UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEM

END-TERM PROJECT REPORT

SUBJECT: DATA WAREHOUSE AND DATA INTEGRATION

MODULE: PRODUCT

&

MODULE: PURCHASING

GROUP 4: MULTIVERSE – Class K20416C

LECTURER: Mr. Nguyen Van Ho

Teaching Assisant: Mr. Le Ba Thien

Ho Chi Minh, December 24th, 2022

GROUP 4: MULTIVERSE

No.	Student's Name	Student's ID	Work evaluation
1	Trương Công Vinh	K204162008	100%
2	Bùi Thu Vân	K204162007	100%
3	Lê Chí Hào	K204161988	100%
4	Trần Triệu Vy	K204162011	100%

ACKNOWLEDGEMENTS

First of all, our team would like to express our sincere thanks to Mr. Nguyen Van Ho (Lecturer of Data Warehouse and Integration) and Mr. Le Ba Thien (Teaching Assistant). During the time doing the research, the team has put knowledge from theory to practice while also learning and researching new information. Under the research effort, maximal application of the obtained information to produce the finest project report. However, due to the short time frame and lack of in-depth expertise, the team was unable to prevent flaws during the implementation process. As a result, the group is looking forward to hearing comments from the teachers on how to enhance the information that the group has acquired in the most correct way and contribute to the group's preparation for future tasks.

COMMITMENT

I now certify that the aforesaid project is the result of our group's research efforts, which were overseen by lecturer Nguyen Van Ho and assistant lecturer Le Ba Thien. The claims in the project are also the product of the author's direct, serious, independent investigation, as well as the foundation for finding, comprehending, and studying scientific publications or translations. Others have already been announced. The initiative will continue to contribute to impartiality, honesty, and science.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACRONYMS	DEFINITION
DWH	Data Warehouse
DBMS	Database Management System
OLAP	On-line Analytical Processing
BI	Business Intelligence
SS	Star Schemas
SF	Snowflake
GS	Galaxy Schema
ETL	Extraction-Transformation-Loading
SQL	Structured Query Language
CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
HR	Human Resources

CHAPTER 1: INTRODUCTION

1.1. Business case for the project

Background: The business produces and sells bicycles made of metal and synthetic materials. North America, Europe, and Asia are the company's main markets. The company's headquarters are in Bothell, Washington.

With a vast scale, a great amount of products, and a range of models, business goods management becomes a major issue. Furthermore, the vast market necessitates better inventory control. Following that, the number of items in each warehouse must be precisely and punctually tallied. In other words, warehouse management is a significant and necessary necessity for the manufacturing organization in general, and the procurement of raw materials from suppliers in particular, all of which must be specified.

In reality, inventory management analysis necessitates determining the ideal amount of merchandise that a company should maintain on hand. It entails comprehending how various variables influence manufacturing, sales, distribution demand, and amount acquired. Decision makers can build a clearer picture of warehouse performance and forecast how many customers will desire by putting all of these elements together.

However, in order to maintain track of these qualities, the corporation normally needs to know what sort of products should go to which warehouse, which goods should be prioritized or when they are anticipated to be sent, and what kind of goods should be prioritized or when they are expected to be dispatched. What items were ordered, what commodities were delivered, and what their exact amounts were. At various stages throughout the supply chain, inventories are categorized differently and become data points that must be tracked to guarantee a smooth transition. Furthermore, anticipating client order demands and stocking enough items to satisfy consumers, as well as the quantity of raw materials to be obtained to meet manufacturing needs, is critical in the production of this process.

In summary, effective approaches are required to maximize the company's products management, particularly in manufacturing firms. It is critical to be able to present comprehensive and accurate product data to the organization so that managers/decision makers may build suitable plans and directions.

1.2. Objectives of the project

1.2.1. General Objective

As a result, our study focus in this project is largely to create a data warehouse to satisfy the demand of providing reports on current inventory levels, current finished products shipments, and quantity has been acquired. A request to examine daily inventory levels: available quantities, shipping quantities, requested quantities, and actually received can also be fulfilled. This will most likely evolve into an hourly request in the future.

1.2.2 Specific Objectives

We want to establish a database where you can simply check up all of the information about your items, from their basic manufacture to how they will be acquired from grant vendors. Look up all the transactions that have been made; for your own production, you will find out what you have produced, how much is produced, their production threshold and quantity success and failure of those products; on the other hand, you will look up how much you bought from the supplier and how much you actually received, and use that difference to calculate the number at the end of term.

1.3. Research Objects

We have two modules, which include multiple fact tables and dimensions. A single fact joins multiple dimension members. Each module represents a solution for a business case.

Module 1: Product. This module is used to show the information of products in production apartments. This module includes 2 fact tables.

FactInventory includes 4 dimensions: Product, Location, Time and Category.

FactTransactionCost include 4 dimensions: WorkOrderRouting, Location, TransactionHistory and Time. This fact is used to show the cost of production in process.

Module 2: Purchasing. This module is used to show the information of products in which we have purchased and audit. This module includes 1 fact table.

FactPurchasing includes 4 dimensions: Vendor, ShipMethod, Product and Time.

1.4. Scope of the project

Time Period: December 10th, 2022 to December 22nd, 2022.

Research population: Data from database AdventureWorks2019.

Theories: Data Warehouse and Data Integration.

Purpose: Implementation of knowledge from theories to resolve the particular business case, using data from a real database.

Limitations: This study only used data from a database with selected tables and columns to solve specific problems mentioned in previous parts.

1.5. Value and desired outcome of the project

In this study, we proposed solutions for business by implementing data warehouse and data integration knowledge to have a better view of solving problems in real cases. We know how to perform data solutions and retrieve insight from the result of this study. When there is a business problem, we can consider applying a solution of data warehouse and data integration to greater results for business. For the purposes of study and analysis, the data may be simply organized into spreadsheets or tables.

1.6. Structure of project

To be able to build a data warehouse to meet the research project, our team divided our project into 5 chapters so that the tasks and research orientation of the group can be transparent.

First, we will present the basic parts of the project: Acknowledgements, commitment, table of content ... and some other information.

Chapter 1, we will talk about the introduction of the project, specifically starting from the business case for the project, finding the right business cases to build the Data Warehouse, besides the main goals and tasks. mentioned in the project to clarify the direction to be studied.

Chapter 2, to supplement the necessary knowledge for building a data warehouse, we will cover basic to advanced theories related to Data Warehouse, schemas, ETL to support project research.

Chapter 3, we will analyze and choose the right modules to build Data Warehouse. From there, approach the analysis of departments in practice, see how it affects the Data Warehouse. Once we have identified the module and the appropriate attributes, we will conduct a data source analysis of data types, attributes, describe, … Finally, there will be Business Requirements Analysis to find answers to the problems that related to.

Chapter 4, after having studied the theory and related business. We will apply them to the practice of building Data Warehouse based on SSIS, SSAS on Visual Studio 2019 platform and SQL Server 18 platform.

Chapter 5, the last part, we will mention the results achieved during the project implementation. The advantages and disadvantages of the project can be drawn. From there, analyze and give a perspective in the future work to contribute to improving the project to be more complete.

CHAPTER 2: THEORETICAL BASIS

2.1 Overview of DWH

2.1.1 What is DWH?

Data warehouse is a central location where consolidated data from multiple locations are stored, not loaded every time when new data is generated. There are timelines determined by the business as to when a Data Warehouse needs to be loaded: Daily, monthly, once in a quarter etc

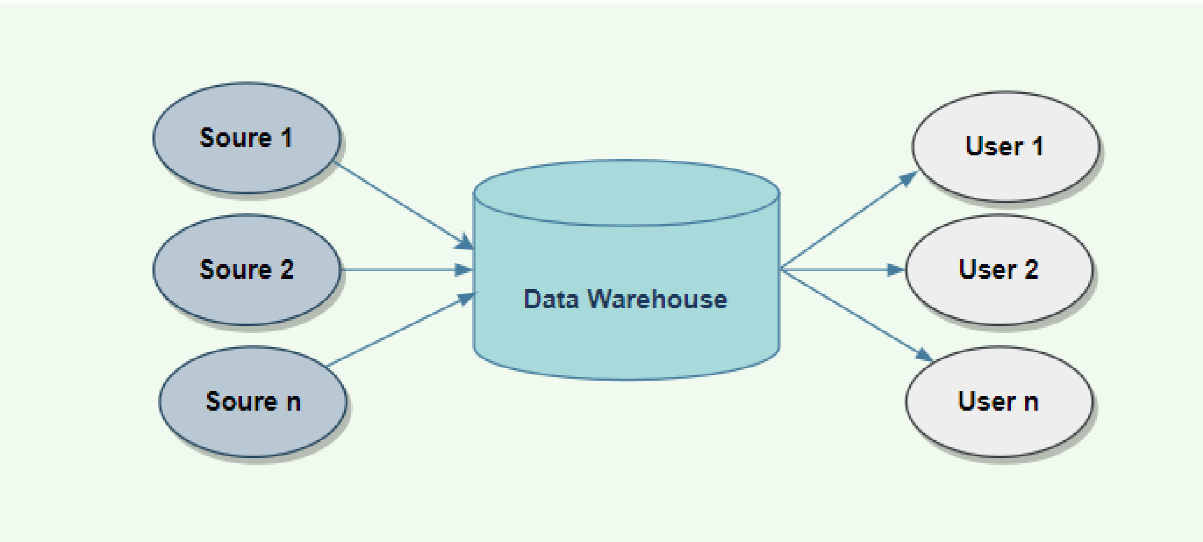


Figure 1: Data Warehouse structure

According to *Inmon*, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. Data warehouses are complicated software environments where data stemming from operational sources are extracted, transformed, cleansed and eventually loaded in fact or dimension tables in the data warehouse. Once this task has been successfully completed, further aggregations of the loaded data are also computed and stored in data marts, reports, spreadsheets, and other formats.

Data warehousing is the process of constructing and using data warehouses. Data warehousing is currently one of the most important applications of database technology in practice. One of the most crucial issues in data warehousing is how to create effective database structures to support end-user queries.

2.1.2 DWH Architecture

A data warehouse is a database which provides a single consistent source of management information for reporting and analysis across the organization (Inmon, 1993; Love, 1994). This is desirable as the main center for storing historical data for a company and is well-suited for reporting and analysis.

Data warehousing is based on a supply chain metaphor. The data "product" is obtained from data "suppliers" (operational systems or external sources) and is temporarily stored in a central data "warehouse". The data is then delivered via data "marts" to data "consumers" (end users). Figure 1 shows a generic architecture for a data warehouse (rectangles indicate data stores, while circles indicate processes).

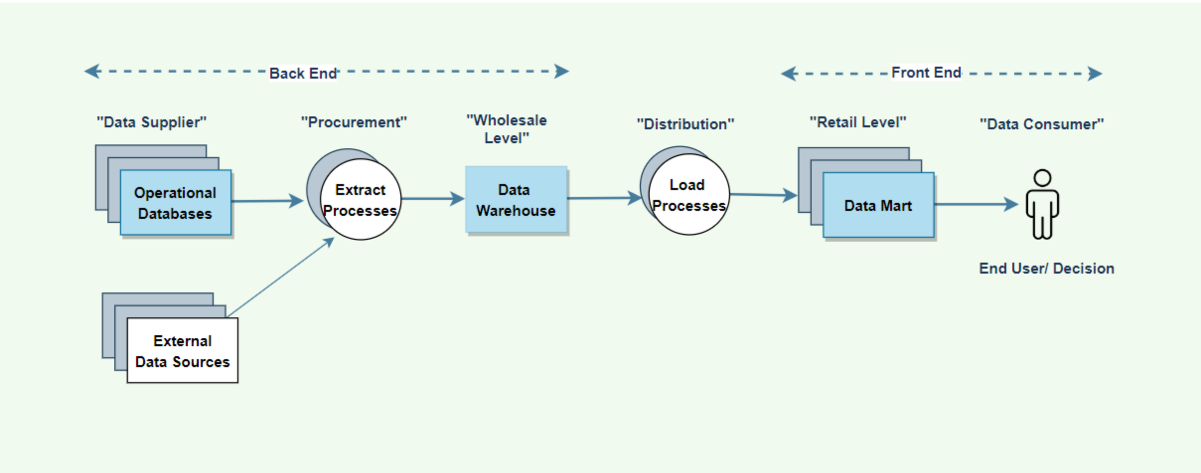


Figure 2: Data warehouse Architecture

The architecture consists of the following components:

- *Operational systems:*

These are systems that keep track of the specifics of business transactions. This is where the majority of the data necessary for decision support is generated.

- *External sources:*

To assist analysis, data warehouses frequently include data from other sources (e.g., census data, economic data).

- *Extract processes:*

On a regular basis (daily, weekly, monthly), these procedures "stock" the data warehouse with data. Data is retrieved from many sources, aggregated and reconciled, then saved in a consistent format. This is equivalent to a procurement function.

- *Central data warehouse:*

This acts as the central source of decision support data across the enterprise. This forms the “wholesale level” of the data warehouse environment and is used to supply data marts. The central data warehouse is usually implemented using a traditional relational DBMS

- *Load processes:*

These mechanisms are responsible for distributing data from the central data warehouse to the data marts. This is equivalent to a distribution function.

- *Data marts:*

These perform as the data warehouse's "retail outlets," offering data in usable form for end users to analyze. Data marts are often designed to meet the requirements of a particular user group or decision-making process. They may be “real” (stored as actual tables populated from the central data warehouse) or virtual (defined as views on the central data warehouse). Utilizing OLAP technologies or conventional relational DBMS, data marts may be constructed.

- *End users:*

Write queries and analyses against data stored in data marts using “user friendly” query tools.

2.1.3 Advantage of BI in enterprises

Business intelligence assists businesses in tracking trends, adapting to changing market conditions, and improving decision making at all levels of the company. The BI tools that a firm employs are determined by its objectives. Some businesses want to learn about customer purchasing habits, while others want to improve staff efficiency or identify top performers. A business intelligence system may be deployed in an endless number of ways. Business Intelligence (BI) software assists businesses in developing new strategies by studying existing market trends and staying ahead of their competition. It enables businesses to capitalize on new prospects and provide comprehensive customer satisfaction based on data-driven reports. It optimizes customer service, boosts productivity, and monitors all data in real time from anywhere.

2.2 Data warehouse and Data mart

2.2.1 What are Data warehouse and Data mart?

- *Data warehouse:*

A data warehouse is a significant central data repository that houses data from several sources inside a business. Through the use of analysis, reporting, and data mining technologies, the collected data is utilized to inform business choices.

Inmon vs. KimBall

Two data warehouse pioneers, Bill Inmon and Ralph Kimball differ in their views on how data warehouses should be designed from the organization's perspective.

- Bill Inmon:

The data warehouse is the primary part of an organization's data systems according to Bill Inmon's top-down strategy, which favors a centralized data repository.

The data warehouse is considered as the physical representation of the comprehensive corporate data model created under the Inmon method. When needed, dimensional data marts pertaining to certain business lines can be produced from the data warehouse.

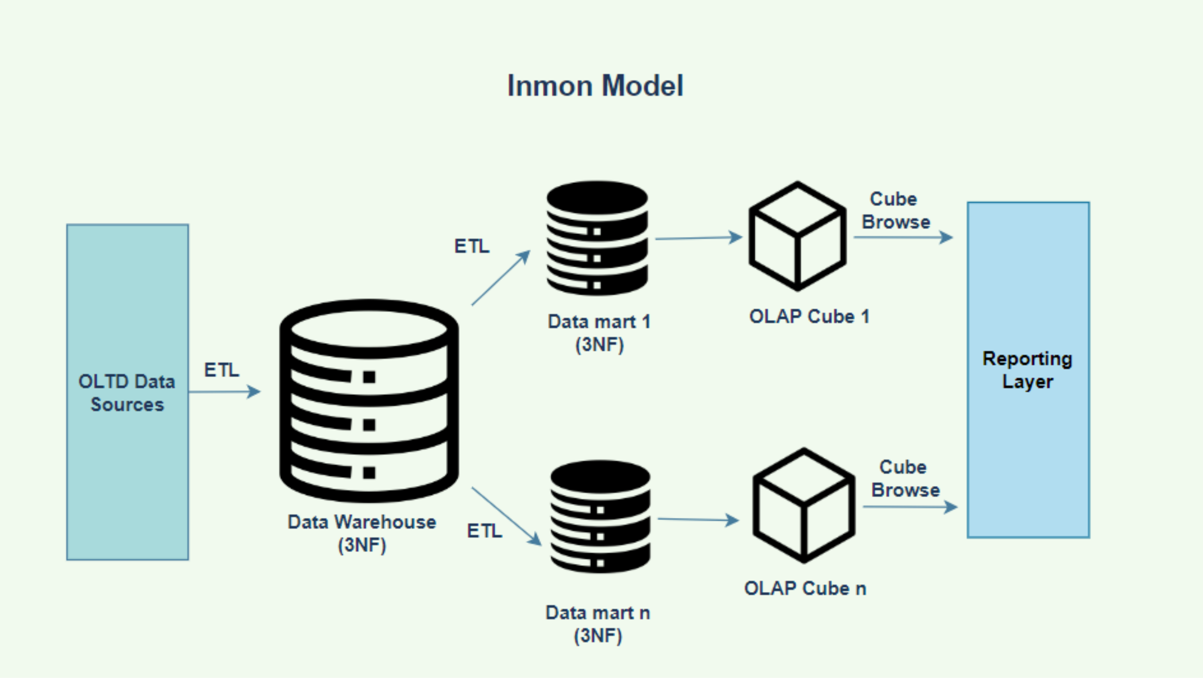


Figure 3: Inmon Model

In the Inmon model, data in the data warehouse is integrated, meaning the data warehouse is the source of the data that ends up in the different data marts. This ensures data integrity and consistency across the organization.

- Ralph Kimball:

In Ralph Kimball's data warehouse architecture, the most crucial business operations come first. With this approach, a firm created data marts that compile essential information on specific topic areas. The organization's several data marts are combined into the data warehouse.

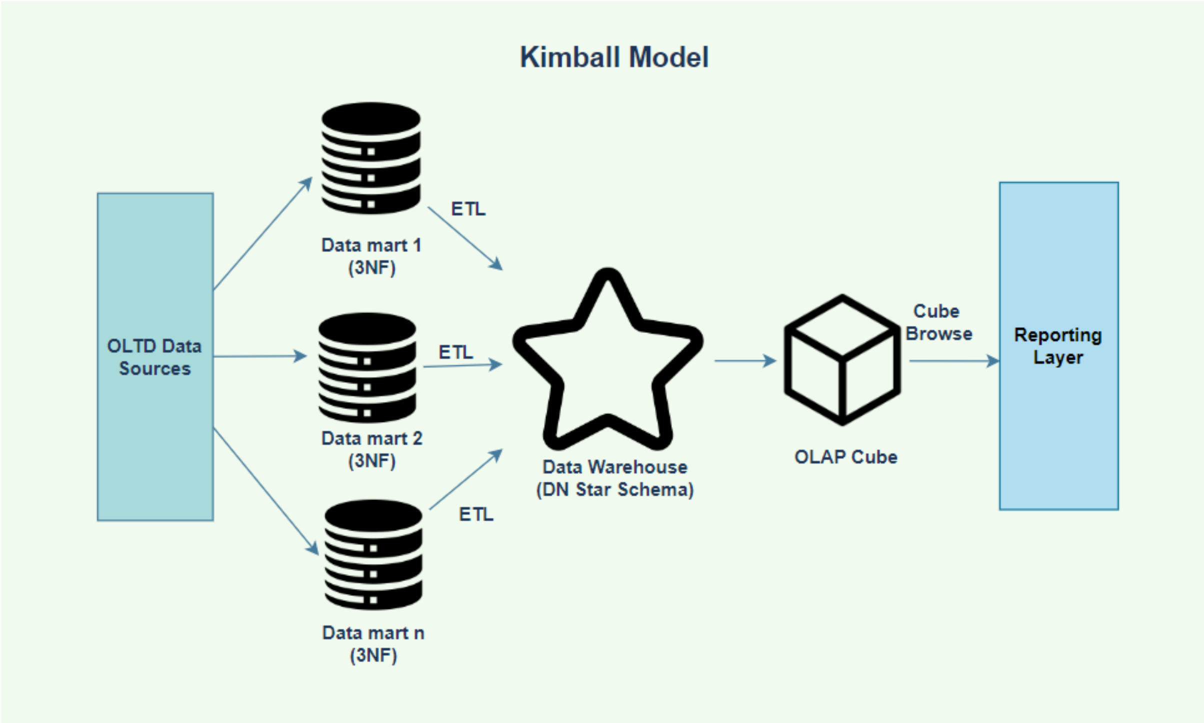


Figure 4: Kimball Model

The Kimball method combines many data marts into a particular data warehouse. As opposed to Inmon's strategy, which builds data marts using information from the warehouse. As Kimball said in 1997, "the data warehouse is nothing more than the union of all data marts."

- *Data mart:*

A data mart is a portion of a data warehouse that is focused on a certain business sector. Data marts are repositories for summarized data that have been gathered for analysis on a separate department or unit inside a company such as the sales department.

2.2.2 Who needs Data warehouse and Data mart?

DWH is needed for all types of users like:

- Those who make decisions based on a lot of data
- Users who collect information from several data sources using complicated, specialized techniques.
- People that prefer straightforward technologies to get info also utilize it.
- It is also crucial for individuals who wish to make judgments in a methodical manner.

- Data warehouses are helpful if the user needs quick performance on a large quantity of data that is required for reports, grids, or charts

Business analysts, other end users, and the BI and data analysts assigned to the business unit may all access and use a data mart easily. Data marts guide important business decisions at a departmental level. For example, a marketing team may use data marts to analyze consumer behaviors, while sales staff could use data marts to compile quarterly sales reports. As these tasks happen within their respective departments, the teams don't need access to all enterprise data.

2.2.3 Advantages and disadvantages of Data warehouse

● *Advantages of DWH):*

- Business users may easily access crucial data from a variety of sources using data warehouses.
- Consistent data on multiple cross-functional operations is provided via data warehouse. Ad hoc reporting and querying are also supported.
- To lessen the strain on the production system, data warehouses assist in integrating several data sources.
- Using a data warehouse can speed up analysis and reporting overall.
- The user may utilize it more easily for reporting and analysis thanks to restructuring and integration.
- Users may obtain crucial data from several sources in a single location thanks to data warehouses. As a result, it saves users' time while obtaining data from various sources.
- A substantial quantity of historical data is kept in data warehouses. Users may use this to evaluate various time periods and patterns to forecast the future.

● *Disadvantages of DWH:*

- An unsuitable choice for unstructured data
- The development and implementation of a data warehouse are undoubtedly time-consuming tasks.
- Data Warehouse can easily become out of date.
- Changes to data types and ranges, data source structure, indexes, and searches are challenging.
- Although the data warehouse may appear simple, most consumers would find it to be excessively complicated.
- The scope of a data warehousing project will constantly expand, even the finest project management efforts.

- Users of warehouses may occasionally create unique business rules.
- Organizations must invest a significant amount of their resources in training and implementation.

2.3 Data Cube in Data Warehouse

A data cube is a multidimensional data model that stores optimized, summarized, or aggregated data, facilitating quick and simple analysis using OLAP tools. Data cube facilitates online analytical processing by storing the precomputed data. We all think of a cube as a three-dimensional structure, however an n-dimensional data cube can be used in data warehousing.

Facts and dimensions are used to represent the data that is stored in a data cube. The entities or attitudes that the business intends to store the data in relation to are the dimensions of the data cube.

The entities or attitudes that the business intends to store the data in relation to are the dimensions of the data cube. A dimension table for each dimension gives more information about that dimension.

A multidimensional data model, such as the data cube, is always built on a concept known as fact. The fact table contains data in numerical forms that indicate numerical measures like the quantity of an item sold, the amount of money sold at a specific branch in a given year, etc. Data cube knowledge enabled us to proceed to data cube classification.

2.4 Star Schemas and Snowflake

Dimensional Modeling: is a data structure method designed for data warehouse data storage. Dimensional modeling is used to enhance databases for quicker data retrieval. The "fact" and "dimension" tables that make up the Dimensional Modeling idea were created by Ralph Kimball.

Star Schemas (SS):

The basic building block used in dimensional modeling is the star schema. A star schema consists of one large central table called the fact table, and a number of smaller tables called dimension tables which radiate out from the central table

There is a one-to-many relationship between each dimension table and the fact table. The fact table usually represents business transactions or events, or a snapshot summary of the transactions/events

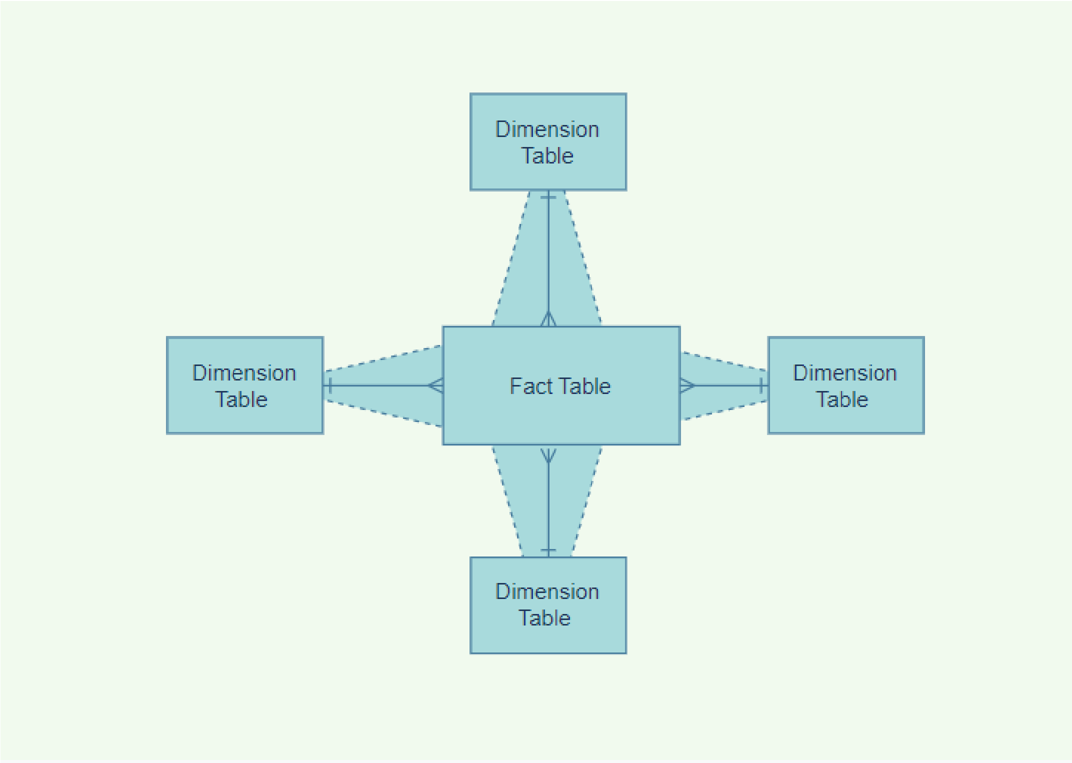


Figure 5: Star Schema

Snowflake (SF):

The snowflake schema is an extension of the star schema. In a snowflake schema, each dimension is normalized and connected to more dimension tables.

