DataEng: Data Transport Activity

# *[this lab activity references tutorials at confluence.com]*

Make a copy of this document and use it to record your results. Store a PDF copy of the document in your git repository along with your code before submitting for this week. For your code, you create several producer/consumer programs or you might make various features within one program. There is no one single correct way to do it. Regardless, store your code in your repository.

The goal for this week is to gain experience and knowledge of using a streaming data transport system (Kafka). Complete as many of the following exercises as you can. Proceed at a pace that allows you to learn and understand the use of Kafka with python.

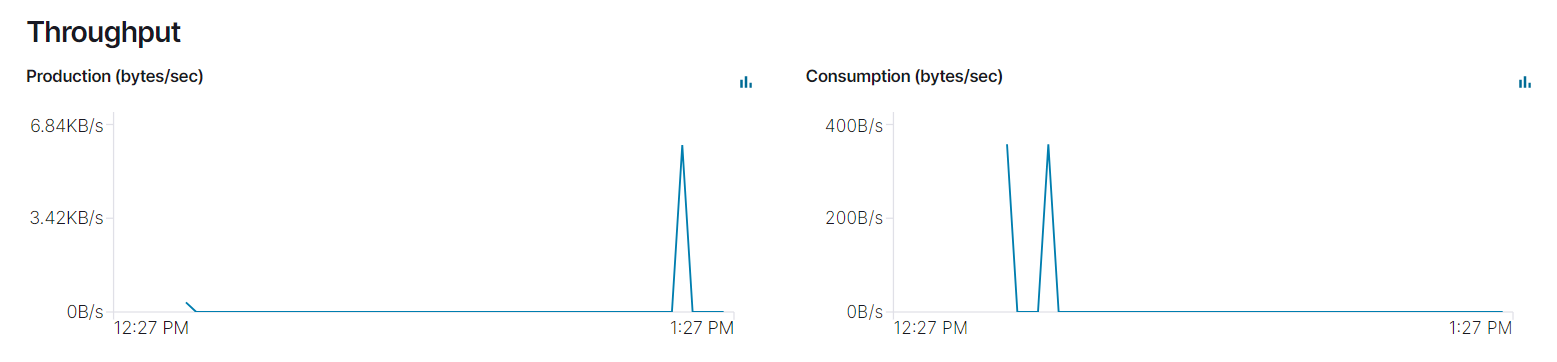
Submit: [In-class Activity Submission Form](https://forms.gle/EYmcAh9mqrzNSU7d7)

