DataArtsFabric

Performance White Paper

Issue	01
Date	2025-07-18

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 Inference Performance White Paper

This document describes how to use the performance test platform to test the performance of the DataArts Fabric inference services and provides the test data report.

Test Environment

- Site: Huawei Cloud DataArts Fabric test environment
- Test time: November 30, 2024
- Inference services and related resources:

Inference Service Name	Model Type	Flavor	Compute (MU)	Instances
LLama-3-8B	LLAMA_3_8B	mu.llama3.8b	2	1
LLama-3-70B	LLAMA_3_70B	mu.llama3.70b	8	1
LLama-3.1-8B	LLAMA_3.1_8B	mu.llama3.1.8b	2	1
LLama-3.1-70B	LLAMA_3.1_70B	mu.llama3.1.70b	8	1
QWEN-2-72B	QWEN_2_72B	mu.qwen2.72b	8	1
GLM-4-9B	GLM_4_9B	mu.glm4.9b	2	1

Test Tools

Apache JMeter is used for the test. It is an open-source software used for performance testing. It can simulate servers and clients across multiple protocols, such as HTTP, FTP, and SMTP. Apache JMeter allows users to execute tests on web applications, database connections, and FTP servers. It supports custom and predefined scripts as well as distributed testing to simulate different loads.

JMeter depends on the JDK. Ensure that the JDK has been installed on the current computer and the environment variables have been configured. To download Apache JMeter, see **Download Apache JMeter**.

Test Methods

1. Log in to the DataArts Fabric console, select the target workspace, and click **Access Workspace**.

   If there is no available workspace, click **Create Workspace** to create one.
2. In the navigation pane, choose **Resources and Assets** > **Model**. In the upper right corner of the page, click **Create Model**. Enter basic model information, including the name and description, select the OBS path of the model file, and click **Create Now**.
3. In the navigation pane, choose **Resources and Assets** > **Inference Endpoint**. In the upper right corner of the page, click **Create Inference Endpoint**. Enter the endpoint name, resource specifications, and quantity, and click **Create Now**.
4. In the navigation pane, choose **Development and Production** > **Inference Services**. In the upper right corner of the page, click **Create Inference Service**. On the displayed page, enter the basic information such as the name and description of the inference service, select the inference endpoint and mode, and configure the minimum and maximum values of resources. After the configuration is complete, click **Create Now**.

   Model Type can be set to **My models** or **Public models**.
5. In the navigation pane, choose **Development and Production** > **Playgrounds**, and select the target inference service.
6. Use the test tool to perform concurrent inference.

Test Metrics

Request Per Minute (RPM) is an important metric for measuring system performance. It indicates the number of requests that can be processed by the system per minute.

Test Data

- **Data 1:**

  It is a short question and **max_tokens** is **256**.

```
{
  "type":"ChatCompletionRequest",
  "messages": [
    {
      "role": "user",
      "content": "What is LLM? What is different between different LLM?"
    }
  ],
  "max_tokens": 256,
  "stream":true
}
```

- **Data 2:**

  It is a medium-length question and **max_tokens** is **2048**.

```
{
  "type":"ChatCompletionRequest",
```



```
 "messages": [
  {
    "role": "user",
     "content": "Please write a novel and the word size should more than 2000,
requirements:1.Setting: Village, ancient forest, bustling city, forgotten island, futuristic metropolis, enchanted castle. 2.Protagonist: Orphaned child, disgraced knight, brilliant scientist, secret agent, reclusive artist, adventurous explorer.3.Antagonist: Shadowy figure, corrupt politician, malevolent sorcerer, rival adventurer, robotic overlord, vengeful ghost.4.Conflict: Quest for revenge, search for a lost artifact, battle for power, love triangle, struggle against fate, resistance against tyranny.2000-Word Requirement Guideline: Writing a 2000-word novel can be challenging, but it's also a great way to hone your writing skills and tell a concise, compelling story. Here are some tips to help you meet the word count while maintaining quality:1.Outline Your Story: Before you start writing, take some time to outline your story. Decide on your main plot points, character arcs, and the overall theme you want to explore. This will help you stay focused and ensure that your story has a clear structure.2.Focus on Key Scenes: With a limited word count, you need to prioritize the most important scenes. Focus on the scenes that drive the plot forward, reveal character development, and create tension. Avoid unnecessary descriptions and subplots that don't contribute to the overall story.3.Show, Don't Tell: Use vivid, sensory details to bring your story to life. Instead of telling readers what's happening, show them through dialogue, actions, and internal monologue. This will make your writing more engaging and help you use your words more effectively.4.Edit Ruthlessly: As you write, be prepared to cut out anything that doesn't add value to your story. This might include redundant descriptions, unnecessary characters, or scenes that don't move the plot forward. Remember, every word should count."
  }
 ],
 "max_tokens": 2048,
 "stream":true
}
```

Test Results

- The test is based on **data 1** with the concurrency of 64. The following table shows the test results.

Table 1-1 Test results of data 1

Model Name	Test Type	Concurrency	maxtoken	Test Time (s)	Success Rate	StatusCode	Total Requests	Average Latency (ms)	TP99 Latency (ms)	TPS	RPM
LLama-3-8B	Concurrency	64	256	300	100%	200	2090	9231	32615	7.01	420.6
LLama-3-70B	Concurrency	64	256	300	100%	200	420	43072	68082	1.79	107.4

Model Name	Test Type	Concurrency	maxtoken	Test Time (s)	Success Rate	StatusCode	Total Requests	Average Latency (ms)	TP99 Latency (ms)	TPS	RPM
LLama-3.1-8B	Concurrency	64	256	300	100%	200	960	20453	51011	3.27	196.2
LLama-3.1-70B	Concurrency	64	256	300	100%	200	679	29975	44826	2.29	137.4
QWEN-2-72B	Concurrency	64	256	300	100%	200	8706	2212	4915	29.02	1741.2
GLM-4-9B	Concurrency	64	256	300	100%	200	578	35655	66167	1.93	115.8

- The test is based on **data 2** with the concurrency of 16. The table below shows the test results.

  NOTE

  The response duration of an inference request varies depending on the input token, output token, and parameters of the request. The values in the following table are for reference only. The actual values may vary greatly.

Table 1-2 Test results of data 2

Model Name	Test Type	Concurrency	maxtoken	Test Time (s)	Success Rate	StatusCode	Total Requests	Average Latency (ms)	TP99 Latency (ms)	TPS	RPM
LLama-3-8B	Concurrency	16	2048	300	100%	200	96	51636	96797	0.32	19.2
LLama-3-70B	Concurrency	16	2048	300	100%	200	82	64296	74727	0.27	16.2
LLama-3.1-8B	Concurrency	16	2048	300	100%	200	192	26072	38645	0.68	40.8
LLama-3.1-70B	Concurrency	16	2048	300	100%	200	64	85552	103198	0.22	13.2
QWEN-2-72B	Concurrency	16	2048	600	100%	200	197	51260	75031	0.33	19.8
GLM-4-9B	Concurrency	16	2048	300	100%	200	137	37630	52302	0.46	27.6

