



**Developing Open LLM
applications with**



Apache OpenServerless

Lesson 1

First Steps

Agenda

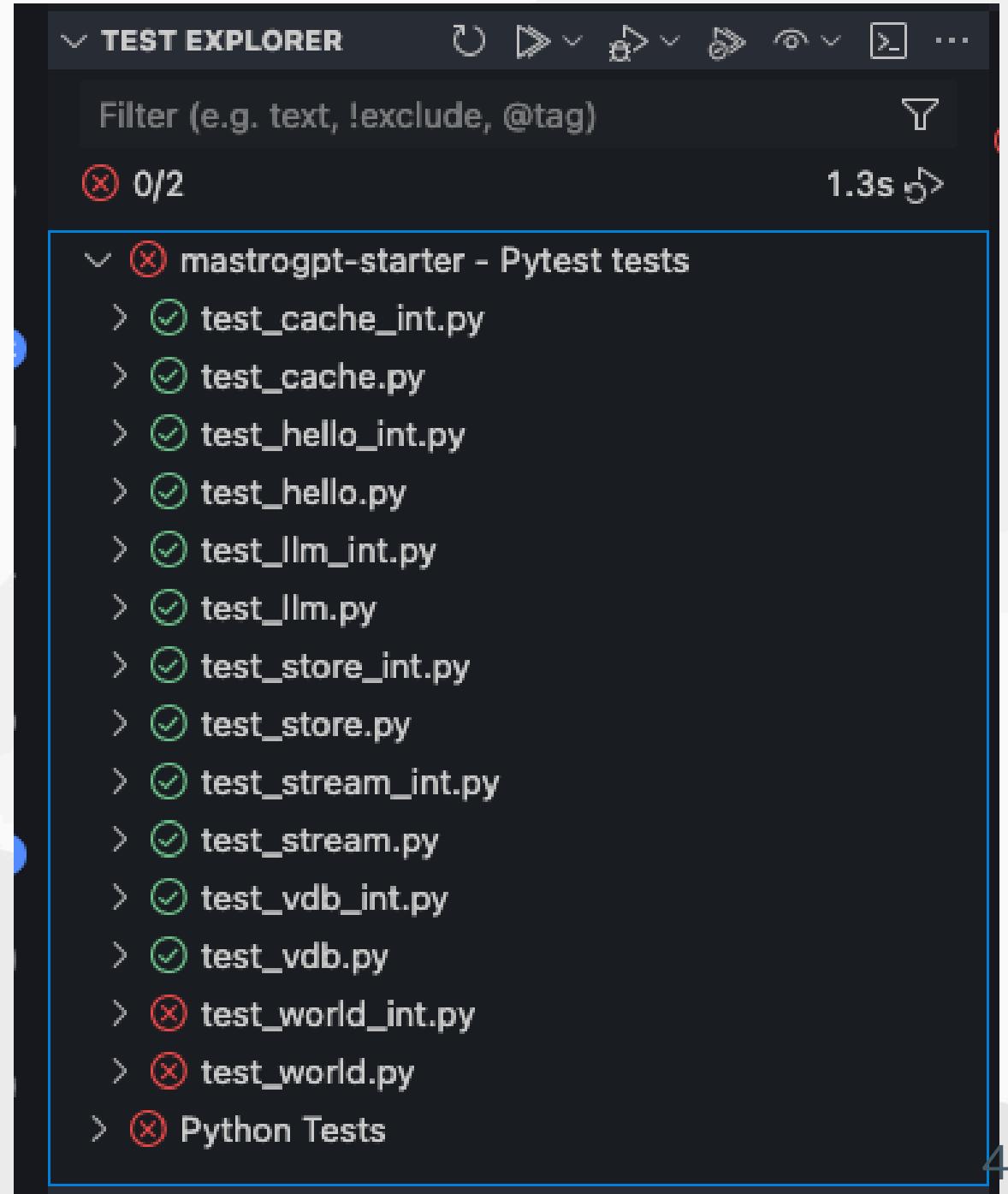
- Integrated Services
- Examples: the `hello`s
- CLI tools
- Exercise: reverse
- About Nuvolaris
- What is next?



