ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN

A picture containing logo

Description automatically generated

**ĐỒ ÁN THỰC HÀNH**

**(LẦN 1)**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

*GVHD: Hồ Thị Hoàng Vy*

|  |  |  |
| --- | --- | --- |
| **Mã nhóm** | **MSSV** | **Họ và tên** |
| **TTKD-24** | 18120227 | Phạm Văn Minh Phương |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

**BẢNG PHÂN CÔNG**

**TH1**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Phân loại thuộc tính cần thiết, kiểm và sửa lỗi, làm báo cáo | 100% |
| 18120299 | Trương Công Quốc Cường | Giải thích thuộc tính từ bảng Accidents | 100% |
| 18120035 | Đoàn Nguyễn Tấn Hưng | Giải thích thuộc tính từ bảng Casualties và bảng LSOA-Postcode Mapping | 100% |
| 18120534 | Hoàng Công Sơn | Giải thích thuộc tính từ bảng Vehicles và bảng Postcodes | 100% |

**TH2**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | ETL dữ liệu từ nguồn sang stage. Sửa các lỗi phát sinh. Viết báo cáo | 100% |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

Mục lục

[I. Giải thích ý nghĩa thuộc tính nguồn dữ liệu 4](#_Toc85856737)

[II. Quá trình ETL dữ liệu từ nguồn vào stage 9](#_Toc85856738)

# **Giải thích ý nghĩa thuộc tính nguồn dữ liệu**

1. UK Car Accidents (2011-2014)

* Bảng Accidents:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | Location Easting OSGR | mã OSGR (**Ordnance Survey Grid Reference**) phía Đông |
| 2 | Location Northing OSGR | mã OSGR phía Bắc |
| 3 | Longitude | Kinh độ |
| 4 | Latitude | Vĩ độ |
| 6 | Accident Severity | Mức độ độ nghiêm trọng của tai nạn |
| 7 | Number of Vehicles | Số phương tiện trong tai nạn |
| 8 | Number of Casualties | Số thương vong |
| 9 | Date | Ngày xảy ra tai nạn |
| 10 | Day of Week | Ngày thứ mấy trong tuần |
| 11 | Time | Giờ xảy ra tai nạn |
| 12 | Road type | Loại đường giao thông |
| 13 | Speed limit | Giới hạn tốc độ |
| 14 | Junction Detail | Chi tiết giao lộ |
| 15 | Junction Control | Đèn giao thông (giá trị) |
| 16 | Light Conditions | Điều kiện ánh sáng |
| 17 | Weather Conditions | Điều kiện thời tiết |
| 18 | Road Surface Conditions | Điều kiện mặt đường |
| 19 | Special Conditions at Site | Điều kiện đặc biệt tại địa điểm |
| 20 | Carriageway Hazards | Mối nguy hiểm trên đường |
| 21 | Urban or Rural Area | Vùng đô thị hay ngoại ô |
| 22 | LSOA of Accident Location | Mã LSOA (**Lower Layer Super Output Areas**) của vị trí tai nạn |

* Bảng Casualties:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | Accident\_Index | Index của bảng |
| 2 | Vehicle\_Reference | Tham chiếu tới bảng Vehicle |
| 3 | Casualty\_Reference | Tham chiếu tới bảng Casualty |
| 4 | Casualty\_Class | Cấp tai nạn (cấp 1, 2 hoặc 3) |
| 5 | Sex\_of\_Casualty | Giới tính người bị nạn (1 hoặc 2) |
| 6 | Age\_of\_Casualty | Tuổi người bị nạn |
| 7 | Age\_Band\_of\_Casualty | Nhóm tuổi người bị nạn |
| 8 | Casualty\_Severity | Mức độ tai nạn (mức 1, 2 hoặc 3) |
| 9 | Car\_Passenger | Hành khách xe hơi |
| 10 | Bus\_or\_Coach\_Passenger | Hành khách xe buýt hoặc xe khách |
| 11 | Casualty\_Type | Loại tai nạn |

