**Xử lý sự kiện thời gian thực trong NiFi, SAM, Schema Registry và Superset**

**Giới thiệu**

-Trong hướng dẫn này, bạn sẽ học cách triển khai một ứng dụng xử lý dữ liệu dòng thời gian thực hiện đại. Ứng dụng này được sử dụng như một khung tham chiếu để phát triển một đường ống dữ liệu big data, bao gồm một loạt các trường hợp sử dụng và các thành phần cốt lõi có thể tái sử dụng mạnh mẽ. Bạn sẽ khám phá ứng dụng NiFi Dataflow, các chủ đề Kafka, Schemas, cấu trúc SAM và hiển thị các phần từ Superset.

**Yêu cầu tiên quyết**

-Tải xuống và triển khai Hortonworks DataFlow (HDF) Sandbox.(1)

-Bắt đầu với HDF.(2)

-Kích hoạt Connected Data Architecture (CDA):

+Kích hoạt CDA cho VirtualBox.(3)

+Kích hoạt CDA cho VMware.(4)

+Kích hoạt CDA cho Docker.(5)

**Tổng Quan**

-Khái niệm

-Tổng quan về ứng dụng tham chiếu IoT vận tải hàng hóa

+Bước 1: Cấu hình SAM cho CDA

+Bước 2: Cấu hình HDFS cho CDA

+Bước 3: Triển khai phiên bản CDA của Topology SAM

+Bước 4: Hiển thị dữ liệu vận tải hàng hóa qua Superset

-Summary

-Đọc thêm

-Phụ lục A: Trucking IoT GitHub Repo

**Các khái niệm**

**SuperSet**

SuperSet là một nền tảng khám phá dữ liệu trực quan, trực quan và tương tác. Nền tảng này cung cấp một cách nhanh chóng để tạo và chia sẻ bảng điều khiển với bạn bè và khách hàng kinh doanh của bạn từ các bộ dữ liệu được trực quan hóa. Có nhiều tùy chọn trực quan hóa khác nhau để phân tích và diễn giải dữ liệu. Semantic Layer cho phép người dùng kiểm soát cách các cửa hàng dữ liệu được hiển thị trên giao diện người dùng. Mô hình này an toàn và cho phép người dùng xác định các quy tắc phức tạp trong đó chỉ có một số tính năng cụ thể được truy cập bởi các cá nhân được chọn. SuperSet có thể tích hợp với Druid hoặc các cửa hàng dữ liệu khác (SQLAlchemy, Python ORM, v.v.) theo sự lựa chọn của người dùng, mang lại tính linh hoạt để tương thích với nhiều hệ thống.

**Druid**

Druid là một cơ sở dữ liệu phân tích mã nguồn mở được phát triển cho các truy vấn thông tin kinh doanh trên dữ liệu. Druid cung cấp khả năng tiếp nhận dữ liệu thời gian thực với độ trễ thấp, khả năng khám phá và tổng hợp dữ liệu linh hoạt. Các triển khai thường xuyên xử lý hàng nghìn tỷ sự kiện liên quan đến hàng petabyte dữ liệu.

**Tổng quan về Ứng dụng Tham chiếu IoT Vận tải hàng hóa**

Ứng dụng Tham chiếu IoT Vận tải hàng hóa được xây dựng bằng cách sử dụng Nền tảng Hortonworks DataFlow.

Dữ liệu IoT vận tải hàng hóa được tạo ra từ một bộ mô phỏng sự kiện của xe tải, được tiếp nhận bởi Apache NiFi. NiFi gửi dữ liệu đến các chủ đề Kafka, sau đó được tiếp nhận bởi Stream Analytics Manager (SAM) để lưu trữ vào Druid. Superset được sử dụng để tạo một biểu diễn trực quan của các nguồn dữ liệu Druid. Một giải thích chi tiết hơn về đường ống sẽ được giải thích khi bạn khám phá ứng dụng NiFi Dataflow, Schema Registry, SAM, Druid và Superset.

**Bước 1: Cấu hình SAM cho CDA**

1.Mở Stream Analytics Manager (SAM) tại địa chỉ <http://sandbox-hdf.hortonworks.com:7777/>

**Thêm một Service Pool cho HDP**

SAM cần được thông báo về vị trí của nút HDP để cho phép truyền dữ liệu giữa các nút HDF và HDP.

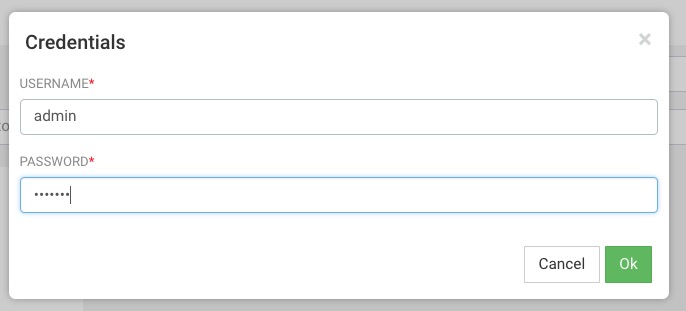
1.Di chuột qua biểu cặp tua-vít để mở danh sách thả xuống Cấu hình, chọn Service Pool, để chúng ta có thể thêm một service pool mới cho HDP.

