LESSON HANDOUT

Data storage and databases

Databases

A database is an **organised collection of data** and exists to enable the management of large volumes of data. The collection of software applications and programs which allow users to interact with and manage databases are called **database management systems (DBMS)**. For example, Oracle is a database management system.

Databases offer the following advantages:

- **Security**. Databases allow access to be restricted and scaled according to security and privacy needs.
- **Processing speed**. Databases store data in an efficient system which reduces storage cost and optimises performance.
- **Data quality**. Databases put restrictions on what can or cannot be recorded under columns, ensuring a level of data validity. For example, it will give you a rap on the knuckles if you try to store text in a number column.
- **Single accessible source**. Databases allow data from multiple sources to be consolidated into a single place where it can be accessed by multiple people at the same time.

Databases can be divided into two groups: structured and unstructured databases.

Structured databases

Structured databases, or relational databases, are the most popular type of database. The data that they hold, structured data, is nearly always the format required for analysis.

In a structured database, data is stored in a manner where every record represents a single instance and every column represents an attribute or variable about the record. This means that all the variables within a record are strictly related to one another because they describe the same record) - hence the term relational.

The programming language that we use to interact with structured databases is called Structured Query Language (SQL).

Unstructured Database

Unstructured databases are the answer to storing what is called **big data**. A data analyst will typically have little interaction with unstructured data other than to extract, or mine, data into a structured format for analysis. This transformation of unstructured to structured data is generally the responsibility of the Data Engineer.

Examples of unstructured databases are Hadoop and Apache Spark.

Big data

Big data is more than just a large amount of data. Three things define big data, **the three V'S:**

- **Volume.** The scale or quantity of data involved.
- **Variety.** The different varieties of data that are required to be stored such as video, text or transactional data.
- **Velocity.** The speed at which the data must be processed.