Big Data Technologies
Assignment - 9

Exercise-1

a)

The Kappa Architecture improves data processing by utilizing the exact same architecture to perform real-time data processing and historic processing by batches, reducing the need for distinct systems. It effortlessly incorporates current business logic into the Lambda Architecture, reducing the impact on the speed layer for the two operating modes. With previous techniques, which recompute all data anytime business logic changes, Kafka simply calculates data impacted by the alteration, resulting in more efficient resource utilization and responsive processing of data.

b)

Advantage: The capacity of pure streaming systems to process data in real time accounts for their low latency.
Drawback: The disadvantage of stream-only devices is that each data item incurs significant overhead costs, particularly those associated with communications.
Because batch processing systems value resource economy over latency, they are inconsistent with applications that run in real time.

c)

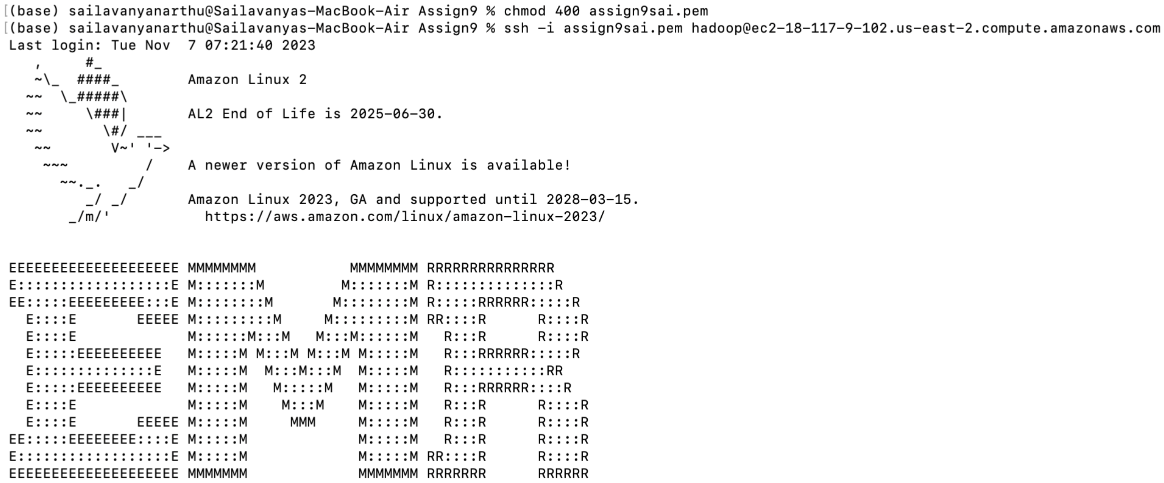
For controlling data processing, Storm uses a framework known as a "topology." 'Bolts' are processing nodes in a topology, assigned with modifying data, storing it outside, and perhaps transferring data to downstream modules. The data flow starts with spouts,' which are in charge of breaking down the information that is provided.

d)

Spark Streaming separates constant information streams into controllable chunks and changes them into Resilient Distributed Datasets (RDDs) for processing using Spark's conventional batch processing algorithms to enable real-time processing. For optimal data distribution and flow, Spark Streaming handles data distribution and flow efficiently.

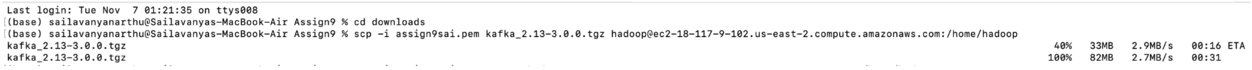
Exercise-2

Starting an EMR Cluster



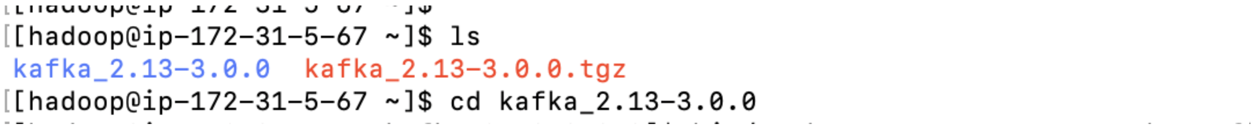
```
[(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % chmod 400 assign9sai.pem
[(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % ssh -i assign9sai.pem hadoop@ec2-18-117-9-102.us-east-2.compute.amazonaws.com
 Last login: Tue Nov  7 07:21:40 2023
    ,     #_
    ~\_  ####_        Amazon Linux 2
   ~~  \_#####\
   ~~     \###|       AL2 End of Life is 2025-06-30.
   ~~       \#/ ___
    ~~       V~' '->
     ~~~         /    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
```

