Lời cảm ơn

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn đến Thầy – Th.S Nguyễn Văn Thành – Giảng viên phụ trách môn Kho Dữ Liệu – Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh vì những đóng góp của thầy trong thời gian qua. Thầy đã cung cấp những kiến thức, hướng dẫn, đóng góp, xây dựng trong quá trình học để nhóm em có thể hoàn thành đồ án môn học một cách tốt nhất.

Kho dữ liệu là một trong những kiến thức đang ngày càng phổ biến cùng với OLAP, nhóm chúng em đã lên ý tưởng thực hiện xây dựng một kho dữ liệu cùng với các quy trình nghiệp vụ. Nhóm chúng em quyết định xây dựng đồ án “Xây dựng và khai thác kho dữ liệu về lĩnh vực hàng không”. Trong quá trình thực hiện đồ án, vận dụng những kiến thức học được từ lý thuyết đến thực hành trong quá trình học tập, cùng với sự tìm hiểu các kiến thức, công cụ mới, nhóm chúng em đã cố gắng hoàn thiện đồ án một cách tốt nhất.

Nếu có những chỗ chưa hoàn thiện, nhóm chúng em mong nhận được sự đóng góp từ thầy nhằm cải thiện và coi đó là một kinh nghiệm để có thể hoàn thành tốt hơn.

Nhóm chúng em một lần nữa xin chân thành cảm ơn thầy!

Lời Nhận Xét Của Giảng Viên

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………….………………………………………………………………………………..………

Chương 1: Lý do chọn đề tài và giới thiệu tổng quan Dataset

* 1. Lý do hình thành đồ án

1.1.1 Vấn đề đặt ra

Ngành hàng không là một trong những ngành trọng điểm đối với thế giới cũng như Việt Nam. Trải qua lịch sử hơn 100 năm đầy biến động, ngành hàng không đã có những biến động cả về công nghệ lẫn dịch vụ.

Từ năm 1994 đến năm 2002 nổi lên với sự phát triển của mô hình hàng không giá rẻ, có ý nghĩa quan trọng với ngành hàng không ngày nay. Các công ty con và trung tâm phân phối được mở rộng toàn thế giới cùng với sự phát triển của thương mại điện tử cho các hệ thống bán vé máy bay trực tuyến.

Tháng 12/2019, tâm dịch Covid-19 nhanh chóng lan rộng ra toàn thế giới khiến tất cả các nước phải áp dụng các biện pháp hạn chế di chuyển, điểu này đặc biệt nghiêm trọng cho ngành hàng không thế giới. American Airlines cắt giảm 75% số chuyến bay quốc tế, Korean Air giảm 80% lịch trình,… tóm lại, hầu như các chuyến bay trên thế giới đều bị cắt giảm đáng kể.

Trong 2 năm vừa qua, ngành hàng không quốc tế cũng như Việt Nam đang có dấu hiệu phục hồi trở lại

Các kiến thức về ngành hàng không rất đa dạng như thông tin về các hãng hàng không trên toàn thế giới, các sân bay, chuyến bay,… và phải kể đến quy trình đặt vé máy bay hay còn gọi là booking. Những data kể trên rất phù hợp để xây dựng một kho dữ liệu về lịch trình chuyến bay cũng như booking.

Vì vậy, nhóm chúng em quyết định xây dựng một kho dữ liệu về các chuyến bay và đặt vé máy bay của ngành hàng không thế giới

<https://thinksmart.com.vn/hon-100-nam-ra-doi-va-phat-trien-nganh-hang-khong/>

1.1.2 Giải pháp

Dựa trên nhu cầu phân tích, thống kê và khai thác dữ liệu về ngành hàng không thế giới. Giải pháp được đưa ra là xây dựng một kho dữ liệu nhằm mục đích khai thác, phân tích và tạo báo cáo tổng về các chuyến bay và booking.

1.1.3 Mục tiêu và ý nghĩa dự án

Đầu tiêu là xây dựng một kho dữ liệu và OLAP

Data Warehouse hay kho dữ liệu lưu trữ data từ nhiều nguồn, giúp tăng cường hiệu suất báo cáo và phân tích. Dựa vào điều đó cũng như các kiến thức đã học được trong quá trình học tập, nhóm chúng em quyết định xây dựng một kho dữ liệu OLAP để thấy được hiệu quả của nó.

Thứ hai là khai thác và phân tích data về hàng không trong thực tế. Khai thác các thông tin của hàng không quốc tế nhằm có cái nhìn chung về kiến thức cũng như quy trình vận hành cơ bản của một hệ thống đặt vé máy bay.

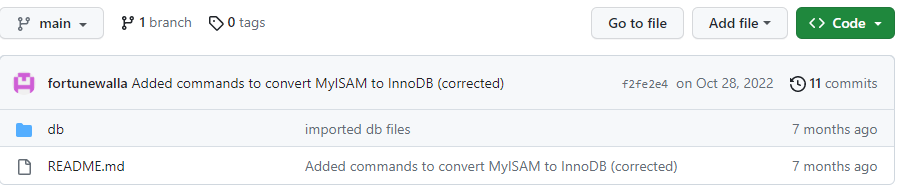
* 1. Giới thiệu tổng quan dataset

1.2.1 Nguồn dữ liệu sử dụng

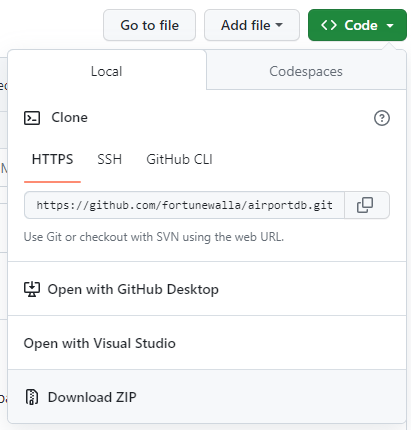
Nguồn dataset được lấy từ github.com/fortunewalla/airportdb

