**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:

Tên đề tài Nhóm: **Xây dựng Kho Dữ Liệu Doanh Số Game Khai Thác Tại Công Ty Phân Phối Game**

Lớp:

**NHÓM 10**

**THÀNH VIÊN**: **MSSV**

Trần Minh Đức 20133037

Trần Văn Thắng 20133121

Trần Văn Trọng 20133103

**TP.HCM, ngày 17 Tháng 4 năm 2023**

**LỜI CẢM ƠN**

Chúng em xin chân t hành cảm ơn sự hướng dẫn tận tình của thầy ***Nguyễn Văn Thành*** đã giúp chúng em hoàn thành bài báo cáo cuối kì bộ môn Kho dữ liệu. Tuy đã cố gắng học tập, tìm hiểu, phân tích thiết kế và xây dựng hệ thống nhưng ắt hẳn thiếu sót là điều không thể tránh khỏi. Chúng em rất mong nhận được sự thông cảm và góp ý chân thành của Thầy. Chúng em xin chân thành cảm ơn.

**MỤC LỤC**

[Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 5](#_Toc135176520)

[1.1. Tổng quan về ĐỒ ÁN HỌC PHẦN 5](#_Toc135176521)

[1.2. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN 5](#_Toc135176522)

[1.3. Mục tiêu của ĐỒ ÁN HỌC PHẦN 5](#_Toc135176523)

[1.4. Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN 5](#_Toc135176524)

[1.5. Sản phẩm của ĐỒ ÁN HỌC PHẦN 5](#_Toc135176525)

[Chương 2. TỔNG QUAN VỀ KHO DỮ LIỆU GAME SALES KHAI THÁC TẠI CÔNG TY PHÂN PHỐI GAME 6](#_Toc135176526)

[2.1. Ý tưởng hình thành DWH Game Sales khai thác tại công ty phân phối game 6](#_Toc135176527)

[2.2. Giới thiệu chung về Kho dữ liệu 6](#_Toc135176528)

[2.2.1. Mô tả tổng quan về DWH 6](#_Toc135176529)

[2.2.2. Các dữ liệu nguồn hình thành DWH 6](#_Toc135176530)

[2.3. Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH 6](#_Toc135176531)

[Chương 3. PHÂN TÍCH, THIẾT KẾ KHO DỮ LIỆU DOANH SỐ GAME KHAI THÁC TẠI CÔNG TY PHÂN PHỐI GAME 7](#_Toc135176532)

[3.1. Mô tả hệ thống các DB gốc liên quan phân hệ DWH 7](#_Toc135176533)

[3.2. Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ 10](#_Toc135176534)

[3.3. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ 10](#_Toc135176535)

[Chương 4. TÍCH HỢP DỮ LIỆU VÀO KHO(SSIS) 12](#_Toc135176536)

[4.1. Xử lí dữ liệu 12](#_Toc135176537)

[4.2. Quá trình đổ dữ liệu vào Stage 22](#_Toc135176538)

[4.2.1. Quá trình tạo Control Flow 22](#_Toc135176539)

[4.2.2. Kết quả thực thi 31](#_Toc135176540)

[4.2.3. Kết quả trong SQL Server 33](#_Toc135176541)

[4.3. Quá trình đổ dữ liệu từ Kho Stage sang Kho Ware Hosue 41](#_Toc135176542)

[4.3.1. Quá trình tạo Control Flow 41](#_Toc135176543)

[4.3.2. Kết quả thực thi 50](#_Toc135176544)

[4.3.3. Kết quả trên SQL Server 52](#_Toc135176545)

[Chương 5. PHÂN TÍCH DỮ LIỆU(SSAS) 57](#_Toc135176546)

[5.1. Quá trình xây dựng mô hình 57](#_Toc135176547)

[5.1.1. Tạo SSAS Project 57](#_Toc135176548)

[5.2. Thực hiện các câu truy vấn (SSAS, Pivot Table, T-SQL) 64](#_Toc135176549)

[5.2.1. Câu 1: Tổng doanh thu của các thể loại game 64](#_Toc135176550)

[5.2.2. Câu 2: Cho biết Doanh thu của các thể loại game theo từng năm? 65](#_Toc135176551)

[5.2.3. Câu 3: Cho biết doanh thu của các thể loại game theo từng khu vực? 67](#_Toc135176552)

[5.2.4. Câu 4: Cho biết Doanh thu của các thể loại game theo từng nền tảng? 68](#_Toc135176553)

[5.2.5. Câu 5: Doanh thu của các thể loại game theo từng năm và tổng doanh thu của toàn thị trường? 70](#_Toc135176554)

[5.2.6. Câu 6: Số lượng doanh thu của các nền tảng game theo từng năm? 71](#_Toc135176555)

[5.2.7. Câu 7: Số lượng doanh thu của các nền tảng game theo từng khu vực? 72](#_Toc135176556)

[Chương 6. TRỰC QUAN HÓA BẰNG POWER BI 73](#_Toc135176557)

[6.1. Tổng doanh thu của các thể loại game 73](#_Toc135176558)

[6.2. Tổng doanh thu theo từng năm(1980 - 2017) 73](#_Toc135176559)

[6.3. Tổng doanh thu theo khu vực 74](#_Toc135176560)

[Phân chia công việc 75](#_Toc135176561)

1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN
   1. Tổng quan về ĐỒ ÁN HỌC PHẦN

Đồ án Học phần môn Kho dữ liệu: **Phát Triển Mô Hình Kho Dữ Liệu Doanh số về game Khai Thác Tại Công Ty Phân Phối Game**

