# 1 Superset in Trucking IoT

## 1.1 ****Introduction****

Superset là một công cụ Business Intelligence được trang bị nhiều tính năng để thiết kế, duy trì và cho phép kể chuyện bằng dữ liệu thông qua các hình ảnh hóa dữ liệu có ý nghĩa. Công ty vận tải mà bạn đang làm việc có một Ứng dụng IoT cho xe tải xử lý dữ liệu về xe tải và giao thông mà nó nhận được từ các cảm biến, nhưng các nhà lãnh đạo doanh nghiệp không thể hiểu được dữ liệu. Họ đã thuê bạn làm Nhà phân tích trực quan hóa dữ liệu để kể một câu chuyện thông qua việc trực quan hóa dữ liệu của ứng dụng này, chẳng hạn như mức độ tắc nghẽn giao thông ảnh hưởng đến hiệu suất của người lái xe tải như thế nào, điều này cuối cùng sẽ ảnh hưởng đến công ty. Do đó, việc bạn truyền đạt những hiểu biết của mình cho các nhà lãnh đạo doanh nghiệp sẽ ảnh hưởng đến họ để hành động dựa trên các khuyến nghị của bạn.

## 1.2 Objective

* Học về khái niệm Trực quan hóa Dữ liệu
* Làm quen với Apache Superset
* Học cách Thiết kế Trực quan hóa với Superset

## 1.3 ****Prerequisites****

* Tải xuống và triển khai [Hortonworks Data Platform](https://www.cloudera.com/campaign/open-source-platform-self-supporters.html?utm_source=mktg-tutorial) (HDP) Sandbox
* Nếu bạn không có 32GB RAM riêng cho HDP Sandbox, vui lòng tham khảo Hướng dẫn triển khai [Hortonworks Sandbox](https://www.cloudera.com/tutorials.html) trên Microsoft Azure
* Kích hoạt Kiến trúc Dữ liệu Kết nối:

1. [Kích hoạt CDA cho VirtualBox](https://www.cloudera.com/tutorials.html#enable-connected-data-architecture-cda---advanced-topic)
2. [Kích hoạt CDA cho VMware](https://hortonworks.com/tutorial/sandbox-deployment-and-install-guide/section/2/#enable-connected-data-architecture-cda---advanced-topic)
3. [Kích hoạt CDA cho Docker](https://www.cloudera.com/tutorials.html#enable-connected-data-architecture-cda---advanced-topic)

## 1.4 ****Outline****

* [Khái niệm về Superset](https://hortonworks.com/tutorial/superset-in-trucking-iot/section/1/) - Bao gồm các khái niệm cơ bản về Trực quan hóa Dữ liệu và Superset.
* [Thiết lập Môi trường Phát triển](https://hortonworks.com/tutorial/superset-in-trucking-iot/section/1/) - Cài đặt ánh xạ tên máy chủ thành địa chỉ IP, thiết lập mật khẩu quản trị Ambari, bật các dịch vụ cần thiết cho Superset và bật Superset.
* [Trực quan hóa Dữ liệu Vận tải hàng hóa](https://www.cloudera.com/tutorials.html) - Hướng dẫn cách trực quan hóa dữ liệu bằng cách sử dụng Superset.

## 1.5 T****utorial Reference Application****

Loạt bài hướng dẫn này sử dụng [Ứng dụng IoT Vận tải hàng hóa](https://github.com/orendain/trucking-iot/tree/hadoop-summit-2017) của chúng tôi, bao gồm nhiều dự án phụ. Bạn sẽ xây dựng dự án phụ trợ của Superset cho việc trực quan hóa.

# 2 Superset Concepts

## 2.1 ****Objective****

Để hiểu tầm quan trọng của Trực quan hóa Dữ liệu, hãy bắt đầu ý tưởng về một số biểu đồ trực quan mà bạn có thể sử dụng để truyền thông tin về dữ liệu vận tải hàng hóa và làm quen với Superset.

## 2.2 Outline

* Cơ bản về Trực quan hóa Dữ liệu
* Lịch sử của Trực quan hóa Dữ liệu
* Tại sao sử dụng Superset?
* Thuật ngữ trong Superset
* Tóm tắt
* Đọc thêm

## 2.3 ****Data Visualization Fundamentals****

**Trực quan hóa Dữ liệu là gì?**

*"Trực quan hóa dữ liệu liên quan đến việc truyền đạt một câu chuyện hoặc ý tưởng một cách hiệu quả nhất có thể. Thường được nói rằng một bức tranh có giá trị ngang với nghìn từ. Trực quan hóa dữ liệu cũng hoạt động theo cách đó. Các nhà khoa học dữ liệu sử dụng trực quan hóa dữ liệu để khám phá các mẫu trong dữ liệu của họ và cuối cùng truyền đạt kết quả."*

*Ryan Orban (Giảng viên Udacity - Trực quan hóa Dữ liệu và D3.js, Bài học 1: Cơ bản về Trực quan hóa, 1. Trực quan hóa Dữ liệu là gì?)*

Nhiều người coi nó như một nghệ thuật và khoa học, có nhiều ứng dụng khác nhau:

