BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG
ĐẠI HỌC TÀI NGUYÊN VÀ MÔI TRƯỜNG TP.HCM
KHOA HỆ THỐNG THÔNG TIN VÀ VIỄN THÁM

ĐỒ ÁN MÔN HỌC
HỆ QUẢN TRỊ CƠ SỞ DỮ LIỆU

ĐỀ TÀI 10: BIG DATA & NOSQL

GVHD: THS.PHẠM TRỌNG HUYNH

NHÓM 3 THỰC HIỆN:
1. TRƯƠNG THỊ THUỲ LINH
2. PHẠM QUỲNH GIANG
3. LÂM THỊ NGỌC MINH

TP.HCM, 11/2022

MỤC LỤC

LỜI MỞ ĐẦU

Một thách thức công nghệ lớn mà thế giới đang phải đối mặt là quản lý và lưu trữ dữ liệu, trong đó hàng triệu dữ liệu trong thời gian gần đây đang được tạo ra với thời gian gián đoạn dưới nano giây. Do đó, việc xử lý một lượng lớn dữ liệu là một thách thức đáng kể và do đó với sự tăng trưởng dân số, cần có công nghệ thu thập và quản lý dữ liệu hiện đại hơn.

Với nhu cầu xử lý và tạo dữ liệu nhanh chóng trong thời gian gần đây đã dẫn đến là hơn 2,6 nghìn tỷ dữ liệu đang được sản xuất hàng ngày. Họ dự đoán thêm rằng trong tương lai sẽ có sự gia tăng theo cấp số nhân hơn về việc sử dụng và tạo dữ liệu nữa trong tương lai, đặc biệt hơn cách nó được sử dụng ngày nay còn thú vị hơn.

I. GIỚI THIỆU BIG DATA VÀ SQL:

Dữ liệu thường không có cấu trúc, nó có thể được tạo ra từ nhiều nguồn khác nhau, chẳng hạn như đăng bài trên phương tiện truyền thông xã hội, nội dung đa phương tiện với kho lưu trữ tự động. Email, truy vấn công cụ tìm kiếm, kho tài liệu quản lý nội dung, dữ liệu cảm biến các loại khác nhau, sàn giao dịch chứng khoán, hình ảnh vệ tinh, hệ thống giám sát và ứng dụng e-health.etc.

Nó là một kiến trúc lưu trữ phân tán rộng rãi, có cấu trúc hệ thống quản lý cơ sở dữ liệu hữu ích cơ bản. Dữ liệu thực tế của các giá trị chính được lưu trữ theo cặp, cột hoặc họ cột, tài liệu và đồ thị.

Bigdata giúp lưu trữ và quản lý lượng thông tin vô hạn được tạo ra mỗi giây, mỗi ngày.

NoSQL hoạt động để giúp giải quyết các yêu cầu về khối lượng, sự đa dạng và vận tốc dữ liệu lớn (Andreas & Michael, 2019) giải thích rằng Dữ liệu lớn vẫn chưa có định nghĩa ràng buộc. Tuy nhiên, nhiều chuyên gia dữ liệu sẽ đồng ý về ba chữ V (3V): V olume cho khối lượng dữ liệu mở rộng, Variety nhiều định dạng, dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc, và Velocity để xử lý dữ liệu tốc độ cao và thời gian thực.

Hình 1: Dữ liệu lớn 3V(Nguồn: https://marketingai.vn/big-data-la-gi/)

1. **Volume**: khối lượng
2. **Vrlocity** : tốc độ
3. **Variety**: Tính đa dạng

1. **Đặc trưng 5V của dữ liệu lớn là gì?**

Đó chính là **Volume** *(khối lượng)*, **Velocity** *(vận tốc)*, **Variety** *(đa dạng)*, **Veracity** *(tính xác thực)*, **Value** *(giá trị)*

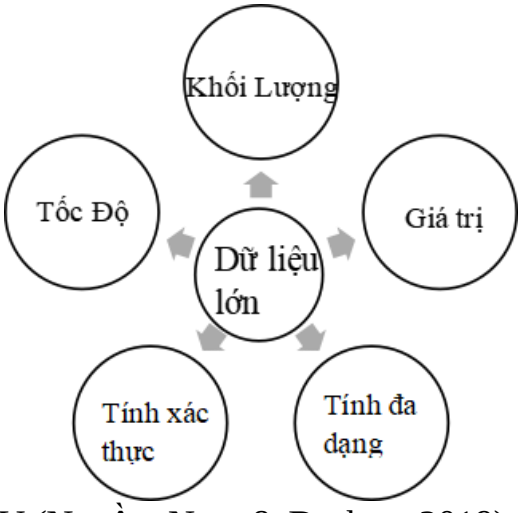
Volume: hiển thị lượng dữ liệu khổng lồ, chẳng hạn như dữ liệu cho thiết bị di động, được sử dụng cho các chức năng khác nhau.

