Trần Nhật Tân MSSV: 19522177

Nguyễn Viết Trường MSSV: 19522443

BÀI TẬP A1.1

**Phân tích nhu cầu báo cáo và đề xuất thiết kế một kho dữ liệu theo chủ đề tùy chọn**

Mục lục bài tập

[**1. Phân tích dữ liệu chất lượng nguồn nước.** 2](#_Toc84443573)

[**2.** **Nội dung đề tài** 3](#_Toc84443574)

[**3.** **Mô tả dữ liệu gốc:** 3](#_Toc84443575)

[**4.** **Thiết kế kho dữ liệu** 7](#_Toc84443576)

[**5.** **10 câu truy vấn** 12](#_Toc84443577)

**1. Phân tích dữ liệu chất lượng nguồn nước.**

Nước là nhu phẩm thiết yếu nhất trong cuộc sống của con người. Nước chiếm tỉ lệ 70% cơ thể con người, nước có khả năng cung cấp khoáng chất, vận chuyển chất dinh dưỡng, oxy cần thiết cho các tế bào, nuôi dưỡng tế bào trong mọi hoạt động của cơ thể. Nước còn có tác dụng điều hòa nhiệt độ cơ thể, đào thải các độc tố thông qua mồ hôi, nước tiểu và đại tiện, giúp xương khớp hoạt động trơn tru. 

Hình 1.1. Nguồn nước cạnh một con kênh ô nhiễm rác

Nguồn nước sạch có ý nghĩa quan trọng đến đời sống kinh tế xã hội của một quốc gia hay toàn thể thế giới. Khi nguồn nước được đảm bảo, những nguy cơ đến sức khỏe cũng được giảm thiểu đến mức tối đa. Nhờ vậy mà các hoạt động sản xuất được đẩy mạnh góp phần tăng trưởng nền kinh tế . Quan trong hơn cả khi nguồn nước được đảm bảo những đối tượng có hệ miễn dịch yếu ớt sẽ hạn chế nguy cơ mắc bệnh hơn.

Với tầm ảnh hưởng quan trọng, WHO luôn đánh giá cao giá trị của nguồn nước. Chính vì vậy các vấn đề ô nhiễm nước đều cần được quan tâm để tìm ra nguyên nhân mà khắc phục kịp thời. Nếu nước là bắt nguồn của sự sống thì nước sạch là một nguồn tài nguyên quý giá hữu hạn và không gì sánh bằng.

Cùng với sự phát triển của loài người, nhu cầu sử dụng tài nguyên càng lớn hơn, việc đó đã tác động qua lại đối với các nguồn tài nguyên. Việc phát triển nhanh chóng của nền công nghiệp đã vô tình tạo ra 1 lượng lớn chất thải độc hại, việc khai thác 1 cách bất hợp lý tài nguyên cũng đã làm mất cân bằng hệ sinh thái, tệ hơn là chiến tranh.

Chính vì thế, việc phân tích dữ liệu cha là rất cần thiết để nhanh chóng tìm ra giải pháp cải thiện chất lượng nước, cũng là cải thiện chất lượng sống của nhân loại.

1. **Nội dung đề tài**

Cơ sở của bộ dữ liệu này được lấy từ dữ liệu chất lượng nước WaterBase được chia sẻ trên ‎[‎EAA‎](https://www.eea.europa.eu/)‎. Sau khi hầu hết các cột ở đó bị bỏ, dữ liệu mới đã được tạo ra với sự trợ giúp của Worldbank, OSM, Foursquare, SEDAC. Sau đó, khoảng cách của một số loại đường gần tọa độ đó cũng được thêm vào với OSM. Người ta cho rằng những thông tin như vậy đóng một vai trò quan trọng trong việc ô nhiễm nước.‎

Nguồn dữ liệu: [Water Quality Dataset | Kaggle](https://www.kaggle.com/ozgurdogan646/water-quality-dataset)   
https://www.kaggle.com/ozgurdogan646/water-quality-dataset