* Bảng Vehicles:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** |
| 1 | Vehicle Reference | Tham chiếu phương tiện |
| 2 | Vehicle Type | Loại phương tiện |
| 5 | Vehicle Location-Restricted Lane | Vị trí phương tiện - tại làn đường hạn chế |
| 6 | Junction Location | Vị trí giao lộ |
| 7 | Journey Purpose of Driver | Lý do tham gia giao thông của tài xế |
| 8 | Sex of Driver | Giới tính tài xế |
| 9 | Age of Driver | Tuổi của tài xế |

1. LSOA – Postcode Mapping

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** |
| 1 | oa11cd | Mã code OA (Output Area) năm 2011 |
| 2 | lsoa11cd | Mã code LSOA (Lower Layer Output Area) năm 2011 |
| 3 | msoa11cd | Mã code MSOA (Middle Layer Output Area) năm 2011 |
| 4 | ladcd | Mã code LAD (Local authority districts) |
| 5 | lsoa11nm | Tên của LSOA |
| 6 | msoa11nm | Tên của MSOA |

1. Postcode

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | postcode | Mã bưu chính |
| 2 | easting | Tọa độ x trong hệ tham chiếu lưới |
| 3 | northing | Tọa độ y trong hệ tham chiếu lưới |
| 4 | latitude | Vĩ độ |
| 5 | longitude | Kinh độ |
| 6 | city | Thành phố |
| 7 | county | Quận |
| 8 | country\_code | Mã quốc gia |
| 9 | country\_name | Tên quốc gia |
| 10 | iso3166-2 | GB-Mã quốc gia (Great Britain) |
| 11 | region\_code | Mã vùng |

# **Quá trình ETL dữ liệu từ nguồn vào stage**

B1. Tạo database mới để chứa dữ liệu từ dữ liệu UK Car Accidents (2011-2014), LSOA-Postcode Mapping, Postcode

A screenshot of a computer

Description automatically generated

B2. Tạo project Integration Process Project để chạy các Data Flow Task nhằm nạp dữ liệu

A screenshot of a computer

Description automatically generated with medium confidence

B3. Mở tab Data Flow Task để tạo task nạp dữ liệu. Sau đó, kéo biểu tượng “Flat File Source” (Nguồn dữ liệu từ file .csv) và “OLE DB Destination” (Đích dữ liệu là SQL Database) vào trong giao diện.

A screenshot of a computer

Description automatically generated with medium confidence

B4. Double click vào biểu tượng Flat File Source trong giao diện Data Flow Task và tiến hành tạo kết nối tới file .csv dữ liệu gốc bằng cách nhấn vào “New”. Chọn “Browse” tới file dữ liệu gốc cần nạp. Nên đổi tên kết nối mặc định để dễ phân biệt.

A screenshot of a computer

Description automatically generated

B5. Nhấp vào tab “Column” để kiểm tra xem dữ liệu được đọc lên đúng hay chưa. Nếu đã đúng, nhấn “OK”

A screenshot of a computer

Description automatically generated

Chú ý tích vào ô “Retain null values from the source as null values in the data flow” để giữ lại giá trị null trong kho dữ liệu.

