

AWSの基盤を活用した ディープラーニング系ハンズオンの取り組み

2023/6/30

朝日放送テレビ株式会社

技術局 技術戦略部 荒木 優

自己紹介

- 荒木 優 (Yu Araki)

- 朝日放送テレビ株式会社 技術局技術戦略部
- 2014年入社 10年目
- 2014年5月～2018年5月 制作技術センター 音声担当
- 2018年6月～2022年5月 技術管理部 送信担当
 - タリー伝送システムAirTallyの開発も担当
- 2022年6月～ 技術戦略部
 - 新技術開発やその推進、クラウド活用推進、若手人材育成などを担当



- 好きなAWSサービス

- AWS IoT Core, Amplify, Cloud9

タリー伝送システムAirTallyの詳細はこちらから！

AirTally



AI研究会

- 技術戦略部が主体となって取り組む若手人材育成の一環
- メンバー9人
 - 若手メンバー： マスター、送信、制作技術
 - 中堅メンバー： マスター、技術戦略
 - 技術局内の有志メンバーで構成

AI研究会

- AI・ディープラーニングの業務での活用を目指し、ハンズオンを中心に、理解度向上・知見蓄積のための活動を行う
- 様々なツールや手法の調査・検証、そして社内のニーズ調査とアプリケーションの試作を実施し、映像・音声を扱う企業として知見獲得に取り組む。

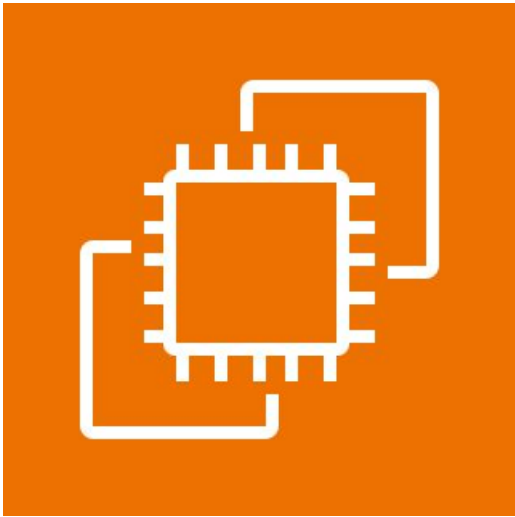
今後システム構築をする際のAI活用のベースになれば

AI研究会

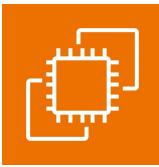
- とりあえずハンズオン形式でAIに触ることからスタート
 - AIの概念だけを座学で勉強してもおもしろくない
 - 実際に触ってみてAIで何ができるのかを体感することにより真の理解につながる

でもハンズオンで使えるGPU付きのマシンが人数分無い！

**せっかくなのでクラウドやLinux,Web系技術を絡めた
ハンズオンもしたい**



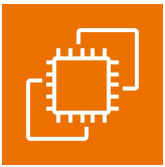
**使いたい時だけ使える
EC2のGPU付き
インスタンスを
使おう！**



EC2を使用

- G系インスタンスを使用
- GPUはNVIDIA T4 Tensor Core(GPUメモリ 16 GiB)

| インスタンスサイズ | GPU | vCPU | メモリ (GiB) | インスタンスストレージ(GB) | ネットワーク帯域幅 (Gbps) | EBS帯域幅 (Gbps) | オンデマンド料金/時間 |
|---------------|-----|------|-----------|-----------------|------------------|---------------|-------------|
| g4dn.xlarge | 1 | 4 | 16 | 1x125 NVMe SSD | 最大25 | 最大3.5 | 0.526 USD |
| g4dn.2xlarge | 1 | 8 | 32 | 1x225 NVMe SSD | 最大25 | 最大3.5 | 0.752 USD |
| g4dn.4xlarge | 1 | 16 | 64 | 1x225 NVMe SSD | 最大25 | 4.75 | 1.204 USD |
| g4dn.8xlarge | 1 | 32 | 128 | 1x900 NVMe SSD | 50 | 9.5 | 2.176 USD |
| g4dn.16xlarge | 1 | 64 | 256 | 1x900 NVMe SSD | 50 | 9.5 | 4.352 USD |
| g4dn.12xlarge | 4 | 48 | 192 | 1x900 NVMe SSD | 50 | 9.5 | 3.912 USD |
| g4dn.metal | 8 | 96 | 384 | 2x900 NVMe SSD | 100 | 19 | 7.824 USD |



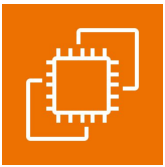
G系インスタンスを使用するには . . .

- デフォルトでvCPUのアサイン上限が0
- このままでは一切使えない

Service Quotas > AWS のサービス > Amazon Elastic Compute Cloud (Amazon EC2) > Running On-Demand G and VT instances

Running On-Demand G and VT instances

| 詳細 | | | |
|------------|---|------------------|------|
| 説明 | Maximum number of vCPUs assigned to the Running On-Demand G and VT instances. | | |
| クォータコード | クォータ ARN | | |
| L-DB2E81BA | arn:aws:servicequotas:ap-northeast-1:210034172202:ec2/L-DB2E81BA | | |
| 使用率 | 適用されたクォータ値 | AWS のデフォルトのクォータ値 | 調整可能 |
| 0 | 0 | 0 | はい |



G系インスタンスを使用するには . . .

