















⚡ Data Engineering Quick Reference ⚡

💠 Databases









- └  **Relational Database** : A database that stores data in tables with a defined schema
- └  **NoSQL Database** : A database that does not use the traditional relational database model
- └  **SQL** : A language used to interact with relational databases
- └  **MongoDB** : A popular NoSQL database that stores data in JSON-like documents
- └  **Cassandra** : A popular NoSQL database that is designed for high scalability and availability
- └  **Redis** : An in-memory key-value store used for caching and other high-performance use cases
- └  **Amazon RDS** : A managed relational database service provided by AWS

💠 Data Warehousing







- └  **Data Warehouse** : A large, centralized repository of data from various sources used for business intelligence and decision-making
- └  **OLAP** : Online Analytical Processing, used for analyzing data from a data warehouse
- └  **Star Schema** : A type of data model used in data warehousing that consists of a central fact table surrounded by dimension tables
- └  **Snowflake Schema** : A variation of the star schema that uses normalized dimension tables
- └  **Slowly Changing Dimensions (SCD)** : A technique used for managing changes to dimensional data over time
- └  **ETL** : Extract, Transform, Load, the process of moving data from source systems into a data warehouse
- └  **Amazon Redshift** : A cloud-based data warehousing service provided by AWS

⚡ Data Engineering Quick Reference ⚡

💠 Big Data Technologies







- └  **Hadoop** : An open-source framework for distributed storage and processing of large data sets
- └  **Spark** : An open-source distributed computing system used for big data processing and analytics
- └  **Hive** : A data warehousing system built on top of Hadoop for querying and analysis of large data sets
- └  **Pig** : A high-level platform for creating MapReduce programs used for large-scale data processing
- └  **MapReduce** : A programming model for processing large data sets across clusters of computers
- └  **Impala** : A distributed SQL query engine for processing big data sets stored in Hadoop
- └  **Kafka** : A distributed streaming platform used for building real-time data pipelines and streaming applications
- └  **Amazon EMR** : A managed big data processing service provided by AWS

💠 Data Processing






- └  **Data Pipeline** : A set of processes used to extract, transform, and load data from various sources into a destination system
- └  **ETL Tools** : Tools used to automate the extraction, transformation, and loading of data
- └  **Apache Airflow** : An open-source platform used for creating, scheduling, and monitoring data pipelines
- └  **AWS Glue** : A fully-managed ETL service provided by AWS
- └  **Talend** : A popular open-source ETL tool used for data integration and management
- └  **Data Governance** : The process of managing the availability, usability, integrity, and security of data

⚡ Data Engineering Quick Reference ⚡

💠 Data Streaming

- └  **Data Stream** : A continuous flow of data that is processed in real-time
- └  **Apache Kafka** : A distributed streaming platform used for building real-time data pipelines and streaming applications
- └  **Kinesis** : A fully-managed data streaming service provided by AWS
- └  **Flume** : A distributed system for collecting, aggregating, and moving large amounts of log data from different sources to a centralized data store
- └  **Spark Streaming** : An extension of the Spark API used for processing real-time data streams
- └  **Flink** : An open-source distributed stream processing framework used for real-time data processing

💠 Data Visualization

- └  **Tableau** : A popular data visualization tool used for creating interactive dashboards and reports
- └  **Power BI** : A business analytics service provided by Microsoft used for creating interactive visualizations and reports
- └  **D3.js** : A JavaScript library used for creating interactive data visualizations in the browser
- └  **ggplot2** : A popular data visualization package for R
- └  **matplotlib** : A popular data visualization package for Python

⚡ Data Engineering Quick Reference ⚡

💠 Cloud Technologies

- └ ☁ **AWS** : Amazon Web Services, a cloud computing platform provided by Amazon
- └ ☁ **Azure** : A cloud computing platform provided by Microsoft
- └ ☁ **GCP** : Google Cloud Platform, a cloud computing platform provided by Google
- └ ☁ **Docker** : A containerization platform used for packaging and deploying applications
- └ ☁ **Kubernetes** : An open-source container orchestration platform used for automating the deployment, scaling, and management of containerized applications

💠 Data Governance

- └ 🔒 **Data Security** : The process of ensuring data privacy and confidentiality
- └ 🔍 **Data Quality** : The process of ensuring data accuracy, consistency, and completeness
- └ 📄 **Data Lineage** : The process of tracking data from its source to its destination
- └ 🕵️ **Data Discovery** : The process of identifying data assets and their relationships
- └ 📁 **Data Stewardship** : The process of managing data assets and their use

⚡ Data Engineering Quick Reference ⚡

💠 Data Modeling

- └ 📖 **Entity-Relationship Model** : A data modeling technique used to represent the relationships between entities in a system
- └ 📖 **Dimensional Modeling** : A data modeling technique used in data warehousing for creating optimized data structures
- └ 📖 **Data Flow Diagrams** : A diagrammatic representation of the flow of data through a system
- └ 📖 **UML** : Unified Modeling Language, a standardized language used for object-oriented modeling
- └ 📖 **ERD Tools** : Tools used for creating entity-relationship diagrams and other data modeling diagrams

💠 Data Integration

- └ 🔄 **Data Federation** : The process of combining data from multiple sources into a single virtual view
- └ 🔄 **Data Replication** : The process of copying data from one database to another in near-real time
- └ 🔄 **Data Synchronization** : The process of ensuring that data is consistent across multiple systems
- └ 🔄 **Extract, Load, Transform (ELT)** : A data integration approach where data is extracted from source systems, loaded into a staging area, and transformed before being loaded into a target system
- └ 🔄 **Change Data Capture (CDC)** : A data integration technique where changes in source systems are captured and propagated to target systems in near-real time

⚡ Data Engineering Quick Reference ⚡

📦 Data Architecture








- └ 📦 **Data Lake** : A storage repository that holds a vast amount of raw, unstructured data in its native format
- └ 📦 **Data Mart** : A subset of a data warehouse that is designed for a specific business function or department
- └ 📦 **Data Hub** : A centralized repository of data that serves as a single source of truth for an organization
- └ 📦 **Data Virtualization** : A data integration technique that allows data to be accessed and manipulated in real-time without copying or moving it
- └ 📦 **Master Data Management (MDM)** : The process of creating and maintaining a single, trusted view of key business data

📦 Machine Learning






- └ 🤖 **Supervised Learning** : A type of machine learning where the algorithm is trained on labeled data
- └ 🤖 **Unsupervised Learning** : A type of machine learning where the algorithm is trained on unlabeled data
- └ 🤖 **Reinforcement Learning** : A type of machine learning where the algorithm learns from feedback in an environment
- └ 🤖 **Deep Learning** : A type of machine learning that uses neural networks to model complex relationships in data
- └ 🤖 **TensorFlow** : An open-source machine learning framework developed by Google
- └ 🤖 **PyTorch** : An open-source machine learning framework developed by Facebook
- └ 🤖 **Scikit-learn** : A popular machine learning library for Python

⚡ Data Engineering Quick Reference ⚡

📊 Data Science

- └  **Statistical Analysis** : The process of analyzing data to uncover relationships and patterns
- └  **Data Exploration** : The process of identifying patterns and trends in data
- └  **Predictive Modeling** : The process of using data to make predictions about future events
- └  **Time Series Analysis** : The process of analyzing data that is collected over time
- └  **Spatial Analysis** : The process of analyzing data that is related to geographic locations
- └  **Data Visualization** : The process of representing data graphically
- └  **Data Mining** : The process of discovering patterns and relationships in large datasets

📊 Programming Languages

- └  **Python** : A popular programming language used for data engineering and machine learning
- └  **Java** : A popular programming language used for building enterprise-level applications and big data technologies
- └  **Scala** : A programming language used for building big data technologies and data streaming applications
- └  **SQL** : A language used for interacting with relational databases
- └  **R** : A programming language used for statistical computing and data analysis

⚡ Data Engineering Quick Reference ⚡

💠 Cloud Computing Services

- └ 📣 **EC2** : Elastic Compute Cloud, a virtual server provided by AWS
- └ 📣 **S3** : Simple Storage Service, a scalable object storage service provided by AWS
- └ 📣 **Lambda** : A serverless compute service provided by AWS
- └ 📣 **CloudFormation** : A service provided by AWS for modeling and setting up cloud resources
- └ 📣 **Azure VM** : A virtual machine provided by Azure
- └ 📣 **Azure Blob Storage** : A scalable object storage service provided by Azure
- └ 📣 **Azure Functions** : A serverless compute service provided by Azure
- └ 📣 **Azure Resource Manager** : A service provided by Azure for modeling and setting up cloud resources
- └ 📣 **GCE** : Google Compute Engine, a virtual machine provided by GCP
- └ 📣 **Cloud Storage** : A scalable object storage service provided by GCP
- └ 📣 **Cloud Functions** : A serverless compute service provided by GCP
- └ 📣 **Cloud Deployment Manager** : A service provided by GCP for modeling and setting up cloud resources.

⚡ Data Engineering Quick Reference ⚡

💠 Resources

- └ 📖 Data Engineering with Python by Paul Crickard III
- └ 📖 Designing Data-Intensive Applications by Martin Kleppmann
- └ 📖 Data Engineering Cookbook by Andreas Kretz
- └ 📖 Streaming Systems by Tyler Akidau, Slava Chernyak, and Reuven Lax
- └ 📖 AWS Certified Data Analytics Study Guide by Richard Wentk

💠 Useful Technologies

- └ 📦 **Apache Airflow** : A platform used for creating, scheduling, and monitoring data pipelines
- └ 📦 **Apache Kafka** : A distributed streaming platform used for building real-time data pipelines and streaming applications
- └ 📦 **Spark** : An open-source distributed computing system used for big data processing and analytics
- └ 📦 **Docker** : A containerization platform used for packaging and deploying applications
- └ 📦 **Kubernetes** : An open-source container orchestration platform used for automating the deployment, scaling, and management of containerized applications