Velocity: chỉ định tốc độ hoặc tần suất tạo, cập nhật, xử lý và truy cập dữ liệu.

Variety: dữ liệu được truy cập thông qua nhiều loại thiết bị khác nhau, chẳng hạn như video, ảnh, v.v.

Veracity: giải thích cách rút ra kiến thức hữu ích từ các tập dữ liệu khổng lồ. Khía cạnh quan trọng nhất của bất kỳ công cụ dữ liệu lớn nào là giá trị, vì nó cho phép tạo ra kiến thức có giá trị

Value: Đề cập đến độ chính xác và giá trị thông tin rất lớn



Hình 2: Dữ liệu lớn 5 V (Nguồn: Nzar & Dashne, 2019)

1. **Khối lượng**: Hiển thị lượng dữ liệu khổng lồ, chẳng hạn như dữ liệu cho thiết bị di động, được sử dụng cho các chức năng khác nhau.
2. **Vận tốc:** chỉ định tốc độ hoặc tần suất tạo, cập nhật, xử lý và truy cập dữ liệu.
3. **Đa dạng**: dữ liệu được truy cập thông qua nhiều loại thiết bị khác nhau, chẳng hạn như video, ảnh, v.v.
4. **Giá trị:** giải thích cách rút ra kiến thức hữu ích từ các tập dữ liệu khổng lồ. Khía cạnh quan trọng nhất của bất kỳ công cụ dữ liệu lớn nào là giá trị, vì nó cho phép tạo ra kiến thức có giá trị.
5. **Tính xác thực**: Đề cập đến độ chính xác và giá trị thông tin rất lớn.

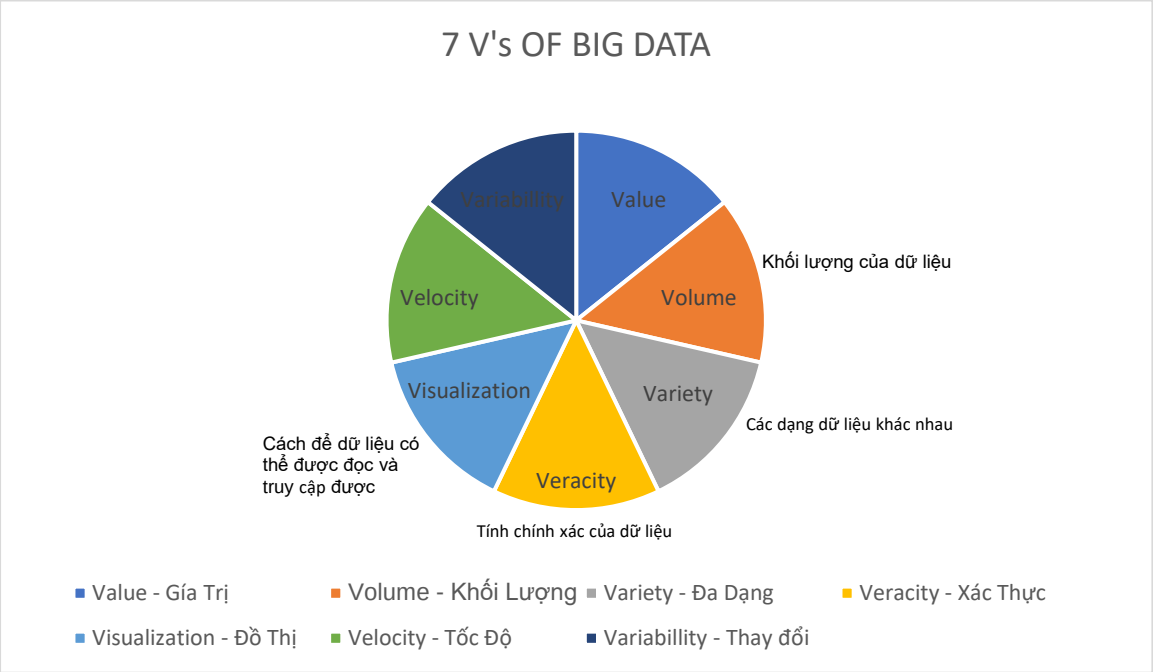
2. **Phân loại 1 - 6 v dữ liệu lớn, phát triển thành giá trị dữ liệu, làm cho nó trở thành 7 V của dữ liệu lớn.**

Bảy chữ V là: Khối lượng, Vận tốc, Sự đa dạng, Tính biến đổi, Tính xác thực, Trực quan hóa và Giá trị.

Tầm quan trọng của dữ liệu lớn không nằm ở lượng dữ liệu mà chúng ta có, nó nằm ở việc chúng ta làm gì với những dữ liệu đó. Ta có thể sử dụng nguồn dữ liệu lớn phân tích để tìm ra câu trả lời cho các câu hỏi: giảm chi phí, giảm thời gian, phát triển sản phẩm mới và dịch vụ tối ưu, ra quyết định thông minh. Khi việc phân tích nguồn dữ liệu lớn được hỗ trợ tối đa ta có thể hoàn thành tốt một số tác vụ như: xác định nguyên nhân gốc rễ của những thất bại, tạo các chương

trình khuyến mại hợp lí dựa trên thói quen của khách hàng đối với công việc kinh doanh, tính toán được những rủi ro gặp phải, phát hiện hành vi gian lận trước khi nó có ảnh hưởng đến chúng ta.

Tóm lại, Big data chính là thách thức đặt ra cho các doanh nghiệp trong thời đại công nghệ số. Một khi tận dụng được tối đa nguồn dữ liệu lớn thì có hội thành công sẽ lớn hơn nhiều lần. Tuy vẫn còn những chỉ trích về Big data nhưng đây là một lĩnh vực còn rất mới nên chúng ta hãy chờ đón sự tiến hóa của Big data trong tương lai.



Hình 3: Dữ liệu lớn 7V (Nguồn từ bài báo)

1. **Value**: Giá trị
2. **Volume**: Khối lượng
3. **Variety:** Đa dạng
4. **Veracity**: Xác thực
5. **Visualization**: Đồ thị
6. **Velocity:** Tốc độ
7. **Variabillity:** Thay đổi

II. NGÔN NGỮ TRUY VẤN CÓ CẤU TRÚC SQL

(MySQL) Nơi dữ liệu được lưu trữ trong một bảng có hàng và cột. Các nhà phát triển hồi đó chủ yếu triển khai các thiết kế của họ theo mô hình phát triển phần mềm thác nước. Điều này có nghĩa là mọi giai đoạn phát triển phần mềm đều được lên kế hoạch tốt trước khi quá trình phát triển bắt đầu bằng cách sử dụng mối quan hệ thực thể phức tạp kỹ lưỡng, đảm bảo rằng tất cả những gì cần thiết trong cơ sở dữ liệu đã được suy nghĩ và cung cấp cẩn thận (Schaefer, 2015)

Một cách tiếp cận quản lý mới được coi là cần thiết để hỗ trợ các ứng dụng như xem xét thời gian thực các tệp nhật ký, giao dịch thương mại điện tử và dữ liệu được đăng lên phương tiện truyền thông xã hội có khối lượng khổng lồ

III. DỮ LIỆU LỚN VÀ CƠ SỞ DỮ LIỆU

Dữ liệu lớn có thể được lưu trữ bằng cách sử dụng cả cơ sở dữ liệu có cấu trúc (MySQL là cơ sở dữ liệu quan hệ) và cơ sở dữ liệu phi cấu trúc (MongoDB là cơ sở dữ liệu phi quan hệ). Có tính đến sự thay đổi về thời gian phản hồi của từng loại cơ sở dữ liệu, các thuật toán khác nhau cần được phân tích để nâng cao hiệu suất trong việc giám sát hệ thống theo thời gian thực liên quan đến cả cập nhật SQL và NoSQL và chèn dữ liệu lớn.

Hệ thống cơ sở dữ liệu NoSQL được gọi là hệ thống lưu trữ dựa trên web miễn là chúng đáp ứng các yêu cầu sau:

- **Mô hình**: Mô hình cơ sở dữ liệu cơ bản không phải là quan hệ.
- **Ít nhất 3 V's:** Một lượng lớn dữ liệu (khối lượng), cấu trúc dữ liệu linh hoạt (đa dạng) và xử lý thời gian thực được đưa vào hệ thống cơ sở dữ liệu (vận tốc).
- **Lược đồ:** Lược đồ cơ sở dữ liệu tập hợp không bị ràng buộc bởi hệ thống quản lý cơ sở dữ liệu.
- **Kiến trúc:** Kiến trúc cơ sở dữ liệu hỗ trợ mở rộng quy mô ngang và các ứng dụng web được phân phối đầy đủ.
- **Nhân rộng:** Hệ thống quản lý cơ sở dữ liệu hỗ trợ sao chép dữ liệu.

