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

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN LÝ THUYẾT**

Phạm Ngọc Thùy Trang 18127022

Ngô Thị Thanh Thảo 18127053

Hoàng Nguyên Trúc 18127055

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

**Thành phố Hồ Chí Minh**

**Ộ MÔN HỆ**

# Mục lục

[I. Mục lục 1](#_Toc86915995)

[II. Bảng thông tin chi tiết nhóm 2](#_Toc86915996)

[III. Phân công công việc 3](#_Toc86915997)

[IV. ETL từ Source sang Stage 4](#_Toc86915998)

[V. Stage sang NDS 6](#_Toc86915999)

[VI. Tài liệu tham khảo 22](#_Toc86916000)

# Bảng thông tin chi tiết nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã nhóm:** | Nhóm 2 | | |
| **Số lượng:** | 3 | | |
| **MSSV** | **Họ tên** | **Email** | **SĐT** |
| 18127022 | Phạm Ngọc Thùy Trang | 18127022@  student.hcmus.edu.vn |  |
| 18127053 | Ngô Thị Thanh Thảo | 18127053@  student.hcmus.edu.vn |  |
| 18127055 | Hoàng Nguyên Trúc | 18127055@  student.hcmus.edu.vn |  |

# Phân công công việc

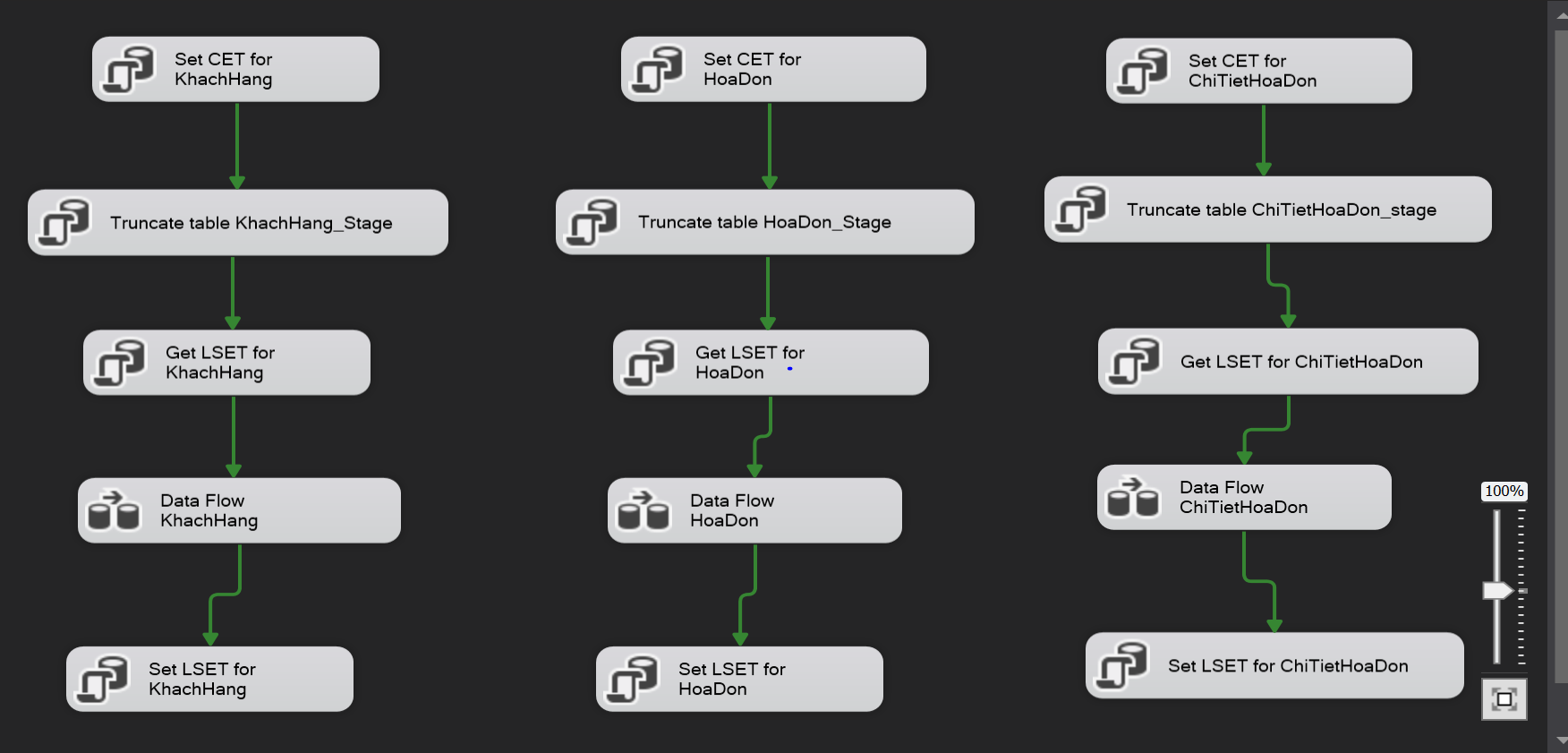
|  |  |  |
| --- | --- | --- |
| **Họ tên thành viên** | **Nội dung công việc** | **Phần trăm hoàn thành (%)** |
| Phạm Ngọc Thùy Trang |  |  |
| Ngô Thị Thanh Thảo |  |  |
| Hoàng Nguyên Trúc |  |  |

