**Tài liệu Thiết kế Kiến Trúc Hệ Thống**

|  |  |
| --- | --- |
| **Dự án** | *Dự án triển khai hệ thống Datawarehouse* |
| **Phiên bản** |  |
| **Ngày cập nhật** |  |
| **Người viết** | Hồ Hoàng Quân |

| Thay đổi bản ghi | | | |
| --- | --- | --- | --- |
| Ngày | Người viết | Phiên bản | Nội dung sửa đổi |
| 25/08/2023 | Hồ Hoàng Quân |  | Cập nhật mô tả kiến trúc giải pháp và xây dựng data warehouse |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Content

[1 Sơ đồ kiến trúc giải pháp 3](#_Toc144021414)

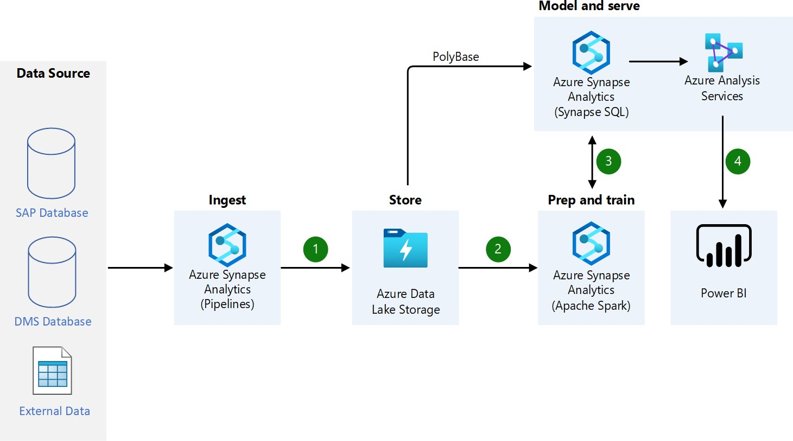
[2 Xây dựng data warehouse 4](#_Toc144021415)

[2.1 Cấu trúc data warehouse 4](#_Toc144021416)

[2.2 Theo dõi hoạt động trên Data Warehouse 9](#_Toc144021417)

[2.3 Kiểm tra và đối chiếu dữ liệu nguồn 10](#_Toc144021418)

# 1 Sơ đồ kiến trúc giải pháp



​

* Triển khai hệ thống ETL dựa trên nền tảng của Azure Synapse Analytics​
* Triển khai hệ thống Data WareHouse dựa trên nền tảng của Azure Synapse Analytics​
* Hệ thống ghi nhận dữ liệu tự động và hệ thống báo cáo thông minh được xây dựng trên nền tảng Microsoft Power BI

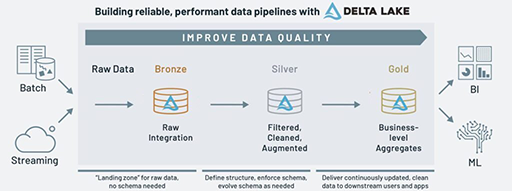
# Xây dựng data warehouse

## 2.1 Cấu trúc data warehouse

Sử dụng kiến trúc Medallion để xây dựng data warehouse.

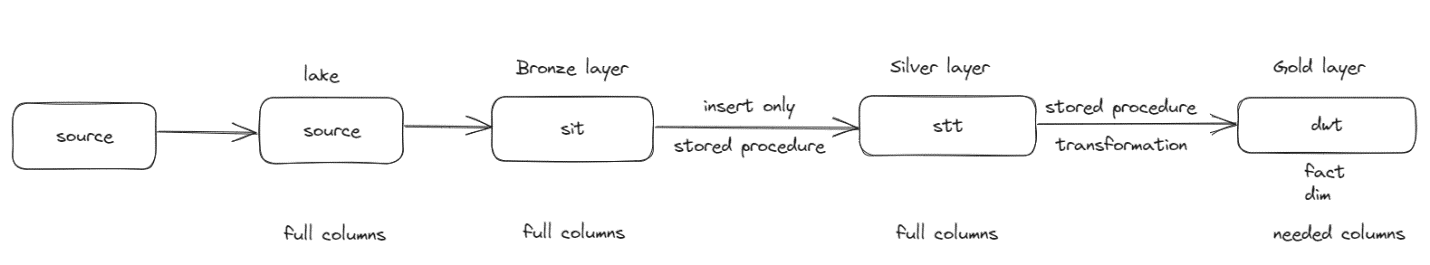
Medallion được sử dụng để tổ chức dữ liệu một cách hợp lý với mục tiêu cải thiện dần dần cấu trúc và chất lượng của dữ liệu khi nó chảy qua từng lớp của kiến ​​trúc (từ bảng lớp Đồng (Bronze) ⇒ Bạc (Silver) ⇒ Vàng(Gold) ). Việc phát triển dần qua từng lớp sẽ giúp người sử dụng:

* Dễ dàng phát hiện vấn đề khi phát sinh lỗi.
* Dễ sử dụng và phát triển trong tương lai. Lớp Bronze và Silver chứa đầy đủ tất cả dữ liệu và được lược bỏ khi qua lớp Gold là lớp tiêu thụ dữ liệu (Power BI lấy data ở lớp Gold). Nên người sử dụngchỉ cần lấy thêm data từ lớp Silver thay vì phải bắt đầu lấy lại và chỉnh sửa từ nguồn.

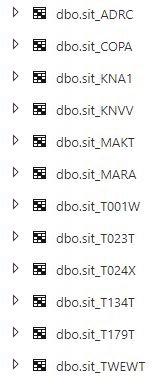


(Nguồn: <https://www.databricks.com/glossary/medallion-architecture>)

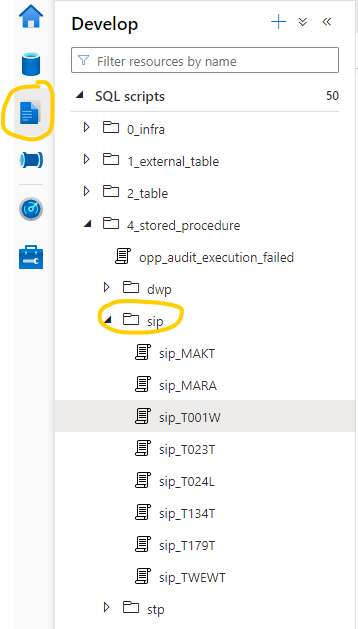
Phương pháp triển khai medallion architect sử dụng dịch vụ cloud Azure Synapse Analytics.



1. Dữ liệu được lấy từ source (SAP, DMS,OneDrive) và lưu lại dưới dạng flat file hoặc parquet trên Azure Datalake.
2. Tạo External table (dùng để đọc các file vật lý như các table bằng SQL) trên Datawarehouse (Dedicated SQL Pool) nằm trong Azure Synapse. Các External Tables này sẽ store dữ liệu hệt như source nguồn và đẩy qua các bảng sit là lớp Bronze. Các bảng sit này sẽ tương tự như các staging tables. Hằng ngày sẽ có script chạy lấy data mới nhất từ các external tables.

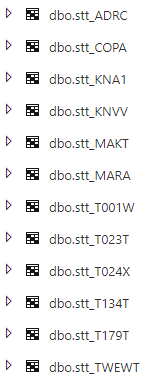
Data ở lớp Bronze sẽ được lưu dưới dạng sit\_TableName. Ví dụ:  


Code của script chạy lấy data từ external tables qua lớp Bronze nằm trong phần Develop > 4.stored\_procedure > sip

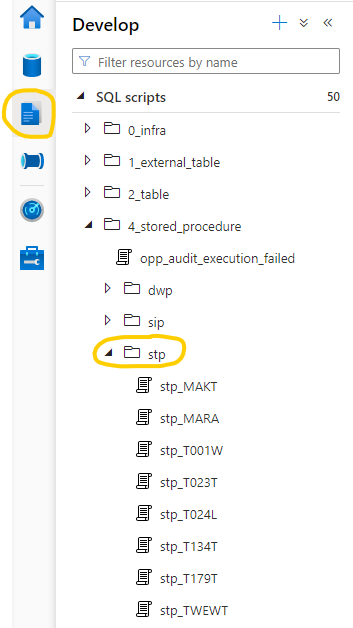


1. Data từ lớp Bronze (External Tables) sẽ được đẩy lên SIT Tables tức là lớp Silver. Lớp Silver sẽ lưu trữ dữ liệu history cho đến hiện tại (lớp Bronze chỉ lưu dữ liệu mới nhất). Hằng ngày hệ thống sẽ tự động chạy script insert để thêm vào các dòng dữ liệu từ file mới nhất trên hệ thống.

