



LLM TRAINING

Build and Productionize
LLM-Powered Applications
with Ray & Anyscale





Meet the tutorial team!



Marwan

marwan@anyscale.com



Adam

adamb@anyscale.com



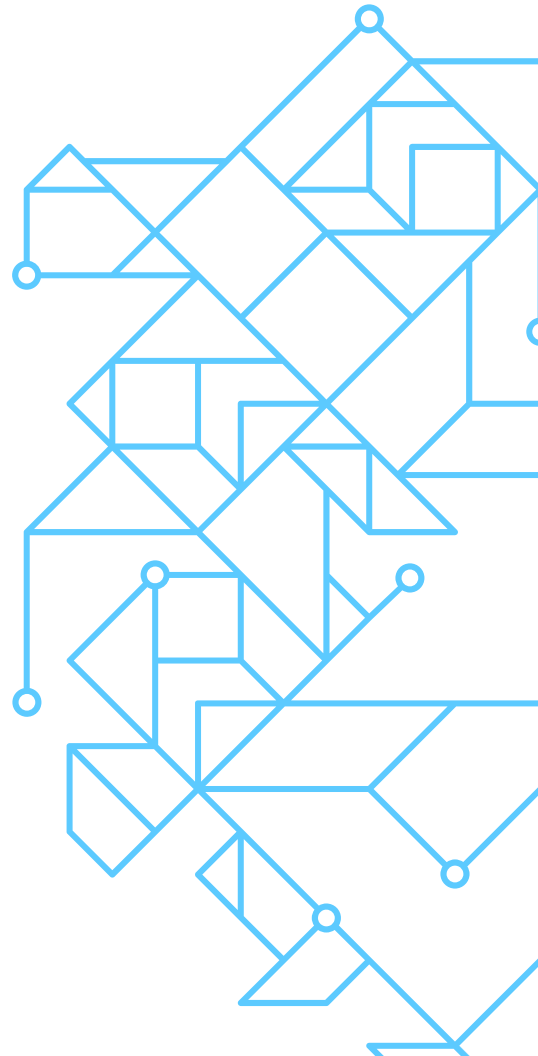
Kamil

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.

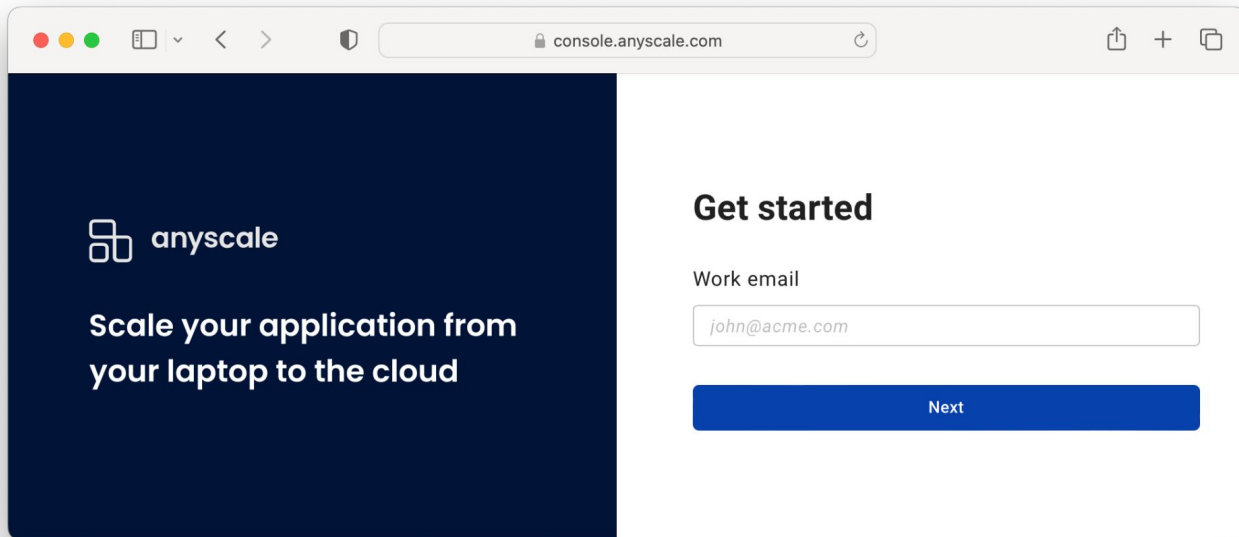


Accessing Anyscale clusters.

