Data Engineering Lifecycle

- Schema: Defines the hierarchical organization of data
- Schemaless: Application defines the schema as data is written
- Fixed schema: Enforced in the DB to which application writes must conform
- Ingestion:
  - Push: Source system writes data out to a target
  - Pull: Data is retrieved from a source system
- Featurization: Extract and enhance data features useful for training ML models
- BI:
  - Describe a business's past and current state
  - Data is stored in a clean but fairly raw form with minimal postprocessing business logic
- Operational Analytics:
  - Fine-grained details of operations, e.g., live view of inventory
  - Focused on present and doesn't concern historical trends
- Security: Principle of least priviledge
- Data Management: Encompasses the set of best practices DE will use to accomplish the task of managing the data lifecycle technically and strategically
- Data Governance: Ensure quality, integrity, security, and usability of the data collected by an organization
- Metadata:
  - Business: Non-technical questions about who, what, where, and how and provides a DE with the right context and definitions to properly use the data
  - Technical: Describes the data created and used by systems across the DE lifecycle
  - Data Lineage: Tracks the origin and changes to data
  - Schema: Describes structure of data stored in a system
  - Operational: Used to determine whether a process succeeded or failed and the data involved in the process
- Data Accountability: Assigning an individual to govern a portition of data
- Data Quality:
  - Quality tests, ensuring data conformance to schema expectations, data completions, and precision
  - Accuracy: Is the data factuall correct? Duplicated values? Are the numeric values accurate?
  - Completeness: Are the records complete? Do all required fields contain valid values?
- Master Data: Business entities (employees, customers etc.)
- Master Data Management: Practice of building consistent entity definitions known as golden records
- Data Modeling and Design: Process for converting data into usable form
- Data Lineage:
  - Provides a trail of data's evolution as it moves through various systems and workflows
  - Tracks both the systems that process the data and the upstream data it depends on
- Data Integration: Process of integrating data across tools and processes. Happens through general-purpose APIs rather than DB connections
- Orchestration: Coordinating many jobs to run as quickly and efficiently as possible on a scheduled cadence (DAGs)