Integrated Services

hello package

- Collection samples for all the services
- Launching the tests verifies all the services
- Also useful for interacting and debugging
- Services are all both local and remote



Exercise: fixing a failing test (trivial bug)

- Search for the `TODO:` string
- Investigate why a test is failing
- Fix it and run the unit test
- Deploy and run the integration test

The screenshot shows a code editor interface with a search results panel on the left and a code editor tab on the right.

Search Results Panel (Left):

- SEARCH
- TODO:** (highlighted)
- Replace
- 1 result in 1 file - Open in editor
- world.py packages/hello/world
- # TODO: the expected output is "Hello, ..."

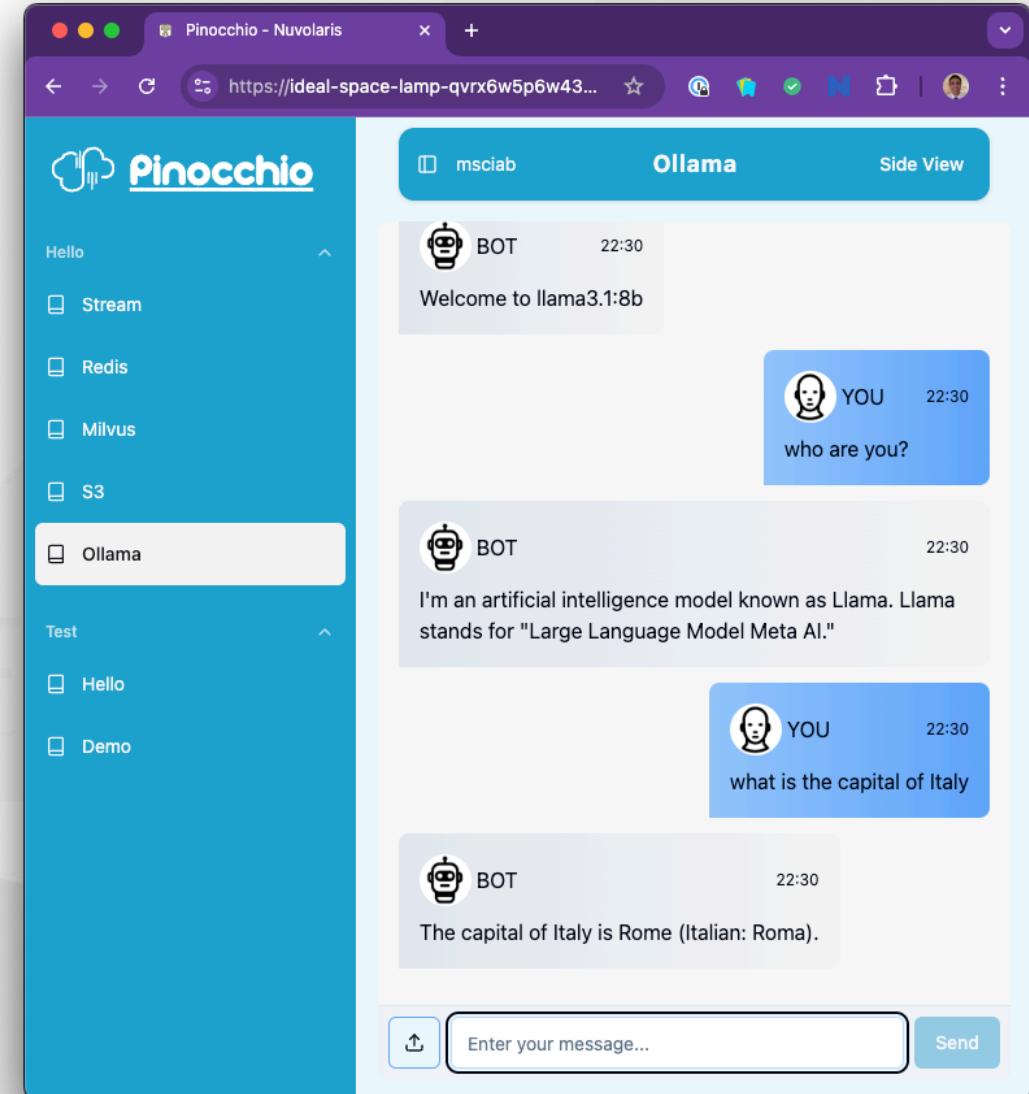
Code Editor Tab (Right):

```
packages > hello > world > world.py > world
1 def world(args):
2     # TODO: the expected output is "Hello, <name>" - fix the bug
3     name = args.get("input", "world")
4     return { "output": f"Hi, {name}" }
```

