QCon by InfoQ

LLM TRAINING

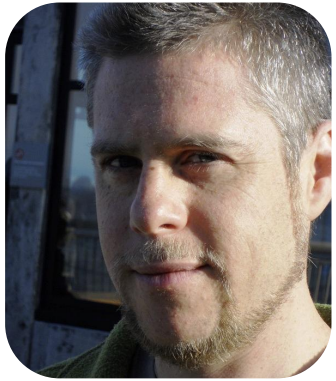
Build and Productionize LLM-Powered Applications with Ray & Anyscale

RAY

anyscale

Meet the tutorial team!

Marwan
marwan@anyscale.com



Adam
adamb@anyscale.com

Kamil
kamil@anyscale.com

📋 The Plan

Here’s what to expect today.

Today’s agenda.

- Retrieval-augmented generation from first principles
- Why host open source LLMs with Ray?
- Accelerate AI inference for higher speed, lower cost
- Combine Vector DBs + Ray for semantic search at scale
- Productionize an end-to-end LLM application with Ray

✅ Tech check.

Key ingredients:

- Anyscale Workspace (most important) - login here first!

- Pinecone free tier token (optional)
  - sign up at app.pinecone.io
  - place in **`pinecone.txt`** next to the notebooks

- Optional: if you have an OpenAI API key, place it in **`openai.txt`** next to the notebooks (only used in 1 demo)

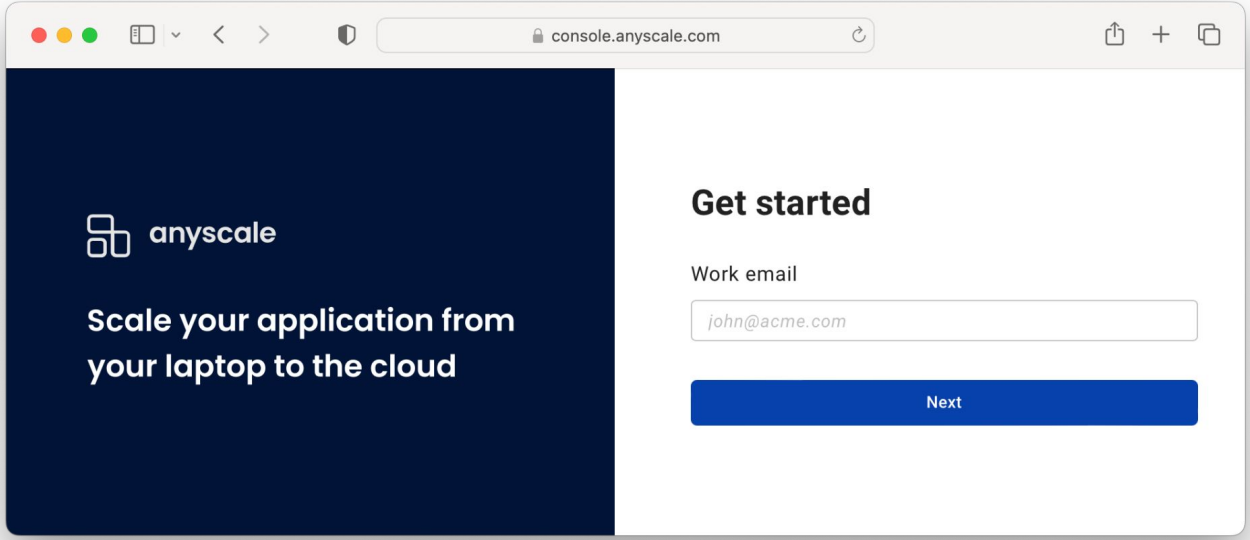
✅ Tech check.

Accessing Anyscale clusters.

- All work will be in Anyscale provisioned clusters.
- Our GitHub repo will be mounted automatically.
- Access begins now.
  - Check your email for login information.
  - Step-by-step instructions to follow.

