APPLYING AI MODEL MANAGEMENT ON THE CSC'S AIOT PLATFORM

BY

I-CHUN HUNG[*], CHIE-HSIUNG KUO[**], YEN-TING CHEN[**], YAN-TING LIN[***], CHUN-CHIEH WANG[***], LI TE TU[****], CHIH KUI CHEN[*****]

SYNOPSIS:

China Steel Corporation (CSC) promotes smart manufacturing and digital transformation including constructing the 5G AIoT hybrid cloud infrastructure in recent years. CSC builds the AIoT platform as a hub for operating various intelligent applications and serving as a smart manufacturing foundation.

The development of each intelligent application is usually a time-consuming process and requires different resource investments. Hence, a suitable management approach for increasing intelligent applications is vital to keep them functioning normally.

This paper describes that CSC implements the AI model operationalization support system (AIMOSS) based on the concept of ModelOps for the AIoT platform. The AIMOSS manages the online AI models and continuously deploys them to the AIoT platform when model drift occurs.

The aim is to keep the online AI models running normally and contribute to effective production processes. It provides key functionalities for AI models such as version control, runtime monitoring, quality validation, and automatic retraining & redeployment.

A pilot case was conducted to test the effectiveness of the AIMOSS. Operation engineers can receive support from AIMOSS to manage intelligent applications efficiently and their feedback shed a light on the direction of future improvements to the AIoT platform.

Keywords: Smart Manufacturing, 5G AIoT, AI Model Operationalization, Intelligent Application, China Steel Corporation (CSC)

[*] Engineer, China Steel Corporation, Kaohsiung, Taiwan
[**] Scientist, China Steel Corporation, Kaohsiung, Taiwan
[***] Engineer, China Steel Corporation, Kaohsiung, Taiwan

[****] Project Manager, InfoCHAMP Systems Corporation, Kaohsiung, Taiwan
[*****] Manager, InfoCHAMP Systems Corporation, Kaohsiung, Taiwan

Introduction

AI models are the foundation of smart manufacturing solutions. Their purpose is to solve specific production problems. Companies must put significant efforts into such resource-intensive tasks for developing corresponding intelligent applications (Hechler, Oberhofer, & Schaeck, 2020).

At the very beginning of the development process, AI models require well-preserved data for selecting appropriate features. Production processes keep generating a large amount of data. The governance of valuable digital assets becomes crucial (Alhassan, Sammon, & Daly, 2016).

After an AI model is ready to use, the next step is to deploy it on an online platform and manage real-time inference requests. However, data drift inevitably occurs along with the changes in production. The detrimental impact on the model would affect its performance over time. The proper maintenance and management can help operation engineers make the online AI models function accurately without deteriorating by accumulated biases. Suitable management of daily operations on a production platform emphasizes the quality and accuracy of the online AI models.

This paper demonstrates the practice of applying AI model management on the CSC's AIoT platform which serves as the central hub of intelligent manufacturing solutions. A quality monitoring mechanism is beneficial to form the foundation of trustworthy AI with open data, processes, and algorithms (Janssen, Brous, Estevez, Barbosa, & Janowski, 2020). Hence, a standard deployment process and supporting tools is necessary to improve the AI model operationalization.

AI Model deployment is a continuous process that requires the AI models to provide accurate results persistently and switch to an adequate version AI model timely. Improve availability of the online AI models can help operation engineers manage the platform of intelligent applications more efficiently. Literature suggests that applying a suitable management approach for the increasing intelligent applications is vital to keep them functioning normally and overall workflow can become more efficient such as XOps (DataOps, MLOps, ModelOps, etc.) (Gartner, 2021). When accumulating intelligent applications continuously, the company and its operation engineers are going to face the challenges of scaling up the business including AI models deployment and quality monitoring.

