**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**VIỆN CNTT & TT**

A close up of a sign

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN**

**Đề tài**

**PHÂN TÍCH DATA GOOGLE TREND VỀ ĐẠI DỊCH COVID-19**

**Nhóm sinh viên thực hiện:**

| Đỗ Hà Thủy | MSSV: 20175246 |
| --- | --- |
| Bùi Minh Hiếu | MSSV: 20173114 |
| Trần Hữu Hiếu | MSSV: 20180078 |
| Nguyễn Như Hoàng | MSSV: 20164850 |

***Hà Nội, 06/2021***

# Mục lục

# 

[**Mục lục**](#_ja8xea8k5av3) **1**

[**Chương 1: Giới thiệu đề tài**](#_7krj6tfuffeu) **2**

[**Chương 2: Triển khai cài đặt**](#_h2ig1e8mhv4h) **2**

[Khái quát mô hình hệ thống](#_vkldt3x1f3hz) 2

[Dữ liệu đầu vào](#_i38neucqy0ia) 3

[Worker](#_9kax656jgwsv) 3

[MongoDB](#_t1ko5zvxrsaf) 3

[Apache Spark, Jupyter Notebook](#_romk6tkt6moj) 4

[Elastic Cloud, Kibana](#_35o54yxkqhvk) 5

[**Chương 3: Kết quả cài đặt**](#_wbe5sejslsa) **6**

[Trực quan hóa dữ liệu](#_ybbtivtmgupa) 6

[Visualizing “data trend” by Kibana](#_ckbdpkjmrkhr) 6

[Visualizing “Real covid data” by kibana](#_lcilxceolw0a) 9

[Spark Machine Learning](#_74srbcweqnf1) 11

[Mô tả bài toán](#_etyszehcdrkv) 11

[Các bước tiến hành:](#_1722irxgntof) 12

[**Tài liệu tham khảo**](#_j87apzkvgrbb) **16**

# Chương 1: Giới thiệu đề tài

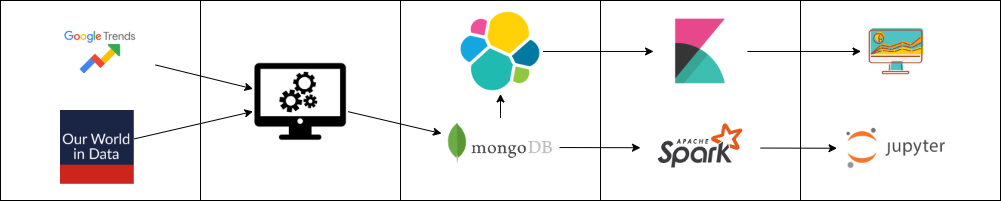
Kể từ khi bùng phát coronavirus vào tháng 12 năm 2019 ở Vũ Hán, Trung Quốc, nó đang lây lan theo cấp số nhân và đã ảnh hưởng đến gần như mọi quận trên thế giới, lây nhiễm cho hàng triệu người và gây ra hơn hàng chục nghìn ca tử vong trên khắp thế giới (tính đến ngày 16 tháng 3, 2020). Nó đã gây ra thiệt hại kinh tế và xã hội vô cùng thảm khốc trên khắp thế giới. Theo ước tính của Cục Dự trữ Liên bang, Coronavirus có thể mất việc làm tổng cộng 47 triệu người, tỷ lệ thất nghiệp có thể lên tới 32%. Để dự đoán số lượng bệnh nhân bị nhiễm là rất quan trọng đối với sự chuẩn bị của cả cá nhân và người ra quyết định, và để làm phẳng các đường cong. Tuy nhiên, làm thế nào để dự đoán chính xác số lượng bệnh nhân mắc bệnh không bao giờ là một việc đơn giản. Có rất nhiều yếu tố góp phần vào sự lan truyền của vi rút này, chẳng hạn như khả năng di chuyển của quần thể, nhiệt độ và tình trạng y tế.

Ngày nay, ngày càng có nhiều người truy cập internet và tìm kiếm những thông tin liên quan mật thiết đến cuộc sống, cảm xúc và suy nghĩ hàng ngày của họ. Người ta ước tính rằng có khoảng 63.000 lượt tìm kiếm trên Google mỗi giây. Một người bình thường thực hiện khoảng ba hoặc bốn lần tìm kiếm mỗi ngày. Google Trend là một trang web được Google tài trợ phân tích mức độ phổ biến của các truy vấn tìm kiếm hàng đầu trong Google Search trên nhiều khu vực và ngôn ngữ khác nhau. Trang web đáng tin cậy để tìm kiếm thông tin mà người dùng quan tâm. Do đó, các xu hướng của Google đang bộc lộ và có thể tạo cơ hội để xem xét mối quan tâm của mọi người cũng như các chủ đề nóng mà họ quan tâm. Các nhà nghiên cứu đã sử dụng dữ liệu của Google Trend để điều tra một số nghiên cứu, trong đó có dự đoán sự bùng phát dịch bệnh.

Trong bài báo này, ta khám phá dữ liệu của Google Trend để tìm ra mối quan hệ của nó với sự lan truyền COVID-19. Thay vì tập trung vào dự đoán dựa trên mô hình, chúng tôi đề xuất sử dụng dữ liệu của Google Trend và kết hợp với chuỗi thời gian lịch sử để dự đoán các ca mắc trong tương lai. Cách tiếp cận của chúng tôi là theo hướng dữ liệu thuần túy và bỏ qua mô hình toán học phức tạp, điều này làm giảm đáng kể độ phức tạp của thuật toán. Chúng tôi đã thực hiện các thử nghiệm toàn diện và áp dụng nhiều mô hình dự đoán phổ biến trên dữ liệu toàn cầu để xem mối tương quan giữa xu hướng tìm kiếm và các trường hợp bị nhiễm bệnh. Các thử nghiệm của chúng tôi đã chứng minh rằng có mối quan hệ chặt chẽ giữa các trường hợp bệnh nhân bị nhiễm bệnh và dữ liệu xu hướng của Google, và có thể được sử dụng với các kỹ thuật phân tích khác để hiểu rõ hơn về sự lây lan của căn bệnh này.

# Chương 2: Triển khai cài đặt

## Khái quát mô hình hệ thống



## Dữ liệu đầu vào

Hai nguồn dữ liệu chính được sử dụng để giải quyết bài toán đã đặt ra bao gồm:

* Google Trends: Google Trends là một dịch vụ web công cộng của Google. Nó giúp cung cấp thống kê về kết quả tìm kiếm Google Search trên toàn cầu kể từ năm 2004.
* Our World in Data (<https://ourworldindata.org/>): là một ấn phẩm trực tuyến khoa học tập trung vào các vấn đề toàn cầu lớn như nghèo đói, bệnh tật, đói kém, biến đổi khí hậu, chiến tranh, rủi ro hiện hữu và bất bình đẳng.Đây là một dự án của Phòng thí nghiệm Dữ liệu Thay đổi Toàn cầu, một tổ chức từ thiện đã đăng ký ở Anh và xứ Wales, và được thành lập bởi Max Roser, một nhà sử học xã hội và nhà kinh tế phát triển. Nhóm nghiên cứu có trụ sở tại Đại học Oxford.

## Worker

Nhiệm vụ của Worker là đọc dữ liệu từ các nguồn dữ liệu nêu trên và insert vào database MongoDB. Đối với dữ liệu Covid thực tế, ta có thể đọc trực tiếp từ các file dữ liệu được cung cấp và insert lần lượt vào MongoDB. Tuy nhiên đối với Google Trends API, vấn đề nằm ở giới hạn về số lượng request trên mỗi IP của Google, giải pháp của nhóm là sử dụng các dịch vụ VPS.

Source code: <https://github.com/Mhieu14/covid_trend_data_collector>

## MongoDB

MongoDB được triển khai trên dịch vụ MongoDB Atlas. Dữ liệu trong mỗi Cluster ở Atlas được lưu trữ theo cơ chế Replication, với 3 nodes: 1 master (primary) và 2 slaves (secondary). Tổ chức lưu trữ trong MongoDB bao gồm các collection như sau:

Dữ liệu về quốc gia (country):

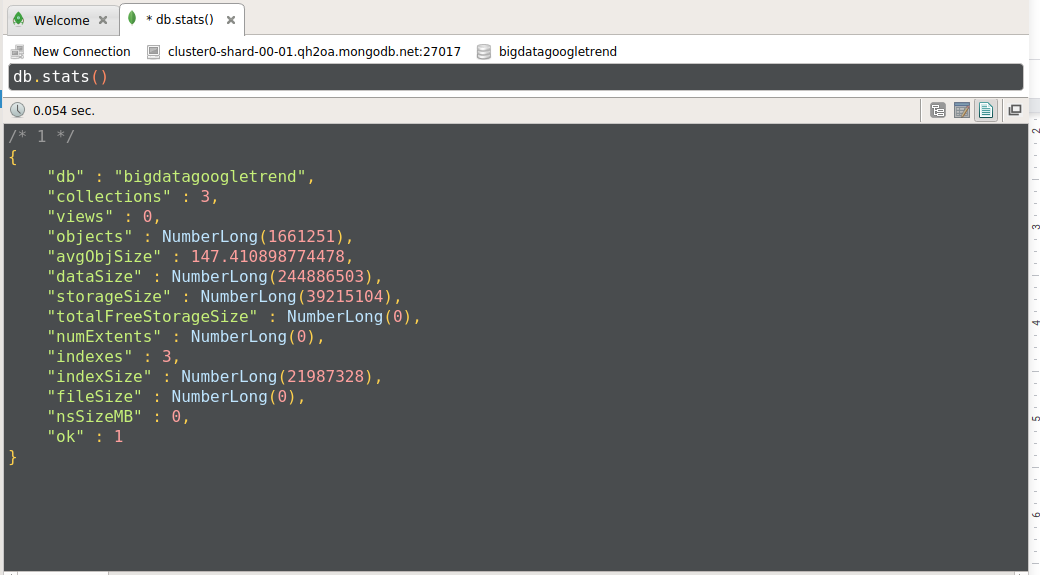
| const country = Schema({  Code\_2: String,  Code\_3: String,  Code\_num: String,  Crawled: { type: Number, default: 0 },  }, { timestamps: false, collection: 'country' }); |
| --- |

Dữ liệu thống kê của Google Trend (data\_trend):

| const dataTrend = Schema({  country\_code\_2: String,  country\_code\_3: String,  date\_statistic: Date,  key\_word: String,  value: { type: Number, default: 0 },  }, { timestamps: false, collection: 'data\_trend' }); |
| --- |

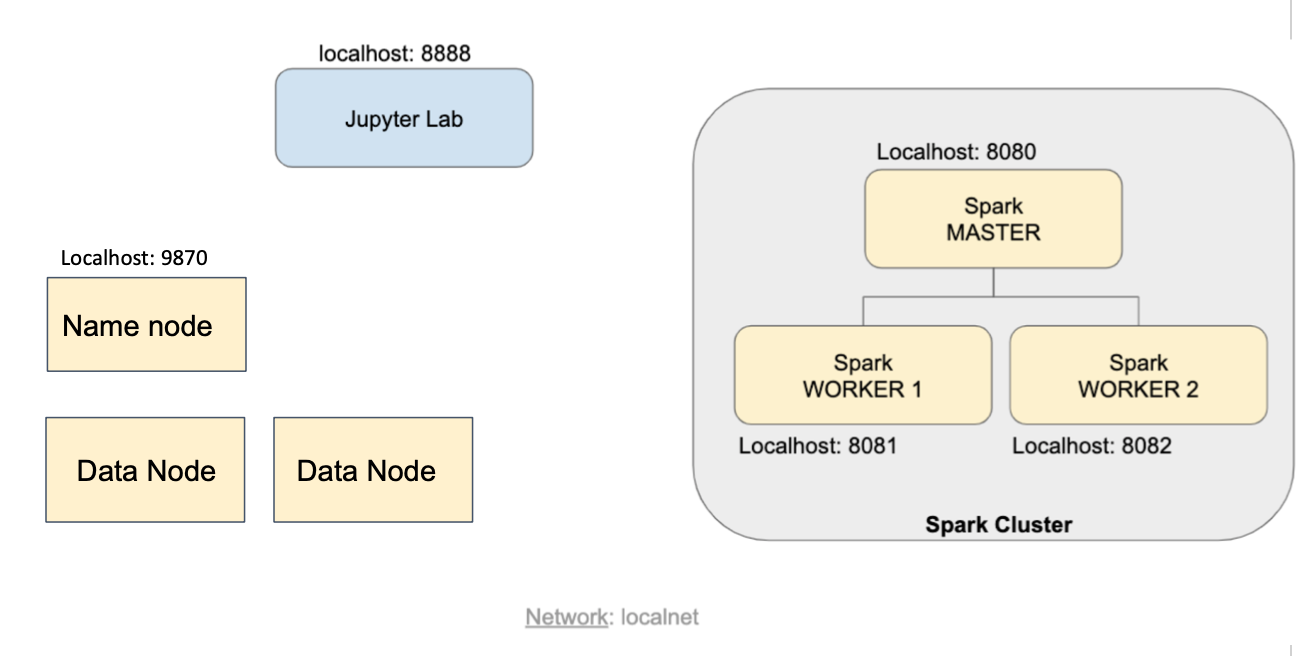
Dữ liệu về các ca nhiễm Covid thực tế (data\_covid):

| const dataCovid = Schema({  country\_code\_2: String,  country\_code\_3: String,  date\_statistic: Date,  total\_cases: Number,  new\_cases: Number,  total\_cases\_per\_million: Number,  new\_cases\_per\_million: Number,  stringency\_index: Number,  }, { timestamps: false, collection: 'data\_covid' }); |
| --- |

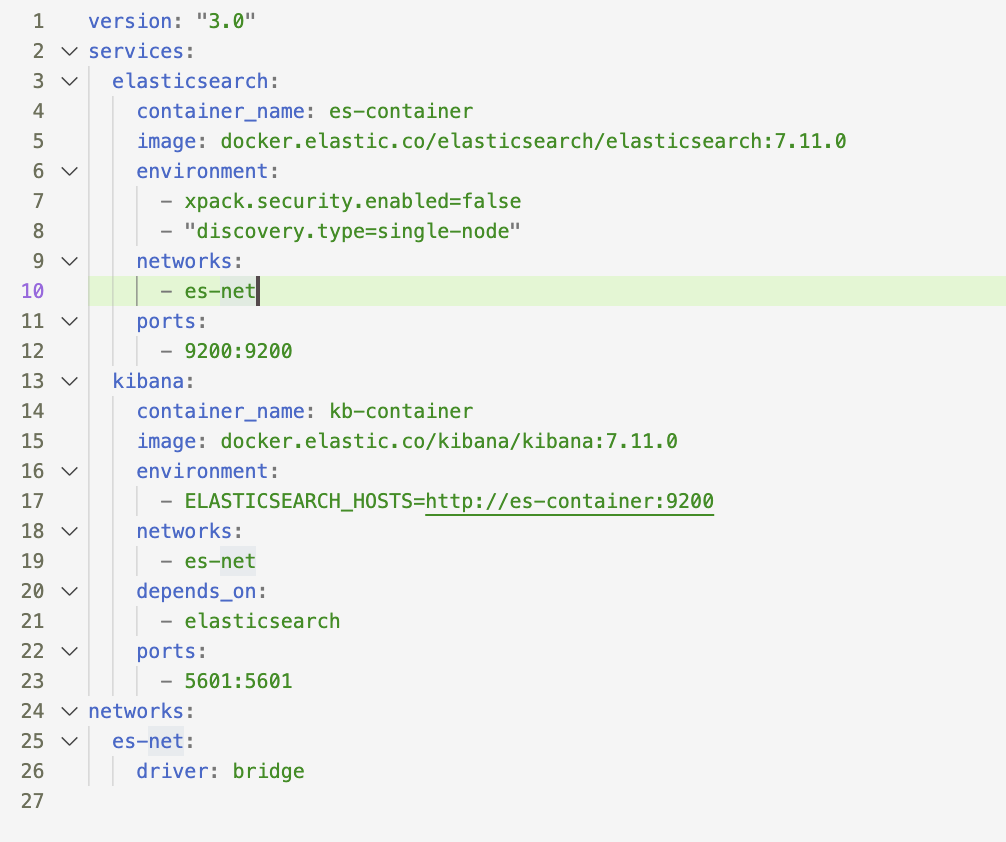


## Apache Spark, Jupyter Notebook vs Hadoop

Cụm Spark cài đặt trên docker bao gồm 4 container là JupyterLab, một master, hai worker, 1 name node, 2 data node



## Elastic Cloud, Kibana

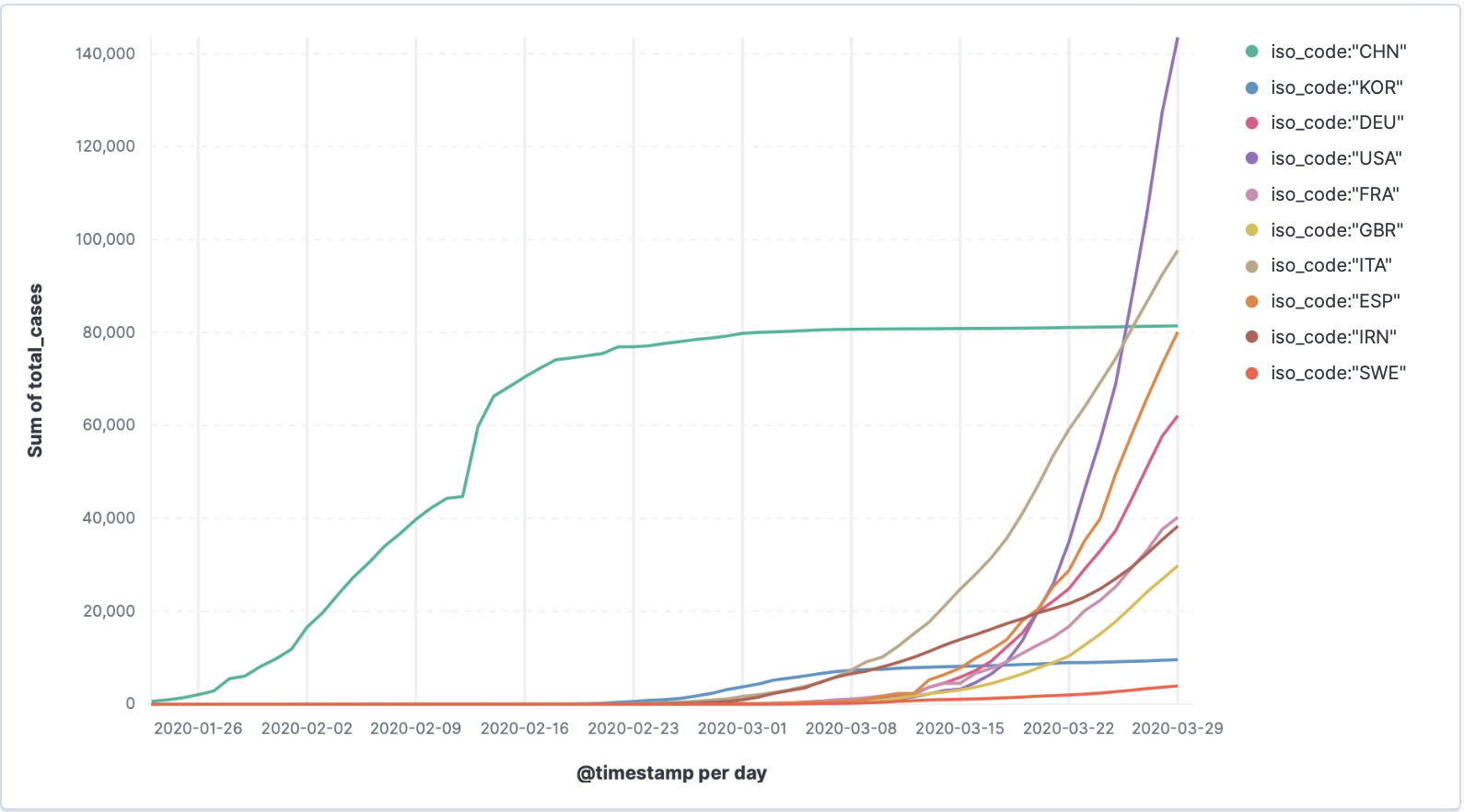


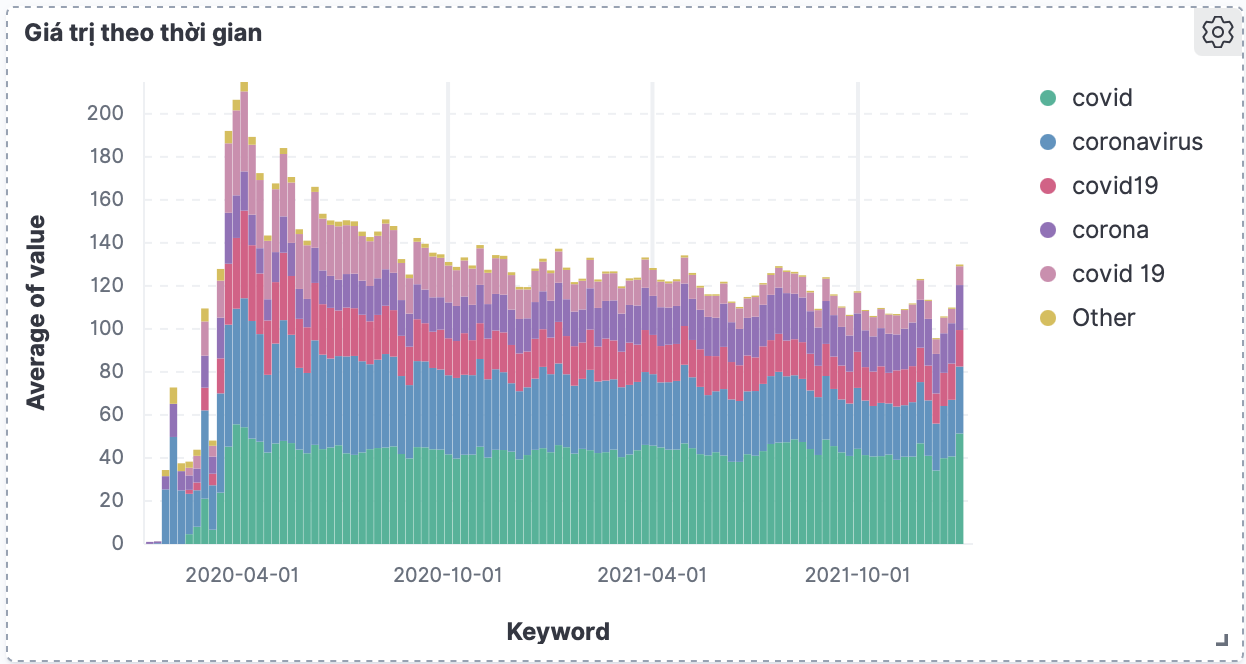
# Chương 3: Kết quả cài đặt

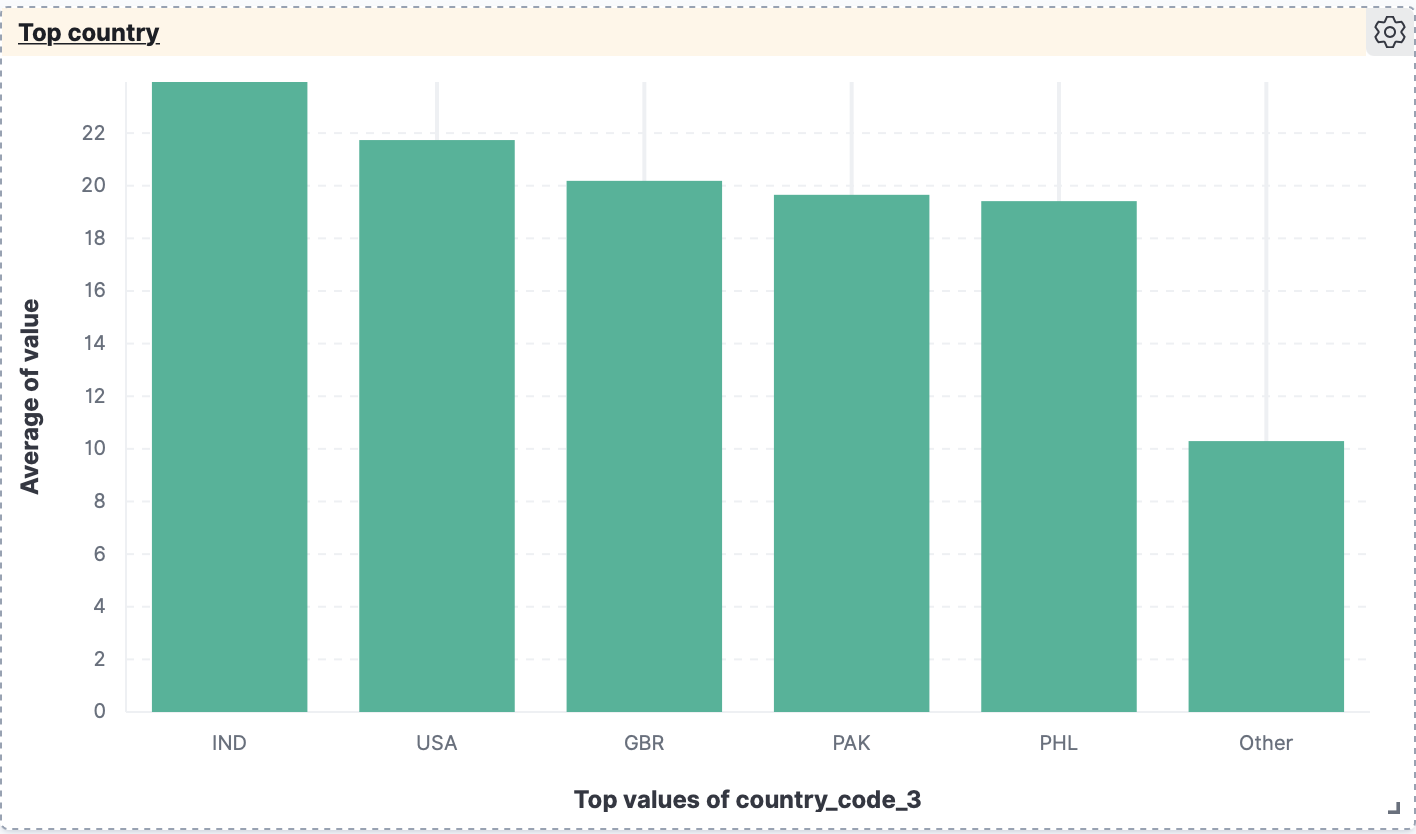
## Trực quan hóa dữ liệu

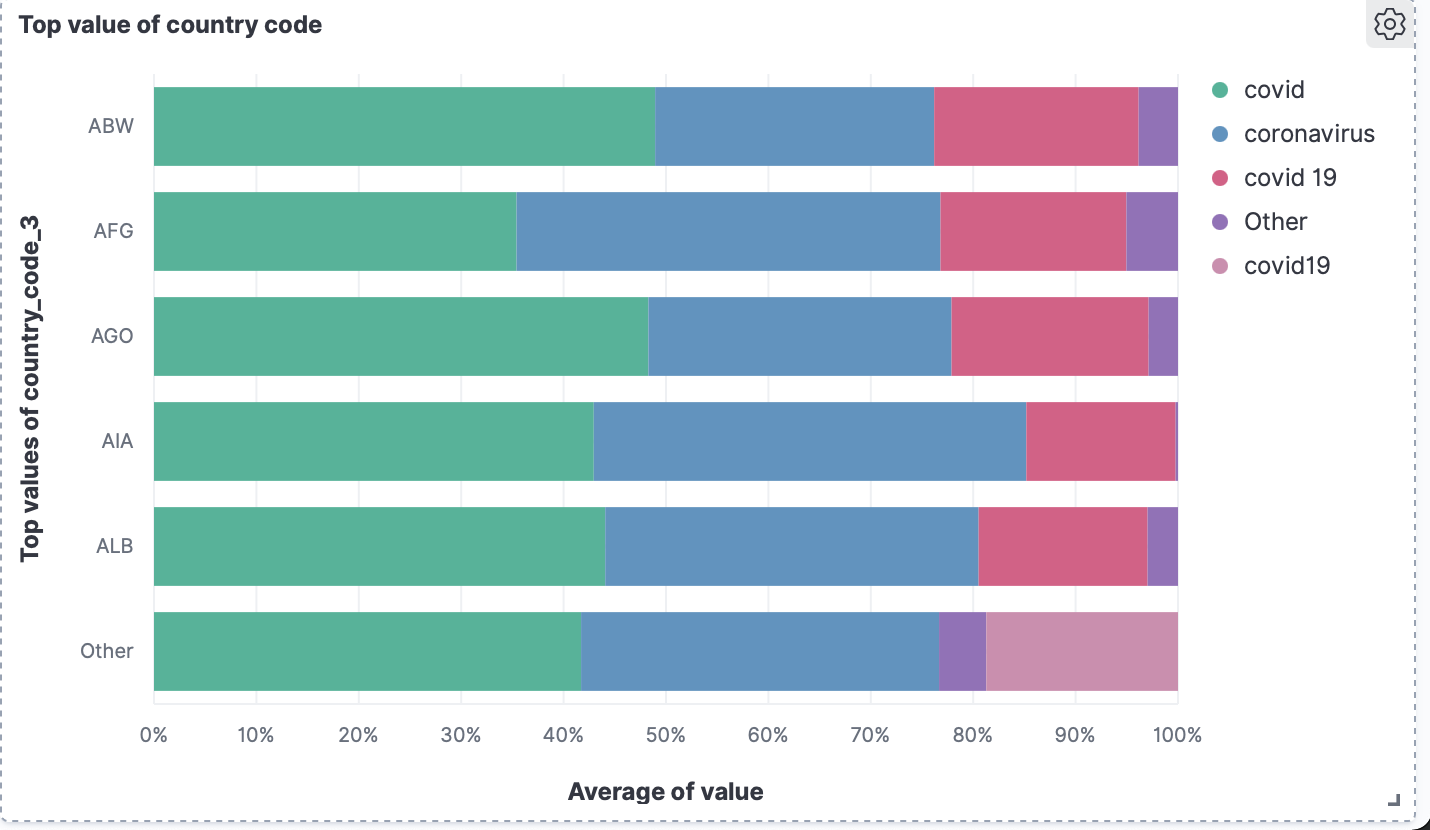
### Visualizing “**data trend”** by Kibana

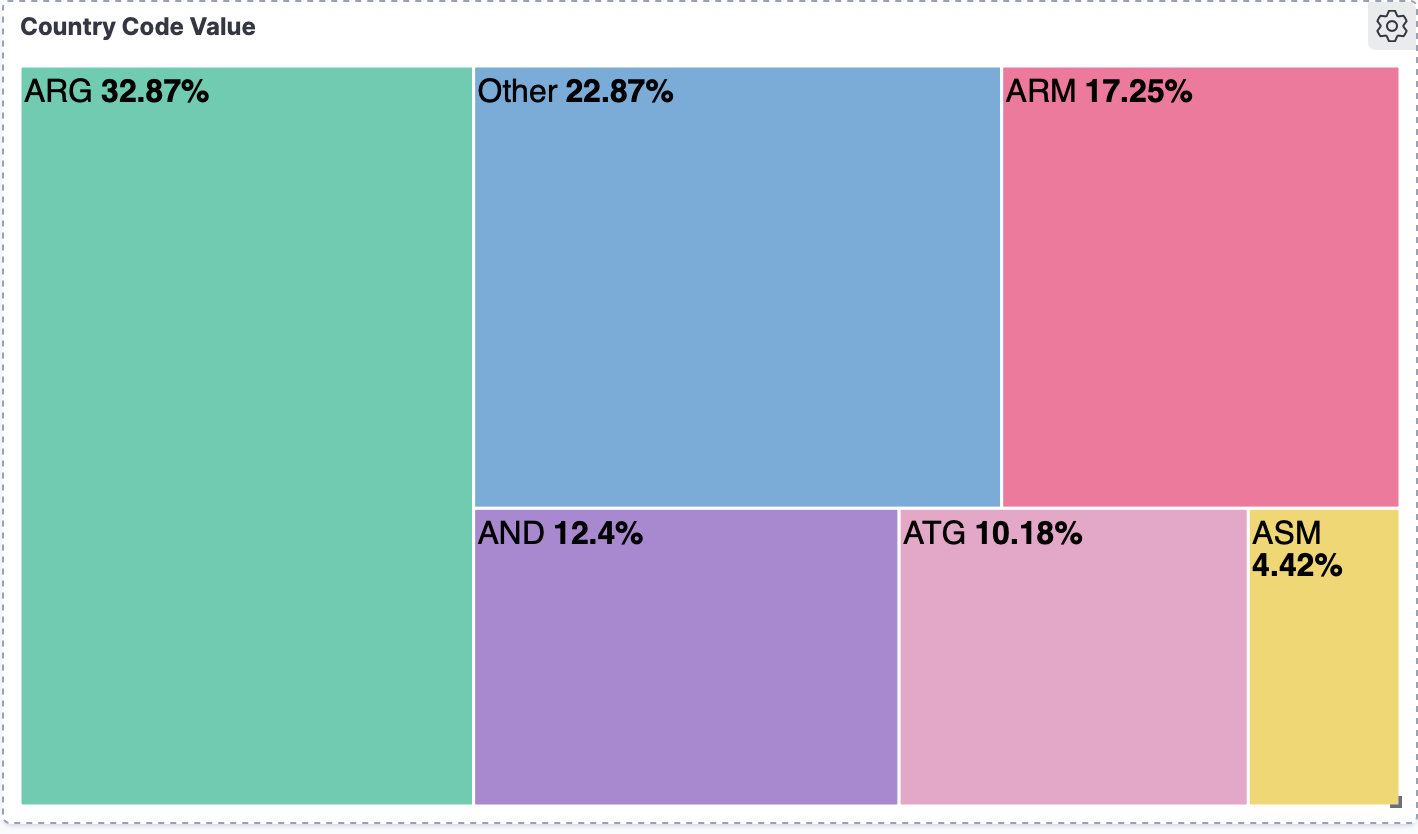


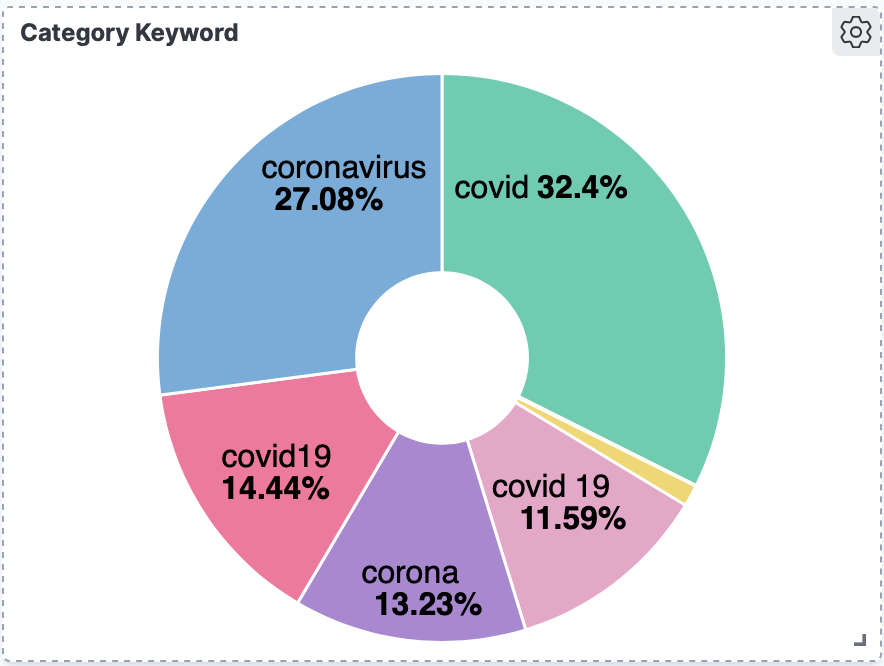


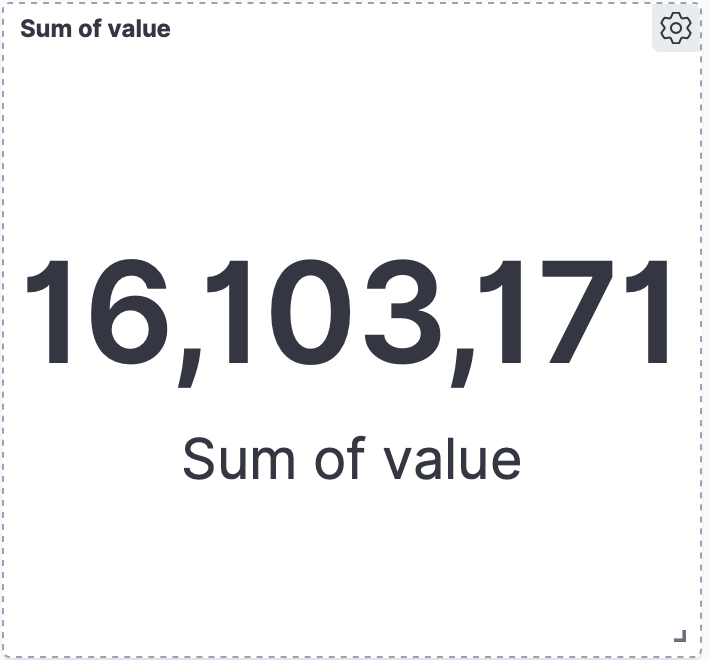


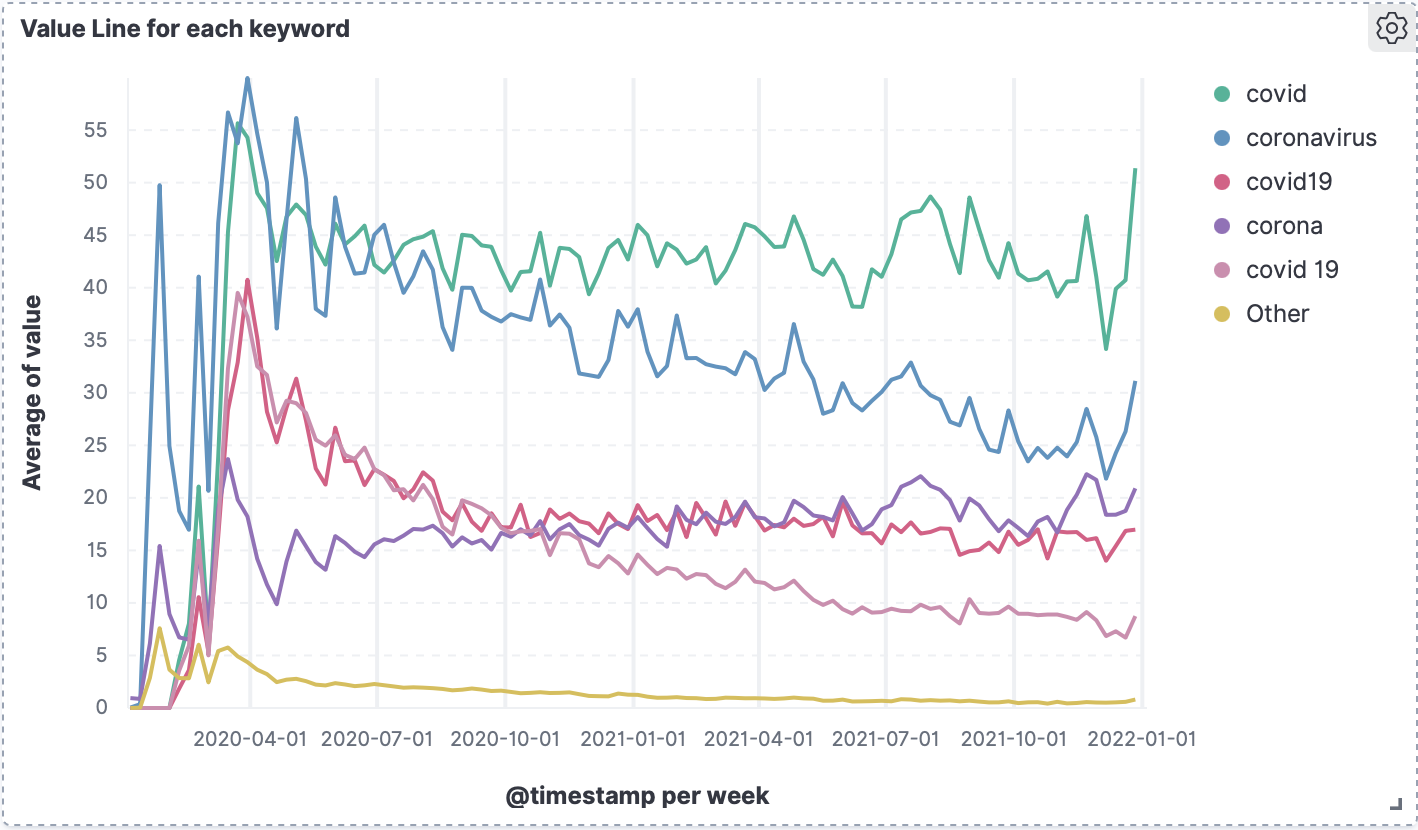


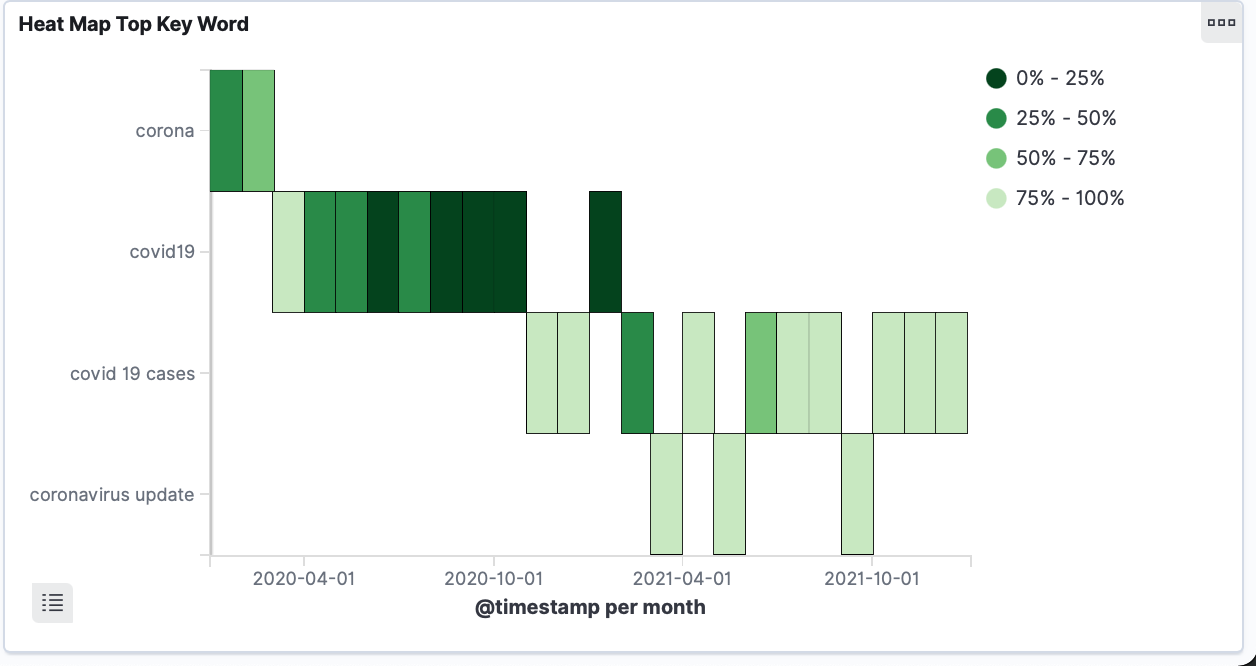




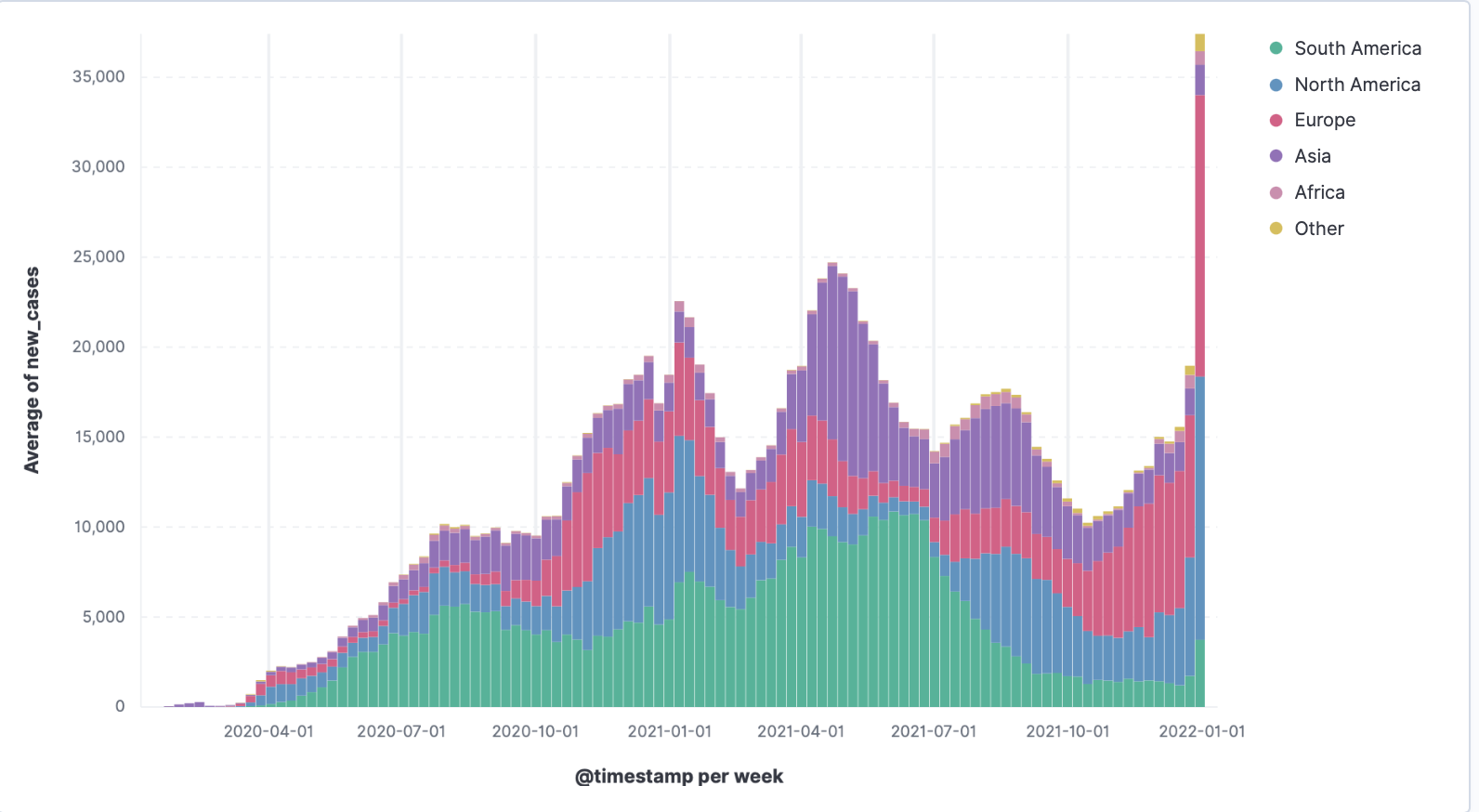


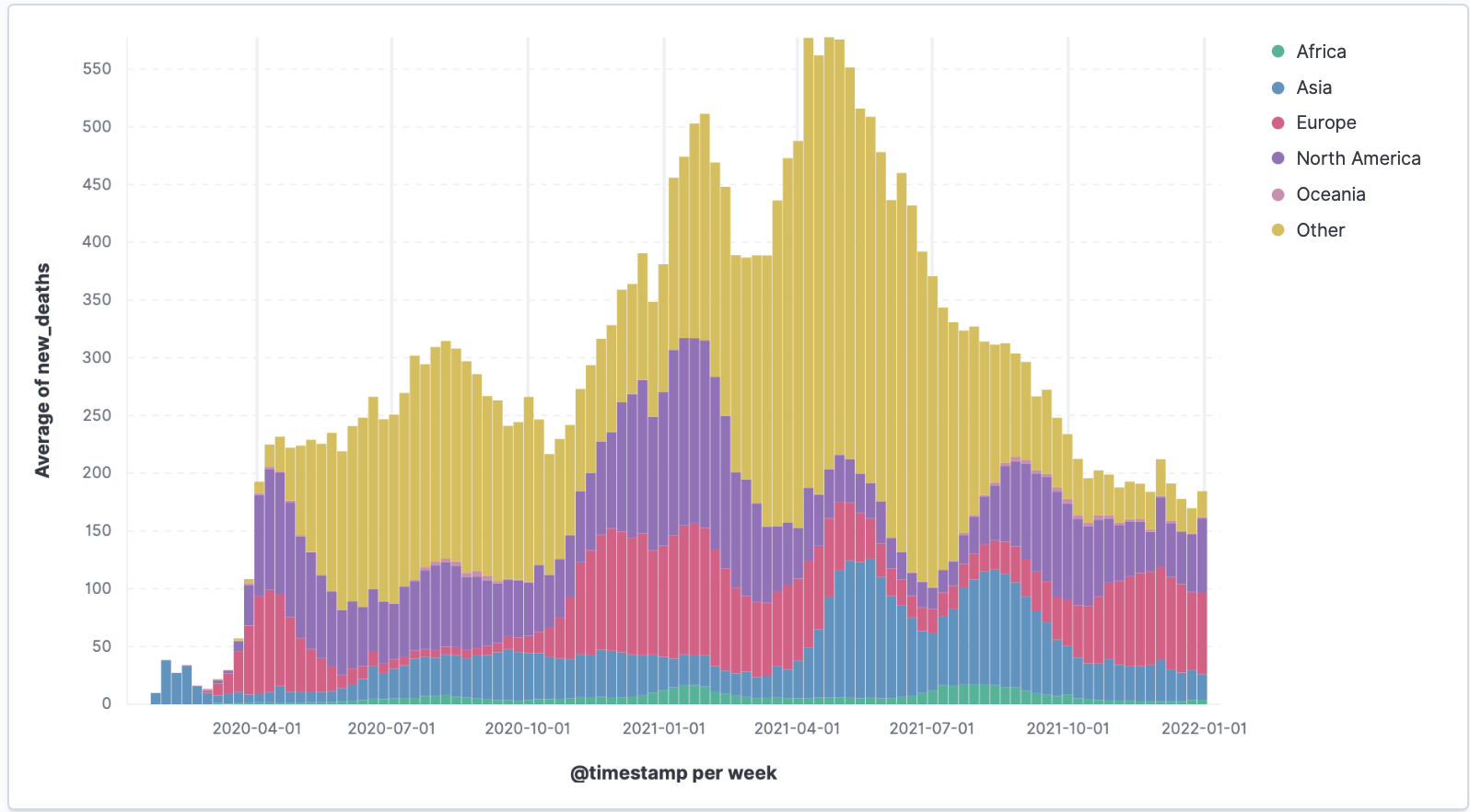


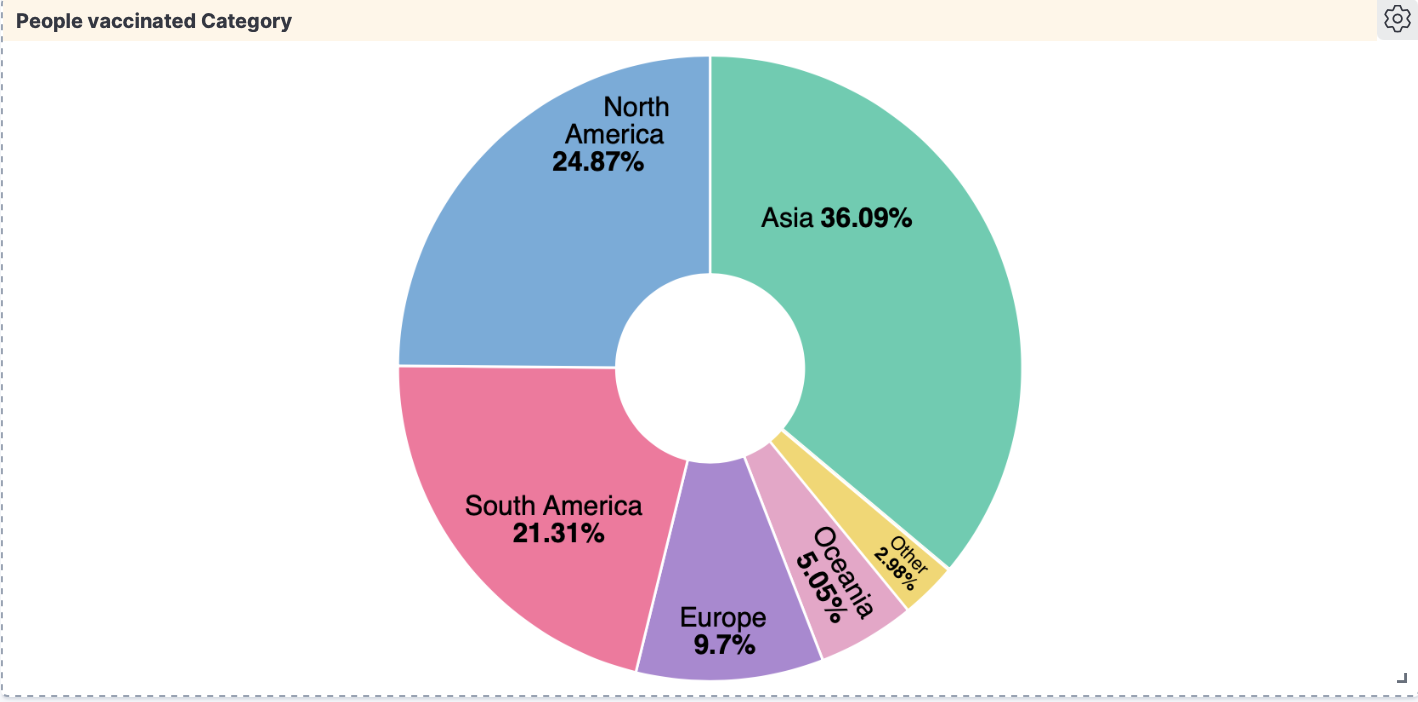


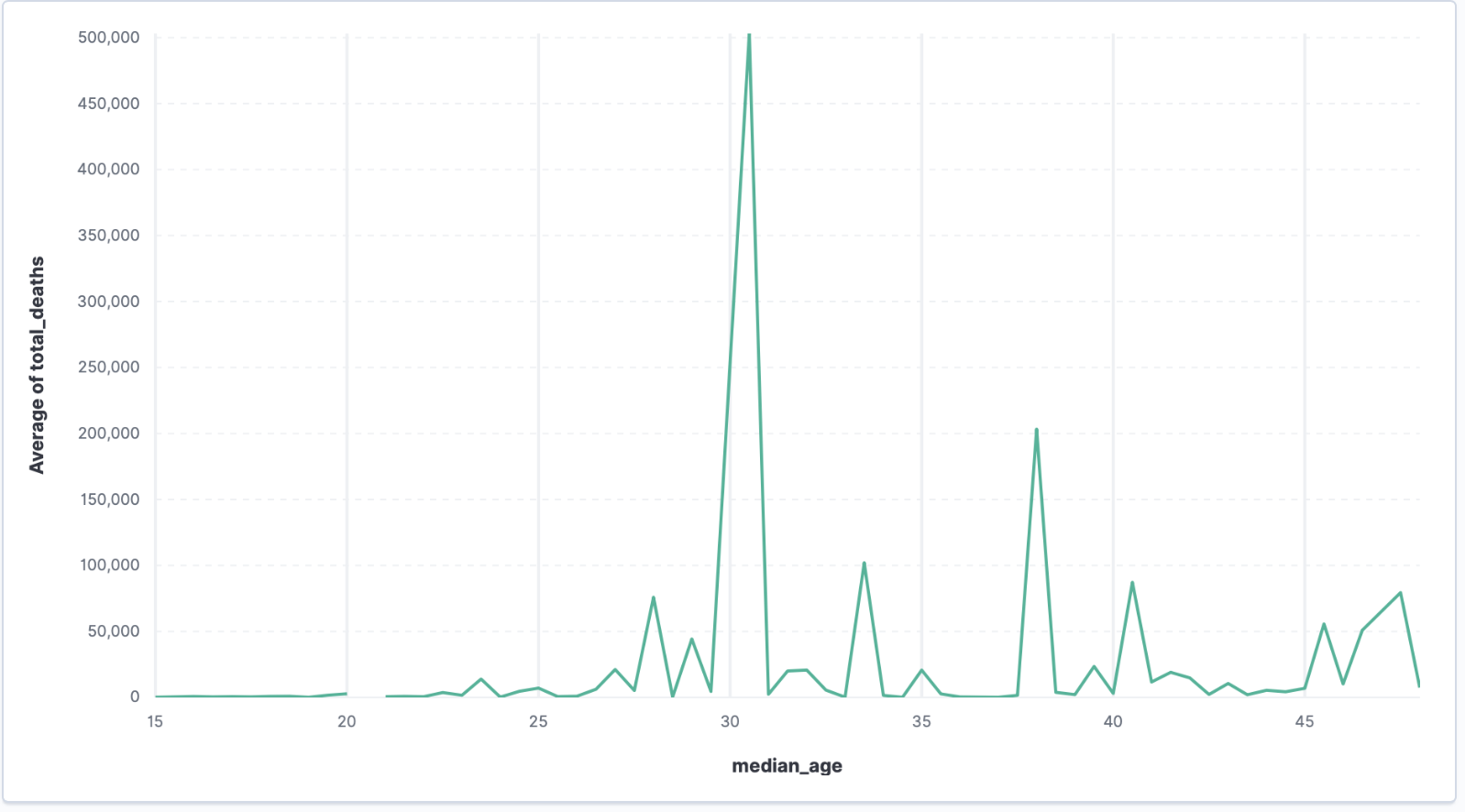


### Visualizing “**Real covid data”** by kibana

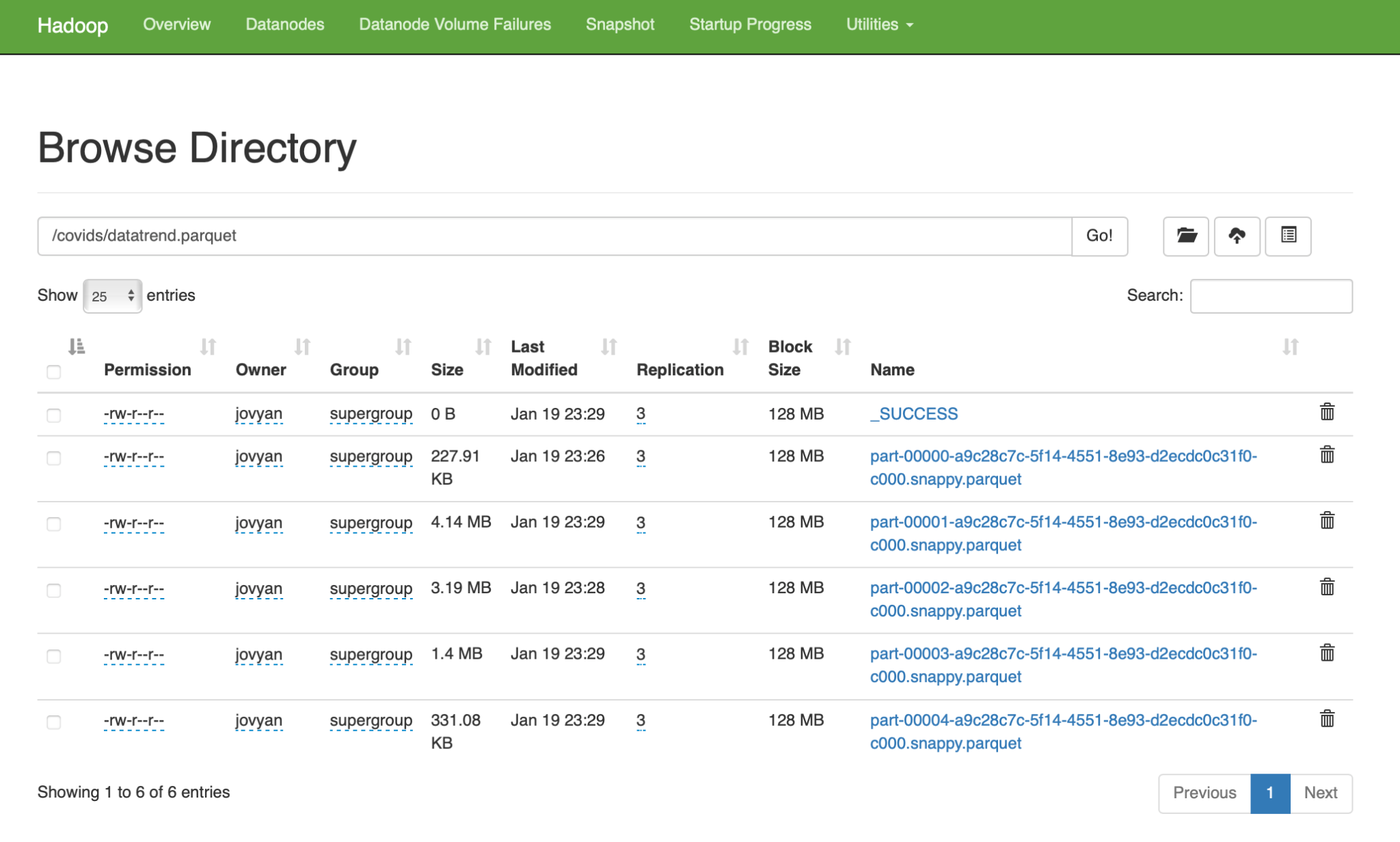








Cuối cùng đưa lưu trữ dữ liệu vào Hadoop.