Examples: the "hello"s

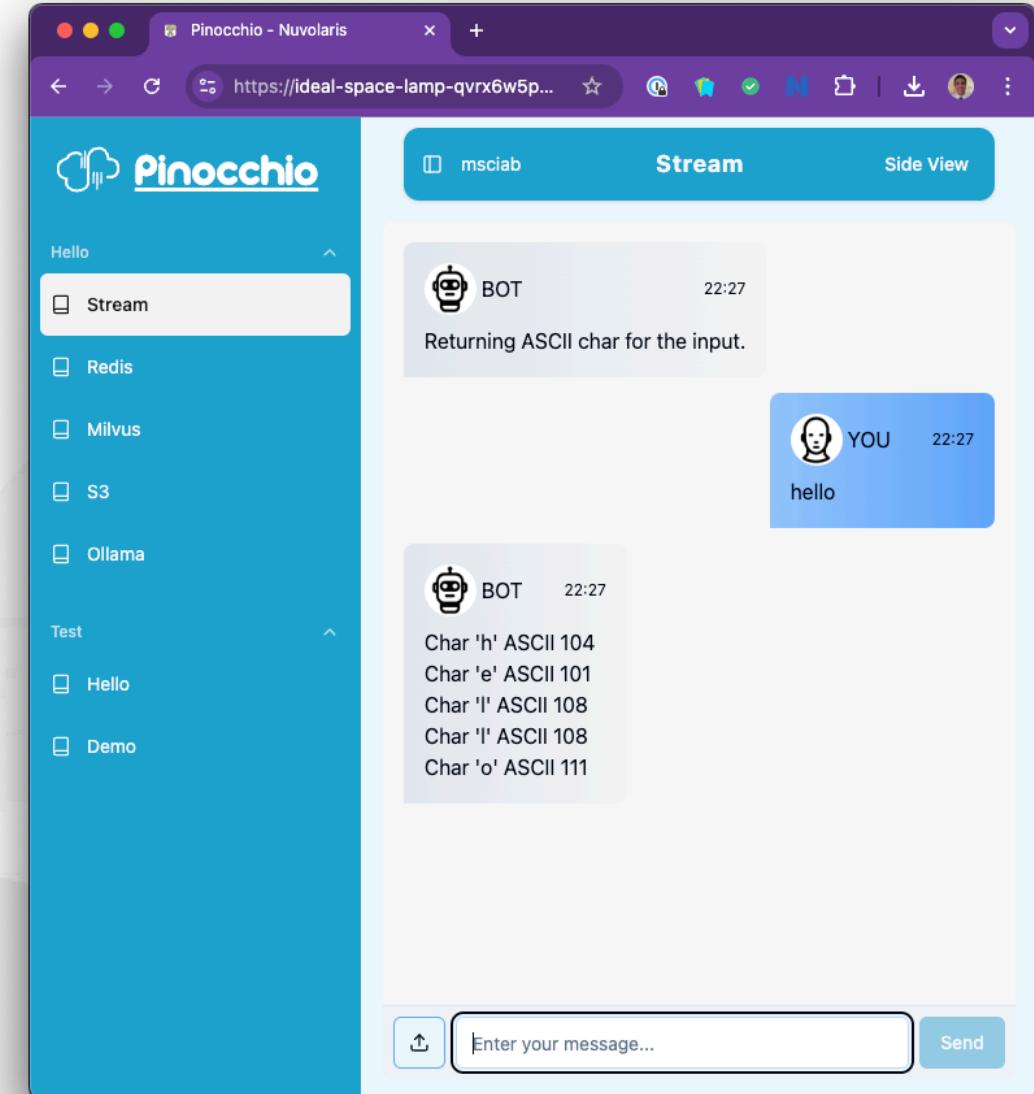
hello/llm

- Access to the LLM
- Ollama with
 - llama3.1:8b
 - powerful small model
 - llama3.2-vision:11b
 - with vision capabilities
 - mxbai-embed-large:latest
 - embedding model



