

Th.S LÊ NGỌC THẠCH

**ỨNG DỤNG
PHÂN TÍCH DỮ LIỆU
VÀ
TRÍ TUỆ NHÂN TẠO
VỚI PYTHON**

2021

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Lời nhắn

eBook "ỨNG DỤNG PHÂN TÍCH DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO VỚI PYTHON" này dự kiến phát hành vào tháng 12/2021. Bạn có thể đặt hàng ngay bây giờ với ưu đãi giảm 50% bằng 2 cách sau:

Cài **App MinePI** cho điện thoại tại theo link:

<https://minepi.com/thachln>

Sử dụng invitation code: **thachln**

Thử dùng điện thoại để đào Pi Coin. eBook được chấp nhận thanh toán Pi Coin với giá tương đương 399K đồng (Xem hình thức thanh toán tiền mặt).

Phiên bản bạn đang nhận là bản nháp trong quá trình hoàn thiện.

Bạn được gửi riêng để tham khảo hoặc để góp ý. Vì thế bạn được toàn quyền sử dụng và **KHÔNG** chia sẻ với bất kỳ ai khác nhé, **KHÔNG** lưu trữ trên internet nói chung để hạn chế đến tay người không thật sự cần nó!

Về nội dung bạn thu lượm được từ eBook dưới dạng các bài tóm tắt, đánh giá, hoặc đề nghị bổ sung thì rất được **KHUYẾN KHÍCH** chia sẻ công khai.



Đặc biệt khuyến khích bạn chia sẻ link:

<https://ThachLN.github.io>

Lê Ngọc Thạch

Hãy cài app [MinePI](https://minepi.com/thachln) ngay với Invitation Code là **thachln** để nhận ngay bản nháp (hơn 600 trang) nhé!

Hình thức thanh toán tiền mặt – Đặt hàng ngay bây với 199K, tiết kiệm 200K qua:

① MoMo	② Chuyển khoản
<div><p>Thanh toán qua MoMo</p><p>0908550642 Lê Ngọc Thạch</p><p>Nội dung tin nhắn: email sdt Ví dụ: abc@gmail.com 0908550642 AIPYTHON Email và sdt của người nhận eBook.</p><p>Trường hợp tặng bạn bè thì ghi thông tin email và sdt của bạn.</p><p>Quét mã QR thanh toán 199K.</p><p>199.000đ</p></div>	<div><p>Thanh toán qua NH Tiên Phong</p><p>Lê Ngọc Thạch, Ngân Hàng Tiên Phong, CN HCM Số tài khoản: 00002888001 Nội dung tin nhắn: email sdt AIPYTHON Vd tin nhắn: abc@gmail.com 0908456321 AIPYTHON</p><p>Quét mã QR để thanh toán cho:</p><p>Quét mã vạch này để giao dịch</p></div>

Mục lục

Quy ước	7
Ngày 1 – Chủ đề: Giới thiệu về ngôn ngữ thống kê, ngôn ngữ lập trình	10
Bài 1: Tóm tắt về thống kê (Statistics)	12
Bài 2: Ngôn ngữ lập trình Python	19
Bài 3: Ngôn ngữ Python và phần mềm Anaconda	27
Bài 4: Cài đặt thêm phần mềm	48
Bài 5: Nhập liệu, biên tập, lưu trữ dữ liệu với Python	53
Bài tập ngày 1	70
Thử thách cho bạn!	72
Ngày 2 – Chủ đề: Biểu đồ	73
Bài 6: Các loại biểu đồ	75
Bài 7: Vẽ biểu đồ trong Python	81
Bài 8: Nguyên tắc soạn biểu đồ	96
Bài 9: Giới thiệu Matplotlib	98
Bài 10: Giới thiệu Bokeh	112
Bài 11: Khai phá Bokeh	120
Ngày 3 – Phân tích mô tả	146
Bài 12: Phân tích mô tả dữ liệu Bank Marketing	148
Bài 13: Phân tích dữ liệu Marketing #2	159
Bài 14: So sánh 2 tỉ lệ	166
Bài 15: Mô hình kiểm định giả thuyết	177
Bài 16: Ứng dụng minh họa kiểm định giả thuyết	178
Bài 17: Phân tích mối tương quan	188
Ngày 4 – Chủ đề: Dữ liệu lớn	196
Bài 18: Cách xử lý tập hợp dữ liệu lớn	197
Bài 19: Sử dụng Ubuntu	232
Bài 20: Cài đặt Hadoop 3.2	241
Bài 21: Trải nghiệm Hadoop với Python	249
Ngày 5 – Chủ đề: Dự báo bằng mô hình hồi qui tuyến tính	255

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Bài 22: Giới thiệu mô hình hồi qui tuyến tính	256
Bài 23: Diễn giải mô hình hồi qui tuyến tính.....	260
Bài 24: Mô hình hồi qui tuyến tính đa biến.....	273
Bài 25: Dự báo bằng mô hình hồi qui tuyến tính	277
Ngày 6 – Chủ đề: Dự báo bằng mô hình hồi qui logistic	281
Bài 26: Giới thiệu mô hình hồi qui logistic.....	282
Bài 27: Mô hình hồi qui logistic đa biến (Multiple logistic regression model).....	286
Bài 28: So sánh mô hình.....	290
Bài 29: Dự báo bằng mô hình hồi qui logistic	296
Ngày 7 – Chủ đề: Phân tích đa biến	303
Bài 30: Xử lý giá trị trống	304
Bài 31: Mô hình phân tích phân định (Linear discriminant analysis) ..	308
Bài 32: Mô hình thành phần (Principal Component Analysis)	316
Bài 33: Mô hình phân tích cụm/nhóm (cluster analysis)	324
Ngày 8 – Chủ đề: Machine Learning	332
Bài 34: Giới thiệu Machine learning	333
Bài 35: Mô hình SVM.....	335
Bài 36: Mô hình Random Forest	343
Bài 37: Mô hình Artificial Neural Network	347
Bài 38: Machine Learning với Python Tensorflow.....	353
Ngày 9 – Chủ đề: Recommendation.....	382
Bài 39: Giới thiệu phương pháp gợi ý Collaborative filtering	383
Bài 40: Triển phương pháp gợi ý Collaborative filtering bằng R	393
Ngày 10 – Chủ đề: Natural Language Processing.....	399
Bài 41: Các kỹ thuật cơ bản	400
Bài 42: Trích đặc trưng (Feature extraction).....	405
Bài 43: Giới thiệu ứng dụng phân tích cảm xúc (Sentiment Analysis) ..	415
Bài 44: Giới thiệu ứng dụng phân tích từ vựng (Word Embedding) ...	426
Bài 45: Giới thiệu ứng dụng xác định chủ đề (Topic Modeling)	438
Ngày 11 – Chủ đề: Computer Vision	449
Bài 46: Giới thiệu Face recognition	450