Data ở lớp silver sẽ được lưu dưới dạng stt\_TableName. Ví dụ:

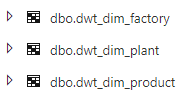


Code của script chạy lấy data từ Bronze (sit) qua Silver (stt) nằm trong phần Develop > 4.stored\_procedure > stp

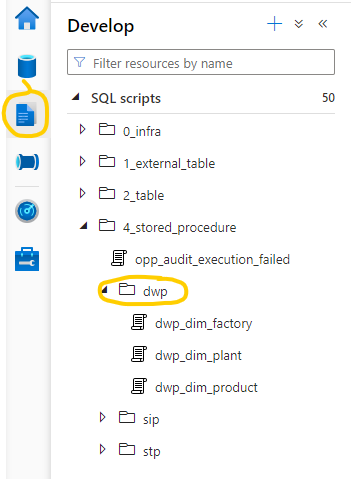


1. Tất cả dữ liệu ở lớp Silver sẽ được biến đổi theo các business logic tương ứng để ra được tables cuối cùng là lớp Gold. Lớp Gold sẽ là nơi data được tiêu thụ. Hằng ngày sẽ có script chạy tự động để cập nhật dữ liệu cũ, đưa vào dữ liệu mới và biến đổi các dữ liệu mới này theo các business logic để có thể sử dụng được.

Data ở lớp gold sẽ đượclưu dưới dạng dwt\_dim(fact)\_name. Ví dụ:

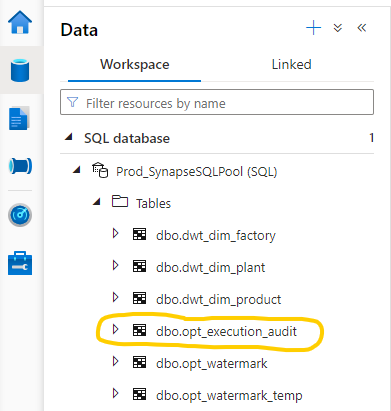


Code của script chạy lấy data từ Silver (stt), biến đổi và đẩy qua lớp Gold nằm trong phần Develop > 4.stored\_procedure > dwp

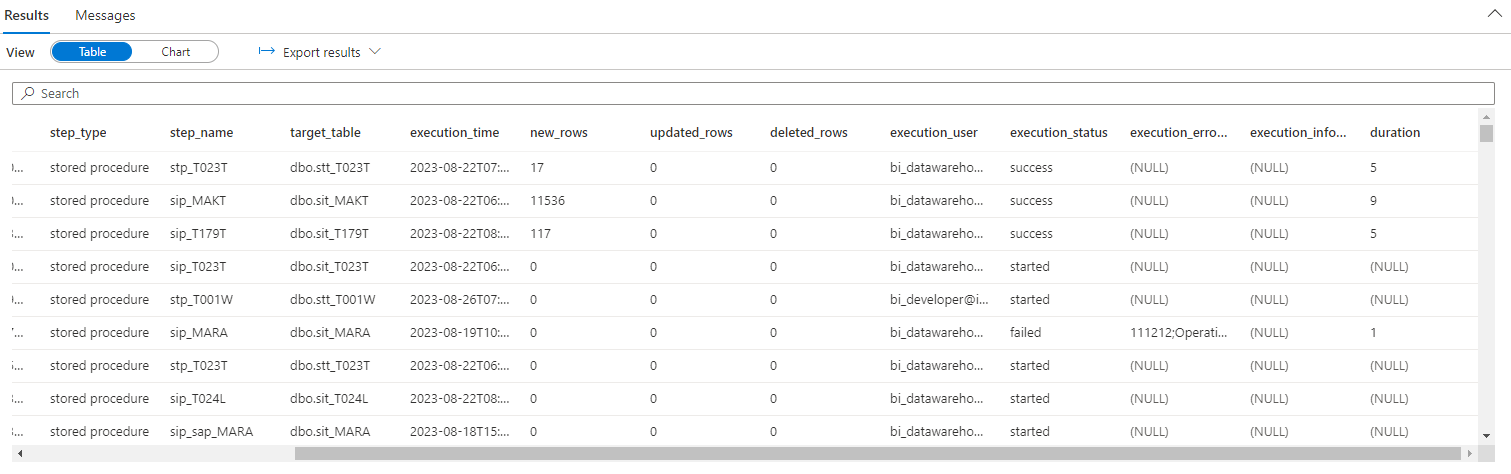


## 2.2 Theo dõi hoạt động trên Data Warehouse

Các hoạt động thêm, sửa, xóa, cập nhật và biến đổi dữ liệu hằng ngày đều được lưu lại trên Data Warehouse ở bảng dbo.opt\_executive\_audit



Kết quả sau khi query:



Một số cột cần lưu ý:

-step\_name: tên của hoạt động. Ví dụ stp\_T023T tức là chạy script lấy data từ lớp Bronze qua Silver của bảng T0023T. Xem mục 4.1 để hiểu quy tắc đặt tên.

