# Big Data Technologies-CSP554

## Assignment-9

**Exercise 1) 5 points**

**Read the article "Real-time stream processing for Big Data" available on the blackboard in the 'Articles' section and then answer the following questions:**

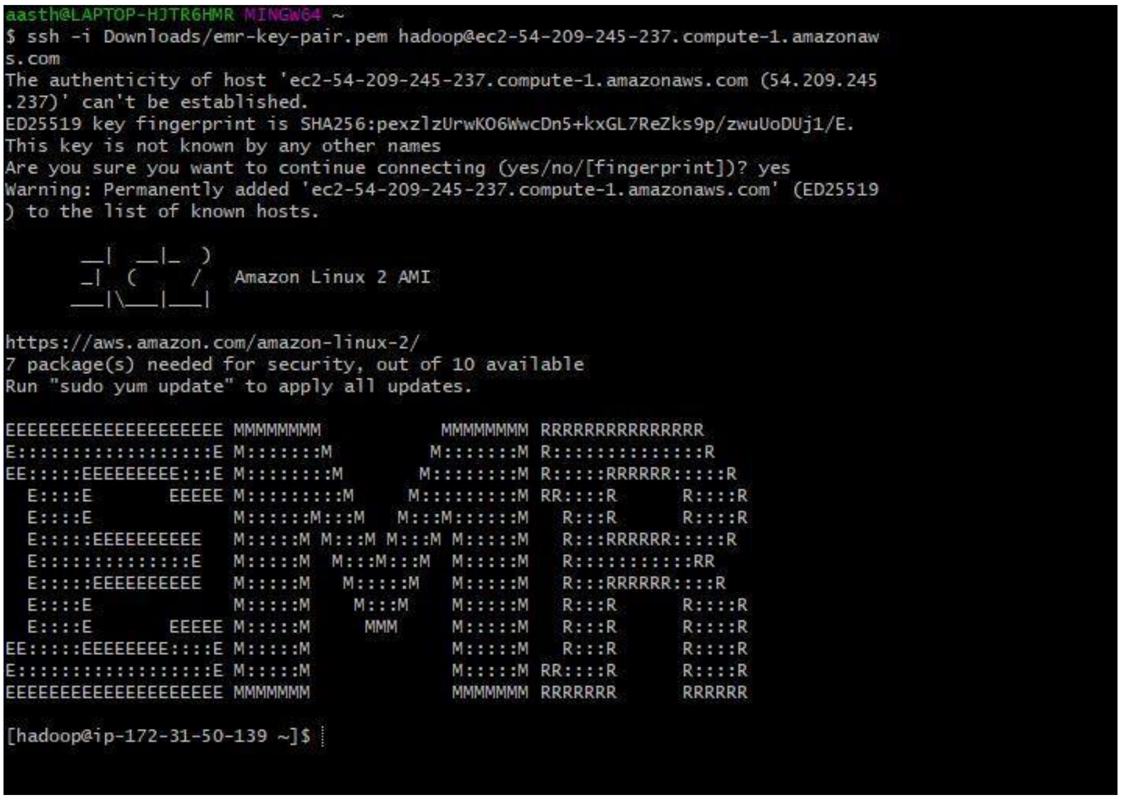
a) **(1.25 points) What is the Kappa architecture and how does it differ from the lambda architecture?**
As mentioned in the article on the kappa architecture, the precomputation of data does not happen on regular basis in the batch layer. All the computation is done in the stream processing system. The re-computation is performed only when the business logic changes by replaying past data. For this, the Kappa Architecture uses a powerful stream processor which can deal with data at a faster speed than with what the data comes in.
On the other hand, Lambda architecture is a system consisting of three layers: Batch, Speed, and Serving layer. It targets both the Volume and Velocity challenge of big data at the same time. It has both a batch-oriented system and a real-time system.

b) **(1.25 points) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?**
Storm and Samza are pure stream-oriented systems with very low latency and somewhat high per-item costs, on the other hand, batch-oriented systems offer unmatched resource efficiency at the tradeoff of unreasonably high latency for real-time applications. Data is buffered and processed in batches in micro-batch real-time processing systems. It improves efficiency while also increasing the time an individual item spends in the data flow. Storm Trident and Spark Streaming are two examples of this type.

c) **(1.25 points) In few sentences describe the data processing pipeline in Storm.**
A topology in Storm refers to a data pipeline or application. Spouts are the nodes that take data and so start the data flow in the topology. Spouts output tuples to bolts, which execute processing, write data to external storage and may transmit tuples further downstream. Data flow between nodes is controlled by storm groupings. Storm distributes spouts and bolts in a round-robin fashion by default, but the scheduler can be modified to accommodate cases in which a specific processing step must be performed on a specific node. Storm offers the option of at least one processing via an acknowledgment feature that maintains the processing status of every single tuple as it travels through the topology. It does not ensure the sequence in which tuples are processed.

