10Alytics

# DATA ORCHESTRATION & VERSION CONTROL

Chu Ngwoke

# DIVY TRIPS

**Divy Trips** is a fictitious cab hailing company in New York City. Divy has been collecting individual trip data from 2009 to 2016. The data is large and is stored in a **clickhouse database** (https://github.demo.trial.altinity.cloud:8443/play). The business seeks to generate monthly aggregated data and load in a DWH to facilitate business analytics and informed decision-making.

# DIVY TRIPS

You have been hired as a data engineer, and your first task is to orchestrate a pipeline to achieve this objective. You are to use **modular coding** and appropriate **version control** for the solution you develop, to facilitate code maintenance, and collaboration with your future team members.

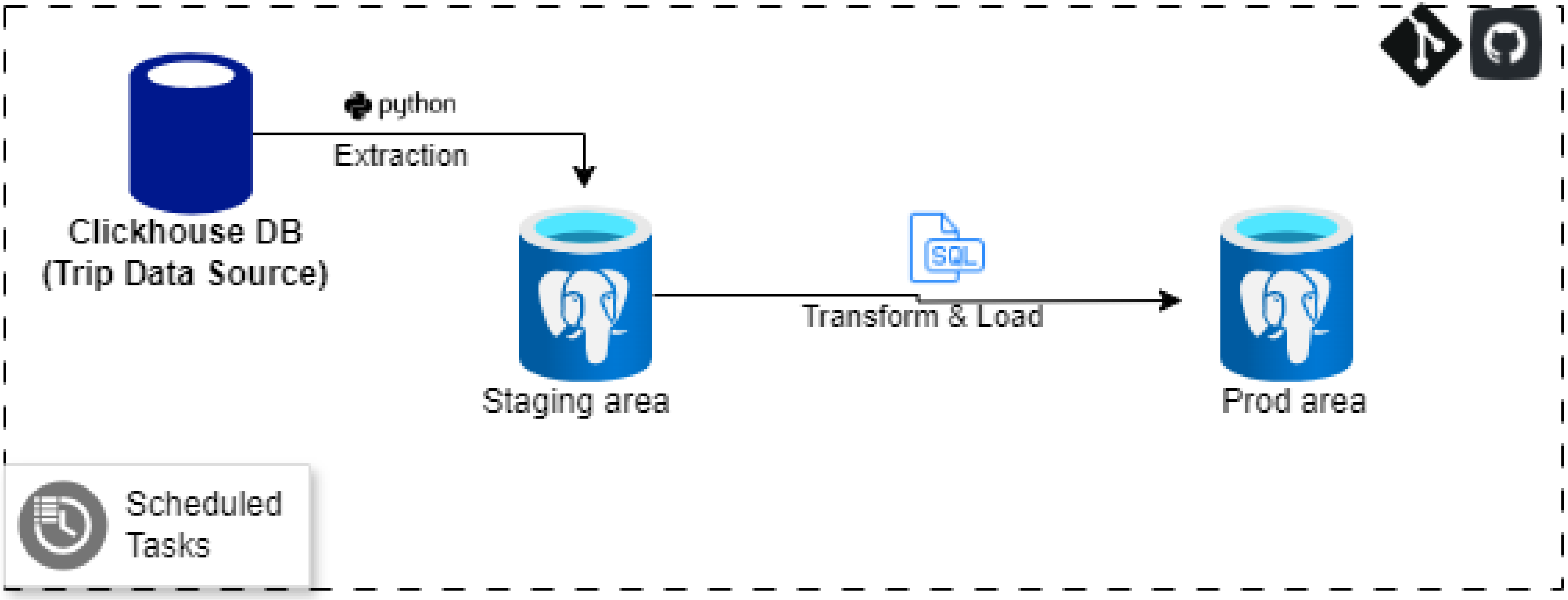
# DIVY TRIPS - TASKS

- Explore the trip data on Clickhouse
- Extract the trip data and load to a staging area on an on-prem database.
- Create a production environment and develop procedures to load aggregate tables to show the following monthly metrics (average trip count, average trip duration, average trip fare)
- Push your code to a new Github repository

# DIVY TRIPS - TASKS

- Pull the codebase from the github repository.
- Create a branch and refactor the code to load data incrementally and implement an orchestration to load the data daily.
- Commit your changes to your branch and create a pull request to merge your changes to the main branch (codebase).
- Orchestrate the incremental load pipeline using Windows task Scheduler.

# SOLUTION ARCHITECTURE



# THANK YOU

Chu Ngwoke

-

Chu.Ngwoke@gmail.com