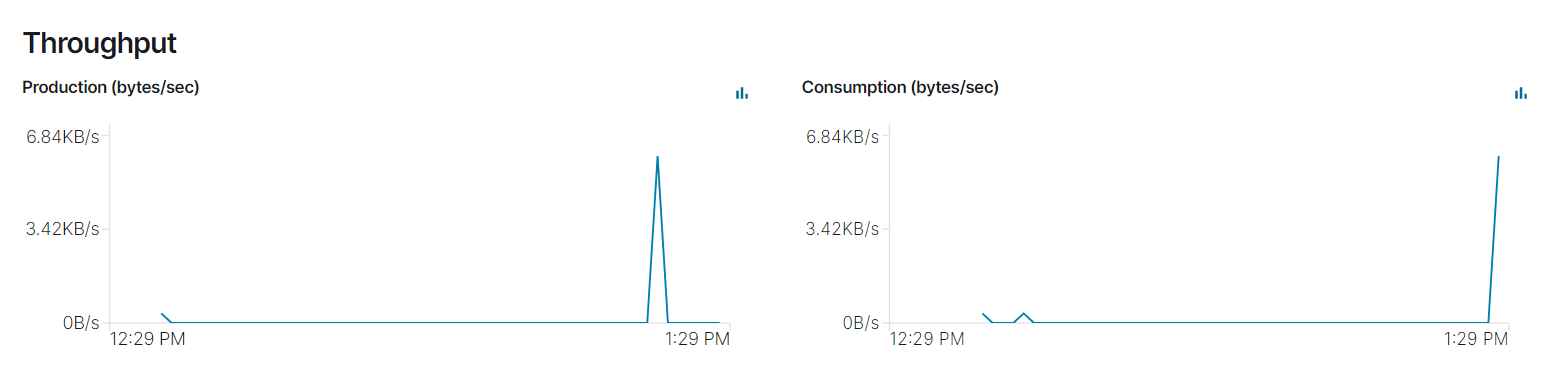
## A. Initialization

1. Get your cloud.google.com account up and running
   1. Redeem your GCP coupon
   2. Login to your GCP console
   3. Create a new, separate VM instance
2. Follow the Kafka tutorial from project assignment #1
   1. Create a separate topic for this in-class activity
   2. Make it “small” as you will not want to use many resources for this activity. By “small” I mean that you should choose medium or minimal options when asked for any configuration decisions about the topic, cluster, partitions, storage, anything. GCP/Confluent will ask you to choose the configs, and because you are using a free account you should opt for limited resources where possible.
   3. Get a basic producer and consumer working with a Kafka topic as described in the tutorials.
3. Create a sample breadcrumb data file (named bcsample.json) consisting of a sample of 1000 breadcrumb records. These can be any records because we will not be concerned with the actual contents of the breadcrumb records during this assignment.
4. Update your producer to parse your sample.json file and send its contents, one record at a time, to the kafka topic.
5. Use your consumer.py program (from the tutorial) to consume your records.

## B. Kafka Monitoring

1. Find the Kafka monitoring console for your topic. Briefly describe its contents. Do the measured values seem reasonable to you?

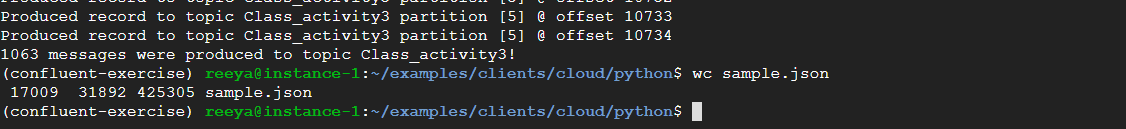




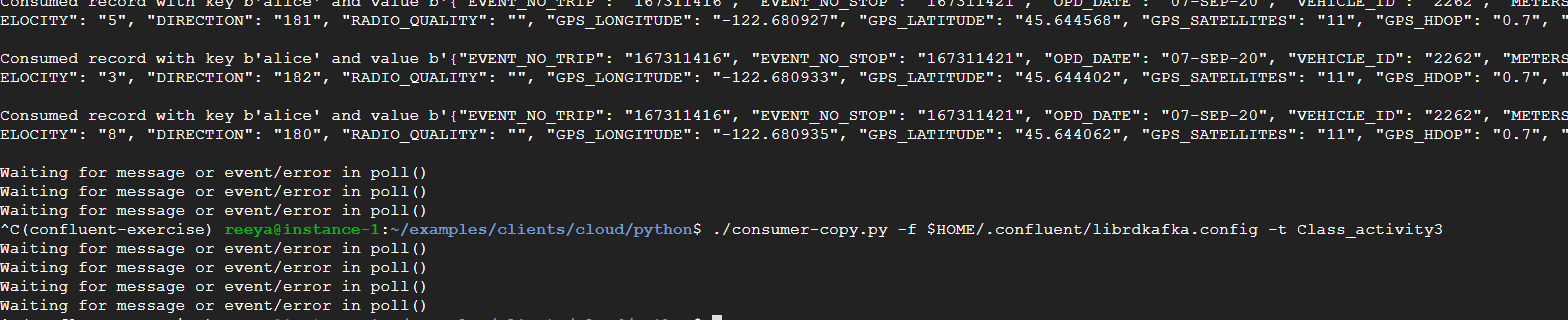
1. Use this monitoring feature as you do each of the following exercises.

## C. Kafka Storage

1. Run the linux command “wc bcsample.json”. Record the output here so that we can verify that your sample data file is of reasonable size.



