Building Your Dataset in CDW:

Joining tables within a domain

by Margaret Gonsoulin, Ph.D.

Acknowledgements

- Richard Pham, BISL
- Feedback from colleagues at VIReC

CDW Cyberseminar Series

- This talk assumes that you have seen the previous cyberseminars:
  - First Time User's Guide to CDW: Getting Started with this Relational Database
  - Seeing the Data When You Can't See the Data: A Tour of Documentation of the CDW
  - Getting the Information You Need from CDW: Starter SQL Language

VIReC
RESEARCHERS' GUIDE TO VA DATA

http://vaww.virec.research.va.gov/CDW/Overview.htm
http://www.hsrd.research.va.gov/cyberseminars/catalog-archive-virec.cfm

Poll #1: About You

Which of the following best describes your role in the VA? (Check all that apply)

- ❑ Research Investigator / PI
- ❑ Career Development Awardee
- ❑ Data Manager / Analyst
- ❑ Project Coordinator
- ❑ Operations / Partnered Research / QI
- ❑ Other

By the end of this talk,

We hope that a new CDW user will:

- Be able to understand the logic of joining tables/ views within a single CDW domain
- Be able to write a query that includes an “inner join” statement
- Be able to incorporate some “query best practices” for working with CDW

VIReC
RESEARCHERS' GUIDE TO VA DATA

Topics Covered

- Review background concepts relavent to CDW:
  - Relational databases concepts
  - SQL Basics
  - SQL Best Practices
- The logic of an INNER JOIN using a simple example
- Step-by-step process of using CDW to join:
  - Two tables/views from a domain
  - Three tables/views from a domain

Background

A Conceptual Overview of Relational Data & SQL

Tables v. Views

- A table is a set of columns and rows that contain data elements.
- A view is the result of a procedure that pulls information out of a database into a virtual table; in simple terms it is a “virtual” table.

Sta3n	State	County	Country	PeriodOfService	MaritalStatus	InsuranceCoverageFlag	Religion
676	WISCONSIN	MILWAUKEE	UNITED STATES	PERSIAN GULF WAR	NEVER MARRIED	N	LUTHERAN
629	ARKANSAS	JEFFERSON	UNITED STATES	VIETNAM ERA	DIVORCED	U	BAPTIST
586	MISSISSIPPI	TIPPAH	UNITED STATES	PERSIAN GULF WAR	MARRIED	U	UNKNOWN/NO PREFERENCE
629	TEXAS	SMITH	UNITED STATES	POST-VIETNAM	SEPARATED	U	BAPTIST
636	NEBRASKA	LINCOLN	UNITED STATES	POST-KOREAN	*Missing*	U	*Missing*
623	OKLAHOMA	TULSA	UNITED STATES	PERSIAN GULF WAR	NEVER MARRIED	Y	UNKNOWN/NO PREFERENCE
676	PENNSYLVANIA	LANCASTER	UNITED STATES	POST-VIETNAM	NEVER MARRIED	NULL	ROMAN CATHOLIC CHURCH
623	OKLAHOMA	TULSA	UNITED STATES	VIETNAM ERA	MARRIED	Y	BAPTIST
676	WISCONSIN	MONROE	UNITED STATES	PERSIAN GULF WAR	MARRIED	Y	BAPTIST
636	CALIFORNIA	CALAVERAS	UNITED STATES	PERSIAN GULF WAR	MARRIED	U	CHRISTIAN (NON-SPECIFIC)
676	WISCONSIN	PORTAGE	UNITED STATES	OTHER NON-VETERANS	*Missing*	NULL	*Missing*
674	TEXAS	ANDERSON	UNITED STATES	WORLD WAR II	WIDOWED	Y	UNKNOWN/NO PREFERENCE
695	WISCONSIN	WAUKESHA	UNITED STATES	VIETNAM ERA	MARRIED	Y	*Missing*
674	TEXAS	HARRIS	UNITED STATES	VIETNAM ERA	DIVORCED	U	PROTESTANT, NO DENOMINATION

SELECT and FROM

- SELECT allows a programmer to list the columns (variables) that they would like to see in the results of their query
  - Each column name should be followed by a comma except the last one in the list
  - This limits the size of your request (a best practice)
- FROM identifies the table/view from which the columns will be collected
  - The name of the view should be written in this format “database.schema.table”

WHERE statements

- The command WHERE allows a user to limit their search to include only a chosen subset of the data.
- When working with large fact tables in CDW, you will want limit the size of your requests for information (another best practice).

```
SELECT column1, column2, column3
FROM Database.Schema.Table
WHERE column1 = X  ;  -- note this is just an example of the kind of
                      -- criteria one might specify in a where
                      -- statement.
```

TOP

- When first drafting your SQL query, another best practice is to test the logic of your query on a small number of rows (a.k.a., cases).
- The command "TOP" is added to the SELECT statement and is followed by the number of rows selected by the programmer.

```
SELECT TOP 100 column1, column2, column3
FROM    Database.Schema.Table1
WHERE  column2 > 10 ; -- new example
```

USE & GO

- Another helpful shortcut is **USE**.
- It is followed by **GO**.
- It allows the user to choose their database at the beginning of the query, so there is no need to repeat the name of the database throughout the query.

```
USE Database
GO
SELECT TOP 25 column1, column2, column3
FROM Database.Schema.Table
WHERE column3 IS NOT NULL ; -- new example
```

Dimension & Fact Tables

- Dimension tables are typically smaller tables holding non-sensitive, supporting information that is meant to be accessed repeatedly
  - Fact tables tend to be large tables containing substantive data about the topic of interest and include sensitive information

Joining Keys

- Primary Key –A column in every table that uniquely identifies each row.
- Foreign Key –These are column(s) in a table that correspond to or reference a primary key in another table.

Simple Example, Joining Keys

Store.CustomerTable

CustomerKey (PK)	LastName	FirstName	Address	CityState	Zipcode
1	Jones	Marianna	123 Oak St	Bee, AR	70788
2	Frank	Josie	11 Pine Ave	Flip, OK	30032
3	Plank	Bill	230 5th St.	Miner, TX	11201

Dim.ItemTable

ItemKey (PK)	Item
1	Shirt
2	Pants
3	Skirt
4	Sweater

Store.PurchasesTable

PurchaseKey (PK)	ItemSoldKey (FK)	CustomerKey (FK)	Date
1	1	1	1/1/2014
2	2	1	1/1/2014
3	3	2	2/2/2014
4	4	2	2/2/2014
5	1	3	3/3/2104
6	1	3	3/3/2014

VIReC
RESEARCHERS' GUIDE TO VA DATA

Poll #2: Your CDW Experience

Rate your level of experience with CDW data on a scale of 1 to 5…

1 Not worked with it at all

2

3

4

5 Very experienced with CDW

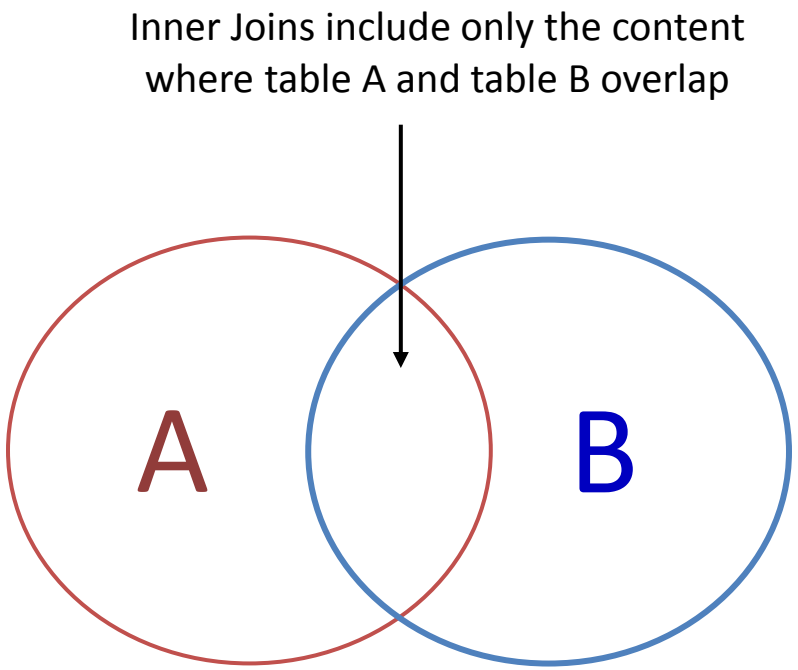
VIReC
RESEARCHERS' GUIDE TO VA DATA

The Logic of an Inner Join

Using a simple example.

Inner Join

This lecture will focus on the SQL command, INNER JOIN, that keeps only those records/ rows that match/exist in both of the tables being joined.



INNER JOIN & ON in SQL

- INNER JOIN clause(s) follow the FROM clause
  - INNER JOIN Database.Schema.Table
- ON should be followed by the name of the two linking keys with an equal sign in the middle.
  - ON LinkingKey1 = LinkingKey2

```
SELECT column1, column2, column3
FROM Database.Schema.Table1
INNER JOIN Database.Schema.Table2
        ON LinkingKey1 = LinkingKey2;
```

Joining, Best Practice

- Join the dimension tables to the fact tables when possible
- Put the fact table into the FROM statement and the dimension table into the JOIN statement

```
USE Database
GO
SELECT column1, column2, column3
FROM Schema. Table1  -- fact table here
INNER JOIN Schema.Table2   -- dimension table here
                ON LinkingKey1 = LinkingKey2;
```

Example 1

```
SELECT Item , Date
FROM Store.PurchasesTable
INNER JOIN  Dim.ItemTable
            ON  ItemSoldKey = ItemKey ;
```

Dim.ItemTable

ItemKey (PK)	Item
1	Shirt
2	Pants
3	Skirt
4	Sweater

Store.PurchasesTable

PurchaseKey (PK)	ItemSoldKey (FK)	CustomerKey (FK)	Date
1	1	1	1/1/2014
2	2	1	1/1/2014
3	3	2	2/2/2014
4	4	2	2/2/2014
5	1	3	3/3/2104
6	1	3	3/3/2014

VIReC
RESEARCHERS' GUIDE TO VA DATA

Example 2, Fails

Store.CustomerTable

CustomerKey (PK)	LastName	FirstName	Address	CityState	Zipcode
1	Jones	Marianna	123 Oak St	Bee, AR	70788
2	Frank	Josie	11 Pine Ave	Flip, OK	30032
3	Plank	Bill	230 5th St.	Miner, TX	11201

Store.PurchasesTable

PurchaseKey (PK)	ItemSoldKey (FK)	CustomerKey (FK)	Date
		1	1/1/2014
		1	1/1/2014
		2	2/2/2014
		2	2/2/2014
		3	3/3/2104
		3	3/3/2014

```
SELECT LastName , FirstName, Date
FROM Store.CustomerTable
INNER JOIN Store.PurchasesTable
    ON CustomerKey = CustomerKey ;
```

Why Example 2 Fails

```
SELECT LastName ,
       FirstName,
       Date
FROM Store.CustomerTable
INNER JOIN Store.PurchasesTable
    ON CustomerKey = CustomerKey ;
```

These column names may not be unique to these tables in the database

The software is confused by the duplicate names

How to distinguish the keys and columns

- In SQL, you can add the name of the table in front of the column or key to distinguish it from others of the same name.
- For columns :
  - Store.CustomerTable.LastName
  - Store.PurchasesTable.Date
- For joining keys:
  - Store.CustomerTable.CustomerKey
  - Store.PurchasesTable.CustomerKey

Repairing the problem with table names…

```
SELECT Store.CustomerTable.LastName ,
       Store.CustomerTable.FirstName,
       Store.PurchasesTable.Date
FROM Store.CustomerTable
INNER JOIN Store.PurchasesTable
    ON Store.CustomerTable.CustomerKey =
       Store.PurchasesTable.CustomerKey ;
```

Alias (Shortened Names)

- You may provide a shortened name to substitute for the name of the table/view by assigning an "alias" using "AS" in SQL:
  - FROM Store.CustomerTable AS A
  - INNER JOIN Store.PurchasesTable AS B
- Use that alias on the columns and joining keys instead of the table name
  - A.LastName , A.FirstName , B.Date
  - A.CustomerKey , B.CustomerKey

Example 2, Succeeds

Store.CustomerTable

CustomerKey (PK)	LastName	FirstName	Address	CityState	Zipcode
1	Jones	Marianna	123 Oak St	Bee, AR	70788
2	Frank	Josie	11 Pine Ave	Flip, OK	30032
3	Plank	Bill	230 5th St.	Miner, TX	11201

Store.PurchasesTable

PurchaseKey (PK)	ItemSoldKey (FK)	CustomerKey (FK)	Date
		1	1/1/2014
		1	1/1/2014
		2	2/2/2014
		2	2/2/2014
		3	3/3/2104
		3	3/3/2014

```
SELECT A.LastName , A.FirstName, B.Date
FROM Store.CustomerTable AS A
INNER JOIN Store.PurchasesTable AS B
    ON A.CustomerKey = B.CustomerKey ;
```

An Inner Join in CDW

A step-by-step guide.

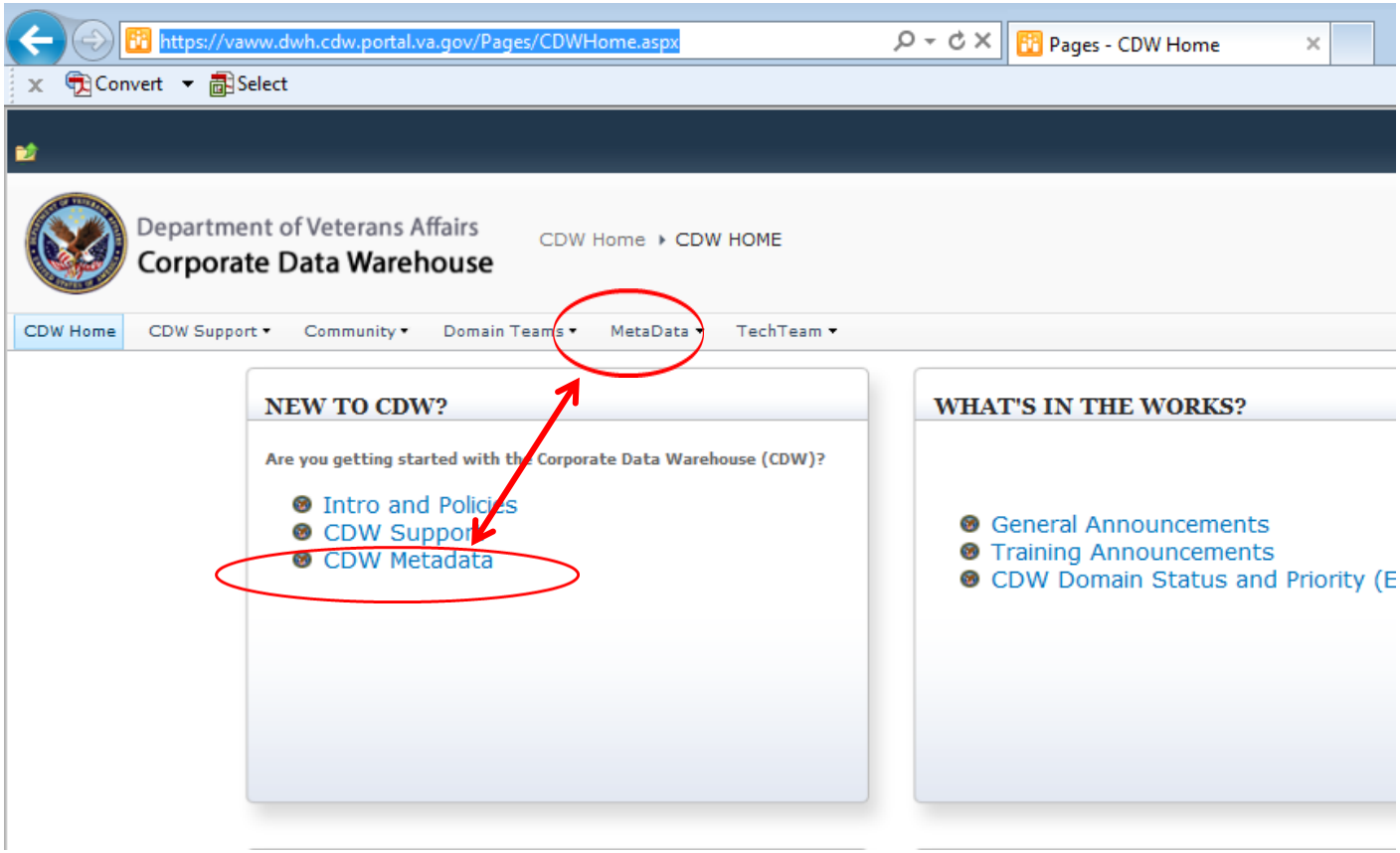
Preparing to use the data

For today's example, we will look at CPT codes associated with outpatient procedures.

1. Go to the CDW Metadata Report to find information on your concepts of interest.
2. Look at "table level" metadata report to find columns.
3. Look at the ER Diagram to discover relationships between tables.

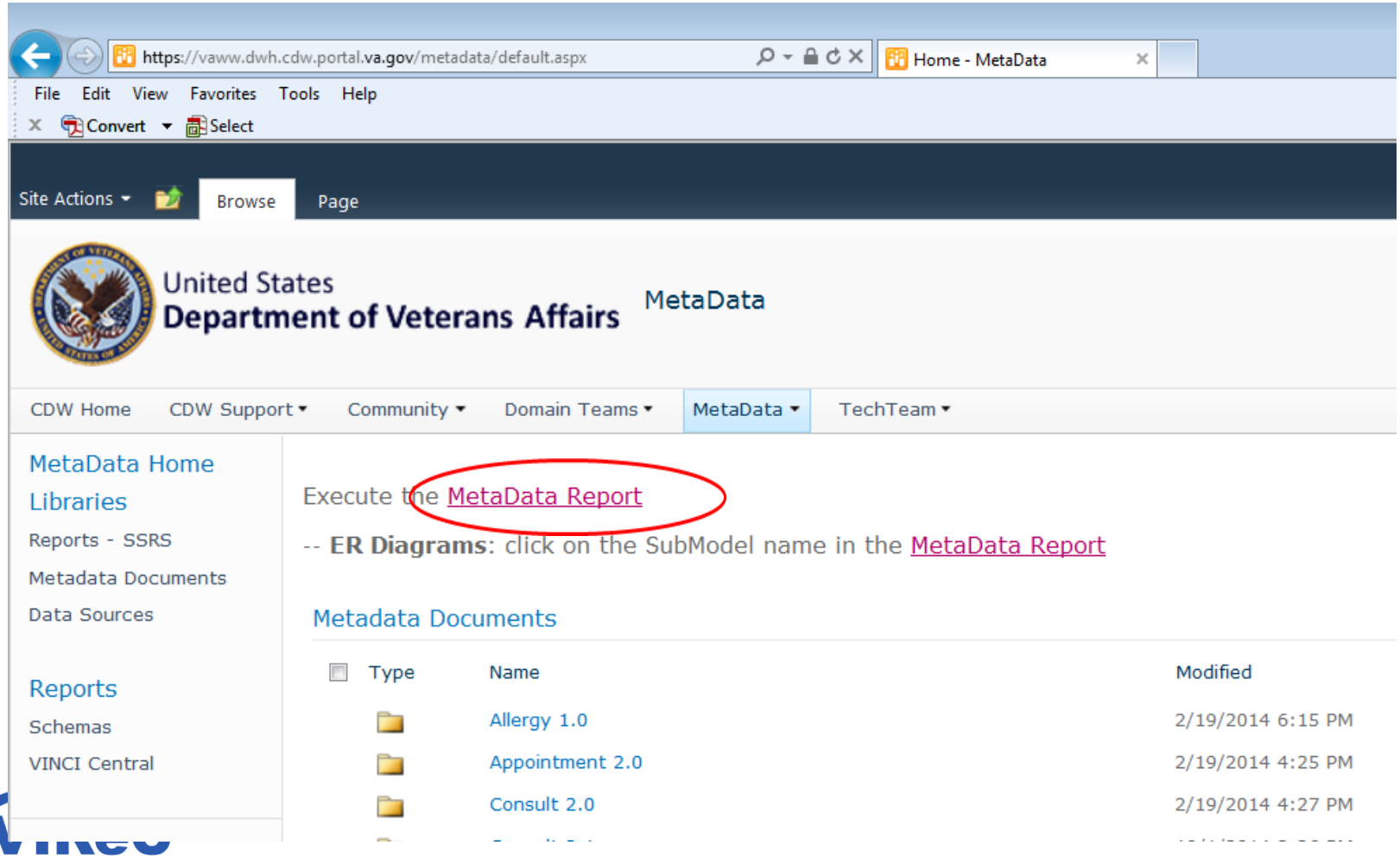
https://vaww.dwh.cdw.portal.va.gov/metadata/default.aspx

1. Go to the CDW Metadata Report



VIReC RESEARCHERS' GUIDE TO VA DATA https://vaww.dwh.cdw.portal.va.gov/Pages/CDWHome.aspx

1. Go to the CDW Metadata Report (Con't)



2. Metadata Reports for Each Domain

CDW Metadata

Contains a grouped list of available CDW ER Diagrams and members.

ImageDescription	
⊞ Allergy 1.0	Image Date: 01 Feb 2014
⊞ Appointment 2.0	Image Date: 03 Mar 2015
⊞ Consult 2.1	Image Date: 03 Oct 2014
⊞ CPRSOrder 1.0	Image Date: 11 Aug 2014
⊞ Data Profiling 1.0	Image Date: 21 Feb 2014
⊞ Dental 1.0 Diagram 1 of 2	Image Date: 11 Aug 2014
⊞ Dental 1.0 Diagram 2 of 2 for Analytics	Image Date: 11 Aug 2014
⊞ Dimensions A Through D 10/3/2014	Image Date: 03 Oct 2014
⊞ Dimensions E Through K 10/3/2014	Image Date: 03 Oct 2014
⊞ Dimensions L Through R	Image Date: 17 Dec 2014
⊞ Dimensions L Through R 10/3/2014	Image Date: 03 Oct 2014
⊞ Dimensions S Through Z 10/3/2014	Image Date: 03 Oct 2014
⊞ Encounter 1.0	Image Date: 29 Oct 2013
⊞ Health Factor 2.0	Image Date: 11 Mar 2015
⊞ HealthFactor 1.0	Image Date: 20 Nov 2013
⊞ Immunization 2.1	Image Date: 24 Feb 2015

- Each domain is listed in alphabetical order
- Scroll down to find the Outpatient Domain
- Click plus sign in the box to the left of the domain name to see metadata reports for each table/view

VIReC
RESEARCHERS' GUIDE TO VA DATA

2. Metadata for each table

⊞ Inpatient 2.1 Diagram 1 of 5	Image Date: 20 Aug 2014
⊞ Lab Microbiology 1.0	Image Date: 11 Aug 2014
⊞ LabChem 2.0	Image Date: 11 Aug 2014
⊞ Mental Health 1.0	Image Date: 21 Feb 2014
⊟ Outpatient 2.0	

Expand to see table level metadata reports on the right

DWViewName	Field Count	FileMan File Data Source	View
Dim.AppointmentStatus	9	APPOINTMENT STATUS (409.63)	DWVi View
Dim.AppointmentType	9	APPOINTMENT TYPE (409.1)	DWVi View
Dim.ClinicalTerm	17	EXPRESSIONS (757.01)	DWVi View
Dim.ClinicalTermType	6	EXPRESSION TYPE (757.011)	DWVi View
Dim.CPT	16	CPT (81)	DWVi View
Dim.CPTCategory	10	CPT CATEGORY (81.1)	DWVi View
Dim.EducationTopic	6	EDUCATION TOPICS (9999999.09)	DWVi View
Dim.Exam	7	EXAM (9999999.15)	DWVi

To Outpat.VProcedure

VIReC
RESEARCHERS' GUIDE TO VA DATA

What's in Outpat.VProcedure?

- By reading the table level metadata reports on CDW's SharePoint site, I can see that Outpat.VProcedure contains:
  - ✓the date of the procedure
  - ✓the date of the encounter
  - ✓comments related a procedure performed during an encounter

What's in Dim.CPT ?

- The metadata report tells us that this table contains:
  - ✓Current Procedural Terminology (CPT) Code, name and description
  - ✓A CPT Category and Major Category
  - ✓The dates that the CPT code became active and went inactive
  - ✓The appropriate age range and gender for applicable CPT codes

3. Entity Relationship Diagrams

⊞ Inpatient 2.1 Diagram 1 of 3	Image Date: 20 Aug 2014
⊞ Lab Microbiology 1.0	Image Date: 11 Aug 2014
⊞ LabChem 2.0	Image Date: 11 Aug 2014
⊞ Mental Health 1.0	Image Date: 21 Feb 2014
⊟ Outpatient 2.0	

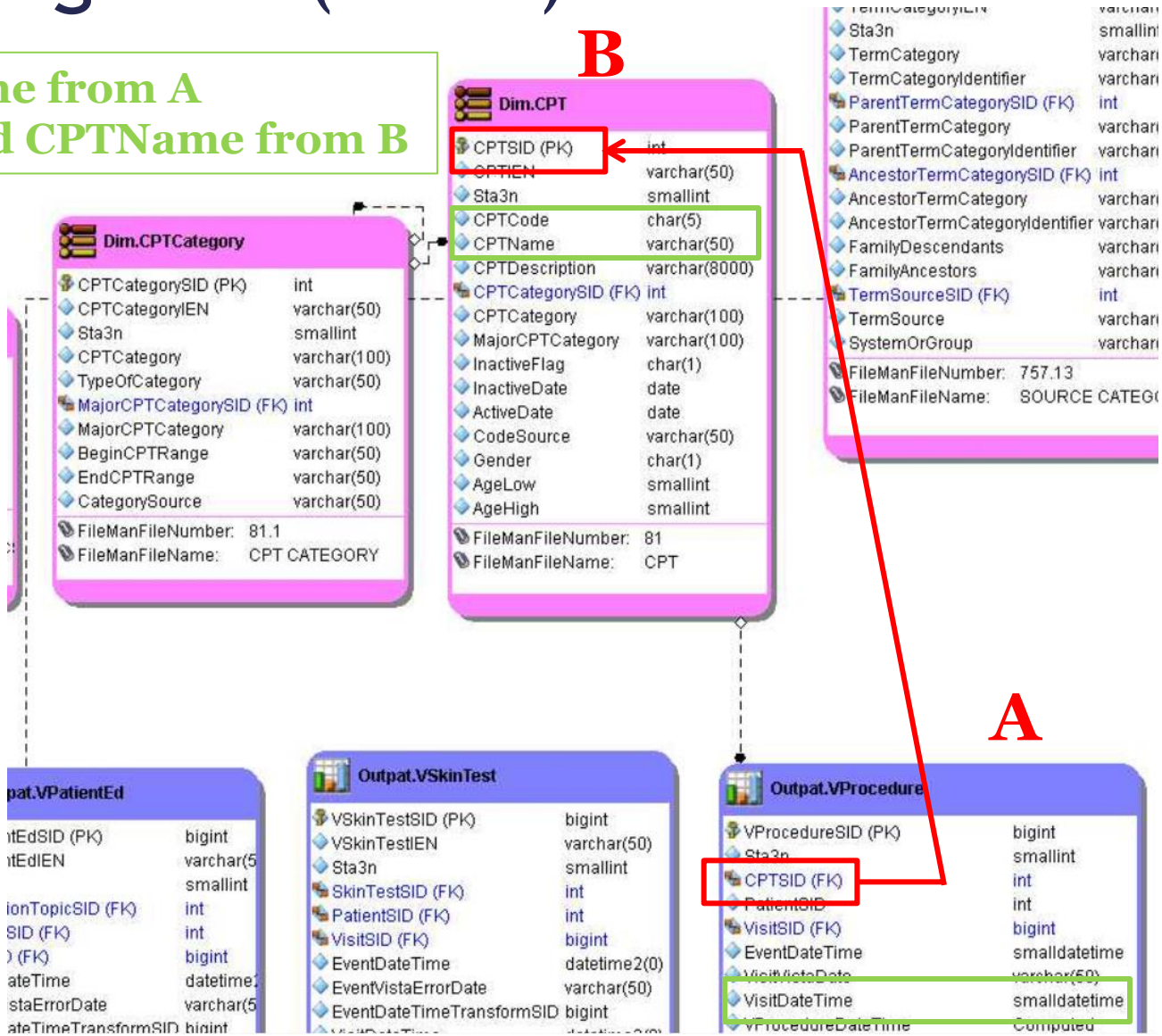
Click the domain name to open the ER Diagram

DWViewName	Field Count	FileMan File Data Source	View
Dim.AppointmentStatus	9	APPOINTMENT STATUS (409.63)	DWVi View
Dim.AppointmentType	9	APPOINTMENT TYPE (409.1)	DWVi View
Dim.ClinicalTerm	17	EXPRESSIONS (757.01)	DWVi View
Dim.ClinicalTermType	6	EXPRESSION TYPE (757.011)	DWVi View
Dim.CPT	16	CPT (81)	DWVi View
Dim.CPTCategory	10	CPT CATEGORY (81.1)	DWVi View
Dim.EducationTopic	6	EDUCATION TOPICS (9999999.09)	DWVi View
Dim.Exam	7	EXAM (9999999.15)	DWVi

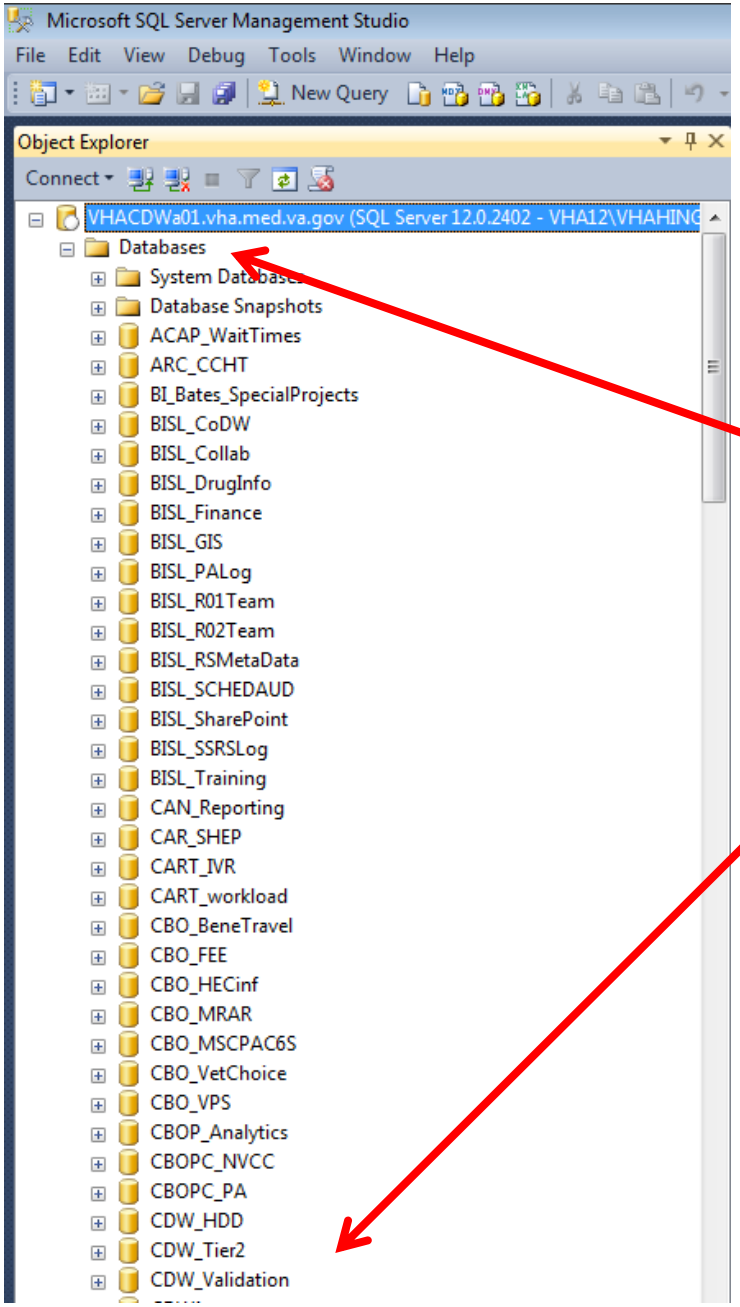
VIReC
RESEARCHERS' GUIDE TO VA DATA

3. ER Diagrams (con't)

VisitDateTime from A
CPTCode and CPTName from B



VIReC
RESEARCHERS' GUIDE TO VA DATA

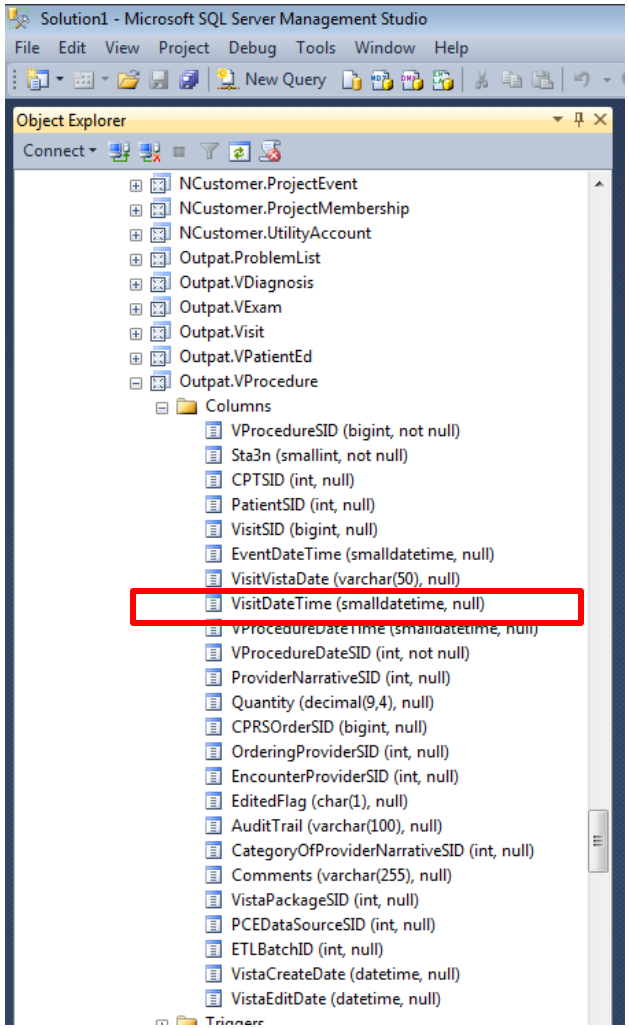


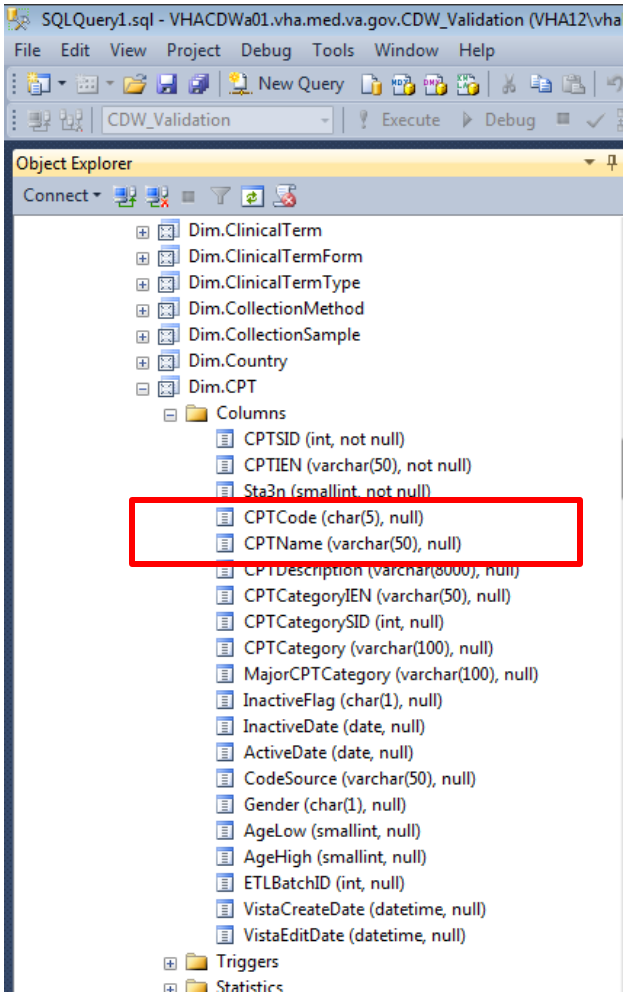
In SSMS,
access your server and
expand “Databases”

Then, scroll down to find
“CDWWork”

Note: this example uses this example accesses the VHACDWa01 server, but data users will have access to a variety of servers.

Finding Your Views in CDWWork





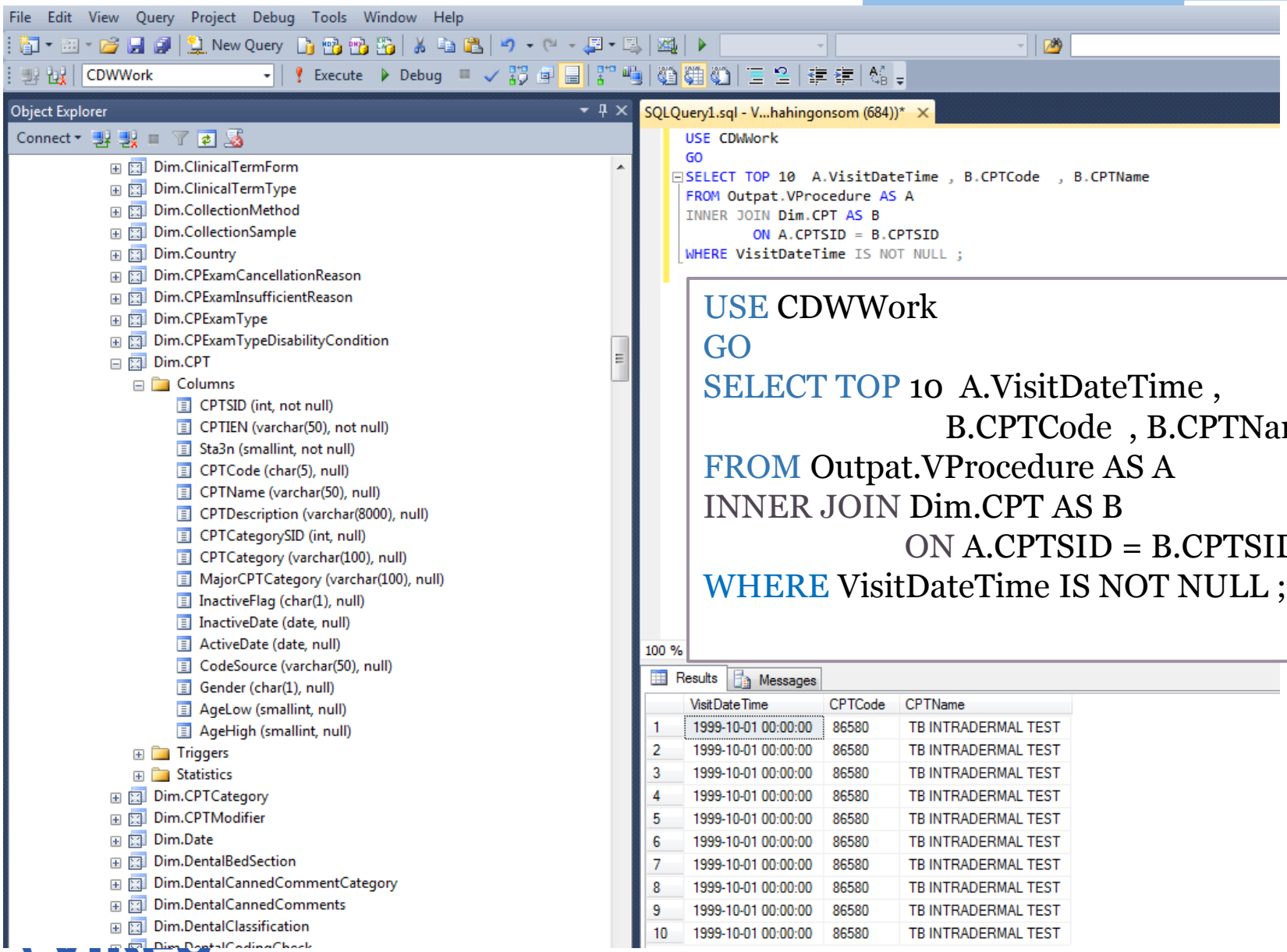
VIReC RESEARCHERS' GUIDE TO VA DATA

Note: this example shows the Outpat.Vprocedure fact table in the CDWWork folder, but a research extract of a fact table may be placed in a research folder with a project-specific name.

Joining 2 CDW Views/Tables

- Use INNER JOIN/ON clause after FROM clause
- Use AS to alias tables, columns and joining keys
- Put the fact table into the FROM statement and the dimension table into the JOIN statement

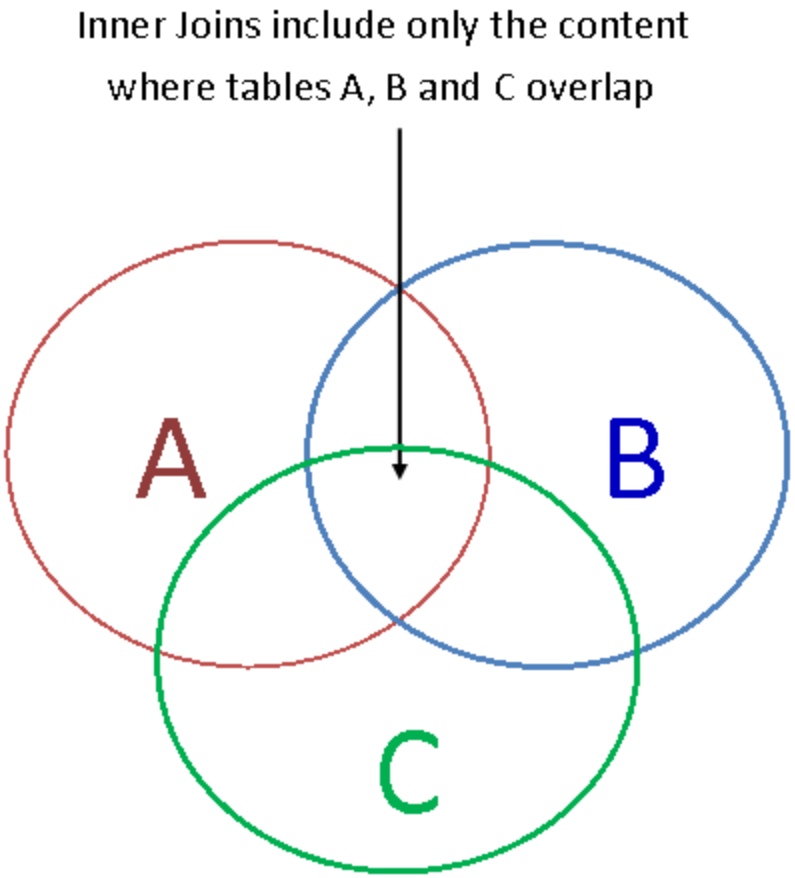
```
USE Database
GO
SELECT A.column1, B.column2
FROM Schema. View1  AS A -- fact table here
INNER JOIN Schema.View2 AS B  -- dimension table here
        ON A.LinkingKey1 = B.LinkingKey2;
```




```
USE CDWWork
GO
SELECT TOP 10  A.VisitDateTime ,
                        B.CPTCode  , B.CPTName
FROM Outpat.VProcedure AS A
INNER JOIN Dim.CPT AS B
                ON A.CPTSID = B.CPTSID
WHERE VisitDateTime IS NOT NULL ;
```

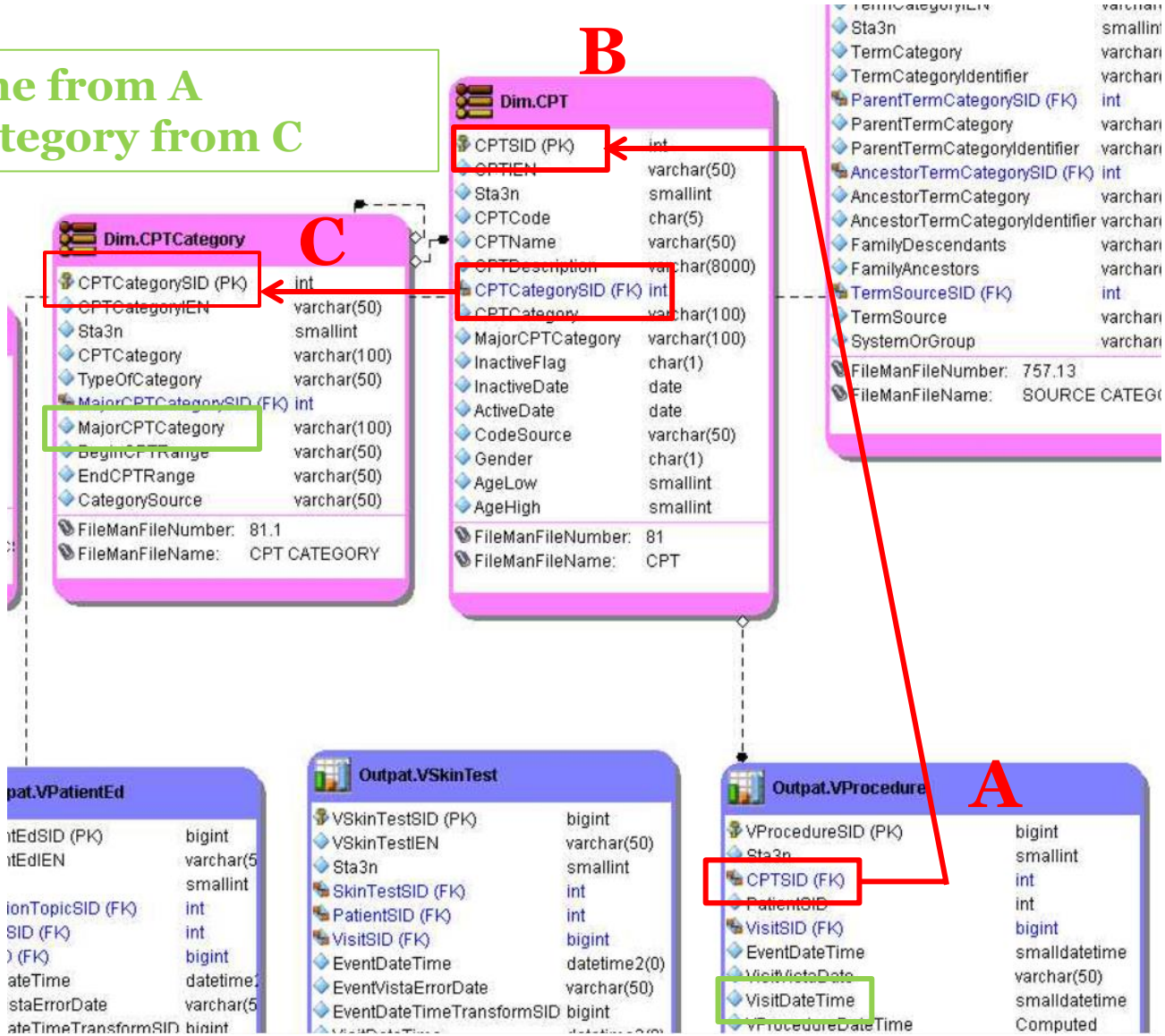
	VisitDateTime	CPTCode	CPTName
1	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
2	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
3	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
4	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
5	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
6	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
7	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
8	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
9	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST
10	1999-10-01 00:00:00	86580	TB INTRADERMAL TEST

Inner Join with three tables/views



Joining 3 views/tables

VisitDateTime from A
MajorCPTCategory from C



What's in Dim.CPTCategory?

- By reading the metadata report for Dim.CPTCategory, I can see that it contains:

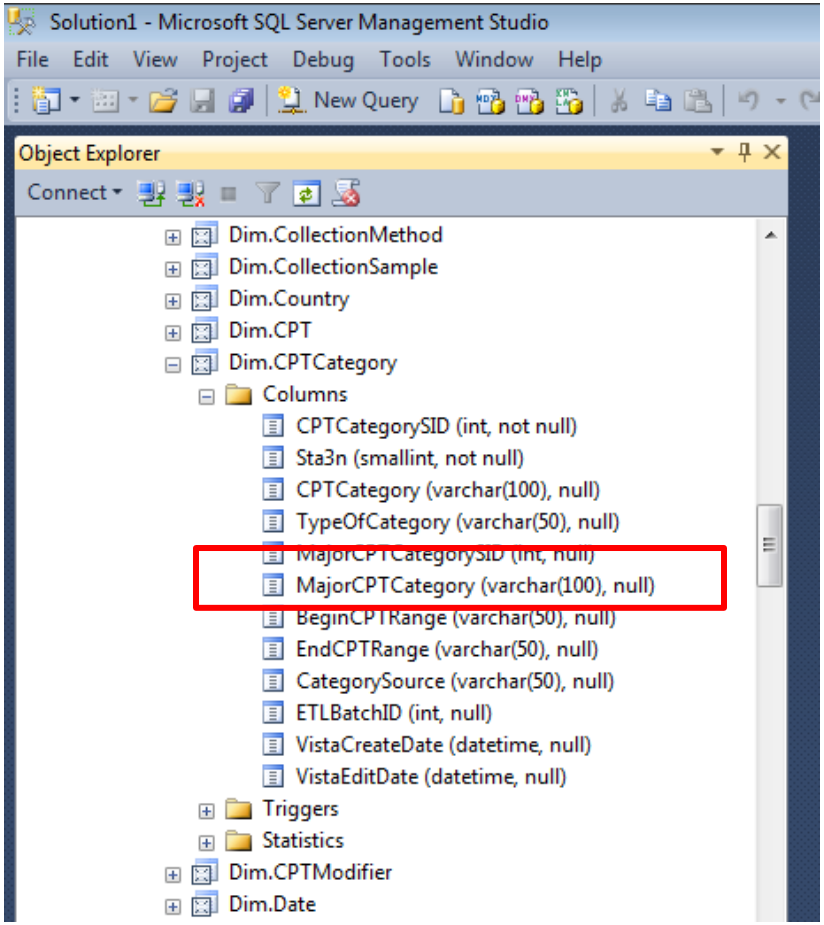
✓ the CPT category

✓ the CPT major category

✓ the type of category (major or sub-category)

✓ the source of the category (CPT or HCPCS)

✓ restrictions on the range of codes that are appropriate



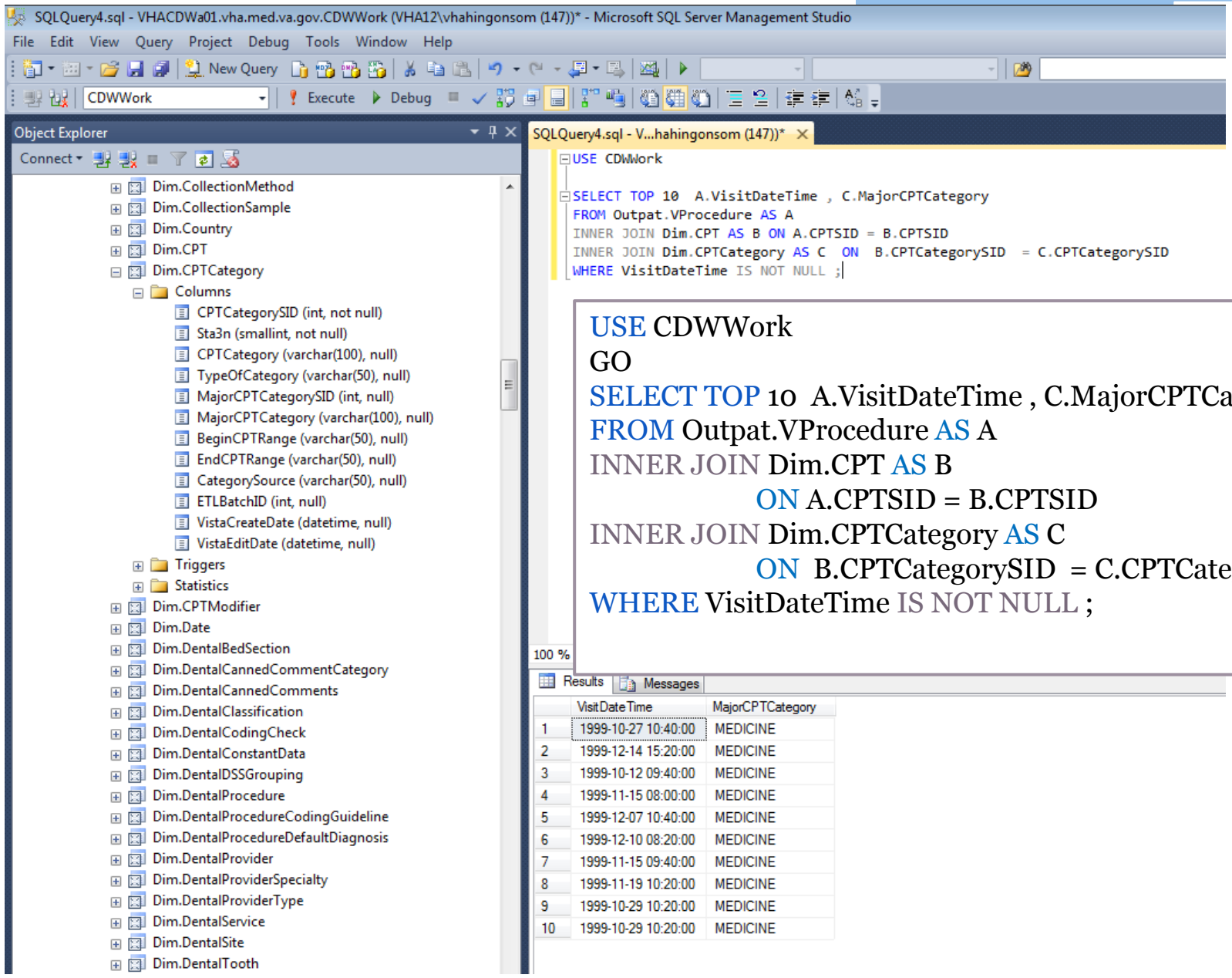
VIReC RESEARCHERS' GUIDE TO VA DATA

CPT = Current Procedural Terminology; HCPCS =Healthcare Common Procedure Coding System

Joining 3 CDW Views/Tables

- Add an additional INNER JOIN/ON clause after the first INNER JOIN clause

```
USE Database
GO
SELECT A.column1, B.column2, C.column3
FROM Schema. View1  AS A                    -- 1st table
INNER JOIN Schema.View2 AS B                -- 2nd table
        ON A.LinkingKey1 = B.LinkingKey2
INNER JOIN Schema.View3 AS C                -- 3rd  table
        ON B.LinkingKey3 = C.LinkingKey4;
```




```
USE CDWWork
GO
SELECT TOP 10  A.VisitDateTime , C.MajorCPTCategory
FROM Outpat.VProcedure AS A
INNER JOIN Dim.CPT AS B
                ON A.CPTSID = B.CPTSID
INNER JOIN Dim.CPTCategory AS C
                ON  B.CPTCategorySID  = C.CPTCategorySID
WHERE VisitDateTime IS NOT NULL ;
```

Summary/Conclusions

- INNER JOIN can be used to pull together fields/columns from many tables/views.
- Best practices such as joining dimension tables to fact tables, using aliases and reducing the size of query with WHERE will lead to greater success in working with CDW.

Best Practice: INNER JOIN (not JOIN)

- SQL Server allows the use of just the word JOIN. If your code ever needs to be exported to any other database system now or in the future, using JOIN by itself IS NOT recommended.
- Not recommended

```
SELECT TOP 10  A.VisitDateTime , B.MajorCPTCategory
FROM Outpat.VProcedure AS A
JOIN Dim.CPT AS B
                ON A.CPTSID = B.CPTSID
```

- Recommended

```
SELECT TOP 10  A.VisitDateTime , B.MajorCPTCategory
FROM Outpat.VProcedure AS A
INNER JOIN Dim.CPT AS B
                ON A.CPTSID = B.CPTSID
```

Best Practice: INNER JOIN & ON (not FROM, WHERE, =)

- Statistical programs such as SAS, SPSS, and Stata permit using comma (,), equal (=), star-equal (*=), and equal-star (=*) for joins. This is not recommended in SQL Server.
- Not recommended (Permitted in SAS and SQL Server 2008)

```
SELECT TOP 10  A.VisitDateTime , B.MajorCPTCategory
FROM Outpat.VProcedure AS A, Dim.CPT AS B
WHERE A.CPTSID = B.CPTSID;
```

- Recommended

```
SELECT TOP 10  A.VisitDateTime , B.MajorCPTCategory
FROM Outpat.VProcedure AS A
INNER JOIN Dim.CPT AS B
                ON A.CPTSID = B.CPTSID;
```

Contact Information

Margaret Gonsoulin

VIReC@va.gov

708-202-2413

VIReC
RESEARCHERS' GUIDE TO VA DATA

Questions?