- **Đảm bảo tính nhất quán:** tính nhất quán có thể được đảm bảo với độ trễ để ưu tiên tính khả dụng cao và khả năng chịu đựng của các phân vùng. (2019 Andreas & Michael)

Nguồn: từ bài báo BIGDA và NOSQL

IV. CẤU TRÚC BIGDATA VÀ NOSQL

NoSQL còn có nghĩa là Non-Relational (NoRel) - không ràng buộc. Tuy nhiên, thuật ngữ đó ít phổ dụng hơn và ngày nay người ta thường dịch NoSQL thành Not Only SQL - Không chỉ SQL. NoSQL ám chỉ đến những cơ sở dữ liệu không dùng mô hình dữ liệu quan hệ để quản lý dữ liệu trong lĩnh vực phần mềm. Các tính năng của NoSQL:

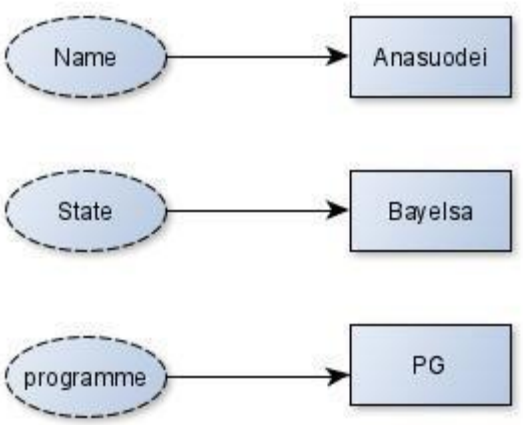
- Lược đồ miễn phí
- **Sự nhất quán**: tính nhất quán của dữ liệu không cần phải đảm bảo ngay tức khắc sau mỗi phép write. Một hệ thống phân tán chấp nhận những ảnh hưởng theo phương thức lan truyền và sau một khoảng thời gian (không phải ngay tức khắc), thay đổi sẽ đi đến mọi điểm trong hệ thống, tức là cuối cùng (eventually) dữ liệu trên hệ thống sẽ trở lại trạng thái nhất quán.
- Sao chép các kho dữ liệu để loại bỏ một điểm thất bại duy nhất. Có khả năng xử lý một loạt dữ liệu và khối lượng lớn dữ liệu.

Nguồn: https://viblo.asia/p/tim-hieu-ve-nosql-Zzb7vDNYMjKd

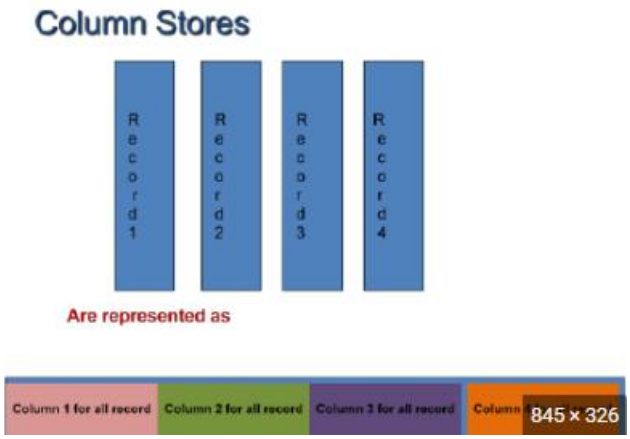
V. CÁC CẤU TRÚC CỦA NOSQL

- **Key-value stores:** là cơ sở dữ liệu NoSQL đơn giản nhất. Mỗi mục trong cơ sở dữ liệu được lưu trữ dưới dạng tên thuộc tính (hoặc 'khóa'), cùng với giá trị của nó. Ví dụ về Key-value stores là Riak, Berkeley DB, Amazon DynamoDB...

