Identity and Access Management (IAM) and Quotas

Omar Ismail

Solutions Developer, Google Cloud

Agenda

Course Intro

Beam and Dataflow Refresher

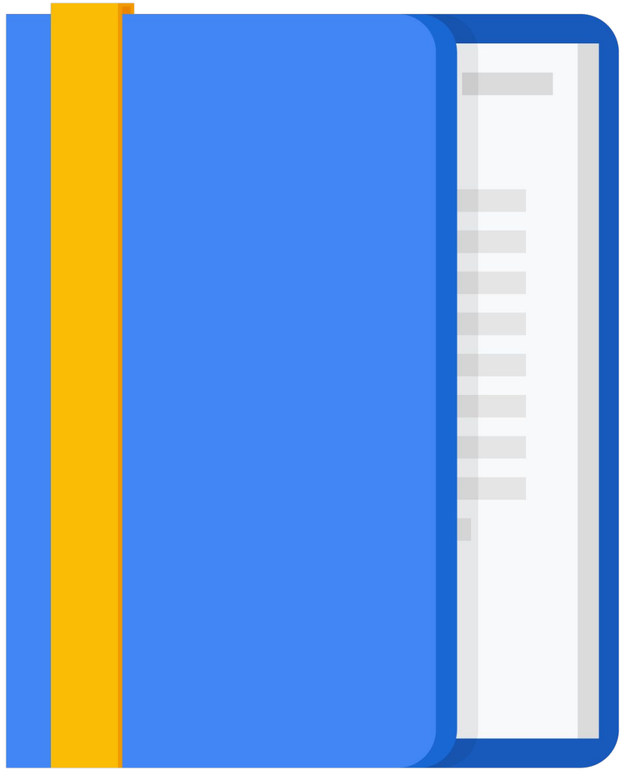
Beam Portability

Separating Compute and Storage

IAM, Quotas, and Permissions

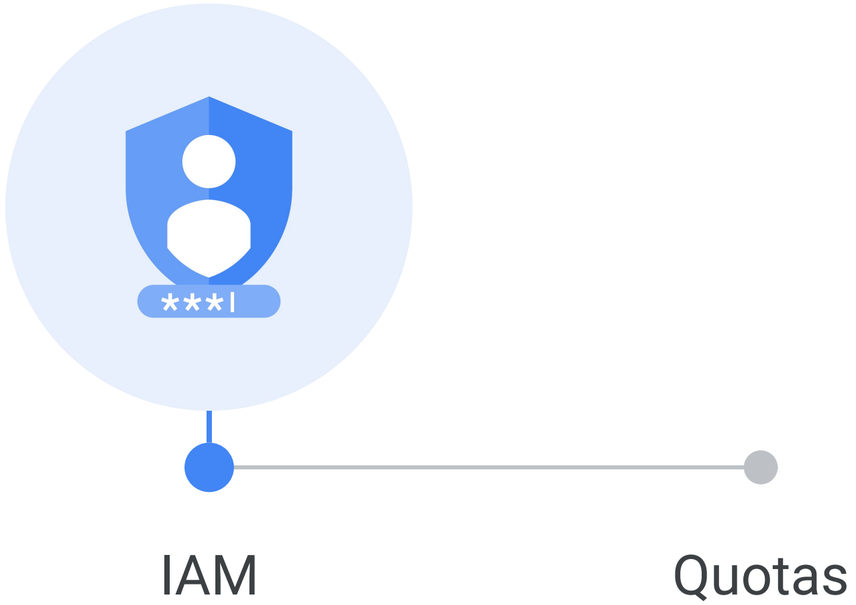
Security

Summary



IAM, quotas, and permissions

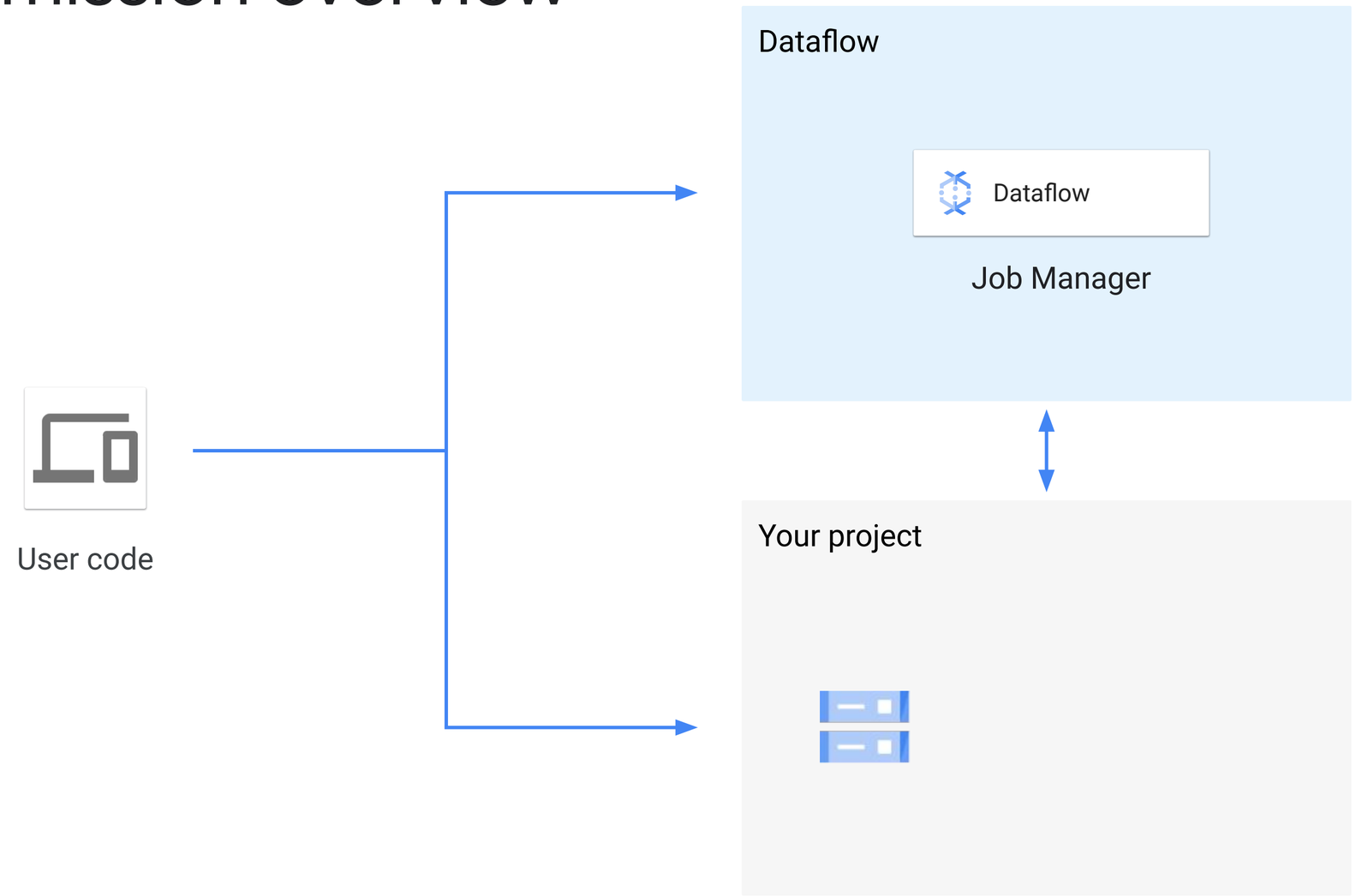
Agenda



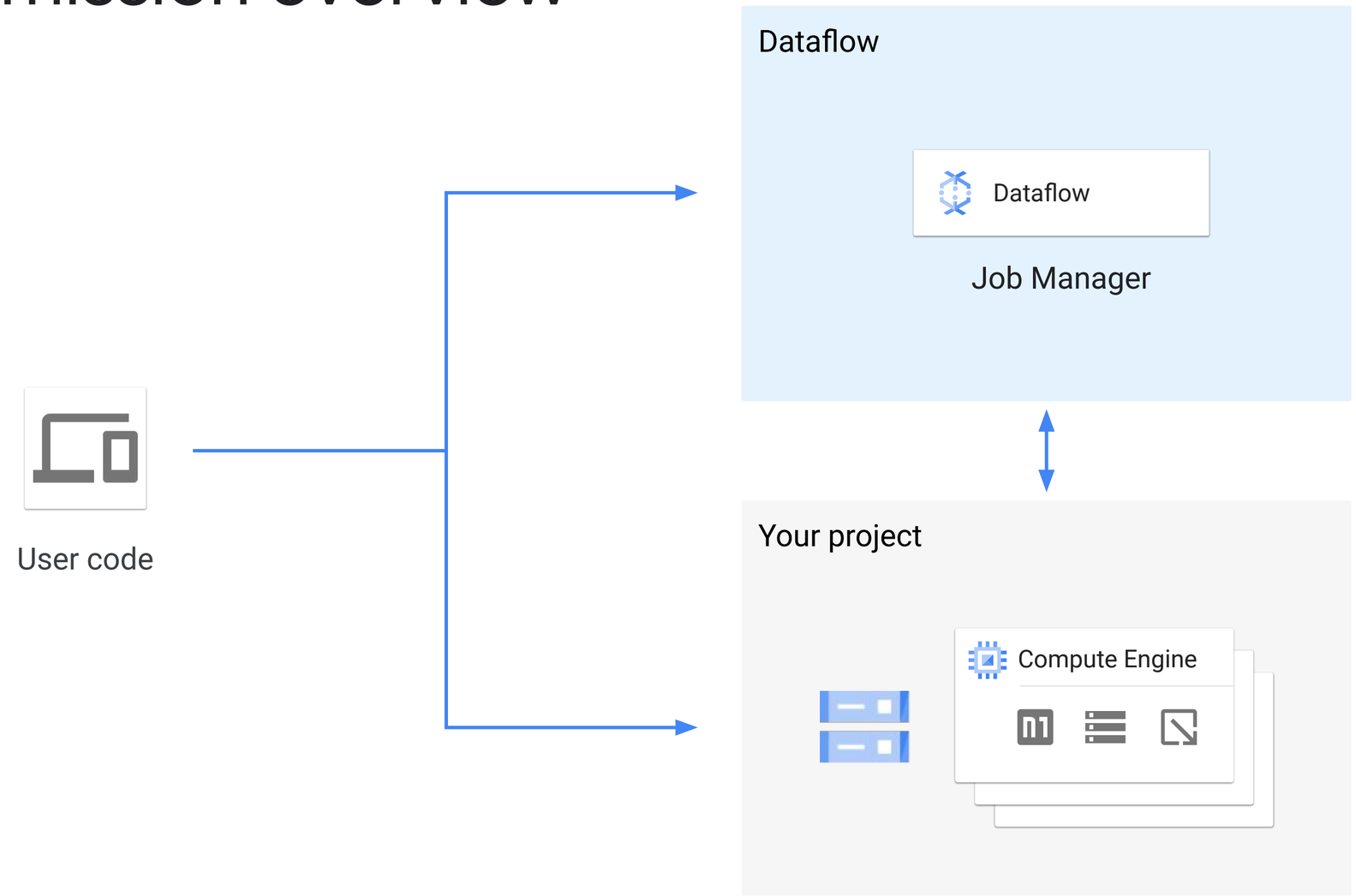
Job submission overview

User code