Now, moving kafka file to home/hadoop directory



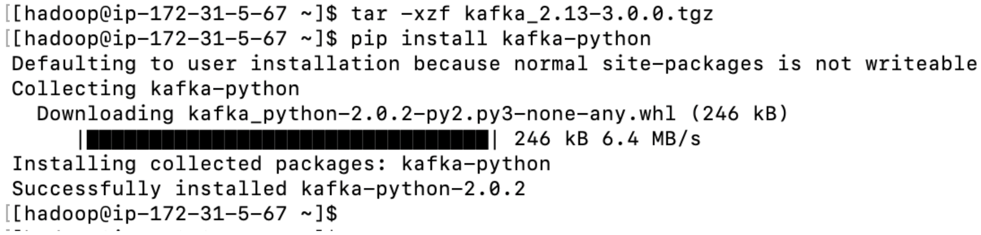
```
Last login: Tue Nov  7 01:21:35 on ttys008
[(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % cd downloads
[(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % scp -i assign9sai.pem kafka_2.13-3.0.0.tgz hadoop@ec2-18-117-9-102.us-east-2.compute.amazonaws.com:/home/hadoop
kafka_2.13-3.0.0.tgz                                                                                                                    40%   33MB   2.9MB/s   00:16 ETA
kafka_2.13-3.0.0.tgz                                                                                                                   100%   82MB   2.7MB/s   00:31
```

Extracting the kafka Package



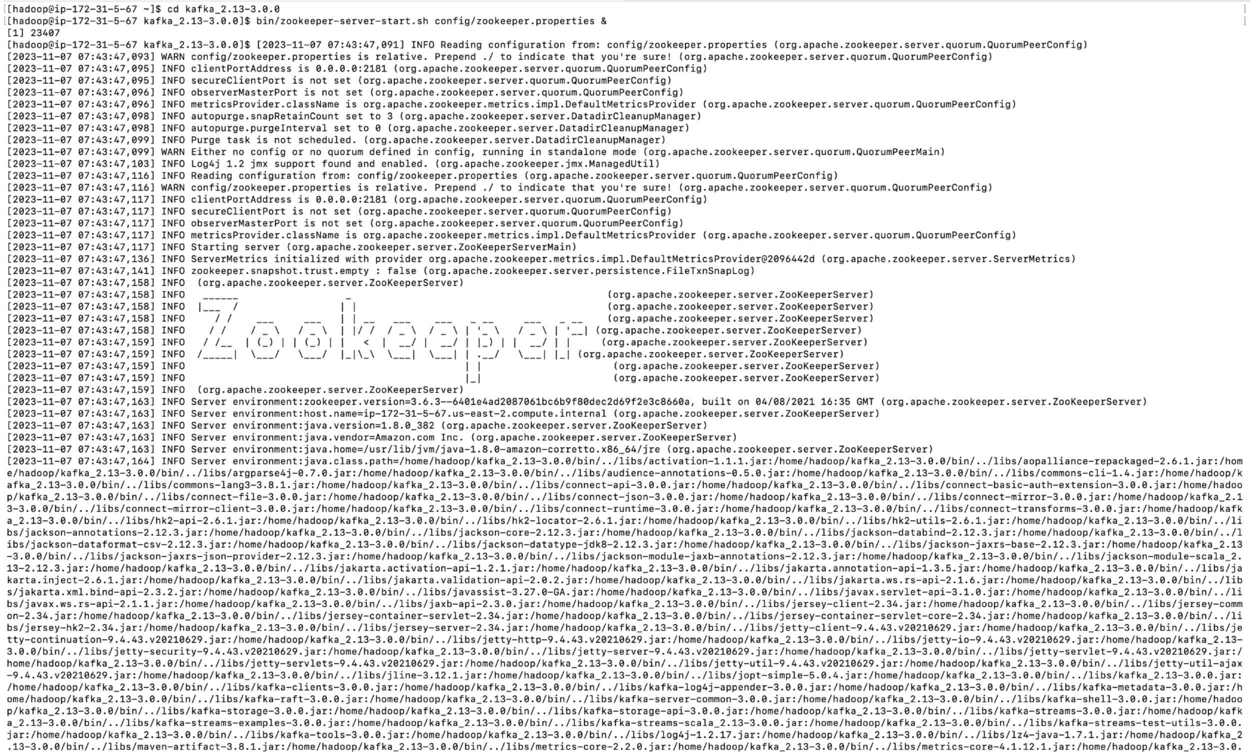
```
[[hadoop@ip-172-31-5-67 ~]$ ls
 kafka_2.13-3.0.0  kafka_2.13-3.0.0.tgz
[[hadoop@ip-172-31-5-67 ~]$ cd kafka_2.13-3.0.0
```