Trong trường bên trái của AUTO ADD, bao gồm các thông tin sau:

<http://sandbox-hdp.hortonworks.com:8080/api/v1/clusters/Sandbox>

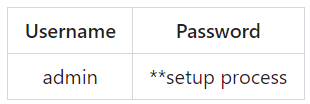
và sau đó nhấp vào AUTO ADD.

2.Sẽ xuất hiện một cửa sổ pop-up yêu cầu tên người dùng/mật khẩu như dưới đây:

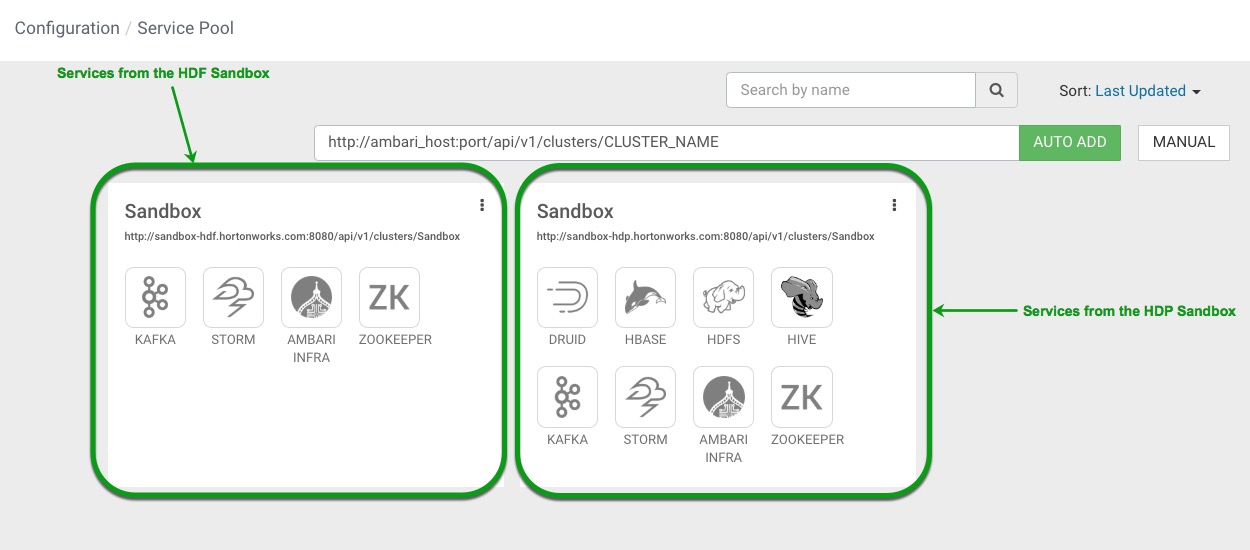


Sử dụng thông tin đăng nhập quản trị HDP của bạn để đăng nhập.

Bảng 1: Thông tin đăng nhập Ambari



Bây giờ bạn sẽ thấy các dịch vụ từ HDP Sandbox trên Bảng điều khiển Service Pool của SAM, nằm bên phải Service Pool của HDF Sandbox. Bây giờ chúng ta đã chia sẻ địa chỉ web mà SAM có thể tìm thấy các dịch vụ cho HDP Ambari Stack như được thấy trong hình ảnh dưới đây:

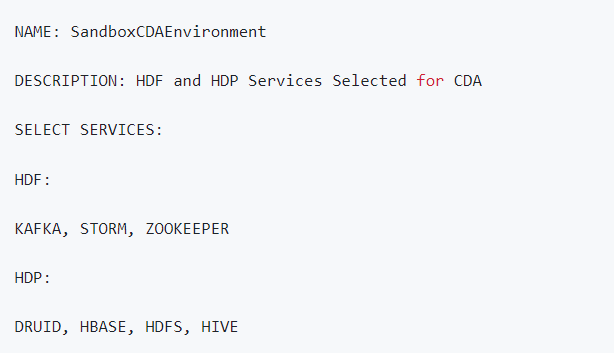


**Tạo một môi trường đa nút**

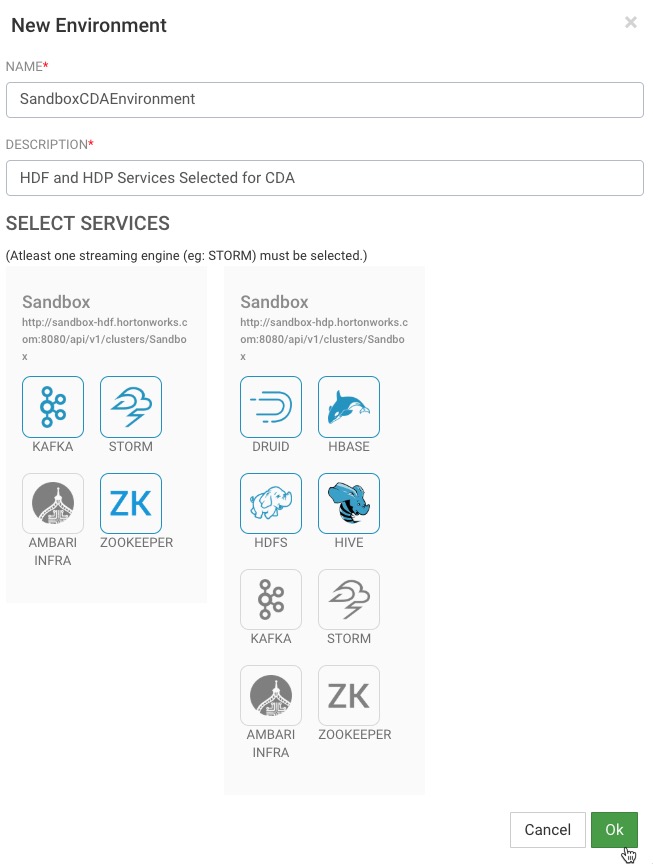
Với môi trường SAM, chúng ta có thể chọn các dịch vụ mà ứng dụng SAM của chúng ta muốn truy cập. Vì vậy, chúng ta sẽ thêm các dịch vụ cần thiết từ HDF và HDP Service Pools và tạo một môi trường đa nút từ những dịch vụ đó.

