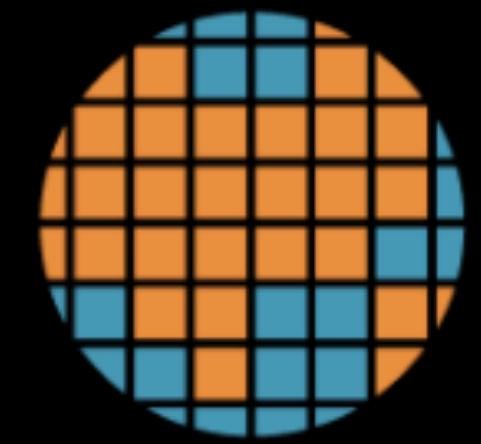


CHASING A WHALE  
DOWN A  
RABBIT HOLE



# PyData Global



**NOVEMBER 11TH - 15TH**

# ABOUT ME:

Data Science Lead @ Xccelerated

Co-chair @ PyData Amsterdam

Vice-chair @ PyData Global

Maintainer of scikit-lego, timeseers

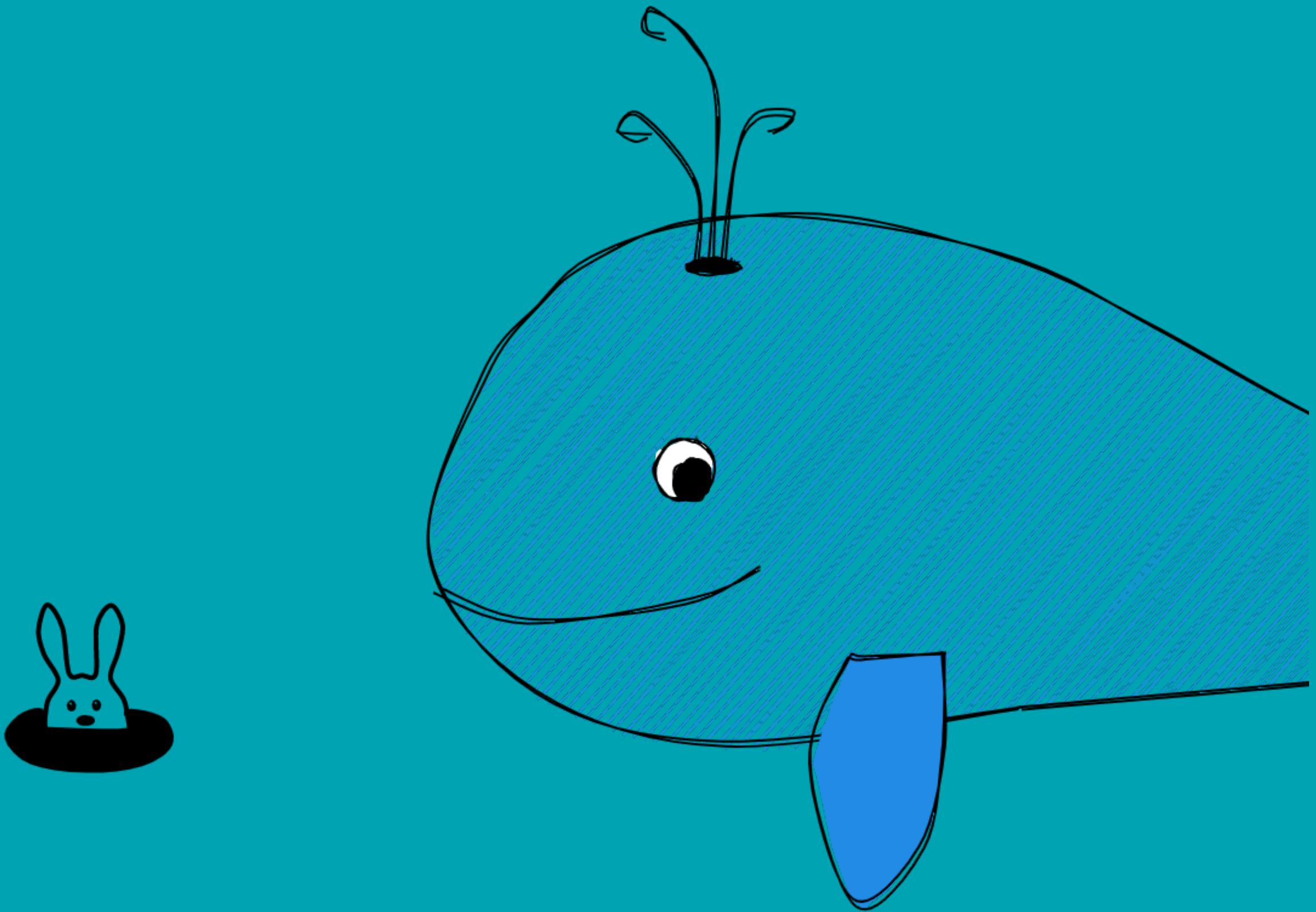


**Jason Ash**  
@ashjasont



docker question: how small could a container for a python app (pandas, numpy, scipy, and friends) reasonably get? I'm sitting at 580mb after a few hours of tinkering with layers. Any examples of a small data science image you could point me to?

9:05 PM · Apr 3, 2020 · Twitter for iPhone



```
matplotlib==3.2.1
numpy==1.19.2
pandas==1.0.3
scikit-learn==0.22.0
cycler==0.10.0
# via matplotlib
kiwisolver==1.2.0
# via matplotlib
pyparsing==2.4.6
# via matplotlib
python-dateutil==2.8.1
# via matplotlib, pandas
pytz==2019.3
# via pandas
joblib==0.16.0
# via scikit-learn
scipy==1.5.2
# via scikit-learn
six==1.14.0
# via cycler, python-dateutil
```

# THE DOCKERFILE IN QUESTION

```
FROM python:3.8
```

```
COPY requirements.txt .
```

```
RUN pip install -r requirements.txt
```

```
COPY . . .
```

```
RUN pip install .
```

```
RUN ["python3", "test.py"]
```

# THE DOCKERFILE IN QUESTION

```
> docker build -f py38.Dockerfile -t size-py38 .
> docker images | grep size
size-py38    latest          d193ec6318c7        1 minute ago      1.69GB
```

STEP 1:

THINK ABOUT YOUR  
APPLICATION'S NEEDS

# Project folder

```
•
├── .git
│   └── ...
├── .idea
│   └── ...
├── data
