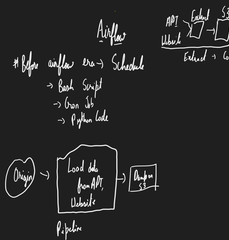
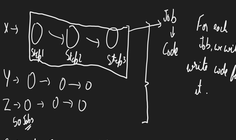
## Airflow




* Most airflow are → Schedule
  → Bash Script
  → Cron Job
  → Python code



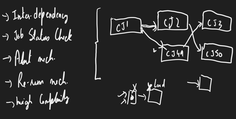

① Getting data from API, website (Extract)
② Dumping into AWS S3 (Load)
→ Bash, Cron Job, Python Code

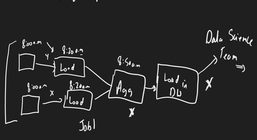



Scale up 50 cron jobs



→ Data dependency
→ Job Status Check
→ Alert mech
→ Re-run mech
→ High Complexity



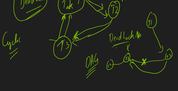

DAG

→ Monitoring: Job Status, error, the previous jobs
→ Manage Failure: alerts on Failure, retries, timeouts
→ Dependencies: Check upstream data, run jobs only after Job1
→ Backfill: rerun the historic jobs
→ Scalability: Centralized System to manage the jobs
→ Deployment: version Control orchestration System
Logs all the previous jobs


# Technical Terms of Airflow:
→ Task: It's the smallest unit of Job



# Core Components of Airflow:


Scheduler → Triggers & schedules workflows and submits task to the executor

Executor → Runs the tasks/jobs provided by scheduler

Workers → Runs the tasks (machines)

Web-server → User Interface to inspect, trigger, debug the DAGs/tasks

Metadata Store → Stores info about DAGs & tasks state (queue, in progress, completed, failed)

DAGs folder → This is directory where all the DAG's code is provided, read by scheduler and ran by the executor.