UNIT - I

Types of Digital Data

1) Structured Data
2) Semi structured data
3) Unstructured data

Structured Sources of structured data

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.

Characteristics of Structured Data

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. Anything that couldn't fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

Source of structured data :

1) What is meta-data? (Australian National Data Service)
2) Library Catalogues (date, author, place, subject, etc)
3) Census records (birth, income, employment, place etc.)
4) Federal and State Hansard
5) Legal records: Old Bailey Online (1674-1913)

6) Economic data (GDP, PPI, ASX etc.)
7) FaceBook like button (big-data collection!)
8) Phone numbers (and the phone book)
9) Databases (structuring fields)
10) XML-TEI (bringing structure to the text through tagging particular elements like versions of the word ”canal’ in 17th C Dutch.

Ease with Structured data

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. Anything that couldn't fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

Semi-Structured

Semi-structured data is data that has not been organized into a specialized repository, such as a database, but that nevertheless has associated information, such as metadata, that makes it more amenable to processing than raw data

sources of semi-structured data

CSV but XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured.

But 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 & sources of unstructured data :

Unstructured data has not been organized into a format that makes it easier to access and process. In reality, very little data is completely unstructured. Even things that are often considered unstructured data, such as documents and images, are structured to some extent. Structured data is basically the opposite of unstructured: It has been reformatted and its elements organized into a data structure so that elements can be addressed, organized and accessed in various combinations to make better use of the information.

Semi-structured data lies somewhere between the two. It is not organized in a complex manner that makes sophisticated access and analysis possible; however, it may have information associated with it, such as metadata tagging, that allows elements contained to be addressed.

Here's an example: A Word document is generally considered to be unstructured data. However, you can add metadata tags in the form of keywords and other metadata that represent the document content and make it easier for that document to be found when people search for those terms -- the data is now semi-structured. Nevertheless, the document still lacks the complex organization of the database, so falls short of being fully structured data.

In reality, there is considerable overlap between the boundaries of the three categories, which are sometimes described collectively as the data continuum

Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure.

For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.

Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. XML and other markup languages are often used to manage semi-structured data.

Issues with terminology :

A lack of tools that easily manage unstructured data. Tools need to provide efficient text parsing and analytics, taxonomy and metadata management.

Difficulty integrating unstructured data with existing information systems. The two are often seen as apples and oranges when it comes to analytics and decision making.

Shortage of skills in existing staff

Missing sense of urgency for managing unstructured data

Despite our best efforts to corral the unstructured beast, this kind of data continues to grow larger and presents a real problem for organizations that want to automate and improve their ability to understand their business, anticipate what’s coming and act quickly on risk and opportunity. There are certainly tools that are maturing and providing the beginnings of a solution. The challenge, however, will be in finding the urgency and getting our organizations to see the value of getting data out of its various hiding places and into a place that it can be used and valued.

Dealing with unstructured data Place me in the basket

A big part of the problem is identifying the unstructured data in order to manage it. For example, if you're looking at bitmap images, seismic data, audio or video, there is no way to really identify the data other than the filename and extension -- there is no way to "look" at the data and know that a given piece of data comprises an image or other data type. This makes essential management tasks, like data identification, classification, legal discovery and even basic searches, very challenging for the enterprise. Just consider your own home PC. We pack so much data onto our systems that it eventually becomes a waste of time to even look for a file, especially if we've forgotten the filename.

And with growth, there is also additional cost associated with unstructured data. The inability to actually manage that data is becoming costly in terms of wasted time, wasted storage capacity and potential legal exposures if data cannot be located.

a large percentage of which contain what could be considered sensitive information -- strewn across the network on every storage device imaginable. There are literally thousands and thousands of files containing sensitive text strewn across practically every network unaccounted for, unclassified and unprotected. Sensitive information spread around the enterprise is not necessarily a problem in and of itself, but once you've thrown in all the information protection requirements mandated by governments and industry groups from all levels, you've got yourself quite an issue to manage.