

User Guide

Data Preparation- 4.2

Contents

1.	About this Guide	3
1.1.	Document History.....	3
1.2.	Overview	3
1.3.	Target Audience	3
2.	System Requirments Specification.....	3
3.	Getting Started with BDB Data Preparation	4
3.1.	Forgot Password Option	6
3.2.	Force Login.....	8
4.	Data Grid.....	9
4.1.	Data Grid Header.....	9
4.2.	Data Types.....	9
4.3.	Panel to List the Selected Filters.	9
4.4.	Data Quality Bar in the Grid.....	10
4.5.	Pagination	10
5.	Summary Pane	11
5.1.	Charts.....	11
5.2.	Info: Value/Statistics.....	12
5.3.	Pattern	13
5.4.	Transforms	13
5.4.1.	Columns.....	14
5.4.2.	Conversions	17
5.4.3.	Data Cleansing	18
5.4.4.	Dates	24
5.4.5.	Integer	29
5.4.6.	ML	30
5.4.7.	Numbers.....	30
5.4.8.	String	33
5.5.	Steps	36
6.	Navigation Pane.....	36
7.	Signing Out.....	37

1. About this Guide

1.1. Document History

Product Version	Date (Release date)	Description
BDB Data Preparation 4.0	December 31 st , 2018	First Release of the Document
BDB Data Preparation 4.2	March 25 th , 2019	Updated Document

1.2. Overview

This guide covers:

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

1.3. Target Audience

This guide is aimed at users who wish to use BDB Data Preparation option to prepare and transform their business data.

2. End User System Requirements Specification

This section provides information on the hardware and software parts to install and run the BDB Data Preparation.

Hardware Requirements

Processor	A 64-bit processor is required.
Allocated Memory	1GB minimum
Disk Space	500MB minimum + datasets = 5 GB + recommended

Software Requirements:

Operating System	Windows 7 64-bits or later version Mac OS X 10.7 Lion or later version Ubuntu 14.04 and above
------------------	---

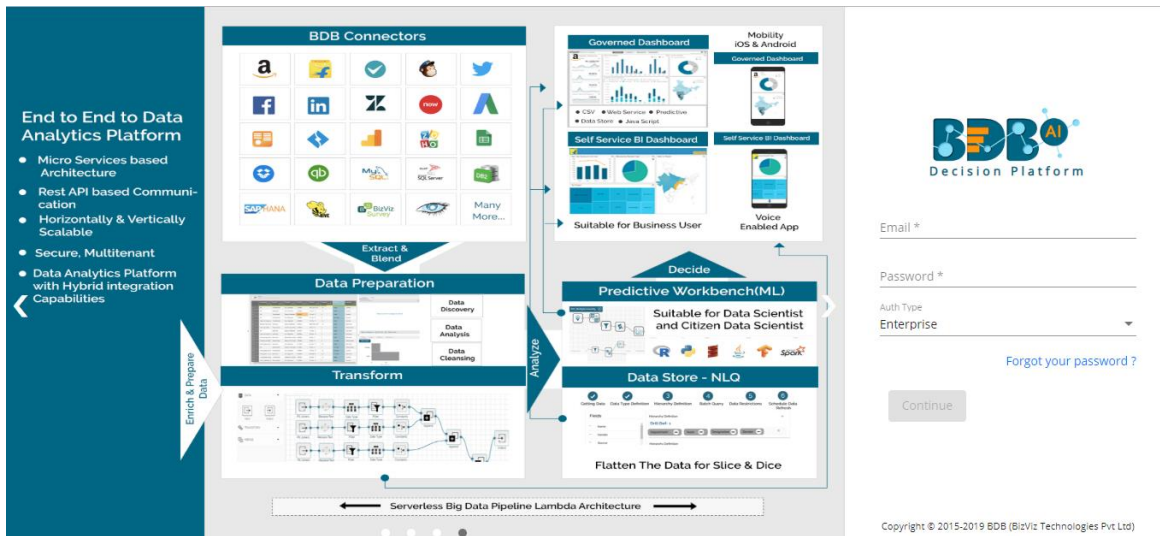
Compatible Web Browsers:

Mozilla Firefox/ Firefox ESR	Latest Version
Microsoft Internet Explorer	11
Microsoft Edge	Latest Version
Apple Safari	10
Google Chrome	Latest Version

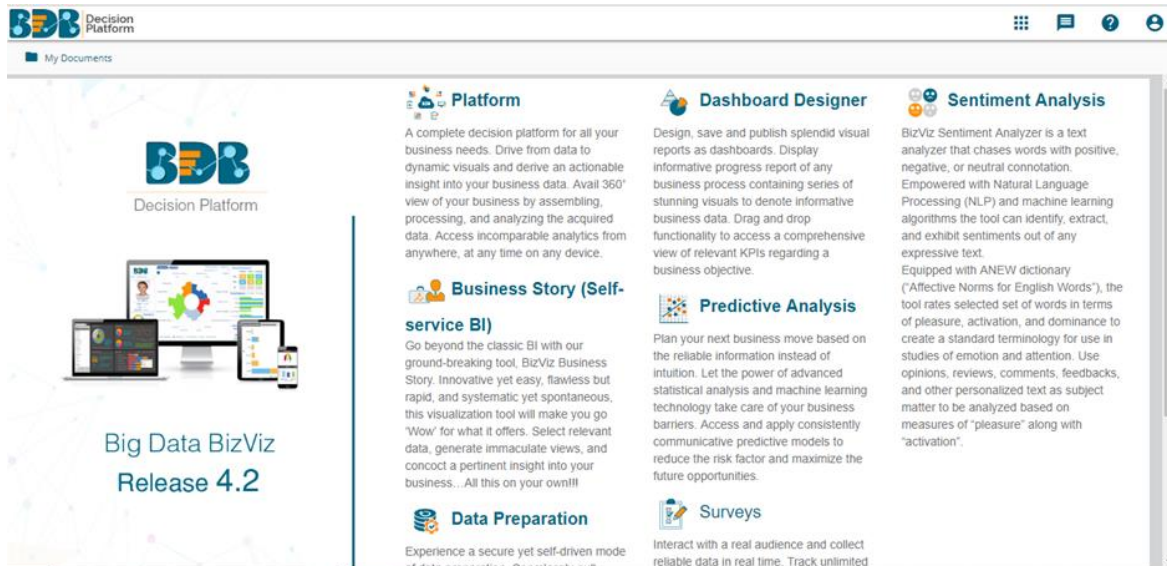
3. Getting Started with BDB Data Preparation

This section covers initial steps to access the BDB Dashboard Designer plugin using the BDB Platform.