Job submission overview



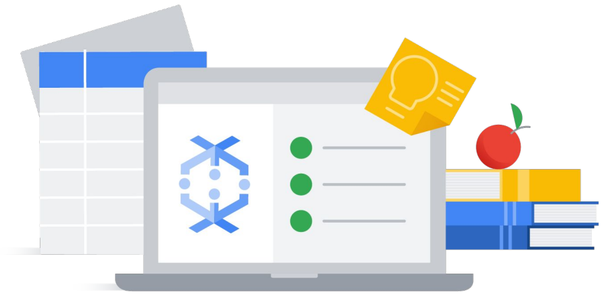
Job submission overview



Three credentials

1. User roles
2. Dataflow service account
3. Controller service account

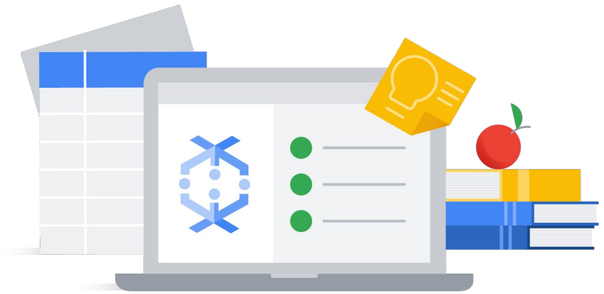
User roles



Dataflow Viewer

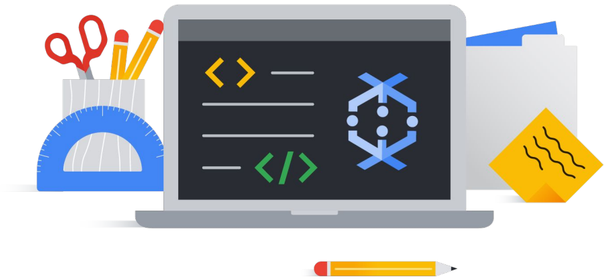
Provides read-only access to all Dataflow-related resources.

User roles



Dataflow Viewer

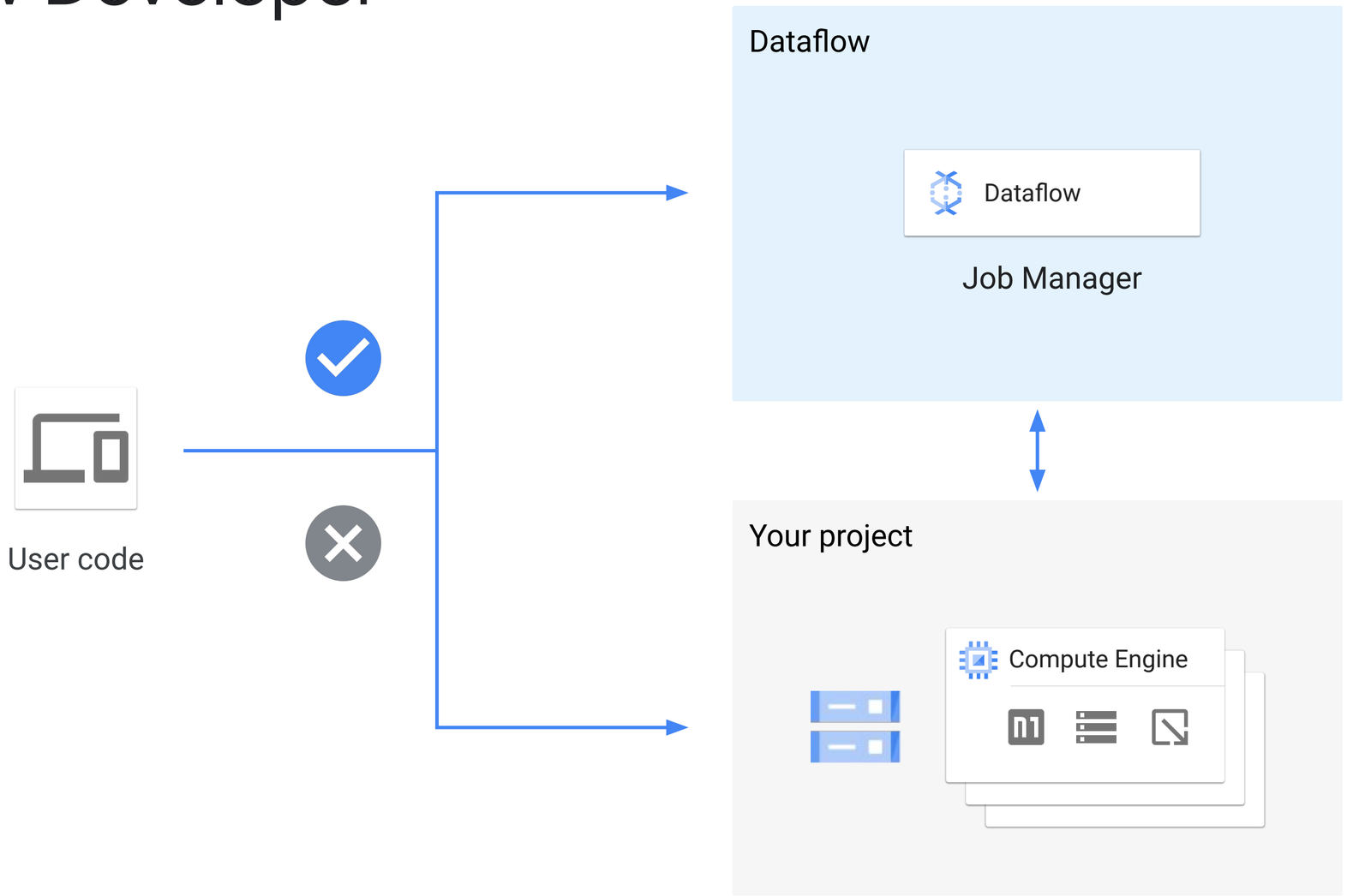
Provides read-only access to all Dataflow-related resources.