- サービスクォータで引き上げのリクエストが必要
- リージョンごとにリクエストが必要
 - 東京・大阪で100ずつ申請してみました

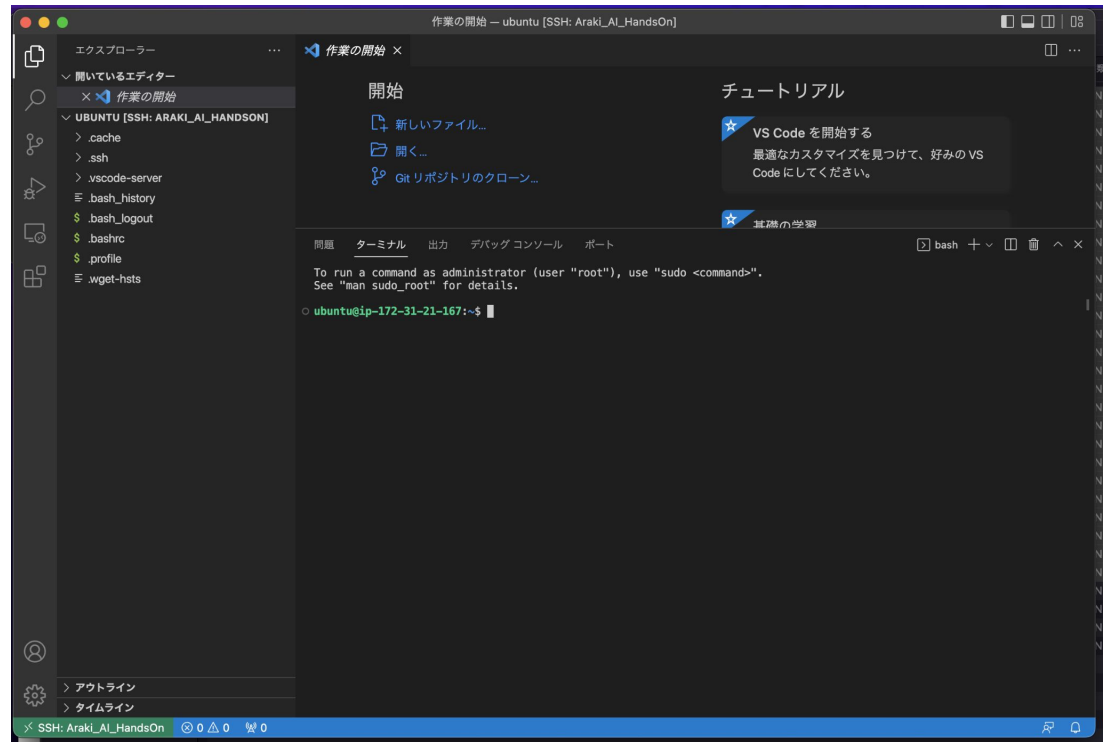
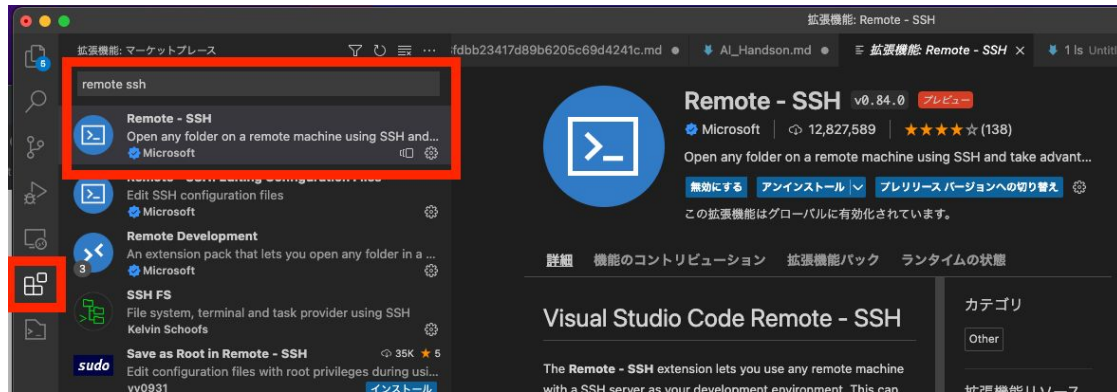
Service Quotas > AWS のサービス > Amazon Elastic Compute Cloud (Amazon EC2) > Running On-Demand G and VT instances

Running On-Demand G and VT instances

| 詳細 | | | |
|---|--|------------------|------|
| 説明 Maximum number of vCPUs assigned to the Running On-Demand G and VT instances. | | | |
| クォータコード | Quota ARN | | |
| L-DB2E81BA | arn:aws:servicequotas:ap-northeast-1:093759980587:ec2/L-DB2E81BA | | |
| 使用率 | 適用されたクォータ値 | AWS のデフォルトのクォータ値 | 調整可能 |
| 0 | 100 | 0 | はい |

