Assignment 01: Data Profiling

Part 1: Overview

One critical activity of data warehousing is the ability to profile your organization's data and business processes. Data profiling is the activity of learning about the characteristics, capabilities, and quality of your data and metadata (schema). Getting intimate with your data is an important step in building the data warehouse because you must understand what you have to work with before you can leverage it as part of your data warehouse solution. This assignment will help you discover how to understand your organization's data so that you can figure out how to properly build a data warehouse around it.

Goals

Specifically the goals are to

- Get you familiar with how to use the tools provided in the course.
- Help you understand how to profile data—reading unknown database schemas/data and deducing intent behind table designs and data.
- Get you comfortable writing SQL SELECT queries against a relational database, as a means to understand the capabilities of the data and metadata that you have.

Effort

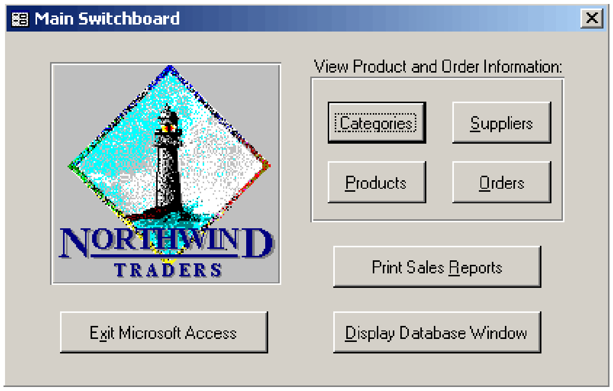
This assignment should be done individually. The work you complete should represent your own ability.

Technical Requirements

To complete this assignment you will need the following:

- Microsoft SQL Server Management Studio.
- Access to the **ist-cs-dw1.ad.syr.edu** Microsoft SQL Server. The connection procedure is explained in another document.

The Northwind Traders Case Study

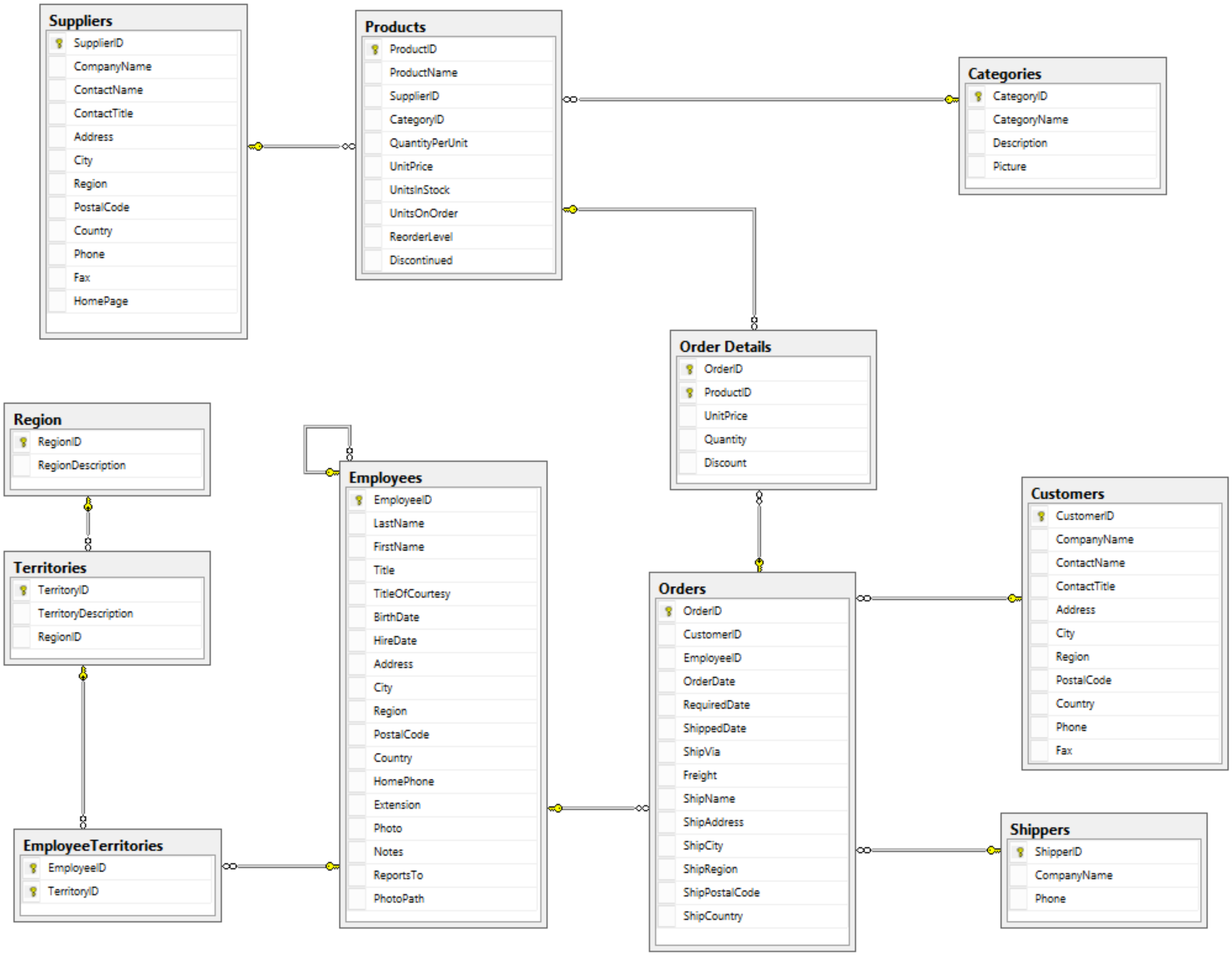


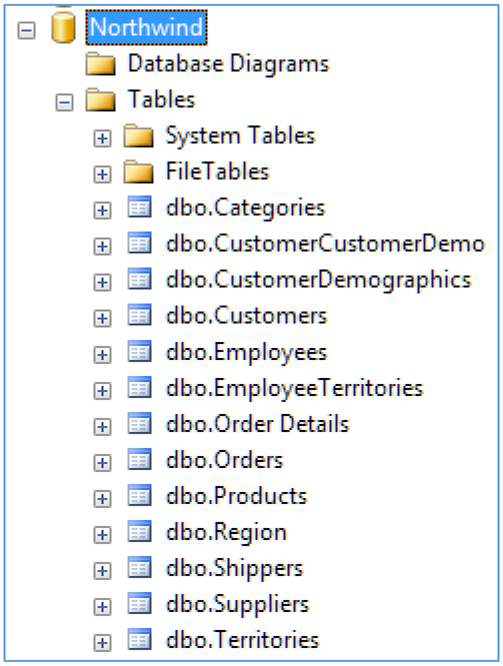
Northwind Traders is a fictitious importer and exporter of specialty foods from around the world. It was created by Microsoft as a sample operational database to use with their database products, such as Microsoft Access and SQL Server.

We'll use this database as a case study for building a data warehouse. You will use this same database throughout each of your assignments in this course and will get very intimate with the data and table design.

The Northwind Data Model

Below is a screen shot of the internal model for the Northwind Traders database. Use this diagram as a reference for understanding the structure of the Northwind data and building your data warehouse.





```
⊟ Northwind
    Database Diagrams
  ⊟ Tables
    ⊞ System Tables
    ⊞ FileTables
    ⊞ dbo.Categories
    ⊞ dbo.CustomerCustomerDemo
    ⊞ dbo.CustomerDemographics
    ⊞ dbo.Customers
    ⊞ dbo.Employees
    ⊞ dbo.EmployeeTerritories
    ⊞ dbo.Order Details
    ⊞ dbo.Orders
    ⊞ dbo.Products
    ⊞ dbo.Region
    ⊞ dbo.Shippers
    ⊞ dbo.Suppliers
    ⊞ dbo.Territories
```

When you open the **Northwind** database in SQL Server Management Studio's Object Explorer, you will see the following tables, which make up the database.

Some of these tables, such as the **Customer*** tables, represent master data. They exist to store information about a key business entity, in this example, *customers*.

Other tables, like **Orders**, represent business activities. These are transactions, events, or actions. From a data warehouse perspective, all source data falls into one of these two categories.

It should be noted that the boundary of what are master data or activities does not always align with the database tables. You must look at the business processes themselves and how data flows through the organization. The tables are just an implementation, and like any implementation they are at the discretion of the database designer.

Of course, in this class we cannot ask someone who works at Northwind Traders about customers or orders. We can only make assumptions about how they would use the data, and for that reason we will rely heavily on the existing database schema.

Assignment Structure

All technical assignments in this course are structured in the same manner. Their purpose is to give you hands-on experience "doing" data warehousing, allowing you to put into practice the concepts you learned through the class sessions. You will learn quite a bit as you work through these assignments. This is by design. Every assignment begins with an overview section, which explains the activity, its goals, and its requirements.

The second part of the assignment will walk you through the process of completing the activity. In this particular assignment, it's profiling data from the Northwind Traders database.

The third part of the assignment you must complete on your own. It is expected you will take what you learned from your studies, class sessions, and your experiences with Parts 1 and 2, then apply them toward the problems you face in the third part.

At the end of the assignment it is stated what you should hand in as a deliverable to demonstrate you completed the work to satisfaction.

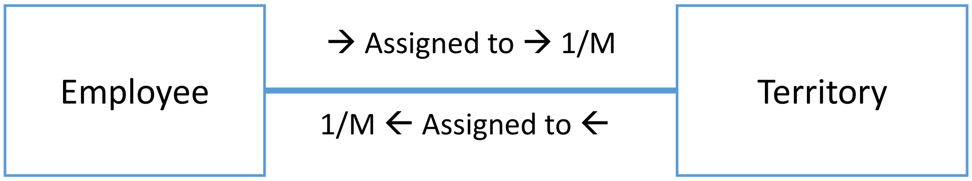
Part 2: Walk-Through

In this part we will walk you through the process of data profiling. We will profile two types of sources: a master data source and a business process.

First Step: Is That Master Data or a Business Process?

In our walk-through, our master data will be employees and our business process will be the assignment of sales territories to the employee.

How do you know which source is master data and which is a business process? The trick is to think of your data at the conceptual model level:



- An **Employee** is *assigned to* one or more **Territories.**
- A **Territory** is *assigned to* one or more **Employees.**

In this example Employee and Territory are master data. They represent categorical business data. The many-to-many relationship between them is a business process representing who gets assigned to which territory.

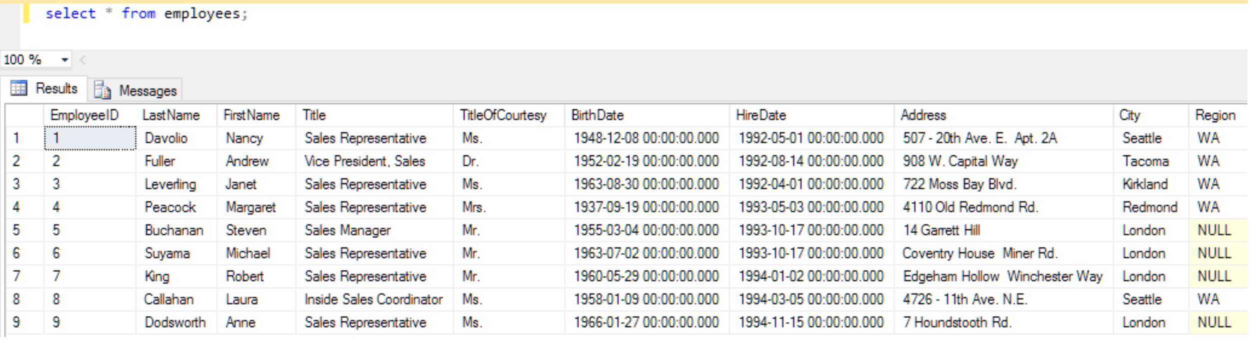
In the Kimball method of data warehousing your master data become ***dimensions***, and the business processes, events, or transactions become ***fact tables***. Honestly, that is a gross oversimplification, but for now it works.

Next: Understanding Your Data at Their Atomic Level

Now that you know which are master and which are business processes, it's time to understand the following:

- What are the business or natural keys? We're not interested in the primary key, which is internal to the RDMBS implementation. We're looking for what makes each entity unique from the business user's vantage point. There should always be a natural key for master data, but there might not be one for business processes.
- What does "one row" of data mean? The answer to this question often reveals itself when you discover the business key or identify the business process.

The easiest way to figure this out is by using the select statement. Observe:



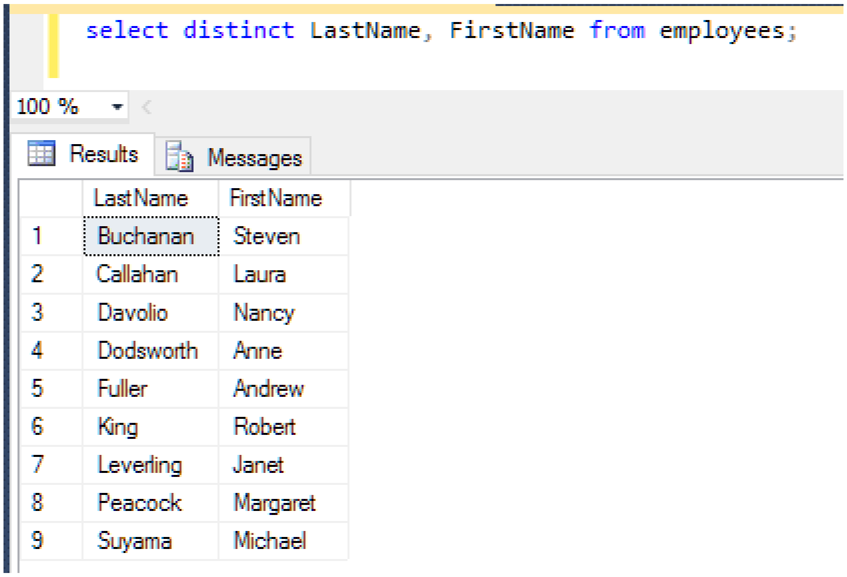
```
select * from employees;
```

	EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	BirthDate	HireDate	Address	City	Region
1	1	Davolio	Nancy	Sales Representative	Ms.	1948-12-08 00:00:00.000	1992-05-01 00:00:00.000	507 - 20th Ave. E. Apt. 2A	Seattle	WA
2	2	Fuller	Andrew	Vice President, Sales	Dr.	1952-02-19 00:00:00.000	1992-08-14 00:00:00.000	908 W. Capital Way	Tacoma	WA
3	3	Leverling	Janet	Sales Representative	Ms.	1963-08-30 00:00:00.000	1992-04-01 00:00:00.000	722 Moss Bay Blvd.	Kirkland	WA
4	4	Peacock	Margaret	Sales Representative	Mrs.	1937-09-19 00:00:00.000	1993-05-03 00:00:00.000	4110 Old Redmond Rd.	Redmond	WA
5	5	Buchanan	Steven	Sales Manager	Mr.	1955-03-04 00:00:00.000	1993-10-17 00:00:00.000	14 Garrett Hill	London	NULL
6	6	Suyama	Michael	Sales Representative	Mr.	1963-07-02 00:00:00.000	1993-10-17 00:00:00.000	Coventry House Miner Rd.	London	NULL
7	7	King	Robert	Sales Representative	Mr.	1960-05-29 00:00:00.000	1994-01-02 00:00:00.000	Edgeham Hollow Winchester Way	London	NULL
8	8	Callahan	Laura	Inside Sales Coordinator	Ms.	1958-01-09 00:00:00.000	1994-03-05 00:00:00.000	4726 - 11th Ave. N.E.	Seattle	WA
9	9	Dodsworth	Anne	Sales Representative	Ms.	1966-01-27 00:00:00.000	1994-11-15 00:00:00.000	7 Houndstooth Rd.	London	NULL

If you want to try it for yourself, open a new query window in the Northwind database by pressing **CTRL+N**, then type the select statement, and press the **!Execute** button in the toolbar.

It appears as if there's one row for each employee. That stands to reason, but to be honest it's not always that simple.

What about the business key? What is used to uniquely identify the employee? Normally this would be a Social Security number or tax payer ID number because each row would have a unique value. With this data it's not clear. You might think of using a name:



```
select distinct LastName, FirstName from employees;
```

	LastName	FirstName
1	Buchanan	Steven
2	Callahan	Laura
3	Davolio	Nancy
4	Dodsworth	Anne
5	Fuller	Andrew
6	King	Robert
7	Leverling	Janet
8	Peacock	Margaret
9	Suyama	Michael

It seems to work as there are still nine rows so we didn't lose an employee.

It's not a good business key because there could be two employees with the same name. In this case there isn't, but we should always think *beyond the data* when we're profiling.

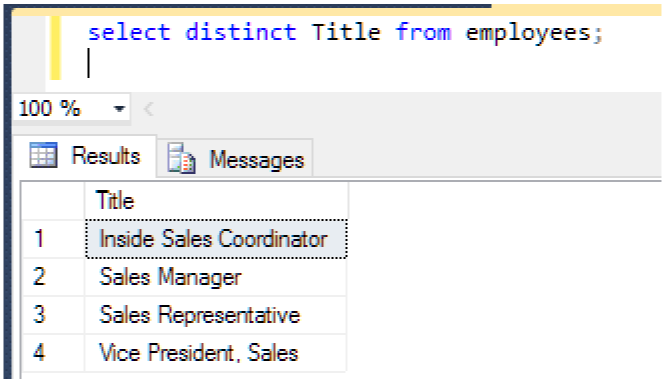
In this case we have no choice but to use the Primary key **EmployeeID**.

Finally: Understanding the Characteristics of Your Data

In addition to knowing what one row means, you might need to discover other characteristics of your data. These are commonly dictated by the business requirements. It's crucial to understand "what you have" so that when it is delivered to the data warehouse you can determine if the data are still accurate. We only want quality data in our data warehouse.

Here are some sample queries that might be asked as part of a functional business requirement and the SQL statements that satisfy them.

What are the different job titles for each employee?



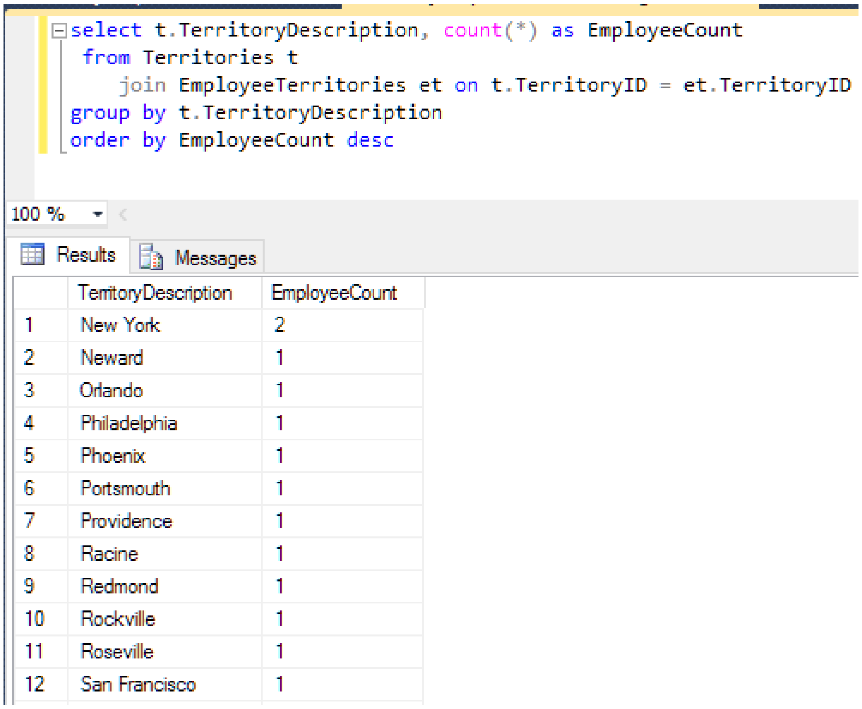
```
select distinct Title from employees;
```

	Title
1	Inside Sales Coordinator
2	Sales Manager
3	Sales Representative
4	Vice President, Sales

There are four distinct job titles for nine employees. That's a decent category, as well as a possibility for drill-down in our analytics as we'll see in future assignments.

Example: How many salespeople per territory?

From the screenshot we now know there is only one territory with more than one employee assigned: **New York**.



```
select t.TerritoryDescription, count(*) as EmployeeCount
 from Territories t
    join EmployeeTerritories et on t.TerritoryID = et.TerritoryID
group by t.TerritoryDescription
order by EmployeeCount desc
```

	TerritoryDescription	EmployeeCount
1	New York	2
2	Neward	1
3	Orlando	1
4	Philadelphia	1
5	Phoenix	1
6	Portsmouth	1
7	Providence	1
8	Racine	1
9	Redmond	1
10	Rockville	1
11	Roseville	1
12	San Francisco	1

Let's take a moment to break down this SQL statement because writing queries like this is common practice in data profiling.

We are joining the **Territories** table to the **EmployeeTerritories** table so we can get a count (**count(*)**) of employees for each **TerritoryDescription** (that's the **group by**). The **t** and **et** are table aliases. These are needed since both tables have a column called Territory ID.

We **order by EmployeeCount** to see the territories with the most employees first.

Part 3: On Your Own

Profile the following table in the Northwind Traders database. The first two have been done for you in the previous part.

Table	**Type**	**Row Count**	**PK(s) Used**	**One Row Is**
Employees	Master Data	9	EmployeeID	An employee
EmployeeTerritories	Business Process	49	EmployeeID, TerritoryID	An employee assigned to a territory.
Customers				
Suppliers				
Products				
Shipments (of Orders)				
Details (of an Order)				

Write SQL queries to answer the following questions that might be associated with functional business requirements in a data warehouse. For each of the following provide a screenshot of the SQL query and its output, making sure your name or NetID appears in the screenshot.

1. List the customer contact names and titles sorted by company name.
2. Factoring in discounts, what is total amount of product sold?
3. Provide a list of product category names with counts of products in each category.
4. Select a specific customer, and display that customer's orders with total amount of product sold for each order.
5. Select a specific employee and each order, how it was shipped (shipvia), the company who shipped it, and the total number of days elapsed from order date to shipped date.

Turning It In

Please turn in a Word document with your name, NetID, and date at the top. Copy and paste your completed Part 3. Be sure you include screenshots as directed.

Do not submit a copy of this assignment file. I only need the Part 3.