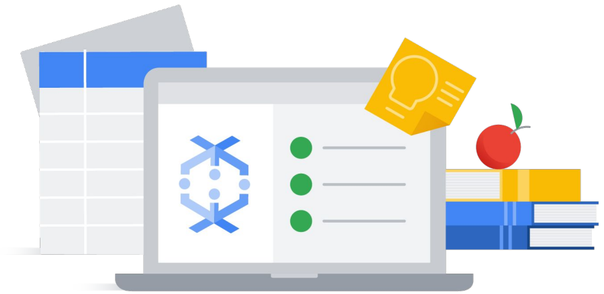
Dataflow Developer

Provides access to view, update, and cancel Dataflow jobs.

Dataflow Developer

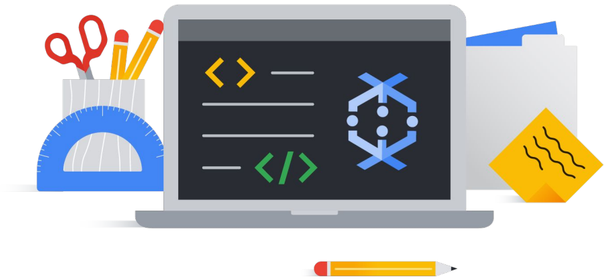


User roles



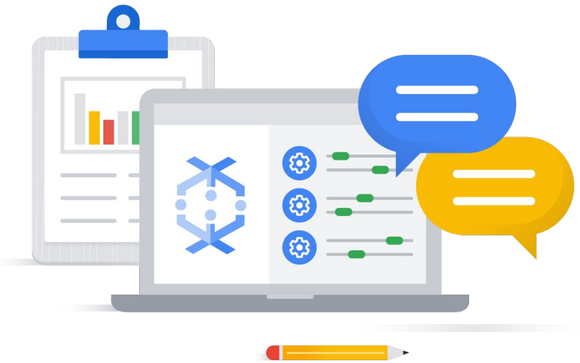
Dataflow Viewer

Provides read-only access to all Dataflow-related resources.



Dataflow Developer

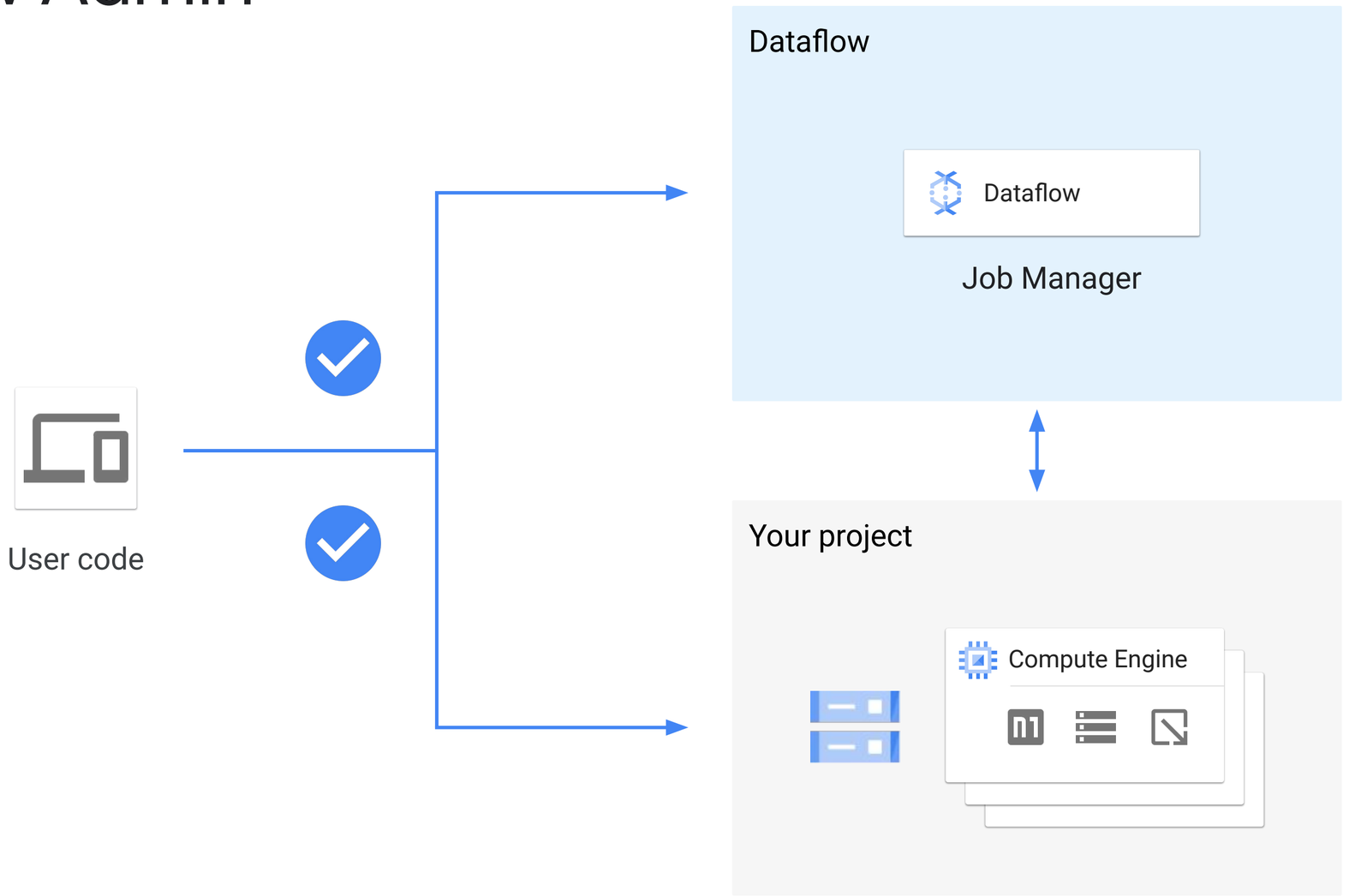
Provides access to view, update, and cancel Dataflow jobs.



Dataflow Admin

Provides access for creating and managing Dataflow jobs.

Dataflow Admin



Question

A Dataflow job (2021-01-31_14_30_00-9098096469011826084) must be canceled. You are assigned the Dataflow Developer role. Can you run the following command:

```
$ gcloud dataflow jobs cancel
2021-01-31_14_30_00-9098096469011826084  --region=$REGION
```

