Quick Tutorial All Pipeline Workflows

Contents

BDB 8.2: Data Pipeline Workflow – 1

Jobs: Create a job to migrate data from source to destination with basic data transformation.

Workflow 1 allows you to connect to the Data Center Plugin and Data Pipeline Plugin to create a job for migrating data from a sandbox file to the ClickHouse database with data transformations.

In this job, you can start by uploading your data file to the Data Center Plugin's data sandbox. The sandbox provides a secure environment for managing and working with data files. Once the file is uploaded, you can utilize the Data Pipeline Plugin to configure the necessary steps for data migration and transformation.

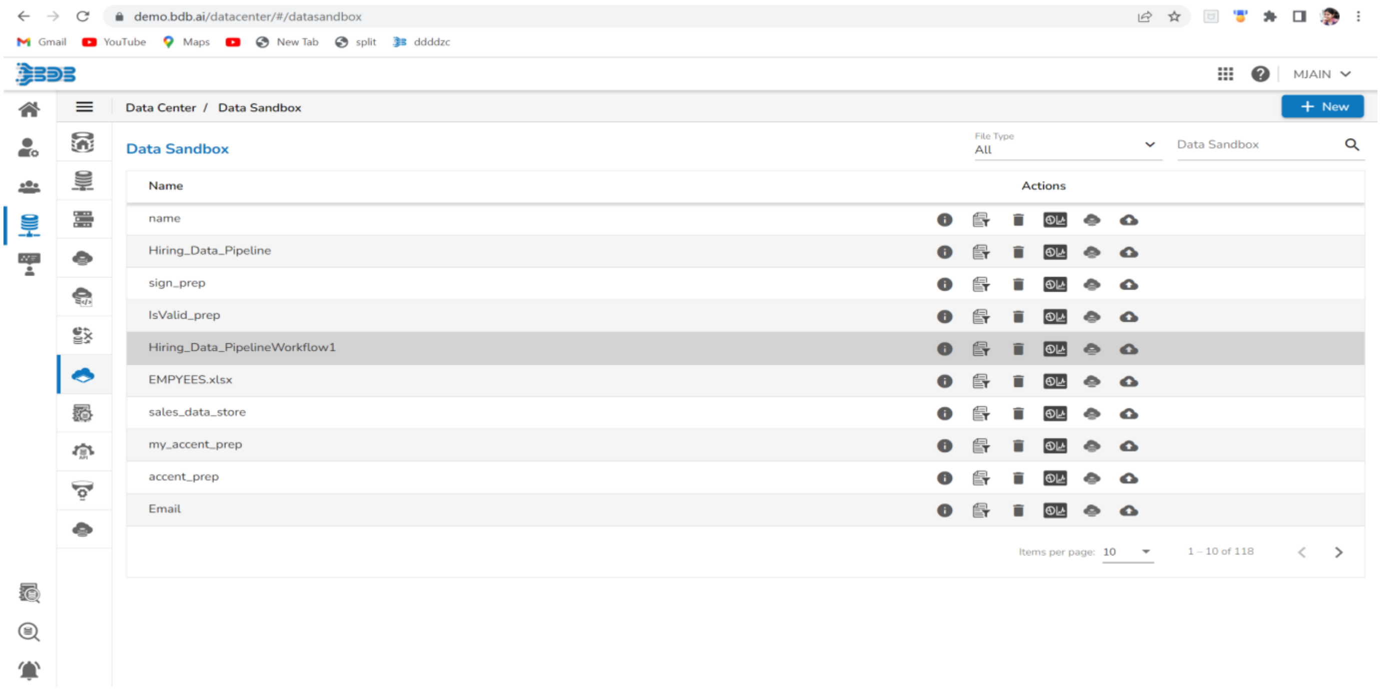
The Data Pipeline Plugin offers a range of components that enable you to read data from the sandbox file, apply various transformations, and then write the transformed data to the ClickHouse database.

Upload File in Sandbox:

Let's Focus on the process of creating a new data sandbox and performing tasks on the data

To begin,

- click on the 'Apps' menu, which will display a list of available modules.
- From the menu, choose the 'Data Center' module. This will take you to the Data Center page, where you can manage your data.
- Now, let's navigate to the Sandbox section, which is dedicated to managing sandboxes for your data.
- Look for the 'Sandbox' section or tab within the Data Center page.
- Once you've found the Sandbox section, you should see an option to create a new sandbox. Click on that.
- Now, let's give the sandbox a name. Choose a name that will help you identify it later. For example, let's name it 'Hiring_data_PipelineWorkflow1'.
- Provide an appropriate description if needed. This can help provide more context or details about the purpose of the sandbox.
- Next, you'll need to choose your data file. Look for an option to upload or select a data file for the sandbox.



Data Center / Data Sandbox

Data Sandbox

File Type: All

Data Sandbox

Name	Actions
name	
Hiring_Data_Pipeline	
sign_prep	
IsValid_prep	
Hiring_Data_PipelineWorkflow1	
EMPYEES.xlsx	
sales_data_store	
my_accent_prep	
accent_prep	
Email	

Items per page: 10 1 – 10 of 118

Once you've uploaded the data file to the sandbox, you're ready to proceed with creating jobs to perform specific tasks or operations on the data.

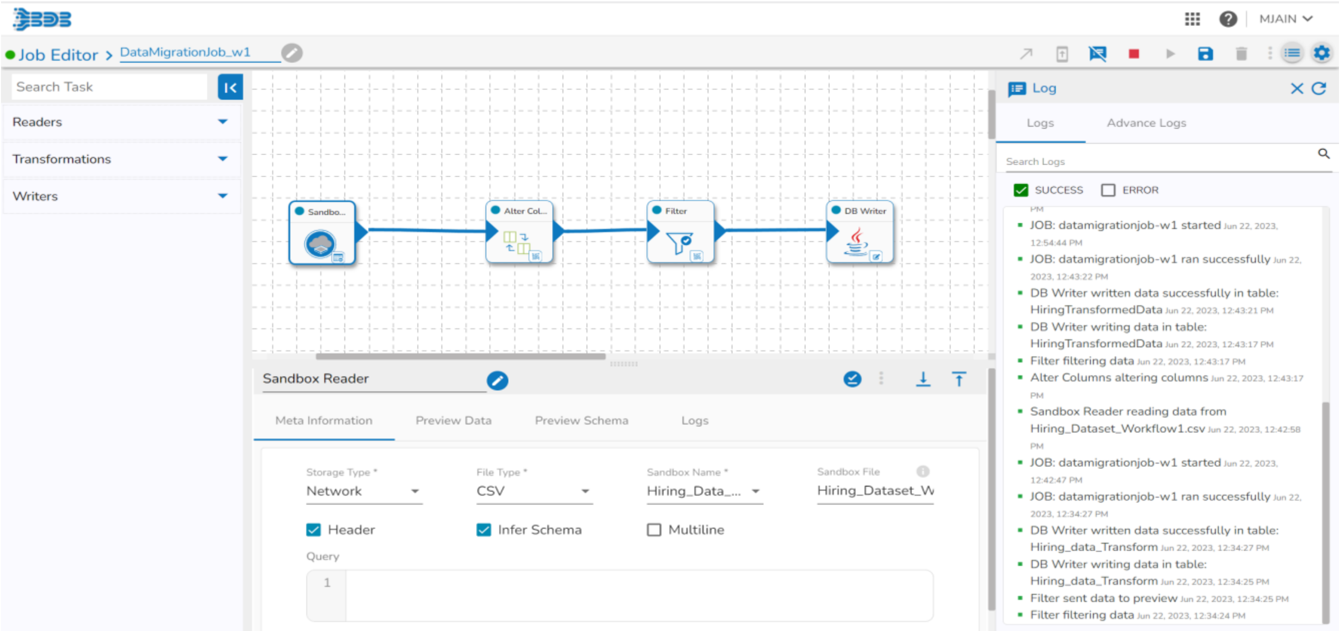
Let's walk through the steps to create and configure the job:

- Start by navigating to the Data Pipeline Plugin from the app menu. This will take you to the pipeline home page.
- Look for the create icon and click on it to create a new job. Click on the '+' icon to proceed.
- In the new job window, specify a suitable name for your job, such as 'DataMigrationJob_W1'. Provide a brief description of the workflow, such as 'Migrate data from source to destination'.
- In the job base info section, select the Spark job option to leverage the power of Spark for efficient data processing.
- Choose the appropriate resource allocation for your job based on the volume and velocity of data. Options include Low, Medium, and High-end configurations.
- If you want to schedule the job to run automatically, select the 'isSchedule' checkbox and specify the desired timestamp. Note that the job must be scheduled according to UTC.
- Just uncheck IsSchedule Checkbox and Once you have configured the job details, click on the save button to save your job.