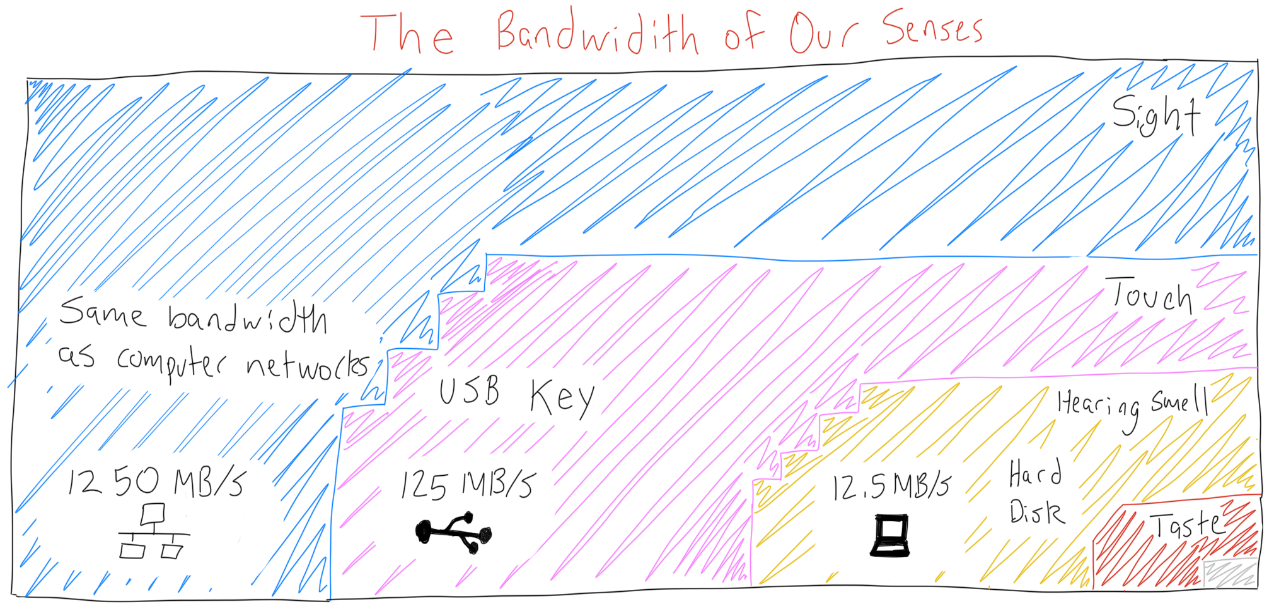
* Infographics (Đồ họa thông tin)
* Information Visualization (Trực quan hóa thông tin)
* Scientific Visualization (Trực quan hóa khoa học)
* Exploratory Data Analysis (Phân tích Dữ liệu khám phá)
* Statistical Graphics (Đồ họa thống kê)

**Tại sao sử dụng Trực quan hóa Dữ liệu?**

Chúng ta nên sử dụng trực quan hóa dữ liệu thay vì biểu diễn dữ liệu dưới dạng số vì con người có khả năng tốt hơn trong việc hiểu quan hệ trong dữ liệu khi nhìn thấy nó trực tiếp dưới dạng biểu đồ.

**Trực quan hóa Dữ liệu và Nhận thức của con người**

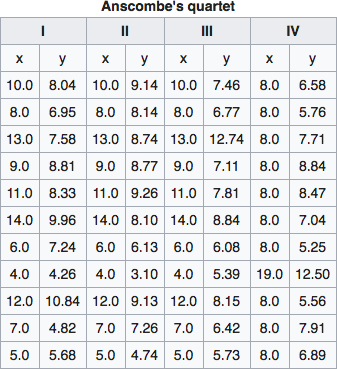
Trực quan hóa dữ liệu giúp não bộ của chúng ta nhận thức dữ liệu. Theo nhà vật lý học Tor Norretrander, thị giác chiếm phần lớn không gian trong băng thông của các giác quan của chúng ta và có thể xử lý thông tin với tốc độ của các mạng máy tính hoặc cáp Ethernet, như được minh họa trong sơ đồ sau (Khóa học Udacity - Trực quan hóa Dữ liệu với Tableau, Bài học 1: Cơ bản về Trực quan hóa Dữ liệu).



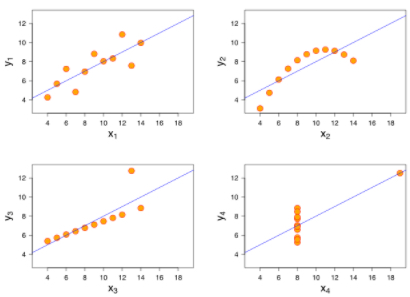
Hình 1: **Sơ đồ vẽ tay dựa trên "Khóa học Udacity - Trực quan hóa dữ liệu với Tableau, Bài 1: Nguyên tắc cơ bản về trực quan hóa dữ liệu, 3. Tại sao nên sử dụng trực quan hóa dữ liệu"**

**Thống kê mô tả của Anscombe so với Biểu đồ**

Nhà thống kê Francis Anscombe đã tạo ra một tình huống trong đó ông có 4 bộ dữ liệu với 11 điểm (x, y) có cùng các thống kê mô tả: trung bình, phương sai, hệ số tương quan và đường thẳng phù hợp nhất. Nếu không có trực quan hóa dữ liệu, có vẻ như các bộ dữ liệu không có sự khác biệt đáng kể như được chỉ ra trong bảng.



**Nhóm dữ liệu của Anscombe** cho thấy luôn nên vẽ biểu đồ dữ liệu của bạn thay vì chỉ dựa vào thống kê mô tả. Trong trường hợp của bảng trên với 4 bộ dữ liệu, các thống kê tóm tắt không khác nhau, nhưng biểu đồ lại khác nhau.



Những gì có thể thấy từ các biểu đồ khiến thống kê tóm tắt của chúng ta bị sai lệch?

Tác động của:

* Độ cong: Nếu dữ liệu không được phân bố đều, các thống kê tóm tắt như trung bình và độ lệch chuẩn có thể không phản ánh chính xác giá trị thực của dữ liệu.
* Giá trị ngoại lai: Những giá trị nằm ngoài phạm vi thông thường của dữ liệu có thể làm sai lệch các thống kê tóm tắt.

Ví dụ trên củng cố lý do tại sao chúng ta nên sử dụng trực quan hóa dữ liệu thay vì chỉ hiển thị các con số. Đồ họa cho phép chúng ta thấy rõ ràng các độ lệch hoặc mô hình trong dữ liệu trong khi bảng và số liệu thống kê không thể làm được điều đó.

**Tại sao đề cập đến Kiểu dữ liệu?**

Kiểu dữ liệu được sử dụng khi tạo hoặc tái tạo hình ảnh. Chúng ta phải hiểu các kiểu dữ liệu khác nhau được tìm thấy trong trực quan hóa dữ liệu:

1. Dữ liệu định lượng (Quantitative Data)

* Là bất kỳ biến nào có số chính xác, có thể là biến rời rạc (discrete) hoặc biến liên tục (continuous).
* Biến rời rạc: là một biến có thể đếm được.
* Ví dụ: số đơn vị bán được, số ngôn ngữ biết, số email nhận được.
* Biến liên tục: là một biến nằm trong một khoảng.

Ví dụ: thời gian, chiều cao, cân nặng, tiền, lãi suất, nhiệt độ.

