Big Data—Governance and Compliance

The phrase* big data *is used to describe structured and unstructured data, significantly imposing in volume, which is too large to be processed by the traditional database and software tools and methods. Mostly, this data exceeds the capacity of data processing. Vendors use this terminology to refer to the technology which is needed to manage the data and storage. From a pure governance perspective, when the executive management decides to take all decisions on the basis of quantitative analysis and refuse to go by the gut feeling, it is said that the management has taken a big-data approach. The author in this article explores the concept of big data and tries to look at that from the perspective of governance and compliance. Read on…

"Big data is often described as extremely large datasets that have grown beyond the ability to be managed and analysed with traditional data-processing tools." This definition, given by a Mckinsey study, is relative in the sense that the data to be processed is larger than the capacity of the processing tools.

It is assumed that as the technology advances, the size of the datasets that qualify as big data will also increase. From a pure governance perspective, when the executive management decides to take all decisions on the basis of quantitative analysis and refuses to go by its gut feeling, we can say that it has followed the *big data approach*. So far, companies have been doing quantitative research in order to take their decisions. So, what's new then? Well, the answer lies in the fact that, due to data explosion, *i.e.*, data gets doubled every two years, an average executive has to sift through large datasets than that of the past to decide on things of similar nature. Normally, the present day top executives cannot afford to wait for the structured and completed information to take their decisions. French aviator, Antoine de Saint-Exupery rightly says,"*As for the future, your task is not to foresee it, but to enable it.*" Instead of waiting for the structured data, they opt for the unstructured one, since most of the data available is unstructured. It is established that enterprises that wait to take a perfect decision based on the structured information only, are at a disadvantage. At this point, we should take note of the Gartner's analysis which says that more than 90% of universal data have been created in the last two years and about 80 % of the enterprise data comprises unstructured data.

Ravikumar Ramachandran

(The author is an Account Security Officer at Hewlett-Packard India Sales Pvt. Ltd., Mumbai and he may be contacted at ravikumar1993@yahoo.co.in.)

Big Data and The Human Brain

It may be interesting to note that the human brain's storage capacity as per the recent research reports is 2.5 petabytes or 1 million gigabytes, which means that it has a capacity of a video recorder which will run for more than 300 years continuously.

Big Data and The Internet

The size of the internet, the world's largest library, can be estimated in *yottabytes* as on date, which would take 11 trillion years to download using the fastest internet connectivity. It had been estimated as 5 lakh TB in size as in 2003, and it has expanded by 20 lakh times in the last 10 years. Scientific philosopher, Karl Popper, has argued in his paper *Conjectures and Refutations: The growth of scientific knowledge* in 2002 that every scientific theory exists on insufficient premises and thus it can be challenged. He then noted,"*The Internet emphasises the depth of our ignorance because our knowledge can only be finite, while our ignorance must necessarily be infinite*."

The IDC's digital universe study states that, between 2009 and 2020, digital data will grow 44 fold to 35 zetta bytes per year. Refer to the following table:

- 1000 Bits = 1 Kilobyte
- 1000 Kilobytes = 1 Megabyte
- 1000 Megabytes = 1 Gigabyte
- 1000 Gigabytes = 1 Terabyte
- 1000 Terabytes = 1 Petabyte
- 1000 Petabytes = 1 Exabyte
- 1000 Exabytes = 1 Zettabyte
- ……..Yottabyte..Brontobyte…GEOPBYTE!!

There are two types of data: structured and unstructured. Structured data is any data capable of being entered in a data field and all the reports of MIS, Profit & Loss account and Balance Sheet are examples of structured data. Unstructured data is information that does not have any predefined data model and/or does not fit well into a relational database. This data typically is text heavy and includes dates, numbers and other details like that of audio, video, image, geospatial, click streams and log files. All new data entering into an organisation typically consists more of unstructured data (90-95%). Opportunity and challenges lie before an organisation that successfully mines the unstructured data and converts that into opportunities, and eventually wealth maximisation for its stakeholders.

There are two types of data: structured and unstructured. Structured data is any data capable of being entered in a data field and all the reports of MIS, Profit & Loss account and Balance Sheet are examples of structured data. Unstructured data is information that does not have any predefined data model and/or does not fit well into a relational database.

The three dimensions of Big Data are: volume, variety and velocity. **Volume** is the sheer size of the data, or the amount of transactions created every day. This is an obvious constraint for the organisation which cannot process huge data sets due to storage or computational processing limitations. **Variety** consists of structured and unstructured data. **Velocity** is the speed with which the data is created, accumulated, and processed. The increasing pace of the world has put demands on business to process information in real-time to arrive at real-time decisions.

Essential Facts for Businesses

1. Businesses cannot wait to take decision on the completed, processed and structured data.
2. They need to take decisions on unstructured data.
3. However, not all unstructured data are useful.
4. Business houses that ignore unstructured data are doomed.

Factors Enabling Big Data

1. Internet and digitisation of opinions and behaviour
2. Mobile computing
3. Social networking
4. Moore's law (power of microprocessor doubles itself in every 18 months)
5. Cloud computing

Data explosion and Knowledge Management

Data explodes every two years. On a reasonable assumption that huge intellectual knowledge is created, what happens to the proprietary knowledge of the big corporations that spend a lot of money on research and development? It gets diluted and until the organisation continues with its rigours of research, it will find itself obsolete and would lose its relevance. So, as a consequence of big data, innovation and new discovery will pay a great role in the coming years.

Effect of Big Data on Data Processing

1. Data needs to be stored in the system in which hardware is infinitely scalable.
2. Storage and network cannot be a bottleneck.
3. Data must be processed into business intelligence where it is.
4. Move the code to the data and not the other way.
5. Data sits in one place; never move that around.

Challenges in Protection of Big Data

1. There is a risk of permanent loss of data stored in monitoring devices, *e.g.* surveillance cameras.
2. Too much of logs generated needs to be analysed in real time, or else it is of no use and the storage poses a problem.
3. Because of the uniqueness of data, deduplication cannot be applied, resulting in constraints on storage.
4. Huge CPU processing power is required to process large files.
5. No good back-up solution is available.
6. RDBMS is not suitable to handle big data.
7. HIPAA and PCI compliance becomes difficult.
8. It is a very risky business in the medical industry.

Enterprise Governance and Big Data

1. *Strategic Alignment:* Identify the business priorities and define the problems to be solved within a specified time frame with measurable and achievable outcomes.
2. *Management Sponsorship:* Management should be willing to go for fact-based decision making. Identify champions for consumption of data analytics and ensure that benefits realisation happens from various reports and statistical models.
3. *Appropriate Human Resources:* Human resources should be well-skilled in data analytics and inferences. Good knowledge of mathematics, business and technology is a must.
4. *Key Performance Indicators:* Ensure business effectively uses analytics to make better business decisions. Ensure that investment is made in right type of analytics and it happens in the right type of people, process and technology.

Protection of Big Data

Security for big data is not an easy task. One needs to have balance between access, availability, performance and liability from unwanted disclosure.

1. *Keep only needed data:* Now, what is needed data? It is tough to ascertain that in practical terms. For example, logs, if we keep it and analyse it, we get to know the performance of large data systems in terms of scale, use and efficiency. At the same time, it poses a risk of disclosure. Deleting the logs eliminates the risk, but we also lose the benefits of big data analytics. A balance needs to be taken between both the options
2. *Classifying data:* This is one obvious solution. Classify data in terms of sensitivity, high importance and low importance. Sensitivity is related to privacy and other confidential data. High importance is data to be analysed in a smaller duration as it may lose importance as the time elapses. The example is number of website hits and the demography of customers taken from cookies. Rest of the data can be classified as low importance. Typically, sensitive data is long-term, while other two are short-term, meaning it can be considered for deletion, after the analysis is done.
3. *Backing up data:* Big data typically contains unique data, *i.e.,* data taken from monitoring devices like that of traffic movement, soil pH, *etc.,* which are accumulated frequently and in real time. All data are unique to the moment, and if they are lost, it is impossible to recreate them. Finding a good back-up solution may be challenging, but with evolution and advancement of technology, something can be expected soon.
4. *Big data and compliance:* Big data is not easily handled by relational databases, hence the traditional way of managing data is not possible here and it may be difficult to understand how compliance affects the data. To put it simply, relational databases took time to give you inputs for decisions, but the inputs were accurate. Now in the realm of big data, horizontal and unstructured databases which are good at solving business problems have arrived. Here, inputs for decision-making are faster, but not accurate. But, we need to comply with the law irrespective of the data methodology. ■

On a reasonable assumption that huge intellectual knowledge is created, what happens to the proprietary knowledge of the big corporations that spend a lot of money on research and development? It gets diluted and until the organisation continues with its rigours of research, it will find itself obsolete and would lose its relevance.