Rakshith Churchagundi Amarnath

Assignment 9

Exercise 1:

1. **(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.

1. **(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.

1. **(1.25 points) In few sentences describe the data processing pipeline in Storm.**

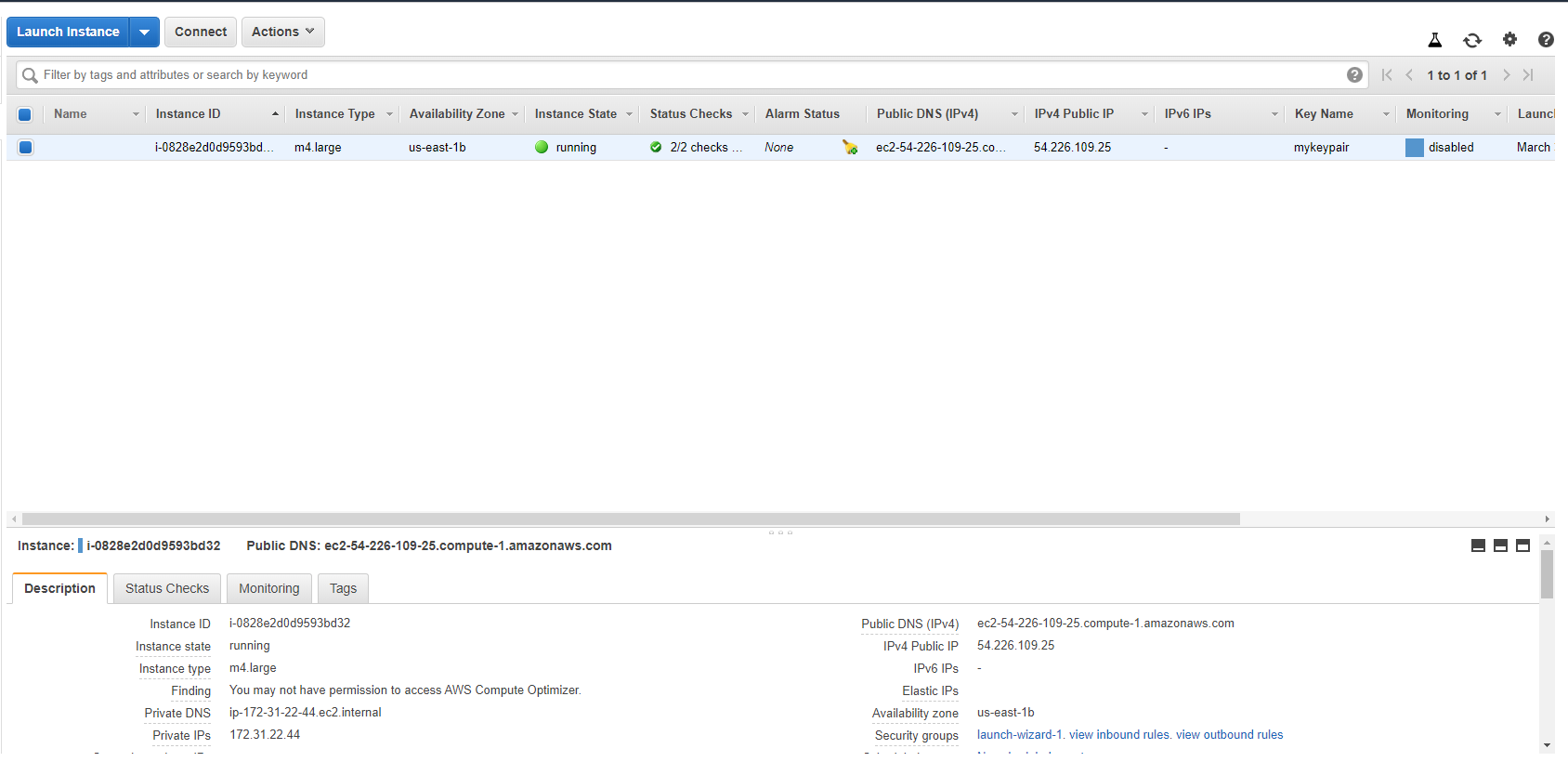
The data processing pipeline in Strom (topology) is a directed graph that represents data flow as directed edges between nodes. The nodes in turn represents the individual processing steps. The data is ingested from the streaming layer then passed between the Storm Components. The node that ingest and 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.

