# **Batch Pipeline**

## Background

You have been asked to create a batch pipeline to take into consideration 2 sources of ingested data, join them and deliver a denormalised table/model into Snowflake (or any other database of your choice).

#### Users

One of the sources consists of an operational database containing information about users. A user has 2 properties:

- An id, uniquely identifying each user. Example: 1234
- A postcode, indicating where a user is at the moment. This attribute may change regularly based on the user's location. Example: SW19

We have an *extract* process which consumes the users' data **on a daily basis** around midnight (00:00). The process **fully** extracts users data, landing the data on a table within the Data Warehouse named "users\_extract". This table is fully truncated/reloaded on each execution.

Volume: we have up to 20M users

#### **Pageviews**

The other source consists of an operational database containing information about pageviews to a website. A pageview has 3 properties:

- A user\_id, uniquely identifying a user. This matches the id on the users table. Example: 1234
- An url of the page being visited. Example: www.website.com/index.html
- A pageview\_datetime when the pageview occurred. Example: 2019-10-11 14:55:23

We have an *extract* process which consumes the pageviews' data **on an hourly basis**. The process **incrementally** extracts pageviews data, landing the data on a table within the Data Warehouse named "pageviews\_extract". On each execution of the extract process, this table is fully truncated and subsequently loaded only with the pageviews data relative to the previous hour.

Volume: on any given day, we may have 100M website pageviews

## **Data Warehouse Model/Pipeline**

Our end goal is to build the Data Warehouse tables/structures which will allow our BI tool to easily and in a performant way answer 2 questions:

- Number of pageviews, on a given time period (hour, day, month, etc), per postcode based on the current/most recent postcode of a user.
- Number of pageviews, on a given time period (hour, day, month, etc), per postcode based on the postcode a user was in at the time when that user made a pageview.

# Requirements

### Part 1

- Design the Data Warehouse schemas and tables which will hold all the data. Focus on logical and physical layers.
- Implement a *Transform* pipeline which will define the schemas and potentially
  materialise any data allowing us to answer our queries. If you have dbt experience,
  please use a dbt project.
- Implement or propose a mechanism for scheduling the *Transform* pipeline
- Source code should be provided via a publicly accessible Github repository
- Provide basic documentation to run the pipeline, along with any other documentation you think is appropriate.
- Be prepared to explain your choices.