Welcome

Berlin dbt November Meetup

Hosts: Kumar Aman & Sowmia Naraynan Siva Subramanian / *Tier Mobility SE*

Join dbt community Slack
getdbt.com/community
#local-berlin

dbt

Kumar Aman

Team Lead Analytics Platform, Tier Mobility SE

(he/him)

I let my mind wander and it never came back so I like going wandering myself to try finding where it went.

kumar.aman@tier.app

Linkedin.com/in/kumar-aman-1011

Sowmia Naraynan Sivasubramanian

Head of Analytics Platform, Tier Mobility SE

(he/him)

I believe in 80/20 rule. Yes, it took 80% of my time to fill up the fun fact here and you know the rest..

sowmia.subramanian@tier.app
linkedin.com/in/sowmianaraynan

Agenda

- 6 - 6:50 pm - Networking, Pizza and Drinks
- 6:50 - 7 pm - Welcome Remarks by Kumar & Sowmi
- 7 - 7:30 pm - From Metadata to Metrics: Utilizing dbt API to create actionable metrics and SLOs for your data org with Marian & Jerry
- 7:30 - 8 pm - Minimum Viable Stacks - Archetyping the Modern Data Stack for Startups with Saman
- 8 - 9 pm - Follow-up Q&A, drinks and networking

Icebreaker

Two truths and a lie

- Tier was founded in 2019 by Lawrence Leuschner, Matthias Laug & Julian Blessin
- Tier is operating in more than 33 countries and 550 cities
- Tier has been climate neutral since 2020

TIER IS THE LEADING MICRO-MOBILITY PROVIDER GLOBALLY

founded in **2018** by Lawrence Leuschner, Matthias Laug & Julian Blessin

active in >**550 cities in 33+ countries,** HQ in Berlin

300,000+ e-scooters, e-mopeds, bikes, cargo bikes & e-bikes

More than **48 million car trips saved** since 2018

100% climate neutral through sustainable ops & long vehicle lifetime

Our focus:

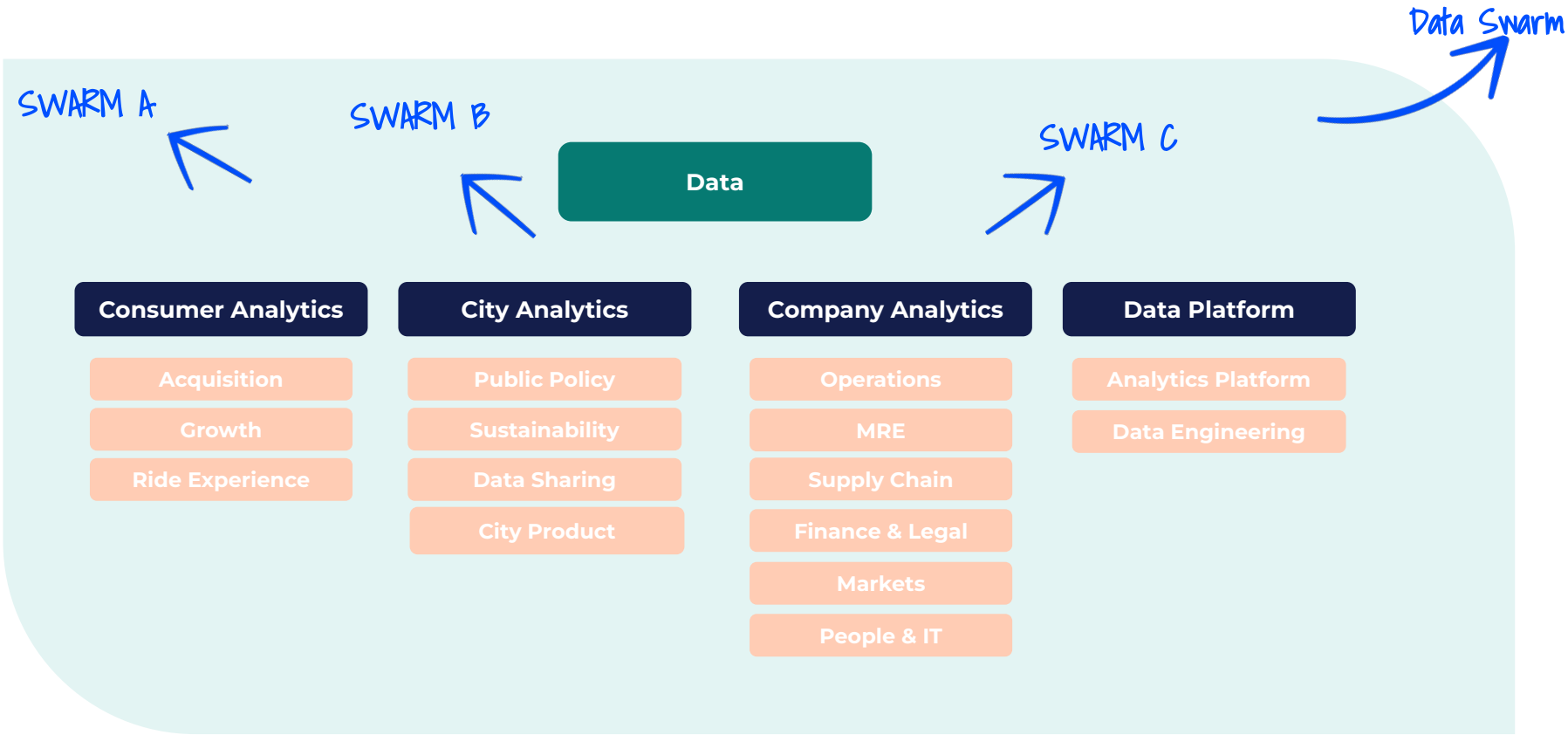
Providing the safest, most equitable & most sustainable mobility solution.

