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。

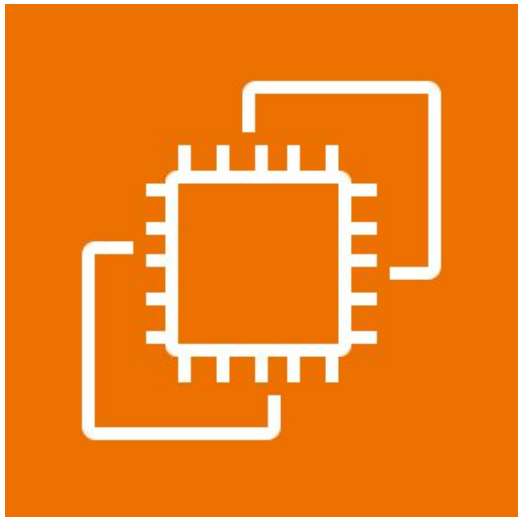
今後システム構築をする際のAI活用のベースになれば

AI研究会

- とりあえずハンズオン形式でAIを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・・・

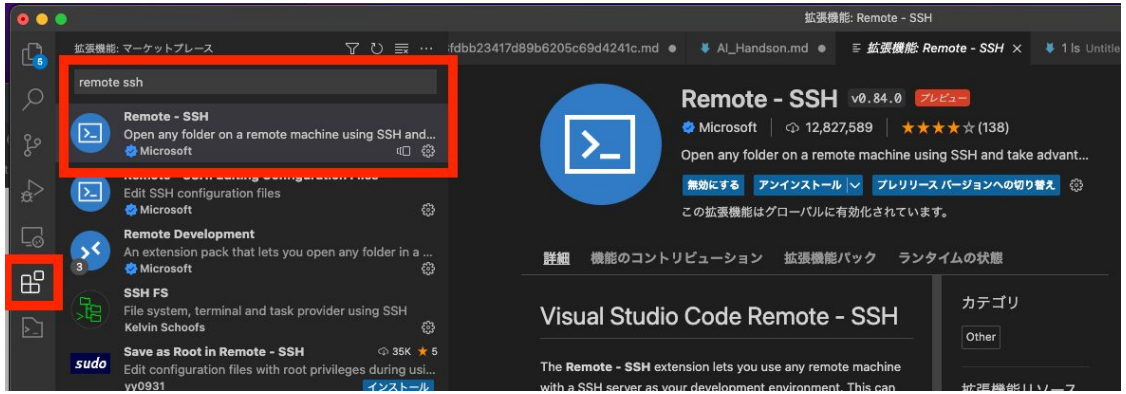
- デフォルトでvCPUのアサイン上限が0
- このままでは一切使えない

G系インスタンスを使用するには・・・

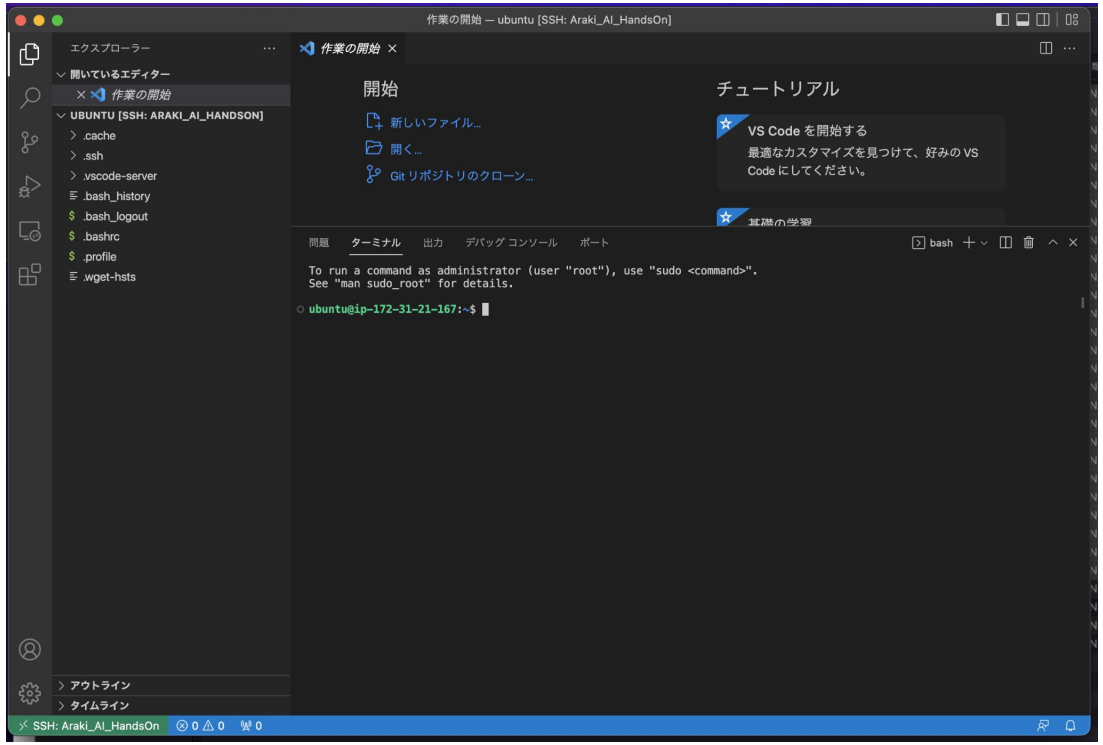
- サービスクォータで引き上げのリクエストが必要
- リージョンごとにリクエストが必要
  - 東京・大阪で100づつ申請してみました

