BIG DATA - THE REVOLUTION THAT COULD CHANGE EVERYTHING

Prof. Rhoda D'Sa

The English philosopher Sir Francis Bacon famously wrote in the 16th century that knowledge is power. The big data revolution will enhance the quality and quantity of our knowledge base in new ways, giving us the power to dramatically improve our lives.

The big data revolution is driven by two factors. Firstly, the volume, velocity and variety of information being generated and stored in this decade far exceeds anything the world has ever known. Secondly, computing hardware and software have developed stupendous capacities to efficiently analyze these exabytes of data.

Those who seek power, desire tools that give them the upper hand over their competitors, the so-called winning edge. Knowledge derived from data analytics is a significant aspect of this quest. Corporations, governments, universities, scientists and everybody else, are focused on leveraging the big data revolution that is poised to change everything.

So what is this big data revolution? This paper tries to find out.

Flashback

To appreciate the big data revolution, we need some background. We begin with a flashback of the changing face of data management.

More than 5000 years ago, people used tally sticks and the abacus to keep track of things. Civilised societies maintained data records on clay tablets, palm leaves, scrolls and so on. Repositories such as the great library of Alexandria and the nine-storey library of Nalanda were famous. Population censuses were conducted in the Roman Empire. Later, the invention of paper and printing technology revolutionized the storage and transmission of information. Over the course of centuries, data management gradually became more sophisticated.

From the middle of the 20th century, dramatic developments have occurred at an accelerated pace. In the '50s, Hollerith cards and magnetic storage of data were invented and the first computers were operational. In the '60s, data centers in the US had digitized and stored millions of tax returns and fingerprints. In the '70s IBM 370 and pocket calculators were on the market; RDBM systems were running and the first email was sent. Supercomputers and PCs appeared in the '80s. In the '90s the Internet was born and the first search engine launched. In a short span of fifty years, the world of information management was transformed.

In the 21st century, the thirst for large-scale data is seemingly insatiable both in the public and private sectors. The Government of India conducts a demographic census every 10 years; through the CSO, RBI and NSSO, primary data are collected on a wide variety of socio-economic indicators of national importance. In the corporate world, business and financial data is routinely gathered and treated as a valuable resource. The market for business information out of databases that store millions of records is flourishing. Google, the Web giant, has created the most sophisticated information-gathering and data-management apparatus in history.

The business of data management comprises four broad processes which convert data into knowledge.

A. Observational Evidence

B. Organisation and Storage

C. Retrieval and Transmission

D. Analysis and Interpretation

A. Observational Evidence

Natural and experimental Data: We collect data by observing (a) natural phenomena and (b) experimentally generated objects and effects. There is a distinction between the two. Observation of what is happening 'naturally' is different from conducting an experiment to generate data. Each has its own set of problems commonly resolved by employing standard statistical techniques.

Heterogeneity: When data are collected by observing natural phenomena – and not from controlled experiments – more often than not, we are dealing with a heterogeneous population. In such situations, a stratified sampling design is useful.

Criteria: The observed data must be closely connected to the phenomena one wishes to study. For example, to study employee productivity in a sales outlet, the criterion could be the time spent at work or the number of customers handled or the transactions completed. If one chooses 'transactions', should one measure the number of transactions per day or the daily average value?

Measurement: What is observed is as important as how these observations are perceived and recorded. Making an inventory of finished goods coming off an assembly line is straightforward, whereas it is difficult to observe marine life in a flowing river. When physicists study particles in a bubble chamber, they are investigating something that is invisible to the human eye! These examples illustrate the kind of difficulties encountered in obtaining reliable data. In particular, the physicist's experiment involves non-perceptual phenomena that do not register on the human senses, but by some recording equipment; which raises the question – can this still be termed observational evidence?

The Human Factor: The subjective element kicks in when human perception and bias affect observation. It will suffice to recall the tale of the elephant and the five blind men! When data are collected by many people using multiple instruments at different times and diverse locations, imagine what could happen to the consistency and uniformity of the whole set of observations.

