

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

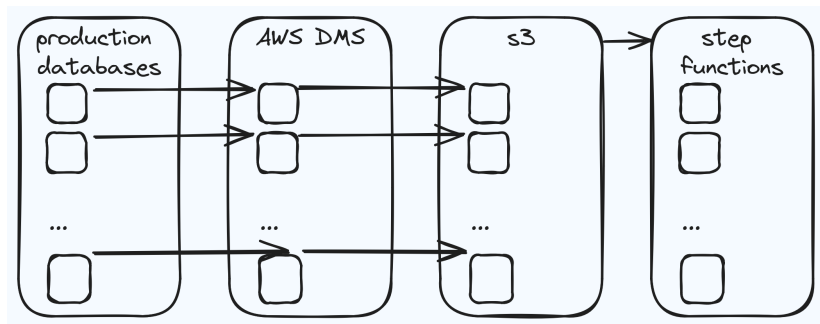


Figure: Legacy DWH product

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

Legacy product(s) - Reporting

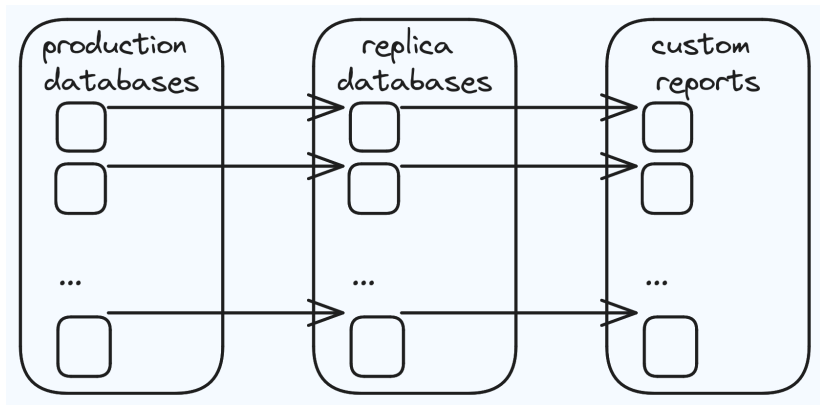


Figure: Legacy reporting product

- ▶ 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

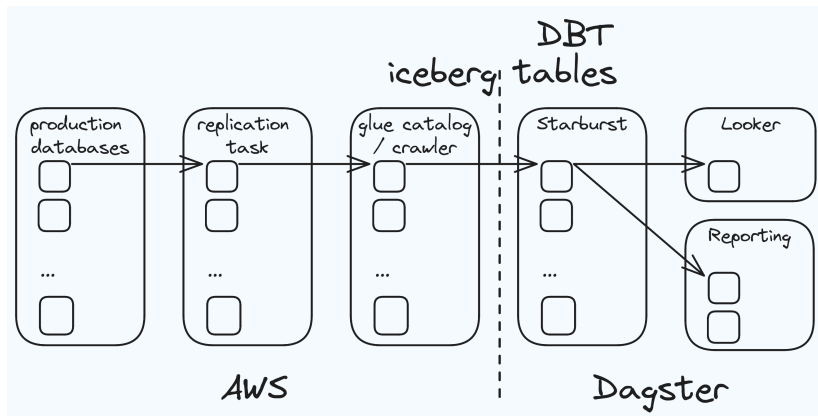


Figure: New architecture to accomplish reporting and DWH requirements

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

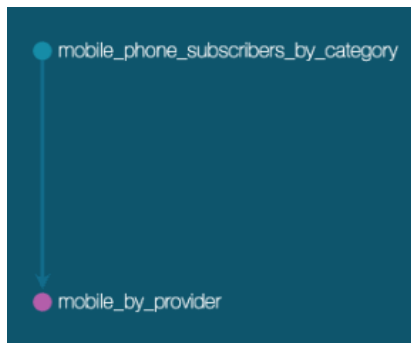


Figure: Example: Example Lineage

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:

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 avantelsas.mobile_phone_subscribers_by_category  
WHERE PROVEEDOR = 'AVANTEL S.A.S'
```

DBT and multiple clients

```
select * from avantel_gas.mobile_by_provider limit 10;
```

year int32	trimester int32	provider varchar	lines_in_service int32	prepaid_lines int32	postpaid_lines int32	enabled_lines int32	retired_lines int32
2021	2	AVANTEL S.A.S	975633	203663	08995	485757	691970
2020	2	AVANTEL S.A.S	1477461	499518	119254	892563	977951
2021	3	AVANTEL S.A.S	626875	94965	53769	482427	532810
2021	4	AVANTEL S.A.S	438817	45146	16399	214557	383271
2022	1	AVANTEL S.A.S	322457	28688	8057	115417	293769
2021	1	AVANTEL S.A.S	1308395	514014	393559	286012	786381
2022	2	AVANTEL S.A.S	273774	10338	3933	52616	263444
2022	3	AVANTEL S.A.S	269632	9127	807	4949	268595
2017	2	AVANTEL S.A.S	1422458	358443	432928	333797	1064815
2019	4	AVANTEL S.A.S	2303177	597276	106771	259678	1705981
10 rows							8 columns

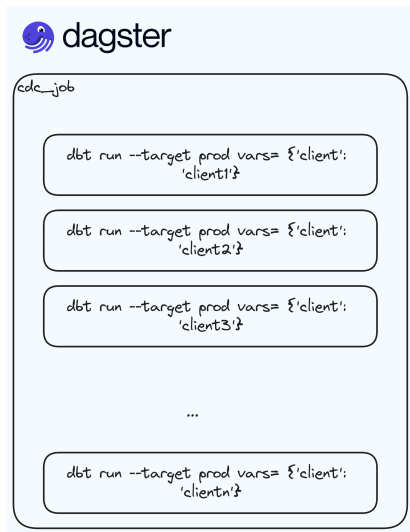
Figure: Example: client = AVANTEL S.A.S

```
select * from uff_movil_gas.mobile_by_provider limit 10;
```

year int32	trimester int32	provider varchar	lines_in_service int32	prepaid_lines int32	postpaid_lines int32	enabled_lines int32	retired_lines int32
2017	3	UFF MOVIL SAS	103452	0	5912	13846	103452
2015	4	UFF MOVIL SAS	190413	0	33297	128048	190413
2017	4	UFF MOVIL SAS	97634	0	3271	9889	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	28271	80638
2016	2	UFF MOVIL SAS	99144	0	17535	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



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

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

Orchestration tool: Dagster

The screenshot shows the Dagster web interface at localhost:3000. The 'Assets' page for 'mobile_by_provider' is active. A list of 18 partitions is on the left, with 'AVANTEL S.A.S.' highlighted. The right panel shows details for 'AVANTEL S.A.S.', which is 'Materialized'. It includes a 'Latest materialization' run with ID '9b8a0455' from June 26, 11:18 AM. A 'Metadata' table is also present.

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

Below the metadata table, the 'Source data' section shows the path 'telephony_sources / mobile_phone_subscribers_by_category' and the date 'Jun 25, 2024'. There are also 'Tags' listed at the bottom.


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?



 Andrea Montes - Senior Data Engineer




 mamontesp

Figure: Andrea Montes

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

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

QRs!




Figure: Andrea Montes in 



Figure: Code Example