Congratulations! You have successfully created the job.

Now, let's proceed with configuring the components within the job.

- Start by configuring the Reader component to extract data from the source. Locate the Sandbox Reader component in the Reader section and drag and drop it onto the canvas or workspace.
- Click on component and configure all the mandatory Fields, select the storage type as 'network' since you'll be accessing the data source over the network. Choose 'CSV' as the file type.
- From the dropdown menu, select the appropriate sandbox name that corresponds to your data source. Check the 'Header' option if your CSV file contains a header row with column names. Enable the 'Infer Schema' checkbox to automatically determine the schema based on the data in the CSV file. Save the component once configured.
- Next, apply transformations to the source data using the Alter Columns Transform. Drag and drop it onto the canvas and connect it with the Sandbox Reader component.
- Configure the Alter Transform component to modify column names as needed. You can rename columns and specify their types using the configuration window. Let me rename gender column and specify alias name and type of the column.
- Also rename monthly salary and specify alias name and select type of the column from the dropdown and save the component after configuring each column modification.



Storage Type * Network
File Type * CSV
Sandbox Name * Hiring_Data_...
Sandbox File Hiring_Dataset_W
☑ Header
☑ Infer Schema
☐ Multiline
Query

☑ SUCCESS ☐ ERROR

```
PM
- JOB: datamigrationjob-w1 started Jun 22, 2023, 12:54:44 PM
- JOB: datamigrationjob-w1 ran successfully Jun 22, 2023, 12:43:22 PM
- DB Writer written data successfully in table: HiringTransformedData Jun 22, 2023, 12:43:21 PM
- DB Writer writing data in table: HiringTransformedData Jun 22, 2023, 12:43:17 PM
- Filter filtering data Jun 22, 2023, 12:43:17 PM
- Alter Columns altering columns Jun 22, 2023, 12:43:17 PM
- Sandbox Reader reading data from Hiring_Dataset_Workflow1.csv Jun 22, 2023, 12:42:58 PM
- JOB: datamigrationjob-w1 started Jun 22, 2023, 12:42:47 PM
- JOB: datamigrationjob-w1 ran successfully Jun 22, 2023, 12:34:27 PM
- DB Writer written data successfully in table: Hiring_data_Transform Jun 22, 2023, 12:34:27 PM
- DB Writer writing data in table: Hiring_data_Transform Jun 22, 2023, 12:34:25 PM
- Filter sent data to preview Jun 22, 2023, 12:34:25 PM
- Filter filtering data Jun 22, 2023, 12:34:24 PM
```

- Now, Drag and drop the Filter Transform onto the canvas and connect it with the Alter Transform component. Configure the Filter component to filter the data based on specific conditions, such as monthly salary greater than a certain value and choose type of the column from the column type dropdown.

- Once the data transformation steps are configured, select the appropriate Writer component for the destination. Drag and drop the DB Writer component onto the canvas and connect it with the previous component.
- Configure the DB Writer component with the necessary connection details for your ClickHouse database.
- specify host, Port number, username, password, database name.
- Specify the destination table or collection where you want to write the transformed data.
- select driver clickhouse from the driver dropdown.
- Choose "Append" from the save mode dropdown menu.
- Click on the "Save" button to save the configured settings.
- Great job! You have successfully configured the DB Writer component for your ClickHouse database. It's now ready to write the transformed data to the specified destination table or collection in ClickHouse

Note: For DB Credentials you can contact Devops or internal BDB Team

Update and Execute JOB

- Click on the 'Update Job' icon on the job interface to save the entire job Workflow.
- Now, click on deployment mode then confirmation message will appear just click on yes then Job execution process will start. it's important to monitor its progress and ensure that the components are running successfully.

Here's how you can do it:

  - Navigate to the logs section. Look for the Log Panel and click on it to access the advanced logs. you will see the pods associated with each component in advance log section.
  - Pods are containers that hold the execution environment for the job. Check if the pods for each component have come up and are running. This indicates that the components are successfully deployed and ready to execute their tasks.

