Security

Omar Ismail

Solutions Developer, Google Cloud

Agenda

Course Intro

Beam and Dataflow Refresher

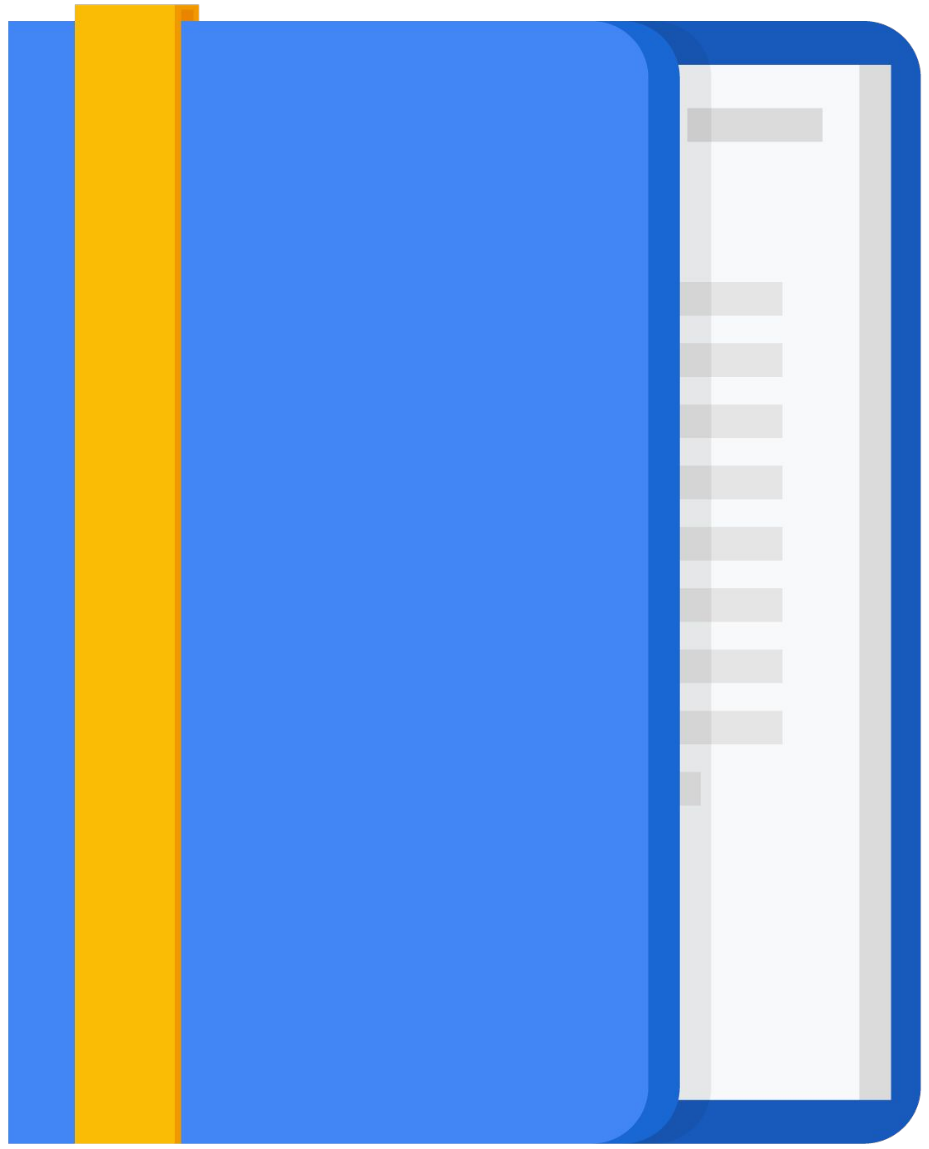
Beam Portability

Separating Compute and Storage

IAM, Quotas, and Permissions

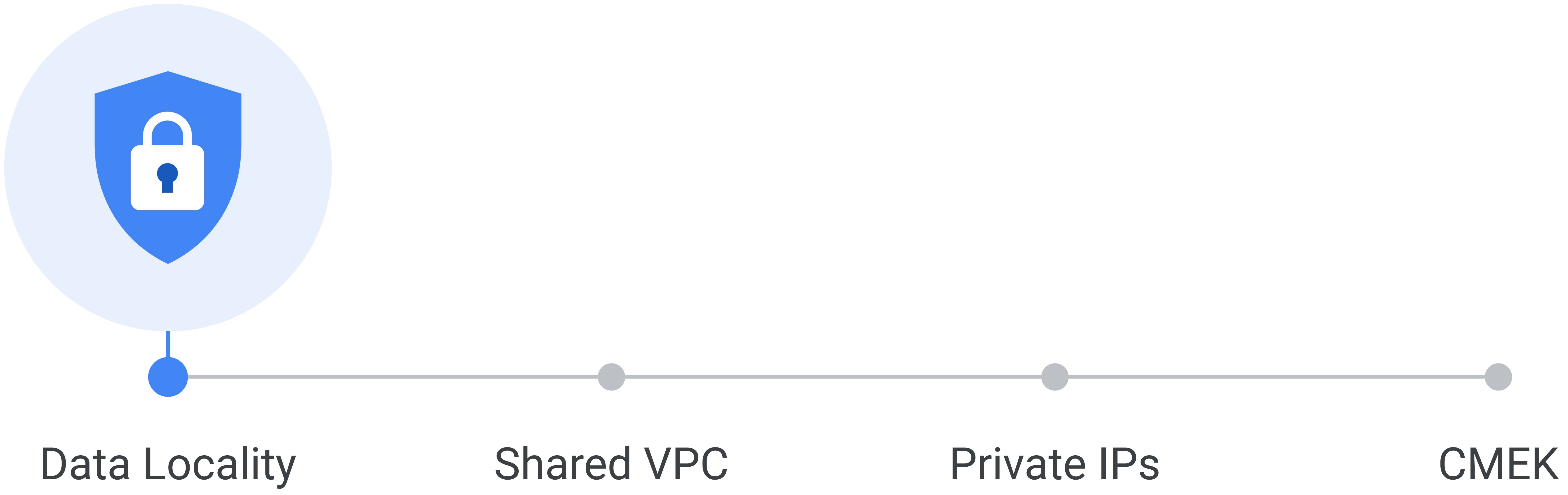
Security

Summary



Security

Agenda



Data Locality

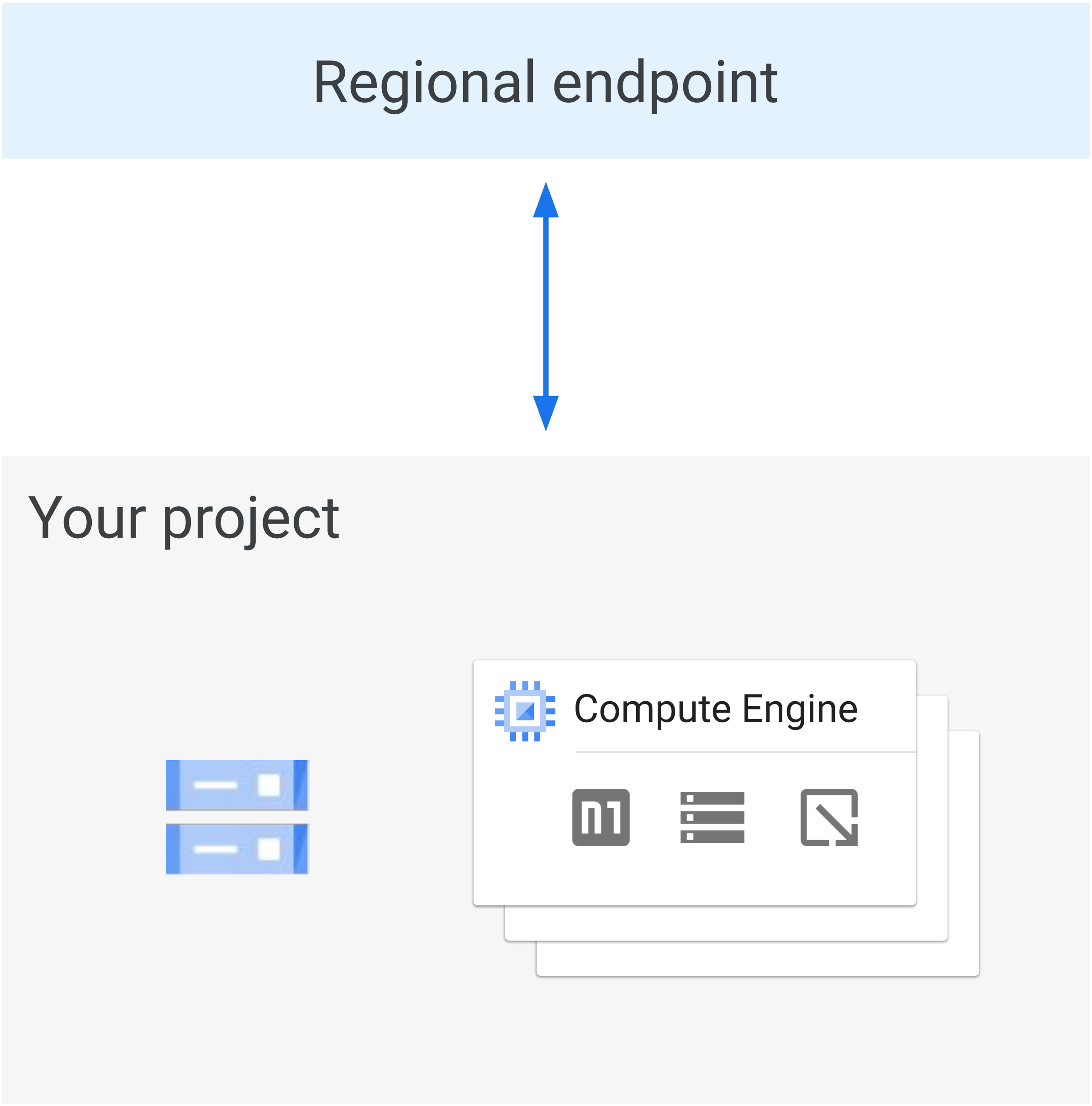
Shared VPC

Private IPs

CMEK

Data locality

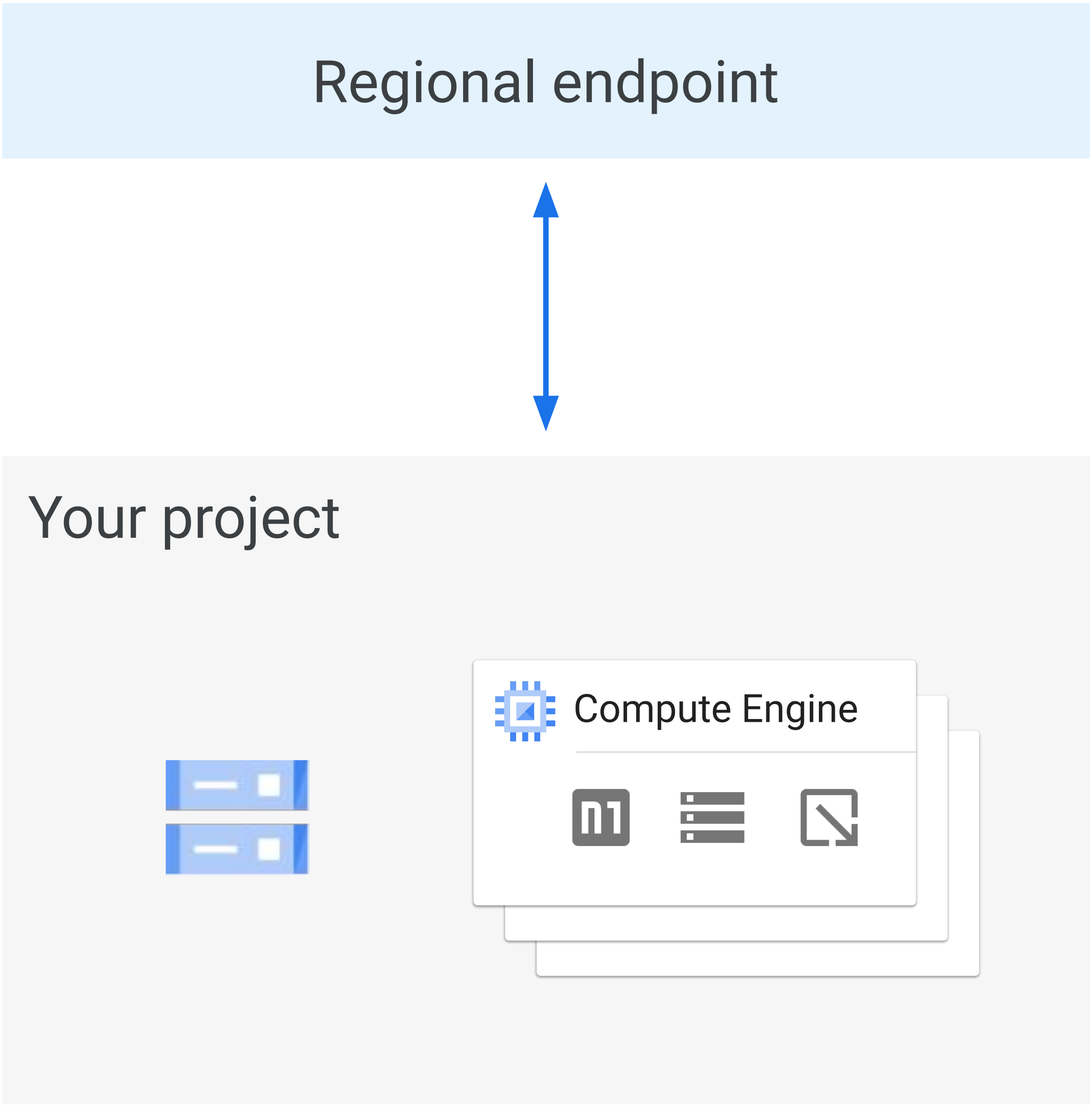
What is a regional endpoint?



Data locality

What is a regional endpoint?

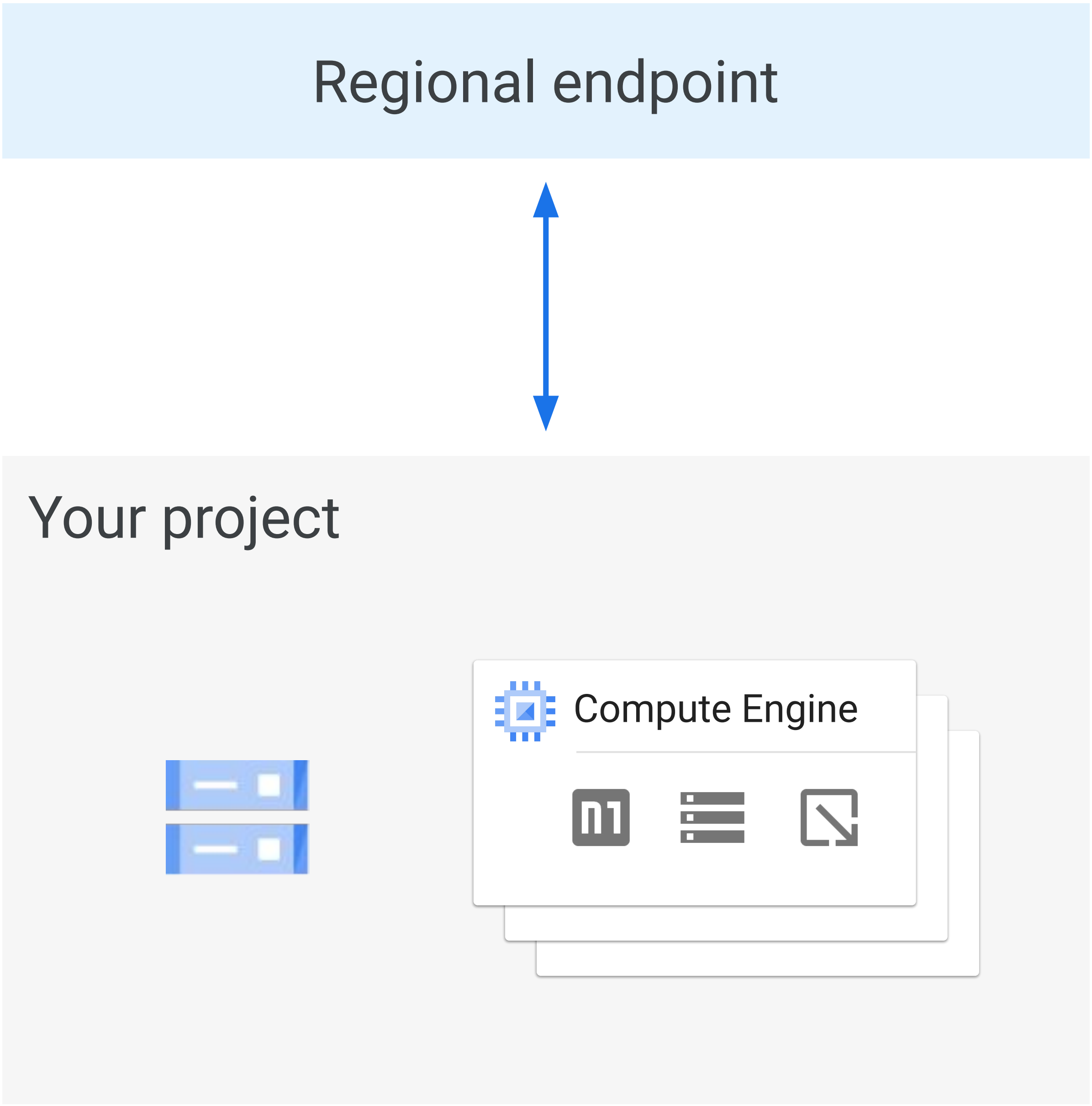
- Backend that deploys and controls your Dataflow workers



Data locality

What is a regional endpoint?