1.2.2 Tiền xử lý dữ liệu

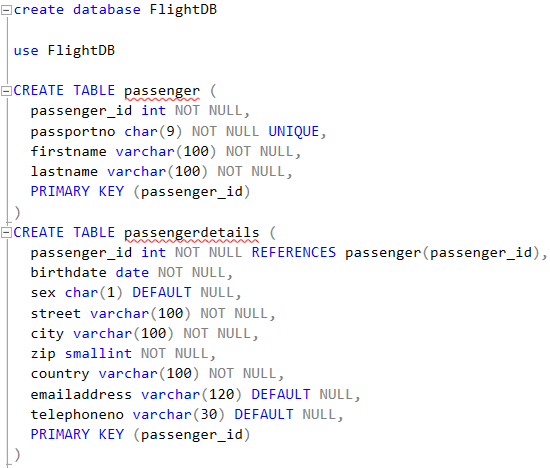
Vào đường dẫn github.com/fortunewalla/airportdb

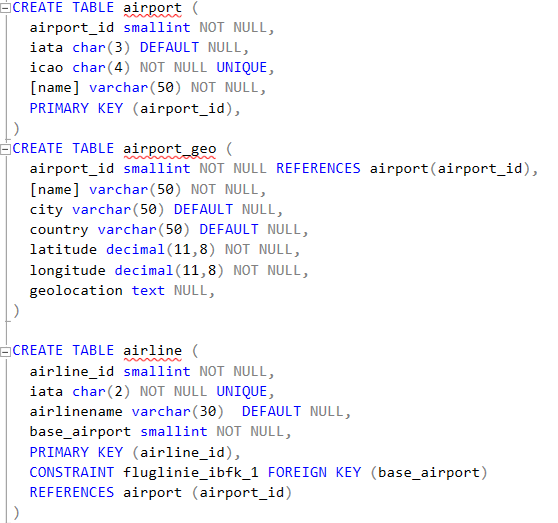


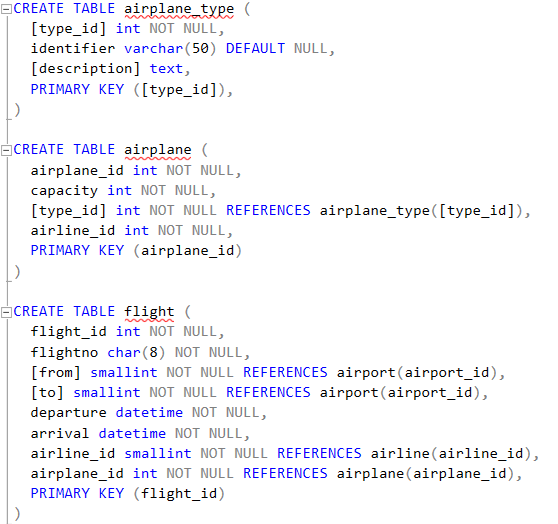
Nhấn vào **Code** rồi **Download ZIP** để tải toàn bộ tài liệu về dataset

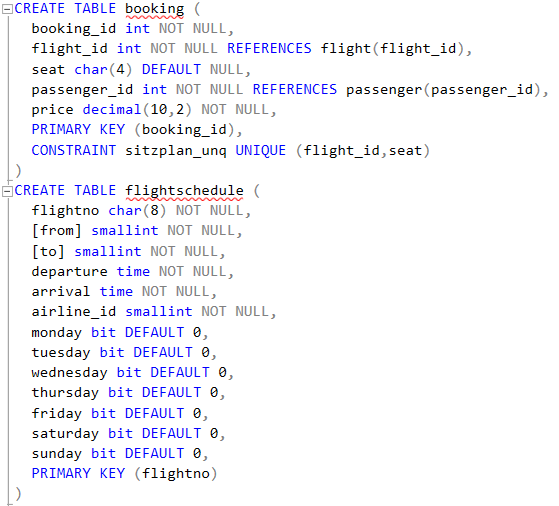


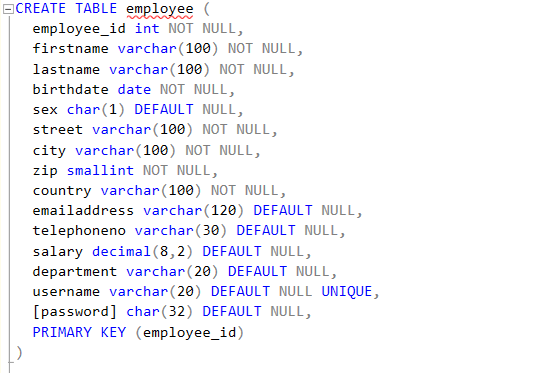
Vào SQL Server, tạo **New Query**, lần lượt tạo database và các table



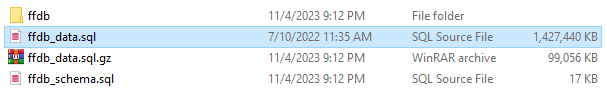




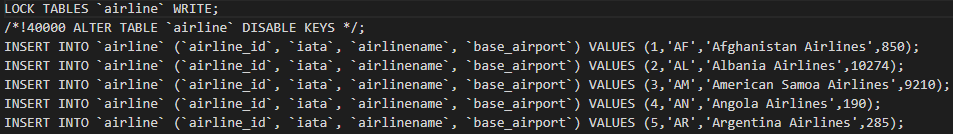




Mở file ffdb\_data.sql



Copy những câu lệnh insert tương ứng với các table



1.2.2 Mô tả chi tiết dữ liệu

1.2.2.1 Tổng quát

Dataset có tổng cộng 12 tables với hơn 450.000 dòng mô tả các hoạt động của các chuyến bay trên toàn thế giới vào năm 2015, mô tả hoạt động đặt vé máy bay cùng với một tập khách hàng, mô tả các sân bay, hãng hàng không, máy bay trên toàn thế giới, ngoài ra dataset còn thống kê các nhân viên làm việc cho các phòng ban khác nhau cùng với mức lương của họ.