- i) Open the BDB Enterprise Platform Link: <https://app.bdb.ai>
- ii) Enter your credentials to log in to the platform.
- iii) Click the ‘Continue’ option.

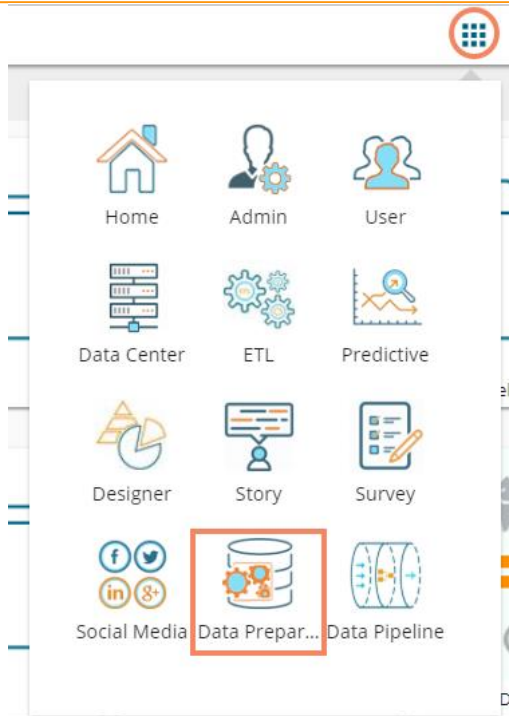


- iv) BDB Platform homepage opens (The below page appears only for the first time when the user login. Once the user creates some document, he gets directed to the homepage by default).

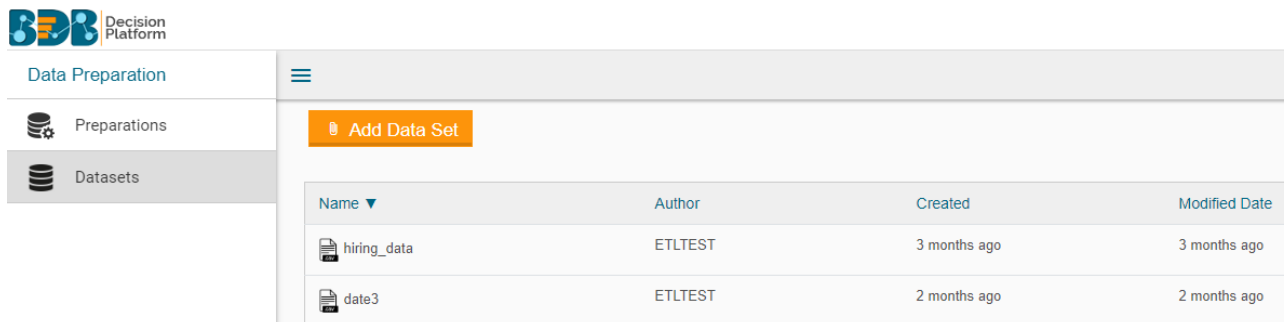


Note: The above screen opens only for those newly created users who have not yet created any document/folder using the BDB Platform.

- v) Click on the ‘App’ menu button.
- vi) Select the ‘Data Preparation’ plugin from the app menu.



vii) A new window opens displaying the landing page for the Data Preparation.



viii) The landing page of data preparation has two menus.

a. Preparations

It lists all the available preparations, when it was created, who created when it was last modified and on which data set.



The users also get an option to add a new preparation. The users can continue adding more steps to the existing preparations.

b. Datasets

The 'Datasets' section lists the data/input which was added to the system. The users can create a new preparation on any dataset.

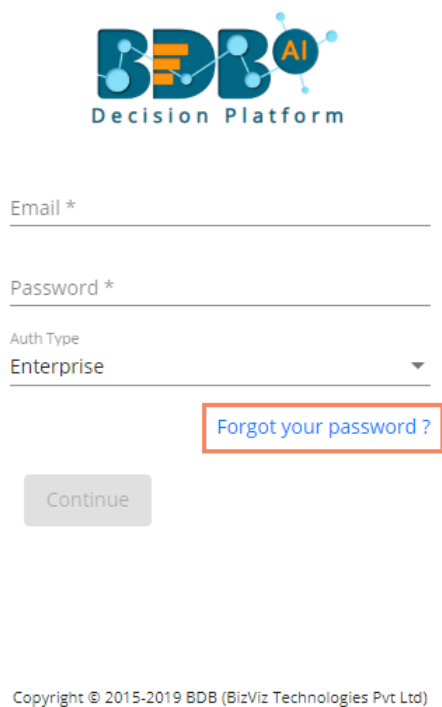
The window also provides an option to add new datasets.

Note: The standalone version of 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

3.1. 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 of the BDB Platform.
- ii) Click the '**Forgot your password?**' option.



BDB AI
Decision Platform

Email *

Password *

Auth Type
Enterprise

[Forgot your password ?](#)

Continue

Copyright © 2015-2019 BDB (BizViz Technologies Pvt Ltd)

- iii) Users get redirected to a new window.
- iv) Provide the email id that is registered with BDB to send the reset password link.
- v) Click the '**Continue**' option.



Having trouble signing in?

To reset your password, enter the email address you use to sign in to BizViz. This can be your email address associated with your account.

