Real-time Early Heart Disease Detection using Apache Spark

FALL 2022 CPSC 531-02, PROFESSOR TSENG-CHING SHEN

Venkat Reddy (CWID: 885187195), Andres Medina (CWID: 891867624), Brijesh Prajapati (CWID: 885909267)

Github link to the project

https://github.com/Brij98/cpsc531-project

Problem Statement

According to the World Health Organization (WHO), stroke and heart attack account for 85% of the 35% of global deaths caused by cardiovascular diseases (CVD).

Due to the severity of heart attack and stroke, early and instant detection of heart diseases can allow patients to react in advance to a probable heart ailment.

Hence, our project aims to develop a real-time heart-disease monitoring system that will allow people to detect any heart abnormality instantaneously.

Dataset for the Project

The heart disease dataset used for this project is available on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/heart+disease). The dataset contains a total of 900 records collected from four different locations namely Budapest, Zurich, Basel, Long Beach.

The dataset has the following 14 attributes:

1. age: The person's age in years
2. sex: The person's sex (1 = male, 0 = female)
3. cp: The chest pain experienced (Value 1:typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4:asymptomatic)
4. trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5. chol: The person's cholesterol measurement in mg/dl
6. fbs: The person's fasting blood sugar (>120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. thalach: The person's maximum heart rate achieved
9. exang: Exercise induced angina (1 = yes; 0 = no)

10. oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
11. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. ca: The number of major vessels (0-3)
13. thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. target: Heart disease (0 = no, 1 = yes)

```
32,1,1,95,0,?,0,127,0,.7,1,?,?,1
34,1,4,115,0,?,?,154,0,.2,1,?,?,1
35,1,4,?,0,?,0,130,1,?,?,?,7,3
```

Fig1. Shows a sample dataset

Functionalities

The system has the following functionalities:

- **Kafka**
  - Generate random heart diseases data for prediction.
  - Using the random heart diseases data produce a message.
  - Push the message to Kafka cluster.
- **Spark Streaming**
  - Read real-time data from Kafka cluster.
  - Process input data to be predicted.
  - Load the ML model.
  - Predict the data.
  - Save the predicted data to Cassandra DB.
- **Spark ML**
  - Read data from the CSV file.
  - Process data by removing any records with null fields.
  - Train Random Forest classifier.
  - Test the Random Forest model.
  - Save the model.

- **Cassandra DB**
    - Generate queried data for analysis.

Architecture and Design

The purpose of this system is to predict and store heart-disease data in real-time. The system consists of three main parts data source, Apache Spark application, and data storage.

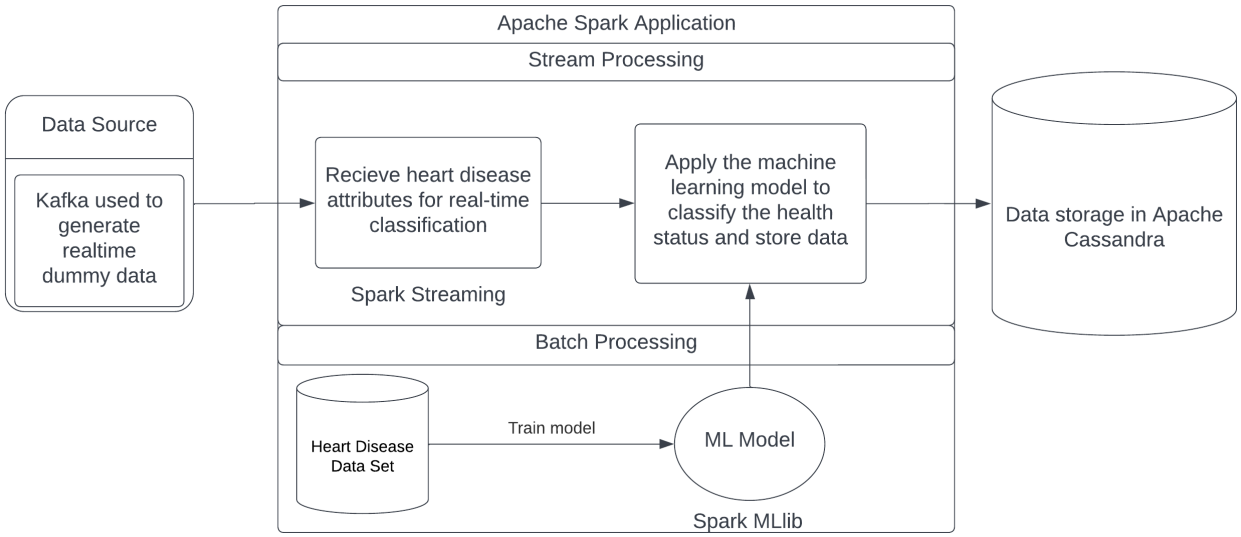


Fig 2. demonstrates the architecture of the system.

Data Source

Uses Apache Kafka which is an open-source event streaming platform to stream randomly generated heart-diseases data in real-time for predictions to spark streaming.

Apache Spark Application

The spark streaming is a module that supports scalable, fault-tolerant processing of live data streams. The spark streaming is a consumer that consumes the heart-diseases data to be predicted from Kafka streaming server. The spark application then uses the trained machine learning model to predict the heart-disease data and then store it in Cassandra DB.

The Spark application uses pyspark ml to train a random forest classifier. Random forest classifier is an ensemble of decision trees that use votes to decide the final

prediction. To train the random forest classifier we included 10 decision trees with max depth of 7.

Cassandra DB

Cassandra DB an open source, distributed database which is used to store the predicted heart-disease data that is received in real-time from spark streaming. The stored data in Cassandra is than later queried to find useful information of the stored data.

Technologies used

All the development was done using Python3 and PySpark.

- Apache Kafka
  - Version: 3.3.1
  - Prerequisites: java 11, kafka manager (CMAK) to configure the topics in cluster.
- Apache Spark
  - Version: 3.3.1
  - Prerequisites: java 11 and python3
- Cassandra (note: Cassandra3 only works on java 1.8 and python2)
  - Version: Cassandra 4.0.7
  - Prerequisites: java 11 and python3

Steps to Run the Application and Deployment

Kafka

First start Zookeeper

Configurations:

Zookeeper default port : 2181

In cmd run : bin/zookeeper-server-start.sh config/zookeeper.properties\

```
csuftitan@LTMacAir-M1-165166 kafka_2.13-3.3.1 % bin/zookeeper-server-start.sh
config/zookeeper.properties
```

