**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu lương thực và nông nghiệp toàn cầu khai thác sử dụng tại các cơ quan quản lý và kinh doanh lương thực và nông nghiệp**

Lớp: **181330A, 181330B, 181330C**

**NHÓM : 2**

4. Trần Gia Nguyên [29]: Phân hệ phân tích kho dữ liệu Cây trồng

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác}

**Ngày nộp: 15/06/2021**

**TP.HCM, ngày 15 Tháng 06 năm 2021**

**NHẬN XÉT CỦA GIẢNG VIÊN ĐỐI VỚI NHÓM:**

**………………………………………………………………..**

**………………………………………………………………..**

**……………………………………………………………….**

**NHÓM : 2**

**ĐIỂM ĐÁNH GIÁ**

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác: NẾU DS DÀI = Trang riêng: 1 lần :BỎ DS BÌA & BÌA LÓT}

1. Trần Gia Nguyên [29]: Phân hệ phân tích kho dữ liệu Cây trồng **Ngày 15/06/2021**

**Giảng viên Ký tên**

THÀNH VIÊN:

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác}

1. Trần Gia Nguyên [29]: Phân hệ phân tích kho dữ liệu Cây trồng

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu lương thực và nông nghiệp toàn cầu khai thác sử dụng tại các cơ quan quản lý và kinh doanh lương thực và nông nghiệp**

**Giảng viên giảng dạy: VÕ XUÂN THỂ \_ Giảng viên chính**

Lớp: **181330A, 181330B, 181330C (Sáng thứ 4)**

**Tên sản phẩm đề tài: 29.Nx.18133036.TranGiaNguyen.DAHP.DWH.rar** (Microsoft SQL Server 2014)

**DB Nguồn: FAOSTAT**

**DB DWH: FAOSTAT\_CayTrong\_DWH**

**Tên máy chủ DB SQL: DESKTOP-CLC1LVD**

**Tên đăng nhập DB SQL: sa**

**Mật khẩu (password) đăng nhập SQL: 123**

**ProjectName BIDS: FAOSTAT\_CayTrong VD: 05N1VXTheQLKCBBHYT\_DWH**

**NHÓM : 2**

1. THÀNH VIÊN: 1 Trần Gia Nguyên [29]: Phân hệ phân tích kho dữ liệu Cây trồng

. . . .

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác}

**TP.HCM, ngày 15 Tháng 06 năm 2021**

# LỜI CẢM ƠN

Nhóm chúng em xin gửi lời cảm ơn chân thành đến thầy Võ Xuân Thể - giảng viên bộ môn Kho dữ liệu trong khoa Công Nghệ Thông Tin đã trang bị cho nhóm chúng em những kiến thức, kĩ năng cơ bản cần có để hoàn thành project này.

Tuy nhiên trong quá trình nghiên cứu đề tài, do kiến thức chuyên ngành còn hạn chế nên em vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá và trình bày về đề tài. Rất mong nhận được sự quan tâm, góp ý của các thầy/ cô giảng viên bộ môn để đề tài của em được đầy đủ và hoàn chỉnh hơn.

Xin chân thành cảm ơn.

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| HCMUTE | HCM University |
| DWH | Datawarehouse = Kho dữ liệu |
| DB | Database = Cơ sở dữ liệu |
| HCMUTE | HCM University |
| DWH | Datawarehouse = Kho dữ liệu |
|  |  |
| BIDS | Công cụ hỗ trợ tổ chức, quản lý và khai thác kho dữ liệu (Hệ thống tổ chức dữ liệu thông minh cho các doanh nghiệp): Business Intelligence Development Studio của Microsoft |
| DWH | Kho dữ liệu (Nhà kho dữ liệu): DataWareHouse |
| ERD | Sơ đồ thực thể kết hợp: Entity Relationship Diagram |
| MS | Công ty Microsoft |
| NF | Dạng chuẩn của CSDL: Normal Form |
| SQL | Ngôn ngữ vấn tin có cấu trúc: Structured Query Language |

# DANH MỤC CÁC THUẬT NGỮ

|  |  |
| --- | --- |
| Giao diện người dùng  (User-Interface) | Là hệ thống các màn hình giao tiếp cho phép người sử dụng tương tác với các thành phần phần mềm, điều khiển phần mềm hoạt động theo yêu cầu của người dùng - tương ứng các chức năng hiện có của phần mềm. |
| Người dùng (User):  Tài khoản (Account) | Là một quyền làm việc trên hệ thống phần mềm được cấp phát cho một cá nhân thông qua tên tài khoản (username) và mật khẩu (password). |
|  |  |
| BigData | Dữ liệu lớn: là một tập hợp dữ liệu rất lớn và phức tạp, không thể xử lý dữ liệu bằng các phương pháp truyền thống. |

# DANH MỤC HÌNH ẢNH

[Hình 1: Diagram Database gốc 25](#_Toc74662725)

[Hình 2: Diagram DataWarehouse phân hệ cá nhân 26](#_Toc74662726)

[Hình 3: Diagram FAOSTAT\_Caytrong\_DWH 30](#_Toc74662727)

[Hình 4: Tạo dbo.DimCountry 34](#_Toc74662728)

[Hình 5: Tạo dbo.DimElement 34](#_Toc74662729)

[Hình 6: Tạo dbo.DimItem 35](#_Toc74662730)

[Hình 7: Tạo dbo.DimYear 35](#_Toc74662731)

[Hình 8: Tạo Fact 36](#_Toc74662732)

[Hình 9: Solution explorer 36](#_Toc74662733)

[Hình 10: Deploy thành công 37](#_Toc74662734)

[Hình 11: Process thành công 37](#_Toc74662735)

[Hình 12: Phân tích sản lượng theo quốc gia – năm 38](#_Toc74662736)

[Hình 13: Phân tích sản lượng 1 vài loại cây trồng tại Việt Nam giai đoạn 2015-2019 39](#_Toc74662737)

[Hình 14: Phân tích sản lượng 1 vài loại cây trồng tại 5 nước trong giai đoạn 2015-2019 40](#_Toc74662738)

[Hình 15: Tổng quan các Chart 50](#_Toc74662739)

[Hình 16: Tạo thống kê đơn giản cho kho 50](#_Toc74662740)

[Hình 17: Ví dụ về thống kê tổng sản lượng một vài loại nông trong giai đoạn 2015-2019 51](#_Toc74662741)

[Hình 18: Ví dụ về thống kê tổng sản lượng 1 vài loại nông sản ở Việt Nam năm 2015-2019 52](#_Toc74662742)

[Hình 19: Biểu đồ Donut thể hiện tỉ lệ sản lượng nông sản tại Việt Nam năm 2019 52](#_Toc74662743)

[Hình 20: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (1) 53](#_Toc74662744)

[Hình 21: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (2) 54](#_Toc74662745)

[Hình 22: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (3) 55](#_Toc74662746)

[Hình 23: Biểu đồ cột thể hiện tổng sản lượng các nước năm 2019 55](#_Toc74662747)

[Hình 24: Báo cáo về tỉ lệ sản lượng các loại nông sản tại Việt Nam năm 2019 56](#_Toc74662748)

[Hình 25: Chú thích biểu đồ Donut 57](#_Toc74662749)

[Hình 26: Biểu đồ cột thể hiện sản lượng các loại cây trồng tại Việt Nam năm 2019 58](#_Toc74662750)

# MỤC LỤC

[LỜI CẢM ƠN 4](#_Toc74662751)

[DANH MỤC CÁC TỪ VIẾT TẮT 5](#_Toc74662752)

[DANH MỤC CÁC THUẬT NGỮ 6](#_Toc74662753)

[DANH MỤC HÌNH ẢNH 7](#_Toc74662754)

[MỤC LỤC 8](#_Toc74662755)

[Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 12](#_Toc74662756)

[1.1. Tổng quan về ĐỒ ÁN HỌC PHẦN 12](#_Toc74662757)

[1.2. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN 12](#_Toc74662758)

[1.3. Mục tiêu của ĐỒ ÁN HỌC PHẦN 13](#_Toc74662759)

[1.4. Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN 13](#_Toc74662760)

[1.5. Sản phẩm của ĐỒ ÁN HỌC PHẦN 13](#_Toc74662761)

[1.6. Bố cục của báo cáo 13](#_Toc74662762)

[Chương 2. CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN 14](#_Toc74662763)

[2.1. Tổng quan về kho dữ liệu (DWH) 14](#_Toc74662764)

[2.1.1. Các khái niệm cơ bản về Kho dữ liệu 14](#_Toc74662765)

[2.1.2. Kiến trúc cơ bản của kho dữ liệu: 17](#_Toc74662766)

[2.1.3. Mô hình CSDL đa chiều (Dimensional Modeling) 17](#_Toc74662767)

[2.2. Phân tích và thiết kế Kho dữ liệu 18](#_Toc74662768)

[2.2.1. Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling) 18](#_Toc74662769)

[2.2.2. Các dạng (phương pháp) tích hợp dữ liệu 18](#_Toc74662770)

[2.2.3. Các bước cơ bản trong tiến trình tích hợp dữ liệu 18](#_Toc74662771)

[2.2.4. Các công cụ tích hợp dữ liệu: 19](#_Toc74662772)

[2.2.5. Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows) 19](#_Toc74662773)

[2.2.6. Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến) 19](#_Toc74662774)

[2.3. Vấn tin khai thác sử dụng Kho dữ liệu 20](#_Toc74662775)

[2.3.1. Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu 20](#_Toc74662776)

[2.3.2. Vấn tin với các hàm phân tích của SQL trên kho dữ liệu 20](#_Toc74662777)

[2.3.3. Vấn tin dùng Materialized View để phân tích trên kho dữ liệu 20](#_Toc74662778)

[2.3.4. Hệ thống OLAP 21](#_Toc74662779)

[Chương 3. TỔNG QUAN VỀ KHO DỮ LIỆU PHÂN TÍCH VÀ THIẾT KẾ KHO DỮ LIỆU LƯƠNG THỰC VÀ NÔNG NGHIỆP TẠI CÁC CƠ QUAN QUẢN LÍ VÀ KINH DOANH LƯƠNG THỰC VÀ NÔNG NGHIỆP 22](#_Toc74662780)

[3.1. Ý tưởng hình thành DWH phân tích và thiết kế kho dữ liệu lương thực và nông nghiệp tại các cơ quan quản lí và kinh doanh lương thực và nông nghiệp 22](#_Toc74662781)

[3.2. Giới thiệu chung về Kho dữ liệu 22](#_Toc74662782)

[3.2.1. Mô tả tổng quan về DWH 22](#_Toc74662783)

[3.2.2. Các dữ liệu nguồn hình thành DWH 22](#_Toc74662784)

[3.3. Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH 23](#_Toc74662785)

[3.4. Các thành phần phân hệ Kho dữ liệu phân công các cá nhân phụ trách 23](#_Toc74662786)

[4.1.1. Mô tả tổng quan về phân hệ DWH Phân hệ phân tích kho dữ liệu Cây trồng 24](#_Toc74662787)

[4.1.2. Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu 24](#_Toc74662788)

[4.1.3. Xác định ý tưởng hình thành phân hệ của Kho dữ liệu 25](#_Toc74662789)

[4.1.4. Mô tả hệ thống các DB gốc liên quan phân hệ DWH 25](#_Toc74662790)

[4.1.5. Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ 26](#_Toc74662791)

[4.1.6. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ 30](#_Toc74662792)

[4.1.7. Thực hiện các thủ tục tích hợp các DB gốc vào phân hệ DWH 31](#_Toc74662793)

[4.1.8. Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH 34](#_Toc74662794)

[4.1.9. Triển khai phân hệ DWH lên BIDS 36](#_Toc74662795)

[4.1.10. Khai thác sử dụng phân hệ DWH dùng các vấn tin SQL 41](#_Toc74662796)

[4.1.11 Xây dựng ứng dụng Winforms để trực quan hóa dữ liệu và tạo các báo cáo nhằm mục đích phân tích cơ bản kho dữ liệu: 47](#_Toc74662797)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN 59](#_Toc74662798)

[5.1. Kết luận 59](#_Toc74662799)

[5.1.1. Những kết quả đạt được 59](#_Toc74662800)

[5.1.2. Hạn chế 59](#_Toc74662801)

[5.2. Hướng phát triển 59](#_Toc74662802)

[5.2.1. Hướng khắc phục các hạn chế 59](#_Toc74662803)

[5.2.2. Hướng mở rộng ĐỒ ÁN HỌC PHẦN 59](#_Toc74662804)

[DANH MỤC TÀI LIỆU THAM KHẢO 61](#_Toc74662805)

[CÁC PHỤ LỤC 62](#_Toc74662806)

[ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN 63](#_Toc74662807)

# Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN

## Tổng quan về ĐỒ ÁN HỌC PHẦN

Đồ án Học phần Kho dữ liệu (DWH)

* Đồ án Học phần: Phát triển mô hình kho dữ liệu lương thực và nông nghiệp toàn cầu
* Với tình hình nông nghiệp phức tạp và tình hình dịch covid-19 đang khó khăn, các tổ chức nông nghiệp buộc phải thường xuyên đưa ra nhiều quyết định đúng đắn trong cùng một lúc một cách nhanh chóng. Do đó vấn đề trợ giúp quyết định trở lên rất cần thiết. Cần phải thu thập, tổng hợp và phân tích dữ liệu từ nhiều nguồn khác nhau một cách nhanh và hiệu quả thì mới có thể đưa ra được những quyết định nhanh chóng và phù hợp. Điều này dẫn đến việc cần phát triển những hệ thống thông minh, biết cách trích chọn và phân tích dữ liệu cho người sử dụng.
* Do đó những năm gần đây đã phát triển mạnh mẽ một loạt các lĩnh vực nghiên cứu về tổ chức các kho dữ liệu và thông tin (data wharehouse, information warehouse), các hệ thống trợ giúp quyết định, các phương pháp phát hiện tri thức và khai phá dữ liệu (data mining). Trong đó khai phá dữ liệu và phát hiện tri thức đã trở thành đề tài một nghiên cứu sôi động áp dụng rất nhiều lĩnh vực khác nhau như các hệ cơ sở dữ liệu, thống kê, học máy, trí tuệ nhân tạo…
* Còn trong lĩnh vực nông nghiệp, hàng ngày chúng ta tiếp nhận thông tin từ nhiều nguồn khác nhau, kho dữ liệu doanh nghiệp tổ chức ngày càng lớn, đòi hỏi ban lãnh đạo các tổ chức cần phải tiến hành phân tích để thấy được thuận lợi cũng như khó khăn để có hướng đi đúng đắn.
* Kho dữ liệu về lương thực và nông nghiệp có dữ liệu trải trên nhiều vùng lãnh thổ quốc gia. Nhóm chúng em thực hiện tạo kho dữ liệu này để có cái nhìn tổng quan về tình hình nông nghiệp cũng như thực phẩm

## Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN

* Nâng cao chất lượng dữ liệu bằng các phương pháp làm sạch và loại bỏ dữ liệu theo những dữ liệu không cần thiệt hoặc bị thiếu.
* Tích hợp kho dữ liệu
* Duy trì kho dữ liệu từ nhiều nguồn dữ liệu khác nhau.
* Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp như là các công cụ chuẩn để phục vụ cho DWH.
* Quản lí DWH.
* Cung cấp thông tin được tích hợp, tóm tắt hoặc được liên kết, tổ chức theo các chủ đề
* Dùng trong các hệ thống hỗ trợ quyết định (Decision support system – DSS), các hệ thống thông tin tác nghiệp hoặc hỗ trợ cho các truy vấn đặc biệt.

## Mục tiêu của ĐỒ ÁN HỌC PHẦN

* Khả năng đáp ứng mọi yêu cầu về thông tin của người sử dụng
* Hỗ trợ để các tổ chức thực hiện tốt, hiệu quả công việc của mình, cũng như có những quyết định hợp lý, nhanh chóng hơn,…
* Giúp cho tổ chức, xác định và phân tích các nghiệp vụ một các hiệu quả và chính xác
* Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau

## Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN

Các công cụ chính sử dụng trong đồ án này là

+ Microsoft SQL Server

+ SQL Server Business Intelligence Development Studio (BIDS)

+ Winforms, Devexpress, C#

## Sản phẩm của ĐỒ ÁN HỌC PHẦN

* Kho dữ liệu về lương thực và nông nghiệp toàn cầu.
* Ứng dụng trực quan hóa và phân tích dữ liệu cơ bản.

## Bố cục của báo cáo

Báo cáo gồm những nội dung như sau:

Chương 1: Giới thiệu đồ án của học phần DWH

Chương 2: Các cơ sở lý thuyết của HP phục việc thực hiện đề tài HP

Chương 3: Giới thiệu tổng quan về Kho dữ liệu (DWH) Kho dữ liệu lương thực và nông nghiệp toàn cầu khai thác sử dụng tại các cơ quan quản lý lương thực và nông nghiệp.

Chương 4: Phân tích, thiết kế và tổ chức khai thác sử dụng từng phân hệ của Kho dữ liệu

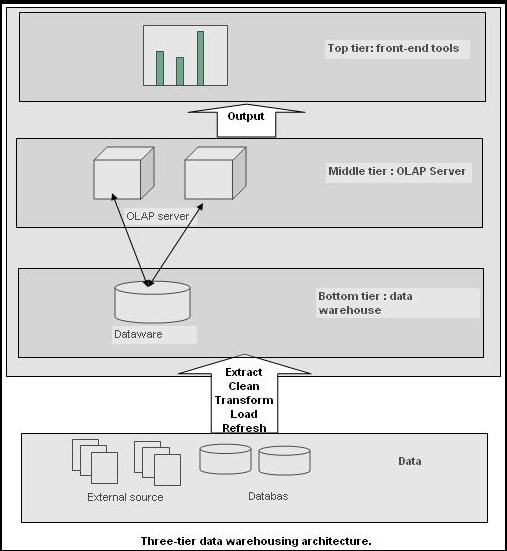
Chương 5: Tổng kết các kết quản đạt được và còn hạn chế của đồ án, đồng thời đề xuất hướng khắc phục hạn chế và phát triển Đồ án.

# Chương 2. CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN

## Tổng quan về kho dữ liệu (DWH)

### Các khái niệm cơ bản về Kho dữ liệu

* + - 1. **Kho dữ liệu (DWH)?**
* **Kho dữ liệu** ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): *data warehouse*) là kho lưu trữ dữ liệu lưu trữ bằng thiết bị điện tử của một tổ chức. Các kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và lập báo cáo.
* Định nghĩa cổ điển này về kho dữ liệu tập trung vào việc lưu trữ dữ liệu. Tuy nhiên, các phương tiện cho việc lấy và phân tích, trích rút, biến đổi, nạp dữ liệu, và quản lý dữ liệu từ điển cũng được coi là các thành phần cốt yếu của một hệ thống kho dữ liệu. Nhiều người sử dụng thuật ngữ "kho dữ liệu" với ngữ cảnh rộng hơn. Một định nghĩa mở rộng cho kho dữ liệu bao gồm cả các công cụ thông minh, các công cụ để trích, biến đổi và nạp dữ liệu vào kho, và các công cụ để quản lý và lấy siêu dữ liệu (meta data).
  + - 1. **Nguyên lý hình thành DWH**
* Nguyên lý hình thành database được thể hiện bởi hình sau:



* Cơ sở dữ liệu nguồn(Database, External Source): Một CSDL tác nghiệp, trung tâm dữ liệu, hệ thống sẵn có, dữ liệu từ nhiều nguồn làm đầu vào cho DWH
* Xác định hướng phân tích, chủ đề cần phân tích và tiến hành ETL(Extract, Transform, Load) để tích hợp dữ liệu từ nhiểu nguồn vào DWH(bottom tier)
* Sử dụng OLAP(OnLine Analytical Processing) để rút trích dữ liệu và triển khai thành nhiều khối dữ liệu(cubes) khác nhau tạo thành middle tier
* Sử dụng các công cụ front end để trực quan hóa, phân tích, tạo báo cáo từ các khối dữ liệu và tạo thành top tier
  + - 1. **Các đặc điểm cơ bản của DWH**
* Hướng chủ đề (subject-oriented)
* Được tích hợp (integrated)
* Có gán nhãn thời gian (time variant)
* Bất biến (non-volatile)
  + - 1. **Phân biệt [SS] DWH với CSDL truyền thống**

|  |  |  |
| --- | --- | --- |
| Chỉ tiêu so sánh | Cơ sở dữ liệu truyền thống | Kho dữ liệu |
| Mục đích | Để ghi và truy vấn dữ liệu | Để xử lý và phân tích dữ liệu |
| Chức năng | Hỗ trợ các hoạt động hàng ngày | Hỗ trợ quyết định mang tính chiến lược |
| Phương pháp xử lý | Cơ sở dữ liệu sử dụng Xử lý giao dịch trực tuyến (OLTP) | Kho dữ liệu sử dụng Xử lý phân tích trực tuyến (OLAP). |
| Các bảng và phép nối | Có độ phức tạp cao vì chúng được chuẩn hóa (cho RDMS) để giảm dữ liệu thừa, tối ưu hóa dung lượng lưu trữ | Bảng và phép nối rất dễ dàng trong kho dữ liệu vì chúng không được chuẩn hóa. |
| Tính chất dữ liệu | Chi tiết, được cập nhật thường xuyên | Có tính lịch sử và thống kê, được thêm mới chứ không cập nhật |
| Sử dụng | Thường xuyên | Trong những trường hợp đặc biệt |
| Đơn vị công việc | Đơn vị công việc | Các câu truy vấn phức tạp |
| Độ đo | Thông lượng giao dịch, có thể thực hiện nhiều giao dịch cùng một lúc | Thông lượng truy vấn và trả lời |
| Sự định hướng | Định hướng ứng dụng | Định hướng chủ đề |
| Mô hình sử dụng | Mô hình quan hệ – thực thể | Mô hình dữ liệu đa chiều |
| Loại truy vấn | Những truy vấn giao dịch đơn giản được sử dụng. | Những truy vấn phức tạp được áp dụng cho mục đích phân tích. |
| Hiệu suất truy vấn phân tích | Thấp | Cao |
| Lưu trữ dữ liệu | Phương pháp tiếp cận quan hệ phẳng, nhiều dữ liệu khác nhau được tích hợp vào một nguồn | Phương pháp tiếp cận đa chiều và chuẩn hóa, nhiều nguồn dữ liệu khác nhau được tích hợp và định dạng lại |

Kho dữ liệu có kích thước lớn hơn nhiều so với cơ sở dữ liệu, thậm chí nó có thể bao gồm cả cơ sở dữ liệu trong đó, hay nói cách khác kho dữ liệu là một loại cơ sở dữ liệu khổng lồ được thiết kế để tối ưu hóa quy trình phân tích và báo cáo. Cơ sở dữ liệu thường là một ứng dụng, chương trình hoặc hệ thống để chứa các thông tin trong một nguồn. Kho dữ liệu là tập hơn các nguồn, hệ thống thông tin khác nhau để sắp xếp, phân tích và xuất báo cáo theo truy vấn người dùng. Điều quan trọng cần lưu ý là kho dữ liệu có thể được lấy từ nhiều cơ sở dữ liệu khác nhau (hoặc có thể không lấy).

### Kiến trúc cơ bản của kho dữ liệu:

#### Tầng đáy: Nạp dữ liệu

Thu thập thông tin, hình thành DB: Ví dụ như nguồn DB của một DWH về y tế xuất phát từ những hóa đơn khám chữa bệnh của bệnh nhân, những file excel nhập thông tin bệnh nhân, các loại thuốc, các file word,..

#### Tầng giữa: OLAP (OLAP server): xử lý DWH

* Duy trì hoạt động của DWH, xử lí, khai thác sử dụng, khai thác nhờ vào vấn tinh phân tích là chính.
* Các nhóm vấn tin phân tích gồm 3 nhóm: Subtotal, Analytic, Materialize View.
* OLAP (OLAP server) dạng:
  + Relational OLAP
  + Multidimentional OLAP
  + Hybrid OLAP
  + Minh họa: Từ DWH về y tế trên phân tích trích xuất ra các nhóm bệnh phổ biến, các phương pháp điều trị tích cực, độ hài lòng của bệnh nhân,...

#### Tầng trên cùng: ứng dụng = khai thác sử dụng DWH

Cung cấp thông tin sử dụng cho người dùng: Hiển thị ra thông tin các biểu đồ, các thông số, theo yêu cầu của người sử dụng

### Mô hình CSDL đa chiều (Dimensional Modeling)

* Một cơ sở dữ liệu đa chiều là một loại hình cụ thể của cơ sở dữ liệu đã được tối ưu hóa cho kho dữ liệu và OLAP (xử lý phân tích trực tuyến). Một cơ sở dữ liệu đa chiều được cấu trúc bởi sự kết hợp dữ liệu từ các nguồn khác nhau mà công việc giữa các cơ sở dữ liệu đồng thời và rằng các mạng cung cấp, phân cấp, mảng, và các dữ liệu khác định dạng phương pháp. Trong một cơ sở dữ liệu đa chiều, các dữ liệu được trình bày cho người dùng của mình thông qua các mảng đa chiều, và mỗi giá trị riêng biệt của dữ liệu được chứa trong một tế bào có thể được truy cập bởi nhiều chỉ số.
* Mô hình CSDL đa chiều là mở rộng của mô hình CDSL quan hệ theo hai hay nhiều chiều hơn nhằm tăng tốc độ xử lý và sử dụng cho mục đích xây dựng các quan hệ phức tạp giữa nhiều nguồn dữ liệu.

## Phân tích và thiết kế Kho dữ liệu

### Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling)

Phân tích và thiết kế kho dữ liệu bao gồm việc kết hợp dữ liệu không đồng nhất trong các nguồn khác nhau vào một lược đồ duy nhất và có thể truy vấn, cung cấp cho người dùng một cái nhìn thống nhất về chúng.

### Các dạng (phương pháp) tích hợp dữ liệu

- Ghép nối chặt chẽ (Tight Coupling): kho dữ liệu

Phương pháp ghép nối chặt chẽ thường được thực hiện thông qua kho dữ liệu, dữ liệu được lấy từ nhiều nguồn khác nhau đưa vào một vị trí vật lý duy nhất thông qua quá trình ETL (Extraction, Transformation, Loading). Lớp ETL giúp ánh xạ dữ liệu từ các nguồn để cung cấp một kho dữ liệu thống nhất. Vị trí vật lý, cung cấp một giao diện đồng nhất để truy vấn dữ liệu.Cách tiếp cận này được gọi là ghép nối chặt chẽ vì trong cách tiếp cận này dữ liệu được kết hợp chặt chẽ với kho lưu trữ vật lý tại thời điểm truy vấn.

- Ghép nối lỏng lẻo (Loose Coupling): lược đồ trung gian ảo:

Ở đây một lược đồ trung gian ảo cung cấp một giao diện nhận truy vấn từ người dùng, biến đổi nó theo cách mà cơ sở dữ liệu nguồn có thể hiểu và gửi truy vấn trực tiếp tới cơ sở dữ liệu nguồn để thu được kết quả. Trong phương pháp này, dữ liệu chỉ nằm trong cơ sở dữ liệu nguồn thực tế. Mô hình của phương pháp ghép nối lỏng lẻo được mô phỏng như hình dưới.

### Các bước cơ bản trong tiến trình tích hợp dữ liệu

* Bước 1: Trích xuất dữ liệu = thu thập dữ liệu lớn từ nhiều nguồn
* Minh họa: Trích từ csdl bệnh nhân, các bản chép tay, word, excel
* Bước 2: Làm sạch = tìm lỗi và sửa sai. Khắc phục các giá trị NULL, xử li các giá trị trống trong database. Việc trống quá nhiều dữ liệu dẫn đến thông tin bị lệch, dự đoán dự báo bị sai, dẫn đến hệ thống dự báo dự đoán sai
* Minh họa: tập dữ liệu về thông tin các nạn nhân trong một trận sóng thần thì có rất nhiều thông tin thiếu xót bằng việc lấy giá trị max để thay cho giá trị thiếu, lấy hồi qui hoặc xóa bỏ bộ dữ liệu đó nếu không ảnh hưởng quá nhiều là các cách để xử lí những giá trị trống
* Bước 3: So khớp mẫu = Lọc và chuyển đổi dữ liệu = chuyển đổi sang định dạng kho
* Minh họa: chuyển dữ liệu từ dạng xml, json sang dạng có cấu trúc như SQL server
* Bước 4: Lưu trữ dữ liệu theo cấu trúc = sắp xếp, hợp nhất và tổng hợp,..