2. Dữ liệu định tính (Qualitative Data)

* Là bất kỳ biến nào có nhãn hoặc thể loại, có thể là biến danh mục (categorical) hoặc biến thứ tự (ordinal).
* Biến danh mục (Categorical) (Nominal): là một biến dán nhãn hoặc phân loại dữ liệu thành các nhóm.

Ví dụ: giới tính, màu tóc, quốc gia, ngành nghề, giống mèo.

Biến thứ tự (Ordinal): là một biến xếp hạng các thể loại.

Ví dụ: xếp hạng, mức độ khó của lớp học, các câu hỏi khảo sát như "Bạn cảm thấy thế nào về nhím?"

* Tôi yêu chúng
* Tiêu cực
* Trung lập
* Tích cực

Nguồn: Udacity Course - Data Visualization with Tableau, Bài 1: Fundamentals of Data Visualization

**Đặc điểm của hình ảnh hóa hiệu quả**

"Hiển thị đồ họa nên:

1. Hiển thị dữ liệu
2. Dẫn dắt người xem suy nghĩ về nội dung thay vì về phương pháp luận, thiết kế đồ họa, công nghệ sản xuất đồ họa hoặc một thứ gì khác
3. Tránh bóp méo những gì dữ liệu muốn nói
4. Trình bày nhiều số trong một không gian nhỏ
5. Làm cho các bộ dữ liệu lớn trở nên mạch lạc
6. Khuyến khích người xem so sánh các phần dữ liệu khác nhau
7. Tiết lộ dữ liệu ở nhiều mức độ chi tiết, từ tổng quan rộng đến cấu trúc chi tiết
8. Phục vụ một mục đích rõ ràng: mô tả, khám phá, bảng hóa hoặc trang trí
9. Được tích hợp chặt chẽ với các mô tả thống kê và bằng lời của một tập dữ liệu.

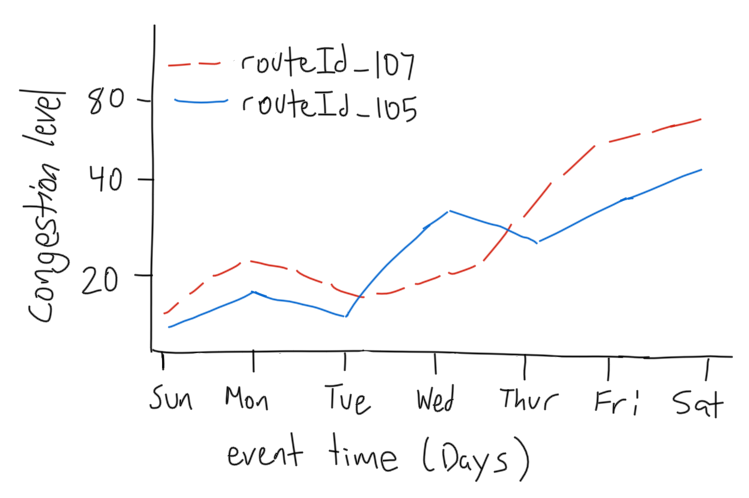
Đồ họa tiết lộ dữ liệu." - Wikipedia

**Biểu đồ thường dùng để truyền đạt thông điệp về dữ liệu**

Khi bạn thử nghiệm các cách để trực quan hóa dữ liệu hiệu quả, các biểu đồ bạn có thể muốn sử dụng để truyền đạt thông điệp định lượng về trường hợp sử dụng của mình (Trucking IoT) bao gồm:

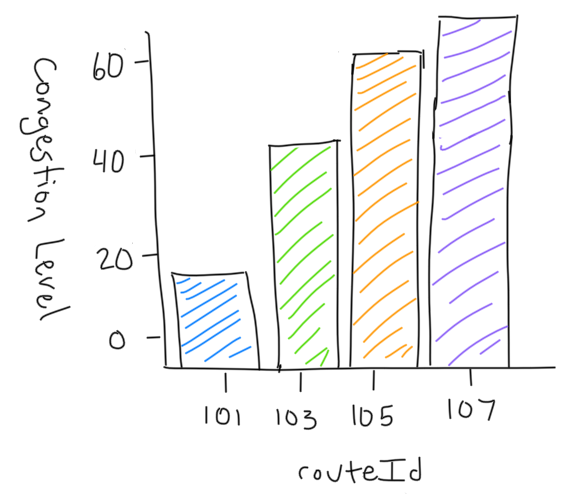
1. Dòng thời gian (Time Series): Biến đơn được ghi lại trong một khoảng thời gian.

* Biểu đồ đường (Line Chart): có thể hiển thị mức độ tắc nghẽn trên một tuyến đường cụ thể trong khoảng thời gian của các ngày trong tuần, tháng, năm, v.v. (Vui lòng đưa hình ảnh vào khi chúng nâng cao nội dung.)



