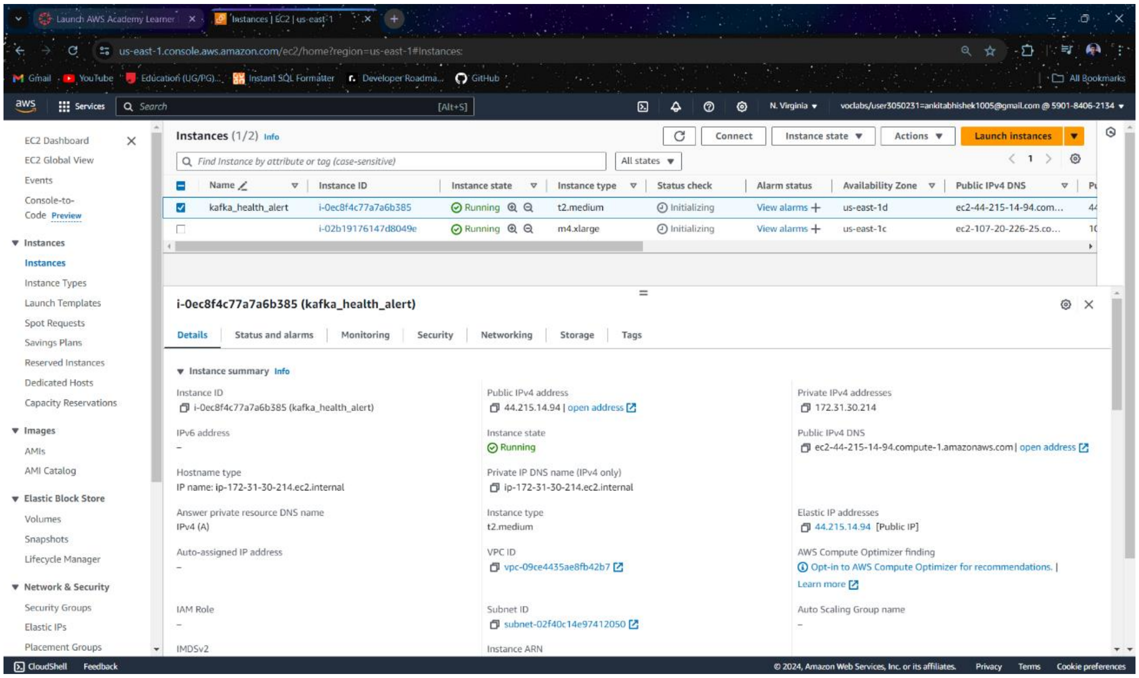
CREATION OF KAFKA CLUSTER AND EMR CLUSTER

1. Create a kafka cluster with the help of ec2 instance. (Referred with PDF attached in the Apache Kafka modules). Make the required configuration needed to run kafka instance. Kafka is pre-installed on below mentioned ec2 cluster with the selection of **ami-06c41d8b5a6ddd3c2** while creating **Amazon Machine Image** as pdf within modules.



us-east-1.console.aws.amazon.com/ec2/home?region=us-east-1#Instances:

voclabs/user3050231=ankitabhishek1005@gmail.com @ 5901-8406-2134

Instances (1/2)

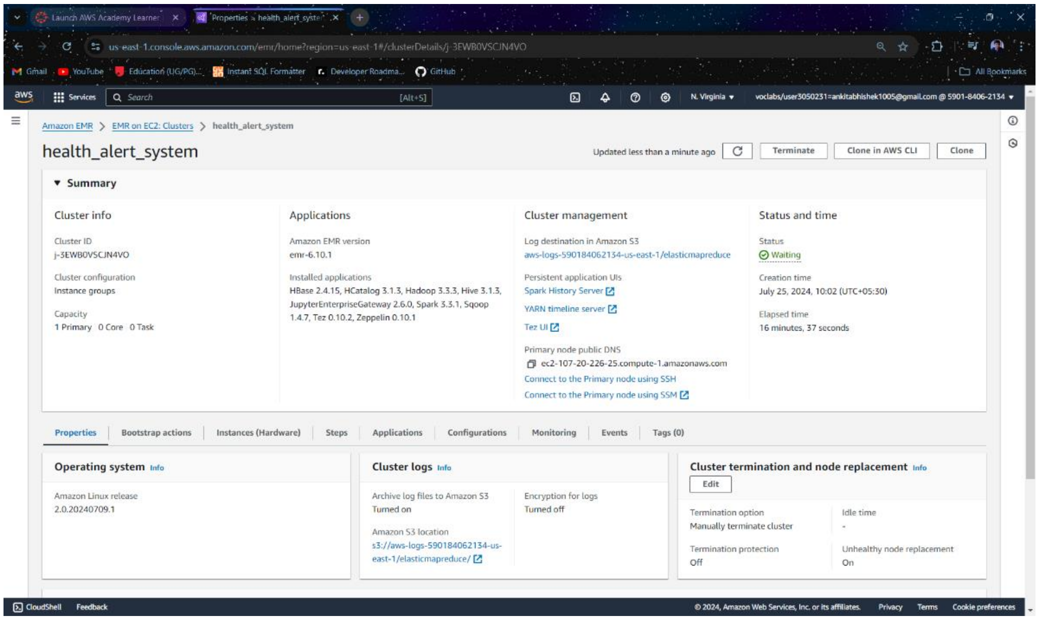
☐	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
☑	kafka_health_alert	i-0ec8f4c77a7a6b385	Running	t2.medium	Initializing	View alarms +	us-east-1d	ec2-44-215-14-94.com...
☐		i-02b19176147d8049e	Running	m4.xlarge	Initializing	View alarms +	us-east-1c	ec2-107-20-226-25.co...

i-0ec8f4c77a7a6b385 (kafka_health_alert)

Instance summary

- Instance ID: i-0ec8f4c77a7a6b385 (kafka_health_alert)
- Public IPv4 address: 44.215.14.94 | open address
- Private IPv4 addresses: 172.31.30.214
- IPv6 address: –
- Instance state: Running
- Public IPv4 DNS: ec2-44-215-14-94.compute-1.amazonaws.com | open address
- Hostname type: IP name: ip-172-31-30-214.ec2.internal
- Private IP DNS name (IPv4 only): ip-172-31-30-214.ec2.internal
- Answer private resource DNS name: IPv4 (A)
- Instance type: t2.medium
- Elastic IP addresses: 44.215.14.94 [Public IP]
- Auto-assigned IP address: –
- VPC ID: vpc-09ce4435ae8fb42b7
- AWS Compute Optimizer finding: Opt-in to AWS Compute Optimizer for recommendations. | Learn more
- IAM Role: –
- Subnet ID: subnet-02f40c14e97412050
- Auto Scaling Group name: –
- IMDSv2
- Instance ARN

2. Create an EMR instance with required below mentioned libraries (Referred with PDF attached in the modules).
   Spark 3.3.1, Sqoop 1.4.7, HBase 2.4.15, HCatalog 3.1.3, Hadoop 3.3.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Zeppelin 0.10.1



us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusterDetails/j-3EWB0VSCJN4VO

voclabs/user3050231=ankitabhishek1005@gmail.com @ 5901-8406-2134

Amazon EMR > EMR on EC2: Clusters > health_alert_system

health_alert_system

Updated less than a minute ago

Summary

Cluster info

- Cluster ID: j-3EWB0VSCJN4VO
- Cluster configuration: Instance groups
- Capacity: 1 Primary 0 Core 0 Task

Applications

- Amazon EMR version: emr-6.10.1
- Installed applications: HBase 2.4.15, HCatalog 3.1.3, Hadoop 3.3.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Spark 3.3.1, Sqoop 1.4.7, Tez 0.10.2, Zeppelin 0.10.1

Cluster management

- Log destination in Amazon S3: aws-logs-590184062134-us-east-1/elasticmapreduce
- Persistent application UIs: Spark History Server, YARN timeline server, Tez UI
- Primary node public DNS: ec2-107-20-226-25.compute-1.amazonaws.com
- Connect to the Primary node using SSH
- Connect to the Primary node using SSM

Status and time

- Status: Waiting
- Creation time: July 25, 2024, 10:02 (UTC+05:30)
- Elapsed time: 16 minutes, 37 seconds

