 80% of the data was used to train the model and tested on the remaining 20% of the data.
4. From `pyspark.ml.regression` used `GBTRegressor` for Gradient Boosted Tree Regression with `maxIter` of 10.
5. Implemented Decision Tree Regression using `DecisionTreeRegressor` function.
6. Elastic Net Linear Regression was implemented with the `LinearRegression` function, using parameters of `maxIter` of 10.
7. `RandomForestRegressor` was used to implement Random Forest Regression. All the functions for the models were implemented from `pyspark.ml.regression`.
8. In the pipeline, initialized each model with `t_avg` as the response variable, which is the average temperature for each state.
9. The models were evaluated on the test set. Used `RegressionEvaluator` from `pyspark.ml.evaluation` to calculate RMSE and Rsquared for each model. The evaluation results are shown in figure 2.



```
RMSE for Gradient Boosted Tree Regression: 4.368317036438087
R2 for Gradient Boosted Tree Regression: 0.9641361790204939

RMSE for Decision Tree Regression: 3.2463124624985618
R2 for Decision Tree Regression: 0.99271165503378

RMSE for Elastic Net Linear Regression: 1.1654150696058916
R2 for Elastic Net Linear Regression: 0.967518117142999

RMSE for Random Forest Regression: 3.8560456735155126
R2 for Random Forest Regression: 0.8254559576937875
```

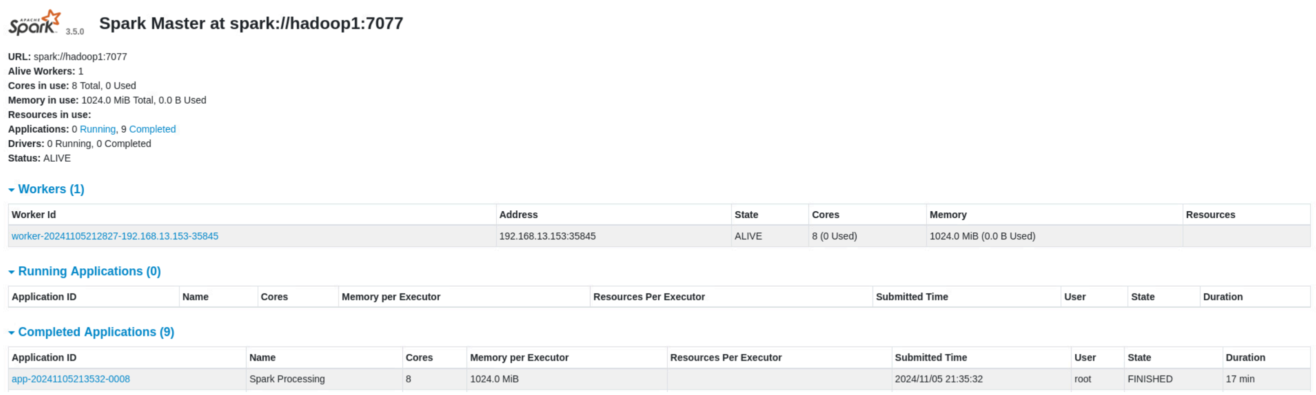
Figure 2. Model Evaluation Results on the Test Set.

From figure 2 we can see that Elastic Net Linear Regression has the lowest RMSE value of 1.16541, and a very good RMSE of 0.9675.

Performance of PySpark:

The performance of Spark to implement all the mentioned ML algorithms for each combination of virtual machines is mentioned below.

- The discussed code with preprocessing and ML algorithms using 10,000 rows of the data when run on 1 machine took 17 mins to run, as shown in figure 3.



