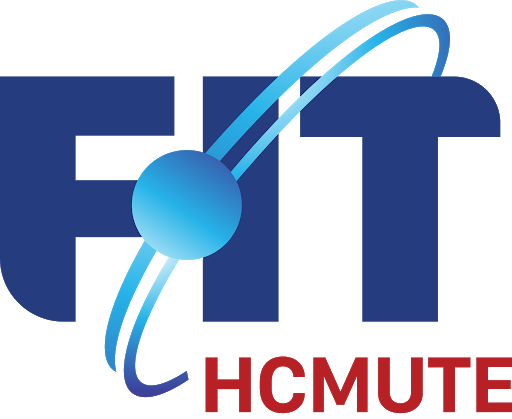
**.**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN ĐIỆN TOÁN ĐÁM MÂY**

-----🙞🙜🕮🙞🙜-----

****

**Nguyễn Phạm Duy Khiêm – 19133027**

**Ngô Phi Lít - 18133024**

**Đề Tài:**

**TÌM HIỂU VÀ SỬ DỤNG APACHE SUPERSET**

**GIẢNG VIÊN HƯỚNG DẪN**

**TS. HUỲNH XUÂN PHỤNG**

**KHÓA 2019 - 2023**

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP. HCM**  **KHOA CNTT**  \*\*\*\*\*\*\* | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  \*\*\*\*\*\*\* |

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Họ và tên Sinh viên 1: **Nguyễn Phạm Duy Khiêm** MSSV 1: **19133027**

Họ và tên Sinh viên 2: **Ngô Phi Lít** MSSV 2: **18133024**

Ngành: **Công nghệ Thông tin**

Tên đề tài: **TÌM HIỂU VÀ SỬ DỤNG APACHE SUPERSET**

Họ và tên Giáo viên hướng dẫn: **TS. Huỳnh Xuân Phụng**

**NHẬN XÉT**

1. Về nội dung đề tài & khối lượng thực hiện:

2. Ưu điểm:

3. Khuyết điểm:

4. Đề nghị cho bảo vệ hay không?

5. Đánh giá loại:

6. Điểm:

|  |  |  |  |
| --- | --- | --- | --- |
| *Tp. Hồ Chí Minh, ngày tháng năm 2021* | | | |
|  | | Giáo viên hướng dẫn  *(Ký & ghi rõ họ tên)* |
| **ĐH SƯ PHẠM KỸ THUẬT TP. HCM**  **KHOA CNTT**  \*\*\*\*\*\*\* | | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh Phúc**  \*\*\*\*\*\*\* | | | |

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên Sinh viên 1: **Nguyễn Phạm Duy Khiêm** MSSV 1: **19133027**

Họ và tên Sinh viên 2: **Ngô Phi Lít** MSSV 2: **18133024**

Ngành: **Công nghệ Thông tin**

Tên đề tài: **TÌM HIỂU VÀ SỬ DỤNG APACHE SUPERSET**

Họ và tên Giáo viên phản biện: **TS. Huỳnh Xuân Phụng**

**NHẬN XÉT**

1. Về nội dung đề tài & khối lượng thực hiện:

2. Ưu điểm:

3. Khuyết điểm:

4. Đề nghị cho bảo vệ hay không?

5. Đánh giá loại:

6. Điểm:

*Tp. Hồ Chí Minh, ngày tháng năm 2021*

Giáo viên phản biện

(Ký & ghi rõ họ tên)

# **LỜI CẢM ƠN**

*Lời đầu tiên nhóm xin phép được gửi lời cảm ơn chân thành và sâu sắc nhất đến với Khoa Công Nghệ Thông Tin – Trường Đại Học Sư Phạm Kỹ Thuật Thành Phố Hồ Chí Minh đã tạo điều kiện cho nhóm chúng em được học tập, phát triển nền tảng kiến thức sâu sắc và thực hiện đề tài này.*

*Bên cạnh đó nhóm chúng em xin gửi đến thầy Huỳnh Xuân Phụng lời cảm ơn sâu sắc nhất. Trải qua một quá trình dài học tập và thực hiện đề tài trong thời gian qua. Thầy đã tận tâm chỉ bảo nhiệt tình nhóm chúng em trong suốt quá trình từ lúc bắt đầu cũng như kết thúc đề tài này.*

*Với sự hướng dẫn nhiệt tình, giảng dạy tận tình đầy đủ kiến thức của thầy Huỳnh Xuân Phụng, chúng em đã học tập và hiểu được những kiến thức cơ bản về Apache Super.et– một ứng dụng phần mềm mã nguồn mở để khám phá dữ liệu và trực quan hóa dữ liệu có thể xử lý dữ liệu ở quy mô petabyte. Qua đó tụi em biết cách cài đặt và sử dụng Apache Superset.*

*Tuy nhiên lượng kiến thức là vô tận và với khả năng hạn hẹp chúng em đã rất cố gắng để hoàn thành một cách tốt nhất. Chính vì vậy việc xảy ra những thiếu sót là điều khó có thể tránh khỏi. Chúng em hi vọng nhận được sự góp ý tận tình của quý thầy (cô) qua đó chúng em có thể rút ra được bài học kinh nghiệm và hoàn thiện và cải thiện nâng cấp lại sản phẩm của mình một cách tốt nhất có thể.*

*Chúng em xin chân thành cảm ơn!*

***Nhóm thực hiện***

*Nguyễn Phạm Duy Khiêm 19133027*

*Ngô Phi Lít 18133024*

**MỤC LỤC**

[LỜI CẢM ƠN 4](#_Toc104420318)

[MỤC LỤC 5](#_Toc104420319)

[PHẦN MỞ ĐẦU 7](#_Toc104420320)

[**1. Tính cấp thiết của đề tài** 7](#_Toc104420321)

[**2. Đối tượng nghiên cứu** 7](#_Toc104420322)

[**3. Phạm vi nghiên cứu** 7](#_Toc104420323)

[**4. Kết quả dự kiến đạt được** 7](#_Toc104420324)

[PHẦN NỘI DUNG 8](#_Toc104420325)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VỀ APACHE SUPERSET 8](#_Toc104420326)

[**1.1** **Tổng quan về APACHE SUPERSET** 8](#_Toc104420327)

[**1.2** **Các tính năng của APACHE SUPERSET ?** 8](#_Toc104420328)

[**1.3** **APACHE SUPERSET là nền tảng gì?** 9](#_Toc104420329)

[CHƯƠNG 2: CÀI ĐẶT VÀ SỬ DỤNG APACHE SUPERSET 10](#_Toc104420330)

[**2.1 Cài đặt, cấu hình và thực hiện một vài trực quan đơn giản:** 10](#_Toc104420331)

[**2.2 Sử dụng Apache Superset:** 17](#_Toc104420332)

[**2.2.1 Tạo cơ sở dữ liệu và tải tập dữ liệu lên Superset** 17](#_Toc104420333)

[**2.2.1.1 Cài đặt MySQL:** 17](#_Toc104420334)

[**2.2.1.2 Cài đặt MySQL Workbench:** 18](#_Toc104420335)

[**2.2.1.3 Tạo cơ sở dữ liệu MySQL trên AWS RDS:** 18](#_Toc104420336)

[**2.2.1.4 Kết nối đến cơ sở dữ liệu MySQL:** 26](#_Toc104420337)

[**2.2.1.5 Tạo cơ sở dữ liệu và tải tập dataset lên Superset:** 28](#_Toc104420338)

[**2.2.2Lọc dữ liệu bên ngoài sử dụng python và thực hiện một vài truy vấn đơn giản sử dụng SQL** 33](#_Toc104420339)

[**2.2.2.1 Xử lý dữ liệu bằng python** 33](#_Toc104420340)

[**2.2.2.2 Sử dụng SQL trên Superset:** 35](#_Toc104420341)

[**2.2.3Sử dụng Jinja Template để lọc dữ liệu trục tiếp trên Superset.** 38](#_Toc104420342)

[PHẦN KẾT LUẬN 40](#_Toc104420343)

[**1.** **Kết quả đạt được** 40](#_Toc104420344)

[*1.1. Kiến thức tìm hiểu được* 40](#_Toc104420345)

[*1.2. Chương trình đã làm được* 40](#_Toc104420346)

[- Cài đặt và sử dụng Superset. 40](#_Toc104420347)

[- Tải tệp dữ liệu lên Superset sử dụng MySQL. 40](#_Toc104420348)

[- Tạo biểu đồ trực quan hóa, cũng như tạo các filter cho việc trực quan được rõ ràng, cụ thể hơn. 40](#_Toc104420349)

[- Lọc dữ liệu từ bên ngoài sử dụng Python . 40](#_Toc104420350)

[- Sử dụng sql để thực hiện truy vấn và tạo biểu đồ. 40](#_Toc104420351)

[- Sử dụng jinja để lọc các dữ liệu. 40](#_Toc104420352)

[**2.** **Ưu điểm** 40](#_Toc104420353)

[**3.** **Nhược điểm** 40](#_Toc104420354)

[**4.** **Hướng phát triển** 40](#_Toc104420355)

[TÀI LIỆU THAM KHẢO 41](#_Toc104420356)

# **PHẦN MỞ ĐẦU**

## **1. Tính cấp thiết của đề tài**

Hiện nay, các tập dữ liệu khổng lồ (big data) đang ngày càng phổ biến, nhưng chúng thì lại có khối lượng lớn, phức tạp và nhiều loại dữ liệu đang tăng lên từng ngày. Nếu xử lý được các tập dữ liệu lớn này ta có thể giải quyết được nhiều vấn đề trong kinh doanh, giáo dục, xã hội. Mà để có thể xử lý các tập dữ liệu lớn đó và thực hiện các thao tác trực quan trên nó ta cần một ứng dụng có thể khám phá và trực quan nó trên quy mô petabyte. Do đó, Apache Superset là ứng viên sáng giá nhất, vì nó là một ứng dụng web thông minh phục vục cho kinh doanh hiện đại, sẵn sàng cho doanh nghiệp. Nó nhanh, nhẹ, trực quan và được tải với các tùy chọn giúp người dùng thuộc tất cả các bộ kỹ năng dễ dàng khám phá và hình dung dữ liệu của họ, từ biểu đồ hình tròn đơn giản đến biểu đồ không gian địa lý deck.gl rất chi tiết.

## **2. Đối tượng nghiên cứu**

Đối với đề tài này, đối tượng nghiên cứu là Big Data. Đồng thời kèm theo đó là các công nghệ áp dụng để phục vục việc trực quan hóa trên tạp dữ liệu được đơn giản hơn

* Apache Superset: đối tựng nghiên cứu chính, phục vục trực quan tập dữ liệu lớn.
* MySQL: một hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở (RDBMS), phục vục mục đích tương tác với tập dữ liệu cũng như là để giúp tải lên superset.
* Visual Studio 2019: là một môi trường phát triển tích hợp (IDE) từ Microsoft, phục vục mục đích lọc dữ liệu từ bên ngoài trước khi truyền vào trong superset.
* Jinja Template: một công cụ tạo khuôn mẫu nhanh chóng tiện lợi, phục vục mục đích lọc dữ liệu và hỗ trợ truy vấn trên tập dữ liệu. .

## **3. Phạm vi nghiên cứu**

Đề tài này chủ yếu tập trung vào trực quan hóa với các tập dữ liệu lớn cũng như thực hiện việc truy vấn trên chúng.

## **4. Kết quả dự kiến đạt được**

* Cài đặt được Apache Superset
* Tải được tập dữ liệu đã qua xử lý lên superset
* Trực quan, truy vấn trên tập dữ liệu.
* Tạo một ứng dụng tự động cập nhật tập dữ liệu theo thời gian.

# **PHẦN NỘI DUNG**

# **CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VỀ APACHE SUPERSET**

## **Tổng quan về APACHE SUPERSET**

Apache Superset là một ứng dụng web thông minh kinh doanh hiện đại, sẵn sàng cho doanh nghiệp. Nó nhanh, nhẹ, trực quan và được tải với các tùy chọn giúp người dùng thuộc tất cả các bộ kỹ năng dễ dàng khám phá và hình dung dữ liệu của họ, từ biểu đồ hình tròn đơn giản đến biểu đồ không gian địa lý deck.gl rất chi tiết.

Dưới đây là một **số cách khác nhau để ta có thể bắt đầu với Superset** :

* Tải xuống [nguồn từ trang web của Apache Foundation](https://dist.apache.org/repos/dist/release/superset/1.4.1/)
* Tải xuống phiên bản Superset mới nhất từ [​​Pypi tại đây](https://pypi.org/project/apache-superset/)
* Thiết lập Superset cục bộ bằng một lệnh sử dụng [Docker Compose](https://superset.apache.org/docs/installation/installing-superset-using-docker-compose)
* Tải xuống [hình ảnh Docker](https://hub.docker.com/r/apache/superset) từ Dockerhub
* Cài đặt phiên bản mới nhất của Superset [từ GitHub](https://github.com/apache/superset/tree/latest)



## **Các tính năng của APACHE SUPERSET ?**

Superset cung cấp:

* Giao diện trực quan để trực quan hóa tập dữ liệu và tạo bảng điều khiển tương tác
* Một loạt các hình ảnh trực quan đẹp mắt để giới thiệu dữ liệu của ta
* Trình tạo trực quan hóa không cần mã để trích xuất và trình bày bộ dữ liệu
* Một SQL IDE đẳng cấp thế giới để chuẩn bị dữ liệu để hiển thị, bao gồm một trình duyệt siêu dữ liệu phong phú
* Một lớp ngữ nghĩa nhẹ cho phép các nhà phân tích dữ liệu nhanh chóng xác định các thứ nguyên và chỉ số tùy chỉnh
* Hỗ trợ out-of-the-box cho hầu hết các cơ sở dữ liệu nói tiếng SQL
* Bộ nhớ đệm và truy vấn không đồng bộ liền mạch, trong bộ nhớ
* Một mô hình bảo mật có thể mở rộng cho phép cấu hình các quy tắc rất phức tạp về việc ai có thể truy cập các tính năng và bộ dữ liệu của sản phẩm.
* Tích hợp với các phụ trợ xác thực chính (cơ sở dữ liệu, OpenID, LDAP, OAuth, REMOTE\_USER, v.v.)
* Khả năng thêm các plugin trực quan hóa tùy chỉnh
* Một API để tùy chỉnh có lập trình
* Một kiến ​​trúc gốc đám mây được thiết kế từ đầu cho quy mô

## **APACHE SUPERSET là nền tảng gì?**

Superset là nền tảng đám mây và được thiết kế để có tính khả dụng cao. Nó được thiết kế để mở rộng ra các môi trường lớn, phân tán và hoạt động rất tốt bên trong các thùng chứa. Mặc dù ta có thể dễ dàng kiểm tra ổ đĩa Superset trên một thiết lập khiêm tốn hoặc đơn giản trên máy tính xách tay của mình, nhưng hầu như không có giới hạn về việc mở rộng nền tảng.

Superset cũng có nguồn gốc đám mây theo nghĩa là nó linh hoạt và cho phép ta chọn:

* máy chủ web (Gunicorn, Nginx, Apache),
* công cụ cơ sở dữ liệu siêu dữ liệu (MySQL, Postgres, MariaDB, v.v.),
* hàng đợi tin nhắn (Redis, RabbitMQ, SQS, v.v.),
* phụ trợ kết quả (S3, Redis, Memcached, v.v.),
* lớp bộ nhớ đệm (Memcached, Redis, v.v.),

Superset cũng hoạt động tốt với các dịch vụ như NewRelic, StatsD và DataDog, đồng thời có khả năng chạy khối lượng công việc phân tích dựa trên hầu hết các công nghệ cơ sở dữ liệu phổ biến.Superset hiện đang được chạy trên quy mô lớn tại nhiều công ty.

