**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

**Logo

Description automatically generated**

**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

***Môn học: Kho dữ liệu và OLAP***

**Đề tài:**

**PHÂN TÍCH DỮ LIỆU KINH DOANH CÁC TRÒ CHƠI VIDEOS GAME TỪ 2016- 2018**

**Giảng viên hướng dẫn: Đỗ Thị Minh Phụng**

**Lớp: IS217.M22.HTCL**

**Sinh viên thực hiện:**

Lê Tiến Vinh MSSV: 19522521

Lê Tuấn Khanh MSSV: 19521681

***TPHCM, 2022***

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

**Logo

Description automatically generated**

**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

***Môn học: Kho dữ liệu và OLAP***

**Đề tài:**

**PHÂN TÍCH DỮ LIỆU KINH DOANH CÁC TRÒ CHƠI VIDEOS GAME TỪ 2016- 2018**

**Giảng viên hướng dẫn: Đỗ Thị Minh Phụng**

**Lớp: IS217.M22.HTCL**

**Sinh viên thực hiện:**

Lê Tiến Vinh MSSV: 19522521

Lê Tuấn Khanh MSSV: 19521681

***TPHCM, 2022***

**MỤC LỤC**

[**LỜI CẢM ƠN** 6](#_Toc99910544)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 8](#_Toc99910545)

[**CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU** 9](#_Toc99910546)

[**I.** **Phát biểu dữ liệu** 9](#_Toc99910547)

[**1.1** **Lý do chọn đề tài** 9](#_Toc99910548)

[**1.2** **Mô tả dữ liệu** 10](#_Toc99910549)

[**1.3** **Thuộc tính của dữ liệu** 10](#_Toc99910550)

[***Chú thích cụ thể:*** 11](#_Toc99910551)

[**1.4 Kho dữ liệu đã thực hiện tiền xử lí** 13](#_Toc99910552)

[**2. Xây dựng kho dữ liệu** 15](#_Toc99910553)

[**2.1 Lược đồ kho dữ liệu ( Star Schema)** 15](#_Toc99910554)

[**2.2 Chi tiết các bảng** 16](#_Toc99910555)

[**2.3 Nội dung 15 câu hỏi truy vấn** 17](#_Toc99910556)

[**CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU ( SSIS )** 18](#_Toc99910557)

[**2.1 Tổng quan mô hình SSIS** 18](#_Toc99910558)

[**2.2 Tạo project SSIS và thiết lập kết nối** 18](#_Toc99910559)

[**2.2.1 Tạo project SSIS** 18](#_Toc99910560)

[**2.2.2 Tạo cơ sở dữ liệu và thiết lập kết nối** 22](#_Toc99910561)

[**a. Tạo cơ sở dữ liệu** 22](#_Toc99910562)

[**b.1 Thiết lập kết nối đến database Datalake\_ Videogames\_Sales** 23](#_Toc99910563)

[**b.2 Thiết lập kết nối đến database DataWarehouse\_VideoGames\_Sales** 28](#_Toc99910564)

[**2.3 Làm sạch và nạp dữ liệu vào Data Lake** 30](#_Toc99910565)

[**Bước 1:** Kéo thả để tạo mới Sequence Container từ thanh công cụ SSIS Toolbox -> Đổi tên thành Clean Data Lake and Load Dataset 31](#_Toc99910566)

[**Bước 2:** Kéo thả công cụ Execute SQL Task vào Container -> Đổi tên thành Clean Data Lake 32](#_Toc99910567)

[**Bước 3:** Chuột phải vào Clean Data Lake -> Edit -> Định nghĩa Connection là tới DataLake\_VideoGames\_Sales -> Chỉnh sửa SQLStatement -> OK 33](#_Toc99910568)

[**Bước 4:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Load Data 35](#_Toc99910569)

[**Bước 5:** Click chuột trái vào Load Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ Flat File Source để định nghĩa dữ liệu nguồn 35](#_Toc99910570)

[**Bước 6:** Click chuột phải và chọn Edit -> Chọn New -> Browse đến đường dẫn chứa file dữ liệu với định dạng csv 36](#_Toc99910571)

[**Bước 7:** Chọn Columns -> Kiểm tra xem tình trạng dữ liệu đã đúng theo mong muốn hay chưa -> Chọn tiếp Advanced -> Chọn Suggest Types để Visual Studio tự động format lại kiểu dữ liệu của các cột -> OK 36](#_Toc99910572)

[**Bước 8:** Click chọn Retain Null để giữ lại các dòng dữ liệu có giá trị Null cho các bước xử lý sau -> OK 39](#_Toc99910573)

[**Bước 9:** Kéo thả công cụ OLE DB Destination để đưa dữ liệu vào bảng Original\_Data -> Đổi tên thành Original\_Data 39](#_Toc99910574)

[**Bước 10:** Click chuột phải vào Original\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Chỉnh sửa data type của các cột thuộc tính theo đúng báo cáo ngay trên đoạn mã đã được tự động sinh ra -> Click chọn OK 40](#_Toc99910575)

[**Bước 11:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 42](#_Toc99910576)

[**Bước 12:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Preprocess Data 43](#_Toc99910577)

[**Bước 13:** Click chuột trái vào Preprocess Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn 43](#_Toc99910578)

[**Bước 14:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Original\_Data 44](#_Toc99910579)

[**Bước 15:** Click chọn Columns để lấy ra những cột cần thiết -> OK 45](#_Toc99910580)

[**Bước 16:** Kéo thả công cụ Conditional Split để lọc dữ liệu 46](#_Toc99910581)

[**Bước 17:** Click chuột phải vào Conditional Split -> Chọn Edit -> Chỉnh sửa Output Name thành Null\_Output -> Thêm điều kiện lọc Null trên toàn bộ cột -> Chỉnh sửa Default Output Name thành NotNull\_Output -> OK 46](#_Toc99910582)

[**Bước 18:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu Null vào -> Đổi tên thành Null\_Data Table -> Nối Data Flow -> Chọn Null\_Output -> OK 48](#_Toc99910583)

[**Bước 19:** Click chuột phải vào Null\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK 49](#_Toc99910584)

[**Bước 20:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 50](#_Toc99910585)

[**Bước 21:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu 51](#_Toc99910586)

[**Bước 22:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK 52](#_Toc99910587)

[**Bước 23:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu đã Clean vào -> Đổi tên thành Clean\_Data Table 53](#_Toc99910588)

[**Bước 24:** Click chuột phải vào Clean\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK 54](#_Toc99910589)

[**Bước 25:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 55](#_Toc99910590)

[Kết quả thu được sau khi hoàn thành định nghĩa các Data Flow Task cho Sequence Container Clean Data Lake and Load Dataset. 56](#_Toc99910591)

[**2.4 Định nghĩa và nạp dữ liệu vào các bảng Dimension** 56](#_Toc99910592)

[**Bước 1:** Kéo thả để tạo mới Sequence Container từ thanh công cụ SSIS Toolbox -> Đổi tên thành ETL data to Dimentsion Table 57](#_Toc99910593)

[**Bước 2:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Dim\_VideoGames 57](#_Toc99910594)

[**Bước 3:** Click chuột trái vào Dim\_VideoGames để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn 58](#_Toc99910595)