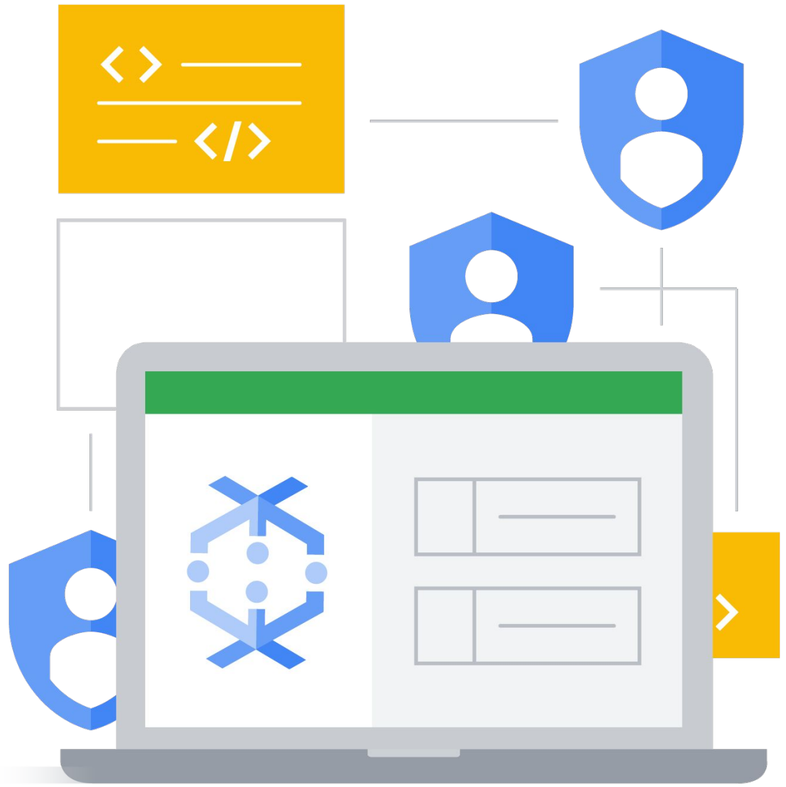
Three credentials

1. User roles
2. Dataflow service account
3. Controller service account

Dataflow service account

The Orchestrator

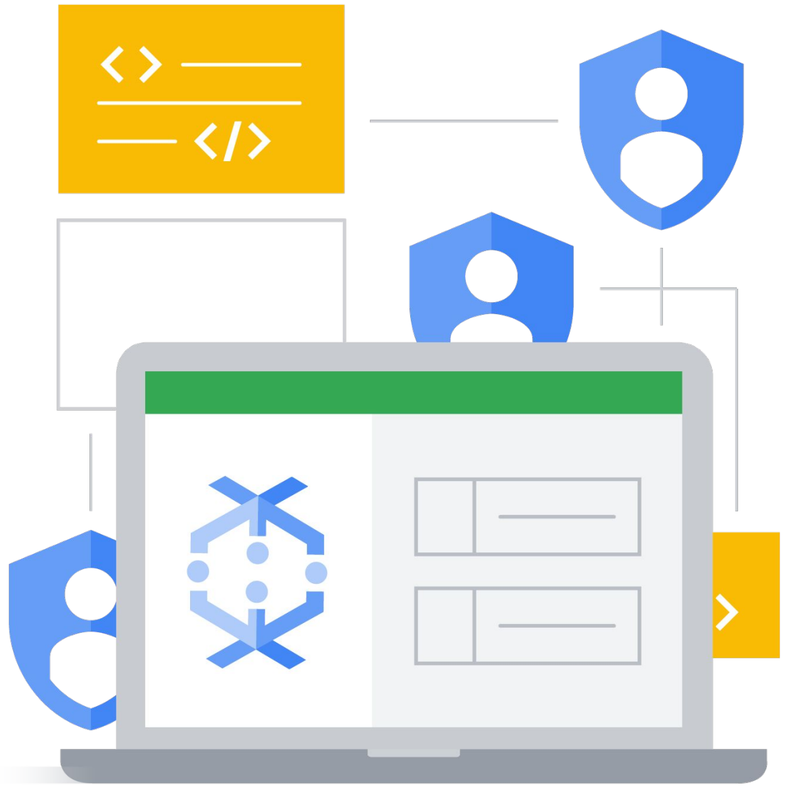
- Interacts between your project and Dataflow



Dataflow service account

The Orchestrator

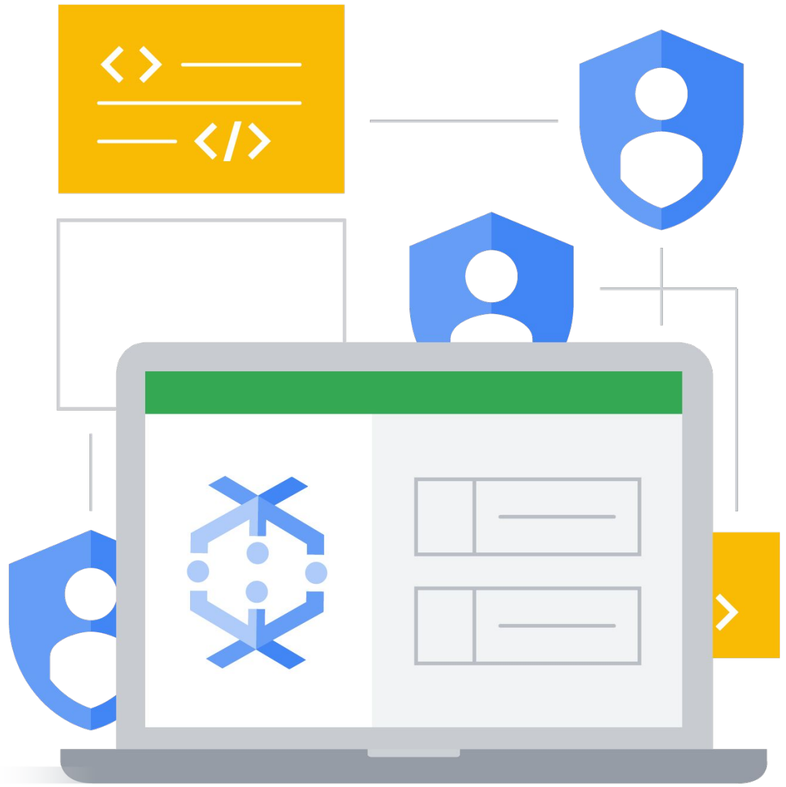
- Interacts between your project and Dataflow
- Used for worker creation and monitoring



Dataflow service account

The Orchestrator

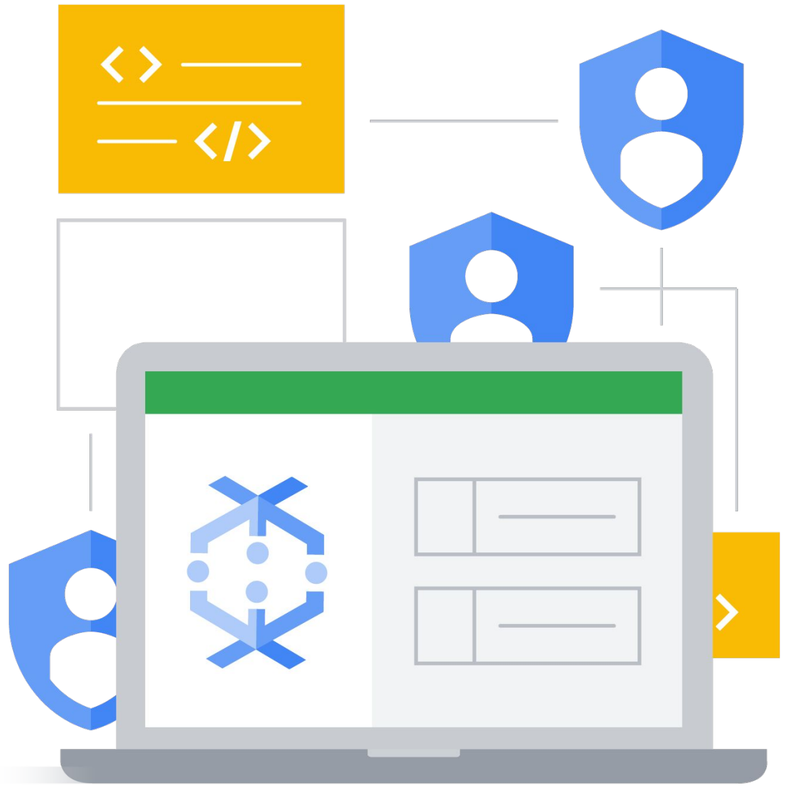
- Interacts between your project and Dataflow
- Used for worker creation and monitoring
- service-<project_number>@dataflow-service-producer-prod.iam.gserviceaccount.com



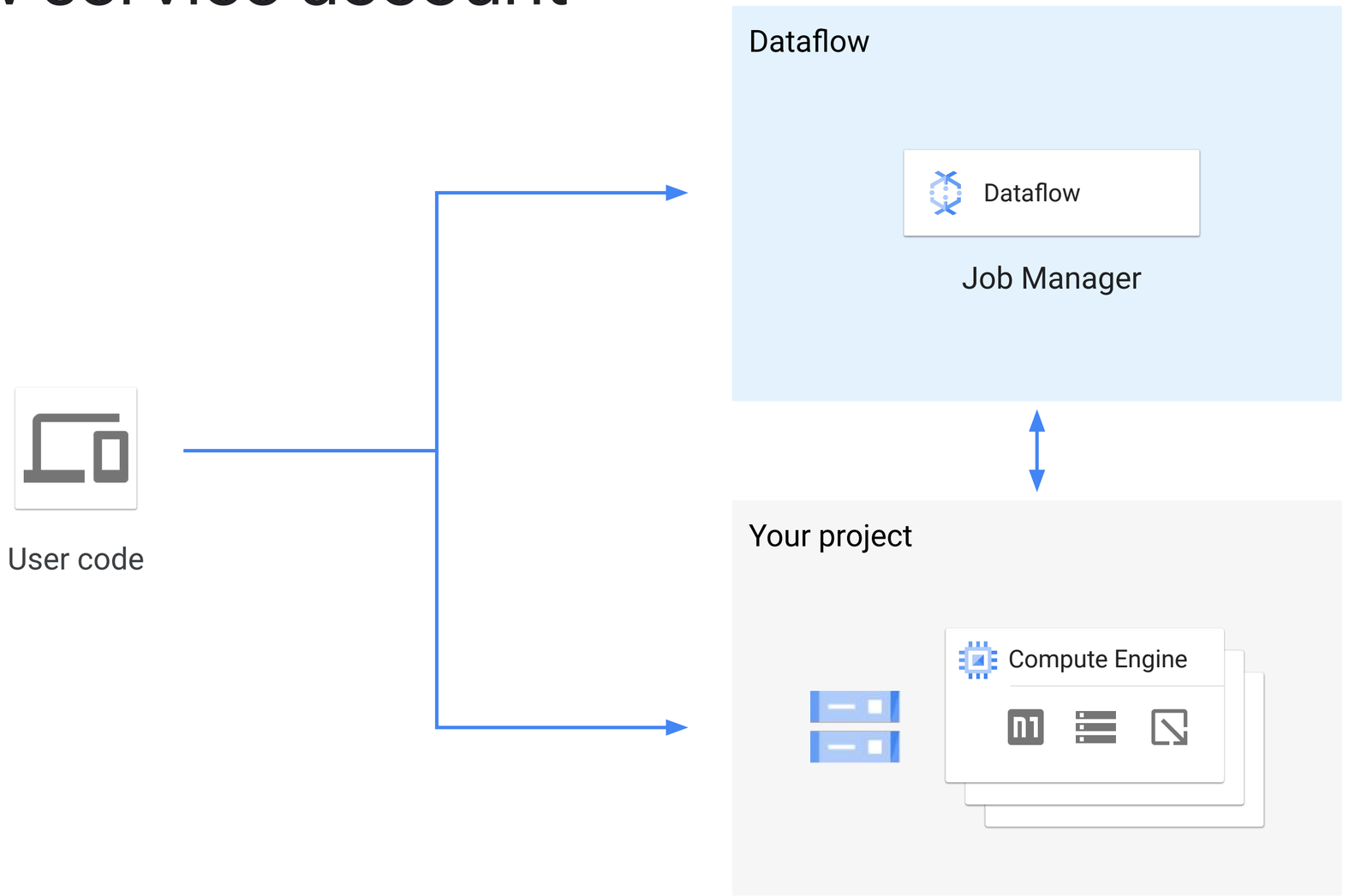
Dataflow service account

The Orchestrator

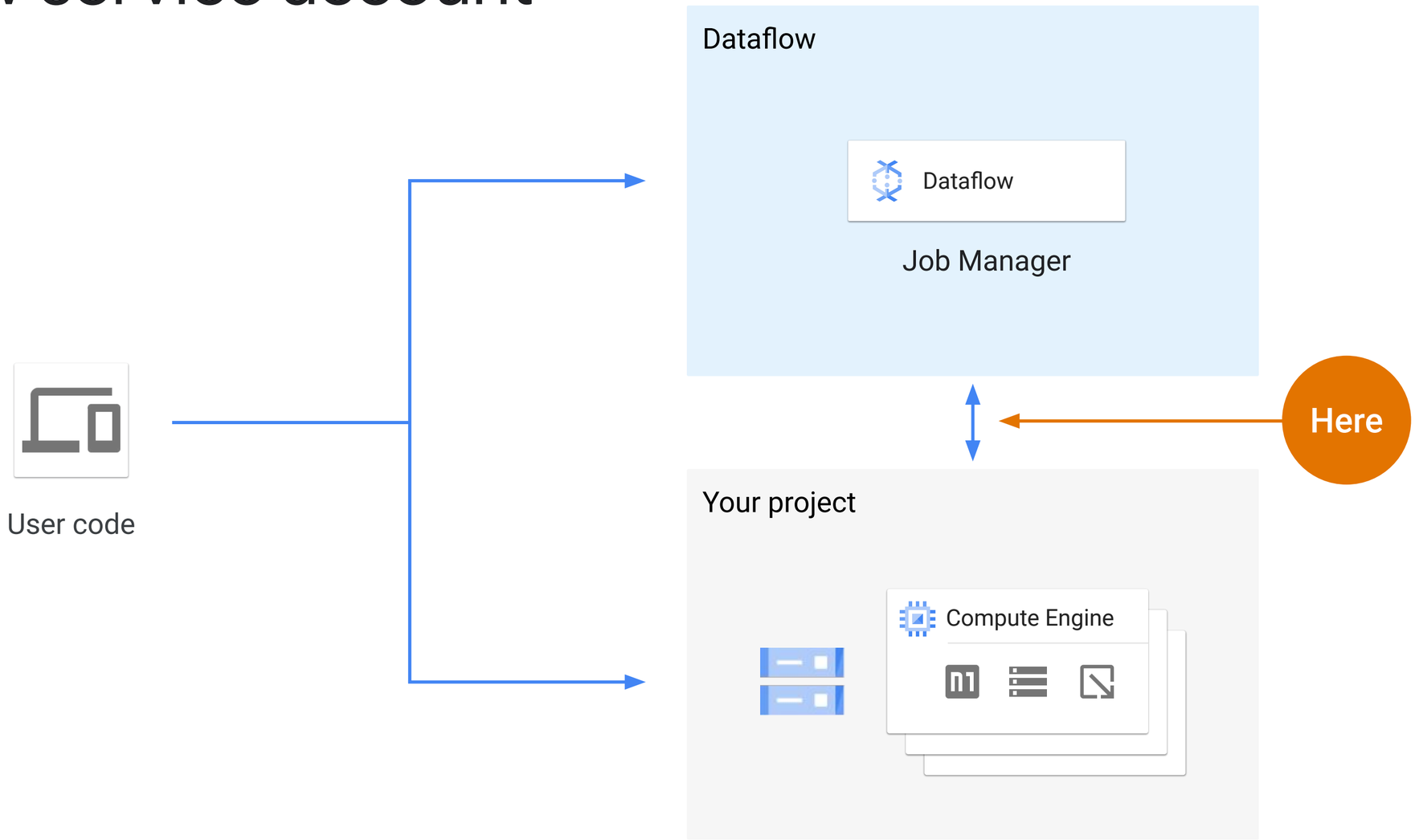
- Interacts between your project and Dataflow
- Used for worker creation and monitoring
- service-<project_number>@dataflow-service-producer-prod.iam.gserviceaccount.com
- Assigned the Dataflow Service Agent role