Ví dụ: Superset được chạy trong môi trường sản xuất của Airbnb bên trong Kubernetes và phục vụ hơn 600 người dùng hoạt động hàng ngày xem hơn 100 nghìn biểu đồ mỗi ngày.

# **CHƯƠNG 2: CÀI ĐẶT VÀ SỬ DỤNG APACHE SUPERSET**

## **2.1 Cài đặt, cấu hình và thực hiện một vài trực quan đơn giản:**

Để dễ thử các tính năng, ta có thể dùng Docker hoặc Docker Compose.

*docker run -d -p 8080:8088 --name superset apache/superset*

Đợi một tí để Docker kéo về là có thể bắt đầu sử dụng. Sau khi Docker ở trạng thái chạy, ta sẽ cấu hình Superset.

Đầu tiên là admin account

*docker exec -it superset superset fab create-admin --username admin --firstname Superset --lastname Admin --email admin@superset.com --password admin*

Nâng cấp CSDL

docker exec -it superset superset db upgrade

Tải vài ví dụ

docker exec -it superset superset load\_examples

Chạy một lần cuối để khởi động

*docker exec -it superset superset init*

Bản Docker gốc đã hỗ trợ nhiều kết nối đến các loại CSDL thông dụng, tuy nhiên, bạn hoàn toàn có thể cài đặt thêm một cách dễ dàng bằng cách chạy lệnh trong Docker. Ví dụ: muốn cài driver để kết nối đến SQL server.

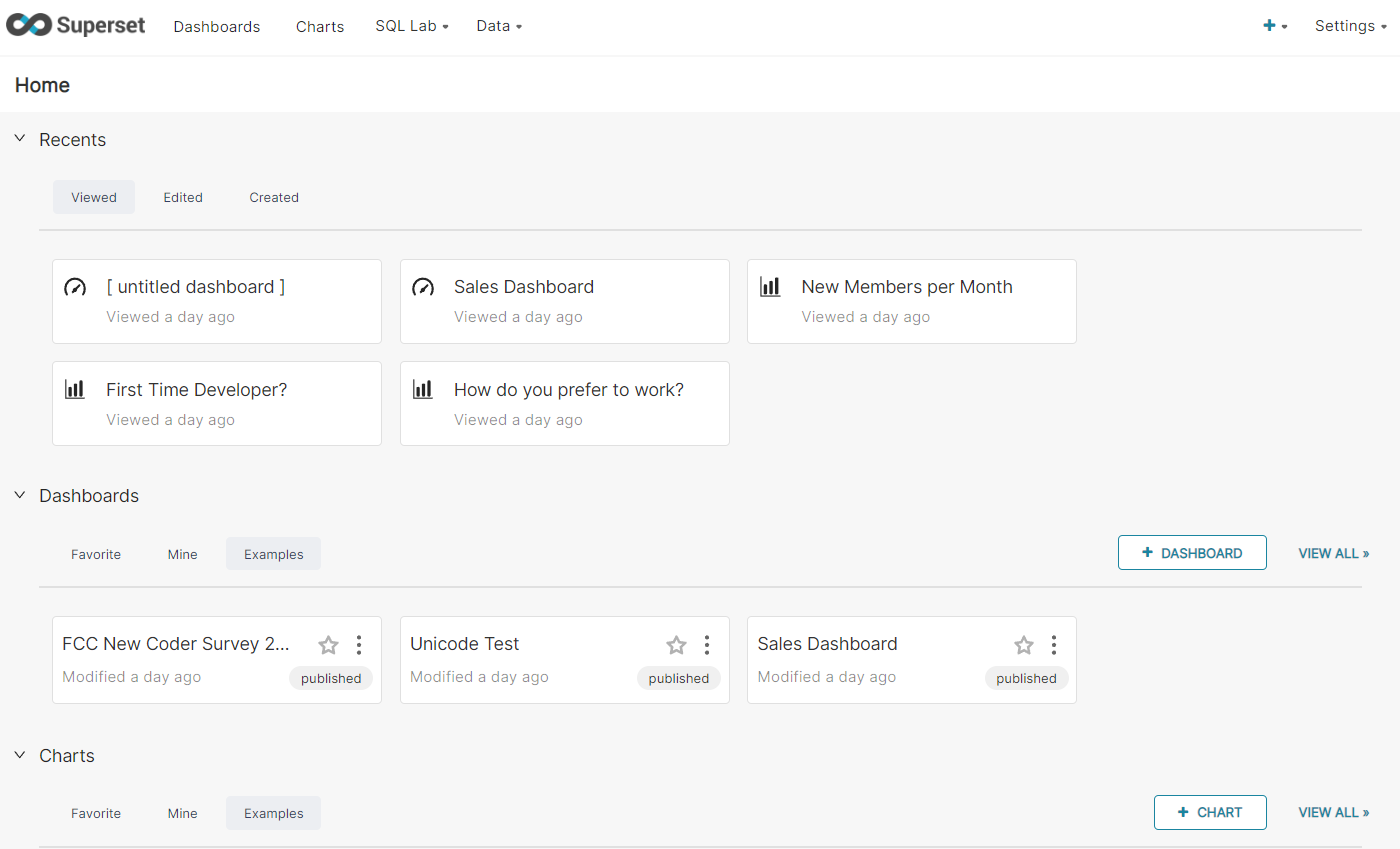
*docker exec -it superset pip install pymssql*

Bây giờ ta có thể bắt đầu!

*http://localhost:8080/login/*

Đăng nhập bằng tài khoản đã khai báo ở trên: [admin/admin]

Màn hình welcome của Superset liệt kê các bảng biểu, đồ thị, báo cáo đã được thao tác trước đó trong hệ thống để dễ dàng truy cập.

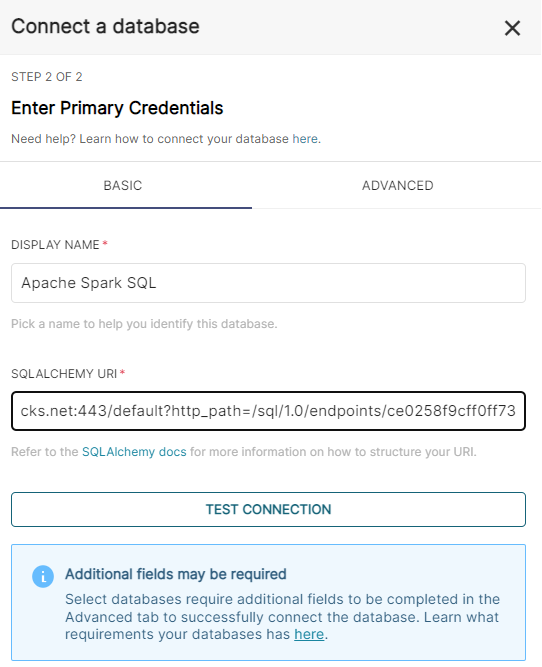


**Kết nối dữ liệu**

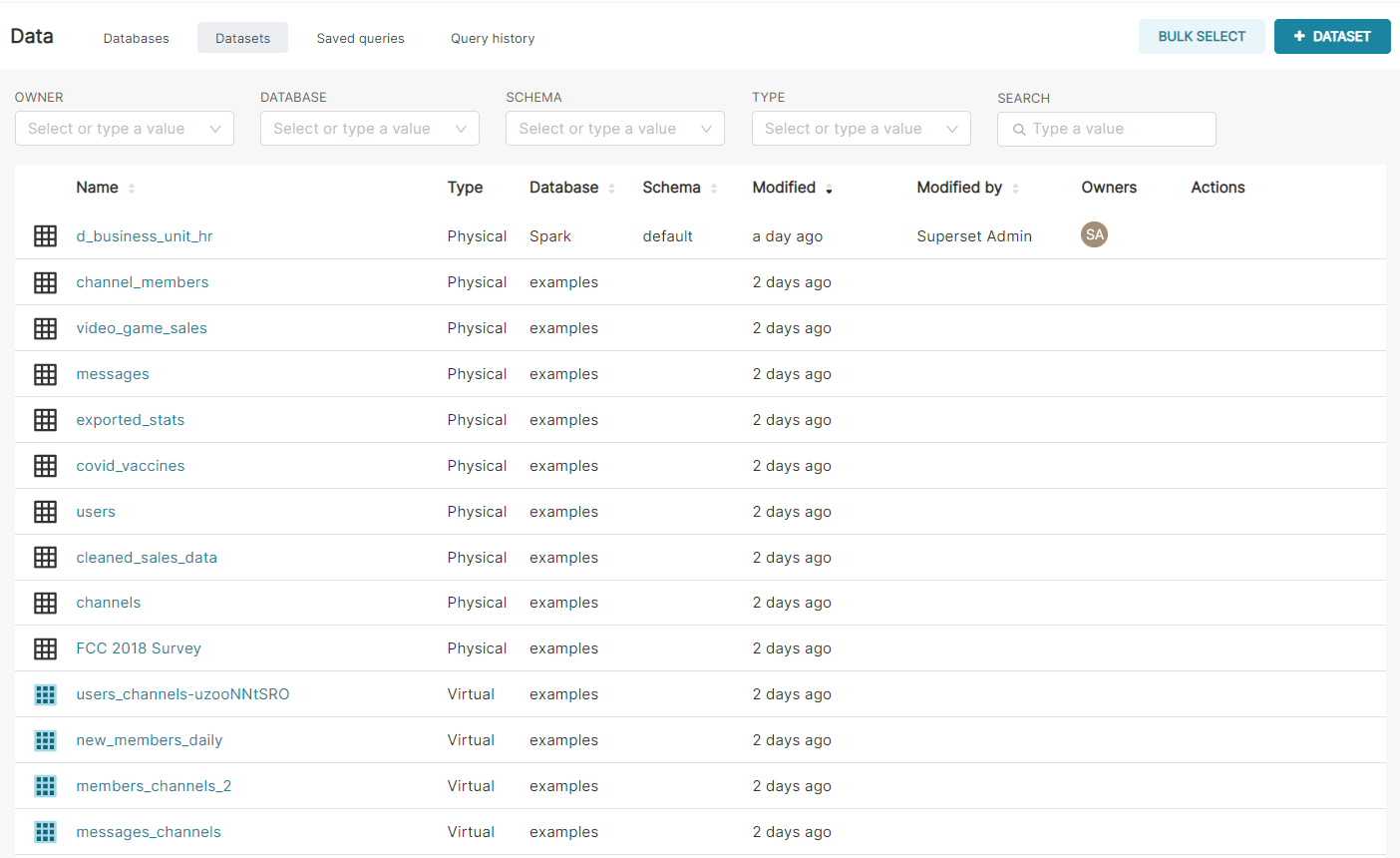
Apache Superset không lưu trữ dữ liệu nên cần phải kết nối đến một CSDL. Superset cung cấp một công cụ trực quan để thao tác việc kết nối. Nhìn chung thao tác kết nối khá đơn giản tuỳ vào loại dữ liệu nguồn mà ta cung cấp các thông tin kết nối tương ứng như username, password.



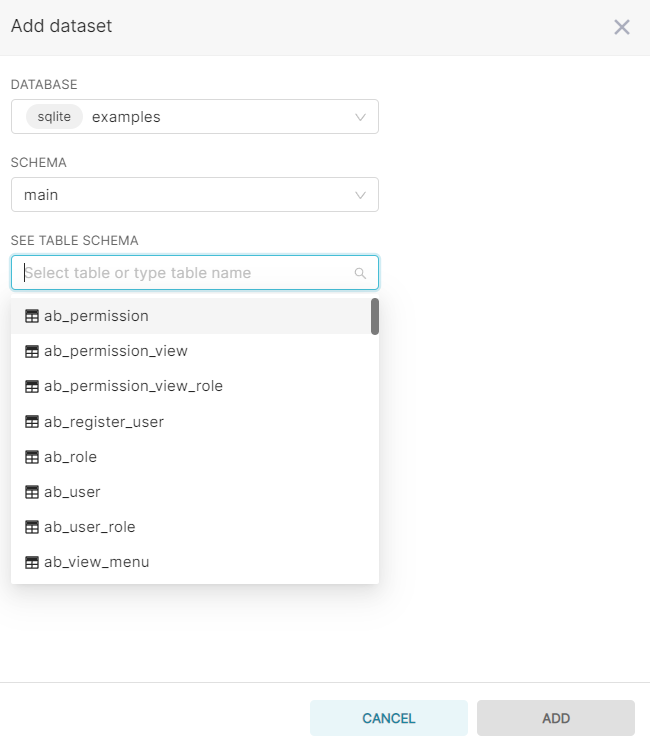
Ta cũng có thể kết nối đến CSDL bằng cách cung cấp URI theo cấu trúc SQLAlchemy.



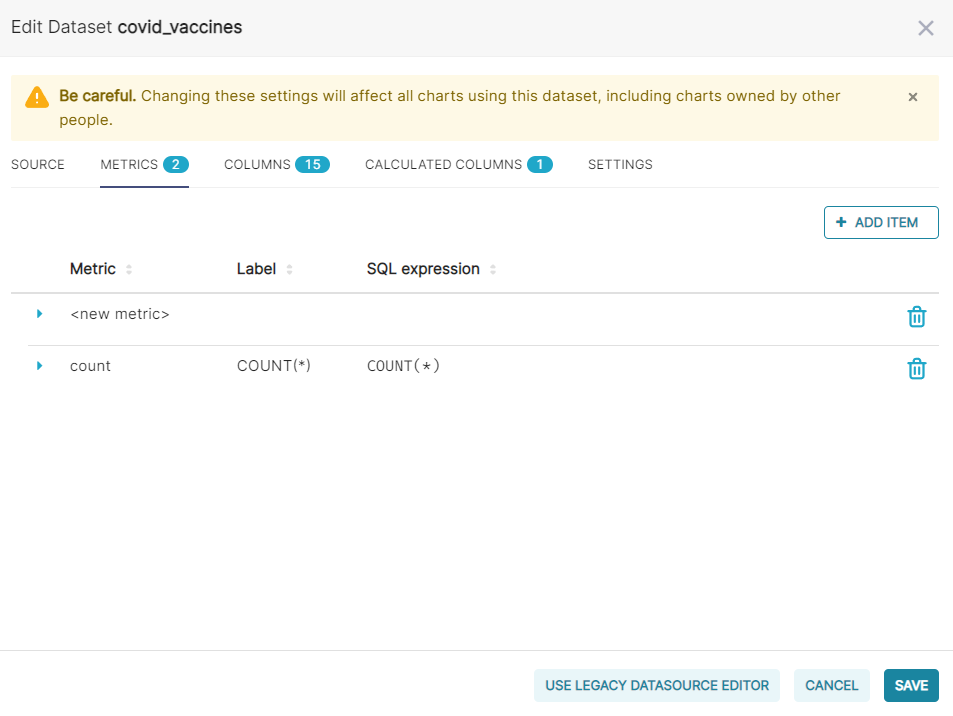
Sau khi kết nối, ta có thể bắt đầu tiến hành tạo các dataset để vẽ biểu đồ. Superset cung cấp giao diện để nhìn tổng quan các dataset có thể thao tác.