Methods

CSC actively promotes smart manufacturing and digital transformation. By developing intelligent applications and optimizing production activities, CSC learns from the experience and defines the intelligent application procedure with five steps, namely defined problem, manageable information, analytic modeling, improved verification, and continued management (DMAIC).

AI models deployed in conventional IT infrastructure usually use different runtime environments which results in the increasing complexity of operation management. For example, operation engineers need to monitor multiple online AI modes during their duty shifts. However, this situation becomes insufficient when the number of online AI models starts increasing. CSC then built the AIoT platform with the hyper-converged infrastructure (HCI). It plays a critical role to support daily operations of the online/ready-to-use AI models. HCI brings significant benefits such as efficient IT, optimized cost, increased scalability, and simplified management. Developers of intelligent applications in a large team utilize the technology of continuous integration (CI) and continuous deployment (CD) to improve productivity and scalability (Hummer et al., 2019).

CSC refers to the design of ModelOps and focuses on the last two steps of DMAIC (i.e., improved verification and continued management) for constructing the platform. The CSC's AIoT platform aims to accommodate all the existing online and upcoming deployable AI models, especially for the widespread deployment of intelligent applications. Since the AI models contribute to the strategic decision support of manufacturing processes, the CSC's AIoT platform help corresponding AI model operations such as feature collection, training dataset preparation, performance monitoring, retraining, validation, and deployment.

Operationalizing AI models on the CSC's AIoT platform is based on two main functionalities, namely continuous deployment and model management. CSC implements an AI model operationalization support system (AIMOSS) to provide necessary tools and mechanisms for the whole workflow and operation engineers (Figure 1).

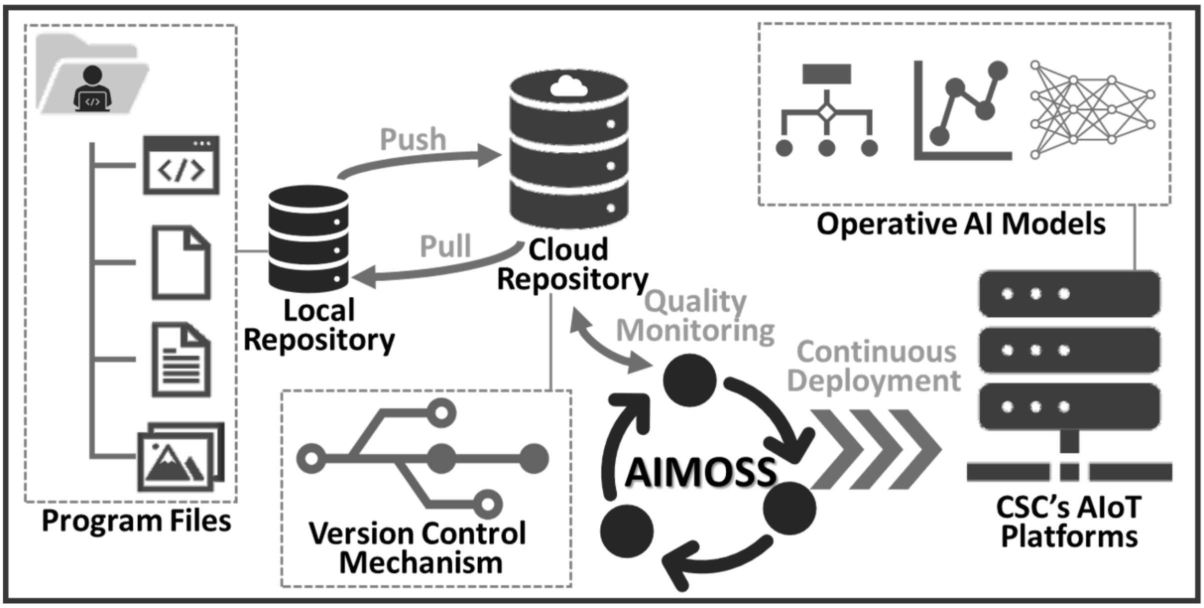

Figure 1. AI model operationalization support system (AIMOSS)