Dataflow service account



Dataflow service account



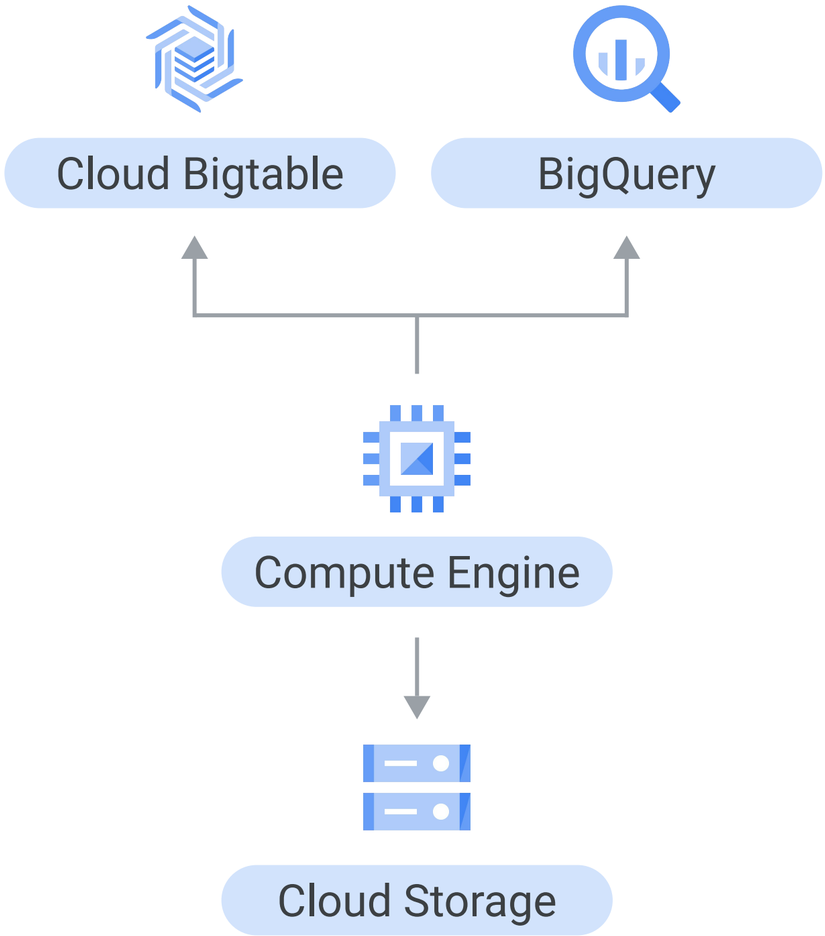
Three credentials

1. User roles
2. Dataflow service account
3. Controller service account

Controller service account

The Worker

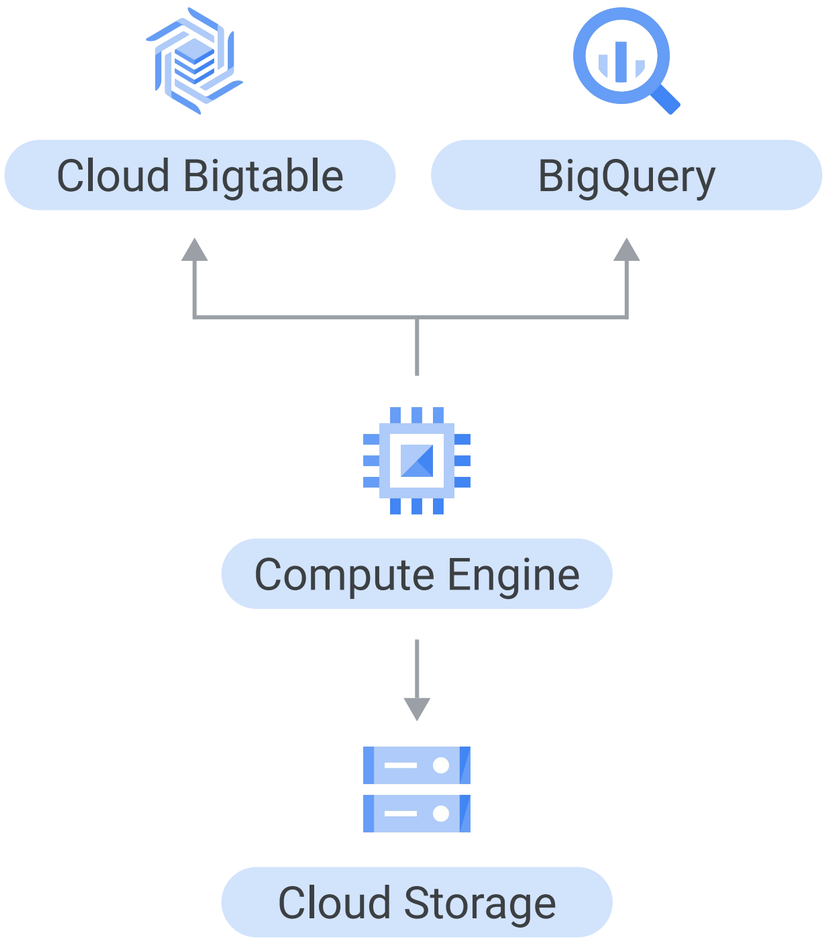
- Used by the workers to access resources needed by the pipeline



Controller service account

The Worker

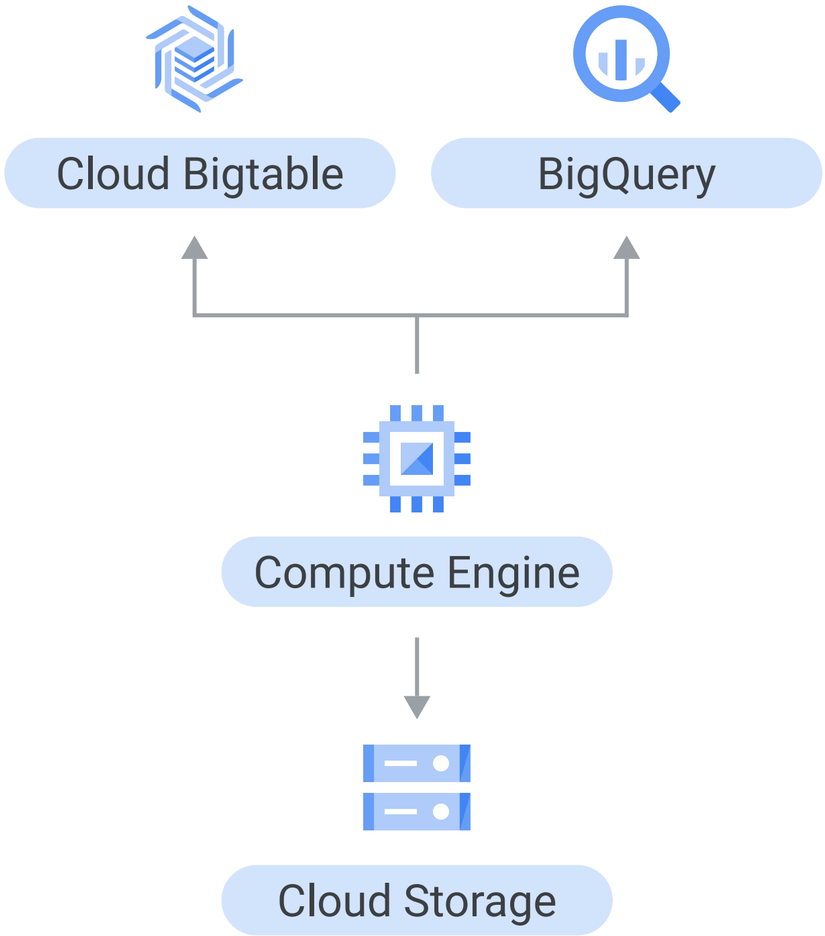
- Used by the workers to access resources needed by the pipeline
- <project-number>-compute@developer.gserviceaccount.com



Controller service account

The Worker

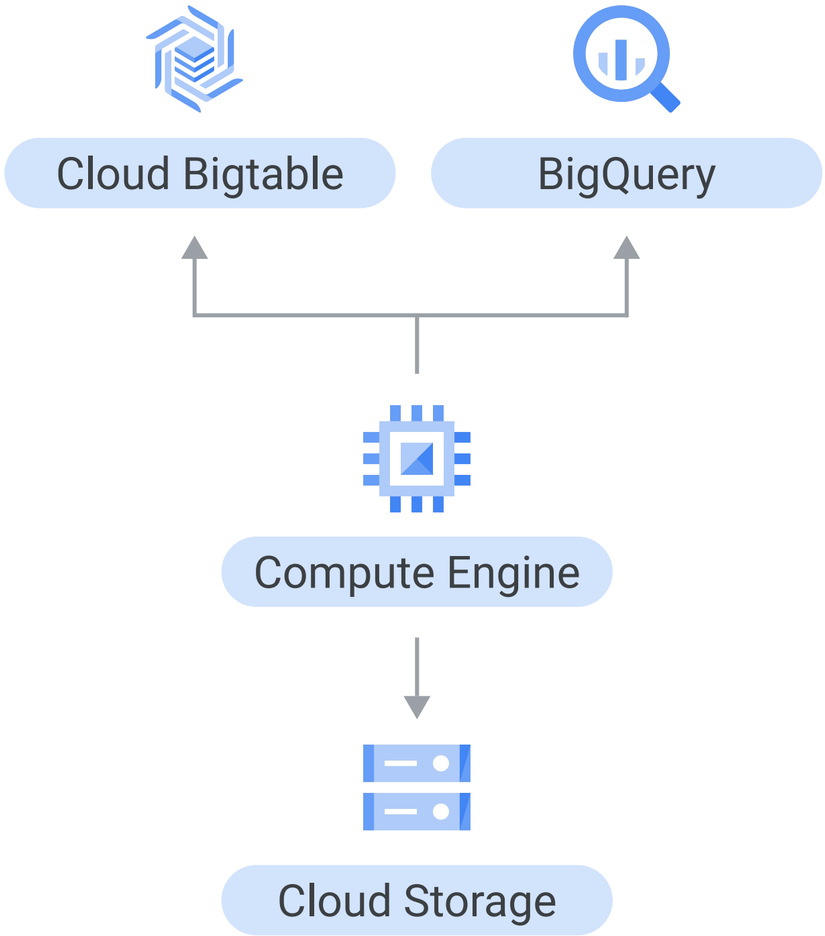
- Used by the workers to access resources needed by the pipeline
- <project-number>-compute@developer.gserviceaccount.com
- Flag to override default:
  - Python: **--service_account_email**
  - Java: **--serviceAccount**