1. What happens if you run your consumer multiple times while only running the producer once?



1. Before the consumer runs, where might the data go, where might it be stored?

In topic

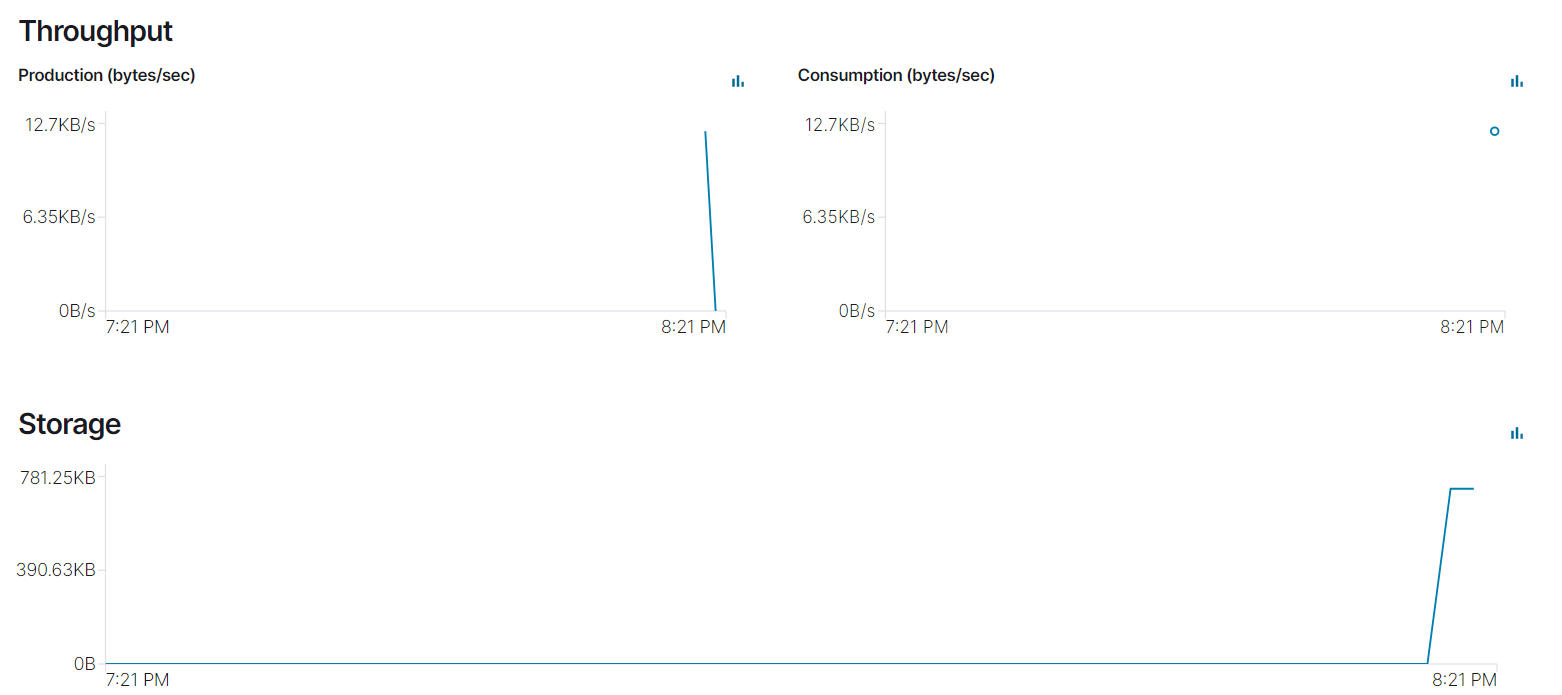
1. Is there a way to determine how much data Kafka/Confluent is storing for your topic? Do the Confluent monitoring tools help with this?



1. Create a “topic\_clean.py” consumer that reads and discards all records for a given topic. This type of program can be very useful during debugging.