A screenshot of a computer

Description automatically generated

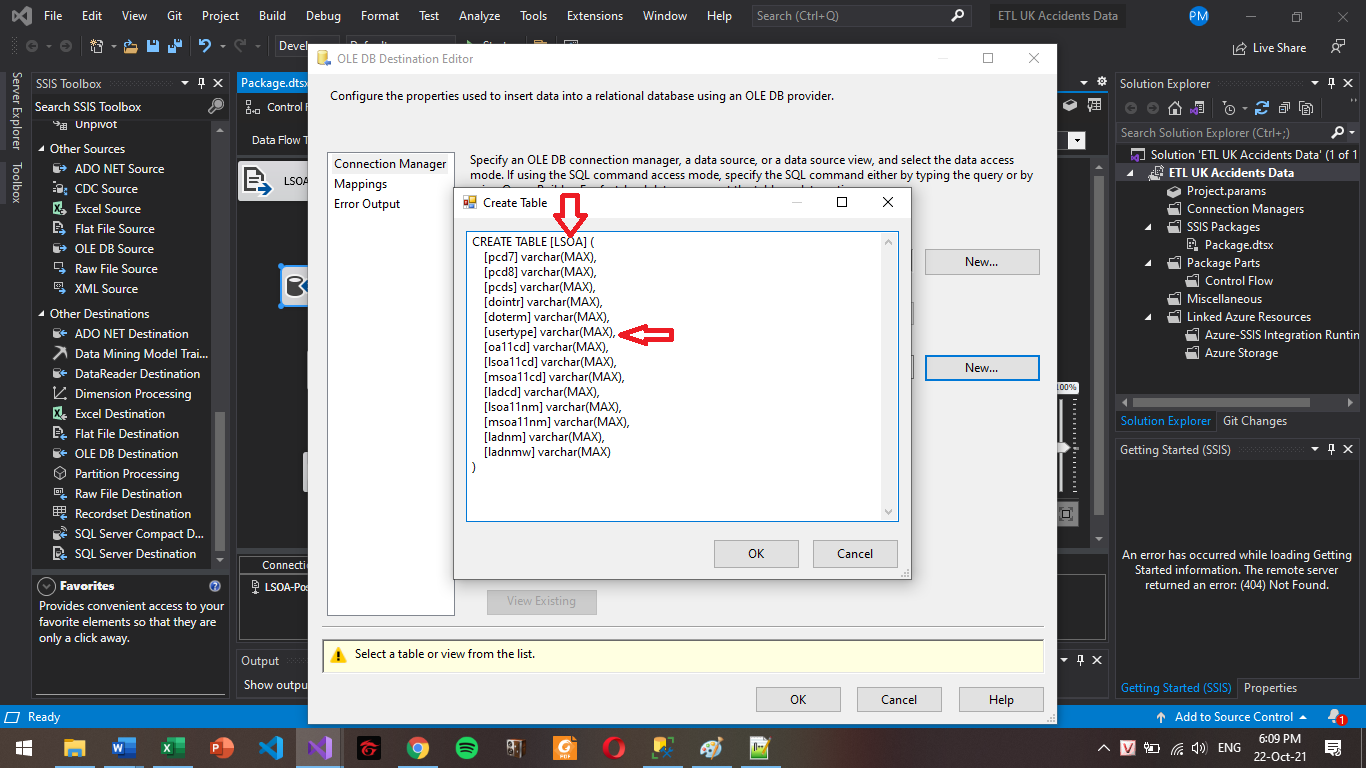
B6. Double click vào tên của biểu tượng Flat File Source trên giao diện để sửa đổi tên tương ứng với nguồn dữ liệu, tránh nhầm lẫn. Sau đó, kéo mũi tên màu xanh (mũi tên output) tới biểu tượng của đích dữ liệu (OLE DB Destination)

