INTRODUCTION

Data engineering is developing, implementing, and maintaining systems and processes that take in raw data and produce high-quality, consistent information supporting downstream use cases, such as analysis and machine learning.

Data Engineer is an individual who manages the entire life cycle of data engineering from getting the data from the source and ending to serving the data for analysis and use cases for business purposes.

Data Engineer is a hub between the data producers and the data consumers.

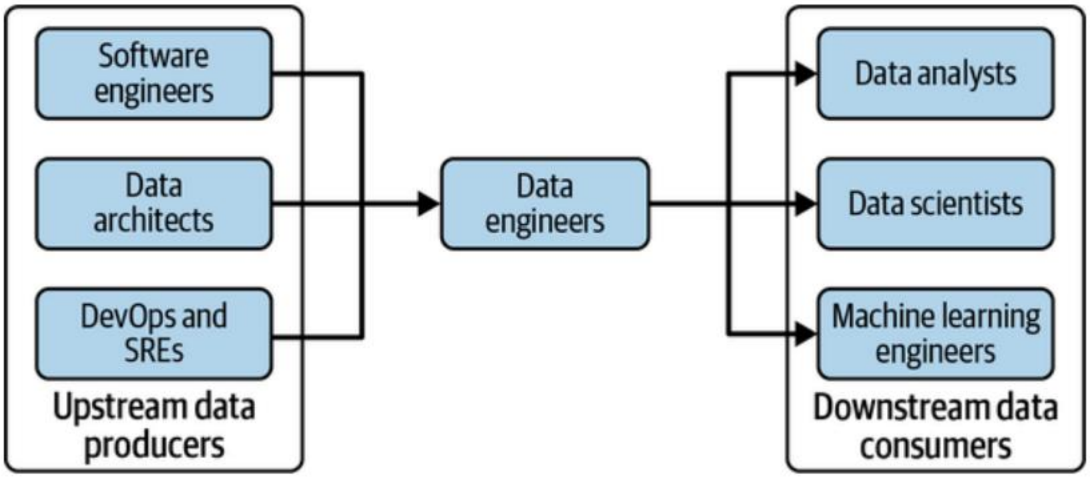
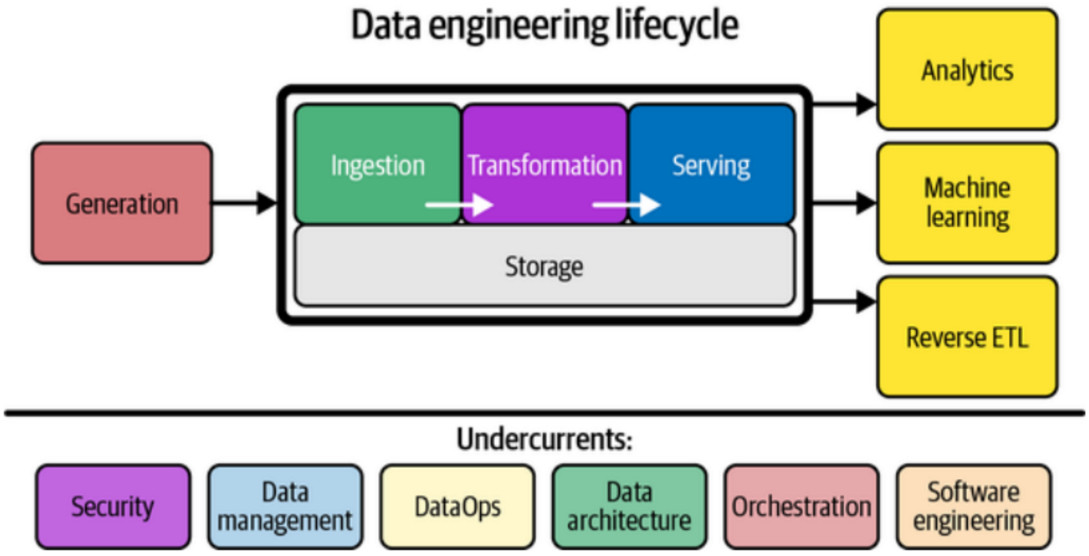


Figure: Stakeholders of Data Engineering

Data Engineering Lifecycle:



The different stages of the data engineering lifecycle are.

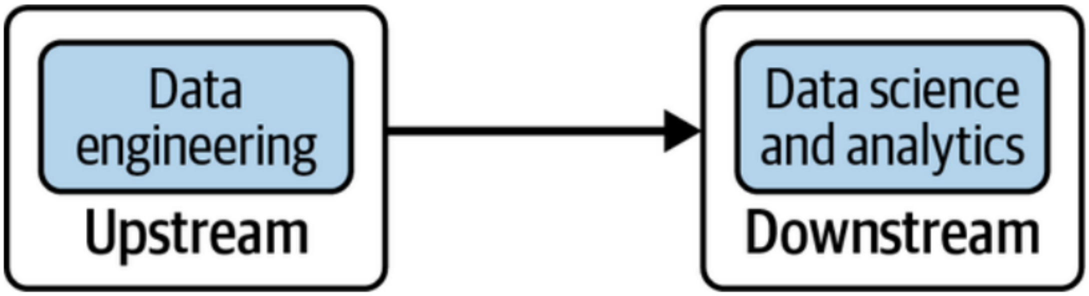
i. **Generation:** This is the first stage where the data is born. Here the data is generated or collected from various resources and can be in any format structured, unstructured, or semi-structured.

ii. **Ingestion:** Here the collected data is ingested into a storage system or a pipeline which involves moving the data from a source to the centralized location for further processing.

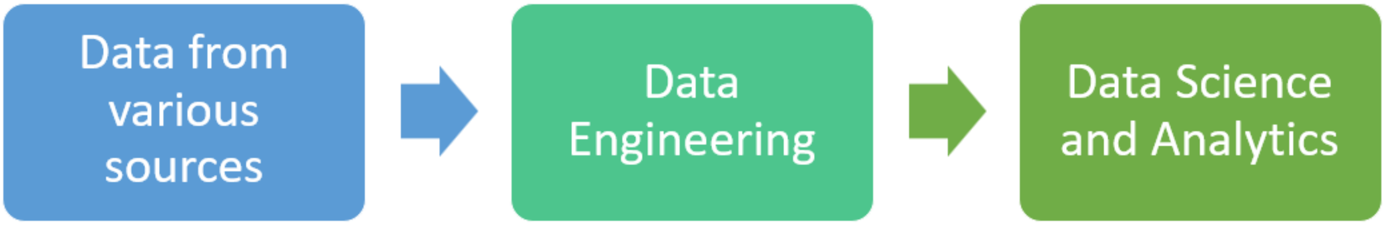
iii. **Transformation:** The ingested data requires transformation to make it suitable for downstream processing which includes data cleaning, filtering, aggregating, and normalizing.

iv. **Serving:** This is the final stage which involves serving the transformed data to the downstream applications or end-users for analysis or consumption.

Data Engineering and Data Science:



Data Engineering is separate from Data Science and Analytics. Data Engineering sits upstream from Data Science. Data Engineers provide the inputs that are used by the Data Scientists who convert these inputs into useful information.



Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud, offering over 200 fully featured services from data centers globally.

Why AWS???...

- **Easy to Use** -- AWS Management Console is used to allow application providers and vendors to host any applications quickly and securely.
- **Flexible** -- It allows the user to select which operating system, database, programming language, and any other services they need.
- **Cost Effective** -- AWS offers flexible pricing models, which allow data engineers to optimize costs based on their needs. It has pay-as-you-go pricing, where the users pay only for the resources they use, reducing upfront infrastructure costs and providing cost efficiency for data engineering projects.
- **Reliable** -- AWS ensures high reliability through multiple availability zones, fault-tolerant architecture, and automated failover mechanisms, ensuring continuous operation of services.
- **Secure** -- Provides features like encryption, access control, and identity management that are crucial for handling sensitive data.
- **Scalable** -- Provides a highly scalable infrastructure that would allow data engineers to easily scale up or scale down their resources based on the demands of their data processing tasks, also accommodates fluctuating workloads.
- **High Performance** -- Using Elastic Load Balancing which distributes the load evenly across multiple resources reducing the bottleneck on a single resource. This way it optimizes the resource utilization thereby increasing the performance.