MLOps Process to manage full lifecycle for a Large Language Model
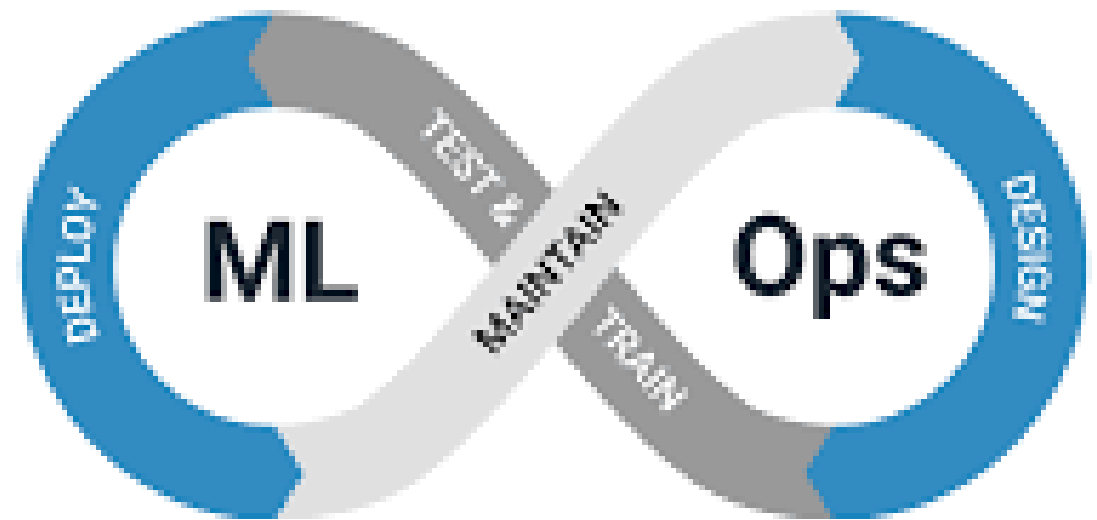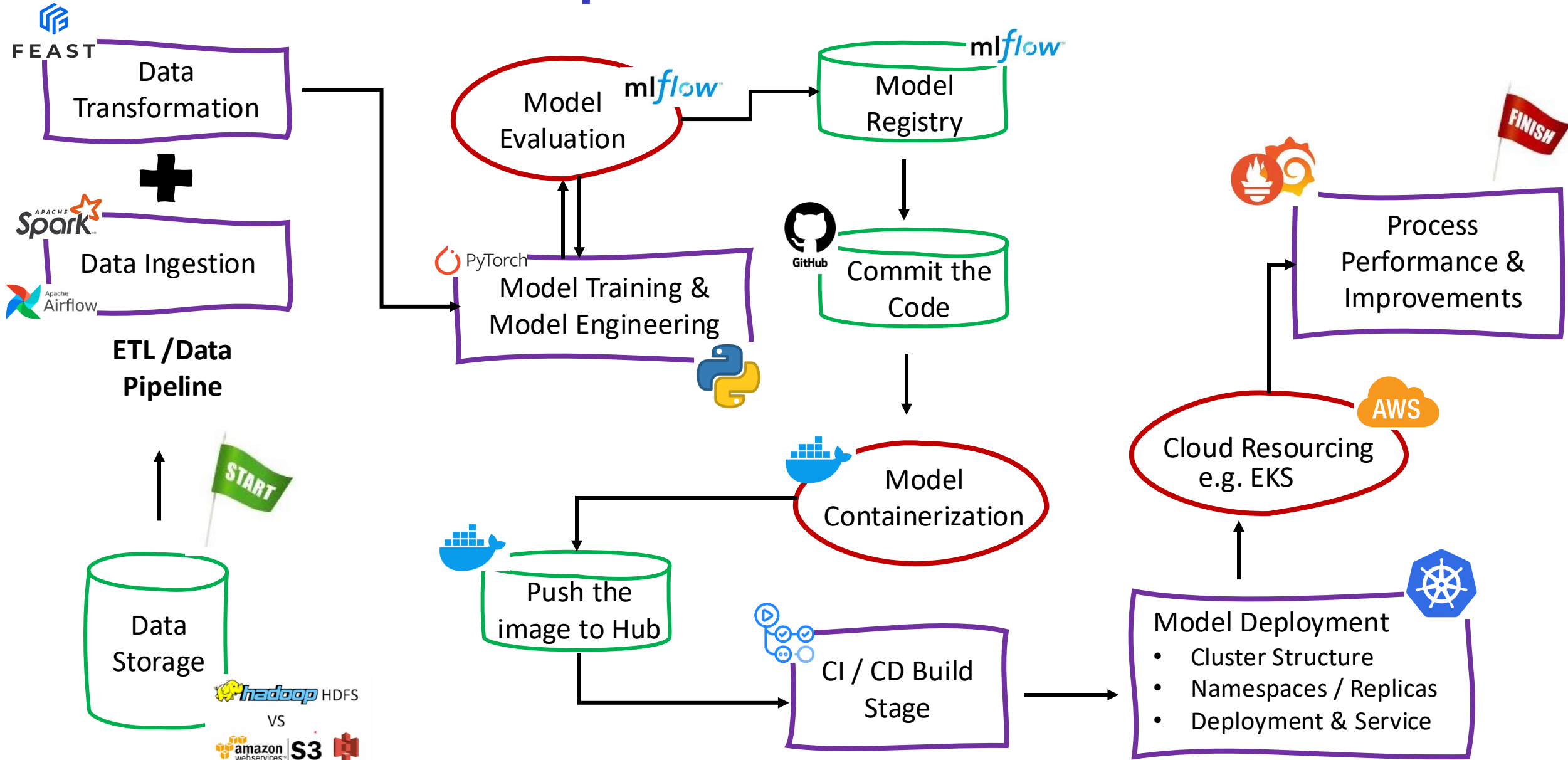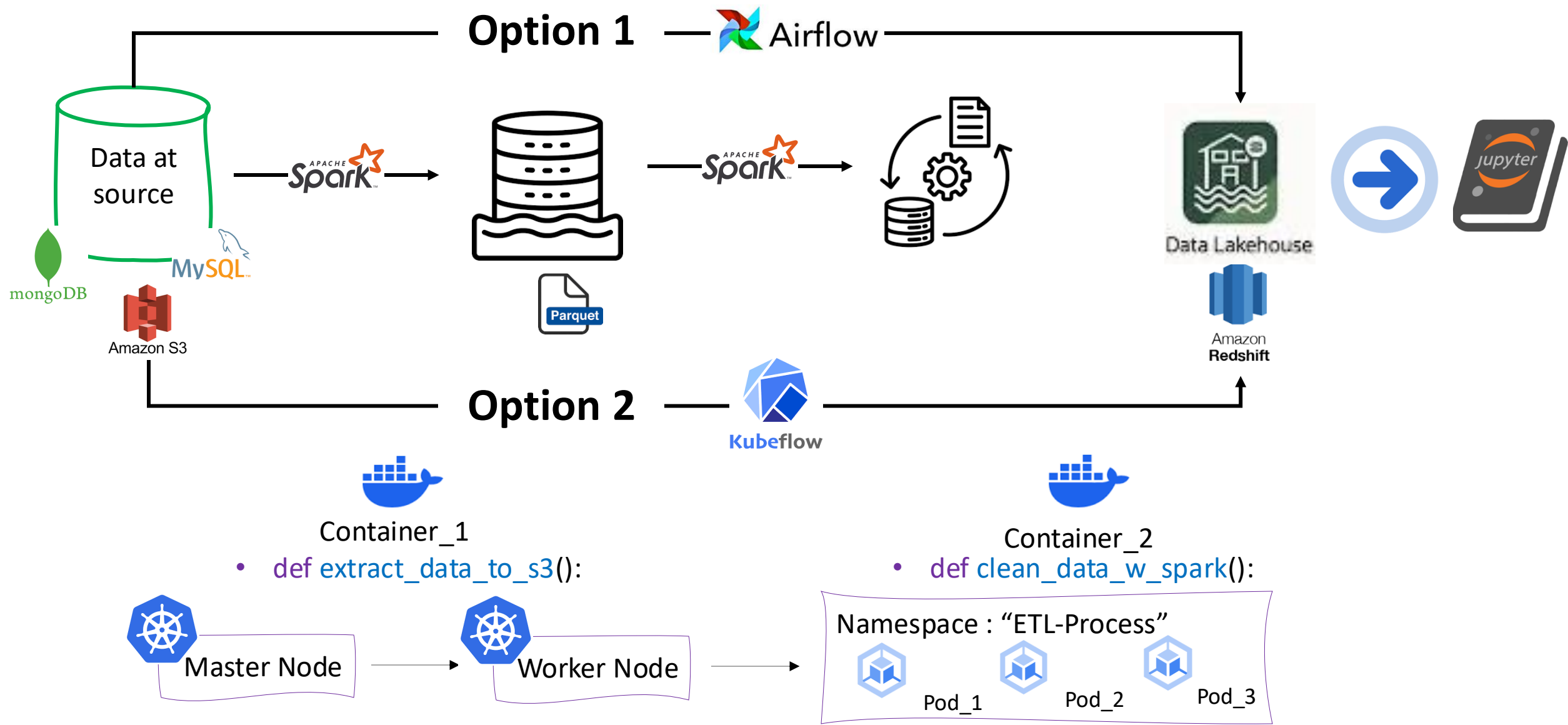
MLOps Process Overview
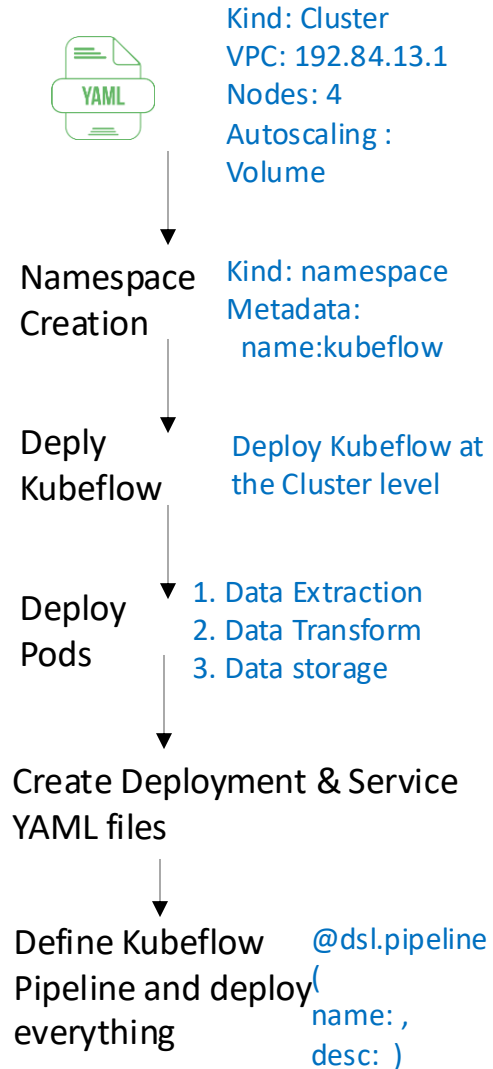
Step 1 – Automate Data Pipelines & Transformation Process
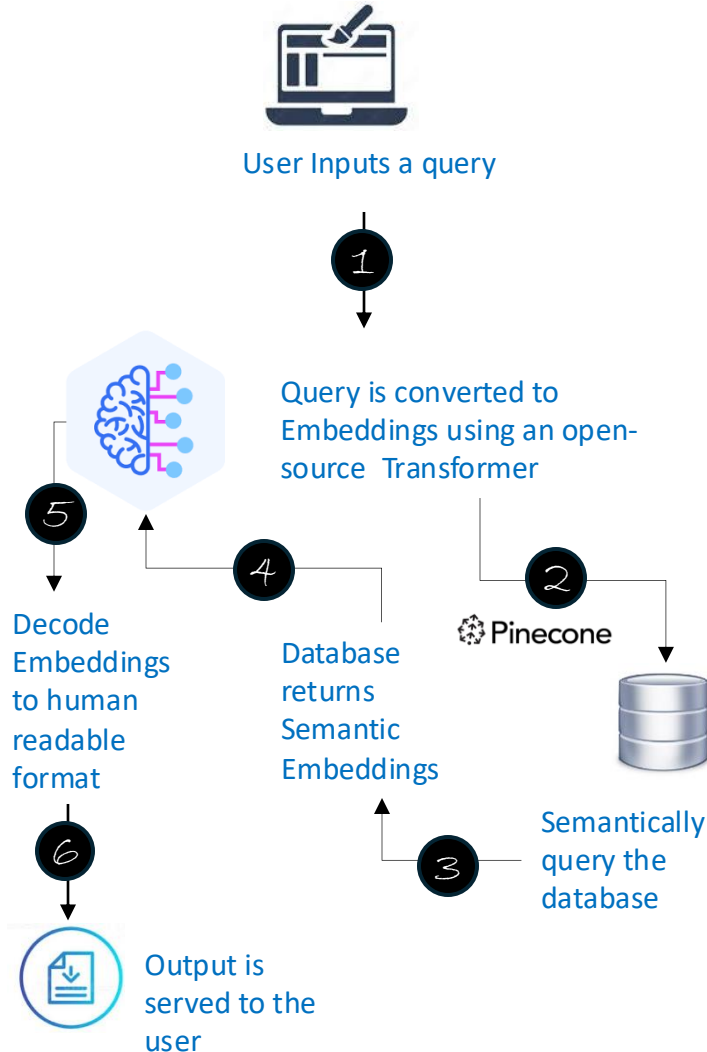
Step 2 – Building, Testing & Deploying the model manually

From Step 1 - Create K8s Set-up

Kind: Cluster
VPC: 192.84.13.1
Nodes: 4
Autoscaling :
Volume

Namespace Creation

Kind: namespace
Metadata:
 name:kubeflow

Deply Kubeflow

Deploy Kubeflow at the Cluster level

Deploy Pods

1. Data Extraction
2. Data Transform
3. Data storage

Create Deployment & Service YAML files

Define Kubeflow Pipeline and deploy everything

@dsl.pipeline
(
name: ,
desc:)

Step 1 – Data Extraction and Transformation via Kubeflow route

Model Building

User Inputs a query

1

Query is converted to Embeddings using an open-source Transformer

2

Pinecone

Semantically query the database

3

Database returns Semantic Embeddings

4

Decode Embeddings to human readable format

5

6

Output is served to the user

Model Deployment

Data Tests

mlflow

ML Infrastructure

Code

Model Test

Running System

Unit Tests

Integration Tests

Model Registry

mlflow

Model Containerization

Model Registry

Amazon ECR

Deployment

Pods   Pods   Pods   Pods

Service

Step 3 – Automation… (CI / CD Pipeline)

Level of automation of Data, ML Model, & Code pipelines determines the maturity of the ML process.