1.2.2.2 Mô tả các bảng dataset

|  |  |
| --- | --- |
| Tên bảng | Mô tả |
| airport | Mô tả các sân bay theo Hiệp hội Vận tải Hàng không Quốc tế IATA và theo Tổ chức Hàng không Dân dụng Quốc tế ICAO |
| airport\_geo | Mô tả chi tiết thành phố, đất nước, vị trí các sân bay |
| airline | Mô tả các hãng hàng không theo Hiệp hội Vận tải Hàng không Quốc tế IATA |
| airplane | Mô tả các máy bay thuộc các hãng hàng không |
| airplane\_type | Mô tả các kiểu máy bay |
| flight | Mô tả hoạt động các chuyến bay bao gồm số hiệu chuyến bay, nơi đi, nơi đến, giờ cất cánh, hạ cánh,… |
| flight\_schedule | Mô tả lịch trình bay của các chuyến bay |
| booking | Mô tả việc đặt vé máy bay của các khách hàng |
| passenger | Mô tả khách hàng |
| passenger\_detail | Mô tả chi tiết thông tin khách hàng |
| employee | Mô tả thông tin nhân viên |

1.2.3 Các công cụ được sử dụng

1.2.3.1 Visual Studio 2022

Visual studio là một phần mềm hỗ trợ đắc lực hỗ trợ công việc lập trình website. Công cụ này được tạo lên và thuộc quyền sở hữu của ông lớn công nghệ Microsoft. Năm 1997, phần mềm lập trình nay có tên mã Project Boston. Nhưng sau đó, Microsoft đã kết hợp các công cụ phát triển, đóng gói thành sản phẩm duy nhất.

Phần mềm lập trình Visual Studio của Microsoft hỗ trợ sử dụng trên nhiều nền tảng khác nhau. Không giống như các trình viết code khác, Visual Studio sử dụng được trên cả Windows, Linux và Mac Systems. Điều này cực kỳ tiện lợi cho lập trình viên trong quá trình ứng dụng.

Không chỉ hỗ trợ đa nền tảng, Visual Studio cũng cho phép sử dụng nhiều ngôn ngữ lập trình khác nhau từ C#, F#, C/C++, HTML, CSS, Visual Basic, JavaScript,… Bởi vậy, Visual Studio có thể dễ dàng phát hiện và thông báo cho bạn khi các chương trình có lỗi.

Mặc dù Visual Studio có hệ thống các ngôn ngữ hỗ trợ lập trình khá đa dạng. Nhưng nếu lập trình viên muốn sử dụng một ngôn ngữ khác, bạn có thể dễ dàng tải xuống các tiện ích mở rộng. Tính năng hấp dẫn này được hoạt động như một phần chương trình độc lập nên không lo làm giảm hiệu năng của phần mềm.

Cho đến nay, nhà phát triển Microsoft vẫn đang không ngừng cải tiến và ứng dụng các công nghệ mới cho Visual Studio. Bởi vậy, đây là “cánh tay đắc lực” hỗ trợ cho các developers, data administrators, testers và bất kỳ ai có nhu cầu tiếp cận, tìm hiểu về lập trình đều có thể sử dụng Visual Studio.

https://visualcpp.net/visual-studio-la-gi-gioi-thieu-phan-mem-visual-studio/

1.2.3.2 Microsoft SQL Server 2019 và Management Studio 18

SQL Server là một hệ quản trị cơ sở dữ liệu quan hệ (Relational Database Management System (RDBMS) ) sử dụng câu lệnh SQL (Transact-SQL) để trao đổi dữ liệu giữa máy Client và máy cài SQL Server. Một RDBMS bao gồm databases, database engine và các ứng dụng dùng để quản lý dữ liệu và các bộ phận khác nhau trong RDBMS.

Microsoft đã xây dựng SQL Server 2019 dựa trên các tính năng xử lý truy vấn thông minh đã được triển khai trong các phiên bản SQL Server trước đó. Đồng thời, Microsoft cũng đã mở rộng SQL Server sang các phần mới của công cụ cơ sở dữ liệu với những ưu điểm nổi bật hơn.

SQL Server 2019 mang đến các tính năng tuân thủ và bảo mật sáng tạo, hiệu suất hàng đầu trong ngành, tính khả dụng quan trọng của sứ mệnh và phân tích nâng cao cho tất cả khối lượng công việc dữ liệu của bạn, hiện đã có hỗ trợ cho dữ liệu lớn được tích hợp sẵn, giúp người dùng tiết kiệm rất nhiều thời gian khi bắt đầu một dự án.

SSMS là một ứng dụng phần mềm thiết kế bởi Microsoft, ra mắt lần đầu năm 2005. Ứng dụng này cho phép lập trình viên cấu hình, quản lý và quản trị bộ máy cơ sở dữ liệu (database engine) SQL Server.

https://microsoft.fptcloud.com/kien-thuc/uu-diem-sql-server-2019/

1.2.3.3 SQL Server Integration Services (SSIS)

SSIS là viết tắt của SQL Server Integration Services – một công cụ tích hợp và quản lý dữ liệu đi kèm với cơ sở dữ liệu Microsoft SQL Server.

SSIS giúp lưu trữ dữ liệu linh hoạt và nhanh chóng hơn với các tính năng như 1 công cụ ETL (trích xuất lượng lớn dữ liệu từ nhiều nguồn dữ liệu trong các định dạng khác nhau, chỉnh sửa và làm sạch dữ liệu sau đó tải chúng vào kho dữ liệu đã được thiết kế).

Một trong những tính năng vượt trội nhất của SSIS là khả năng trích xuất và đồng bộ dữ liệu nhanh chóng và linh hoạt. SSIS cung cấp một nền tảng thuận tiện để di chuyển dữ liệu một cách dễ dàng từ nguồn này sang nguồn khác.

Có một số thành phần cốt lõi tạo nên SSIS GUI (Đồ họa giao diện người dùng) trong cấu trúc SSIS: Control Flow (Luồng điều khiển), Data Flow (Luồng dữ liệu), Event Handlers (Xử lý sự kiện), Package Explorer

SSIS mang một tính linh hoạt cao giống như một công cụ ETL hay một công cụ chuyển đổi dữ liệu. Việc kết nối giao diện người dùng với cấu hình SSIS vô cùng đơn giản. SSIS cung cấp một số tính năng cho phép bạn tạo tài liệu mới khi bạn chuyển đổi dữ liệu ban đầu.

<https://blog.cole.vn/sql-server-integration-services-ssis/>

1.2.3.4 SQL Server Analysist Services (SSAS)

SSAS trong là một trong những dịch vụ chính của SQL Server dùng để xây dựng các chiều và cubes cho kho dữ liệu, trong phiên bản mới này còn hỗ trợ một số thuật toán khai phá dữ liệu điều này cung cấp cho người quản lý có cái nhìn sâu sắc về dữ liệu của họ hơn. SSAS là một phần của nền tảng quản lý doanh nghiệp thông minh (BI), nó không chỉ là một thành phần của SQL Server, nó còn được sử dụng trên .NET Framework và môi trường phát triển Visual Studio.

Dịch vụ SSAS của SQL Server cung cấp cho người dùng các tính năng mạnh mẽ trong phân tích dữ liệu. Đây là công cụ mạnh để xây dựng các hệ thống xử lý giao dịch trực tuyến.

Chương 2: Thiết kế kho dữ liệu dùng kỹ thuật mô hình hóa dữ liệu theo chiều (Dimensional Modeling)

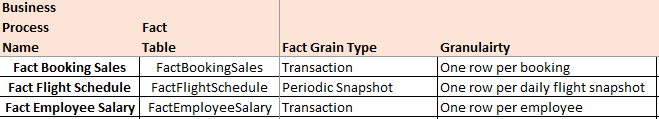
2.1 Thiết kế kho dữ liệu mức cao (High Level Design)

2.1.1 Xây dựng Business Process

Các yêu cầu business đặt ra:

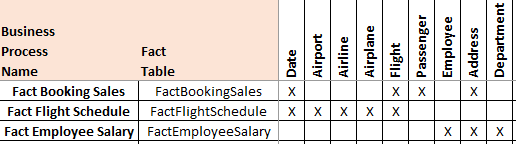
* Báo cáo doanh số đặt vé máy bay (Booking Sales Reporting): theo dõi doanh số bán vé máy bay (booking) dựa trên khách hàng (passenger), chuyến bay (flight) để biết được rằng khách hàng nào đặt vé nhiều nhất, chuyến bay nào được đặt nhiều nhất.
* Lịch trình chuyến bay (Flight Schedule Monitoring): ghi lại hoạt động của các chuyến bay(flight) theo ngày để biết được rằng chuyến bay bay nhiều nhất, hãng hàng không (airline) nào có nhiều chuyến bay nhất, máy bay (airplane) nào bay nhiều nhất, điểm đến (airport) nhiều nhất.
* Phân tích lương nhân viên (Employee Salary Analysis): thống kê lương nhân viên (employee) theo tháng làm việc tại ngành hàng không để biết được rằng nhân viên thuộc phòng ban (department) nào được trả lương cao nhất.

2.1.2 Xác định mức độ chi tiết (Grain) cho các business process



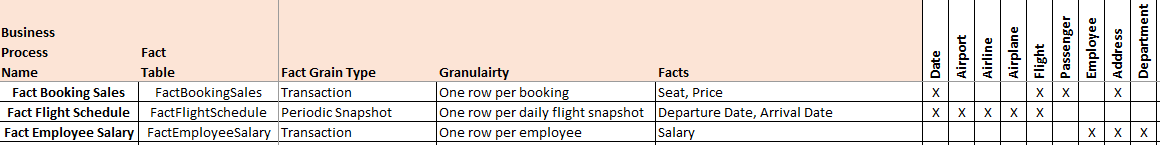
* Hoạt động Booking Sales Reporting gắn với Fact Booking Sales ghi lại mỗi dòng thể hiện việc bán một vé máy bay, thuộc loại Transaction.
* Hoạt động Flight Schedule Monitoring gắn với Fact Flight Schedule ghi mỗi dòng cho hoạt động bay của một chuyến bay, thuộc loại Periodic Snapshot.
* Hoạt động Employee Salary Analysis gắn với Fact Employee Salary ghi lại mỗi dòng với mỗi lương của một nhân viên, thuộc loại Transaction.

2.1.3 Xác định các Dimension



* Fact Booking Sales liên quan đến bảng Flight, Passenger, phân cấp khách hàng theo bảng Address, bảng Date.
* Fact Flight Schedule liên quan đến bảng Airport, Airline, Airplane, Flight và Date.
* Fact Employee Salary liên quan đến bảng Employee, phân cấp Employee theo Address, Department.

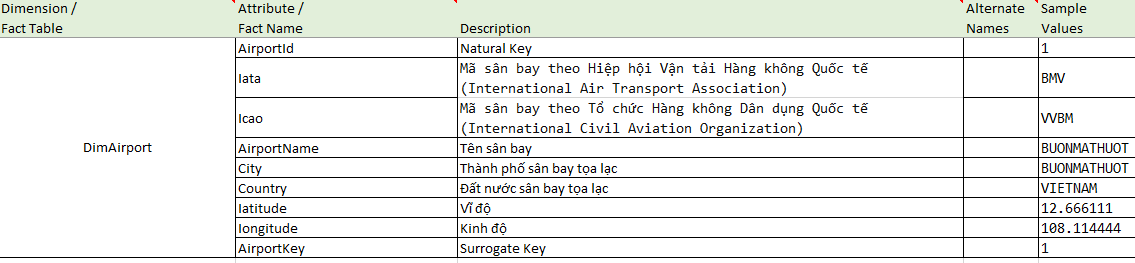
2.1.4 Xác định các Fact Measure

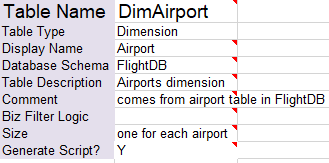


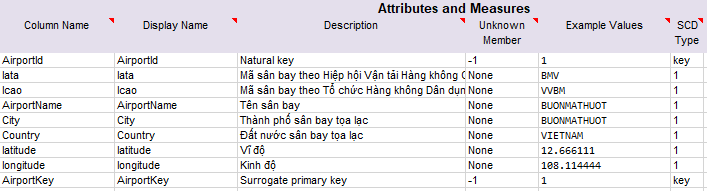
* Fact Booking Sales ghi lại Seat (ghế ngồi) và Price (giá vé theo USD) của một khách hàng khi đặt vé.
* Fact Flight Schedule ghi lại Departure Date (ngày bay) và Arrival Date (ngày hạ cánh) của một chuyến bay
* Fact Employee Salary ghi lại Salary (lương theo tháng) ứng với mỗi nhân viên

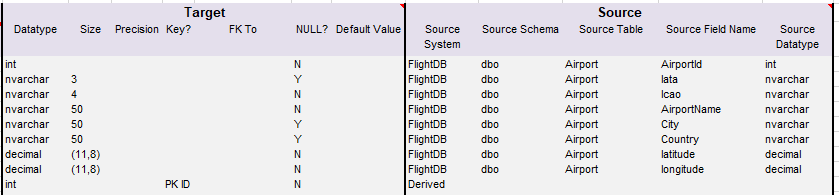
2.2 Thiết kế kho dữ liệu mức chi tiết (Detailed Level Design)

2.2.1 Mô tả các bảng Dimension

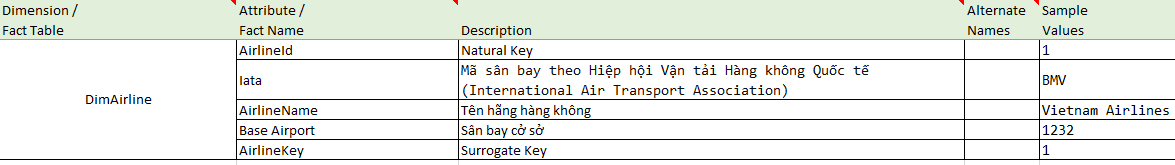


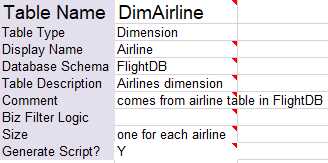


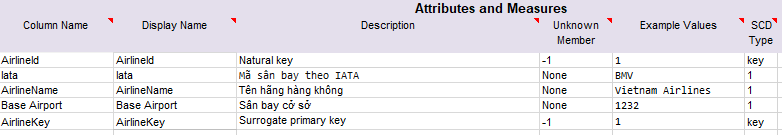


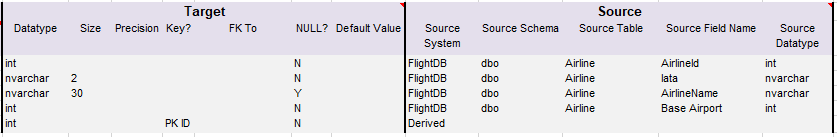


* Bảng DimAirport gồm các trường AirportId, Iata, Icao, AirportName, City, Country, Iatitude, Iongitude, AirportKey với mô tả và giá trị ví dụ

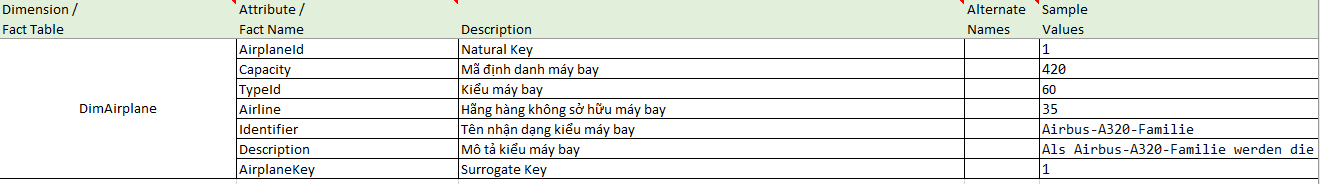


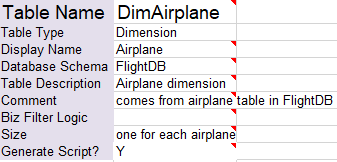


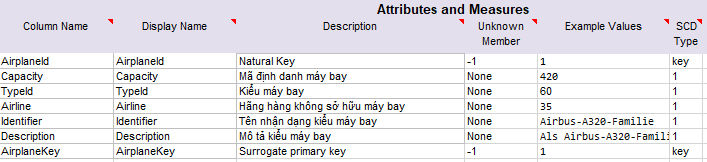


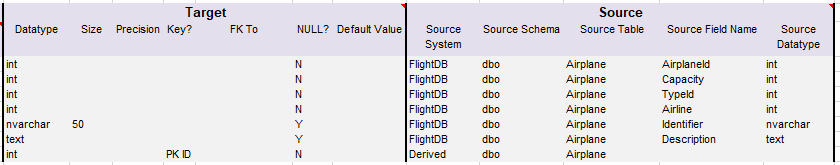


* Bảng DimAirline gồm các trường AirlineId, Iata, AirlineName, BaseAirport, AirlineKey với mô tả và giá trị ví dụ

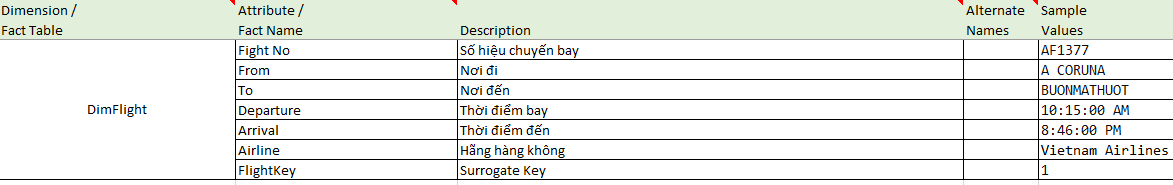


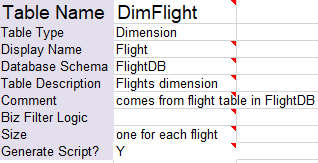


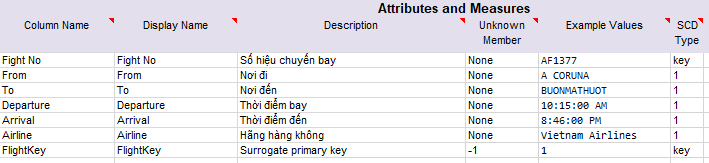


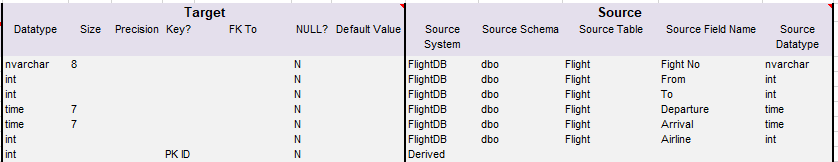


* Bảng DimAirplane gồm các trường AirplaneId, Capacity, TypeId, Airline, Identifier, Description, AirplaneKey với mô tả và giá trị ví dụ

