# Data Engineering

What is Data Engineering?

Data Engineering is the practice of designing, building and maintain systems that enable the efficient collection, storage and processing of large volumes of data. It focuses on creating robust data pipelines that transform raw data into usable formats for data analysis, machine learning and business intelligence.

**Essential Data Engineering Tools & Technologies**

**1️) SQL (Structured Query Language)**

* Used to query, manipulate, and manage structured data in relational databases.
* Common databases: MySQL, PostgreSQL, Snowflake, and BigQuery.

**2️) Python**

* A powerful programming language for data processing, automation, and building ETL pipelines.
* Popular libraries: Pandas, NumPy, PySpark.

**3️) Hadoop**

* A distributed storage and processing framework for big data.
* Uses HDFS (Hadoop Distributed File System) for data storage and MapReduce for parallel processing.

**4️) Apache Spark**

* A fast, distributed computing engine for big data processing and analytics.
* Supports batch and real-time data processing with APIs in Python, Scala, and Java.

**5️) Apache Hive**

* A data warehouse infrastructure built on Hadoop for querying and managing large datasets.
* Uses SQL-like queries (HiveQL) to simplify working with big data.

**6️) Apache Kafka**

* A real-time data streaming platform used for event-driven architectures.
* Enables high-throughput, low-latency messaging between distributed systems.

**7️) Docker**

* A containerization tool that allows applications to run in isolated environments.
* Helps package data pipelines and deploy them consistently across different platforms.

**8️) Apache Airflow**

* A workflow orchestration tool for scheduling and managing complex ETL pipelines.
* Uses Directed Acyclic Graphs (DAGs) to automate data workflows.

**9️) AstroDB**

* A managed cloud-native database service optimized for Apache Cassandra.
* Designed for scalable, high-performance applications with real-time data access.

**10) Databricks**

* A cloud-based data engineering and AI platform built on Apache Spark.
* Supports data lakes, real-time processing, and machine learning workflows.

**1️1) Azure / GCP / AWS**

* **Azure** (Microsoft) – Cloud services for computing, storage, and analytics.
* **GCP** (Google Cloud Platform) – BigQuery, Dataflow, and AI services for scalable data processing.
* **AWS** (Amazon Web Services) – Redshift, S3, Glue, and Lambda for big data and cloud computing.