The continuous deployment regards how to manage various intelligent applications in a production field when operation engineers take over the upcoming deployable AI models. The AIMOSS implements a source code management (SCM) mechanism to provide repositories for storing and tracking each version of a deployed AI model (Figure 2). The mechanism makes the switch in between different versions of a deployed AI model more easily. The AIMOSS becomes capable of carrying out the switching procedure automatically without being intervened by operation engineers. Besides, it also provides more capability for collaboration among operation engineers and developers. Each version is given a tag and helps operation engineers identify a target version of an AI model. Operation engineers and developers can obtain the source code of the AI model promptly when requiring further modifications that the AIMOSS cannot handle.

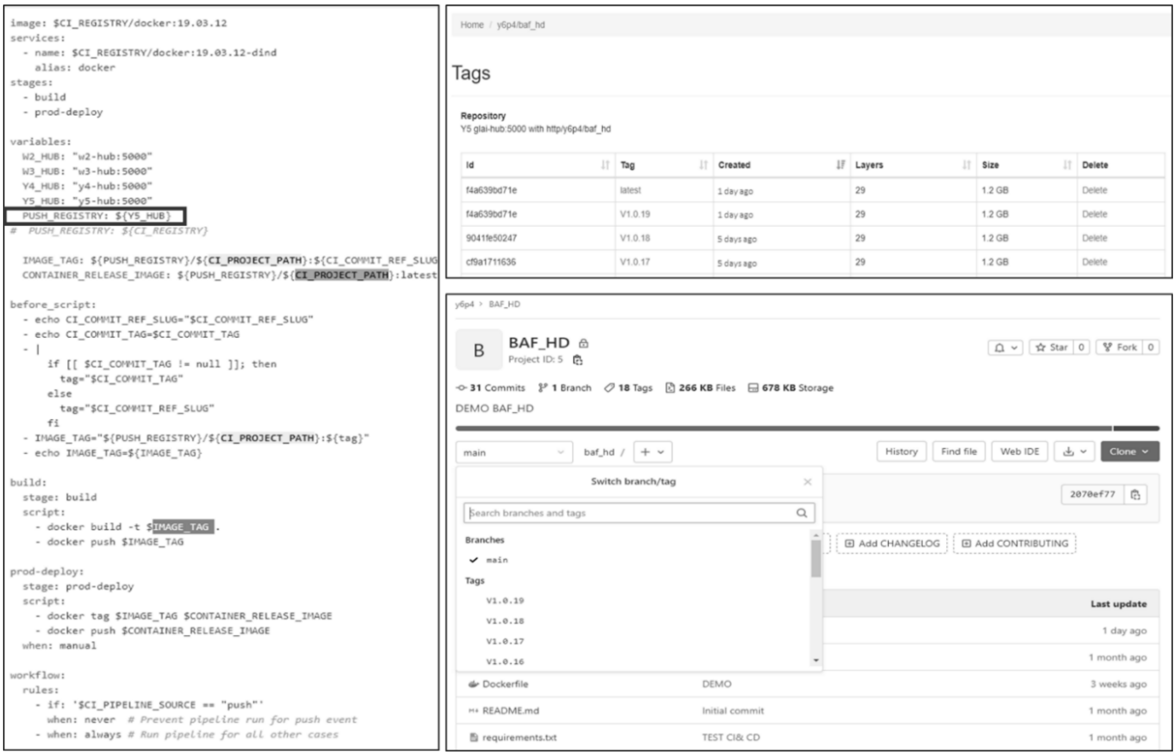

Figure 2. The configuration and source code versions of an AI model operation

The model management regards how to manage outdated AI model and replace it with another retrained AI model when operation engineers maintain the existing online AI models. The AIMOSS monitors the quality of AI model performance by utilizing three components, namely predicting, validating, and retraining (Table 1). When an AI model is deployed on the CSC's AIoT platform, the intelligent application starts providing corresponding services such as the trend of production parameters. Through the AIMOSS, operation engineers can manually call a specific online AI model, input a test dataset, and check the output results for the model performance inspection and other test purposes. Operation engineers can also set designated conditions for automatically validating the online AI model performance by the AIMOSS. For example, the MAE of an online AI model must be greater than 1 and the performance indicator needs to be checked repeatedly every 1 day based on a designated source as the training dataset.