Installing the kafka Package

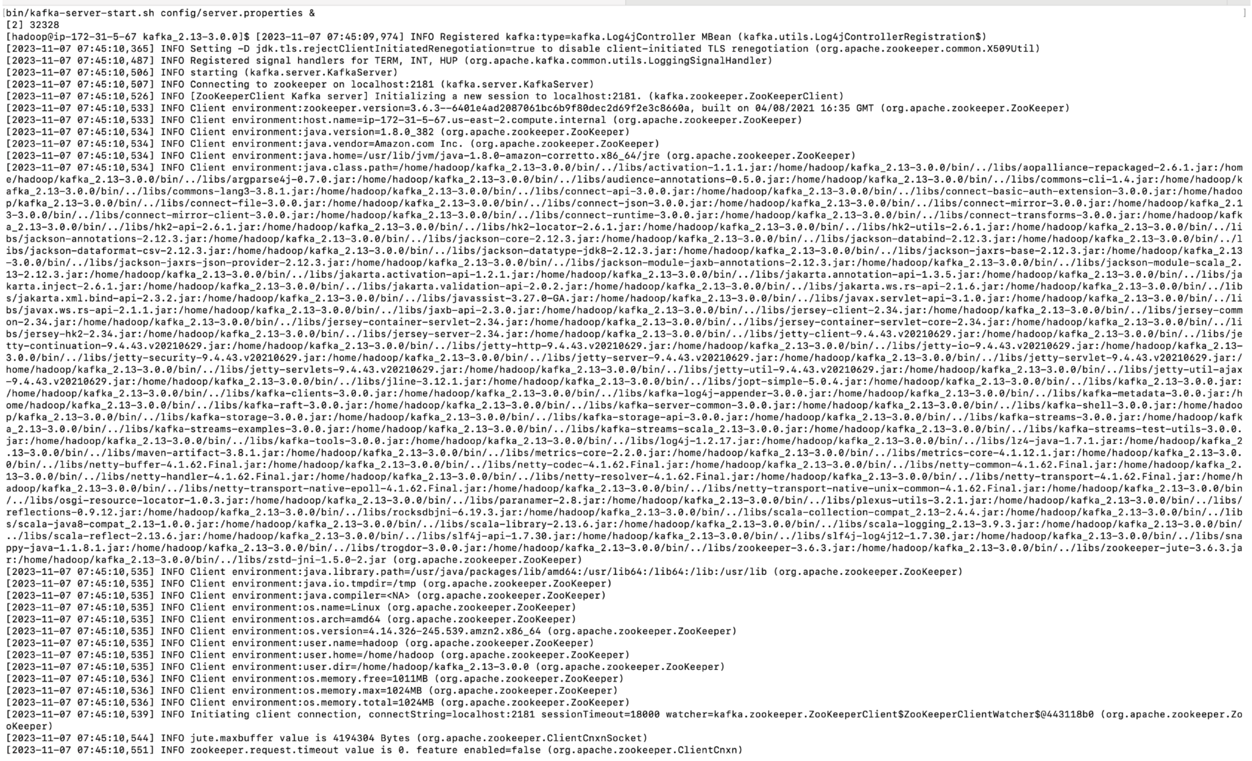


```
[[hadoop@ip-172-31-5-67 ~]$ tar -xzf kafka_2.13-3.0.0.tgz
[[hadoop@ip-172-31-5-67 ~]$ pip install kafka-python
 Defaulting to user installation because normal site-packages is not writeable
 Collecting kafka-python
   Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
      |████████████████████████████████| 246 kB 6.4 MB/s
 Installing collected packages: kafka-python
 Successfully installed kafka-python-2.0.2
[[hadoop@ip-172-31-5-67 ~]$
```