1.Điều hướng đến danh sách thả xuống Cấu hình, chọn Môi trường.

Nhấp vào biểu tượng cộng màu xanh lá cây để tạo một môi trường mới, thêm các thông tin sau:



Các cấu hình sẽ trông như thế này:



Bây giờ khi các cấu hình này đã hoàn tất, chúng ta đã sẵn sàng để nhập và chạy topology SAM của chúng ta cho CDA.

**Bước 2: Cấu hình HDFS cho CDA**

1.Mở HDP Shell-in-a-Box.

2.Đăng nhập với tên người dùng/mật khẩu: root/hadoop.

*Lưu ý: Nếu đây không phải lần đầu tiên bạn đăng nhập vào Shell-in-a-Box, hãy sử dụng mật khẩu trước đó bạn đã thiết lập cho root.*

3.Nhập các lệnh sau:

su hdfs

hdfs dfs -chmod 777 /apps

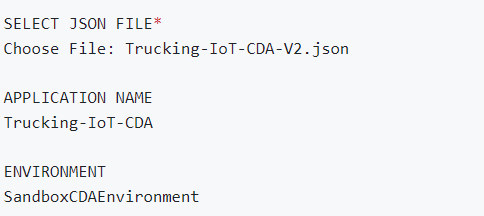
Các lệnh trên sẽ cho phép dịch vụ Storm trong HDF ghi vào HDFS nằm trong HDP.

**Bước 3: Triển khai Topology SAM phiên bản CDA**

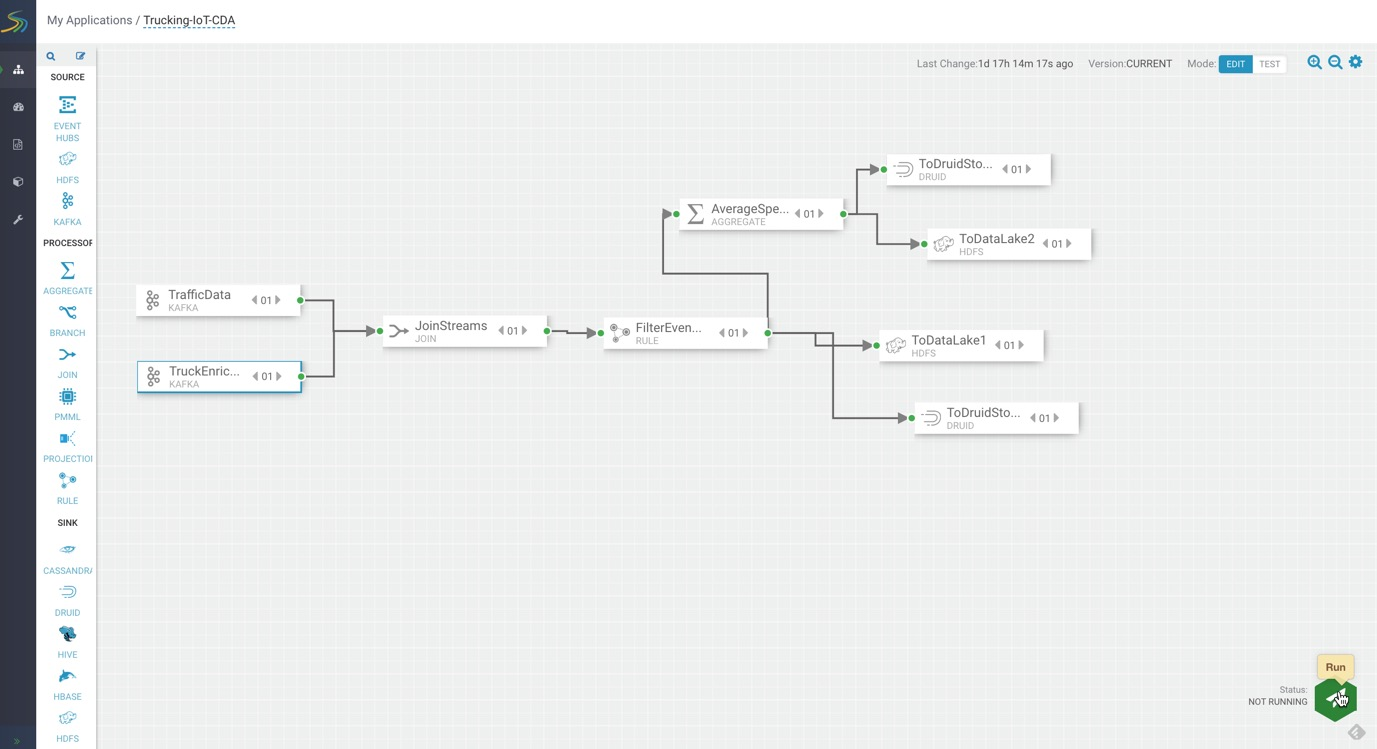
1.Mở Stream Analytics Manager (SAM) tại địa chỉ http://sandbox-hdf.hortonworks.com:7777/

2.Tải xuống topology SAM cho CDA và lưu nó ở một vị trí nào đó trên máy tính của bạn với tên "Trucking-IoT-CDA-V2.json".

