ORCHESTRATE ML WORKFLOWS: RETAIL FORECASTING, INVENTORY MANAGEMENT IN POS AND SUPPLY CHAIN



Predicting point-of-sale (POS) sales across stores, coordinated with inventory and supply chain data, is table stakes for retailers. This blog explains this use case leveraging PySpark for data and machine learning (ML) pipelines on Databricks, orchestrated with Control-M to predict POS and forecast inventory items. This blog has two main parts. In the first section, we will cover the details of a retail forecasting use case and the ML pipeline defined in Azure Databricks. The second section will cover the integration between Control-M and Databricks.

Developing the use case in Azure Databricks

Note: All of the code used in this blog is available at this github repo.

In real life, data would be ingested from sensors and mobile devices, with near-real-time inventory measurements and POS data across stores. The data and ML pipeline is coordinated with Control-M to integrate the different components and visualize the results in an always-updated dashboard.

The data lands in the Databricks Intelligent Data Platform and is combined, enriched, and aggregated with PySpark Jobs. The resulting data is fed to different predictive algorithms for training and forecasting sales and demand with the results visualized in:

- Graphical dashboards
- Written as delta files to a data repository for offline consumption

In this post, we will also walk through the architecture and the components of this predictive system.

Data set and schema

The project uses real-world data, truncated in size and width to keep things simple. A simplified and abridged version of the schema is shown in Figure 1. The location table is a reference table, obtained from public datasets. The color coding of the fields shows the inter-table dependencies. Data was obtained partially from <u>Kaggle</u> and other public sources.

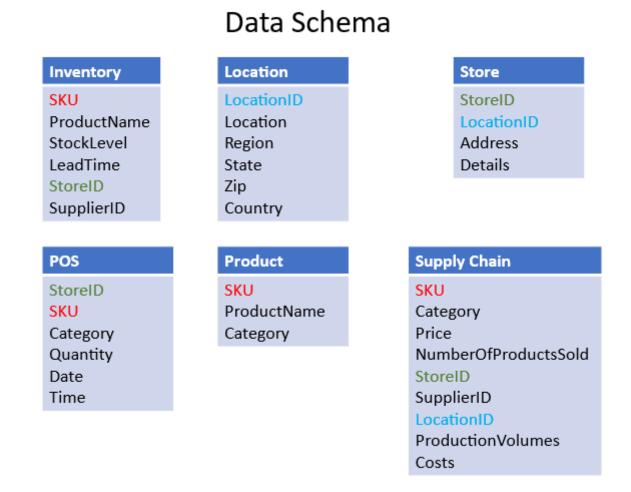


Figure 1. Example of schema.

Platform and components

The following components are used for implementing the use case:

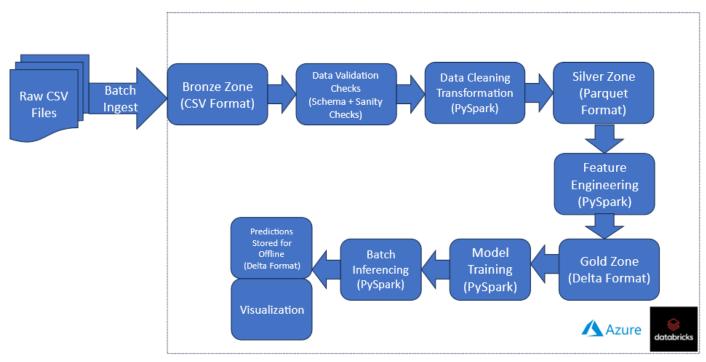
- Databricks Intelligent Data Platform on Azure
- PySpark
- Python Pandas library
- Python Seaborn library for data visualization
- Jupyter Notebooks on Databricks
- Parquet and Delta file format

Project artifacts

- Working environment on Azure
- Code for data ingestion, processing, ML training, and serving and saving forecasted results to Databricks Lakehouse in delta format
- Code for workflow and orchestration with Control-M to coordinate all the activities and tasks and handle failure scenarios

High-level architecture and data flow

Current architecture assumes that data lands in the raw zone of the data lakehouse as a csv file with a pre-defined schema as a batch. The high-level overview of the data flow and associated processes is showed in Figure 2.



Architecture and Data Flow

Figure 2. Overview of data flow and associated processes.