│   └── *.csv
├── models
│   └── *.pkl
├── notebooks
│   ├── .ipynb_checkpoints
│   │   └── *.ipynb
│   └── ...
├── readme.md
├── setup.py
└── src
    ├── small_docker
    │   ├── __pycache__
    │   │   └── *.py
    │   └── small_docker.egg-info
    │       └── ...
└── venv
    └── ...
└── tests
    └── ...
```

# Project folder

```
.  
├── .git  
│   └── ...  
├── .idea  
│   └── ...  
├── data  
│   └── *.csv  
├── models  
│   └── *.pkl  
├── notebooks  
│   └── .ipynb_checkpoints  
│       └── *.ipynb  
├── readme.md  
└── setup.py
```

src

```
└── small_docker  
    ├── __pycache__  
    │   └── *.py  
    └── small_docker.egg-info  
        └── ...
```

venv

```
└── ...
```

tests

```
└── ...
```

# .dockerignore

```
.git  
.idea  
**/__pycache__/  
**/*.*.egg-info  
*.pyc  
.pytest_cache/  
venv/  
  
data/  
models/  
notebooks/
```

# DOCKERIGNORE

```
> docker build -f py38.Dockerfile -t size-py38 .
```

```
> docker images | grep size
```

size-py38	latest	15ecbfe3c323	3 minute ago	1.69GB
size-py38	latest	d193ec6318c7	1 minute ago	1.29GB

DON'T IGNORE

DOCKERTIGNORE

> docker history size-py38			
IMAGE	CREATED	CREATED BY	SIZE
d193ec6318c7	1 day ago	python3 test.py	44.8kB
5b1e1cda3241	1 day ago	/bin/sh -c #(nop) COPY file:08513e72b1abb5c7...	174B
3a2eb08ee1ab	1 day ago	/bin/sh -c pip install -r requirements.txt	361MB
34a02dc176cd	1 day ago	/bin/sh -c #(nop) COPY file:89534a70ba7662fd...	165B
7f5b6ccd03e9	1 day ago	/bin/sh -c #(nop) CMD ["python3"]	0B
<missing>	1 day ago	FROM python:3.8	929MB