ハンズオン内容

- AIのハンズオンと言いつつ、まずはEC2インスタンスを立ち上げてSSHでの接続からスタート



**インスタンス起動→SSH接続
の流れを毎回訓練！**

環境構築

- NVIDIAのドライバやCUDAのインストール
 - 最初は勉強のためにみんなで苦労しながら手動でインストール
 - インストールに時間がかかるのでめんどくさい



```
ubuntu@ip-172-31-26-33:~$ nvidia-smi
Thu Oct 20 05:19:19 2022

+-----+
| NVIDIA-SMI 515.76      Driver Version: 515.76      CUDA Version: 11.7     |
+-----+
| GPU  Name           Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|                                           MIG M.         |
+-----+-----+
|  0  Tesla T4             Off          | 00000000:00:1E:0 Off |   0%      Default  |
| N/A   72C    P8      21W / 70W   |  2MiB / 15360MiB |           N/A      |
+-----+-----+

Processes:
GPU  GI  CI          PID  Type  Process name          GPU Memory
   ID  ID  ID                               Usage
+-----+-----+
| No running processes found |
+-----+-----+
```

環境構築

- 途中からはAMIを事前に作って保存しておいて環境構築の時間短縮
- AWSで事前に用意されているDeep Learning AMIでもOK (Ubuntu 22.04は非対応)

バージョン問題があるので自分で作った方がいい場合も

▼ アプリケーションおよび OS イメージ (Amazon マシンイメージ) 情報

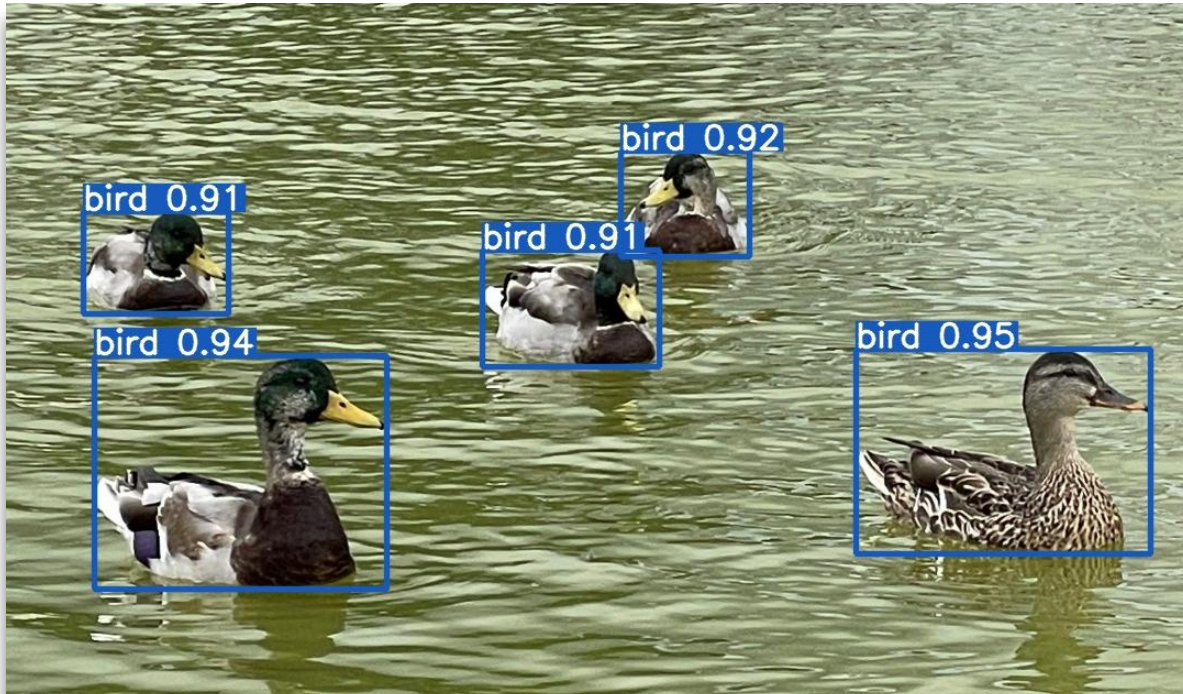
AMIは、インスタンスの起動に必要なソフトウェア設定 (オペレーティングシステム、アプリケーションサーバー、アプリケーション)

