

Robuste Datenpipelines mit Open-Source-Software erstellen

Sascha Dittmann
Customer Engineer (Smart Analytics)
Google Germany GmbH





How did we get here?

Foto von Fallon Michael auf Unsplash



Foto von Dmitry Dreyer auf Unsplash

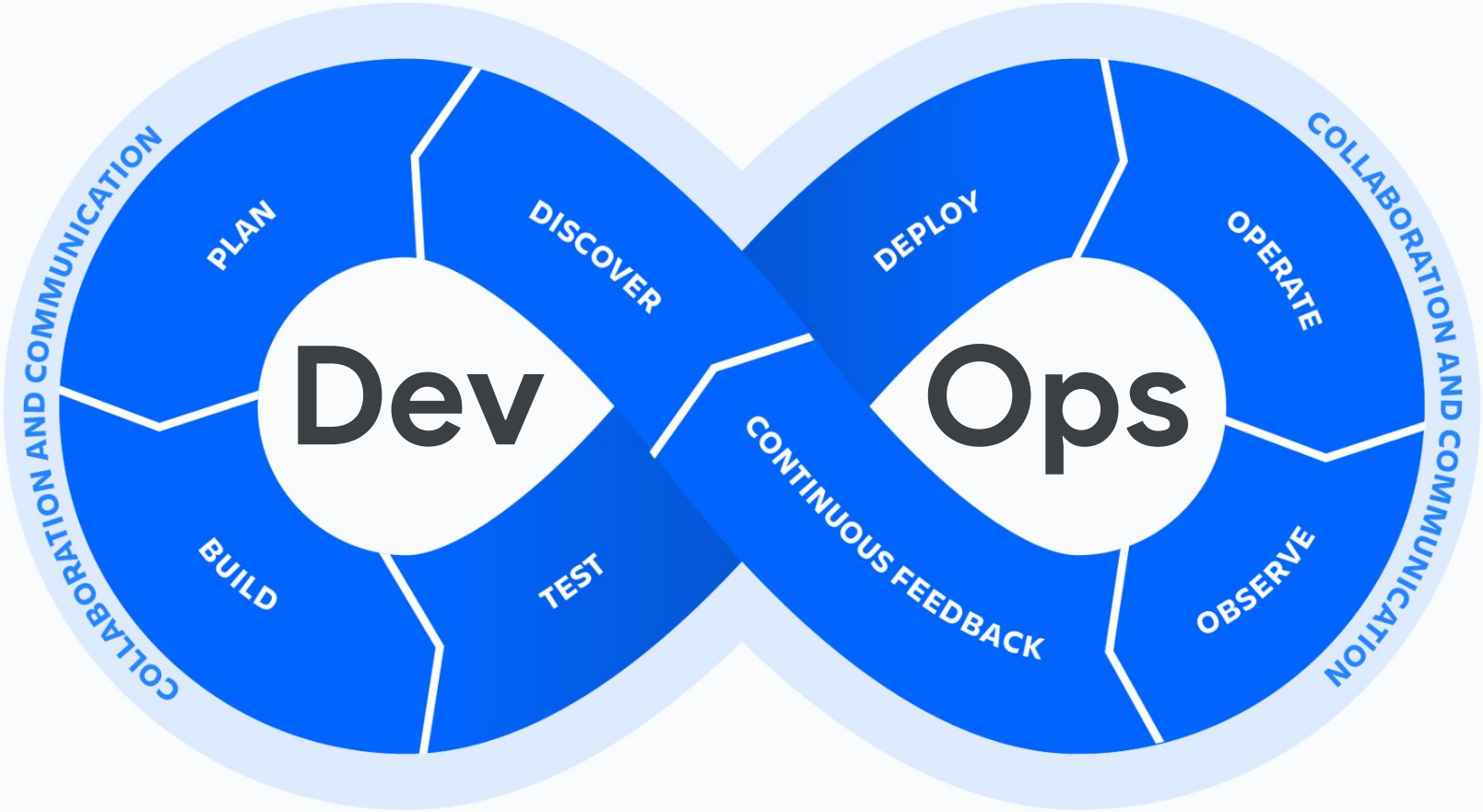


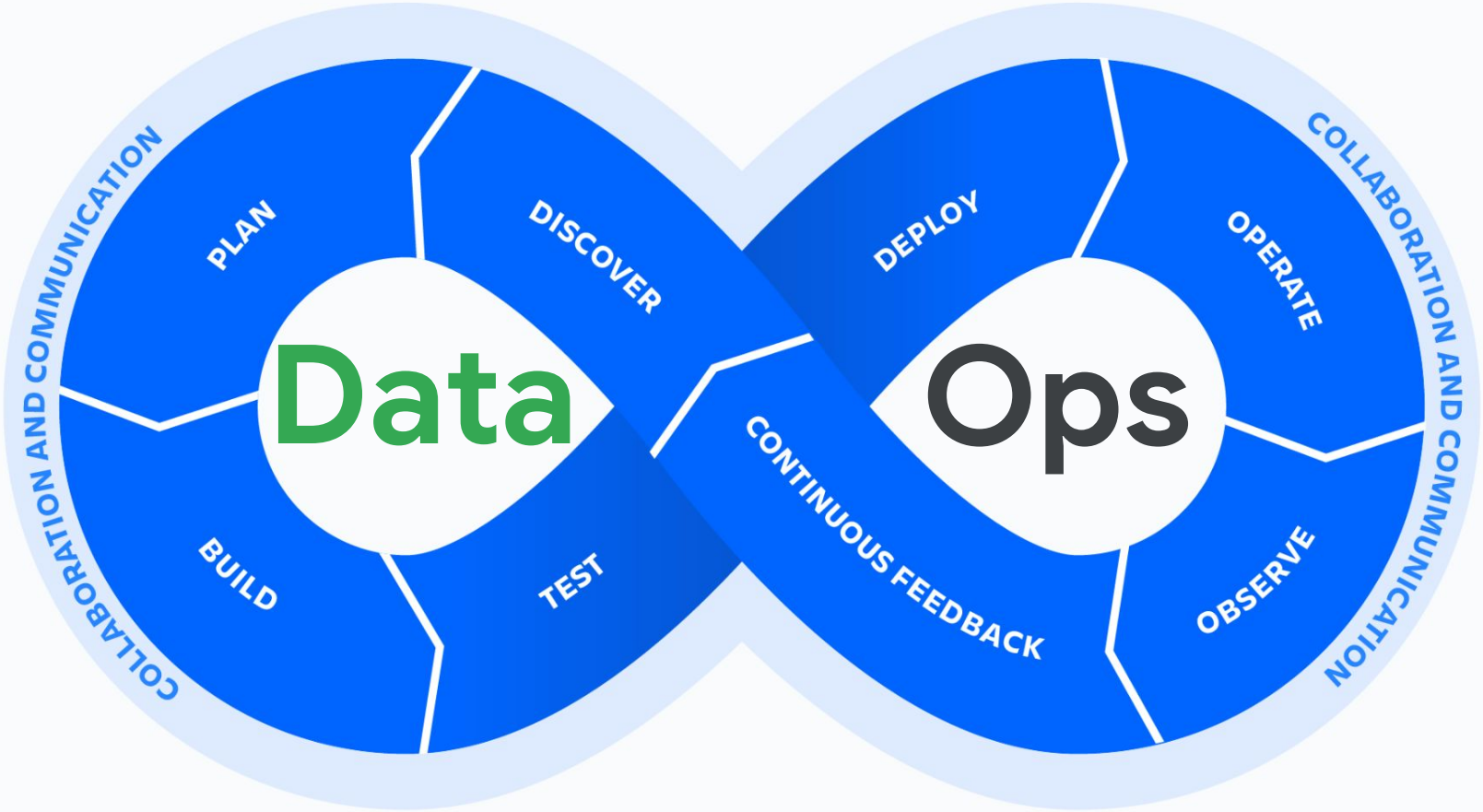
