

Name: Abishek Vellineni

Email Id: av739@njit.edu

Module 07 Assignment 01: Programming Assignment 2

Wine Quality Prediction AWS Spark Application

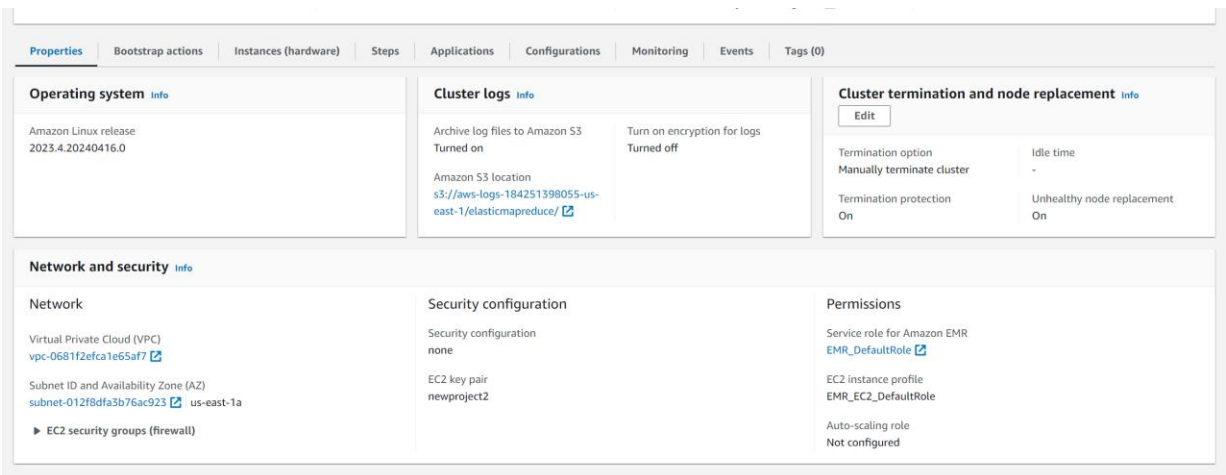
The application is deployed on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. The primary objective is to parallelly train a machine learning model on EC2 instances for predicting wine quality using publicly available data. Subsequently, the trained model is employed to predict the quality of wine. Docker is utilized to create a container image for the trained machine learning model, streamlining the deployment process.

Link for GitHub - https://github.com/Abishek183/wine_prediction

Link for Docker - https://hub.docker.com/repository/docker/abishek183/wine_predict/tags

Cluster Creation AWS SPARK:

1. Navigate to the EMR console and then click on create new instance.
2. Provide the name for your cluster.
3. In cluster Termination, change to manual from automatic.
4. Provide the key_pair which is a .pem file in security configuration



5. In instance creation, provide 1 for core and 4 for tasks as we need to run on 4 EC2 instances.
6. Then select default roles for the IAM roles.

Properties | Bootstrap actions | **Instances (hardware)** | Steps | Applications | Configurations | Monitoring | Events | Tags (0)

Instance group settings info Edit cluster scaling option

Cluster scaling option
Manually set cluster size

Core
Name and maximum core nodes in the cluster
Core: 1 instances

Task
Name and maximum task nodes in the cluster
Task - 1: 4 instances

Instance groups (3) info Terminate instance Resize instance group Add task instance group

With the instance groups configuration, each node type consists of the same instance type and the same purchasing option for instances: on-demand or spot.

Find instances by status Find resources by ID or type; or search for text within loaded results

Type and name	ID	Status Last state change reason	Instances	Purchasing option and p...	EBS size (GiB)	EC2 instance ID
Primary	ig-21YXG8XL97N	Running	1	On-demand	-	-
Core	ig-27S23Z7KOU018	Running	1	On-demand	-	-
Task (Task - 1)	ig-3OUBUOCQM2B6	Running	4	On-demand	-	-

- Create a S3 queue to upload the python and csv files.
s3://winepredcit

aws Services Search [Alt+S] N. Virginia voclabs/user3025972-av/739@njit.edu @ 1842-5139-8055

Amazon S3 > Buckets > winepredcit

winepredcit Info

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (5) Info Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