Q |

| | | | | | | |
|---|-----------------------|--------------------------|----------|--------------|-----------------|---|
| 20230427_Docker_Tensorflow2.2.0_JupyterLab | ami-06c36a5f7d1a2f736 | 2023-04-27T03:29:52.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | |
| Ubuntu22.04LTS_CUDAv11.7_cuDNNv8.6.0_Cloud9 | ami-08ad60f1907de4ecf | 2023-06-05T11:47:13.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | |
| Ubuntu22.04LTS_CUDAv11.7_cuDNNv8.6.0_nnabla-ext-cuda114 | ami-0a02571de4fe758e4 | 2023-01-26T06:11:29.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | |
| Ubuntu22.04LTS_CUDAv11.7_cuDNNv8.6.0 | ami-07725fd4e07dce589 | 2023-01-26T06:00:39.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | ✓ |
| 20230324_Docker_Tensorflow_JupyterLab | ami-0ca4213ae777e4b36 | 2023-03-23T22:25:39.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | |
| Ubuntu22.04LTS_CUDAv11.7_cuDNNv8.6.0 | ami-07725fd4e07dce589 | 2023-01-26T06:00:39.000Z | 仮想化: hvm | ENA 有効: true | ルートデバイスタイプ: ebs | ▲ |

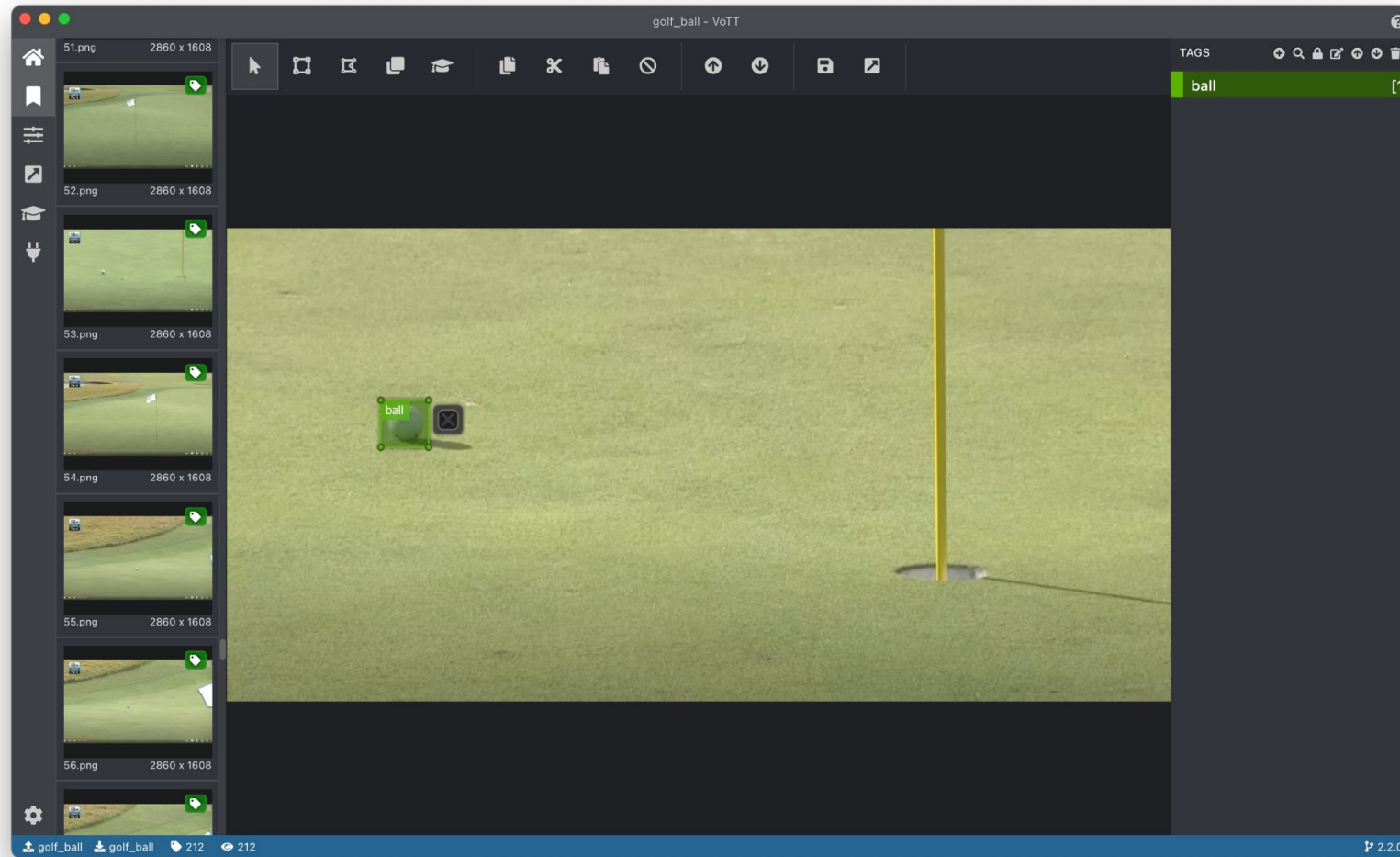
ハンズオン内容の一例

- 画像認識
 - よくある学習済みモデルを使った推論



ハンズオン内容の一例

- 全員でタグ付け大会&転移学習



ball 0.91



ハンズオン内容の一例

- 顔認識&自動モザイク
- 音源分離
- 大規模言語モデル
 - 簡易なAPI化なども
- Dockerでの環境構築





**Amplifyでハンズオン
サイトを作ろう！**

個人的にはパワポ作りがあまり好きではないので . . .