Email *

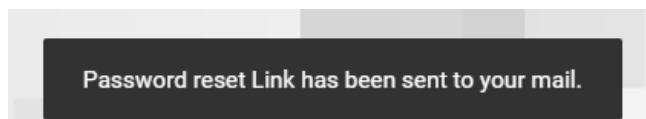
admin@bdb.ai

[Sign in](#)

Continue

Copyright © 2015-2019 BDB (BizViz Technologies Pvt Ltd)

- vi) Users may be redirected to select a space in case of multiple areas under one server link; they need to choose a space and click the '**Continue**' option once again. Otherwise, a message will pop-up to notify that the password reset link has been sent to the 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.



Reset Password

You've confirmed ownership of the BizViz Account, Reset your password now to regain access.

New Password *

.....

Confirm New Password *

.....

Continue

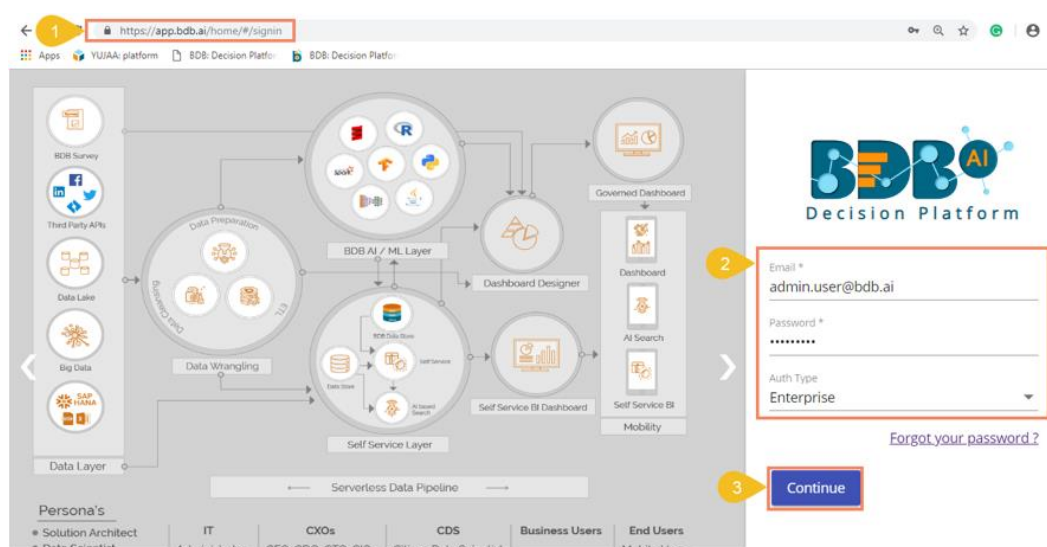
Copyright © 2015-2019 BDB (BizViz Technologies Pvt Ltd)

- xii) The new password gets updated for the selected BDB account, and the user gets redirected back to the 'Log In' page of the BDB Platform.

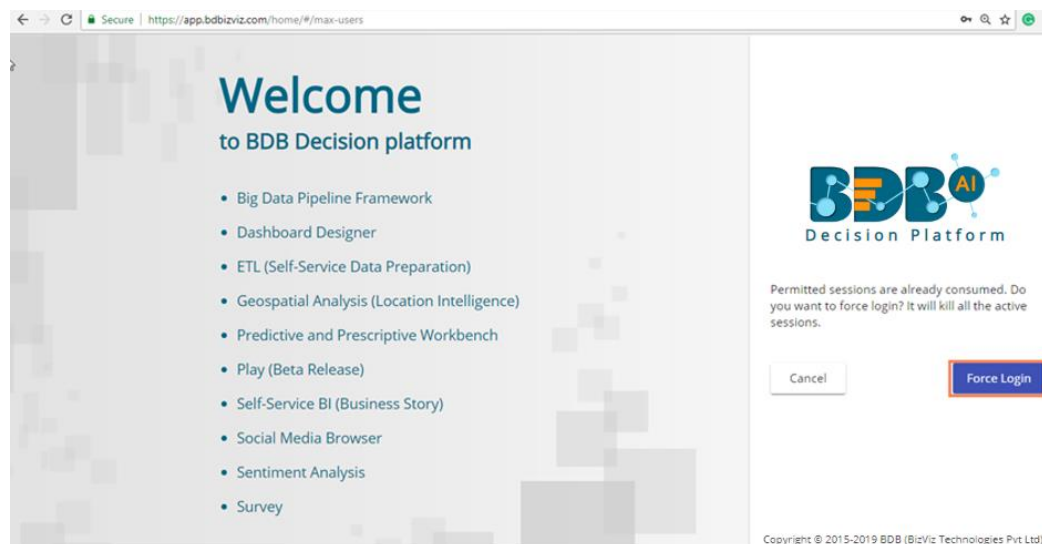
3.2. Force Login

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 will get 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. 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 displayed data in the grid changes based on the number of transforms performed on it.

4.1. Data Grid Header

The grid has a header which displays the column name from the dataset. The context menu in the header has an option to rename the column and delete the column. 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 a 10000 rows sample has 9000 integers and 1000 string values, the selected The datatype is Integer, and the 1000 rows will be detected as invalid rows.

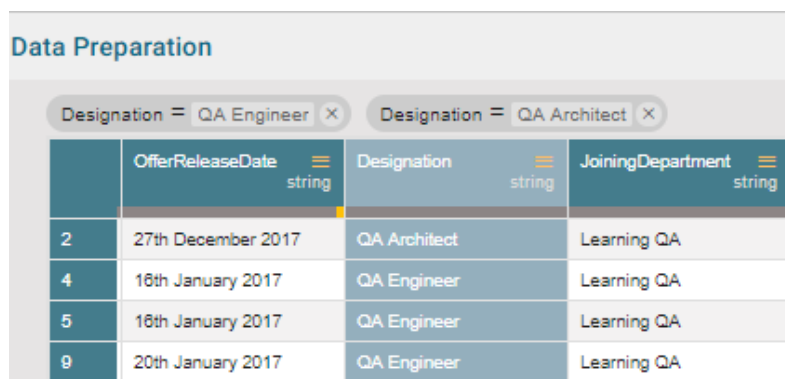
4.2. Data Types

The BDB Data Preparation supports the following data types:

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

4.3. Panel to List the Selected Filters.

When a filter is selected, it gets added to the filter panel on top of the grid. The added filter has an option to remove it by clicking the ‘**Close**’ (X) mark.



