Technical Assessment [External]

HubSpot Analytics Engineer

Overview

This document outlines the process for the technical assessment of an Analytics Engineer at HubSpot. This assessment includes a data modeling exercise and several stakeholder questions for you to use to prepare for the live technical presentation that will be scheduled after the assessment is returned.

Please attach or link any materials you'd like to use for the technical presentation along with your answers by the deadline. You can share additional materials during the presentation, even if they weren't in your initial submission. As a rough approximation, we expect you to spend no more than 3 hours on the problems and presentation preparation. Submit whatever you come up with! A partial solution is much better than nothing at all.

Technical Presentation Agenda

This presentation is your chance to demonstrate 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)
- **Candidate-led stakeholder presentation** (15 min)
- **Candidate questions and wrap-up** (10 min)

Toolkit

This assessment is open-book so feel free to refer to any appropriate resources. You can upload the source data to the environment of your choice. Please include code comments as necessary to illustrate your thought process. The following are examples of the toolkit we are looking to see **(use of dbt is required for Principal and Senior II levels)**:

- **Outstanding toolkit:** Build and return a dbt project to stage the source data and make any necessary transformations to solve the problems.
- **Great toolkit:** Return dbt-styled code (with refs, jinja expressions, etc.)
- **Good toolkit:** Return raw SQL (in any dialect)

Data Modeling Exercise

You are an analytics engineer at a rental property management company. Several new data sources have just become available from the source system and business analysts are eager to start using this data to make business decisions.

The data source for this assessment uses rental property listings, reviews, a calendar, and an amenity changelog. You will be provided .csv files for each of the source tables. We've included table descriptions below for your reference.

A. Create staging and intermediate (if necessary) data layers to organize and transform the data so it can be used for further analysis.

Tips:

- *Please follow dbt Lab's best practices for structuring your solution*
- *Apply any necessary data cleanup transformations on the source*
- *Choose appropriate column/table names (even if different than the source)*
- *Choose appropriate materializations*
- *Apply any helpful business transformations/joins to aid analysis*
- *Review problems 1-3 below to inform your design*

B. Develop a mart table that will be surfaced at the reporting layer to enable business analysts to perform period over period analysis across revenue and occupancy (how frequently a listing has a reservation) or to perform amenity and review score analysis per listing.

Tips:

- *The table should be at the day/listing grain*
- *Review problems 1-3 below to inform your design*
- *Consider how your stakeholder could use or misuse your data set*

C. Prepare to present your data modeling approach in a live code review.

The audience for this portion of the presentation is a panel of your peers. Think of it like a live code-review. We will ask questions along the way digging into the details of your approach/code or to provide feedback. We want to hear about your understanding of the data set, tooling choices, data modeling approach (design and stakeholder considerations, scalability, etc.), and to understand your overall thought process. You can share any materials (commented code, pseudocode, slides, sketches, etc.) that will help drive the conversation.

Stakeholder Presentation

Prepare for the live stakeholder presentation.

The audience for this portion of the presentation is a group of analysts in the business that you support. We want to hear how you would communicate with stakeholders (distilling technical solutions into business terms) to explain the considerations you made in your design for their ease-of-use. Please dive into sample queries for the problems below showing how they could use your mart table from the data modeling exercise to answer their questions.

#1 - Amenity Revenue

Write a query to find the total revenue and percentage of revenue by month segmented by whether or not air conditioning exists on the listing.

Tip: *For example, only 21.2% of revenue in July 2022 came from listings without air conditioning.*

#2 - Neighborhood Pricing

Write a query to find the average price increase for each neighborhood from July 12th 2021 to July 11th 2022.

Tip: *For example, the Back Bay neighborhood only has one listing, so the difference of $44 is the average for the whole neighborhood based solely on listing 10813.*

#3 - Long Stay / Picky Renter

A) **Write a query to find the maximum duration one could stay in each of these listings, based on the availability and what the owner allows.**

   Tip: *For example, listing 863788 is heavily booked. The largest timespan for which it is available is four days from September 18th to 21st in 2021. The correct solution should show that three listings are tied for the longest possible stay.*

B) **Write a variation of the maximum duration query above for listings that have both a lockbox and a first aid kit listed in the amenities.**

   Tip: *For example, listing 10986 has a lockbox. The correct result should show that*

across the results, the longest possible stay is much shorter than the answer to #3.

Table Descriptions

LISTINGS

Column Name	Data Type Description
ID	INTEGER Unique ID for this listing. Primary Key.
NAME	VARCHAR Display name of listing.
HOST_ID	INTEGER Unique ID for the Host who owns this property.
HOST_NAME	VARCHAR Display name of Host.
HOST_SINCE	DATETIME When the Host signed up.
HOST_LOCATION	VARCHAR Where the Host is based.
HOST_VERIFICATIONS NEIGHBORHOOD PROPERTY_TYPE ROOM_TYPE	VARCHAR (Parseable as JSON) Array of methods the Host can use to verify. VARCHAR The neighborhood where this listing is located. VARCHAR Description of the type of property. VARCHAR Description of the type of room.

ACCOMMODATES BATHROOMS_TEXT BEDROOMS BEDS	INTEGER Number of guests this room can accommodate. VARCHAR Number and types of bathrooms available. VARCHAR Number of bedrooms available for use. INTEGER Number of beds available for use.
AMENITIES	VARCHAR (Parseable as JSON) Array of amenities available for guests.
PRICE	VARCHAR The price of this listing as of the start of the date range in CALENDAR.
NUMBER_OF_REVIEWS	INTEGER The number of reviews this listing has ever received.
FIRST_REVIEW	VARCHAR The date of the first review this listing received.
LAST_REVIEW	VARCHAR The date of the most recent review this listing received.
REVIEW_SCORES_RATING	VARCHAR The average review score of this listing.

CALENDAR

Column Name	Data Type Description
LISTING_ID	INTEGER Unique ID for the listing to this which this row applies. Part of the Primary Key.
DATE	DATETIME Date of availability this row describes. Part of the Primary Key.
AVAILABLE	VARCHAR Contains ‘t’ if this property is available on this date. Contains ‘f’ if not.
RESERVATION_ID	INTEGER Unique ID for that DATE’s reservation. Foreign key. If NULL, there was no reservation on that date.
PRICE	VARCHAR The USD price to rent this property on DATE.
MINIMUM_NIGHTS	INTEGER The minimum number of nights that must be booked consecutively for this property.
MAXIMUM_NIGHTS	INTEGER The maximum number of nights that may be booked consecutively for this property.

GENERATED_REVIEWS

Column Name	Data Type Description
ID	INTEGER Auto-incrementing ID for the dummy reviews data
LISTING_ID	INTEGER Unique ID for the listing to this which this row applies
REVIEW_SCORE	INTEGER Generated score of the review, integer 1 to 5
REVIEW_DATE	DATE Generated date of the review

AMENITIES_CHANGELOG

Column Name	Data Type Description
LISTING_ID	INTEGER Unique ID for the listing to this which this row applies. Part of the Primary Key.
CHANGE_AT	DATETIME When the amenities list changed.
AMENITIES	VARCHAR (Parseable as JSON) Array of the amenities available as of the change.