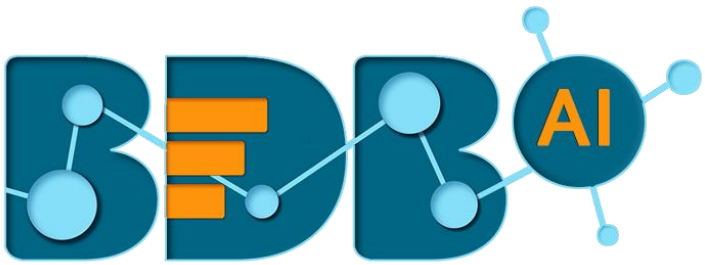


User Guide

ETL R-4.0

Contents

1. About this Guide

1.1. Document History

Product Version	Release Date	Description
BDB Data Preparation (ETL) 1.0	August 31st, 2017	First Release of the document
BDB Data Preparation (ETL) 1.1	December 11th, 2017	Updated document
BDB Data Preparation (ETL) 1.2	April 15th, 2018	Updated document
BDB ETL 3.8	December 1st , 2018	Updated document
BDB ETL 4.0	December 31st, 2018	Updated document

Note: ETL 3.8 onwards the plugin Product Version reflects the BDB Platform release version.

1.2. Overview

This guide covers:

- Introduction and steps to use various components of the BDB ETL plugin
- Configuration details of the ETL components

1.3. Target Audience

This guide is aimed at business users of all skill levels who deal with vast amounts of data and requires data preparation to be attempted before getting informative insights from the collated business datasets.

2. Introduction

2.1. Introducing the BDB ETL

The BDB ETL is a self-service data preparation tool that empowers data-driven Business users with powerful capabilities to extract, transform, and load new data sources. The tool offers a range of

components to transform and merge the selected dataset. Users can get analytics-ready data faster to generate valuable insights in less time.

2.2. Supported Web Browsers

The BDB Platform is a web browser-based application. The users can run the BDB Platform and its various plugins on the below given versions of the browsers:

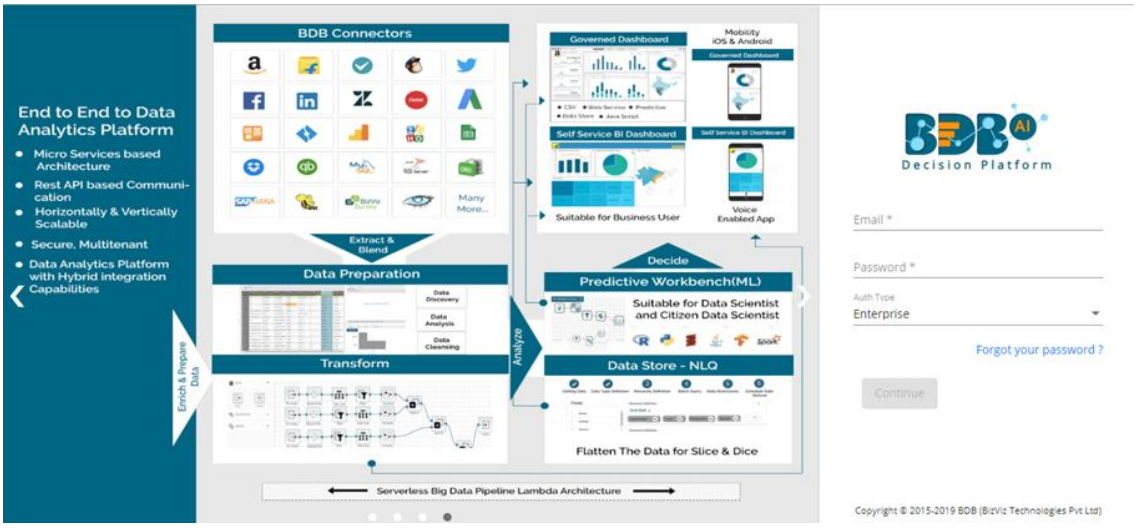
Mozilla Firefox/ Firefox ESR	Latest Version
Microsoft Internet Explorer	11
Microsoft Edge	Latest Version
Apple Safari	10
Google Chrome	Latest Version (recommended web browser)

3. Getting Started with the BDB ETL

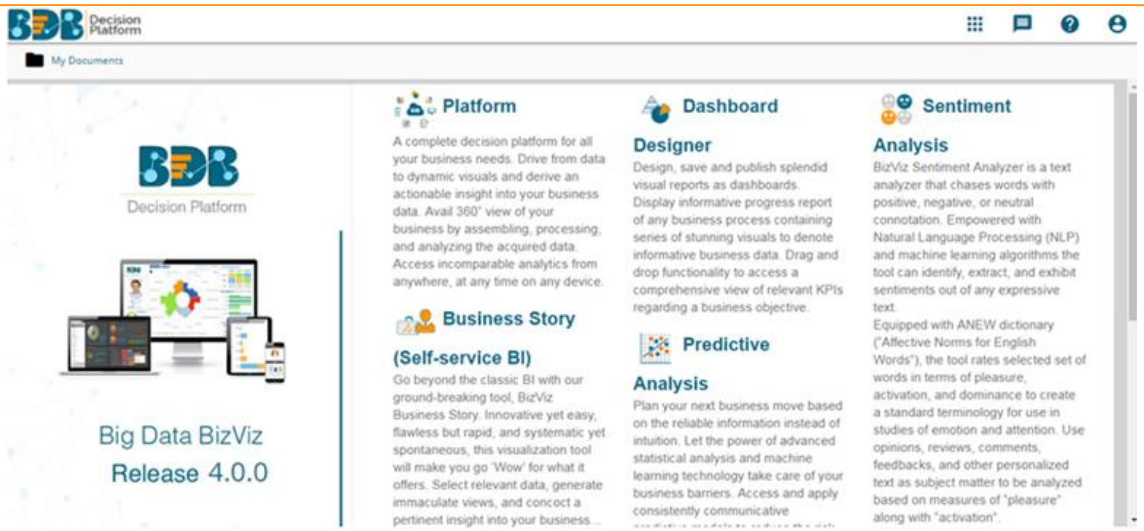
3.1. Accessing the ETL Plugin

This section explains how to access the BDB Platform and a variety of plugins that it offers:

i) Open BDB Enterprise Platform Link: https://app.bdb.ai
ii) Enter your credentials to log in.
iii) Click the '**Continue**' option.



iv) BDB Platform homepage opens.
v) Users get redirected to the BDB Platform homepage.



Note:

a. The above screen opens only for those newly created users who have not yet created any document using the BDB Platform.
b. If the user has created some documents previously, then the Platform homepage opens displaying the '**My Documents**' page by default.

vi) Click the '**Apps**' icon.
vii) All the available plugin applications get displayed.
viii) Select the '**ETL**' plugin.

ix) The ETL landing page opens.
x) Four major modules get displayed on the ETL landing page:

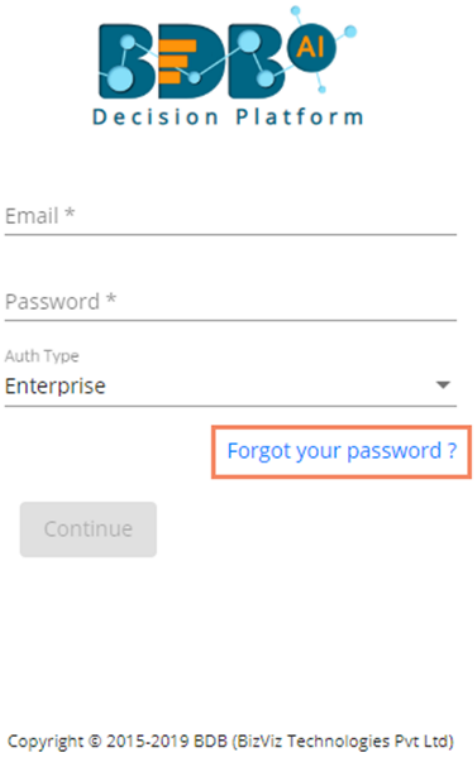
a. My Workspace (Default Component)
b. Job
c. Trash
d. Scheduler

This document aims to describe all the significant components and the related workflows at details.

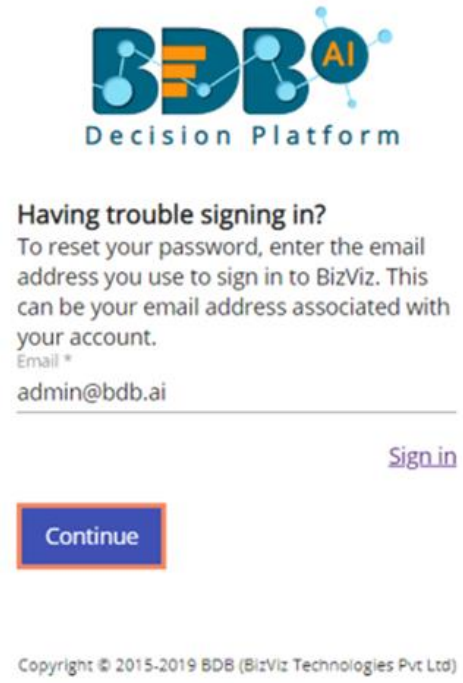
3.1.1. Forgot Password Option

Users are provided with a choice to change the password on the Login page of the platform.

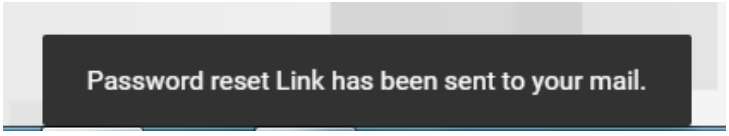
i) Navigate to the Login page.
ii) Click '**Forgot your password?**' option.



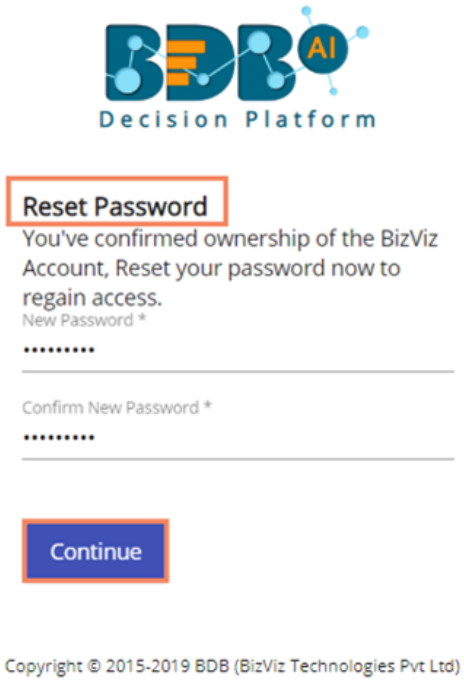
iii) A new window opens.
iv) Provide the email id that is registered with BDB to send the reset password link.
v) Click the '**Continue**' option.



vi) Users may be redirected to select a space in case of multiple spaces under one server link(They need to select a space and click the 'Continue' option once again). If users do not have multiple spaces then, a message appears to notify the user that about the password reset link (The users receive the reset link via their registered email.)



vii) Click the link from your registered email.
viii) Users get redirected to the '**Reset Password**' page to set a new password.
ix) Set a new password.
x) Confirm the newly set password.
xi) Click the '**Continue**' option.

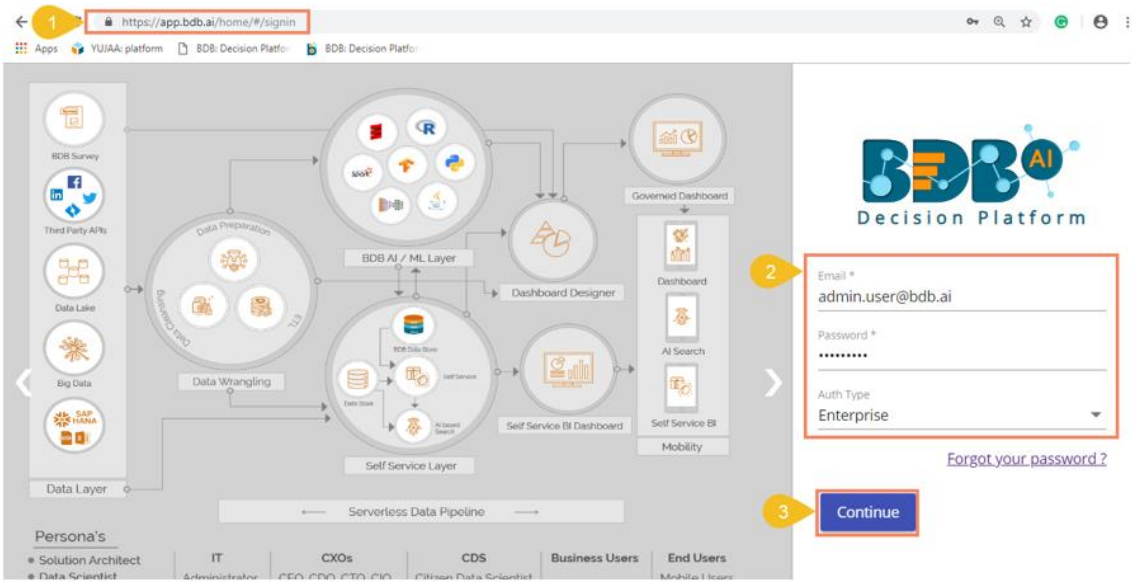


xii) The password for the selected BDB account gets reset.

3.1.2. Force Login

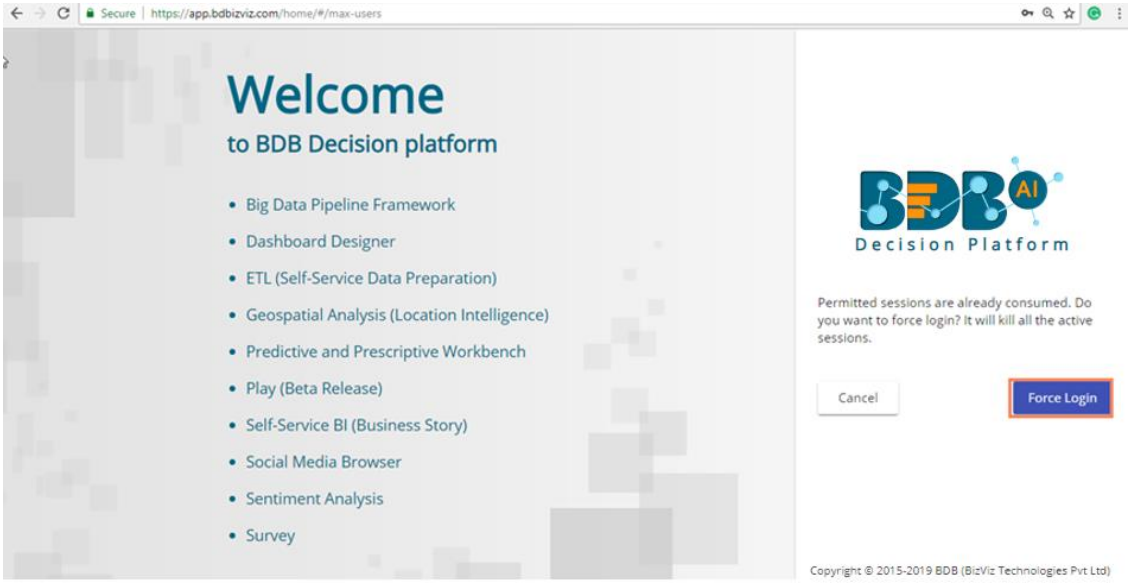
The '**Force Login**' functionality has been introduced to control the number of active sessions up to three. The users can access only 3 sessions at a time when they try to access 4th session a warning message displays to inform that the user has consumed the permitted sessions and a click on the '**Force Login**' would kill all those active sessions.

i) Navigate to the BDB Platform Login page.
ii) Enter the valid credentials to log in.
iii) Click the '**Continue**' option.



iv) The user gets the following message if the user already consumes the permitted active sessions (3 sessions at a time).

v) Click the '**Force Login**' option.



vi) A warning message appears that the currently active sessions get killed for the user and the user has redirected to the log in a page of the BDB Platform.

Note: The user can successfully login to the BDB Platform after selecting the '**Force Login'** option to log in the platform.

4. Basic Features

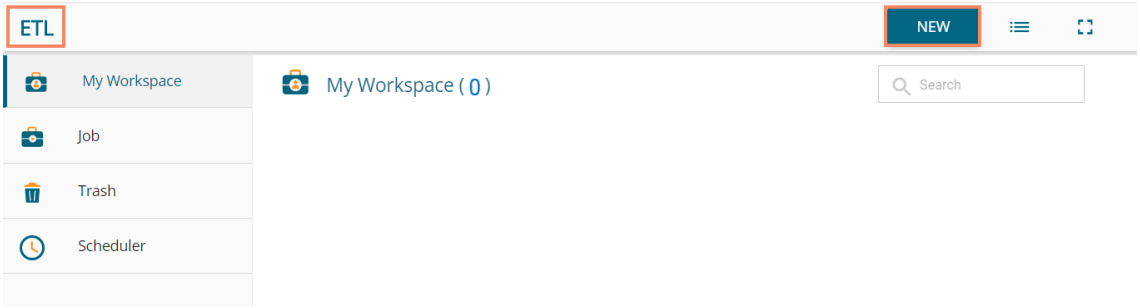
The landing page of ETL launches workspace view. '**My Workspace**' will be displayed by default.

4.1. Workflow Editor

'**My Workspace**' is a placeholder for the workflows which are created using various ETL components. Users can create workflows using the workflow editor.

i) Navigate to the ETL landing page.

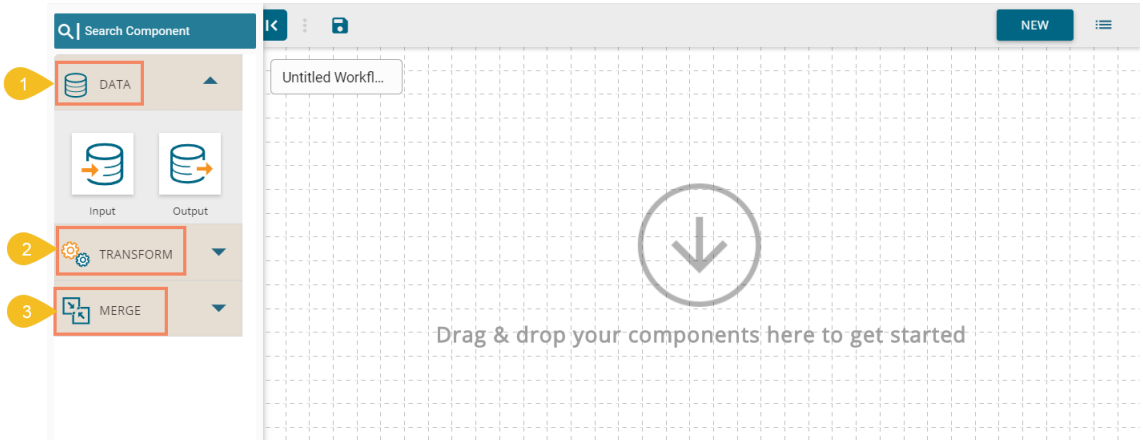
ii) Click the '**NEW**' option.



iii) The user gets redirected to the **Workflow Editor.**

iv) The Workflow editor exposes the user to the following aspects to autonomously prepare data:

a. Data
b. Transform
c. Merge

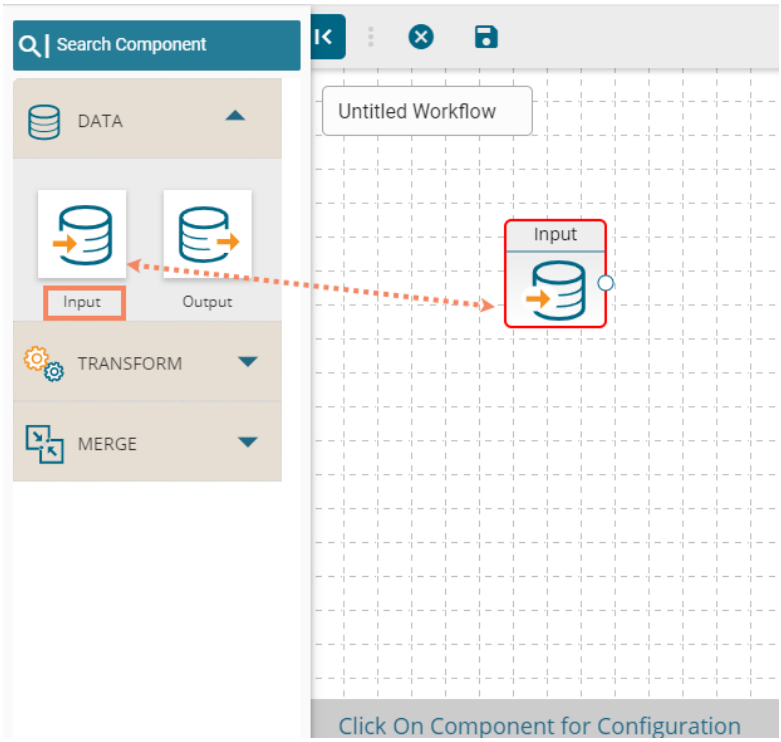


4.2. Extracting Data: Full and Incremental

i) Navigate to the Workflow Editor.

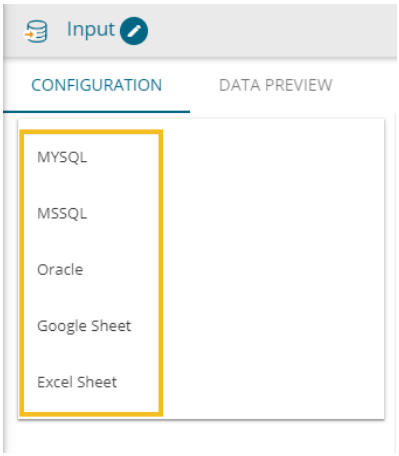
ii) The '**Data**' option gets selected by default.

iii) Drag and drop the '**Input**' component onto the workflow editor.



iv) Click on the dragged Input component to get the configuration tab.

v) Select a database type using the '**Select Source Type'** drop-down menu (At present only MYSQL, MSSQL, Oracle, Google Sheet, and Excel Sheet are supported).



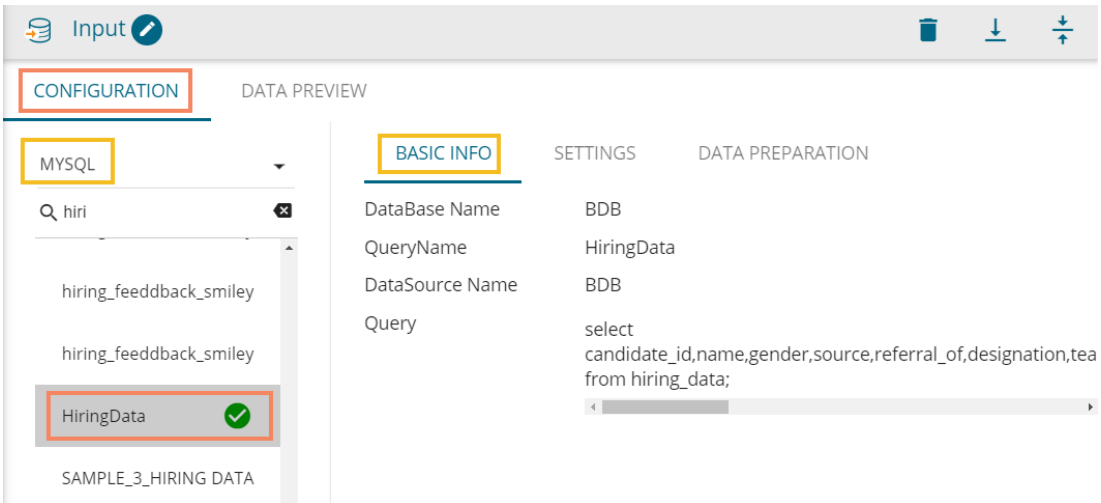
E.g., In this case, the selected data source type is **MySQL**.

vi) A list of Query Services based on the MySQL database appears.

vii) Select a Query Service from the list.

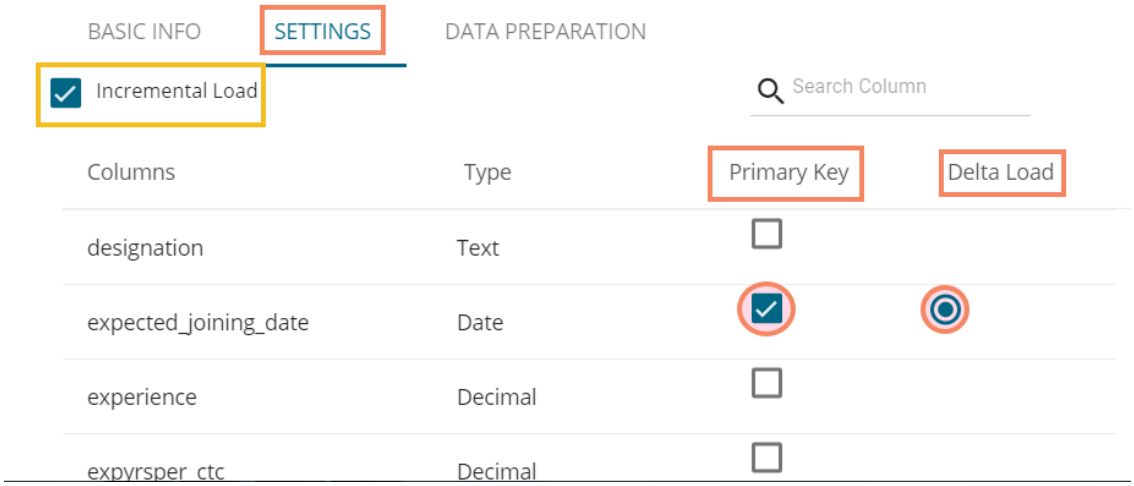
a. BASIC INFO

i. The '**BASIC INFO**' tab opens (by default) describing the basic information of the selected Database and Query Service.

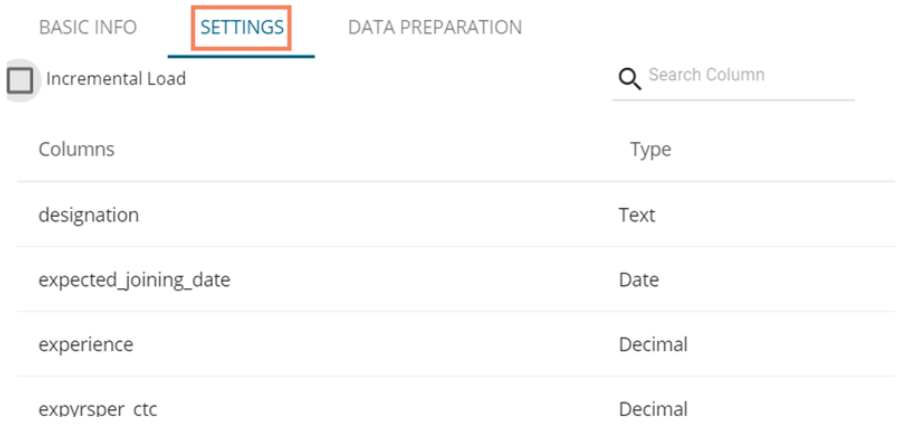


b. SETTINGS

i. Click the '**SETTINGS**' tab.
ii. The '**Incremental Load**' option appears.
iii. Enable the Incremental Load to get the recently updated data.
iv. The user needs to configure the following options after enabling the Incremental Load:
    1. **Primary Key**- Select a primary key of the data source by using a checkmark in the given box.
    2. **Delta Load**-Select a column of type timestamp or date or long which is updated whenever a new row is inserted or updated in the data source. This column will be used to load the **Incremented data.** Use the radio button to select a '**Delta Load**' column.



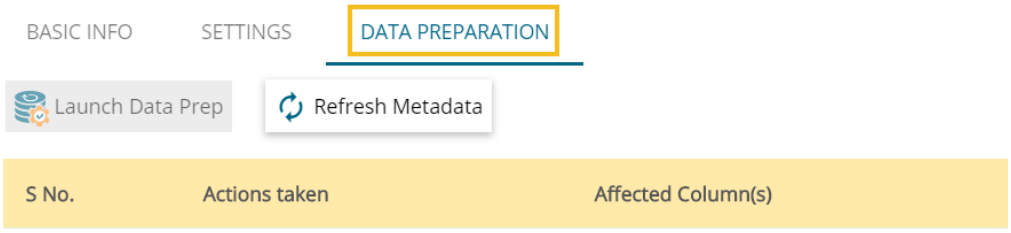
Note: The users can choose not to enable the '**Incremental Load**' option. In this case, the following details get displayed, and the complete data gets extracted.



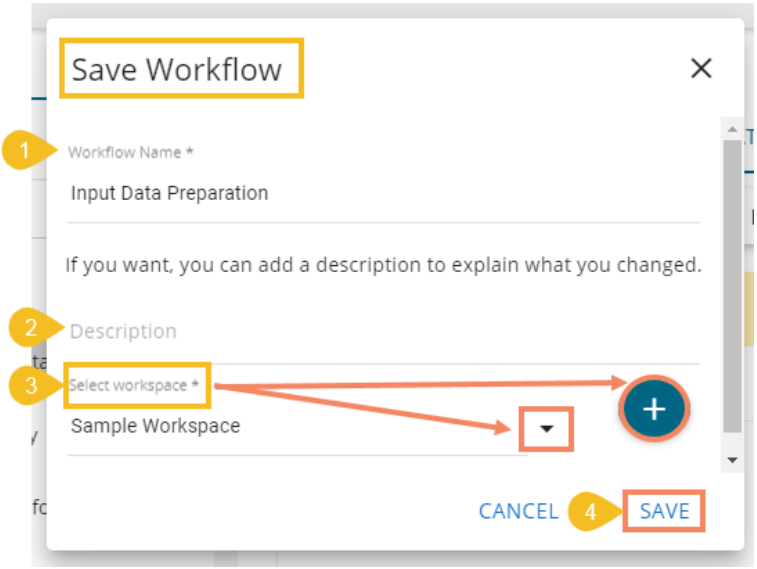
c. DATA PREPARATION

i. Click the '**DATA PREPARATION**' tab.
ii. Two options get displayed to proceed with the action of data preparation.
    1. Launch Data Prep

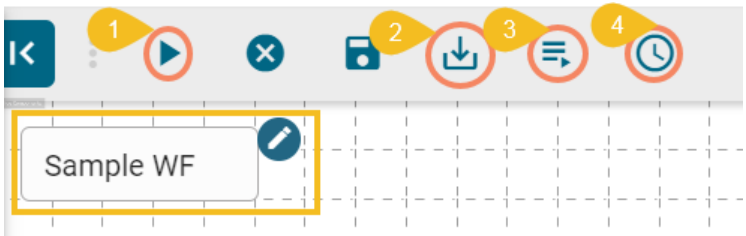
2. Refresh Metadata



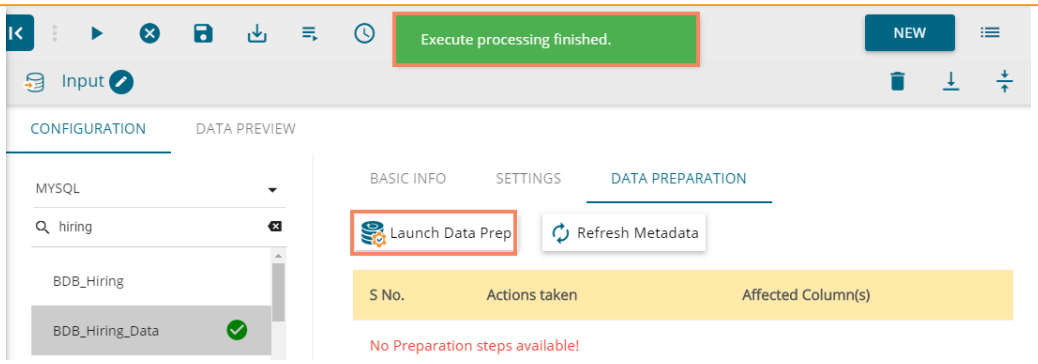
iii. A window opens prompting to save the workflow.
   1. Provide a Workflow Name.
   2. Describe the workflow (Optional).
   3. Select a Workspace from the drop-down menu or create a new workspace.
   4. Click the '**SAVE**' option.



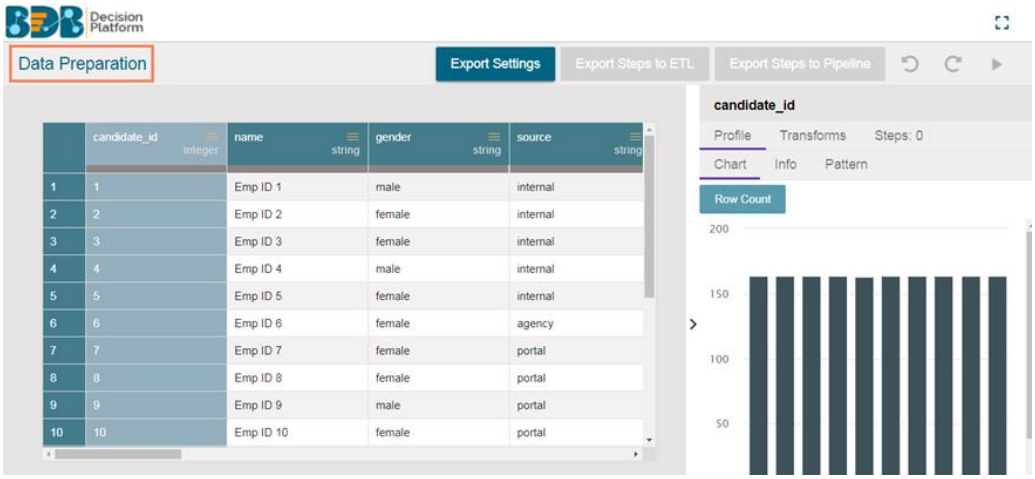
iv. The workflow gets saved by the given name, and it displays on the workspace.
v. The following icons appear in the workflow editor taskbar:
   1. Save & Run Preview
   2. Save As
   3. Save and Execute
   4. Schedule



vi. The '**Launch Data Prep**' gets enabled only after saving the workflow.
vii. Click the '**Launch Data Prep**' option.

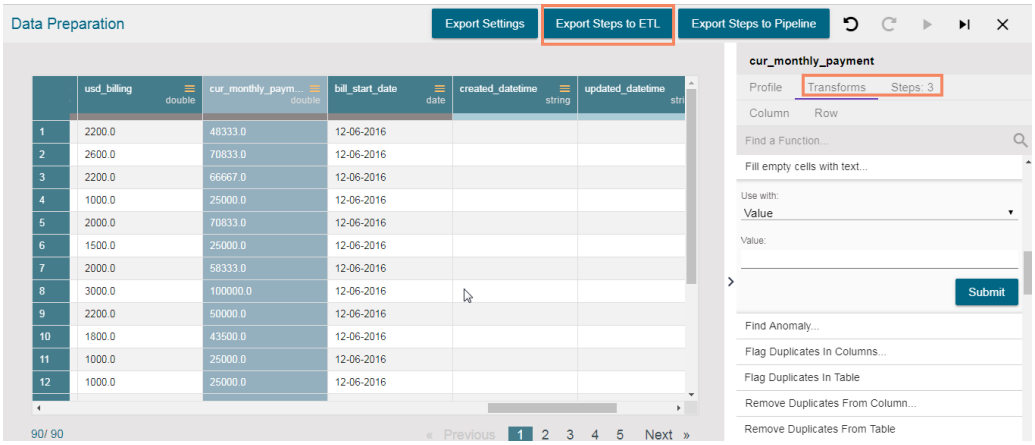


viii. The Data Preparation plugin gets launched.



ix. The user can perform desired transform using the '**Transforms**' tab provided on the Data Preparation.

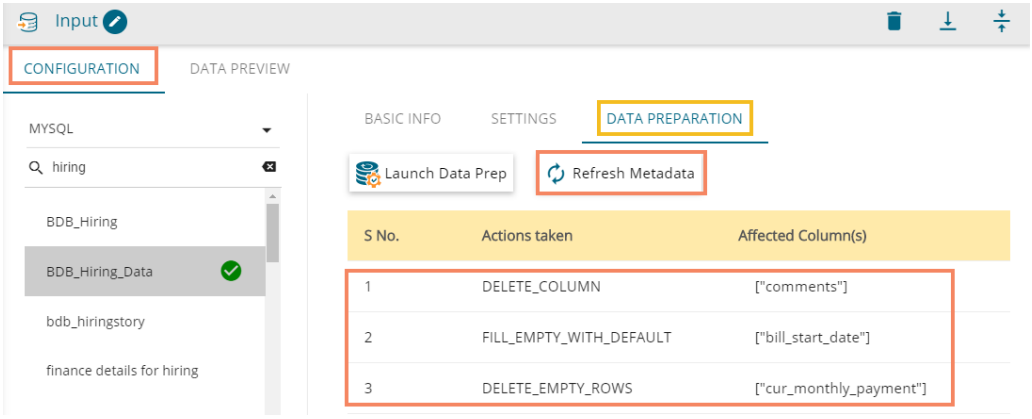
x. Click the '**Export Steps to ETL**' option.



xi. A success message appears to inform about the completion of the export action.

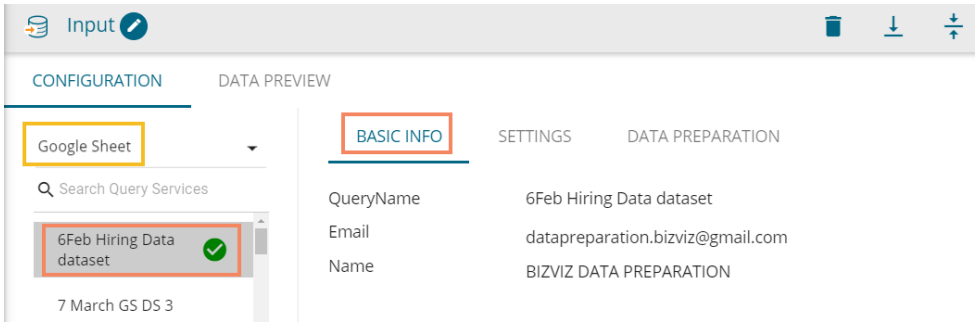
xii. Click the 'Refresh Metadata' option using the 'DATA PREPARATION' tab.

xiii. All the performed transforms get listed at the bottom of the tab.

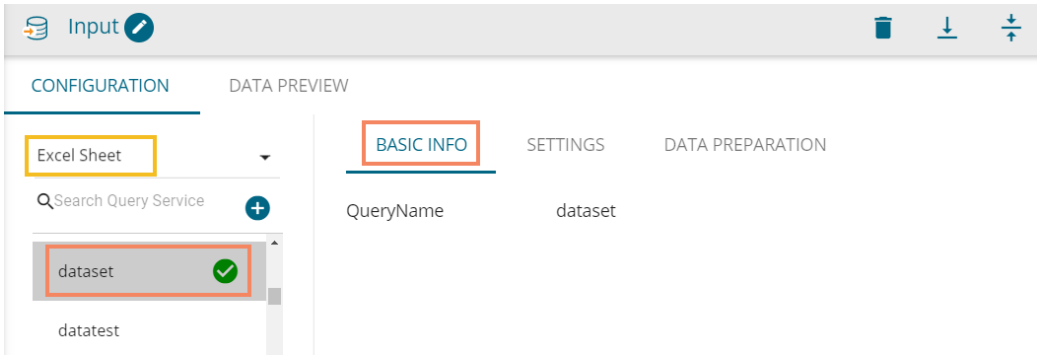


Note:

a. The '**BASIC INFO**' tab varies based on the selected Input data source types:
   i. The '**BASIC INFO**' tab opens the same information about database and query service for the MSSQL and Oracle data sources types as displayed for the MySQL data source type.
   ii. The '**BASIC INFO**' tab displays the query information and the concerned email address for the Google Sheet data source type.



iii. The '**BASIC INFO**' tab for the Excel Sheet displays only query name.



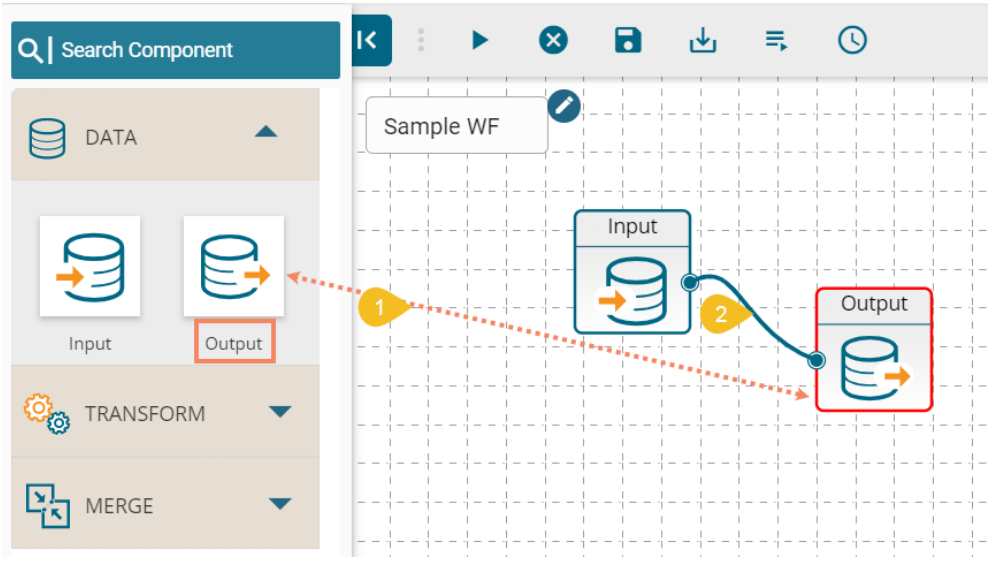
b. The '**SETTINGS**' and '**DATA PREPARATION**' tabs display the same set of information as described above for all the Input data source types.
c. The user can prepare the input data using the Data Preparation module. Refer the Data Preparation UG for more details.

4.3. Loading Data

The user can load the extracted data into Output Connectors for visualization via the output component.

i) Drag and drop the '**Output**' component on the Workflow editor.

ii) Connect it with the configured '**Input**' component.

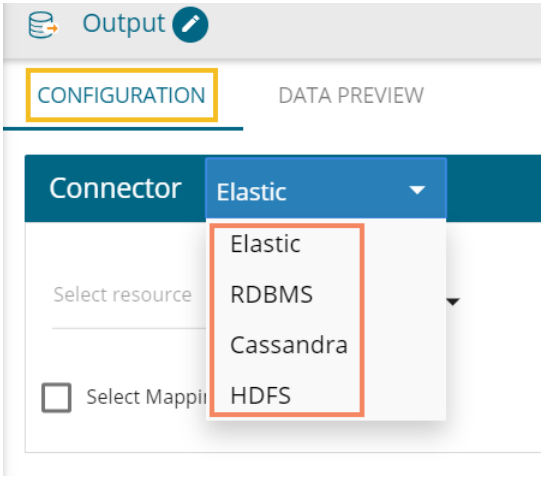


iii) Click on the '**Output**' component to display the '**CONFIGURATION**' option.

iv) The following options get displayed:

a. Elastic
b. RDBMS
c. Cassandra
d. HDFS

v) Select any one option to access the CONFIGURATION tab.



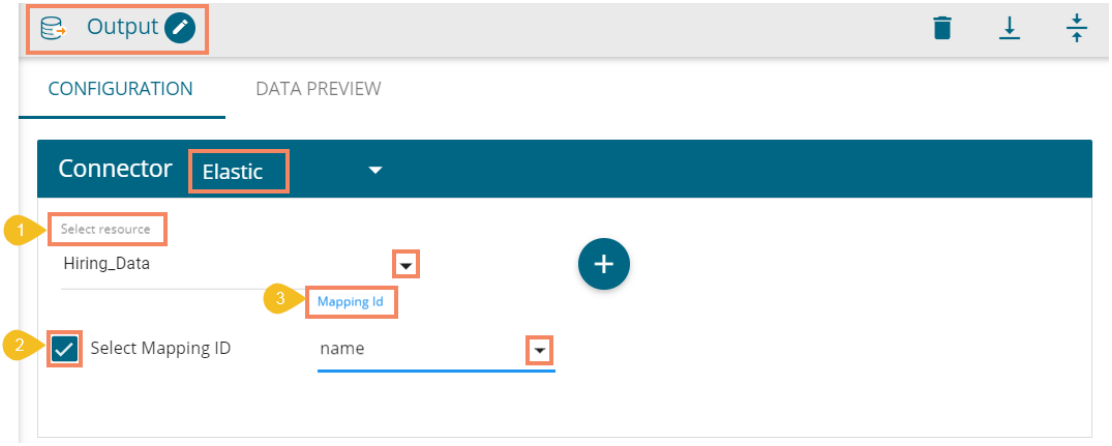
4.3.1. Configuring Elastic

The following configuration fields open when the selected Output option is Elastic.

i) Select a resource using the drop-down menu for the Elastic writer.

ii) Enable '**Select Mapping ID**' option.

iii) After enabling the '**Select Mapping ID'** option, the user gets redirected to select a mapping id from the '**Mapping id**' drop-down menu.



Note:

a. The '**Mapping Id**' drop-down menu appears when the user enables the '**Select Mapping ID**' option with a checkmark.

b. Click the '**Create Meta Data**' icon to open the configuration window for creating metadata.

4.3.2. Configuring RDBMS

The following configuration fields appear for the RDBMS output connector.

i) Select a Data Source Type from the drop-down menu.

ii) Select a Data Source Name from the drop-down menu.

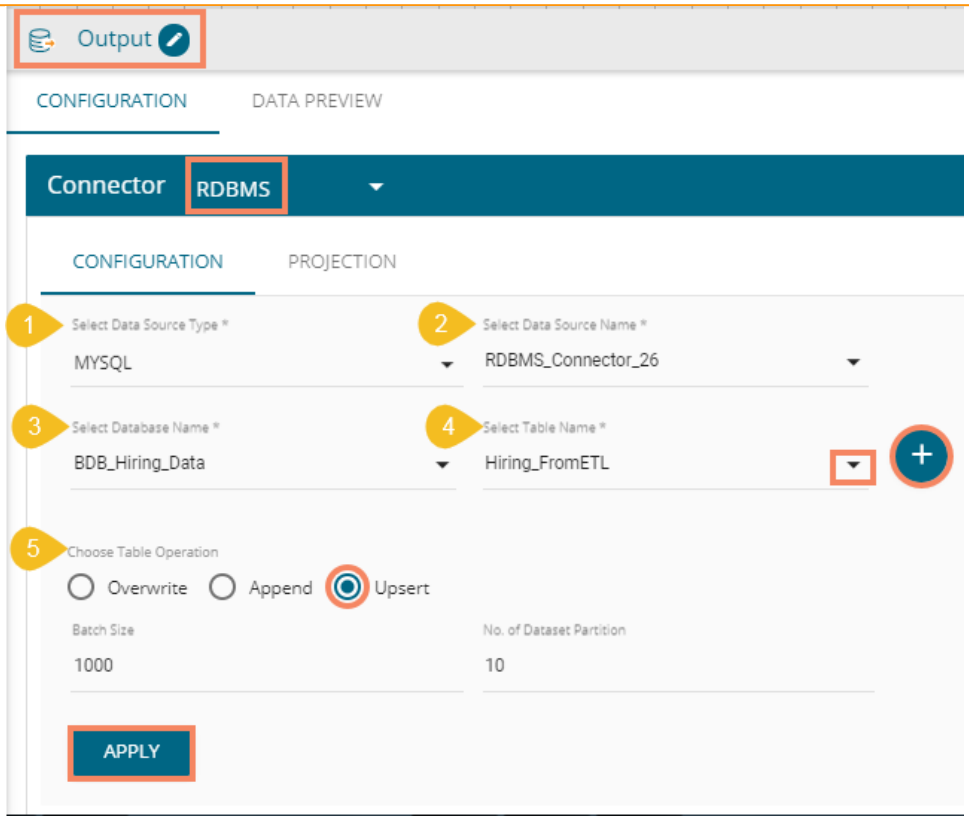
iii) Select a Database Name from the drop-down menu.

iv) Select a Table Name from the drop-down menu or Click the '**Add New Table**' icon to Create a New Table.

v) Choose a Table Operation from the given choices:

1. Overwrite: By choosing Overwrite as operation, the existing records gets overwritten in the selected table.
2. Append: By choosing Append as operation, the extracted records get added at the end of the existing records in the columns of the selected file or table.
3. Upsert: By choosing Upsert as operation, only new records get added to the file or selected table.

vi) Click the '**APPLY**' option.

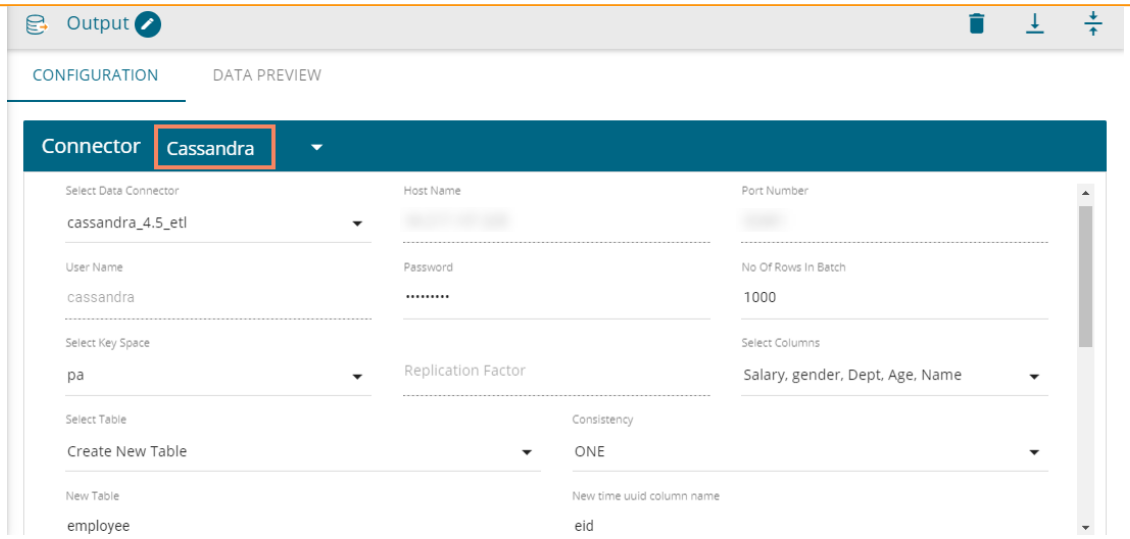


Note: The user gets '**Batch Size**' and '**No. of Dataset Partition**' options only when 'Upsert' is selected as the table operation.

4.3.3. Configuring Cassandra

The following configuration fields appear for the Cassandra output connector.

i) Select a Data Connector from the drop-down menu.
ii) Host Name: Displays a preselected Host Name based on the selected data connector.
iii) Port Number: Displays a preselected Port Number based on the selected data connector.
iv) User Name: Displays a preselected username based on the selected data connector.
v) Password: Enter the Password
vi) No. of Rows in Batch: Set number of Rows that you wish to allow in one Batch (the default value for this field is 1000)
vii) Select Key Space: Select a Key Space from the drop-down menu
viii) Replication Factor: Enter the Replication Factor
ix) Select Columns: Select Columns using the drop-down menu
x) Select Table: Select an existing table from the drop-down menu or choose the '**Create New Table**' option to create a new table.
xi) Consistency: Select a Consistency option from the drop-down menu
xii) New Table: Provide a title to the newly created table using the '**New Table**' field
xiii) New time uuid column name: Provide a name for the new Time UUID Column



xiv) **Headers**: All the columns from the data set will be listed.

xv) **Partition Key**: The Partition Key determines which node stores the data. It is responsible for data distribution across the nodes.

a. The UUID Column name gets displayed under the '**Partition Key**' window.

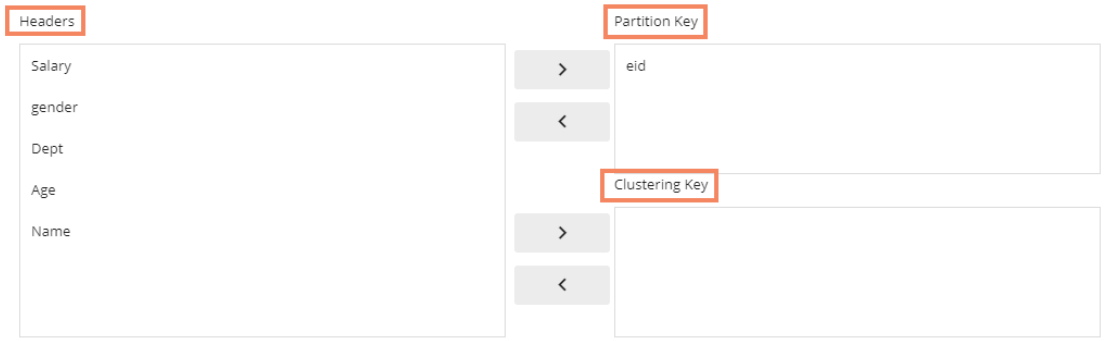
b. The user can select and move any column from '**Header**' (Select Column) to '**Partition Key**' space.

c. The sequence of the columns listed under Partition Key can be arranged by using '**Up**' or '**Down**' options.

xvi) **Clustering Key**: The Clustering Key is a storage engine process that sorts data within the partition. It determines per-partition clustering.

a. The items listed under the Clustering Key box can be arranged by using '**Up**' or '**Down**' options.

b. Users can select any column from '**Headers**' (Select Column) to '**Clustering Key**' space.



4.3.4. Configuring HDFS

The following fields appear for the HDFS output connector.

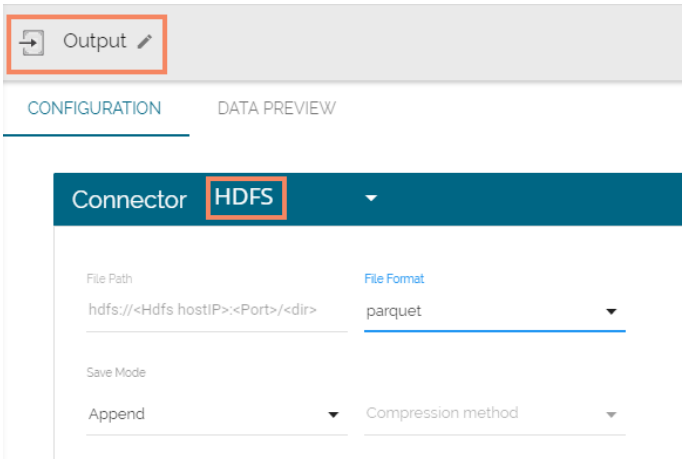
i) Provide file path

ii) Select a File Format from the below given choices in the drop-down menu.

1. Parquet
2. Json
3. Avro
4. CSV

iii) Select a Save Mode from the below given choices in the drop-down menu.
1. Append
2. Overwrite
3. Error
4. Ignore

iv) Select a Compression Method from the below given options in the drop-down menu.
1. Gzip
2. Snappy
3. None

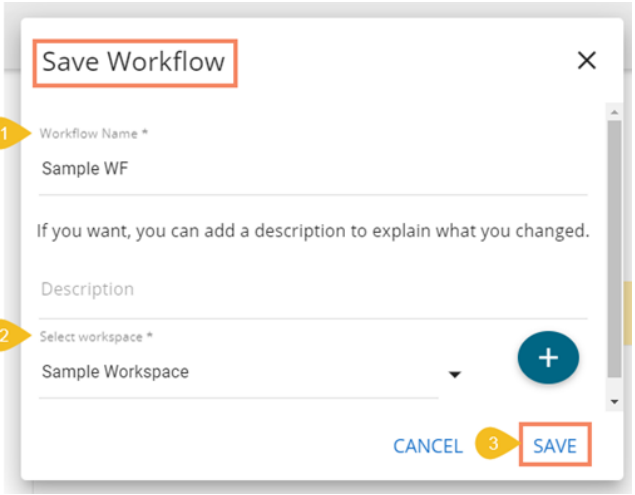


Note: The user needs to run or execute the workflow after configuring the Cassandra and HDFS output connectors.

4.4. Saving a Workflow

Users are provided with two options to save a workflow.

i) Click the '**Save**' icon from the workflow editor header.
ii) A new window appears to redirect the users to save the workflow.
   a. Workflow Name: Provide a name for the Workflow (mandatory field)
   b. Description: Enter Description for the Workflow (Optional)
   c. Select Workspace: Select a workspace from the drop-down menu or Add a new Workspace by clicking the '**Add**' icon.
iii) Click the '**SAVE**' option.

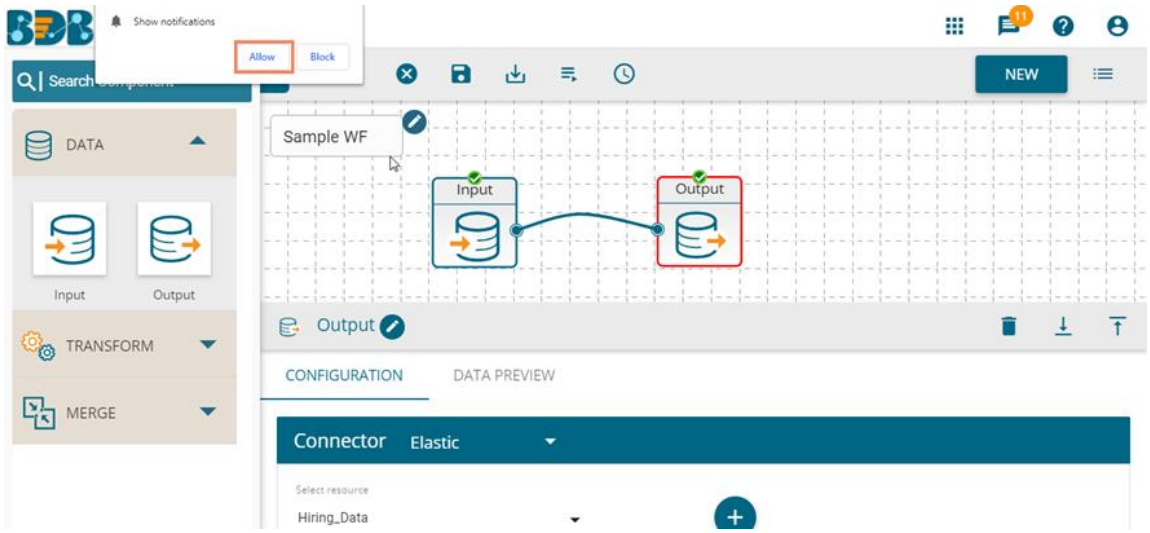


Note: The user can use the '**Save As**' icon if they wish to save it in another workspace.

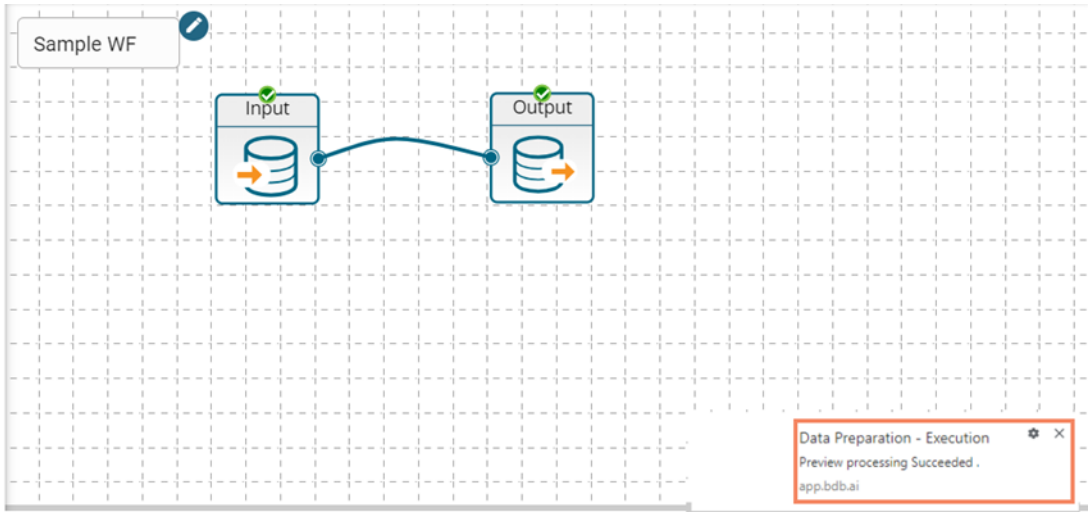
4.5. Save & Run Preview

The '**Save & Run Preview**' icon appears only for the saved workflows. The users can get the data preview only after running the workflow using this option.

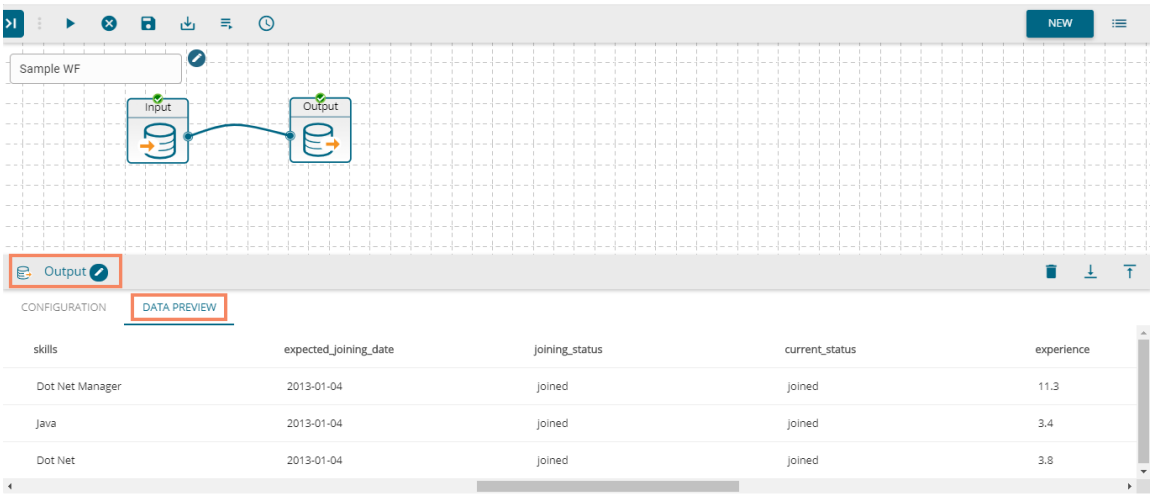
i) After saving a workflow, Users will be able to access the '**Save & Run Preview**' option on the workflow editor toolbar.
ii) Click the '**Save & Run Preview**' option.
iii) The ongoing execution process will be displayed through a continuous blue line.
iv) Users will get notified about the beginning and end of the execution process by pop-up messages.
v) After the execution gets completed, a green tick mark appears on the top of each component in the workflow. The input data with a green checkmark is ready to preview.
vi) A pop-up message appears asking permission for showing the notifications.
vii) Click the '**Allow**' option to display the notification.



viii) A notification appears to inform about the preview process status.



ix) Open the '**Data Preview**' tab by clicking the input or output component to view the preview of the extracted or loaded data.

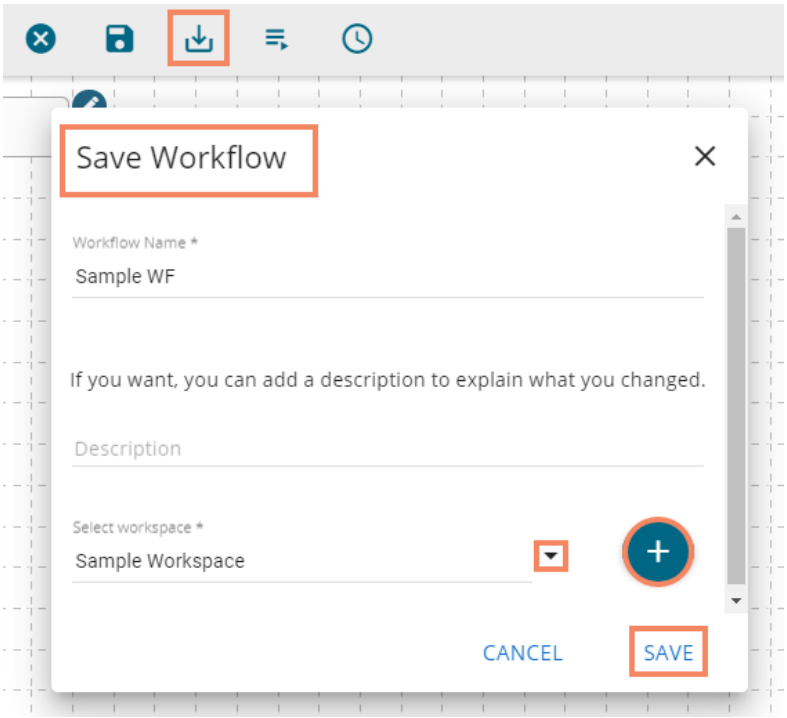


Note: The user will get notifications on the screen for success or failure of the preview processing.

4.6. Save As

By using the '**Save As**' icon, the user can save the workflow to a different Workspace.

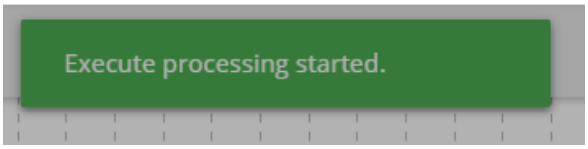
i) Click the '**Save As**' icon.
ii) Modify the workflow name.
iii) Modify the workflow description.
iv) Change the Workspace or create new workspace using the '**Add**' icon.



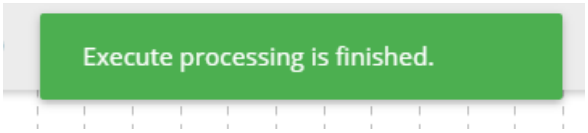
4.7. Save and Execute

By using the '**Save and Execute**' option users can save and write a workflow in the metadata to create a datastore out of it.

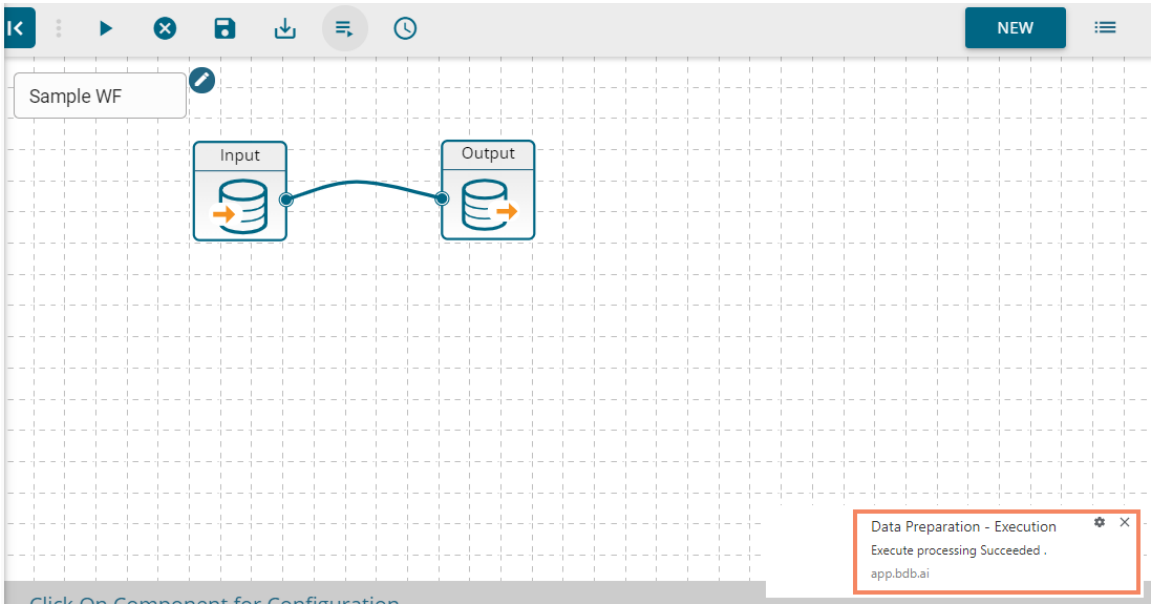
i) Click the '**Save and Execute**' icon from the workflow editor taskbar.
ii) A message appears to inform the user that the execution process has been started.



iii) The user gets another message to obtain confirmation about the completion of the execution process.



iv) Another notification appears at the bottom of the page displaying the status of the execution process.

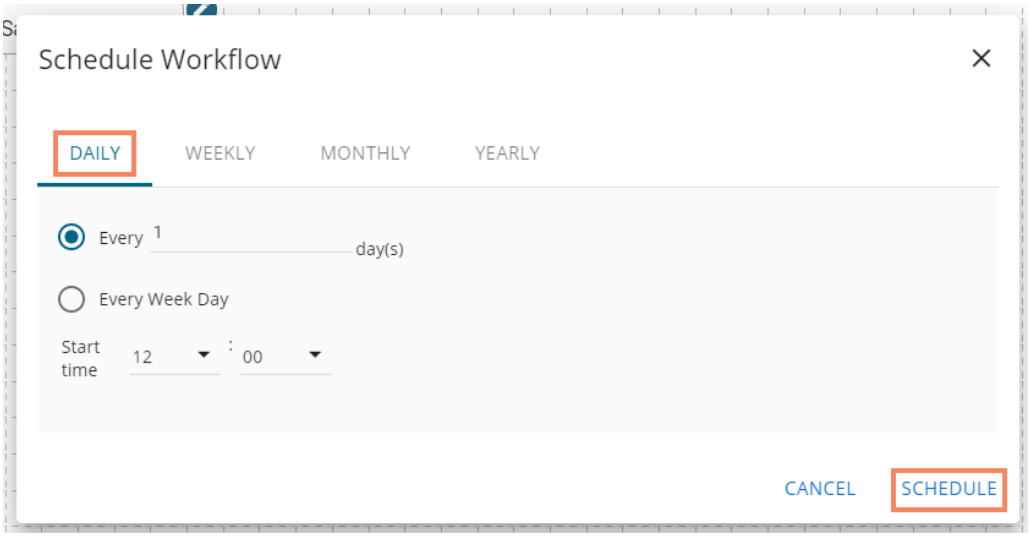


4.8. Schedule a Workflow

Users can schedule a created workflow for data refresh. The Schedule option appears only for the saved workflows.

i) Create a workflow.
ii) Save and run the workflow.
iii) Click the '**Schedule**' icon.
iv) Click a range of time.
v) Set the required information asked for the selected time range. E.g., The below-given image displays Schedule Workflow configuration details for the '**DAILY**' option.

vi) Click the '**SCHEDULE**' option.



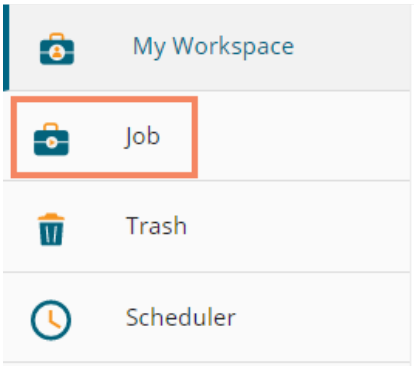
vii) The selected workflow gets scheduled for the data refresh as per the configured information.

4.9. Job

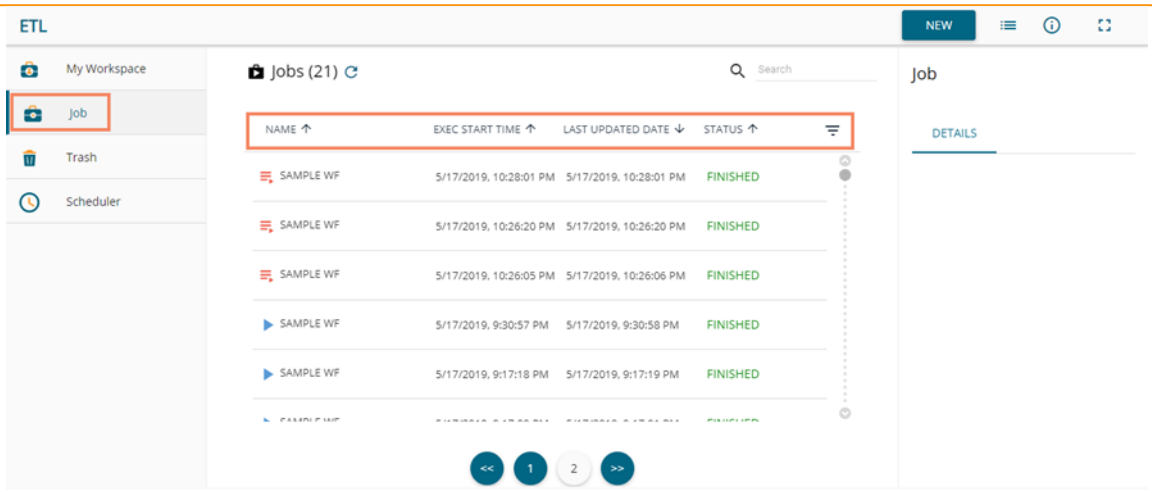
Users can see the job status for the saved workflows.

i) Navigate to the ETL landing page.
or
Click the icon from the workflow editor.

ii) Select the '**Job**' option from the menu list.

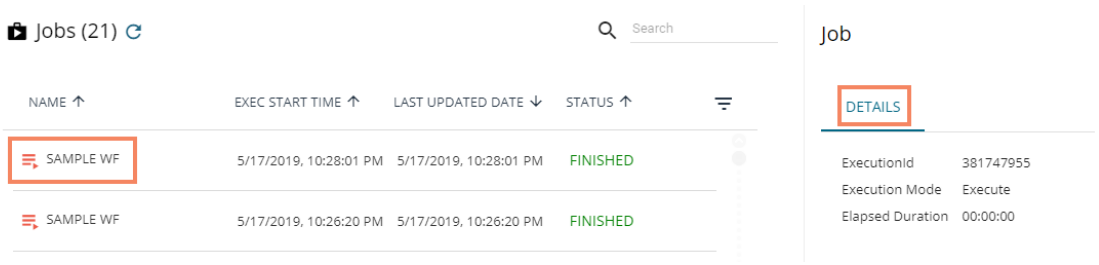


iii) A table appears displaying a list of jobs.



iv) Click on a job from the list.

v) The '**DETAILS**' tab opens, displaying the execution details for the selected job on the right-hand side.



Note: The execution details will be displayed on the right-hand side of the '**Job**' page. Users need to click on the 'STATUS' of a job using the list of jobs.

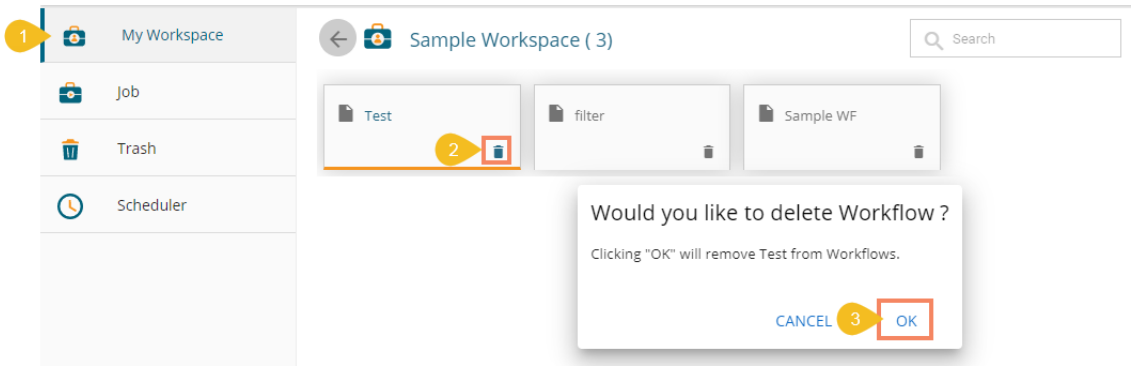
4.10. Trash

The '**Trash**' folder is provided to store all the deleted workflows and workspaces. Users can restore the deleted workflows and workspaces using this folder.

i) Navigate to 'My Workspace' page.

ii) Select a workflow and click the '**Delete**' icon.

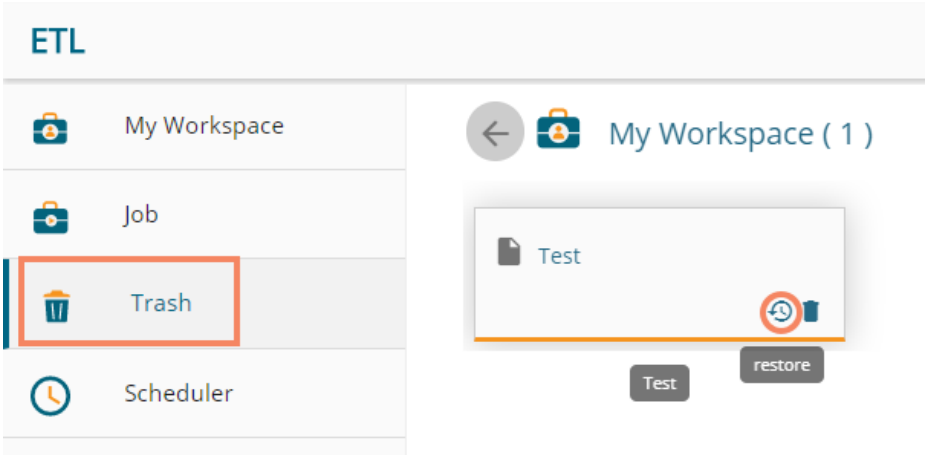
iii) A pop-up window opens to assure about the action of deletion. Click the '**OK**' option from the pop-up window.



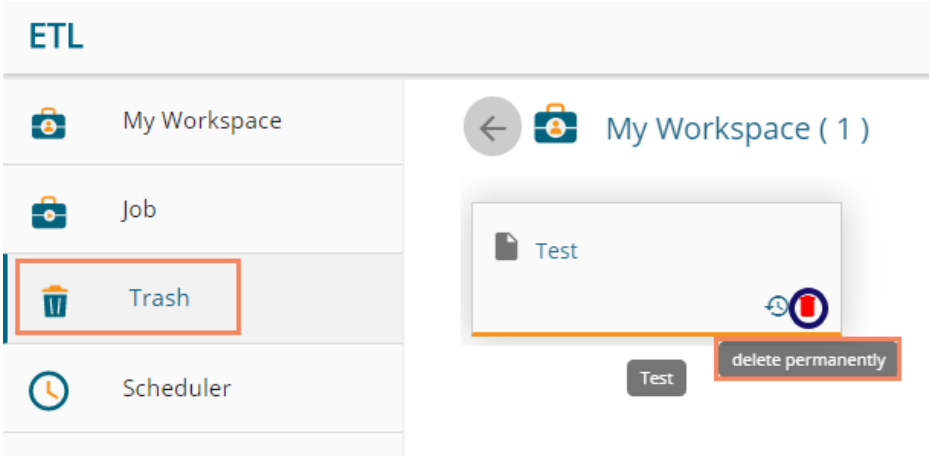
iv) The Workflow gets deleted and moved to the Trash folder.

v) Click on the '**Trash**' option.
vi) The Trash folder displays the deleted workflows.
vii) The '**Details**' tab displays information about the selected workflow.

viii) Click '**Restore**' to restore the selected workflow/workspace.



ix) Click '**delete permanently**' icon to permanently delete the selected workflow/workspace.



Note:

a. Users can check out all the essential features of the ETL module on a relevant input dataset.
b. Other options provided on the workflow editor are as described below:

Icons	Name	Description
or	Hide and Show Components	Hides or shows the components on the left-hand side.

	Clear Workflow	Clears the current workflow from the workflow editor.
	Navigator	Redirects the users to the following hyperlinks: 1. Workspace 2. Job 3. Trash 4. Scheduler

5. Transform

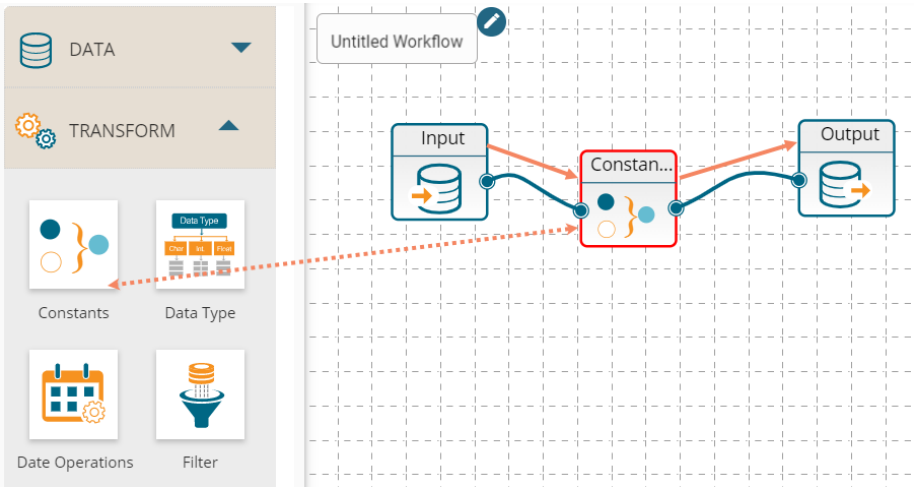
The Transformation components have both input ports and output ports when used in a workflow, and they apply specific transformations to the input data in the transformation stream.

5.1. Constants

Users can give a corresponding valid constant value for each type of column.

i) Navigate to the Workflow editor.

ii) Connect the '**Constants**' component to the configured input dataset. Connect the output node of the dragged Constants component with the Output component and create a workflow.



iii) Click on the '**Constants**' component to open the '**CONFIGURATION**' fields.

iv) Configure the required details for the '**Constants**' component:

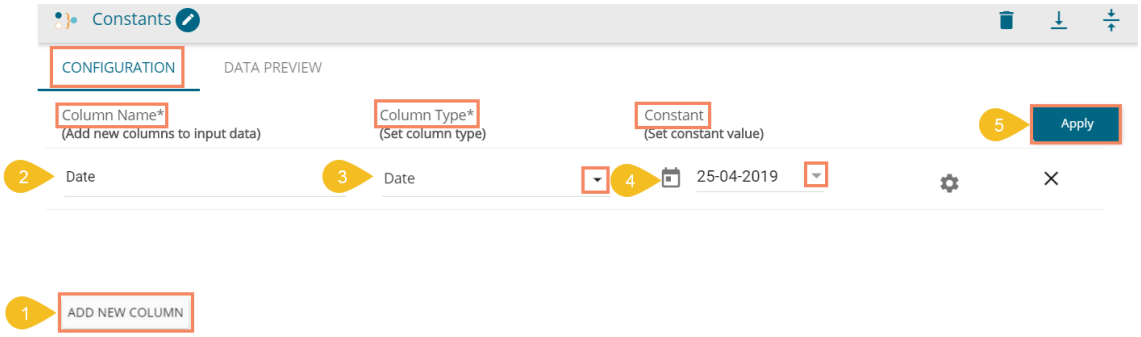
a. Click the '**ADD NEW COLUMN**' option

b. Column Name: Add new columns to the input dataset

c. Column Type: Set column type from the drop-down menu

d. Constant: Set constant value using the drop-down calendar.

e. Click the '**Apply**' option.