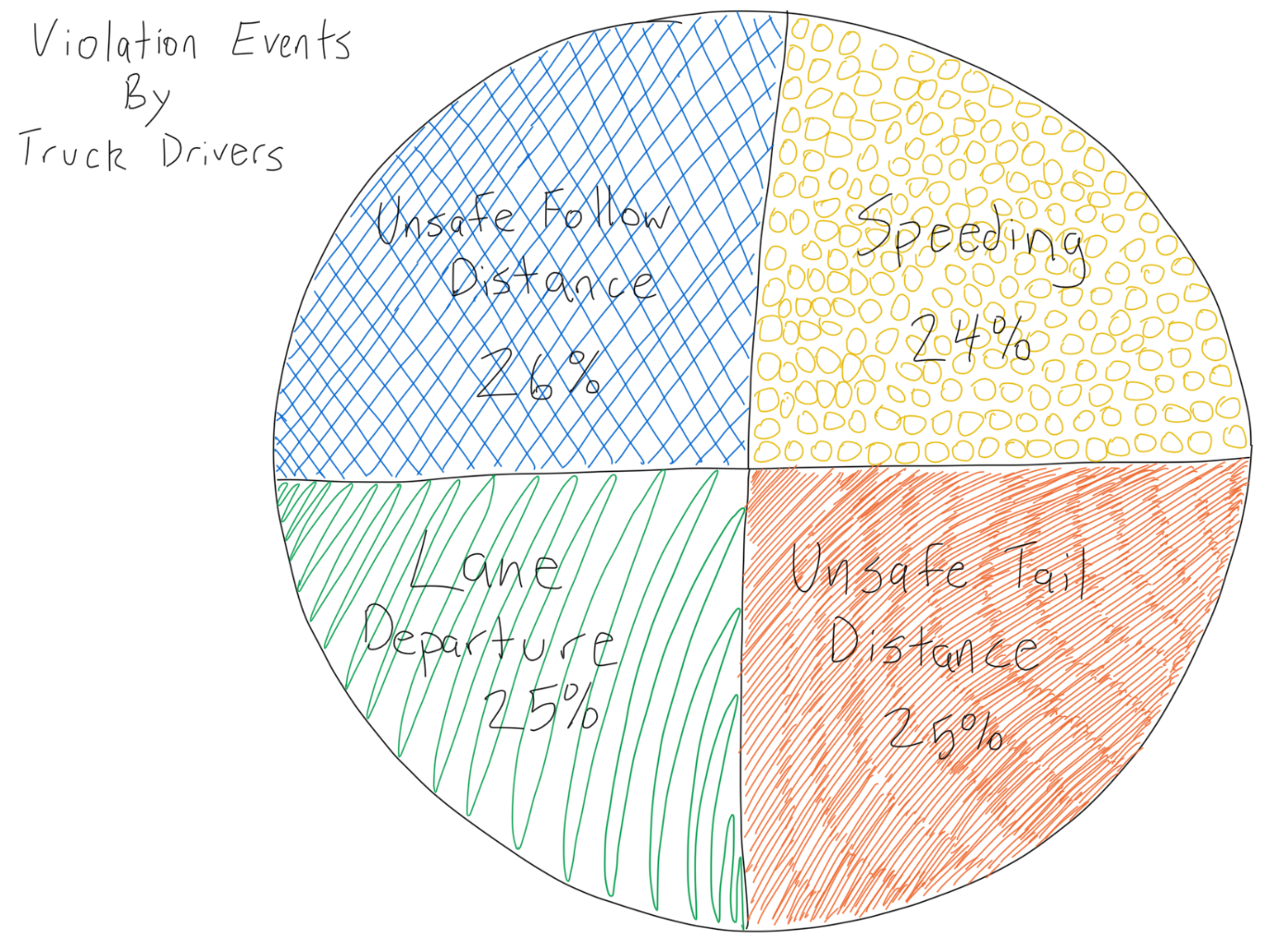
2. Xếp hạng (Ranking)

* Phân loại theo danh mục được xếp hạng theo thứ tự tăng dần hoặc giảm dần.
* Biểu đồ thanh (Bar Chart): có thể xếp hạng theo thứ tự tăng dần mức độ tắc nghẽn giao thông (biện pháp) theo tuyến đường (thể loại) trong một khoảng thời gian duy nhất.



3. Tỷ lệ phần-toàn bộ (Part-to-whole)

* Phân loại theo danh mục được đo lường như một tỷ lệ so với tổng thể (tỷ lệ phần trăm của 100%).
* Biểu đồ tròn (Pie Chart): hoặc Biểu đồ thanh (Bar Chart): có thể hiển thị tỷ lệ phần trăm của các sự kiện vi phạm của người lái xe tải: Chệch làn đường (Lane Departure), Khoảng cách theo đuôi không an toàn (Unsafe Follow Distance), Quá tốc độ (Speeding), Khoảng cách đuôi không an toàn (Unsafe Tail Distance) so với tổng thể.

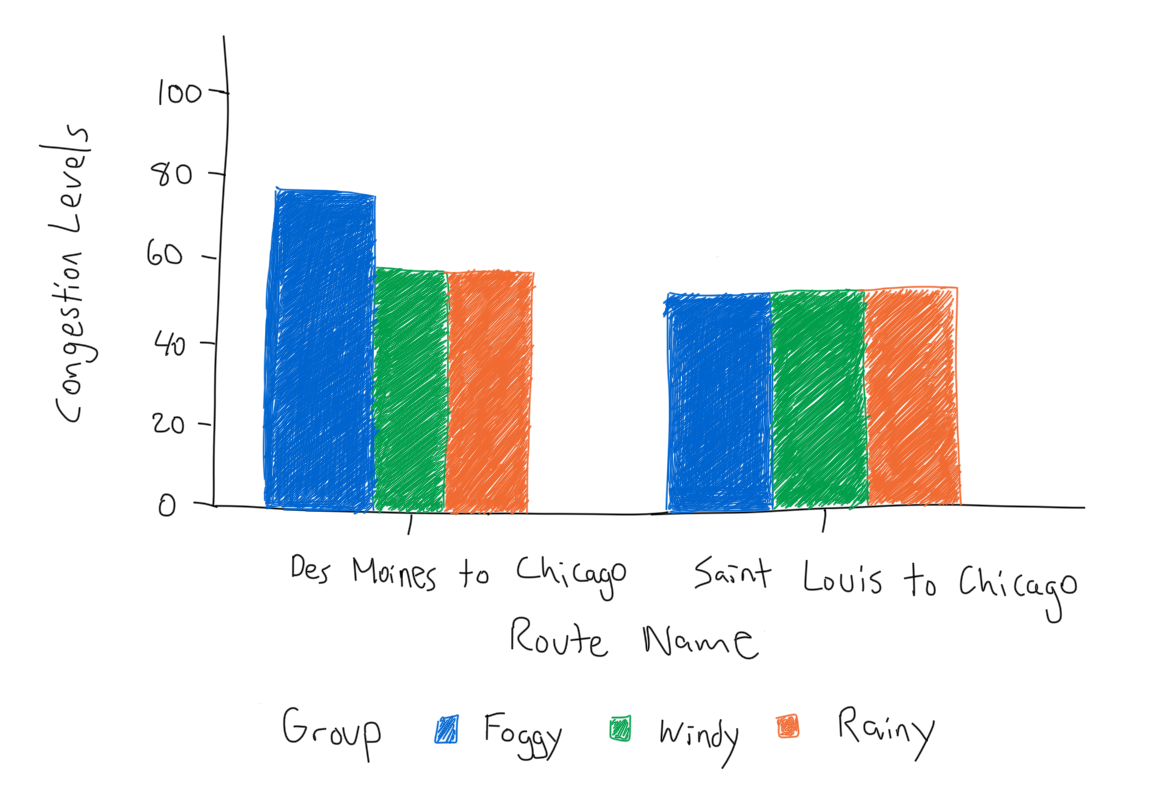


4. Độ lệch (Deviation):

* Phân loại theo danh mục được so sánh với một tham chiếu.
* Biểu đồ thanh (Bar Chart): có thể hiển thị chi phí thực tế so với ngân sách cho các bộ phận kinh doanh trong một khoảng thời gian.