Operating system

- Amazon Linux release: 2.0.20240709.1

Cluster logs

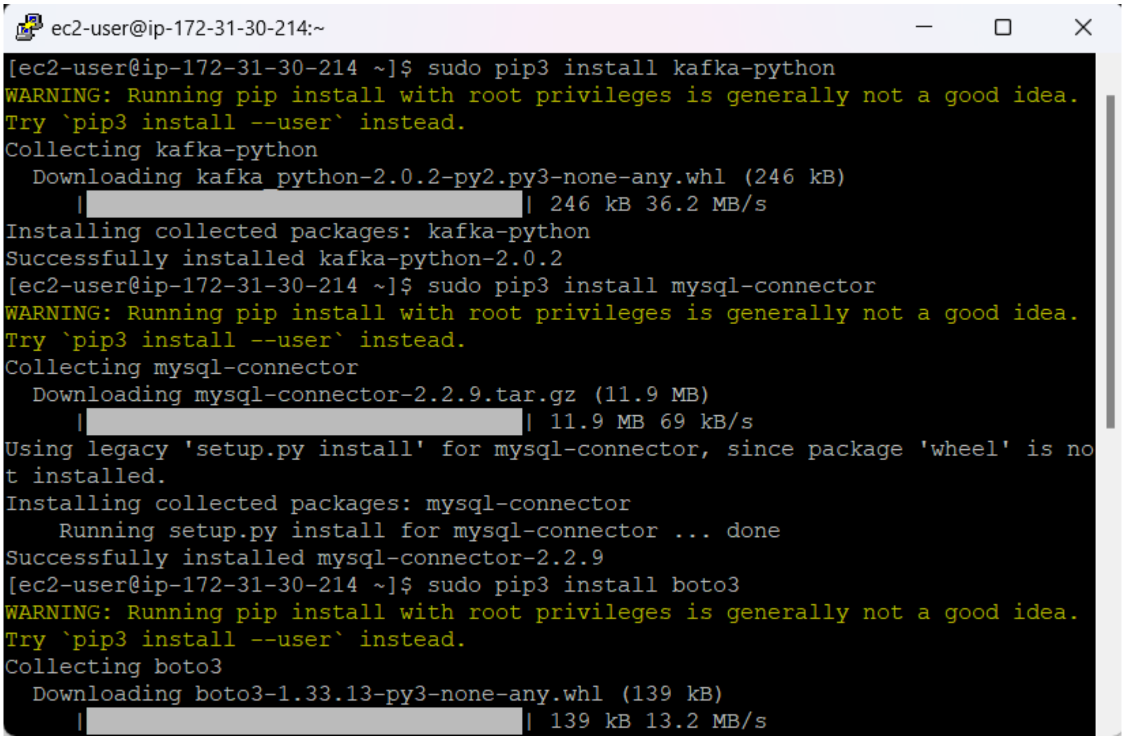
- Archive log files to Amazon S3: Turned on
- Encryption for logs: Turned off
- Amazon S3 location: s3://aws-logs-590184062134-us-east-1/elasticmapreduce/

Cluster termination and node replacement

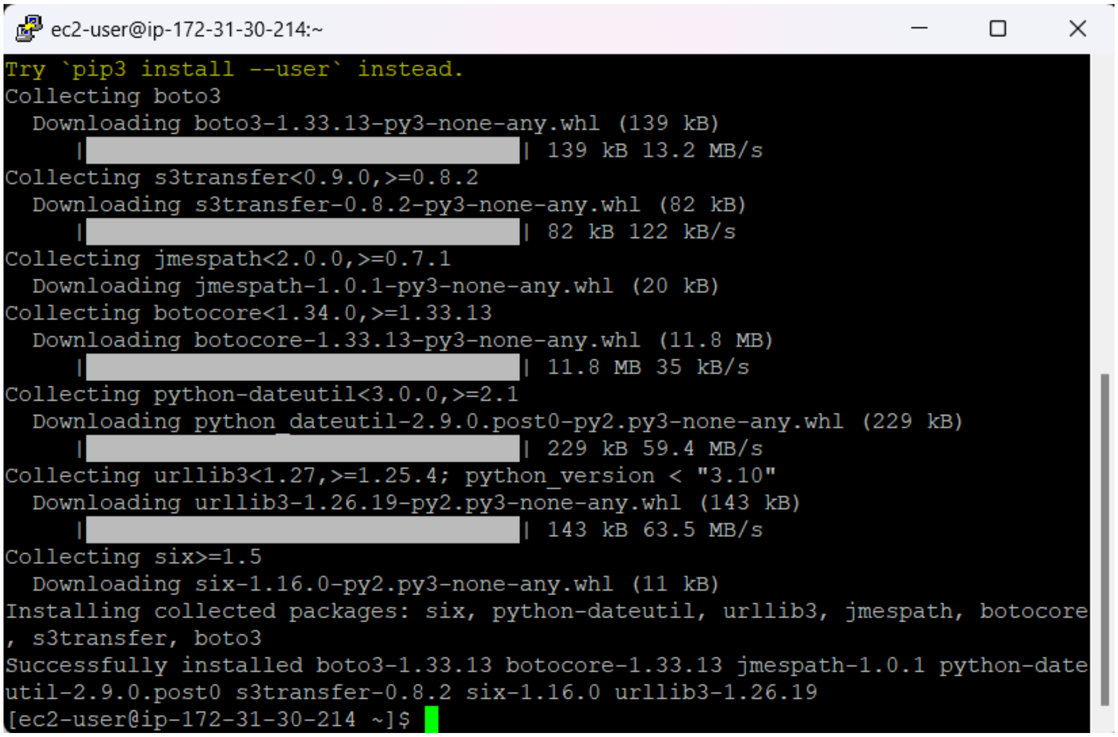
- Termination option: Manually terminate cluster
- Idle time: -
- Termination protection: Off
- Unhealthy node replacement: On

INSTALLING REQUIRED PACKAGES ON KAFKA CLUSTER

Sudo pip3 install kafka-python
Sudo pip3 install mysql-connector
Sudo pip3 install boto3



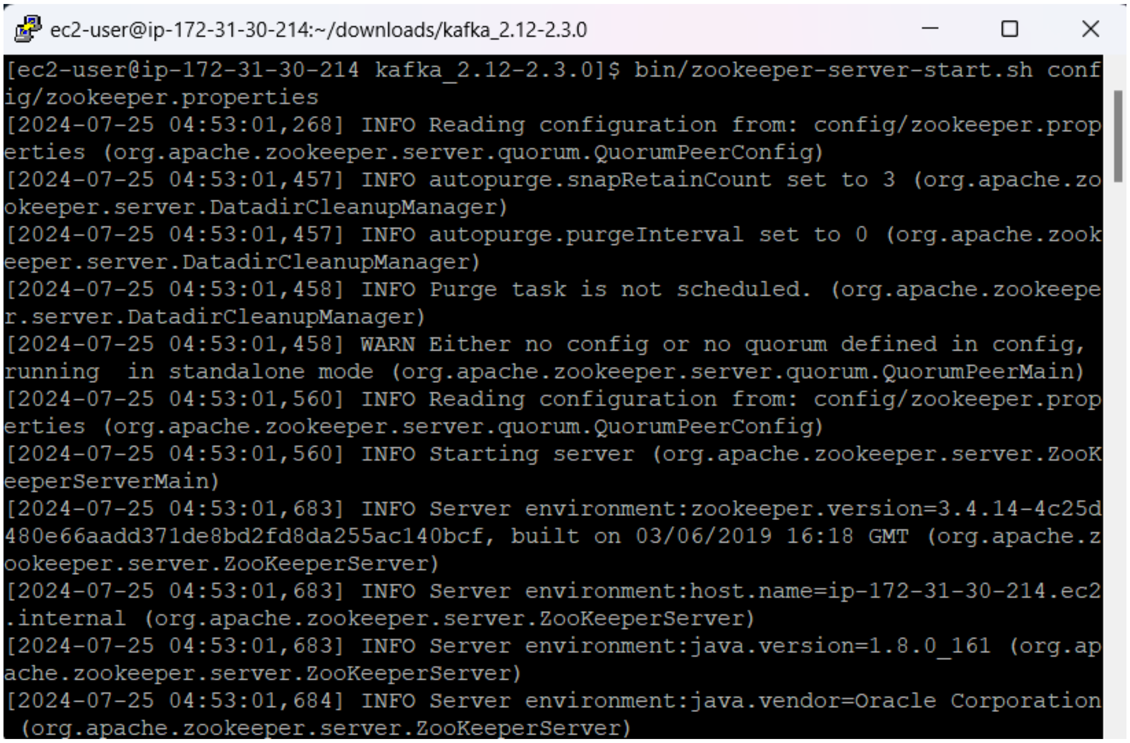
```
[ec2-user@ip-172-31-30-214 ~]$ sudo pip3 install kafka-python
WARNING: Running pip install with root privileges is generally not a good idea.
Try `pip3 install --user` instead.
Collecting kafka-python
  Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
     |████████████████████████████████| 246 kB 36.2 MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
[ec2-user@ip-172-31-30-214 ~]$ sudo pip3 install mysql-connector
WARNING: Running pip install with root privileges is generally not a good idea.
Try `pip3 install --user` instead.
Collecting mysql-connector
  Downloading mysql-connector-2.2.9.tar.gz (11.9 MB)
     |████████████████████████████████| 11.9 MB 69 kB/s
Using legacy 'setup.py install' for mysql-connector, since package 'wheel' is no
t installed.
Installing collected packages: mysql-connector
    Running setup.py install for mysql-connector ... done
Successfully installed mysql-connector-2.2.9
[ec2-user@ip-172-31-30-214 ~]$ sudo pip3 install boto3
WARNING: Running pip install with root privileges is generally not a good idea.
Try `pip3 install --user` instead.
Collecting boto3
  Downloading boto3-1.33.13-py3-none-any.whl (139 kB)
     |████████████████████████████████| 139 kB 13.2 MB/s
```