STEP 2:

USE A SMALL BASE IMAGE

# PYTHON:3.8-SLIM

```
FROM python:3.8-slim
```

```
COPY requirements.txt .
```

```
RUN pip install -r requirements.txt
```

```
COPY test.py .
```

```
RUN ["python3", "test.py"]
```

# PYTHON:3.8-SLIM

```
> docker build -f py38-slim.Dockerfile -t size-py38-slim .
```

```
> docker images | grep size
```

size-py38	latest	15ecbfe3c323	3 minute ago	1.69GB
size-py38	latest	d193ec6318c7	1 minute ago	1.29GB
size-py38-slim	latest	097d88af2244	1 minute ago	489mb

DUTWELTIN

# LET'S REMOVE THOSE

```
FROM python:3.8-slim

COPY requirements.txt .

RUN pip install --user -r requirements.txt

RUN rm -rf /root/.cache/pip
RUN find /usr/local/bin/lib/python3.8/site-packages/ -follow -type f ( \
    -name '*.a' -name '*.txt' -name '*.md' -name '*.png' \
    -name '*.jpg' -name '*.jpeg' -name '*.js.map' -name '*.pyc' \
    -name '*.c' -name '*.pxc' -name '*.pyd' \
) -delete
RUN find /usr/local/lib/python3.8 -name '__pycache__' | xargs rm -r

COPY test.py .
RUN ["python3", "test.py"]
```

# LET'S REMOVE THOSE

```
> docker build -f py38-slim-no-cache.Dockerfile -t py38-slim-no-cache .
> docker images | grep size
size-py38           latest    15ecbfe3c323    3 minute ago    1.69GB
size-py38           latest    d193ec6318c7    8 minutes ago   1.29GB
size-py38-slim      latest    097d88af2244    5 minutes ago   489mb
size-py38-slim-cache latest    097d88af2244    1 minute ago   503mb
```

WHY DIDN'T THIS SLIM OUR  
IMAGE

9d84edf35a0a

FROM Python:3.8-slim

113 mb



9d84edf35a0a.tar

3da088a4ef01

COPY requirements.txt .

+ requirements.txt

114 b



3da088a4ef01.tar

9d84edf35a0a

FROM Python:3.8-slim

113 mb



9d84edf35a0a.tar

fe8725920948

RUN pip install -r requirements.txt

+ /root/.cache/pip  
+ /root/.local/lib/python3.8/site-packages

324 mb



fe8725920948.tar

3da088a4ef01

COPY requirements.txt .

+ requirements.txt

114 b



3da088a4ef01.tar

9d84edf35a0a

FROM Python:3.8-slim

113 mb



9d84edf35a0a.tar

8c8cla79aaaf6

RUN rm -rf /root/.cache/pip

- /root/.cache/pip

0 b

8c8cla79aaaf6.tar

fe8725920948

RUN pip install -r requirements.txt

+ /root/.cache/pip

+ /root/.local/lib/python3.8/site-packages

324 mb

fe8725920948.tar

32a088a4efaf

COPY requirements.txt .

+ requirements.txt

114 b

32a088a4efaf.tar

9d84edf35a0a

FROM Python:3.8-slim

113 mb

9d84edf35a0a.tar

b83fd35df863

RUN find /root/.local/ -follow -type f ... -delete

- /root/.local/ lib/python3.8/site-packages

0 b



b83fd35df863.tar

8c8cla79aaf6

RUN rm -rf /root/.cache/pip

- /root/.cache/pip

0 b



8c8cla79aaf6.tar

fe8725920948

RUN pip install -r requirements.txt

+ /root/.cache/pip  
+ /root/.local/ lib/python3.8/site-packages

324 mb



fe8725920948.tar

32a088a4ef01

COPY requirements.txt .

