The sources of data that you will extract from.

Data source 1: Google, Amazon and more Employee Reviews web scraped from Glassdoor

Data source 2: Google Job Skills

The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

* Data source 1
  + has data for 5 companies, since we are joining it with a Google specific dataset, we filtered out all other companies
* We wanted to treat data source 2 like a key (Ex: get the “job –title” from data source 1 and return the corresponding preferred qualifications from data source 2). In order to make it a cleaner data set:
  + We removed all the duplicates (duplicate job postings meant same information across all rows except for location)
  + Dropped null values

The type of final production database to load the data into (relational or non-relational).

The final tables or collections that will be used in the production database.

Sources:

https://www.kaggle.com/niyamatalmass/google-job-skills