Zookeeper running:

```
[2022-12-10 22:12:19,641] INFO Created server with tickTime 3000 minSessionTim
eout 6000 maxSessionTimeout 60000 clientPortListenBacklog -1 datadir /tmp/zook
eeper/version-2 snapdir /tmp/zookeeper/version-2 (org.apache.zookeeper.server.
ZooKeeperServer)
[2022-12-10 22:12:19,648] INFO Using org.apache.zookeeper.server.NIOServerCnxn
Factory as server connection factory (org.apache.zookeeper.server.ServerCnxnFa
ctory)
[2022-12-10 22:12:19,648] WARN maxCnxns is not configured, using default value
 0. (org.apache.zookeeper.server.ServerCnxnFactory)
[2022-12-10 22:12:19,649] INFO Configuring NIO connection handler with 10s ses
sionless connection timeout, 2 selector thread(s), 16 worker threads, and 64 k
B direct buffers. (org.apache.zookeeper.server.NIOServerCnxnFactory)
[2022-12-10 22:12:19,653] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apach
e.zookeeper.server.NIOServerCnxnFactory)
[2022-12-10 22:12:19,667] INFO Using org.apache.zookeeper.server.watch.WatchMa
nager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2022-12-10 22:12:19,667] INFO Using org.apache.zookeeper.server.watch.WatchMa
nager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2022-12-10 22:12:19,668] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache
```

Second start kafka:

Kafka default port : 9092

JMX_PORT=8004 bin/kafka-server-start.sh config/server.properties

```
csuftitan@LTMacAir-M1-165166 kafka_2.13-3.3.1 % JMX_PORT=8004 bin/kafka-server
-start.sh config/server.properties
```

```
3b94a3ee-a40d-45e6-b821-ae370d8714ec, groupInstanceId=None, clientId=consumer-
spark-kafka-source-6530eed1-6bfd-43ca-b332-36e232f67ebb--189583945-driver-0-1,
 clientHost=/127.0.0.1, sessionTimeoutMs=10000, rebalanceTimeoutMs=300000, sup
portedProtocols=List(range)) in group spark-kafka-source-6530eed1-6bfd-43ca-b3
32-36e232f67ebb--189583945-driver-0 with generation 1. (kafka.coordinator.grou
p.GroupMetadata$)
[2022-12-10 22:20:15,821] INFO [GroupMetadataManager brokerId=0] Finished load
ing offsets and group metadata from __consumer_offsets-36 in 33 milliseconds f
or epoch 0, of which 32 milliseconds was spent in the scheduler. (kafka.coordi
nator.group.GroupMetadataManager)
[2022-12-10 22:20:15,823] INFO Loaded member MemberMetadata(memberId=consumer-
spark-kafka-source-1ec15e6f-0e24-4838-9c34-6c546b902861-984749602-driver-0-1-a
0f2c219-b3c9-4284-846b-5ab71b95749b, groupInstanceId=None, clientId=consumer-s
park-kafka-source-1ec15e6f-0e24-4838-9c34-6c546b902861-984749602-driver-0-1, c
lientHost=/127.0.0.1, sessionTimeoutMs=10000, rebalanceTimeoutMs=300000, suppo
rtedProtocols=List(range)) in group spark-kafka-source-1ec15e6f-0e24-4838-9c34
-6c546b902861-984749602-driver-0 with generation 1. (kafka.coordinator.group.G
roupMetadata$)
```

KAFKA MANAGER : CMAK

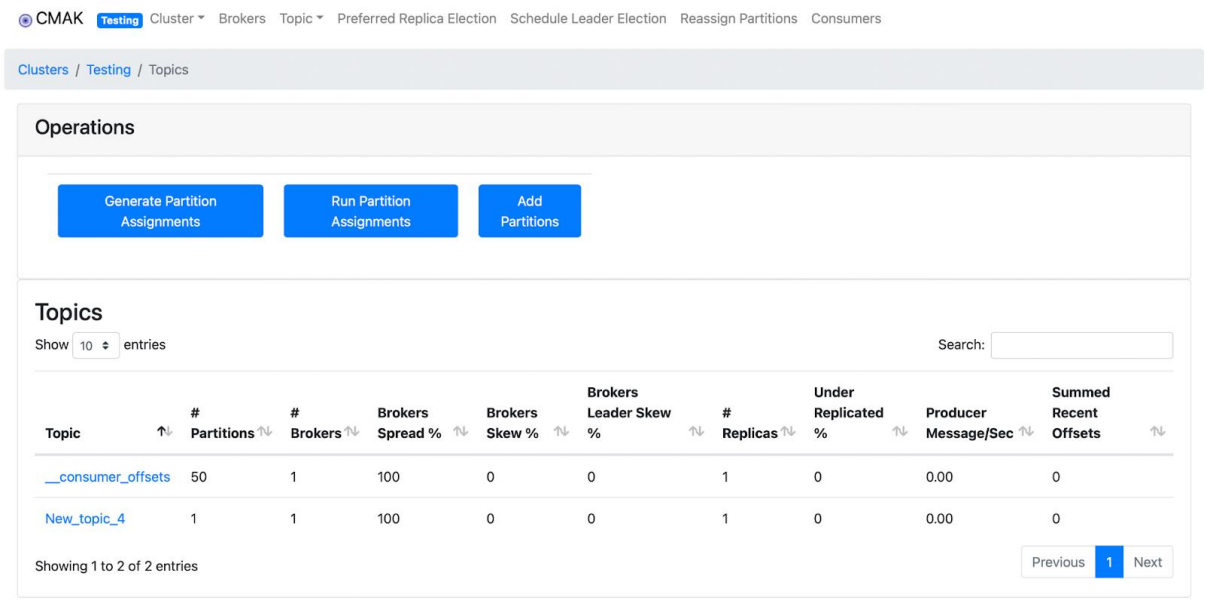
Kafka Manager Default : 8090

bin/cmak -Dconfig.file=conf/application.conf -Dhttp.port=8090

```
csuftitan@LTMacAir-M1-165166 cmak-3.0.0.6 % bin/cmak -Dconfig.file=conf/applic
ation.conf -Dhttp.port=8090
```

```
2022-12-10 22:27:11,475 - [INFO] k.m.a.DeleteClusterActor - Started actor akka
://kafka-manager-system/user/kafka-manager/delete-cluster
2022-12-10 22:27:11,475 - [INFO] k.m.a.KafkaManagerActor - Started actor akka:
//kafka-manager-system/user/kafka-manager
2022-12-10 22:27:11,475 - [INFO] k.m.a.KafkaManagerActor - Starting delete clu
sters path cache...
2022-12-10 22:27:11,475 - [INFO] k.m.a.DeleteClusterActor - Starting delete cl
usters path cache...
2022-12-10 22:27:11,481 - [INFO] k.m.a.DeleteClusterActor - Adding kafka manag
er path cache listener...
2022-12-10 22:27:11,481 - [INFO] k.m.a.KafkaManagerActor - Starting kafka mana
ger path cache...
2022-12-10 22:27:11,481 - [INFO] k.m.a.DeleteClusterActor - Scheduling updater
 for 10 seconds
2022-12-10 22:27:11,485 - [INFO] k.m.a.KafkaManagerActor - Adding kafka manage
r path cache listener...
2022-12-10 22:27:11,871 - [INFO] p.c.s.AkkaHttpServer - Listening for HTTP on
/0:0:0:0:0:0:0:0:8090
2022-12-10 22:27:12,506 - [INFO] k.m.a.KafkaManagerActor - Updating internal s
tate...
2022-12-10 22:27:22,501 - [INFO] k.m.a.KafkaManagerActor - Updating internal s
tate...
```

CMAK:



Spark Streaming

sbin/start-all.sh

```
csuftitan@LTMacAir-M1-165166 spark-3.3.1-bin-hadoop3 % sbin/start-all.sh
```

```
csuftitan@LTMacAir-M1-165166 spark-3.3.1-bin-hadoop3 % sbin/start-all.sh
org.apache.spark.deploy.master.Master running as process 2438.  Stop it first.
localhost: org.apache.spark.deploy.worker.Worker running as process 2488.  Sto
p it first.
csuftitan@LTMacAir-M1-165166 spark-3.3.1-bin-hadoop3 % sbin/start-all.sh
org.apache.spark.deploy.master.Master running as process 2438.  Stop it first.
localhost: org.apache.spark.deploy.worker.Worker running as process 2488.  Sto
p it first.
csuftitan@LTMacAir-M1-165166 spark-3.3.1-bin-hadoop3 % jps
27043 Kafka
2438 Master
17975
26583 QuorumPeerMain
2488 Worker
28216 Jps
28026 ProdServerStart
20302 RemoteJdbcServer
20303 RemoteJdbcServer
csuftitan@LTMacAir-M1-165166 spark-3.3.1-bin-hadoop3 %
```

Cassandra :

./bin/cassandra -f

```
csuftitan@LTMacAir-M1-165166 apache-cassandra-4.0.7 % ./bin/cassandra -f
```

```
 Compacted (ff8d3f20-791e-11ed-a2a3-bbf8a0e26096) 5 sstables to [/Users/csufti
tan/apache-cassandra-4.0.7/data/data/system/local-7ad54392bcdd35a684174e047860
b377/nb-16-big,] to level=0.  1.038KiB to 0.668KiB (~64% of original) in 585ms
.  Read Throughput = 1.772KiB/s, Write Throughput = 1.141KiB/s, Row Throughput
 = ~2/s.  5 total partitions merged to 1.  Partition merge counts were {5:1, }
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,776 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/compaction_
history-b4dbb7b4dc493fb5b3bfce6e434832ca/nb-8-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,780 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/compaction_
history-b4dbb7b4dc493fb5b3bfce6e434832ca/nb-5-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,783 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/compaction_
history-b4dbb7b4dc493fb5b3bfce6e434832ca/nb-7-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,787 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/local-7ad54
392bcdd35a684174e047860b377/nb-15-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,789 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/local-7ad54
392bcdd35a684174e047860b377/nb-14-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,793 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/local-7ad54
392bcdd35a684174e047860b377/nb-12-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,796 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/local-7ad54
392bcdd35a684174e047860b377/nb-13-big
INFO  [NonPeriodicTasks:1] 2022-12-10 22:42:36,798 SSTable.java:111 - Deleting
 sstable: /Users/csuftitan/apache-cassandra-4.0.7/data/data/system/local-7ad54
392bcdd35a684174e047860b377/nb-11-big
```

To Start the Cassandra Shell :

./bin/cqlsh

- then navigate into the keyspace or create the new key space
- Then open or create the new table as per the schema

```
csuftitan@LTMacAir-M1-165166 apache-cassandra-4.0.7 % ./bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.0.0 | Cassandra 4.0.7 | CQL spec 3.4.5 | Native protocol v5]
Use HELP for help.
cqlsh> use cas
   ... ;
cqlsh:cas>
```

Test Results

- As we are Generating the Simulating real time data using Kafa.
- And using Apache PySpark ML for machine learning model
- And also using Apache PySpark Streaming to using the ML model to predict the heart disease.
- And final using Cassandra to store final prediction data.
- And perform base queries on top it.

We have First Run the PySpark Streaming :

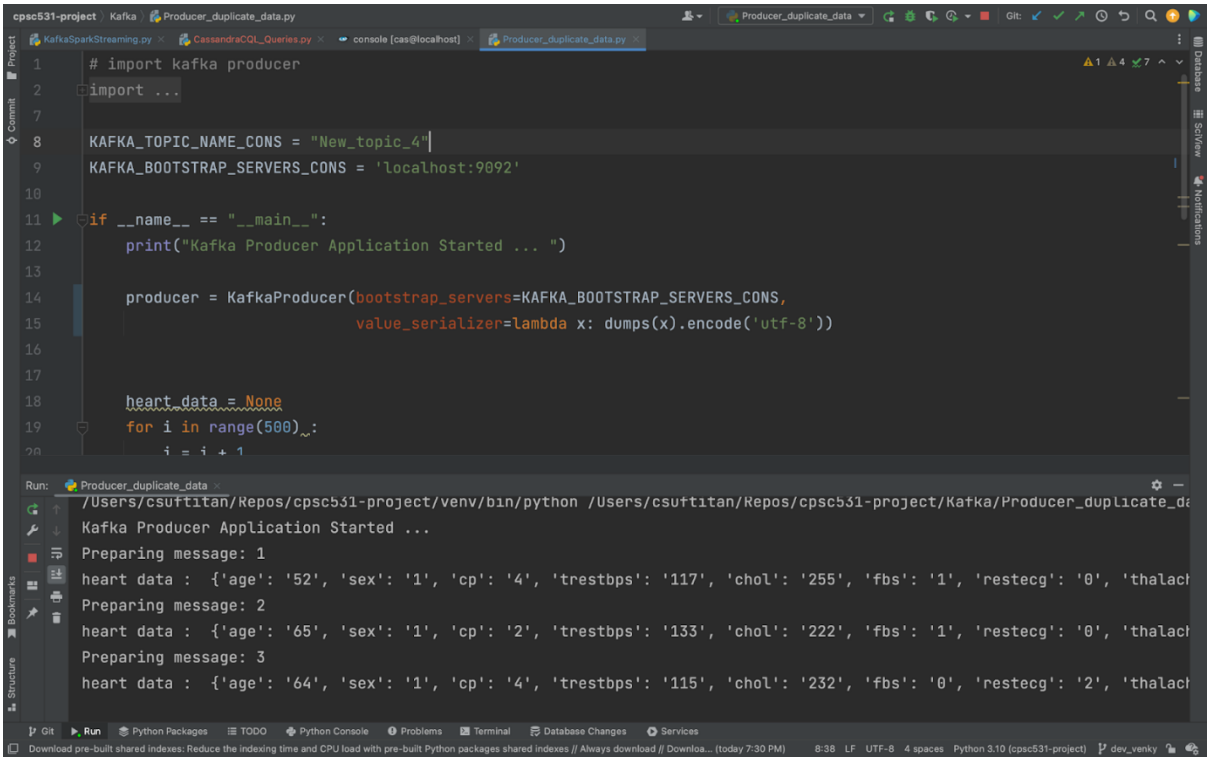
Path :

