

**BDB Data Preparation- High Level Introduction**

**AI, Data Science & Advanced Analytics Platform**

**Unique Disruptive Platform by Ex-SAP BO R&D Team**

**One Platform, End to End Solution**

**Removes industry pain points**

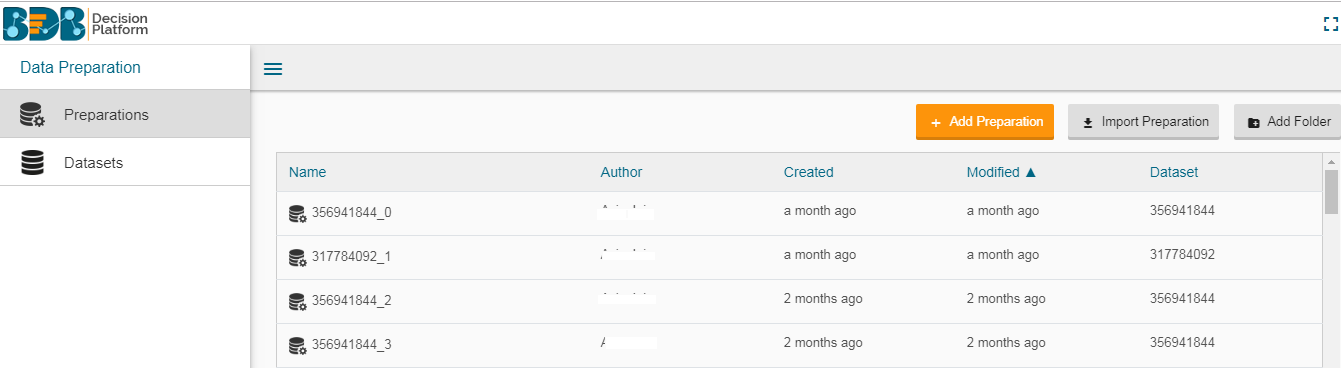
**Leads to exponential benefits realization**

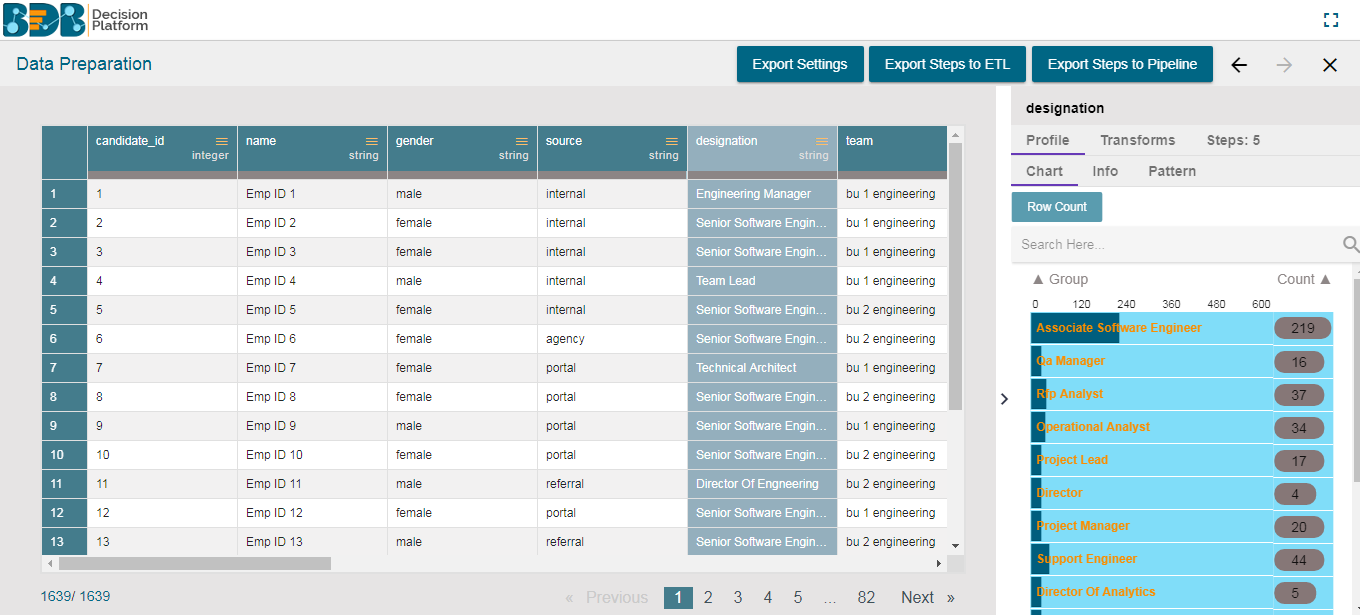
BDB Data Preparation-

**BDB Data Preparation is a self-service data cleansing tool for cleaning and transforming raw data prior to processing and analysis.**

This tool lets you add data sets for transformation, view your data preparations. It profiles your data into a grid like format, the data displayed in the grid is a sample from the actual data set or complete data based on the data volume. You can perform 50+ transformations on this data set, visualize profile chart, info & pattern of the data.