The AIMOSS can automatically address that the performance indicator of an online AI model drops down through repeated validation cycles. A pre-determined script is then activated for conducting the procedure of retraining the AI model affected by the data drift. At the same time, operation engineers receive related notifications. If the event about the quality of the online AI model cannot be automatically solved by the AIMOSS, operation engineers can promptly intervene in the retraining. All functions and components are designed to keep the existing online AI models functioning normally.

Table 1. Three main components of AIMOSS

Component	Description
Predicting	Services for providing prediction results of the online AI

	models
Validating	Quality monitoring by automatically validating the online AI models
Retraining	Model drift correction by automatically retraining the online AI models

Results

A pilot case was selected to evaluate the effectiveness of the AIMOSS. In this case, an AI model was designed to provide predictions of soaking time in a batch annealing furnace (BAF) (Figure 3). The process in the BAF provides a heat treatment for cold-rolled steel coils to relieve mechanical stress and achieve specific mechanical properties. Production engineers can refer to the result of this intelligent application and adjust corresponding production parameters for meeting customers' requirements.

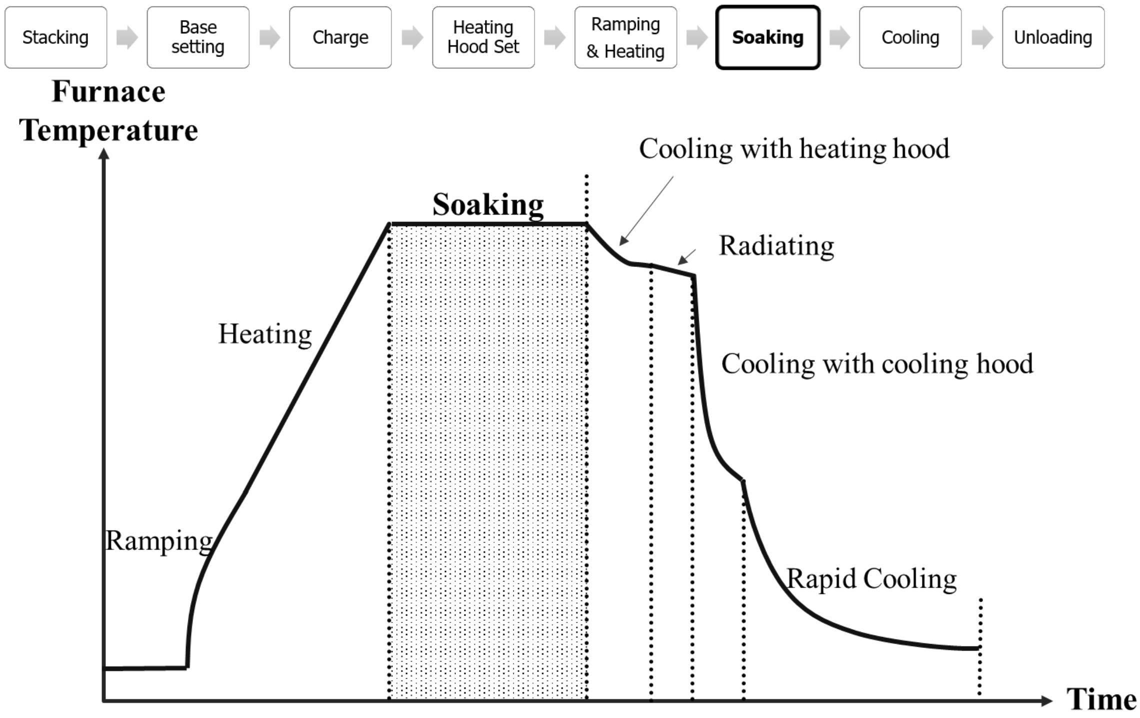

Figure 3. The production process of batch annealing furnace (BAF)

