> Text version for visually impaired users

Data Engineer in 2021

- CS fundamentals
  - Basic terminal usage [general recommendation]
  - Data structures & algorithms [general recommendation]
  - APIs [general recommendation]
  - REST [general recommendation]
  - Structured vs unstructured data [general recommendation]
  - Serialisation
  - Linux [general recommendation]
    - CLI
    - Vim
    - Shell scripting
    - Cronjobs
  - How does the computer work? [general recommendation]
  - How does the Internet work? [general recommendation]
  - Git — Version control [general recommendation]
  - Math & statistics basics [general recommendation]

Note: Git is used for tracking changes in source code and coordinating work among programmers. In your day to day work you will use Git server as a service like GitHub, GitLab or Bitbucket.

- Learn a programming language
  - Python [personal recommendation]
  - Java [general recommendation]
  - Scala
  - Go

Note: Learn how to write clean, extensibile code. Spend some time understanding programming paradigms (functional vs. OOP) and best practices (design patterns, YAGNI, stateful vs stateless applications). Get familiar with an IDE or code editor like VSCode.

- Testing
  - Unit testing [general recommendation]
  - Integration testing [general recommendation]
  - Functional testing [general recommendation]
- Database fundamentals
  - SQL [general recommendation]
  - Normalisation [general recommendation]
  - ACID transactions [general recommendation]
  - CAP theorem [general recommendation]
  - OLTP vs OLAP [general recommendation]
  - Horizontal vs vertical scaling [general recommendation]
  - Dimensional modeling [general recommendation]
- Relational databases
  - MySQL [general recommendation]
  - PostgreSQL [general recommendation]
  - MariaDB
  - Amazon Aurora
- Non-relational databases
  - Document databases
    - MongoDB [general recommendation]

        - Elasticsearch [general recommendation]
        - Apache CouchDB
        - Azure CormosDB
    - Wide column databases
        - Apache Cassandra [general recommendation]
        - Apache HBase [general recommendation]
        - Google Cloud Bigtable [personal recommendation]
    - Graph databases
        - Neo4j
        - Amazon Neptune
    - Key-value stores
        - Redis [personal recommendation]
        - Memcached
        - Amazon DynamoDB [general recommendation]

Note: Understand the difference between Document, Wide column, Graph and Key-value NoSQL databases. We recommend mastering one database from each category.

- Data warehouses
    - Snowflake [general recommendation]
    - Presto
    - Apache Hive
    - Apache Impala
    - Amazon Redshift [general recommendation]
    - Google BigQuery [personal recommendation]
    - Azure Synapse
    - ClickHouse
- Object storage
    - AWS S3 [general recommendation]
    - Azure Blob Storage
    - Google Cloud Storage
    - Apache Ozone
- Cluster computing fundamentals
    - Apache Hadoop [general recommendation]

    - HDFS [general recommendation]
    - MapReduce [general recommendation]
    - Lambda & Kappa architectures
    - Managed Hadoop [general recommendation]
        - Amazon EMR
        - Google Dataproc
        - Azure Data Lake

Note: Most modern data processing frameworks are based on Apache Hadoop and MapReduce to some extent. Understanding these concepts can help you learn modern data processing frameworks much quicker.

- Data processing
    - Batch
        - Apache Pig [general recommendation]
        - Apache Arrow
        - data build tool [personal recommendation]
    - Hybrid
        - Apache Spark [general recommendation]
        - Apache Beam [personal recommendation]
        - Apache Flink [general recommendation]
        - Apache NiFi
    - Streaming
        - Apache Kafka [personal recommendation]
        - Apache Storm [general recommendation]
        - Apache Samza
        - Amazon Kinesis

Note: Hybrid frameworks are able to process both batch and streaming data. Batch data processing is often done by analytical data warehouse applications. See Data warehouses section for more.

- Messaging

    - RabbitMQ [general recommendation]
    - Apache ActiveMQ
    - Amazon SNS & SQS
    - Google PubSub
    - Azure Service Bus

- Workflow scheduling
    - Apache Airflow [personal recommendation]
    - Google Composer
    - Apache Oozie
    - Luigi

Note: Cloud Composer is a managed Apache Airflow service on Google Cloud Platform.

- Monitoring and observability for data pipelines
    - Prometheus [general recommendation]
    - Datadog [general recommendation]
    - Sentry [general recommendation]
    - Monte Carlo
    - Datafold
    - Soda Data
    - StatsD
- Networking
    - Protocols [general recommendation]
        - HTTP / HTTPS
        - TCP
        - SSH
        - IP
        - DNS
    - Firewalls [general recommendation]
    - VPN [general recommendation]
    - VPC [general recommendation]
- Infrastructure as Code
    - Containers
        - Docker [personal recommendation]
        - LXC
    - Container orchestration
        - Kubernetes [general recommendation]
        - Docker Swarm
        - Apache Mesos

        - Google Kubernetes Engine (GKE) [general recommendation]
    - Infrastructure provisioning
        - Terraform [personal recommendation]
        - Pulumi
        - AWS CDK [general recommendation]

- CI/CD
    - GitHub Actions [general recommendation]
    - Jenkins [general recommendation]
- Identity and access management
    - Active Directory [general recommendation]
    - Azure Active Directory
- Data security & privacy
    - Legal compliance [general recommendation]
    - Encryption [general recommendation]
    - Key management [general recommendation]
    - Data governance & integrity