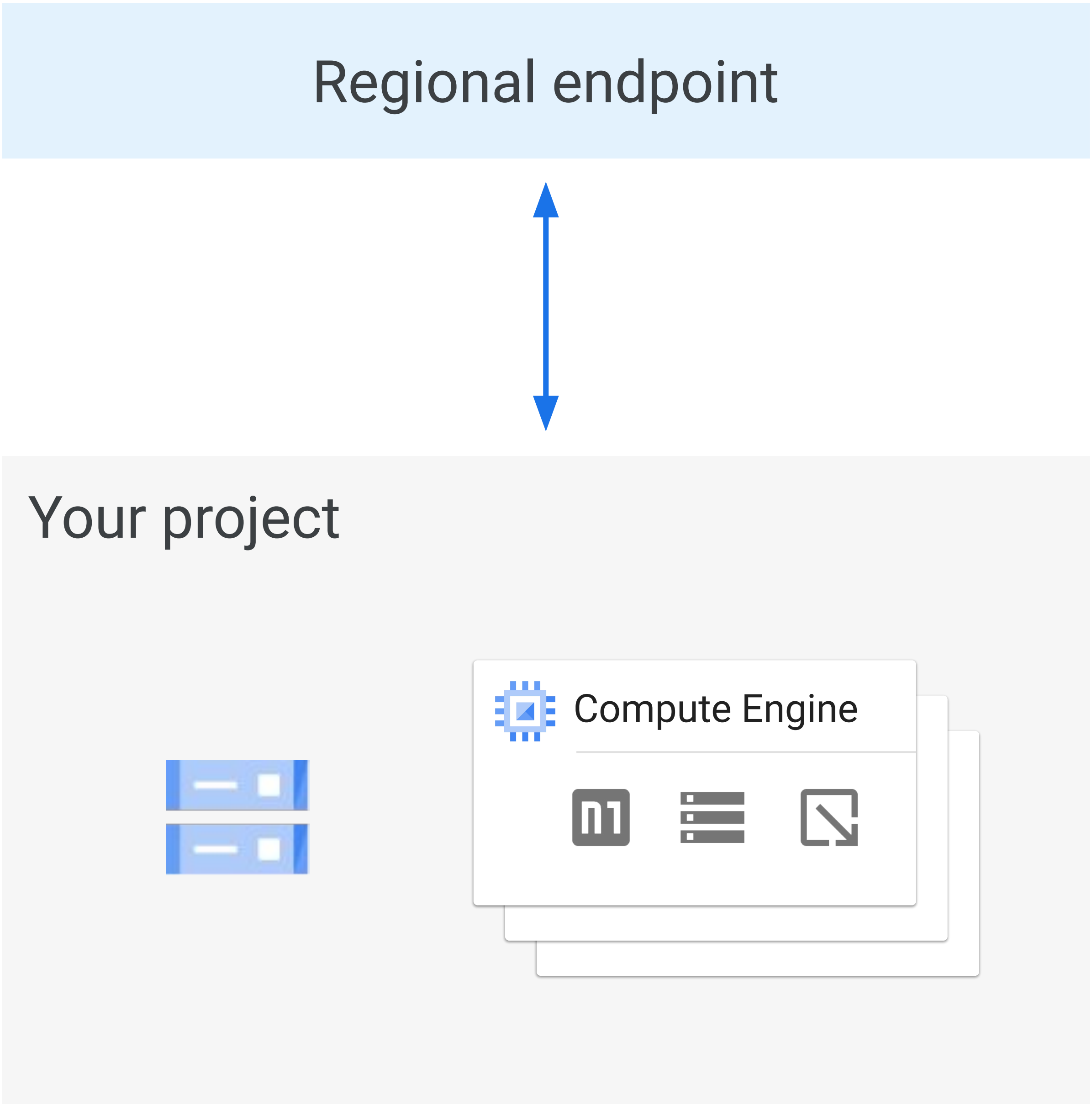
- Backend that deploys and controls your Dataflow workers
- Talks with the Dataflow service account in your project



Data locality

What is a regional endpoint?

- Backend that deploys and controls your Dataflow workers
- Dataflow service account in your project talks with the regional endpoint
- Stores and handles metadata about your Dataflow job



Data locality

Why specify a regional endpoint?

Data locality

Why specify a regional endpoint?

- Security and compliance

Data locality

Why specify a regional endpoint?

- Security and compliance
- Minimize network latency and network transport costs

Data locality: How to specify a regional endpoint

No zone preference

```
$ 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
```

```
$ 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"
```

Data locality: How to specify a regional endpoint

Run worker in a specific zone in a region with a regional endpoint

```
$ 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_zone $WORKER_ZONE
```

```
$ 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 --workerZone=$WORKER_ZONE"
```

Data locality: How to specify a regional endpoint

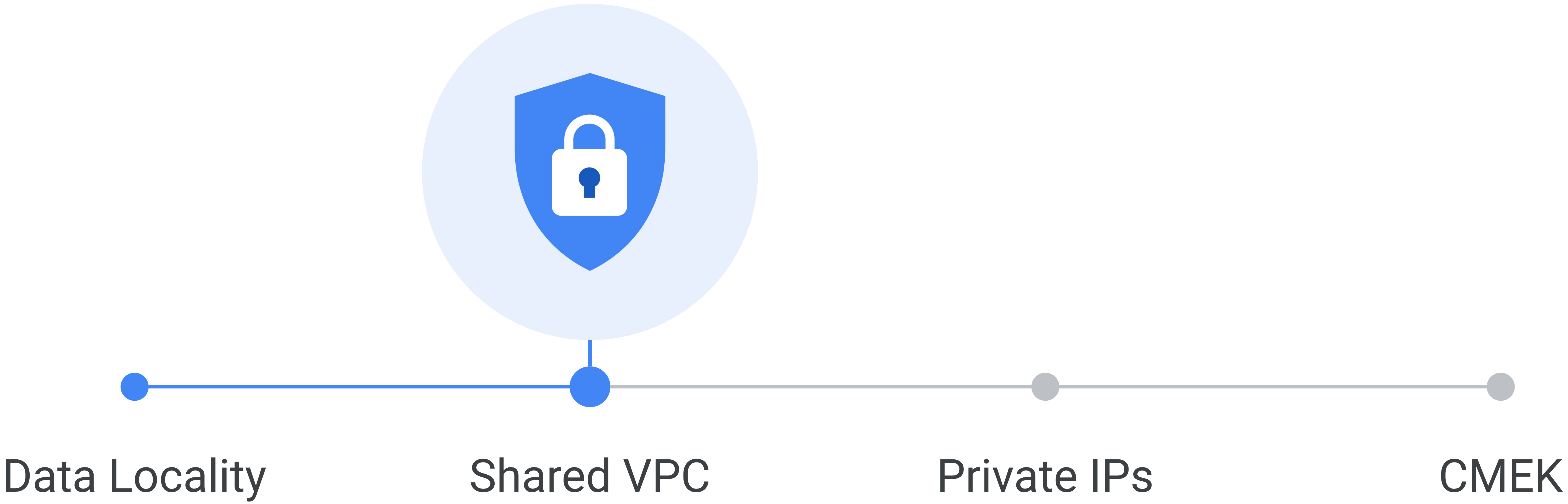
Run worker in a region with no regional endpoint

```
$ 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_region $WORKER_REGION
```

```
$ 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 --workerRegion=$WORKER_REGION"
```