[**Bước 4:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data 59](#_Toc99910596)

[**Bước 5:** Click chọn Mappings -> Click chọn Columns để lấy ra những column cần thiết cho bảng Dim\_VideoGames -> OK 59](#_Toc99910597)

[**Bước 6:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu 61](#_Toc99910598)

[**Bước 7:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK 61](#_Toc99910599)

[**Bước 8:** Kéo thả công cụ OLE DB Destination để định nghĩa Dim\_VideoGames-> Đổi tên thành Dim\_VideoGames 62](#_Toc99910600)

[**Bước 9:** Click chuột phải vào Dim\_VideoGames -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK 63](#_Toc99910601)

[**Bước 10:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 64](#_Toc99910602)

[**Bước 11:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Dim\_Date 64](#_Toc99910603)

[**Bước 12:** Click chuột trái vào Dim\_Date để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn 65](#_Toc99910604)

[**Bước 13:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data 66](#_Toc99910605)

[**Bước 14:** Click chọn Columns để lấy ra những column cần thiết cho bảng Dim\_Date -> OK 67](#_Toc99910606)

[**Bước 15:** Kéo thả công cụ Aggerate để Group by dữ liệu, xóa những dữ liệu bị trùng 67](#_Toc99910607)

[**Bước 16:** Click chuột phải vào Aggerate -> Chọn Edit -> Click chọn group by theo cột Year\_of\_Release -> Đổi tên Output Alias thành Year\_of\_Release theo đúng thiết kế kho dữ liệu -> OK 69](#_Toc99910608)

[**Bước 15:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu 70](#_Toc99910609)

[**Bước 20:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK 71](#_Toc99910610)

[**Bước 21:** Kéo thả công cụ OLE DB Destination để định nghĩa Dim\_Date -> Đổi tên thành Dim\_Date 71](#_Toc99910611)

[**Bước 22:** Click chuột phải vào Dim\_Date -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK 72](#_Toc99910612)

[**Bước 23:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 73](#_Toc99910613)

[Kết quả thu được sau khi hoàn thành định nghĩa các Data Flow Task cho Sequence Container ETL data to Dimension Table 73](#_Toc99910614)

[**2.5 Định nghĩa và nạp dữ liệu vào bảng Fact** 74](#_Toc99910615)

[**Bước 1:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Fact\_VideoGames\_Sales 74](#_Toc99910616)

[**Bước 2:** Click chuột trái vào Fact\_VideoGames\_Sales để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn 74](#_Toc99910617)

[**Bước 3:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data 76](#_Toc99910618)

[**Bước 5:** Kéo thả công cụ Lookup để link bảng Fact và các bảng Dim\_VideoGames -> Đổi tên thành VideoGames\_Lookup 77](#_Toc99910619)

[**Bước 6:** Click chuột phải vào VideoGames\_Lookup -> Chọn Edit -> Chuyển specify how to handle rows thành Redirect rows to no match output 78](#_Toc99910620)

[**Bước 7:** Click chọn Connection -> Đổi connection đến Data Warehouse -> Chọn table Dim\_VideoGames 79](#_Toc99910621)

[**Bước 8:** Chọn columns -> Nối cột Name -> OK 80](#_Toc99910622)

[**Bước 9:** Kéo thả công cụ Lookup để link bảng Fact và các bảng Dim\_Date -> Đổi tên thành Date\_Lookup -> Kéo Data Flow -> Chọn Lookup Match Output -> OK 81](#_Toc99910623)

[**Bước 10:** Click chuột phải vào Date\_Lookup -> Chuyển specify how to handle rows thành Redirect rows to no match output 82](#_Toc99910624)

[**Bước 11:** Click chọn Connection -> Đổi connection đến Data Warehouse -> Chọn table Dim\_Date 83](#_Toc99910625)

[**Bước 12:** Chọn columns -> Nối cột Date\_ID ->  Đổi Lookup Operation thành Replace Year\_of\_Release -> OK 84](#_Toc99910626)

[**Bước 13:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu -> Kéo Data Flow -> Chọn Lookup Match Output -> OK 85](#_Toc99910627)

[**Bước 14:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo cột Name, Date\_ID và phương thức là Ascending -> OK 86](#_Toc99910628)

[**Bước 15:** Kéo thả công cụ OLE DB Destination để định nghĩa Fact\_VideoGames\_Sales   -> Đổi tên thành Fact\_VideoGames\_Sales 87](#_Toc99910629)

[**Bước 16:** Click chuột phải vào Fact\_VideoGames\_Sales -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK 88](#_Toc99910630)

[**Bước 17:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK 89](#_Toc99910631)

[**2.6 Định nghĩa và Set các khóa ngoại cho bảng Fact** 89](#_Toc99910632)

[**Bước 1:** Kéo thả công cụ Execute SQL Task vào Container -> Đổi tên thành Set Foreign Keys 89](#_Toc99910633)

[**Bước 2:** Chuột phải vào Set Foreign Keys -> Edit -> Định nghĩa Connection là tới DataWarehouse\_ VideoGames\_Sales -> Chỉnh sửa SQLStatement -> OK 90](#_Toc99910634)

[**2.7 Thực thi Package** 91](#_Toc99910635)

[Control flow 93](#_Toc99910636)

[**2.8 Dữ liệu sau khi hoàn thành** 94](#_Toc99910637)

[Dữ liệu bảng FACT\_VideoGames\_Sales 94](#_Toc99910638)

[Dữ liệu bảng Dim\_VideoGames 94](#_Toc99910639)

[Dữ liệu bảng Dim\_Date 95](#_Toc99910640)

# **LỜI CẢM ƠN**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, đầu tiên nhóm chúng em xin gởi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới Cô Đỗ Thị Minh Phụng - giảng viên lý thuyết môn Kho dữ liệu và OLAP đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn nhóm trong suốt quá trình làm đồ án môn học. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đồ án. Nếu không có những lời hướng dẫn, dạy bảo của cô thì nhóm chúng em nghĩ đồ án này của nhóm rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn Cô. Ngoài ra, để đồ án được hoàn thành thì không thể nào cám ơn những người đã làm ra đó, cám ơn các bạn các thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ.

Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới để ứng dụng vào thực hiện đề tài “Phân tích dữ liệu kinh doanh các trò chơi video games từ năm 2016 - 2018”. Tuy nhiên, do kiến thức và kinh nghiệm còn hạn hẹp, nhóm tác giả sẽ khó tránh khỏi những thiếu sót. Chính vì vậy, nhóm rất mong nhận được những sự góp ý từ phía các Thầy nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn Cô !

Nhóm sinh viên thực hiện

# **NHẬN XÉT CỦA GIẢNG VIÊN**

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

# **CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU**

1. **Phát biểu dữ liệu**
   1. **Lý do chọn đề tài**

Bên cạnh sự phát triển về công nghệ vào thời đại 4.0 như hiện nay, các trò chơi video games ngày càng được giới trẻ cũng như thanh niên trên toàn thế giới yêu thích như một công cụ giải trí vào những thời gian rảnh, hoặc sau những giờ làm việc căng thẳng. Cũng chính vì lí do đó, thị trường dần xuất hiện đa dạng các loại sản phẩm giúp con người có thể nâng cao chất lượng trải nghiệm với trò chơi. Với lượng sản phẩm đa dạng như vậy, nhiều thử thách được đặt ra với các doanh nghiệp hoạt động trong lĩnh vực buôn bán và thương mại, đặc biệt nhất có thể kể tới là việc phải cân nhắc, chọn lọc và đưa ra được những sản phẩm phù hợp với nhu cầu của khách hàng.