In this case, the input data for the AI model comes from different IoT devices and databases across various production fields. The target mechanical properties, for example, are hardness and yield strength. The job of the AI model is to infer the doable soaking time of cold-rolled steel coils for meeting the requirement of hardness and yield strength. By creating a pipeline of the AI model, its source code and required environment configuration were transferred to the AIMOSS first (Figure 4). Then, the AIMOSS automatically prepares the corresponding runtime environment image and creates a container on the CSC's AIoT platform.

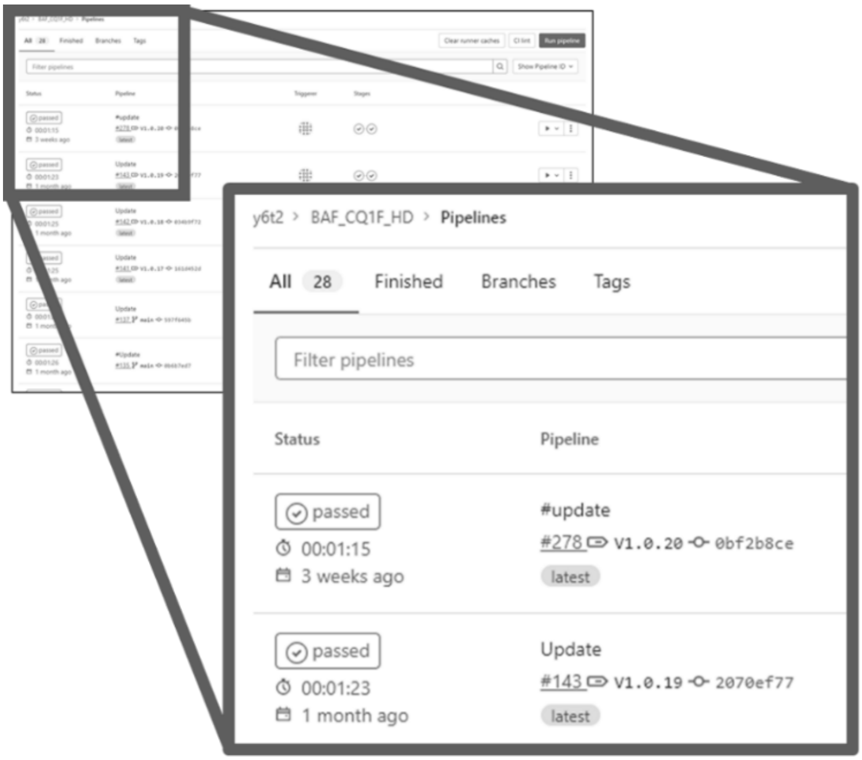

Figure 4. AI model operationalization – Build a pipeline

Next, the AIMOSS runs the contain up according to the AI model operation parameters determined by operations engineers including the training dataset, the validation dataset, validation rules, and notification settings. The container runtime is a component in a Kubernates (K8S) node. Operation engineers can monitor the status of the container via Rancher's web-based tools. Each container of validating and retraining the AI model showed in a list of historical records (Figure 5). Predicting results can be presented in real-time execution logs with the web-based command line interface (Figure 6).