* 1. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN

Phân tích, thiết kế, chuyển đổi dữ liệu database gốc sang nhà Kho dữ liệu

* 1. Mục tiêu của ĐỒ ÁN HỌC PHẦN

Nắm được cách phân tích thiết kế Data Warehouse theo nhu cầu phân tích dữ liệu

Nắm được kỹ thuật tích hợp dữ liệu từ các nguồn về Data Warehouse

Nắm được kỹ thuật phân tích bằng SSAS

* 1. 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

+ SSAS

+SSIS

+Power BI

* 1. Sản phẩm của ĐỒ ÁN HỌC PHẦN

Kho dữ liệu “video\_games\_DW”

SSIS Project

SSAS Project

Các view phân tích, truy vấn Database gốc, DW

1. TỔNG QUAN VỀ KHO DỮ LIỆU GAME SALES KHAI THÁC TẠI CÔNG TY PHÂN PHỐI GAME
   1. Ý tưởng hình thành DWH Game Sales khai thác tại công ty phân phối game

Ý tưởng: Xây dựng 1 Data Warehouse dưới dạng mô hình Star Schema có các Dim table (chứa các đối tượng cần phân tích) và Fact table (chứa các id của các đối tượng và 1 cột chứa giá trị được group by từ các id)

* 1. Giới thiệu chung về Kho dữ liệu
     1. Mô tả tổng quan về DWH

1 tập đoàn chuyên phân phối game cho các thị trường trên thế giới có 1 hệ thống database lưu trữ doanh số bán game qua hằng năm.

* + 1. Các dữ liệu nguồn hình thành DWH

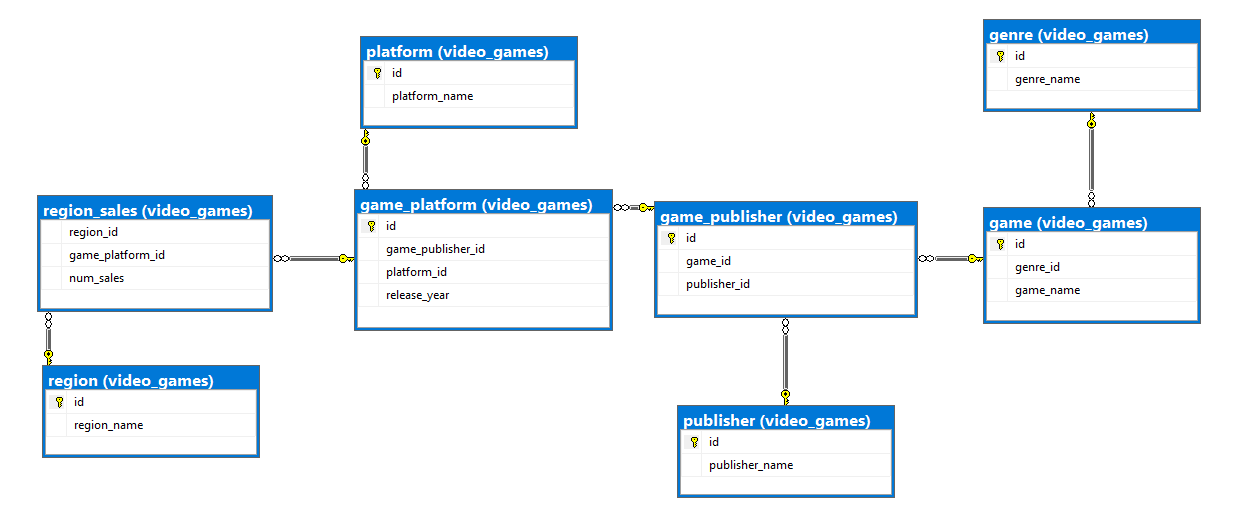
Nguồn dữ liệu lấy trên Kaggle là Database video\_games đặt tại trụ sở của công ty phân phối game.

* 1. Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH

Tập đoàn có nhu cầu phân tích các doanh số bán hàng đã thu thập qua các năm để hiểu thị trường, hiểu thể loại game nào được ưa thích, nhà phát hành, thể loại game nào nào đang làm mưa làm gió trên thị trường…

1. PHÂN TÍCH, THIẾT KẾ KHO DỮ LIỆU DOANH SỐ GAME KHAI THÁC TẠI CÔNG TY PHÂN PHỐI GAME
   1. Mô tả hệ thống các DB gốc liên quan phân hệ DWH

Hệ thống Database lưu trữ



Mô tả DB:

**region**(id, region\_name)

**genre**(id, genre\_name)

**publisher**(id, publisher\_name)

**platform**(id, platform\_name)

**game**(id, genre\_id, game\_name)

**game\_publisher**(id, game\_id, publisher\_id)

**game\_platform**(id, game\_publisher\_id, platform\_id, release\_year)

**region\_sales**(region\_id, game\_platform\_id, num\_sales)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt vùng |
| 2 | region\_name | nvarchar(150) | NOT NULL | Tên vùng |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Genre** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt thể loại |
| 2 | genre\_name | nvarchar(150) | NOT NULL | Tên thể loại |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Publisher** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt nhà phát hành |
| 2 | publisher\_name | nvarchar(150) | NOT NULL | Tên nhà phát hành |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Platform** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt nền tảng |
| 2 | platform\_name | nvarchar(150) | NOT NULL | Tên nền tảng |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Game** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt game |
| 2 | game\_name | nvarchar(150) | NOT NULL | Tên game |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **game\_publisher** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt game và nhà phát hành |
| 2 | game\_id | int | NOT NULL | ID game |
| 3 | publisher\_id | int | NOT NULL | ID nhà phát hành |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **game\_platform** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | id | int | PRIMARY KEY | Khoá chính phân biệt game và nền tảng |
| 2 | game\_publisher\_id | int | NOT NULL | ID game và nhà phát hành |
| 3 | platform\_id | int | NOT NULL | ID nền tảng |
| 4 | release\_year | int | NOT NULL | Năm phát hành |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **region\_sale** | | | | |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** | **Diễn giải** |
| 1 | region\_id | int | PRIMARY KEY | ID khu vực |
| 2 | game\_platform\_id | int | PRIMARY KEY | ID game và nền tảng |
| 3 | num\_sales | Decimal(5,2) | NOT NULL | Đơn vị bán (1.000.000 USD) |

* 1. Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ

Dim:

DimRegion: Dim về khu vực/thị trường bán game

DimGerne: Dim về thể loại game

DimPublisher: Dim về nhà phát hành

DimPlatform: Dim về nền tảng phát triển

DimGame: Dim về game

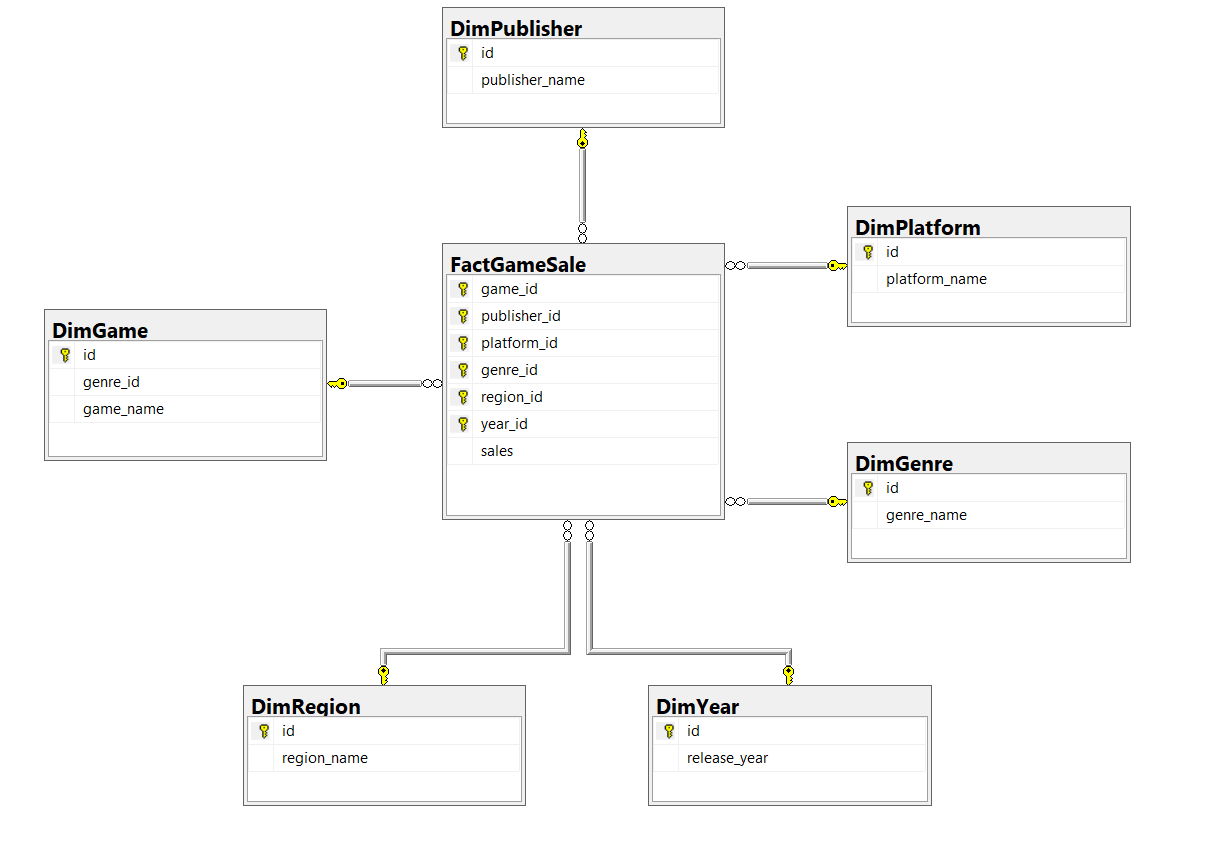
DimYear: Dim về năm phát hành

Có 1 fact:

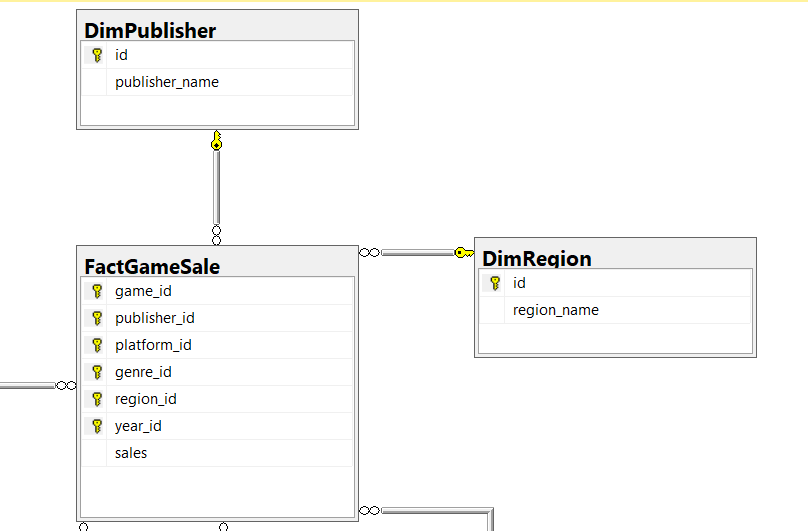
Fact doanh thu = sum(num\_sales) group by (id DimRegion, id DimGerne, id DimPublisher, id DimPlatform, id DimGame, id DimYear)