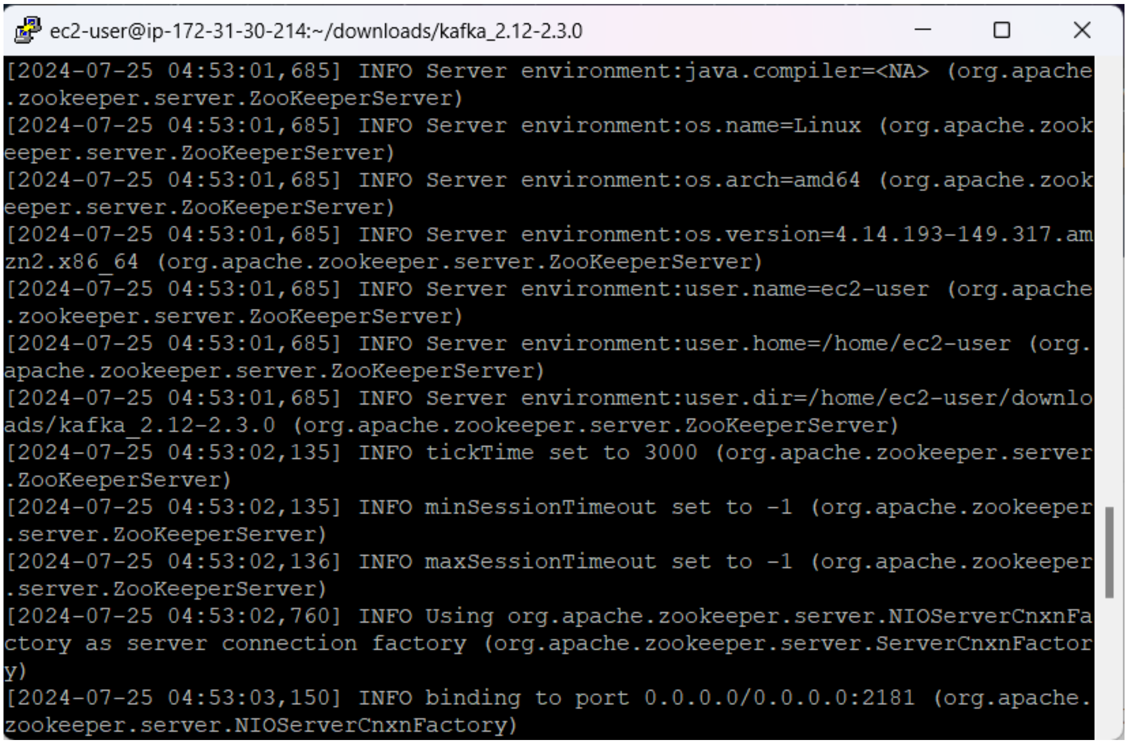
```
Try `pip3 install --user` instead.
Collecting boto3
  Downloading boto3-1.33.13-py3-none-any.whl (139 kB)
     |████████████████████████████████| 139 kB 13.2 MB/s
Collecting s3transfer<0.9.0,>=0.8.2
  Downloading s3transfer-0.8.2-py3-none-any.whl (82 kB)
     |████████████████████████████████| 82 kB 122 kB/s
Collecting jmespath<2.0.0,>=0.7.1
  Downloading jmespath-1.0.1-py3-none-any.whl (20 kB)
Collecting botocore<1.34.0,>=1.33.13
  Downloading botocore-1.33.13-py3-none-any.whl (11.8 MB)
     |████████████████████████████████| 11.8 MB 35 kB/s
Collecting python-dateutil<3.0.0,>=2.1
  Downloading python_dateutil-2.9.0.post0-py2.py3-none-any.whl (229 kB)
     |████████████████████████████████| 229 kB 59.4 MB/s
Collecting urllib3<1.27,>=1.25.4; python_version < "3.10"
  Downloading urllib3-1.26.19-py2.py3-none-any.whl (143 kB)
     |████████████████████████████████| 143 kB 63.5 MB/s
Collecting six>=1.5
  Downloading six-1.16.0-py2.py3-none-any.whl (11 kB)
Installing collected packages: six, python-dateutil, urllib3, jmespath, botocore
, s3transfer, boto3
Successfully installed boto3-1.33.13 botocore-1.33.13 jmespath-1.0.1 python-date
util-2.9.0.post0 s3transfer-0.8.2 six-1.16.0 urllib3-1.26.19
[ec2-user@ip-172-31-30-214 ~]$ █
```

STATEMENT FOR STARTING KAFKA SERVER

1. **STARTING ZOOKEEPER SERVER**:
   Inside cd downloads/kafka_2.12-2.3.0 run
   bin/zookeeper-server-start.sh config/zookeeper.properties



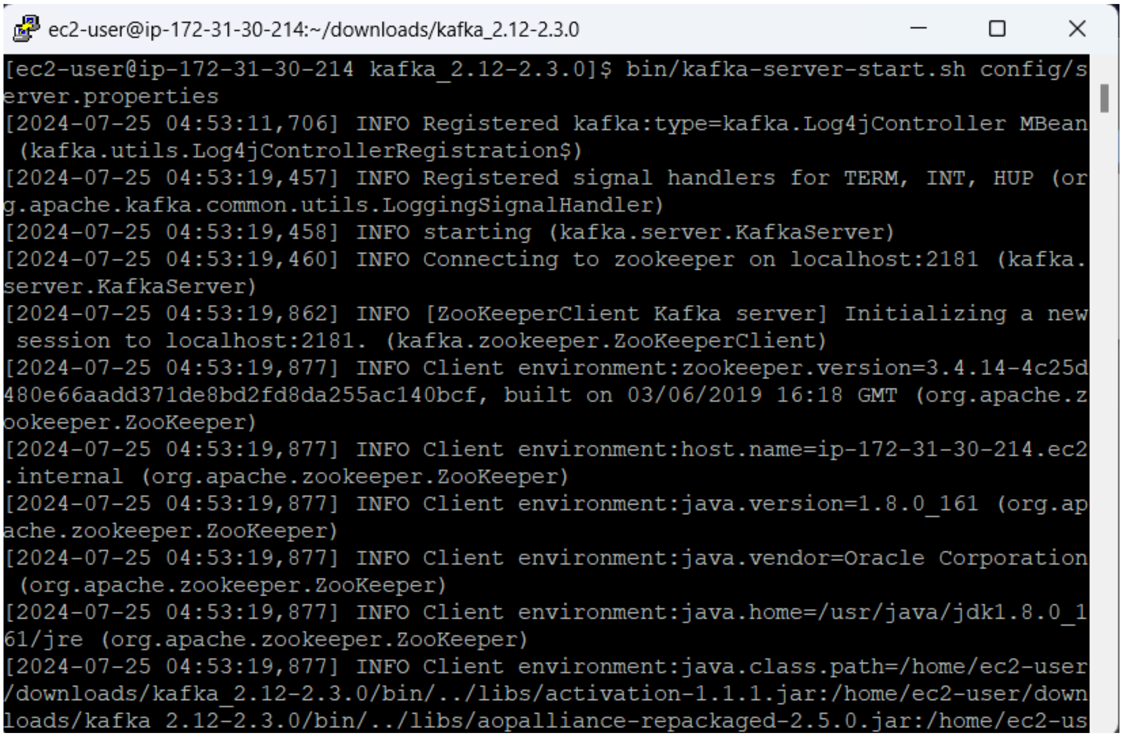
```
[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$ bin/zookeeper-server-start.sh conf
ig/zookeeper.properties
[2024-07-25 04:53:01,268] INFO Reading configuration from: config/zookeeper.prop
erties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-07-25 04:53:01,457] INFO autopurge.snapRetainCount set to 3 (org.apache.zo
okeeper.server.DatadirCleanupManager)
[2024-07-25 04:53:01,457] INFO autopurge.purgeInterval set to 0 (org.apache.zook
eeper.server.DatadirCleanupManager)
[2024-07-25 04:53:01,458] INFO Purge task is not scheduled. (org.apache.zookeepe
r.server.DatadirCleanupManager)
[2024-07-25 04:53:01,458] WARN Either no config or no quorum defined in config,
running  in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2024-07-25 04:53:01,560] INFO Reading configuration from: config/zookeeper.prop
erties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-07-25 04:53:01,560] INFO Starting server (org.apache.zookeeper.server.ZooK
eeperServerMain)
[2024-07-25 04:53:01,683] INFO Server environment:zookeeper.version=3.4.14-4c25d
480e66aadd371de8bd2fd8da255ac140bcf, built on 03/06/2019 16:18 GMT (org.apache.z
ookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,683] INFO Server environment:host.name=ip-172-31-30-214.ec2
.internal (org.apache.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,683] INFO Server environment:java.version=1.8.0_161 (org.ap
ache.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,684] INFO Server environment:java.vendor=Oracle Corporation
 (org.apache.zookeeper.server.ZooKeeperServer)
```



```
[2024-07-25 04:53:01,685] INFO Server environment:java.compiler=<NA> (org.apache
.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:os.name=Linux (org.apache.zook
eeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:os.arch=amd64 (org.apache.zook
eeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:os.version=4.14.193-149.317.am
zn2.x86_64 (org.apache.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:user.name=ec2-user (org.apache
.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:user.home=/home/ec2-user (org.
apache.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:01,685] INFO Server environment:user.dir=/home/ec2-user/downlo
ads/kafka_2.12-2.3.0 (org.apache.zookeeper.server.ZooKeeperServer)
[2024-07-25 04:53:02,135] INFO tickTime set to 3000 (org.apache.zookeeper.server
.ZooKeeperServer)
[2024-07-25 04:53:02,135] INFO minSessionTimeout set to -1 (org.apache.zookeeper
.server.ZooKeeperServer)
[2024-07-25 04:53:02,136] INFO maxSessionTimeout set to -1 (org.apache.zookeeper
.server.ZooKeeperServer)
[2024-07-25 04:53:02,760] INFO Using org.apache.zookeeper.server.NIOServerCnxnFa
ctory as server connection factory (org.apache.zookeeper.server.ServerCnxnFactor
y)
[2024-07-25 04:53:03,150] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.
zookeeper.server.NIOServerCnxnFactory)
```