### Các công cụ tích hợp dữ liệu:

Bao gồm các công cụ và các dịch vụ có liên quan đến SQL

### Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows)

* **Tích hợp dữ liệu** là quá trình kết hợp **dữ liệu** từ nhiều nguồn khác nhau để trích xuất giá trị bổ sung. Quá trình này thường bao gồm **các hành động** như tham gia, làm sạch, xác thực và làm phong phú **dữ liệu** trong quá trình thực hiện.
* Data Integration Workflows cho phép mọi nhóm sử dụng các ứng dụng phù hợp nhất với mục đích, bao gồm cơ sở dữ liệu, kỹ thuật, hoạt động và tài chính, trong khi vẫn giữ mọi thứ được kết nối. ... **Dữ liệu** được tự động xác nhận và tải vào mô hình **dữ liệu** của kho để nhóm GIS sử dụng
* Data Integration Workflows của chúng ta sẽ được thực hiện như sau:
  + Thu thập các yêu cầu kinh doanh
  + Xác định nhu cầu về dữ liệu và chất lượng
  + Lập hồ sơ dữ liệu hoặc hiểu các nguồn dữ liệu và chất lượng liên quan cả trong hệ thống nguồn và trên nhiều hệ thống nguồn, nếu có
  + Thực hiện đánh giá chất lượng dữ liệu dựa trên các chỉ số mà doanh nghiệp đã yêu cầu
  + Xác định khoảng cách giữa dữ liệu có sẵn và chất lượng của dữ liệu đó so với những gì doanh nghiệp đã yêu cầu
  + Sửa đổi kỳ vọng kinh doanh hoặc chi phí dự án và xác định giải pháp dữ liệu đã chọn
  + Mô hình hóa các kho dữ liệu cần thiết - các khu vực tổ chức, kho dữ liệu, kho dữ liệu hoạt động và (các) kho dữ liệu - cả từ góc độ logic (để xác nhận các yêu cầu kinh doanh) và góc độ vật lý (để cho phép thực hiện).
  + Xác định dữ liệu và nguồn phù hợp từ các hoạt động trên.
  + Chuẩn bị dữ liệu
  + Nhượng quyền dữ liệu
  + Quản lý dữ liệu và metadata

### Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến)

* OLTP viết tắt từ **On-line transactional processing** có nghĩa là xử lý giao dịch trực tuyến. Các hệ thống OLTP là các hệ thống cổ điển chúng xử lý dữ liệu giao dịch ngắn xung quanh chúng ta.
* OLTP có liên quan đến hoạt động của một hệ thống cụ thể. OLTP được đặc trưng bởi một số lượng lớn các giao dịch trực tuyến ngắn (XÁC NHẬN, CẬP NHẬT, XÓA). Điểm nhấn chính cho các hệ thống OLTP được xử lý truy vấn rất nhanh, duy trì tính toàn vẹn dữ liệu trong môi trường đa truy cập và hiệu quả được đo bằng số lượng giao dịch mỗi giây. Trong cơ sở dữ liệu OLTP có dữ liệu chi tiết và hiện tại và lược đồ được sử dụng để lưu trữ cơ sở dữ liệu giao dịch là mô hình thực thể (thường là 3NF). Nó liên quan đến Truy vấn truy cập hồ sơ cá nhân như Cập nhật Email của bạn trong cơ sở dữ liệu Công ty.

## Vấn tin khai thác sử dụng Kho dữ liệu

### Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu

* Các mở rộng của các hệ quản trị cơ sở dữ liệu để hỗ trợ cho kho dữ liệu
  + Mệnh đề SQL CUBE
  + Mệnh đề SQL ROLLUP
  + Mệnh đề SQL GROUPING SETS
  + Kết hợp các phép toán SUBTOTAL
* Các vấn tin này phục vụ mục đích phân tích nhất định và luôn có cấu trúc rất phức tạp, tuy nhiên tốc độ xử lý sẽ nhanh hơn câu truy vấn thường.

### Vấn tin với các hàm phân tích của SQL trên kho dữ liệu

* SQL Analytic
  + Mô hình xử lí và cú pháp cơ bản
  + Cú pháp mở rộng và các hàm xếp hạng
  + So sánh cửa sổ (Window Comparisons)
  + Các hàm so sánh tỷ lệ

### Vấn tin dùng Materialized View để phân tích trên kho dữ liệu

Phân biệt : Traditional View và Materialized View

|  |  |
| --- | --- |
| Vấn tin truyền thống | Vấn tin phân tích |
| * Chủ yếu là hỏi đáp, sử dụng các câu lệnh như: Select..from…where * Vấn tin trên cơ sở dữ liệu | * Là vấn tin trên kho dữ liệu, phân tích, rút ra kết luận nào đó từ kho dữ liệu, từ DWH: SQL Subtotal, SQL Analytic, Materialized View * Phân tích để xác định các qui luật có sẵn trong DWH, chứ không đơn giản chỉ là hỏi đáp thông thường |
| * Ví dụ: Lấy danh sách các mặt hàng thuộc thương hiệu Xmen có trong CSDL của một cữa hàng tiện lợi ( các loại hỏi đáp) | * Ví dụ: từ DWH thông tin khách hàng mua hàng của một tiệm bán điện thoại dùng vấn tin phân tích cho biết qui luật mua hàng của khách hàng * Nếu khách hàng đã mua điện thoại A thì có xu hướng sẽ mua kèm theo loại linh kiện nào và xác suất là bao nhiêu % |

### Hệ thống OLAP

OLAP (Xử lý phân tích trực tuyến) liên quan đến Dữ liệu lịch sử hoặc Dữ liệu lưu trữ. OLAP được đặc trưng bởi khối lượng giao dịch tương đối thấp. Truy vấn thường rất phức tạp và liên quan đến tập hợp. Đối với OLAP hệ thống, thời gian phản hồi là thước đo hiệu quả. Các ứng dụng OLAP được sử dụng rộng rãi bởi các kỹ thuật Khai thác dữ liệu. Trong cơ sở dữ liệu OLAP có dữ liệu lịch sử tổng hợp, được lưu trữ trong các lược đồ đa chiều (thường là lược đồ sao). Thỉnh thoảng truy vấn cần truy cập một lượng lớn dữ liệu trong hồ sơ Quản lý như lợi nhuận của công ty bạn trong năm ngoái.

# Chương 3. TỔNG QUAN VỀ KHO DỮ LIỆU PHÂN TÍCH VÀ THIẾT KẾ KHO DỮ LIỆU LƯƠNG THỰC VÀ NÔNG NGHIỆP TẠI CÁC CƠ QUAN QUẢN LÍ VÀ KINH DOANH LƯƠNG THỰC VÀ NÔNG NGHIỆP

## Ý tưởng hình thành DWH phân tích và thiết kế kho dữ liệu lương thực và nông nghiệp tại các cơ quan quản lí và kinh doanh lương thực và nông nghiệp

* Tổ chức Nông nghiệp lương thực (FAO) là một cơ quan chuyên môn của Liên hợp quốc, lãnh đạo các nỗ lực quốc tế nhằm đánh bại nạn đói. Mục tiêu của tổ chức là đạt được an ninh lương thực cho tất cả mọi người và đảm bảo rằng mọi người được tiếp cận thường xuyên với đủ thực phẩm chất lượng cao để có cuộc sống năng động và khỏe mạnh.
* Với mong muốn góp một phần công sức vào việc chấm dứt nạn đói, phát triển nông nghiệp cũng như giảm thiểu ô nhiễm môi trường do nông nghiệp, chúng ta sẽ tổ chức một hệ thống kho dữ liệu dựa vào dữ liệu của hơn 194 quốc gia trên thế giới(là thành viên của FAO) nhằm góp phần phân tích và đưa ra các kết luận hữu ích cho các cơ quan quản lý, kinh doanh lương thực nông nghiệp để giúp họ đưa ra các hoạch định chính xác trong việc quản lý các hoạt động liên quan đến nông/lâm nghiệp.
* Tất cả nguồn dữ liệu được lấy từ trang web chính thức của FAO: http://www.fao.org/

## Giới thiệu chung về Kho dữ liệu

### Mô tả tổng quan về DWH

* Kho dữ liệu sẽ được tổ chức trên nền tảng SQL Server và sử dụng công cụ phân tích là SQL Server Analysis Service.
* Ban đầu, chúng ta sẽ tổ chức một database để lưu trữ một cách logic các dữ liệu mà chúng ta thu thập được sau đó triển khai quá trình tích hợp dữ liệu vào kho dữ liệu và phát triển các ứng dụng nhằm khai thác sử dụng kho dữ liệu.

### Các dữ liệu nguồn hình thành DWH

* Nguồn dữ liệu của chúng ta sẽ bao gồm các file excel đơn lẻ và rời rạc, không liên kết với nhau được thu thập từ trang web của tổ chức FAO.
* Chúng ta sẽ phải xử lý để làm sạch nguồn dữ liệu này và tiến hành tổ chức một CSDL quan hệ để lưu trữ chúng.

## Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH

* Các tổ chức kinh doanh trong lĩnh vực nông nghiệp cần biết rằng họ nên đầu tư vào việc xuất khẩu hàng hóa của họ đi đâu, lĩnh vực nào đang thiếu hụt hoặc khu vực nào đang cần đến mặt hàng của họ vì thế phân tích về giá cả thị trường nông nghiệp ở các nước trên thế giới sẽ đưa ra một cái nhìn tổng quan hơn về thị trường và giúp họ đưa ra các quyết định kinh doanh đúng đắn.
* Các cơ quan quản lý nông nghiệp cũng như môi trường sẽ quan tâm đến các lĩnh vực khí thải hoặc lâm nghiệp để hoạch định các chính sách thích hợp để phục hồi rừng hoặc giảm thiểu ô nhiễm môi trường

## Các thành phần phân hệ Kho dữ liệu phân công các cá nhân phụ trách

* Nguyễn Ngọc Hòa [11]\_Nhóm trưởng: Phân hệ phân tích kho dữ liệu Khí thải nông nghiệp
* Lê Đình Khang [18]: Phân hệ phân tích kho dữ liệu Giá nông sản
* Lương Uy Long [22]: Phân hệ phân tích kho dữ liệu Độ che phủ đất
* Trần Gia Nguyên [29]: Phân hệ phân tích kho dữ liệu Cây trồng
* Trần Như Thuận[44]: Phân hệ phân tích kho dữ liệu Lâm nghiệp

**Chương 4. PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC PHÂN HỆ KHO DỮ LIỆU LƯƠNG THỰC VÀ NÔNG NGHIỆP TẠI CÁC CƠ QUAN QUẢN LÍ VÀ KINH DOANH LƯƠNG THỰC VÀ NÔNG NGHIỆP**

* 1. 29\_ Trần Gia Nguyên\_Phân hệ phân tích kho dữ liệu Cây trồng

### Mô tả tổng quan về phân hệ DWH Phân hệ phân tích kho dữ liệu Cây trồng

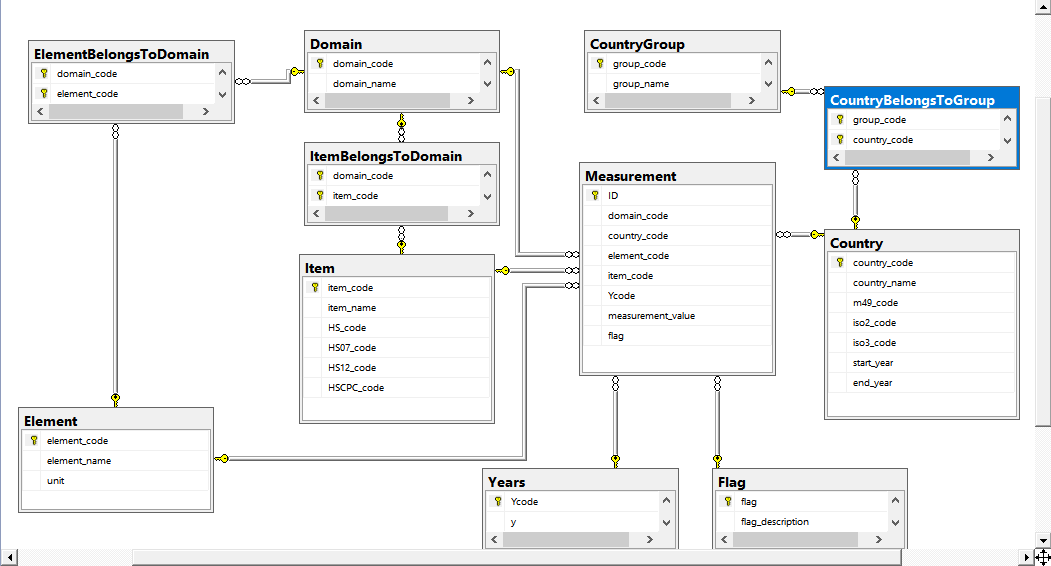
Vấn đề đang gặp phải của rất nhiều nông dân, những người kinh doanh, sản xuất nông sản hiện tại là dữ liệu, kể cả khi có dữ liệu rồi cũng không không thể nào làm cho dữ liệu có ý nghĩa. Vậy dữ liệu có ý nghĩa khi nào. Khi nó được phân tích, áp dụng vào thực tế, giúp người trồng trọt, người thu mua, người kinh doanh đạt được nhiều lợi ích hơn.

Kho dữ liệu Cây trồng này sẽ cho một cái nhìn toàn cục về sản lượng của cây trồng trên thế giới. Kết hợp so sánh với giá, qua đó đưa ra một vài nhận định về mối tương quan giữa sản lượng và giá. Hoặc từ nhu cầu của thị trường về loại nông sản mà có thể đưa ra gợi ý thay đổi sản lượng cây trồng. Xem xét xu hướng tăng giảm sản lượng theo element (loại đơn vị) và so sánh chúng với nhu cầu thị trường. So sánh sản lượng các loại cây trên 1 đất nước. Để có thể tập trung chú ý vào những loại cây có nhu cầu trong nước nhưng sản lượng thấp.

Thống kê sản lượng các loại cây trồng được trên Châu Á, thống kê sản lượng của chúng theo năm, theo khu vực, theo đất nước, theo thời gian.

### Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu

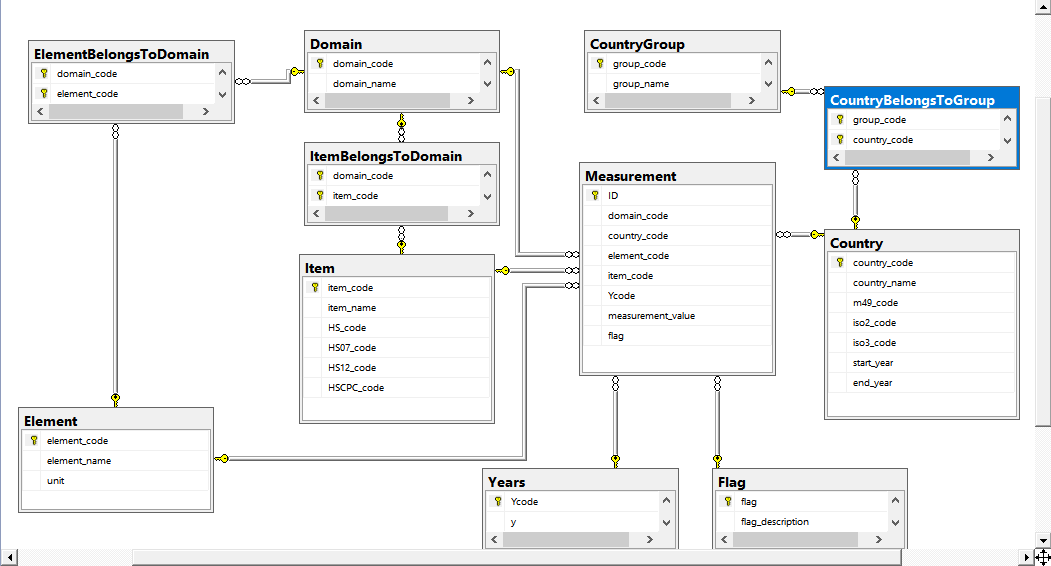
Dữ liệu nguồn là FAOSTAT, được lấy ở tổ chức Lương thực và Nông nghiệp Liên Hợp Quốc



### Xác định ý tưởng hình thành phân hệ của Kho dữ liệu

Ý tưởng hình thành phân hệ: Chủ yếu phân tích các số liệu liên quan tới cây trồng như Sản lượng cây trồng.

### Mô tả hệ thống các DB gốc liên quan phân hệ DWH



Hình 1: Diagram Database gốc

**\* Hệ thống các DB gốc liên quan :**

Domain : Lĩnh vực được thống kê

Item : Mục (Chứa tên các loại cây trồng được thống kê)

Element : Yếu tố (diện tích thu hoạch, số lượng sản xuất và năng suất)

Country : Tên nước (Cụ thể gồm : South Africa, South Sudan, Spain, Svalbard and Jan Mayen Islands, Tajikistan, Thailand, VietNam, …)

CountryGroup : Các vùng quốc gia, châu lục hoặc các tổ chức

Years : Năm

Flag : Cờ chỉ nguồn dữ liệu

Measurement : Các chỉ số được thống kê

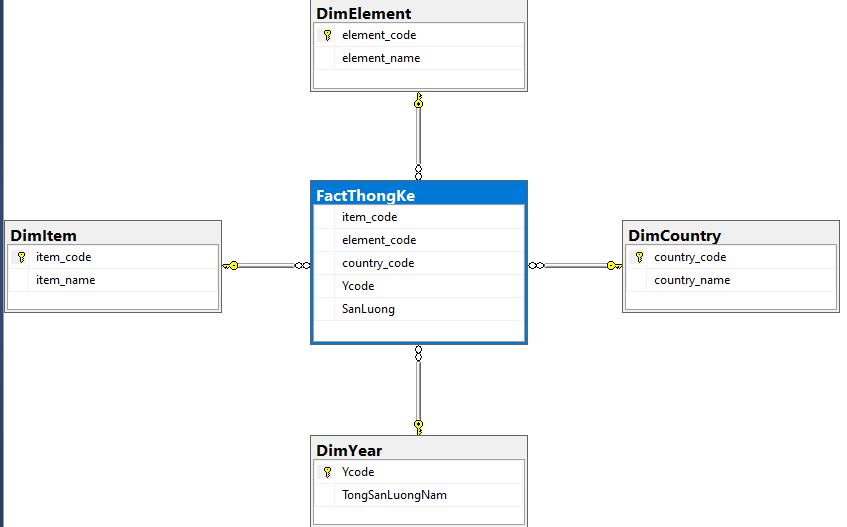
CountryBelongsToGroup : Thể hiện mối quan hệ M-N của country và CountryGroup

ItemBelongsToDomain : Thể hiện mối quan hệ M-N của Item và Domain

ElementBelongsToDomain : Thể hiện mối quan hệ M-N của Element và Domain

### Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ

Xác định các Facts và Dims cho DWH của phân hệ cá nhân:



Hình 2: Diagram DataWarehouse phân hệ cá nhân

**Fact :**

* FactThongKe : tính tổng sản lượng từng loại cây trồng

**Dim :**

* DimCountry : chứa tên nước và mã nước
* DimItem : chứa tên cây trồng, mã cây trồng
* DimElement : chứa mã yếu tố và tên yếu tố
* DimYear : chứa mã năm và số năm
* DimThongKeSanLuongTheoNam : chứa mã năm và tổng sản lượng theo năm
* DimItemValue : Chứa tổng sản lượng theo mã tên cây trồng, mã yếu tố, mã nước

|  |
| --- |
| USE [FAOSTAT\_CayTrong\_DWH]  GO  /\*\*\*\*\*\* Object: Table [dbo].[DimCountry] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[DimCountry](  [country\_code] [int] NOT NULL,  [country\_name] [varchar](70) NULL,  CONSTRAINT [PK\_\_DimCount\_\_3436E9A4A8702CAA] PRIMARY KEY CLUSTERED  (  [country\_code] ASC  )WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]  ) ON [PRIMARY]  GO  /\*\*\*\*\*\* Object: Table [dbo].[DimElement] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[DimElement](  [element\_code] [int] NOT NULL,  [element\_name] [varchar](100) NULL,  CONSTRAINT [PK\_\_DimEleme\_\_36BE77F3D77B44DF] PRIMARY KEY CLUSTERED  (  [element\_code] ASC  )WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]  ) ON [PRIMARY]  GO  /\*\*\*\*\*\* Object: Table [dbo].[DimItem] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[DimItem](  [item\_code] [int] NOT NULL,  [item\_name] [varchar](150) NULL,  CONSTRAINT [PK\_\_DimItem\_\_4A67201F2AEC1522] PRIMARY KEY CLUSTERED  (  [item\_code] ASC  )WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]  ) ON [PRIMARY]  GO  /\*\*\*\*\*\* Object: Table [dbo].[DimItemValue] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[DimItemValue](  [item\_code] [int] NOT NULL,  [Ycode] [int] NULL,  [measurement\_value] [money] NULL,  [element\_code] [int] NULL,  [country\_code] [int] NULL,  CONSTRAINT [PK\_\_DimItemV\_\_4A67201F5947C4E1] PRIMARY KEY CLUSTERED  (  [item\_code] ASC  )WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]  ) ON [PRIMARY]  GO  /\*\*\*\*\*\* Object: Table [dbo].[DimYear] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[DimYear](  [Ycode] [int] NOT NULL,  [TongSanLuongNam] [money] NULL,  PRIMARY KEY CLUSTERED  (  [Ycode] ASC  )WITH (PAD\_INDEX = OFF, STATISTICS\_NORECOMPUTE = OFF, IGNORE\_DUP\_KEY = OFF, ALLOW\_ROW\_LOCKS = ON, ALLOW\_PAGE\_LOCKS = ON, OPTIMIZE\_FOR\_SEQUENTIAL\_KEY = OFF) ON [PRIMARY]  ) ON [PRIMARY]  GO  /\*\*\*\*\*\* Object: Table [dbo].[FactThongKe] Script Date: 5/21/2021 12:19:13 AM \*\*\*\*\*\*/  SET ANSI\_NULLS ON  GO  SET QUOTED\_IDENTIFIER ON  GO  CREATE TABLE [dbo].[FactThongKe](  [item\_code] [int] NULL,  [element\_code] [int] NULL,  [country\_code] [int] NULL,  [Ycode] [int] NULL,  [SanLuong] [money] NULL  ) ON [PRIMARY]  GO  ALTER TABLE [dbo].[FactThongKe] WITH CHECK ADD FOREIGN KEY([country\_code])  REFERENCES [dbo].[DimCountry] ([country\_code])  GO  ALTER TABLE [dbo].[FactThongKe] WITH CHECK ADD FOREIGN KEY([element\_code])  REFERENCES [dbo].[DimElement] ([element\_code])  GO  ALTER TABLE [dbo].[FactThongKe] WITH CHECK ADD FOREIGN KEY([item\_code])  REFERENCES [dbo].[DimItem] ([item\_code])  GO  ALTER TABLE [dbo].[FactThongKe] WITH CHECK ADD FOREIGN KEY([Ycode])  REFERENCES [dbo].[DimYear] ([Ycode])  GO |

**Xác định các Dim và Fact:**

**Factor…**

* FactThongKe=SUM(Measurement.measurement\_value): tính tổng sản lượng cây trồng

**Dim…**

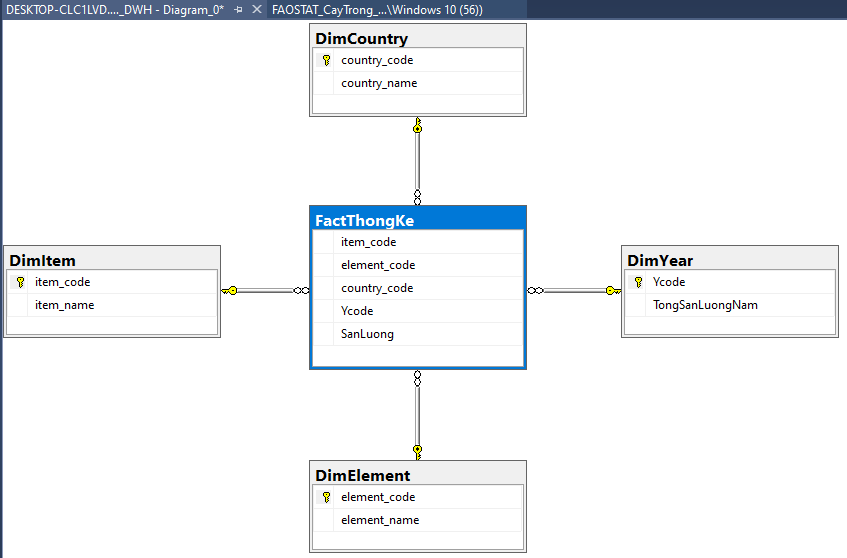
* DimItem: Loại cây trồng
* DimElement: Loại đơn vị
* DimYear: Năm (năm trùng mã năm)
* DimCountry: Quốc gia

### Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ

Thao khảo Buổi: 6 và 4,5

theo yêu đã xác định trong 3

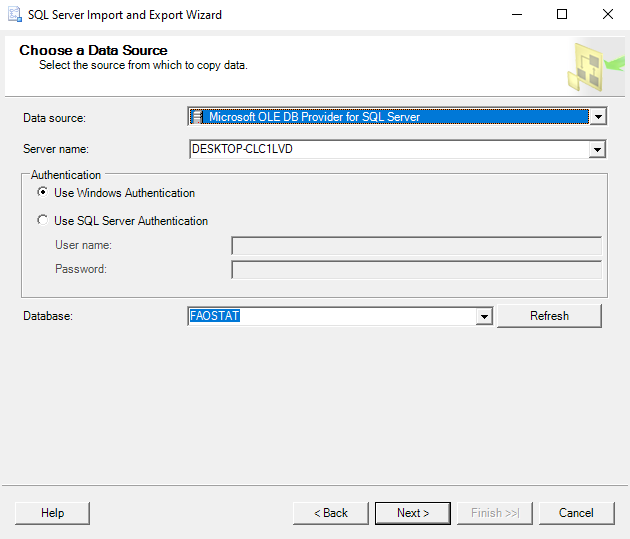
**FAOSTAT = lập DWH ==> FAOSTAT\_Caytrong\_DWH**

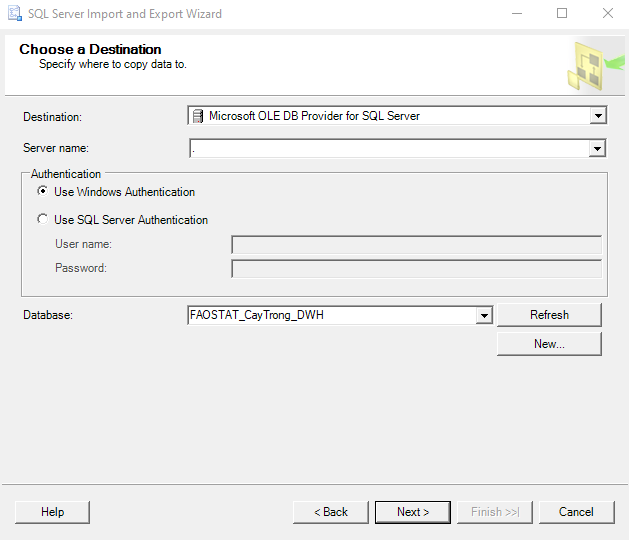


Hình 3: Diagram FAOSTAT\_Caytrong\_DWH

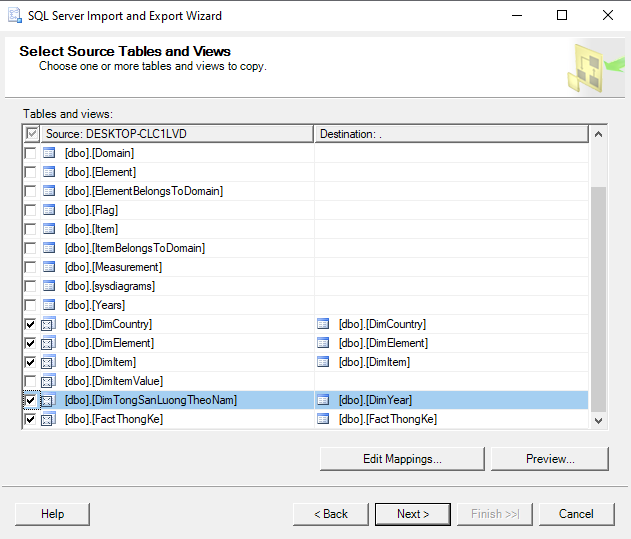
Phân tích + thiết kế DWH: **FAOSTAT\_Caytrong\_DWH**

