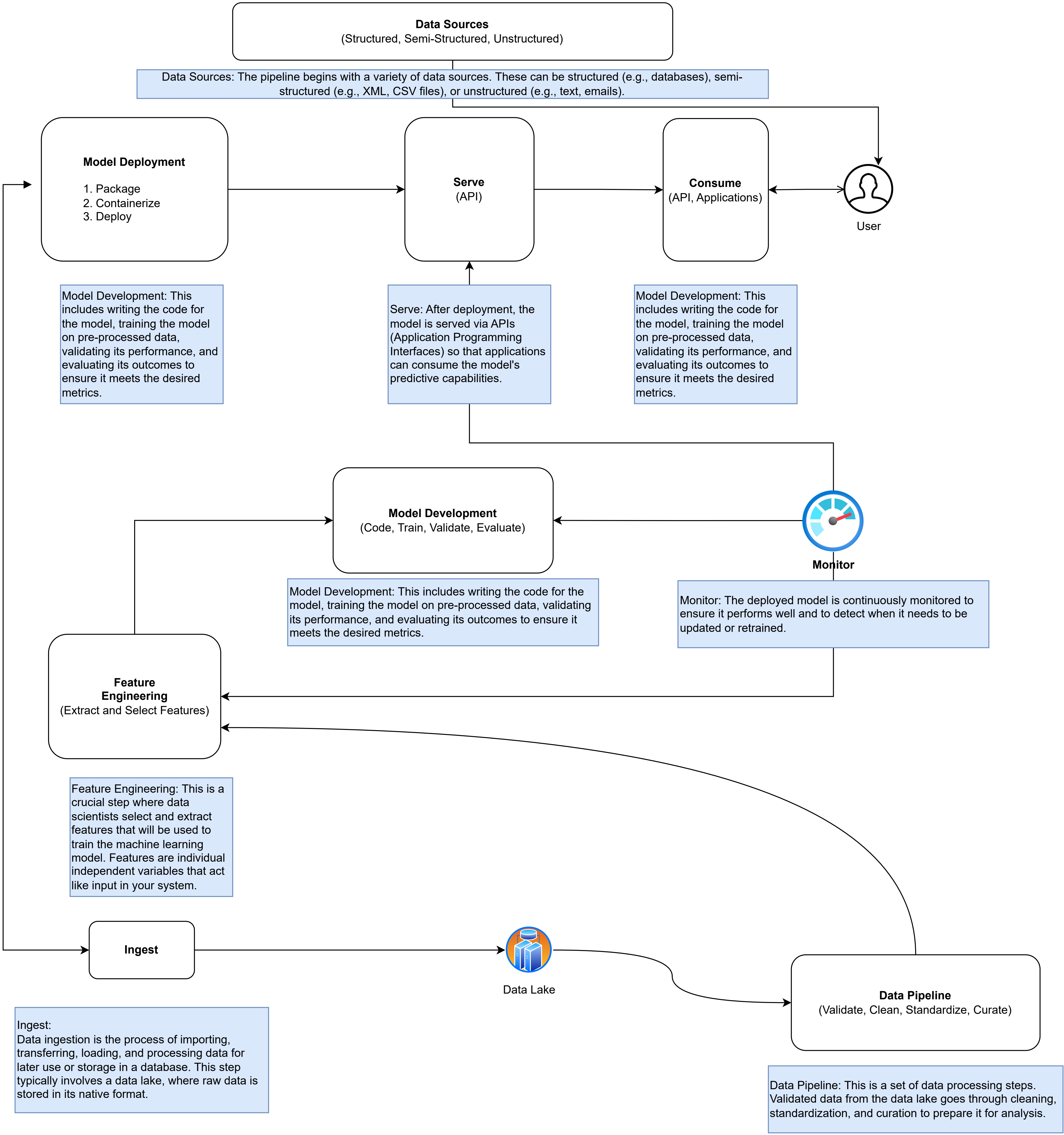
Data Science Workflow



Data Sources
(Structured, Semi-Structured, Unstructured)

Data Sources: The pipeline begins with a variety of data sources. These can be structured (e.g., databases), semi-structured (e.g., XML, CSV files), or unstructured (e.g., text, emails).

Model Deployment

1. Package
2. Containerize
3. Deploy

Serve
(API)

Consume
(API, Applications)

User

Model Development: This includes writing the code for the model, training the model on pre-processed data, validating its performance, and evaluating its outcomes to ensure it meets the desired metrics.

Serve: After deployment, the model is served via APIs (Application Programming Interfaces) so that applications can consume the model's predictive capabilities.

Model Development: This includes writing the code for the model, training the model on pre-processed data, validating its performance, and evaluating its outcomes to ensure it meets the desired metrics.

Model Development
(Code, Train, Validate, Evaluate)

Monitor

Model Development: This includes writing the code for the model, training the model on pre-processed data, validating its performance, and evaluating its outcomes to ensure it meets the desired metrics.

Monitor: The deployed model is continuously monitored to ensure it performs well and to detect when it needs to be updated or retrained.

Feature Engineering
(Extract and Select Features)

Feature Engineering: This is a crucial step where data scientists select and extract features that will be used to train the machine learning model. Features are individual independent variables that act like input in your system.

Ingest

Data Lake

Data Pipeline
(Validate, Clean, Standardize, Curate)

Ingest:
Data ingestion is the process of importing, transferring, loading, and processing data for later use or storage in a database. This step typically involves a data lake, where raw data is stored in its native format.

Data Pipeline: This is a set of data processing steps. Validated data from the data lake goes through cleaning, standardization, and curation to prepare it for analysis.