A screenshot of a computer

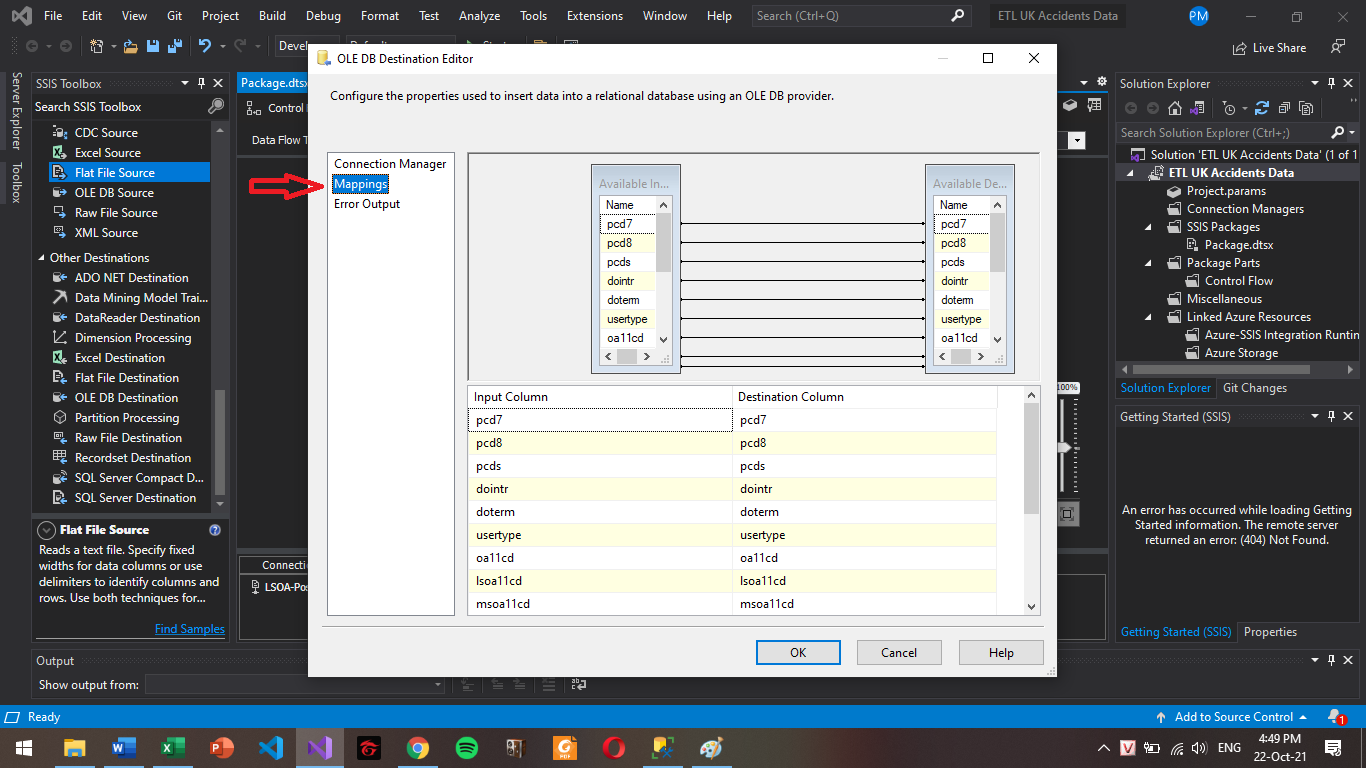
Description automatically generated with medium confidence

B7. Double click vào biểu tượng OLE DB Destination, chọn “New” để tạo kết nối tới Database đích tương ứng với nguồn dữ liệu được chọn. Nhấn “OK” sau khi chọn đúng giá trị ở mỗi tabA screenshot of a computer

Description automatically generated

B8. Tạo bảng trong Database để chứa dữ liệu từ nguồn. Chú ý loại kiểu dữ liệu và kích cỡ phải khớp với dữ liệu gốc để tránh xảy ra lỗi. Ở đây chúng em sử dụng nvarchar(MAX) vì nvarchar có cơ chế quản lý bộ nhớ động nên nvarchar chỉ thực sự chiếm số byte tương ứng với số kí tự.