### Thực hiện các thủ tục tích hợp các DB gốc vào phân hệ DWH

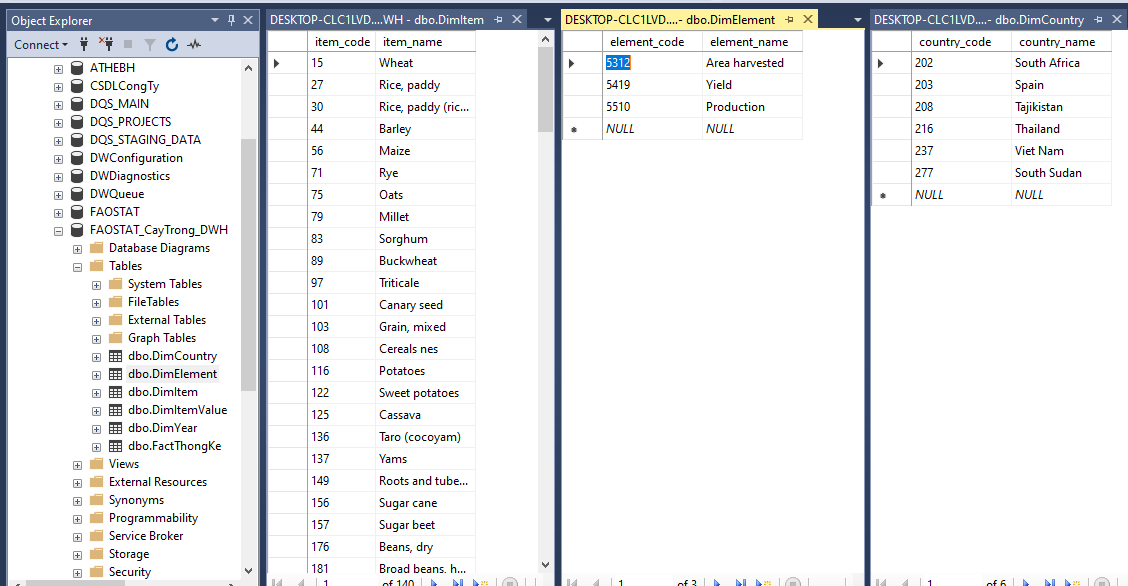
****

****

Tại **Specify Table Copy or Query** chọn **Copy data from onr or more tables or views**



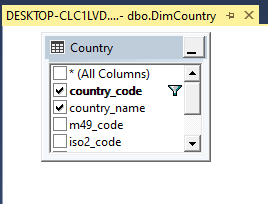
**Kết quả:**

****

### Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH

* DimCountry :

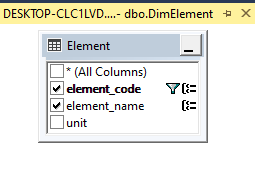
|  |
| --- |
| SELECT country\_code, country\_name  FROM dbo.Country  WHERE (country\_code = 202) OR  (country\_code = 277) OR  (country\_code = 203) OR  (country\_code = 208) OR  (country\_code = 216) OR  (country\_code = 237) |



Hình 4: Tạo dbo.DimCountry

* DimElement

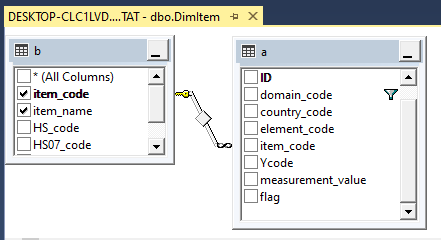
|  |
| --- |
| SELECT DISTINCT element\_code, element\_name  FROM dbo.Element  WHERE (element\_code = 5419) OR  (element\_code = 5510) OR  (element\_code = 5312)  GROUP BY element\_code, element\_name |



Hình 5: Tạo dbo.DimElement

* DimItem

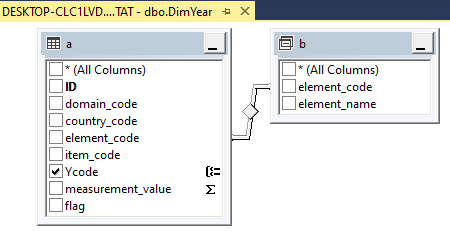
|  |
| --- |
| SELECT DISTINCT b.item\_code, b.item\_name  FROM dbo.Measurement AS a INNER JOIN  dbo.Item AS b ON a.item\_code = b.item\_code  WHERE (a.domain\_code = 'QC') |



Hình 6: Tạo dbo.DimItem

* DimYear

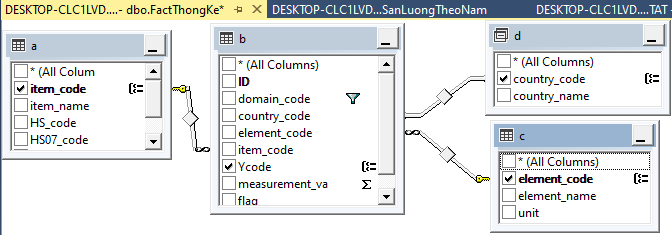
|  |
| --- |
| SELECT TOP (100) PERCENT a.Ycode, SUM(a.measurement\_value) AS TongSanLuongNam  FROM dbo.Measurement AS a INNER JOIN  dbo.DimElement AS b ON a.element\_code = b.element\_code  GROUP BY a.Ycode |



Hình 7: Tạo dbo.DimYear

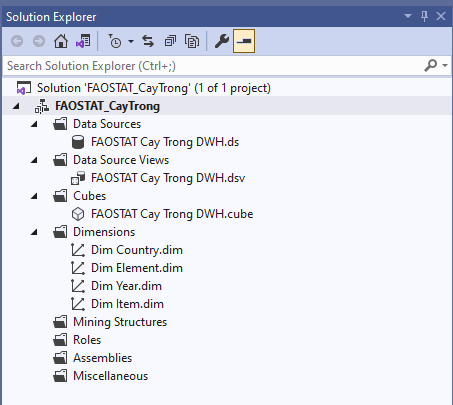
* FactThongKe

|  |
| --- |
| SELECT TOP (100) PERCENT a.item\_code, c.element\_code, d.country\_code, b.Ycode, SUM(b.measurement\_value) AS SanLuong  FROM dbo.Item AS a INNER JOIN  dbo.Measurement AS b ON a.item\_code = b.item\_code INNER JOIN  dbo.Element AS c ON b.element\_code = c.element\_code INNER JOIN  dbo.DimCountry AS d ON b.country\_code = d.country\_code  WHERE (b.domain\_code = 'QC')  GROUP BY a.item\_code, c.element\_code, d.country\_code, b.Ycode |

