1Y-05

低費用化により中小規模組織の機械学習活用をめざす ハイブリッドMLOps基盤の提案

平見　修司†　　竹原　一駿†　　北　健志‡　　喜田　弘司†　　亀井　仁志†

香川大学†　　株式会社 STNet‡

1. はじめに

中小規模程度の組織において，機械学習の活用が進められている[1]．例えば機械学習は，過去のデータに基づいて，入力を分類できるため，製品の目視検査のような単純な作業の自動化に活用されている．

機械学習を使用するには，過去のデータを用いて機械学習モデル（以下，モデル）を作成する．そのモデルに対して，データを入力すると分類結果を出力する．一方，モデルを運用し続けると，モデルの精度が低下することが多い[2]．モデル精度を維持するために，モデルの継続的な再学習が必要である．この一連のプロセスは，運用を開発にフィードバックする MLOps を利用するのが一般的である． MLOps を実現するには複数のコンピュータを用いた MLOps 基盤を構築する必要がある．しかし，中小規模組織の機械学習活用では，高い月額費用が問題になる[3]．

本稿は，低費用で機械学習活用を目指すハイブリッド MLOps 基盤の開発と，費用における有効性について述べる．

2. MLOpsにおける課題

MLOps は，モデルの開発から運用までをスムーズに行うための手法であり，DevOps の考えを取り入れている． DevOps は， CI(Continuous Integration)/CD(Continuous Delivery)を採用することにより，開発と運用の間でフィードバックを受け渡し，アプリケーションの開発とリリースを迅速に行う考え方及び手法である[4]．

MLOps の実現には，MLOps 基盤の費用が問題になる． 文献[3]には，AI 導入の障害として月々の費用に関する懸念が上位にあり，許容できる額は，「〜10 万円/月」が回答の 80%を占めている．そのため，中小規模組織の機械学習活用を促進するには，月額費用を抑えるシステムの実現が課題になる．

3. 提案システム

本研究は，MLOps 基盤の構成要素をクラウドとオンプレミスに分けて配置し，月額費用を削減するハイブリッド MLOps 基盤を提案する．

ハイブリッド MLOps 基盤を構成する要素は，以下の 3 つである．

A) 学習向け高性能計算リソース

B) 運用向け計算リソース

C) 大容量の学習データを管理するストレージ

A)は，機械学習時のみ稼働でき，クラウドの従量課金制に適する．また，機械学習には高性能な計算リソースが必要なため，オンプレミスで用意すると初期費用が高くなる．B)は，高性能な計算リソースを必要とせず，既存の計算リソースを割り当てられると考えられる．C)は，クラウドだと料金が高くなる．以上から，A)をクラウドに，B)と C)をオンプレミスに配置することで，単一の場所に用意する場合よりも費用を削減できる．

4. システム構成

図 1 に，OSS を用いて構築したシステム構成を示す．本システムは，クラウドとして Microsoft Azure のマネージド Kubernetes サービスである AKS(Azure Kubernetes Service)[5]と，中小規模組織のオンプレミス物理マシンを使用する 2 つのプラットフォームで構成される．クラウド上にある A)は，オンプレミス上の C)から学習データを取得し，機械学習を用いてモデルを開発する．開発したモデルは，オンプレミス上の B)でテストし，デプロイする．

クラウドとオンプレミスはパブリックネットワークで接続している．

4.1. クラウドとオンプレミス連携のモデル開発

クラウドは，システムノードプールとユーザーノードプールで構成する．常駐するシステムノ

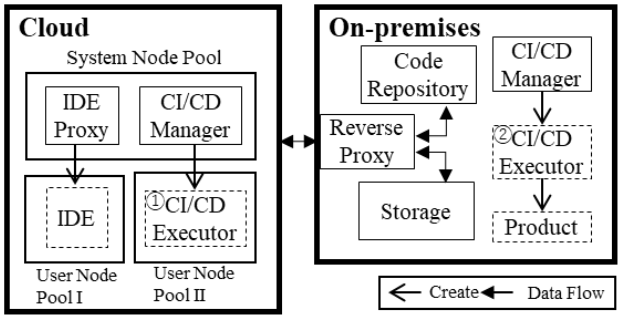

図1 システム構成

A Proposal of Hybrid MLOps Infrastructure toward Expanding Machine Learning for Small and Medium Size Organizations by Reducing Costs

†Shuji Hirami, Ichitoshi Takehara, Koji Kida, Hitoshi Kamei・Kagawa University

‡Kenji Kita・STNet, Inc.

