**Files for project execution**

* **config file for docker-compose to spin-up containers of all services**
  + docker-compose.yml
* **Shell Script to setup Airflow environment**
  + airflow\_init.sh
* **Python scripts to test Airflow DAGs**
  + hive\_connection.py
  + hive\_script.py
* **Nifi template to execute data pipelines before and after Spark Job**
  + NifiCovidTemplate.xml
* **HQL scripts to create external hive tables for raw and processed data**
  + corana\_data.hql
  + processed\_data.hql
* **Python Script to run Spark Job**
  + test.py
* **Create t2.2xlarge ec2 instance with Amazon Linux 2 AMI (HVM)**
  + Storage- 96 GB
  + Allow ports- 0-10000
* **Connect to the instance using SSH**
  + ssh -i "D:\path\to\private\key.pem" user@Public\_DNS

Example: *ssh -i "D:\Users\pyerravelly\Desktop\twitter\_analysis.pem" ec2-user@ec2-34-235-133-172.compute-1.amazonaws.com*

**Install and setup Docker and Docker-compose**

*sudo yum update -y*

*sudo yum install docker*

*sudo curl -L "https://github.com/docker/compose/releases/download/1.29.1/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose*

*sudo chmod +x /usr/local/bin/docker-compose*

*sudo gpasswd -a $USER docker*

*newgrp docker*

**Start/Stop Docker**

Start*- sudo systemctl start docker*

Stop*- sudo systemctl stop docker*

* **Copy folder from local to ec2 and give required permissions**

*scp -r -i "dezyre.pem" docker\_exp ec2-user@ec2-34-235-133-172.compute-1.amazonaws.com:/home/ec2-user/docker\_exp*

*sudo chmod -R 755 docker\_exp*

* **Change directory to docker\_exp**
  + *cd docker\_exp*
* **List files with permissions**
  + *ls -l*
* **Run shell script airflow\_init.sh after starting the docker**
* *./airflow\_init.sh*
* 3 new folders named ‘dags’, ‘logs’ and ‘plugins’ will be created
* **move airflow DAG scripts to dags folder**
* *cd docker\_exp* (run if not in the right directory)
* *mv -t dags hive\_script.py hive\_connection.py*
* **Start/Stop docker containers**
* *cd docker\_exp* (run if not in the right directory)
* *docker-compose up*
* *docker-compose stop*
* *docker-compose down* (stop and remove containers)
* **Port Forwarding to access services locally**

*ssh -i "dezyre.pem" ec2-user@ec2-34-235-133-172.compute-1.amazonaws.com -o "ServerAliveInterval 30" -L 2081:localhost:2041 -L 4888:localhost:4888 -L 4889:localhost:4889 -L 2080:localhost:2080 -L 8050:localhost:8050 -L 8051:localhost:8051 -L 4141:localhost:4141 -L 4090:localhost:4090 -L 3180:localhost:3180 -L 50075:localhost:50075 -L 50070:localhost:50070 -L 50010:localhost:50010 -L 3077:localhost:3077 -L 4080:localhost:4080 -L 9870:localhost:9870 -L 8188:localhost:8188 -L 9864:localhost:9864 -L 8042:localhost:8042 -L 8088:localhost:8088 -L 8080:localhost:8080 -L 8081:localhost:8081 -L 10000:localhost:10000 -L 6080:localhost:6080 -L 8998:localhost:8998*

* **Check status of all running containers and get their ports and names**

*docker ps*

**To connect to different services locally after port forwarding**

* <http://localhost:4888/lab?-> jupyterlab
* <http://localhost:6080/> - airflow (username & password- **airflow**)
* <http://localhost:8080/> -spark master
* <http://localhost:2080/nifi/>- nifi

**To get into bash shell of different containers**

* docker exec -i -t <container name> bash
* *docker exec -i -t hdp\_namenode bash (for hadoop )*
* *docker exec -i -t hdp\_spark-master bash*
* *docker exec -i -t hdp\_hive-server bash*

**To exit from any container’s bash shell**

* ctrl+D

**To execute both the nifi pipelines (1. Before spark job for HDFS and Kafka and**

**2. After spark job)**

* Goto <http://localhost:2080/nifi/>
* Upload and use template NifiCovidTemplate.xml (covid\_proj\_final\_nifi)
* **Open Process group “EncryptionTemplate-v1.3” goto “EncryptContent” -> Configure -> Properties, and set Password to a 10-character string (eg. ‘dezyre1234’)**

**To access data pushed to hadoop from nifi**

* *docker exec -i -t hdp\_namenode bas*
* *hadoop fs -ls /dezyre\_data/corona-table*
* *hadoop fs -ls /dezyre\_data/dezyre\_kafka\_out* (after executing spark job and 2nd nifi pipeline)

**To access data published to kafka topic from nifi**

* *docker exec -i -t hdp\_kafka bash*
* *cd /opt/kafka\_2.12-2.5.0*
* *bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --from-beginning --topic dezyre\_data\_csv* (change kafka topic name accordingly)

**To run spark-submit job to publish data to kafka topic**

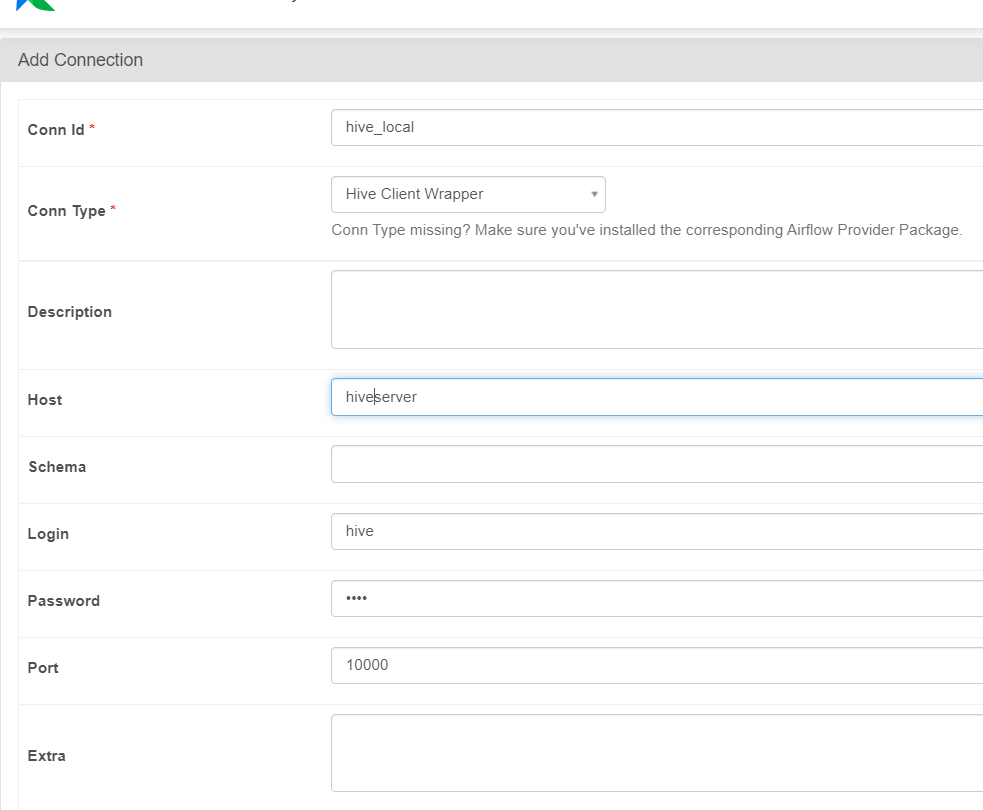
* *docker cp /home/ec2-user/docker\_exp/test.py hdp\_spark-master:/test.py*
* *docker exec -i -t hdp\_spark-master bash*
* *chmod 755 test.py*
* *./spark/bin/spark-submit --packages org.apache.spark:spark-sql-kafka-0-10\_2.12:3.0.0 --master local[2] test.py*

**To create external hive table or run queries**

* *docker exec -i -t hdp\_hive-server bash*
* *hive*
* Run the queries from corona\_data.hql and processed\_data.hql to create external tables for both

**Setup up Airflow for hive access**

* Open airflow dashboard through localhost



(Note: Host= hiveserver)