Published in Dev Genius

Amit Singh Rathore Follow
Aug 2 · 4 min read · Listen

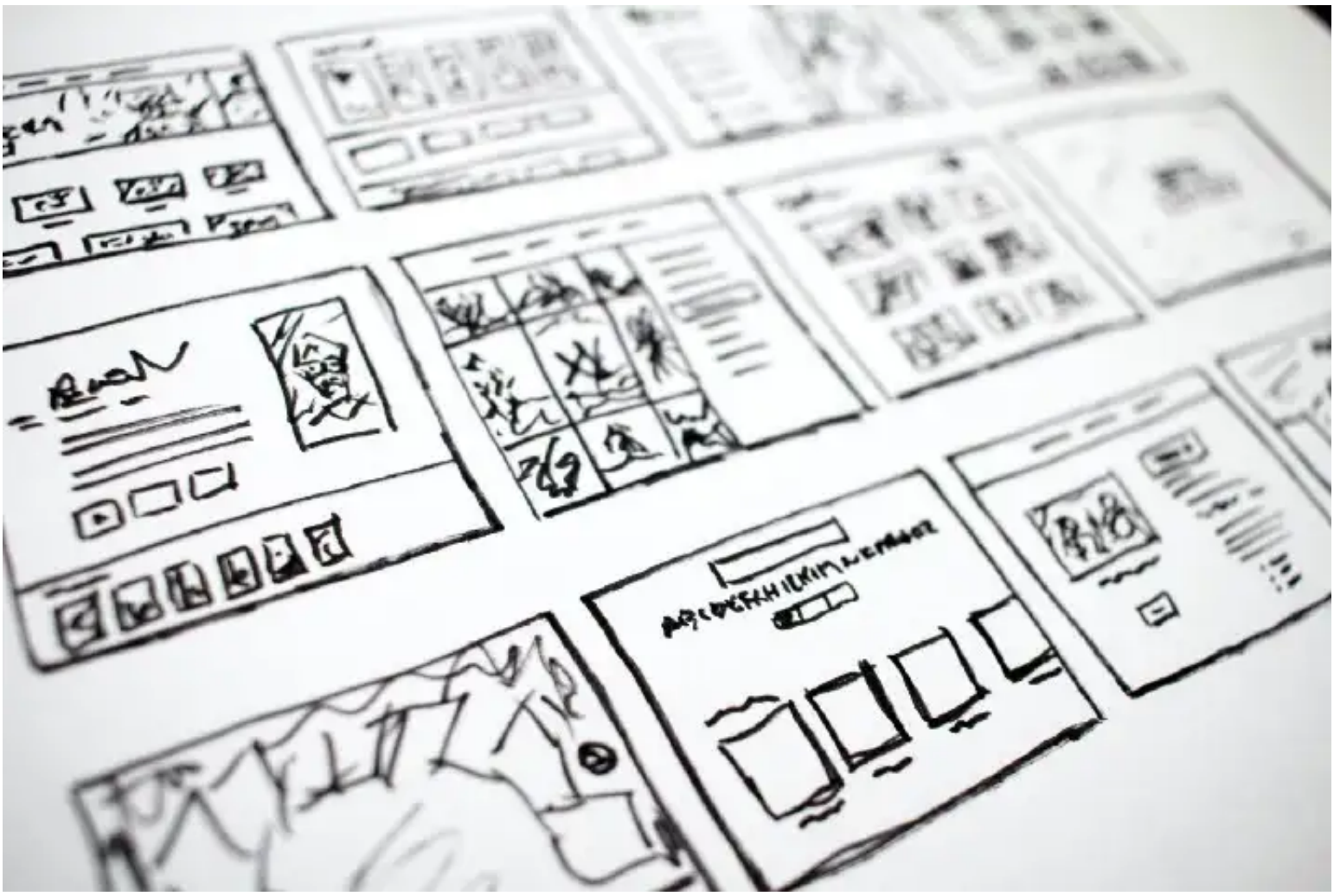

Photo by Hal Gatewood on Unsplash

Data Engineer — Learning Path

Skills you need to become a data engineer

Typically a data engineer performs

3. Write algorithms to transform data into actionable information
4. Automate the above tasks using data pipelines
5. Create new data validation methods for tests
6. Ensure compliance with the company's data governance and security policies
7. Expose data as a service/product
8. Provide technical assistance to BI/BAs in their queries & dashboards

To perform these tasks Data Engineer needs to have certain skill sets. Let us go through the common & basic skills one would need in a data engineer role.

Programming Language

Data engineering leads to various roles in the future like Data Analyst, Data Scientist, ML Engineer, and BI Developer. So choosing the right language is important. Python is more diverse in the sense that it is used widely in all the previously mentioned job profiles.

- **Python** / Scala / Java
- For python Pandas, Numpy & Matplotlib, Scipy
- YAML / JSON

Note: If you want to stick to data engineering in long run as well then you will need to work with both Java (Kafka, Flink, Nifi, Storm, and Datahub are java based)and Scala (spark is written in Scala). So it will be good to know that.

Operating System Basics

Data engineers use tools that are distributed and almost always run on linux systems. So knowing Linux is a must to debug and understand the process flow. Also, Data engineers deal with lots of files and hence file manipulation is an important skill.

- Linux (filesystem vs Object storage, NFS, process, SIGTERM vs SIGKILL)
- Shell scripting (vi, jq)




- Networking — TCP, IP, DNS, SSH, Firewall

Newer data platforms are now being hosted in cloud/virtualized environments hence knowing about virtualization is a good skill to have.

Virtualization

- **Docker** /LXC
- **Kubernetes** / Mesos

DSA

As a data engineer, you are supposed to write fast and scalable solutions. That can be achieved if you have a good understanding of data structure and algorithms. Generally lean towards algorithms that can be distributed.

- Array, Linked List, String, Stack, Queue, Heap, Tree
- DP, Graph (BFS & DFS)

Data storage

As a data engineer, you will often work with databases as a source or sink. So it is required to know the basics of DBMS. These topics also help in the way you store or organize data (data modeling) in databases.

DBMS

- SQL & DBMS concepts
- **MySQL** / Postgres

Key Topics

- JOINS, Aggregation (Group By), Window Function
- ACID & Normalization
- Index, Vertical and horizontal partitioning
- Subquery, Correlated subquery, CTE (recursive)

query patterns. So as a Data Engineer you will need to know NoSQL databases. It is based on the requirement.

- **Cassandra**
- MongoDB
- Elasticsearch

Data Warehouse

Data engineers will build and organize the data for consumption and the majority of that output is stored in data warehouses since it is suited for analysis and aggregated results. So knowing the basics of Data warehousing is needed to build the right kind of warehouse.

- Star vs Snowflake schema
- SCD (Types)
- **Hive**/BigQuery/Redshift/Snowflake

Data Processing/Storage Framework

Data engineers read, clean, enrich, transform and store the data. To do this activity on the large dataset we need a distributed computation framework and for that, we have many frameworks. We need to understand how these frameworks operate.

- Hadoop & HDFS
- **Spark** / Dask / Ray

Queue / Message broker

As Data engineers work on large data and failure in between causes recomputation of large scale, we need a buffer or queue system to decouple certain subsystems. Also in many cases, data engineers have to process huge numbers of small events for those use cases we need a robust messaging platform.

- **Kafka** / Pulsar / Redpanda
- Kinesis




correct schedule and should be automated. For that, we have many orchestration engines that data engineer needs to understand.

- **Airflow**
- Apache Nifi
- Azkaban
- Oozie (Some big enterprises still have these pipelines although the trend is to replace these with other alternatives)
- Step function
- Camunda

Design (system)

As a data engineer, you need to build solutions that are scalable, reliable & fault tolerant. So knowing system design skills come in handy.

- Decoupling
- Backoff / retry/ backfill
- Sharding / Network partitioning
- Distributed designs
- CAP & PACELC
- Lambda & Kappa architectures

Cloud

Nowadays most data platforms are hosted in the cloud and use its offering as auxiliary services to enrich the platform's features. Knowing the cloud adds to the employability of Data engineers.

- **AWS** / GCP / Azure

Every individual has a different learning process. These skills are not needed all at once on day




Sign up for DevGenius Updates

By Dev Genius

Get the latest news and update from DevGenius publication Take a look.

Emails will be sent to kamaljp@gmail.com. Not you?

Get this newsletter