

# Xử lý dữ liệu

Môn: Phát triển ứng dụng hệ thống thông tin  
hiện đại

GV: Ths. Phạm Minh Tú



# Nội dung

- **Giới thiệu**
- Hướng dẫn sử dụng SSIS
- Chương trình tích hợp
- Một số lưu ý khi viết câu truy vấn



# Giới thiệu

- Bất kì ứng dụng nào cũng cần dữ liệu để báo cáo, khai thác ra quyết định phục vụ doanh nghiệp.
- Quá trình xử lý dữ liệu nền bao gồm các giai đoạn sau:
  - Phân tích – thiết kế - cài đặt mô hình dữ liệu
  - ETL - **Extracts - Transforms - Load**



# Giới thiệu

- **Extracts dữ liệu** - tức là đi thu gom dữ liệu từ nhiều nguồn khác nhau - doanh nghiệp của bạn sẽ có một vài phần mềm với mỗi phần mềm đảm nhiệm một công việc nào đó như quản trị nhân sự (HCM), quản lý quan hệ khách hàng (CRM) và đây là công việc đi thu gom dữ liệu từ các nguồn của các phần mềm này



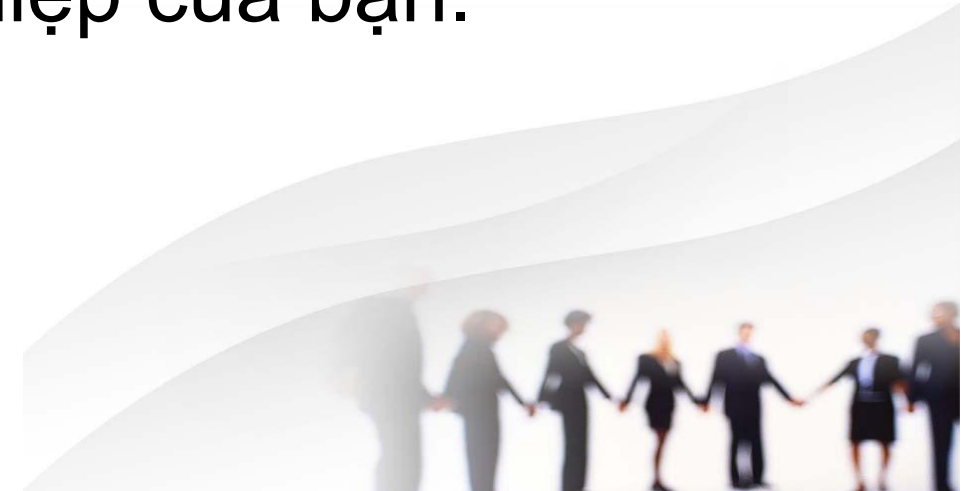
# Giới thiệu

- **Transforms dữ liệu** - tức là chuyển đổi dữ liệu, việc chuyển đổi này có mục đích hẳn hoi, đó là chuyển đổi từ các **dữ liệu nghiệp vụ** của các phần mềm thành **dữ liệu phân tích** của các nhà quản trị, đồng thời phải tối ưu hóa cho mục đích phân tích dữ liệu này. Ngoài ra, chuyển đổi dữ liệu còn tham gia vào một mục đích khác nữa là làm sạch dữ liệu



# Giới thiệu

- **Load** dữ liệu - như bạn thấy ở hình trên, sau khi được chuyển đổi thì toàn bộ các dữ liệu này được đưa vào một nơi lưu trữ mới, mà người ta gọi là DataWarehouse (tạm dịch là kho dữ liệu). Và đến đây là kết thúc giai đoạn ETL dữ liệu, giai đoạn đầu tiên để bạn triển khai giải pháp Business Intelligence cho doanh nghiệp của bạn.

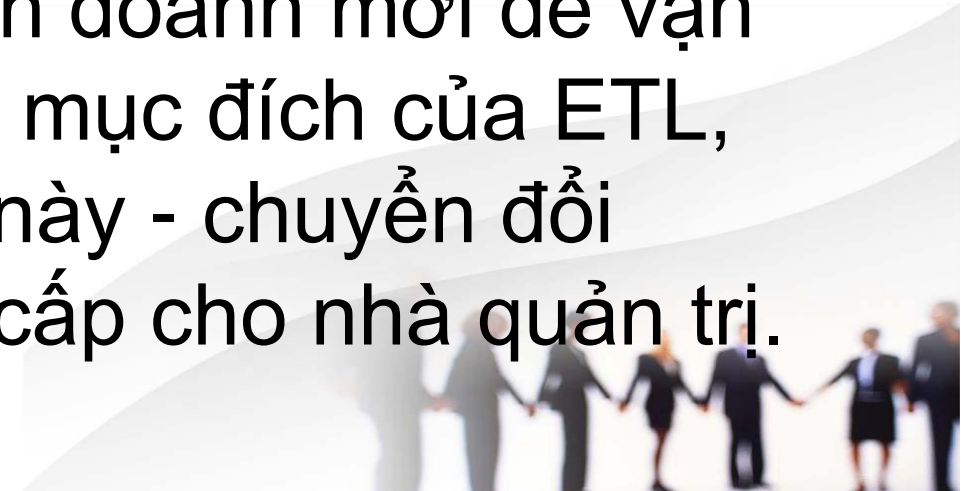


# Giới thiệu



# Giới thiệu

- Tại sao cần ETL?
- Chuyển mục đích, và tối ưu hóa mục đích sử dụng dữ liệu của các phần mềm từ **ghi nhận các nghiệp vụ phát sinh** hàng ngày, sang mục đích **khai thác, vận hành, và phân tích các dữ liệu** này để các nhà quản trị tìm ra các cơ may phát triển, các hoạt động kinh doanh mới để vận hành doanh nghiệp - và đây chính là mục đích của ETL, và là nguyên nhân bạn cần công cụ này - chuyển đổi công năng sử dụng dữ liệu để cung cấp cho nhà quản trị.





# Giới thiệu

- Công cụ ETL

– **Pentaho Kettle** - là công cụ Open Source, thành lập 2001 và sử dụng công cụ GUI để bạn xây dựng và vận hành ETL dữ liệu của mình - họ có phiên bản Community và phiên bản thương mại, và bạn có thể sử dụng Java để phát triển Engine của sản phẩm này. Đây là công cụ tương đối đầy đủ cho việc ETL, tổ chức Warehouse, và xây dựng các báo cáo phân tích BI. Phiên bản Community hiện đang có 13,500 Register



# Giới thiệu

- Công cụ ETL

- **Talend** - thành lập tháng 10, 2006 - tập trung vào ETL dữ liệu và là một opensource cho ETL dữ liệu

- **Informatica PowerCenter** - Đây là công cụ rất tốt cho ETL dữ liệu (bản thương mại) - được thành lập năm 1993 và cho tới nay đã có khoảng 2,600 khách hàng tin dùng và đang vận hành công cụ này trên toàn cầu. Informatica tập trung mạnh vào ETL dữ liệu



# Giới thiệu

- Công cụ ETL

- **Inaplex Inaport** - ETL dữ liệu với một số các phần mềm nguồn là CRM thì Inplex Inaport có thể là một lựa.

- **SSIS**: ETL dữ liệu từ nhiều nguồn khác nhau, sản phẩm từ microsoft.



# Giới thiệu

- SSIS

- SQL Server Integration Service (SSIS) được đưa vào từ bản 2005, là phiên bản tiếp theo của DTS trong SQL Server 2000 trở về trước. Công cụ này dùng để thực hiện các tác vụ tích hợp dữ liệu (Data integration), là thành phần chính trong các ứng dụng data warehouse. Nhiệm vụ gom dữ liệu từ các nguồn khác nhau và tổ chức lại theo cách thích hợp cho các mục đích báo cáo nhất định.



# Nội dung

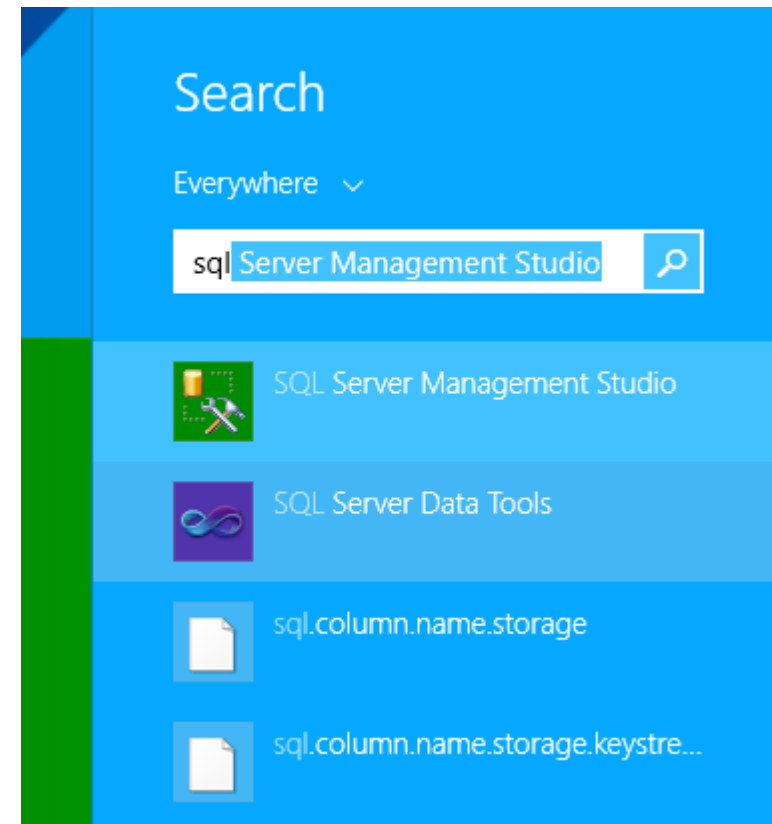
- Giới thiệu
- **Hướng dẫn sử dụng SSIS**
- Chương trình tích hợp
- Một số lưu ý khi viết câu truy vấn



# Hướng dẫn sử dụng SSIS

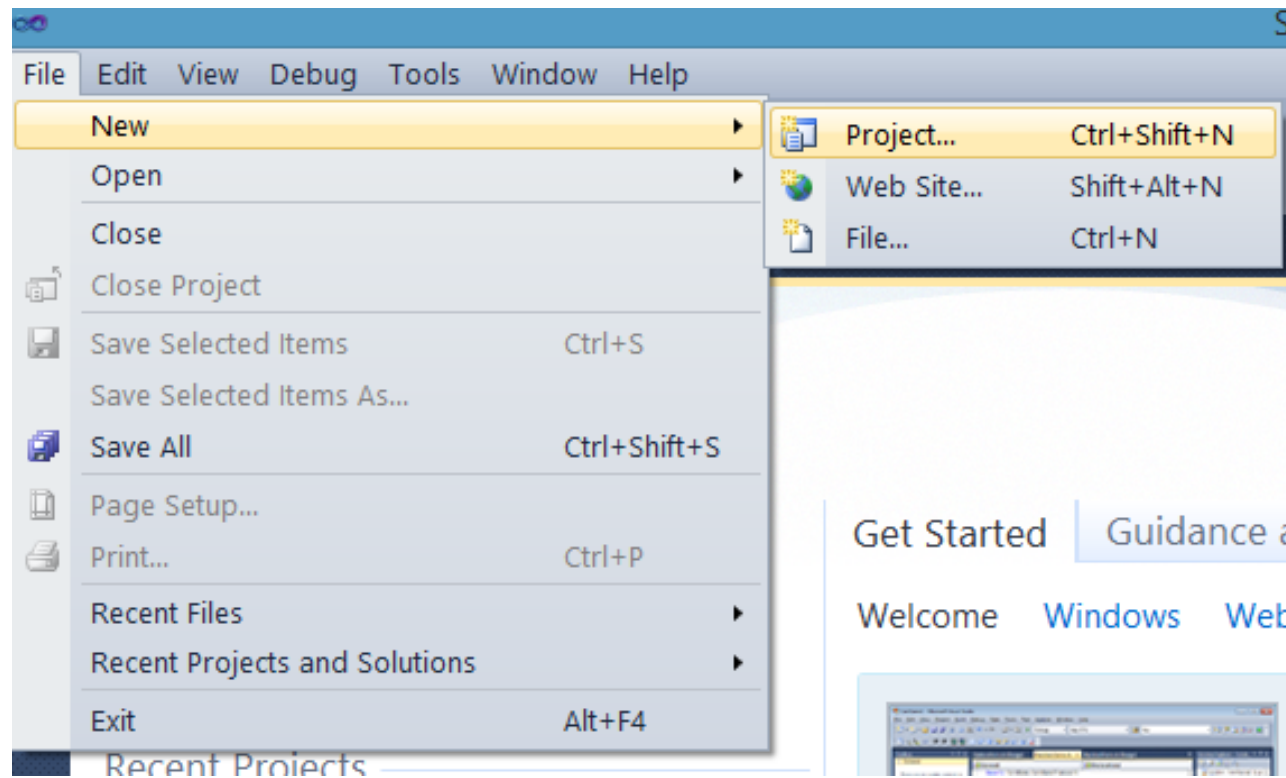
- Yêu cầu

- Download và cài Microsoft SQL Server Data Tools - Business Intelligence for Visual Studio 2013



# Hướng dẫn sử dụng SSIS

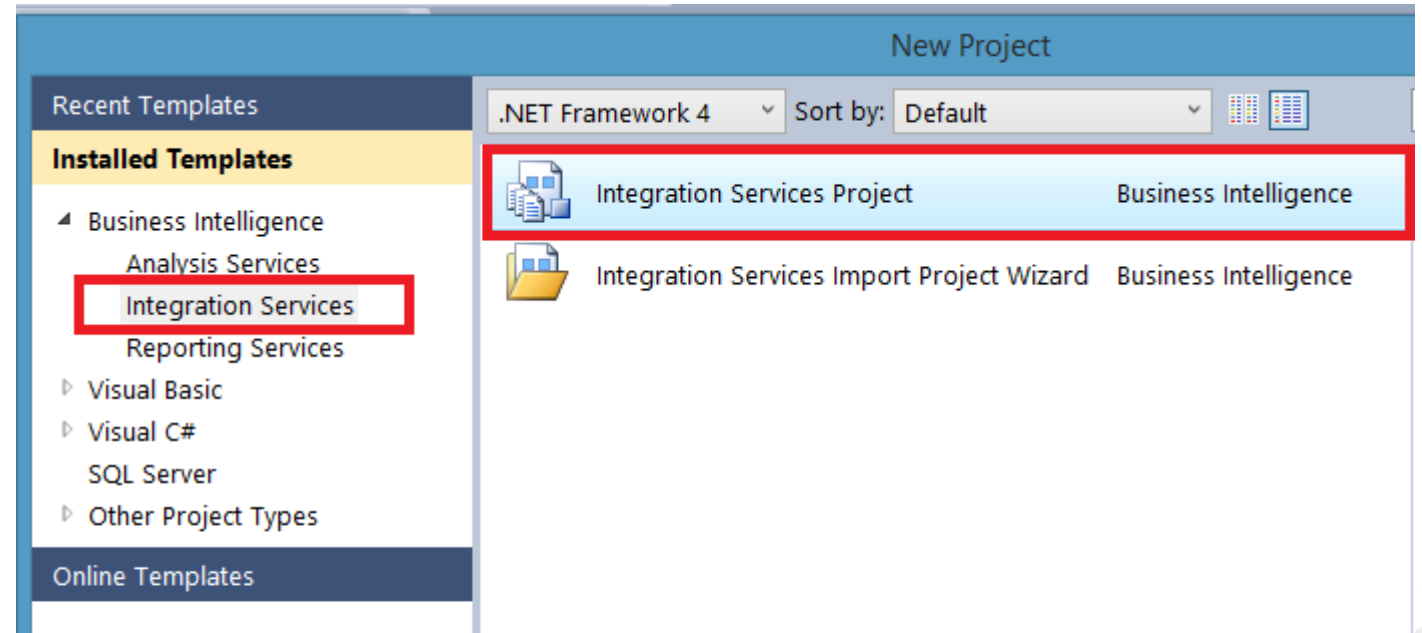
–Tạo mới một project



# Hướng dẫn sử dụng SSIS

- Chọn template ->
- > Integration Services
- > Integration Services Project

Đặt tên project



Name:	demoSSIS	
Location:	F:\Research\Projects\NET\	Browse...
Solution name:	demoSSIS	<input checked="" type="checkbox"/> Create directory for solution
		OK Cancel

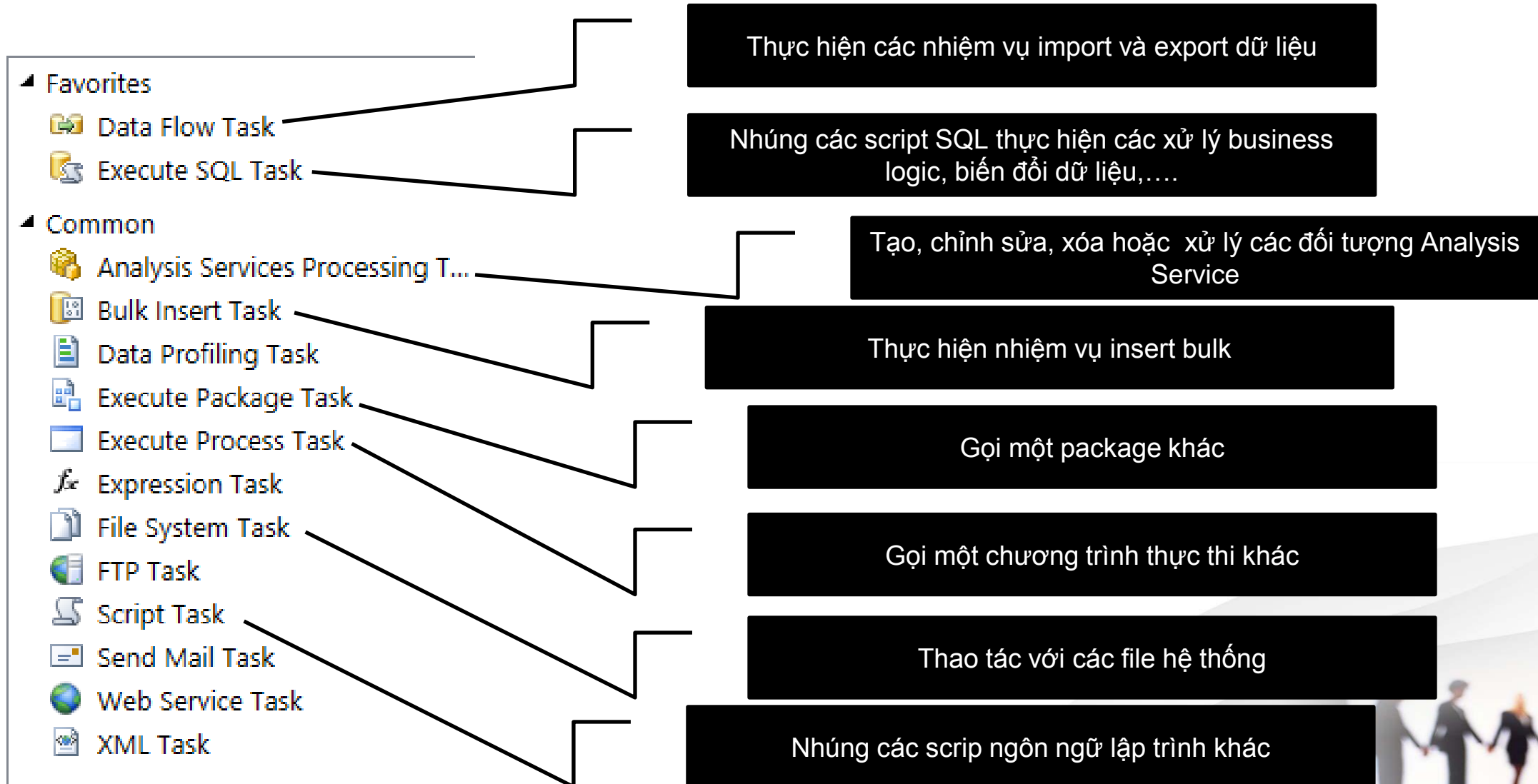


# Hướng dẫn sử dụng SSIS

- Các tool trong SSIS
  - Control Flow (Containers and Tasks)
  - Data Flow (Source, Transformations, Destinations)
  - Event Handler (Sending of messages, Emails)
  - Package Explorer (A single view for all in package)
  - Parameters (User interaction)





# Hướng dẫn sử dụng SSIS















# Hướng dẫn sử dụng SSIS

## ▲ Favorites

-  Data Flow Task
-  Execute SQL Task

## ▲ Common

-  Analysis Services Processing T...
-  Bulk Insert Task
-  Data Profiling Task
-  Execute Package Task
-  Execute Process Task
-  Expression Task
-  File System Task
-  FTP Task
-  Script Task
-  Send Mail Task
-  Web Service Task
-  XML Task