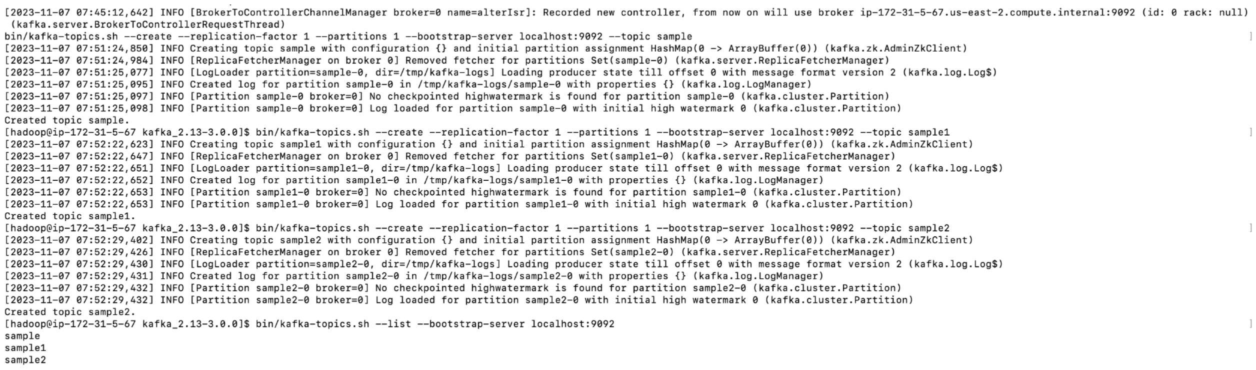
Commands



```
[hadoop@ip-172-31-5-67 ~]$ cd kafka_2.13-3.0.0
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ bin/zookeeper-server-start.sh config/zookeeper.properties &
[1] 23407
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ [2023-11-07 07:43:47,091] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,093] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,095] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,095] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,096] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,096] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,098] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-11-07 07:43:47,098] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-11-07 07:43:47,099] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-11-07 07:43:47,099] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2023-11-07 07:43:47,103] INFO Log4j 1.2 jmx support found and enabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2023-11-07 07:43:47,116] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,116] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,117] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,117] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,117] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,117] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-11-07 07:43:47,117] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2023-11-07 07:43:47,136] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@2096442d (org.apache.zookeeper.server.ServerMetrics)
[2023-11-07 07:43:47,141] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2023-11-07 07:43:47,158] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,158] INFO   ______                  _                                          (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,158] INFO  |___  /                 | |                                         (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,158] INFO     / /    ___     ___   | | __   ___    ___   _ __     ___   _ __   (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,158] INFO    / /    / _ \   / _ \  | |/ /  / _ \  / _ \ | '_ \   / _ \ | '__| (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,159] INFO   / /__  | (_) | | (_) | |   <  |  __/ |  __/ | |_) | |  __/ | |     (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,159] INFO  /_____|  \___/   \___/  |_|\_\  \___|  \___| | .__/   \___| |_| (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,159] INFO                                               | |                     (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,159] INFO                                               |_|                     (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,159] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,163] INFO Server environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,163] INFO Server environment:host.name=ip-172-31-5-67.us-east-2.compute.internal (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,163] INFO Server environment:java.version=1.8.0_382 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,163] INFO Server environment:java.vendor=Amazon.com Inc. (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,163] INFO Server environment:java.home=/usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre (org.apache.zookeeper.server.ZooKeeperServer)
[2023-11-07 07:43:47,164] INFO Server environment:java.class.path=/home/hadoop/kafka_2.13-3.0.0/bin/../libs/activation-1.1.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/aopalliance-repackaged-2.6.1.jar:/hom
e/hadoop/kafka_2.13-3.0.0/bin/../libs/argparse4j-0.7.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/audience-annotations-0.5.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/commons-cli-1.4.jar:/home/hadoop/k
afka_2.13-3.0.0/bin/../libs/commons-lang3-3.8.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-api-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-basic-auth-extension-3.0.0.jar:/home/hadoo
p/kafka_2.13-3.0.0/bin/../libs/connect-file-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-json-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-mirror-3.0.0.jar:/home/hadoop/kafka_2.1
3-3.0.0/bin/../libs/connect-mirror-client-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-runtime-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-transforms-3.0.0.jar:/home/hadoop/kafk
a_2.13-3.0.0/bin/../libs/hk2-api-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/hk2-locator-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/hk2-utils-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/jackson-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-core-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-databind-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../l
ibs/jackson-dataformat-csv-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-datatype-jdk8-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-jaxrs-base-2.12.3.jar:/home/hadoop/kafka_2.13
-3.0.0/bin/../libs/jackson-jaxrs-json-provider-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-module-jaxb-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-module-scala_2.
13-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.activation-api-1.2.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.annotation-api-1.3.5.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/ja
karta.inject-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.validation-api-2.0.2.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.ws.rs-api-2.1.6.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../lib
s/jakarta.xml.bind-api-2.3.2.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/javassist-3.27.0-GA.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/javax.servlet-api-3.1.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/javax.ws.rs-api-2.1.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jaxb-api-2.3.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-client-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-comm
on-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-container-servlet-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-container-servlet-core-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/jersey-hk2-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-server-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-client-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/je
tty-continuation-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-http-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-io-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-
3.0.0/bin/../libs/jetty-security-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-server-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-servlet-9.4.43.v20210629.jar:/
home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-servlets-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-util-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-util-ajax
-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jline-3.12.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jopt-simple-5.0.4.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka_2.13-3.0.0.jar:
/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-clients-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-log4j-appender-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-metadata-3.0.0.jar:/h
ome/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-raft-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-server-common-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-shell-3.0.0.jar:/home/hadoo
p/kafka_2.13-3.0.0/bin/../libs/kafka-storage-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-storage-api-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-3.0.0.jar:/home/hadoop/kafk
a_2.13-3.0.0/bin/../libs/kafka-streams-examples-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-scala_2.13-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-test-utils-3.0.0.
jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-tools-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/log4j-1.2.17.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/lz4-java-1.7.1.jar:/home/hadoop/kafka_2
.13-3.0.0/bin/../libs/maven-artifact-3.8.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/metrics-core-2.2.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/metrics-core-4.1.12.1.jar:/home/hadoop/kafka_2.13-3.0.
```