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Bài 47: Giới thiệu mô hình CNN	465
Ngày 12 – Chủ đề: Nhận diện tiếng nói (Speech Recognition)	484
Bài 48: Giới thiệu đặc trưng của âm thanh	485
Bài 49: Các thao tác cơ bản với file âm thanh	491
Bài 50: Mô hình Chuyển giọng nói thành văn bản	495
Ngày 13 – Chủ đề: Phân tích dữ liệu theo trường phái Bayes	498
Bài 51: Nhập môn.....	499
Tạm kết thúc	499
Phụ lục	500
Quan sát giao dịch cổ phiếu VNM (Vinamilk)	501
Đọc và vẽ tín hiệu âm thanh.....	511
Tải sách nói “Từ tốt đến vĩ đại”	514
Đọc ảnh y khoa DiCOM.....	517
Áp dụng biến đổi Fourier cho ảnh.....	520
Sử dụng Git.....	524
Khảo sát ảnh và ma trận	553
Phát triển ứng dụng với Python.....	555
Xử lý file pdf	562
Khảo sát file âm thanh.....	570
Phân tích âm thanh với thư viện mutagen	573
Khám phá Python trong WSL2	574
Crawl dữ liệu bằng Selenium	576
Sử dụng OpenCV để phân tích dữ liệu ảnh.....	577
Cài đặt OpenCV	578
Đóng gói chương trình Python	579
Tải file video từ Youtube	582
Sinh code Restful API từ database	583
Trải nghiệm Restful API với Flask	585
Trải nghiệm Kafka.....	586
Trải nghiệm Apache NiFi.....	589
Giới thiệu superset.....	594
Cài đặt.....	595

Đây là eBook của riêng bạn – đề nghị không chia sẻ cho ai khác nhé!

Khởi động lại superset.....	600
Cài đặt và cấu hình Nginx	601
Truy cập nginx trên Ubuntu từ máy Windows.....	602
Khai thác superset	604
Cài đặt Ubuntu Server 20.04.2	608
Superset introduction.....	609
Giới thiệu PowerBI.....	609
Import data	612
Tạo biểu đồ.....	615
Bài bổ sung	617
Bài 101: Trải nghiệm ứng dụng với Flask	618
Bài 102: Xử lý dữ liệu	624

Quy ước

Một số nội dung trong tài liệu được trình bày với các định dạng khác nhau thì có ý nghĩa của nó, bạn đọc nên nắm thông tin này để tiện theo dõi.

Mã nguồn

Mã lệnh được viết và đóng khung với font chữ Courier New như sau:

```
print('Xin chào!')
print('Welcome!')
print('{} + {} = {}'.format(1, 2, (1 + 2)))
print('%d + %d = %d' % (1, 1, 4))
```

Bạn có thể sao chép và dán (đôi khi trong tài liệu viết luôn tiếng Anh: copy & paste) vào phần mềm để chạy.

Kết quả của lệnh, tùy theo phần mềm bạn sử dụng để chạy mã nguồn thì kết quả sẽ hiển thị ở các vị trí khác nhau. Phần văn bản kết xuất của phần mềm sẽ được trình bày theo khung màu đỏ gạch bên dưới:

```
Xin chào độc giả của ebook Chạm tới AI trong 10 ngày.
welcome to ebook Touch on AI in ten days.
1 + 1 = 4
```

Lệnh thực thi trong hệ điều hành

Trường hợp các lệnh thực thi trong môi trường hệ điều hành (phân biệt với các lệnh, hoặc mã nguồn của chương trình thực thi trong môi trường của R hoặc Python như RStudio hoặc Spyder như đã qui ước ở mục Mã nguồn) thì dấu hiệu như sau:

Đối với lệnh thực thi trong dấu nhắc lệnh của Anaconda hoặc trong cửa sổ lệnh CMD của Windows, hoặc trong Terminal của Linux/MacOS thì khung màu vàng có 2 vạch đậm ở cạnh trái và phải như sau:

```
pip install python-docx
```

Cặp dấu nháy

Các dữ liệu dạng chuỗi (string, text, char nói chung là có nghĩa giống nhau trong Python) được bao đóng trong **dấu nháy đơn** hoặc **dấu nháy đôi**. Trên bàn phím máy tính thì dấu **nháy trái** và **phải** là giống nhau. Tuy nhiên trong phần mềm soạn thảo văn bản như Microsoft Word thì gập dấu nháy đơn và đôi được thay thế bằng ‘, “” để tăng tính thẩm mỹ. Các dấu nháy thẩm mỹ này khác với kí tự ' và " trên bàn phím (phím bên trái phím Enter).

Đôi khi bạn copy & paste mã nguồn vào các phần mềm như Microsoft Word thì các dấu nháy có thể bị “trang trí” lại như trên. Vì vậy khi copy mã nguồn từ Microsoft vào các phần mềm chạy R hoặc Python thì hãy thay thế lại cho đúng.

Ứng dụng Phân tích dữ liệu và Trí tuệ nhân tạo với Python

Một qui ước khác liên quan đến dấu nhảy đôi là khi dùng trong văn bản để bao đóng danh từ riêng, hoặc lệnh như: *Bạn hãy thử gõ lệnh “exit()” trong cửa sổ console để thoát chương trình Python.* Trong câu hướng dẫn này thì lệnh `exit()` được gõ vào console **KHÔNG** bao gồm cặp dấu nhảy.

Cách viết thông tin lặp lại với dấu ba chấm

Khi cần mô tả một lệnh có nhiều thông tin lặp lại thì dùng dấu ba chấm như ví dụ sau.