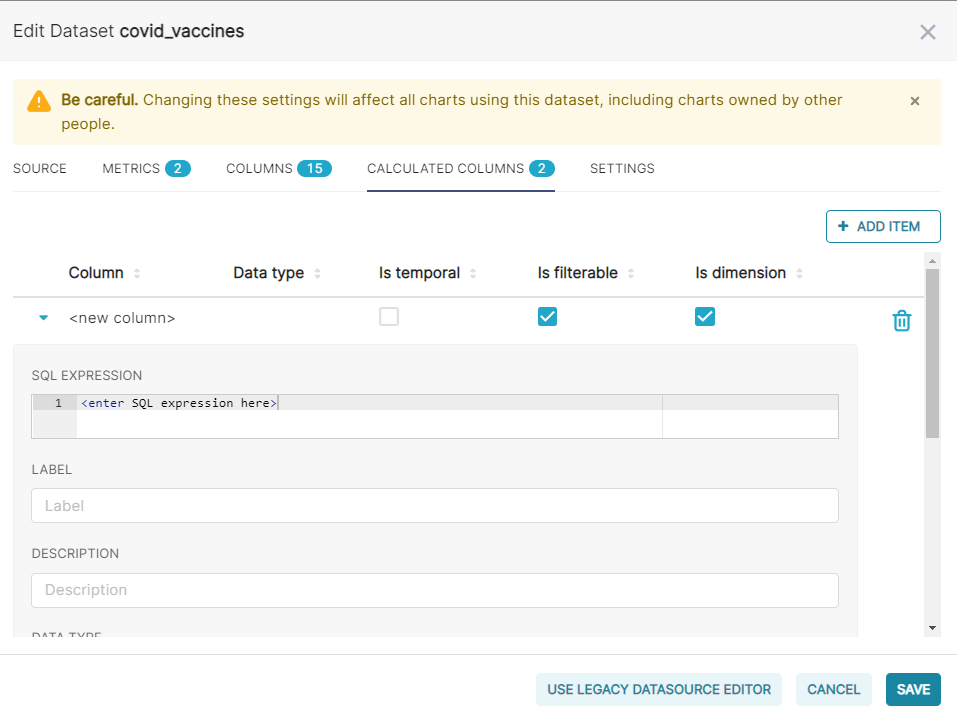
Ta có thể tiếp tục làm việc với các dataset đang có hoặc tạo một dataset mới bằng cách kết nối đến CSDL, chọn schema và bảng.



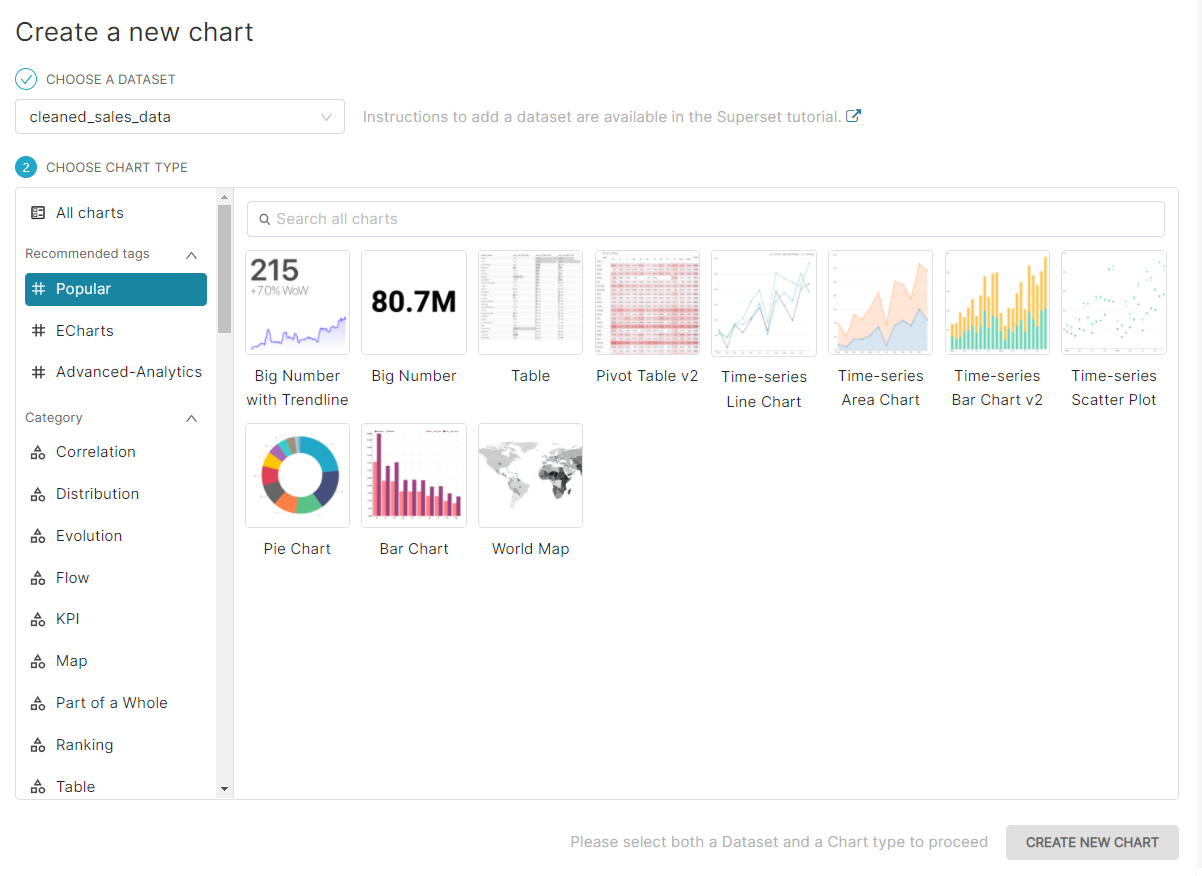
Sau khi tạo dataset, ta có thể điều chỉnh cấu hình của dataset như thêm các metric, điều chỉnh cấu hình của dữ liệu như: cột được phép lọc, cột dữ liệu, hoặc thêm các cột tính toán.



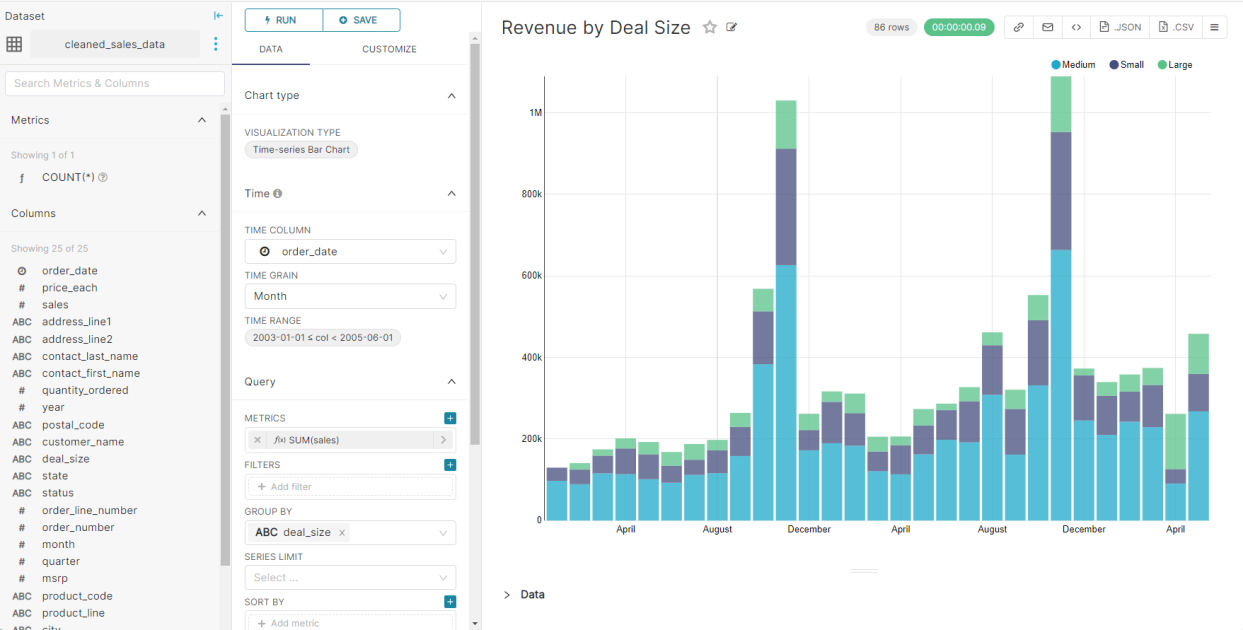




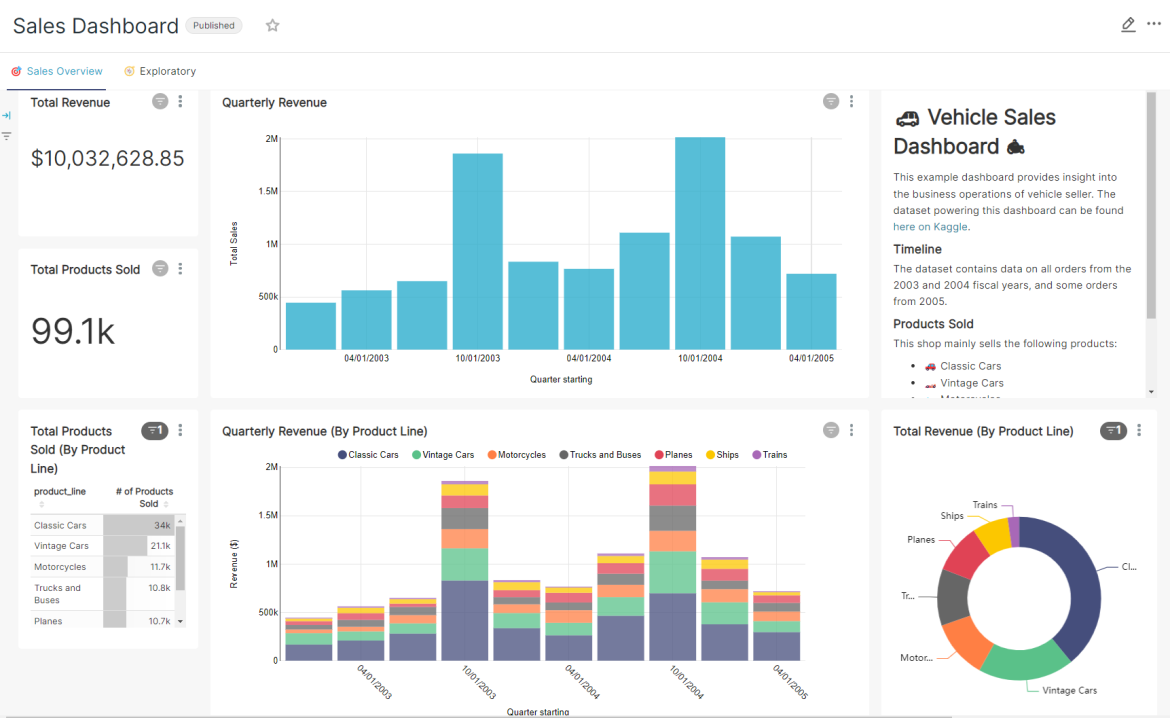
Sau khi có tập dữ liệu, ta có thể tiến hành tạo biểu đồ. Superset cung cấp rất nhiều loại biểu đồ, công việc cần làm là chọn dataset và loại biểu đồ muốn hiển thị.



Sau đó, ta có thể điều chỉnh lại tuỳ theo nhu cầu cần hiển thị, Superset cung cấp rất nhiều tuỳ chỉnh tuỳ theo biểu đồ và dữ liệu.



Bước cuối cùng là đưa tất cả các biểu đồ lên dashboard.



## **2.2 Sử dụng Apache Superset:**

## **2.2.1 Tạo cơ sở dữ liệu và tải tập dữ liệu lên Superset**

## **2.2.1.1 Cài đặt MySQL:**

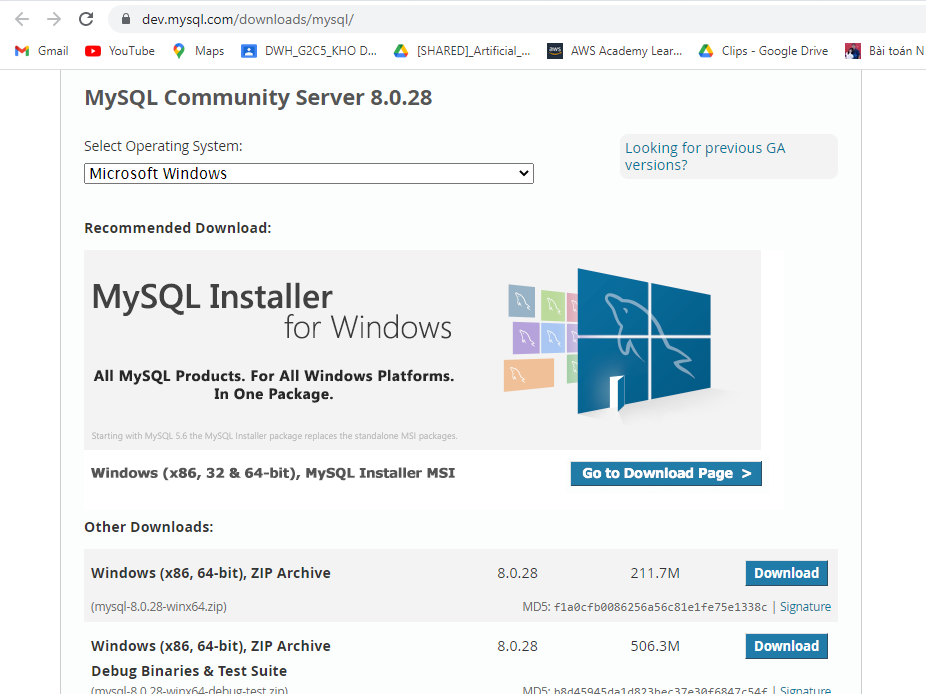
-Chúng ta truy cập vào : https://dev.mysql.com/downloads/mysql/

-Sau đó ta tải bản phù hợp với máy tính. Ở đây ta sẽ chọn phiên bản MySQL Community Server 8.0.28 (Windows x86, 64 bit). Nên chọn phiên bản có kích cỡ file lớn hơn cho đầy đủ.

- Sau khi giải nén file thì ta tiến hành cài đặt.

