Open in app Get started

Published in CodeX

Saikat Dutta Follow
Apr 16 · 8 min read · Listen

Save

DATA ENGINEERING 101

How to Become a Data Engineer : Complete Roadmap

A complete roadmap on how you can learn Data Engineering in 2022

In October 2012, HBR predicted that Data Science will be the sexiest job in the 21st astronomical demand for people who could fix these issues, AKA **Data Engineers (DE in short)**.

A lot of people have asked me on LinkedIn to guide them about How to Become a Data Engineer?

So, you are here because you want to become a Data Engineer, but why? Let me answer that first.

Why Become a "Data Engineer" in 2022?

So, supply for quality data engineers are extremely low at the moment and demand is astronomical. And as normal economics will tell you when supply can not match the demand the prices are bound to go up.

> *"With great demand comes great rewards"*

As per Glassdoor, Ambitionbox and Payscale the average Data Engineer Salary in India is 8–9lakhs per annum. However the salary can range from 3–4lakhs for freshers to upwards of 30lakhs for people with 10+ years of experience.

More people are even considering moving away from other data roles to a Data Engineer role. Its a great move, even if you are not into data field.

4 reasons why Data Engineering is a Great career move in 2022?

Data Engineering is the new Sexiest Job of the century. Here are 4 reasons to explain why and whether its a good time…

withsaikat.medium.com

Great, now that we have addressed the WHY, let us go deep into **HOW?**

What are the skills needed to become a Data Engineer?

However, not everything is needed just to start or break into the role.

Note to beginners

> **Beginners shouldn’t feel overwhelmed by the huge set of tools and topics needed to learn.**
>
> **There are several stages of learning involved, and as a beginner you should only concentrate on perfecting the fundamentals.**
>
> **Once you feel comfortable you can move into the advanced topics with time and experience you will feel at home.**

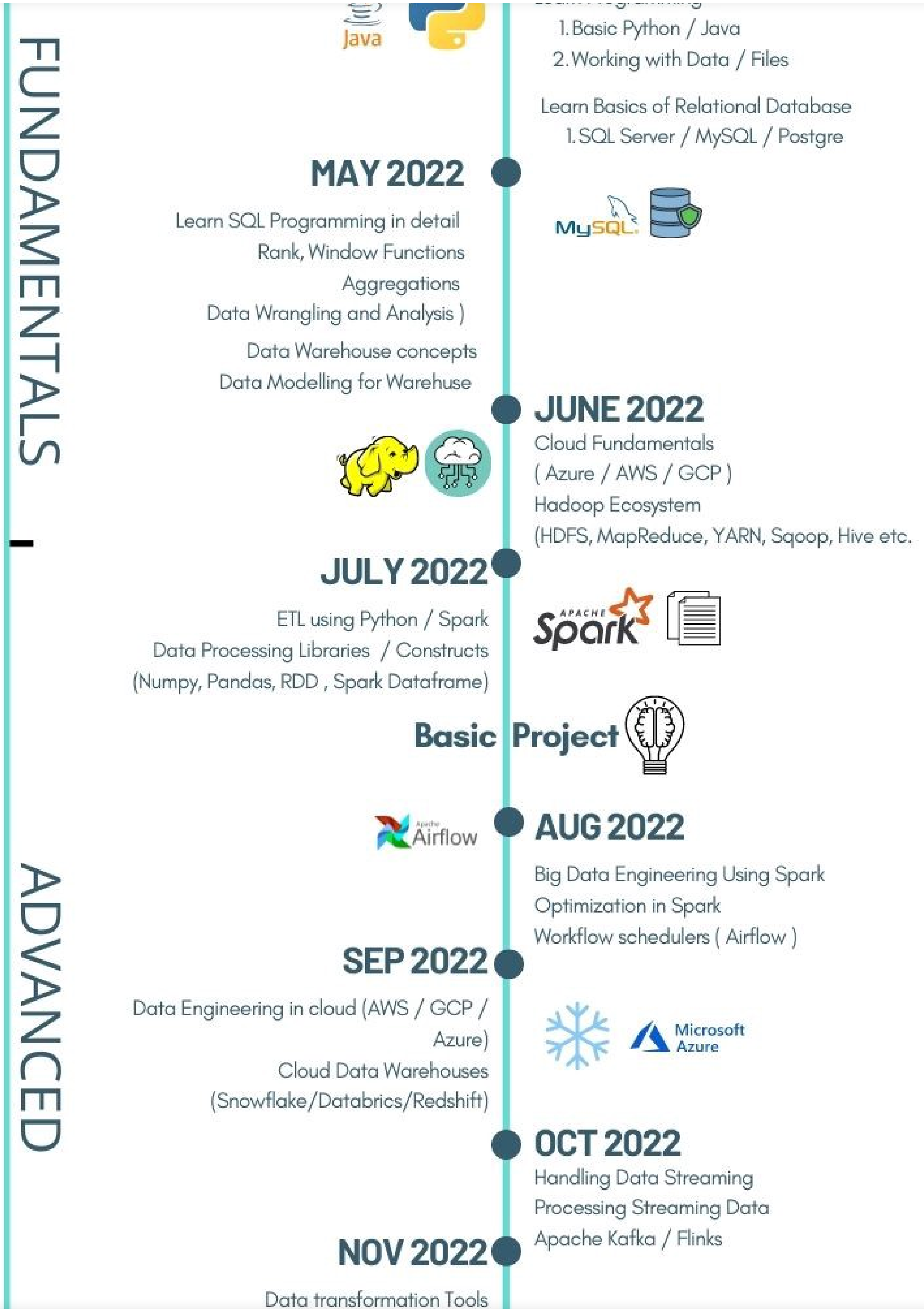
As I have noted above I will divide the complete set of skills and subjects into **Fundamentals, Advanced topics & Good To Have .**

Fundamentals

The Base is the most important part of any building, and its here that any construction starts. Hence, its important to build it better.

Its easy to get distracted. However its important that 3–4 months are spent building the fundamentals.

Once this part is mastered the next phase of learning will be much easier.



FUNDAMENTALS

Learn Programming

1. Basic Python / Java

2. Working with Data / Files

Learn Basics of Relational Database

1. SQL Server / MySQL / Postgre

MAY 2022

Learn SQL Programming in detail

Rank, Window Functions

Aggregations

Data Wrangling and Analysis)

Data Warehouse concepts

Data Modelling for Warehuse

JUNE 2022

Cloud Fundamentals

(Azure / AWS / GCP)

Hadoop Ecosystem

(HDFS, MapReduce, YARN, Sqoop, Hive etc.

JULY 2022

ETL using Python / Spark

Data Processing Libraries / Constructs

(Numpy, Pandas, RDD , Spark Dataframe)

Basic Project

ADVANCED

AUG 2022

Big Data Engineering Using Spark

Optimization in Spark

Workflow schedulers (Airflow)

SEP 2022

Data Engineering in cloud (AWS / GCP / Azure)

Cloud Data Warehouses

(Snowflake/Databrics/Redshift)

OCT 2022

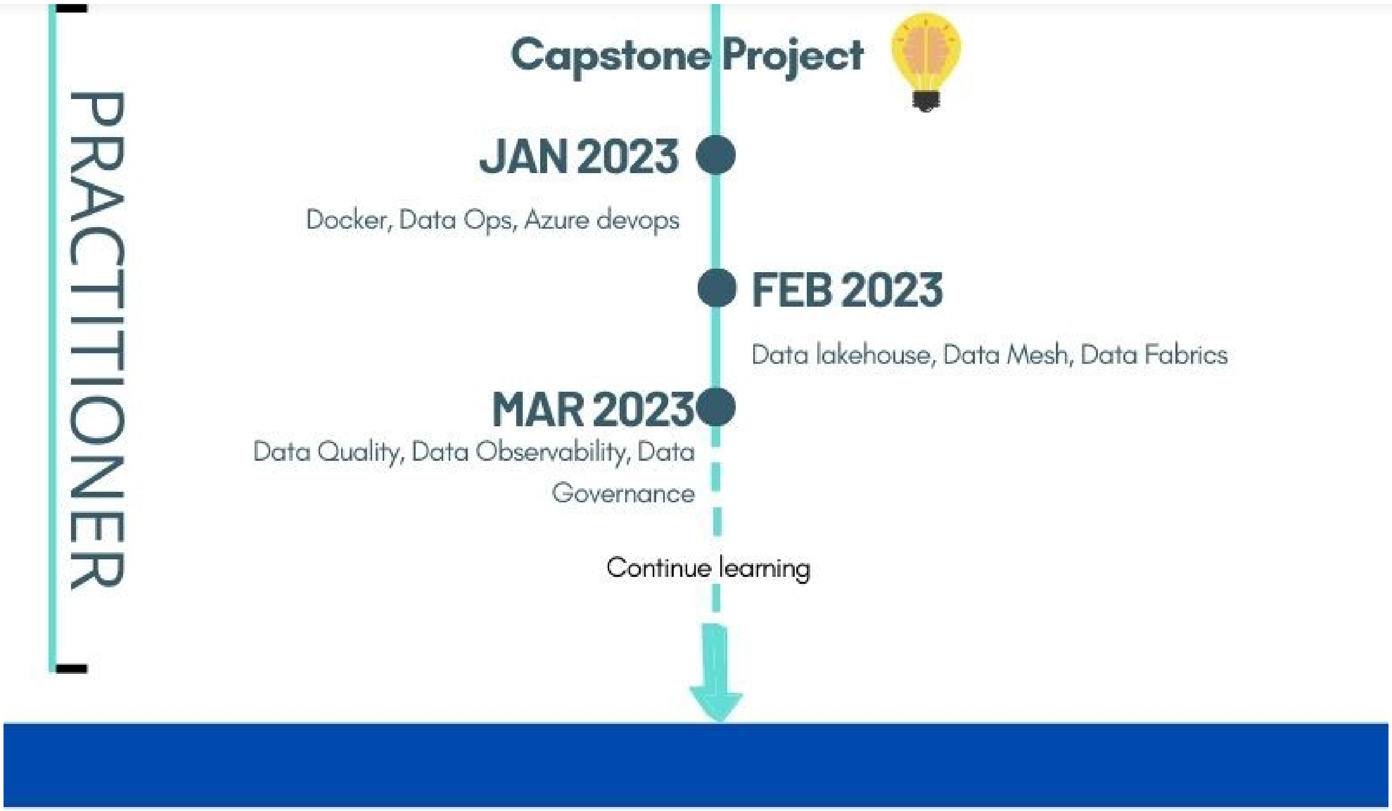
Handling Data Streaming

Processing Streaming Data

Apache Kafka / Flinks

NOV 2022

Data transformation Tools


PRACTITIONER
Capstone Project
JAN 2023
Docker, Data Ops, Azure devops
FEB 2023
Data lakehouse, Data Mesh, Data Fabrics
MAR 2023
Data Quality, Data Observability, Data Governance
Continue learning

Below are the **fundamental** topics to cover, in no specific order of sequence.

1. **Database Concepts:**

Basic Database concepts, normalization, keys, constraints, database storage etc.

2. **Programming**

Basic syntaxes, working with files, connecting to databases, building basic APIs, working with structured (database and tables)and unstructured(xml,json etc.) data.

Python on Youtube.

Java on Udemy

3. SQL

b. SQL from khan academy

c. Scenario based Hands on SQL series from Mentorskool.

4. Data Warehouse and Data Modelling

Basic Data Warehouse Concepts, Data Modelling for Data Warehouse, Star-Snowflake schema, Facts and Dimension tables etc.

5. Cloud Fundamentals

Learn about basics of Cloud computing, SAAS, PAAS, IAAS offerings, distributed computing, Capex vs Op-ex, Elastic scalability, Storage and Compute in the cloud, Data Stacks in cloud.

6. Hadoop Eco-System & Spark

History of Hadoop, Hadoop 1,2,3 , HDFS, MapReduce, YARN, Sqoop, Hive, PIG, HBase, Oozie, zookeeper, SPARK basics

Basic MapReduce programming with Python / Java

Spark with Python in Udemy, With Scala

Spark Dedicated Course

1**st End2ERnd project:** At this point you have all the required skills to create your first basic DE project. Concentrate on the below as you build it:

a. Scrape or collect free data from web.

b. Convert the data into csv / json and read the data using Python

You can not miss the Zoomcamp ser e best free course on Data Engineering, I have found.

Advanced topics

1. **ETL using Python / Scala in Spark**

Creating ETL code in Python / Scala, PySpark, Spark SQL, Spark Context, Spark Jobs, Spark submit, Optimizing Spark Jobs.

2. Data Processing Libraries / Constructs

RDDs, Data Sets, DataFrame etc. , Numpy, Pandas

Different file type (**CSV, JSON, AVRO, Protocol Buffers, Parquet, and ORC**.)

3. NOSQL Db

Pick any (Casandra / MongoDB) \ Graph DB is rarely needed, but good to have.

4. Workflow Management and Schedulers

This is a very important component in the modern Data Stack. Pick between **AirFlow** (most preferred and market leader) or anything else (**Luigi, Prefect)**

Great Airflow Tutorial

5. Data Streaming

Data Velocity is one of the key parameters for Big Data and Data Engineering.

We all want real time analysis and feedback on what's working and what not, Reverse ETL and real-time analytics have become a must have in new business.

6. DE in cloud (AWS / GCP / Azure)

Cover the complete Data Engineering lifecycle in any of the major cloud providers. Complete either one from 1–3 below and complete point 4. As Data offerings of Azure / Google / AWS are conceptually not very different and one can easily pickup the other, once they are comfortable with one.

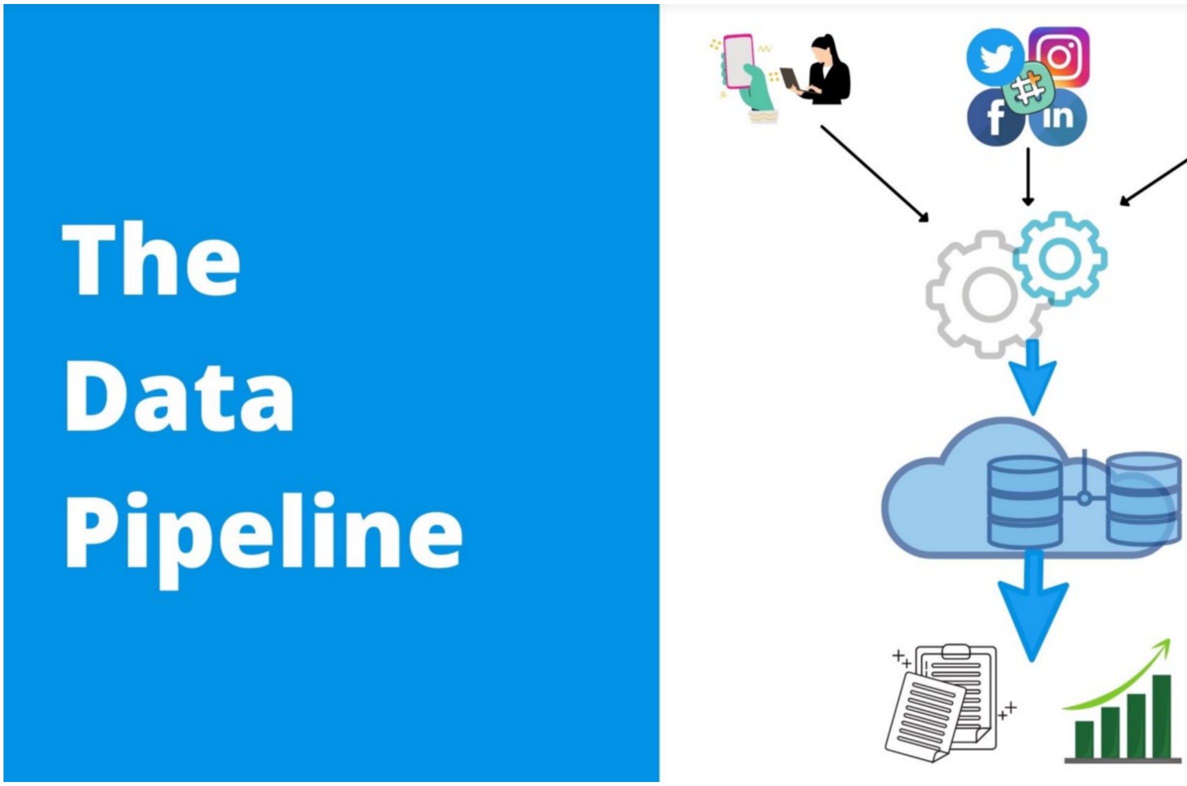
EX:

1. **Azure Stack:** Azure Data Lake, Azure Synapse, Azure Data Factory, Azure Cosmos DB, Azure Event Hub, Power BI.

Refer this course by Ramesh Retnaswami on Data Factory And also on Spark and Databricks here.

1. **Google Stack:** Big Query, Pub-Sub, Dataflow, Dataproc, Looker
2. **AWS Stack:** AWS S3, AWS Kinesis, AWS Glue, Redshift, AWS Athena, Lambda, AWS RDS
3. **Cloud Data Warehouses / Lakes:** Databricks, Snowflake

2**nd End2End project:** The second project should cover more hands on knowledge and should be built more like a real time project.



Data Engineering Project End to End

Good To Have

1. **Dashboarding Tools :** In depth knowledge of any specific Dashboarding tool is not a must have for a Data Engineering role. However it is extremely critical and good to have. Dashboarding can really help identify potential Data Quality issues, impact of bad data and can really save a lot of time for developers.

Power BI / Tableu / Looker are the primary players in the segment.

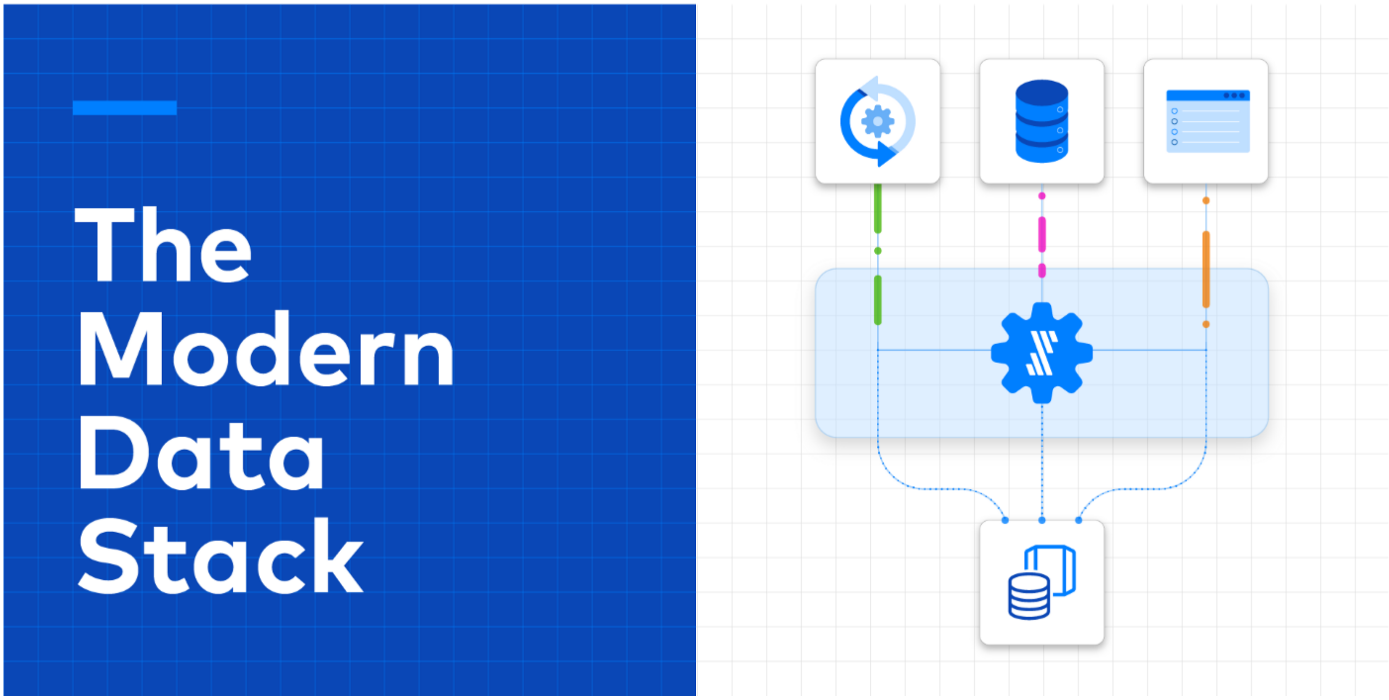
2. Docker : Docker helps to keep the infrastructure related complexity away. This helps to independently and easily setup a Data environment.

3. Devops / Data Ops

4. Modern Data Stack : Modern Data Stack refers to a set of independent mostly open source toolset. These tools provides flexibility to business . Even SMBs and Start up can now easily setup a modern Data Architecture, without worrying for vendor lock in and

Its good to have an understanding of the different tools in the stack and how do they fit in the whole DE roadmap.

Fivetran for ETL, **Airflow** for orchestration, Any cloud warehouse / lake, **DBT** for Data Transformation, **Hightouch** for reverse ETL, **Monte Carlo** for Data Observability etc.



https://www.fivetran.com/blog/what-is-the-modern-data-stack

Final Project : Now that all the important lessons are done, its important to use the learnings and creating an end to end pipeline as a Capstone Project. The important topics to be incorporated are :

1. Building containers to run the ETL/ELT pipelines

2. Creating pipeline with python to Load data in lake.

3. Creating orchestration to run the codes

4. Running jobs on Spark, Batch and Stream processing

5. Data Modelling for Warehouse

8. Data visualization and building the dashboard.
9. Documentation

Conclusion

We might not need each of these skills in the day to day job as a Data Engineer. However you might need one or many of these frequently based on the role.

Learning most of these well will take time. So, keep learning every day. Compounded learning will ensure that with time you get better. There is no shortcut, so don’t believe people who claim to make a Data Engineer in one or two months.

Stay Up To Date:

Are you already someone like me, have been working in the industry as GUI based ETL developer or Data Modeler? or even a code based Data Engineer?

The only secret to stay relevant is to stay updated and up-to-date about all the changes happening in the industry. Follow Data Leaders on LinkedIn, read about blogs and news letters. And most important keep learning everyday.

Feeling Left Out : here’s How to Stay Relevant in IT sector

The world is fast changing. The technology is changing even faster. Here is how you can navigate through this, and…

medium.com

Free Planner

Follow the below link to get access to a free planner of items to cover as part of your Data Engineer preparation. You can tick the items you already covered and track progress.

I want the DE2022 Study plan.

Data Engineering Roadmap in 2022 : Study Plan

Hi learners,With this study plan, you will know the nitty-gritty behind what goes into the making of a Data…

withsaikatdt.gumroad.com

You can either follow along or make your own timeline, but ticking off all the items will give you confidence to face any interviews and also personal satisfaction to have covered the necessary skillsets.

I hope both the blog and the planner inspires you to study along. Go Crush Your Data Engineering dreams in 2022…..

If you still feel lost, don’t hesitate to book some time with me here.

Book a time with Saikat on topmate.io

I Help you discover the career that defines you.

topmate.io

I will be sharing more stories, writings, and experiences in the data industry. You can follow me for more posts like this.

Thanks *for reading! If you want to get in touch with me, feel free to reach me at* withsaikatdt@gmail.com *or my LinkedIn Profile.*