Highlight Note

Pipeline Step Reuse

Pipelines with multiple long-running steps can take a significant time to complete, so Azure Machine Learning includes some default caching and reuse features to reduce this time.

Managing Step Output Reuse

By default the step output from a previous pipeline run is reused without re-running the step as long as the script, source directory, and other parameters for the step have not changed. This can significantly reduce the time it takes to run a pipeline; however it can lead to stale results when changes to downstream data sources have not been accounted for.

To control reuse for an individual step, you can set the **allow_reuse** parameter in the step configuration, like this:

```
step1 = PythonScriptStep(name = 'prepare data',
                         source_directory = 'scripts',
                         script_name = 'data_prep.py',
                         compute_target = 'aml-cluster',
                         runconfig = run_config,
                         inputs=[raw_ds.as_named_input('raw_data')],
                         outputs=[prepped_data],
                         arguments = ['--folder', prepped_data]),
                         # Disable step reuse
                         allow_reuse = False)
```

Forcing All Steps to Run

When you have multiple steps, you can force all of them to run regardless of individual reuse configuration by setting the **regenerate_outputs** parameter when submitting the pipeline experiment:

```
pipeline_run = experiment.submit(train_pipeline, regenerate_outputs=True)
```