Awesome Data Engineering

Learning path and resources to become a data engineer

Best books, best courses and best articles on each subject.

Other sections: Data engineering best books

How to read it: First, not every subject is required to master. Look for the "essentiality" measure. Then, each resource standalone for its measurements. "coverage" and "depth" are relative to the subject of the specific resource, not the entire category.

48M

SQL

essentiality
■■■

Querying data using SQL is an essential skill for anyone who works with data

📖 **Head First SQL: Your Brain on SQL - Lynn Beighley**

$ coverage ■■■■■ depth ■■■■□

📖 **Learning SQL: Generate, Manipulate, and Retrieve Data 3rd Edition - Alan Beaulieu**

$ coverage ■■■■■ depth ■■■■■

📖 **Tutorials Point SQL tutorial - tutorials point**

FREE coverage ■■■■□ depth ■■□□□

↓ Show more

Programming language

essentiality
■■■

As a data engineer you'll be writing a lot of code to handle various business cases such as ETLs, data pipelines, etc. The de facto standard language for data engineering is Python (not to be confused with R or nim that are used for data science, they have no use in data engineering).

📖 **Python Crash Course: A Hands-On, Project-Based Introduction to Programming - Eric Matthes**

💲 coverage ■■■■■ depth ■■■■□

📖 **Learning Python, 5th Edition - Mark Lutz**

💲 coverage ■■■■■ depth ■■■■■

📖 **The Python Tutorial - Python documentation**

FREE coverage ■■■■■ depth ■■■■■

↓ Show more

Relational Databases - Design & Architecture

essentiality ■■■

RDBMS are the basic building blocks for any application data. A data engineer should know how to design and architect their structures, and learn about concepts that are related to them.

🎥 **Database Design Course (freeCodeCamp) - Caleb Curry**

FREE coverage ■■■■■ depth ■■■■□

📖 **Designing Data-Intensive Applications - Martin Kleppmann**

💲 coverage ■■■■■ depth ■■■■■

📖 **Normalization of Database - studytonight.com**

FREE coverage ■■■■□ depth ■■■■□

↓ Show more

noSql

essentiality ■■□

noSQL is a term for any non-relational database model: key-value, document, column, graph, and more. A basic acquaintance is required, but going deeper into any model depends on the job (except columnar, in the next section).

NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence - Martin Fowler, Pramod J. Sadalage

$ coverage ■■■■■ depth ■■■□□

NoSQL for Mere Mortals - Dan Sullivan

$ coverage ■■■■■ depth ■■■□□

NoSQL systems (Coursera) - María del Pilar Ángeles

FREE coverage ■■■■□ depth ■■■□□

↓ Show more

Columnar Databases

essentiality ■■□

Column databases are a kind of nosql databases. They deserve their own section as they are essential for the data engineer as working with Big Data online (as opposed to offline batching) usually requires a columnar back-end.

Data Processing Holy Grail? Row vs. Columnar Databases - Joao Sousa

FREE coverage ■■■□□ depth ■■□□□

Why We Built Our Own Distributed Column Store (42:40) - Sam Stokes at strangeloop

FREE coverage ■■■■□ depth ■■■■□

Why Column Stores? - John Schulz

FREE coverage ■■■□□ depth ■■□□□

Data warehouses

essentiality ■■■

Understand the concepts behind data warehouses and familiarize youself with common data warehouse solutions

What is Data Warehouse? Types, Definition & Example - Guru99

FREE coverage ■■■□□ depth ■■■□□

The Data Warehouse Toolkit - Ralph Kimball

$ coverage ■■■■■ depth ■■■■■

Building the Data Warehouse - W. H. Inmon

$ coverage ■■■■■ depth ■■■■■

↓ Show more

OLAP Data modeling

essentiality ■■□

OLAP (analytical) databases (used in data warehouses) data modeling concepts, modeling the data correctly is essential for a functioning data warehouse

The Data Warehouse Toolkit - Ralph Kimball

$ coverage ■■■■■ depth ■■■■■

Data warehouse schema design - dimensional modeling and star schema - Snir David

FREE coverage ■■■□□ depth ■■■■□

Slowly changing dimensions in depth - Snir David

FREE coverage ■■■□□ depth ■■■■□

↓ Show more

(A note)

Data processing - Batch, MapReduce, Streaming

The next 2 categories are all about data processing mechanisms. We'll start with batch processing and MapReduce, typically with Hadoop. This is considered the first gen of data processing. From there we'll go to stream processing, typically done with Spark. These subjects are deeply connected. For example, Spark can operate on HDFS which is the file system for Hadoop. Even though it would seem outdated to learn about batch processing with Hadoop, it is essential to understand the subject even if you plan to live the streaming data life.

Batch data processing & MapReduce

essentiality
■□□

The "first" generation of data processing, using Hadoop and Spring. Everyone should know how it works, but going deep into the details and operations are recommended only if necessary. Focus more on streaming with tools like Spark today.

📄 **Beginner’ s Guide to Batch Processing - talend**

FREE coverage ■■■□□ depth ■■□□□

📄 **What is MapReduce? How it Works - Hadoop MapReduce Tutorial - Guru99**

FREE coverage ■■■□□ depth ■■■□□

📄 **How do Hadoop and Spark Stack Up? - Amir Kalron**

FREE coverage ■■■□□ depth ■■■□□

↓ Show more

Stream data processing

essentiality
■■□

The "next" generation of data processing. Suggested to get a good grasp of the subject from the "Streaming Systems" book and then dive deep into a specific tool like Kafka, Spark, Flink, etc.

📖 **Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing - Tyler Akidau, Slava Chernyak, Reuven Lax**

$ coverage ■■■■■ depth ■■■■■

📖 **Spark: The Definitive Guide: Big Data Processing Made Simple - Bill Chambers, Matei Zaharia**

$ coverage ■■■■■ depth ■■■■■

📖 **Stream Processing with Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications - Fabian Hueske, Vasiliki Kalavri**

$ coverage ■■■■■ depth ■■■■■

↓ Show more

Pipeline / Workflow Management

essentiality
■■□

Scheduling tools for data processing. Airflow is considered to be the defacto standard, but any understanding of DAGs - directed acyclical graphs for tasks will be good.

Directed Acyclic Graph (DAG) - hazelcast
FREE coverage ■■■□□ depth ■■■□□

Airflow docs - Concepts (Demonstrated with Airflow but the concepts are generic) - Airflow documentation
FREE coverage ■■■■□ depth ■■■□□

Data Pipelines with Apache Airflow - Bas P. Harenslak, Julian Rutger de Ruiter
$ coverage ■■■■■ depth ■■■■■

↓ Show more

Security and privacy

essentiality
■□□

How to manage sensitive data, compliance with regulation (GDPR) and more

What is data security? - IBM
FREE coverage ■■■□□ depth ■■■□□

Show All | Free resources | Books | Courses

IT Governance: An International Guide to Data Security and ISO 27001/ISO 27002 - Alan Calder, Steve Watkins
$ coverage ■■■■■ depth ■■■■■

↓ Show more

Made with ♥ by Snir David