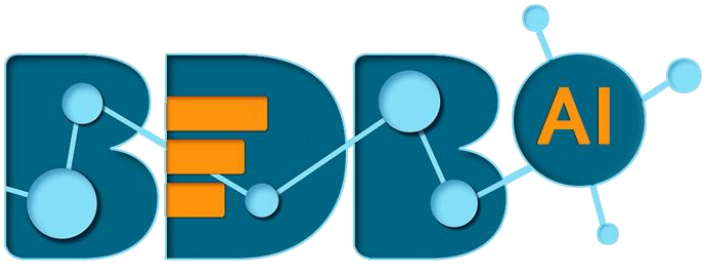


User Guide

Data Preparation- 6.0

Contents

1. Introduction

1.1. Introducing Data Preparation

The Data Preparation module can turn any Business data into a cost-effective and custom-made experience. The Data Analysts can instantly detect anomalous records (rows with invalid or empty values) and purge the unwanted data sets in a few clicks using Machine-Learning based smart techniques and sampling. The users can identify errors and apply changes to data set from any source and export the analysis-ready data in minutes. Automated detection of groups and categories in your data can be viewed through a frequency table. The user can filter the group in a single click and transform data matching the filter conditions and get intelligent Data Transformation suggestions based on data type and quality.

1.2. Document History

Product Version	Date (Release Date)	Description
Data Preparation 4.0	December 31st, 2018	First Release of the document
Data Preparation 4.2	March 25th, 2019	Updated document
Data Preparation 4.3	April 24th, 2019	Updated document
Data Preparation 4.4	June 7th, 2019	Updated document
Data Preparation 4.5	August 5th, 2019	Updated document
Data Preparation 4.6	November 15th, 2019	Updated document
Data Preparation 5.2.0	August 21st, 2020	Updated document
Data Preparation 6.0	February 26th, 2021	Updated document

Note: The Product Version column displays Release Version of the BDB Platform.

1.3. Overview

This guide covers:

- Explanation and usage of all the Data Preparation options
- Explanation and usage of the Transforms
- Integration with Data Pipeline

1.4. Target Audience

The document is targeted to the following audience:

- Data Engineers
- Citizen Data Scientists

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:

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

The supported browser versions are driven by the capabilities the UI employs and the dependencies it uses. UI features will be developed and tested against the supported browsers.

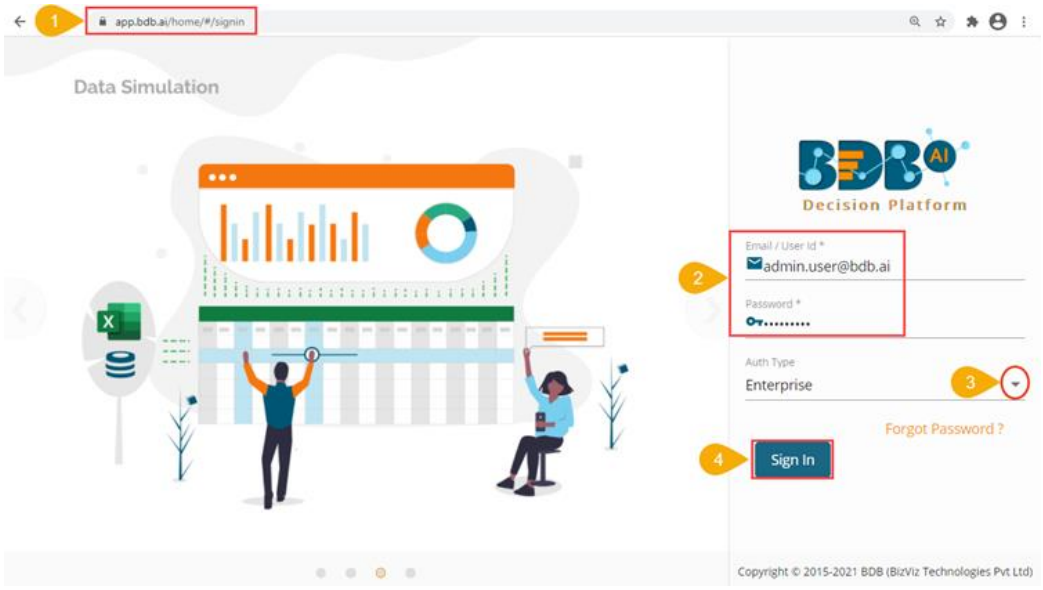
2.1. Unsupported Browsers

While the UI may run successfully in unsupported browsers, it is not actively tested against them. Additionally, the UI is designed as a desktop experience and is not currently supported in mobile browsers.

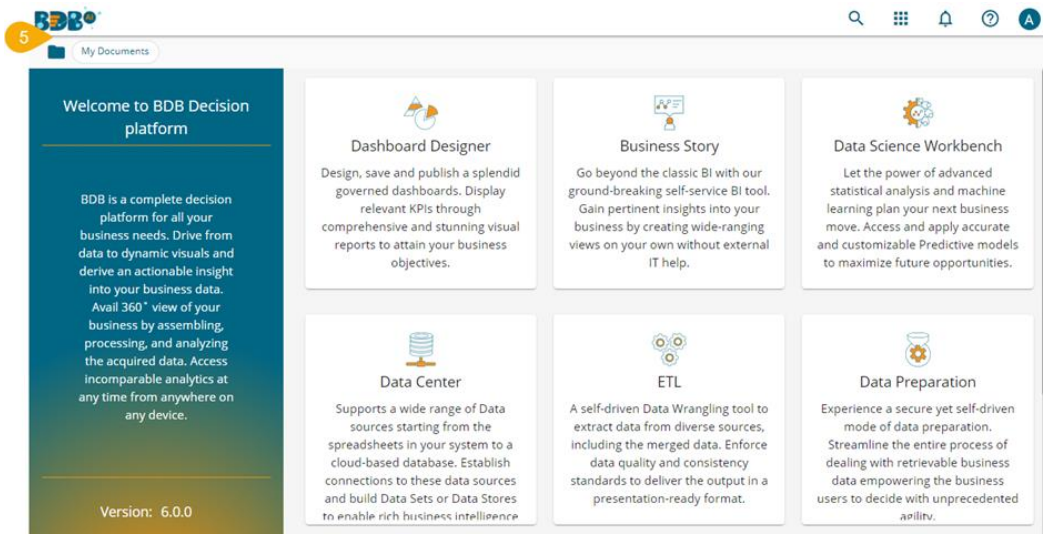
3. Getting Started with BDB Data Preparation

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.
iii) Select an Auth Type from the drop-down menu.
iv) Click the '**Sign In**' option.



v) The Platform homepage opens.



Note:

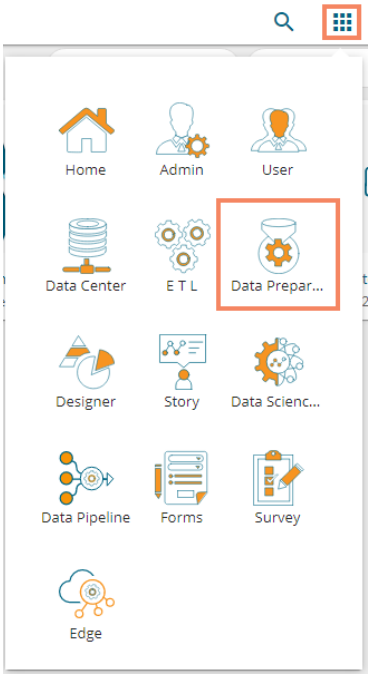
a. The above screen opens only for those newly created users who have not yet created any document using the BDB Platform.

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 '**Data Preparation**' plugin.



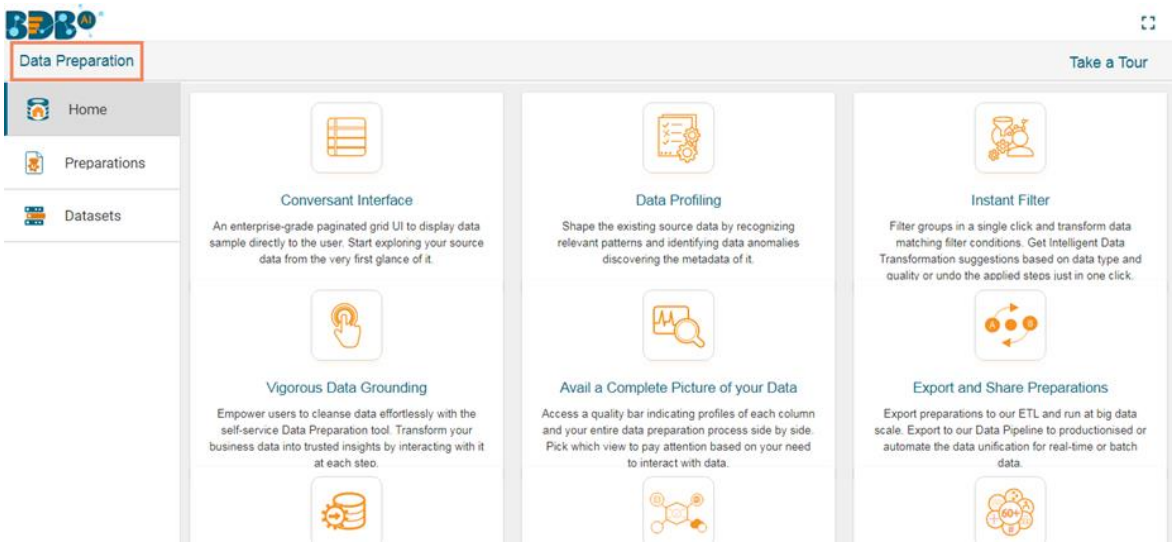
ix) The Data Preparation landing page opens.

x) The major Data Preparation modules get displayed on the landing page:

a. Home (Default module)

b. Preparations

c. Datasets

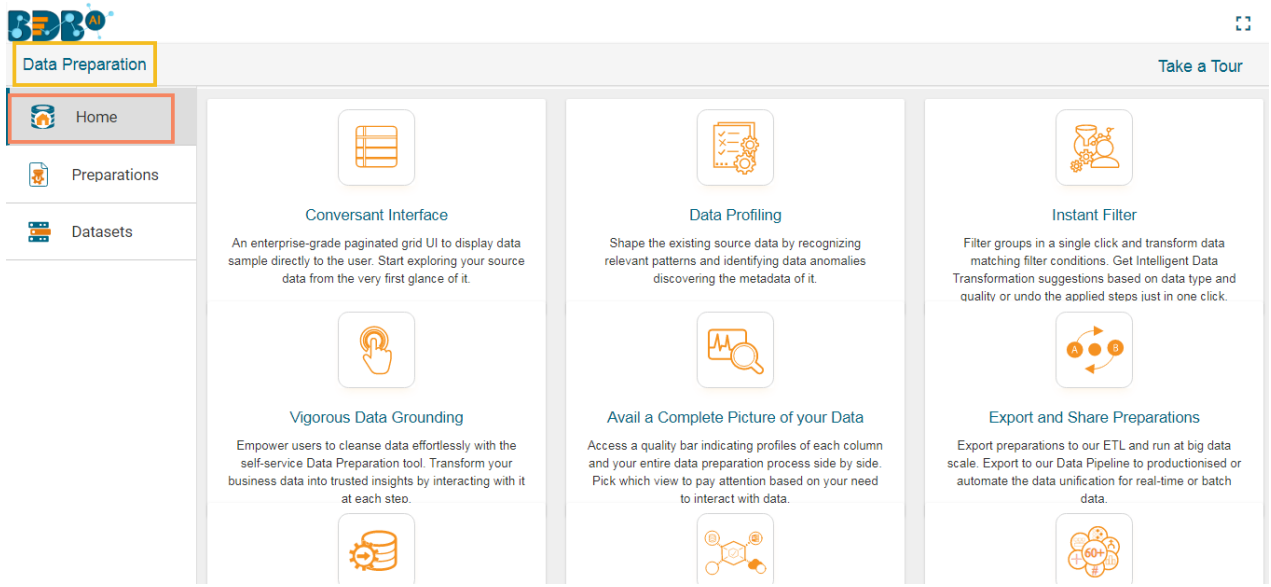


xi) The user can start a guided tour of the plugin by clicking on the '**Take a Tour**' option.
xii) Click the '**Next**' option.

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

4. Data Preparation Landing Page

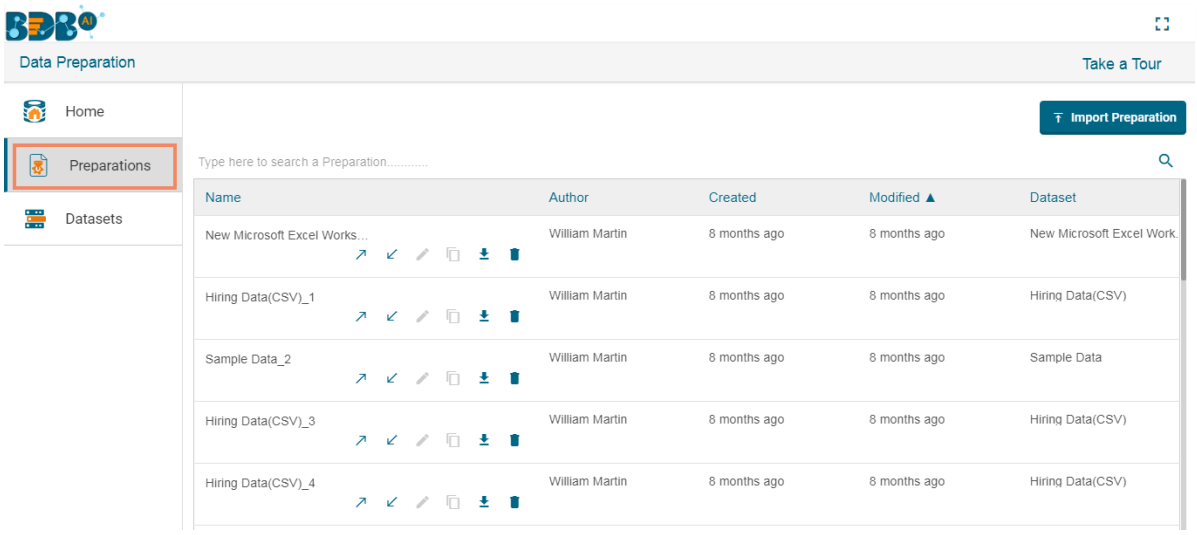
The landing page of the data preparation has three menus: 1) Home, 2) Preparations, and 3) Datasets. The '**Home**' page opens by default while selecting '**Data Preparation**' plugin from the Apps menu.



The user can start the data preparation process by uploading a dataset, and the newly created preparation gets saved under the '**Preparations**' tab.

4.1. Preparations

The '**Preparations**' tab lists all the available preparations displaying Name, Author, when it was created, when it was last modified and using which data`1set it was created.

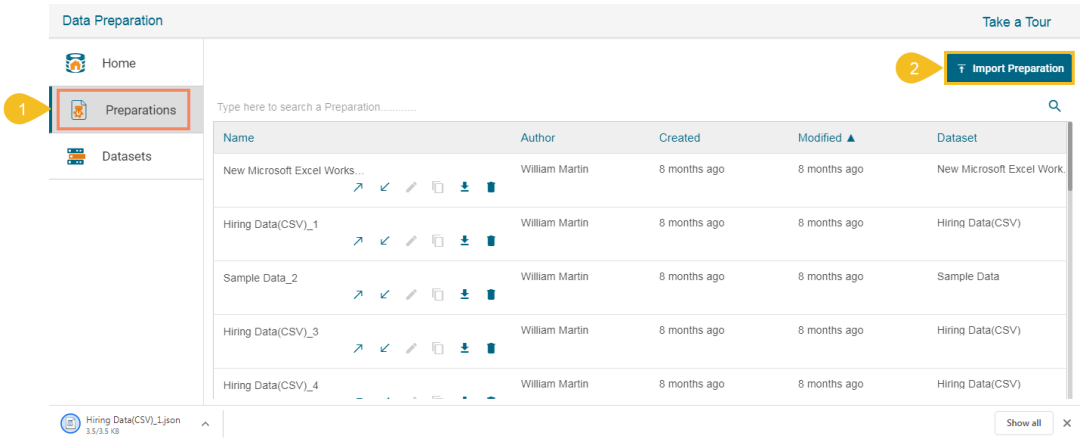


The user can continue adding more steps to the existing preparations. The user can import an existing preparation using the '**Import Preparation**' option.

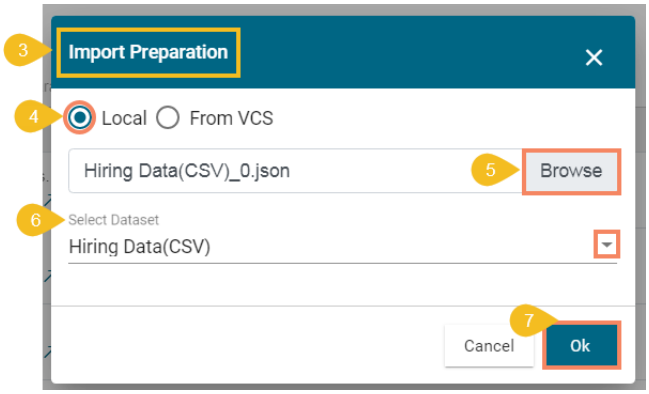
4.1.1. Importing a Preparation

This feature can be used to apply a set of cleansing steps on a dataset with similar metadata.

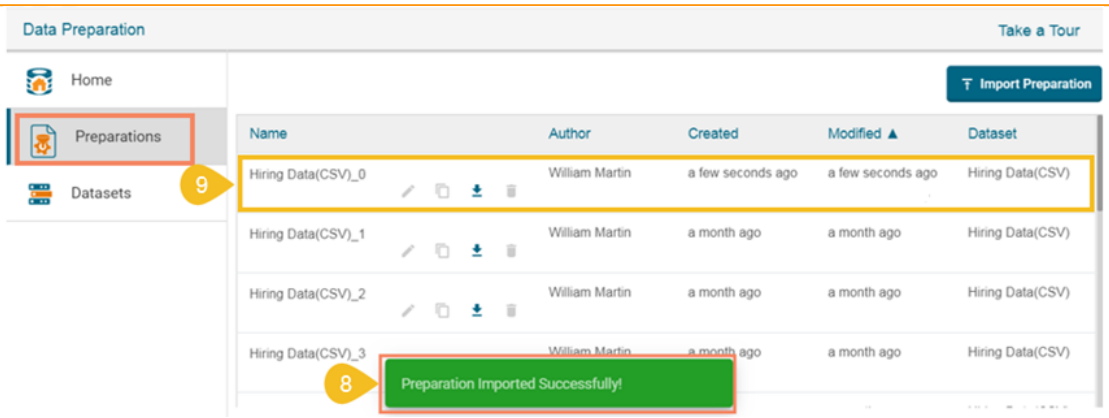
i) Navigate to the '**Preparations**' list.
ii) Click the '**Import Preparation**' option.



iii) The '**Import Preparation**' window opens.
iv) Select an option by marking the checkbox out of Local and VCS options.
v) Browse a downloaded JSON file.
vi) Select a dataset of similar metadata from the drop-down menu.
vii) Click the '**Ok**' option.



viii) A success message appears.
ix) The Preparation gets imported and applied to the selected dataset.

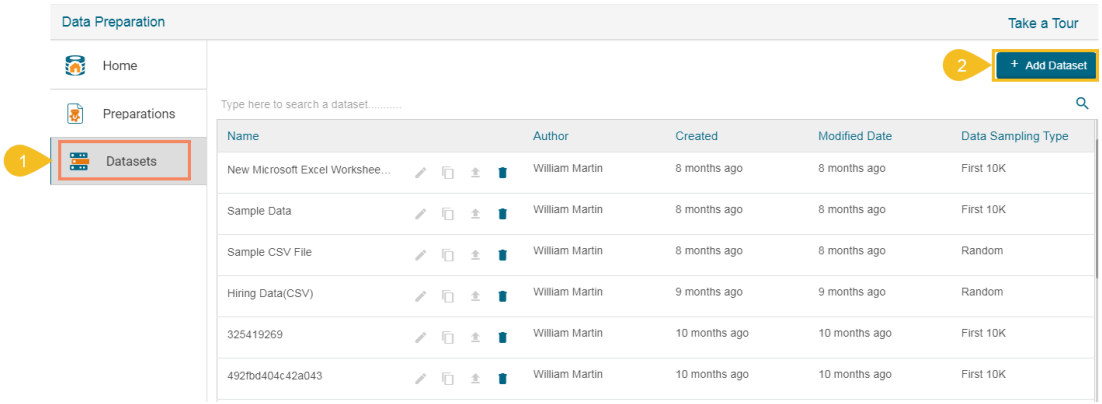


4.2. Datasets

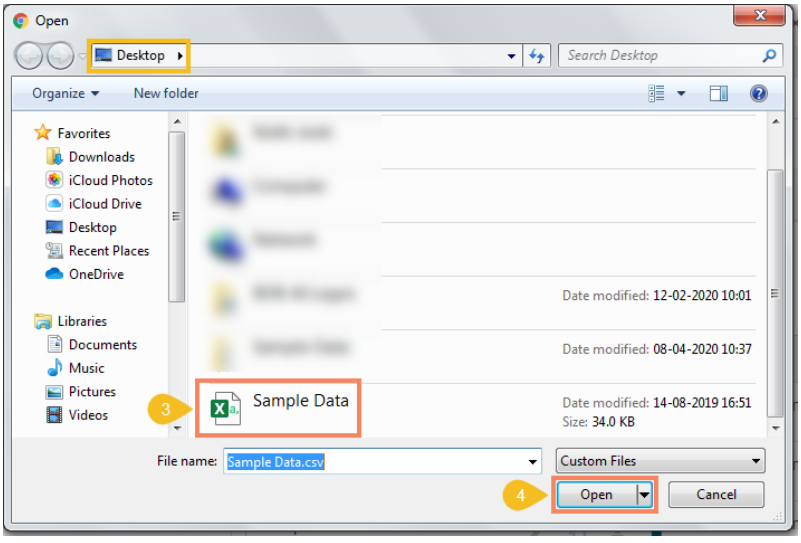
The '**Datasets**' section lists the data/inputs added to the system. The user can create a new preparation by selecting any of the listed datasets. The Datasets window also provides an option to add new datasets.

4.2.1. Adding a new Dataset

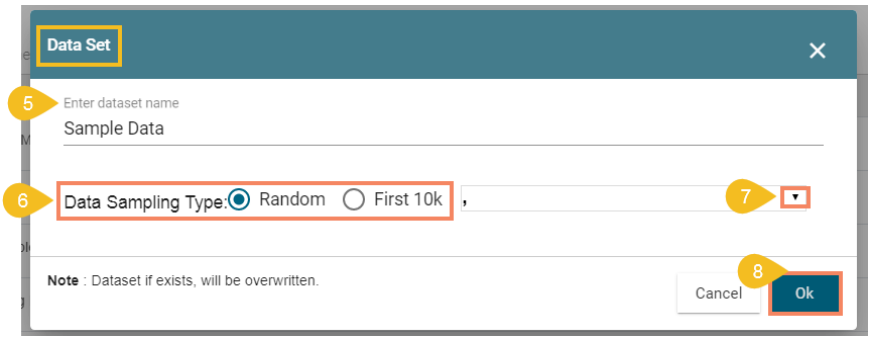
i) Navigate to the Datasets option.
ii) Click the '**Add Dataset**' option.



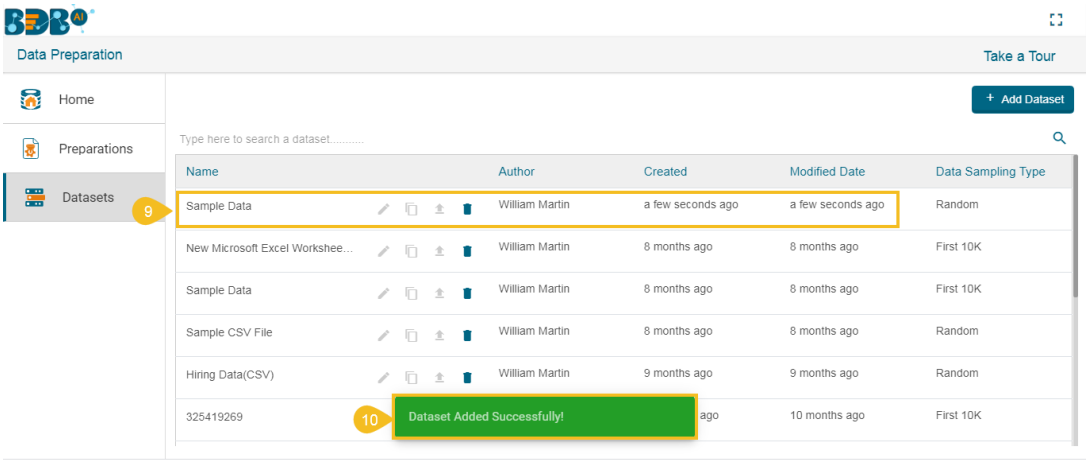
iii) A new window opens redirecting the user to select a CSV file.
iv) Browse the file and upload it.



v) The Data Set window appears with the selected CSV file.
vi) The user can select a **Data Sampling Type** out of **'Random'** or **'First 10k'** options by marking the radio button.
vii) The user can set an option to separate the data values in the selected file using the drop-down menu (E.g., Comma is selected in the following image).
viii) Click the '**Ok**' option.



ix) A success message appears.
x) The selected CSV File gets added to the Datasets page.

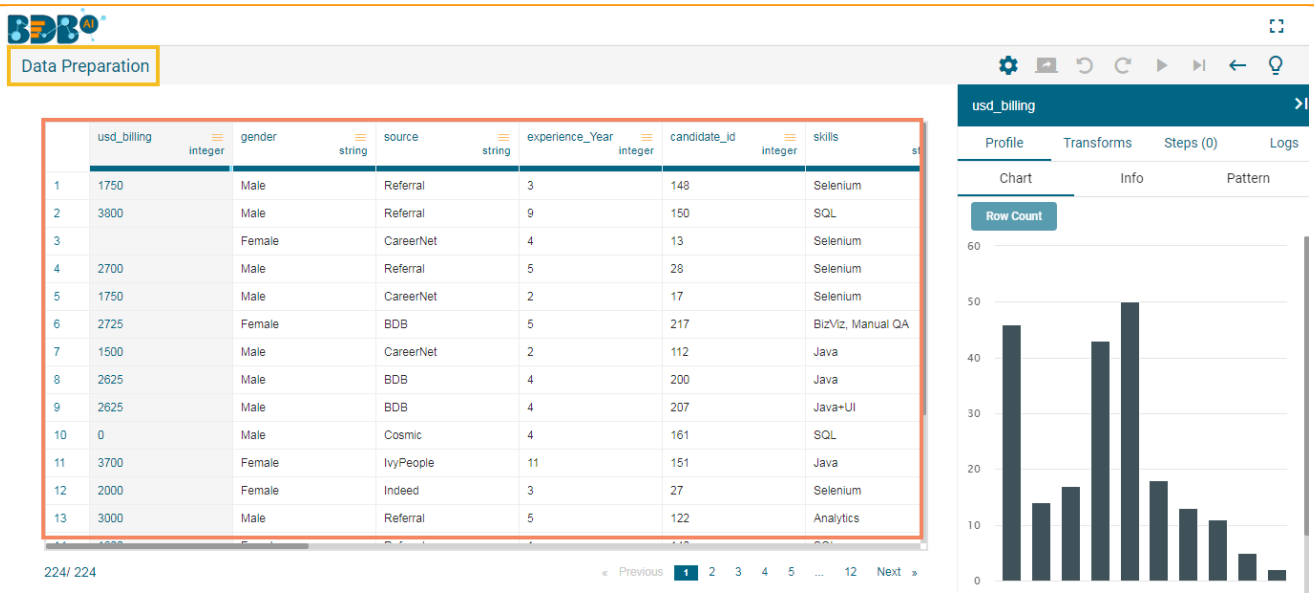


Note: The standalone version of the Data Preparation supports only CSV input of max 10k records. To work on other data sources and colossal volume, please use the ETL integrated version of data cleansing.

5. Data Grid

The Data Grid in the BDB Data Preparation is used for visualizing the data. The data displayed in the grid is a sample from the actual data set or complete data based on the data volume. The grid always shows the first 10 K rows in the dataset.

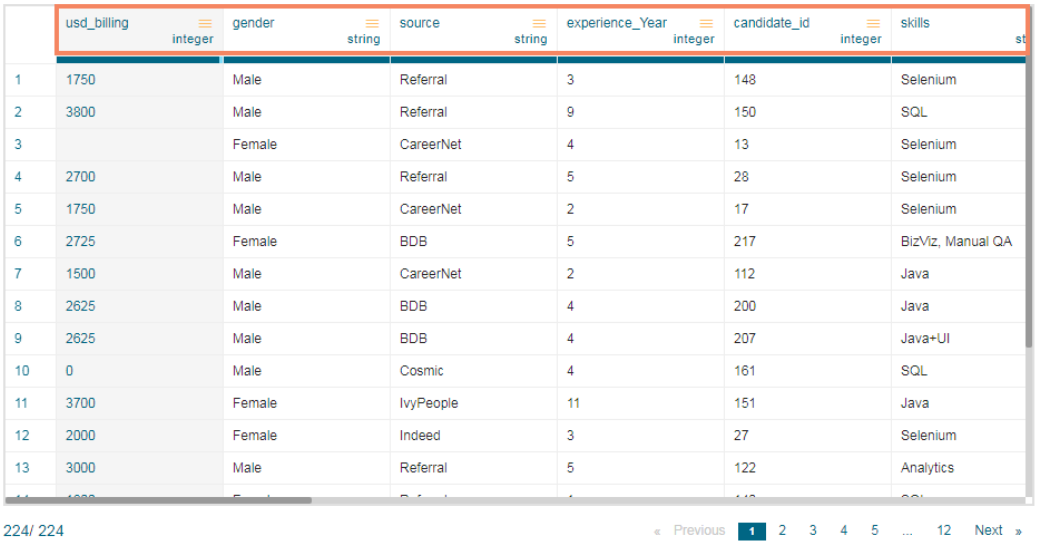
The user can access the Data Grid view of the selected dataset or data preparation by clicking on it. The displayed data in the grid changes based on the number of transforms performed on it.



Note: The above image displays the data grid page opened when clicked on a dataset from the Datasets page.

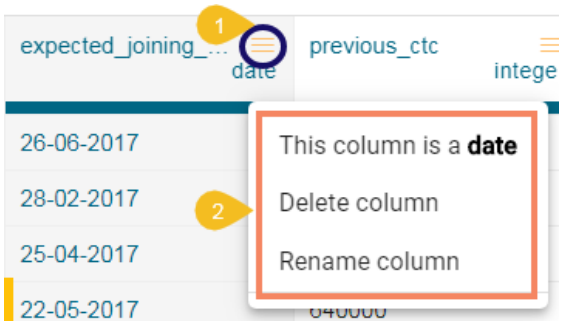
5.1. Data Grid Header

The grid has a header that displays the column name and column type from the selected dataset.



	usd_billing integer	gender string	source string	experience_Year integer	candidate_id integer	skills
1	1750	Male	Referral	3	148	Selenium
2	3800	Male	Referral	9	150	SQL
3		Female	CareerNet	4	13	Selenium
4	2700	Male	Referral	5	28	Selenium
5	1750	Male	CareerNet	2	17	Selenium
6	2725	Female	BDB	5	217	BizViz, Manual QA
7	1500	Male	CareerNet	2	112	Java
8	2625	Male	BDB	4	200	Java
9	2625	Male	BDB	4	207	Java+UI
10	0	Male	Cosmic	4	161	SQL
11	3700	Female	IvyPeople	11	151	Java
12	2000	Female	Indeed	3	27	Selenium
13	3000	Male	Referral	5	122	Analytics

Each Column Header has a Context menu ≡ icon. The Context menu displays the Column Type, option to Delete the column, and option to Rename.



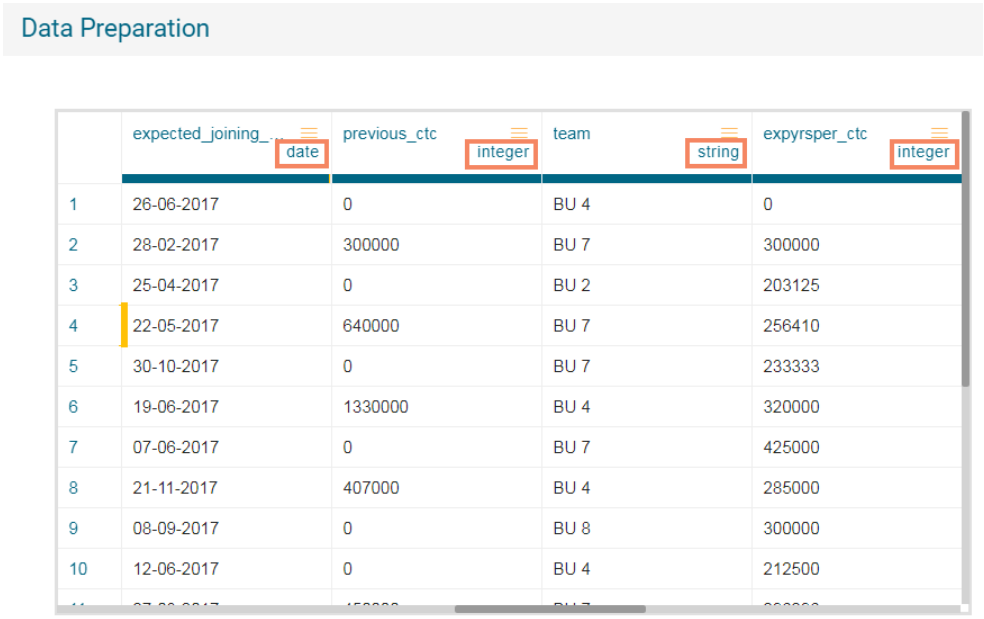
It also presents the data type of the column. It is analyzed based on the max match to any data type in the first 10K records.

Consider that out of 10000 rows sample, there are 9000 integers and 1000 string values, the selected data type is Integer. The 1000 string rows get detected as invalid rows.

5.1.1. Data Types

The Data Grid header displays Data Types. It supports the following data types:

1. Integer
2. Double
3. String
4. Date
5. Timestamp



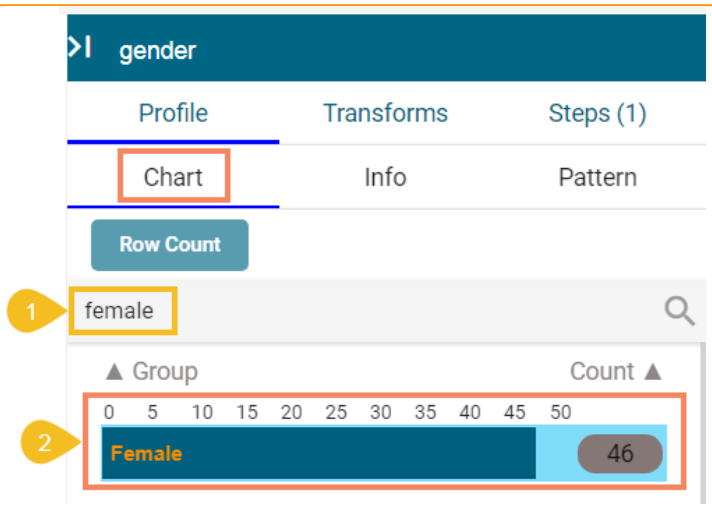
5.2. Panel to List the Selected Filters.

The user can insert a filter condition by using the 'Search' option provided under the '**Chart**' tab. When a filter is selected, it gets added to the filter panel on the top of the data grid. The added filter has an option to remove it by clicking the '**Close**' (X) mark.

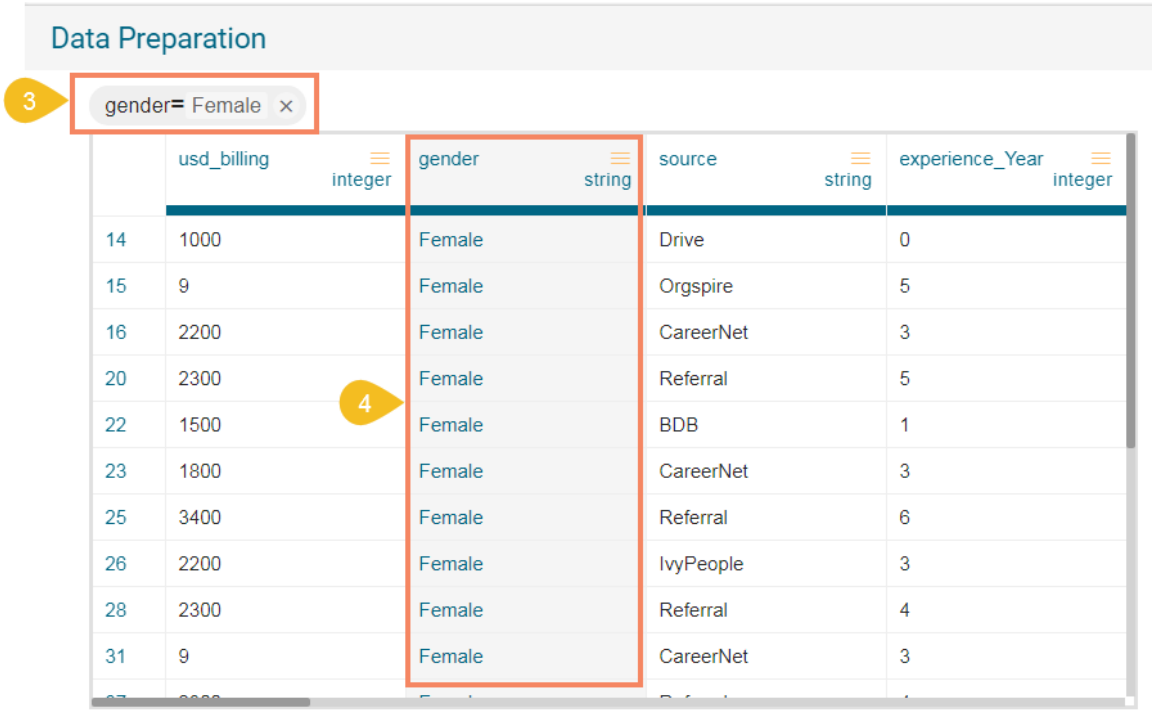
Steps to Apply Filter Condition

i) Type the filter condition using the '**Search**' bar.

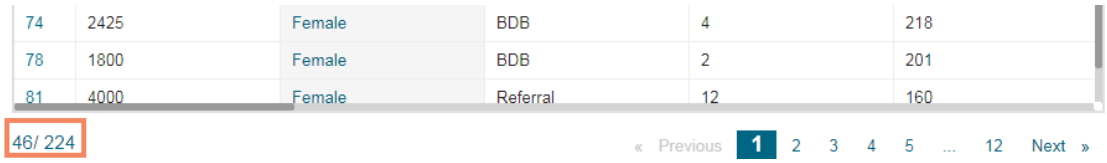
ii) The Bar chart displays the searched value. Click on it to apply the filter.



iii) The applied filter condition displays on the top of the data grid table.
iv) The selected column displays the filtered data.



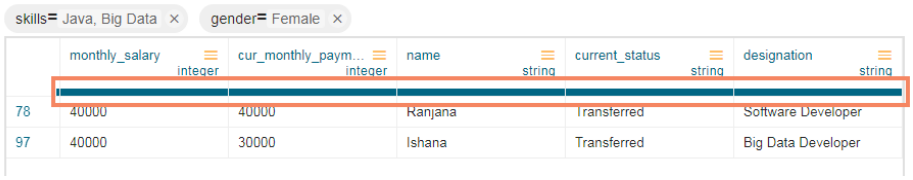
v) The rows meeting the filter condition out of the total number of rows gets displayed on the bottom-left part of the page.



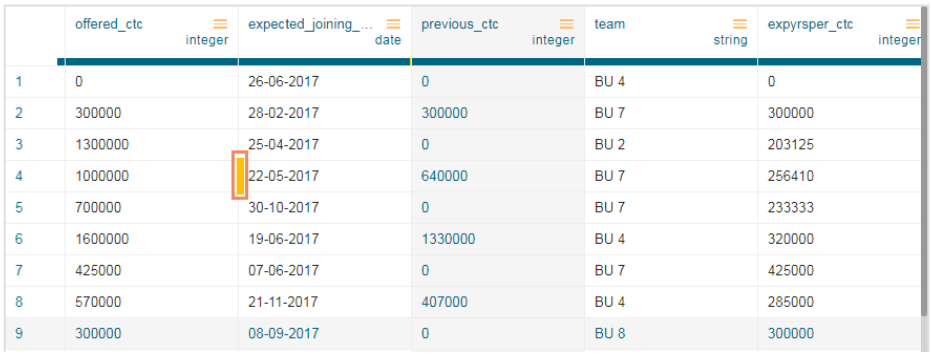
5.3. Data Quality Bar in the Grid

A Data Quality Bar appears in the header of the grid. The Data Quality is indicated through color-coding, as explained below:

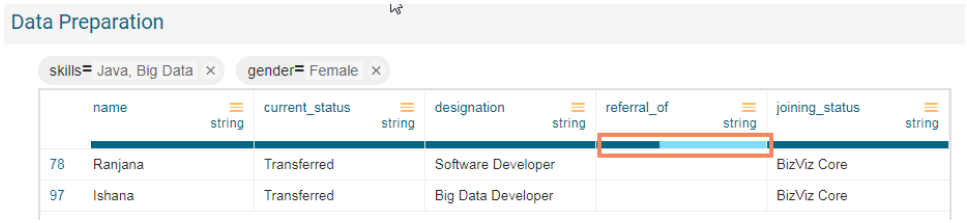
- Dark Blue-Valid Data



- Orange– Invalid data

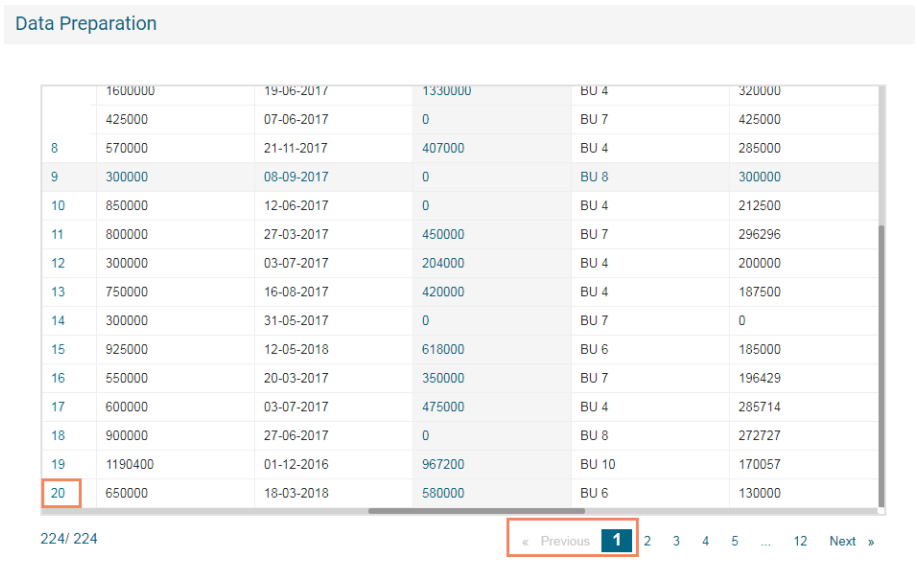


- Light Blue -Blank data



5.4. Pagination

Pagination is implemented for the grid data. The tool displays 20 records on each page.



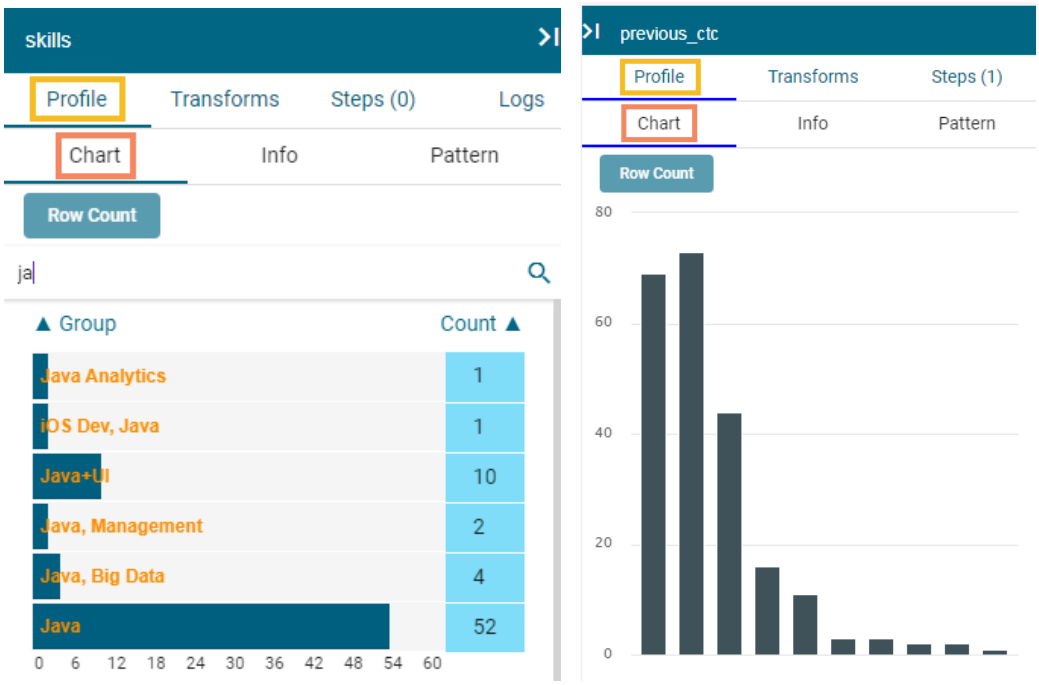
Note: The maximum rows displayed for sampling is always 10k.

6. Summary Pane

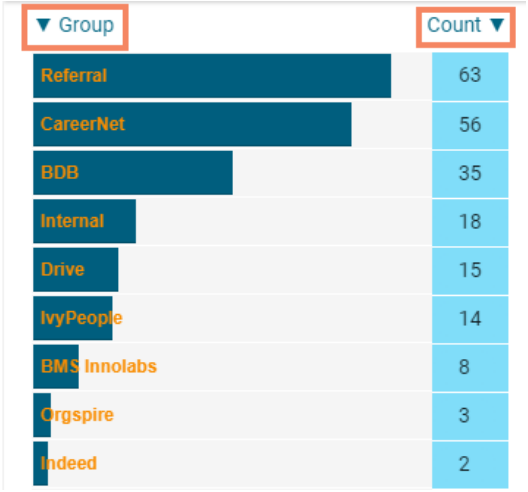
The summary pane gives an overview of the data profile like different patterns of data, distinct values, and occurrences.

6.1. Charts

The in-built charts (Column and Bar charts) display the occurrence of each value. The Bar appears to display string value. The Column chart projects numeric value columns and dates.



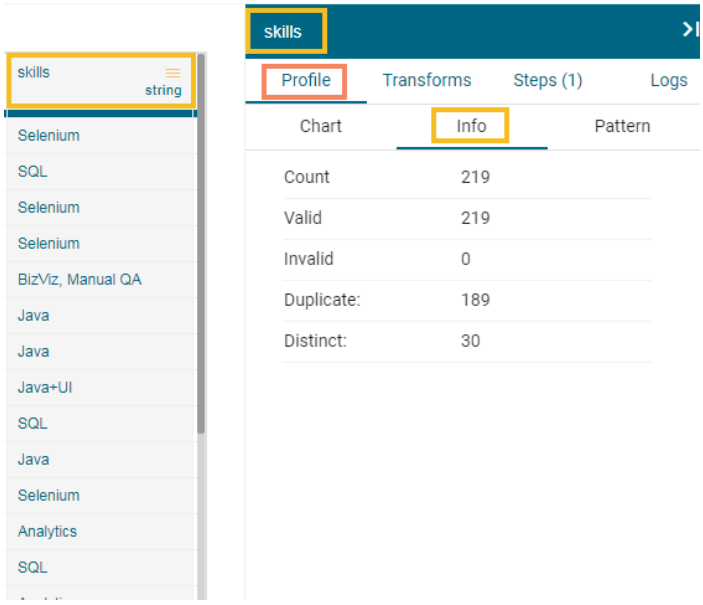
The chart can be sorted based on the group or the count of occurrence of a group. The sorted chart displays values from Ascending to Descending manner.



6.2. Info: Value/Statistics

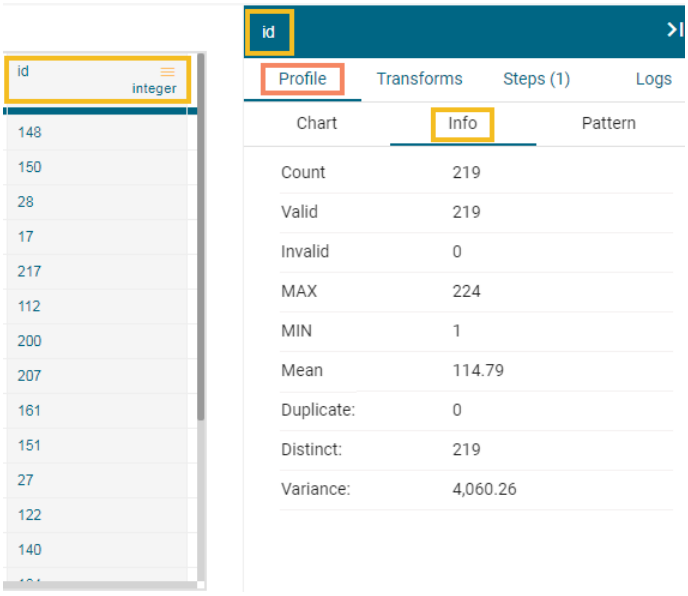
The information tab displays the value or statistics of the data. The following aspects are displayed about the chosen data when the column is of string type:

- Count: Count of Rows
- Valid: Count of Valid Data
- Invalid: Count of Invalid Data
- Duplicate: Count of Duplicates
- Distinct: Distinct Values



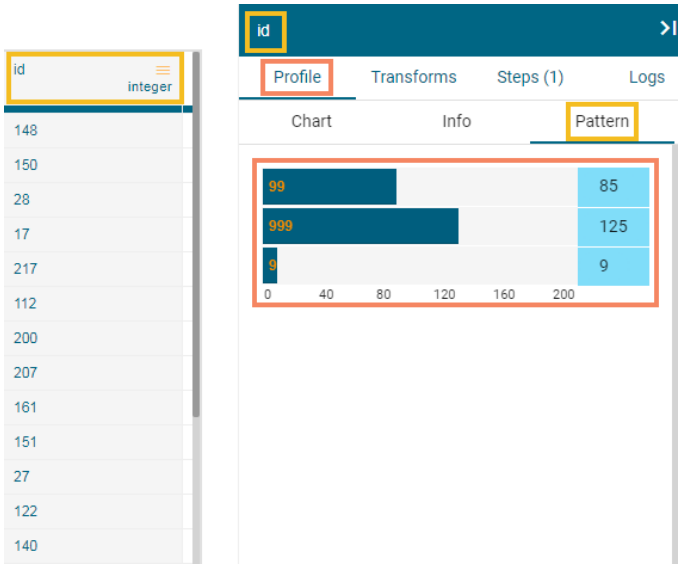
When the selected column is of numeric type, other than the above-mentioned details the additional displayed information under the '**Info**' tab is based on aggregation functions as mentioned below:

- Minimum
- Maximum
- Mean
- Variance



6.3. Pattern

This section focuses on how data patterns and occurrences of each pattern in the dataset sample get plotted in a chart for the selected column.



Note: The value displayed is not the actual value, and it's just a pattern of the value.

6.4. Transforms

Data Preparation module provides a list of transforms that can be performed on the data to clean/prepare the data for insightful visualization.

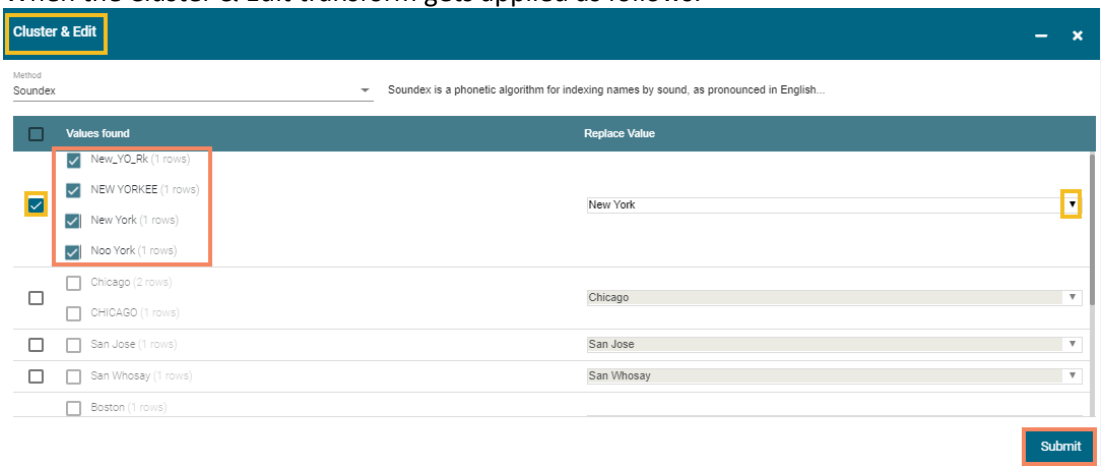
This section explains the details of the transforms.

6.4.1. Advanced

6.4.1.1. Cluster & Edit

The '**Cluster & Edit**' transform when applied groups the words with similar phonetic (Speech sound/Pronunciation) into a cluster. The user can apply this transform to replace function on that bucket to replace all those words at once. We can also exclude some value when replacing it with the new value. It works on the Soundex algorithm to cluster the data.

When the Cluster & Edit transform gets applied as follows:



The existing column '**Location**' with the following Phonetic variations as displayed below:

	Location
1	New_York_City
2	Bostn
3	Chicago
4	Nu Yorkk Sity
5	New York
6	Noo York
7	NewYorkCity
8	San Jose
9	Chicago
10	Boston
11	Nu Yorkk Sity,
12	New_YO_Rk
13	San Whosay
14	NEW YORKEE
15	Noo York City
16	CHICAGO

it gets converted into

	Location
1	New York
2	Bostn
3	Chicago
4	New York
5	New York
6	New York
7	New York
8	San Jose
9	Chicago
10	Boston
11	New York
12	New York
13	San Whosay
14	New York
15	New York
16	CHICAGO

6.4.1.2. Expression Editor

The Expression Editor transform has a collection of different functions to manipulate the data like absolute, to date, from Unix time.

i) Select the '**Expression Editor**' transform option using the '**Transforms**' tab

ii) The Expression Editor window opens with the following information:

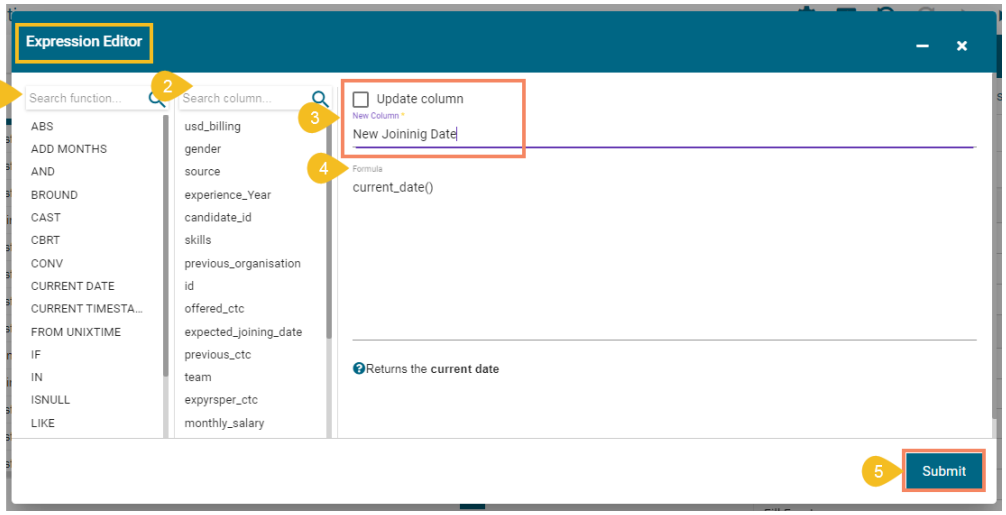
a. Search Function- Use double click to search/select a function from the displayed list. The selected function appears under the '**Formula**' space.

b. Search Column- Use double click to search/select a column from the displayed list. The selected column appears under the '**Formula**' space.

c. The user can select an existing column by enabling the '**Update column**' option or create a New Column by entering the column name for the new column. (E.g., the following image displays New Column creation)

d. The selected function and column appear under the '**Formula**' space.

e. Click the '**Submit**' option.

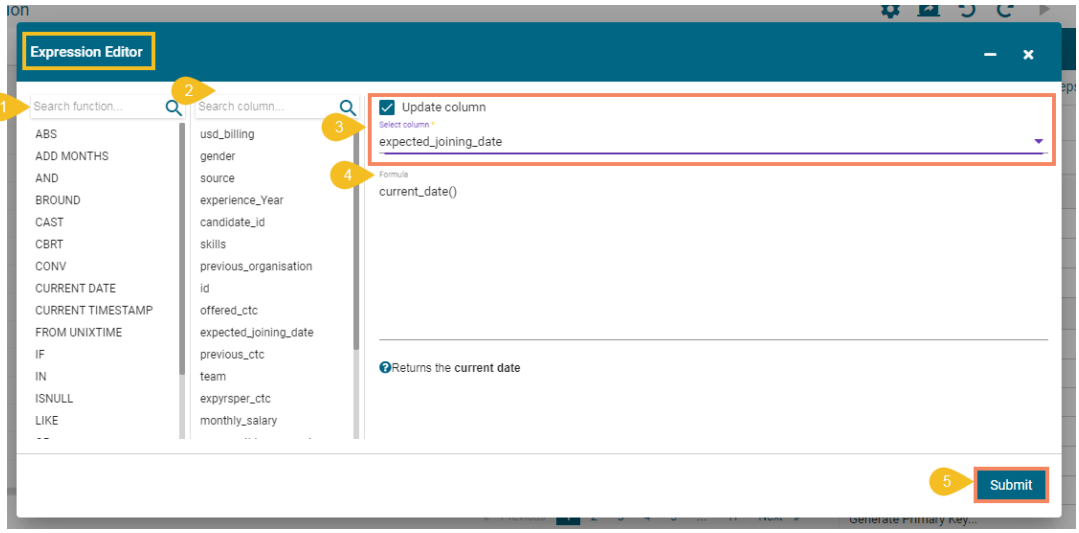


The new column gets added to the data grid with the updated data based on the applied formula.

joining_status (string)	New Joining Date (date)
Joined	2020-04-13
Joined	2020-04-13
Joined	2020-04-13
Joined	2020-04-13
BizViz Core	2020-04-13
Joined	2020-04-13
BizViz Core	2020-04-13
BizViz Core	2020-04-13
Declined	2020-04-13
Joined	2020-04-13
Joined	2020-04-13
Joined	2020-04-13
Joined	2020-04-13

Note: In the case of selecting an existing column, the data gets updated as per the applied formula in the column.

E.g., the following image displays selecting an existing column.



The original data

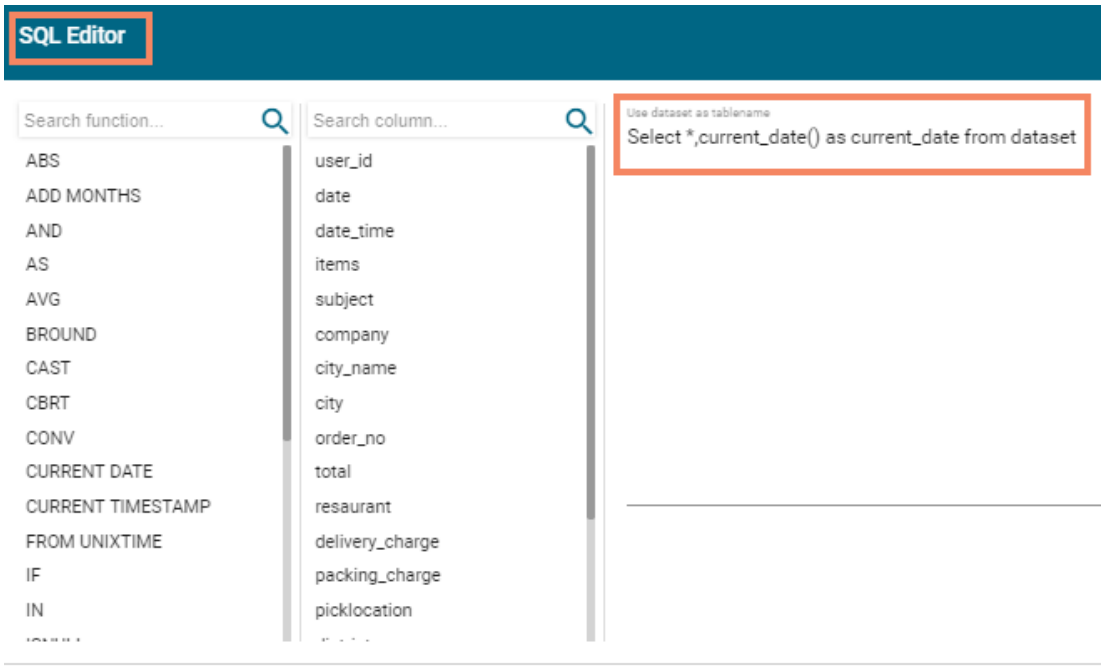
expected_joining_... (date)
24-07-2017
20-09-2017
06-03-2017
20-05-2018
01-12-2016
03-07-2017
01-12-2016
01-12-2016
14-08-2017
27-07-2017
17-04-2017
26-06-2017
07-08-2017

gets converted into

expected_joining_... (date)
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13
2020-04-13

6.4.1.3. SQL Transform

This transform allows the user to write SQL Query against the table as we can write in any SQL editor. This transform requires the table name to be mentioned as '**InputDS**' in the query.



Use syntax : "Select [*,function as new_column_name] from dataset "
Example : Select *,abs(column1) as abs_value from dataset

It adds a new column along with the existing ones in the grid. When '*' is not specified in the query It just returns the current date.

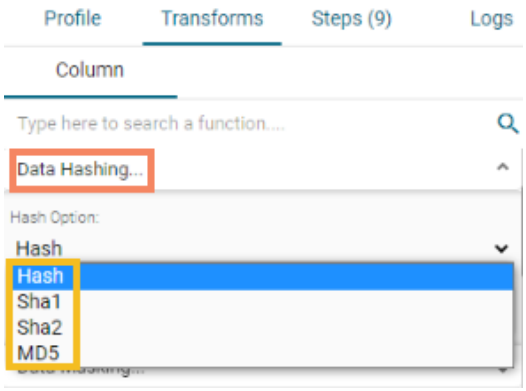
region (string)	current_date (date)
Bangalore	2019-08-16
Kochi	2019-08-16
Bhopal	2019-08-16
Hyderabad City	2019-08-16
Bangalore	2019-08-16
South Bengal	2019-08-16
Hyderabad City	2019-08-16
Kochi	2019-08-16
Chennai Region	2019-08-16
Calcutta	2019-08-16
Chennai Region	2019-08-16
Bangalore	2019-08-16
Hyderabad City	2019-08-16
Delhi	2019-08-16
Calcutta	2019-08-16
Chennai Region	2019-08-16

6.4.2. Anonymization

Anonymization is a type of information sanitization whose intent is privacy protection. It is a data processing technique that removes or modifies personally identifiable information.

6.4.2.1. Data Hashing

Data Hashing is a technique of using an algorithm to map data of any size to a fixed length. Every hash value is unique. The supported Hash options are Hash, Sha-1, Sha-2 and MD-5.



a. Select the '**Data Hashing**' transform from the '**Transforms**' tab.
b. Select a column from data grid for transformation
c. Select the required '**Hash Option**'- Hash, Sha-1, Sha-2, MD-5.
d. Click the '**Submi**t' option.

The following image displays the transformed data after the 'Data Hashing' transform is applied.

converts to

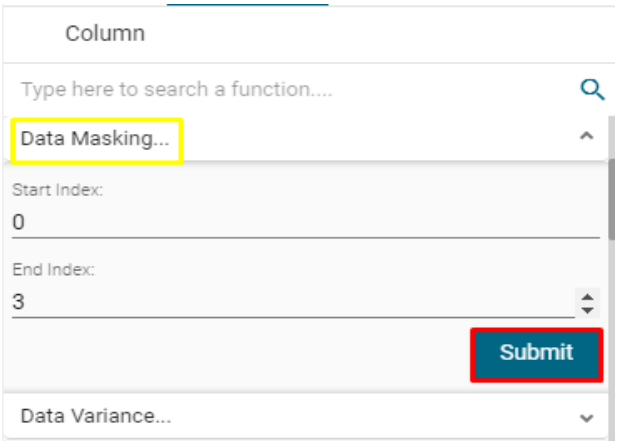
name string
fbd03a5ff8294d6a498...
738e182b34464a1230...
41bd6796e9c7c377cb...
3227457b2cb0d45c67...
a98c61ff35dd6c88660...
315238028befdf2a322...
b40e23a080b9a4d999...
8d7a9dbd35ae2b7d5f...
5ce7ba03febb5be376...
e92752e8705d7947bc...
57fe5722b9448f15170...
1329db5b386e9a5764...

6.4.2.2. Data Masking

Data masking transform is the process of hiding original data with modified content. It is a method of creating a structurally similar but inauthentic version of an actual **data.**

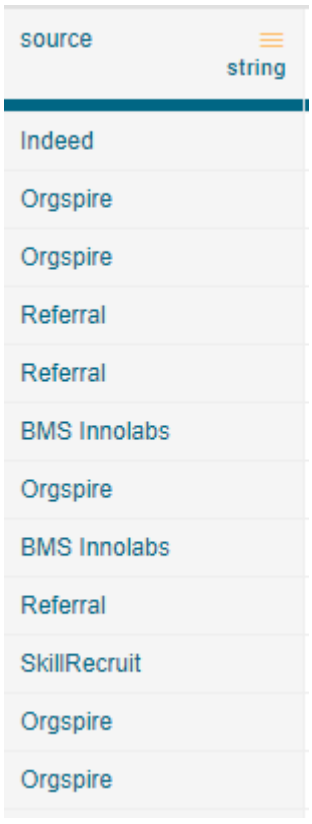
Select the '**Data Masking**' transform from the '**Transforms**' tab.

a. Select a column from data grid for transformation.



b. Select a start index and end index which you want to mask.
c. Click the '**Submit**' option.

Below is the snapshot of how the 'Data Masking' transform when applied converts the selected data:

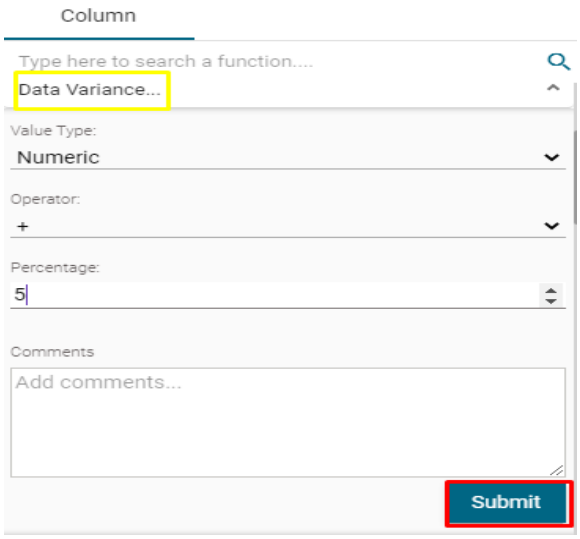


gets converted to

6.4.2.3. Data Variance

Select the '**Data Variance**' transform from the '**Transforms**' tab.

a. Select a column from data grid for transformation.
b. Select the required Value Type-Numeric/Date.
c. Select an adequate operator from the options and a Percentage value.
d. Click the '**Submit**' option.



Below is the snapshot of how the 'Data Variance' transform when applied converts the selected data:

current_status (string)	experience (integer)
Transferred	15
Resigned	10
Terminated	4
Transferred	5
Transferred	2
Declined	4
Absconded	3
Declined	3
Declined	2
Declined	2
Absconded	5
Declined	3

gets converted to

experience (double)
15.75
10.5
4.2
5.25
2.1
4.2
3.15
3.15
2.1
2.1
5.25
3.15

6.4.3. Columns

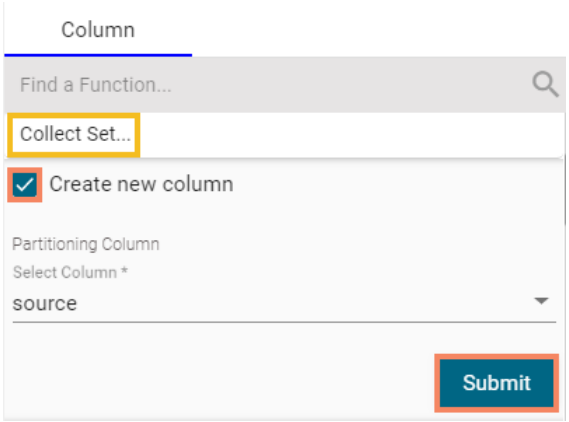
6.4.3.1. Cast to Types

It is a table-based operation. The profiling of a column is done based on the data type present in the majority. Let's say in column A; we have four integer values and one string value, then the data type of column gets profiled as the integer despite one string value present in it. The 'Cast to Types' transform removes the value with the invalid data type. In this case, it converts data with a string data type to the null value.

Note: Cast to types is a lossy transformation. There is a possibility of some data loss.

6.4.3.2. Collect Set

The '**Collect Set**' transform generates the list of all the unique values of the column based on the selected column. It performs group concatenation.
Configure the Transform and click the 'Submit' option.



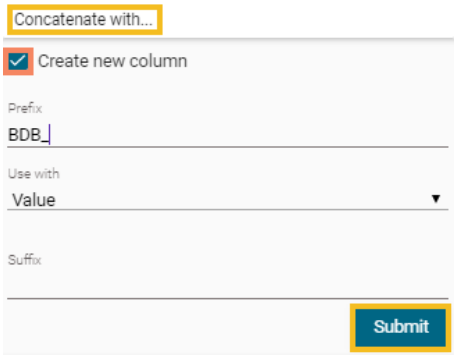
It generates a list of all unique values as displayed in the below image:

team (string)	team_set_1 (string)
BU 8	[BU 8]
BU 4	[BU 4]
BU 7	[BU 7]
BU 4	[BU 4]
BU 4	[BU 4]
BU 7	[BU 7]
BU 6	[BU 6]
BU 7	[BU 7]
BU 4	[BU 4]
BU 8	[BU 8]
BU 10	[BU 10]
BU 6	[BU 6]

6.4.3.3. Concatenate with

The users can concatenate a column value with some other column or with some prefix/suffix. To perform the transform, select the column to which data must be concatenated and select the '**concatenate with**' transform. The available options are:

a. **Prefix:** Specify the value to be prefixed to the selected column value
b. **Use with:**
   i. Select the '**Value**' to add a Prefix/Suffix
   ii. Select '**Other column**' to concatenate two columns
c. **Suffix:** Specify the value to be suffixed to the selected column value returns when performed on the selected column.



The above configuration provides 'BDB_' as the prefix for the new column, 'Candidate_id_concat_1'.

	candidate_id integer	candidate_id_conc... string
1	105	BDB_105
2	192	BDB_192
3	62	BDB_62
4	70	BDB_70
5	170	BDB_170
6	92	BDB_92
7	121	BDB_121
8	169	BDB_169
9	182	BDB_182
10	102	BDB_102
11	21	BDB_21
12	112	BDB_112
13	97	BDB_97

The users must select '**Use with Other column**' option to concatenate a value with another column and select the '**Use with Value**' option to add prefix/suffix.

6.4.3.4. Delete Column

It deletes any selected column.
To perform the transform, select the column and click on the '**Delete Column**' transform.

6.4.3.5. Duplicate Columns

The '**Duplicate Columns**' transform creates another column containing the duplicate data of the selected column.

skills (string)
Java
Java
Scripting
Java
Java
DotNet
Java
Java
Selenium
Java
Java
Java
DotNet

returns

skills_duplicate_1 (string)
Java
Java
Scripting
Java
Java
DotNet
Java
Java
Selenium
Java
Java
Java
DotNet

6.4.3.6. Fill Empty

The '**Fill Empty**' transform is used to fill the null/empty value of cell using either above or below values available in the column.

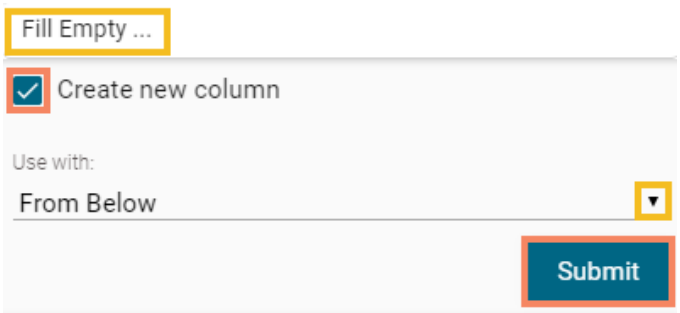
Configure the '**Fill Empty**' transform:

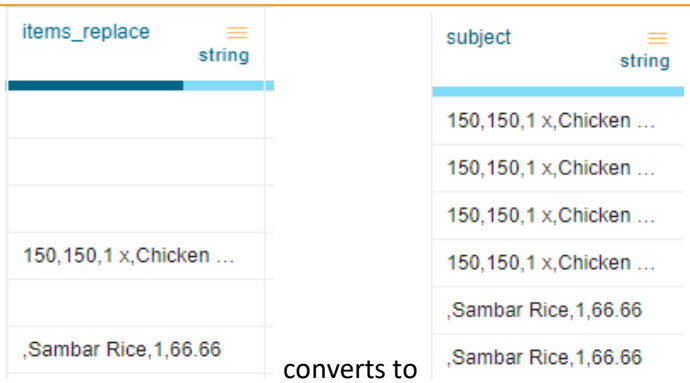
i) **Create new column**- Click the checkbox to create a new column or else the currently selected column gets updated.

ii) **Use with-**The user can use either of the options from the provided choices:

a. From Above: To fill the empty cells and replace them by the value of the cells given above the empty cells.

b. From Below: To fill the empty cells and replace them by the value of the cells given below the empty cells.





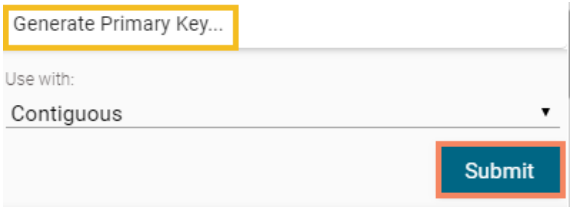
6.4.3.7. Generate Primary Key

It generates the primary key for the table. It is a table-based operation.

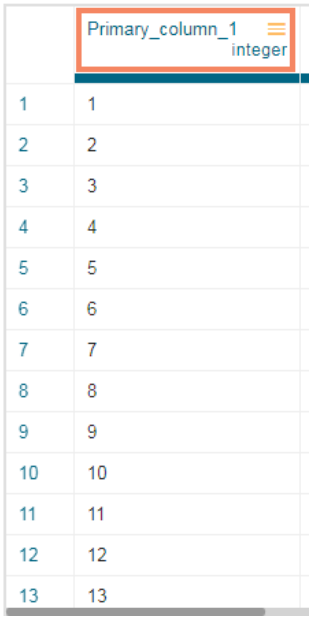
Use with: The user gets two options to generate the primary key:

i) Contiguous- it generates the auto-incremented value starting from 1.

ii) Non_contiguous- it generates a unique and random integer value.



A new column with primary values gets added to the data grid.



6.4.3.8. Get Character Length

The transform '**Get Character Length**' when applied adds a new column with numbers displaying the length of character present in that cell.

designation (string)	designation_length_1 (integer)
QA Manager	10
QA Architect	12
Senior Software Engin...	24
QA Engineer	11
QA Engineer	11
Senior Software Engin...	24
AWS Consultant	14
Senior Software Engin...	24
QA Engineer	11
Business Analyst	16
Senior QA Engineer	18
QA Engineer	11
Senior QA Engineer	18

The empty cells are kept as it is in the column.

items (string)
1,Special Hyderabadi ...
150,150,1 x,Chicken ...
,Sambar Rice,1,66.66
170,85,2,Masala Dosa

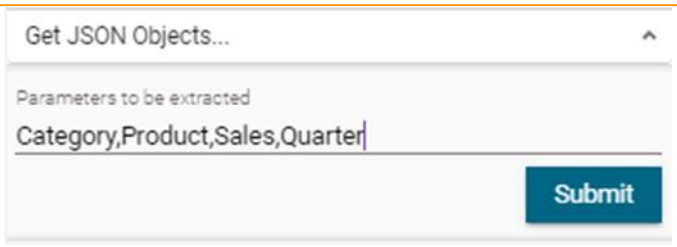
it returns

items_length_1 (integer)
49
26
20
20

6.4.3.9. Get JSON Objects

The '**Get JSON Objects**' transform extracts any parameter from a given column with JSON data.

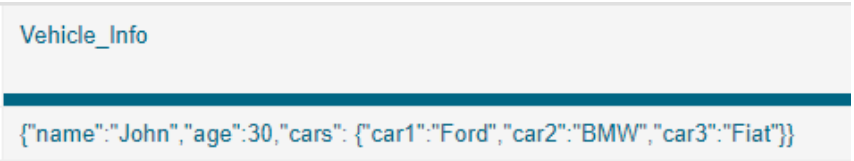
Sales_detials
{"Category":"Beverages","Product":"Chai","Sales":"705.6","Quarter":"Qtr 1"}
{"Category":"Beverages","Product":"Chai","Sales":"878.4","Quarter":"Qtr 2"}
{"Category":"Beverages","Product":"Chai","Sales":"1174.5","Quarter":"Qtr 3"}
{"Category":"Beverages","Product":"Chai","Sales":"2128.5","Quarter":"Qtr 4"}
{"Category":"Beverages","Product":"Ipoh Coffee","Sales":"1398.4","Quarter":"Qtr 1"}
{"Category":"Beverages","Product":"Ipoh Coffee","Sales":"4496.5","Quarter":"Qtr 2"}
{"Category":"Beverages","Product":"Ipoh Coffee","Sales":"1196","Quarter":"Qtr 3"}
{"Category":"Beverages","Product":"Ipoh Coffee","Sales":"3979","Quarter":"Qtr 4"}
{"Category":"Beverages","Product":"Juices","Sales":"1141.92","Quarter":"Qtr 1"}
{"Category":"Beverages","Product":"Juices","Sales":"1774.08","Quarter":"Qtr 2"}
{"Category":"Beverages","Product":"Juices","Sales":"3261.6","Quarter":"Qtr 3"}
{"Category":"Beverages","Product":"Juices","Sales":"1705.5","Quarter":"Qtr 4"}
{"Category":"Condiments","Product":"Aniseed Syrup","Sales":"544","Quarter":"Qtr 1"}
{"Category":"Condiments","Product":"Aniseed Syrup","Sales":"600","Quarter":"Qtr 2"}
{"Category":"Condiments","Product":"Aniseed Syrup","Sales":"140","Quarter":"Qtr 3"}
{"Category":"Condiments","Product":"Aniseed Syrup","Sales":"440","Quarter":"Qtr 4"}



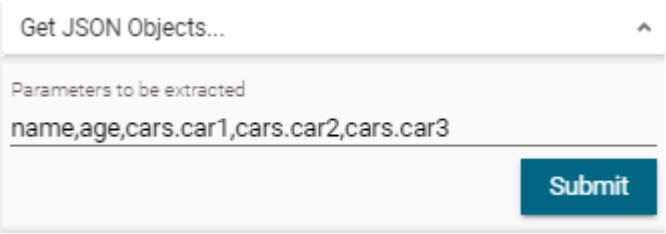
When the parameter in the JSON is specified, the transform extracts all parameter values as columns into a tabular format.

Sales_detials (string)	Sales_detials_Cat... (string)	Sales_detials_Pro... (string)	Sales_detials_Sales (double)	Sales_detials_Qua... (string)
{"Category":"Beverages","Produ...	Beverages	Chai	705.6	Qtr 1
{"Category":"Beverages","Produ...	Beverages	Chai	878.4	Qtr 2
{"Category":"Beverages","Produ...	Beverages	Chai	1174.5	Qtr 3
{"Category":"Beverages","Produ...	Beverages	Chai	2128.5	Qtr 4
{"Category":"Beverages","Produ...	Beverages	Ipoh Coffee	1398.4	Qtr 1
{"Category":"Beverages","Produ...	Beverages	Ipoh Coffee	4496.5	Qtr 2
{"Category":"Beverages","Produ...	Beverages	Ipoh Coffee	1196	Qtr 3
{"Category":"Beverages","Produ...	Beverages	Ipoh Coffee	3979	Qtr 4
{"Category":"Beverages","Produ...	Beverages	Juices	1141.92	Qtr 1
{"Category":"Beverages","Produ...	Beverages	Juices	1774.08	Qtr 2
{"Category":"Beverages","Produ...	Beverages	Juices	3261.6	Qtr 3
{"Category":"Beverages","Produ...	Beverages	Juices	1705.5	Qtr 4
{"Category":"Condiments","Prod...	Condiments	Aniseed Syrup	544	Qtr 1
{"Category":"Condiments","Prod...	Condiments	Aniseed Syrup	600	Qtr 2
{"Category":"Condiments","Prod...	Condiments	Aniseed Syrup	140	Qtr 3

To extract the nested Json the '.' mark can be used to specify the path.



The above data can be extracted using the transform by specifying the parameters as given below



The data is converted to a tabular structure as below

Vehicle_Info (string)	Vehicle_Info_name (string)	Vehicle_Info_age (integer)	Vehicle_Info_cars_... (string)	Vehicle_Info_cars_... (string)	Vehicle_Info_cars_... (string)
{"name":"John","age":30,"...	John	30	Ford	BMW	Fiat

6.4.3.10. Pivot

When applied, the '**Pivot**' transform converts the data into a Pivot table based on the selected Pivot

Column and Group of the selected columns.

Sample Data

Category (string)	Product (string)	Sales (double)	Quarter (string)
Condiments	Hot Pepper Sauce	1347.36	Qtr 1
Beverages	Ipoh Coffee	3979.00	Qtr 4
Beverages	Ipoh Coffee	1196.00	Qtr 3
Dairy Products	Yogurt	242.50	Qtr 2
Condiments	Grandma's Boysenberry Spread	1750.00	Qtr 3
Condiments	Hot Pepper Sauce	2150.77	Qtr 2
Condiments	Aniseed Syrup	440.00	Qtr 4
Confections	Chocolade	162.56	Qtr 2
Condiments	Northwoods Cranberry Sauce	1300.00	Qtr 2
Condiments	Hot Pepper Sauce	3857.41	Qtr 4
Beverages	Juices	1141.92	Qtr 1
Condiments	Aniseed Syrup	544.00	Qtr 1
Dairy Products	Yogurt	150.00	Qtr 4
Dairy Products	Mik	3329.28	Qtr 1

In the given example, the 'Quarter' is selected as the Pivot Column, and Sum aggregation is selected for the Sales column. The Selected group columns are Category and Product.

Transform Selection:

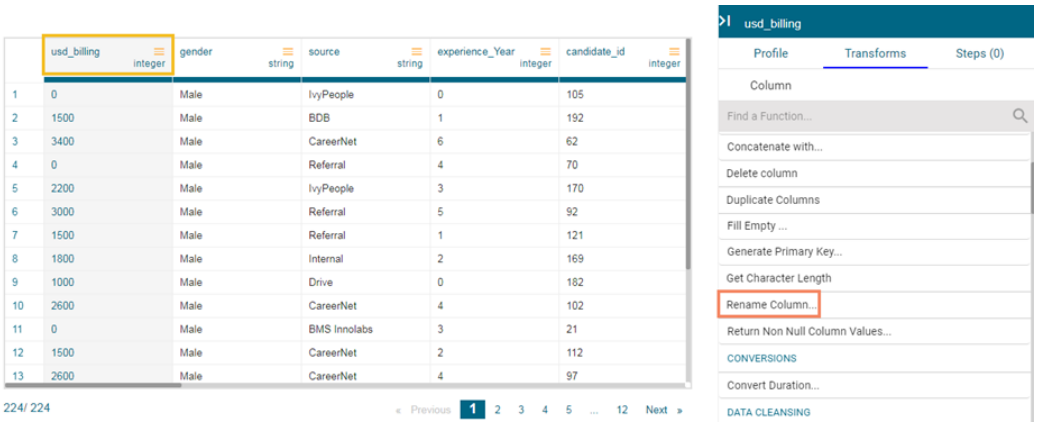
Thus, it returns the sum of Sales is aggregated based on Category and Product columns for each Quarter.

Category (string)	Product (string)	Qtr 1 (double)	Qtr 2 (double)	Qtr 3 (double)	Qtr 4
Beverages	Chai	705.6	878.4	1174.5	2128.5
Beverages	Ipoh Coffee	1398.4	4496.5	1196.0	3979.0
Condiments	Aniseed Syrup	544.0	600.0	140.0	440.0
Dairy Products	Cheese	4454.8	174.15	2541.29	2472.5
Condiments	Northwoods Cranberry Sa...		1300.0		2960.0
Dairy Products	Probiotics	487.0	2993.12	1458.75	2681.87
Condiments	Grandma's Boysenberry S...			1750.0	750.0
Dairy Products	Yogurt	294.0	242.5	99.5	150.0
Confections	Chocolade	744.6	162.56	68.85	306.0
Dairy Products	Mik	3329.28	3989.9	10273.1	3060.0
Beverages	Juices	1141.92	1774.08	3261.6	1705.5
Condiments	Hot Pepper Sauce	1347.36	2150.77	1975.54	3857.41
Condiments	Vegie-spread	3202.87	263.4	842.88	2590.1
Condiments	Gula Malacca	1994.85	1753.62	1093.09	1701.87

6.4.3.11. Rename Column

The Rename Column transform allows the user to rename the selected column.

a. Select a column from the data grid that needs to be renamed.
b. Choose the '**Rename Column**' transform from the **Transforms** tab.



c. The Rename Column dialog box opens. Provide a name that you wish to use as a rename for the selected column.
d. Click the '**Submit**' option.

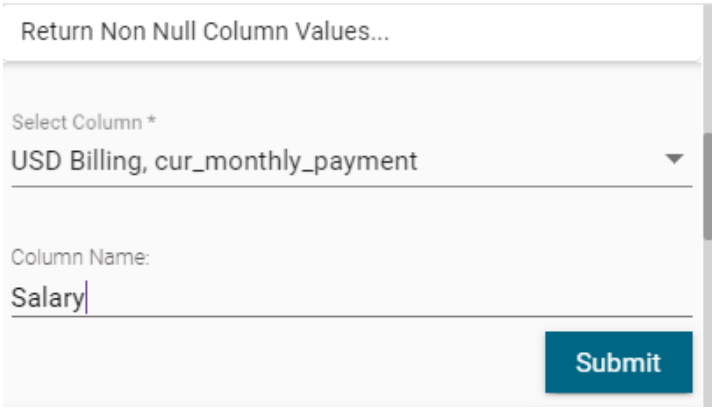
e. The column gets renamed.

	USD Billing integer	gender string
1	0	Male
2	1500	Male
3	3400	Male
4	0	Male
5	2200	Male
6	3000	Male
7	1500	Male
8	1800	Male
9	1000	Male
10	2600	Male
11	0	Male
12	1500	Male
13	2600	Male

6.4.3.12. Return Non-Null Column Values

The transform returns the first non-null value from the list of columns specified to a new column. To perform the transform, select the columns which must be checked for null and specify a column name for the result.

a. **Select Column:** Select the columns to be checked for null
b. **Column name:** The name for the new result column returns



	usd_billing integer	cur_monthly_paym... integer
5	1750	43333
6	1750	
7	2300	
8	2000	
9	2000	
10		52000
11		52000
12		52000
13		43333
14	4000	141666
15		0
16	1800	70833
17	1750	38333

returns the new result column

Salary integer
1750
1750
2300
2000
2000
52000
52000
52000
43333
4000
0
1800
1750

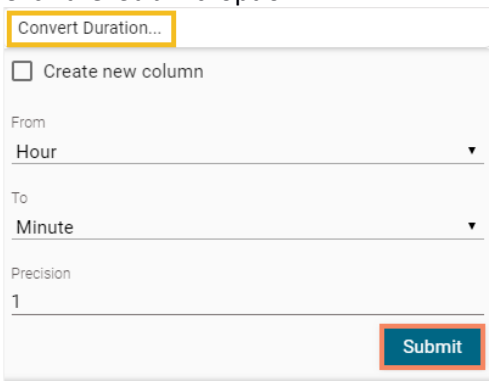
6.4.4. Conversions

6.4.4.1. Convert Duration

The transform converts any duration (day, hour, minute, seconds, milliseconds) to any specified duration.

To perform the transform, select the column which has the duration to be converted and specify the duration type.

a. **From:** The type of source interval
b. **To:** The type of destination interval
c. **Precision:** The decimal points to be retained
d. Click the **'Submit'** option.



Below is the snapshot of how the 'Convert Duration' transform when applied converts the selected data:

Duration_hrs	double
2.8	
3.6	
5.4	
6.2	
7.4	
9.1	
4.4	
6.7	
8.1	
4.5	
9.2	

converts to

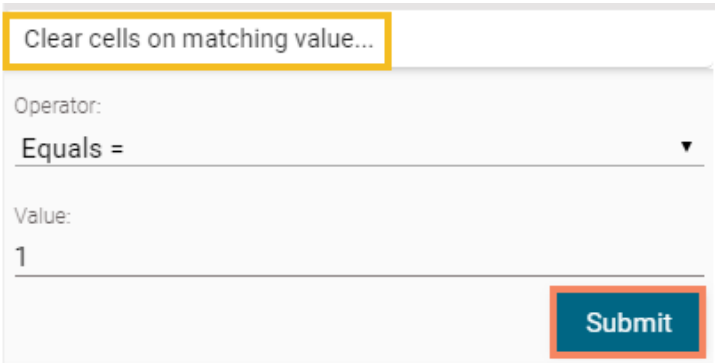
Duration_hrs	double
168.0	
216.0	
324.0	
372.0	
444.0	
546.0	
264.0	
402.0	
486.0	
270.0	
552.0	

6.4.5. Data Cleansing

6.4.5.1. Clear Cells on Matching Value

Clear the cell value on matching the condition specified. Operators include contains, equals, starts with, end with, and regex match. Transform applies in the same column.

- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column



The value selected in the form clears the cell with 1 in the selected column.

returns data like this


6.4.5.2. Delete Rows on Matching Value

Delete the rows on matching the condition specified for that column. Operators include contains, equals, starts with, ends with, and regex match.

- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column

The value selected in the form deletes the row with any numbers from 0-9 in the selected column.

turns to


when the above transform is applied.

6.4.5.3. Delete Rows with Empty Cell

a. The transform deletes any row which has a blank value in the selected column. The transform does not have a form.

Team string	Designation string	Referral_of string
B3	QA Lead	
B4	Software Eng.	
B6	Sr. Sofware Eng.	EMP 9
B8	Sr. Software Eng.	
B4	QA Eng.	
B1	Project Manager1	
B4	Executive Manager	EMP5
B5	BI Lead	
B6	Sr. Sofware Eng.	
B1	Project Manager2	
B8	Sr. Sofware Eng.	

b. When we perform the transform on column “referral_of” it deletes all the rows which have an empty value in that column returning the data as below:

Team string	Designation string	Referral_of string
B6	Sr. Sofware Eng.	EMP 9
B4	Executive Manager	EMP5

6.4.5.4. Delete Rows with Invalid Cell

a. The transform deletes any row which has an invalid value in the selected column. The transform does not have form.

b. When we do the transform on the ‘gender’ column, it deletes all rows marked invalid as displayed below:

Gender	string
male	
female	
male	
female	
female	
1	
1	
female	
0	
1	
male	

returns

Gender	string
male	
female	
male	
female	
female	
female	
male	

6.4.5.5. Delete Rows with Negative Values

1. It deletes the rows which have a negative value in the selected column. This transform does not have a form.
2. When this transform is applied to experience column, it deletes all rows with negative, as displayed below:

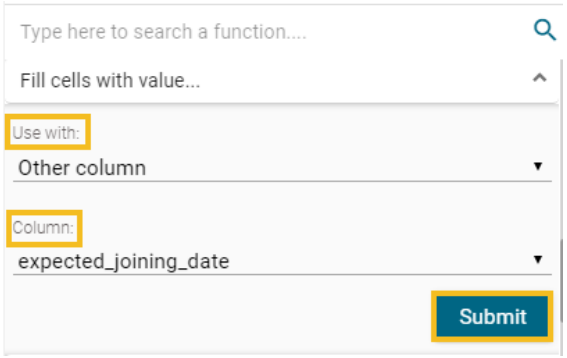
Referral_of	string	Experience	integer
		4	
		6	
EMP 9		6	
		7	
		-1	
		8	
EMP5		5	
		5	
		6	
		9	
		6	

3. It deletes the row with the negative value and returns the data as displayed below:

Referral_of string	Experience integer
	4
	6
EMP 9	6
	7
	8
EMP5	5
	5
	6
	9
	6

6.4.5.6. Fill Cells with Value

It fills the selected column with a value or a value from another column.



- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be filled, or the column with which the value must be replaced

When the above transform is applied to the below data on the column 'created_date,' it copies the value from the 'expected_joining_date' column to the 'created_date' column.

expected_joining_... date	created_date string
24-07-2017	
03-07-2017	
14-08-2017	
08-09-2017	
15-06-2017	
21-08-2017	
10-07-2017	
04-05-2017	
10-04-2017	
01-12-2016	
19-06-2017	
01-12-2016	
20-11-2017	

converts into

expected_joining_... date	created_date date
24-07-2017	24-07-2017
03-07-2017	03-07-2017
14-08-2017	14-08-2017
08-09-2017	08-09-2017
15-06-2017	15-06-2017
21-08-2017	21-08-2017
10-07-2017	10-07-2017
04-05-2017	04-05-2017
10-04-2017	10-04-2017
01-12-2016	01-12-2016
19-06-2017	19-06-2017
01-12-2016	01-12-2016
20-11-2017	20-11-2017

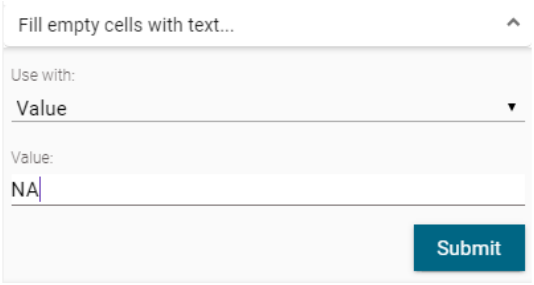
Note: The user can also fill the column with a specific value if the selected option is 'Value'. E.g., the following image mentions 30 as the selected value for the 'created_date' column.

the column displays

created_date integer
30
30
30
30
30
30
30
30
30
30
30
30

6.4.5.7. Fill Empty Cells with Text

It helps to fill the empty cells of a selected column with a value or a value from another column if the destination column is empty.