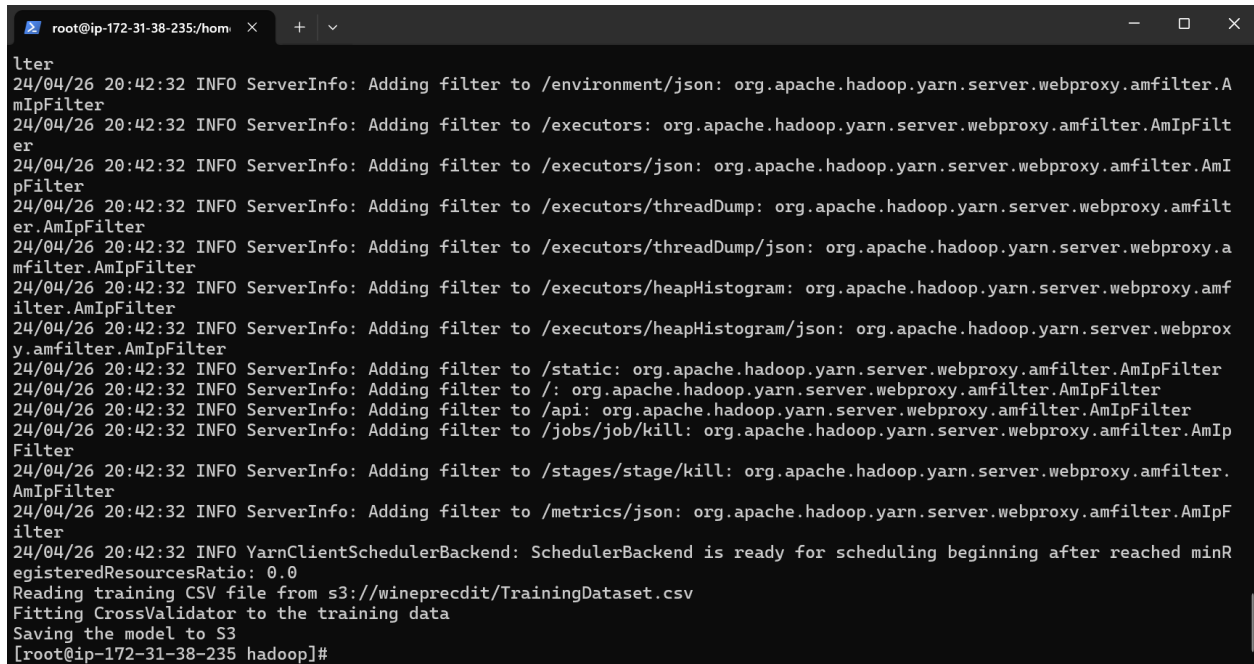
Name	Type	Last modified	Size	Storage class
prediction.py	py	April 25, 2024, 13:57:56 (UTC-04:00)	2.0 KB	Standard
trainedmodel/	Folder	-	-	-
training.py	py	April 25, 2024, 13:57:57 (UTC-04:00)	3.5 KB	Standard
TrainingDataset.csv	csv	April 24, 2024, 20:57:07 (UTC-04:00)	67.2 KB	Standard
ValidationDataset.csv	csv	April 24, 2024, 20:57:07 (UTC-04:00)	8.6 KB	Standard

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

- Open terminal and use the below command to connect to the cluster.
ssh -i ~/newproject2.pem hadoop@ec2-52-201-250-228.compute-1.amazonaws.com

Execution without Docker

- Run “sudo su” command to change user.
- Install numpy by using ‘pip install numpy --user’
- Then run “spark-submit s3://wineprecidit/training.py”. It runs the file from S3 bucket and creates a ML model

A terminal window with a dark background and light text. The window title bar shows 'root@ip-172-31-38-235/hom'. The terminal output consists of multiple log lines from the Spark framework. It starts with 'lter' on a new line. Then, several lines show 'INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter'. This is followed by similar log entries for '/executors', '/executors/threadDump', and '/executors/heapHistogram'. After these, there are log entries for '/static', '/', and '/api'. Then, there are log entries for '/jobs/job/kill' and '/stages/stage/kill'. Finally, there are log entries for '/metrics/json' and 'YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0'. The last few lines of the log show 'Reading training CSV file from s3://wineprecidit/TrainingDataset.csv', 'Fitting CrossValidator to the training data', and 'Saving the model to S3'. The prompt '[root@ip-172-31-38-235 hadoop]#' is visible at the bottom.

```
lter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:42:32 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Reading training CSV file from s3://wineprecidit/TrainingDataset.csv
Fitting CrossValidator to the training data
Saving the model to S3
[root@ip-172-31-38-235 hadoop]#
```