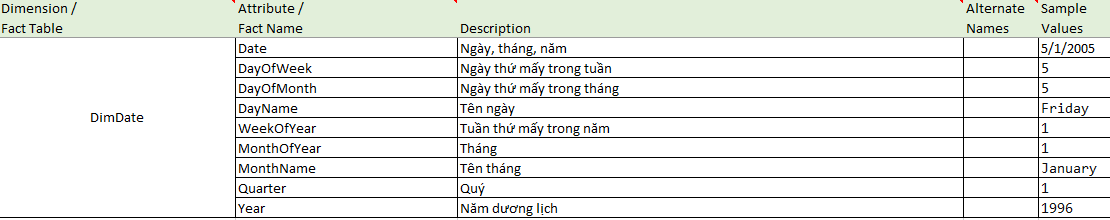


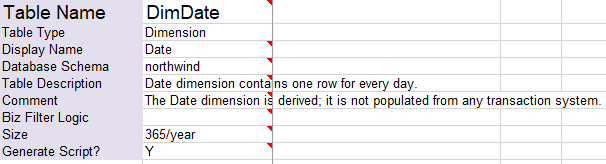


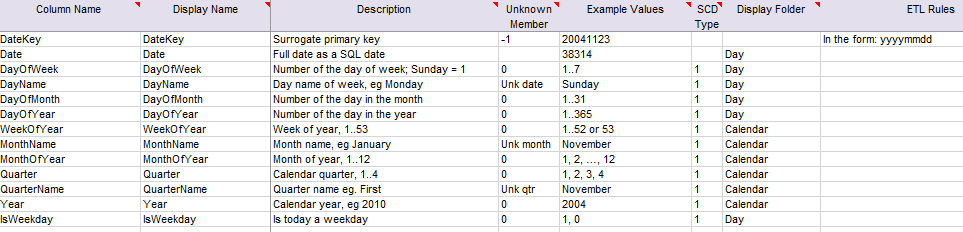


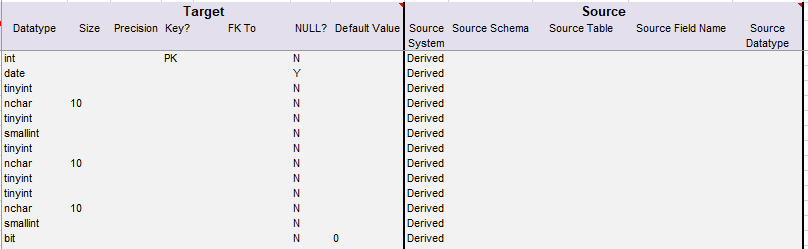


* Bảng DimFlight gồm các trường FlightNo, From, To, Departure, Arrival, Airpline, FlightKey với mô tả và giá trị ví dụ

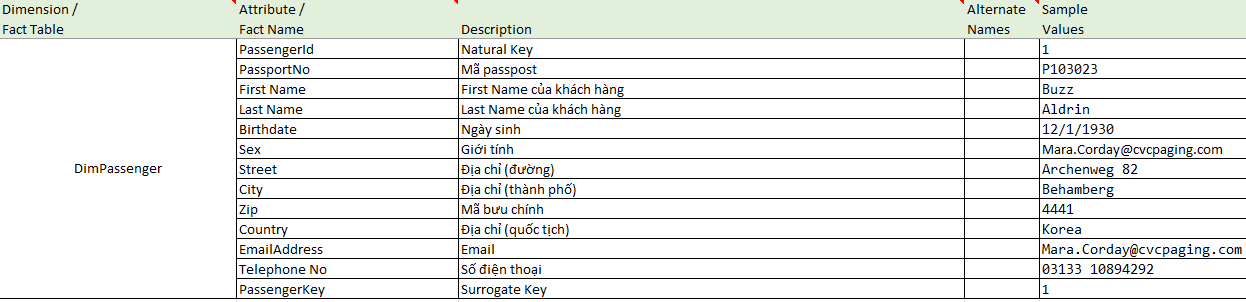


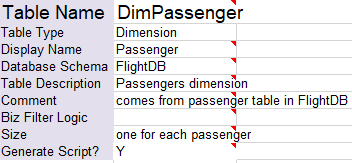


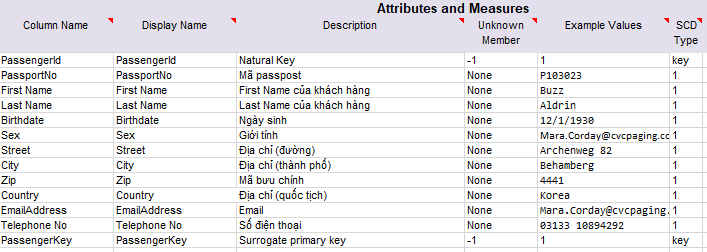


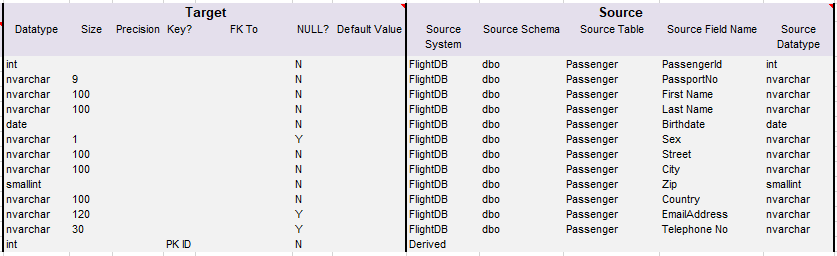


* Bảng DimDate mô tả một dòng là một ngày

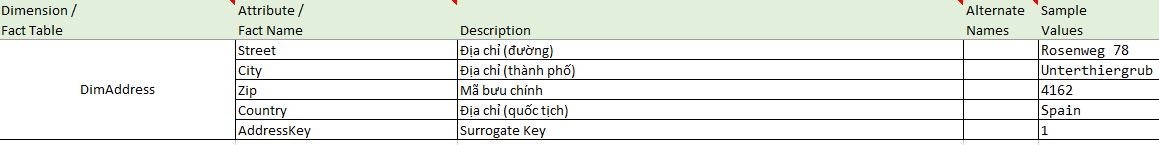


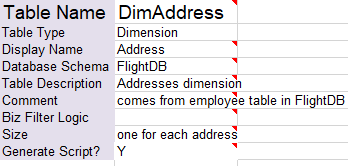


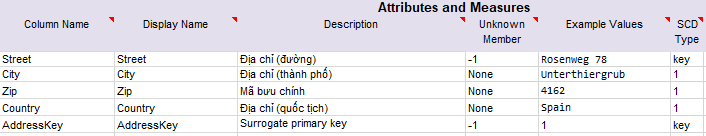


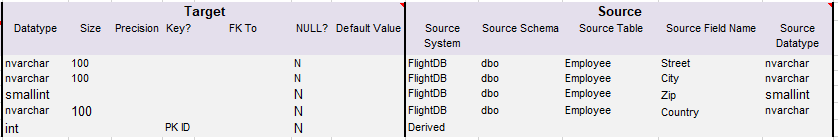


* Bảng DimPassenger gồm các trường PassengerId, PassportNo, FirstName, LastName, Birthdate, Sex, Street, City, Zip, Country, EmailAddress, TelephoneNo, PassengerKey với mô tả và giá trị ví dụ

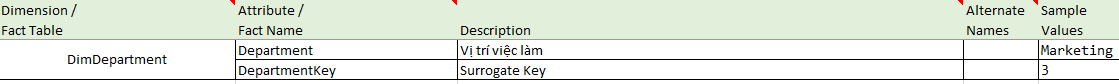


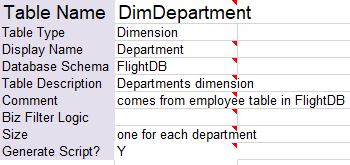


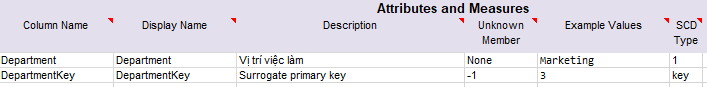


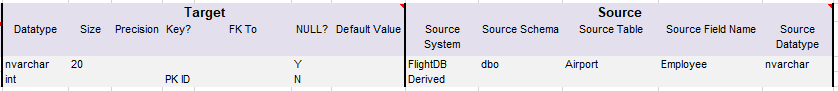


* Bảng DimAddress gồm các trường Street, City, Zip, Country, AddressKey với mô tả và giá trị ví dụ

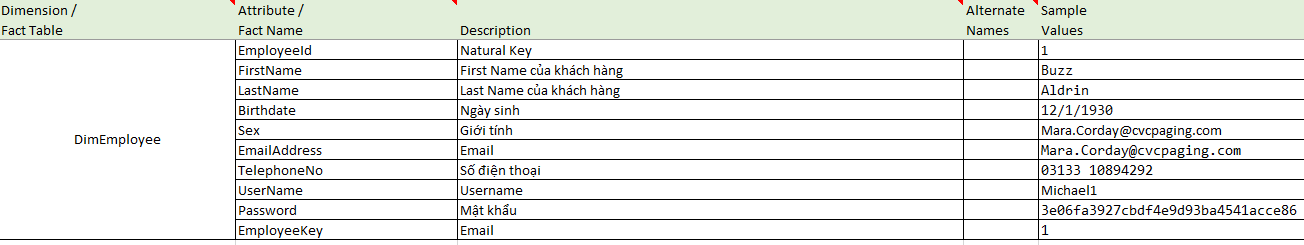


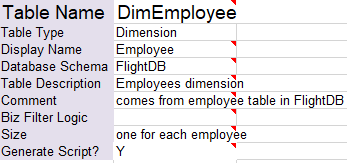


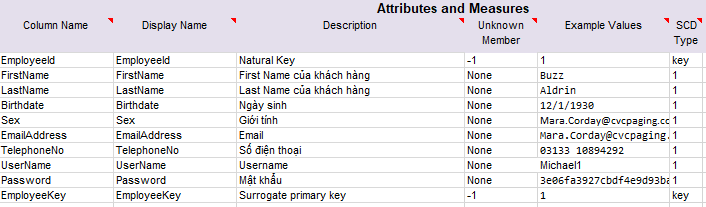


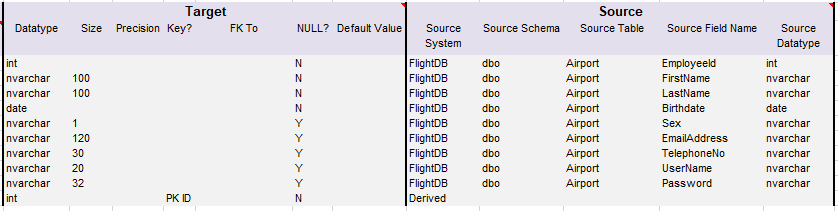


* Bảng DimDepartment gồm các trường Department, DepartmentKey với mô tả và giá trị ví dụ



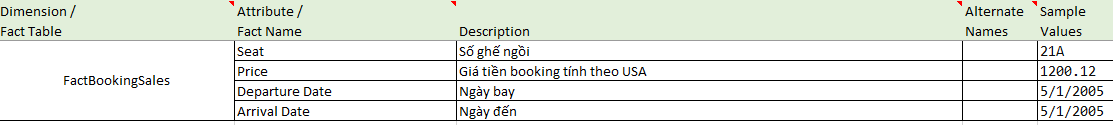


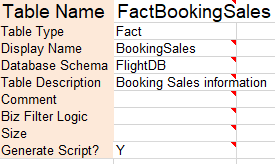


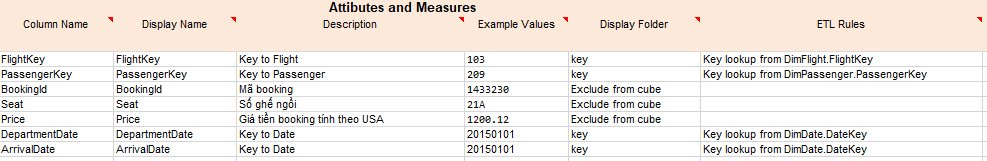


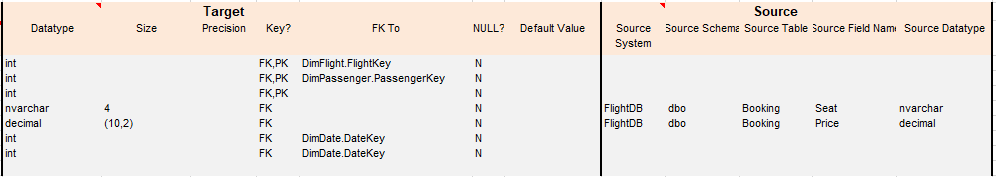
* Bảng DimEmployee gồm các trường EmployeeId, FirstName, LastName, BirthDate, Sex, EmailAddress, TelephoneNo, UserName, Password, EmployeeKey với mô tả và giá trị ví dụ

