3장_1 Apache Airflow 실행시간 이해하기

#airflow #schedule #dataengineer

Summary

- 일배치면 하루 전 기준으로 돌고, 시간배치면 시간 전 기준으로 도는 컨셉
- Airflow는 UTC 시간 기준으로 수행

- 예시)
  - 하루에 한번 0 15 * * * 로 도는 배치 -> 2023-01-04 15:00에 2023-01-03 15:00 기준으로 배치가 수행

1. start_date

- DAG 가 시작되는 기준 시점. (start_date에 시작되는 것이 아님.)
- start_date가 2023-01-04이면 DAG는 2023-01-04 00:00 기준으로 시작되는 것으로 스케쥴링 된다. 그리고 매 10분 기준마다 돌 것이다. 2023-01-04 00:10, 2023-01-04 00:20, ...

Warning

현재 시간이 start_date 이전이면 수행되지 않음.

2023-01-04 start_date로 일배치를 돌리면 첫 수행 날짜는 2023-01-05가 된다.

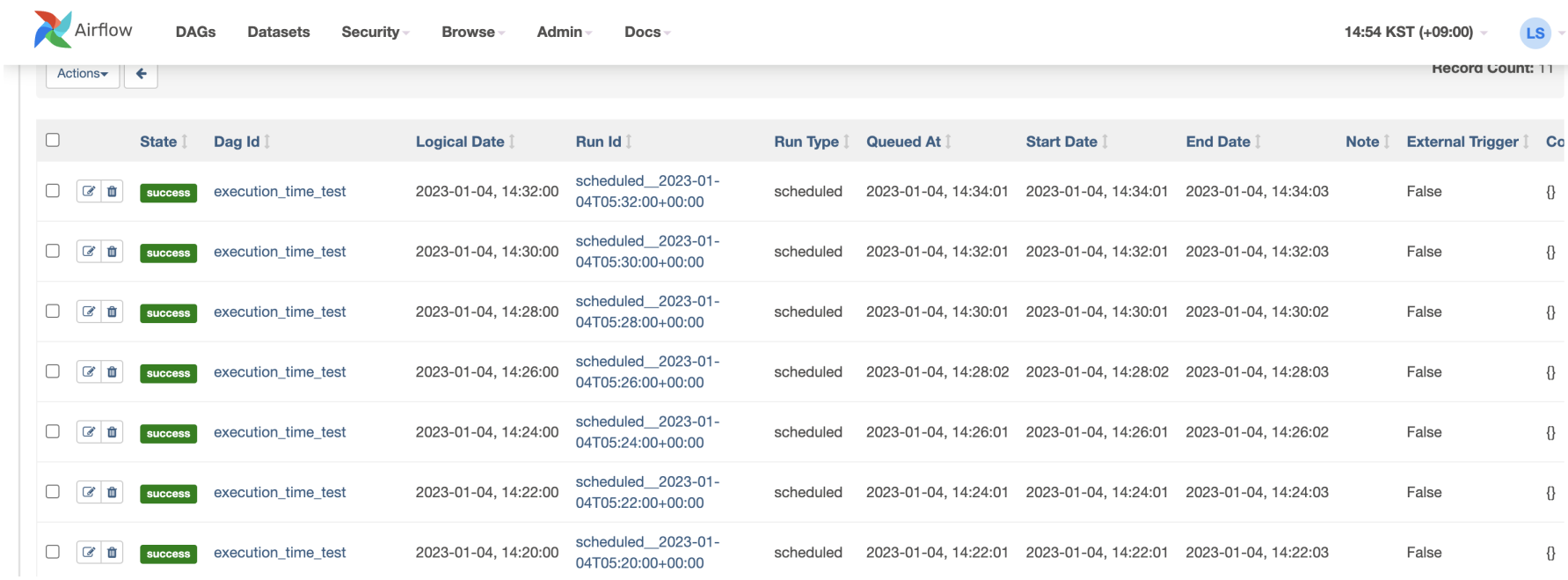
2.execution_date

- **실제 실행날짜가 아님. Logical Date**를 의미
- 일종의 주문번호. DAG 실행될 때마다 바뀜
- Airflow DAG 실행시 세팅할 수 있는 유일한 parameter이다.
- context variable을(ds,yesterday_ds, tomorrow_ds 등) execution_date로 사용할 수 있다.
- run, test, backfill, trigger 등 CLI 환경에서는 execution_date를 반드시 명시해야 한다.
- 스케쥴러로 DAG를 실행할 때는 execution_date가 자동으로 들어간다.

예시 코드

```
import datetime as dt
from airflow import DAG
from airflow.operators.python import PythonOperator

dag = DAG(
        dag_id="execution_time_test",
        schedule_interval="*/2 * * * *",
        start_date=dt.datetime(2023, 1, 4, 14, 10),
)

def _print_hello():
        print("Hello World")


print_hello = PythonOperator(
task_id="print_hello", python_callable=_print_hello, dag=dag

)



print_hello
```

- 2023-01-04일 14:10 분 기준 start_date
- interval은 2분마다 배치 수행 - 첫 배치 수행 시간 : 20230104 14: 12 분 첫 배치 수행



State	Dag Id	Logical Date	Run Id	Run Type	Queued At	Start Date	End Date	Note	External Trigger	Co
success	execution_time_test	2023-01-04, 14:32:00	scheduled__2023-01-04T05:32:00+00:00	scheduled	2023-01-04, 14:34:01	2023-01-04, 14:34:01	2023-01-04, 14:34:03		False	{}
success	execution_time_test	2023-01-04, 14:30:00	scheduled__2023-01-04T05:30:00+00:00	scheduled	2023-01-04, 14:32:01	2023-01-04, 14:32:01	2023-01-04, 14:32:03		False	{}
success	execution_time_test	2023-01-04, 14:28:00	scheduled__2023-01-04T05:28:00+00:00	scheduled	2023-01-04, 14:30:01	2023-01-04, 14:30:01	2023-01-04, 14:30:02		False	{}
success	execution_time_test	2023-01-04, 14:26:00	scheduled__2023-01-04T05:26:00+00:00	scheduled	2023-01-04, 14:28:02	2023-01-04, 14:28:02	2023-01-04, 14:28:03		False	{}
success	execution_time_test	2023-01-04, 14:24:00	scheduled__2023-01-04T05:24:00+00:00	scheduled	2023-01-04, 14:26:01	2023-01-04, 14:26:01	2023-01-04, 14:26:02		False	{}
success	execution_time_test	2023-01-04, 14:22:00	scheduled__2023-01-04T05:22:00+00:00	scheduled	2023-01-04, 14:24:01	2023-01-04, 14:24:01	2023-01-04, 14:24:03		False	{}
success	execution_time_test	2023-01-04, 14:20:00	scheduled__2023-01-04T05:20:00+00:00	scheduled	2023-01-04, 14:22:01	2023-01-04, 14:22:01	2023-01-04, 14:22:03		False	{}

- 위 예시에 따르면,
  - 작업 시작 기준 일자 : 2023-01-04 14:20
  - `Start Date` : 작업이 실제 수행된 시간
  - `Logical Date` : Execution Date, 실제 수행되는 시간

주의점

- Airflow 는 UTC 기준으로 동작하므로 DAG 마다 KST 시간 설정 필요

```
import pendulum

# Korea Time Zone
kr_tz = pendulum.timezone("Asia/Seoul")

dag = DAG(
        dag_id="execution_time_test",
        schedule_interval="*/2 * * * *",
        start_date=dt.datetime(2023, 1, 4, 14, 20, tzinfo=kr_tz),
)
```