## Methodology

The way the database will first be created is to scrape the indeed website. This happens in two phases.

The information shall be stored on a back end cloud database. At this time it’s AWS RDS database.

# Searching Indeed

## Generating a list of results

Is done by the function indeed\_search\_function.py. It takes in two parameters (keywords, zip\_codes) and returns a DataFrame with the following

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data Type** | **Comments** |
| guid | Object | The guid is a unique id assigned to a job by indeed |
| job\_title\_row | Object | This group of data resides in one tag under the title header. Consists usually of Job Title – Company – Location. |
| real\_job\_title | Object | This is the result of splitting before the first hyphen. |
| company | Object | The second split on the hyphen |
| in\_location\_zip | Object | The third split on the hyphen. |
| Listing\_source | Object | Not captured at this time |
| publish\_date | Object | An unparsed date time string (Tue, 06 Aug 2019 04:21:46 GMT) |
| short\_description | Object | The one line description from the search results. |
| lat | Float64 | Latitude from the listing |
| L ongitude | Float64 | Longitude from the listing |
| extracted\_url | Object | The URL with only the essential information to retrieve the job found. No external tracking |
| scraped | Bool | A flag created as false and to be used in the future to be see to positive when the job has been scraped. |