Chirag Khandhar

A20438926 | CSP 554 - Big Data Technologies | Fall 2020 | Assignment 9

Exercise 1:

a) What is the Kappa architecture and how does it differ from the lambda architecture?
  - The basic idea of Kappa Architecture is to **not periodically** recompute all data in the batch layer, but to do all computation in the stream processing system alone and only perform recomputation when the business logic changes by replaying historical data.
  - As against in Lambda Architecture, data is stored in a persistence layer like HDFS from which it is ingested and processed by the batch layer **periodically**, while the speed layer handles the portion of the data that has not-yet been processed by the batch layer, and the serving layer consolidates both by merging the output of the batch and the speed layer.

b) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?
  - Pure streaming processing systems like Storm and Samza provide a very low latency and relatively high per-item cost.
  - Pure Streaming processing systems provides low latency, but does not offer ordering guarantees
  - Micro-batch processing systems like Trident, groups tuples into batches to relax the one-at-a-time processing model in favour of increased throughput.
  - Micro-batch processing systems introduces batch-size as a parameter to increase throughput at the cost of latency.
  - Trident provides its own API for fault-tolerant state management with exactly-once processing semantics.

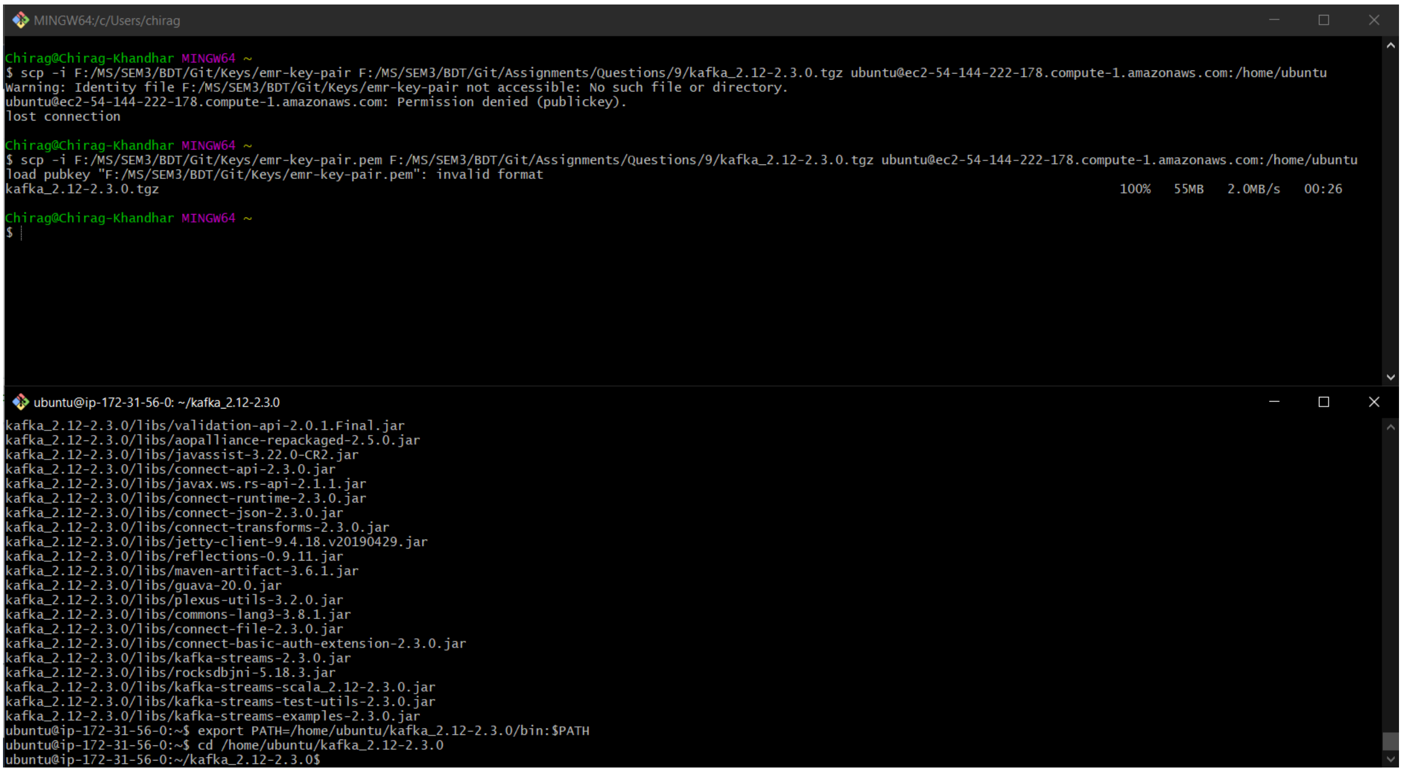
c) In few sentences describe the data processing pipeline in Storm.
  - A data pipeline or application in Storm is called a topology.
  - The nodes that ingest data and thus initiate the data flow in the topology are called **spouts** and emit tuples to the nodes downstream which are called **bolts** and do processing, write data to external storage and may send tuples further downstream themselves.
  - Storm comes with several groupings that control data flow between nodes.
  - By default, Storm distributes spouts and bolts across the nodes in the cluster in a round-robin fashion.
  - The application logic is encapsulated in a manual definition of data flow and the spouts and bolts which implement interfaces to define their behaviour during start-up and on data ingestion or receiving a tuple, respectively.
  - Storm does not provide any guarantee on the order in which tuples are processed.

