UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

– & –

PROJECT REPORT

ADVENTURE WORKS SALES & PERSON

Data Warehouse and Integration

Subject : Data Warehouse and Integration

Instructor : M.A. Ho Nguyen Van

Group : Sỏi Thận

Ho Chi Minh City, December 2022

GROUP MEMBER

NO.	FULLNAME	STUDENT ID	SIGNATURE
1	Vy Tram Viet Tuong (Leader)	K204162010	
2	Anh Dao Phuong	K204160660	
3	Cuong Ta Quoc	K204160661	
4	Mai Dang Xuan	K204160667	

ACKNOWLEDGEMENTS

First of all, I would like to thank M.S. Nguyen Van Ho, the person in charge of the subject, helped us with the background knowledge in the database subject to do the midterm project and gave our class more time to be able to complete the project better. I would like to sincerely thank T.A Le Ba Thien for sharing her knowledge, experience and preparing mini-tests during the tutoring sessions, helping us consolidate our knowledge. Although there are many efforts in the learning process, as well as in the process of making the final project, the team cannot avoid shortcomings. I look forward to your valuable comments as well as your friends to improve my results.

Once again, I sincerely thank you.

Soi Than

COMMITMENT

Our group hereby declares that the contents of this project were carried out by the group itself.

In addition to the cited references, the information used in this study is self researched and implemented by the group.

The data is collected and processed honestly.

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

1.1. Business case for the project

Adventure Works is a fictional bicycle wholesaler. The company has 97 different brands of bikes that are grouped into three categories: mountain bikes, road bikes, and touring bikes. Moreover, Adventure Works also manufacture some of its own components. Several components, accessories and clothing are purchased from outside vendors.

Adventure Works is not only selling bicycles, but it also provides accessories, clothing, and components. Many of those things are made by vendors, so AdventureWorks stands as a reseller. Adventure Works serve customers globally, including Australia, Canada, France, Germany, the United Kingdom, and the United States. There are 2 business models in Adventure Works: retail stores that sell bikes, and internet sales that serve individual customers. Usually, Adventure Works sells in bulk to retail stores, which act as resellers for its products.

Following a successful fiscal year, Adventure Works Cycles now wants to broaden its market share by targeting advertising to its best customers, extending product availability through an external Web site, and reducing the cost of sales by reducing production costs. Therefore, the company also faced the following challenges:

More advanced users need direct query access to the database for interactive queries and specialized reports. However, due to the complexity of the database, such users need too much time to learn how to create effective queries.

Users in different business units are interested in different views of data. Each group is distracted and confused by data elements unrelated to them.

It is very difficult to combine related sets of information, such as sales and sales quotas, which are difficult for business users to build. Such queries overload the database.

That means, senior managers need the facility to support the data analysis needs of the sales and marketing departments. Adventure Works sales data warehouse will be the best solution to the above problems.

1.2. Objectives of the project

1.2.1. General Objective

- Understand the process and operations of AdventureWorks Cycles
- Apply knowledge of technology and software: Microsoft SQL, Power BI to build a data warehouse to make raw data structured and analyzable, visualize them as dashboards for better and clearer reports
- Data integration, data querying and reporting supporting more effective decision making

1.2.2 Specific Objectives

- Successfully building a data storage platform specializing in Sales and Customers
- Follow up AdventureWorks Sales activities.
- Combine sets of relevant information on customer demographics and purchasing activity
- Analyze customer behavior based on purchase activity.
- From the analysis activity based on the Sales data warehouse, come up with specific sales plans or strategies.

1.3. Research Objects

- Sales and Customer Department
- Integration and ETL Process
- SQL Server Business Intelligence
- AdventureWorks database

1.4. Scope of the project

After analyzing the company and the business operations of Adventure Works, we define the scope of this project as follows:

1. Product sales analysis
2. Sales analysis by location
3. Sales analysis by customer

The scope of this development is only in Adventure Works company. The dashboard will be built in order to be used by the Executive General or Sales Manager.

1.5. Value and desired outcome of the project

About theory:

- Understand Sales Business Process
- Understanding the Data Warehouse Design Process
- Understanding the ETL Process
- Understand data analytics models and business intelligence systems

In reality:

- Create a data warehouse about Sales
- The data warehouse can reach the requirements of building intelligent business models and data analysis with business reports for Sales
- To help administrators control order information, customer information, sales performance as well as making the right decisions

1.6. Structure of the project

Report includes 5 chapters:

1. Introduction
2. Theoretical Basis
3. Requirements Analytics
4. Building Data Warehouse and Integrating Data
5. Conclusion and Future Works

CHAPTER 2: THEORETICAL BASIS

2.1. Overview of Data Warehouse

2.1.1 What is Data Warehouse?

A data warehouse is an information system which stores historical and commutative data from single or multiple sources. It is designed to analyze, report, and integrate transaction data from different sources.

Data Warehouse eases the analysis and reporting process of an organization. It is also a single version of truth for the organization for decision making and forecasting processes. (Taylor, 2022)

Data is gathered from many different sources: relational databases, flat files, online dashboards. This will lead to the process of data collection having to perform cleaning, sorting, and data reduction.

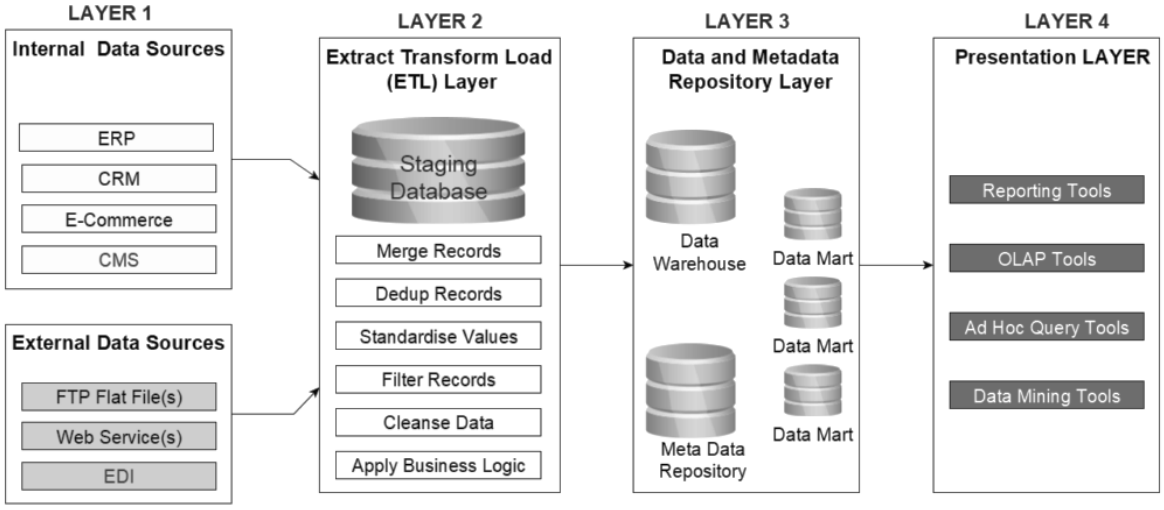


Figure 2.1 *Description of structure class of Data Warehouse*

2.1.2 DWH Architecture

A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise. Each data warehouse is different, but all are characterized by standard vital components.

Data warehouses and their architectures very depending upon the elements of an organization's situation.

Three common architectures are:

- Data Warehouse Architecture: Basic
- Data Warehouse Architecture: With Staging Area
- Data Warehouse Architecture: With Staging Area and Data Marts

2.1.2.1. Data Warehouse Architecture: Basic

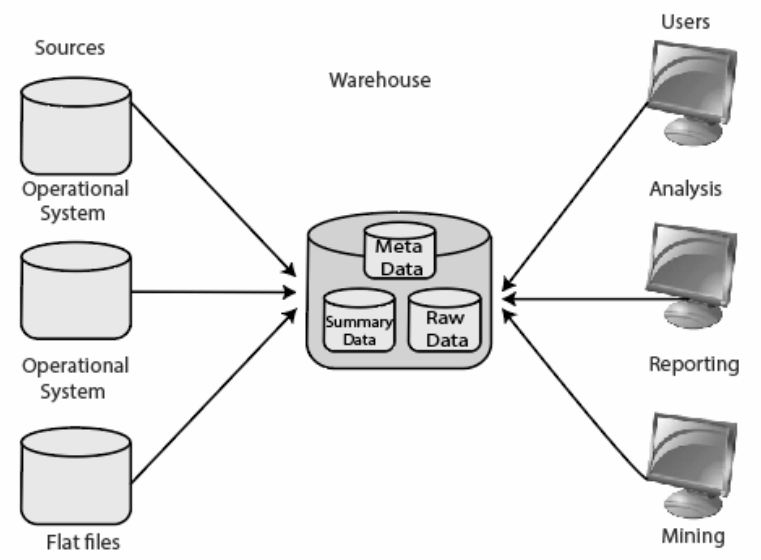


Figure 2.2 *Architecture of a Data Warehouse*

Operational System: An operational system is a method used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization.

Flat Files: A flat file system is a system of files in which transactional data is stored, and every file in the system must have a different name.

Meta Data: A set of data that defines and gives information about other data. Metadata used in Data Warehouse for a variety of purpose, including:

- Meta Data summarizes necessary information about data, which can make finding and work with particular instances of data more accessible. For example, author, data build, and data changed, and file size are examples of very basic document metadata.
- Metadata is used to direct a query to the most appropriate data source.

Lightly and highly summarized data: The area of the data warehouse saves all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.

The goals of the summarized information are to speed up query performance. The summarized record is updated continuously as new information is loaded into the warehouse.

End-User access Tools: The principal purpose of a data warehouse is to provide information to the business managers for strategic decision-making. These customers interact with the warehouse using end-client access tools.

2.1.2.2. Data Warehouse Architecture: With Staging Area

We must clean and process your operational information before putting it into the warehouse.

We can do this programmatically, although data warehouses uses a staging area (A place where data is processed before entering the warehouse).

A staging area simplifies data cleansing and consolidation for operational methods coming from multiple source systems, especially for enterprise data warehouses where all relevant data of an enterprise is consolidated.

Data Warehouse Staging Area is a temporary location where a record from source systems is copied.

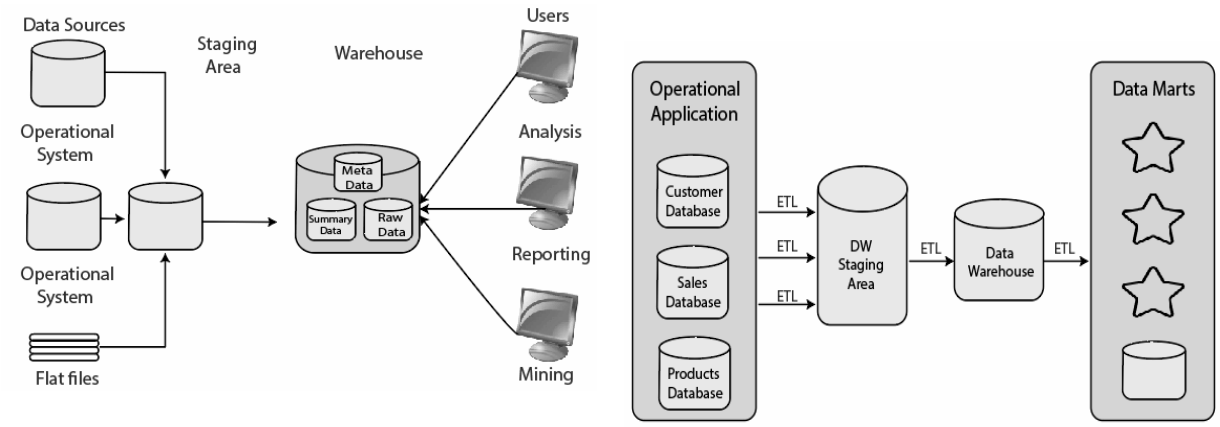


Figure 2.3 *Data Warehouse Architecture: With Staging Area*

2.1.2.3. Data Warehouse Architecture: With Staging Area and Data Marts

We may want to customize our warehouse's architecture for multiple groups within our organization. We can do this by adding data marts.

