


Learn Python, SQL and Command Lines

SQL

- Declarative language used to store and query structured data.
- A common interface for databases and datalakes
- *Super underrated*. Transactional properties make it fast and powerful to use.

Python

- Open source language with third party libraries everywhere
- Virtual environments allow snapshot at point of time
 -  *Key for Reproducibility. Isolated for each system*
- Popular data tools:
 - Pandas, NumPy, sci-kit learn, Tensorflow, PySpark

Command Lines


- Data Engineers often work on remote machines
- Moving around files, searching through logs, version control
 - i.e. useful when building an Extract Transform Load (ETL) pipeline
- Command Lines help facilitate that interaction and improve developer productivity

Resources

Mindset: Find one content that you resonate with the most and stick to it.

Timeframe: 3 months plus. Don't rush it and take more time if you need to.

Suggested Materials (*non-definitive, will be shared in link below*)

	Material	Notes
Python	O'reilly textbooks	One of my all time favourite books from Columbia Professor Andreas C. Muller: Introduction to Machine Learning with Python
	educative.io	Learn Data Structures and Algorithms in Python https://www.educative.io/courses/grokking-the-coding-interview
	Datacamp	Learn Pandas https://www.datacamp.com/tutorial/pandas
	Stackoverflow	Stackoverflow everytime you get stuck
	learnpython.org	Advanced Python concepts
SQL	Leetcode Premium	Low key fire  Have to pay but so worth it IMO
	Codecademy	Basic introduction to SQL https://www.codecademy.com/catalog/language/sql
	Youtube	One of my favourite youtubers for advanced SQL. Checkout

		@techTFQ
	Columbia University (COMS4111) Intro to Databases	Comprehensive slides that cover theory: https://github.com/w4111/w4111.github.io/blob/main/files/lectures/lec5.pdf One of my fav classes when studying in Columbia. Go Lions! 🦁
Command Lines	learnenough.com	Learn enough has other awesome tutorials on Git too https://www.learnenough.com/command-line-tutorial/basics
	The Missing Semester at MIT	The class that taught me how to be a programmer after graduating https://missing.csail.mit.edu/
	The Linux Command Line	For the experts who want to dive deep https://linuxcommand.org/tlcl.php

Learn about Data Storage and Orchestration

Data Storage

- Storage underpins every aspect of Data Engineering
 - ingestion, transformation, queries
- Many Cloud Computing Services already exists [**AWS, GCP, Azure**]
 - Pick the one your company uses

Storage	Description
Object Stores	Gold standard for data lakes
	Inherently key-values stores
	Ideal for unstructured data such as images, audio, text
Relational Databases	The most widely deployed DB in the world is not PostgreSQL or MySQL...it's SQLite! Surprise surprise
	Often times the solution to most Data Engineering problems
	Extremely fast lookup time

Object Stores using AWS S3

Say we're storing a CSV file `s3://like-and-subscribe/jayzern.csv`

Our bucket name is `like-and-subscribe` and our key is `jayzern.csv`

Tons of useful info stored: versioning, metadata, backups

Data Orchestration

- When you talk about Orchestration, usually related to Data Warehousing.
 - The process of getting data into a Data Warehouse is **Extract Transform Load (ETL)**
- Why do we need Data Orchestration?
 - Schedule jobs at specific **cadence** (minutes, days, weeks) for **tasks** (ingestion, processing, transforming)
 - Data Provenance. "*Where did the data originate from?*"
 - Directed Acyclic Graphs (DAGs) help us organize workflows programmatically
- **Apache Airflow** is the gold standard in the industry
 - Design complex ETL tasks
 - Logging capabilities
 - Sleek UI for visualization

Resources

Mindset: Focus on learning concepts and building projects.

Time frame: 2 months

Suggested Materials (*non-definitive*)

- Skim through these documentation quickly to get a high level overview
 - **AWS S3:** <https://aws.amazon.com/s3/>
 - **PostgreSQL:** <https://www.postgresql.org/docs/>
 - **Airflow:** <https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/index.html>



Project idea: Create an Airflow DAG that extracts data from your favorite source (finance, sports, music) and uploads it to a Data Warehouse (any relational database of your choice)

Batch and Stream Processing

Batch Processing

- Heavily intertwined with Data Processing technologies. **Apache Spark** is the gold standard
- Core idea is we have data growing at unmanageable sizes. We have tools that enable us to process them in a scalable way.
- **MapReduce** is a distributed framework within the Hadoop ecosystem
 - **Map**: split data between parallel tasks
 - **Reduce**: aggregate data
 - *The order does not matter*
- Cloud Infrastructure already exists: Amazon EMR, Databricks
- Don't suggest learning specific tools. Instead, focus on understanding concepts and create projects using tool that fits your needs.



Biggest misconception is you'll need to be an expert in Big Data technologies, which is totally wrong. You'll learn these on the job, and these tools are constantly evolving.

Stream Processing

- Build real-time applications where multiple sources are feeding into a system
 - Publish + Subscribe model. Producers publish messages to Topics. Consumers subscribe to topics for information.
- **Apache Kafka** is one of the most popular frameworks.
 - High throughput. Process millions of messages per second.
 - High scalability. Thousands brokers
 - Maintains a commit log that is immutable
- Real world example
 - Netflix uses Apache Kafka for their messaging and streaming needs
 - Literally Billions of data points processed using Kafka each day

Resources

Mindset: Batch/streaming technologies are constantly evolving. By the time you master a tool, it will be obsolete soon. Focus on learning concepts instead, and work on a project that you can showcase to interviews

Time frame: 2 months for reading and project work

Starting Materials

Material	Notes
MapReduce paper	A must read paper for all Data Engineers: https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

How does Netflix use Apache Kafka?	For general purpose understanding. Learning about how tools are used in the real-world. https://www.confluent.io/blog/how-kafka-is-used-by-netflix/
Intro to Distributed Systems Lecture 1 (MIT)	I love this class but the material is grad school level. The first lecture gives a great umbrella overview. Try to pick up the high-level gems in this video. https://www.youtube.com/watch?v=cQP8WApzIQQ
Confluent	Learn about Kafka from the creators of Kafka themselves https://developer.confluent.io/learn-kafka/



Project Idea: Implement Wordcount using MapReduce in the Hadoop environment.

Count the number of words given a book

It doesn't matter if I count from the first or last page.