## D. Multiple Producers

1. Clear all data from the topic
2. Run two versions of your producer concurrently, have each of them send all 1000 of your sample records. When finished, run your consumer once. Describe the results



## E. Multiple Concurrent Producers and Consumers

1. Clear all data from the topic
2. Update your Producer code to include a 250 msec sleep after each send of a message to the topic.
3. Run two or three concurrent producers and two concurrent consumers all at the same time.
4. Describe the results.

As all 5 processes are operating on same time, they are running parallel. Because of sleep time, producers are sleeping for given time, meanwhile other producers might be producing or sleeping parallelly. At the same time consumers are also waiting for message from topic. As soon as message gets delivered by any of the producer, consumer starts consuming it. However consumer also has poll(1) which means it waits for 1 sec before polling next record.

## F. Varying Keys

1. Clear all data from the topic

So far you have kept the “key” value constant for each record sent on a topic. But keys can be very useful to choose specific records from a stream.

1. Update your producer code to choose a random number between 1 and 5 for each record’s key.
2. Modify your consumer to consume only records with a specific key (or subset of keys).
3. Attempt to consume records with a key that does not exist. E.g., consume records with key value of “100”. Describe the results

* As per my code, there is no record polled by consumer having key 100, as producer is producing records with key range 1-5, the consumer will show message “Record does not have key 100” and it will wait for record with key 100.

1. Can you create a consumer that only consumes specific keys? If you run this consumer multiple times with varying keys then does it allow you to consume messages out of order while maintaining order within each key?

* I ran consumer 3 times, for first 2 times it consumed messages in order or key, messages with 2 first and then with key 5. After running 3rd time, I got messages in out of order of key.

I believe, kafka does not provide facility to consume the record with specific key. Consumer will consumer all the available records but further processing we can handle for specific key.

## G. Producer Flush

The provided tutorial producer program calls “producer.flush()” at the very end, and presumably your new producer also calls producer.flush().

1. What does Producer.flush() do?

* flush() makes producer synchronous by blocking it until previous sent messages in queue have been delivered or errored to the topic. And it increases run time of the process.

1. What happens if you do not call producer.flush()?

* Without flush() I got - 0 messages were produced to topic Class\_activity3! Message. I think, no message got pushed into topic. I checked dashboard as well, producer graph has 0B of data.

1. What happens if you call producer.flush() after sending each record?

* The run time increased because for every message record process waits until that record gets delivered to topic.

1. What happens if you wait for 2 seconds after every 5th record send, and you call flush only after every 15 record sends, and you have a consumer running concurrently? Specifically, does the consumer receive each message immediately? only after a flush? Something else?

* Only after producer flush 15 records, the consumer would be able to receive 15 records.

## H. Consumer Groups

1. Create two consumer groups with one consumer program instance in each group.
2. Run the producer and have it produce all 1000 messages from your sample file.
3. Run each of the consumers and verify that each consumer consumes all of the 1000 messages.
4. Create a second consumer within one of the groups so that you now have three consumers total.
5. Rerun the producer and consumers. Verify that each consumer group consumes the full set of messages but that each consumer within a consumer group only consumes a portion of the messages sent to the topic.

## I. Kafka Transactions

1. Create a new producer, similar to the previous producer, that uses transactions.
2. The producer should begin a transaction, send 4 records in the transactions, then wait for 2 seconds, then choose True/False randomly with equal probability. If True then finish the transaction successfully with a commit. If False is picked then cancel the transaction.

* I have written this code in ‘producer-new.py’, but somehow getting error while committing transaction.

1. Create a new transaction-aware consumer. The consumer should consume the data. It should also use the Confluent/Kaka transaction API with a “read\_committed” isolation level. (I can’t find evidence of other isolation levels).
2. Transaction across multiple topics. Create a second topic and modify your producer to send two records to the first topic and two records to the second topic before randomly committing or canceling the transaction. Modify the consumer to consume from the two queues. Verify that it only consumes committed data and not uncommitted or canceled data.