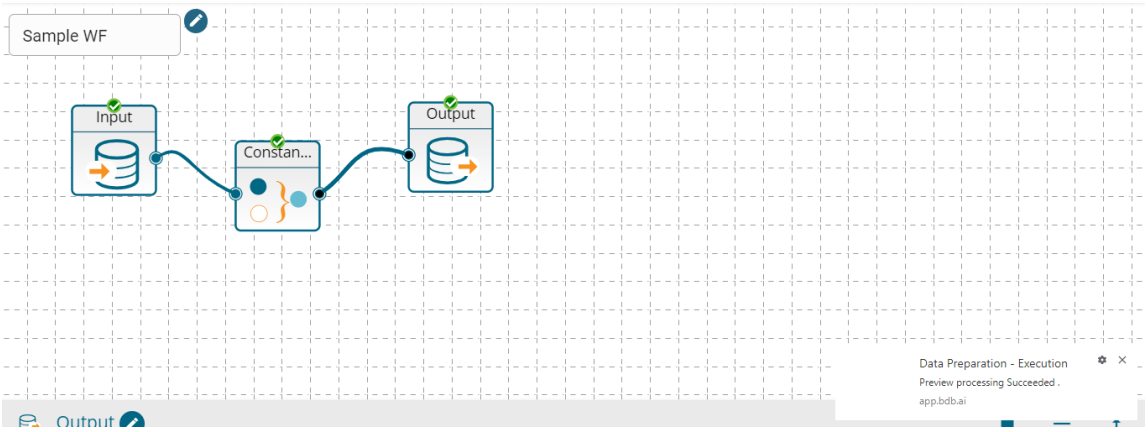


v) A message appears stating that the Constant fields are successfully configured.

vi) Save the workflow.

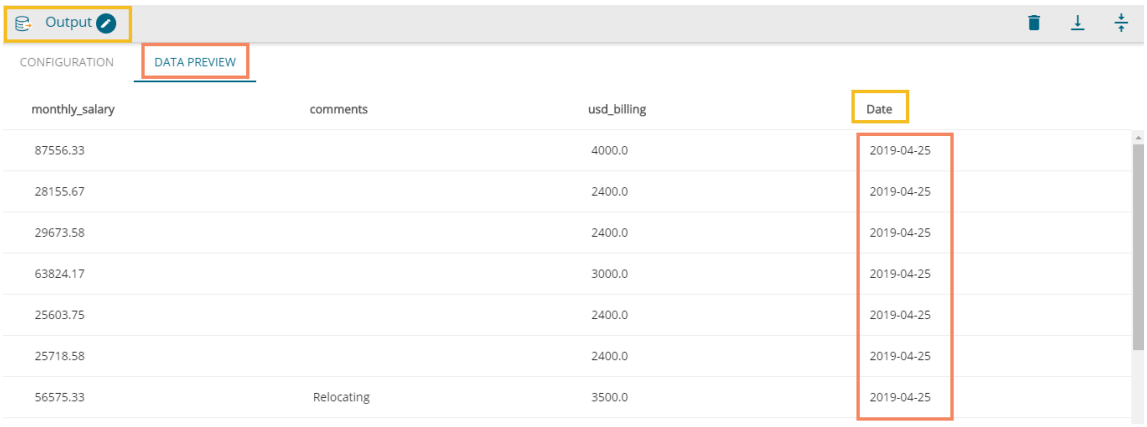
vii) Run/Execute the workflow.

viii) A notification appears at the bottom of the screen to inform about the process execution status. If the process gets completed successfully, then the components in the selected workflow will display a green checkmark.



Note: The user can also use the 'DATA PREVIEW' tab given in the '**Constants**' component.

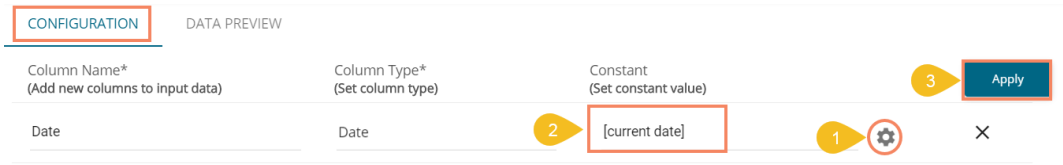
ix) The set constant value gets applied to the selected column in the output dataset.



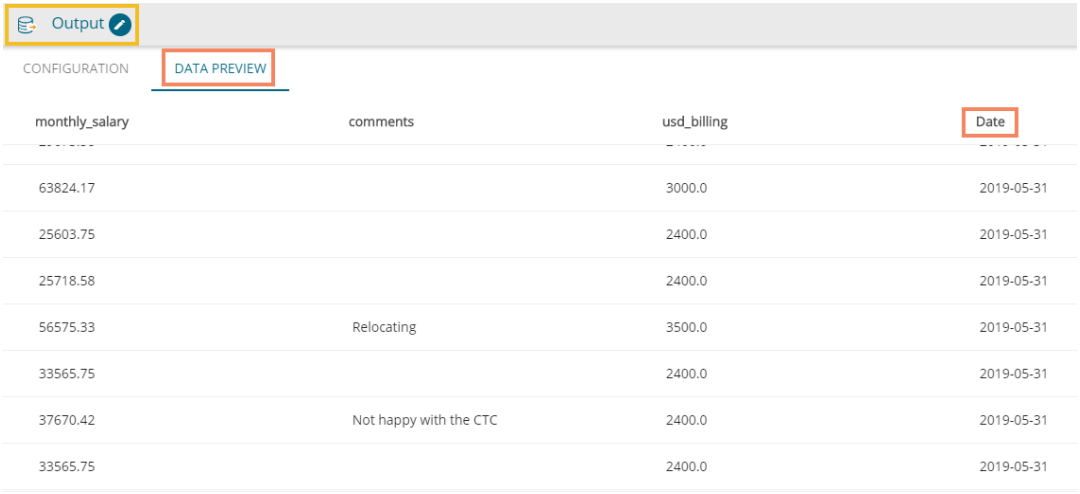
Note:

a. Click the '**Remove**' ✕ icon to remove the added constant information.

b. The Constants component can take current date by clicking the icon.
   i. Click the icon.
   ii. The '**current date**' option appears for the constant value field.
   iii. Click the '**Apply**' option.



   iv. Run the workflow and open the data preview.
   v. The set data gets added to the column created using the '**Constants**' component.

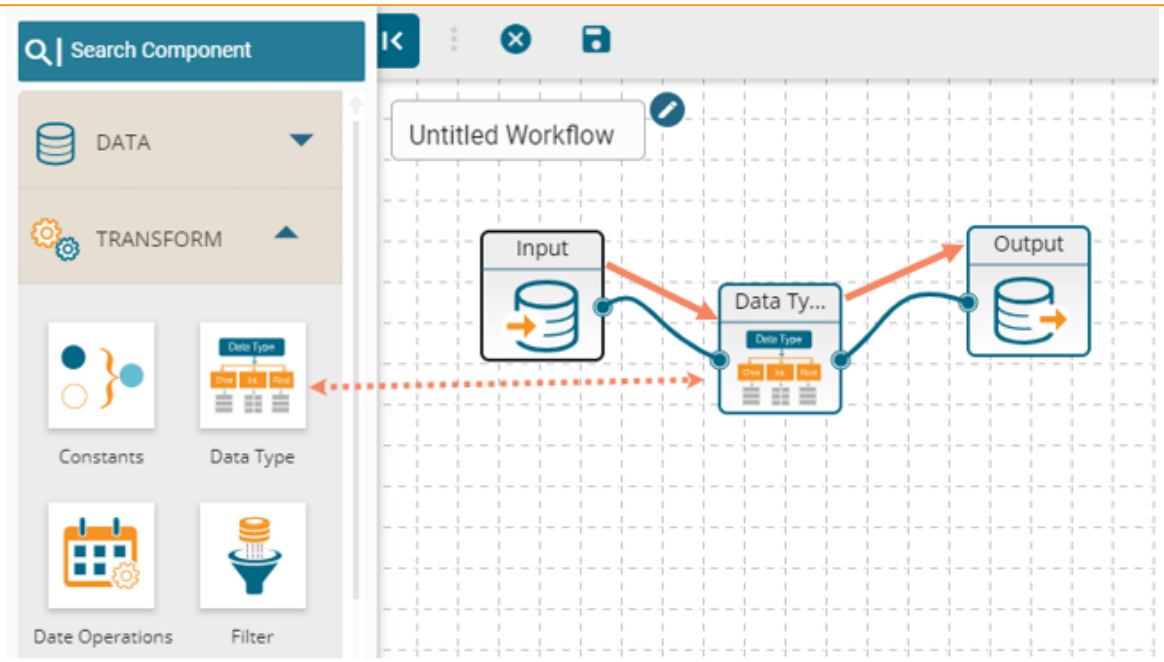


monthly_salary	comments	usd_billing	Date
63824.17		3000.0	2019-05-31
25603.75		2400.0	2019-05-31
25718.58		2400.0	2019-05-31
56575.33	Relocating	3500.0	2019-05-31
33565.75		2400.0	2019-05-31
37670.42	Not happy with the CTC	2400.0	2019-05-31
33565.75		2400.0	2019-05-31

5.2. Data Type

Users can change the data type of the selected columns by using the '**Date Type**' transform.

i) Navigate to the Workflow editor.
ii) Connect the '**Data Type**' component to the configured input dataset and output component.



iii) Click on the Data Type component to open the CONFIGURATION tab.

iv) Click the '**ADD COLUMN**' option to add a new column.

v) Select the columns and change the column data type using the drop-down menu.

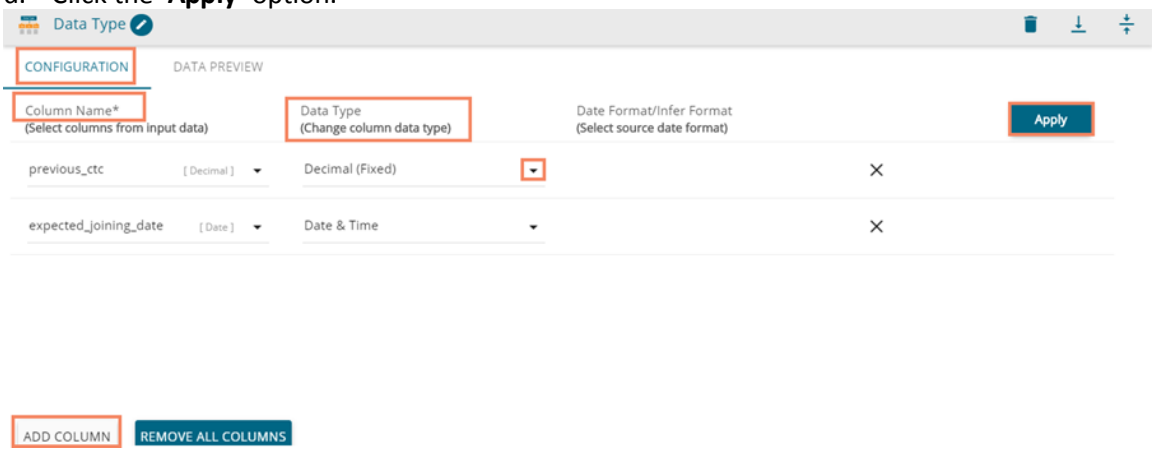
a. Column Name: Select columns from input data

b. Data Type: Change column data type

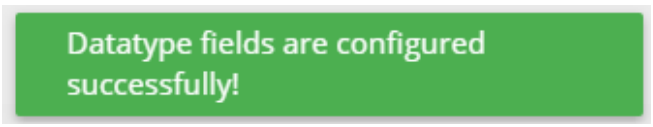
c. Date Format: Select source date format

E.g. In the following instance, the column data type for the selected columns has been changed from '**Decimal**' to '**Decimal (Fixed)**' and '**Date**' to '**Date & Time**.'

d. Click the '**Apply**' option.



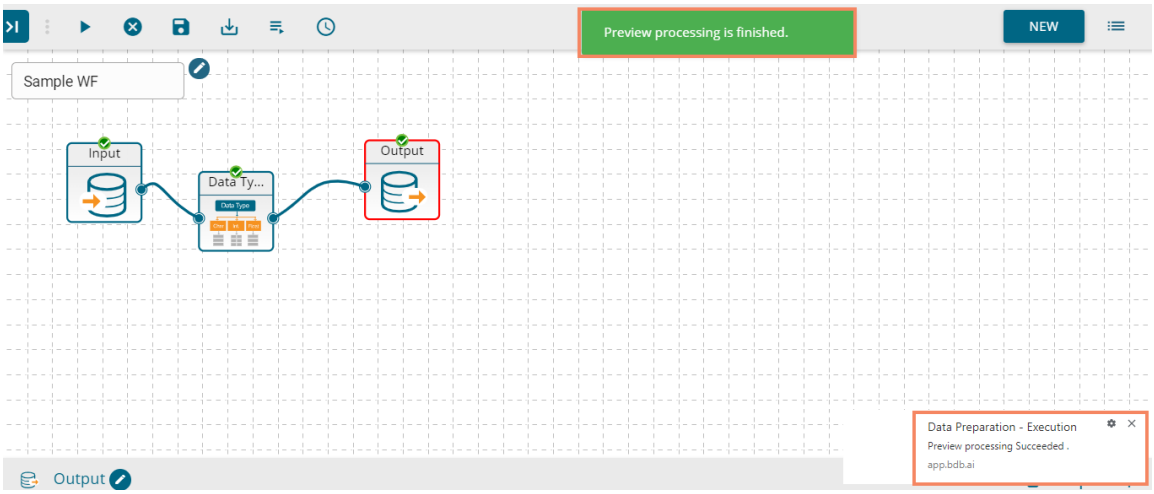
vi) A success message appears to confirm the Data Type field configuration.



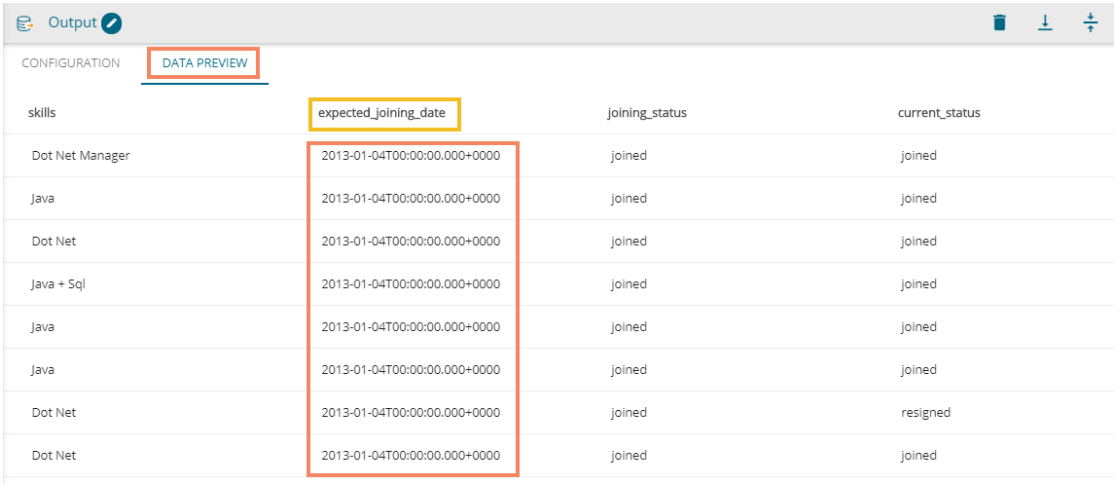
vii) Save the workflow.

viii) Run/Execute the workflow.

ix) The success notification will appear, and the components in the workflow will get green checkmarks at the top.

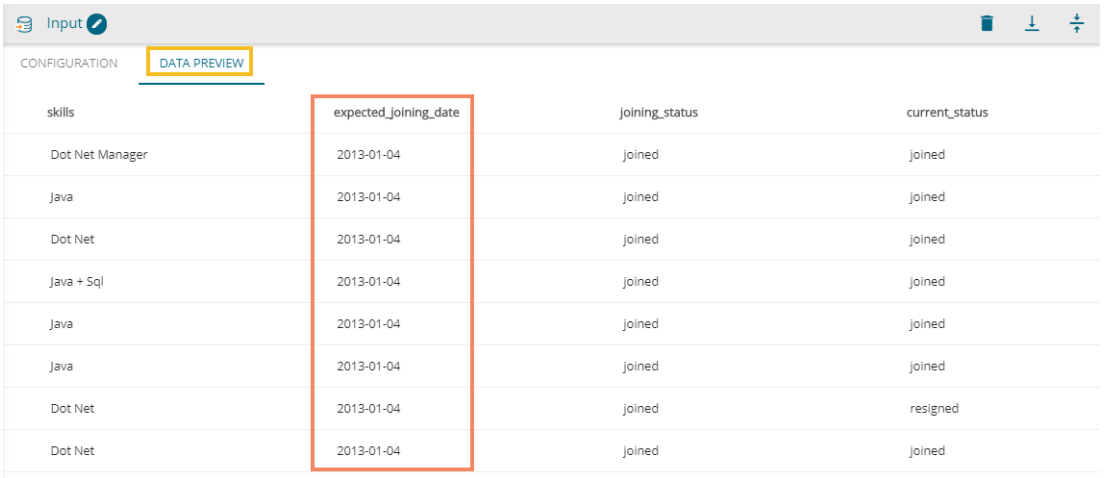


x) Click the '**DATA PREVIEW**' tab for the Output component to see the transform result.



skills	expected_joining_date	joining_status	current_status
Dot Net Manager	2013-01-04T00:00:00.000+0000	joined	joined
Java	2013-01-04T00:00:00.000+0000	joined	joined
Dot Net	2013-01-04T00:00:00.000+0000	joined	joined
Java + Sql	2013-01-04T00:00:00.000+0000	joined	joined
Java	2013-01-04T00:00:00.000+0000	joined	joined
Java	2013-01-04T00:00:00.000+0000	joined	joined
Dot Net	2013-01-04T00:00:00.000+0000	joined	resigned
Dot Net	2013-01-04T00:00:00.000+0000	joined	joined

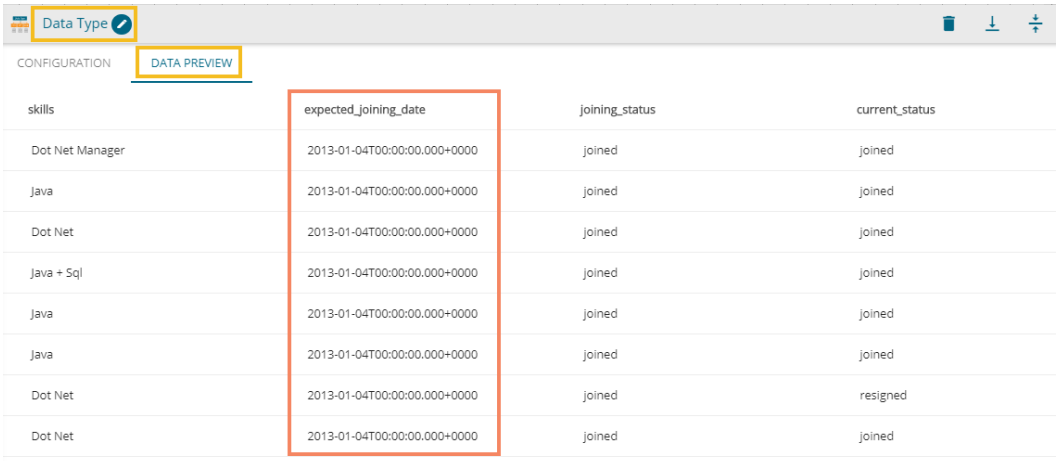
xi) Users can compare the data previews of the Input and Data Type modules (E.g., the selected input, in this case, contains the following column types)



skills	expected_joining_date	joining_status	current_status
Dot Net Manager	2013-01-04	joined	joined
Java	2013-01-04	joined	joined
Dot Net	2013-01-04	joined	joined
Java + Sql	2013-01-04	joined	joined
Java	2013-01-04	joined	joined
Java	2013-01-04	joined	joined
Dot Net	2013-01-04	joined	resigned
Dot Net	2013-01-04	joined	joined

Note:

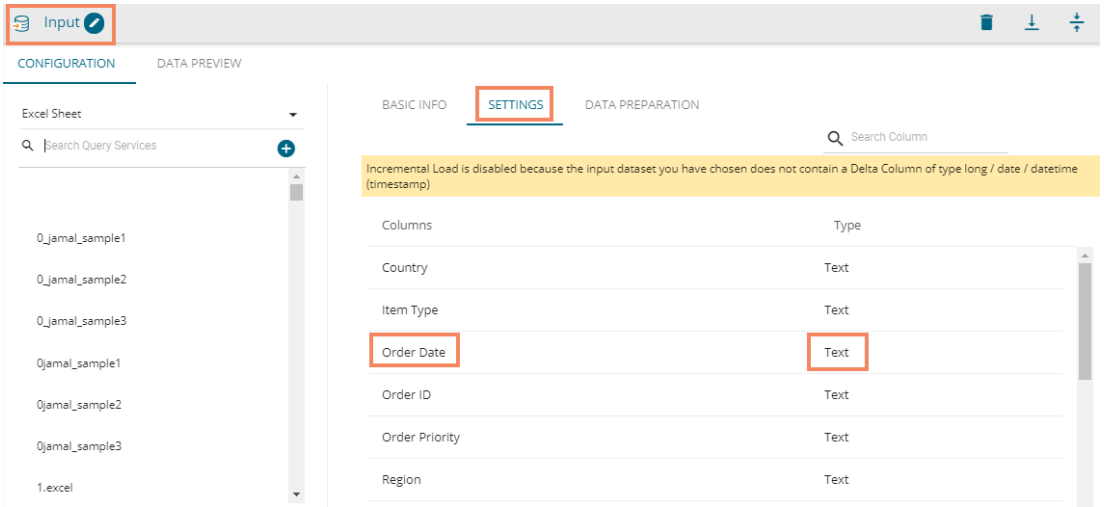
a. Users can get the same Data Preview as Output dataset while opening the '**DATA PREVIEW**' tab from any selected transform component. E.g., The '**DATA PREVIEW**' tab for the '**Data Type**' Transform component is as displayed below.



5.2.1. Inferring Date & Date Time Formats

The Infer Date/Data Time functionality is provided for users to include various Date/Date Time formats which are not provided by ETL plugin. This functionality works only on the Text type of columns.

i) Select an input data set with some columns in the text data type (preferably the Input data set should have a column displaying Date in the Text data type).

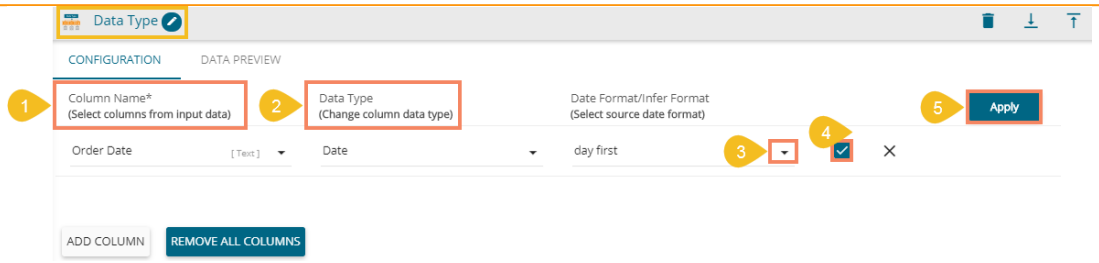


ii) Connect the configured dataset with the Data Type transform component.

iii) Provide the required information to configure the '**Data Type**' component.

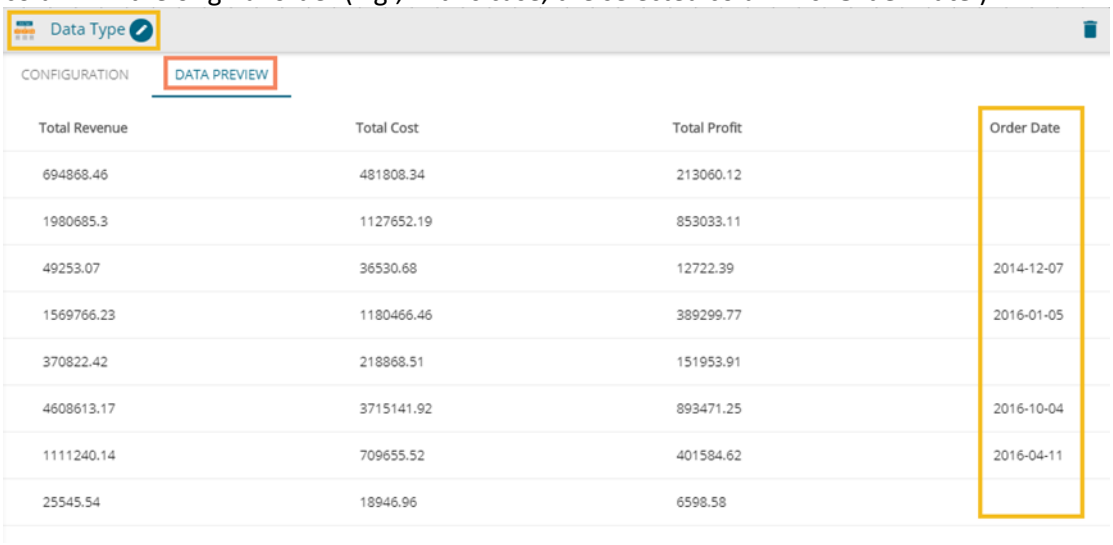
a. Pass the Date column(s) that is in Text data type from the input dataset.
b. Change the column data type using the drop-down icon provided for the '**Data Type**' field.
c. The Date Format/Infer Format will display a drop-down icon to select an option for the source data format.
d. Use a checkmark in the given box to enable the Infer Format.
e. Click the '**Apply**' icon.

iv) Run and execute the workflow.



v) Open the '**DATA PREVIEW'** tab of the Data Type component.

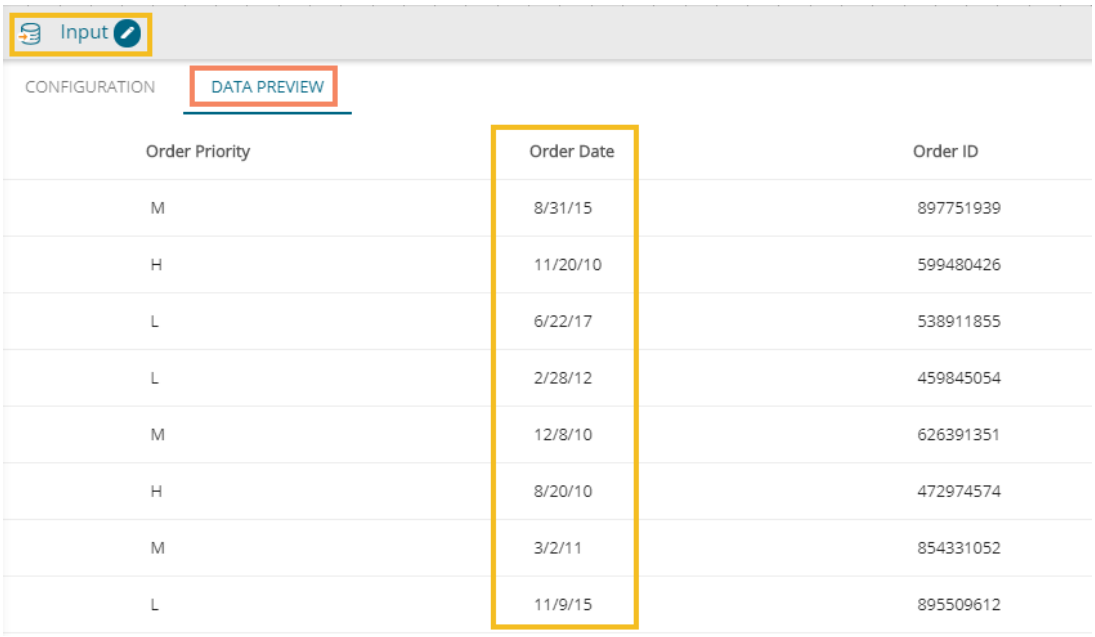
vi) Check the column provided for the Data Infer Format function. It will display the selected date column in the original order (E.g., In this case, the selected column is '**Order Date'**).



