Data Engineering Fundamentals to Master

📌 **Prepared By:**
Ankita Gulati
linkedin.com/in/ankita-gulati-de
Pooja Jain
linkedin.com/in/pooja-jain-898253106

🚀 The Beginner-Friendly Starter Stack

Instead of learning 20 tools, master these 4 categories:

🔧 Language: Python

- Easy to learn, widely used
- Great for data manipulation and automation

🗄️ Database: PostgreSQL

- Open source, well-documented
- Teaches you real SQL, not simplified versions

⚙️ Orchestration: Apache Airflow

- Industry standard, lots of resources
- Concepts transfer to any workflow tool

📊 Transformation: dbt

- SQL-based, beginner-friendly
- Teaches modern data modeling practices

✅ Tools Come into Play When Needed:

Learn Tools in Context: Once you understand the problem you need to solve (e.g., "I need to move data from this database to that database securely and efficiently"), then you can learn the specific tool best suited for that task (like a specific ETL tool, a specific database connector, etc.).

You learn the tool because you need it, not just because it exists.

🎓 The Learning Path That Actually Works

Month 1-2: Foundations

- SQL basics and intermediate concepts
- Python for data manipulation (pandas, requests)
- Understanding ETL concepts

Month 3-4: First Tool Deep Dive

- Pick ONE orchestration tool (Airflow recommended)
- Build 2-3 simple pipelines
- Focus on best practices, not advanced features

Month 5-6: Data Modeling

- Learn dimensional modeling concepts
- Try dbt with your existing SQL knowledge
- Build a simple data warehouse

After Month 6: Expand Gradually

- Add cloud platform (AWS/GCP/Azure) basics
- Learn streaming concepts (Kafka/Kinesis)
- Only add tools that solve specific problems

📌 **Prepared By:**
Ankita Gulati
linkedin.com/in/ankita-gulati-de
Pooja Jain
linkedin.com/in/pooja-jain-898253106

🛑 Common Beginner Mistakes to Avoid

❌ Tool Hopping

"I need to learn Spark, Kafka, Kubernetes, Docker, Terraform..."

Result: Surface-level knowledge of everything, mastery of nothing

❌ Certification Chasing

"I'll get AWS, GCP, and Azure certified first"

Result: Lots of badges, can't build a simple pipeline

❌ Tutorial Hell

"I've watched 47 YouTube videos on different tools"

Result: Knowledge without practical application

💡 Why Logic > Tools: Real Examples

Why Understanding 'Why' Matters More Than 'What'

Example 1: Data Validation

Bad Approach: "I need to learn Great Expectations, Deequ, and Monte Carlo"

Good Approach: "I need to understand how to check for nulls, duplicates, and data freshness"

The logic works in any tool. The syntax is just Google-able.

📌 **Prepared By:**
Ankita Gulati
linkedin.com/in/ankita-gulati-de
Pooja Jain
linkedin.com/in/pooja-jain-898253106

Example 2: Pipeline Design

Bad Approach: "Should I use Airflow, Prefect, or Dagster?"

Good Approach: "How do I handle dependencies, retries, and monitoring?"

Once you understand these concepts, switching tools is just learning new syntax.

Example 3: Data Storage

Bad Approach: "PostgreSQL vs MySQL vs MongoDB syntax"

Good Approach: "ACID properties, indexing strategies, query optimization"

Switch tools in days, not months!

✅ Avoid Tool Fatigue & Build Real Things:

Build Projects: Instead of just reading about tools, build small projects! Start with a simple ETL pipeline (maybe even on your own data!). Solve a real (or simulated) problem.

Iterate, Don't Wait: Don't wait until you're a "tool master" to start. Start building, encounter the need for a tool, learn that specific piece, iterate, and repeat. This builds real, practical knowledge.

Develop Transferable Skills: The logic you learn is transferable. Knowing the why behind a process is far more valuable long-term than knowing the what of every tool.

📍 Learn Python & SQL for Data Engineering, with these resources:

1. Python for beginners by Guido van Rossum- https://lnkd.in/dzfrg5Hv
2. Python data cleaning reference guide - https://lnkd.in/dGjp2MNC
3. Python MySQL tutorial by Tutorialspoint - https://lnkd.in/djsg7Wqd
4. Python for data analysis - https://lnkd.in/dbDzpKAM
5. Python cheat sheet for Data Science by DataCamp - https://lnkd.in/dw6ATq7e

6. Research paper on Python tools for big Data Analytics - https://lnkd.in/d6FseYZf
7. Data Engineering with Python by Packt - https://lnkd.in/dP9hVDcB
8. Data Visualization cookbook - https://lnkd.in/dpstdRfu
9. Python for Artificial Intelligence - https://lnkd.in/dhXWTpcZ
10. Python Programming - Real Python by Dan Bader
11. DataLemur 🐵 (Ace the SQL & Data Interview) & Nick Singh 📕🐒 - https://lnkd.in/gCxQjeRD
12. Ankit Bansal's Youtube channel for SQL interview preparation: https://lnkd.in/dZQq33uq
13. Explore Udemy course on SQL by Harshit Bhadiyadra - https://lnkd.in/gyYMzrAg
14. Explore Udemy course on SQL by Harshit Bhadiyadra - https://lnkd.in/gyYMzrAg
15. SQL Tutorial by Tutorialspoint : https://lnkd.in/dka4_dFX
16. Free SQL for Data Analysis course by Mode on Udacity : https://lnkd.in/dBhveAC9
17. Data Analysis using SQL and excel by Gordon S. Linoff : https://lnkd.in/diGj8sVN
18. SQL on Big Data - Technology, Architecture, and Innovation: https://lnkd.in/dyhahYqh

---

📌 **Prepared By:**
Ankita Gulati
linkedin.com/in/ankita-gulati-de
Pooja Jain
linkedin.com/in/pooja-jain-898253106