The transformation recipe prepared in BDB data prep can be exported in BDB ETL & BDB Data Pipeline.





**BDB Data Preparation Purpose-**

##### **Vigorous Data Grounding**

##### **Speed Up Insight Discovery**

##### **Reliable Self-Service Access**

##### **Transforms Raw Data into Meaningful Data**

**High-Level Features of BDB Data Preparation**

* View sample data (first 10 K) in a paginated grid.
* Get statistical profile of the data
* A quality bar that indicates the percent of valid, invalid and blank rows.
* 50+ transforms that can be performed on the data.
* View the changes in data, after each transform.
* Undo/redo the transforms if changes are not acceptable.
* Write nested transforms using SQL transform.
* View the list of transforms performed on the current data.
* Filter data by clicking on the profiling charts.
* Perform transforms on the filtered data.
* Export the steps to ETL /Pipeline so that it can be performed on full data.

**Deep-Dive into Features**

As, BDB Data Prep profiles data set into a grid format, below are the properties of data grid and how it functions –

# 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.

## Data Grid Header

The grid has a header which displays the column name from the dataset.

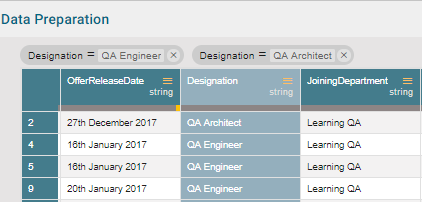
## Data Types

The BDB Data Preparation supports the following data types:

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

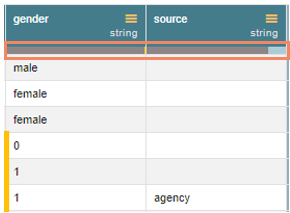
## Panel to List the Selected Filters.

When a filter is selected, it gets added to the filter panel on top of the grid.



## 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

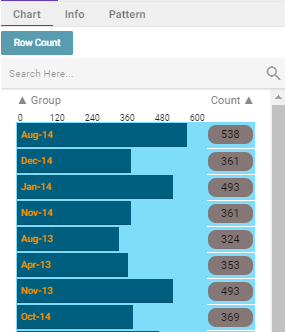
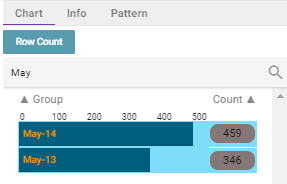


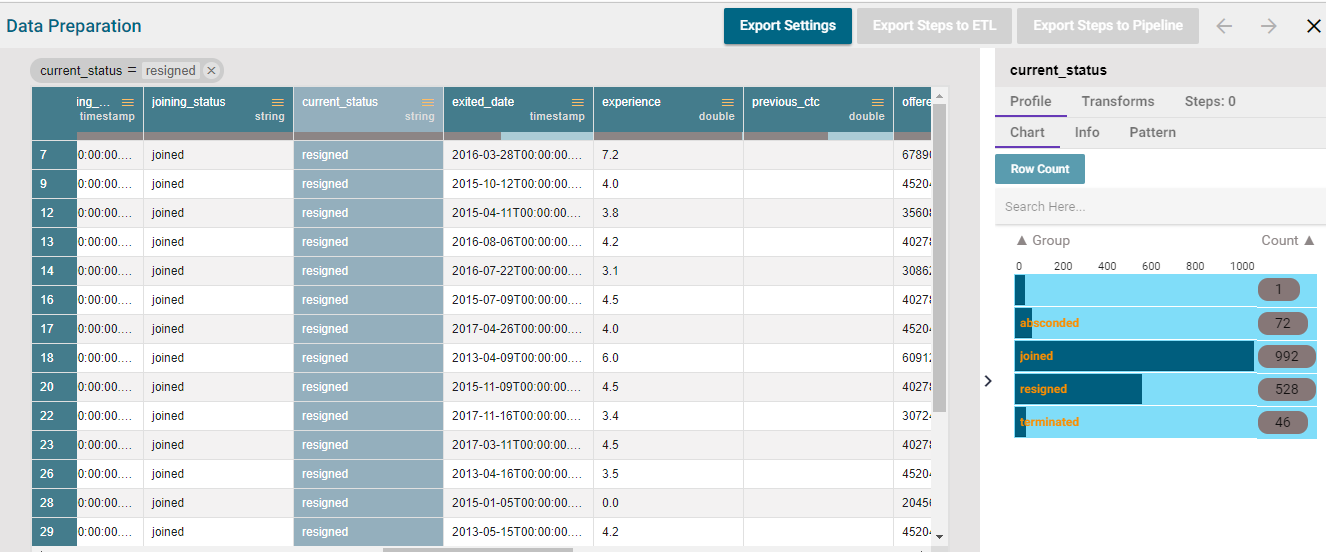
# Summary Pane

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

## Charts

The in-built charts (Column and Bar charts) display the occurrence of each value.