Data and feature engineering

Currently the data and ML pipelines are modeled as a batch process. Initial exploratory data analysis (EDA) was done to understand the datasets and relevant attributes contributing to predicting the inventory levels and POS sales. Initial EDA indicated that it is useful to transform the dates to "day of the week"- and "time of day"- type categories for best predictive results. The data pipelines included feature engineering capabilities for datasets that had time as part of the broader dataset.

Figure 3 shows a sample data pipeline for POS dataset. Figure 4 shows another similar data pipeline for an inventory dataset. Post data transformation, the transformed tables were joined to form a denormalized table for model training. This is shown in Figure 5 for enriching the POS and the inventory

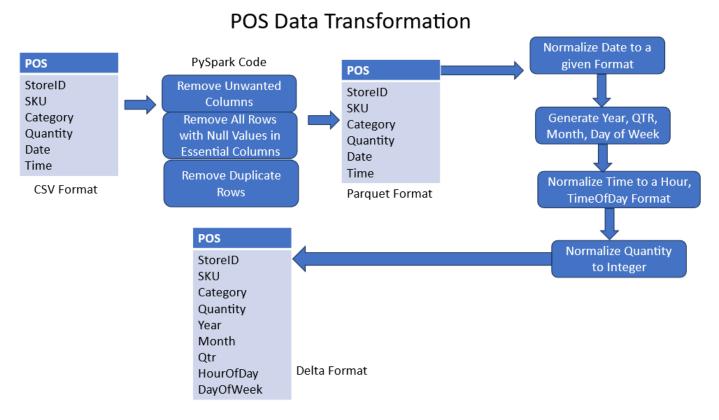


Figure 3. A sample data pipeline for POS dataset.

Inventory Data Transformation

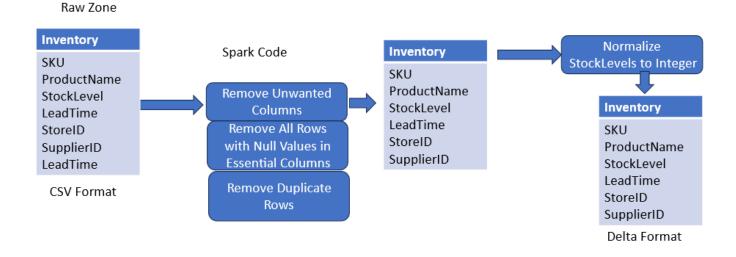


Figure 4. Sample data pipileine for inventory dataset.

Enrich POS Data

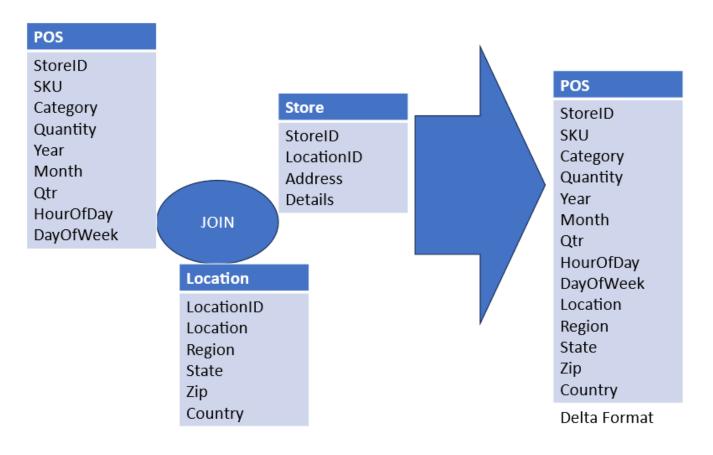


Figure 5. The joining of transformed tables to form a de-normalized table for model training.

Model training

The ML training pipeline used random forest and linear regression to predict the sales and inventory levels. The following modules from PySpark were used to create the ML pipeline and do one-hot encoding on the categorical variables.

- pyspark.ml.Pipeline
- pyspark.ml.feature.StringIndexer, OneHotEncoder
- pyspark.ml.feature.VectorAssembler
- pyspark.ml.feature.StandardScaler
- pyspark.ml.regression.RandomForestRegressor
- pyspark.ml.evaluation.RegressionEvaluator
- pyspark.ml.regression.LinearRegression

The enriched data was passed to the pipelines and the different regressor models were applied to the data to generate the predictions.