Khi cần mô tả hàm xóa cột dữ liệu trong tham số thứ nhất của hàm `drop` như:

```
df.drop(['cột 1', 'cột 2', ...], axis =1)
```

thì phần in đậm có nghĩa là có thể gồm 1 hoặc nhiều tên cột dữ liệu. Ví dụ lệnh sau có nghĩa là xóa cột `Fullname` khỏi DataFrame `df`.

```
df.drop(['Fullname'], axis =1)
```

Hoặc lệnh sau sẽ xóa 2 cột `Fullname` và `Year` khỏi DataFrame `df`:

```
df.drop(['Fullname', 'Year'], axis =1)
```

Kí hiệu optional (không bắt buộc)

Khi sử dụng hàm số thì có nhiều tham số (argument, parameter) không bắt buộc (optional) thì sử dụng cặp dấu ngoặc vuông `[]`. Ví dụ hàm `plot` bên dưới không bắt buộc tham số `x` và `format`:

```
plot([x], y, [format])
```

Cách viết in nghiêng cho các biến

Thông thường các biến được mô tả trong các câu lệnh sẽ để trong cặp dấu ngoặc nhọn `<>`. Ví dụ lệnh sau có nghĩa là khi gõ lệnh bạn phải thay nội dung *<tên cột>* thành tên cột cụ thể trong data frame của bạn:

```
df[df['<tên cột>'].notnull()]
```

Trong tài liệu này đôi lúc sẽ không dùng cặp dấu ngoặc nhọn để mô tả lệnh chung như sau:

```
df[df['tên cột'].notnull()]
```

Cách viết trình tự bấm chọn menu

Khi cần trình bày thứ tự các nút bấm, hoặc các mục cần bấm trong các thao tác thì sẽ dùng dấu lớn hơn `>`. Ví dụ khi hướng dẫn bạn vào trang web

“<https://github.com/vncorenlp/VnCoreNLP>”, bấm vào nút “Clone”, sau đó bấm tiếp vào nút hoặc link “Download Zip” thì sẽ viết gọn như sau:

Bấm vào nút Clone > nút Download Zip, hoặc nút Clone > Download Zip.

Các từ tiếng Anh viết tắt thường xuyên được sử dụng trong sách

AI: Artificial Intelligent - **Trí thông minh nhân tạo**. Nhiều người dịch là Trí Tuệ Nhân Tạo. Trong sách này tôi muốn dùng đúng nghĩa Intelligent có nghĩa là Trí thông minh thôi vì khoảng cách từ Thông Minh đến Tuệ thì rất rất là xa. Trí thông minh nhân tạo tôi cho là phù hợp nhất trong bối cảnh hiện nay. Có thể bạn và cả tôi quen với cách đọc Trí Tuệ Nhân Tạo vừa gọn và vừa sang. Tuy nhiên nếu khi cần nói thì vẫn nên dùng từ “Thông minh” để phản ánh đúng mức độ của nó để mà còn phân đầu đến mức “Tuệ”. Đẳng nào thì tôi cũng viết là AI thay vì viết tiếng Việt nên chắc không nhầm lẫn.

Đường dẫn thư mục (Path)

Trong Windows thì dấu cách thư mục là dấu xuyệt trái (back slash). Ví dụ: D\ai2021\data.

Tuy nhiên ngôn ngữ R hoặc Python được thiết kế tương thích với các hệ điều hành khác như Macintosh, Linux. Các hệ điều hành thì dùng dấu xuyệt phải (right slash) để phân cách thư mục. Ví dụ: /mnt/d/ai2021.

Vì vậy khi trình bày đường dẫn thư mục trong câu văn thì đôi lúc dùng \, hoặc đôi lúc dùng / do dữ liệu được minh họa trên Windows hoặc Linux.

Nhưng trong mã nguồn (R hoặc Python) thì đều thống nhất là dùng dấu xuyệt phải / như sau:

```
read.csv("D:/ai2021/data/test.csv")
```

Trong Windows, code R hoặc Python có một cách khác là dùng hai (double) dấu \. Ví dụ:

```
read.csv("D:\\ai2021\\data\\test.csv")
```

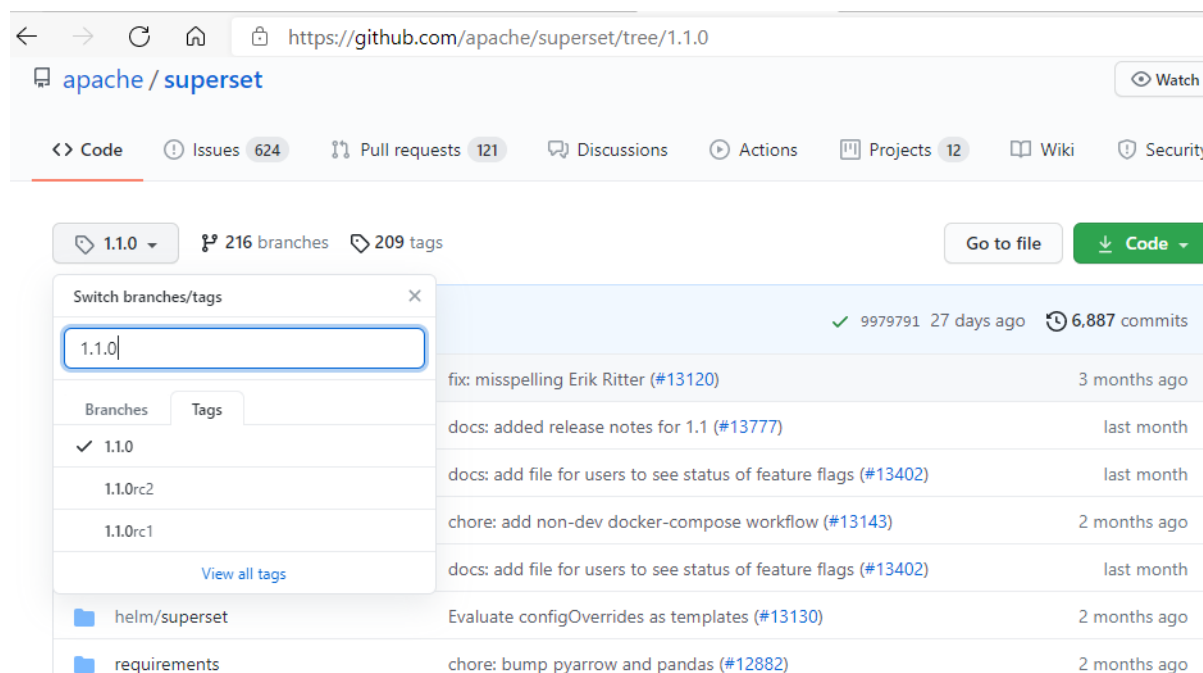
Tuy nhiên code này không tương thích trong Python trên Linux và cả MacOS nên **không** khuyến khích dùng.

Giới thiệu superset

Bài này giúp bạn khai thác dự án <https://github.com/apache/superset> để làm Data Visualization.