# ETL từ Source sang Stage

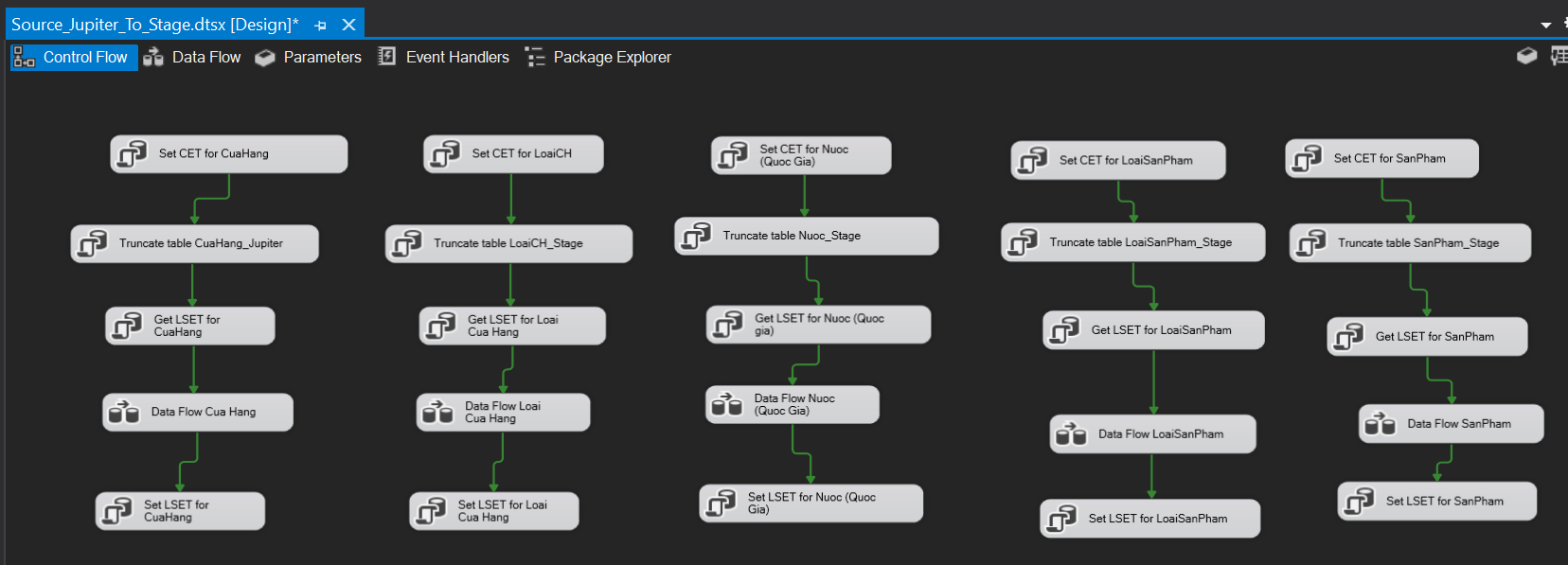
* Theo đồ án lý thuyết, quá trình ETL dữ liệu từ Source sang Stage sử dụng 3 source là Jade, Jupiter và Webtower.
* **Chiến lược rút trích dữ liệu từ cơ sở dữ liệu**: Incremental extract
* Cơ bản về đặc điểm của chiến lược Incremental extract là chỉ rút trích những thay đổi so với lần rút trích gần đây nhất (Dòng mới thêm, dòng mới xoá, dòng được cập nhật) dựa vào thuộc tính nhãn thời gian -> Trong đồ án này sử dụng hai nhãn thời gian là NgayTao (ngày tạo) và NgayCapNhat (ngày cập nhật mới nhất). Mỗi khi có một dòng thay đổi (update, delete) thì cập nhật lại nhãn thời gian.
* Khi đổ dữ liệu vào nguồn không có sự thay đổi về cấu trúc dữ liệu (tức là không thêm hay chỉnh sửa cột trong mỗi bảng).
* Tạo một cơ sở dữ liệu metadata gồm bảng có các cột Tên bảng, LSET, CET. Trong đó:
  + LSET: Thời điểm rút trích hiện tại
  + CET: Thời điểm rút trích thành công gần đây nhất
* Tạo một cơ sở dữ liệu Stage có cấu trúc giống với cơ sở dữ liệu nguồn
* Các bước thực hiện đổ nguồn vào stage:

1. Lấy thông tin LSET được lưu trong metadata
2. Lấy CET bằng thời gian khởi động package
3. Truncate bảng Stage hiện tại: TRUNCATE TABLE <TenBang>
4. Rút trích dữ liệu từ bảng Source sang bảng Stage: select \* from BangSource where (NgayTao >= LSET and NgayTao < CET) or (NgayCapNhat >= LSET and NgayCapNhat < CET)
5. Cập nhật LSET = CET

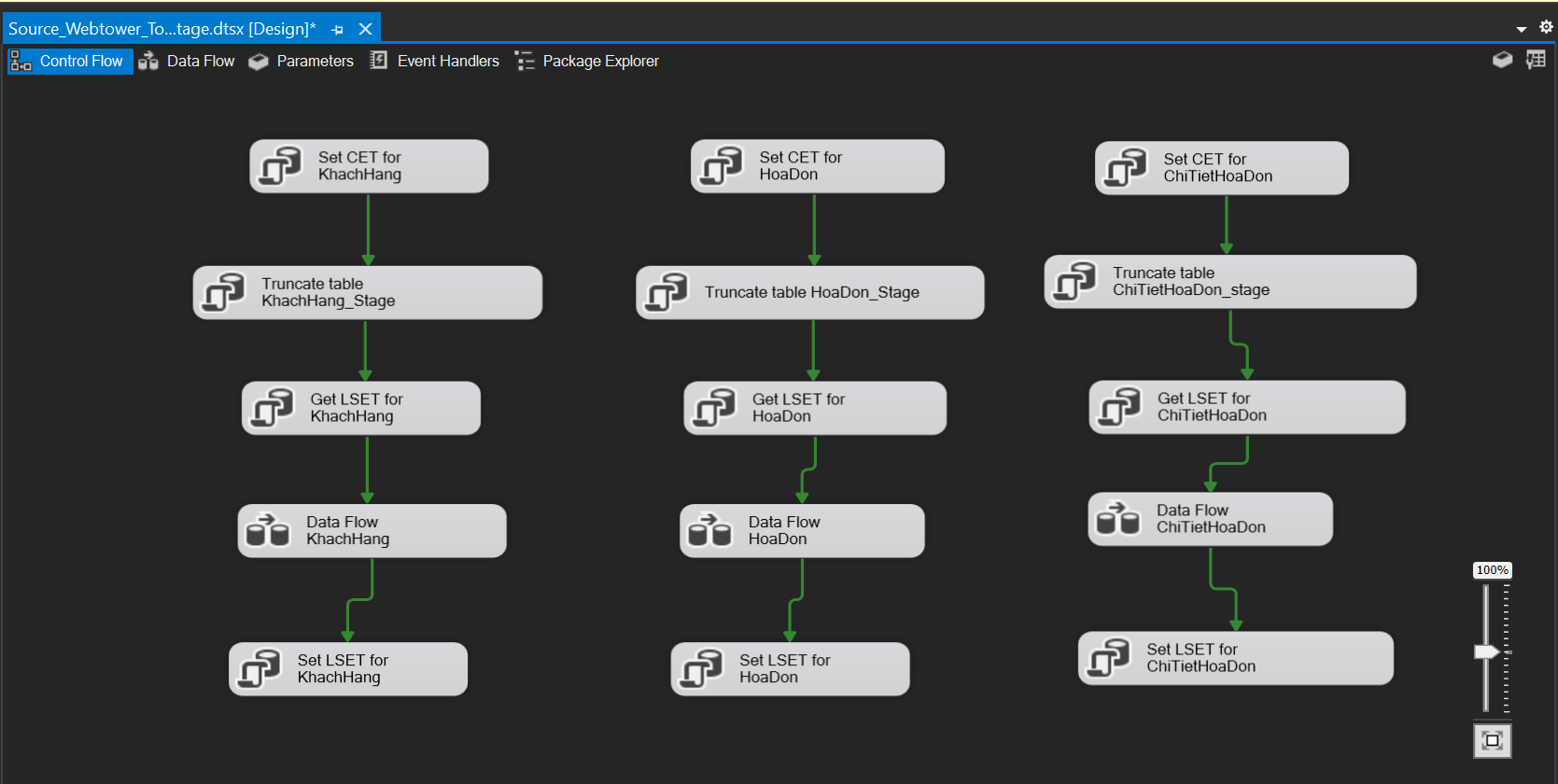
* Flow khi thực hiện trên Visual Studio Code
  + ETL từ Source Jade sang Stage

