A20424771

Rakshith Churchagundi Amarnath

Assignment 9

Exercise 1:

a) **(1.25 points) What is the Kappa architecture and how does it differ from the lambda architecture?**

Kappa Architecture is a simplification of Lambda Architecture. A Kappa Architecture system is like a Lambda Architecture system with the batch processing system removed. To replace batch processing, data is simply fed through the streaming system quickly.

The basic idea 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 re-computation when the business logic changes by replaying historical data. Whereas, in lambda architecture, periodically re-computation of all data is done in the batch layer.

b) **(1.25 points) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?**

Purely stream-oriented systems provide very low latency and comparatively high per-item cost. Whereas, batch-oriented systems achieve unparalleled resource-efficiency at the expense of latency that's prohibitively large in real-time. The micro-batching strategy is to trade latency against throughput: Groups tuples into lots to relax the one on one processing model in favor of increased throughput, whereas Spark Streaming restricts batch size in an exceedingly native batch processor to cut back latency.

c) **(1.25 points) In few sentences describe the data processing pipeline in Storm.**

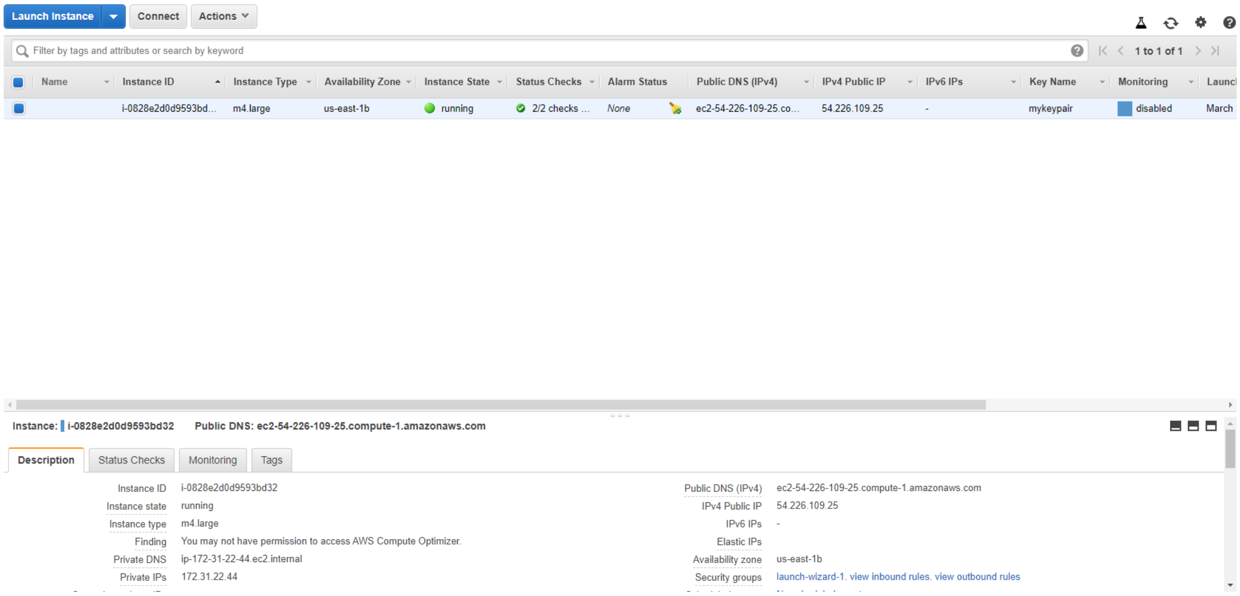
The data processing pipeline in Strom (topology) is a directed graph that represents data flow as directed edges between nodes. The nodes in turn represents the individual processing steps. The data is ingested from the streaming layer then passed between the Storm Components. The node that ingest and initiate the data flow in the topology is termed spouts. The spouts emit tuples to the nodes downstream called bolts. Bolts process the data, write data to storage and even may send tuples further downstream. The ultimate output is downstream to the serving layer.

d) **(1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?**

Spark streaming shift the Spark batch processing approach towards real-time data streams by chunking the stream of incoming data items into small batches, then transforming them into RDDs and processing them as usual.

Extra Credit:

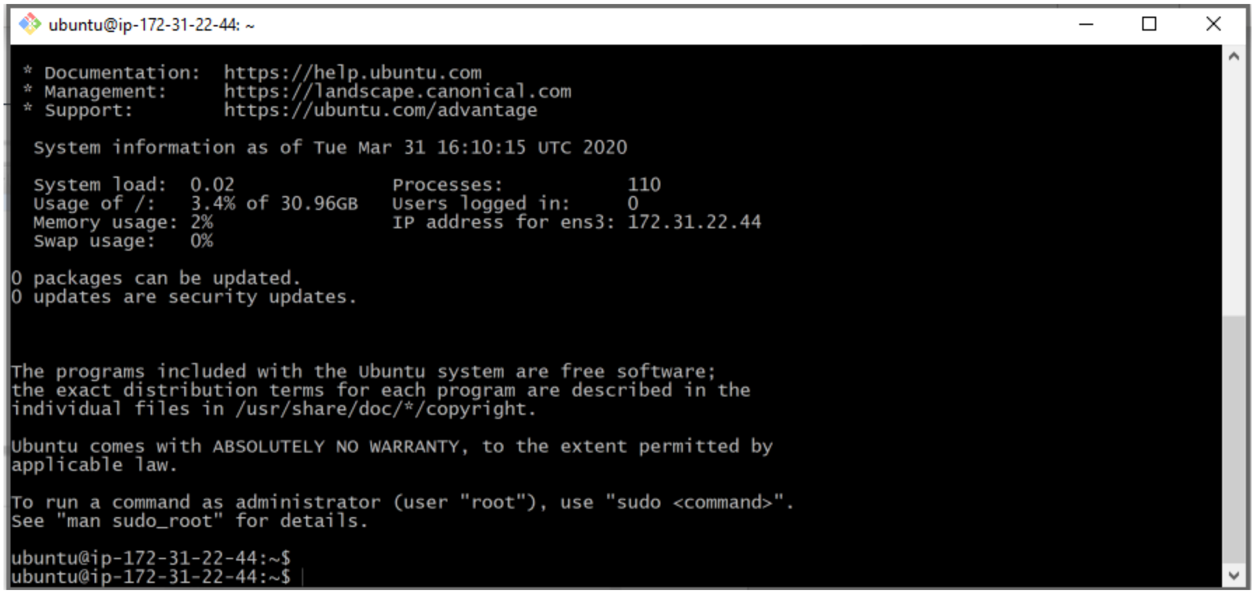
Section 1:



	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs	Key Name	Monitoring	Launc
■		i-0828e2d0d9593bd...	m4.large	us-east-1b	● running	✓ 2/2 checks ...	*None*	ec2-54-226-109-25.co...	54.226.109.25	-	mykeypair	disabled	March

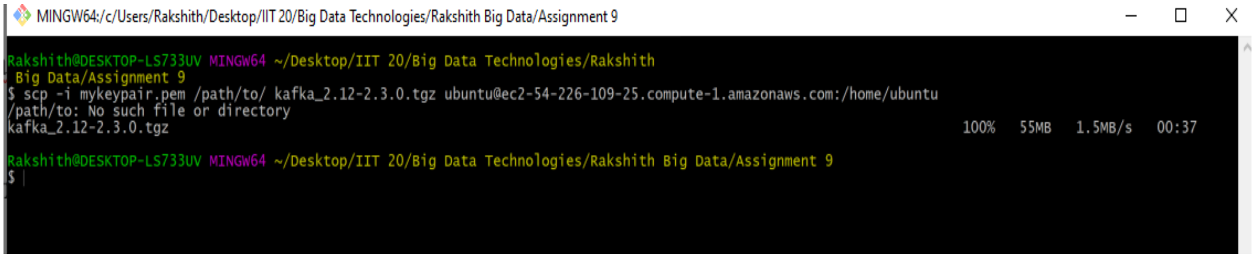
Instance: i-0828e2d0d9593bd32 **Public DNS:** ec2-54-226-109-25.compute-1.amazonaws.com