+ requirements.txt

114 b



32a088a4ef01.tar

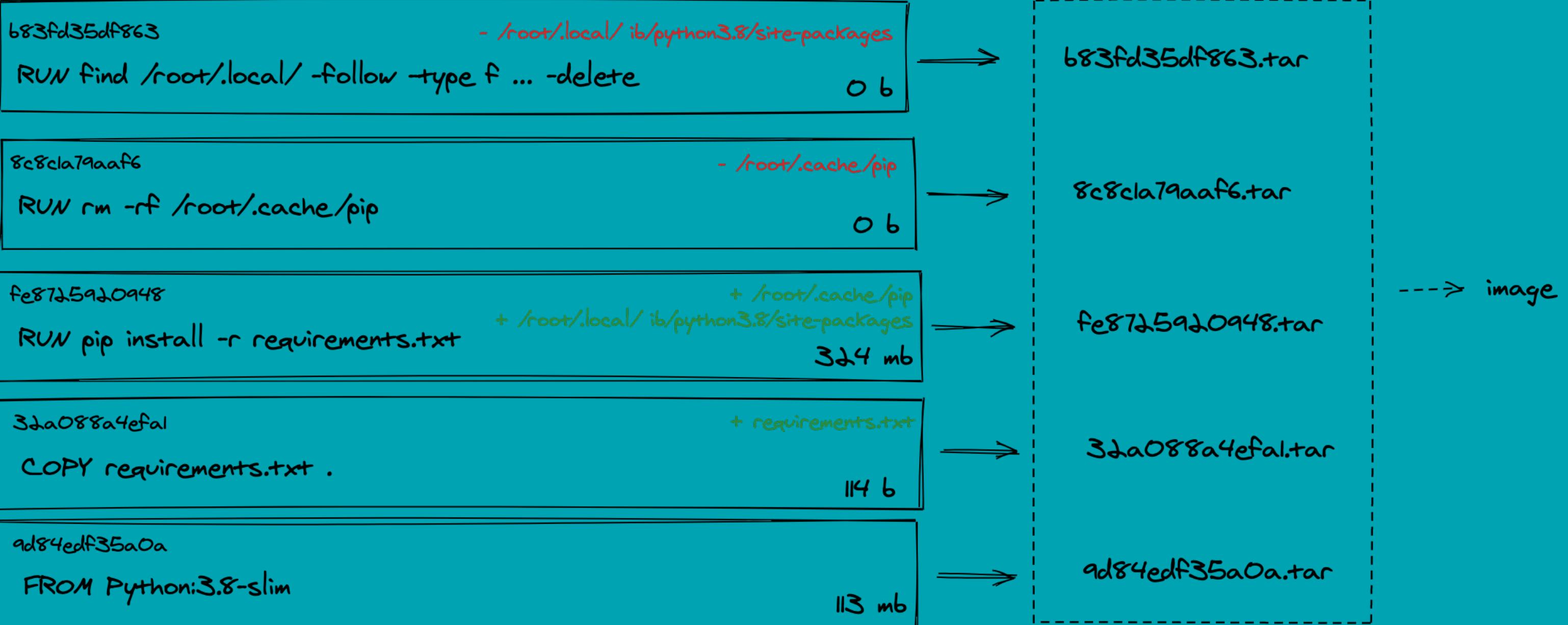
9d84edf35a0a

FROM Python:3.8-slim

113 mb



9d84edf35a0a.tar



STEP 3:

USE AS FEW LAYERS AS POSSIBLE

# SQUASHING LAYERS

```
FROM python:3.8-slim

COPY requirements.txt .

RUN pip install --user --no-cache-dir -r requirements.txt \
&& find /root/.local/ -follow -type f ( \
    -name '*.a' -name '*.txt' -name '*.md' -name '*.png' \
    -name '*.jpg' -name '*.jpeg' -name '*.js.map' -name '*.pyc' \
    -name '*.c' -name '*.pxc' -name '*.pyd' \
) -delete \
&& find /usr/local/lib/python3.8 -name '__pycache__' | xargs rm -r

COPY test.py .
RUN ["python3", "test.py"]
```

