Setting Up a Laptop/Desktop for a Data Engineer with 4 Years of Experience

Operating System (OS)

- **Preferred OS:** Ubuntu 22.04 LTS (or any Linux-based system)
- **Alternative:** Windows 11 with WSL (Windows Subsystem for Linux)

Hardware Specifications

- **CPU:** Minimum Intel i7 or AMD Ryzen 7
- **RAM:** Minimum 32GB
- **Storage:** 1TB SSD
- **GPU:** Optional, but preferred if working with Machine Learning/Deep Learning

Core Technologies

Programming Languages

- **Python**: Primary language for data engineering tasks (ETL, data pipelines, scripting).
- **SQL**: For database querying and management.
- **Scala**: For Apache Spark development.
- **Bash**: For scripting and automation.

Big Data Frameworks

- **Apache Spark**: For distributed data processing.
- **Apache Hadoop**: For distributed storage and processing (HDFS, MapReduce).
- **Apache Kafka**: For real-time data streaming.
- **Apache Flink**: For stream processing.

Databases

- **Relational Databases**: PostgreSQL, MySQL.
- **NoSQL Databases**: MongoDB, Cassandra, Redis.
- **Cloud Databases**: Amazon RDS, Google BigQuery, Snowflake.

Cloud Platforms

- **AWS**: S3, Redshift, Glue, EMR, Lambda.
- **Google Cloud Platform (GCP)**: BigQuery, Dataflow, Pub/Sub.
- **Microsoft Azure**: Azure Data Lake, Azure Synapse, Databricks.

Data Orchestration

- **Apache Airflow**: For workflow orchestration.
- **Prefect**: Modern alternative to Airflow.

Containerization & Orchestration

- **Docker**: For containerizing applications.
- **Kubernetes**: For container orchestration.

Version Control

- **Git**: For version control (GitHub, GitLab, Bitbucket).

Integrated Development Environments (IDEs)

Primary IDE

- **PyCharm Professional**: Best for Python development with advanced features like database tools, remote development, and scientific mode.
- **IntelliJ IDEA**: For Scala and Java development (Apache Spark).

Alternative IDEs

- **VS Code**: Lightweight and highly customizable with a vast library of extensions.
- **Jupyter Notebooks**: For interactive data analysis and prototyping.

Plugins for IDEs

- **PyCharm/IntelliJ Plugins**:
    - **Big Data Tools**: For connecting to Hadoop, Spark, and databases.
    - **Database Navigator**: For managing databases.
    - **GitToolBox**: Enhanced Git integration.
    - **Rainbow CSV**: Color-coded CSV files.
    - **Scala**: For Scala development.
- **VS Code Extensions**:
    - **Python**: For Python development.

    - **SQLTools**: For database management.
    - **Docker**: For Docker integration.
    - **Prettier**: Code formatting.
    - **Remote - SSH**: For remote development.

Docker Setup

- Install Docker for containerizing applications
- Use Docker Compose for multi-container applications

Monitoring and Logging

- **Prometheus + Grafana** for monitoring
- **ELK Stack (Elasticsearch, Logstash, Kibana)** for logging
- **Datadog** as an alternative

Learning Resources

- **Books:**
    - Designing Data-Intensive Applications by Martin Kleppmann
    - The Data Engineering Cookbook by Andreas Kretz
- **Online Platforms:**
    - Coursera
    - Udemy
    - DataCamp

Additional Tools

- **Postman** for API testing
- **Insomnia** as an alternative
- **Slack** for team communication
- **Trello** or **Jira** for task management
- **Notion** for documentation