DATA SCIENCE LIFECYCLE

INTRODUCTION

Data science is a structured process that can be divided into five key stages:

1. Capturing
2. Processing
3. Analysis
4. Communication
5. Maintenance

CAPTURING DATA

The capturing stage is crucial because it sets the foundation for the entire data science process. This stage involves two primary activities:

Defining the Purpose and Problems

- **Identify Stakeholders**: Understand who needs the problem solved (e.g., business stakeholders, project sponsors).
- **Define Goals**: Establish clear, measurable, and quantifiable goals for the project.

Acquiring Data

- **Identify Data Sources**: Determine what data is needed and where it can be obtained.
- **Evaluate Data Quality**: Ensure the data is sufficient and of acceptable quality.
- **Explore Data**: Conduct preliminary data exploration to confirm its suitability for the project goals.

PROCESSING DATA

The processing stage involves discovering patterns in the data and building models to understand these patterns. This stage often uses statistical methods and machine learning techniques.

Techniques Used:

- **Classification**: Categorize data for efficient use.
- **Clustering**: Group similar data points.
- **Regression**: Predict or forecast values based on relationships between variables.

MAINTAINING DATA

Maintenance is an ongoing process that involves managing, storing, and securing data throughout the project's lifecycle.

Storing Data

- **On-Premise vs. Off-Premise**: Decide whether to store data on your own equipment or in a data center.
- **Public vs. Private Cloud**: Choose between shared public cloud services or a private cloud for more control over data security.

Data Storage Types

- **Cold Data**: Rarely accessed data, cheaper to store.
- **Hot Data**: Frequently accessed data, more expensive but quicker to retrieve.

Managing Data

- **Data Cleaning**: Use automated tools to cleanse, aggregate, and compress data for consistency and quality.
- **Example Tool**: Azure Data Factory.

Securing Data

- **Encryption**: Ensure all data is encrypted.
- **Access Control**: Limit data access to necessary personnel.
- **Compliance**: Adhere to local laws and regulations and maintain ethical standards.

Key Security Practices:

1. Confirm data encryption.
2. Inform customers about data usage.
3. Remove access for those who leave the project.
4. Restrict data alterations to certain project members.