From the above, it is obvious that human as well as other sorts of errors are an inherent aspect of collecting observational evidence. Fortunately, there are statistical methods to deal with it.

B. Organisation and Storage

The history of the replacement of manual methods by electronic devices and digital records such as magnetic tapes to siloed data to remote cloud-based storage is closely linked to the evolution of computers and telecom technology. Initially,

when statistics and computers met, the focus was on electronic data processing, tabulations, DBMS and MIS. This symbiotic relationship soon became increasingly complex, driven by the explosive growth of information, rapid advances in computer technology and the Internet. Where storage is concerned, we have figuratively reached the clouds!

We know that data is exploding, but just how much data is out there? The Harvard Business Review states that more data passes through the Internet every second than was stored in the entire Internet 20 years ago. We shall return to this when we discuss Big Data later.

C. Retrieval and Transmission

Maintaining a massive database is one aspect. Finding the information one wants, when needed and in a usable form is the another. The latter depends on the field of study and the nature of queries on the database. Databases are widely used for research, financial and administrative purposes. The following is an indicative list of disciplines that rely heavily on data analytics.

Government: demography, fiscal policy, social welfare

Defence: transportation, logistics, weaponry

Space: astronomy, rocket science

Science and Academia: physics, economics, computer technology

Medicine: biology, genetics, healthcare

Business: finance, commerce, industry, management

Agriculture and Environment: farming, climate, geology

The retrieval and transmission of data is a specialized discipline, in which intensive research and innovation are taking place. With every passing year, the boundaries are pushed farther.

D. Analysis and Interpretation

What is the bearing of observed evidence on the theories one wishes to evaluate? This is the crux of analysis. Statistical theory rests on a strong mathematical foundation. Concepts such as probability, random variables and distributions are at its core. While the tools of quantitative analysis are plentiful, the ability to use them effectively requires a mix of three disciplines - statistical theory, mathematics and computing techniques. Just as one does not need a course in automobile engineering in order to drive a car, so it is with the application of statistics. Even without a rigorous grasp of the theoretical intricacies, one is able to apply basic techniques using standard computer software.

Interpretation goes hand in hand with analysis. For this, it is helpful to have some knowledge of the field under investigation, be it finance, commerce or industry. Together with the scientific approach, a dose of common sense is mandatory. To give a trivial example: Scientifically, we classify tomato as a fruit. Common sense tells us not to put it in a fruit salad.

In the digital world, cognitive computing systems learn and interact naturally with people. They help experts make better decisions by penetrating the complexity of big data. This is the underlying system that Google successfully deploys to know what we mean when we search for something.

The Scientific Approach

We have touched upon four broad areas related to data management. The discussion is incomplete without reference to the scientific approach. Reasoning from observations has been important to scientific studies since the time of Aristotle, if not earlier. There are fairly well established methods to scientific discovery and the application of theories to practical problems. Scientific study is usually not a solitary pursuit. Therefore, if an individual's subjective perceptions can be evaluated and validated by other individuals, it increases the reliability of observational evidence. In other words, replications of an empirical experiment should yield comparable results.

Great care must be exercised by investigators to use both perceptual and non-perceptual evidence to evaluate data and to reject 'false' hypotheses. Fortunately, the scientific method has ways and means to do so. One can, eventually, make corrections and understand the significance of data that had not originally been salient to them.

Sometimes, the scientific approach gets sidelined by the human mindset. The most hard-headed businessman has, consciously or not, a philosophy inside him. Does he lean towards Empiricism which believes that all knowledge of facts derives from experience and the mind is not furnished with a set of concepts in advance? Is he a Rationalist who thinks that reason alone - independent of experience - is a source of knowledge? The first rejects reasoning which is not backed by factual evidence. The second places reason far above mere observational perceptions which are merely subjective experience. Notice that empiricism and rationalism seem to be firmly opposed to each other. A third philosophical approach could be Positivism, when he is solely concerned with positive facts and phenomena. He excludes speculation, but is not in a position to draw conclusions about causes and origins. If these three types of managers jointly attempt to study a set of business reports, there could be a colourful confrontation but no consensus.

Monte Carlo Methods

Consequent to the invention and proliferation of computers, our capabilities for number-crunching have grown apace. Not only that, inter-disciplinary collaboration has created synergies in unexpected ways. The development of the Monte Carlo methods is a dramatic example.

Monte Carlo experiments are a class of algorithms that use repeated random sampling to obtain numerical results. They are the only alternative when it is impossible to use other mathematical methods. Monte Carlo simulations are most useful for three distinct types of problems: optimization, numerical integration and generation of draws from a probability distribution.

A small digression to know how the method got its name. Its origin is in Monte Carlo, the place where European princes and nobility used to rub shoulders with

gamblers. Fortunes were made or lost in games of chance, which gave rise to the calculation of probabilities and odds for placing bets.

The period was the 1940's, when the World War II raged. It was a fortuitous combination of mathematics, statistical techniques and electronic computing applied to the study of neutron diffusion in nuclear physics, led by three brilliant scientists John Von Neumann, Stan Ulam and Roberto Fermi. The computers that they used were the historic ENIAC and MANIAC!

Today, statistical methods are being subjected to rigorous empirical scrutiny in the form of simulations, to understand their limitations and strengths. With the combination of powerful built-in statistical procedures and versatile capabilities of modern computers, Monte Carlo research is popular among quantitative researchers. In physics, the methods are useful to simulate systems such as fluids, disordered materials and cellular structures. Other examples include modeling phenomena with significant uncertainty in inputs such as the calculation of risk in business.

In any discipline, analysis and interpretation relies on quantitative methods. Some aspects of this topic are discussed in the author's papers in earlier issues of this journal. In what follows, we look at analytics in the context of huge volumes, high speeds and diversity of data.

Big Data

Data management is a highly specialized field wherein one sees the symbiosis of information technology, telecommunications and the internet. Business Intelligence (BI) is a set of techniques and tools that use state-of-the-art computer technology to transform raw data into meaningful and useful reports for business purposes. When the reports are ready, the analysts step in, to interpret and make sense of the reports. These two stages – report generation and analysis - remained distinct till the advent of Big Data Analytics (BDA).

BDA is the process of collecting, organizing and analyzing large sets of data to discover patterns and information for enabling an organization

- to better understand the information contained within the masses of data and

- to identify the most crucial data for present and future business decisions.

In short, businesses want the knowledge that comes from analyzing big data. One might ask, "So what's new? It is true that the big data movement, like BI before it, seeks to extract intelligence from data and translate that into business advantage. However, there are three key differences, as defined by Doug Laney:

1. Volumes: We generated more than 2 trillion GBs of information in 2012. Google is believed to have data storage capacity of 15 exabytes@ - the largest in the world.

2. Velocity: Billions of business transactions happen on the internet every day. The forecast is that the world's information now doubles every 18 months. World-wide, 5 billion people are using mobile phones 24/7 - calling, texting, tweeting and browsing.

3. Variety: In a medium-sized organisation, a variety of information - documents, presentations, databases, graphics, spreadsheets, e-mails, audio and video files - streams continually like channels of flowing water. Microsoft supposedly has a million servers to handle their diverse data operations across the globe.

Taken together, these three dimensions transform data management into an entirely new space, the space of Big Data. What will help make sense of this sort of information and provide a realistic look into what is happening?

With advances in computer technology, automated number crunching and complex algorithms to process information have become commonplace. Triggered by the three big V's, big data is another leap forward . The new trends are cloud services, cognitive computing and the Internet. Enterprise technology is morphing into one big, inter-linked ecosystem, wherein objects and people are connecting without human intervention.

The Challenges

Peter Drucker said "You can't manage what you can't measure." , which underlines the need for big data analytics in today's Information Age. The volumes are huge, the processes collaborative, the information diverse and the technology complex. Because of this, existing problems of data management have grown in complexity; and new problems have surfaced. For a business enterprise, big data presents three types of challenges:

- The first is the massive volume and the variety of data coming in many different formats - structured and unstructured - across the entire organization. These different types of data can be combined and analyzed in many ways, to search for patterns and other useful insights. The volume and variety is so large that traditional database and software methods fail to cope.
- The second challenge is to break down separate data silos to access all the data that an organization stores in different geographic locations and possibly in different systems.
- Thirdly, it is a challenge to create platforms that can pull in unstructured data as easily as structured data in a meaningful way.

Distributed Data Systems: A new challenge is to build software applications with distributed data systems, which are a collection of interconnected nodes that share digital data. Typically, a distributed data system comprises many networks, each having several servers and clients. The nature of the connectivity among them defines their inter-relationship. The nodes could be located geographically anywhere – that is, they are split across space and time. The CAP theorem states that in a distributed system only two out of the following three are guaranteed across a write/read pair:

1. Consistency - A read will return the most recent write for a given client.
2. Availability - A non-failing node will return a reasonable response within a reasonable time.

3. Partition Tolerance - The system will continue to function when network partitions occur.

In other words, one of them must be sacrificed. We know that unpredictable network failures or outages can and do occur. When the partitioned node fails to respond, the system must continue to function (partition tolerance). The software application has the option of sacrificing availability to maintain consistency or vice versa. This is the trade-off implicit to the CAP theorem. Without building such safeguards, the application is doomed to fail.

Scalability: Internet-based applications are particularly vulnerable to bottlenecks of scalability. For example, the Slashdot Effect. A sudden, temporary surge in traffic on a website occurs when a high-traffic source posts a popular high-interest story that directs large number of visitors to it. The unusual surge in traffic will slow down the site or make it impossible to reach. The problem here is that the scalability requirement is hard to predict in advance. The traffic can change instantly, based on sudden popularity. The website has a difficult choice: either over-invest in costly infrastructure, hoping for growth or under-invest and hope to quickly add capacity as and when required.

Computing Environment: Vendors are changing/updating platforms fast. User organisations have to face migrations to new platforms. Working on multiple data management platforms with multiple versions of each platform creates major management overheads and headaches for operational personnel. Data Recovery is an old problem which re-appears with a new twist in this computing environment. Increasing data volumes make the problem more difficult.

Inefficient use of Computer Resources: In a large organization, half the servers are only having at most 10% of their available resources used. Overall, capacity utilization is about one-third. This is due to high capacities installed with an eye on future developments and to the slow rate of skills up-gradation.

Lack of Trained People: Rapid changes in technology combined with expansion, create demand for more trained people than are available in the job market. We are witnessing the paradox of high unemployment and unmet demand for skilled

people. Skills shortage in an organization hampers productivity and holds back growth.

Decentralized Data: Your database may not be stored on your own computers, but on a cloud - a server located in an unknown location. You may not even own a database, but just subscribe to a data service. The advantages are evident: You need not maintain costly individual databases and you can pull timely information from the data service provider as and when you need it. As always, there is a downside: how is consistency, recoverability and availability of the information ensured? Which leads to the following two issues.

1. Auditing: Auditing a live database is not easy. There is increased focus on risk, accountability and avoidance of fraud. Most breaches occur by authorized users who are either negligent or malicious. Moreover, the auditing process adds a costly burden to the corporation and generates its own volume of data and reports.

2. Data Security: In an environment of global electronic distribution networks and cloud-based computing, data security has become a hydra-headed problem - think hacking, spyware and bots. One needs to handle threat prediction, detection, deterrence and prevention - a huge challenge and critical too. Big data analytics tools themselves are being harnessed for this purpose, with a combination of machine learning, text mining and modeling to provide integrated security.

Hadoop: Hadoop is one of the most talked about technologies. It offers solutions to many of the challenges we have discussed above. The USP is its ability to handle huge amounts of any kind of data at incredible speeds at low cost. With volumes and varieties of data growing each day, especially from social media and automated sensors, that is a key consideration which is attracting corporate giants to harness Hadoop technology.

The Way People Work

Remote Work: Communicating asynchronously is a hallmark of networking. This applies to people too. In recent years the percentage of man-hours of employees

working from home or anywhere outside the office has risen by 70 per cent in the United States. This upward trend is global and telecommuters are found in every emerging workplace. Undoubtedly, the way people work in the globalised environment has changed.

New Skills: Consider this scenario. You have arrived at a critical point in your project. and urgently need to interact with your team. The team members may not be online or situated in the same time zone as yours or are busy doing something else. So you communicate your concerns and updates via your project's online collaboration software: create a screen cast to describe your status, highlight a feature, raise a question in the team's chat room or send a global email. The team members will get to it and eventually respond in a similar manner, in their own time. To successfully remote work, you and your organization have to learn an entirely new set of personal, technical, and managerial skills.

Culture of Decision Making: The managerial challenges of using big data are real. No doubt that data-driven decisions can be better decisions. Will leaders acknowledge this? Senior people are accustomed to rely more on experience and intuition and not enough on data. This can change, provided project leaders and senior executives are genuinely interested in a big data transition. If you are a team leader faced with an important decision, adopt three simple techniques:

(1) Start by asking "What do the data say?". Follow up with specifics such as "Where did the data come from?" "What analyses was done?" and "How confident are you of the reports?". Your team will quickly get the message.

(2) Allow yourself to be overruled by the data. When your team members see that you accept that data have disproved a hunch, it is a powerful incentive to use analytics.

(3) Maximize cross-functional cooperation. People who understand the problems need access to the right data as well as the expertise of people who know problem-solving techniques that can effectively exploit them.

Cherry Picking: There are managers who shoot first, and then call whatever they hit, the target. Some managers like to spice up their reports with lots of data and impressive graphics to support decisions they had already made using the HiPPO (Highest Paid Person's Opinion) approach. Only afterwards are junior executives instructed to find the numbers that would justify the decision. With access to software tools and enormous databases, this sort of cherry picking is very easy, and at the same time, dangerous.

Causality: Managers unfamiliar with quantitative methods make genuine mistakes. For example, it is very easy to confuse correlation with causation. When two unrelated indicators happen to trend together over a finite time period, the analyst wrongly concludes that one causes the other. Similar pitfalls await the unwary manager who finds misleading patterns in the data.

The Benefits

Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right platforms in place, an enterprise can boost sales, increase efficiency, improve operations, customer service and risk management. The application of big data in industrial settings is driving a productivity revolution. GE recently announced that they gained around $45 billion in revenue last year through the use of analytics to automate processes, optimize performance, eliminate downtime, and predict when a machine or component will fail. It is being said that the industrial internet is bigger than industry. IDC predicts that the BDA market will reach $125 billion worldwide in 2015. Concerns of data integrity, privacy and security are going to become more significant. But the underlying trends, both in the technology and in the business pay-off are unmistakable.

The people who can reach into captured data and discover new knowledge are the ones who will be the most sought after; That is, those who can "provide you with information, discoveries, trends, factoids, patterns, visualizations and needles you didn't even know were in the haystack."

Big data enables corporations to use the collective knowledge hidden in their massive information systems to understand and serve their customers better. The customer is still king and business is Darwinian.

While companies are at the forefront of leveraging big data for big profits, big data can also expand the frontiers of knowledge in the social sciences, medical research and life sciences. Big data has the potential to change the way we live, work and play on planet earth.

Footnotes

1. @ 1 exabyte = 10**18 bytes
2. # For this quote and some of the big data challenges discussed, the author has drawn on the views of Tony Bain, an international expert in big data technologies.

* * *