To verify that the job has executed successfully, and the data has been read, transformed, and written to the ClickHouse table, follow these steps:

- Go to the logs section and check the logs for any errors or issues. The logs will provide detailed information about the execution process.
- Look for a confirmation message in the logs indicating that the job ran successfully and completed the data migration process.

If you want to preview the processed data

- Click on Component
- Navigate to the data preview option inside the component. This will display a sample of the records that the component has produced as its output during the job execution.

Analyse the sample records to verify that the component has generated the expected output. You can examine the data structure, values, and any applied transformations or filters.

Hope you will be able to create your own Job Workflows. Thankyou!

BDB 8.2: Data Pipeline Workflow – 2

To Read Hiring CSV/Excel files from an SFTP location, split the files, ensuring data quality through data preparation, leveraging AutoML techniques to extract valuable insights, and storing the processed data into a databases (ClickHouse and MongoDB)

In today's fast-paced hiring environment, managing and analyzing real-time hiring data is crucial. That's why we've designed a powerful workflow to streamline the process of ingesting and processing hiring data from SFTP. Our workflow begins with the SFTP Monitor component. This component allows us to pick up hiring data files in CSV and Excel formats directly from the SFTP location. Once the files are picked up, we use the File Splitter component to split them based on their file formats. This allows us to handle CSV and Excel files separately for further processing.

Next, we utilize the SFTP Reader component to read the content of the files. We have both the SFTP Reader and SFTP Stream Reader components available to handle different scenarios.

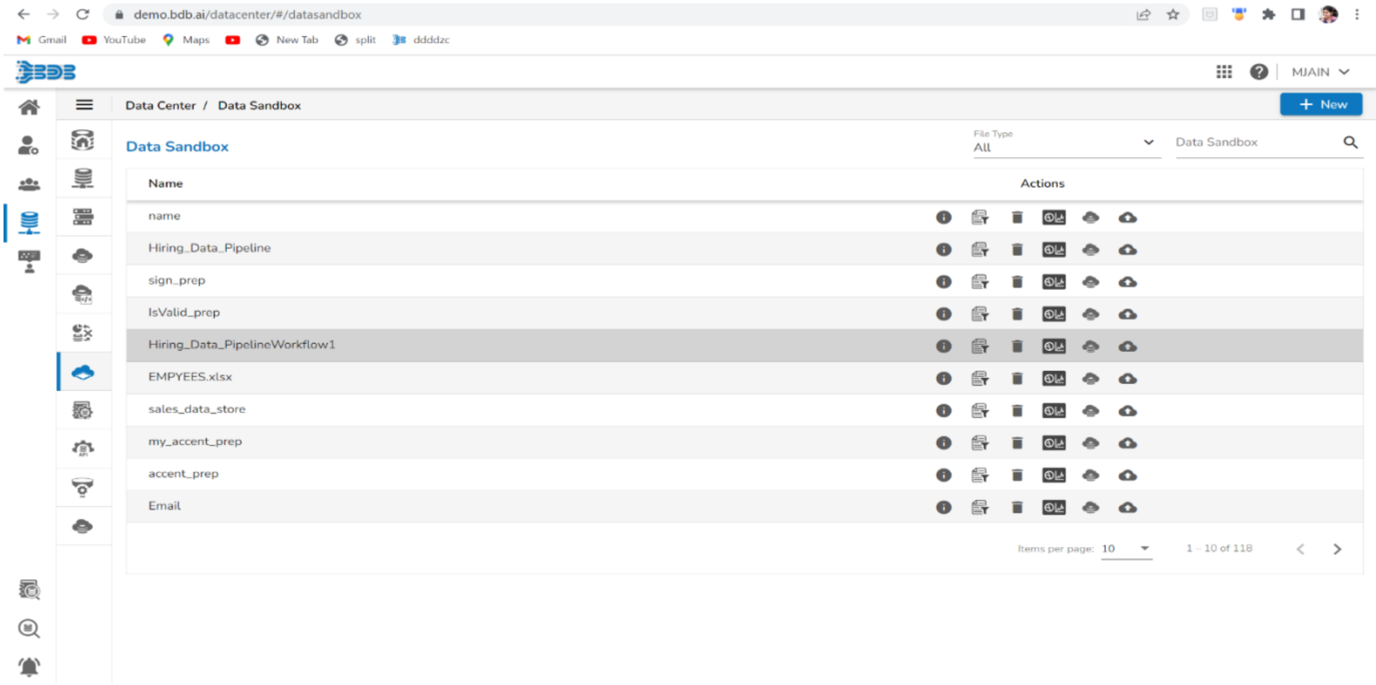
To ensure the data is clean and ready for analysis, we incorporate the Data Preparation plugin. This powerful tool enables us to transform and cleanse the hiring data, including job listings, candidate profiles, and recruitment metrics.