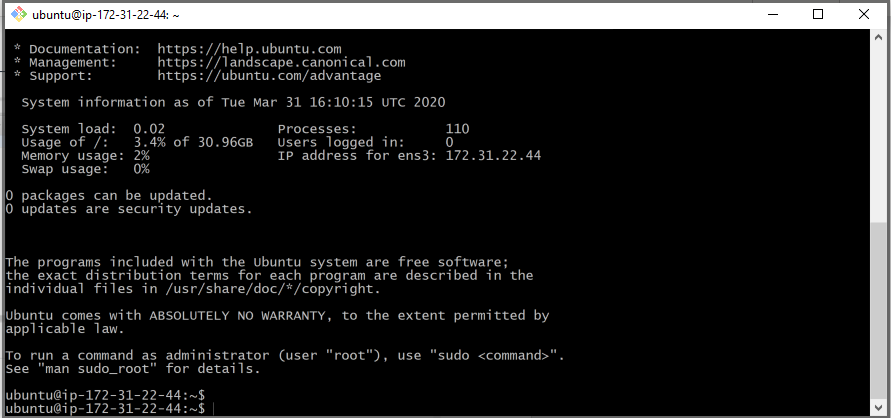
1. **(1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?**

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.

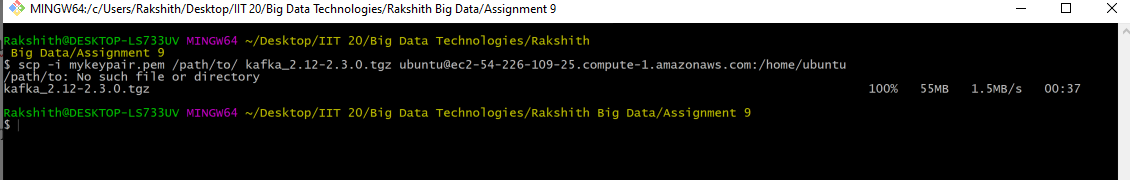