A picture containing graphical user interface

Description automatically generated

Do số lượng người chơi và số lượng nhà phát triển ngày càng tăng, việc phân tích dữ liệu là vô cùng cần thiết để có thể có cái nhìn tổng quát về chất lượng cũng như sự phù hợp của trò chơi với từng lứa tuổi. Bên cạnh đó, việc phân tích dữ liệu về các trò chơi video games có sự tham gia của các nhà phê bình cũng như của chính những khách hàng sử dụng trò chơi cũng giúp những nhà phát triển có được cái nhìn, có được số điểm khách quan về trò chơi của mình, qua đó có thể sửa chữa, cải thiện, làm mới và phát triển thế giới trò chơi video games hơn.

* 1. **Mô tả dữ liệu**
* Kho dữ liệu Videos Game Sales with Rating là một kho dữ liệu thu nhập thông tin kinh doanh của các tựa trò chơi video games với rating và các số liệu khác trong khoảng thời gian từ 2016 – 2018.
* Thông qua kho dữ liệu người dùng có thể biết được thông tin về tựa game, nền tảng phát triển của game, năm mà game được phát hành, nhà phát hành tựa game, nhà phát triển game, rating của games ở các quốc gia hàng đầu, và nhiều thông tin khác,…
* Kho dữ liệu gồm 16708 dòng và 16 thuộc tính.
* Link gốc: <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>.
  1. **Thuộc tính của dữ liệu**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên Thuộc Tính** | **Kiểu** | **Ý Nghĩa Thuộc Tính** | **Gía trị** |
| 1 | Video\_Name | string | Tên của trò chơi | Khoảng hơn 60 tựa game. |
| 2 | Platform | string | Nền tảng mà trò chơi sử dụng | 3DS, DC, DS, GBA, GC, PS, PC, PS2, PS3, PS4, PSP, PSV, Wii, WiiU, X360, XB, X1. |
| 3 | Year\_of\_Realease | date | Năm mà trò chơi được phát hành | Trong khoảng 1985 tới 2016 |
| 4 | Gener | string | Hay còn gọi là Category, thể loại của trò chơi. | Action, Adventure, Fighting, Misc, Platform, Puzzel, Racing, Role-Playing, Shooter, Simulation, Sports, Stagery. |
| 5 | Publisher | string | Nhà phát hành | Khoảng hơn 20 nhà phát hành. |
| 6 | NA\_sales | int | Số liệu doanh thu ở khu vực NA.( Bắc Mĩ) | Trong khoảng 0 đến 41.36 |
| 7 | EU\_sales | int | Số liệu doanh thu ở khu vực EU. ( Châu Âu) | Trong khoảng 0 đến 28.96 |
| 8 | JP\_sales | int | Số liệu doanh thu ở Nhật bản | Trong khoảng 0 đến 6.5 |
| 9 | Other sales | int | Các số liệu doanh thu khác | Trong khoảng 0 đến 10.57 |
| 10 | Global sales | int | Số liệu doanh thu toàn cần | Trong khoảng 0 đến 82.53 |
| 11 | Critic\_Score | int | Điểm phê bình của trò chơi | Trong khoảng 13 đến 98 |
| 12 | Critic\_Count | int | Tổng sổ phê bình của trò chơi | Trong khoảng 3 đến 113 |
| 13 | User\_Score | int | Điểm của người dùng cho trò chơi | Trong khoảng 0.5 đến 9.6 |
| 14 | User\_Count | int | Tổng số người dùng tham gia khảo sát | Trong khoảng 4 đến 10665 |
| 15 | Developer | string | Nhà phát triển | Khoảng hơn 20 nhà phát triển trò chơi. |
| 16 | Rating | decimal | Đánh giá trò chơi theo độ tuổi của người chơi( từ trẻ em cho đến người lớn, theo các thang mức E,T, E10+,…) | AO, E, E10+, K-A, M, RP, T. |

***Chú thích cụ thể:***

Khi các bài đánh giá về một bộ phim nhất định được tích lũy, điểm Rotten Tomatoes đo lường tỷ lệ phần trăm tích cực hơn tiêu cực và ấn định đánh giá tổng thể mới hoặc thấp cho bộ phim. Điểm trên 60 phần trăm được coi là tốt, và điểm từ 59 phần trăm trở xuống là xấu.

ESRB sử dụng 7 xếp hạng khác nhau cho các game. 4 xếp hạng trong số này rất phổ biến, trong khi 2 tùy chọn khác khá hiếm gặp và loại xếp hạng cuối cùng chỉ mang tính "giữ chỗ".

* **Early Childhood (EC)** là xếp hạng thấp nhất. Nó biểu thị các game dành cho khán giả nhí. Do đó, những tựa game này không có nội dung phản cảm và có thể không mấy thú vị đối với khán giả nói chung, vì chúng dành cho trẻ nhỏ. Xếp hạng này không phổ biến lắm. Các game ví dụ bao gồm Dora the Explorer: Dance to the Rescue và Bubble Guppies.
* **Everyone (E)** là xếp hạng cơ sở. Các game với xếp hạng này có nội dung “phù hợp với mọi lứa tuổi”. Trước năm 1998, xếp hạng này được gọi là Kids to Adults (KA). Các game được xếp hạng E bao gồm Mario Kart 8 Deluxe và Rocket League.
* **Everyone 10+ (E10+)** biểu thị các game phù hợp cho trẻ em từ 10 tuổi trở lên. So với một game được xếp hạng E, những tựa game này có thể chứa một số nội dung mang tính khêu gợi, hài hước hoặc bạo lực nhiều hơn. Đáng chú ý, đây là xếp hạng duy nhất mà ESRB đã thêm kể từ khi thành lập. Một số game có đánh giá này là Super Smash Bros. Ultimate và Kingdom Hearts III.
* **Teen (T)** là cấp độ tiếp theo. Đánh giá này phù hợp cho người chơi từ 13 tuổi trở lên. Các tiêu đề có thể có nội dung khơi gợi tính dục, ngôn ngữ mạnh hơn và có những cảnh liên quan đến máu me. Bạn sẽ tìm thấy xếp hạng Teen trên các game như Apex Legends và Fortnite.
* **Mature (M)** là xếp hạng “thông thường” ở mức cao nhất. Các game xếp hạng M được coi là phù hợp cho những người từ 17 tuổi trở lên. So với các tựa game Teen, chúng có thể chứa những cảnh bạo lực dữ dội hơn, nội dung tình dục mạnh hơn, thậm chí cả hình ảnh khoả thân và ngôn ngữ khá mạnh. Một số “kho” game không bán các game được xếp hạng M cho trẻ vị thành niên, nhưng đây không phải là một tiêu chuẩn hợp pháp. Các tựa game ví dụ được xếp hạng M bao gồm Red Dead Redemption II và Assassin’s Creed Odyssey.
* **Adults Only (AO)** là xếp hạng 18+ của ESRB. Mức xếp hạng này được áp dụng cho các game có nội dung tình dục đồ họa hoặc những game cho phép đánh bạc bằng tiền thật. Tuy nhiên, nó thực sự là một đánh giá khập khiễng. Không có nhà sản xuất console lớn nào cho phép các game AO trên hệ thống của họ và rất ít nhà bán lẻ bán những game AO trong “kho” game của mình.