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 |
| N/A   72C    P8    21W /  70W |      2MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

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

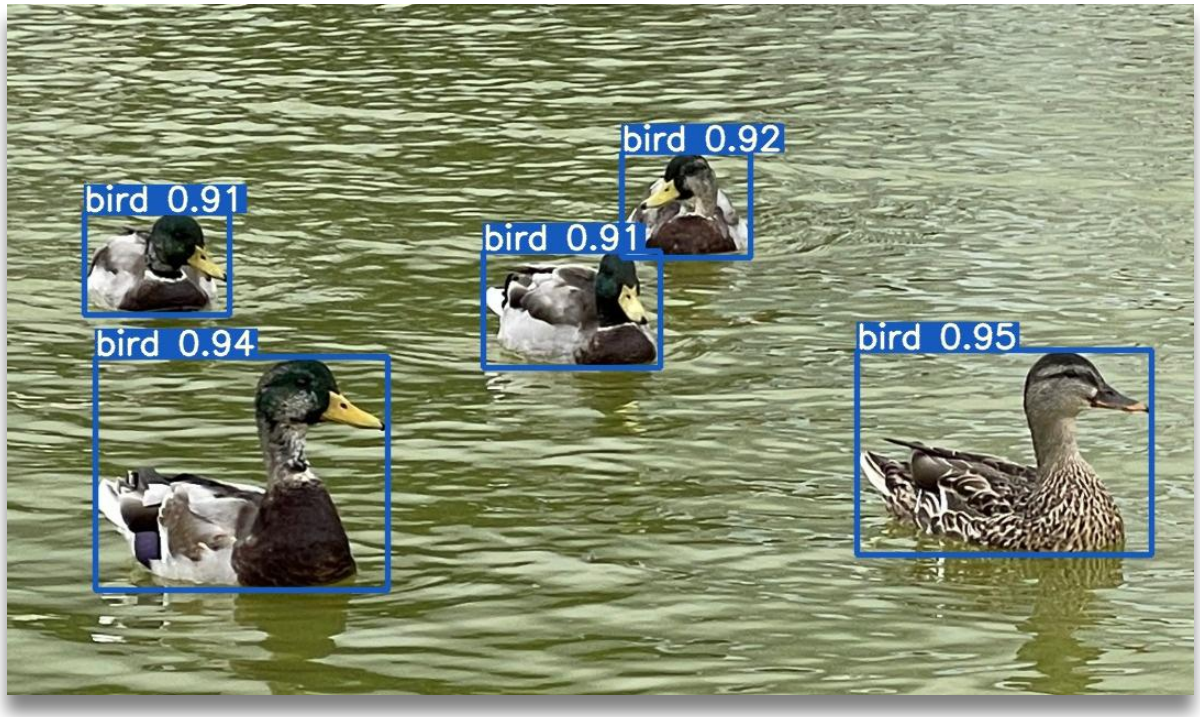
環境構築

- 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

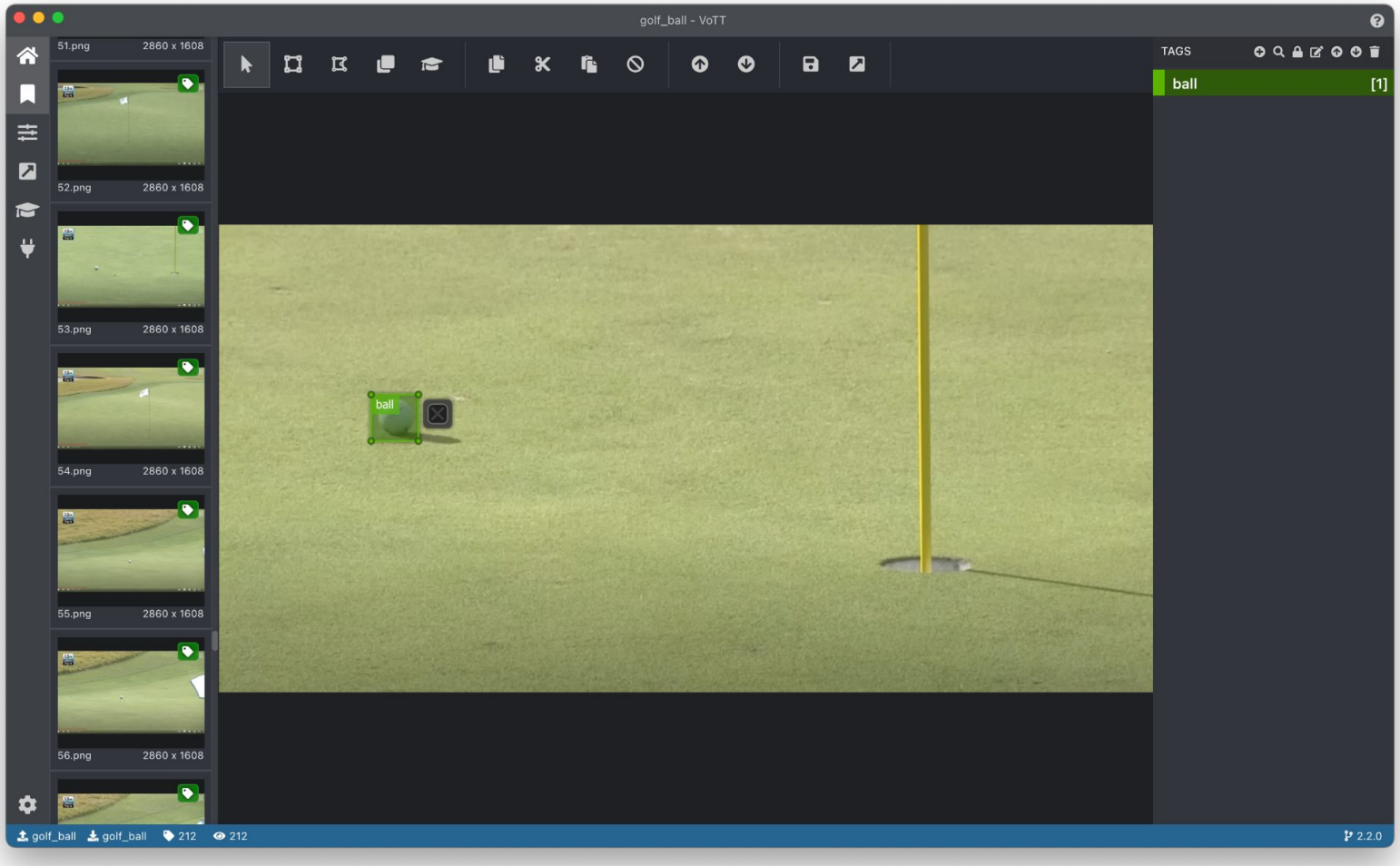
ハンズオン内容の一例

- 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

