Governed MLOps Workshop
Introduction

Document version: May 2023

DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift is a trademark of Red Hat, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Table of Contents

Introduction

As enterprises invest in developing and adopting AI in their business, they need a framework or an approach to follow so they can extract value from their AI investment. The AI Ladder, described in Figure 1, is a prescriptive approach on how to achieve business value from AI and consists of the four rungs:

- Collect: Collect data of every type (structured/unstructured) regardless of where it resides enabling access to ever-changing data sources in a hybrid cloud environment where data may exist on-prem, in a public cloud platform, or may even be 3rd party data.
- Organize: Organize all data into a trusted, business-ready foundation to meet enterprise governance and compliance requirements.
- Analyze: Build and scale AI models with trust and transparency.
- Infuse: Operationalize AI throughout the business by infusing AI models in business processes and customer interactions.

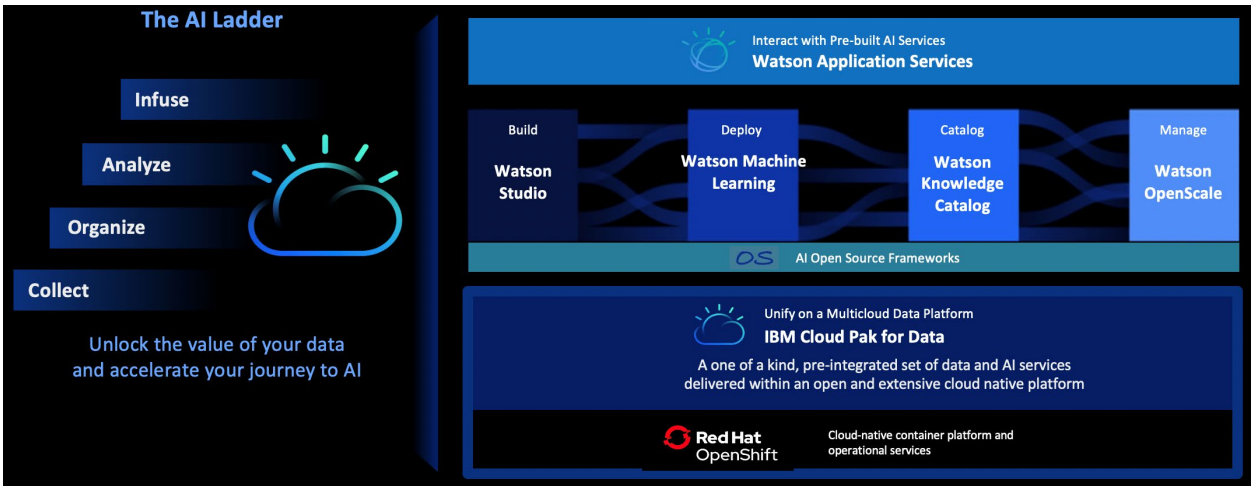

Figure 1: The AI Ladder - a prescriptive approach for organizations to transform their business by connecting data and AI

Cloud Pak for Data is IBM's Data and AI integrated platform that enables organizations to execute on the AI Ladder approach and build AI models powered by a solid data foundation. It is composed of several containerized software services as detailed in the documentation.

MLOps is a set of processes, best practices, and technologies to enable the development, deployment, maintenance, and management of machine learning (ML), and more generally AI, models in production reliably and efficiently.

Governed MLOps is applying governance across the MLOps lifecycle to meet the governance and compliance requirements of the enterprise. Note that data governance is only one component of delivering a completely governed MLOps solution. This workshop describes several governance requirements that are required to implement a comprehensive Governed MLOps approach.

In this workshop, you will learn how Cloud Pak for Data supports a governed MLOps methodology enabling enterprises to adopt the AI Ladder prescriptive approach to develop, trust and scale AI adoption. The workshop steps through the various modules of the Governed MLOps methodology to address the business use case of Telco Customer Churn prediction using Cloud Pak for Data and the integrated data and AI portfolio of offerings.

Use Case

XYZ, a Telco company, is interested in developing an insights platform to better understand and minimize churn rate among its customers and deliver personalized customer experience. Churn rate is the rate at which customers stop doing business with an entity either by cancelling their service or changing their providers. XYZ company collects data about their customers and would like to train AI models for predicting churn and personalizing their customers' experience based on churn likelihood.