The figure illustrates an example where purchasing, sales, and stocks are separated. In this example, a financial analyst wants to analyze historical data for purchases and sales or mine historical information to make predictions about customer behavior.

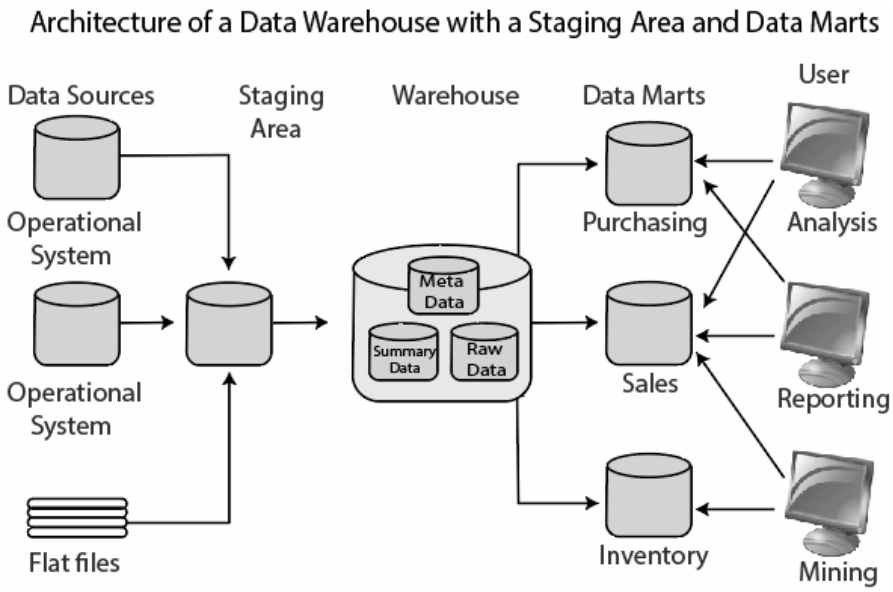


Figure 2.3 *Data Warehouse Architecture: With Staging Area and Data Mart*

2.1.2.4. Properties of Data Warehouse Architectures

The following architecture properties are necessary for a data warehouse system:

1. Separation: Analytical and transactional processing should be keep apart as much as possible.

2. Scalability: Hardware and software architectures should be simple to upgrade the data volume, which has to be managed and processed, and the number of user's requirements, which have to be met, progressively increase.

3. Extensibility: The architecture should be able to perform new operations and technologies without redesigning the whole system.

4. Security: Monitoring accesses are necessary because of the strategic data stored in the data warehouses.

5. Administrability: Data Warehouse management should not be complicated.

2.1.2.5. Types of Data Warehouse Architectures

Single-Tier Architecture

Single-Tier architecture is not periodically used in practice. Its purpose is to minimize the amount of data stored to reach this goal; it removes data redundancies.

The figure shows the only layer physically available is the source layer. In this method, data warehouses are virtual. This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer

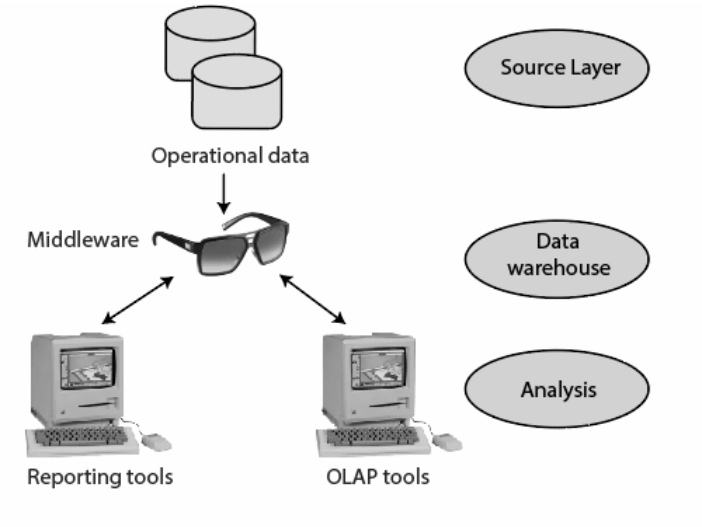


Figure 2.4 *Single-Tier Architecture*

The vulnerability of this architecture lies in its failure to meet the requirement for separation between analytical and transactional processing. Analysis queries are agreed to operational data after the middleware interprets them. In this way, queries affect transactional workloads.

Two-Tier Architecture

The requirement for separation plays an essential role in defining the two-tier architecture for a data warehouse system, as shown in fig:

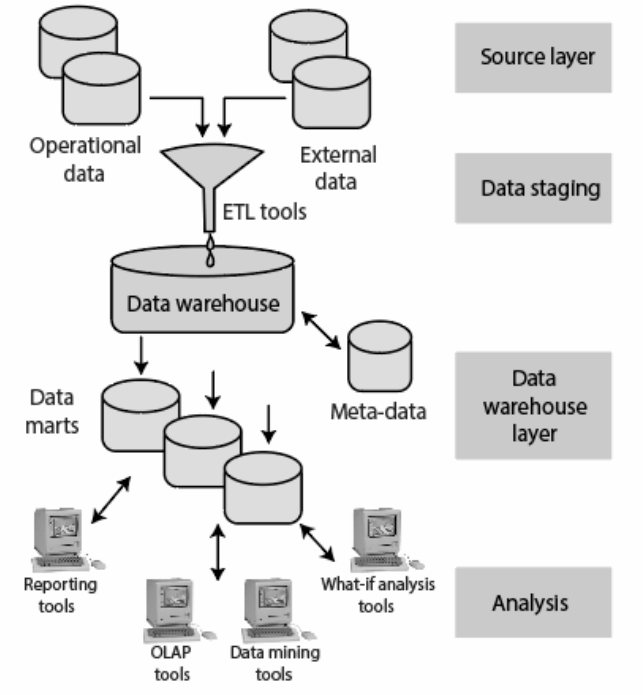


Figure 2.5 *Two-Tier Architecture*

Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses, in fact, consists of four subsequent data flow stages:

1. Source layer: A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.

2. Data staging: The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. The so-named Extraction, Transformation, and Loading Tools (ETL) can combine heterogeneous schemata, extract, transform, cleanse, validate, filter, and load source data into a data warehouse.

3. Data Warehouse layer: Information is saved to one logically centralized individual repository: a data warehouse. The data warehouses can be directly

accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments. Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.

4. Analysis: In this layer, integrated data is efficiently, and flexible accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. It should feature aggregate information navigators, complex query optimizers, and customer-friendly GUIs.

Three-Tier Architecture

The three-tier architecture consists of the source layer (containing multiple source systems), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts). The reconciled layer sits between the source data and data warehouse.

The main advantage of the reconciled layer is that it creates a standard reference data model for a whole enterprise. At the same time, it separates the problems of source data extraction and integration from those of data warehouse population. In some cases, the reconciled layer is also directly used to accomplish better some operational tasks, such as producing daily reports that cannot be satisfactorily prepared using the corporate applications or generating data flows to feed external processes periodically to benefit from cleaning and integration.

This architecture is especially useful for the extensive, enterprise-wide systems. A disadvantage of this structure is the extra file storage space used through the extra redundant reconciled layer. It also makes the analytical tools a little further away from being real-time.

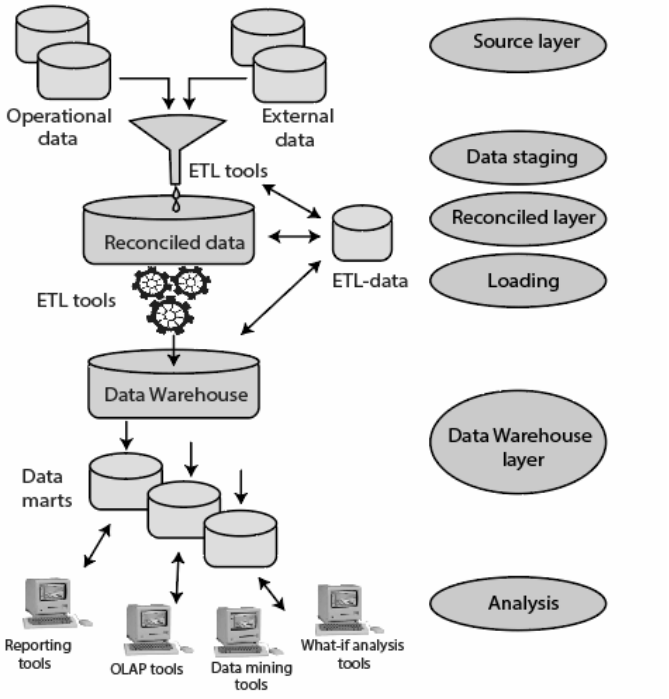


Figure 2.6 *Three-Tier Architecture*

2.2 Data warehouse and Data mart

2.2.1 What are Data warehouse and Data mart?

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

A data mart is a curated subset of data often generated for analytics and business intelligence users. Data marts are often created as a repository of pertinent information for a subgroup of workers or a particular use case.

2.1.2. Difference between data warehouses and data mart.

Parameter	Data mart	Data Warehouse
Usage	It helps to take tactical decisions for the business.	It helps to take a strategic decision.
Data Handling	Data warehousing includes a large area of the corporation, it takes a long time to process it.	Data marts are easy to use, design and implement as it can only handle small amounts of data.
Designing	Easy	Quite difficult.
Source	comes from very few sources.	comes from many sources.
Scope	Line-of-Business.	Enterprise-wide.
Size	Less than 100 GB.	Range from 100 GB to 1 TB+.
Implementation time	Data Mart is restricted to a few months.	Data Warehouse can be extended from months to years.

2.2.3. Who needs Data warehouse and Data mart?

2.2.3.1. Who needs the Data Warehouse?

- Decision-makers who rely on the mass amount of data
- Users who use customized, complex processes to obtain information from multiple data sources.
- It is also essential for those people who want a systematic approach for making decisions.
- If the user wants fast performance on a huge amount of data which is a necessity for reports, grids or charts, then Data warehouse proves useful.

2.2.3.2. Who needs Data Mart?

- Long sales cycles can use a data mart to help different departments align on the needs of the sales process.
- Who have many years of data can use a data mart to help departments focus on current data while still maintaining the historical information.
- Data marts enable them to manage large product sets better by focusing each product group on its individual mission.