- It does provide the option of at-least-once processing through an acknowledgement feature.

d) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?

- Spark Streaming shifts Spark's batch-processing approach towards real-time requirements by chunking the stream of incoming data items into small batches, transforming them into RDDs and processing them as usual. It further takes care of data flow and distribution automatically. Data is ingested and transformed into a sequence of RDDs which is called DStream (discretised stream) before processing through workers. All RDDs in a DStream are processed in order, whereas data items inside an RDD are processed in parallel without any ordering guarantees.

Exercise 2:

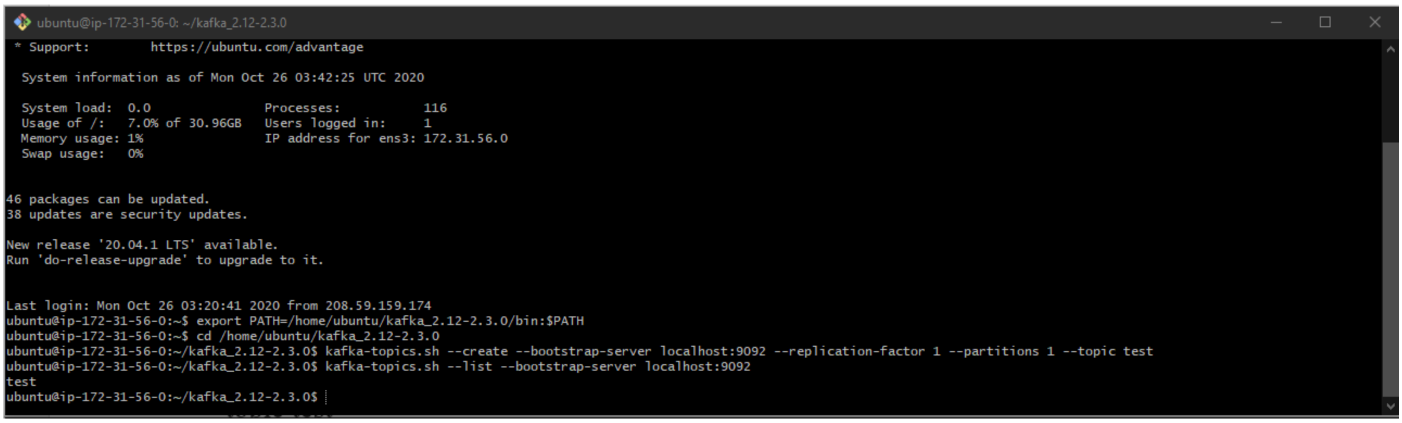


```
Chirag@Chirag-Khandhar MINGW64 ~
$ scp -i F:/MS/SEM3/BDT/Git/Keys/emr-key-pair F:/MS/SEM3/BDT/Git/Assignments/Questions/9/kafka_2.12-2.3.0.tgz ubuntu@ec2-54-144-222-178.compute-1.amazonaws.com:/home/ubuntu
Warning: Identity file F:/MS/SEM3/BDT/Git/Keys/emr-key-pair not accessible: No such file or directory.
ubuntu@ec2-54-144-222-178.compute-1.amazonaws.com: Permission denied (publickey).
lost connection

Chirag@Chirag-Khandhar MINGW64 ~
$ scp -i F:/MS/SEM3/BDT/Git/Keys/emr-key-pair.pem F:/MS/SEM3/BDT/Git/Assignments/Questions/9/kafka_2.12-2.3.0.tgz ubuntu@ec2-54-144-222-178.compute-1.amazonaws.com:/home/ubuntu
load pubkey "F:/MS/SEM3/BDT/Git/Keys/emr-key-pair.pem": invalid format
kafka_2.12-2.3.0.tgz                                                                                                              100%   55MB   2.0MB/s   00:26

Chirag@Chirag-Khandhar MINGW64 ~
$
```

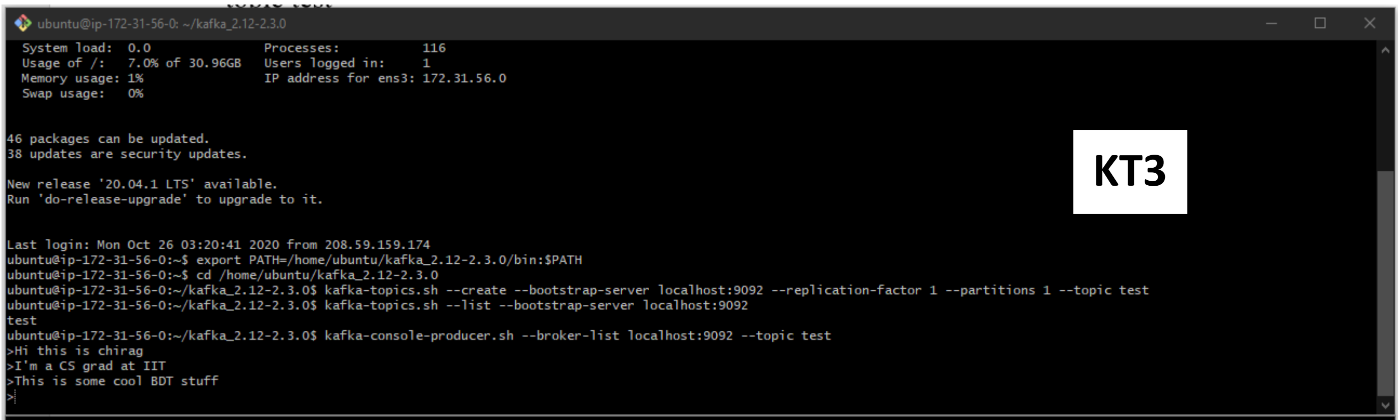
```
kafka_2.12-2.3.0/libs/validation-api-2.0.1.Final.jar
kafka_2.12-2.3.0/libs/aopalliance-repackaged-2.5.0.jar
kafka_2.12-2.3.0/libs/javassist-3.22.0-CR2.jar
kafka_2.12-2.3.0/libs/connect-api-2.3.0.jar
kafka_2.12-2.3.0/libs/javax.ws.rs-api-2.1.1.jar
kafka_2.12-2.3.0/libs/connect-runtime-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-json-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-transforms-2.3.0.jar
kafka_2.12-2.3.0/libs/jetty-client-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/reflections-0.9.11.jar
kafka_2.12-2.3.0/libs/maven-artifact-3.6.1.jar
kafka_2.12-2.3.0/libs/guava-20.0.jar
kafka_2.12-2.3.0/libs/plexus-utils-3.2.0.jar
kafka_2.12-2.3.0/libs/commons-lang3-3.8.1.jar
kafka_2.12-2.3.0/libs/connect-file-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-basic-auth-extension-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-2.3.0.jar
kafka_2.12-2.3.0/libs/rocksdbjni-5.18.3.jar
kafka_2.12-2.3.0/libs/kafka-streams-scala_2.12-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-test-utils-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-examples-2.3.0.jar
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$
```

Section 4: Working with Kafka

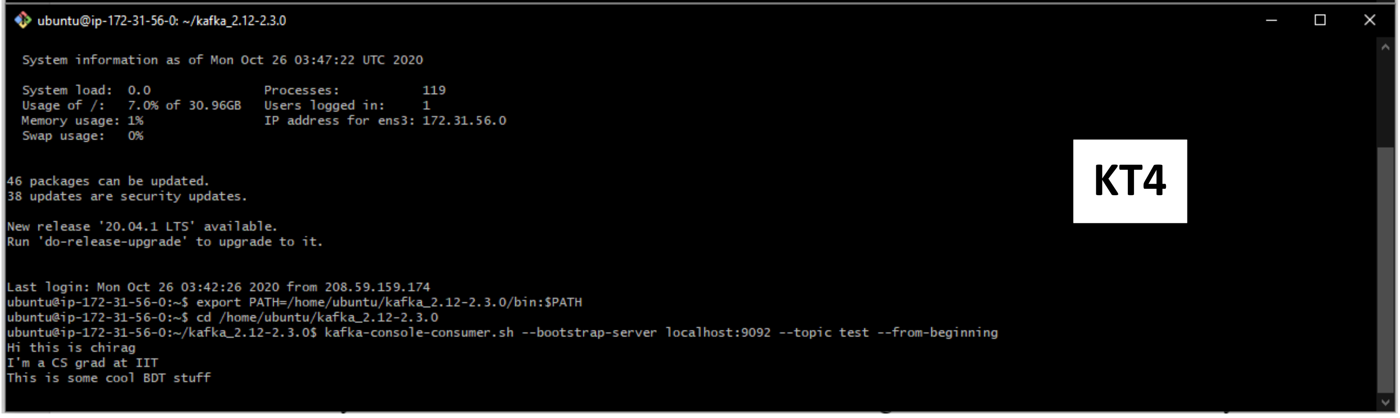
Step 2 Output:



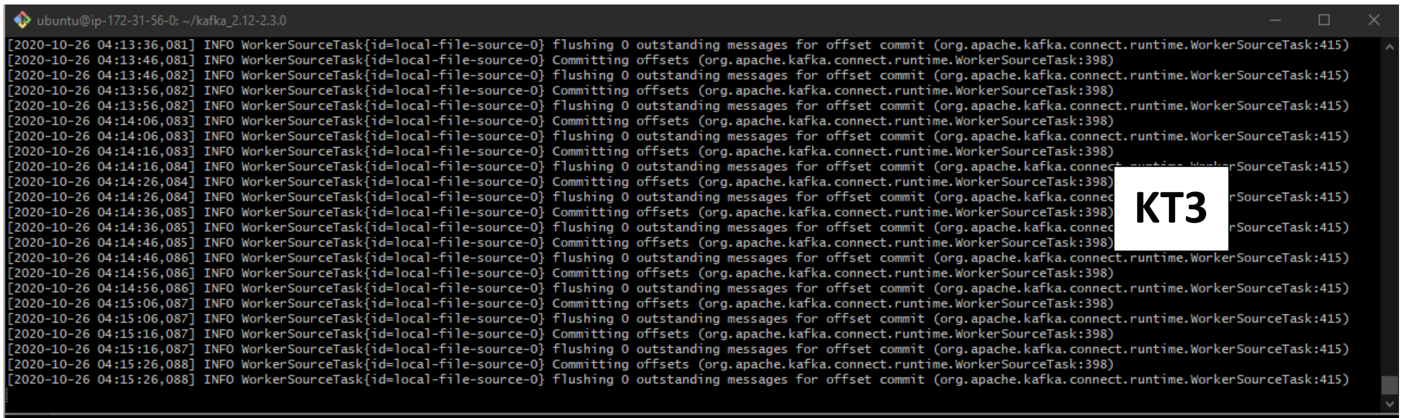
```
 * Support:        https://ubuntu.com/advantage

  System information as of Mon Oct 26 03:42:25 UTC 2020

  System load:  0.0               Processes:           116
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.56.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:20:41 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic test
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$
```



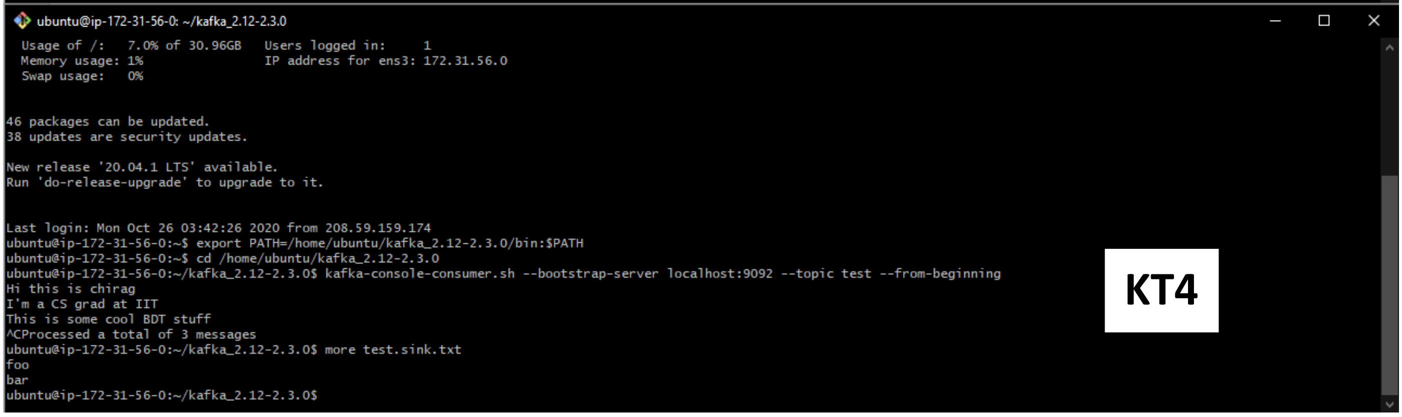
```
  System load:  0.0              Processes:           116
  Usage of /:   7.0% of 30.96GB  Users logged in:     1
  Memory usage: 1%               IP address for ens3: 172.31.56.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:20:41 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic test
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-producer.sh --broker-list localhost:9092 --topic test
>Hi this is chirag
>I'm a CS grad at IIT
>This is some cool BDT stuff
>
```



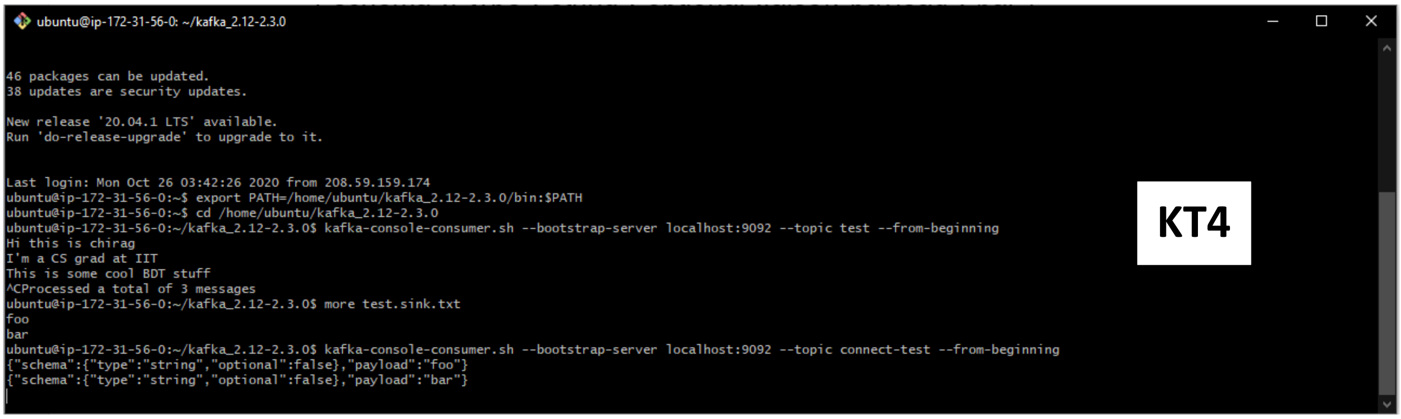
```
  System information as of Mon Oct 26 03:47:22 UTC 2020

  System load:  0.0              Processes:           119
  Usage of /:   7.0% of 30.96GB  Users logged in:     1
  Memory usage: 1%               IP address for ens3: 172.31.56.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:42:26 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning
Hi this is chirag
I'm a CS grad at IIT
This is some cool BDT stuff
```



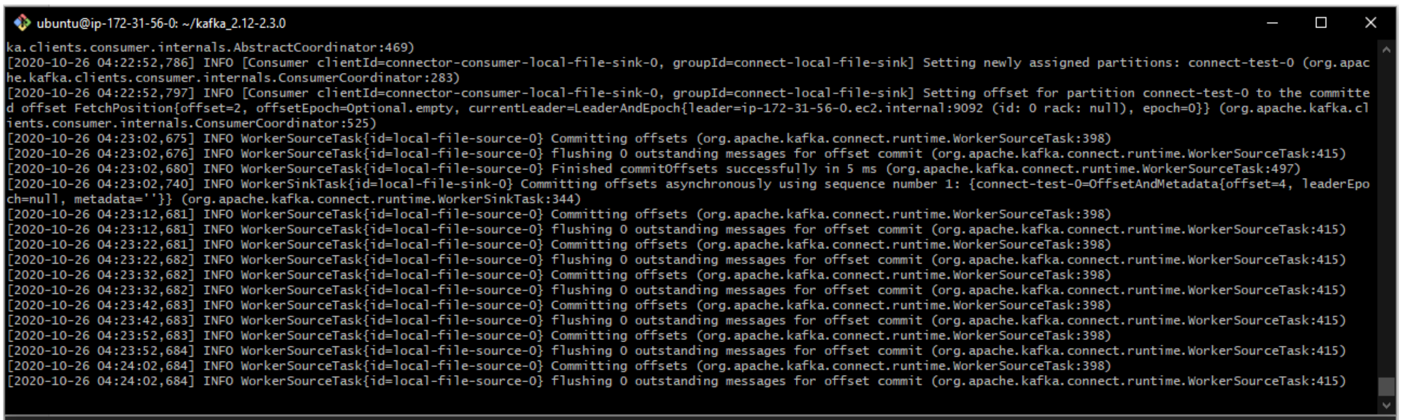
```
[2020-10-26 04:13:36,081] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:13:46,081] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:13:46,082] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:13:56,082] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:13:56,082] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:14:06,083] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:06,083] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:14:16,083] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:16,084] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:14:26,084] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:26,084] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connec rSourceTask:415)
[2020-10-26 04:14:36,085] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:36,085] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connec rSourceTask:415)
[2020-10-26 04:14:46,085] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:46,086] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:14:56,086] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:14:56,086] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:15:06,087] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:15:06,087] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:15:16,087] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:15:16,087] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:15:26,088] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:15:26,088] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
```



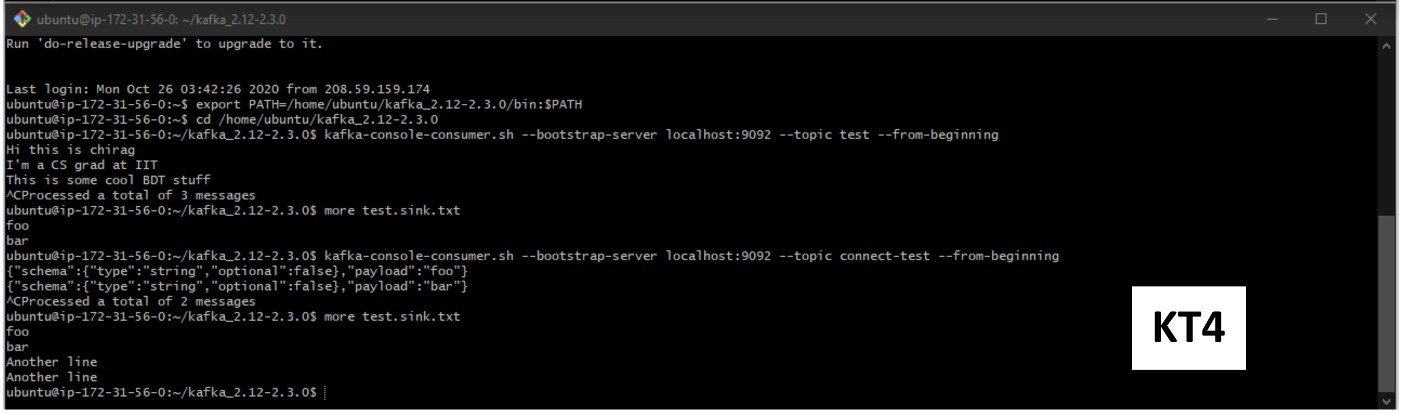
```
  Usage of /:   7.0% of 30.96GB  Users logged in:     1
  Memory usage: 1%               IP address for ens3: 172.31.56.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:42:26 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning
Hi this is chirag
I'm a CS grad at IIT
This is some cool BDT stuff
^CProcessed a total of 3 messages
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$
```



```
46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:42:26 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning
Hi this is chirag
I'm a CS grad at IIT
This is some cool BDT stuff
^CProcessed a total of 3 messages
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic connect-test --from-beginning
{"schema":{"type":"string","optional":false},"payload":"foo"}
{"schema":{"type":"string","optional":false},"payload":"bar"}
```



```
ka.clients.consumer.internals.AbstractCoordinator:469)
[2020-10-26 04:22:52,786] INFO [Consumer clientId=connector-consumer-local-file-sink-0, groupId=connect-local-file-sink] Setting newly assigned partitions: connect-test-0 (org.apac
he.kafka.clients.consumer.internals.ConsumerCoordinator:283)
[2020-10-26 04:22:52,797] INFO [Consumer clientId=connector-consumer-local-file-sink-0, groupId=connect-local-file-sink] Setting offset for partition connect-test-0 to the committe
d offset FetchPosition{offset=2, offsetEpoch=Optional.empty, currentLeader=LeaderAndEpoch{leader=ip-172-31-56-0.ec2.internal:9092 (id: 0 rack: null), epoch=0}} (org.apache.kafka.cl
ients.consumer.internals.ConsumerCoordinator:525)
[2020-10-26 04:23:02,675] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:02,676] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:23:02,680] INFO WorkerSourceTask{id=local-file-source-0} Finished commitOffsets successfully in 5 ms (org.apache.kafka.connect.runtime.WorkerSourceTask:497)
[2020-10-26 04:23:02,740] INFO WorkerSinkTask{id=local-file-sink-0} Committing offsets asynchronously using sequence number 1: {connect-test-0=OffsetAndMetadata{offset=4, leaderEpo
ch=null, metadata=''}} (org.apache.kafka.connect.runtime.WorkerSinkTask:344)
[2020-10-26 04:23:12,681] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:12,681] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:23:22,681] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:22,682] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:23:32,682] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:32,682] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:23:42,683] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:42,683] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:23:52,683] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:23:52,684] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:24:02,684] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.WorkerSourceTask:398)
[2020-10-26 04:24:02,684] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apache.kafka.connect.runtime.WorkerSourceTask:415)
```



```
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:42:26 2020 from 208.59.159.174
ubuntu@ip-172-31-56-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-56-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning
Hi this is chirag
I'm a CS grad at IIT
This is some cool BDT stuff
^CProcessed a total of 3 messages
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic connect-test --from-beginning
{"schema":{"type":"string","optional":false},"payload":"foo"}
{"schema":{"type":"string","optional":false},"payload":"bar"}
^CProcessed a total of 2 messages
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
Another line
Another line
ubuntu@ip-172-31-56-0:~/kafka_2.12-2.3.0$
```