How to Guide

Data Preparation Integration (DP)

Version: Release 1.1

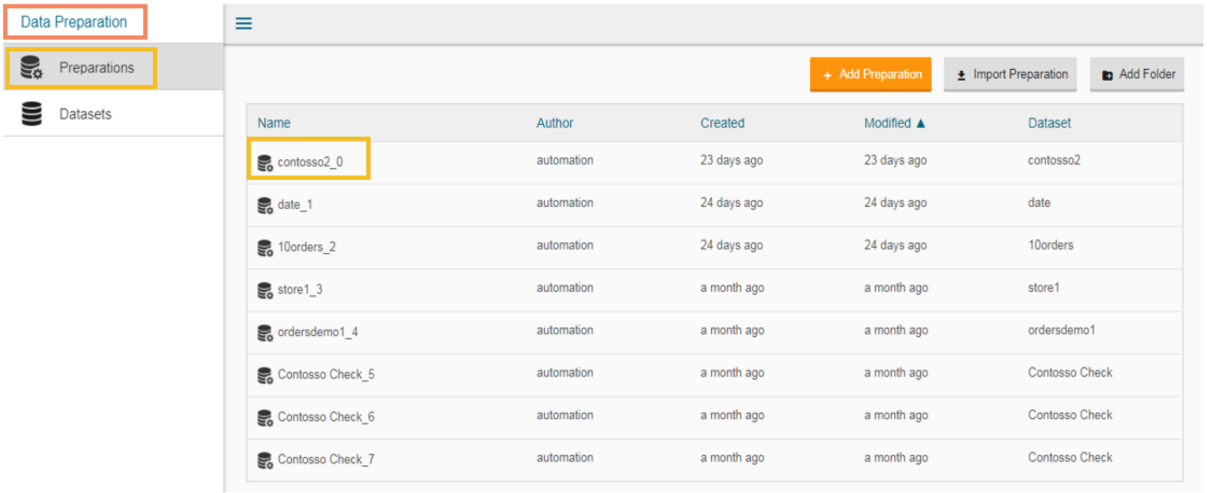
Contents

Data Preparation

Data Preparation is a BDB plugin that is used for cleansing data to prepare it for further analysis. The users can integrate the Data Preparation scripts with Data Pipeline to facilitate the data streaming.

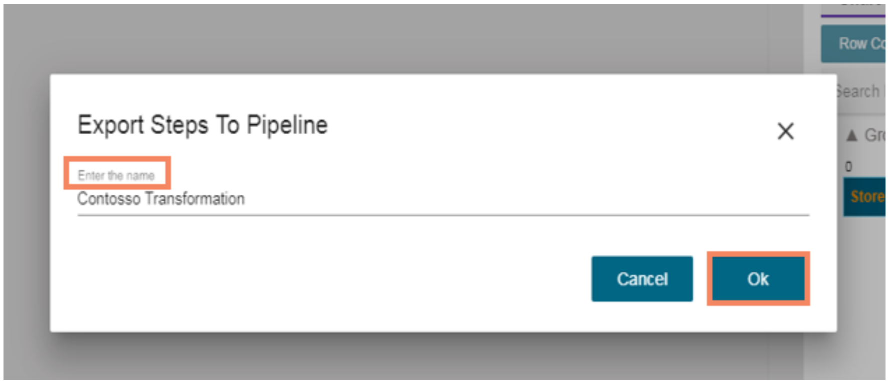
1.1 Deploying a Data Preparation Model to Data Pipeline

1. Navigate to the Data Preparation landing page.
2. A list of all the preparation scripts appears under the '**Preparations**' option
3. Select a Preparation Model from the list



Name	Author	Created	Modified ▲	Dataset
contosso2_0	automation	23 days ago	23 days ago	contosso2
date_1	automation	24 days ago	24 days ago	date
10orders_2	automation	24 days ago	24 days ago	10orders
store1_3	automation	a month ago	a month ago	store1
ordersdemo1_4	automation	a month ago	a month ago	ordersdemo1
Contosso Check_5	automation	a month ago	a month ago	Contosso Check
Contosso Check_6	automation	a month ago	a month ago	Contosso Check
Contosso Check_7	automation	a month ago	a month ago	Contosso Check

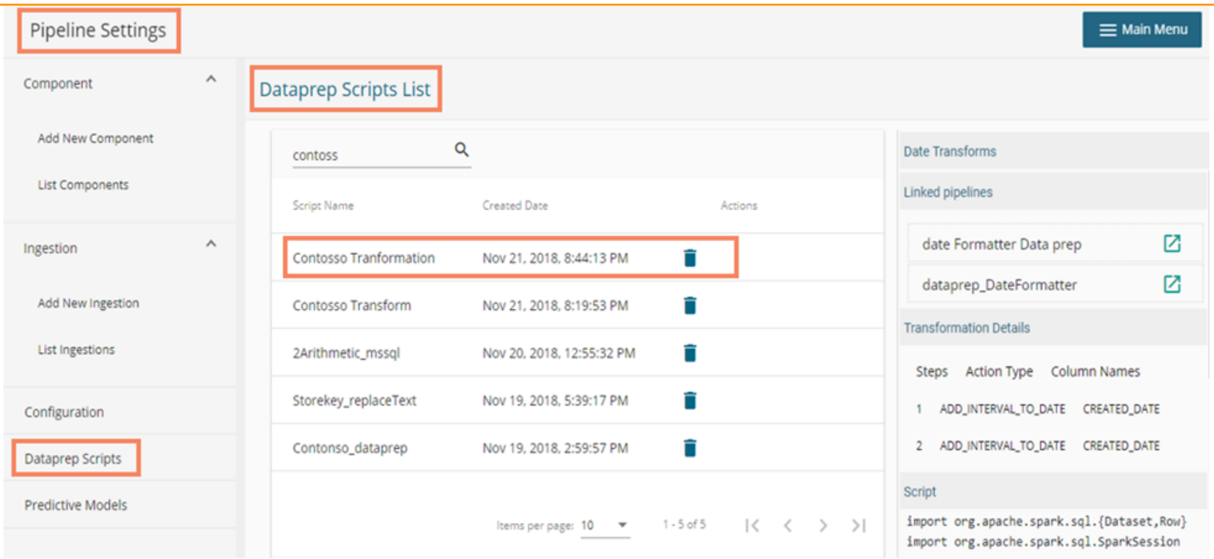
4. A new window '**Export Steps to Pipeline**' appears
   a. Enter a specific title for the Data Preparation model
   b. Click the '**Ok**' option to deploy the model to the Data Pipeline.



Enter the name: Contosso Transformation

1.2 Using the Deployed Data Preparation Model in Data Pipeline

1. Navigate to the Pipeline Settings page
2. Select the Dataprep Scripts from the left panel
3. The '**Dataprep Scripts List**' appears with the deployed Data Preparation script.



contoss

Script Name	Created Date	Actions
Contosso Tranformation	Nov 21, 2018, 8:44:13 PM	
Contosso Transform	Nov 21, 2018, 8:19:53 PM	
2Arithmetic_mssql	Nov 20, 2018, 12:55:32 PM	
Storekey_replaceText	Nov 19, 2018, 5:39:17 PM	
Contonso_dataprep	Nov 19, 2018, 2:59:57 PM	

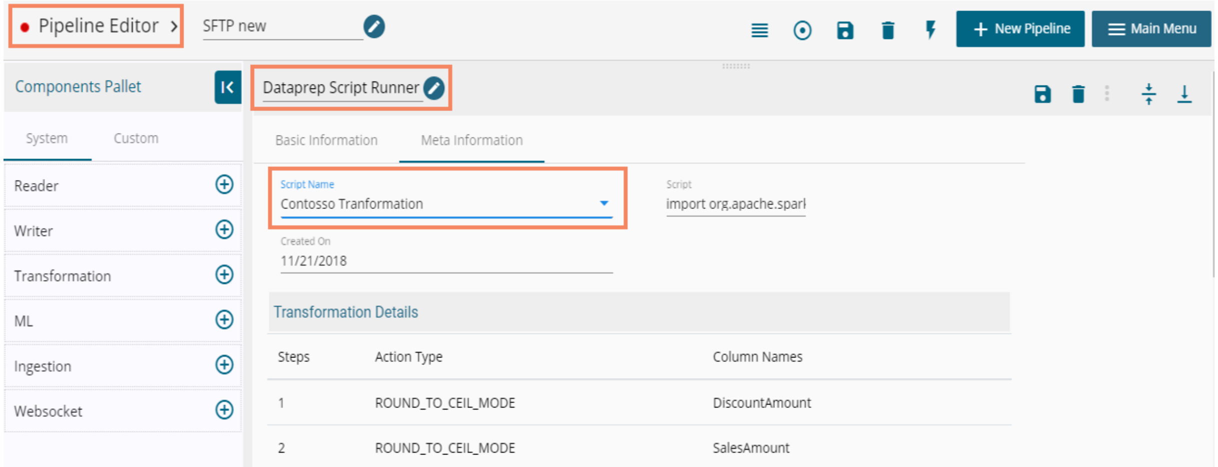
Transformation Details

Steps	Action Type	Column Names
1	ADD_INTERVAL_TO_DATE	CREATED_DATE
2	ADD_INTERVAL_TO_DATE	CREATED_DATE

Script

```
import org.apache.spark.sql.{Dataset,Row}
import org.apache.spark.sql.SparkSession
```

4. The users can access the deployed Data Preparation script from the '**Script Name**' drop-down menu that appears inside the Dataprep Script Runner component of the Data Pipeline.



Script Name: Contosso Tranformation

Script: import org.apache.sparl

Created On: 11/21/2018

Transformation Details

Steps	Action Type	Column Names
1	ROUND_TO_CEIL_MODE	DiscountAmount
2	ROUND_TO_CEIL_MODE	SalesAmount