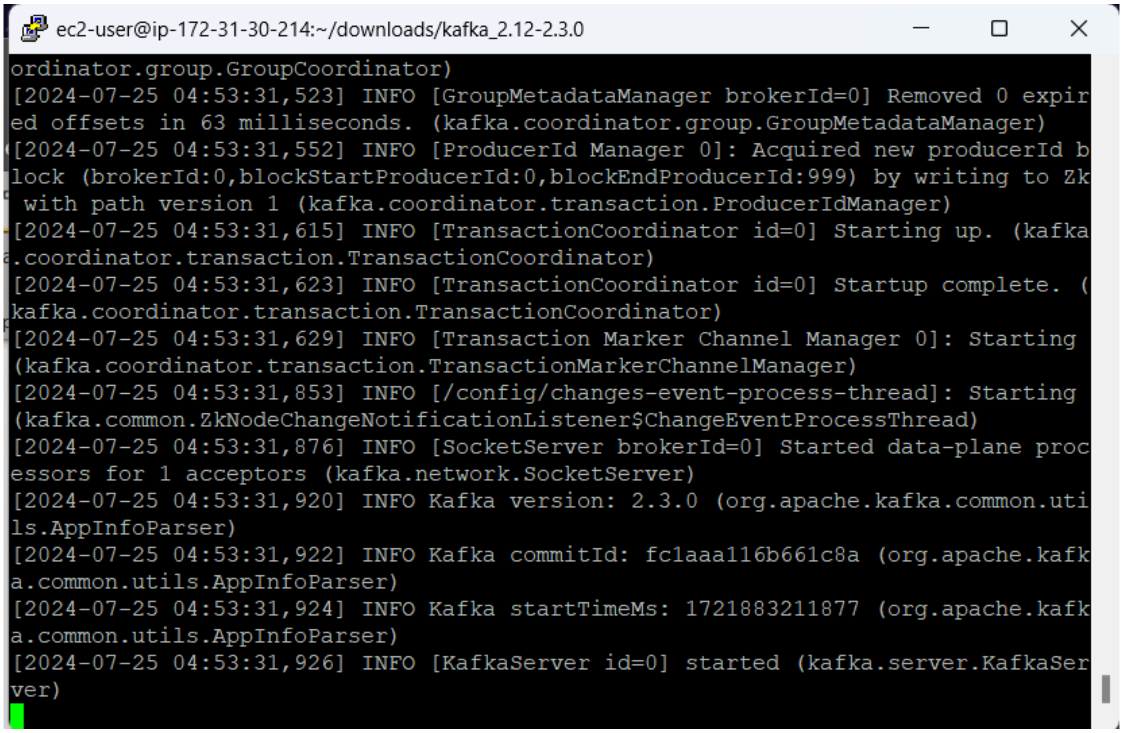
2. **STARTING KAFKA SERVER:**

   Into another putty Session of kafka cluster inside cd downloads/kafka_2.12-2.3.0 run **bin/kafka-server-start.sh config/server.properties**



ec2-user@ip-172-31-30-214:~/downloads/kafka_2.12-2.3.0

```
[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$ bin/kafka-server-start.sh config/s
erver.properties
[2024-07-25 04:53:11,706] INFO Registered kafka:type=kafka.Log4jController MBean
 (kafka.utils.Log4jControllerRegistration$)
[2024-07-25 04:53:19,457] INFO Registered signal handlers for TERM, INT, HUP (or
g.apache.kafka.common.utils.LoggingSignalHandler)
[2024-07-25 04:53:19,458] INFO starting (kafka.server.KafkaServer)
[2024-07-25 04:53:19,460] INFO Connecting to zookeeper on localhost:2181 (kafka.
server.KafkaServer)
[2024-07-25 04:53:19,862] INFO [ZooKeeperClient Kafka server] Initializing a new
 session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2024-07-25 04:53:19,877] INFO Client environment:zookeeper.version=3.4.14-4c25d
480e66aadd371de8bd2fd8da255ac140bcf, built on 03/06/2019 16:18 GMT (org.apache.z
ookeeper.ZooKeeper)
[2024-07-25 04:53:19,877] INFO Client environment:host.name=ip-172-31-30-214.ec2
.internal (org.apache.zookeeper.ZooKeeper)
[2024-07-25 04:53:19,877] INFO Client environment:java.version=1.8.0_161 (org.ap
ache.zookeeper.ZooKeeper)
[2024-07-25 04:53:19,877] INFO Client environment:java.vendor=Oracle Corporation
 (org.apache.zookeeper.ZooKeeper)
[2024-07-25 04:53:19,877] INFO Client environment:java.home=/usr/java/jdk1.8.0_1
61/jre (org.apache.zookeeper.ZooKeeper)
[2024-07-25 04:53:19,877] INFO Client environment:java.class.path=/home/ec2-user
/downloads/kafka_2.12-2.3.0/bin/../libs/activation-1.1.1.jar:/home/ec2-user/down
loads/kafka_2.12-2.3.0/bin/../libs/aopalliance-repackaged-2.5.0.jar:/home/ec2-us
```