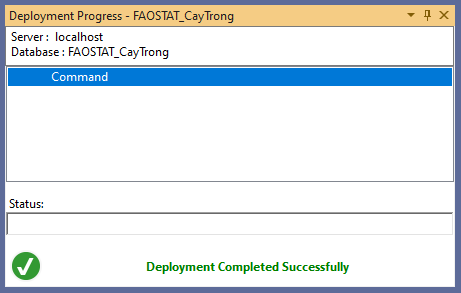


Hình 8: Tạo Fact

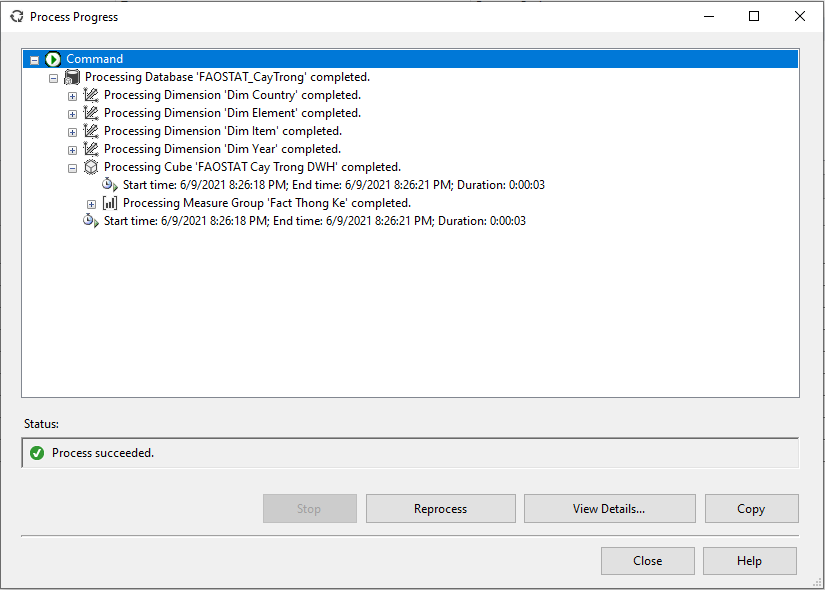
### Triển khai phân hệ DWH lên BIDS



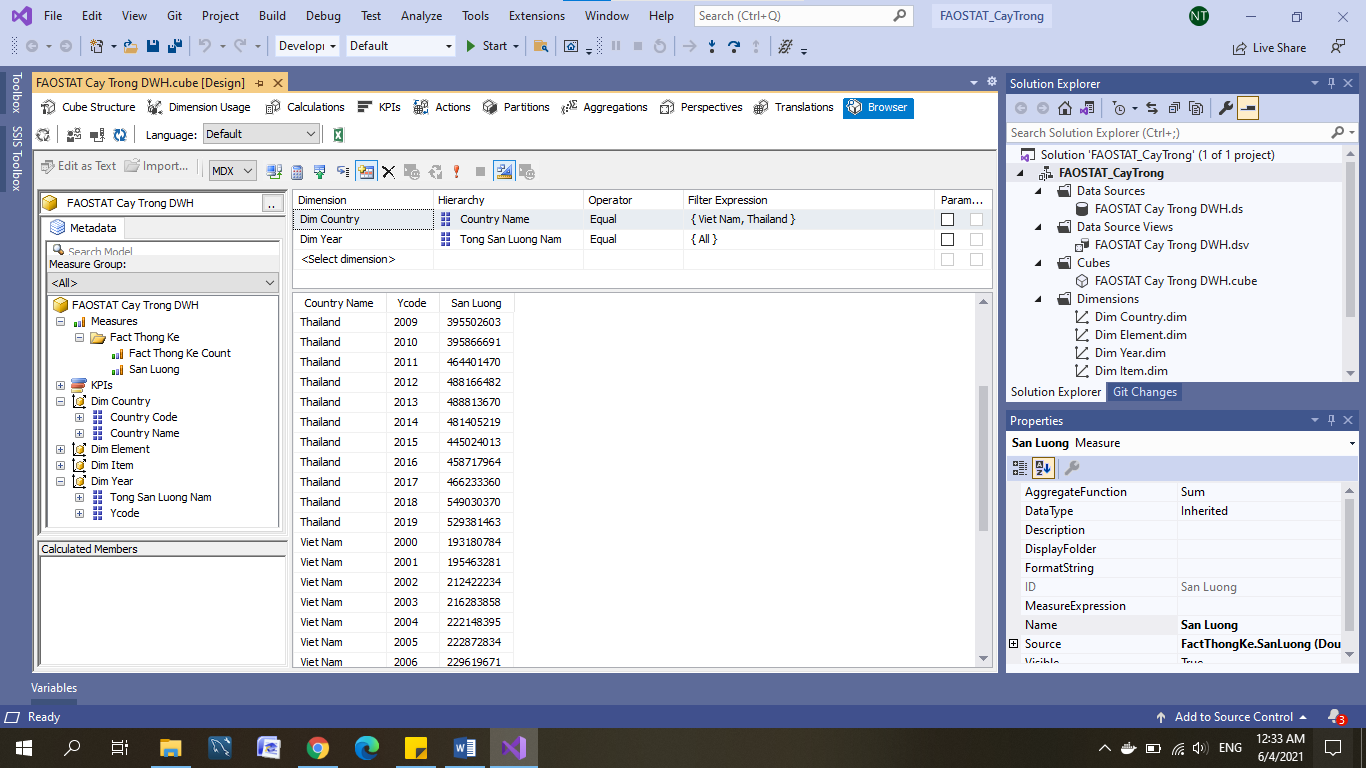
Hình 9: Solution explorer



Hình 10: Deploy thành công

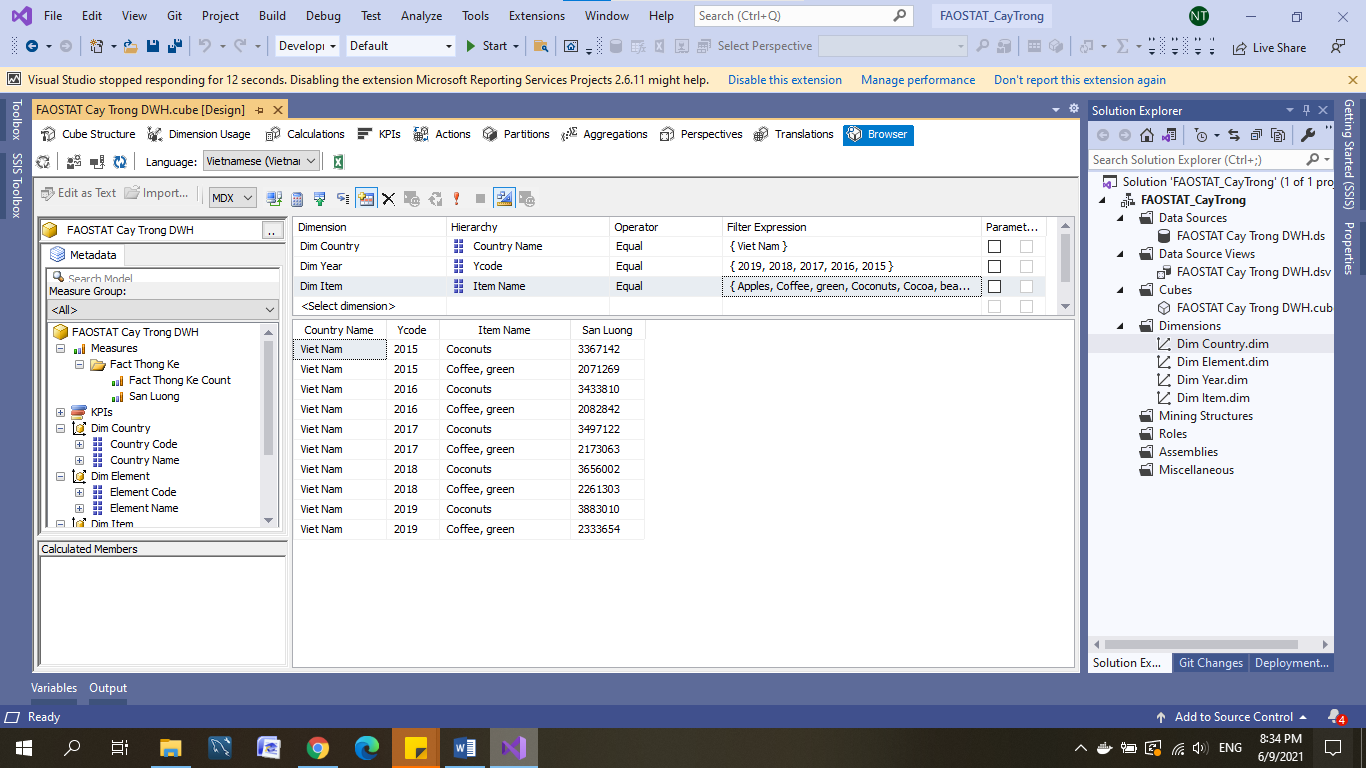


Hình 11: Process thành công



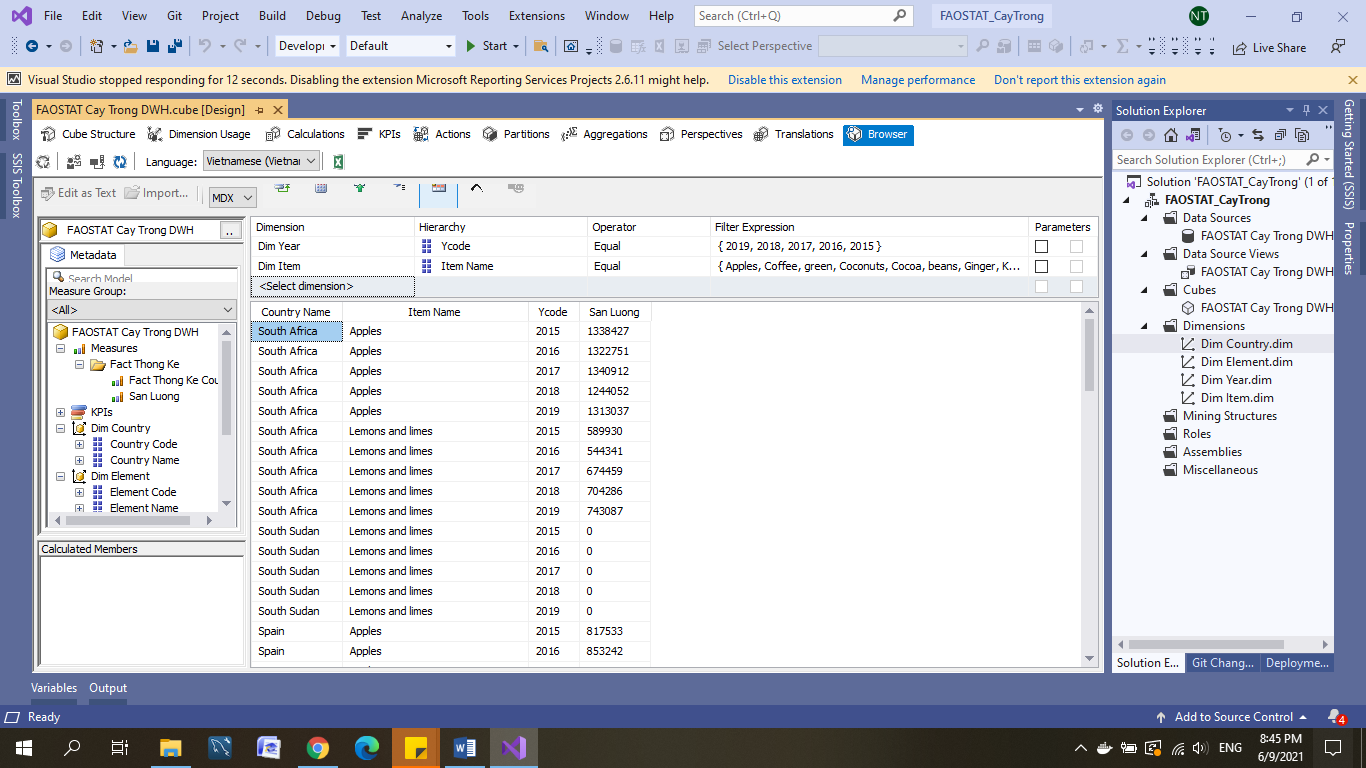
Hình 12: Phân tích sản lượng theo quốc gia – năm

Phân tích ở hình 4 cho thấy Sản lượng của từng quốc gia trong từng năm ở giai đoạn 2000-2019.

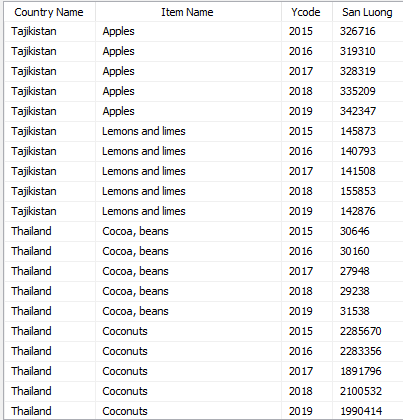


Hình 13: Phân tích sản lượng 1 vài loại cây trồng tại Việt Nam giai đoạn 2015-2019

Chọn ngẫu nhiên vài loại cây trồng tại Việt Nam trong khoảng năm 2015-1029, nhưng kết quả hiển thị chỉ có 2 loại cây. Điều này cho thấy dù nước ta là nước cận xích đạo, có khí hậu phù hợp với nhiều loại cây, nhưng vẫn còn rất nhiều loại cây chưa được trồng nhiều như Táo, ca cao, đậu,… Số liệu này có vẻ đúng bởi táo nước ta đang ăn hiện nay chủ yếu nhập khẩu từ các nước như Trung quốc, Mỹ, … .



Hình 14: Phân tích sản lượng 1 vài loại cây trồng tại 5 nước trong giai đoạn 2015-2019

Các loại cây trong phân tích này gồm: Táo, ca cao và đậu, dừa, cà phê và rau xanh, gừng, kiwi, chanh.

Tại South Africa:

* Sản lượng Táo những năm gần đây có xu hướng giảm nhẹ.
* Sản lượng chanh thì ngược lại, có sự tăng nhẹ
* Không có số liệu về ca cao và đậu, dừa, cà phê và rau xanh, gừng, kiwi.

Tại South Sudan:

* Sản lượng chanh là 0. Có thể xem xét nhập khẩu chanh từ quốc gia có sản lượng chanh cao vào South Sudan
* Không có số liệu về Táo, ca cao và đậu, dừa, cà phê và rau xanh, gừng, kiwi.

Tại Spain:

* Sản lượng táo và Kiwi có xu hướng tăng nhẹ, nhưng với tốc độ này thì có lẽ còn lâu mới đuổi kịp sản lượng táo của South Africa
* Chanh giữ mức ổn định trong suốt 5 năm
* Không có số liệu về ca cao và đậu, dừa, gừng.

Tajikistan:

* Táo và chanh giữ mức ổn định trong suốt 5 năm, trong đó sản lượng táo gần như gấp đôi chanh.
* Không có số liệu về ca cao và đậu, dừa, cà phê và rau xanh, gừng, kiwi.

Thái Lan

* Ca cao và đậu, dừa, gừng, chanh giữ mức sản lượng ổn định trong 5 năm liền.
* Không có số liệu về Táo, cà phê và rau xanh, kiwi.

Việt Nam

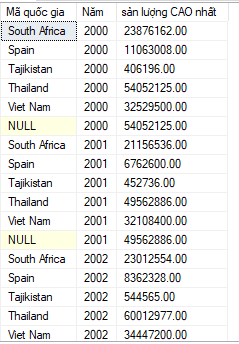
* Dừa và Cà phê đều có xu hướng tăng nhẹ

Chọn ngẫu nhiên 7 loại cây trồng, thì đa số các nước đều có khoảng 2-3 loại cây trong số này. Có những nước không có sự tương đồng về loại cây trồng. Chủ yếu sản lượng cây trong 5 năm gần đây đều tăng nhẹ, ít có giảm, và gần như không có loại cây nào có sản lượng tăng cao đột ngột.

### Khai thác sử dụng phân hệ DWH dùng các vấn tin SQL

* Dùng ROLLUP để vấn tin sản lượng cao nhất theo quốc gia, theo sản phẩm

|  |
| --- |
| SELECT b.country\_name, a.Ycode, MAX(a.SanLuong) AS [Sản lượng cao nhất]  FROM dbo.FactThongKe AS a INNER JOIN  dbo.DimCountry AS b ON a.country\_code = b.country\_code  GROUP BY ROLLUP (b.country\_name), ROLLUP (a.Ycode) |



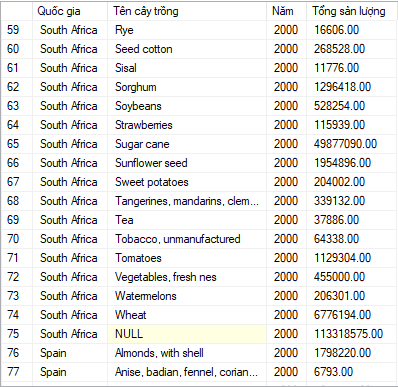
Kết quả thu được cho thấy: năm 2000, nước có sản lượng cao nhất là Thái Lan; năm 2001 cũng là Thái Lan….

Điều này có thể tạm đưa ra kết luận là Thái Lan luôn có sản lượng nông sản cao nhất.

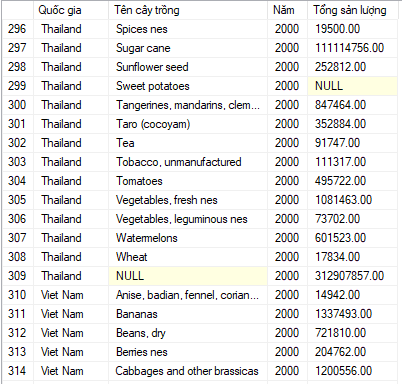
Nhưng giả sử, Thái Lan chỉ có 1 loại nông sản đạt sản lượng cao, còn các loại nông sản khác thì ngược lại. Để chứng thực giả thuyết này, ta có thể Rollup tất cả 3 thuộc tính (Quốc gia, Năm-ycode, Loại nông sản) để xem sản lượng các loại nông sản khác của Thái Lan nói riêng và các nước khác nói chung.

* Dùng ROLLUP để vấn tin tổng sản lượng từng loại nông sản theo quốc gia và năm

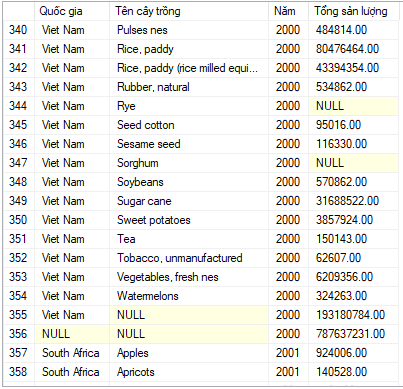
|  |
| --- |
| SELECT a.country\_name, c.item\_name, b.Ycode, SUM(b.SanLuong) AS [Tổng sản lượng]  FROM dbo.DimCountry AS a INNER JOIN  dbo.FactThongKe AS b ON a.country\_code = b.country\_code INNER JOIN  dbo.DimItem AS c ON b.item\_code = c.item\_code  GROUP BY ROLLUP (a.country\_name, c.item\_name), ROLLUP(b.Ycode) |



Ảnh trên chỉ ra 1 phần kết quả của vấn tin: Tổng sản lượng tất cả các loại cây trồng của South Africa năm 2000 là 62 901 838



Ảnh trên chỉ ra 1 phần kết quả của vấn tin: Tổng sản lượng tất cả các loại cây trồng của Thái Lan năm 2000 là 167 977 895



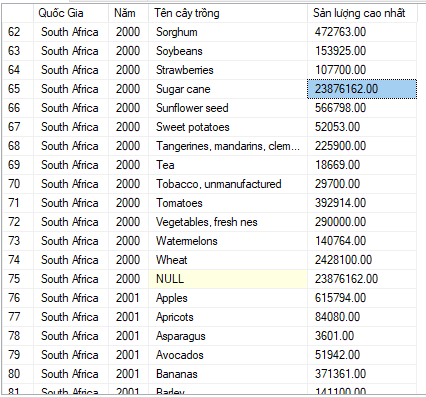
Ảnh trên chỉ ra 1 phần kết quả của vấn tin: Tổng sản lượng tất cả các loại cây trồng của Việt Nam năm 2000 là 104 373 320

Dòng 356 chỉ ra rằng tổng sản lượng tất cả các loại cây trồng tại các quốc gia là 440 497 048

Có thể dựa vào kết quả vấn tin trên để thấy sản lượng cây trồng của từng quốc gia ở từng năm có tăng hay giảm. Từ các so sánh tăng giảm, qua đó có thể tạm đưa ra kết luận về loại nông sản nào trong những năm gần đây được chú ý gieo trồng, loại nào được trồng nhiều ở quốc gia nào, hay quốc gia nào có sản lượng loại nông sản này cao để hợp tác xuất khẩu, …

* Dùng ROLLUP để truy vấn quốc gia có 1 loại nông sản có sản lượng cao nhất qua từng năm

|  |
| --- |
| SELECT b.country\_name [Quốc Gia], a.Ycode [Năm], c.item\_name [Tên cây trồng], MAX(a.SanLuong) AS [Sản lượng cao nhất]  FROM dbo.FactThongKe AS a INNER JOIN  dbo.DimCountry AS b ON a.country\_code = b.country\_code INNER JOIN  dbo.DimItem AS c ON a.item\_code = c.item\_code  GROUP BY ROLLUP (b.country\_name, a.Ycode, c.item\_name) |



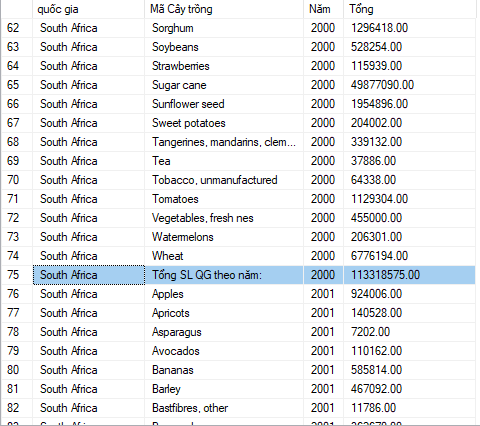
Từ một phần kết quả trên, ta có thể nhìn ra sản lượng các loại cây trồng tại từng quốc gia, năm. Đặc biệt, ta có thể thấy được sản lượng cao nhất của một quốc gia trong một năm nào đó.

