

ETL PROCESS EXPLANATION

Extraction

The `contact_collector` function performs the extraction process in the ETL (Extraction, Transformation, and Loading) pipeline by collecting contact records from the HubSpot CRM.

Here's a step-by-step explanation of the extraction process based on the function:

- 1. **Initialization:** The function initializes variables such as `data_limit`, `all_records`, and `paging_count`. It also sets the access token and headers required for making API requests.
- 2. **Data Query:** The function constructs a JSON payload `data` that specifies the desired properties to retrieve from the CRM. It includes properties like "raw_email", "country", "phone", "technical_test__create_date", "industry", "address", and "hs_object_id". The payload also includes a filter to only collect records where the "allowed_to_collect" property is set to `True`.
- 3. **API Requests:** The function sends POST requests to the specified URL ('https://api.hubapi.com/crm/v3/objects/contacts/search') using the access token and headers. It passes the JSON payload as the request data. The function retrieves the response and checks if the status code is 200 (indicating a successful response).
- 4. **Pagination:** The function checks if the response includes a "paging" key. If not, it breaks the loop as all records have been collected. If the "paging" key is present, it updates the 'paging_count' and the JSON payload 'data' with the new "after" value to fetch the next page of results.
- 5. **Data Collection:** The function extracts the contact properties from the response JSON and appends them to the `all_records` list.
- 6. **Return:** Finally, the function creates a new list 'record_properties' that contains dictionaries representing the contact properties extracted from all the collected records. It returns this list as the result of the extraction process.

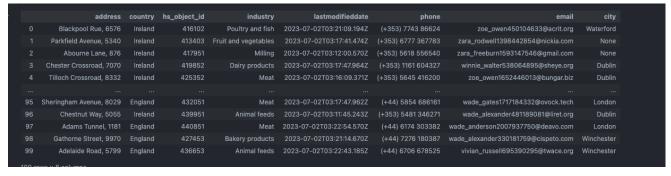
Subsequently, a Data Frame is generated from the data obtained.



	address	country	createdate	hs_object_id	industry	lastmodifieddate	phone	raw_email
0	Blackpool Rue, 6576	Waterford	2023-05-15T02:39:02.002Z	416102	Poultry and fish	2023-07-02T03:21:09.194Z	0-774-386-624	Zoe <zoe_owen450104633@acrit.org> Contact Info.</zoe_owen450104633@acrit.org>
1	Parkfield Avenue, 5340	Ireland	2023-05-15T02:39:02.003Z	413403	Fruit and vegetables	2023-07-02T03:17:41.474Z		Zara <zara_rodwell1398442854@nickia.com> Conta</zara_rodwell1398442854@nickia.com>
2	Abourne Lane, 876	Ireland	2023-05-15T02:39:02.003Z	417951	Milling	2023-07-02T03:12:00.570Z	5-618-556-540	Zara <zara_freeburn1593147546@gmail.com> Conta</zara_freeburn1593147546@gmail.com>
3	Chester Crossroad, 7070	Dublin	2023-05-15T02:39:02.003Z	419852	Dairy products	2023-07-02T03:17:47.964Z	1-161-604-327	Winnie <winnie_walter538064895@sheye.org> Cont</winnie_walter538064895@sheye.org>
4	Tilloch Crossroad, 8332	Dublin	2023-05-15T02:39:02.003Z		Meat	2023-07-02T03:16:09.371Z	5-645-416-200	Zoe <zoe_owen1652446013@bungar.biz> Contact Info.</zoe_owen1652446013@bungar.biz>
6931	Cato Hill, 5704	Waterford	2023-05-15T02:39:02.968Z	463801	Fruit and vegetables	2023-07-02T03:11:02.685Z	2-301-310-740	Rhea <rhea_grady231044837@deons.tech> Contact</rhea_grady231044837@deons.tech>
6932	Howard Crossroad, 9998	London	2023-05-15T02:39:02.969Z	454595	Bakery products	2023-07-02T03:22:55.259Z	6-747-016-018	Hailey <hailey_farrell934796609@brety.org> Con</hailey_farrell934796609@brety.org>
6933	Walnut Rue, 3859	Dublin	2023-05-15T02:39:02.969Z	454631	Animal feeds	2023-07-02T03:21:09.195Z	5-516-171-174	Caleb <caleb_purvis1251615808@iatim.tech> Cont</caleb_purvis1251615808@iatim.tech>
6934	Kimberley Pass, 3455	Plymouth	2023-05-15T02:39:02.969Z	457622	Bakery products	2023-07-02T03:22:49.581Z	0-605-727-343	Gina <gina_weasley83573127@naiker.biz> Contact</gina_weasley83573127@naiker.biz>
6935	Yorkshire Alley, 5843	Cork	2023-05-15T02:39:02.969Z	461395	Fruit and vegetables	2023-07-02T03:22:56.134Z	3-814-518-751	Aeris <aeris_walsh769266811@gompie.com> Contac</aeris_walsh769266811@gompie.com>
6936 rows x 8 columns								

Transformation

- 1 . **Detecting country and city:** The process applies the `country_recognition` function. This function recognizes the country based on the city. The result of this transformation is then split into two columns, 'country' and 'city'.
- 2. **Applying Functions to Email Column:** The function applies the `found_emails` function to the 'raw_email' column. This function extracts the email address from a raw email string.
- 3. **Formatting phones:** This method formats the phone numbers based on the country code.
- 4. **Duplicates Management:** The function performs duplicates management on the transformed DataFrame using the `duplicates_management` function. This function manages duplicate records based on email values.
- 5. **Return**: Finally, the function returns the transformed DataFrame, which contains the result of the transformation process.





Load

After the Transformation process it is necessary to upload the transformed records into the HubSpot platform using its API.