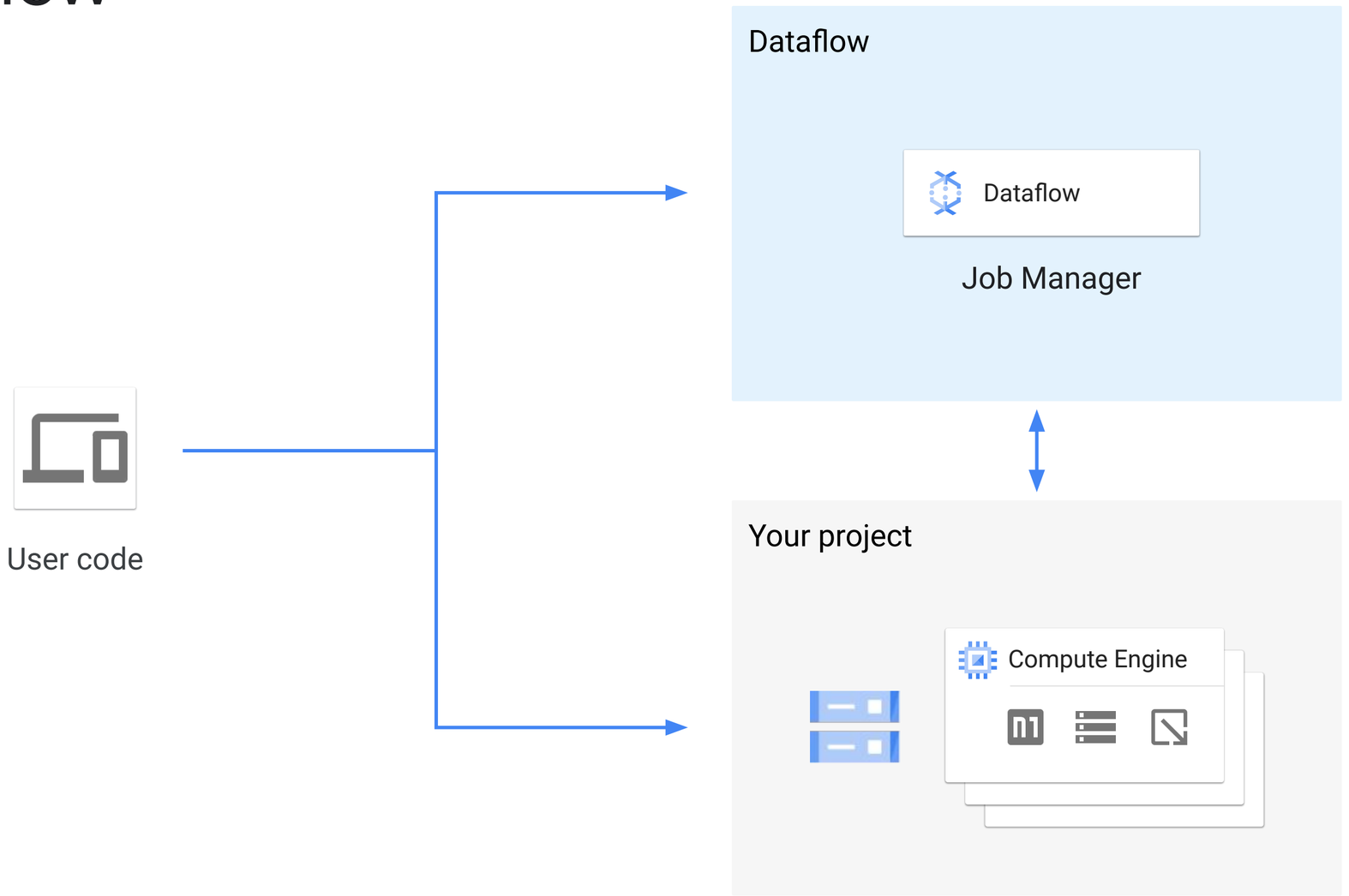
Controller service account

The Worker

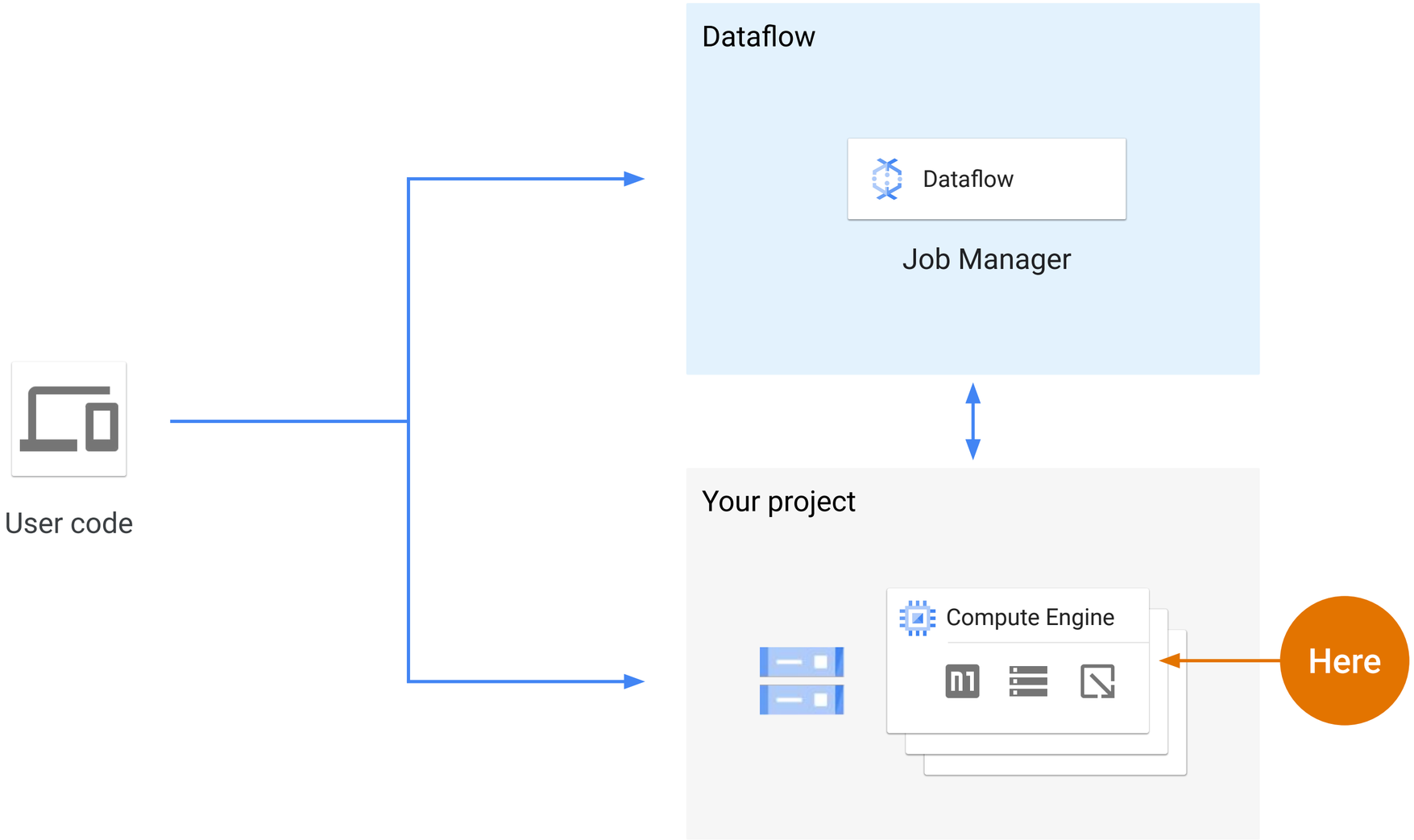
- Used by the workers to access resources needed by the pipeline
- <project-number>-compute@developer.gserviceaccount.com
- Flag to override default:
  - Python: **--service_account_email**
  - Java: **--serviceAccount**
- At a minimum, your custom service account must have the Dataflow Worker role.