## Spark Machine Learning

### Mô tả bài toán

Trong phần này, ta sẽ xây dựng một mô hình dùng đề dự đoán số ca mắc Covid-19 của các nước dựa trên các tham số đầu vào.

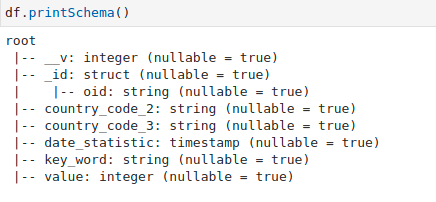
Các tham số đầu vào bao gồm:

* Keyword: “corona”, “covid19”,...
* Value
* Số ca mắc

Đầu ra sẽ là số ca mắc covid-19 của tháng sau.

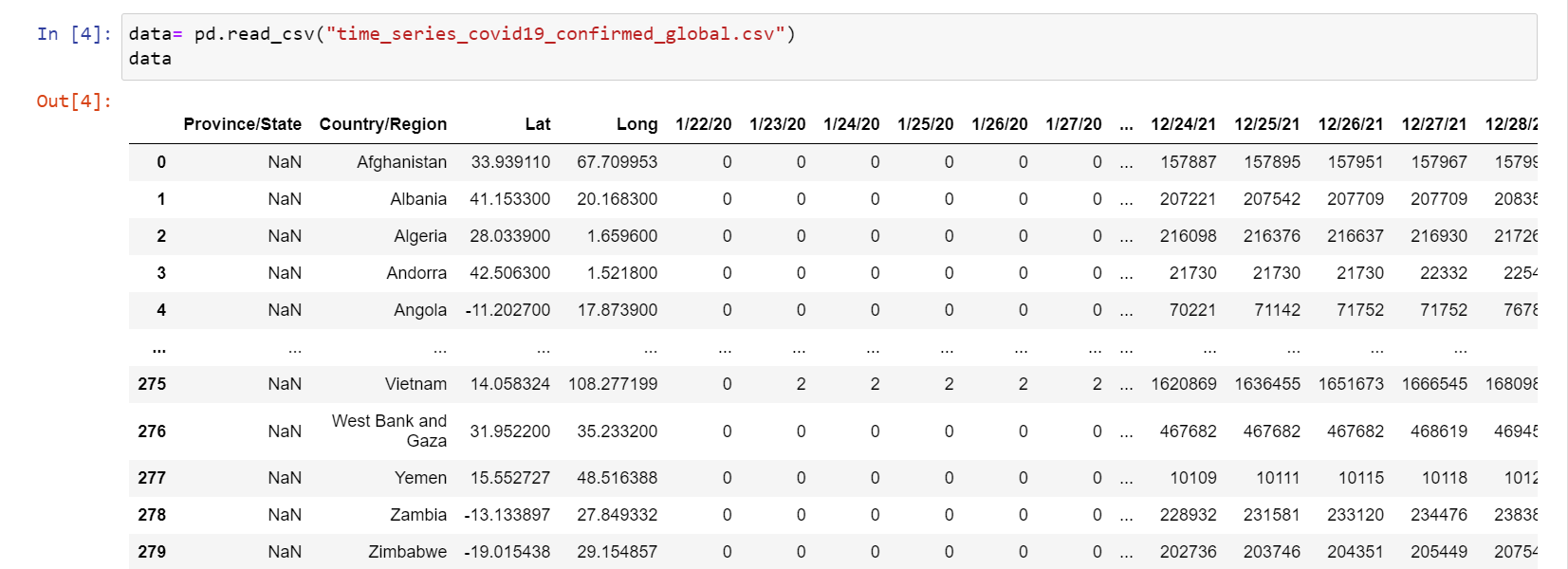
### Các bước tiến hành:

Đầu tiên tiến hành load dữ liệu:



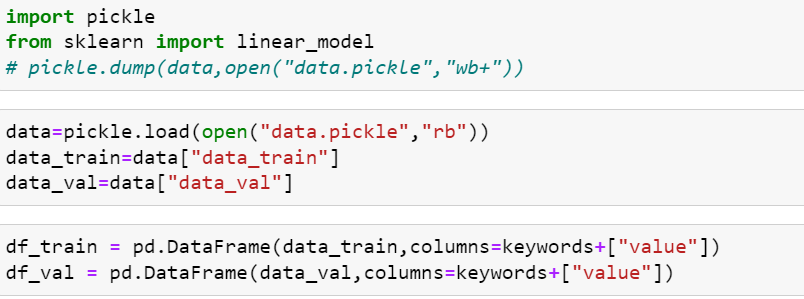
Tiếp theo, ta tiến hành xử lý dữ liệu:

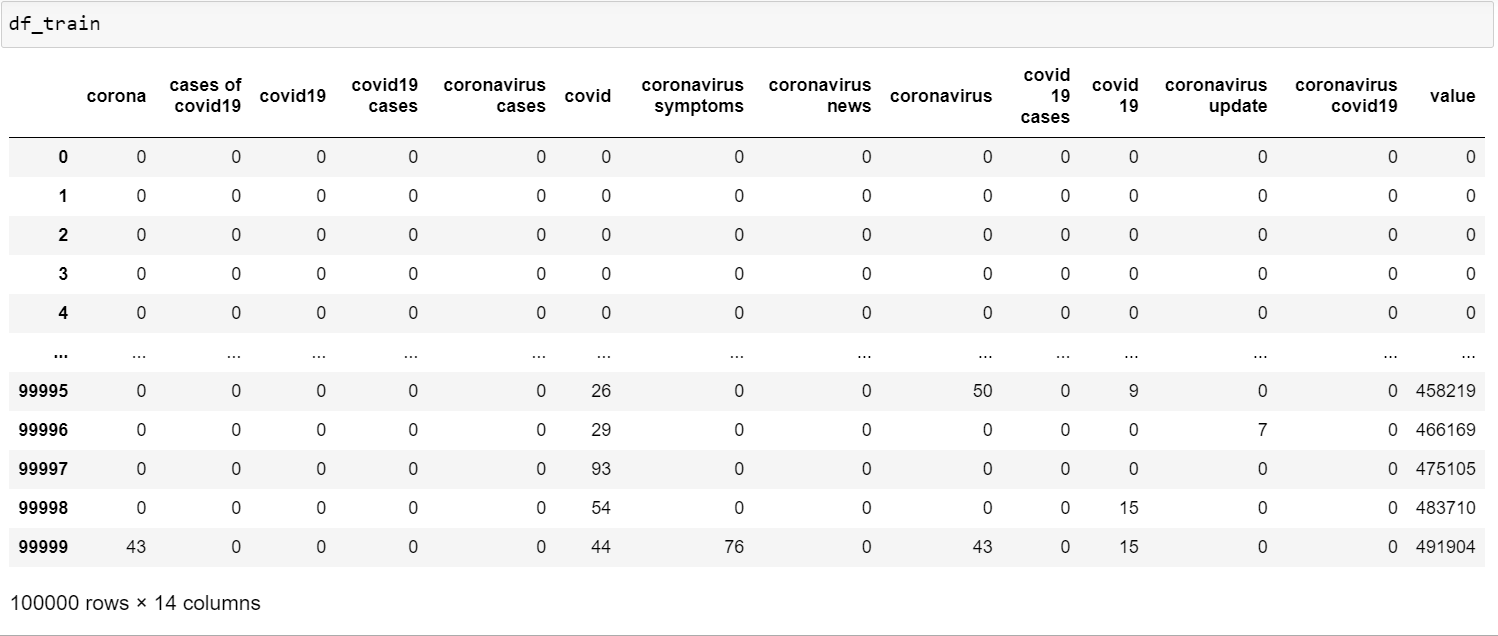


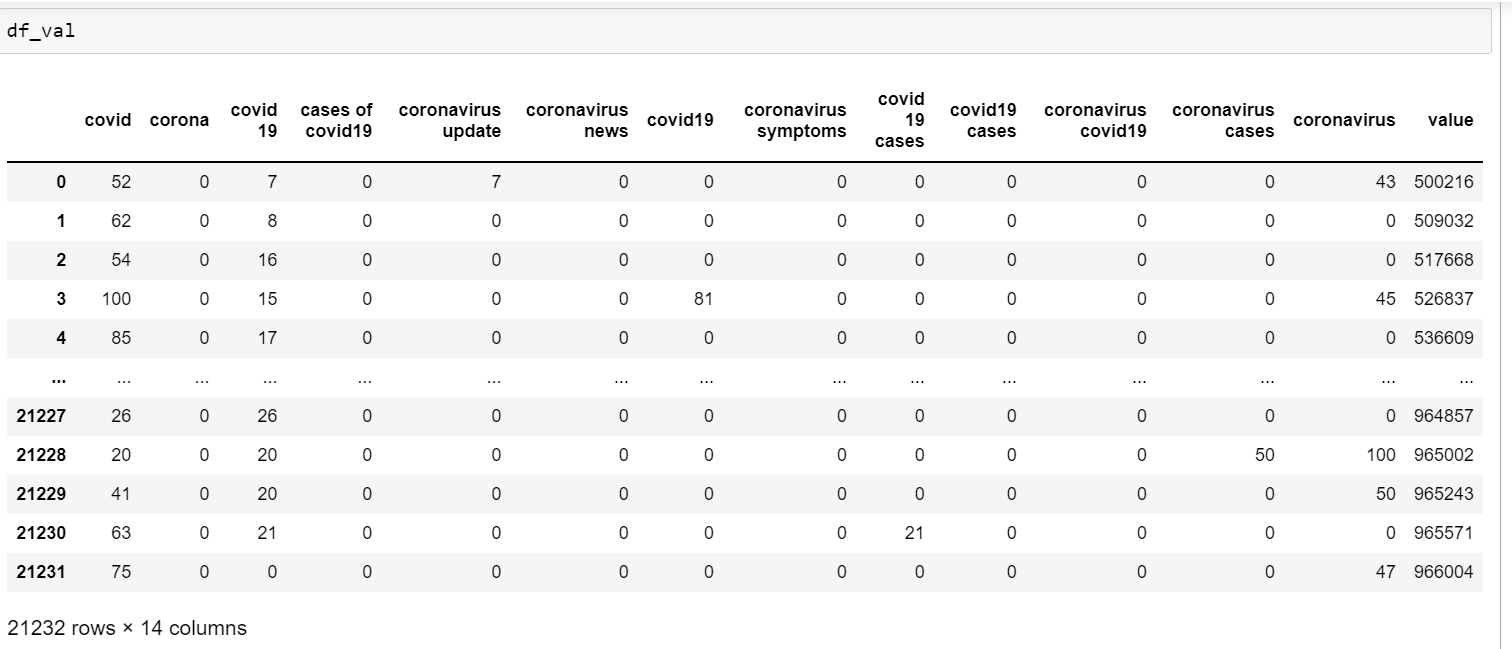




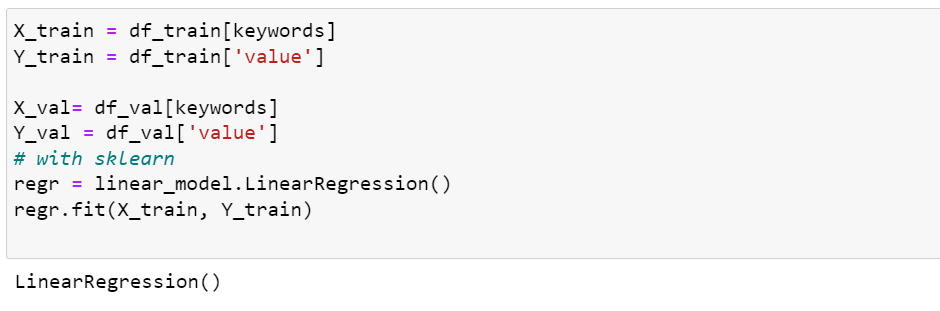
Chia dữ liệu thành 2 tập trainingData và testVal:



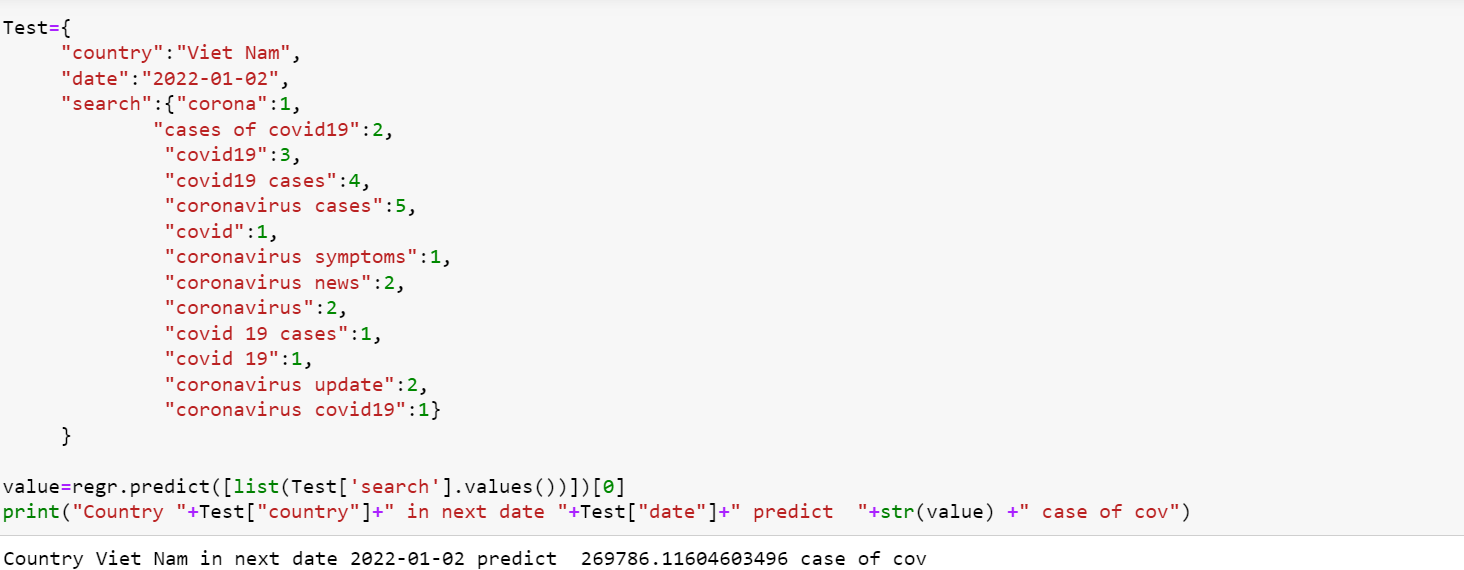




Training:



Dự đoán số ca mắc của 1 nước, vd Việt Nam:



# Tài liệu tham khảo

[1]. <https://ieeexplore.ieee.org/abstract/document/9377852>

[2]. <https://www.mongodb.com/blog/post/getting-started-with-mongodb-pyspark-and-jupyter-notebook>