- 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



A screenshot of a web browser window showing the Anyscale console login page. The browser's address bar displays 'console.anyscale.com'. The page is split into two main sections. The left section has a dark blue background and features the Anyscale logo, the text 'anyscale', and the slogan 'Scale your application from your laptop to the cloud'. The right section has a white background and is titled 'Get started'. It contains a 'Work email' label, a text input field with the placeholder 'john@acme.com', and a blue 'Next' button.

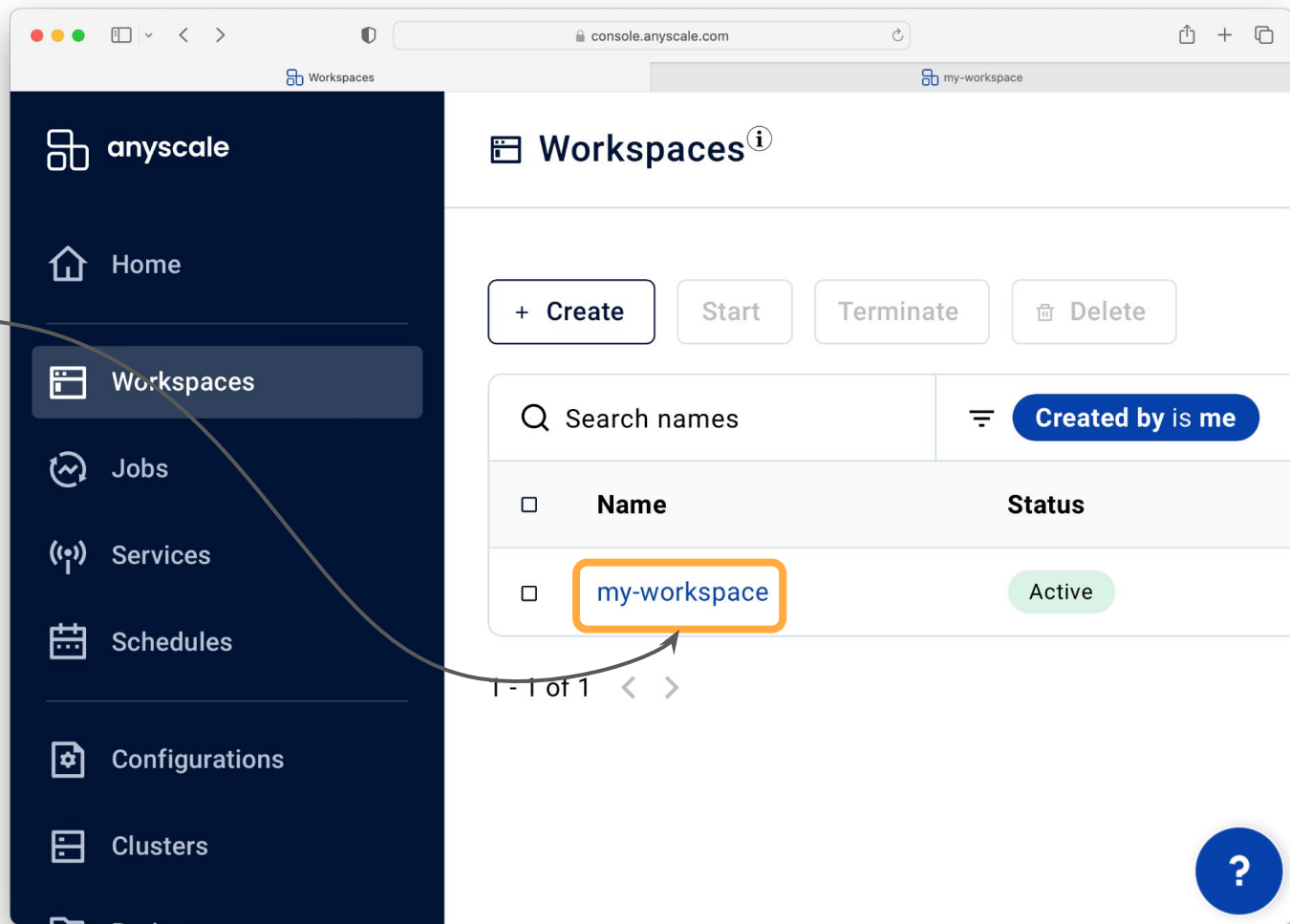
Enter the
**unique
credentials**
sent to your
email!