Chính vì vậy, chỉ có một số ít các game nhận được mức xếp hạng này. Hầu hết các game AO nhận được đánh giá này do nội dung nặng về tình dục. Các nhà xuất bản sẽ thay đổi game của mình để tránh bị xếp hạng ở mức này, vì về cơ bản, đó giống như một bản án tử hình. Các game có xếp hạng AO bao gồm Seduce Me và Ef: A Fairy Tale of the Two.

* **Rating Pending (RP)** chỉ mang tính “giữ chỗ”. Nó xuất hiện cùng với quảng cáo cho các game chưa được đánh giá.

A screenshot of a computer

Description automatically generated with medium confidence

## **1.4 Kho dữ liệu đã thực hiện tiền xử lí**

* Sau khi dữ liệu được xử lí, sẽ bao gồm 16708 dòng và 16 cột.

Table

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên Thuộc Tính** | **Kiểu** | **Ý Nghĩa Thuộc Tính** | **Gía trị** |
| 1 | Video\_Name | string | Tên của trò chơi | Khoảng hơn 60 tựa game. |
| 2 | Platform | string | Nền tảng mà trò chơi sử dụng | 3DS, DC, DS, GBA, GC, PS, PC, PS2, PS3, PS4, PSP, PSV, Wii, WiiU, X360, XB, X1. |
| 3 | Year\_of\_Realease | date | Năm mà trò chơi được phát hành | Trong khoảng 1985 tới 2016 |
| 4 | Gener | string | Hay còn gọi là Category, thể loại của trò chơi. | Action, Adventure, Fighting, Misc, Platform, Puzzel, Racing, Role-Playing, Shooter, Simulation, Sports, Stagery. |
| 5 | Publisher | string | Nhà phát hành | Khoảng hơn 20 nhà phát hành. |
| 6 | NA\_sales | int | Số liệu doanh thu ở khu vực NA.( Bắc Mĩ) | Trong khoảng 0 đến 41.36 |
| 7 | EU\_sales | int | Số liệu doanh thu ở khu vực EU. ( Châu Âu) | Trong khoảng 0 đến 28.96 |
| 8 | JP\_sales | int | Số liệu doanh thu ở Nhật bản | Trong khoảng 0 đến 6.5 |
| 9 | Other sales | int | Các số liệu doanh thu khác | Trong khoảng 0 đến 10.57 |
| 10 | Global sales | int | Số liệu doanh thu toàn cần | Trong khoảng 0 đến 82.53 |
| 11 | Critic\_Score | int | Điểm phê bình của trò chơi | Trong khoảng 13 đến 98 |
| 12 | Critic\_Count | int | Tổng sổ phê bình của trò chơi | Trong khoảng 3 đến 113 |
| 13 | User\_Score | int | Điểm của người dùng cho trò chơi | Trong khoảng 0.5 đến 9.6 |
| 14 | User\_Count | int | Tổng số người dùng tham gia khảo sát | Trong khoảng 4 đến 10665 |
| 15 | Developer | string | Nhà phát triển | Khoảng hơn 20 nhà phát triển trò chơi. |
| 16 | Rating | decimal | Đánh giá trò chơi theo độ tuổi của người chơi( từ trẻ em cho đến người lớn, theo các thang mức E,T, E10+,…) | AO, E, E10+, K-A, M, RP, T. |

# **2. Xây dựng kho dữ liệu**

## **2.1 Lược đồ kho dữ liệu ( Star Schema)**

Diagram

Description automatically generated

## **2.2 Chi tiết các bảng**

* **Bảng FACT:**

|  |  |  |
| --- | --- | --- |
| Tên thuộc tính | Kiểu dữ liệu | Ràng buộc |
| Fact\_ID | int | Khóa chính |
| Video\_ID | int | Khóa ngoại |
| Year\_of\_Release | string | Khóa ngoại |
| Name | String |  |
| NA\_Sales | Int |  |
| EU\_Sales | Int |  |
| JP\_Sales | Int |  |
| Other\_Sales | Int |  |
| Global\_Sales | Int |  |
| Critic\_Score | Int |  |
| Critic\_Count | Int |  |
| User\_Score | Int |  |
| User\_Count | Int |  |

* **Bảng Dim\_Date :**

|  |  |  |
| --- | --- | --- |
| Tên thuộc tính | Kiểu dữ liệu | Ràng buộc |
| Year\_of\_Realease | string | Khóa chính |

* **Bảng Dim\_VideoGames:**

|  |  |  |
| --- | --- | --- |
| Tên thuộc tính | Kiểu dữ liệu | Ràng buộc |
| Video\_ID | int | Khóa chính |
| Name | String |  |
| Gener | String |  |
| Publisher | String |  |
| Developer | String |  |
| Rating | decimal |  |
| Platform | String |  |

## **2.3 Nội dung 15 câu hỏi truy vấn**

**15 câu truy vấn :**

1. Tìm ra trò chơi thuộc nền tảng PS có đánh giá là M ( mature), được phát hành năm 20xx và đạt mức doanh thu bán ra lớn hơn 5. tại châu âu (EU)
2. Tìm ra tên trò chơi, năm phát hành, nhà sản xuất, nhà phát triển của những tựa game có đánh giá là E10.
3. Tìm nền tảng có trò chơi với đánh giá thuộc E, được phát hành năm 2007 và có số điểm đánh giá phê bình lớn hơn 7, bé hơn 8 và số người đánh giá phê bình >100 người.
4. Tìm ra thông tin top 10 trò chơi có doanh thu lớn nhất và nhỏ nhất trên toàn thể giới.
5. Liệt kê những trò chơi có đánh giá phê bình lớn hơn 50 với số người đánh giá trong khoảng 70-80 người.
6. Đưa ra thông tin tựa game, năm phát hành, nhà xuất bản, nhà phát triển, tổng doanh thu toàn cầu của từng trò chơi thuộc thể loại SPORTS.
7. Đưa ra top 3 trò chơi có mức thu nhập thấp nhất tại Châu Âu.
8. Cho biết thông tin của nhà phát hành nào có tựa game thuộc đánh giá T và sở hữu điểm đánh giá phê bình từ người sử dụng lớn hơn 8.5.
9. Thống kê doanh thu ở toàn thế giới và phần trăm đánh giá từng loại nền tảng trò chơi sắp xếp theo thứ tự phần trăm đánh giá giảm dần.
10. Cho biết thông tin nhà phát triển có ít sản phẩm nhất và có nhiều sản phẩm nhất.
11. Liệt kê những trò chơi thuộc top 5 doanh thu cao nhất.
12. Với mỗi nền tảng, đưa ra top 3 trò chơi có doanh thu cao nhất, tính trên toàn thế giới.
13. Đưa ra doanh thu của những trò chơi được phát hành trong khoảng thời gian từ 1999- 20008 với doanh thu bán ra tại Nhật Bản lớn hơn 1.42.
14. Liệt kê top 10 những trò chơi được phát hành trong năm 2016 có doanh thu cao nhất tại các châu lục còn lại ( tức là trừ Nhật Bản, Châu Âu, và Bắc Mĩ).
15. Liệt kê top 3 nền tảng có những sản phẩm được phát hành năm 2014 có đánh giá thuộc loại T, có đánh giá từ người dùng phải lớn hơn hoặc bằng 7 và tổng doanh thu trên toàn thế giới của nền tảng đó với tổng các trò chơi phải lớn hơn 700.