Thực hiện nhiệm vụ gửi email

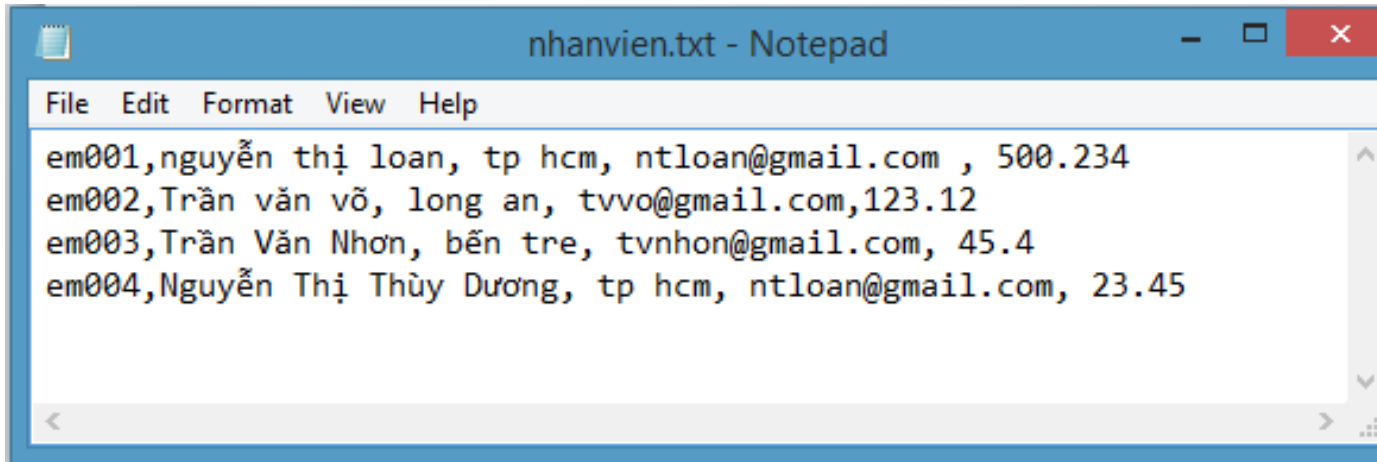
Tương tác với một web services, nhận kết quả trả về và lưu trong một Destination

Thao tác với tập tin XML



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Import dữ liệu từ một tập tin vào một table trong một CSDL**
- Ta có một tập tin có cấu trúc định dạng như sau:



```
File Edit Format View Help
em001,nguyễn thị loan, tp hcm, ntloan@gmail.com , 500.234
em002,Trần văn võ, long an, tvvo@gmail.com,123.12
em003,Trần Văn Nhon, bến tre, tvnhon@gmail.com, 45.4
em004,Nguyễn Thị Thùy Dương, tp hcm, ntloan@gmail.com, 23.45
```



# Hướng dẫn sử dụng SSIS

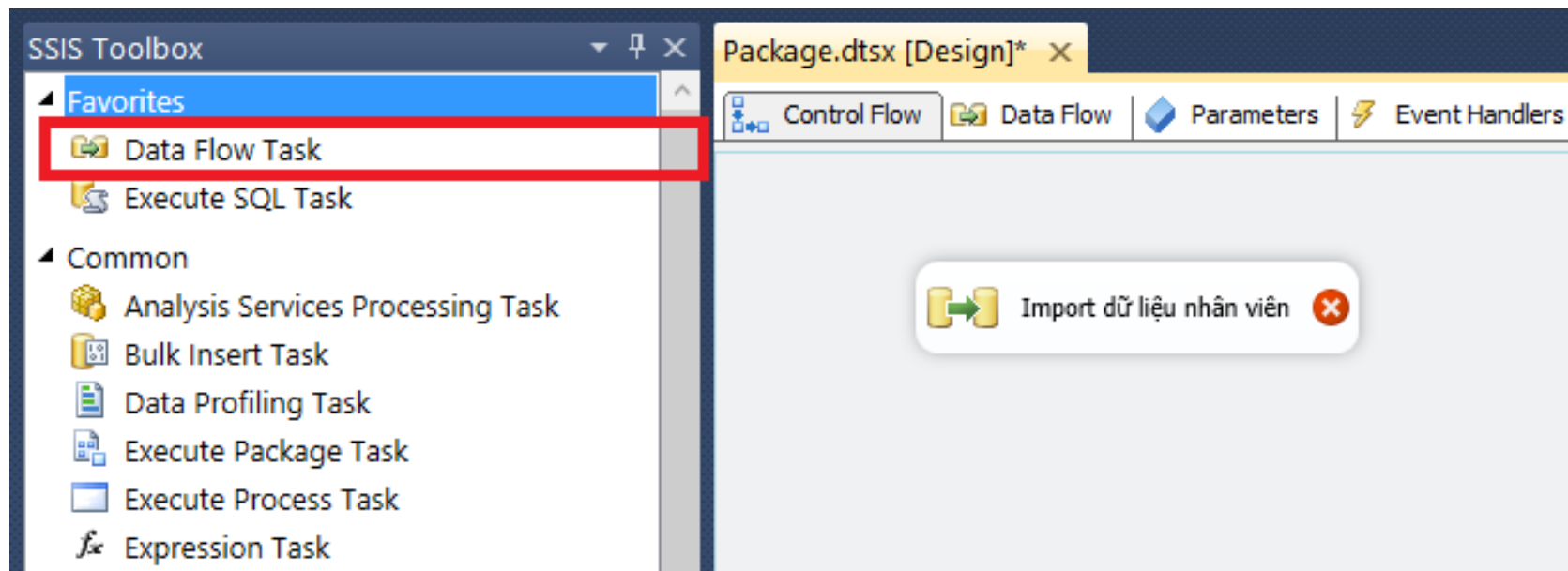
- **Ví dụ: Import dữ liệu từ một tập tin vào một table trong một CSDL**
- Ta có database với table NhanVien

```
SQLQuery1.sql -...(home\pmtu (54))* x
CREATE DATABASE demo_ssis
GO
USE demo_ssis
GO
CREATE TABLE nhanvien
(
    id INT PRIMARY KEY IDENTITY,
    manv VARCHAR(50),
    hoten NVARCHAR(50),
    diachi NVARCHAR(100),
    email VARCHAR(50),
    luong money
)
```



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Import dữ liệu từ một tập tin vào một table trong một CSDL**
- Dùng tool Data Flow Task để thực hiện import



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Import dữ liệu từ một tập tin vào một table trong một CSDL**
- Dùng tool Data Flow Task để thực hiện import



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Export dữ liệu từ table trong CSDL ra một tập tin**
- Dùng tool Data Flow Task để thực hiện import
  - Cần xác định lấy dữ liệu từ table như thế nào
    - Dùng câu truy vấn
    - Toàn bộ dữ liệu của table
  - Cần xác định nơi lưu tập tin cần export





# Hướng dẫn sử dụng SSIS

- **Ví dụ: Export dữ liệu từ table trong CSDL ra một tập tin**
- Dùng tool Data Flow Task để thực hiện import



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Gọi package từ command line**

- Dùng tool dtexec

- Cú pháp

- dtexec /option [value] [/option [value]]...

- Ví dụ

- dtexec /f E:\test\_package\test\_package\Package.dtsx /set \package.variables[id];1**



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Gọi package từ command line**
- Dùng tool dtexec



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Dùng biến trong package**
- `dtexec /f F:\Research\Projects\NET\demoSSIS\demoSSIS\Import-Export.dtsx /set \package.variables[file_path_import];F:\\nhanvien.txt /set \package.variables[file_path_export];F:\\export\\nhanvien_new.txt`



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Một quá trình xử lý dữ liệu đơn giản**
- Đọc dữ liệu từ file
- Lưu bảng chi tiết
- Summary dữ liệu
- Lưu vào bảng cần thiết



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Một quá trình xử lý dữ liệu đơn giản**
- Cho lược đồ CSDL

```
CREATE TABLE ChiTietHoaDon
(
    MaCTHD INT PRIMARY KEY IDENTITY,
    NgayMuaHang DATE,
    SoLuong INT,
    Gia MONEY,
    MaSanPham VARCHAR(20),
    MaKH VARCHAR(20),
    XuLy INT
)
GO
```

```
CREATE TABLE ChiTietHoaDon_TMP
(
    NgayMuaHang DATE,
    SoLuong INT,
    Gia MONEY,
    MaSanPham VARCHAR(20),
    MaKH VARCHAR(20),
    Xuly INT DEFAULT(0) NOT NULL
)
```

```
CREATE TABLE HoaDon
(
    MaHoaDon INT PRIMARY KEY IDENTITY,
    NgayMuaHang DATE,
    NgayLapHD DATE,
    MaKH VARCHAR(20),
    TongTien MONEY
)
GO
```