The RegressionEvaluator module was used to evaluate the results and the median absolute error (MAE), root mean squared error (RMSE), and R-squared metrics were generated to evaluate the predicted results. Feature weights were used to understand the contribution weights of each of the

features to the predictions.

Orchestrating the end-to-end predictions

Data orchestration of the different PySpark notebooks uses a Databricks Workflows job while the production orchestration is performed by Control-M using its Databricks plug-in. This approach enables our Databricks workflow to be embedded into the larger supply chain and inventory business workflows already managed by Control-M for a fully automated, end-to-end orchestration of all related processing. Furthermore, it gives us access to advanced scheduling features like the ability to manage concurrent execution of multiple Databricks workflows that may require access to constrained shared resources such as public IP addresses in our Azure subscription.

Figure 6 shows the different orchestrated tasks to generate the offline predictions and dashboard. The orchestration was kept simple and does not show all the error paths and error handling paths.

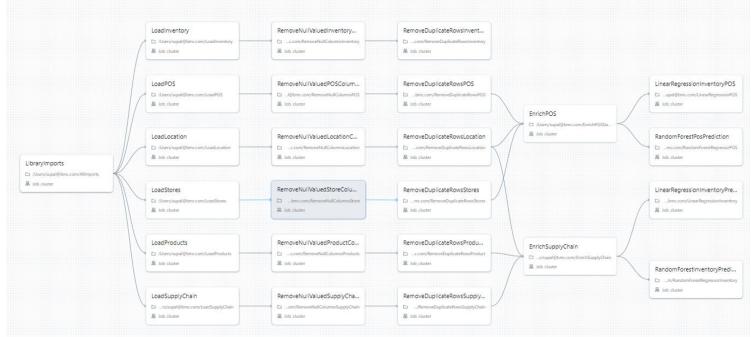


Figure 6. The different orchestrated tasks to generate the offline predictions and dashboard.

Control-M Integration with Databricks

Before creating a job in Control-M that can execute the Databricks workflow, we will need to <u>create</u> <u>a connection profile</u>. A connection profile contains authorization credentials—such as the username, password, and other plug-in-specific parameters—and enables you to connect to the application server with only the connection profile name. Connection profiles can be created using the web interface and then retrieved in json format using Control-M's Automation API. Included below is a sample of the connection profile for Azure Databricks in json format. If you create the connection profile directly in json before running the job, it should be deployed using the <u>Control-M Automation API CLI</u>.

1 ~	{
2 🗸	"JOG-AZ-DATABRICKS": {
3	"Type": "ConnectionProfile:Azure Databricks",
4	"Tenant ID": " <insert azure="" tenant",<="" th=""></insert>
5	"Databricks Resource": " <databricks id",<="" resource="" th=""></databricks>
6	"Client Secret": "****",
7	"Application ID": "Azure Application ID",
8	"Azure Login url": "https://login.microsoftonline.com",
9	"Databricks url": " <url>",</url>
10	"Connection Timeout": "50",
11	"Description": "",
12	"Centralized": true
13	}
14	}
15	

Creating a Databricks job

The job in Control-M that will execute the Databricks workflow is defined in json format as follows:

```
{
"jog-databricks" : {
"Type" : "Folder",
"ControlmServer" : "smprod",
"OrderMethod" : "Manual",
"SiteStandard" : "jog",
"SubApplication" : "Databricks",
"CreatedBy" : "joe_goldberg@bmc.com",
"Application" : "jog",
"DaysKeepActiveIfNotOk" : "1",
"When" : {
"RuleBasedCalendars" : {
"Included" : ,
"EVERYDAY" : {
"Type" : "Calendar:RuleBased",
"When" : {
"DaysRelation" : "OR",
"WeekDays" : ,
"MonthDays" :
}
}
}
},
"jog-azure-databricks-workflow" : {
"Type" : "Job:Azure Databricks",
"ConnectionProfile" : "JOG-AZ-DATABRICKS",
"Databricks Job ID" : "674653552173981",
```

```
"Parameters" : "\"params\" : {}",
"SubApplication" : "Databricks",
"Host" : "azureagents",
"CreatedBy" : "joe_goldberg@bmc.com",
"RunAs" : "JOG-AZ-DATABRICKS",
"Application" : "jog",
"When" : {
"WeekDays" : ,
"MonthDays" :
"DaysRelation" : "OR"
},
"AzEastPublicIPs" : {
"Type" : "Resource:Pool",
"Quantity" : "8"
}
}
}
}
```

Running the Databricks workflow

To run the job we will use the <u>run service</u> within Automation API.