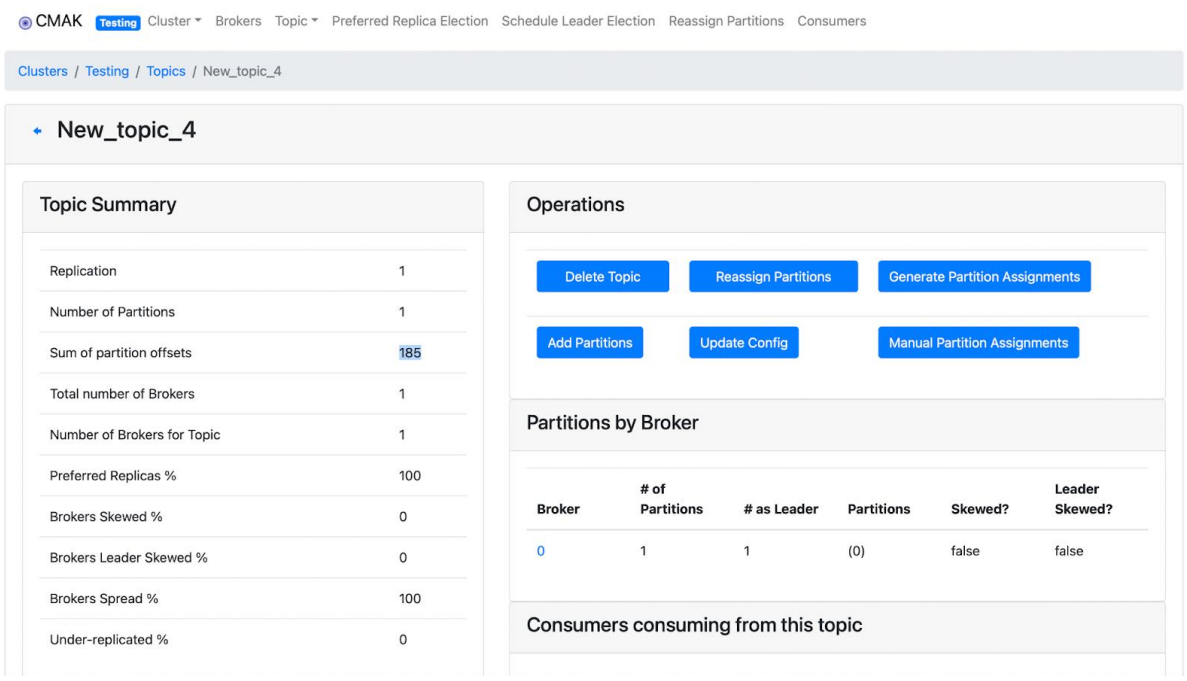
cmd : ./bin/spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.3.1,com.datastax.spark:spark-cassandra-connector_2.12:3.2.0,com.datastax.oss:java-driver-core:4.15.0 /Users/csuftitan/Repos/cpsc531-project/SparkStreaming/KafkaSparkStreaming.py

```
root
 |-- key: binary (nullable = true)
 |-- value: binary (nullable = true)
 |-- topic: string (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- timestampType: integer (nullable = true)

DataFrame[age: int, sex: int, cp: int, trestbps: int, chol: int, fbs: int, res
tecg: int, thalach: int, exang: int, oldpeak: float, slope: int, ca: int, thal
: int]
root
 |-- age: integer (nullable = true)
 |-- sex: integer (nullable = true)
 |-- cp: integer (nullable = true)
 |-- trestbps: integer (nullable = true)
 |-- chol: integer (nullable = true)
 |-- fbs: integer (nullable = true)
 |-- restecg: integer (nullable = true)
 |-- thalach: integer (nullable = true)
 |-- exang: integer (nullable = true)
 |-- oldpeak: float (nullable = true)
 |-- slope: integer (nullable = true)
 |-- ca: integer (nullable = true)
 |-- thal: integer (nullable = true)

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.util.SizeEstimator$ (fi
le:/Users/csuftitan/spark-3.3.1-bin-hadoop3/jars/spark-core_2.12-3.3.1.jar) to
 field java.math.BigInteger.mag
WARNING: Please consider reporting this to the maintainers of org.apache.spark
.util.SizeEstimator$
WARNING: Use --illegal-access=warn to enable warnings of further illegal refle
ctive access operations
WARNING: All illegal access operations will be denied in a future release
```

Second Run the Kafka Producer Script :

Path :





Checking the Results in Cassandra Data Base :

```
cqlsh:cas> select * from Heart_Prediction_data;

 id                                   | age | ca | chol | cp | exang | fbs | o
ldpeak | prediction | restecg | sex | slope | thal | thalach | trestbps
--------------------------------------+-----+----+------+----+-------+-----+--
-------+------------+---------+-----+-------+------+---------+----------
 bf91e9dc-52d4-40f9-a489-4521ebd38ac0 |  55 |  0 |  171 |  3 |     0 |   1 |
   3.5 |          0 |       1 |   1 |     2 |    2 |     135 |      117
 a98187a6-f95e-43af-9148-87189d584948 |  58 |  3 |  147 |  3 |     1 |   0 |
   0.2 |          0 |       1 |   0 |     0 |    3 |      98 |      115
 f43dca05-8af3-4f8f-95de-3235384b05b2 |  64 |  3 |  241 |  4 |     1 |   0 |
   1.6 |          0 |       1 |   1 |     0 |    1 |     137 |      117
 792e0382-5556-4372-ae8c-30b68a3b3ea8 |  55 |  1 |  152 |  4 |     0 |   0 |
   3.8 |          0 |       1 |   1 |     1 |    2 |     147 |       92
 760a4d5e-0b0b-42b9-884d-e75cd6ab4ffa |  79 |  0 |  133 |  1 |     0 |   0 |
   2.5 |          1 |       0 |   1 |     0 |    1 |     172 |      146
 f18c53b1-b28d-4917-8379-f273db10e31b |  54 |  0 |  245 |  1 |     1 |   0 |
   2.9 |          0 |       0 |   0 |     0 |    1 |      91 |      126
 105204f6-3027-4964-97c6-2d0986f1574a |  45 |  0 |  205 |  3 |     1 |   1 |
   2.2 |          0 |       1 |   1 |     2 |    3 |     110 |      126
 a6fe9ade-80ab-4157-b0bc-e4b7bd172799 |  45 |  1 |  134 |  3 |     0 |   0 |
   3.9 |          0 |       1 |   0 |     1 |    2 |      83 |      144
 0df0c464-cfc6-41bd-83f6-b14068cb5cfb |  53 |  3 |  170 |  2 |     1 |   0 |
   3.0 |          0 |       0 |   1 |     0 |    3 |     162 |      120
 765130ed-64f3-40f9-be8e-3235631b94c5 |  60 |  1 |  202 |  2 |     1 |   0 |
   2.1 |          0 |       2 |   0 |     2 |    3 |     123 |      137
 16dfccac-e547-42c0-a4c7-32d19051cefa |  72 |  3 |  230 |  4 |     1 |   1 |
   0.4 |          0 |       0 |   1 |     0 |    1 |     158 |      138
 e8e8b84c-eb1b-4588-b7fc-1d6db86dec46 |  65 |  2 |  194 |  3 |     0 |   1 |
   0.6 |          1 |       0 |   0 |     0 |    2 |     161 |      110
 eae57388-dfe2-4bca-838f-48594d2d2a19 |  59 |  2 |  231 |  1 |     1 |   1 |
   2.0 |          0 |       0 |   0 |     2 |    2 |     146 |       95
 8509e442-5733-457b-a328-5d8bfe2198f3 |  69 |  2 |  204 |  1 |     1 |   0 |
   2.5 |          1 |       0 |   0 |     2 |    1 |     151 |      149
 71a231bc-7746-4efe-92ac-928a4d7baa58 |  41 |  3 |  214 |  3 |     1 |   0 |
   2.2 |          0 |       0 |   0 |     2 |    1 |      87 |      120
 391858ad-9268-4538-8811-e53265b3f1b8 |  41 |  3 |  243 |  3 |     1 |   0 |
   3.8 |          0 |       1 |   0 |     0 |    3 |     108 |      127
 8e554491-a829-4251-be4b-c31607bb7bd5 |  64 |  3 |  244 |  3 |     1 |   1 |
   4.0 |          0 |       2 |   0 |     1 |    3 |      85 |      103
 585c7bbe-4a94-4d66-be93-6abb407f67bc |  47 |  1 |  176 |  3 |     0 |   0 |
   3.7 |          0 |       0 |   0 |     2 |    3 |     149 |      106
 6e46ecfc-91fd-42ee-9877-af17ee7fd17a |  73 |  2 |  197 |  4 |     0 |   1 |
   2.3 |          0 |       1 |   0 |     1 |    1 |     101 |      131
```