**Extra Credit:**

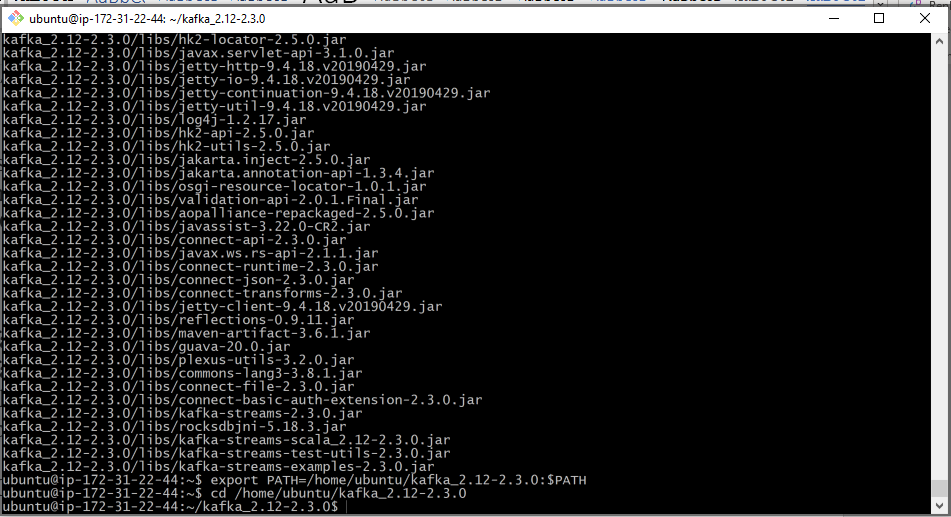
**Section 1:**



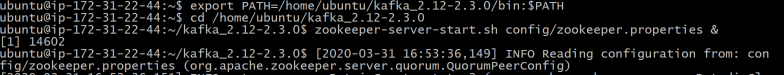


**Section 2:**



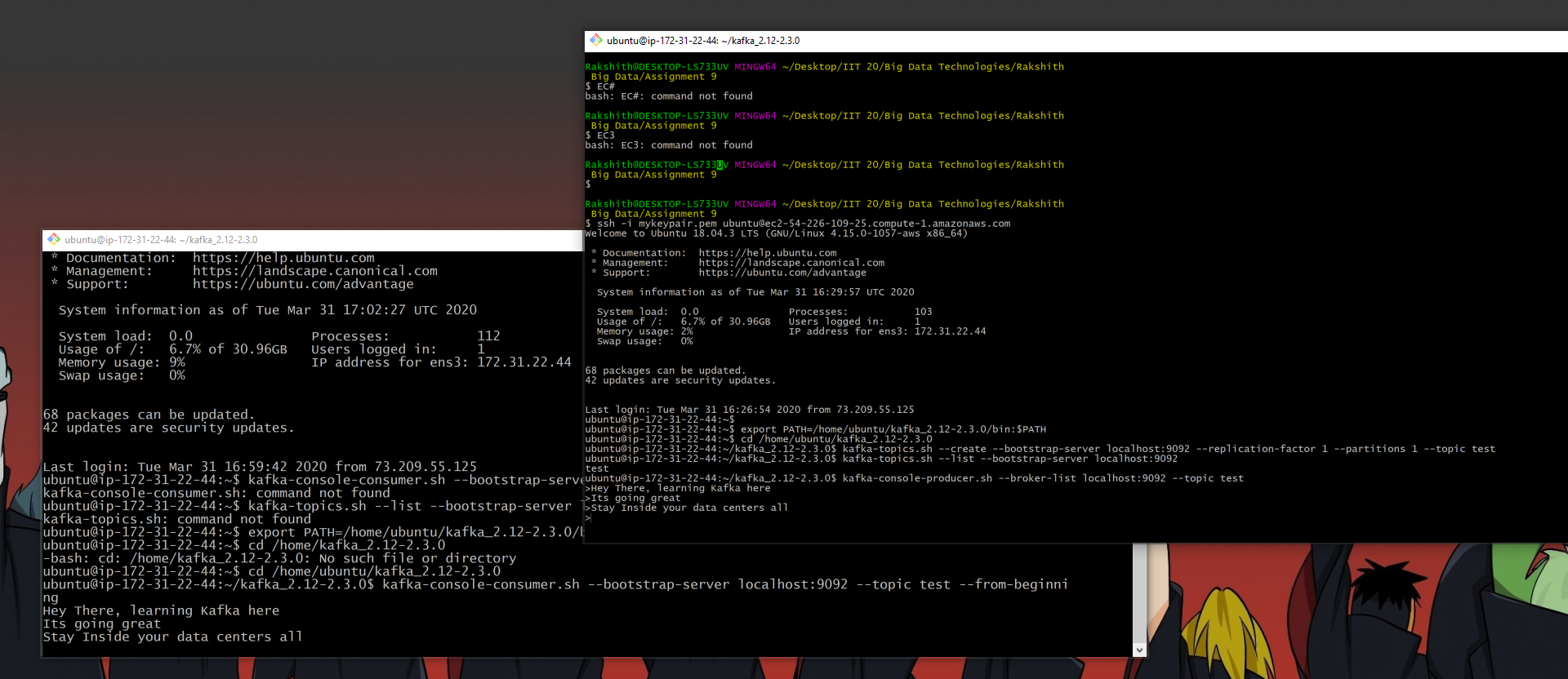


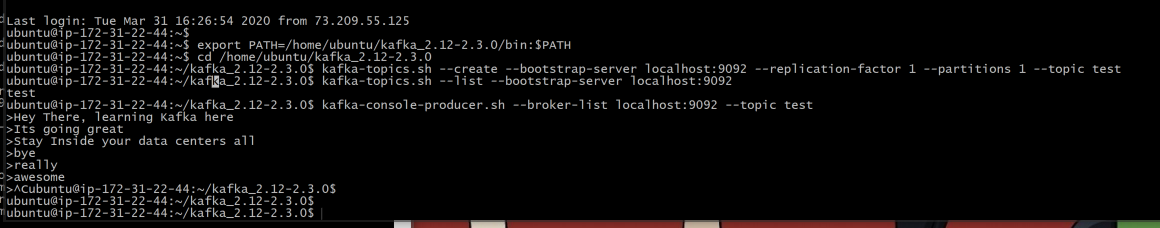
