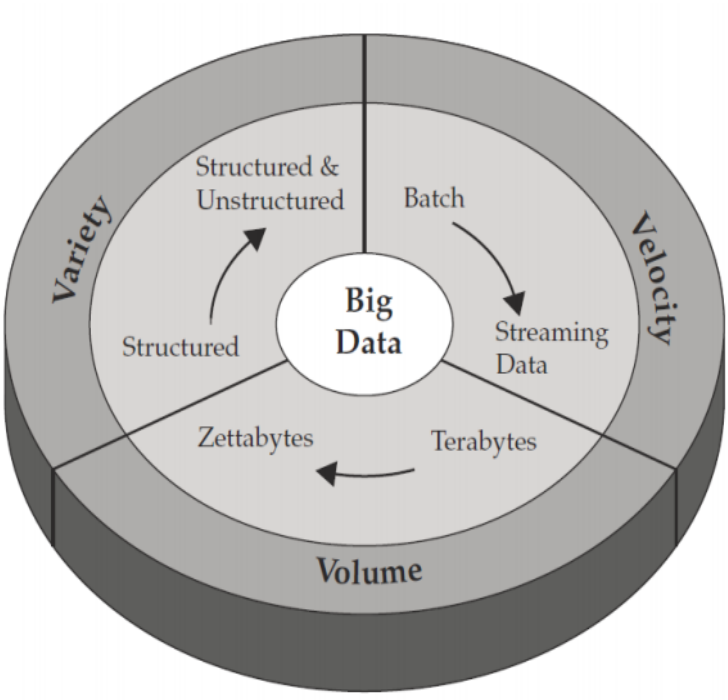
Big Data Assignment

Question 1: Big Data Introduction?

Answer: Big Data is a broad term for data sets so large or complex that they are difficult to process using traditional data processing applications. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy.

Characterization of big data



VOLUME

The volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying. Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it.

Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements theMapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage.

To store data, Hadoop utilizes its own distributed filesystem, HDFS, which makes data available to multiple computing nodes. A typical Hadoop usage pattern involves three stages:

- loading data into HDFS,
- MapReduce operations, and
- retrieving results from HDFS.

This process is by nature a batch operation, suited for analytical or non-interactive computing tasks. Because of this, Hadoop is not itself a database or data warehouse solution, but can act as an analytical adjunct to one.
One of the most well-known Hadoop users is Facebook, whose model follows this pattern. A MySQL database stores the core data. This is then reflected into Hadoop, where computations occur, such as creating recommendations for you based on your friends' interests. Facebook then transfers the results back into MySQL, for use in pages served to users.

Velocity
The importance of data's velocity — the increasing rate at which data flows into an organization — has followed a similar pattern to that of volume. Problems previously restricted to segments of industry are now presenting themselves in a much broader setting. Specialized companies such as financial traders have long turned systems that cope with fast moving data to their advantage.

VARIETY
Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in big data systems is that the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a sensor source. None of these things come ready for integration into an application. A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application.

Question 2: State of the Practice in Analytics.
Answer:
Advanced analytics is fairly common today. Roughly three quarters (74%) of organizations surveyed have adopted some form of analytics today, regardless of the analytic method or tool type, whether with big data or not. This reveals a strong adoption of advanced analytics, which isn't a surprise, given that it's been around for at least 15 years. Analytics doesn't require big data. The two get jammed into the same sentence so much lately that we forget that they don't have to go together. In fact, 40% of survey respondents practice advanced analytics without big data.

One-third of organizations (34%) do big data analytics today, although it's new. In other words, they practice some form of advanced analytics, and they apply it to big data.

This is a respectable presence for big data analytics, given the newness of the combination of advanced analytics and big data.

Benefits of Big Data Analytics

We just saw that user organizations have adopted big data analytics in appreciable numbers. To determine the potential benefits that are driving the adoption, TDWI's survey asked: "Which of the following benefits would ensue if your organization implemented some form of big data analytics?"

The most likely benefits are those most often selected by survey respondents, and the likelihood of a benefit declines as the list proceeds downward.

Anything involving customers could benefit from big data analytics. Near the top of the list (in Figure 2), this includes better-targeted social-influencer marketing (61%), customer-base segmentation (41%), and recognition of sales and market opportunities (38%). Recent economic changes worldwide have changed consumer behaviors. Big data analytics can help develop definitions of churn and other customer behaviors (35%), as well as an understanding of consumer behavior from clickstreams (27%).