# **CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU ( SSIS )**

# **2.1 Tổng quan mô hình SSIS**

**Bước 1 - Clean Data Lake and Load Dataset:** Thực hiện xóa toàn bộ dữ liệu trong Data Lake, tiến hành đưa toàn bộ dữ liệu thô vào bảng tương ứng trong Data Lake, thực hiện tiền xử lý dữ liệu và đổ vào các bảng dữ liệu tương ứng

**Bước 2 – ETL data to Dimension Table:** Thực hiện đổ dữ liệu vào các bảng Dimension trong Data warehouse

**Bước 3 – Fact VideoGames\_Sales:** Thực hiện đổ dữ liệu vào bảng Fact trong Data warehouse

**Bước 4 – Set Foreign Keys:** Cài đặt các khóa ngoại tương ứng theo thiết kế lược đồ

# **2.2 Tạo project SSIS và thiết lập kết nối**

## **2.2.1 Tạo project SSIS**

**Bước 1:** Mở visual studio 2019 -> Create a new project

Graphical user interface, application

Description automatically generated

**Bước 2:** Tìm Integration Services Project -> Click chọn và nhấn Next

Graphical user interface, text, application, email

Description automatically generated

**Bước 3:** Đặt tên cho project -> chọn location đặt project -> Create

Graphical user interface, application, email

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Giao diện chính của SSIS project, các chức năng chính nằm ở cột bên trái SSIS Toolbox.

## **2.2.2 Tạo cơ sở dữ liệu và thiết lập kết nối**

### **a. Tạo cơ sở dữ liệu**

Bước 1: Vào MS SQL Server Management -> tạo New Query

Graphical user interface, application, Word

Description automatically generated

**Bước 2:** Viết lệnh tạo database cần thiết -> Run

Graphical user interface, text, application

Description automatically generated

Kết quả sau khi chạy lệnh tạo database

Graphical user interface, text, application

Description automatically generated

Datalake\_Videogames\_Sales là database chứa các table sau

* Original\_Data: Table chứa dữ liệu gốc, dữ liệu chưa qua các bước tiền xử lý
* Clean\_Data: Table chứa dữ liệu đã trải qua các bước tiền xử lý (lọc Null, đổi tên column, thêm column,…)
* Null\_Data: Table chứa những dòng dữ liệu chứa giá trị Null được lọc từ bảng Original\_Data

DataWarehouse\_Videogames\_Sales là database đóng vai trò là một kho dữ liệu, chứa các table theo đúng thiết kế star schema ở mục Xây dựng lược đồ kho dữ liệu

* Dim\_VideoGames: Bảng chiều chứa dữ liệu thông tin về trò chơi video games.
* Dim\_Date: Bảng chiều chứa dữ liệu thông tin về ngày tháng năm
* Fact\_Videogames\_Sales: Bảng fact chứa thông tin và các thuộc tính dùng để tính toán

b. Thiết lập kết nối

### **b.1 Thiết lập kết nối đến database Datalake\_ Videogames\_Sales**

**Bước 1:** Mở Project SSIS trong Visual Studio -> Chuột phải Connection Managers -> New Connection Manager

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

Description automatically generated

**Bước 2:** Chọn OLEDB -> Nhấn Add

Graphical user interface, text, application

Description automatically generated

**Bước 3:** Chọn New -> Vào MS SQL Server Management để lấy được Server Name -> Điền Server Name -> Tại bước Select database name, chọn Datalake\_VideoGames\_Sales -> Chọn Test Connection -> Chọn OK nếu test connection thành công

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Kết quả thu được

Graphical user interface, text

Description automatically generated

### **b.2 Thiết lập kết nối đến database DataWarehouse\_VideoGames\_Sales**

**Bước 1:** Thực hiện tương tự 2.2.2.1

**Bước 2:** Thực hiện tương tự 2.2.2.1

**Bước 3:** Chọn New -> Vào MS SQL Server Management để lấy được Server Name -> Điền Server Name -> Tại bước Select database name, chọn DataWarehouse\_VideoGames\_Sales -> Chọn Test Connection -> Chọn OK nếu test connection thành công

Graphical user interface, application

Description automatically generated

Kết quả thu được:

Graphical user interface, text

Description automatically generated

**2.3 Làm sạch và nạp dữ liệu vào Data Lake**

**Mục tiêu:** Xóa tất cả dữ liệu đã được insert vào 3 bảng (Original\_Data, Clean\_Data, Null\_Data) trong Datalake\_VideoGames\_Sales, tiến hành nạp dữ liệu vào bảng Original\_Data, làm sạch dữ liệu và nạp dữ liệu vào 2 bảng (Clean\_Data, Null\_Data) tương ứng.

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

Description automatically generated

### **Bước 1:** Kéo thả để tạo mới Sequence Container từ thanh công cụ SSIS Toolbox -> Đổi tên thành Clean Data Lake and Load Dataset

Graphical user interface, text

Description automatically generated

### **Bước 2:** Kéo thả công cụ Execute SQL Task vào Container -> Đổi tên thành Clean Data Lake

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

Description automatically generated

### **Bước 3:** Chuột phải vào Clean Data Lake -> Edit -> Định nghĩa Connection là tới DataLake\_VideoGames\_Sales -> Chỉnh sửa SQLStatement -> OK

Graphical user interface, application

Description automatically generated

Câu lệnh SQL Statements

Graphical user interface, text, application, email

Description automatically generated

### **Bước 4:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Load Data

Graphical user interface, application

Description automatically generated

### **Bước 5:** Click chuột trái vào Load Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ Flat File Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 6:** Click chuột phải và chọn Edit -> Chọn New -> Browse đến đường dẫn chứa file dữ liệu với định dạng csv

Graphical user interface, application

Description automatically generated

### **Bước 7:** Chọn Columns -> Kiểm tra xem tình trạng dữ liệu đã đúng theo mong muốn hay chưa -> Chọn tiếp Advanced -> Chọn Suggest Types để Visual Studio tự động format lại kiểu dữ liệu của các cột -> OK

Thực hiện ở Columns:

Graphical user interface, application

Description automatically generated

Thực hiện ở Advanced:

Graphical user interface, application

Description automatically generated

### **Bước 8:** Click chọn Retain Null để giữ lại các dòng dữ liệu có giá trị Null cho các bước xử lý sau -> OK

Graphical user interface, text, application

Description automatically generated

### **Bước 9:** Kéo thả công cụ OLE DB Destination để đưa dữ liệu vào bảng Original\_Data -> Đổi tên thành Original\_Data