# SQUASHING LAYERS

```
> docker build -f py38-slim-no-cache2.Dockerfile -t py38-slim-no-cache2 .
> docker images | grep size
size-py38           latest  15ecbfe3c323    3 minute ago  1.69GB
size-py38           latest  d193ec6318c7    13 minutes ago 1.29GB
size-py38-slim       latest  097d88af2244    10 minutes ago 489mb
size-py38-slim-cache latest  b44f2fefd038    5 minute ago  503mb
size-py38-slim-cache2 latest  5acbbfa56210    1 minute ago  326mb
```

# SQUASHING LAYERS

```
> docker build -f py38-slim-no-cache.Dockerfile -t py38-slim-no-cache-sq --squash .
> docker images | grep size
size-py38
size-py38
size-py38-slim
size-py38-slim-cache
size-py38-slim-cache2
size-py38-slim-cache-sq
          latest  15ecbfe3c323    3 minute ago  1.69GB
          latest  d193ec6318c7    13 minutes ago  1.29GB
          latest  097d88af2244   10 minutes ago  489mb
          latest  b44f2fefd038   5 minute ago   503mb
          latest  5acbbfa56210   1 minute ago   326mb
          latest  d3311ad46d86   5 seconds ago  325MB
```

# ALPINE

```
> docker images | grep python
```

python	3.8	7f5b6ccd03e9	1 day ago	934MB
python	3.8-slim	6cf621cb1327	1 day ago	113MB
python	3.8-alpine	a24bb4013296	1 day ago	5.57MB

# ALPINE

```
FROM python:3.8-alpine
```

```
COPY requirements.txt .
```

```
RUN pip install -r requirements.txt
```

```
COPY test.py .
```

```
RUN ["python3", "test.py"]
```

# UH-OH

```
> docker build -f py38-alpine.Dockerfile -t py38-alpine
sending build context to Docker daemon 99.33kB
Step 1/5 : FROM python:3.8-alpine
--> 44fce565b2a
Step 2/5 : COPY requirements.txt .
--> 38124d1c9617
Step 3/5 : RUN pip install -r requirements.txt
--> Running in 93cb1d202652
Collecting matplotlib==3.2.1
  Downloading matplotlib-3.2.1.tar.gz (40.3 MB)
    ERROR: Command errored out with exit status 1:
    command: /usr/local/bin/python -c ...
      ...
    ERROR: Command errored out with exit status 1: python setup.py egg_info
Check the logs for full command output.
```

# WHAT'S THE DIFFERENCE

```
> FROM python:3.8
Collecting cycler==0.10.0
  Downloading cycler-0.10.0-py2.py3-none-any.whl (6.5 kB)
Collecting matplotlib==3.2.1
  Downloading matplotlib-3.2.1-cp38-cp38-manylinux1_x86_64.whl (12.4 MB)

> FROM python:3.8-alpine
Collecting cycler==0.10.0
  Downloading cycler-0.10.0-py2.py3-none-any.whl (6.5 kB)
Collecting matplotlib==3.2.1
  Downloading matplotlib-3.2.1.tar.gz (40.3 kB)
```

# WHAT'S THE DIFFERENCE

```
> FROM python:3.8
Collecting cycler==0.10.0
  Downloading cycler-0.10.0-py2.py3-none-any.whl (6.5 kB)
Collecting matplotlib==3.2.1
  Downloading matplotlib-3.2.1-cp38-cp38-manylinux1_x86_64.whl (12.4 MB)

> FROM python:3.8-alpine
Collecting cycler==0.10.0
  Downloading cycler-0.10.0-py2.py3-none-any.whl (6.5 kB)
Collecting matplotlib==3.2.1
  Downloading matplotlib-3.2.1.tar.gz (40.3 kB)
```

