Stream Data

April 29, 2019

Molecules Example

- `run-local`: runs a typical set of python scripts that uses tensorflow and tensorflow transform plus beam functions
- it does not use anything related with the GCP
- uses only resources in your local machine
- tensorflow is a python module that implements functions to transform data and to build models based on deep learning (multi-layered neural networks)
- apache beam allows for creating pipelines to be executed in parallel

Molecules Example

- `run-cloud`: runs a typical set of python scripts using a google cloud storage (gs://*)
- the only step that runs **on the cloud** is the training step
- script uses commands `gcloud...` to start jobs
- specifically:

```
gcloud ml-engine jobs submit training $JOB \
--module-name trainer.task \
--package-path trainer \
--staging-bucket $BUCKET \
--runtime-version $RUNTIME \
--stream-logs \
-- \
--work-dir $WORK_DIR
```

- the preprocessing and prediction steps use a specific beam runner: `--runner DataflowRunner`

Molecules Example

- the whole process described before runs in **batch** mode
- we can transform the whole process in a **streamed** mode and instead of waiting for all predictions to be ready, to collect them **as they are produced**
- need: publisher, subscriber, **stream** prediction

Molecules Example: simulating stream data