Chapter 1

Q1: What is big data? Discuss the factors that led to the era of big data. Compare and contrast the paradigm shift from traditional operational databases to the big data analytical data warehousing.

Big data is defined as a combination of structured, in-house operational databases with external databases, with automatically captured and often non-structured data from social networks, web server logs, banking transactions, content of web pages, financial market data, and so on. Big data is usually characterized by three Vs: volume, velocity and variety. All this data, coming from a wide variety of sources is combined into non-normalized data warehouse schema.

Business organizations have been using standard databases for more than three decades. However, big data demands new techniques and many of them are still in the developmental stages. Acquiring the new tools requires a radical change in underlying beliefs or theory: they require a new way of thinking. It requires, for example, that more people think probabilistically rather than anecdotally. It also requires that managers learn to focus on the signals and do not get lost in the noise.

Q2: What is business analytics? What is the difference between analytics and statistics? Discuss the major fields of study which compound business analytics.

From a practitioner's perspective, business analytics is defined as set of tools and techniques that are used to retrieve, process, transform, and analyze data in order to generate insights for better decisions in the business world. Wayne Winston, a prominent scholar and

consultant in management science and prescriptive analytics, defines analytics as simply “using data for better decision making”.

The major fields of study which compound business analytics are database and data warehouses, descriptive analytics, predictive analytics, and prescriptive analytics. While data structures are used to effectively store and efficiently retrieve information, descriptive analytics can be used to report the past. While predictive analytics uses past data to create models that predict the future, prescriptive analytics utilizes optimization, heuristics, or simulation models that can specify optimal solutions and prescribe the best courses of action. Descriptive analytics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. Some of the common tools used in descriptive statistics include: sampling, mean, mode, median, standard deviation, range, variance, stem and leaf diagram, histogram, interquartile range, quartiles, and frequency distributions. The results of descriptive statistics are often displayed via graphics/charts, tables, and summary statistics such as single numbers.

As descriptive statistics are considered a straightforward presentation of facts, predictive analytics uses statistical modeling to draw conclusions and predict future behavior based on the assumption that what has happened in the past will continue to happen in the future.

Q3: Compare and contrast descriptive, predictive, and prescriptive analytics in terms of tools and techniques used, data input and output, and their use in the decision-making process. Which type of analytics is more important for an organization?

Business analytics integrate tools and techniques from four major fields: information management, descriptive analytics, predictive analytics, and prescriptive analytics. Information management deals with storing, extracting, transforming, and loading data and information from operational databases into data warehouses. Once the information is made available in data warehouses and data marts, business analysts can use a series of descriptive analytics tools to understand what has happened in the organization regarding its key performance indicators.

Further, predictive analytics tools can be used to forecast and estimate future behavior based on past performance. Finally, optimization and other management science models are used as prescriptive analytics to identify the best courses of actions and optimal decisions. The nature of management science has changed to accommodate the need to process large amounts of data sets.

Q4: The traditional management science techniques have been revitalized in the era of big data and have become the basis for many prescriptive analytics models. Explain the change in the nature of management science to accommodate the need to process large amounts of data.

Optimization and other management science models are used as prescriptive analytics to identify the best courses of actions and optimal decisions. The nature of management science has changed to accommodate the need to process large amounts of data sets. An obvious feature of today's management science is its heavy reliance on spreadsheet modeling and other data analysis software programs.

Q5: The majority of companies today are using data analytics to gain a competitive advantage. Describe various ways in which business analytics can be used to lower costs, improve customer

experience, and increase productivity. Make sure to support your argument with examples reported in newspapers, magazines, and online sources.

There is no doubt that organizations are becoming more competitive by using business analytics and practitioners, from both large and small companies, are eager to learn more about data analytics and how to implement it in their everyday decision making. The examples of Target, LinkedIn, First Tennessee, and many other companies which have implemented data analytics, can be used to derive practical steps for implementing analytics in organizational settings.

LinkedIn, for example, uses information about its members, which is housed in operational databases. This information is then organized into a data warehouse and the analysts can use this information to explore the browsing history of LinkedIn members. Furthermore, LinkedIn uses descriptive statistics to generate reports and discover patterns. These patterns let the data analytical team at LinkedIn use predictive analytics to discover that speed is very important in receiving positive responses. Specifically, LinkedIn analysts were able to determine that "adaptation exponentially increases as the response time goes towards sub-seconds." Finally, prescriptive analytics is used to generate appropriate actions. For example, LinkedIn could use optimization techniques to identify the best mix of companies or individuals which maximizes the number of prospects or product sales.

Q6: Select a company where you are a regular customer. Think of the potential data this company stores about you. Also, consider other data sources the company might use (demographics, market information, consumer information such as credit score). How do you think the business uses this information to make you a more valuable customer? For example,

how does your bank know to send emails to you with discount offers for a loan product? How does your favorite restaurant generate a free appetizer coupon to be used during your next visit?

The answer to this question will depend on the specific organization. Consider the Fandango case described in the chapter. Information about you, the customer is combined with movie theaters, ticket sales, and show times and stored into structured databases. This information is extracted, transformed, and loaded into data warehouses or data marts, which mostly reside in distributed servers. Fandango data scientists will then use descriptive analytics. For example, using a sample of movie titles, the analysts can investigate the correlations among total sales for different movies. Using a sample of movie goers, they can calculate the average ticket sales for a week, the most popular movie, distribution of customers among movie genres, the busiest hours of the day in the movie theater, age distribution of movie goers, gender distribution, and so on. This type of data analysis helps Fandango set ticket prices, offer discounts for certain movies or show times, and display show times of the same movie in different theaters.

Q7: Using the same company from the previous question, brainstorm ideas about how the company can improve its operations by lowering costs or improving productivity. For example, how does the bank identify the optimal interest rates for their new loan offerings? How does the restaurant decide the menu prices and happy hour discounts?

Using predictive analytics tools, Fandango can crunch data to determine that while you may like science fiction movies, you have not seen the latest sci-fi movie, which has been in the theaters recently. Prescriptive tools allow for ticket price offerings to change every hour. Fandango has learned when the most desirable movie times are by sifting through millions and millions of show times instantaneously. This information is then used to set an optimal price at any given

time, based on the supply of show times and the demand for movie tickets, thus maximizing their profits

Q8: What is volume in the big data definition? How does the high volume of big data impact descriptive, predictive, and prescriptive analytics? Provide examples to illustrate your ideas.

The volume of big data is larger than the volume processed by conventional relational databases. Today, the high volume of business transactions is automatically captured by advanced enterprise information systems. Descriptive and predictive analytics benefit from the high volume of data. After all, statistical analysis and reliability of predictions is better when the population size increases. A forecasting method with hundreds of factors can predict better than the one with only a few input factors. Prescriptive models also benefit from big data. They are based on aggregated inputs, which are the result of descriptive analytics: contribution coefficients, average processing times, mean of distributions, etc. The validity of these aggregate values improves with high volume data.

Q9: What is velocity in the big data definition? How does high velocity of big data impact descriptive, predictive, and prescriptive analytics? Provide examples to illustrate your ideas.

Velocity is defined as the rate at which data flows into an organization. Online sales, mobile computing, smart phones, and social networks have significantly increased the information flow for the organization. Organizations can analyze customer behavior, sales history, and buying patterns. They are able to quickly produce operational business intelligence and recommend additional purchases or customized marketing strategies. The velocity of system output is also important. The recommendations must be delivered in a timely manner and must

be included as part of business operations. A loan officer, for example, could compare the information in a loan application against business rules and mining models, and make a recommendation to the applicant or make a decision about the loan.

Q10: What is variety in the big data definition? How does a high variety of big data impact descriptive, predictive, and prescriptive analytics? Provide examples to illustrate your ideas.

Variety of data is defined as mix of different data sources in different formats. As mentioned earlier, big data input arrives in the form of a text from social networks or an image from a camera sensor. Even when the data source is structured, the format can be different. Different browsers generate different data. Different users may withhold information, or different vendors may send different information based on the type of software they use. Of course, every time humans are involved, there may be errors, redundancy, and inconsistency. Management science models require the input data to be uniform. As such, the implementation of these models in the era of big data normally requires an additional layer between the source data and the prescriptive model.

Q11: Discuss challenges that organizations face when trying to analyze big data. Make sure to include in your discussion one or more challenges such as privacy invasion, financial exposure, mistaking noise for the signal, and poorly defining business problems.

There is a greater potential for privacy invasion, greater financial exposure in fast-moving markets, greater potential for mistaking noise for true insight, and a greater risk of spending lots of money and time chasing poorly defined problems or opportunities. The everyday use of mathematical modeling and other techniques requires that business managers or

other practitioners have a good understanding of numeracy and mathematical skills. However, there is a lack of such skills, especially for medium sized or small organizations. Information based decisions across organizational boundaries can upset traditional power relationships.

Q12: How important is it to have optimal versus good but practical solutions? Discuss the importance of spreadsheet modeling in this comparison. List other modeling software and compare it to spreadsheets.

It is important to consider a trade-off between less than optimal but time feasible and practical solution and optimal but complex and often delayed solutions. Traditional techniques are modified to better process large volumes of data, offer simpler and practical models, utilize spreadsheet modeling techniques, and offer practical solutions, which can be implemented in real time. Today, several optimization software programs exist, which are able to model and to solve a large number of constraints and decision variables. Solver is an excellent program, licensed to Excel as an add-in from Frontline Systems. Solver can be used by practitioners to solve mathematical programming models and perform what-if analysis and optimizations to determine the best product mix, optimal shipping routes, maximize profit, or minimize costs.

Q13: Explore advantages and disadvantages of Excel modeling for prescriptive analytics. Consider in your discussion the ability of spreadsheets to process large amount of data as well as the potential use of add-ins.

Business analytics, particularly prescriptive analytics, can become more popular with the use of spreadsheet modeling. Spreadsheet modeling is widely used in colleges and universities for teaching mathematical programming. Instead of heavy modeling, which seeks optimal

solutions, spreadsheet modeling techniques include simpler formulations, which seek practical solutions. Several optimization software programs exist, which are able to model and to solve a large number of constraints and decision variables. Solver is an excellent program, licensed to Excel as an add-in from Frontline Systems. Solver can be used by practitioners to solve mathematical programming models and perform what-if analysis and optimizations to determine the best product mix, optimal shipping routes, maximize profit, or minimize costs.

However, the spreadsheets have limitations in the amount of data they can store. They cannot store data about millions of transactions in a bank or the details of federal spending on transportation projects

Q14: Good analytics models can sometimes lead to bad business results, conclusions, and recommendations. List at least three reasons why this might happen. For each reason, offer practical recommendations to avoid erroneous conclusions. Provide examples to illustrate your ideas.

One important step when building mathematical models is the process of abstraction. Through this process, the modeler eliminates or suppresses any unnecessary details and allows only the relevant information to enter the model. When good information goes in the model, a good model will produce good results. The opposite is known as GIGO (garbage in-garbage out). Very often, valid models produce poor results, which lead to the wrong decisions. In the era of big data, this happens very often. A recent story reports how ten volunteers checked the accuracy of their information on AboutTheData.com and they each found inaccuracies. In one specific case, a volunteer found that "she had two teens, at 26." Interestingly, a CNN team found

that Acxiom, the company which runs the database, was more accurate specifying the interests and less accurate in demographic data (marriage status, number of children).

Wrong assumptions can lead to wrong decisions. If you are a company purchasing this database, you know the interests of your future customers, but very likely you may be sending out 2 million direct mails pieces on baby products to people who may not even have children.

Q15: Discuss challenges faced by practitioners when exploring big data with management science models. Suggest practical solutions to these challenges.

Organizations face different types of challenges while implementing optimization models in the era of big data. Management scientists are modifying traditional techniques to better process large volumes of data, offer simpler and practical models, utilize spreadsheet modeling techniques, and offer practical solutions, which can be implemented in real time. High volume of big data requires that decision scientists have the capability to store and process a large amount of data. Cloud computing technology, which has risen over the past few years, has dramatically increased the ability for the businesses to store and process information. This technology offers dynamic and large distributed platforms for organizations to process input parameters and solve models at a large scale. These platforms can be used to run advanced optimization models which engage multiple clusters. The use of a declarative programming approach to model and solve mathematical programming models is still at an early stage.

Apache Hadoop is another good example of using advanced technology to handle high variety of data in optimization models. Hadoop, an open source platform, offers distributed

computing, which places no condition on the structure of the data it can process. As such, Hadoop can be used as a great platform to mitigate the variety components of big data.