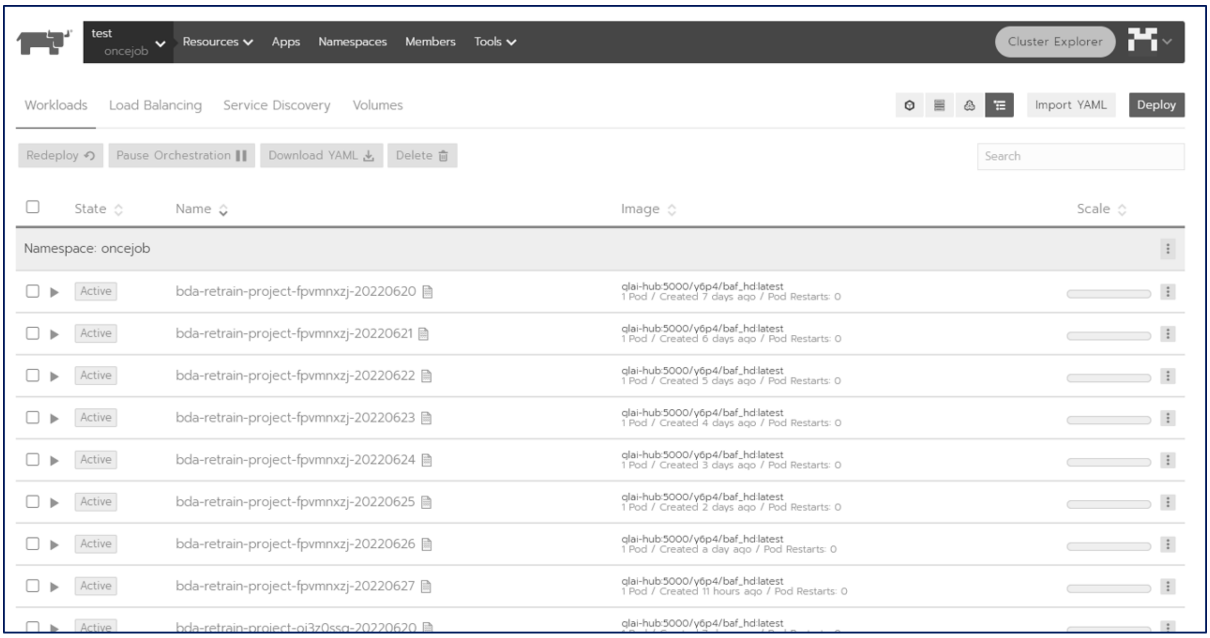

Figure 5. AI model operationalization – Logs of retraining

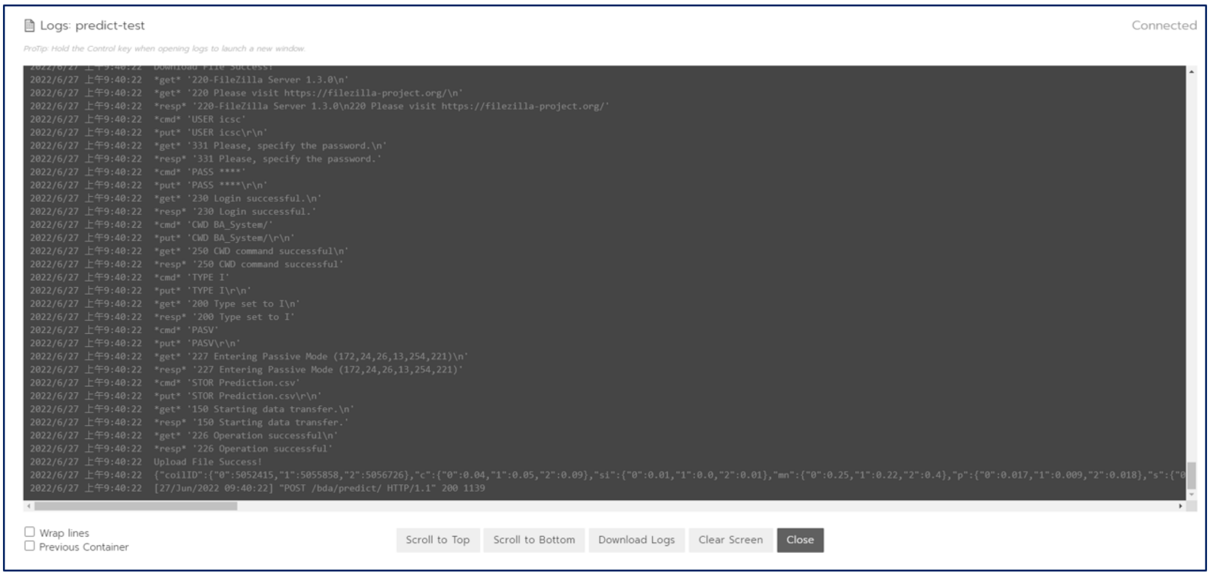

Figure 6. AI model operationalization – Logs of predicting

Discussion

The online AI models were deployed on workstations of different IT facilities located in production fields. The computation resource of each workstation is limited which makes a retraining process usually takes up to one hour. Operation engineers need to monitor the online AI models on their own. However, they would sometimes be occupied by other tasks and might takes more time to be aware of AI model performance changes. When they identify that a performance indicator of an online AI model is dropping down, a retraining process still must be conducted by them and followed by a manual deployment to the platform.

After applying AI model management on the CSC's AIoT platform, the HCI provides more flexibility in computation resources and shortens the retraining process down to 10 minutes approximately. When the AIMOSS finds out an outdated AI model because of performance issues, it can automatically use the latest training dataset to retrain the AI model. A new version of the AI model is then created and deployed on the CSC's AIoT platform through the pre-defined pipeline. The AIMOSS monitors the online AI models more comprehensively. Each version of an online AI model is recorded clearly, and the version control mechanism makes the switching process in between different versions more easily.

Based on interview results, the feedback of operation engineers can be summarized into two major points. The technologies used in the AIMOSS need to be delineated in more detail. For example, understanding the limitations of the HCI and containers can help operation engineers determine the AI model operation parameters. Moreover, making the web-based command line interface interactive brings benefits of executing more management commands. The efficiency of debugging could be further improved with more flexibility.

