The Challenges of Teaching Business Analytics: Finding Real Big Data for Business Students

Alexander Y. Yap
ayyap@ncat.edu

Sherrie L. Drye
sldrye@ncat.edu

Department of Business Education
North Carolina A&T State University
Greensboro, North Carolina 27411, USA

Abstract

This research shares the challenges of bringing in real-world big business data into the classroom so students can experience how today's business decisions can improve with the strategic use of data analytics. Finding a true big data set that provides real world business transactions and operational data has been a challenge for academics developing a data analytics course or curriculum, because in the past academics use to rely on 'fictitious small data' to teach students the basics of analytics. The ideal scenario for business academics who wish to bring a more cutting-edge experience to business students is to show them how evolutionary tools in data mining analytics can interpret real world big business data. This research emphasizes the need for students to have more exposure to big data analytics. This paper presents how a real world big data set has been utilized in a business course.

Keywords: Education in Data Analytics, Real World Big Data, Data Warehousing, Data Mining, Business Analytics

1. INTRODUCTION

Many business schools have shown interest in offering courses in Big Data Analytics due to the significant demand of companies for data scientists (Davenport & Patil, 2012; O'Connor, 2012). There is a severe shortage of skilled data scientists graduating into the workforce, with reports stating that by 2018, the supply will exceed the demand by anywhere from 200,000 to over a million jobs (Brown, 2016; Overly, 2013). Employers readily admit that they need "numerate employees" with "quantitative aptitude," "data literacy" skills, and a "data-driven mindset" (Overly, 2013, p. 2; Harris, 2012, p. 2). Most employees will be working with big data in some way. Some will need primarily skills and insights to interpret the results of data analysis. These are called citizen data scientists by Gartner (Marr, How The Citizen Data Scientist Will Democratize Big Data, 2016). Other jobs will entail a deeper knowledge of statistics, programming, data cleansing, data management, visualization, and data analysis techniques (Thibodeau, 2012).

Companies often advertise information technology and related jobs based on their "ideal" qualifications, with so many qualifications listed that no person really has all of the skills advertised. Usually they are able to choose the closest candidate to what the ideal would be.

However, big data jobs are different for a number of reasons. First, job descriptions for big data jobs are difficult to write since it is such a new field (Thibodeau, 2012). What skills do you include when the job description is unclear? Companies need to examine what their needs are and not seek the traditionally "ideal" candidate. For big data jobs, they cannot necessarily be

entrenched in requiring knowledge of certain technology as they could be with typical information technology jobs (Brown, 2016). Second, because universities are not graduating enough students with data science skills, companies are often training them in-house (Tschakert, Kokina, Koziowski, & Vasarhelyi, 2016; Brown, 2016). The shortage of university graduates is critical. The state of Illinois actually set up an alliance, IoT Talent Consortium (Internet of Things Talent Consortium) with Microsoft to increase the number of skilled workers in the state. They set up an online certification program (www.iottalent.org/illinois-datascience) where any resident (and up to 100 prisoners in the state penal system) could take nine courses in "analytics, predictive analytics, various coding languages, and other digital disciplines" (Shueh, 2017, p. 1). Lastly, it may not be a matter of choosing one ideal person, but matching people with varying skills into teams to overcome the skills gap and to address the complexity of doing big data well (Brown, 2016).

The skills gap is real and it is in the face of companies and universities. Although universities are struggling to meet the demand, they are falling short of meeting the needs of the job market. One reason is that they may not have the technology and resources to teach big data. It is a good idea to start early with statistical analyses, Excel, and Access, but bigger tools and data are needed (Tschakert, Kokina, Koziowski, & Vasarhelyi, 2016; Thibodeau, 2012). There are analytic tools such as R, which are free and readily available (Bisson, 2017), but there are only a few offered and used in universities. Even if R is free, the trick is how to teach the analytical and interpretive side of the big data analysis. IBM released the Watson and Other Natural Language Processing platform for widely available use in universities and even has a student version of its Watson Analytics platform (Marr, How IBM is Hoping to Close the Massive Big Data and Analytics Skills Gap, 2016). The natural language processing can skip the statistical middleman so that users do not have to know all the statistical techniques and can go straight to interpretation after machine analysis (Marr, How IBM is Hoping to Close the Massive Big Data and Analytics Skills Gap, 2016). What academia needs is more tools and big data to teach these critical skills.

2. THE SEARCH FOR BIG BUSINESS DATA IN ACADEME

At academic conferences and seminars on "big data", the demo data or the sample data in these academic gatherings ironically end up as "small data" samples. The noble intent is to showcase some academic exercises which can be used in the classroom, but the data is not "big" by any means. Software vendors who partnered with academia so their big data analytics software can be employed in Big Data Business Analytics classes have not really offered "big" business data samples, either. What they offer are good small data samples with probably up to a few hundred entries that would give students an appreciation of organizational data, but the data sample were not true samples of real world big data.

It appears that real world business organizations have not been that willing to publicly share their real big data samples to the academic community due to fear of privacy and fear that their competition will have access to publicly-shared data. At a conference attended by one of the authors, a lecturer, who admitted frustration for not getting access to big business data, suggested that big data academicians in business schools can use the free weather big data that meteorologists, oceanographers, and climate scientists use to predict the weather using predictive analytics. However, as instructors of big data analytics classes, we want to find real big business data that is relevant for business decisions instead of predicting the weather. Business students need to see big business data and not big weather data.

3. OBJECTIVE OF RESEARCH

This objective of this case study paper is threefold: (1) To share our experience why using real world big business data can pose a challenge for business academics teaching big data. (2) To share our quest to offer students a real big data set in a big data course curriculum. Where can we find real big business data that can be used in the classroom to teach big data business analytics? (3) Lastly, we want to showcase the software tools we used to teach big data analytics.

4. THE CASE OF SAM'S CLUB BIG DATA

Sam's Club has donated its big data set to Sam Walton's College of the University of Arkansas. The data is more than a few years old but it has transaction data dating back to the 1990s. While it is not the most current set of data, it is a real Big Data warehouse that contains millions of historic data entry and captured detailed retail transactions and store information of Sam's Club across the United States. This big data can be

used to teach Big Data Analytics specifically for business students.

As a member of the academic community, an instructor can apply for access to the big data set with the University of Arkansas (https://walton.uark.edu/enterprise/). The *Teradata University Network* (TUN) is the organization managing the partnership between different technology providers and access to real word big business data. The big data set is stored in a Teradata warehouse with 20 terabytes of storage and with 576 Gigabytes of RAM memory. Teradata is an independent solutions provider of Data warehouse, Data mining, and Big Data analytics solutions. Teradata provide big data solutions to companies like PayPal, Proctor and Gamble, Wells Fargo, T-Mobile, and other well-known Fortune 500 companies. The biggest data set in TUN is from Sam's Club, but the TUN network also has big data for other retailers, like Dillards, for example. Other software vendors like SAP, SAS, and Microsoft software platforms can access the Data warehouse through universal ODBC connections, given the correct password credentials. Figure 1 shows the connections in a network diagram.

Challenges

There are several challenges to instructors in using this system. These are discussed based on our experience of teaching this system for the past two years.

(1) **Connecting to the TUN Network** - Sam's Club/Walmart has an agreement with the Sam Walton's College of Business in the University of Arkansas that they will donate the big data set for academic use, but it cannot be taken out of the University's system. So, if you are an instructor in another university, you and your students can only access the big business data via Remote Desktop connection to the TUN network (See Figure 2). Occasionally, this poses a problem for laboratory classroom computers that do not have a huge amount of RAM memory. And if the lab classrooms are running on VMWare clients connected to a virtual machine (VM) server, the administrators should also allocate an adequate amount of RAM memory for the lab computers clients functioning as virtual machines. Our experience is that with limited RAM memory in VM clients, remote desktop can be slow and can cause the VM clients to reboot. In fact, one University administrator suggested that the VM clients need to be rebooted if the memory cache is full. Students complain that their lab exercises have been interrupted by these freezing and reboot problems. However, once rebooted, the access and connection to the University of Arkansas TUNS server tends to be more stable. So if the class being taught is after 2:00 pm and the laboratory computers have been used by other classes before that, it is best to reboot the laboratory computers before all students go into their remote desktop sessions.

In addition to the challenges of using remote desktop on laboratory computers that may not cooperate with students, the other challenge falls on the tenacity of students to set up the remote desktop connection. It is normal to see a couple of students who read the instructions quickly and cannot get any remote desktop connection for more than 30 minutes into the laboratory exercises, and they get frustrated. Most of the time, it has to do with spelling the server information wrong. It is common to see students type the server address wrong like "waltoncollege.urak.edu" instead of "waltoncollege.uark.edu" and they think it is the TUN server that is not letting them in. Sometimes it does take up to 7-15 minutes for the TUN server to negotiate and recognize the username and password credentials when it is used for the very first time. When everything does not work out, it is best to give the student an alternative username and password. As instructors, it is wise to ask the TUN administrator for more usernames and passwords than the number of students in the class, so if there is a non-working username/password you have extra ones on hand. Based on our experience, this rarely happens, but the extra credentials come in handy.

(2) **The Lack of Metadata and Data Content descriptions** - For instructors wishing to learn about the big data content, the negative aspect is the lack of metadata descriptions about the data fields. There are a few metadata description samples, but does not cover everything in the Big Data set. The administrators of the TUNs server and data warehouse system have done an excellent job making sure that the data warehouse and all the Teradata software are running in excellent condition. But since they are not employees of Sams' Club, they cannot answer questions about the data content itself or what some of the data field names stand for. Therefore, calling them for assistance is not very helpful at times. Instructors wishing to use this big data need to explore the content and determine what the data content is all about, because the metadata descriptions are mostly lacking from the brief manual and instructions.

In Figure 3, when one opens the *"item_desc"* table and one can see several data fields like *Item_Nbr, Category_Nbr, Sub_Category_Nbr, Primary_Desc, and UPC*, a person without any business background would probably not immediately recognize these data field syntaxes. However, if a user is familiar with a Product table, these fields begin to look familiar. Most business organizations selling physical artifacts will have a data table containing all information about their products, including description, color, weight, price, and vendor information. Once we extracted the data from this table, we confirmed that the "item_desc" table is fundamentally a Product table that contains information about various products that Sams Club sells. Item_Nbr is essentially the unique Product identification number. And since each product belong to a certain product category, like "flat screen television" belongs to the product category "Electronics", Sam's Club assigns a Product Category identification number for each Product category. There is also a Product sub-category, like "Televisions and Video Products", under the larger product category of "Electronics". So the data set also has a Product Sub-Category identification number. The data field "Primary_Desc" actually contains the name of product, hence the wording 'primary description'. There is also another data field called "Secondary_Desc", or secondary description. This field provides additional or more detailed product description, in addition, to the product name. If the product has a color and size, the data fields "color_desc" and "size_desc" will provide information about that. And the UPC data field is the "universal product code (UPC)" that we find in 'bar codes' and which the barcode scanners pick up.

For instructors who wish to use this big data set, it is a plus if the instructor has some familiarity with business data, because the data field descriptions are not completely provided in the sparse 'user's manual' and you may have to spend some time going over the big data content and understanding what the data field names stand for and what contents are inside these data fields.

(3) **Learning SQL** – A common question from both business students and faculty inquiring about teaching this big data system is whether there are some programming skills involved. The answer is that students will need a basic understand of the SQL language and how data sets are created, structured, and connected to other data sets in order to extract data from the data warehouse.

In Figures 3, using the basic SQL "select" and "from" allows one to determine how many Sam's Club members there are and how many unique products Sam's Club carry in this data set. Running the SQL script, we found that the big data set has recorded a total of about 5.66 million Sam's club members and more than 432,000 unique product items, as the reference for all retail transactions in this data set.

Having previously taught 'Introduction to Database' classes, students in those classes are shown sample data sets that had less than a hundred items or records. With a true big data class, students slowly begin to realize that we are going through hundreds of thousands of items to millions of data records. Occasionally, extracting and filtering such a huge data set could take 20 – 30 minutes to execute, before it displays the results. Students often ask "Why is it still not finished? The query has been already running for 20 minutes? "

More complex SQL scripts can answer questions like 'What is the average sales of electronic items for Sam's Club stores located in Wisconsin from January 1st to January 30th ' The more one 'slices and dices' the data set, the more SQL skills one needs to have using the Teradata system. In teaching this course, we need to give students some sample SQL scripts when we try to extract a specific data set in order for them to appreciate (1) how data is extracted to answer specific questions; (2) what type of data set they are extracting; and (3) how that is relevant to making certain business decisions if one were to assume the role of a Sam's Club manager who decides how much inventory to carry, what is the best price mark-up for each item, and what colors and sizes are more demanded by customers, and what coupons to issue for various products. Using big data analytics can help business manager with such decisions.

Most business students do not have a background in SQL, nor are they required to take SQL in AACSB business degrees. So, in a business school environment where SQL is not a required course, there is a need to start introducing students to basic SQL scripting for extracting data from a Teradata warehouse. Figure 6 shows a sample of how to extract data using a SQL query from the big data set to display product information.

(4) **Big Data Analytics Software** - Teradata has two main software packages that work with the Data warehouse. The first is the Teradata SQL Assistant that allows users to write

SQL scripts in order to extract specific information from Sam's Club big data (Figure 6). The other software is the Teradata Analytics Miner, which can perform several analytic techniques that run on intelligent algorithms that search for unique data patterns. The challenge of teaching students analytics is to put these statistical tools in the perspective of making better management decisions with these data and tools.

Teaching Analytics Techniques
Some examples of analytics techniques we use for this course are as follows:

Cluster Analysis
This analytics tool allows the users to search for spending patterns of Sam's Club members. For example, one could search for "how much does each customer spend every time they visit a Sam's Club store?" The dataset we look at is called "Total Visit Amount". In Figure 4, we then ask the analytics software to look for 5 distinct clusters and we also ask the clustering algorithm to go through the entire big data set in 25 iteration cycles. This means that the algorithm combs through the big data set 25 times to find 5 cluster data patterns. The more iteration the algorithm does, the more it will accurately detect and identify cluster patterns. Of course, we can ask the software to look for 4 or 6 clusters, instead of 5 clusters, and we could ask it to perform 50 iterations. It is mostly a trial and error process to determine how many clusters will best represent the data analysis we are looking for.

The result (Figure 5) shows that the clustering algorithm has identified five clusters of consumer spending. The first cluster is $95.32 and 77% of Sam's Club customers spend that average amount of money every time they visit a Sam's Club store. The second cluster is $255.97 and 20% of customers spend that average amount of money every store visit. The third cluster shows that about 3% of Sam's Club customers spend $552.49 every time they visit the store. And there are two other clusters that are probably a fraction of 1%, namely $2,811.18 and $94,356.48. These are large amounts because Sam's Club has corporate members. And corporations spend more than individual consumers.

It is important for students to know that a large percentage of Sam's Club members (77%) spend no more than $100 every time they visit a store, so managers need to make a decision what kind of products and brand they carry to fit that type of shopper. And there are 20% of members that spend about $256 every time they visit a store, and they may have preferences for high end product brands and their spending level needs to be met by carrying certain high-end quality products. As a side note, these dollar numbers are from a decade ago at the time of this writing, so in today's dollar value – these dollar amounts could be 30%-50% higher.

Association Analytics
This analytics tool can search through big data sets generated by each Sam's Club store and see if there is any unique consumption pattern linking two products. For example, a large majority of people who buy hamburgers in grocery stores may also buy hotdogs. Consumers who buy paper towels may also buy toilet paper. This is called associative analytics, which many retailers use to determine which related products consumers may be interested in. Many retailers issue coupons to customers who may be interested in products that are highly associated with other products in terms of consumption habits. Most Dads who pick up a six-pack of beer could also pick up diapers for their kids, and the next time they get beer they could get a $2 discount coupon for diapers.

In Figure 6, the Associative Analytics software of Teradata allows users to choose the Product Name contained in the data field "Primary_Desc" and the number of times people buy products every time they visit a Sam's Club store (Visit_Nbr). The analytics algorithm tries to find one-on-one patterns between one product and another product. Would people buy Product A and then also buy Product B? How strong is the association between Product A and B? This is what the association tool wants to find out.

In Figure 7, we show an example of how at one particular Sam's Club store, people who buy 'Bounty Paper Towels' also buy 'Charmin 24 Double Roll Toilet papers' and this comes out with a high association showed at a ZScore of almost 68 (red color chart). We can also see in this chart that people who buy Extra large 18 pieces of eggs also buy 2% Fat Milk and 1% Fat Milk (green and blue color).

Decision Tree Analytics
A third popular analytics tool is the decision tree. Using this tool, we want to determine why consumers continue to become active or inactive members of Sam's Club. The data field "Member_Status_CD" shows whether a member has an 'active', 'deactivated' or 'inactive', or 'expired' status. The dataset has a code (A, D, and E). Deactivated members are members who have been inactive for a long time. Their

information is never deleted. The expired membership status is for members whose membership was just recently expired (maybe a few days or weeks) and they could renew and become active members again. In this exercise, students try to find out what causes members to stay 'active' with Sam's Club or stay 'inactive' and not renew their membership. Is it because they like the products being sold (Item_nbr), is it because they like to buy a certain quantity of items every time they visit (Item_Quantity), membership type (retail or corporate accounts), the different product categories available (category_nbr), the product sub-categories (Sub_category_desc) available, the location (store zip code), what store they registered as members (Register_Nbr), how much they spend or the amount scanned at the counter (Total_Scan_Amount) and what type of payment they use (Tender_type) which could be cash or credit card.

The results (Figure 8) show that the product categories (Category_Nbr), the products items being sold at the store (Item_Nbr), member type - retail and corporate memberships (member_type), the location of store (store_zip), the membership number which identifies location of member registration (Membership_Nbr), the payment method (Tender type), the amount spent by member (Total Scan Amount), are significant factors that determines a member's decision to be 'active' members or 'inactive members'. The analytics software also has a graphic representation of the decision tree, depicting how one variable affects the decision of people to be an active or inactive member.

5. STUDENT FEEDBACK

Student feedback about this analytics course has been very positive. For example, students rated "the practical application of the course" at 4.81 out of a 5.0 perfect score. In comparison, the average rating for this criterion for the entire University is at 4.22 and, for the Business College, the average is at 4.29. The comments were very positive. One student, for example, said, "I enjoyed the opportunity to work with the different data technologies used in the class. I was exposed to new programs that I feel will be very helpful to my career."

6. SUMMARY

The paper shares our experience in finding and teaching big data analytics in the hope that other business school instructors who are interested in the fast evolving field of big data analytics are able to get some insights into our experience. While we have met some challenges in teaching real world big data, the advantages for student getting exposure to real world big data sets, data warehouse systems, and data mining analytics prepares their mindset for the world of business where big data is becoming an invaluable asset for strategic decision making, planning, and forecasting.

While the availability of real world big business data for academic consumption is still limited, the availability of Sam's Club big data set is a good start for the academe to get a feel of what big business data really looks and feels like. Hopefully, more business organizations in the future can share their old big business data to the academic community and allow students to discover how they make more accurate business decision by performing data mining analytics on big data sets, and how they align their business goals in response to discoveries made by big data analytics.

7. REFERENCES

Bisson, S. (2017, January 11). *Microsoft's R Tools Bring Data Science to the Masses*. Retrieved May 26, 2017, from InfoWorld: http://www.infoworld.com/article/3156544/big-data/microsofts-r-tools-bring-data-science-to-the-masses.html

Brown, M. S. (2016, June 27). What Analytics Talent Shortage? How to Get and Keep the Talent You Need. *Forbes*. Retrieved May 26, 2017, from https://www.forbes.com/sites/metabrown/2016/06/27/what-analytics-talent-shortage-how-to-get-and-keep-the-talent-you-need/#5df4ab3018da

Davenport, T., & Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. Retrieved May 27, 2017

Harris, J. (2012, September 13). Data is useless without the Skills to Analyze It. *Harvard Business Review*. Retrieved May 26, 2017, from https://hbr.org/2012/09/data-is-useless-without-the-skills

Marr, B. (2016, February 29). How IBM is Hoping to Close the Massive Big Data and Analytics Skills Gap. *Forbes*. Retrieved May 26, 2017, from https://www.forbes.com/sites/bernardmarr/2016/02/29/how-ibm-is-hoping-to-close-

the-massive-big-data-and-analytics-skills-gap/#13f2ea14df87

Marr, B. (2016, April 1). *How The Citizen Data Scientist Will Democratize Big Data*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2016/04/01/how-the-citizen-data-scientist-will-democratize-big-data/#6bb5a9d365b8

O'Connor, F. (2012, October 10). *Colleges Incorporate Data Science into Curriculums*. Retrieved May 26, 2017, from Computer World: http://www.computerworld.com/article/2492245/it-careers/colleges-incorporate-data-science-into-curriculums.html

Overly, S. (2013, September 15). As Demand for Big Data Analysts Grows, Schools Rush to Graduate Students with Necessary skills. *The Washington Post*. Retrieved May 26, 2017, from https://www.washingtonpost.com/business/capitalbusiness/as-demand-for-big-data-analysts-grows-schools-rush-to-graduate-students-with-necessary-skills/2013/09/13/afbafb3e-1a66-11e3-82ef-a059e54c49d0_story.html?utm_term=.62ad6830497d

Shueh, J. (2017, April 20). *Illinois Fights Technical Skills Gap with Courses in Data Science, Analytics, Coding*. Retrieved May 26, 2017, from Daily Scoop: http://statescoop.com/illinois-fights-technical-skills-gap-with-courses-in-data-science-analytics-coding

Thibodeau, P. (2012, October 24). *Q&A: What's Needed to Get a Big Data Job?* Retrieved May 26, 2017, from ComputerWorld: http://www.computerworld.com/article/2492880/big-data/q-a--what-s-needed-to-get-a-big-data-job-.html

Tschakert, N., Kokina, J., Koziowski, S., & Vasarhelyi, M. (2016). The Next Frontier in Data Analytics. *Journal of Accountancy*. Retrieved May 26, 2017, from http://www.journalofaccountancy.com/issues/2016/aug/data-analytics-skills.html

Editor's Note:

This paper was selected for inclusion in the journal as an EDSIGCON 2017 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2017.

2. APPENDIX 1 (FIGURES)

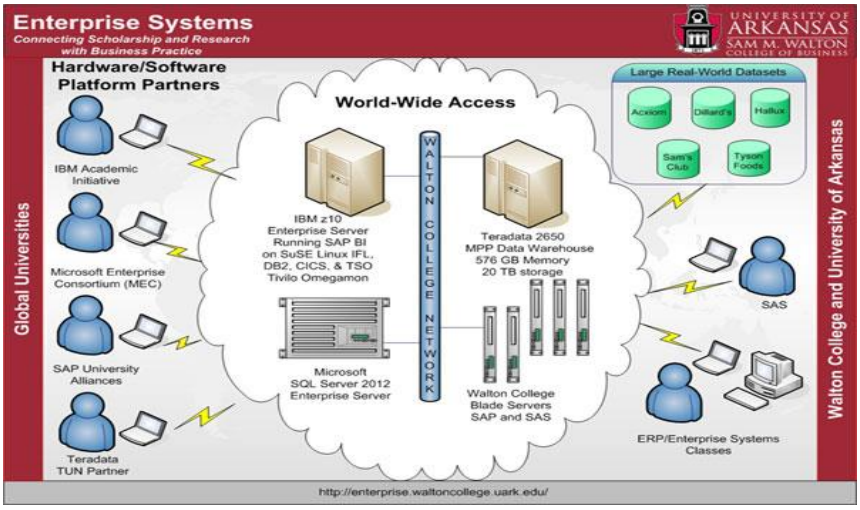


Figure 1. Connecting Universities to real world large datasets

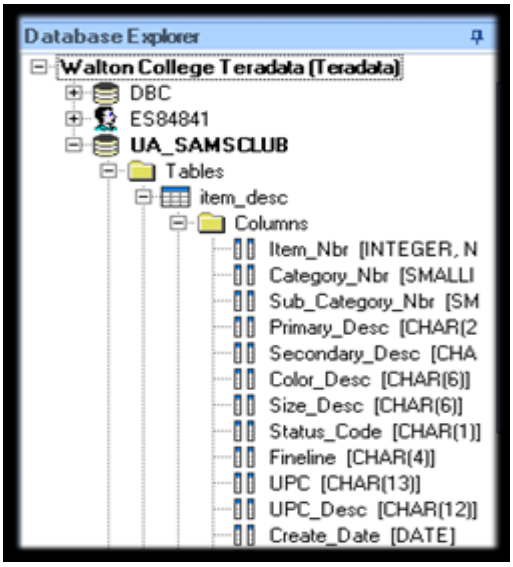


Figure 2. Deciphering the data field descriptions

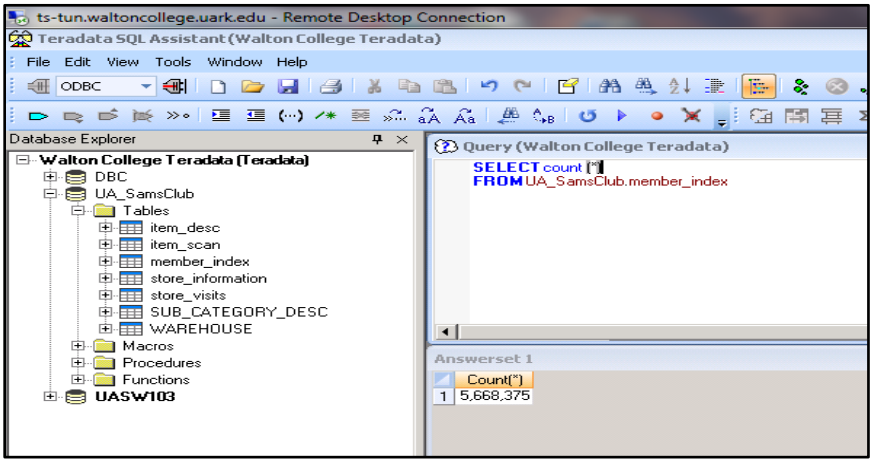


Figure 3. Getting a count of how many unique Sam's Club members there are (about 5.66 million)

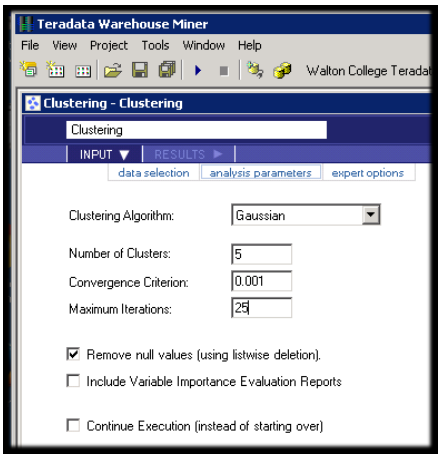
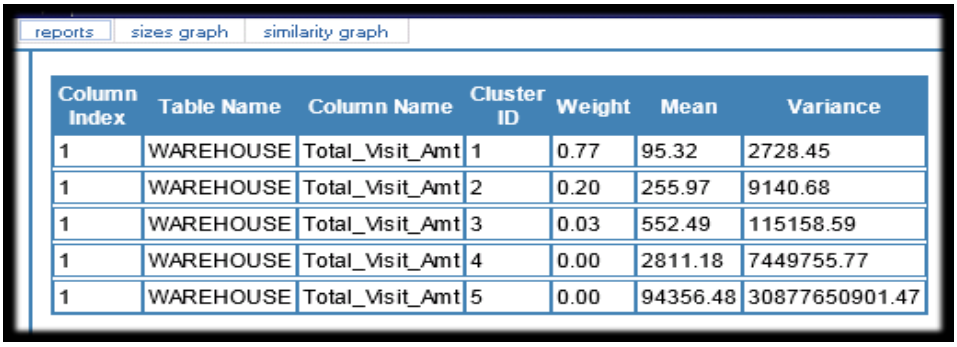


Figure 4. Defining the number of clusters and iteration



Column Index	Table Name	Column Name	Cluster ID	Weight	Mean	Variance
1	WAREHOUSE	Total_Visit_Amt	1	0.77	95.32	2728.45
1	WAREHOUSE	Total_Visit_Amt	2	0.20	255.97	9140.68
1	WAREHOUSE	Total_Visit_Amt	3	0.03	552.49	115158.59
1	WAREHOUSE	Total_Visit_Amt	4	0.00	2811.18	7449755.77
1	WAREHOUSE	Total_Visit_Amt	5	0.00	94356.48	30877650901.47

Figure 5. Table showing the cluster spending (mean) and the weight shows percentages

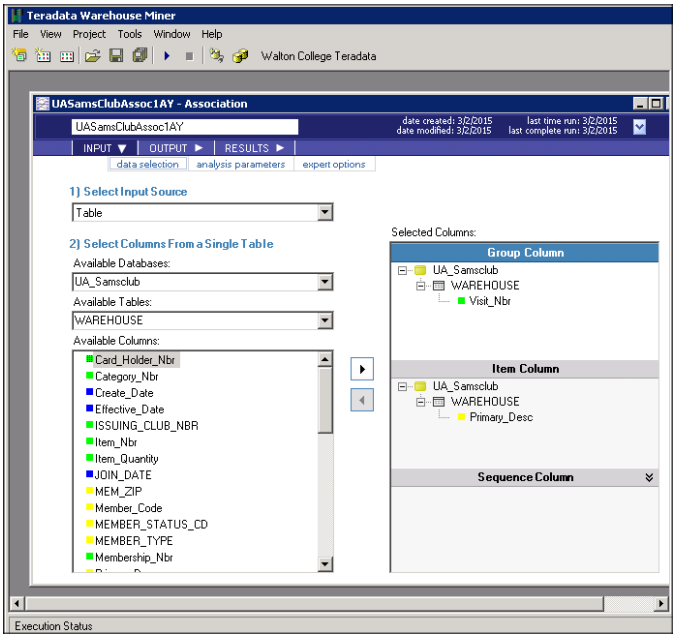


Figure 6. Choosing the Product (Primary_Desc) and the frequency people buy various products on every visit (Visit_Nbr)

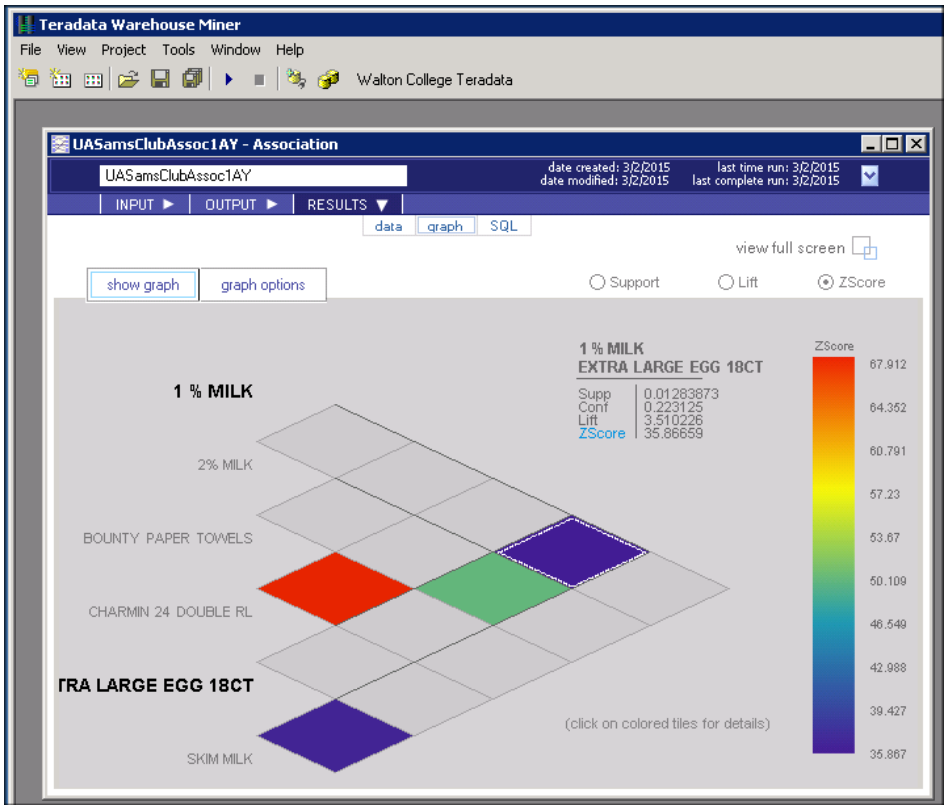


Figure 7. Associative Analytics show close association between paper towels and toilet papers and eggs and milk for people who shop at this Sam's Club store.

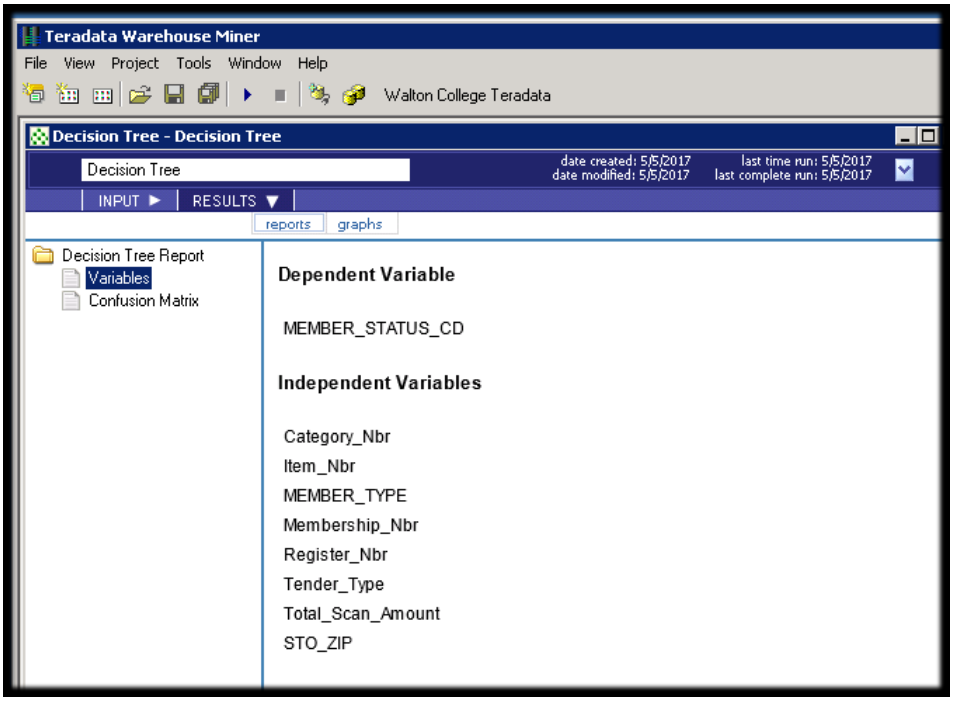


Figure 8. Results shows Independent Variables that affects the Dependent Variable