A screenshot of a computer

Description automatically generated

### **Bước 10:** Click chuột phải vào Original\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Chỉnh sửa data type của các cột thuộc tính theo đúng báo cáo ngay trên đoạn mã đã được tự động sinh ra -> Click chọn OK

Graphical user interface, application

Description automatically generated

### **Bước 11:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, table

Description automatically generated

### **Bước 12:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Preprocess Data

Graphical user interface, application

Description automatically generated

### **Bước 13:** Click chuột trái vào Preprocess Data để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 14:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Original\_Data

Graphical user interface, application

Description automatically generated

### **Bước 15:** Click chọn Columns để lấy ra những cột cần thiết -> OK

Graphical user interface, table

Description automatically generated

### **Bước 16:** Kéo thả công cụ Conditional Split để lọc dữ liệu

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 17:** Click chuột phải vào Conditional Split -> Chọn Edit -> Chỉnh sửa Output Name thành Null\_Output -> Thêm điều kiện lọc Null trên toàn bộ cột -> Chỉnh sửa Default Output Name thành NotNull\_Output -> OK

Graphical user interface, application

Description automatically generated

Condition :

ISNULL( [Name] ) || ISNULL( [Platform] ) || ISNULL( [Year\_of\_Release] ) || ISNULL( [Genre] ) || ISNULL( [Publisher] ) || ISNULL( [NA\_Sales] ) || ISNULL( [EU\_Sales] ) || ISNULL( [JP\_Sales] ) || ISNULL( [Other\_Sales] ) || ISNULL( [Global\_Sales] ) || ISNULL( [Critic\_Score] ) || ISNULL( [Critic\_Count] ) || ISNULL( [User\_Score] ) || ISNULL( [User\_Count] ) || ISNULL( [Developer] ) || ISNULL( [Rating] )

### **Bước 18:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu Null vào -> Đổi tên thành Null\_Data Table -> Nối Data Flow -> Chọn Null\_Output -> OK

A screenshot of a computer

Description automatically generated

### **Bước 19:** Click chuột phải vào Null\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK

Graphical user interface, application

Description automatically generated

### **Bước 20:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, table

Description automatically generated

### **Bước 21:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu

Diagram

Description automatically generated

### **Bước 22:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK

Table

Description automatically generated

### **Bước 23:** Kéo thả công cụ OLE DB Destination để định nghĩa bảng sẽ đưa dữ liệu đã Clean vào -> Đổi tên thành Clean\_Data Table

Diagram

Description automatically generated

### **Bước 24:** Click chuột phải vào Clean\_Data Table -> Chọn Edit -> Chọn connection đến Data Lake -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Click chọn OK

Graphical user interface, application

Description automatically generated

### **Bước 25:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, table

Description automatically generated

### Kết quả thu được sau khi hoàn thành định nghĩa các Data Flow Task cho Sequence Container Clean Data Lake and Load Dataset.

Graphical user interface, application

Description automatically generated

Sau khi chạy thử ta được kết quả như sau:

A screenshot of a computer

Description automatically generated with medium confidence

## **2.4 Định nghĩa và nạp dữ liệu vào các bảng Dimension**

**Mục tiêu:** Nạp dữ liệu đã được tiền xử lý nằm trong bảng Clean\_Data của Data Lake vào các bảng Dimension tương ứng theo đúng thiết kế lược đồ kho dữ liệu

### **Bước 1:** Kéo thả để tạo mới Sequence Container từ thanh công cụ SSIS Toolbox -> Đổi tên thành ETL data to Dimentsion Table

Rectangle

Description automatically generated with low confidence

### **Bước 2:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Dim\_VideoGames

Graphical user interface, text, application

Description automatically generated

### **Bước 3:** Click chuột trái vào Dim\_VideoGames để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated

### **Bước 4:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data

Graphical user interface, text, application

Description automatically generated

### **Bước 5:** Click chọn Mappings -> Click chọn Columns để lấy ra những column cần thiết cho bảng Dim\_VideoGames -> OK

Graphical user interface

Description automatically generated

### **Bước 6:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 7:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK

Table

Description automatically generated

### **Bước 8:** Kéo thả công cụ OLE DB Destination để định nghĩa Dim\_VideoGames-> Đổi tên thành Dim\_VideoGames

A screenshot of a computer

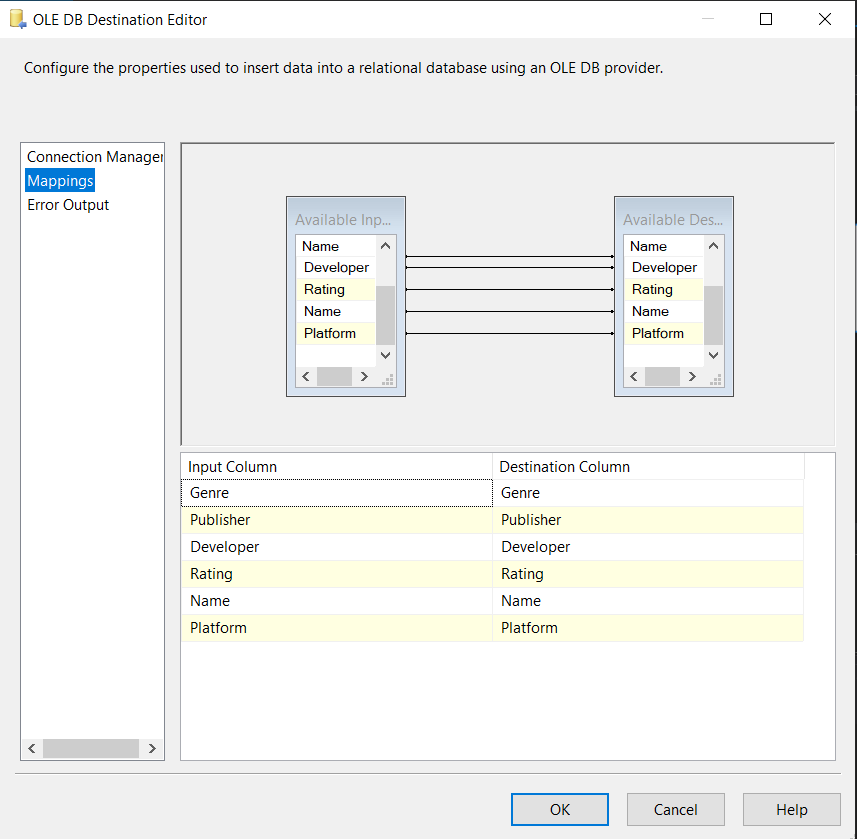
Description automatically generated with medium confidence

### **Bước 9:** Click chuột phải vào Dim\_VideoGames -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK

Graphical user interface, application

Description automatically generated

### **Bước 10:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK



### **Bước 11:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Dim\_Date

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

Description automatically generated

### **Bước 12:** Click chuột trái vào Dim\_Date để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 13:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data

Graphical user interface, text, application

Description automatically generated

### **Bước 14:** Click chọn Columns để lấy ra những column cần thiết cho bảng Dim\_Date -> OK

Graphical user interface

Description automatically generated

### **Bước 15:** Kéo thả công cụ Aggerate để Group by dữ liệu, xóa những dữ liệu bị trùng

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 16:** Click chuột phải vào Aggerate -> Chọn Edit -> Click chọn group by theo cột Year\_of\_Release -> Đổi tên Output Alias thành Year\_of\_Release theo đúng thiết kế kho dữ liệu -> OK