- AmplifyとMkDocsでハンズオン用Webサイトを作ってみた

≡ AI研究会
🔍 検索

第1回 ディープラーニング環境構築とYOLOv7による物体認識

本日のゴール

- AWS EC2上のインスタンスにディープラーニングを行う環境を構築する
- Pytorchを用いてYOLOv7を実行できる環境を構築する
- YOLOv7による物体認識を行う

• YOLOv7による姿勢推定を行う

目次

本日のゴール

事前準備

1. Visual Studio Codeのインストール
2. 拡張機能「Japanese Language Pack for Visual Studio Code」のインストール
3. 拡張機能「Remote-SSH」のインストール
4. Cyberduckのインストール

AWS EC2上でのインスタンス構築

1. G4インスタンスの詳細
2. AWSへのログイン
3. インスタンスの作成

Macの場合のみ、下記の作業を実施する

Visual Studio CodeのSSH接続設定

1. Remote-SSHの設定
2. Remote-SSHの起動

EC2上でのディープラーニング環境の構築

1. NVIDIA製GPUドライバーとCUDAのインストール

GPUドライバー

CUDA

インストール

2. cuDNNのインストール
- cuDNNのダウンロード
- EC2へのcuDNNのアップロード
- EC2へのcuDNNのアップロード

≡ 第1回 ディープラーニング環境構築とYOLOv7による物体認識
🔍 検索

1. NVIDIA製GPUドライバーとCUDAのインストール

GPUドライバー

GPU付きのインスタンスを立ち上げただけでは、NVIDIA製GPUを使用するための純正ドライバーが入っておらず、GPUの性能を十分に発揮することができない。そのため、NVIDIA純正のドライバーをインストールする。

CUDA

CUDAは、NVIDIAが自社製のGPU向けに提供しているソフトウェアの開発・実行環境。GPUを用いて汎用的な並列計算を行うためのライブラリである。GPUを本来の画像処理のために使用する場合はCUDAは不要だが、ディープラーニングに必要な汎用的な並列計算を行う際に必要になる。

インストール

GPUドライバーとCUDAは一括でインストールされるため、ターミナルで、下記のコマンドを順次実行してインストールを行う。

インストールにはaptというubuntuのパッケージマネージャーを使用する。

```

1 | wget https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64/cuda-keyring_1.0-1_all.deb
1 | sudo dpkg -i cuda-keyring_1.0-1_all.deb
1 | sudo apt update
1 | sudo apt -y install cuda-11-7
    
```

途中、下記のような紫の画面が出るが、気にせずエンターキー押下で進んでOKです。

Package configuration

Daemons using outdated libraries

目次

本日のゴール

事前準備

1. Visual Studio Codeのインストール
2. 拡張機能「Japanese Language Pack for Visual Studio Code」のインストール
3. 拡張機能「Remote-SSH」のインストール
4. Cyberduckのインストール

