PDF / SOLUTIONS™

2023 PDF Users Conference:

AI for tomorrow's manufacturing & R&D

Location:
Santa Clara Marriott - 2700 Mission College
Boulevard Santa Clara, California 95054 USA

PDF Edge Inference for Test

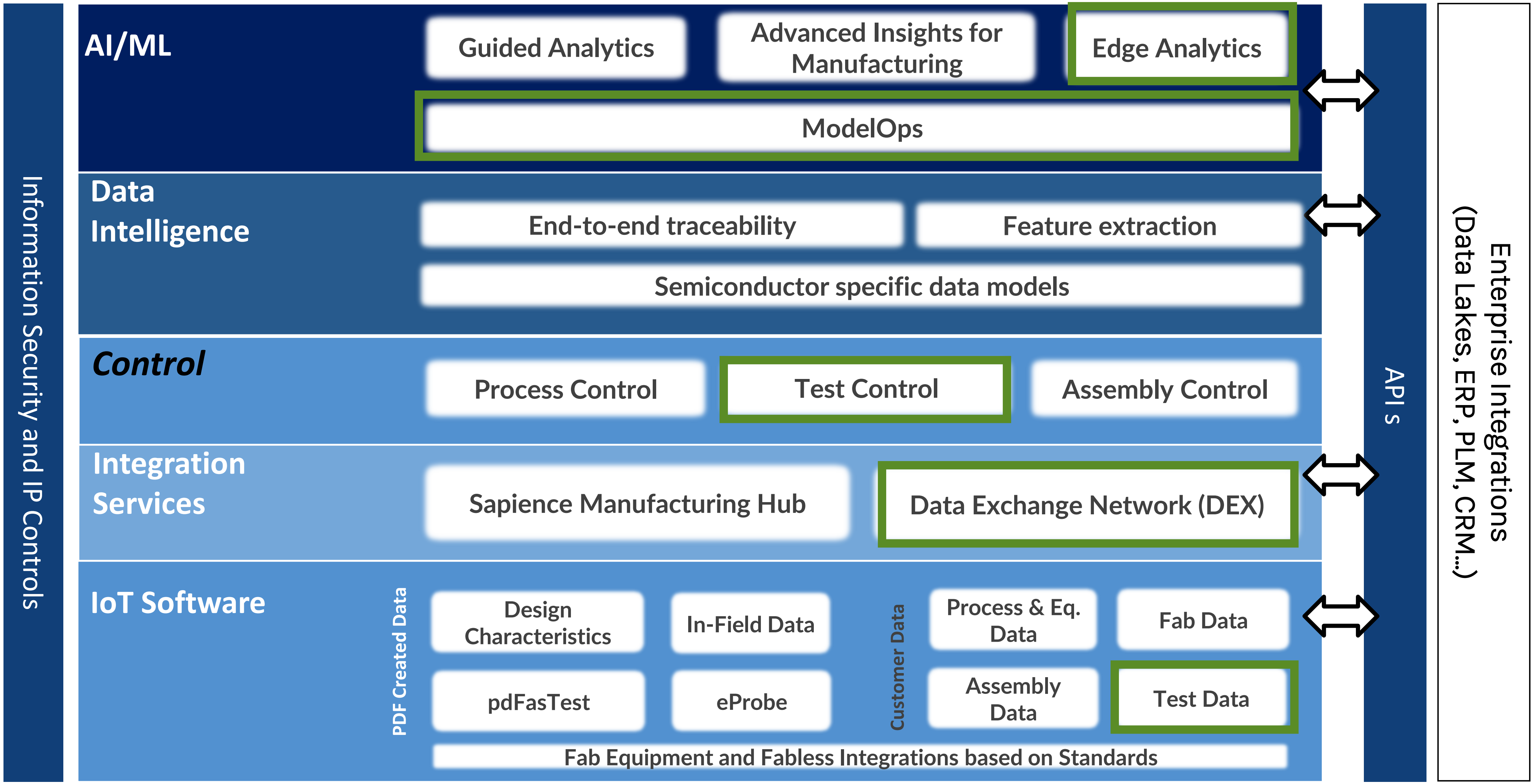
Data Feed Forward – Model Management – Edge Deployment

25th October 2023

This presentation and discussions resulting from it may include future product features or fixes, or the expected timing of future releases. This information is intended only to highlight areas of possible future development and current prioritizations. Nothing in this presentation or the discussions stemming from it are a commitment to any future release, new product features or fixes, or the timing of any releases. Actual future releases may or may not include these product features or fixes, and changes to any roadmap or timeline are at the sole discretion of PDF Solutions, Inc. and may be made without any requirement for updating. For information on current prioritizations and intended future features or fixes, contact sales@pdf.com.

PDF Solutions, Exensio, CV, Cimetrix, the PDF Solutions logo, and the Cimetrix logo are registered trademarks of PDF Solutions, Inc. or its subsidiaries. All other trademarks cited in this document are the property of their respective owners.

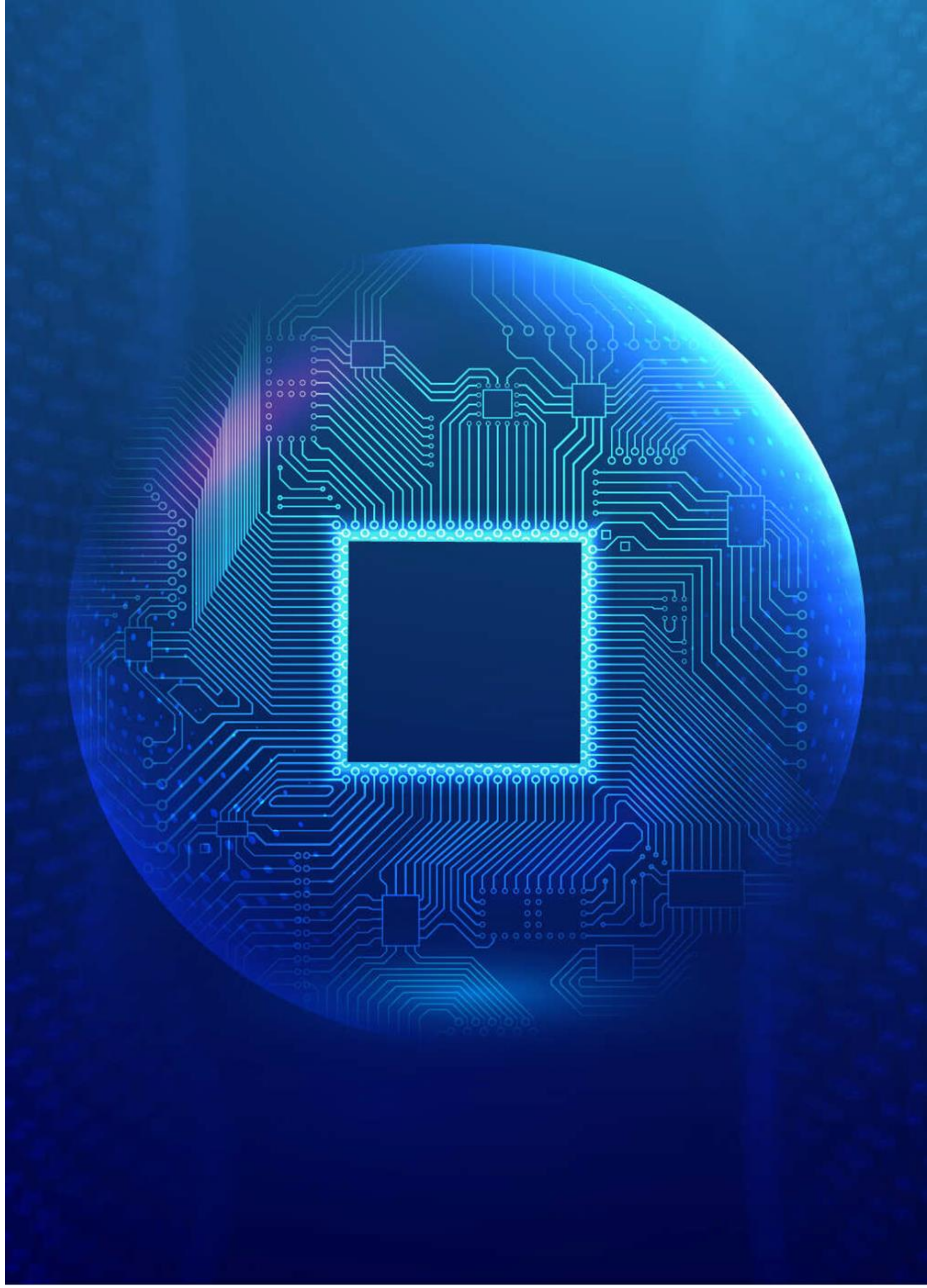
End-to-end Integrated Platform for Semiconductor Analytics



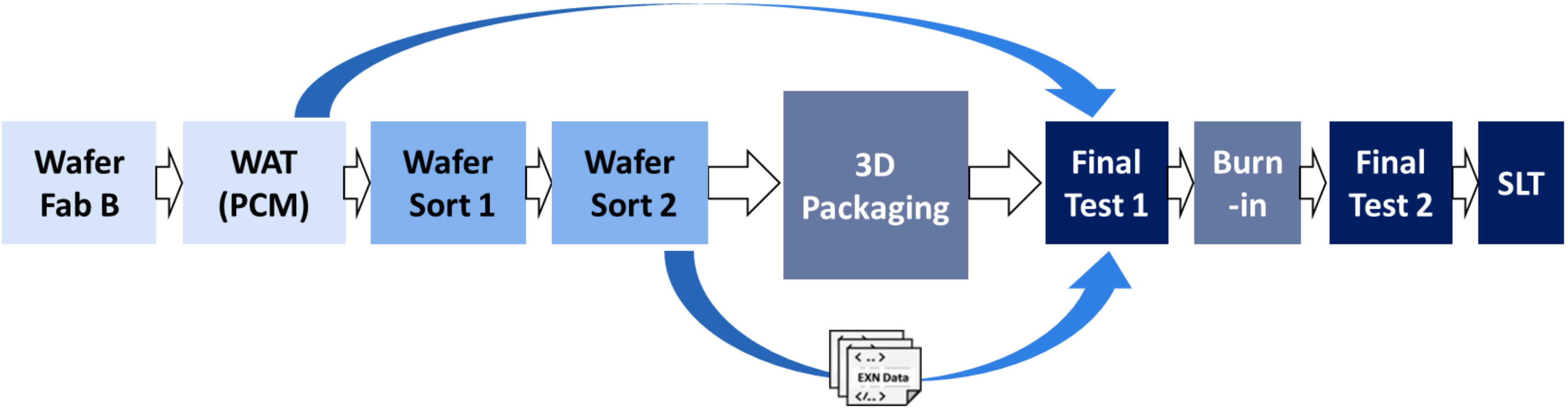
Fully integrated solution to accelerate production ramp, improve overall yield and quality for Semiconductors

Overview

- Application Scenario
- Edge Deployment Architecture
- Model Training and Distribution
- Exensio Data Feed Forward Integration
- Demo: Data Feed Forward Orchestration
- Demo: Edge Model Execution



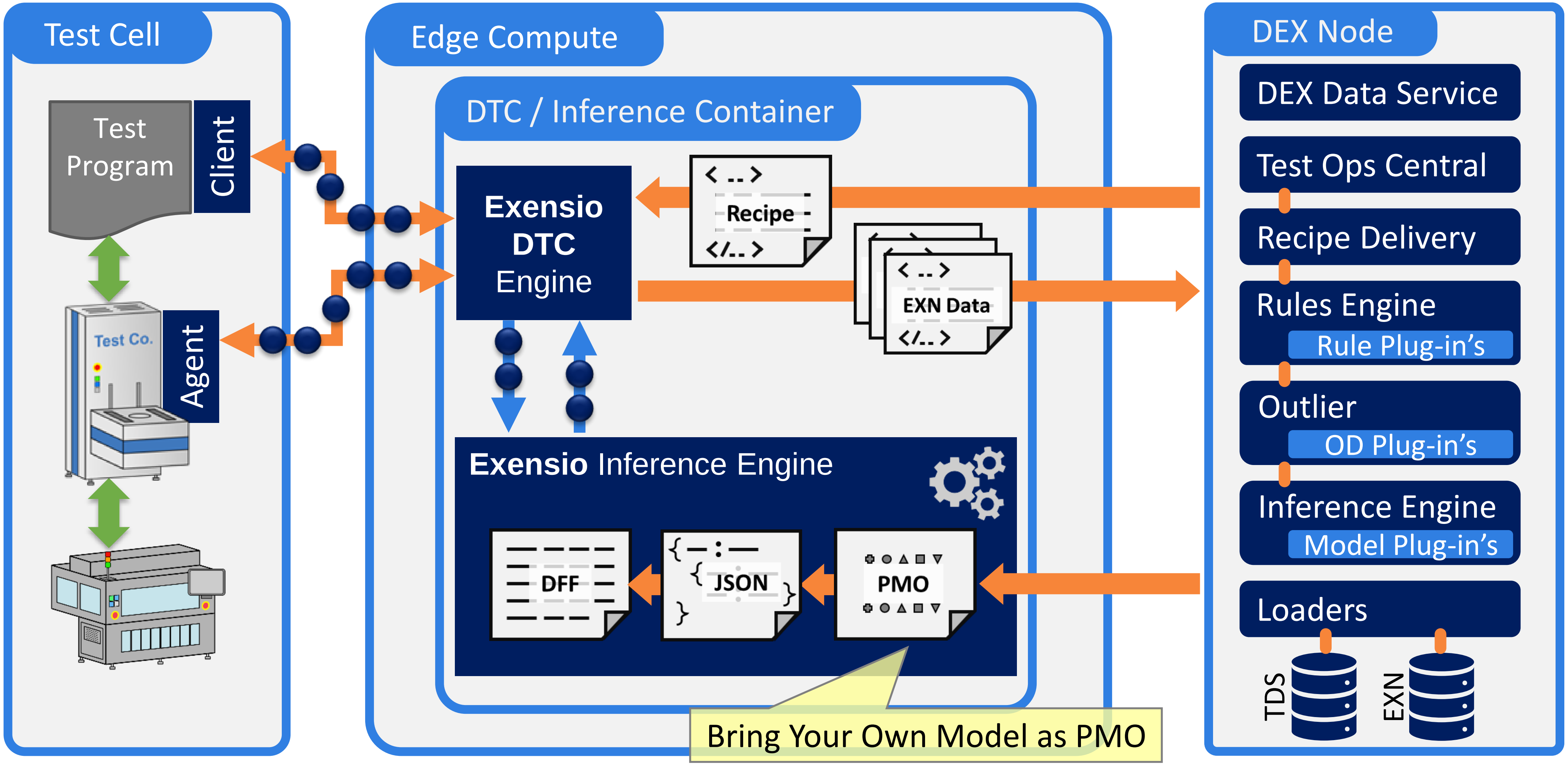
Backdrop Scenario



- **Deploy a feed-forward-loop test operation for a multi-chip device**
  - Reliable and timely capture of wafer sort test operation data
  - Automate feature extraction from PCM/WAT & WS to produce feed-forward data
  - Mechanism to train & periodically retrain models on current data
  - Deploy models across supply chain to the package test operation
  - Monitor test data and model performance

Edge Integrated Inference Container

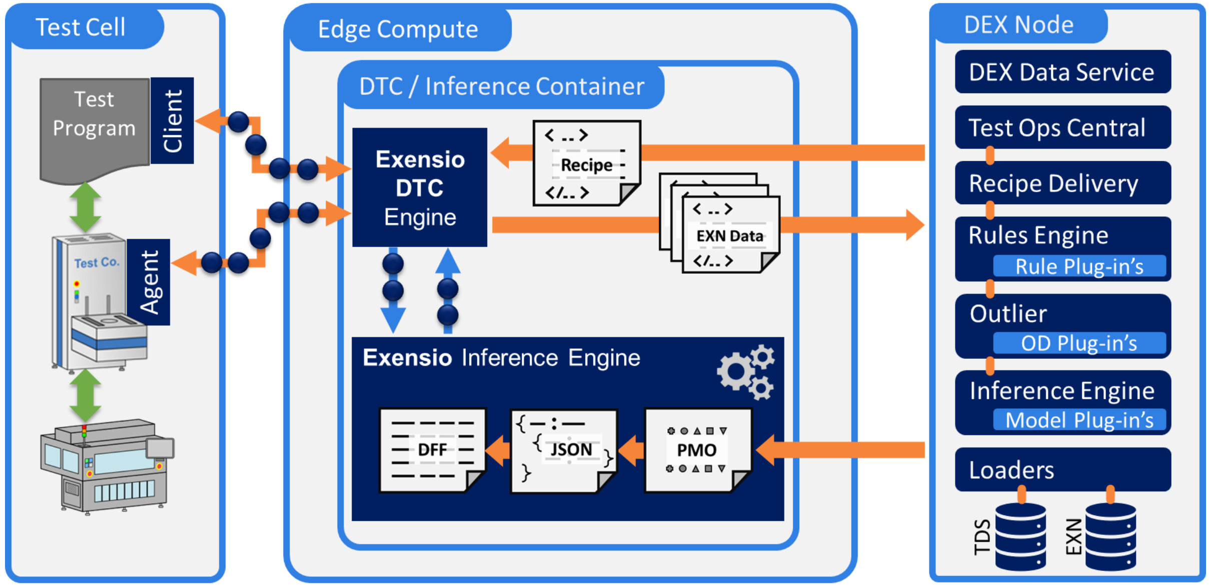
Your model with Exensio rules & model management deployed to the edge