**~~~~**

* + ETL từ Source Jupiter sang Stage

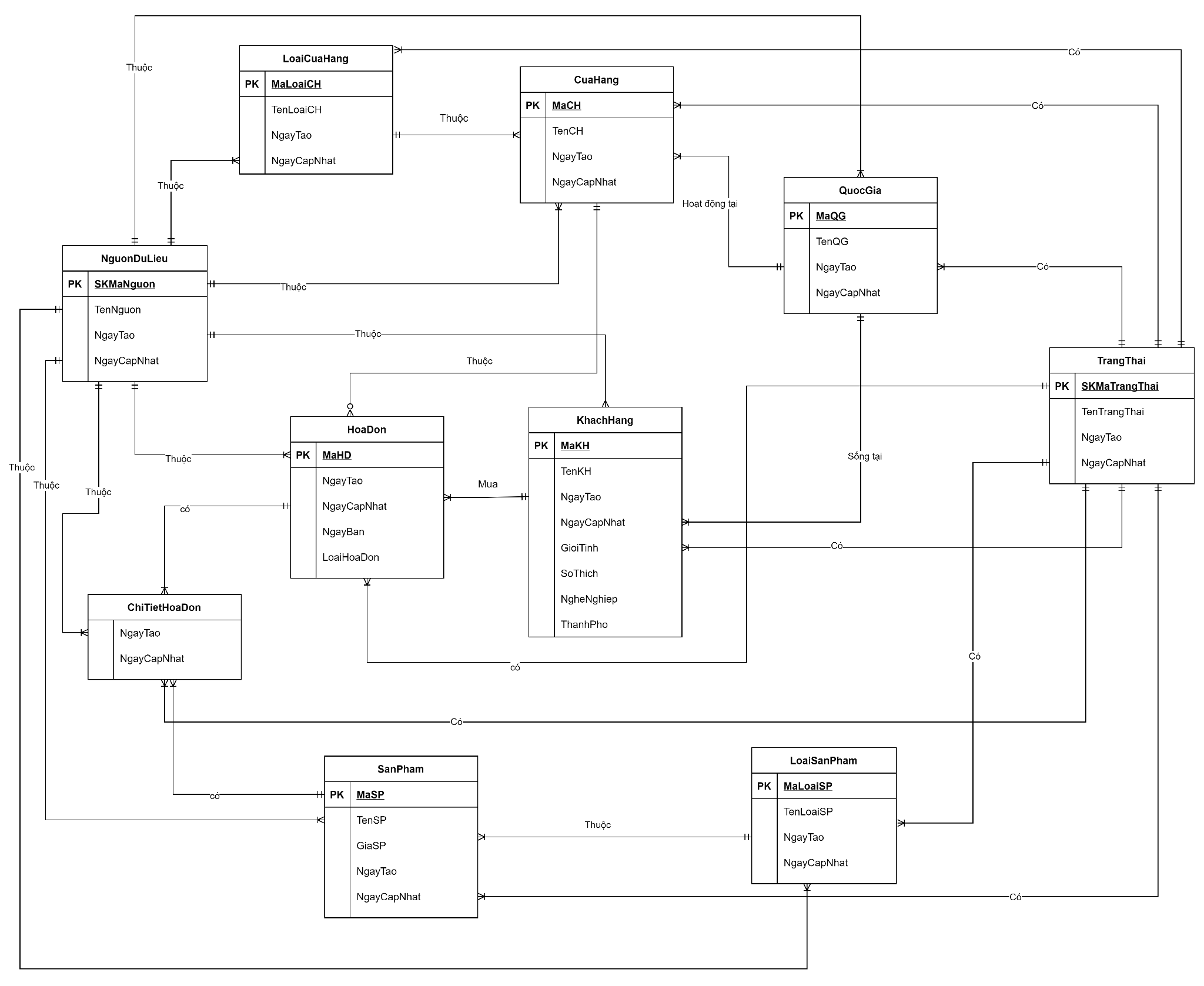
****

* + ETL từ Source Webtower sang Stage

****

# Stage sang NDS

* Khi đổ cả ba nguồn (Jade, Jupiter, Webtower) vào Stage, dữ liệu trong cơ sở dữ liệu DALT\_Stage sẽ gồm các bảng
  + KhachHang\_Stage
  + CuaHang\_Stage
  + LoaiCuaHang\_Stage
  + HoaDon\_Stage
  + ChiTietHoaDon\_Stage
  + SanPham\_Stage
  + LoaiSanPham\_Stage
  + Nuoc\_Stage
  + ChiTietHoaDon\_STAGE\_Webtower
  + HoaDon\_STAGE\_Webtower
  + KhachHang\_STAGE\_Webtower
* Ở giai đoạn từ source sang stage, nhóm vẫn sẽ chưa trộn dữ liệu của KhachHang, HoaDon và ChiTietHoaDon của 2 nguồn Jade và Webtower lại với nhau vì nếu làm như vậy thì sẽ rất khó để phân biệt được cơ sở dữ liệu nguồn của dữ liệu khi dữ liệu đã được gộp chung**.** Với nguồn dữ liệu đến từ Webtower sẽ đặt tên table có đuôi là STAGE\_Webtower. Vì hai nguồn dữ liệu từ Jade và Jupiter có dữ liệu khác nhau thì chỉ cần để tên fotmat là <Tên bảng>\_Stage
* Theo khái niệm, NDS được chuẩn hoá từ 3NF trở lên. Ta có dữ liệu gồm ba nguồn (Jade, Webtower, Jupiter) chưa được chuẩn hoá vì vậy cần chuẩn hoá dữ liệu để đáp ứng cấu trúc trong NDS.
* Một vài sự thay đổi dữ liệu chuyển từ giai đoạn Stage sang NDS:
* Bảng KhachHang\_Stage và KhachHang\_Webtower9:
* Tách thuộc tính DiaChi ra thành 2 cột Thành phố và MaQG, trong đó, dữ liệu cho cột ThanhPho nằm ở phần sau dấu “-” của giá trị cột DiaChi gốc, và giá trị của cột MaQG sẽ nằm ở phần trước dấu “-” của giá trị DiaChi gốc và tham chiếu tới khóa chính MaQG của bảng QuocGia. Chẳng hạn, với dòng dữ liệu có MaKH là ‘008V’ thì ở cột DiaChi là ‘Kazakhstan - Alabama’, sau khi chuyển từ giai đoạn stage sang NDS thì phần sau “-” là ‘Alabama’ được bỏ vô cột ThanhPho, phần nằm trước dấu “-” thì tham chiếu tới khóa chính MaQG của bảng QuocGia có TenQG là “Kazakhstan”.
* Bảng Nuoc\_Stage:
* Đổi tên thành QuocGia, hai thuộc tính MaNuoc, TenNuoc -> MaQG, TenQG.
* Thêm bảng TrangThai: bổ sung thêm bảng để làm rõ trangthai (ở đây đang xét ngữ cảnh là dòng dữ liệu đã bị xóa “soft deleted” hay chưa, false là khi dòng dữ liệu đó chưa xóa.
* Thêm bảng NguonDuLieu: bổ sung thêm bảng để nắm rõ nguồn dữ liệu đến từ Webtower, Jupiter hay Jade.
* **Hình vẽ dưới thể hiện rõ hơn cơ sở dữ liệu sau khi chuẩn hóa thành NDS dưới dạng mô hình ER,** ta sẽ dễ dàng thấy rõ sự thay đổi dữ liệu sau khi chuẩn hóa đã được liệt ở trên.

****

* **Mô tả dữ liệu:**

1. CuaHang (MaCH, TenCH, NgayTao, NgayCapNhat, MaLoaiCH, MaQG, SKMaNguon, SKMaTrangThai)

PHỤ THUỘC HÀM F = { MaCH -> TenCH, NgayTao, NgayCapNhat, MaLoaiCH, MaQG, SKMaNguon, SKMaTrangThai }

DẠNG CHUẨN: BCK

1. LoaiCuaHang (MaLoaiCH, TenLoaiCH, NgayTao, NgayCapNhat, SKMaTrangThai, SKMaNguon)

PHỤ THUỘC HÀM F = {MaLoaiCH -> TenLoaiCH, NgayTao, NgayCapNhat, SKMaTrangThai, SKMaNguon}

DẠNG CHUẨN: BCK

1. QuocGia (MaQG, TenQG, NgayTao, NgayCapNhat, SKMaTrangThai, SKMaNguon)

PHỤ THUỘC HÀM F = {MaQG -> TenQG, NgayTao, NgayCapNhat, SKMaTrangThai, SKMaNguon}

DẠNG CHUẨN: BCK

1. KhachHang ( MaKH, TenKH, NgayTao, NgayCapNhat, GioiTinh, SoThich, NgheNghiep, ThanhPho, MaQG, SKMaNguon, SKMaTrangThai)

PHỤ THUỘC HÀM F = {MaKH -> TenKH, NgayTao, NgayCapNhat, GioiTinh, SoThich, NgheNghiep, ThanhPho, MaQG, SKMaNguon, SKMaTrangThai }

DẠNG CHUẨN: BCK

1. SanPham ( MaSP, TenSP, GiaSP, NgayTao, NgayCapNhat, MaLoaiSP, SKMaNguon, SKMaTrangThai)

PHỤ THUỘC HÀM F = {MaSP -> TenSP, GiaSP, NgayTao, NgayCapNhat, MaLoaiSP, SKMaNguon, SKMaTrangThai}

DẠNG CHUẨN: BCK

1. LoaiSanPham (MaLoaiSP, TenLoaiSP, NgayTao, NgayCapNhat, SKMaNguon, SKMaTrangThai)

PHỤ THUỘC HÀM F = { MaLoaiSP > TenLoaiSP, NgayTao, NgayCapNhat, SKMaNguon, SKMaTrangThai}

DẠNG CHUẨN: BCK

1. HoaDon (MaHD, NgayBan, NgayCapNhat, NgayTao, LoaiHoaDon, MaCH, MaKH, SKMaNguon, SKMaTrangThai )

PHỤ THUỘC HÀM F = {MaHD -> NgayBan, NgayCapNhat, NgayTao, LoaiHoaDon MaCH, MaKH, SKMaNguon, SKMaTrangThai }

DẠNG CHUẨN: BCK

1. ChiTietHoaDon (MaHD, MaSP, NgayTao, NgayCapNhat, SKMaNguon, SKMaTrangThai)

PHỤ THUỘC HÀM F = {MaHD, MaSP -> NgayTao, NgayCapNhat, SKMaNguon, SKMaTrangThai}

DẠNG CHUẨN: BCK

1. NguonDuLieu(SKMaNguon, TenNguon, NgayTao, NgayCapNhat)

PHỤ THUỘC HÀM F = {SKMaNguon -> TenNguon, NgayTao, NgayCapNhat}

DẠNG CHUẨN: BCK

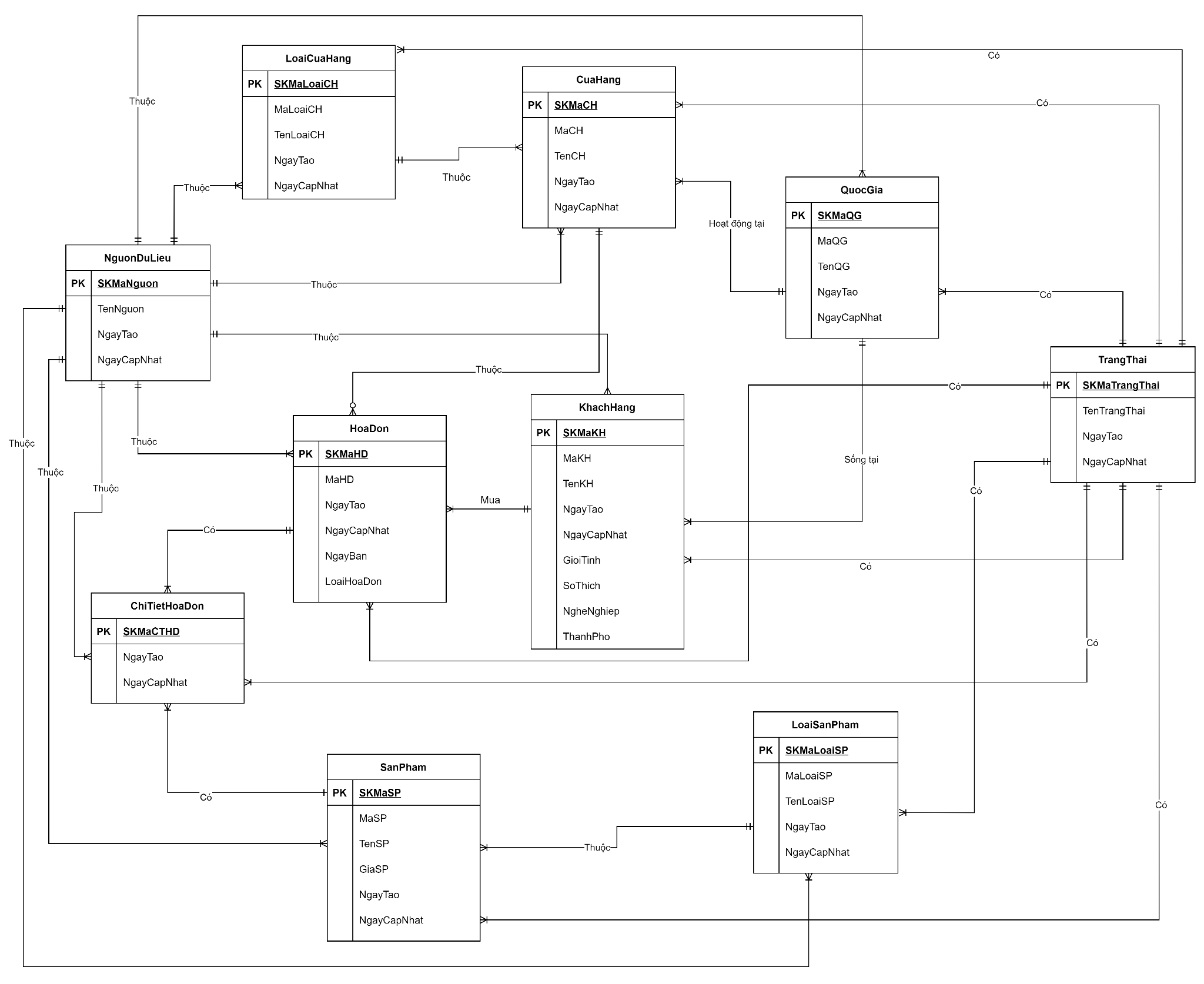
1. TrangThai(SKMaTrangThai, TenTrangThai, NgayTao, NgayCapNhat)

PHỤ THUỘC HÀM F = {SKMaTrangThai -> TenTrangThai, NgayTao, NgayCapNhat}

DẠNG CHUẨN: BCK

**=> Kết luận: Như vậy, lược đồ tổng thể đạt dạng chuẩn BCK, đáp ứng cấu trúc trong NDS**

* **Lược đồ quan hệ dưới dạng physical:** Khi cài đặt ở mức vật lý, để tránh tình trạng sai sót và thất thoát dữ liệu trong tương lai khi các mã cửa hàng, mã quốc gia hay mã loại cửa hàng bị thay đổi sẽ ảnh hưởng đến tham chiếu khóa ngoại. Nhóm chúng em quyết định tạo thêm một trường khóa tự tăng có tiền tố là SK ở mỗi bảng với kiểu dữ liệu là int. Khi đó, các bảng khác nếu có tham chiếu sẽ tham chiếu tới các giá trị có kiểu dữ liệu là int này, và ta cũng không cần phải lo lắng nếu trong tương lai các mã thực thể kia có thay đổi.

****

* Cấu trúc bảng CuaHang (WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Mô tả** | **Nguồn** | **Transformation** |
| SKMaCH | Khóa tự tăng của bảng CuaHang |  |  |
| MaCH | Mã cửa hàng trong source system | Ja.HoaDon\_Stage.MaCH  Wt.HoaDon\_STAGE\_Webtower.MaCH  Ju.CuaHang\_Stage.MaCH |  |
| TenCH | Tên của cửa hàng | Ju.CuaHang\_Stage.TenCH |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaLoaiCH | Mã loại cửa hàng cho biết cửa hàng thuộc loại cửa hàng nào | Tham chiếu khóa ngoại từ bảng LoaiCuaHang |  |
| SKMaQG | Mã quốc gia mà khách hàng đang sống | Tham chiếu khóa ngoại từ bảng QuocGia |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng LoaiCuaHang (WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Mô tả** | **Nguồn** | **Transformation** |
| SKMaLoaiCH | Khóa tự tăng của bảng LoaiCuaHang |  |  |
| MaLoaiCH | Mã loại cửa hàng trong source system | Ju.LoaiCuaHang\_Stage.MaLoaiCH |  |
| TenLoaiCH | Tên của loại cửa hàng | Ju.LoaiCuaHang\_Stage.TenLoaiCH |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng KhachHang(WebTower (Wt), Jade (Ja), Jupiter (Ju) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaKH | Khoá tự tăng của bảng KhachHang |  |  |
| MaKH | Mã khách hàng trong source system | Ja.KhachHang\_Stage.MaKH  Wt.KhachHang\_STAGE\_Webtower.MaKH |  |
| TenKH | Tên của khách hàng | Ja.KhachHang\_Stage.TenKH  Wt.KhachHang\_STAGE\_Webtower.TenKH |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| GioiTinh | Giới tính của khách hàng | Ja.KhachHang\_Stage.GioiTinh  Wt.KhachHang\_STAGE\_Webtower.GioiTinh |  |
| SoThich | Sở thích của khách hàng | Ja.KhachHang\_Stage.SoThich  Wt.KhachHang\_STAGE\_Webtower.SoThich |  |
| NgheNghiep | Nghề nghiệp của khách hàng | Ja.KhachHang\_Stage.NgheNghiep  Wt.KhachHang\_STAGE\_Webtower.NgheNghiep |  |
| ThanhPho | Thành phố mà khách hàng đang sống | Ja.KhachHang\_Stage.DiaChi  Wt.KhachHang\_STAGE\_Webtower.DiaChi | Đổi từ fomat “Nuoc-ThanhPho” thành fomat “ThanhPho” |
| SKMaQG | Mã quốc gia mà khách hàng đang sống | Tham chiếu khóa ngoại từ bảng QuocGia |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng QuocGia(WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaQG | Khoá tự tăng của bảng QuocGia |  |  |
