How To ?

This project was exclusively done in Centos and was demonstrated in a virtual box.

Steps to complete the entire project’s process are:

1. First install virtual box (Oracle Vm was used )
2. Then import the file VM 2\_tb\_analysis.ova which has the state of all the configurations as well versions
3. use, hadoop user with password ‘***hadoop***’
4. Open konsole
5. Being on root, open fume directory with command : *cd /usr/local/flume/flume-1.9.0*
6. Start flume to ingest data from file in location /Home/Documents/project\_mtb , into HDFS
7. Run this command to run flume : *./bin/flume-ng agent --name agent2 --conf ./conf --conf-file ./conf/project.conf -Dflume.root.logger=INFO,console*
8. The data in HDFS will be saved to /usr/hadoop/mtb\_data
9. Open Jupyter Notebook in Documents folder where we have Tuberclosis\_data\_processing\_bigdatapipeline.ipynb file to run
10. In terminal just run : *jupyter notebook*
11. Inside the notebook run all cells
12. The scripts inside the notebook will do loading, preprocessing, creating a database and saving the processed data into a specific table.
13. You can see tables created in HDFS at location : /usr/hive/warehouse/my\_databasemtb.db/mtb\_table
14. Database can also be viewed by running command in terminal : *hive*
15. To run superset in terminal go to : /Home/Documents/superset and run command: *sudo docker compose -f docker-compose-image-tag.yml up*
16. In browser search : *localhost/8089/superset*  (credentials, user: ***admin*** and password: ***admin****)*
17. The analysis is presented in the **Tuberculosis Analysis** Dashboard.

For more detailed information please refer to demo provided along with this file

Thank you for your time!!!