Business intelligence in general can benefit from big data analytics. This could result in more numerous and accurate business insights (45%), an understanding of business change (30%), better planning and forecasting (29%), and the identification of root causes of cost (29%).

Specific analytic applications are likely beneficiaries of big data analytics. For example, consider analytic applications for the detection of fraud (33%), the quantification of risks (30%), or market sentiment trending (30%). At the leading edge, big data analytics might help automate decisions for real-time business processes such as loan approvals or fraud detection (37%). Potential benefits entered by survey respondents selecting "other" include customer loyalty, service experience optimization, healthcare delivery optimization, and supplier performance based on cost and quality.

Question 3: When to consider Big Data solutions?

Answer: The term Big Data can be interpreted in many different ways. The Big Data solutions aren't a replacement for the existing warehouse solutions. But have their own advantages over the warehouses.

A few principles when we should consider Big Data solutions include:

1. Big Data solutions are ideal for analyzing not only raw structured data, but semistructured and unstructured data from a wide variety of sources.
2. Big Data solutions are ideal when all, or most, of the data needs to be analyzed versus a sample of the data; or a sampling of data isn't nearly as effective as a larger set of data from which to derive analysis.
3. Big Data solutions are ideal for iterative and exploratory analysis when business measures on data are not predetermined.
4. Big Data solution not only leverage s data not typically suitable for a traditional warehouse environment, and in massive amounts of volume, but it gives up

some of the formalities and “strictness” of the data. The benefit is that you can preserve the fidelity of data and gain access to mountains of information for exploration and discovery of business insights.

5. Big Data is well suited for solving information challenges that do not natively fit within a traditional relational database approach for handling the problem at hand.

Since there exists some class of problems that don’t natively belong in traditional databases and there’s data that we’re not sure we want in warehouse, because perhaps we don’t know if it’s rich in value, it’s unstructured, or it’s too voluminous. In many cases, we can’t find out the value per byte of the data until after we spend the effort and money to put it into the warehouse; but we want to be sure that data is worth saving and has a high value per byte before investing in it. Big Data comes into action in such cases.

Question 4. Big Data For Market and Research, Examples.

Answer: For Market -

1. Monitor Google Trends to Inform Your Global/Local Strategy

Google Trends is probably the most approachable method of utilizing Big Data. Google Trends showcases trending topics by quantifying how often a particular search-term is entered relative to the total search-volume. Global marketers can use Google Trends to assess the popularity of certain topics across countries, languages, or other constituencies they might be interested in, or, stay informed on what topics are cool, hip, top-of-mind or relevant to their buyers.

2. Use Digital Information to More Clearly Define Your ICP

Use heaps of analytics to learn more about your target buyers than you’ve ever known before.

Whereas in years past, marketers would make educated guesses at the age, demographics, and work profile of their target buyer, modern marketers have vats of data intelligence to prove their intuitions, and shed light on a more granular level of detail, such as: which web sites a user frequents most often, which social media profiles they have and use, and even which buttons they click on a given website.

ICP (or Ideal Customer Profiles) can be extremely targeted, while also data-backed.

For instance, in an Avis Budget case study, Tim Doolittle, vice president of CRM and marketing science, said adding Big Data to understand their customer profile

“...increased the effectiveness of our contact strategy, in many cases above 30% over control.”

3. Create Real-Time Personalization to Buyers

Marketers need to send the right message at the right time. Timeliness and relevancy aren’t just qualities of the fourth estate; they’re also the foundation of successful marketing campaigns, email click-through rates, and consumer engagement with your brand.

Big Data gives marketers the most timely insights into who is interested or engaging with their product or content in real time. Tying buyer digital behavior into your CRM systems and marketing automation software allows you to track the topics that your buyers are most interested in and send them content that makes the most sense to develop those ideas or extrapolate on those topics.

On average, companies collect customer and prospect data from 3.4 channels. Most commonly, this include the company's website, followed by the sales team and then the call center.

4. Identify the Specific Content that Moves Buyers Down the Sales Funnel

How successful was a singular blog or social post at generating revenue? Before Big Data that was an unanswerable question. We executed on social media strategies and content creation because we had a feeling that it was working, but we had no way to back that claim.