Total Revenue	Total Cost	Total Profit	Order Date
694868.46	481808.34	213060.12	
1980685.3	1127652.19	853033.11	
49253.07	36530.68	12722.39	2014-12-07
1569766.23	1180466.46	389299.77	2016-01-05
370822.42	218868.51	151953.91	
4608613.17	3715141.92	893471.25	2016-10-04
1111240.14	709655.52	401584.62	2016-04-11
25545.54	18946.96	6598.58	

vii) Check the selected Date column from the Input '**DATA PREVIEW'** tab.

viii) The data gets displayed in the selected Date Format.



Order Priority	Order Date	Order ID
M	8/31/15	897751939
H	11/20/10	599480426
L	6/22/17	538911855
L	2/28/12	459845054
M	12/8/10	626391351
H	8/20/10	472974574
M	3/2/11	854331052
L	11/9/15	895509612

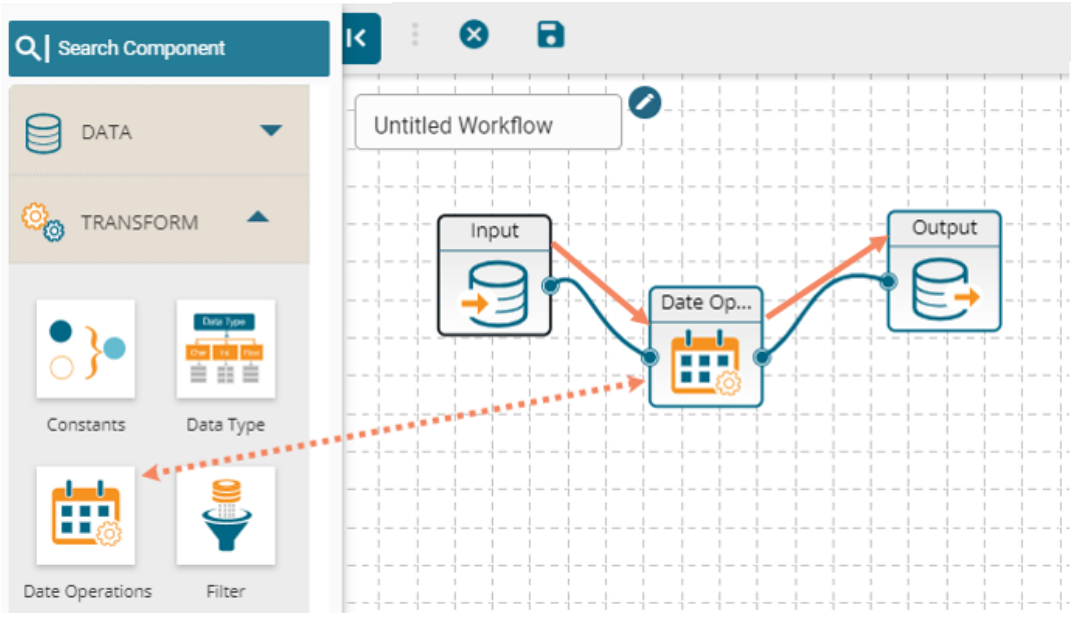
Note:

a. The functionality only works for the '**Text**' type of column.
b. If the source data format does not befit in the selected infer format, then those entries will not be listed in the output.
c. Use the '**ADD COLUMN**' option to add a new column for the Data Type change

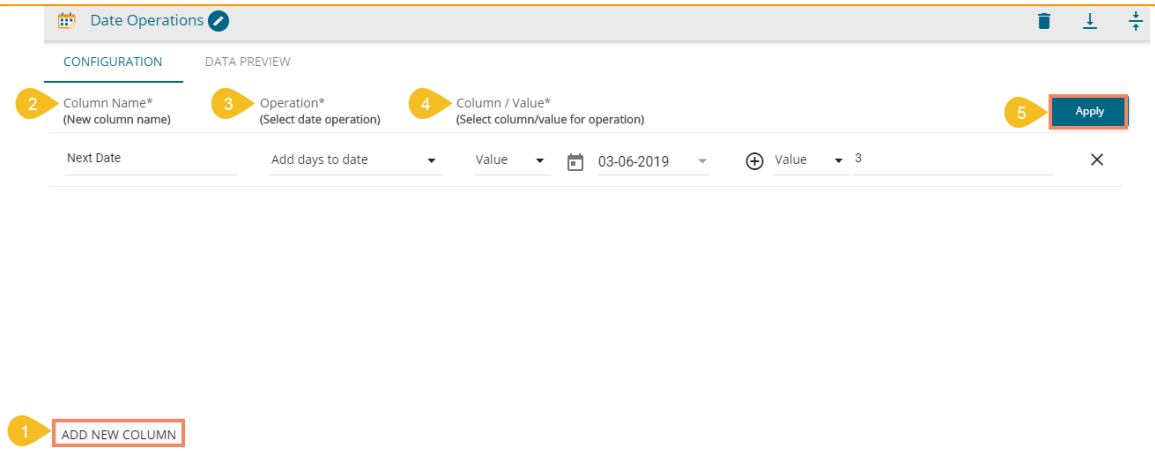
5.3. Date Operations

Users can perform various operations of dates addition/subtraction with integers or other dates. It also allows extraction of parts of times like day-part, month part, etc.

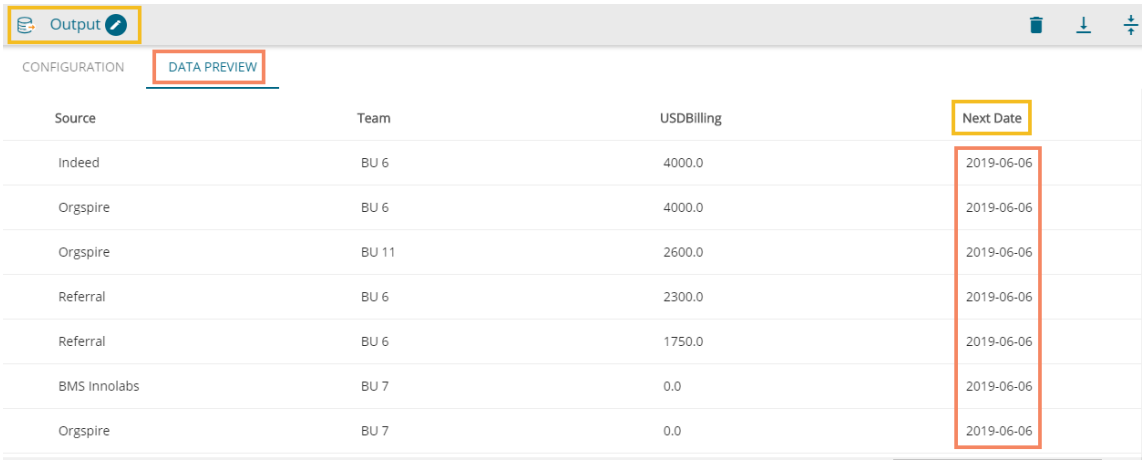
i) Navigate to the Workflow editor.
ii) Connect the '**Date Operations**' component to the configured input dataset and output component.



iii) Configure the '**Date Operations**' component as described below:
a. Click the '**ADD NEW COLUMN**' option.
b. Column Name: Provide a name for the New Column.
c. Operation: Select a data operation option from the drop-down menu.
d. Column/Value: Select a column or value for operations.
   i. By selecting the '**column**' option, a drop-down menu appears.
   ii. By selecting the '**value**' option, users will be redirected to enter a value.
   E.g., In the following image, the selected option is Value for the operation, and three as a constant value will be added to the selected date.
e. Click the '**Apply**' option.



iv) Save the workflow.
v) Run/Execute the workflow.
vi) Open the '**DATA PREVIEW**' tab of Date Operation or Output component.
vii) The newly configured Date Column '**Next Date**' gets added to the data with the set value.

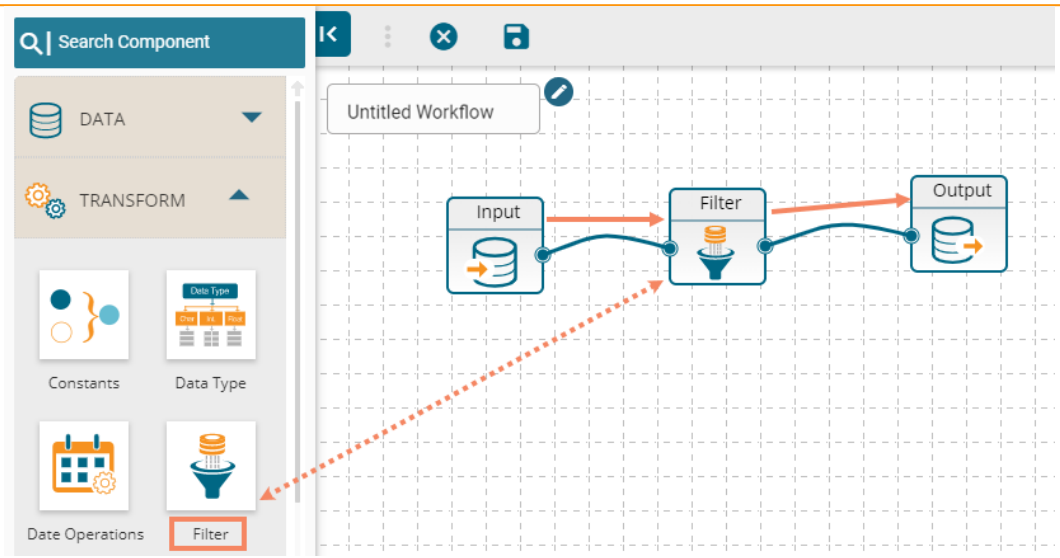


Source	Team	USDBilling	Next Date
Indeed	BU 6	4000.0	2019-06-06
Orgspire	BU 6	4000.0	2019-06-06
Orgspire	BU 11	2600.0	2019-06-06
Referral	BU 6	2300.0	2019-06-06
Referral	BU 6	1750.0	2019-06-06
BMS Innolabs	BU 7	0.0	2019-06-06
Orgspire	BU 7	0.0	2019-06-06

5.4. Filter

Users can filter the input dataset by specifying conditional expressions using the '**Filter**' transform. Multiple filter conditions can be imposed in the same transform. The following table lists the map of data types and permissible filter conditions.

i) Navigate to the Workflow editor.
ii) Drag the '**Filter**' component.
iii) Connect the '**Filter**' component to the configured input dataset and output component.

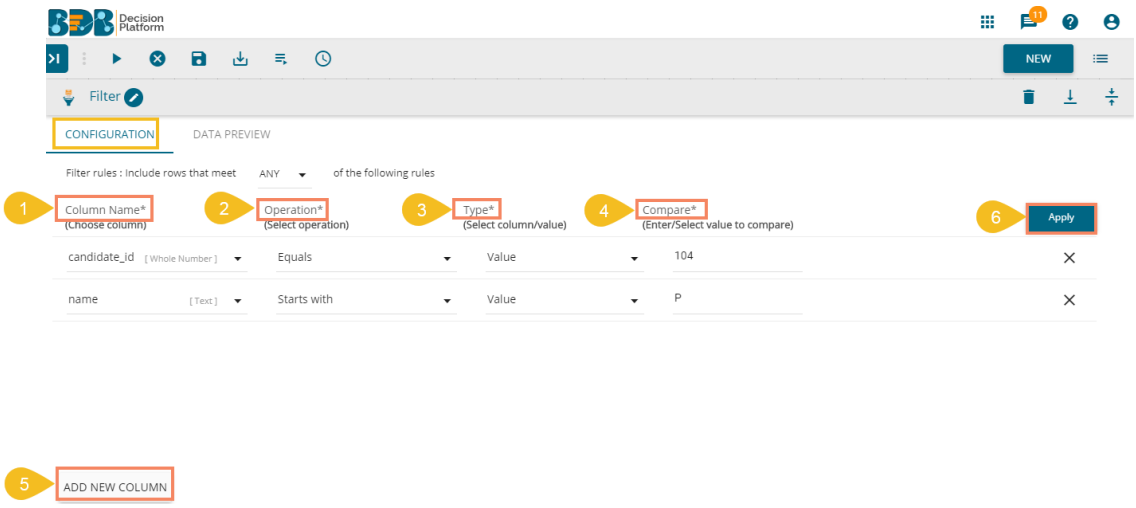


iv) Configure the '**Filter**' Component as described below:

a. Select a filter rule from the drop-down
   i. **ALL**: By selecting this option filter gets applied only if all the added conditions are true
   ii. **ANY**: By choosing this option filter gets applied even if any one condition is true

   E.g., in this case, the selected Filter Rule is 'ANY' so the data preview displays data satisfying either of the filter conditions.

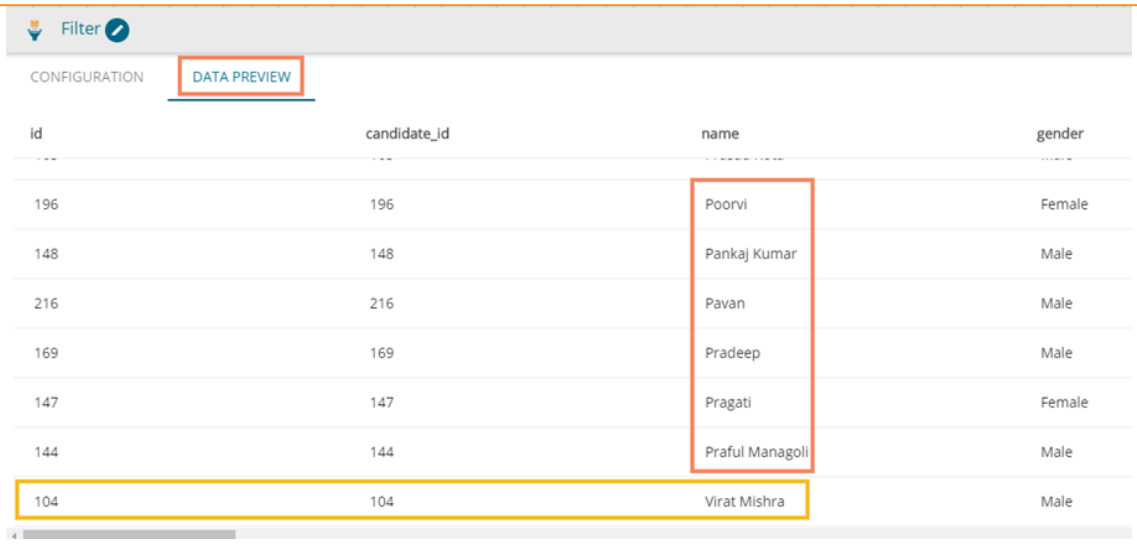
b. Column Name: Choose a column from the drop-down menu
c. Operation: Select an operation from the drop-down menu
d. Type: Select one option out of '**Column**' or '**Value.**'
e. Compare: Enter a value/Select a column from the list to compare with
f. Click the '**ADD NEW COLUMN**' option to insert fields for the new column
g. Click the '**Apply**' option.



v) Save the workflow.

vi) Run the workflow.

vii) The input data will be filtered as per the applied conditions.

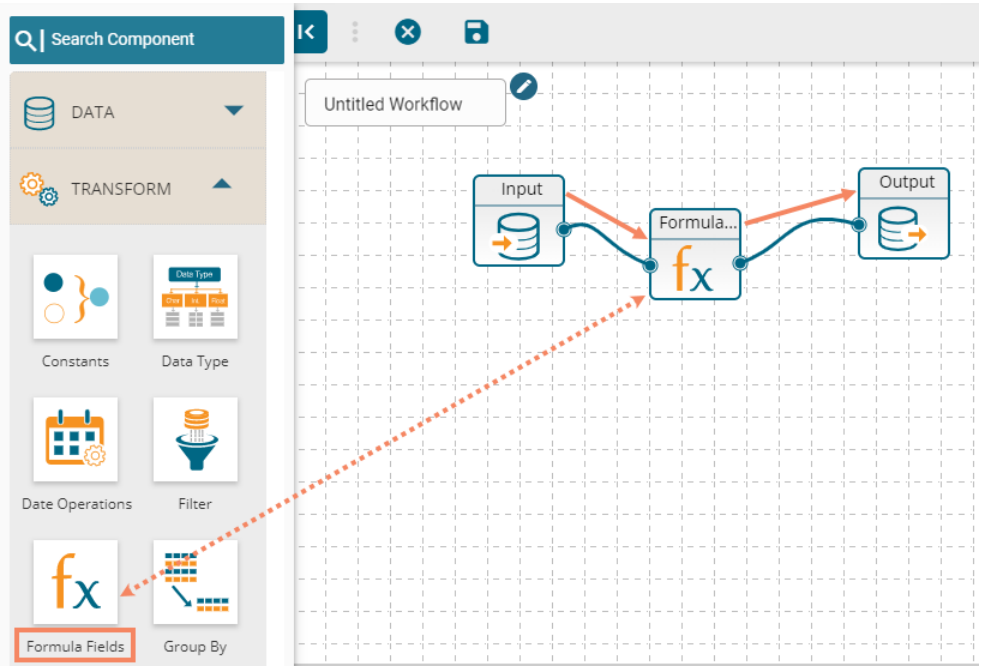


5.5. Formula Fields

Users can perform most common arithmetic operations (add, subtract, multiply and divide) on constants and columns.

i) Navigate to the Workflow editor.

ii) Connect the '**Formula Fields**' to the configured input dataset and output component.

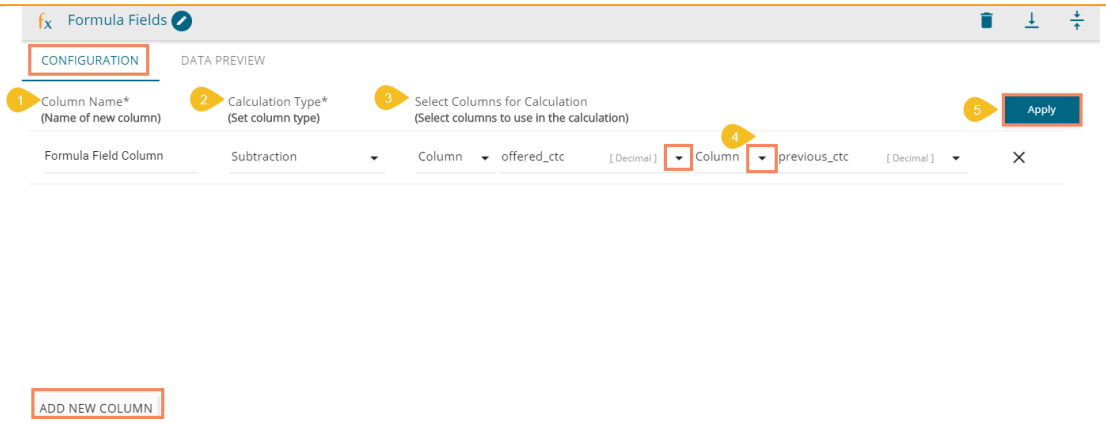


iii) Configure the '**Formula Fields**' component as described below:

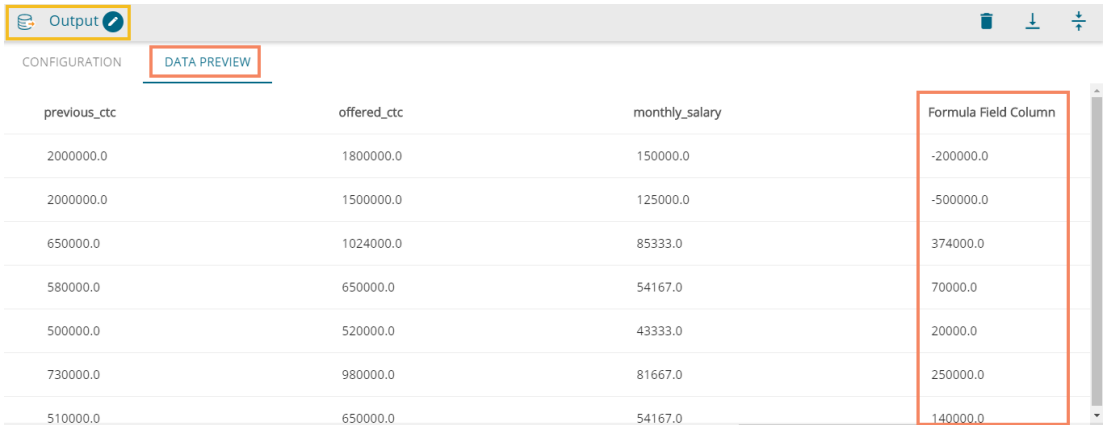
a. Column Name: Enter a name for the formula column

b. Calculation Type: Select a calculation type using the drop-down menu

c. Select Columns for Calculation: Select columns to be used in the calculation. Users can choose either a column or enter a value to complete the calculation process.
E.g. The 'Value' option is chosen in the below given example.



iv) Save the workflow.
v) Run the workflow.
vi) The calculated column will be added in the output dataset.

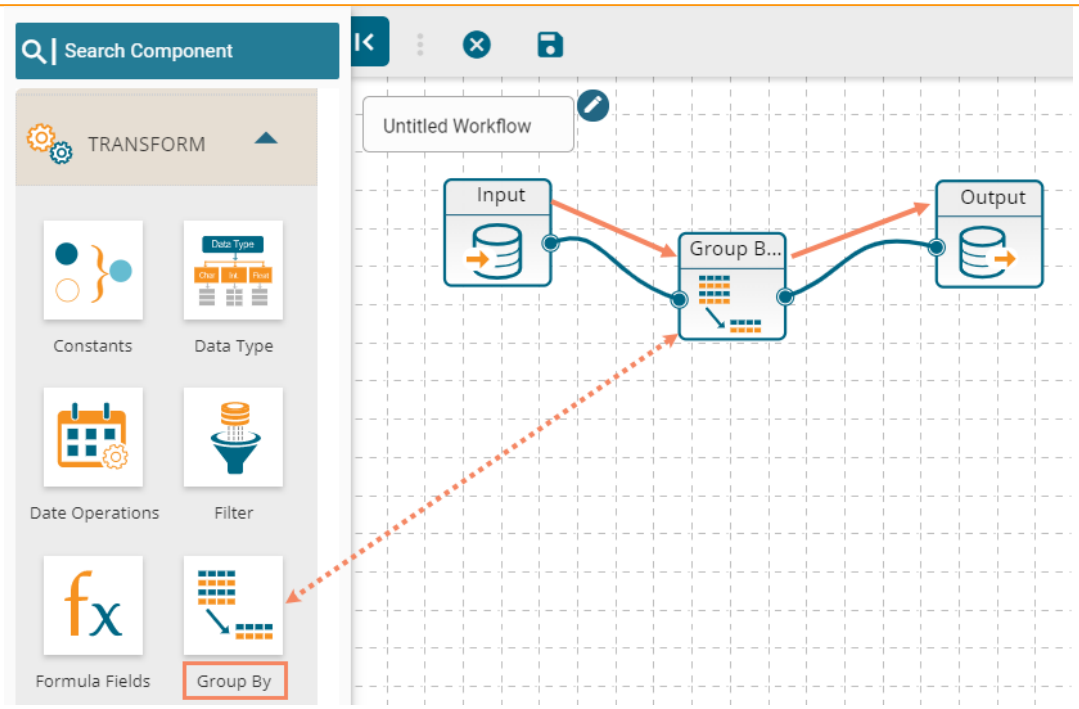


previous_ctc	offered_ctc	monthly_salary	Formula Field Column
2000000.0	1800000.0	150000.0	-200000.0
2000000.0	1500000.0	125000.0	-500000.0
650000.0	1024000.0	85333.0	374000.0
580000.0	650000.0	54167.0	70000.0
500000.0	520000.0	43333.0	20000.0
730000.0	980000.0	81667.0	250000.0
510000.0	650000.0	54167.0	140000.0

5.6. Group By

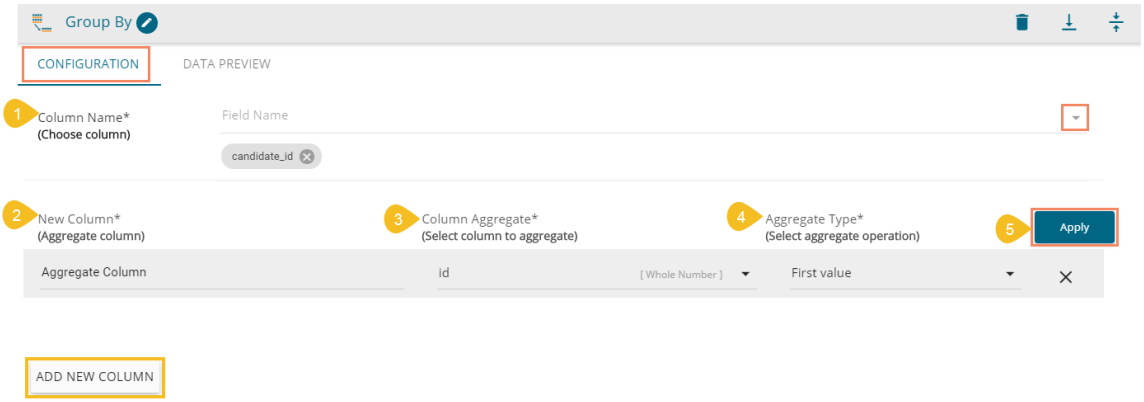
The '**Group By**' feature allows multiple aggregations on the same or different columns. Users can obtain numerous aggregations in the same transform. The aggregated values are added to a new column.

i) Navigate to the Workflow editor.
ii) Connect the '**Group By**' component to the configured input dataset and output component.