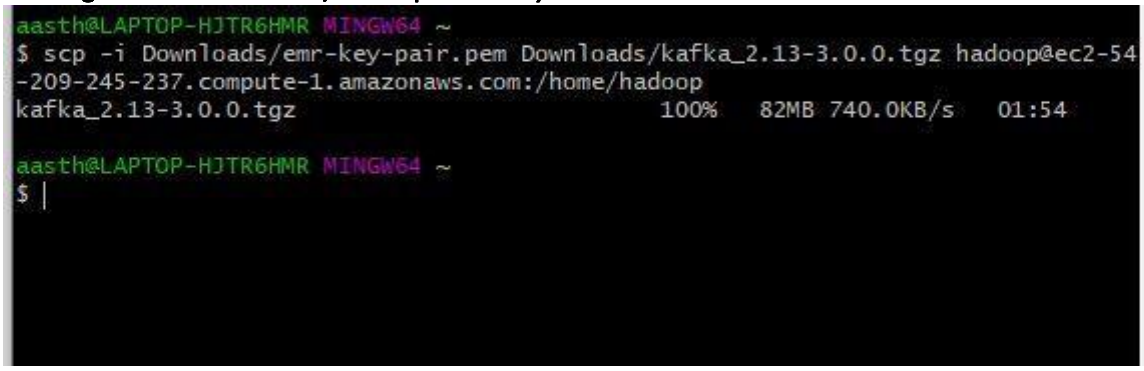
d) **(1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?**
It does so by chunking the stream of incoming data items into small batches, changing them into DDs, and processing them as usual, Spark streaming shifts the batch-processing method towards real-time requirements. It also automatically manages data flow and distribution.

**Exercise 2)**
**Starting an EMR Cluster**



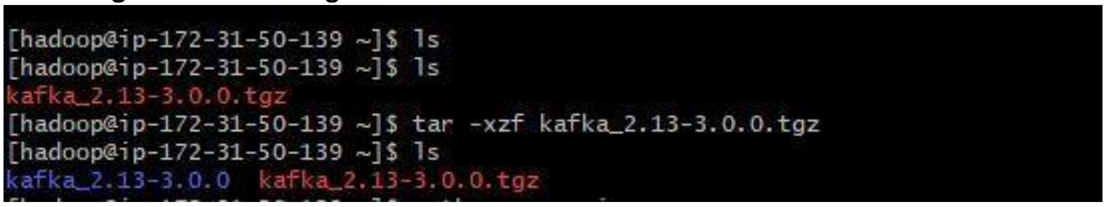
```
aasth@LAPTOP-HJTR6HMR MINGW64 ~
$ ssh -i Downloads/emr-key-pair.pem hadoop@ec2-54-209-245-237.compute-1.amazonaw
s.com
The authenticity of host 'ec2-54-209-245-237.compute-1.amazonaws.com (54.209.245
.237)' can't be established.
ED25519 key fingerprint is SHA256:pexzlzUrwKO6WwcDn5+kxGL7ReZks9p/zwuUoDUj1/E.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-54-209-245-237.compute-1.amazonaws.com' (ED25519
) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
7 package(s) needed for security, out of 10 available
Run "sudo yum update" to apply all updates.

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-172-31-50-139 ~]$
```

**Moving kafka file to home/hadoop directory**



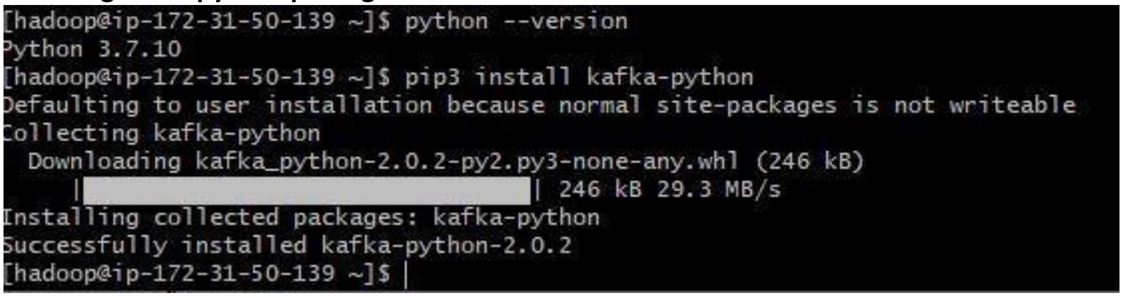
```
aasth@LAPTOP-HJTR6HMR MINGW64 ~
$ scp -i Downloads/emr-key-pair.pem Downloads/kafka_2.13-3.0.0.tgz hadoop@ec2-54
-209-245-237.compute-1.amazonaws.com:/home/hadoop
kafka_2.13-3.0.0.tgz                          100%   82MB 740.0KB/s   01:54

aasth@LAPTOP-HJTR6HMR MINGW64 ~
$ |
```

