Cloud Composer/Airflow

What is Airflow and Cloud Composer

Airflow is open source project which is used for orchestrating and scheduling the workloads.

Cloud Composer is fully managed workflow orchestration service offered by Google that is built on Airflow. It deploys multiple components on GCP to run Airflow and relies on certain configurations to successfully execute workflows. We need to create environment first to setup cloud composer in GCP.

Core Components in Composer

- **Metadata Database**: Airflow uses a SQL database to store metadata about the data pipelines being run. In case of Cloud Composer CloudSQL is used.
- **Web Server** and **Scheduler**: The Airflow web server and Scheduler are separate processes running on the airflow machine and interact with the metadata database. In case of Cloud Composer both these processes are deployed on App Engine.
- **Executor** This process runs with the scheduler.
- **Worker(s)** are separate processes which also interact with the other components of the Airflow architecture and the metadata repository. In case of Cloud Composer all the workers are deployed in Google Kubernetes Engine in containers which autoscales based on throughput.
- **airflow.cfg** is the Airflow configuration file which is accessed by the Web Server, Scheduler, and Workers.
- **DAGs** refers to the DAG files containing Python code, representing the data pipelines to be run by Airflow.The location of these files is specified in the Airflow configuration file, but they need to be accessible by the Web Server, Scheduler, and Workers. In case of Composer the dags are stored in Google Cloud Storage.
- **Logs** are stored in Stackdriver incase of Cloud Composer.