**VIETNAM-KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**COMPUTER SCIENCE DEPARTMENT**



**DATA WAREHOUSE**

**Topic:**

**BUILDING AND ANALYZING E-COMMERCE SYSTEM DATA WAREHOUSE**

Student group implementation: **21IT345**– Huỳnh Thị Hoa

**21IT506** – Nguyễn Văn Phó

**21IT335**– Võ Khắc Đoài

**21IT362** – Lê Quang Nghĩa

Instructor: **Ths. Trần Thanh Liêm**

***Da Nang, November 2024***

**VIETNAM-KOREA UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**COMPUTER SCIENCE DEPARTMENT**



**DATA WAREHOUSE**

**Topic:**

**BUILDING AND ANALYZING E-COMMERCE SYSTEM DATA WAREHOUSE**

Student group implementation: **21IT345**– Huỳnh Thị Hoa

**21IT506** – Nguyễn Văn Phó

**21IT335** – Võ Khắc Đoài

**21IT362** – Lê Quang Nghĩa

Instructor: **Ths. Trần Thanh Liêm**

***Da Nang, November 2024***

**FOREWORD**

In the current digital age, information technology and electronic applications have become an important foundation for governments, organizations and businesses. In particular, data warehouse systems play an indispensable role in supporting the process of storing and analyzing information, contributing to optimizing business operations and making strategic decisions. The strong development of e-commerce has promoted the need to build modern data warehouses to store large volumes of data from transaction activities, customer management and related support services.

The topic "Building and analyzing e-commerce system data warehouse" will provide an overview of how to deploy a data warehouse specifically for the e-commerce sector. By using advanced tools and methods, this data warehouse is not only a place to store information but also an analytical tool to help businesses better understand customer behavior, optimize processes and improve business efficiency. Through this project, the research team hopes to provide a useful solution to support e-commerce businesses in accessing information in a scientific and organized manner, thereby improving their competitiveness in today's digital market.

**ACKNOWLEDGEMENTS**

To create this product, our group received a lot of support and help from our instructor - MD. Tran Thanh Liem. With deep and sincere feelings, please allow me to express my deep gratitude to the teacher and all my friends in the group who have helped and researched together during the process of implementing the topic. With limited time and experience, this report cannot avoid shortcomings. We look forward to receiving guidance and comments from teachers, from which we can supplement and improve our awareness, to better serve future projects. Our group would like to sincerely thank!

**LIST OF ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| **ID** | **Phrase** | **Abbreviation** |
| **1** | SSMS | SQL Server Management Studio |
| **2** | SSIS | SQL Server Integration Services |
| **3** | SSRS | SQL Server Reporting Services |
| **4** | API | Application Programming Interface |
| **5** | GUI | Graphical User Interface |
| **6** | IDE | Integrated Development Environment |
| **7** | UI | User Interface |
| **8** | SQL | Structured Query Language |
| **9** | ETL | Extract, Transform, Load |

CATALOG

[CHAPTER 1: DATASET AND PROJECT OVERVIEW 1](#_Toc182237916)

[1.1 Reason for project formation 1](#_Toc182237917)

[1.1.1 Problems noticed 1](#_Toc182237918)

[1.1.2 Solution 1](#_Toc182237919)

[1.1.3 Project objectives and significance 2](#_Toc182237920)

[1.2. Dataset Overview 3](#_Toc182237921)

[1.2.1 Data sources used 3](#_Toc182237922)

[1.2.2 Describe the data in detail. 4](#_Toc182237923)

[1.3 Tool introduction 6](#_Toc182237924)

[1.3.1 Visual Studio 6](#_Toc182237925)

[1.3.1 SSMS 8](#_Toc182237926)

[1.3.3 SSIS 9](#_Toc182237927)

[1.3.4 SSRS 11](#_Toc182237928)

[CHAPTER 2. ANALYSIS AND DESIGN OF DATA WAREHOUSE 14](#_Toc182237929)

[2.1 Logical Modeling 14](#_Toc182237930)

[2.1.1 Dimensions 14](#_Toc182237931)

[2.1.2 Facts 17](#_Toc182237932)

[2.1.3 Constellation Schema 19](#_Toc182237933)

[2.2 Physical Modeling 19](#_Toc182237934)

[2.2.1 Dimensions 19](#_Toc182237935)

[2.2.2 Facts 21](#_Toc182237936)

[2.2.3 Physical Model 25](#_Toc182237937)

[CHAPTER 3. ETL PROCESS 26](#_Toc182237938)

[3.1 Prepare data 26](#_Toc182237939)

[3.2 Create new project ssis 27](#_Toc182237940)

[3.3 ETL 30](#_Toc182237941)

[3.3.1 New Source assistant from Sql Server 30](#_Toc182237942)

[3.3.2 New Destination Assistant from Sql Server 31](#_Toc182237943)

[3.3.3 Create data flow for Dimension table 31](#_Toc182237944)

[3.3.4 Create data flow for Fact table 33](#_Toc182237945)

[3.4 SSAS (Microsoft SQL Server Analysis Services) 35](#_Toc182237946)

[3.4.1 Fact\_Product 35](#_Toc182237947)

[3.4.2 Fact\_Shipping 35](#_Toc182237948)

[3.4.3 Fact\_Sale 36](#_Toc182237949)

[3.5 SSRS (SQL Server Reporting Services) 37](#_Toc182237950)

[3.5.1 Product Sales and Volume by City and Country 37](#_Toc182237951)

[3.5.2 Revenue and product quantity by SKU and product type 38](#_Toc182237952)

[3.5.3 Revenue and Product Quantity by Sales Channel and Order Status 39](#_Toc182237953)

[3.5.4 Revenue by Day and Order Status 40](#_Toc182237954)

[3.5.5 Product distribution by sales channel and country 41](#_Toc182237955)

[3.5.6 Distribution of product quantity by product type and order status 42](#_Toc182237956)

[3.5.7 Analysis of revenue and product quantity by shipping service level 43](#_Toc182237957)

[CONCLUDE 44](#_Toc182237958)

[1. Results not achieved 44](#_Toc182237959)

[2. Development direction 44](#_Toc182237960)

[3. Lessons learned 45](#_Toc182237961)

[REFERENCES 46](#_Toc182237962)

**IMAGE CATALOG**

[Hình 1.3.1.1 Visual Studio 6](#_Toc182234565)

[Hình 2.3.2.1 SQL Server Management Studio 8](#_Toc182234566)

[Hình 3.2.3 Physical Model 25](#_Toc182234567)

[Hình 4.1 Import flat file to Sql Server 26](#_Toc182234568)

[Hình 5.4.1 Fact\_Product 37](#_Toc182234569)

[Hình 3.5.6 Revenue by Day and Order Status 42](#_Toc182234570)

[Hình 3.5.7 Product distribution by sales channel and country 43](#_Toc182234571)

# CHAPTER 1: DATASET AND PROJECT OVERVIEW

## Reason for project formation

### 1.1.1 Problems noticed

* Huge and Complex Data from E-commerce Transactions

E-commerce platforms like Amazon have to manage a huge amount of data from millions of transactions every day. This data is complex, diverse in types (products, customers, revenue, profits, etc.) and often grows continuously.Processing big data without a suitable management system will lead to difficulties in storing, organizing and extracting information, making it difficult for businesses to have an overall and detailed view of their business operations.

* Lack of information to support strategic decision making

To optimize business strategies, businesses need accurate data and in-depth analysis of consumer behavior, shopping trends, and factors affecting sales. However, manually analyzing and extracting this information from raw data is very difficult and prone to errors.Lack of reporting and analytics can make it difficult for businesses to forecast demand, manage inventory, optimize pricing, and develop effective marketing strategies.

* Competitive Needs and Improved Customer Experience

In a highly competitive environment, e-commerce businesses must leverage data to optimize processes, develop products, and improve customer experiences.Without an effective data warehouse system, companies may miss out on important opportunities to create personalized shopping experiences, quickly respond to customer needs, and maintain a competitive advantage.

### 1.1.2 Solution

Building a dedicated data warehouse for e-commerce

* Design and build a dedicated data warehouse for e-commerce systems, helping to collect, store and organize data from many different sources (sales transactions, customers, products, etc.).
* The data warehouse will provide a reliable source of historical data, which can be quickly queried, supporting analysis and reporting according to business requirements.

Integrate Business Intelligence (BI) tools to analyze data

* Combine the data warehouse with BI tools such as Power BI, Tableau or SQL Server Reporting Services (SSRS) to build dashboards, automated reports, and in-depth analysis.
* BI tools will help businesses easily access and analyze data visually, thereby supporting data-driven decision making.

Apply data analytics to optimize business operations

* Using data warehouses to exploit factors such as customer shopping behavior, product consumption trends, and cost efficiency, helps optimize marketing strategies, predict inventory needs, and optimize supply chains.
* Data analytics will help businesses make decisions based on forecasts and trends, instead of just reacting to events that have occurred, thereby increasing proactive management.

Enhance the ability to personalize customer experiences

* Data warehouses can store and analyze the behavior of each customer, thereby helping businesses better understand their individual needs and preferences.
* Based on this information, businesses can personalize marketing strategies and suggest suitable products, improve user experience and increase customer satisfaction. liệu.

### 1.1.3 Project objectives and significance

* Project Objectives

Build a complete data warehouse for e-commerce systems : The main objective is to build a data warehouse dedicated to e-commerce systems, which can effectively store and manage data from transactions, products and customer information. This data warehouse will be a solid foundation for analysis and reporting activities in the enterprise.

Support for creating reports and in-depth analysis :The project aims to provide a system that helps businesses easily perform reports and in-depth analysis to grasp operational efficiency, business trends and optimize business decisions.The data warehouse will support complex queries, provide real-time information or historical data quickly, meeting the diverse analysis needs of businesses.

Create a foundation for Business Intelligence (BI) applications :The data warehouse will support the integration of BI tools to build dashboards and automatic reports. Thereby, businesses can easily access, visualize and analyze data without having to perform complex operations.

* Significance of the project

Improve decision-making efficiency : The data warehouse provides a comprehensive information platform, allowing managers and professional teams in the business to make accurate and timely decisions based on real data instead of predictions or emotions. .

Optimize business operations and increase competitiveness : With effectively organized data, businesses can optimize processes such as inventory management, adjust pricing strategies and improve customer service. This contributes to reducing costs and improving business efficiency.