Proposal

To help XYZ achieve their business goals of personalizing customer interactions and minimizing churn, we propose the following solution outlined in Figure 2. Effectively, as a customer engages with the virtual assistant, the assistant personalizes its responses by leveraging real time predictions for the likelihood of the customer churning. The virtual assistant accesses these predictions by calling an AI model trained in Watson Studio and deployed at scale using Watson Machine Learning; both services available in Cloud Pak for Data.
Note that the workshop does not include the technical instructions to setup and deploy the virtual assistant component but that is something you can leverage Watson Assistant tutorials for.

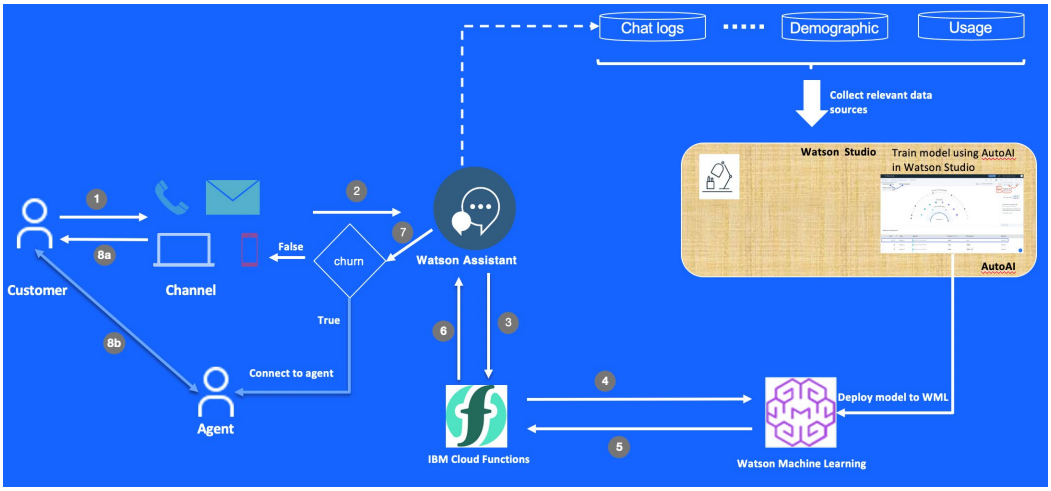

Figure 2: Personalized Customer Experiences Based on Churn Prediction

PreRequisites

Please make sure you review and complete the prerequisites before starting with the workshop steps below. NOTE: If you're following these instructions as part of a scheduled workshop, you can skip Prerequisite 1 as the environment will be provided to you by the instructors.

Prerequisite 1 – Cloud Pak for Data Environment TechZone: Reserve a Cloud Pak for Data environment on TechZone. You can use your own Cloud Pak for Data environment provided it has all the requires services (Data Virtualization, Db2, WKC, Watson Studio, AutoAI, Watson OpenScale, OpenPages with Watson)

PreRequisites MLOps Workshop - Setup Users Catalogs Data: Execute the instructions in this prerequisite to configure and setup your Cloud Pak for Data instance with the right users, catalogs, and permissions to be able to run through the different steps in the workshop.

Figure 3 outlines a more detailed view of how the various components of Cloud Pak for Data enable enterprises to establish a governed data foundation to leverage for developing and deploying trustworthy AI models that can be infused in customer interactions for personalizing customer experience. Left to right, Figure 3 depicts how data is collected from various sources using data integration capabilities (DataStage) and organized in a catalog to meet quality, privacy, lineage, and governance requirements (Watson Knowledge Catalog). At that point, data is consumed to train, deploy, and monitor AI models (Watson Studio, Watson Machine Learning, OpenScale). Lastly, deployed AI models are integrated with other applications such as Watson Assistant to personalize customer experience.

