LOG206: E-Business

Spring 2018

Notes

Module 8: Introduction to Big Data Analytics

Data, Information, Knowledge and Decision making

Business owners and managers make decisions on a daily basis, addressing everything from day-to-day operational issues to long-range strategic planning. Indeed, decision making is the bread and butter of managers and executives, who make about three billion decisions each year. Bain researchers found that decision effectiveness is 95% correlated with financial performance. In order to make good decisions managers require knowledge.

Knowledge

Knowledge is what we know. Think of this as the map of the World we build inside our brains. Like a physical map, it helps us know where things are – but it contains more than that. It also contains our beliefs and expectations. "If I do this, I will probably get that." Crucially, the brain links all these things together into a giant network of ideas, memories, predictions, beliefs, etc. It is from this "map" that we base our decisions, not the real world itself. Our brains constantly update this map from the signals coming through our eyes, ears, nose, mouth and skin. Knowledge comes from information and information comes from data.

Data

Data is a collection of figures and facts, and is raw, unprocessed, and unorganized. The Latin root of the word "data" means "something given", which is a good way to look at it. Individuals and organizations can't do much with unprocessed data because it's so random. Once data is given structure, organized in a cohesive way, and is able to be interpreted or communicated, it becomes information. In other words, data is/are the facts of the World, for example, take yourself. You may be 5.5ft tall, have brown hair and blue eyes. All of this is "data". You have brown hair whether this is written down somewhere or not. In many ways, data can be thought of as a description of the World. We can perceive this data with our senses, and then the brain can process this.Human beings have used data as long as we have existed to form knowledge of the world.

Information

Information is not just data that's been neatly filed away, it has to be ordered in a way that gives meaning and context. This is what allows people to use data for reasoning, calculations, and other processes. With that said, data's importance lies in the fact that it's a building block. Without it, information cannot be created.

In summary, data has no meaning until it's turned into information. In order for people to interpret data or make any use of it, it must be understood. For instance, a company's sales figure for one month is a piece of data that's meaningless because it has no context. It tells nothing, and there's little that anyone can do with it as is. However, you take a business's sales figures from three months and average that number, we'd be able to derive many bits of information from that data. When one has incomplete data, it's highly likely that it will be misinterpreted and lead to the development of misinformation. For example, suppose someone saw that his business's sales were up by 4%, and he drew the conclusion that his current marketing campaign was working well. However, if he found out that a competitor who sold the same products had a sales increase of 16% during the same time period, he'd start to question just how well his campaign really performed and would want to gather more facts (data) to analyze the situation again.

Big Data

We're living in an era of massive data generated by web transactions, our mobile devices, social media and even our refrigerators and cars. Each and every one of us is constantly producing and releasing data about ourselves. We do this either by moving around passively -- our behaviour being registered by cameras or card usage -- or by logging onto our PCs and surfing the net.

Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

Big data would typically be too expensive to store, manage, and analyze using traditional database systems. Usually, such systems are cost-inefficient because of their inflexibility for storing unstructured data (such as images, text, and video), accommodating "high-velocity" (real-time) data, or scaling to support very large (petabyte-scale) data volumes.

The Three Primary Sources of Big Data

Human generated data is comprised of the emails, Word documents, spreadsheets, presentations, images, audio, and video files that we create and share with other people every day. There is massive explosion of human generated data in the enterprise. It is one of the fastest-growing data sets, one of the most valuable and the most relevant for digital collaboration, mobile or otherwise. This kind of data provides invaluable insights into consumer behavior and sentiment and can be enormously influential in marketing analytics. The public web is another good source of social data, and tools like Google Trends can be used to good effect to increase the volume of big data.

Machine generated data is data that is automatically generated by a computer process, application, or other mechanism without the active intervention of a human. It includes data generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behavior. This type of data is expected to grow exponentially as the internet of things grows ever more pervasive and expands around the world. Sensors such as medical

devices, smart meters, road cameras, satellites, games and the rapidly growing Internet Of Things will deliver high velocity, value, volume and variety of data in the very near future.

Business generated data is data generated from all the daily transactions that take place in organizations both online and offline. Invoices, payment orders, storage records, delivery receipts – all are characterized as transactional data yet data alone is almost meaningless, and most organizations struggle to make sense of the data that they are generating and how it can be put to good use. For example UPS, delivers 16 million shipments per day. They get around 40 million tracking requests. An estimate of how much data UPS has on its operations is 16 petabytes. Walmart is a big organization that gets 250 million customers in 10,000 stores. They collect data on Twitter tweets, local events, local weather, in-store purchases, online clicks and many other sales, customer and product related data. In total, Walmart collects 2.5 petabytes of data per hour.

Characteristics of Big data

Volume: Volume is probably the best known characteristic of big data; this is no surprise, considering more than 90 percent of all today's data was created in the past couple of years. The current amount of data can actually be quite staggering. Here are some examples: 300 hours of video are uploaded to YouTube every minute. An estimated 1.1 trillion photos were taken in 2016. Therefore, Big data implies enormous volumes of data. It used to be employees created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive

Velocity: Velocity refers to the speed at which data is being generated, produced, created, or refreshed. For example Facebook claims 600 terabytes of incoming data per day; Google alone processes on average more than "40,000 search queries every second," which roughly translates to more than 3.5 billion searches per day. I other words, Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI if you are able to handle the velocity.

Variety: Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

Veracity: Big Data veracity refers to the biases, noise and abnormality in data. Although there is widespread agreement about the potential value of Big Data, the data is virtually worthless if it’s not accurate. What’s crucial to understanding Big Data is the messy, noisy

nature of it, and the amount of work that goes in to producing an accurate dataset before analysis can even begin.
Veracity is one of the unfortunate characteristics of big data. As any or all of the above properties increase, the veracity (confidence or trust in the data) drops.

Valence: Valence simply means connectedness. The more connected data is, the higher it's valences. Data items are often directly connected to one another. As there are more and more connections among the data the complexity of the analysis increases.

The Sixth V, Value
The five characteristics described above are considered dimensions of big data. However, at the heart of the big data challenge is turning all of the other dimensions into truly useful business value. The idea behind processing all this big data in the first place is to bring value to the problem at hand. Value is the ability to convert Big Data information into a financial reward. For example, if you find a relationship between two products at a point of sale, you can recommend them to customers at a website or put the products next to each in a store.

Categories Of 'Big Data'

Big data' could be found in three forms: structured, semi-structured data, and unstructured data.

Structured
Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, now days, we are foreseeing issues when size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabyte. Data stored in a relational database management system is one example of a 'structured' data.

The data that has a structure and is well organized either in the form of tables or in some other way and can be easily operated is known as structured data. Searching and accessing information from such type of data is very easy. Structured data represent only 5 to 10% of all informatics data.

Semi-Structured Data
Semi-structured data is information that doesn’t reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process you can store them in relation database (it could be very hard for somme kind of semi structured data), but the semi structure exist to ease space, clarity or compute…

Examples of semi-structured: CSV but XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured. As structured data, semi structured data represents a few parts of data (5 to 10%) so the last data type is the strong one : unstructured data.

Unstructured Data

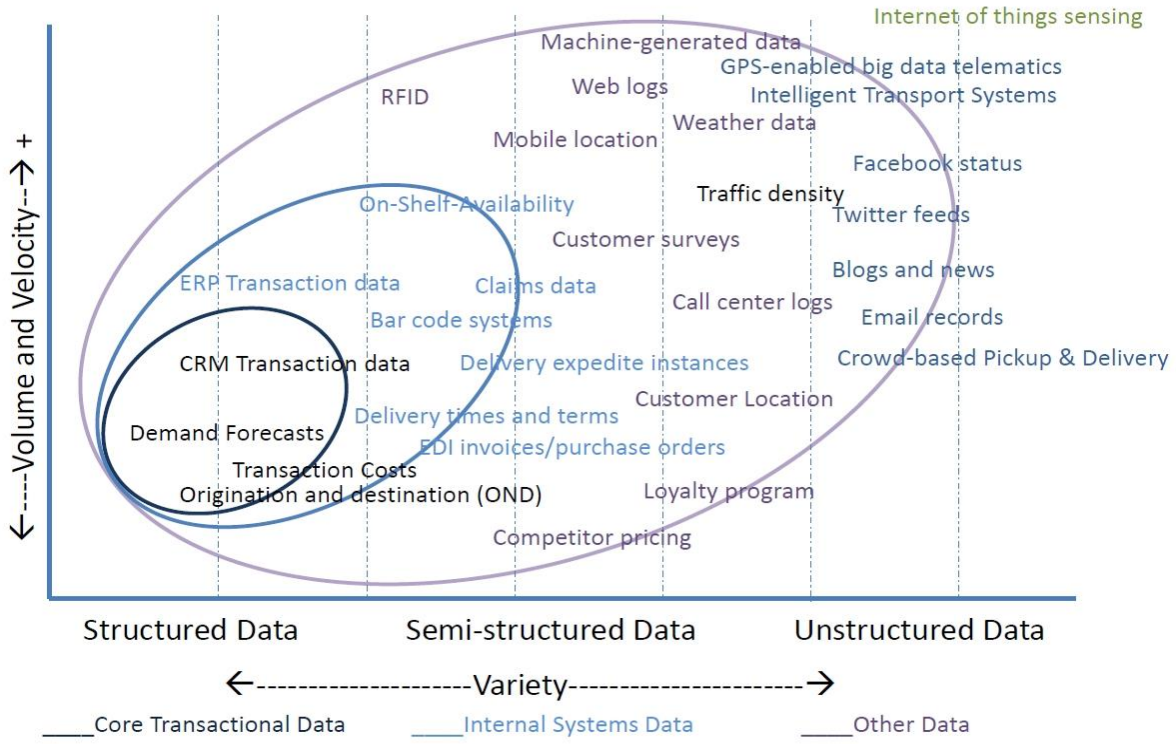
The data that is unstructured or unorganized. Operating such type of data becomes difficult and requires advance tools and softwares to access information.
Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos etc. Now a day organizations have wealth of data available with them but unfortunately, they do not know how to derive value out of it since this data is in its raw form or unstructured format. Examples Of Un-structured Data: Output returned by 'Google Search'

Unstructured data represent around 80% of data. It often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered « unstructured » because the data they contain does not fit neatly in a database. Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.

Here are some examples of machine-generated unstructured data:

- Satellite images: This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
- Scientific data: This includes seismic imagery, atmospheric data, and high-energy physics.
- Photographs and video: This includes security, surveillance, and traffic video.
- Radar or sonar data: This includes vehicular, meteorological, and oceanographic seismic profiles.
- The following list shows a few examples of human-generated unstructured data:
- Text internal to your company: Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
- Social media data: This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- Mobile data: This includes data such as text messages and location information.
- Website content: This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

The list goes on. The unstructured data growing quickiest than the other, and their exploitation could help in business decision.



Types of analytics

Descriptive analytics: Descriptive analytics answers the question of what happened. For instance, a healthcare provider will learn how many patients were hospitalized last month; a retailer – the average weekly sales volume; a manufacturer – a rate of the products returned for a past month, etc.
Descriptive analytics juggles raw data from multiple data sources to give valuable insights into the past. However, these findings simply signal that something is wrong or right, without explaining why. For this reason, highly data-driven companies do not content themselves with descriptive analytics only, and prefer combining it with other types of data analytics.

Diagnostic analytics: At this stage, historical data can be measured against other data to answer the question of why something happened. Thanks to diagnostic analytics, there is a possibility to drill down, to find out dependencies and to identify patterns. Companies go for diagnostic analytics, as it gives a deep insight into a particular problem. At the same time, a company should have detailed information at their disposal otherwise, data collection may turn out to be individual for every issue and time-consuming.

Predictive: Predictive analytics tells what is likely to happen.. Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known

attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. For example, Predictive models analyze past performance to assess how likely a customer is to exhibit a specific behavior in the future. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision.

Thanks to predictive analytics and the proactive approach it enables, a telecom company, for instance, can identify the subscribers who are most likely to reduce their spend, and trigger targeted marketing activities to remediate; a management team can weigh the risks of investing in their company's expansion based on cash flow analysis and forecasting.

Prescriptive analytics: Prescriptive analytics represent the highest form of business analytics. It takes the logical step of switching from reactive analysis to proactive action. The purpose of prescriptive analytics is to literally prescribe what action to take to eliminate a future problem or take full advantage of a promising trend. An example of prescriptive analytics from our project portfolio: a multinational company was able to identify opportunities for repeat purchases based on customer analytics and sales history.

This state-of-the-art type of data analytics requires not only historical data, but also external information due to the nature of statistical algorithms. Besides, prescriptive analytics uses sophisticated tools and technologies, like machine learning, business rules and algorithms, which makes it sophisticated to implement and manage. That is why, before deciding to adopt prescriptive analytics, a company should compare required efforts vs. an expected benefit.

In sum, prescriptive analytics tools help business leaders determine the best course of action to achieve specific business objectives. Prescriptive analytics provide business leaders with specific recommended actions, whereas other forms of analytics provide information.

Machine learning

"Learning" means that the algorithms analyze sets of data to look for patterns and/or correlations that result in insights. Those insights can become deeper and more accurate as the algorithms analyze new data sets. The models created and continuously updated by machine learning can be used as input to decision logic or to improve the decision logic automatically.

Supervised Learning

Here the human expert acts as the teacher where he/she feeds the computer with training data containing the input/predictors and show it the correct answers (output) and from the data, the computer should be able to learn the patterns. Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features such that we can predict the output values for new data based on those relationships that it learned from the previous data sets.

Unsupervised Learning

The computer is trained with unlabeled data. Here there's no teacher at all, actually the computer might be able to teach you new things after it learns patterns in data, these algorithms a particularly useful in cases where the human expert doesn't know what to look for in the data.

Thus, unsupervised learning occurs when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of uncorrelated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.

This is mainly used in pattern detection and descriptive modeling. However, there are no output categories or labels here based on which the algorithm can try to model relationships. These algorithms try to use techniques on the input data to mine for rules, detect patterns, and summarize and group the data points that help in deriving meaningful insights and describe the data better to the users.

As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

Dashboarding

Ever-increasing volumes of information are captured and stored across myriad systems and in hundreds of formats. This information is vital for a business; but providing a single source of truth is difficult.

A dashboard is a page that provides a single view of a component or complete business. With interactive visualisations that link to each other, users are able to uncover behaviours and insights in their data. A dashboard is a simple-to-consume format for information extraction; can be shareable and modular; allowing for functionality and advanced analytics to be added as the business need requires.

Dashboards can be utilised by day-to-day managers, to provide visibility over people, projects or reportable measures. Dashboards also provide a simple-to-consume format for executive level decision makers – providing an interactive overview of their business, in real time, on any device, anywhere in the world.

Figure 2: An example of a dashboard

Advantages of Dashboards

- Dashboards are valuable because they transform business data into critical information that jumps out to the user, who can then make sense and act on it immediately.

- Fast and effective decision-making – Gives executives, managers and analysts convenient immediate access to key performance metrics, which help them monitor performance and processes for a greater understanding of the business.

- On demand, accurate and relevant information in line with business priorities – Dashboards clearly communicate business objectives throughout the organization and allow users to see progress towards those goals. This keeps everyone focused and informed. With a personalized layout, users only see the information that is most important to them, and they can filter out information that is not relevant.

- Focused identification of problems, inefficiencies or negative trends for immediate action and improved performance Users can immediately see any problems and drill down on charts and links to explore detailed information and analyze data in real time, to determine root causes and to correct negative trends.

NB: See slides for examples of how companies benefit from Big Data