TIER

OUR VEHICLES

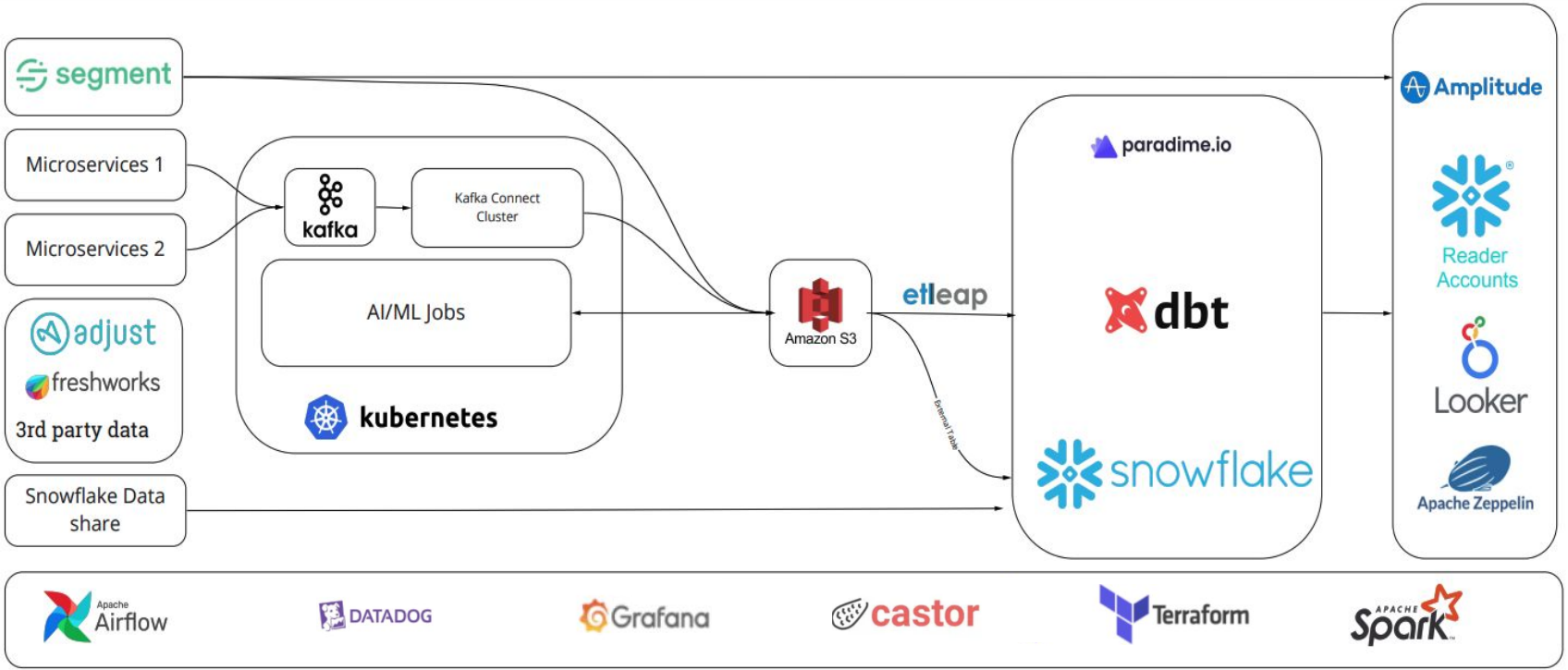
TIER

DATA TEAM SETUP



TIER

DATA WORKFLOW



TIER

WE ARE HIRING, COME JOIN THE RIDE!

- ➔ **We are hiring and offer remote work from the many countries such as Germany, Spain, Netherlands, United Kingdom, Poland, Hungary, Greece, Austria.**
- ➔ **Being part of the team means you're part of its success.**
  - ◆ **Our ESOP program gives you an opportunity to have a share in the company**
  - ◆ **We support your growth with an annual development budget**
- ➔ **Diversity and inclusivity is super important to us, join initiatives & programs such as #WomenofTIER**
- ➔ **Profit from competitive compensation combined with perks like memberships with Urban Sports Club, Blinklist and Headspace.**
- ➔ **Surf your city with free rides on our e-scooters and e-bikes**

TIER

TIER

EN | Get the app

All open positions

WE ARE HIRING !!

Link to job board!

-- Data | All offices

POSITION	DEPARTMENT	LOCATION
Data Engineer (m/f/d)	Data	Athens, Berlin, Budapest, Kraków, Madrid, Remote - Germany
Head of Data Management (m/f/d)	Data	Athens, Berlin, Budapest, Kraków, Madrid
Head of Data & ML Engineering (m/f/d)	Data	Athens, Berlin, Budapest, Kraków, Madrid
Machine Learning Engineer (m/w/d)	Data	Athens, Berlin, Budapest, Kraków, Madrid
Product Data Analyst - City (m/f/d)	Data	Athens, Budapest, Kraków, Madrid
(Senior) Data Analyst - Growth (m/f/d) - REMOTE	Data	Athens, Barcelona, Berlin, Budapest, Madrid
(Senior) Product Data Analyst - Consumer (m/f/d)	Data	Athens, Barcelona, Berlin, Budapest, Madrid

TIER

Join dbt Community Slack

getdbt.com/community

#local-berlin

Speaking Today

Marian Moreno Corral

Tier Mobility SE

Jerry Nwabuilo

Tier Mobility SE

Saman Arefi

Project A Ventures

Maria de los Angeles Moreno Corral

Analytics Engineer,
Tier Mobility SE

(she/her)

"I spent more time thinking about a fun fact than preparing the presentation"

maria.moreno@tier.app
linkedin.com/in/maria-morenocorral/

Jerry Nwabuilo

Senior Analytics Engineer, Tier Mobility SE

(he/him)

“Working for a mobility company but can’t ride a scooter.”

jerry.nwabuilo@tier.app
Linkedin.com/in/jerry-nwabuilo

From Metadata to Metrics:

Utilizing dbt API to create actionable metrics and SLOs for your data org

dbt November Meetup 2022

TIER

Agenda

- Overview
- Business Objectives
- Data Collection Workflow
- dbt Performance Dashboard
- SLOs Overview
- Data Pipeline For SLO Reporting

TIER

Overview

Overview

Are there tangible ways to measure the reliability of your data pipeline and the efficiency of your analytics engineers?

At Tier Mobility, we use the dbt API to fetch metadata of our models and take it all the way to our visualisation tool Looker conveying the state of our infrastructure to Data Management.

Metadata analysis helps our data teams to optimize their models, track the impact of their initiatives, and reduce costs.

TIER

Business Objectives

Business Objectives

Data Ownership

- ★ foster data ownership by assigning models to domain and individual
- ★ decentralize cost based on usage per domain

Quality Benchmarks

- ★ % data assets documented
- ★ % models incremental
- ★ % models with tests

How does getting this information **generate value for the team?**

Performance Management

- ★ identify when WH needs to be scaled up
- ★ identify when perf. improvements are required

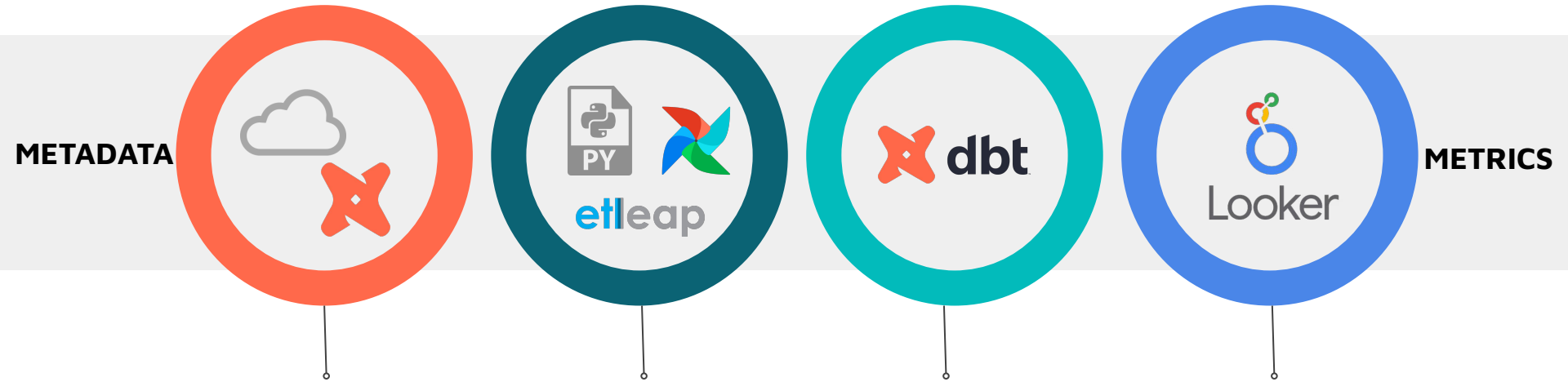
SLOs Tracking

- ★ monitor models' downtime > 3h to ensure the team SLO is being met

TIER

Data Collection Workflow

Data Collection Workflow



TIER

Data Collection Workflow

dbt Cloud API → Orchestration → Data Transformation → Looker Explores

The dbt Cloud API allows us to fetch **metadata from our dbt Cloud account**

Each dbt run generates **Artifacts** which can be downloaded as json files in the Artifacts section of dbt cloud jobs.