B9. Kiểm tra mapping dữ liệu

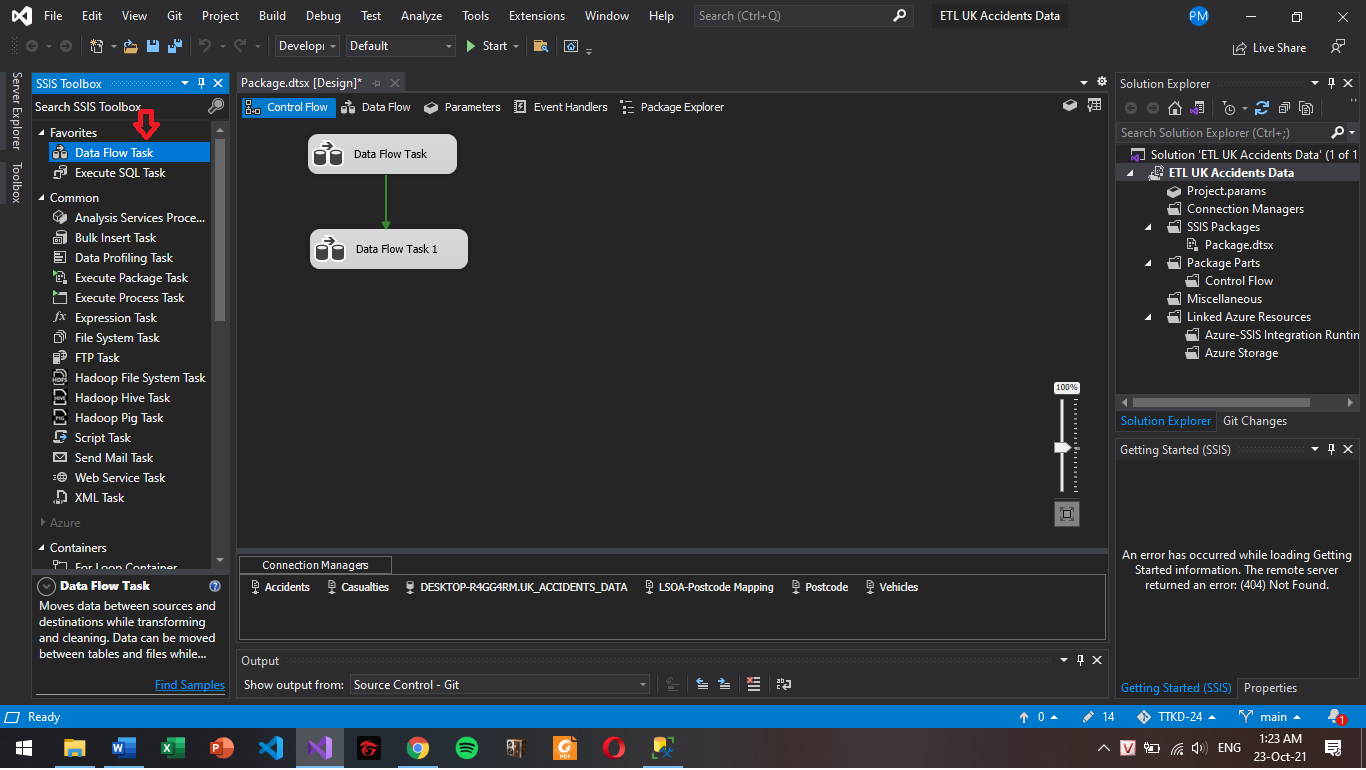


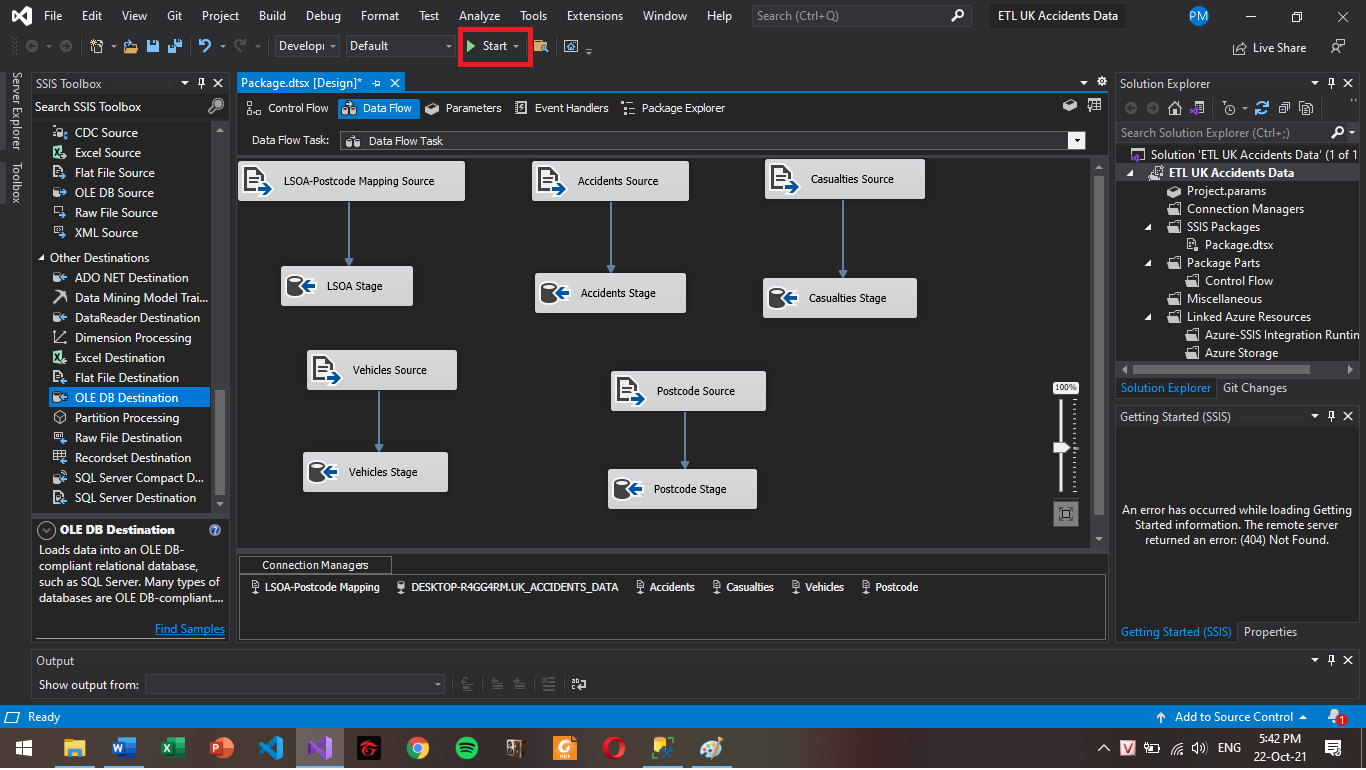
B10. Double click vào tên của biểu tượng OLE DB Destination trên giao diện để đổi lại tên tương ứng, tránh nhầm lẫn

A screenshot of a computer

Description automatically generated with medium confidence

B11. Tạo một Data Flow Task khác và lặp lại tương tự từ B3 tới B10 với các file nguồn dữ liệu và đích dữ liệu còn lại. Nối Data Flow Task mới tạo này vào Data Flow Task ban đầu để cho Data Flow Task này chạy trước





**LƯU Ý ĐẶC BIỆT:** Trong quá trình nạp dữ liệu, chúng em phát hiện rằng file .csv LSAO-Postcode Mapping có chứa giá trị lớn, tốn nhiều byte để lưu trữ nhưng dữ liệu gốc lại được lưu ở dạng non-unicode, dẫn đến việc khi sử dụng trực tiếp file .csv này để nạp dữ liệu thông qua Data Flow Task sẽ dẫn đến 1 trong 2 lỗi:

* Nếu bảng LSAO tương ứng ở trong Database sử dụng kiểu dữ liệu nvarchar(max) cho tất cả thuộc tính thì sẽ gặp lỗi không thể convert giá trị non-unicode (từ file .csv) sang dạng Unicode character được sử dụng trong bảng trong database

A screenshot of a computer

Description automatically generated

* Nếu bảng LSAO tương ứng ở trong Database không sử dụng kiểu dữ liệu nvarchar(max) mà sử dụng varchar(max) cho tất cả thuộc tính thì sẽ gặp lỗi truncation do không chứa được toàn bộ dữ liệu khi Data Flow Task tự động chuyển đổi kiểu dữ liệu về non-unicode character. Khả năng do trong nguồn dữ liệu có 1 dòng quá dài để lưu theo dạng non-unicode hoặc có chứa kí tự unicode.

Graphical user interface

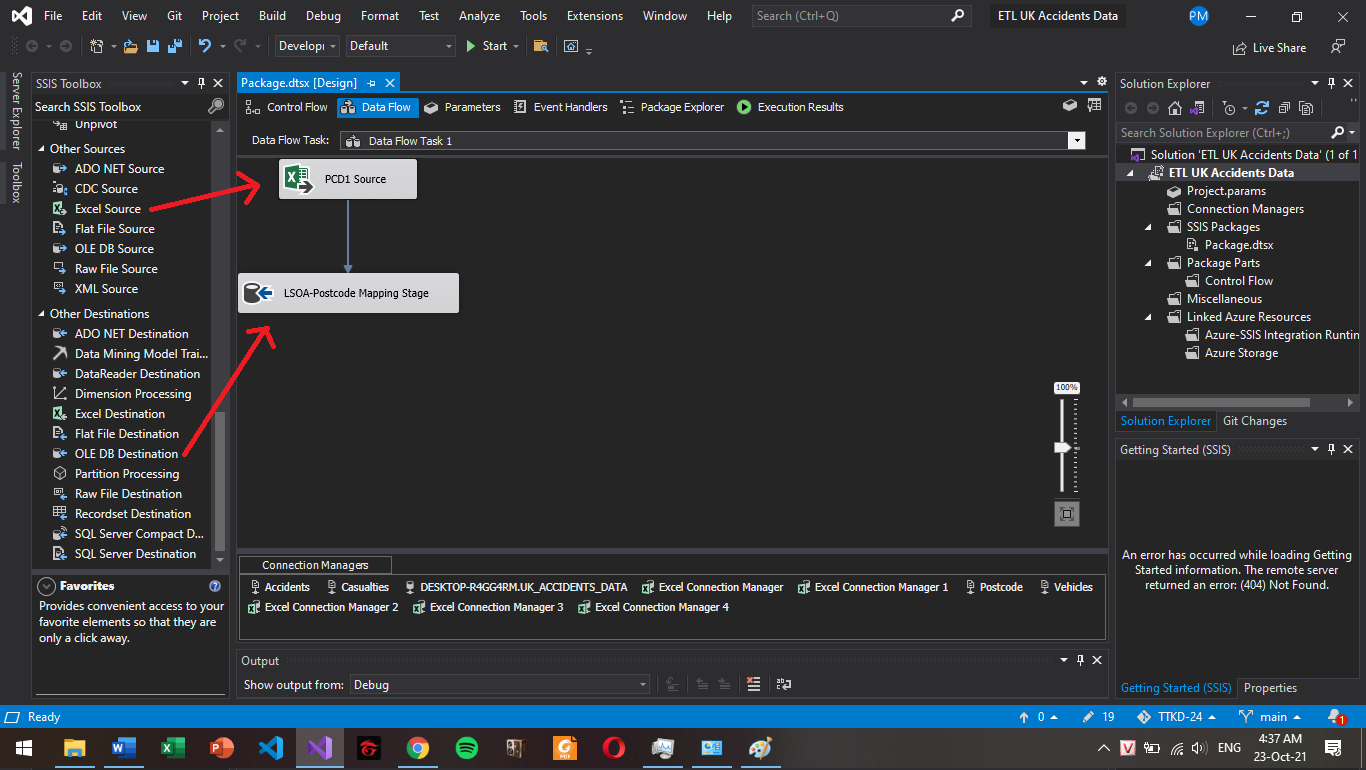
Description automatically generated

Điều đó dẫn tới giải pháp là phải đảm bảo được 2 điều:

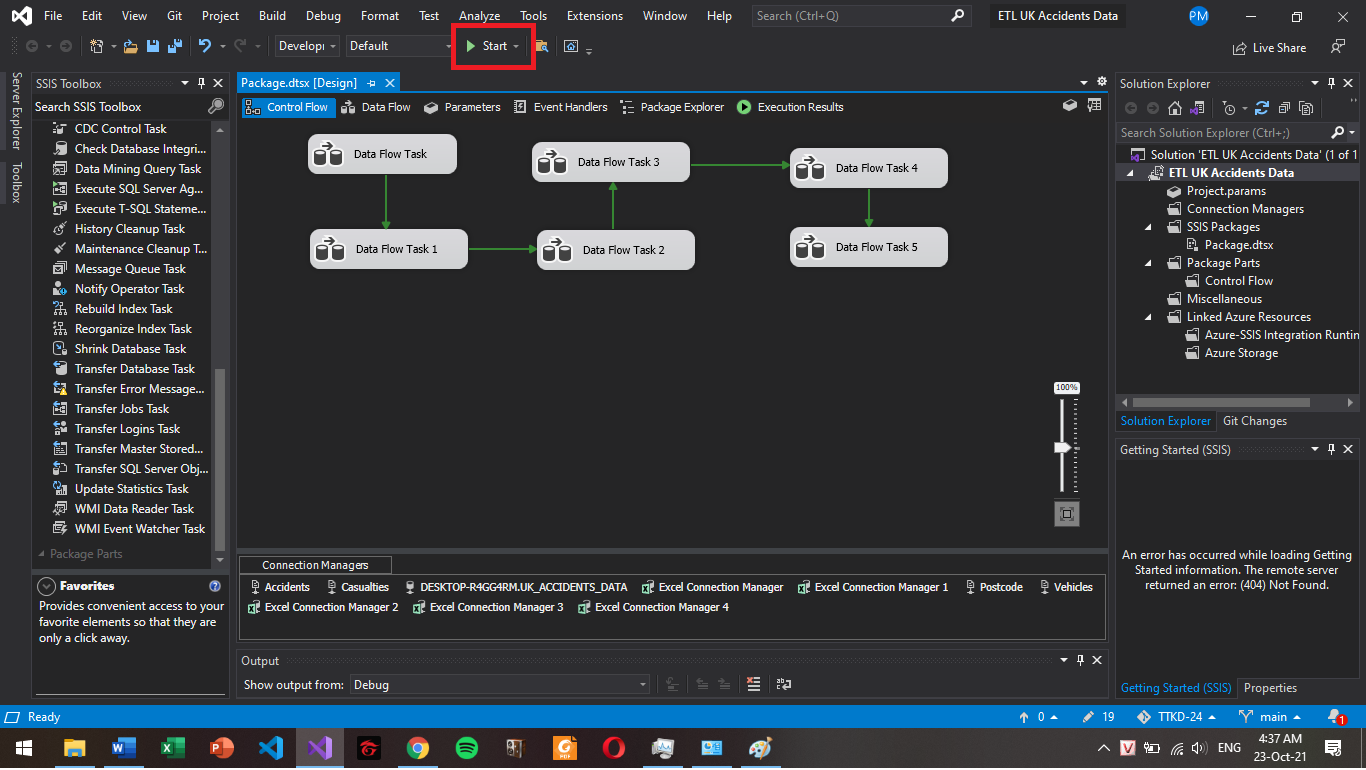
* Các cột trong bảng dữ liệu LSAO (tương ứng với nguồn dữ liệu LSAO-Postcode Mapping) phải là dạng nvarchar(max)
* File nguồn dữ liệu gốc .csv được chuyển sang định dạng mới của Excel là .xlsx để sử dụng nguồn dữ liệu “Excel Source” thay vì sử dụng “Flat File Source” như ban đầu vì sẽ gặp phải lỗi truncation.

Ngoài ra còn nên chia file .csv gốc ra làm nhiều file nhỏ để tiện chuyển đổi qua dạng .xlxs bằng Excel.

B12. Thực hiện tương tự B3-B10 nhưng với source là “Excel Source” cho các file .xlxs là nguồn dữ liệu LSOA-Postcode mapping được chia nhỏ và convert.



B13. Nối thứ tự các task và click “Start”



**Kết quả chạy thành công:**

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated