Keeping Up with Big Data - Designing an Introductory Data Analytics Class

Dr. Sam Hijazi
Texas Lutheran University
1000 West Court Street
Seguin, Texas 78130
shijazi@tlu.edu

Abstract

Universities need to keep up with the demand of the business world when it comes to Big Data. The exponential increase in data has put additional demands on academia to meet the big gap in education. Business demand for Big Data has surpassed 1.9 million positions in 2015. Big Data, Business Intelligence, Data Analytics, and Data Mining are the four main branches of data analysis. These areas intertwine, overlap, and clearly depend on each other. This study endeavors to examine the concepts, tools, and techniques of these topics through an introductory class in Big Data. After serious efforts and examination, this study found Data Analytics to be the most suitable topic. One main reason for making this choice included the need to teach the students how to ask questions when they manipulate large amounts of data. The other reason is the availability of PowerPivot and DAX as an Add-In to MS Excel. Most students are familiar with MS Excel, since it is readily available to them. The tools, the techniques, the built-in functions, PivotTables, and DAX, as a formula language, will allow students to experiment with a million rows of data resulting in a rich and rewarding learning environment. The class will also cover the other areas of Big Data and their relationships. Students should become Big Data literate by the time they finish the class successfully.

Descriptors: Big Data, Data Analytics, MS Excel, PowerPivot, PivotTables, DAX, Class Design, Introductory, Learning, Business.

Introduction and Problem Statement

The exponential increase in data size is not a hidden fact in our daily activities. We can hardly measure the amount of data we capture, process and store every day. “The idea of data creating business value is not new; however, the effective use of data is becoming the basis of competition” (Insights, 2014). The author added that businesses have always wanted to derive some insights in order to make more informed, real time, factual, and smarter decisions. Big Data is hitting organizations from all directions, internally and externally. Data is not limited to machine data but also found as unstructured, online and on mobile as well. The Insights article added that statistical data (historical) and predictive (forward thinking) are needed to make helpful decisions. One of the problems that has been noticed, according to the Insight article, is that we are able to capture and store massive amount of data, but we still lack the technical capacity to aggregate and analyze unpredictable volumes of data.

Marr (2016) has summarized some intriguing facts with a full invitation for us to examine them closely. His intention is clearly to make us realize that Big Data is not only a problem but also a great oppor-

tunity, if we decide to take it seriously as educators and business people. Here are some facts that might help some of the hesitant business educators to think over and see why a class related to Big Data with data analytics is not only an option but a must have:

- Everything we do has a digital trace for us to analyze and use.
- Every two days we create as much data as we did from the beginning of our civilization until 2003.
- "Over 90% of all the data in the world was created in the past 2 years."
- "The total amount of data being captured and stored by industry doubles every 1.2 years."
- It will take 15 years to watch all the uploaded videos to YouTube every day.
- In the Big Data field, there were 1.9 million IT jobs created in the US by the year 2015. Knowing that every job needs 3 new jobs created outside of IT to support the one in Big Data. This will result in 6 million new jobs caused by Big Data.
- By the year 2020, the value of the Hadoop (the open source for big data technologies market) will soar from $2 billion in 2013 to $50 billion.

The above facts and findings do not require any additional proof in order to decide our need to offer additional classes in Big Data in the business curriculum. If we don't react accordingly, students will have to invest in additional trainings, seminars, and online classes to catch up with other peers in a very competitive market. The problem requires an immediate solution. It is clear that we need to incorporate the concepts and applications of Big Data in our academic curricula.

Why Teach Big Data Related Class?

These days when accountants were responsible for crunching historical data are behind us (Meyer, 2016). The world expects CPAs and accounting students to manipulate a much larger scale and volume of data, that is, Big Data. Meyer also quoted Wenger, an assistant professor at the University of Mississippi, who stated that data exploded in all sort of businesses. Future business students need to possess the ability to sift through massive amount of data by using different techniques and tools in order to evaluate data effectively. Students need "basic exposure to Big Data and data analytics" by adding this class to the accounting information systems degree curriculum. It is general belief that all students in business, especially information systems ones, need a deep understanding of Big Data concepts, techniques, and tools.

Deciding on the Best Approach to Offer an Introductory Data Analytics class

It is clear there are some differences in the areas we hear about everyday concerning Big Data. However, if we probe closely, a person could notice, regardless of the name of the area that they all intertwine, overlap, and clearly depend on each other. "The difference between Big Data & Business Intelligence (BI) is synonymous to fishing in the sea versus fishing in the lake. Your target is the same but the tools are decided by the scale" (Mohanasundaram, 2015). In addition to Big Data and BI, there are two fields related to Big Data. These are data analytics and data mining. Obviously, it is very hard to come up with a quick conclusion concerning which approach is best for your students. Having said that, we need to act to find a solution. There are factors we have to consider such as cost, size, and support. Junk (2015) discussed the lack of boundaries between all the areas of Big Data. To help us navigate the complexity of business data concepts, he discussed the most common terms in this field and their relationship:

1. Business Intelligence. Junk stated that BI is the broadest category that encompasses the three other areas according to how the business world uses them these days. BI is based on decision-making and concentrates on the generation, aggregation, analysis, and visualization of data. It is not only about the data and the tools but also about the policy and procedures that support all the activities that convert the data into actionable results.

2. Analytics. Junk stated if BI is about making decisions, then Analytics is about asking questions. Here you can break down the data, create an assessment over time, and compare one trend to another, just to name a few activities. You can compare sales from this month to other periods. Data Analytics is about opening a wide door to be inquisitive .Business today needs to use both historical and predictive data.

3. Big Data. Junk stated this typically refers to the incredible volumes of data from internal and external sources. In addition to volume, data usually is completely raw and in many cases unstructured. Usually businesses use Key Performance Indicators (KPIs) as the main key to turn their questions into answers. Junk added that "Big Data is the library you visit when the information to answer your questions isn't readily at hand. And like a real library it allows you to look for answers to questions you didn't even know you had."

4. Data Mining. Finding an answer to a question you never thought about is what data mining is (Junk). Data mining allows users to sift through lots of data to find unrecognized trends or patterns among the noise. Further, data mining works closely with Data Analytics. The difference is analytics is about measuring data while data mining is about sifting through data.

No matter which class would be taught, it should cover the four areas above conceptually. However, it would be impossible to cover the four areas in depth. For the sake of this paper, the decision was made, with multiple reasons in mind, to teach a Data Analytic class. Reasons to teach the class include affordability, portability, and more importantly the transferability of the gained skilled and knowledge by students to the business world. Regardless, the main reason is to prepare our students for the real world in order to enhance their chance to compete effectively in the job market.

Given that a limited budget is an issue in most academic institutions, there was a need to find the most affordable tools without jeopardizing the learning outcomes. In the process of deciding on which techniques and tools are the best options for this class, there were a few to consider. This study looked at these areas:

1. SAS sells proprietary software for data analysis and management, BI. It is robust data analytics and comes with machine learning, statistics, Econ, forecasting and others. SAS can work with Hadoop and R (Hall, n. d.), discussed below. The problem with SAS clearly is the cost associated with licensing the software for the academic area. It is not cost-effective at all, where the only price I was able to find at their website direct was a whooping tag price of $9200.00 for an individual license. In addition to the prohibitive price, educators need to worry about the learning curve students must go through to be effective in using the software, in addition to the learning of the tool for data analysis and predication. SAS is

very powerful and sophisticated software; however, it is more suitable for high-end data analysis for big corporations.

2. Hadoop. Hadoop is an open source system for processing massive amounts of data (Hall). In order to effectively use Hadoop, the students need to learn a programming language by the name of mapReduce. It is used for large computations and multimedia types of data. Hadoop would be the ideal choice if time is not a constraint. To go through the multiple technologies it encompasses, it requires a time consuming effort. To teach Hadoop effectively probably would require two consecutive classes. The recommendation is to look for a two additional elective classes for Big Data using Hadoop. If the reader wishes to know more about Hadoop, visit this website discussing the top 25 points about Hadoop: http://www.bigdataeducation.in/top-25-things-about-hadoop/.

3. R. R is an open source language used for machine learning and mainly used for statistical analysis (Hall). R is a computer language. A full class must be dedicated to it in order to benefit from its strong syntax. R is not the best choice for an introductory class in Data Analytics. The author of this paper has personally experimented with R and found that it requires a steep learning curve.

4. PowerPivot. PowerPivot is a Data Analytics/BI powerful extension to MS Excel. After exprimenting with PowerPivot, PowerView (the graphical side of PowerPivot), and Pivot Tables for over three weeks with different file sizes, it was decided this is the right tool to teach in an introductory Data Analytics class. Since the decision was made to use PowerPivot as the main tool, this paper will explore the reasons for making this decision in a separate heading below.

Why PowerPivot?

PowerPivot is added to MS Excel as an add-in feature. We Know that:

1. MS Excel is easily accessible and has been around for many years. The learning curve is minimal for most students. Most universities provide MS Office for free or minimal fee.

2. PowerPivot lets management use their reporting. Most users are familiar with PivotTables already. The combination of these powerful applications will result in a very effective outcome.

3. No special IT resources are needed, including servers or unique software.

4. PowerPivot works with Pivot Tables like magic. “As the name implies, PowerPivot is a PivotTable on steroids. With PowerPivot, you can pull into Excel large amounts of data from multiple database tables, databases or other sources of data, and sort and filter them almost instantly” (Jackson, 2010).

5. There are no limits on the numbers of the rows and column especially if PowerPivot works with an SQL server. Of course there is a limit to how large the number of records is before Excel starts to slow down.

6. PowerPivot allows the integration of multiple entities (tables) in a similar fashion to a traditional relational database. This is a great chance to explore with the students the concept of normalization and how to avoid data anomalies. Relational databases are not going to disappear tomorrow and it is more likely our students will have to handle some database normalization in the real world. PowerPivot requires students to understand how to link different tables based on their relations. Up to this point and, just like

MS Access, PowerPivot allows only for one-to-one and one-to-many relationships between two or more entities. This will encourage students to think ahead before they start to build their data model to avoid the pitfalls when they use Excel as a plain spreadsheet.

7. PowerPivot allows the use of Data Analysis Expression (DAX). DAX's functions use similar functions to EXCEL. In addition to these functions, DAX is not a programming language such as C or Java; rather it is a formula language. The most important aspects of DAX are its ability to work with relational data and perform dynamic aggregation. While Excel uses cell references and cell addresses such as G2=F2+C2, DAX deals with columns and tables. Additionally, Excel allows one to change the cells values, while DAX only allows refreshing the data. This means the students will learn how to be aware of the needed changes before modifying the data.

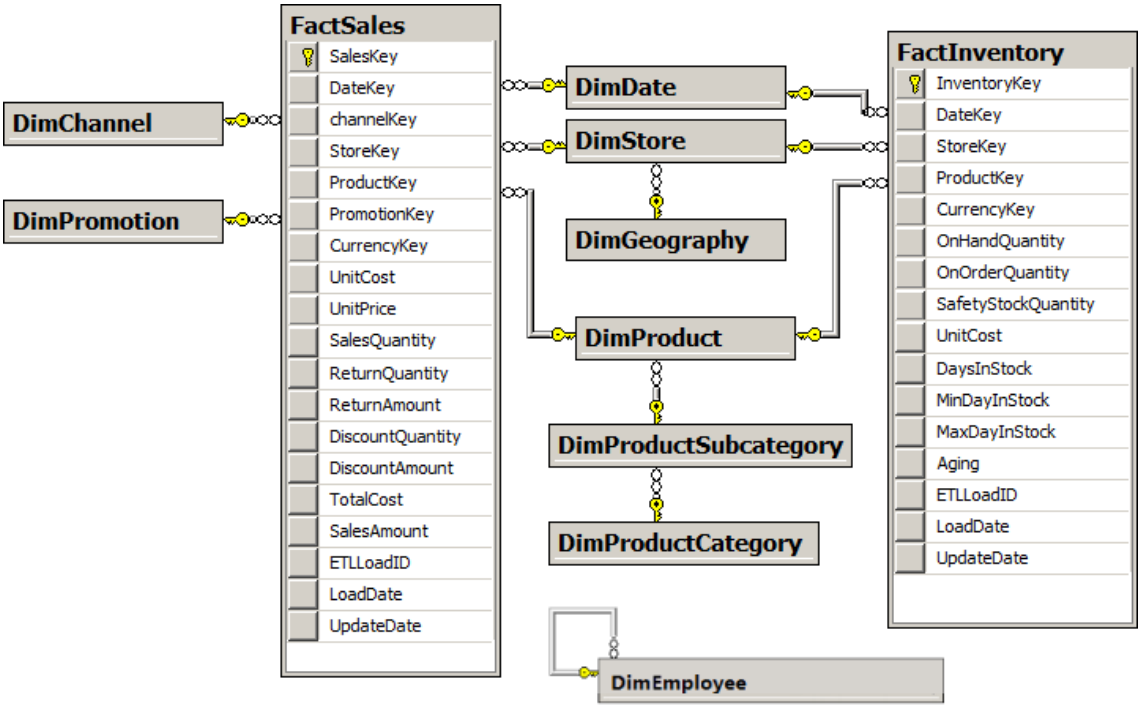
Experimenting with PowerPivot and DAX

In order to come to this conclusion, there was a need to seriously experiment with PowerPivot and DAX to see whether they are suitable for an introductory class on Data Analytics/BI. As usual, there were many useful sites and tutorials that offered excellent help. In addition to watching numerous videos and reading multiple whitepapers, I decided to test Contoso DAX formula Sample. It is a large file with 159.8 MB and the main spreadsheet contains over 3 million rows of data. Contoso file comes with comprehensive whitepaper, 69 pages of instructions and very useful external links, to walk a new learner through the different aspects of PowerPivot and DAX.

The first reading was ambiguous and hard to relate to, except for a few discussions of similar functions in Excel. The moment I decided to sit by my computer to read and follow the instructions in the manual, a whole new door of understanding was open. Learning about PowerPivot and DAX was interesting and rewarding.

Dickerman, H., & Myers, P. (2011) produced a friendly and easy to follow training paper. The authors use samples to illustrate the use of PowerPivot by importing a relational database from the "Contoso" SQL server, a relational database, where any reader can down download it from this URL: https://www.microsoft.com/en-us/download/details.aspx?id=28572 and it comes with a graphical representation of the data model as seen in the below graph. There two main transactional tables, FactSales (3.4 million sales) and FactInventory (8 million inventory) which are related to eight other tables. There is one standalone table by the name DimEmployee as seen in the graph 1.

Graph 1: All the tables with relationship in Contoso database, page 6



Having the ability to work with multiple tables within a PivotTable in Excel is a new experience. To manipulate much larger numbers than the typical 1,048,576 rows as the maximum number in Excel spreadsheet makes a better choice.

Dickerman & Myers stated that the goal from the tabular models is provided to ease the use of data analysis. Students will not be intimidated by this layout. Many students and business professionals will benefit from their previous experience with Excel. The difference between DAX and a typical Excel spreadsheet is that DAX works on columns and not ranges of cells. Columns have captions and are used as variables in calculation. For example, to create a new column by the name Margin based on subtracting totalCost from SalesAmount, the syntax would be = [SalesAmount] – [TotalCost]. The columns SalesAmount and TotalCost are already part of the table. It took less than 7 seconds to do the calculation for 3.4 million rows. Once the table moves into PowerPivot from Excel, it will be compressed and will reside in RAM, the main memory. This will result in slow calculations initially but in time, a user will feel the difference. Most of the calculations that were made took less than 7 seconds.

The intention of this section is not to offer an exhaustive discussion of PowerPivot and DAX, rather to give the reader a taste of the working environment. If you are interested in quick familiarity, I suggest the following link: https://support.office.com/en-us/article/QuickStart-Learn-DAX-Basics-in-30-Minutes-51744643-c2a5-436a-bdf6-c895762bec1a. Students will have no problems learning the many features that come with DAX in a regular, three credit class. Also, it should be noted that the list of functions below is a quick list and by no means will cover these functions in depth:

1. Simple DAX functions. These 80 functions resemble Excel functions such as ISBLANK, Average, AND, OR, etc. DAX uses FORMAT function instead of TEXT function. Also, DAX uses aggregate

functions such SUMX, COUNTX, and AVERAGEX since, as stated earlier, DAX works on columns and tables.

2. Some of the most powerful approaches to DAX are found in Row Context and Filter Context. Both require some serious applications in the classroom setting to understand them.
3. A Data Analytics must covers these functions in DAX:
    a. CALCULATE
    b. VALUES
    c. FILTER
    d. ALLEXCEPT
    e. RANKX
    f. RANK.EO
    g. TOPN
    h. LOOKVALUE
    i. Time intelligence functions
    j. Parent-Child Functions
    k. DAX Query

There is much more to PowerPivot and DAX than what was mentioned above. The whole purpose was to make sure that the materials in this section are substantial enough to cover an introductory class in Data Analytics. As the author of this paper, I completely believe this would be an informative class to break the barriers of my students' understanding to this timely topic. It will include the conceptual, logical, and the practical aspects of a new technology that is not accessible to all.

The Importance of the Study

This study attempts to show the impact of Big Data as an unavoidable phenomenon and the need to prepare our students to deal with the high demands of this field. This study admits the lack of clarity between four areas including Big Data, Business Intelligence, Data Analytics, and Data Mining. The finding of the study suggested the adaption of Data Analytics as an introductory topic to introduce business students to these areas. Selecting PowerPivot and DAX will provide students with the needed accessibility to practice and learning, since Excel is ubiquitous. PowerPivot has the strength to handle millions of rows with powerful filtering and manipulation functions. The new students to this area should be able to break the barriers in this field and build the needed self-confidence to pursue additional training, such as Hadoop technologies.

Conclusion

This study found that Big Data is a lasting and increasing phenomenon. As educators, we need to respond to the changes in the business world. After reviewing keys areas in this field, this research found that the best approach is to offer an introductory class in Data Analytics using PowerPivot found MS Excel. This Application software is accessible by all students and many are already familiar with it. The time saved initially can be used to learn additional functions found in DAX, a power formula language.

Adapting PowerPivot will invite the students to focus on their problem solving-skills. The main reason for a Data Analytic class is to ask the right questions and discuss the possibilities to answer them. As we know these days, it is easy to find the answer for almost everything, but still the hardest task is to come up with the right question.

The study shows an alarming rate of increase in data. However, this gives us the opportunity to respond positively to this increase and take advantage of the hidden treasures in the data by using PowerPivot to filter, calculate, and manage columns and tables very easily.

References

Dickerman, H. & Myers, P. (2011). Data Analysis Expressions (DAX) In the Tabular BI Semantic Model. Retrieved online from https://www.microsoft.com/en-us/download/details.aspx?id=28572

Hall, P. (n. d.). What are Hadoop, SAS, and R, and what are the relationships between them? Is SAS a development of Hadoop? Which is better? Retrieved online from https://www.quora.com/What-are-Hadoop-SAS-and-R-and-what-are-the-relationships-between-them-Is-SAS-a-development-of-Hadoop-Which-is-better

Insights on governance, risk and compliance (2014). Big data: Changing the way businesses compete and operatehttp://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf

Jackson, J. (2010). Excel PowerPivot Disrupts Business Intelligence. Retrieved online from http://www.pcworld.com/article/205260/how_microsoft_powerpivot_will_disrupt_bi.html

Junk, D. (2015). Business Intelligence vs Analytics vs Big Data vs Data Mining. Retrieved from http://blog.apterainc.com/business-intelligence/business-intelligence-vs-analytics-vs-big-data-vs-data-mining

Marr, B. (2016). Big Data - 25 Amazing Facts Everyone. Retrieved online from http://www.slideshare.net/BernardMarr/big-data-25-facts

Meyer, C. (2016). 8 tips for teaching Big Data. Retrieved online from http://www.aicpa.org/InterestAreas/AccountingEducation/NewsAndPublications/Pages/how-to-teach-big-data.aspx

Mohanasundaram, M. (2015). What is the difference between big data & business intelligence? Retrieved online from https://www.quora.com/What-is-the-difference-between-big-data-business-intelligence