Open-Source Data Reliability

Using dbt & re_data

What is dbt?

- Framework for developing data application
- React for data people

What dbt lets you do?

- Write just SELECT statements
- Make SQL much more powerful
- Schedule execution of different models
- Macros - don’t repeat yourself
- Write tests for data
- Easily add packages with certain functionality

Models

```
{% set metrics_tables =  ['table_name','table_name2'] %}
{%- for table_name in metrics_tables %}
    select
        table_name,
        column_name,
        metric,
        value as last_value,
        interval_length_sec,
        computed_on
    from
        {{ ref(table_name) }}
    where
        time_window_end = {{- time_window_end() -}}

    {%- if not loop.last %} union all {%- endif %}

{% endfor %}
```

Macros

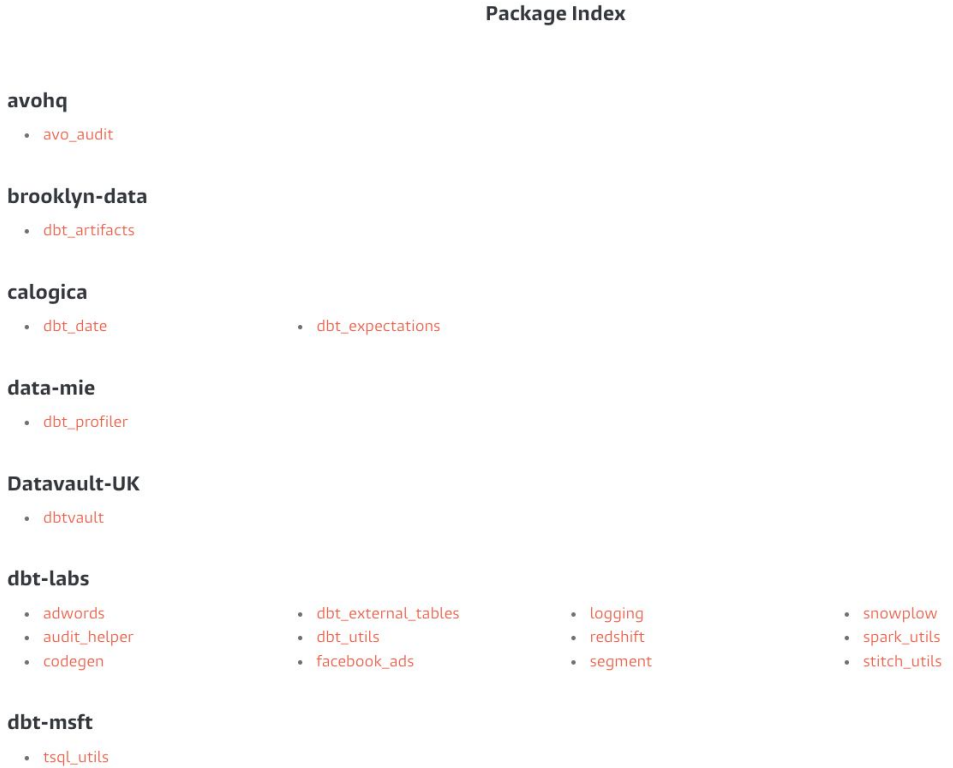
```
{% macro cents_to_dollars(column_name, precision=2) %}
    (({{ column_name }} / 100)::numeric(16, {{ precision }})
{% endmacro %}

{% macro filter_remove_duplicates(relation, unique_cols, sort_columns) %}
    (
        with with_row_num as (
            {{re_data.add_duplication_context(relation, unique_cols, sort_columns)}}
        ),
        one_row_num as (
            select * from with_row_num where re_data_duplicate_group_row_number = 1
        )
        select {{ dbt_utils.star(from=relation) }}
        from one_row_num
    )

{% endmacro %}
```

Data tests

```
models:
  - name: orders
    columns:
      - name: order_id
        tests:
          - unique
          - not_null
      - name: status
        tests:
          - accepted_values:
              values: ['placed', 'shipped', 'completed', 'returned']
      - name: customer_id
        tests:
          - relationships:
              to: ref('customers')
              field: id
```

Packages hub

```
packages:
  - package: dbt-labs/dbt_utils
    version: [">=0.7.0", "<0.9.0"]
  - package: re-data/re_data
    version: [">=0.6.0", "<0.7.0"]
```



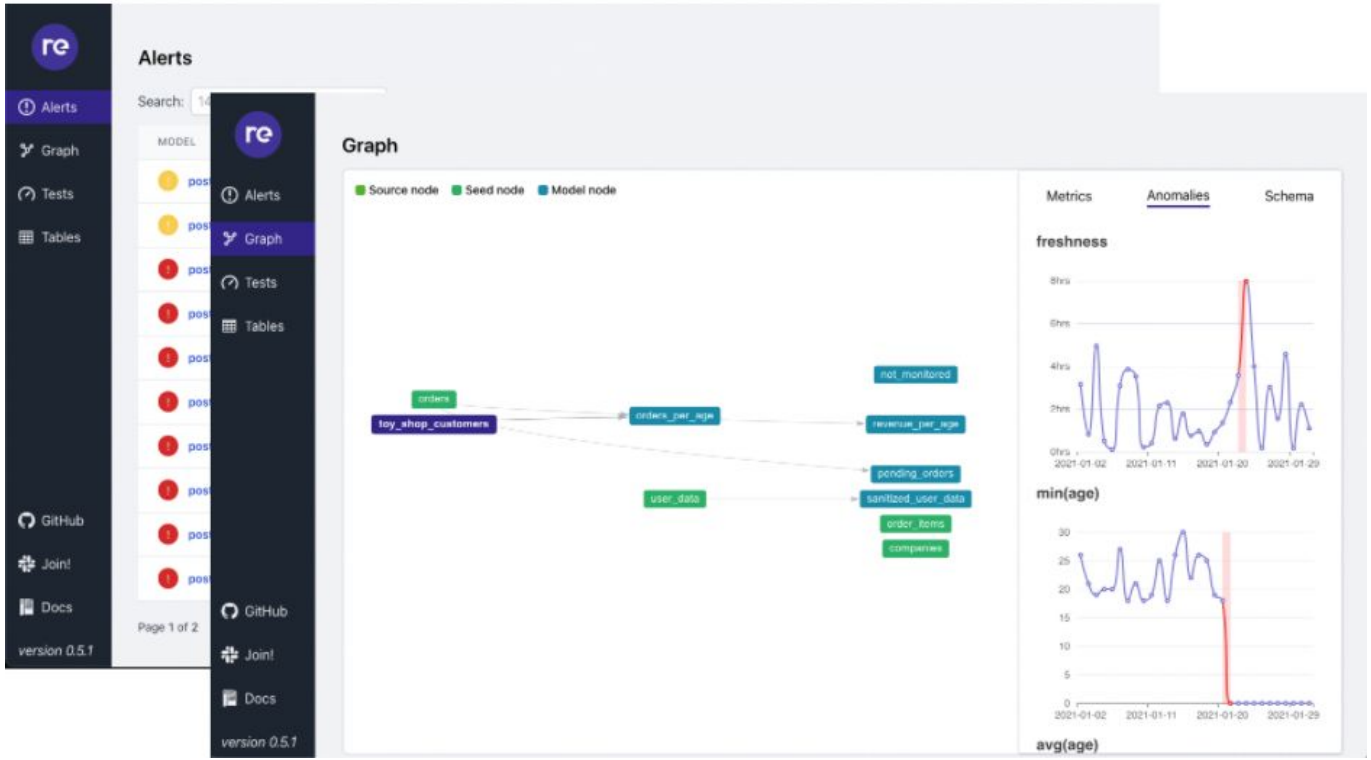
What re_data is?

- Tool helping you find & debug errors in data
- DataDog/Grafana for data people

What does re_data do?

- Compute data quality metrics
- Anomaly detection
- Detect schema changes
- dbt tests history
- Data cleaning macros
- Integrate Slack alerting
- UI to navigate your data 😊

re_data demo! 😊


Alerts
Search:
MODEL
Alerts
Graph
Tests
Tables
GitHub
Join!
Docs
version 0.5.1
Page 1 of 2
Graph
Source node
Seed node
Model node
orders
toy_shop_customers
orders_per_age
not_monitored
revenue_per_age
pending_orders
user_data
sanitized_user_data
order_items
companies
Metrics
Anomalies
Schema
freshness
8hrs
6hrs
4hrs
2hrs
0hrs
2021-01-02
2021-01-11
2021-01-20
2021-01-29
min(age)
30
25
20
15
10
5
0
avg(age)

Z-score & Modified Z-score

$$Z\text{-score} = \frac{x - mean}{Standard\ Deviation}$$

$$Modified\ Z\text{-score} = \frac{0.6745*(x - median)}{MAD}$$

MAD (Median Absolute Deviation) for:
(1, 1, 2, **2**, 4, 6, 9).
Deviations from media: (1, 1, 0, 0, 2, 4, 7)
Median of that deviations: (0, 0, 1, **1**, 2, 4, 7)

So the median absolute deviation for this data is 1.

Questions?