Intro to AWS SageMaker

Sign In & Check Setup

***Etherpad* — go.wisc.edu/fd65q4**

MLM25: Oct. 9

WISCONSIN INSTITUTE FOR DISCOVERY

data science HUB

ML+X

Slides: go.wisc.edu/54nt66

MLM25 — Looking Ahead

1. **10/16 (Thur), 4:30-6:30PM**: Sprint 3 + RAG with Watsonx.ai
    a. Request access to Watsonx.ai (15/50 seats filled so far):
2. **10/23 (Thur), 4:30-6:30PM**: Sprint 4
3. **10/30 (Thur), 4:30-7:30PM**: Intro to GCP and Vertex AI

Adjusted 11/13 & 12/11 to run from 5-7:30pm (prev. 5:30-7:30) to give us an extra 30 minutes for presentations

Full schedule: ml-marathon.wisc.edu/schedule/

About the Workshop

- Developed the Intro to SageMaker workshop last year to train participants of MLM24.
- **Open source!** Available for independent study.
- Adding additional materials and as (research) needs arise
    - At the end of the workshop, please fill out our feedback survey: https://forms.gle/mTUhuuKvHfnowzik6

Slides: go.wisc.edu/54nt66

Code Along Workshop: Available To Assist

- **Zekai Otles** — RCI Consultant, UW-Madison
- **Shashank Tanksali** — Senior Solutions Architect, AWS
- **Abrar Hussain** — Assoc IT Professional, JRP

Flag a helper (raise your hand) if you need help!

Lesson Materials — Follow Along
https://carpentries-incubator.github.io/ML_with_AWS_SageMaker/index.html

Shared AWS Account — Expectations

- Please stick to the materials (within reason) in the lesson to avoid surprise charges. Credits are used for training researchers & students.
  - **Do not not use any services beyond SageMaker and S3**
  - **This applies for your projects as well**!
    - Avoid additional tools such as Bedrock, Glue, Lambda, etc., before discussing with Chris first
- **AWS Access**: All participants will retain access for 48 hours after the workshop (*to be used for completing workshop materials only*)
  - Extended access available upon request (next slide)

Extended AWS Access Available Upon Request (for Projects)

- **Max $50/person — sufficient for moderate GPU usage, model training (millions of params), LLM inference, RAG, etc.** Can pool credits across team members if you want to delegate AWS usage to one person (e.g., 1 person uses $150 max in a 3 person team). Make sure the whole team is on board!
  - **NO TRAINING LLMS**
  - **NO FINETUNING LLMs** — some *exceptions apply (talk to Chris)*

- **To retain/gain access to AWS after the workshop, please complete the AWS credit request form**
  - You will need to describe your planned experiments briefly
  - If additional experiments are needed later (new models, new pipelines), please re-submit the form

Resource Stewardship

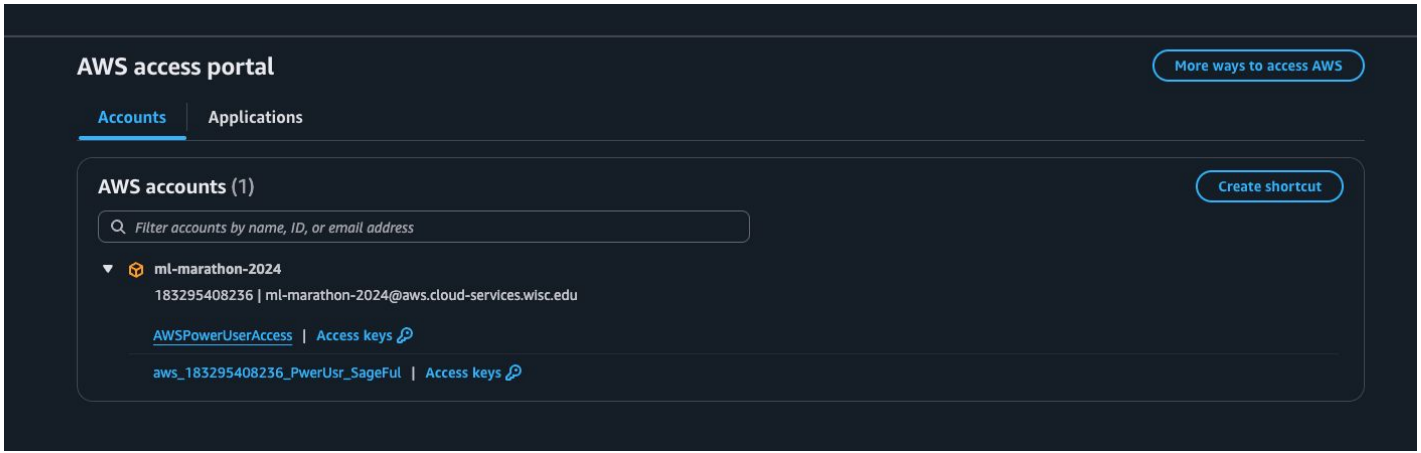
- **Tagging**: Tag all S3 buckets and Jupyter notebooks to include...
  - **Team:** My Awesome Team
  - **Name:** John Doe
  - **Purpose:** Train, Tune, Hyperparameter Search, RAG, etc.
  - **Model size:** 1M, 10M, 1B, 3B, etc.
- **Monitoring via Cost Explorer**. 24-48 hours to see costs reflected :(
- **Model size limits**
  - 20 million parameters for training
  - 8 B parameter for inference or generation
- **Delete data/buckets after work is complete. Pause notebooks when not in use.**
- **No sensitive or restricted data** (PHI, HIPAA, FERPA, or proprietary data)

Slides: go.wisc.edu/54nt66

Logging In

uw-madison-dlt3.awsapps.com/start/#/?tab=accounts

Select “aws_*_PwerUsr_SageFul”



Slides: go.wisc.edu/54nt66

Why AWS SageMaker?

- **High-performance compute (e.g., GPUs) on demand** – pay only for what you need
- **Flexible compute options** — easy to adjust GPU selection, use additional GPUs, etc.
- **Simplified and scalable ML/AI pipelines**
  - Avoids manual job orchestration / DAGs for common ML procedures (e.g., cross-validation)
  - Easy to parallelize both training and tuning
- **Support for custom scripts**
- **Cost management and monitoring**