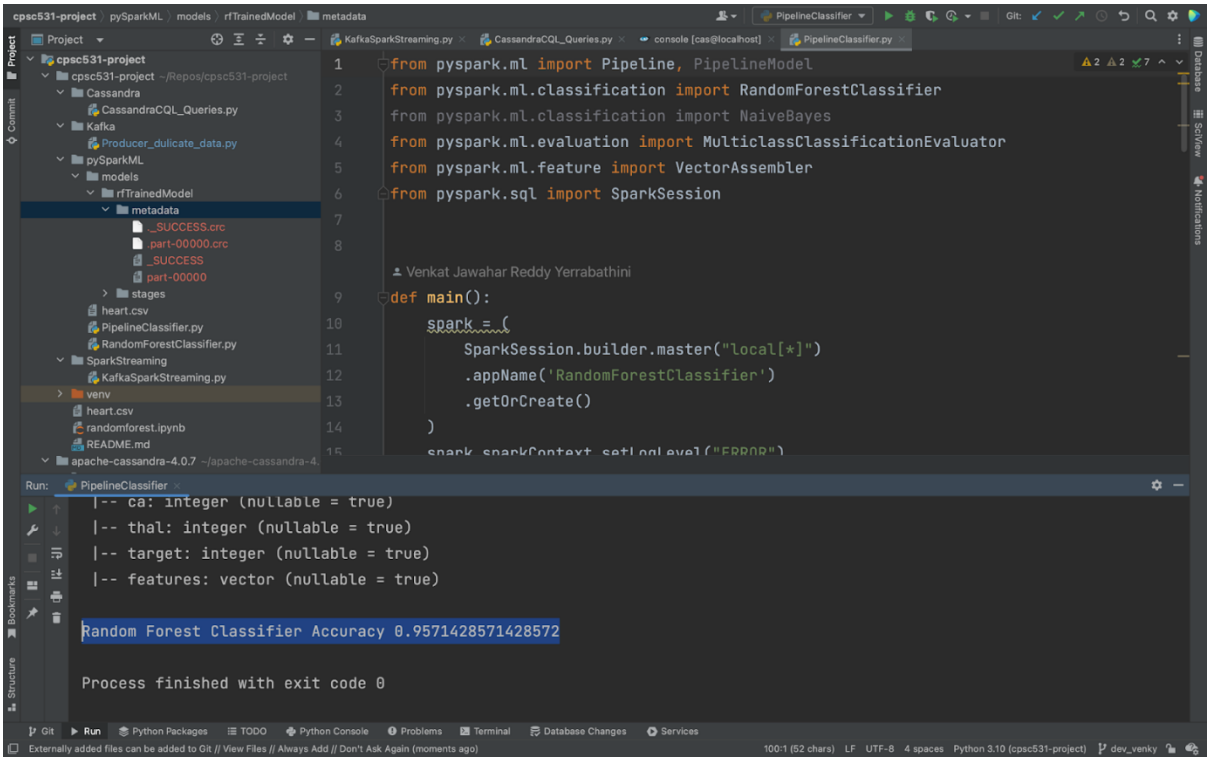
```
 0cbfc510-05db-47f0-8231-58ac4d6d42a4 |  51 |  3 |  212 |  2 |     0 |   1 |
   2.8 |          0 |       2 |   0 |     2 |    2 |     125 |      127
 00b5ded1-3384-46e4-adbd-a205240136b0 |  80 |  2 |  220 |  1 |     1 |   0 |
   3.3 |          0 |       0 |   1 |     0 |    1 |     116 |      130
 bd455e39-531e-487a-833b-830fbe898d95 |  40 |  1 |  205 |  2 |     1 |   0 |
   0.3 |          0 |       2 |   0 |     0 |    3 |      97 |       92

(136 rows)
cqlsh:cas>
```

```
 076e168b-f0cc-42f4-a1e5-d6460014ef80 |  50 |  0 |  157 |  4 |     0 |   0 |
   1.3 |           1 |        2 |   1 |     2 |     2 |      177 |        95
 fb0951be-73be-46ee-b1be-cff0b8cf61be |  58 |  0 |  174 |  3 |     1 |   1 |
   2.4 |           0 |        1 |   0 |     1 |     1 |      163 |       106
 0cbfc510-05db-47f0-8231-58ac4d6d42a4 |  51 |  3 |  212 |  2 |     0 |   1 |
   2.8 |           0 |        2 |   0 |     2 |     2 |      125 |       127
 00b5ded1-3384-46e4-adbd-a205240136b0 |  80 |  2 |  220 |  1 |     1 |   0 |
   3.3 |           0 |        0 |   1 |     0 |     1 |      116 |       130
 bd455e39-531e-487a-833b-830fbe898d95 |  40 |  1 |  205 |  2 |     1 |   0 |
   0.3 |           0 |        2 |   0 |     0 |     3 |       97 |        92

(138 rows)
cqlsh:cas>
```