Now, marketers can distill the effectiveness of a marketing push down the to tweet. Tools like content scoring illuminate which individual content assets were successful to a closed / won deal, and which were inefficient. The allows marketers to hone the strategies around the content topics or types that resonate with their buyers the most, and truly compel them to purchase.

5. Integrate Predictive Analytics Into Your Lead Scoring Strategy

Predictive analytics is one of the most progressive (and maybe aggressive) strategies marketers can employ with Big Data.

In particular, marketers are seeing high rates of success in predictive lead scoring, which uses a company’s base CRM data and other third party Internet data to generate a model that successfully predicts future lead behavior. It pools and analyzes historical

data around successful leads (leads that became closed won), thereby giving marketers clear indications about which digital behaviors are hand-raising activities or should be weighed more heavily in lead scoring.

Already, companies excelling at traditional lead nurturing generate 50% more sales-ready leads at 33% lower cost. The possibility that predictive lead scoring dwarfs that, is likely.

For Research -

Will big data make primary research obsolete?

This was the usual thought process in what could be called the “before big data” world. Whether the research objective was to segment consumer needs to improve targeting or to evaluate the impact of advertising on brand health, it was reasonable to assume that we would need to generate and analyze a new set of data on each occasion. But that assumption is no longer valid. In today’s big data world, nearly everything is passively observed and managed in a digitized fashion; thus we have the ability to use data assets that were previously untapped or nonexistent to quickly and deeply address these same topics. Big Data isn’t really a brand-new phenomenon; for years now, large data sources have included information on customer purchases, credit scores, and lifestyle information. And for years, data scientists have used this data to help businesses evaluate risk and anticipate customer needs. The difference today is twofold: more sophisticated tools and methods are available to analyze and combine various datasets, and these analytic tools are now augmented by an avalanche of new data sources ignited by the digitization of nearly all data collection and measurement. The range of content now available is both inspiring and intimidating to researchers raised in the structured survey environment. Consumer sentiment is captured on websites and the variety of social media outlets. Exposure to advertising is recorded not only by set-top boxes but also by digital tags and mobile devices communicating with TVs

Behavioral outcomes such as call volume, shopping patterns, and purchases are now available in real time. Thus many of the insights that were previously provided by survey research can now be discerned through big data sources. And all of these data assets are generated on an ongoing basis, independent of any research process. These are the changes that motivate the question of whether big data will replace market research.

The Benefits of Liberated Research

This brings us back to how big data is not replacing research but rather is liberating it. Researchers are liberated from having to generate a new survey on each new learning occasion; ongoing big data assets can be leveraged for many topics, allowing subsequent primary research to go deeper and fill in the gaps. Researchers are liberated from needing to rely upon bloated surveys and instead can keep surveys short

and focused on those variables that they are ideally suited for, resulting in better data quality

Once liberated, researchers can use their established first principles and insights to impart accuracy and meaning into the big data assets, leading to new areas of survey-based exploration. This cycle should lead to deeper insights across a range of strategic issues, ultimately moving toward what should always be our primary objective—to inform and improve brand and communications decisions.

Question 6: Why Small Data Rather Than Big Data
Answer:

The idea that Big Data can improve your marketing efforts rests on the belief that data in general is essential for making the right decisions, and this cannot be denied. However, the unending quest for more information and more tools may be preventing your marketing teams from understanding the true value of using some of the small data. It’s not about having a lot of data. Rather, it's about having the right kind of data, and more importantly, it's about the ability to extract actionable insights for effective marketing.
The idea that Big Data can improve your marketing efforts rests on the belief that data in general is essential for making the right decisions, and this cannot be denied. However, the unending quest for more information and more tools may be preventing your marketing teams from understanding the true value of using some of the small data. It’s not about having a lot of data. Rather, it's about having the right kind of data, and more importantly, it's about the ability to extract actionable insights for effective marketing.
For most companies, this means having marketing leaders in place who know how to analyze the data. The ability to transform it from a big pile of unfiltered information into smaller, more focused metrics and KPIs is essential.
A mountain of data about a company’s overall leads or marketing’s contribution to the sales pipeline could be more useful than having tons of reports about hundreds of lead sources and campaigns. Marketing leaders need smaller data or just a few key metrics and KPIs to prevent getting lost in too much Big Data. This allows marketing execs to prioritize their efforts quickly and focus on campaigns that have the best ROI. In the long run, focusing on activities that carry the most impact on growing sales is a better strategy than just producing more leads that may not convert well.
In the modern world of Big Data, it may seem attractive to gather as much as you can, but the reality is that less is often more. Big Data may be stealing the limelight for its novelty and sheer size, but small data with a few key insights is much more likely to help your marketing become more effective.
Small Data sets are best implemented over **traditional databases** whereas Big Datasets are best implemented over **HDFS** and other big data file systems.
Moreover, it all depends of what our needs are. The requirements and constraints are as follows and so are the choices:

1. Is the data being analyzed structured or unstructured?

Structured Data: Data that resides within the fixed confines of a record or file is known as structured data. Owing to the fact that structured data – even in large volumes – can be entered, stored, queried, and analyzed in a simple and straightforward manner, this type of data is best served by a **traditional database**.

Unstructured Data: Data that comes from a variety of sources, such as emails, text documents, videos, photos, audio files, and social media posts, is referred to as unstructured data. Being both complex and voluminous, unstructured data cannot be handled or efficiently queried by a traditional database. Hadoop's ability to join, aggregate, and analyze vast stores of multi-source data without having to structure it first allows organizations to gain deeper insights quickly. Thus **Hadoop** is a perfect fit for companies looking to store, manage, and analyze large volumes of unstructured data.

2. Is a scalable analytics infrastructure needed?

Companies whose data workloads are constant and predictable will be better served by a **traditional database.**

Companies challenged by increasing data demands will want to take advantage of **Hadoop's** scalable infrastructure. Scalability allows servers to be added on demand to accommodate growing workloads. As a cloud-based Hadoop service, Qubole offers more flexible scalability by spinning virtual servers up or down within minutes to better accommodate fluctuating workloads.

3. Will a Hadoop implementation be cost-effective?

Cost-effectiveness is always a concern for companies looking to adopt new technologies. When considering a Hadoop implementation, companies need to do their homework to make sure that the realized benefits of a Hadoop deployment outweigh the costs. Otherwise it would be best to stick with a **traditional database** to meet data storage and analytics needs.

All things considered, **Hadoop** has a number of things going for it that make implementation more cost-effective than companies may realize. For one thing, Hadoop saves money by combining open source software with commodity servers. Cloud-based Hadoop platforms such as Qubole reduce costs further by eliminating the expense of physical servers and warehouse space.

Hybrid systems, which integrate Hadoop platforms with traditional relational databases, are gaining popularity as cost-effective ways for companies to leverage the benefits of both platforms.

4. Is fast data analysis critical?

Hadoop was designed for large distributed data processing that addresses every file in the database. And that type of processing takes time. For tasks where fast performance isn't critical, such as running end-of-day reports to review daily transactions, scanning historical data, and performing analytics where a slower time-to-insight is acceptable, **Hadoop** is ideal.

On the other hand, in cases where organizations rely on time-sensitive data analysis, a **traditional database** is the better fit. That's because shorter time-to -insight isn't about

analyzing large unstructured datasets, which Hadoop does so well. It's about analyzing smaller data sets in real or near-real time, which is what traditional databases are well equipped to do.
Hybrid systems are also a good fit to consider, as they allow companies to use traditional databases to run small, highly interactive workloads while using Hadoop to process huge, complex data sets.

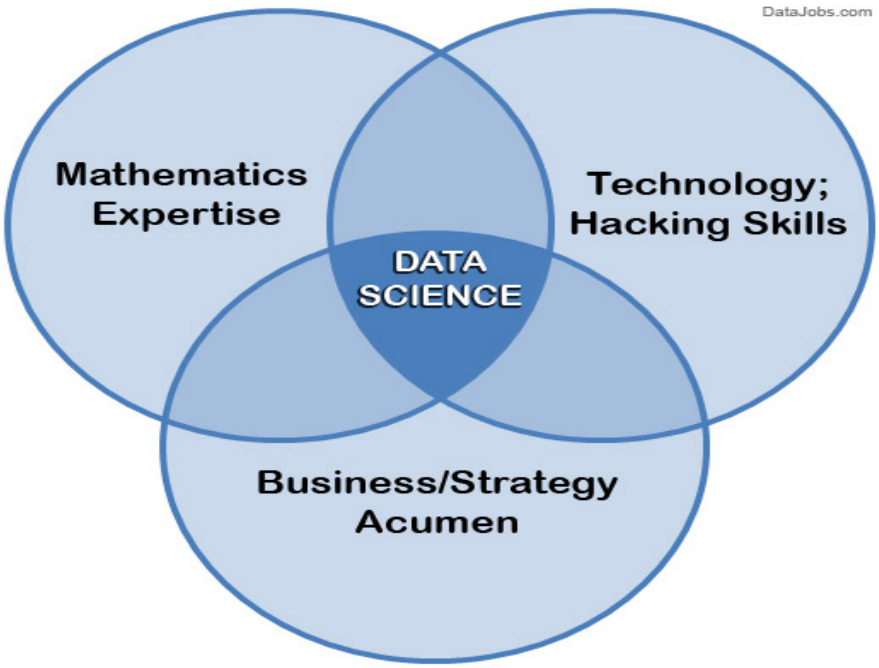
Question 7: **The Data Science and Big Data and**
Question 8: Considering Big Data Independent of Data Science and Vice Versa

Data science is deep knowledge discovery through data inference and exploration. This discipline often involves using mathematic and algorithmic techniques to solve some of the most analytically complex business problems, leveraging troves of raw information to figure out hidden insight that lies beneath the surface. It centers around evidence-based analytical rigor and building robust decision capabilities. Ultimately, data science matters because it enables companies to operate and strategize more intelligently. It is all about adding substantial enterprise value by learning from data.

The variety of projects that a data scientist may be engaged in is incredibly broad. Here are few examples:

- tactical optimization – improvement of marketing campaigns, business processes, etc
- predictive analytics – anticipate future demand, future events, etc
- nuanced learning – e.g. developing deep understanding of consumer behavior
- recommendation engines – e.g. Amazon product recs, Netflix movie recs
- automated decision engines – e.g. automated fraud detection, and even self-driving cars

Data science is multidisciplinary; the skill set of a data scientist lies at the intersection of 3 main competencies:



Mathematics Expertise

At the heart of deriving insight from data is the ability to view the data through a quantitative lens. There are textures, patterns, dimensions, and correlations in data that can be expressed numerically, and discovering inference from data becomes a brain teaser of mathematical techniques. Solutions to many business problems often involve building analytic models that are deeply grounded in the hard math theory, and being able to understand how models work is as important as knowing the process to build them

Technology and Hacking

Data scientists absolutely need to leverage *technology* in order to wrangle enormous data sets and work with complex algorithms, and it requires using tools far more sophisticated than Excel. Examples of such tools are SQL, SAS, and R, all of which require technical/coding ability. With these high-performance tools, a true 'hacker' is a technical ninja, able to use ingenious problem solving ability to achieve mastery in data exploration – piecing together unstructured information and teasing out golden nuggets of insight.

Strong Business Acumen

It is very important to note that a data scientist is first and foremost a strategy consultant. Data science teams have become invaluable resources within companies because by being able to learn from data in ways no one else can, they are extraordinarily well-positioned to figure out how to add substantial business value. But this means having a keen sense of how to dissect and approach business problems becomes as important as having a keen sense of how to approach algorithmic problems. Ultimately, the value doesn't come from numbers; it comes from strategic thinking based on those numbers.

Difference between an analyst and a data scientist

"Analyst" is somewhat of an ambiguous term that can represent many different types of roles (marketing analyst, operations analyst, portfolio analyst, financial analyst, etc). Is an analyst the same as a data scientist? We've discussed pretty strict canon around what is a data scientist – as an expert's role with requisite talents in math, technology, and strategy consulting. Let's just say that some analysts are definitely data-scientists-in-training. As represented in this visual, there is a place in the middle where the distinction can blur a bit.

Here are examples of growth from analyst to veritable data scientist:

- An analyst who has previously only mastered Excel, learns how to dive into raw warehouse data using SQL and R
- An analyst who previously only knew enough stats to report the results of an A/B test, gains the expertise to build a predictive model with latent variable analysis and cross-validation

Overall point is that moving in the direction of "data scientist" requires motivation to learn many new skills. Many companies have actually found success cultivating their own home-grown data scientists, by giving their analysts the resources and training to take their abilities to the next level.

Big Data independence from Data Science

Re-develop your products

Big Data can also help you understand how others perceive your products so that you can adapt them, or your marketing, if need be. Analysis of unstructured social media text allows you to uncover the sentiments of your customers and even segment those in different geographical locations or among different demographic groups.

On top of that, Big Data lets you test thousands of different variations of computer-aided designs in the blink of an eye so that you can check how minor changes in, for instance, material affect costs, lead times and performance. You can then raise the efficiency of the production process accordingly.

Perform risk analysis

Success not only depends on how you run your company. Social and economic factors are crucial for your accomplishments as well. Predictive analytics, fueled by Big Data allows you to scan and analyze newspaper reports or social media feeds so that you permanently keep up to speed on the latest developments in your industry and its environment. Detailed health-tests on your suppliers and customers are another goodie that comes with Big Data. This will allow you to take action when one of them is in risk of defaulting.

Keeping your data safe

You can map the entire data landscape across your company with Big Data tools, thus allowing you to analyze the threats that you face internally. You will be able to detect potentially sensitive information that is not protected in an appropriate manner and make sure it is stored according to regulatory requirements. With real-time Big Data analytics you can, for example, flag up any situation where 16 digit numbers – potentially credit card data - are stored or emailed out and investigate accordingly.

Create new revenue streams

The insights that you gain from analyzing your market and its consumers with Big Data are not just valuable to you. You could sell them as non-personalized trend data to large industry players operating in the same segment as you and create a whole new revenue stream.

One of the more impressive examples comes from Shazam, the song identification application. It helps record labels find out where music sub-cultures are arising by monitoring the use of its service, including the location data that mobile devices so conveniently provide. The record labels can then find and sign up promising new artists or remarket their existing ones accordingly.

Customize your website in real time

Big Data analytics allows you to personalize the content or look and feel of your website in real time to suit each consumer entering your website, depending on, for instance, their sex, nationality or from where they ended up on your site. The best-known example is probably offering tailored recommendations: Amazon’s use of real-time, item-based, collaborative filtering (IBCF) to fuel its ‘Frequently bought together’ and

‘Customers who bought this item also bought’ features or LinkedIn suggesting ‘People you may know’ or ‘Companies you may want to follow’. And the approach works: Amazon generates about 20% more revenue via this method.

Reducing maintenance costs

Traditionally, factories estimate that a certain type of equipment is likely to wear out after so many years. Consequently, they replace every piece of that technology within that many years, even devices that have much more useful life left in them. Big Data tools do away with such unpractical and costly averages. The massive amounts of data that they access and use and their unequalled speed can spot failing grid devices and predict when they will give out. The result: a much more cost-effective replacement strategy for the utility and less downtime, as faulty devices are tracked a lot faster.

Offering tailored healthcare

We are living in a hyper-personalized world, but healthcare seems to be one of the last sectors still using generalized approaches. When someone is diagnosed with cancer they usually undergo one therapy, and if that doesn’t work, the doctors try another, etc. But what if a cancer patient could receive medication that is tailored to his individual genes? This would result in a better outcome, less cost, less frustration and less fear.

With human genome mapping and Big Data tools, it will soon be commonplace for everyone to have their genes mapped as part of their medical record. This brings medicine closer than ever to finding the genetic determinants that cause a disease and developing drugs expressly tailored to treat those causes — in other words, personalized medicine.

Offering enterprise-wide insights

Previously, if business users needed to analyze large amounts of varied data, they had to ask their IT colleagues for help as they themselves lacked the technical skills for doing so. Often, by the time they received the requested information, it was no longer useful or even correct. With Big Data tools, the technical teams can do the groundwork and then build repeatability into algorithms for faster searches. In other words, they can develop systems and install interactive and dynamic visualization tools that allow business users to analyze, view and benefit from the data.

Question 9: Five repositories for downloading large data sets.

Answer:

- The 1000 Genomes project makes 260 TB of human genome data available
- The Internet Archive makes an 80 TB web crawl available for research
- CNetS at Indiana University makes a 2.5 TB click dataset available
- ICWSM made a large corpus of blog posts available for their 2011 conference. You'll have to register, but it's free. It's about 2.1 TB compressed.
- The Proteome Commons makes several large data sets available. The largest, the Personal Genome Project, is 1.1 TB in size. There are several others over 100 GB in size.