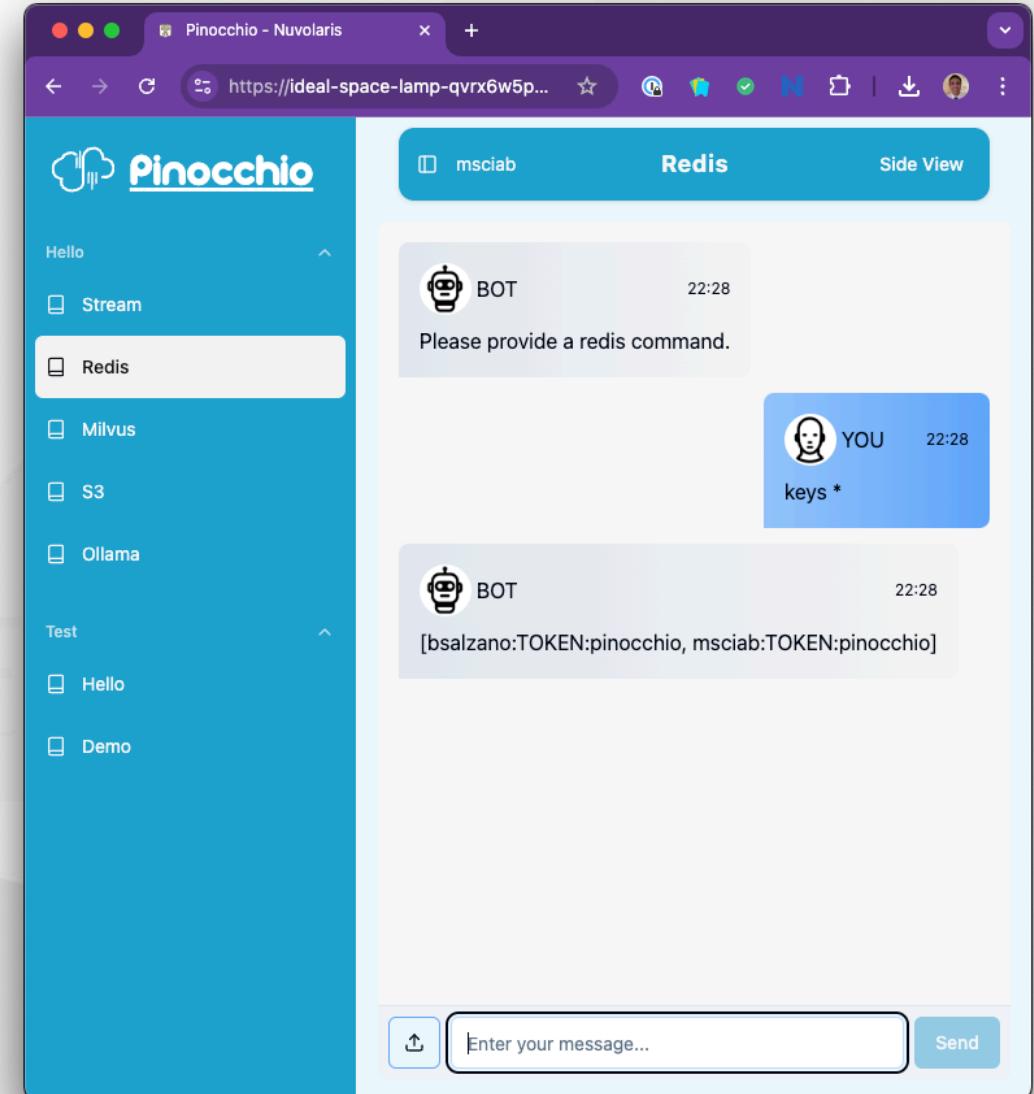
hello/stream

- An example of the streamer
- Return the ASCII of each character
- Stream the input in 1 second interval