The screenshot shows the 'Data Preparation' interface. At the top, there are two filter panels: 'Designation = QA Engineer' and 'Designation = QA Architect', each with a close (X) button. Below the filters is a data grid with the following columns: OfferReleaseDate (string), Designation (string), and JoiningDepartment (string). The grid displays four rows of data:

	OfferReleaseDate <small>string</small>	Designation <small>string</small>	JoiningDepartment <small>string</small>
2	27th December 2017	QA Architect	Learning QA
4	16th January 2017	QA Engineer	Learning QA
5	16th January 2017	QA Engineer	Learning QA
9	20th January 2017	QA Engineer	Learning QA

The left bottom of the grid displays the number of rows meeting the filter condition out of the total.

30	017	QA Engineer
35	2017	QA Engineer
37	2017	QA Architect
39	2017	QA Engineer

19/ 183

4.4. 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:

- Brown-Valid Data
- Orange- Invalid data
- Light blue -Blank data

gender	source
string	string
male	
female	
female	
0	
1	
1	agency

4.5. Pagination

Pagination is implemented for the grid data. The tool displays 20 records on each page. The maximum rows displayed for sampling is always 10k.



Note: The users can get information about the Column Type, option to Delete the column and option to Rename the column by clicking the 'Column Menu' icon provided next to the column names in the data grid.

Name	age
string	integer
Ahsan	
Rajive Raveendra Pai	
Amit Kumar Soni	

This column is a **string**

Delete column

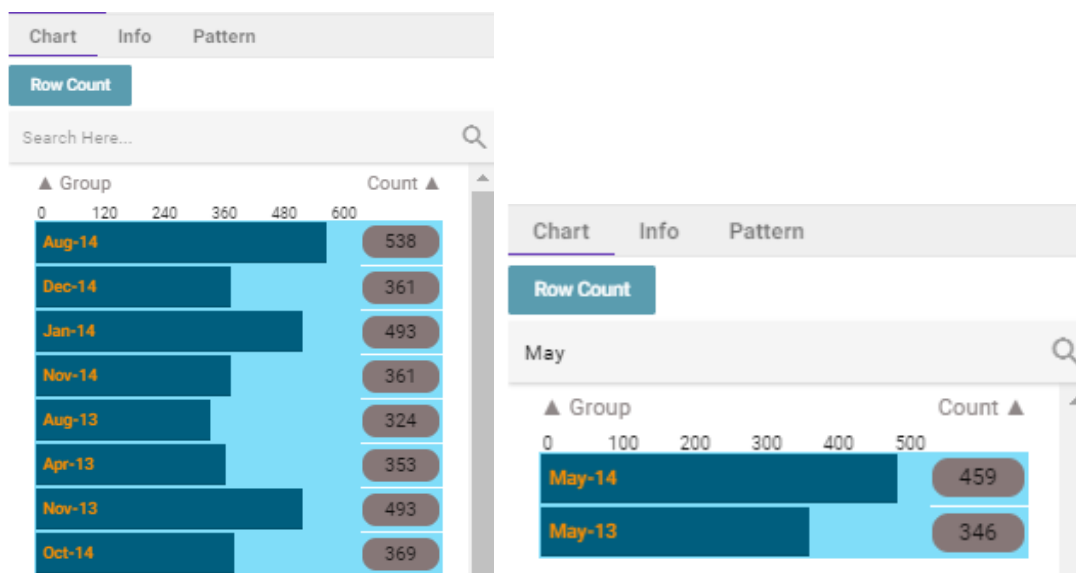
Rename column

5. Summary Pane

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

5.1. 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 graph is interactive. When the user clicks on any bar, it will add a filter in the filter pane and filters the data displayed in the grid. Later the transform can be performed on the filtered data.

The chart can be sorted based on the group or the count of occurrence of a group.

The screenshot shows the 'Data Preparation' interface. On the left is a data grid with columns: ing_timestamp, joining_status, current_status, exited_date, experience, previous_ctc, and offered_salary. The 'current_status' column is filtered to 'resigned'. On the right is a summary pane for 'current_status' with a 'Row Count' chart. The chart shows counts for 'absconded' (72), 'joined' (992), 'resigned' (528), and 'terminated' (46).

ing_timestamp	joining_status	current_status	exited_date	experience	previous_ctc	offered_salary	
7	0:00:00...	joined	resigned	2016-03-28T00:00:00...	7.2		67891
9	0:00:00...	joined	resigned	2015-10-12T00:00:00...	4.0		4520
12	0:00:00...	joined	resigned	2015-04-11T00:00:00...	3.8		35601
13	0:00:00...	joined	resigned	2016-08-06T00:00:00...	4.2		40271
14	0:00:00...	joined	resigned	2016-07-22T00:00:00...	3.1		3086
16	0:00:00...	joined	resigned	2015-07-09T00:00:00...	4.5		40271
17	0:00:00...	joined	resigned	2017-04-26T00:00:00...	4.0		4520
18	0:00:00...	joined	resigned	2013-04-09T00:00:00...	6.0		6091
20	0:00:00...	joined	resigned	2015-11-09T00:00:00...	4.5		40271
22	0:00:00...	joined	resigned	2017-11-16T00:00:00...	3.4		3072
23	0:00:00...	joined	resigned	2017-03-11T00:00:00...	4.5		40271
26	0:00:00...	joined	resigned	2013-04-16T00:00:00...	3.5		4520
28	0:00:00...	joined	resigned	2015-01-05T00:00:00...	0.0		20451
29	0:00:00...	joined	resigned	2013-05-15T00:00:00...	4.2		4520