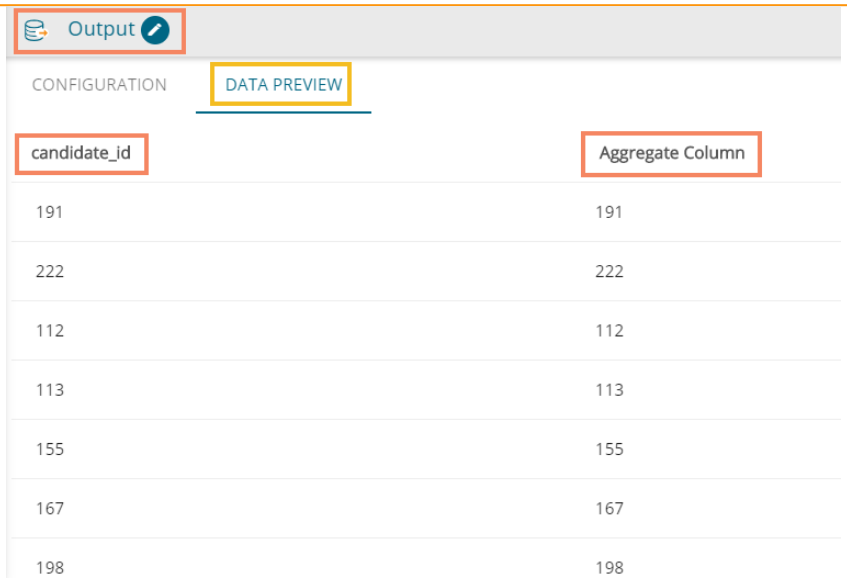


iii) Configure the '**Group By**' component as described below:

a. Column Name: Select a column from the drop-down menu
b. New Column: Enter a title for the aggregate column
c. Column Aggregate: Select a column from the drop-down menu to apply aggregation
d. Aggregate Type: Select an aggregation operation from the drop-down menu



iv) Save the workflow
v) Run the workflow
vi) The aggregated column will be displayed in the output data preview



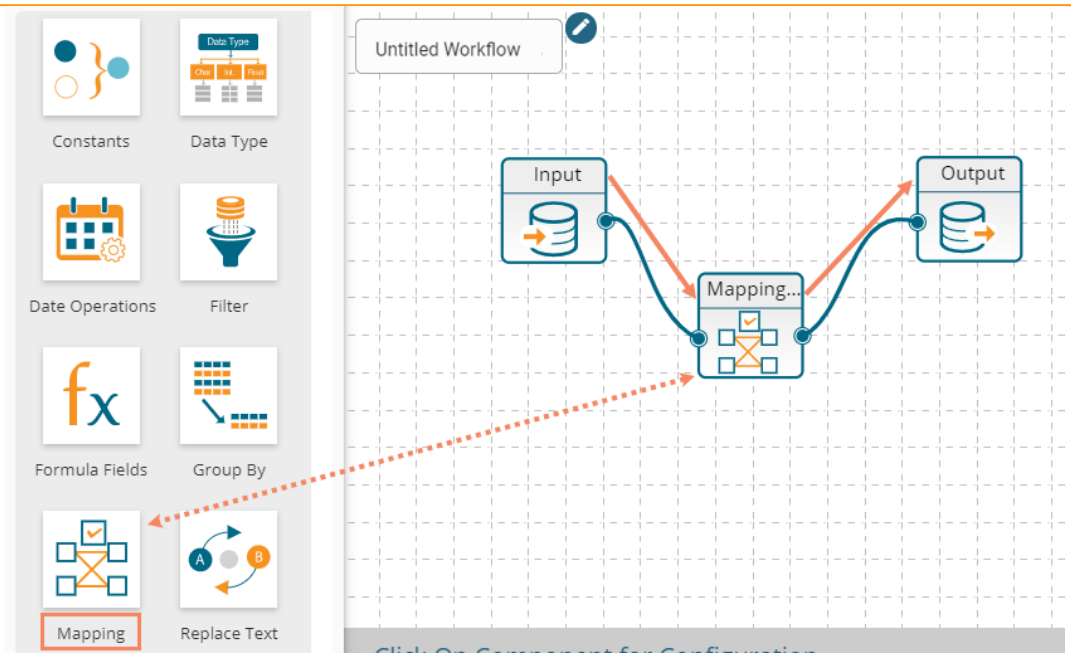
Note: The supported data types and aggregate operations are displayed in the following table:

Data Type	Aggregate
Text Date Date Time	Count First Value Last Value
Whole Number Decimal Decimal (Fixed)	Sum Average Minimum Maximum Standard Deviation Count First Value Last Value

5.7. Mapping

Users should be able to select, remove or rename columns in the input dataset to fit the structure of the sink.

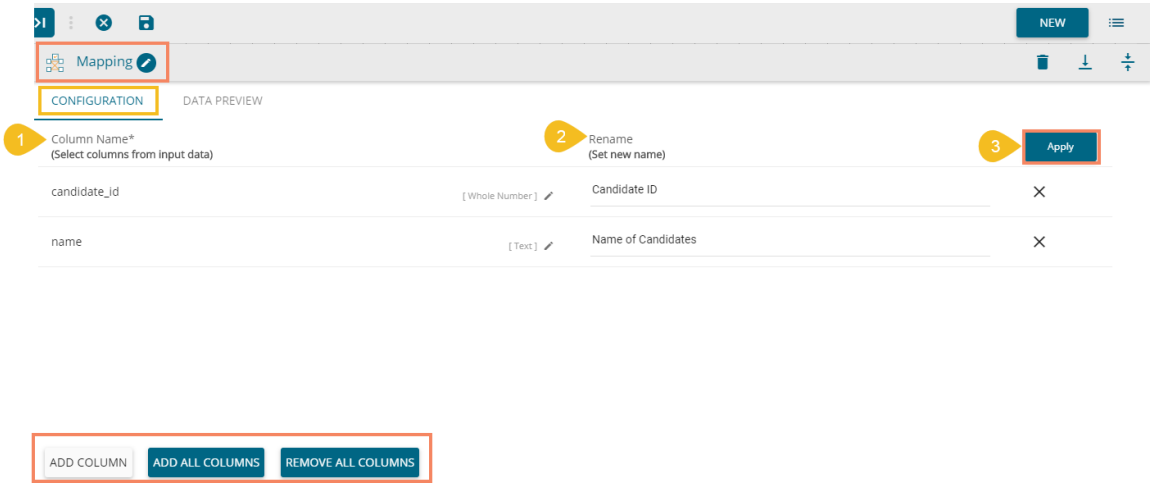
i) Navigate to the Workflow editor.
ii) Connect the '**Mapping**' component to the configured input dataset and output component.



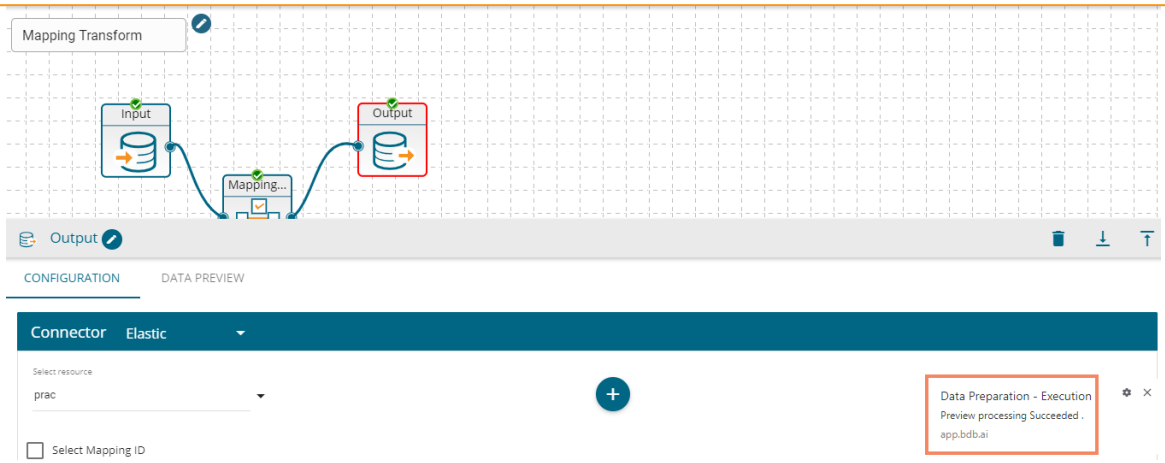
iii) Configure the '**Mapping**' component:

a. Column Name: Select a Column from the input data using the drop-down menu
b. Rename: Rename the selected column of the input data
c. Choose either of the options from the below given choices:
   i. ADD COLUMN: Click this option to add one more column from the input dataset
   ii. ADD ALL COLUMNS: Click this option to map all the columns from the input dataset
d. Click the '**Apply**' option.

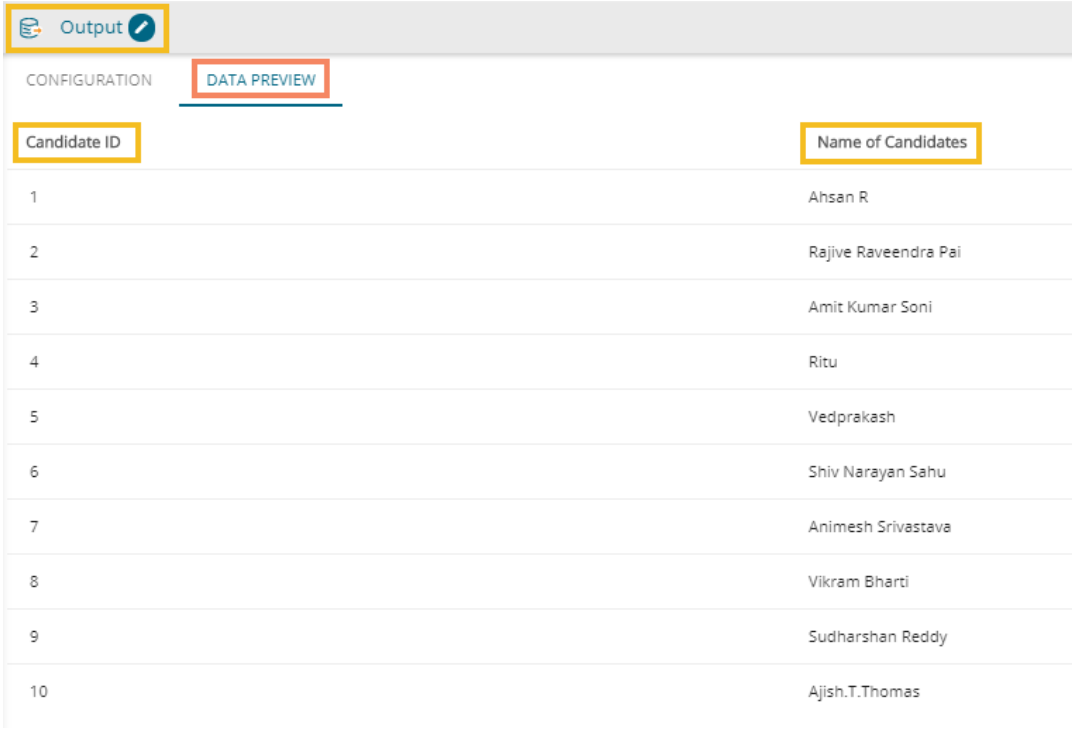
Note: The '**REMOVE ALL COLUMNS**' gets enabled after clicking on the '**ADD ALL COLUMNS'** option. Click this option to remove all the added columns for mapping.



iv) A success message appears to confirm the Mapping Component configuration.
v) Save the workflow.
vi) Run the workflow.
vii) The notification messages appear to display the update of the data preview process.
viii) A notification appears to confirm the successful completion of the Data Preview process. All the components in the workflow get green checkmark at the top.



ix) The aggregated column gets displayed in the 'DATA PREVIEW' tab of the Mapping or Output component.

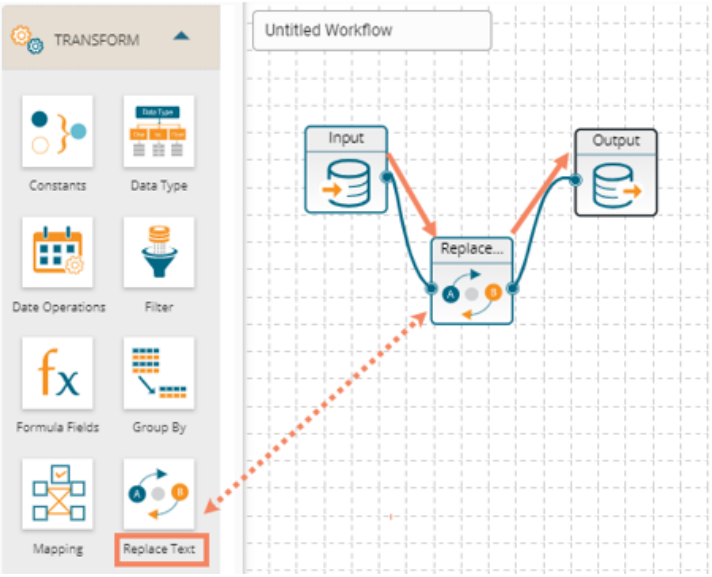


Candidate ID	Name of Candidates
1	Ahsan R
2	Rajive Raveendra Pai
3	Amit Kumar Soni
4	Ritu
5	Vedprakash
6	Shiv Narayan Sahu
7	Animesh Srivastava
8	Vikram Bharti
9	Sudharshan Reddy
10	Ajish.T.Thomas

5.8. Replace Text

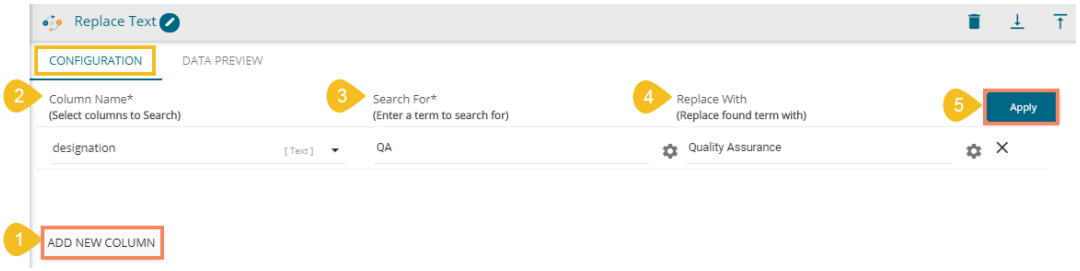
Users can search by whole word, sensitive to case, search for particular values like NULL or empty strings, or use regular expressions, and then replace with any given constant values or even empty strings. Only text columns can be transformed using this component. Users can replace text for the multiple text columns.

i) Navigate to the Workflow editor.
ii) Connect the '**Replace Text**' component with the configured Input dataset and Output component.

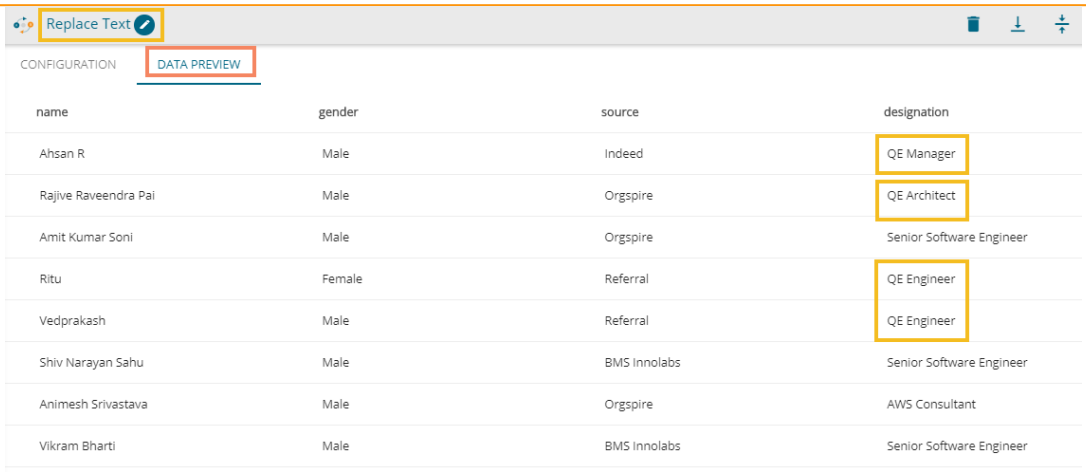


iii) Configure the '**Replace Text**' component as described below:

a. Click the '**ADD NEW COLUMN**' option to get the configuration fields.
b. Column Name: Select a column from the input data set.
c. Search For: Enter a term to search from the selected column.
d. Replace With: Enter a term to replace the searched term in the input data.
e. Click the '**Apply**' option.



iv) Run the workflow.
v) Save the workflow.
vi) The notification messages appear to display the update of the data preview process.
vii) A notification appears to confirm the successful completion of the Data Preview process. All the components in the workflow get green checkmark at the top.
viii) Open the '**DATA PREVIEW**' tab from Replace Text or Output to see the replacement of the selected text in the column.



Note:

a. The users can click on the '**ADD NEW COLUMN**' option to add multiple columns for any transform component.
b. The users can also see data preview of the various transform components by clicking the '**DATA PREVIEW**' tab of the respective options.

6. Merge

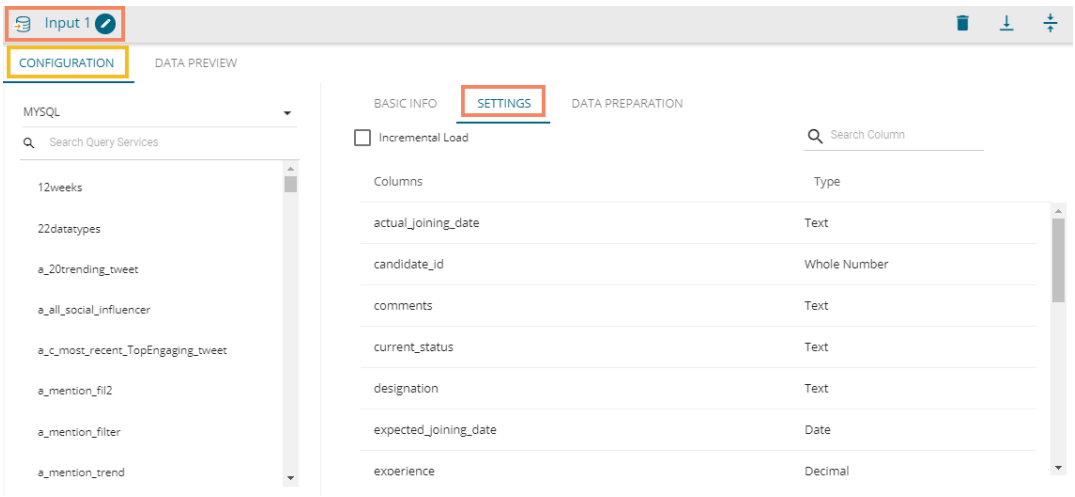
The users can use the '**Merge**' components to combine input data sets and get the required output.

6.1. Append

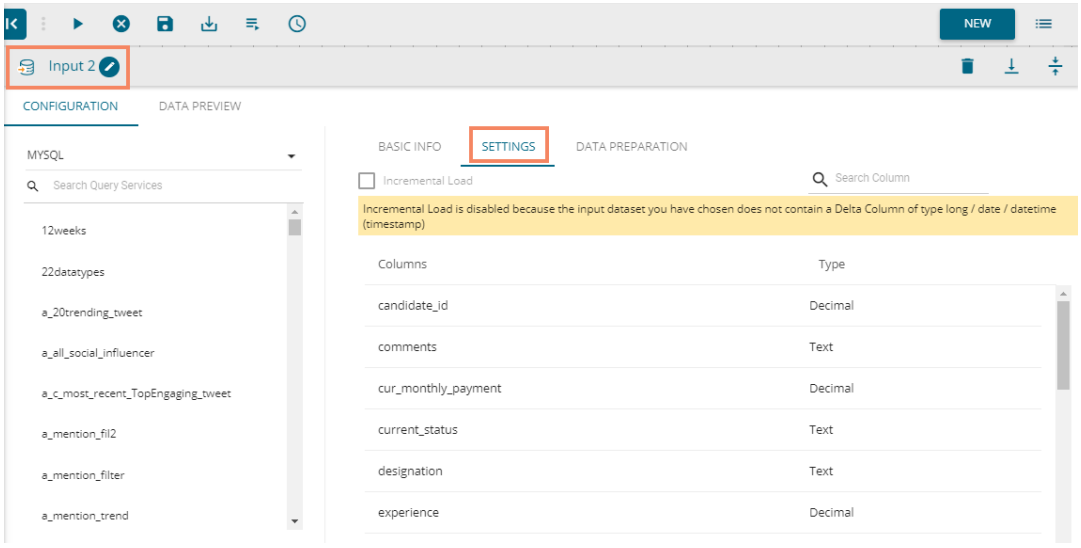
The '**Append**' feature combines one dataset on top of another. If the datasets are of different structures, still the union is possible, and the output will be a unified more massive structure with NULL values populated wherever data is missing. Users can choose whether to include only shared columns or all columns to append.

6.1.1. Append All Columns

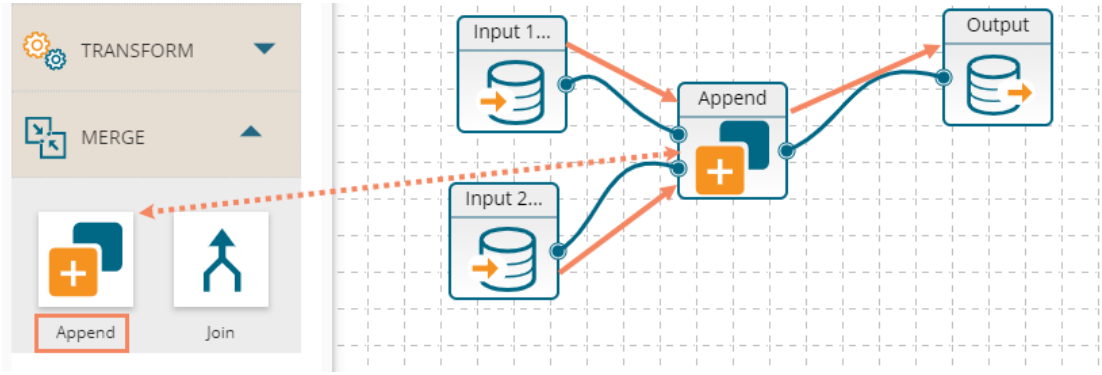
i) Navigate to the Workflow editor.
ii) Configure two input datasets.
iii) Open the '**SETTINGS**' tab to see the available columns in the respective input dataset.
   a. Input 1



b. Input 2

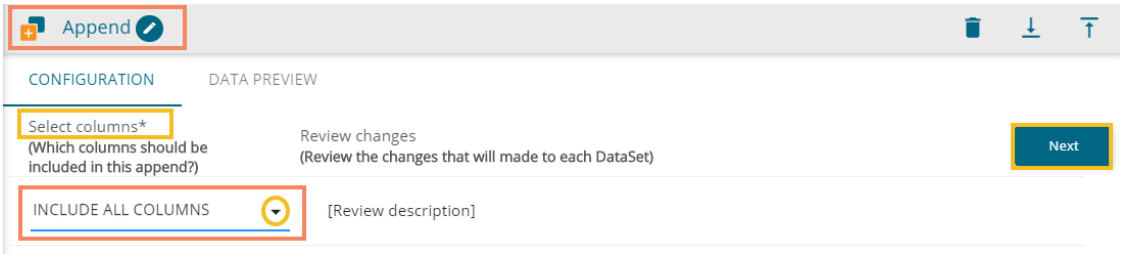


iv) Connect the '**Append**' component with the configured Input datasets and an Output component.



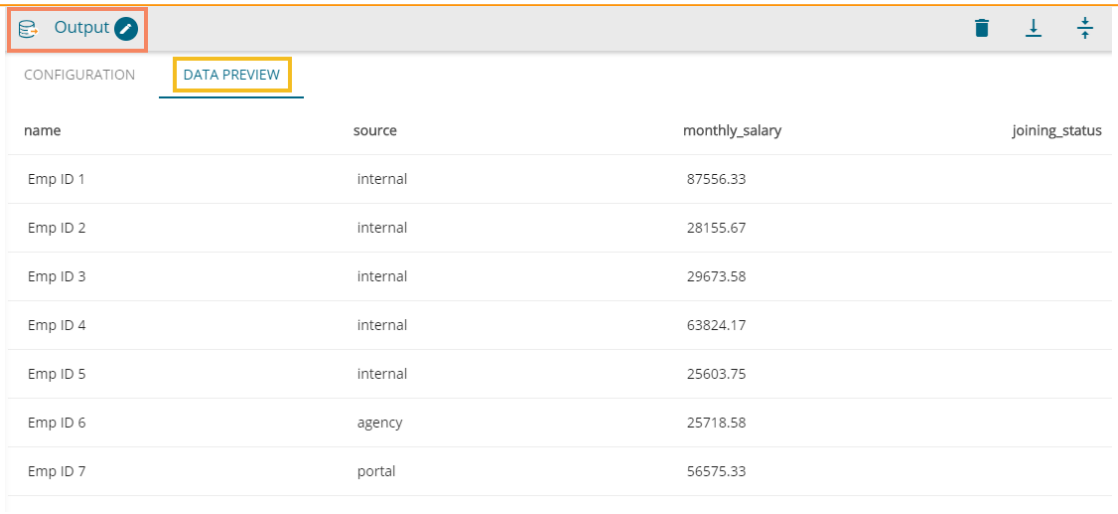
v) Click on the dragged Append component to get the configuration fields.

vi) Select '**INCLUDE ALL COLUMNS**' option using the '**Select Columns**' drop-down menu.



vii) Click the '**Next**' option.

viii) The columns from both the selected input data sets get displayed.

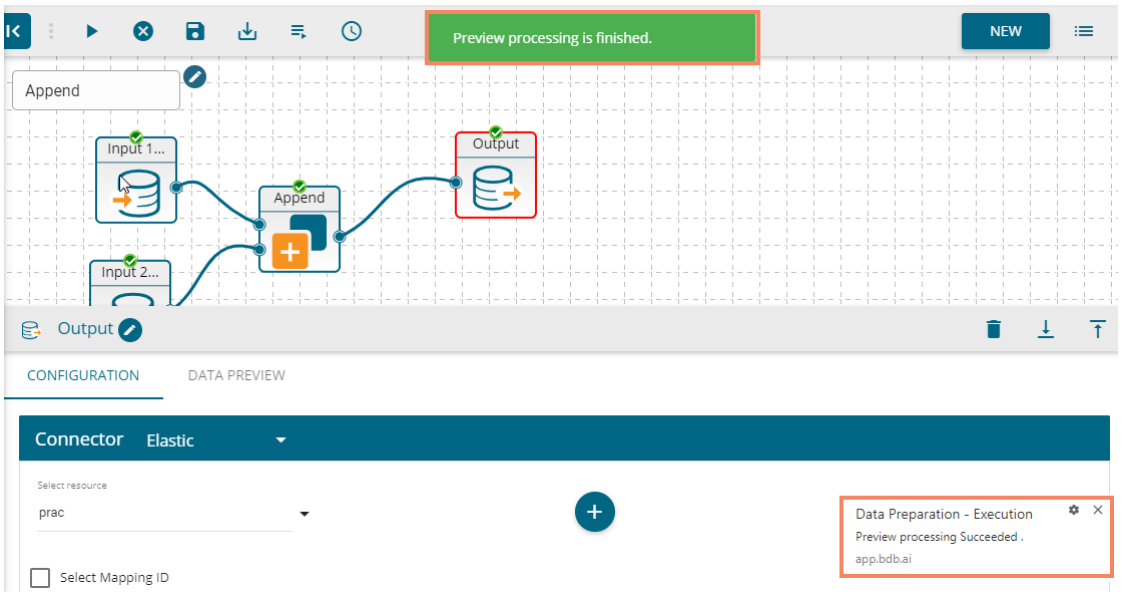


ix) Save the workflow.

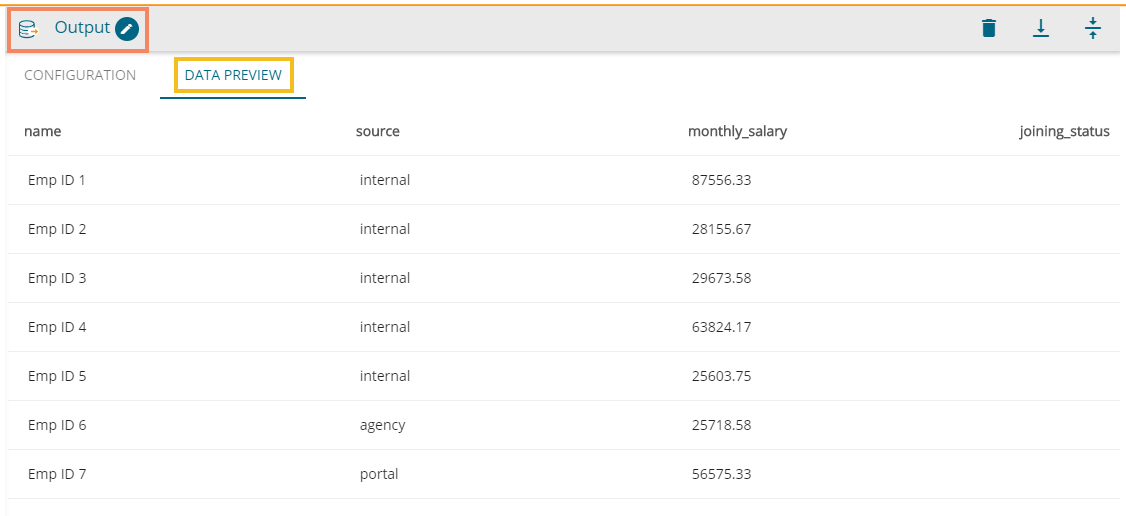
x) Run the workflow.

xi) The notification messages appear displaying update of the data preview process.

xii) A notification appears to confirm the successful completion of the Data Preview process. All the components in the workflow get green checkmark at the top.

