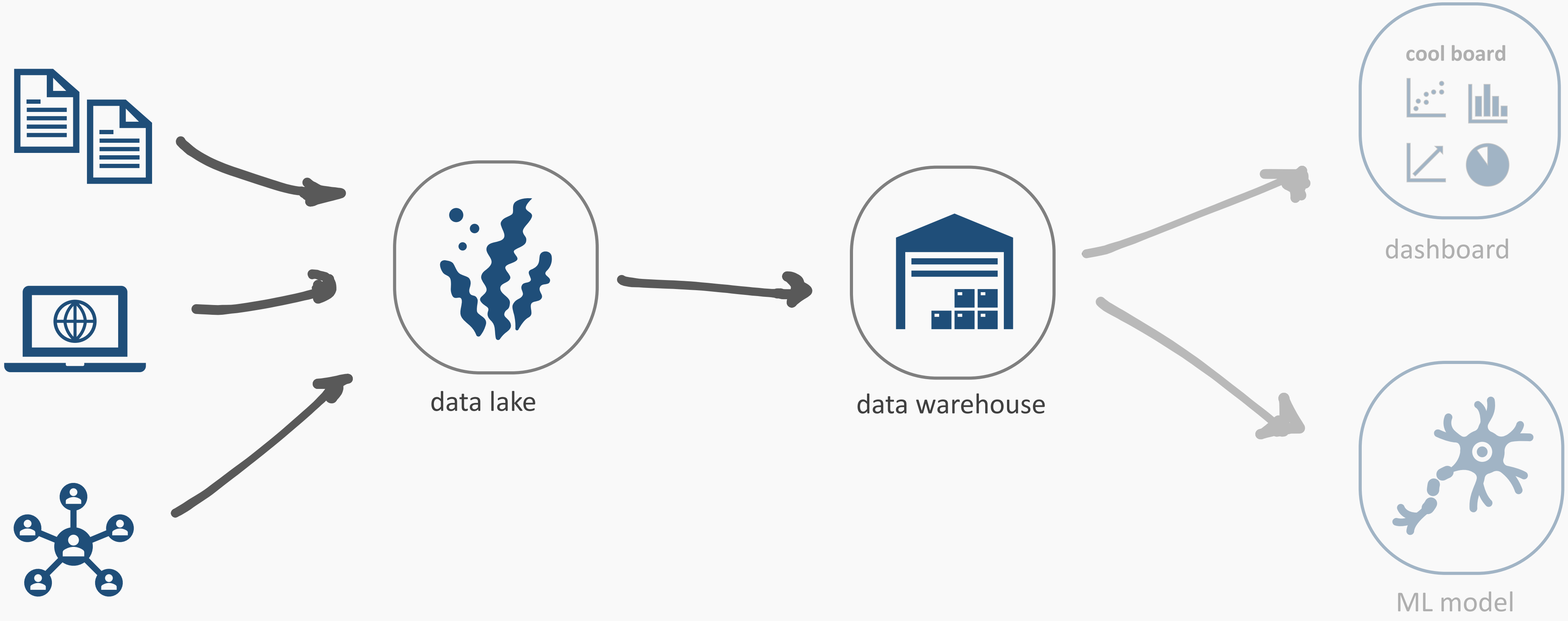
kokchun giang

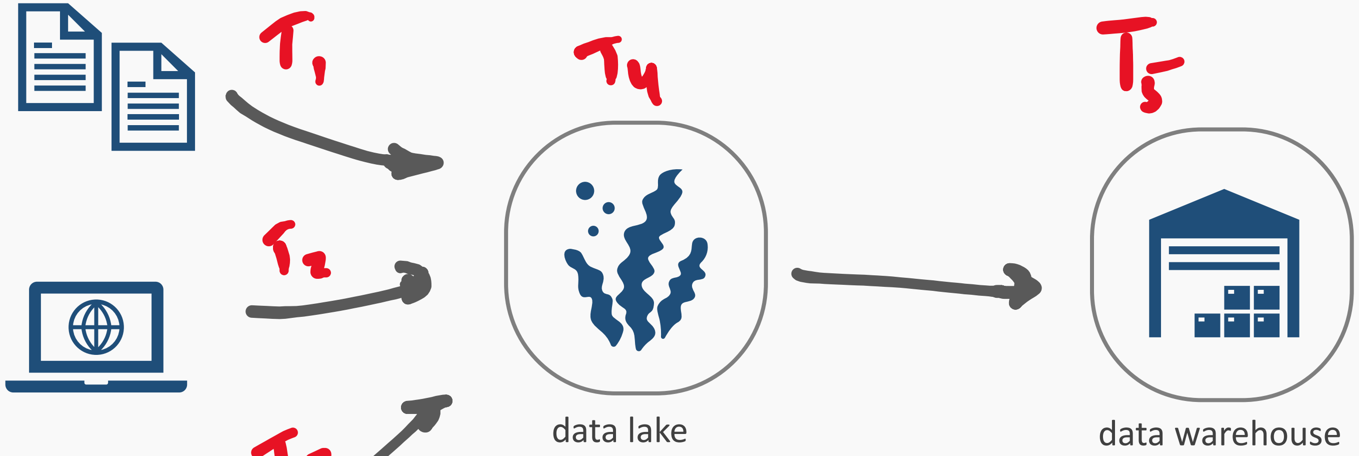
put your tasks in directed acyclic graphs **(DAGs)** in **airflow** to create data pipelines

let’s take a look at a **data pipeline**



different data