- Then run “spark-submit s3://wineprecidit/prediction.py s3://wineprecidit/ValidationDataset.csv”. It uses the model created and validates the data from the csv file and provides the result.
- We can infer that it provides an accuracy of 95.479% from the below image.

```
root@ip-172-31-38-235/hom x + v
mIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /api: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /jobs/job/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /stages/stage/kill: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:45:15 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Accuracy = 0.96875
/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/context.py:158: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
F1 Score = 0.9547916666666667
Exiting Spark Application
[root@ip-172-31-38-235 hadoop]#
```

Execution with Docker


- Run the below commands to start the docker in the EC2.
 - `sudo systemctl start docker`
 - `sudo systemctl enable docker`
- Get the image from docker repo using the below command.
 - `sudo docker pull abishek183/wine_predict:train`
 - `sudo docker pull abishek183/wine_predict:predict`
- Run the train tag image to create a ML model using the below command.
 - `sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:train`

```
root@ip-172-31-38-235:/hom x + v
40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/fetchFileTemp1858610424113670995.tmp
24/04/26 20:48:16 INFO Utils: /tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/fetchFileTemp1858610424113670995.tmp has been previously copied to /tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/joda-time_joda-time-2.12.7.jar
24/04/26 20:48:16 INFO Executor: Adding file:/tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/joda-time_joda-time-2.12.7.jar to class loader default
24/04/26 20:48:16 INFO Executor: Fetching spark://2c2b3eab375a:46493/jars/org.apache.directory.api_api-asn1-api-1.0.0-M20.jar with timestamp 1714164494437
24/04/26 20:48:16 INFO Utils: Fetching spark://2c2b3eab375a:46493/jars/org.apache.directory.api_api-asn1-api-1.0.0-M20.jar to /tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/fetchFileTemp8265105440123927859.tmp
24/04/26 20:48:16 INFO Utils: /tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/fetchFileTemp8265105440123927859.tmp has been previously copied to /tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/org.apache.directory.api_api-asn1-api-1.0.0-M20.jar
24/04/26 20:48:16 INFO Executor: Adding file:/tmp/spark-ed2c40aa-c9d3-44db-8852-c1676e841cd1/userFiles-f600e761-6cd8-4fb1-a177-7630d193d908/org.apache.directory.api_api-asn1-api-1.0.0-M20.jar to class loader default
24/04/26 20:48:16 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 42989.
24/04/26 20:48:16 INFO NettyBlockTransferService: Server created on 2c2b3eab375a:42989
24/04/26 20:48:16 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 20:48:16 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 2c2b3eab375a, 42989, None)
24/04/26 20:48:16 INFO BlockManagerMasterEndpoint: Registering block manager 2c2b3eab375a:42989 with 366.3 MiB RAM, BlockManagerId(driver, 2c2b3eab375a, 42989, None)
24/04/26 20:48:16 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 2c2b3eab375a, 42989, None)
24/04/26 20:48:16 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 2c2b3eab375a, 42989, None)
Reading training CSV file from TrainingDataset.csv
Fitting CrossValidator to the training data
Saving the model
[root@ip-172-31-38-235 hadoop]# |
```

- Run the image with predict tag to get the accuracy.
 - `sudo docker run -v /home/ec2-user/:/job abishek183/wine_predict:predict ValidationDataset.csv`
- we can infer from the below image the accuracy is 95%.

```
root@ip-172-31-38-235:/hom x + v
6ed0/fetchFileTemp5806784952250657780.tmp has been previously copied to /tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/com.sun.jersey_jersey-server-1.9.jar
24/04/26 20:50:47 INFO Executor: Adding file:/tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/com.sun.jersey_jersey-server-1.9.jar to class loader default
24/04/26 20:50:47 INFO Executor: Fetching spark://d1c88003a569:35613/jars/org.codehaus.jackson_jackson-mapper-asl-1.9.13.jar with timestamp 1714164644878
24/04/26 20:50:47 INFO Utils: Fetching spark://d1c88003a569:35613/jars/org.codehaus.jackson_jackson-mapper-asl-1.9.13.jar to /tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/fetchFileTemp6958102296460489263.tmp
24/04/26 20:50:47 INFO Utils: /tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/fetchFileTemp6958102296460489263.tmp has been previously copied to /tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/org.codehaus.jackson_jackson-mapper-asl-1.9.13.jar
24/04/26 20:50:47 INFO Executor: Adding file:/tmp/spark-c82b1bdb-732a-405a-a190-b1704aac1dc5/userFiles-00b14209-4e8b-4a21-9a0e-dfd90d536ed0/org.codehaus.jackson_jackson-mapper-asl-1.9.13.jar to class loader default
24/04/26 20:50:47 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44469.
24/04/26 20:50:47 INFO NettyBlockTransferService: Server created on d1c88003a569:44469
24/04/26 20:50:47 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 20:50:47 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, d1c88003a569, 44469, None)
24/04/26 20:50:47 INFO BlockManagerMasterEndpoint: Registering block manager d1c88003a569:44469 with 366.3 MiB RAM, BlockManagerId(driver, d1c88003a569, 44469, None)
24/04/26 20:50:47 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, d1c88003a569, 44469, None)
24/04/26 20:50:47 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, d1c88003a569, 44469, None)
Accuracy = 0.96875
/opt/spark/python/lib/pyspark.zip/pyspark/sql/context.py:160: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
F1 Score= 0.9547916666666667
Exiting Spark Application
[root@ip-172-31-38-235 hadoop]# |
```

