Rate Limiting

When using DocETL, you might have rate limits based on your usage tier with various API providers. To help manage these limits and prevent exceeding them, DocETL allows you to configure rate limits in your YAML configuration file.

Configuring Rate Limits

You can add rate limits to your YAML config by including a `rate_limits` key with specific configurations for different types of API calls. Here's an example of how to set up rate limits:

```yaml
rate_limits:
  embedding_call:
    - count: 1000
      per: 1
      unit: second
  llm_call:
    - count: 1
      per: 1
      unit: second
    - count: 10
      per: 5
      unit: hour
  llm_tokens:
    - count: 1000000
      per: 1
      unit: minute
```

Your YAML configuration should have a `rate_limits` key with the config as shown above. This example sets limits for embedding calls and language model (LLM) calls, with multiple rules for LLM calls to accommodate different time scales.

You can also use rate limits in the Python API, passing in a `rate_limits` dictionary when you initialize the `Pipeline` object.