**Section 3:**

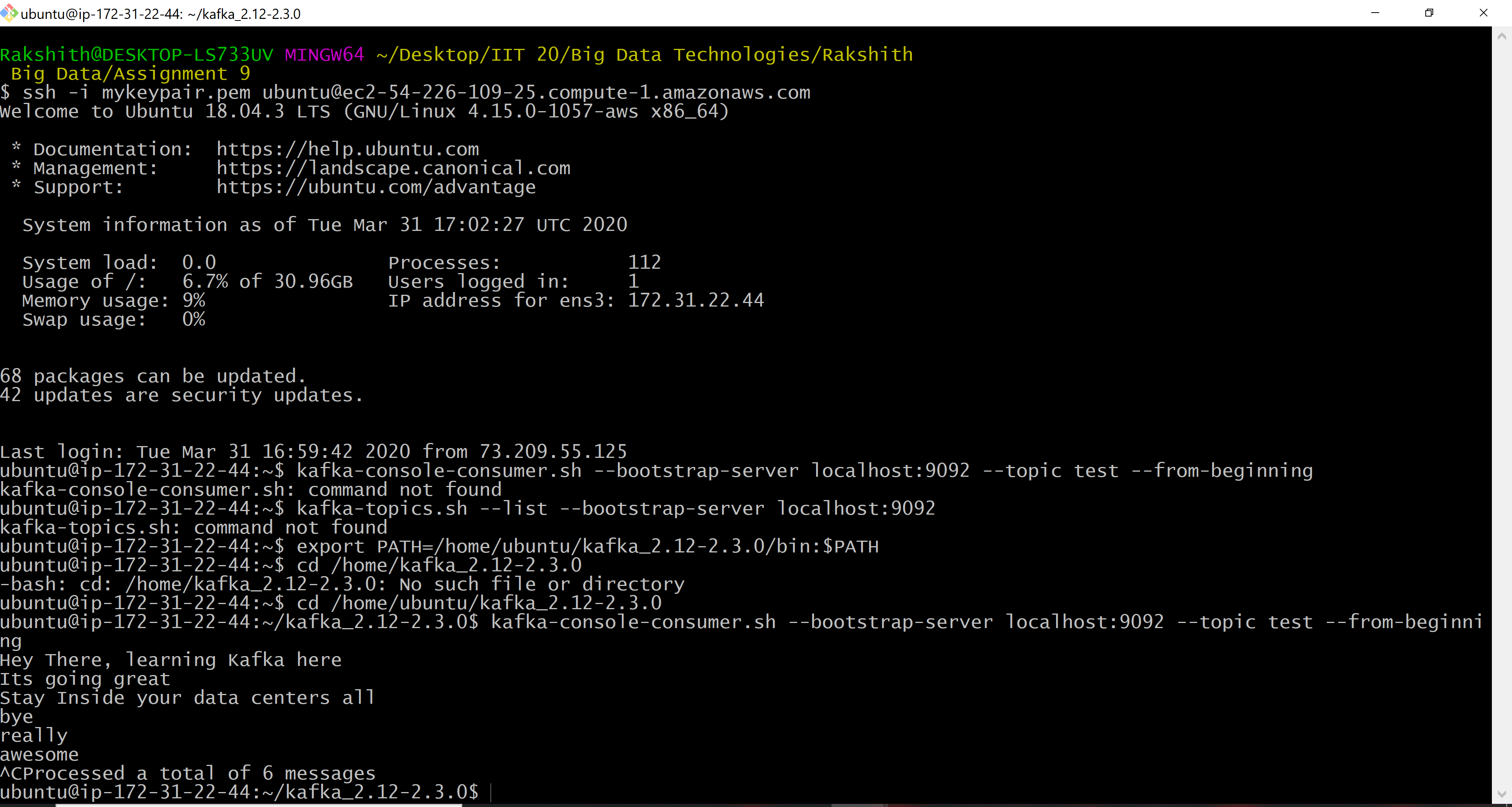


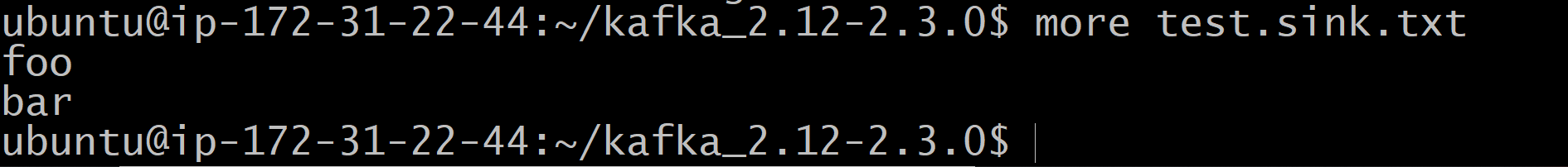


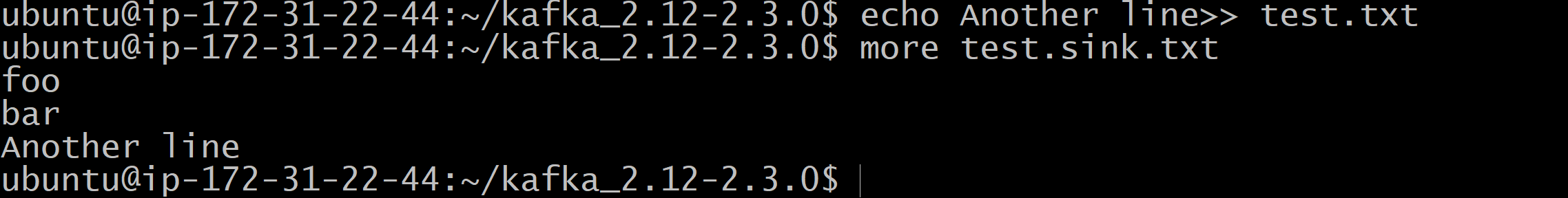
**Section 4:**











Terminating:

