Big Data Analytics: Lab 10

Note: For this lab, we will be using the external libraries in Spark, *graphframes* and *spark-nlp.* To this end, please follow the steps to prepare your cluster. Note that once installed, you can clone cluster to start them more efficiently later.

1. Start a Cluster (For *Spark-NLP* to work, select the “ML” cluster type)
2. Either go to “Compute”, select the cluster and click on “libraries” or open a notebook, select the cluster and select “Configuration”
3. For *graphframes*, either search in *Maven* or upload the JAR from Moodle to *DBFS.* The correct version for the default version of the cluster is *graphframes* 0.8.2, for Spark 3.2 and Scala 2.12.
4. For Spark NLP, go to the PyPi tab and add “spark-nlp==5.3.3”.
5. Go to *Maven* and enter the following coordinates: *com.johnsnowlabs.nlp:spark-nlp_2.12:5.3.3*