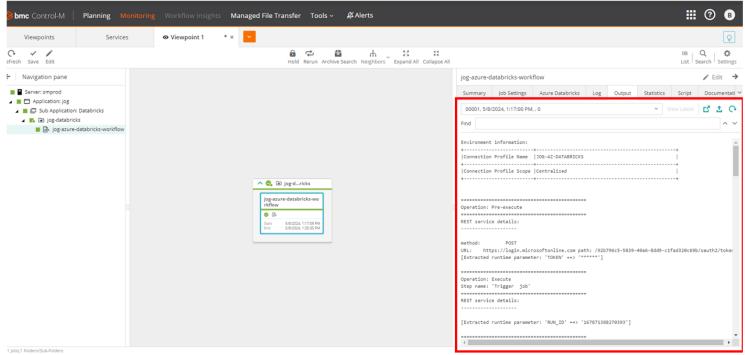
```
PS C:\Users\bfaruqui> ctm run C:\users\bfaruqui\Downloads\Workspace_38.json
{
    "runId": "f84ba1f8-8dd3-4614-bcd0-a0234b8bc0cd",
    "statusURI": "https://smprod.ctmdemo.com:8443/automation-api/run/status/f84ba1f8-8dd3-4614-bcd0-a0234b8bc0cd"
```

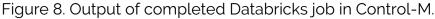
Visualizing the workflow

mc Control-M Planning M		Managed File Transfer	Tools ~	🛱 Alerts							₩ ?	
Viewpoints Services	• Viewpoint 1	* × 🔽										ç
h Save Edit		Hold	Kill Archive Sear	ch Neighbors Expand	보호 All Collapse All						i≣ Q List Search	Settin
Navigation pane						jog-azure-data	bricks-workflow				/ E	dit
Server: smprod						Summary)	ob Settings Azu	ure Databricks	Log O	utput Statistics	s Script	Do
 Application: jog Sub Application: Databricks Sub jog-databricks 						Run Time Paran	neters					
 jog-azure-databricks-workflow 						Execution Log						
						Run Number			001			
						Job Type			ujon@bmc.cor	~		
						Support group na Support group er			ujon@bmc.cor			
		🔨 👶 🗈 jog-0	d ricke			Subbour Broop e	nan	1.501	6,0112,01110,000			
						Log Details						
		jog-azure-dat rkflow	abricks-wo			Run #	Time	Source		Step	Log Messag	e
		0 B				00001	08-05-2024 1.	Applica	tion I	Pre-execute	Retrieve RES	JT re
		Start 5/8/2 End 5/8/2	024, 1:17:09 PM 2024, 1:23:50 PM			00001	08-05-2024 1.	Azure D	atabr	Pre-execute	REST respon	sage REST re ponse REST re ponse REST re ponse ior next
						00001	08-05-2024 1.	Applica	tion I	Execute (Trig	Retrieve RES	iT re
						00001	08-05-2024 1.	Azure D	atabr	Execute (Trig	REST respon	H Seiter
						00001	08-05-2024 1.	Applica	tion I	MonitorCom	Retrieve RES	iT re
						00001	08-05-2024 1.	Azure D	atabr	MonitorCom	REST respon	ise
						00001	08-05-2024 1.	Applica	tion I	Execute (Trig	Waiting for r	next
						00001	08-05-2024 1.	Applica	tion I	MonitorCom	Retrieve RES	iT re
						00001	08-05-2024 1.	Azure D	atabr	MonitorCom	REST respon	ise

Figure 7. Databricks workflow executing in the Control-M monitoring domain.

Output of completed Databricks job in Control-M





Task workflow in Azure Databricks

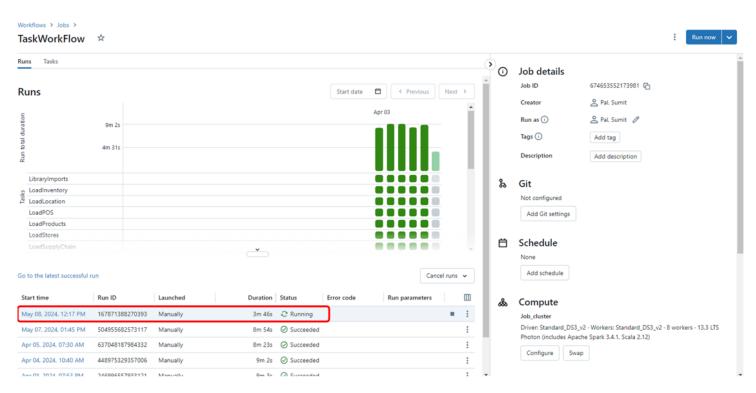


Figure 9. Task workflow in Azure Databricks.

