Data Engineering Test Assignment

Background

At Matic, our BI stack can be simplified to the following:
- All products store data in Postgres DB
- We use ETL Service to replicate data from Postgres to Snowflake
- We use raw SQL to write aggregations and getDBT.com to run SQL queries
- We use Metabase and Tableau to present aggregated results in a form of reports and graphs

Responsibility of Data Engineer is to create and monitor connections to new data sources, build a full process of data aggregations, and work on the data design process. This includes working with business and engineering to help define and ensure accurate representation of business processes, metrics, segments, etc.

Assignment

For this test assignment, we ask you to apply your skill and experience to build two datasets that represent one of our core business entities - Policies. Dataset should be build using data aggregation framework - dbt. You can find more technical info in readme.md and framework documentation.

First one has basic data about each policy sale, including dimensions and facts. Second build based on active policies and should have records for each day when the policy was active.

Below you can find the desired structure for datasets:

POLICIES_AGGREGATED

policy_id	unique identificator for policy record
policy_type	type of policy which describes what type of property was insured : home, auto, fire,

	condo, renters, flood, earthquake, umbrella, moto
carrier	insurance carrier that holds the policy
premium	amount that should be paid for one policy term
sale_date	date the initial policy was sold
effective_date	start date of the policy term
expiration_date	expiration date of the policy term
cancellation_date	cancellation date of the policy
status	current status of the policy: active, bound, rewritten, cancelled, never_bound, pending, pending_cancellation, renewed, up_for_renewal
lead_id	lead that linked to the the policy record
x_is_sold	indicates initial policy sale record
x_is_renewed	indicates that policy was renewed from the initial policy sale record
property_state	state code in which this policy is effective
active_policies_count	number of active policies for a current sale date
renewed_policy_id	policy_id of the previous policy in the case of a renewal

POLICIES_PER_DAY_AGGREGATED

date	should include each calendar day starting from 2014-01-01
policy_id	identificator for policy record that was active on date
sale_date	date the initial policy was sold

property_state	state code in which this policy is effective
status	current status of the policy

If you will have any other recommendations regarding data structures, aggregations or business terms - feel free to describe them as well or bring up during feedback interview.
As part of the feedback interview, we expect to discuss all your findings and learn more about your process.
We ask you to spend not more than 4-8 hours on this test assignment and happy to assist with any questions you might have as part of working on it.

Input

In the folder data under dbt_project, you can source data that could be used in final datasets.
The data is anonymized extraction from a real database.

Aggregation

For this assignment, we selected Policies (short of Insurance Policy) aggregation (transformation) as main entity of our business. We aggregate only valid policies.

Definitions:

- **Valid policy** - a policy that has any status other than `never_bound` and not created as part of lead with disposition type `test_request`

- **Active policy** - a policy that is active on a certain date. The date is in the range between the effective date and less than cancellation or expiration date

- **X is sold** - is this a new policy that was sold by an agent. Sale date should not be empty and transaction type is one of `new_business`, `cross_sale` or `undefined`

- **X is renewed** - is this a policy that was renewed from the previously sold policy. The policy should have renewed_policy_id value

- **Property state** - State code in which this policy is effective. Selected from policy asset, lead asset or customer address

- **Active policies count** - Number of active policies for a current sale date. Used to display a number of active policies over time

Data Structure

Raw data has consist of the following tables

- **Assets** - list of customer assets
    - Belongs to customer
    - Has one home details if an asset is a home
    - Has many policies

- **Carriers** - list of insurance carriers
- Has many policies

- **Customers** - list of customers
    - Has many assets
    - Has many leads
    - Has many policies

- **Home Details** - details about a particular home
    - Belongs to Asset

- **Leads** - Attempts to sell a new policy to a new customer
    - Belongs to Customer
    - Has many Opportunities
    - Has many Policies

- **Opportunity** - tracking of an attempt to sell specific policy type
    - Belongs to Lead
    - Has many Opportunity Assets

- **Opportunity Asset** - relation between opportunity and asset
    - Belongs to Opportunity
    - Belongs to Asset

- **Policies** - list of policies
    - Belongs to Customer

- Belongs to Carrier
- Belongs to Lead
- Belongs to Asset