- **Use with:** Specify whether to fill with a value or another column value.
- **Column/ Value:** The value with which the column must be filled, or the column with which the value must be replaced.

When the transform is applied to the below data on column 'referral_of,' it fills the value '**NA**' for all the empty cells of that column.

designation string	referral_of string
Software Engineer	
Software Engineer	
Lead Software Engineer	
Senior Software Engin...	Mahendra
Senior Software Engin...	
Senior Software Engin...	Manish Jaiswal
Associate Software En...	Ritesh
Software Engineer	Tripura
Associate QA Engineer	
Senior Software Engin...	
Senior Software Engin...	
Associate Software En...	
Senior Software Engin...	

converts to

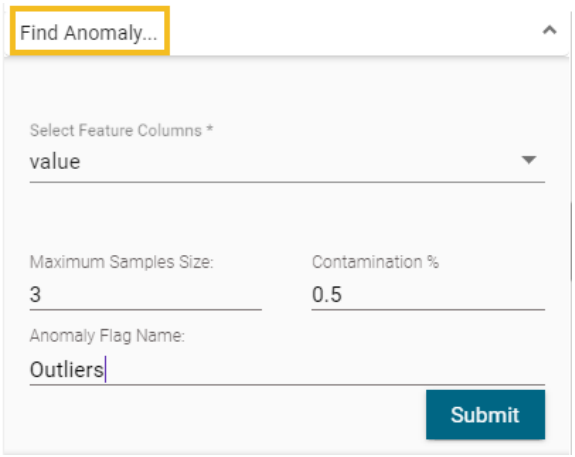
designation string	referral_of string
Software Engineer	NA
Software Engineer	NA
Lead Software Engineer	NA
Senior Software Engin...	Mahendra
Senior Software Engin...	NA
Senior Software Engin...	Manish Jaiswal
Associate Software En...	Ritesh
Software Engineer	Tripura
Associate QA Engineer	NA
Senior Software Engin...	NA
Senior Software Engin...	NA
Associate Software En...	NA
Senior Software Engin...	NA

6.4.5.8. Find Anomaly

Anomaly detection is used to identify any anomaly present in the data. i.e., Outlier. Instead of looking for usual points in the data, it looks for any anomaly. It uses the **Isolation Forest** algorithm.

The '**Find Anomaly**' transform takes four parameters:

1. **Select Feature Columns:** We can select one or more columns where we want to find the anomaly.
2. **Maximum Sample Size:** Isolation forest takes the training data of a given sample size to find out the normal value in the dataset. The sample size can vary from 1 to 250 (both inclusive).
3. **Contamination (%):** It is the percentage of observations we believe to be outliers. It varies from 0 to 1 (both inclusive).
4. **Anomaly Flag Name:** The result is either 0 or 1. 0 means the data is standard, and 1 means data is an outlier. This information gets stored in the new column given in the anomaly flag name.
5. Click the '**Submit**' option to detect anomaly from the selected data.

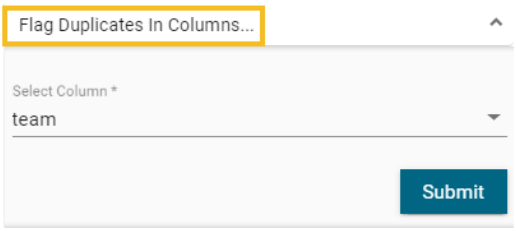


The anomaly gets stored in the new column under the anomaly flag name (In this case, it is displayed under the '**outlier**' column).

value integer	outlier double
1	0.0
2	0.0
3	0.0
4	0.0
21	1.0
6	0.0
1000	1.0
1200	1.0
1000	1.0
	1.0
	1.0
	1.0
	1.0

6.4.5.9. Flag Duplicates in Columns

This transform adds a new Boolean column based on duplicate values in the column. For original value it gives false, and for the duplicate value, it provides true value.



team (string)
BU 4
BU 7
BU 2
BU 7
BU 7
BU 4
BU 7
BU 4
BU 8
BU 4
BU 7
BU 4

returns

IsDuplicate_team (boolean)
false
false
false
true
true
true
true
true
false
true
true
true

6.4.5.10. Flag Duplicates in Tables

This transform adds a new Boolean column based on duplicate rows in the table. For original value it gives false, and for the duplicate value, it provides true value.

6.4.5.11. Remove Duplicates from Column

It removes duplicate values from the selected columns. This transform can be performed on a single as well as on multiple columns.

team (string)
BU 4
BU 7
BU 2
BU 7
BU 7
BU 4
BU 7
BU 4
BU 8

converts to

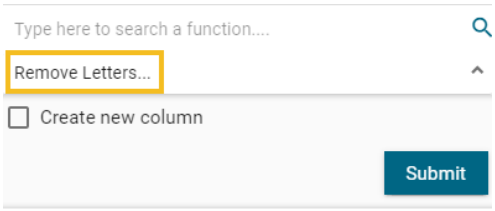
team (string)
BU 4
BU 7
BU 2
BU 8
BU 6
BU 10
BU 5
BU 1
BU 9

6.4.5.12. Remove Duplicates from Table

It removes all duplicate rows from the table.

6.4.5.13. Remove Letters

It removes any letter present in the selected column. The users can either add a new column with the transformed value or overwrite the same column.



The selected column

Employee ID string
EMP ID 1
EMP ID 2
EMP ID 3
EMP ID 4
EMP ID 5
EMP ID 6
EMP ID 7
EMP ID 8
EMP ID 9
EMP ID 10
EMP ID 11

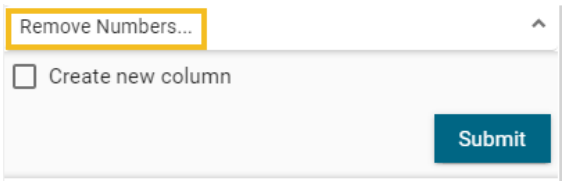
converts into

Employee ID string
1
2
3
4
5
6
7
8
9
10
11

after transformation.

6.4.5.14. Remove Numbers

It removes any number present in the selected column. We can either add a new column with the transformed value or overwrite the same column.



When the '**Remove Numbers**', transform gets performed on a selected column,

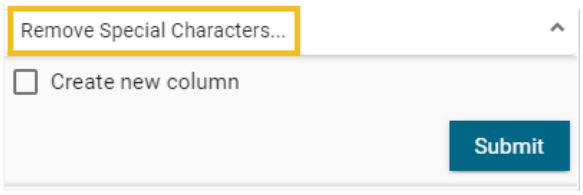
Qualification string
BE 1
BE 2
BE 3
BE 4
BE 5
ME 6
MTech 7
BTech 8
BE 9
BTech 10
MTech 11

it removes numbers from the selected column

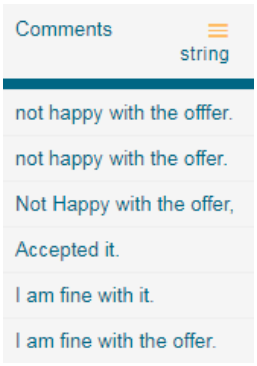
Qualification string
BE
BE
BE
BE
BE
ME
MTech
BTech
BE
BTech
MTech

6.4.5.15. Remove Special Characters

It removes any special character present in the selected column. Only letters, numbers, and spaces are retained. We can either add a new column with the transformed value or overwrite the same column.



When the transform '**Remove Special Characters**' gets performed on the selected column,



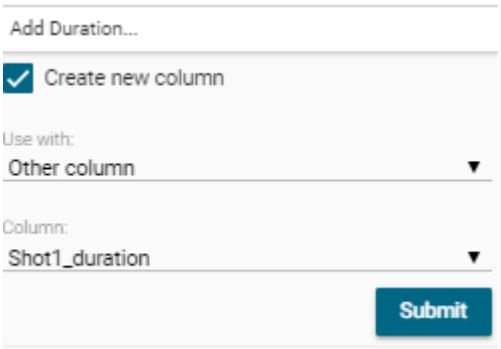
the punctuations get removed from the column

6.4.6. Dates

6.4.6.1. Add Duration

The transform adds two-time values. It can either add the selected column with a time value or time from another column. The transform supports adding time into '**hh:mm:ss.mmm'** and '**hh:mm:ss'** formats.

- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be added, or the column with which the selected column value must be added.



The transform when performed on the data selecting 'Shot1_duration', it adds

Shot1_duration and Shot2_duration and gives a new column with the result.

Shot 1 string	Shot 2 string
00:00.0	00:00.0
00:00.0	00:00.0
00:00.0	00:00.0
00:01.0	00:01.0
00:02.0	00:00.0
00:03.1	00:00.0
00:00.0	00:00.0
00:00.0	00:02.0
00:01.0	00:02.0
00:02.1	00:00.0

converts to

Shot 1 string	Shot 2 string	Shot 2_addtime_1 string
00:00.0	00:00.0	00:00:00.000
00:00.0	00:00.0	00:00:00.000
00:00.0	00:00.0	00:00:00.000
00:01.0	00:01.0	00:02:00.000
00:02.0	00:00.0	00:02:00.000
00:03.1	00:00.0	00:03:00.001
00:00.0	00:00.0	00:00:00.000
00:00.0	00:02.0	00:02:00.000
00:01.0	00:02.0	00:03:00.000
00:02.1	00:00.0	00:02:00.001

6.4.6.2. Add Interval to Date

It adds the time duration specified to the selected datetime column.

- **Input Format:** It is used to specify the format of the selected Date column format. It can have values 'Year first', 'Month first', and 'Day first.'
- **Value Type:** It specifies the type of duration which acts as the operand for the addition. The value type can be years, months, days, weeks, hours, minutes or milliseconds
- **Value:** The value or the operand that must be added with the selected column

Note: The transform supports the datetime column of '**yyyy-mm-dd'** into the '**hh:mm:ss'** format.

6.4.6.3. Extract Time

Extract the time units from a selected column with a time value. The time units that get extracted include hours, minutes, seconds, milliseconds, and time to milliseconds.

- **Hours:** Extracts hours from a time
- **Minutes:** Extracts minutes from a time
- **Seconds**: Extracts seconds from a time
- **MilliSeconds:** Extracts milliseconds from a time
- **Time to MilliSeconds:** Converts the time given to milliseconds

Note : The transform supports time format like- hh:mm:ss:mmm, hh:mm:ss, hh:mm

6.4.6.4. Extract Date

It extracts the date part from a selected column with a date value.
The date parts that can be extracted include day, month, year, the day of the week, the day of the year and the week of the year.

- **Day:** It extracts day from a date
- **Month:** It extracts the month from a date/datetime. We can specify the pattern in which the month value has to be returned. Month pattern can be 0-12, Jan - Dec or January - December
- **Year**: It extracts the year from a date. We can specify the pattern in which the year has to be returned. The year pattern can be in the 'yy' or 'yyyy' format.

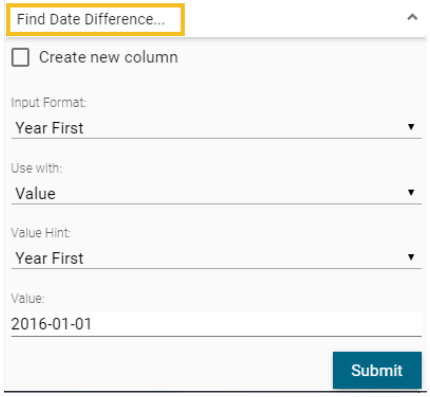
- **Day of Week:** It returns the day of the week for the selected date. Day of week pattern can also be specified. The pattern can be 1-7, Sun-Sat or Sunday-Saturday
- **Day of Year:** It returns a number between 1 and 365, which indicates the sequential day number starting with day one on January 1st.
- **Week of Year:** It replaces a number between 1 and 53, which indicates the sequential week number beginning with 1 for the week January 1st falls.

Note: The transform supports Date and DateTime format (date hh:mm:ss)

6.4.6.5. Find Date Difference

The transform finds the difference between two date values. It can either subtract the selected column with a date value or date from another column. The transformed value can replace the existing column value or can be added as a new column.

- **Input Format**: Specifies the format of the given date column
- **Use with**: Specify whether to fill with a value or another column value
- **Value Hint**: Specifies format of value from which we want to find the difference
- **Value**: Pass the date value from where you want to find the date difference



This transform gives the number of days by finding out the difference between the given date and value/date column which we have used.
Here value used is: 2016-01-01

expected_joining_... date
2017-05-22
2017-06-19
2017-07-06
2017-11-21
2017-06-27
2018-03-18
2017-06-03
2017-10-08
2017-06-26
2017-09-10
2017-06-26
2017-08-30
2017-09-08

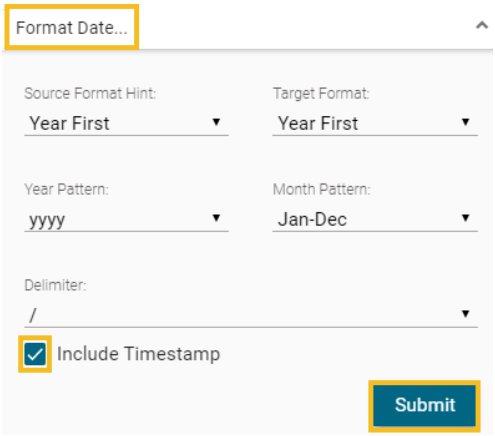
converts to

expected_joining_... integer
507
535
552
690
543
807
519
646
542
618
542
607
616

6.4.6.6. Format Date

The users can change the format of a date column by using this transform.

- **Source Format Hint:** Specifies the current format of the date column.
- **Target Format:** Specifies what we want first(Year, Month, Day) in our output format of the date column
- **Year Pattern:** Specifies the format of the year (yyyy or yy) in the output date column.
- **Month Pattern:** It specifies the format of the month (number, Jan-Dec, January-December) in the output date column.
- **Delimiter:** Specifies Delimiter(like- slash, a hyphen, comma, full stop, space) for the output date column.
- **Include Timestamp:** It adds a timestamp to the current date format if enabled with a tick mark.



expected_joining_... date
2017-05-22
2017-06-19
2017-07-06
2017-11-21
2017-06-27
2018-03-18
2017-06-03
2017-10-08
2017-06-26
2017-09-10
2017-06-26
2017-08-30
2017-09-08

converts to

expected_joining_... timestamp
2017/May/22 00:00:00
2017/Jun/19 00:00:00
2017/Jul/06 00:00:00
2017/Nov/21 00:00:00
2017/Jun/27 00:00:00
2018/Mar/18 00:00:00
2017/Jun/03 00:00:00
2017/Oct/08 00:00:00
2017/Jun/26 00:00:00
2017/Sep/10 00:00:00
2017/Jun/26 00:00:00
2017/Aug/30 00:00:00
2017/Sep/08 00:00:00

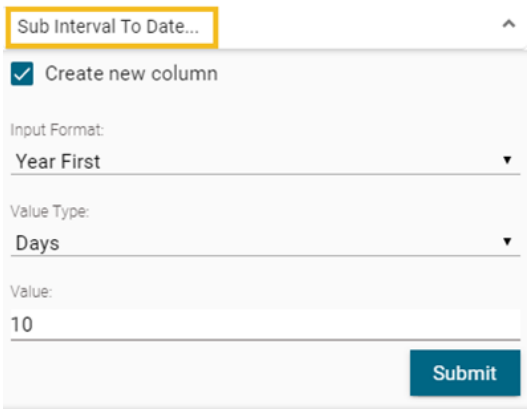
6.4.6.7. From Unix Time

The '**From Unix Time**' transform converts the Unix time into a specified format (1349862300 – 2012 10-10 both date and datetime).

6.4.6.8. Sub Interval to Date

The '**Sub Interval to Date**' transform subtracts specified value(interval) from the given date column. The transformed value can replace the existing column value or can be added as a new column.

- **Input Format**- Format of date column(given) should be specified here.
- **Value Type**-specifies what we want to subtract like years, months, days, weeks, etc.
- **Value**- specifies how many years/months/days(value type) we want to subtract.



This transform when performed subtracts four months from the date column and gives this new column having the date which is 10 days back from the given date.

expected_joining_... date
2017-05-22
2017-06-19
2017-07-06
2017-11-21
2017-06-27
2018-03-18
2017-06-03
2017-10-08
2017-06-26
2017-09-10
2017-06-26
2017-08-30
2017-09-08

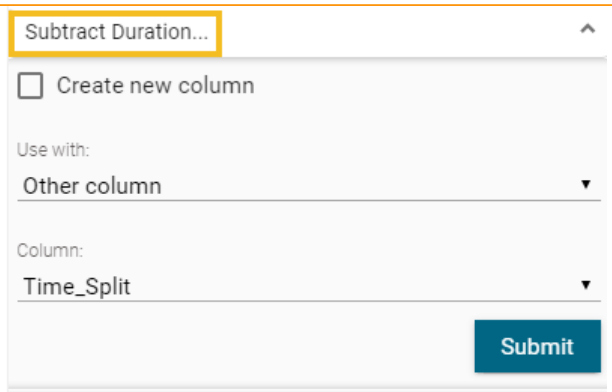
converts to

expected_joining_... date
2017-05-12
2017-06-09
2017-06-26
2017-11-11
2017-06-17
2018-03-08
2017-05-24
2017-09-28
2017-06-16
2017-08-31
2017-06-16
2017-08-20
2017-08-29

6.4.6.9. Subtract Duration

The transform 'Subtract Duration' deducts the time values in two ways. It can either subtract the selected column with a time value or time from another column. The transform supports subtracting time into '**hh:mm:ss.mmm' ,'hh:mm:ss'** and **'hh:mm'** formats. The transformed value can replace the existing column value or can be added as a new column.

- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be subtracted, or the column with which the selected column value must be subtracted.



This transform when performed on Time1_split1 for subtracting 01:00:00 from this column provides a new column having values after deducting 01:00:00.

Time_Split string
01:00:00
02:00:00
03:00:00
04:00:00
05:00:00
06:00:00
07:00:00
08:00:00
09:00:00
10:00:00
11:00:00

converts to

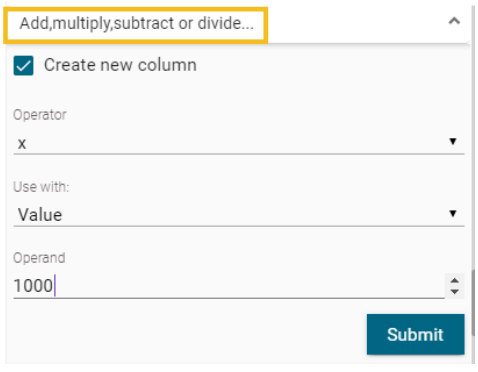
Time_Split string
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000
00:00:00.000

6.4.7. Integer

6.4.7.1. Add, Multiply, Subtract or Divide

It performs the arithmetic operation on the selected numerical column.

- **Operator:** There are four arithmetic operations to choose (+, -, / , *).
- **Use with:** The operation can be performed between column-column and column-value.
- **Operand/Column:** The arithmetic operation needs two operands. The first operand is one on which the operation is being performed. The second operation can either be a value or other numerical column based on the choice of use with an option.