Let's review


Dataflow
Dataflow
Job Manager
User code
Your project
Compute Engine

Answer


Dataflow
Dataflow
Job Manager
User code
Your project
Compute Engine
Here

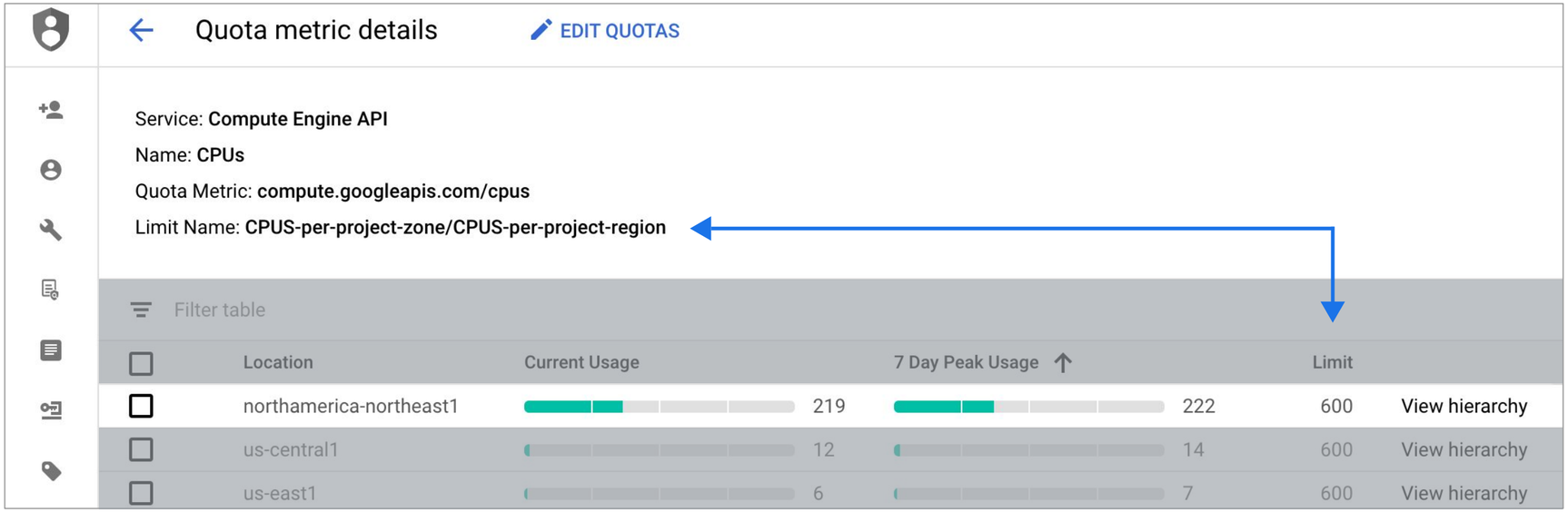
IAM, quotas, and permissions

Agenda

Quotas

CPUs

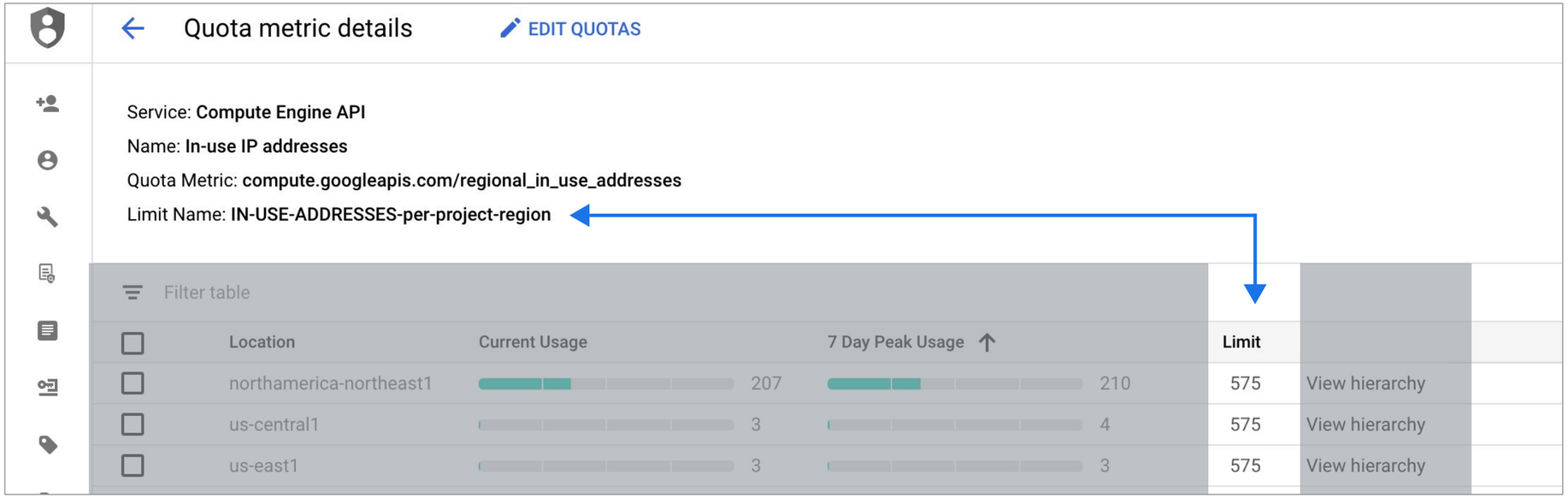
CPU quota based on number of cores



Quotas

IPs

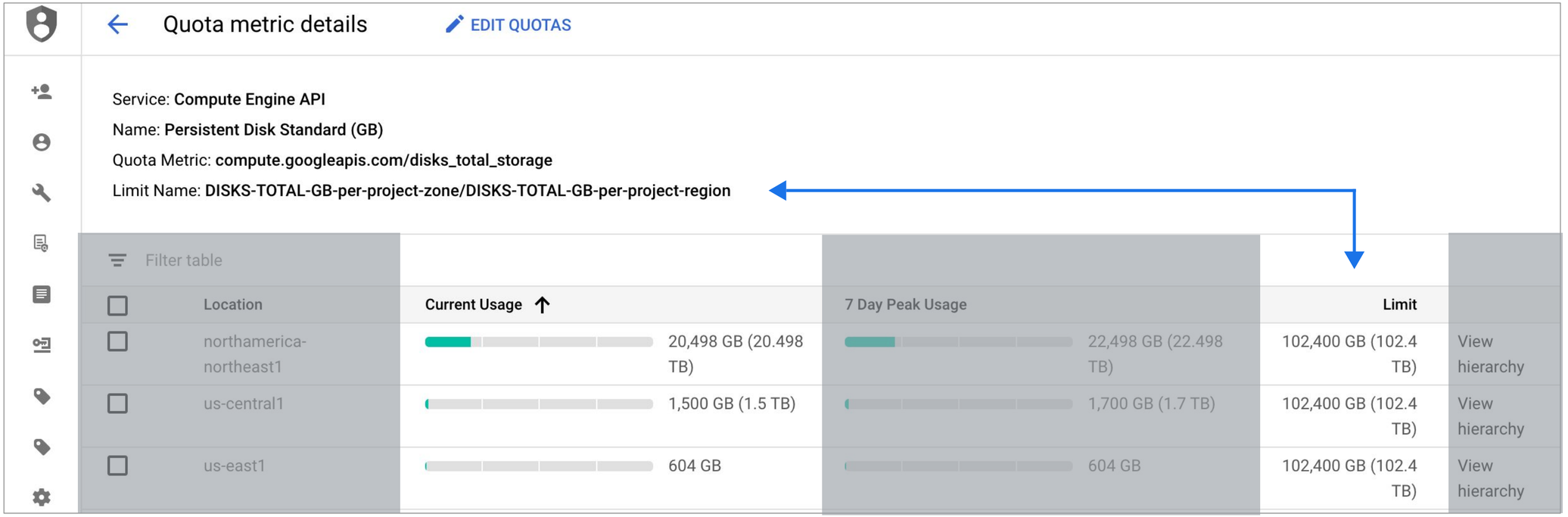
In-use IP addresses must be sufficient to accommodate the desired number of instances



Quotas

Persistent Disks

Choose either HDD or SSD



Quotas

Persistent Disks

Python: Use the **--worker_disk_type** flag

```
$ python3 -m apache_beam.examples.wordcount \
  --input gs://dataflow-samples/shakespeare/kinglear.txt \
  --output gs://$BUCKET/results/outputs --runner DataflowRunner \
  --project $PROJECT --temp_location gs://$BUCKET/tmp/  --region $REGION \
  --worker_disk_type compute.googleapis.com/projects/$PROJECT/zones/$ZONE/diskTypes/pd-ssd
```

Java: Use the **--workerDiskType** flag

```
$ gradle clean execute -DmainClass=org.apache.beam.examples.WordCount -Dexec.args="\
  --inputFile=gs://apache-beam-samples/shakespeare/kinglear.txt \
  --output=gs://$BUCKET/results/outputs --runner=DataflowRunner \
  --project=$PROJECT --tempLocation=gs://$BUCKET/tmp/  --region=$REGION \
  --workerDiskType=compute.googleapis.com/projects/$PROJECT/zones/$ZONE/diskTypes/pd-ssd"
```

Quotas

Persistent Disks - Batch Pipeline

- VM to PD ratio is 1:1 for Batch

Quotas

Persistent Disks - Batch Pipeline

- VM to PD ratio is 1:1 for Batch
- Size if Shuffle on VM: 250 GB

Quotas

Persistent Disks - Batch Pipeline

- VM to PD ratio is 1:1 for Batch
- Size if Shuffle on VM: 250 GB
- Size if Shuffle Service: 25 GB

Quotas

Persistent Disks - Batch Pipeline

- VM to PD ratio is 1:1 for Batch
- Size if Shuffle on VM: 250 GB
- Size if Shuffle Service: 25 GB
- Flag to override default:
  Python: **--disk_size_gb**
  Java: **--diskSizeGb**

Quotas

Persistent Disks - Streaming Pipeline

- Fixed number of PDs

Quotas

Persistent Disks - Streaming Pipeline

- Fixed number of PDs
- Default size if shuffle on VM: 400 GB

Quotas

Persistent Disks - Streaming Pipeline

- Fixed number of PDs
- Default size if shuffle on VM: 400 GB
- Default size if Streaming Engine: 30 GB

Quotas

Persistent Disks - Streaming Pipeline

- Fixed number of PDs
- Default size if shuffle on VM: 400 GB
- Default size if Streaming Engine: 30 GB
- Flag to override default:

  Python: **--disk_size_gb**

  Java: **--diskSizeGb**

Quotas

Persistent Disks - Streaming Pipeline

- Amount of disk allocated == Maximum number of workers

Quotas

Persistent Disks - Streaming Pipeline

- Amount of disk allocated == Maximum number of workers
- Flag to set maximum number of workers:

  Python: --max_num_workers
  Java: --maxNumWorkers
- Flag required for streaming with shuffle on VMs
- Maximum number of workers that can be launched is 1000