Spark 3.5.0 **Spark Master at spark://hadoop1:7077**

URL: spark://hadoop1:7077
Alive Workers: 1
Cores in use: 8 Total, 0 Used
Memory in use: 1024.0 MiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 9 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241105212827-192.168.13.153-35845	192.168.13.153:35845	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

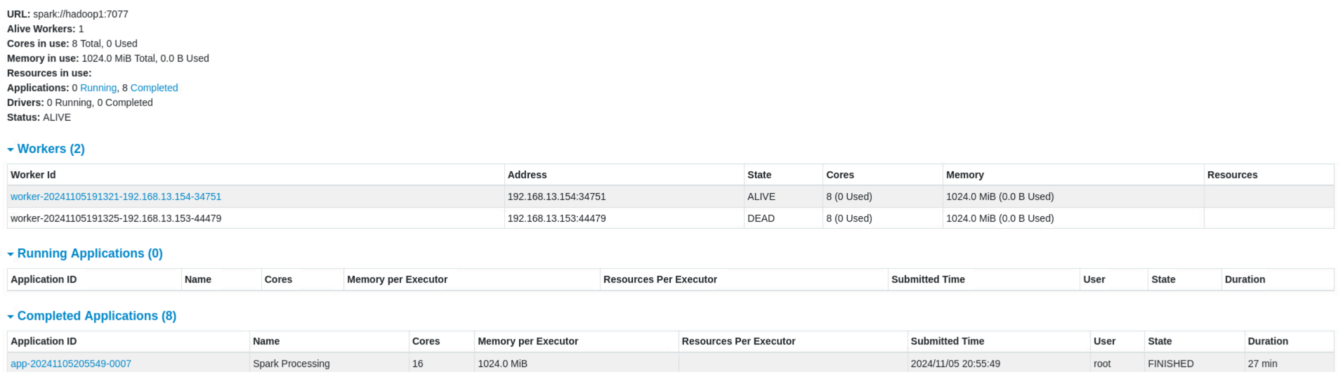
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (9)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241105213532-0008	Spark Processing	8	1024.0 MiB		2024/11/05 21:35:32	root	FINISHED	17 min

Figure 3. Execution time of 1 VM, 8 cores.

- The discussed code with preprocessing and ML algorithms using 10,000 rows of the data when run on 2 machines took 27 mins to run, as shown in figure 4. This was not expected with two VMs.



URL: spark://hadoop1:7077
Alive Workers: 1
Cores in use: 8 Total, 0 Used
Memory in use: 1024.0 MiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 8 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241105191321-192.168.13.154-34751	192.168.13.154:34751	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191325-192.168.13.153-44479	192.168.13.153:44479	DEAD	8 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

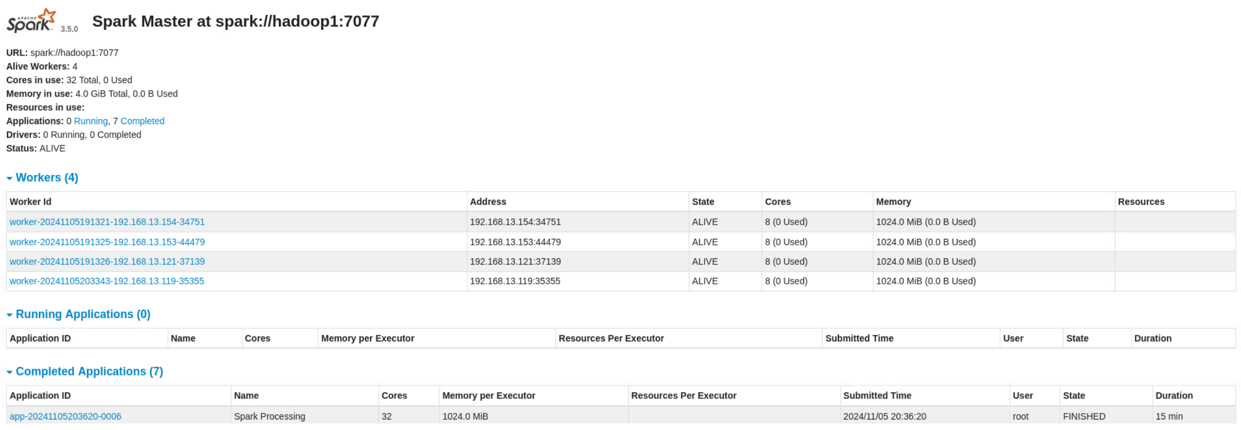
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (8)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241105205549-0007	Spark Processing	16	1024.0 MiB		2024/11/05 20:55:49	root	FINISHED	27 min

Figure 4. Execution time of 2 VMs, 16 cores.

- The discussed code with preprocessing and ML algorithms using 10,000 rows of the data when run on 4 machines took 15 mins to run, as shown in figure 5.



3.5.0 **Spark Master at spark://hadoop1:7077**

URL: spark://hadoop1:7077
Alive Workers: 4
Cores in use: 32 Total, 0 Used
Memory in use: 4.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 7 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241105191321-192.168.13.154-34751	192.168.13.154:34751	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191325-192.168.13.153-44479	192.168.13.153:44479	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191326-192.168.13.121-37139	192.168.13.121:37139	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105203343-192.168.13.119-35355	192.168.13.119:35355	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

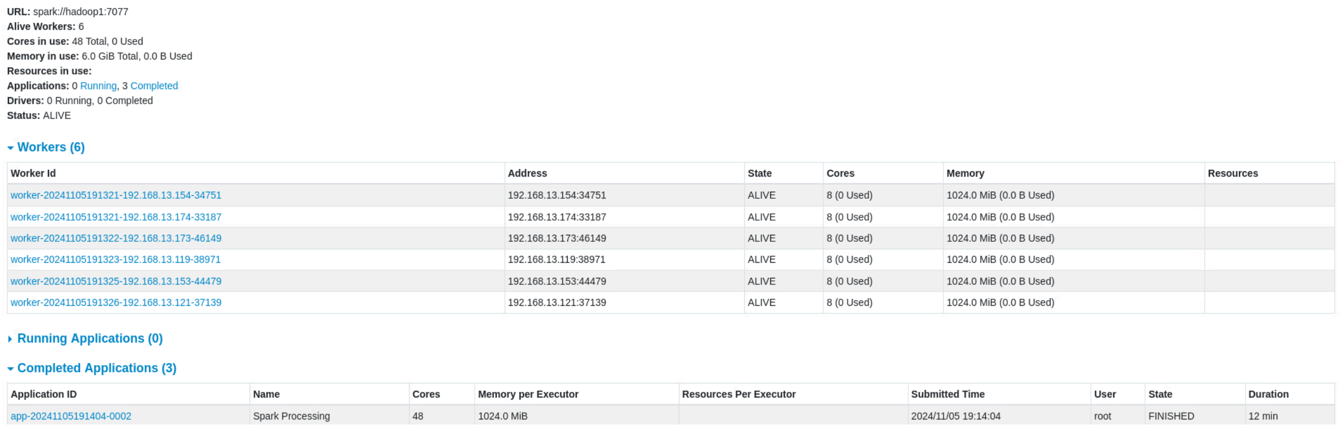
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (7)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241105203620-0006	Spark Processing	32	1024.0 MiB		2024/11/05 20:36:20	root	FINISHED	15 min

Figure 5. Execution time of 4 VMs, 32 cores

- The discussed code with preprocessing and ML algorithms using 10,000 rows of the data when run on all 6 machines took 12 mins to run, as shown in figure 6.



URL: spark://hadoop1:7077
Alive Workers: 6
Cores in use: 48 Total, 0 Used
Memory in use: 6.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 3 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (6)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241105191321-192.168.13.154-34751	192.168.13.154:34751	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191321-192.168.13.174-33187	192.168.13.174:33187	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191322-192.168.13.173-46149	192.168.13.173:46149	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191323-192.168.13.119-38971	192.168.13.119:38971	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191325-192.168.13.153-44479	192.168.13.153:44479	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	
worker-20241105191326-192.168.13.121-37139	192.168.13.121:37139	ALIVE	8 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

Completed Applications (3)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241105191404-0002	Spark Processing	48	1024.0 MiB		2024/11/05 19:14:04	root	FINISHED	12 min

Figure 6. Execution time of 6 VMs, 48 cores.

Performance summary to run the ML methods on 10,000 records of data:

Number of VMs Used	Time Taken	Cores Used
One	17 minutes	8
Two	27 minutes	16
Four	15 minutes	32
Six	12 minutes	48