Price(K) integer
34
176
324
74
109
111

converts to

Price(K)_multiply_1 integer
34000
176000
324000
74000
109000
111000

6.4.8. ML

6.4.8.1. Binarizer

It converts the value of a numerical column to zero when the value in the column is less than or equals to the threshold value and one if the value in the column is greater than threshold value.

Screen Size double
13.3
13.3
15.6
15.4
13.3
15.6
15.4
13.3

converts to

Screen Size_binari... double
0.0
0.0
1.0
1.0
0.0
1.0
1.0
0.0

6.4.9. Numbers

6.4.9.1. Max

It gives the maximum value from the selected columns row-wise. The selected column should be numerical and more than one.

6.4.9.2. Mean

It gives the average value of the selected columns row-wise. The selected column should be numerical and more than one.

6.4.9.3. Min

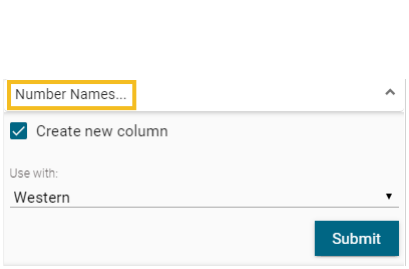
It gives the minimum value from the selected columns row-wise. The selected column should be numerical and more than one.

6.4.9.4. Negate

It complements the sing of a numeric value. If the value is positive, then a negative value comes and vice-versa.

6.4.9.5. Number Name

It converts the value of the selected column into words. The column must be of integer type.
Use with: It gives the users an option to convert word into either western format or Indian format.



Price Euros (integer)
34900
176900
324000
74900
109900

converts to

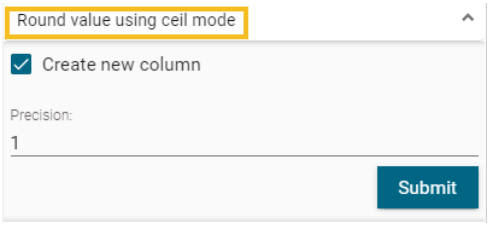
Price Euros_In_W... (string)
Thirty Four Thousand Nine Hundred
One Hundred and Seventy Six Thousand Nine Hundred
Three Hundred and Twenty Four Thousand
Seventy Four Thousand Nine Hundred
One Hundred and Nine Thousand Nine Hundred

6.4.9.6. Remove Fractional Part

It removes the fractional part from the numerical column. The float column is converted into the integer data type.

6.4.9.7. Round Value using Ceil Mode

It replaces the number with a greater integer value if the number is between two integer values. The transformed value can replace the existing column value or can be added as a new column.



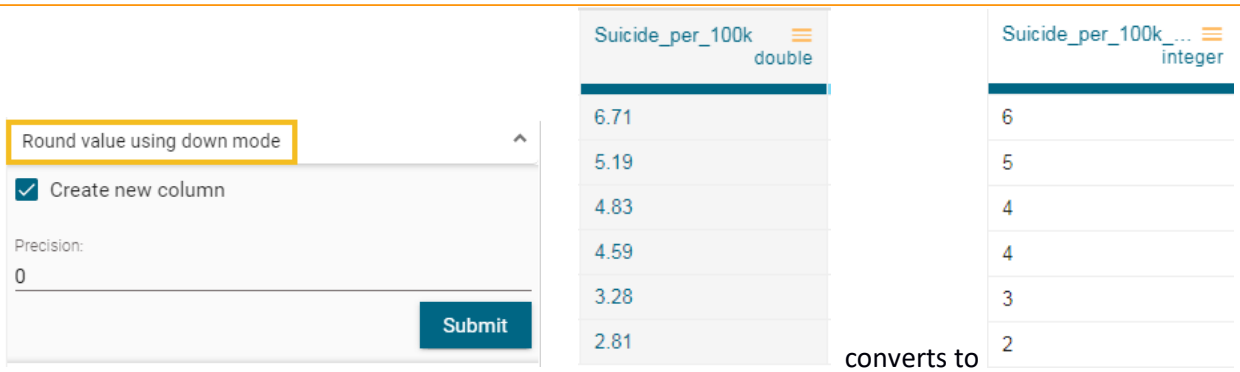
Suicide_per_100k (double)
6.71
5.19
4.83
4.59
3.28
2.81

converts to

Suicide_per_100k_... (double)
6.8
5.2
4.9
4.6
3.3
2.9

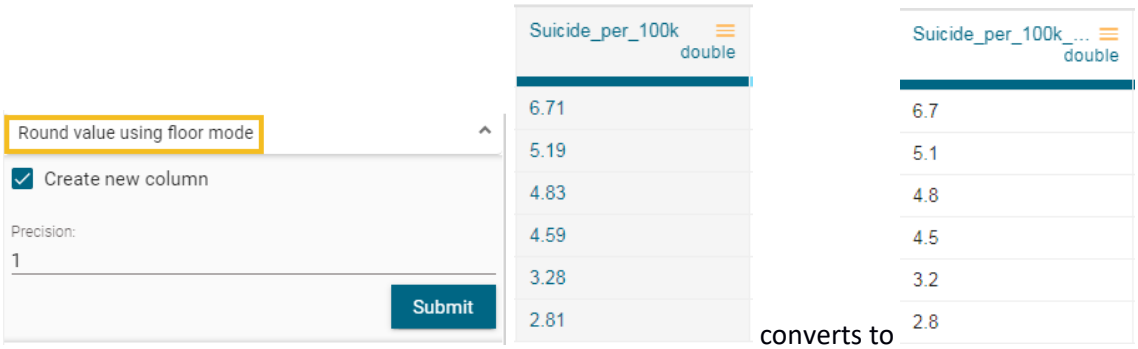
6.4.9.8. Round Value using Down Mode

It rounds the number down to a specified digit or gives the specified number of decimals without any change in value. The transformed value can replace the existing column value or can be added as a new column.



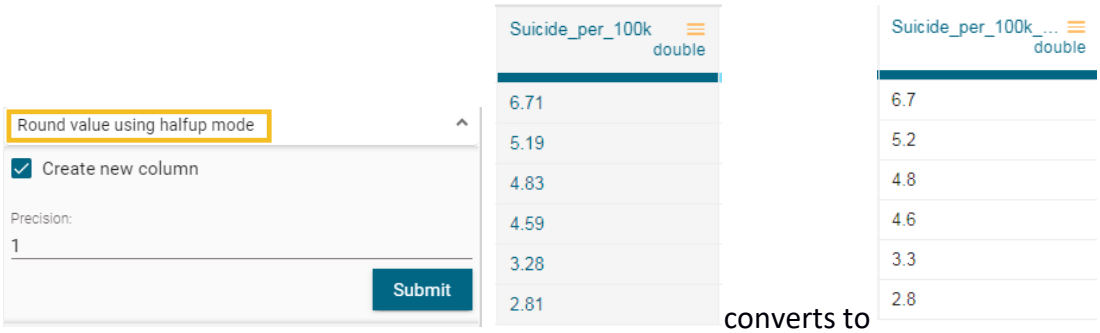
6.4.9.9. Round Value using Floor Mode

It replaces a number with the lesser integer value, if the number is between two integer value, or it rounds the number down to the nearest multiple of Specified significance. It does not consider whether the next digit is 5 or less than or greater than 5. The transformed value can replace the existing column value or can be added as a new column.



6.4.9.10. Round Value using Half-up mode

It replaces a number with the next integer value if its next digit is 5 or greater than 5. The transformed value can replace the existing column value or can be added as a new column.



6.4.10. String

6.4.10.1. Change to lower case

It converts the selected column value to the small case. The transformed value can replace the existing column value or can be added as a new column.

6.4.10.2. Change to Title Case

It converts the selected column value to the title case. The transformed value can replace the existing column value or can be added as a new column.

6.4.10.3. Change to Upper Case

It converts the selected column value to capital letters. The transformed value can replace the existing column value or can be added as a new column.

6.4.10.4. Extract Substring at Position

It extracts the substring from the selected column based on the starting position and the length of the extract. The transformed value can replace the existing column value or can be added as a new column.

- **Position:** This value is required and is the start position. It can be both a positive or negative number. If it is a positive number, this function extracts from the beginning of the string. If it is a negative number, this function extracts from the end of the string.
- **Length:** This value is optional. It specifies the number of characters to extract. If omitted, the whole string is returned starting from the given position.

6.4.10.5. Extract Substring before Delimiter

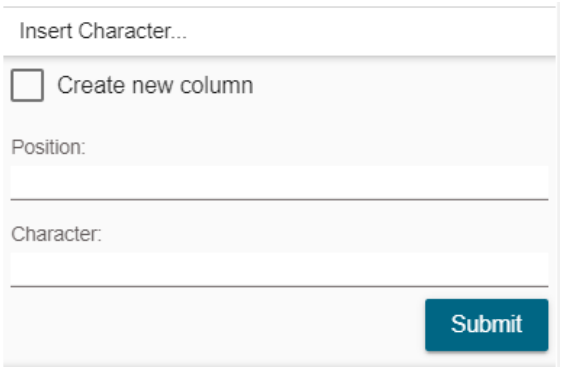
It extracts the substring from the selected column, before the '$n^{th'}$ occurrence of the delimiter specified where 'n' is the count. The transformed value can replace the existing column value or can be added as a new column.

- **Delimiter:** The delimiter on whose occurrence the extract should happen.
- **Count:** This value is mandatory and specifies the count of occurrence of the delimiter before which the extract should happen.

6.4.10.6. Insert Character

It inserts the character entered after a specified position. The transformed value can replace the existing column value or can be added as a new column.

- **Position:** The position in the cell value, after which the character must be inserted. We can even pass comma separated values. E.g., 2,4,6 insert the specified character after position 2, 4 & 6 of the cell values
- **Character**: The character that should be inserted after the specified positions



6.4.10.7. Remove Consecutive Characters

The transform removes the repeated whitespace or character and modifies the selected column /adds the result to a new column. It removes only the repetition.

- **Separator**: it has values whitespace /other. If whitespace, the transform searches for multiple white spaces and returns a single-spaced value.
- **Custom repeated Character:** When a repeated character is '**Other**,' this provides an option to give the character whose consecutive occurrence must be searched.

6.4.10.8. Remove Part of Text

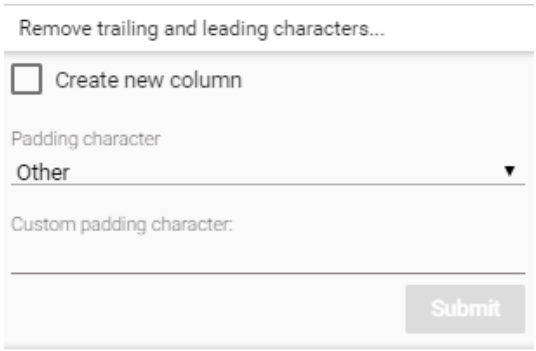
It matches and removes the matching part or entire value based on the condition. The transformed value can replace the existing column value or can be added as a new column.

- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column

6.4.10.9. Remove Trailing and Leading Characters

It removes trailing and leading characters from the column. The transformed value can replace the existing column value or can be added as a new column.

- **Padding character:** Specify whether to remove whitespace or another character using the drop-down menu.
- **Custom padding character -** If '**other**' is selected as a padding character, specify which is the character to be removed.

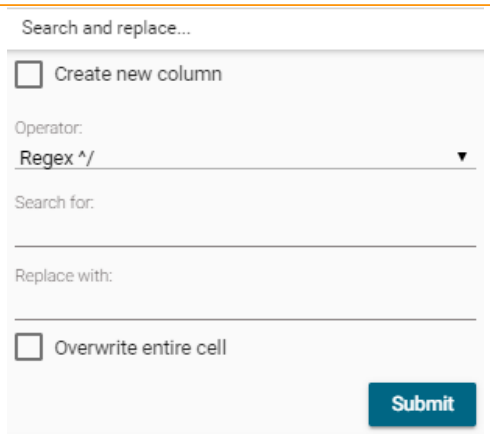


6.4.10.10. Search and Replace

It searches and replaces the matching part or entire value based on the option selected. The transformed value can replace the existing column value or can be added as a new column.

Operator- Select the operator required for matching from the list. Operators include contains, equals, starts with, end with, and regex match.

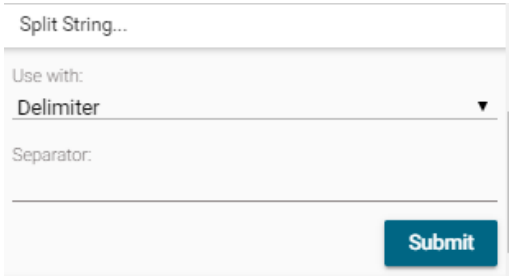
Value: It is the value or pattern to be searched for in the selected column.



6.4.10.11. Split String

It splits the string based on condition. It displays new columns based on the number of delimiter and on position.

- **Use With:** Specify whether to split with a delimiter or at position
- **Delimiter:** The delimiter on whose occurrence the split should happen
- **Position:** After which position split should happen if use with is 'position.'



Here splitting of the column is done based on position (after the 5th character)

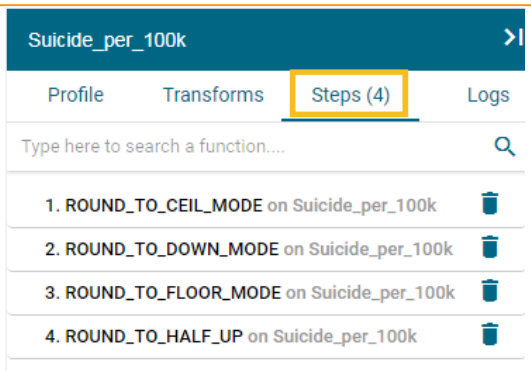
age (string)
15-24 years
35-54 years
15-24 years
75+ years
25-34 years
75+ years
35-54 years

converts to

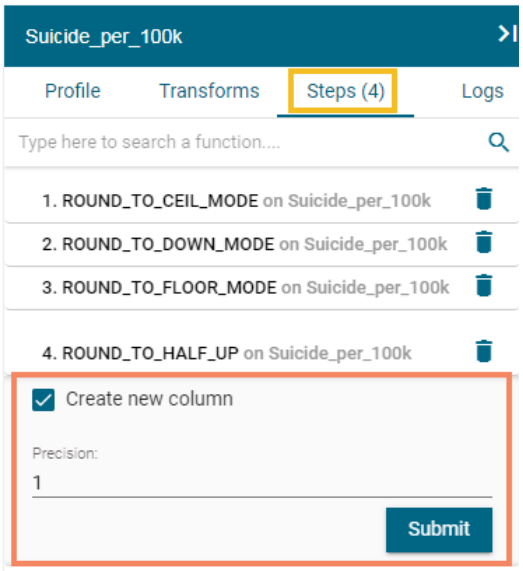
age (string)	age_split_1 (integer)	age_split_2 (string)
15-24 years	15	24 years
35-54 years	35	54 years
15-24 years	15	24 years
75+ years	75+ years	
25-34 years	25	34 years
75+ years	75+ years	
35-54 years	35	54 years

6.5. Steps

This tab lists all the transforms that were performed on the data. It also gives a count of steps performed.



The user can open any performed transform and edit it using the '**Steps**' tab.

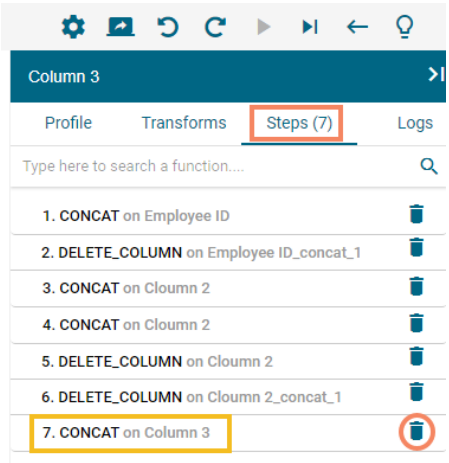


6.5.1. Deleting or Undeleting Steps

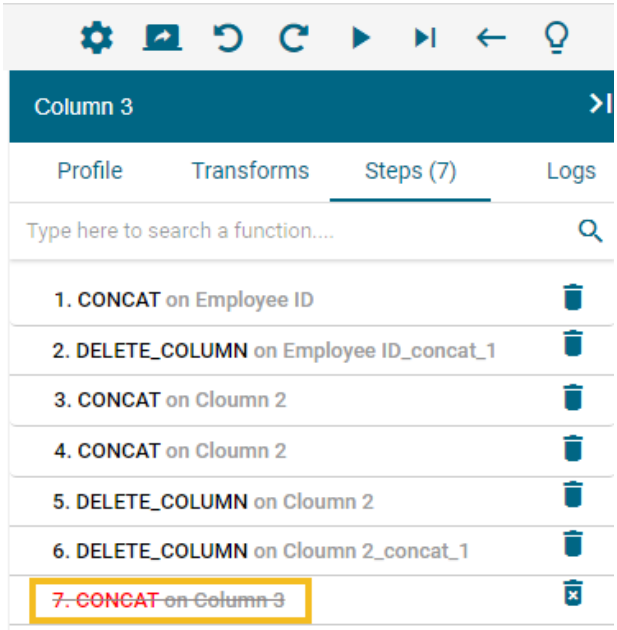
The '**Steps**' tab has option to delete and undelete options to delete or undelete performed steps.

The user can follow the below given steps to use these options:

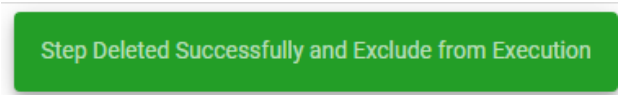
i) Navigate to the '**Steps**' tab (Make sure there are some steps listed on the tab).
ii) Click the '**Delete**' icon for a listed steps.



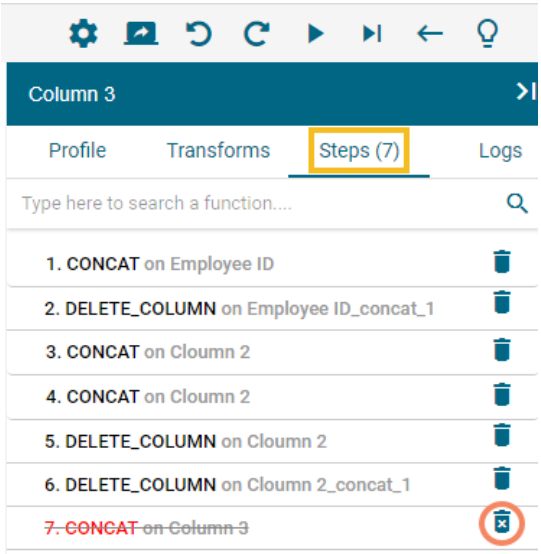
iii) The selected steps get deleted.



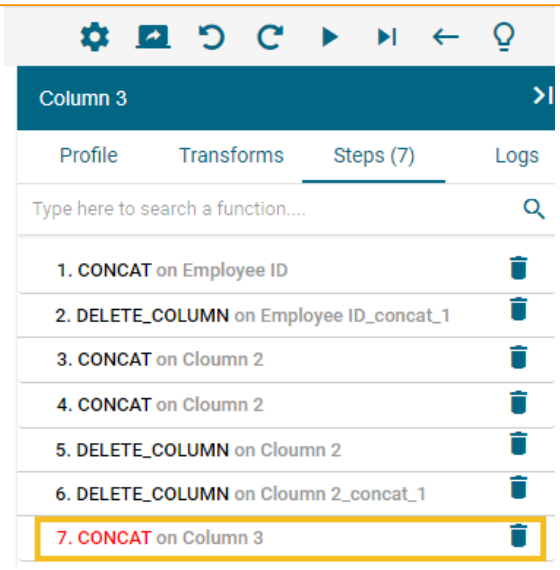
iv) A message appears to notify the same.