-target\_table: tên của bảng hứng kết quả từ hoạt động thêm / xóa / chỉnh sửa /biến đổi

-execution\_time: thời gian hoạt động bắt đầu chạy

- new rows: số dòng mới được thêm vào từ hoạt động

- updated rows: số dòng được cập nhật từ hoạt động

- deleted rows: số dòng bị xóa từ hoạt động

- execution\_user: tên của người chạy hoạt động

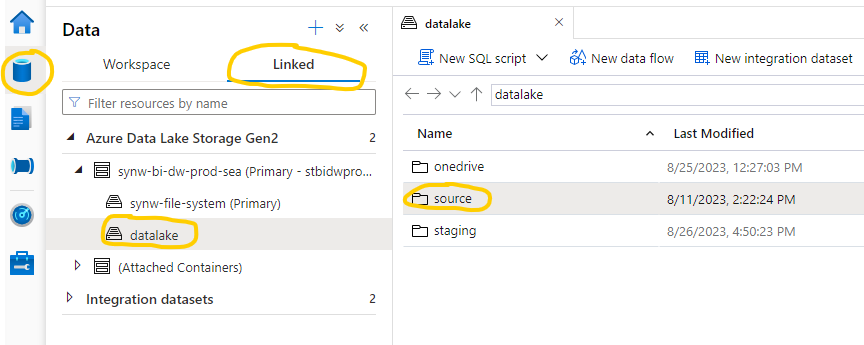
- execution\_status: kết quả của hoạt động (success hay failed)

- execution\_error: câu lỗi phát sinh nếu hoạt động failed

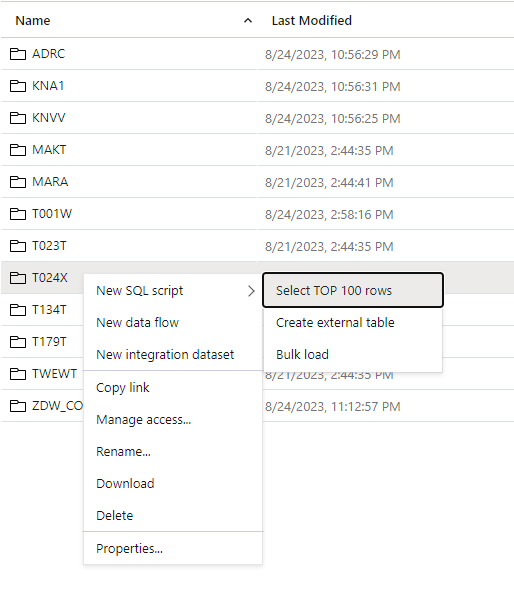
## 2.3 Kiểm tra và đối chiếu dữ liệu nguồn

Công nghệ của Azure Synapse Analytics cho phép người dùng truy vấn các file vật lý được lưu dưới data lake bằng SQL đế có thể dễ dàng đối chiếu với dữ liệu ở các hệ thống nguồn cũng như dữ liệu ở các lớp trên Data Warehouse.

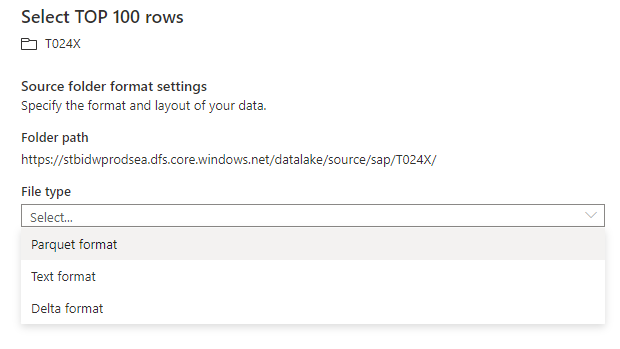
Vào mục Data > Linked > datalake > source/onedrive



Trong phần source/onedrive, tìm đến file hoặc thư mục muốn truy vấn và bấm chuột phải chọn New SQL script > Select TOP 100 rows



Chọn định dạng là Parque format



Viết câu SQL và chạy xem kết quả truy vấn

