Cyber Seminar Transcript
Date: 10/13/2016
Series: VINCI
Session: Full Grid Project Development with Enterprise Guide and GSUB, and Enterprise Miner Analysis
Presenter: Mark Ezzo
 *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 www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm.*

Heidi: We’re just at the top of the hour here, so we’ll get things started. Again, thank you everyone for joining us for today’s VINCI cyber seminar. This is part of the VA Informatics and Computing Infrastructure Cyber Seminar series, and today’s session is Full Grid Project Development with Enterprise Guide and Gsub and Enterprise Miner Analysis.

Today’s presenter is Mark Ezzo. Mark is a VINCI SAS administrator at the Salt Lake City, Utah, VA, OINT Servicing and Engineering. And, Mark, can I turn things over to you?

Mark: With pleasure. Let me show my screen.

Heidi: Perfect.

Mark: And, you put this in presentation mode, which that was not it. Ooh, that’s too much.

Heidi: You’re actually, if you click just to the left of that 100%, there’s that little icon there, that one right there, that’s what you want, just like that.

Mark: Cool beans. All right. I’m going to begin now. We have, as Heidi said, we’re doing Full Grid Development with Enterprise Guide, which is what we’re going to use for data preparation and what we recommend that everyone uses for data preparation. And, you can also use it quite a bit for analysis, too, but we’re going to do our analysis today in Enterprise Miner. We’re also going to look at Gsub, which essentially is a grids batch facility. Now, Gsub is very, very useful just as a preliminary, because it takes the Windows environment away and allows you to just run within the grid. So, you can run large batch jobs and you put them in overnight and have no worries. It’ll give you a login of this file.

Let’s go down. As I said, Complete Project Development, Enterprise Guide, Gsub, Enterprise Miner. I encourage you to ask questions at any time. Heidi, you want to put the poll questions up?

Heidi: Sorry about that, the first software I have, the first poll question I have here, the software, oh, which software do you have experience using? I can’t write them here, but let us know which ones you have experience using and we will—responses are coming in nicely. I know this one may take an extra second or two, because everybody’s got to evaluate each one. But, I’ll give everyone just a few more moments to respond and we will..

Mark: Right. We just want to get an idea of, of what everyone’s using out there and I personally encourage you to use whatever tool works best, or use them in tandem, which is what a lot of folks are doing. And, if we have a little bit of time, I’m going to give you a preview of how we use R [PH] within SAS, too, a very, very quick, small demo.

Heidi: Very cool. So, what we’re seeing for the audience here, we’re seeing 61% saying that they use SAS, 24% SQL, 10% SPSF and 5% are zero data users.

Mark: Okay. Cool. Let’s advance and then we have another question I believe.

Heidi: And, here we are looking for your level of expertise with SAS, and then we have one last poll question once we get through one here. Looks like we finished off here, so what we are seeing is 21% experts—fantastic, you guys can send in great questions for Mark—48% intermediate, and 30% beginner. And, our last poll question that we have, we’re wondering do you use the grid or just desktop SAS.

Mark: Or, both.

Heidi: Oh, sorry, that is an option, so grid only, desktop only, or both. Looks like the majority of people are actually using both.

Mark: All right. And, that’s as would be expected, sure.

Heidi: So, we’ll close that out, and what we are seeing is 11% saying grid only, 42% desktop only, and 47% both. Thank you, everyone.

Mark: Wonderful. Okay, let’s talk about Enterprise Guide as preliminary. What Enterprise Guide uses are, program development and testing. Essentially, you can, I always encourage the use of samples rather than full dataset, but you can, it’s a very iterate tool. You can develop programs as the Enterprise Guide users know, you can develop them within projects, you can do flows. It has wonderful wizards where, allow you to develop the code automatically. It’s a very robust tool. You can do quite a bit of analysis on that also. I mean, it can be used for either, but we’re going to use it mostly, essentially today for program development testing and data. Data development and interpolation, essentially bring your data down. I do this all the time. You pull the data that you want and you look at it, you make sure, you create your segmentations in there, whatever you need to do. It’s something, and we have much, much more room on the grid than you folks will out in the Windows world. I think you’re limited to 100 to 300 gigabytes. We have about 76 terabytes out on the grid and we will continue to grow. Again, as I mentioned, it’s best used with smaller samples rather production data. Save your production data for later. Program development, program debugging and I like to, when I have a process that works, something that’s, say production ready or something I’m going to use over and over, I like to put that into Gsub, which we will see. And, now let’s do a bit of a live demo.

Now, I do this on App 15, this will work very, very well in VINCI. And, the way I do it, and I’m sure all you know, is I essentially just bring up—I like, and people who are in VINCI, if you’re going to use Enterprise Guide , do not click on the icons. Go, go in there and then RDP over to VHA CDW App 06. This cuts out quite a bit of, of the network and allows you to work, to work on a VN that’s exclusively devoted to Enterprise Guide. It’s far better. So, again, I click here and this would bring this up. And, after you click into Enterprise Guide, the thing you want to do—no one will have as much data as I do, so don’t worry about that—is you just want to expand SAS App, because that’s where you’re going to run all of your programs. And, then you essentially have a process flow, which shows you all the programs that you’ve done. And, you can run them in any order that you like, essentially. As I said before, we have wizards here, so if you the intermediate or, or beginning folks, I strongly encourage you to use some of the wizards. Because, what the wizards do, they create code for you, and it does things like describe, you can do graphs. Graphs, for me, is, and to be quite candid, and that SAS people won’t like me saying this, but graphs, to me, is not the strongest portion of SAS. It’s robust, but it’s not as easy as other products to use. And, we have all sorts of analysis here, and it’s a point and click environment where you can just fill in the blanks and it’ll produce code for you. And, if you have any question, and then you can use that over and over again. Now, if you desire to do it old school, which many of us do, you can say new program and code to your heart’s content, as if you were desktop SAS, and it will work just as well. So, there’s no issues to do it either way.

So, getting back, what have we done so far? We know that we have tasks on here that can do many, many things. We can control our environment, for example, just by going to options and you can, for example, let’s say if I wanted to put out a PDF. I merely click here and it will, I can download PDFs, I can download RTF, however I want to do it. Or, I can use ODS, which is quite powerful, too. I don’t want to go through all the options here, I’m just more interested in doing a project at this point. But, I encourage you all to take a look at the options here, and then you’ll get a lot of power out of that. There’s many things you can control. You can put custom code in per project, etc., etc., but a very, very powerful tool. And, just as a heads-up, SAS, since Version 9 came out, essentially is not going to do any desktop enhancement. We’ll always have the desktop tool, but what you see is essentially, at this point, about four or five years old and will never, never move forward. They’re putting out everything into the UI’s like SAS Enterprise Guide, Enterprise Miner and all their products. Okay.

Now, after you click on SAS App, you can expand your libraries, and the way we set things up, we encourage you to—we have to set the project up in a SAS Metadata for the grid. We essentially enable all the people, we give it a name, we set the authority on it, and so you only see the libraries for which you have access. I, of course, have access to practically everything. And, that way we have very, very strong security we have here within SAS and Metadata, and we also have it in the Linux OS environment via Accels [PH]. So, it’s about as security, really, as you can get at this point. So, contact us if you, and to get it set up, because we have to add you to the security zones, and we have to enable the project and create the project groups so that no one else but you can see the data that you have access to.

Now, let’s go into programs. Up here, this is all data, this is data that I’ve looked at and I can build query builders off of them, which we’ll take a look at in a little bit. Or, these are programs I started. I always, for a project, have a little start project here where I can put in all my options. Like, in this case, I’m putting compress equals yes, and full S timer, that, not necessary to do. I do that so that I can benchmark. This is a little bit more unique. This is for extreme benchmarking, and this sets up my library. This compress equals yes, please, everyone, use that option, and I’ll tell you why. You’ll see that when I pull data, this can shrink by about 80% in many cases. But, you can double or triple the amount of storage, use of the storage that you have just with that very simple option. Okay? I encourage everyone to use it and we may force it in the future, who knows. So, first things first, I set up my environment. Now, I know I have my live name, AAA, and why do we call this AAA? So, let’s just run this, because when I refresh my libraries, it comes to the top and I have, don’t have to go searching for it. You can call it anything you like, this is just something that I do. And, that will show me all the data I have within that library, and I can just click on it to bring it up if I so chose. So, I click here, it comes up and it’s wide open. So, what have we done here? We have set up our options and we have initialized our library.

Now, second part, I needed some data. So, how do we want to do data? You probably, I select the top two million rows of the pass-through. Now, we encourage, Kevin, Tony, myself, constantly, that you use SQL pass-through and you can generate that automatically just on a query off of a dataset, too. But, this essentially is how it works. It connects to SQL Server, or whatever database, it has the options all ready to go for you, passes your credentials over. Okay? And, then when I say create table AAA, as you saw before, AAA is my live name and I’m creating a table, B Diagnosis 2, from, and in this case, I put select top two million. Now, if you’re going to limit the number of observations using pass-through, anything you put in the SAS, obs equals two million, in obs equals two million, out obs equals two million will not work until all of the data comes down. So, I’m using the, quite a large table here, Out Pat B Diagnosis, I mean, that’s a lot, a few billion rows. So, if I don’t put top two million records, it’s going to wait until all the billions of rows come down with all the columns before it begins to add any sort of SAS culling or filtering. So, be very, very wary of that. But, what we want to see from people essentially is that you use things for SQL within SQL. In this case, I say select top two million rows. Once it’s in the SAS world, everything in SAS is able. Now, if I run that, and I’ll show you exactly what I mean, I ran that and it ran in just seconds, actually. Yep. About 19 seconds it dropped down two million rows. And, as you can also see, it decreased size by 81.26%, compressed. If it was uncompressed, so it’s essentially 5431 pages, uncompressed would be almost 29,000. Strongly encourage usage of that.

So, here I’ve dropped my data down. Now, this essentially is code from the query builder. When I build it, all you need to do is go into the query builder, and let’s do one as an example. If I come here, it shows me my input data, my code, and to save it as code in the program, just do anything, just click on it, this goes do you want, it’s read only, do you want to create a copy that can be modified, you say yes. It now adds it to your project as a SAS program. Very, very useful. And, that’s pretty much how I do it. Now, make sure you put some filtering in there. Don’t, don’t just drag everything over like it’s a select star. That’s where a lot of issues and the SQL Server people will rightfully yell at you.

So, I have code for query, and I could enable this code and it, just to point it out, really, because when you grid enable, you essentially go to analyze program, analyze for grid computing, and just \_\_\_\_\_ [0:15:39] on grid and begin analysis. Since this is already grid enabled, I don’t, and then I merely just added to, to the project. And, this is only a count of one, but I’m going to show you something that’s multiple, that you can do an ETL. One very useful feature of a grid if not the most useful feature, besides security and failover, is that you can run in parallel. I am not forced to run in a linear manner. I don’t have to run A, B, C, D and E, I run everything as A. So, if I have five programs and they’re all pulling tables and each table takes one hour, if I do it in a linear manner and I don’t grid enable, then essentially I have a five-hour job. If I grid enable, it takes as long as the longest job. So, if each one takes one hour, I go from a five-hour job to a one-hour job.

Now, let’s look at things we’ve done here. In here and I’m going to create a sample of 100,000 randomly selected records, and I’m going to lit up and then join it together. So, I look to see what I had up there and again, I just \_\_\_\_\_ [0:16:48] that we have our options compress equals yes, we have our live name. I put those in sometimes if I’m going to run the program standalone. I’ve prepped datasets and I wanted to know the, I wanted to get rid of a table just to make sure up into SAS task, which is my own area on SQL Server. You can use prox datasets or you can use proxSQL with an exact pass-through and a drop, either one will work exactly the same. Then, because I have defined, and you folks in VINCI know this in your default areas, I can just load up data, just say data SAS \_\_\_\_\_ [0:17:26] which is the live name, B Diagnosis, and I load my SAS data up there completely. Okay? So, now I load the original table up there.

Now, I want to get 100,000 records. So, I do a prox sort on this and all I want are patient SIDs and I don’t want any duplicates, so I say no dup key, by patient SID and I output a SAS work dataset called SRS. And, there’s a reason for that. I create a sample of 100,000 randomly selected patient SIDs and I’m using prox survey select, data SRS, which is what I have here. So, essentially, I have about 1.08 and change million, but all I want are a thousand, and this is a simple random selection. And, I create an output dataset that’s sample SRS. So, after I run that, it essentially does a simple random sample, and selects 100,000 unique records. Okay. I verify that, because I’m from Missouri, and I put no dup key and no dups were dropped, essentially, I just sorted it by patient SID.

Now, what happens, I created a table upon SAS \_\_\_\_\_ [0:18:45] and you can do that in your default area, too, the same way as a lead table. Meaning, I am going to join against a lot of data, but I don’t want to have to, I don’t want to drop 200 billion rows, I just really want to get these 100,000 as a sample. So, I load this up and you can either do it, if you’re doing it on like in CDW, then this would be a temp table, otherwise, I just loaded it as a table called TP. And, then, because I now have that table up there, I now can look, and this is all done within SQL Server, you recognize the pass-through. I now have pass-through here and I’m going to create a SAS dataset of 100,000 randomly selected by joining, I’m going to pick all, everything I have up here, because I selected that earlier to the coms I wanted, and I’m going to do an inner join on TP by patient ID equals patient ID. And, when I’m done, I have, I have 100,000 randomly selected patients and the records. Now, obviously, because these are outpatient records, one patient can generate 5, 10, 15 records, so I’m getting quite a bit more. I actually average about 9.6 records per patient.

So, let’s review what we’ve done here very quickly. We have, we are going to put a table up there, we had that all set up in a live name, or you can just use the live name off of Metadata, which this SAS has it’s the same thing. We essentially got rid of the old one, put in the new one, sorted so we have unique patient SIDs, selected 100,000, verified that we had 100,000 unique ones. Then we just keep it clean, got rid of the old table and make sure that that’s all very good. We could put a table up there and then we joined against the two.

So, if you have, if you’re going to do a study and you’ve selected people like \_\_\_\_\_ [0:20:56] ICD 9 codes and you want to check them individually, this is really the way to do it. You pick under, I use patient SID here, it could be anything. This way you present a lead table up in SQL Server, it’ll save you tons of time and processing and storage, and because I’m only returning those records that I wish.

Any questions at this point? All right. Very good.

Now, let’s go to, see what I had for program two. I was just looking at things, so no problems there. Now, I also wish, very similar, I’m creating the table again from my, from my sample. Now I wish to add information, essentially demographic information from a very large table called S Patient. All right? But, if I were to drop that down and use it in SAS, I’m waiting here quite a while and using quite a bit of storage, even if I’m only doing it in the work area. So, exactly the same way, I have S Patient, I’m doing a right join onto the table again, and this gives me precisely what I’m looking for. So, I created, I called it S Patient SRS, so now I have my S Patient data there, same, same 100,000 people. Now, I want to, for demographics, I have B Diag Demonk [PH], where I’m calling in everything I have there that I want to keep. I’ve culled it down a little bit from the first table, and everything from the second table, which is S Patient table. So, I have all my information and I create that into the SAS world. All right? And, then I figure, I put in age grouping, which I can do in SAS, or I could’ve done in case. I actually, my, my thing is I like to bring it down into SAS and use it there, because I have all the robustness of SQL in SAS and I also have all the ability of SAS datasets. So, here I put my grouping in and then I format my ages, and now I do a little—and, now when we say we’re preparing the data, this is one of the great uses of EG 2. I can look at my data in an iterative process, so now I have, I can make a dataset A, I want to see what kind of agent orange exposure I have. I’m essentially looking to see how it all works out, and I make decisions from there. So, I’m essentially making, I’m going to add the ages to VD age and then I can join A, what I made out for A, so I get the correct period of service that I’m looking for. In this case, because we’re, we, I decided to do an agent orange study, we concentrated on Korean War, Vietnam War and post-Vietnam War. And, then I did a little bit of checking out there, and if you look at the results, I actually ran a very, very simple prox summary. And, I used end way, which means me a highest classification combination. I want to include missing data, so I want to look at period of service, age group and agent orange exposure flag. And, that’s what X was. And, I can see here that we have a lot of missing variable, we find it in Korea. By age group, we get to see who the yes and the no’s are all the way down the line. All right? So, this enables me to get a good idea of what I’ll be seeing as far as agent orange. For example, people who served in Vietnam area were between 55 and 59. We’re looking to see a very large, several more instances of agent orange than we would in other places. And, you’ll see, obviously, Vietnam is, is the big winner here, or loser, really, to be quite honest.

So, when I do that and I have that in my \_\_\_\_\_ [0:25:15]. Just as an example, and I’m not going to do a class on hash objects right now—Kevin has given several classes and I believe he’s bookmarked a lot of them—but, hash objects are very, very useful. You can do merging, you can do sorting, you can create output datasets, all in memory with the hash object. So, here, just to display it, here I create this. Again, I just took the top two million rows, very simple, and then I do a sort. So, if I want to do a sort by patient SID, event date/time, and then this essentially is a hash object which emulates the sort. It’s setting up—excuse me while I kill that—it’s essentially setting up every—I create the hash object. I can order it, I say yes. I put multi-data in, because if you don’t, you may not get everything returned. I suggest using multi-data if you have any questions about it. Say, you started with two million and you get less, if you put multi-data in, then it will, it will guarantee you. I define my keys as I did up here. I want all the data, and I outputted that. And, that all worked out. See if I’ve saved a log on that. So, this dropped down, again, we, this \_\_\_\_\_ [0:26:45] in under 13 seconds, and then I did the sort. Now, you haven’t seen this case. The sort actually outperformed the hash object. Not unusual, because if, if, this, it can pretty much go either way, but it really performs very, very well in merges as opposed to this. But, the sort only took about nine seconds, and the hash object took about 12 seconds, you know. And, it’s, that’s not going to be a 25% difference going forward, but, so I just wanted to do that as an example. Both datasets are identical.

So, what have we learned here? We learned that we can pull the data, we can do, we can sort it if we like. Now, be careful about order by sometimes in SQL Server. I mean, they work fine and they do precisely the same thing, and I can do a first \_\_\_\_\_ [0:27:39] or last \_\_\_\_\_ [0:27:40] processing with an order by. But, depending upon the usage of SQL Server in a temps table space, that might take significantly longer than just by bringing it down in SAS and doing a sort that way. Just, I find that to be quite frequently true. And, another reason is also when you, why you want to pull the data is SQL Server uses a lot of views, but views have to adjudicate in the background. So, if I’m always going against a SQL Server table, I’m hitting that view continuously, and if the view is not optimized, I can add hours or days onto a study if I just, instead of just selecting my data and bringing it down to the SAS world. Okay, folks?

Now, again, going forward, just to prepare, I do certain things like prox freaks [PH] to see what, how many combat friends I have. I do summaries again, and then I take a look at my data to see how many I have, and what we did earlier with the 100,000 records turned out to be that. So, at this point, I am just interpolating my data. So, I’m looking at, I’m looking at everything, and this is where you make the preparations. So, I’m just selecting data here, and then I put, and I do something here I call data prep. And, this is something I do quite commonly. I essentially look at my data and prepare my data accordingly in a small module called data prep. So, I’m looking at everything, checking it out, and then I want to make, because I want to make sure that my data is relevant for when I’m doing modeling. And, then after doing everything I do with prox summaries, very simple SAS procedures, descriptive statistics, I determine that all I want for period of service is Korean, Persian Gulf War and Vietnam and I make that a one and a zero. So, for all you statisticians, I think you know where I’m going with the binary. And, then when we look at our results, I see, I have mostly all ones here and zeros a little bit. And, now I’m ready to go. Okay?

\_\_\_\_\_ [0:30:09] time for it, and what this is, this is a temp, this is something I think I encourage also very strongly. This is a temporary table and do you folks hear noise in the background? I have some contractors over here. I apologize for that. So, this is a temporary table actually going into the CDW world. Hold on a minute, let me \_\_\_\_\_ [0:30:34]. And, I can create a temporary table, which will only stay there for the session. And, again, I’m doing it with my sample SRS, and I create my tunnel and I do creation here. And, I want to do this twice. Now, again, to the point we brought up earlier about gridding, if I were to run this in this mode, it will run one after the other. First, it’ll run the first table, then it’ll run the second table. And, let’s say these, these don’t, but let’s say these tables took about two or three hours to run, and essentially I have a six, six-hour job. So, essentially, I just grid enable it, this is count two and I run this, and it will run then both simultaneously. I create my first table, SR1, and I create my second table, they’re identical tables, SR2. And, the grid brought them up, so instead of me running on job, I run two jobs on the grid. You can run several jobs on the grid and we strongly encourage that. And, this was one I was just running and going through. This was code for query builder.

So, do I have any questions about Enterprise Guide and the usage that I had for it in creating data? I use it solely here to create data in very simple descriptive statistics.

Heidi: I do not see any on that right now. We do have some other pending questions, but we can hold those until later if you would prefer.

Mark: No, let’s, let’s take a few of them now.

Heidi: Okay. One that just came in, will we be able to get any of these sample code for reference?

Mark: Oh, sure. I’m going to make this available for everybody. We’ll put this out on the website.

Heidi: Okay. Next one here, we just had a suggestion or an FYI sent in. Your hash object had two million records. You’d get faster results if you increased the storage bucket \_\_\_\_\_ [0:32:41] statement up to something higher than the default eight bucket. Hash X 20 is the current highest you can go.

Mark: Okay. No problem. I actually put in, I actually put in data all, I, I, I’m, so I missed one. But, we’ll discuss, we’ll take that offline. That individual can contact me. I’d be happy to talk about hash objects with him at any time.

Now, \_\_\_\_\_ [0:33:09] let me go back…

Heidi: \_\_\_\_\_ [0:33:11]?

Mark: Go, fire away.

Heidi: Okay. If you have a data library set up for you with the project, do you need to use the pass-through?

Mark: No, you do not. Pass-through is used, you can use that to, for work data or you can pull it out to your permanent data, which is what the data library is there. If I just wanted to see this data and connect it to something else, that’s fine. I can use a temporary work area, or I can save it. I strongly suggest you save it, because you can use it over. I equally strongly \_\_\_\_\_ [0:33:46] you clean up.

So, that was Enterprise Guide and then \_\_\_\_\_ [0:33:54] progress a bit. Now, let’s talk about Gsub, which we’re going to look at next. Gsub uses are large jobs, long jobs, especially overnight, production quality, meaning you really don’t want to test in Gsub too much unless you’re doing very large sets of data and testing and you don’t want to clog everything up. We use it with Linux access products, like Putty, I use MobaXterm and we also use WinSCP to look at our results. It produces log and LST output files, and can run anything you do in EG, ODS reports, parallel jobs, etc., and let’s do another live demo.

So, because, I did not leave this up, because you only get about 10 minutes to get into it. So, I’m going to come here and I get in as myself. Now, most of the time I’ll just say history and look for my last Gsub here, it was my last one up. The history \_\_\_\_\_ [0:35:11] will show you anything you’ve put in, so if you ever are in Linux and you want to remember something that you did, type in history and you can also type that over into a file, and that will show you everything you’ve done in the past. So, essentially, here I up-arrowed twice and I say let’s run this, sasgsub and the way this works is we say grid snip program. You have to show it where it’s at even if you happen to be in the directory. So, you have to \_\_\_\_\_ [0:35:45] say to my programs area, that’s the name of the program. Then you hit enter and I type in my password and it submits. Now, it’s going to show me the job directory and everything in Gsub goes out to our grid share area. So, here, and it’s also under your username. So, all I \_\_\_\_\_ [0:36:12] do we haven’t saved it, is just come here, copy and I call up WinSEP and \_\_\_\_\_ [0:36:24]. I go to a node and I have to reconnect, because it always forces you in that particular node \_\_\_\_\_ [0:36:33]. And, now if I refresh and if I go here, here’s my grid share area, here’s my latest one, which as you can see, 137, 136, and \_\_\_\_\_ [0:36:57] job ID 154, I’m running it on this host and it’s creating a log for me, which is in the process of running. So, that’s what we do then.

Now, just to show you what happened earlier, here’s the log and it’s, if I go down to the end, because there’s a lot of stuff \_\_\_\_\_ [0:37:21] here. It’s showing me all the \_\_\_\_\_ [0:37:21] I have here. It shows me the complete log here and it ran both jobs simultaneously, and ended those and closed down all the sessions. So, and the time it took to create one table, I create two. Now, this could be 10 tables, this could be 15 tables, very, very powerful \_\_\_\_\_ [0:37:47]. I, without kidding, I have had people who have had jobs, ETL jobs and fiscal jobs run for days, a couple of days, and we knocked it down to a couple of hours with a few modifications, which parallel was the center of it. So, how do we get to it? We got in here, we went to, in this case, I went to Gsub 2 and in VINCI you’d got to Gsub 3B. I put in the command, and I’ll show you this in our presentation, and then when everything was ready, I went out and came out here. And, I get a log file and if I put any data then I got \_\_\_\_\_ [0:38:32] or if I did, if I created a graph or anything, or \_\_\_\_\_ [0:38:35] statement, that would come out as LST file. All right?

And, again, let’s go back and read a few things. If we’re looking here, this is the command we use for batch processing. \_\_\_\_\_ [0:38:56] this is the simplest. There’s several options. We have that all documented on our SharePoint site. Sasgsub –GRIDSUBMITPGM, whatever program you call it. It breaks down as follows. The sasgsub is the actual command that initiates this. –GRIDSUBMITPGM, I’m saying I want to submit it. My location, my program name, and we did our live demo. And, we did a good one, because what we did, we enable it in batch here on this one, grid enabled, saved the job. And, because it runs a little bit longer, we ran it out here and this way I can continue working in SAS EG without blocking it, blocking the session, having to worry about the communication between Windows and Linux being interrupted. This is fully within the enclosed Linux network. Okay? So, many of you folks who have said I read a job overnight, it didn’t run, if you’re running it in EG, that’s quite likely the reason. If there’s any sort of network interruption, it can hold everything. If you do it here, that’s not going to happen.

Now, we have about 20 minutes left, so I want to go into Enterprise Miner. Let’s come back to our little guy here. Now, we’re going to talk about Enterprise Miner. Now, this is a tool that is out there growing leaps and bounds. The last few years the product market for it increased about 100% eash year. Extremely popular, very, very useful. It, essentially, it’s a drag and drop environment and all you need to do is understand statistics, fill in the blanks and select the methods you want to use and the flow of the program, and you can create data model, statistical model, extremely powerful. And, the uses for Enterprise Miner are analysis, model creation, model testing, interpolation again, model publishing. You can share these models with your project group. It has been proclaimed the leading edge tool as far as data modeling and data mining, etc., etc. It’s…

[audio interrupted with music]

\_\_\_\_\_ [0:41:30] diagram that you created before. Now, to get your data in, now, I don’t even do that with one of these. But, say I wanted to do this one all over again, I merely drag it and drop it. Now, my data is in my diagram. I can delete it, I say yes, or do anything on it. Now, everything in Enterprise Miner is controlled here in properties. Oh, one very, very important thing, because I’m using SAS data as we talked about before, I have to put my library in my project start code. So, if I click on my ellipsis, this is how I—I would not recommend going directly against SQL Server, but preparing all your data as we did in EG, or Gsub. And, essentially, I want to compress it, I want to put my same live name up there, and I just say okay, because I’ve run this before. I, and then I, essentially I run that so that when I go and if I want to create a new data source, it will open my data source wizard and I’m going for a SAS table. I can browse and I have everything on Metadata, each and every single \_\_\_\_\_ [0:43:07] here, but I just want to use my SAS data there. So, I click on AA and I can select anything from here. Okay? And, it goes through a series of questions. I can—and, we’ll look at a little bit of that later—I can select whatever I like and then answer questions about how I wish to use the data, targets, binaries, output, etc. I’m just, I’m going to cancel out of here, and we say yes and I did that already with our friend agent orange out pack [PH]. And, the output role, I can, I put it up to rerun anytime. I’m not going to summarize. If I want to change anything in there, I can edit my variables.

Since we’re approaching 12 minutes before the hour, I’m going to do less of that and just, just run through this a little more quickly, because I want to save a little bit of time for questions. I, essentially, as I showed you before when I prepared my data. I want to make agent orange a target. I want to make period of service a target. This is input, patient SID is an interval. I want to use AO as a prediction. I can edit this using SAS code. I can pretty much do anything I like. I drop, obviously, I really don’t care about patient SID, so I drop that out. So, \_\_\_\_\_ [0:44:33]. Now, you have several options up here. We can sample and manipulate data and just run over and it’ll tell you what is, give you a brief explanation. We can explore the data, which I did here, and that means I can do association, clusters, compute \_\_\_\_\_ [0:44:54] statistics, graph explore, link analysis, market basket, multiply, which I use here, path analysis, \_\_\_\_\_ [0:45:06] if you folks know this, stat explore, which I always use. I can do variable clustering and variable selection.

Now, in this case I decided to do a stat explore on this and I ran it earlier, and to run something you just click there and say run at any point. And, I can run the entire model by clicking here. Or, I can just run a piece of the model. But, I ran that and now I want to look at my results and all the results came up for how it worked previously. I have output, separate windows, so my \_\_\_\_\_ [0:45:51] my targets, deviation, standard deviation. I can even view up here. I can look at summaries and I can even look at interval variables like class variables, interval variables. And, this is very simple and, you know, this is not made to be that \_\_\_\_\_ [0:46:21]. But, the point of is, I can look at this, I can explore, I can run this several ways, change the data and do a simple static score.

Now, I did a multi-plot, too, and let’s get the results of that. And, again, it’s same method, all these work in the same method. I can choose types of charts, both, it could be selection of bar charts or scattered plots, so I chose both. I make it vertical, I can include, I can do mean, I could do some or none. Okay? Statistic is free, reduction type is linear, and it gives me information about the status of it. Same thing here, and stat explore, it gives me a lot of ways to do this. I wanted all the observations, I don’t care about validation or test. I can answer yes or no here, of course, we said yes. Number of bins I can make, I decide to make it five, it shows a \_\_\_\_\_ [0:47:25] correlations, I can choose one or the other, I decide to choose both. And, that was to here, same thing down here. So, if we look at our results here, I can even go forward and look at everything. It shows me graphs of everything that came through. And, again, it also shows me output.

Now, SAS also has some HP, high performance and in several here, like a high performance GLM, which I ran over on the side, principle component, that’s, that’s working \_\_\_\_\_ [0:48:14] in memory. Right. That’s an \_\_\_\_\_ [0:48:18] I have an \_\_\_\_\_ [0:48:17] down here and I did an HP explore there. I’m, in essence of time, I’m not going to go through all that. So, I did a data partition where I divided my dataset allocation. I want 60% training, 30% validation, 10% testing. And, from there, once I have everything divided up, I did a \_\_\_\_\_ [0:48:47]. I called this one profit and loss, I could say average error and misclassification by, did a profit and loss, because I want to see how it comes out. Residuals, I can apply everything I need right there. Optimization, network, yada, yada, yada. I did a regression. Okay? And, I’m doing a simple logistic regression backwards. I define all my inputs here, how I want to use it. And, I did a decision tree, and the same thing here. I don’t how many of you folks have used decision trees. They’re very, very interesting. You can do cell pruning or you can prune it yourself. I actually did an interactive one, so it does a lot, does a lot of pruning itself.

Now, from there I hook it all to a control point, because I wanted to do model comparison. And, if I look at my results it’s telling me that how I define this, my regression is my best model out of \_\_\_\_\_ [0:50:03] decision tree. Okay? Train, validate, etc., max predicted \_\_\_\_\_ [0:50:14]. So, I can really look at this and get into, I can save the output and we can run this, because it only takes a moment. And, because I, regression was it, I chose regression as my, as my model example. Now, I’ll show you that this is grid enable, let’s just run it. And, when I do it here, I’m going to run the entire path from here. I asked it to rerun the data, then I’ll go to data partition, it’ll go up \_\_\_\_\_ [0:50:42] regression, decision tree and come down to my control point, which brings it all back together, and then it’ll come down here to model comparison. All right. That’s going to take a minute or two to run.

Do we have, since we only have five minutes, do we have questions?

Heidi: We did get a question wondering if you’re still going to go over \_\_\_\_\_ [0:51:07]?

Mark: Oh, okay. While this is running, you do R through proc [PH] IML. This is very, very simple. So, I say proc IML, I’m making a matrix. I’m making it in SAS, I’m making it here. I can do this under IML and essentially, I’ll say run. And, it will give me, these are results from \_\_\_\_\_ [0:51:36], it’s results from SAS. I can include any R package I want down here. I can write it, I can include it, whatever I need to do. So, let’s say I want to prepare the data in SAS, but I want to run a little procedure in R and I want to use R graphics. I can just put it down here, this is just a very simple on. And, this will, work day later this month, I’m going to go into this with a lot more detail. All right? So, what it does, it does the weaknesses of R, which is memory and I/O and handling large bits of data, SAS handles that and I can pass the data back and forth between R and SAS and run any R procedure I want. I believe we’re using R 3.25 here.

Now, let’s go back to Enterprise Miner, and it came through, it ran data partition because I have it grid enabled, and I do that just up here in my options. Use grid processing when available. It shows me that it’s running on which node, here we’re only to use node two and four. And, it already ran these together and ran this one last, and it comes down to our control point, and I’ve come down to our model comparison. So, these all ran at the same time. I’ve run very, very complex models and I thank God I had the grid, because I had to do it on large datasets and it ran, it can run three, four, five, six, seven at a time. I can bring everything together and it will tell me, did a lot of validation here, which is my best model to use.

So, to finish up in the last few minutes, here are some tips. I want to say I also use SQL pass-through queries. SAS works faster and is more robust than SQL Server. And, I don’t mean to denigrate SQL Server, I’m just saying that when you bring in the SAS world and you have both SQL Server, all the SQL Server code you want and you have SAS code. It gives you the ability to SQL Server Proc SQL or SAS data functions. Work with the appropriate venues as we stated, and that’s EG, Gsub and EM. You can incorporate other technologies like R. We use SAS quite a bit. For people who want to use data or SPSS. We prepare the data in SAS and then transform it in SAS into \_\_\_\_\_ [0:54:11] data or SPS data. As we said, SAS can create data for other folks.

And, more tips, parallel processing can greatly reduce time in both ETL and analysis. Far more storage in the Linux Grid that on Windows per project, and not even in the ballpark. Standardized environment makes efforts simple and efficient, and that’s EG and that’s EM. And, we have pre-prepared security and data via Metadata, which we enabled for you. And, always I suggest having your project enabled with Metadata. Okay?

And, I’ll take any questions if we have time. And, again, if you have any other questions, we do this. I want to thank you all for attending. And, please contact the VINCI SAS administrators, that’s Tony Silik [PH], Kevin Mine [PH] and myself. And, we’re happy to answer any questions or give any training, etc., etc. As always, it’s a pleasure to work and present to you folks. And, I thank you very much for attending. Please feel free to put any questions into the VINCI SAS admins afterwards, or call me directly. Okay?

So, if there aren’t any questions, then I thank you very kindly and wish you a wonderful day.

Heidi: Mark, actually, \_\_\_\_\_ [0:55:29] I’m just going to try to sneak in one question and then we will log everyone off for today.

Mark: \_\_\_\_\_ [0:55:34].

Heidi: So, with the discontinuation or limitations of use of the remote desktop, \_\_\_\_\_ [0:55:43] different or preferred way of obtaining access to SAS Grid?

Mark: The best way to get access to SAS Grid is two ways. One, when you’re setting up your project, there’s, there’s actually a box you can click and that automatically will put you into all the right groups] in Active Directory, because we do it by Active Directory and then Metadata and then Accles [PH] and Linux. But, if you’re already using desktop, essentially, you can call us, the SAS administrators, and we’ll get you set up and we’ll enable your project in Metadata and give you a little bit of training. And, you’re off to the races.

Heidi: Fantastic. Sounds great. With that, we will wrap things up for today. I want to thank the audience for joining us today. We really do appreciate you joining us for today’s cyber seminar. Mark, again, thank you so much for presenting for today. I know people really appreciate and really jump on any information that you are willing to present with us. You are definitely a very popular speaker here in this series. So, we appreciate you continuing to take the time to present for us.

Mark: I’m very flattered. And, enjoy doing it.

Heidi: Wonderful. Good to hear. When I close the meeting out here, the audience, you will be prompted with a feedback form. Please take a few moments to fill that out. We really would love to hear if you have ideas for other things that you would like to hear about. We really do read through all of that feedback that you submit. Thank you, everyone, for joining us for today’s HSR&D cyber seminar, and we look forward to seeing you at a future session. Thank you.

[End of audio]