2.2: Transcript: Lecture 1: Data Characteristics of Big Data

As you already know, the Big data is
Large Volume of Data with Varying degrees of formats, complexities, and ambiguities (Variety), generated at different Velocities.
This cannot be processed using traditional technologies, methods, algorithms or any off-theshelf solutions.
You also know the big data technologies
Provide the ability to access a large volume of data to gain critical and useful insights The learning process is machine managed with minimum human intervention; hence the analysis is simpler and error-free.
So why is it in huge demand now? We can say,
Current technologies along with new data processing frameworks and platforms like NoSQL, Hadoop provide cost-effective and scalable solutions in comparison to the traditions data management platforms.
Data, algorithms and methods have already existed with the traditional system**,** but scalability and flexibility of the processing architecture was a limitation.
Traditional system supported the data analysis with a lot of human processing and analytic refinement
Big Data introduced the automated data processing capability, which is extremely fast, scalable and has flexible processing.

Each organisation will have its own data requirements for big data processing. Here are some examples: Weather Data:
Lots of weather data reported by governments agencies around the world (met office UK), scientific organisations (World Meteorological Organisation –WMO a UN organisation), and consumers like farmers. Television channels or radio channels need Key Performance Indicators of temperatures and forecasted conditions based on several factors Clinical Trials Data:
Pharmaceutical companies wanted to minimise the life cycle of processing for clinical trials data and manage the same with rules-based processing
“New vaccines” are in urgent need especially in a situation similar to COVID 19, and also generally in the society, in which economic growth, globalization, and immigration are leading to the emergence/re-emergence of old and new infectious agents at the animal-human interface. Big data technologies could play a key role in moving toward a tailored and personalized vaccine design and administration Epidemiology:
Big data could play a vital role in providing the analytical capacity while dealing with the incidence, distribution, and possible control of diseases.
Data analysis could help to find other factors relating to health for shaping the policy decisions as well as an evidence-based practise by identifying risk factors for disease and targets for preventive healthcare. Appropriate and accurate Data analysis could become the cornerstone of public health, however, there could be many pitfalls which should be carefully prevented from happening.

A very simple example:
A fast-food restaurant chain wants to know the correlation between its sales and consumer traffic based on weather conditions.
Historic pattern and current pattern need to be integrated and analysed, social media sharing of consumer perspectives about the shopping experience and where the weather is mentioned is a key factor.

Here is a possible set of sample data for example in the previous slide, covering weather, customer sentiment, product competition, location and campaign details etc. highlighting the features, source and complexity.
Pause the video and examine the details carefully.

After determining various sample data sets for analysis in previous slide, they need to integrate as a large volume of data obtained from different sources and in varied format.
The company observing the correlation between weather and food sales can answer the following types of questions more effectively

What sales occurred across the US for a given day/week/month/quarter/year, and under various weather conditions?
Did people prefer drive-thru in extreme weather conditions irrespective of the geography? Did restaurants along the highway get more traffic in drive-thru during regular versus abnormal weather?
Did service interruptions occur due to weather?
What is the tendency for business impact in abnormal weather?
Does the restaurant need to staff more in different weather conditions? What is the budget impact in such situations?
Does coffee sell more than burgers in the winter in Boston, and during the same time, do more cold beverages sell in Orlando?
What are the customer expectations of pricing? Do they give feedback by phone, email or social media?
What is the competition comparison by customers in social media?

When all of the data listed in the previous slide is integrated and processed with appropriate business intelligent systems, it has provided the fast food centre with a better understanding of the following subject areas.
Customers
Markets
Products
Vendors/Suppliers
Contracts

Staff Management
Campaign
Location Management
The organisation can get a better insight into what drives the business, and how the weather can affect (power of the weather) the running the business, how it affects the consumers in general and what is the impact of that on the business.

AS you can see from the example, data is the valuable raw material for producing meaningful information. Data Growth is characterised by Volume (amount of Data), Velocity (speed of data in and out), and Variety (range of data type and sources). Innovation in technologies transformed the way we engage in business and provide services and the associated value and profitability. The business model transformation has a key role in data growth.
Organisations are now service oriented rather than product oriented.
The success of the organisation is measured by customer satisfaction (service and effectiveness), not by the usefulness of the product. Fundamental data required for the business remains the same, however huge amount of supporting data is generated through various mediums which need to be captured and analysed (Volume)
Globalisation: key trend, changed the variety and the format of the data. (Variety)
Personalisation of services: Businesses are measured by the personalisation of their services. (Velocity)

Data Volume is characterised by the amount of data that is generated continuously. Data can be generated from various internal and external sources.
Machine Data:
Steady Patterns of numbers and Text occurs in a rapid fire fashion.
Examples
Sensor Data: Sensors in the building to control the heating and cooling, sensors in the automobile. Similar structure but values depends on many factors. For example Robotic Arm in an assembly line sending signals for every movement. Application Log: Different devices generated application log
Click Stream Data: Clickstream log from internet portals and sites.
External of third party data:

Data feed collected from external parties, for example, weather data
Emails generated by its employees, customers and others
Data could be generated and collected form Contracts
Different type of contracts generated by the company, classified as Human resources, legal, vendor, supplier, customer etc.

Other contributing factors to increase the data volume, especially with the advances in the digital technologies are:
Geo-Spatial Data Mobile devices use GPS for navigation and personalisation.
GPS added to the camera's and camcorders to set the location of the picture was taken along with the date and other information.
Social media:
The amount of data generated by the customers every minute provide extremely important insights into choices, trends, competitions, opinions, influences, connections, brand loyalty, brand management and much more.
Companies must make use of the available data Different format, large volume, increased velocity etc.….

Velocity is the speed and direction of motion of an (physical) object.
With current requirements data needs to be processed in a continuous stream, not as a batch.
For example, Amazon, Facebook, Yahoo, Google etc. operates by tracking the customer clicks and
navigation on the websites and providing personalised browsing experience and millions of clicks every second resulting in a large amount of data.
The speed of the data generated by the user clicks on any website is a good example for Big Data Velocity.

As you can see from the examples listed in the slide, the data source like sensor data, mobile networks, and social media creates data in huge volume with high velocity.
Data velocity along with the data volume will be tricky for RDBMS, creating silos of solutions to process the data type means scalability will be an issue.

Big Data comes in different data formats Emails, Tweets, Social Media, Sensor Data, Video, Audio….
Every form of data is important
Availability of appropriate metadata to identify what is in the actual data is very critical.
Absence of metadata or partial metadata means processing delays

So far, we have discussed in detail about the three V's associated with Big Data,
Volume, Velocity and Variety. This can be extended with Vagueness:
Ambiguity, lack of metadata for example in a photograph or graph M and F could mean Male/Female, or Monday/Friday
Virality:
How quickly the data is shared between people-to-people through a peer network. Rate of spread is measure in time, re-tweets. But understanding the context is important…
Viscosity:
Measures the resistance (slow down) to flow in the volume of the data.
Resistance can manifest in data flows, business rules or even be the limitation of the technology.

Further studies suggest there are many more of Vs big data. Investigate the Vs associated with Big data. Get involved in the discussion forum asking the question, which Vs are the most important? Post your finding and comment on your peers' view with justification.

In Summary, we have seen that the recent advances in digital technologies and data collection led to data science and the emergence of big data. Big data and big data analytics unquestionably have the potential to improve our lives. However, many of them have valid concerns too.
How do Facebook and other social media platforms target advertising towards you? Is this something you welcome or not? As we go through this module, we will be asking these questions quite often.