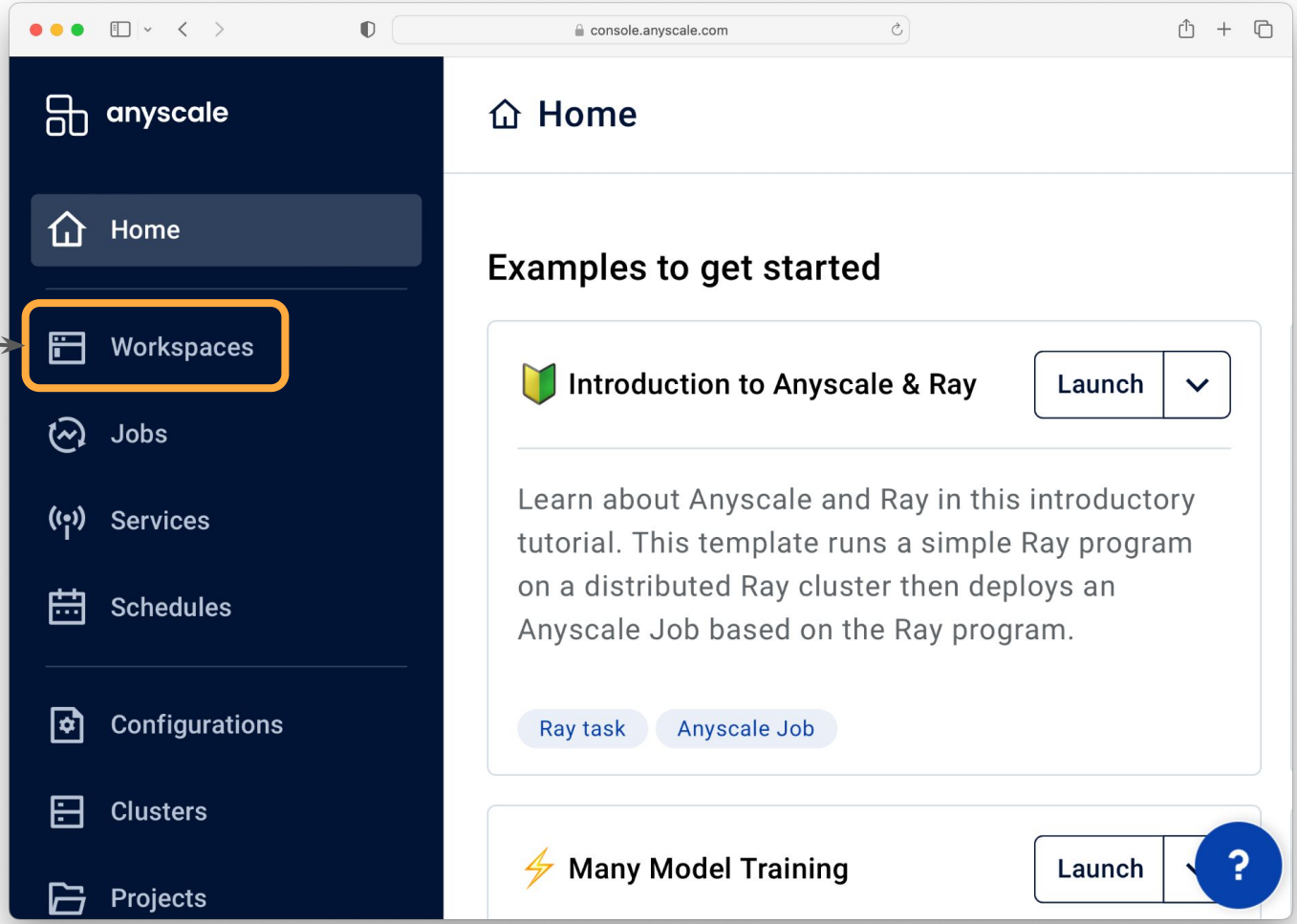
Anyscale login

Link to Anyscale cluster: console.anyscale.com

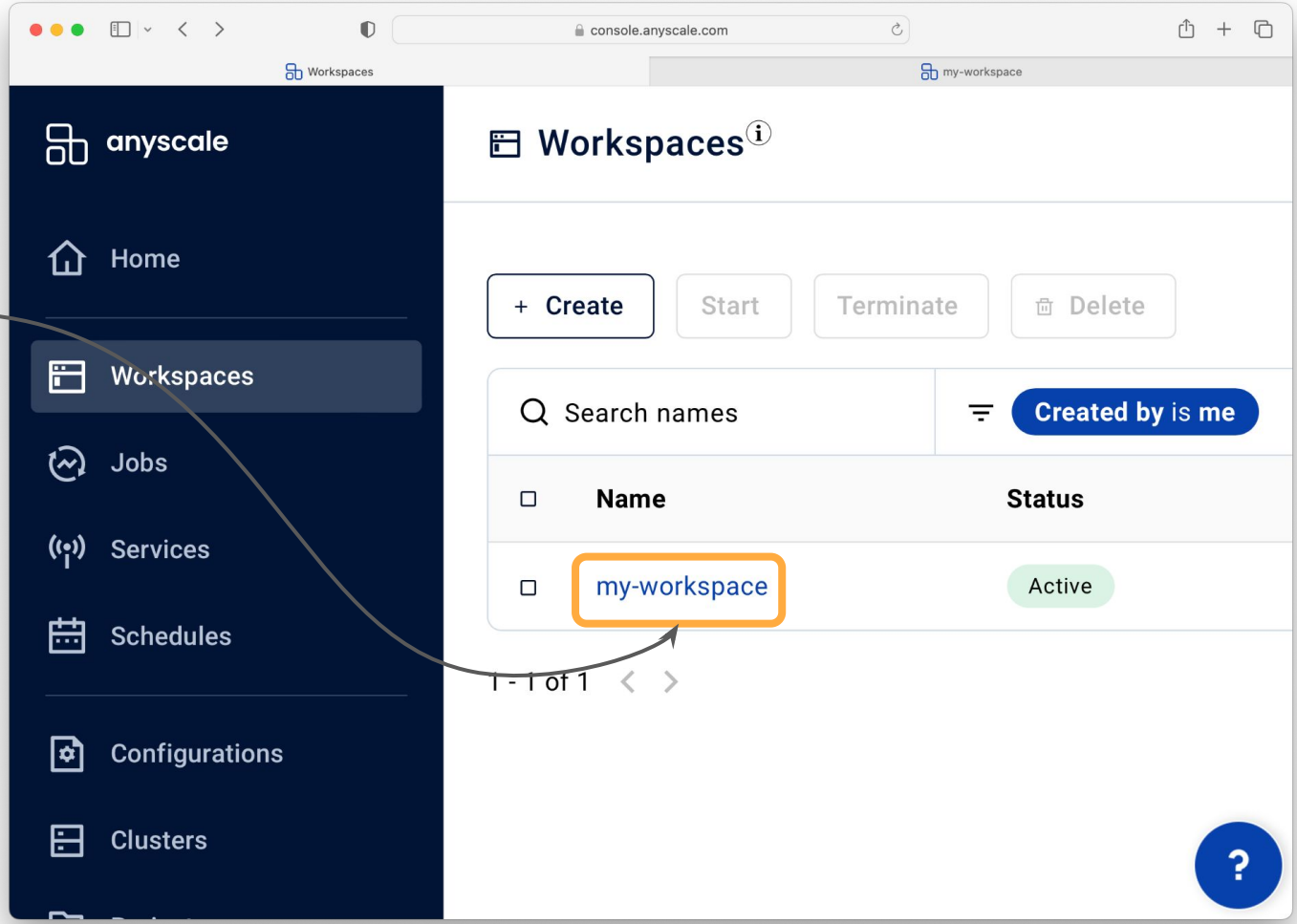


Enter the **unique credentials** sent to your email!

1. Select Workspaces



2. Select Your Workspace



3. Click on Jupyter icon

console.anyscale.com

anyscale

Home

Workspaces

Jobs

Services

Schedules

Configurations

Clusters

Projects

Emmy

Help

Feedback

m...workspace Active (Ray)

Terminate

Tools

About | Files | Terminal | Logs | Serve deployments

Status
Active (Ray)