Docker Image:



abishek183/wine_predict:train

MANIFEST DIGEST: sha256:d5eb9d5ee75a24457b9b7a78d25f93a2c0f6de47b3240313a0c24804899636

OS/ARCH

linux/amd64

COMPRESSED SIZE

815.25 MB

LAST PUSHED

an hour ago by [abishek183](#)

TYPE

Image

MANIFEST DIGEST

sha256:d5eb9d5ee...


Delete Tag

Image Layers

Vulnerabilities

IMAGE LAYERS

1	ADD file ... in /	72.57 MB	Command
2	LABEL org.label-schema.schema-version=1.0 org.label-	0 B	ADD file:b3ebbe8bd384723d43b7b44ade998cd57b63d93de2af293983a30bfc1dfa53 in /
3	CMD ["bin/bash"]	0 B	
4	RUN /bin/sh -c yum -y	212.55 MB	
5	RUN /bin/sh -c python3 -V	93 B	
6	ENV PYTHON_DRIVER_PYTHON-python3	0 B	
7	ENV PYTHON_PYTHON-python3	0 B	
8	RUN /bin/sh -c pip3 install	4.48 MB	
9	RUN /bin/sh -c pip3 install	26.77 MB	
10	RUN /bin/sh -c pip3 install	53.36 MB	
11	WORKDIR /opt	32 B	
12	RUN /bin/sh -c wget --no-verbose	382.41 MB	
13	RUN /bin/sh -c yum -y	52.13 MB	
14	ENV SPARK_HOME=/opt/spark	0 B	
15	ENV PATH=/opt/spark/bin:/usr/local/sbin:/usr/local/..	0 B	
16	RUN /bin/sh -c wget https://repo.maven.org/m...	18.69 MB	
17	RUN /bin/sh -c wget https://repo.maven.org/m...	278.38 KB	
18	ADD training.py . # buildkit	1.35 KB	



abishek183/wine_predict:predict

MANIFEST DIGEST: sha256:46d233c6c234306723f92d78209484e5d2d485d70e5b80243267eeb95ba35

OS/ARCH

linux/amd64

COMPRESSED SIZE

815.24 MB

LAST PUSHED

2 hours ago by [abishek183](#)

TYPE

Image

MANIFEST DIGEST

sha256:46d233dc...

Delete Tag

Image Layers

Vulnerabilities

IMAGE LAYERS

1	ADD file ... in /	72.57 MB	Command
2	LABEL org.label-schema.schema-version=1.0 org.label-	0 B	ADD file:b3ebbe8bd384723d43b7b44ade998cd57b63d93de2af293983a30bfc1dfa53 in /
3	CMD ["bin/bash"]	0 B	
4	RUN /bin/sh -c yum -y	212.55 MB	
5	RUN /bin/sh -c python3 -V	93 B	
6	ENV PYTHON_DRIVER_PYTHON-python3	0 B	
7	ENV PYTHON_PYTHON-python3	0 B	
8	RUN /bin/sh -c pip3 install	4.48 MB	
9	RUN /bin/sh -c pip3 install	26.77 MB	
10	RUN /bin/sh -c pip3 install	53.36 MB	
11	WORKDIR /opt	32 B	
12	RUN /bin/sh -c wget --no-verbose	382.41 MB	
13	RUN /bin/sh -c yum -y	52.13 MB	
14	ENV SPARK_HOME=/opt/spark	0 B	
15	ENV PATH=/opt/spark/bin:/usr/local/sbin:/usr/local/..	0 B	
16	RUN /bin/sh -c wget https://repo.maven.org/m...	18.69 MB	
17	RUN /bin/sh -c wget https://repo.maven.org/m...	278.38 KB	
18	ADD prediction.py . # buildkit	1.82 KB	