-Chúng ta sẽ cài đặt dựa vào link sau:

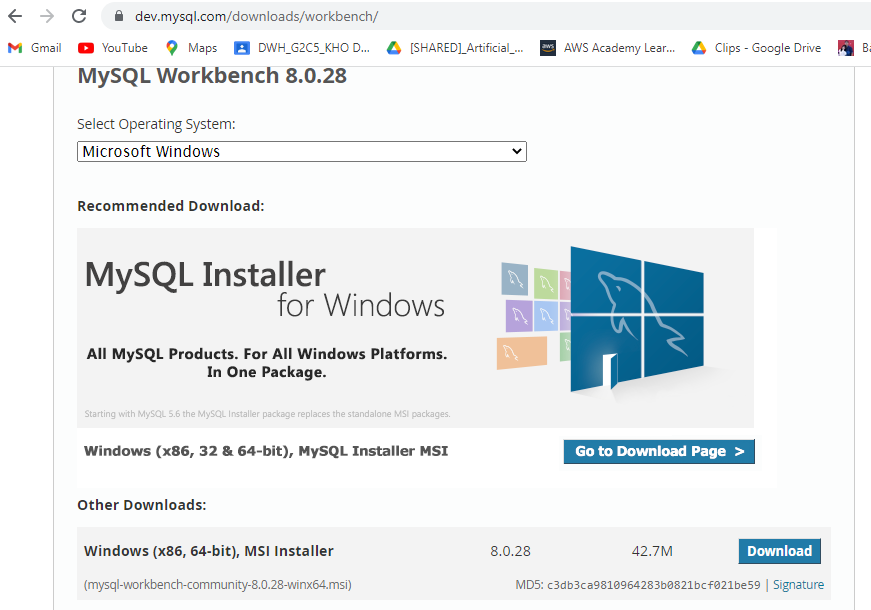
https://blogchiasekienthuc.com/thu-thuat-hay/cai-dat-mysql-workbench-tren-windows-10.html



## **2.2.1.2 Cài đặt MySQL Workbench:**

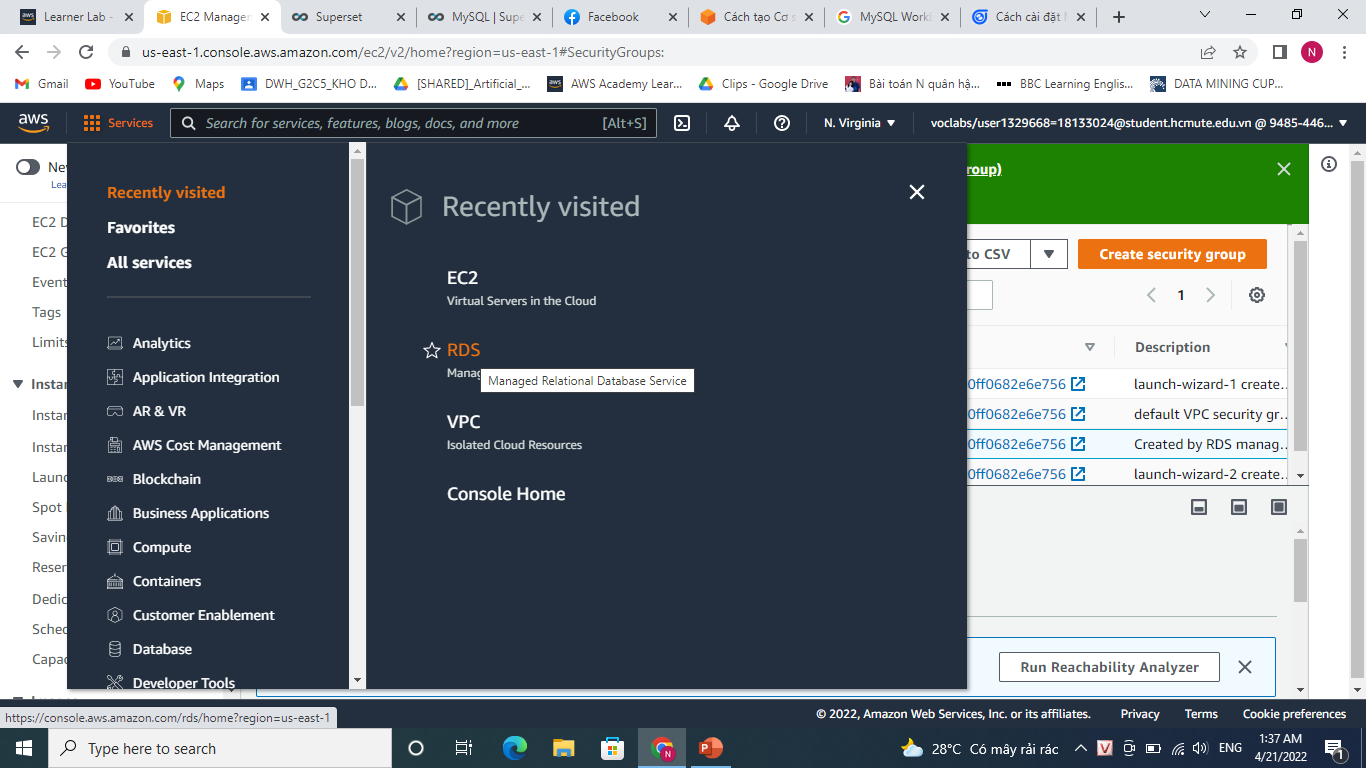
-Chúng ta truy cập vào : <https://dev.mysql.com/downloads/workbench/>

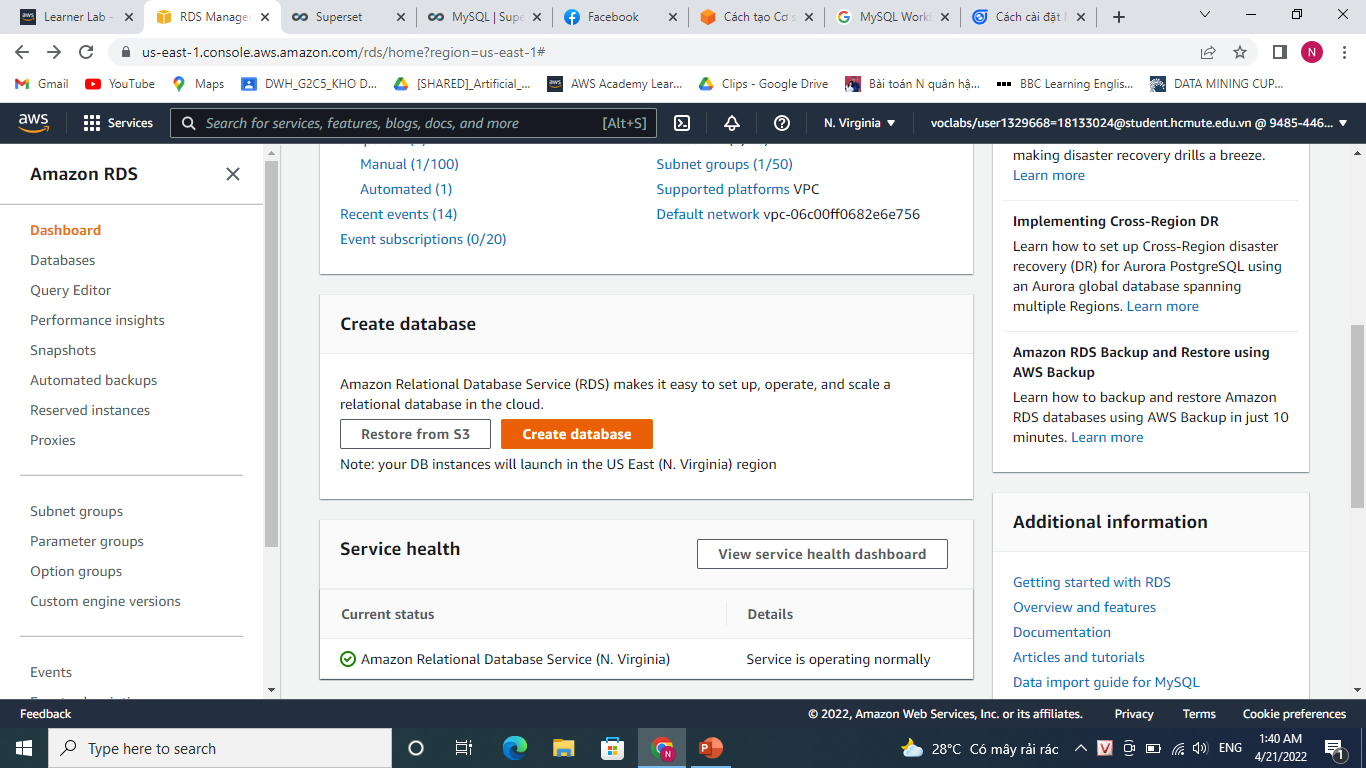
Sau đó chúng ta tải về tương tự như MySQL, giải nén file và tiến hành cài đặt.



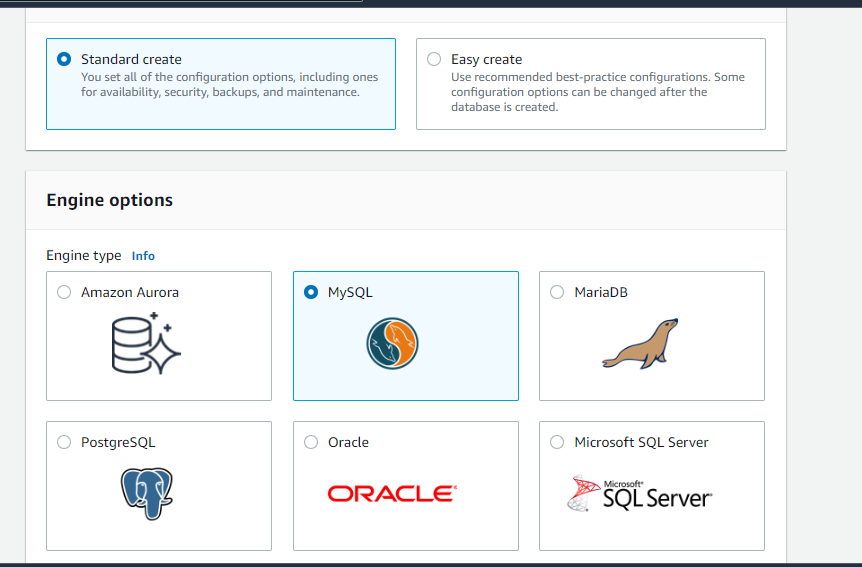
## **2.2.1.3 Tạo cơ sở dữ liệu MySQL trên AWS RDS:**

- Chúng ta vào service và chọn RDS.

 - Chọn Create database.

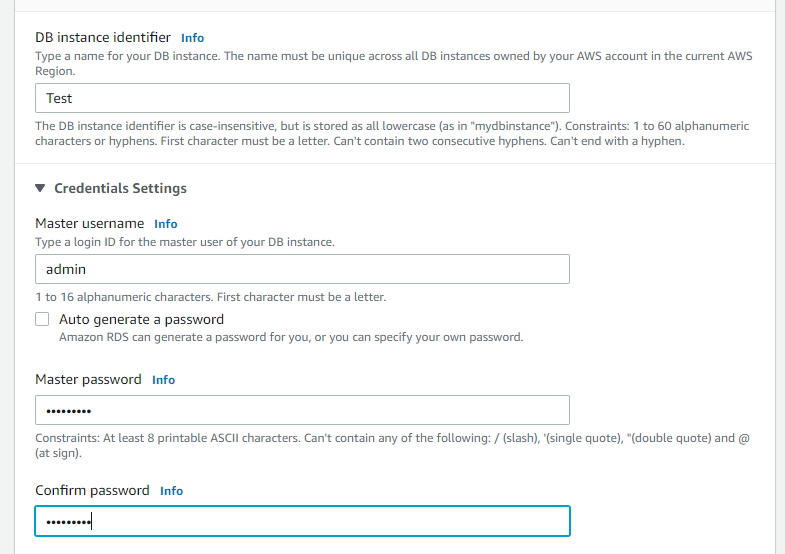


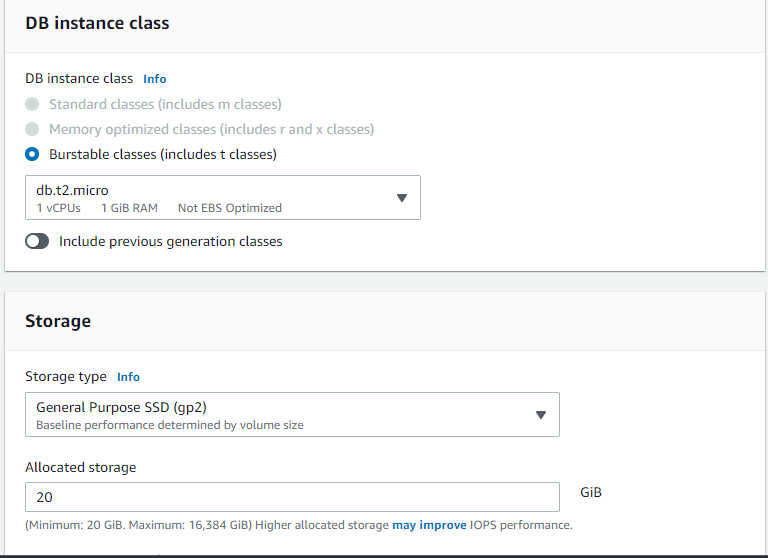
- Chọn MySQL



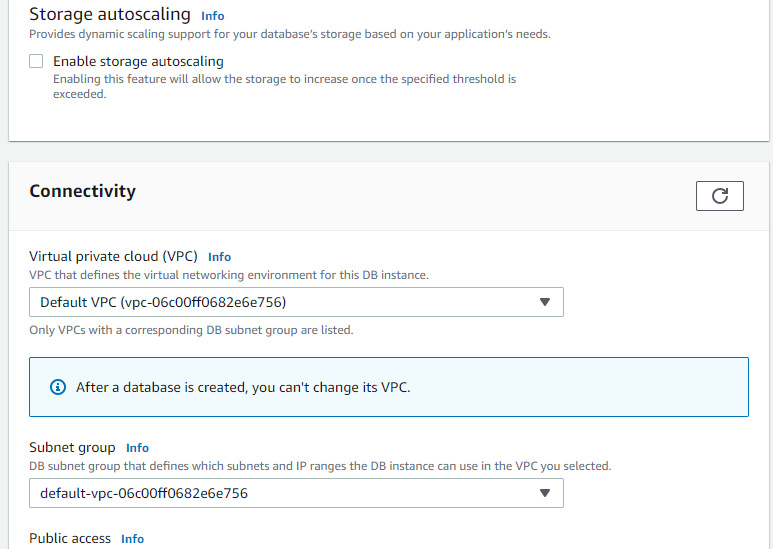
- Chọn Templates là Free tier.

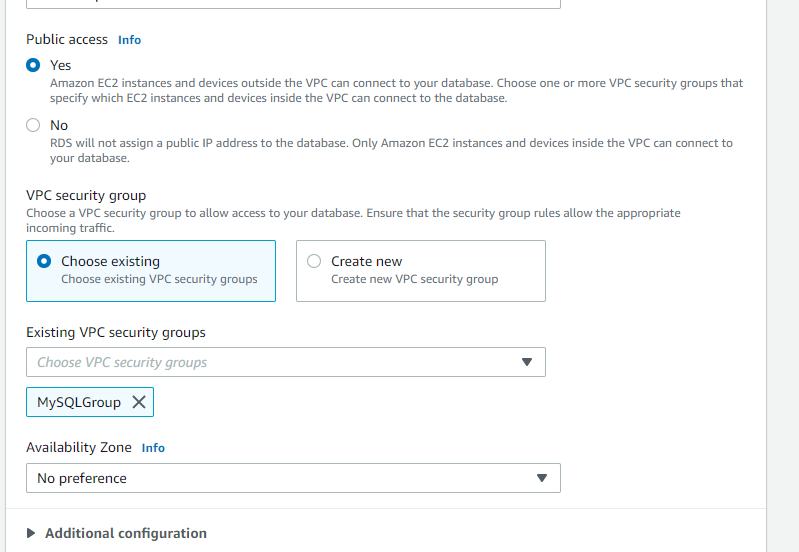
 - Set tên Database, master username và password.

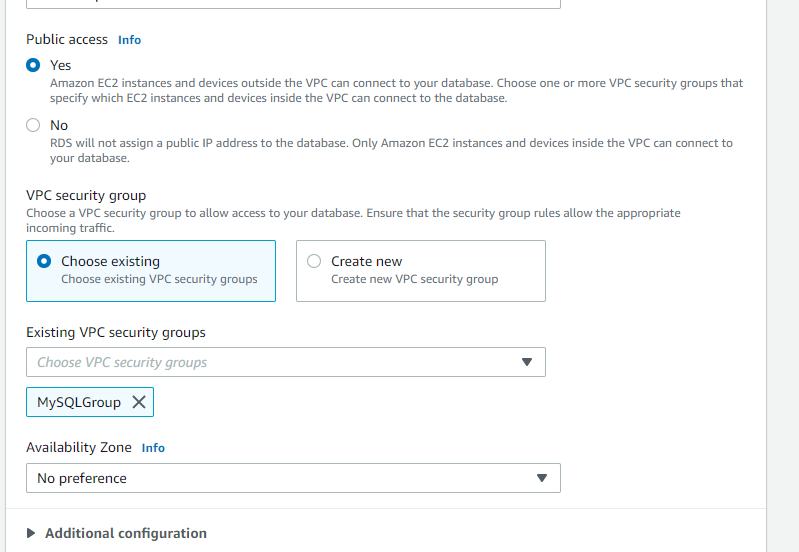
 - Chọn DB instance class và Storage.



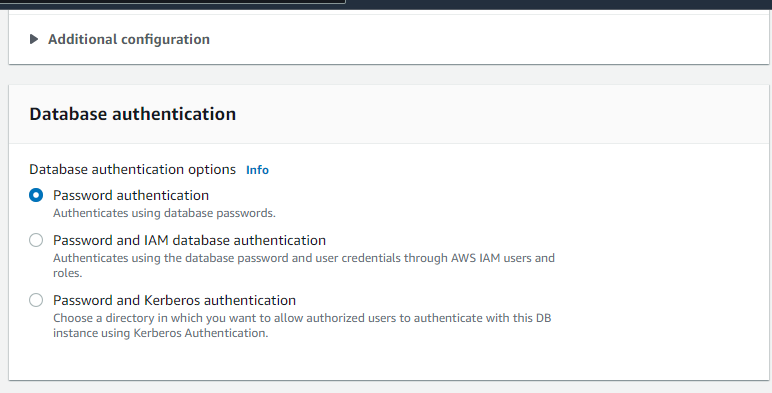
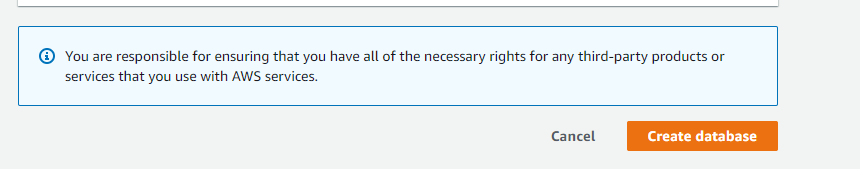
- Bỏ tích Storage autoscaling.

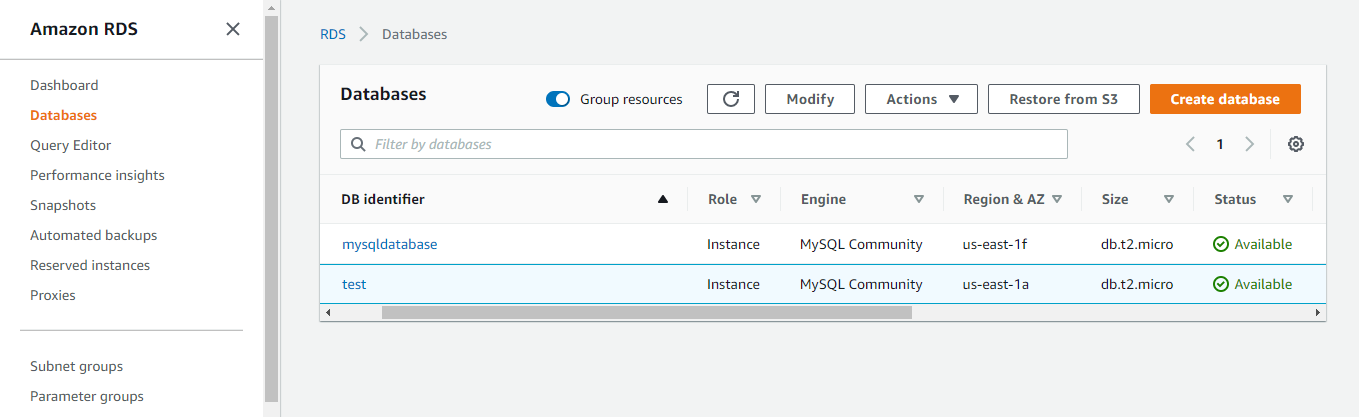
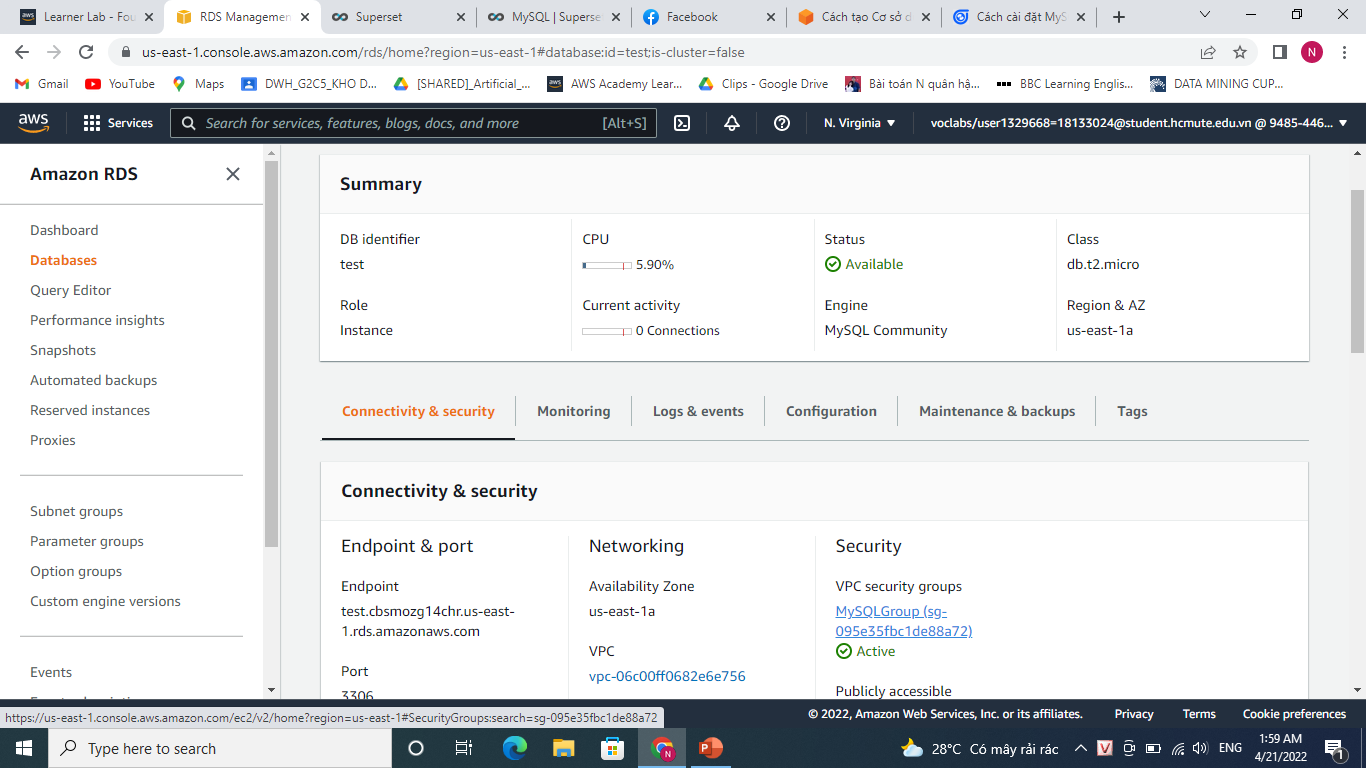




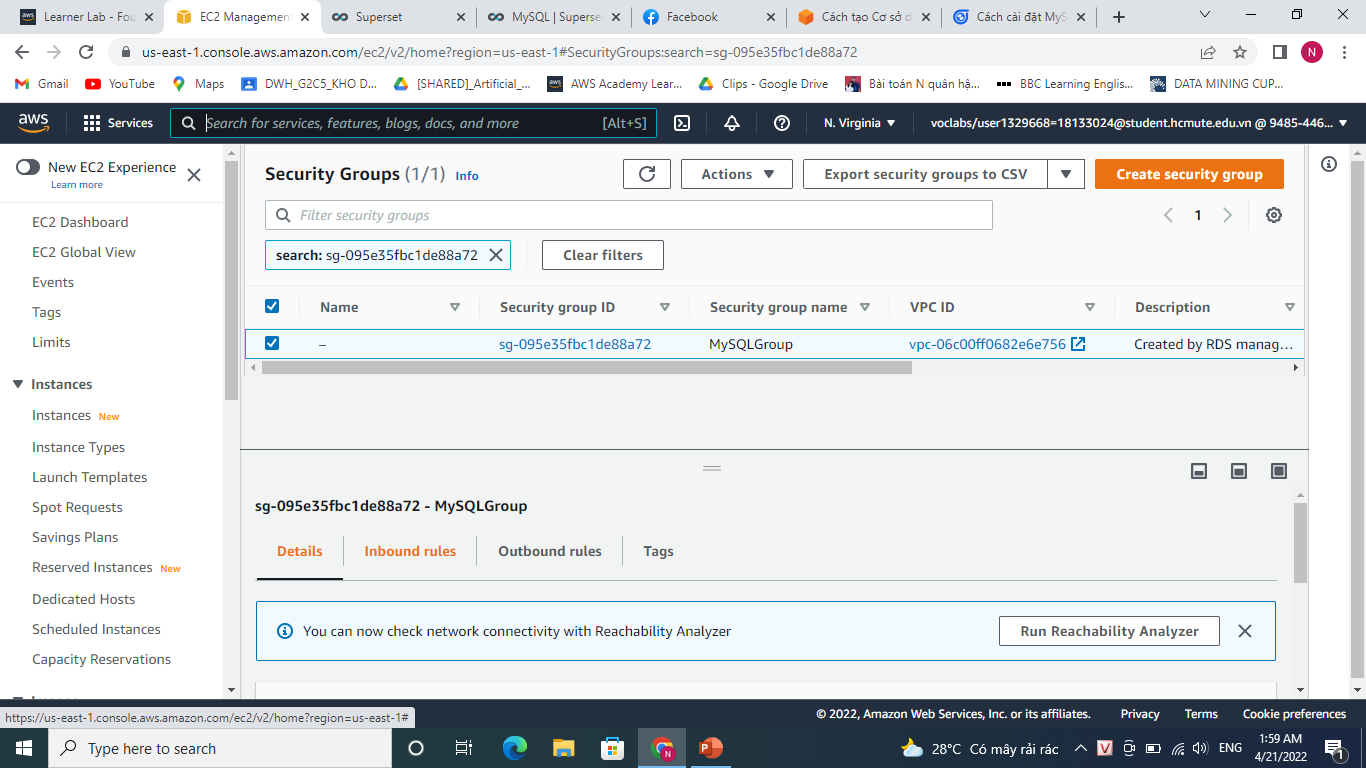


- Chọn Create database để tạo cơ sở dữ liệu.

  - Sau khi tạo ta có database test.

 - Giờ ta vào database test vừa tạo, sau đó vào vpc security groups của test. 

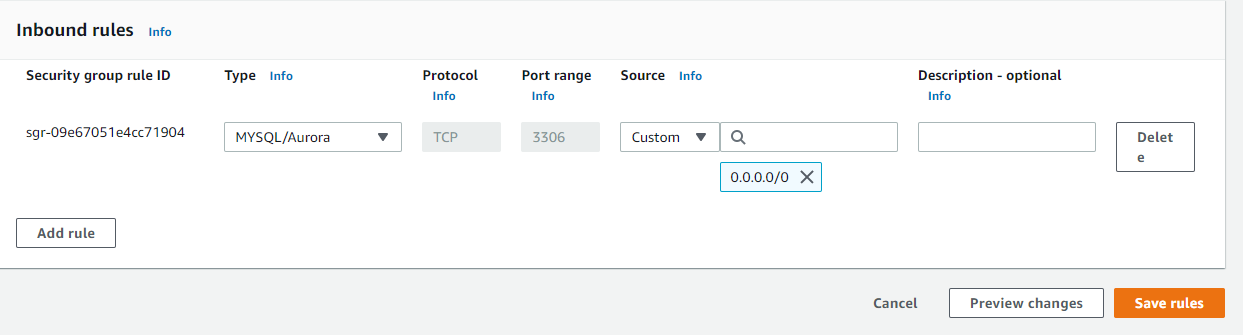
- Chọn phần inbound rules.



- Chọn edit inbound rules.

