PROJECT PROPOSAL V3 - Group 5

INFO 7290 Data Warehousing and Business Intelligence [Spring 2021]

Sakshi Patil, Reema Yadav, Ashi Chaturvedi, Shouvik Ash

TABLE OF CONTENTS

1. INTRODUCTION

Adventure Works 2019 is a sample database for Microsoft SQL Server. The database is about a fictitious, multinational sales/product/customer data called AdventureWorks and the database schema is designed to showcase SQL Server features.

The AdventureWorks database supports standard online transaction processing scenarios for a fictitious bicycle manufacturer (Adventure Works Cycles). Scenarios include Manufacturing, Sales, Purchasing, Product Management, Contact Management, and Human Resources.

Adventure Works 2019 is a vast database and we decide to work on the following tables:

Person.BusinessEntity
HumanResources.Department
Sales.SalesOrderHeader
Sales.Store
Sales.SalesPerson
Sales.SalesOrderDetail
Sales.SalesTerritory
Sales.Currency
Sales.SalesOrderHeaderSalesReason
Sales.SpecialOffer
Sales.SpecialOfferProduct
Production.ProductInventory
Purchasing.PurchaseOrderHeader
Purchasing.PurchaseOrderDetail
Further, data processing will be achieved as per the following flow diagram:-

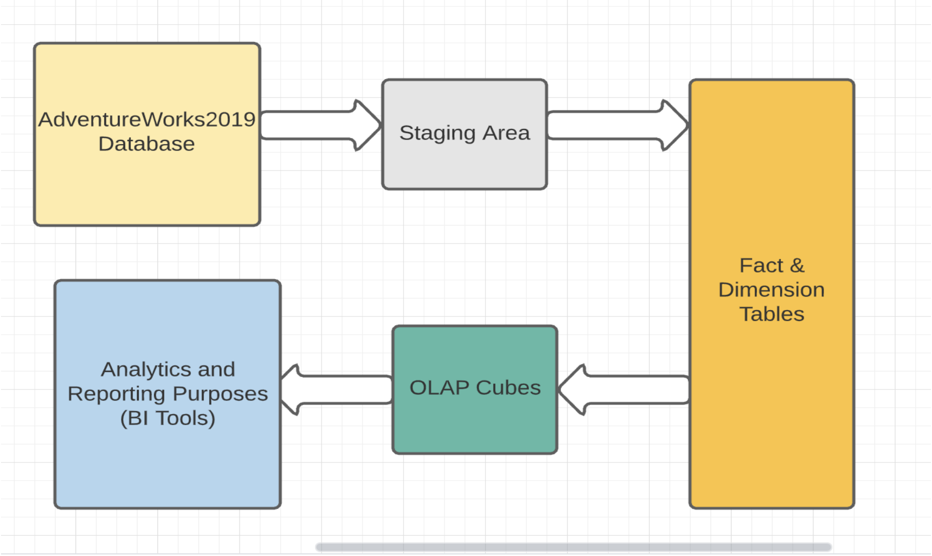

AdventureWorks2019 Database
Staging Area
Fact & Dimension Tables
OLAP Cubes
Analytics and Reporting Purposes (BI Tools)

Figure 1- Data Process Flow Diagram

2. DATA FLOW

We will load the data from multiple table sources listed below into the staging tables in order to carry out necessary data transformations in the staging area.

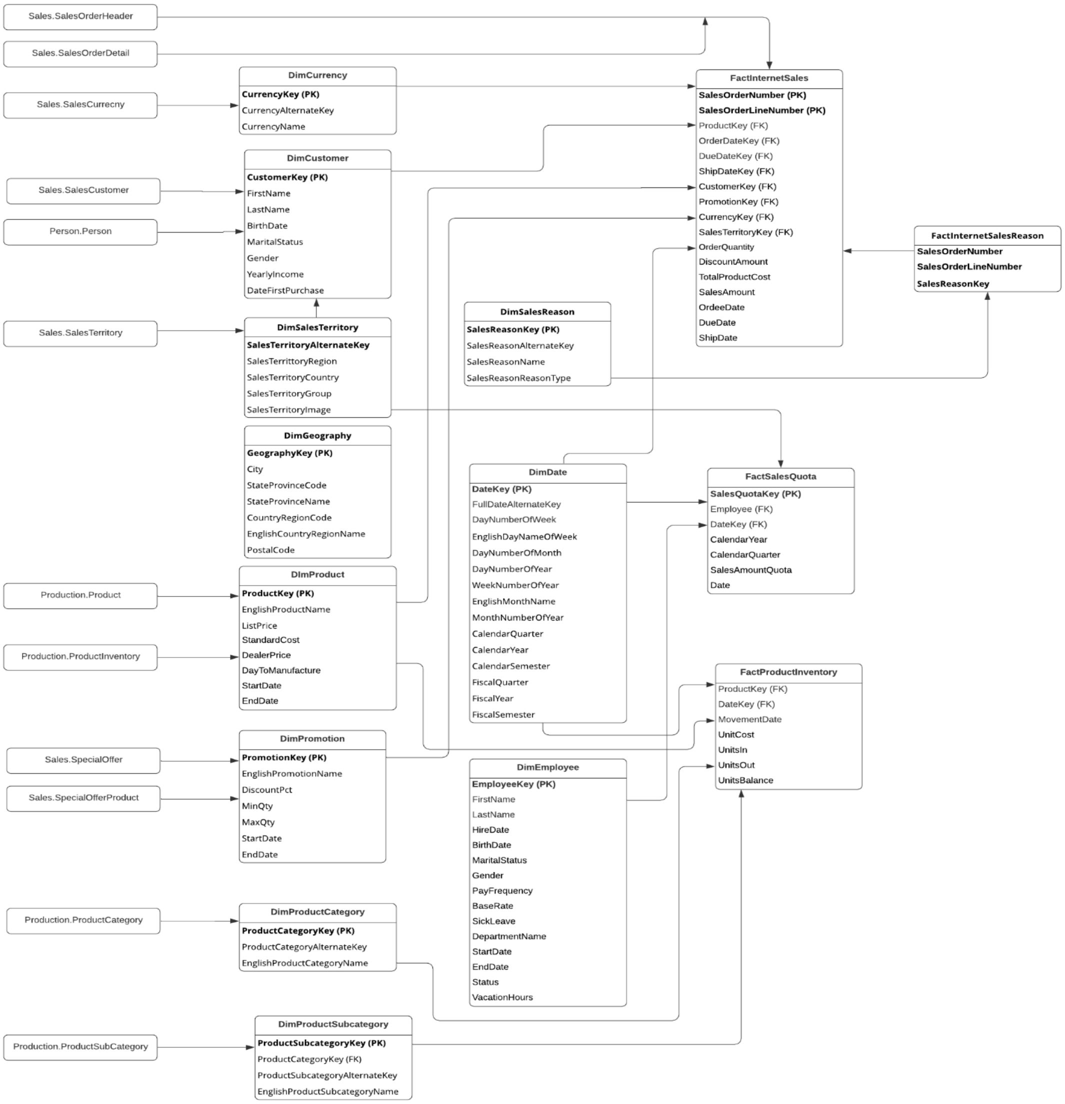


Figure 2- Data Flow Diagram for FactProductInventory, FactInternetSales and FactSalesQuota

Source Table Selection:

TABLE NAME	COLUMN NAMES	TABLE DESCRIPTION
Person.BusinessEntity	BusinessEntityID, rowguid, ModifiedDate	Source of the ID that connects vendors, customers, and employees with address and contact information
Person.Person	BusinessEntityID, PersonType , NameStyle, Title, FirstName , MiddleName, LastName, Suffix, EmailPromotion, AdditionalContactInfo, Demographics, rowguid, ModifiedDate	Human beings involved with AdventureWorks: employees, customer contacts, and vendor contacts
HumanResources.Department	DepartmentID, Name, GroupName, ModifiedDate	Lookup table containing the departments within the Adventure Works Cycles company
HumanResources.Employee		
Sales.SalesOrderHeader	SalesOrderID, RevisionNumber, OrderDate, DueDate, ShipDate, Status, OnlineOrderFlag, SalesOrderNumber, PurchaseOrderNumber, AccountNumber, CustomerID, SalesPersonID, TerritoryID, BillToAddressID, ShipToAddressID, ShipMethodID, CreditCardID, CreditCardApprovalCode, CurrencyRateID, SubTotal, TaxAmt, Freight, TotalDue, Comment, rowguid, ModifiedDate	General sales order information
Sales.Store	BusinessEntityID, Name, SalesPersonID, Demographics, rowguid, ModifiedDate	Customers (resellers) of Adventure Works products
Sales.SalesPerson	BusinessEntityID, TerritoryID, SalesQuota, Bonus, CommissionPct, SalesYTD,	Sales representative current information

	SalesLastYear, rowguid, ModifiedDate	
Sales.SalesOrderDetail	SalesOrderID, SalesOrderDetailID, CarrierTrackingNumber, OrderQty, ProductID, SpecialOfferID, UnitPrice, UnitPriceDiscount	Individual products associated with a specific sales order
Sales.SalesTerritory	TerritoryID, Name, CountryRegionCode, Group, SalesYTD, SalesLastYear, CostYTD, CostLastYear, rowguid, ModifiedDate	Sales territory lookup table
Sales.Currency	CurrencyCode, Name, ModifiedDate	Lookup table containing standard ISO currencies
Sales.SalesOrderHeaderSalesReason	SalesOrderID, SalesReasonID, ModifiedDate	Cross-reference table mapping sales orders to sales reason codes
Sales.SalesCustomer	CustomerID, PersonID, StoreID, TerritoryID, AccountNumber, rowguid, ModifiedDate	Current customer information
Sales.SpecialOffer	SpecialOfferID, Description, DiscountPct, Type, Category, StartDate, EndDate, MinQty, MaxQty, rowguid ModifiedDate	Sale discounts lookup table
Sales.SpecialOfferProduct	SpecialOfferID, ProductID, rowguid, ModifiedDate	Cross-reference table mapping products to special offer discounts
Production.ProductInventory	ProductID, LocationID, Shelf, Bin, Quantity, rowguid, ModifiedDate	Product inventory information
Production.Product	ProductID, Name, ProductNumber, MakeFlag, FinishedGoodsFlag, Color, SafetyStockLevel, ReorderPoint, StandardCost, ListPrice, Size, SizeUnitMeasureCode, WeightUnitMeasureCode, Weight, DaysToManufacture, ProductLine, Class, Style,	Products sold or used in the manufacturing of sold products.

	ProductSubcategoryID, ProductModelID, SellStartDate, SellEndDate, DiscontinuedDate, rowguid, ModifiedDate	
Production.ProductCategory	ProductCategoryID, Name, rowguid, ModifiedDate	High-level product categorization.
Production.ProductSubcategory	ProductSubcategoryID, ProductCategoryID, Name, rowguid, ModifiedDate	Product subcategories. See ProductCategory table.
Purchasing.PurchaseOrderHeader	PurchaseOrderID, RevisionNumber, Status, EmployeeID, VendorID, ShipMethodID, OrderDate, ShipDate, SubTotal, TaxAmt, Freight, TotalDue, ModifiedDate	General purchase order information
Purchasing.PurchaseOrderDetail	PurchaseOrderID, PurchaseOrderDetailID, DueDate, OrderQty, ProductID, UnitPrice, LineTotal, ReceivedQty	Individual products associated with a specific purchase order

Fact Table Selection:

TABLE NAME	COLUMN NAMES	TABLE DESCRIPTION
FactSalesQuota	SalesQuotaKey, DateKey, EmployeeKey, CalendarYear, CalendarQuarter, SalesAmountQuota, Date	Employee information combined with date details and Sales information
FactProductInventory	ProductKey, DateKey, MovementDate, UnitCost, UnitsIn, UnitsOut, UnitsBalance	Inventory details combined with product and date details
FactInternetSales	ProductKey, OrderDateKey, DueDateKey, ShipDateKey, CustomerKey, PromotionKey, CurrencyKey, SalesTerritoryKey, SalesOrderNumber,SalesOrderLineNumber, OrderQuantity, DiscountAmount,	It has summary of product, date, customer information, promotion details, currency and territory details

	TotalProductCost,SalesAmount, OrderDate, DueDate, ShipDate	

Types of facts for the Fact tables:-

FactSalesQuota

- Additive: SalesAmountQuota
- Semi-additive: Date

FactProductInventory

- Semi-additive: MovementDate, UnitsIn, UnitsOut, UnitsBalance
- Nonadditive: UnitCost

FactInternetSales

- Additive: SalesAmount, TotalProductCost
- Semi-additive: OrderQuantity, OrderDate, ShipDate, DueDate,
- Nonadditive: DiscountAmount

<u>Dimension Table Selection With Type of SCDs:</u>

TABLE NAME	COLUMN NAMES	TABLE DESCRIPTION	TYPE OF SCDs
DimEmployee	FirstName, LastName, HireDate, BirthDate, MaritalStatus, Gender, PayFrequency, BaseRate, SickLeave, DepartmentName, Vacation Hours	Employee basic information including some basic company related information	Type 1
	Status, StartDate, EndDate		Type 2
DimDate	FullDateAlternateKey, DayNumberOfWeek, EnglishDayNameOfWeek, DayNumberOfMonth, DayNumberOfYear, WeekNumberOfYear, EnglishMonthName, MonthNumberOfYear, CalendarQuarter,	All basic date attributes that define a particular date in an year	Type 0

	CalendarYear, CalendarSemester, FiscalQuarter, FiscalYear, FiscalSemester		
DimProduct	EnglishProductName, ListPrice, StandardCost, DealerPrice,	Basic Product attributes that are associated with a particular product	Type 1
	DayToManufacture, Status StartDate, EndDate		Type 2
DimProductSubcategory	ProductSubcategoryKey, ProductSubcategoryAlternateKey, EnglishProductSubcategoryName, ProductCategoryKey	Product Sub Category attributes	Type 1
DimProductCategory	ProductCategoryKey, ProductCategoryAlternateKey, EnglishProductCategoryName	Product Category attributes	Type 1
DimCurrency	CurrencyKey, CurrencyAlternateKey, CurrencyName	Currency Type and its other attributes	
DimPromotion	EnglishPromotionName	Product and it's discount attributes	Type 1
	DiscountPct, MinQty, MaxQty, StartDate, EndDate		Type 2
DimCustomer	FirstName, LastName, BirthDate, Gender, MaritalStatus	Customer basic information including some personal information	Type 1
	YearlyIncome, DateFirstPurchase		Type 2
DimSalesReason	SalesReasonAlternateKey, SalesReasonName, SalesReasonReasonType	The reason of sale details	Type 1

DimGeography	City, StateProvinceCode, StateProvinceName, CountryRegionCode, EnglishCountryRegionName, PostalCode	Fully defined Geography	Type 1
DimPromotion	EnglishPromotionName	Promotion with discount attributes	Type 1
	DiscountPct, MinQty, MaxQty, StartDate, EndDate		Type 2

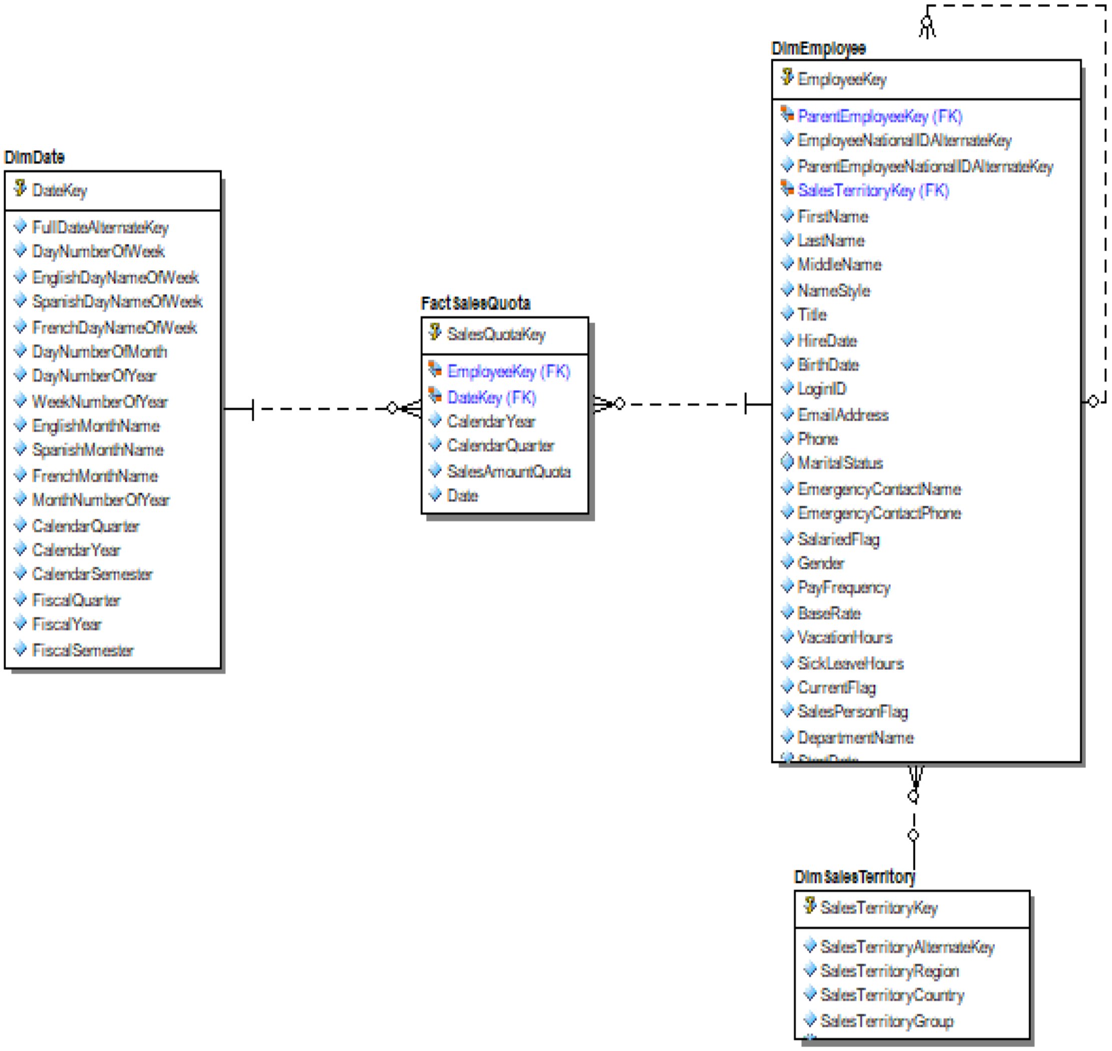


Figure 3- Schema for the FactSalesQuota

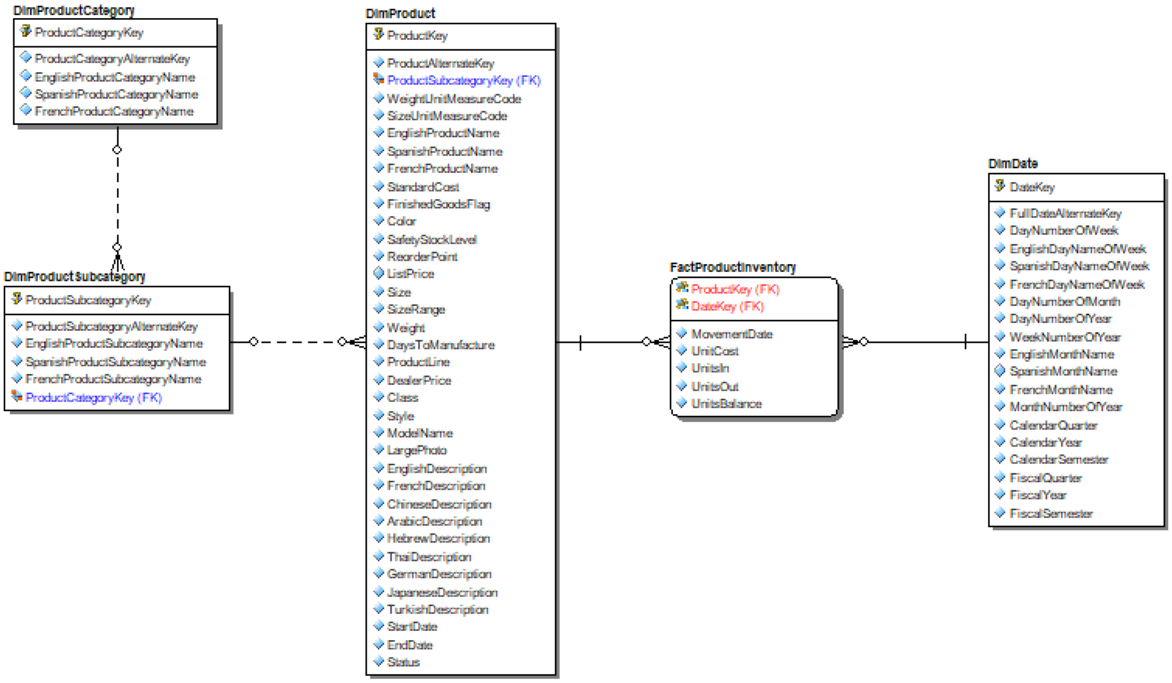


Figure 4- Schema for the FactProductInventory

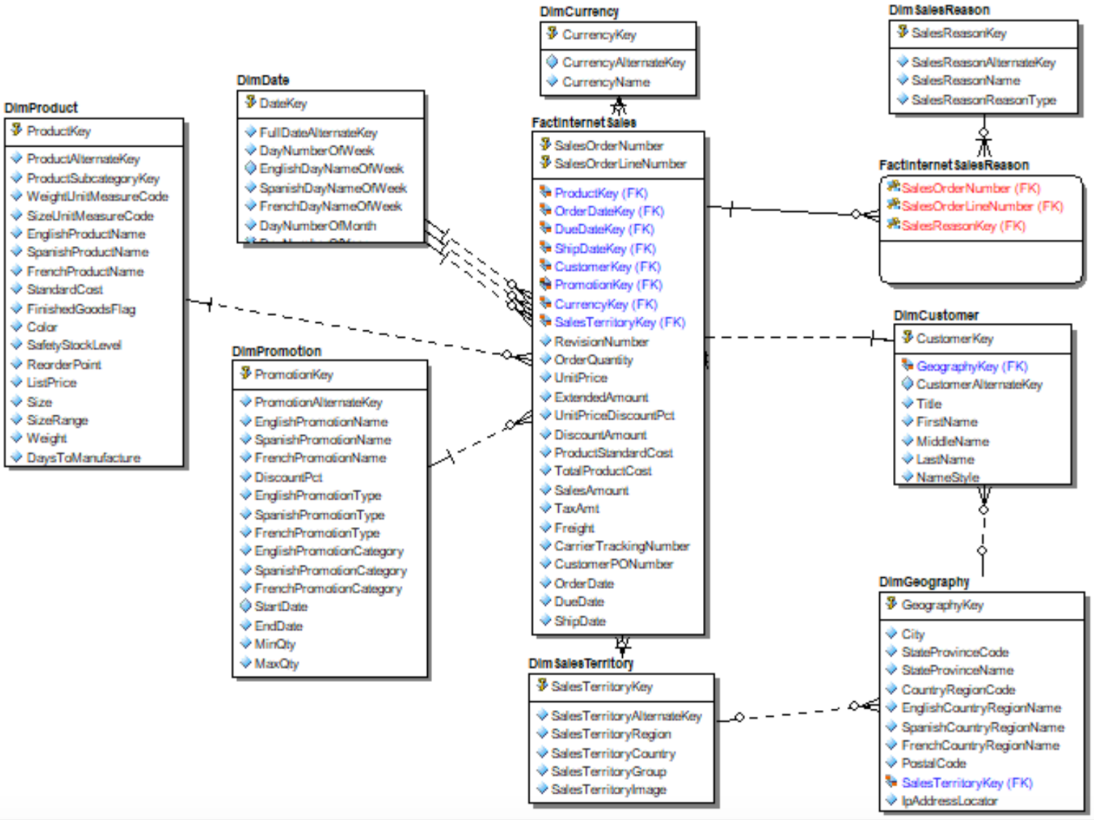


Figure 5- Schema for the FactInternetSales

3. DATA TRANSFORMATION:

Data transformation is the process of changing the structure,format or, values of the data or we can say that it is the mapping and conversion of data from one format to another. Basic data transformation are Cleaning: Mapping NULL to 0 or "Male" to "M" and "Female" to "F," date format consistency, etc. Deduplication: Identifying and removing duplicate records. Format revision: Character set conversion, unit of measurement conversion, date/time conversion.

In our case, we will perform the data transformation by checking if there are any duplicate or null entries in the table. Then,we will separate multivalued attributes into different entries and will remove any special characters present in the data columns. Further, we will eliminate the entries that do not satisfy the referential constraints.

The data is taken from the OLE DB source and the data conversion function is performed if required. The Slowly Changing Dimension (SCD) is performed on the dimension tables for the data. A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse.

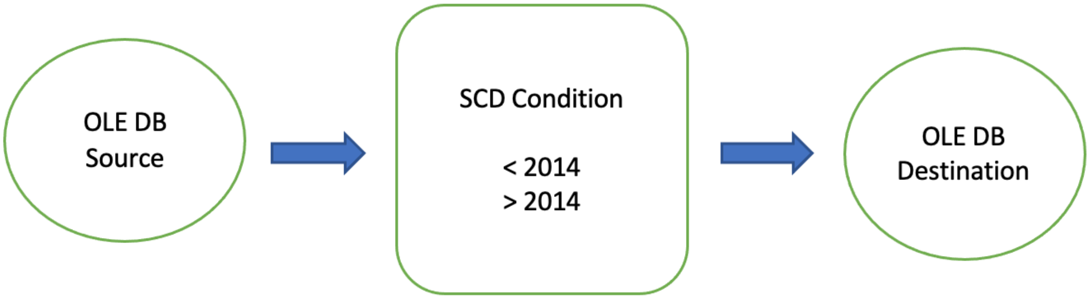
Using the SSIS data flow we will load the data from multiple table sources into the staging tables and avoid loading duplicates in the database tables.

Source: Tables from MS SQL Database
Destination: OLE DB Destination

4. EXTRACT TRANSFORM LOAD DATA

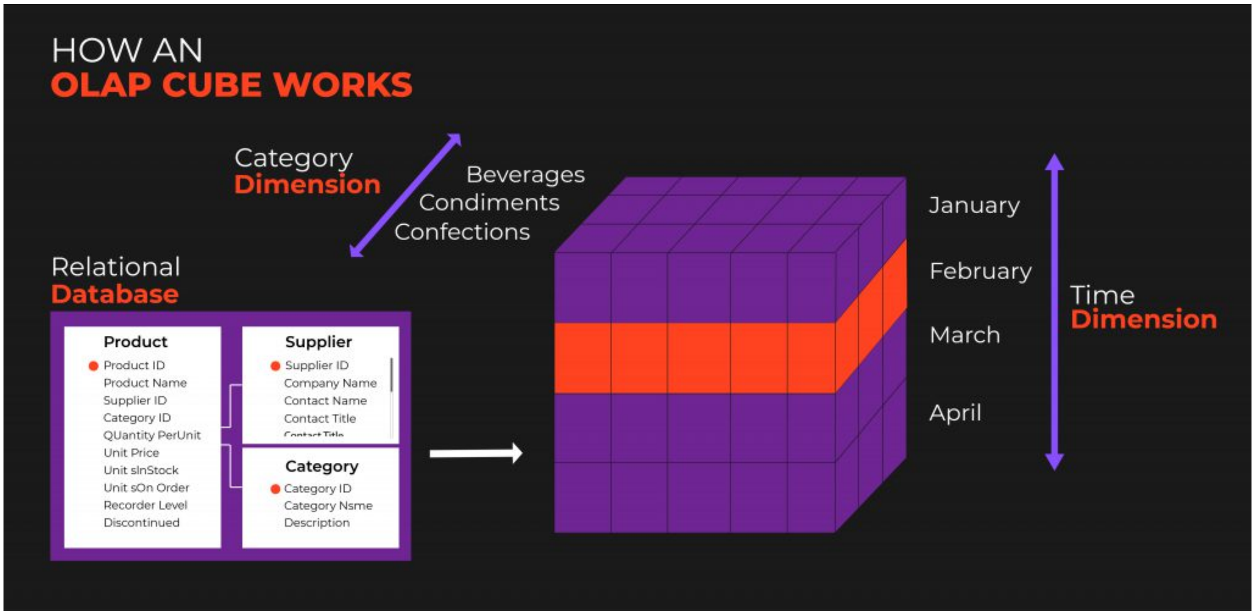
ETL is a method of data integration that uses the three steps (extract, transform, load) to blend data from multiple sources. It is often used to build a data warehouse. The data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse or other system. Extract, load, transform (ELT) is an alternate but related approach designed to push processing down to the database for improved performance.

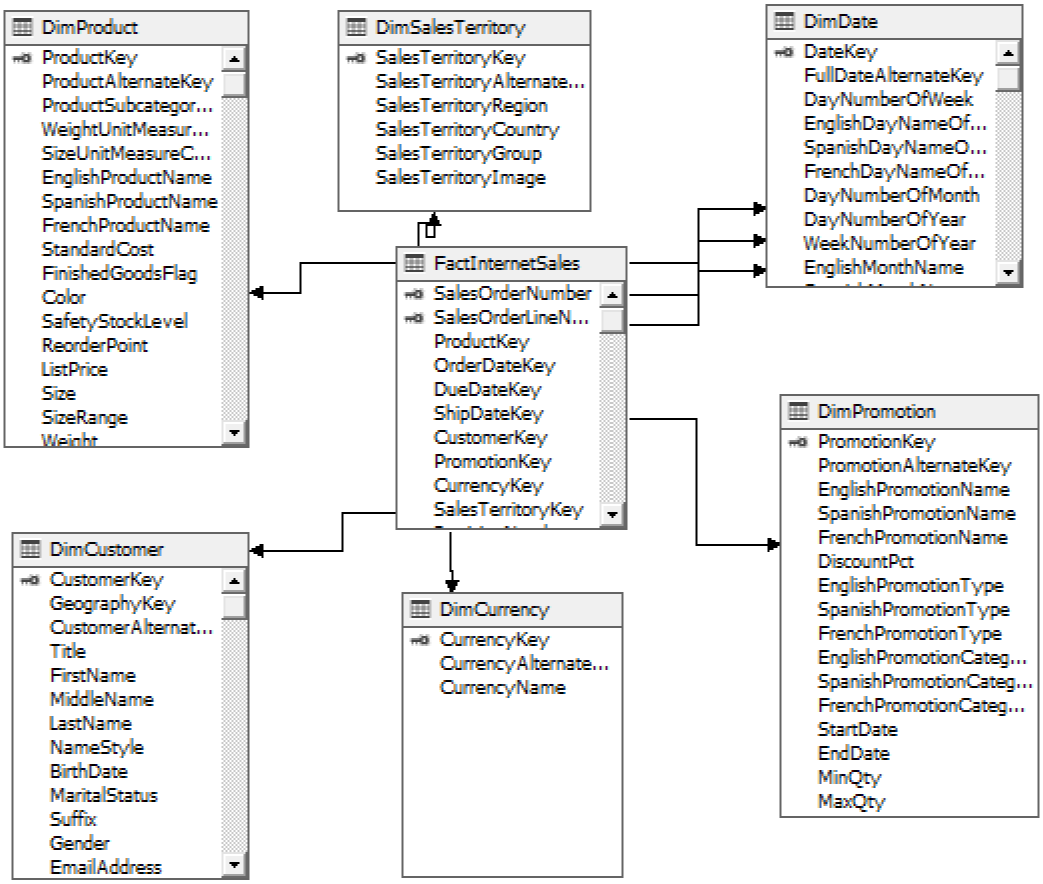
With initial load we will load the data before year 2014 into the destination Fact table. We will retrieve all the related data for the dimension table, apply SCD and populate them. We will do the incremental load by processing data daily after 2014. We will pull the data from present day and load them into the staging area.



5. OLAP CUBE

An OLAP cube is a data structure that overcomes the limitations of relational databases by providing rapid analysis of data. Cubes can display and sum large amounts of data while also providing users with searchable access to any data points. These cubes are stored in SQL Server Analysis Services (SSAS).





6. VISUALIZATIONS

From the analytical data which we obtain from the OLAP cubes, we will develop some dashboards on Tableau with appropriate visualizations for uncovering some hidden insights from the data accumulated in the cubes. For eg. a map visualization can showcase the region wise or country wise sales on monthly basis and even on yearly basis. Further, we can analyze the demand of a particular product in a region or country by visualizing the analytical data of sales volume of that particular product. Also, we can compare the monthly sales patterns on a time-series graph for the AdventureWorks company and identify the business challenges which need to be addressed to increase their sales.

7. PROFESSOR COMMENTS

Professors Comments 1st Draft

- Intro paragraph is a good idea, you need to expand upon it more
- As document grows a table of contents will be needed
- You describe the sources which is good but you should describe them further
- You will need a data model at some time
- High level flow is good but you more description of what will be happening
- You don’t need to use flat files for input but that is acceptable if you want
- I am not sure what the purpose of the sort process is for.

Professor Comments 2nd Draft

- A table of contents usually has page numbers
- Why are you oshowing a picture of adventureworks 2008
- You describe the source tables but I don’t know what the destination of them will be
- Your section on data transformation should be more detailed… what attributes do you need to modify
- Why does your picture show two data marts?
- The flows don’t tell me anything
- You need to include a modification history in your document
- The source section is really the only part of the document I consider complete

8. MODIFICATION HISTORY

1. Table of content modified, page numbers added.
2. Added the content in introduction, data transformation, ETL by explaining the meaning, and usage in our case.
3. Added ETL diagram, explained before and after 2014 data loading.
4. Table description added and the formatting of fact and dimension table changed.
5. Data flow diagram added along with schema diagrams of fact tables.
6. Added a section on OLAP cubes and the structure of the cube which we plan to process.
7. Discussed about the visualizations which we plan to develop based on our analytical data.