5. Phân bố tần suất (Frequency Distribution):

* Hiển thị các quan sát của một biến cho một khoảng nhất định.
* Biểu đồ cột (Histogram) hoặc/và biểu đồ hộp (Box Plot)
* Biểu đồ cột (Histogram): có thể hiển thị mức độ tắc nghẽn giao thông trong dữ liệu cho tuyến đường Des Moines đến Chicago và Saint Louis đến Chicago trong những ngày có sương mù, gió hoặc mưa. Biểu đồ cột sau minh họa thông tin này.
* Biểu đồ hộp (Box Plot): có thể hình dung các thống kê về mức độ tắc nghẽn giao thông trên các tuyến đường này: trung vị, tứ phân vị, giá trị ngoại lai, v.v.



6. Tương quan (Correlation):

* So sánh giữa hai biến (x, y) để xác định xem chúng có chuyển động cùng hướng hay ngược hướng.
* Biểu đồ phân tán (Scatter Plot): có thể biểu thị thất nghiệp (x) và lạm phát (y) cho một mẫu tháng.

7. Địa lý hoặc Không gian địa lý (Geographical or Geospatial):

* So sánh một biến trên bản đồ hoặc bố cục.
* Biểu đồ bản đồ (Cartogram): có thể hiển thị số người trên các tầng khác nhau của một tòa nhà.

Khi bạn thử nghiệm các cách trực quan hóa dữ liệu để truyền đạt thông điệp và mối tương quan có ý nghĩa, bạn sẽ dabbble vào Phân tích dữ liệu khám phá (Exploratory Data Analysis), cung cấp thông tin chi tiết về tập dữ liệu thông qua thử nghiệm và lỗi.

**Áp dụng nhận thức của con người để thiết kế biểu đồ trực quan**

Kiến thức về nhận thức và nhận thức của con người là điều cần thiết để truyền đạt thông điệp mà bạn muốn truyền tải từ biểu đồ vì tất cả các hình ảnh đều được tạo ra để con người sử dụng. Nhận thức của con người là cách con người xử lý thông tin trực quan. Do đó, quá trình xử lý thị giác của con người bao gồm phát hiện những thay đổi và thực hiện so sánh giữa các số lượng, kích thước, hình dạng và sự khác biệt về độ sáng. Miễn là các tính chất của dữ liệu biểu tượng được ánh xạ đến các tính chất trực quan được biểu diễn bởi dữ liệu đó, con người có thể duyệt qua khối lượng lớn dữ liệu một cách hiệu quả. Do đó, trực quan hóa dữ liệu cho phép người dùng thực hiện các khám phá dữ liệu.

## 2.4 ****History of Data Visualization****

Dấu vết được ghi lại đầu tiên về trực quan hóa dữ liệu là Bản đồ Papyrus Turin và có thể bắt nguồn từ 1160 TCN. Bản đồ minh họa thông tin về vị trí của các tài nguyên địa chất và cách khai thác các tài nguyên đó. Bản đồ cũng có thể hiển thị bản đồ theo chủ đề, là một minh họa địa lý được thiết kế để thể hiện một chủ đề liên quan đến một khu vực địa lý cụ thể. Con người cũng đã tìm thấy các hình thức trực quan hóa dữ liệu khác nhau thông qua các bức vẽ trên vách hang và các hiện vật vật lý từ Mesopotamia, Inca và các nhóm bản địa khác.

Phát minh ra giấy và giấy da là một bước đột phá lớn cho phép phát triển thêm các hình ảnh trực quan. Vào thế kỷ 16, các kỹ thuật và dụng cụ đã được phát triển để đo lường chính xác các đại lượng vật lý, vị trí địa lý và thiên thể. Với sự phát triển của hình học giải tích và hệ thống tọa độ hai chiều bởi Rene Descartes và Pierre de Fermat, họ đã ảnh hưởng đến các phương pháp chúng ta sử dụng để hiển thị và tính toán các giá trị dữ liệu.

Fermat hợp tác với Blaise Pascal và công trình của họ về thống kê và lý thuyết xác suất đã đặt nền móng cho những gì chúng ta khái niệm hóa là dữ liệu. Công việc của John Tukey và Edward Tufte trong việc tinh chỉnh các kỹ thuật trực quan hóa dữ liệu đã cho phép lĩnh vực này không chỉ dành cho những người trong lĩnh vực Thống kê. Cùng với sự phát triển của công nghệ này, các hình ảnh trực quan vẽ tay đã phát triển thành các hình ảnh trực quan bằng phần mềm. Ứng dụng trực quan hóa phần mềm có khả năng visualize trong 2D, 3D, thực tế ảo và thực tế tăng cường. Có rất nhiều chương trình để trực quan hóa đã được phát triển: Tableau, Zeppelin, AtScale, R, Python, Superset.

## 2.5 Why use Superset?

Superset là một công cụ có giao diện dễ sử dụng để khám phá và trực quan hóa dữ liệu, tích hợp với nhiều cơ sở dữ liệu khác nhau (trong trường hợp này, chúng ta sẽ sử dụng Druid). Superset cung cấp một bộ tính năng trực quan phong phú và cho phép người dùng tạo các bảng điều khiển để phân loại các bảng dữ liệu của bạn. Superset cũng cho phép người dùng tạo các phân đoạn trực quan, đại diện cho việc trực quan hóa thực tế của bảng dữ liệu, và chọn bảng điều khiển phù hợp cho phân đoạn của họ. Superset đã sẵn sàng cho doanh nghiệp, có khả năng mở rộng và có mô hình bảo mật/quyền hạn chi tiết, cho phép thiết lập các quy tắc phức tạp về việc ai có thể truy cập các tính năng riêng lẻ và tập dữ liệu.

Để dự đoán hỏng xe một cách trực quan và phân tích thói quen lái xe của tài xế trên các bản đồ riêng biệt, nhằm ngăn chặn tài xế gặp tai nạn.

