Introduction

At Street Group, we handle and process a lot of data on properties and the property market. This often involves data pipeline design, implementation and management. This technical task will ask you to design and build a data transformation pipeline.

We don't expect comprehensive testing / documentation for your solution, a follow up conversation around your approach can be discussed in the technical interview so please don't feel it's necessary to invest time outside the core task.

Task Brief

The business wants to collect and analyse data on property transactions (the sale and purchase of a property). The data for UK property transactions is publically available for download in CSV format from the Land Registry here
https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Each row of the data is a unique property transaction, if a property has been bought/sold multiple times in the timeseries the data covers there will be multiple rows for the same property, one per unique transaction.

We would like you to convert the data to newline delimited JSON format grouping the transactions by property with the output having a single JSON object per property with each property object containing an array of transaction objects for the property.

The exact format of the property and transaction json objects is not fixed and can be defined by you but should include all appropriate fields from the dataset relevant to the property / transaction with relevant field naming.

The dataset does not include a unique property ID (the unique IDs given are for the transaction) and you will therefore need to construct an appropriate unique property ID from the property details, you may choose the appropriate method to construct a property ID from the property details.

Our preferred method (and the one we use in production) for defining data transformation pipelines is Apache Beam and we would like to see solutions which define the above transformation using Apache Beam in Python 3, however, we would be happy to see any other appropriate implementation which would be suitable for a dataset of this size (the full dataset is 4.3GB)

Non Functional Requirements and Tips

- Please use a public github repository to share your code.
- It's ok to use a sample of the data, not the entire 20+ million transactions. We would like your solution to scale to processing the full dataset if necessary (you may want to use a subset over a longer time period than using a single month to ensure you have some properties with multiple transactions) but we would prefer a solution which works on a smaller subset of data than no solution.
- Tests are a good idea if they add value but it's more important to us that you produce a working solution than extensive tests / documentation for this task.

Resources

Data Definition Information: https://www.gov.uk/guidance/about-the-price-paid-data#download-option
Data Downloads: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads
Apache Beam Documentation: https://beam.apache.org/documentation/sdks/python/