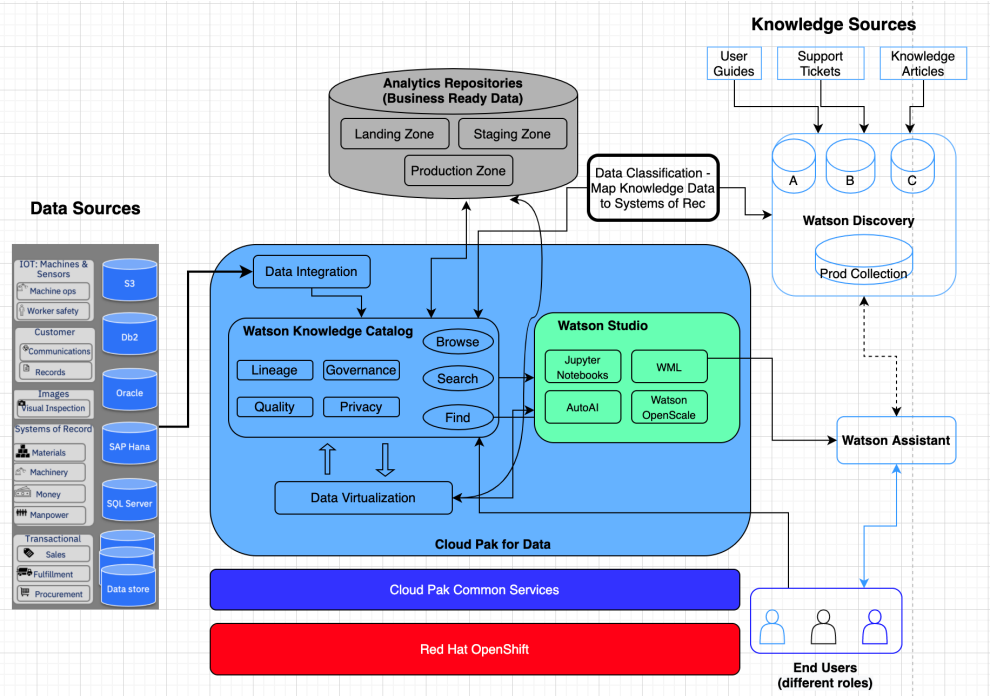

Figure 3: Governed MLOps Workshop Architecture

While the focus of this workshop is Governed MLOps, it is important to understand the activities required to access, prepare, and organize data so it is business ready because having high quality and well-governed data is critical for training useful and trusted AI models. To best represent that, we consider the roles and tasks of data providers and data consumers:

- Data providers: Data providers focus on collecting and organizing the data so it is business ready. Data providers typically include the roles of data engineer, data steward, and data quality analyst. Data engineers collect data from various data sources. Data stewards define governance artifacts (business terms, data classes, rules, policies, ...) to organize and govern the accessed data according to the enterprise regulatory and compliance requirements. Data quality analysts discover and analyze the data to evaluate its quality and readiness for business use. Once data quality meets the required specification, data is published to the catalog where it is available for data consumers to find, access, and leverage for various purposes.
- Data consumers: Data consumers leverage the business ready data from the catalog to gather business insights, train AI models and embed in applications. Data consumers typically includes the roles of data scientists, business analysts, and developers. Business analysts leverage cataloged data to create business intelligence dashboards and reports to highlight business insights. Data scientists use the business ready data to train AI models for various purposes. Developers build applications that consume the data for different use cases.

Additionally, IBM's Data Fabric point of view supports multiple entry points including:

- Data Governance and Privacy
- Trustworthy AI

This workshop is designed to cover both entry points as outlined in Figure 4 and allows participants to focus on one or the other or both.