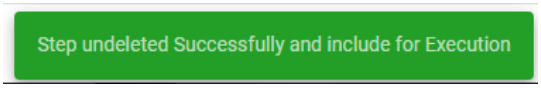
v) Click the '**Undelete**' icon for the deleted step.



vi) The deleted step gets undeleted.

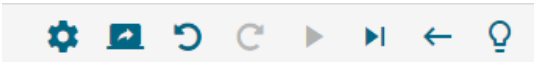


vii) A message appears to notify the same.



7. Navigation Pane

The navigation pane provides options to export the preparation steps in Elastic settings, move the steps out of the BDB Data Preparation. The navigation panel also has icons to perform Undo, Re-do, Replay Dirty, and Replay All options.



7.1. Export Steps to a Data Store Meta Data

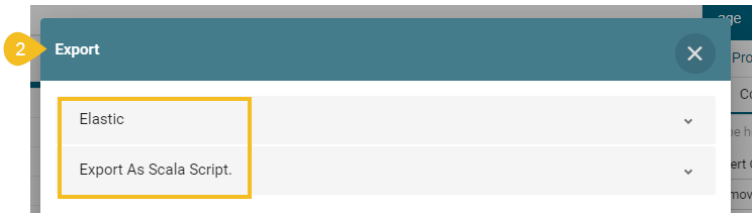
The '**Export Steps to data store**' option redirects the user to specify the settings into which the cleansed data must be moved.

viii) Click the **'Export Steps to data store'** icon using the Navigation Pane**.**

ix) The Export window opens with two options.

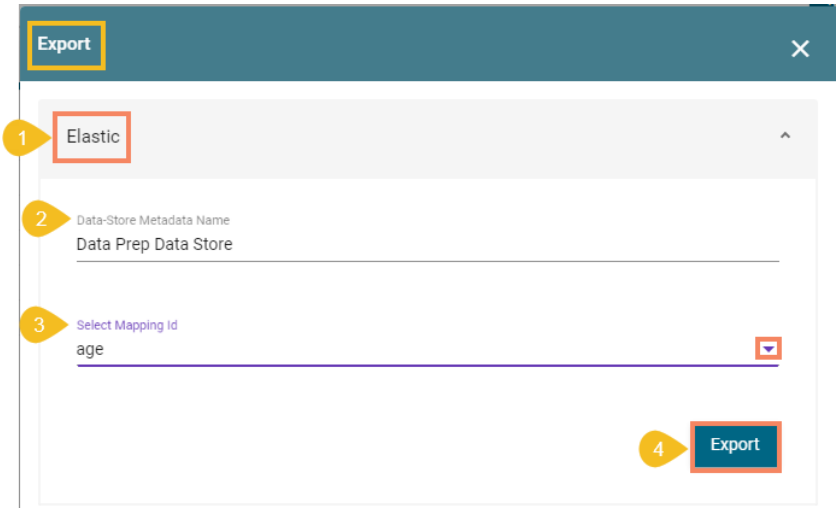
a. Elastic

b. Export as Scala Script



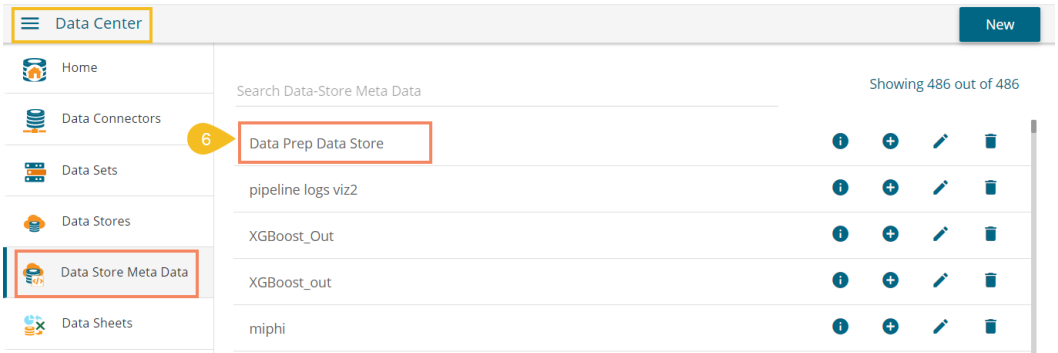
- Steps when the '**Elastic**' option is selected.

i) Select the '**Elastic**' option and provide the following details.
ii) Data-Store Metadata Name: Provide a name for the data store metadata.
iii) Select Mapping Id: Select a matching column from the drop-down menu.
iv) Click the '**Export**' option.

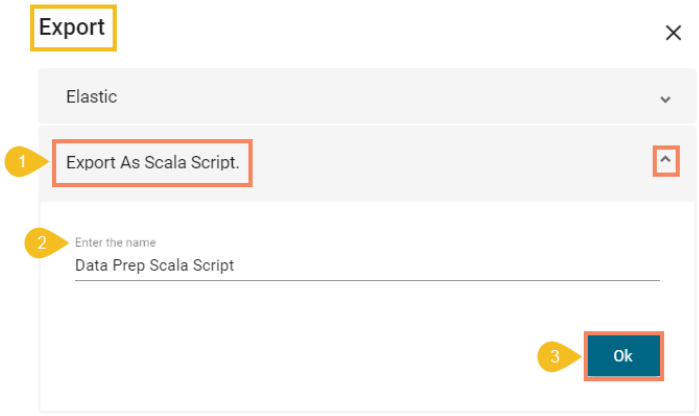


v) A Success message appears to confirm.

vi) The settings get exported to the selected Elastic Settings. The user can see it under the Data Store Meta Data list of the Data Center module.



- Steps when the '**Export as Scala Script**' option is selected.
  i) Select the '**Export as Scala Script**' option and provide the following details.
  ii) Enter the name of the Script.
  iii) Click the '**Ok**' option.



iv) A success message appears, and the Scala Script gets downloaded to the system.

7.2. Export Steps to Pipeline

This option provides an option to specify the name in which the steps/transforms created as part of cleansing must be exposed to the pipeline module of the platform.

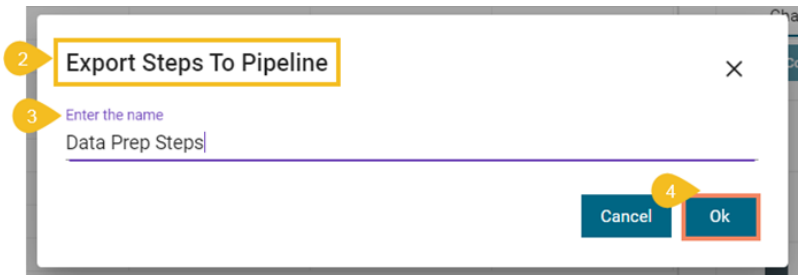
i) Click the '**Export Steps to Pipeline**' icon.

ii) The '**Export Steps to Pipeline**' window opens.

iii) Provide the name for the Data Prep script.

iv) Click the 'Ok' option.

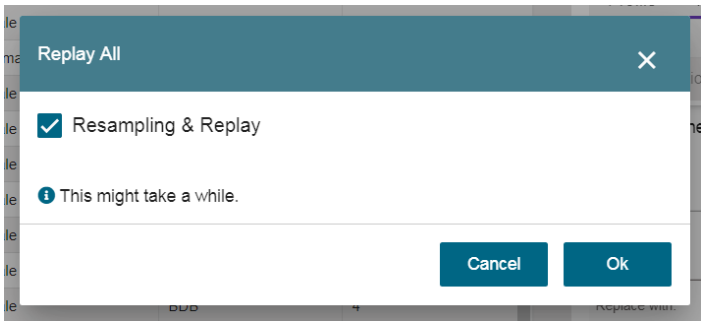
v) A success message appears to assure the step.



vi) The Data Prep Script gets listed under the Data Scripts section of the Data Pipeline plugin.

7.3. Other options in the Navigation Pane

1. **Undo** : The user can Undo a list of the last few transforms. This button gets enabled only if we have applied at least one transform on the data.
2. **Redo** : Redo a list of last few transforms, that was undone. If we have not undone any transform, then the '**redo**' icon gets disabled.
3. **Replay dirty** **:** The '**Replay Dirty**' option when applied on the data from a specific step it replays all the transforms which are listed after the selected transform in the list of steps.
   - The '**Replay Dirty**' option gets enabled only when the user edits some transform step using the '**Steps**' tab.
   - To indicate what all transform steps will be affected, the listed steps get colored in red.
   - After the '**Replay Dirty**' function gets applied, all the steps that were colored in red become black and all the transforms get applied to the dataset.
4. **Replay All** **:** The Replay All option allows the user to resample the data and replay the steps on the new data sample. It is useful when there is a change in the underlying dataset. It updates the data in the grid applying all the steps (In case of edit or steps added after edit).
   - Click the '**Replay All**' icon from the navigation pane.
   - The '**Replay All**' window appears.
   - Select the '**Resampling & Replay'** option using the checkbox (if required).
   - Click the '**Ok**' option.



4. **Close the Preparation**: The user can exit from the preparation window and reach the landing page of data preparation.

Note: The standalone version of data preparation provides an option to export the prepared data to elastic so that that visualization modules can consume it.

8. Signing Out

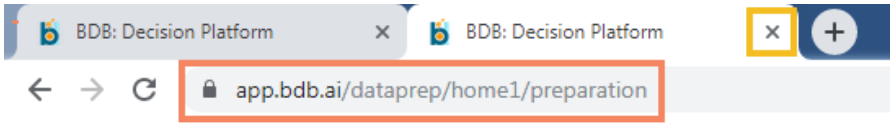
The users can Sign-out from the Data Preparation tab at any given stage, but preferable is that the users should complete all the preparation tasks they wish to perform and save it before closing the tab or singing out from the Platform.

The Signing Out process for the Data Preparation has two steps:

1. Closing the BDB Data Preparation

Once you have completed the Data Preparation tasks, save your work and close the Data Preparation tab.

Click the **'Close'** button (the 'X' on the right edge) from the Data Preparation tab.

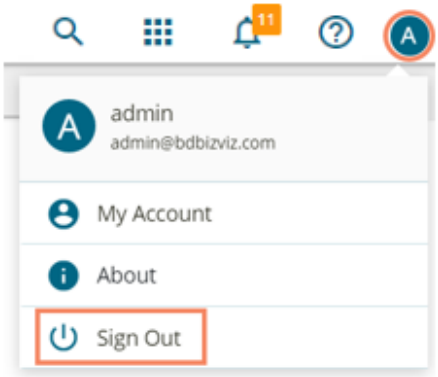


2. Sign Out from the BDB Platform

The following steps describe how to Sing-off from the BDB Platform.

i) Click the '**User Profile**' icon on the Platform homepage.

ii) Click the '**Sign Out**' option.

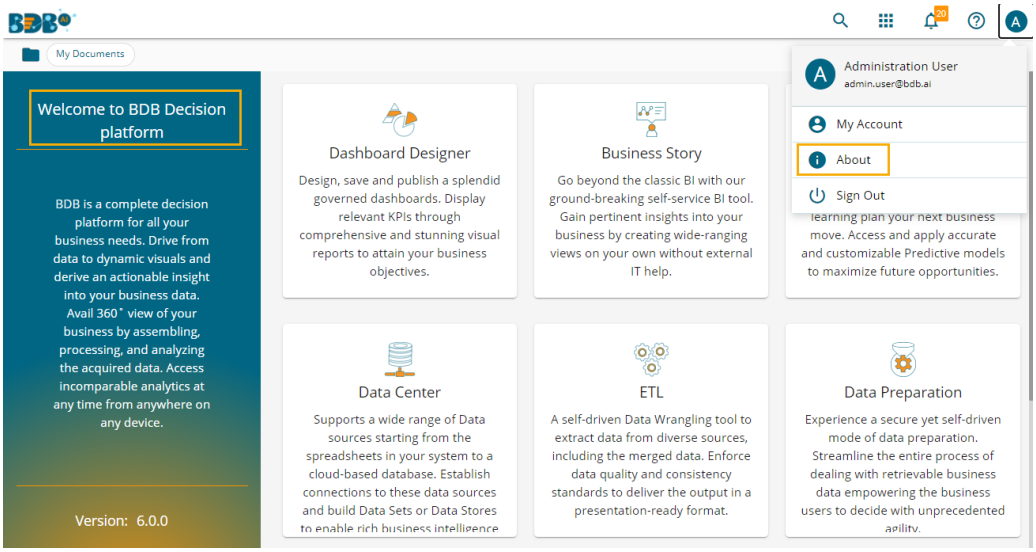


iii) The user successfully signs off from the **BDB Platform.**

Note:

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

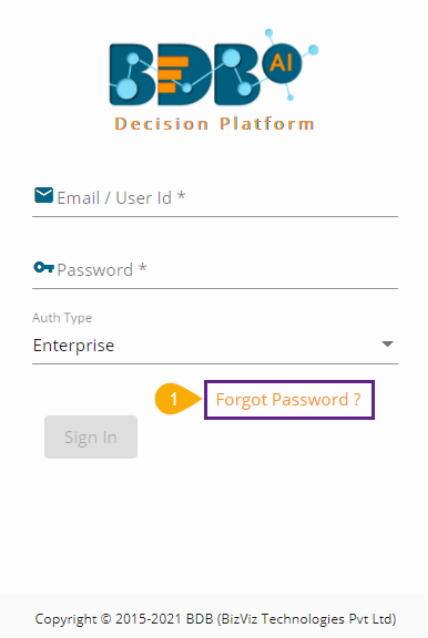
b. Click the '**About**' option to open the default homepage for the BDB Platform.



8.1. Forgot Password Option

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

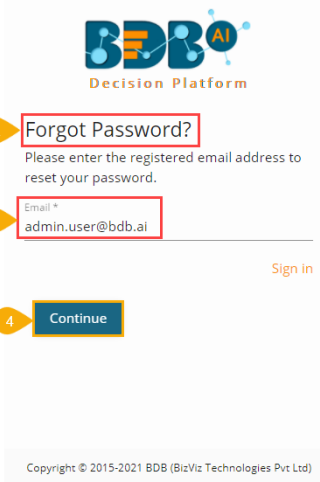
i) Click the '**Forgot Password?**' option from the Sign In page.



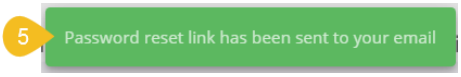
ii) The '**Forgot Password?**' page opens.

iii) Provide the email id that is registered with BDB to send the reset password link.

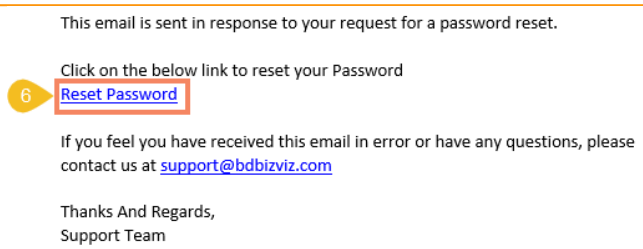
iv) Click the '**Continue**' option.



v) The user may be redirected to select a space in case of multiple spaces under one server link (The user needs to select a space and click the '**Continue**' option once again). In case if a user does not have multiple spaces, a message appears to notify the user that the password reset link (The users receive the reset link via their registered email.)

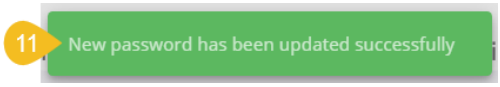


vi) Click the link from your registered email.



vii) The user gets redirected to the '**Reset Password**' page to set a new password.
viii) Set a new password.
ix) Confirm the newly set password.
x) Click the '**Continue**' option.

xi) The password for the selected BDB account gets reset and a message appears to inform the user.

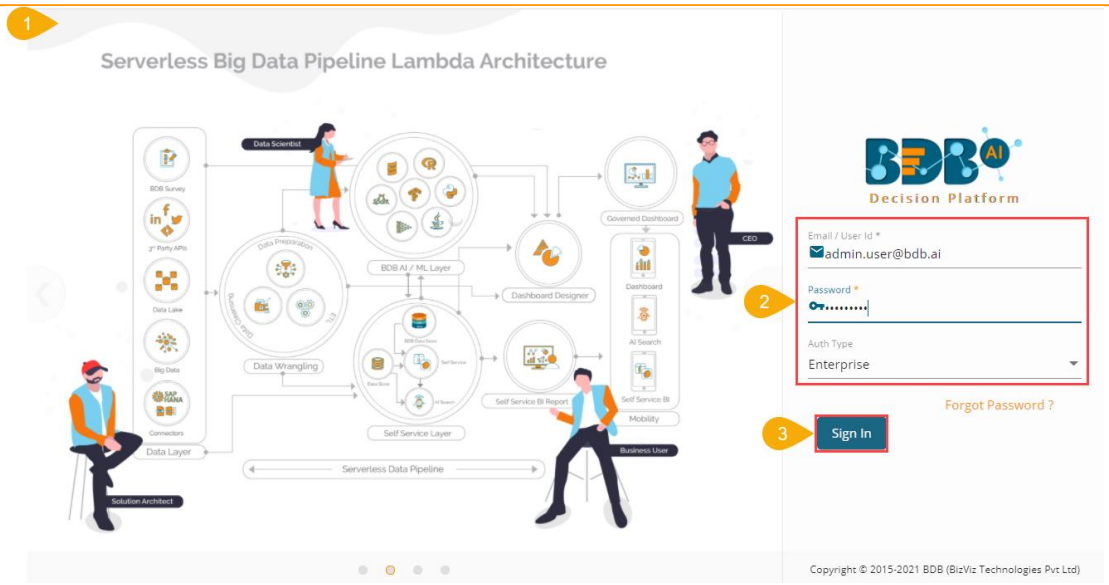


Note: The user gets redirected back to the Sign In page after successfully resetting the password.

8.2. Force Login

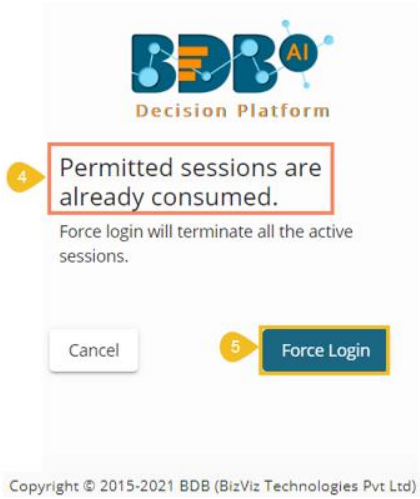
The '**Force Login**' functionality has been introduced to control the number of active sessions up to three. The users can access only 3 sessions at a time when they try to access the 4th session, a warning message displays to inform that the user has consumed the permitted sessions and, a click on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



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

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



vi) A warning message appears the currently active sessions get killed, and the user gets redirected to the BDB Platform Sign In page.

vii) The user needs to provide valid credentials once again and click the '**Continue**' option to access the platform.