AWS EC2上でのインスタンス構築

1. G4インスタンスの詳細
2. AWSへのログイン
3. インスタンスの作成

Macの場合のみ、下記の作業を実施する

Visual Studio CodeのSSH接続設定

1. Remote-SSHの設定
2. Remote-SSHの起動

EC2上でのディープラーニング環境の構築

1. NVIDIA製GPUドライバーとCUDAのインストール

GPUドライバー

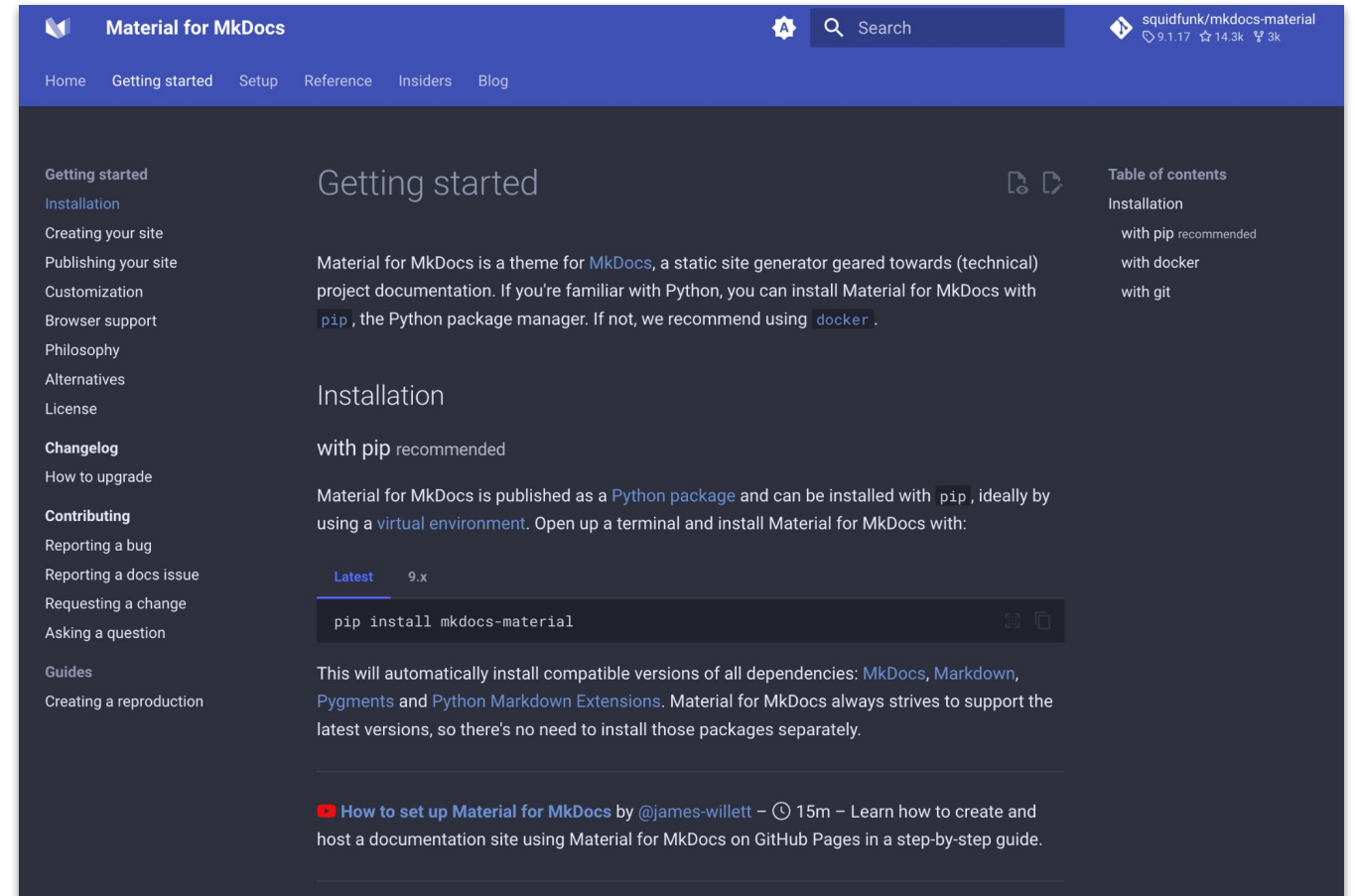
