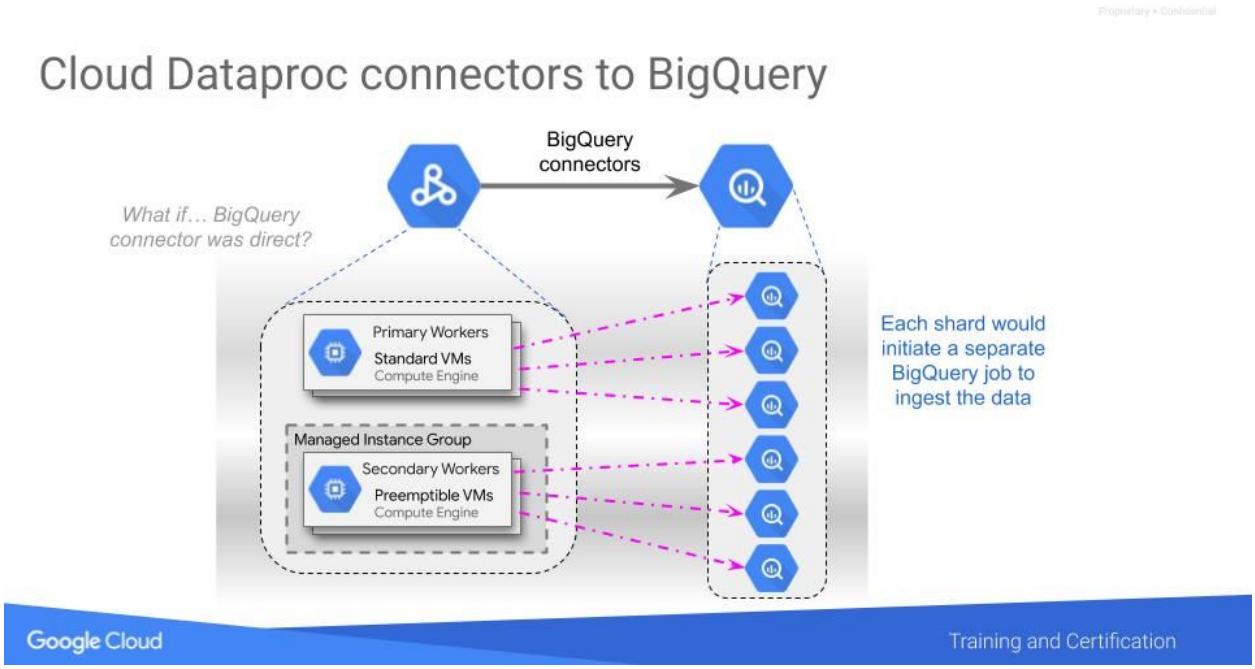
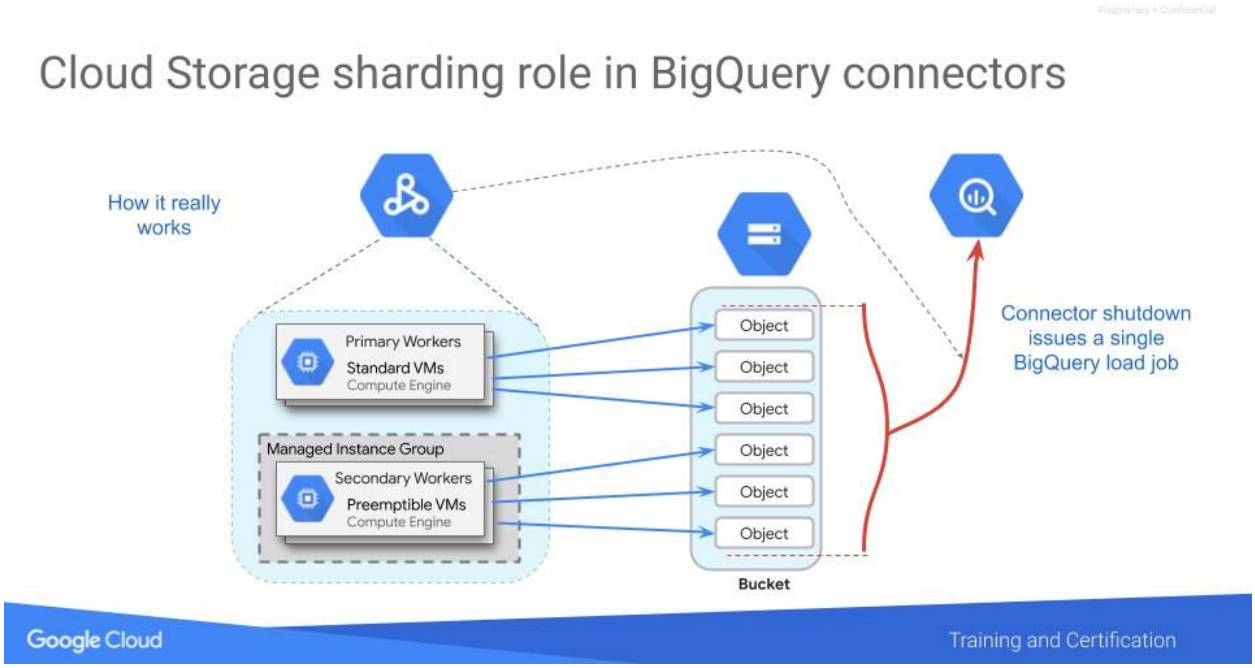
Cloud Dataproc to BigQuery Connectors



Cloud Dataproc has a collection of connectors to BigQuery from Map Reduce, Hadoop, Spark Scala, and Spark Pyspark.

The way you would probably expect this to work is shown. And it kind of does. But there are some more details involving Cloud Storage.

The main issue has to do with sharding. Cloud Dataproc uses distributed parallel processing. The workers generate separate shards of the output. If BigQuery was triggered to ingest each shard, it would cause a separate BigQuery job to run for each shard of data.



To overcome this problem, Cloud Dataproc's BigQuery Connectors use Cloud Storage.
The workers are able to write their shards as objects in Cloud Storage.
When the work is complete, as part of the connector shutdown process, a single load job is issued in BigQuery.

BigQuery job submission works slightly differently depending on the Dataproc client.
For example Spark Scala is able to directly generate the request via the API to BigQuery. But PySpark does not support the correct objects, so it issues a bq shell command to run the job.