Let's Focus on the process of creating a new data sandbox and performing tasks on the data.

To begin,

- click on the 'Apps' menu, which will display a list of available modules.
- From the menu, choose the 'Data Center' module. This will take you to the Data Center page, where you can manage your data.
- Now, let's navigate to the Sandbox section, which is dedicated to managing sandboxes for your data.
- Look for the 'Sandbox' section or tab within the Data Center page.
- Once you've found the Sandbox section, you should see an option to create a new sandbox. Click on that.

- Now, let's give the sandbox a name. Choose a name that will help you identify it later. For example, let's name it 'Hiring_data_PipelineWorkflow1'.
- Provide an appropriate description if needed. This can help provide more context or details about the purpose of the sandbox.
- Next, you'll need to choose your data file. Look for an option to upload or select a data file for the sandbox.



Data Center / Data Sandbox

Data Sandbox

File Type: All

Data Sandbox

Name	Actions
name	
Hiring_Data_Pipeline	
sign_prep	
IsValid_prep	
Hiring_Data_PipelineWorkflow1	
EMPYEES.xlsx	
sales_data_store	
my_accent_prep	
accent_prep	
Email	

Items per page: 10   1 – 10 of 118

Great! You've successfully uploaded the data file to the sandbox. Now, let's proceed to the next step, which is Preparation. In this step, we'll prepare the data for further analysis or processing.

Create Preparation Using Data Preparation:

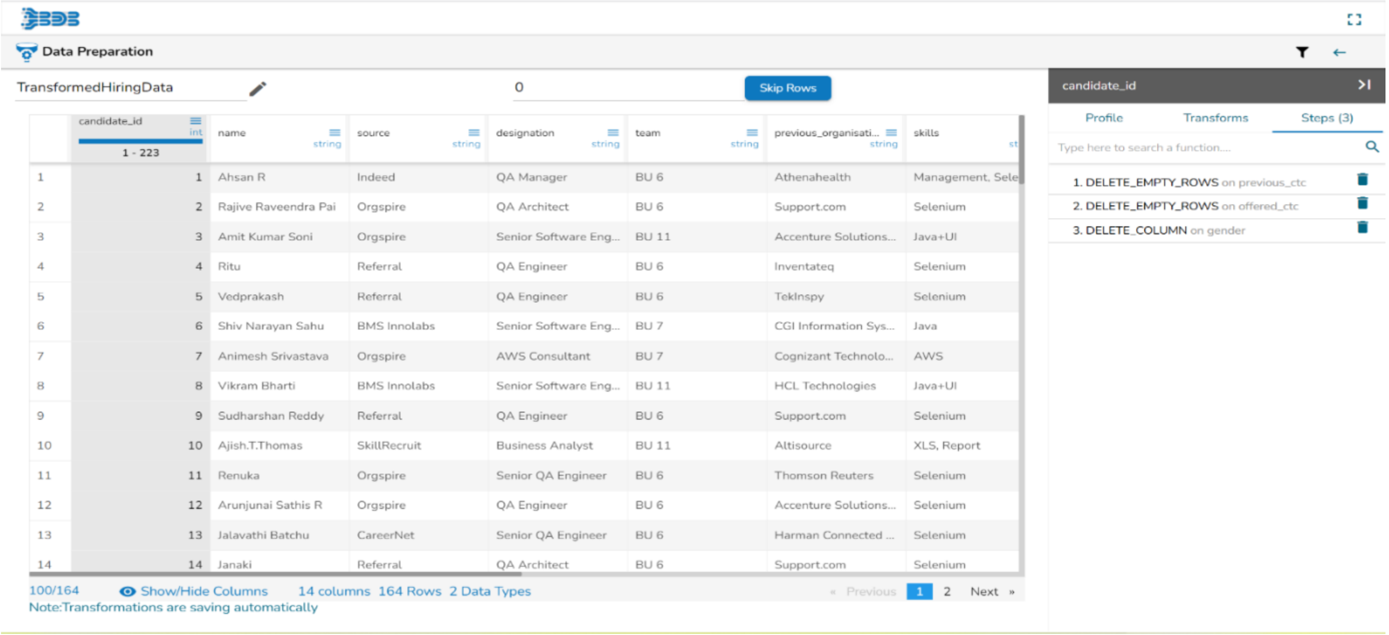
Data Preparation is used to clean the data.