ードには IDE Proxy(Jupyter Hub)を構築し，開発者がユーザー認証を行った後，ユーザーノードで IDE(Jupyter Notebook)を起動する．開発を行わないときはユーザーノードを起動しないため，クラウドに掛かる費用を削減できる．これにより，クラウド上にて A)を実現する．オンプレミスには，コードリポジトリ(GitLab)とストレージ(MinIO)を構築し，オンプレミス上で C)を実現する．開発者は，オンプレミスからコードと学習データを取得し，IDE 上で試行錯誤する．完成したコードを GitLab へ Push すると，CI によってモデルを作成し，CD により運用へデプロイする．

4.2. 開発したモデルのデプロイフロー

4.1.節同様，CI/CD でもユーザーノードは必要なときのみ起動するため，費用を削減できる．システムノードプールとオンプレミスに CI/CD Manager(GitLab Runner)を構築しており，CI/CD パイプラインに従って CI/CD Executor を立ち上げ，ジョブを実行する．流れを以下に示す．

I. クラウド上の CI/CD Executor①で機械学習し，モデルを作成
II. オンプレミス上の CI/CD Executor②でモデルの精度をテスト
III. モデルと API 用コードをイメージにビルドし，GitLab のコンテナレジストリに push
IV. IIIでビルドしたイメージを pull し，コンテナを作成・展開

II. ～IV. により，オンプレミス上で，3 章の B)を実現する．

5. 評価

本章は，ハイブリッド MLOps 基盤の費用見積もり式Cと，想定ユースケースの費用見積もり結果を述べる．ユースケースは，従業員数 100 人未満で，目視検査に本基盤を利用する場合を想定する．クラウドに掛かる費用は，以下の 4 つである．なお，①から④は全て従量課金制である．

① ノードプール代
② ロードバランサー(LB)代
③ 静的 IPv4 アドレス代
④ その他

①の費用C_{np}は，1 時間あたりの VM 費をC_{vm}としノード n 台で稼働した累計時間をT_n，ノード数の上限値を k とする場合，(1)式で算出できる．

$$C_{np} = C_{vm} \sum_{n=0}^{k} nT_n \quad \cdots\cdots(1)$$

②と③は，それぞれ掛かる費用をC_{lb}，C_{ip}，1 時間あたりの費用をL，Pとし，累計稼働時間tとする場合，(2),(3)式で算出できる．

$$C_{lb} = Lt \quad \cdots\cdots(2)$$
$$C_{ip} = Pt \quad \cdots\cdots(3)$$

④は，Pod に一時的に割り当てられるディスクや，データ通信などの費用である．これらの費用を e と置く．ただし，e が占める割合は十分小さい．

表1 見積もり値

Items		Value			Ratio
C_{np}		C_{vm}	k	T_n	81%
	C_{np1} [B2s]	11.52 yen / h	3	0 h, 650 h, 60 h, 10 h	
	C_{np2} [NV6ads_A10_v5]	74.7 yen / h	3	675 h, 40 h, 4 h, 1 h	
	C_{np3} [NV18ads_A10_v5]	263.61 yen / h	3	710 h, 5 h, 1 h, 0.2 h	
C_{lb}	L	3.36 yen / h			13%
	t	720 h			
C_{ip}	P	0.67 yen /h			5%
	t	720 h			
e		200 yen			1%

[]内は使用する VM サイズ，B2s は OS ディスク代込
価格はリージョン：Japan East（参照2022-01-04）

本システムは，システムノードプールとユーザーノードプール 2 つで構成されるため，それぞれC_{np1}，C_{np2}, C_{np3}とする．さらに，Public IP は，LB と IDE Proxy の 2 つ必要である．よって，MLOps 基盤の費用見積もりは(4)式となる．

$$C = C_{np1} + C_{np2} + C_{np3} + C_{lb} + 2C_{ip} + e \quad \cdots(4)$$

想定するユースケースに対する式の変数の値を表 1 に示す．結果，約 18,388.8 円 / 月 の費用で利用できることが分かった．本システムの費用は，10 万円/月 以内であることから，中小規模組織が機械学習活用を検討する際に有効である．

6. おわりに

本稿は，中小規模組織が機械学習の活用を進めるための費用を抑えたハイブリッド MLOps 基盤を提案して開発し，その費用を評価した．今後，ハイブリッド MLOps 基盤の有効範囲を明らかにする．

参考文献

[1] “中小企業の AI 活用促進について”, 経済産業省, https://www.meti.go.jp/policy/it_policy/jinzai/AIutilization.html, (参照 2022-12-18)
[2] T. Yamada and H. Matsutani, 2022, “A Sequential Concept Drift Detection Method for On-Device Learning on Low-End Edge Devices.”, https://arxiv.org/abs/2212.09637, arXiv
[3] “予算は 10 万円まで！？ / 5.5%にとどまる中小企業の AI 導入率 -日本の中小企業の AI 導入状況-”, AINOW, https://ainow.ai/2019/03/27/165636/#_AI, (参照 2022-12-18)
[4] 澁井雄介, ”AI エンジニアのための機械学習システムデザインパターン”, 株式会社翔泳社, 2021, pp. 004‐005.
[5] “Azure Kubernetes Service”, https://azure.microsoft.com/ja-jp/products/kubernetes-service/, (参照 2022-12-18)