PYPI WHEELS ARE NOT  
COMPATIBLE WITH ALPINE

```
● ● ●
```

```
> python setup.py sdist  
> tar -xzvf dist/pandas-1.1.0.dev0+2019.g33f3fb7a4.tar.gz  
> tree -L 2 dist/pandas  
pandas  
├── __init__.py  
├── _config  
├── _libs  
│   ├── __init__.py  
│   ├── algos.c  
│   ├── algos.pxd  
│   ├── algos.pyx  
│   ├── ...  
│   └── writers.c  
     └── writers.pyx  
├── _testing.py  
├── _typing.py  
└── _version.py  
api  
arrays  
compat  
conftest.py  
core  
errors  
io  
plotting  
testing.py  
tests  
tseries  
util
```

```
● ● ●
```

```
> python setup.py bdist_wheel  
> tar -xzvf dist/pandas-1.1.0.<...>.whl  
> tree -L 2 dist/pandas  
pandas  
├── __init__.py  
├── _config  
├── _libs  
│   ├── __init__.py  
│   ├── algos.cpython-37m-darwin.so  
│   ├── ...  
│   └── writers.cpython-37m-darwin.so  
├── _testing.py  
├── _typing.py  
└── _version.py  
api  
arrays  
compat  
conftest.py  
core  
errors  
io  
plotting  
testing.py  
tests  
tseries  
util
```

WHY ARE THEY  
INCOMPATIBLE?

# RETRY: ALPINE

```
FROM python:3.8-alpine
```

```
COPY requirements.txt .
```

```
RUN apk --update --no-cache add \
    gcc build-base freetype-dev libpng-dev openblas-dev g++ musl-dev
RUN ln -s /usr/include/locale.h /usr/include/xlocale.h
```

```
COPY requirements.txt .
```

```
RUN pip install -r requirements.txt
```

```
RUN pip install cython==0.29.21 scipy==1.5.2 scikit-learn==0.22.1
```

```
COPY test.py .
```

```
RUN ["python3", "test.py"]
```

# RETRY: ALPINE

```
> docker build -f py38-alpine.Dockerfile -t py38_alpine .
```

...

Building wheel for numpy (PEP 517): Started

...

completed after 59 minutes 31 seconds

```
> docker images | grep size
```

size-py38

latest 15ecbfe3c323

3 minute ago 1.69GB

size-py38

latest d193ec6318c7

1 hour ago 1.29gb

size-py38-slim

latest 097d88af2244

1 hour ago 489mb

size-py38-slim-cache

latest 097d88af2244

1 hour ago 503mb

size-py38-slim-cache2

latest 5acbbfa56210

1 hour ago 326mb

size-py38-alpine

latest edac1e447cd6

1 minute ago 747mb

STEP 4:

REMOVE BUILD-TIME  
DEPENDENCIES

# REMOVE BUILD DEPENDENCIES

```
FROM python:3.8-alpine

COPY requirements.txt .
RUN apk --update --no-cache add openblas-dev freetype-dev
RUN apk --update --no-cache add --virtual .builddeps gcc build-base g++ musl-dev libpng-dev

RUN ln -s /usr/include/locale.h /usr/include/xlocale.h

COPY requirements.txt .
RUN pip install -r requirements.txt

RUN pip install cython==0.29.21 scipy==1.5.2
RUN pip install scikit-learn==0.22.1
RUN apk del --no-cache .builddeps

COPY test.py .
RUN ["python3", "test.py"]
```

# USE MULTI-STAGE BUILDS

```
FROM python:3.8-alpine AS builder
```

```
RUN apk --update --no-cache add openblas-dev freetype-dev  
RUN apk --update --no-cache add --virtual .builddeps gcc build-base g++ musl-dev libpng-dev  
RUN ln -s /usr/include/locale.h /usr/include/xlocale.h
```

```
COPY requirements.txt .
```

```
RUN pip install --user -r requirements.txt  
RUN pip install --user cython==0.29.21 scipy==1.5.2  
RUN pip install --user scikit-learn==0.22.1
```

```
FROM python:3.8-slim AS runner
```

```
RUN apk --update --no-cache add openblas freetype  
COPY --from=builder /root/.local /root/.local
```

# USE MULTI-STAGE BUILDS

```
> docker build -f py38-alpine-multistage.Dockerfile -t py38_alpine-multistage .
> docker images | grep size
size-py38
size-py38
size-py38-slim
size-py38-slim-cache
size-py38-slim-cache2
size-py38-alpine
size-py38-alpine-multistage latest
```

	latest	15ecbfe3c323	3 minute ago	1.69GB
size-py38	latest	d193ec6318c7	2 hours ago	1.29gb
size-py38-slim	latest	097d88af2244	2 hours ago	489mb
size-py38-slim-cache	latest	097d88af2244	2 hours ago	503mb
size-py38-slim-cache2	latest	5acbbfa56210	2 hours ago	326mb
size-py38-alpine	latest	edac1e447cd6	1 hour ago	747mb
size-py38-alpine-multistage	latest	38e7175e17d7	1 minute ago	487mb

WHY DO WE STILL NEED  
OPENBLAS AND FREETYPE?

