CURRENT TOPICS IN COMPUTER SCIENCE

Business Intelligence Systems and Analytics

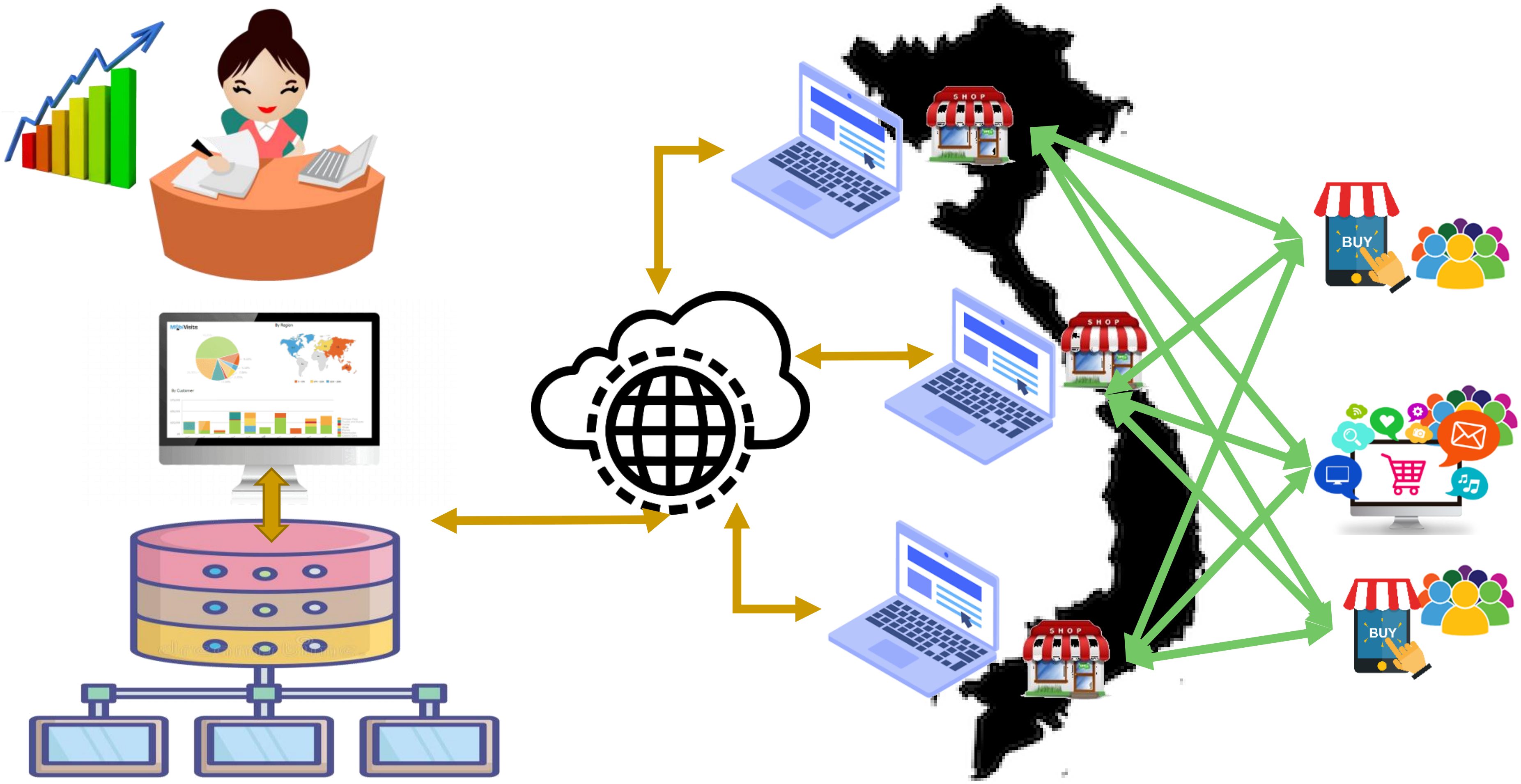
DATASET INVESTIGATION

Trong Nhan Phan, PhD

OUTLINE

- Introduction
- OLTP vs. OLAP
- Dataset Investigation
- Summary
- References

INTRODUCTION



http://www.dinh.vip/small-store-icon-psd-free-psd; https://iconbug.com/data/06/512/f91d30a37d1b631de0005b1402442924.png; https://static.pakwheels.com/2015/06/graph.gif; https://image.shutterstock.com/image-vector/high-detailed-vector-map-vietnam-260nw-150472901.jpg; https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcQeeOZor21HxfRcdDaFqBdsNqBknZoYjo1b1h8Lv0FAmwXwLjqr&usqp=CAU; https://st2.depositphotos.com/1001911/7684/v/450/depositphotos_76840879-stock-illustration-depressed-emoticon.jpg; https://www.onlinewebfonts.com/icon/504308; https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcRxN_w4kEKxFlrNl_P7sJrrIaKDbPgnxeleV3aZO7uCuRPPibJ8&usqp=CAU; https://telegramchannels.me/stickers/s-thevirus; https://thumbs.dreamstime.com/b/centralized-database-fill-vector-icon-which-can-easily-modify-edit-centralized-database-fill-vector-icon-which-can-easily-185306424.jpg; https://e7.pngegg.com/pngimages/56/552/png-clipart-dashboard-business-intelligence-report-analytics-computer-icons-analysis-company-display-advertising.png

OLTP VS. OLAP

ALICE' SYSTEMS

OLAP

- Data warehouse
- BI system
- Product quality assurance system
- New store development system
- Market analysis
- Data mining system
- …

- ERP
- CRM
- Tracking and Analytics
- …

OLTP

POS system
Delivery system
Reservation system
Feedback system
Kitchen management system
Handy system
Tablet order system
Human resource system
Finance system
Supply chain management system
Customer relationship management system
Marketing system
…

http://www.dinh.vip/small-store-icon-psd-free-psd; https://iconbug.com/data/06/512/f91d30a37d1b631de0005b1402442924.png; https://static.pakwheels.com/2015/06/graph.gif; https://image.shutterstock.com/image-vector/high-detailed-vector-map-vietnam-260nw-150472901.jpg; https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcQeeOZor21HxfRcdDaFqBdsNqBknZoYjo1b1h8Lv0FAmwXwLjqr&usqp=CAU; https://st2.depositphotos.com/1001911/7684/v/450/depositphotos_76840879-stock-illustration-depressed-emoticon.jpg; https://www.onlinewebfonts.com/icon/504308; https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcRxN_w4kEKxFlrNl_P7sJrrIaKDbPgnxeleV3aZO7uCuRPPibJ8&usqp=CAU; https://telegramchannels.me/stickers/s-thevirus; https://thumbs.dreamstime.com/b/centralized-database-fill-vector-icon-which-can-easily-modify-edit-centralized-database-fill-vector-icon-which-can-easily-185306424.jpg; https://e7.pngegg.com/pngimages/56/552/png-clipart-dashboard-business-intelligence-report-analytics-computer-icons-analysis-company-display-advertising.png

OLTP

- On-Line Transactional Processing (OLTP)
- Focusing much on
    - High volume of transactions
    - Data manipulation (SELECT, INSERT, UPDATE, DELETE)
        - Add product to shopping cart
        - Update item price
        - Display product information
    - Fast and incomplex query processing
    - Data integrity in multi-access environments
    - Current and detailed data
    - High normal forms of databases

OLAP

- On-Line Analytical Processing (OLAP)
- Focusing much on
  - Low volume of transactions
  - Data selection (SELECT)
    - Report total sales in each areas for each month
    - Display “super hero” products
    - Indetify VIP customers
  - Complex and aggregated queries
  - Data integration and data quality
  - Historical and summarized data
  - Low normal forms of database

OLTP vs. OLAP IN GENERAL

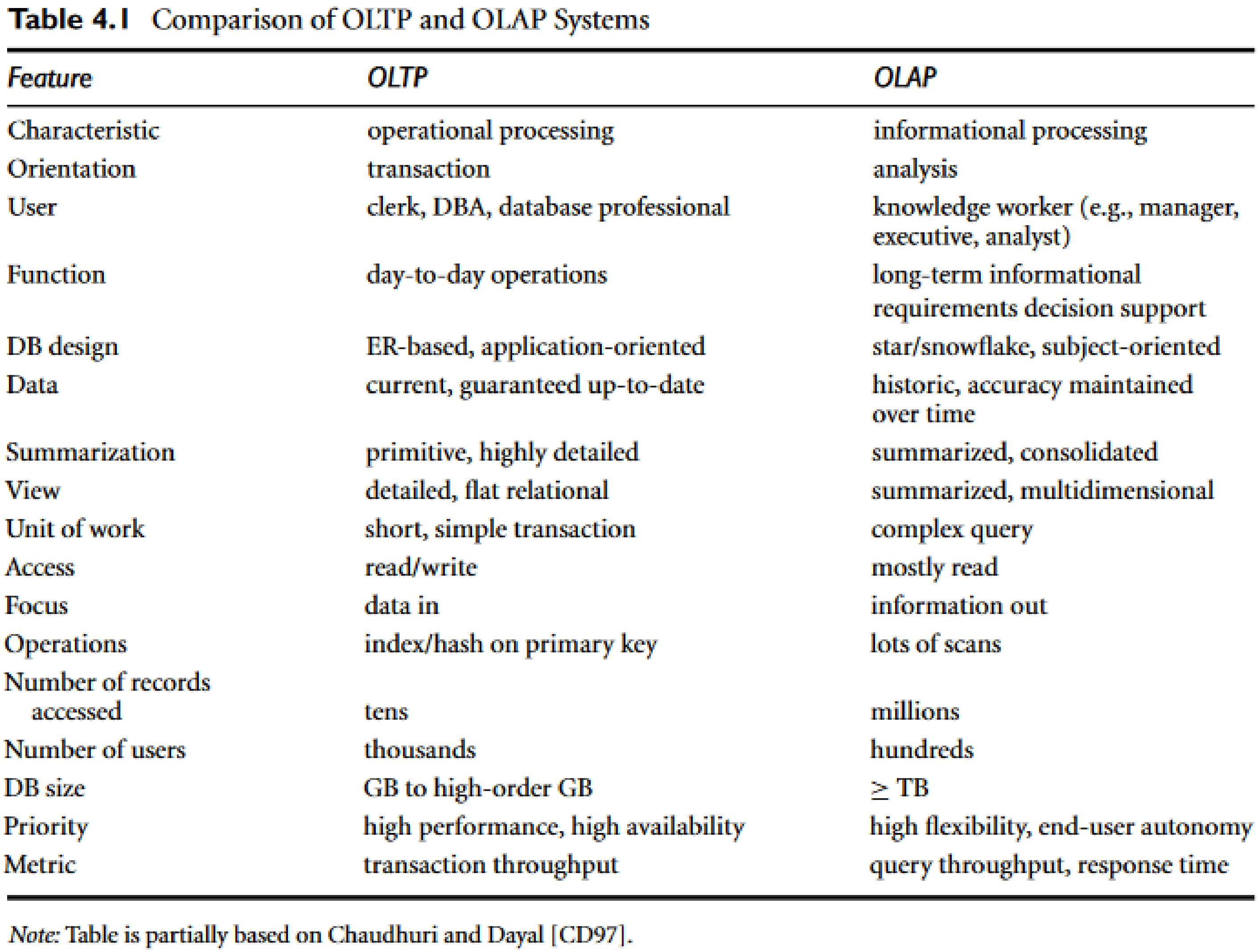


Table 4.1 Comparison of OLTP and OLAP Systems

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Note: Table is partially based on Chaudhuri and Dayal [CD97].

[2]

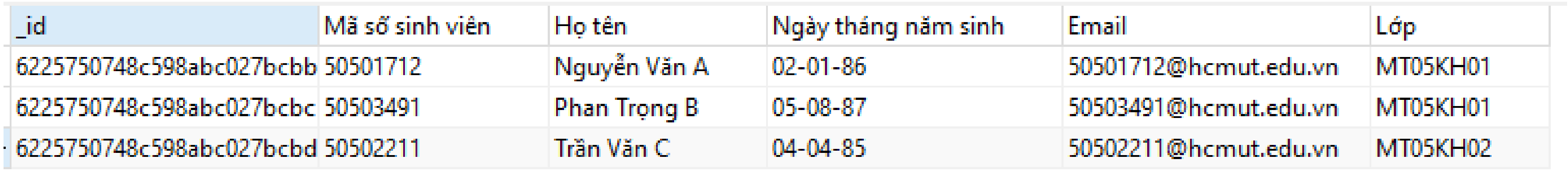
OLTP vs. OLAP: WHAT'S MORE

Characteristic	OLTP	OLAP
Performance	Fast response time is important (normally < 1 second)	Response time may be longer (hours to days and more)
Data model	Complex models (SQL vs. NoSQL) Normalized database	Simplified models Denormalized database
Data	Up-to-date and consistent data at all times Current data	High quality and integrated data Current and historical data
Process	A particular process (e.g., ordering items)	Integrated processes (e.g., NET sales)
Data source	One	Many

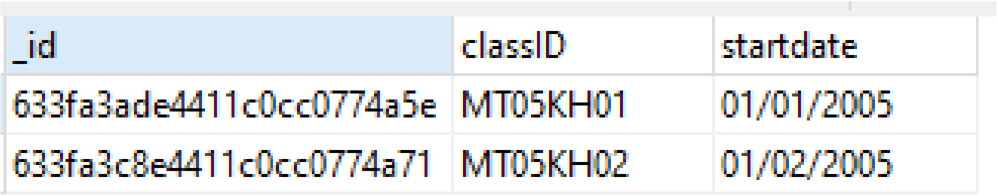
DISCUSSION

ANALYTICAL DATABASE VS. TRANSACTIONAL DATABASE

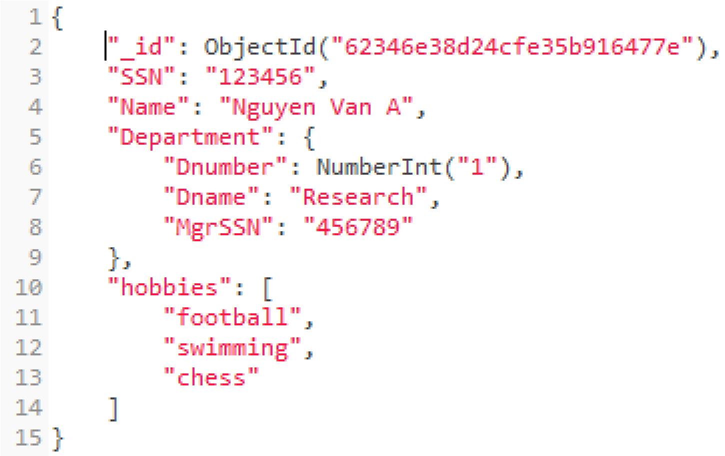
TRANSACTIONS VS. ANALYTICS



_id	Mã số sinh viên	Họ tên	Ngày tháng năm sinh	Email	Lớp
6225750748c598abc027bcbb	50501712	Nguyễn Văn A	02-01-86	50501712@hcmut.edu.vn	MT05KH01
6225750748c598abc027bcbc	50503491	Phan Trọng B	05-08-87	50503491@hcmut.edu.vn	MT05KH01
6225750748c598abc027bcbd	50502211	Trần Văn C	04-04-85	50502211@hcmut.edu.vn	MT05KH02



_id	classID	startdate
633fa3ade4411c0cc0774a5e	MT05KH01	01/01/2005
633fa3c8e4411c0cc0774a71	MT05KH02	01/02/2005



```
{
    "_id": ObjectId("62346e38d24cfe35b916477e"),
    "SSN": "123456",
    "Name": "Nguyen Van A",
    "Department": {
        "Dnumber": NumberInt("1"),
        "Dname": "Research",
        "MgrSSN": "456789"
    },
    "hobbies": [
        "football",
        "swimming",
        "chess"
    ]
}
```

DISCUSSION

NORMALIZATION VS.
DENORMALIZATION

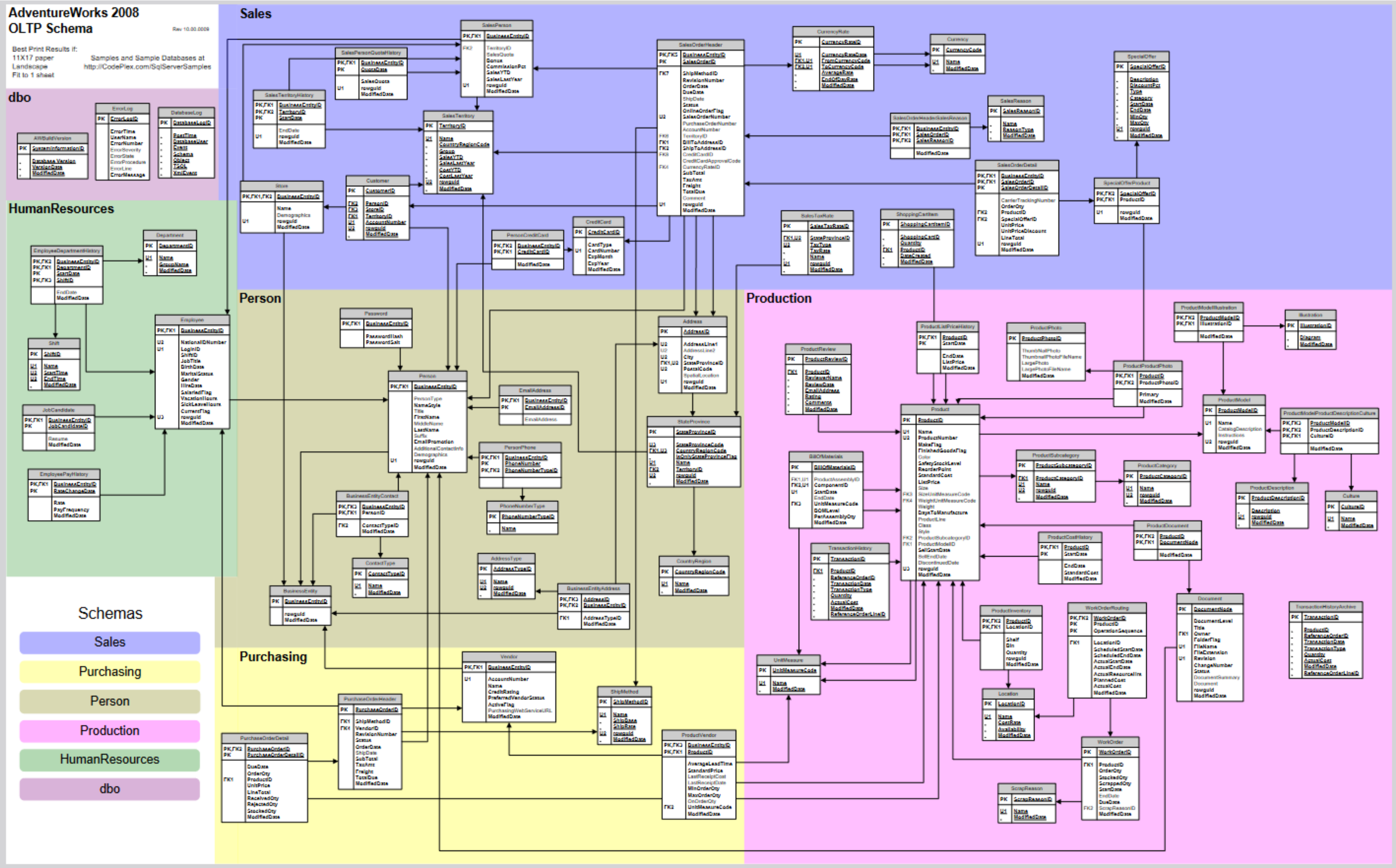
FOR INSTANCE

AgencyID	AgencyName	ProductID	ProductName	ProductPrice	Quantity	Date
1	A	101	Beauty Soap	7	120	01/01/2022
1	A	102	Tooth Brush	5	100	01/01/2022
2	B	103	Tooth Paste	4	80	01/02/2022
3	C	103	Tooth Paste	4	110	01/02/2022
…	…	…	…	…	…	…

StudentID	StudentName
1001	NVA
1002	NVB
…	…

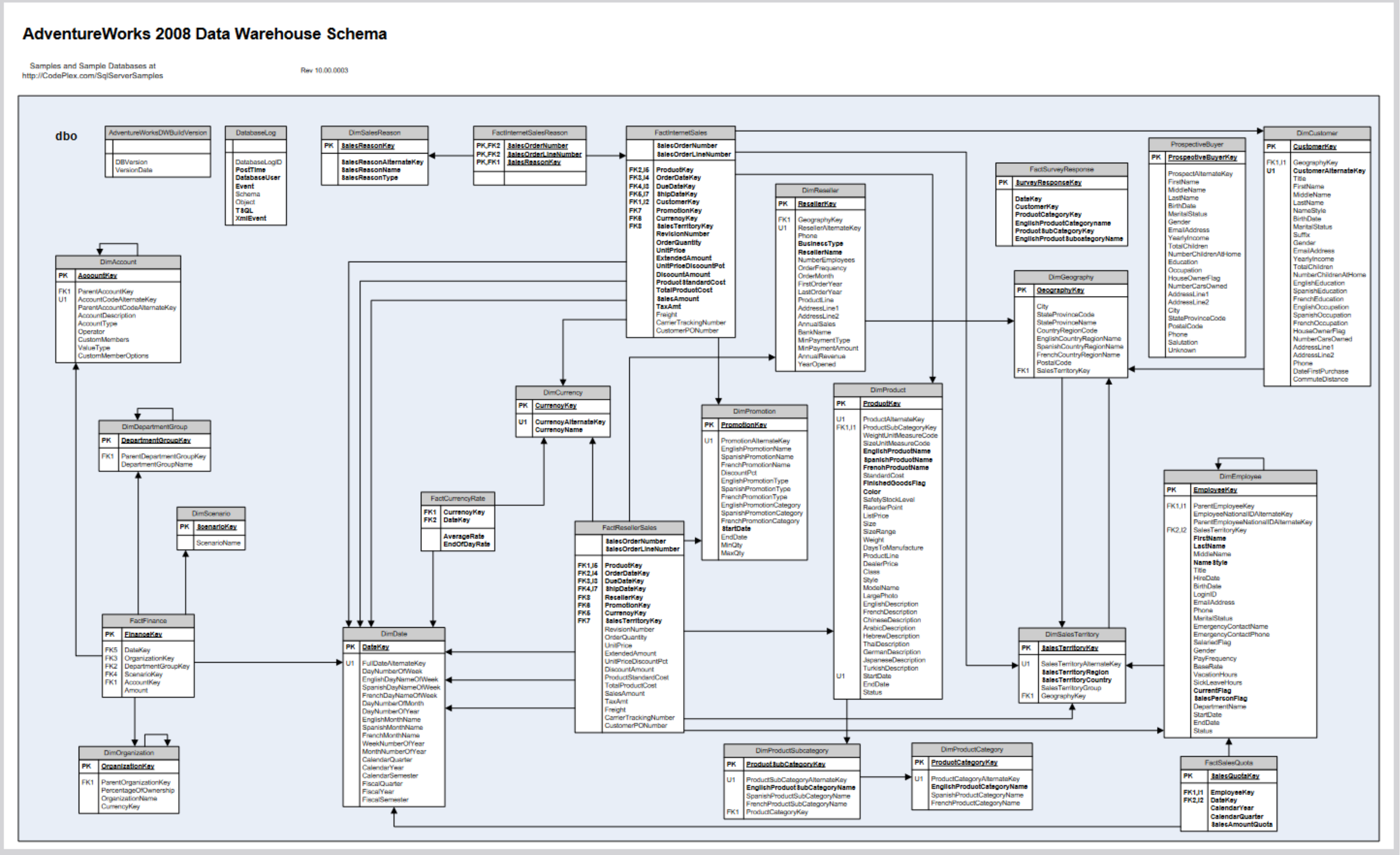
StudentID	Course
1001	Database Systems
1001	E-commerce
1002	E-commerce
…	…

SAMPLE OLTP DATABASE



https://akela.mendelu.cz/~jprich/vyuka/db2/AdventureWorks2008_db_diagram.pdf;

SAMPLE DATA WAREHOUSE



AdventureWorks 2008 Data Warehouse Schema

Samples and Sample Databases at
http://CodePlex.com/SqlServerSamples

Rev 10.00.0003

dbo

AdventureWorksDWBuildVersion

	DBVersion VersionDate

DatabaseLog

	DatabaseLogID PostTime DatabaseUser Event Schema Object TSQL XmlEvent

DimSalesReason

PK	SalesReasonKey
	SalesReasonAlternateKey SalesReasonName SalesReasonType

FactInternetSalesReason

PK,FK2 PK,FK2 PK,FK1	SalesOrderNumber SalesOrderLineNumber SalesReasonKey

FactInternetSales

	SalesOrderNumber SalesOrderLineNumber
FK2,I6 FK3,I4 FK4,I3 FK5,I7 FK1,I2 FK7 FK6 FK8	ProductKey OrderDateKey DueDateKey ShipDateKey CustomerKey PromotionKey CurrencyKey SalesTerritoryKey RevisionNumber OrderQuantity UnitPrice ExtendedAmount UnitPriceDiscountPct DiscountAmount ProductStandardCost TotalProductCost SalesAmount TaxAmt Freight CarrierTrackingNumber CustomerPONumber

FactSurveyResponse

PK	SurveyResponseKey
	DateKey CustomerKey ProductCategoryKey EnglishProductCategoryname ProductSubCategoryKey EnglishProductSubcategoryName

ProspectiveBuyer

PK	ProspectiveBuyerKey
	ProspectAlternateKey FirstName MiddleName LastName BirthDate MaritalStatus Gender EmailAddress YearlyIncome TotalChildren NumberChildrenAtHome Education Occupation HouseOwnerFlag NumberCarsOwned AddressLine1 AddressLine2 City StateProvinceCode PostalCode Phone Salutation Unknown

DimCustomer

PK	CustomerKey
FK1,I1 U1	GeographyKey CustomerAlternateKey Title FirstName MiddleName LastName NameStyle BirthDate MaritalStatus Suffix Gender EmailAddress YearlyIncome TotalChildren NumberChildrenAtHome EnglishEducation SpanishEducation FrenchEducation EnglishOccupation SpanishOccupation FrenchOccupation HouseOwnerFlag NumberCarsOwned AddressLine1 AddressLine2 Phone DateFirstPurchase CommuteDistance

DimAccount

PK	AccountKey
FK1 U1	ParentAccountKey AccountCodeAlternateKey ParentAccountCodeAlternateKey AccountDescription AccountType Operator CustomMembers ValueType CustomMemberOptions

DimReseller

PK	ResellerKey
FK1 U1	GeographyKey ResellerAlternateKey Phone BusinessType ResellerName NumberEmployees OrderFrequency OrderMonth FirstOrderYear LastOrderYear ProductLine AddressLine1 AddressLine2 AnnualSales BankName MinPaymentType MinPaymentAmount AnnualRevenue YearOpened

DimGeography

PK	GeographyKey
        FK1	City StateProvinceCode StateProvinceName CountryRegionCode EnglishCountryRegionName SpanishCountryRegionName FrenchCountryRegionName PostalCode SalesTerritoryKey

DimCurrency

PK	CurrencyKey
U1	CurrencyAlternateKey CurrencyName

DimPromotion

PK	PromotionKey
U1	PromotionAlternateKey EnglishPromotionName SpanishPromotionName FrenchPromotionName DiscountPct EnglishPromotionType SpanishPromotionType FrenchPromotionType EnglishPromotionCategory SpanishPromotionCategory FrenchPromotionCategory StartDate EndDate MinQty MaxQty

DimProduct

PK	ProductKey
U1 FK1,I1 U1	ProductAlternateKey ProductSubCategoryKey WeightUnitMeasureCode SizeUnitMeasureCode EnglishProductName SpanishProductName FrenchProductName StandardCost FinishedGoodsFlag Color SafetyStockLevel ReorderPoint ListPrice Size SizeRange Weight DaysToManufacture ProductLine DealerPrice Class Style ModelName LargePhoto EnglishDescription FrenchDescription ChineseDescription ArabicDescription HebrewDescription ThaiDescription GermanDescription JapaneseDescription TurkishDescription StartDate EndDate Status

DimDepartmentGroup

PK	DepartmentGroupKey
FK1	ParentDepartmentGroupKey DepartmentGroupName

DimScenario

PK	ScenarioKey
	ScenarioName

FactCurrencyRate

FK1 FK2	CurrencyKey DateKey
	AverageRate EndOfDayRate

FactResellerSales

	SalesOrderNumber SalesOrderLineNumber
FK1,I6 FK2,I4 FK3,I3 FK4,I7 FK8 FK6 FK5 FK7	ProductKey OrderDateKey DueDateKey ShipDateKey ResellerKey PromotionKey CurrencyKey SalesTerritoryKey RevisionNumber OrderQuantity UnitPrice ExtendedAmount UnitPriceDiscountPct DiscountAmount ProductStandardCost TotalProductCost SalesAmount TaxAmt Freight CarrierTrackingNumber CustomerPONumber

DimEmployee

PK	EmployeeKey
FK1,I1 FK2,I2	ParentEmployeeKey EmployeeNationalIDAlternateKey ParentEmployeeNationalIDAlternateKey SalesTerritoryKey FirstName LastName MiddleName NameStyle Title HireDate BirthDate LoginID EmailAddress Phone MaritalStatus EmergencyContactName EmergencyContactPhone SalariedFlag Gender PayFrequency BaseRate VacationHours SickLeaveHours CurrentFlag SalesPersonFlag DepartmentName StartDate EndDate Status

FactFinance

PK	FinanceKey
FK5 FK3 FK2 FK4 FK1	DateKey OrganizationKey DepartmentGroupKey ScenarioKey AccountKey Amount

DimDate

PK	DateKey
U1	FullDateAlternateKey DayNumberOfWeek EnglishDayNameOfWeek SpanishDayNameOfWeek FrenchDayNameOfWeek DayNumberOfMonth DayNumberOfYear EnglishMonthName SpanishMonthName FrenchMonthName WeekNumberOfYear MonthNumberOfYear CalendarQuarter CalendarYear CalendarSemester FiscalQuarter FiscalYear FiscalSemester

DimSalesTerritory

PK	SalesTerritoryKey
U1 FK1	SalesTerritoryAlternateKey SalesTerritoryRegion SalesTerritoryCountry SalesTerritoryGroup GeographyKey

DimOrganization

PK	OrganizationKey
FK1	ParentOrganizationKey PercentageOfOwnership OrganizationName CurrencyKey

DimProductSubcategory

PK	ProductSubCategoryKey
U1 FK1	ProductSubCategoryAlternateKey EnglishProductSubCategoryName SpanishProductSubCategoryName FrenchProductSubCategoryName ProductCategoryKey

DimProductCategory

PK	ProductCategoryKey
U1	ProductCategoryAlternateKey EnglishProductCategoryName SpanishProductCategoryName FrenchProductCategoryName

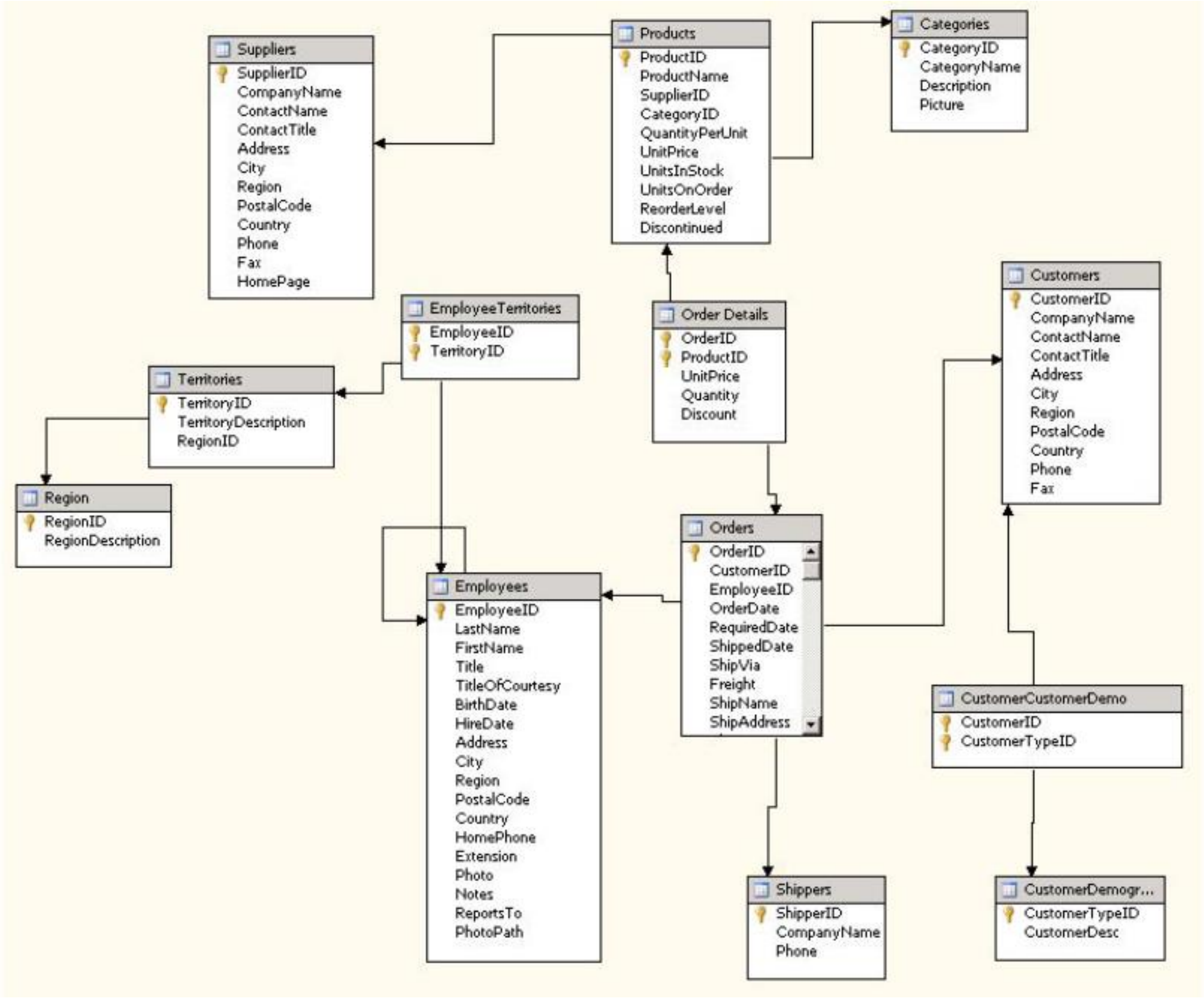
FactSalesQuota

PK	SalesQuotaKey
FK1,I1 FK2,I2	EmployeeKey DateKey CalendarYear CalendarQuarter SalesAmountQuota

https://moodle.usth.edu.vn/pluginfile.php/5907/mod_resource/content/1/AdventureWorksDW2008.pdf

DATASET INVESTIGATION

A SIMPLE DATABASE SCHEMA



Suppliers
- SupplierID
- CompanyName
- ContactName
- ContactTitle
- Address
- City
- Region
- PostalCode
- Country
- Phone
- Fax
- HomePage

Products
- ProductID
- ProductName
- SupplierID
- CategoryID
- QuantityPerUnit
- UnitPrice
- UnitsInStock
- UnitsOnOrder
- ReorderLevel
- Discontinued

Categories
- CategoryID
- CategoryName
- Description
- Picture

EmployeeTerritories
- EmployeeID
- TerritoryID

Order Details
- OrderID
- ProductID
- UnitPrice
- Quantity
- Discount

Customers
- CustomerID
- CompanyName
- ContactName
- ContactTitle
- Address
- City
- Region
- PostalCode
- Country
- Phone
- Fax

Territories
- TerritoryID
- TerritoryDescription
- RegionID

Region
- RegionID
- RegionDescription

Orders
- OrderID
- CustomerID
- EmployeeID
- OrderDate
- RequiredDate
- ShippedDate
- ShipVia
- Freight
- ShipName
- ShipAddress

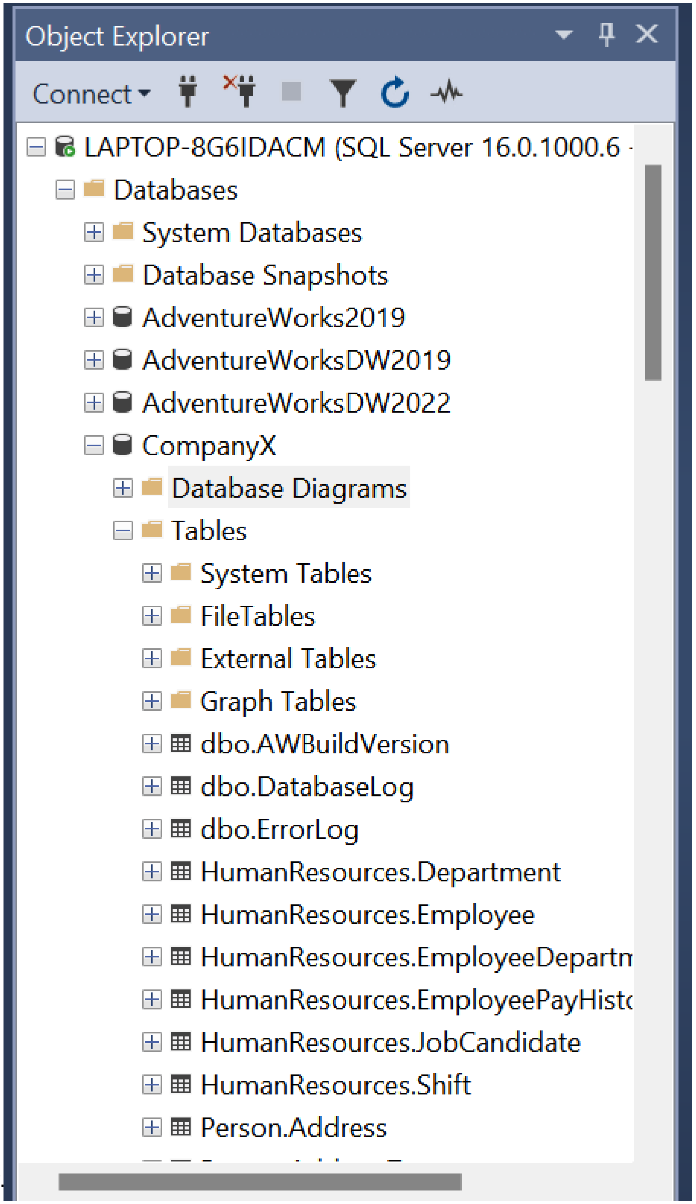
Employees
- EmployeeID
- LastName
- FirstName
- Title
- TitleOfCourtesy
- BirthDate
- HireDate
- Address
- City
- Region
- PostalCode
- Country
- HomePhone
- Extension
- Photo
- Notes
- ReportsTo
- PhotoPath

CustomerCustomerDemo
- CustomerID
- CustomerTypeID

Shippers
- ShipperID
- CompanyName
- Phone

CustomerDemogr...
- CustomerTypeID
- CustomerDesc

COMPANY X



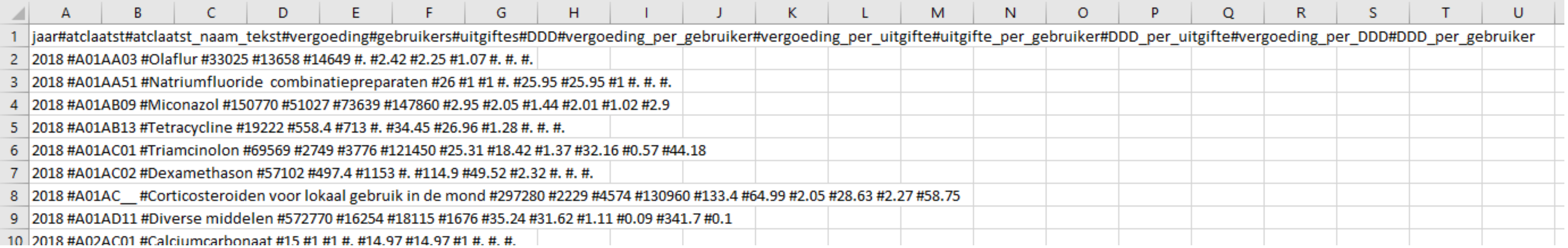
```
⊟ LAPTOP-8G6IDACM (SQL Server 16.0.1000.6 -
  ⊟ Databases
    ⊞ System Databases
    ⊞ Database Snapshots
    ⊞ AdventureWorks2019
    ⊞ AdventureWorksDW2019
    ⊞ AdventureWorksDW2022
    ⊟ CompanyX
      ⊞ Database Diagrams
      ⊟ Tables
        ⊞ System Tables
        ⊞ FileTables
        ⊞ External Tables
        ⊞ Graph Tables
        ⊞ dbo.AWBuildVersion
        ⊞ dbo.DatabaseLog
        ⊞ dbo.ErrorLog
        ⊞ HumanResources.Department
        ⊞ HumanResources.Employee
        ⊞ HumanResources.EmployeeDepartm
        ⊞ HumanResources.EmployeePayHist
        ⊞ HumanResources.JobCandidate
        ⊞ HumanResources.Shift
        ⊞ Person.Address
```

https://quotefancy.com/quote/1733770/T-B-Joshua-A-beautiful-life-does-not-just-happen-it-is-built; https://www.

WHAT

- What is it?
- What is the data about?
- What does it mean?
- …

➔ Understand the data objects (e.g., table, column, data) and the business

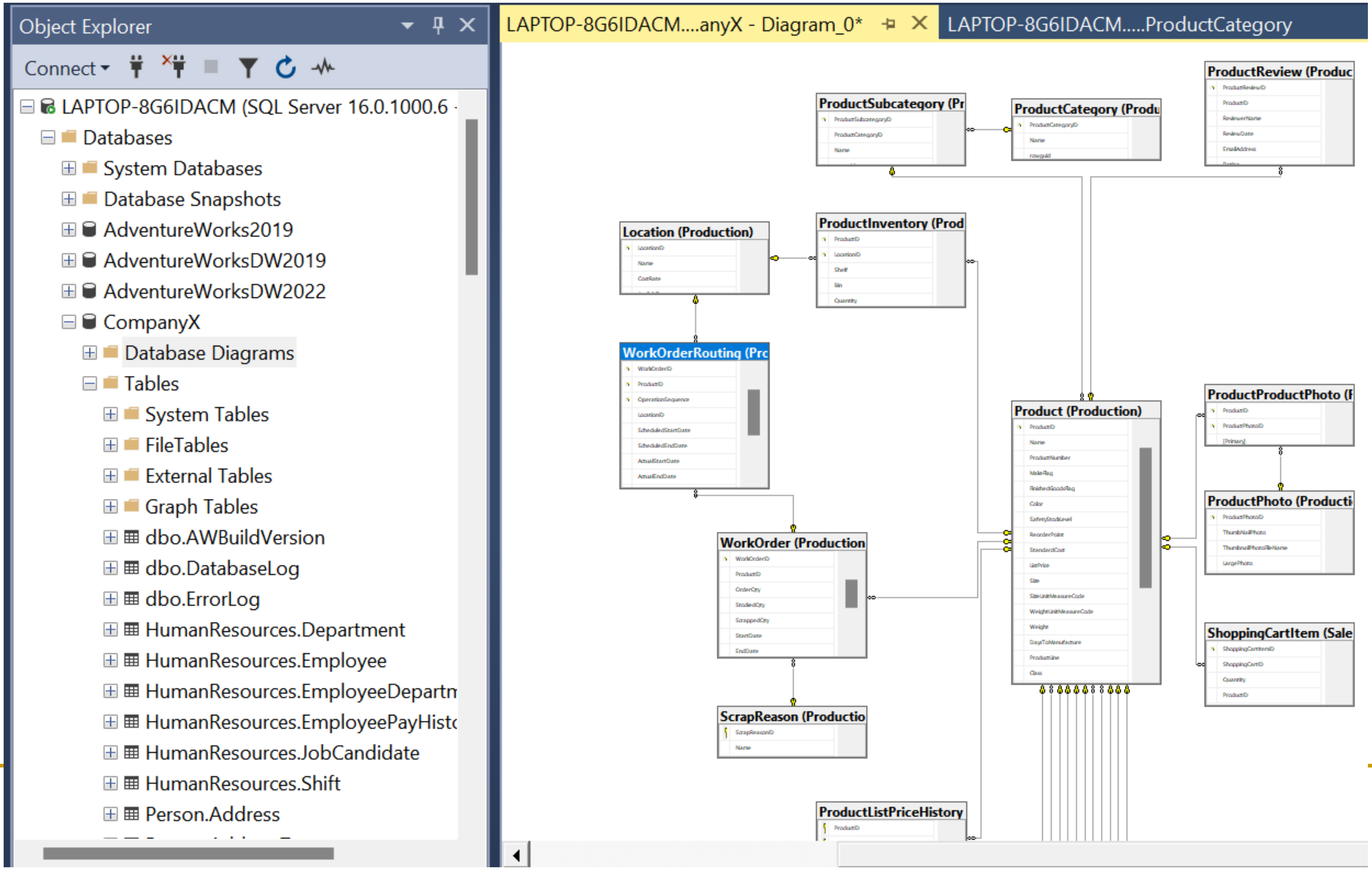


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	jaar#atclaatst#atclaatst_naam_tekst#vergoeding#gebruikers#uitgiftes#DDD#vergoeding_per_gebruiker#vergoeding_per_uitgifte#uitgifte_per_gebruiker#DDD_per_uitgifte#vergoeding_per_DDD#DDD_per_gebruiker																				
2	2018 #A01AA03 #Olaflur #33025 #13658 #14649 #. #2.42 #2.25 #1.07 #. #. #.																				
3	2018 #A01AA51 #Natriumfluoride combinatiepreparaten #26 #1 #1 #. #25.95 #25.95 #1 #. #. #.																				
4	2018 #A01AB09 #Miconazol #150770 #51027 #73639 #147860 #2.95 #2.05 #1.44 #2.01 #1.02 #2.9																				
5	2018 #A01AB13 #Tetracycline #19222 #558.4 #713 #. #34.45 #26.96 #1.28 #. #. #.																				
6	2018 #A01AC01 #Triamcinolon #69569 #2749 #3776 #121450 #25.31 #18.42 #1.37 #32.16 #0.57 #44.18																				
7	2018 #A01AC02 #Dexamethason #57102 #497.4 #1153 #. #114.9 #49.52 #2.32 #. #. #.																				
8	2018 #A01AC__ #Corticosteroiden voor lokaal gebruik in de mond #297280 #2229 #4574 #130960 #133.4 #64.99 #2.05 #28.63 #2.27 #58.75																				
9	2018 #A01AD11 #Diverse middelen #572770 #16254 #18115 #1676 #35.24 #31.62 #1.11 #0.09 #341.7 #0.1																				
10	2018 #A02AC01 #Calciumcarbonaat #15 #1 #1 #. #14.97 #14.97 #1 #. #. #.																				

https://www.gipdatabank.nl/servicepagina/open-data

FOR EXAMPLE

- It would be great if we have a database schema


Object Explorer
Connect
LAPTOP-8G6IDACM....anyX - Diagram_0*
LAPTOP-8G6IDACM.....ProductCategory

- LAPTOP-8G6IDACM (SQL Server 16.0.1000.6 -
  - Databases
    - System Databases
    - Database Snapshots
    - AdventureWorks2019
    - AdventureWorksDW2019
    - AdventureWorksDW2022
    - CompanyX
      - Database Diagrams
      - Tables
        - System Tables
        - FileTables
        - External Tables
        - Graph Tables
        - dbo.AWBuildVersion
        - dbo.DatabaseLog
        - dbo.ErrorLog
        - HumanResources.Department
        - HumanResources.Employee
        - HumanResources.EmployeeDepartm
        - HumanResources.EmployeePayHisto
        - HumanResources.JobCandidate
        - HumanResources.Shift
        - Person.Address

ProductSubcategory (Pr
ProductSubcategoryID
ProductCategoryID
Name

ProductCategory (Produ
ProductCategoryID
Name
rowguid

ProductReview (Produc
ProductReviewID
ProductID
ReviewerName
ReviewDate
EmailAddress

Location (Production)
LocationID
Name
CostRate

ProductInventory (Prod
ProductID
LocationID
Shelf
Bin
Quantity

WorkOrderRouting (Prc
WorkOrderID
ProductID
OperationSequence
LocationID
ScheduledStartDate
ScheduledEndDate
ActualStartDate
ActualEndDate

WorkOrder (Production
WorkOrderID
ProductID
OrderQty
StockedQty
ScrappedQty
StartDate
EndDate

ScrapReason (Productio
ScrapReasonID
Name

ProductListPriceHistory
ProductID

Product (Production)
ProductID
Name
ProductNumber
MakeFlag
FinishedGoodsFlag
Color
SafetyStockLevel
ReorderPoint
StandardCost
ListPrice
Size
SizeUnitMeasureCode
WeightUnitMeasureCode
Weight
DaysToManufacture
ProductLine
Class

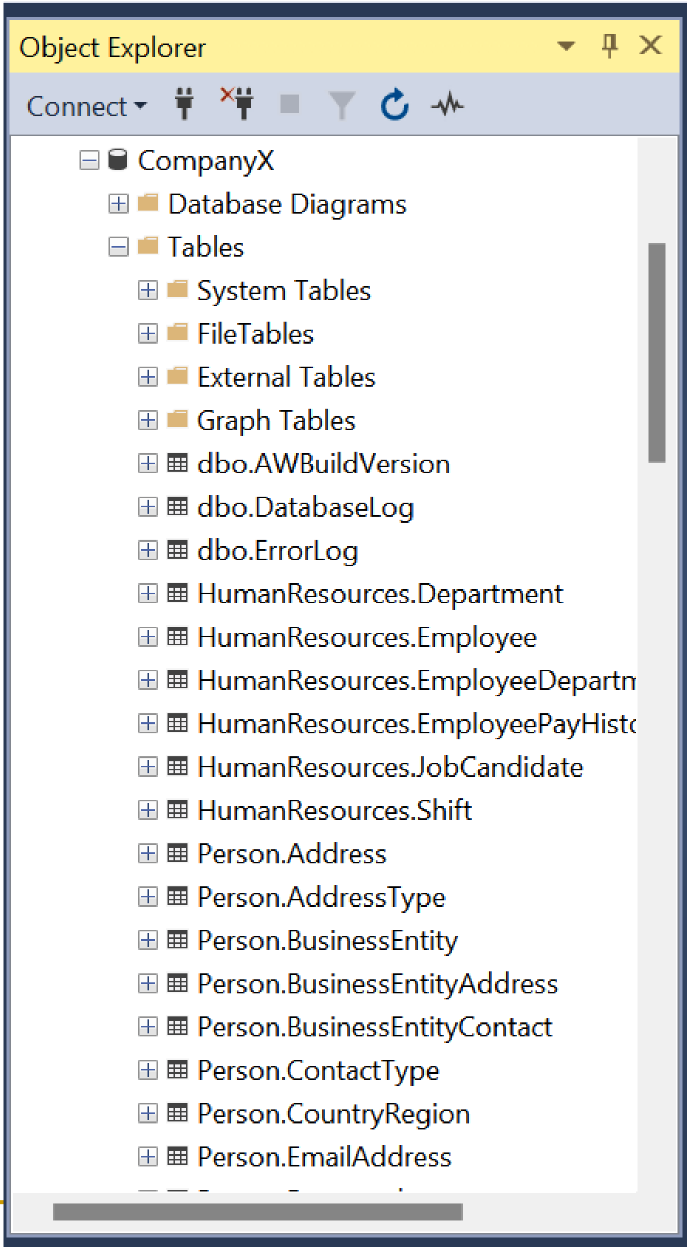
ProductProductPhoto (F
ProductID
ProductPhotoID
[Primary]

ProductPhoto (Producti
ProductPhotoID
ThumbNailPhoto
ThumbnailPhotoFileName
LargePhoto

ShoppingCartItem (Sale
ShoppingCartItemID
ShoppingCartID
Quantity
ProductID

FOR EXAMPLE

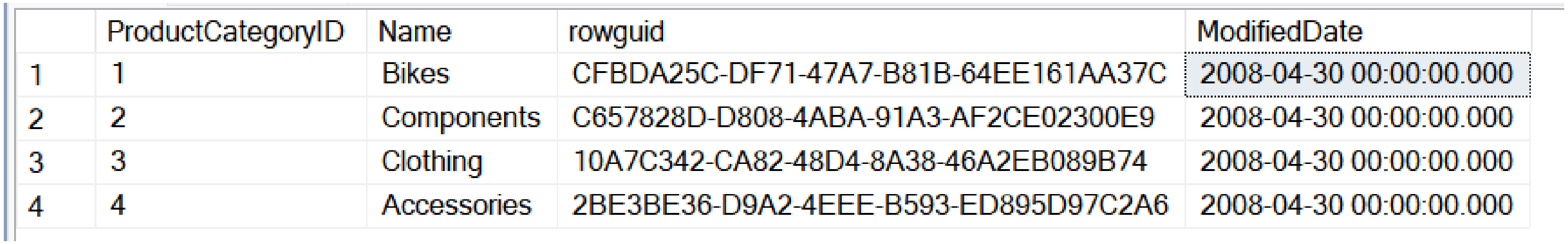
- Tables (71) with main schemas
  - HumanResources
  - Person
  - Production
  - Purchasing
  - Sales
- Views (20)
- Procedures (10)
- Functions (11)



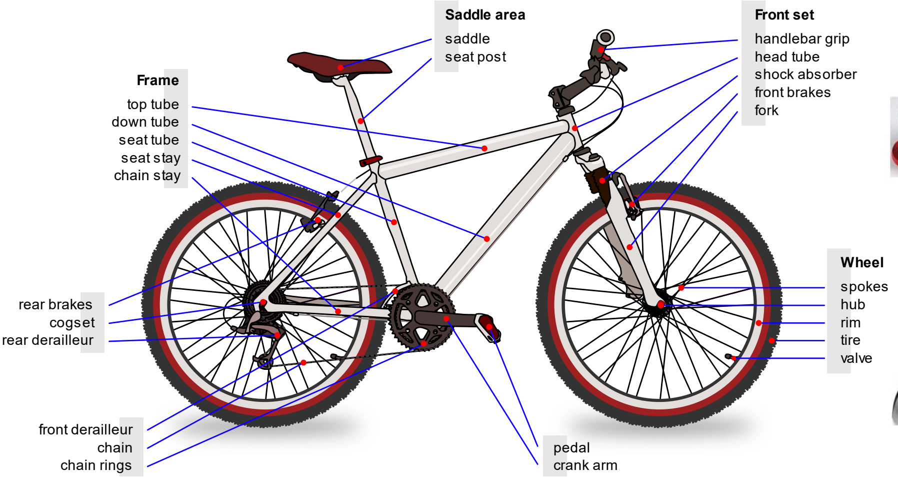
- CompanyX
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.AWBuildVersion
    - dbo.DatabaseLog
    - dbo.ErrorLog
    - HumanResources.Department
    - HumanResources.Employee
    - HumanResources.EmployeeDepartn
    - HumanResources.EmployeePayHisto
    - HumanResources.JobCandidate
    - HumanResources.Shift
    - Person.Address
    - Person.AddressType
    - Person.BusinessEntity
    - Person.BusinessEntityAddress
    - Person.BusinessEntityContact
    - Person.ContactType
    - Person.CountryRegion
    - Person.EmailAddress

FOR EXAMPLE

- The company X sells which products?

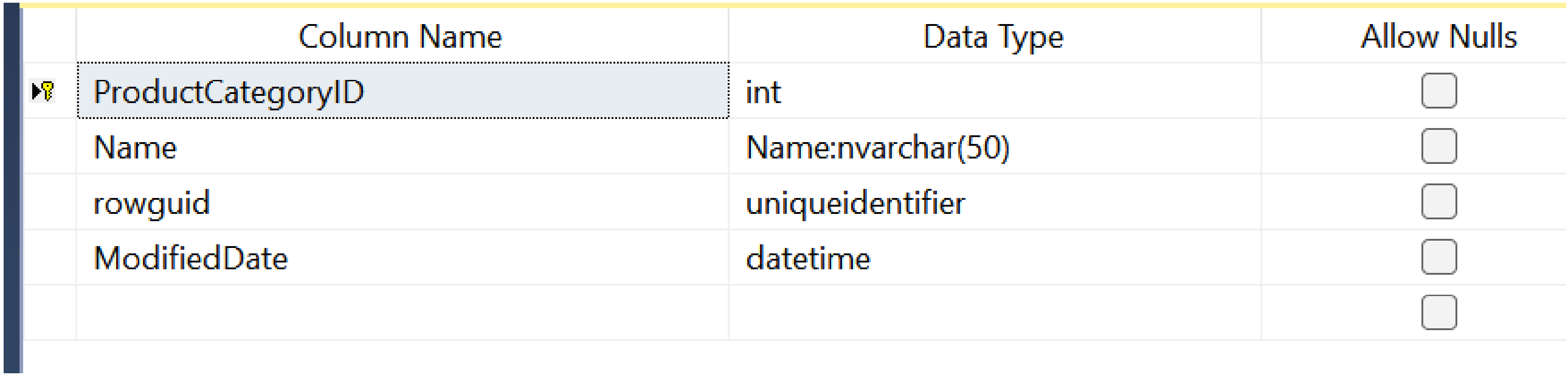


	ProductCategoryID	Name	rowguid	ModifiedDate
1	1	Bikes	CFBDA25C-DF71-47A7-B81B-64EE161AA37C	2008-04-30 00:00:00.000
2	2	Components	C657828D-D808-4ABA-91A3-AF2CE02300E9	2008-04-30 00:00:00.000
3	3	Clothing	10A7C342-CA82-48D4-8A38-46A2EB089B74	2008-04-30 00:00:00.000
4	4	Accessories	2BE3BE36-D9A2-4EEE-B593-ED895D97C2A6	2008-04-30 00:00:00.000

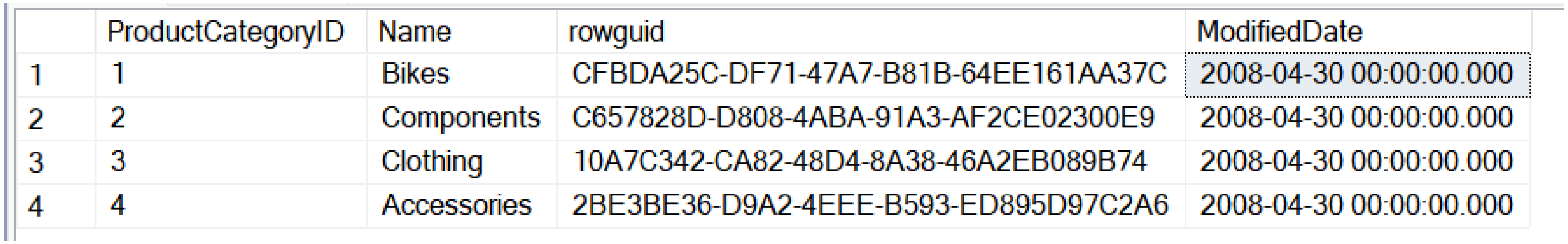


https://en.wikipedia.org/wiki/List_of_bicycle_parts; https://shop.northparkbikeshop.com/accessories; https://www.mtb-gear.nl/en/mountain-bike-clothing/

GUESS THE MEANING OF ATTRIBUTES



	Column Name	Data Type	Allow Nulls
▶🔑	ProductCategoryID	int	□
	Name	Name:nvarchar(50)	□
	rowguid	uniqueidentifier	□
	ModifiedDate	datetime	□
			□



	ProductCategoryID	Name	rowguid	ModifiedDate
1	1	Bikes	CFBDA25C-DF71-47A7-B81B-64EE161AA37C	2008-04-30 00:00:00.000
2	2	Components	C657828D-D808-4ABA-91A3-AF2CE02300E9	2008-04-30 00:00:00.000
3	3	Clothing	10A7C342-CA82-48D4-8A38-46A2EB089B74	2008-04-30 00:00:00.000
4	4	Accessories	2BE3BE36-D9A2-4EEE-B593-ED895D97C2A6	2008-04-30 00:00:00.000

OUR HYPOTHESIS

- Company X is a bicycle manufacturer, whose scenarios include
  - Manufacturing
  - Sales
  - Purchasing
  - Product Management
  - Contact Management
  - Human Resources

PRACTICE

- What are the sales markets of Company X?

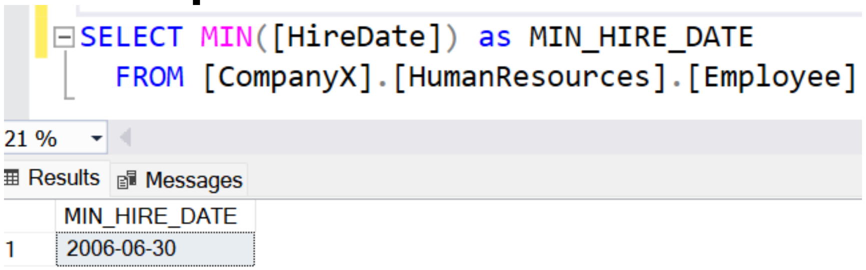
WHEN

- When was the data stored?
- Til when do we have the data?
  - Date/Timestamp columns
- …

➔ Understand the data periods

FOR EXAMPLE

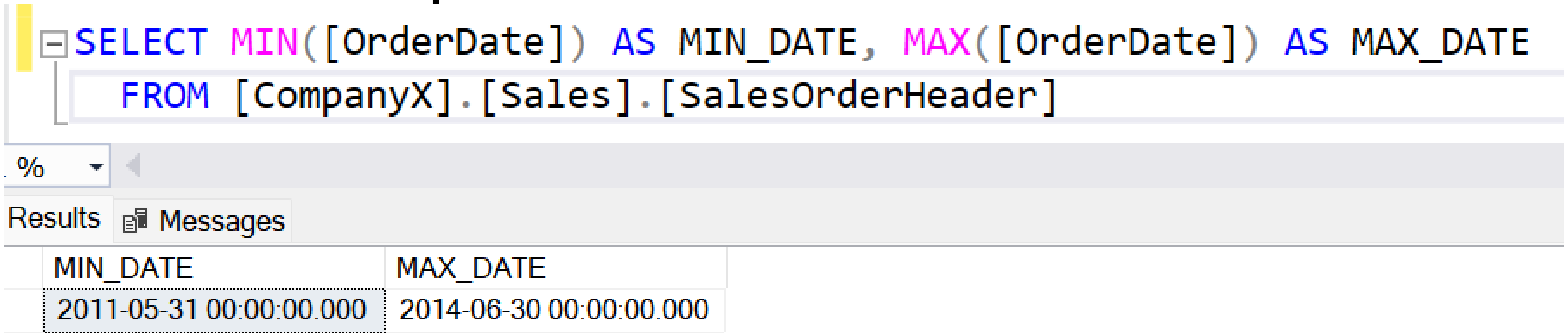
- The operation time in database



```
SELECT MIN([HireDate]) as MIN_HIRE_DATE
  FROM [CompanyX].[HumanResources].[Employee]
```

	MIN_HIRE_DATE
1	2006-06-30

- The sales periods



```
SELECT MIN([OrderDate]) AS MIN_DATE, MAX([OrderDate]) AS MAX_DATE
  FROM [CompanyX].[Sales].[SalesOrderHeader]
```

MIN_DATE	MAX_DATE
2011-05-31 00:00:00.000	2014-06-30 00:00:00.000

PRACTICE

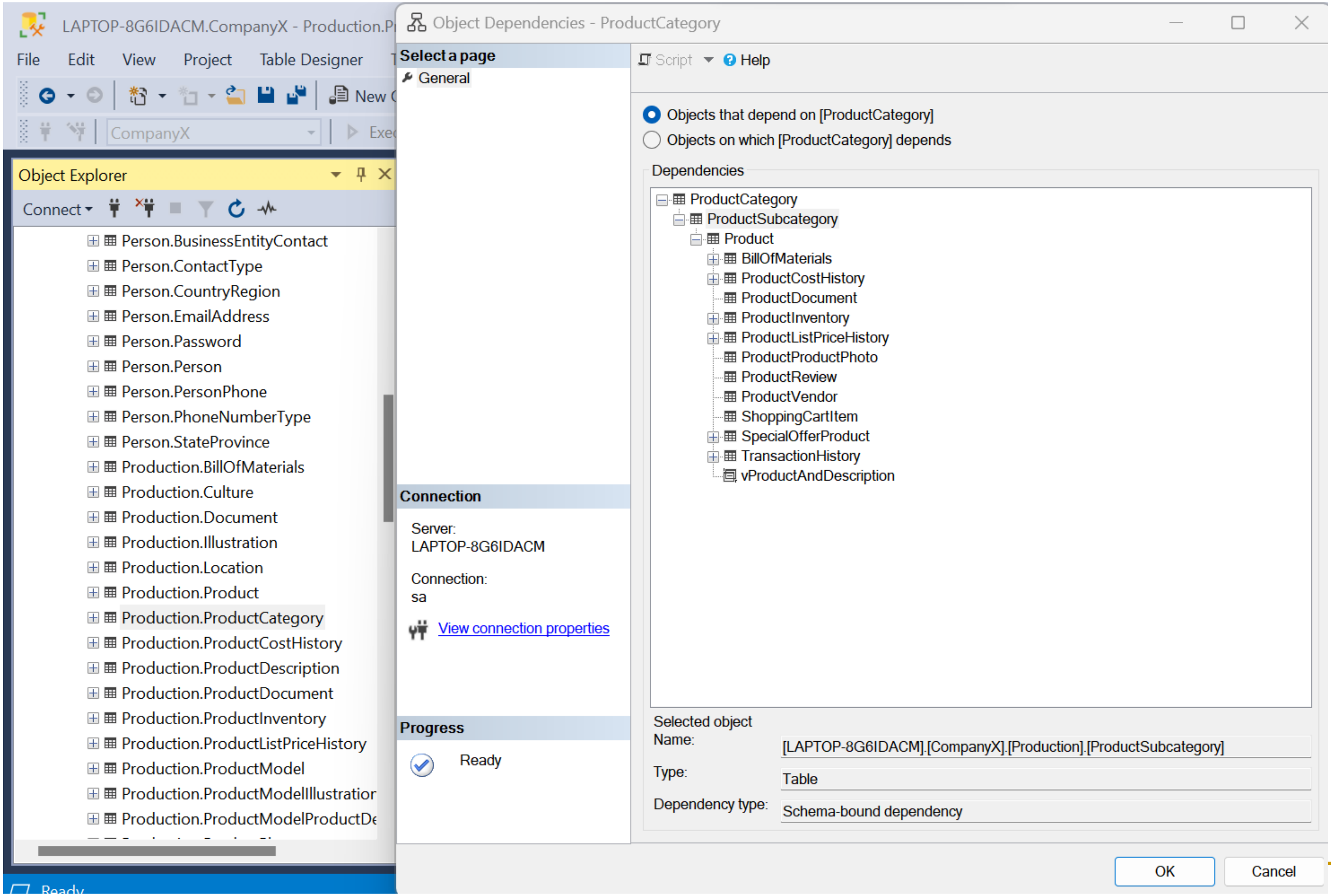
- When is the fist product sell start date?

HOW

- How are data related to one another?
- How do you know about data?
- How does the business work?
- …

➔ Understand the data relationships and further information such as business models and processes

TABLE DEPENDENCY



Object Explorer:

- Person.BusinessEntityContact
- Person.ContactType
- Person.CountryRegion
- Person.EmailAddress
- Person.Password
- Person.Person
- Person.PersonPhone
- Person.PhoneNumberType
- Person.StateProvince
- Production.BillOfMaterials
- Production.Culture
- Production.Document
- Production.Illustration
- Production.Location
- Production.Product
- Production.ProductCategory
- Production.ProductCostHistory
- Production.ProductDescription
- Production.ProductDocument
- Production.ProductInventory
- Production.ProductListPriceHistory
- Production.ProductModel
- Production.ProductModelIllustratior
- Production.ProductModelProductDe

● Objects that depend on [ProductCategory]

○ Objects on which [ProductCategory] depends

Dependencies

- ProductCategory
  - ProductSubcategory
    - Product
      - BillOfMaterials
      - ProductCostHistory
      - ProductDocument
      - ProductInventory
      - ProductListPriceHistory
      - ProductProductPhoto
      - ProductReview
      - ProductVendor
      - ShoppingCartItem
      - SpecialOfferProduct
      - TransactionHistory
      - vProductAndDescription

Selected object

Name:	[LAPTOP-8G6IDACM].[CompanyX].[Production].[ProductSubcategory]
Type:	Table
Dependency type:	Schema-bound dependency

Server:
LAPTOP-8G6IDACM

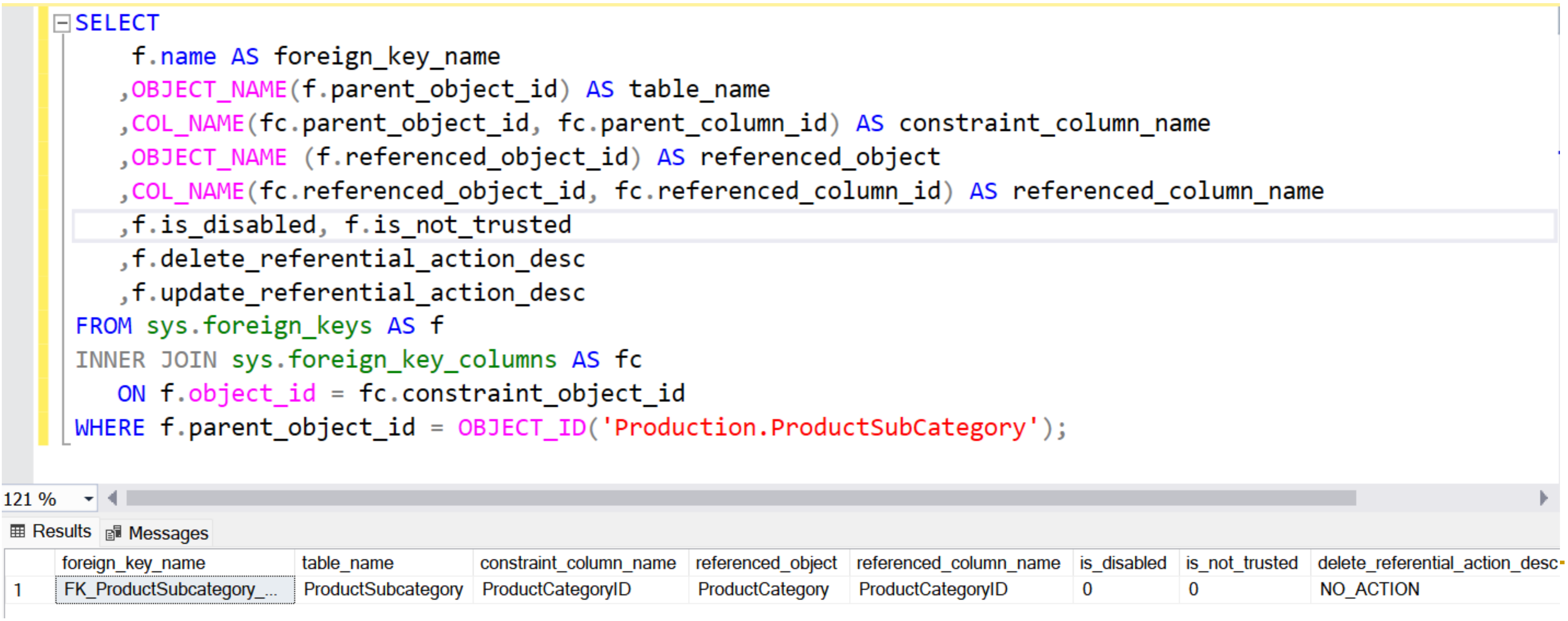
Connection:
sa

View connection properties

Ready

TABLE DEPENDENCY

- Linked to
  - None
- Linked from
  - Production.ProductSubCategory

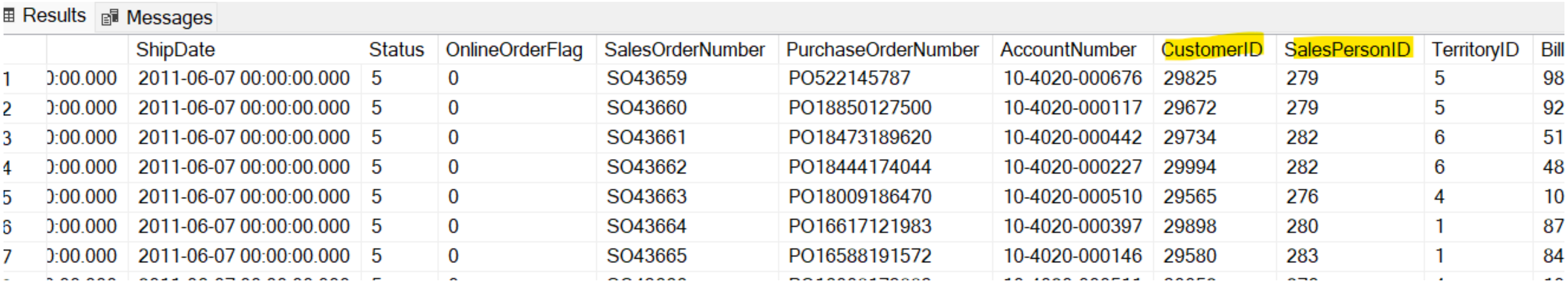


```
SELECT
     f.name AS foreign_key_name
    ,OBJECT_NAME(f.parent_object_id) AS table_name
    ,COL_NAME(fc.parent_object_id, fc.parent_column_id) AS constraint_column_name
    ,OBJECT_NAME (f.referenced_object_id) AS referenced_object
    ,COL_NAME(fc.referenced_object_id, fc.referenced_column_id) AS referenced_column_name
    ,f.is_disabled, f.is_not_trusted
    ,f.delete_referential_action_desc
    ,f.update_referential_action_desc
FROM sys.foreign_keys AS f
INNER JOIN sys.foreign_key_columns AS fc
   ON f.object_id = fc.constraint_object_id
WHERE f.parent_object_id = OBJECT_ID('Production.ProductSubCategory');
```

	foreign_key_name	table_name	constraint_column_name	referenced_object	referenced_column_name	is_disabled	is_not_trusted	delete_referential_action_desc
1	FK_ProductSubcategory_...	ProductSubcategory	ProductCategoryID	ProductCategory	ProductCategoryID	0	0	NO_ACTION

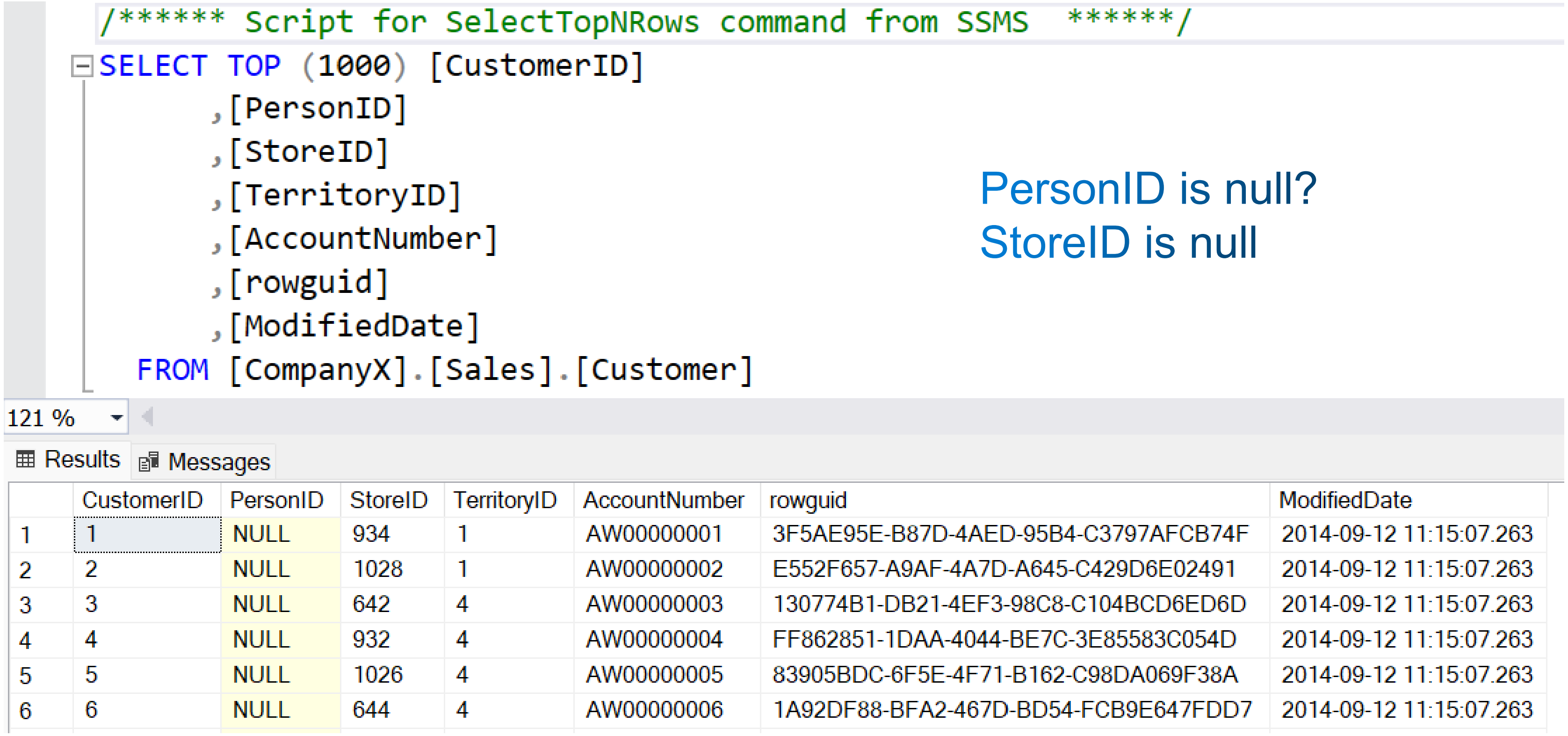
FOR EXAMPLE

- Who is the customers of company X?
  - Customer?
  - Store?
  - Salesperson?
  - And then... Business Entity?



		ShipDate	Status	OnlineOrderFlag	SalesOrderNumber	PurchaseOrderNumber	AccountNumber	CustomerID	SalesPersonID	TerritoryID	Bill
1	):00.000	2011-06-07 00:00:00.000	5	0	SO43659	PO522145787	10-4020-000676	29825	279	5	98
2	):00.000	2011-06-07 00:00:00.000	5	0	SO43660	PO18850127500	10-4020-000117	29672	279	5	92
3	):00.000	2011-06-07 00:00:00.000	5	0	SO43661	PO18473189620	10-4020-000442	29734	282	6	51
4	):00.000	2011-06-07 00:00:00.000	5	0	SO43662	PO18444174044	10-4020-000227	29994	282	6	48
5	):00.000	2011-06-07 00:00:00.000	5	0	SO43663	PO18009186470	10-4020-000510	29565	276	4	10
6	):00.000	2011-06-07 00:00:00.000	5	0	SO43664	PO16617121983	10-4020-000397	29898	280	1	87
7	):00.000	2011-06-07 00:00:00.000	5	0	SO43665	PO16588191572	10-4020-000146	29580	283	1	84

SALES.CUSTOMER



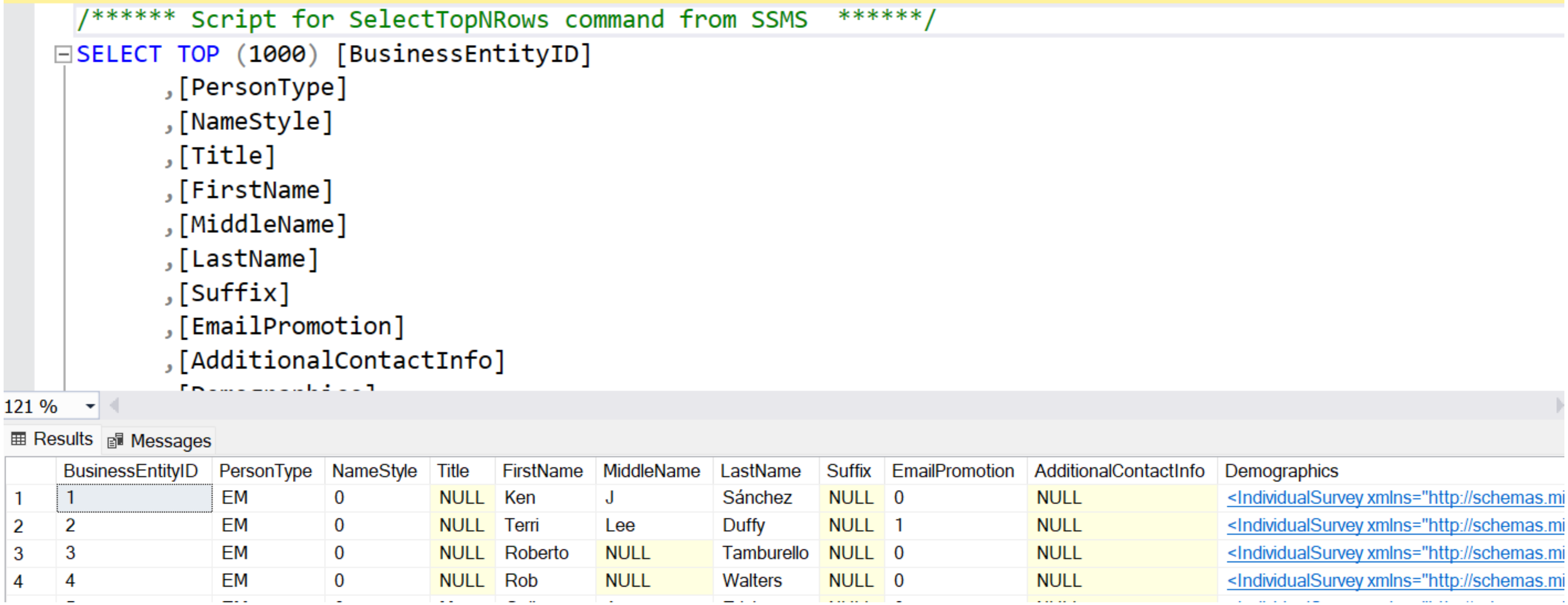
```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [CustomerID]
      ,[PersonID]
      ,[StoreID]
      ,[TerritoryID]
      ,[AccountNumber]
      ,[rowguid]
      ,[ModifiedDate]
  FROM [CompanyX].[Sales].[Customer]
```

	CustomerID	PersonID	StoreID	TerritoryID	AccountNumber	rowguid	ModifiedDate
1	1	NULL	934	1	AW00000001	3F5AE95E-B87D-4AED-95B4-C3797AFCB74F	2014-09-12 11:15:07.263
2	2	NULL	1028	1	AW00000002	E552F657-A9AF-4A7D-A645-C429D6E02491	2014-09-12 11:15:07.263
3	3	NULL	642	4	AW00000003	130774B1-DB21-4EF3-98C8-C104BCD6ED6D	2014-09-12 11:15:07.263
4	4	NULL	932	4	AW00000004	FF862851-1DAA-4044-BE7C-3E85583C054D	2014-09-12 11:15:07.263
5	5	NULL	1026	4	AW00000005	83905BDC-6F5E-4F71-B162-C98DA069F38A	2014-09-12 11:15:07.263
6	6	NULL	644	4	AW00000006	1A92DF88-BFA2-467D-BD54-FCB9E647FDD7	2014-09-12 11:15:07.263

PersonID is null?

StoreID is null

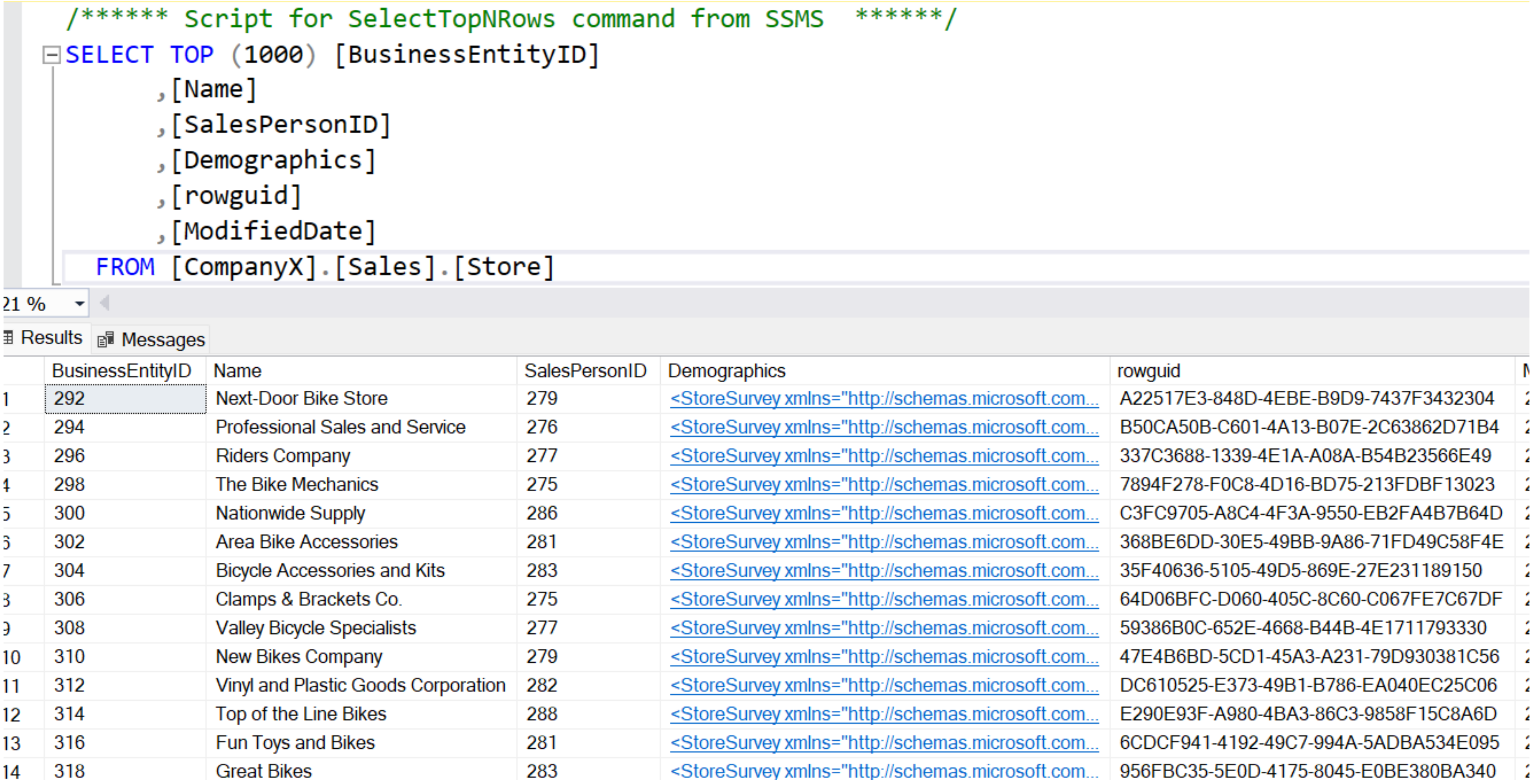
PERSON.PERSON



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [BusinessEntityID]
      ,[PersonType]
      ,[NameStyle]
      ,[Title]
      ,[FirstName]
      ,[MiddleName]
      ,[LastName]
      ,[Suffix]
      ,[EmailPromotion]
      ,[AdditionalContactInfo]
      ,[Demographics]
```

	BusinessEntityID	PersonType	NameStyle	Title	FirstName	MiddleName	LastName	Suffix	EmailPromotion	AdditionalContactInfo	Demographics
1	1	EM	0	NULL	Ken	J	Sánchez	NULL	0	NULL	<IndividualSurvey xmlns="http://schemas.mi
2	2	EM	0	NULL	Terri	Lee	Duffy	NULL	1	NULL	<IndividualSurvey xmlns="http://schemas.mi
3	3	EM	0	NULL	Roberto	NULL	Tamburello	NULL	0	NULL	<IndividualSurvey xmlns="http://schemas.mi
4	4	EM	0	NULL	Rob	NULL	Walters	NULL	0	NULL	<IndividualSurvey xmlns="http://schemas.mi

SALES.STORE



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [BusinessEntityID]
      ,[Name]
      ,[SalesPersonID]
      ,[Demographics]
      ,[rowguid]
      ,[ModifiedDate]
  FROM [CompanyX].[Sales].[Store]
```

	BusinessEntityID	Name	SalesPersonID	Demographics	rowguid
1	292	Next-Door Bike Store	279	<StoreSurvey xmlns="http://schemas.microsoft.com...	A22517E3-848D-4EBE-B9D9-7437F3432304
2	294	Professional Sales and Service	276	<StoreSurvey xmlns="http://schemas.microsoft.com...	B50CA50B-C601-4A13-B07E-2C63862D71B4
3	296	Riders Company	277	<StoreSurvey xmlns="http://schemas.microsoft.com...	337C3688-1339-4E1A-A08A-B54B23566E49
4	298	The Bike Mechanics	275	<StoreSurvey xmlns="http://schemas.microsoft.com...	7894F278-F0C8-4D16-BD75-213FDBF13023
5	300	Nationwide Supply	286	<StoreSurvey xmlns="http://schemas.microsoft.com...	C3FC9705-A8C4-4F3A-9550-EB2FA4B7B64D
6	302	Area Bike Accessories	281	<StoreSurvey xmlns="http://schemas.microsoft.com...	368BE6DD-30E5-49BB-9A86-71FD49C58F4E
7	304	Bicycle Accessories and Kits	283	<StoreSurvey xmlns="http://schemas.microsoft.com...	35F40636-5105-49D5-869E-27E231189150
8	306	Clamps & Brackets Co.	275	<StoreSurvey xmlns="http://schemas.microsoft.com...	64D06BFC-D060-405C-8C60-C067FE7C67DF
9	308	Valley Bicycle Specialists	277	<StoreSurvey xmlns="http://schemas.microsoft.com...	59386B0C-652E-4668-B44B-4E1711793330
10	310	New Bikes Company	279	<StoreSurvey xmlns="http://schemas.microsoft.com...	47E4B6BD-5CD1-45A3-A231-79D930381C56
11	312	Vinyl and Plastic Goods Corporation	282	<StoreSurvey xmlns="http://schemas.microsoft.com...	DC610525-E373-49B1-B786-EA040EC25C06
12	314	Top of the Line Bikes	288	<StoreSurvey xmlns="http://schemas.microsoft.com...	E290E93F-A980-4BA3-86C3-9858F15C8A6D
13	316	Fun Toys and Bikes	281	<StoreSurvey xmlns="http://schemas.microsoft.com...	6CDCF941-4192-49C7-994A-5ADBA534E095
14	318	Great Bikes	283	<StoreSurvey xmlns="http://schemas.microsoft.com...	956FBC35-5E0D-4175-8045-E0BE380BA340

PRACTICE

- How about SalesPerson?
- How can you find the customer demographic?

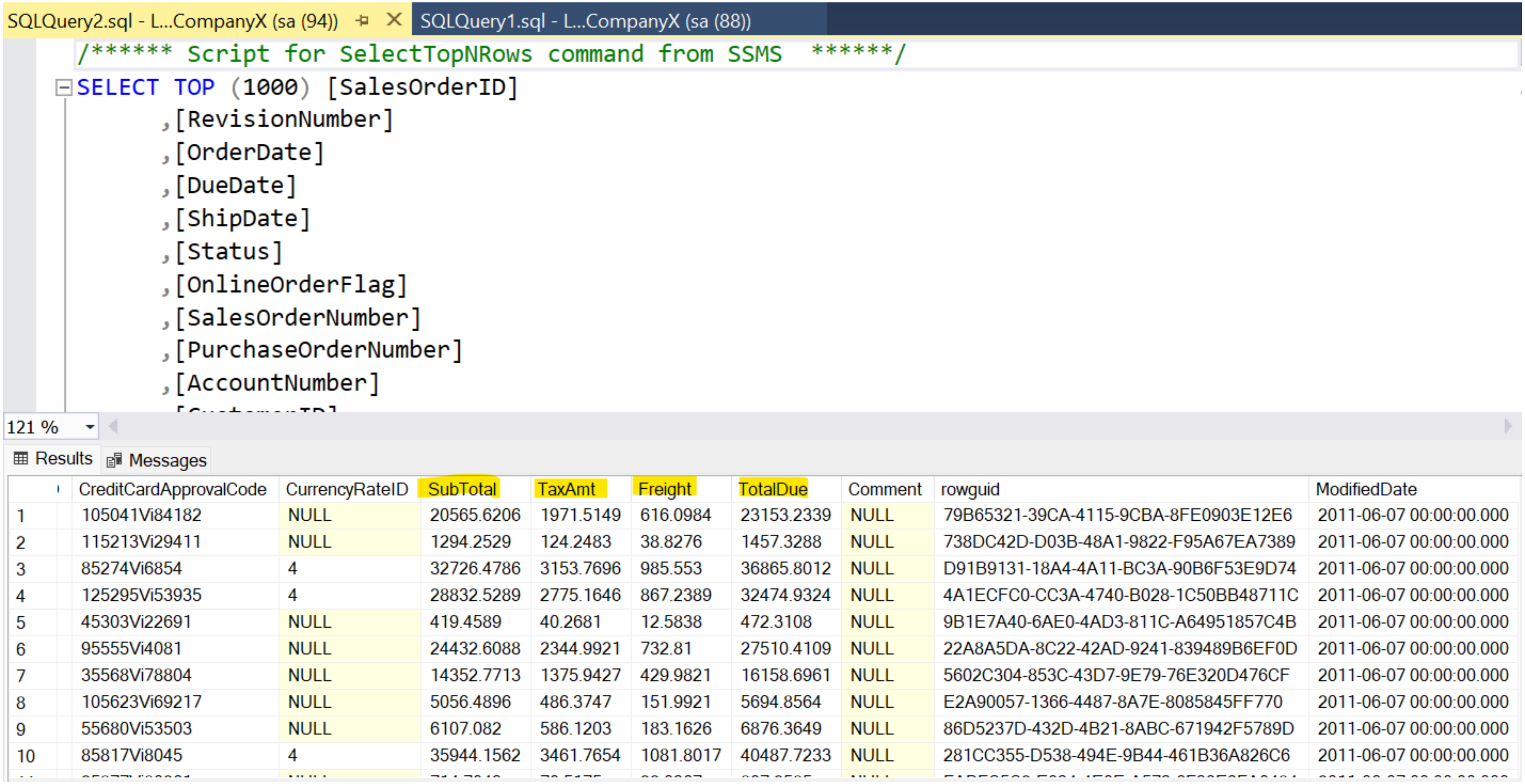
WHY

- Why do we have that value?
- Why would we have that assumption?
- …

➔ Understand the semantics behind like the rules, formulae, and consolidate our hypothesis about business models and processes

FOR EXAMPLE

- In Sales.OrderHeader
  - Can you guess why we would have the highlight values?

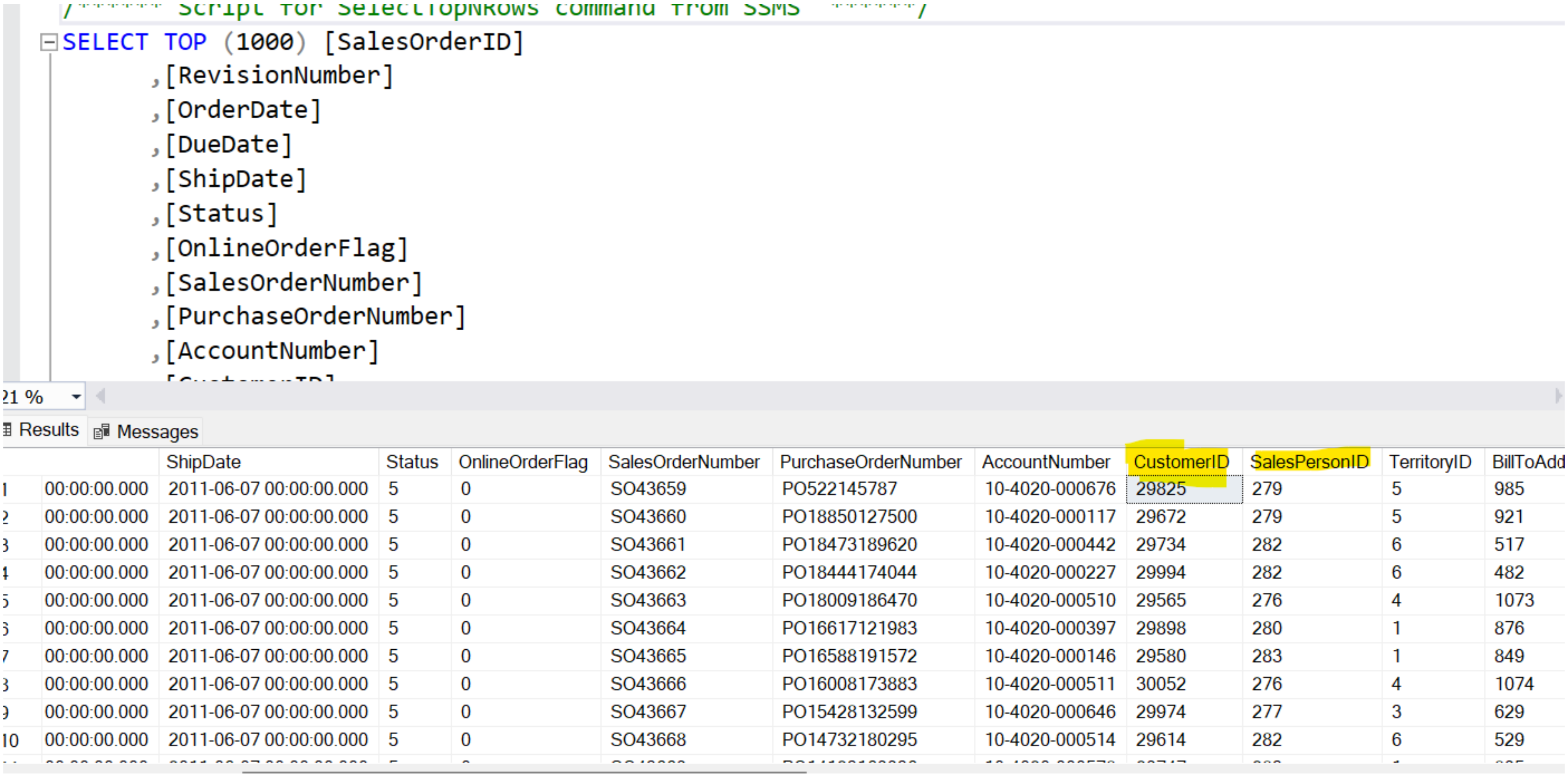

SQLQuery2.sql - L...CompanyX (sa (94))
SQLQuery1.sql - L...CompanyX (sa (88))
121 %
Results
Messages

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [SalesOrderID]
      ,[RevisionNumber]
      ,[OrderDate]
      ,[DueDate]
      ,[ShipDate]
      ,[Status]
      ,[OnlineOrderFlag]
      ,[SalesOrderNumber]
      ,[PurchaseOrderNumber]
      ,[AccountNumber]
      ,[CustomerID]
```

	CreditCardApprovalCode	CurrencyRateID	SubTotal	TaxAmt	Freight	TotalDue	Comment	rowguid	ModifiedDate
1	105041Vi84182	NULL	20565.6206	1971.5149	616.0984	23153.2339	NULL	79B65321-39CA-4115-9CBA-8FE0903E12E6	2011-06-07 00:00:00.000
2	115213Vi29411	NULL	1294.2529	124.2483	38.8276	1457.3288	NULL	738DC42D-D03B-48A1-9822-F95A67EA7389	2011-06-07 00:00:00.000
3	85274Vi6854	4	32726.4786	3153.7696	985.553	36865.8012	NULL	D91B9131-18A4-4A11-BC3A-90B6F53E9D74	2011-06-07 00:00:00.000
4	125295Vi53935	4	28832.5289	2775.1646	867.2389	32474.9324	NULL	4A1ECFC0-CC3A-4740-B028-1C50BB48711C	2011-06-07 00:00:00.000
5	45303Vi22691	NULL	419.4589	40.2681	12.5838	472.3108	NULL	9B1E7A40-6AE0-4AD3-811C-A64951857C4B	2011-06-07 00:00:00.000
6	95555Vi4081	NULL	24432.6088	2344.9921	732.81	27510.4109	NULL	22A8A5DA-8C22-42AD-9241-839489B6EF0D	2011-06-07 00:00:00.000
7	35568Vi78804	NULL	14352.7713	1375.9427	429.9821	16158.6961	NULL	5602C304-853C-43D7-9E79-76E320D476CF	2011-06-07 00:00:00.000
8	105623Vi69217	NULL	5056.4896	486.3747	151.9921	5694.8564	NULL	E2A90057-1366-4487-8A7E-8085845FF770	2011-06-07 00:00:00.000
9	55680Vi53503	NULL	6107.082	586.1203	183.1626	6876.3649	NULL	86D5237D-432D-4B21-8ABC-671942F5789D	2011-06-07 00:00:00.000
10	85817Vi8045	4	35944.1562	3461.7654	1081.8017	40487.7233	NULL	281CC355-D538-494E-9B44-461B36A826C6	2011-06-07 00:00:00.000

PRACTICE

- Why would we have CustomerID and SalesPersonID in Sales.OrderHeader?



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP (1000) [SalesOrderID]
      ,[RevisionNumber]
      ,[OrderDate]
      ,[DueDate]
      ,[ShipDate]
      ,[Status]
      ,[OnlineOrderFlag]
      ,[SalesOrderNumber]
      ,[PurchaseOrderNumber]
      ,[AccountNumber]
      ,[CustomerID]
```

		ShipDate	Status	OnlineOrderFlag	SalesOrderNumber	PurchaseOrderNumber	AccountNumber	CustomerID	SalesPersonID	TerritoryID	BillToAdd
1	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43659	PO522145787	10-4020-000676	29825	279	5	985
2	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43660	PO18850127500	10-4020-000117	29672	279	5	921
3	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43661	PO18473189620	10-4020-000442	29734	282	6	517
4	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43662	PO18444174044	10-4020-000227	29994	282	6	482
5	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43663	PO18009186470	10-4020-000510	29565	276	4	1073
6	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43664	PO16617121983	10-4020-000397	29898	280	1	876
7	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43665	PO16588191572	10-4020-000146	29580	283	1	849
8	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43666	PO16008173883	10-4020-000511	30052	276	4	1074
9	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43667	PO15428132599	10-4020-000646	29974	277	3	629
10	00:00:00.000	2011-06-07 00:00:00.000	5	0	SO43668	PO14732180295	10-4020-000514	29614	282	6	529

PRACTICE

- How to get product information such as name, price, category?

WHAT'S MORE...

- Do you find any weird data values?

Game on !!!

https://www.clipartmax.com/middle/m2i8Z5i8m2N4i8G6_smiley-faces-emoticon-smileys-emojis-video-games-gaming-smiley-face-png/

SUMMARY

- The need of business with BI systems
- Two different systems: OLTP and OLAP
- Initial dataset investigation
  - Self-exploration with WH questions and hypotheses

QUESTIONS AND ANSWERS

Picture from: http://philadelphiasculpturegym.blogspot.com/2013/09/save-date-free-talk-and-q-on-affordable.html

REFERENCES

1. Tobias Zwingmann, “AI-Powered Business Intelligence,” Kindle Edition, O'reilly Press, 2022
2. Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques,” Third Edition, Morgan Kaufmann Publishers, 2012.
3. David L. Olson, Dursun Delen, “Advanced Data Mining Techniques,” Springer-Verlag, 2008.
4. Jeen Su Lim, John Heinrichs, “Digital Business Intelligence Management with Big Data Analytics," Kindle Edition, O'reilly Press, 2021.
5. William H. Inmon, “Building the Data Warehouse,” Fourth Edition, Wiley Publishing, Inc., 2005.
6. R. Kimball, M. Ross, "The Data Warehouse ToolKit," 3rd Edition, Wiley Publishing, Inc., 2013.
7. Turban, E., Aronson,J.E., “Decision Support Systems and Intelligent Systems” - 7th Edition, Prentice-Hall, 2005.
8. Ramesh Sharda, Dursun Delen, Efraim Turban, “Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support,” 7th Edition, Pearson Education, Inc., 2020.