Summary

Thank you for joining us in this module on DAG scheduling.

1. We learnt about the states of a DAGRun which initially is Queued, and changes to Running as soon as the first task runs. Once all tasks are completed, the final state of the DAGRun is determined by the state of the leaf tasks, which could be Success or Failure.
2. The start date and scheduling interval are two important parameters to consider when scheduling a DAG. DAGRuns are created based on these parameters and have a data interval start and end corresponding to the time frame for which the DAG was executed.
3. The start date parameter defines the timestamp from which the scheduler will attempt to backfill, and the scheduling interval determines how often a DAG runs.
4. The catchup mechanism in Airflow allows running all non-triggered DAGRuns between the start date and the last time the DAG was triggered. The backfilling mechanism allows running historical DAGRuns or rerun already existing DAGRuns.
5. We also explored the Airflow CLI commands to backfill our DAG for any data intervals, regardless of the start date of our data pipeline.
6. Users can define the scheduling interval using CRON expressions or timedelta object, and can use a timetable to schedule a DAG more precisely.
7. It is important to set the start date parameter properly to avoid confusion with Airflow, and use a static date.

Thank you for tuning in. See you in the next one.

Was this page helpful?

😠 ☹️ 😐 🙂 😍

Want feedback like this? Try Hotjar