2.2.4. Advantages and disadvantages of Data warehouse

2.2.4.1. Advantages of data warehouse

- The Data Warehouse helps to combine multiple data sources. It also saves time for users to access data from various sources.
- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- The Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and queries.
- A large amount of historical information is stored in a data center. This helps users to compare different times and trends to construct possible predictions.

2.2.4.2. Disadvantages of data warehouse

- Not an ideal option for unstructured data.
- Data warehouses' maintenance usually generates significant costs.
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- Data Warehouse can be outdated relatively quickly
- Difficulty in Compatibility: the use of data warehouse technology could likely require a helping hand from an independent BI team. With the intricacies of operating systems, it can be difficult for a business owner.
- Organizations need to spend lots of their resources on training and implementation purposes.

2.3 Snowflake and Star schemas

Star schema is the type of multidimensional model which is used for data warehouses. In the star schema, The fact tables and the dimension tables are contained. In this schema fewer foreign-key joins are used. This schema forms a star with a fact table and dimension tables.

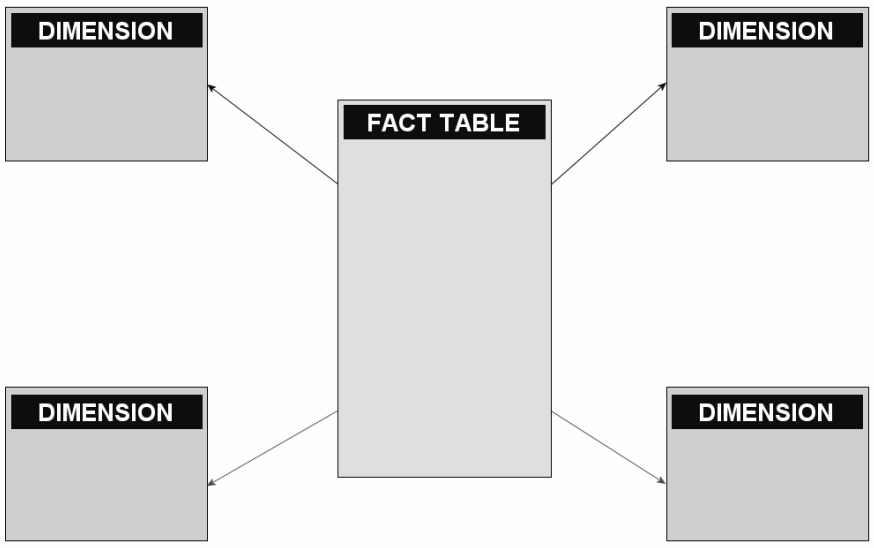


Figure 2.7 *Star schemas*

The snowflake schema represents a dimensional model which is also composed of a central fact table and a set of constituent dimension tables which are further normalized into sub-dimension table**s**

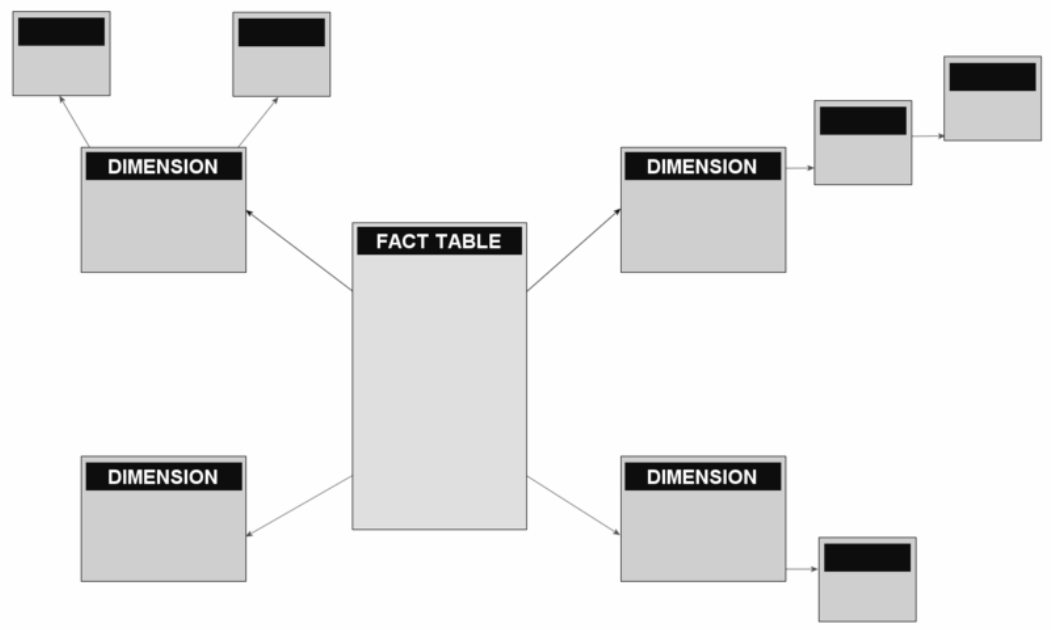


Figure 2.8 *Snow Schema*

Star Schema	**Snowflake Schema**
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
Only a single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.

Star schema is a top-down model.	While it is a bottom-up model.
It takes less time for the execution of queries.	It takes more time than star schema for the execution of queries.
Normalization is not used.	Both normalization and denormalization are used.
It's design is very simple.	it's design is complex.
It has high data redundancy.	While it has low data redundancy.

2.4 Integration and ETL Process

2.4.1 What is ETL?

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

- Extraction- Data is extracted from the available internal and external sources.
- Transformation- The goal of the cleaning and transformation phase is to improve the quality of data extracted from different sources.
- Loading- After extraction and transformation, data are loaded into the tables of the data warehouse.

2.4.2 Why do we need ETL?

Time-Efficiency: An ETL tool allows you to collect, transform, and consolidate data in an automated way. As a result, you can save plenty of time and effort otherwise spent on importing data manually

Handle Complex Data Easily: An ETL tool streamlines the tedious data cleansing tasks

Reduced Error Probability: ETL tools automate several parts of a data process, reducing manual intervention and lowering error probability.

Improved Business Intelligence And ROI: ETL tools helps companies to analyze their business data for taking critical business decisions.

2.4.3 ETL Process

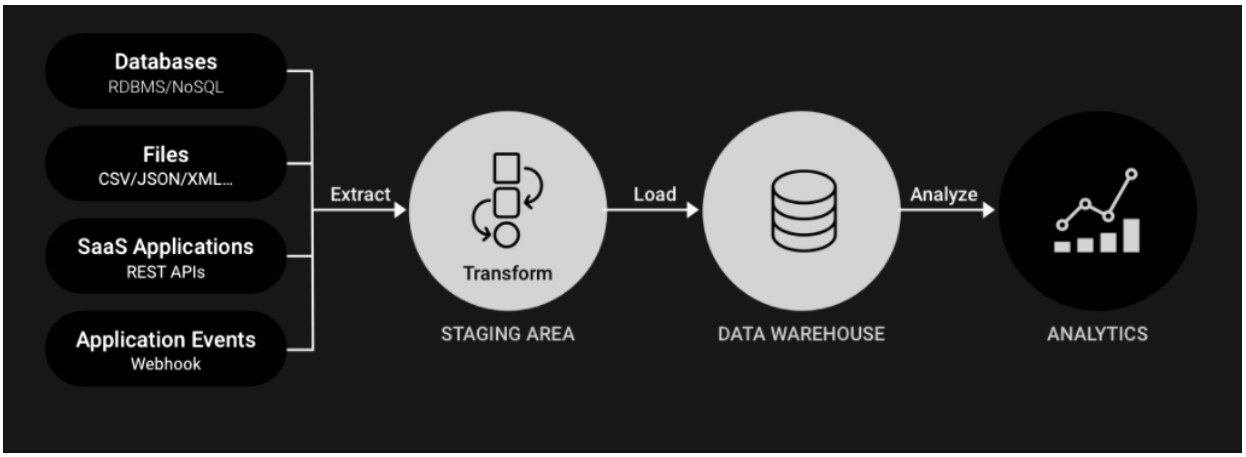


Figure 2.9 *ETL Process*

The ETL process consists of 3 steps that enable data integration from source to destination: data extraction, data transformation, and data loading.

Step 1: Extraction

To create a data warehouse, extraction typically involves combining data from these various sources into a single data set and then validating the data with invalid data flagged or removed. Extracted data may be several formats, such as relational databases, XML, JSON, and others. It is important to extract the data from various source systems and store

it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult.

Step 2: Transformation

Data extracted from the source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. The process of data transformation is comprised of several sub-processes:

- Cleansing - inconsistencies and missing values in the data are resolved.
- Cleansing: inconsistencies and missing values in the data are resolved.
- Standardization: formatting rules are applied to the dataset.
- Deduplication: redundant data is excluded or discarded.
- Verification: unusable data is removed and anomalies are flagged.
- Sorting: data is organized according to type.
- Other tasks: any additional/optional rules can be applied to improve data quality.

Step 3: Loading

This process is to load the newly transformed data into a new destination (data lake or data warehouse.)

In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, the load process should be optimized for performance.

In case of load failure, recovery mechanisms should be configured to restart from the point of failure without data integrity loss.

Three types of Loading:

- Initial Load: populating all the Data Warehouse tables
- Incremental Load: applying ongoing changes as needed periodically.
- Full Refresh: erasing the contents of one or more tables and reloading with fresh data.

Often, the three ETL phases are run in parallel to save time. For example, while data is being extracted, a transformation process could be working on data already received and preparing it for loading, and a loading process can begin working on the prepared data, rather than waiting for the entire extraction process to complete.

CHAPTER 3: REQUIREMENTS ANALYTIC

3.1. Business processes

3.1.1. Sales Department

A Sales department shows customers and sales-related data like sales orders and sales territories. A sales department is responsible for selling products or services for a company. The department comprises a sales team that works together to make sales, increase profitability and build and maintain relationships with customers to encourage repeat purchases and brand loyalty.

The customers of Adventure Works include over 700 stores and over 19000 individuals worldwide and its vendors are quantified around 100 vendors companies that supply raw materials, accessories, clothing, and components. Person schema is this department data mart including names and addresses of individual customers, vendors, and employees.

Primary type of Person schema:

- SC = Store Contact
- IN = Individual (retail) customer
- SP = Sales person
- EM = Employee (non-sales)
- VC = Vendor contact
- GC = General contact

3.1.2. The purpose of Sales

When a company's sales department has a clearly defined objective, the team members understand what actions to take to achieve it. These objectives are realistic and measurable, which improves the chances of achieving them.

Achieving a high conversion rate

The goal of a sales department is to gain a high conversion rate. The sales manager and other sales executives are responsible for researching methods to increase the conversion rate. A higher rate means that the company is spending less money and producing higher profits.

Retaining customers

Keeping existing customers rather than continuously spending funds to attract new customers can increase profitability. Customer retention can only occur when customers are happy with the company's products or services.

Growing the business

Like other departments in a company, the sales department exists for business growth and development. The ability to satisfy customers encourages repeat purchases and referrals, which helps a business grow. The sales department is one of the most important departments for business growth.

3.1.3. Sales process

Adventure Works Cycles company has 2 main sales processes are retail or wholesales stores that buy products for resale from Adventure Works Cycles sales representatives, and internet sales, which the company sells directly to customers through the company website, serving individual customers. Usually Adventure Works sells in bulk to retail stores, which acts as resellers for its products.