PDF/SOLUTIONS™

Edge Integrated Inference Container

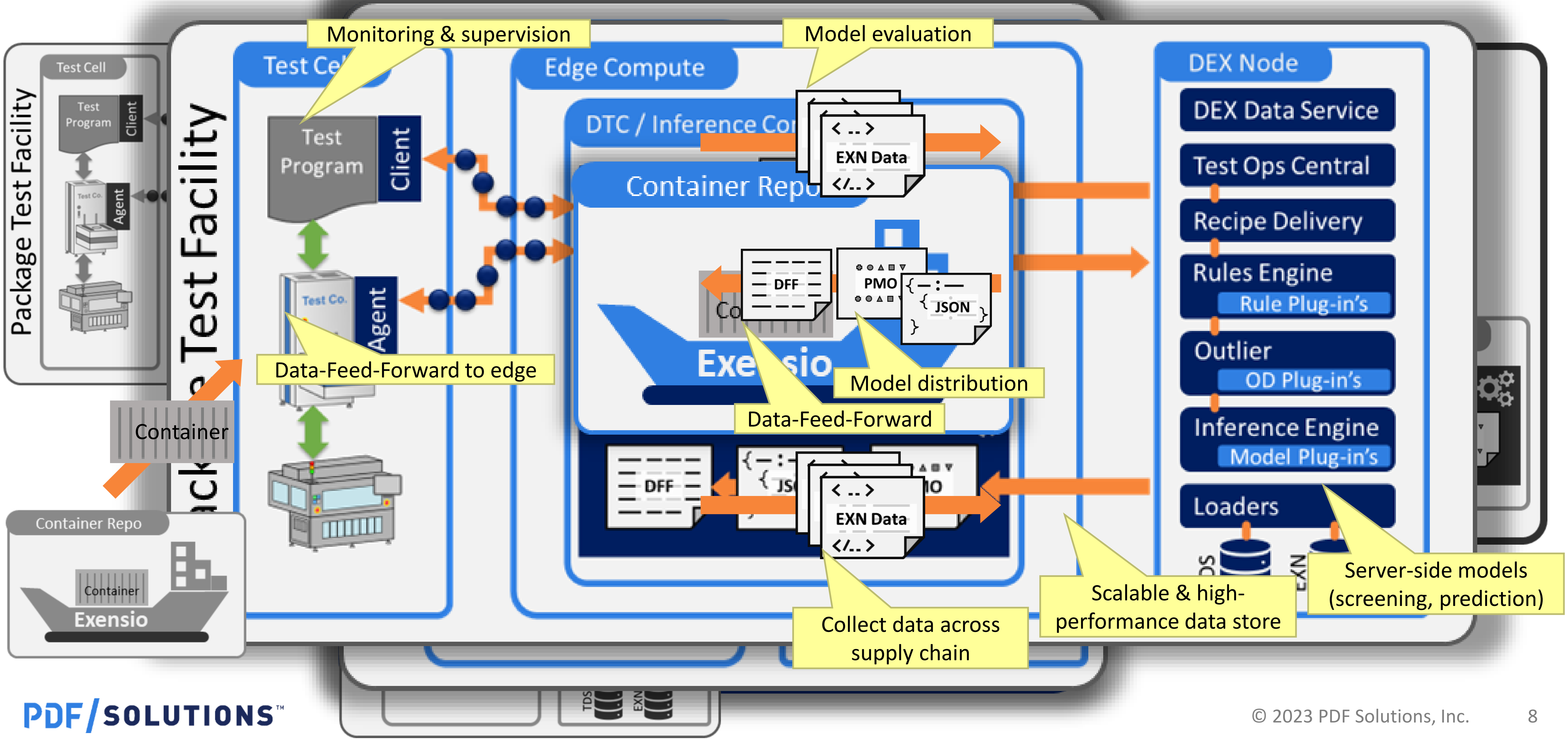
Your model with Exensio rules & model management deployed to the edge



- Synchronous inline inference per test flow enables adaptive test
- High-speed prediction and bin override (<200ms roundtrip)
- At-scale deployment architecture
- Bring Your Own (BYO) model
- Full spectrum data feed
- Compliment model with rules
- Secure execution environment

Exensio ML Model Deployment

Automatic dynamically scaled prediction model training, deployment and model management



Monitoring & supervision

Model evaluation

Data-Feed-Forward to edge

Model distribution

Data-Feed-Forward

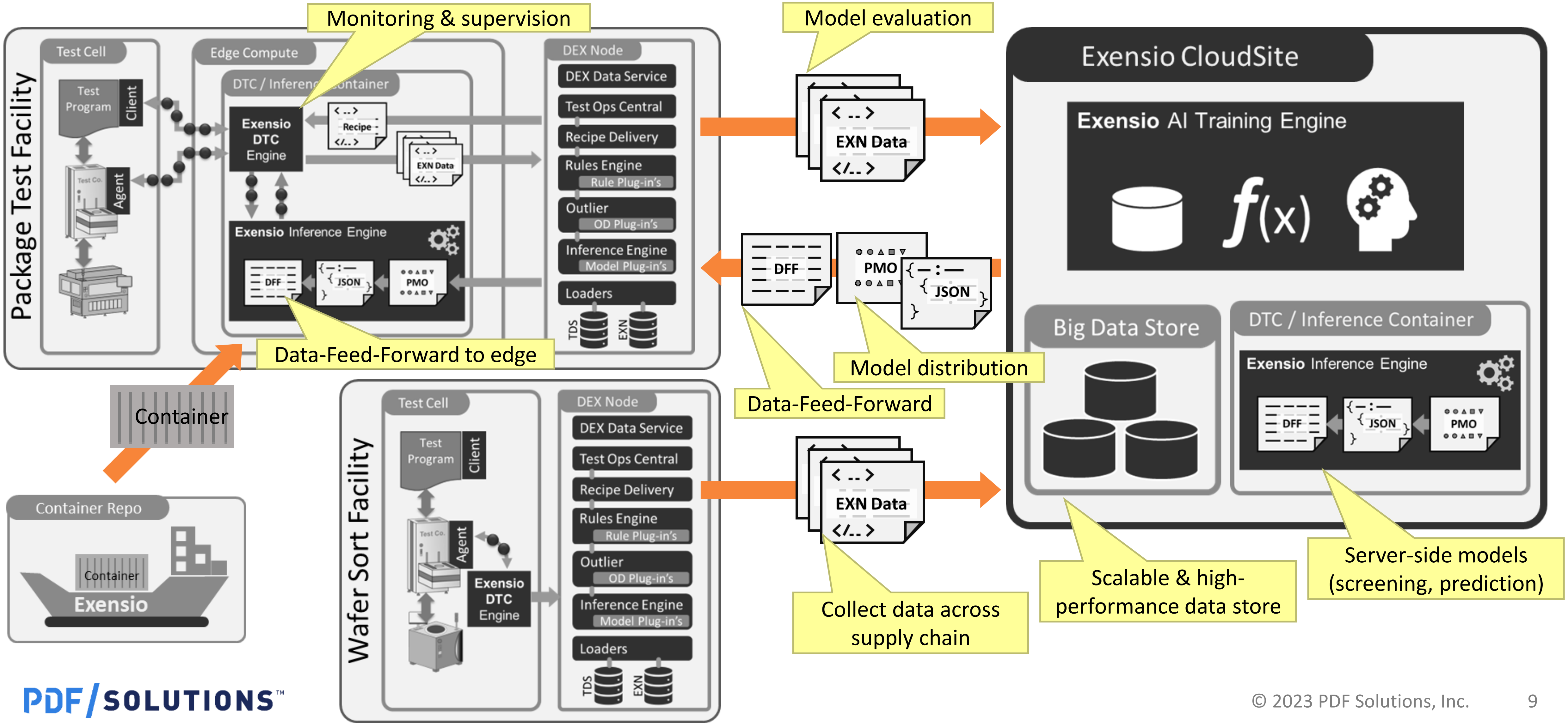
Collect data across supply chain

Scalable & high-performance data store

Server-side models (screening, prediction)

Exensio ML Model Deployment

Automatic dynamically scaled prediction model training, deployment and model management



Monitoring & supervision

Model evaluation

Data-Feed-Forward to edge

Model distribution

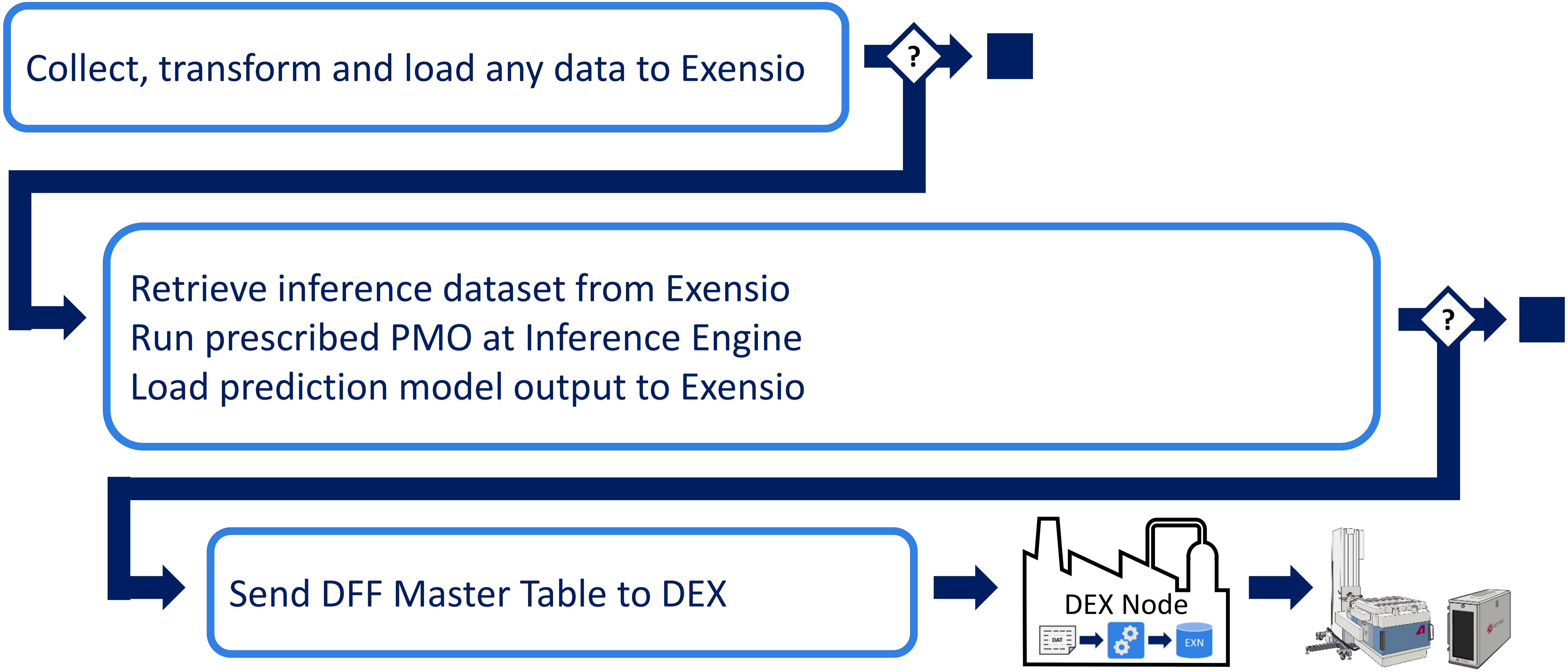
Data-Feed-Forward

Collect data across supply chain

Scalable & high-performance data store

Server-side models (screening, prediction)

ML Inference Engine Prediction and Data-Feed-Forward

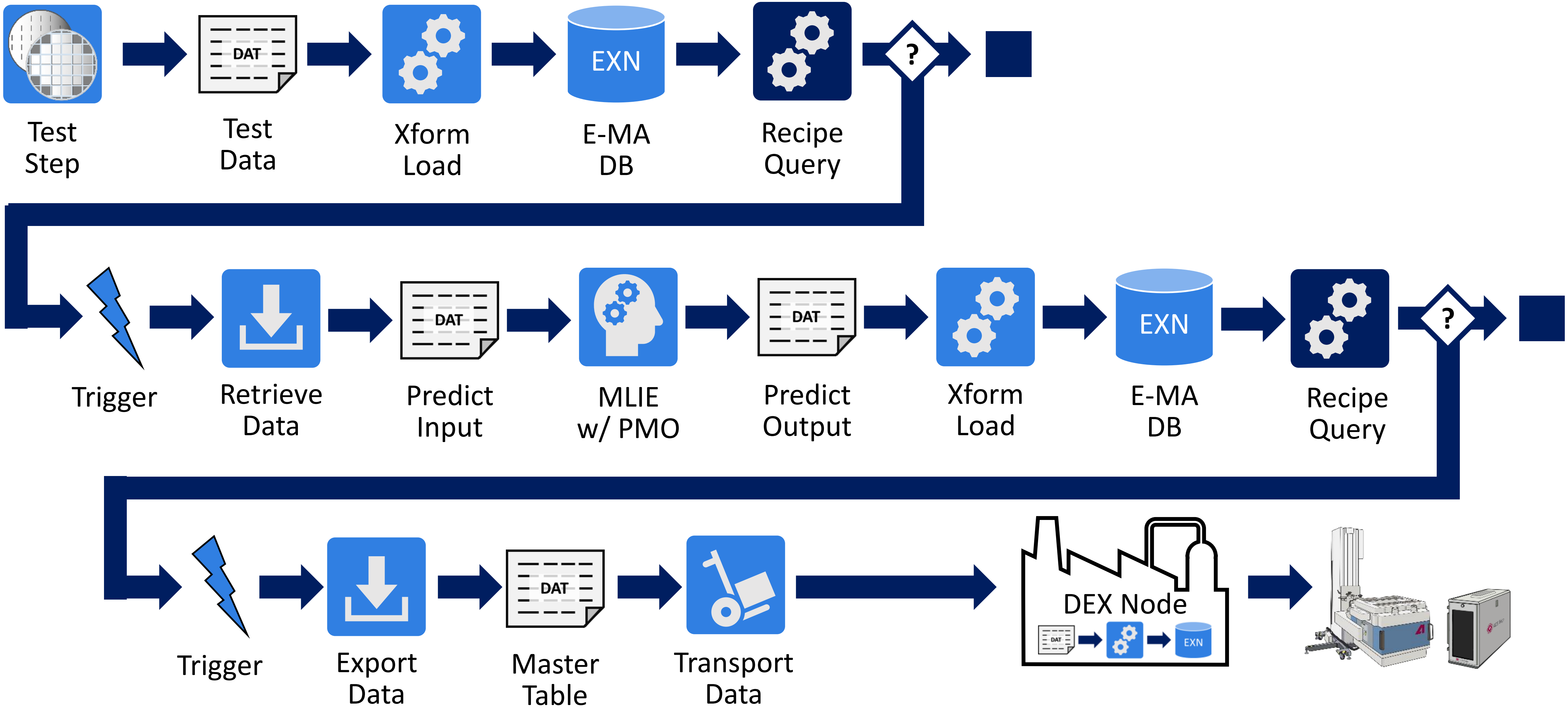


Collect, transform and load any data to Exensio

Retrieve inference dataset from Exensio
Run prescribed PMO at Inference Engine
Load prediction model output to Exensio

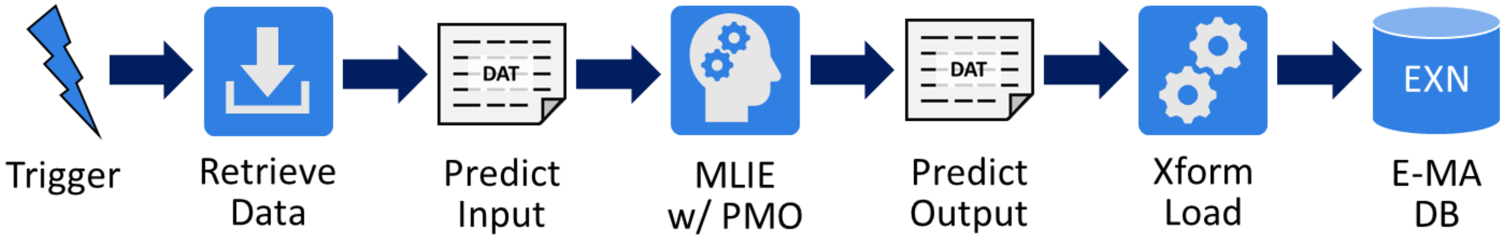
Send DFF Master Table to DEX

ML Inference Engine Prediction and Data-Feed-Forward

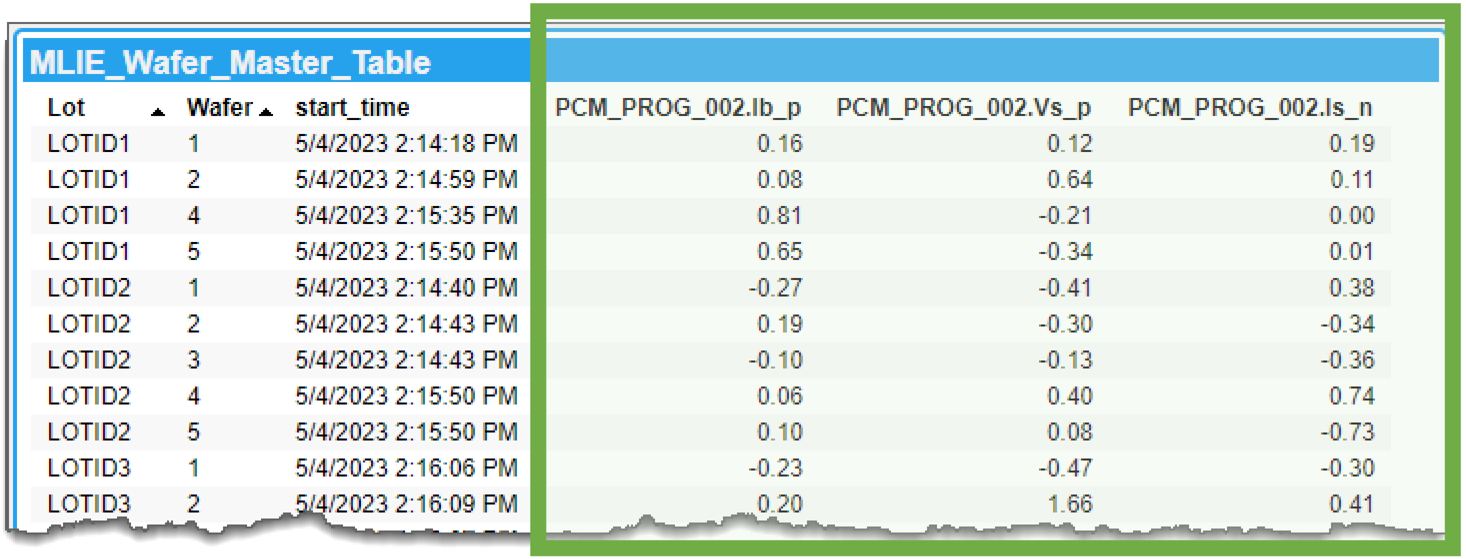


PDF/SOLUTIONS™

Per-Wafer DFF Model Outputs



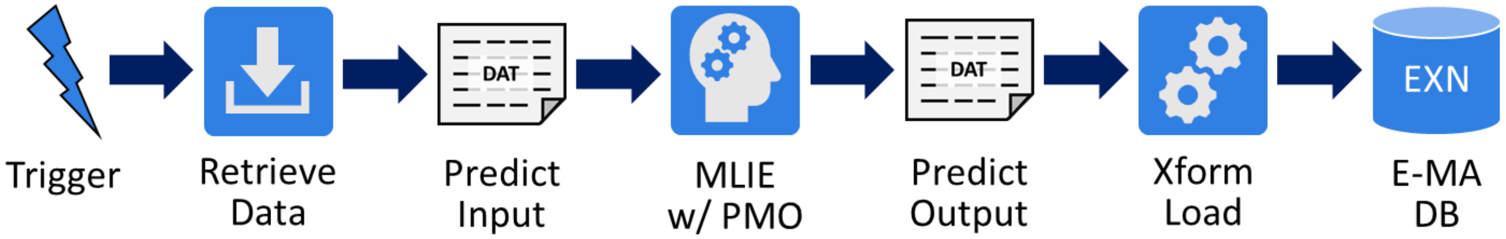
Wafer-level statistics for parameters used in die-level prediction model



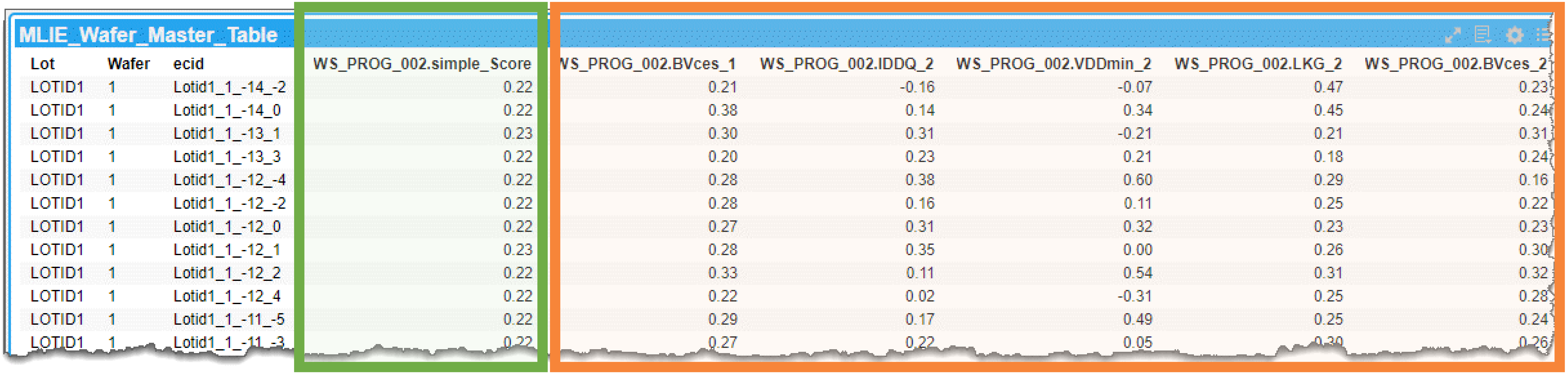
MLIE_Wafer_Master_Table

Lot	Wafer	start_time	PCM_PROG_002.Ib_p	PCM_PROG_002.Vs_p	PCM_PROG_002.Is_n
LOTID1	1	5/4/2023 2:14:18 PM	0.16	0.12	0.19
LOTID1	2	5/4/2023 2:14:59 PM	0.08	0.64	0.11
LOTID1	4	5/4/2023 2:15:35 PM	0.81	-0.21	0.00
LOTID1	5	5/4/2023 2:15:50 PM	0.65	-0.34	0.01
LOTID2	1	5/4/2023 2:14:40 PM	-0.27	-0.41	0.38
LOTID2	2	5/4/2023 2:14:43 PM	0.19	-0.30	-0.34
LOTID2	3	5/4/2023 2:14:43 PM	-0.10	-0.13	-0.36
LOTID2	4	5/4/2023 2:15:50 PM	0.06	0.40	0.74
LOTID2	5	5/4/2023 2:15:50 PM	0.10	0.08	-0.73
LOTID3	1	5/4/2023 2:16:06 PM	-0.23	-0.47	-0.30
LOTID3	2	5/4/2023 2:16:09 PM	0.20	1.66	0.41

Per-Die DFF Model Outputs



Die-level engineered feature from raw DFF wafer sort parametric inputs

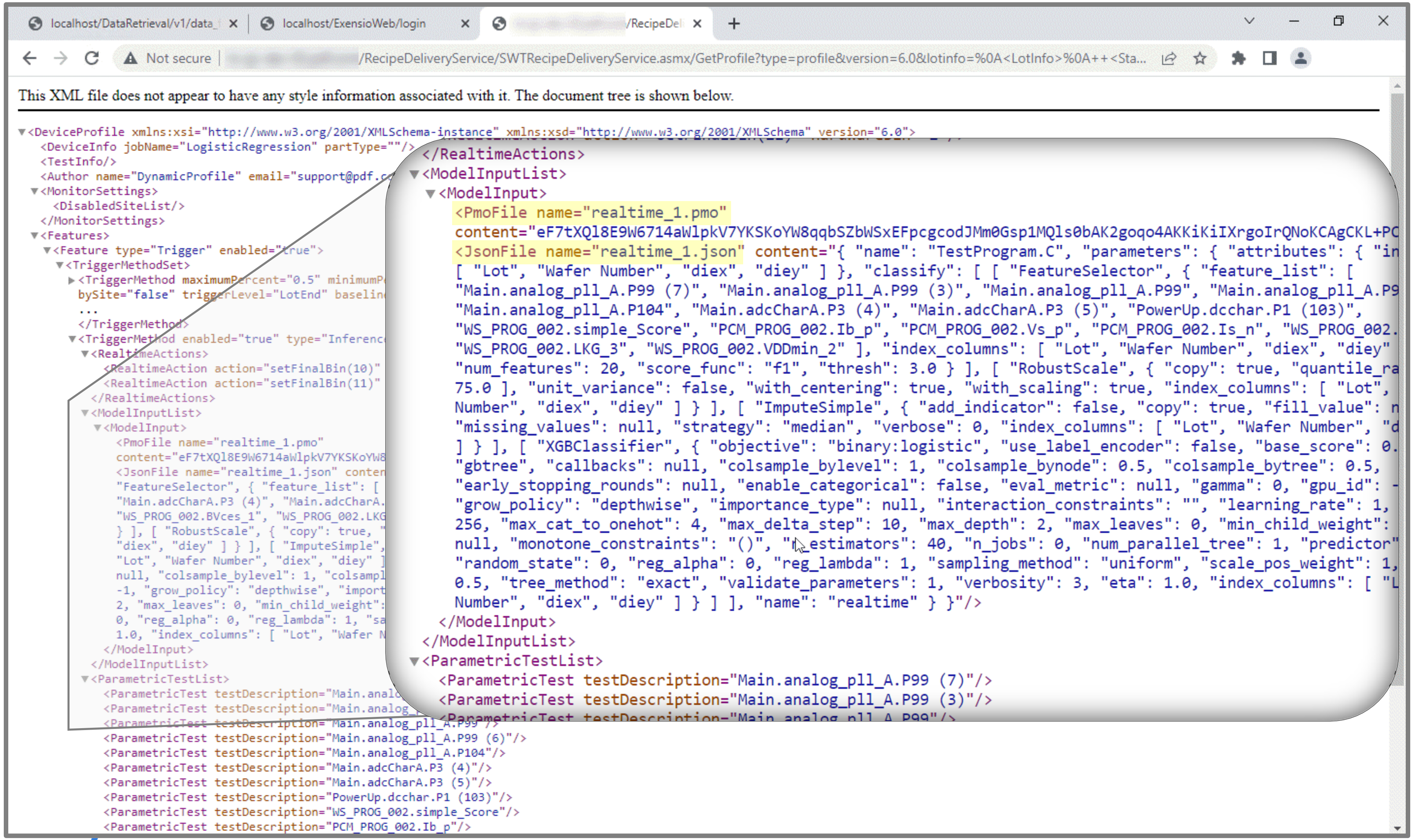


MLIE_Wafer_Master_Table

Lot	Wafer	ecid	WS_PROG_002.simple_Score	WS_PROG_002.BVces_1	WS_PROG_002.IDDQ_2	WS_PROG_002.VDDmin_2	WS_PROG_002.LKG_2	WS_PROG_002.BVces_2
LOTID1	1	Lotid1_1_-14_-2	0.22	0.21	-0.16	-0.07	0.47	0.23
LOTID1	1	Lotid1_1_-14_0	0.22	0.38	0.14	0.34	0.45	0.24
LOTID1	1	Lotid1_1_-13_1	0.23	0.30	0.31	-0.21	0.21	0.31
LOTID1	1	Lotid1_1_-13_3	0.22	0.20	0.23	0.21	0.18	0.24
LOTID1	1	Lotid1_1_-12_-4	0.22	0.28	0.38	0.60	0.29	0.16
LOTID1	1	Lotid1_1_-12_-2	0.22	0.28	0.16	0.11	0.25	0.22
LOTID1	1	Lotid1_1_-12_0	0.22	0.27	0.31	0.32	0.23	0.23
LOTID1	1	Lotid1_1_-12_1	0.23	0.28	0.35	0.00	0.26	0.30
LOTID1	1	Lotid1_1_-12_2	0.22	0.33	0.11	0.54	0.31	0.32
LOTID1	1	Lotid1_1_-12_4	0.22	0.22	0.02	-0.31	0.25	0.28
LOTID1	1	Lotid1_1_-11_-5	0.22	0.29	0.17	0.49	0.25	0.24
LOTID1	1	Lotid1_1_-11_-3	0.22	0.27	0.22	0.05	0.30	0.26

Filtered subset of raw wafer sort parameters used by prediction model

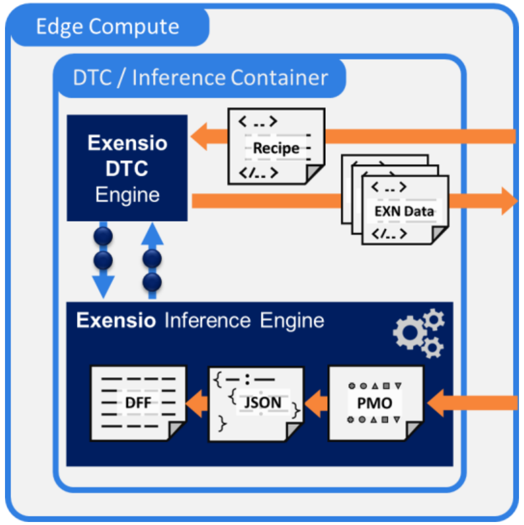
Recipe Delivery Web Service Model Integration


localhost/DataRetrieval/v1/data_
localhost/ExensioWeb/login
/RecipeDel
Not secure
/RecipeDeliveryService/SWTRecipeDeliveryService.asmx/GetProfile?type=profile&version=6.0&lotinfo=%0A<LotInfo>%0A++<Sta...

This XML file does not appear to have any style information associated with it. The document tree is shown below.

