Deploy

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/predict endpoint, providing a straightforward method for deploying and using machine learning capabilities as an API endpoint. As we are usingEZKL we need to add--framework EZKL (or-f EZKL for short) to the command:

```

Copy

    giza endpoints deploy --model-id 1 --version-id 1 --framework EZKL ▰▰▰▰▰▱▱Creatingendpoint! [giza][2024-02-0712:31:02.498] Endpoint is successful ✅ [giza][2024-02-0712:31:02.501] Endpoint created with id ->1 ✅ [giza][2024-02-0712:31:02.502] Endpoint created with endpoint URL: https://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app

```

Example Request

Now the model is available to generate predictions and generate proofs of those predictions. The schema of the data is the same as used to create theinput.json needed to create version, for a linear regression it would be:

```

Copy { "input_data":[ [ 0.12177091836929321, 0.7048522233963013 ] ] }

```

To execute a prediction usingcURL :

```

Copy curlhttps://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app/predict\ -H"Content-Type: application/json"-d'{ "input_data": [ [ 0.12177091836929321, 0.7048522233963013 ] ] }'|jq

```

This yields the following response:

```

Copy { "prediction":[ [ 4.53125 ] ], "request_id":"d0564505755944b8bef9292d980f3e27" }

```

There is an extra args,job_size , that can be used in each request to specify the size of the proving job so it has more CPU and memory available to generate the proof. An example:

```

Copy curlhttps://deployment-gizabrain-38-1-53427f44-dagsgas-ew.a.run.app/predict\ -H"Content-Type: application/json"-d'{ "input_data": [ [ 0.12177091836929321, 0.7048522233963013 ] ], "job_size": "M" }'|jq

```

Available sizes areS ,M ,L, andXL , each with different usage limits.

List the proving jobs for an endpoint

To list the proving jobs for an endpoint, we can use thelist-jobs command available for the endpoints. This command will return a list of all the proving jobs for the endpoint with therequest_id for easier tracking.

```

Copy giza endpoints list-jobs --endpoint-id 1 [giza][2024-03-06 18:13:50.485] Getting jobs from endpoint 1 ✅ [ { "id": 1, "job_name": "proof-ezkl-20240306-979342e7", "size": "S", "status": "Completed", "elapsed_time": 120., "created_date": "2024-03-06T16:12:31.295958", "last_update": "2024-03-06T16:14:29.952678", "request_id":
```

"979342e7b94641f0a260c1997d9ccfee" }, { "id": 2, "job_name": "proof-ezkl-20240306-f6559749", "size": "S", "status": "COMPLETED", "elapsed_time": 120.0, "created_date": "2024-03-06T16:43:27.531250", "last_update": "2024-03-06T16:45:17.272684", "request_id": "f655974900d8479c9bb662a060bc1365" }]

```

List the proofs for an endpoint

To list the proofs for an endpoint, we can use thelist-proofs command available for the endpoints. This command will return a list of all the proofs for the endpoint with therequest_id for easier tracking.

```

Copy giza endpoints list-proofs --endpoint-id 1 [giza][2024-03-06 18:15:23.146] Getting proofs from endpoint 1 ✅ [{ "id": 1, "job_id": 1, "metrics": { "proving_time": 0.03023695945739746 }, "created_date": "2024-03-06T16:44:46.196186", "request_id": "979342e7b94641f0a260c1997d9ccfee" }, { "id": 1, "job_id": 2, "metrics": { "proving_time": 0.07637895945739746 }, "created_date": "2024-03-06T16:44:46.196186", "request_id": "f655974900d8479c9bb662a060bc1365" }]

```

Download the proof

We can download the proof using thedownload-proof command available for the endpoints:

```

Copy ❯gizaendpointsdownload-proof--proof-id"d0564505755944b8bef9292d980f3e27" [giza][2024-02-2015:40:48.560] Getting proof from endpoint 1 ✅ [giza][2024-02-2015:40:49.288] Proof downloaded to zk.proof ✅

```

Theproof id used is therequest_id returned in the response.

Previous Transpile Next Prove

Last updated1 day ago
```