3.Bây giờ trên giao diện SAM, nhấp vào biểu tượng cộng màu xanh lá cây và chọn Import Application.



4.Nhấp vào nút Run để triển khai topology:



Sẽ xuất hiện một cửa sổ hỏi bạn có muốn tiếp tục triển khai không, nhấp vào Ok.

5.Bạn sẽ nhận được thông báo rằng ứng dụng topology SAM đã triển khai thành công và topology của bạn sẽ hiển thị Trạng thái Hoạt động ở góc dưới bên phải.

**Tổng quan về SAM Canvas :**

-My Applications: Các dự án Topology khác nhau

-Thanh bên trái thứ nhất: My Applications, Dashboard, Schema Registry, Model Registry, Configuration

-Thanh bên trái thứ hai: Các thành phần luồng khác nhau (nguồn, xử lý, đích)

-Biểu tượng Gear: cấu hình thiết lập topology

-Biểu tượng Trạng thái: Bắt đầu hoặc Dừng Topology

**Tổng quan về SAM topology:**

-Nguồn **TrafficData** dữ liệu từ chủ đề Kafka "trucking\_data\_traffic"

-Nguồn **TruckEnrichedData** dữ liệu từ chủ đề Kafka "trucking\_data\_truck\_enriched"

-Dòng **JoinStreams** kết hợp dữ liệu TruckEnrichedData và TrafficData theo "routeId"

-Bộ lọc **FilterEvents** kiểm tra nếu eventType không phải là "Normal", sau đó sẽ phát ra các sự kiện đó

-Bộ tính trung bình **AverageSpeed** tính toán tốc độ trung bình cho driverId trên routeId

**-ToDruidStore1** lưu trữ các sự kiện vi phạm từ bộ xử lý FilterEvents vào Druid

**-ToDataLake1** lưu trữ các sự kiện vi phạm từ bộ xử lý FilterEvents vào HDFS

**-ToDruidStore2** lưu trữ các sự kiện tốc độ trung bình vào Druid

**-ToDataLake2** lưu trữ các sự kiện tốc độ trung bình vào HDFS

**Bước 4: Trực quan hóa dữ liệu xe tải thông qua Superset**

1.Mở Ambari tại địa chỉ http://sandbox-hdp.hortonworks.com:8080/. Thông tin đăng nhập người dùng là tên người dùng/mật khẩu = raj\_ops/raj\_ops.

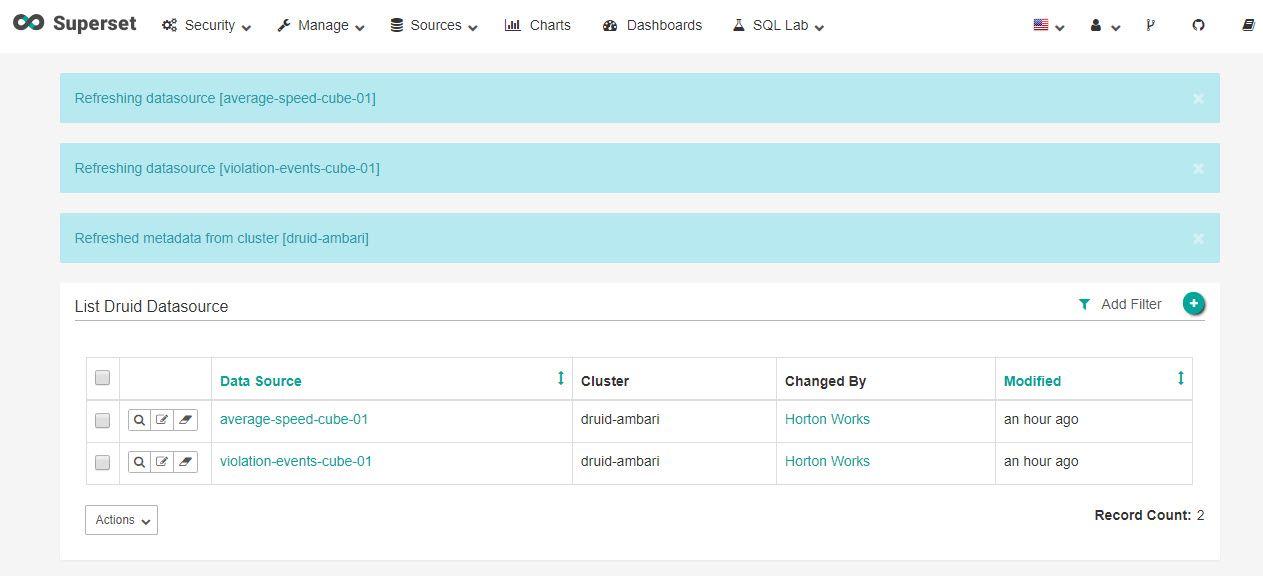
2.Đảm bảo rằng các dịch vụ HDFS, YARN, Druid và Superset đang hoạt động và tắt chế độ bảo trì.

*Ví dụ, để bật dịch vụ HDFS, nhấp vào tên dịch vụ trong Ambari, nhấp vào danh sách thả xuống Service Actions và nhấp vào Start. Trong cửa sổ, bạn sẽ được hỏi liệu bạn có muốn bắt đầu không, xác nhận và cũng nhấp vào ô kiểm để tắt chế độ bảo trì.*

3.Mở Superset tại địa chỉ http://sandbox-hdp.hortonworks.com:9089/

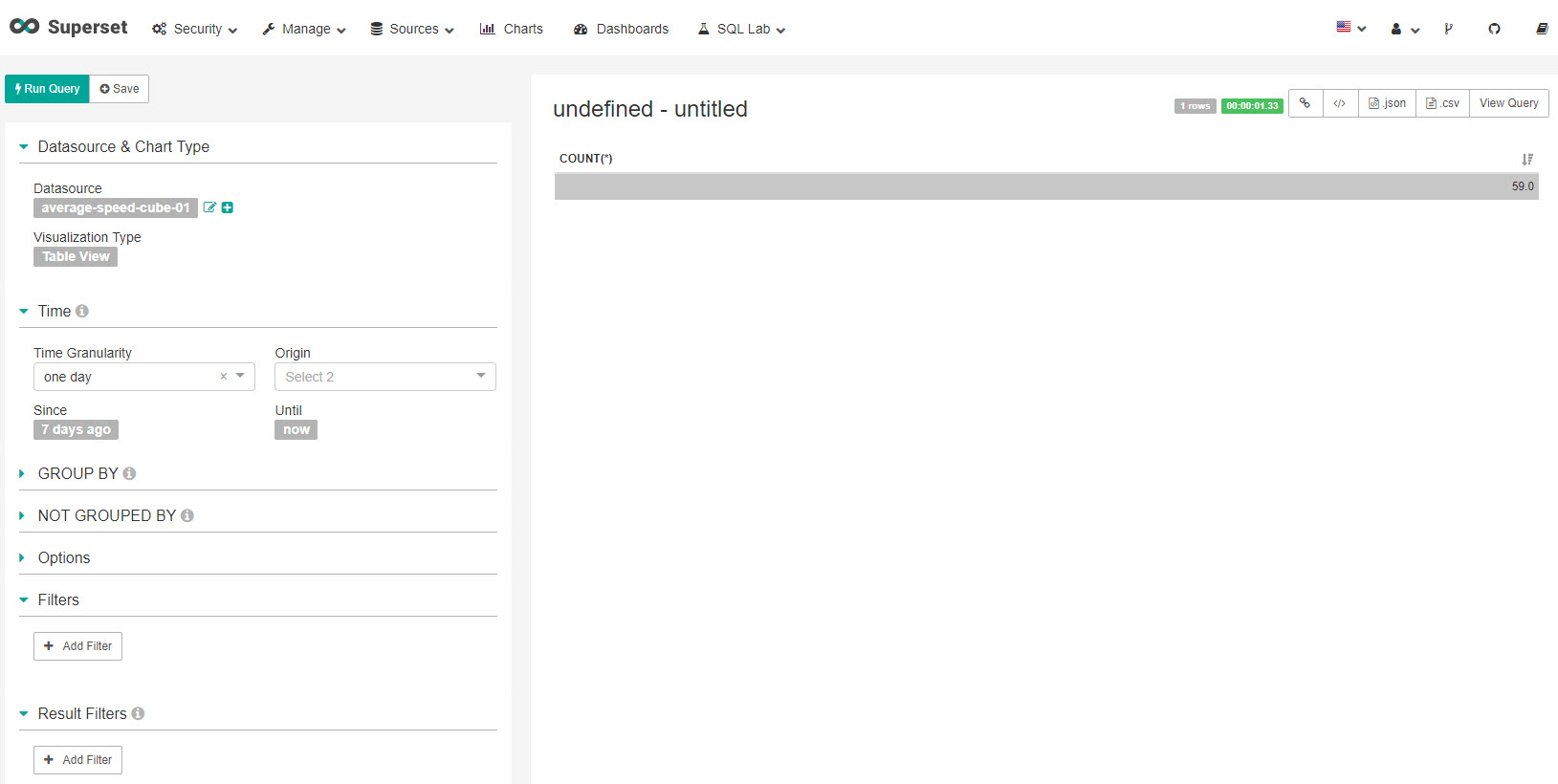
Lưu ý: Tên người dùng và mật khẩu cho Superset là admin/admin.

4.Chờ khoảng 25 - 30 phút để dữ liệu từ Kafka được tiêu thụ, sau đó định kỳ, chọn danh sách thả xuống Sources và nhấp vào Refresh Druid Metadata. Cuối cùng, hai nguồn dữ liệu Druid sẽ xuất hiện.