Instance ID	i-0828e2d0d9593bd32	Public DNS (IPv4)	ec2-54-226-109-25.compute-1.amazonaws.com
Instance state	running	IPv4 Public IP	54.226.109.25
Instance type	m4.large	IPv6 IPs	-
Finding	You may not have permission to access AWS Compute Optimizer.	Elastic IPs	
Private DNS	ip-172-31-22-44.ec2.internal	Availability zone	us-east-1b
Private IPs	172.31.22.44	Security groups	launch-wizard-1. view inbound rules. view outbound rules

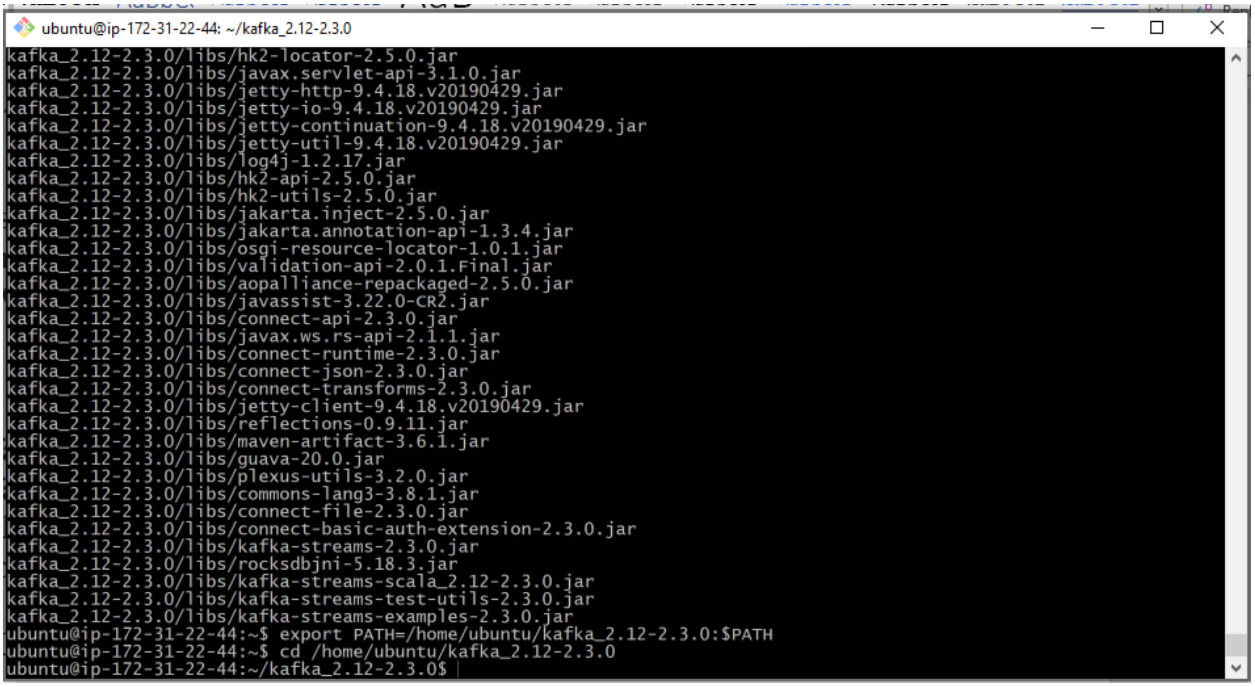

ubuntu@ip-172-31-22-44: ~

```
 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Mar 31 16:10:15 UTC 2020

  System load:  0.02              Processes:           110
  Usage of /:   3.4% of 30.96GB   Users logged in:     0
  Memory usage: 2%                IP address for ens3: 172.31.22.44
  Swap usage:   0%

0 packages can be updated.
0 updates are security updates.



The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-22-44:~$
ubuntu@ip-172-31-22-44:~$
```

