Data Engineer Skills & Tools Cheat Sheet

A one-page reference of the tools, languages, and concepts every data engineer needs.

Core Languages

- SQL — non-negotiable. Joins, window functions, query optimization.
- Python — for scripting, pipeline logic, and data transformation.
- Bash — for automation scripts and working comfortably in Linux environments.

Pipeline & Orchestration Tools

- Apache Airflow — the most widely used workflow orchestration tool for scheduling and monitoring pipelines.
- dbt (data build tool) — for transforming data already loaded into a warehouse (the "T" in ELT).
- Apache Spark — for distributed processing of large datasets.
- Kafka — for real-time streaming data pipelines.

Storage & Warehousing Concepts

- Data warehouses: Snowflake, BigQuery, Redshift — understand at least one well.
- Data lakes vs. data warehouses — when to use each.
- Star schema and dimensional modeling for analytics-friendly data design.
- Partitioning and indexing strategies for query performance.

Concepts You Must Understand Conceptually

- ETL vs. ELT — and when each approach makes sense.
- Batch vs. streaming processing.
- Idempotency — why pipelines should produce the same result if run twice.
- Data quality checks — handling nulls, duplicates, schema drift.

A Realistic Learning Order

1. SQL fundamentals

Master joins, aggregations, and window functions before anything else.

2. Python for data

Learn pandas and basic scripting for transformation logic.

3. One orchestration tool

Build a simple Airflow pipeline end-to-end.

4. One warehouse

Get hands-on with Snowflake or BigQuery’s free tier.

5. Data quality practices

Deliberately test your pipeline against bad data to build engineering judgment.