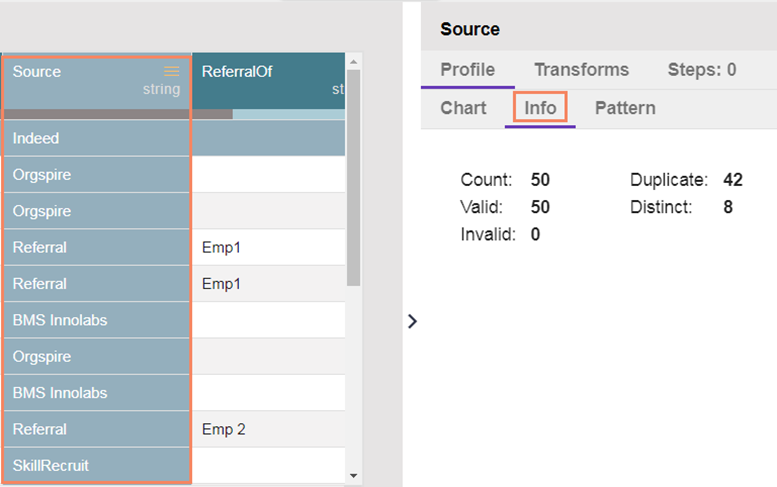
 



## Info: Value/Statistics

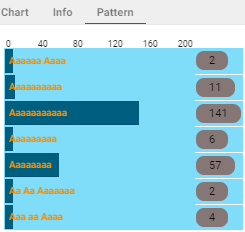
The information tab displays value or statistics of the data. The following aspects are displayed about the chosen data when the column is of string type:

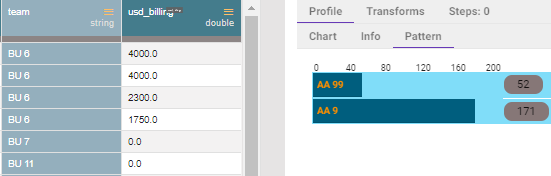
* Count of Rows
* Count of Duplicates
* Count of Valid Data
* Distinct Values
* Count of Invalid Data



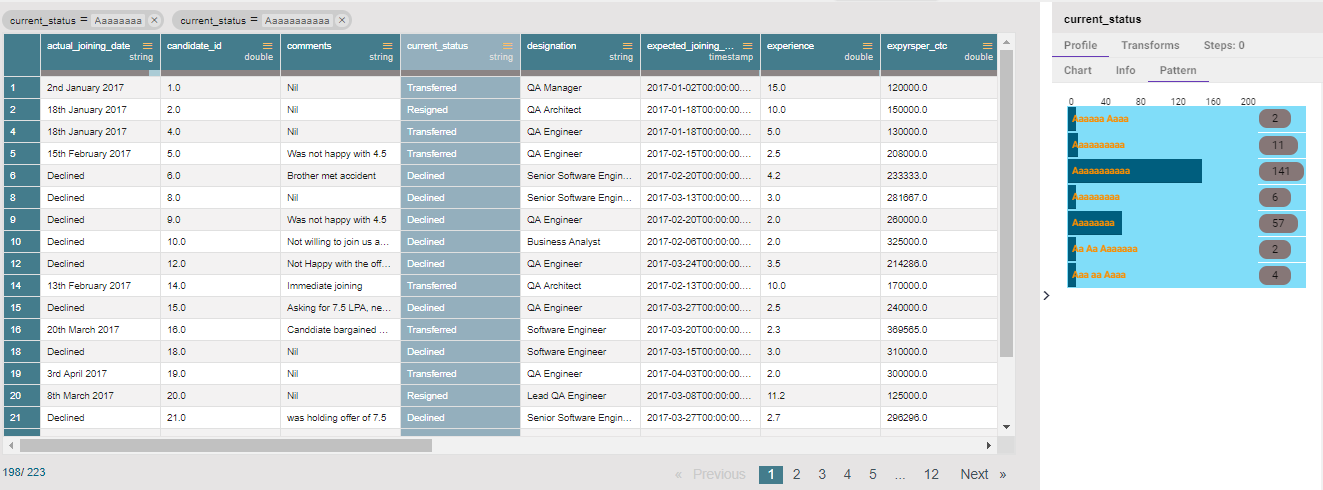
## Pattern

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





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



## 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.

### Columns

#### Cast to Types

It is a table-based operation. Cast to type can remove the value with the invalid data type.

