Data Engineer in 2023

- Personal recommendation
- General recommendation
- Cloud based

CS fundamentals

- Basic Terminal usage
- Data structures & algorithms
- APIs
- REST
- Structured vs unstructured data
- Serialisation
- Linux
- Math & statistics

- CLI
- Vim
- Shell scripting
- Cronjobs

- How does the computer work?
- How does the Internet work?
- Git — Version control

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
- Java
- Scala
- Go

Learn how to write clean, extensibile code. Spend some time understanding **programming paradigms and best practices**. Get familiar with an IDE or code editor like **VSCode**.

Testing

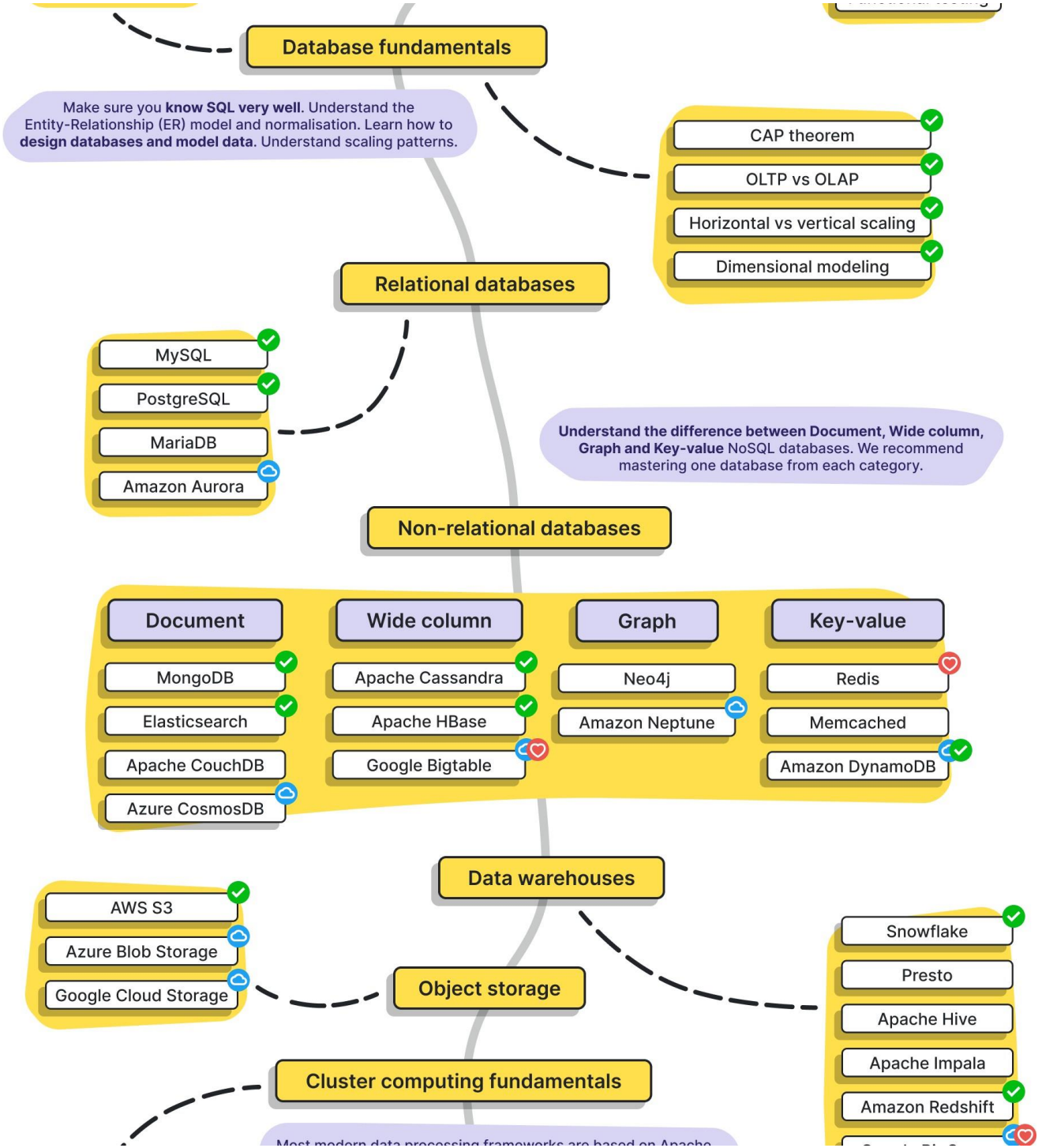
- Unit testing
- Integration testing
- Functional testing

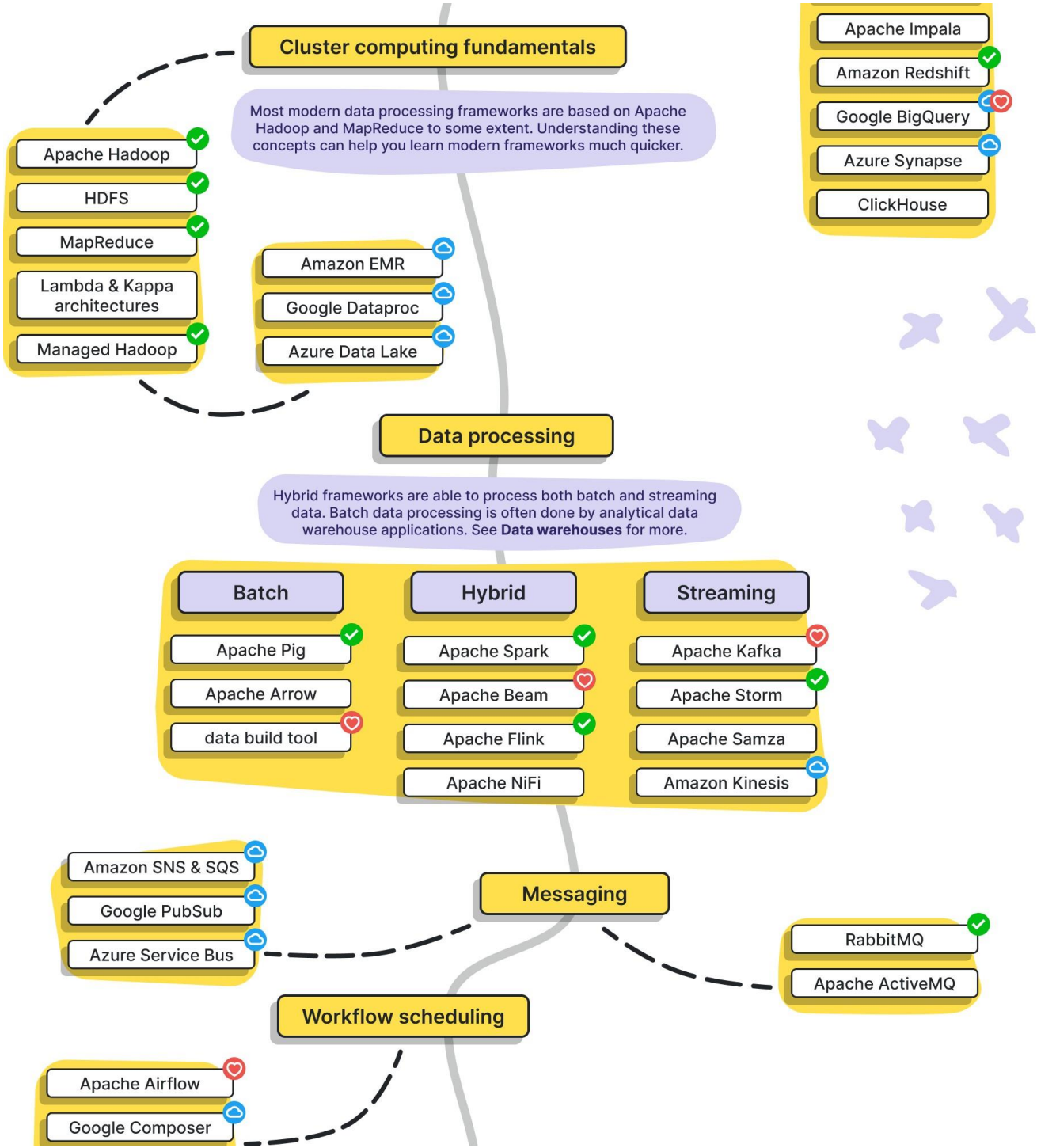
Database fundamentals

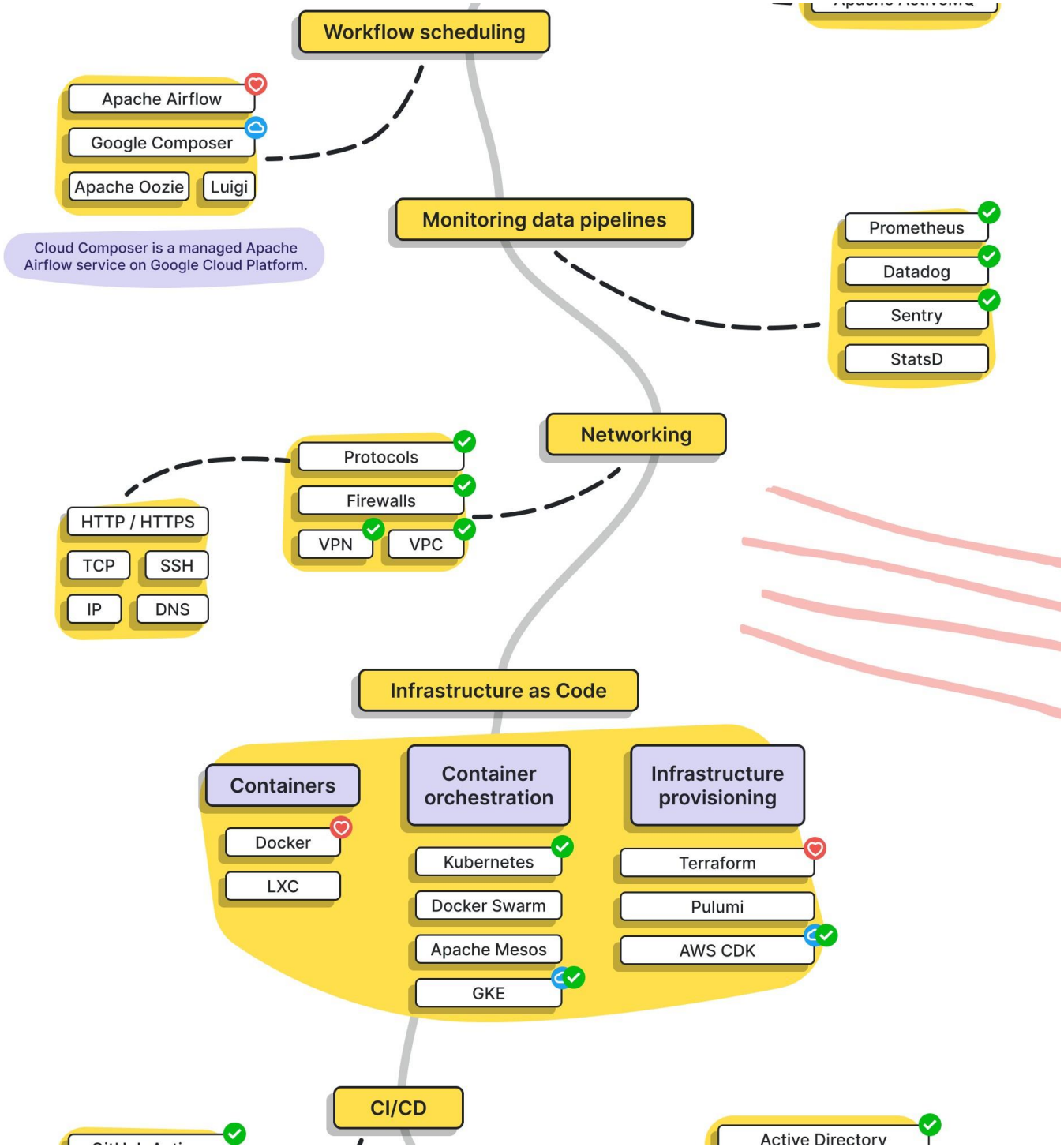
- SQL
- Normalisation
- ACID transactions

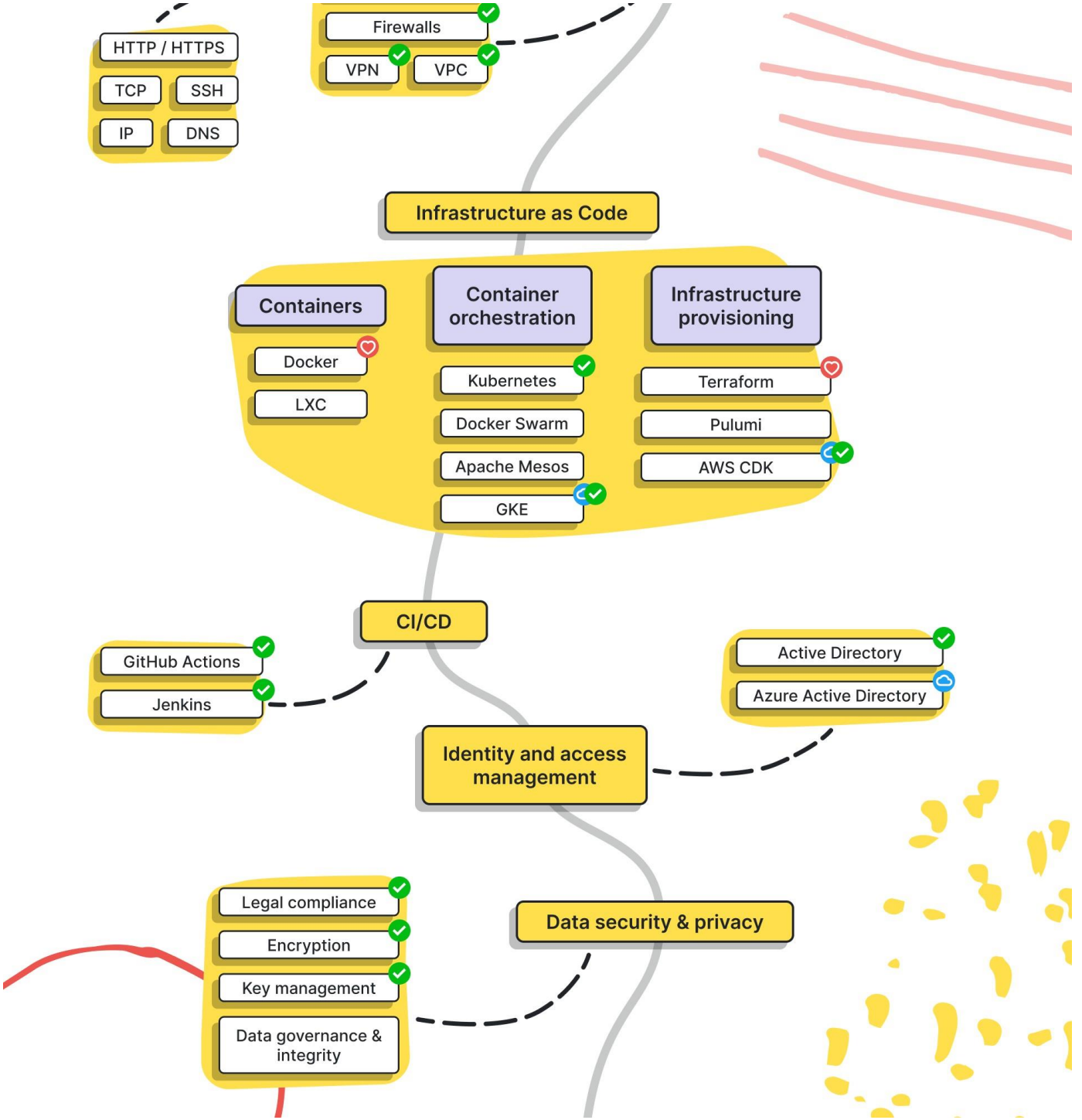
Make sure you **know SQL very well**. Understand the Entity-Relationship (ER) model and normalisation. Learn how to

- CAP theorem


Database fundamentals
Make sure you know SQL very well. Understand the Entity-Relationship (ER) model and normalisation. Learn how to design databases and model data. Understand scaling patterns.
CAP theorem
OLTP vs OLAP
Horizontal vs vertical scaling
Dimensional modeling
Relational databases
MySQL
PostgreSQL
MariaDB
Amazon Aurora
Understand the difference between Document, Wide column, Graph and Key-value NoSQL databases. We recommend mastering one database from each category.
Non-relational databases
Document
MongoDB
Elasticsearch
Apache CouchDB
Azure CosmosDB
Wide column
Apache Cassandra
Apache HBase
Google Bigtable
Graph
Neo4j
Amazon Neptune
Key-value
Redis
Memcached
Amazon DynamoDB
Data warehouses
AWS S3
Azure Blob Storage
Google Cloud Storage
Object storage
Snowflake
Presto
Apache Hive
Apache Impala
Amazon Redshift
Cluster computing fundamentals


Cluster computing fundamentals
Most modern data processing frameworks are based on Apache Hadoop and MapReduce to some extent. Understanding these concepts can help you learn modern frameworks much quicker.
Apache Hadoop
HDFS
MapReduce
Lambda & Kappa architectures
Managed Hadoop
Amazon EMR
Google Dataproc
Azure Data Lake
Apache Impala
Amazon Redshift
Google BigQuery
Azure Synapse
ClickHouse
Data processing
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 for more.
Batch
Apache Pig
Apache Arrow
data build tool
Hybrid
Apache Spark
Apache Beam
Apache Flink
Apache NiFi
Streaming
Apache Kafka
Apache Storm
Apache Samza
Amazon Kinesis
Amazon SNS & SQS
Google PubSub
Azure Service Bus
Messaging
RabbitMQ
Apache ActiveMQ
Workflow scheduling
Apache Airflow
Google Composer


Workflow scheduling
Apache Airflow
Google Composer
Apache Oozie
Luigi
Cloud Composer is a managed Apache Airflow service on Google Cloud Platform.
Monitoring data pipelines
Prometheus
Datadog
Sentry
StatsD
Networking
Protocols
Firewalls
VPN
VPC
HTTP / HTTPS
TCP
SSH
IP
DNS
Infrastructure as Code
Containers
Docker
LXC
Container orchestration
Kubernetes
Docker Swarm
Apache Mesos
GKE
Infrastructure provisioning
Terraform
Pulumi
AWS CDK
CI/CD
Active Directory


HTTP / HTTPS
TCP
SSH
IP
DNS
Firewalls
VPN
VPC
Infrastructure as Code
Containers
Docker
LXC
Container orchestration
Kubernetes
Docker Swarm
Apache Mesos
GKE
Infrastructure provisioning
Terraform
Pulumi
AWS CDK
CI/CD
GitHub Actions
Jenkins
Active Directory
Azure Active Directory
Identity and access management
Legal compliance
Encryption
Key management
Data governance & integrity
Data security & privacy