1. Select Workspaces



The screenshot shows the Anyscale console interface. On the left is a dark blue sidebar with the 'anyscale' logo at the top. Below the logo are several menu items: 'Home' (with a house icon), 'Workspaces' (with a document icon and highlighted by an orange border), 'Jobs' (with a circular arrow icon), 'Services' (with a radio tower icon), 'Schedules' (with a calendar icon), 'Configurations' (with a gear icon), 'Clusters' (with a document icon), and 'Projects' (with a folder icon). On the right is the main content area, which has a light gray header with a 'Home' link and a house icon. Below the header, the section is titled 'Examples to get started'. There are two example cards. The first card is titled 'Introduction to Anyscale & Ray' with a green book icon. It has a 'Launch' button and a dropdown arrow. Below the title, there is a paragraph of text: 'Learn about Anyscale and Ray in this introductory tutorial. This template runs a simple Ray program on a distributed Ray cluster then deploys an Anyscale Job based on the Ray program.' At the bottom of the card are two buttons: 'Ray task' and 'Anyscale Job'. The second card is titled 'Many Model Training' with a yellow lightning bolt icon. It has a 'Launch' button and a dropdown arrow. A blue circle with a white question mark is overlaid on the dropdown arrow of the 'Many Model Training' card. The browser's address bar at the top shows 'console.anyscale.com'.