## 2.6 Superset Terminology

1. Nguồn (Sources): Cho phép người dùng xem cơ sở dữ liệu, bảng, nhóm Druid, nguồn dữ liệu Druid.
2. Phân đoạn (Slices): Biểu đồ đồ họa của các nguồn dữ liệu Druid, truyền tải một thông điệp cụ thể về dữ liệu trong bảng cho khách hàng.
3. Bảng điều khiển (Dashboards): Nhóm các phân đoạn trực quan hóa cùng nhau thành các danh mục cụ thể, chẳng hạn như Vận tải IoT, Thời tiết IoT, Nông nghiệp IoT, v.v.

## 2.7 Summary

Bây giờ chúng ta đã quen thuộc với cách Apache Superset liên kết với Ứng dụng IoT xe tải của mình, hãy cùng xem Superset hoạt động như thế nào và tạo một số phân đoạn trực quan hóa dữ liệu của chúng ta.

## 2.8 Further Reading

* Udacity - Data Visualization with Tableau
* Udacity - Data Visualization and D3.js
* Udacity - Intro to Data Science
* Data Visualization - Wikipedia
* Apache Superset

# 3 Setting up the Development Environment

## 3.1 ****Introduction****

Trong hướng dẫn này, bạn sẽ xác minh địa chỉ IP sandbox của bạn được ánh xạ đến tên máy chủ mong muốn, mật khẩu quản trị của bạn được thiết lập và các dịch vụ cần thiết đã được kích hoạt.

## 3.2 Prerequisites

* Học về khái niệm Trực quan hóa Dữ liệu
* Làm quen với Apache Superset
* Học cách Thiết kế Trực quan hóa với Superset

## 3.3 ****Prerequisites****

* Ánh xạ IP của HDP Sandbox thành tên máy chủ:
* Nếu bạn cần trợ giúp, hãy tham khảo phần ENVIRONMENT SETUP trong hướng dẫn Learning the Ropes of the HDP Sandbox (Hướng dẫn tìm hiểu về HDP Sandbox), cụ thể là phần Map Sandbox IP To Your Desired Hostname In The Hosts File (Ánh xạ IP Sandbox thành tên máy chủ mong muốn trong tệp Hosts).
* Ánh xạ IP của HDF Sandbox thành tên máy chủ:
* Nếu bạn cần trợ giúp, hãy tham khảo phần ENVIRONMENT SETUP trong hướng dẫn Learning the Ropes of the HDF Sandbox (Hướng dẫn tìm hiểu về HDF Sandbox), cụ thể là phần Map Sandbox IP To Your Desired Hostname In The Hosts File (Ánh xạ IP Sandbox thành tên máy chủ mong muốn trong tệp Hosts).
* Đặt mật khẩu quản trị Ambari cho HDP:
* Nếu bạn cần trợ giúp, hãy tham khảo phần Admin Password Reset (Đặt lại mật khẩu quản trị) trong hướng dẫn Learning the Ropes of the HDP Sandbox (Hướng dẫn tìm hiểu về HDP Sandbox).
* Đặt mật khẩu quản trị Ambari cho HDF:
* Nếu bạn cần trợ giúp, hãy tham khảo phần Learning the Ropes of the HDF Sandbox (Hướng dẫn tìm hiểu về HDF Sandbox) trong hướng dẫn Learning the Ropes of the HDF Sandbox (Hướng dẫn tìm hiểu về HDF Sandbox).
* Dữ liệu cần thiết trong Druid:
* Để có dữ liệu trong Druid, hãy tham khảo hướng dẫn Real-Time Event Processing In NiFi, SAM, Schema Registry and SuperSet (Xử lý sự kiện thời gian thực trong NiFi, SAM, Schema Registry và SuperSet) để thiết lập đường ống dữ liệu SAM lưu trữ dữ liệu vào Druid. Bạn chỉ cần thực hiện các bước từ 1 đến 3.

## 3.4 ****Outline****

* Started up all required services for HDF and HDP
* Setup Druid for Superset
* Summary
* Further Reading

## 3.5 Started up all required services for HDF and HDP

Nếu bạn không chắc chắn, hãy đăng nhập vào Bảng điều khiển quản trị Ambari:

Đối với HDF:

* Truy cập http://sandbox-hdf.hortonworks.com:8080
* Kiểm tra xem Zookeeper, Storm, Kafka, NiFi, Schema Registry, Streaming Analytics Manager đã khởi động hay chưa. Nếu chưa, hãy khởi động chúng với chế độ Bảo trì (Maintenance Mode) bị tắt.

Đối với HDP:

* Truy cập http://sandbox-hdp.hortonworks.com:8080
* Kiểm tra xem HDFS, YARN, Druid và Superset đã khởi động hay chưa. Nếu chưa, hãy khởi động chúng với chế độ Bảo trì (Maintenance Mode) bị tắt.

Ví dụ:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động

Sau khi khởi động Druid và Superset, Tác vụ nền (Background Operations) của bạn sẽ trông tương tự như sau:

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Phông chữ

Mô tả được tạo tự động

## Setup Druid for Superset

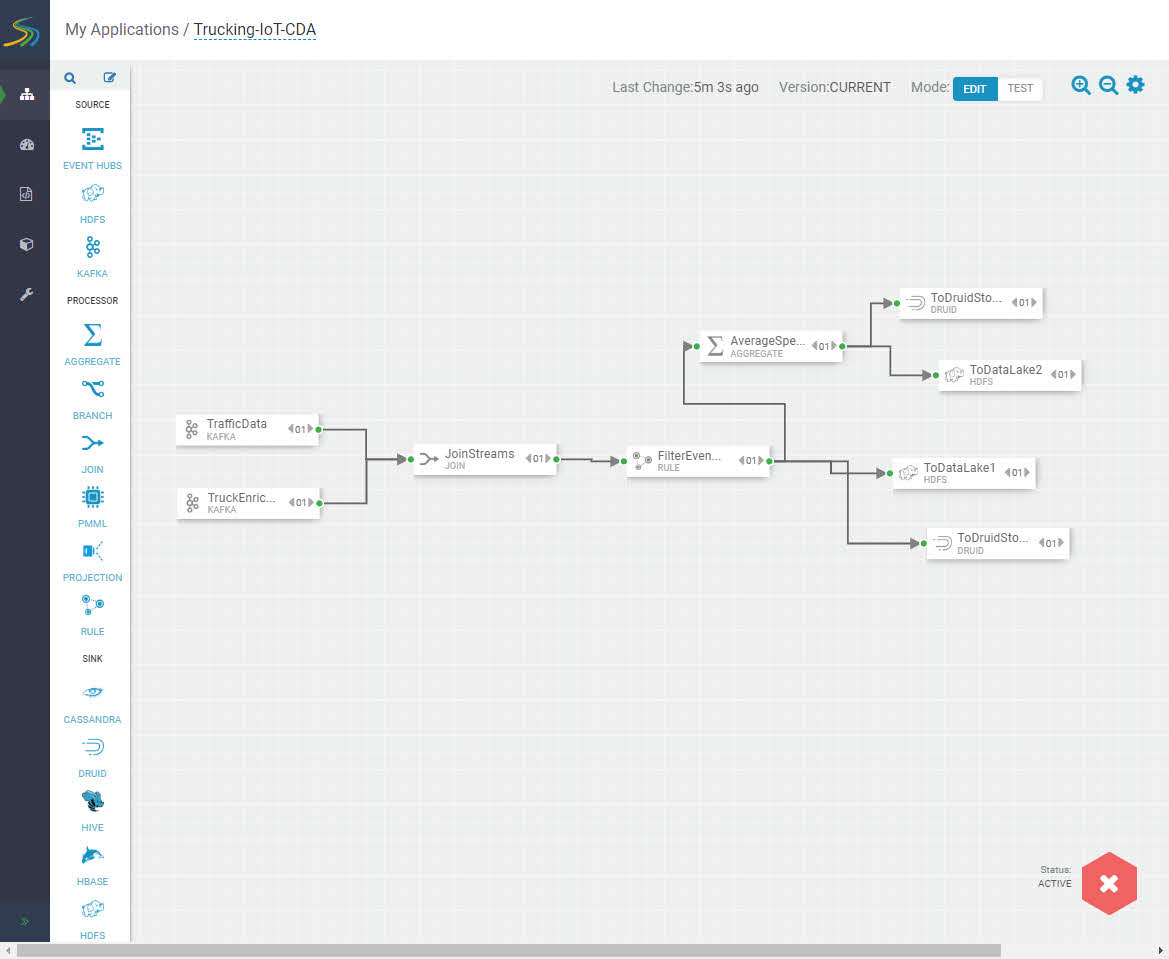
Chúng tôi bao gồm hình ảnh tham khảo về những gì cần được bắt đầu trong đường ống dữ liệu để đưa dữ liệu vào Druid. Bạn đã hoàn thành bước này, được nêu trong các yêu cầu tiên quyết.

1. Trong canvas NiFi [http://sandbox-hdf.hortonworks.com:9090/nifi], hãy bắt đầu NiFi DataFlow bằng cách nhấn nút bắt đầu màu xanh lá cây trong bảng điều khiển vận hành (operate panel).

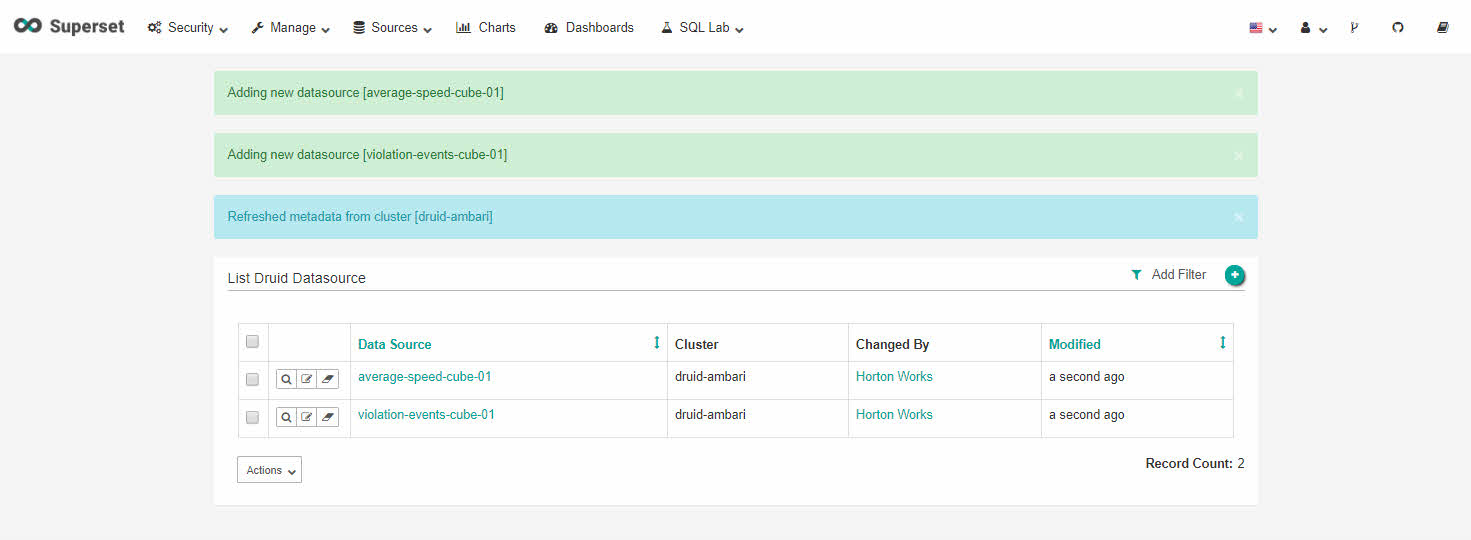
Ảnh có chứa văn bản, biên lai, biểu đồ, Song song

Mô tả được tạo tự động

1. Trên bảng SAM canvas http://sandbox-hdf.hortonworks.com:7777/, hãy bắt đầu topology SAM bằng cách nhấn vào nút bắt đầu màu xanh lá cây ở góc dưới bên phải của bảng.



1. Trên giao diện người dùng của Superset http://sandbox-hdp.hortonworks.com:9089, đăng nhập bằng thông tin đăng nhập admin/admin, đợi khoảng 5 - 10 phút để dữ liệu Kafka được tiêu thụ, sau đó định kỳ, chọn menu Sources và nhấp vào Refresh Druid Metadata. Cuối cùng, hai nguồn dữ liệu Druid sẽ xuất hiện.



## 3.7 Summary

Chúng ta có thể thấy các nguồn dữ liệu trong Superset. Giờ đây, chúng ta đã sẵn sàng bắt đầu tạo các hình ảnh trực quan về dữ liệu.