Sales by retail stores process has 4 steps:

- Order processing: after receiving the agent's order, the company will check the sales conditions (inventory, credit limit...). If the sales conditions are satisfied, the company proceeds to receive the order, otherwise the company announces the refusal to the customer.

- Packing and delivery of goods: After the order is approved, the company will pack and deliver the goods to the agent.
- Invoicing and recording debts: Based on the information on orders and warehouses, the company makes sales invoices and sends them to customers, and at the same time records receivables from customers.
- Payment: When it is due, the customer makes payment for the order. The Company accepts payment and records a corresponding decrease in receivables.

Internet Sales process has 2 steps:

- Order, payment: customers order payment directly through the company's website.
- Delivery, invoicing: the company delivers and issues invoices to customers.

3.2. Data source and challenges

From the determination of the requirements data, the necessary information dimensions to assist the manager in capturing the sales performance corresponding to the identified requirements are summarized as follows:

In order to observe and capture the situation and situation in each business period, as well as compare the business between time periods, the dimension of information that needs to be obtained includes: information about products, time and the number of quantity and value of products.

In order to forecast business opportunities by capturing the shopping behavior of customers in each region, through agents in the same area, the information that needs to be obtained includes: information about customers, information about regions, information about agents with specific numbers.

In addition, in order to determine the appropriate marketing, promotion and sales decisions, it is necessary to have information about the reasons for choosing to buy products, special promotions, discounts, and discounts.

Schema	Contains objects related to	Number of table
Human Resources	Employee of Adventure Works Cycles	6
Person	Names and addresses of customers who are individuals, suppliers and employees	13
Production	Products manufactured and sold by the company Adventure Works Cycles	25
Purchasing	Suppliers of parts and products purchased by the company	5
Sales	Customers and data related to the purchase	18

3.3. Business Requirements Analysis

3.3.1. Person

3.3.1.1. Demography

The analysis of demographic factors helps the company understand the characteristics of the people who buy its products and services. The right demographic data helps define your target audience more accurately thereby reducing your cost per lead and increasing sales.

We have posed specific questions to analyze demographic values related to the following:

- Which platforms do the customers choose to buy (Location - Territory- Area)?
- How does the Income variable affect purchases?
- Factors: getting married, having a house, having a car, education level affect the choice of products or not.
- Is there a difference when buying products between men and women?
- How much money, on average, a customer will spend in a given month across all product categories?

3.3.2. Sales

3.3.2.1. Product profitability analysis

Product is the connection between businesses, customers and profit. Analyzing product data to consolidate existing product development, strengthen sub-categories, especially support the sales purpose of increasing sales.

- Which products are the most profitable?
- How much Average revenue per unit? Sales amount by years? Profit growth (%)?

3.3.2.2. Market Segment

- Which markets have the most profitable products?
- Which brands are the core income for the company?
- Profit Margin of the different Regions and Sub-Categories?

3.4. IT requirements Analysis (IT & Infrastructure)

3.4.1. The Data Warehouse project

Requirements analysis and capacity planning: The first process in Data Warehouse includes identifying business needs, defining architecture, performing capacity planning, and selecting hardware and software tools. This step will involve consultation with senior management as well as various stakeholders and be done consistently.

Hardware integration: Once hardware and software have been selected, they require to be included by integrating the server, storage methods, and user software tools.

Modeling: Modeling is an important phase involved in the design of the warehouse schema. This may include using a modeling tool if the Data Warehouse is complex.

Source: Information for the Data Warehouse can come from a number of data sources. This step includes identifying and connecting the sources using ports, ODBC drives, or other wrappers.

ETL: Data from the source system will require an ETL stage. The ETL design and implementation process may include identifying the right ETL tool vendor and implementing those tools. Includes tool customization to fit your business needs.

Fill in the data warehouse: Once the ETL tools have been agreed, testing the tools will be necessary. Once everything is working properly, the ETL tools can be used to populate the repositories provided by the schema and view definitions.

User Applications: For a Data Warehouse to be useful, an end-user application is required. This step includes the design and implementation of applications according to the end-user requirements.

Deployment of repositories and applications: Once the Data Warehouse has been disseminated and the end clients are tested, the repository system and operations can be deployed for public use.

3.4.2. Data Management

The data management process includes a combination of different functions that collectively aim to make sure that the data in corporate systems is accurate, available, and accessible to ensure greater visibility and more accurate results overall, combining data from multiple data sets and different file formats from disparate sources to create a single, data warehouse or dataset ready for processing or analysis:

- Data Modeling.
- OLAP and Multi-Dimensional Analysis

3.4.3. Data Querying

Data querying is a request for a database's data so we can retrieve or manipulate it, perform calculations, automate tasks, or dig deeper through data mining, which uncovers hidden trends and relationships between data points.

3.4.4. Data Analytics

- Tools: Power BI, SSIS, SSAS, Microsoft SQL
- Dashboarding and Data Visualization

Data dashboards are progress reports as a data visualization provides an objective view of performance metrics and serves as an effective foundation for further dialogue immediately understood. A dashboard is a business intelligence tool used to display data visualizations in a way that is immediately understood.

Data visualization is a way of representing data under visual images, charts, and tables. From there, convey information to viewers more vividly and understandably.

Dashboard incorporates a Customers & Sales analytics platform that makes it easy to combine data from all systems and deep-explore this data directly in the dashboard. Thereby, the Sales department will optimize the business growth, improve customer performance and ensure customer retention.

In this project, we use related BI platforms such as Power BI to provide visualize the following cleaning and integration process with the following features:

- Dashboard.
- Interactive data visualization.
- Filtering
- Geospatial visualization and Maps.

CHAPTER 4: BUILDING DATA WAREHOUSE AND INTEGRATING DATA

4.1. Designing Data Warehouse

4.1.1. Bus Matrix

4.1.1.1 What is Bus Matrix?

Bus Matrix is the tool for designing and communicating the enterprise data warehouse bus architecture. Besides the technical design considerations, the bus matrix is used as input to prioritize DW/BI projects with business management as teams should implement one row of the matrix at a time.

4.1.1.2. Structure

Consists of rows and columns where the rows of the matrix are business

processes and columns are dimensions.

4.1.1.3. Benefits

- There is a common overall architectural design framework.
- Bus Matrix allows us to easily identify which facts have the same dimension.
- It is possible based on Bus Matrix to evaluate which tables can be combined into one report.

4.1.1.4. Design of Bus Matrix

	FactInternetSales	FactResellerSales
DimCustomer	x	x
DimDate	x	x
DimProduct	x	x
DimProductSubCategory	x	x
DimProductCategory	x	x
DimTerritoy	x	x

4.1.2 Master Data

4.1.2.1 What is master data?

Master data is the set of identifiers that provides context to business data. It is the core data that is essential for running operations within a business enterprise or unit.

4.1.2.2. Master data description

Object	**Description**
Customer	Contains customer information
Person	Detailed information about customers and employees

Sales Territory	Information about territories, regions
CountryRegion	Information about countries
Product	Date and time the record was last updated.
ProductSubCategory	Information about the company's products
ProductCategory	Information about product type

4.1.3 Transaction Data

4.1.3.1 What is Transaction data?

Transaction Data is data describing an event and is usually described with verbs. Transaction data always has a time dimension, a numerical value and refers to one more object.

4.1.3.2. Transaction data description

SalesOrderHeader	Details of each customer's order, transactions are shown through the total amount that the customer pays for the order, the total tax value that the customer has to pay, which staff sells the goods, and how to ship the goods. The delivery of goods to customers is shown in each specific region.
SalesOrderDetail	Contains detailed data about the sales process, transactions are shown with the quantity, price, payment amount of each item in the most specific time.

4.1.4. Fact and dimension tables

4.1.4.1. Dimension tables

Object	**Description**
Dim Customer	Customer detail information
Dim Product	Product detail information
Dim ProductSubCategory	Product type detail information
Dim ProductCategory	Product group type detail information
Dim Territory	Information about territories, regions
Dim Date	Time information of Sales

4.1.4.2. Fact tables

Object	**Description**
FactInternetSales	All data of customer buying online information
FactResellerSales	All data of company buying offline information

DimCustomer

SQL Command	Create Table
SELECT * FROM AdventureWorks2019.Sales.Customer C JOIN (SELECT P.BusinessEntityID , P.[FirstName] , P.[MiddleName] , P.[LastName] , P.[FirstName] + ' ' + P.[MiddleName] + ' ' + P.[LastName] as Fullname , D.[BirthDate] , D.[MaritalStatus] , D.[Gender] , D.[Education] , D.[Occupation] , D.NumberCarsOwned , D.NumberChildrenAtHome , D.HomeOwnerFlag , D.TotalChildren , D.YearlyIncome FROM AdventureWorks2019.Person.Person P JOIN AdventureWorks2019.Sales.vPersonDemographics D on P.BusinessEntityID = D.BusinessEntityID) as PersonDemo on C.PersonID = PersonDemo.BusinessEntityID	CREATE TABLE DimCustomer (**CustomerKey int IDENTITY(1,1) NOT NULL,** [FullName] nvarchar(152), [BirthDate] datetime, [Gender] nvarchar(1), [HomeOwnerFlag] bit, [MaritalStatus] nvarchar(1), [NumberCarsOwned] int, [NumberChildrenAtHome] int, [TotalChildren] int, [Occupation] nvarchar(30), [Education] nvarchar(30), [YearlyIncome] nvarchar(30), [CustomerID] int, [TerritoryID] int)

DimProductCategory

SQL Command	Create Table
SELECT ProductCategoryID , Name as CategoryName FROM AdventureWorks2019.Production.ProductCategory ;	CREATE TABLE DimProductCategoryKey (**ProductCategoryKey int IDENTITY(1,1) NOT NULL,** [ProductCategoryID] int, [CategoryName] nvarchar(50))

DimProductSubCategory

SQL Command	Create Table
SELECT ProductSubcategoryID , ProductCategoryID, Name as Name FROM AdventureWorks2019.Production.ProductSubcategory ORDER BY ProductSubcategoryID;	CREATE TABLE DimProductSubCategory (**ProductSubCategoryKey int IDENTITY(1,1) NOT NULL,** [ProductSubCategoryID] int, [ProductCategoryID] int, [Name] nvarchar(50))

DimProduct

SQL Command	Create Table
SELECT P.ProductID, P.ProductNumber, P.Name as ProductName, P.Color, P.ProductSubCategoryID, PC.ProductCategoryID, PS.Name as SubcategoryName, Pc.Name AS ProductType FROM AdventureWorks2019.Production.Product P JOIN AdventureWorks2019.Production.ProductSubcategory PS ON P.ProductSubcategoryID = PS.ProductSubcategoryID JOIN AdventureWorks2019.Production.ProductCategory PC ON PC.ProductCategoryID = PS.ProductCategoryID;	CREATE TABLE DimProduct (**ProductKey int IDENTITY(1,1) NOT NULL,** [ProductName] nvarchar(50), [Color] nvarchar(15), [ProductSubCategoryID] int, [ProductCategoryID] int, [SubcategoryName] nvarchar(50), [ProductType] nvarchar(50), [ProductID] int, [ProductNumber] nvarchar(25))

DimTerritory

SQL Command	Create Table
SELECT S.TerritoryID , S.Name as Region, R.Name as Country, S.[Group] as CountryGroup FROM AdventureWorks2019.Sales.SalesTerritory S, AdventureWorks2019.Person.CountryRegion R WHERE S.CountryRegionCode = R.CountryRegionCode;	CREATE TABLE DimTerritory (**TerritoryKey int** **IDENTITY(1,1) NOT NULL,** [TerritoryID] int, [Region] nvarchar(50), [Country] nvarchar(50), [CountryGroup] nvarchar(50))

FactResellerSales

SQL Command	Create Table
SELECT SH.SalesOrderID , SD.ProductID , SH.SalesOrderNumber , CustomerID , TerritoryID , TotalDue as TotalAmount , Freight , UnitPrice , OrderQty, CONVERT(datetime,orderdate) as OrderDate, CONVERT(datetime,ShipDate) as ShipDate, CONVERT(datetime,DueDate) as DueDate FROM [AdventureWorks2019].Sales.[SalesOrderHea der] SH JOIN [AdventureWorks2019].Sales.[SalesOrderDeta il] SD on SD.SalesOrderID = SH.SalesOrderID Where SH.OnlineOrderFlag = 0	CREATE TABLE FactResellerSales ([SalesOrderID] int, [ProductID] int, [SalesOrderNumber] nvarchar(25), [CustomerID] int, [TerritoryID] int, [TotalAmount] money, [Freight] money, [UnitPrice] money, [OrderQty] smallint, [OrderDate] datetime, [ShipDate] datetime, [DueDate] datetime, **[CustomerKey] int,** **[ProductKey] int,** **[TerritoryKey] int,** **[LookupOrderDate.DateKey] int,** **[LookupShipdate.DateKey] int,** **[LookupDueDate.DateKey] int**)

FactInternetSales

SQL Command	Create Table
SELECT SH.SalesOrderID , SD.ProductID , SH.SalesOrderNumber , CustomerID , TerritoryID , TotalDue as TotalAmount , Freight , UnitPrice , OrderQty, CONVERT(datetime,orderdate) as OrderDate, CONVERT(datetime,ShipDate) as ShipDate, CONVERT(datetime,DueDate) as DueDate FROM [AdventureWorks2019].Sales.[SalesOrderHeader] SH JOIN [AdventureWorks2019].Sales.[SalesOrderDetail] SD on SD.SalesOrderID = SH.SalesOrderID Where SH.OnlineOrderFlag = 1	CREATE TABLE FactResellerSales ([SalesOrderID] int, [ProductID] int, [SalesOrderNumber] nvarchar(25), [CustomerID] int, [TerritoryID] int, [TotalAmount] money, [Freight] money, [UnitPrice] money, [OrderQty] smallint, [OrderDate] datetime, [ShipDate] datetime, [DueDate] datetime, **[CustomerKey] int,** **[ProductKey] int,** **[TerritoryKey] int,** **[LookupOrderDate.DateKey] int,** **[LookupShipdate.DateKey] int,** **[LookupDueDate.DateKey] int**)

4.1.5. Data Warehouse model

4.1.5.1. Fact Constellation Schema

Fact Constellation Schema is defined as two or more fact data tables that share one or more dimensions. It is also known as Galaxy schema. The actual constellation diagram depicts the logical structure of a data warehouse or data mart.

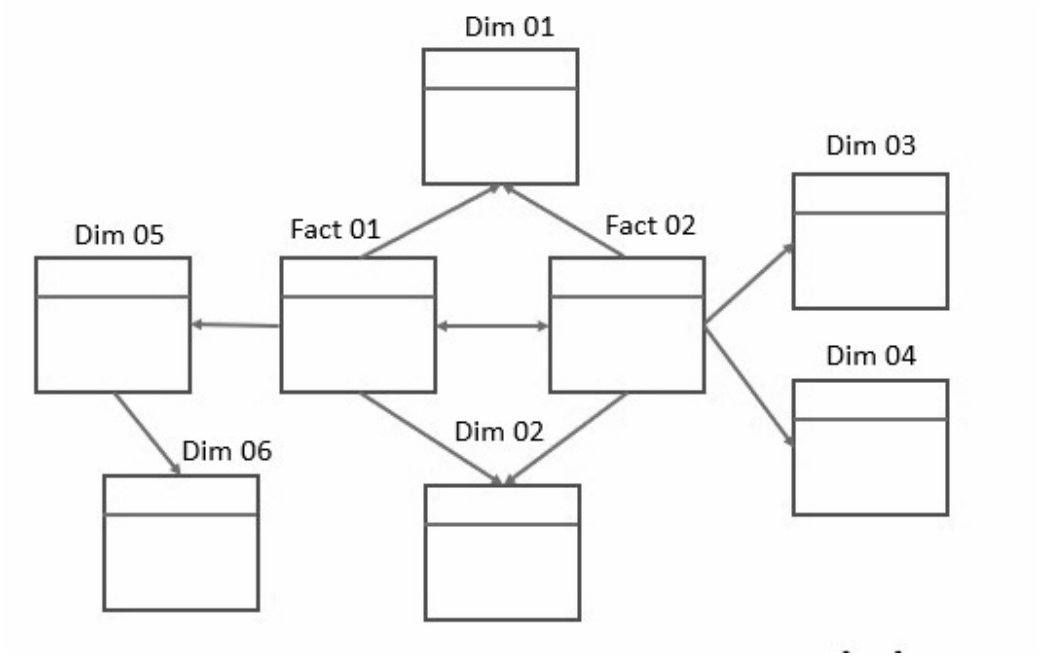


Figure 4.1 *Galaxy schema*

FactInternetSales and FactResellerSales use sharing 4 Dimensions:

- DimProduct
- DimCustomer
- DimTerritory
- DimDate

4.1.5.2. Sales Person data warehouse schema

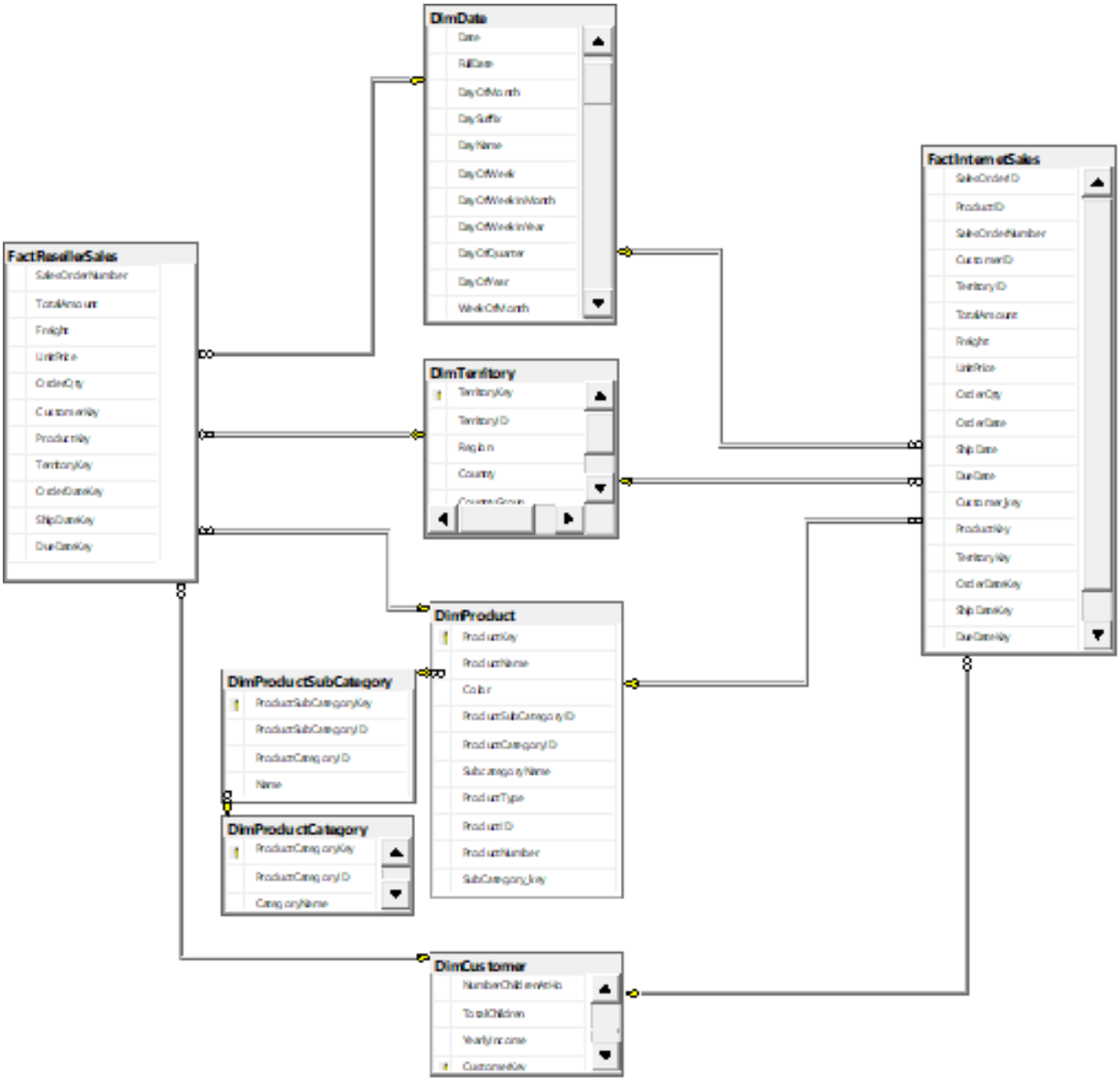
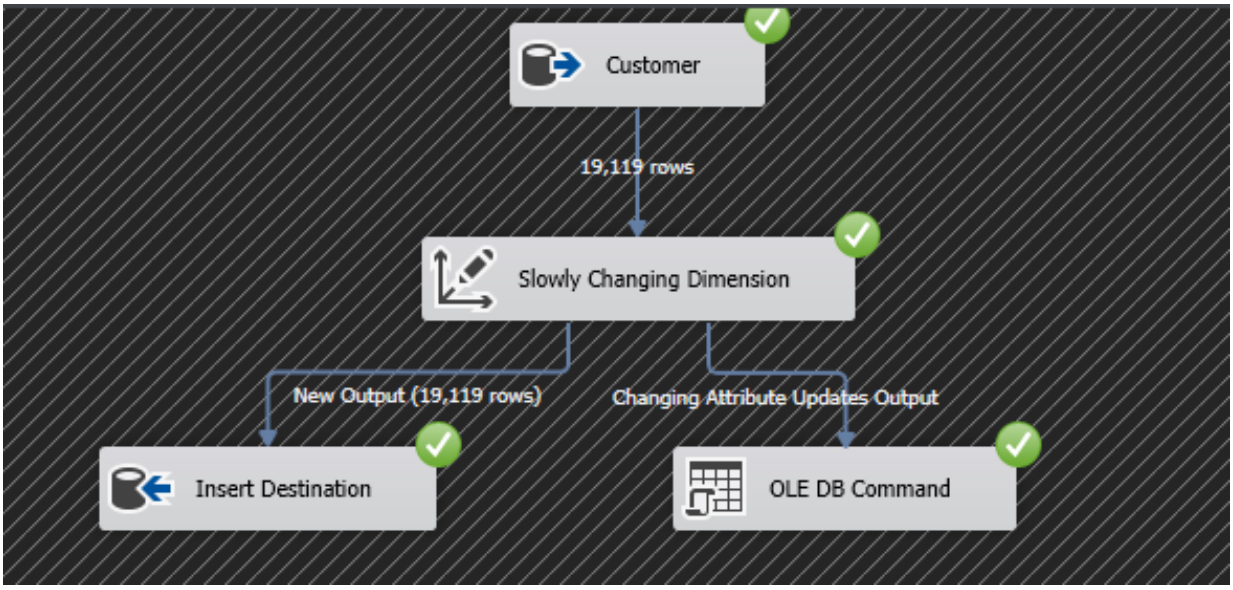