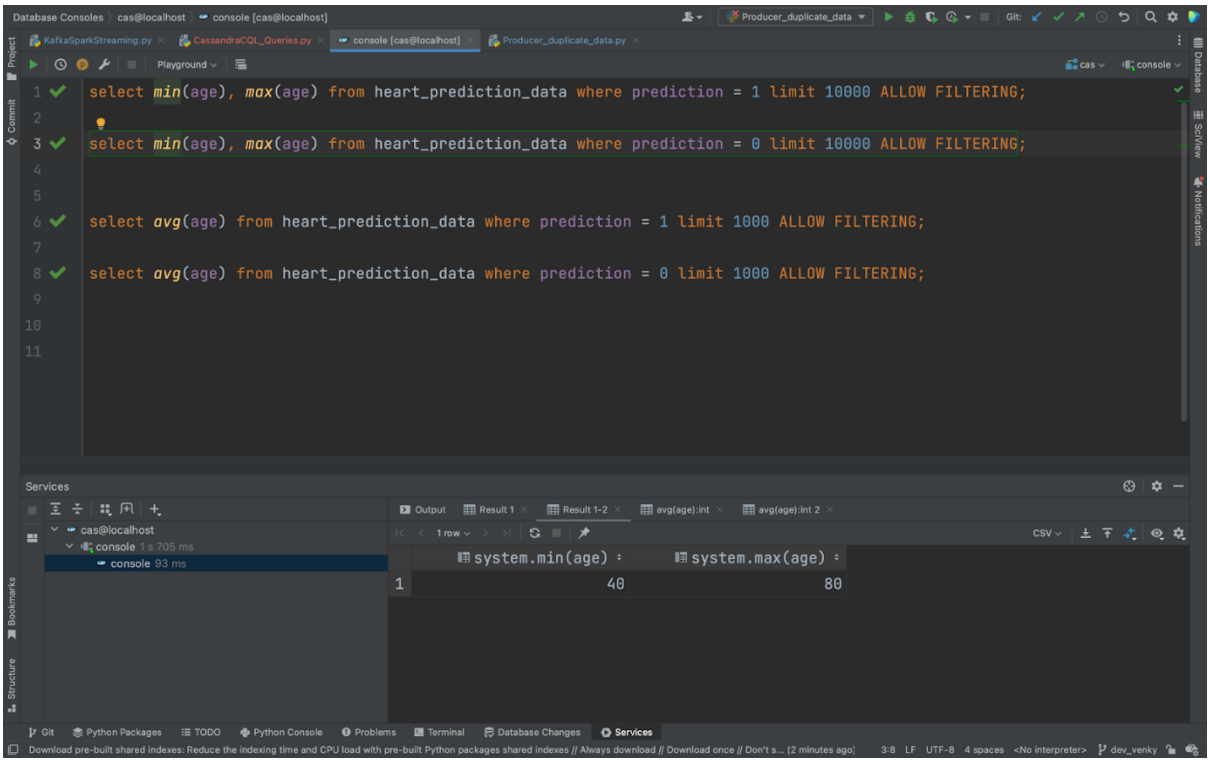
```
...he-cassandra-4.0.7 — cqlsh.py   ...31-project/Cassandra — -zsh   ~ — -zsh   +
  3.6 |           0 |        0 |   1 |     2 |     1 |      162 |       117
 5ad94690-8a14-4ceb-80d9-c7f9a94530bf |  43 |  1 |  260 |  2 |     1 |   0 |
  2.9 |           0 |        1 |   1 |     1 |     3 |      121 |       143
 8b7ffd16-3779-432e-a149-303f8ad605f6 |  67 |  1 |  231 |  4 |     1 |   1 |
  0.5 |           1 |        1 |   0 |     0 |     2 |      139 |        92
 5efd09a1-ebfd-42d9-b45a-483f06fa43d5 |  79 |  3 |  250 |  3 |     1 |   0 |
  1.8 |           0 |        0 |   0 |     1 |     2 |      104 |       132
 3cfad339-9e81-4ae4-baee-0d657a574c36 |  74 |  2 |  241 |  4 |     1 |   0 |
  1.3 |           0 |        0 |   0 |     0 |     3 |      166 |       111
 34c0bab7-c968-4978-aefb-00ecaa05f63e |  69 |  0 |  186 |  2 |     0 |   0 |
  3.5 |           1 |        1 |   1 |     0 |     2 |       80 |        95
 39dbf6d1-1f61-473a-af87-50300a22132b |  57 |  2 |  206 |  2 |     1 |   0 |
  2.8 |           0 |        0 |   1 |     2 |     2 |      116 |        93
 c0799275-6eec-449a-ad06-95d8e805b4be |  70 |  0 |  228 |  1 |     1 |   0 |
  1.2 |           1 |        2 |   0 |     1 |     1 |      129 |       147
 29fbec55-6950-4522-9423-e5ed477ad5f1 |  77 |  0 |  204 |  3 |     1 |   1 |
  3.7 |           1 |        0 |   0 |     2 |     1 |      117 |       100
 06b67461-fdf3-4cef-bd75-e6a11b0609ae |  78 |  1 |  146 |  3 |     0 |   1 |
  3.2 |           0 |        2 |   1 |     2 |     2 |      161 |       122
 62c42824-a838-4649-90ae-06cc57731a3d |  60 |  1 |  126 |  4 |     1 |   1 |
  0.9 |           1 |        2 |   1 |     2 |     3 |      131 |        92
 805c2bbe-effc-4f99-895c-2d85f244de46 |  40 |  0 |  248 |  1 |     1 |   1 |
  1.8 |           0 |        1 |   0 |     1 |     3 |       95 |       122
 322f3f3e-524d-4f26-8052-ea5df547030b |  76 |  1 |  249 |  3 |     1 |   1 |
  0.3 |           1 |        1 |   0 |     0 |     3 |      150 |       101
 0153be71-ca0e-45b2-9fe0-b3835ef9a8d2 |  42 |  0 |  161 |  4 |     0 |   1 |
  2.3 |           1 |        1 |   0 |     0 |     2 |      170 |       143
 34e111fc-ded6-43a6-826a-8cea5ad60727 |  78 |  0 |  132 |  4 |     1 |   0 |
  2.1 |           1 |        2 |   1 |     2 |     3 |       86 |       116
 9a375148-0bf2-46b7-a70c-666b71e3ddce |  57 |  0 |  209 |  1 |     1 |   1 |
  2.0 |           0 |        1 |   1 |     1 |     1 |      170 |       149
 35dece06-372b-42ec-828b-759d410c5db2 |  51 |  2 |  242 |  1 |     1 |   1 |
  3.6 |           0 |        1 |   0 |     2 |     2 |      124 |       109
 e68b3fa5-3443-4587-aaca-79ca68b74e0d |  60 |  1 |  141 |  2 |     0 |   1 |
  3.0 |           0 |        1 |   1 |     0 |     2 |      127 |        94
 076e168b-f0cc-42f4-a1e5-d6460014ef80 |  50 |  0 |  157 |  4 |     0 |   0 |
  1.3 |           1 |        2 |   1 |     2 |     2 |      177 |        95
 fb0951be-73be-46ee-b1be-cff0b8cf61be |  58 |  0 |  174 |  3 |     1 |   1 |
  2.4 |           0 |        1 |   0 |     1 |     1 |      163 |       106
 0cbfc510-05db-47f0-8231-58ac4d6d42a4 |  51 |  3 |  212 |  2 |     0 |   1 |
  2.8 |           0 |        2 |   0 |     2 |     2 |      125 |       127
 00b5ded1-3384-46e4-adbd-a205240136b0 |  80 |  2 |  220 |  1 |     1 |   0 |
  3.3 |           0 |        0 |   1 |     0 |     1 |      116 |       130
 bd455e39-531e-487a-833b-830fbe898d95 |  40 |  1 |  205 |  2 |     1 |   0 |
  0.3 |           0 |        2 |   0 |     0 |     3 |       97 |        92

(193 rows)
cqlsh:cas>
```