ec2-user@ip-172-31-30-214:~/downloads/kafka_2.12-2.3.0

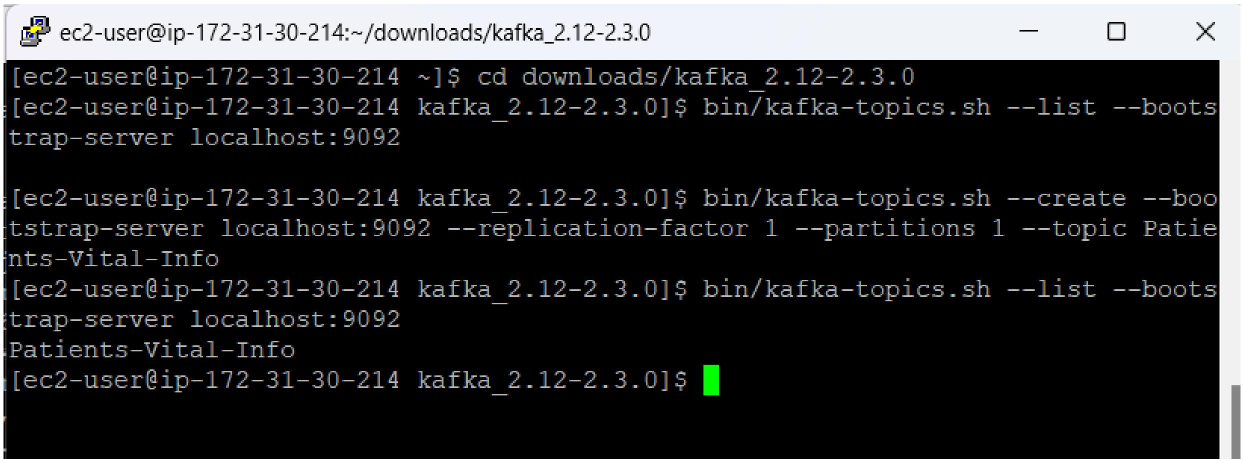
```
ordinator.group.GroupCoordinator)
[2024-07-25 04:53:31,523] INFO [GroupMetadataManager brokerId=0] Removed 0 expir
ed offsets in 63 milliseconds. (kafka.coordinator.group.GroupMetadataManager)
[2024-07-25 04:53:31,552] INFO [ProducerId Manager 0]: Acquired new producerId b
lock (brokerId:0,blockStartProducerId:0,blockEndProducerId:999) by writing to Zk
 with path version 1 (kafka.coordinator.transaction.ProducerIdManager)
[2024-07-25 04:53:31,615] INFO [TransactionCoordinator id=0] Starting up. (kafka
.coordinator.transaction.TransactionCoordinator)
[2024-07-25 04:53:31,623] INFO [TransactionCoordinator id=0] Startup complete. (
kafka.coordinator.transaction.TransactionCoordinator)
[2024-07-25 04:53:31,629] INFO [Transaction Marker Channel Manager 0]: Starting
(kafka.coordinator.transaction.TransactionMarkerChannelManager)
[2024-07-25 04:53:31,853] INFO [/config/changes-event-process-thread]: Starting
(kafka.common.ZkNodeChangeNotificationListener$ChangeEventProcessThread)
[2024-07-25 04:53:31,876] INFO [SocketServer brokerId=0] Started data-plane proc
essors for 1 acceptors (kafka.network.SocketServer)
[2024-07-25 04:53:31,920] INFO Kafka version: 2.3.0 (org.apache.kafka.common.uti
ls.AppInfoParser)
[2024-07-25 04:53:31,922] INFO Kafka commitId: fc1aaa116b661c8a (org.apache.kafk
a.common.utils.AppInfoParser)
[2024-07-25 04:53:31,924] INFO Kafka startTimeMs: 1721883211877 (org.apache.kafk
a.common.utils.AppInfoParser)
[2024-07-25 04:53:31,926] INFO [KafkaServer id=0] started (kafka.server.KafkaSer
ver)
```

STATEMENT TO CREATE TOPICS

To create topic in kafka server, the command used is **bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic Patients-Vital-Info**

STATEMENT TO LIST TOPICS

To list the created topic inside cd downloads/kafka_2.12-2.3.0, the command used is **bin/kafka-topics.sh --list --bootstrap-server localhost:9092**



```
[ec2-user@ip-172-31-30-214 ~]$ cd downloads/kafka_2.12-2.3.0
[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --list --boots
trap-server localhost:9092

[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --create --boo
tstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic Patie
nts-Vital-Info
[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$ bin/kafka-topics.sh --list --boots
trap-server localhost:9092
Patients-Vital-Info
[ec2-user@ip-172-31-30-214 kafka_2.12-2.3.0]$
```

EXECUTING PRODUCER APPLICATION AND CONSUMER APPLICATION:

Producer application which is file named as **kafka_produce_patient_vitals.py** is built on the **python language** which will consume data residing on rds with below mentioned credentials:
Hostname = "upgraddetest.cyaielc9bmnf.us-east-1.rds.amazonaws.com"
username = "student"
password = "STUDENT123"
dbname = "testdatabase".

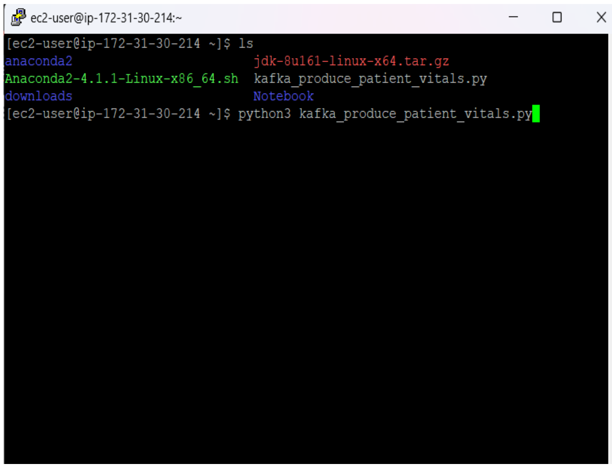
Consumer Application which is file named as **kafka_spark_patient_vitals.py** is built on the **Apache PySpark** language which will consume data being produced with the help of above mentioned producer application

NOTE: Run the producer application on ec2 Kafka cluster after starting the consumer application on EMR cluster created with Spark, Hive and another libraries

STATEMENT FOR EXECUTING PRODUCER APPLICATION AND CONSUMER APPLICATION

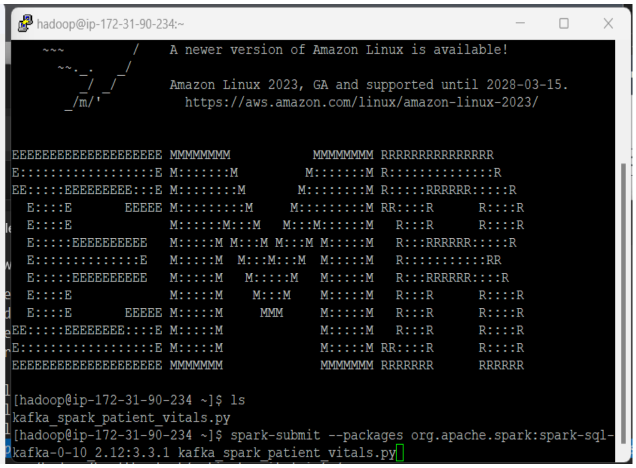
Spark Submitting Job to Consume Message from The Topic Patients-Vital-Info And Stored To HDFS Location

For Producer application: **python3 kafka_produce_patients_vitals.py**



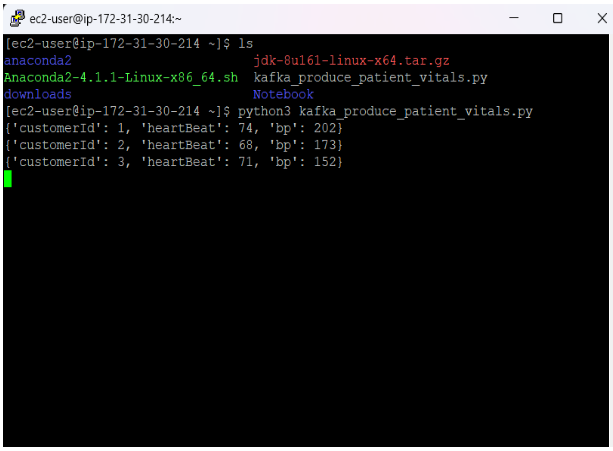
ec2-user@ip-172-31-30-214:~

```
[ec2-user@ip-172-31-30-214 ~]$ ls
anaconda2                         jdk-8u161-linux-x64.tar.gz
Anaconda2-4.1.1-Linux-x86_64.sh   kafka_produce_patient_vitals.py
downloads                         Notebook
[ec2-user@ip-172-31-30-214 ~]$ python3 kafka_produce_patient_vitals.py
```