Figure 4.2 *Sales Person data warehouse schema*

4.2. ETL processes

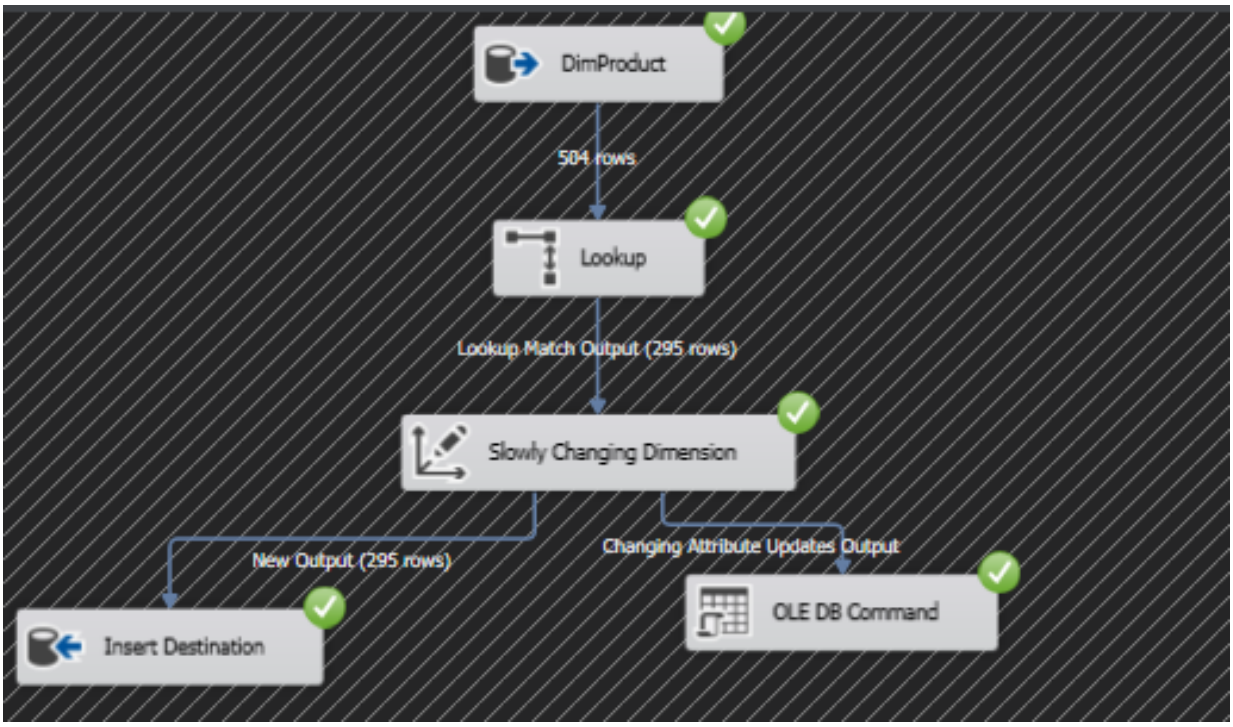
4.2.1. Dimension Table's ETL Process

Dim Customer



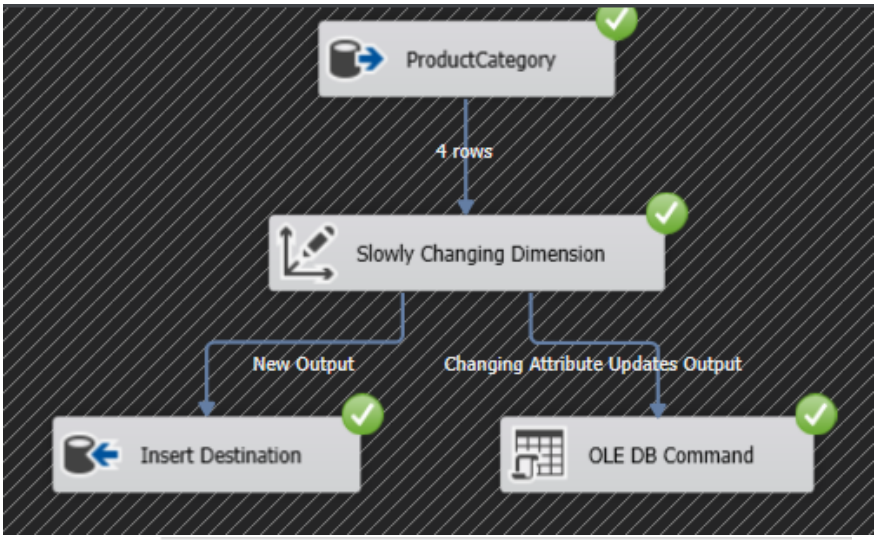
BusinessEntityID	FirstName	MiddleName	LastName	FullName	HomeOwnerFlag	NumberCarsOwned	NumberChildrenAtHome	TotalChildren	YearlyIncome	CustomerKey
2470	Erica	J	Zhang	Erica J Zhang	1	2	0	5	50001-75000	812
2471	Erica	J	Wang	Erica J Wang	0	2	0	3	25001-50000	813
2472	Isabella	L	Price	Isabella L Price	1	0	0	2	75001-100...	814
2473	Erica	P	Chen	Erica P Chen	1	1	0	1	50001-75000	815
2474	Francis	J	Navarro	Francis J Navarro	0	4	1	4	greater tha...	816
2475	Erica	A	Liu	Erica A Liu	1	4	0	0	75001-100...	817
2476	Erica	NULL	Yang	NULL	0	2	0	0	0-25000	818
2477	Erica	C	Huang	Erica C Huang	1	1	1	2	50001-75000	819
2478	Erica	NULL	Wu	NULL	1	0	1	1	25001-50000	820
2479	Erica	E	Lin	Erica E Lin	1	0	0	0	0-25000	821
2480	Erica	M	Zhou	Erica M Zhou	1	2	1	4	75001-100...	822
2481	Erica	E	Ye	Erica E Ye	1	2	3	3	50001-75000	823
2482	Erica	NULL	Zhao	NULL	1	2	1	2	25001-50000	824

DimProduct



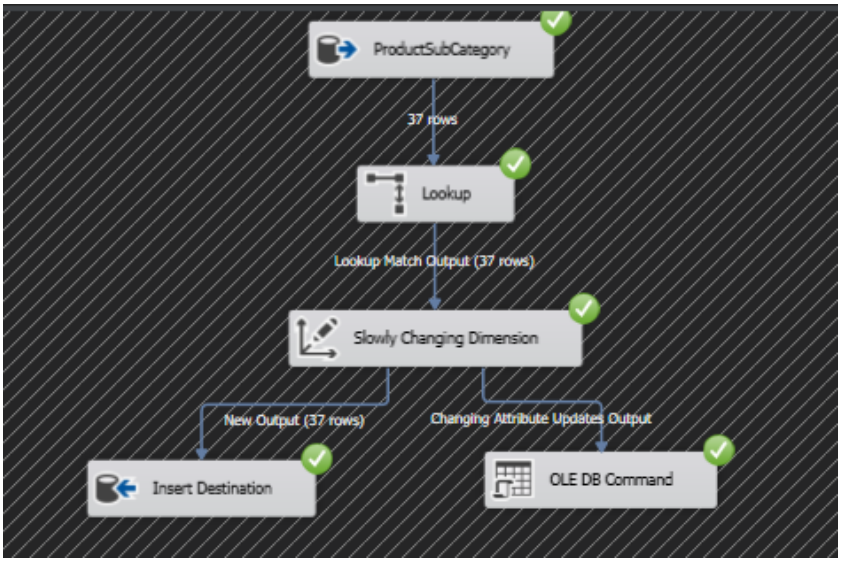
	ProductKey	ProductName	Color	ProductSubCategoryID	ProductCategoryID	SubcategoryName	ProductType	ProductID	ProductNumber	SubCategory_key
1	1	HL Road Frame - Black, 58	Black	14	2	Road Frames	Components	680	FR-R92B-58	14
2	2	HL Road Frame - Red, 58	Red	14	2	Road Frames	Components	706	FR-R92R-58	14
3	3	Sport-100 Helmet, Red	Red	31	4	Helmets	Accessories	707	HL-U509-R	31
4	4	Sport-100 Helmet, Black	Black	31	4	Helmets	Accessories	708	HL-U509	31
5	5	Mountain Bike Socks, M	White	23	3	Socks	Clothing	709	SO-B909-M	23
6	6	Mountain Bike Socks, L	White	23	3	Socks	Clothing	710	SO-B909-L	23
7	7	Sport-100 Helmet, Blue	Blue	31	4	Helmets	Accessories	711	HL-U509-B	31
8	8	AWC Logo Cap	Multi	19	3	Caps	Clothing	712	CA-1098	19
9	9	Long-Sleeve Logo Jersey, S	Multi	21	3	Jerseys	Clothing	713	LJ-0192-S	21
10	10	Long-Sleeve Logo Jersey, M	Multi	21	3	Jerseys	Clothing	714	LJ-0192-M	21
11	11	Long-Sleeve Logo Jersey, L	Multi	21	3	Jerseys	Clothing	715	LJ-0192-L	21
12	12	Long-Sleeve Logo Jersey, XL	Multi	21	3	Jerseys	Clothing	716	LJ-0192-X	21
13	13	HL Road Frame - Red, 62	Red	14	2	Road Frames	Components	717	FR-R92R-62	14
14	14	HL Road Frame - Red, 44	Red	14	2	Road Frames	Components	718	FR-R92R-44	14
15	15	HL Road Frame - Red, 48	Red	14	2	Road Frames	Components	719	FR-R92R-48	14
16	16	HL Road Frame - Red, 52	Red	14	2	Road Frames	Components	720	FR-R92R-52	14
17	17	HL Road Frame - Red, 56	Red	14	2	Road Frames	Components	721	FR-R92R-56	14

