Module 2

DATA LIFE CYCLE

Plan → Capture → Manage → Analyze → Archive → Destroy

1. **Plan:** Decide what kind of data is needed, how it will be managed, and who will be responsible for it.
2. **Capture:** Collect or bring in data from a variety of different sources.
3. **Manage:** Care for and maintain the data. This includes determining how and where it is stored and the tools used to do so.
4. **Analyze:** Use the data to solve problems, make decisions, and support business goals.
5. **Archive:** Keep relevant data stored for long-term and future reference.
6. **Destroy:** Remove data from storage and delete any shared copies of the data.

Variations of Data Life Cycle

1. Plan
2. Acquire
3. Maintain
4. Access
5. Evaluate
6. Archive

U.S Fish & Wildlife service

1. Plan
2. Acquire
3. Process
4. Analyze
5. Preserve
6. Publish/share

Several cross-cutting or overarching activities are also performed during each stage of their life cycle:

- Describe (metadata and documentation)
- Manage quality
- Backup and secure

U.S Geological Survey

1. Capture
2. Qualify
3. Transform
4. Utilize
5. Report
6. Archive
7. Purge

Financial Institutions

1. Generation
2. Collection
3. Processing
4. Storage
5. Management
6. Analysis
7. Visualization
8. Interpretation

Harvard Business School

More on the phases of data analysis

The ask phase

At the start of any successful data analysis, the data analyst:

- Takes the time to fully understand stakeholder expectations
- Defines the problem to be solved
- Decides which questions to answer in order to solve the problem

Qualifying stakeholder expectations means determining who the stakeholders are, what they want, when they want it, why they want it, and how best to communicate with them. Defining the problem means looking at the current state and identifying the ways in which it's different from the ideal state. With expectations qualified and the problem defined, you can derive quest that will help achieve these goals.

The prepare phase

In the prepare phase, the emphasis is on identifying and locating data you can use to answer your questions. This is where data analysts collect and store data they'll use for the upcoming analysis process.

The process phase

In this phase, the aim is to refine the data. Data analysts find and eliminate any errors and inaccuracies that can get in the way of results. This usually means:

- Cleaning data (spreadsheets, SQL used)
- Transforming data into a more useful format
- Combining two or more datasets to make information more complete
- Removing outliers (data points that could skew the information)

After data analysts process data, they check the data they prepared to make sure it's complete and correct. This phase is all about getting the details right. Accordingly, the data analyst will refine strategies for verifying and sharing their data cleaning with stakeholders.

The analyze phase

With a solid foundation of well-defined questions and clean data, you'll delve into the analyze phase. This is when you turn the data you've gathered, prepared, and processed into actionable information. Data analysts use many powerful tools in their work. (SQL, R, etc.)

The share phase

This phase is exactly what it sounds like: It's time to share what you've learned with your stakeholders. data analysts interpret results and share them with others to help stakeholders make effective, data-driven decisions. In the share phase, visualization (often w R) is key.

The act phase

The data analysis journey culminates in the act phase, when data insights are put to work.

DATA ANALYSIS TOOLBOX

Most common tools used:

- → Spreadsheets (Microsoft Excel, Google Sheets)
- → Query Languages (SQL)
- → Visualization tools (Tableau, Looker)

SPREADSHEETS

- Collect, store, organise, & sort information
- Identify patterns, piece data together for specific project
- Create excellent data visualizations, like graphs and charts

DATABASES & QUERY LANG.

- collection of structured data stored in a system
- allow analysts to select, create, add, or download data from a database for analysis
- allow isolation of specific info from database
- makes it easy to learn & understand requests made to databases

VISUALISATION TOOLS

- turn complex nos. into a story people can understand
- help stakeholders come up with conclusions
- Tableau → drag-and-drop feature
- Looker → communicates with database directly

Spreadsheets	Databases
Accessed through a software application	Database accessed using a query language
Structured data in a row and column format	Structured data using rules and relationships
Organizes information in cells	Organizes information in complex collections
Provides access to a limited amount of data	Provides access to huge amounts of data
Manual data entry	Strict and consistent data entry
Generally one user at a time	Multiple users
Controlled by the user	Controlled by a database management system

Database: a collection of data stored in a computer system

formula: set of instructions used to perform a calculation using data in a spreadsheet

function: preset command that automatically performs a specified process or task using data in a spreadsheet