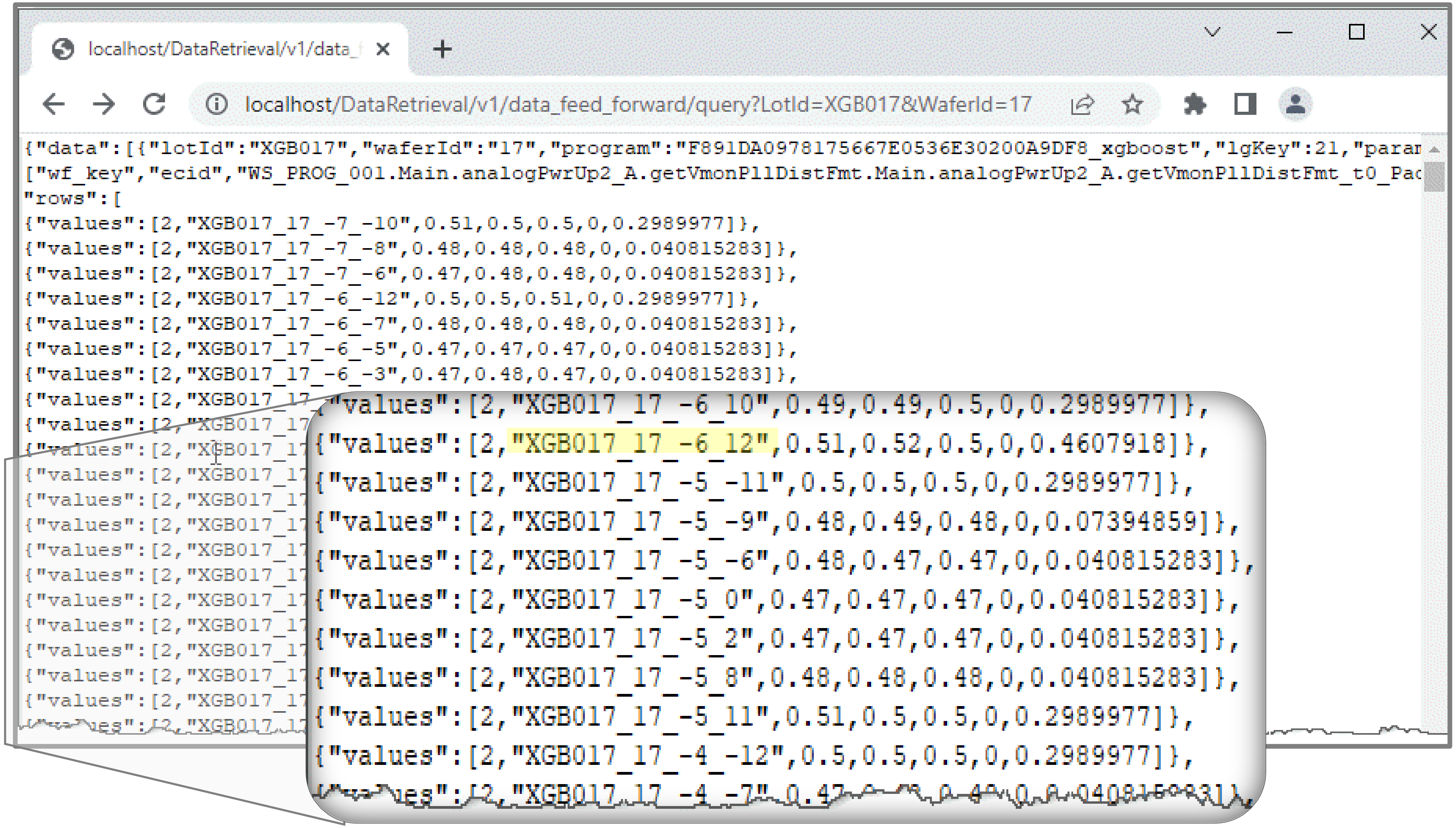
```
▼<DeviceProfile xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema" version="6.0">
   <DeviceInfo jobName="LogisticRegression" partType=""/>
   <TestInfo/>
   <Author name="DynamicProfile" email="support@pdf.c
 ▼<MonitorSettings>
     <DisabledSiteList/>
   </MonitorSettings>
 ▼<Features>
   ▼<Feature type="Trigger" enabled="true">
     ▼<TriggerMethodSet>
       ▶<TriggerMethod maximumPercent="0.5" minimumP
         bySite="false" triggerLevel="LotEnd" baselin
         ...
         </TriggerMethod>
       ▼<TriggerMethod enabled="true" type="Inferenc
         ▼<RealtimeActions>
             <RealtimeAction action="setFinalBin(10)"
             <RealtimeAction action="setFinalBin(11)"
           </RealtimeActions>
         ▼<ModelInputList>
           ▼<ModelInput>
               <PmoFile name="realtime_1.pmo"
               content="eF7tXQl8E9W6714aWlpkV7YKSKoYW8
               <JsonFile name="realtime_1.json" conten
               "FeatureSelector", { "feature_list": [
               "Main.adcCharA.P3 (4)", "Main.adcCharA.
               "WS_PROG_002.BVces_1", "WS_PROG_002.LKG
               } ], [ "RobustScale", { "copy": true, "
               "diex", "diey" ] } ], [ "ImputeSimple",
               "Lot", "Wafer Number", "diex", "diey" ]
               null, "colsample_bylevel": 1, "colsampl
               -1, "grow_policy": "depthwise", "import
               2, "max_leaves": 0, "min_child_weight":
               0, "reg_alpha": 0, "reg_lambda": 1, "sa
               1.0, "index_columns": [ "Lot", "Wafer N
             </ModelInput>
           </ModelInputList>
         ▼<ParametricTestList>
             <ParametricTest testDescription="Main.analo
             <ParametricTest testDescription="Main.analog
             <ParametricTest testDescription="Main.analog_pll_A.P99"/>
             <ParametricTest testDescription="Main.analog_pll_A.P99 (6)"/>
             <ParametricTest testDescription="Main.analog_pll_A.P104"/>
             <ParametricTest testDescription="Main.adcCharA.P3 (4)"/>
             <ParametricTest testDescription="Main.adcCharA.P3 (5)"/>
             <ParametricTest testDescription="PowerUp.dcchar.P1 (103)"/>
             <ParametricTest testDescription="WS_PROG_002.simple_Score"/>
             <ParametricTest testDescription="PCM_PROG_002.Ib_p"/>
```

```
 </RealtimeActions>
▼<ModelInputList>
  ▼<ModelInput>
     <PmoFile name="realtime_1.pmo"
     content="eF7tXQl8E9W6714aWlpkV7YKSKoYW8qqbSZbWSxEFpcgcodJMm0Gsp1MQls0bAK2goqo4AKKiKiIXrgoIrQNoKCAgCKL+PC
     <JsonFile name="realtime_1.json" content="{ "name": "TestProgram.C", "parameters": { "attributes": { "in
     [ "Lot", "Wafer Number", "diex", "diey" ] }, "classify": [ [ "FeatureSelector", { "feature_list": [
     "Main.analog_pll_A.P99 (7)", "Main.analog_pll_A.P99 (3)", "Main.analog_pll_A.P99", "Main.analog_pll_A.P9
     "Main.analog_pll_A.P104", "Main.adcCharA.P3 (4)", "Main.adcCharA.P3 (5)", "PowerUp.dcchar.P1 (103)",
     "WS_PROG_002.simple_Score", "PCM_PROG_002.Ib_p", "PCM_PROG_002.Vs_p", "PCM_PROG_002.Is_n", "WS_PROG_002.
     "WS_PROG_002.LKG_3", "WS_PROG_002.VDDmin_2" ], "index_columns": [ "Lot", "Wafer Number", "diex", "diey"
     "num_features": 20, "score_func": "f1", "thresh": 3.0 } ], [ "RobustScale", { "copy": true, "quantile_ra
     75.0 ], "unit_variance": false, "with_centering": true, "with_scaling": true, "index_columns": [ "Lot",
     Number", "diex", "diey" ] } ], [ "ImputeSimple", { "add_indicator": false, "copy": true, "fill_value": n
     "missing_values": null, "strategy": "median", "verbose": 0, "index_columns": [ "Lot", "Wafer Number", "d
     ] } ], [ "XGBClassifier", { "objective": "binary:logistic", "use_label_encoder": false, "base_score": 0.
     "gbtree", "callbacks": null, "colsample_bylevel": 1, "colsample_bynode": 0.5, "colsample_bytree": 0.5,
     "early_stopping_rounds": null, "enable_categorical": false, "eval_metric": null, "gamma": 0, "gpu_id": -
     "grow_policy": "depthwise", "importance_type": null, "interaction_constraints": "", "learning_rate": 1,
     256, "max_cat_to_onehot": 4, "max_delta_step": 10, "max_depth": 2, "max_leaves": 0, "min_child_weight":
     null, "monotone_constraints": "()", "n_estimators": 40, "n_jobs": 0, "num_parallel_tree": 1, "predictor"
     "random_state": 0, "reg_alpha": 0, "reg_lambda": 1, "sampling_method": "uniform", "scale_pos_weight": 1,
     0.5, "tree_method": "exact", "validate_parameters": 1, "verbosity": 3, "eta": 1.0, "index_columns": [ "L
     Number", "diex", "diey" ] } ] ], "name": "realtime" } }"/>
   </ModelInput>
 </ModelInputList>
▼<ParametricTestList>
   <ParametricTest testDescription="Main.analog_pll_A.P99 (7)"/>
   <ParametricTest testDescription="Main.analog_pll_A.P99 (3)"/>
```



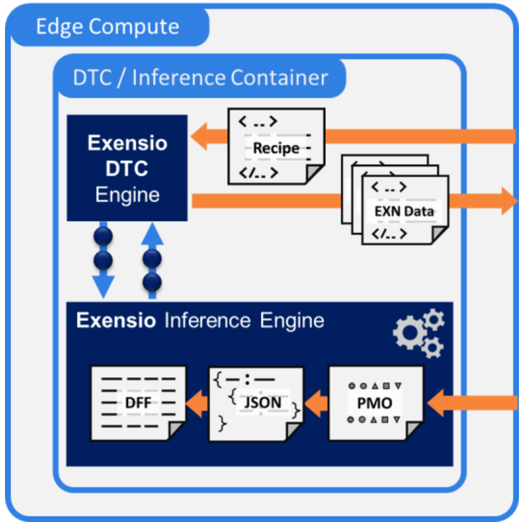
realtime_1.pmo
and
realtime_1.json
included and delivered within the test session recipe and sourced to Inference Engine container

Edge Data Feed Forward Query Web API



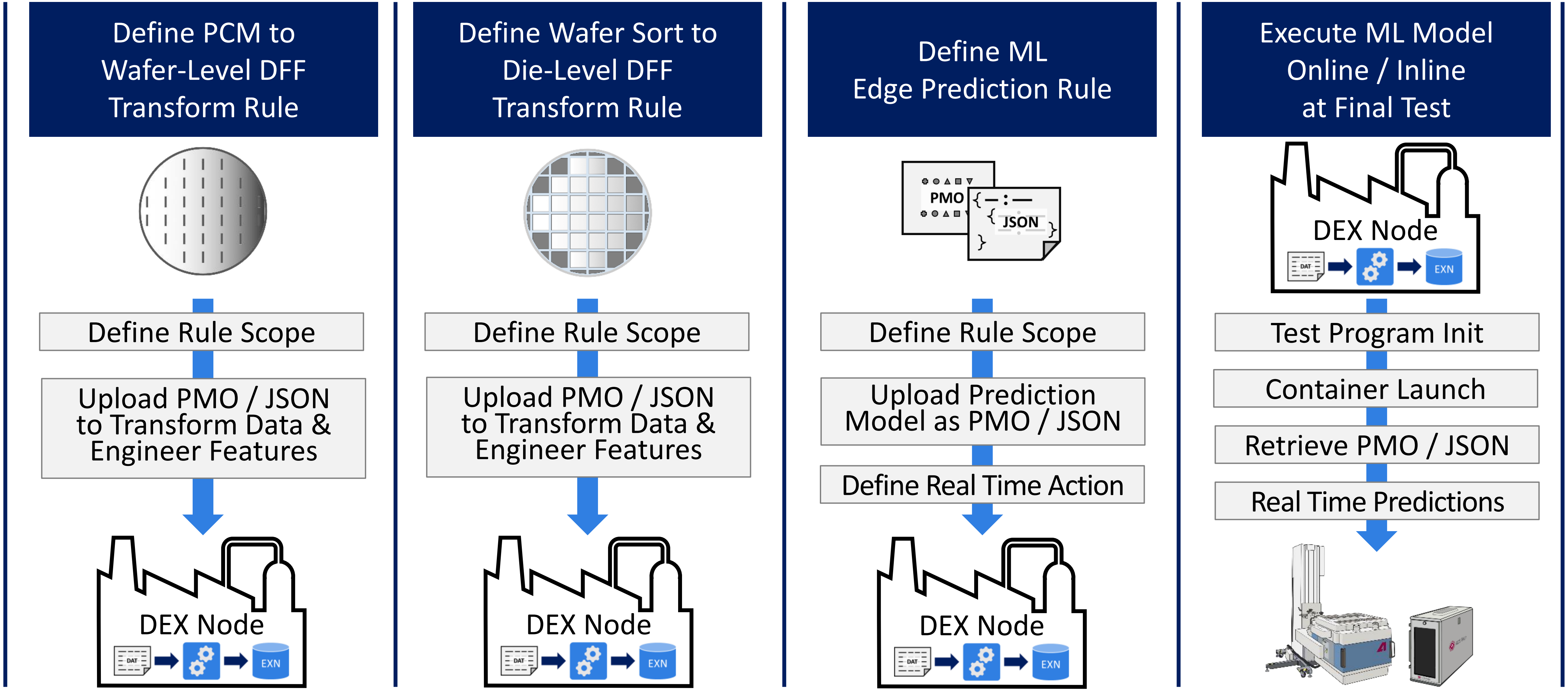
localhost/DataRetrieval/v1/data_f ×
+
localhost/DataRetrieval/v1/data_feed_forward/query?LotId=XGB017&WaferId=17

```
{"data":[{"lotId":"XGB017","waferId":"17","program":"F891DA0978175667E0536E30200A9DF8_xgboost","lgKey":21,"paran
["wf_key","ecid","WS_PROG_001.Main.analogPwrUp2_A.getVmonPllDistFmt.Main.analogPwrUp2_A.getVmonPllDistFmt_t0_Pac
"rows":[
{"values":[2,"XGB017_17_-7_-10",0.51,0.5,0.5,0,0.2989977]},
{"values":[2,"XGB017_17_-7_-8",0.48,0.48,0.48,0,0.040815283]},
{"values":[2,"XGB017_17_-7_-6",0.47,0.48,0.48,0,0.040815283]},
{"values":[2,"XGB017_17_-6_-12",0.5,0.5,0.51,0,0.2989977]},
{"values":[2,"XGB017_17_-6_-7",0.48,0.48,0.48,0,0.040815283]},
{"values":[2,"XGB017_17_-6_-5",0.47,0.47,0.47,0,0.040815283]},
{"values":[2,"XGB017_17_-6_-3",0.47,0.48,0.47,0,0.040815283]},
{"values":[2,"XGB017_17_-6_10",0.49,0.49,0.5,0,0.2989977]},
{"values":[2,"XGB017_17_-6_12",0.51,0.52,0.5,0,0.4607918]},
{"values":[2,"XGB017_17_-5_-11",0.5,0.5,0.5,0,0.2989977]},
{"values":[2,"XGB017_17_-5_-9",0.48,0.49,0.48,0,0.07394859]},
{"values":[2,"XGB017_17_-5_-6",0.48,0.47,0.47,0,0.040815283]},
{"values":[2,"XGB017_17_-5_0",0.47,0.47,0.47,0,0.040815283]},
{"values":[2,"XGB017_17_-5_2",0.47,0.47,0.47,0,0.040815283]},
{"values":[2,"XGB017_17_-5_8",0.48,0.48,0.48,0,0.040815283]},
{"values":[2,"XGB017_17_-5_11",0.51,0.5,0.5,0,0.2989977]},
{"values":[2,"XGB017_17_-4_-12",0.5,0.5,0.5,0,0.2989977]},
{"values":[2,"XGB017_17_-4_-7",0.47
```



Data Feed Forward dataset retrieved from web service API with array of multiple parameters **per-device** by **ECID** and sourced to Inference Engine container

Demo Scenario



Define PCM to Wafer-Level DFF Transform Rule

- Define Rule Scope
- Upload PMO / JSON to Transform Data & Engineer Features

Define Wafer Sort to Die-Level DFF Transform Rule

- Define Rule Scope
- Upload PMO / JSON to Transform Data & Engineer Features

Define ML Edge Prediction Rule

- Define Rule Scope
- Upload Prediction Model as PMO / JSON
- Define Real Time Action

Execute ML Model Online / Inline at Final Test

- Test Program Init
- Container Launch
- Retrieve PMO / JSON
- Real Time Predictions

Demo Videos

PDF/SOLUTIONS™

MLIE Data-Feed-Forward Workflow

Recorded Demonstration May 2023

03:51

PDF/SOLUTIONS™

Confidential Information

Exensio Inference Container for ACS Edge

Recorded Demonstration August 23, 2022

04:50 – 07:54

Statistical and ML Apps for ACS Edge™

Statistical Applications

- ACS Outlier Screening: Real-time outlier screening using a variety of algorithms
- ACS Adaptive Test: Real-time adaptive test; Test More or Test Less
- ACS Tester Control: Real-time statistical process control (SPC) to avoid quality excursions and escapes
- ACS Statistical Binning: Dynamically bin devices via statistical rules and without test program modification

ML Based Applications

- ACS Predictive Binning: Dynamically bin devices with upstream data (Data Feed Forward) and via an ML model decision
- ACS Outlier Screening ML: Real-time outlier screening using ML
- ACS Adaptive Test ML: Real-time adaptive test using ML
- ACS Custom ML: Bring-Your-Own ML model and leverage the Exensio Data Feed Forward/Backward infrastructure

PDF/SOLUTIONS™