Resources

Access
Everyone in your org

Created
Sep 7, 2023 at 4:26:50 PM, by emmy+education@anyscale.com

Cluster environment
summit:9

Compute config
ray-summit-2023-gc

Network access
Public with auth token

Job submissions
None

Ports
36075

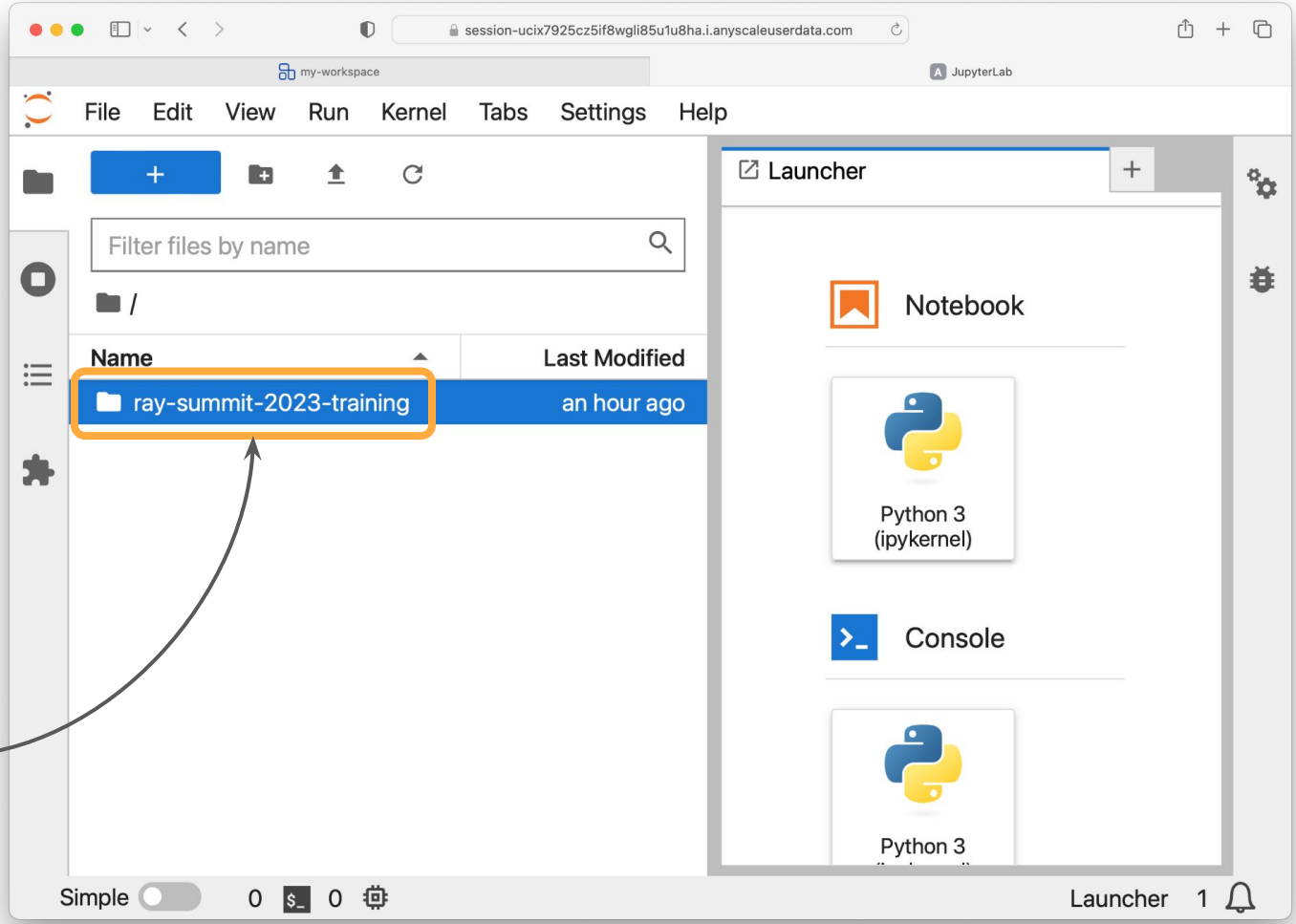
README

Workspaces

A Workspace is a fully managed development environment focused on developer productivity. W enable ML practitioners and ML platform developers to quickly build distributed Ray applications from research to development to production easily, all within a single environment.

Workspaces provide a remote experience for programming your cluster while working with JupyterLab notebooks or Visual Studio Code.

4. Find the content for your class here.



☕ Time for a Break!

15 minutes.

🍎 Today we learned...

🚂 The principles behind practical LLM RAG apps

🔬 How to build and orchestrate RAG with Ray

🔍 Why Ray simplifies LLM apps in prod, at scale

More Resources

For further exploration with Ray, Anyscale, and LLMs.

🍏 Sneak Peek: Self-Paced Ray & Anyscale Education

📧 Online at training.anyscale.com

🔍 Preview special technical content releases from the whole team!

Reading list.

Ray Education GitHub

Access bonus notebooks and scripts about Ray.

Ray documentation

API references and user guides.

Anyscale Blogs

Real world use cases and announcements.

YouTube Tutorials

Video walkthroughs about learning LLMs with Ray.

👩‍💻 Connect with the community.

👋 Join the community

Attend events, *subscribe to newsletter*, *follow on Twitter*.

🙋‍♀️ Get support

Join Ray Slack, *ask questions on forum*, *open an issue*.

🔭 Contribute to Ray

Read contributor guide, *create a pull request*.

Thank you!

We hope to meet again.