**MedSynora DW – A Comprehensive Synthetic Hospital Patient Data Warehouse  
Group No:15**

**Higher National Diploma in Software Engineering 24.2F**

**Datawarehouse & Business Intelligence**

**Project Report**

**Submitted By:**

|  |  |  |
| --- | --- | --- |
| **Index Number** | **Name** | **Contribution** |
| COHNDSE242F-041 | Ilma M H F | ETL Process |
| COHNDSE242F-042 | Azward M A | Power BI page 01 – 03 |
| COHNDSE242F-092 | Ishini B G U | Schema Design Database Implementation Power BI page 04 |
| COHNDSE242F-104 | Jayamanne S D | Power BI page 01 – 03 |
| COHNDSE242F-114 | Chathama M A L | Schema Design Database Implementation Power BI page 04 |

****

**School of Computing and Engineering**

**National Institute of Business Management**

**Colombo 7**

Contents

[Chapter 1: Introduction 3](#_Toc203121471)

[1.1 Overview of the Business Domain 3](#_Toc203121472)

[1.2 Objective and Purpose 3](#_Toc203121473)

[1.3 Key Deliverables 4](#_Toc203121474)

[Chapter 2: Data Source Identification 5](#_Toc203121475)

[2.1 Description of Data 5](#_Toc203121476)

[2.2 Data Structures 6](#_Toc203121477)

[2.3 Justification for Data Source Selection 14](#_Toc203121478)

[Chapter 3: Data Warehouse Design 15](#_Toc203121479)

[3.1 Galaxy Schema Overview 15](#_Toc203121480)

[3.2 Fact Tables: Grain and Business Process 15](#_Toc203121481)

[3.3 Data Warehouse Design 16](#_Toc203121482)

[Chapter 4: ETL Process Using Apache Hop 22](#_Toc203121483)

[4.1 Tools Used 22](#_Toc203121484)

[4.2 ETL Workflow 22](#_Toc203121485)

[4.3 Source and Target Setup 30](#_Toc203121486)

[Chapter 5: Data Visualization Using Power BI 31](#_Toc203121487)

[5.1 Data Connection 31](#_Toc203121488)

[5.2 Visualizations Created 31](#_Toc203121489)

[5.3 Dashboard Layout 31](#_Toc203121490)

[5.4 Insights Discovered 34](#_Toc203121491)

[Chapter 6: Findings & Recommendations 35](#_Toc203121492)

[6.1 Key Business Insights 35](#_Toc203121493)

[6.2 Recommendations 35](#_Toc203121494)

[Chapter 7: Challenges & Solutions 36](#_Toc203121495)

[Chapter 8: Conclusion 37](#_Toc203121496)

[References 38](#_Toc203121497)

# Chapter 1: Introduction

## 1.1 Overview of the Business Domain

MedSynora DW is a corporate synthetic hospital data warehouse for the emulation of realistic clinical and administrative activities in a huge hospital. The dataset is a fine-grained health care activity, attention is emphasized on 'patient journey'. It comprises data of patient visits, diagnoses, labs, vitals, treatments, procedures or costs and spans throughout the year 2024.

This data warehouse further includes patient-centered activities, such as admissions, doctor encounters, room utilization, insurance status, and chronic diseases. The dataset is available on Kaggle and has been designed for data warehousing and analysis applications. It is of particular interest to data engineers, healthcare analysts and system designers who are interested in building clinical dashboards and healthcare intelligence systems. [**https://www.kaggle.com/datasets/mebrar21/medsynora-dw**](https://www.kaggle.com/datasets/mebrar21/medsynora-dw)

## 1.2 Objective and Purpose

**Objective:**

The goal of this work is to develop a database Data Warehouse to facilitate the efficient storage and retrieval of meaningful health-related data with the help of concepts, principles, and techniques of the Galaxy Schema model. The system is expected to facilitate decision-making and performance measurement in a hospital.

**Purpose:**

* To unify data from various sources into a single ' data warehouse '-capable of separating ' data-load ' from ' analysis ' and providing all the latest performance management facilities.
* To make data more available so that doctors, analysts and hospital administrators can actually use it.
* To be able to perform multi-dimensional data analysis using well defined fact and dimension tables.
* To monitor KPIs like cost per treatment, readmission rate, average LOS, etc.
* For reporting and strategic planning, to have similar and standardized data throughout the campus.

## Key Deliverables

1. Data Warehouse

Constructed with SQL Server Management Studio, tables created and populated from multiple sources (SQL, CSV, JSON, XML, Excel).

1. ETL Pipeline

Created in “Apache Hop”, in charge of data ETL (extract, transform, load)

1. Dashboard

Developed with “Power BI” for visualization of data i.e., KPI s, charts, filters, interactive document.

# Chapter 2: Data Source Identification

## 2.1 Description of Data

The MedSynora Data set is an integrated data collection of different structured data sources in a hospital. These include:

* Patient information and encounter history.
* Doctor assignments and room details.
* Diagnoses and treatments.
* Vitals and lab tests.
* Insurance claims.
* Cost breakdowns by treatment stage.

Sample attributes include:

* Encounter\_ID, Patient\_ID, Disease\_ID, ResponsibleDoctorID, InsuranceKey, RoomKey, CheckinDate, CheckoutDate, CheckinDateKey, CheckoutDateKey, Patient\_Severity\_Score, RadiologyType, RadiologyProcedureCount, EndoscopyType, EndoscopyProcedureCount, CompanionPresent

## 2.2 Data Structures

**Dimension Tables** (11 total)

* DimPatient (SQL)

A screenshot of a computer

AI-generated content may be incorrect.

* DimDoctor (CSV)

A screenshot of a computer

AI-generated content may be incorrect.

* DimDisease (XML)

A screenshot of a computer

AI-generated content may be incorrect.

* DimChronicDisease (Excel)

A screenshot of a computer

AI-generated content may be incorrect.

* DimAllergy (Excel)

A screenshot of a computer

AI-generated content may be incorrect.

* DimInsurance (CSV)

A computer screen shot of a computer

AI-generated content may be incorrect.

* DimRoom (CSV)

A screenshot of a computer

AI-generated content may be incorrect.

* DimDate (SQL)

A screenshot of a computer

AI-generated content may be incorrect.

* DimSpecialTest (JSON)

A screen shot of a computer

AI-generated content may be incorrect.

* DimTreatment (JSON)

A screen shot of a computer

AI-generated content may be incorrect.

* DimAdditionalService (XML)

A screenshot of a computer

AI-generated content may be incorrect.

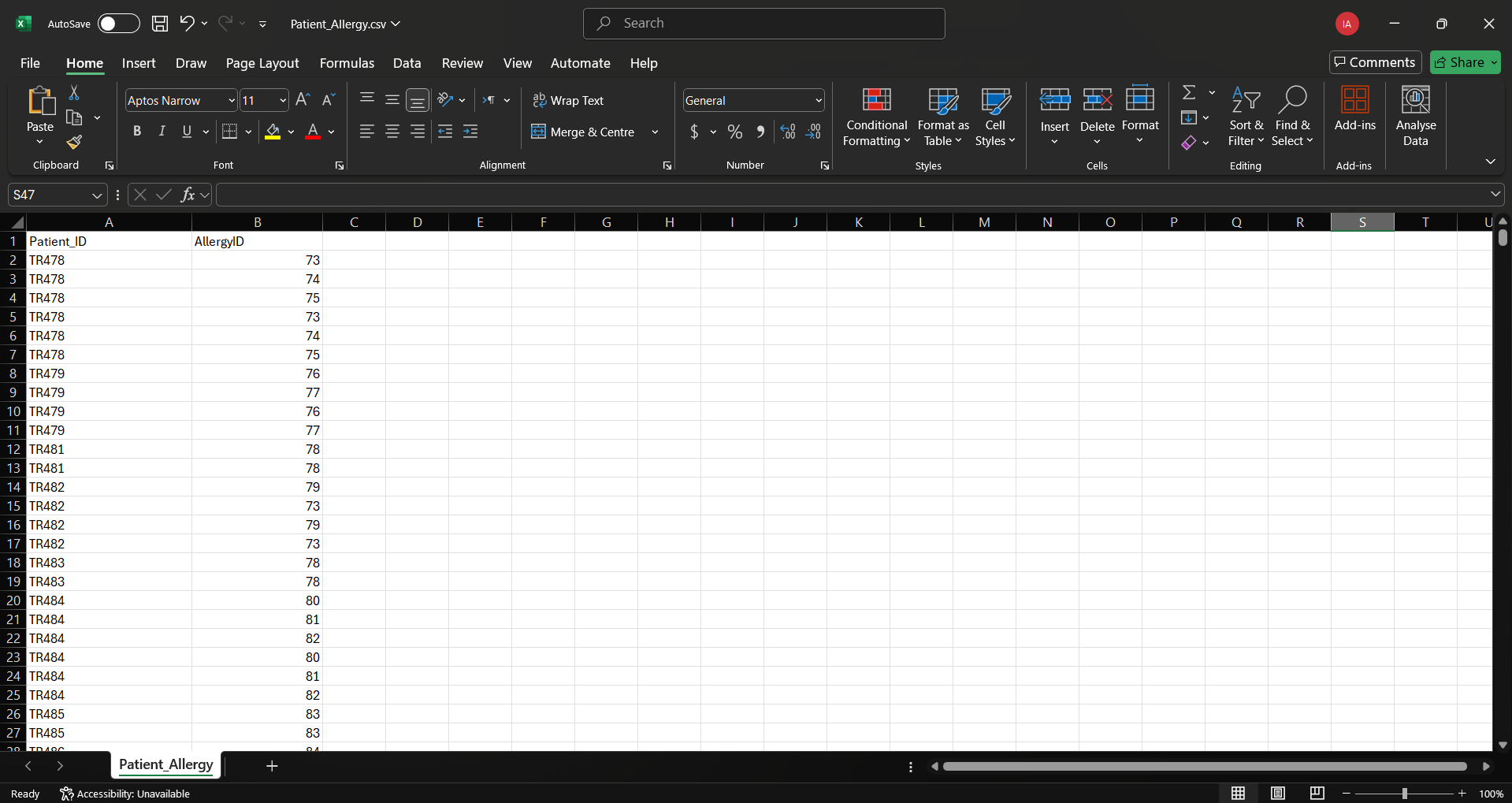
**Bridge Tables** (4 total, CSV format):

* Patient\_ChronicDisease

A screenshot of a computer

AI-generated content may be incorrect.

* Patient\_Allergy



* BridgeEncounterDoctor

A screenshot of a computer

AI-generated content may be incorrect.

* BridgeEncounterAdditionalService

A screenshot of a computer

AI-generated content may be incorrect.

**Fact Tables** (7 total):

* FactEncounter
* FactTreatment
* FactLabTests
* FactVitals
* FactCost
* FactProcedures
* FactPatientHealthService

## 2.3 Justification for Data Source Selection

The diversity of data formats (SQL, CSV, JSON, XML, Excel) mirrors the complexity of healthcare systems in real life, where data comes from EHRs, insurance systems, lab results and financial systems. This combination makes the warehouse powerful, practical, and scalable.

# Chapter 3: Data Warehouse Design

## 3.1 Galaxy Schema Overview

The warehouse design follows a **Galaxy Schema** (also called a fact constellation schema), which supports multiple fact tables sharing dimension tables. This model is ideal for healthcare data, where different types of measures (e.g., encounters, treatments, costs) share common dimensions like patient, date, doctor, and disease.

## 3.2 Fact Tables: Grain and Business Process

|  |  |  |
| --- | --- | --- |
| **Table Name** | **Grain** | **Business Process** |
| **FactEncounter** | One record per patient admission | Monitor check-in/out, when patients were in a room, insurance, doctor, and disease |
| **FactTreatment** | One record per treatment event | Records drug use, surgery price, therapy visits |
| **FactLabTests** | One record is generated for each laboratory test performed for each patient. | Tracks test name, result, and phase |
| **FactVitals** | One record for each vital sign per date | Records heart rate, temperature, O2 levels, BP |
| **FactCost** | One record for each cost entry and patient and day | Tracks billing costs across categories |
| **FactProcedures** | One record per procedure event | Monitors radiology, endoscopy and other procedures |
| **FactPatientHealthService** | One record per patient | Summarizes chronic diseases, allergies, services |

The fact tables enable stakeholders (decision makers) to evaluate hospital performance from different dimensions: clinical, operational, and financial.

## 3.3 Data Warehouse Design

In terms of data warehouse design, the following “Galaxy Schema” can be shown.

A diagram of a company

AI-generated content may be incorrect.

After the Galaxy Schema Design, the following database schema has designed using “Microsoft SQL Server Management Studio (SSMS).” The following queries are used to create the mentioned database schema.

1. Database Creation



1. Dimension Table Creation

A white background with green border

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Bridge Table Creation

A white background with text

AI-generated content may be incorrect.

1. Fact Table Creation

A computer screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

# Chapter 4: ETL Process Using Apache Hop

## 4.1 Tools Used

Apache Hop Lightweight open-source platform for orchestrating data pipelines to build ETL-runnable workloads on a scale. It is a GUI driven tool which also supports the designing of different pipelines and different data sources.

## 4.2 ETL Workflow

* **Extract:** Read data from the source (SQL, CSV, XML etc) and store the raw data in one place (Persistent storage).

Step 1: Extract dimension data from sources (SQL, CSV, XML, etc.)

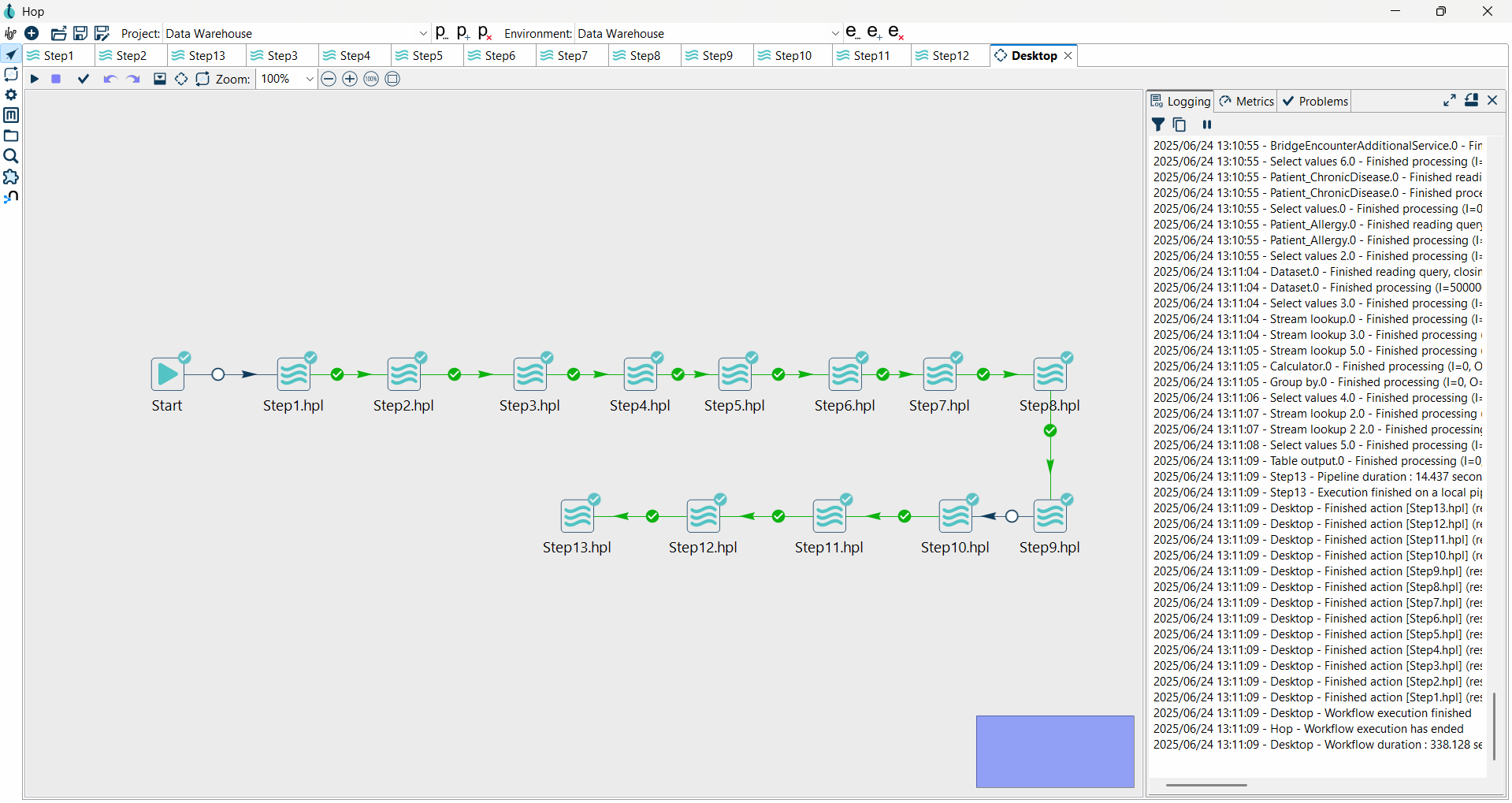
Step 2-4: Load Bridge tables (Patient\_ChronicDisease, Patient\_Allergy, Patient\_AdditionalService…...)

Step 5-6: Create bill and service assignment bridge tables by the doctor and service on encounter.

Step 7-13: Provide data into fact tables in sequence, along with keys and data range/integrity.

* **Transform:** Techniques such as data cleaning, aggregation, validation using ETL transforms were done to ensure the quality of data.
* **Load:** Transformed data were loaded to SQL Server Management Studio (SSMS) which is selected as a Database Schema.

**Workflow -**



**PIPLINES**

**STEP 1 –**

DimDate

DimPatient

DimDoctor

DimDisease

DimInsurance

DimRoom

DimChronicDisease

DimAllergy

DimSpecialTest

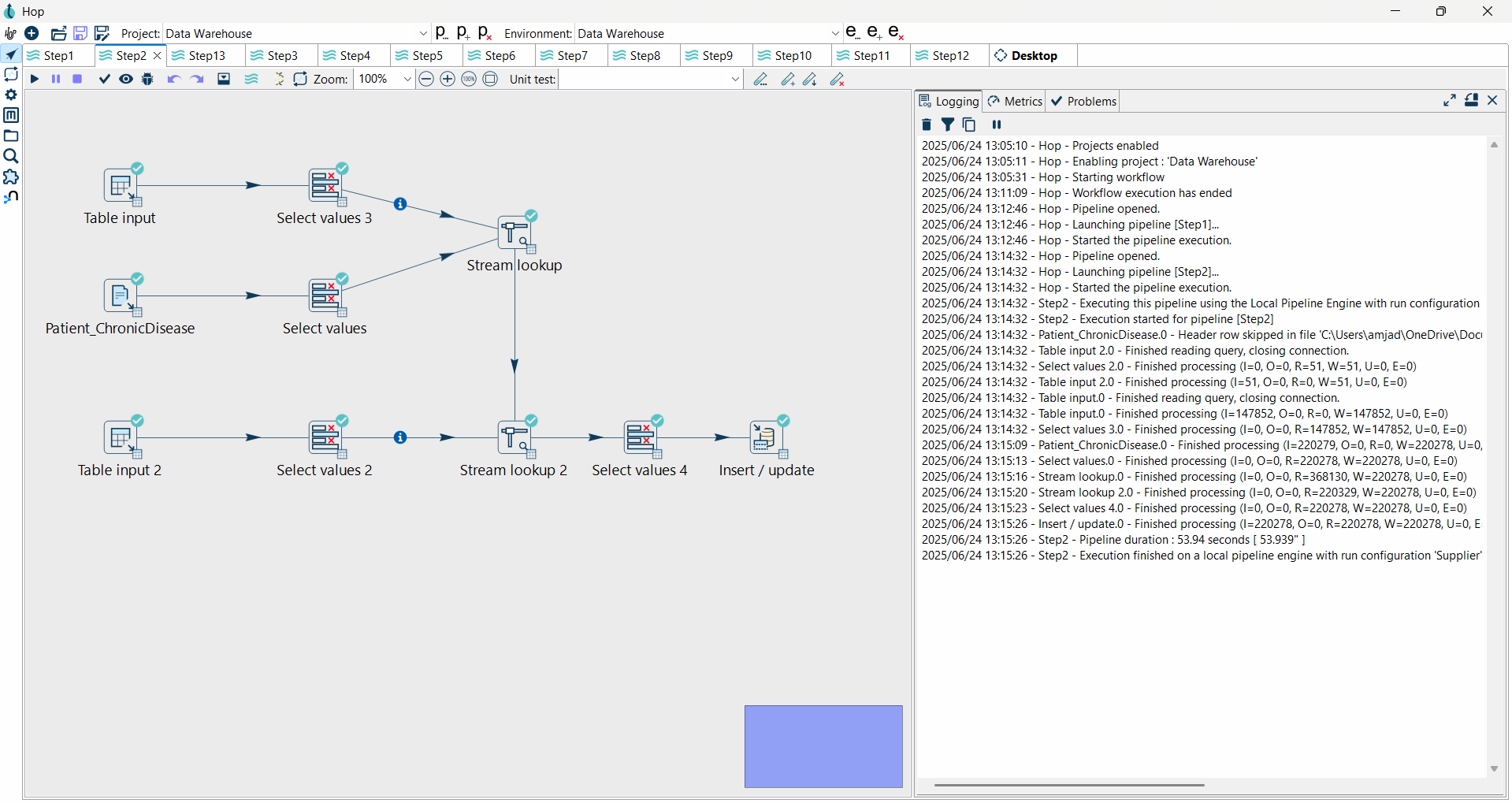
DimTreatment

DimAdditionalService

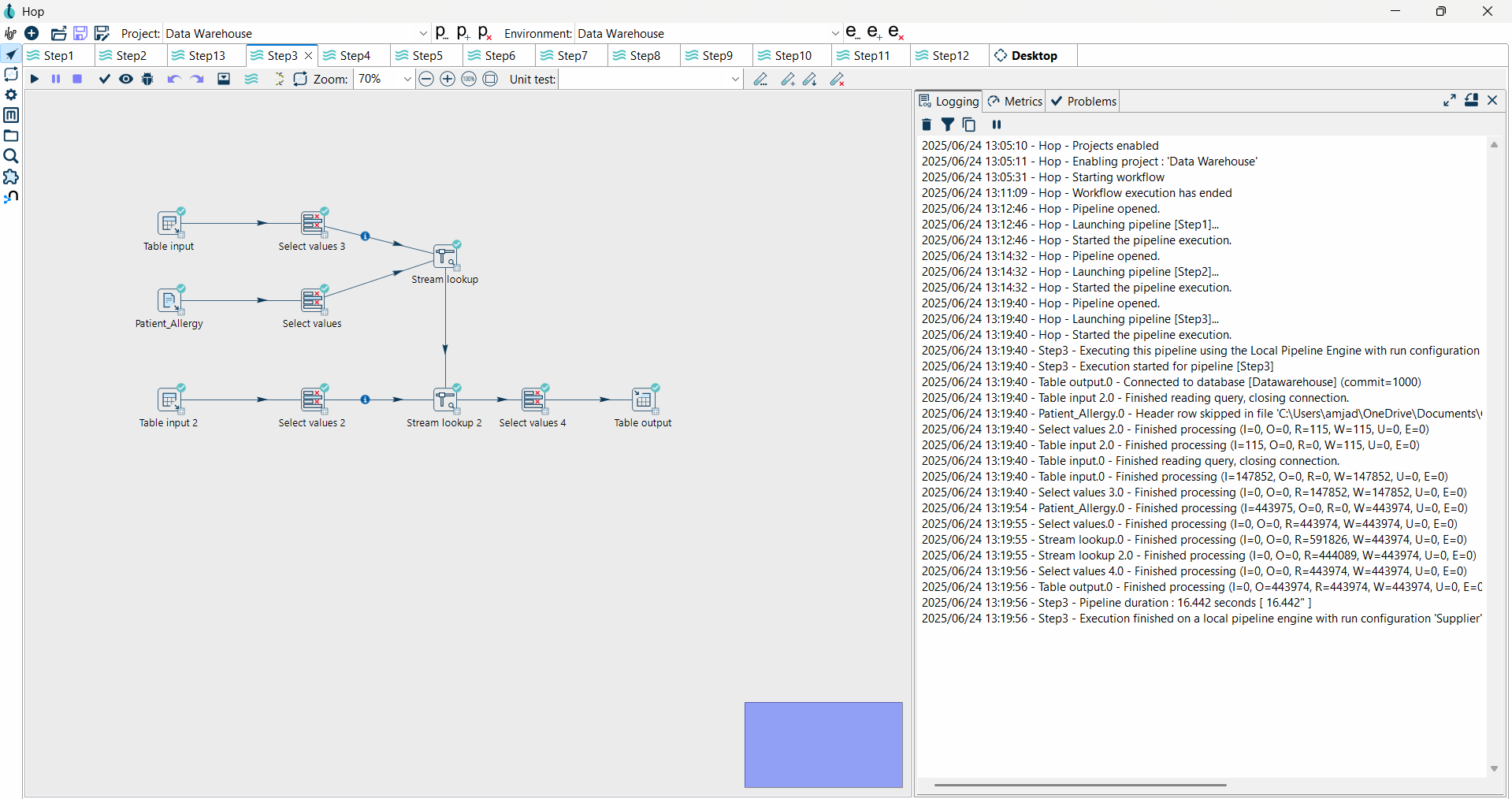
**A screenshot of a computer

AI-generated content may be incorrect.**

**STEP 2 –** Patient\_ChronicDisease

****

**STEP 3 -** Patient\_Allergy

****

**STEP 4 -** Patient\_AdditionalService

**A computer screen shot of a computer program

AI-generated content may be incorrect.**

**STEP 5 -** BridgeEncounterAdditionalService

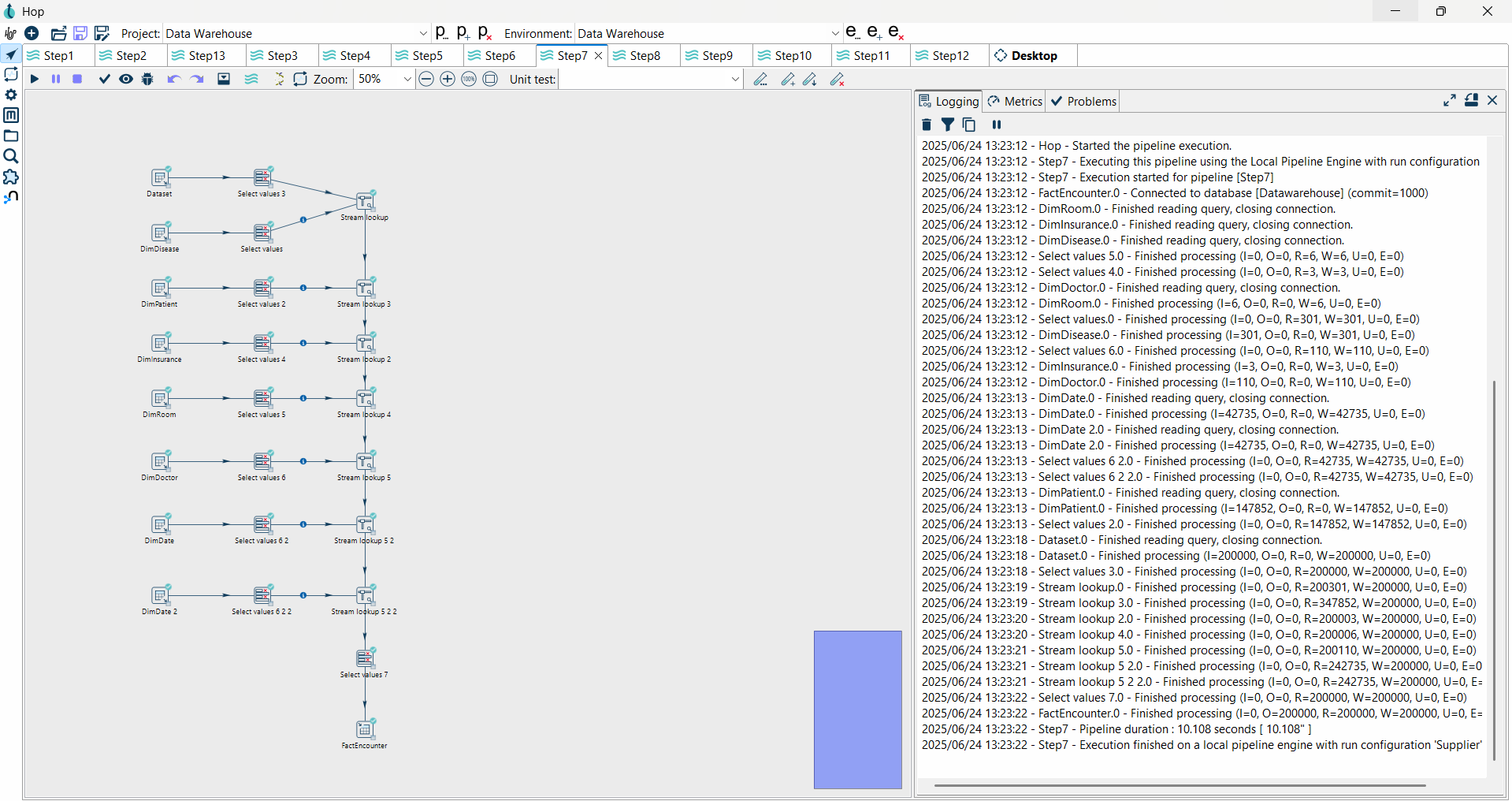
**A computer screen shot of a computer

AI-generated content may be incorrect.**

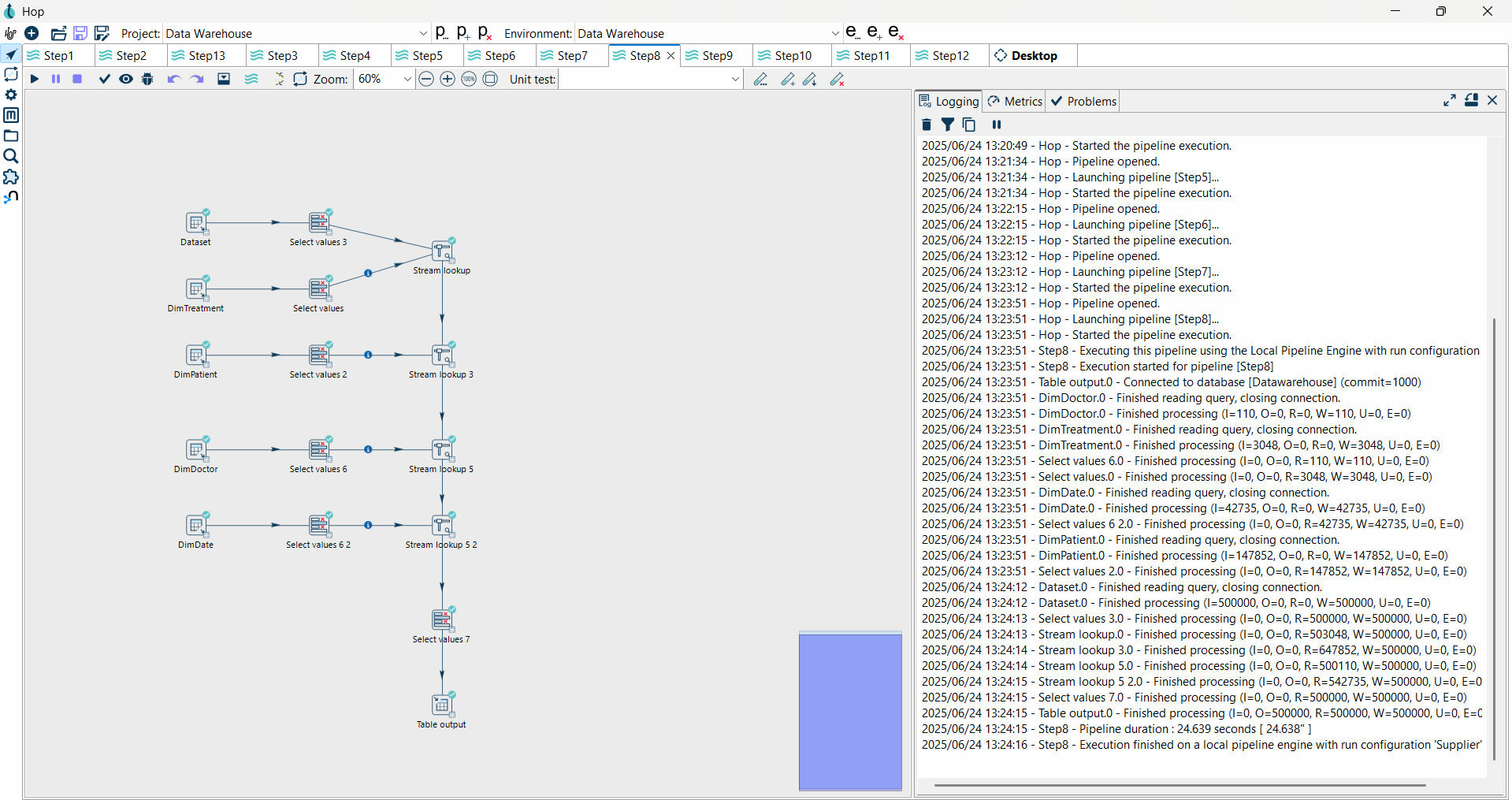
**STEP 6 -** BridgeEncounterDoctor

**A screenshot of a computer

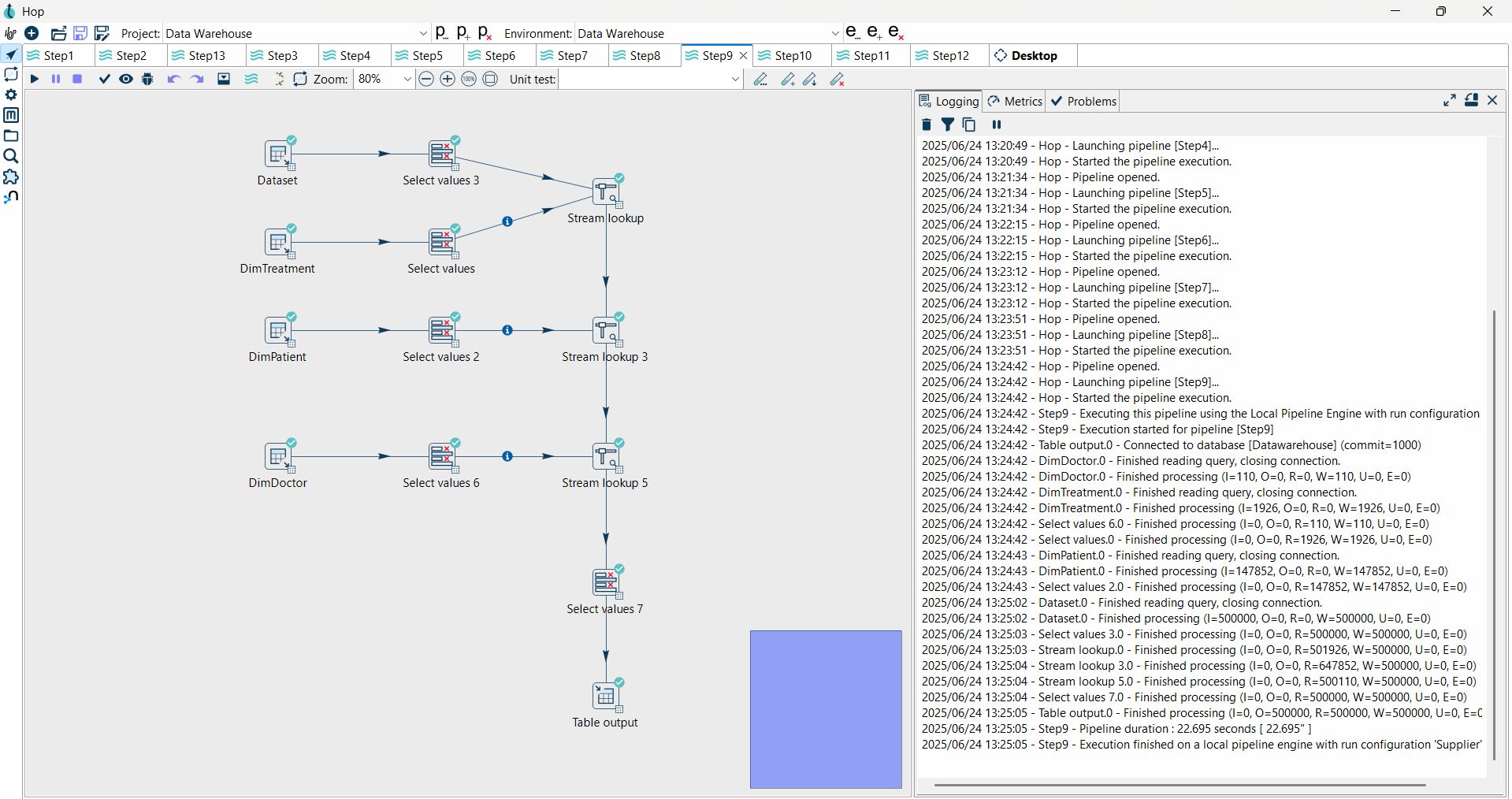
AI-generated content may be incorrect.**

**STEP 7 –** FactEncounter ****

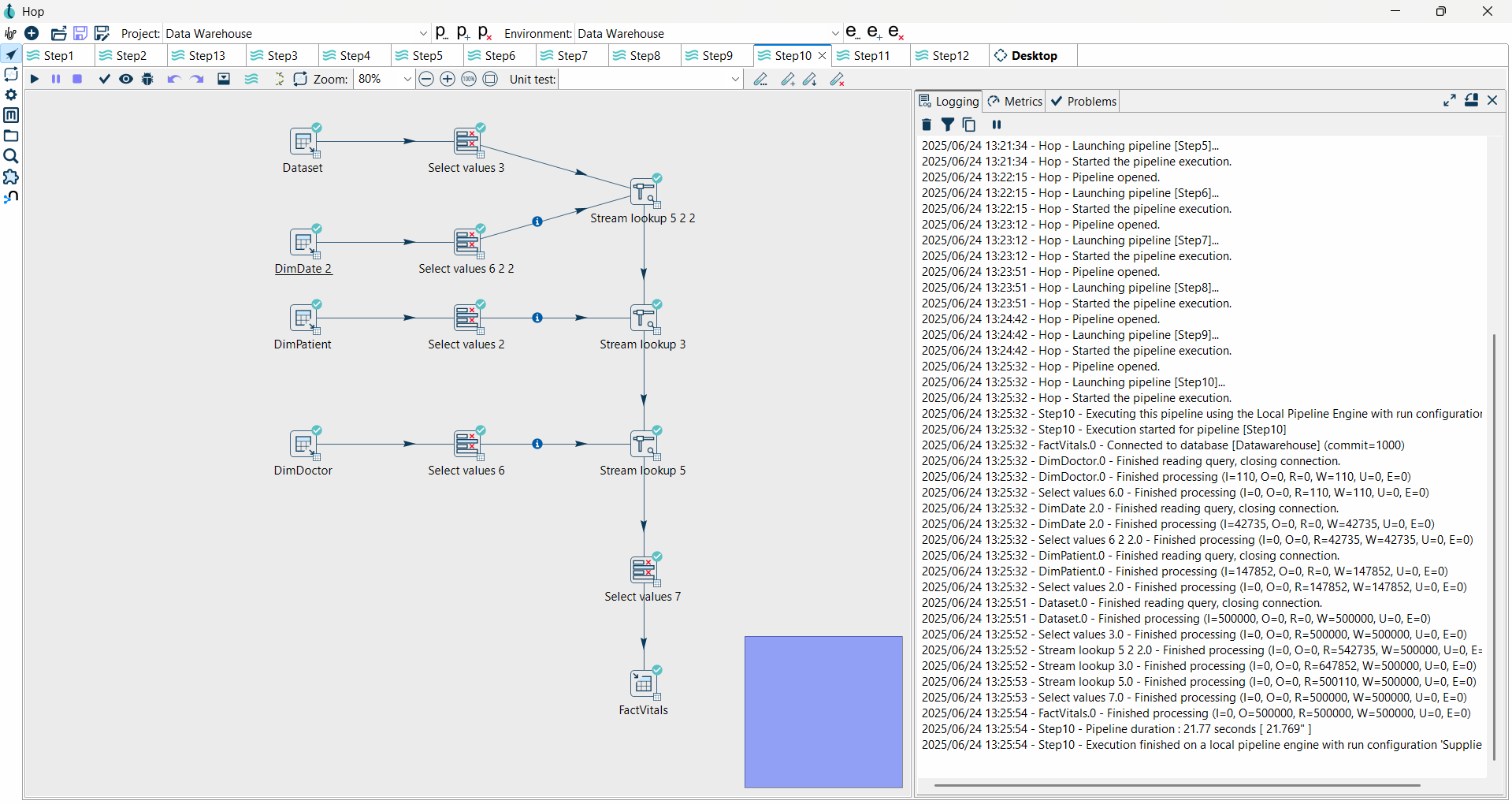
**STEP 8 -** FactTreatment

****

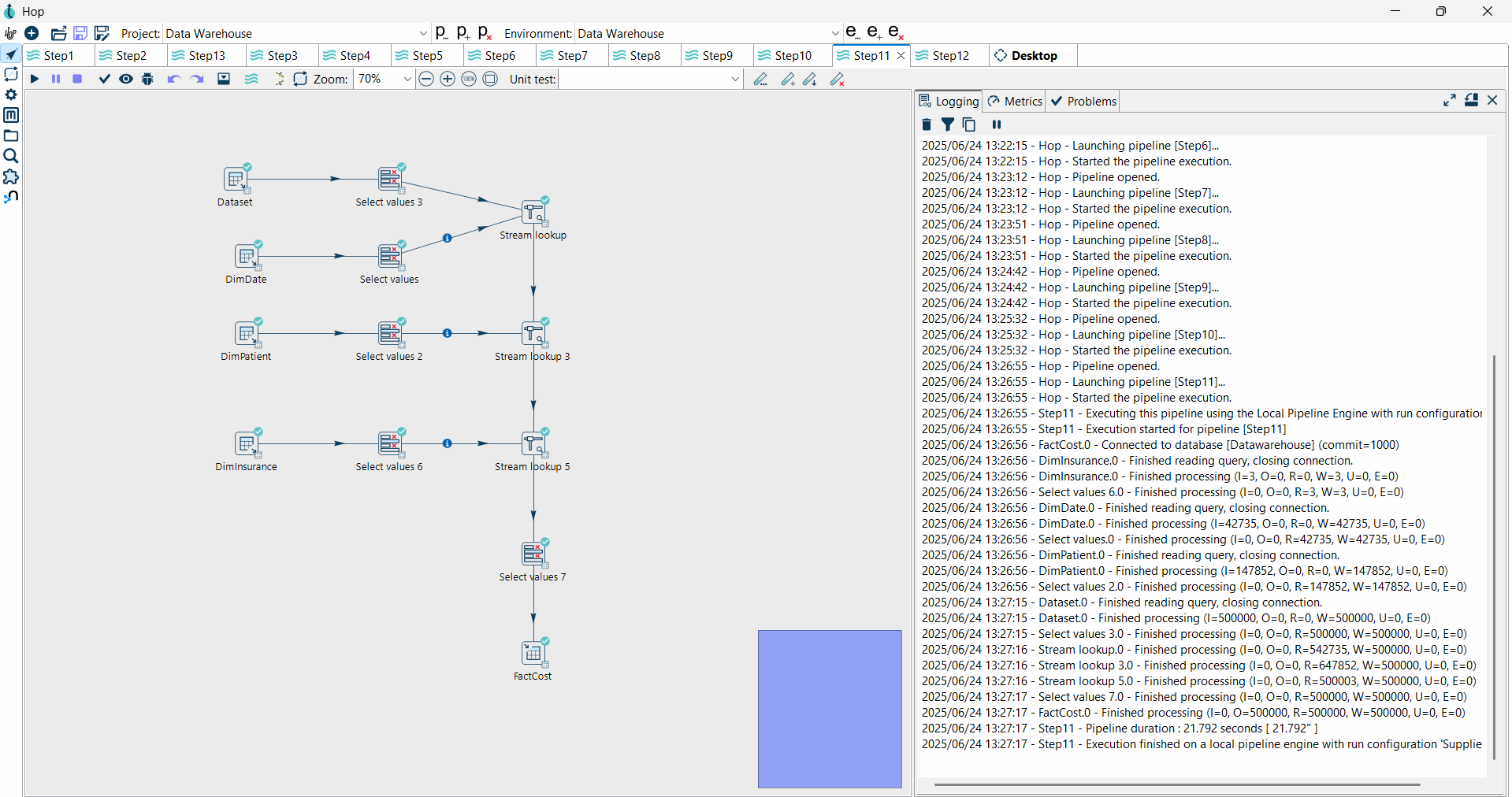
**STEP 9 -** FactLabTests

****

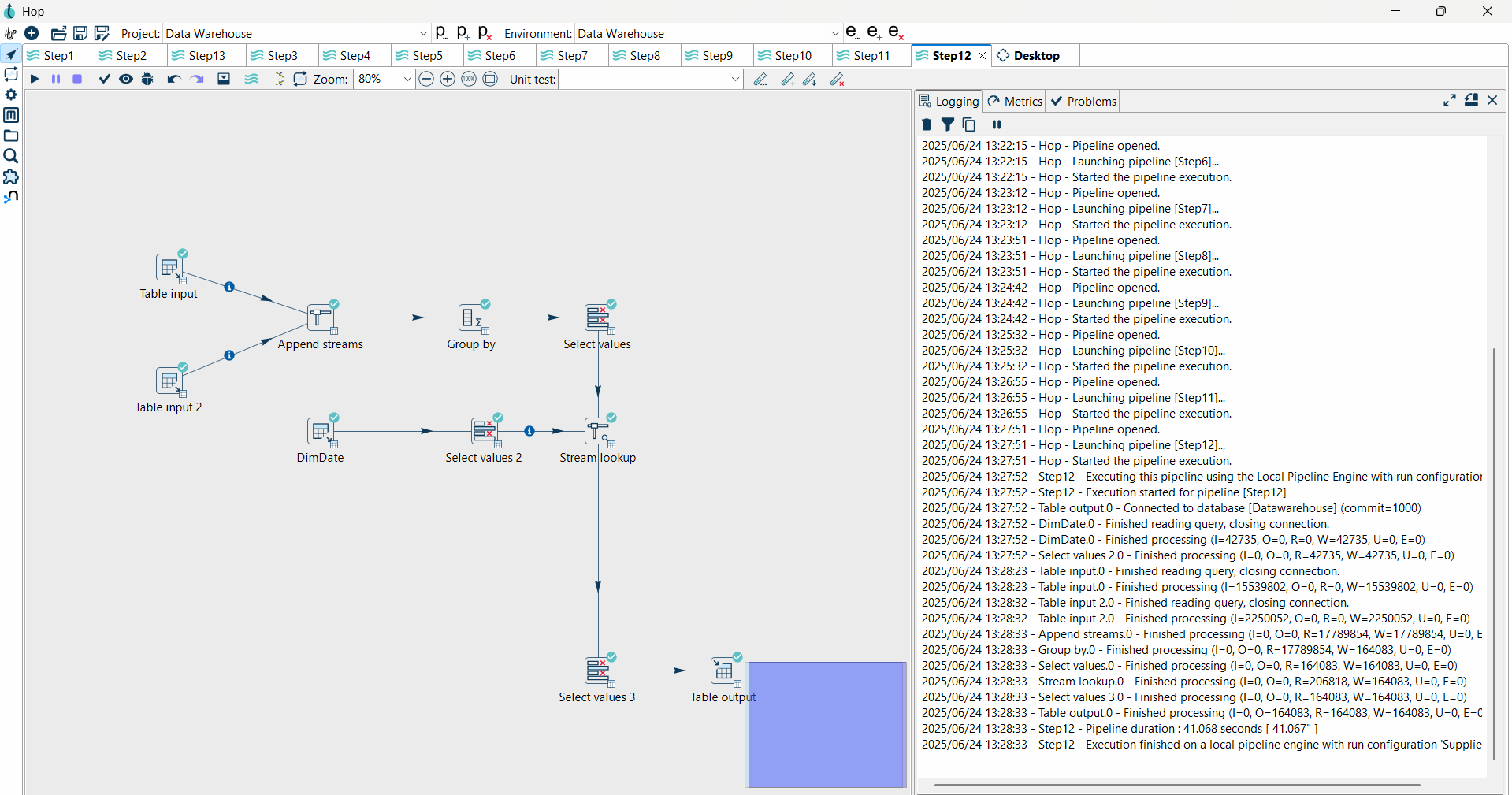
**STEP 10 -** FactVitals

****

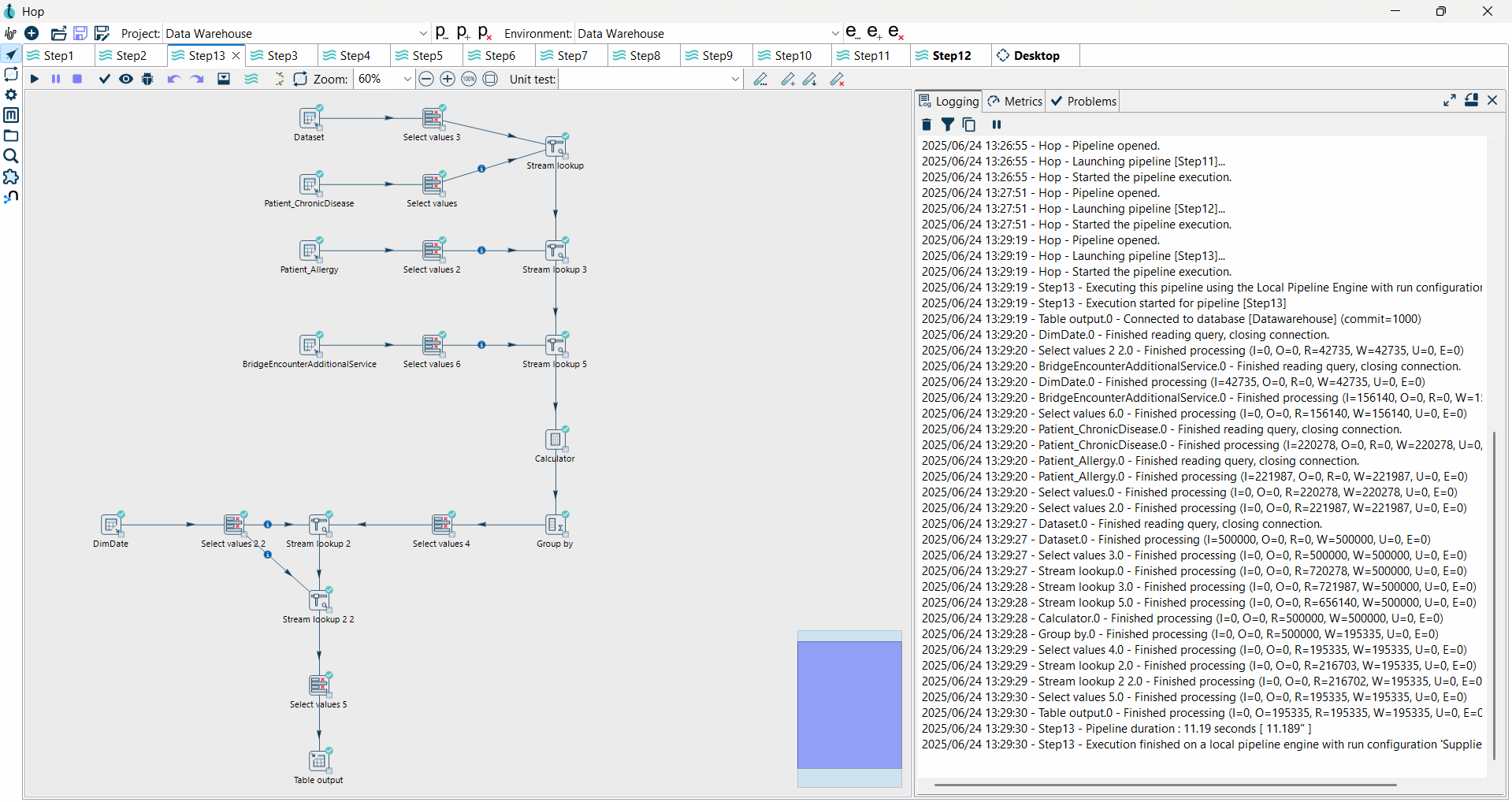
**STEP 11-** FactCost

****

**STEP 12 -** FactProcedures

****

**STEP 13 -** FactPatientHealthService

****

## 4.3 Source and Target Setup

Apache Hop connected to flat files and SQL Server Mana. Transformation pipeline included joins, nulls handling, formatting, data cleaning and loading into DW normalized tables. Screenshots of these runs are shown in the Appendix.

# Chapter 5: Data Visualization Using Power BI

## 5.1 Data Connection

Power BI wasn’t importing the data from the SQL Server database, but was connecting

directly to it. Tables were uploaded preserving relationships as in Galaxy Schema.

## 5.2 Visualizations Created

* Slicers are used to filter the dashboard report
* Line and stacked column chart for value comparison and identify trends
* Cards and Gauge used to display values
* Area chart and line chary used to identify trends over time or any other factor
* 100% stacked area chart and 100% stacked bar chart used to show percentages over time
* Clustered bar chart for comparison of factors
* Pie chart and Donut chart for showing percentages as a whole
* Stacked bar chart used to see the relationship between 02 factors where one factor displays in a bar which is split into parts with its subcategories
* The waterfall chart shows an increase and decreases of values.
* Key influences for analyzing the factors which influence any outcomes.
* Decomposition tree used as a step by step breakdown to understand the result

## 5.3 Dashboard Layout

The dashboard included multiple pages:

* Patient Insights
* Patient Condition
* Clinical encounters and diagnostics
* Cost, insurance & Service utilization

1. Patient Insights

A screenshot of a data analysis dashboard

AI-generated content may be incorrect.

1. Patient Condition

A screenshot of a computer dashboard

AI-generated content may be incorrect.

1. Clinical encounters and diagnostics

A screenshot of a computer

AI-generated content may be incorrect.

1. Cost, insurance & Service utilization

A screenshot of a computer screen

AI-generated content may be incorrect.

## 5.4 Insights Discovered

Expensive Surgery was Associated with Long Length of Stay.

• Elderly patients had more chronic diseases.

• Some doctors simply saw a higher proportion of sickest patients.

• Testing on labs affected the treatment results.

• Prices varied by season, with the highest prices during high admission periods.

# Chapter 6: Findings & Recommendations

## 6.1 Key Business Insights

* Admissions: Seasonal with peaks in Q1 and Q4.
* Vitals: Low O2 was associated with higher EDRs
* Effect on Treatment: Less follow-up if intervention was early
* Economical: Education and rehab expenditure were quite high although often uncounted.
* Workload: Some doctors, particularly those in chronic care, were overburdened

## Recommendations

* Recognition of complications.
* Invest in chronic disease units.
* Enhance monitoring of costs associated with education and postsurgery.
* Leveraging seasonal patterns to optimize both staffing and inventory.

# Chapter 7: Challenges & Solutions

|  |  |
| --- | --- |
| **Challenge** | **Solution** |
| Handling multiple data formats | Used Apache Hop’s multi-format connectors |
| Missing or inconsistent data | Nullable handling and validation logic are in use |
| Complex joins in bridge tables | Composite keys and intermediary tables were used |
| ETL Errors - by Data Types | Standardized data before transformation |
| Dashboard refresh lags | Also improve DAX queries and design of the tables. |

**Lessons Learned -**

* Data cleanup is what makes ETL efficient and normalizing data is key to this.
* Minimize transformation overhead by pre-cleaning data sources.
* Galaxy Schema allows more analytical richness star schema.

# Chapter 8: Conclusion

This work has successfully established a robust health care data warehouse with the MedSynora DW data set. By applying a Galaxy Schema, federating a broad range of source systems and modern approaches like Apache Hop or PowerBI, the project provided a scalable solution to support clinical, operational and financial decisions in a hospital. The lessons learned/discovery is that data is king in healthcare management as can be seen from the Power BI dashboard. The modular ETL process is also designed so that next editions for data will be simple to add in, and this keeps the warehouse up to date.

# References

* Dataset - Kaggle: <https://www.kaggle.com/datasets/mebrar21/medsynora-dw>
* Microsoft SQL Server Studio
* Apache Hop Official Documentation
* Power BI Documentation – Microsoft