2.2.2 Mô tả các bảng Fact

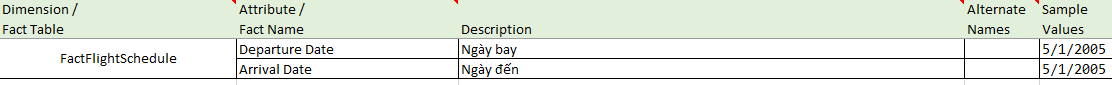


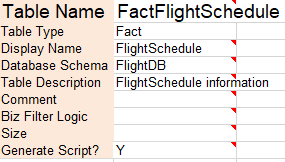


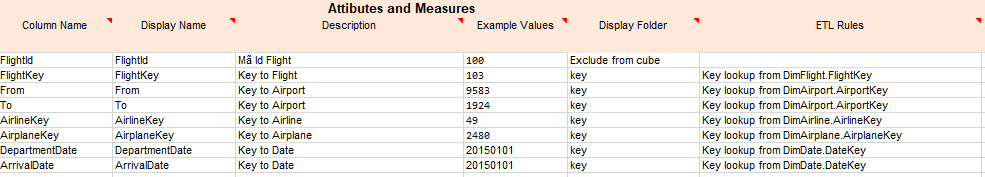


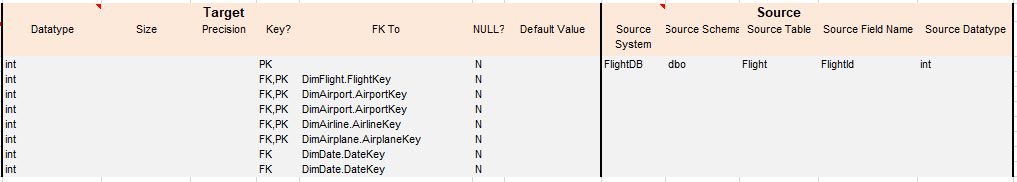


* FactBookingSales ghi lại Seat, Price, DepartureDate, ArrivalDate cho mỗi booking

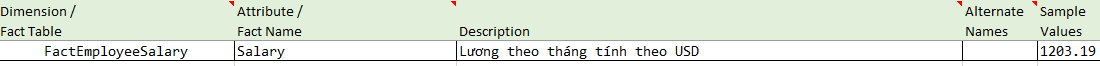


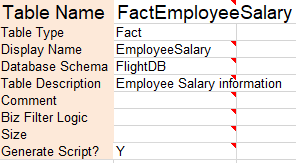


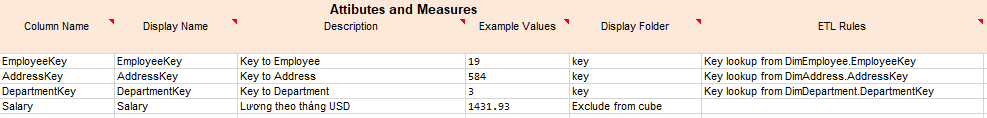


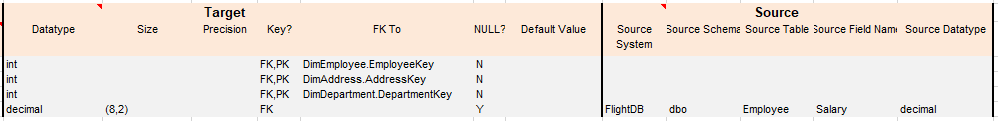


* FactFlightSchedule ghi lại DepartureDate và ArrivalDate cho mỗi chuyến bay









* FactEmployeeSalary ghi lại Salary cho mỗi nhân viên

Chương 3: Tích hợp dữ liệu vào kho (SSIS)

Chương 4: Phân tích dữ liệu (SSAS)

Chương 5: Report

Chương 6: Kết luận

6.1 Kết quả đạt được

Trong quá trình học tập cho tới thời điểm hoàn thành đồ án, nhóm đã tìm hiểu và ứng dụng những kiến thức về kho dữ liệu và OLAP để đạt được những kết quả:

* Ôn lại kiến thức về câu lệnh truy vấn SQL và hệ quản trị cơ sở dữ liệu SQL Server
* Hiểu rõ các khái niệm cơ bản về một kho dữ liệu Datawarehouse
* Hiểu được quy trình xây dựng một kho dữ liệu theo kỹ thuật mô hình hóa theo chiều Dimensional Modeling từ level cao đến level chi tiết
* Xây dựng được quy trình tích hợp dữ liệu vào kho thông qua câu lệnh truy vấn SQL và công cụ hỗ trợ SSIS
* Phân tích được những câu hỏi, vấn đề được đặt ra trong kho dữ liệu đã xây dựng
* Từ kho dữ liệu giúp nắm rõ và hình dung được một hệ thống quản lý bán vé máy bay trên thực tế diễn ra như thế nào

6.2 Hạn chế

Trong quá trình làm đồ án cho tới khi hoàn thành, nhóm nhận thấy có những hạn chế như sau

6.3 Phân công nhiệm vụ

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Công việc | Chi tiết | Nguyễn Khoa Quang Thắng | Nguyễn Hữu Thắng | Đỗ Trọng Phi | Hoàng Uyên | Mức độ hoàn thành |
| Tiền xử lí Dataset | Tạo database và các table |  |  |  |  |  |
|  | Tạo scripts và thêm data từ dataset |  |  |  |  |  |
|  | Xử lí các giá trị lỗi và không cần thiết |  |  |  |  |  |
| Thiết kế kho dữ liệu | Đặt ra yêu cầu business và xác định mức độ chi tiết grain, đưa ra các fact measure |  |  |  |  |  |
|  | Xác định các dimension |  |  |  |  |  |
|  | Thiết kế chi tiết definition, target, source các dimension |  |  |  |  |  |
|  | Thiết kế chi tiết definition, target, source các fact |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Tích hợp dữ liệu vào kho SSIS |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Phân tích dữ liệu SSAS |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Tạo report |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Xây dựng báo cáo |  |  |  |  |  |  |