DimProductCategory



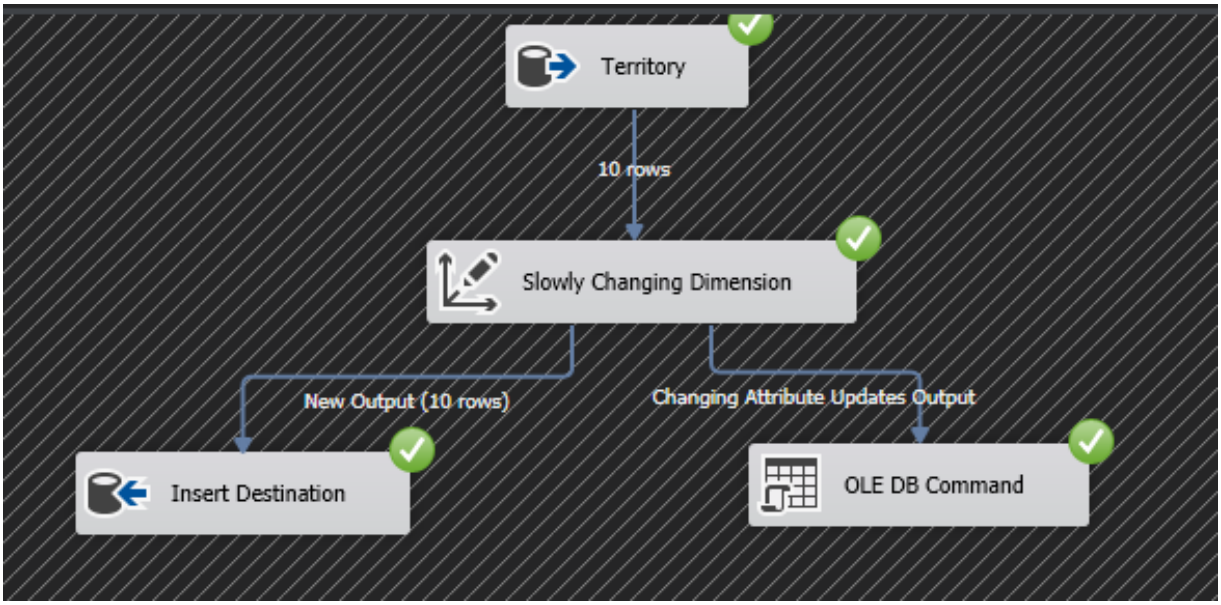
	ProductCategoryKey	ProductCategoryID	CategoryName
1	1	4	Accessories
2	2	1	Bikes
3	3	3	Clothing
4	4	2	Components

DimProductSubCategory



	ProductSubCategoryKey	ProductSubCategoryID	ProductCategoryID	Name
1	1	1	1	Mountain Bikes
2	2	2	1	Road Bikes
3	3	3	1	Touring Bikes
4	4	4	2	Handlebars
5	5	5	2	Bottom Brackets
6	6	6	2	Brakes
7	7	7	2	Chains
8	8	8	2	Cranksets
9	9	9	2	Derailleurs

DimTerritory



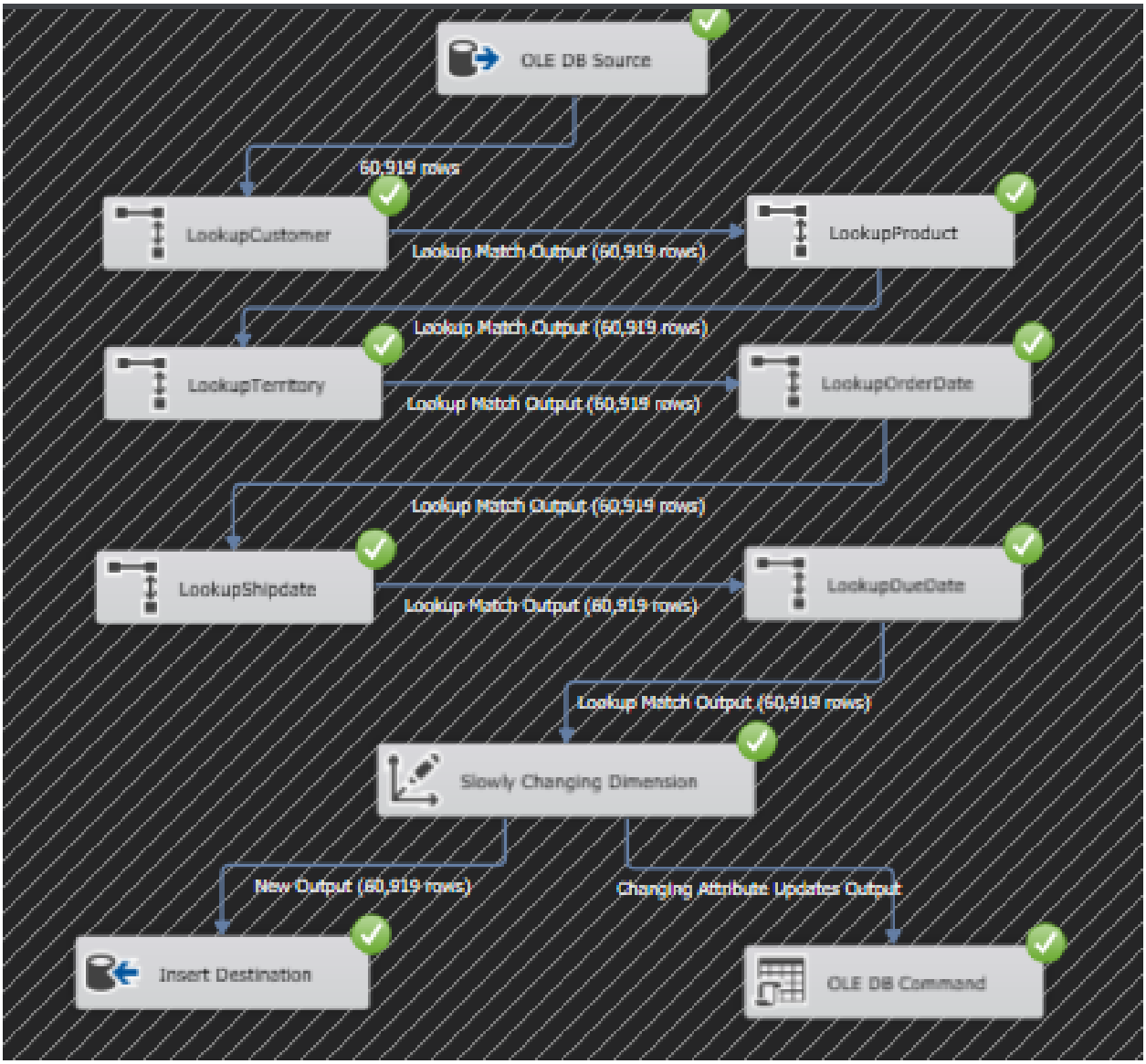
	TerritoryKey	TerritoryID	Region	Country	CountryGroup
1	1	1	Northwest	United States	North America
2	2	2	Northeast	United States	North America
3	3	3	Central	United States	North America
4	4	4	Southwest	United States	North America
5	5	5	Southeast	United States	North America
6	6	6	Canada	Canada	North America
7	7	7	France	France	Europe
8	8	8	Germany	Germany	Europe
9	9	9	Australia	Australia	Pacific
10	10	10	United Kingdom	United Kingdom	Europe

4.2.2. Fact Table's ETL Process

FactInternetSales

	SalesOrderID	SalesOrderNumber	TotalAmount	Freight	UnitPrice	OrderQty	Customer_key	ProductKey	TerritoryKey	OrderDateKey	ShipDateKey	DueDateKey
1	43697	SO43697	3953.9884	89.4568	3578.27	1	12420	45	6	20110531	20110607	20110612
2	43698	SO43698	3756.989	84.9998	3399.99	1	13861	69	7	20110531	20110607	20110612
3	43699	SO43699	3756.989	84.9998	3399.99	1	8895	69	1	20110531	20110607	20110612
4	43700	SO43700	772.5036	17.4775	699.0982	1	9553	63	4	20110531	20110607	20110612
5	43701	SO43701	3756.989	84.9998	3399.99	1	9700	69	9	20110531	20110607	20110612
6	43702	SO43702	3953.9884	89.4568	3578.27	1	11856	46	4	20110601	20110608	20110613
7	43703	SO43703	3953.9884	89.4568	3578.27	1	13239	45	9	20110601	20110608	20110613
8	43704	SO43704	3729.364	84.3748	3374.99	1	5332	74	9	20110601	20110608	20110613
9	43705	SO43705	3756.989	84.9998	3399.99	1	16600	67	9	20110601	20110608	20110613
10	43706	SO43706	3953.9884	89.4568	3578.27	1	2571	47	4	20110602	20110609	20110614
11	43707	SO43707	3953.9884	89.4568	3578.27	1	10195	47	4	20110602	20110609	20110614
12	43708	SO43708	772.5036	17.4775	699.0982	1	7716	60	10	20110602	20110609	20110614