Improve customer experience : The data warehouse allows for deeper analysis of customer behavior and preferences, helping to personalize the experience and increase customer satisfaction. From there, businesses can develop products and services that suit customer needs, creating a sustainable competitive advantage.

Building a Foundation for Future Advanced Analytics Projects : The e-commerce data warehouse will be the foundation for future advanced analytics projects such as demand forecasting, fraud detection, and supply chain optimization, helping businesses continue to grow. grow in the digital data age.

## 1.2. Dataset Overview

### 1.2.1 Data sources used

*Dataset name:* "Amazon Sale Report"

This is a dataset named "Amazon Sale Report", simulating sales data from the Amazon e-commerce platform.

*Origin:*

This dataset is built from aggregated information of sales transactions, including products, sales volume, revenue and profit.

The dataset is designed to provide an overview and details of Amazon's online business, facilitating in-depth analysis of sales trends, profits and customer behavior.

*Time and scope:*

The dataset includes information from multiple transactions and product categories, helping to reflect the overall factors affecting Amazon's sales, revenue and profit over a certain period of time.

*Purpose of use:*

This dataset serves the purpose of building and analyzing data warehouses, supporting multi-dimensional queries and reports. It provides a rich source of information, helping students and businesses better understand the factors that determine the success of an e-commerce platform. Robin.

### 1.2.2 Describe the data in detail.

The "Amazon Sale Report" dataset includes many important information fields, reflecting details about transactions and sales performance. Below are the main data fields in the dataset and the meaning of each field:

* Product id :

A unique identifier for each product, helping to distinguish different products in the dataset.

The product code allows for product tracking and product-specific analysis.

* Product name :

The full name of the product, helping users easily identify and classify products in the report.

The product name provides additional descriptive information for each transaction, helping to perform analysis by product type.

* Product category :

The category or product group that each product belongs to, for example: electronics, fashion, home appliances, etc.

This field helps to classify products and support category-based analysis, helping businesses better understand the sales performance of each product type.

* Sales Quantity:

The number of products sold in each transaction.

This field helps calculate the total sales quantity by product, category, or time, supporting demand analysis and inventory planning adjustments.

* Unit Price

The selling price of each product per unit.

Unit price allows calculating revenue and profit based on the quantity sold, thereby helping to analyze price and profit factors.

* Total Revenue

The total revenue from each product transaction, calculated by multiplying the quantity sold by the unit price.

This field helps businesses evaluate the revenue efficiency of each product or category, thereby making decisions about pricing strategies.

* Cost

The total cost required to provide that product, including manufacturing, shipping, and other costs.

The cost field provides information about actual costs, supporting profit calculations and cost optimization in business strategies.

* Profit

The profit earned from each product transaction, calculated by subtracting costs from total revenue.

The profit field is an important factor that helps businesses identify high-profit products or categories and optimize sales strategies.

* Overall meaning of the data

The data set provides a comprehensive view of the key factors affecting e-commerce business operations, thereby supporting analysis such as revenue forecasting, profit assessment, and better understanding of customer shopping trends.

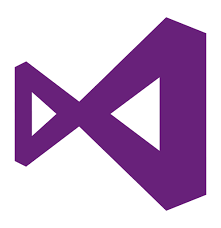
These data fields will be the basis for building a data warehouse and creating analytical reports, helping businesses make smart and effective business decisions. Linux..

## 1.3 Tool introduction

### 1.3.1 Visual Studio

Visual Studio is a powerful Integrated Development Environment (IDE) created by Microsoft for building various types of applications. It provides a comprehensive set of tools and features that support the entire development lifecycle, from coding and debugging to testing and deployment.

#### 1.3.1.1 What is Visual Studio?



Hình 1.3.1.1 Visual Studio

Visual Studio is an IDE that provides a single platform for developers to create various types of applications, including:

* Desktop applications: Develop applications with a graphical user interface (GUI) for Windows.
* Web applications: Build web applications and services using technologies like ASP.NET, Node.js, and ASP.NET Core.
* Mobile applications. Create native mobile applications for Android, iOS, and Windows using tools like Xamarin and React Native.
* Cloud applications: Build and deploy applications and services to Azure, AWS, and other cloud platforms
* Games: Develop games using Unity, Unreal Engine, and other game development tools
* Data science and machine learning Visual Studio provides tools and libraries for data analysis, machine learning, and artificial intelligence.

#### Key features of Visual Studio

Here are some of the key features that make Visual Studio a popular choice for developers:

• Code editor: Visual Studio provides a powerful code editor with syntax highlighting, code completion, and code refactoring features.

• Debugger. The debugger allows you to step through your code line by line, set breakpoints, and examine variables to identify and fix errors.

• Testing tools: Visual Studio integrates with various testing frameworks like NUnit, MSTest, and xUnit to help you write and run unit tests, integration tests, and UI tests.

• Version control: Visual Studio supports popular version control systems like Git and Mercurial to help you manage your code history and collaborate with other developers.

Extensibility: Visual Studio is highly extensible, allowing you to add new features and functionality through extensions.

Support for multiple languages: Visual Studio supports a wide range of programming languages, including C#, C++, VB.NET, F#, Python, JavaScript, and more.

#### Benefits of using Visual Studio

There are many benefits to using Visual Studio for your development

projects, including:

• Increased productivity: The various features and tools in Visual Studio can help you write code faster and more efficiently.

• Improved quality: Visual Studio helps you identify and fix errors early in the development process, resulting in higher quality applications.

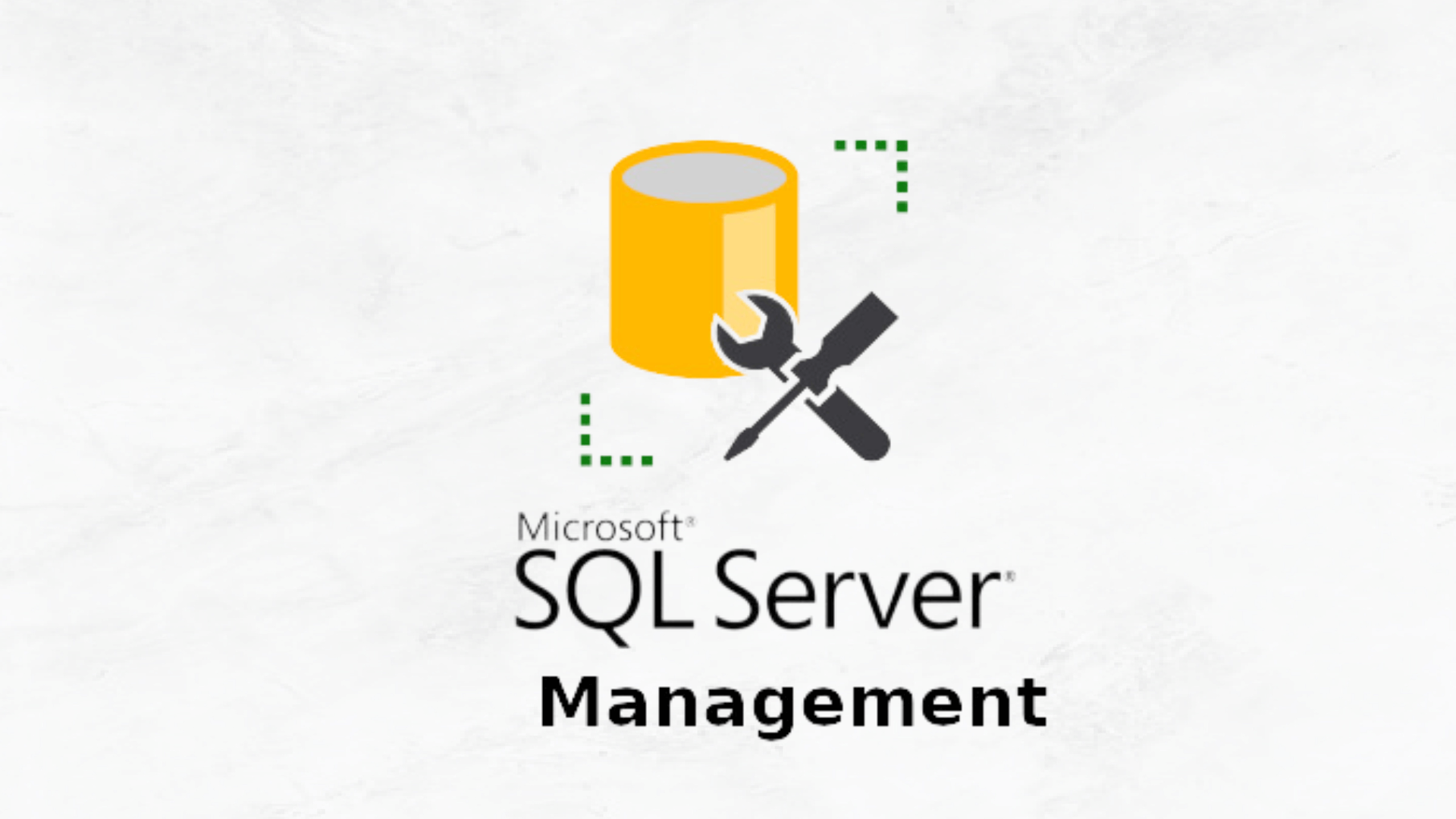
• Collaboration: Visual Studio provides features that make it easier to collaborate with other developers on your projects.

• Scalability: Visual Studio can be used to develop small, simple applications as well as large, complex applications.

• Customization: Visual Studio can be customized to meet your specific workflow and preferences.

### SSMS

#### What is the SSMS?



Hình 1.3.2.1 SQL Server Management Studio

SQL Server Management Studio (SSMS) is a powerful and versatile graphical user interface (GUI) tool used to manage and administer Microsoft SQL Server databases. It provides a comprehensive set of features for working with databases, including:

• Connecting to and managing SQL Server instances

• Writing and executing Transact-SQL (T-SQL) queries

• Creating and managing database objects (tables, views, stored procedures, etc.)

• Designing and modifying database schemas

• Administering user permissions and security

• Monitoring and optimizing database performance

SSMS is a crucial tool for anyone who works with SQL Server databases, from developers and database administrators to business analysts and data scientist

#### **Features of SSMS**

Here are some of the key features of SSMS:

• Object Explorer: A hierarchical view of all the objects in a SQL Server instance.

Query Editor: A powerful editor for writing and executing T-SQL queries.

• Table Designer. A graphical tool for designing and modifying database tables.

• Data Viewer: A grid-based interface for viewing and editing data in tables.

• Index Tuning Wizard: A tool for assisting with the creation and optimization of indexes.

