Mastering Multitenant Orchestration with dbt and Dagster

Life of Data Engineers does not have to be that hard

Andrea Montes

DBT Bogotá Meetup

2024-06

Audience questions

Problem

We need to re-design

- Requirements
- New architecture
- Stage: Extract
- Stage: Load
- Stage: Transformation
- Stage: Transformation - Leveraging DBT for multitenancy
- Stage: Transformation - but for +250 clients?

Takeaways

Audience questions

1. Airflow familiarity?
2. Crazy tools difficult to debug?

What is needed?

- Daily updates to client dashboards. +250 different clients
- Custom reports per client
- Product and business questions

Legacy product(s) - Data Warehouse

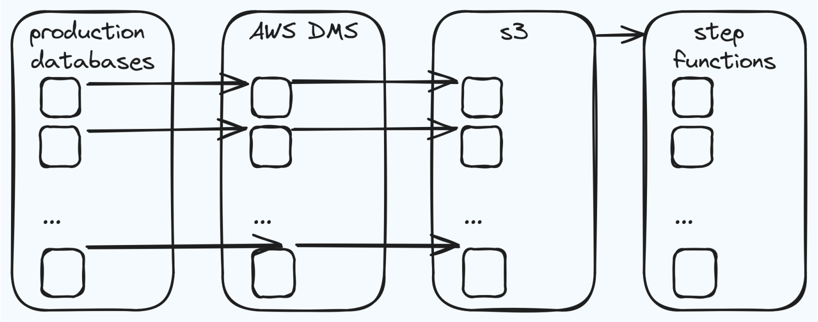


Figure: Legacy DWH product

- Daily refresh
- Failed 2 or 4 times a week for heavy clients

Legacy product(s) - Reporting

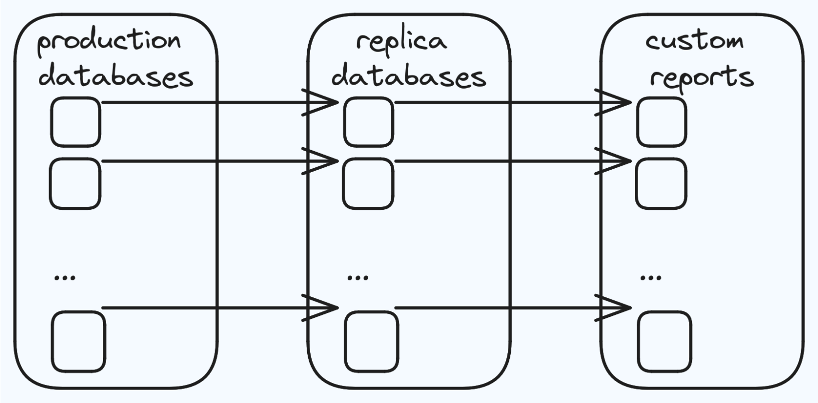


Figure: Legacy reporting product

- Replicated once a week
- Raw sql
- Stored procedures

We need to re-design

Different users, different requirements

Dashboards

Clients need their dashboards updated to know events statuses

Business questions

What are the most used features?, how virtual vs in-person events attendance has changed after pandemic?

Client questions

How long is taking a candidate to become an applicant? Do I have a diverse pipeline?