5.2. Info: Value/Statistics

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

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

The screenshot shows a data table with two columns: 'Source' (string) and 'ReferralOf' (string). The 'ReferralOf' column is selected, and its 'Info' statistics are displayed in a side panel. The statistics are as follows:

Source		
Profile	Transforms	Steps: 0
Chart	Info	Pattern
Count:	50	Duplicate: 42
Valid:	50	Distinct: 8
Invalid:	0	

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

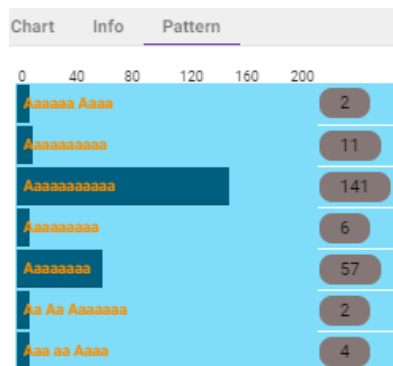
- Minimum
- Maximum
- Mean
- Variance

The screenshot shows a data table with two columns: 'Experience' (double) and 'PreviousCTC' (double). The 'Experience' column is selected, and its 'Info' statistics are displayed in a side panel. The statistics are as follows:

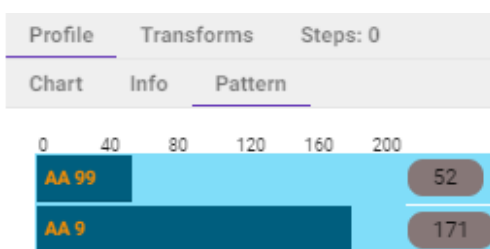
Experience		
Profile	Transforms	Steps: 0
Chart	Info	Pattern
Count:	50	Duplicate: 20
Valid:	25	Distinct: 30
Invalid:	25	Variance: 14.02
MAX:	20.0	
MIN:	1.0	
Mean:	4.86	

5.3. Pattern

This section focuses on how data pattern and occurrences of each pattern in the dataset sample are plotted in a chart.



team	string	usd_billing	double
BU 6		4000.0	
BU 6		4000.0	
BU 6		2300.0	
BU 6		1750.0	
BU 7		0.0	
BU 11		0.0	



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

	actual_joining_date	candidate_id	comments	current_status	designation	expected_joining_date	experience	expysper_ctc
1	2nd January 2017	1.0	Nil	Transferred	QA Manager	2017-01-02T00:00:00...	15.0	120000.0
2	16th January 2017	2.0	Nil	Resigned	QA Architect	2017-01-18T00:00:00...	10.0	150000.0
4	18th January 2017	4.0	Nil	Transferred	QA Engineer	2017-01-18T00:00:00...	5.0	130000.0
5	15th February 2017	5.0	Was not happy with 4.5	Transferred	QA Engineer	2017-02-19T00:00:00...	2.5	208000.0
6	Declined	6.0	Brother met accident	Declined	Senior Software Engin...	2017-02-20T00:00:00...	4.2	233333.0
8	Declined	8.0	Nil	Declined	Senior Software Engin...	2017-03-13T00:00:00...	3.0	281967.0
9	Declined	9.0	Was not happy with 4.5	Declined	QA Engineer	2017-02-20T00:00:00...	2.0	260000.0
10	Declined	10.0	Not willing to join us a...	Declined	Business Analyst	2017-02-09T00:00:00...	2.0	325000.0
12	Declined	12.0	Not Happy with the off...	Declined	QA Engineer	2017-03-24T00:00:00...	3.5	214266.0
14	13th February 2017	14.0	Immediate joining	Transferred	QA Architect	2017-02-13T00:00:00...	10.0	170000.0
15	Declined	15.0	Asking for 7.5 LPA, ne...	Declined	QA Engineer	2017-03-27T00:00:00...	2.5	240000.0
16	20th March 2017	16.0	Candidate bargained ...	Transferred	Software Engineer	2017-03-20T00:00:00...	2.3	369665.0
18	Declined	18.0	Nil	Declined	Software Engineer	2017-03-15T00:00:00...	3.0	310000.0
19	3rd April 2017	19.0	Nil	Transferred	QA Engineer	2017-04-03T00:00:00...	2.0	300000.0
20	8th March 2017	20.0	Nil	Resigned	Lead QA Engineer	2017-03-08T00:00:00...	11.2	125000.0
21	Declined	21.0	was holding offer of 7.5	Declined	Senior Software Engin...	2017-03-27T00:00:00...	2.7	296296.0

5.4. 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.

5.4.1. Columns

5.4.1.1. 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 value and one string value, then the data type of column will be profiled as the integer despite one string value present in it. Cast to type will remove the value with the invalid data type. In this case, it will convert 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.*

5.4.1.2. Collect Set

It will generate the list of all the unique values of the column based on the selected column. It will perform group concatenation.

CPU	RAM
AMD A12-Series 9720...	12GB
AMD A12-Series 9720...	12GB
AMD A12-Series 9720...	8GB
AMD A12-Series 9720...	6GB
AMD A12-Series 9720...	6GB

CPU	RAM
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]

generates the list of all unique value

5.4.1.3. 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 'candidate_id' column.

Concatenate with...	candidate_id integer	BDB_candidate_id string
<input checked="" type="checkbox"/> Create new column		
Prefix BDB_	1	BDB_1
	2	BDB_2
	3	BDB_3
Use with Value	4	BDB_4
	5	BDB_5
Suffix	6	BDB_6
	7	BDB_7
<input type="button" value="Submit"/>	8	BDB_8

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.

5.4.1.4. Delete Column

It deletes any selected column.

To perform the transform, select the column and click on the 'Delete Column' transform.

5.4.1.5. Duplicate Columns

It will create a duplicate of the selected column.

name string	name_duplicate_1 string
Ritu	Ritu
Vedprakash	Vedprakash
Ajish.T.Thomas	Ajish.T.Thomas
Amit Kumar Soni	Amit Kumar Soni
Animesh Srivastava	Animesh Srivastava
Ahsan R	Ahsan R

gives

5.4.1.6. Generate Primary Key

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

Use with: Here we have two options to generate the primary key. Contiguous will generate the auto-incremented value starting from 1.

