Endpoints

Endpoints in our platform provide a mechanism for creating services that accept predictions via a designated endpoint. These services, based on existing platform versions, leverage Cairo under the hood to ensure provable inferences. Using the CLI, users can effortlessly deploy and retrieve information about these machine learning services.

Deploying a model as an endpoint

To deploy a model, you must first have a version of that model. If you have not yet created a version, please refer to theversions documentation.

To create a new service, users can employ thedeploy command. This command facilitates the deployment of a machine learning service ready to accept predictions at the/cairo_run endpoint, providing a straightforward method for deploying and utilizing machine learning capabilities.

```

Copy

    giza endpoints deploy --model-id 1 --version-id 1 model.sierra ▰▰▰▰▰▱▱ Creating endpoint! [giza][2024-02-07 12:31:02.498] Endpoint is successful ✅ [giza][2024-02-07 12:31:02.501] Endpoint created with id -> 1 ✅ [giza][2024-02-07 12:31:02.502] Endpoint created with endpoint URL: https://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app

```

If a model is fully compatible the sierra file is not needed and can be deployed without using it in the command:

```

Copy

    giza endpoints deploy --model-id 1 --version-id 1 ▰▰▰▰▰▱▱ Creating endpoint! [giza][2024-02-07 12:31:02.498] Endpoint is successful ✅ [giza][2024-02-07 12:31:02.501] Endpoint created with id -> 1 ✅ [giza][2024-02-07 12:31:02.502] Endpoint created with endpoint URL: https://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app

```

For a partially compatible model the sierra file must be provided, if not an error will be shown.

Example request

Now our service is ready to accept predictions at the provided endpoint URL. To test this, we can use thecurl command to send a POST request to the endpoint with a sample input.

```

Copy

    curl -X POST https://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app/cairo_run \ -H "Content-Type: application/json" \ -d '{ "args": "[\"2\", \"2\", \"2\", \"4\", \"1\", \"2\", \"3\", \"4\"]" }' | jq { "result": [ { "value": { "val": [ 1701737587, 1919382893, 1869750369, 1852252262, 1864395887, 1948284015, 1231974517 ] } } ] }

```

Listing endpoints

The list command is designed to retrieve information about all existing endpoints. It provides an overview of the deployed machine learning services, allowing users to monitor and manage multiple endpoints efficiently.

```

Copy giza endpoints list [giza][2024-01-17 17:19:00.631] Listing endpoints ✅ [ { "id": 1, "status": "COMPLETED", "uri": "https://deployment-gizabrain-1-1-53427f44-dagsgas-ew.a.run.app", "size": "S", "service_name": "deployment-gizabrain-1-1-53427f44", "model_id": 1, "version_id": 1, "is_active": true }, { "id": 2, "status": "COMPLETED", "uri": "https://deployment-gizabrain-1-2-53427f44-dagsgas-ew.a.run.app", "size": "S", "service_name": "deployment-gizabrain-1-2-53427f44", "model_id": 1, "version_id": 2, "is_active": false } ]

```

Executing this command will display a list of all current endpoints, including relevant details such as service names, version numbers, and endpoint status.

To list only active endpoints you can use the flag--only-active/-a so only active ones are shown.

Retrieving an endpoint

For retrieving detailed information about a specific endpoint, users can utilize the get command. This command allows users to query and view specific details of a single endpoint, providing insights into the configuration, status, and other pertinent information.

```

Copy

    giza endpoints get --endpoint-id 1 { "id": 1, "status": "COMPLETED", "uri": "https://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app", "size": "S", "service_name": "deployment-gizabrain-38-1-53427f44", "model_id": 38, "version_id": 1, "is_active": true }

```

Delete an endpoint

For deleting an endpoint, users can use the delete command. This command facilitates the removal of a machine learning service, allowing users to manage and maintain their deployed services efficiently.

```

Copy

    giza endpoints delete --endpoint-id 1 [giza][2024-03-06 18:10:22.548] Deleting endpoint 1 ✅ [giza][2024-03-06 18:10:22.830] Endpoint 1 deleted ✅

```

The endpoints are not fully deleted, so you can still access the underlying proofs generated by them.

Previous Workspaces Next Cairo

Last updated4 days ago