* 1. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ

DW chung :



DW phân hệ phân tích **doanh thu theo nhà phát hành và khu vực**



1. TÍCH HỢP DỮ LIỆU VÀO KHO(SSIS)
   1. Xử lí dữ liệu

-Dùng python để chuyển dữ liệu từ file csv thành các bảng trong DB Stage

-Sau khi xử lí xong ta tạo được các file csv với nội dung là dữ liệu của các bảng:

\*Region

Graphical user interface, application, table, Excel

Description automatically generated

\*Year

Graphical user interface, application, table, Excel

Description automatically generated

\*Platform

Graphical user interface, application, table, Excel

Description automatically generated

\*Publisher

Graphical user interface, application, table, Excel

Description automatically generated

\*Genre

Graphical user interface, application, table, Excel

Description automatically generated

\*Game

Graphical user interface, application, table, Excel

Description automatically generated

\*Game\_Publisher

Graphical user interface, application, table, Excel

Description automatically generated

\*Game\_Platform

Graphical user interface, application, table, Excel

Description automatically generated

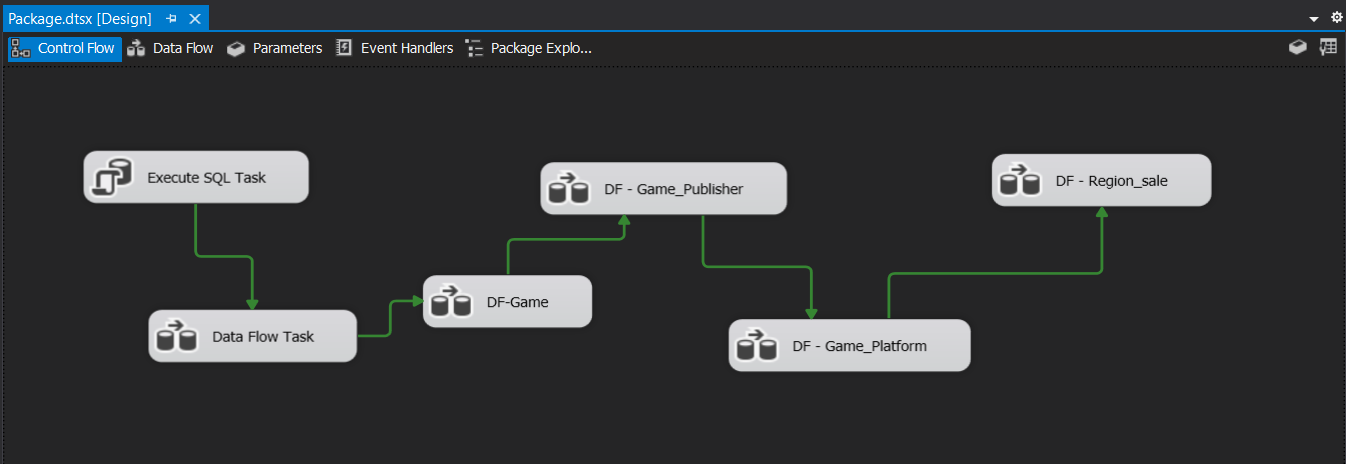
\*Region\_sale

Graphical user interface, application, table, Excel

Description automatically generated

* 1. Quá trình đổ dữ liệu vào Stage
     1. Quá trình tạo Control Flow

Kéo một Execute SQL Task và 5 Data Flow vào Control Flow và đặt tên chúng như sau:



**Data Flow Task:**

Kéo Source Assistant và chọn Flat File:

Graphical user interface, text, application, email

Description automatically generated

Bấm vào New để tạo một Flat File Connection Manager sau đó chọn file csv của bảng Year

Graphical user interface, text, application, email

Description automatically generated

Bấm qua cột Columns để xem thông tin các cột trong bảng

Graphical user interface

Description automatically generated

Sau đó bấm OK ta được một Flat File source

Graphical user interface, application

Description automatically generated

Kéo một Destination Assistant vào chọn DB stage đã tạo trước đó sau đó nhấn OK và đổi tên thành Year Staging:

Graphical user interface, application

Description automatically generated

Kéo một kết nối giữa đích và nguồn

Graphical user interface

Description automatically generated

Ở Destination ta chọn bảng Year

Graphical user interface, application

Description automatically generated

Sau đó Mapping các columns cho phù hợp

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generated

Làm tương tự đối với các file csv khác ta thu được các Data Flow task như sau:

Data Flow task:

Graphical user interface

Description automatically generated

DF - Game

Graphical user interface, application

Description automatically generated

DF – Game\_Publisher

Graphical user interface, text

Description automatically generated

DF – Game\_Platform

Graphical user interface, application, Teams

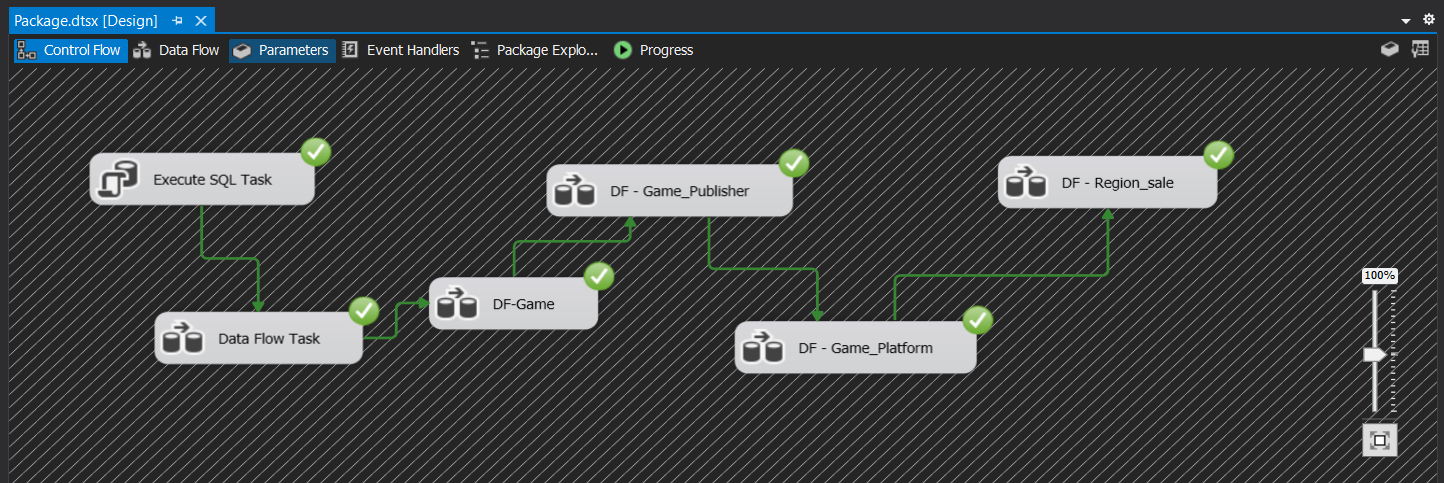
Description automatically generated

DF – Region\_Sale

Graphical user interface, application

Description automatically generated

* + 1. Kết quả thực thi



Data Flow Task

A screenshot of a computer

Description automatically generated with medium confidence

DF- Game

A screenshot of a computer

Description automatically generated with medium confidence

DF – Game\_publisher

A screenshot of a video game

Description automatically generated

DF – Game\_platform

A screenshot of a video game

Description automatically generated

DF – Region sale

Graphical user interface, application

Description automatically generated

* + 1. Kết quả trong SQL Server

Bảng Year:

A picture containing table

Description automatically generated

Bảng Region

Table

Description automatically generated

Bảng Genre

Table

Description automatically generated

Bảng Publisher

A picture containing chart

Description automatically generated

Bảng Platform

A screenshot of a computer

Description automatically generated with low confidence

Bảng Game



Bảng Game\_Publisher

A picture containing chart

Description automatically generated

Bảng Game\_Platform

Table

Description automatically generated with medium confidence

Bảng Region\_Sale

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Quá trình đổ dữ liệu từ Kho Stage sang Kho Ware Hosue
     1. Quá trình tạo Control Flow

Kéo một Execute SQL task và 3 data flow task vào và setup như sau:

Graphical user interface, text, application, chat or text message

Description automatically generated

Ở Data Flow Task ta kéo một Source Assistant vào sau đó chọn SQL Server và bấm New:

Graphical user interface, application

Description automatically generated

Chọn DB Stage để làm Source

Graphical user interface, text, application, email

Description automatically generated

Ở Source Editor chọn bảng year

Graphical user interface, text, application

Description automatically generated

Và đổi tên thành year source

Text

Description automatically generated

Sau đó kéo một Destination vào và chọn kết nối tới database DWH

Graphical user interface, text, application

Description automatically generated

Và đổi tên thành DimYear

Text

Description automatically generated with medium confidence

Ở Destination Editor chọn bảng DimYear

Graphical user interface, application

Description automatically generated

Sau đó Mapping cho phù hợp

Graphical user interface, application

Description automatically generated

Làm tương tự như vậy ta được như sau

Graphical user interface, text, application, chat or text message

Description automatically generated

Data Flow Task

Graphical user interface

Description automatically generated

Data Flow Task 1

Graphical user interface

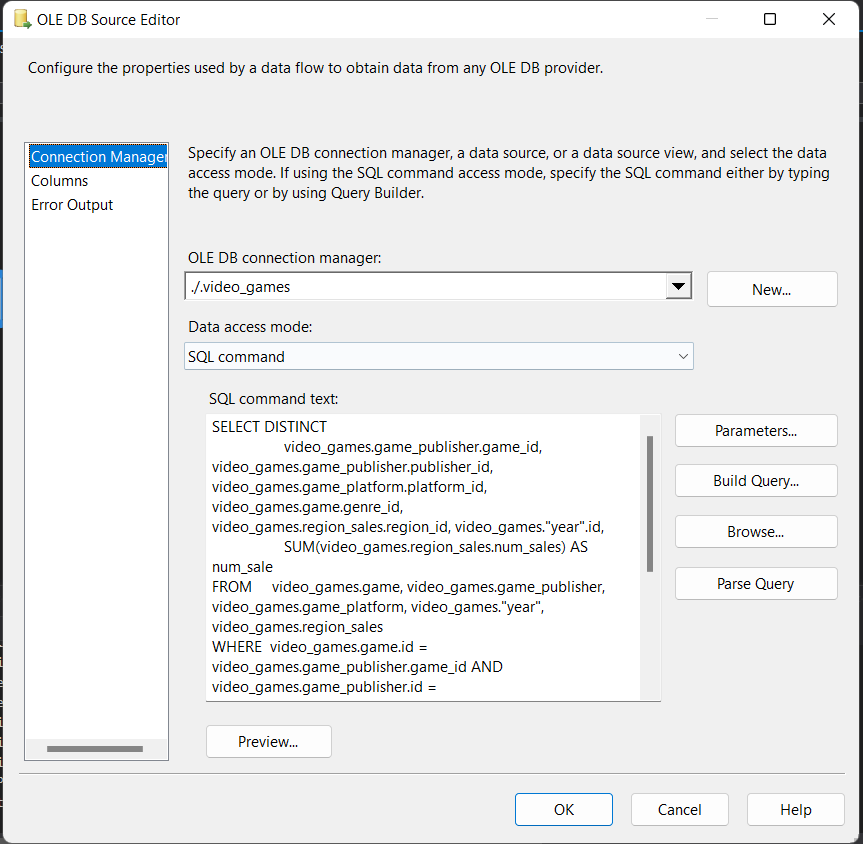
Description automatically generated

Data Flow Task 2:

Ở đây Sales Source phải cần dùng đến Sql Command để lấy dữ liệu từ các bảng cần thiết

Graphical user interface, application, Word

Description automatically generated

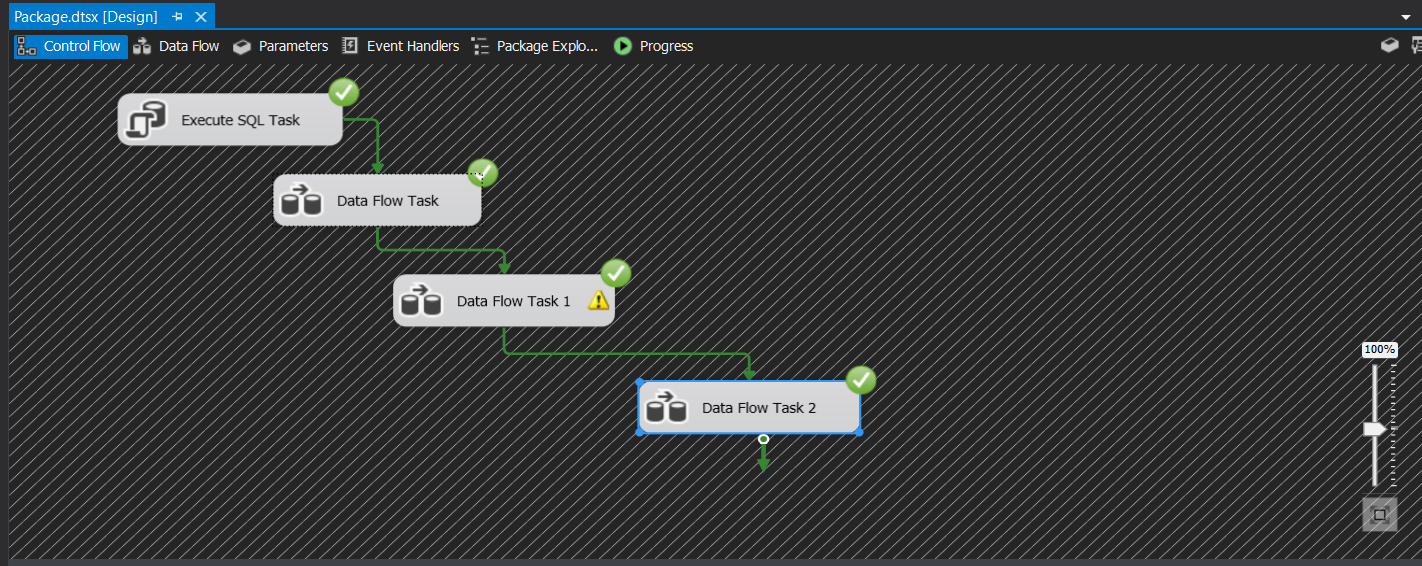


Graphical user interface, application

Description automatically generated

* + 1. Kết quả thực thi

Sau khi chạy SSIS project ta được:



Data Flow Task

A screenshot of a computer

Description automatically generated with medium confidence

Data Flow Task 1

A screenshot of a computer

Description automatically generated with medium confidence

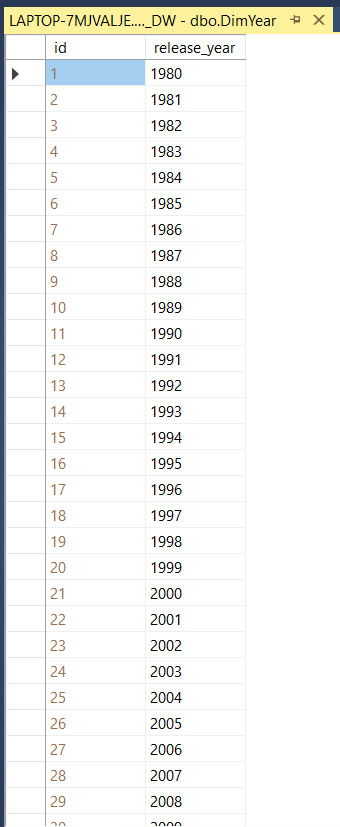
Data Flow Task 2

Graphical user interface, text

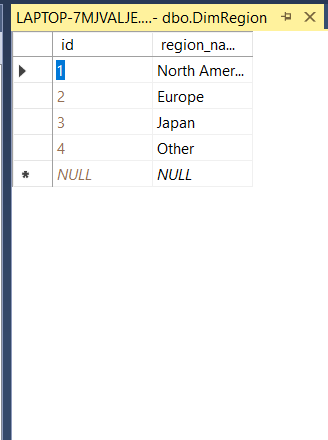
Description automatically generated

* + 1. Kết quả trên SQL Server

**DimYear**

****

Dim Region



DimPublisher

Application

Description automatically generated with medium confidence

DimPlatform

Chart

Description automatically generated with medium confidence

DimGenre

Table

Description automatically generated

DimGame

A picture containing chart

Description automatically generated

FactGameSale

Table

Description automatically generated with medium confidence

1. PHÂN TÍCH DỮ LIỆU(SSAS)
   1. Quá trình xây dựng mô hình
      1. Tạo SSAS Project

B1: Tạo Data Source

Graphical user interface, text, application

Description automatically generated

B2: Tạo Data Source View

Graphical user interface, application

Description automatically generated

Diagram

Description automatically generated

B4: Tạo Data Cube

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

B5: Cấu hình các Dimension

DimGenre

Graphical user interface, Word

Description automatically generated

DimGame

Graphical user interface, Word

Description automatically generated

DimRegion

Graphical user interface, text, Word

Description automatically generated

DimYear

Graphical user interface, Word

Description automatically generated

B6: Deploy and Process

A screenshot of a computer

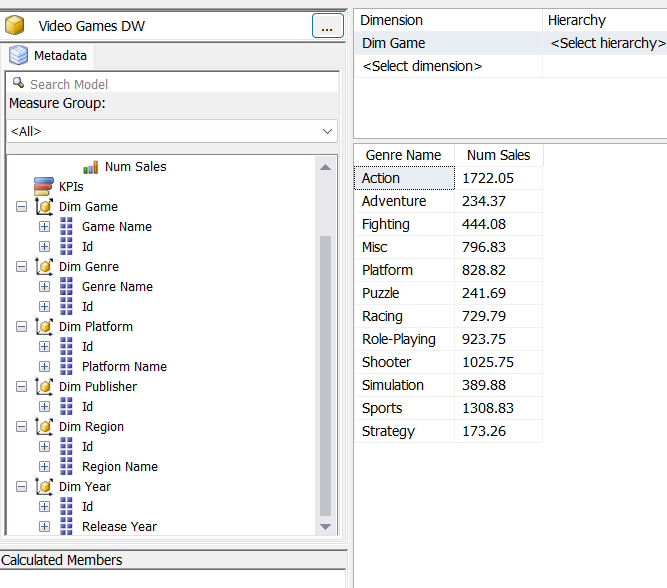
Description automatically generated with medium confidence

A picture containing text, screenshot, indoor, computer

Description automatically generated

* 1. Thực hiện các câu truy vấn (SSAS, Pivot Table, T-SQL)
     1. Câu 1: Tổng doanh thu của các thể loại game

-Sử dụng SSAS



-Pivot table trên Excel

Table

Description automatically generated

-Truy vấn SQL

Graphical user interface, application, Word

Description automatically generated

* + 1. Câu 2: Cho biết Doanh thu của các thể loại game theo từng năm?

- Sử dụng SSAS

Graphical user interface

Description automatically generated

-Pivot Table trên excel

A picture containing calendar

Description automatically generated

-Truy vấn SQL

Graphical user interface, application

Description automatically generated

* + 1. Câu 3: Cho biết doanh thu của các thể loại game theo từng khu vực?

-Sử dụng SSAS

Graphical user interface

Description automatically generated with medium confidence

-Pivot Table trên Excel

Table

Description automatically generated

* + 1. Câu 4: Cho biết Doanh thu của các thể loại game theo từng nền tảng?

-Sử dụng SSAS

Graphical user interface

Description automatically generated

-Pivot Table

Calendar

Description automatically generated

-Truy vấn SQL

Graphical user interface, application, Word

Description automatically generated

* + 1. Câu 5: Doanh thu của các thể loại game theo từng năm và tổng doanh thu của toàn thị trường?

-Truy vấn SQL

Graphical user interface, application, Word

Description automatically generated

Table

Description automatically generated

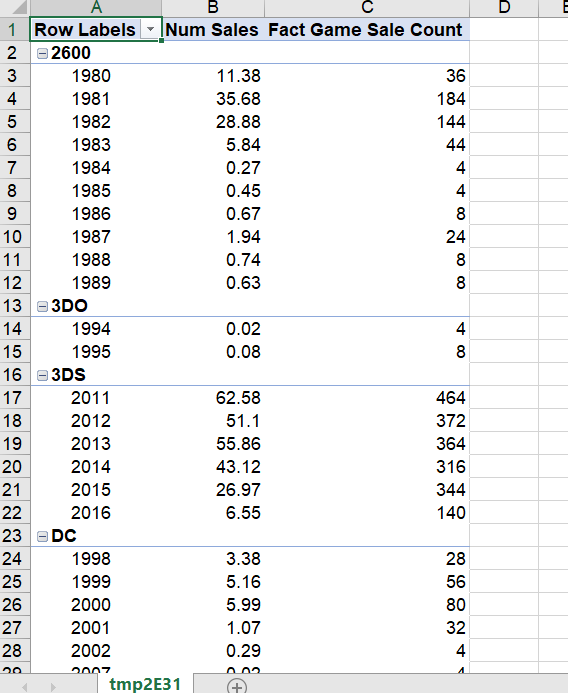
* + 1. Câu 6: Số lượng doanh thu của các nền tảng game theo từng năm?