The 'Non_contiguous' option will generate the unique and random integer value.

	Primary_column_1
	integer
Generate Primary Key...	1
Use with:	2
Contiguous	3
	4
	5

Submit

5.4.1.7. 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

Return Non Null Column Values...

Select Column
 usd_billing, cur_monthly_payment

Column Name:
 salary

Submit

usd_billing	cur_monthly_paym...
double	double
3000.0	63824.17
2400.0	25603.75
2400.0	25718.58
3500.0	56575.33
2400.0	33565.75
2400.0	37670.42
2400.0	33565.75
	200000.0
2400.0	29673.58
2400.0	33565.75

salary
double
3000.0
2400.0
2400.0
3500.0
2400.0
2400.0
2400.0
200000.0
2400.0
2400.0

returns the new result column

5.4.1.8. SQL Transform

This transform allows us 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.

designation	length(designation)
QA Engineer	11
QA Engineer	11
Business Analyst	16
Senior Software Engin...	24
AWS Consultant	14
QA Manager	10

5.4.2. Conversions

5.4.2.1. 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

Below is the snapshot of how the transform converts data:

Duration_hrs	Converted Value
11.3	678.00
3.4	204.00
3.8	228.00
6.7	402.00
3.4	204.00
3.1	186.00
7.2	432.00
4.2	252.00
4.0	240.00
4.2	252.00

converts to

5.4.3. Data Cleansing

5.4.3.1. 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 on 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

Clear cells on matching value...

Operator:

Value:

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

gender
male
female
female
0
1
1
female
1
male

turns

gender
male
female
female
0
female
male

when above transformation is applied

5.4.3.2. 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

Delete rows on matching value...

Operator:
 Regex ^/ ▼

Value:

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

gender
male
female
female
0
1
1
female
1
male

turns to

gender
male
female
female
female
male

when the above transform is applied.

5.4.3.3. Delete Rows with Empty Cell

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

name	gender	source	referral_of
Emp ID 1	male	internal	
Emp ID 2	female	internal	
Emp ID 3	female	internal	
Emp ID 4	0	internal	
Emp ID 5	1	internal	
Emp ID 6	1	agency	
Emp ID 7	female	portal	
Emp ID 8	1	portal	
Emp ID 9	male	portal	
Emp ID 10	1	portal	
Emp ID 11	male	referral	
Emp ID 12	1	portal	
Emp ID 13	male	referral	Emp ID 9
Emp ID 14	male	referral	Emp ID 1

- 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:

	name <small>string</small>	gender <small>string</small>	source <small>string</small>	referral_of <small>string</small>
1	Emp ID 13	male	referral	Emp ID 9
2	Emp ID 14	male	referral	Emp ID 1

5.4.3.4. Delete Rows with Invalid Cell

- a. The transform deletes any row which has 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 <small>string</small>
male
female
female
0
1
1
female
1
male

returns

gender <small>string</small>
male
female
female
female
female
male

5.4.3.5. 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:

	string	exited_date <small>timestamp</small>	experience <small>double</small>
5			3.4
6			3.1
7		2016-03-28T00:00:00....	7.2
8			4.2
9		2015-10-12T00:00:00....	4.0
10			4.2
11			-1
12		2015-04-11T00:00:00....	3.8
13		2016-08-06T00:00:00....	4.2

3. It returns the transformed column as displayed below:

	string	exited_date timestamp	experience double
5			3.4
6			3.1
7		2016-03-28T00:00:00....	7.2
8			4.2
9		2015-10-12T00:00:00....	4.0
10			4.2
11		2015-04-11T00:00:00....	3.8
12		2016-08-06T00:00:00....	4.2

5.4.3.6. Fill Cells with Value

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

DATA CLEANSING

Fill cells with value...

Use with:
Other column ▼

Column:
bill_start_date ▼

Submit

- **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_datetime,' it copies the value from the 'bill_start_date' column to the 'created_datetime' column.

bill_start_date timestamp	created_datetime string
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	

converts into

bill_start_date timestamp	created_datetime timestamp
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....

5.4.3.7. 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.

Fill empty cells with text...

Use with:
Value

Value:
NA

Submit

- **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.

	source	referral_of
	string	string
81	agency	
82	drive	
83	referral	Emp ID 7
84	referral	Emp ID 2
85	portal	
86	portal	
87	internal	

converts to

	source	referral_of
	string	string
81	agency	NA
82	drive	NA
83	referral	Emp ID 7
84	referral	Emp ID 2
85	portal	NA
86	portal	NA
87	internal	NA

5.4.3.8. Flag Duplicates in Columns

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

Flag Duplicates In Columns...

Select Column
team

Submit

team	IsDuplicate_team
string	boolean
BU 6	false
BU 6	true
BU 11	false
BU 11	true
BU 7	false
BU 6	true

returns

5.4.3.9. Flag Duplicates in Tables

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

5.4.3.10. 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.

Remove Duplicates From Column...

Select Column
team

Submit

team
BU 6
BU 6
BU 11
BU 11
BU 7
BU 6

converts to

team
BU 6
BU 11
BU 7

5.4.3.11. Remove Duplicates from Table

It Removes all duplicate rows from the table.

5.4.3.12. 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.

Remove Letters...

Create new column

Submit

The selected column

Emp ID 9
Emp ID 1
Emp ID 13
Emp ID 7
Emp ID 9

converts into

9
1
13
7
9

after transformation.