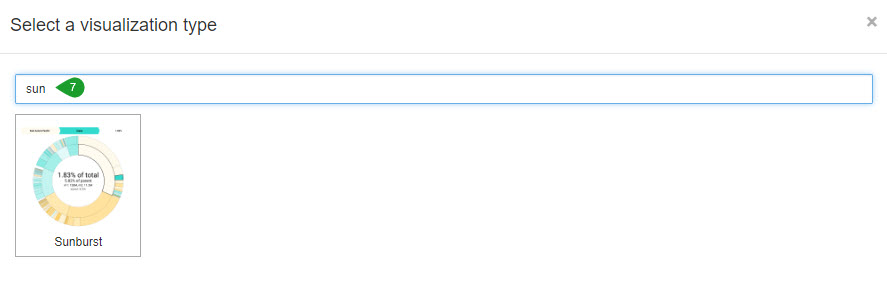


5.Chọn nguồn dữ liệu druid average-speed-cube-01.

6.Bạn sẽ được chuyển đến phần mảnh cắt trực quan hóa Superset, nơi bạn có thể trực quan hóa nguồn dữ liệu druid đó.

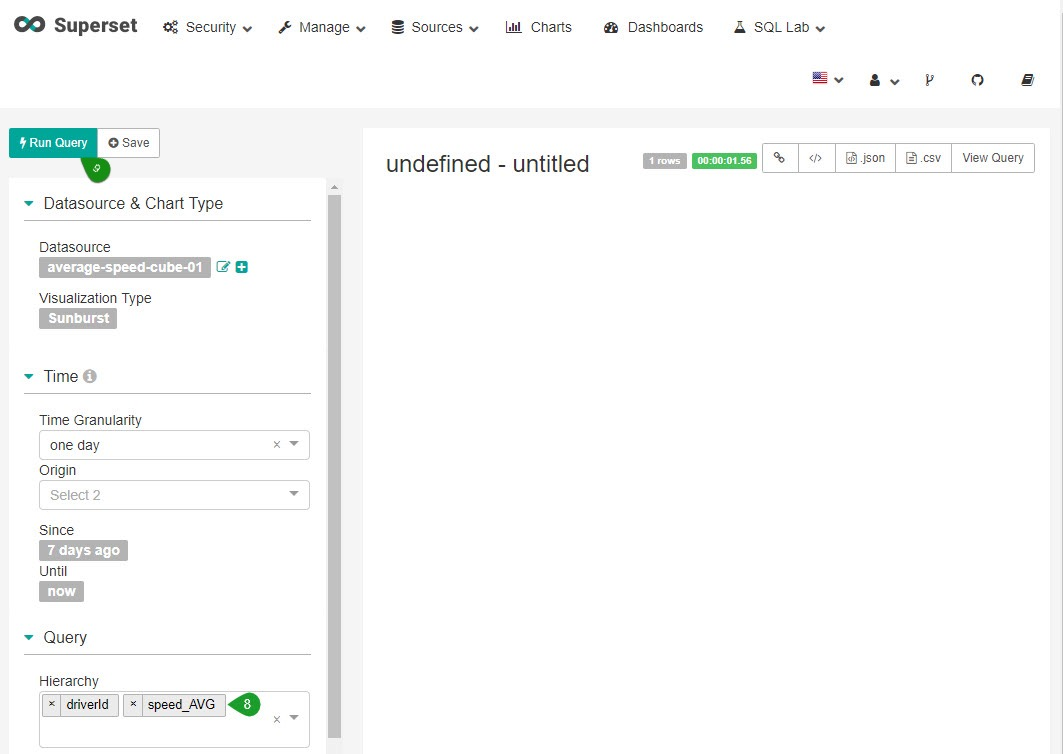


7.Dưới phần Datasource & Chart Type, chọn Loại trực quan hóa: Sunburst.

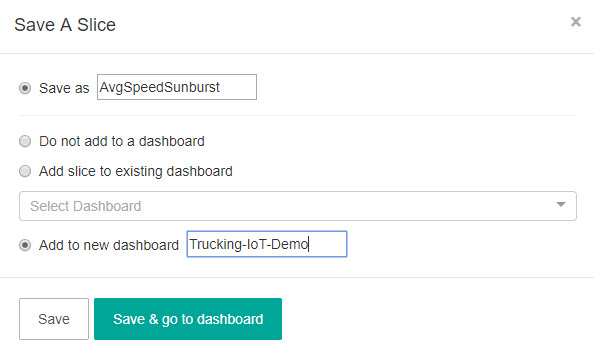


8.Dưới phần Hierarchy, thêm driverId, speed\_AVG.