Security

Agenda



Data Locality

Shared VPC

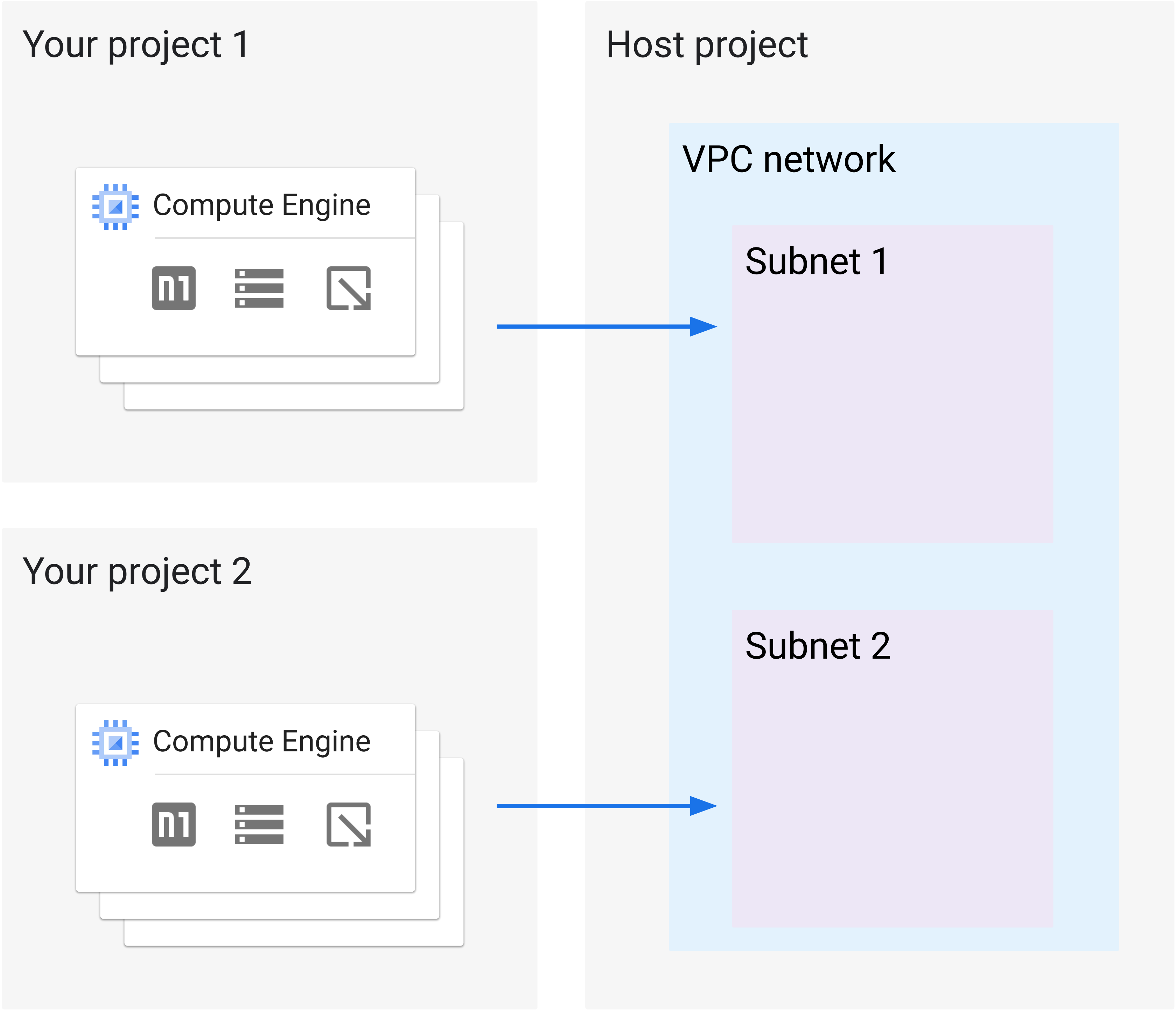
Private IPs

CMEK

Shared VPC

Hosts and services

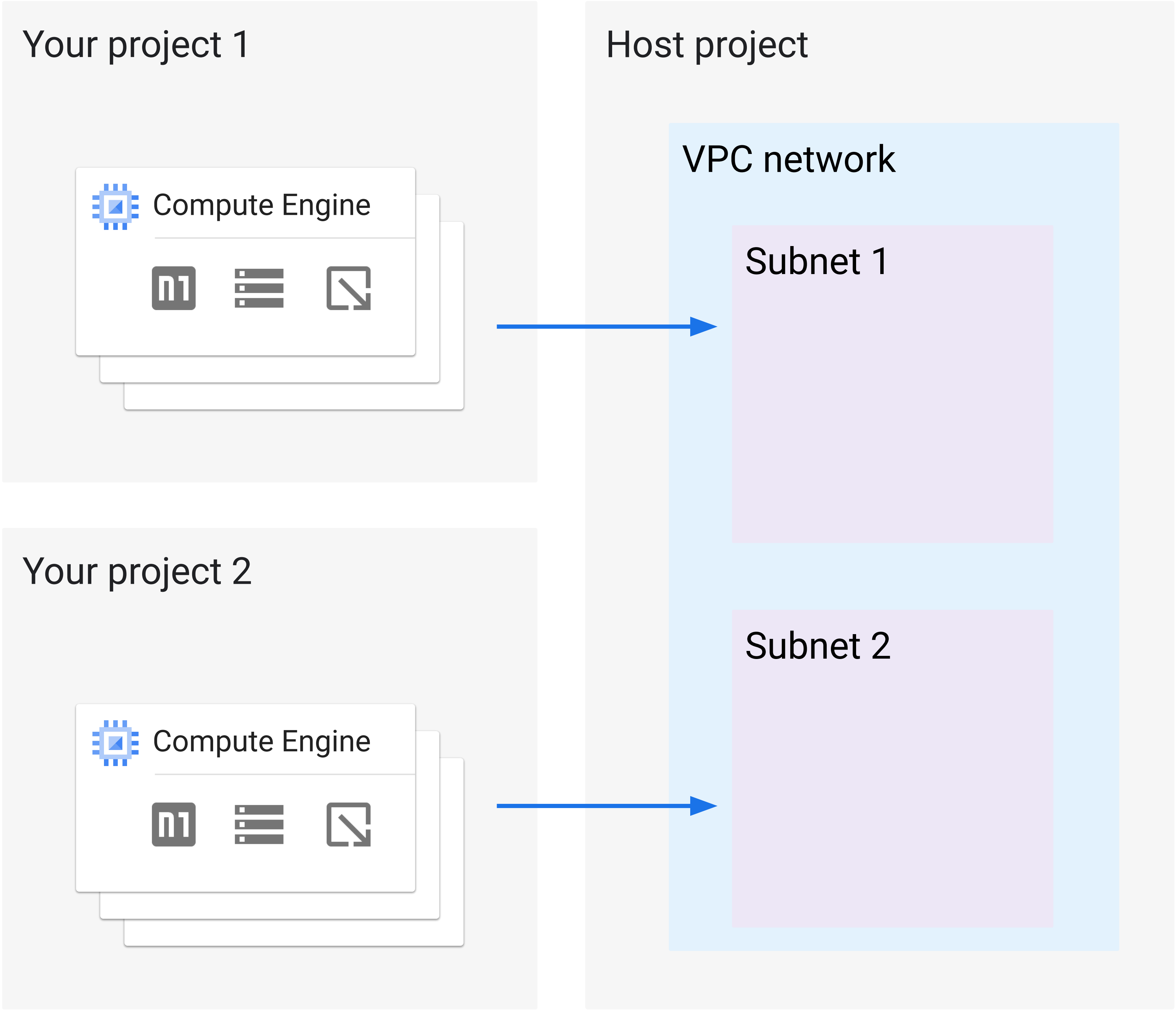