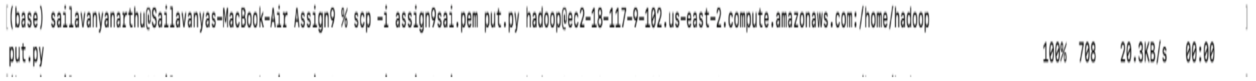


```
bin/kafka-server-start.sh config/server.properties &
[2] 32328
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ [2023-11-07 07:45:09,974] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2023-11-07 07:45:10,365] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2023-11-07 07:45:10,487] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2023-11-07 07:45:10,506] INFO starting (kafka.server.KafkaServer)
[2023-11-07 07:45:10,507] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2023-11-07 07:45:10,526] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2023-11-07 07:45:10,533] INFO Client environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,533] INFO Client environment:host.name=ip-172-31-5-67.us-east-2.compute.internal (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,534] INFO Client environment:java.version=1.8.0_382 (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,534] INFO Client environment:java.vendor=Amazon.com Inc. (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,534] INFO Client environment:java.home=/usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,534] INFO Client environment:java.class.path=/home/hadoop/kafka_2.13-3.0.0/bin/../libs/activation-1.1.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/aopalliance-repackaged-2.6.1.jar:/hom
e/hadoop/kafka_2.13-3.0.0/bin/../libs/argparse4j-0.7.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/audience-annotations-0.5.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/commons-cli-1.4.jar:/home/hadoop/k
afka_2.13-3.0.0/bin/../libs/commons-lang3-3.8.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-api-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-basic-auth-extension-3.0.0.jar:/home/hadoo
p/kafka_2.13-3.0.0/bin/../libs/connect-file-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-json-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-mirror-3.0.0.jar:/home/hadoop/kafka_2.1
3-3.0.0/bin/../libs/connect-mirror-client-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-runtime-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/connect-transforms-3.0.0.jar:/home/hadoop/kafk
a_2.13-3.0.0/bin/../libs/hk2-api-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/hk2-locator-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/hk2-utils-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/jackson-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-core-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-databind-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../l
ibs/jackson-dataformat-csv-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-datatype-jdk8-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-jaxrs-base-2.12.3.jar:/home/hadoop/kafka_2.13
-3.0.0/bin/../libs/jackson-jaxrs-json-provider-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-module-jaxb-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-module-scala_2.
13-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.activation-api-1.2.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.annotation-api-1.3.5.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/ja
karta.inject-2.6.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.validation-api-2.0.2.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jakarta.ws.rs-api-2.1.6.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../lib
s/jakarta.xml.bind-api-2.3.2.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/javassist-3.27.0-GA.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/javax.servlet-api-3.1.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/javax.ws.rs-api-2.1.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jaxb-api-2.3.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-client-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-comm
on-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-container-servlet-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-container-servlet-core-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../li
bs/jersey-hk2-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jersey-server-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-client-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/je
tty-continuation-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-http-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-io-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-
3.0.0/bin/../libs/jetty-security-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-server-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-servlet-9.4.43.v20210629.jar:/
home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-servlets-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-util-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jetty-util-ajax
-9.4.43.v20210629.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jline-3.12.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jopt-simple-5.0.4.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka_2.13-3.0.0.jar:
/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-clients-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-log4j-appender-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-metadata-3.0.0.jar:/h
ome/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-raft-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-server-common-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-shell-3.0.0.jar:/home/hadoo
p/kafka_2.13-3.0.0/bin/../libs/kafka-storage-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-storage-api-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-3.0.0.jar:/home/hadoop/kafk
a_2.13-3.0.0/bin/../libs/kafka-streams-examples-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-scala_2.13-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-streams-test-utils-3.0.0.
jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/kafka-tools-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/log4j-1.2.17.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/lz4-java-1.7.1.jar:/home/hadoop/kafka_2
.13-3.0.0/bin/../libs/maven-artifact-3.8.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/metrics-core-2.2.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/metrics-core-4.1.12.1.jar:/home/hadoop/kafka_2.13-3.0.
0/bin/../libs/netty-buffer-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-codec-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-common-4.1.62.Final.jar:/home/hadoop/kafka_2.
13-3.0.0/bin/../libs/netty-handler-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-resolver-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-transport-4.1.62.Final.jar:/home/h
adoop/kafka_2.13-3.0.0/bin/../libs/netty-transport-native-epoll-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-transport-native-unix-common-4.1.62.Final.jar:/home/hadoop/kafka_2.13-3.0.0/bin
/../libs/osgi-resource-locator-1.0.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/paranamer-2.8.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/plexus-utils-3.2.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/
reflections-0.9.12.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/rocksdbjni-6.19.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/scala-collection-compat_2.13-2.4.4.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../lib
s/scala-java8-compat_2.13-1.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/scala-library-2.13.6.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/scala-logging_2.13-3.9.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/
../libs/scala-reflect-2.13.6.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/slf4j-api-1.7.30.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/slf4j-log4j12-1.7.30.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/sna
ppy-java-1.1.8.1.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/trogdor-3.0.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/zookeeper-3.6.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/zookeeper-jute-3.6.3.ja
r:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/zstd-jni-1.5.0-2.jar (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:java.io.tmpdir=/tmp (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:java.compiler=<NA> (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:os.name=Linux (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:os.arch=amd64 (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:os.version=4.14.326-245.539.amzn2.x86_64 (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:user.name=hadoop (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:user.home=/home/hadoop (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,535] INFO Client environment:user.dir=/home/hadoop/kafka_2.13-3.0.0 (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,536] INFO Client environment:os.memory.free=1011MB (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,536] INFO Client environment:os.memory.max=1024MB (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,536] INFO Client environment:os.memory.total=1024MB (org.apache.zookeeper.ZooKeeper)
[2023-11-07 07:45:10,539] INFO Initiating client connection, connectString=localhost:2181 sessionTimeout=18000 watcher=kafka.zookeeper.ZooKeeperClient$ZooKeeperClientWatcher$@443118b0 (org.apache.zookeeper.Zo
oKeeper)
[2023-11-07 07:45:10,544] INFO jute.maxbuffer value is 4194304 Bytes (org.apache.zookeeper.ClientCnxnSocket)
[2023-11-07 07:45:10,551] INFO zookeeper.request.timeout value is 0. feature enabled=false (org.apache.zookeeper.ClientCnxn)
```