#### 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.

#### Concatenate with

The users can concatenate a column value with some other column or with some prefix/suffix.

#### Delete Column

It deletes any selected column.

#### Duplicate Columns

It will create a duplicate of the selected column.

#### Generate Primary Key

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

#### Return Non-Null Column Values

The transform returns the first non-null value from the list of columns specified to a new column.

#### SQL Transform

This transform allows us to write SQL Query against the table as we can write in any SQL editor.

### Conversions

#### Convert Duration

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

### Data Cleansing

#### 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.

#### 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.

#### Delete Rows with Empty Cell

1. The transform deletes any row which has a blank value in the selected column.

#### Delete Rows with Invalid Cell

1. The transform deletes any row which has invalid value in the selected column.

#### Delete Rows with Negative Values

1. It deletes the rows which have a negative value in the selected column.

#### Fill Cells with Value

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

#### 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

#### Flag Duplicates in Columns

This transform adds a new Boolean column based on duplicate values in the column.

#### Flag Duplicates in Tables

This transform adds a new Boolean column based on duplicate rows in the table.

#### 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 Table

It Removes all duplicate rows from the table.

#### 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 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 Special Characters

It removes any special character present in the selected column. Only letters, numbers and spaces are retained.

### Dates

#### Add Duration

The transform adds two-time values. The transform supports adding time into

‘**hh:mm:ss.mmm’** and ‘**hh:mm:ss’** formats.

#### Add Interval to Date

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

#### 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.

#### Extract Date

It extracts the date part from a selected column with a date value

#### 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.

#### Format Date

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

* **Source Format Hint**
* **Target Format**
* **Year Pattern**
* **Month Pattern**
* **Delimiter**
* **Include Timestamp**

#### Sub Interval to Date

The ‘Sub Interval to Date’ transform subtracts specified value(interval) from the given date column.

#### Subtract Duration

The ‘Subtract Duration’ subtracts a duration

### Integer

#### Add, Multiply, Subtract or Divide

It performs the arithmetic operation on the selected numerical column.

### ML

#### 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.

### Numbers

#### Max

It gives the maximum value from the selected columns row-wise

#### Mean

It gives the average value of the selected columns row-wise.

#### Min

It gives the minimum value from the selected columns row-wise.

#### Negate

It will complement the sign of a numeric value.

#### Number Name

It will convert the value of the selected column into words.

#### Remove Fractional Part

It removes the fractional part from the numerical column.

#### Round Value using Ceil Mode

It replaces the number with a greater integer value if the number is between two integer value

#### 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.

#### 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.

#### Round Value using Half-up mode

It replaces a number with next integer value if its next digit is 5 or greater than 5.

### String

#### Change to lower case

It converts the selected column value to the small case.

#### Change to Title Case

It converts the selected column value to title case.

#### Change to Upper Case

It converts the selected column value to capital letters.

#### Extract Substring at Position

It extracts the substring from the selected column based on the starting position and the length of the extract.

#### 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.

#### Insert Character

It inserts the character entered after specified position.

#### Remove Consecutive Characters

The transform removes the repeated whitespace or character

#### Remove Part of Text

It matches and removes the matching part or entire value based on the condition.

#### Remove Trailing and Leading Characters

It removes trailing and leading characters from the column.

#### Search and Replace

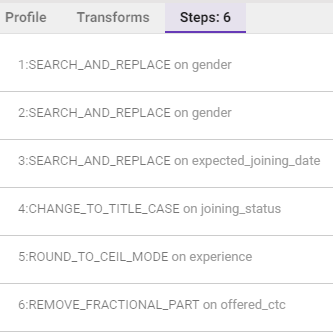
It searches and replaces the matching part or entire value based on the selected option.

#### Split String

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

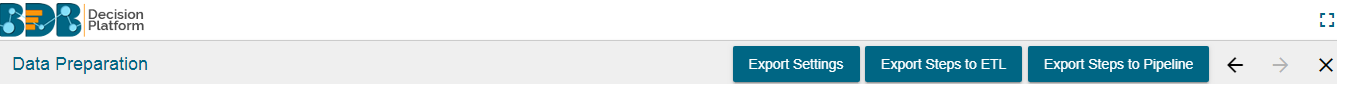
## Steps

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



# Export Options

The navigation pane lets you export the prepared data to ETL, BDB Data Pipeline, move out of the BDB Data Preparation and Perform Undo or Re-do options.



* *Export to ETL*

As BDB Data Preparation is embedded in ETL, you can perform number of transformations via data prep console and export the prepared data to ETL workflows

* *Export to BDB Data Pipeline*

You can also export the transformed data in BDB Data Pipeline where this data transformation recipe can be used as a data prep model(component) in Pipeline workflow. This component runs the same set of transformation on incoming data in pipeline which were performed during its preparation in BDB data prep.