- Dataflow jobs can run in either VPC or Shared VPC



Shared VPC

Hosts and services

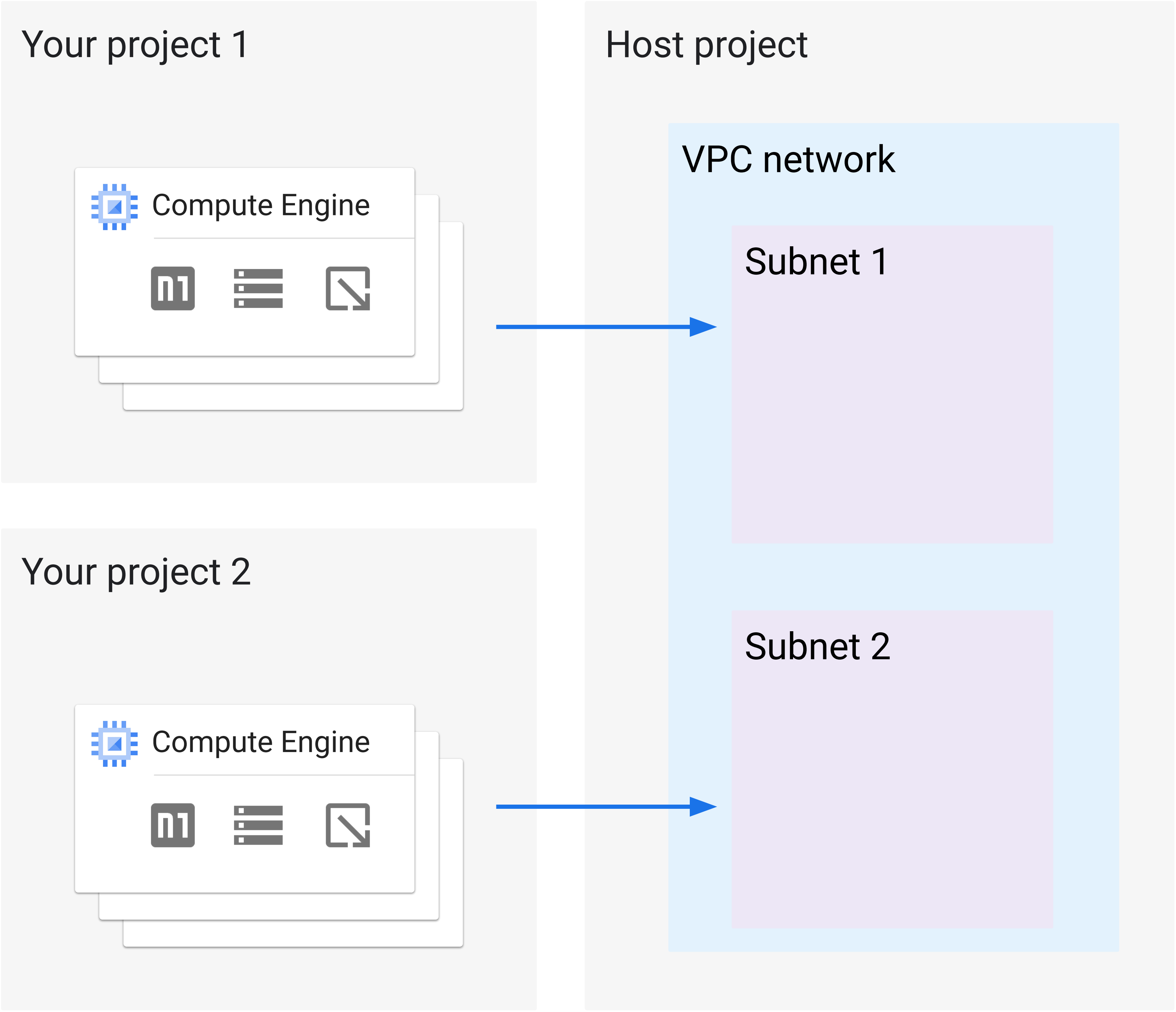
- Dataflow jobs can run in either VPC or Shared VPC
- Works for both default and custom networks