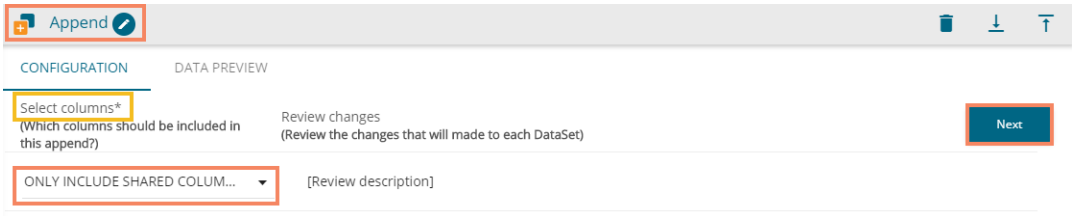


xiii) The selected columns of both the input data sets get appended in the '**DATA PREVIEW**' tab of the Append and Output components.

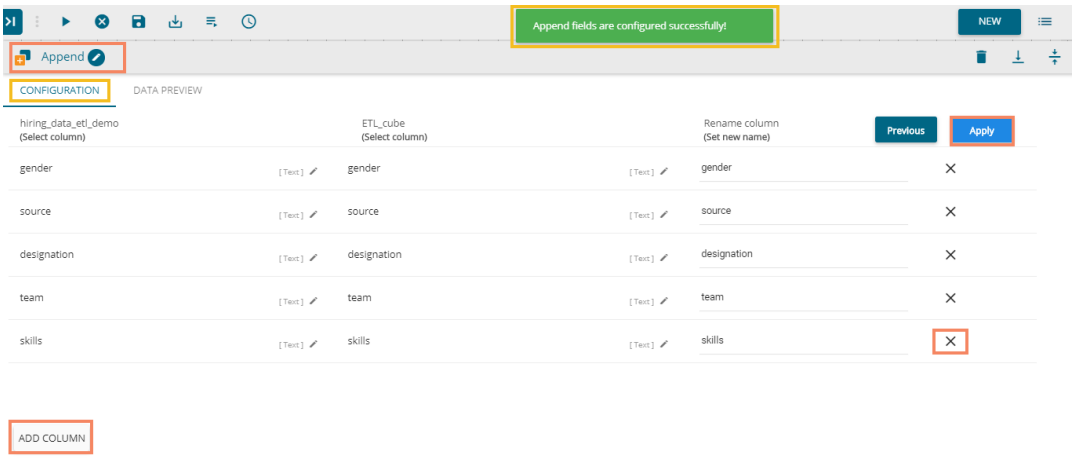


6.1.2. Append Only Shared Columns

i) Connect the '**Append**' component to the configures input datasets and an output component (As described in section 6.1.1.).
ii) Click on the '**Append**' component from the workflow.
iii) The '**CONFIGURATION**' tab opens.
iv) Choose '**ONLY INCLUDE SHARED COLUMNS**' as an option to append the datasets.
v) Click the '**Next**' option.



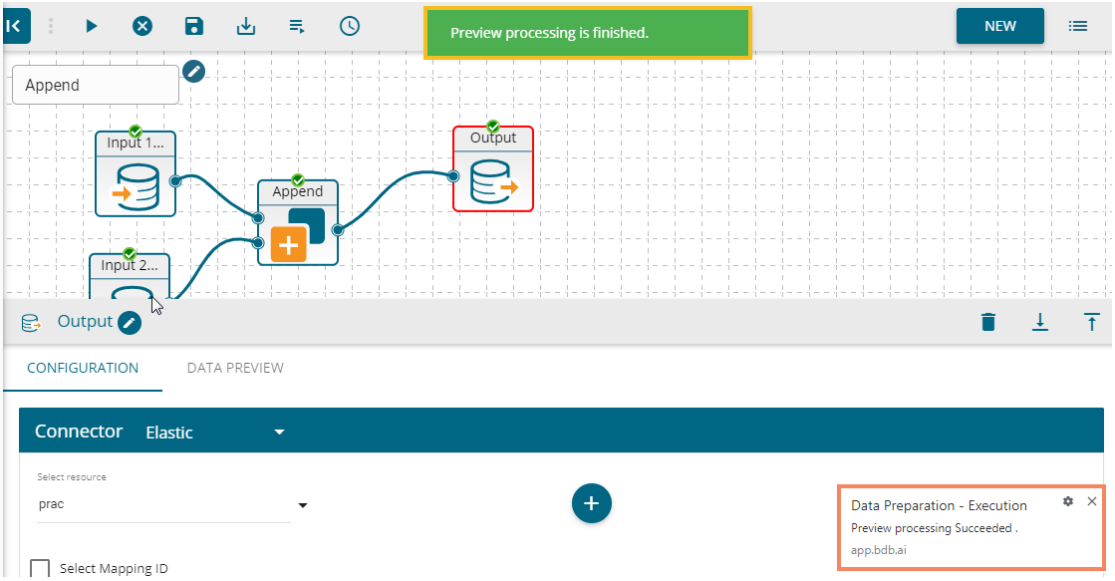
vi) All the columns from both the input data sets get displayed for the user to select the columns from both the datasets. If the user does not select any column, it automatically takes all the shared columns between both the selected datasets.
vii) Click the '**Apply**' option to configure the Append component.
viii) A success message appears if the configuration is successful.



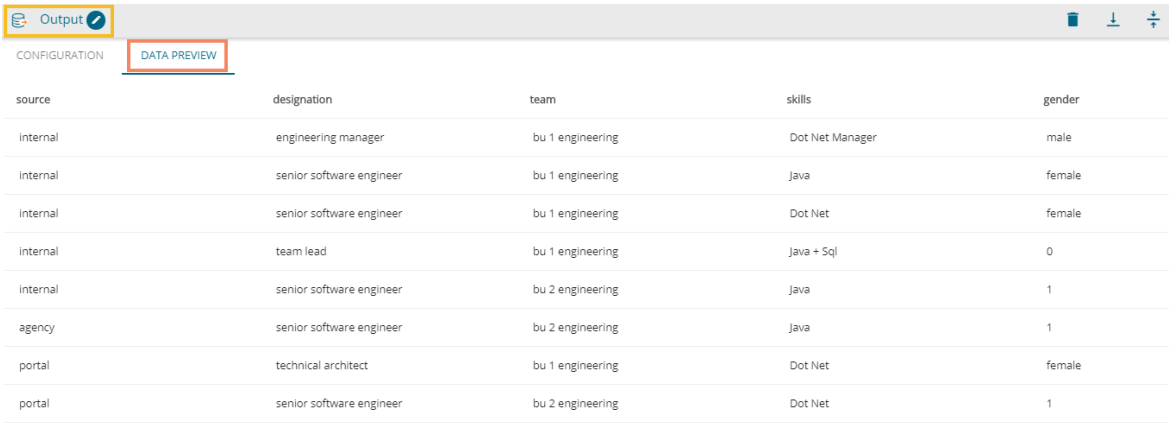
ix) Save the Workflow.

x) Run the Workflow.

xi) After successful execution of the process, all the components in the workflow get green checkmarks, and a success message appears to confirm the successful completion of the process. Note: In case of the failure in the execution process, a notification comes to inform the users about the same. The user can see the reason behind the collapse of the process by opening the '**Job**' UI.



xii) Open the '**DATA PREVIEW**' tab from the Append or Output components to see the appended columns.



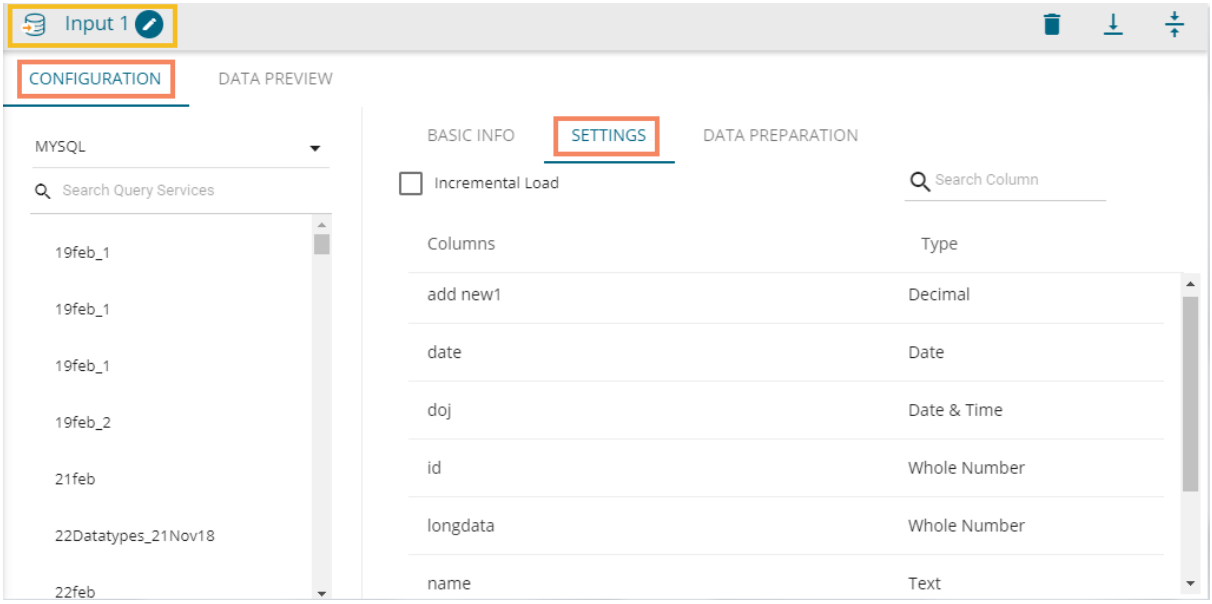
source	designation	team	skills	gender
internal	engineering manager	bu 1 engineering	Dot Net Manager	male
internal	senior software engineer	bu 1 engineering	Java	female
internal	senior software engineer	bu 1 engineering	Dot Net	female
internal	team lead	bu 1 engineering	Java + Sql	0
internal	senior software engineer	bu 2 engineering	Java	1
agency	senior software engineer	bu 2 engineering	Java	1
portal	technical architect	bu 1 engineering	Dot Net	female
portal	senior software engineer	bu 2 engineering	Dot Net	1

6.2. Join

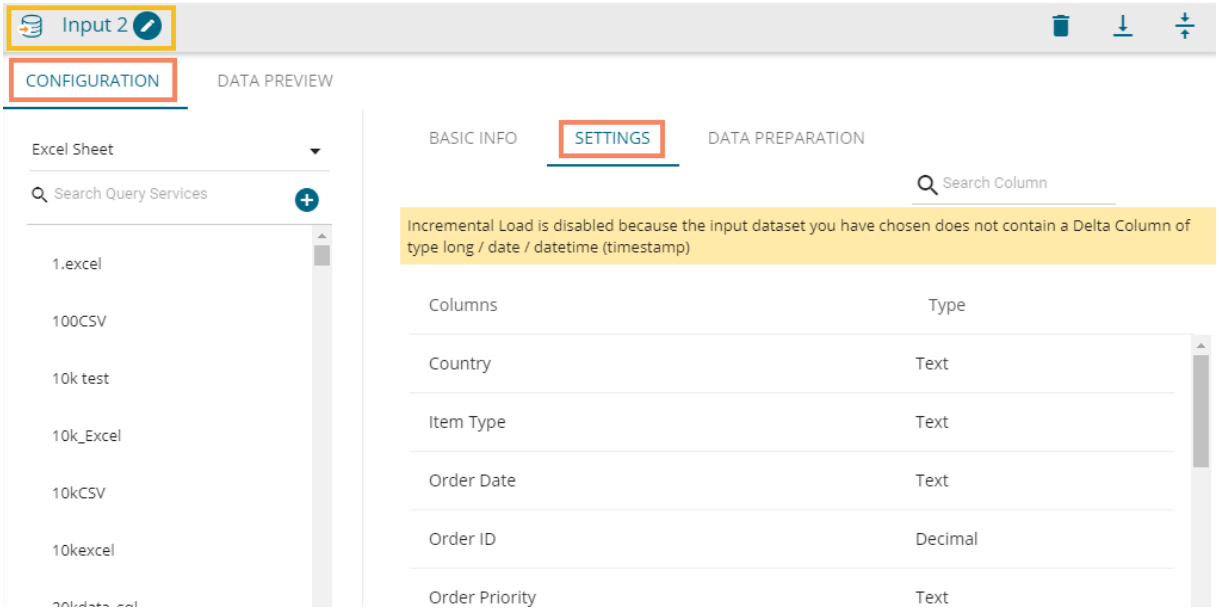
The users can join two datasets and use the merged output to write the workflow in the selected metadata.

i) Drag two input datasets and configure them to see the dataset preview.

Input Data Set 1

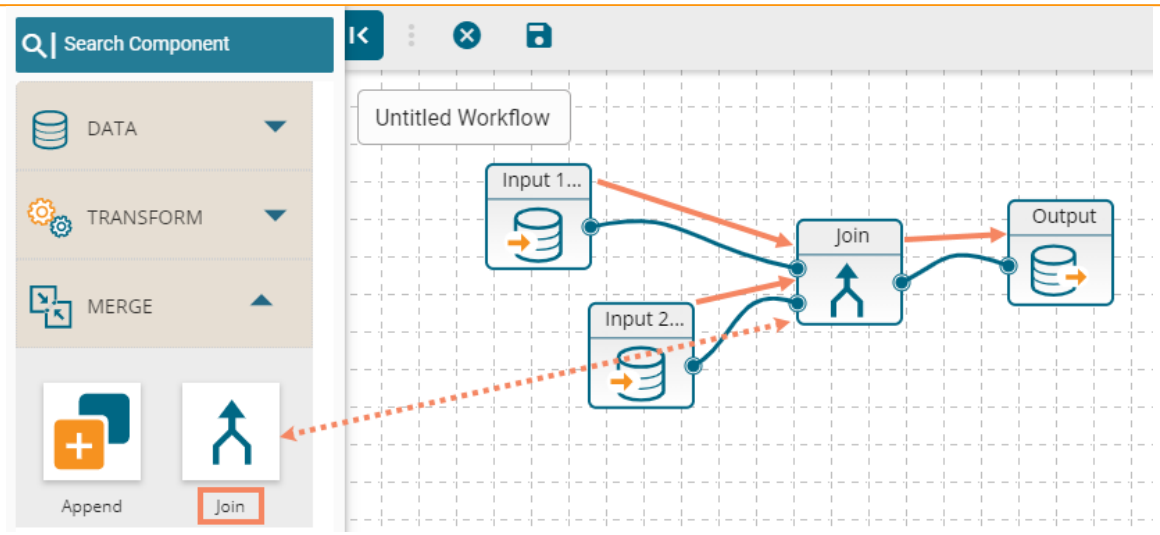


Input Data Set 2



ii) Drag the '**Join**' component to the workspace.

iii) Connect the '**Join**' component to the above-given input datasets and one output component to complete the workflow.



iv) Configure the '**Join**' component as described below:

a. Identify Column: Identify a column from the first Input dataset.

b. Join Type: Choose a join type to merge the selected datasets out of the given choices

   i. Inner

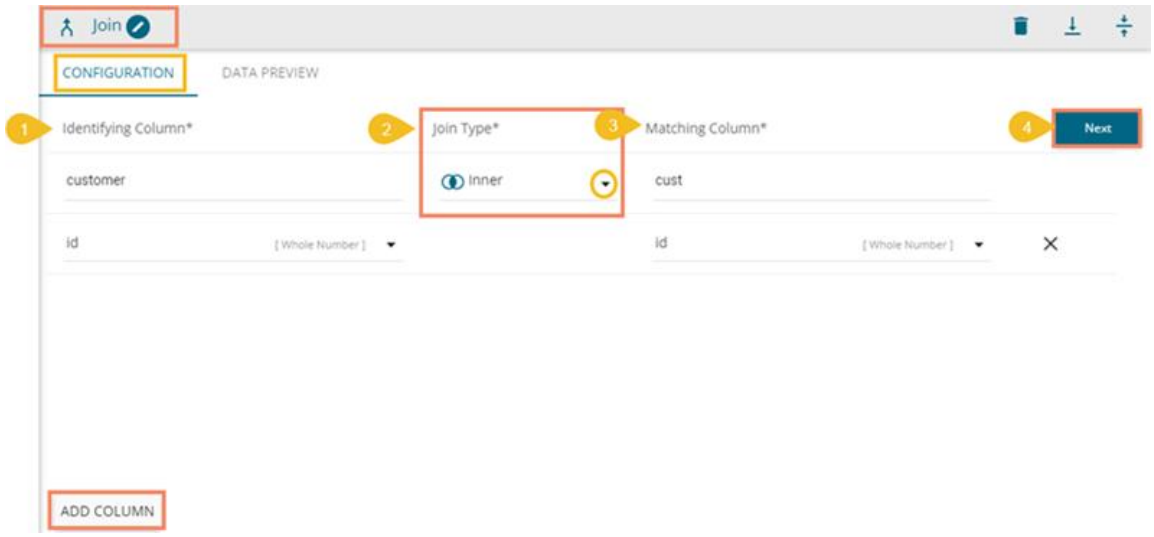
   ii. Left Outer

   iii. Right Outer

   iv. Full Outer

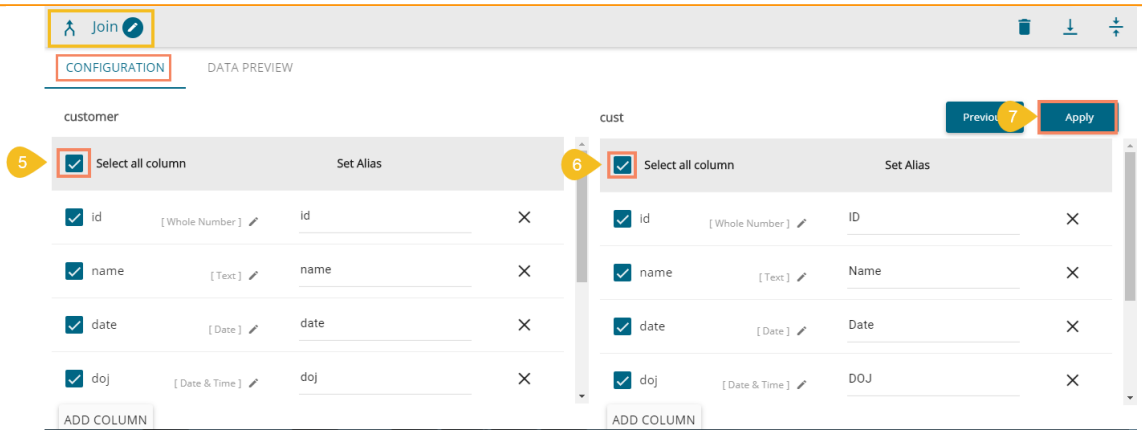
c. Matching Column: Select a column from the second Input dataset.

d. Click the '**Next**' option.



e. The next screen appears prompting the user to select columns from both the input datasets. The users can use the '**Select all column'** option to select all the columns. However, they can also deselect any column by removing the checkmark from the box given next to the column name.

f. Click the '**Apply**' option to configure the selection of columns.

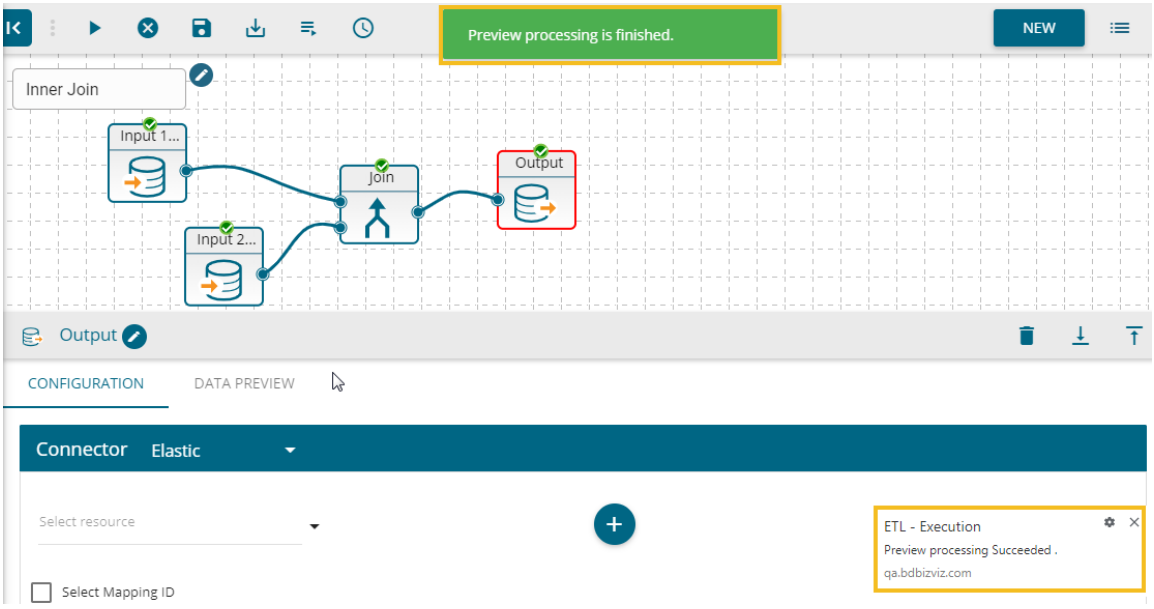


g. A notification appears to confirm the status of the configuration process. (Success or error status is communicated through the color-coded pop-up messages.)

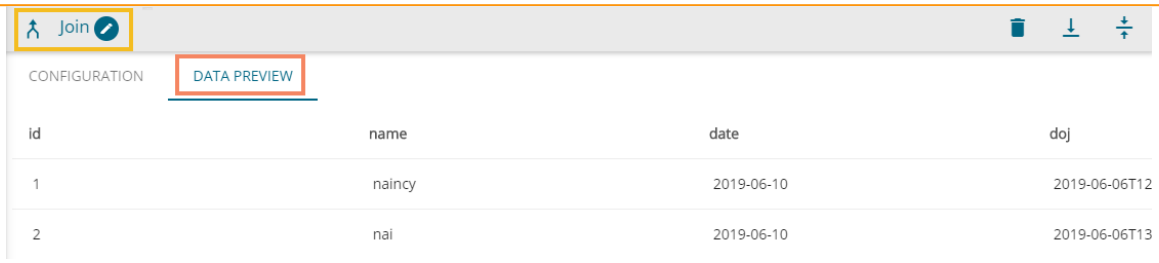
Note:

a. By default, the '**Inner**' join type gets selected.
b. Users can apply multiple joins by using the '**ADD COLUMN**' tab.

v) Save the workflow.
vi) Run/Execute the workflow for data preview.
vii) The notification messages appear displaying update of the data preview process.
viii) A notification appears to confirm the successful completion of the Data Preview process. All the components in the workflow get green checkmark at the top.
Note: The notification message also gets displayed for an error in the data preview process.



ix) Open the '**DATA PREVIEW**' tab using the Join or Output component to see the merged data based on the selected Join type.

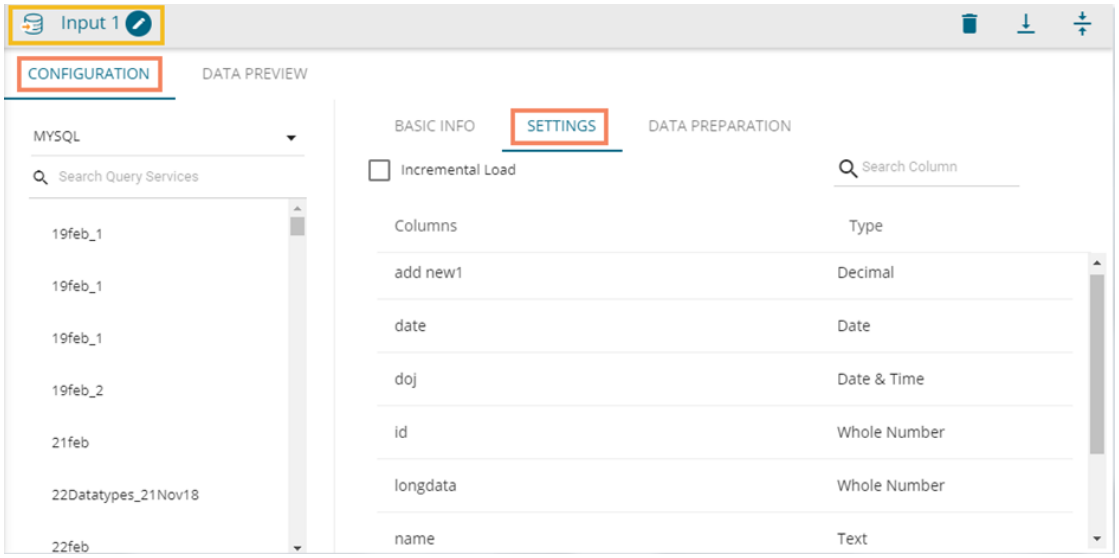


6.2.1. Join Types

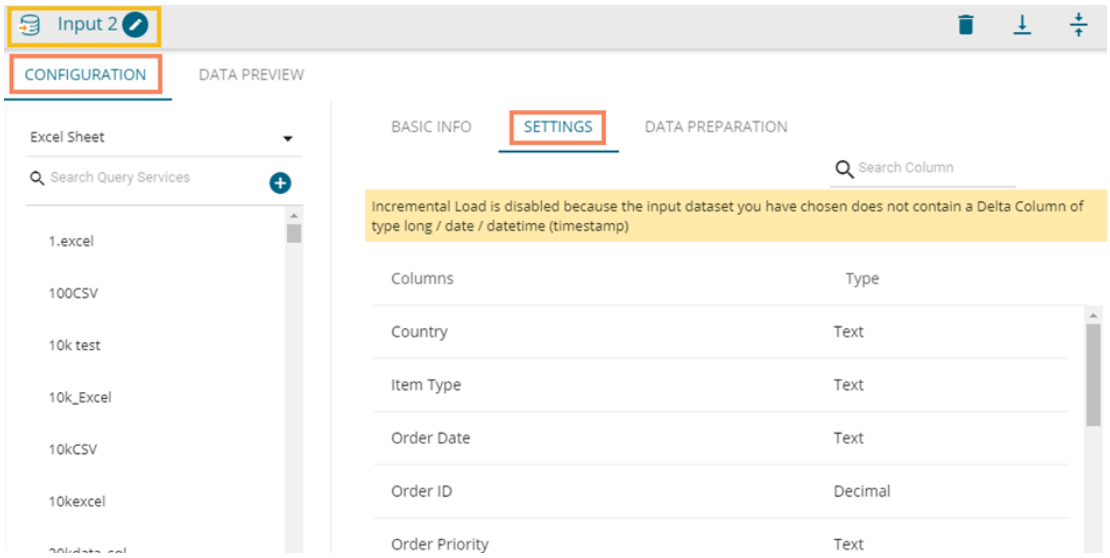
The '**Join**' feature offers four types of join to merge datasets.