# FIXING THE WHEEL

```
FROM python:3.8-alpine

RUN apk --update --no-cache add openblas-dev gcc build-base g++ musl-dev libpng-dev unzip patchelf

RUN pip wheel -w /root/wheels/ numpy

RUN pip install auditwheel
RUN auditwheel show /root/wheels/numpy-1.19.2-cp38-cp38-linux_x86_64.whl

> docker build .
...
The following external shared libraries are required by the wheel:
{
    "libc.musl-x86_64.so.1": "/lib/ld-musl-x86_64.so.1",
    "libgcc_s.so.1": "/usr/lib/libgcc_s.so.1",
    "libstdc++.so.6": "/usr/lib/libstdc++.so.6.0.28",
    "libopenblas.so": "/usr/lib/libopenblas-p0-52d5fd6e.3.9.so"
}
```

# FIXING THE WHEEL

```
FROM python:3.8-alpine

RUN apk --update --no-cache add openblas-dev gcc build-base g++ musl-dev libpng-dev unzip patchelf

RUN pip wheel -w /root/wheels/ numpy

RUN pip install auditwheel
RUN auditwheel repair /root/wheels/numpy-1.19.2-cp38-cp38-linux_x86_64.whl

> docker build .
...
INFO:auditwheel.main_repair:Repairing numpy-1.19.2-cp38-cp38-linux_x86_64.whl
INFO:auditwheel.wheeltools:Previous filename tags: linux_x86_64
INFO:auditwheel.wheeltools>New filename tags: linux_x86_64
INFO:auditwheel.wheeltools:Previous WHEEL info tags: cp38-cp38-linux_x86_64
INFO:auditwheel.wheeltools>No WHEEL info change needed.
INFO:auditwheel.main_repair:
Fixed-up wheel written to numpy-1.19.2-cp38-cp38-linux_x86_64.whl
```

# FIXING THE WHEEL

```
FROM python:3.8-alpine AS builder

RUN apk --update add gcc build-base freetype-dev libpng-dev openblas-dev
RUN apk --no-cache add unzip patchelf
RUN pip3 install auditwheel==3.1.0

COPY requirements.txt .
RUN pip wheel -w /root/wheels/ -r requirements.txt scipy==1.5.2
RUN find /root/wheels -name '*.whl' | xargs auditwheel repair --plat=linux_x86_64 -w /root/wheels
RUN pip install --user --no-index --find-links=/root/wheels -r requirements.txt scipy==1.5.2

RUN pip install cython==0.29.21
RUN pip wheel -w /root/wheels/ scikit-learn==0.22.1

RUN find /root/wheels -name '*scikit_learn*' | xargs auditwheel repair --plat=linux_x86_64 -w /root/wheels
RUN pip install --user --no-index --find-links=/root/wheels scikit-learn

FROM python:3.8-alpine AS runner
WORKDIR app

COPY --from=builder /root/.local /root/.local
COPY test.py .
RUN ["python3", "test.py"]
```

# FIXING THE WHEEL

```
FROM python:3.8-alpine AS builder

RUN apk --update add gcc build-base freetype-dev libpng-dev openblas-dev
RUN apk --no-cache add unzip patchelf
RUN pip3 install auditwheel==3.1.0

COPY requirements.txt .
RUN pip wheel -w /root/wheels/ -r requirements.txt scipy==1.5.2
RUN find /root/wheels -name '*.whl' | xargs auditwheel repair --plat=linux_x86_64 -w /root/wheels
RUN pip install --user --no-index --find-links=/root/wheels -r requirements.txt scipy==1.5.2

RUN pip install cython==0.29.21
RUN pip wheel -w /root/wheels/ scikit-learn==0.22.1

RUN find /root/wheels -name '*scikit_learn*' | xargs auditwheel repair --plat=linux_x86_64 -w /root/wheels
RUN pip install --user --no-index --find-links=/root/wheels scikit-learn

FROM python:3.8-alpine AS runner
WORKDIR app

COPY --from=builder /root/.local /root/.local
COPY test.py .
RUN ["python3", "test.py"]
```

# FIXING THE WHEEL

```
> docker build -f py38-alpine-wheel.Dockerfile -t py38_alpine-wheel .  
> docker images | grep size
```

size-py38	latest	15ecbfe3c323	3 minute ago	1.69GB
size-py38	latest	d193ec6318c7	3 hours ago	1.29gb
size-py38-slim	latest	097d88af2244	3 hours ago	489mb
size-py38-slim-cache	latest	097d88af2244	3 hours ago	503mb
size-py38-slim-cache2	latest	5acbbfa56210	3 hours ago	326mb
size-py38-alpine	latest	edac1e447cd6	2 hours ago	747mb
size-py38-alpine-multistage	latest	38e7175e17d7	1 hour ago	487mb
size-py38-alpine-wheel	latest	0e3b53d950d4	1 minute ago	288mb

# COMPILER FLAGS

FLAG	PURPOSE	SCIPY SIZE
		35.5M
-Os	Optimize disk space	30.5M
-g0	Remove debug information	29.2M
both		26.4M

# COMPILER FLAGS

```
RUN CFLAGS="-g0 -Os -DNDEBUG -Wl,--strip-all" \
pip wheel \
--no-cache-dir \
-w /root/wheels/ \
--global-option=build_ext \
--global-option="-j 4" \
-r requirements.txt
```

# COMPILER FLAGS

> docker build -f py38-alpine-wheel.Dockerfile -t py38_alpine-wheel .						
> docker images   grep size						
size-py38	latest	15ecbfe3c323	3 minute ago		1.69GB	
size-py38		latest	d193ec6318c7	3 hours ago	1.29gb	
size-py38-slim		latest	097d88af2244	3 hours ago	489mb	
size-py38-slim-cache		latest	097d88af2244	3 hours ago	503mb	
size-py38-slim-cache2		latest	5acbbfa56210	3 hours ago	326mb	
size-py38-alpine		latest	edac1e447cd6	2 hours ago	747mb	
size-py38-alpine-multistage	latest	38e7175e17d7	1 hour ago		487mb	
size-py38-alpine-wheel	latest	0e3b53d950d4	1 minute ago		288mb	
size-py38-alpine-compile	latest	b1fb9a4c91fc	1 minute ago		232mb	

# SHARED LIBRARIES

```
RUN cd /root/.local/lib/python3.8/site-packages && \  
rm numpy.libs/libopenblas-p0-52d5fd6e.3.9.so && \  
ln -s scipy.libs/libopenblas-p0-52d5fd6e.3.9.so \  
numpy.libs/libopenblas-p0-52d5fd6e.3.9.so
```

# SHARED LIBRARIES

```
> docker build -f py38-alpine-wheel.Dockerfile -t py38_alpine-wheel .  
> docker images | grep size
```

size-py38	latest	15ecbfe3c323	3 minute ago	1.69GB
size-py38	latest	d193ec6318c7	3 hours ago	1.29gb
size-py38-slim	latest	097d88af2244	3 hours ago	489mb
size-py38-slim-cache	latest	097d88af2244	3 hours ago	503mb
size-py38-slim-cache2	latest	5acbbfa56210	3 hours ago	326mb
size-py38-alpine	latest	edac1e447cd6	2 hours ago	747mb
size-py38-alpine-multistage	latest	38e7175e17d7	1 hour ago	487mb
size-py38-alpine-wheel	latest	0e3b53d950d4	1 minute ago	288mb
size-py38-alpine-compile	latest	b1fb9a4c91fc	1 minute ago	232mb
size-py38-alpine-shared	latest	1a65abab1254	1 minute ago	188mb

# SHOULD YOU DO THIS?

# 326 mb, < 5 min

**188 mb, >> 1 hour**

# IN CONCLUSION

- Think About Your Application's Needs
- Don't ignore `dockerignore`
- Use a Small Base Image
- Use as Few Layers As Possible
- Remove build-time dependencies