• Backup and Restore Wizard: A tool for backing up and restoring databases.

• Security Management: Tools for managing user permissions and roles.

• Performance Dashboard: A graphical interface for monitoring database performance

#### **Benefits of using SSMS**

There are many benefits to using SSMS for your database administration tasks, including:

• Increased productivity: The graphical interface and intuitive features of SSMS make it easy to manage and administer databases.

• Improved efficiency: SSMS automates many tasks, such as creating scripts and managing permissions, saving you time and effort.

• Reduced errors: SSMS helps to prevent errors by providing features such as syntax highlighting and IntelliSense.

• Enhanced security: SSMS provides tools for managing user permissions and roles, helping to keep your databases secure.

• Improved performance: SSMS provides tools for monitoring and optimizing database performance, ensuring that your databases run smoothly

### 1.3.3 SSIS

#### **What is SSIS ?**



Hình 1.3.3.1 SQL Server Integration Services

SQL Server Integration Services (SSIS), a powerful tool within the Microsoft SQL Server suite, transcends the limitations of a mere data extraction, transformation, and loading (ETL) tool. It's a complete platform, an ecosystem, empowering developers and data professionals to build robust, scalable, and maintainable data integration solutions.

#### **Features of SSIS**

SSIS offers a vast array of features and functionalities, enabling you to:

• Extract data from diverse sources: Connect to various data sources, including relational databases, flat files, APIs, and cloud services, for comprehensive data acquisition.

• Transform and cleanse data: Apply transformations to restructure, clean, and enrich your data, ensuring its accuracy and consistency for further analysis

• Load data into target destinations: Load your transformed data into a multitude of target destinations, including relational databases, data warehouses, cloud storage, and data lakes.

• Orchestrate complex data flows: Build intricate data pipelines with control flow tasks and data flow components, automating data workflows and ensuring data integrity.

• Monitor and manage your data pipelines: Leverage the SSIS Management Studio for comprehensive monitoring, logging, and troubleshooting, guaranteeing smooth data integration processes.

• Extend functionality with custom components: Develop custom components to address unique integration needs and extend the capabilities of SSIS beyond its core functionality.

• Leverage the power of the cloud: Integrate with Azure services for scalable data processing and leverage cloud-based data storage solutions for cost-effective and efficient data management

#### **Benefits of Choosing SSIS**

Embrace a world of advantages when you choose SSIS:

• Increased Efficiency: Automate data integration tasks, eliminating manual work and boosting productivity.

• Improved Data Quality: Cleanse and transform data, ensuring its accuracy and consistency for reliable analysis.

• Scalability and Flexibility. Design data pipelines to handle large volumes of data and adapt to changing needs.

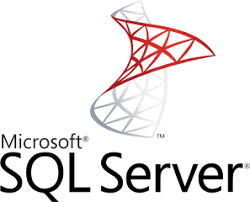
• Reduced Costs: Optimize data processing and leverage cloud services for cost-effective solutions.

• Enhanced Security: Integrate with existing security infrastructure and ensure data privacy.

• Ease of Use: Utilize a user-friendly interface and visual tools for intuitive data integration design.

### SSRS

#### **1.3.4.1 What is SSRS?**



Hình 1.3.4.1 SQL Server Reporting Services

SQL Server Reporting Services (SSRS), crafted by Microsoft, transcends the limitations of a mere reporting tool. It's a robust platform, a complete ecosystem, empowering developers and business users to create and manage comprehensive and insightful reports.

#### **1.3.4.2 Features of SSRS**

SSRS offers a vast array of features and functionalities, enabling you to:

• Design and build interactive reports. Create visually appealing and interactive reports with tables, charts, graphs, and other visualizations for effective data presentation

• Connect to diverse data sources: Access data from various sources, including relational databases, flat files, APIs, and cloud services, for consolidated reporting.

• Develop complex reports with rich features: Utilize powerful features like parameters, filters, and drill-down capabilities to allow users to explore data in-depth and uncover hidden insights.

• Deliver reports in various formats: Export reports in multiple formats, including PDF, Word, Excel, and HTML, for convenient distribution and access

• Schedule and automate report delivery: Schedule reports to run automatically at specific times or intervals, ensuring timely and consistent delivery to stakeholders.

• Manage and secure your reports: Utilize the web-based Report Manager for centralized report management, access control, and security configuration.

• Extend functionality with custom extensions: Develop custom extensions to address unique reporting needs and extend the capabilities of SSRS beyond its core functionality.

• Leverage the power of the cloud: Integrate with Azure services for scalable reporting and leverage cloud-based storage solutions for cost- effective and efficient report management

#### **1.3.4.3 Benefits of Choosing SSRS**

SSRS offers a vast array of features and functionalities, enabling you to:

• Design and build interactive reports: Create visually appealing and interactive reports with tables, charts, graphs, and other visualizations for effective data presentation.

Connect to diverse data sources: Access data from various sources, including relational databases, flat files, APIs, and cloud services, for consolidated reporting.

• Develop complex reports with rich features: Utilize powerful features like parameters, filters, and drill-down capabilities to allow users to explore data in-depth and uncover hidden insights

• Deliver reports in various formats: Export reports in multiple formats, including PDF, Word, Excel, and HTML, for convenient distribution and access.

• Schedule and automate report delivery: Schedule reports to run automatically at specific times or intervals, ensuring timely and consistent delivery to stakeholders.

• Manage and secure your reports: Utilize the web-based Report Manager for centralized report management, access control, and security configuration.

• Extend functionality with custom extensions: Develop custom extensions to address unique reporting needs and extend the capabilities of SSRS beyond its core functionality

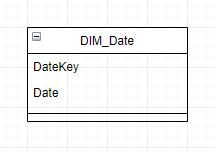
• Leverage the power of the cloud: Integrate with Azure services for scalable reporting and leverage cloud-based storage solutions for cost- effective and efficient report management

