Deeque Vial\_Anastasiya

1. Install Docker on Windows.
2. Clone repository <https://github.com/jupyter/docker-stacks>.

Chose any location on computer and clone repo  
A screenshot of a computer

Description automatically generated with low confidence

C:\Users\Anastasiya\_Vial\deequ\clone\_repo\docker-stacks\pyspark-notebook – place with dockerfile

1. Rebuild docker "pyspark-notebook" image to support spark 3.0.0 version for PyDeequ needs according to instructions: <https://jupyter-docker-stacks.readthedocs.io/en/latest/using/specifics.html>  
   sample:  
   "docker build --rm --force-rm  
   -t jupyter/pyspark-notebook:spark-3.0.0 ./pyspark-notebook   
   --build-arg spark\_version=3.0.0   
   --build-arg spark\_checksum=bfe45406c67cc4ae00411ad18cc438f51e7d4b6f14eb61e7bf6b5450897c2e8d3ab020152657c0239f253735c263512ffabf538ac5b9fffa38b8295736a9c387"

My code:

docker build --rm --force-rm  
-t jupyter/pyspark-notebook:spark-3.0.0 C:\Users\Anastasiya\_Vial\deequ\clone\_repo\docker-stacks\pyspark-notebook   
--build-arg spark\_version=3.0.0   
--build-arg hadoop\_version=3.2   
--build-arg scala\_version=2.12  
--build-arg spark\_checksum=BFE45406C67CC4AE00411AD18CC438F51E7D4B6F14EB61E7BF6B5450897C2E8D3AB020152657C0239F253735C263512FFABF538AC5B9FFFA38B8295736A9C387  
--build-arg openjdk\_version=8

docker build --rm --force-rm -t jupyter/pyspark-notebook:spark-3.0.0 C:\Users\Anastasiya\_Vial\deequ\clone\_repo\docker-stacks\pyspark-notebook --build-arg spark\_version=3.0.0 --build-arg hadoop\_version=3.2 --build-arg scala\_version=2.12 --build-arg spark\_checksum=bfe45406c67cc4ae00411ad18cc438f51e7d4b6f14eb61e7bf6b5450897c2e8d3ab020152657c0239f253735c263512ffabf538ac5b9fffa38b8295736a9c387 --build-arg openjdk\_version=8

(!я меняла ссылку на загрузку спарка в докер файле, тк версия 3.0.0 лежит в архиве)

A picture containing screenshot

Description automatically generated

1. Run rebuilt docker image:

docker run -v %cd%:/home/jovyan/work -p 8888:8888 -p 4040:4040 --user root -e JUPYTER\_ENABLE\_LAB=yes --name pyspark jupyter/pyspark-notebook:spark-3.0.0

этот код запустит контейнер запустит контейнер пайспарк в докере. Там же в докере можно посмотреть все креденшиалс  
A picture containing text, screenshot, font

Description automatically generated  
To access the server, open this file in a browser:

file:///home/jovyan/.local/share/jupyter/runtime/jpserver-18-open.html

Or copy and paste one of these URLs:

http://cf1525079410:8888/lab?token=bbad3045a4a5f17301eaca19165df381429151c434e43e8f

<http://127.0.0.1:8888/lab?token=bbad3045a4a5f17301eaca19165df381429151c434e43e8f>

token=bbad3045a4a5f17301eaca19165df381429151c434e43e8f

1. Copy necessary JDBC driver to running Docker container for the further usage.

We need driver for sql server (as we are supposed to connect with SQL Server)

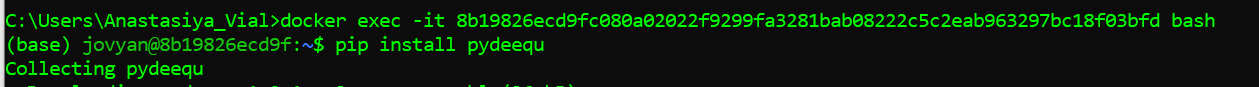
docker cp sqljdbc42.jar f00e97ec1c30:/home/jovyan/work

A picture containing screenshot, green

Description automatically generated

1. Upload Deequ\_puSpark\_skeleton and run command from Deequ\_pySpark\_skeleton\_Vial.ipynb

install pydeequ:  
docker exec -it 8b19826ecd9fc080a02022f9299fa3281bab08222c5c2eab963297bc18f03bfd bash  
pip install pydeequ

  
  
restart Kernel (go to Kernel => restart)