Kafka + Spark

Kafka is installed on our [Cluster](https://coursys.sfu.ca/2018fa-cmpt-732-g1/pages/Cluster). We will be using it for an assignment, and it could be used for your [Project](https://coursys.sfu.ca/2018fa-cmpt-732-g1/pages/Project).

Spark

If you're doing console output, it will probably get lost in the Spark debugging info. You may want to start by turning that down by doing this early in your code:

spark.sparkContext.setLogLevel('WARN')

In Spark, you can read a Kafka stream from our cluster, and convert the byte strings Kafka produces ihnto usable strings like this:

messages = spark.readStream.format('kafka') \

.option('kafka.bootstrap.servers', '199.60.17.210:9092,199.60.17.193:9092') \

.option('subscribe', topic).load()

values = messages.select(messages['value'].cast('string'))

Please be cautious about how long your job runs. Likely, you don't want an infinite-listening process and should start the stream like this:

stream = streaming\_df.….start()

stream.awaitTermination(600)

And you can run with the necessary Spark package with a command line like:

spark-submit --packages org.apache.spark:spark-sql-kafka-0-10\_2.11:2.3.1 job.py

Non-Spark Python

You can also access Kafka with the [kafka-python](https://pypi.org/project/kafka-python/) package like this:

from kafka import KafkaConsumer

consumer = KafkaConsumer(topic, bootstrap\_servers=['199.60.17.210', '199.60.17.193'])