CHAPTER 2. ANALYSIS AND DESIGN OF DATA WAREHOUSE

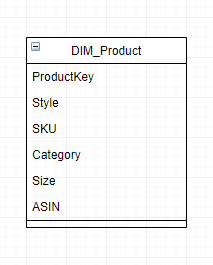
## 2.1 Logical Modeling

### 2.1.1 Dimensions

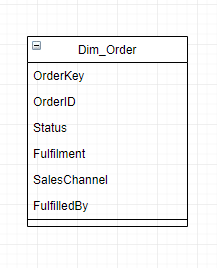
* **Dim\_Date**



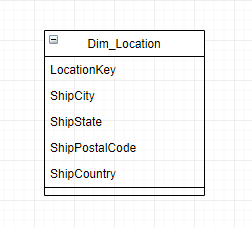
* **Dim\_Product**



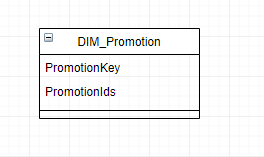
* **Dim\_Order**



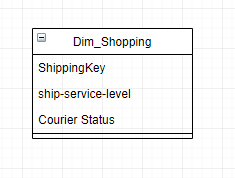
* **Dim\_Location**

****

* **Dim\_Promotion**

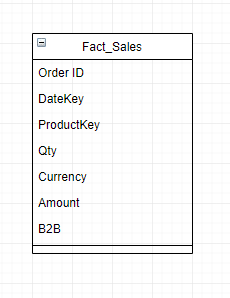
****

* **Dim\_Shopping**

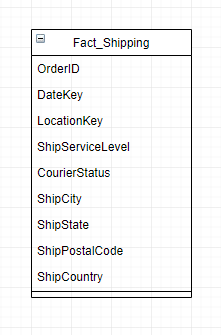
****

### 2.1.2 Facts

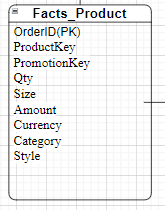
* Fact\_Sales



* Fact\_Shipping

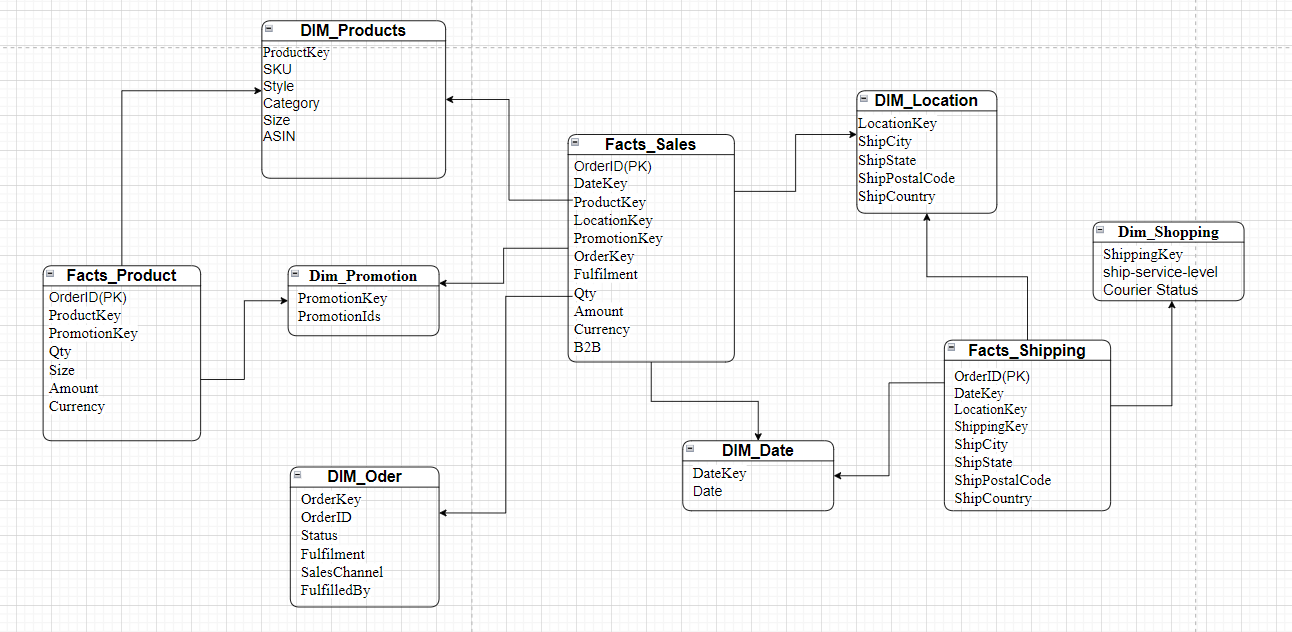


* Fact\_Product



### 

### 2.1.3 Constellation Schema



## 2.2 Physical Modeling

### 2.2.1 Dimensions

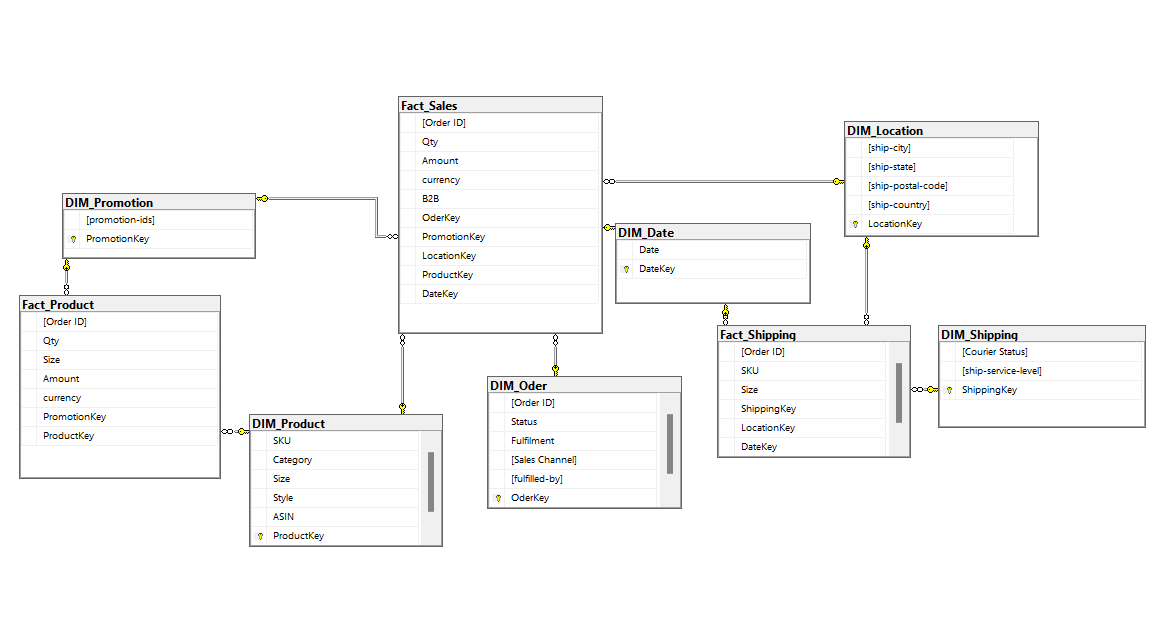
|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Constraint | Comment |
| **DIM\_Order** |  |  | The DIM\_Order table contains information about orders. |
| OrderKey | int | PRIMARY KEY | Unique ID for each order |
| OrderID | varchar(50) |  | Order identifier |
| Status | varchar(50) |  | Order status (e.g., shipped) |
| Fulfilment | varchar(50) |  | Fulfilment status |
| SalesChannel | varchar(50) |  | Sales channel used (e.g., online) |
| FulfilledBy | varchar(50) |  | Fulfilled by (e.g., self, Amazon) |
|  |  |  |  |
| **DIM\_Product** |  |  | The DIM\_Product table stores detailed information about products. |
| ProductKey | int | PRIMARY KEY | Unique ID for each product |
| SKU | varchar(50) |  | Stock Keeping Unit |
| Category | varchar(50) |  | Product category |
| Style | varchar(50) |  | Product style |
| Size | varchar(20) |  | Product size |
| ASIN | varchar(20) |  | Amazon Standard Identification |
|  |  |  |  |
| **DIM\_Date** |  |  | The DIM\_Date table stores information about dates. |
| DateKey | int | PRIMARY KEY | Unique ID for each date |
| Date | date |  | Calendar date |
|  |  |  |  |
| **DIM\_Location** |  |  | The DIM\_Location table stores information about delivery locations. |
| LocationKey | int | PRIMARY KEY | Unique ID for each location |
| ShipCity | varchar(50) |  | Shipping city |
| ShipState | varchar(50) |  | Shipping state |
| ShipPostalCode | varchar(20) |  | Shipping postal code |
| ShipCountry | varchar(50) |  | Shipping country |
|  |  |  |  |
| **DIM\_Promotion** |  |  | The DIM\_Promotion table stores information about promotions. |
| PromotionKey | int | PRIMARY KEY | Unique ID for each promotion |
| PromotionIds | varchar(100) |  | List of associated promotions |
|  |  |  |  |

.

### 2.2.2 Facts

|  |  |  |
| --- | --- | --- |
| Name | Type | Comment |
| **FACT\_Product** |  | FACT table stores information about sales transactions |
| Order ID | INT | Unique identifier for each sales order |
| ProductKey | int | Reference to the DIM\_Product table to identify the products |
| PromotionKey | int | Reference to the DIM\_Promotion table to identify promotions |
| Qty | int | Quantity of the product sold in the order. |
| Size | varchar(20) | Product size |
| Amount | decimal | Total amount of the product |
| Currency | varchar(20) | Currency used in the transaction |
| Category | varchar(50) | Product category |
| Style | varchar(50) | Product style |
|  |  |  |
| **FACT\_Shipping** |  | FACT table stores shipping information of orders |
| OrderID | int | Unique identifier for each order |
| DateKey | int | Reference to the DIM\_Date table to specify the shipping date |
| LocationKey | int | Reference to the DIM\_Location table to identify the shipping location |
| ShipServiceLevel | varchar(50) | Shipping service level (e.g., standard, expedited) |
| CourierStatus | varchar(50) | Status of the courier service (e.g., in transit, delivered) |
| ShipCity | varchar(50) | Shipping destination city |
| ShipState | varchar(50) | Shipping destination state/province |
| ShipPostalCode | varchar(20) | Postal code of the shipping destination |
| ShipCountry | varchar(50) | Country of the shipping destination |
|  |  |  |
| **FACT\_Sales** |  | FACT table stores information about sales transactions |
| OrderID | int | Unique identifier for each sales order |
| DateKey | int | Reference to the DIM\_Date table to specify the sales date |
| ProductKey | int | Reference to the DIM\_Product table to identify the product |
| Currency | varchar(20) | Currency used in the transaction |
| Amount | decimal | Total amount of the sales transaction |
| B2B | boolean | Identifies if the transaction is B2B (business-to-business) |

### 2.2.3 Physical Model

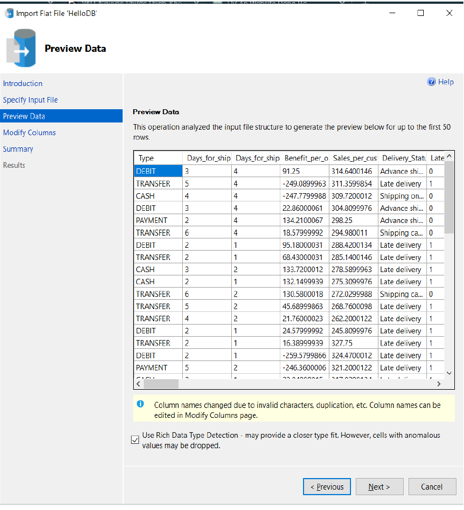


Hình 3.2.3 Physical Model

# CHAPTER 3. ETL PROCESS

## 3.1 Prepare data

* Import flat file to Sql Server

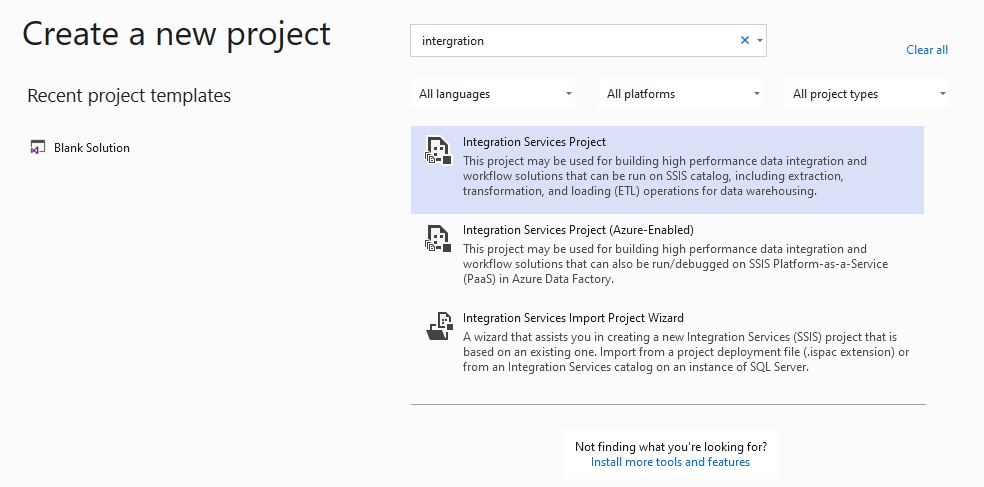


Hình 4.1 Import flat file to Sql Server

## 

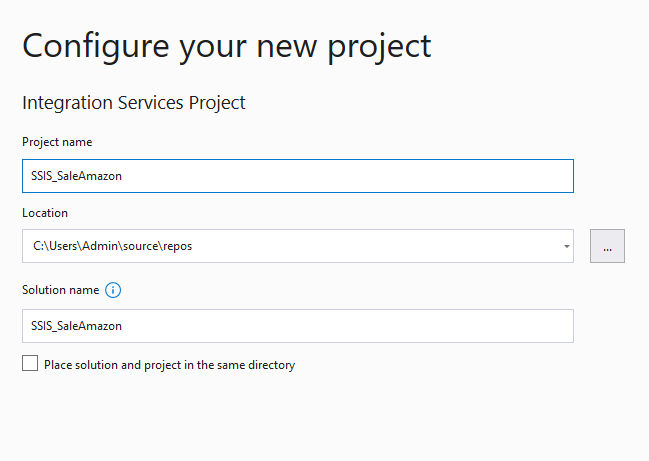
## 3.2 Create new project ssis

Select Integration Services Project



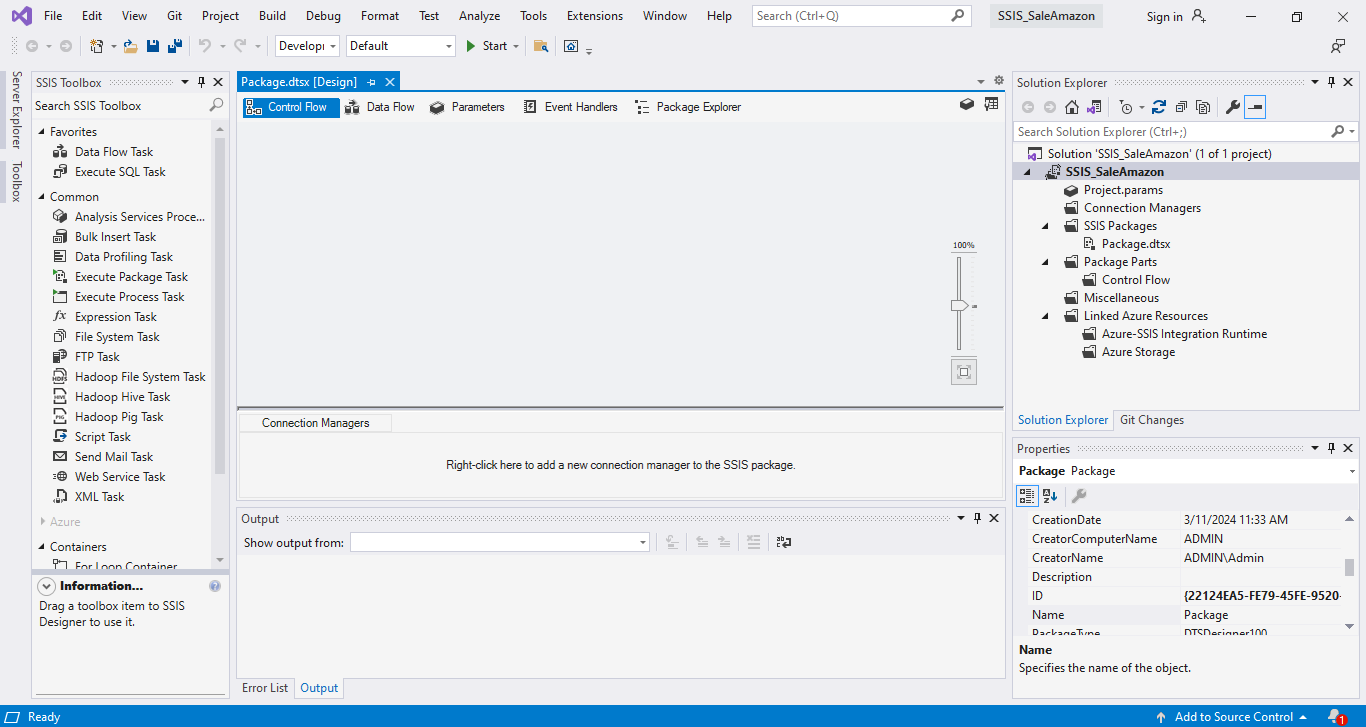
Hình 3.2a Create new project ssis

Configure your new project



Hình 3.2.b Configure your new project

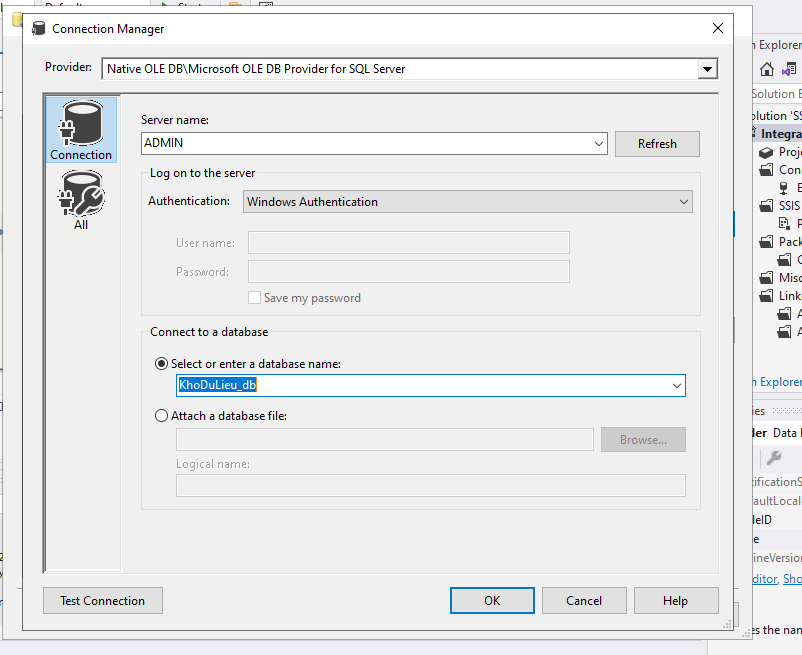
Project after creation



Hình 3.2.c Project after creation

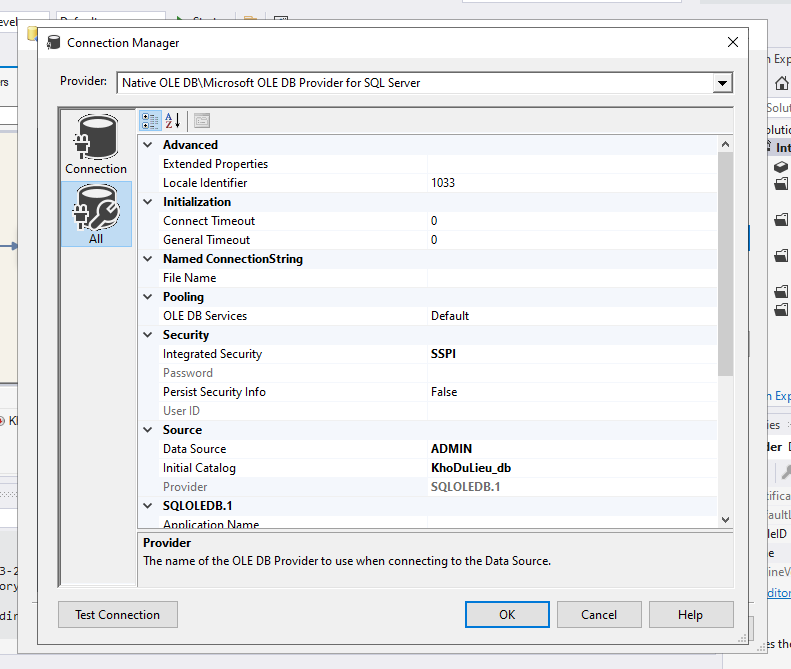
## 3.3 ETL

### 3.3.1 New Source assistant from Sql Server



Hình 3.3.1 New Source assistant from Sql Server

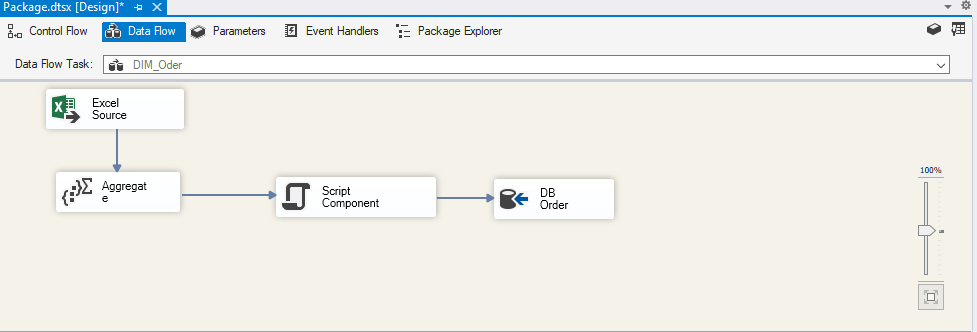
### 3.3.2 New Destination Assistant from Sql Server



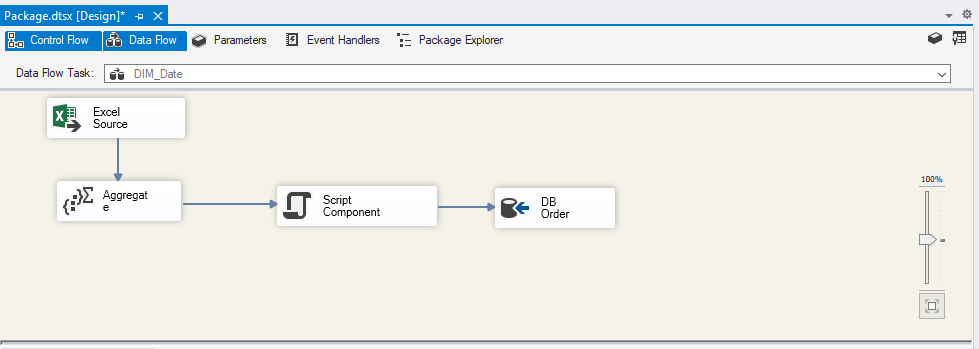
Hình 3.3.2 New Destination Assistant from Sql Server