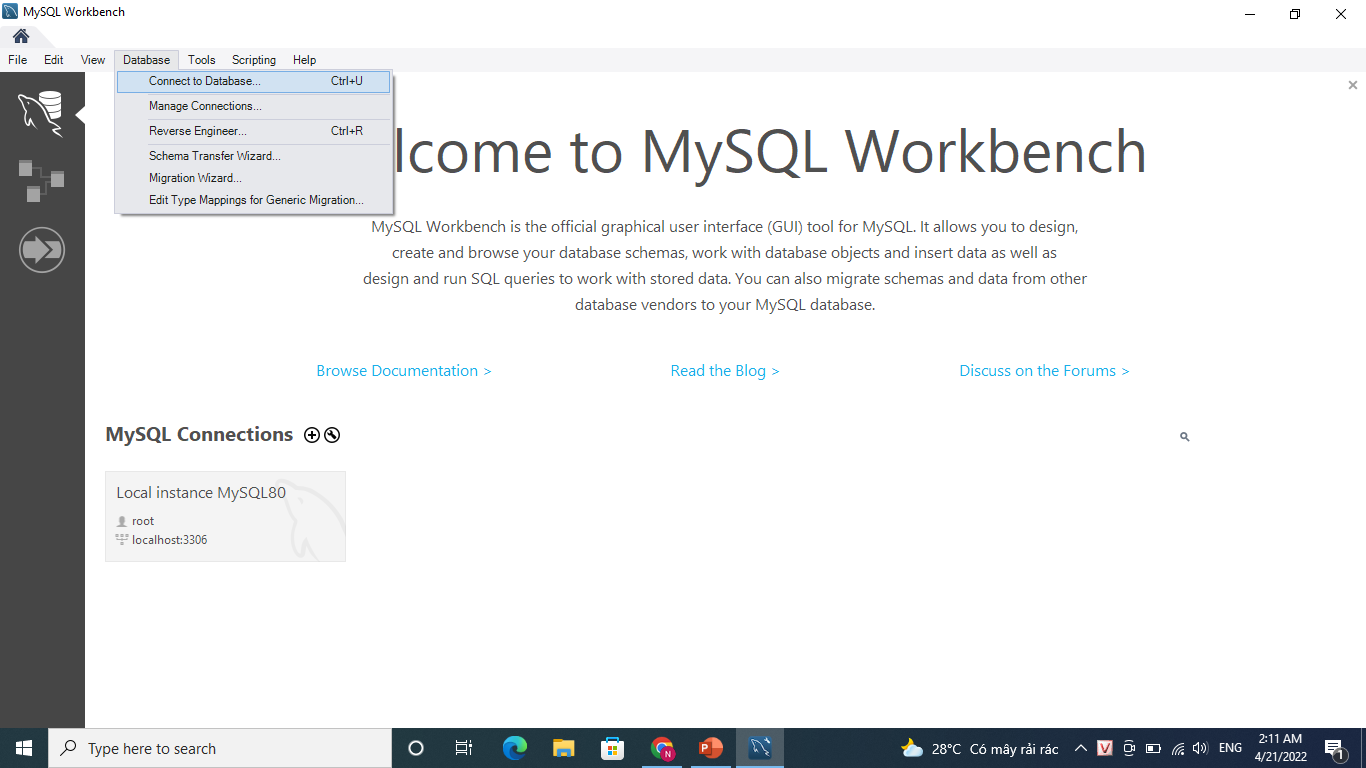


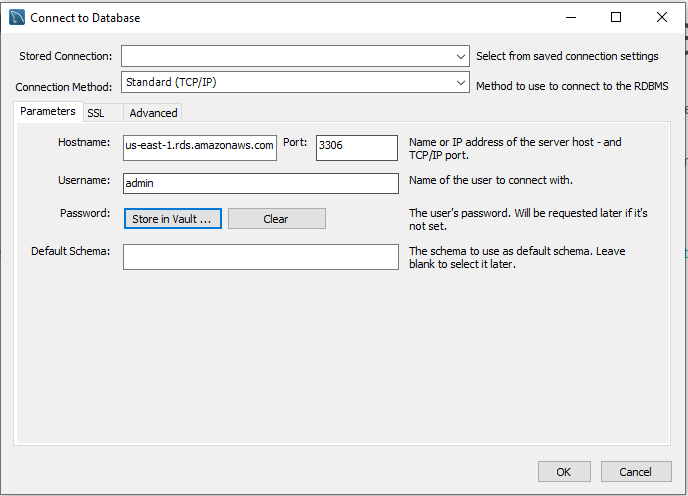
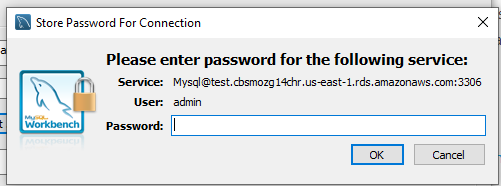
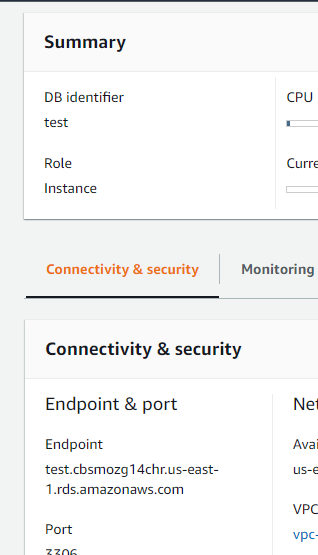
- Thêm rule và thiết lập cổng port là 3306. Sau đó save rules



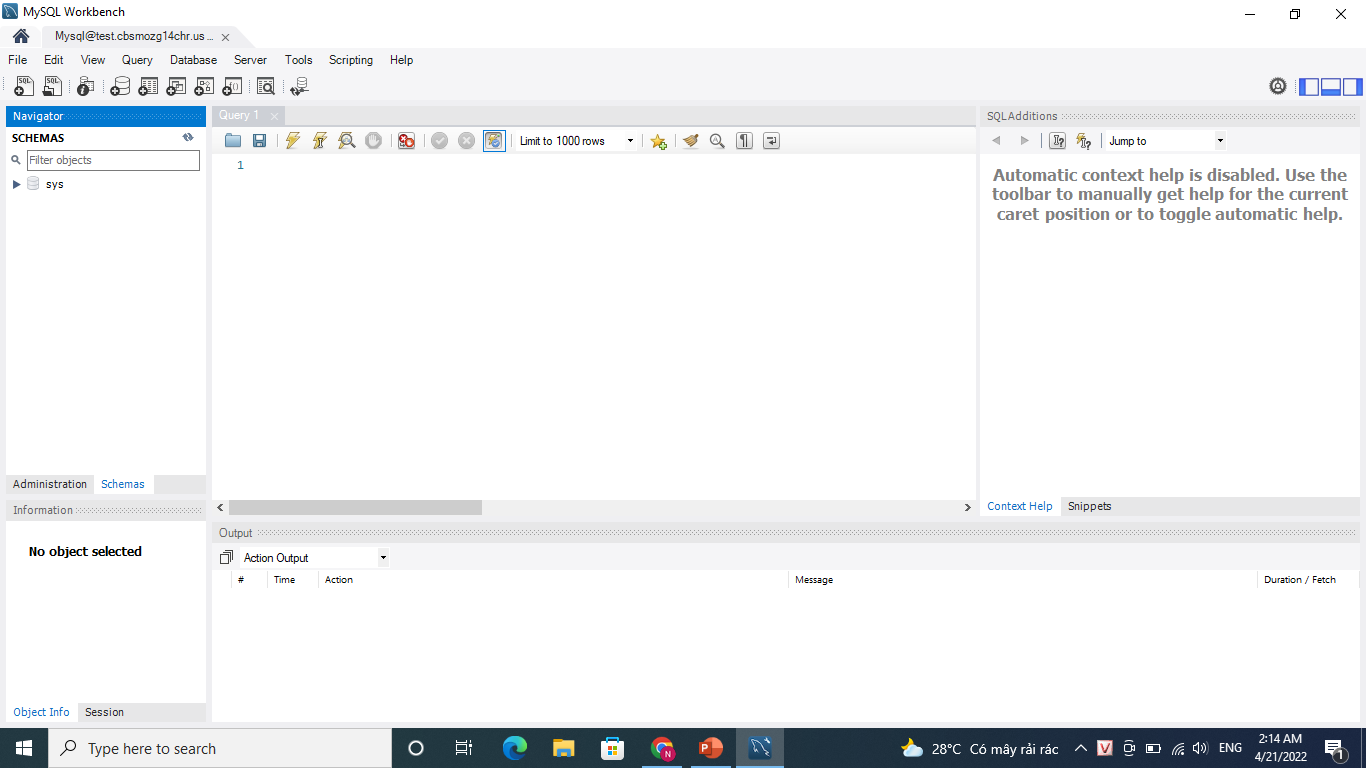
## **2.2.1.4 Kết nối đến cơ sở dữ liệu MySQL:**

- Ta mở MySQL Workbench. Sau đó vào mục databse chọn Connect to Database.

 - Phần host ta điền địa chỉ endpoint của database test. Port ta điền 3306. Và điền username và password đã thiết lập . Sau đó nhấn oke.

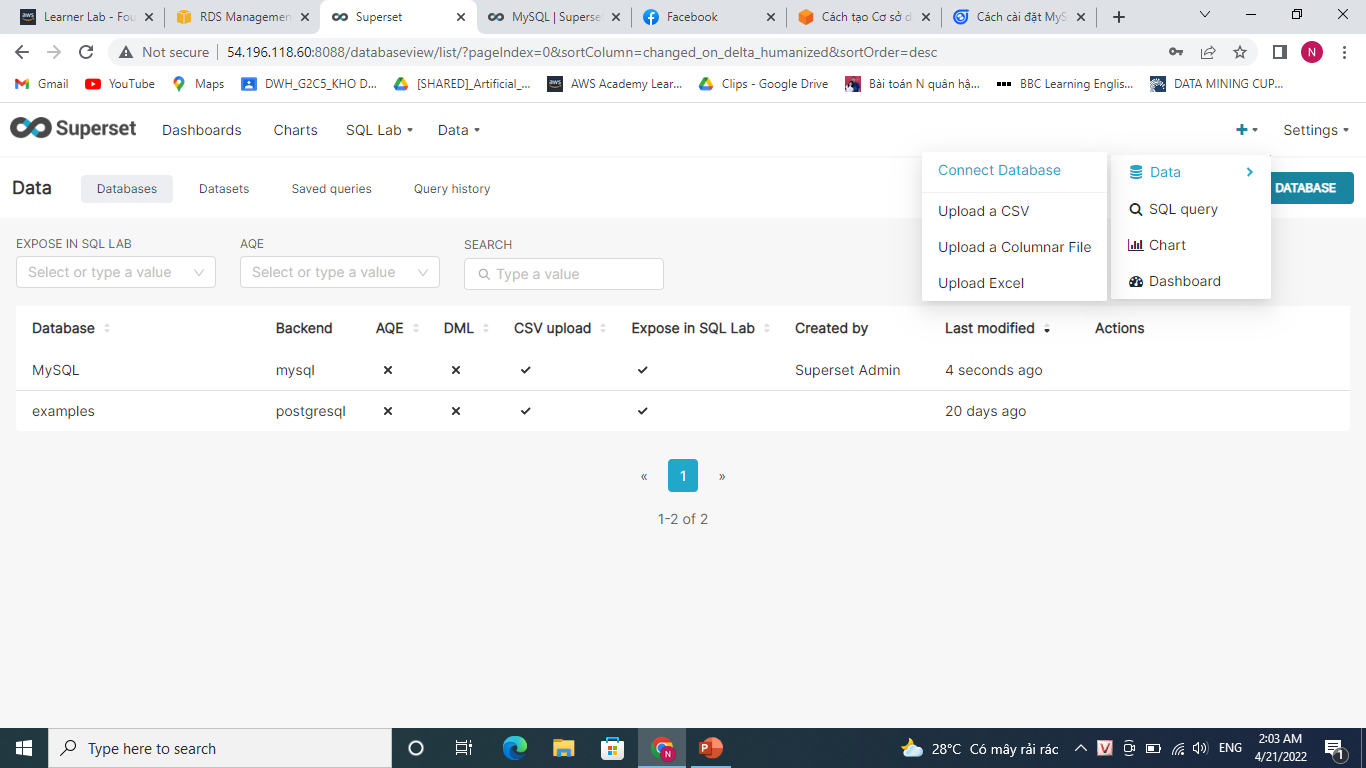
  

- MySQL Workbench hiện ra như thế này là đã kết nối thành công.

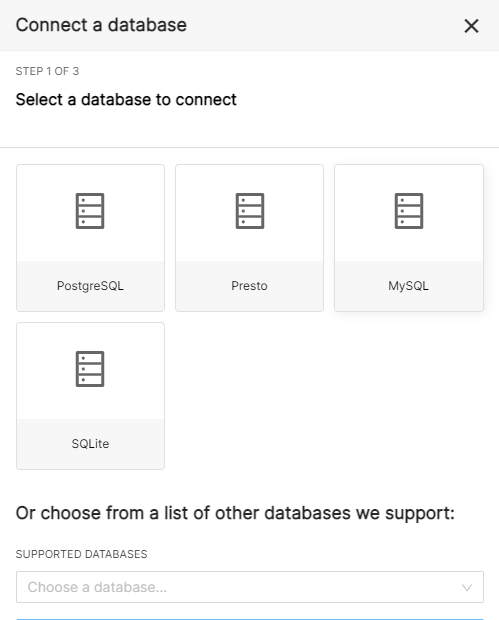


## **2.2.1.5 Tạo cơ sở dữ liệu và tải tập dataset lên Superset:**

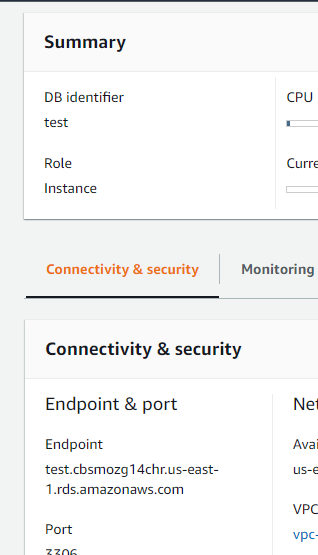
Ta nhấn vào dấu cộng ở trên phải của màn hinh rồi chọn Data->Connect Database



Tiếp theo ta chọn kiểu data base để kết nối, với tụi em là sử dụng MySQL

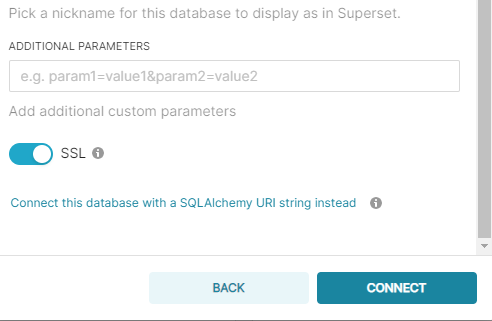
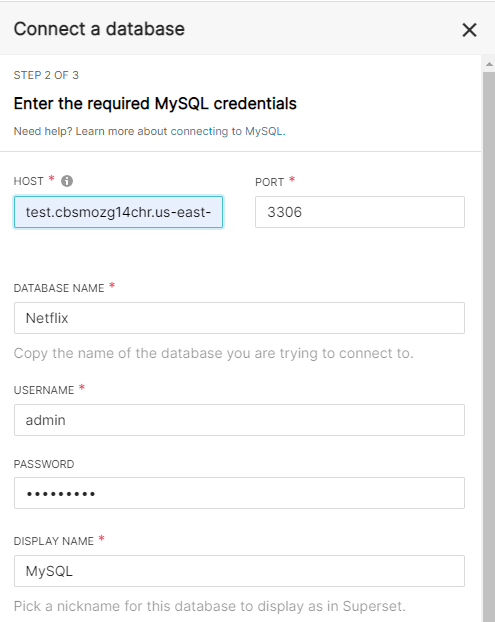


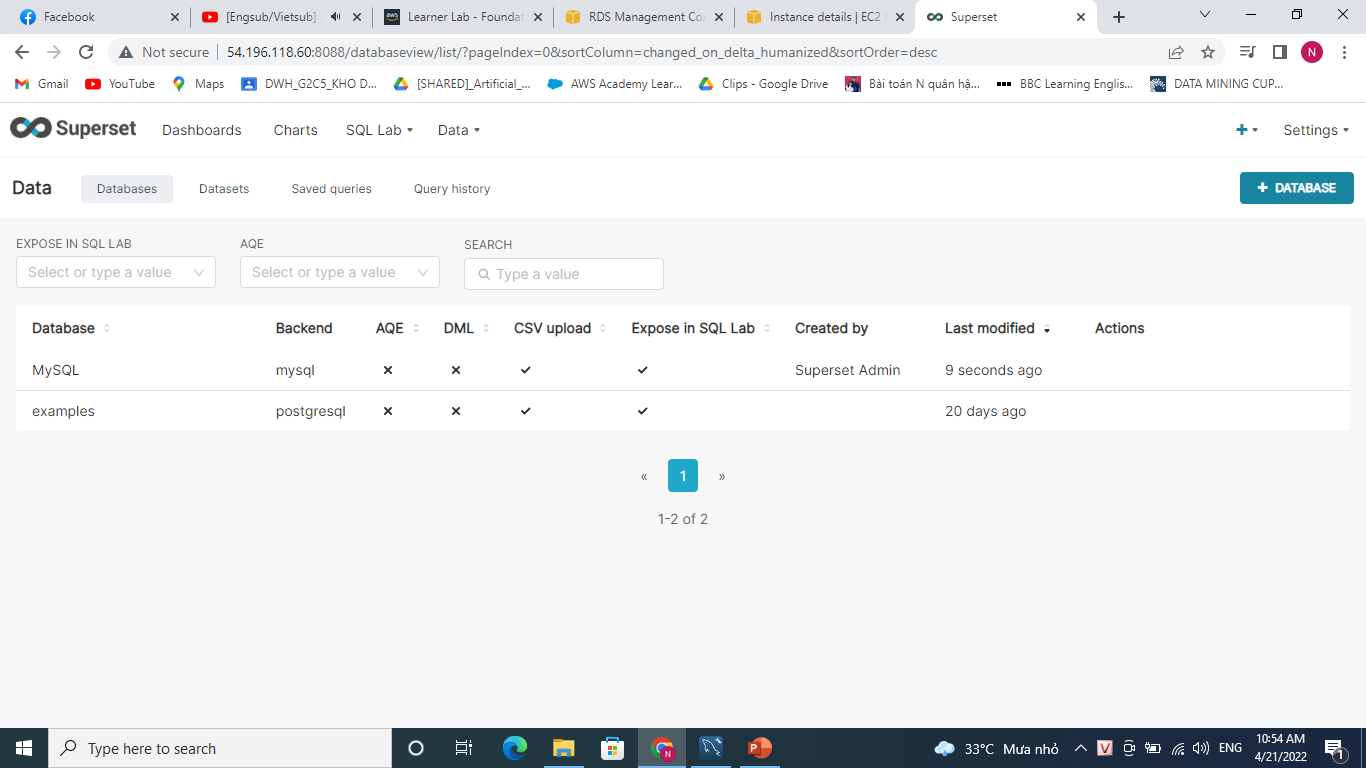
- Ta chọn host là endpoint của database đã tạo trên RDS