Năm 2000, cây trồng có sản lượng cao nhất tại South Africa là Sugar cane, có sản lượng là 23876162.00

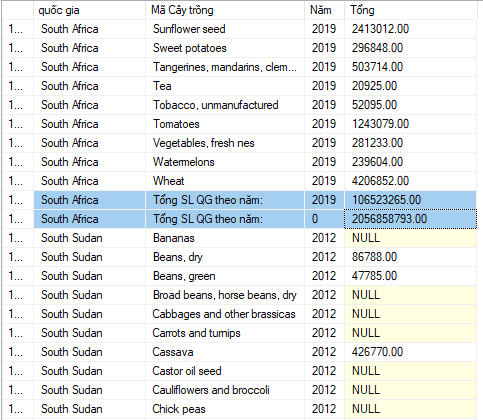
Chọn thế mạnh của từng nước để so sánh (lấy nông sản có sản lượng cao nhất để so sánh).

* Sản lượng theo quốc gia: dùng CASE WHEN … THEN … ELSE … END => Thêm “Tổng”/”Tổng SL QG theo năm”/0 theo từng dòng thống kê

|  |
| --- |
| SELECT CASE WHEN Grouping(a.country\_name) = 1 THEN N'Tổng' ELSE a.country\_name END AS [quốc gia],  CASE WHEN Grouping(c.item\_name) = 1 THEN N'Tổng SL QG theo năm: ' ELSE c.item\_name END AS [Mã Sản Phẩm],  CASE WHEN Grouping(b.Ycode) = 1 THEN 000 ELSE b.Ycode END AS Năm, SUM(b.SanLuong) AS Tổng  FROM dbo.DimCountry AS a INNER JOIN  dbo.FactThongKe AS b ON a.country\_code = b.country\_code INNER JOIN  dbo.DimItem AS c ON b.item\_code = c.item\_code  GROUP BY ROLLUP (a.country\_name, b.Ycode, c.item\_name) |



Năm 2000, tổng sản lượng các loại cây trồng của South Africa là 113318575.00



Năm 2019, tổng sản lượng các loại cây trồng của South Africa là 106523265.00

Tổng sản lượng trong 20 năm từ 2000-2019 của South Africa là 2056858793.00

South Sudan không có dữ liệu về cây trồng mãi cho tới năm 2012. Nhìn vào số liệu có thể thấy loại cây trồng của quốc gia này chưa được phong phú lắm,…



South Sudan thiếu rất nhiều loại cây trồng, ta có thể xem xét vấn đề nhập khẩu các loại nông sản mà quốc gia này không có.

Năm 2019, tổng sản lượng các loại cây trồng của South Sudan là 9493887.00

Từ năm 2012-2019, tổng sản lượng các loại cây trồng của South Sudan là 76043953.00

Tỉ trọng sản lượng các loại cây trồng năm 2019 chiếm khoảng 12,48% trên tổng sản lượng các loại cây trồng trong vòng 20 năm tại South Sudan. Con số này cho thấy dù có vẻ như nông nghiệp của South Sudan bắt đầu chậm hơn các nước khác, nhưng nó có tốc độ tăng trưởng cao. Đây là quốc gia có nền nông nghiệp tiềm năng.

### 4.1.11 Xây dựng ứng dụng Winforms để trực quan hóa dữ liệu và tạo các báo cáo nhằm mục đích phân tích cơ bản kho dữ liệu:

Xây dựng một phần mềm Winforms sử dụng cubes vừa tạo từ SSAS ở trên

Các Chart trong bài báo cáo này bao gồm 1 biểu đồ bánh Donut, 1 biểu đồ Line, 1 biểu đồ cột.

View được dùng để hỗ trợ vẽ các biểu đồ trên:

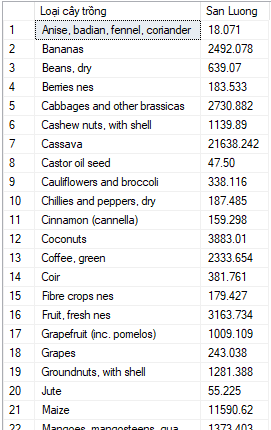
**W\_PieChart:** hỗ trợ vẽ biểu đồ Donut thể hiện tỉ lệ sản lượng các loại cây trồng tại Việt Nam năm 2019

|  |
| --- |
| SELECT dbo.DimItem.item\_name AS [Loại cây], FactThongKe\_1.Ycode AS Năm, SUM(FactThongKe\_1.SanLuong) \* 100 /  (SELECT SUM(SanLuong) AS TongSL  FROM dbo.FactThongKe  WHERE (country\_code = 237) AND (Ycode = 2019)) AS [Phần trăm SL]  FROM dbo.FactThongKe AS FactThongKe\_1 INNER JOIN  dbo.DimItem ON FactThongKe\_1.item\_code = dbo.DimItem.item\_code  WHERE (FactThongKe\_1.Ycode = 2019) AND (FactThongKe\_1.country\_code = 237)  GROUP BY FactThongKe\_1.Ycode, dbo.DimItem.item\_name |

****

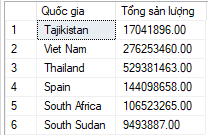
**W\_chartSL\_VietNam\_2019:** hỗ trợ vẽ biểu đồ thể hiện sản lượng các loại cây trồng tại Việt Nam năm 2019

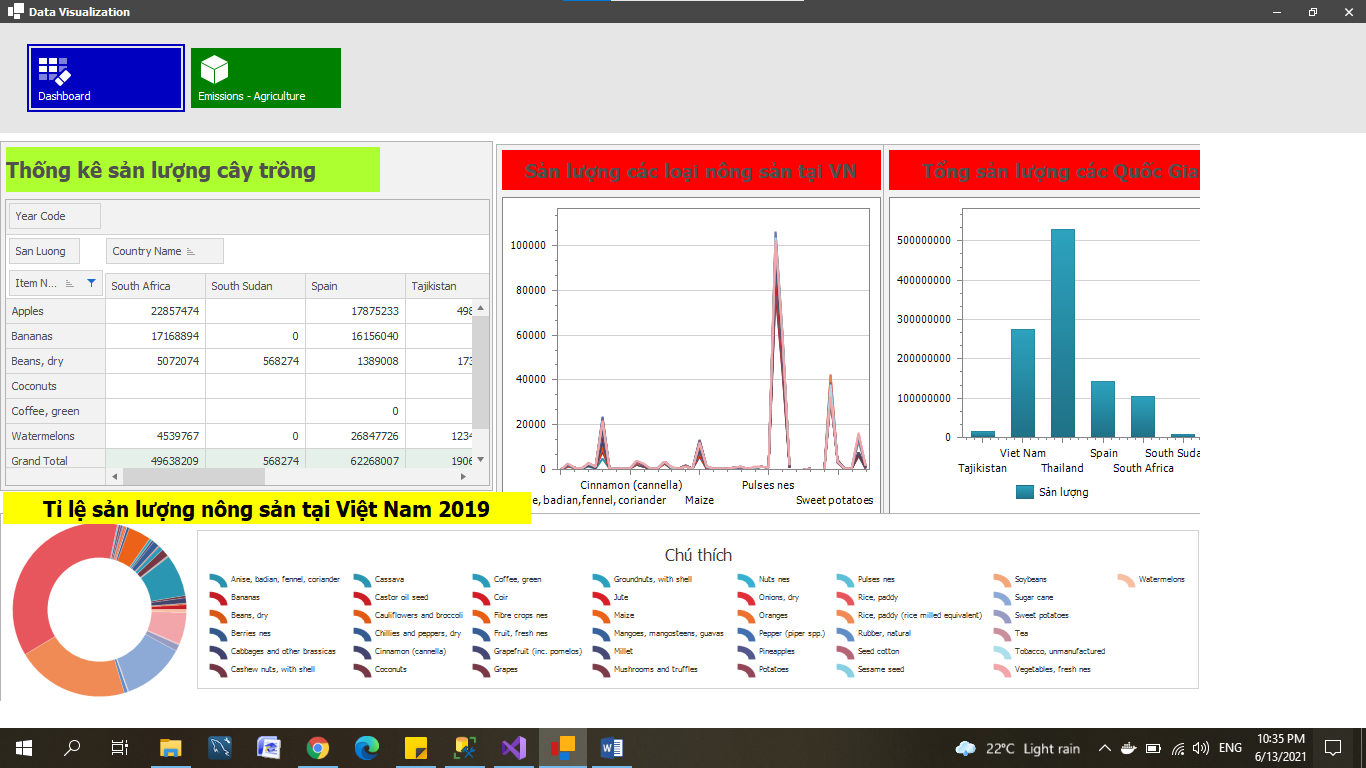
|  |
| --- |
| SELECT DimItem.item\_name AS [Loại cây trồng], SUM(FactThongKe.SanLuong / 1000) AS [San Luong]  FROM dbo.FactThongKe AS FactThongKe INNER JOIN  dbo.DimItem AS DimItem ON DimItem.item\_code = FactThongKe.item\_code  WHERE (FactThongKe.country\_code = 237) AND (FactThongKe.Ycode = 2019)  GROUP BY DimItem.item\_name |

****

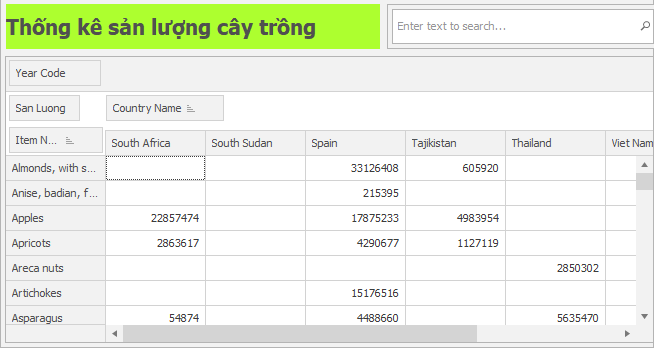
**W\_QG\_SumSL:** hỗ trợ vẽ biểu đồ cột nhằm so sánh tổng sản lượng của các quốc gia năm 2019

|  |
| --- |
| SELECT a.country\_name AS [Quốc gia], SUM(b.SanLuong) AS [Tổng sản lượng]  FROM dbo.DimCountry AS a INNER JOIN  dbo.FactThongKe AS b ON a.country\_code = b.country\_code INNER JOIN  dbo.DimItem AS c ON b.item\_code = c.item\_code  WHERE (b.Ycode = 2019)  GROUP BY a.country\_name |

****

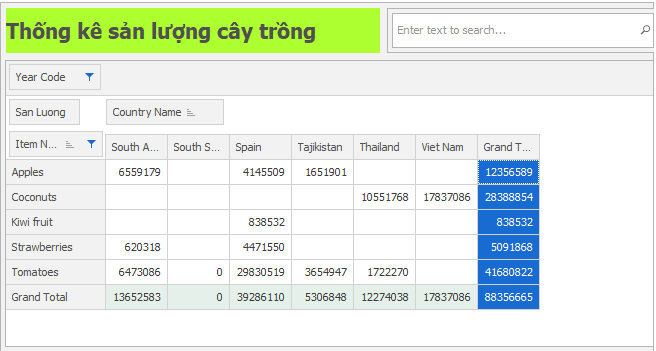


Hình 15: Tổng quan các Chart



Hình 16: Tạo thống kê đơn giản cho kho

Tùy mục đích mà tiến hành lọc kết quả. Thống kê này có thể lọc 3 yếu tố: Năm (Year Code), Loại mặt hàng nông sản (Item Name), Tên quốc gia (Country Name).



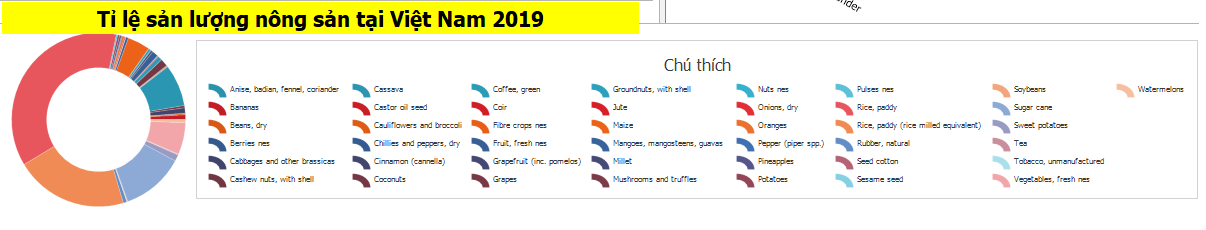
Hình 17: Ví dụ về thống kê tổng sản lượng một vài loại nông trong giai đoạn 2015-2019

Ví dụ trên tiến hành lọc Item Name (Loại mặt hàng nông sản), chỉ lấy kết quả của 5 loại nông sản là Táo, Dừa, trái Kiwi, Dâu, Cà chua. Nhìn sơ ta có thể thấy 5 loại nông sản này không phải quốc gia nào cũng có đủ. Trừ Spain có 4/5 mặt hàng. Còn lại các quốc gia khác chỉ có 1-2 loại. Thậm chí South Sudan không có bất kì loại nông sản nào ở trên. Tại Việt Nam Táo, trái Kiwi, Dâu, Cà chua cũng có nhưng sản lượng không cao, chủ yếu phải nhập khẩu. Qua đó ta có thể dựa vào các thống kê, tìm các mặt hàng mà nước khác không có, nhưng nước mình có để phục vụ việc xuất khẩu các mặt hàng nông sản của nước ta.



Hình 18: Ví dụ về thống kê tổng sản lượng 1 vài loại nông sản ở Việt Nam năm 2015-2019

Tiếp tục lọc một vài loại nông sản của nước ta trong 5 năm 2015-2019. Các mặt hàng lần lượt là hạt điều, bắp cải, quả mọng, đậu - khô, chuối, hồi badian. Có thể thấy bắp cải và chuối là 2 mặt hàng có sản lượng cao, hồi có sản lượng thấp. Hồi có sản lượng thấp có lẽ vì nước ta còn ít sử dụng mặt hàng nông sản này. Nếu ta có số liệu về nhu cầu sử dụng hồi của nước ta hoặc các nước khác, ta có thể thông qua đó thúc đẩy sản xuất hoặc giảm bớt.

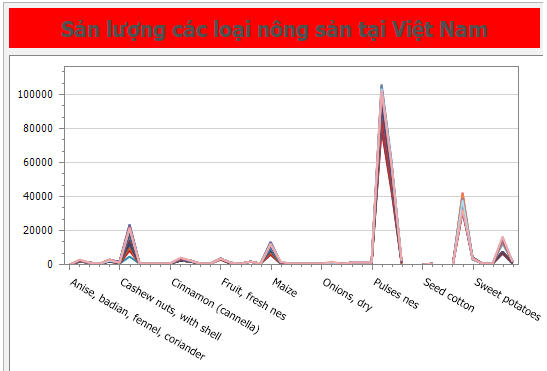


Hình 19: Biểu đồ Donut thể hiện tỉ lệ sản lượng nông sản tại Việt Nam năm 2019

Biểu đồ Donut hay biểu đồ tròn là loại biểu đồ thích hợp dùng để thể hiện cơ cấu hay tỷ lệ các thành phần trong một tổng thể chung hoặc cũng có thể vẽ biểu đồ tròn khi tỷ lệ % trong bảng số liệu cộng lại tròn 100. Vì đó mà em chọn biểu đồ tròn nhằm muốn xem tỷ lệ sản lượng nông sản tại Việt Nam trong năm gần đây (2019).

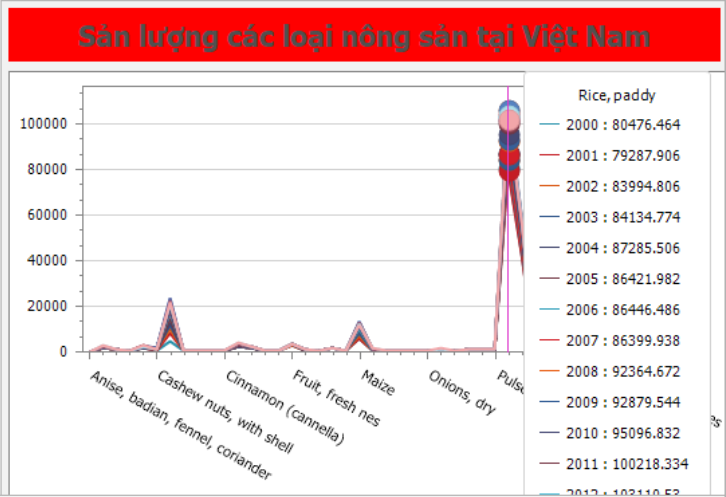
Nhìn vào biểu đồ có thể thấy được có 6/43 mặt hàng nông sản chiếm tỷ lệ cao trong tổng sản lượng nông sản Việt Nam năm 2019. Cao nhất trong đó là Rice and paddy (lúa, thóc). Nhì là rice and paddy (rice milled equivalent) (gạo và thóc loại gạo xay xát tương đương).

Sản lượng lúa gạo cao như vậy đã đưa Việt Nam trở thành nước xuất khẩu gạo đứng thứ 2 trên thế giới.



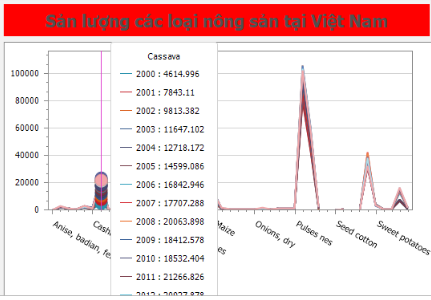
Hình 20: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (1)

Để xem được mặt bằng sản lượng các mặt hàng nông sản, và sự thay đổi của sản lượng trong một khoảng thời gian dài, Line Graph là một lựa chọn.



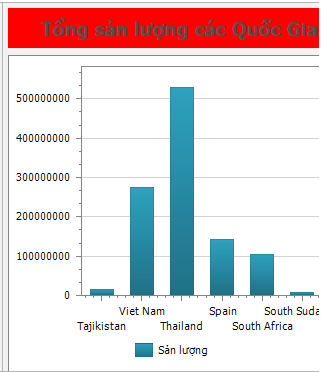
Hình 21: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (2)

Mỗi đường cong trong biểu đồ thể hiện sản lượng của 1 năm. Trong biểu đồ có 20 đường cong thể hiện sản lượng của từng loại nông sản trong 20 năm đó, nhưng các đường cong quá khít khiến ta nhìn không rõ. Dù vậy nhìn vào biểu đồ ta cùng có thể nhìn ra có một vài loại nông sản có sản lượng cực kì cao, cao hơn mặt bằng chung rất nhiều lần như: rice and paddy (lúa và thóc), Pulses (đậu), sugar canne (mía),… Các đường cong quá gần nhau, gần như chồng lên nhau cũng cho thấy trong 20 năm, sản lượng các mặt hàng nông sản không có sự thay đổi nhiều.



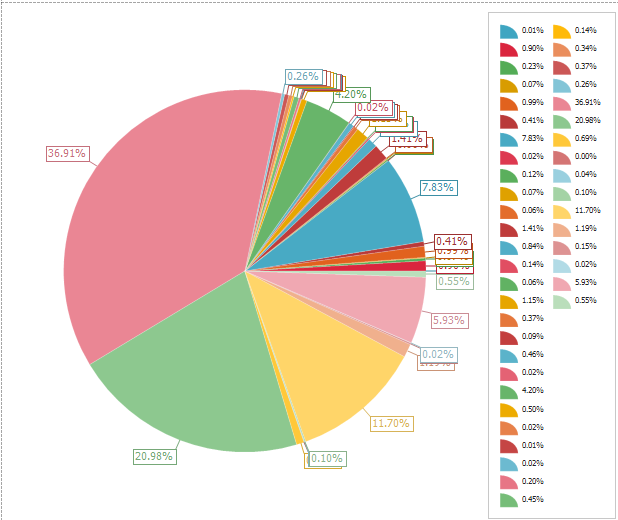
Hình 22: Line Graph thể hiện sản lượng của các mặt hàng nông sản tại Việt Nam trong giai đoạn 2000-2019 (3)

Nhưng cũng có một vài ngoại lệ như: Cassava (khoai mì) – sản lượng tăng khoảng 5 lần trong vòng 10 năm kể từ năm 2000.

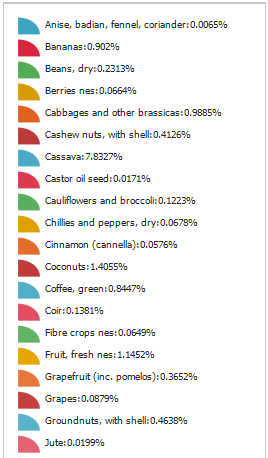
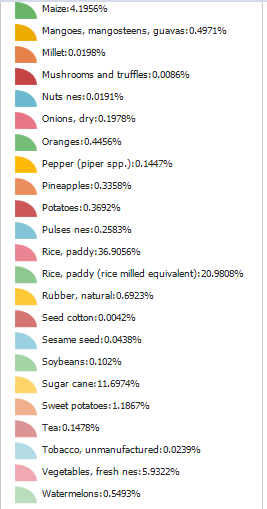


Hình 23: Biểu đồ cột thể hiện tổng sản lượng các nước năm 2019

Biểu đồ cột này cho thấy chênh lệch sản lượng giữa các nước. Thái Lan có sản lượng cao nhất, cao gấp đôi nước có sản lượng cao thứ nhì là Việt Nam. Tajikistan, Sudan là hai nước có sản lượng thấp nhất, thấp hơn khoảng 15 lần so với Việt Nam.



Hình 24: Báo cáo về tỉ lệ sản lượng các loại nông sản tại Việt Nam năm 2019

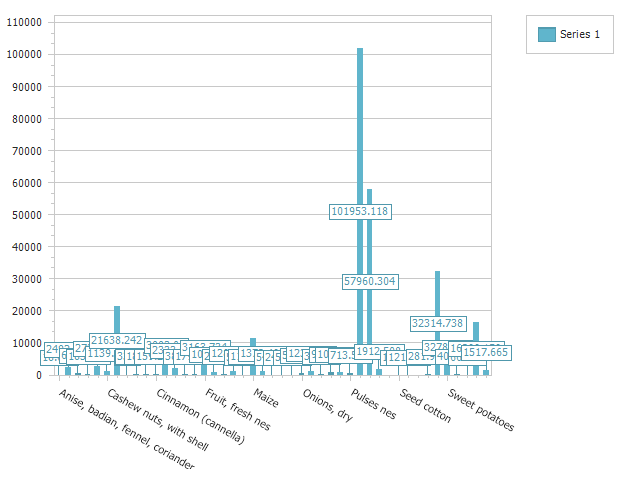
Hình 25: Chú thích biểu đồ Donut

Nhìn vào biểu đồ có thể thấy được có 6/43 mặt hàng nông sản chiếm tỷ lệ cao trong tổng sản lượng nông sản Việt Nam năm 2019. Cao nhất trong đó là Rice and paddy (lúa, thóc). Nhì là rice and paddy (rice milled equivalent) (gạo và thóc loại gạo xay xát tương đương).

Sản lượng lúa gạo cao như vậy đã đưa Việt Nam trở thành nước xuất khẩu gạo đứng thứ 2 trên thế giới.

Ngoài gạo, đường cũng chiếm tỉ lệ sản lượng cao, chiếm 11,7974% tổng sản lượng nông sản cả nước.

Còn các mặt hàng khác đa số chiếm tỉ lệ từ 0 đến 1%.



Hình 26: Biểu đồ cột thể hiện sản lượng các loại cây trồng tại Việt Nam năm 2019

Các cột thể hiện sản lượng của nông sản trong năm 2019, trong biểu đồ có 2 cột cao vượt trội, 2 cột này đều thể hiện sản lượng của gạo.

# Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN

## Kết luận

### Những kết quả đạt được

- Hiểu được mô hình của một kho dữ liệu trên thực tế và mô hình CSDL đa chiều.

- Hiểu được các khái niệm cơ bản về Data Integration Workflows, OLTP, OLAP

**-** Hiểu được cách xây dựng một kho dữ liệu và duy trì hoạt động cũng như khai thác nó.

- Biết cách sử dụng các vấn tin phân tích để phân tích một CSDL quan hệ.

- Hiểu được cơ chế hoạt động của các công cụ đã sử dụng đặc biệt là SQL Server và các dịch vụ đi kèm theo nó.

- Biết cách trực quan hóa để đưa ra cái nhìn tổng quát về dữ liệu cũng như phân tích các yếu tố ảnh hưởng đến số liệu thống kê.

- Học được cách xây dựng Dashboard, các Report thống kê phục vụ cho mục đích phân tích, báo cáo.

### Hạn chế

- Tốc độ truy vấn trên DWH khá chậm vì quy mô dữ liệu rất lớn và cần được cải tiến.

- Chưa hiện thực hóa được việc sử dụng Materialize View

- Do thời gian có hạn nên các file báo cáo cũng như sản phẩm đính kém có thể chưa chỉnh chu.

## Hướng phát triển

### Hướng khắc phục các hạn chế

- Cải thiện tốc độ truy vấn trên kho dữ liệu

- Triển khai các Meterialize View trên kho dữ liệu

- Cố gắng chỉnh chu các file báo cáo cũng như snar phẩm.

### Hướng mở rộng ĐỒ ÁN HỌC PHẦN

- Mở rộng kho dữ liệu ra nhiều lĩnh vực hơn và phạm vi nghiên cứu lớn hơn vì dữ liệu của tổ chức FAO vô cùng dồi dào và phong phú.

- Sử dụng các công cụ khác để tích hợp và triển khai kho dữ liệu để so sánh hiệu suất, có thể sử dụng các nền tảng đám mây phổ biến như là Google Cloud, Amazon Web Service, Azure hoặc các nền tảng kho dữ liệu mã nguồn mở như Apache Hive, postgresql,…

- Mở rộng bài toán phân tích bằng cách sử dụng các thuật toán mining như phân loại, hồi quy,.. để đưa ra cái nhìn chính xác hơn về dữ liệu mà chúng ta có.

- Sử dụng các công cụ front end để khai thác, trực quan hóa và phân tích kho dữ liệu để giúp các nhà quản lý đưa ra các quyết định chính xác hơn.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Võ Xuân Thể (2021), *Bài giảng học phần Kho dữ liệu*, TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM (lưu hành nội bộ).

# CÁC PHỤ LỤC

* Toàn bộ sourecode được lưu trữ trong folder đính kèm
* Yêu cầu phần mềm: Visual Studio 2019, Microsoft SQL Server Management Studio 18, DevExpress, SQL Server Analysis Services và SQL Server Data Tool
* Hướng dẫn sử dụng:

+ Cài đặt các phần mềm cần thiết.

+ Sử dụng SSMS18 để restore database từ file .bak đã đính kèm.

+ Mở project SSAS và deploy cubes lên server SSAS server.

+ Mở project bằng Visual Studio 2019 và chạy ứng dụng.

# ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN

ĐỒ ÁN HP THAY CHO BÀI THI KẾT THÚC HỌC PHẦN

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hình thức KT** | **Nội dung** | **Chuẩn đầu ra đánh giá** | **Trình độ năng lực** | **Phương pháp đánh giá** | **Công cụ đánh giá** | **Tỉ lệ (%)** |
| **BÀI TẬP LẬP TRÌNH** | | | | | |  |
| BL#1 | Bài tập lập trình theo từng Buổi thực hành = tính bình quân  Kiểm tra viết | G1.1 | 3 | Bài tập  Kiểm tra | Bài tập  Kiểm tra | 50 |
| G2.1 | 3 |
| G3.1  G3.2 | 4  3 |
| G4.1 | 5 |
| **ĐỒ ÁN HỌC PHẦN (Project)** | | | | | |  |
| ĐA#1 | Phát triển một hệ thống KHO DỮ LIỆU đơn giản:  + Đề tài Nhóm có phân công phân hệ cụ thể cho từng thành viên thuộc nhóm  + Xây dựng và thực nghiệm Hệ thống Kho dữ liệu theo nhóm và phân hệ cá nhân  => báo cáo vào ngày thi. | G1.1 | 3 | Báo cáo, Demo và thuyết trình | Rubric | 50 |
| G2.1 | 3 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |
| G2.1 | 4 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |