Technical Assessment [External]

HubSpot MDM Engineer

Overview

This document outlines the process for the technical assessment of an MDM Engineer at HubSpot. This assessment includes an entity resolution exercise and an opportunity to document your approach. These will be discussed during a feedback session that will be scheduled after the assessment is returned.

Please return a zip file or repository link containing your solution, README, and output data by the deadline. As a rough approximation, we expect you to spend no more than 3 hours on the exercise. Submit whatever you come up with! A partial solution is much better than nothing at all.

Technical Feedback Session Agenda

This feedback session is a chance to discuss your thought process, approach, and skills after completing this assessment. The following is a general timeline of the technical presentation for your reference.

- **Introductions** (5 min)
- **Candidate-led code review** (30 min)
- **Interviewer-led feedback session** (15 min)
- **Candidate questions and wrap-up** (10 min)

Toolkit

This assessment is open-book so feel free to refer to any appropriate resources. The following are examples of the toolkit we are looking to see **(use of dbt, Python, or Java is required for Principal and Senior II levels)**:

- **Outstanding toolkit:** Build and return a repository that uses a modern data pipeline or application framework (python/java/dbt/Airflow/Dagster, etc.) to load data and create the expected output.
- **Great toolkit:** Return a repository that contains an application that loads source data into a database and generates the expected output.

- **Good toolkit:** Return a SQL query (in any dialect) or script that generates the expected output.

Entity Resolution Exercise

You are an MDM engineer at InsightGrid, a B2B SaaS tech company. To aid in sales prospecting and outbound sales motions, the company has acquired contact data from two data brokers, which will enrich the existing contact data in their CRM. Sales teams are eager to start using this contact data as part of their sales motion, and want a clean, accurate, and up-to-date contact list based on all the contact data available.

The data sources for this assessment contain contact data for a synthetic population, each containing common fields as well as source-specific details. You will be provided .csv files for each of the source tables. We've included table descriptions below for your reference.

A. Create a tool that resolves contact records across all data sources and outputs an up-to-date list of Contacts.

Tips:

- *This tool should work for any other contact CSVs provided by the three contact data sources.*
- *Consider that data cleanliness includes preventing duplicates, avoiding dropping records, and providing the most up-to-date data available.*
- *There are many falsehoods about names, phone numbers, and other common pieces of information that should be scrutinized during data cleaning.*

B. Generate a CSV file that represents a likely current state of all contacts. The file must have the following format

Column Name	Data Type	Description
contact_id	String	A unique ID for each contact
name	String	The contact's full name
email_address	String	The contact's work email address
phone_number	String	The contact's primary phone number
country	String	The country the contact lives in
favorite_color	String	The contact's favorite color
title	String	The contact's business title with their employer

company_name	Date	The name of the contact's employer
company_domain	Date	The domain name for the contact's employer
company_revenue	Integer	The estimated amount of revenue generated by the contact's employer
company_employees	Integer	The estimated number of employees working at the contact's employer
company_industry	String	The industry that this company is a part of
intent_signals	JSON	A list of all intent signals received for this contact.
do_not_call	Boolean	A boolean indicator that the Contact is part of the US Do Not Call Registry
created_at	Date	The date that the Contact was first created
updated_at	Date	The date that the Contact was most recently updated

Tips:

- *The CSV should use the provided field names*
- *Consider the data types that you will use for each field.*
- *This process requires trading off timeliness for quality data, so be sure to document assumptions made along the way.*

C. Write a README document to describes your approach for this exercise

This document is your opportunity to explain how your solution works, your thought process, and ways to improve this approach in the future. We want to hear about your understanding of the data set, tooling choices, entity resolution approach (design and stakeholder considerations, scalability, etc.), and to understand your overall thought process. *Please feel empowered to share any other materials (pseudocode, slides, diagrams, etc.) that will help us understand your solution.*

Table Descriptions

crm__contacts

Contacts sourced from the InsightGrid's CRM software. These are both generated automatically based on sign-ups for marketing content and events, and freemium products, as well as manually curated during conversations with sales reps.

Column Name	Data Type	Description
name	String	The full name of the contact.
email_address	String	The contact's work email address
phone_number	String	The contact's work phone number
favorite_color	String	The contact's favorite color.
title	String	The contact's business title with their employer
company_name	String	The name of the contact's employer
company_domain	String	The domain name for the contact's employer's website.
created_at	Date	The date that this contact record was created
updated_at	Date	The date that this contact record was most recently updated.

acme__contacts

Contacts sourced from Acme Quality Data Co, a data broker known for high quality contact data. These contacts are more complete than other sources, but this data set will contain fewer contacts than other sources due to the time and effort required to manually curate data.

Column Name	Data Type	Description
name	String	The full name of the contact.
email_address	String	The contact's work email address
phone_number	String	The contact's work phone number
country	String	The name of the country that contact lives in
title	String	The contact's business title with their employer

company_name	String	The name of the contact's employer
company_domain	String	The domain name for the contact's employer's website.
company_industry	String	The industry that the contact's employer operates in
company_employees	Integer	The estimated number of employees that work for the contact's employer.
company_revenue	Integer	The estimated revenue generated by the contact's employer.
created_at	Date	The date that this contact record was created
updated_at	Date	The date that this contact record was las updated.

rapid_data__contacts

Contacts sourced from RapidData, a data broker known for sourcing contacts using web tracking pixels, strategic partner's web traffic data, and myriad other automated means. These contacts tend to have higher error rates and duplicated records, but provide for a large sample of contacts.

Column Name	Data Type	Description
name	String	The full name of the contact.
email_address	String	The contact's work email address
phone_number	String	The contact's work phone number
ip_address	String	The public IP address at the time the contact viewed marketing content.
title	String	The contact's business title with their employer
company_name	String	The name of the contact's employer
company_domain	String	The domain name for the contact's employer's website.
intent_signals	JSON	A list of signals inferred by RapidData about the contact's intent to buy.
do_not_call	Boolean	An indicator that the contact does not want to be called.
created_at	Date	The date that this contact record was created
updated_at	Date	The date that this contact record was most recently updated.