Office of Sustainability

Basin: the MITOS Data Platform

Yu Cheng

June 11, 2024

Warm Up Poll

Yu Cheng 鄭嵎

Sustainability Data Scientist

MIT Office of Sustainability

- Five years working as data scientist/consultant in a few startups.
- Tackling climate change with a unique blend of backgrounds
- Joined MIT Sep. 2023

1 Brief intro of MITOS

Get to know the Office of Sustainability

2 Data Challenges

Observations and learnings from existing data practice

3 Basin: walkthrough

Introducing ‘Basin’, the MITOS data platform, featuring data lineage, cataloging, testing and scheduling.

4 Looking Ahead

Outlining forthcoming goals and potential enhancements to ’Basin’.

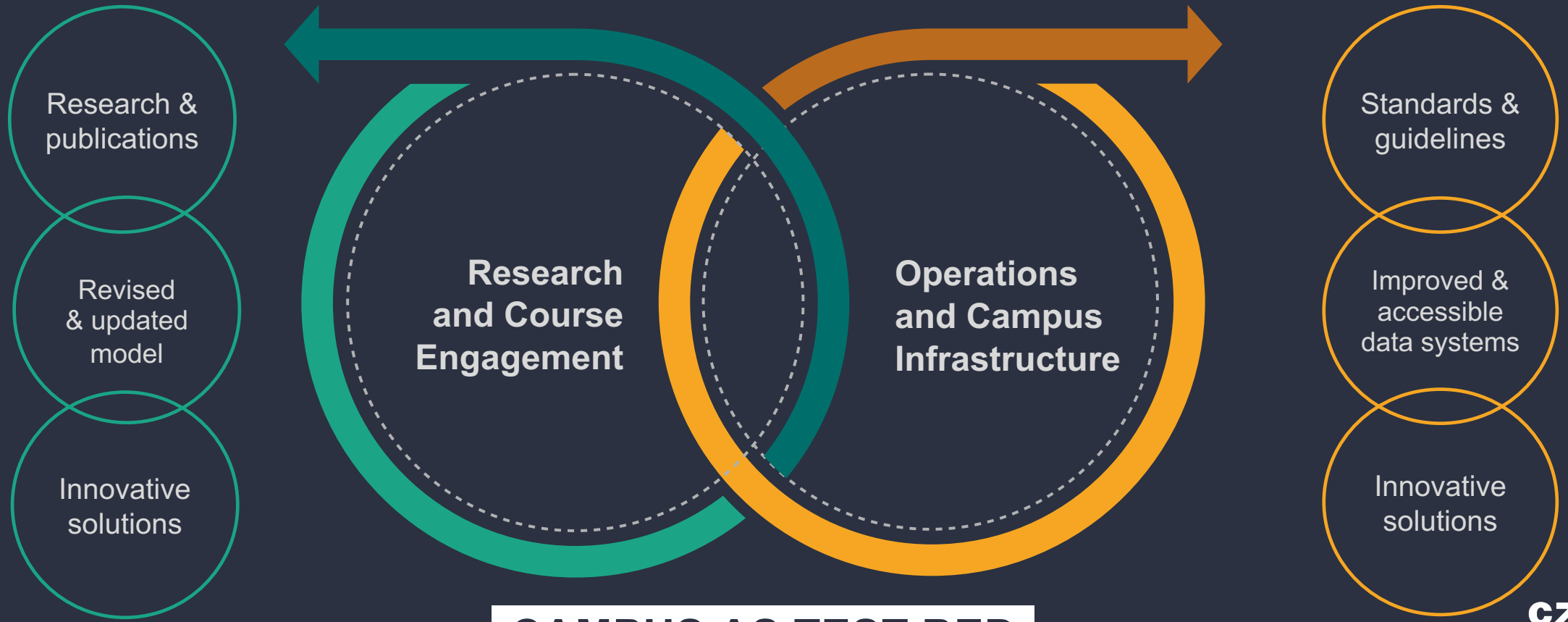
Brief Intro of MITOS

Overview

- Founded in 2013 under Executive Vice President & Treasurer's Office (EVPT) to integrate sustainability across all levels of our campus operations.
- Organizational lead for campus goals on Fast Forward
- E38 Third floor, above the MIT Welcome Center

MITOS Team 2024

How do we solve for sustainability at MIT?



CAMPUS AS TEST BED

campus climate action

Scales of Impact

Projects & Events

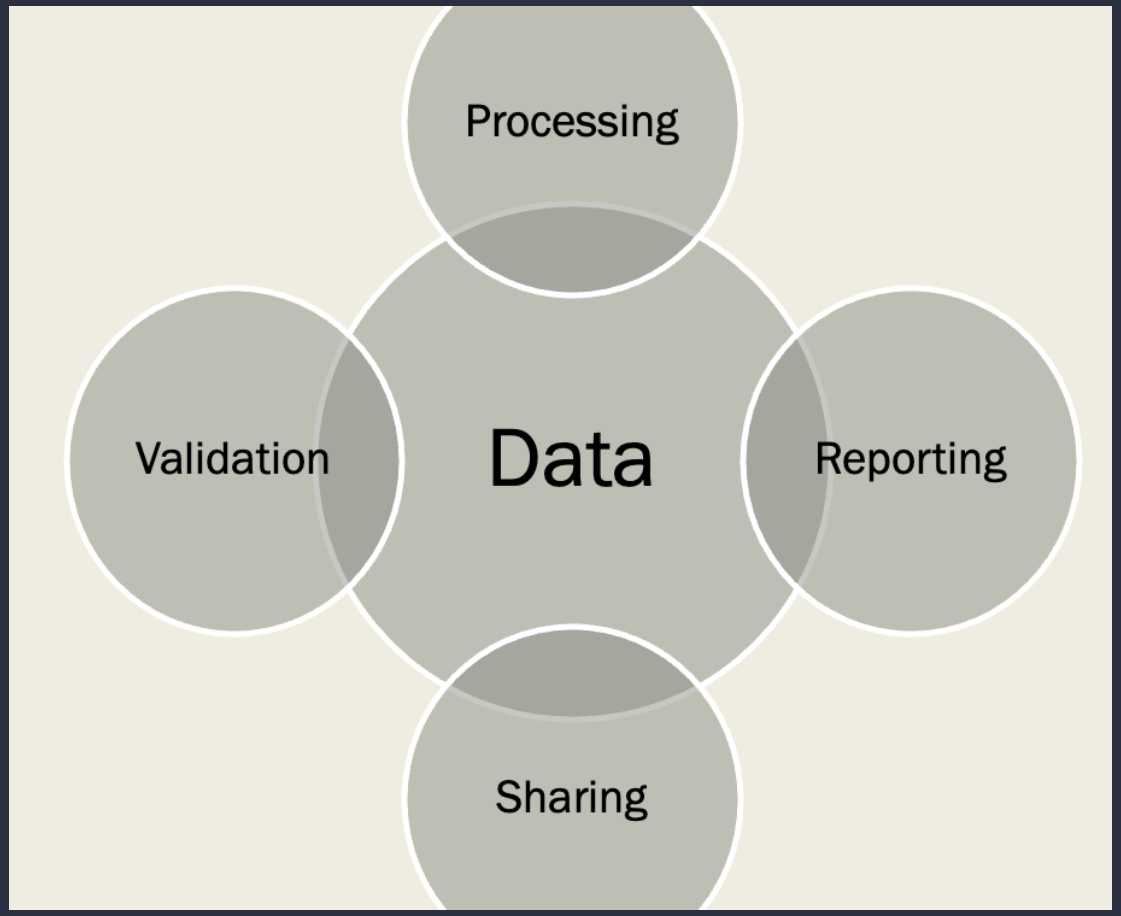
- Annual Sustainability Connect conference
- Food waste fighters, Choose-to-reuse etc.
- Offsite Renewable Energy Project development
- Co-lead Fast-Forward Workstreams and the Decarbonization working group
- **Sustainability Datapool**

Data Challenges

Sustainability Datapool

Challenges

- Excel files
- One-off projects
- Multiple versions and locations of data artifacts
- Limited data quality control, and documentation
- Datahub, Datapool, and Dashboards are not in sync.



Case Study: Scope 3 Emission

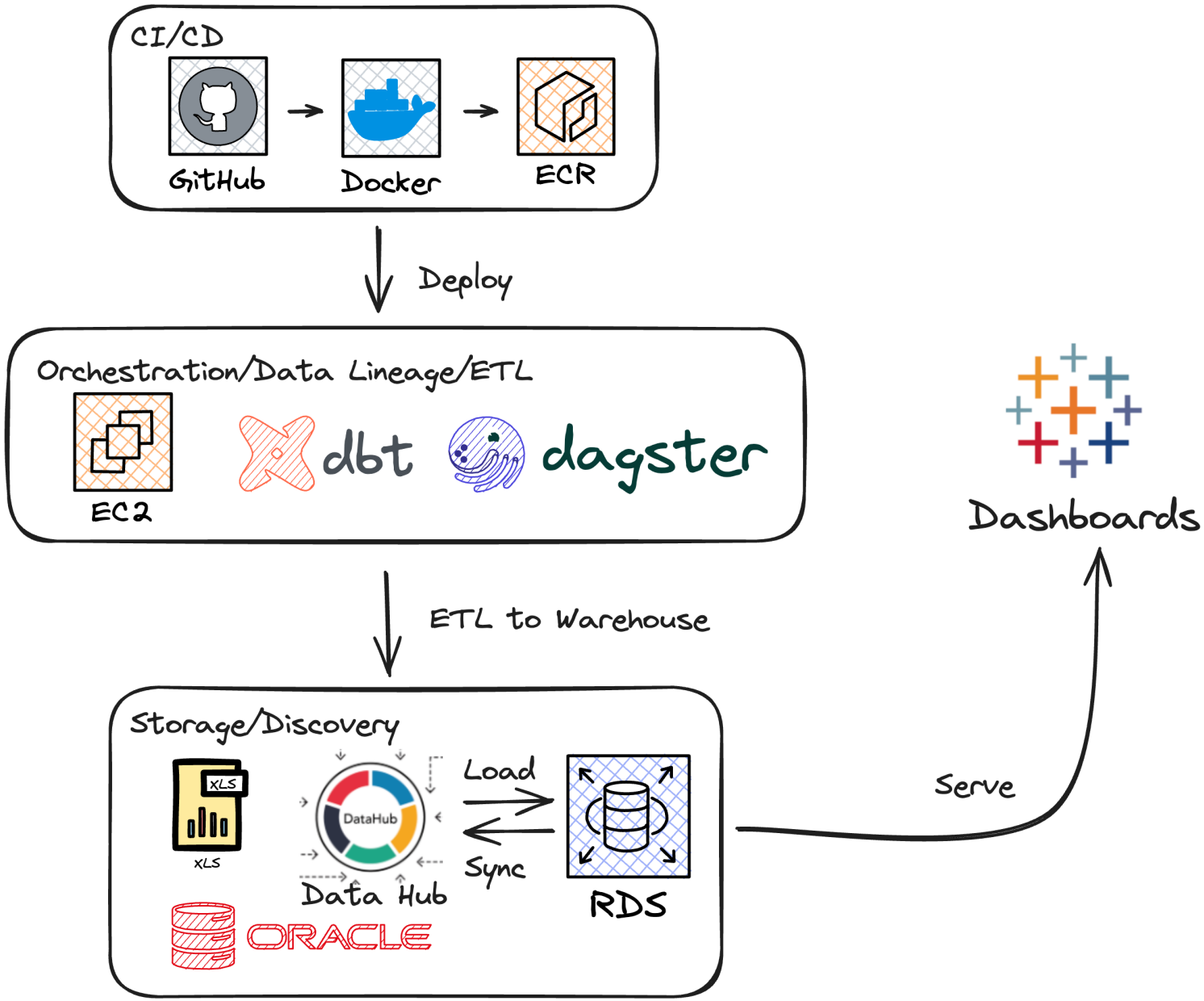
- “tableau_all_scope.xlsx” in multiple folders
- Lacking documentation how to update
- Category Mapping
- Emission Factors are not consistent.

Basin: walkthrough

A river basin consists of many streams.

Objectives

- A centralized repository to track all pipelines
- Transparency in data lineage and quality
- Consistency between reported and shared data
- Modular and easy to build on top.


CI/CD
GitHub
Docker
ECR
Deploy
Orchestration/Data Lineage/ETL
EC2
dbt
dagster
ETL to Warehouse
Storage/Discovery
XLS
XLS
DataHub
Data Hub
Load
Sync
RDS
ORACLE
Serve
Dashboards

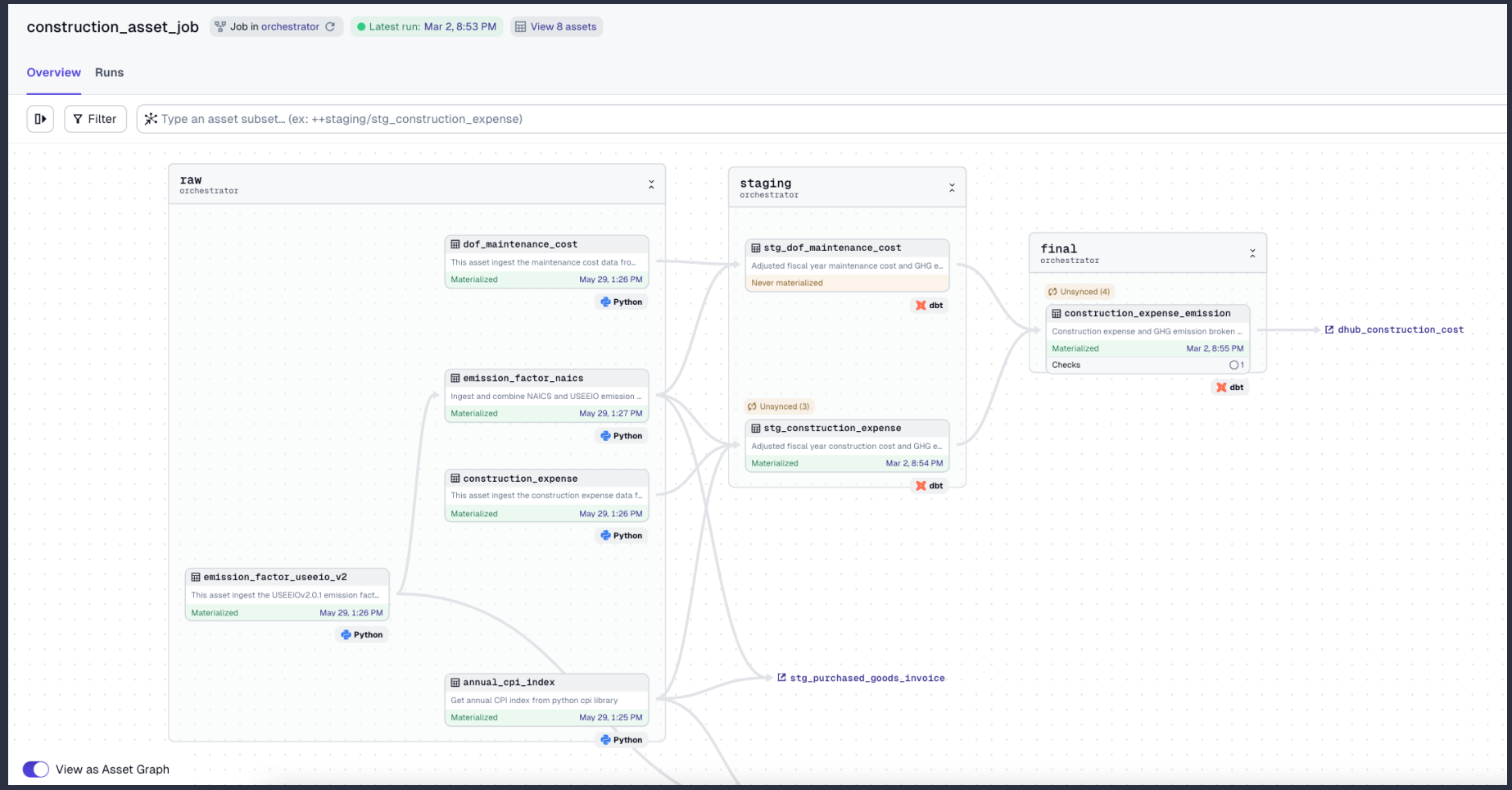
Dagster

- An open-source data orchestrator
- Declarative pipeline using Python
- Scheduling, monitoring, and observability
- Easy integration with other tools and solutions

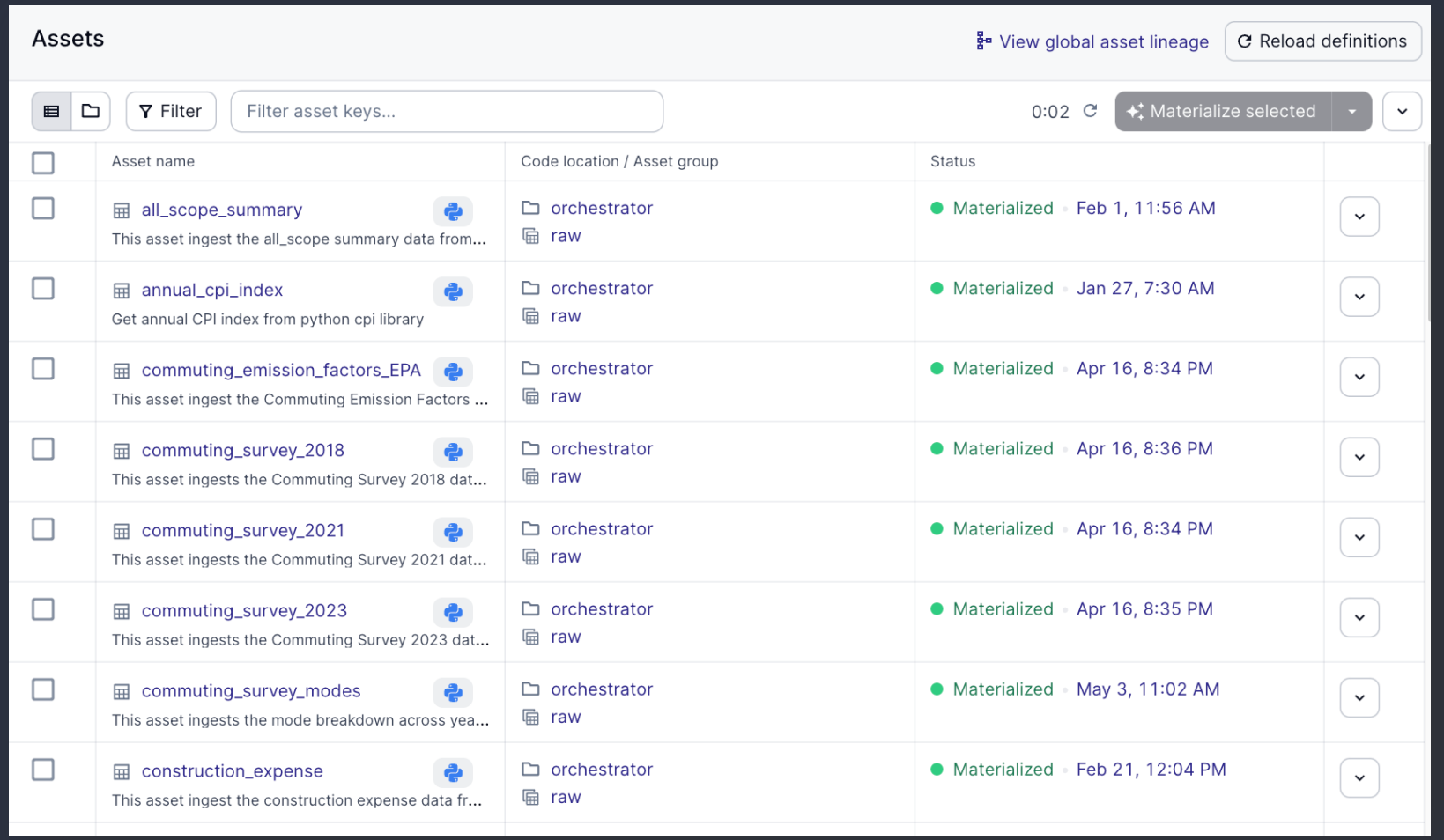
Asset Definition

```python
@asset(
    io_manager_key="postgres_replace",
    compute_kind="python",
    group_name="raw",
)
def expense_emission_mapper(dhub: ResourceParam[DataHubResource]):
    """This asset ingest the expense_type_to_emissions.json from the Data Hub"""
    project_id = dhub.get_project_id("Scope3 Business Travel")
    logger.info(f"Found project id: {project_id}!")
    download_links = dhub.search_files_from_project(project_id, "expense_type_to_emissions.json")
    if download_links is None:
        logger.info("No download links found!")
        return pd.DataFrame()
    response = requests.get(download_links[0], timeout=10)
    if response.status_code == 200:
        payload = json.loads(response.text)
    mapper = {v: key for key, value in payload.items() for v in value}
    df = pd.DataFrame(list(mapper.items()), columns=["expense_type", "emission_category"])
    return df
```

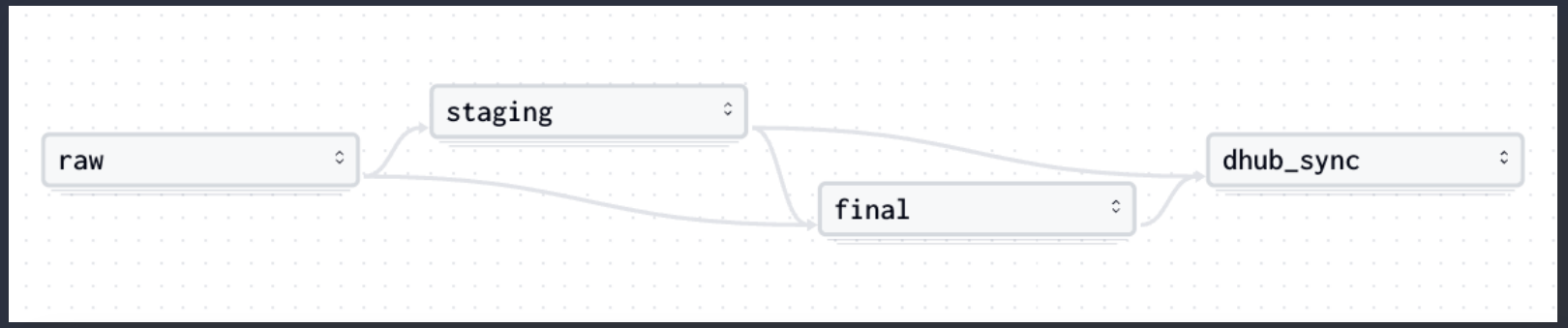
Data Lineage and Observability

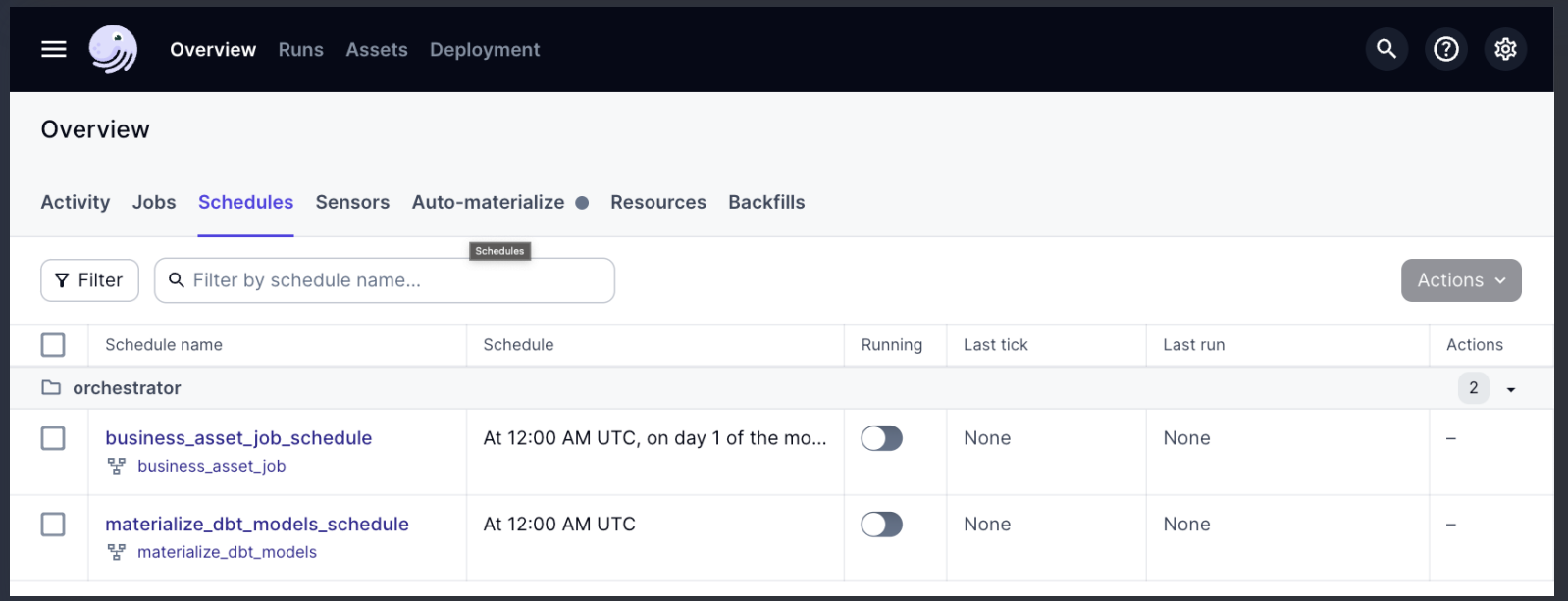


Run History


Assets
View global asset lineage
Reload definitions
Filter
Filter asset keys...
0:02
Materialize selected
Asset name
Code location / Asset group
Status
all_scope_summary
This asset ingest the all_scope summary data from...
orchestrator
raw
Materialized
Feb 1, 11:56 AM
annual_cpi_index
Get annual CPI index from python cpi library
orchestrator
raw
Materialized
Jan 27, 7:30 AM
commuting_emission_factors_EPA
This asset ingest the Commuting Emission Factors ...
orchestrator
raw
Materialized
Apr 16, 8:34 PM
commuting_survey_2018
This asset ingests the Commuting Survey 2018 dat...
orchestrator
raw
Materialized
Apr 16, 8:36 PM
commuting_survey_2021
This asset ingests the Commuting Survey 2021 dat...
orchestrator
raw
Materialized
Apr 16, 8:34 PM
commuting_survey_2023
This asset ingests the Commuting Survey 2023 dat...
orchestrator
raw
Materialized
Apr 16, 8:35 PM
commuting_survey_modes
This asset ingests the mode breakdown across yea...
orchestrator
raw
Materialized
May 3, 11:02 AM
construction_expense
This asset ingest the construction expense data fr...
orchestrator
raw
Materialized
Feb 21, 12:04 PM

Processing Stages and Scheduling





dbt (Data Build Tool)

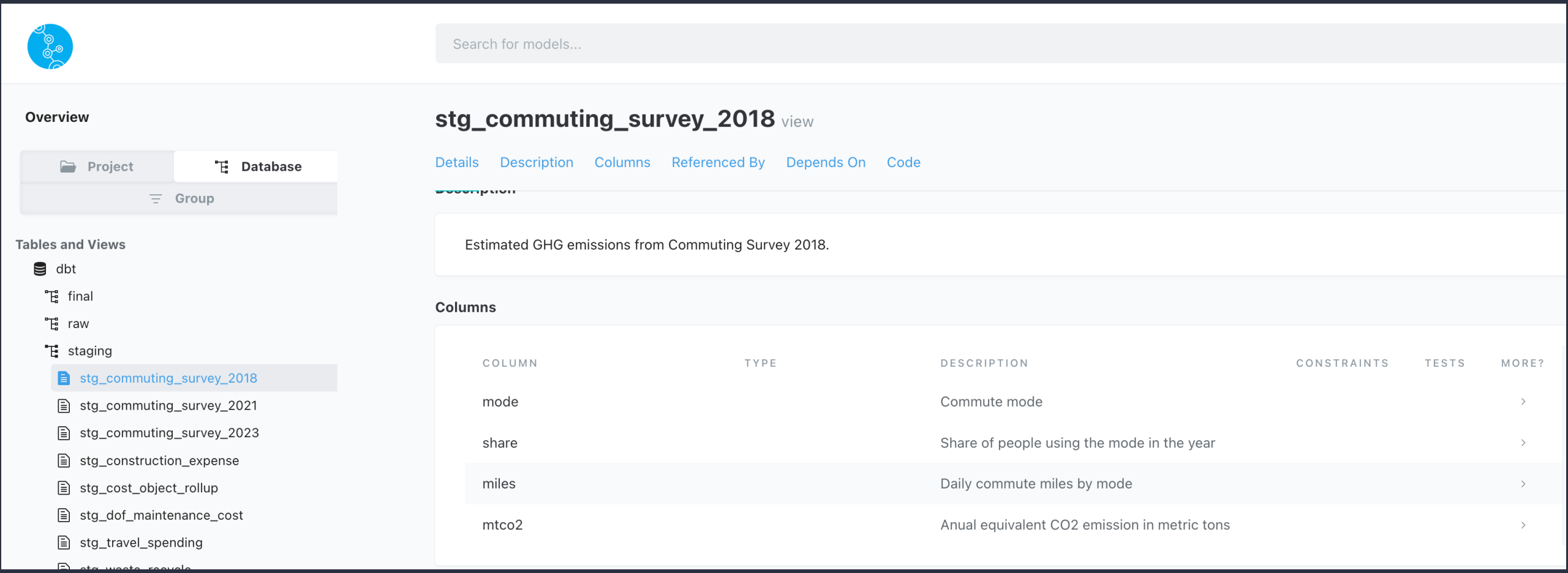
- Open-source tool for SQL-based data modeling
- Versioned controlled SQL queries, reusable macros.
- Support various SQL flavors, i.e. BigQuery, Redshift, Oracle, and Postgres
- Automatically generation of documentation and data lineage.
- Easy integration with orchestrator like Dagster and Airflow

SQL-based Data Modeling

```sql
-- set static variables using jinja2 syntax
{% set reply_rate = 0.33 %}
{% set work_week = 50 %}
{% set remote_rate = 0.21 %}

WITH distance AS (
    SELECT
        "drove alone" * {{ var('car_speed') }} * commute_time_average_hours AS drove_alone,
        "carpooled(2-6)"
        * {{ var('car_speed') }}
        * {{ var('car_share_ratio') }}
        * commute_time_average_hours AS carpooled,
        "vanpooled(7+)"
        * {{ var('car_speed') }}
        * {{ var('van_share_ratio') }}
        * commute_time_average_hours AS vanpooled,
        shuttle * {{ var('bus_speed') }} * commute_time_average_hours AS shuttle,
        "public transportation"
        * {{ var('t_ratio') }}
        * {{ var('t_speed') }}
        * commute_time_average_hours AS subway,
```

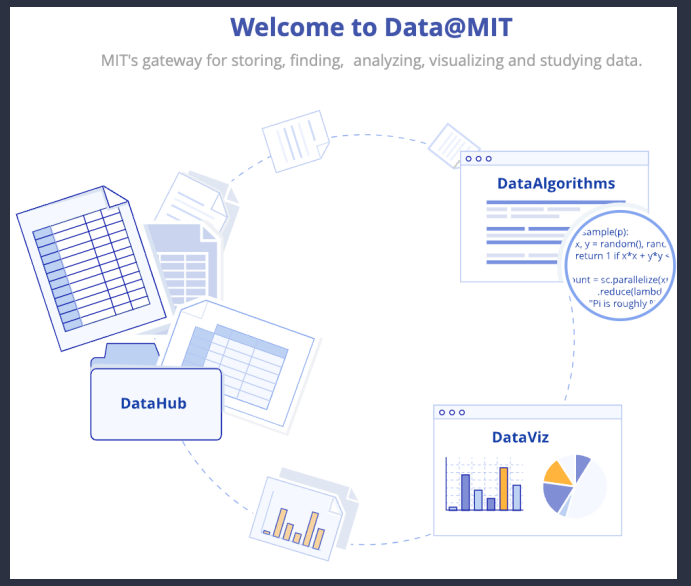
Documentation



Data Warehouse

- Use a RDS managed Postgres instance
- Processing stages managed by “schema”
- Isolated local development and production environments
- All dashboards serve from the Postgres warehouse.

IS&T DataHub



- MIT's "data lake" powered by AWS
- Web UI to interact with projects and files
- Project-based data access management
- API support to handle files programmatically.

CI/CD

- Use Github Actions to run tests, build image and documentation, push to ECR, and deploy.
- Use Docker to ensure portability
- Infrastructure as code: With the AWS Cloud Formation template to setup cloud resources in minutes.

Demo

- Dagster
- Dbt documentation
- Github Repo

Looking Ahead

Benefits

- Version controlled pipelines
- Automated build/deploy process
- Transparent data lineage and run history
- Enhanced data quality control
- Accelerated development cycle
- Integrated with MIT Data Hub
- Single source of truth

Improvement Ideas

- Enhance data validation and quality checks: explore tools like Great Expectations and Pandera
- Add more unit tests
- Explore scalable structures, i.e. Redshift and ECS.
- Improve on Infrastructure as code
- Implement more pipelines

Looking for Collaborators

- Github Repo
- This project is open-source. As the sole developer in the office, I need help to improve the code base.
- The data-warehouse/data-platform approach is a great first step for small teams striving for a data-driven culture.
- Join me ☺