2. Select Your Workspace



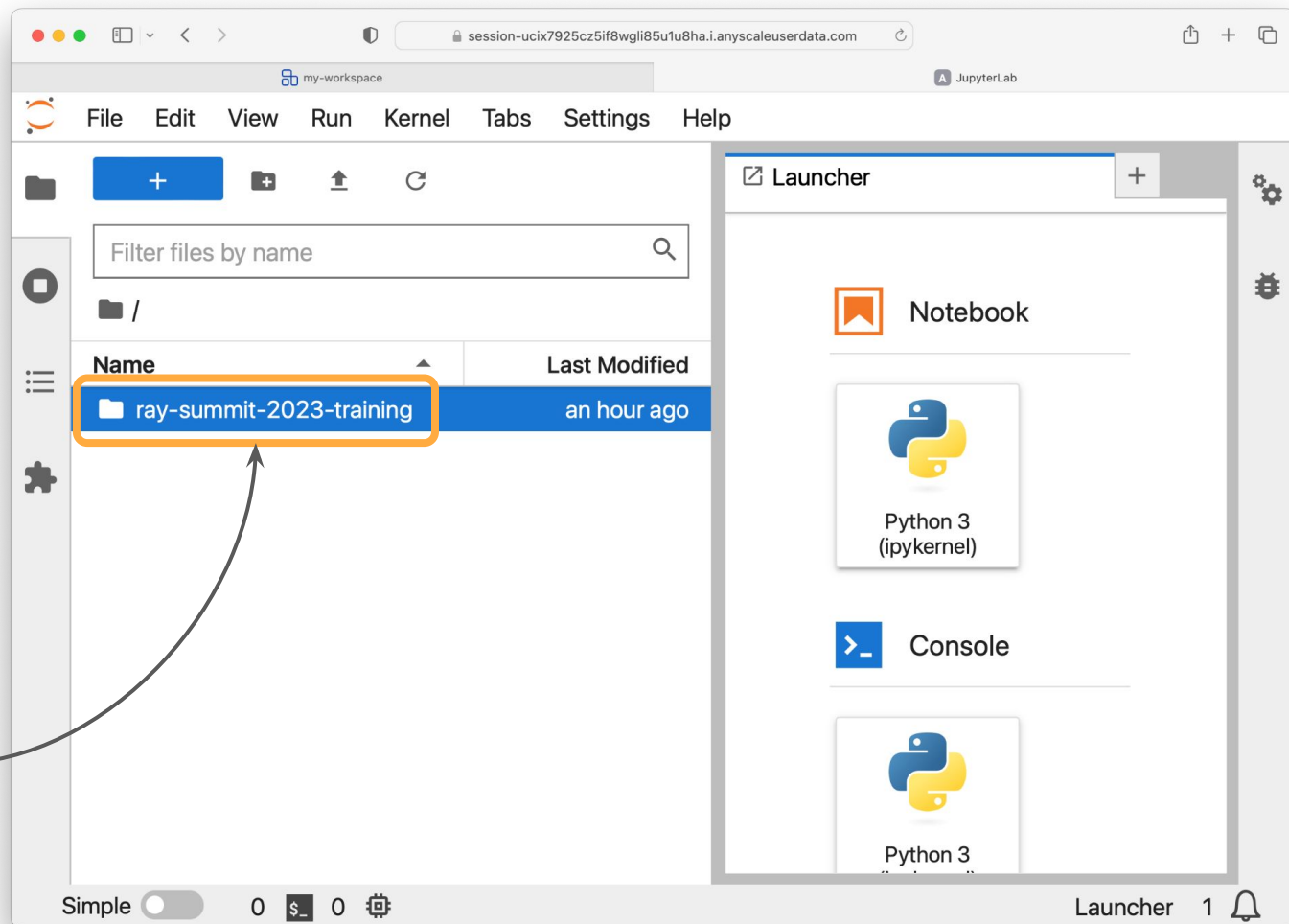
The screenshot shows the Anyscale console interface. The left sidebar contains the following menu items: Home, Workspaces (highlighted), Jobs, Services, Schedules, Configurations, and Clusters. The main content area is titled 'Workspaces' and includes buttons for '+ Create', 'Start', 'Terminate', and 'Delete'. Below these is a search bar labeled 'Search names' and a filter button labeled 'Created by is me'. A table lists the workspaces with columns 'Name' and 'Status'. The table contains one entry: 'my-workspace' with a status of 'Active'. An orange box highlights the 'my-workspace' entry, and a curved arrow points from the 'Workspaces' menu item in the sidebar to this entry. The bottom of the table shows '1 - 1 of 1' and navigation arrows. A blue help button with a question mark is in the bottom right corner.

Name	Status
my-workspace	Active

3. Click on
Jupyter
icon

The screenshot shows the Anyscale console interface. On the left is a dark blue sidebar with navigation links: Home, Workspaces (selected), Jobs, Services, Schedules, Configurations, Clusters, Projects, Emmy, Help, and Feedback. The main content area displays details for a workspace named 'm...workspace' which is 'Active (Ray)'. At the top right of this section are icons for JupyterLab (highlighted with an orange box), Visual Studio Code, and Databricks, along with 'Terminate' and 'Tools' buttons. Below the workspace name is a tabbed interface with 'About', 'Files', 'Terminal', 'Logs', and 'Serve deployments'. The 'About' tab is active, showing metadata like 'Created' (Sep 7, 2023) and 'Network access' (Public with auth token). It also lists 'Resources' such as 'Cluster environment' (summit:9) and 'Job submissions' (None). A 'README' section titled 'Workspaces' explains that they are fully managed development environments for building distributed Ray applications. A blue question mark icon is visible in the bottom right corner.

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.