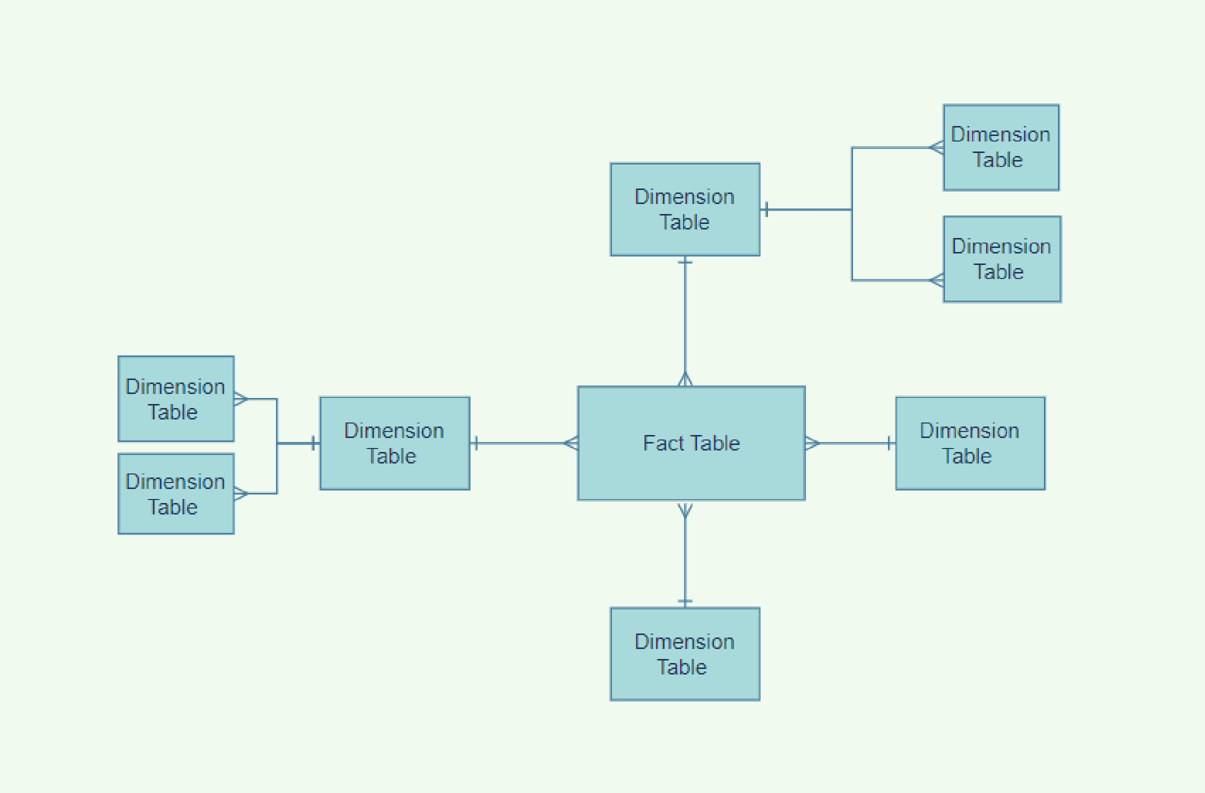


Figure 6: Snowflake Schema

Galaxy Schema: (GS)

The Galaxy Data Warehouse Schema, also known as a Fact Constellation Schema, is the data warehouse schema's next generation. The Galaxy Schema, unlike the Star and Snowflake Schemas, employs numerous fact tables linked by shared normalized dimension tables. Galaxy Schema is a collection of interconnected and properly normalized star schema that avoids data redundancy and inconsistencies.

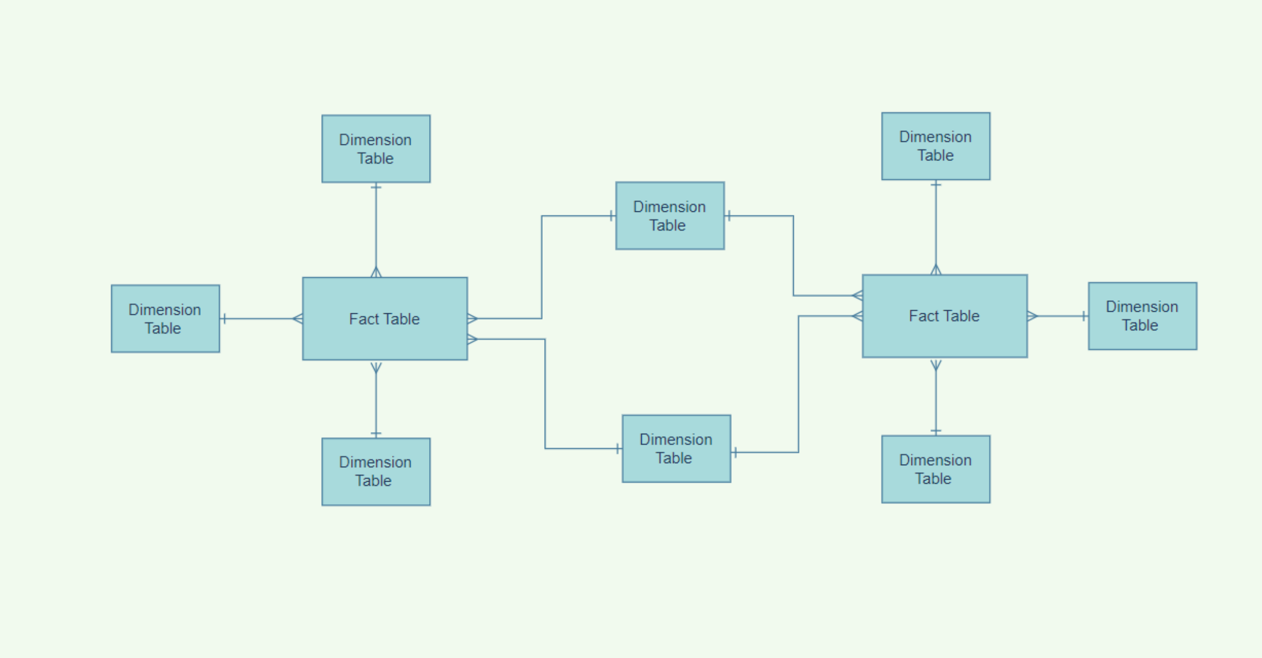


Figure 7: Galaxy Schema

2.5 *Integration and ETL Process*

2.5.1 *What is ETL?*

In data warehouse development, the extraction-transformation-loading (ETL) process extracts data from diverse sources, transforms it into an appropriate format, and loads it into data warehouse storage.

As databases became more prominent in the 1970s, ETL was developed as a mechanism for integrating and loading data for calculation and analysis, eventually becoming the predominant method for data processing in data warehousing operations.

The cornerstone for data analytics and machine learning workstreams is ETL. ETL cleanses and organizes data using a set of business rules to fulfill particular business intelligence needs, such as monthly reporting, but it may also handle more complex analytics to improve back-end operations or end user experiences. An organization will frequently employ ETL to:

- Extract data from legacy systems
- Cleanse the data to increase its quality and consistency.
- Load data into the specified database.

2.5.2 *Why do we need ETL?*

Legacy systems with conventional databases cannot answer complicated business queries, only ETL can do so.

- It enables organizations to recover historical data that may be used to provide context and an in-depth understanding of the firm through time.
- It improves and integrates corporate intelligence solutions for decision-making.
- It generates relevant patterns and insights and converts disparate data into a uniform structure.
- It incorporates data collected from external suppliers, partners, and recent company mergers and acquisitions.
- It provides for the creation of a centralized hub or data repository for easier access, as well as sample data comparison between the source and destination systems.
- It increases efficiency by codifying and reusing without requiring extra technical expertise. Instead of devoting time in developing specialized tools for analysis, developers and analysts may now devote time to company analysis and growth.

2.5.3 *ETL Process*

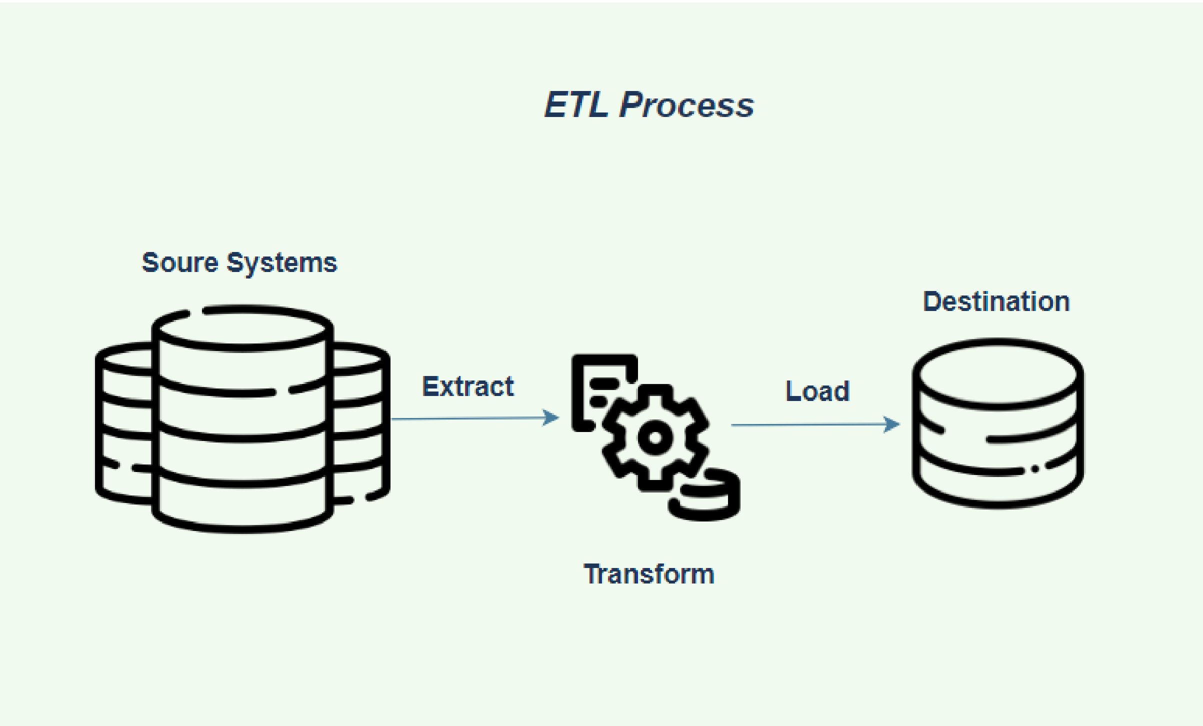


Figure 8: The ETL process

- *Extraction:*

Raw data is transferred or exported from source locations to a staging area during data extraction. Data management teams can extract data from a range of structured and unstructured data sources. Among these sources include, but are not limited to:

- SQL or NoSQL servers
- CRM and ERP systems
- Flat files
- Email
- Web pages

Extraction identifies all relevant sources and extracts data as effectively as feasible. This procedure involves looking through a file or database, picking data using various criteria, locating appropriate data, and then transporting the data into a file or another database.

- *Transformation:*

Transformation is the process of converting data from source system to another in a data warehouse or data mart in order to make it relevant.

The raw data undergoes data processing in the staging area. The data is converted and consolidated in this step for its intended analytical use case. This stage may include the following tasks:

- Filtering, cleansing, de-duplicating, validating, and authenticating the data.
- Performing calculations, translations, or summarizations based on the raw data. Changing row and column headings for uniformity, converting currencies or other units of measurement, modifying text strings, and more are all examples of this.
- Conducting audits to ensure data quality and compliance
- Data removal, encryption, or protection mandated by business or government regulators
- Formatting the data into tables or joined tables to match the schema of the target data warehouse.

- *Loading:*

The final ETL step is to load data into the target multidimensional structure. In this phase, the extracted and converted data is written into the dimensional structures that end users and application systems will actually access. Loading dimension tables and loading fact tables are both included in the loading stage

CHAPTER 3: REQUIREMENTS ANALYTICS

3.1. Business processes (Purchasing, Production, Sales, or HR,…)

3.1.1 Purchasing & Production department

The transformation of raw materials and other inputs into completed products or services is the responsibility of the production department. Between production steps, the department works to increase the production line's effectiveness so it can reach output goals established by the business management and guarantee that the finished goods provide customers with the best value and quality.

The component of a company that is in charge of acquiring the products and services that the company needs to operate, which is known as the purchasing department. The terms procurement departments and buying departments are sometimes used in business. These divisions frequently play a significant role in assisting businesses in achieving both short- and long-term strategic objectives. Depending on the size of the business, a department may be given more or less authority, but in order to maintain the effectiveness of the business' operations, they frequently assist in managing vendor contracts and supply chain monitoring.

3.2 Data source and challenges

This is a large-scale multinational manufacturing company. The company manufactures and sells bicycles made of metal and synthetic materials. The company's market includes North America, Europe and Asia. While the company's headquarters are located in Bothell, Washington.

In On-line transactional processing (OLTP) AdventureWorks illustrates, objects such as tables, views and procedures are organized into schemas. We use 2 main module in this project :

Production: Products that are manufactured and sold by Adventure Works Cycles. Total tables: 25

Purchasing: Vendor or supplements and other products that company brought. Total table: 5.

Data source:

Key	Name	Data type	Null	Attributes	Description
Primary key	LocationID	smallint		Identity	Primary key for Location records.
	Name	nvarchar(50)			Location description.
	CostRate	smallmoney		Default: 0.00	Standard hourly cost of the manufacturing location.
	Availability	decimal(8, 2)		Default: 0.00	Work capacity (in hours) of the manufacturing location.
	ModifiedDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 1: Production.Location

Key	Name	Data type	Null	Attributes	Description
Primary key	ProductID	int		Identity	Primary key for Product records.
	Name	nvarchar(50)			Name of the product.
	Product Number	nvarchar(25)			Unique product identification number.
	MakeFlag	bit		Default: 1	0 = Product is purchased, 1 =

					Product is manufactured in-house.
	Finished GoodsFlag	bit		Default: 1	0 = Product is not a salable item. 1 = Product is salable.
	Color	nvarchar(15)	x		Product Color.
	Size	nvarchar(5)	x		Product size.
	SizeUnit MeasureCode	nchar(3)	x		Unit of measure for Size column.
	WeightUnit MeasureCode	nchar(3)	x		Unit of measure for Weight column.
	Weight	decimal(8,2)	x		Product weight.
	Class	nchar(2)	x		H = High, M = Medium, L = Low
	Style	nchar(2)	x		W = Womens, M = Mens, U = Universal
Foreign Key	ProductSubcategoryID	int			Product is a member of this product subcategory. Foreign key to ProductSubCategory.ProductSubCategoryID.
	ModifiedDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 2: Production.Product

Key	Name	Data type	Null	Attributes	Description
Primary key	ProductCategoryID	int		Identity	Primary key for ProductCategory records.
	Name	nvarchar (50)			Category description.
	ModifiedDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 3: Production.ProductCategory

Key	Name	Data type	Null	Attributes	Description
Primary key	ProductSubcategoryID	int		Identity	Primary key for ProductSubcategory records.
Foreign key	Product CategoryID	int			Product category identification number. Foreign key to ProductCategory.ProductCategoryID.
	Name	nvarchar (50)			Subcategory description.
	ModifiedDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 4: Production.ProductSubCategory

Key	Name	Data type	Null	Attributes	Description
Primary key	TransactionID	int		Identity	Primary key for TransactionHistory records.
Foreign key	ProductID	int			Product identification number. Foreign key to Product.ProductID.
	TransactionDate	datetime		Default: getdate()	Date and time of the transaction.
	Quantity	nvarchar(50)			Product quantity.
	ActualCost	money			Product Cost.
	ModifiedDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 5: Production.TransactionHistory

Key	Name	Data type	Null	Attributes	Description
Primary key	WorkOrderID	int			Primary key. Foreign key to WorkOrder.WorkOrderID.
Primary Key	ProductID	int			Primary key. Foreign key to Product.ProductID.
Primary	Operation	smallint			Primary key. Indicates

Key	Sequence				the manufacturing process sequence.
	LocationID	smallint			Manufacturing location where the part is processed. Foreign key to Location.LocationID.
	ActualStartDate	datetime	x		Actual start date.
	ActualEndDate	datetime	x		Actual end date.
	ActualResourceHrs	decimal(9,4)	x		Number of manufacturing hours used.
	Actual Cost	money	x		Actual manufacturing cost.
	ModifieDate	datetime		Default: getdate()	Date and time the record was last updated.

Table 6: Production.WorkOrder

Key	**Name**	**Data type**	**Null**	**Attributes**	**Description**
Primary key	ProductID	int			Primary key. Foreign key to Product.ProductID.
Primary Key	BusinessEntityID	int			Primary key. Foreign key to Vendor.BusinessEntityID.
	LastReceiptCost	money			The selling price when last purchased

	MinOrderQty	int	x		The maximum quantity that should be ordered.
	MaxOrderQty	int	x		The minimum quantity that should be ordered.

Table 7: Purchasing.ProductVendor

Key	Name	Data type	Null	Attributes	Description
Primary key	PurchaseOrderID	int			Primary key. Foreign key to PurchaseOrderHeader.PurchaseOrderID
	OrderQty	smallint			Quantity ordered..
	ProductID	int			Product identification number. Foreign key to Product.ProductID
	UnitPrice	money			Vendor's selling price of a single product
	LineTotal	money			Per product subtotal. Computed as OrderQty * UnitPrice. Computed: isnull([OrderQty]*[UnitPrice],(0.00))
	ReceivedQty	decimal(8, 2)			Quantity actually received from the vendor
	StockedQty	decimal(9, 2)			Date and time the record was last updated. Default: getdate()

Table 8: Purchasing.PurchaseOrderDetail

Key	Name	Data type	Null	Attributes	Description
Primary key	ShipMethodID	int		Identity/ Auto increment column	Primary key for ShipMethod records.
	Name	nvarchar(50)			Shipping company name
	ShipBase	money		Default: 0.00	Minimum shipping charge
	ShipRate	money		Default: 0.00	Shipping charge per pound.

Table 9: Purchasing.ShipMethod

Key	Name	Data type	Null	Attributes	Description
Primary key	PurchaseOrderID	int		Identity/Auto increment column	Primary key
	VendorID	int			Vendor with whom the purchase order is placed. Foreign key to Vendor.BusinessEntityID

					.
	ShipMethodID	int			Shipping method. Foreign key to ShipMethod.ShipMethodID.
	Freight	money		Default: 0.00	Shipping cost.

Table 10: Purchasing.PurchaseOrderHeader

Key	Name	Data type	Null	Attributes	Description
Primary key	BusinessEntityID	int			Primary key for Vendor records. Foreign key to BusinessEntity.BusinessEntityID
	AccountNumber	nvarchar(15)			Vendor account (identification) number.
	Name	nvarchar(50)			Company name.
	CreditRating	tinyint			1 = Superior, 2 = Excellent, 3 = Above average, 4 = Average, 5 = Below average
	PreferredVendorStatus	bit		Default: 1	0 = Do not use if another vendor is available. 1 = Preferred over other vendors supplying the same product

	ActiveFlag	bit		Default: 1	0 = Vendor no longer used. 1 = Vendor is actively used.

Table 11: Purchasing.Vendor

Challenges:

The number of tables in a schema (which can be considered as modules) is quite large, and the name of the table can be misleading in choosing fact and dim tables. Our goal is to be able to produce a quantity report in the manufacturing process and then there is another transactional fact that shows costs. But in the data source, the clear separation of the tables for the analysis that the team wanted was not evident.

There are lots of other dimensions that we've found out in the source system, such as the Supplier, Delivery Method, Delivery Vessel, Departure Port, Delivery Port, the PO, the Receipt Number etc. This makes it difficult to determine the elements needed for daily counting.

Furthermore, the large number of transactions and various manufacturing locations, this leads to a large number of rows in our fact table.

3.3. Business Requirements Analysis

E-commerce is forcing companies to become more agile every day. This reflects the need for data-driven storage to meet the ever-expanding needs of our customers. Increased revenue and fierce competition will make it necessary to run warehouses at optimum utilization. Warehouse reporting plays an important role in supporting the supply chain. Above all, learning from past actions, in order to improve problems in the present and in the future, is an evolutionary trait that also applies to warehouses. Warehouse reports support both operations and improvisation. Generally, the inventory management system generates these reports unless they are prepared manually.

3.3.1. *Make a report on the number of products produced, damaged products in a day*

While manual counting is essential, it is done periodically for obvious reasons. That data cannot be used to fulfill orders. Therefore, you will need an inventory report available that gives you real-time information. If you don't have it, significant failures like oversold and oversold will occur. They have an extremely detrimental effect on your reputation in the market. These reports serve as actionable insights into commodity prevalence and damage. Responding to seasonal demand and product bundles also requires an estimate of current inventory levels. What I find most important in terms of inventory management is that it helps with decisions regarding inventory liquidation.

3.3.2. *Make reports on costs incurred in the production department*

Calculating product value is not a simple matter of an enterprise. This requires an accurate collection of costs incurred in the business. If one cost source is wrong, this leads to a series of errors later. Therefore, it is necessary to closely manage these sources of costs on a regular basis. Thus, in the production process, the costs incurred need to be accurately calculated and managed to avoid errors.

3.3.3. *Make a report on the quantity of purchased goods and check the difference with the actual amount of goods in stock*

Although it is a manufacturing company, purchasing operations still play an important role in the business. However, buying and selling is not always smooth, there are products damaged in transit or substandard upon receipt. Therefore, it is necessary to check the quantity of goods purchased when entering the warehouse in reality compared to that on paper. This also applies to product management in manufacturing.

CHAPTER 4: BUILDING DATA WAREHOUSE AND INTEGRATING DATA

4.1. Designing Data Warehouse

4.1.1. Bus Matrix

COMMON DIMENSIONS

BUSINESS PROCESSES	Date	Production.Product	Production.ProductCategory	Production.ProductSubCategory	Production.Location	Production.TransactionHistory	Production.WorkOder	Purchasing.Vendor	Purchasing.Product	Purchasing.ShipMethod
Number of products in stock	x	x								
Classification of products in stock	x	x	x	x						
Rejected product	x	x	x	x	x					
Warehouse location	x				x					
Transactions made in date	x	x				x	x			
Number of transaction	x					x				
Checking product	x							x	x	
Ship method	x								x	x

Figure 9: Bus matrix

4.1.2 Master Data

Master data represents the actual, critical business objects upon which said transactions are performed, also taking into account the parameters on which data analysis is conducted.

Application for Production and Purchasing module, master data involve:

- Parties: Vendor
- Products: Products traded among the parties
- Locational concepts: office Location

4.1.3 Transaction Data

Transactional data is information obtained through transactions. It captures the time of the transaction, the location of the transaction, the prices of the things purchased, the payment method used, any discounts, and other quantities and qualities related with the transaction. Typically, transactional data is collected at the point of sale. Transaction data is presented in this study:

- DimLocation has CostRate column.

- DimTransaction has ActualCost column.
- DimWorkOrder has ActualResourceHrs column.

4.1.4. Fact and dimension tables

4.1.4.1. Module Production

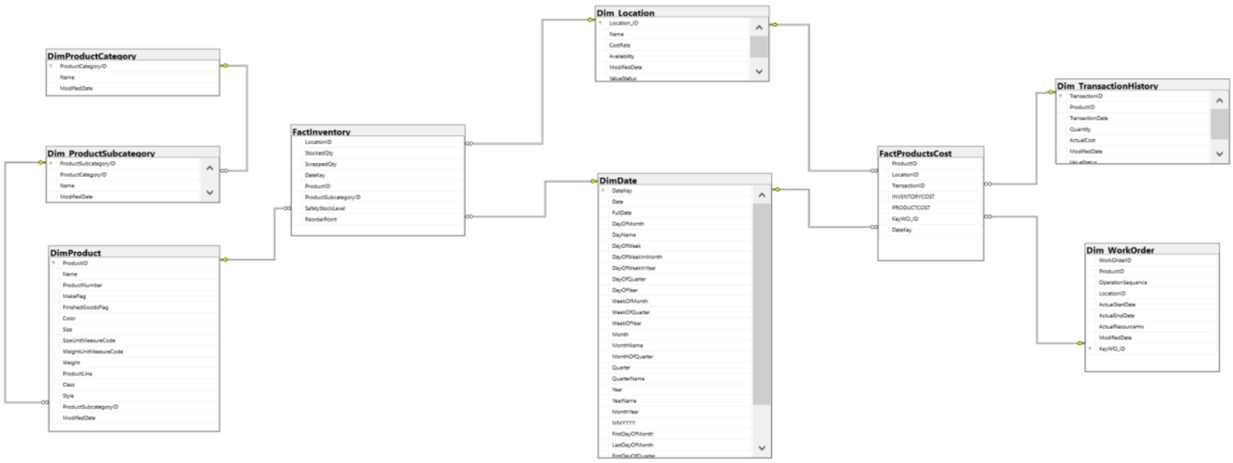


Figure 10: Galaxy schema (Module Production)

This schema include two fact table and seven dimensions:

FactInventory – This table contains reference to dimension tables and 5 additional key of information: Optimal Quantity (appropriate storage level for 1 product), Quantity On Hand (the number of product is in inventory in real time), Quantity on order (quantity that failed inspection), Failed Quantity và Reorder Threshold (inventory level that triggers a purchase order or work order). There are 3 foreign keys concurrent with the primary keys of the table.

Dim_product – This table stores basic information and status of product (In process; goods; sold).

Dim_Date – This table stores time. The lowest level is Date. It includes features as illustrated above. We use ModifiedDate which was set in the source data tables to reference time in data.

Dim_Location – Each inventory of the company will record the name, productivity of the manufacturing and their costs.

Dim_ProductCategory và Dim_ProductSubcategory: Show name and code of Category and Subcategory.

In the fact table name Fact Inventory, we can retrieve information about product quantity in inventory in detail. We can know the location of inventory and its address such as name and product stored in that inventory. We have the time of data recorded from a dim date. This fact table includes product category ID, which is used for retrieved category detail in other dimensions. Transactions will be divided into each record based on ProductID, for example: 1 Customer Order will have many different product codes, the more product codes there will be more trans records.

From Dim Product, we can get information about products including color, size, weights,… and category. Category of product is extracted from dimension product category, and each category has sub categories. This shows more details of the product.

From Dim Location, we have standard cost of manufacturing in each inventory corresponding to each location ID. This can be used in another fact table.

Modified date presented the time that information was recorded.

In this module, we also want to extract the cost of manufacturing products. It is shown in Fact Products Cost. Inventory Cost (factory cost) will be calculated by actual manufacturing cost in the Work Order Routing table multiple actual resource hours. Product Cost will be calculated by dividing the actual cost with quantity in table Transaction History.

Information about order is presented in both dimensions: transaction history and work order. Work order show details of manufacturing of each manufacturing ordering process.

4.1.4.2. Module Purchasing

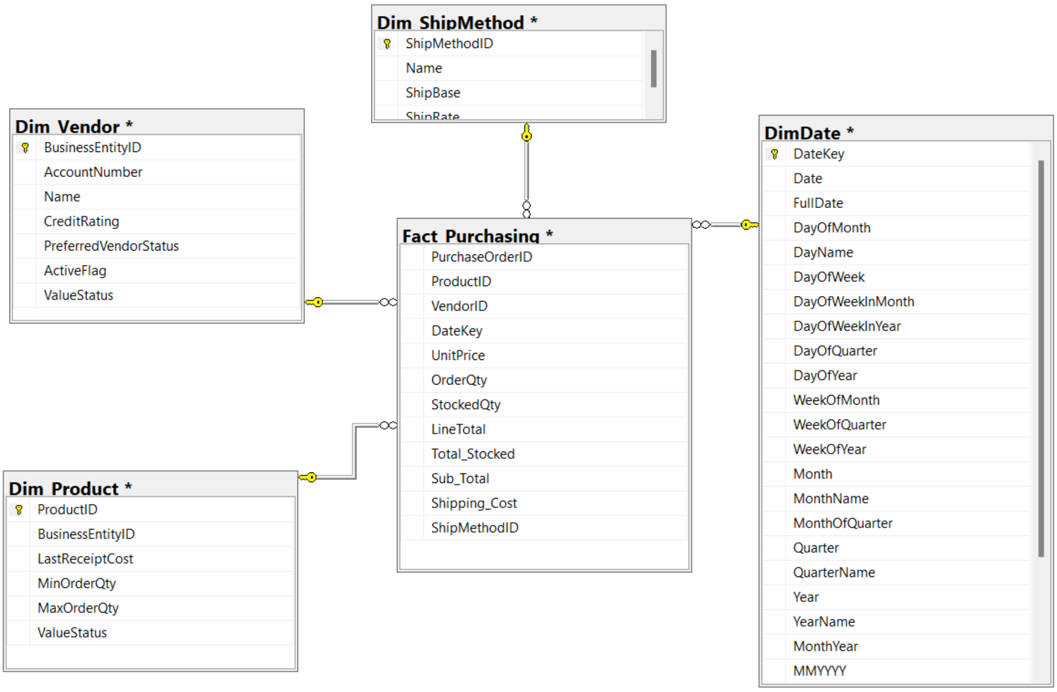


Figure 11: Star schema (Module Purchasing)

After selecting dimensions and the fact table, we represent it in figure 11. A fact name Fact_Purchasing includes all information from 4 dimensions, which show ID of purchase order, vendor and ship method. This information can be retrieved from the primary key in dimension tables. It also shows the time of the purchase order when this company purchases supplements from their vendors. From these orders, we can calculate and show the quantity of product which was ordered, the products which were in stock and total of products. We also retrieve information of the shipping stage, including shipping cost for audit.

This module uses the same Dim Date as module Product.

From Dim Vendor, we can know information about that vendor. It includes vendor ID, which is a number unique for each vendor. Each vendor is recorded with account number, name, credit rating, preferred vendor status and active flag. From this Dim, we selected only information that is useful for creating fact table and for users.

These dimensions have a column that shows value status, indicating current or expired information. When an information is changed, that row will be expired and new information will be set as current status.

From Dim Product, we can know about information when purchasing products. In this dim, business entity ID is known as vendor ID. With each product, there is a limitation when ordering it. This is shown in Max Order Quantity and Min Order Quantity, a range of quantities of products for ordering.

From the Shipping Method dimension, we can calculate the shipping cost in detail. This dimension presents the cost per pound and the minimum shipping charge that we have to pay.

4.1.5. Data Warehouse model (Snowflake or Star)

The model shown in Figure 10 contains two fact tables and seven dimensions tables. We can look at the above model as 2 data marts, 1 is for warehouse management, 2 is for the accounting department to collect costs incurred in the production module. Each data mart consists of only one fact table and several dimension tables. Specifically, as shown below:

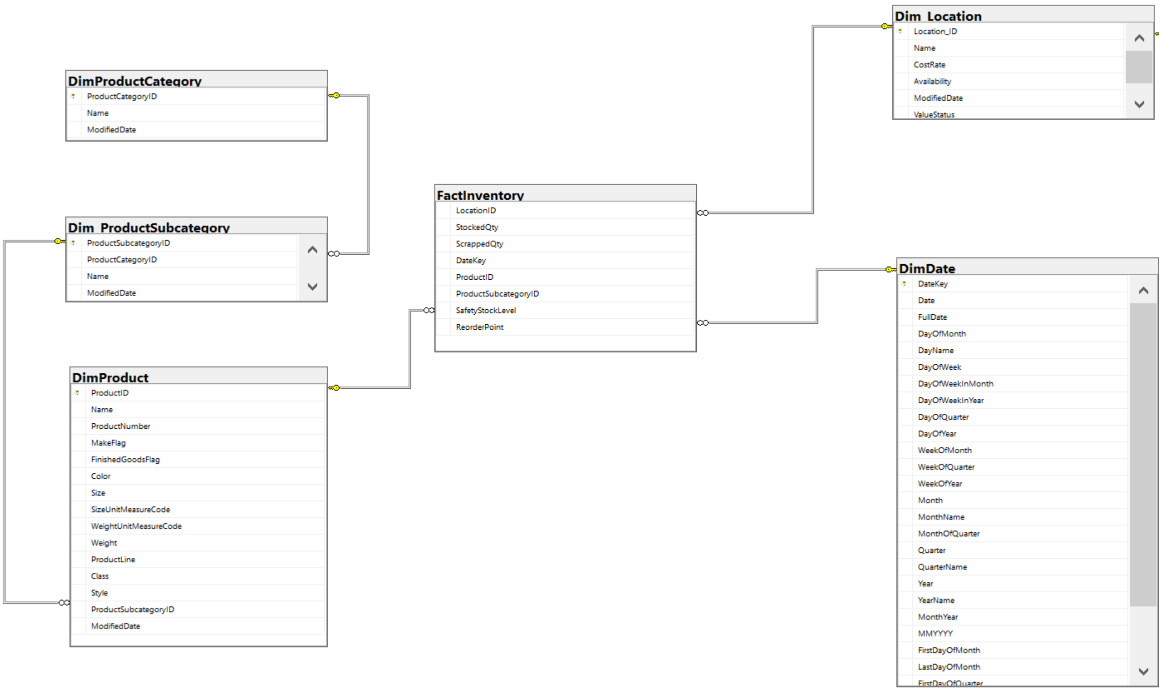


Figure 12: Data mart Inventory

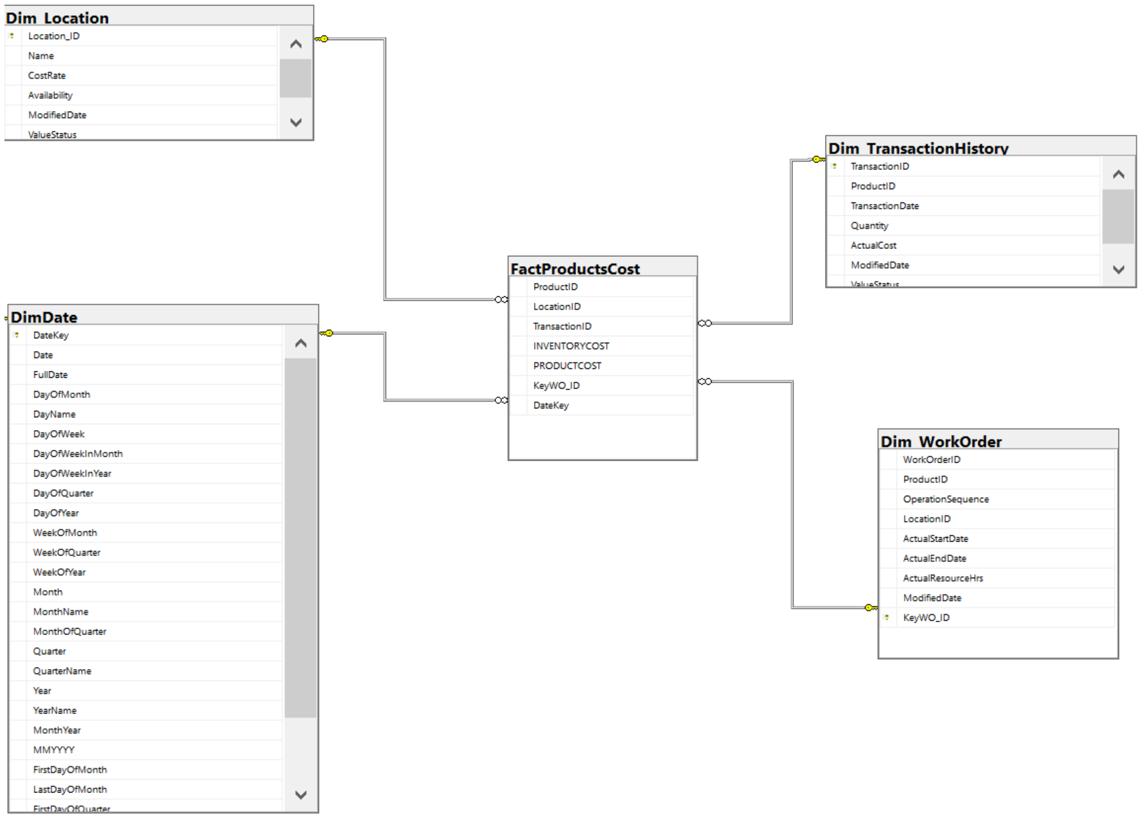


Figure 13: Data mart Production cost

We have chosen a design that combines 2 data marts into one and is called the galaxy model (galaxy schema) as introduced in chapter 2. Sharing dimension tables can reduce the size of the database data especially when sharing dimension tables with multiple values. The idea is that the dimension tables in both data marts share a common approach. So there are attributes in the Dim_Location table that, although not used for both data marts at the same time, are still included in the table to match both.

Besides, another reason that we chose this design is because the storage capacity of the data in this module is quite large. Creating a fact table requires joining (join on) data many times, so the implementation will become difficult and slow, which is not suitable for star schema.

In contrast to the Production module, we design the model so that the Purchasing module follows the purpose of the proposed project. As shown in Figure 11, the model has 1 fact table and 4 dimensions tables. There are many strengths of the star pattern. The fact table is related to each dimension table by a relation, and we don't need anything from additional points to describe the dimension tables. Simplify the query and reduce execution time. This makes it possible to quickly execute reports directly from the OLTP system. For example, we can quickly query the number of products the company has ordered on the invoice, the actual quantity of products in stock and the difference in the amount between them:

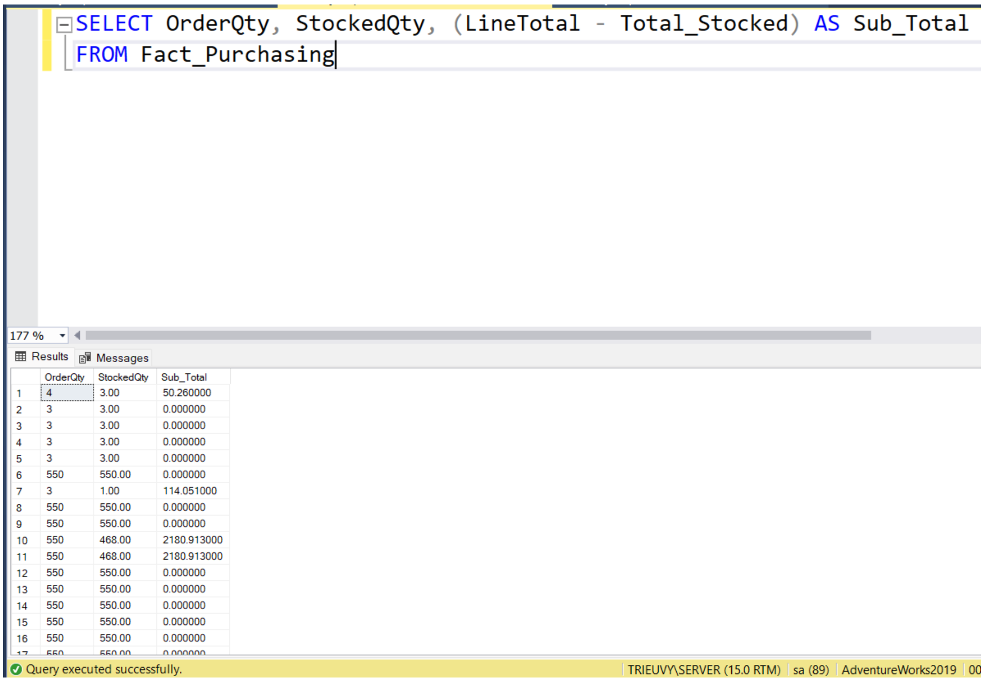


Figure 14: Query

However, the weakness of the star model is data redundancy. Each dimension table is stored separately, and this is the cause of denormalization. This means that we will consume a lot of disk space and there is a risk of data integrity.

4.2. ETL processes

4.2.1. Dimension Table

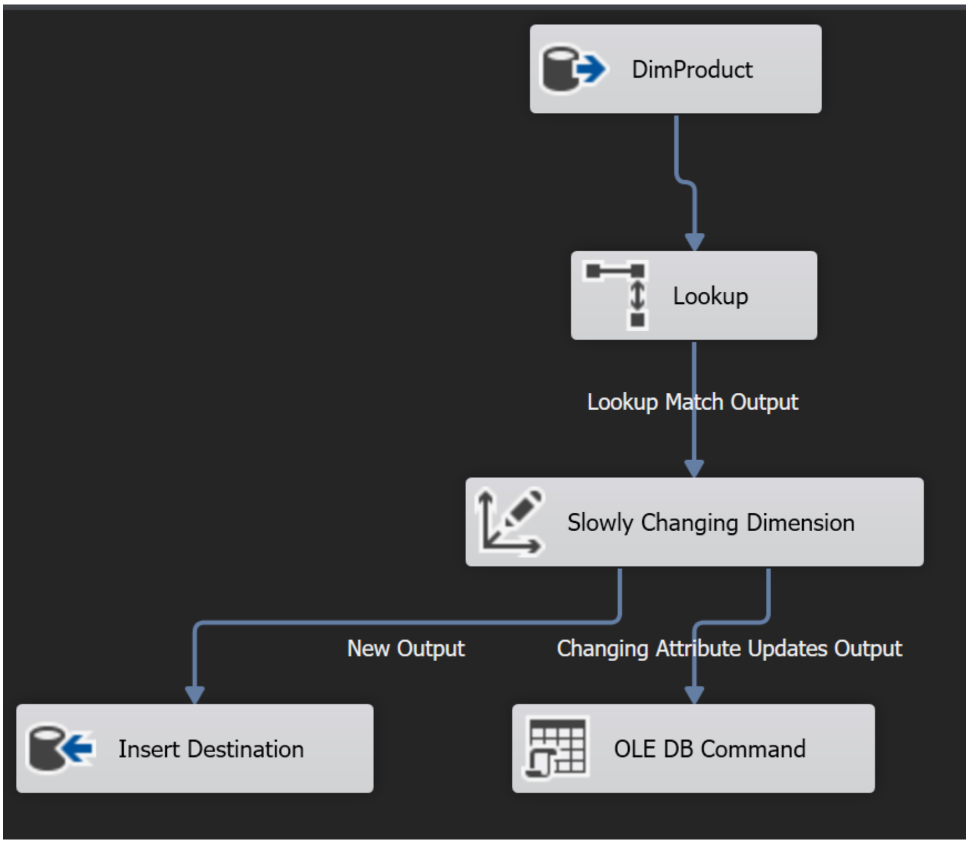


Figure 15: SSIS Dim table

Before the data source pours into the data warehouse, the data will have to go through the ETL process. In this project, we use two tools: Lookup and Slowly Changing Dimension to serve the ETL process. Lookup works to filter out the right data and avoid reloading duplicate data. Slowly Changing Dimension aims to update the Data warehouse according to the settings if the data from the datasource has any changes compared to the data from the Data warehouse.

4.2.2. Fact Table

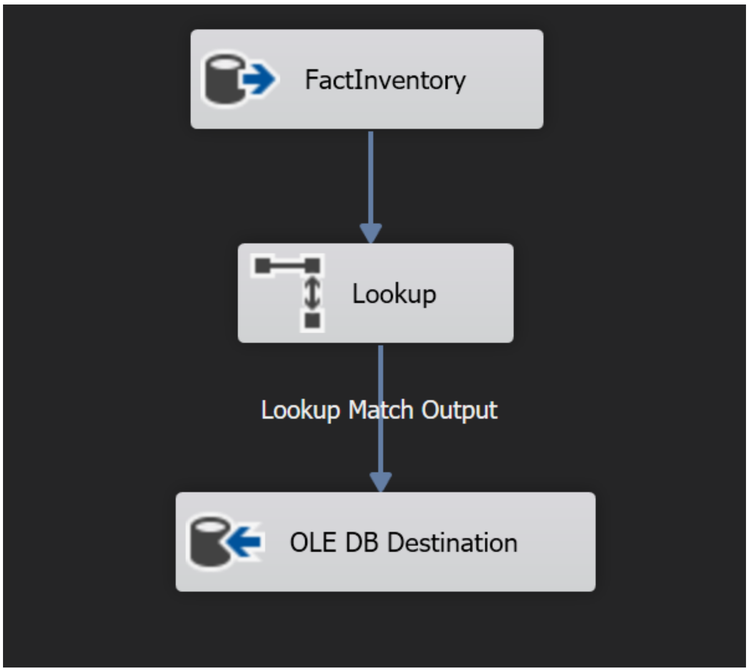


Figure 16: SSIS Fact table

Similarly, before pouring data into the Fact table, data will have to go through the ETL process. In this table, we will use Lookup to filter out matching data and avoid reloading duplicate data. We do not use Slowly Changing Dimension here because the measures in the fact table use mainly calculations from data, so there is little change. Moreover, according to hands-on research with data warehouses, we found that when using Slowly Changing Dimension for data dump into fact tables, some problems are likely to occur during data dump. Therefore, we decided not to use Slowly Changing Dimension.

CHAPTER 5: CONCLUSION AND FUTURE WORKS

5.1. Results

5.1.1. Create Cube

Cube allows to merge or aggregate related data and then drill down, slicing and dicing, or pivoting data to view it from different angles. It is essentially a container to store the data needed for the report. Therefore, the data structure in the cube will be optimized for quick analysis and more efficient report generation.

The special feature of cube is that users do not have to perform too complicated interactions to create reports, users only need to drag and drop the appropriate data to create a report that suits their requirements.

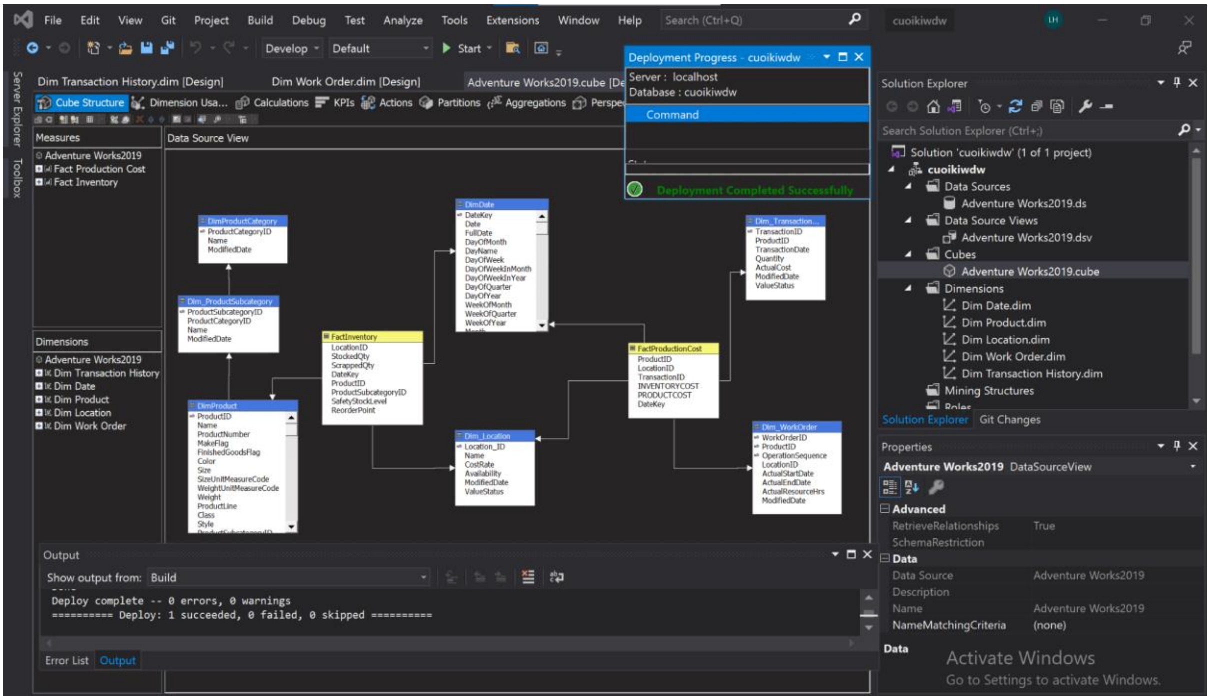


Figure 17: Cube Production

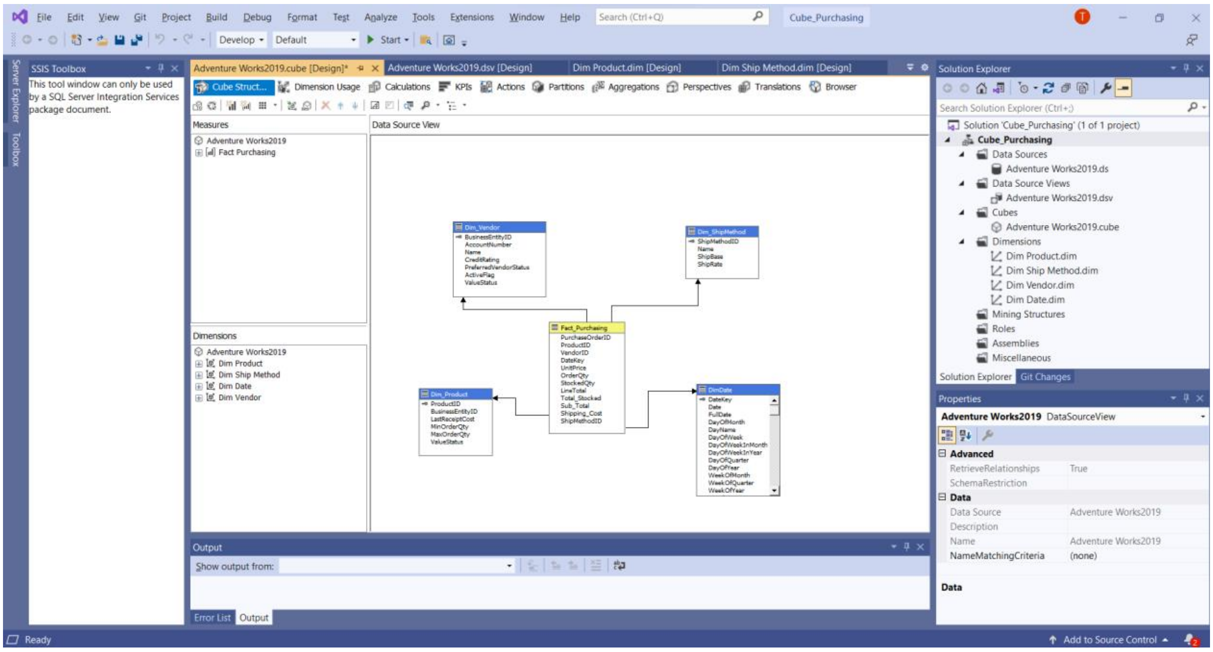


Figure 18: Cube Purchasing

5.1.2 Deploy cube to server

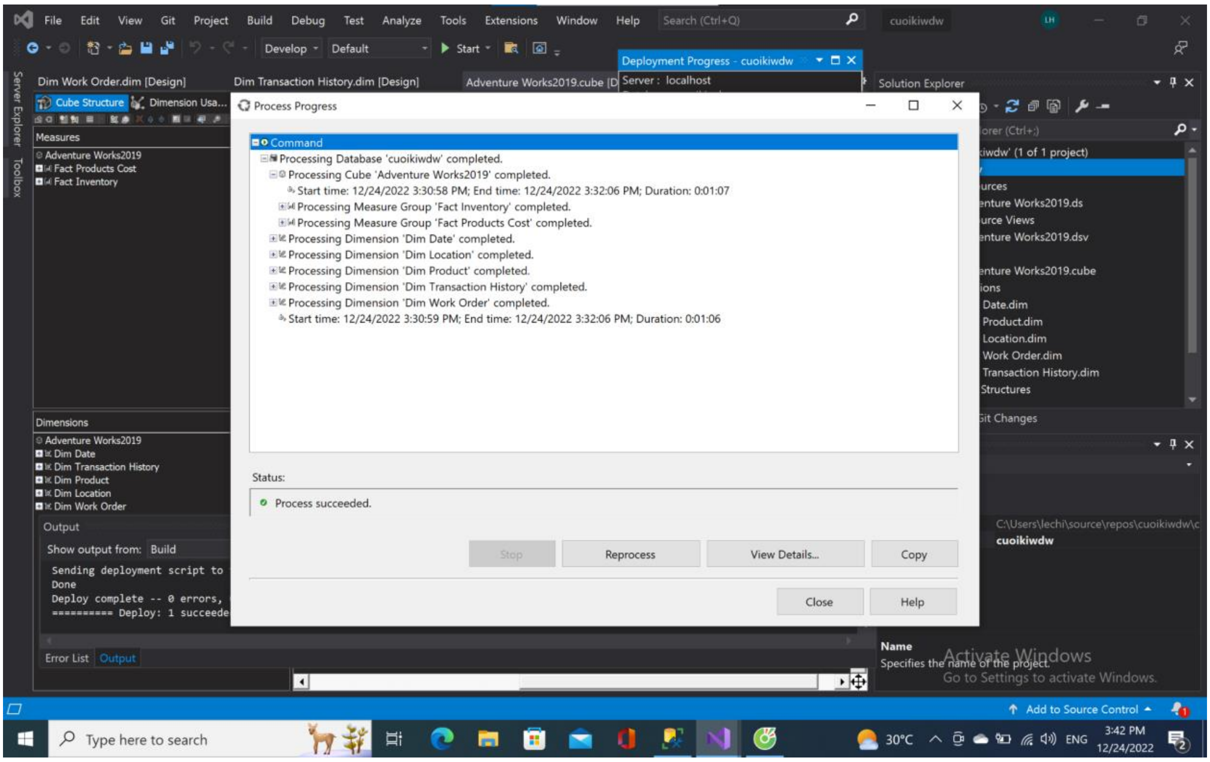


Figure 19: Deploy successful on Production module

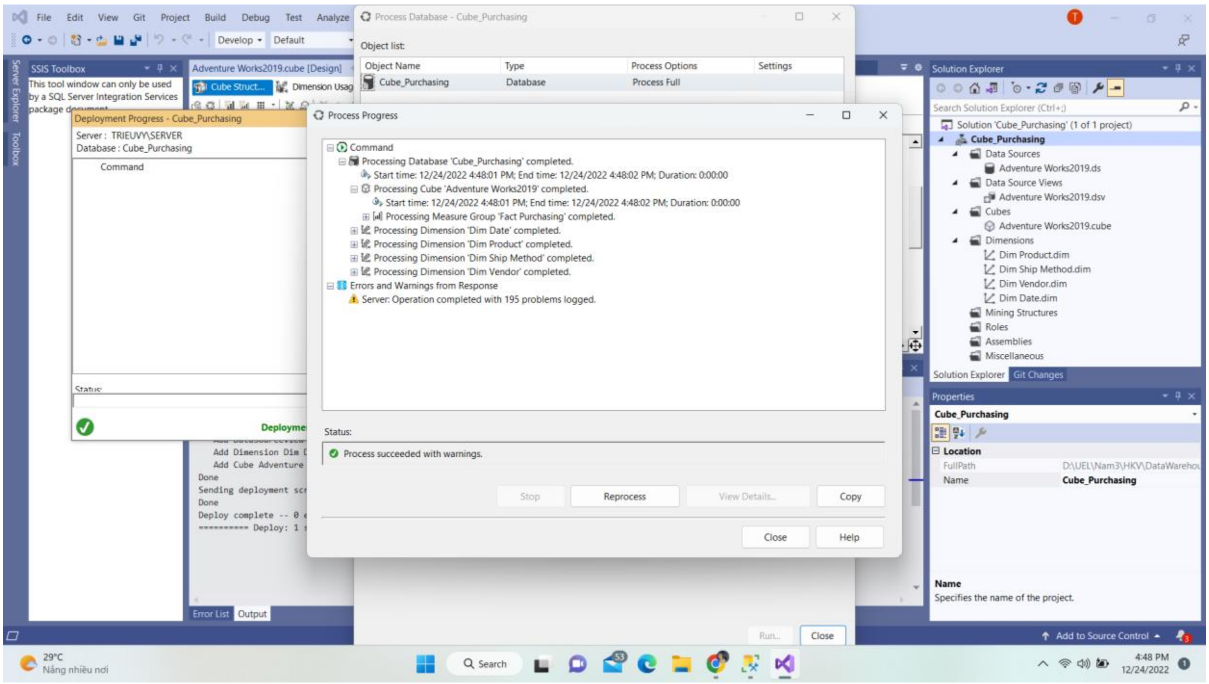


Figure 20: Deploy successful on Purchasing module

5.1.3 Some reports by Cube

We can produce reports based on the requirements and intended uses thanks to the cube. After dragging the Measures elements into the Detail Fields section, drop the Dim attributes from the Dimension Attributes into the row or column that will serve as the measure's representation.

5.1.3.1 Production Module

The report page reports the total number of goods produced in a day.

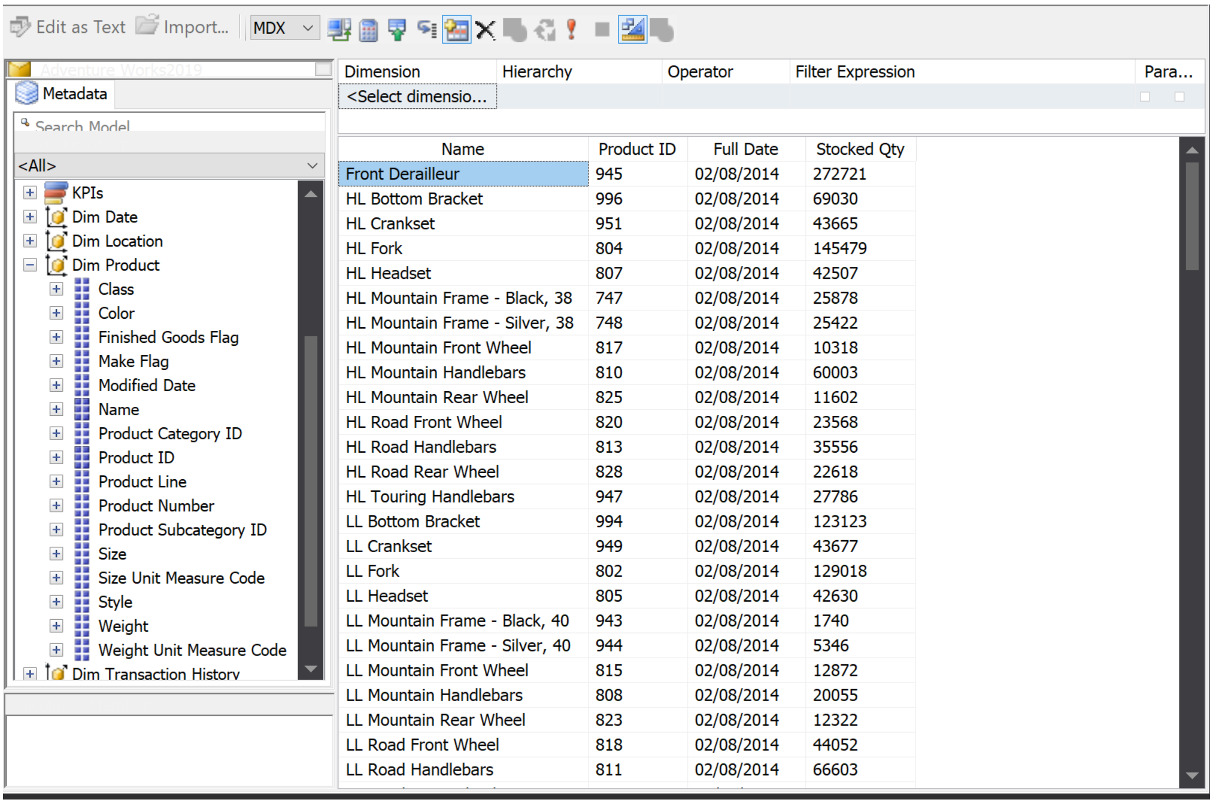
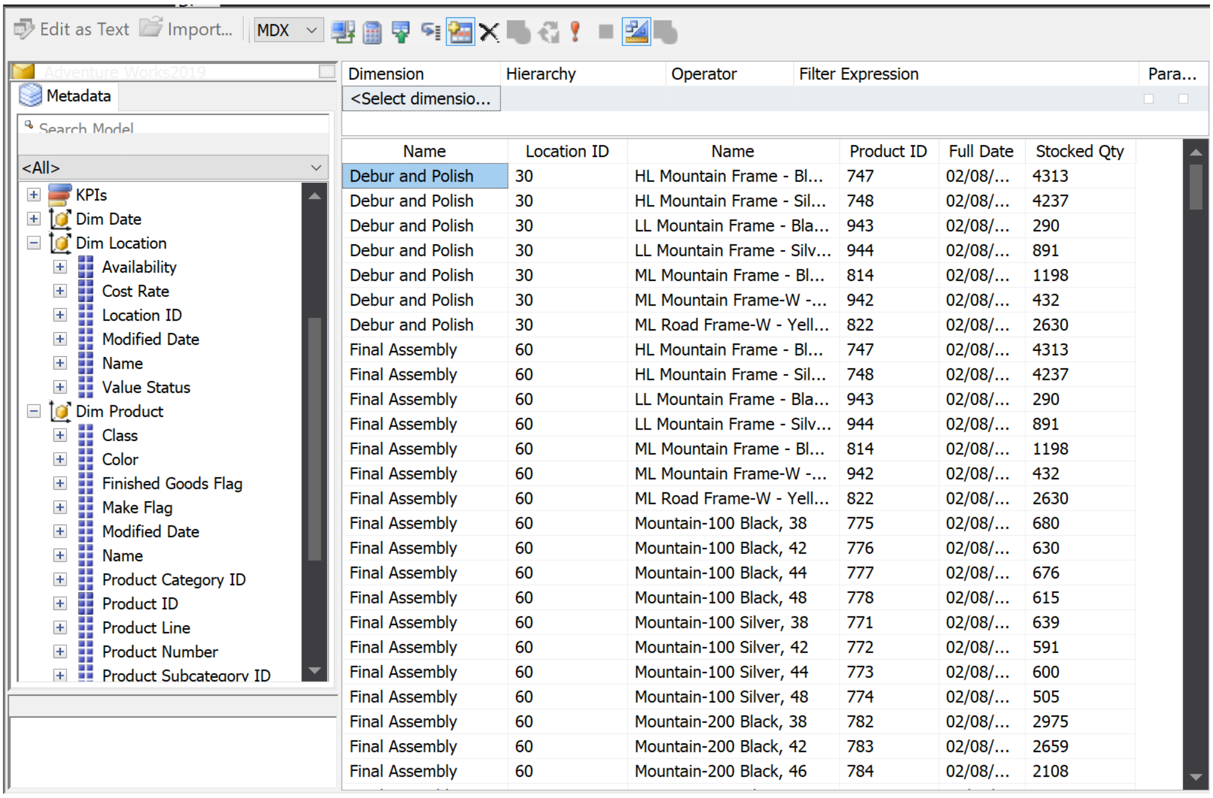


Figure 21: Report on Production module

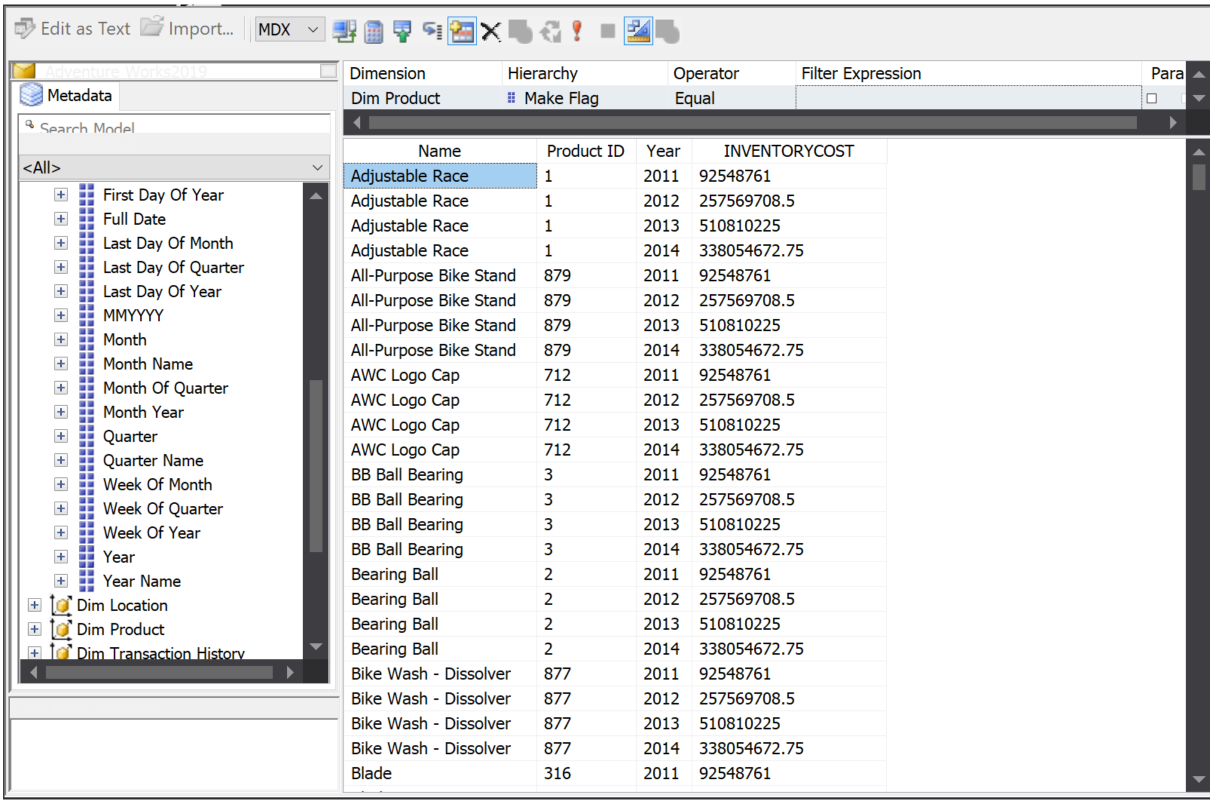
The report page statistics the total quantity of goods by factory in days.



Name	Location ID	Name	Product ID	Full Date	Stocked Qty
Debur and Polish	30	HL Mountain Frame - Bl...	747	02/08/...	4313
Debur and Polish	30	HL Mountain Frame - Sil...	748	02/08/...	4237
Debur and Polish	30	LL Mountain Frame - Bla...	943	02/08/...	290
Debur and Polish	30	LL Mountain Frame - Silv...	944	02/08/...	891
Debur and Polish	30	ML Mountain Frame - Bl...	814	02/08/...	1198
Debur and Polish	30	ML Mountain Frame-W -...	942	02/08/...	432
Debur and Polish	30	ML Road Frame-W - Yell...	822	02/08/...	2630
Final Assembly	60	HL Mountain Frame - Bl...	747	02/08/...	4313
Final Assembly	60	HL Mountain Frame - Sil...	748	02/08/...	4237
Final Assembly	60	LL Mountain Frame - Bla...	943	02/08/...	290
Final Assembly	60	LL Mountain Frame - Silv...	944	02/08/...	891
Final Assembly	60	ML Mountain Frame - Bl...	814	02/08/...	1198
Final Assembly	60	ML Mountain Frame-W -...	942	02/08/...	432
Final Assembly	60	ML Road Frame-W - Yell...	822	02/08/...	2630
Final Assembly	60	Mountain-100 Black, 38	775	02/08/...	680
Final Assembly	60	Mountain-100 Black, 42	776	02/08/...	630
Final Assembly	60	Mountain-100 Black, 44	777	02/08/...	676
Final Assembly	60	Mountain-100 Black, 48	778	02/08/...	615
Final Assembly	60	Mountain-100 Silver, 38	771	02/08/...	639
Final Assembly	60	Mountain-100 Silver, 42	772	02/08/...	591
Final Assembly	60	Mountain-100 Silver, 44	773	02/08/...	600
Final Assembly	60	Mountain-100 Silver, 48	774	02/08/...	505
Final Assembly	60	Mountain-200 Black, 38	782	02/08/...	2975
Final Assembly	60	Mountain-200 Black, 42	783	02/08/...	2659
Final Assembly	60	Mountain-200 Black, 46	784	02/08/...	2108

Figure 22: Report on Production module

The report page reports factory operating costs by year of each product.



Name	Product ID	Year	INVENTORYCOST
Adjustable Race	1	2011	92548761
Adjustable Race	1	2012	257569708.5
Adjustable Race	1	2013	510810225
Adjustable Race	1	2014	338054672.75
All-Purpose Bike Stand	879	2011	92548761
All-Purpose Bike Stand	879	2012	257569708.5
All-Purpose Bike Stand	879	2013	510810225
All-Purpose Bike Stand	879	2014	338054672.75
AWC Logo Cap	712	2011	92548761
AWC Logo Cap	712	2012	257569708.5
AWC Logo Cap	712	2013	510810225
AWC Logo Cap	712	2014	338054672.75
BB Ball Bearing	3	2011	92548761
BB Ball Bearing	3	2012	257569708.5
BB Ball Bearing	3	2013	510810225
BB Ball Bearing	3	2014	338054672.75
Bearing Ball	2	2011	92548761
Bearing Ball	2	2012	257569708.5
Bearing Ball	2	2013	510810225
Bearing Ball	2	2014	338054672.75
Bike Wash - Dissolver	877	2011	92548761
Bike Wash - Dissolver	877	2012	257569708.5
Bike Wash - Dissolver	877	2013	510810225
Bike Wash - Dissolver	877	2014	338054672.75
Blade	316	2011	92548761

Figure 23: Report on Production module

The page reports the statistics of factory costs and production costs by location of the factory.

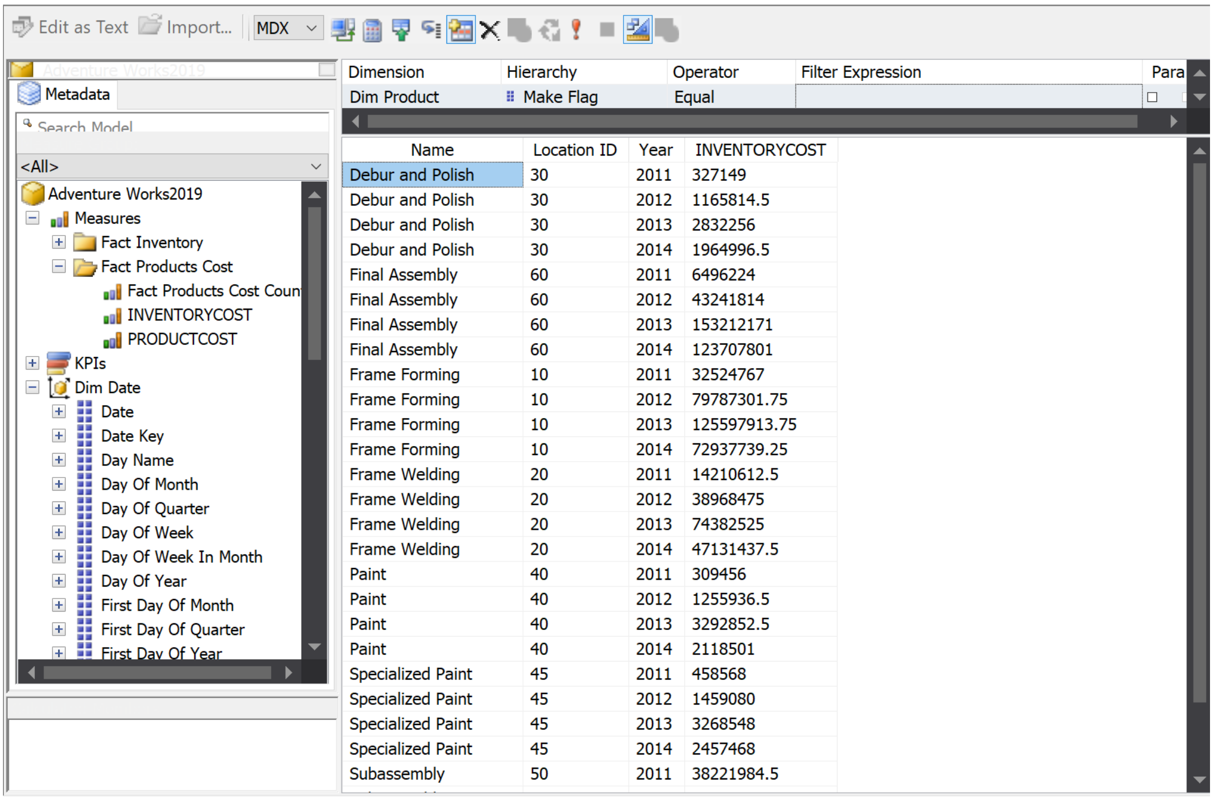


Figure 24: Report on Production module

The page reports production cost statistics based on product by year.

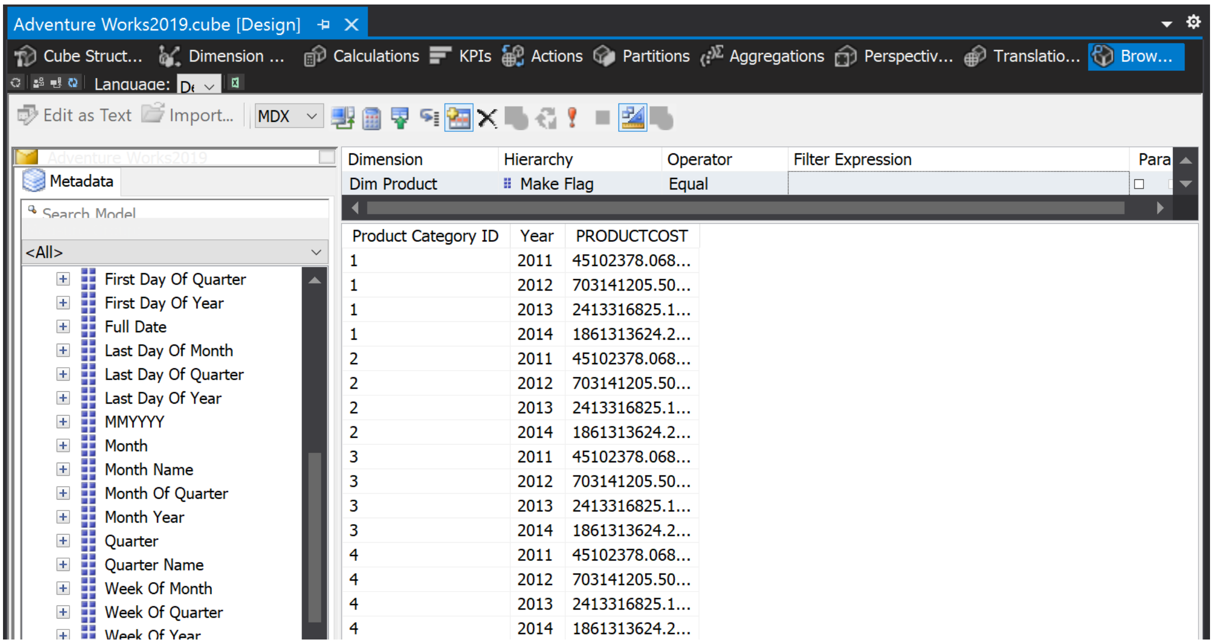


Figure 25: Report on Production module

5.1.3.2 Purchasing Module

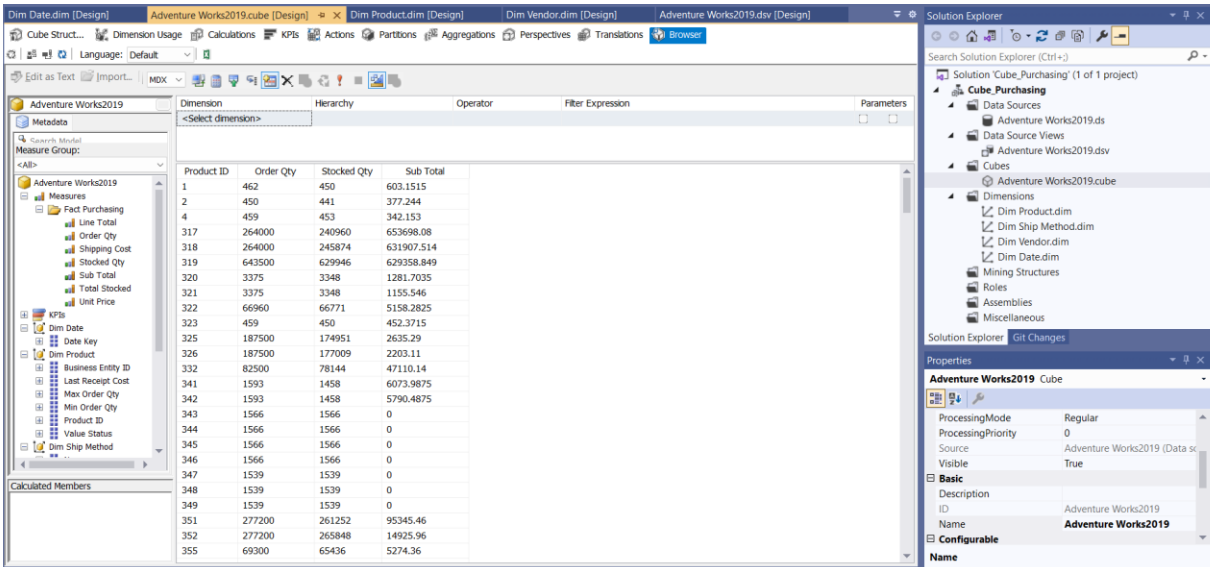


Figure 26: Statistics of the total difference of goods

For this Purchasing Module, we will create a report to list the difference between the quantity of product we ordered and the quantity of product received in stock. Take into account how much the total difference of that product is to be able to transfer costs most accurately. Having said that, Cube helps us to visualize data quickly.

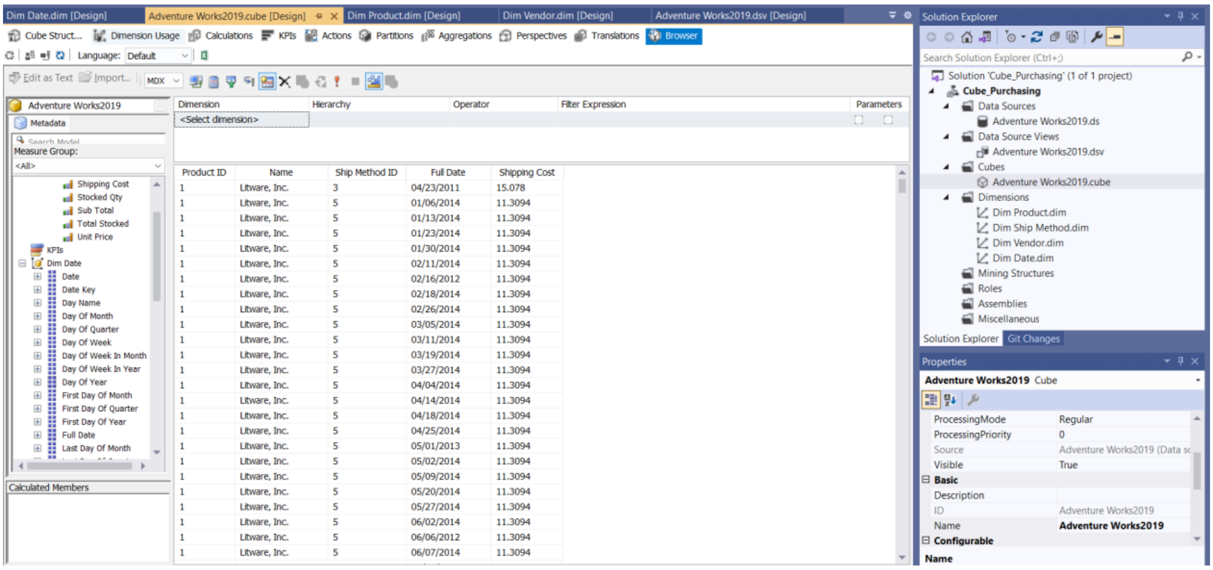


Figure 27: Statistics on products by supplier and delivery costs

For this report, we want to extract goods with data such as the vendor's name, the method of delivering it, the date it was delivered, and the shipping cost of the product. From there, it is feasible to generalize and form an opinion on the product.

5.2. Limitations

For implementation purposes built on a sample database, we only have two modules for two particular cases. They have a complex and multidisciplinary design. It is not very useful for making decisions in real time due to the long processing time it may require. In any case, the tendency of current products (together with hardware advances) is to solve this problem by converting the disadvantage into an advantage. Once implemented, it can be difficult to add new data sources.

In this study, although we have built a pretty detailed data warehouse, we still lose some information. In the Production module, we have not extracted data of products in production. The cost is extracted by days, weeks, months but in detail, we have not extracted the cost of production in hours.

In module Purchasing, we can not extract the reason why the order was canceled, which is important for the company to resolve the problem from the shipping process. The fact Inventory from the module Product is considered a snapshot fact table but in this case this table received a large amount of data, which might be duplicated.

5.3. Future works

The project's objective will be expanded as we carry out more research. Actually, there are numerous different business analyses required for a module. As a result, we anticipate improving and developing our project and two modules selected in the next few years.

Besides, we have seen that the key to data storage is data design. A good data design creates a viable data warehouse. Therefore, we need to check and complete the necessary attributes in the modules to ensure adequate information storage. From there, avoid missing information for future research and analysis.

References:

[1] Bajwa, I. S., Sibalija, T., & Jawawi, D. N. A. (Eds.). (2020). *Intelligent Technologies and Applications. Communications in Computer and Information Science*.doi:10.1007/978-981-15-5232-8.

[2] Nikita Sachdeva, (2022), What is an ETL Tool and Why is it Necessary for Business Growth?, https://bom.so/ytF2dr

[3] Gustavo du Mortier (2021), What Are Facts and Dimensions in a Data Warehouse?, https://bom.so/aLm2Qe

[4] Moody, Daniel L., and Mark AR Kortink (2000). "From enterprise models to dimensional models: a methodology for data warehouse and data mart design." *DMDW*.

[5] Rifaie, Mohammad, et al. "Data warehouse architecture and design." (2008) *IEEE International Conference on Information Reuse and Integration*. IEEE, 2008.

[6] Sahama, Tony, and Peter Croll. "A data warehouse architecture for clinical data warehousing." (2007) *ACSW Frontiers 2007: Proceedings of 5th Australasian Symposium on Grid Computing and e-Research, 5th Australasian Information Security Workshop (Privacy Enhancing Technologies), and Australasian Workshop on Health Knowledge Management and Discovery*. Australian Computer Society.

[7] Vaisman, Alejandro, and Esteban Zimányi (2014) "Data warehouse systems." *Data-Centric Systems and Applications* .

[8] Ariyachandra, Thilini, and Hugh J. Watson. (2006) "Which data warehouse architecture is most successful?." *Business intelligence journal* 11.1.

[9] El-Sappagh, Shaker H. Ali, Abdeltawab M. Ahmed Hendawi, and Ali Hamed El Bastawissy. (2011) "A proposed model for data warehouse ETL processes." *Journal of King Saud University-Computer and Information Sciences* 23.2: 91-104.