We use 2 Artifacts files:

dbt

manifest.json

Metadata about all the **resources available** in the dbt project (models, tests, macros, seeds,..) along with **their properties and configurations**

run_results.json

Information about the **resources that are executed** in a complete dbt invocation, and **the timing and status for each**

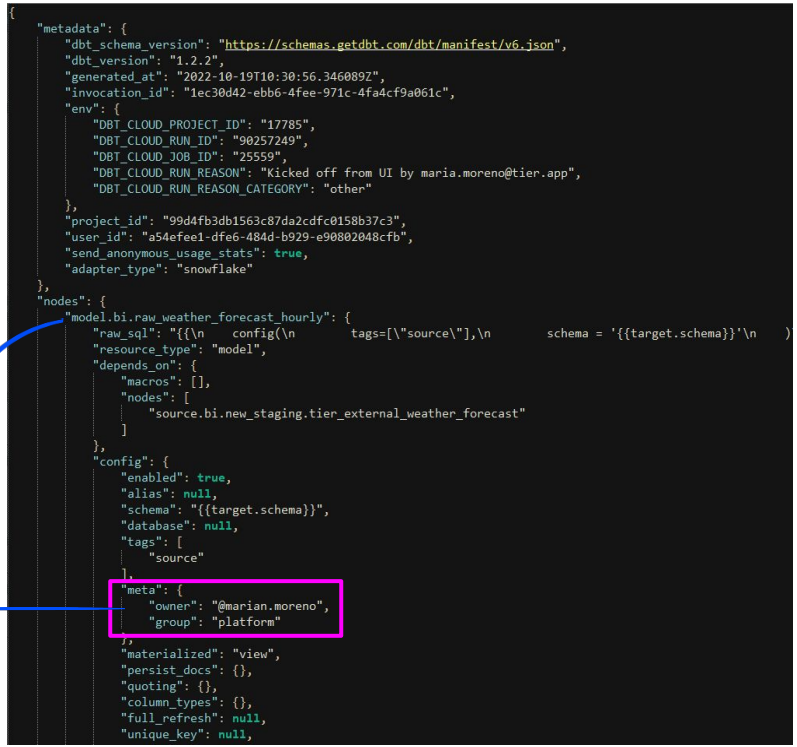
TIER

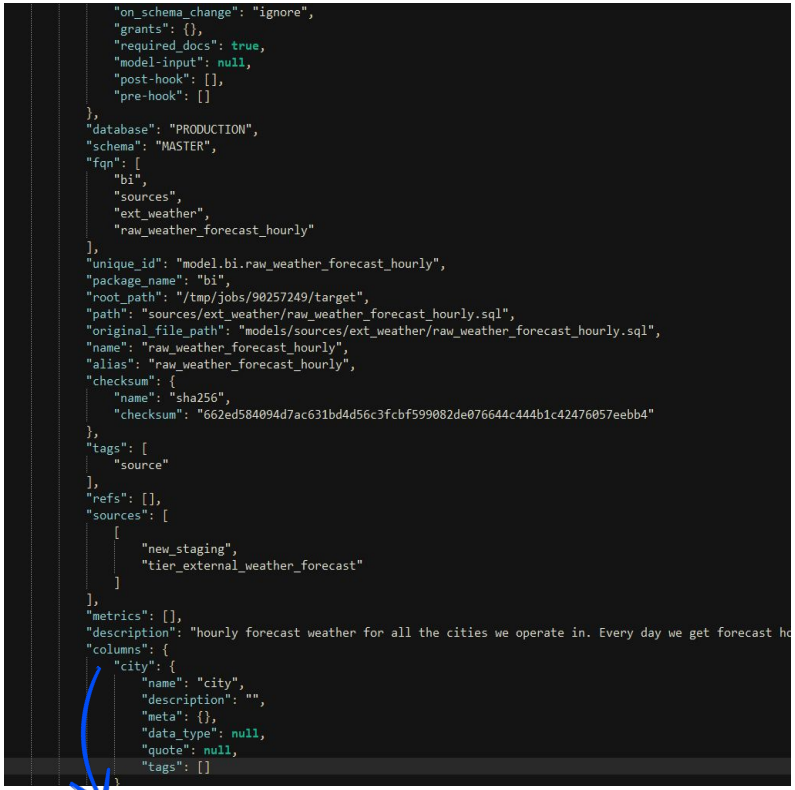
manifest.json

Models metadata

Customised metadata:: Model ownership

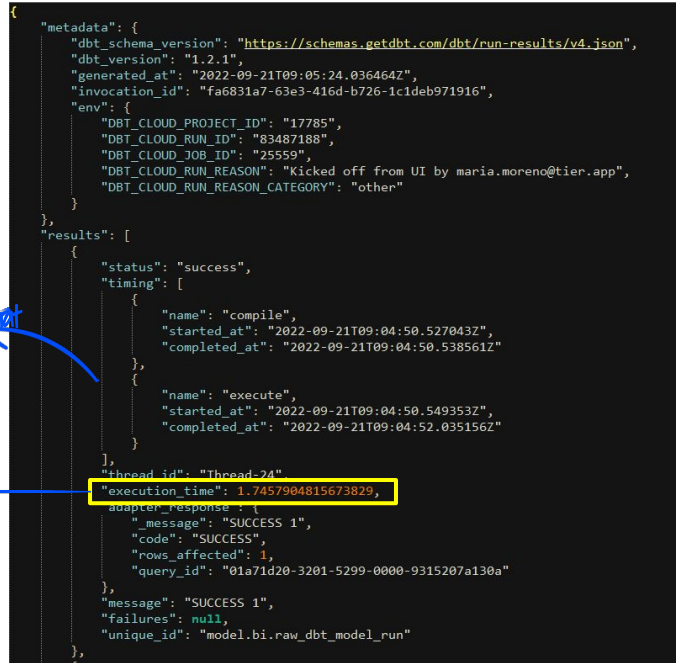
Columns metadata





TIER

run_results.json



Results for each model that ran

Timing of the first executed model

Second model that ran

Execution time in seconds of the whole run, all the models in total

Run command

TIER

Data Collection Workflow

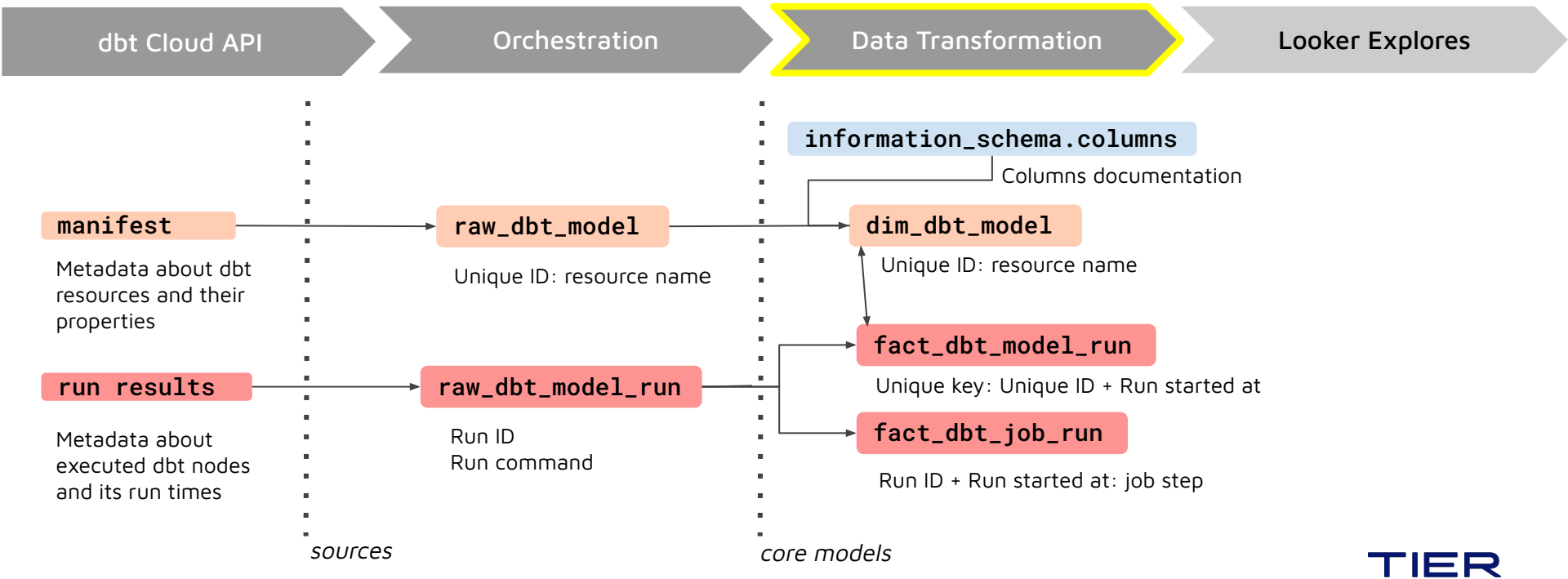
dbt Cloud API | **Orchestration** | Data Transformation | Looker Explores

- One **Python script** for pulling each Artifact file
- ***Run_results*** script runs **hourly**, and ***manifest*** **daily** in Airflow
- Both scripts push data to **ETLeap pipelines**

Scripts DAGs Pipelines

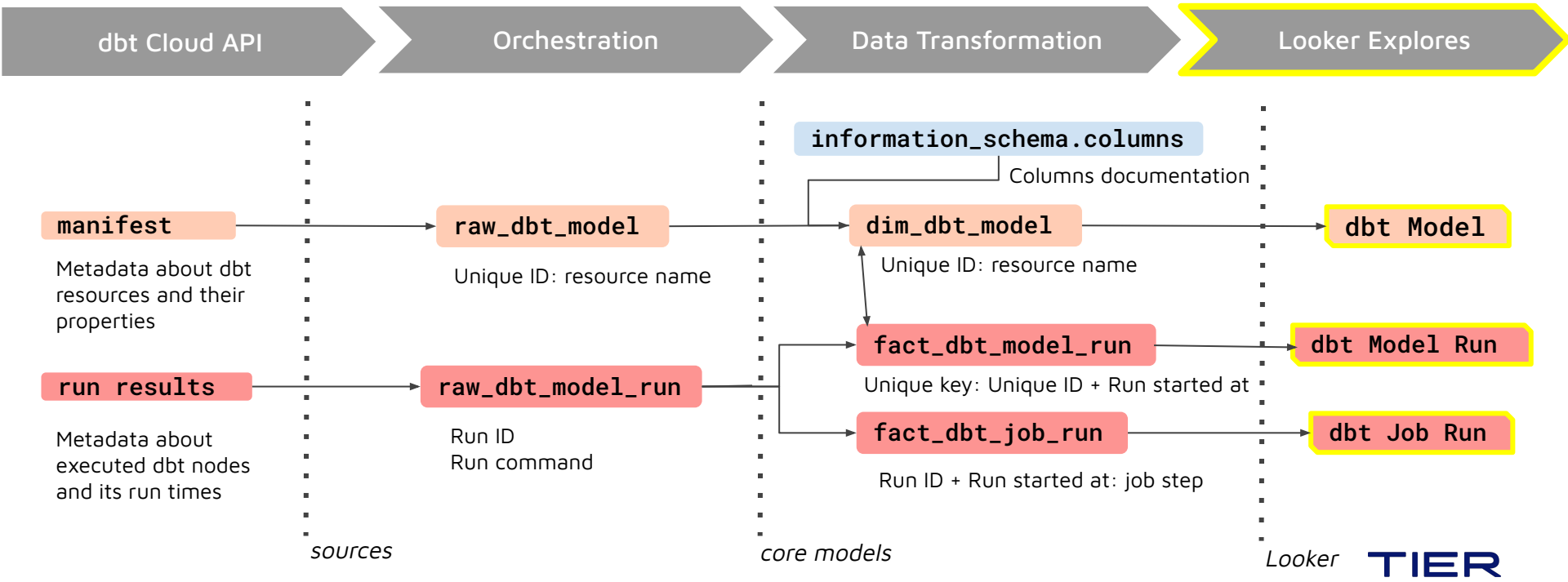
TIER

Data Collection Workflow



TIER

Data Collection Workflow



TIER

Data Collection Workflow - Summary

manifest.json

Metadata about the dbt resources (models, tests and seeds) and their properties - same data available in dbt docs.

dim_dbt_model

Each row represents an available resource in our repository: model, test or seed, with its respective information.

dbt Model explore

Available fields:

- → Resource type: model, test or seed
- → Name and alias
- → Materialization: view, table or incremental model
- → Tags: nightly, weekly,...
- → Description
- → Customised metadata: owner and analytics group
- → Unique key
- → Pre- and post-hook

TIER

Data Collection Workflow - Summary

run_results.json

Metadata about the execution times and results of executed models and scheduled jobs.

fact_dbt_model_run

Each row represents a different run of a dbt resource (model, test, seed), its timing and status

- ➔ Run id, timestamp and command
- ➔ Run reason: scheduled or kicked off by user
- ➔ Result status and execution times
- ➔ Downtime duration

dbt Model Run explore

fact_dbt_job_run

Each row represents an executed step within an job, with its timing and total status (success, error, warn,..)

- ➔ Job id, job started at, job name, run command
- ➔ Job execution times
- ➔ Number of models or tests which got an error, success, pass or warn within each job

dbt Job Run explore

TIER

dbt Performance Dashboard

TIER

dbt Performance Dashboard

This dashboard allows us to **monitor our dbt project's performance** and, therefore, to:

- have a better overview of data **ownership**
- monitor **documentation and test coverage**
- track **downtimes and errors**
- identify **worst-performing** dbt resources to prioritise their refactoring or fix

TIER

dbt Performance Dashboard

Models:

- Number of models by owner and group
- Costs by group
- Materializations
- Resources with description, owner, unique key,...
- Model run times

dbt Performance Dashboard

Scheduled Jobs:

- Errors or warnings in the jobs running on schedule
- Run times by scheduled job
- Job timeseries

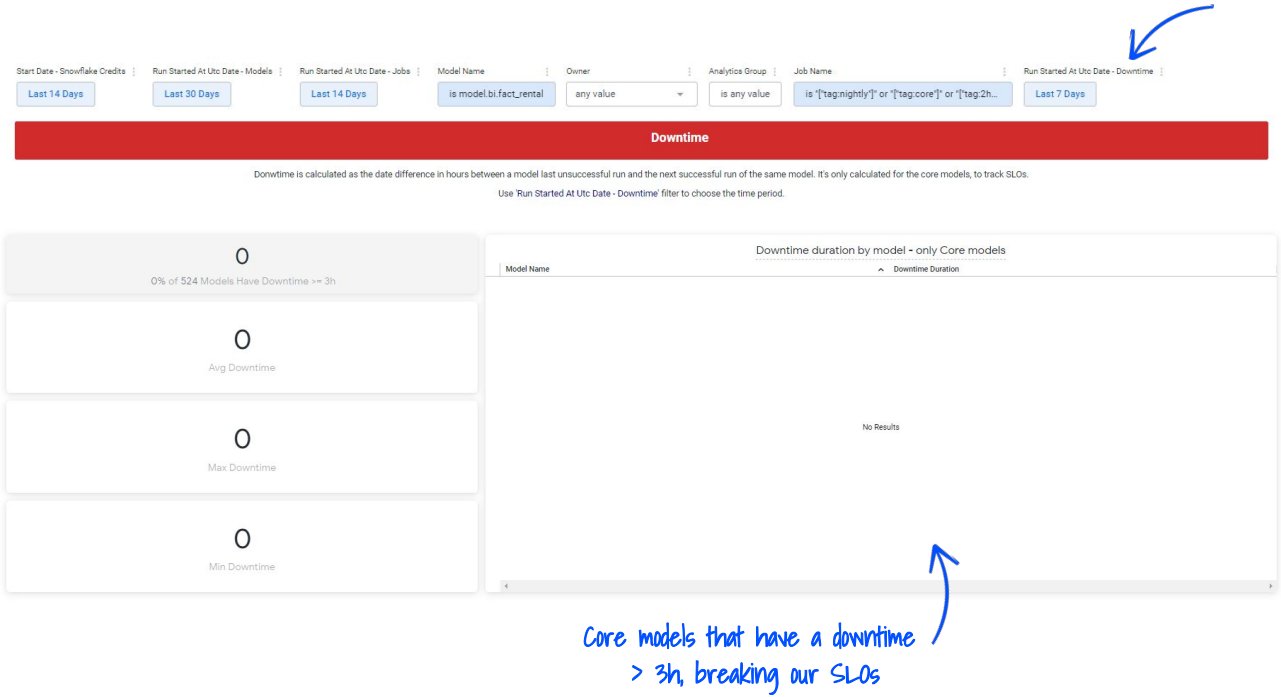
TIER

dbt Performance Dashboard

Downtime:

Models with downtime, minimum, maximum and average downtime, and downtime duration by model.

Filtered by last 7 days.



Core models that have a downtime > 3h, breaking our SLOs

TIER

dbt Performance Dashboard

Filter by your user to check how your models are performing

Start Date - Snowflake Credits	Run Started At Utc Date - Models	Run Started At Utc Date - Jobs	Model Name	Owner	Analytics Group	Job Name	Run Started At Utc Date - Downtime
Last 14 Days	Last 30 Days	Last 14 Days	is model.bi.fact_rental	any value	is any value	is "{"tag:nightly"}" or "{"tag:core"}" or "{"tag:2h...	Last 7 Days

Performance Ranking

Execution times by model - Top 15 Worst-Performing Models (filter by owner or group)

	Model Name	Materialization	Median Execution Times - minutes	Avg Execution Times - minutes	Min Execution Times - minutes	Max Execution Times - minutes
1	model.bi.segments_new_inactive_reactivated_retained_rollup	table	115.61	110.50	93.82	116.94
2	model.bi.dm_idle_time_snapshot	table	84.02	82.47	57.91	112.68
3	model.bi.fact_consumer_app_event	incremental	64.07	61.93	37.43	89.04
4	model.bi.fact_customer_acq_channel_and_revenue	table	40.59	41.34	36.53	50.21
5	model.bi.fact_vehicle_state_snapshot	table	40.24	35.83	15.19	69.02
6	model.bi.dim_zone_hexbin_mapping	table	30.39	30.48	14.70	46.32
7	model.bi.fact_customer_acq_cohorts	incremental	29.19	29.44	23.90	34.71
8	model.bi.raw_consumer_app_preprocess	incremental	23.73	24.48	13.52	35.48
9	model.bi.dim_crm_segmentation_weekly	incremental	22.06	19.32	1.03	27.37
10	model.bi.oba_contact_and_churn	table	13.29	12.95	5.90	22.94
11	model.bi.dm_revenue	table	11.38	12.47	0.00	41.43

Execution times by test - Top 15 Worst-Performing Tests

	Model Name	Median Execution Times - minutes	Min Execution Times - minutes	Max Execution Times - minutes
1	test.bi.unique_segment_tracks_raw_event_id.d5eda4c89d	8.61	5.62	9.79
2	test.bi.unique_fact_consumer_app_event_event_id.ef823673bd	8.44	2.85	12.87
3	test.bi.unique_raw_consumer_app_preprocess_event_id.be02ef0d96	7.60	2.92	16.52
4	test.bi.dbt_utils_unique_combination_of_columns_fact_consumer_app_event_session_id__event_order.92722636b9	6.34	4.44	12.59
5	test.bi.relationships_fact_consumer_app_session_session_id__session_id__ref_fact_consumer_app_event_.3b7cae9303...	3.42	2.16	5.66
6	test.bi.not_null_raw_rental_services_relocation_confirmed_rental_id.45b7633083	1.22	0.02	3.08
7	test.bi.unique_raw_rental_services_relocation_confirmed_rental_id.624c5969ce	1.08	0.02	8.31
8	test.bi.not_null_raw_rental_services_pre_auth_rental_id.1928f01329	0.81	0.02	2.11
9	test.bi.assert_frontend_event_session_referential_integrity	0.67	0.44	1.31
10	test.bi.not_null_dim_rental_geo_route_route_geo_linestring.f7147f2420	0.53	0.40	0.96
11	test.bi.unique_adjust_events_raw_event_id.cfe9fb2859	0.52	0.25	1.12

Execution times by seed - Top 15 Worst-Performing Seeds

	Model Name	Median Execution Times - seconds	Min Execution Times - seconds	Max Execution Times - seconds
1	seed.bi.erp_packages	12.1738	12.1738	12.1738
2	seed.bi.erp_currencies	7.1341	7.1341	7.1341
3	seed.bi.markets_base	3.8105	3.2839	4.1429
4	seed.bi.dm_test_customer_segmentation_crm	2.8486	2.4357	3.2614

Ranking performing:

Top 15 worst-performing models, tests and seeds.

It can be filtered by owner and group, so we can identify our own models.

TIER

SLOs

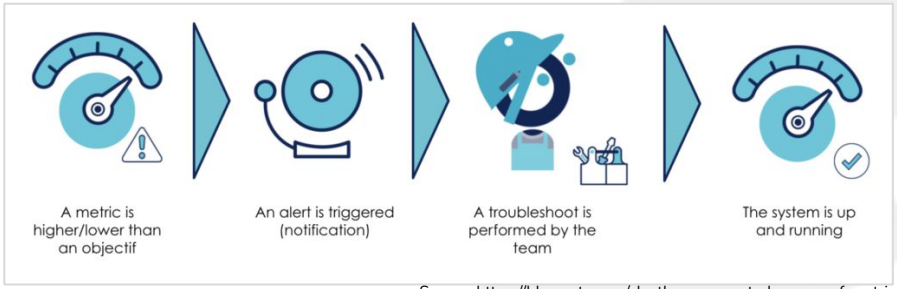
SLOs: Overview

A **metric** is basically a core value/service being measured.

An **SLO** (Service Level Objective) is essentially a level that we want to hit with a particular metric.
It is usually expressed as a percentage over a period of time

Customers place food orders in a restaurant.

- Did the customer receive the order they requested? This is seen as the metric
- The percentage of orders serviced in 10 minutes (Ideally, you would like all orders for food to be serviced/responded to within 10 minutes) - SLO

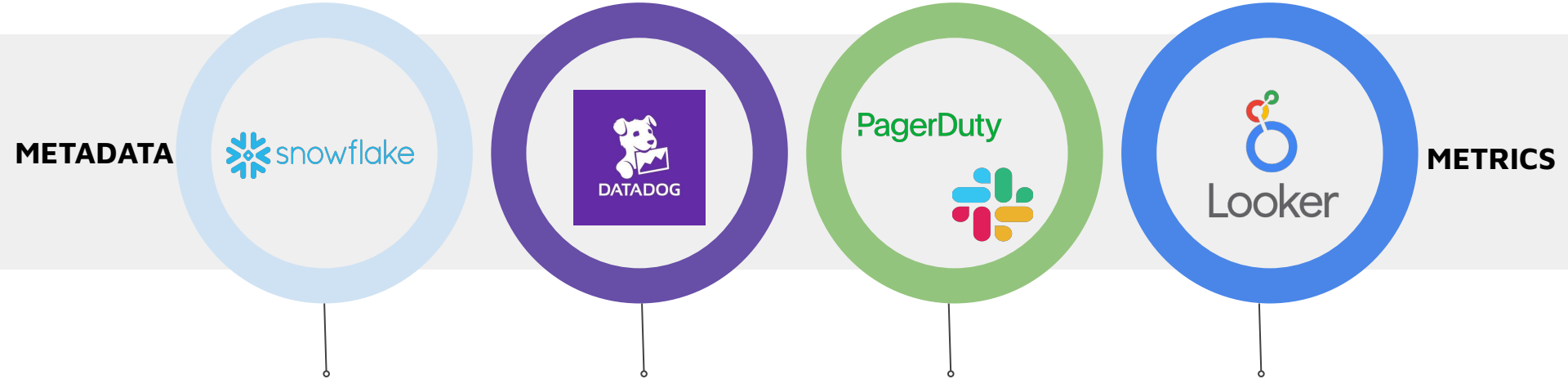


Source: https://blog.octo.com/slo-the-unexpected-power-of-metrics/

Data Pipeline For SLO Reporting

TIER

Data Pipeline For SLO Reporting



Snowflake

Metrics &
SLOs Creation

Incidents & Alerts

Swarm Tracking
dashboard

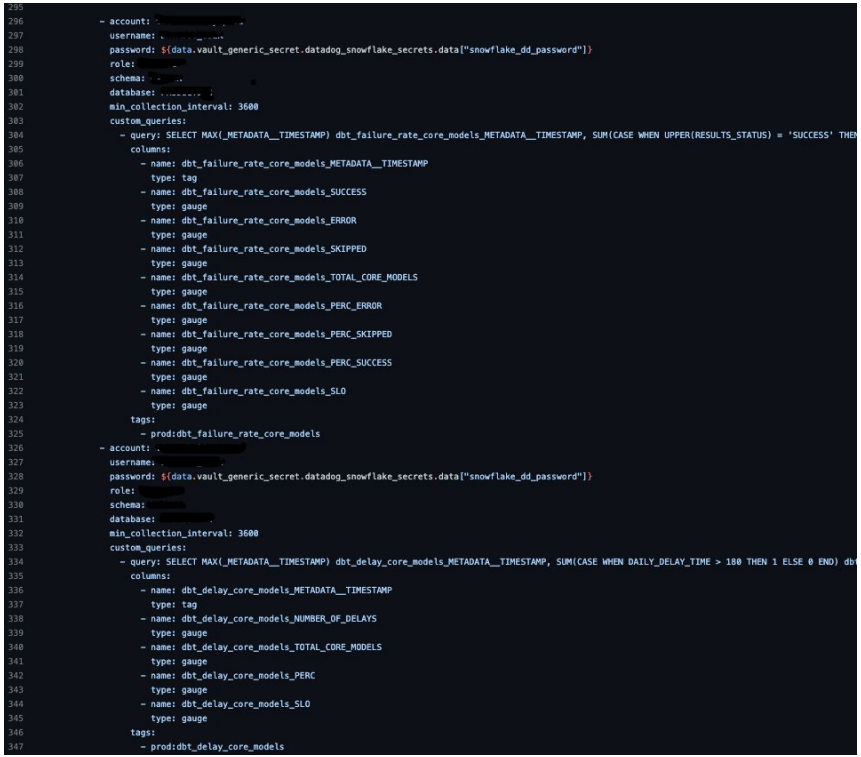
TIER

Data Pipeline For SLO Reporting

Snowflake | Metrics and SLOs | Incidents and alerts | Swarm tracking dashboard

Persisted Model on Snowflake:

The metrics that are being used to track SLOs are then configured in yml files that feed data from Snowflake to Datadog.



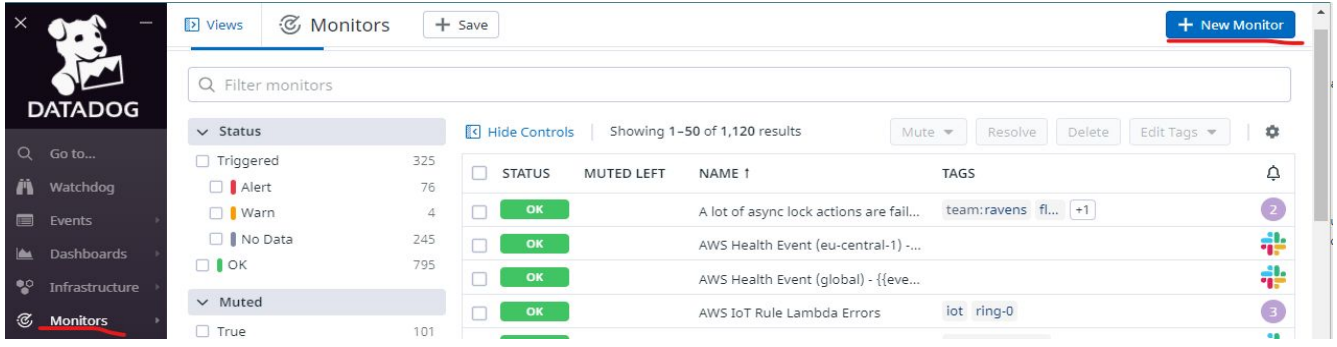
TIER

Data Pipeline For SLO Reporting

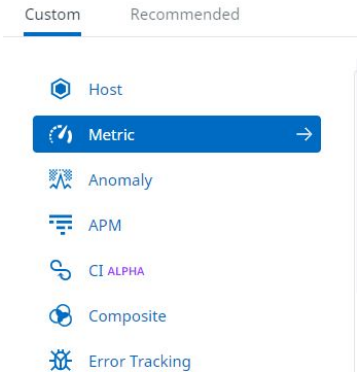
Snowflake → **Metrics and SLOs** → Incidents and alerts → Swarm tracking dashboard

Definition, creation and notification of monitors and SLO thresholds:

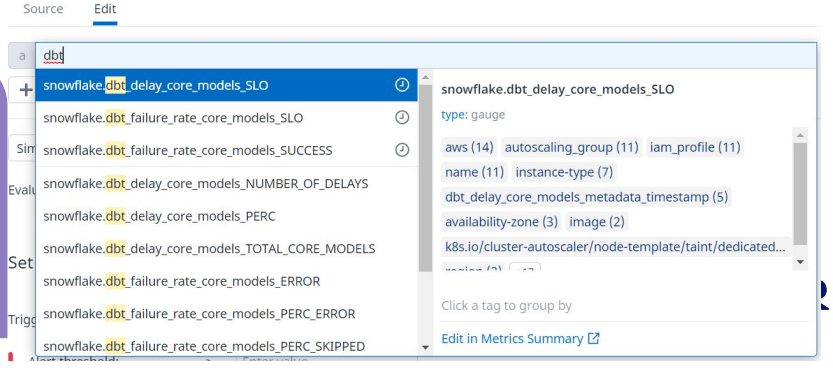
Once in DataDog click Monitors / New Monitor



Then select Metric



In the area we can search for the metric as it was defined on the YAML file

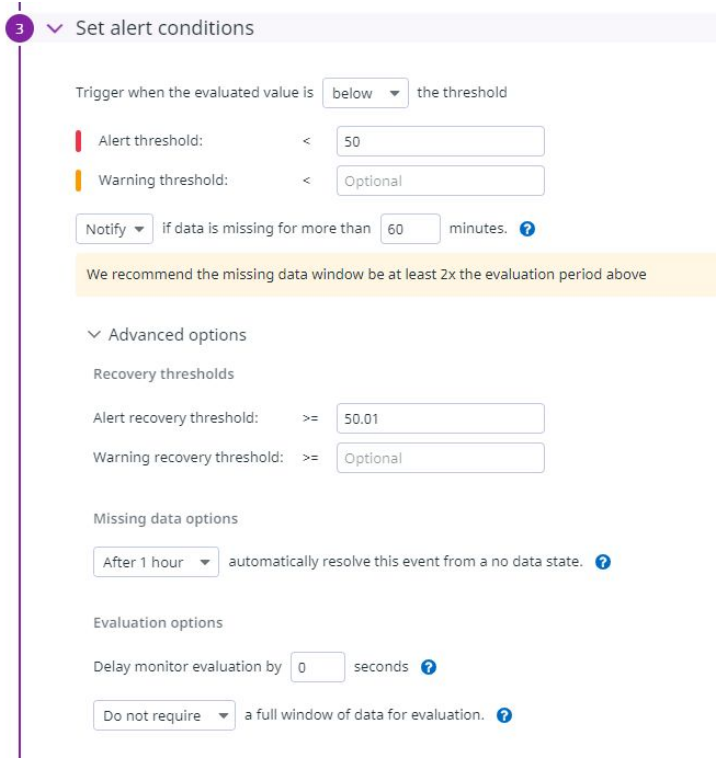


Data Pipeline For SLO Reporting

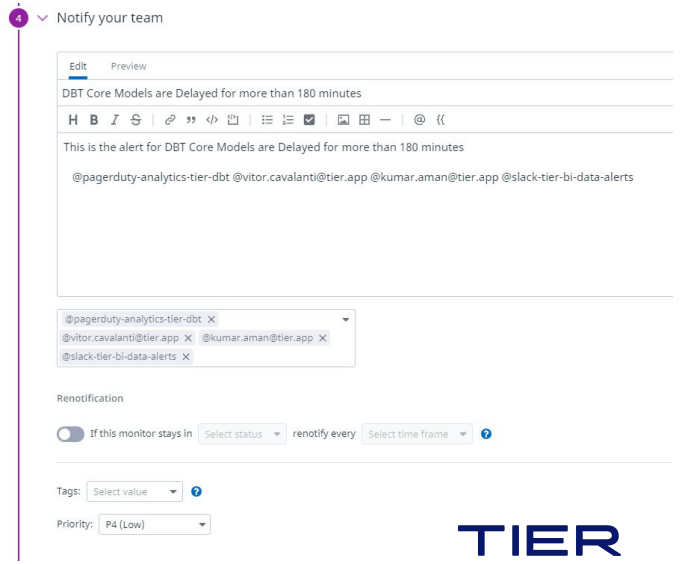
Snowflake | Metrics and SLOs | Incidents and alerts | Swarm tracking dashboard

Next we can define the alert conditions.

- Thresholds for alerts and warnings
- Recovery thresholds
- Missing data alerts



Finally we can define notification options. Who should be notified once there is an alert or warning and also what is the priority (P1 to P5) and what service on PagerDuty should be triggered (for example @pagerduty-analytics-tier-dbt)

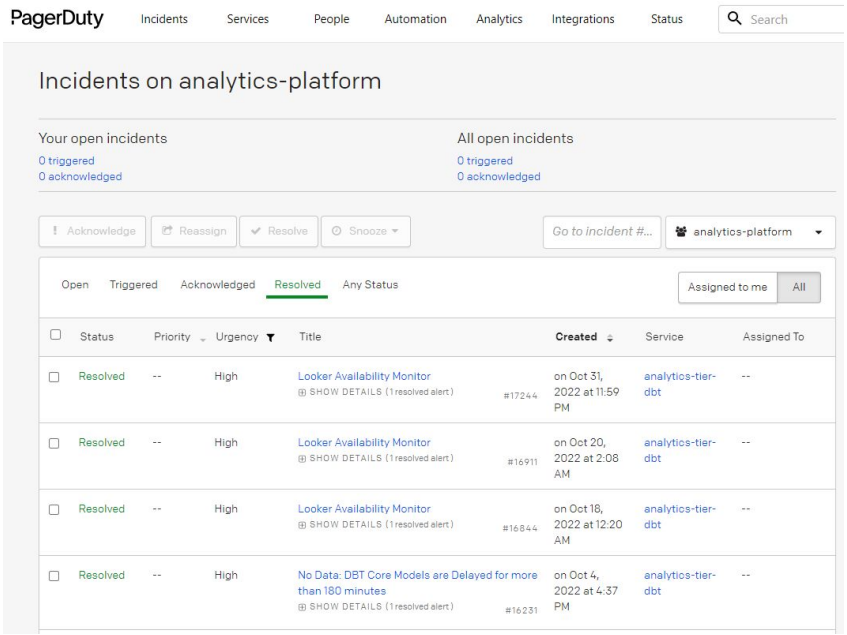


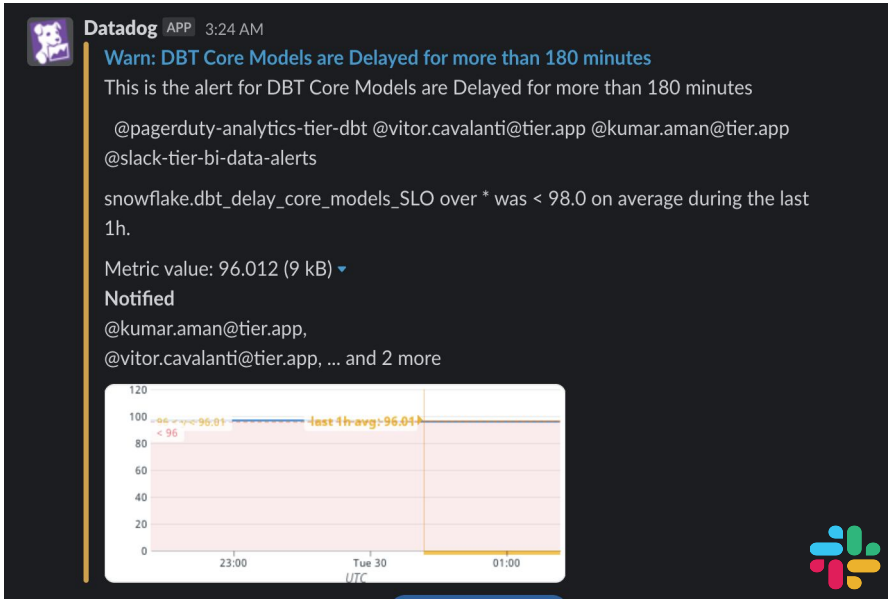
TIER

Data Pipeline For SLO Reporting

Snowflake → Metrics and SLOs → Incidents and alerts → Swarm tracking dashboard

Metrics and SLOs failures/warnings created as incidents on Pagerduty and notifications sent to slack

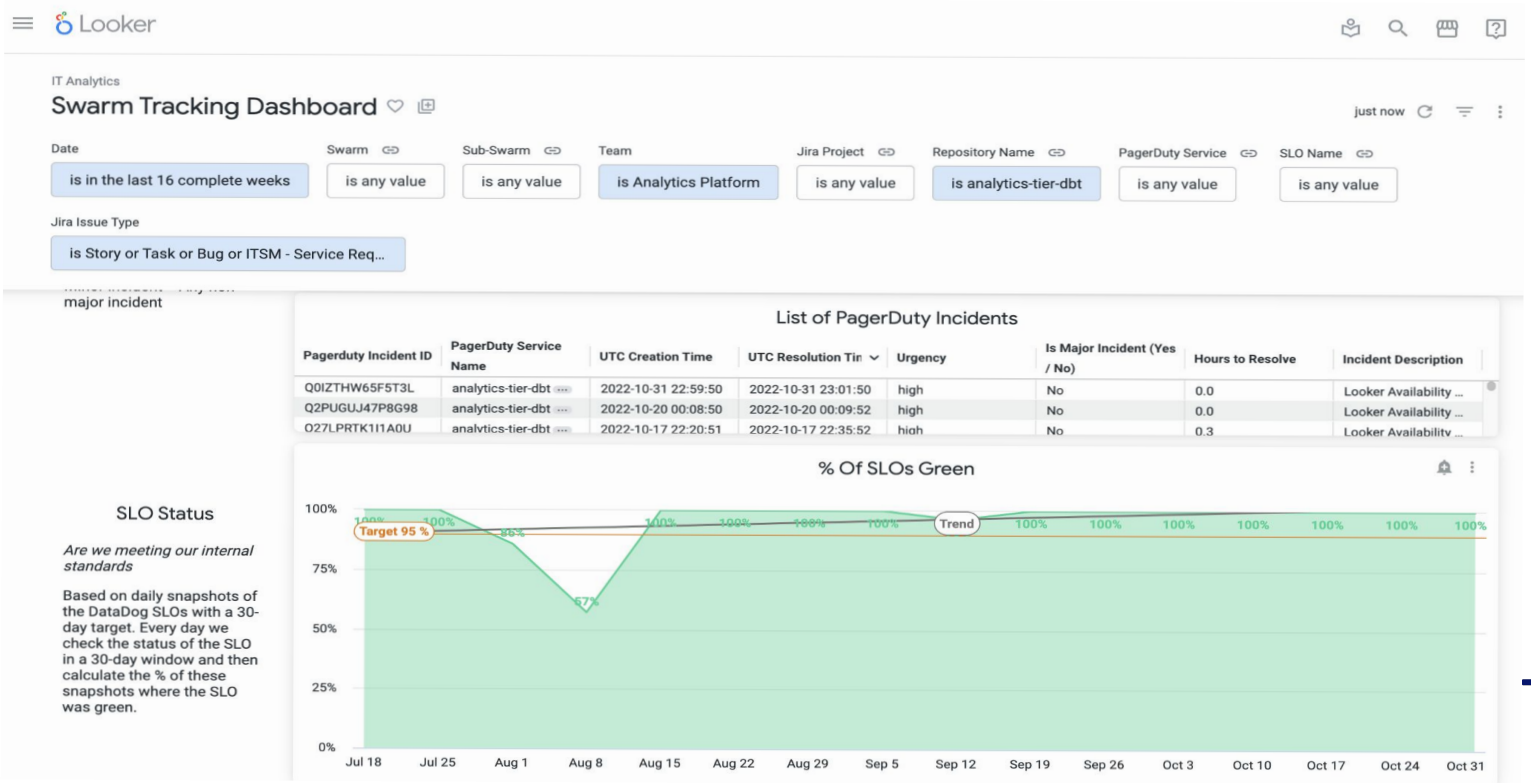




TIER

Data Pipeline For SLO Reporting

Snowflake → Metrics and SLOs → Incidents and alerts → Swarm Tracking Dashboard



TIER

Data Pipeline For SLO Reporting - Summary

Snowflake → Metrics and SLOs → Incidents and alerts → Swarm Tracking Dashboard

The metrics are configured in **yml files** that feed data from Snowflake to Datadog

snowflake

Definition, creation and notification of **SLO thresholds** on DataDog

DATADOG

Metrics and SLOs failures/warnings created as **incidents** on Pagerduty

PagerDuty

Notifications are send to Slack

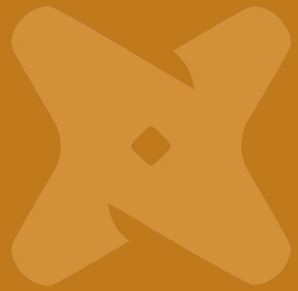
slack

The **Swarm Tracking Dashboard** combines all SLOs for teams in the data organization

Looker

TIER

Thank you for your attention!
Time for questions 👂



TIER

Thank you!

Stay tuned for upcoming events and other content

Berlin dbt Meetup Group
#local-berlin Slack Channel

dbt

Before you leave...

Please share your feedback with us!