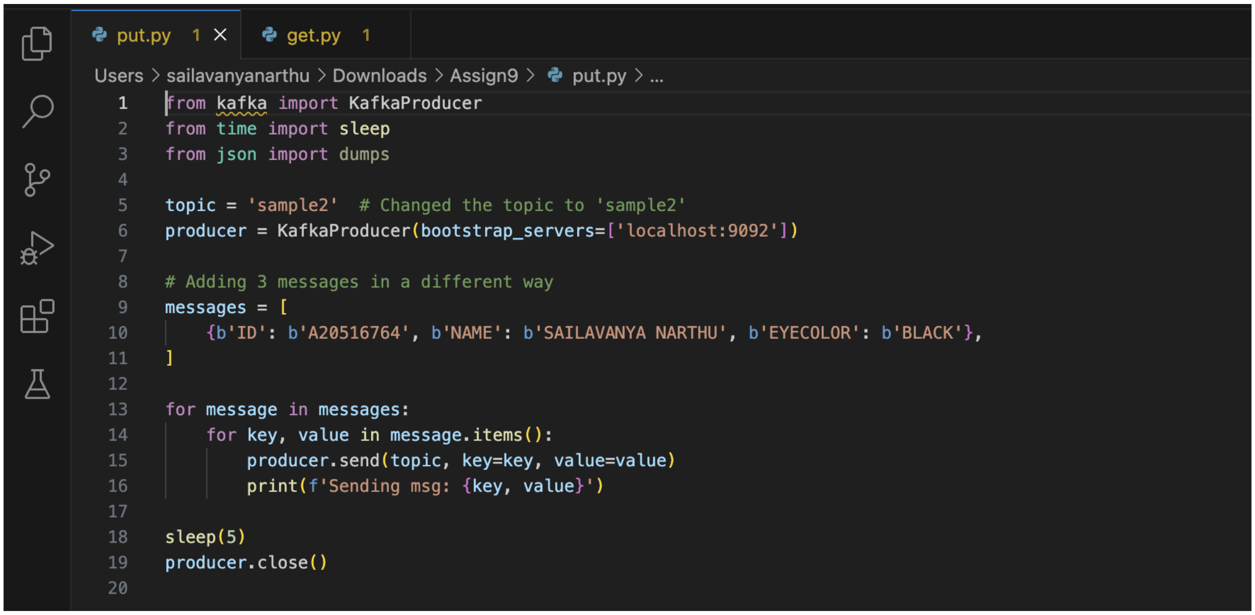


```
[2023-11-07 07:45:12,642] INFO [BrokerToControllerChannelManager broker=0 name=alterIsr]: Recorded new controller, from now on will use broker ip-172-31-5-67.us-east-2.compute.internal:9092 (id: 0 rack: null)
 (kafka.server.BrokerToControllerRequestThread)
bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic sample
[2023-11-07 07:51:24,850] INFO Creating topic sample with configuration {} and initial partition assignment HashMap(0 -> ArrayBuffer(0)) (kafka.zk.AdminZkClient)
[2023-11-07 07:51:24,984] INFO [ReplicaFetcherManager on broker 0] Removed fetcher for partitions Set(sample-0) (kafka.server.ReplicaFetcherManager)
[2023-11-07 07:51:25,077] INFO [LogLoader partition=sample-0, dir=/tmp/kafka-logs] Loading producer state till offset 0 with message format version 2 (kafka.log.Log$)
[2023-11-07 07:51:25,095] INFO Created log for partition sample-0 in /tmp/kafka-logs/sample-0 with properties {} (kafka.log.LogManager)
[2023-11-07 07:51:25,097] INFO [Partition sample-0 broker=0] No checkpointed highwatermark is found for partition sample-0 (kafka.cluster.Partition)
[2023-11-07 07:51:25,098] INFO [Partition sample-0 broker=0] Log loaded for partition sample-0 with initial high watermark 0 (kafka.cluster.Partition)
Created topic sample.
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic sample1
[2023-11-07 07:52:22,623] INFO Creating topic sample1 with configuration {} and initial partition assignment HashMap(0 -> ArrayBuffer(0)) (kafka.zk.AdminZkClient)
[2023-11-07 07:52:22,647] INFO [ReplicaFetcherManager on broker 0] Removed fetcher for partitions Set(sample1-0) (kafka.server.ReplicaFetcherManager)
[2023-11-07 07:52:22,651] INFO [LogLoader partition=sample1-0, dir=/tmp/kafka-logs] Loading producer state till offset 0 with message format version 2 (kafka.log.Log$)
[2023-11-07 07:52:22,652] INFO Created log for partition sample1-0 in /tmp/kafka-logs/sample1-0 with properties {} (kafka.log.LogManager)
[2023-11-07 07:52:22,653] INFO [Partition sample1-0 broker=0] No checkpointed highwatermark is found for partition sample1-0 (kafka.cluster.Partition)
[2023-11-07 07:52:22,653] INFO [Partition sample1-0 broker=0] Log loaded for partition sample1-0 with initial high watermark 0 (kafka.cluster.Partition)
Created topic sample1.
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic sample2
[2023-11-07 07:52:29,402] INFO Creating topic sample2 with configuration {} and initial partition assignment HashMap(0 -> ArrayBuffer(0)) (kafka.zk.AdminZkClient)
[2023-11-07 07:52:29,426] INFO [ReplicaFetcherManager on broker 0] Removed fetcher for partitions Set(sample2-0) (kafka.server.ReplicaFetcherManager)
[2023-11-07 07:52:29,430] INFO [LogLoader partition=sample2-0, dir=/tmp/kafka-logs] Loading producer state till offset 0 with message format version 2 (kafka.log.Log$)
[2023-11-07 07:52:29,431] INFO Created log for partition sample2-0 in /tmp/kafka-logs/sample2-0 with properties {} (kafka.log.LogManager)
[2023-11-07 07:52:29,432] INFO [Partition sample2-0 broker=0] No checkpointed highwatermark is found for partition sample2-0 (kafka.cluster.Partition)
[2023-11-07 07:52:29,432] INFO [Partition sample2-0 broker=0] Log loaded for partition sample2-0 with initial high watermark 0 (kafka.cluster.Partition)
Created topic sample2.
[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --list --bootstrap-server localhost:9092
sample
sample1
sample2
```