Cài đặt

Dưới đây là tóm tắt quá trình tôi cài đặt superset từ source code phiên bản 1.1.0 trên <https://github.com/apache/superset>, tag 1.1.0.



Tài liệu tham khảo chính thức ở đây:

<https://github.com/apache/superset/blob/master/CONTRIBUTING.md#fask-server>

Tuy nhiên trong quá trình cài đặt có phát sinh vài lỗi và phải giải quyết bằng cách bổ sung thêm vài thư viện (được **bôi vàng**).

Yêu cầu

Ubuntu 20: Nếu bạn định dùng máy ảo trên Windows để cài Ubuntu thì tham khảo [Bài 14 trong Ngày 4](#).

Trong trường hợp bạn dùng Ubuntu thì có khả năng Python được cài là 3.6.x. Bạn cần nâng cấp lên Python 3.8.x. Phiên bản Python tại thời điểm phần này được viết là 3.8.10. Hãy làm theo tài liệu sau:

<https://linuxize.com/post/how-to-install-python-3-8-on-ubuntu-18-04/>

Tiếp tục thực hiện các lệnh sau để backup file /usr/bin/python3 và ánh xạ python3.8 thành python3.

```
sudo cd /usr/bin
sudo mv ./python3 ./python3.bak
```

```
sudo ln -nsf ./python3.8 python3
```

Kiểm tra Python

Mặc định khi cài Ubuntu thì Python phiên bản 3 được cài. Kiểm tra lại bằng lệnh sau:

```
python3 -V
```

```
Python 3.8.5
```

Cài đặt back-end

Thực hiện các lệnh sau:

```
mkdir ~/projects && cd ~/projects

echo Checkout source code version 1.1.0
git clone https://github.com/apache/superset
cd superset
git checkout -b 1.1.0

export VENV_NAME=venv-superset
echo Create a new virtual environment with name: $VENV_NAME
sudo apt install python3-virtualenv
virtualenv $VENV_NAME
source $VENV_NAME/bin/activate

# Upgrade pip
sudo apt install python3-pip
python3 -m pip install --upgrade pip
sudo apt-get install build-essential libmysqlclient-dev
python3-dev python3-dev libsasl2-dev
pip3 install -r requirements/local.txt

pip3 install -e .
superset fab create-admin
superset db upgrade
superset init
superset load-examples
FLASK_ENV=development superset run -p 8088 --with-threads --
reload -debugger
```

Nếu bạn dùng Ubuntu 18 thì có thể thay dòng màu đỏ ở trên bằng 2 lệnh sau:

```
pip3 install virtualenv
sudo virtualenv $VENV_NAME
sudo apt-get install python-dev python3-dev libpython3.8-dev
```

Lệnh cuối cùng ở trên là để khởi động superset với port 8088. Bạn có thể truy cập <http://localhost:8088> ngay trong Ubuntu. Tuy nhiên lúc này superset chỉ có phần backend nên sẽ không có giao diện đầy đủ. Hãy nhấn Ctrl + C để tắt chương trình superset và chuyển sang phần [Cài đặt front-end](#) bên dưới.

Ghi chú

Nếu vì lý do nào đó mà quá trình làm việc của bạn bị gián đoạn, phải login lại thì thực hiện lệnh sau sau khi login để thực hiện tiếp các lệnh:

```
cd ~/projects/superset
export VENV_NAME=venv-superset
source $VENV_NAME/bin/activate
```

Cài đặt front-end

Chuẩn bị

- NodeJS
- npm

Một trong các cách chuẩn bị NodeJs và npm là dùng nvm (<https://github.com/nvm-sh/nvm>) với lệnh sau:

```
curl -o- https://raw.githubusercontent.com/nvm-sh/nvm/v0.37.0/install.sh | bash
```

Logout và login lại. Sau đó chạy lại các lệnh sau:

```
cd ~/projects/superset
export VENV_NAME=venv-superset
source $VENV_NAME/bin/activate
```

Chuẩn bị môi trường nodejs và npm:

```
cd superset-frontend
nvm install
nvm use
```

Cài đặt thư viện

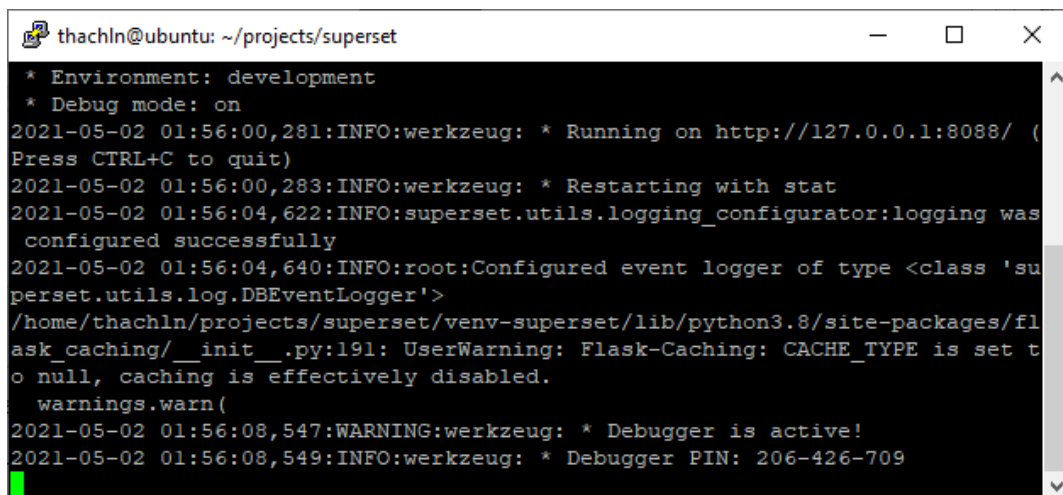
```
cd ~/projects/superset/superset-frontend
# Install dependencies from `package-lock.json`
npm ci
```

Build assets

```
npm run build
```

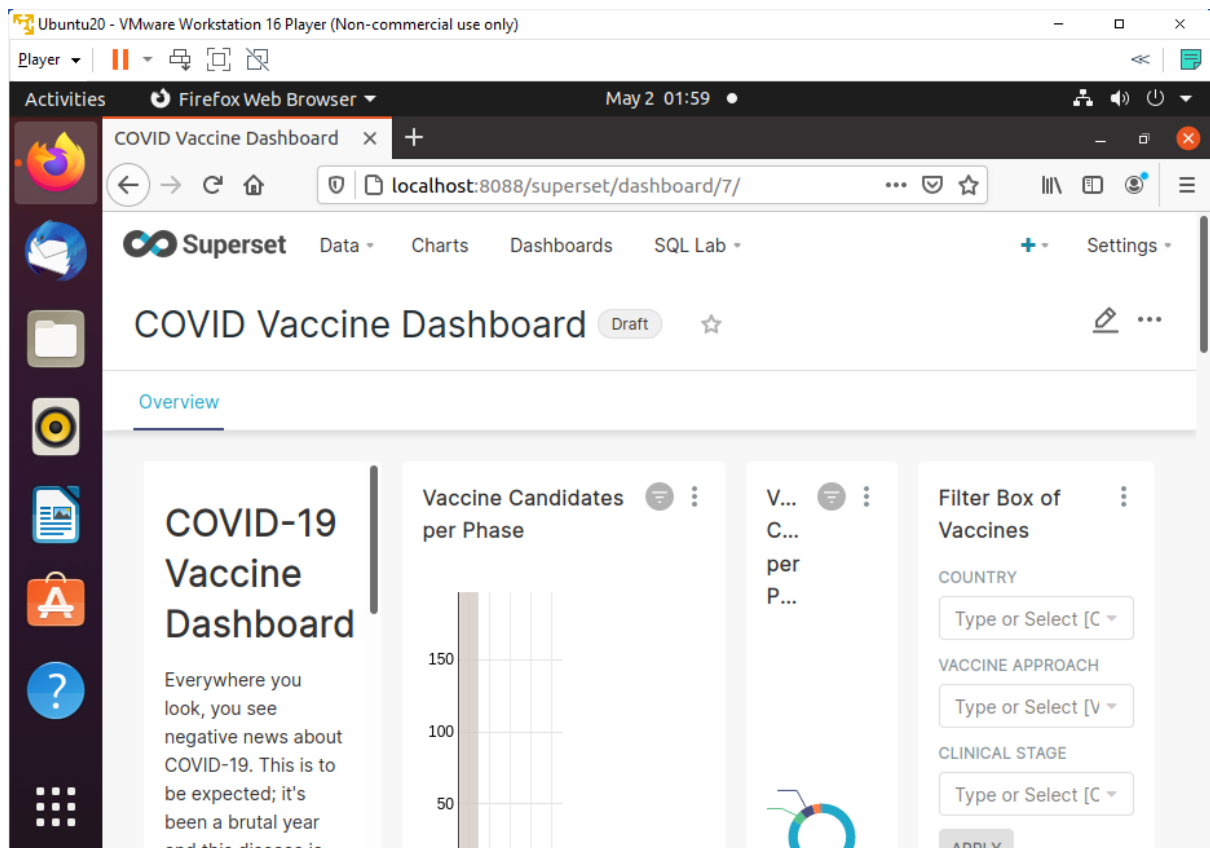
Sau khi biên dịch xong hãy khởi động lại superset bằng lệnh sau:

```
FLASK_ENV=development superset run -p 8088 --with-threads --
reload --debugger
```



```
thachln@ubuntu: ~/projects/superset
* Environment: development
* Debug mode: on
2021-05-02 01:56:00,281:INFO:werkzeug: * Running on http://127.0.0.1:8088/ (
Press CTRL+C to quit)
2021-05-02 01:56:00,283:INFO:werkzeug: * Restarting with stat
2021-05-02 01:56:04,622:INFO:superset.utils.logging_configurator:logging was
configured successfully
2021-05-02 01:56:04,640:INFO:root:Configured event logger of type <class 'su
perset.utils.log.DBEventLogger'>
/home/thachln/projects/superset/venv-superset/lib/python3.8/site-packages/fl
ask_caching/__init__.py:191: UserWarning: Flask-Caching: CACHE_TYPE is set t
o null, caching is effectively disabled.
  warnings.warn(
2021-05-02 01:56:08,547:WARNING:werkzeug: * Debugger is active!
2021-05-02 01:56:08,549:INFO:werkzeug: * Debugger PIN: 206-426-709
```

Nếu Ubuntu của bạn cài đặt ở chế độ có giao diện thì hãy đăng nhập và trải nghiệm superset với link <http://localhost:8088> ngay trên Ubuntu. Thông tin đăng nhập do bạn thiết lập trong quá trình cài đặt back-end.



Khởi động lại superset

Vì tôi chạy Ubuntu với VMware trên Windows khi cần sử dụng superset thì bật máy lên và thực thi các lệnh sau:

```
cd ~/projects/superset
export VENV_NAME=venv-superset
source $VENV_NAME/bin/activate
FLASK_ENV=development superset run -p 8088 --with-threads --
reload --debugger
```


Cài đặt và cấu hình Nginx

Phần trước bạn đã biết cách cài đặt và khai thác superset qua URL <http://localhost:8088> ngay trên máy Ubuntu đã cài superset. Tình huống đặt ra cho bạn là:

① Nếu bạn muốn truy ứng dụng superset thông qua web server như nginx với URL như <http://localhost> thì phải làm sao?

② Mở rộng một chút ý ở trên: Nếu bạn muốn truy cập từ một máy khác thông qua địa chỉ IP của máy đang chạy Ubuntu thì làm sao? Cụ thể trong trường hợp này là nếu dùng trình duyệt từ máy Windows để truy cập vào superset trên máy chạy Ubuntu thì làm sao?

Một trong cách giải quyết cho hai tình huống trên là sử dụng Nginx (như đã gợi ý trong tình huống ①).

Cài đặt Nginx

Chạy lệnh sau trong Ubuntu:

```
sudo apt install nginx
```

Cấu hình Nginx

Chỉnh sửa file cấu hình mặc định của nginx bằng lệnh sau:

```
sudo nano /etc/nginx/sites-available/default
```

```
#         location / {
#             # First attempt to serve request as file, then
#             # as directory, then fall back to displaying a 404.
#             try_files $uri $uri/ =404;
#         }
#         location / {
#             proxy_pass http://127.0.0.1:8088;
#             include /etc/nginx/proxy_params;
#         }
```

Phần **màu đỏ** là chú thích lại (thêm dấu # phía trước mỗi dòng).

Phần **màu xanh** là thêm mới.

Nhấn phím Ctrl+O để lưu file. Ctrl+X để thoát trình soạn thảo nano.