| MaQG | Mã Quốc Gia trong source system | Ju.Nuoc\_Stage.MaNuoc |  |
| TenQG | Tên của quốc gia | Ju.Nuoc\_Stage.TenNuoc |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng SanPham(WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaSP | Khoá tự tăng của bảng SanPham |  |  |
| MaSP | Mã sản phẩm trong source system | Ju.SanPham\_Stage.MaSP |  |
| TenSP | Tên của sản phẩm | Ju.SanPham\_Stage.TenSP |  |
| GiaSP | Giá của sản phẩm | Ju.SanPham\_Stage.GiaSP |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaLoaiSP | Mã loại sản phẩm cho biết sản phẩm thuộc loại sản phẩm nào | Tham chiếu khóa ngoại từ bảng LoaiSanPham |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng LoaiSanPham(WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaLoaiSP | Khoá tự tăng của bảng LoaiSanPham |  |  |
| MaLoaiSP | Mã loại sản phẩm trong source system | Ju.LoaiSanPham\_Stage.MaLoaiSP |  |
| TenLoaiSP | Tên của loại sản phẩm | Ju.LoaiSanPham\_Stage.TenLoaiSP |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng HoaDon (WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaHD | Khoá tự tăng của bảng HoaDon |  |  |
| MaHD | Mã hóa đơn trong source system | Ja.HoaDon\_Stage.MaHD  Wt.HoaDon\_STAGE\_Webtower.MaHD |  |
| Ngày bán | Ngày lập hóa đơn bán hàng | Ja.HoaDon\_Stage.NgayBan  Wt.HoaDon\_STAGE\_Webtower.NgayBan |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| LoaiHoaDon | Mã của loại hóa đơn | Tham chiếu khóa ngoại từ bảng LoaiHoaDon |  |
| SKMaCH | Mã của cửa hàng tạo hóa đơn | Tham chiếu khóa ngoại từ bảng CuaHang |  |
| SKMaKH | Mã của khách hàng mua | Tham chiếu khóa ngoại từ bảng KhachHang |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng ChiTietHoaDon(WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaHD | Khoá tự tăng của bảng HoaDon |  |  |
| SKMaSP | Khoá tự tăng của bảng SanPham |  |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |
| SKMaNguon | Nguồn chứa record | 1 đối với Jupiter, 2 đối với Jade, 3 đối với WebTower |  |
| SKMaTrangThai | Trạng thái của dòng record | Tham chiếu khóa ngoại từ bảng TrangThai |  |

* Cấu trúc bảng NguonDuLieu (WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaNguon | Khóa tự tăng của bảng NguonDuLieu |  |  |
| TenNguon | Tên của nguồn dữ liệu |  |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |

* Cấu trúc bảng TrangThai (WebTower (Wt), Jade (Ja), Jupiter (Jupiter) trong NDS

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| SKMaTrangThai | Khóa tự tăng của bảng TrangThai |  |  |
| TenTrangThai | Tên của trạng thái |  |  |
| NgayTao | Thời gian mà record được tạo trong Dimension Data Store |  |  |
| NgayCapNhat | Thời gian mà record được cập nhật trong Dimension Data Store |  |  |

# Tài liệu tham khảo

<https://drive.google.com/file/d/1N6BIL4k77qrnY2HxsUee2-sBJkfq1wun/view?usp=sharing>