## 2.4: Lecture 2: Transcript - Data Structures and Data Analytics

Big data is Large Volume of Data with Varying degree of formats, complexities, and ambiguities (Variety), generated at different Velocities, which cannot be processed using traditional technologies, methods, algorithms or any off-the-shelf solutions.
The data source may include machine-generated data from sensor networks, clickstream data, scanning devices, machine log, aeroplane engines and consumer-driven data from social media, transactional data etc.

Traditional IT Systems were designed to generate and manage structured data. Created with a well-defined data model and every data handled by the system is expected to fit within the predefined structure.
The system will support Predefined functionalities, closely related to the day to day running of the organisation, also known as transaction processing system. Most of the systems were created to move the organisational activities from a paper-based system to a computerised system.
Modern / 21st Century systems are aimed to handle unstructured data. They don't have any predefined format. It could be text, image, audio, video, or even sensor data.
The systems must deal with human generated data like documents, email, eBook, logs, blogs, social n/w etc. as well as store and interpret machine generated data like sensor data, satellite images, radar data, surveillance data etc.

Data, in general, can be classified as Structured data, semi-structured data and unstructured data.
As you have seen in the previous slide the traditional IT systems support structured Data, where the data has a well-defined structure and most often stored in a database.
The data is designed to support various organisational functionalities and often managed by Structured Query Language.
Customer details, order details etc. can be examples for structured data. Every customer will have a Customer ID, Name, Address, Phone number etc.
Order Details will include Order ID, Customer ID, Order date, Total etc.
Semi-Structured Data has some structure and share common characteristics, but do not have a rigid structure like a tabular form in a relational database, hence it is not feasible to store semistructured data in a database.
The advantage of the semi-structured data is, to a certain extent it is flexible and adaptable and can accommodate variations in the structure.
CSV files, NoSQL documents, HTML, XML etc. are some examples. Unstructured Data, most of the Big Data resource falls into this category.
As the name suggests un-structured data has no predefined format. It can take any type of data and has no constraints imposed by the data architecture. So, it is quite obvious that it cannot be stored in a database.
Any data without the structure can be an example, hence the list will be very long and will grow all the time.

Just get a feel of what can be considered here is a brief list of unstructured data, image, videos, clickstream data, sensor data, log files etc.

The diagram shown on the slide provides some basics comparison between structured and unstructured data. Please pause the video and have a read.
Slide 6:The diagram shown on the slide provides some basics comparison between structured and unstructured data. Structured data v/s Unstructured data in terms of appearance, general data types, the quantity of data generated, storage requirements and management issues etc.

The details on this slide are also very similar to the previous slide, comparing the two on slightly different factors as well as examples for application and data itself. Please pause the video and read the details.
This diagram provides a comparison between the structured data and unstructured data in terms of its characteristics, where is it stored, how it is generated, applications of them along with some examples.
Here you can see why the unstructured is fall under the big data umbrella.

Unstructured data is easily generated and can also be easily stored (volume, velocity and variety).
It is certainly a large repository of hidden information.
However, processing the unstructured data is much more complex in comparison to structured or even semi-structured data.
Using appropriate tools and algorithms will help to harvest the power of unstructured data.
Some of the techniques that are used in the data analysis are Natural Language Processing, Pattern Sensing, Clustering and Classification, Text-mining, Sentiment Analysis etc.
Processing may also require imposing some structure on the unstructured data. The following slide will explain a simple example.

For example, processing free text to get retrieve some information may involve imposing a structure to the free text.
To accomplish this, the first step is translating the text into the preferred language and parse the text into sentences which is the fundamental unit of a natural language.
Then extract and normalise the conceptual terms from the text, normalise could also be interpreted as minimise.
Map the identified terms into standard nomenclature and then annotate or tag them.
Now the data will have some descriptions and values.
The next step is associating the data into an appropriate classification.
The classified data can then be assigned to a data storage and retrieval system. Finally indexing the data item in the system. Think about this in large-scale, in relation to Big data resources and big data analytics.

The example in previous slide showed that each data type requires different and specific types of processing or analyses by big data analytics in order to turn data into usable information. By further analysis, this process can possibly continue to convert information into Knowledge or Wisdom (small data) described as a “knowledge pyramid”. This is in turn true for those data analytics that provide visualisation techniques for various types of data, i.e., each data type requires different and specific types of visualization. For example, visualisation for numerical data are different than social network. What big data brings is the ability to process, analyse and visualize all types of data, in their original form, by integrating new analysis methods and new ways of creative working resulted from better data visualization and capabilities to find patterns in complex data.

As another example, in the past, data were usually defined in a quantitative form (i.e., data being values that describe a measurable quantity) and qualitative form (i.e., describing of qualities or characteristics). These types of data will be considered as structured data because they require a simple transformation before issuing their meaning. But this is absolutely not the same case when we are faced with unstructured or semi structured data (web page, pdf, video, geolocation or sensors, etc.). We may be wondering what this has to do with the types of data that were previously presented. In fact, by analysing such data, we apply a treatment aimed to reduce data (i.e., data reduction) inside the computer to a sequence of numbers that can be interpreted for example by a “machine learning algorithm”. Thus, it is easy to understand the need for powerful algorithms to process these types of complex data.

You have to distinguish data analytics from data analysis: Data analysis as subset of the processes in data analytics, which brings the extensive use of data to make decisions and resultant actions. Turning data into information and then turning that information into knowledge which is the target paradigm of “knowledge discovery”, remains a key factor for business success. Information is a message or relationship amongst data with a higher level of meaning than data. Knowledge has a wider and deeper meaning than data or information. It is created from the use, analysis and productive utilization of data through a cognitive or intellectual operation of data analytics. Ackoff (1989) defines data as symbols, information as data that are processed to be useful and knowledge as the application of data and information in order to have the ability to understand why and how. This idea can be seen in the knowledge pyramid which illustrates that data are essential to build knowledge in order to make a good decision and generate value.