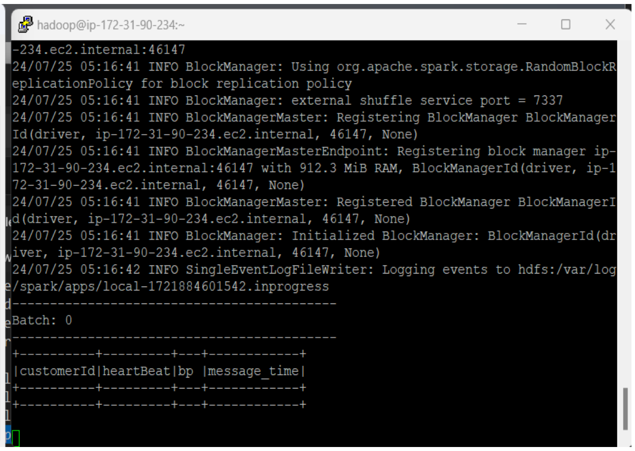
hadoop@ip-172-31-90-234:~

```
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/


EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-90-234 ~]$ ls
kafka_spark_patient_vitals.py
[hadoop@ip-172-31-90-234 ~]$ spark-submit --packages org.apache.spark:spark-sql-
kafka-0-10_2.12:3.3.1 kafka_spark_patient_vitals.py
```



ec2-user@ip-172-31-30-214:~

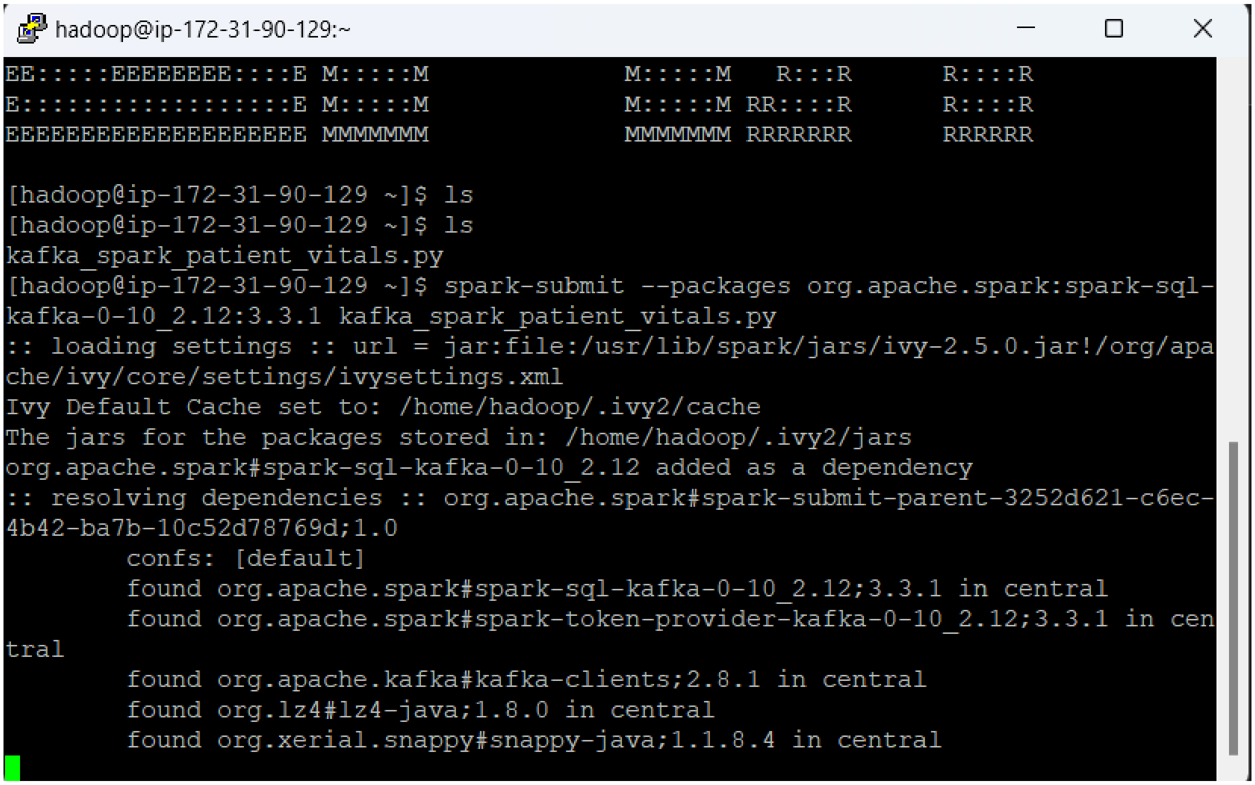
```
[ec2-user@ip-172-31-30-214 ~]$ ls
anaconda2                         jdk-8u161-linux-x64.tar.gz
Anaconda2-4.1.1-Linux-x86_64.sh   kafka_produce_patient_vitals.py
downloads                         Notebook
[ec2-user@ip-172-31-30-214 ~]$ python3 kafka_produce_patient_vitals.py
{'customerId': 1, 'heartBeat': 74, 'bp': 202}
{'customerId': 2, 'heartBeat': 68, 'bp': 173}
{'customerId': 3, 'heartBeat': 71, 'bp': 152}
```



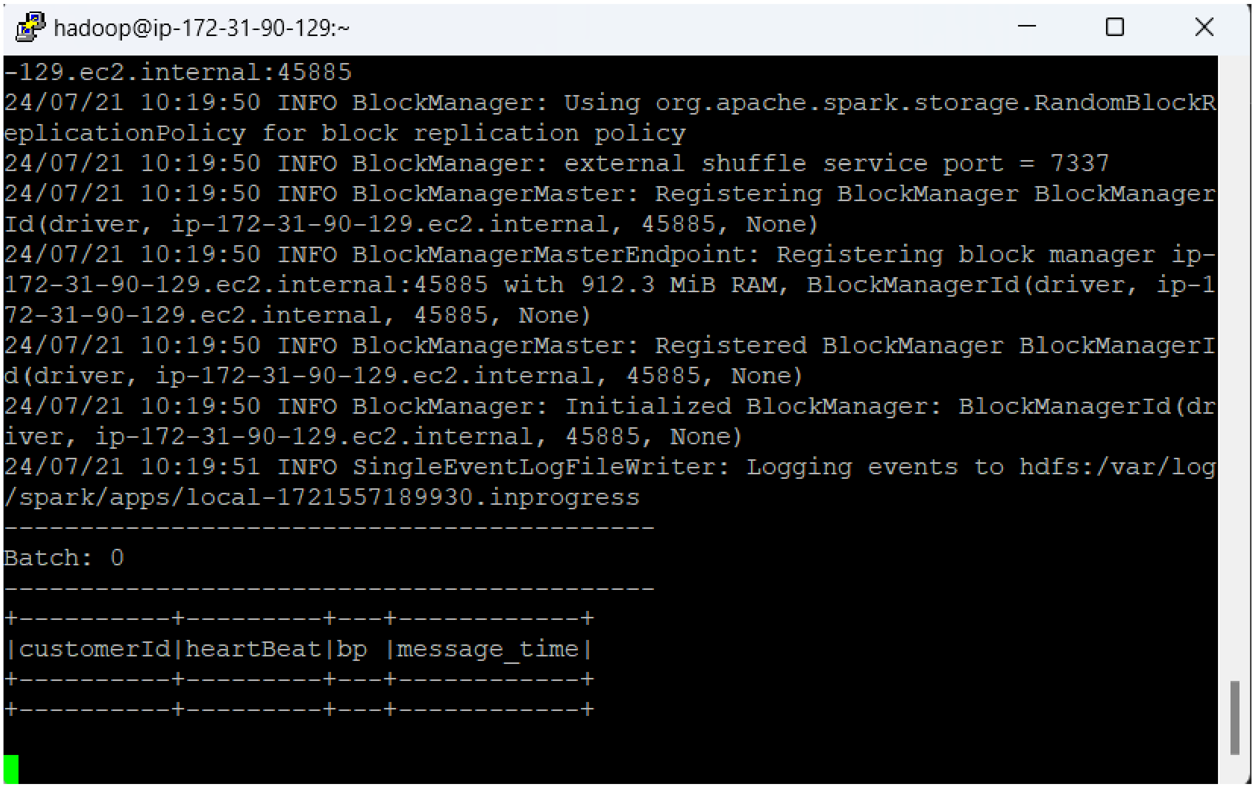
hadoop@ip-172-31-90-234:~

```
-234.ec2.internal:46147
24/07/25 05:16:41 INFO BlockManager: Using org.apache.spark.storage.RandomBlockR
eplicationPolicy for block replication policy
24/07/25 05:16:41 INFO BlockManager: external shuffle service port = 7337
24/07/25 05:16:41 INFO BlockManagerMaster: Registering BlockManager BlockManager
Id(driver, ip-172-31-90-234.ec2.internal, 46147, None)
24/07/25 05:16:41 INFO BlockManagerMasterEndpoint: Registering block manager ip-
172-31-90-234.ec2.internal:46147 with 912.3 MiB RAM, BlockManagerId(driver, ip-1
72-31-90-234.ec2.internal, 46147, None)
24/07/25 05:16:41 INFO BlockManagerMaster: Registered BlockManager BlockManagerI
d(driver, ip-172-31-90-234.ec2.internal, 46147, None)
24/07/25 05:16:41 INFO BlockManager: Initialized BlockManager: BlockManagerId(dr
iver, ip-172-31-90-234.ec2.internal, 46147, None)
24/07/25 05:16:42 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log
/spark/apps/local-1721884601542.inprogress
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+---------+---+------------+
|customerId|heartBeat|bp |message_time|
+----------+---------+---+------------+
+----------+---------+---+------------+
```

For Consumer Application: **spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.3.1 kafka_spark_patient_vitals.py**

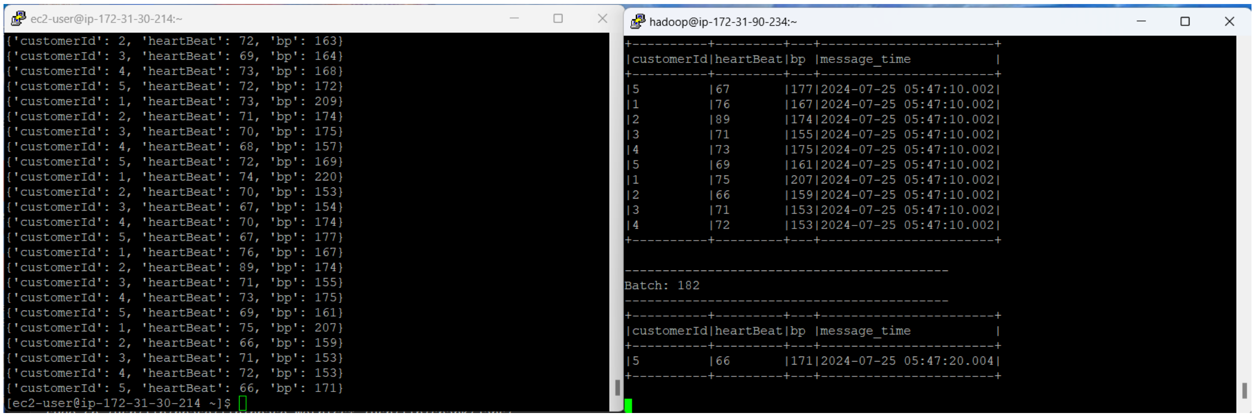


```
EE:::::EEEEEEEE::::E M:::::M                   M:::::M   R:::R       R::::R
E::::::::::::::::::E M:::::M                   M:::::M RR::::R       R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM                   MMMMMMM RRRRRRR       RRRRRR

[hadoop@ip-172-31-90-129 ~]$ ls
[hadoop@ip-172-31-90-129 ~]$ ls
kafka_spark_patient_vitals.py
[hadoop@ip-172-31-90-129 ~]$ spark-submit --packages org.apache.spark:spark-sql-
kafka-0-10_2.12:3.3.1 kafka_spark_patient_vitals.py
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.5.0.jar!/org/apa
che/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-3252d621-c6ec-
4b42-ba7b-10c52d78769d;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.3.1 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.3.1 in cen
tral
        found org.apache.kafka#kafka-clients;2.8.1 in central
        found org.lz4#lz4-java;1.8.0 in central
        found org.xerial.snappy#snappy-java;1.1.8.4 in central
```



```
-129.ec2.internal:45885
24/07/21 10:19:50 INFO BlockManager: Using org.apache.spark.storage.RandomBlockR
eplicationPolicy for block replication policy
24/07/21 10:19:50 INFO BlockManager: external shuffle service port = 7337
24/07/21 10:19:50 INFO BlockManagerMaster: Registering BlockManager BlockManager
Id(driver, ip-172-31-90-129.ec2.internal, 45885, None)
24/07/21 10:19:50 INFO BlockManagerMasterEndpoint: Registering block manager ip-
172-31-90-129.ec2.internal:45885 with 912.3 MiB RAM, BlockManagerId(driver, ip-1
72-31-90-129.ec2.internal, 45885, None)
24/07/21 10:19:50 INFO BlockManagerMaster: Registered BlockManager BlockManagerI
d(driver, ip-172-31-90-129.ec2.internal, 45885, None)
24/07/21 10:19:50 INFO BlockManager: Initialized BlockManager: BlockManagerId(dr
iver, ip-172-31-90-129.ec2.internal, 45885, None)
24/07/21 10:19:51 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log
/spark/apps/local-1721557189930.inprogress
-------------------------------------------
Batch: 0
-------------------------------------------
+----------+---------+---+------------+
|customerId|heartBeat|bp |message_time|
+----------+---------+---+------------+
+----------+---------+---+------------+
```

After 30 minutes when all 1800 data being streamed and saved to Parquet file of the required HDFS location

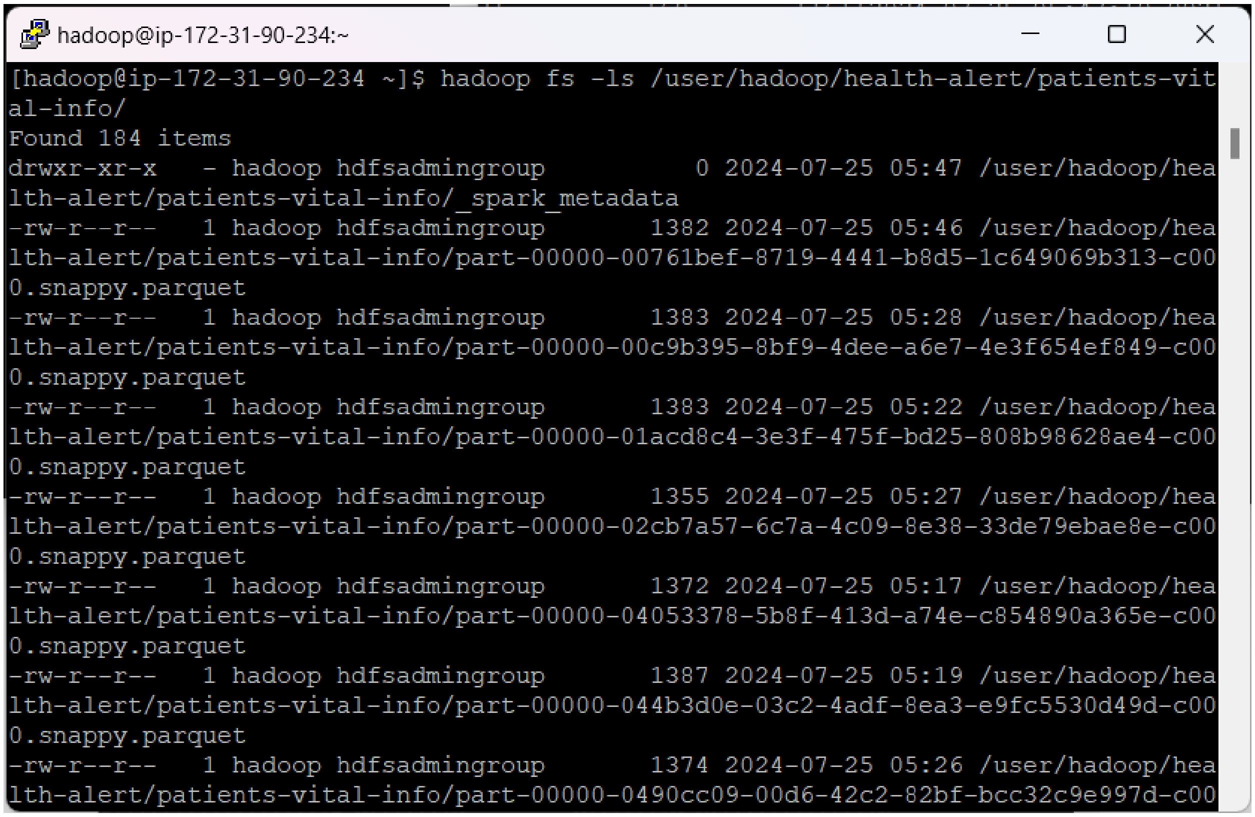


```
{'customerId': 2, 'heartBeat': 72, 'bp': 163}
{'customerId': 3, 'heartBeat': 69, 'bp': 164}
{'customerId': 4, 'heartBeat': 73, 'bp': 168}
{'customerId': 5, 'heartBeat': 72, 'bp': 172}
{'customerId': 1, 'heartBeat': 73, 'bp': 209}
{'customerId': 2, 'heartBeat': 71, 'bp': 174}
{'customerId': 3, 'heartBeat': 70, 'bp': 175}
{'customerId': 4, 'heartBeat': 68, 'bp': 157}
{'customerId': 5, 'heartBeat': 72, 'bp': 169}
{'customerId': 1, 'heartBeat': 74, 'bp': 220}
{'customerId': 2, 'heartBeat': 70, 'bp': 153}
{'customerId': 3, 'heartBeat': 67, 'bp': 154}
{'customerId': 4, 'heartBeat': 70, 'bp': 174}
{'customerId': 5, 'heartBeat': 67, 'bp': 177}
{'customerId': 1, 'heartBeat': 76, 'bp': 167}
{'customerId': 2, 'heartBeat': 89, 'bp': 174}
{'customerId': 3, 'heartBeat': 71, 'bp': 155}
{'customerId': 4, 'heartBeat': 73, 'bp': 175}
{'customerId': 5, 'heartBeat': 69, 'bp': 161}
{'customerId': 1, 'heartBeat': 75, 'bp': 207}
{'customerId': 2, 'heartBeat': 66, 'bp': 159}
{'customerId': 3, 'heartBeat': 71, 'bp': 153}
{'customerId': 4, 'heartBeat': 72, 'bp': 153}
{'customerId': 5, 'heartBeat': 66, 'bp': 171}
[ec2-user@ip-172-31-30-214 ~]$
```

```
+----------+---------+---+-----------------------+
|customerId|heartBeat|bp |message_time           |
+----------+---------+---+-----------------------+
|5         |67       |177|2024-07-25 05:47:10.002|
|1         |76       |167|2024-07-25 05:47:10.002|
|2         |89       |174|2024-07-25 05:47:10.002|
|3         |71       |155|2024-07-25 05:47:10.002|
|4         |73       |175|2024-07-25 05:47:10.002|
|5         |69       |161|2024-07-25 05:47:10.002|
|1         |75       |207|2024-07-25 05:47:10.002|
|2         |66       |159|2024-07-25 05:47:10.002|
|3         |71       |153|2024-07-25 05:47:10.002|
|4         |72       |153|2024-07-25 05:47:10.002|
+----------+---------+---+-----------------------+

-------------------------------------------
Batch: 182
-------------------------------------------
+----------+---------+---+-----------------------+
|customerId|heartBeat|bp |message_time           |
+----------+---------+---+-----------------------+
|5         |66       |171|2024-07-25 05:47:20.004|
+----------+---------+---+-----------------------+
```