Graph view of task workflow in Azure Databricks

Workflows > Jobs > TaskWorkFlow >

h T	limeline						
		(Loadinventory Secceeded - 55s Diserviupal@bmc.com/.codinvent	0 m	RemoveNullValuedInventorØ Sociodod 21s Dcom/RemoveNuColumnitiventory B .ob.duter	RemoveDupEctARRowsInve0 Invested 18 Dstantemacligatextensionentry B = at_stant	
			LoadPOS Succeeded - 56s D /Juer/Jupak@bmc.com/LoadPOS & Job_cluster	•	RemoveNullValuedPOSCol @ Successed 16s DJ@tmc.com/RemoveNullColumnuPOS R. Job_dumer	RemoveDuplicateRowsPOS 0 Summer 2ts Descender 2ts DseconfiltemenDuplication Descender 1ts 21 B. Interview EnrichPOS	
	LibraryImports © Succeeded -4n 18s	\int	LoadLocation Succeeded - 57s D /Users/supakitienc.com/LoadLocate & 100_cluster		RemoveNullValuedLocationØ Successed 16s Dc.com/femoveNuBColumnsLocation 元 Job_chutter	RemoveOuplicateRovsLoca0 D. //savbagdbmccandbush000a Summed: 16 D. //savbagdbmccandbush000a	
	口 /Vuer/Jupalityber.com/Allimports 最 iob_duster	P	LoadStores Succeeded - 56s D /Juer/Jupakijbmc.com/LoadStores & Job_cluster	• 	RemoveNullValuedStoreColØ Socceedid 16s Distancion/RemoveNullColumeStore R Job_dutter	RemoveDuplicateRoviStoret@ UnearRegressionInventory@ Investment*18s Investments/second/s	
			LoadProducts Succeeded - 52s D /Juer/supaktformc.com/LoadProduc B Job_cluster	21	RemoveNullValuedProduct Ø Socceded 31s Dccom/RemoveNullColumesProducts R Job_duster	RemoveQuplicateRoverProd@ EnrichSupplyChain @ Issuedied 105	
			LoadSupplyChain Succeeded - 56s 	ain	RemoveNullValuedSupplyC @ Succeeded -21s	RemoveDup/scateRowsSup_0 Screeted 2b	

Repair ru

Q []

Figure 10. Graph view of task workflow in Azure Databricks.

Outputs and visualization

Two forms of output were generated during the project.

- Predicted results from the model for POS sales and inventory predictions needed based on demand—these were stored as Delta format files on the lakehouse for offline viewing and analysis.
- Visualizations of the feature weights that contributed to the predictions for POS and inventory data for both random forest and linear regression algorithms.

The four figures below show the feature weights for each of the above algorithms across the different features for POS and inventory predictions from the feature engineered attributes.

