5. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
6. What are the common business problems addressed by Big Data analytics?
7. Who is a data scientist? What makes them so much in demand?
8. What are the common characteristics of data scientists? Which one is the most important?
9. In the era of Big Data, are we about to witness the end of data warehousing? Why?
10. What are the use cases for Big Data/Hadoop and data warehousing/RDBMS?
11. What is stream analytics? How does it differ from regular analytics?
12. What are the most fruitful industries for stream analytics? What is common to those industries?
13. Compared to regular analytics, do you think stream analytics will have more (or fewer) use cases in the era of Big Data analytics? Why?

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Go to **teradatauniversitynetwork.com** and search for case studies. Read cases and white papers that talk about Big Data analytics. What is the common theme in those case studies?
2. At **teradatauniversitynetwork.com**, find the SAS Visual Analytics white papers, case studies, and hands-on exercises. Carry out the visual analytics exercises on large data sets and prepare a report to discuss your findings.
3. At **teradatauniversitynetwork.com**, go to the podcasts library. Find podcasts about Big Data analytics. Summarize your findings.
4. Go to **teradatauniversitynetwork.com** and search for BSI videos that talk about Big Data. Review these BSI videos and answer case questions related to them.
5. Go to the **teradata.com** and/or **asterdata.com** Web sites. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
6. Go to **IBM.com**. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
7. Go to **cloudera.com.** Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
8. Go to **MapR.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
9. Go to **hortonworks.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
10. Go to **marklogic.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
11. Go to **youtube.com**. Search for videos on Big Data computing. Watch at least two. Summarize your findings.
12. Go to **google.com/scholar** and search for articles on stream analytics. Find at least three related articles. Read and summarize your findings.
13. Enter **google.com/scholar** and search for articles on data stream mining. Find at least three related articles. Read and summarize your findings.
14. Search the job search sites like **monster.com**, **careerbuilder.com**, and so forth. Find at least five job postings for data scientist. Identify the key characteristics and skills expected from the applicants.
15. Enter **google.com/scholar** and search for articles that talk about Big Data versus data warehousing. Find at least five articles. Read and summarize your findings.

End-of-Chapter Application Case

Discovery Health Turns Big Data into Better Healthcare

Introduction—Business Context

Founded in Johannesburg more than 20 years ago, Discovery now operates throughout the country, with offices in most major cities to support its network of brokers. It employs more than 5,000 people and offers a wide range of health, life and other insurance services.

In the health sector, Discovery prides itself on offering the widest range of health plans in the South African market. As one of the largest health scheme administrators in the country, its is able to keep member contributions as low as possible, making it more affordable to a wider cross-section of the population. On a like-for-like basis, Discovery's plan contributions are as much as 15 percent lower than those of any other South African medical scheme.

Business Challenges

When your health schemes have 2.7 million members, your claims system generates a million new rows of data daily,

and you are using three years of historical data in your analytics environment, how can you identify the key insights that your business and your members' health depend on?

This was the challenge facing Discovery Health, one of South Africa's leading specialist health scheme administrators. To find the needles of vital information in the big data haystack, the company not only needed a sophisticated data-mining and predictive modeling solution, but also an analytics infrastructure with the power to deliver results at the speed of business.

Solutions—Big Data Analytics

By building a new accelerated analytics landscape, Discovery Health is now able to unlock the true potential of its data for the first time. This enables the company to run three years' worth of data for its 2.7 million members through complex statistical models to deliver actionable insights in a matter of minutes. Discovery is constantly developing new analytical applications, and has already seen tangible benefits in areas such as predictive modeling of members' medical needs and fraud detection.

Predicting and preventing health risks

Matthew Zylstra, Actuary, Risk Intelligence Technical Development at Discovery Health, explains: "We can now combine data from our claims system with other sources of information such as pathology results and members' questionnaires to gain more accurate insight into their current and possible future health.

"For example, by looking at previous hospital admissions, we can now predict which of our members are most likely to require procedures such as knee surgery or lower back surgery. By gaining a better overview of members' needs, we can adjust our health plans to serve them more effectively and offer better value."

Lizelle Steenkamp, Divisional Manager, Risk Intelligence Technical Development, adds: "Everything we do is an attempt to lower costs for our members while maintaining or improving the quality of care. The schemes we administer are mutual funds–non-profit organizations–so any surpluses in the plan go back to the members we administer, either through increased reserves or lowered contributions. "One of the most important ways we can simultaneously reduce costs and improve the well-being of our members is to predict and prevent health problems before they need treatment. We are using the results of our predictive modeling to design preventative programs that can help our members stay healthier."

Identifying and eliminating fraud

Estiaan Steenberg, Actuary at Discovery Health, comments: "From an analytical point of view, fraud is often a small intersection between two or more very large data-sets. We now have the tools we need to identify even the tiniest anomalies and trace suspicious transactions back to their source."

For example, Discovery can now compare drug prescriptions collected by pharmacies across the country with healthcare providers' records. If a prescription seems to have been issued by a provider, but the person fulfilling it has not visited that provider recently, it is a strong indicator that the prescription may be fraudulent. "We used to only be able to run this kind of analysis for one pharmacy and one month at a time," says Estiaan Steenberg. "Now we can run 18 months of data from all the pharmacies at once in two minutes. There is no way we could have obtained these results with our old analytics landscape."

Similar techniques can be used to identify coding errors in billing from healthcare providers–for example, if a provider "upcodes" an item to charge Discovery for a more expensive procedure than it actually performed, or "unbundles" the billing for a single procedure into two or more separate (and more expensive) lines. By comparing the billing codes with data on hospital admissions, Discovery is alerted to unusual patterns, and can investigate whenever mistakes or fraudulent activity are suspected.

The Results—Transforming Performance

To achieve this transformation in its analytics capabilities, Discovery worked with BITanium, an IBM Business Partner with deep expertise in operational deployments of advanced analytics technologies. "BITanium has provided fantastic support from so many different angles," says Matthew Zylstra. "Product evaluation and selection, software license management, technical support for developing new models, performance optimization and analyst training are just a few of the areas they have helped us with."

Discovery is an experienced user of IBM SPSS® predictive analytics software, which forms the core of its data-mining and predictive analytics capability. But the most important factor in embedding analytics in day-to-day operational decision-making has been the recent introduction of the IBM PureData™ System for Analytics, powered by Netezza® technology–an appliance that transforms the performance of the predictive models.

"BITanium ran a proof of concept for the solution that rapidly delivered useful results," says Lizelle Steenkamp. "We were impressed with how quickly it was possible to achieve tremendous performance gains." Matthew Zylstra adds: "Our data warehouse is so large that some queries used to take 18 hours or more to process–and they would often crash before delivering results. Now, we see results in a few minutes, which allows us to be more responsive to our customers and thus provide better care."

From an analytics perspective, the speed of the solution gives Discovery more scope to experiment and optimize its models. "We can tweak a model and re-run the analysis in a few minutes," says Matthew Zylstra "This means we can do more development cycles faster–and release new analyses to the business in days rather than weeks."

From a broader business perspective, the combination of SPSS and PureData technologies gives Discovery the ability to put actionable data in the hands of its decision-makers faster. "In sensitive areas such as patient care and fraud investigation, the details are everything," concludes Lizelle Steenkamp. "With the IBM solution, instead of inferring a 'near enough' answer from high-level summaries of data, we can get the right information,

develop the right models, ask the right questions, and provide accurate analyses that meet the precise needs of the business."

Looking to the future, Discovery is also starting to analyze unstructured data, such as text-based surveys and comments from online feedback forms.

About BITanium

BITanium believes that the truth lies in data. Data does not have its own agenda, it does not lie, it is not influenced by promotions or bonuses. Data contains the only accurate representation of what has and is actually happening within a business. BITanium also believes that one of the few remaining differentiators between mediocrity and excellence is how a company uses its data.

BITanium is passionate about using technology and mathematics to find patterns and relationships in data. These patterns provide insight and knowledge about problems, transforming them into opportunities. To learn more about services and solutions from BITanium, please visit bitanium.co.za.

About IBM Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management, and risk management. Business Analytics solutions enable companies to identify and visualize trends and patterns in areas, such as customer analytics, that can have a profound effect on business performance. They can compare scenarios, anticipate potential threats and opportunities, better plan, budget and forecast resources, balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision-making to achieve business goals. For more information, you may visit ibm.com/business-analytics.

Questions for the End-of-Chapter Application Case

1. How big is Big Data for Discovery Health?
2. What big data sources did Discovery Health use for their analytic solutions?
3. What were the main data/analytics challenges Discovery Health was facing?
4. What were the main solutions they have produced?
5. What were the initial results/benefits? What do you think will be the future of Big Data analytics at Discovery?

Source: **IBM Customer Story, "Discovery Health turns big data into better healthcare"** public.dhe.ibm.com/common/ssi/ecm/en/ytc03619zaen/YTC03619ZAEN.PDF (accessed October 2013).

References

Awadallah, A., and D. Graham. (2012). "Hadoop and the Data Warehouse: When to Use Which." White paper by Cloudera and Teradata. **teradata.com/white-papers/Hadoop-and-the-Data-Warehouse-When-to-Use-Which** (accessed March 2013).

Davenport, T. H., and D. J. Patil. (2012, October). "Data Scientist." *Harvard Business Review,* pp. 70–76.

Dean, J., and S. Ghemawat. (2004). "MapReduce: Simplified Data Processing on Large Clusters." **research.google.com/archive/mapreduce.html** (accessed March 2013).

Delen, D., M. Kletke, and J. Kim. (2005). "A Scalable Classification Algorithm for Very Large Datasets." *Journal of Information and Knowledge Management,* Vol. 4, No. 2, pp. 83–94.

Ericsson. (2012). "Proof of Concept for Applying Stream Analytics to Utilities." Ericsson Labs, Research Topics, **labs.ericsson.com/blog/proof-of-concept-for-applying-stream-analytics-to-utilities** (accessed March 2013).

Issenberg, S. (2012, October 29). "Obama Does It Better" (from "Victory Lab: The New Science of Winning Campaigns"), *Slate.*

Jonas, J. (2007). "Streaming Analytics vs. Perpetual Analytics (Advantages of Windowless Thinking)." **jeffjonas.typepad.com/jeff_jonas/2007/04/streaming_analy.html** (accessed March 2013).

Kelly, L. (2012). "Big Data: Hadoop, Business Analytics and Beyond." **wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond** (accessed January 2013).

Kelly, L. (2013). "Big Data Vendor Revenue and Market Forecast 2012–2017." **wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017** (accessed March 2013).

Romano, L. (2012, June 9). "Obama's Data Advantage." *Politico.*

Russom, P. (2013). "Busting 10 Myths about Hadoop: The Big Data Explosion." TDWI's *Best of Business Intelligence,* Vol. 10, pp. 45–46.

Samuelson, D. A. (2013, February). "Analytics: Key to Obama's Victory." INFORMS' *ORMS Today,* pp. 20–24.

Scherer, M. (2012, November 7). "Inside the Secret World of the Data Crunchers Who Helped Obama Win." *Time.*

Shen, G. (2013, January–February). "Big Data, Analytics, and Elections." INFORMS' *Analytics Magazine.*

Watson, H. (2012). "The Requirements for Being an Analytics-Based Organization." *Business Intelligence Journal,* Vol. 17, No. 2, pp. 42–44.

Watson, H., R. Sharda, and D. Schrader. (2012). "Big Data and How to Teach It." Workshop at AMCIS. Seattle, WA.

White, C. (2012). "MapReduce and the Data Scientist." Teradata Aster white paper. **teradata.com/white-paper/MapReduce-and-the-Data-Scientist** (accessed February 2013).

Zikopoulos, P., D. DeRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. (2013). *Harness the Power of Big Data.* New York: MacGraw Hill Publishing.