-Sử dụng SSAS

Graphical user interface, application

Description automatically generated

-Pivot Table



* + 1. Câu 7: Số lượng doanh thu của các nền tảng game theo từng khu vực?

-SSAS

Graphical user interface

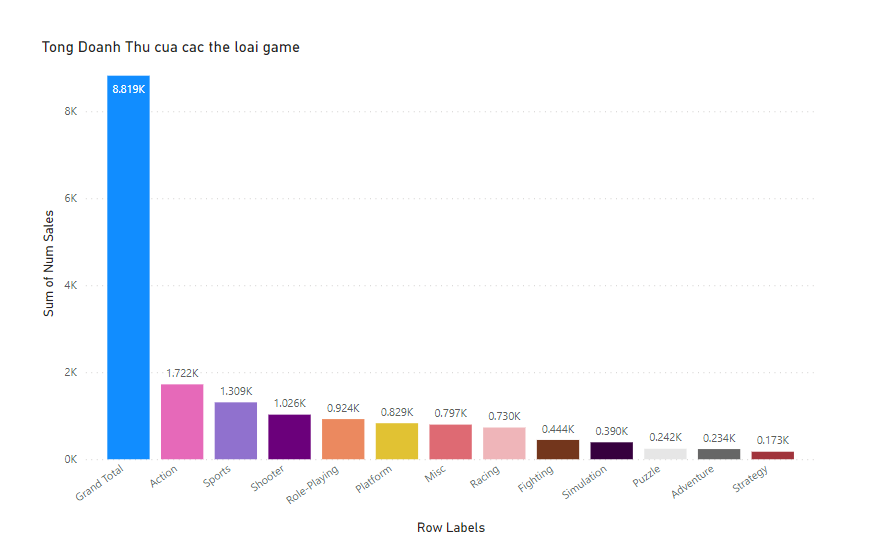
Description automatically generated

-Pivot Table

Table

Description automatically generated

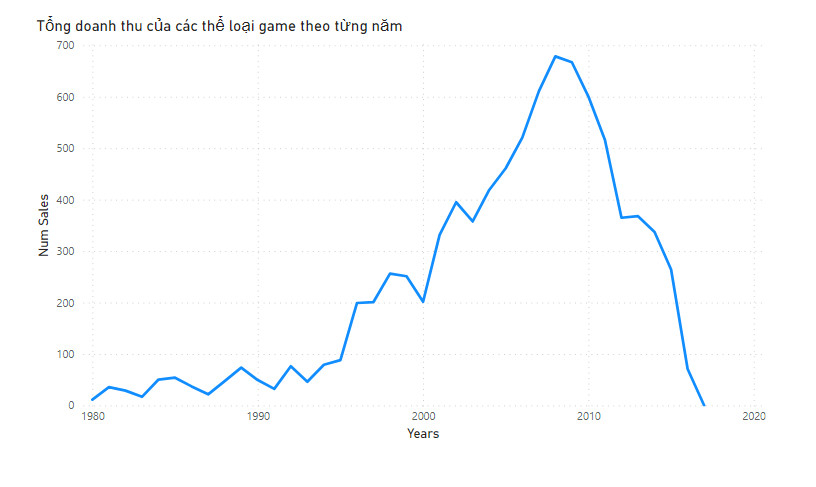
1. TRỰC QUAN HÓA BẰNG POWER BI
   1. Tổng doanh thu của các thể loại game



* 1. Tổng doanh thu theo từng năm(1980 - 2017)

Chart, sunburst chart

Description automatically generated



* 1. Tổng doanh thu theo khu vực

Chart, line chart

Description automatically generated

# Phân chia công việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Công Việc** | **Chi tiết** | **Trần Minh Đức** | **Trần Văn Thắng** | **Trần Văn Trọng** |
| Tìm hiểu về Dataset | Tìm hiểu về nghiệp vụ của tập dữ liệu từ đó đưa ra các câu hỏi cũng như tạo kho |  |  |  |
| Phân tích thiết kế Kho dữ liệu | Thiết kế Kho Stage,Ware House |  |  |  |
| Code tạo kế Kho Stage, Ware House |  |  |  |
| Tích hợp dữ liệu vào kho (SSIS) | Xử lí dữ liệu |  |  |  |
| Đổ dữ liệu vào kho Stage |  |  |  |
| Đổ dữ liệu từ Stage sang Ware House |  |  |  |
| Phân tích dữ liệu bằng SSAS | Tổng doanh thu của các thể loại game |  |  |  |
| Doanh thu của các thể loại game theo từng năm |  |  |  |
| Doanh thu của các thể loại game theo từng khu vực |  |  |  |
| Doanh thu theo nền tảng |  |  |  |
| Theo từng năm và toàn thị trường |  |  |  |
| Doanh thu theo nền tảng theo từng năm |  |  |  |
| Doanh thu theo |  |  |  |
| Trực quan hóa bằng Power BI | Tổng doanh thu của các thể loại game |  |  |  |
| Tổng doanh thu theo từng năm |  |  |  |
| Tổng doanh thu theo khu vực |  |  |  |