1. Install Docker on Windows <https://docs.docker.com/desktop/windows/install/>
2. Clone repository *git clone https://github.com/jupyter/docker-stacks*
3. Edit docker file

*ARG spark\_version="3.0.0"*

*ARG hadoop\_version="3.2"*

*ARG spark\_checksum="bfe45406c67cc4ae00411ad18cc438f51e7d4b6f14eb61e7bf6b5450897c2e8d3ab020152657c0239f253735c263512ffabf538ac5b9fffa38b8295736a9c387"*

*ARG openjdk\_version="8"*

*…*

*# You need to use https://archive.apache.org/dist/ website if you want to download old Spark versions*

*# But it seems to be slower, that's why we use recommended site for download*

*RUN if [ -z "${scala\_version}" ]; then \*

*wget -qO "spark.tgz" "https://archive.apache.org/dist/spark/spark-${APACHE\_SPARK\_VERSION}/spark-${APACHE\_SPARK\_VERSION}-bin-hadoop${HADOOP\_VERSION}.tgz"; \*

*else \*

*wget -qO "spark.tgz" "https://archive.apache.org/dist/spark/spark-${APACHE\_SPARK\_VERSION}/spark-${APACHE\_SPARK\_VERSION}-bin-hadoop${HADOOP\_VERSION}-scala${scala\_version}.tgz"; \*

*fi && \*

*echo "${spark\_checksum} \*spark.tgz" | sha512sum -c - && \*

*tar xzf "spark.tgz" -C /usr/local --owner root --group root --no-same-owner && \*

*rm "spark.tgz"*

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

"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"

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 pysparkHW jupyter/pyspark-notebook:spark-3.0.0 "

1. Check spark version:

*Docker exec -it -u 0 f64da20c6522 /bin/bash*

*export SPARK\_SUBMIT\_OPTS="--add-opens java.base/java.nio=ALL-UNNAMED"*

*spark-submit --version*

Text

Description automatically generated

1. Download and copy nessessary JDBC driver to running Docker container for the further usage.

Docker cp C:\Users\Anastasiya\_Viktarovi\Auto\sqljdbc\_12.2\enu\ pyspark:/home/jovyan/work/

1. Check if file exists on container

Text

Description automatically generated

1. Create and run code in notebook – notebook and report in HW.ipynb and Verification Result Report.html files