vLLM + Prometheus/Grafana

This is a simple example that shows you how to connect vLLM metric logging to the Prometheus/Grafana stack. For this example, we launch Prometheus and Grafana via Docker. You can checkout other methods through [Prometheus](https://prometheus.io/) and [Grafana](https://grafana.com/) websites.

Install:

- [`docker`](https://docs.docker.com/engine/install/)
- [`docker compose`](https://docs.docker.com/compose/install/linux/#install-using-the-repository)

Launch

Prometheus metric logging is enabled by default in the OpenAI-compatible server. Launch via the entrypoint:

```bash
python3 -m vllm.entrypoints.openai.api_server \
    --model mistralai/Mistral-7B-v0.1 \
    --max-model-len 2048 \
    --disable-log-requests
```

Launch Prometheus and Grafana servers with `docker compose`:

```bash
docker compose up
```

Submit some sample requests to the server:

```bash
wget
https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered/resolve/main/ShareGPT_V3_unfiltered_cleaned_split.json

python3 ../../benchmarks/benchmark_serving.py \
    --model mistralai/Mistral-7B-v0.1 \
    --tokenizer mistralai/Mistral-7B-v0.1 \
    --endpoint /v1/completions \
    --dataset ShareGPT_V3_unfiltered_cleaned_split.json \
    --request-rate 3.0
```

Navigating to [`http://localhost:8000/metrics`](http://localhost:8000/metrics) will show the raw Prometheus metrics being exposed by vLLM.

Grafana Dashboard

Navigate to [`http://localhost:3000`](http://localhost:3000). Log in with the default username (`admin`) and password (`admin`).

Add Prometheus Data Source

Navigate to

[`http://localhost:3000/connections/datasources/new`](http://localhost:3000/connections/datasources/new) and select Prometheus.

On Prometheus configuration page, we need to add the `Prometheus Server URL` in `Connection`. For this setup, Grafana and Prometheus are running in separate containers, but Docker creates DNS name for each containers. You can just use `http://prometheus:9090`.

Click `Save & Test`. You should get a green check saying "Successfully queried the Prometheus API.".

Import Dashboard

Navigate to [`http://localhost:3000/dashboard/import`](http://localhost:3000/dashboard/import), upload `grafana.json`, and select the `prometheus` datasource. You should see a screen that looks like the following: