IMLA – A machine learning platform to democratize ML modeling in ML-APC

Duo(Jason) Zhang/Intel Corporation
Runshen Xu/Intel Corporation
Ali Ahmadi/Intel Corporation
Gobind Bisht/Intel Corporation
Rajan Sawhney/Intel Corporation

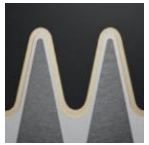
APCSM Conference 2022

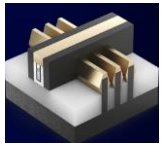
Why ML-APC

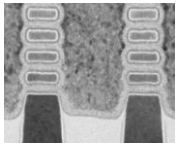
Manufacturing Roadmap


Strained Silicon


Hi-K Metal Gate

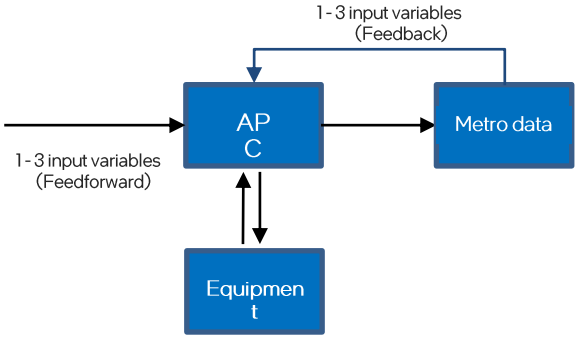

First FinFET


SuperFin


RibbonFET

Leading Edge Process : smaller process margin

Traditional APC

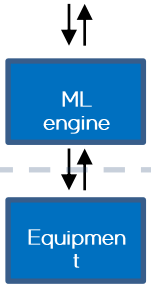


ML-APC

Data Driven: Fully utilize process data

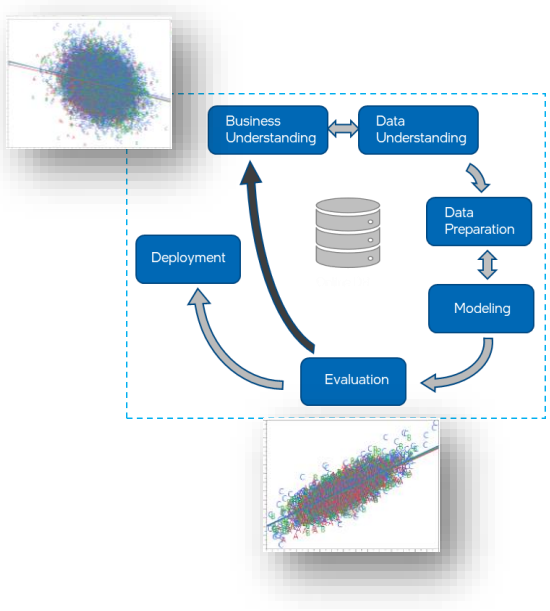
Dry Etch | Litho | Thin Film | Metro | Wet Etch | CMP

MLC (ML control):
- GBT
- REGRESSION
- Random forest



ML-APC Highlights
- Data Driven + ML-APC: Improve process margin.
- Accelerate process dev and optimization.
- Yield and Capacity.

Wafer Level ML-APC Workflow



Problem Statement

Initial Project Review

Problem Statement

Usecase review

ML Model Developmen t

ML-APC App Developmen

ML-APC App Deployment

Wafer Level ML-APC Application

Feedforwards → RequestSettings → Data Shaping → ML Prediction → PostMLCal

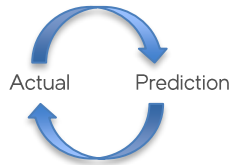
Data validation

spec check

ML Calculation for both sparse and non-sparse data

wafer 1 : var 1, var 2, var 3 ... var 14 --> predicted result for wafer 1
wafer 2 : var 1, var 2, var 3 ... var 14 --> predicted result for wafer 2
wafer 3 : var 1, var 2, var 3 ... var 14 --> predicted result for wafer 3
wafer 4 : var 1, var 2, var 3 ... var 14 --> predicted result for wafer 4

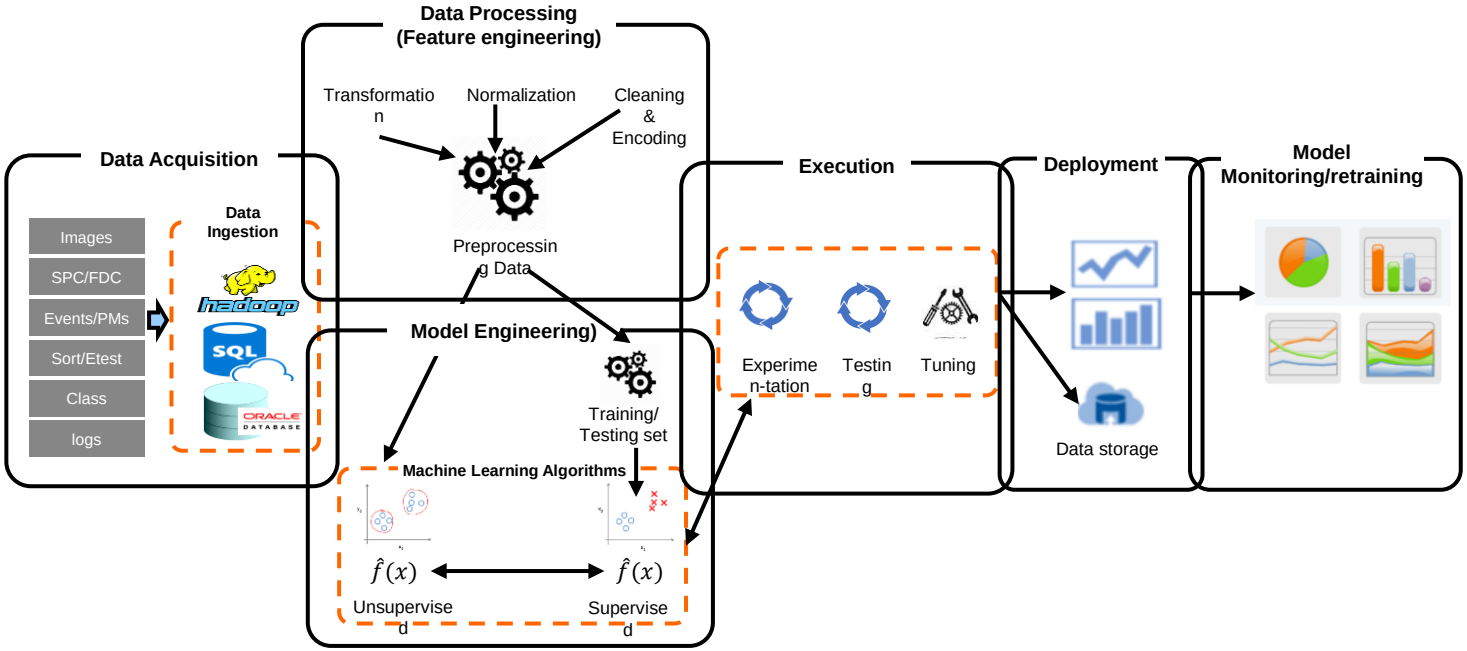
Model Performance Monitor



IMLA: Custom MLOps Platform

- **What**: a set of practices that aims to ***develop***, ***deploy*** and ***maintain*** machine learning models in production ***reliably*** and ***efficiently***
- **Why**: we are seeing a huge surge in using machine learning for decision-making, and business improvement. MLOps means **faster go-to-market** times and **lower operational costs**.
- **How**: an **in-house web-based** solution that automates and streamlines ML workflow and allows data scientists and non-experts to develop and deploy ML models
  - Integrated with Intel POR data sources:
  - Easy integration for custom data pre-processing algorithm
  - Easy integration with both Intel In-House and Open Source ML Algorithm
  - Integrated to Intel online and offline systems

Machine Learning Lifecycle



Machine Learning Lifecycle

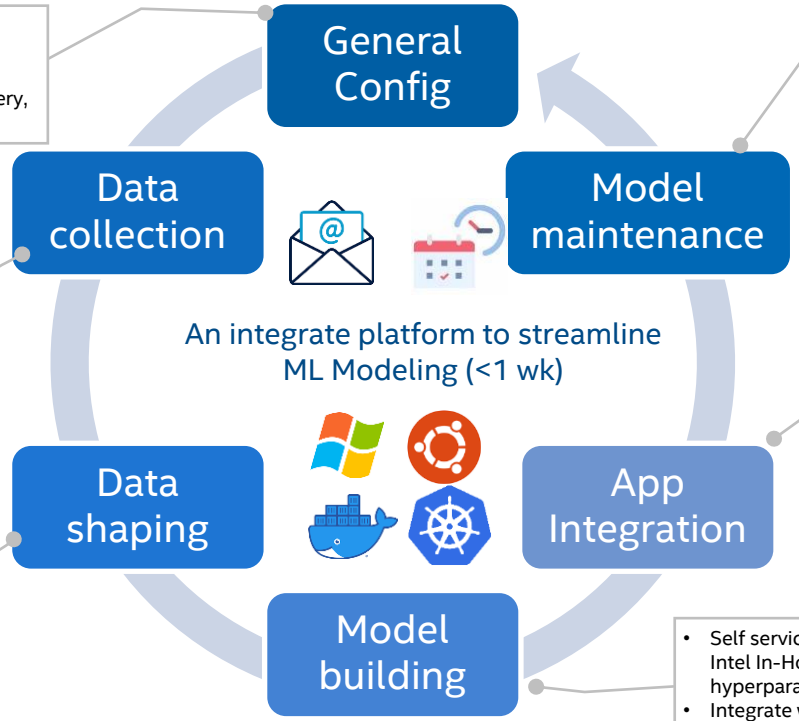
Task (Proportion of Effort)	Subtasks
Problem understanding (5% - 10%)	a) Determine the objective ; b) Define success criteria c) Assess constrains
Data understanding (10% - 25%)	a) Assess data situation; b) Obtain data access c) Explore data
Data preparation (20% - 40%)	a) Clean data ; b) Feature engineering
Modeling (15% - 20%)	a) Select model approach; b) Build models c) Tune models
Evaluations of Results (5% - 10%)	a) Select Model ; b) Validate Model c) Explain Model
Deployment (5% - 10%)	a) Deploy Model; b) Maintain and Monitor
Model monitoring and retraining (10% - 15%)	a) Monitor model and send notification; b)re-train outdated models

~ 50%

~ 20%

~ 20%

IMLA – Highly Extensible and Horizontally Scalable



- Web UIs with security management
- Rapid configuration setup by operations and data category (Process, SPC, APC, etc.)
- Support customized scripts including Python, sql query, json, etc.

- Support both ad-hoc and scheduled prediction requests
- Rich statistics and visualization of prediction results
- Monitor prediction performance
- Model refresh based on the process state and model performance

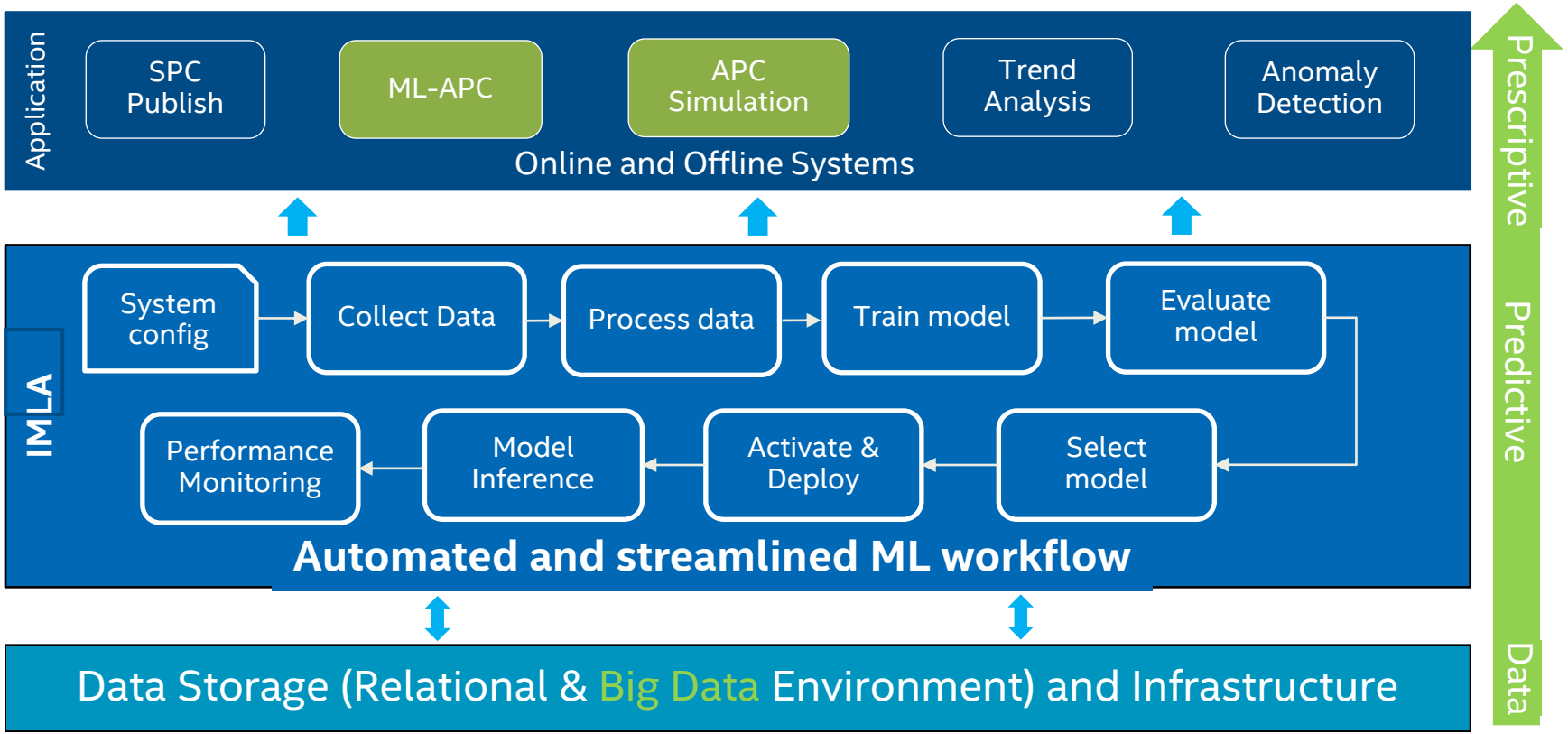
- Extensive data connector to support multiple data sources, including customized csv files, Oracle RDB, MS RDB, NoSQL, etc.
- Automate DB-based data collection and store the dataset to be reused in modeling cases

- Integrate models from IMLA with other applications, including APC online, simulation, SPC, predictive maintenance, etc.

- GUI based canvas for users to build highly customized data processing pipeline with the maximum flexibility
- The created data model is used for prediction as well as data processing in other applications, such as ML-APC.

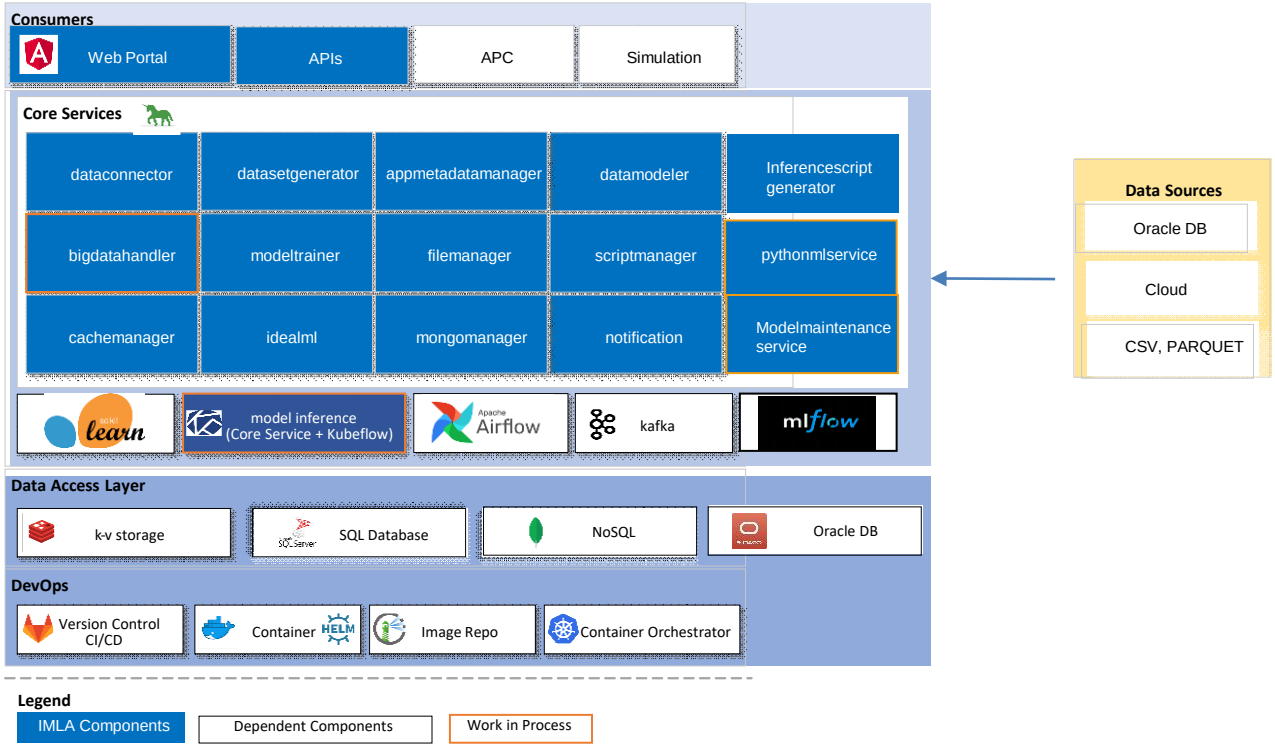
- Self service modeling (Several ML estimators of Intel In-House and open-source packages, hyperparameters, selection criteria, etc.)
- Integrate with MLFlow to perform deep analysis

IMLA Architecture



IMLA Architecture

- Supports Intel In-House Algorithm, Sklearn and any open-source ML library
- Airflow for job scheduling and monitoring
- Several layers of caching
- Model versioning and inference scheduling
- Containerized which for automated deployment
- User-friendly web interface

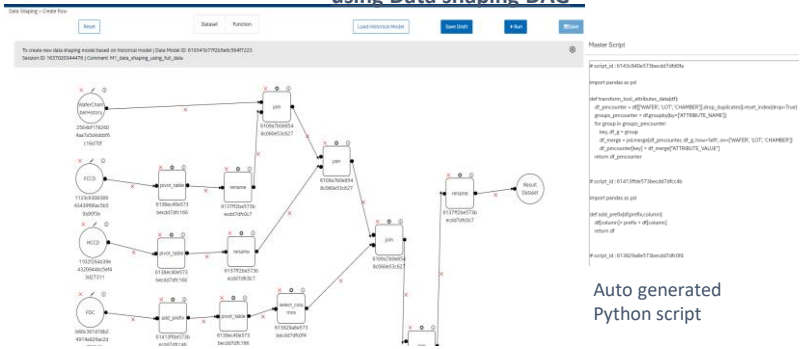


Use Case

Step1: Submit data collection job

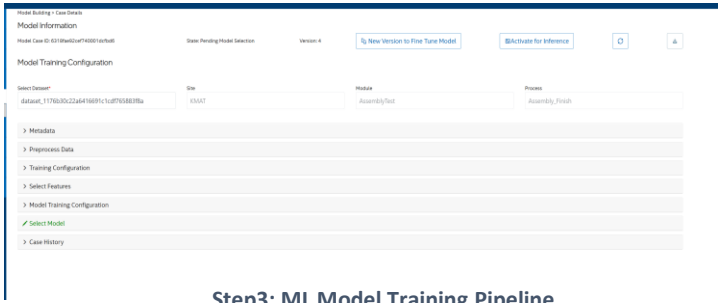
Actions	ID	WorkspaceID	DataCategory	Comment	CreatedDate
	894	M1_PS	FDC		2022-01-12 09:39:49 PM PST
	895	M1_PS	SPCContext		2022-01-12 10:28:27 PM PST
	896	M1_PS	WaferChamberHistory		2022-01-12 10:28:28 PM PST
	903	M1_PS	SPCContext		2022-01-18 05:42:14 PM PST

Step2: Custom data processing using Data shaping DAG

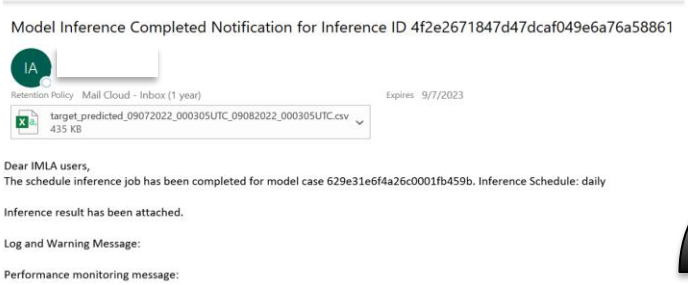


Auto generated Python script

Step3: ML Model Training Pipeline

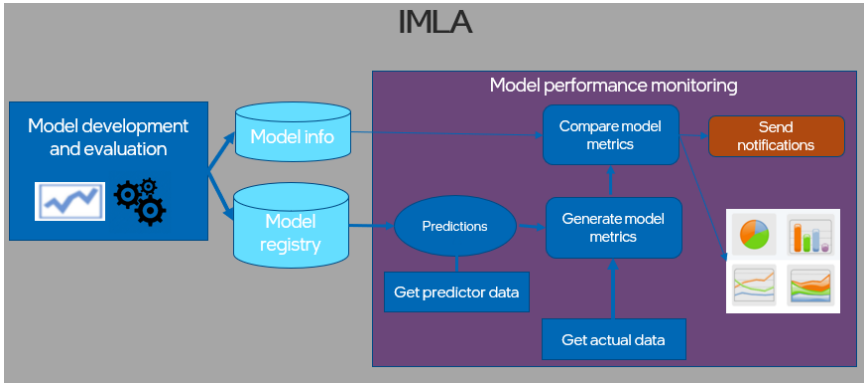


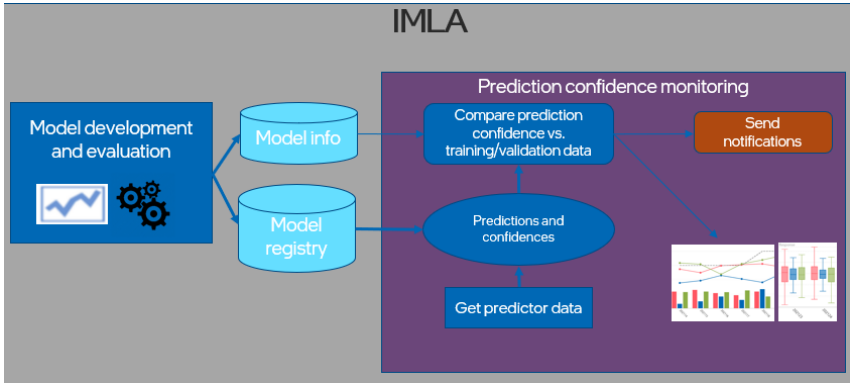
Step4: Schedule Inference to automatically receive latest prediction



Step5: Integrate models in APC app

IMLA Performance Monitoring Flow





- IMLA proactively monitors model performance by prediction confidence even if there is a delay in obtaining actual target values
- When actual target is available (metro target e.g.). IMLA will further monitor the model health via computing the performance metric
- Stakeholder will be automatically notified if system has found the model performance deteriorates

Thank you!