sources

let’s take a look at a **data pipeline**



extract data

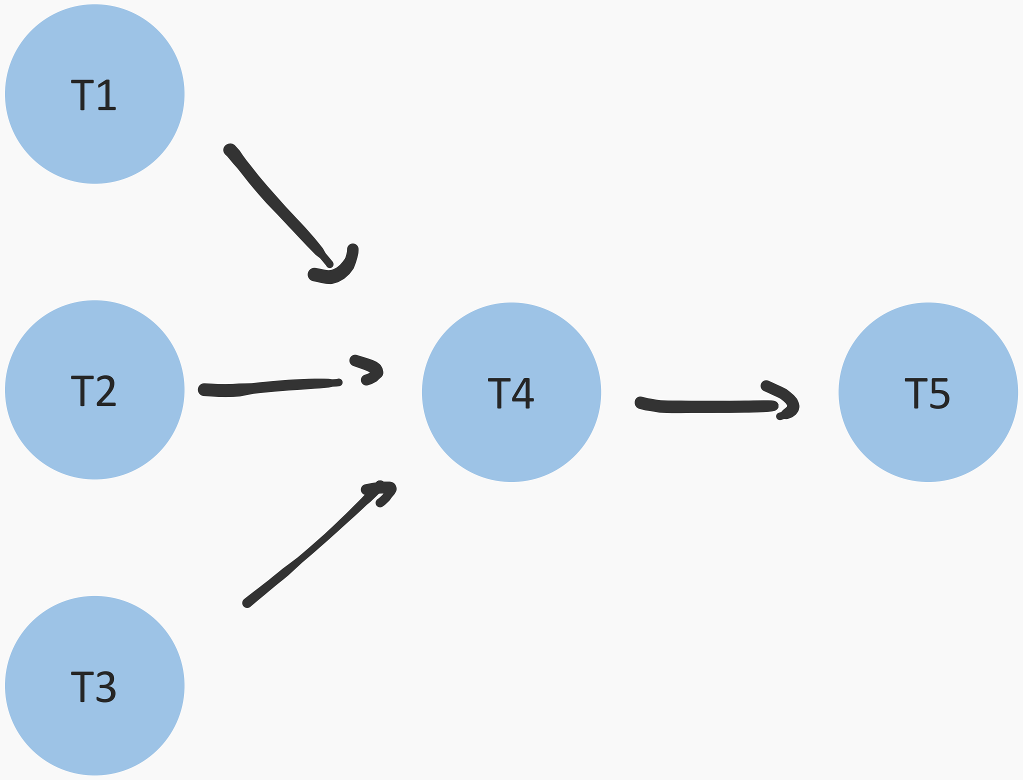
transform

unstructured raw data

structured data ready for ML & dashboard

- how to solve this task?

use **cron** scheduling to solve it?

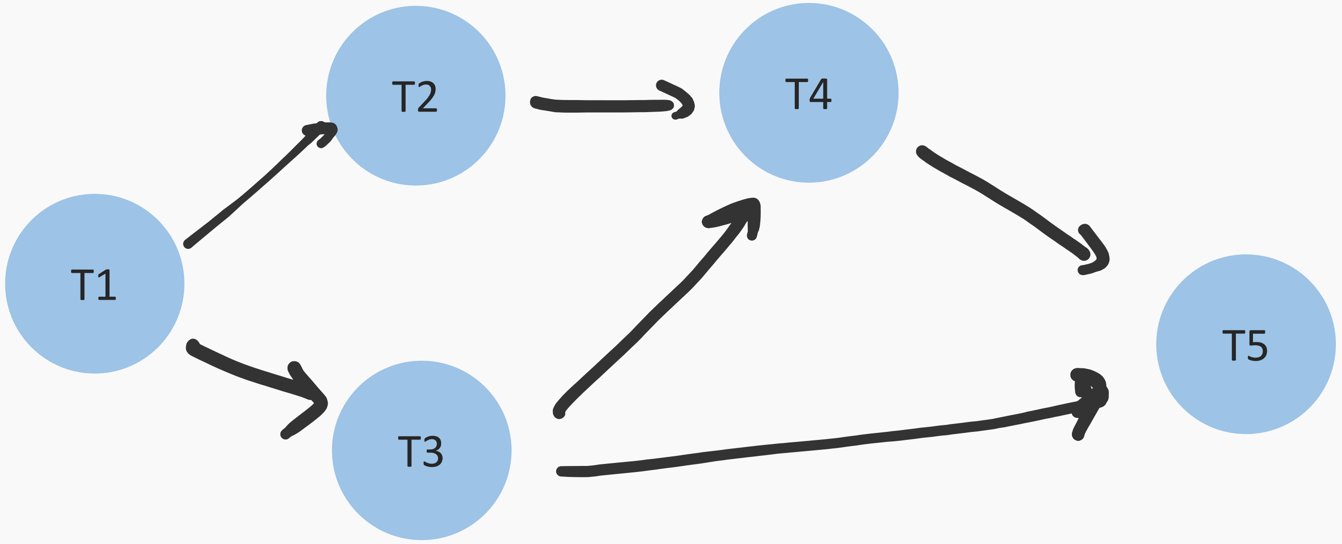


- naive solution:

  5 scripts T1, — T5

- use cron to schedule all scripts with 1h interval
- done?

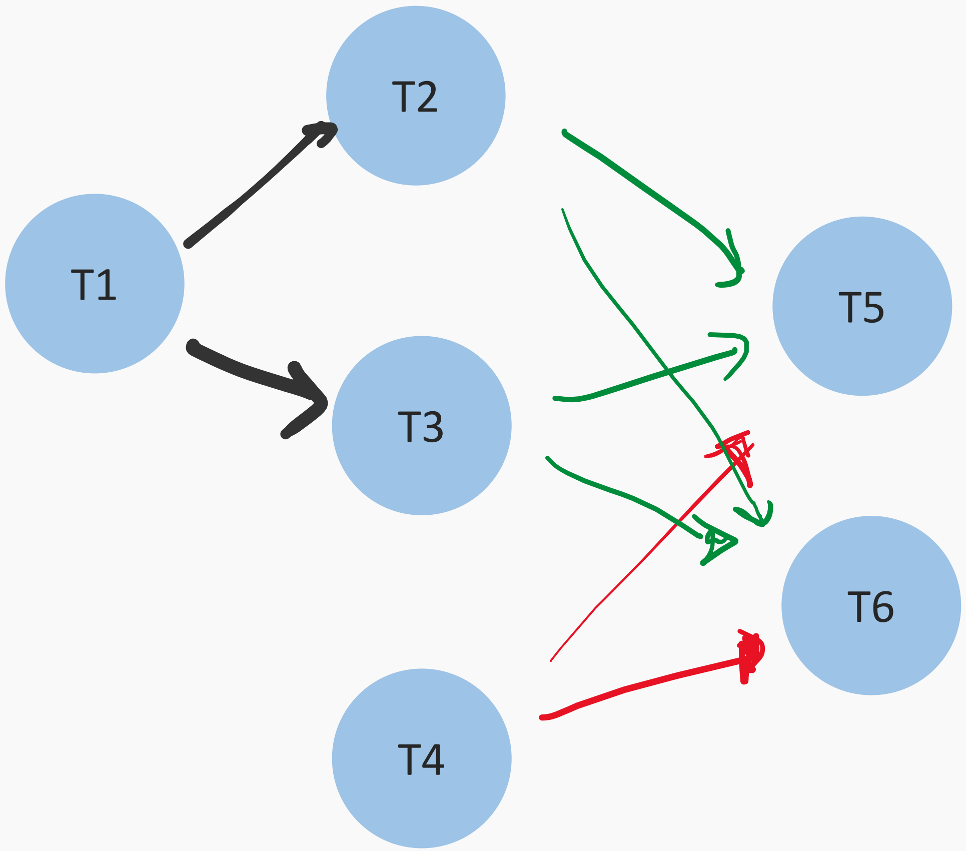
how about when we have **dependencies?**



- so T_4 waits for T_2 & T_3
- T_5 waits for T_3 & T_4
- can we schedule this with cron?

→ maybe but becomes complex!

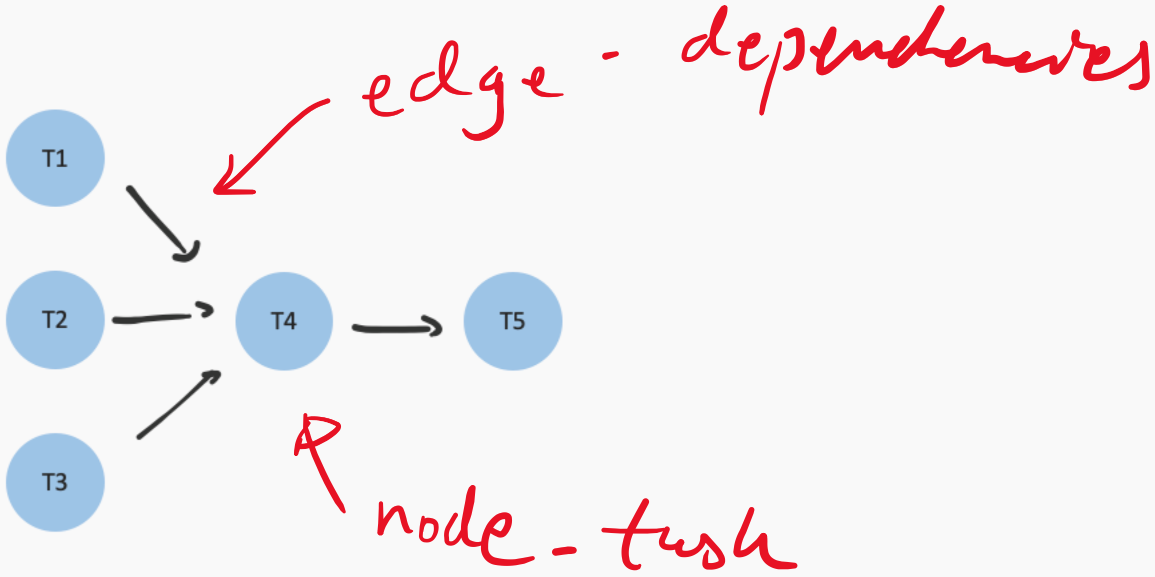
some more examples when **cron** is not enough

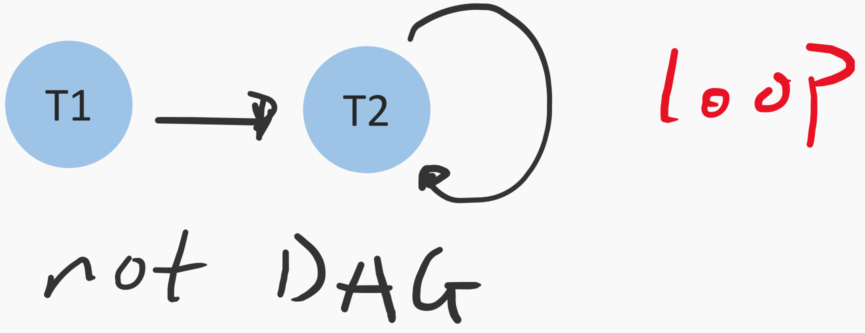


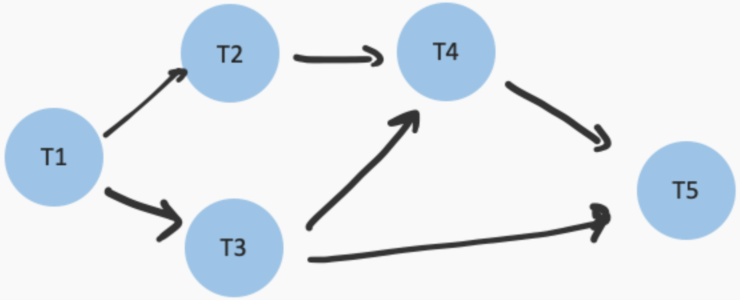
T5 should run when at least two of previous was successful

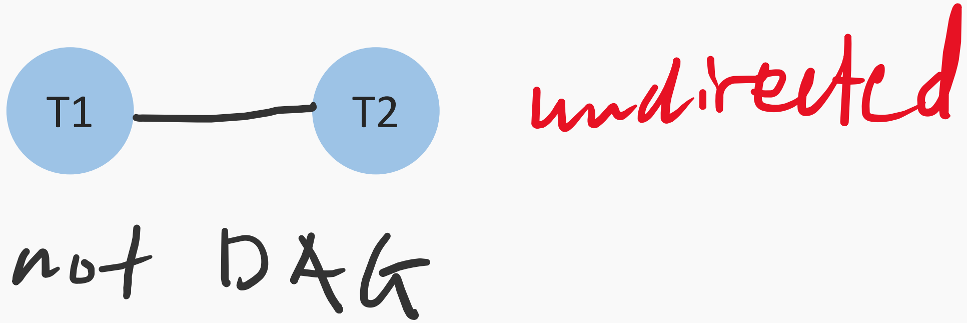
T5 should run when at least one of upstream succeeds, e.g. T6 is an email/discord/slack notification when some upstream tasks fails

what are **DAGs**









what are **DAGs**

Directed $\forall$ tasks ≥ 1 upstream $\vee$ 1 downstre

Acyclic tasks no dependency to itself, $\exists$ no loops

Graphs relationship betw. tasks by nodes & vertices

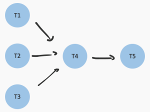
DAGs are data pipelines with collection of tasks and relationsships between tasks

orchestrate workflows with **airflow**

coordination and management of multiple tasks

define a task
(unit of work)

tasks organized
in DAGs



schedule the
execution of tasks

upstream/downstream
dependency

monitor progress in
Airflow UI

integrate with
other tools