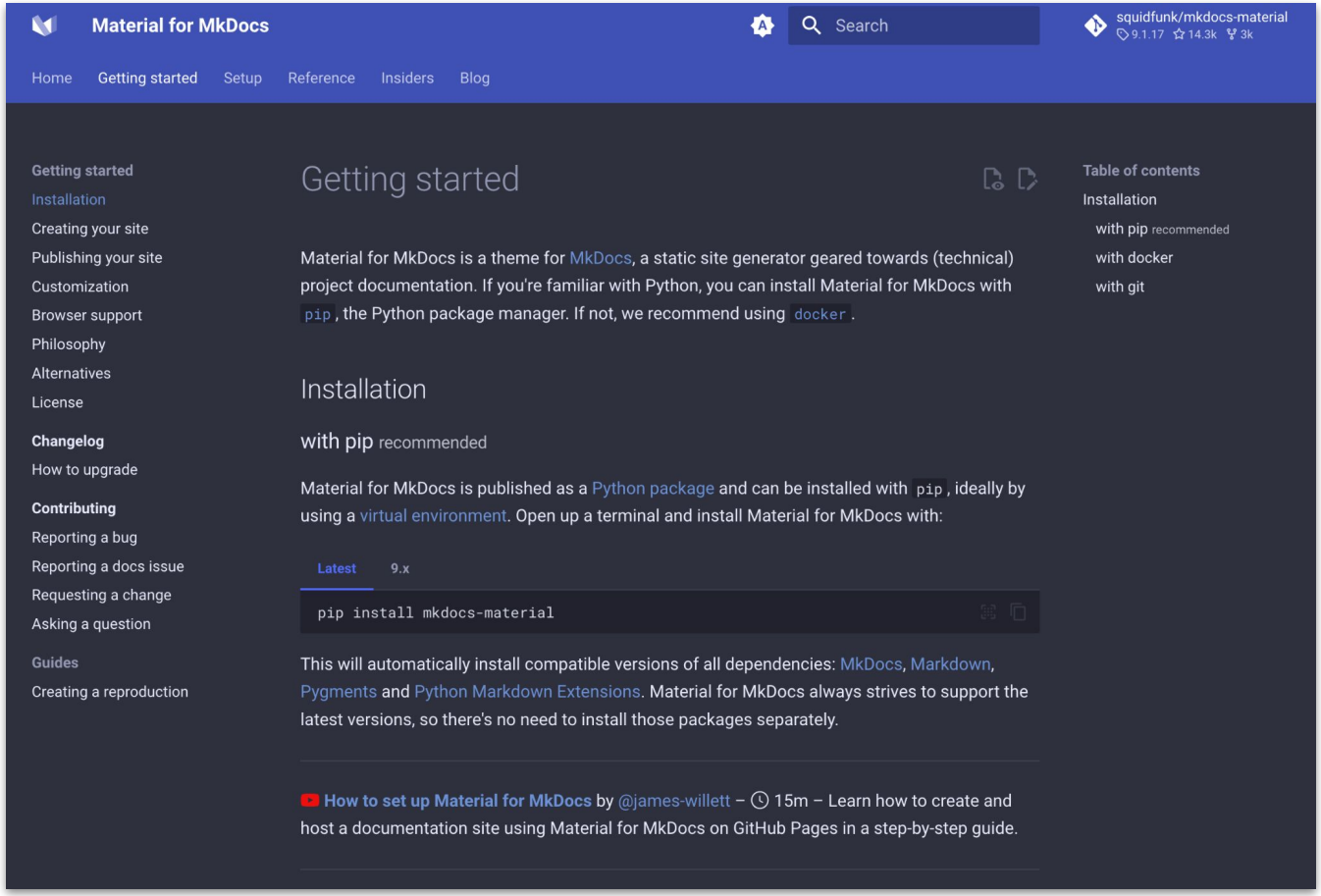
Amplifyでハンズオン サイトを作ろう！

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

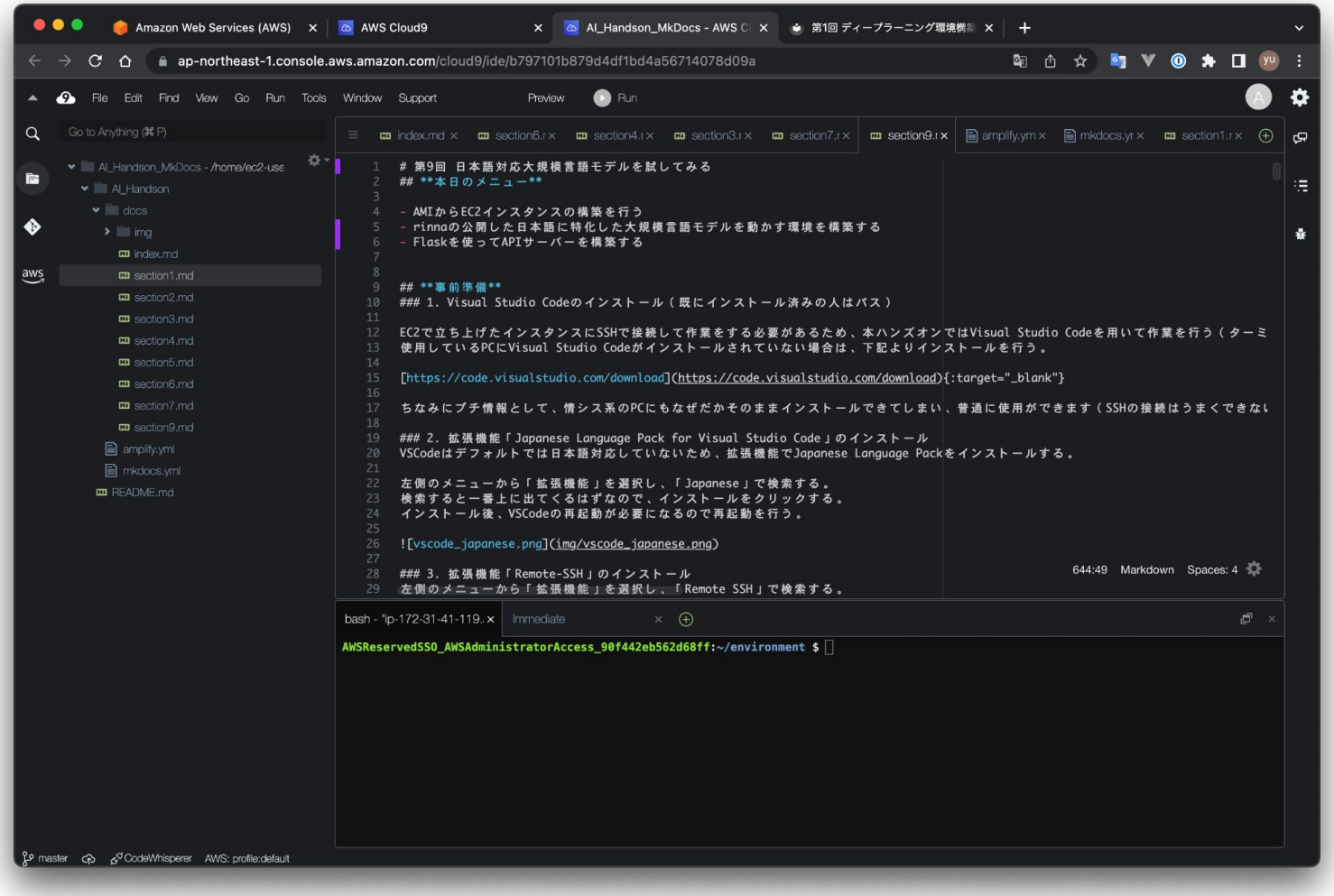
MkDocs

- 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