FactResellerSales

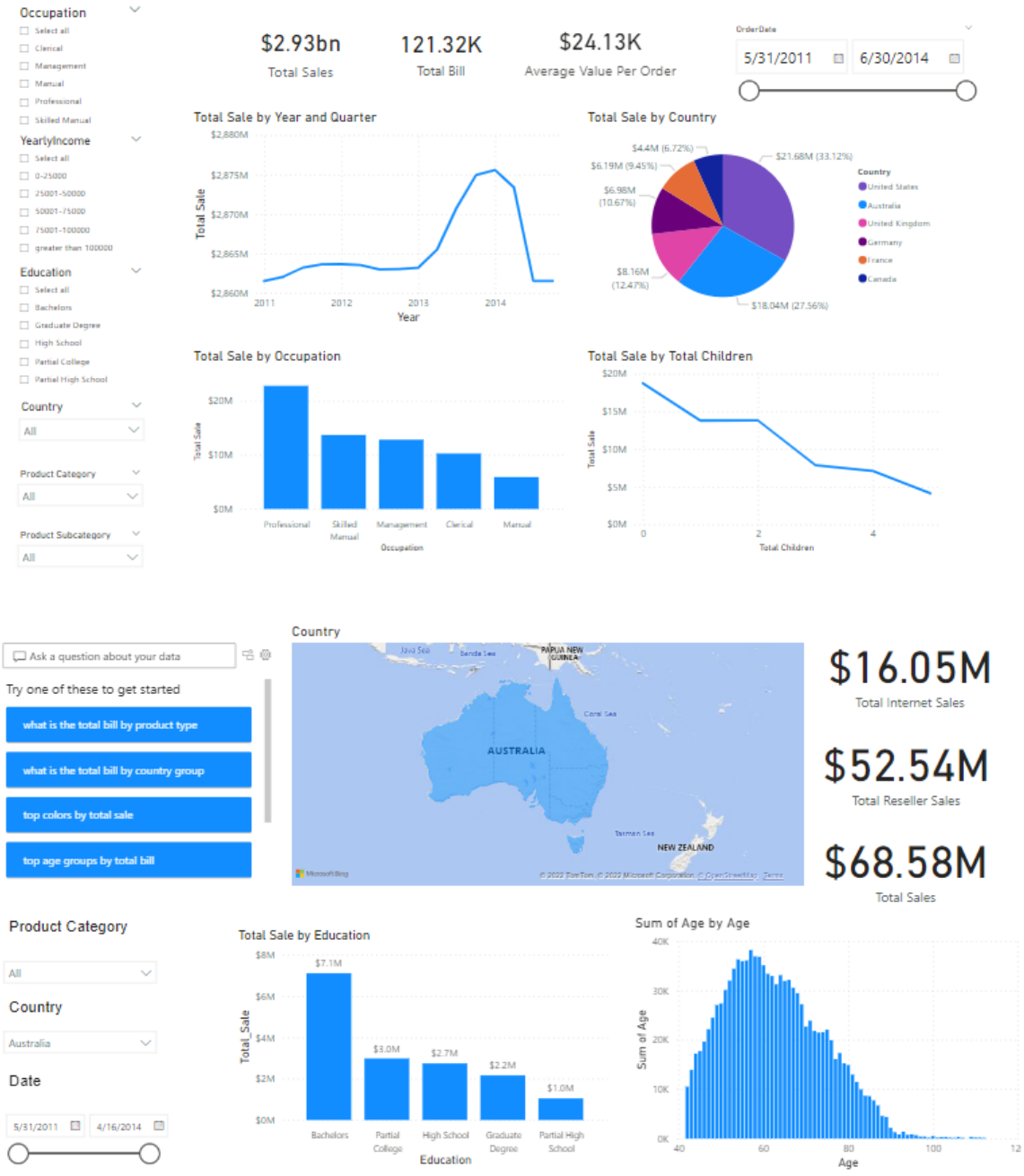


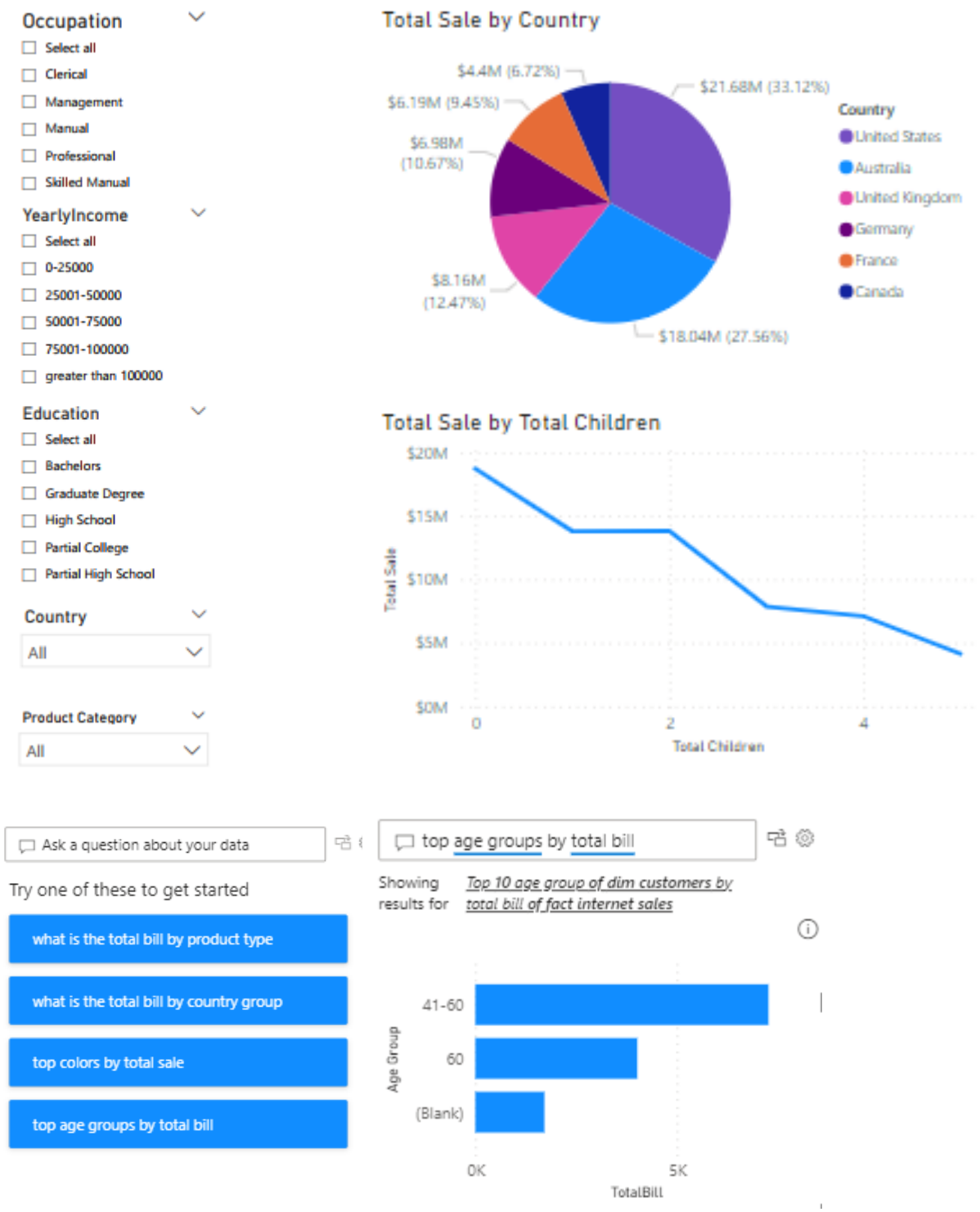
	SalesOrderNumber	TotalAmount	Freight	UnitPrice	OrderQty	CustomerKey	ProductKey	TerritoryKey	OrderDateKey	ShipDateKey	DueDateKey
1	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
2	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
3	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
4	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
5	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
6	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
7	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
8	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
9	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612
10	SO43659	23153.2339	616.0984	20.1865	4	342	7	5	20110531	20110612	20110612

CHAPTER 5: CONCLUSION AND FUTURE WORKS

5.1. Results

5.1.1. Build a BI report from the SalesPerson data warehouse

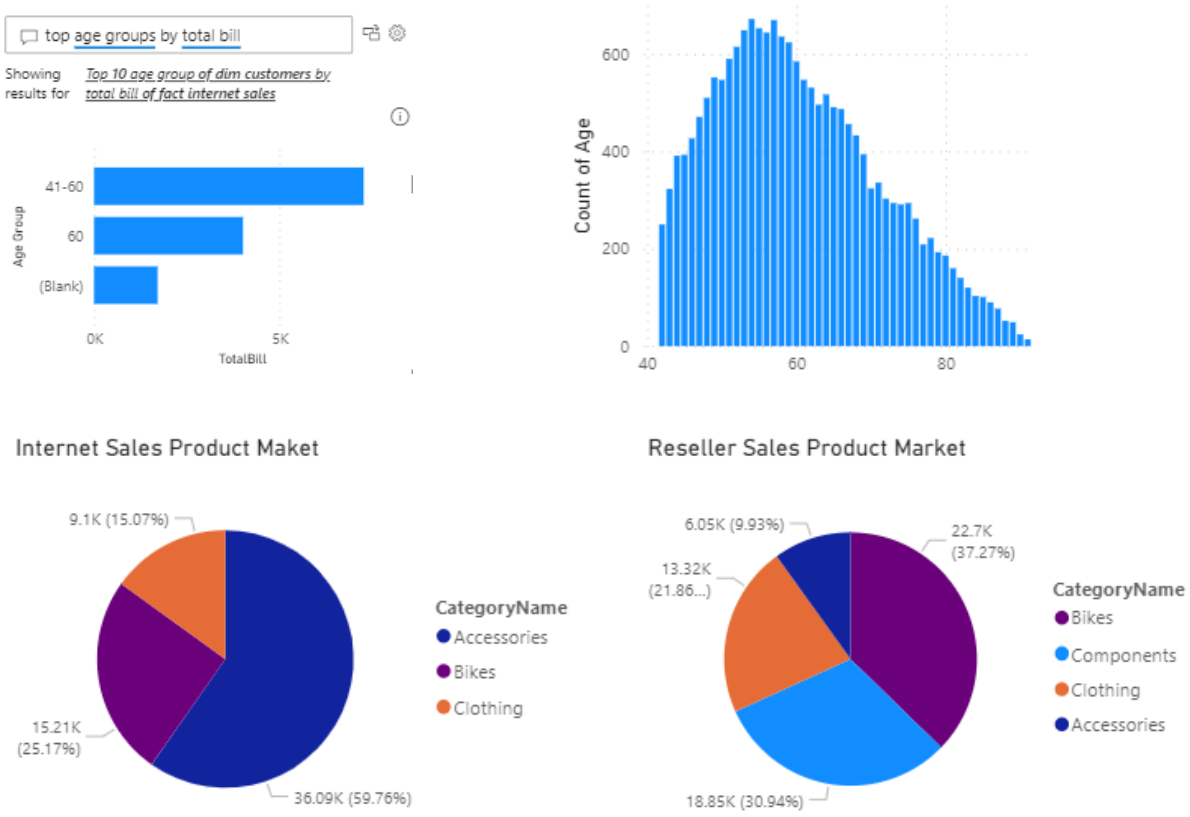




Connect data from OLE DB Sale Person Data Warehouse to Power BI apps to support analytics. Power BI support to build dashboard, interactive data visualization, filtering, geospatial visualization and Maps.

In addition, Power BI also supports automatic questioning for insight analysis.

5.1.1. Data Insight



In terms of age of customers, we know that customers aged 40-60 are particularly interested in Adventureworks Cycle products. The highest is at age 57 with 670 customers.

There is also a difference between the method of sale and the industry in which it is purchased. As shown above, customers buying through the Internet have absolutely no product section about Components.

Customers buying via the Internet prefer Accessories products, on the other hand, at Reseller stores, Bikes is the most chosen product category (accounting for 37.27%).

With the Filter Demography and Auto Q&A from Power BI, we can easily analyze the questions posed related to the age, gender, and education factors that affect customers' purchases and can Create many Sales reports according to business requirements.

5.2. Limitations

Currently, operational management information systems have been widely applied in administrative units as well as enterprises. They range from large-scale organizations with branches in many provinces to a company that only conducts business within the district. These systems have been promoting efficiency to help businesses manage operational processes closely and quickly. The next stage of the "Operational Information Systems (OLTP)" phase is the synthesis and statistical analysis of huge data sources resulting from the use of those systems. Data warehouse technology was born in the early 90s and became an exciting field in the world.

In the country, this technology has begun to be researched and put into application in the past few years, but according to what has been shown outside, the interest in the application of this technology is still quite superficial. With the set goal, the thesis has presented the most basic issues about Data Warehouse Technology: Data Warehouse architecture, dimensional database theory, OLAP services and applications on it. In the thesis, a complete and profound presentation of the main group of objects in the online analytical processing problem has been presented, Fact and Dimension table, Deploy data in Power BI.

5.3. Future works

You can design an enterprise-scale data warehouse from a data mart. This project is scalable across the industry. Due to my limited time, I will be designing and developing in SSAS to create a cube that will be completed in the near future.

In the future, we may integrate the business intelligence window into our collaboration tools. The goal of the proposed design is to support decision makers and perform data mining and data analysis on data stored in the warehouse. This will ultimately allow you to discover important trends and patterns.

Division of work and contributions

	Tram Viet Tuong Vy	**Dao Phuong Anh**	**Ta Quoc Cuong**	**Dang Xuan Mai**
Chapter 1	100%			
Chapter 2		100%		
Chapter 3	10%	20%	20%	50%
Chapter 4	100%			
Chapter 5	5%	50%	45%	

REFERENCES

[1] Ho Trung Thanh, Phan tich kho du lieu trong kinh doanh, Ho Chi Minh: ĐHQG HCM, 2016.

[2] "Data Warehouse Architecture", javatpoint, https://www.javatpoint.com/data-warehouse-architecture. Accessed 22 December 2022.

[3] "Understanding Sales Reporting." *Deskera*, https://www.deskera.com/blog/sales-reporting/. Accessed 22 December 2022.