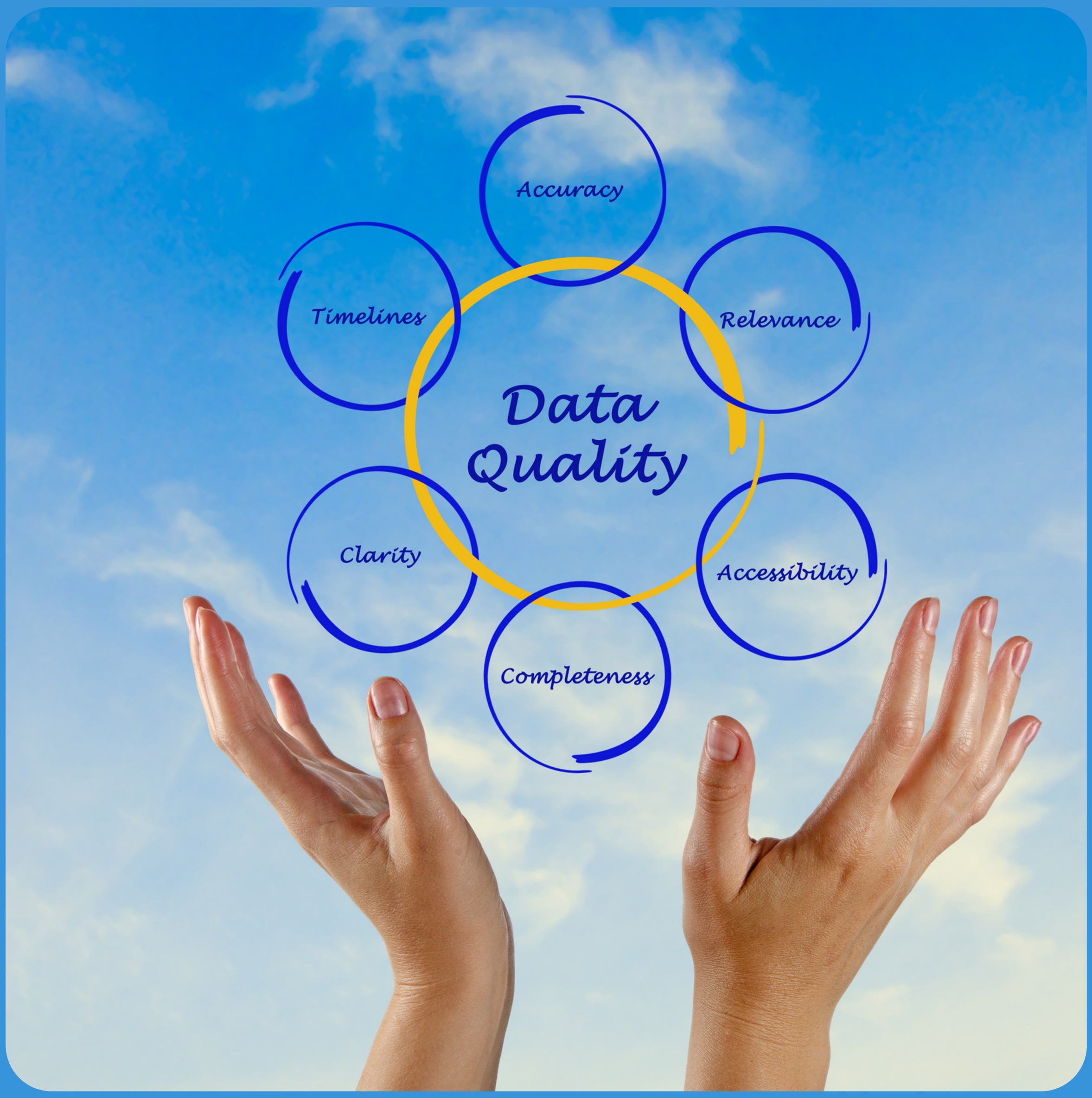

Comprehensive Data Engineering Roadmap for Learners

This roadmap provides a structured overview of essential tools and concepts, **helping learners navigate** the data engineering landscape effectively.

Programming Languages

Data engineering is the backbone of modern data-driven applications. It involves building pipelines, transforming data, managing storage, and enabling analysis. While tools and platforms evolve, programming languages remain the core building blocks.

Python — The King of Data

- Rich ecosystem: pandas, pySpark, Airflow, dbt, etc.
- Widely used in ETL pipelines, data wrangling, and automation

Scala — The Spark Specialist

- Native language of Apache Spark
- Works seamlessly with big data workflows
- High-performance Spark jobs
- Real-time data streaming

Java — Good to have

- Many enterprise data systems are Java-based

DSA & Spark

Data Structures & Algorithms

These are essential for building efficient data pipelines, processing data in memory, and writing performant ETL jobs.

PySpark

PySpark is the Python API for Apache Spark, enabling scalable data processing and big data analytics using Python. It allows you to write distributed computing jobs for tasks like ETL, machine learning, and SQL querying across large datasets.

Scala Spark

- Many enterprise data systems are using Scala & Spark

DBMS, SQL & Shell Scripting

Database Management System

RDBMS: PostgreSQL / MySQL
NoSQL: MongoDB, Redis

SQL (Structured Query Language)

SQL is a language to interact with relational databases. It helps you manage, manipulate, and analyze data stored in tables.

Shell Scripting (Bash)

Shell scripting is used to automate tasks on Unix/Linux systems using command-line commands written in scripts (.sh files).

- Run ETL jobs automatically
- Move and process files (CSV, logs, JSON)
- Schedule tasks using cron
- Monitor data pipelines

Data Engineering concepts

- OLTP vs OLAP
- Star Schema vs Snowflake Schema
- Fact and Dimension Tables
- Slowly Changing Dimensions (SCD Types 1, 2, 3)
- Data partitioning and clustering
- Granularity and grain of fact tables
- Data Modeling

Cloud Platfrom

GCP

AZURE

AWS

Important Tools

Data Validation

- Great Expectations
- Deequ

Scheduling

- Airflow
- cron
- Oozie

Data ingestion

- Sqoop
- Kafka