**Extracting the Kafka Package**



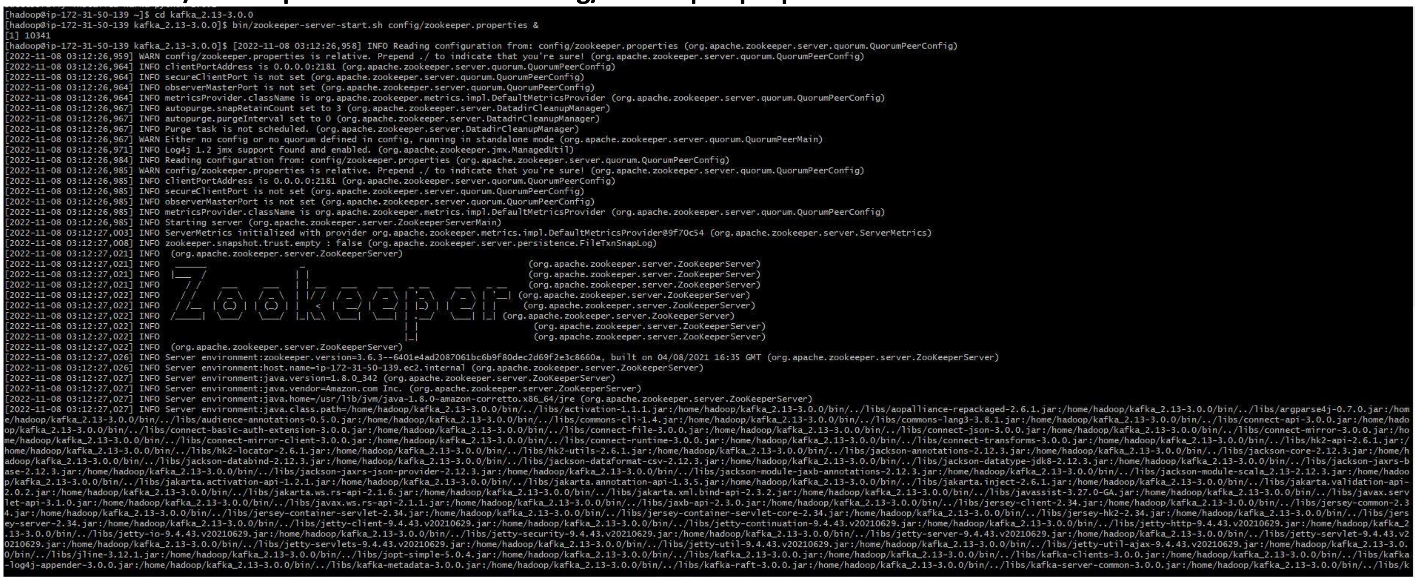
```
[hadoop@ip-172-31-50-139 ~]$ ls
[hadoop@ip-172-31-50-139 ~]$ ls
kafka_2.13-3.0.0.tgz
[hadoop@ip-172-31-50-139 ~]$ tar -xzf kafka_2.13-3.0.0.tgz
[hadoop@ip-172-31-50-139 ~]$ ls
kafka_2.13-3.0.0  kafka_2.13-3.0.0.tgz
```

**Installing Kafka python package**



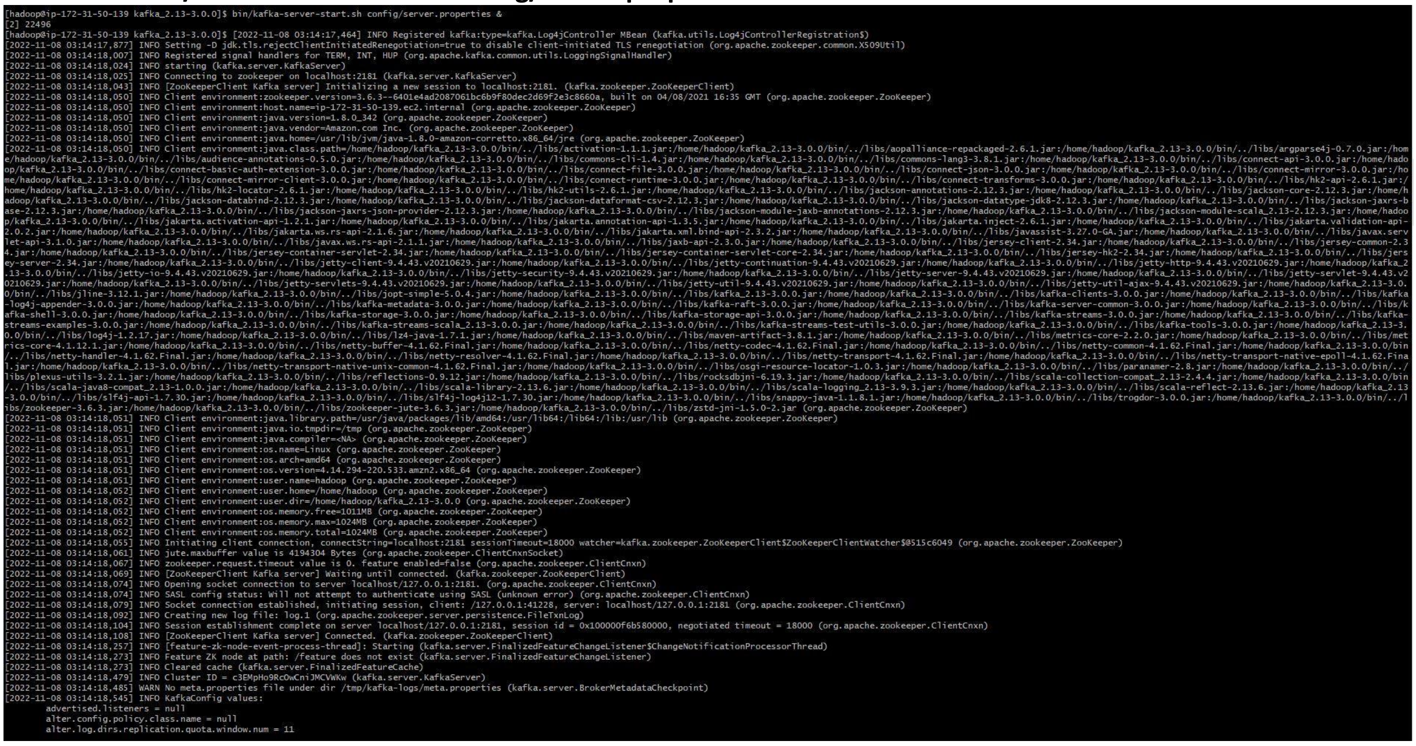
```
[hadoop@ip-172-31-50-139 ~]$ python --version
Python 3.7.10
[hadoop@ip-172-31-50-139 ~]$ pip3 install kafka-python
Defaulting to user installation because normal site-packages is not writeable
Collecting kafka-python
  Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
     |████████████████████████████████| 246 kB 29.3 MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
[hadoop@ip-172-31-50-139 ~]$ |
```

