**Follow the data life cycle**

Vòng đời dữ liệu (***data life cycle***) là quá trình dữ liệu đi qua từ lúc hình thành, xử lý và phá hủy

Mục tiêu xem xét vòng đời dữ liệu để:

* Biết được những pha mà dữ liệu cần đi qua
* Triển khai những thực thi thích hợp ở mỗi pha

***Các pha trong vòng đời dữ liệu:***

A diagram of a diagram

Description automatically generated with medium confidence

1. Lên kế hoạch (***planning***): lập kế hoạch cho bài toán, các quyết định cần làm và loại dữ liệu cần có, cách quản lý xuyên suốt vòng đời, người chịu trách nhiệm và kết quả kỳ vọng từ dữ liệu.
2. Thu thập (***capture***): thu thập dữ liệu từ nơi nào và bằng cách nào, thu thập từ một nguồn hay nhiều nguồn, cách đưa chúng vào hệ thống quản lý.
3. Quản lý (***manage***): cách thức dữ liệu được chăm sóc như lưu trữ dữ liệu như thế nào, cách sử dụng ra sao, bảo mật chúng thế nào và hành động gì để bảo trì, bảo dưỡng chúng.
4. Phân tích (***analyze***): cách thức phân tích dữ liệu, công cụ/công nghệ gì hỗ trợ cho phân tích.
5. Đóng gói (***archive***): dữ liệu có dùng tiếp không, xử lý dữ liệu cũ thế nào, thời điểm nào cần quan tâm xử lý chúng.
6. Phá hủy (***destroy***): dữ liệu có cần phải xóa không? Nếu có thì khi nào và như thế nào? Việc xóa và cắt nhỏ dữ liệu còn giúp bảo vệ thông tin riêng tư

**Outline the data analysis process**

Stakeholders are people who have invested time and resources into a project and are interested in the outcome

## **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 questions that will help achieve these goals.

In an upcoming course, you'll learn how to ask effective questions and define the problem by working with stakeholders. You'll also cover strategies that can help you share what you discover in a way that keeps people interested.

**The prepare phase**

In the prepare phase, the emphasis is on identifying and locating data you can use to answer your questions. In an upcoming course, you'll learn more about the different types of data and how to identify which kinds of data are most useful for solving a particular problem. You'll also discover why it's so important that data and results are objective and unbiased. In other words, any decisions made from an analysis should always be based on facts and be fair and impartial.

**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
* 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. In an upcoming course, you’ll use spreadsheets and structured query language, or SQL, to clean data.

**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. In one upcoming course you'll continue using two of them: spreadsheets and SQL. In another upcoming course you’ll explore using the programming language R to work with and analyze data.

**The share phase**

This phase is exactly what it sounds like: It’s time to share what you’ve learned with your stakeholders! In this part of the program, you'll learn how data analysts interpret results and share them with others to help stakeholders make effective, data-driven decisions. In the share phase, visualization is a data analyst's best friend. So, an upcoming course will highlight why visualization is essential to getting others to understand what your data is telling you. In another upcoming course, you’ll learn how to visualize data with R.

**The act phase**

The data analysis journey culminates in the act phase, when data insights are put to work. For you, this action involves preparing for your job search and having the chance to complete a case study project. It's a great opportunity for you to bring together everything you've worked on throughout this course. Plus, adding a case study to your portfolio helps you stand out from other candidates!

**The data analysis toolbox**

***Công cụ hỗ trợ***

Trong quá trình thực hiện phân tích dữ liệu, có nhiều công cụ hỗ trợ như:

* Bảng tính: Microsoft Excel, Google Sheets
* Truy vấn dữ liệu SQL: MySQL, Microsoft SQL Server
* Trực quan hóa dữ liệu: Tableau, Looker

Công cụ bảng tính (***spreadsheet***) là công cụ lưu trữ dữ liệu ở mức đơn giản và thông dụng.

Các đặc điểm chính:

* Tổ chức dữ liệu dưới dạng cột và dòng
* Sắp xếp và lọc dữ liệu
* Thực hiện tính toán trên dữ liệu thông qua các công thức (formula)
* Thực thi các lệnh, các nhiệm vụ thông qua các hàm (function)

Cơ sở dữ liệu (***database***) là cách thức dữ liệu được tập hợp và lưu trữ một cách chặt chẽ và có tổ chức hơn, và được thực hiện bởi một chương trình máy tính. Dữ liệu lưu trữ trên cơ sở dữ liệu được truy xuất/giao tiếp thông qua ngôn ngữ truy vấn dữ liệu như SQL (***structured query language*** – ngôn ngữ truy vấn có cấu trúc).

* Để thực hiện yêu cầu (request), cập nhật (update) và truy xuất (retrieve) dữ liệu từ cơ sở dữ liệu, ta sử dụng câu truy vấn (query)

Trực quan hóa (***visualization***) là thể hiện dữ liệu dưới dạng đồ thị nhằm tạo sự dễ dàng cho các bên liên quan trong việc hiểu, đưa ra quyết định và chiến lược thực thi. Một số công cụ phổ biến: **Tableau, Looker.**

# Glossary terms from module 2

## **Terms and definitions for Course 1, Module 2**

**Database:** A collection of data stored in a computer system

**Formula:** A set of instructions used to perform a calculation using the data in a spreadsheet

**Function:** A preset command that automatically performs a specified process or task using the data in a spreadsheet

**Query:** A request for data or information from a database

**Query language:** A computer programming language used to communicate with a database

**Stakeholders:** People who invest time and resources into a project and are interested in its outcome

**Structured Query Language:** A computer programming language used to communicate with a database

**Spreadsheet:** A digital worksheet

**SQL:** (Refer to Structured Query Language)