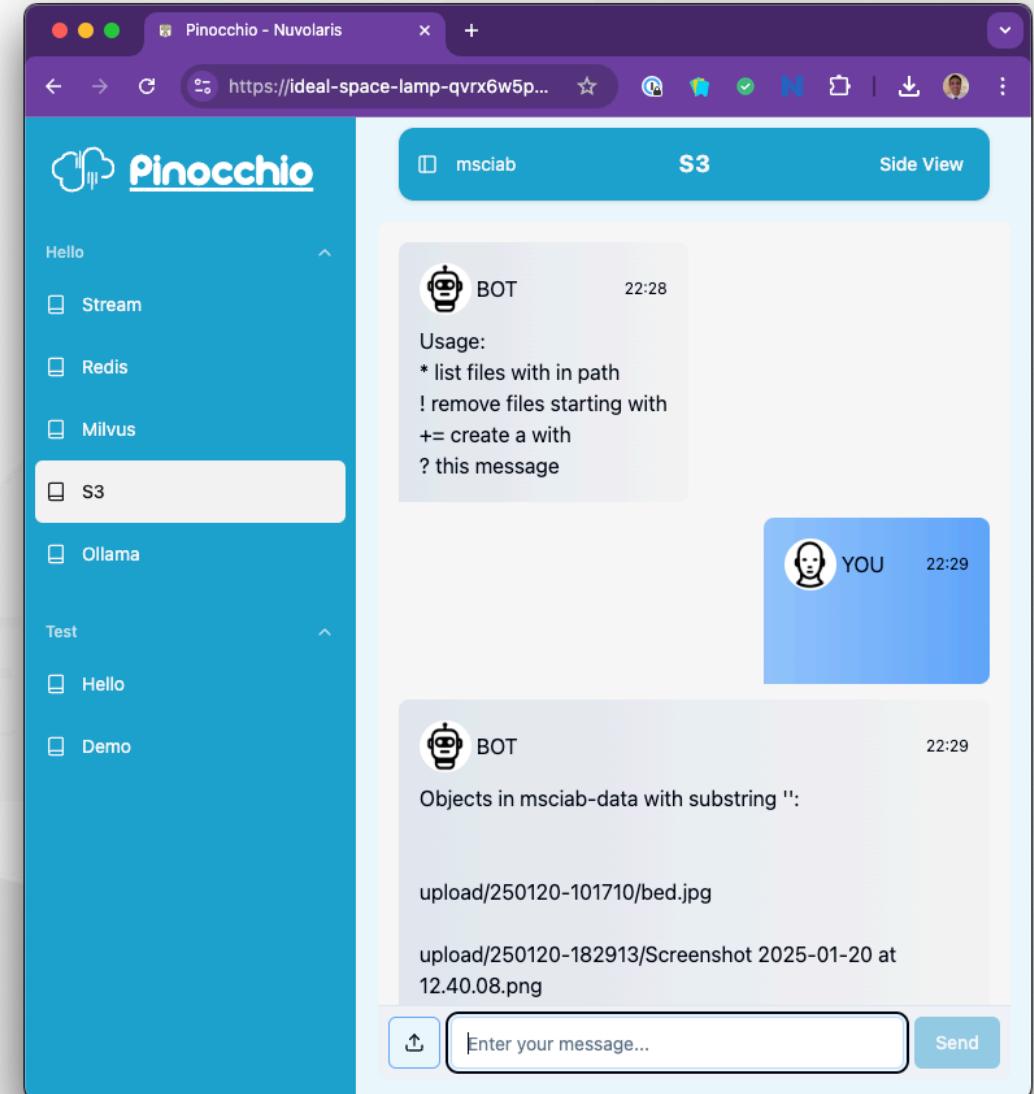
hello/cache

- Talk directly with REDIS
- Useful for debugging
- Remember there is a required **PREFIX** for the keys!
 - <username>:



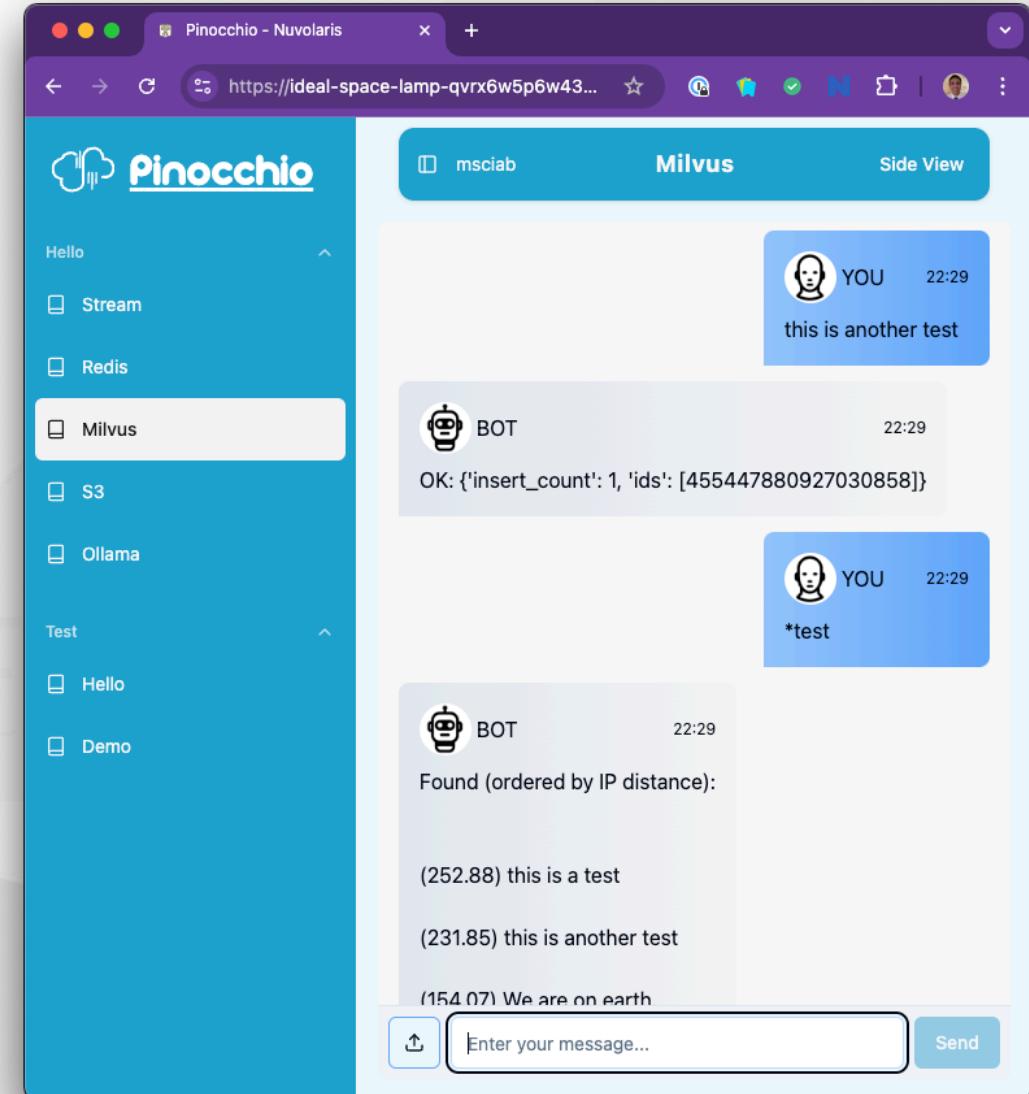
hello/store

- S3 Storage
- Files are uploaded here
- Simple commands:
 - *<prefix> list content by prefix
 - !<prefix> remove content by prefix
 - +<file>=<content> create a file on the fly



hello/vdb

- Milvus Vector Database
- Store what you type
- Simple commands:
 - *<search> vector search
 - !<word> remove entries containing a word



CLI Tools

ops docs on <https://openserverless.apache.org>

- It is also self-documenting:

```
ops                      # main help message  
ops -h                   # list embedded tools  
ops -t                   # list tasks
```

We will use mostly:

- ops ide support
- ops ai A.I. oriented plugin

ops essentials

- basics commands to manage actions

```
ops action list
ops action create reverse lessons/reverse.py
ops invoke reverse
ops invoke reverse input=hello
ops url reverse
curl https://openserverless.dev/api/v1/namespaces/msciab/actions/reverse
ops action update reverse lessons/reverse.py --web true
ops url reverse
curl https://openserverless.dev/api/v1/web/msciab/default/reverse
curl "https://openserverless.dev/api/v1/web/msciab/default/reverse?input=hello"
ops action delete reverse
ops action list
```