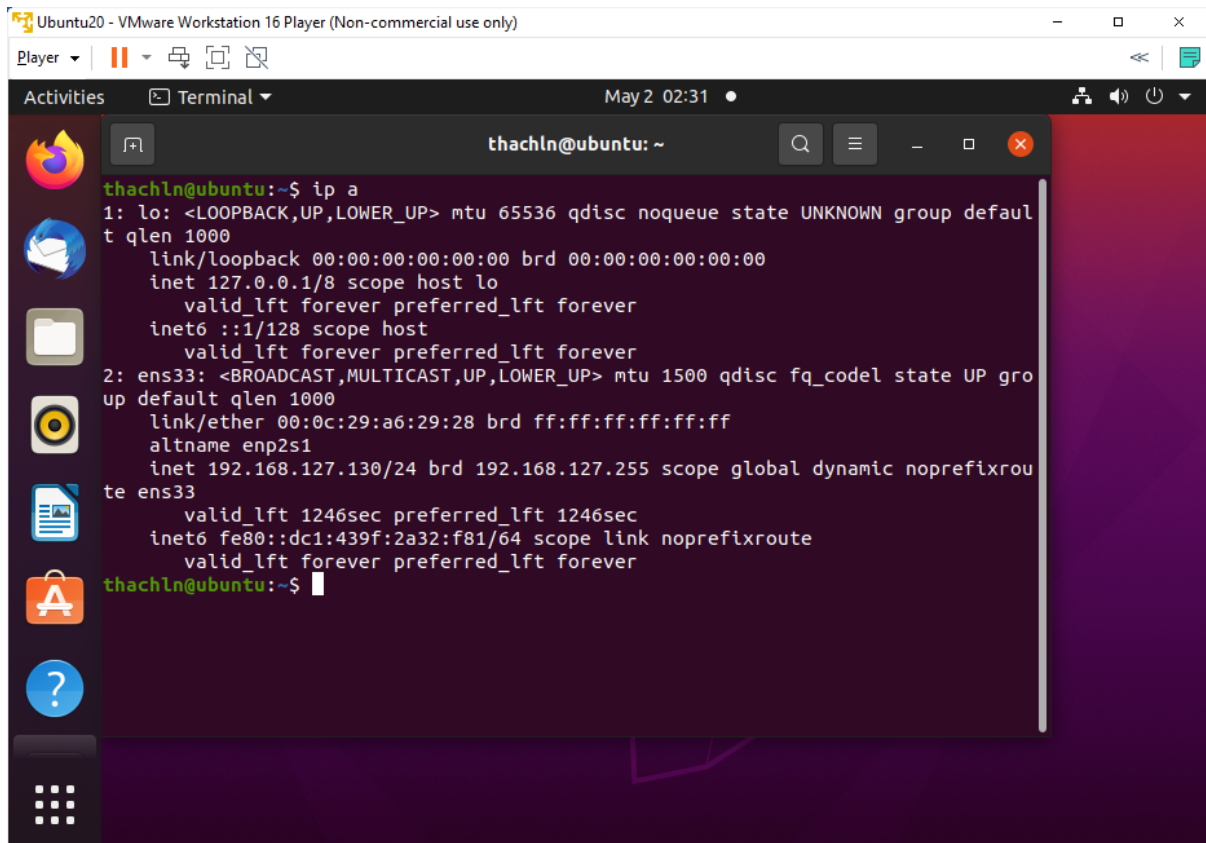
Khởi động lại nginx

```
sudo service nginx restart
```

Truy cập nginx trên Ubuntu từ máy Windows

Xem địa chỉ IP của máy Ubuntu bằng lệnh sau:

```
ip a
```

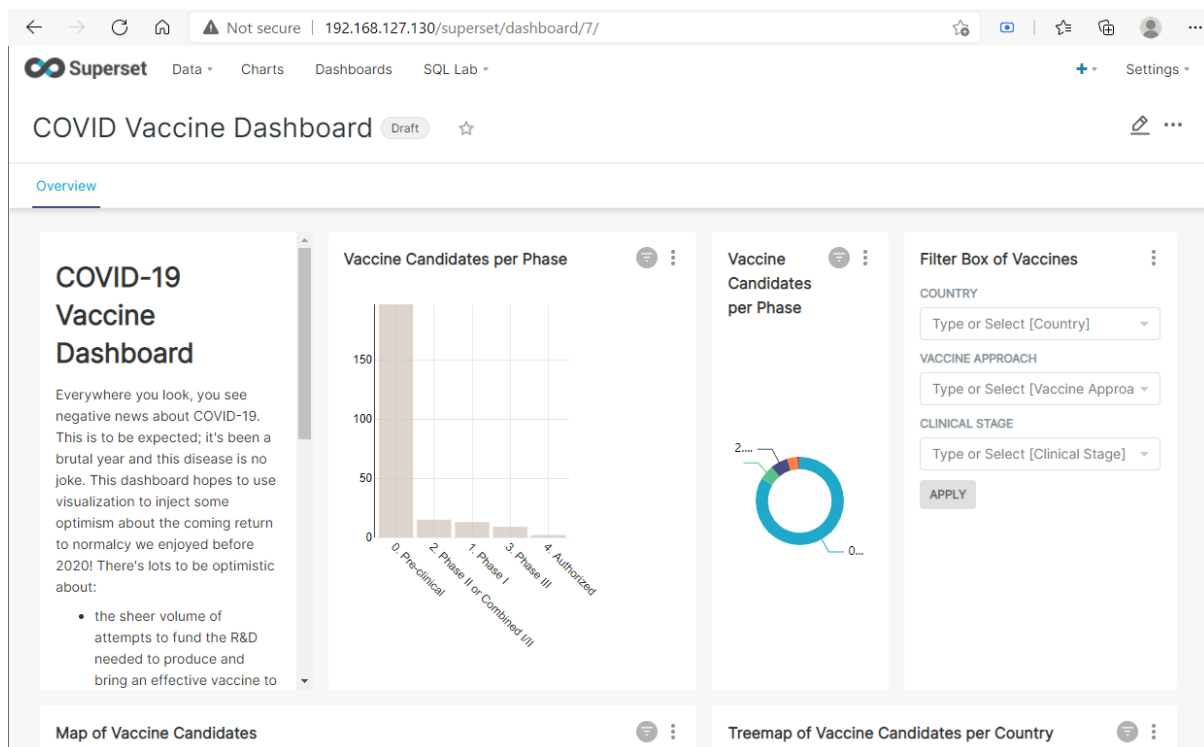


The screenshot shows a terminal window titled 'thachln@ubuntu: ~' within a VMware Workstation 16 Player. The terminal displays the output of the 'ip a' command, showing details for the loopback interface 'lo' and the ethernet interface 'ens33'. The IP address for 'ens33' is 192.168.127.130.

```
thachln@ubuntu:~$ ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: ens33: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default qlen 1000
    link/ether 00:0c:29:a6:29:28 brd ff:ff:ff:ff:ff:ff
    altname enp2s1
    inet 192.168.127.130/24 brd 192.168.127.255 scope global dynamic noprefixroute ens33
        valid_lft 1246sec preferred_lft 1246sec
    inet6 fe80::dc1:439f:2a32:f81/64 scope link noprefixroute
        valid_lft forever preferred_lft forever
thachln@ubuntu:~$
```

Ubuntu trên máy tôi có địa chỉ IP là: 192.168.127.130.

Vì vậy đường dẫn để trải nghiệm superset là <http://192.168.127.130>



Khai thác superset

Cài đặt MySQL

[Tham khảo cài đặt MySQL trên Ubuntu tại đây.](#)

Tạo database

Thực hiện 3 lệnh sau trong cửa sổ lệnh MySQL:

```
CREATE DATABASE mysuperset DEFAULT CHARACTER SET 'UTF8MB4';  
CREATE USER 'mysuperset_user'@'%' IDENTIFIED BY  
'Mysuperset#123';  
GRANT ALL PRIVILEGES ON mysuperset.* TO 'mysuperset_user'@'%'  
WITH GRANT OPTION;
```

Thử kết nối database bằng mysqlclient

Thực hiện lệnh sau trong cửa sổ lệnh của hệ điều hành (Ubuntu):

```
mysql -u mysuperset_user -pMysuperset#123 mysuperset
```

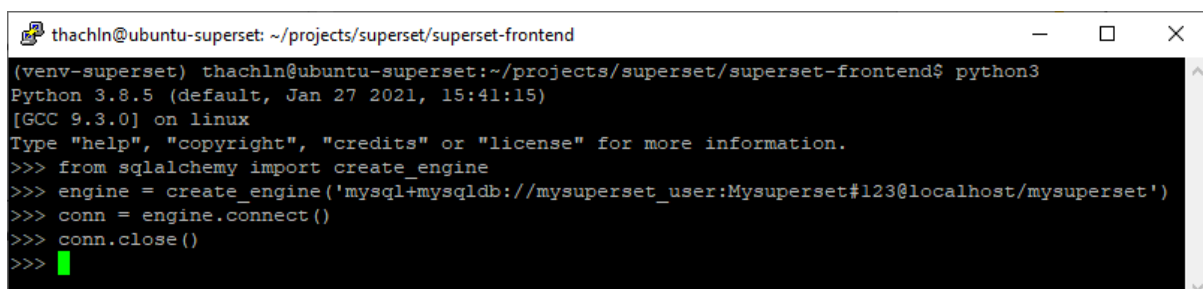
Thử test nối database bằng Python ngay trên máy Ubuntu

Mở cửa sổ lệnh Python bằng lệnh sau:

```
python3
```

Thực hiện các lệnh sau trong cửa sổ lệnh Python:

```
from sqlalchemy import create_engine  
engine =  
create_engine('mysql+mysqldb://mysuperset_user:Mysuperset#123@192.168.  
127.131/mysuperset?charset=utf8mb4')  
conn = engine.connect()  
conn.close()
```



```
thachln@ubuntu-superset: ~/projects/superset/superset-frontend  
(venv-superset) thachln@ubuntu-superset:~/projects/superset/superset-frontend$ python3  
Python 3.8.5 (default, Jan 27 2021, 15:41:15)  
[GCC 9.3.0] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> from sqlalchemy import create_engine  
>>> engine = create_engine('mysql+mysqldb://mysuperset_user:Mysuperset#123@localhost/mysuperset')  
>>> conn = engine.connect()  
>>> conn.close()  
>>>
```

Thử kết nối database bằng Python trên máy Windows

Cấu hình MySQL Server

Mặc định MySQL được cấu hình chỉ chạy trên localhost, tức địa chỉ IP là 127.0.0.1. Để MySQL được truy cập từ máy khác trong mạng thì sửa file cấu hình bằng lệnh sau:

```
sudo nano /etc/mysql/mysql.conf.d/mysqld.cnf
```

Tìm dòng **bind-address** sửa lại như sau:

```
bind-address            = 0.0.0.0
```

Sau đó khởi động lại mysql server bằng lệnh:

```
sudo service mysql restart
```

Cài thư viện cho môi trường python (trên client)

```
pip install mysqlclient
```

Mã nguồn Python

```
from sqlalchemy import create_engine
engine =
create_engine('mysql+mysqldb://mysuperset_user:Mysuperset#123@192.168.
127.131/mysuperset')
engine.connect()
engine.close()
```

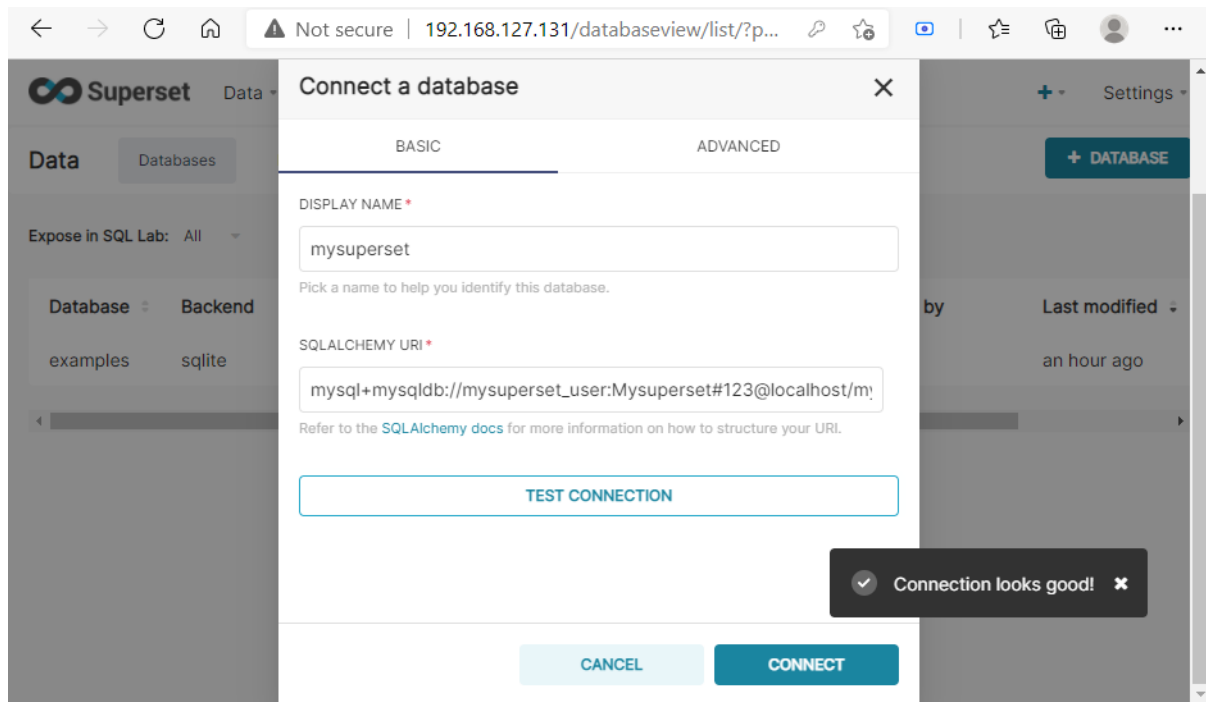
(Hãy thay địa chỉ IP bằng địa chỉ IP đúng trên máy Ubuntu của bạn)

