revest

Overview

The following technical task is split into multiple parts all related to each other. The main objective of the task is to assess the candidate's foundational knowledge and skills in data engineering and MLOps.

There are general requirements and expectations for the task as a whole and for each part separately, but feel free to extend the scope of both or any parts within the given time to submit. Any additional features presented are welcome and positively regarded.

Expectations

1. For each part of the assignment, deliver code, documentation and dependencies necessary to run your solution and evaluate accordingly.
2. The code should well structured, clean, and easy to read.
3. Machine learning model training or fine-tuning is not the goal of this task. For any part of the assignment that involve training or using a pre-trained model, hyperparameter tuning and testing multiple solutions to achieve the highest performance possible is not the goal of the exercise.
4. Use any databases, libraries or frameworks you see fit. The choice of suitable databases or storage solutions is part of the evaluation.
5. We don't mind using existing code or getting some help from LLMs to accelerate your task, but you should understand every line of code submitted.
6. Preparing Dockerfiles and Docker Compose stacks for the deliverables is **highly preferred**.

Tasks

The following parts of the technical tasks are based on the provided `sales.csv` file attached to this technical task. The provided dataset is originally found on Kaggle, and represents sales records from a retail store.

Part I: Data Warehouse

Based on the sales stores records provided, you're asked to setup a data warehouse to store the records for **analytical** purposes. The data warehouse will be eventually used to query for insights and analytics of the store's sales. You are asked to do the following:

1. Create a script that ingests and pre-processes the data based on the different datatypes available into a database of your choice.
2. Write queries that extracts the following information (you can deliver them as raw queries. Executing and visualizing the results is a plus):
    1. What's the average total value of an `order`
    2. What is the average total revenue for each month of the year
3. Write a script that exports the datawarehouse to a `parquet` file.

Part II: Model Deployment

You are now asked to deploy a machine learning service that does the following:

1. The service receives a request with a product ID, the service then predicts what other products might be purchased and recommended to the user to buy as well.
2. The service responds with the recommended products based on the input

Since the recommendation model isn't the goal of this task, always recommend a static set of product IDs, eg: `recommended_product_ids = [1, 2, 3]`

You are asked to implement the following:

1. Create a simple REST API with the framework/package of your choice
2. Create a Dockerfile for the service
3. Implement logging for the service, the logging should include the request data and recommended products (static). You can use any suitable database of your choice.
4. Create a `docker-compose` file to deploy both the service and logging database/solution.
5. Adding a caching layer is considered a plus but not required.