ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGÀNH CÔNG NGHỆ THÔNG TIN



HW2 REPORT

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

# Mục lục

[I. Mục lục 1](#_Toc86265272)

[I. Bảng thông tin chi tiết nhóm 2](#_Toc86265273)

[II. Phân công công việc 3](#_Toc86265274)

[III. ETL từ Source sang Stage 4](#_Toc86265275)

[IV. Stage sang NDS 5](#_Toc86265276)

[V. Tài liệu tham khảo 10](#_Toc86265277)

# 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

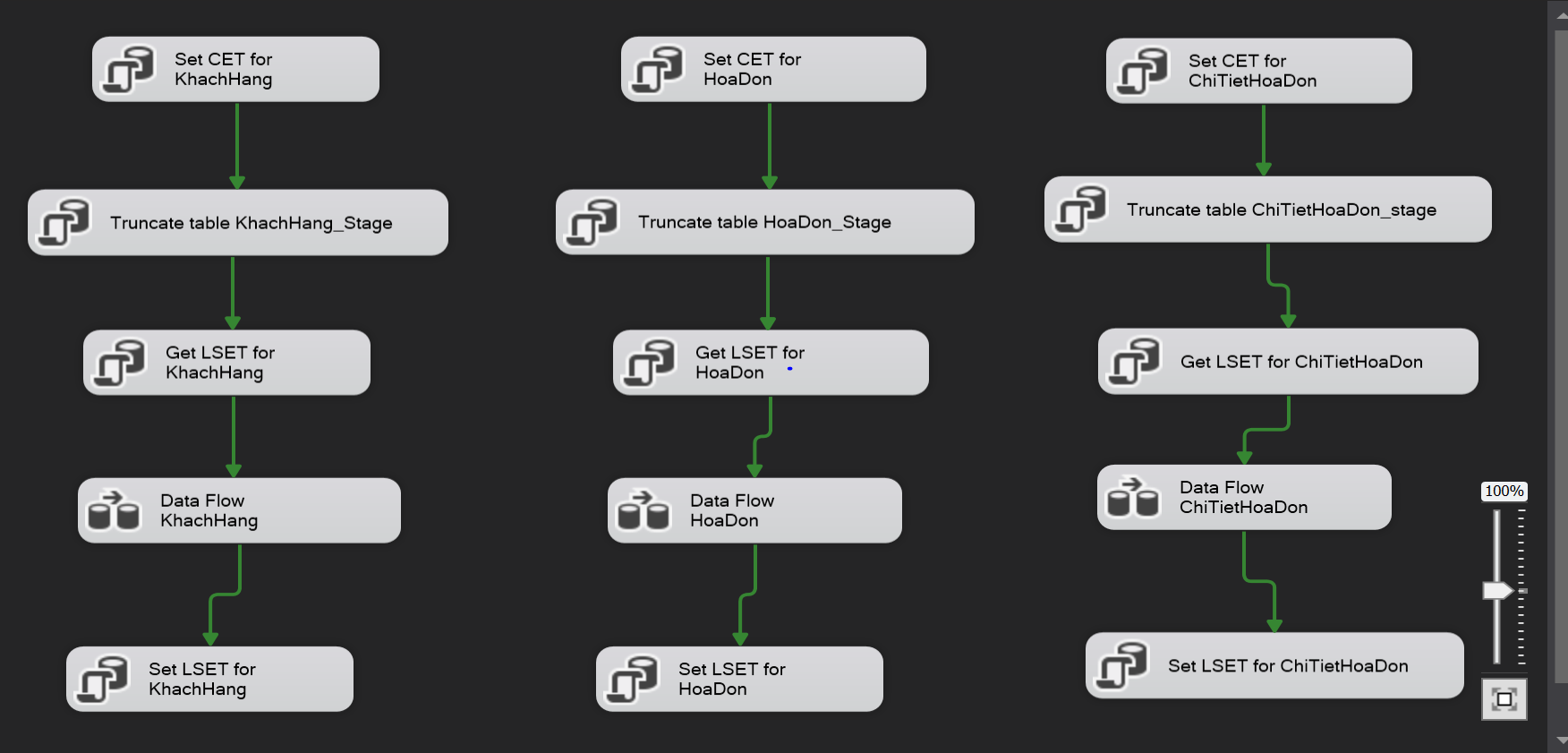
|  |  |  |
| --- | --- | --- |
| Họ tên thành viên | Nội dung công việc | Phần trăm hoàn thành (%) |
| Phạm Ngọc Thùy Trang |  |  |
| Ngô Thị Thanh Thảo |  |  |
| Hoàng Nguyên Trúc |  |  |

# ETL từ Source sang Stage

* Chọn source Jade để phục vụ cho việc ETL dữ liệu từ Source sang State
* Chiến lược rút trích dữ liệu từ cơ sở dữ liệu Jade: 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:

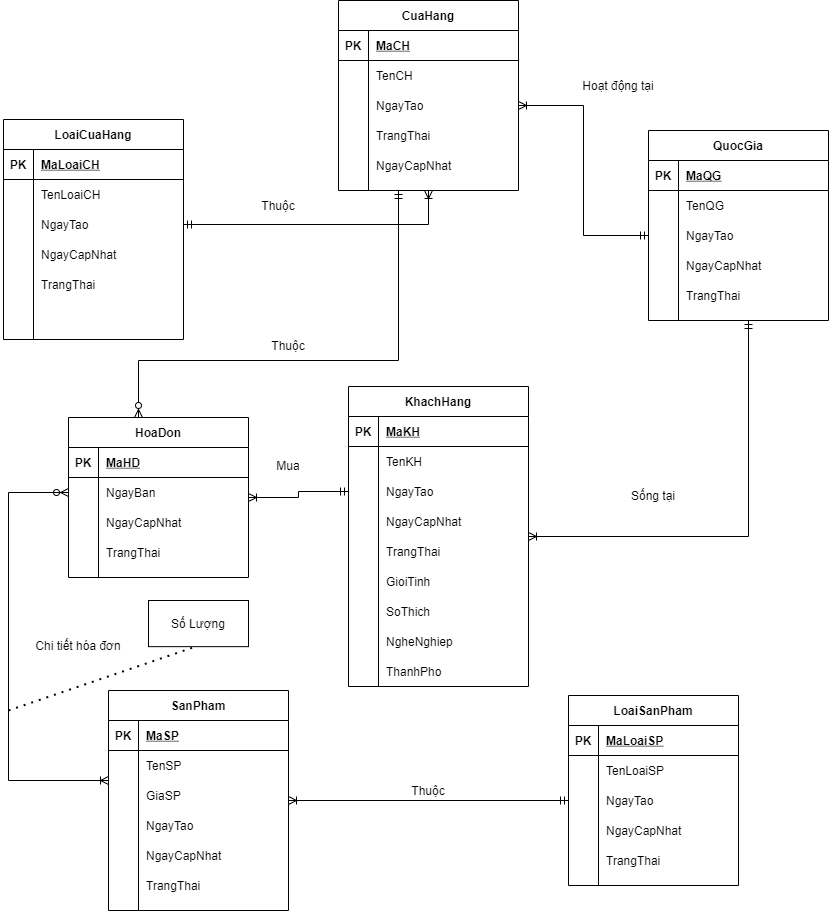
1. Lấy thông tin LSET được lưu trong metadata
2. Lấy CET bằng thời gian khởi động package
3. Truncate bảng Stage hiện tại: TRUNCATE TABLE <TenBang>
4. Rút trích dữ liệu từ bảng Source sang bảng Stage: select \* from BangSource where (NgayTao >= LSET and NgayTao < CET) or (NgayCapNhat >= LSET and NgayCapNhat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

* Flow khi thực hiện trên Visual Studio Code



# Stage sang NDS

* Giả sử khi đổ cả ba nguồn (Jade, Webtower, Jupiter) 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
* 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:
* 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 ChiTietHoaDon: bổ sung thêm thuộc tính SoLuong để làm rõ mỗi sản phẩm trong hóa đơn khách hàng đã mua bao nhiêu, mặc định là 0 do ban đầu nguồn không có thuộc tính này -> Vì vậy bảng gồm các thuộc tính MaHD, MaSP, SoLuong.
* Hình vẽ dưới thể hiện rõ hơn cơ sở dữ liệu sau khi chuẩn hóa thành NDS dưới dạng lược đồ quan hệ, 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, TrangThai, NgayCapNhat, MaLoaiCH, MaQG)

PHỤ THUỘC HÀM F = {MaCH -> TenCH, NgayTao, TrangThai, NgayCapNhat,

MaLoaiCH, MaQG}

DẠNG CHUẨN: BCK

1. LoaiCuaHang (MaLoaiCH, TenLoaiCH, NgayTao, NgayCapNhat, TrangThai)

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

TrangThai}

DẠNG CHUẨN: BCK

1. QuocGia (MaQG, TenQG, NgayTao, NgayCapNhat, TrangThai)

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

DẠNG CHUẨN: BCK

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

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

DẠNG CHUẨN: BCK

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

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

DẠNG CHUẨN: BCK

1. LoaiSanPham (MaLoaiSP, TenLoaiSP, NgayTao, NgayCapNhat, TrangThai)

PHỤ THUỘC HÀM F = {MaLoaiSP -> TenLoaiSP, NgayTao, NgayCapNhat,

TrangThai}

DẠNG CHUẨN: BCK

1. HoaDon (MaHD, NgayBan, NgayCapNhat, TrangThai, MaCH, MaKH)

PHỤ THUỘC HÀM F = {MaHD -> NgayBan, NgayCapNhat, TrangThai, MaCH,

MaKH}

DẠNG CHUẨN: BCK

1. ChiTietHoaDon (MaHD, MaSP, SoLuong)

PHỤ THUỘC HÀM F = {MaHD, MaSP -> SoLuong}

DẠNG CHUẨN: BCK

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

# Tài liệu tham khảo

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