Section 2:


MINGW64:/c/Users/Rakshith/Desktop/IIT 20/Big Data Technologies/Rakshith Big Data/Assignment 9

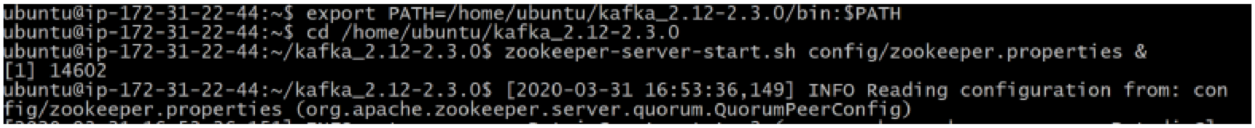
```
Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$ scp -i mykeypair.pem /path/to/ kafka_2.12-2.3.0.tgz ubuntu@ec2-54-226-109-25.compute-1.amazonaws.com:/home/ubuntu
/path/to: No such file or directory
kafka_2.12-2.3.0.tgz                                                                          100%   55MB   1.5MB/s   00:37

Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith Big Data/Assignment 9
$
```



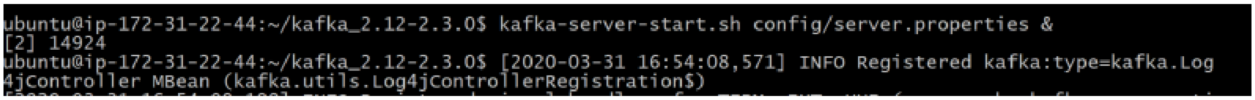
```
kafka_2.12-2.3.0/libs/hk2-locator-2.5.0.jar
kafka_2.12-2.3.0/libs/javax.servlet-api-3.1.0.jar
kafka_2.12-2.3.0/libs/jetty-http-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-io-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-continuation-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-util-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/log4j-1.2.17.jar
kafka_2.12-2.3.0/libs/hk2-api-2.5.0.jar
kafka_2.12-2.3.0/libs/hk2-utils-2.5.0.jar
kafka_2.12-2.3.0/libs/jakarta.inject-2.5.0.jar
kafka_2.12-2.3.0/libs/jakarta.annotation-api-1.3.4.jar
kafka_2.12-2.3.0/libs/osgi-resource-locator-1.0.1.jar
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-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0:$PATH
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
```

A20424771

Section 3:

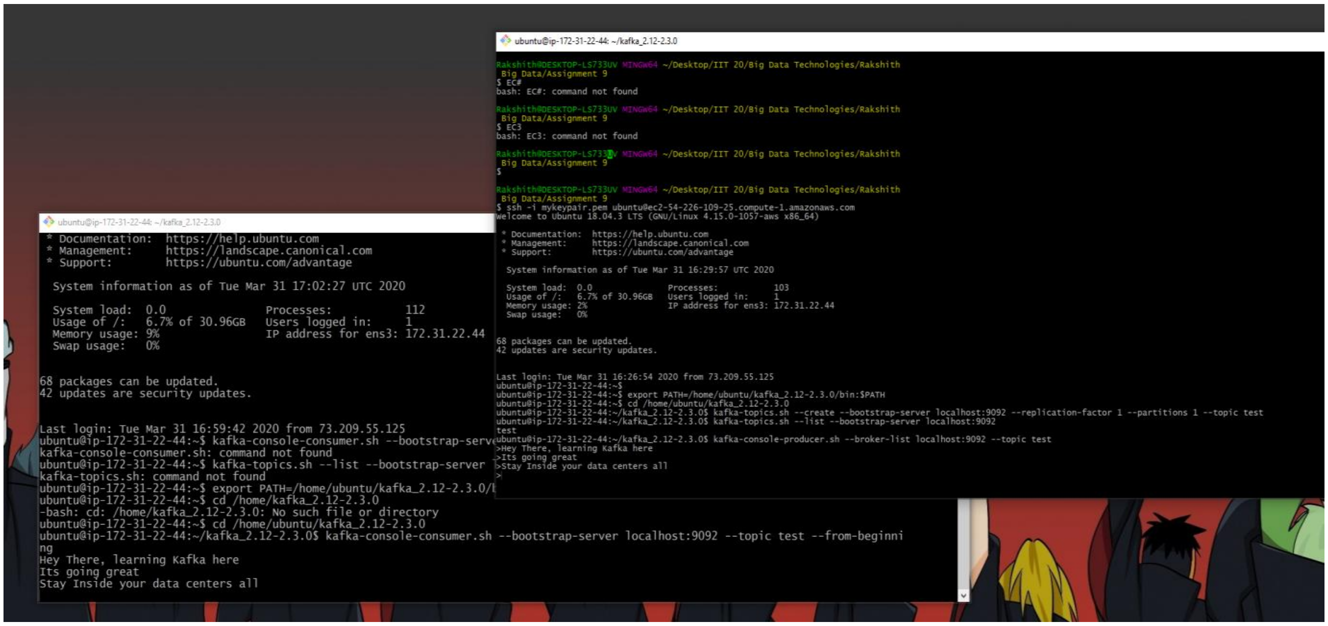


```
ubuntu@ip-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ zookeeper-server-start.sh config/zookeeper.properties &
[1] 14602
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ [2020-03-31 16:53:36,149] INFO Reading configuration from: con
fig/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
```



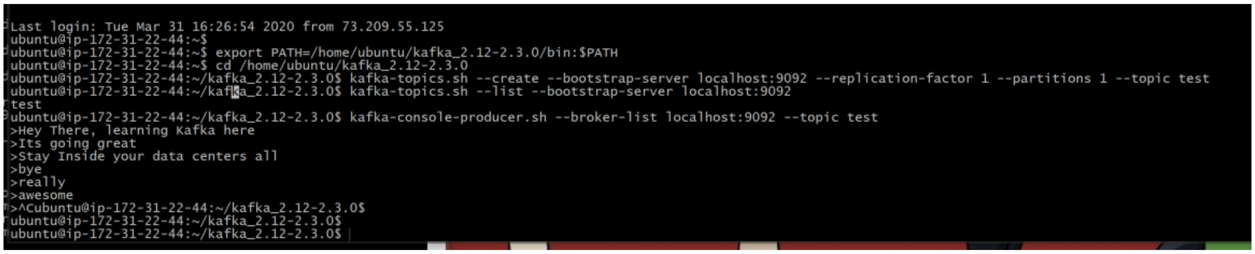
```
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-server-start.sh config/server.properties &
[2] 14924
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ [2020-03-31 16:54:08,571] INFO Registered kafka:type=kafka.Log
4jController MBean (kafka.utils.Log4jControllerRegistration$)
```

Section 4:



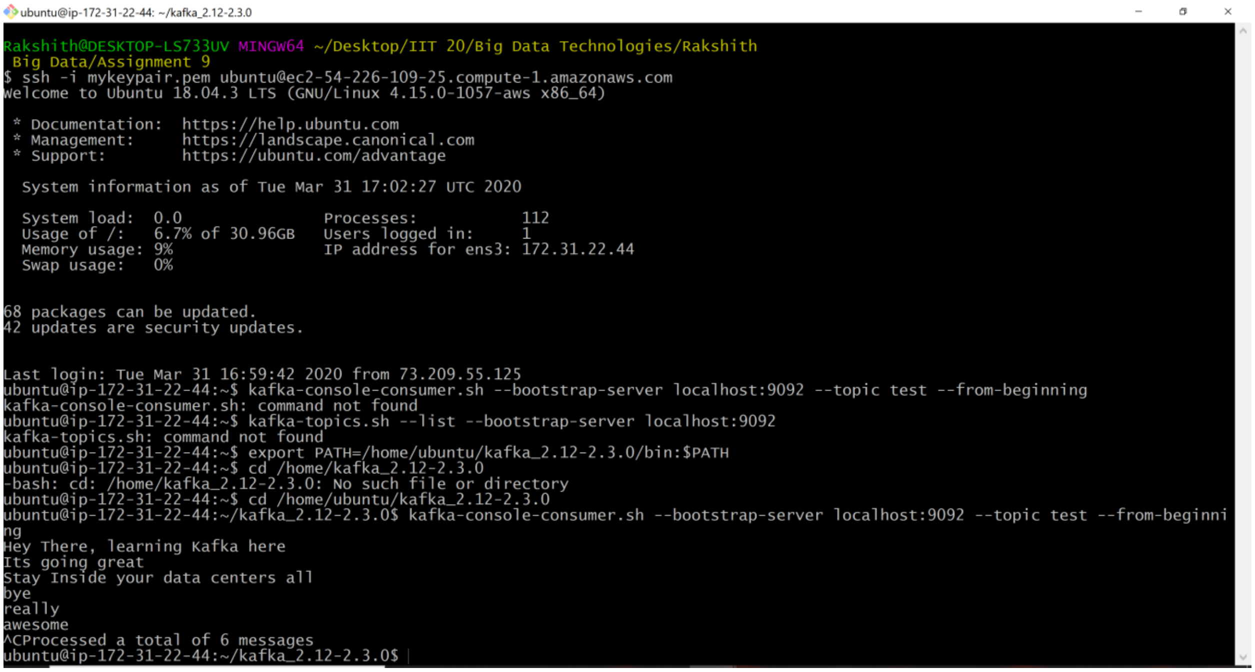
```
Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$ EC#
bash: EC#: command not found

Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$ EC3
bash: EC3: command not found

Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$

Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$ ssh -i mykeypair.pem ubuntu@ec2-54-226-109-25.compute-1.amazonaws.com
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-1057-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Mar 31 16:29:57 UTC 2020

  System load:  0.0               Processes:           103
  Usage of /:   6.7% of 30.96GB   Users logged in:     1
  Memory usage: 2%                IP address for ens3: 172.31.22.44
  Swap usage:   0%


68 packages can be updated.
42 updates are security updates.


Last login: Tue Mar 31 16:26:54 2020 from 73.209.55.125
ubuntu@ip-172-31-22-44:~$
ubuntu@ip-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic test
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-console-producer.sh --broker-list localhost:9092 --topic test
>Hey There, learning Kafka here
>Its going great
>Stay Inside your data centers all
>
```

```
 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Mar 31 17:02:27 UTC 2020

  System load:  0.0               Processes:           112
  Usage of /:   6.7% of 30.96GB   Users logged in:     1
  Memory usage: 9%                IP address for ens3: 172.31.22.44
  Swap usage:   0%


68 packages can be updated.
42 updates are security updates.


Last login: Tue Mar 31 16:59:42 2020 from 73.209.55.125
ubuntu@ip-172-31-22-44:~$ kafka-console-consumer.sh --bootstrap-serve
kafka-console-consumer.sh: command not found
ubuntu@ip-172-31-22-44:~$ kafka-topics.sh --list --bootstrap-server
kafka-topics.sh: command not found
ubuntu@ip-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/b
ubuntu@ip-172-31-22-44:~$ cd /home/kafka_2.12-2.3.0
-bash: cd: /home/kafka_2.12-2.3.0: No such file or directory
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginni
ng
Hey There, learning Kafka here
Its going great
Stay Inside your data centers all
```

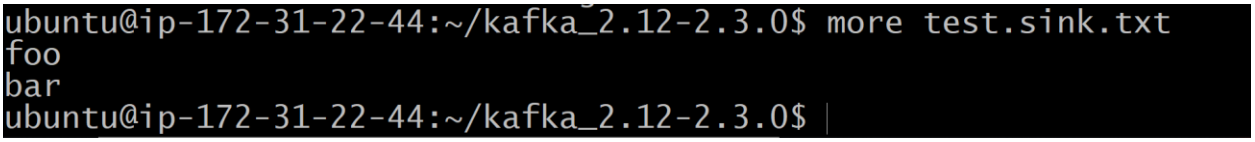


```
Last login: Tue Mar 31 16:26:54 2020 from 73.209.55.125
ubuntu@ip-172-31-22-44:~$
ubuntu@ip-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic test
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-console-producer.sh --broker-list localhost:9092 --topic test
>Hey There, learning Kafka here
>Its going great
>Stay Inside your data centers all
>bye
>really
>awesome
>^Cubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
```

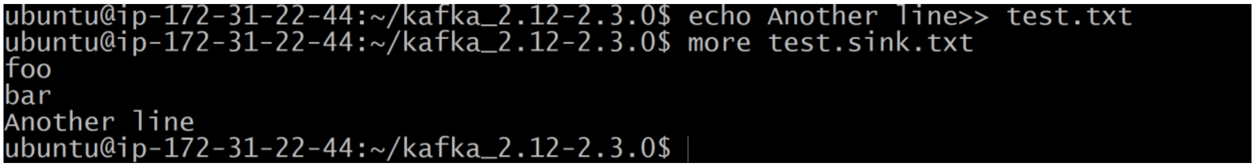
A20424771



```
Rakshith@DESKTOP-LS733UV MINGW64 ~/Desktop/IIT 20/Big Data Technologies/Rakshith
 Big Data/Assignment 9
$ ssh -i mykeypair.pem ubuntu@ec2-54-226-109-25.compute-1.amazonaws.com
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-1057-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Tue Mar 31 17:02:27 UTC 2020

  System load:  0.0                Processes:           112
  Usage of /:   6.7% of 30.96GB    Users logged in:     1
  Memory usage: 9%                 IP address for ens3: 172.31.22.44
  Swap usage:   0%


68 packages can be updated.
42 updates are security updates.


Last login: Tue Mar 31 16:59:42 2020 from 73.209.55.125
ubuntu@ip-172-31-22-44:~$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning
kafka-console-consumer.sh: command not found
ubuntu@ip-172-31-22-44:~$ kafka-topics.sh --list --bootstrap-server localhost:9092
kafka-topics.sh: command not found
ubuntu@ip-172-31-22-44:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-22-44:~$ cd /home/kafka_2.12-2.3.0
-bash: cd: /home/kafka_2.12-2.3.0: No such file or directory
ubuntu@ip-172-31-22-44:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginni
ng
Hey There, learning Kafka here
Its going great
Stay Inside your data centers all
bye
really
awesome
^CProcessed a total of 6 messages
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
```

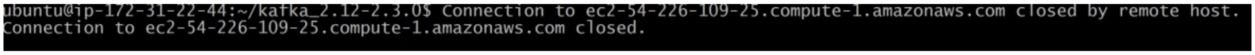


```
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
```



```
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ echo Another line>> test.txt
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
Another line
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$
```

Terminating:



```
ubuntu@ip-172-31-22-44:~/kafka_2.12-2.3.0$ Connection to ec2-54-226-109-25.compute-1.amazonaws.com closed by remote host.
Connection to ec2-54-226-109-25.compute-1.amazonaws.com closed.
```