The sample data sets used to describe the supported join types are:

1. **Input Dataset 1**

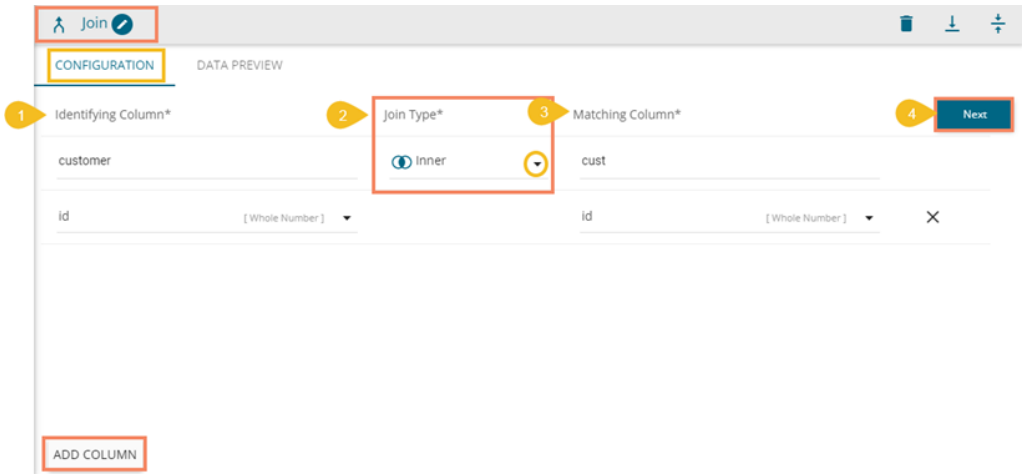


2. **Input Dataset 2**

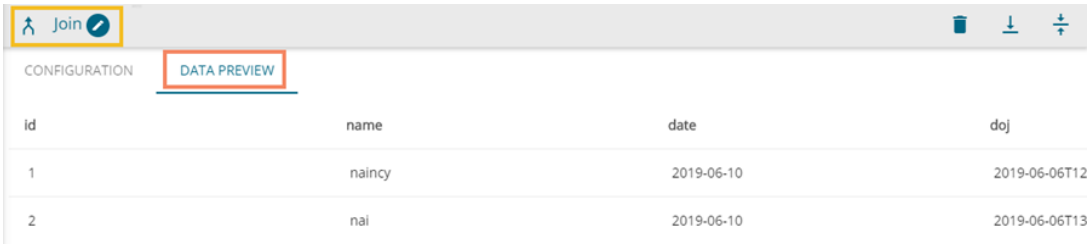


a) **Inner Join**

i. Connect the join component to the configured input datasets and output component to create a workflow.
ii. Configure the '**Join**' component for the '**Inner**' join type.

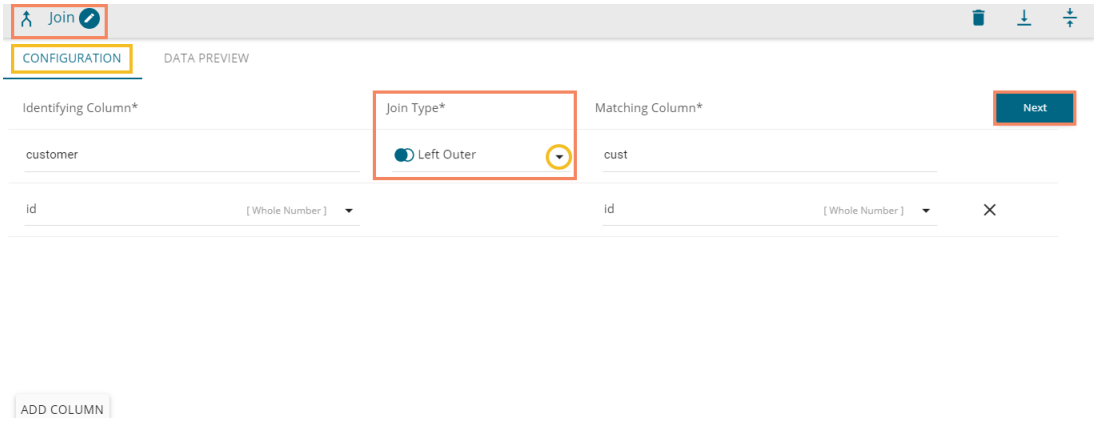


iii. Save and run the workflow to get data preview of the merged data.
iv. Click the '**DATA PREVIEW**' tab using the Join or Output component to view the merged datasets.
v. The DATA PREVIEW displays only matching columns.



b) Left Outer Join

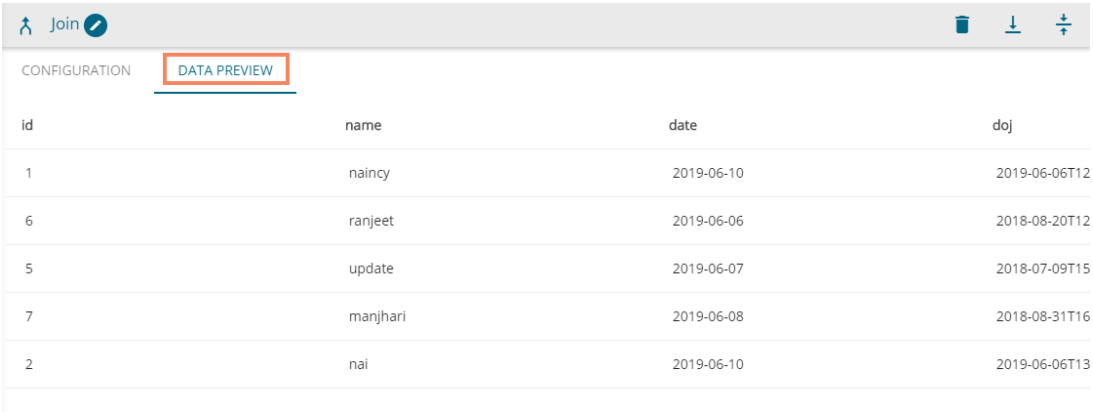
i. Connect the join component to the configured input datasets and output component to create a workflow.
ii. Configure the '**Join**' component for the '**Left Outer**' join type.



iii. Save and run the workflow to get data preview of the merged data.

iv. Click the '**DATA PREVIEW**' tab using the Join or Output component to view the merged datasets.

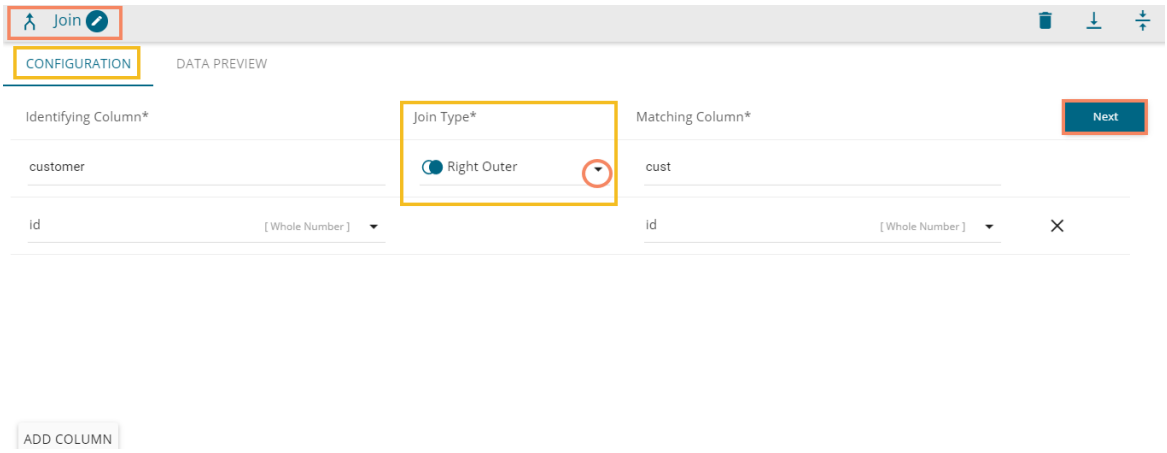
v. The '**DATA PREVIEW**' tab displays the complete data from the left input dataset and matching columns from the other input data set.



c) Right Outer Join

i. Connect the join component to the configured input datasets and output component to create a workflow.

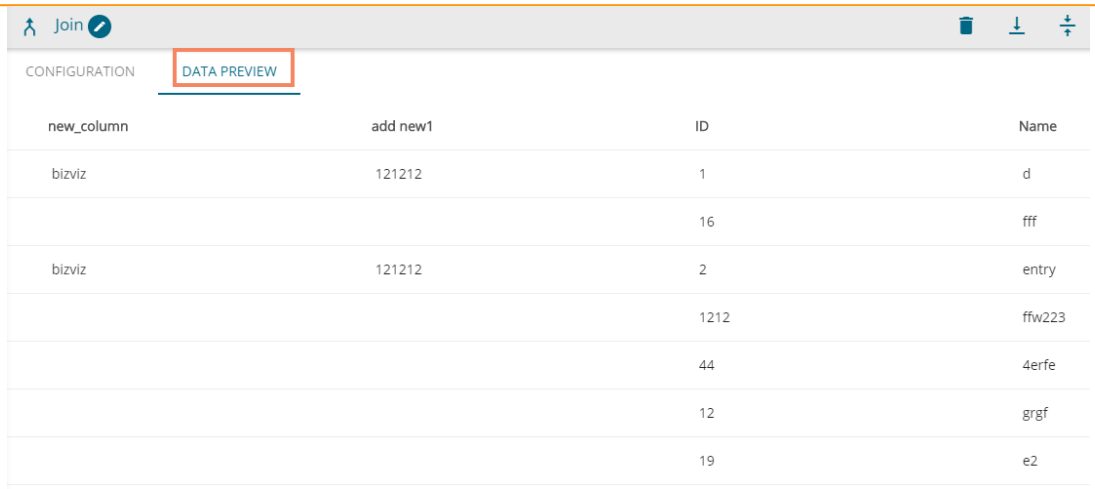
ii. Configure the '**Join**' component for the '**Right Outer**' join type.



iii. Save and run the workflow to get data preview of the merged data.

iv. Click the '**DATA PREVIEW**' tab using the Join or Output component to view the merged datasets.

v. The '**DATA PREVIEW**' tab displays the complete data from the left input dataset and matching columns from the other input data set.

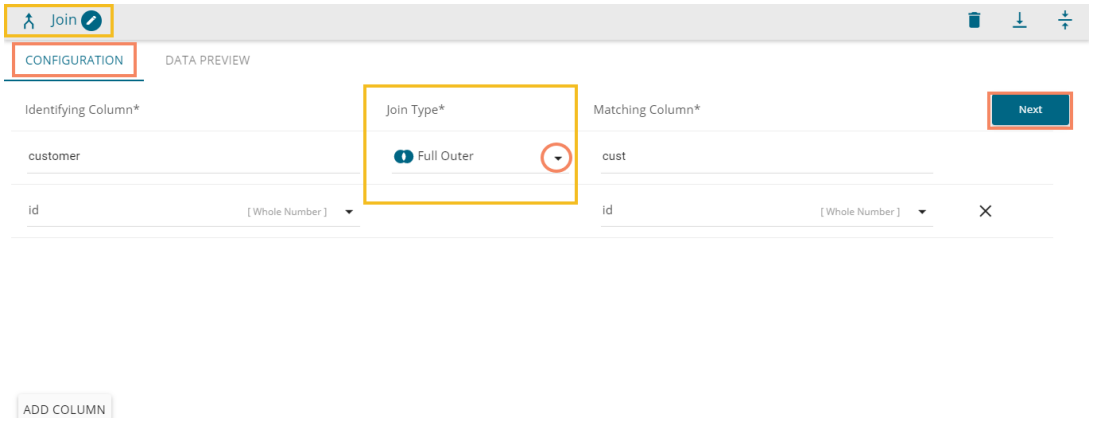


Note: The output data preview will be aligned with the selected left input dataset.

d) Full Outer

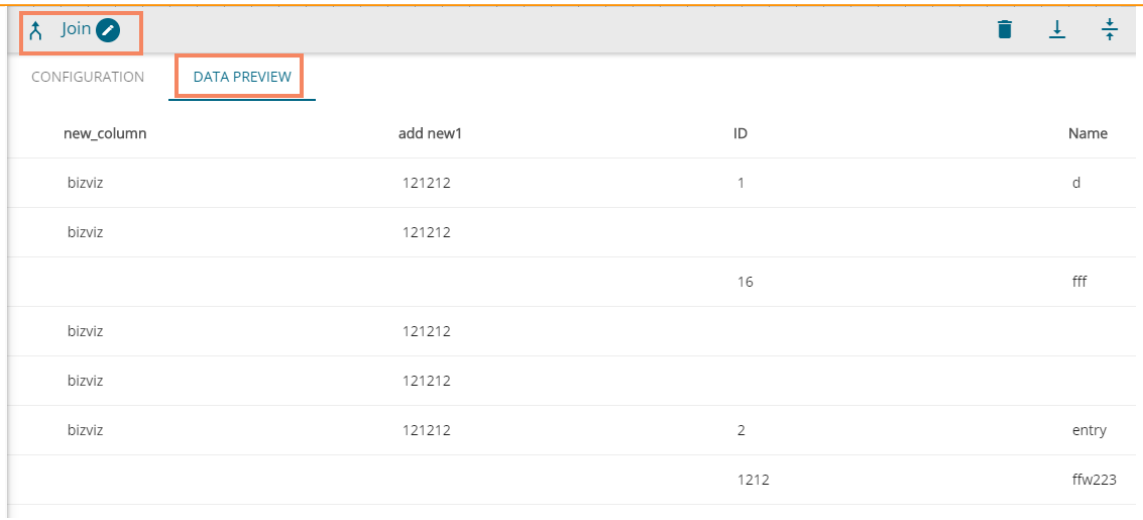
i. Connect the join component to the configured input datasets and output component to create a workflow.

ii. Configure the '**Join**' component for the '**Full Outer**' join type.



iii. Save and run the workflow to get data preview of the merged data.

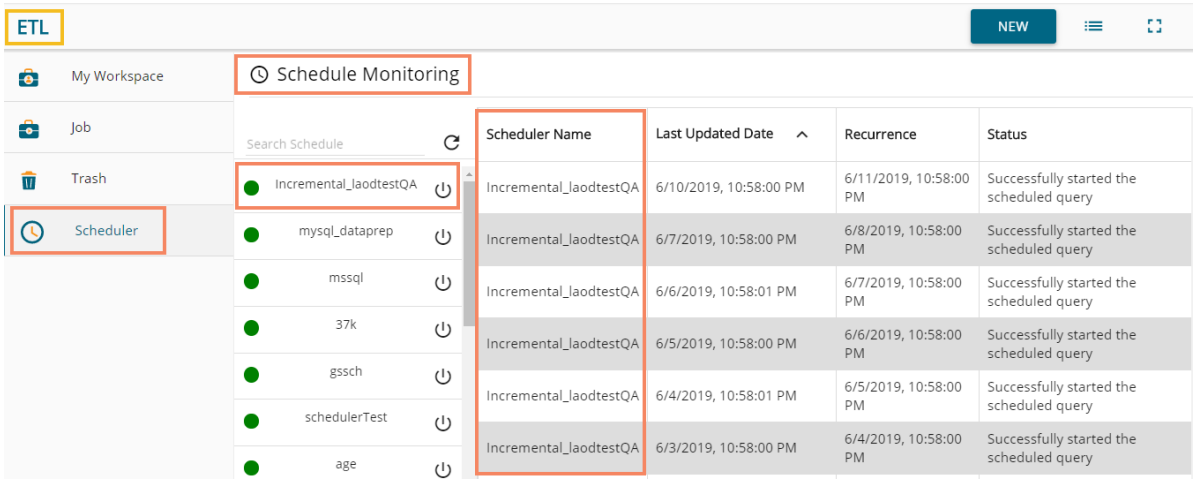
iv. Click the '**DATA PREVIEW**' tab using the Join or Output component to view the merged datasets.



7. Scheduler

The '**Scheduler**' section displays the schedule monitoring details. Users can see a list containing all the scheduled workflows.

i) Click the '**Navigator**' icon.
ii) Select the '**Scheduler**' option from the drop-down menu.
iii) The '**Schedule Monitoring**' page opens.
iv) The scheduled workflow gets listed on the Schedule Monitoring page.
v) By clicking a scheduled workflow from the list, the following schedule details appear:
   a. Scheduler Name
   b. Last Updated Date
   c. Recurrence (Date and Time)
   d. Status

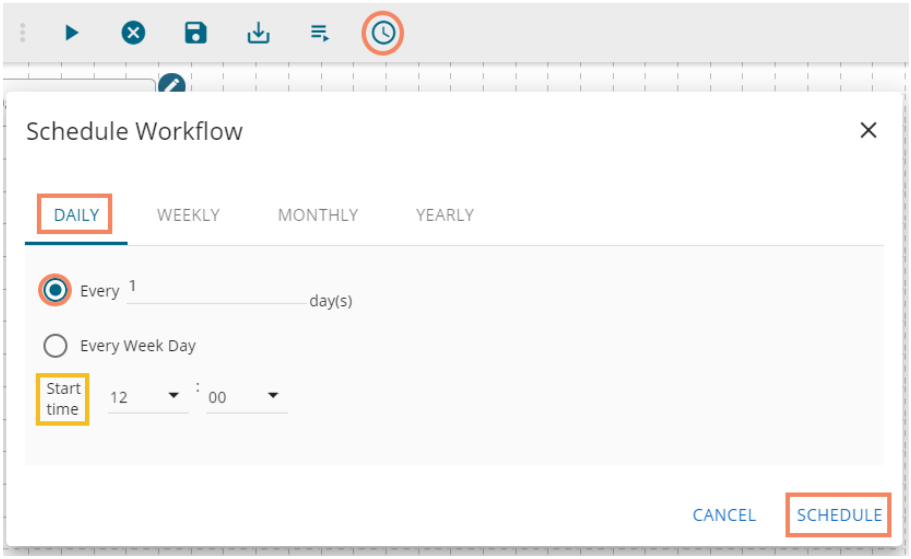


7.1. Schedule Configuration Options

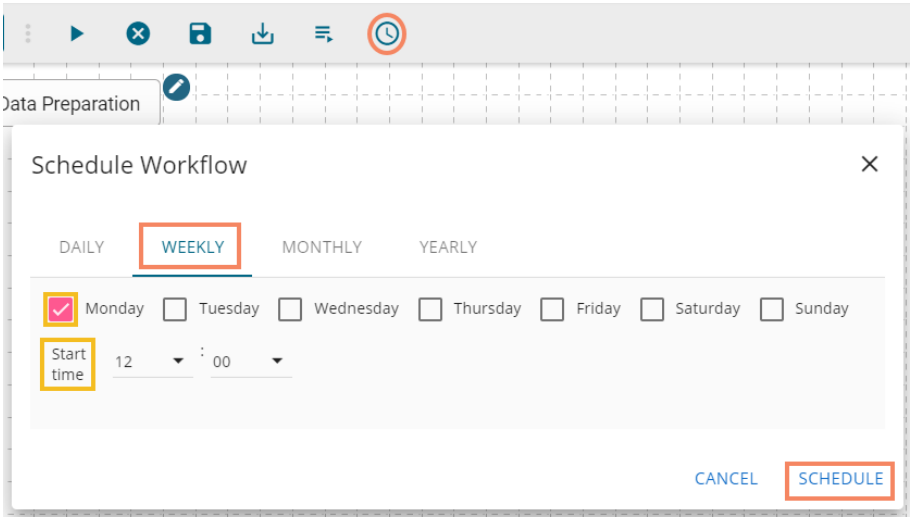
These options are provided to configure a range of time for a scheduled workflow. The user can select only

one option at a time from the given menu.

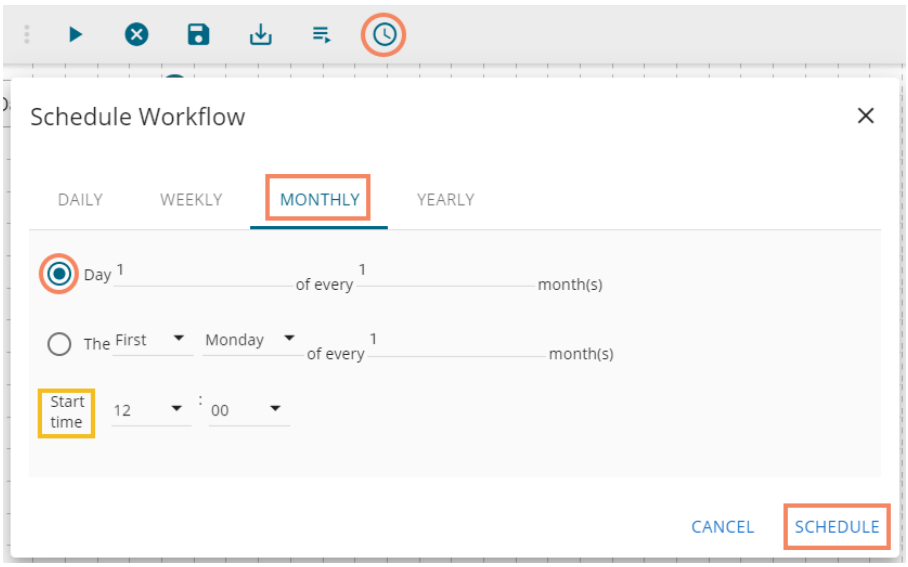
1. **Daily:** User can schedule the job daily by using this option.
   a. Click the '**Scheduler**' icon on the workflow editor.
   b. Choose the '**DAILY**' option from the '**Schedule Workflow**' window (It is a default option).
      i. Select an option out of the given choices
         1. Every __ day(s)
         2. Every Weekday
         3. Set the start time using the drop-down
   c. Click the '**SCHEDULE**' option.



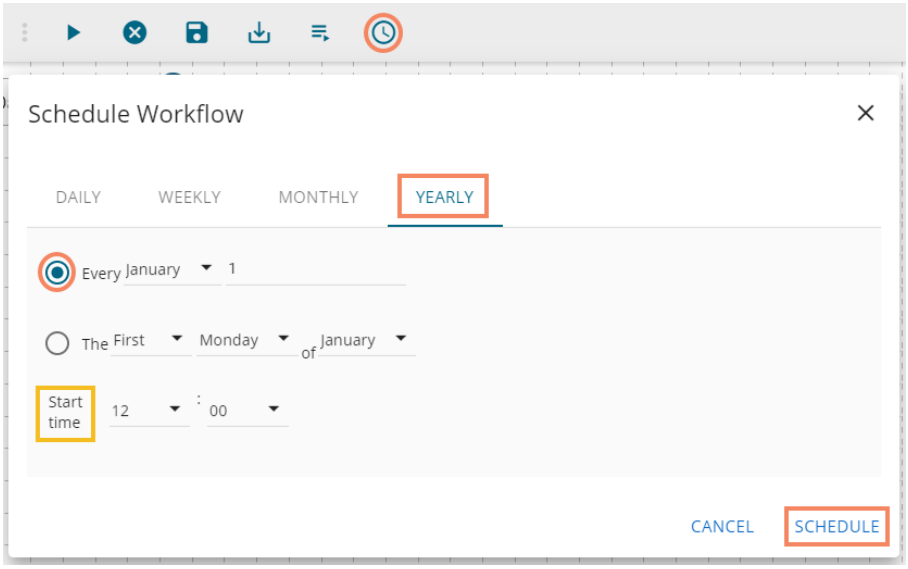
2. **Weekly:** The user can schedule the job weekly by using this option.
   a. Click the '**Scheduler**' icon on the workflow editor.
   b. Choose the '**WEEKLY**' option from the '**Schedule Workflow**' window.
      i. Select an option out of the given choices.
         1. Choose the days of the week by check marking in the box.
         2. Set the start time using the drop-down.
   c. Click the '**SCHEDULE**' option.



3. **Monthly:** User can schedule the job monthly by using this option.
   a. Click the '**Scheduler**' icon on the workflow editor.
   b. Choose the '**MONTHLY**' option from the '**Schedule Workflow**' window.
      i. Select an option out of the given choices to choose a day for each month.
      ii. Set the start time using the drop-down menu.
   c. Click the '**SCHEDULE**' option.

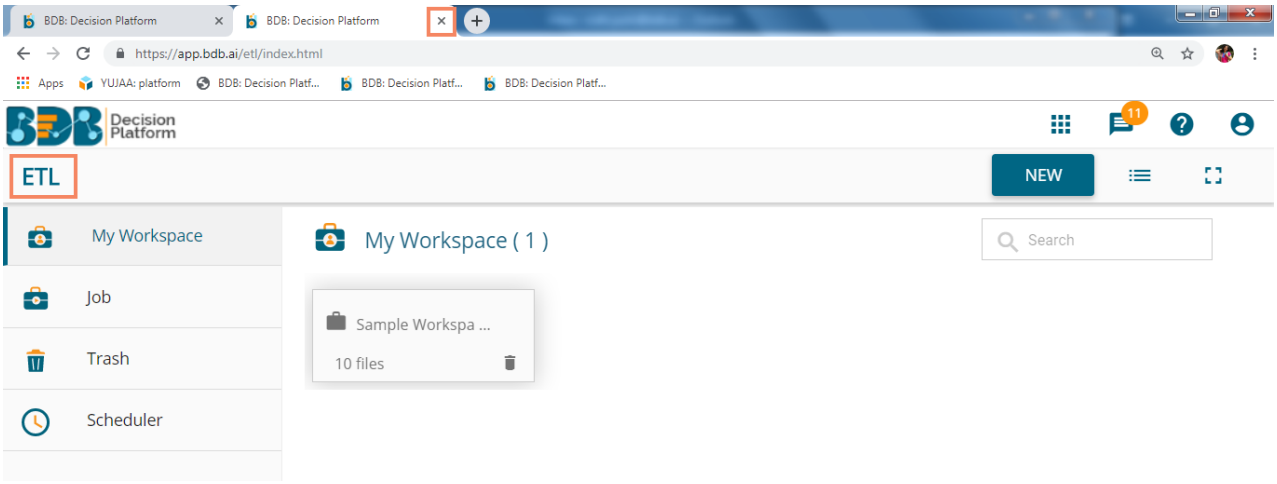


4. **Yearly:** User can schedule the job yearly by using this option.
   a. Click the '**Scheduler**' icon on the workflow editor
   b. Choose the '**YEARLY**' option from the '**Schedule Workflow**' window.
      i. Select an option out of the given choices
         1. Specify either a day or date of a specific month in a year
         2. Set the start time using the drop-down
   c. Click the '**SCHEDULE**' option.



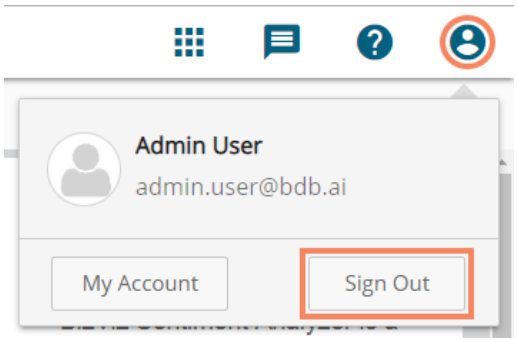
8. Signing Out

It is possible for a user to log out from the BDB ETL plugin at any given stage. Users need to click on the '**Close**' ✕ option to close the ETL page.



The following steps describe how to log out from the BDB Platform.

i) Click the '**User**' icon on the Platform homepage.
ii) The details of the logged in user appears.
iii) Click the '**Sign Out**' option.



iv) The user successfully signs off from the **BDB Platform.**

Note: By clicking the '**Sign Out**' option, the user gets back to the Sign in page of the BDB platform.