# Pre-requisites

## Tutorial

Follow the tutorial Batch Job on Dataflow Quickstart on how to run a simple batch job on Dataflow.

## GCP setup

Ensure `gcloud` CLI is installed and `application\_default\_credentials.json` has been generated.

## Data access

If you want to run an adhoc job with your ldap, you will need access to multiple LDAP groups to read the datasets.

# Running the job

### Running an adhoc job

```bash

export GCP\_PROJECT\_NAME='twttr-recos-ml-prod'

./bazel bundle src/scala/com/twitter/simclusters\_v2/scio/multi\_type\_graph/assemble\_multi\_type\_graph:assemble-multi-type-graph-scio-adhoc-app

bin/d6w create \

${GCP\_PROJECT\_NAME}/us-central1/assemble-multi-type-graph-scio-adhoc-app \

src/scala/com/twitter/simclusters\_v2/scio/multi\_type\_graph/assemble\_multi\_type\_graph/assemble-multi-type-graph-scio-adhoc.d6w \

--jar dist/assemble-multi-type-graph-scio-adho-app.jar \

--bind=profile.project=${GCP\_PROJECT\_NAME} \

--bind=profile.user\_name=${USER} \

--bind=profile.date="2021-11-04" \

--bind=profile.machine="n2-highmem-16"

```

### Scheduling the job on Workflow

Scheduling a job will require a service account as `recos-platform`.

Remember this account will need permissions to read all the required dataset.

```bash

export SERVICE\_ACCOUNT='recos-platform'

export GCP\_PROJECT\_NAME='twttr-recos-ml-prod'

bin/d6w schedule \

${GCP\_PROJECT\_NAME}/us-central1/assemble-multi-type-graph-scio-batch-app \

src/scala/com/twitter/simclusters\_v2/scio/multi\_type\_graph/assemble\_multi\_type\_graph/assemble-multi-type-graph-scio-batch.d6w \

--bind=profile.project=${GCP\_PROJECT\_NAME} \

--bind=profile.user\_name="recos-platform" \

--bind=profile.date="2021-11-04" \

--bind=profile.machine="n2-highmem-16"

```