DATABASE NAME là tên của db mà ta sẽ tạo

User name và password là tài khoản và mật khẩu để đăng nhập vào đường link superset của nhóm em/



Nếu hiện ra kết quả như hình là ta đã làm đúng

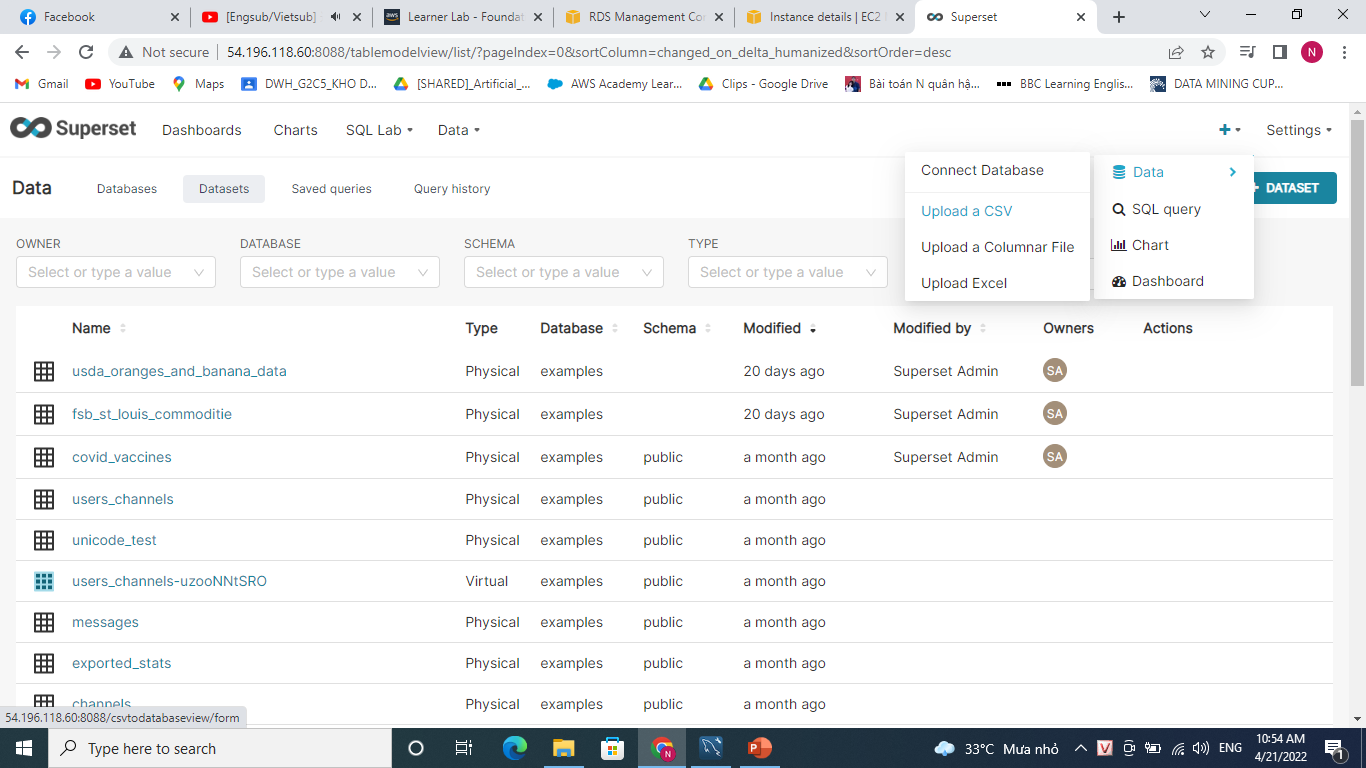
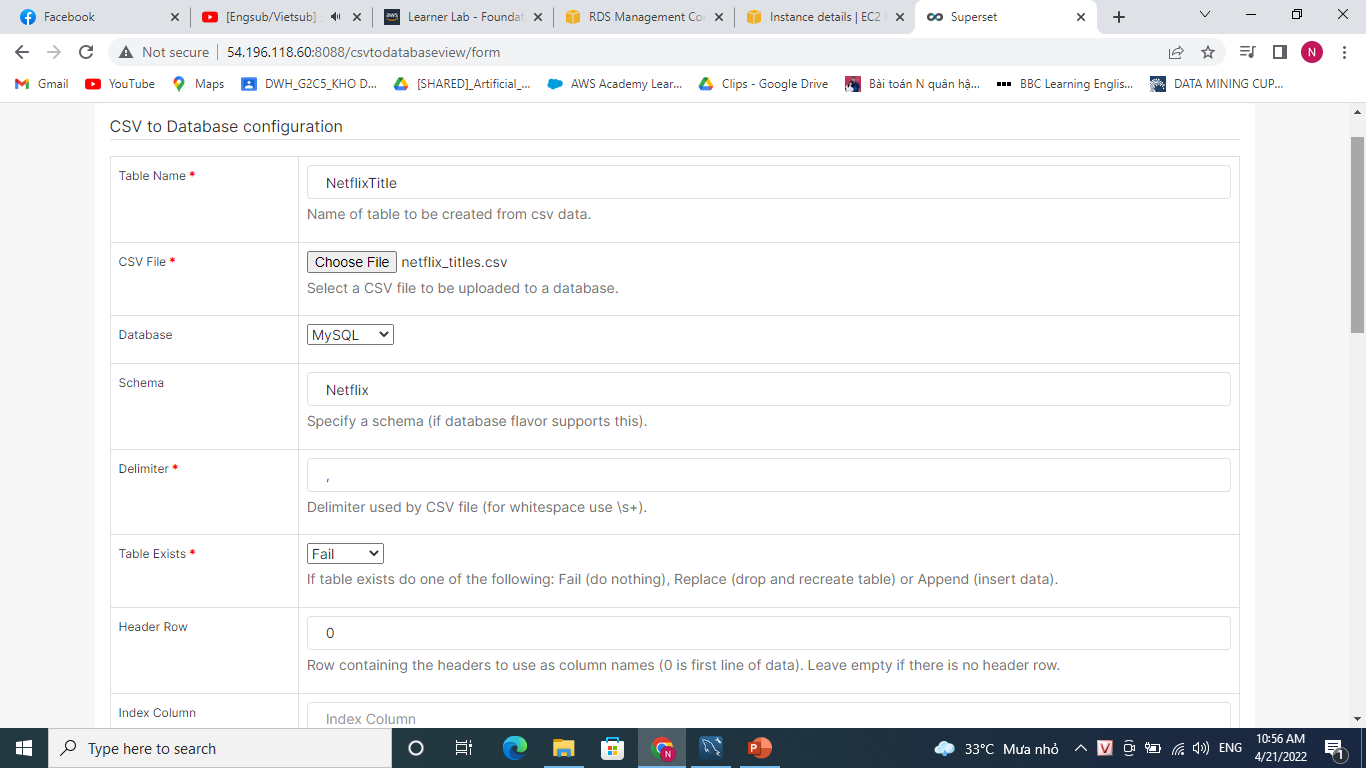
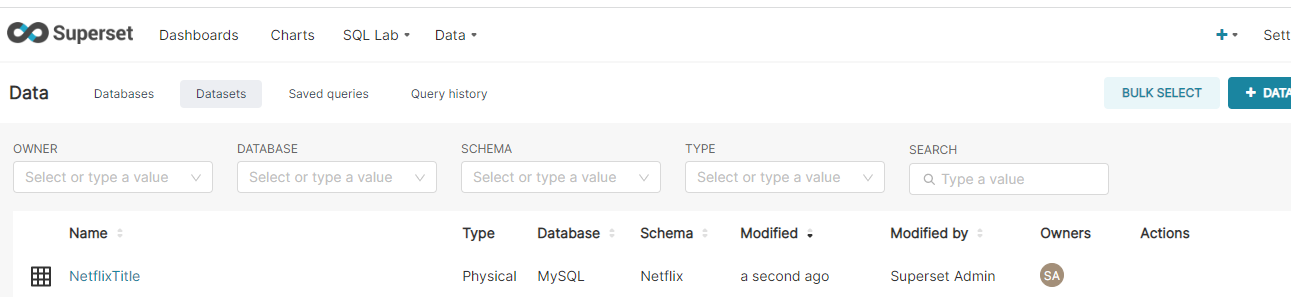
Rồi tiêp theo ta sẽ tải một tệp csv trên datasets theo Data-> Upload a CSV

Table name là tên của table

Csv file: tệp csv mà nhóm sẽ sử dụng

Database: kiểu db mà nhóm sẽ sử dụng

Nếu như ta ra như hình bên dưới tức là ta đã thành công việc kết nối database từ MySQL tới Superset

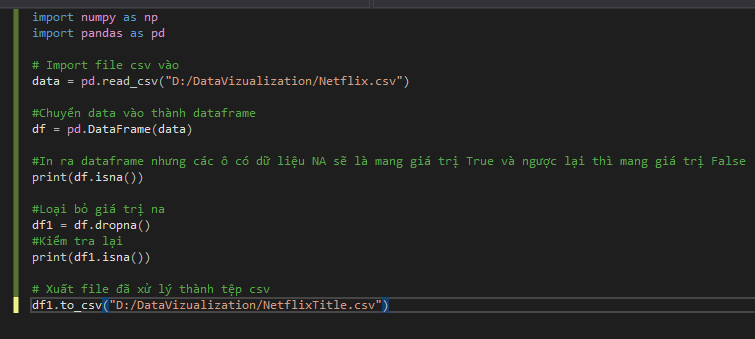
## **2.2.2Lọc dữ liệu bên ngoài sử dụng python và thực hiện một vài truy vấn đơn giản sử dụng SQL**

## **2.2.2.1 Xử lý dữ liệu bằng python**

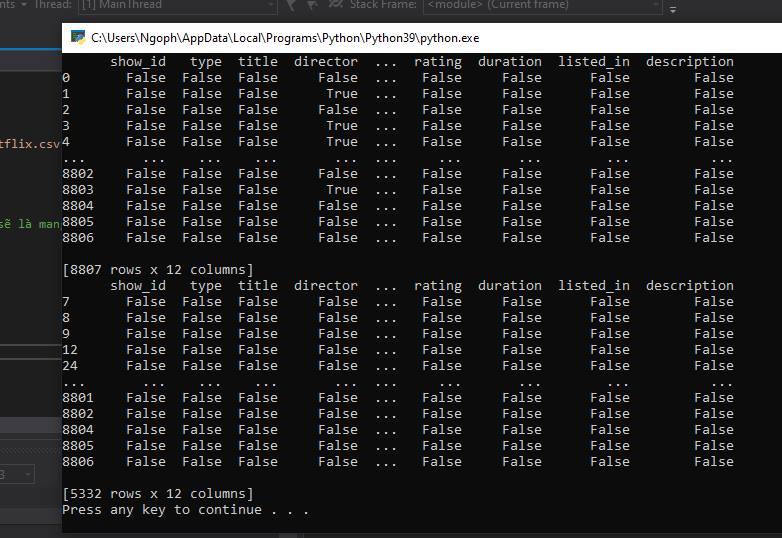
- Đây là file dữ liệu ban đầu của chúng ta.



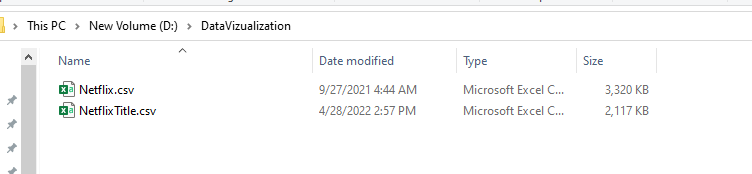
- Chúng ta sẽ sử dụng ngôn ngữ Python trên phần mềm Visual Studio Code để tiến hành xử lý dữ liệu.

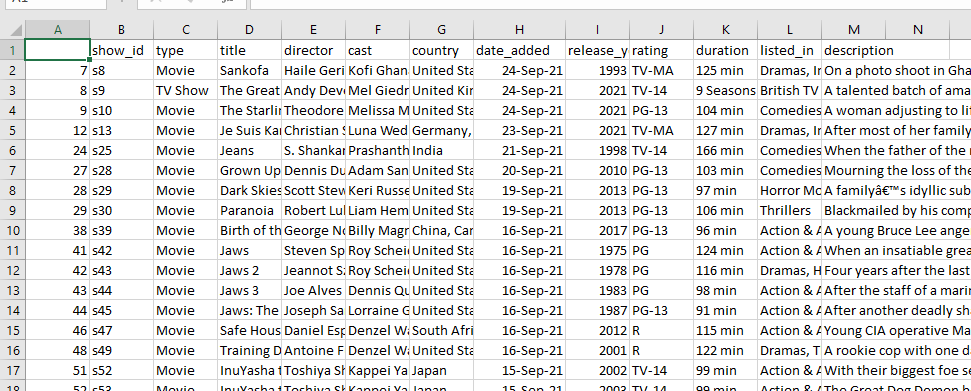


* Sau khi chạy code thì xuất hiện màn hình console như sau.
* Bảng đầu tiên ta thấy có các cột có giá trị True, có nghĩa là ô đó là dữ liệu lỗi.
* Bảng thứ 2 là bảng sau khi ta đã loại bỏ các hàng có dữ liệu lỗi.