Graphical user interface, application

Description automatically generated

### **Bước 15:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu

Diagram

Description automatically generated

### **Bước 20:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo tất cả các cột và phương thức là Ascending -> OK

Graphical user interface, table

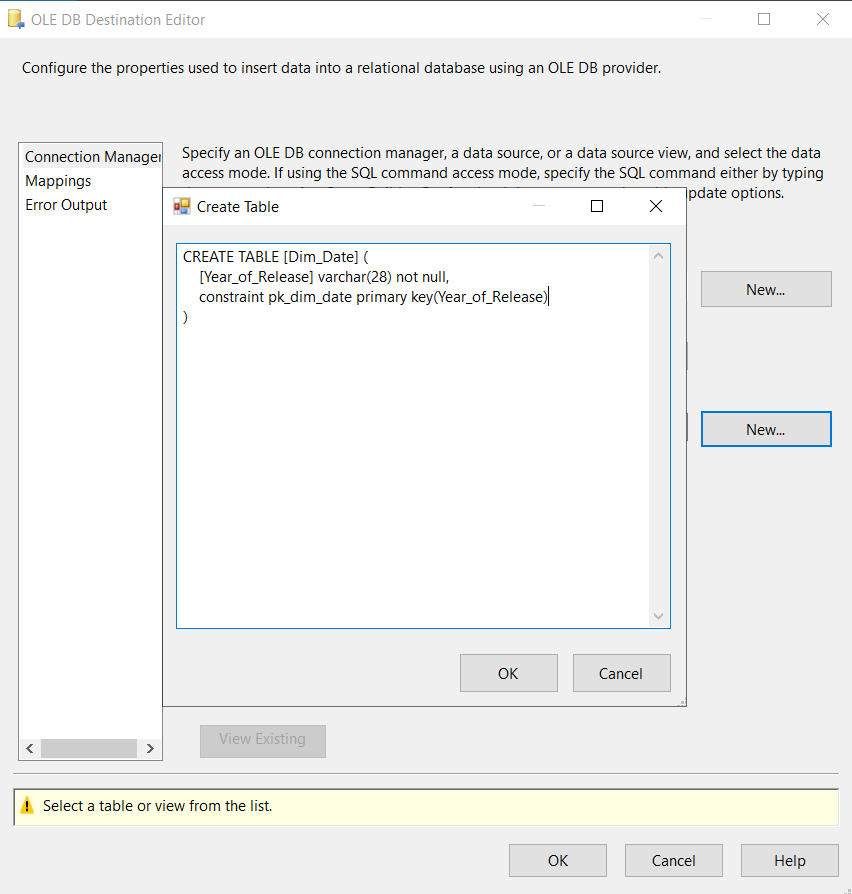
Description automatically generated

### **Bước 21:** Kéo thả công cụ OLE DB Destination để định nghĩa Dim\_Date -> Đổi tên thành Dim\_Date

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 22:** Click chuột phải vào Dim\_Date -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK



### **Bước 23:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Graphical user interface, application

Description automatically generated

### Kết quả thu được sau khi hoàn thành định nghĩa các Data Flow Task cho Sequence Container ETL data to Dimension Table

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

Description automatically generated

## **2.5 Định nghĩa và nạp dữ liệu vào bảng Fact**

**Mục tiêu:** Tiến hành nạp dữ liệu vào bảng Fact\_ VideoGames\_Sales theo đúng thiết kế tương ứng

**Bước 1:** Kéo thả công cụ Data Flow Task vào Container -> Đổi tên thành Fact\_VideoGames\_Sales

Graphical user interface, application

Description automatically generated

**Bước 2:** Click chuột trái vào Fact\_VideoGames\_Sales để chỉnh sửa Data Flow Task -> Kéo thả công cụ OLE DB Source để định nghĩa dữ liệu nguồn

A screenshot of a computer

Description automatically generated with medium confidence

### **Bước 3:** Click chuột phải và chọn Edit -> Chọn connection đến Data Lake -> Chọn table Clean\_Data

Graphical user interface, text, application

Description automatically generated

**Bước 4:** Click chọn Mappings -> Click chọn Columns để lấy ra những column cần thiết cho bảng Fact\_VideoGames\_Sales -> OK

Graphical user interface

Description automatically generated

### **Bước 5:** Kéo thả công cụ Lookup để link bảng Fact và các bảng Dim\_VideoGames -> Đổi tên thành VideoGames\_Lookup

Diagram

Description automatically generated with medium confidence

### **Bước 6:** Click chuột phải vào VideoGames\_Lookup -> Chọn Edit -> Chuyển specify how to handle rows thành Redirect rows to no match output

Graphical user interface, text, application, email

Description automatically generated

### **Bước 7:** Click chọn Connection -> Đổi connection đến Data Warehouse -> Chọn table Dim\_VideoGames

Graphical user interface, application

Description automatically generated

### **Bước 8:** Chọn columns -> Nối cột Name -> OK

Graphical user interface

Description automatically generated

### **Bước 9:** Kéo thả công cụ Lookup để link bảng Fact và các bảng Dim\_Date -> Đổi tên thành Date\_Lookup -> Kéo Data Flow -> Chọn Lookup Match Output -> OK

A screenshot of a computer

Description automatically generated

### **Bước 10:** Click chuột phải vào Date\_Lookup -> Chuyển specify how to handle rows thành Redirect rows to no match output

Graphical user interface, text, application, email

Description automatically generated

### **Bước 11:** Click chọn Connection -> Đổi connection đến Data Warehouse -> Chọn table Dim\_Date

Graphical user interface, application

Description automatically generated

### **Bước 12:** Chọn columns -> Nối cột Date\_ID ->  Đổi Lookup Operation thành Replace Year\_of\_Release -> OK

Graphical user interface, application, Word

Description automatically generated

### **Bước 13:** Kéo thả công cụ Sort để sắp xếp lại các dòng dữ liệu -> Kéo Data Flow -> Chọn Lookup Match Output -> OK

Graphical user interface, text, application

Description automatically generated

### **Bước 14:** Click chuột phải vào Sort -> Chọn Edit -> Click chọn sort theo cột Name, Date\_ID và phương thức là Ascending -> OK