ops ide essentials

- manages packaging and hot-reload

```
ops ide                      # support tools main subcommand
ops ide login                 # login to one openserverless instance
ops ide deploy                # package and deploy all the actions
ops ide deploy hello/llm     # package and deploy one action
ops ide devel                 # incremental development mode
ops ide clean                  # clean temporary files
```

ops ai essential

- our AI-oriented plugin

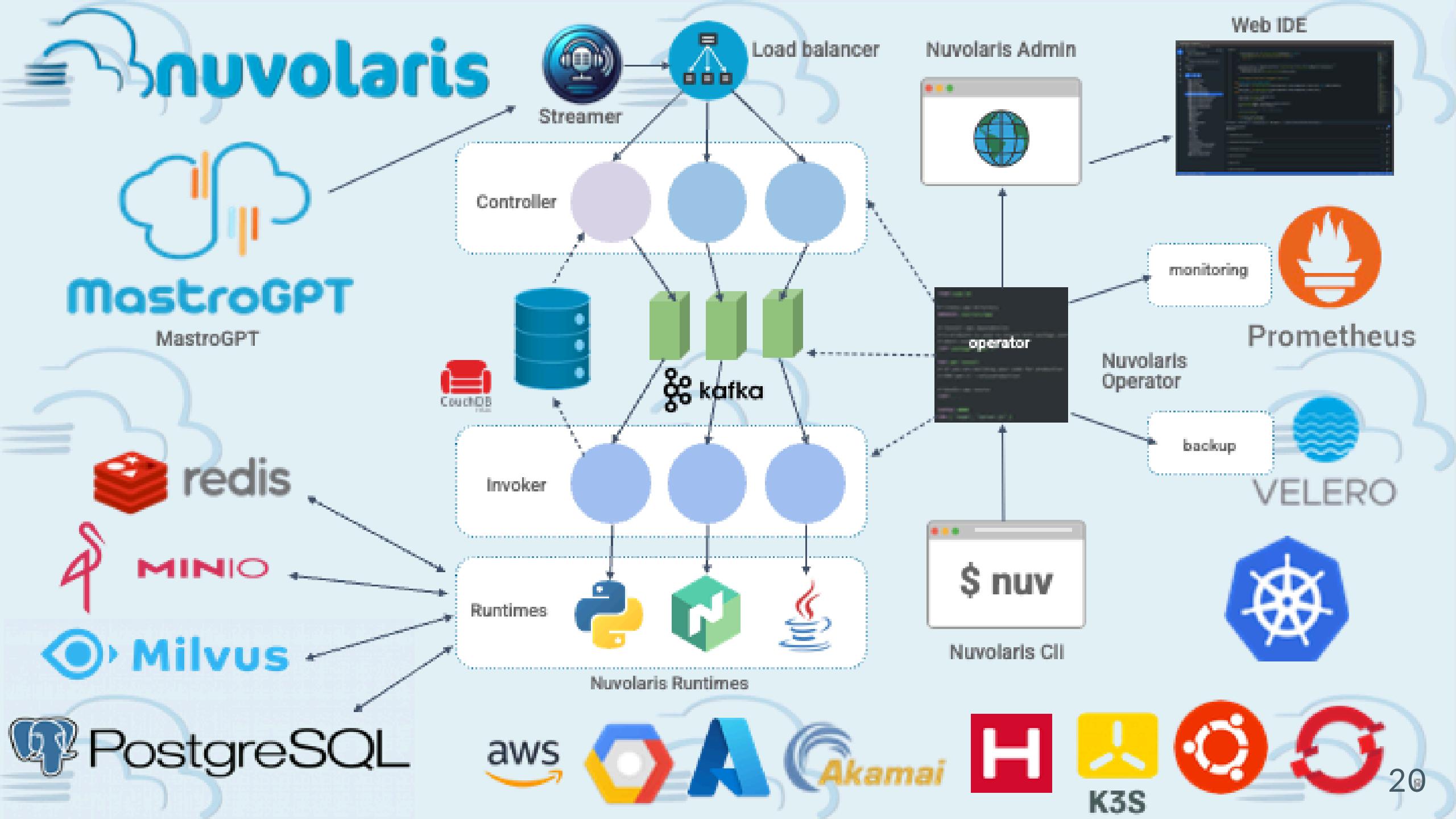
```
ops ai                      # help
ops ai lesson                # download lessons and solutions
ops ai user                  # update users
ops ai chat                  # command line chat
ops ai cli                   # the Python REPL
ops ai new                   # create a new service
```