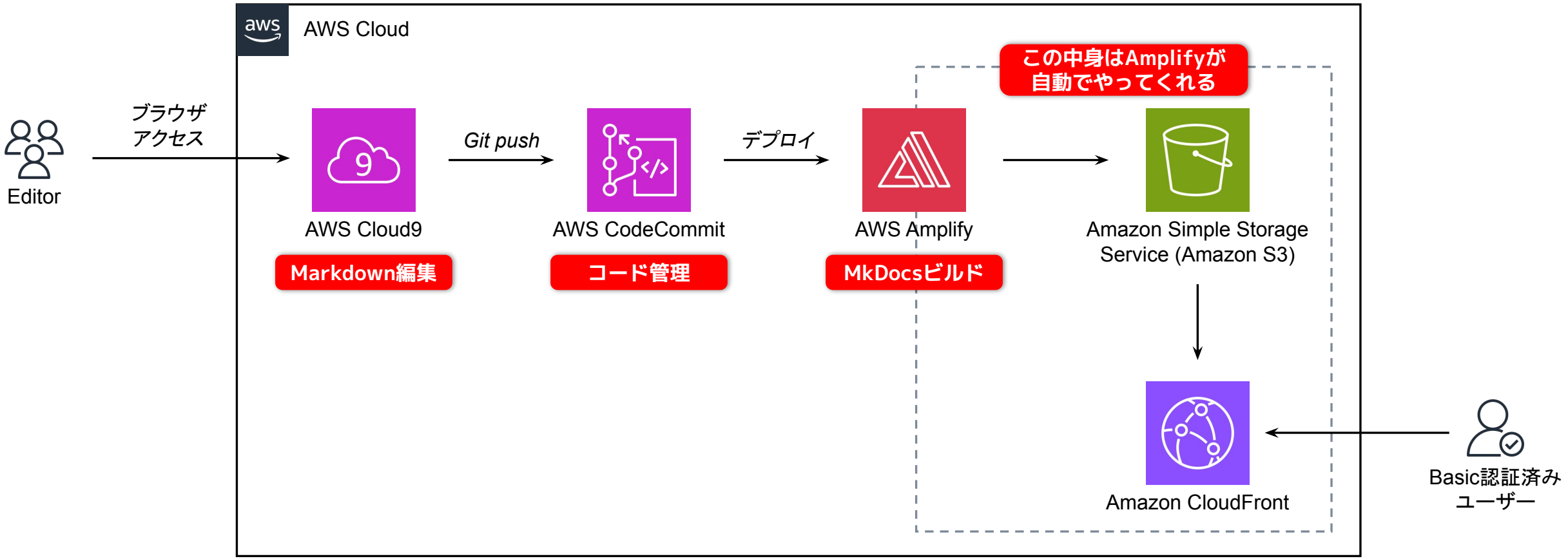


Cloud9を使用してMarkdownで記述

- 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



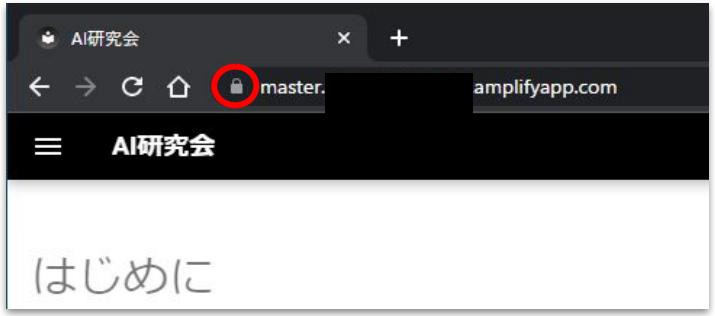
CodeCommitとAmplifyを使用して構築