### 3.3.3 Create data flow for Dimension table

DIM\_Oder

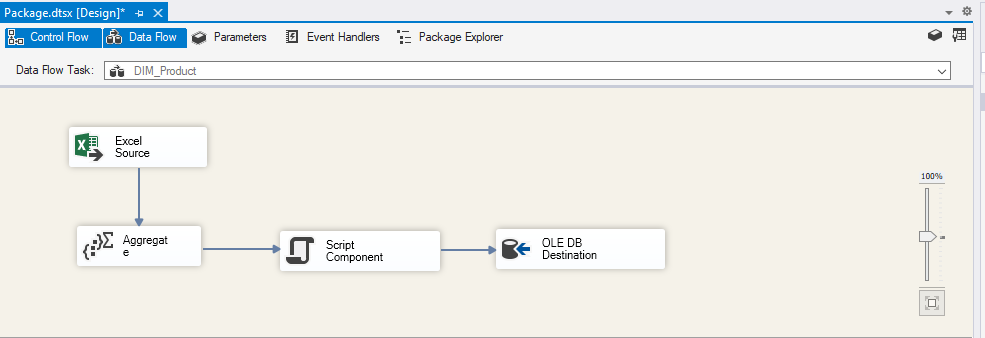


Hình 3.3.3a DIM\_Oder

DIM\_Date

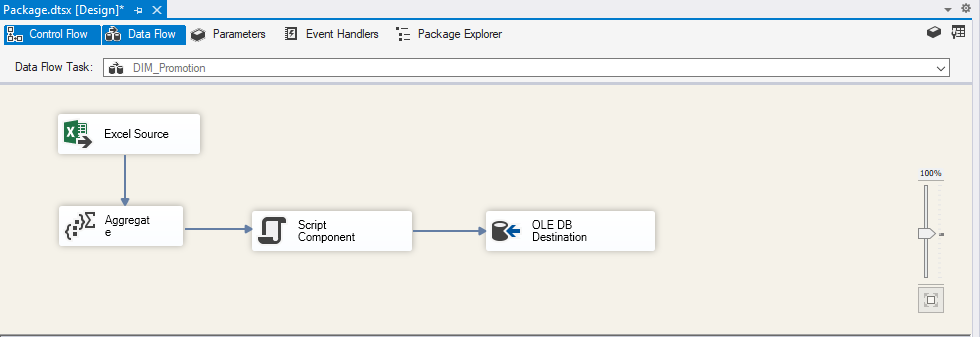
Hình 3.3.3b DIM\_Date

DIM\_Product



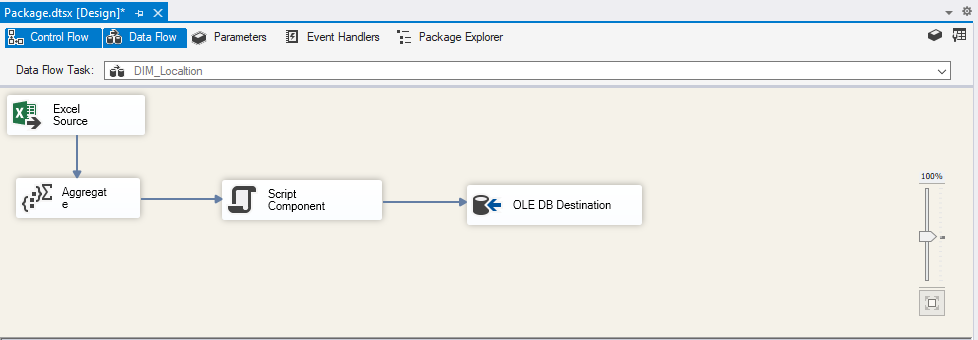
Hình 3.3.3c DIM\_Product

DIM\_Promotion



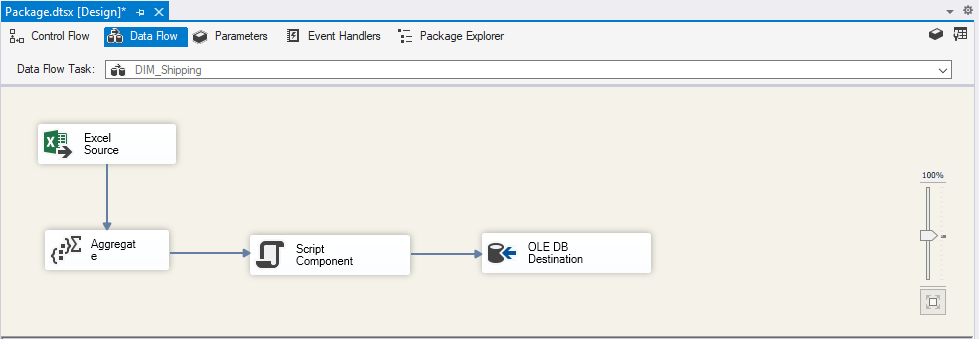
Hình 3.3.3d DIM\_Product

DIM\_Location



Hình 3.3.3e DIM\_Location

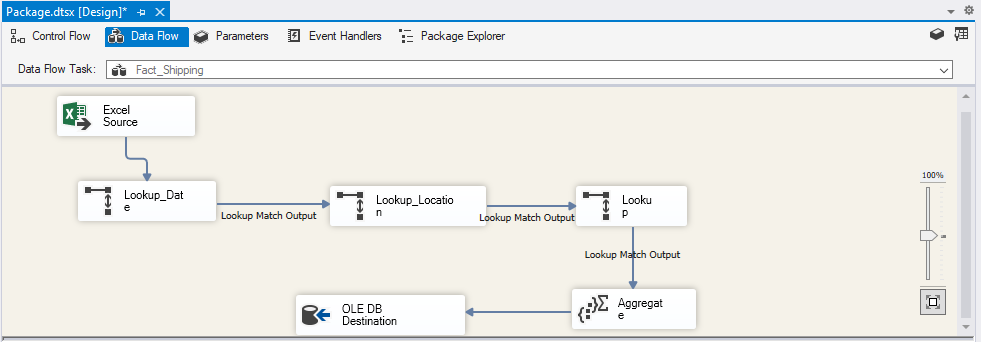
DIM\_Shipping



Hình 3.3.3f DIM\_Shipping

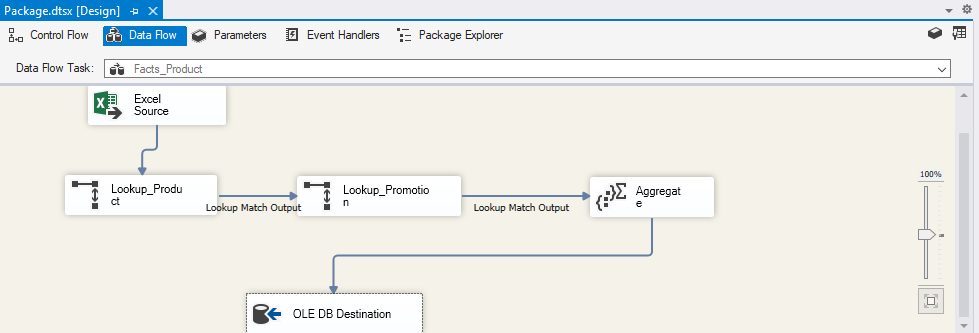
### 3.3.4 Create data flow for Fact table

