IIITH-SDES Capstone Project

NEWS ARTICLES CLASSIFIER

Group Members:
Korrapati Venkata Monika
Smriti Chinta

Project Documentation

Objective-To design a Machine Learning/Deep Learning Project by developing a solution which demonstrates complete end-to-end pipeline and which helps in understanding all stages of Machine Learning Project Lifecycle.

Problem Statement-Classify News Articles into categories:-Design a system which can classify incoming news articles and appropriately tag the corresponding category. Develop a data pipeline which includes the all the following stages of Machine Learning Project Life Cycle –Data Ingestion,Data Preparation,Data segregation ,Model Training ,Model Deployment,Model Prediction.

Project Milestones-1)Data Ingestion-The objective of this project is to source new data to re-train the model.Use a Publisher-Subscriber model to collect data and push it to the store. Create a Kafka topic and use it to queue cleaned responses from all sources. The sink for the stream should be MySQL/ MongoDB.

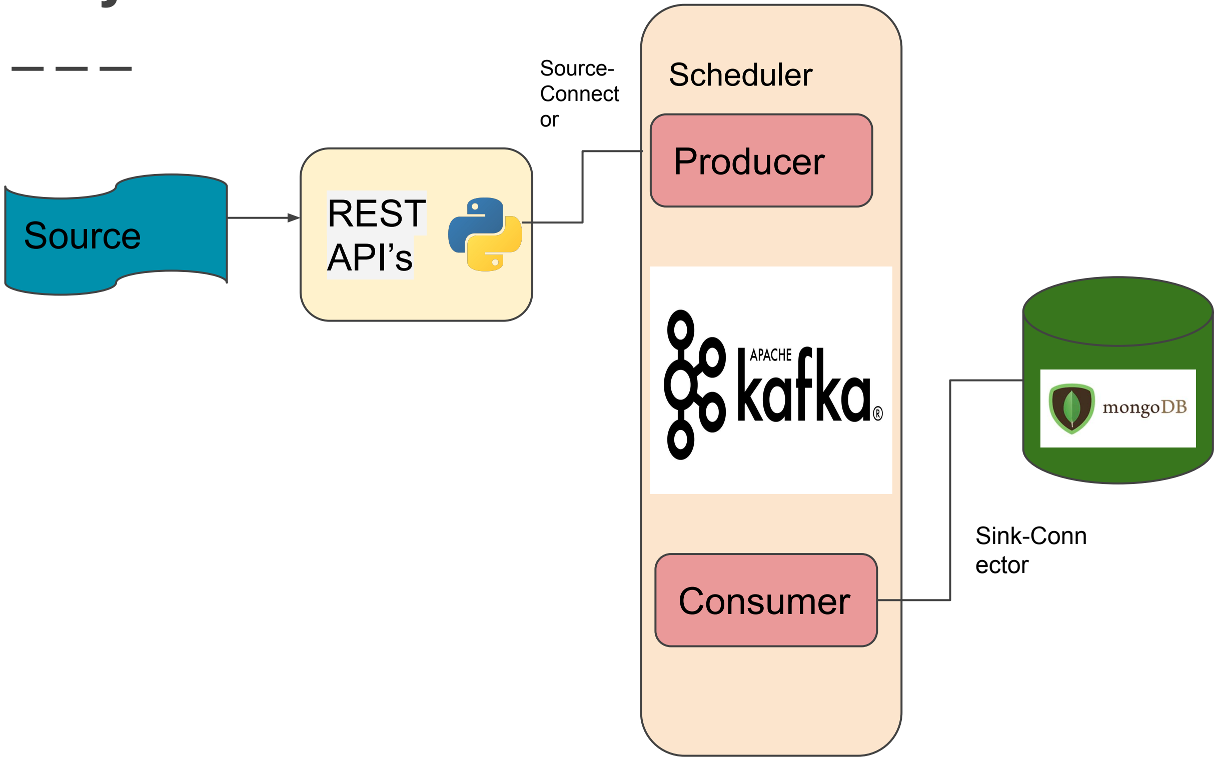
2)Data Preparation,Data Segregation,Model Training-Load the data from “raw_data” source (MySQL/ MongoDB) into Spark by using relevant connector for PySpark.Perform data cleaning and preprocessing, followed by segregation to train and test datasets

3)Model Prediction-A separate classifier project picks up the trained model either from a location or from the model registry, and exposes it for prediction

4)Model Deployment-Deployment can be orchestrated by using docker-compose.

ML Tools Used-1)Pycharm as Python IDE 2)Virtual Environment 3)MongoDb as Database 4)Pyspark for Streaming Process 5)Apache Zookeeper + Kafka for message queue/ streams.

Project Architecture


Source
REST API’s
Source-Connect or
Scheduler
Producer
APACHE kafka
mongoDB
Sink-Conn ector
Consumer