## [CI / CD / CM with ML Works](https://www.tredence.com/blog/ci-cd-cm-with-ml-works/)

Most of us are familiar with Continuous Integration (CI) and Continuous Deployment (CD) which are core parts of MLOps/DevOps processes. However, Continuous Monitoring (CM) may be the most overlooked part of the MLOps process, especially when you are dealing with machine learning models.

CI, CD and CM, together, are an integral part of an end-to-end ML model management framework, which not only helps customers to streamline their data science projects, but to also get full value out of their analytics investments. This blog focuses on the Continuous Monitoring aspect of MLOps and gives an overview of how Tredence is using ML Works, a model monitoring accelerator built on Databricks’ platform, to help customers build a robust model management framework.

Here are a few examples of MLOps customer personas:

**1.Business Org** – A business team, which sponsors an analytics project will have the expectation that machine learning models are running in the background, helping them to get valuable insights from their data. However, these ML models are mostly in a black box and in a lot of cases, the business sponsors are not even sure if the analytics project will lead to a good ROI.

**2.IT/Data Org** – A company’s internal IT team, which supports business teams usually has a team of data engineers and data scientists who build ML pipelines. Their core mandate is to build the best ML model and migrate them to production. When doing so, they’re either too busy building the next best ML model to put it into production or managing production model support is not the right use of their time. Hence, there is a lack of streamlined model monitoring process in production and IT, and data leaders are left wondering how to support their business partners.

**3.Support Org** – A company has an IT support organization, which takes care of supporting all IT issues. This team likely treats all issues the same, including similar SLAs, and may not differentiate between supporting an ML model and a Java web application. Hence, a generic support team may not have the right skills to support ML models and may not be able to meet the expectations of their internal customers

A well-designed MLOps framework will address the challenges of all three personas.

Tredence not only has multiple experiences in end-to end MLOps implementations across tech stacks but has also built MLOps accelerators to help customers gain the full potential of their analytics investments.

Let’s drill down on our model monitoring accelerator in the Continuous Monitoring (CM) space and talk about the offer in more detail.

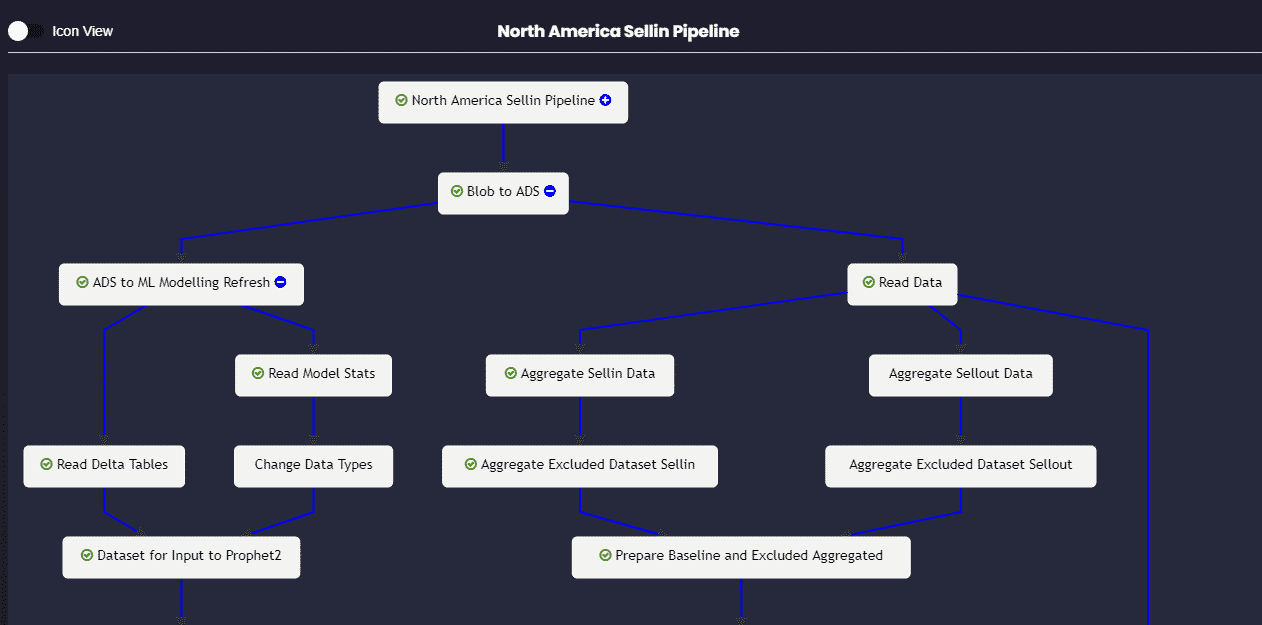
**Model monitoring is not easy!**

Unlike monitoring a BI dashboard or an ETL pipeline, the biggest challenge with ML models is that their results are probabilistic in nature and have their own dependencies like training data, hyper parameters, model drift, and the ability to explain the output of the model results. As a result, complications increase, and model monitoring becomes almost impossible when models are built on unstructured notebook formats that are used across multiple data science teams. This severely impacts Support SLAs and results in business users gradually losing confidence in the model’s predictions.

**ML Works to the rescue**

ML Works is our model monitoring accelerator built on [Databricks’ unified data analytics platform](https://databricks.com/) to augment our MLOps offerings. After evaluating multiple architectural options, we decided to build ML Works on Databricks to leverage Databricks’ offerings like [Managed MLflow](https://databricks.com/product/managed-mlflow) and [Delta Lake](https://databricks.com/product/delta-lake-on-databricks). ML Works is trained on thousands of models and can handle Enterprise scale model monitoring, or it can be used for automated monitoring within a small team of data scientists and analysts. Here is an overview of ML Works core offerings:

**1.Workflow Graph** – Monitoring a ML pipeline along with its relevant data engineering tasks can be a daunting task for a support engineer. ML Works uses Databricks’ managed ML flow framework to build a visual end-to-end workflow monitor for easy and efficient model monitoring. This helps support engineers troubleshoot production issues and narrow down the root cause faster, significantly reducing Support SLAs.

[](https://www.tredence.com/wp-content/uploads/2020/06/workflowgrapph.png)

*Figure 1 – Visual Workflow Graph to monitor end to end Model Pipeline*

**2.Persona-based Monitoring** – We understand that a ML model monitoring process should not only make the life of a support engineer easier but also help other relevant persons like business users, data scientists, ML engineers and data engineers to get visibility into their respective ML model metrics. Hence, we have built a persona-based monitoring journey using Databricks’ managed ML flow to make the model monitoring process easy for all personas.

[](https://www.tredence.com/wp-content/uploads/2020/06/MicrosoftTeams-image-2.png)

*Figure 2 – Persona based model monitoring dashboard*

**3.Lineage Tracker** – Picking up the task of debugging someone else’s ML code is not a pleasant experience, especially when there isn’t good documentation. Our Lineage Tracker uses Databricks’ managed ML flow and helps customers start from a dashboard metric and drill all the way to the base ML model, including the model’s hyper parameter values, training data, etc. thus giving full visibility into every model’s operations. This gets all relevant details about a model in one place, which improves model traceability. This feature is further enhanced when we use Delta Lake’s Time Travel functionality to create snapshots of training data

**3.Drift Analyzer** – Monitoring the model’s accuracy with time is critical for business users to gain trust in the insights. Unfortunately, a model’s accuracy will drift with time for various reasons including production data changing over time; business requirements changing and making original features no longer relevant and acquiring a new business which introduces new data sources and new patterns in the data. Our Drift Analyzer analyzes the Data Drift and Concept Drift automatically by reviewing the data distributions, which triggers alerts if the drift has exceeded a threshold and ensures that production models are continuously monitored for accuracy and relevance.

Using ML Works, business teams are able to monitor and track their relevant metrics on the Persona Dashboard and use Drift Analyzer to understand the impact of model degradation on metrics. This will help them to look at the underlying ML models as a white box solution. Lineage Tracking helps data engineers and data scientists obtain end-to-end visibility into ML models and their relevant data pipelines, which streamlines development cycles by taking care of the dependencies.

Support teams can use Workflow Graph and relevant metrics to troubleshoot production issues faster, significantly reducing Support SLAs. And finally, customers can now get full value from their analytics investments using ML Works, while also ensuring that ML deployments in production really work.

Let’s look at a specific use case of ML Works integration with Azure ML Ops in the next part of this blog series.

Author: [**Pavan Nanjundaiah**](https://www.linkedin.com/in/pavannanjundaiah), *Head of Solutions*