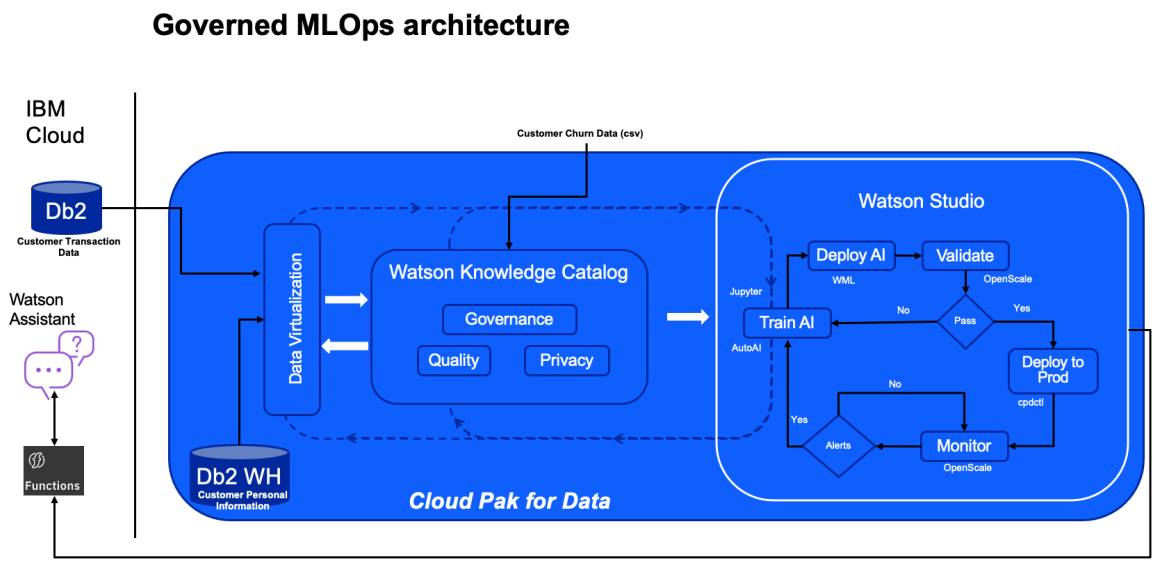

Figure 4: Governed MLOps Architecture

Data Governance and Privacy

Data providers are responsible for collecting and organizing the data so it is business ready for consumption by data scientists, business analysts, and developers. This section describes the hands-on

modules which highlight the tasks involved in collecting and organizing the data which are typically executed by data provides.

1. Setup Data Governance foundation (Data Providers – Data Steward): Data governance is a critical first step in preparing for AI model development. Before Data Stewards and Data Quality Engineers can start creating governance artifacts and curating data, data governance foundation needs to be setup. IBM's data governance tool is Watson Knowledge Catalog (WKC) which is an intelligent data catalog that powers self-service discovery of data, models and more. To setup data governance, you must create governance categories, assign Watson Knowledge Catalog roles to users, add users to categories, and set up workflow configurations. Governance is provided by governance artifacts which are organized by categories. The processes for creating, updating, and deleting governance artifacts are controlled by governance artifact workflow configurations. Governance artifacts are created and assigned to data assets by users who are collaborators in categories. In this lab, you will step through sample workflows to setup governance foundation.

2. MLOps Workshop Data Virtualization (Data Providers – Data Engineer)

   Leverage Data Virtualization capabilities in Cloud Pak for Data to connect to various data sources and virtualize relevant data assets for purposes of data visualization, extracting insights, and training churn prediction models.

   Connect to data sources across a hybrid cloud environment where data resides on a public cloud and/or on-premises.

   Get relevant data that can be helpful in predicting customer churn. In this lab, we use two types of data assets:

   Customer data stored in an on-prem relational database; in our example, it is stored in a Db2 table on the same Cloud Pak for Data cluster.

   Customer transaction data stored in a managed Db2 instance on IBM Cloud.

   These data assets emulate data sources in different environments. Cloud Pak for Data supports a rich set of data sources as referenced in the documentation. Please note that the data is critical for the accuracy of the trained AI models. Different organizations may leverage different data sets to use for churn prediction.
   In this lab, you will learn how to virtualize the data sets referenced above.

3. Data Quality and Data Privacy with Watson Knowledge Catalog (Data Providers – Data Quality Analyst): Collected data assets are governed using Watson Knowledge Catalog capabilities to make sure enterprise governance and compliance rules are enforced. There are use cases where it makes sense to connect to data assets directly and leverage such data for extracting insight and training AI models. However, in general, it is recommended to catalog all data assets and only leverage catalogued data for subsequent tasks such as business intelligence insights and training

AI models. Cataloging the data ensures that the enterprise's governance and compliance rules are being enforced and trusted data is delivered.
In this lab, you will learn how to leverage Watson Knowledge Catalog capabilities for auto discovery of data, assessing data quality and masking data to meet data privacy requirements.

Trustworthy AI

For purposes of the Governed MLOps workshop, our focus is on the Trustworthy AI entry point and we assume that the relevant data assets have been identified, cleansed, quality-verified, and stored in data stores/catalogs; ready to be consumed. If you would like to explore the various tasks involved in organizing and governing the data, which is the focus of the Data Governance and Privacy entry point, please check the modules referenced in the Data Governance and Privacy section of this document.

The workshop consists of the following hands-on modules:

4. AI Governance (Data Consumers – AI Model Owners)

   This module steps through configuration of OpenPages with Watson to define and follow workflows for proposing models, approving models, and triggering development/deployment of models all while meeting enterprise governance and compliance requirements.

5. Train machine learning models to predict likelihood of churn (Data Consumers – Data Scientist)

   Leverage IBM's AutoAI capabilities to speed up exploration and identification of best ML algorithms for the given data set.

   Deploy best performing AutoAI model to a REST API endpoint with Watson Machine Learning.

   Alternatively, train AI models using Jupyter notebook and leveraging open-source libraries.

   In this lab, you will learn how to leverage AutoAI to automate much of the process of training AI models by automatically exploring different data processing methods, feature engineering techniques, and machine learning algorithms to find best performing models. You will select one of the trained models, typically the best performing model, and deploy it in Watson Machine Learning for online inference where predictions can be triggered via REST APIs.

   You will also step through training an AI model using a Jupyter notebook approach which is a very popular approach for data scientists.

6. Automation with Watson Pipelines (Data Consumers – Data Scientists / MLOps)

   In this module, you learn how to leverage Watson Pipelines to automate the tasks of shaping data, training models, deploying models in development deployment spaces and selecting best performing models to propagate to preproduction deployment spaces. Deploying with Watson Pipelines enables automation and scheduling of jobs to run periodically to capture updated data and re-train models accordingly.

7. Validate and Monitor AI models with Watson OpenScale to build and scale AI trust and explainability (Data Consumers – Data Scientists / MLOps)
   In this lab, you will step through configuring Watson OpenScale to monitor the AI models in production for fairness, drift and explainability. To deliver trust in AI models, it is important for the business leaders to have the confidence that the AI models are fair and explainable. Fairness in AI models describes how evenly the model delivers favorable outcomes between the monitored group (the group potentially susceptible to be biased against) and the reference group. Generally, references to fairness and bias are used to describe similar behavior of the AI models but in an inverse manner. A model that is fair is not biased and a model that is biased is not fair. As for explainability, it refers to the ability to describe how the model determined the prediction and what were the most important factors that led to that prediction specifically for the given transaction. In addition to fairness and explainability, Watson OpenScale measures model quality and drift. After AI models are deployed in production, they may drift over time signaling a drop in accuracy or in data consistency and it is important to detect potential drift early and trigger a re-training as needed.

8. Governed MLOps CI/CD (Data Consumers – Data Scientists / MLOps)
   In this lab, you will learn how to leverage cpdctl tool to copy assets from one deployment space (for example, QA deployment space) to another deployment space (for example, prod). As organizations scale adoption of AI models in production, it becomes more important to automate the process for testing, validating, and promoting such models from dev (development) to UAT (user acceptance testing, also known as pre-prod, quality assurance or staging) to prd (production) environments. The cpdctl tool is a command line utility that enables automation of the process of copying assets from one environment to another. In practice, the environments can exist in the same Cloud Pak for Data cluster or in completely different Cloud Pak for Data clusters hosted on different cloud platforms. For this lab, we will use the same Cloud Pak for Data cluster and illustrate how to use the cpdctl tool to propagate assets from the quality assurance (QA) deployment space to the production deployment space. The process is identical to how you'd propagate models from one cluster to another as the cpdctl tool is designed to handle the hybrid multi-cloud environments seamlessly.

9. Infuse AI Model into Cloud Native Application (Infuse – How to consume model into applications – Developer)
   Integrate churn prediction results with a cloud native application developed to run on Openshift using Cloud Native toolkit. Please note that the Openshift licenses that are included with Cloud Pak for Data licenses are restricted licenses which mean that can only be used to leverage Cloud Pak for Data workloads. In this lab, we demonstrate technical solution for developing a cloud native application that consumes the churn prediction AI model. However, deploying such a cloud native application on OpenShift requires separate licensing of OpenShift.

Summary

This introductory module captures the end-to-end view of the use case addressed by this workshop as well as the individual modules and how they align to IBM's Data Fabric entry points. You will gain experience visualizing how the model is tracked through its lifecycle using AI Factsheets. Additionally, you will learn how to leverage OpenPages for tracking the model AI lifecycle and confirming it meets the organization's compliance requirements by defining tasks and owners.