**Big Data, Data Types and Data Formats – U1 Lecturecast**

By simple definition, big data grows exponentially and is boundless (volume), has varied formats (variety), and has high complexity (voracity and velocity).

Big data has grown so large and in complexity that conventional tech can no longer handle **Ohlhurst 2012**  
  
The fusion of the real and virtual worlds has been the core driver of Industry 4.0. Big data is central to this emerging phenomena, digital evolution, and revolution. Big data emerges from a variety of sources comprising of IOTs, factories, enterprise resource planning and customer relations management systems, according to industry experts.  
  
For cleaning, exploring, creating, optimising and evaluating big data, the following are essential in the big data management process.  
  
Method - Cleansing, Standardization, Formatting, Normalisation  
Tools - Python, SQl, web services, GUI , security, SPSs  
Techniques - as methods  
Strategy - data refining, cleaning, machine learning, data modelling, statistical (SPSS)  
  
Varied data - text, video, photos, meta data, html  
  
Structured - highly organised EMG relational database  
  
semi-structured data include: being partially organised and not fully organised;  
not fully conforming with data storage formats and standards;  
being delimited in form, thereby making the process of retrieving data more difficult.  
Examples include HTML, XML and other markup languages.  
  
Quasi structured data - This involves specifically textual data that has a temporal state, having erratic data formats. Sometimes data transforms to a temporal state due to its continuous interaction with other systems. An example could be data streams on a social media page or clickstream data from Google searches  
  
Unstructured - poorly formatted, loosely joined, varied formats - XML, JSON, CSV? Social media posts an example  
  
CSV is a data exchange format that is widely used in business to hold tabular data and transfer data between applications  
  
JavaScript Object Notation, or JSON, is a simple data-interchange format that is easy for machines to parse and generate, and holds data in key and value pairs separated by a colon. This data format is one of the most commonly used formats for data transfers. It is preferred because it is clean, easy to read, and easy to parse. Many websites have JSON-enabled APIs.  
  
Markup languages have specially formatted text using tags and nodes, and include XML (or Extensible Markup Language) and HTML (or Hypertext Markup Language). They use tags (as readable strings) to control the document structure for holding and manipulating data. XML allows data exchange quickly between applications and platforms, but is not meant for presenting the data, while HTML displays data and is used widely for web pages.  
  
Pdf, highly human readable but hard to extract. Require algorithms.  
  
Xlsx - human readable, unstructured. More complex libraries in python needed to read.  
  
Before attempting to handle PDF, Excel and other hard-to-parse formats, check if other raw formats are available. The rawer the format (that is machine-readable), the more likely it is to be accurate and easy to parse with code.  
  
Consider how other corporates interested in the same dataset as you have approached their data management implementation. If Python is the language for the data management application, then identifying the relevant tools and libraries is an essential first step.  
  
Ultimately, consider if the dataset can be converted into simpler machine-readable formats.  
  
The data is available in PDF, Excel and text formats, and the files are included in the textbook repository downloaded in Unit 2 (see under chapters 4 and 5).  
  
The example Python codes mentioned below are included on pages 76-77 in the Kazil textbook.  
  
Python codes mentioned below are included on pages 115-123 in the Kazil textbook.