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Một quá trình xử lý dữ liệu đơn giản**
- Các bước xử lý



# Hướng dẫn sử dụng SSIS

- **Ví dụ: Một quá trình xử lý dữ liệu đơn giản**

Tính toán và đưa dữ liệu vào  
bảng thực

- Thêm dữ liệu bảng thực từ bảng temp

```
DECLARE @NgayMua DATE, @MaKH VARCHAR(20)
SELECT TOP 1 @NgayMua=NgayMuaHang, @Makh = Makh FROM ChiTietHoaDon_TMP
DELETE FROM ChiTietHoaDon WHERE MaKH = @MaKH AND NgayMuaHang = @NgayMua
```

```
INSERT INTO ChiTietHoaDon (NgayMuaHang, MaKH, Gia, SoLuong, MaSanPham)
SELECT NgayMuaHang, MaKH, Gia, SoLuong, MaSanPham
FROM ChiTietHoaDon_TMP
```





# Hướng dẫn sử dụng SSIS

- **Ví dụ: Một quá trình xử lý dữ liệu đơn giản**

Tính toán và đưa dữ liệu vào  
bảng thực

- Tính toán dữ liệu

```
INSERT INTO HoaDon (MaKH, NgayLapHD, NgayMuaHang, TongTien)
SELECT cthd.MaKH,
       GETDATE(),
       cthd.NgayMuaHang,
       SUM (cthd.SoLuong * cthd.Gia)
FROM ChiTietHoaDon cthd
WHERE cthd.XuLy=0
GROUP BY cthd.MaKH, cthd.NgayMuaHang

UPDATE ChiTietHoaDon
SET XuLy = 1
WHERE NgayMuaHang=@NgayMua AND MaKH = @MaKH
```



# Hướng dẫn sử dụng SSIS

- Ví dụ: Một quá trình xử lý dữ liệu đơn giản



# Hướng dẫn sử dụng SSIS

- Vấn đề
  - Quá trình xử lý dữ liệu cần các yếu tố?
    - Validate
    - Log
    - Tính toàn vẹn
    - Độ tin cậy
- Cách giải quyết ?



# Nội dung

- Giới thiệu
- Hướng dẫn sử dụng SSIS
- **Chương trình tích hợp**
- Một số lưu ý khi viết câu truy vấn



# Chương trình tích hợp

- Có bao nhiêu cách gọi package?

Chạy trong VS Studio

Từ command line

Tự viết một chương trình tự động  
gọi package

Từ một công cụ khác cùng hãng

Từ một công cụ khác hãng



# Nội dung

- Giới thiệu
- Hướng dẫn sử dụng SSIS
- Chương trình tích hợp
- **Một số lưu ý khi viết câu truy vấn**



# Một số lưu ý khi viết câu truy vấn

- Comment
  - Nên comment đối với các business logic phức tạp
- Ghi log
  - Nên ghi log để đảm bảo an toàn dữ liệu
- Thông tin phiên bản
  - Nên quản lý phiên bản tốt, do một đối tượng trong CSDL có thể được chỉnh sửa nhiều lần và nhiều tác giả.



# Một số lưu ý khi viết câu truy vấn

- Từ khóa
  - Viết IN HOA

IN HOA từ khóa

```
CREATE TABLE HoaDon
(
    MaHoaDon INT PRIMARY KEY IDENTITY,
    NgayMuaHang DATE,
    NgayLapHD DATE,
    MaKH VARCHAR(20),
    TongTien MONEY
)
GO
```





# Một số lưu ý khi viết câu truy vấn

- Thủ tục, hàm

- Hạn chế gọi thủ tục, hàm trong mệnh đề WHERE
- Các tham số INPUT nên có giá trị mặc định

- Lời khuyên

- Hạn chế dùng Cursor
- Hạn chế dùng câu truy vấn lồng nhiều cấp
- Hạn chế đánh INDEX trên những bảng thường xuyên thay đổi dữ liệu
- Không dùng SELECT \*
- Không dùng IF EXISTS (SELECT \* FROM.....)
- Nên dùng IF EXISTS (SELECT TOP 1 1 FROM...)



# Một số lưu ý khi viết câu truy vấn

- Lời khuyên
  - Viết thủ tục, hàm sao cho tính dừng lại
  - Hạn chế dùng trigger



# Câu hỏi