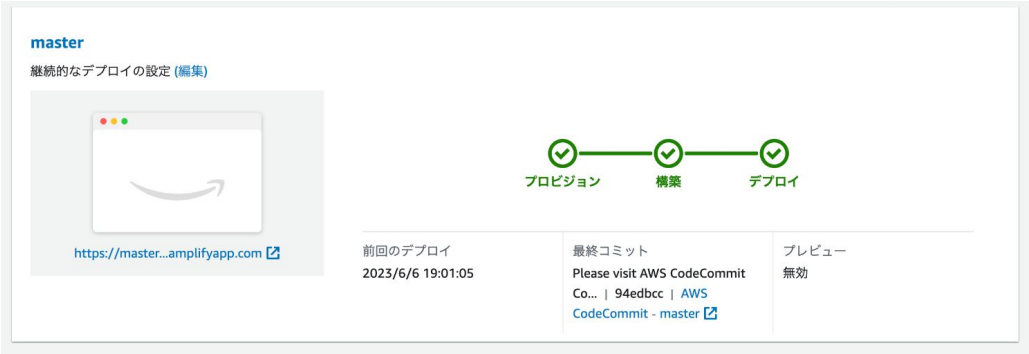


全てAWS上で完結

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

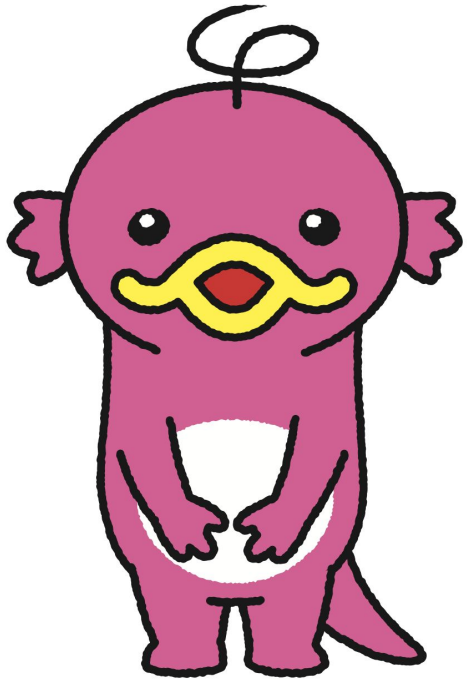
最後に

- 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