New architecture

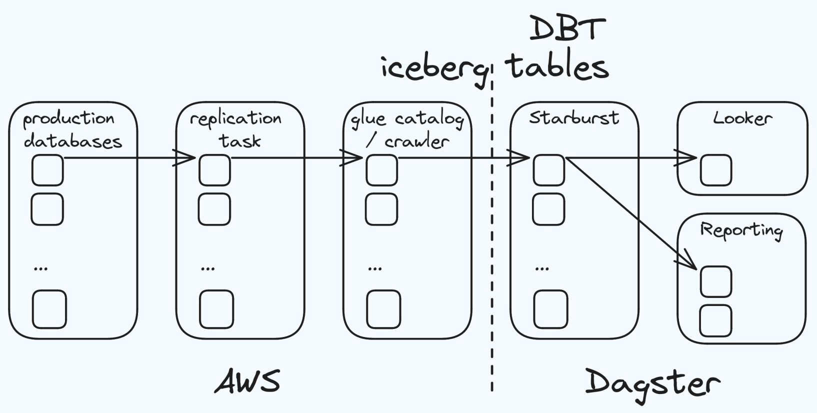


Figure: New architecture to accomplish reporting and DWH requirements

DMS

- Cheap...
- Not reliable as we would like
- Trade-off

Glue catalog and Crawler

- Easy enough to implement
- Our compute engine support it

Starburst - DBT

Pros:

- Starburst is a vendor option for Trino
- Cheaper than snowflake
- Trino has a good community

Cons:

- SQL ANSI
- It's a new product, random changes
- Compute throttling
- We were their first large user

Minimum Example: DBT and multiple clients

Data used:

- Open data Colombia - Mobile Telephony subscribers by category
- One CSV with information for all mobile telephony providers
- We need to split this data in a convenient way

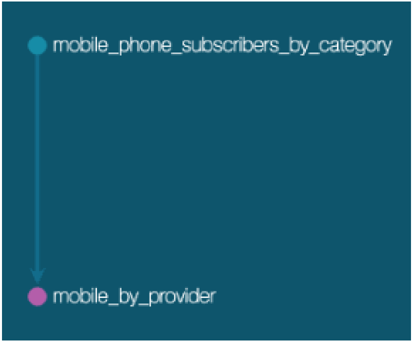


Figure: Example: Example Lineage

DBT and multiple clients

Model: Mobile by provider

```
SELECT
    ANO AS year
    , TRIMESTER AS trimester
    , PROVEEDOR AS provider
    , "LINEAS EN SERVICIO" AS lines_in_service
    , "LINEAS PREPAGO" AS prepaid_lines
    , "LINEAS POSPAGO" AS postpaid_lines
    , "LINEAS ACTIVADAS" AS enabled_lines
    , "LINEAS RETIRADAS" AS retired_lines
FROM {{ ref(’mobile_phone_subscribers_by_category’)}}
WHERE PROVEEDOR = ’{{ var("client")}}’
```

DBT and multiple clients

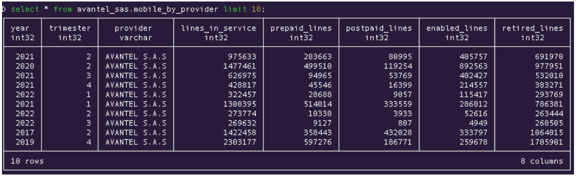
DBT command:

```
dbt run —profiles-dir dbt_project/.dbt —select
    mobile_by_provider —vars '{client: AVANTEL S
    .A.S}'
```

SQL compiled:

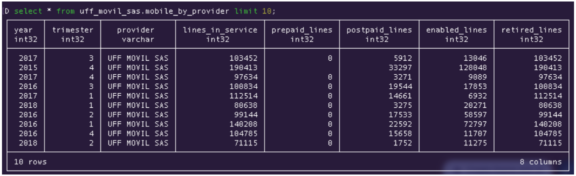
```
CREATE TABLE
      "housing"."avantelsas"."mobile_by_provider"
    AS (
SELECT
    ANO AS year
    , TRIMESTER AS trimester
    , PROVEEDOR AS provider
    , "LINEAS EN SERVICIO" AS lines_in_service
    , "LINEAS PREPAGO" AS prepaid_lines
    , "LINEAS POSPAGO" AS postpaid_lines
    , "LINEAS ACTIVADAS" AS enabled_lines
    , "LINEAS RETIRADAS" AS retired_lines
FROM housing.telephony_sources.
   mobile_phone_subscribers_by_category
WHERE PROVEEDOR = 'AVANTEL S.A.S')
```

DBT and multiple clients



```
D select * from avantel_sas.mobile_by_provider limit 10;
┌───────┬───────────┬──────────────┬──────────────────┬───────────────┬────────────────┬───────────────┬───────────────┐
│ year  │ trimester │   provider   │ lines_in_service │ prepaid_lines │ postpaid_lines │ enabled_lines │ retired_lines │
│ int32 │   int32   │   varchar    │      int32       │     int32     │     int32      │     int32     │     int32     │
├───────┼───────────┼──────────────┼──────────────────┼───────────────┼────────────────┼───────────────┼───────────────┤
│  2021 │         2 │ AVANTEL S.A.S│           975633 │        283663 │          80995 │        405757 │        691970 │
│  2020 │         2 │ AVANTEL S.A.S│          1477461 │        499510 │         119254 │        892563 │        977951 │
│  2021 │         3 │ AVANTEL S.A.S│           626975 │         94965 │          53769 │        402427 │        532010 │
│  2021 │         4 │ AVANTEL S.A.S│           428817 │         45546 │          16399 │        214557 │        383271 │
│  2022 │         1 │ AVANTEL S.A.S│           322457 │         28688 │           9057 │        115417 │        293769 │
│  2021 │         1 │ AVANTEL S.A.S│          1300395 │        514014 │         333559 │        286012 │        786381 │
│  2022 │         2 │ AVANTEL S.A.S│           273774 │         10330 │           3933 │         52616 │        263444 │
│  2022 │         3 │ AVANTEL S.A.S│           269632 │          9127 │            807 │          4949 │        260505 │
│  2017 │         2 │ AVANTEL S.A.S│          1422458 │        358443 │         432028 │        333797 │       1064015 │
│  2019 │         4 │ AVANTEL S.A.S│          2303177 │        597276 │         186771 │        259678 │       1705901 │
├───────┴───────────┴──────────────┴──────────────────┴───────────────┴────────────────┴───────────────┴───────────────┤
│ 10 rows                                                                                                   8 columns │
└─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────┘
```

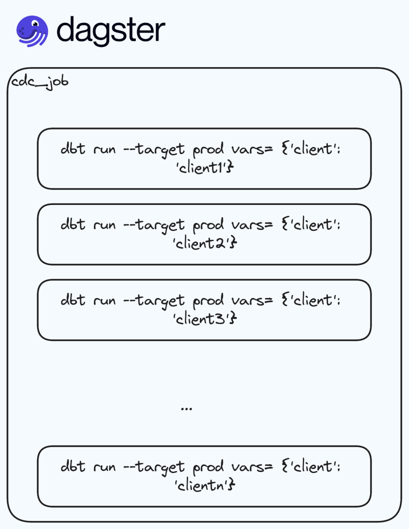
Figure: Example: client = AVANTEL S.A.S



```
D select * from uff_movil_sas.mobile_by_provider limit 10;
┌───────┬───────────┬──────────────┬──────────────────┬───────────────┬────────────────┬───────────────┬───────────────┐
│ year  │ trimester │   provider   │ lines_in_service │ prepaid_lines │ postpaid_lines │ enabled_lines │ retired_lines │
│ int32 │   int32   │   varchar    │      int32       │     int32     │     int32      │     int32     │     int32     │
├───────┼───────────┼──────────────┼──────────────────┼───────────────┼────────────────┼───────────────┼───────────────┤
│  2017 │         3 │ UFF MOVIL SAS│           103452 │             0 │           5912 │         13046 │        103452 │
│  2015 │         4 │ UFF MOVIL SAS│           190413 │               │          33297 │        128048 │        190413 │
│  2017 │         4 │ UFF MOVIL SAS│            97634 │             0 │           3271 │          9089 │         97634 │
│  2016 │         3 │ UFF MOVIL SAS│           100834 │             0 │          19544 │         17853 │        100834 │
│  2017 │         1 │ UFF MOVIL SAS│           112514 │             0 │          14661 │          6932 │        112514 │
│  2018 │         1 │ UFF MOVIL SAS│            80638 │             0 │           3275 │         20271 │         80638 │
│  2016 │         2 │ UFF MOVIL SAS│            99144 │             0 │          17533 │         58597 │         99144 │
│  2016 │         1 │ UFF MOVIL SAS│           140208 │             0 │          22592 │         72797 │        140208 │
│  2016 │         4 │ UFF MOVIL SAS│           104785 │             0 │          15658 │         11707 │        104785 │
│  2018 │         2 │ UFF MOVIL SAS│            71115 │             0 │           1752 │         11275 │         71115 │
├───────┴───────────┴──────────────┴──────────────────┴───────────────┴────────────────┴───────────────┴───────────────┤
│ 10 rows                                                                                                   8 columns │
└─────────────────────────────────────────────────────────────────────────────────────────────────────────────────────┘
```

Figure: Example: client = UFF MOVIL

Orchestration tool: Dagster

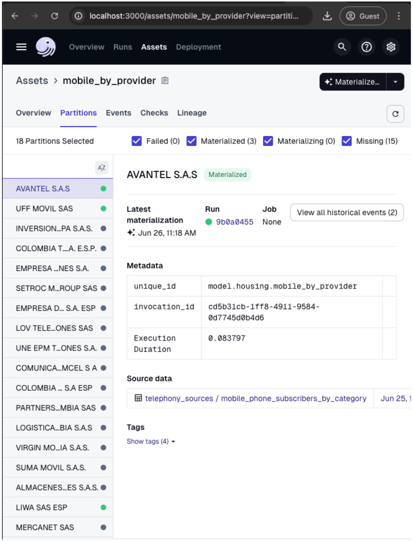


```
dbt run --target prod vars= {'client': 'client1'}
dbt run --target prod vars= {'client': 'client2'}
dbt run --target prod vars= {'client': 'client3'}
dbt run --target prod vars= {'client': 'clientn'}
```

Figure: Dagster: How do we run +250 transformations?

- We run the job daily starting at 2 am CST
- Each job client partition takes about 7 mins
- Job finishes before 7 am CST

Orchestration tool: Dagster



localhost:3000/assets/mobile_by_provider?view=partiti...

Assets > mobile_by_provider

18 Partitions Selected

☑ Failed (0) ☑ Materialized (3) ☑ Materializing (0) ☑ Missing (15)

- AVANTEL S.A.S ●
- UFF MOVIL SAS ●
- INVERSION...PA S.A.S. ●
- COLOMBIA T...A. E.S.P. ●
- EMPRESA ...NES S.A. ●
- SETROC M...ROUP SAS ●
- EMPRESA D... S.A. ESP ●
- LOV TELE...ONES SAS ●
- UNE EPM T...ONES S.A. ●
- COMUNICA...MCEL S A ●
- COLOMBIA ... S.A ESP ●
- PARTNERS...MBIA SAS ●
- LOGISTICA...BIA S.A.S ●
- VIRGIN MO...IA S.A.S. ●
- SUMA MOVIL S.A.S. ●
- ALMACENES...ES S.A.S. ●
- LIWA SAS ESP ●
- MERCANET SAS ●

AVANTEL S.A.S Materialized

Latest materialization	Run	Job
Jun 26, 11:18 AM	● 9b0a0455	None

View all historical events (2)

Metadata

unique_id	model.housing.mobile_by_provider
invocation_id	cd5b31cb-1ff8-4911-9584-0d7745d0b4d6
Execution Duration	0.083797

Source data

telephony_sources / mobile_phone_subscribers_by_category Jun 25, 1

Tags

Show tags (4)

Figure: Dagster: How do we run +250 transformations?

Takeaways

- Not getting used to what is poorly done
- Don't fall in love with a technology product
- Invest time learning new tools!
- Fail fast mindset

Who am I?

Figure: Andrea Montes

Andrea Montes - Senior Data Engineer

mamontesp

Code example: https://github.com/mamontesp/mastering-orchestration-dbt-dagster

My blog: https://mamontesp.ghost.io/

QRs!

Figure: Andrea Montes in in

Figure: Code Example