## Further Reading

* An introduction to Druid
* How Superset and Druid Power Real-Time Analytics at Airbnb | DataEngConf SF '17

# 4 Visualizing Trucking Data

## 4.1 ****Introduction****

Chúng ta hiểu được vai trò của việc trực quan hóa dữ liệu trong các ứng dụng khoa học dữ liệu. Hãy sử dụng Superset để tạo một số biểu đồ trực quan từ dữ liệu của bộ dữ liệu Trucking IoT.

## 4.2 Prerequisites

* Thiết lập môi trường phát triển

## 4.3 ****Outline****

* Create Visualization Slices
* Summary
* Further Reading

## 4.4 ****Create Visualization Slices****

Superset được sử dụng để trực quan hóa các nguồn dữ liệu từ Druid. Từ các đặc điểm của một hiển thị trực quan hiệu quả từ phần khái niệm, chúng ta sẽ áp dụng chúng vào việc trực quan hóa hiệu quả các bộ dữ liệu average-speed-cube-01 và violation-events-cube-01.

**Biểu đồ Sunburst từ Bảng Cubes Tốc độ Trung bình**

Nhấp vào nguồn dữ liệu average-speed-cube-01. Cửa sổ phân đoạn trực quan hóa sẽ xuất hiện:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Biểu tượng máy tính

Mô tả được tạo tự động

Từ trực quan hóa trên, chúng ta nhận được dữ liệu về tổng số hàng cho nguồn dữ liệu của chúng ta. Hãy thay đổi cấu hình truy vấn để tạo một trực quan hóa sunburst mô tả rõ ràng dữ liệu:

Nguồn dữ liệu và loại biểu đồ:

* Nguồn dữ liệu: average-speed-cube-01
* Loại biểu đồ: Table View (Bảng) là lựa chọn mặc định, nhưng chúng ta sẽ thay đổi nó thành Sunburst (Bánh xe mặt trời)

Truy vấn:

* Phân cấp: thêm driverId (mã tài xế), speed\_AVG (tốc độ trung bình)
* Chỉ số chính: COUNT(\*) (đếm số lượng)
* Chỉ số phụ (tùy chọn): không chọn (mặc định)
* Giới hạn hàng: 50000

Ảnh có chứa văn bản, ảnh chụp màn hình, Nhiều màu sắc, biểu đồ

Mô tả được tạo tự động

Từ trực quan hóa trên, khi chúng ta di chuột qua các phần của biểu đồ, chúng ta sẽ thấy driverId và tốc độ trung bình của họ. Ví dụ, driverId là 3 và tốc độ trung bình của họ là 8. Phần trăm tốc độ trung bình mà họ đóng góp vào tốc độ trung bình tổng thể là 12,7%.

**Save the Slice**

* Nhấp vào nút lưu.
* Đặt tên cho phân đoạn là AvgSpeedSunburst.
* Thêm phân đoạn vào bảng điều khiển mới Trucking-IoT-Demo.
* Sau đó nhấp vào Lưu.

Ảnh có chứa văn bản, ảnh chụp màn hình, Nhiều màu sắc, vòng tròn

Mô tả được tạo tự động

**Chuyển sang nguồn dữ liệu "Violation Events Cube"**

* Click vào Datasource bên dưới Datasource & Chart Type.
* Một cửa sổ sẽ xuất hiện. Chọn violation-events-cube-01.
* Bây giờ, biểu đồ sẽ đại diện cho dữ liệu từ nguồn "Violation Events Cube" thay vì "Average Speed Cube".

Ảnh có chứa văn bản, ảnh chụp màn hình, Nhiều màu sắc, biểu đồ

Mô tả được tạo tự động

* Lưu ý rằng truy vấn của "Violation Events" không có trường speed\_AVG vì nó không có trong nguồn dữ liệu này.
* Để cải thiện biểu đồ, hãy thay đổi tên thành "DriverViolationSunburst" để khách hàng biết đây là một nguồn dữ liệu khác.
* Trong phần Query:
* Bên Hierarchy:
* Xóa driverId.
* Thêm driverName và eventType.
* Bên Secondary Metric:
* Thêm COUNT(\*).
* Nhấp vào Run Query.
* Lớp đầu tiên của biểu đồ Sunburst sẽ hiển thị cho chúng ta DriverName và tỷ lệ phần trăm của họ đối với các sự kiện vi phạm.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, số

Mô tả được tạo tự động

Tầng thứ hai hiển thị phần trăm loại vi phạm mà tài xế có khả năng đóng góp trên đường.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, số

Mô tả được tạo tự độngẢnh có chứa văn bản, ảnh chụp màn hình, số, biểu đồ

Mô tả được tạo tự động

Lưu phân đoạn  
Nhấp vào nút Lưu để lưu phân đoạn Sunburst về vi phạm sự kiện. Lưu với tên DriverViolationSunburst. Thêm vào bảng điều khiển hiện có:

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Nhiều màu sắc

Mô tả được tạo tự động

## 4.5 Summary

Xin chúc mừng! Bạn vừa mới học cách tạo các phân đoạn trực quan hóa bằng cách sử dụng Superset. Bạn đã bắt đầu với việc đếm số hàng trong nguồn dữ liệu tốc độ trung bình, nhưng sau đó đã thay đổi trực quan hóa sang một biểu đồ khác, Sunburst, để hiển thị cách nhìn khác về dữ liệu giúp khách hàng hiểu và diễn giải nó tốt hơn. Bạn đã học cách lưu phân đoạn trực quan hóa cho nguồn dữ liệu đó. Vì nhiệm vụ khác của bạn là tạo một phân đoạn trực quan cho nguồn dữ liệu vi phạm sự kiện, bạn đã học cách chuyển đổi nguồn dữ liệu trong quá trình đang ở trong phân đoạn trực quan. Sau đó, bạn đã điều chỉnh cấu hình truy vấn để cho phép khách hàng hiểu được trực quan hóa Sunburst của nguồn dữ liệu vi phạm sự kiện. Bạn cũng đã học cách thêm các phân đoạn vào một bảng điều khiển chung. Hãy tự do thử nghiệm với các khả năng trực quan hóa khác của Superset, ví dụ như thử tạo các biểu đồ từ phần khái niệm.

## 4.5 Further Reading

* Apache Superset