Tạo Database trong superset

Tạo database với 2 thông số sau:

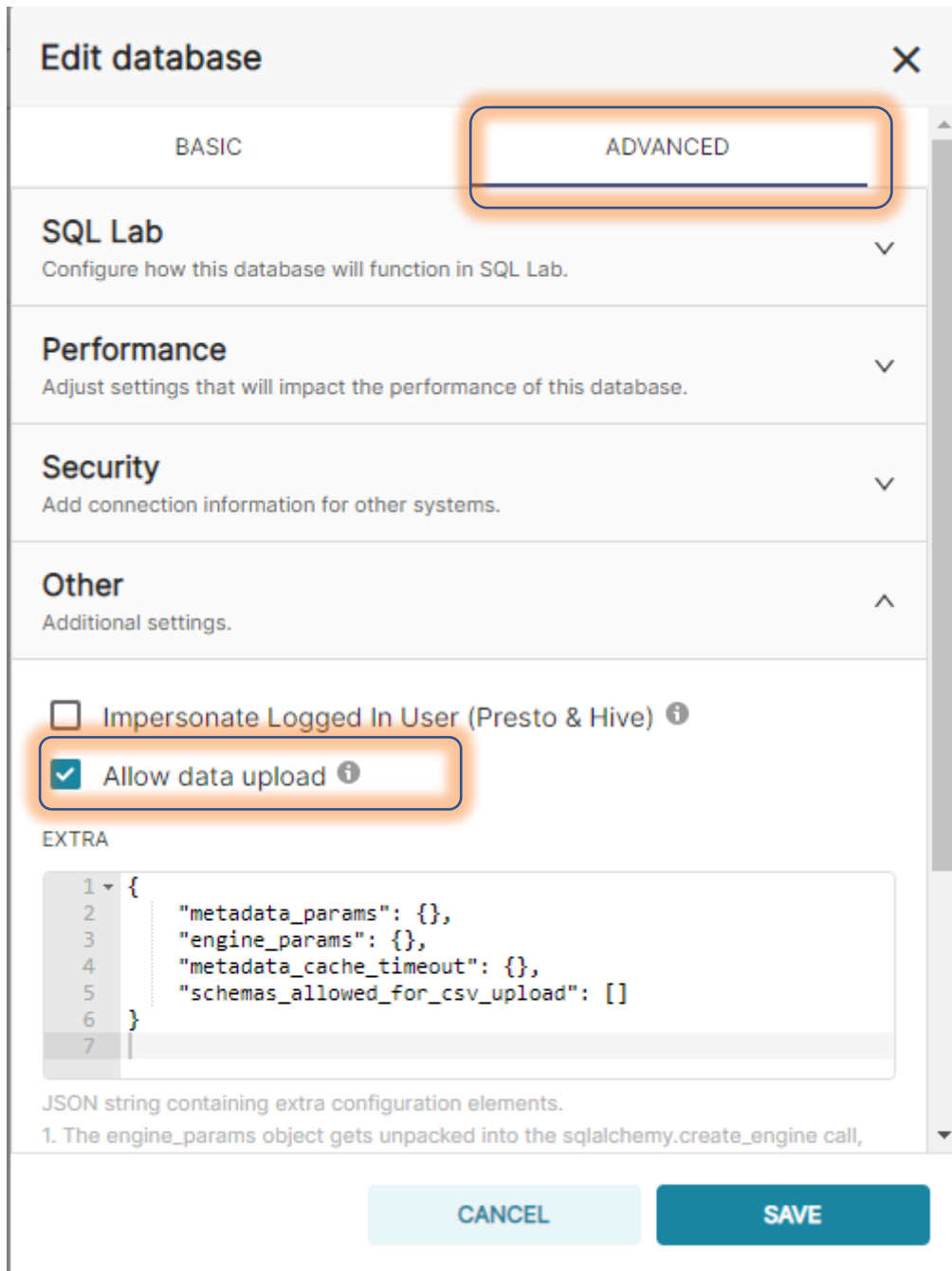
DISPLAY NAME	mysuperset (<i>bạn có thể đặt tên khác</i>)
SQL ALCHEM Y URI	mysql+mysqldb://mysuperset_user:Mysuperset#123@localhost/mysu perset

Bấm nút “TEST CONNECTION” để kiểm tra kết nối.



Thử bậc “Allow data upload”

Vào tab “ADVANCED”, đánh dấu vào mục “Allow data upload”.



Edit database [X]

BASIC ADVANCED

SQL Lab ✓
Configure how this database will function in SQL Lab.

Performance ✓
Adjust settings that will impact the performance of this database.

Security ✓
Add connection information for other systems.

Other ^
Additional settings.

☐ Impersonate Logged In User (Presto & Hive) ⓘ

☒ Allow data upload ⓘ

EXTRA

```
1 {  
2   "metadata_params": {},  
3   "engine_params": {},  
4   "metadata_cache_timeout": {},  
5   "schemas_allowed_for_csv_upload": []  
6 }  
7
```

JSON string containing extra configuration elements.
1. The engine_params object gets unpacked into the sqlalchemy.create_engine call,

CANCEL SAVE

Nếu bạn không gặp lỗi gì thì coi như đã rất may mắn 😊!