**1. Introduction**

This test is designed for candidates to demonstrate their knowledge of good software and data engineering practices. Please take the time to read through this document before starting the coding.

Before the session

* Setup your local Python env and IDE.
* Ensure you are able to run some sample tests in either pytest or unittest
* Read this document and note down any questions you might have on the day.
* Please consider writing proper unit tests using
* Explore the data provided in the attached zip file
* Choose either Apache Spark or Pandas as your solution framework
* This exercise should not take more than 60-90 minnutes.

Instructions

The purpose of this exercise is to complete a small data pipeline that aggregates and transforms input data according to requirements driven by the data science team. The actual format of the output will be shared during the exercise.

The aim of this pairing exercise is to write code that processes existing data sources so that it meets requirements. Attention should be given to good engineering practises like automated testing as far as possible within the time available.

As part of solving this, we are looking to assess:

* Your problem-solving approach.
* Your ability to turn your solution into working code and choosing appropriate technology.
* How you structure and test your code.

**2. Requirement**  
User Story

|  |
| --- |
| As a data scientist I want to be able to consume a data source that contains information about how many times each of our customers buys our products in a given period, so that I can predict what they will buy next. |

Customer Shopping Patterns

The task involves developing a data pipeline to complete the user story above using sample data sources that will be provided.

Our client is a major high street retailer that handles millions of transactions each day. Their data science team has reached out to our data engineering team requesting we pre-process some of the data for them at scale so that they can make better use of it in their downstream algorithms. They would like us to deliver this data weekly.

The input data sources are comprised of customers (in CSV format), transactions (in JSON Lines format) and products (in CSV format). Their details are presented below:

Customers

Contains information about customers such as the customer id and the date when they joined:

|  |  |
| --- | --- |
| customer\_id | loyalty\_score |
| C1 | 7 |

Transactions

Is an ever-increasing data source that currently contains 2 years of transactions.

Each transaction contains the customer id, details of what products they purchased and the date of purchase:

{"customer\_id": "C1", "basket": [{"product\_id": "P3", "price": 506}, {"product\_id": "P4", "price": 121}], "date\_of\_purchase": "2018-09-01 11:09:00"}

Products

Contains information about products such as the product id, product description and category:

|  |  |  |
| --- | --- | --- |
| product\_id | product\_description | product\_category |
| P100 | red trousers | C |

Acceptance Criteria

The output data source should contain information for every customer and has the following fields:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| customer\_id | loyalty\_score | product\_id | product\_category | purchase\_count |
| C1 | 7 | P2 | baw | 11 |
| C1 | 7 | P3 | caw | 5 |
| C2 | 4 | P9 | caw | 7 |

Further Implementation Details

The code and design should meet the above requirements and should consider future extension or maintenance by different members of the team.