Senior Data Engineer Exercise Brief

July/August 2023

The following exercise is designed to test your level of expertise with using APIs to ingest data, querying the dataset using SQL and analytical/critical thinking. During the interview, you will be asked to present how you extracted data using the API, the data structures you setup in the database of your choosing and the SQL you created to address the questions below. This section of the interview should last 30-45 mins.

**API Details** (Extract)

Official documentation:  <https://www.nomisweb.co.uk/api/v01/help>

The page provides the necessary information for getting started with the Nomis API. API key is not necessary as API does allow for guest requests, but these are limited to 25,000 cells. To acquire an API key, you will need to register with Nomis and an API key will be provided in the form of a Unique ID.

*Tip:*

Constructing the API request can present some challenges as the URI takes internal codes to reference data tables e.g. QS201EW (ethnicity data 2011) table - internal Nomis code for this table is NM\_522\_1. The internal code can then be supplies into a ‘summary’ URI that directs you to a page with a detailed look at arguments available to that dataset. <https://www.nomisweb.co.uk/api/v01/dataset/NM_522_1/summary>

Target Dataset: 2021 Census

NOTE: if unable to figure out the internal code reference for 2021 ethnicity data, use the code for 2011 data (NM\_522\_1).

**Database** (Load)

Loading data from JSON/CSV (depending on API query you choose) into a database of your choice, such as PostgreSQL or mysql.

NOTE: BigQuery, which is used by JLR, does offer a free tier, but this is subject to limitations on storage and query capacity and does require billing details in case these limits are exceeded.

**SQL** (Transform)

Use SQL to determine the following:

* Which are the 5 most and least diverse districts in England, Wales, Scotland and Northern Ireland? (The definition of what constitutes a “diverse” district is your decision)
* For each of JLR’s UK sites, in their surrounding areas, what is % of the population by ethnicity?
* What is the ethnicity split for all districts combined?

Also answer the following questions:

* Was any data cleansing required?
* Are there any other observations of interest from the datasets?