CUDA

インストール

2. cuDNNのインストール
- cuDNNのダウンロード
- EC2へのcuDNNのアップロード
- cuDNNのインストール

MkDocs

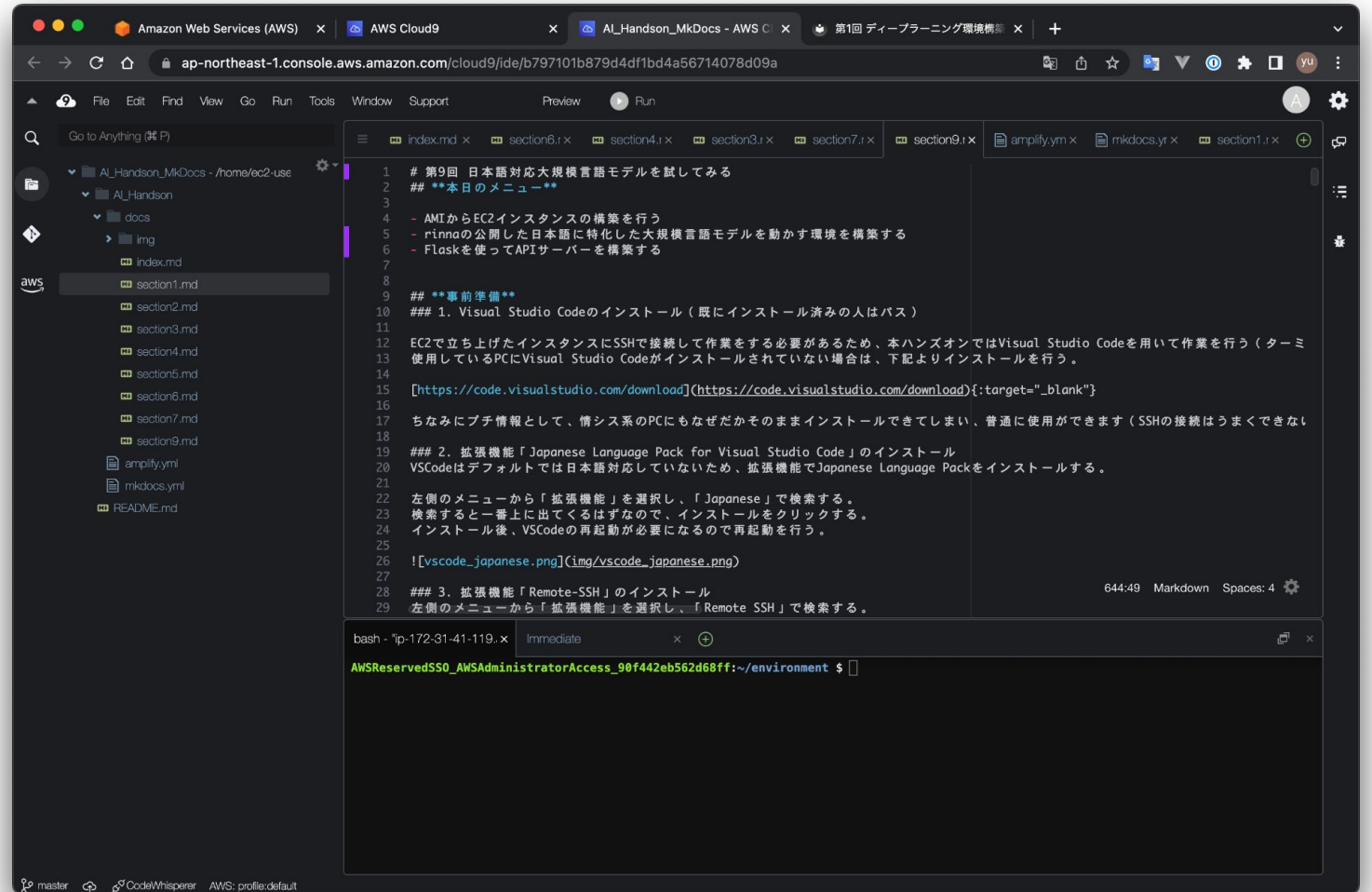
- ドキュメント構築向けの静的サイトジェネレータ
- コンテンツはMarkdownで記述
- レイアウトは勝手にいい感じになるので楽
- 今回はAmplify上でビルドして使う



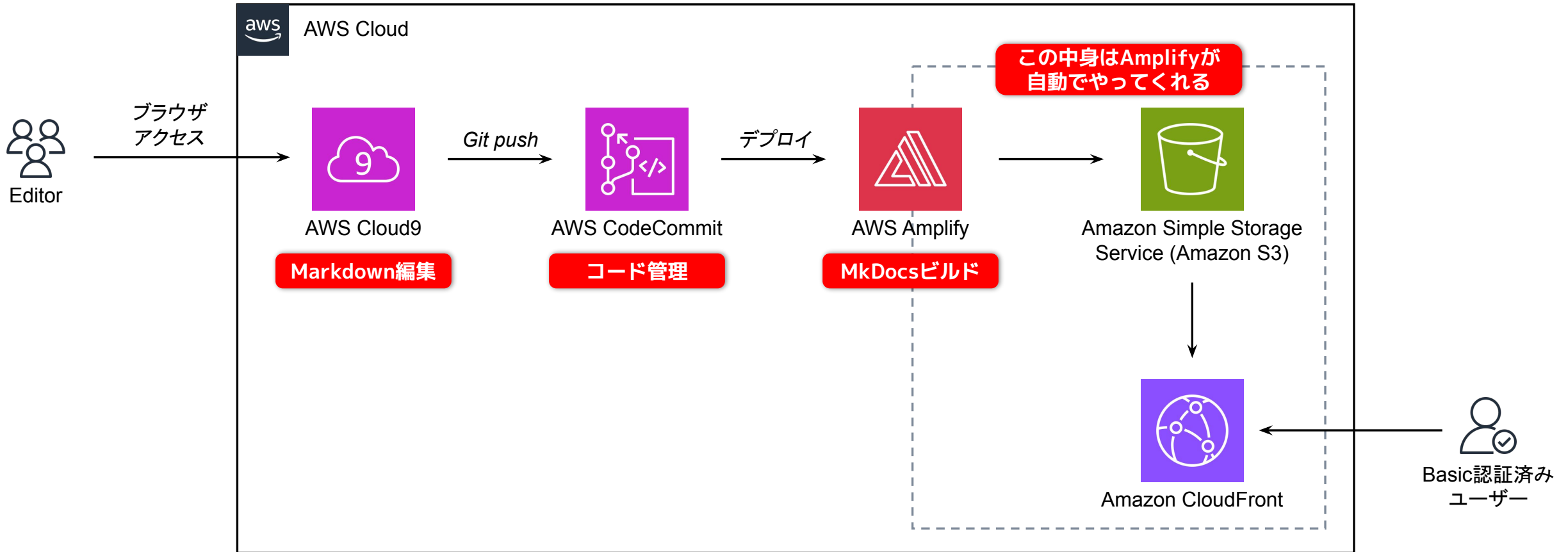


Cloud9を使用してMarkdownで記述

- ブラウザのみでコードを記述、実行、デバッグできるクラウドベースの統合開発環境 (IDE)
- ブラウザさえあればほとんどの環境で使用できる



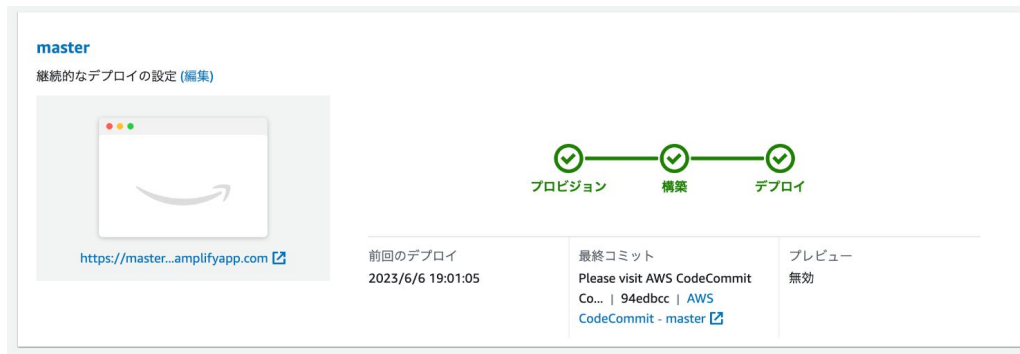
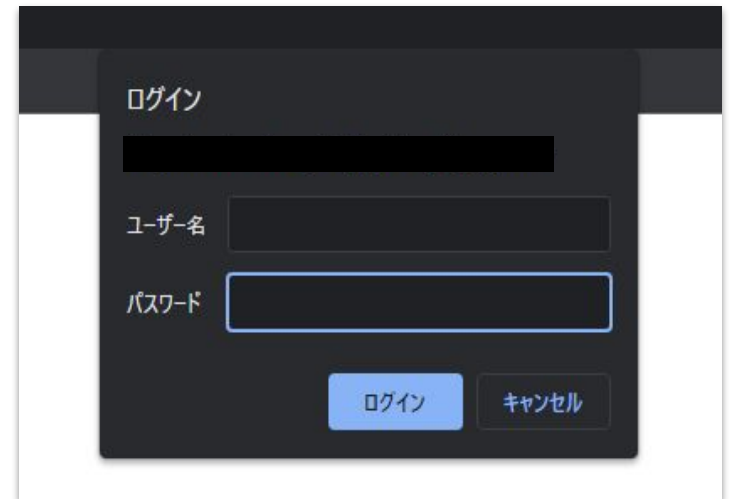
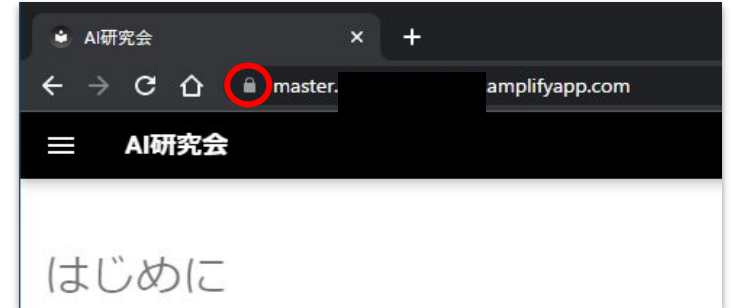
CodeCommitとAmplifyを使用して構築



全てAWS上で完結

Amplifyのいいところ

- 意識しなくてもhttpsのサイト構築が可能
- Basic認証を簡単にかけられる
 - 内部向けサイトを作りたい時に便利
- git pushコマンドだけでページ更新まで自動で行える
 - コード管理とデプロイが一括で行え便利



お金の話

- 毎月かかっている費用
 - EC2インスタンス
 - 事前準備とハンズオンで人数分動かしても現実的な価格
 - Amplify, Cloud9, CodeCommit
 - ごくわずか

年間12回続けてもGPU付きのマシンは1台も購入できないはず

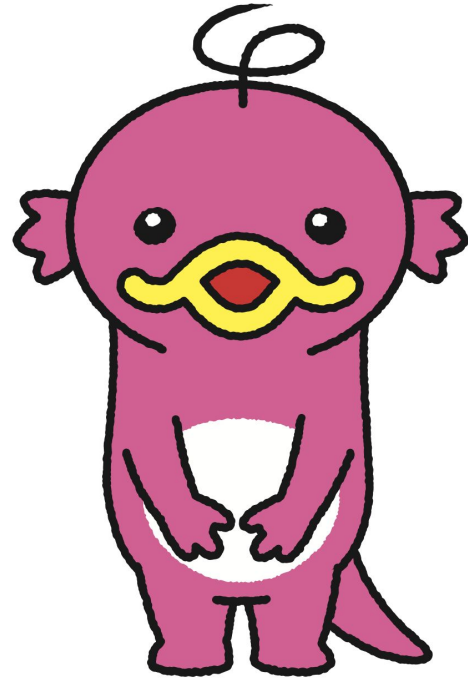
最後に

- AWSの基盤をうまく活用してAI系のハンズオンを実施
- AIの知識ももちろんだが、クラウドやLinux、Web周りの技術についても扱う
- クラウドを活用して低コストに
- 使いこなすにはまず触ってみて体感する機会が必要
- AIも手段の一つ

**メンバーがうまくAIやクラウド、新しい技術を使いこなして
目的とすることをスマートにできるようになるのが今後の目標**

AWSさんへの要望

- Cloud9でG系インスタンスがデフォルトで使えるようになると嬉しいです！



ご清聴ありがとうございました