a)

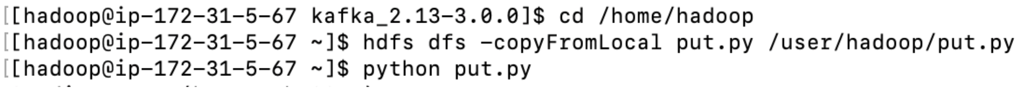


```
[(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % scp -i assign9sai.pem put.py hadoop@ec2-18-117-9-102.us-east-2.compute.amazonaws.com:/home/hadoop
put.py                                                                                              100%  708    20.3KB/s   00:00
```

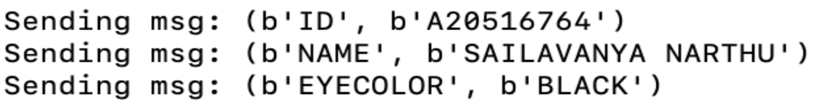

put.py 1
get.py 1
Users > sailavanyanarthu > Downloads > Assign9 > put.py > ...

```python
from kafka import KafkaProducer
from time import sleep
from json import dumps

topic = 'sample2'  # Changed the topic to 'sample2'
producer = KafkaProducer(bootstrap_servers=['localhost:9092'])

# Adding 3 messages in a different way
messages = [
    {b'ID': b'A20516764', b'NAME': b'SAILAVANYA NARTHU', b'EYECOLOR': b'BLACK'},
]

for message in messages:
    for key, value in message.items():
        producer.send(topic, key=key, value=value)
        print(f'Sending msg: {key, value}')

sleep(5)
producer.close()

```

Output:

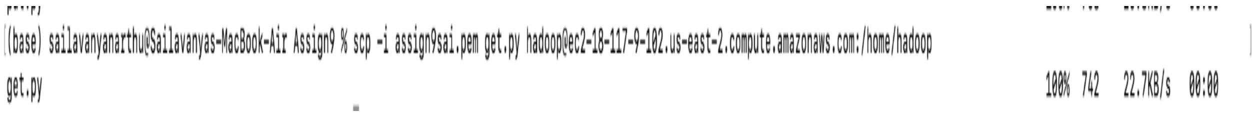


```
[[hadoop@ip-172-31-5-67 kafka_2.13-3.0.0]$ cd /home/hadoop
[[hadoop@ip-172-31-5-67 ~]$ hdfs dfs -copyFromLocal put.py /user/hadoop/put.py
[[hadoop@ip-172-31-5-67 ~]$ python put.py
```

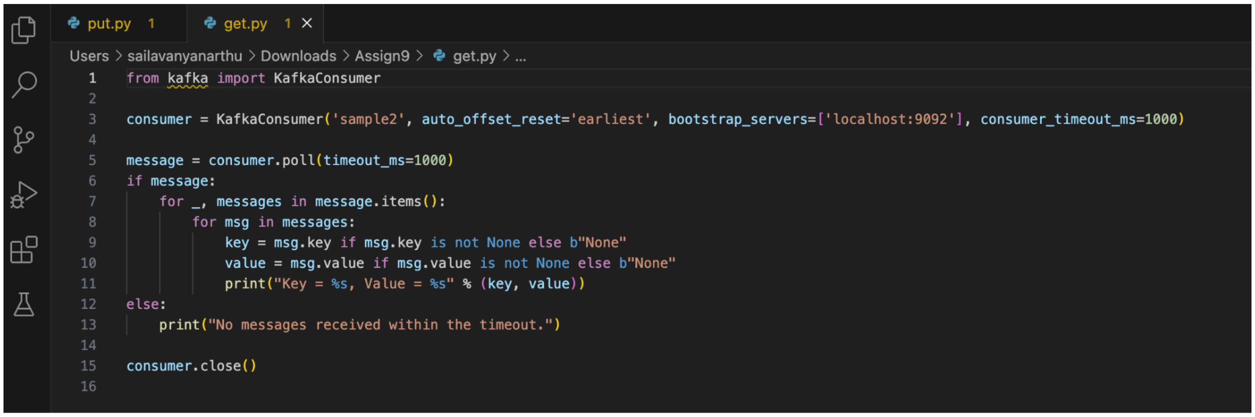


```
Sending msg: (b'ID', b'A20516764')
Sending msg: (b'NAME', b'SAILAVANYA NARTHU')
Sending msg: (b'EYECOLOR', b'BLACK')
```

b)



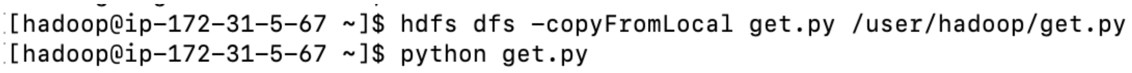
```
(base) sailavanyanarthu@Sailavanyas-MacBook-Air Assign9 % scp -i assign9sai.pem get.py hadoop@ec2-18-117-9-102.us-east-2.compute.amazonaws.com:/home/hadoop
get.py                                                                                    100%  742    22.7KB/s   00:00
```



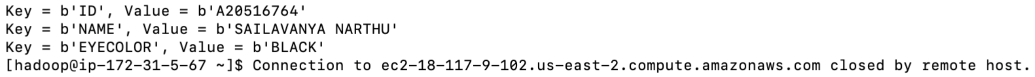
put.py 1
get.py 1 ×
Users > sailavanyanarthu > Downloads > Assign9 > get.py > ...

```python
from kafka import KafkaConsumer

consumer = KafkaConsumer('sample2', auto_offset_reset='earliest', bootstrap_servers=['localhost:9092'], consumer_timeout_ms=1000)

message = consumer.poll(timeout_ms=1000)
if message:
    for _, messages in message.items():
        for msg in messages:
            key = msg.key if msg.key is not None else b"None"
            value = msg.value if msg.value is not None else b"None"
            print("Key = %s, Value = %s" % (key, value))
else:
    print("No messages received within the timeout.")

consumer.close()
```

Output:



```
[hadoop@ip-172-31-5-67 ~]$ hdfs dfs -copyFromLocal get.py /user/hadoop/get.py
[hadoop@ip-172-31-5-67 ~]$ python get.py
```



```
Key = b'ID', Value = b'A20516764'
Key = b'NAME', Value = b'SAILAVANYA NARTHU'
Key = b'EYECOLOR', Value = b'BLACK'
[hadoop@ip-172-31-5-67 ~]$ Connection to ec2-18-117-9-102.us-east-2.compute.amazonaws.com closed by remote host.
```

Submitted by:
Sailavanya Narthu
A20516764
snarthu@hawk.iit.edu