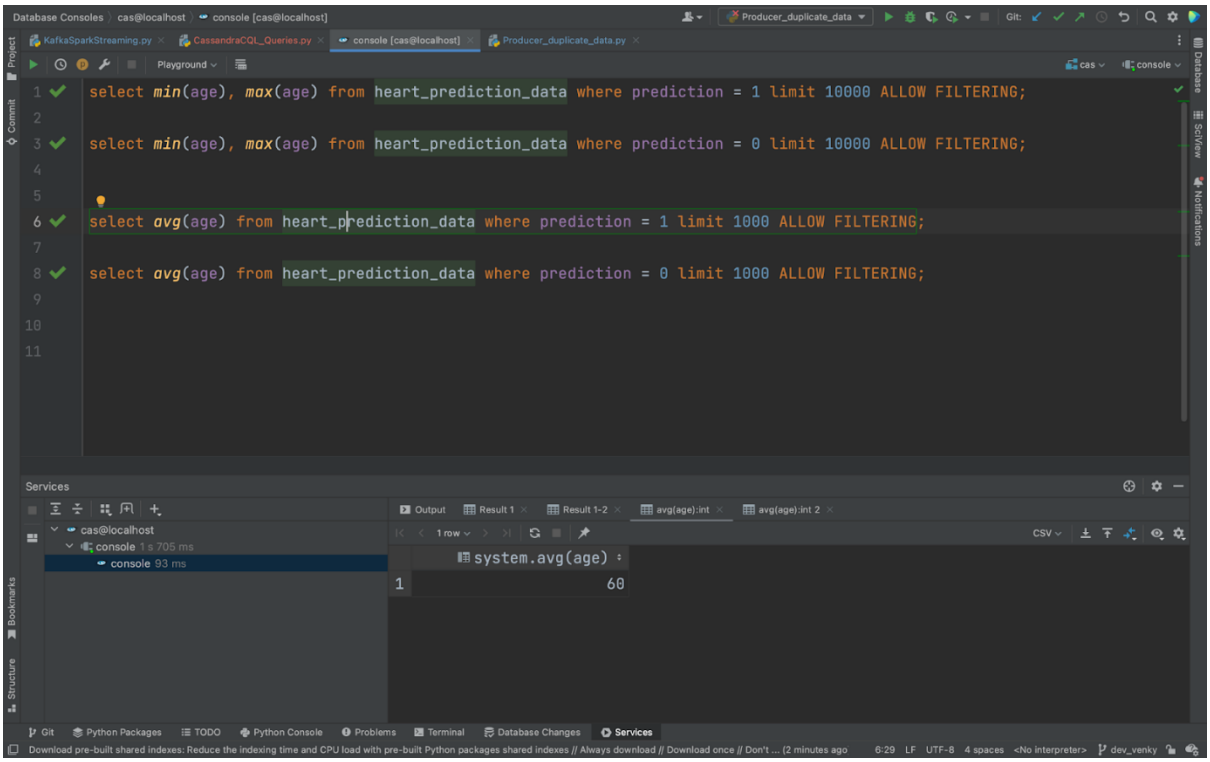
```
2:56727
22/12/10 23:09:27 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/12/10 23:09:27 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 10.67.89.42, 56727, None)
22/12/10 23:09:27 INFO BlockManagerMasterEndpoint: Registering block manager 10.67.89.42:56727 with 434.4 MiB RAM, BlockManagerId(driver, 10.67.89.42, 56727, None)
22/12/10 23:09:27 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 10.67.89.42, 56727, None)
22/12/10 23:09:27 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 10.67.89.42, 56727, None)
root
 |-- key: binary (nullable = true)
 |-- value: binary (nullable = true)
 |-- topic: string (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- timestampType: integer (nullable = true)

DataFrame[age: int, sex: int, cp: int, trestbps: int, chol: int, fbs: int, restecg: int, thalach: int, exang: int, oldpeak: float, slope: int, ca: int, thal: int]
root
 |-- age: integer (nullable = true)
 |-- sex: integer (nullable = true)
 |-- cp: integer (nullable = true)
 |-- trestbps: integer (nullable = true)
 |-- chol: integer (nullable = true)
 |-- fbs: integer (nullable = true)
 |-- restecg: integer (nullable = true)
 |-- thalach: integer (nullable = true)
 |-- exang: integer (nullable = true)
 |-- oldpeak: float (nullable = true)
 |-- slope: integer (nullable = true)
 |-- ca: integer (nullable = true)
 |-- thal: integer (nullable = true)

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.util.SizeEstimator$ (file:/Users/csuftitan/spark-3.3.1-bin-hadoop3/jars/spark-core_2.12-3.3.1.jar) to field java.math.BigInteger.mag
WARNING: Please consider reporting this to the maintainers of org.apache.spark.util.SizeEstimator$
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
```

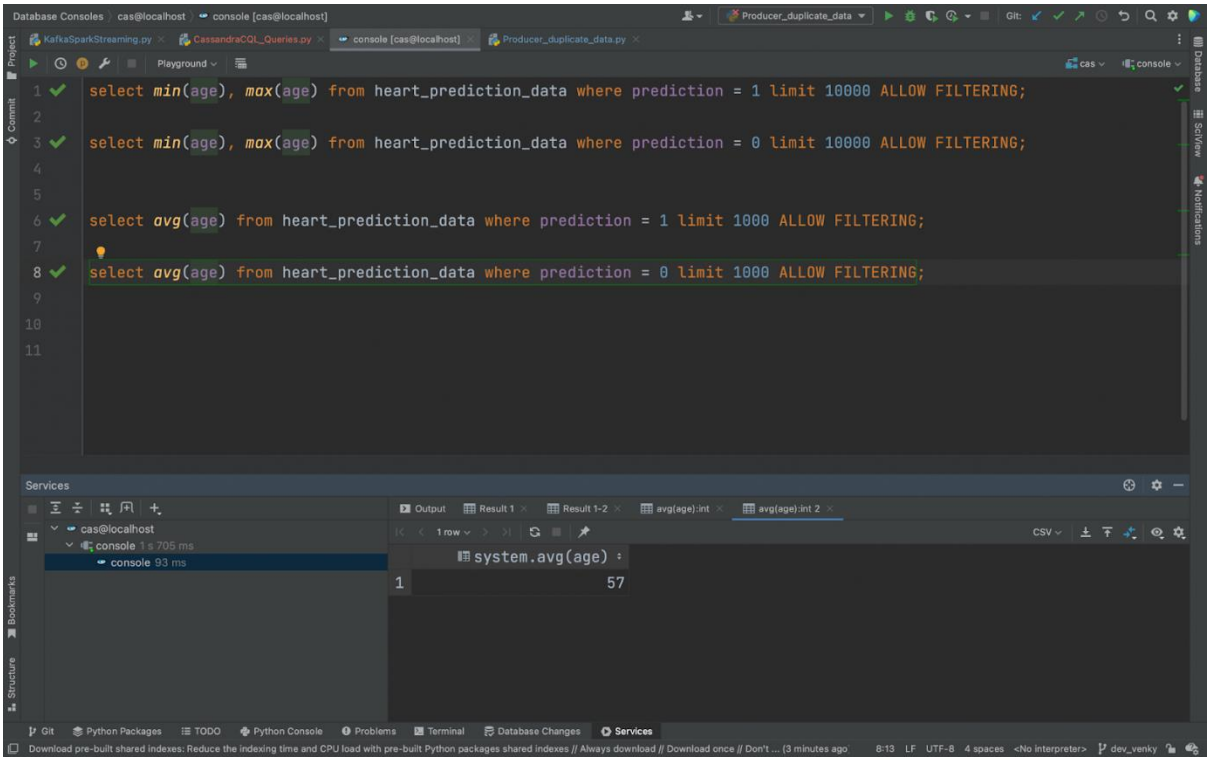
Training the random forest classifier results