Graphical user interface

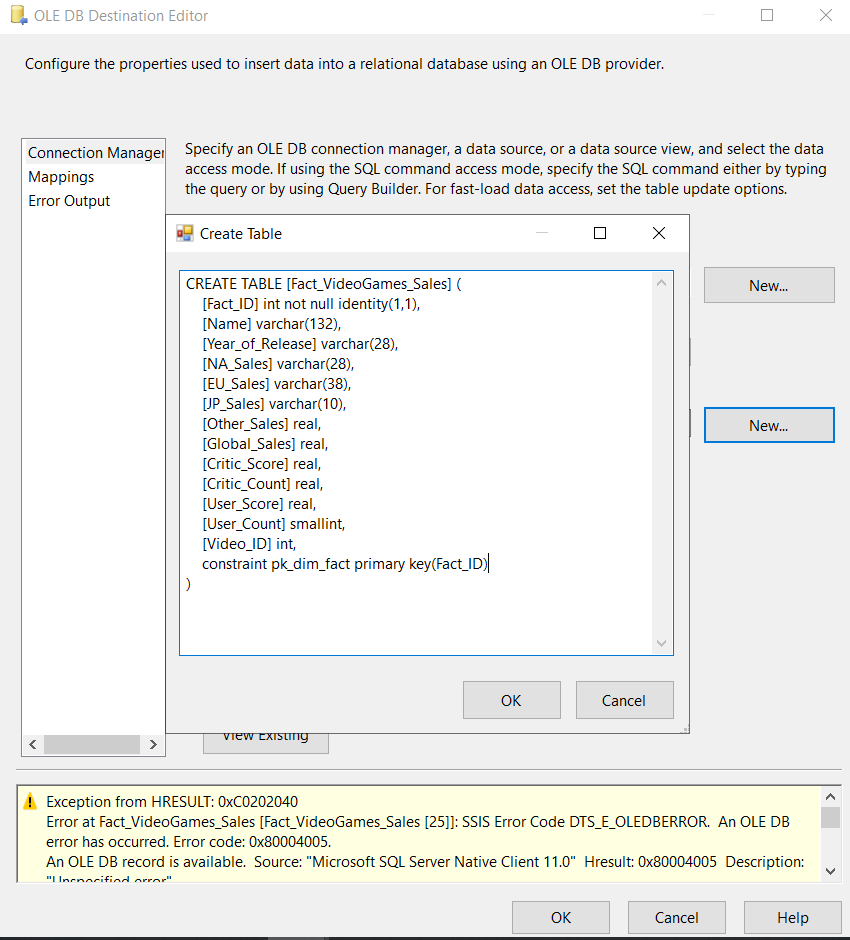
Description automatically generated

### **Bước 15:** Kéo thả công cụ OLE DB Destination để định nghĩa Fact\_VideoGames\_Sales   -> Đổi tên thành Fact\_VideoGames\_Sales

Graphical user interface, application

Description automatically generated

### **Bước 16:** Click chuột phải vào Fact\_VideoGames\_Sales -> Chọn Edit -> Chọn connection đến Data Warehouse -> Chọn New để SSIS tự sinh ra đoạn mã insert data vào table -> Thêm primary key ngay trên đoạn mã được sinh tự động -> OK



### **Bước 17:** Click chọn Mappings -> Check xem những columns đã được map đúng hay chưa -> OK

Table

Description automatically generated

## **2.6 Định nghĩa và Set các khóa ngoại cho bảng Fact**

**Mục tiêu:** Sau khi đã nạp dữ liệu vào bảng Fact, tiến hành tạo các khóa ngoại theo đúng thiết kế

**Bước 1:** Kéo thả công cụ Execute SQL Task vào Container -> Đổi tên thành Set Foreign Keys

Graphical user interface, application

Description automatically generated

### **Bước 2:** Chuột phải vào Set Foreign Keys -> Edit -> Định nghĩa Connection là tới DataWarehouse\_ VideoGames\_Sales -> Chỉnh sửa SQLStatement -> OK

Graphical user interface, text, application

Description automatically generated

Câu lệnh SQL query:

Graphical user interface, text, application

Description automatically generated

## **2.7 Thực thi Package**

Graphical user interface, application

Description automatically generated

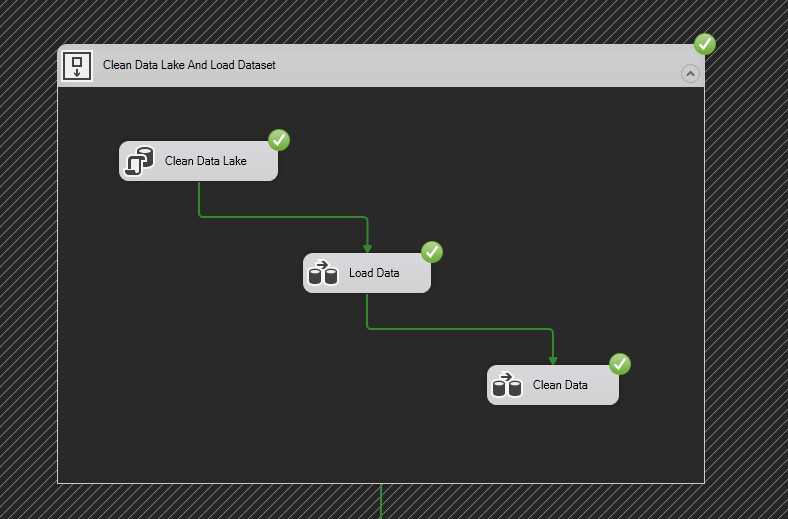
Graphical user interface, application

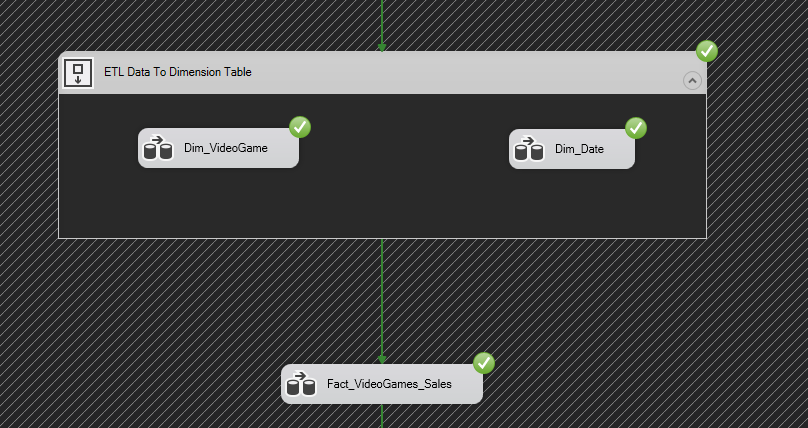
Description automatically generated

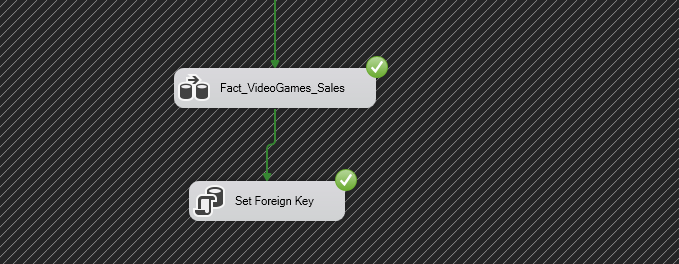
A picture containing graphical user interface

Description automatically generated

## Control flow







## **2.8 Dữ liệu sau khi hoàn thành**

### Dữ liệu bảng FACT\_VideoGames\_Sales

Graphical user interface, application

Description automatically generated

### Dữ liệu bảng Dim\_VideoGames

Graphical user interface, application

Description automatically generated

### Dữ liệu bảng Dim\_Date

Graphical user interface, application

Description automatically generated

**Link drive source SSIS và video :** https://drive.google.com/drive/u/1/folders/1dUnryapOTlKMiEHD5-NhL3P8mmjFstj5?fbclid=IwAR2Tw81\_qF4F827taC8BBmYq7pt5eURv3GPWz69-zm3588\_NrMYF4dKluco