Exercise: reverse

Exercise: implement a reverse chat

- `ops ai new reverse msciab`
- implement the code to a reverse functions
 - read input, return output
 - if empty input, return usage
- `ops ide deploy msciab/reverse`
- Add the service to `packages/mastrogpt/index/90-Tests.json`
- Use it in Pinocchio

About Nuvolaris





Pricelist 2025

Purpose	Name	Requirements	Price/month	Price/year	Hosted	VCPU	HA	GPU
Development	Nuvolaris VM	1 VM 16GB 4VCPU	€100.00	€1,000.00	yes	shared	no	no
Development	Nuvolaris Dedicated Server	1 Server 64GB 12VCPU (3vm)	€200.00	€2,000.00	yes	dedicated	no	no
Development	Nuvolaris Dedicated GPU Server	1 Server 64GB 12VCPU 20GPU (3vm)	€500.00	€5,000.00	yes	dedicated	yes	yes
Production	Nuvolaris Cluster	3 VM 16GB 4VCPU +1 LB	€300.00	€3,000.00	yes	shared	yes	no
Production	Nuvolaris Cluster Pro	6 VM 16GB 4VCPU +1 LB	€600.00	€6,000.00	yes	shared	yes	no
Production	Nuvolaris Cluster Additional VM	1 VM 16GB 4VCPU	€100.00	€1,000.00	yes	shared	yes	no
Production	Nuvolaris Dedicated Cluster	3 Server 64GB 12VCPU (6vm) + 1LB	€1,000.00	€10,000.00	yes	dedicated	yes	no
Production	Nuvolaris Dedicated GPU Cluster	3 Server 64GB 12VCPU 20GPU (6vm) + 1LB	€2,000.00	€20,000.00	yes	dedicated	yes	yes
Production	Nuvolaris Cluster Additional Server	1 Server 64GB 12VCPU (2vm)	€200.00	€2,000.00	yes	dedicated	yes	no
Production	Nuvolaris Cluster Additional GPU Server	1 Server 64GB 12VCPU 20GPU (2vm)	€500.00	€5,000.00	yes	dedicated	yes	yes
Development	Nuvolaris Your VM	1 VM 16GB 4VCPU	€75.00	€750.00	no	yours	no	no
Development	Nuvolaris Your Dedicated Server	1 Server 64GB 12VCPU (3vm)	€150.00	€1,500.00	no	yours	no	no
Development	Nuvolaris Your Dedicated GPU Server	1 Server 64GB 12CPU 20GPU (6vm)	€300.00	€3,000.00	no	yours	yes	yes
Production	Nuvolaris Your Cluster	3 VM 16GB 4CPU + 1 LB	€200.00	€2,000.00	no	yours	yes	no
Production	Nuvolaris Your Cluster Pro	6 VM 16GB 4CPU +1 LB	€400.00	€4,000.00	no	yours	yes	no
Production	Nuvolaris Your Cluster Additional VM	1 VM 16GB 4VCPU	€75.00	€750.00	no	yours	yes	no
Production	Your Dedicated Cluster	3 Server 64GB 12VCPU (6vm) + 1LB	€900.00	€9,000.00	no	yours	yes	no
Production	Your Dedicated GPU Cluster	3 Server 64GB 12VCPU 20GPU (6vm) + 1LB	€1,200.00	€12,000.00	no	yours	yes	yes
Production	Nuvolaris Your Dedicated Additional Server	1 Server 64GB 12VCPU (3vm)	€150.00	€1,500.00	no	yours	yes	no
Production	Nuvolaris Your Dedicated Additional GPU Server	1 Server 64GB 12CPU 20GPU (6vm)	€300.00	€3,000.00	no	yours	yes	yes

What is Next?

Lesson 2 - Streaming Chat

Implementing an LLM chat with streaming support

More lessons

- Lesson 3: Form Support
- Lesson 4: Building an Assistant
- Lesson 5: Vision Support
- Lesson 6: VectorDB
- Lesson 7: Building a RAG