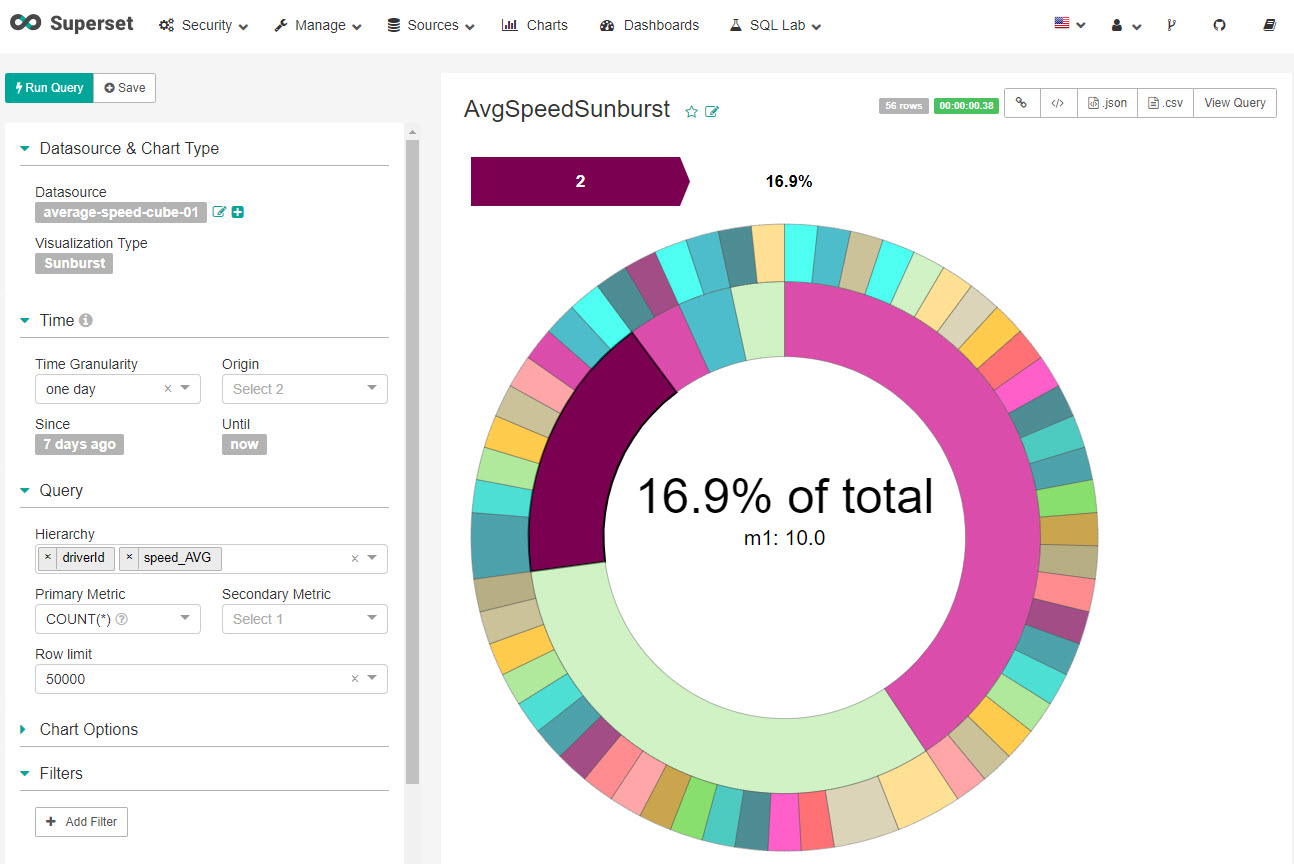
9.Nhấn vào truy vấn để trực quan hóa dữ liệu thành biểu đồ Sunburst.



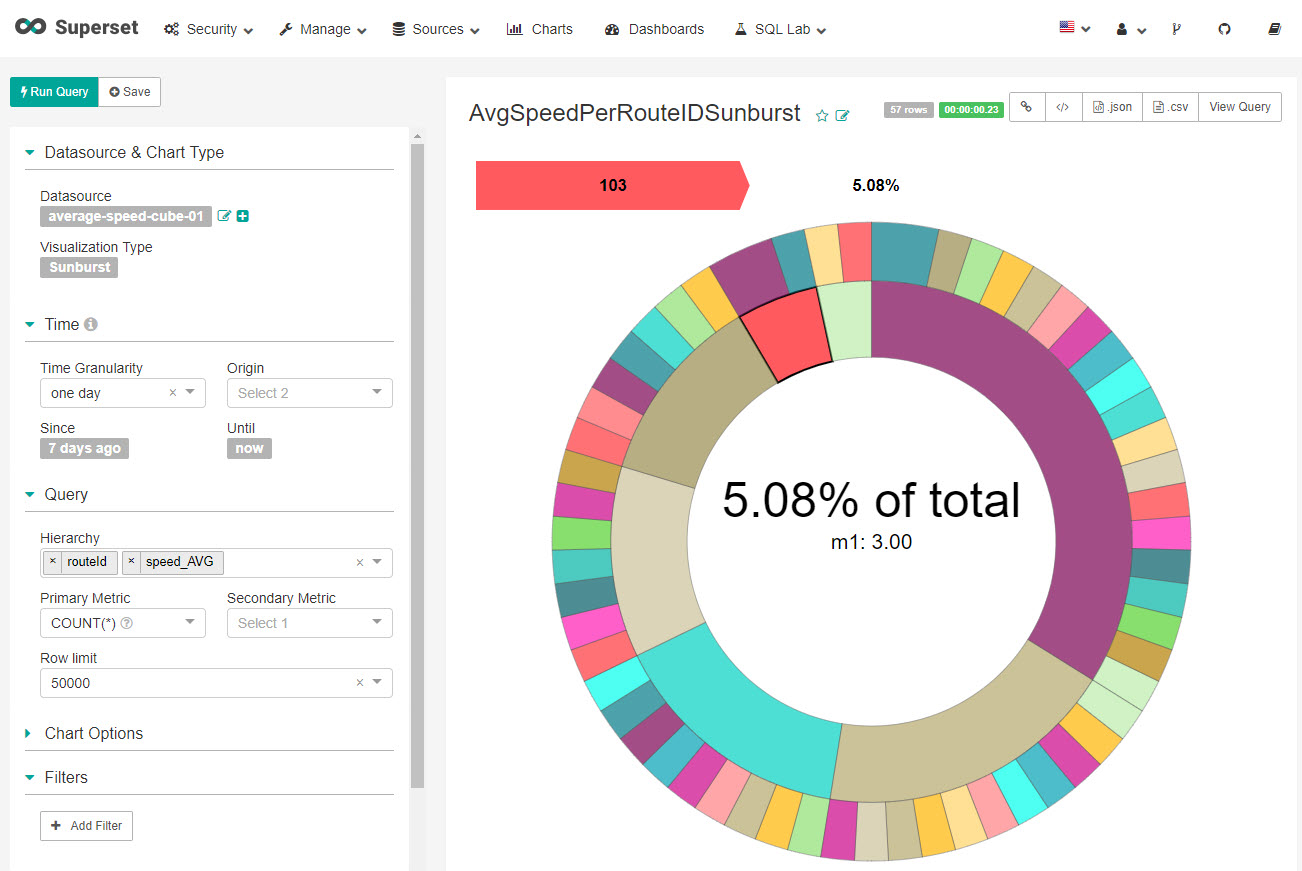
10.Chọn Lưu và đặt tên cho mảnh cắt: AvgSpeedSunburst. Tạo một bảng điều khiển mới và đặt tên cho nó: Trucking-IoT-Demo. Nhấp vào Lưu.



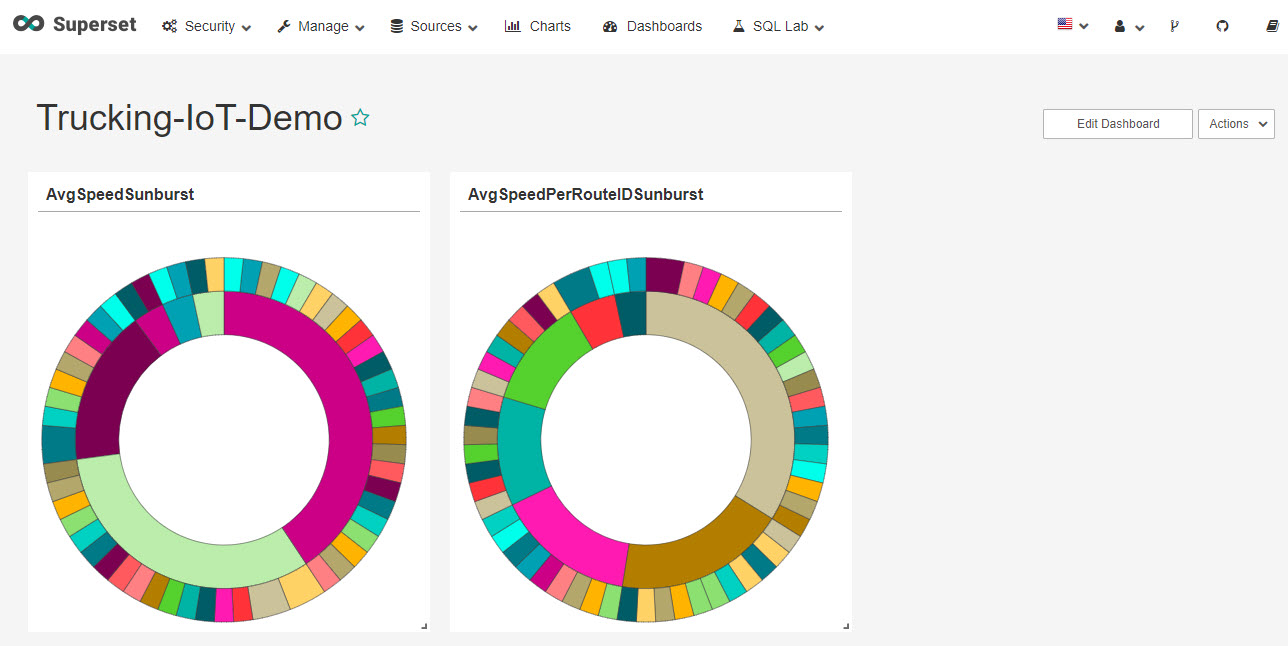
Mảnh cắt trực quan hóa sau đây là một "Sunburst" của nguồn dữ liệu average-speed-cube-01.



Mảnh cắt trực quan hóa sau đây là một "Sunburst" của nguồn dữ liệu violation-events-cube-01.



Tất cả các mảnh cắt trực quan hóa đã tạo sẽ được gửi đến bảng điều khiển mà bạn gán cho chúng, trong hai ví dụ trên, cả hai mảnh cắt đều thuộc bảng điều khiển Trucking-IoT-Demo.



**Summary**

Xin chúc mừng! Bạn đã triển khai demo CDA Trucking IoT xử lý dữ liệu sự kiện xe tải bằng cách sử dụng NiFi, Kafka, SAM, Schema Registry và SuperSet.

**Đọc thêm**

Superset : https://superset.apache.org/

Druid : https://druid.apache.org/

Phụ lục A: Trucking IoT GitHub Repo

https://github.com/orendain/trucking-iot/tree/master