Required Tools & Setup for Hands-On Data Engineering Practice

To effectively kick off your hands-on learning journey in Data Engineering, the following tools and environments must be properly installed and accessible. These tools will be used throughout the practical sessions to simulate real-world data engineering workflows, pipelines, and orchestration systems.

1. Python

- Install the latest stable version of Python (3.10 or higher).
- Python will be the primary language used for:
    - Data extraction, transformation, and loading (ETL/ELT)
    - Building and automating data pipelines
    - Scripting, automation, and integrations

2. Visual Studio Code (VS Code)

- Install Visual Studio Code as the primary development environment.
- Required extensions:
    - Python
    - Docker
    - SQL (PostgreSQL / Microsoft SQL Server)
- VS Code will be used for:
    - Writing and debugging Python applications
    - Developing and executing SQL queries
    - Managing Docker containers and services
    - Source code management using Git

3. PostgreSQL Database

- Install PostgreSQL (latest stable version) locally or run it via Docker.
- PostgreSQL will be used as:
    - A source system for transactional and operational data
    - A destination for transformed and analytical datasets
- Ensure access to:
    - pgAdmin
    - Core SQL concepts including DDL, DML, joins, indexing, and performance basics

4. Microsoft SQL Server (MS SQL)

- Install Microsoft SQL Server (Developer or Express Edition).
- Install SQL Server Management Studio (SSMS) for database administration.
- MS SQL Server will be used to:
    - Simulate enterprise-grade relational database environments
    - Practice data movement, transformation, and synchronization across systems

5. Azure Data Factory (ADF)

- Access Azure Data Factory through the Microsoft Azure Portal.
- Requirements:
    - An active Azure account (Free Tier is sufficient)
    - Basic familiarity with Azure resources and services
- Azure Data Factory will be used for:
    - Designing and executing cloud-based ETL/ELT pipelines
    - Orchestrating data flows between on-premise and cloud systems
    - Scheduling jobs, monitoring executions, and handling failures

6. Apache Airflow (Docker-Based Setup)

- Install Docker and Docker Compose on your machine.
- Deploy Apache Airflow using Docker.
- Apache Airflow will be used for:
    - Workflow orchestration and scheduling
    - Managing task dependencies and retries
    - Monitoring pipeline execution and alerts
- This setup mirrors production-grade orchestration environments used in industry.

Expected Outcome

With this setup in place, you will be able to:

- Build and manage end-to-end data engineering pipelines
- Work seamlessly with both on-premise and cloud-based data platforms
- Orchestrate, monitor, and troubleshoot workflows
- Gain practical, job-ready experience aligned with real-world data engineering roles