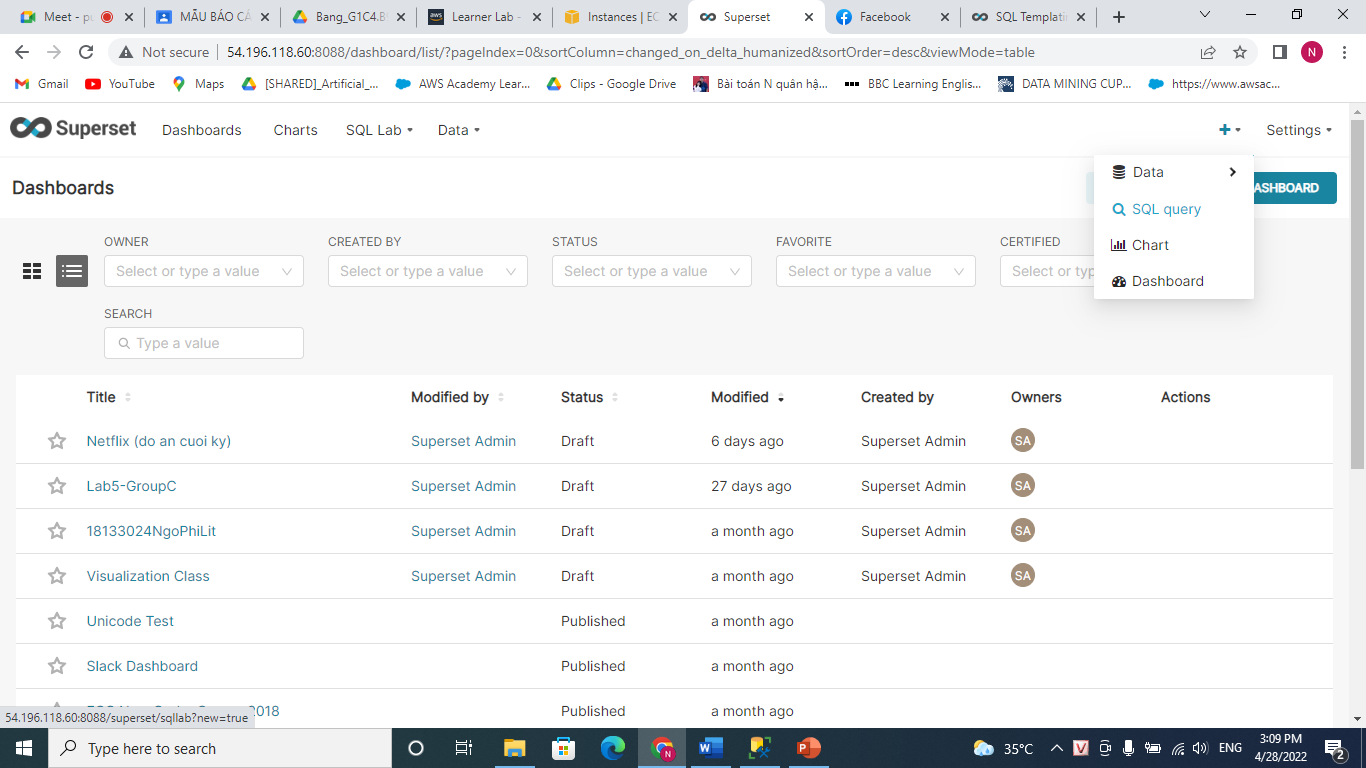
- Ta kiểm tra lại xem file csv đã được xuất ra chưa.



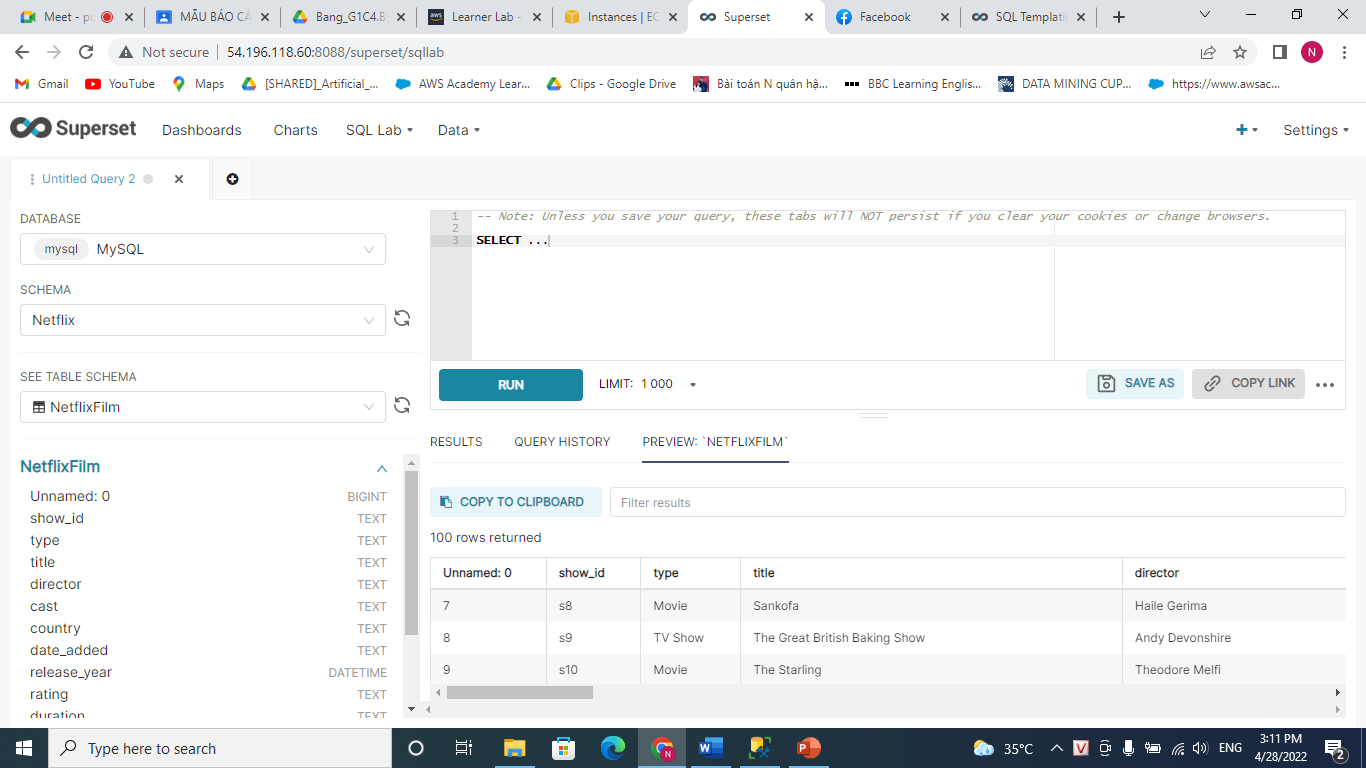


## **2.2.2.2 Sử dụng SQL trên Superset:**

- Ta đăng nhập Superset, sau đó vào mục SQL query.



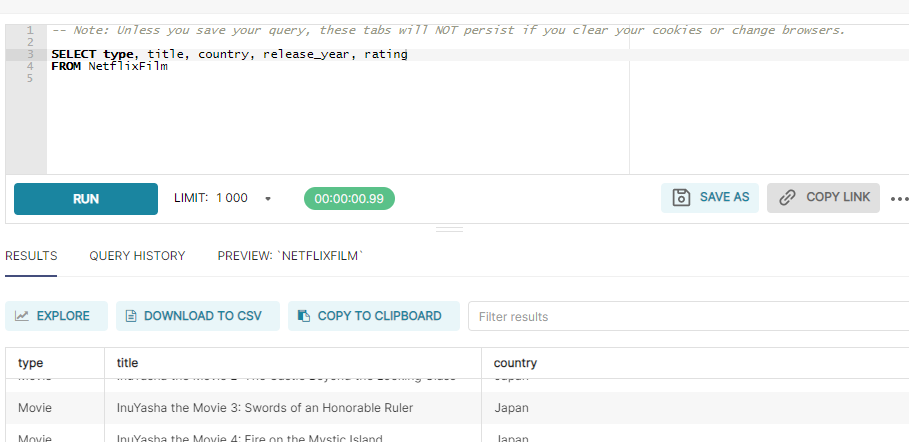
- Ta chọn Database, schema và table. Sau đó ta tiến hành viết SQL.



- Ví dụ đây là câu lệnh để in ra tất cả bộ phim trên Vương quốc Anh.

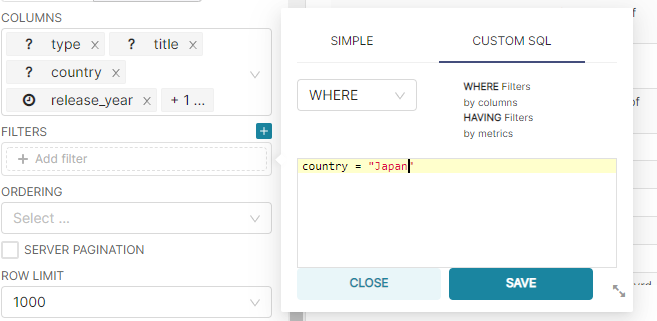


- Còn đây là lấy các cột type, title, country, release\_year, rating.



- Ta áp dụng SQL trong chart.

- Từ data sau khi thực hiện câu lệnh SQL ở trên. Ta tiếp tục viết lệnh SQL để in ra các bộ phim trên lãnh thổ của nước Nhật.



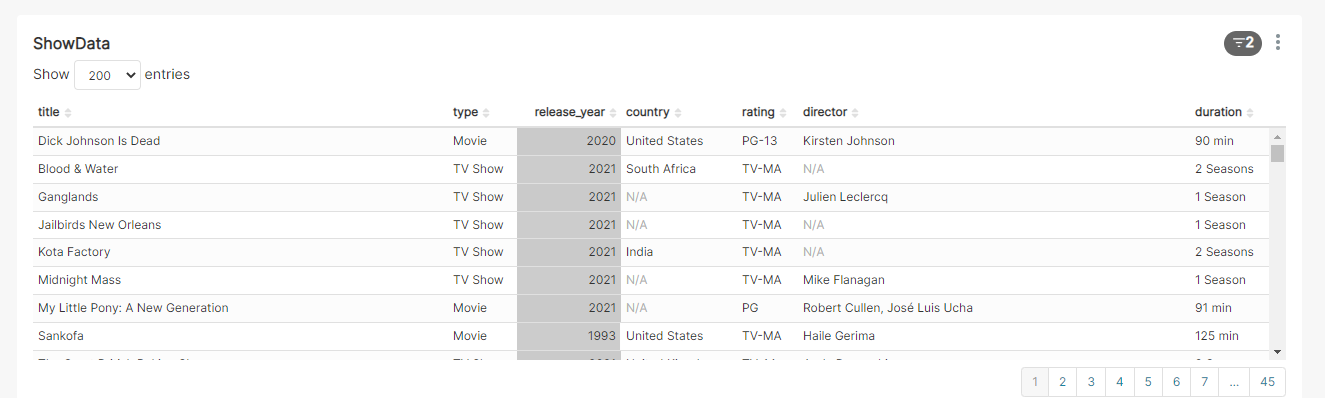
- Và đây là kết quả thu được.



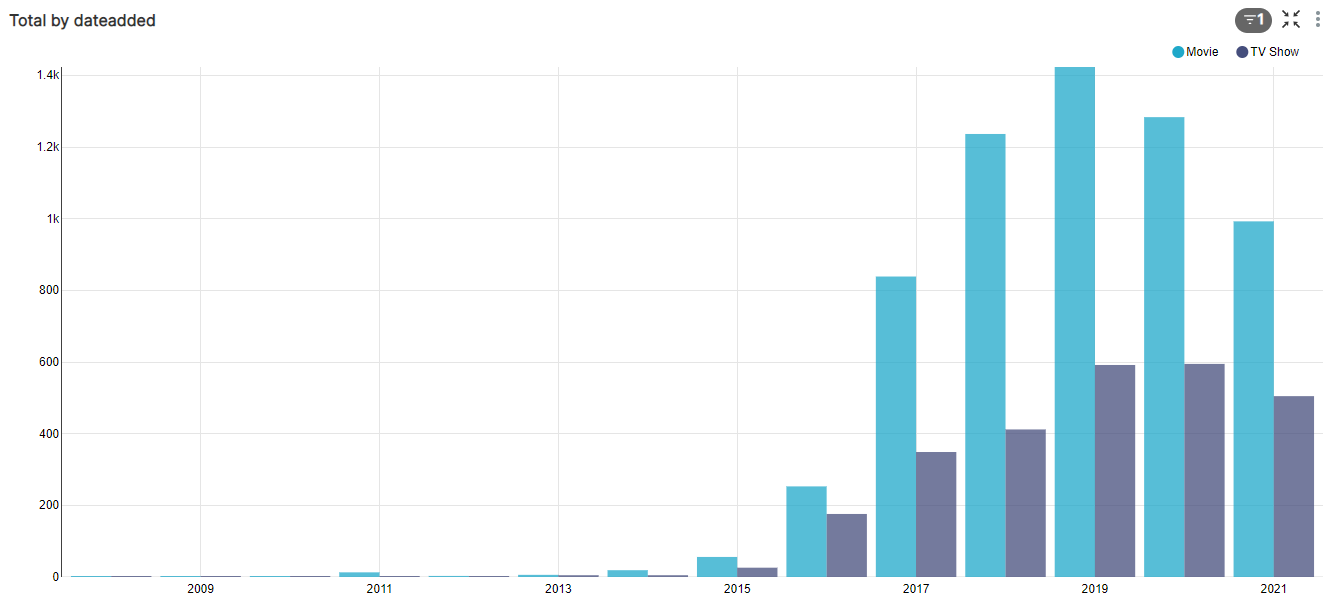
## **2.2.3Sử dụng Jinja Template để lọc dữ liệu trục tiếp trên Superset.**

- Sử dụng Jinja để tạo Filter Box cho Dashboard.



- Sửa đổi lại dashboard. Ở tuần trước nhóm chúng em có thực hiện 1 số biểu đồ về phim hoặc tivi show của một số khu vực( Nhật Bản và Mỹ). Và sau khi hoàn thành Filter Box thì nhóm cảm thấy nên gồm các biểu đồ trên thành 1 biểu đồ và sử dụng Filter Box để lọc ra các khu vực mà mình muốn xem. 

- Bên cạnh đó thì trong tuần qua nhóm cũng đã bổ sung them được một biểu đồ thống kê số lượng phim và chương trình truyền hình được thêm vào Netflix theo các năm.



# **PHẦN KẾT LUẬN**

## **Kết quả đạt được**

Sau một thời gian nghiên cứu và thực hiện đề tài “***Tìm hiểu và sử dụng Apache Superset***”, nhóm chúng em đã đạt được những kết quả như sau:

### *1.1. Kiến thức tìm hiểu được*

Nắm bắt được các kiến thức cũng như những vấn đề liên quan trọng về Apache Superset. Biết được cách tạo các biểu đồ trực quan một cách hệ thống, biết tải dữ liệu lên Superset, hiểu được công dụng và nền tảng của nó. Thực hiện truy vấn với sql và áp dụng kiến thức để thiết kế và trực quan hóa các biểu đồ trên Superset sử dụng SQL.

### *1.2. Chương trình đã làm được*

## Cài đặt và sử dụng Superset.

## Tải tệp dữ liệu lên Superset sử dụng MySQL.

## Tạo biểu đồ trực quan hóa, cũng như tạo các filter cho việc trực quan được rõ ràng, cụ thể hơn.

## Lọc dữ liệu từ bên ngoài sử dụng Python .

## Sử dụng sql để thực hiện truy vấn và tạo biểu đồ.

## Sử dụng jinja để lọc các dữ liệu.

## **Ưu điểm**

* Đã có một dashboard superset hiển thị các biểu đồ trực quan hóa một cách có hệ thống, liên kết và đa dạng.
* Có một vài biểu đồ được tạo từ code SQL.
* .Tập dữ liệu đã được lọc bằng python hoặc jinja.

## **Nhược điểm**

* Tập dữ liệu khá đơn giản nên các biểu đồ trực quan không được nhiều
* Chưa tạo được một ứng dụng có thể tự cập nhật data theo thời gian

## **Hướng phát triển**

* Tiếp tục hoàn thiện các chức năng còn thiếu.
* Cải thiện tập dữ liệu cho phù hợp hơn

# **TÀI LIỆU THAM KHẢO**

https://aws.amazon.com/vi/getting-started/hands-on/create-mysql-db/

https://www.youtube.com/watch?v=dE1QlhunWMs

Netflix Movies and TV Shows https://www.kaggle.com/datasets/shivamb/netflix-shows