## Command- bin/zookeeper-server-start.sh config/zookeeper.properties &



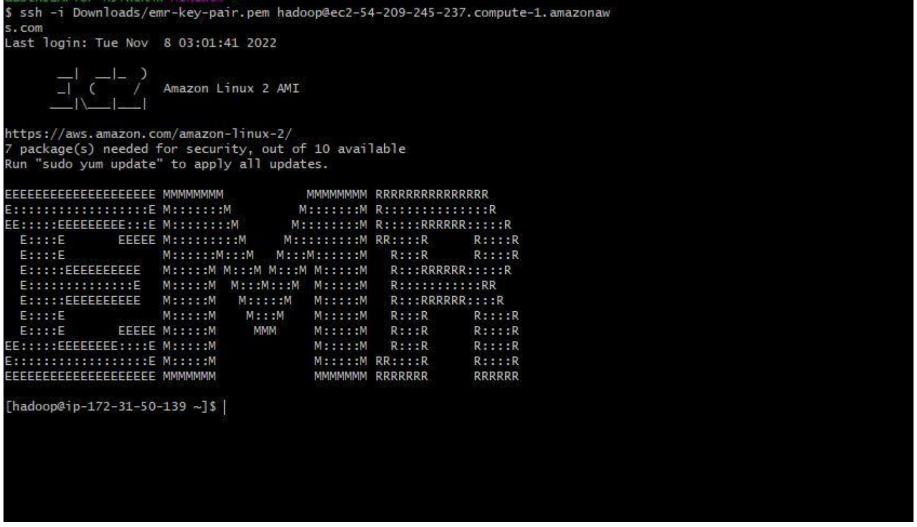
```
[hadoop@ip-172-31-50-139 ~]$ cd kafka_2.13-3.0.0
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/zookeeper-server-start.sh config/zookeeper.properties &
[1] 10341
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ [2022-11-08 03:12:26,958] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,959] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,964] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,964] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,964] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,964] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,967] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-11-08 03:12:26,967] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-11-08 03:12:26,967] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-11-08 03:12:26,967] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2022-11-08 03:12:26,971] INFO Log4j 1.2 jmx support found and enabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2022-11-08 03:12:26,984] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-11-08 03:12:26,985] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2022-11-08 03:12:27,003] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@9f70c54 (org.apache.zookeeper.server.ServerMetrics)
[2022-11-08 03:12:27,008] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2022-11-08 03:12:27,021] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,021] INFO   ______                  _                                          (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,021] INFO  |___  /                 | |                                         (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,021] INFO     / /    ___     ___   | | __   ___    ___   _ __     ___   _ __   (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO    / /    / _ \   / _ \  | |/ /  / _ \  / _ \ | '_ \   / _ \ | '__| (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO   / /__  | (_) | | (_) | |   <  |  __/ |  __/ | |_) | |  __/ | |    (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO  /_____|  \___/   \___/  |_|\_\  \___|  \___| | .__/   \___| |_| (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO                                               | |                     (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO                                               |_|                     (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,022] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,026] INFO Server environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,026] INFO Server environment:host.name=ip-172-31-50-139.ec2.internal (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,027] INFO Server environment:java.version=1.8.0_342 (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,027] INFO Server environment:java.vendor=Amazon.com Inc. (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,027] INFO Server environment:java.home=/usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre (org.apache.zookeeper.server.ZooKeeperServer)
[2022-11-08 03:12:27,027] 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:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/argparse4j-0.7.0.jar:/hom
e/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/kafka_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/hado
op/kafka_2.13-3.0.0/bin/../libs/connect-basic-auth-extension-3.0.0.jar:/home/hadoop/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:/ho
me/hadoop/kafka_2.13-3.0.0/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/kafka_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/../libs/jackson-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-core-2.12.3.jar:/home/h
adoop/kafka_2.13-3.0.0/bin/../libs/jackson-databind-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/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-b
ase-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/hadoo
p/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/jakarta.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/../libs/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.serv
let-api-3.1.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/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-common-2.3
4.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/../libs/jersey-hk2-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jers
ey-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/jetty-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.v2
0210629.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:/home/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/k
```

## Command- bin/kafka-server-start.sh config/server.properties &



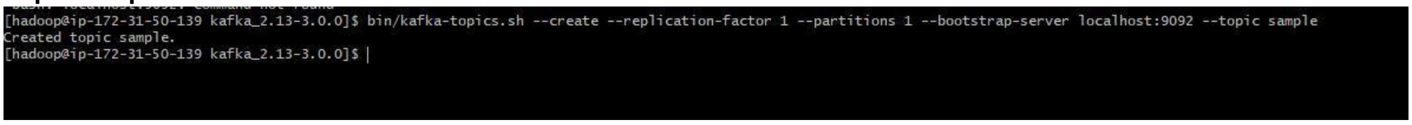
```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/kafka-server-start.sh config/server.properties &
[2] 22496
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ [2022-11-08 03:14:17,464] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2022-11-08 03:14:17,877] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2022-11-08 03:14:18,007] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2022-11-08 03:14:18,024] INFO starting (kafka.server.KafkaServer)
[2022-11-08 03:14:18,025] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2022-11-08 03:14:18,043] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2022-11-08 03:14:18,050] INFO Client environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08/2021 16:35 GMT (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,050] INFO Client environment:host.name=ip-172-31-50-139.ec2.internal (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,050] INFO Client environment:java.version=1.8.0_342 (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,050] INFO Client environment:java.vendor=Amazon.com Inc. (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,050] INFO Client environment:java.home=/usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,050] 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:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/argparse4j-0.7.0.jar:/hom
e/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/kafka_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/hado
op/kafka_2.13-3.0.0/bin/../libs/connect-basic-auth-extension-3.0.0.jar:/home/hadoop/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:/ho
me/hadoop/kafka_2.13-3.0.0/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/kafka_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/../libs/jackson-annotations-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jackson-core-2.12.3.jar:/home/h
adoop/kafka_2.13-3.0.0/bin/../libs/jackson-databind-2.12.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/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-b
ase-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/hadoo
p/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/jakarta.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/../libs/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.serv
let-api-3.1.0.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/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-common-2.3
4.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/../libs/jersey-hk2-2.34.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/jers
ey-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/jetty-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.v2
0210629.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:/home/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/k
afka-shell-3.0.0.jar:/home/hadoop/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/kafka_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/met
rics-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/hadoop/kafka_2.13-3.0.0/bin/../libs/netty-transport-native-epoll-4.1.62.Fina
l.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
/../libs/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/snappy-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/../l
ibs/zookeeper-3.6.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/zookeeper-jute-3.6.3.jar:/home/hadoop/kafka_2.13-3.0.0/bin/../libs/zstd-jni-1.5.0-2.jar (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/usr/lib (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:java.io.tmpdir=/tmp (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:java.compiler=<NA> (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:os.name=Linux (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:os.arch=amd64 (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:os.version=4.14.294-220.533.amzn2.x86_64 (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,051] INFO Client environment:user.name=hadoop (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,052] INFO Client environment:user.home=/home/hadoop (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,052] INFO Client environment:user.dir=/home/hadoop/kafka_2.13-3.0.0 (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,052] INFO Client environment:os.memory.free=1011MB (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,052] INFO Client environment:os.memory.max=1024MB (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,052] INFO Client environment:os.memory.total=1024MB (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,055] INFO Initiating client connection, connectString=localhost:2181 sessionTimeout=18000 watcher=kafka.zookeeper.ZooKeeperClient$ZooKeeperClientWatcher$@515c6049 (org.apache.zookeeper.ZooKeeper)
[2022-11-08 03:14:18,061] INFO jute.maxbuffer value is 4194304 Bytes (org.apache.zookeeper.ClientCnxnSocket)
[2022-11-08 03:14:18,067] INFO zookeeper.request.timeout value is 0. feature enabled=false (org.apache.zookeeper.ClientCnxn)
[2022-11-08 03:14:18,069] INFO [ZooKeeperClient Kafka server] Waiting until connected. (kafka.zookeeper.ZooKeeperClient)
[2022-11-08 03:14:18,074] INFO Opening socket connection to server localhost/127.0.0.1:2181. (org.apache.zookeeper.ClientCnxn)
[2022-11-08 03:14:18,074] INFO SASL config status: Will not attempt to authenticate using SASL (unknown error) (org.apache.zookeeper.ClientCnxn)
[2022-11-08 03:14:18,079] INFO Socket connection established, initiating session, client: /127.0.0.1:41228, server: localhost/127.0.0.1:2181 (org.apache.zookeeper.ClientCnxn)
[2022-11-08 03:14:18,092] INFO Creating new log file: log.1 (org.apache.zookeeper.server.persistence.FileTxnLog)
[2022-11-08 03:14:18,104] INFO Session establishment complete on server localhost/127.0.0.1:2181, session id = 0x100000f6b580000, negotiated timeout = 18000 (org.apache.zookeeper.ClientCnxn)
[2022-11-08 03:14:18,108] INFO [ZooKeeperClient Kafka server] Connected. (kafka.zookeeper.ZooKeeperClient)
[2022-11-08 03:14:18,257] INFO [feature-zk-node-event-process-thread]: Starting (kafka.server.FinalizedFeatureChangeListener$ChangeNotificationProcessorThread)
[2022-11-08 03:14:18,273] INFO Feature ZK node at path: /feature does not exist (kafka.server.FinalizedFeatureChangeListener)
[2022-11-08 03:14:18,273] INFO Cleared cache (kafka.server.FinalizedFeatureCache)
[2022-11-08 03:14:18,479] INFO Cluster ID = c3EMpHo9RcOwCniJMCVWKw (kafka.server.KafkaServer)
[2022-11-08 03:14:18,485] WARN No meta.properties file under dir /tmp/kafka-logs/meta.properties (kafka.server.BrokerMetadataCheckpoint)
[2022-11-08 03:14:18,545] INFO KafkaConfig values:
        advertised.listeners = null
        alter.config.policy.class.name = null
        alter.log.dirs.replication.quota.window.num = 11
```

## At Producer Terminal
## 2nd EMR Connection



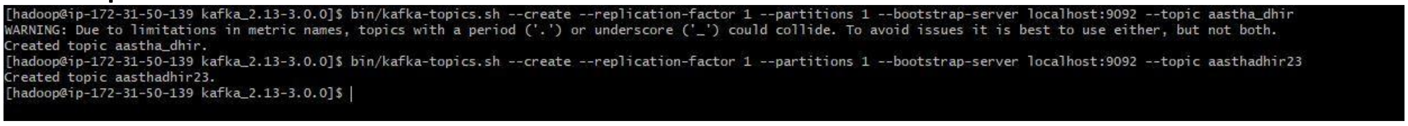
```
aasth@LAPTOP-HJTR6HMR MINGW64 ~
$ ssh -i Downloads/emr-key-pair.pem hadoop@ec2-54-209-245-237.compute-1.amazonaw
s.com
Last login: Tue Nov  8 03:01:41 2022

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
7 package(s) needed for security, out of 10 available
Run "sudo yum update" to apply all updates.

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-172-31-50-139 ~]$ |
```

## Topic Sample Created



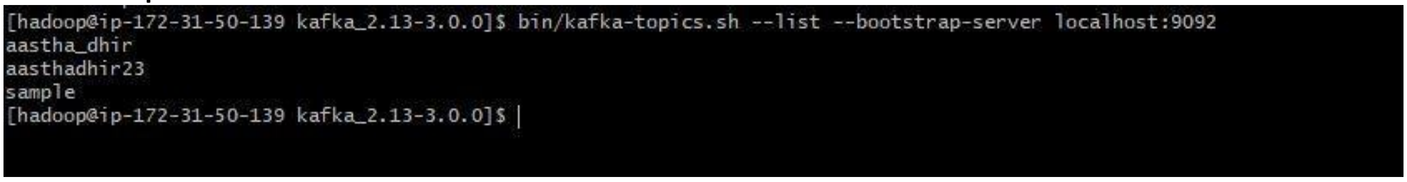
```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic sample
Created topic sample.
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ |
```

## More Topics created



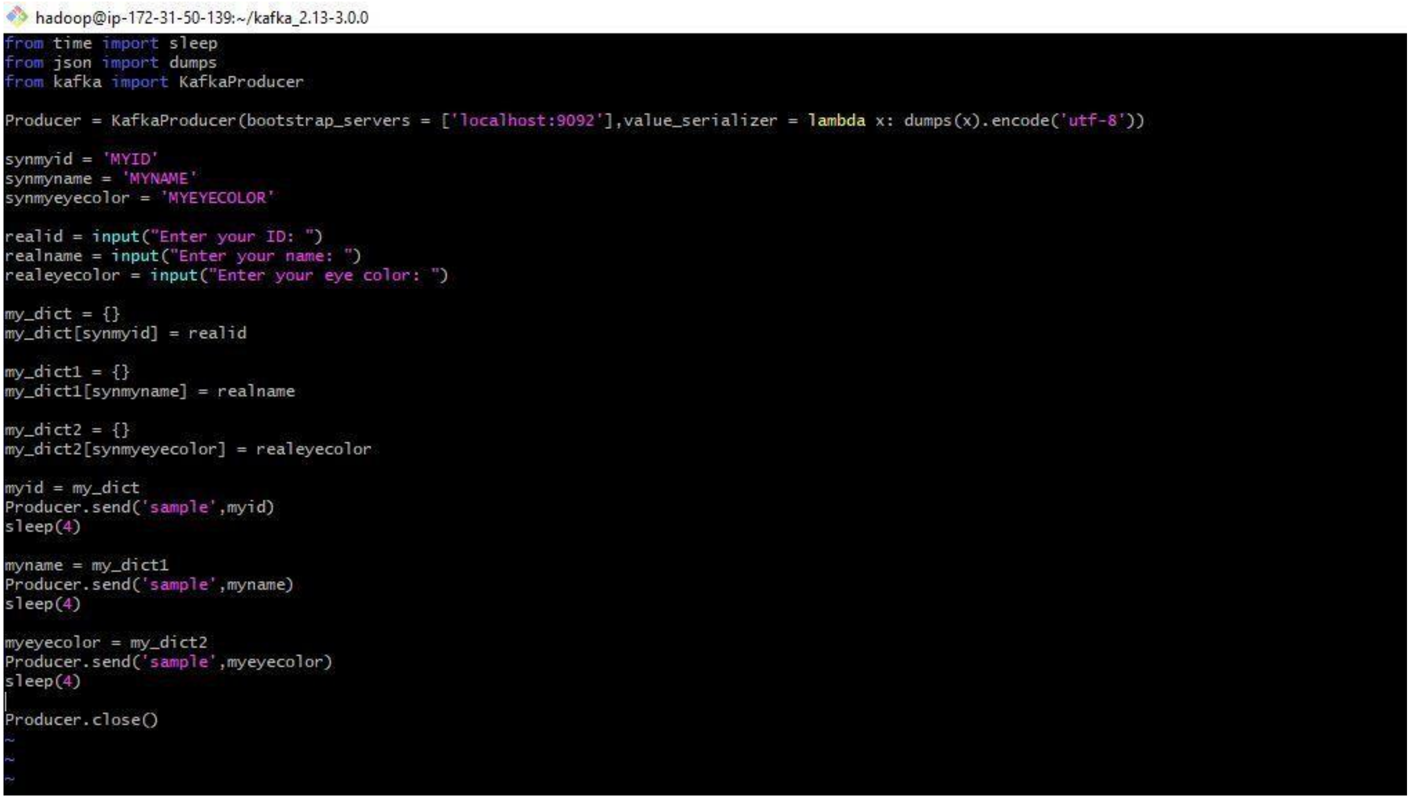
```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic aastha_dhir
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid issues it is best to use either, but not both.
Created topic aastha_dhir.
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --create --replication-factor 1 --partitions 1 --bootstrap-server localhost:9092 --topic aasthadhir23
Created topic aasthadhir23.
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ |
```

## List of all topics created



```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ bin/kafka-topics.sh --list --bootstrap-server localhost:9092
aastha_dhir
aasthadhir23
sample
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ |
```

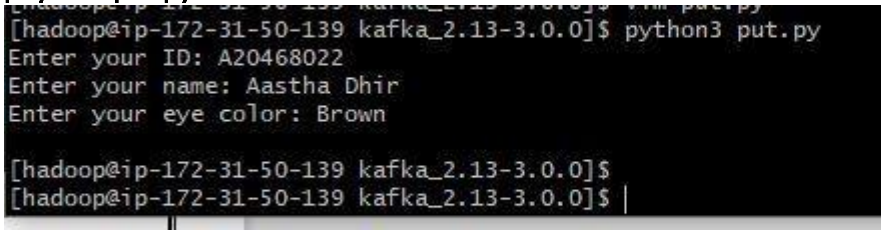
## Producer Terminal
## vim put.py



hadoop@ip-172-31-50-139:~/kafka_2.13-3.0.0

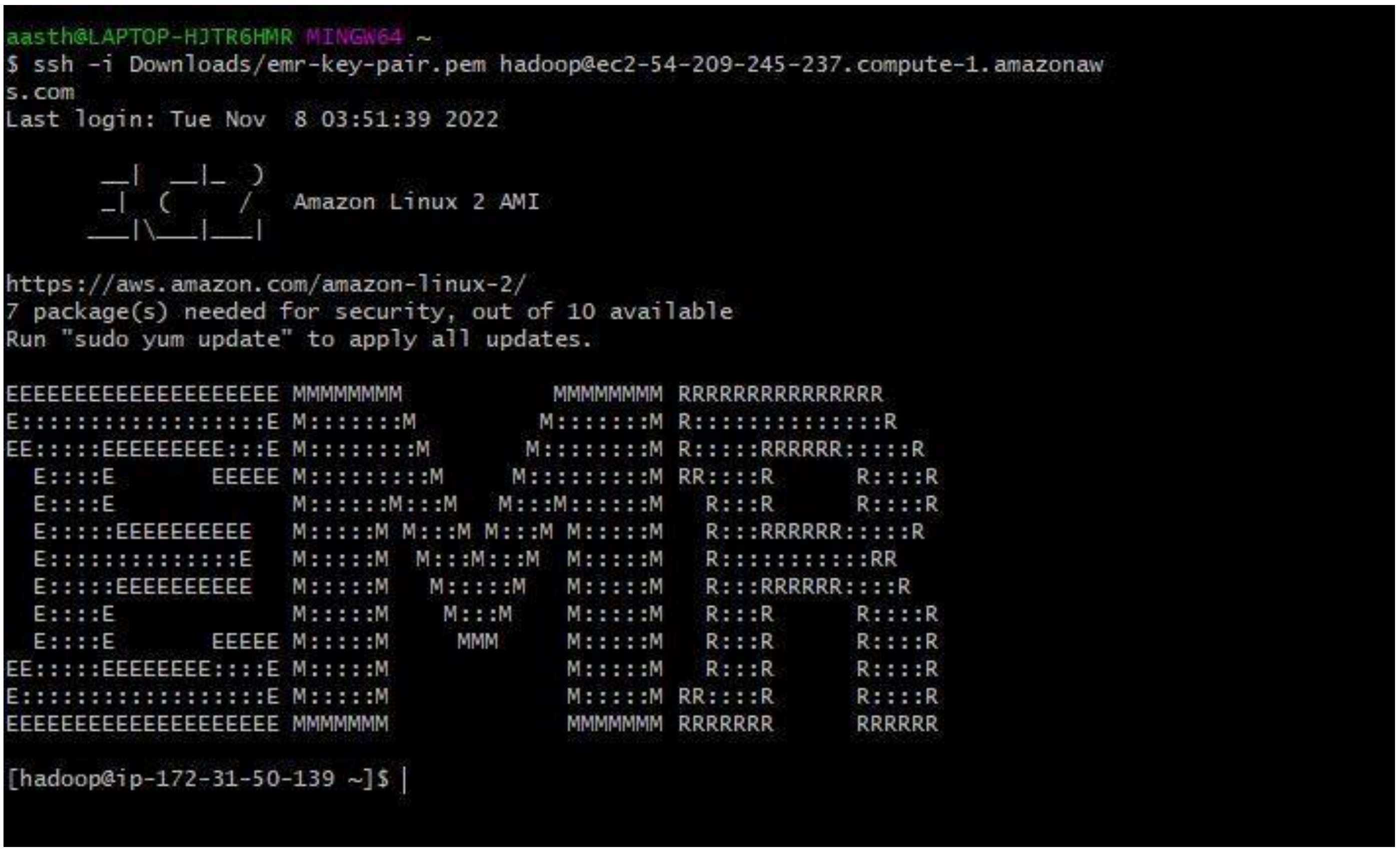
```python
from time import sleep
from json import dumps
from kafka import KafkaProducer

Producer = KafkaProducer(bootstrap_servers = ['localhost:9092'],value_serializer = lambda x: dumps(x).encode('utf-8'))

synmyid = 'MYID'
synmyname = 'MYNAME'
synmyeyecolor = 'MYEYECOLOR'

realid = input("Enter your ID: ")
realname = input("Enter your name: ")
realeyecolor = input("Enter your eye color: ")

my_dict = {}
my_dict[synmyid] = realid

my_dict1 = {}
my_dict1[synmyname] = realname

my_dict2 = {}
my_dict2[synmyeyecolor] = realeyecolor

myid = my_dict
Producer.send('sample',myid)
sleep(4)

myname = my_dict1
Producer.send('sample',myname)
sleep(4)

myeyecolor = my_dict2
Producer.send('sample',myeyecolor)
sleep(4)
|
Producer.close()
~
~
~
```

## phython put.py



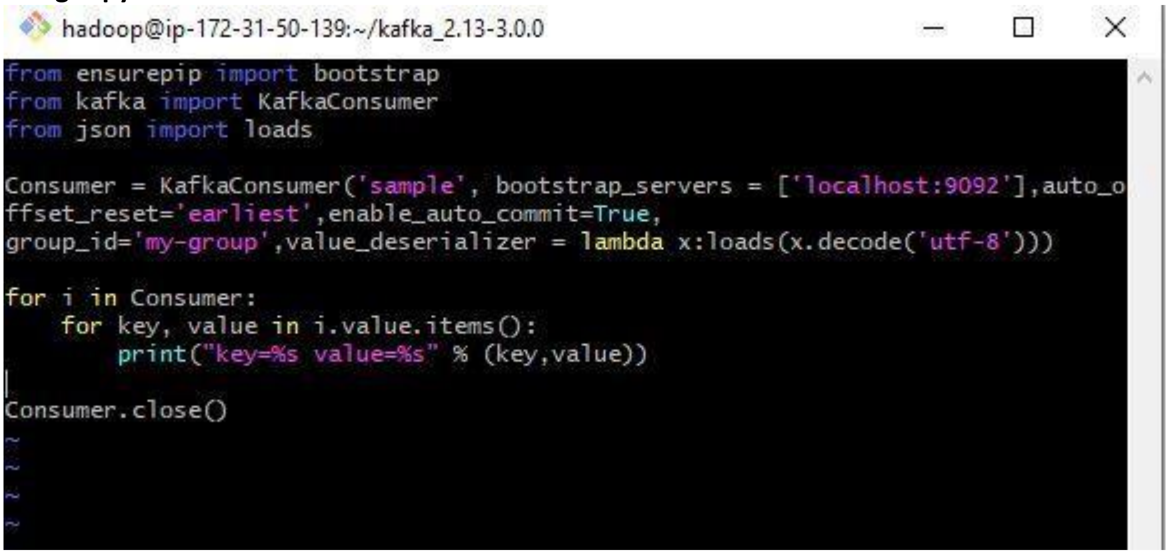
```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ python3 put.py
Enter your ID: A20468022
Enter your name: Aastha Dhir
Enter your eye color: Brown

[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ |
```

## At Consumer Terminal
## 3rd EMR Connection



```
aasth@LAPTOP-HJTR6HMR MINGW64 ~
$ ssh -i Downloads/emr-key-pair.pem hadoop@ec2-54-209-245-237.compute-1.amazonaw
s.com
Last login: Tue Nov  8 03:51:39 2022

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
7 package(s) needed for security, out of 10 available
Run "sudo yum update" to apply all updates.

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-172-31-50-139 ~]$ |
```

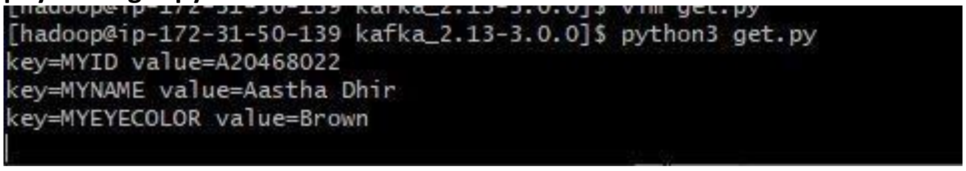
**Consumer Terminal**
**vim get.py**



hadoop@ip-172-31-50-139:~/kafka_2.13-3.0.0

```
from ensurepip import bootstrap
from kafka import KafkaConsumer
from json import loads

Consumer = KafkaConsumer('sample', bootstrap_servers = ['localhost:9092'],auto_o
ffset_reset='earliest',enable_auto_commit=True,
group_id='my-group',value_deserializer = lambda x:loads(x.decode('utf-8')))

for i in Consumer:
    for key, value in i.value.items():
        print("key=%s value=%s" % (key,value))

Consumer.close()
~
~
~
~
```

**phython3 get.py**



```
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ vim get.py
[hadoop@ip-172-31-50-139 kafka_2.13-3.0.0]$ python3 get.py
key=MYID value=A20468022
key=MYNAME value=Aastha Dhir
key=MYEYECOLOR value=Brown
|
```

**Submitted By:-**
**Aastha Dhir**
**CWID- A20468022**
**adhir2@hawk.iit.edu**