STATEMENT TO CHECK DATA STORED IN HDFS LOCATION:

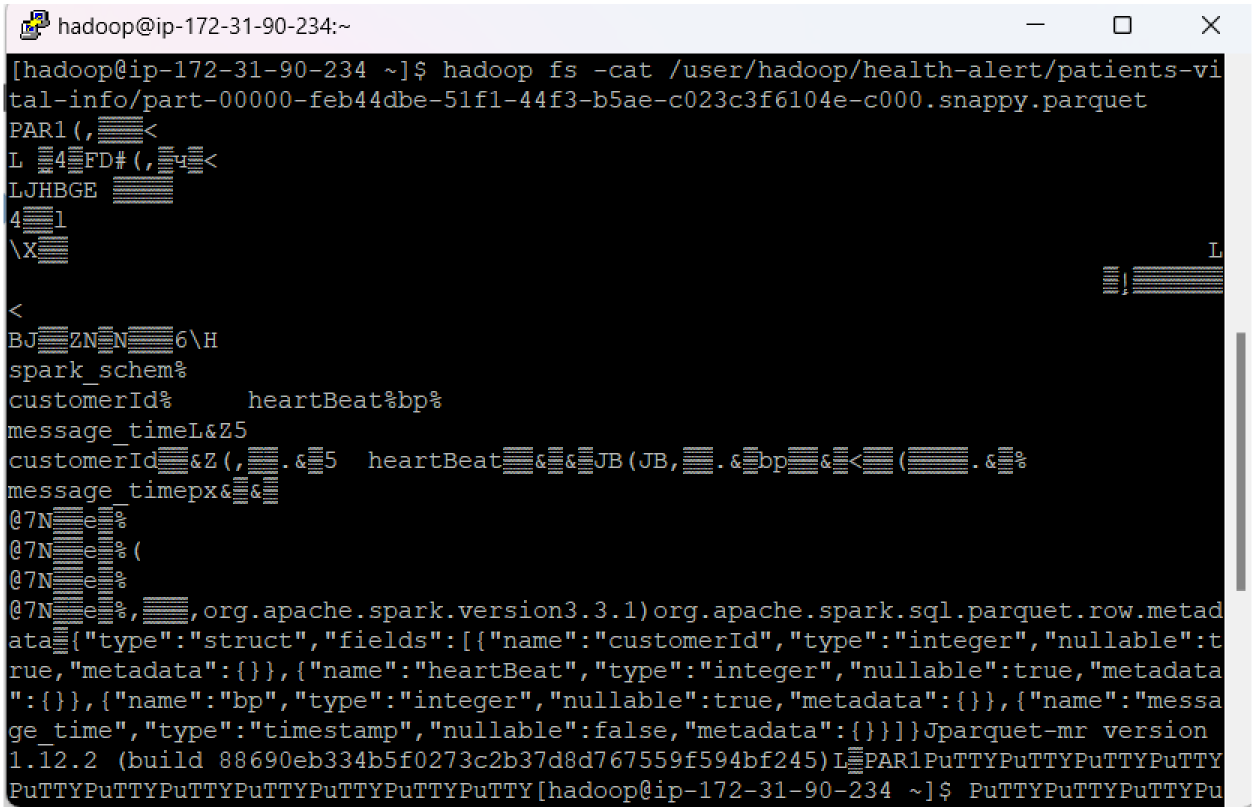
hadoop fs -ls /user/hadoop/health-alert/patients-vital-info/



```
[hadoop@ip-172-31-90-234 ~]$ hadoop fs -ls /user/hadoop/health-alert/patients-vit
al-info/
Found 184 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-07-25 05:47 /user/hadoop/hea
lth-alert/patients-vital-info/_spark_metadata
-rw-r--r--   1 hadoop hdfsadmingroup       1382 2024-07-25 05:46 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-00761bef-8719-4441-b8d5-1c649069b313-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1383 2024-07-25 05:28 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-00c9b395-8bf9-4dee-a6e7-4e3f654ef849-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1383 2024-07-25 05:22 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-01acd8c4-3e3f-475f-bd25-808b98628ae4-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1355 2024-07-25 05:27 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-02cb7a57-6c7a-4c09-8e38-33de79ebae8e-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1372 2024-07-25 05:17 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-04053378-5b8f-413d-a74e-c854890a365e-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1387 2024-07-25 05:19 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-044b3d0e-03c2-4adf-8ea3-e9fc5530d49d-c00
0.snappy.parquet
-rw-r--r--   1 hadoop hdfsadmingroup       1374 2024-07-25 05:26 /user/hadoop/hea
lth-alert/patients-vital-info/part-00000-0490cc09-00d6-42c2-82bf-bcc32c9e997d-c00
```

STATEMENT TO READ ONE OF THE FILES USING '-CAT'

hadoop fs -cat /user/hadoop/health-alert/patients-vital-info/part-00000-ffcd45dc-ef2e-4219-b11e-f26f0e051d73-c000.snappy.parquet



```
[hadoop@ip-172-31-90-234 ~]$ hadoop fs -cat /user/hadoop/health-alert/patients-vi
tal-info/part-00000-feb44dbe-51f1-44f3-b5ae-c023c3f6104e-c000.snappy.parquet
PAR1(,▒▒▒<
L ▒4▒FD#(,▒ч▒<
LJHBGE ▒▒▒▒
4▒▒l
\X▒▒
                                                                                L
                                                                        ▒▒|▒▒▒▒▒▒

<
BJ▒▒ZN▒N▒▒▒6\H
spark_schem%
customerId%     heartBeat%bp%
message_timeL&Z5
customerId▒▒&Z(,▒▒.&▒5  heartBeat▒▒&▒&▒JB(JB,▒▒.&▒bp▒▒&▒<▒▒(▒▒▒▒.&▒%
message_timepx&▒&▒
@7N▒▒e▒%
@7N▒▒e▒%(
@7N▒▒e▒%
@7N▒▒e▒%,▒▒▒,org.apache.spark.version3.3.1)org.apache.spark.sql.parquet.row.metad
ata▒{"type":"struct","fields":[{"name":"customerId","type":"integer","nullable":t
rue,"metadata":{}},{"name":"heartBeat","type":"integer","nullable":true,"metadata
":{}},{"name":"bp","type":"integer","nullable":true,"metadata":{}},{"name":"messa
ge_time","type":"timestamp","nullable":false,"metadata":{}}]}Jparquet-mr version
1.12.2 (build 88690eb334b5f0273c2b37d8d767559f594bf245)L▒PAR1PuTTYPuTTYPuTTYPuTTY
PuTTYPuTTYPuTTYPuTTYPuTTYPuTTYPuTTY[hadoop@ip-172-31-90-234 ~]$ PuTTYPuTTYPuTTYPu
```