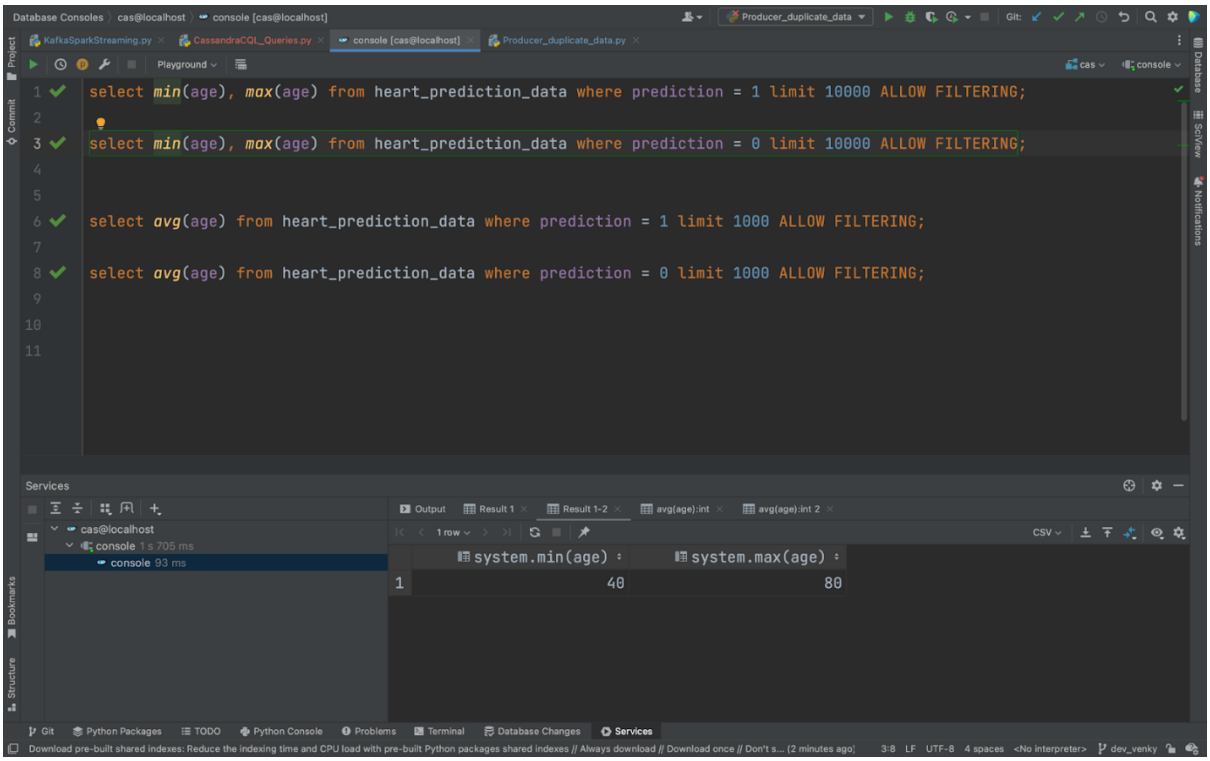


Querying the results in Cassandra




Database Consoles
cas@localhost
console [cas@localhost]
Producer_duplicate_data
KafkaSparkStreaming.py
CassandraCQL_Queries.py
console [cas@localhost]
Producer_duplicate_data.py
Playground
cas
console
select min(age), max(age) from heart_prediction_data where prediction = 1 limit 10000 ALLOW FILTERING;
select min(age), max(age) from heart_prediction_data where prediction = 0 limit 10000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 1 limit 1000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 0 limit 1000 ALLOW FILTERING;
Services
cas@localhost
console 1 s 705 ms
console 93 ms
Output
Result 1
Result 1-2
avg(age):int
avg(age):int 2
1 row
CSV
system.avg(age)
1
60
Git
Python Packages
TODO
Python Console
Problems
Terminal
Database Changes
Services
Download pre-built shared indexes: Reduce the indexing time and CPU load with pre-built Python packages shared indexes // Always download // Download once // Don't ... (2 minutes ago
6:29
LF
UTF-8
4 spaces
<No interpreter>
dev_venky


Database Consoles
cas@localhost
console [cas@localhost]
Producer_duplicate_data
KafkaSparkStreaming.py
CassandraCQL_Queries.py
console [cas@localhost]
Producer_duplicate_data.py
Playground
cas
console
select min(age), max(age) from heart_prediction_data where prediction = 1 limit 10000 ALLOW FILTERING;
select min(age), max(age) from heart_prediction_data where prediction = 0 limit 10000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 1 limit 1000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 0 limit 1000 ALLOW FILTERING;
Services
cas@localhost
console 1 s 705 ms
console 93 ms
Output
Result 1
Result 1-2
avg(age):int
avg(age):int 2
1 row
CSV
system.avg(age)
1
57
Git
Python Packages
TODO
Python Console
Problems
Terminal
Database Changes
Services
Download pre-built shared indexes: Reduce the indexing time and CPU load with pre-built Python packages shared indexes // Always download // Download once // Don't ... (3 minutes ago
8:13
LF
UTF-8
4 spaces
<No interpreter>
dev_venky


Database Consoles
cas@localhost
console [cas@localhost]
Producer_duplicate_data
Git:
KafkaSparkStreaming.py
CassandraCQL_Queries.py
console [cas@localhost]
Producer_duplicate_data.py
Playground
cas
console
select min(age), max(age) from heart_prediction_data where prediction = 1 limit 10000 ALLOW FILTERING;
select min(age), max(age) from heart_prediction_data where prediction = 0 limit 10000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 1 limit 1000 ALLOW FILTERING;
select avg(age) from heart_prediction_data where prediction = 0 limit 1000 ALLOW FILTERING;
Services
cas@localhost
console 1 s 705 ms
console 93 ms
Output
Result 1
Result 1-2
avg(age):int
avg(age):int 2
1 row
CSV
system.min(age)
system.max(age)
1
40
80
Git
Python Packages
TODO
Python Console
Problems
Terminal
Database Changes
Services
Download pre-built shared indexes: Reduce the indexing time and CPU load with pre-built Python packages shared indexes // Always download // Download once // Don't s... (2 minutes ago)
3:8
LF
UTF-8
4 spaces
<No interpreter>
dev_venky