Chủ đề: Con người và xã hội  
Dataset gồm : 29 cột và 20000 dòng dữ liệu

1. **Mô tả dữ liệu gốc:**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ý nghĩa |
| 1 | observedPropertyDeterminandCode | Varchar | Mã duy nhất của định thức được giám sát, như được xác định trong danh sách mã. |
| 2 | resultUom | Varchar | Đơn vị đo kết quả |
| 3 | phenomenonTimeReferenceYear | Int | Năm lấy mẫu |
| 4 | parameterSamplingPeriod | Date | Khoảng thời gian trong năm dữ liệu được sử dụng để tổng hợp được lấy mẫu |
| 5 | resultMeanValue | Float | Giá trị kết quả trung bình |
| 6 | waterBodyIdentifier | Varchar | Nhận dạng quốc tế của cơ quan nước |
| 7 | Country | Varchar | Quốc gia của mẫu nước |
| 8 | PopulationDensity | Float | Mật độ dân số |
| 9 | TouristMean\_1990\_2020 | Int | Tổng khách du lịch từ 1990-2020 |
| 10 | composition\_food\_organic\_waste\_percent | Float | Tỉ lệ thành phần rác hữu cơ thực phẩm |
| 11 | composition\_glass\_percent | Float | Tỉ lệ thành phần thủy tinh |
| 12 | composition\_metal\_percent | Float | Tỉ lệ thành phần kim loại |
| 13 | composition\_other\_percent | Float | Tỉ lệ thành phần khác |
| 14 | composition\_paper\_cardboard\_percent | Float | Tỉ lệ thành phần giấy |
| 15 | composition\_plastic\_percent | Float | Tỉ lệ thành phần nhựa |
| 16 | composition\_rubber\_leather\_percent | Float | Tỉ lệ thành phần cao su |
| 17 | composition\_wood\_percent | Float | Tỉ lệ thành phần gỗ |
| 18 | waste\_treatment\_recycling\_percent | Float | Tỉ lệ rác thải tái chế |
| 19 | composition\_yard\_garden\_green\_waste\_percent | Float | Tỉ lệ rác thải từ cây xanh |
| 20 | parameterWaterBodyCategory | Varchar | Mã phân loại thủy vực, như được xác định trong danh sách mã hóa |
| 21 | procedureAnalysedFraction | Varchar | Đặc điểm kĩ thuật của phần mẫu được phân tích |
| 22 | procedureAnalysedMedia | Varchar | Loại phương tiện được giám sát. Là nước hoặc trầm tích |
| 23 | TerraMarineProtected\_2016\_2018 | Float | Trung bình của các khu vực biển Terra được bảo vệ của Quốc gia Từ năm 2016-2018 |
| 24 | VenueCount | Int | Số lượng địa điểm gần các tọa độ nhất định |
| 25 | netMigration\_2011\_2018 | Float | Tỷ lệ di cư trung bình của một quốc gia nhất định trong giai đoạn 2011-2018 |
| 26 | droughts\_floods\_temperature | Float | Chỉ số hạn hán, lũ lụt, nhiệt độ |
| 27 | literacyRate\_2010\_2018 | Float | Tỷ lệ thu nhập của quốc gia từ năm 2010-2018 |
| 28 | combustibleRenewables\_2009\_2014 | Float | Số lượng có thể tái tạo được ở Quốc gia từ năm 2009-2014 |
| 29 | gdp | Varchar | Thu nhập bình quân đầu người của quốc gia |

1. **Thiết kế kho dữ liệu**

Lược đồ hình sao :

Graphical user interface, text

Description automatically generated

Các bảng chiều:

DIM\_DATE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | Date\_id | Varchar | Khóa chính | Tự động tăng |
| 2 | phenomenonTimeReferenceYear | int |  | Năm lấy mẫu |

DIM\_COUNTRY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | Country | Varchar | Khóa chính | Tên quốc gia |
| 2 | PopularDestiny | int |  | Mật độ dân số |
| 3 | TouristMean | int |  | Số lượng du khách |

DIM\_WATER

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | waterBodyIdentifier | Varchar | Khóa chính | Mã định danh của cơ quan nước của quốc gia lấy mẫu |
| 2 | resultUom | Varchar |  | Đơn vị đo kết quả |
| 3 | observedPropertyDeterminandCode | Varchar |  | Mã duy nhất của định thức được giám sát, như được xác định trong danh sách mã |
| 4 | resultMeanValue | float |  | Giá trị kết quả trung bình |

Bảng FACT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | Date\_id | Varchar | Khóa chính | Năm khảo sát |
| 2 | Country | Varchar |  | Tên quốc gia lấy mẫu |
| 3 | waterBodyIdentifier | Varchar |  | Mã định danh của cơ quan nước của quốc gia lấy mẫu |
| 4 | composition\_food\_organic\_waste\_percent | Float |  | Tỉ lệ thành phần rác hữu cơ thực phẩm |
| 5 | composition\_glass\_percent | Float |  | Tỉ lệ thành phần thủy tinh |
| 6 | composition\_metal\_percent | Float |  | Tỉ lệ thành phần kim loại |
| 7 | composition\_other\_percent | Float |  | Tỉ lệ thành phần khác |
| 8 | composition\_paper\_cardboard\_percent | Float |  | Tỉ lệ thành phần giấy |
| 9 | composition\_plastic\_percent | Float |  | Tỉ lệ thành phần nhựa |
| 10 | composition\_rubber\_leather\_percent | Float |  | Tỉ lệ thành phần cao su |
| 11 | composition\_wood\_percent | Float |  | Tỉ lệ thành phần gỗ |
| 12 | waste\_treatment\_recycling\_percent | Float |  | Tỉ lệ rác thải tái chế |
| 13 | composition\_yard\_garden\_green\_waste\_percent | Float |  | Tỉ lệ rác thải từ cây xanh |

1. **10 câu truy vấn**
2. Thống kê thông tin mẫu nước nào có tỉ lệ thành phần giấy cao nhất ?
3. Thống kê những mẫu nước, thông tin quốc gia nào có nhiều mẫu nước có tỉ lệ nitrat, nitrit lớn hơn tiêu chuẩn ?  
   Ảnh có chứa văn bản, trong nhà, ảnh chụp màn hình

   Mô tả được tạo tự động
4. Thống kê số lượng các mẫu nước, thông tin của quốc gia đó theo từng năm ?
5. Quốc gia có nhiều du khách nhất trong năm 2015 ?
6. Thống kê các mẫu nước có tỉ lệ % plastic giảm dần trong năm 2009 ?
7. Thống trong từng năm, thì loại xét nghiệm thuộc tính nào được thực hiện và liệt kê, tổng số lần ?
8. Thống kê Top 10 các mẫu nước kết quả trung bình cao nhất, liệt kê theo từng thuộc tính xét nghiệm ?
9. Với mỗi quốc gia, hãy liệt kê 2 mẫu nước xét nghiệm nhiều và ít nhất theo các năm ?
10. Tính kết quả trung bình (resultUOM) của các loại xét nghiệm theo từng năm?
11. Thống kê các chỉ số thành phần của 10 mẫu nước mà chỉ số MgNo3 cao nhất ?

**Hết.**