Online prediction on Vertex AI metrics

- Performance
    - Predictions/second
    - Prediction error percentage
    - Requests/second
    - Response codes.
    - Total latency duration
    - Model latency duration.
- Resource usage
    - Replica.
    - CPU usage
    - Memory usage
    - Network bytes sent
    - Network bytes received
    - Accelerator average duty cycle
    - Accelerator memory usage.

HF TGI metrics

- https://huggingface.co/docs/text-generation-inference/en/reference/metrics

tgi_batch_current_max_tokens	Maximum tokens for the current batch
tgi_batch_current_size	Current batch size
tgi_batch_decode_duration	Time spent decoding a batch per method (prefill or decode)
tgi_batch_filter_duration	Time spent filtering batches and sending generated tokens per method (prefill or decode)
tgi_batch_forward_duration	Batch forward duration per method (prefill or decode)
tgi_batch_inference_count	Inference calls per method (prefill or decode)
tgi_batch_inference_duration	Batch inference duration
tgi_batch_inference_success	Number of successful inference calls per method (prefill or decode)
tgi_batch_next_size	Batch size of the next batch
tgi_queue_size	Current queue size
tgi_request_count	Total number of requests
tgi_request_duration	Total time spent processing the request (e2e latency)
tgi_request_generated_tokens	Generated tokens per request
tgi_request_inference_duration	Request inference duration
tgi_request_input_length	Input token length per request
tgi_request_max_new_tokens	Maximum new tokens per request
tgi_request_mean_time_per_token_duration	Mean time per token per request (inter-token latency)
tgi_request_queue_duration	Time spent in the queue per request
tgi_request_skipped_tokens	Speculated tokens per request
tgi_request_success	Number of successful requests
tgi_request_validation_duration	Time spent validating the request

vLLM metrics

- https://docs.vllm.ai/en/stable/serving/metrics.html

vllm:num_requests_running	Number of requests currently running on GPU.
vllm:num_requests_waiting	Number of requests waiting to be processed.
vllm:lora_requests_info	Running stats on lora requests.
vllm:num_requests_swapped	Number of requests swapped to CPU.
vllm:gpu_cache_usage_perc	GPU KV-cache usage. 1 means 100 percent usage.
vllm:cpu_cache_usage_perc	CPU KV-cache usage. 1 means 100 percent usage.
vllm:cpu_prefix_cache_hit_rate	CPU prefix cache block hit rate.
vllm:gpu_prefix_cache_hit_rate	GPU prefix cache block hit rate.
vllm:num_preemptions_total	Cumulative number of preemption from the engine.
vllm:prompt_tokens_total	Number of prefill tokens processed.
vllm:generation_tokens_total	Number of generation tokens processed.
vllm:tokens_total	Number of prefill plus generation tokens processed.
vllm:iteration_tokens_total	Histogram of number of tokens per engine_step.
vllm:time_to_first_token_seconds	Histogram of time to first token in seconds.
vllm:time_per_output_token_seconds	Histogram of time per output token in seconds.
vllm:e2e_request_latency_seconds	Histogram of end to end request latency in seconds.
vllm:request_queue_time_seconds	Histogram of time spent in WAITING phase for request.
vllm:request_inference_time_seconds	Histogram of time spent in RUNNING phase for request.
vllm:request_prefill_time_seconds	Histogram of time spent in PREFILL phase for request.
vllm:request_decode_time_seconds	Histogram of time spent in DECODE phase for request.
vllm:time_in_queue_requests	Histogram of time the request spent in the queue in seconds.
vllm:model_forward_time_milliseconds	Histogram of time spent in the model forward pass in ms.
vllm:model_execute_time_milliseconds	Histogram of time spent in the model execution function in ms.
vllm:request_prompt_tokens	Number of prefill tokens processed.
vllm:request_generation_tokens	Number of generation tokens processed.

- vllm:request_max_num_generation_tokens Histogram of maximum number of requested generation tokens.
- vllm:request_params_n Histogram of the n request parameter.
- vllm:request_params_max_tokens Histogram of the max_tokens request parameter.
- vllm:request_success_total Count of successfully processed requests.
- vllm:spec_decode_draft_acceptance_rate Speulative token acceptance rate.
- vllm:spec_decode_efficiency Speculative decoding system efficiency.
- vllm:spec_decode_num_accepted_tokens_total Number of accepted tokens.
- vllm:spec_decode_num_draft_tokens_total Number of draft tokens.
- vllm:spec_decode_num_emitted_tokens_total Number of emitted tokens.
- vllm:avg_prompt_throughput_toks_per_s Average prefill throughput in tokens/s.
- vllm:avg_generation_throughput_toks_per_s Average generation throughput in tokens/s.