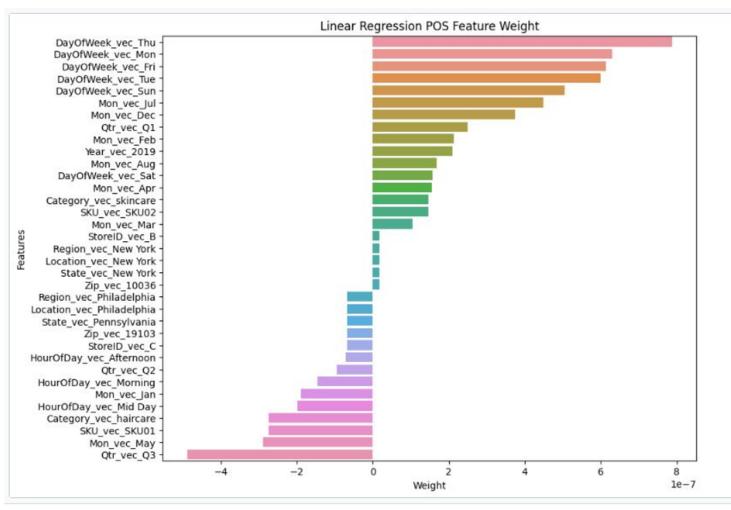
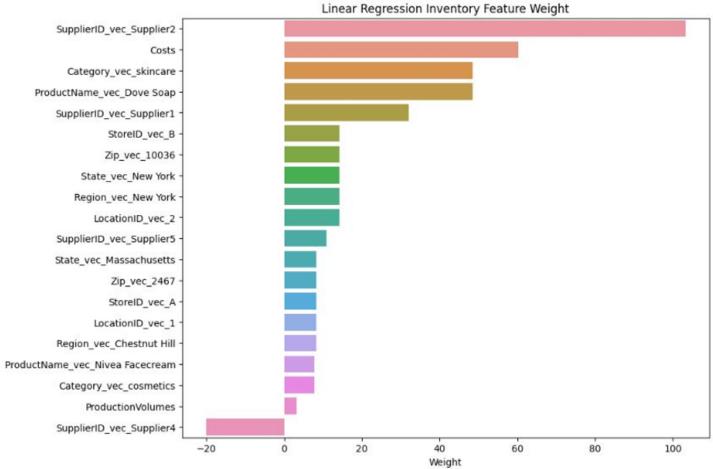
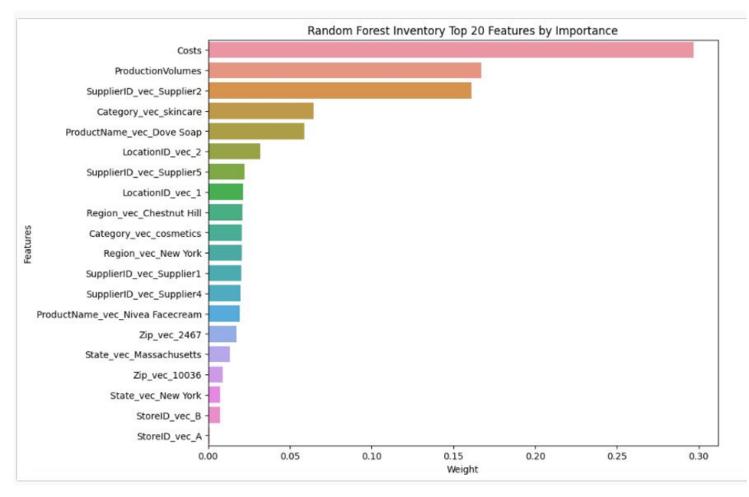


Figure 11. Linear regression POS feature weight.







Features

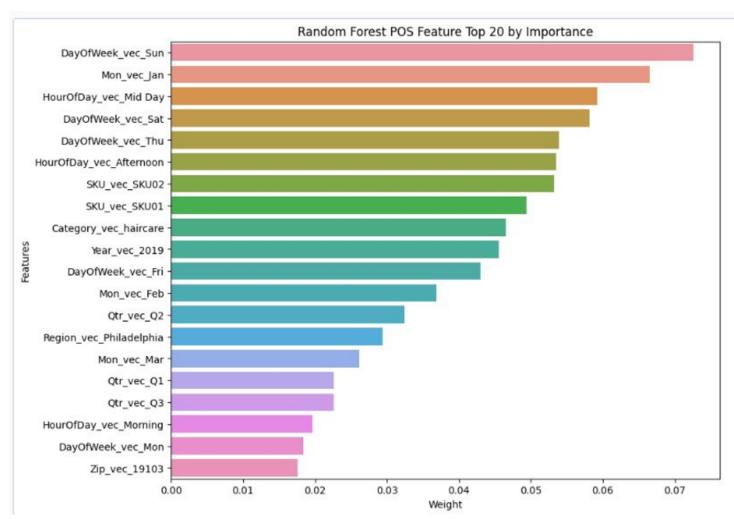
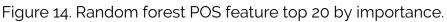


Figure 13. Random forest inventory top 20 features by importance.



Conclusion

This blog demonstrates an ML use case for forecasting of sales and inventory. The ML workflow is likely to be part of a larger orchestration workflow in Control-M. where it is interdependent on workflows running in POS and inventory management applications. However, in this blog, we have maintained the focus on the ML workflow in Databricks and its integration and execution through Control-M.