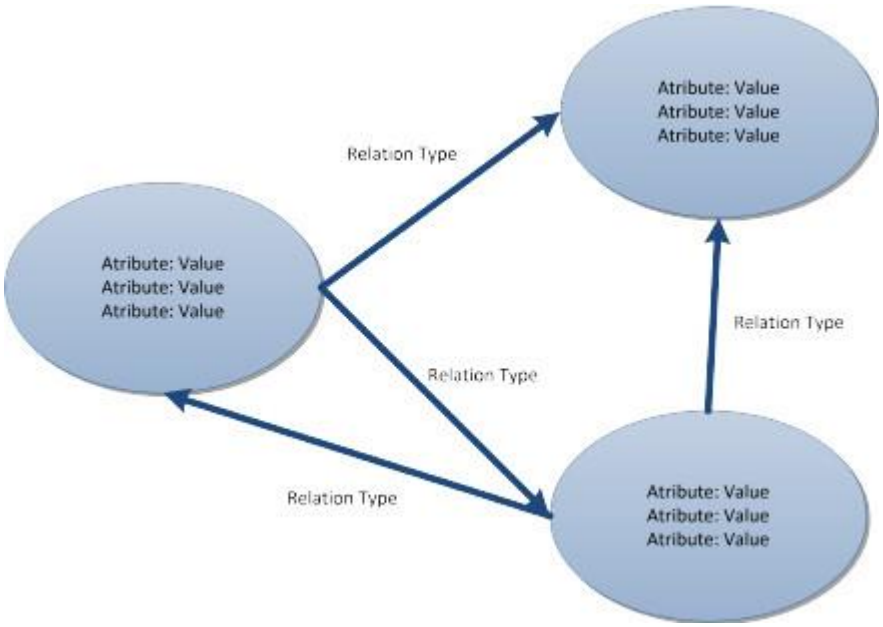
Khoá Giá trị



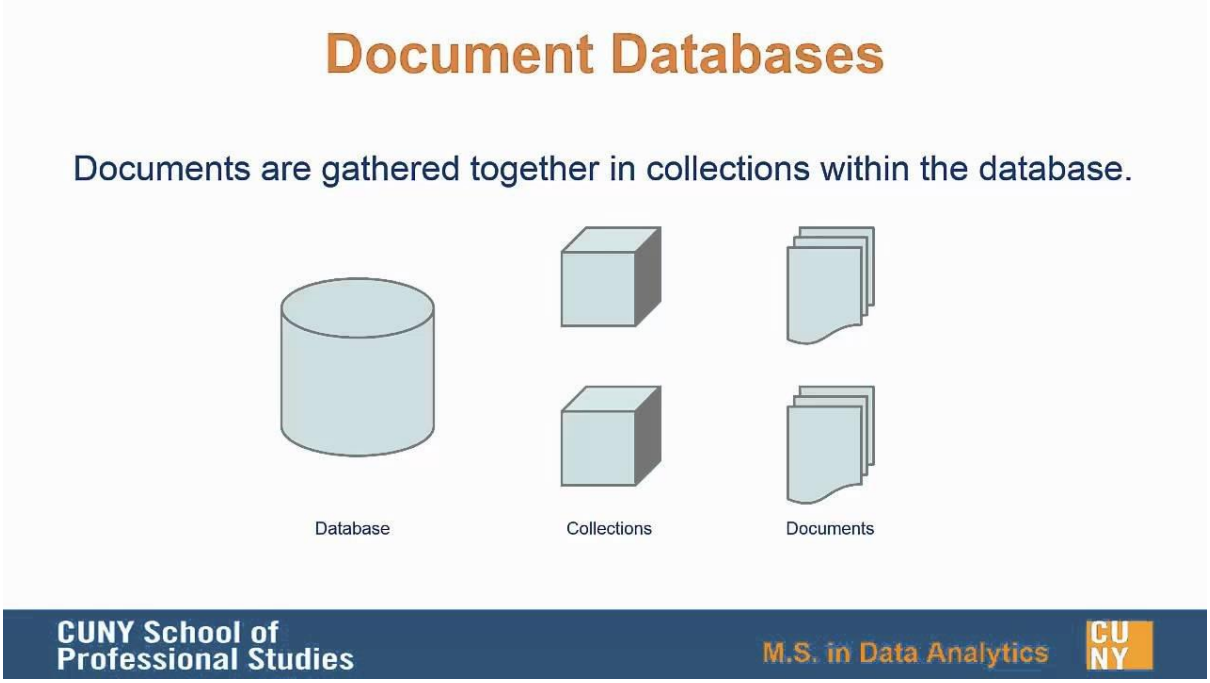
- **Column-oriented stores**: như Cassandra và HBase được tối ưu hóa cho các truy vấn trên các bộ dữ liệu lớn và lưu trữ các cột dữ liệu cùng nhau, thay vì các hàng.



- **Graph stores:** được sử dụng để lưu trữ thông tin về các mạng dữ liệu, chẳng hạn như các kết nối xã hội. Ví dụ Graph stores: Neo4J và Giraph...



- **Document Oriented databases:** ghép từng khóa với cấu trúc dữ liệu phức tạp được gọi là tài liệu. Tài liệu có thể chứa nhiều cặp khóa-giá trị khác nhau hoặc cặp khóa-mảng hoặc thậm chí các tài liệu lồng nhau. Ví dụ: MongoDB, OrientDB, RavenDB



VI. ĐIỂM MẠNH VÀ ĐIỂM YẾU CỦA NOSQL

1. Ưu điểm

- Có một số lợi thế, điểm mạnh khi làm việc với cơ sở dữ liệu NoSQL như MongoDB và Cassandra. Những ưu điểm chính của nosql là khả năng mở rộng và tính sẵn sàng cao.
- NoSQL giải quyết được các vấn đề dữ liệu lớn(big data) về các hệ thống thông tin hoặc là phân tán dữ liệu. Việc mở rộng phạm vi là mềm dẻo: NoSQL thay thế câu thần chú cũ của các nhà quản trị CSDL về việc 'mở rộng phạm vi' với một thứ mới: 'mở rộng ra ngoài'. Thay vì phải bổ sung thêm những máy chủ lớn hơn để tải nhiều dữ liệu hơn, thì CSDL NoSQL cho phép một công ty phân tán tải qua nhiều máy chủ khi tải gia tăng. High availability: Khả năng tự động sao chép trong MongoDB làm cho nó rất tốt trong mọi trường hợp vì trong trường hợp có bất kỳ lỗi nào, dữ liệu sẽ tự động sao chép về trạng thái nhất quán trước đó.

2. Nhược điểm

Bên cạnh những ưu điểm của nó thì NoSQL Database cũng có những nhược điểm sau:

- **Quản lý dữ liệu:** Mục đích của các công cụ dữ liệu lớn là làm cho việc quản lý một lượng lớn dữ liệu trở nên đơn giản nhất. Nhưng quản lý dữ liệu trong NoSQL phức tạp hơn nhiều so với cơ sở dữ liệu quan hệ. Đặc biệt, NoSQL nổi tiếng là khó cài đặt và thậm chí là để quản lý nó hằng ngày cũng tốn khá nhiều thời gian.
- **Sao lưu dữ liệu**: Sao lưu là một điểm yếu lớn đối với một số cơ sở dữ liệu NoSQL như MongoDB. Nó không có cách tiếp cận để làm sao lưu dữ liệu một cách nhất quán.
- **Thiếu tính nhất quán:** NoSQL đánh đổi sự nhất quán để ưu tiên tốc độ, hiệu suất hiệu quả hơn.
- **Trọng tâm hẹp:** Cơ sở dữ liệu NoSQL có trọng tâm rất hẹp vì nó chủ yếu được thiết kế để lưu trữ nhưng nó cung cấp rất ít chức năng.
- Mã nguồn mở: NoSQL là cơ sở dữ liệu mã nguồn mở và không có tiêu chuẩn đáng tin cậy cho NoSQL được nêu ra.
- **Không có lược đồ**: Ngay cả khi bạn lấy dữ liệu ở dạng tự do, bạn hầu như luôn cần áp đặt các ràng buộc để làm cho nó hữu ích. Với NoSQL, trách nhiệm sẽ được chuyển từ cơ sở dữ liệu sang nhà phát triển, lập trình ứng dụng.

- **Kỹ năng NoSQL:** Một hạn chế khác đối với NoSQL là người sử dụng có thể sẽ thiếu các kỹ năng chuyên môn ở mức tương đối vì hệ thống này còn khá mới và không phải ai cũng biết sử dụng nó một cách thành thạo.

Nguồn: https://chiasekinang.com/nosql-la-gi-mot-so-uu-diem-va-nhuoc-diem-can-biet-ve-nosql/

VII. DEMO BÁO CÁO VÀ PHÂN TÍCH DỮ LIỆU DOANH SỐ BÁN HÀNG

Đầu tiên chúng ta demo về doanh số bán hàng, nhập file data để báo cáo và phân tích dữ liệu doanh số của doanh nghiệp.

```
import pandas as pd
import os
from google.colab import drive
drive.mount('/content/drive')
import matplotlib.pyplot as plt
filename =input('Nhập tên tệp tin cần phân tích: ')
df =pd.read_csv(filename)
df

```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Nhập tên tệp tin cần phân tích: drive/MyDrive/data/Sales2019.csv
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	141234	iPhone	1	700	01/22/19 21:25	944 Walnut St, Boston, MA 02215
1	141235	Lightning Charging Cable	1	14.95	01/28/19 14:15	185 Maple St, Portland, OR 97035
2	141236	Wired Headphones	2	11.99	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
3	141237	27in FHD Monitor	1	149.99	01/05/19 20:33	738 10th St, Los Angeles, CA 90001
4	141238	Wired Headphones	1	11.99	01/25/19 11:59	387 10th St, Austin, TX 73301
...	...	...	...	...	...	...
186845	319666	Lightning Charging Cable	1	14.95	12/11/19 20:58	14 Madison St, San Francisco, CA 94016
186846	319667	AA Batteries (4-pack)	2	3.84	12/01/19 12:01	549 Willow St, Los Angeles, CA 90001
186847	319668	Vareebadd Phone	1	400	12/09/19 06:43	273 Wilson St, Seattle, WA 98101
186848	319669	Wired Headphones	1	11.99	12/03/19 10:39	778 River St, Dallas, TX 75001
186849	319670	Bose SoundSport Headphones	1	99.99	12/21/19 21:45	747 Chestnut St, Los Angeles, CA 90001

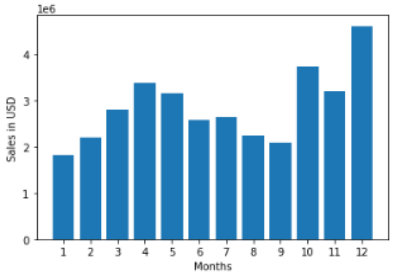
186850 rows × 6 columns

Kết luận: Với bước này chúng ta sẽ xác định được số lượng dữ liệu cần phải phân tích và báo cáo dữ liệu của doanh nghiệp

Demo tiếp theo nói về tháng nào có doanh thu lớn nhất? Doanh số tháng đó là bao nhiêu?

```
[10]  1 df['Quantity Ordered'] = pd.to_numeric(df['Quantity Ordered'], downcast='integer')
      2 df['Price Each'] = pd.to_numeric(df['Price Each'], downcast='float')
      3 df['Sales'] = df['Quantity Ordered'] * df['Price Each']
      4 moving_column = df.pop('Sales')
      5 df.insert(4, 'Sales', moving_column)
      6 df.groupby('Month').sum()['Sales']
      7 sales_value = df.groupby('Month').sum()['Sales']
      8 months = range(1,13)
      9 plt.bar(x=months, height=sales_value)
     10 plt.xticks(months)
     11 plt.xlabel('Months')
     12 plt.ylabel('Sales in USD')
     13 plt.show()
     14
```

```
<ipython-input-10-8038b03d2645>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Quantity Ordered'] = pd.to_numeric(df['Quantity Ordered'], downcast='integer')
<ipython-input-10-8038b03d2645>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Price Each'] = pd.to_numeric(df['Price Each'], downcast='float')
<ipython-input-10-8038b03d2645>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Sales'] = df['Quantity Ordered'] * df['Price Each']
```



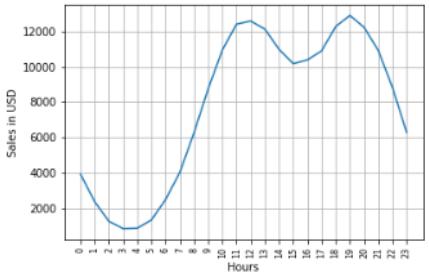
Kết luận: tháng 12 trong năm sẽ có doanh thu cao nhất

Doanh thu tháng 12 là: 4.613443e+06

Demo kế tiếp chúng em sẽ nói về doanh nghiệp cần chiếu quảng cáo vào khung thời gian nào để tăng khả năng mua hàng của khách hàng

```
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Hours'] = df['Order Date'].dt.hour
sales_value_hours = df.groupby('Hours').count()['Sales']
hours = [hour for hour, sales in sales_value_hours.items()]
plt.plot(hours, sales_value_hours)
plt.grid()
plt.xticks(hours, rotation=90, size=8)
plt.xlabel('Hours')
plt.ylabel('Sales in USD')
plt.show()
```

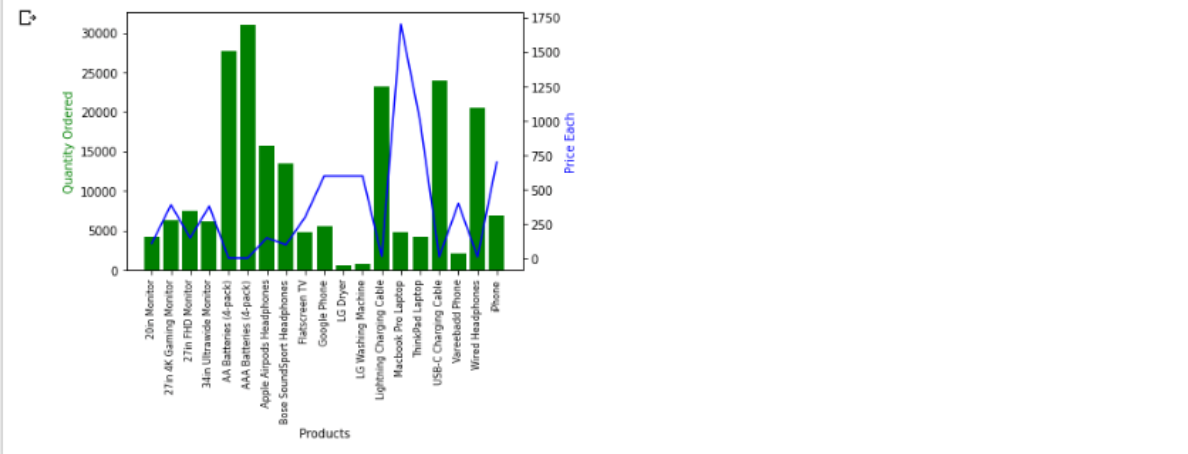
```
<ipython-input-12-c442410627e0>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Order Date'] = pd.to_datetime(df['Order Date'])
<ipython-input-12-c442410627e0>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Hours'] = df['Order Date'].dt.hour
```



Kết luận: Doanh nghiệp cần chiếu quảng cáo vào khung thời gian nào để tăng khả năng mua hàng của khách hàng vào 7 giờ đêm, để tiếp cận đến 12 000 traffic

Demo cuối cùng sẽ nói về sản phẩm nào được bán nhiều nhất? Giả thiết của bạn về lý do sản phẩm này được bán nhiều nhất là gì?

```
all_products = df.groupby('Product').sum()['Quantity Ordered']
prices = df.groupby('Product').mean()['Price Each']
products_ls = [product for product, quant in all_products.items()]

x = products_ls
y1 = all_products
y2 = prices

fig, ax1 = plt.subplots()

ax2 = ax1.twinx()
ax1.bar(x, y1, color='g')
ax2.plot(x, y2, 'b-')

ax1.set_xticklabels(products_ls, rotation=90, size=8)
ax1.set_xlabel('Products')
ax1.set_ylabel('Quantity Ordered', color='g')
ax2.set_ylabel('Price Each', color='b')

plt.show()
```



Giả thuyết: Các sản phẩm đó thì giá thành tốt, dễ phân phối ra thị trường

Kết luận: Sản phẩm aaa patterief (4-pack) là sản phẩm bán chạy nhất

Nguồn:
https://colab.research.google.com/drive/1joavvic4JLScRquPFmw2UtESnc2ukoax?usp=sharing#scrollTo=btrIpYTcLohi

VII. KẾT LUẬN

Ngoài ra, nhờ bài báo cáo này mà nhóm chúng em đã có thể tận dụng ngôn ngữ Python để tiếp cận các dữ liệu, báo cáo &phân tích dữ liệu của gười dung chodoanh nghiệp nắm bắt kịp thời về tình trạng , hiệu suất cũng như tiến độ công việc. Bằng cách phân tích & báo cáo về các dữ liệu doanh thu của bất cứ doanh nghiệp nào với thời gian cực kì nhanh chóng, chúng em sẽ cố gắng phát triển và dùng những kiến từ thừ bài nghiên cứu này để có thể giúp người dung có thể nhập bất kì định dạng file mà không phải csv nào. Chúng em cảm thấy đồ án này rất hay và thú vị, thế nên sau đồ án này hướng phát triển của chúng e trong tương lai là sẽ tạo ra được nhiều truy vấn để hình thành insight của khách hàng và giúp doanh nghiệp thấu hiểu, phát triển sản phẩm của mình lẫn doanh thu một cách nhanh chóng và hoàn thiện nhất.

Qua quá trình học tập và nghiên cứu dữ liệu lớn trong thời gian gần đây có và cách xử lý dữ liệu khối lượng ngày càng nhiều. Đồng thời, cơ sở dữ liệu có cấu trúc gặp khó khăn khi xử lý dữ liệu phi cấu trúc do kích thước của nó. Để hiểu được Dữ liệu lớn, cấu trúc và phương pháp mới luôn là điều cần thiết, việc nghiên cứu này cũng đã kiểm tra từ các cấu trúc cơ sở dữ liệu NoSQL dữ liệu lớn khác nhau, các loại liên quan đến chúng, tầm quan trọng và cách sử dụng

LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn sâu sắc đến Trường Đh Tài Nguyên và Môi Trường đã đưa bộ môn Hệ Quản Trị CSDL vào chương trình giảng dạy. Đặc biệt, chúng em xin bày tỏ lòng biết ơn sâu sắc đến giảng viên bộ môn - thầy Phạm Trọng Huynh, người đã tận tình dạy dỗ và truyền đạt những kiến thức quý báu cho chúng em trong suốt học kỳ vừa qua. Trong thời gian tham dự lớp học của thầy, chúng em đã được tiếp cận với nhiều kiến thức bổ ích và rất cần thiết cho quá trình học tập, làm việc sau này của chúng em.

Bộ môn Hệ Quản Trị CSDL là một môn học thú vị và vô cùng bổ ích. Tuy nhiên, những kiến thức và kỹ năng về môn học này của chúng em vẫn còn nhiều hạn chế. Do đó, bài tiểu luận của chúng em khó tránh khỏi những sai sót. Kính mong thầy cô xem xét và góp ý giúp bài tiểu luận của chúng em được hoàn thiện hơn.

Em xin chân thành cảm ơn!