5.4.3.13. Remove Numbers

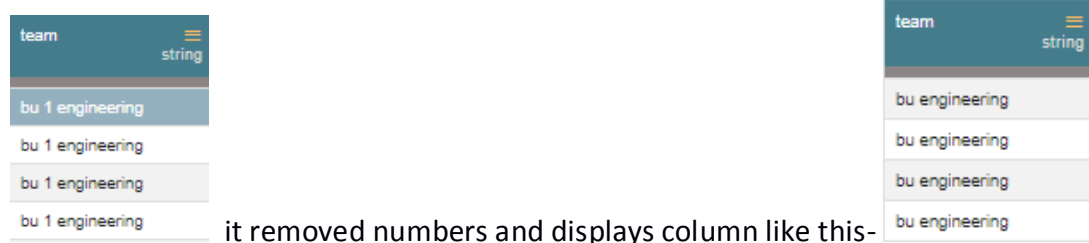
It removes any number present in the selected column. We can either add a new column with the transformed value or overwrite the same column.

Remove Numbers...

Create new column

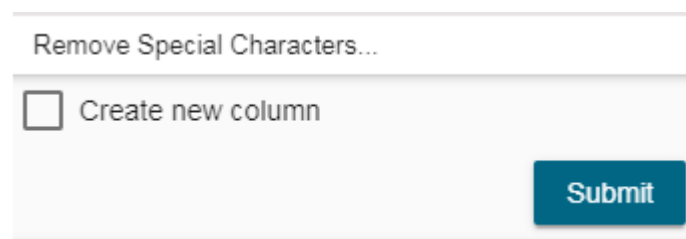
Submit

When the transform is performed on the selected column



5.4.3.14. 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 is performed on the selected column, the punctuations get removed from the column as displayed below:



5.4.4. Dates

5.4.4.1. 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.

Add Duration...

Create new column

Use with:
Other column ▼

Column:
Shot1_duration ▼

Submit

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.

Shot1_duration	Shot2_duration	
string	string	
00:00:00.000	00:00:00.033	
00:00:00.000	00:00:00.033	
00:00:01.033	00:00:01.066	
00:00:01.033	00:00:01.066	
00:00:02.033	00:00:02.066	
00:00:02.033	00:00:02.066	
00:00:02.033	00:00:02.066	
00:00:02.033	00:00:02.066	

converts to

Shot1_duration	Shot2_duration	Shot2_duration_ad...
string	string	string
00:00:00.000	00:00:00.033	00:00:00.033
00:00:00.000	00:00:00.033	00:00:00.033
00:00:01.033	00:00:01.066	00:00:02.099
00:00:01.033	00:00:01.066	00:00:02.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099

5.4.4.2. 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 datetime column of 'yyyy-mm-dd' into the 'hh:mm:ss' format.

5.4.4.3. Extract Time

Extract the time units from a selected column with a time value. The time units that can be 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

5.4.4.4. 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 a 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. Year pattern can be in the 'yy' or 'yyyy' format.
- **Day of Week:** It returns the 'day of 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)

5.4.4.5. 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

Find date difference...

Create new column

Input Format:
Month First ▼

Use with:
Value ▼

Value Hint:
Month First ▼

Value:

[Submit](#)

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	expected_joining_... integer
2017-01-02	367
2017-01-18	383
2017-01-19	384
2017-01-18	384
2017-02-15	383
2017-02-16	411
2017-02-17	412

converts to

5.4.4.6. 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 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, hyphen, comma, full stop, space) for the output date column.
- **Include Timestamp:** It will add a timestamp to the current date format if enabled with a tick mark.

Format Date...

Source Format Hint: Year First ▼	Target Format: Year First ▼
Year Pattern: yyyy ▼	Month Pattern: Jan-Dec ▼
Delimiter: / ▼	
<input type="checkbox"/> Include Timestamp	

Submit

expected_joining_... date	expected_joining_... timestamp
2017-01-02	2017/Jan/02 00:00:00
2017-01-18	2017/Jan/18 00:00:00
2017-01-19	2017/Jan/19 00:00:00
2017-01-18	2017/Jan/18 00:00:00
2017-02-15	2017/Feb/15 00:00:00
2017-02-16	2017/Feb/16 00:00:00

converts to

5.4.4.7. 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(value type) we want to subtract.

Sub Interval To Date...

Create new column

Input Format:
Month First ▼

Value Type:
Years ▼

Value:

Submit

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

expected_joining_... date	expected_joining_... date
2017-01-02	2016-09-02
2017-01-18	2016-09-18
2017-01-19	2016-09-19
2017-01-18	2016-09-18
2017-02-15	2016-10-15
2017-02-16	2016-10-16

converts to

5.4.4.8. Subtract Duration

The 'Subtract Duration' transform 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.

Time1_split_1 ☰ string	Time1_split_1_sub... ☰ string
1:00:00	00:00:00.000
2:00:00	01:00:00.000
3:00:00	02:00:00.000
4:00:00	03:00:00.000
5:00:00	04:00:00.000
6:00:00	05:00:00.000

converts to

5.4.5. Integer

5.4.5.1. Add, Multiply, Subtract or Divide

It performs the arithmetic operation on the selected numerical column.

- **Operator:** There is four arithmetic operation to choose from +, -, / and *.
- **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 be either be a value or other numerical column based on the choice of use with an option.

Price(K)	integer	Price(K)_multiply_1	integer
34		34000	
176		176000	
324		324000	
74		74000	
109		109000	
111		111000	

converts to

5.4.6. ML

5.4.6.1. 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	Screen Size_binari...	double
13.3		0.0	
13.3		0.0	
15.6		1.0	
15.4		1.0	
13.3		0.0	
15.6		1.0	
15.4		1.0	
13.3		0.0	

converts to

5.4.7. Numbers

5.4.7.1. Max

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

5.4.7.2. Mean

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

5.4.7.3. Min

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

5.4.7.4. Negate

It will complement the sign of a numeric value. If the value is positive, then a negative value will come and vice-versa.

5.4.7.5. Number Name

It will convert 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	Price (Euros)_In_...	string
34900		Thirty Four Thousand ...	
176900		One Hundred and Sev...	
324000		Three Hundred and T...	
74900		Seventy Four Thousa...	
109900		One Hundred and Nin...	

converts to

5.4.7.6. Remove Fractional Part

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

5.4.7.7. Round Value using Ceil Mode

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

suicides_per_100k...	double	suicides_per_100k...	double
6.71		6.8	
5.19		5.2	
4.83		4.9	
4.59		4.6	
3.28		3.3	
2.81		2.9	

converts to

5.4.7.8. 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.

Round value using down mode

Create new column

Precision: 0

Submit

suicides_per_100k... double

-6.71
-5.19
-4.83
-4.59
-3.28
-2.81

converts to

-6
-5
-4
-4
-3
-2

5.4.7.9. 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 nearest multiple of Specified significance. It does not consider whether 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.

Round value using floor mode

Create new column

Precision: 1

Submit

suicides_per_100k... double

6.71
5.19
4.83
4.59
3.28
2.81

converts to

6.7
5.2
4.8
4.6
3.3
2.8

5.4.7.10. Round Value using Half-up mode

It replaces a number with 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.

Round value using halfup mode	suicides_per_100k... double	suicides_per_100k... double
<input checked="" type="checkbox"/> Create new column	6.71	6.7
	5.19	5.2
Precision:	4.83	4.8
1	4.59	4.6
	3.28	3.3
<input type="button" value="Submit"/>	2.81	2.8

converts to

5.4.8. String

5.4.8.1. 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.

5.4.8.2. Change to Title Case

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

5.4.8.3. 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.

5.4.8.4. 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 will be returned starting from the given position.

5.4.8.5. Extract Substring before Delimiter

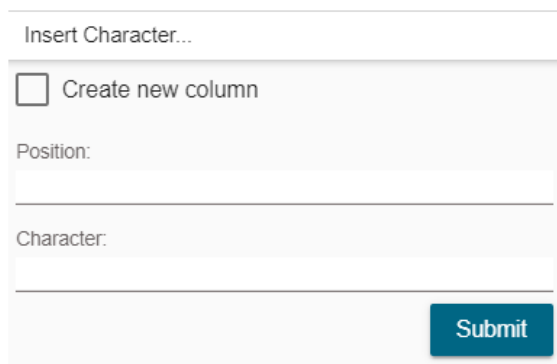
It extracts the substring from the selected column, before the 'nth' occurrence of the delimiter specified where 'n' is the count. The transformed value can replace the existing column value or can be added as a new column.

- **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

5.4.8.6. Insert Character

It inserts the character entered after 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 will insert the specified character after position 2, 4 & 6 of the cell values
- **Character:** The character that should be inserted after the specified positions



5.4.8.7. 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 return 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.

5.4.8.8. 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

5.4.8.9. 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

Remove trailing and leading characters...

Create new column

Padding character
Other ▼

Custom padding character:

Submit

5.4.8.10. 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: The value or pattern to be searched for in the selected column

Search and replace...

Create new column

Operator:
Regex ^/ ▼

Search for:

Replace with:

Overwrite entire cell

Submit

5.4.8.11. Split String

It splits the string based on condition. It will give 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.'

Split String...

Use with:

Delimiter ▼

Separator:

Submit

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

age	string	age	string	age_split_1	string	age_split_2	string
15-24 years		15-24 years		15-24		years	
35-54 years		35-54 years		35-54		years	
15-24 years		15-24 years		15-24		years	
75+ years		75+ years		75+ y		ears	
25-34 years		25-34 years		25-34		years	
75+ years		75+ years		75+ y		ears	
35-54 years		35-54 years		35-54		years	

converts to

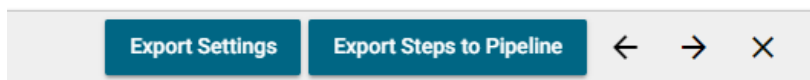
5.5. Steps



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

Profile	Transforms	Steps: 6
		1:SEARCH_AND_REPLACE on gender
		2:SEARCH_AND_REPLACE on gender
		3:SEARCH_AND_REPLACE on expected_joining_date
		4:CHANGE_TO_TITLE_CASE on joining_status
		5:ROUND_TO_CEIL_MODE on experience
		6:REMOVE_FRACTIONAL_PART on offered_ctc

6. Navigation Pane

The navigation pane provides an option to export the data, move out of the BDB Data Preparation and Perform Undo or Re-do options.



- a. **Export Settings:** Export settings provides an option to specify the elastic into which the cleansed data must be moved.
- b. **Export Steps to Pipeline:** This button 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.
- c. **Undo**  : Undo a list of last few transforms. This button will be enabled only if, we have applied some transform on the data.
- d. **Redo**  : Redo a list of last few transforms, that was undone. If we have not undone any transform, then the 'redo' icon will be disabled.
- e. Close the Preparation: We will 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.

7. Signing Out

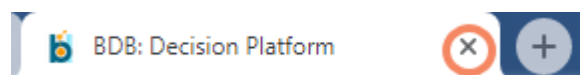
The users can Sign-out from the Data Preparation tab at any given stage, but preferable is that the users should complete all the preparation tasks they wish to perform and save it before closing the tab or signing out from the Platform.

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


1. Closing the BDB Data Preparation

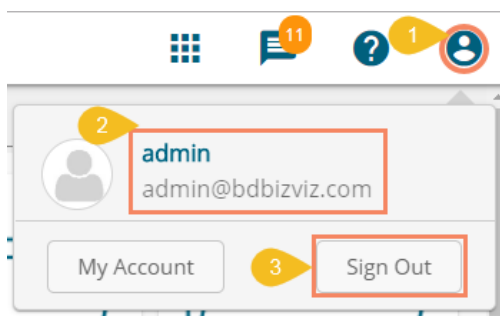
Once you have completed the Data Preparation tasks, save your work and close the Data Preparation tab.

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



2. Sign Out from the BDB Platform

- i) Click the '**User**' icon  on the Platform homepage.
- ii) A menu appears with the logged in user details (User's name and email id).
- iii) Click '**Sign Out.**'



iv) Users successfully log out from the **BDB Platform**.

Note: Clicking on the '**Sign Out**' option will redirect the user back to the login page of the BDB platform.