Storage
has become very cheap



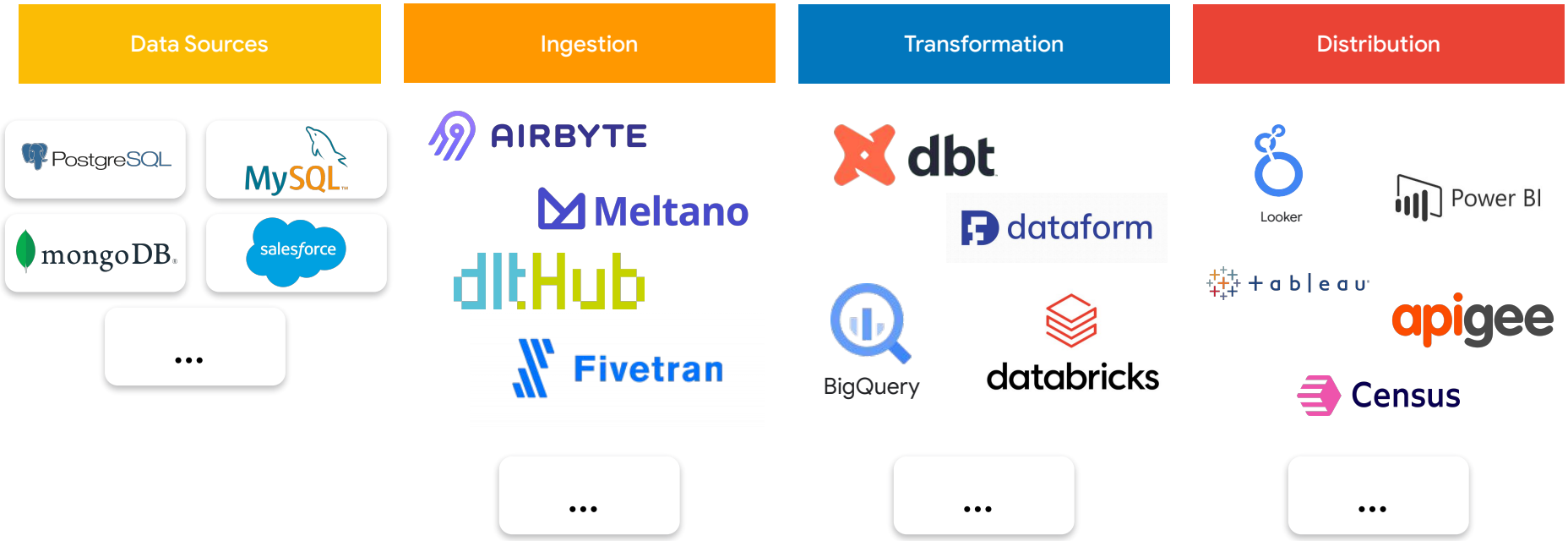
Raw data
is highly valuable

Fotos von Jingming Pan und Ashley auf Unsplash


COLLABORATION AND COMMUNICATION
PLAN
DISCOVER
BUILD
TEST
Dev
DEPLOY
OPERATE
OBSERVE
CONTINUOUS FEEDBACK
Ops
COLLABORATION AND COMMUNICATION


Data
Ops
PLAN
DISCOVER
BUILD
TEST
DEPLOY
OPERATE
OBSERVE
CONTINUOUS FEEDBACK
COLLABORATION AND COMMUNICATION
COLLABORATION AND COMMUNICATION

Overview


Data Sources
Ingestion
Transformation
Distribution
PostgreSQL
MySQL
mongoDB.
salesforce
...
AIRBYTE
Meltano
dltHub
Fivetran
...
dbt
dataform
BigQuery
databricks
...
Looker
Power BI
tableau
apigee
Census
...
Orchestration
Apache Airflow
GitHub Actions
Cloud Composer

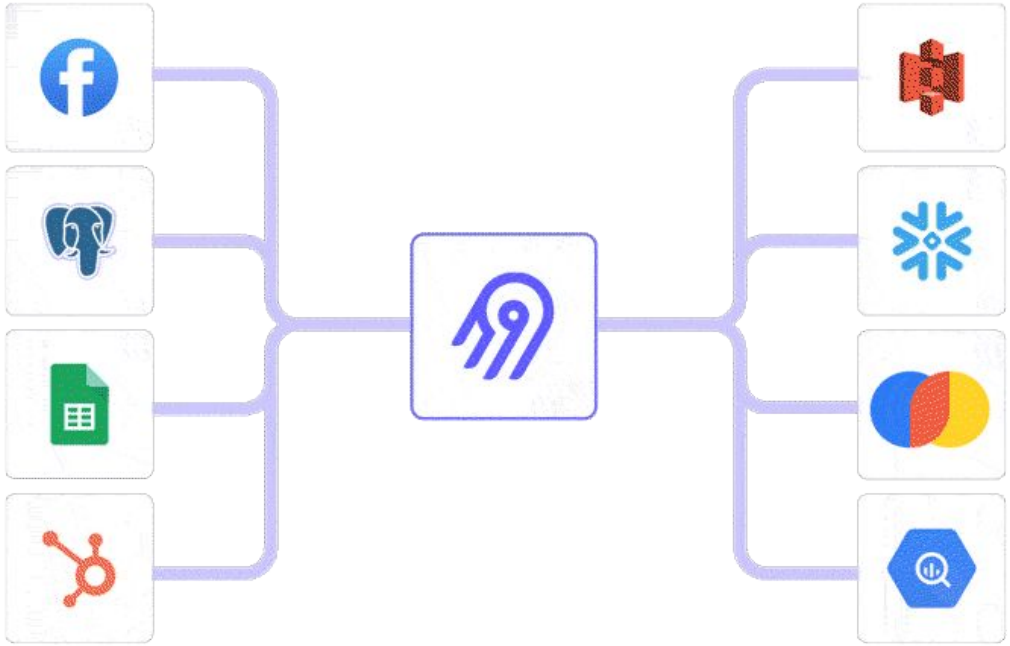
Extract & Load



Foto von wd toro 🇮🇩 auf Unsplash

Airbyte - Open Data Movement Platform

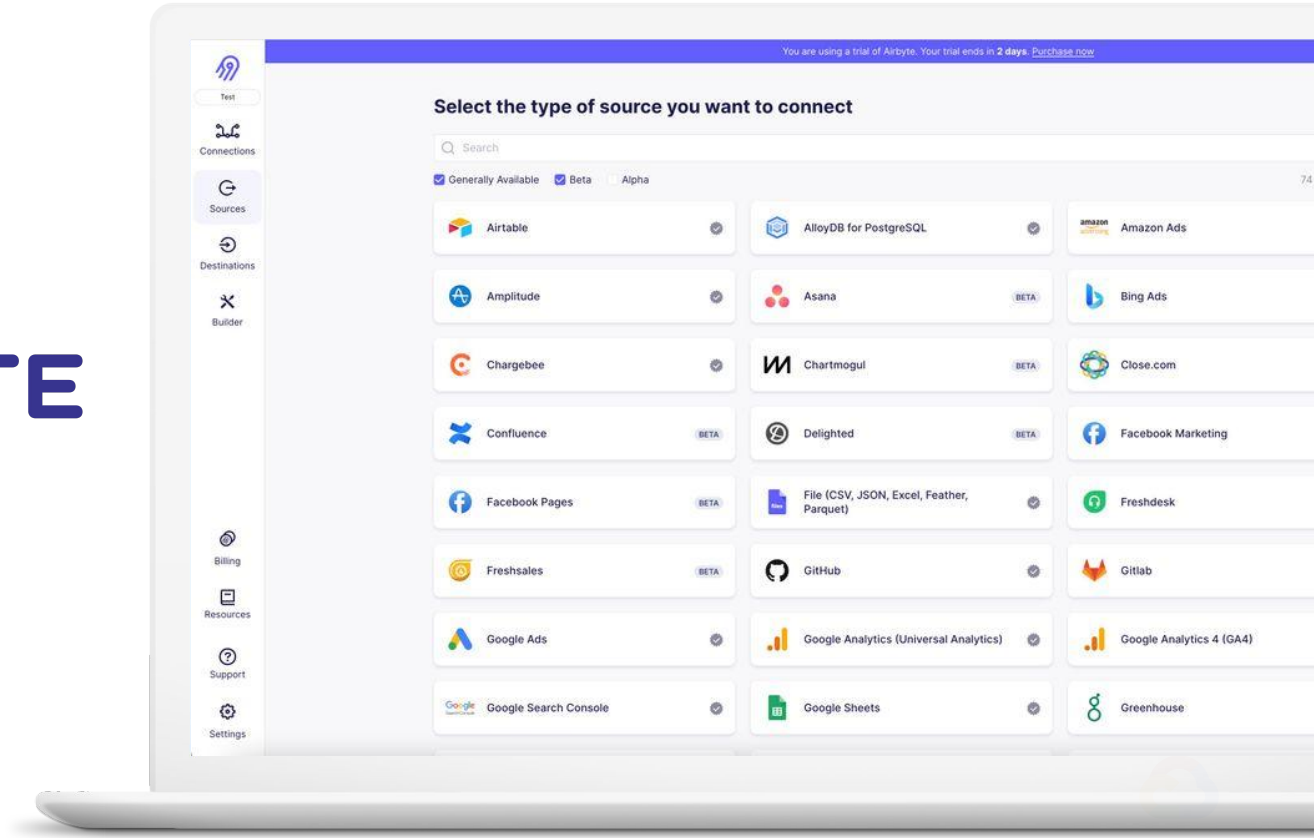
AIRBYTE



https://airbyte.com

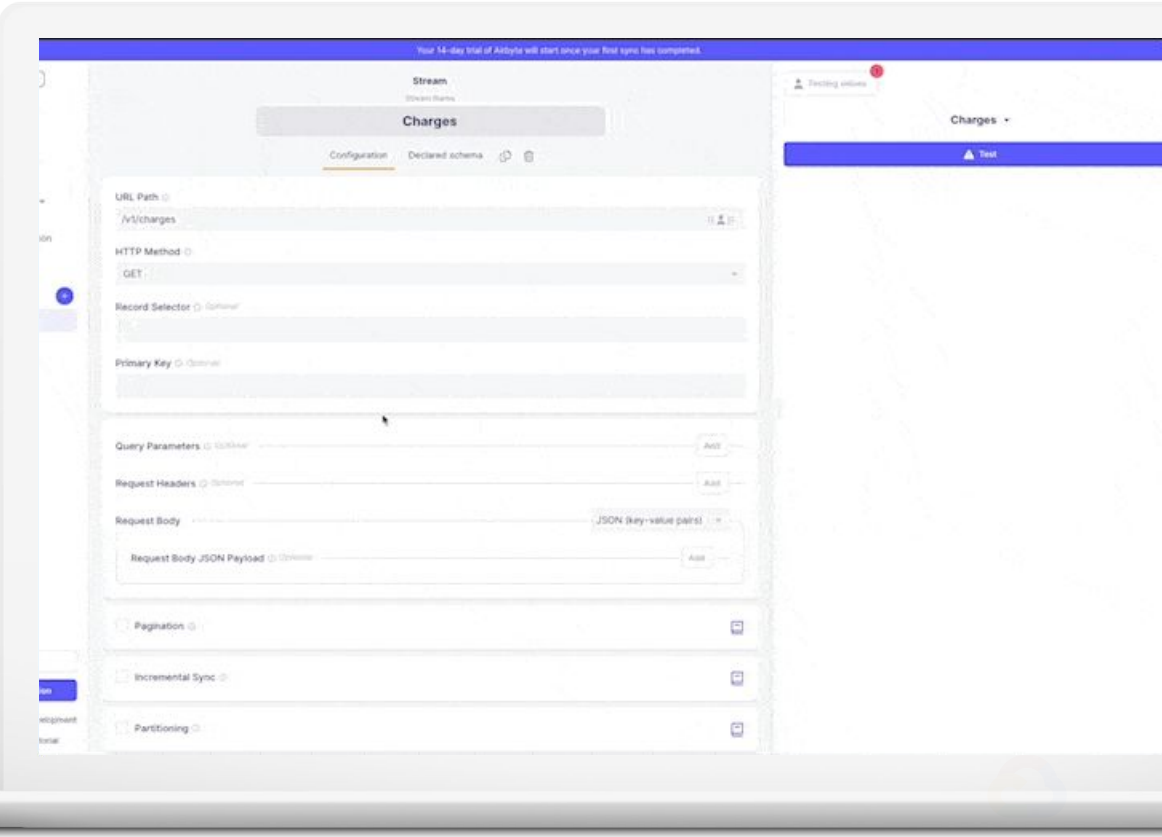
Airbyte - Open Data Movement Platform

AIRBYTE



https://airbyte.com

Airbyte - Open Data Movement Platform



https://airbyte.com

data load tool (dlt)




```python
if __name__=='__main__':
    # configure the pipeline with your destination details
    pipeline = dlt.pipeline(
        pipeline_name='github_api_pipeline',
        destination='duckdb',
        dataset_name='github_api_data'
    )

    # print credentials by running the resource
    data = list(github_api_resource())

    # print the data yielded from resource
    print(data)

    # run the pipeline with your parameters
    load_info = pipeline.run(github_api_source())

    # pretty print the information on data that was loaded
    print(load_info)
```

https://dlthub.com

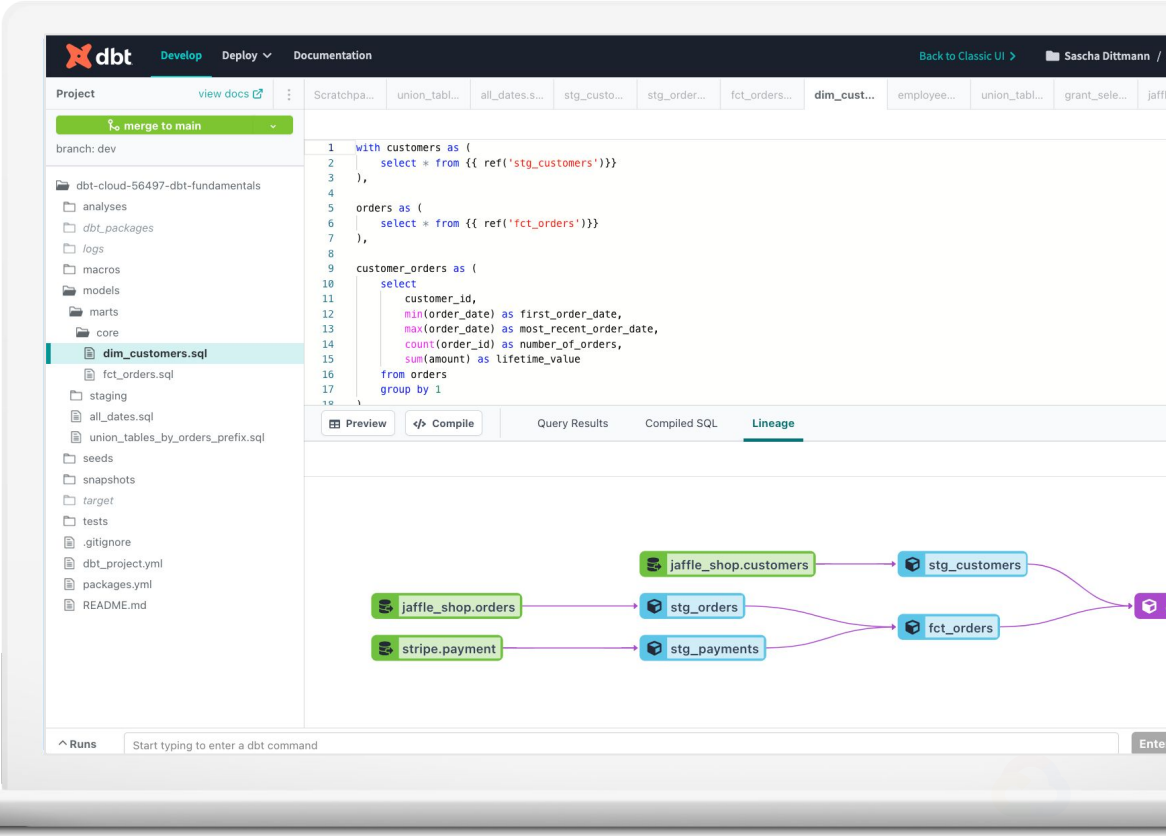
Let’s see that in action





Foto von Richard Horvath auf Unsplash

data build tool (dbt)



https://www.getdbt.com/

Data Build Tool



https://www.getdbt.com/

But there is more...

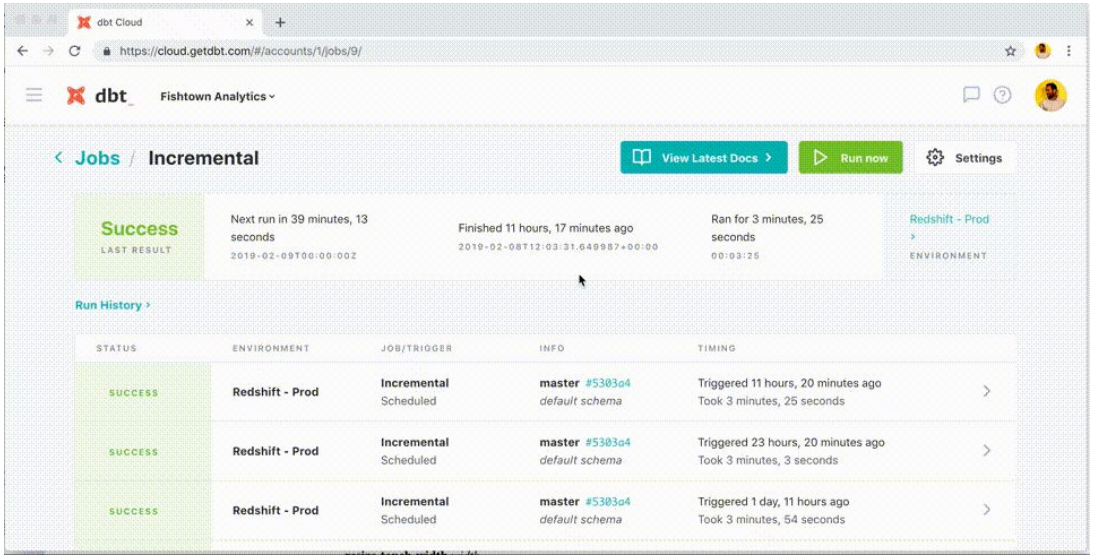
- Tests
- Documentation
- Linting
- Monitoring

Dbt Core vs Dbt Cloud

Problems 2 | Output | whejna-orange-dataops

```
23:03:49  2 of 7 PASS dbt_expectations_expect_row_values_to_have_recent_data_my_first_dbt_model_ingestion_ts__day__3  [PASS in 1.02s]
23:03:49  3 of 7 START test dbt_expectations_expect_table_row_count_to_equal_my_first_dbt_model_2  [RUN]
23:03:50  3 of 7 PASS dbt_expectations_expect_table_row_count_to_equal_my_first_dbt_model_2  [PASS in 0.86s]
23:03:50  4 of 7 START test not_null_my_first_dbt_model_id ..... [RUN]
23:03:51  4 of 7 PASS not_null_my_first_dbt_model_id ..... [PASS in 0.91s]
23:03:51  5 of 7 START test not_null_my_second_dbt_model_id ..... [RUN]
23:03:52  5 of 7 PASS not_null_my_second_dbt_model_id ..... [PASS in 0.87s]
23:03:52  6 of 7 START test unique_my_first_dbt_model_id ..... [RUN]
23:03:53  6 of 7 PASS unique_my_first_dbt_model_id ..... [PASS in 0.84s]
23:03:53  7 of 7 START test unique_my_second_dbt_model_id ..... [RUN]
23:03:54  7 of 7 PASS unique_my_second_dbt_model_id ..... [PASS in 0.84s]
23:03:54
23:03:54  Finished running 7 tests in 6.41s.
23:03:54
23:03:54  Completed successfully
23:03:54
23:03:54  Done. PASS=7 WARN=0 ERROR=0 SKIP=0 TOTAL=7
whejna@cloudshell:~/dbt-quickstart/hello_world (whejna-orange-dataops)$
```

Dbt Core



Dbt Cloud

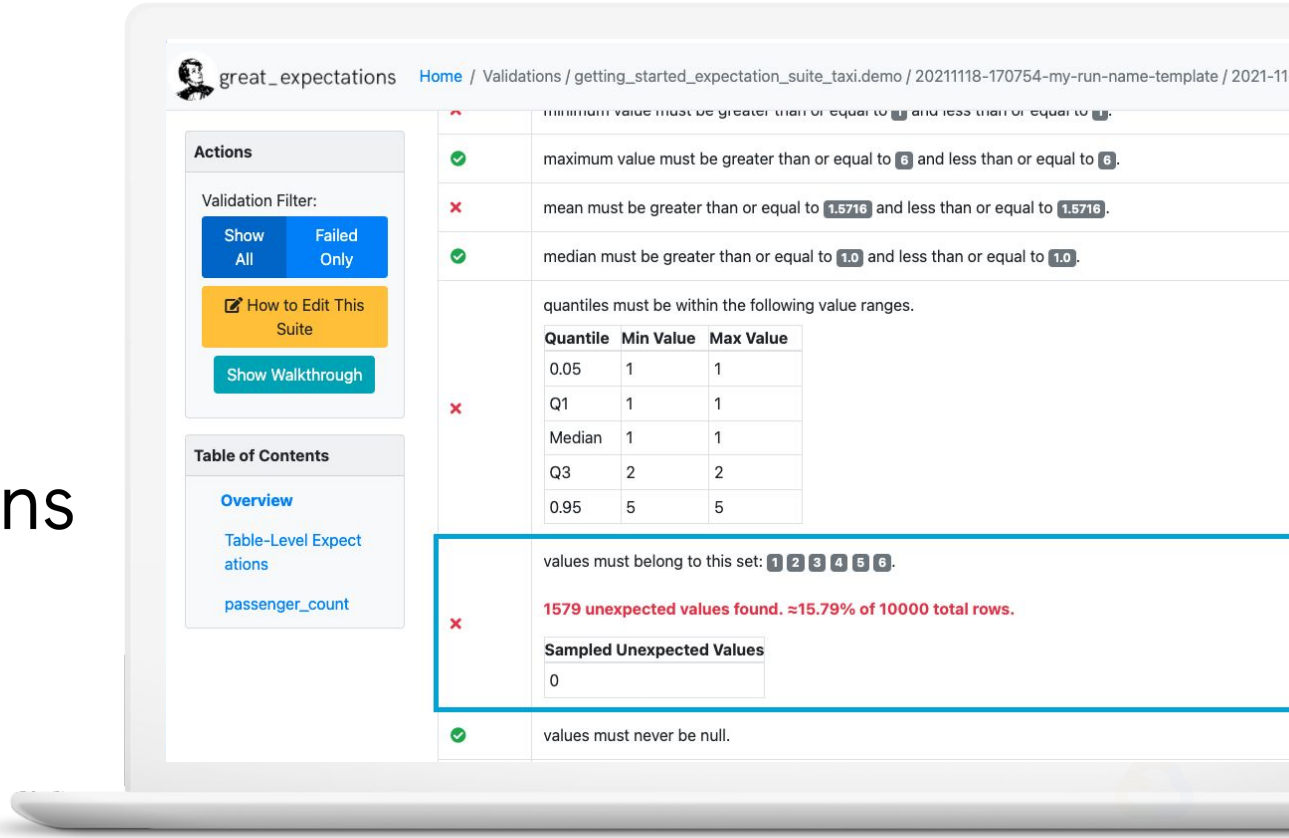
Let’s see that in action



More Tests...

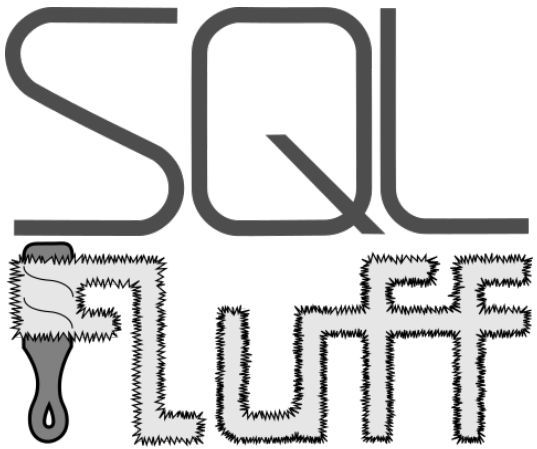


great_expectations



https://greatexpectations.io/

Linter




```sql
SELECT
  this_column,
  my_table.THAT_COLUMN AS this_name,
  some_number *10 AS a_bigger_number
from my_table
```

```
$ sqlfluff lint test.sql
== [test.sql] FAIL
L:   3 | P:   5 | L028 | Qualified reference 'my_table.THAT_COLUMN'
                       | found in a single table select this is
                       | inconsistent with previous references.
L:   3 | P:  14 | L014 | Inconsistent capitalisation of unquoted
                       | identifiers.
L:   4 | P:  18 | L006 | Operators should be followed by a space.
```



https://www.sqlfluff.com/

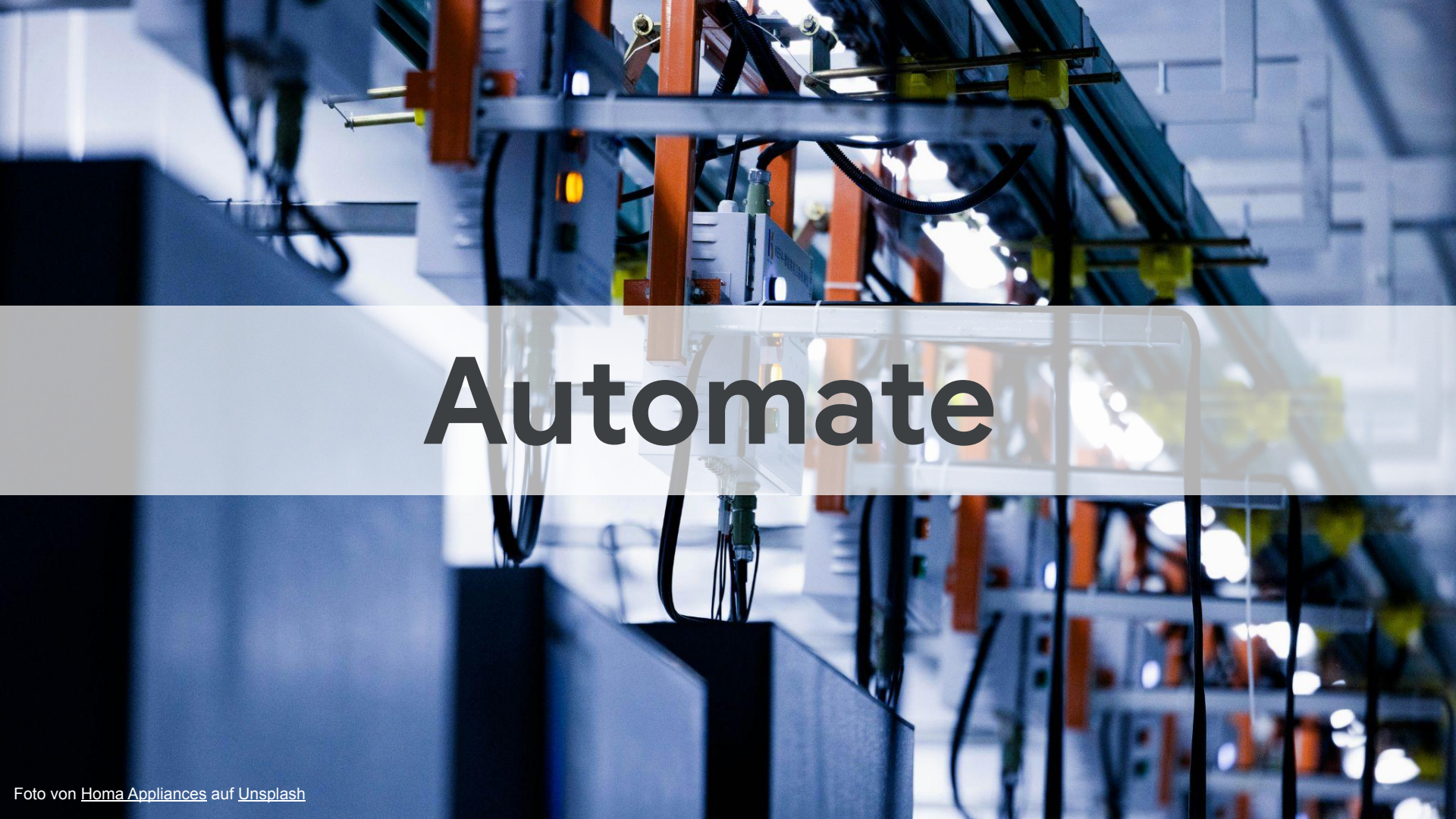
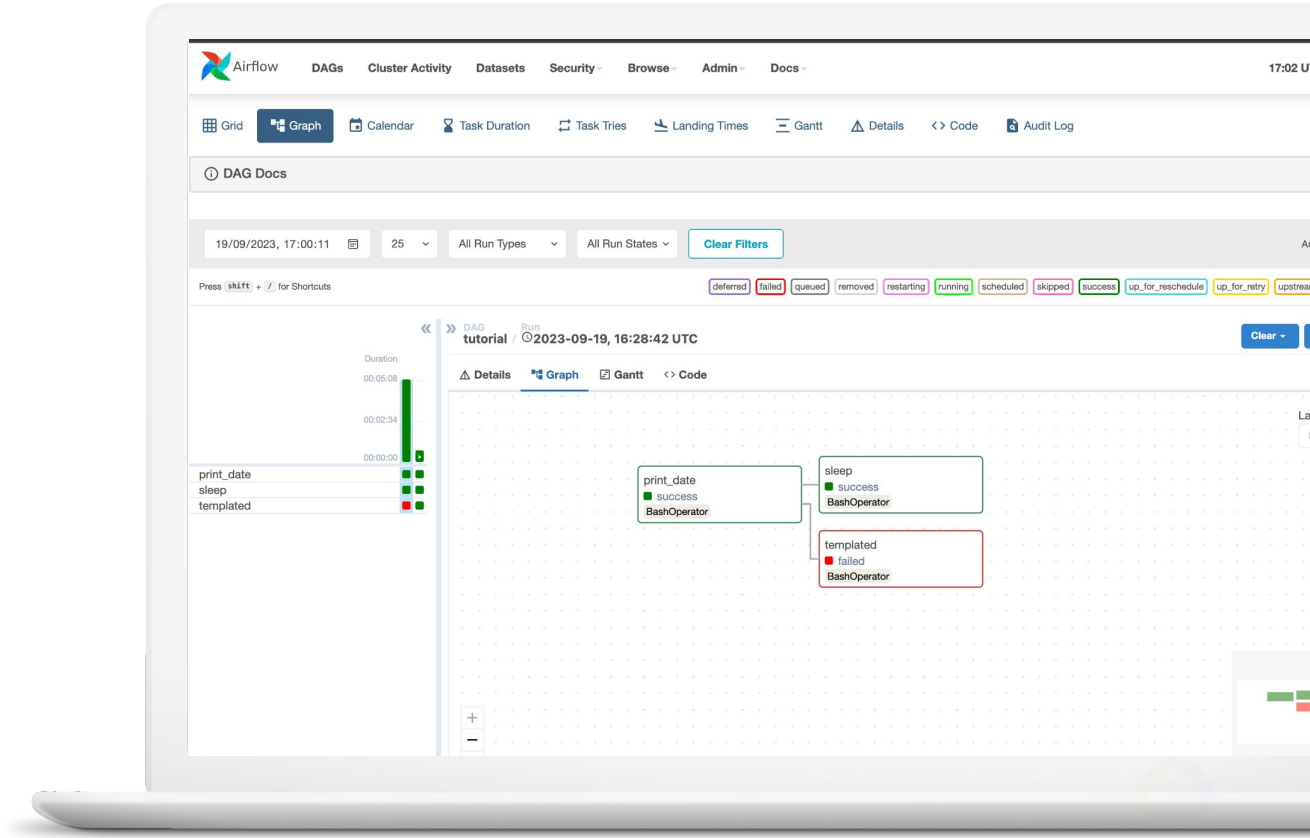


Foto von Homa Appliances auf Unsplash

Apache Airflow





https://www.getdbt.com/

Apache Airflow




```python
from airflow.decorators import dag, task # type: ignore
from datetime import datetime
import random

@dag(
    'random_number_checker',
    start_date=datetime(2024, 1, 1),
    schedule="@daily",
    description='A simple DAG to generate and check random numbers',
    catchup=False
)
def random_number_checker():
    @task()
    def generate_random_number():
        number = random.randint(1, 100)
        print(f"Generated random number: {number}")
        return number

    @task()
    def check_even_odd(number):
        result = "even" if number % 2 == 0 else "odd"
        print(f"The number {number} is {result}.")

    check_even_odd(generate_random_number())

random_number_checker()
```

https://www.getdbt.com/

Let’s see that in action



Feedback



SQL KONFERENZ 2024



THANK YOU