Summary

This paper presents the practice and progress of developing and deploying AI applications on the CSC's AIoT platform. CSC promotes smart manufacturing such as implementing the AIMOSS and utilizing the 5G AIoT hybrid cloud infrastructure. CSC encourages employees to utilize the new enabling technology, enhance their abilities, and improve their outcomes. On the other hand, smart manufacturing accompanies more intelligent applications and more loads of daily operations. With the support of AIMOSS, the CSC's AIoT platform becomes more capable of operationalizing a large number of intelligent applications.

References

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems, 25*(sup1), 64-75. https://doi.org/10.1080/12460125.2016.1187397

Gartner. (2021). Top trends in data and analytics for 2021: XOps Retrieved 22 August 2022, from https://www.gartner.com/document/code/739072

Hechler, E., Oberhofer, M., & Schaeck, T. (2020). The Operationalization of AI *Deploying AI in the Enterprise: IT Approaches for Design, DevOps, Governance, Change Management, Blockchain, and Quantum Computing* (pp. 115-140). Berkeley, CA: Apress. doi: https://doi.org/10.1007/978-1-4842-6206-1_6

Hummer, W., Muthusamy, V., Rausch, T., Dube, P., Maghraoui, K. E., Murthi, A., & Oum, P. (2019). ModelOps: Cloud-based lifecycle management for reliable and trusted AI. In I. Brandic & D. Milojicic (Eds.), *Proceedings of 2019 IEEE International Conference on Cloud Engineering (IC2E)* (pp. 113-120). Prague, Czech Republic: IEEE Computer Society.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly, 37*(3), 101493. https://doi.org/10.1016/j.giq.2020.101493