Fact\_Shipping



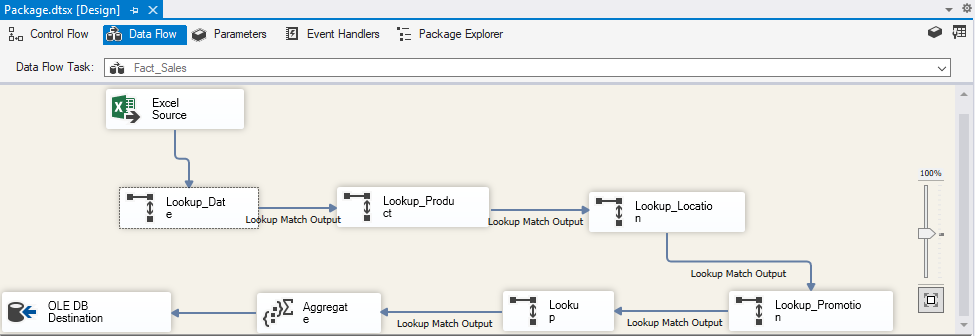
Hình 3.3.4a Fact\_Shipping

Fact\_Product



Hình 3.3.4b Fact\_Product

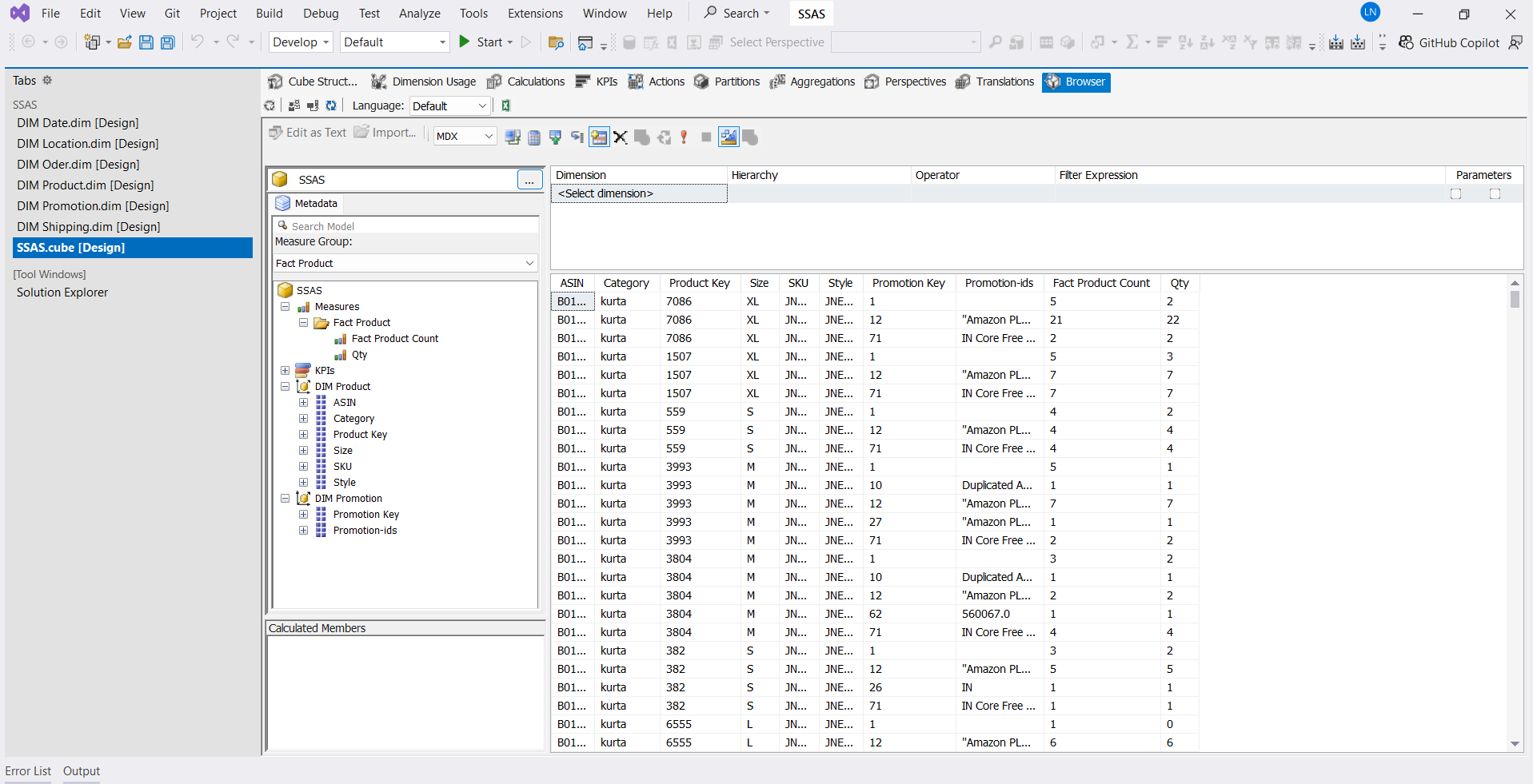
Fact\_Sales



Hình 3.3.4c Fact\_Sales

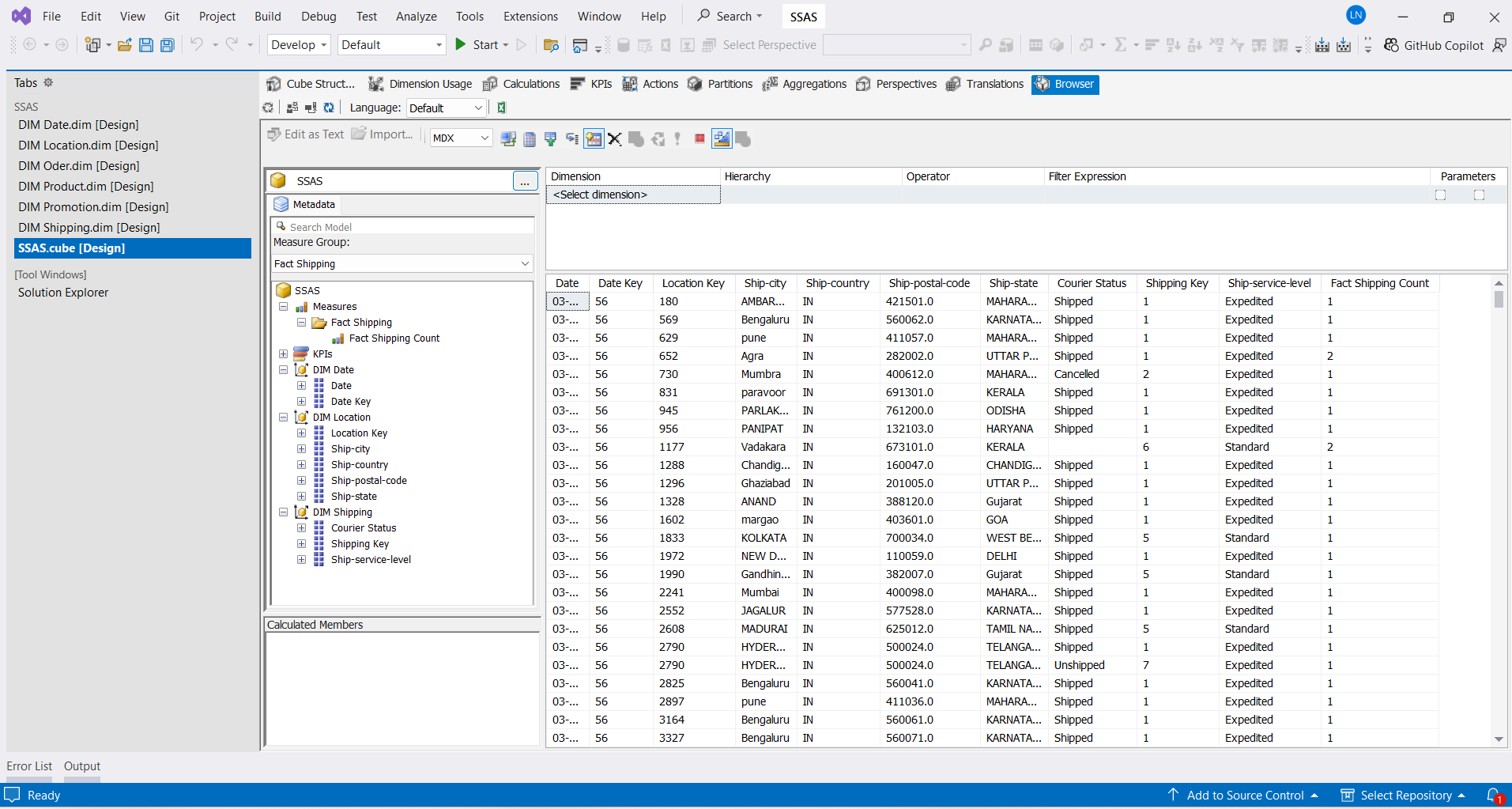
## 3.4 SSAS (Microsoft SQL Server Analysis Services)

### 3.4.1 Fact\_Product



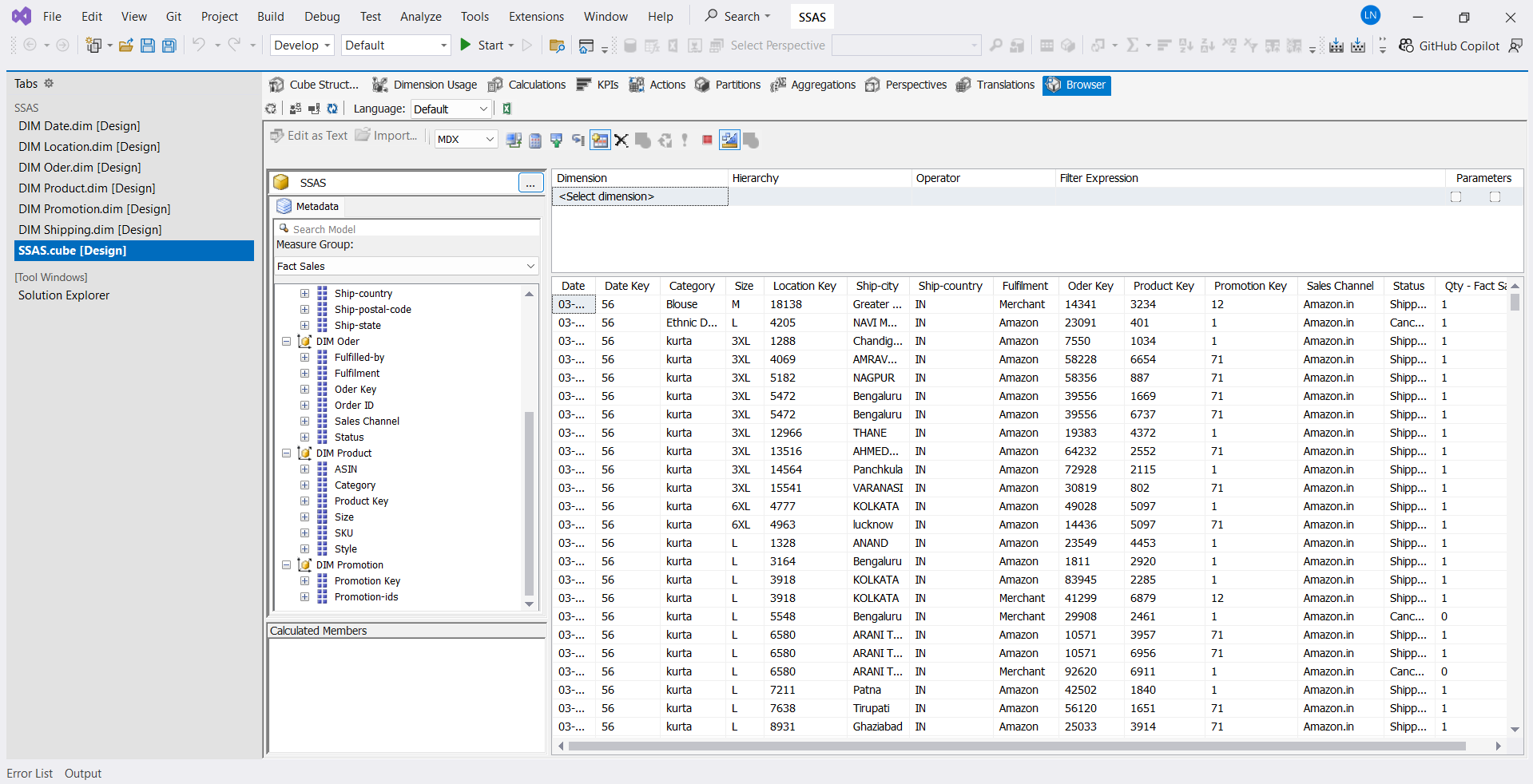
Hình 5.4.1 Fact\_Product

### 3.4.2 Fact\_Shipping



Hình 3.4.2 Fact\_Shipping

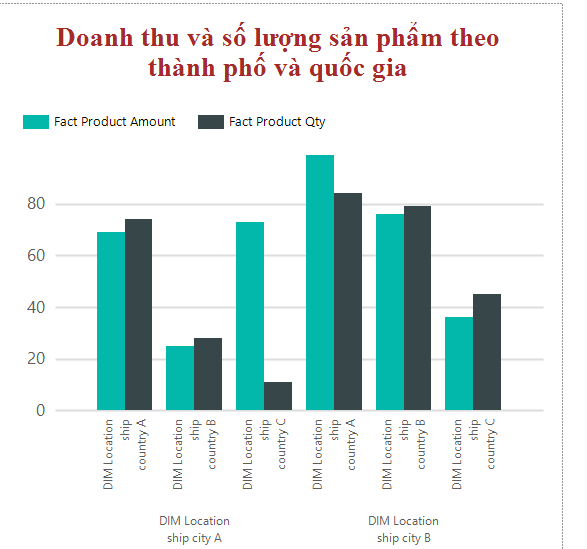
### 3.4.3 Fact\_Sale



Hình 3.4.3 Fact\_Sale

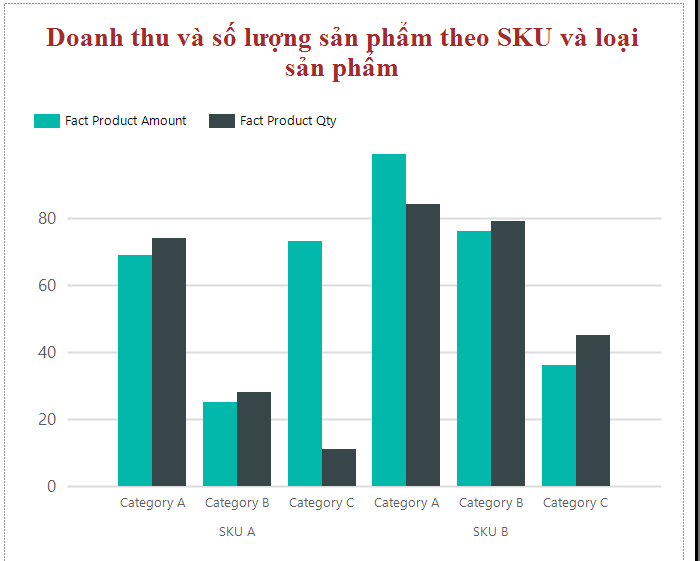
## 3.5 SSRS (SQL Server Reporting Services)

### 3.5.1 Product Sales and Volume by City and Country



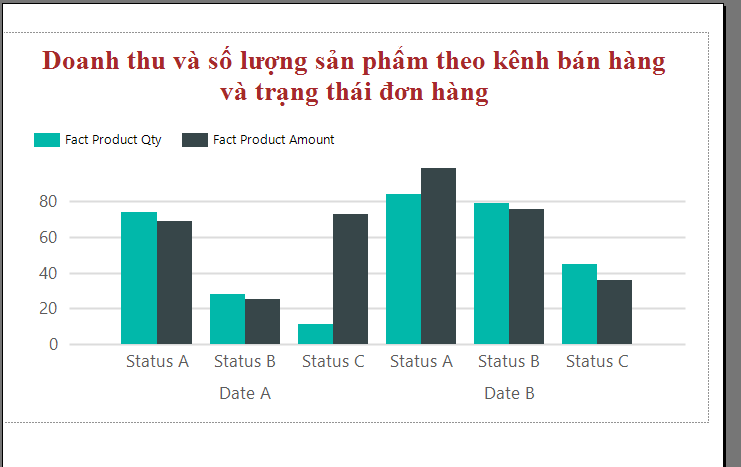
Hình 3.5.1 Product Sales and Volume by City and Country

### 3.5.2 Revenue and product quantity by SKU and product type



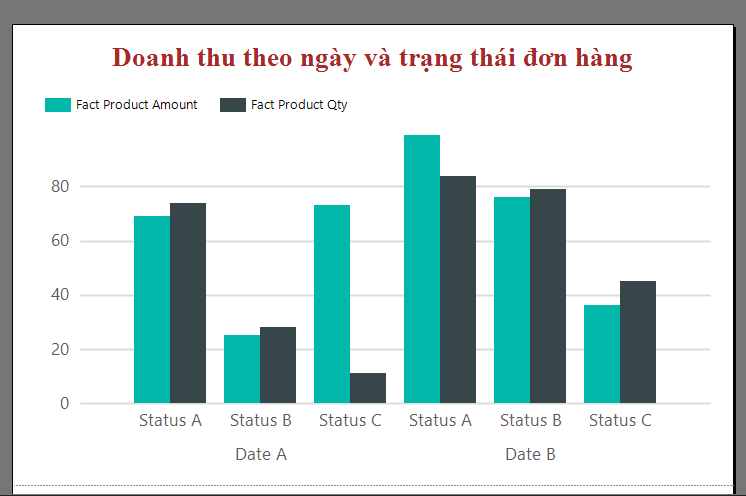
Hình 3.5.2 Revenue and product quantity by SKU and product type

### 3.5.3 Revenue and Product Quantity by Sales Channel and Order Status



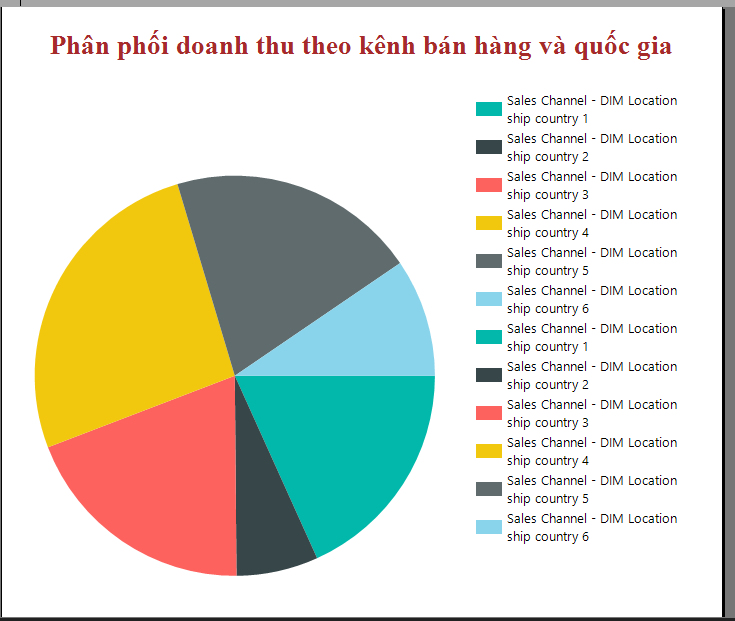
Hình 3.5.3 Revenue and Product Quantity by Sales Channel and Order Status

### 3.5.4 Revenue by Day and Order Status



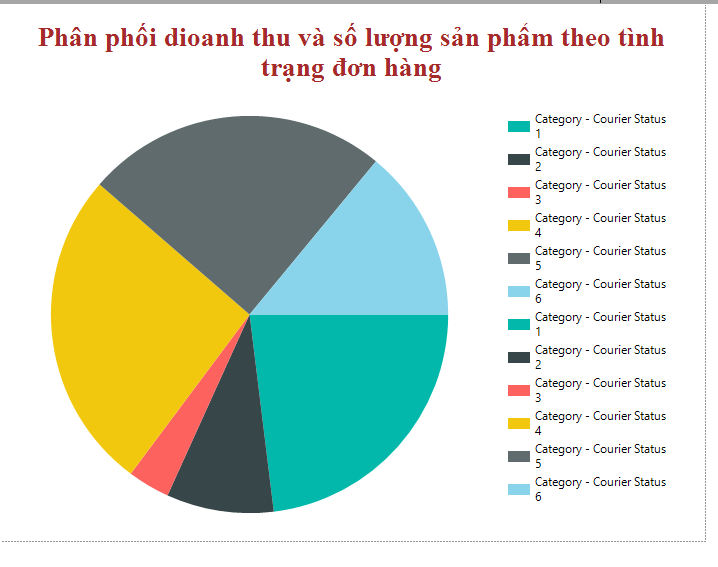
Hình 3.5.6 Revenue by Day and Order Status

### 3.5.5 Product distribution by sales channel and country



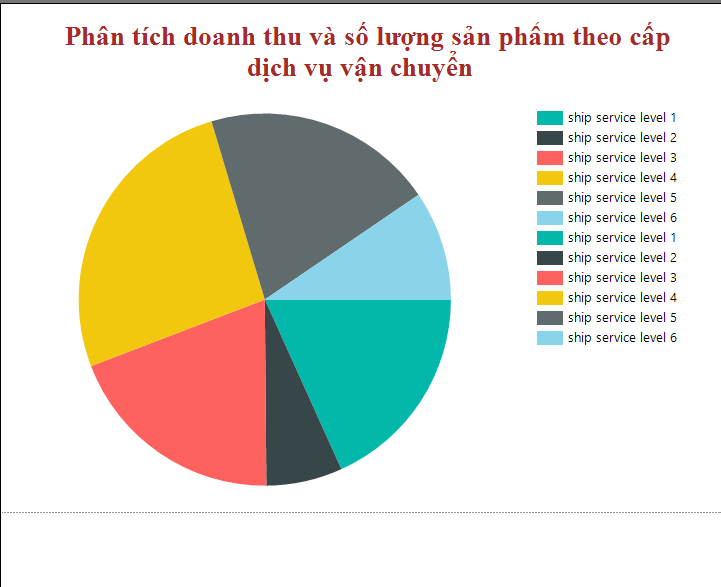
Hình 3.5.7 Product distribution by sales channel and country

### 3.5.6 Distribution of product quantity by product type and order status



Hình 3.5.6 Distribution of product quantity by product type and order status

### 3.5.7 Analysis of revenue and product quantity by shipping service level



Hình 3.5.7 Analysis of revenue and product quantity by shipping service level

# CONCLUDE

- In general, the project of building and analyzing a data warehouse for an e-commerce system has achieved significant results in organizing and providing data to support business analysis and decision making. The data warehouse helps businesses have a more comprehensive view of operational efficiency and customer behavior.

- In addition, there are still some limitations such as lack of real-time data and query performance with big data that need to be improved. Development directions and lessons learned from the project will be an important foundation for the next steps, helping to optimize and expand the data warehouse in the future.

### 1. Results not achieved

- Real-time data analysis: Current data warehouses only focus on historical data, without integrating real-time data to support immediate analysis and decision making. This can limit the effectiveness of responding quickly to changes in the market and user behavior.

- Optimizing query performance with big data: As data grows, the system will face challenges in query performance and management. There is no solution to thoroughly optimize big data processing, especially with complex queries or rapidly expanding data volumes.

- Ability to integrate with external systems: The data warehouse is not fully integrated with other systems in the enterprise, such as customer relationship management (CRM) systems or supply chain management systems, leading to a lack of synchronization and connectivity in analysis.

### 2. Development direction

- Real-time data integration: To enhance real-time analysis and quick decision making, aim to integrate real-time data into the data warehouse. This will allow businesses to respond faster to changes in the market and user behavior.

- Optimize system performance: Apply optimization methods such as indexing, partitioning, and use Big Data support tools (such as Hadoop or Spark) to enhance processing and querying capabilities as data scales.

Expand predictive analytics capabilities: Build predictive analytics models by applying machine learning techniques to predict customer demand, consumption trends, and make automated recommendations. This will help businesses better prepare for market changes and optimize inventory.

- Enhance security and data management: Implement enhanced security measures to protect customer data and comply with data privacy regulations, meeting information security requirements in e-commerce.

### 3. Lessons learned

- Importance of understanding data requirements: Thoroughly identifying and analyzing data requirements from the beginning is a key factor in building an effective data warehouse that serves the analytical goals well.

* Need for in-depth knowledge of data tools and technologies: Using BI tools and Big Data supporting technologies requires technical skills and understanding of these tools. Therefore, the process of learning and mastering the technology is very important to ensure the success of the project.
* Data management and optimization is a continuous process: Data warehouses need to be continuously managed, maintained and optimized as the data grows. This is not just a one-time project but a long-term development process to meet the growing needs of the business.

# REFERENCES

[1] Data Warehousing Fundamentals For It Professionals– Paulraj Ponniah

[2] Data Mining and Data Warehousing – Parteek Bhatia

[3] Data Warehouse Systems Alejandro Vaisman Esteban Zimányi - Design and Implementation - Springer

[4] Ngôn ngữ MDX: https://www.youtube.com/watch?v=d4oNce69GK4&list=PL2vTwB9LeGfDndnObrtpGHr2\_sQLV41oq&index=13

[5] ETL project with SSIS: https://www.youtube.com/watch?app=desktop&v=8JyOFgrhjL0