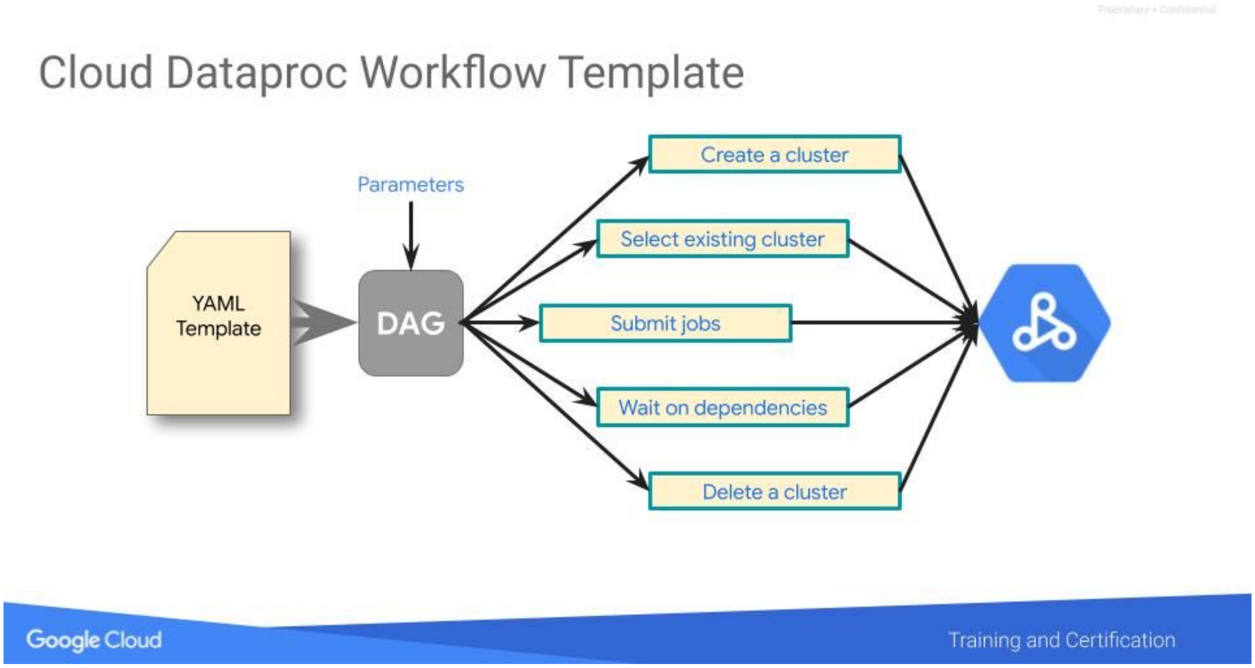
Cloud Dataproc Workflow Automation



Proprietary + Confidential

Cloud Dataproc Workflow Template

Google Cloud

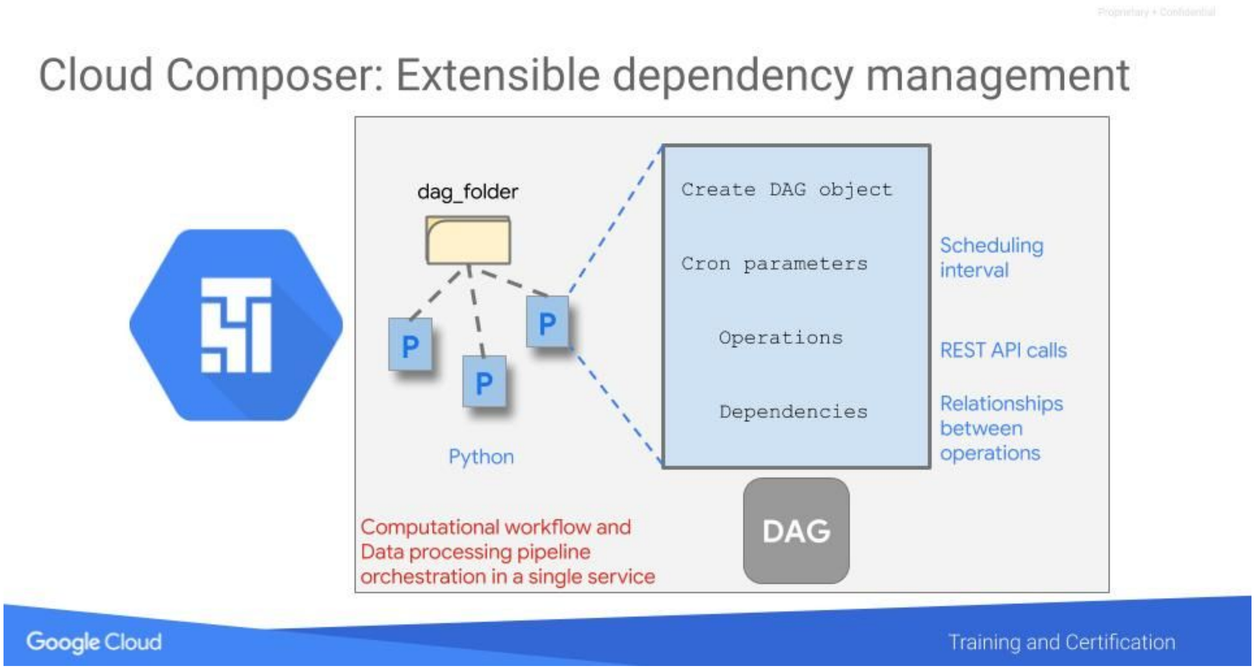
Training and Certification

The Cloud Dataproc Workflow Template is a YAML file that is processed through a Directed Acyclic Graph (DAG). It can create a new cluster, select from an existing cluster, submit jobs, hold jobs for submission until dependencies can complete, and it can delete a cluster when the job is done.

It is currently available through the gcloud command and the REST API, but not through Console.

The Workflow Template becomes active when it is instantiated into the DAG. The Template can be submitted multiple times with different parameter values. You can also write a template inline in the gcloud command, and you can list workflows and workflow metadata to help diagnose issues.

The Cloud Dataproc Workflow Template is a great tool if you have Dataproc jobs to run -- that is if the workflow is confined to Cloud Dataproc. But what if you need to coordinate and automate larger and more complex workflows that involve running jobs on other services and channeling data between them?



Cloud Composer: Extensible dependency management

Computational workflow and Data processing pipeline orchestration in a single service

Google Cloud

Training and Certification

Cloud Composer is a workflow orchestration service based on Apache Airflow.

Cloud Composer can be used to automate Cloud Dataproc jobs and to control clusters.

The DAG in Cloud Composer is defined in a Python file. There can be multiple DAGs. They are contained in a DAG folder. The DAG file is concerned with providing the instructions necessary to complete each step in a workflow.

Each Python file creates a DAG object. Next, it provides parameters that define the scheduling interval. After that are the operations to be performed. And finally, dependencies that determine which operations must complete before others can begin.

The operators are extensible. Many services offer APIs that can be integrated into the Python operators in the DAG. This makes Cloud Composer very versatile. It can be used to orchestrate a data processing pipeline and to orchestrate a computational workflow. And it can connect these together to create rich dynamic solutions.