Không làm được bước 1 - 4 thì pull docker hub này về `docker pull nguyennhattung2003/spark`
và run cái images vừa pull về `docker run -d --name spark --network mynetwork -p 8888:8888 -p 4040:4040 nguyennhattung2003/spark` -> bắt đầu luôn từ bước 5.

1. Tạo Dockerfile với nội dung:

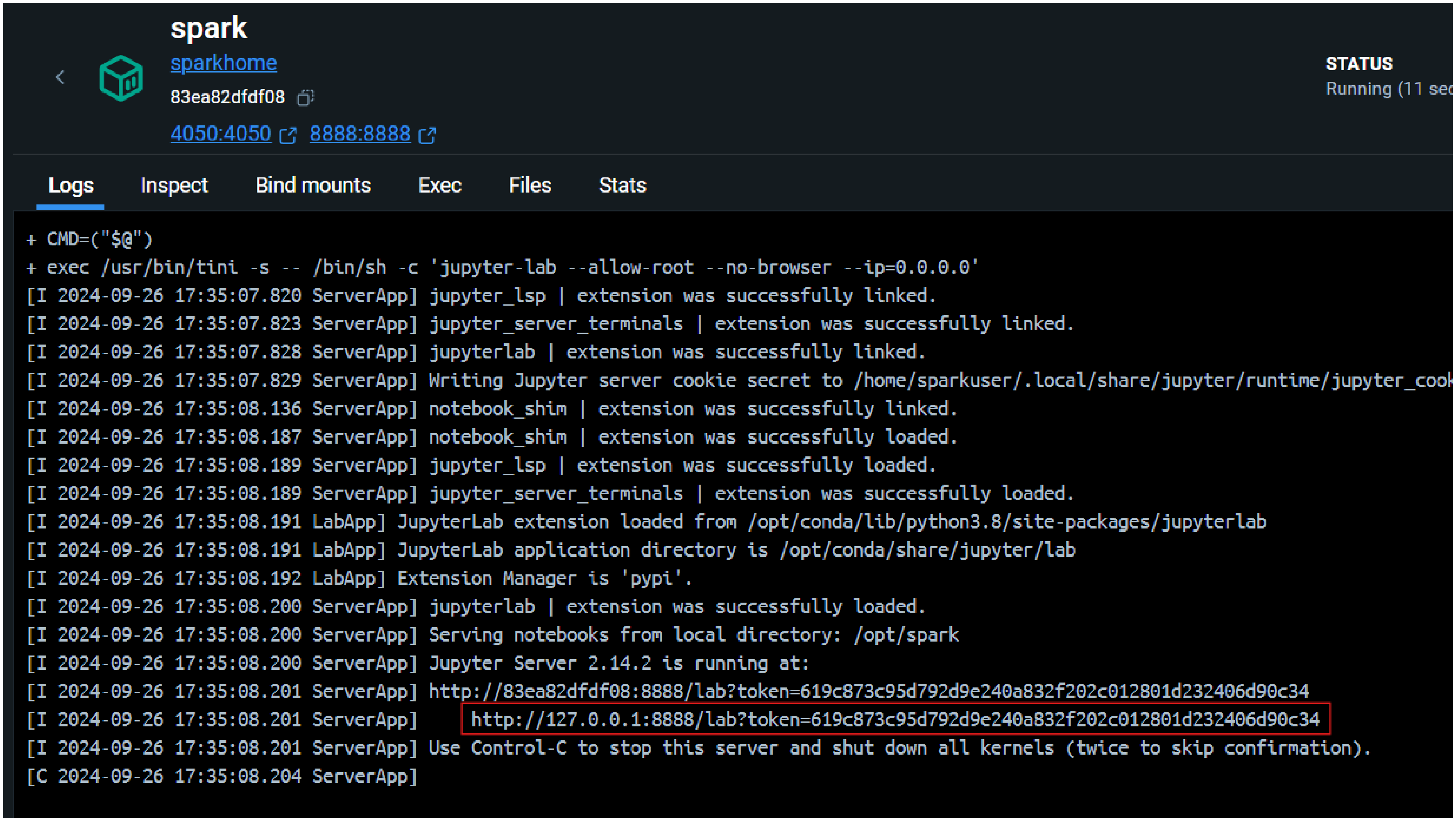
   ```
   FROM apache/spark:latest
   USER root
   WORKDIR /opt/spark
   RUN pip install --upgrade pip
   COPY  requirements.txt .
   RUN pip3 install -r requirements.txt
   CMD jupyter-lab --allow-root --no-browser --ip=0.0.0.0
   ```

2. Tạo requirements.txt với nôi dung:

   ```
   pyspark
   pymongo
   itemadapter
   pandas
   numpy
   jupyterlab
   ```

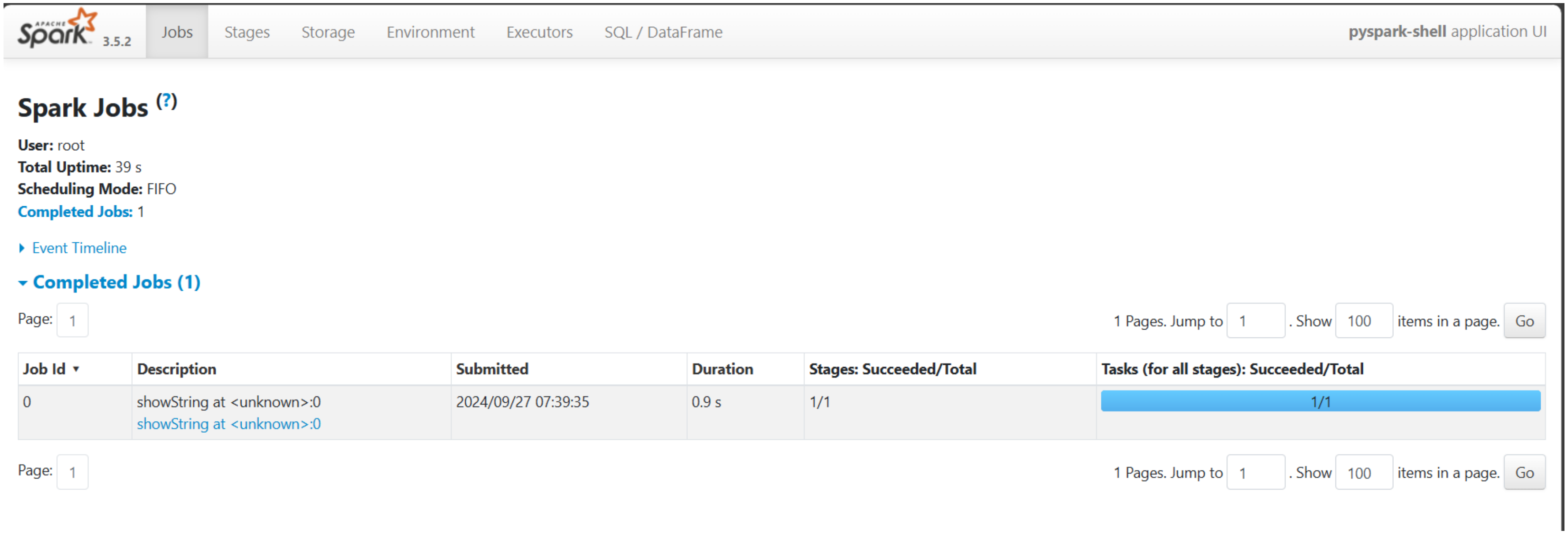
3. Build docker: `docker build . -t sparkhome`
4. Run docker container: `docker run -d --name spark --network mynetwork -p 8888:8888 -p 4050:4050 sparkhome`

5. Truy cập jupyter với `http://127.0.0.1:8888/`



```
+ CMD=("$@")
+ exec /usr/bin/tini -s -- /bin/sh -c 'jupyter-lab --allow-root --no-browser --ip=0.0.0.0'
[I 2024-09-26 17:35:07.820 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2024-09-26 17:35:07.823 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2024-09-26 17:35:07.828 ServerApp] jupyterlab | extension was successfully linked.
[I 2024-09-26 17:35:07.829 ServerApp] Writing Jupyter server cookie secret to /home/sparkuser/.local/share/jupyter/runtime/jupyter_cook
[I 2024-09-26 17:35:08.136 ServerApp] notebook_shim | extension was successfully linked.
[I 2024-09-26 17:35:08.187 ServerApp] notebook_shim | extension was successfully loaded.
[I 2024-09-26 17:35:08.189 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2024-09-26 17:35:08.189 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2024-09-26 17:35:08.191 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.8/site-packages/jupyterlab
[I 2024-09-26 17:35:08.191 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab
[I 2024-09-26 17:35:08.192 LabApp] Extension Manager is 'pypi'.
[I 2024-09-26 17:35:08.200 ServerApp] jupyterlab | extension was successfully loaded.
[I 2024-09-26 17:35:08.200 ServerApp] Serving notebooks from local directory: /opt/spark
[I 2024-09-26 17:35:08.200 ServerApp] Jupyter Server 2.14.2 is running at:
[I 2024-09-26 17:35:08.201 ServerApp] http://83ea82dfdf08:8888/lab?token=619c873c95d792d9e240a832f202c012801d232406d90c34
[I 2024-09-26 17:35:08.201 ServerApp]     http://127.0.0.1:8888/lab?token=619c873c95d792d9e240a832f202c012801d232406d90c34
[I 2024-09-26 17:35:08.201 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 2024-09-26 17:35:08.204 ServerApp]
```

6. Truy cập Spark UI với `http://127.0.0.1:4040/`



Spark Jobs (?)

User: root
Total Uptime: 39 s
Scheduling Mode: FIFO
Completed Jobs: 1

▸ Event Timeline

▾ **Completed Jobs (1)**

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page.

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	showString at <unknown>:0 showString at <unknown>:0	2024/09/27 07:39:35	0.9 s	1/1	1/1

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page.

7. Cleaning data ✅ code: https://drive.google.com/file/d/1Ygonr6XFc8IqI4ppd72TxEMzdWJo9VIr/view?usp=drive_link
8. Data modeling ❌
9. Connect with Postgresql ❌