Cyber Seminar Transcript

Date: March 29, 2017

Series: VIReC Corporate Data Warehouse

Session: CDW: A Conceptual Overview 2017

Presenter: Margaret Gonsoulin, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///%5C%5Cvhaslcnasres.v19.med.va.gov%5Cl%20)

Hira Khan: Hello everyone and welcome to VIReC’s Corporate Data Warehouse Cyberseminar series. Thank you to CIDER for providing the technical and promotional supports. Today’s session is titled CDW: A Conceptual Overview 2017. This seminar will review and summarize terms and concepts associated with relational databases in general and the CDW in particular as it exists in 2017. Today’s speaker is Dr. Margaret Gonsoulin. Dr. Gonsoulin is a sociologist who studied at the University of Virginia and taught for eight and half years in the California State University system before she joined the VA Information Resource Center, VIReC in 2014. Since joining the VA she has focused her efforts on the Corporate Data Warehouse. If you have any questions for Dr. Gonsoulin during the presentation please send them in using the question box and I will present them to her in the end of the session. And now I’m pleased to welcome Dr. Margaret Gonsoulin.

Dr. Gonsoulin: Thank you Hira. Thank you Heidi. And thank you everyone for coming today, welcome. As Hira already told you, today will be an update to the talk that we give annually and we update this because there are changes every year in the Corporate Data Warehouse that need to be addressed in the overview. So with that, I would like to thank Richard Pham, Heidi and Hira for help with this session. I really appreciate it and before we begin, I would like to poll the audience to learn a little bit more about you. How would you describe your level of experience with CDW on a scale from one to five where one is not having worked with it at all and five is very experienced with CDW?

Heidi: And responses are coming in I will give everyone a few more moments to respond before we close the poll question out and go to the results. And it looks like we’ve started slowing down so I’m going to close that out. And what we are seeing is 42% of the audience saying that they have not worked with it at all. Twenty-eight percent rate themselves at a 2. Twenty-three percent rate themselves at a 3. Eight percent at a 4 and zero say that they are very experienced with CDW data. Thank you everyone.

Dr. Gonsoulin: Thank you everyone. So great, it sounds like we have an audience here that should mostly benefit from the talk. And if you’re a little bit higher up on the scale perhaps you could use this talk to help train new members of your team in the future, reducing your work load. So for today’s agenda we are going to go through some basic conceptual issues as they relate to the Corporate Data Warehouse. And the first one will be about all of the acronyms that you see that are associated with the Warehouse and what they mean. Second one will be to start to think in what I’m referring to here as “relational data” terms. I think we were all trained in school to think about non-relational data or flat files. And so this takes some sort of transition time to get used to relational data if you haven’t worked with it in the past. We’ll review that. The third topic for today’s talk is aimed at having you become more familiar with the components of the Corporate Data Warehouse, and I’m talking about the major components or parts that are referred to as production and raw domains, and fact and dimension tables or views. So we’ll cover that. And fourth, we’ll talk on a theoretical level about how to create an analytic dataset when you start with a relational database. In order to have that talk we’ll discuss primary and foreign keys, and the concept of joining tables and the views using those keys.

So first, let’s get to the bottom of all those acronyms. When you see documentation about CDW you will often see xDW. That little ‘x’ is waiting to be filled in with one of three letters. When you see it filled in with a ‘V’ for VDW that is referring to the VISN Data Warehouses. When you see it filled in with the letter ‘R’ it’s referring to the Regional Data Warehouses. And when you see ‘C’ it’s referring to Corporate or our National Data Warehouses at VHA. This is because each level of the organization maintains its own data warehouse for its specific population. And they perform all sorts of reporting roles and operational work and things with these data warehouses. Today’s talk is going to focus exclusively on the CDW or the National Data Warehouse.

So in a general sense, I put together this really rather over-simplified picture of how data are arriving in the Corporate Data Warehouse. So on the far left in the purple circles you see a representation of the 130 or so VISTA systems that serve as the primary source of the data in the CDW. So when these data from these VISTA systems are first brought into the Regional Data Warehouses for the four regions. And from the Regional Data Warehouse it’s brought into the Corporate Data Warehouses and then the VISN Data Warehouses, plural. So if you’re not familiar with VISTA that is the Veterans Health Information System and Technology Architecture that holds the VHA’s administrative and electronic health record data. And it is the source data for most VA databases of all kinds.

So let’s think a little bit about what it means for our data to be stored in a data warehouse. In technical terms, a data warehouse can be defined as a data delivery system that’s intended to give users the information they need to support their business decisions. But this talk is not really aimed at a technical audience, so in ordinary terms, what is a data warehouse? I guess we could think about it as a large storage facility for a whole lot of data or big data.

So what data is the VHA putting into our Corporate Data Warehouse? Well I already mentioned as you can see pictured here in blue the 130 VISTA systems and how specific information is selected and brought into the data warehouse. But that’s not the only data that we’ll find in the data warehouse. A lot of the VA’s data is making its way into the data warehouse now. So you see in the red box to the left, other types of data that you will also find stored in the data warehouse such as MCA or Managerial Cost Accounting data. They put their national data extracts in the data warehouse but the Managerial Cost Accounting office has already created that data set. So in a sense you can refer to that as a derived data set that is now being housed in the data warehouse. You can also find things like CAN scores and the Health Economics Resource Center data in the data warehouse.

So once these data make their way into the Corporate Data Warehouse they’re organized in a particular way and so when you’re first encountering it it’s helpful to understand that basic organization. So at the highest level, the data gets organized into what are referred to as data domains and we’ll discuss all of this in greater detail as the slides progress. And so in this case you can see a red box, a blue box, and an orange box. Each one of those is a data domain. The first domain is for allergy, second for inpatient, the third for surgery. And then within each domain you see that it has multiple tables in each domain, and then you can see all sorts of green arrows in this picture. And these represent the fact that these tables have to able to be regionally or linked back together. So linking keys are built between the tables within a domain and between a table in one domain and a table in another domain, so between the domains. So by using these linking keys you can bring together bits of data about allergies and surgeries and inpatient stays.

Another really common question that people have whenever they first encounter the Corporate Data Warehouse but have been using VA data for a long time is how do the data that I see in the Corporate Data Warehouse compare to VHA data that I’ve been using for a very long time? In the HSR&D community a very common data source is called MedSAS data. In this case I’m using MedSAS outpatient as the example. But people will say ‘well where are the variables that I’m used to seeing in MedSAS data? Where in the warehouse can I find these and how can I tell if it’s the same thing as what I’m used to seeing in the MedSAS data?’ So in order to figure out something like this on a conceptual or theoretical level, these are the steps that are required. So we start with the big blue box on the left, VISTA data from the 130 VISTA systems. These are really the official sort of way of keeping the VA’s data for the last twenty something years. And from these VISTA data for a very long time, National Data Systems, or NDS has been creating the National Patient Care Database. They would select very specific information out of VISTA. They would filter it and apply all sorts of logic to it and they would create the NPCD database. And then from the NPCD database they would apply more logic and they would create the MedSAS outpatient file. Now for the last ten years or so, I think I have that timeline right, if we look at the purple arrow and purple boxes at the bottom, the Corporate Data Warehouse has been getting increasing amounts of data from VISTA and organizing it into the CDW. So in order to figure out the relationship between a database like MedSAS and CDW first you will want to figure out which VISTA files and fields were brought into MedSAS, what logic and filters were applied to it and then you would search for that same field in the Corporate Data Warehouse, recognizing that the logic and the filters that have been applied to MedSAS would not have been applied to the Corporate Data Warehouse. So you would have to apply those same business rules to the data you see in the warehouse to arrive at the same data you saw in MedSAS.

Moving on to the second topic, learning to think in relational data terms. So at this point, what does the word ‘relational’ really mean and how can we come to understand that? Well, in ordinary everyday language you can think of it this way, relational data are data that have been separated out into multiple tables and each one of these tables looks like a spreadsheet with the variables at the top and the records going down the rows. Inside each one of these tables you’ll find linking keys have been added to these tables so that you can bring data together from those various tables in order to create a set of data that you want to use to analyze your question. In order to do that linking, people most often use structured query language or SQL programing. That’s the language that can be used, or one of the languages that can be used to bring together relational databases.

So let’s start with something familiar. This is what some people would refer to as a flat file. In ordinary, everyday statistics and research methods classes it’s just a database that we are familiar with. If we think about the content found in this familiar flat file, we can see that we have three main types of information in it. The first type is patient information. It’s the names of the patients and the addresses of the patients. The second kind of information is information about diagnoses. In this case we’re seeing the ICD-9 codes. And also we see information about their visit or episode of care. We see the location where it occurred in terms of station number and we see the date that it happened on. So how we do go from this flat environment to the relational database? Well, I mentioned earlier that in general it’s about taking the data and breaking it into multiple tables. So each one of the multiple tables would have one type of information in it. So we’ll take these three types, patient, diagnosis and visit information, and create three tables. And now we have a relational database. The first one is devoted exclusively to information about the patient, the second one exclusively to diagnoses information, and the third one exclusively to information about the visit. We’ll keep coming back to this simple example as the talk proceeds.

But before we go into the third part of the talk I would like to ask one more question. And I know I had too many answer options so I may just turn this one over to Heidi.

Heidi: So what we’re looking for here are which of the following best describes your role in the VA? And you can select all that apply here. The options here: operations/quality improvement only, research investigator or PI, research data manager or analyst, project coordinator, both operations and research work. And we know not everyone is going to fit into these categories but I ran out of options for anything else in here. If you do have a different role or different category, please send that in as a questions box. I would love to, we would love to see what it is you are doing and I will happily read through those as I’m going through the pole results here. I’m going to give everyone just a few more minutes, few more moments to respond here, and then I will close this out and we will go through the results that we are seeing here. Okay it looks like we have stopped on the poll. Thank you to those of you who are sending in into the questions box. I’ll go through those in a second. So I’m going to close this out. And what we are seeing is 32% of the audience being operations/quality improvement only. Eighteen percent, saying research investigator or PI. Thirty-four percent, research data manager or analyst. Twenty-four percent, project coordinator. Twenty-four percent, both operation and research role. And then the comments that we’re seeing come in through the question box, we have nursing, informatics, performance improvement, database programmer, clinical applications coordinator, measurement consultant and clinical services analyst. Thank you, everyone.

Dr. Gonsoulin: Yes, thank you so much everyone. So now I’ll move on to the third part of today’s talk, and that’s focused on becoming more familiar with the components or parts of the Corporate Data Warehouse. So I’ve alluded to this in previous slides.

There are two main parts to the Corporate Data Warehouse, and they are referred to as production and raw. So people say production domains and raw domains. So these are classifications of the data domains that exist in the Corporate Data Warehouse. So if we think about what makes domain a production domain, it is referred to in this way because the team of database architects that build the Corporate Data Warehouse have taken the information from VISTA and they’ve modeled or architected it to make it easier to use. And in short, they’ve built in linking keys and they’ve organized the data in a way that is more productive in terms of making it run more quickly, run more easily, join more easily and things like that. And another distinguishing factor of production domains is that they are updated every single night, all of them, consistently. So that’s the key characteristics that will make a domain be referred to as a production domain. So what makes a domain raw? Well the architects are still bringing in the data from VISTA and putting it in the Corporate Data Warehouse, but they’re leaving it in a state that is much more closely aligned to its organization and arrangement in VISTA. They’re not adding in the linking keys and they’re not necessarily updating it on a consistent basis. So from one raw domain to the next you have different rates of update. One may be every two weeks; another one may be once a month, and so on and so forth. So if you’re working in the raw domains you would have to learn slightly different skill sets for working with big data, you’d have to learn different ways of joining, and you would have to look up information about how often they are updated.

So in the general sense, I’d like to kind cement our understanding of what a domain is. I already mentioned some examples so you’ve probably have picked up on the fact that they are based on subject matter that’s found within the groups of tables. So what makes a domain a domain is that all of the tables in it tend to be on the same subject. So some examples of production domains that exist right now in the Corporate Data Warehouse are outpatient, meaning outpatient episodes of care are captured in this domain. Lab microbiology, mental health, patient vital signs, patient insurance, so these are some of the examples but there are about 38, I think right now? And then some examples of domains that are currently among the raw domains are IV meds, oncology, prosthetics, bill claims and radiology. And because the CDW is always evolving and always growing, radiology right now is on its way to becoming a production domain. Perhaps within the month there would be an email updating you that now radiology is no longer raw. You can find it in the production domain.

I’ll also mention that within these domains you will see a set of tables. Every domain has multiple tables. Often times you will hear people use the word table and then other times you will hear people use the word view. When you’re new this can be confusing and often at first made me wonder ‘well what is the difference between a table and a view?’ So I’d like to take a minute to talk about the way that these terms are used and the difference that they make to the end user. A table is what you would expect it to be. It is a set of columns that act as variables in research language and a set of rows or records related to those columns. And it’s like you see right here in this image at the bottom of this screen. You know the Stay3n, and then all of the Stay3n values. State, all of the states. It looks and feels like any database you’ve ever worked with in the past. So what is a view? A view is, in technical terms, the results of a stored procedure that is pulling information out of a database and putting it in a virtual table. So in simple terms it’s like a virtual table or a look at the table. So what’s happening is that what you see here at the bottom of the screen, the Stay3n or station, state, county, country, etcetera, this also a view from the perspective of an end user, right. To us, who are only using the views, because end users are only allowed to use a view, the view feels exactly the same as a table. Because it’s still a set of columns and a set of rows and it’s in a table or a view that feels real to me as an end user. What’s happening is that the table that’s really containing those data are just being kept safely in background so that no accidents can occur when end users are using the view of the table. But ultimately it feels the same to the end user.

So now that we have sorted out that it doesn’t really matter whether we call it a table or view, because it feels the same on the end where people use it, let’s talk about the fact that there are two different kinds tables or views. The first kind of table that we’ll talk about today is called a dimension table. And these are tables that are referred to often as reference tables or supporting tables. And each one of these dimension tables tends to hold a specific kind of information that is accessed over and over again with some kind of list or some kind of reference. They do not tend to have any kind of patient information in them and so that generally means that you don’t need any special kinds of permissions, just basic read access can get you to a place where you can see the information that is stored in a dimension table. And it is also important to realize that this means that these tables are fairly small in size compared to the second kind of table that we’ll talk about next.

The second kind of table is referred to as a fact table. These tables hold what we refer to as measurements. They tend to be quite large in the CDW. They can be so large that, you know, really I’ve given up on ever trying to summarize all the stuff in the table in terms of records. I always have to pick a date range, and so should everyone else I suppose, if they don’t want to take down the computing system, as they’re huge. So you have to be pretty careful with these. Also they tend to have patient and staff identifiers in them, so you’ll need to request permission to be able to see these. And it’s important when you’re working with these very large tables to do as I mentioned, that I try to do, is reduce the size of the request, only look at a small portion of people, only look at a small timeframe, things like that when you’re using them because they are so incredibly large at times.

So if we go back to our simple example that we started with, we can determine whether our three tables are fact tables or dimension tables. If we look at the patient table we can see that it contains facts about the patient such as their name and their address. So that is the reason why we identify it as a fact table. If we take a look at the diagnosis table, we can see that it is really just a list of ICD-9 codes, and so we are going refer back to this set of diagnoses over and over again as new patients come in for visits and are diagnosed with various health conditions. So that list of references that we go back to over and over again is what makes the content of this table a dimension table. When we take a look at the third table about visits, we see that it contains facts such as location and date of visit and that is what defines it as a fact table.

So how would we apply this basic principle for identifying fact tables and dimension tables in the more complicated Corporate Data Warehouse? Well actually the good news is this part is super easy. Each table or view in the Corporate Data Warehouse has two parts to its name. The first part is referred to as a prefix or schema and it is really the clue or the answer as to whether the table is a fact table or a dimension table. So for any given fact table you will see the schema or prefix and in the table name. And that schema or prefix will have a name that refers back to the content of the domain. And for dimension tables you will see that same two part name, with the schema and the table name, but the schema will always be the same. It will always be ‘dim’ for dimension. So here are some examples. On the left you see fact tables and you can see that the prefix or schema in all cases refer back to the name of the domain – dental, CPRS, order, etcetera. And then for dimension tables, no matter which domains they belong to they all have the same schema, ‘dim’.

So taking this to an example of a domain that exists in the CDW, health factor, you can see that health factor contains two tables or views. The one on the left is the fact table and you can see that its schema is ‘HF’ standing for health factor, and then of course its dimension table on the right begins with ‘dim’. If we take a look at the content of these tables we can see examples of why they are fact tables or dimension tables. So for the fact table, HF, health factor content we can see right away that it has patient identifiers and staff identifiers, so sensitive information that would be protected were you would have to request access. We also see facts about the health factor type that was actually being assessed for that patient in that episode of care and we see comments about it. So those are the facts about that interaction. For the dimension table, health factor type content, we start to see lists of factors, lists of categories that the factors might go in. We see whether that factor is gender specific and we see descriptions of the factor. We don’t see any patient information, we don’t see any staff information, so we know that this is indeed what defines it as a dimension table.

Moving into the final section of today’s talk; creating an analytic dataset. On the conceptual level, what would you need to do in order to go from that relational dataset that we’ve been looking at to an analytic dataset that you could use in SAS or SPSS, whatever your statistical software of choice is?

Well, in order to arrive at that analytic dataset you have to be able to utilize joining keys. There are two types of joining keys. The first one is primary key. And a primary key is a column that exists in every table and this column’s purpose is to uniquely identify each row of information found in that table. The second kind of joining key is referred to as a foreign key. These, may be more than one of these in a table, or none in a table sometimes. And these are columns that correspond to, or reference a primary key that exists in another table. So two things to sort of come back to after these concepts are more familiar is that the values of primary keys will be unique, but the values of foreign keys will repeat. The names of a primary key and a foreign key most often match each other, but there are many occasions in the Corporate Data Warehouse where they actually have different names. So you’ll want to be aware of that and we’ll look at a few examples.

So back to our simple example. Let’s take the patient table first. In the very first column you can see that when I created this patient table I added a column called patient key. And I put it first which is the traditional location in the table. And right now I marked it PK for primary key. This is the primary key for this table because it gives a unique number to each patient listed in the table. If we take a look at the diagnosis table you can see the same tradition is being followed. That first column called ‘DxKey’, that’s the primary key for that table and it’s going to give that unique number to each one of the ICD-9 codes in the table. And the tradition is the same again in the first column of the visit table with visit key. With each visit getting its own unique ID number. Now when we take a look at the second column in the visit table called ‘DxKey’, this is our first example of a foreign key in our simple relational database here. And it is intended to allow the user to connect the visit information back to the diagnosis. So if you take a look at the first row you can see for our first visit a diagnosis code of 110.6 was applied. We take a look at the third column in the visit table. It’s another foreign key called patient key, and it is a connection back to the patient information. So if we read across the second row of the visit table we can see that for the second visit stored on this table there was a diagnosis of 377.75, that’s number four from the diagnosis table. And it was for our patient number two, Josie Frank, and it happened at station number 358 on February 2nd, 2014.

So how do you find these linking keys in the more complex circumstances of the VHAs Corporate Data Warehouse? Well, you would primarily start by coming here to the CDW's SharePoint site.

And from their homepage you would want to click CDW metadata at either one of the two links you see circled here.

Then that brings you to the metadata page and you would want to click, launch the CDW metadata report that you see here in the gray circle. And it will bring you to the metadata report. Inside the metadata report there is more than one way to look at these linking keys. The first one I’ll talk about is by clicking on the name of the domain to open what’s referred to as an ER diagram, or an entity relationship diagram. So let’s pretend that I did click on one of these domain names.

And I open up this example here of an entity relationship diagram. So in this case I clicked on the domain called pharmacy patient. And I thought it made a good example because it’s not too big, right. So in each one of these boxes you will see the content of a single table. So we have three tables and the very first column in each one of these is listed first, and each time you see a PK for primary key listed beside it, because the first column in each one of the tables is its primary key to uniquely identify each record in the table. Also you will notice that two of the boxes near the top have a dotted line connecting them. And this means that you should look for linking keys that will allow you to join these two tables. And when you start to do that you’ll find that pharmacy patient status SID is the linking key that you need to be able to accomplish that goal. Let’s look at some more examples.

So in the more complicated patient domain we have, I think it’s near to 30 tables in this domain. And you can see now that we have lots of boxes and lots of dotted lines saying that they should be able to be connected. So you, after some searching, would start to realize’ oh okay here are my keys that will allow me to connect patient to dim.race or patient.patient to dim.country’ and so on and so forth.

Let’s look at another example, going back to our health factors domain. In red you can see that we identified health factor type SID as the way to connect the two tables that are included in this entity relationship diagram. And in green you can see that I’ve highlighted a foreign key call patient SID. And what this is is a connection out to the patient domain and the patient table. And then down below in the brown box you can see another foreign key called encounter staff SID. Now from past experience I’ve come to learn that seeing staff SID at the end of this name means that I will be able to connect this to the staff tables in the staff domain.

So let’s look at the second way of identifying linking keys using the metadata report created by the Corporate Data Warehouse. So instead of actually clicking on the name of the domain we’re going to hit the ‘+’ sign next to the domain and expand the more detailed table level summaries that exist for each domain.

So going back to our example with health factors, we would expand the plus sign and we would see our two now familiar tables are part of the health factors domain. And we would focus our attention to the far right under the column called relationship. And then we would have to pick whether we wanted to click on the linking keys for the dimension table or the linking keys for the health factor, because you can only see them one at a time. In this case I’m going to select the fact table.

That opens up our table showing us all of the linking keys that exist for this table, HF, health factor. So if we focus our attention on the first row you can read across and it will tell you how to connect two tables. So we’re starting with our fact table, HF, health factor and it has a foreign key in it called encounter staff SID. And it will connect you to staff.staff which is in the staff domain on its primary key, staff SID. So notice here that the two joining keys do not have the same name, so the only way to be sure that this was the correct joining would be to check this table. Let’s take a look at another example. We’re connecting health factor fact table using its foreign key patient SID to patient.patient on patient SID.

So to summarize, today’s talk should have illustrated the very large variety of data that we now have in our Corporate Data Warehouse and all of the different kinds of questions that we might explore using these data. Currently most of the data found in this comes from VISTA either directly or indirectly because most of our derived datasets also have VISTA as its primary source like HERC and MCA. They got their data from VISTA turned it in to MCA and then went back into the warehouse as well. So really VISTA is the primary source. As the data are coming in they are being broken into the domains we discussed, the tables we discussed, and some are fact tables, others are dimension tables. And it is our job as end users to learn how to bring these back together using joining keys. And in order to do that we’ll need some SQL skills to recreate our analytic sets of data and do our analyses.

So I’d be happy to take any questions.

Hira: Hi Margaret, thank you for your presentation to the audience. We still have about 15 minutes left in our time. And if you have any additional questions for Margaret please send those in. So Margaret I’ll get started with this list of questions we’ve got. The first question, why are scrambled SSNs versus real SSNs sometimes used in the raw CDW?

Dr. Gonsoulin: I would assume it depends on the source of the data that are found in raw. So depending on who collects the information they may have chosen that identifier as the primary identifier for the data. So like, you know, in MedSAS for example, they primarily chose scrambled SSN as the identifier that they used when they created their dataset. And I believe various offices contribute data into the warehouse and they may have made that choice. So I think primarily that’s where that comes from. Because any VISTA sourced data would have patient ICN and SSN as far as I’ve ever seen. So I’m assuming it’s the office that created it.

Hira: Alright, thank you. There is a couple questions here I’m seeing about people who are having trouble getting to the CDW metadata report. I’m also having trouble getting on there. Do you know if there is some kind of like technical issue going on with the website?

Dr. Gonsoulin: I did hear that. Actually today they are having a CDW insight stay and they mentioned difficulty with their SharePoint site. So I think it’s just temporary and they are aware of it.

Hira: Alright, thank you. Moving on to the next question, is there a document or plan for translating raw to production domains? Do you know what’s next on the list?

Dr. Gonsoulin: Yes. Actually you can visit the CDW when it’s up, the CDW SharePoint site and I’m trying to go back to the, to here. So if you can see my screen still, you see on the top right hand side, there is a box called ‘What’s in the Works’. And there is a link; it’s the third one down called ‘CDW domain status and priority’. You can click on this link and it will open up an excel spreadsheet and it will walk you through what’s in the queue, what they are working on, what’s coming next, how far along they are in terms of progress toward making a domain a production domain versus a raw domain, things like that. So it’s posted on their SharePoint site.

Hira: Alright, thank you. Is there a codebook that identifies or describes what data is in a column?

Dr. Gonsoulin: Yes, that’s the subject of next week’s talk. There are multiple books, if you will. So on the one hand you could go into CDW’s metadata report here that we’re looking at on the screen. You can drill down to each table and each column and I’ll walk you through that next Wednesday if you are able to make it. In CDW: Locating its Documentation. Also here at VIReC we have been producing documentation of some of the domains. Obviously there are many and we have not got a complete set. And also whenever you’re talking about raw, which I’ll cover in next week’s talk as well, the reports look very different and often times if I really want to know what’s in a raw domain I would go to the data architecture repository and look at VISTA’s metadata, so I will step everyone through that process and show them like ‘here’s what you can do’. In addition, I mention in next week talk, but really I’m creating a new talk about searching the meta-views that exist in the data themselves. So in the CDW’s data there are views or tables that have a prefix or schema called ‘meta’ and that’s metadata views and so I’m putting together a talk, but we don’t have a date for it yet, that walks people through the process of querying the metadata to search for information that they need.

Hira: Alright, I had a lot more questions coming in, so I’m just going to continue rolling through these. What are the tables that go into creating the main and bed section inpatient MedSAS data? Since MedSAS datasets will be going away, I’m trying to figure out how I can create them.

Dr. Gonsoulin: Oh, that is a very complicated question. I’m not sure we can really address it in this context, or even that I would be the right person. But I would like to say that lots tables both in MedSAS and inpatient and outpatients, I think I estimated one of them one time was, I think it was IE was 28 different tables, to create MedSAS IE. But I would like to reassure the listeners that I think ‘go away’ is not really an accurate phrase to describe processes that are being considered or that are, or that are being taken action on right now. So right now they are replacing, not going away, but replacing MedSAS outpatient in the context of the Corporate Data Warehouse. There are no plans right now around inpatient that I’ve heard of, and when I say replacing I mean absolutely recreating all of the logic and all of the filtering that went into the creation of MedSAS. But they are doing it in the context of the Corporate Data Warehouse instead of going through National Patient Care Database and the HL7 messaging system that they used to use in the past. They’re just using the Corporate Data Warehouse and recreating the exact variables that you are accustomed to seeing in MedSAS outpatient. So it will be there still.

Hira: Alright, thank you.

Dr. Gonsoulin: Thank you. But I would be happy to talk about that off line if the person would like.

Hira: Okay. So next question, is the patient ICN still the gold standard for linking patient level data, or is patient SID and station still used?

Dr. Gonsoulin: Well, I think ICN is still considered the gold standard in every report that I’ve read. There are perhaps instances where for various reasons it is not available, maybe? So there may be cases where the ideals aren’t available or scenarios where that happens.

Hira: Okay. Is there a way to find out all the names of tables? I think there is a second part to this question. So we can select a specific diagnosis and clinic, etcetera.

Dr. Gonsoulin: So, yes, we do run a quick list of tables. I will show you that in next week’s talk as well. We have a product that periodically just goes and gets all of the names of the production tables. So if you want a list of all the names of both production and raw, I don’t know of a place where that is laid out particularly. But I will show you where you can find the definitive, most up to date list of domains, and for both production and raw in the next talk.

Hira: Okay, are there advantages to using the CDW over the MedSAS files? If what I want is in MedSAS, why would I want to recreate that using CDW?

Dr. Gonsoulin: Well, I mean I think it’s very particular to your subject, you know, what it is that you are trying to discover. So for example with MedSAS it’s only going to include records. Like one of the filtering mechanisms will exclude anything that is not considered work load. So for example if you had some kind of scenario where you wanted to know about records of like, of occasions of service or things that did not end up counting as work load records in MedSAS you could not achieve your goal with MedSAS. You would need to go to CDW’s outpat visit and include all the records. So I think it’s not really that one is better and the other is worse in my humble opinion. It’s that one may be better or worse given your need, what you need to measure. And so just sort of spending the time kind of trying to understand what the underlying relationships are and how those translate to your study needs is really where the answer lies.

Hira: Okay, thank you. We still have five minutes and a few more questions to get through. You mentioned that the raw domain gets updated at varying times. Is there some sort of schedule available that would give us the update schedule per table, and/or the last time the table was updated?

Dr. Gonsoulin: Yes. Again I would just go to the patient now but I know that there are problems with the SharePoint site so I’m just sort of saying, it is in next week’s talk and I will show exactly where to look to find those schedules. And it’s, you start here from the homepage that you see on the web and you would start in the CDW tab at the top and go to the raw domains and then it’s sort of in a table. And it just shows you, ‘this is every two weeks; this is every month, this one’s…’

Hira: Does that also show you who owns a view? So that if someone notices a problem with it they can contact the owner?

Dr. Gonsoulin: Well, I mean I guess the concept of ownership is, I’m not sure if there is a consistent definition of that. You know, with regard to domains from VISTA sometimes there really is a very straightforward group of, or office in the VA that has a very close relationship to the domain and the tables in it like you know, not just to call anybody out but the National Surgery Office is very close to the surgery data from VISTA. But like for example, with the patient domain, you know, I’m not sure that there is a clear group that would be in charge of that. However you can see often times in the documentation from CDW domain teams, groups of people from various offices who put in their expertise in building those domains and they may have information to share. And then on top of that, too, if you see a problem or whatever you it may need to go to the architect team. They may not be aware of it or, there is all sorts of scenarios. So depending on your exact need, like if your office saw a problem and wanted to request additional data be in the Warehouse, then that would track through the governance board for CDW. So it depends on your needs and the particular part of the data that you’re referring to I think.

Hira: Alright, and hopefully we can get through a couple more questions. Where can one find a list of all domains with associated tables and variables within each table to make an assessment about whether VA data is appropriate for a given study? Is this information available publicly?

Dr. Gonsoulin: Well, right now all of the lists of domains and tables and things that I am aware of are on the intranet sites, really, plural. So some on the CDW’s intranet site, some on VIReC’s intranet site. Depending on exactly what you’re after maybe HARPs intranet site, or MCA’s intranet site so I don’t know of any internet material per se. So if by public, it would need to be on the internet, I’m not sure it’s there right now.

Hira: Okay, maybe one last question before we wrap things up. If raw domains and tables do not have linking keys can they be linked at patient level to the production domains and tables?

Dr. Gonsoulin: Yes, they can. There is a great querying CDW raw guide that CDW made so that you can read through this guide and see. But I mean it’s a process and that’s not to mention that raw is on a separate server from production. So all the production data is on server one and all the raw domain is on server two. So you not only have issues of joining but you also have issues of getting the data to the same server I think.

Hira: Alright, Margaret thank you so much for taking the time to present today’s session. We still have several questions that we weren’t able to get to but I will send those to you. And the asker can also contact the VIReC help desk at VIReC@va.gov. If you still have additional questions. As Margaret has mentioned, she will be presenting another CDW session next week on Wednesday April 4th at 12pm Eastern. In this session she will be talking about CDW and locating its documentation. It looks like several of you will be trying to join us and we look forward to seeing you there.

[END OF AUDIO]