Shared VPC

Hosts and services

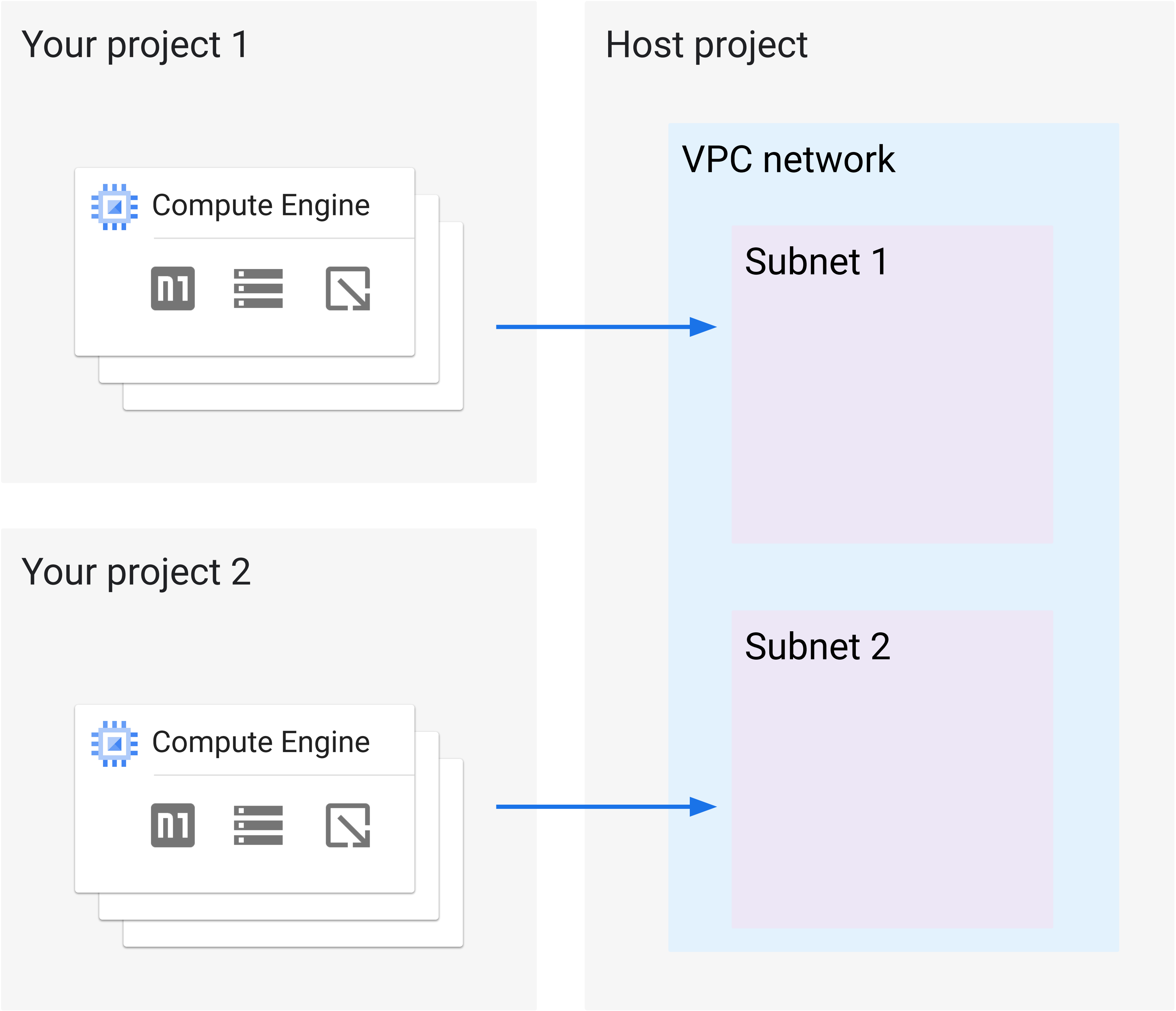
- Dataflow jobs can run in either VPC or Shared VPC
- Works for both default and custom networks
- Number of VMs is constrained by subnet IP block size



Shared VPC

Hosts and services

- Dataflow jobs can run in either VPC or Shared VPC
- Works for both default and custom networks
- Number of VMs is constrained by subnet IP block size
- Dataflow service account needs Compute Network User role in host project



Shared VPC: How to set

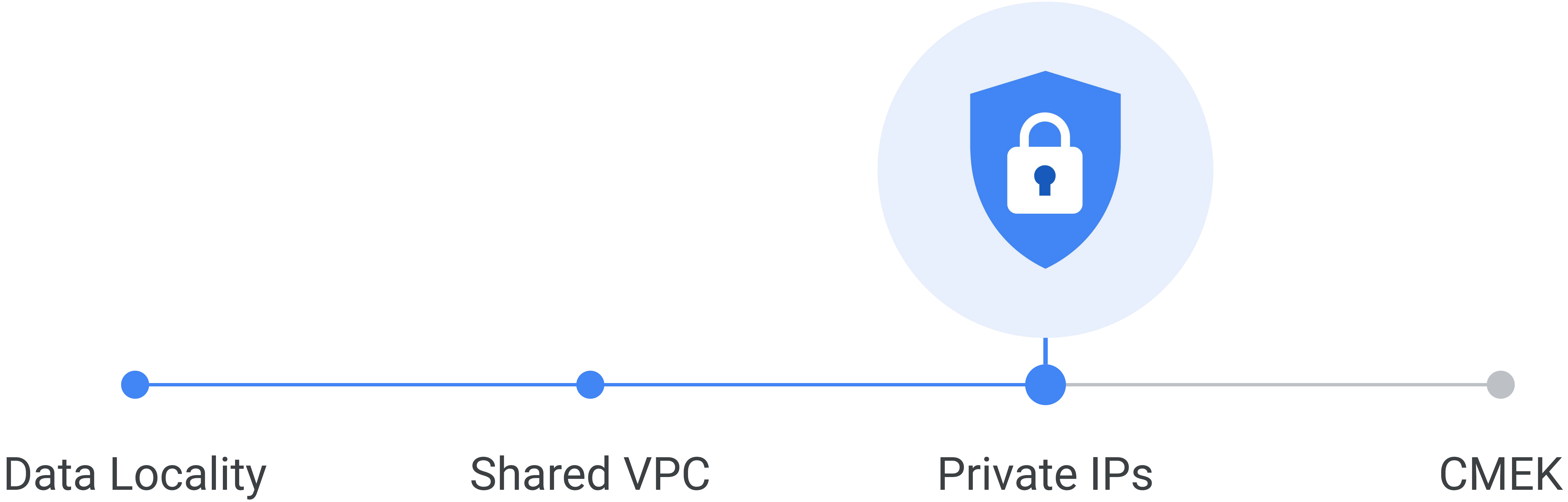
Use --network or --subnetwork flag

```
$ 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 \
  --network default
```

```
$ 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 \
  --subnetwork=https://www.googleapis.com/compute/v1/projects/$HOST_PROJECT_ID/regions/$REG
  ION/subnetworks/$SUBNETWORK
```

Security

Agenda



Data Locality

Shared VPC

Private IPs

CMEK

Private IPs

No external IPs

- Secure your data processing infrastructure

Private IPs

No external IPs

- Secure your data processing infrastructure
- Pipeline cannot access the internet and other Google Cloud networks

Private IPs

No external IPs

- Secure your data processing infrastructure
- Pipeline cannot access the internet and other Google Cloud networks
- Network must have Private Google Access on in order to reach Google Cloud APIs and services

Private IPs: How to set

Python: Use **--network** or **--subnetwork** flag and **--no_use_public_ips** flag

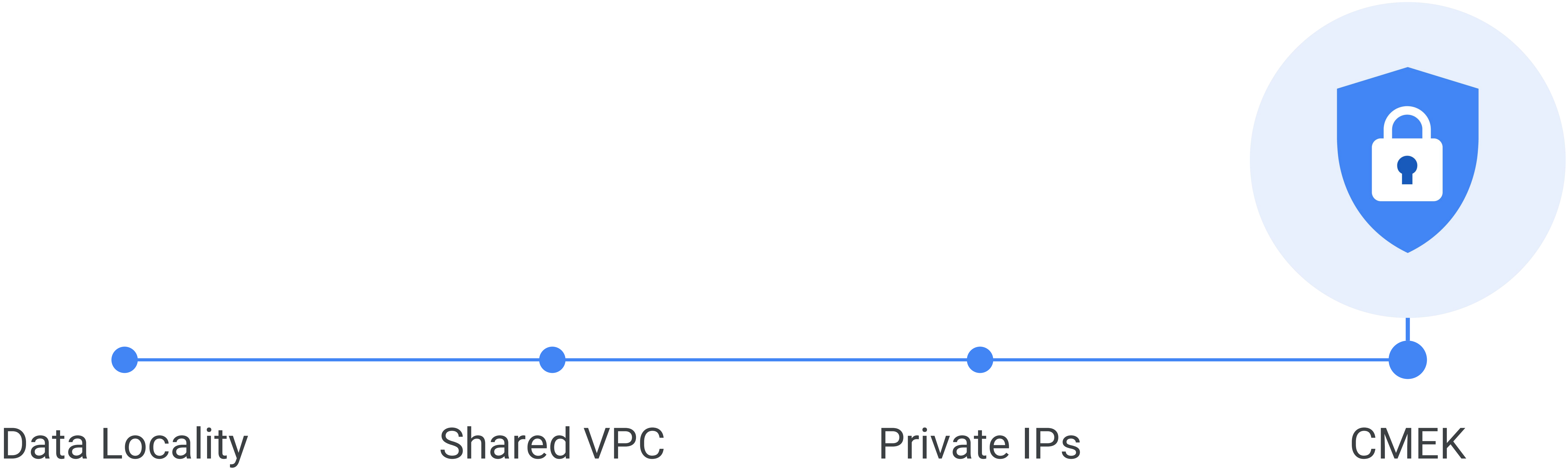
```
$ python3 -m apache_beam.examples.wordcount \
  --input gs://dataflow-samples/shakespeare/kinglear.txt \
  --output gs://$BUCKET/results/outputs --runner DataflowRunner  \
  --project $PROJECT  --temp_location gs://$BUCKET/tmp/  --region $REGION \
  --subnetwork regions/$REGION/subnetworks/$SUBNETWORK \
  --no_use_public_ips
```

Java: Use **--network** or **--subnetwork** flag and **--usePublicIps** 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 \
  --subnetwork=regions/$REGION/subnetworks/$SUBNETWORK \
  --usePublicIps=false”
```

Security

Agenda



Data Locality

Shared VPC

Private IPs

CMEK

CMEK

What is it?

- Where data is stored:
  - Persistent Disk
  - Storage buckets
  - Dataflow Shuffle backend
  - Streaming Engine backend

CMEK

What is it?

- Where data is stored:
  - Persistent Disk
  - Storage buckets
  - Dataflow Shuffle backend
  - Streaming Engine backend
- Data keys in grouping operations are decrypted using CMEK key.

CMEK

What is it?

- Where data is stored:
  - Persistent Disk
  - Storage buckets
  - Dataflow Shuffle backend
  - Streaming Engine backend
- Data keys in grouping operations are decrypted using CMEK key.
- Metadata is protected by Google-managed key encryption.

CMEK

What is it?

- Where data is stored:
  - Persistent Disk
  - Storage buckets
  - Dataflow Shuffle backend
  - Streaming Engine backend
- Data keys in grouping operations are decrypted using CMEK key.
- Metadata is protected by Google-managed key encryption.
- Add Cloud KMS CryptoKey Encrypter/Decrypter role to Dataflow service account and Controller Agent service account.

CMEK: How to set

Python: Use **--temp_location** and **--dataflow_kms_key** flags

```
$ python3 -m apache_beam.examples.wordcount \
  --input gs://dataflow-samples/shakespeare/kinglear.txt \
  --output gs://$BUCKET/results/outputs --runner DataflowRunner  \
  --project $PROJECT --region $REGION --temp_location gs://$BUCKET/tmp/ \
  --dataflow_kms_key=projects/$PROJECT/locations/$REGION/keyRings/$KEY_RING/cryptoKeys/$KEY
```

Java: Use **--tempLocation** and **dataflowKmsKey** flags

```
$ 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 --region=$REGION --tempLocation=gs://$BUCKET/tmp/ \
  --dataflowKmsKey=projects/$PROJECT/locations/$REGION/keyRings/$KEY_RING/cryptoKeys/$KEY”
```