TECKNOWORKS Internship | Technical Project: Data Engineering & AI Intern

Assessment, 1st phase of the process

Overview

This technical project is designed to evaluate your technical skills, creativity, and problem-solving abilities in both Data Engineering and Artificial Intelligence. The project will help us assess how well you understand the concepts and tools you'll be working with during the internship. The project should take no more than 2 to 3 hours to complete.

Project Description

You are tasked with designing a simple data pipeline using Microsoft Fabric and performing data science tasks using Python. Most likely, you don't have experience with Microsoft Fabric. We're interested in how your mind works when you have to face a challenge like this. Microsoft Fabric is a comprehensive data analytics platform that integrates data engineering, data science, data warehousing, and business intelligence in one environment. By using Microsoft Fabric, we expect our interns to gain hands-on experience with a tool that Microsoft expects to be widely adopted in the industry. This exposure not only makes you more competitive in the job market but also prepares you for real-world scenarios where you are likely to encounter similar integrated platforms.

Microsoft Fabric is a paid tool, so in order to solve these exercises, you have to simply read the documentation to have an introduction to the tool, watch a few tutorials if needed and answer the questions based on your findings and your prior knowledge from University courses. You don't need to access Fabric for this basic assignment. You don't have to implement the solution, only to design it based on the documentation that you read. You can use text descriptions or any type of visual representation to answer the questions (e.g. Dataflow diagrams) for Section 1. For Section 2 you can implement the solution for free using Google Colab or any other local python environment.

The project is divided into two main sections: Data Engineering and Artificial Intelligence.

Section 1: Data Engineering with Microsoft Fabric

Scenario

You are a new data engineer at a retail company. The company has a dataset containing information on customer transactions. Your task is to design and implement a data pipeline that ingests, processes, and stores this data using Microsoft Fabric. The dataset is available here: https://www.kaggle.com/datasets/fahadrehman07/retail-transaction-dataset?resource=download

Retail Transaction Dataset

Retail Trends: A Deep Dive into Transactional Data

www.kaggle.com

***** Task 1: Data Ingestion and Storage

Question: First, you have to bring this data into Fabric. How do you envision this? What will be the expected steps? How exactly will it be stored? Is the data in a suitable format and size to ingest in Fabric? What can you do related to storage in order to improve the performance of the future data queries if the size of the dataset is expected to grow in time (hint: e.g. partition the data)?

(Answer with text descriptions or diagrams)

****** Task 2: Data Transformation and Processing

Question: Like most of the datasets you'll find in your practice, this dataset has some data issues that should be corrected. You need to find a way within Microsoft Fabric to perform the following transformations:

Separate TransactionDate column in 2 distinct columns for Date and Time.

Aggregate the TotalAmount spent by each customer per month.

Replace the "Home Decor" values from the ProductCategory column with "Home Products".

Create a new column HighValueCustomer that is a boolean column that assigns True or False based on your own rule. Think about a rule with a logic that makes sense in the context.

Load the transformed data into a new table in your Data Lake.

How do you envision solving this based on your research on Microsoft Fabric? No implementation is needed.

(Answer with text descriptions, images or diagrams).

*** Task 3: Data Visualization

Question: How will you use Microsoft Fabric to create a simple dashboard? The dashboard will show:

A chart for total monthly sales by product category & a chart showing the number of HighValueCustomers over time.

Can you create a simple design/sketch in any form to present to your supervisor and colleagues about how do you evision this Dashboard to look like? Make sure you choose the correct graph type for each chart in order to correctly tell the data story.

(Answer with text descriptions and a visual representstion of the dashboard)

Section 2: Artificial Intelligence with Python

Scenario

Continuing from the previous scenario, you need to develop a simple predictive model to understand customer behavior. Based on their transaction history, the goal is to predict whether a customer is likely to become a High-Value Customer in the next month. Use the same dataset as in Section 1.

* Task 1: Data Preparation

Load the data you created into a Pandas DataFrame. Create new features such as the total amount spent in the last 3 months, the average transaction amount, and the number of distinct product categories purchased, etc.

Create a target variable IsHighValueNextMonth which is True if the customer becomes a HighValueCustomer in the next month, otherwise False. Clean the dataset if needed.

** Task 2: Model Building

Split the data into training and testing sets.

Train a simple model of your choice (e.g., Logistic Regression or Decision Tree) to predict IsHighValueNextMonth.

Evaluate the model using appropriate metrics such as accuracy, precision, and recall.

*** Task 3: Creative Analysis

Propose at least one creative feature that could improve the model's performance. Implement this feature and re-evaluate the model.

Suggest a potential business use case for the predictive model in a retail environment.

Submission Guidelines

Deliverables:

- ✓ For Section 1: A PDF or document containing text descriptions and screenshots of your designs for the workflow, data structure, and dashboard.
- ✓ For Section 2: A Jupyter notebook in any environment (or Python script) containing your code, with clear explanations for each step.

Deadline: **4 days** from the time you receive the project.

Evaluation Criteria

We expect you to have good research skills for this project and care about the originality of your approach in both the data pipeline and AI model.

We want to see that you can quickly understand a tool such as Microsoft Fabric and design a simple solution just by reading the documentation and understanding what and how it is possible. We expect to see clear explanations and documentation of your work and that you think about the subtle aspects of your work. If you want to provide a more in-depth answer or solution or to go the extra mile on any of the scenarios, please don't hesitate to do so.

This project is designed to be challenging but also an opportunity for you to showcase your skills and creativity.
Good luck! 😊