Are you ready to prepare your hiring data pipeline sandbox file? Let's get started with the data preparation process.

- Select Data Preparation Icon. will navigate to the Data preparation home page.
- Here on the Data Preparation home page, you can see your complete dataset displayed in a grid form. The Data Preparation Plugin automatically profiles the data,
- providing valuable insights into its characteristics and detecting any anomaly data. You can also view the data profiling details.
- On the right-hand side, you'll find the selected column's profile. Here, you can explore various details such as charts, information, and patterns associated with the selected column.
- All the Transformations will appear inside transform tab.

Transforms

- Now, let's start preparing and cleaning the data. To remove the empty cells in the 'Previous CTC' column, just select prevoius_ctc column and
- navigate to the 'Transforms' tab and search for the 'Delete Empty Rows with Empty Cells' transform. Click on it to remove all the empty rows from the 'Previous CTC' column."
- You can see that the empty rows in the 'Previous CTC' column have been removed.
- Next, let's perform the same transformation on the 'Offered CTC' column.
- Select the column and search for the 'Delete Rows with Empty Cell' transform. Click on it to remove the empty rows from the 'Offered CTC' column.
- Great! The empty rows in the 'Offered CTC' column have been successfully removed."
- Now, let's delete the 'Gender' column from the dataset. Simply select the column and search for the 'Delete Column' transform. Click on it to remove the 'Gender' column.
- Perfect! The 'Gender' column has been deleted from the dataset
- You can see that all the performed transforms are recorded in the 'Steps' section. This helps you keep track of the changes made to the dataset.
- Now, let's rename the preparation for identification purposes. Simply click on the edit icon and give it a new name, such as 'TransformedHiringData
- Great! The preparation has been renamed to 'TransformedHiringData.



TransformedHiringData

0

	candidate_id int 1 - 223	name string	source string	designation string	team string	previous_organisati... string	skills
1	1	Ahsan R	Indeed	QA Manager	BU 6	Athenahealth	Management, Sele
2	2	Rajive Raveendra Pai	Orgspire	QA Architect	BU 6	Support.com	Selenium
3	3	Amit Kumar Soni	Orgspire	Senior Software Eng...	BU 11	Accenture Solutions...	Java+UI
4	4	Ritu	Referral	QA Engineer	BU 6	Inventateq	Selenium
5	5	Vedprakash	Referral	QA Engineer	BU 6	TekInspy	Selenium
6	6	Shiv Narayan Sahu	BMS Innolabs	Senior Software Eng...	BU 7	CGI Information Sys...	Java
7	7	Animesh Srivastava	Orgspire	AWS Consultant	BU 7	Cognizant Technolo...	AWS
8	8	Vikram Bharti	BMS Innolabs	Senior Software Eng...	BU 11	HCL Technologies	Java+UI
9	9	Sudharshan Reddy	Referral	QA Engineer	BU 6	Support.com	Selenium
10	10	Ajish.T.Thomas	SkillRecruit	Business Analyst	BU 11	Altisource	XLS, Report
11	11	Renuka	Orgspire	Senior QA Engineer	BU 6	Thomson Reuters	Selenium
12	12	Arunjunai Sathis R	Orgspire	QA Engineer	BU 6	Accenture Solutions...	Selenium
13	13	Jalavathi Batchu	CareerNet	Senior QA Engineer	BU 6	Harman Connected ...	Selenium
14	14	Janaki	Referral	QA Architect	BU 6	Support.com	Selenium

100/164 Show/Hide Columns 14 columns 164 Rows 2 Data Types « Previous 1 2 Next »

Note:Transformations are saving automatically

1. DELETE_EMPTY_ROWS on previous_ctc
2. DELETE_EMPTY_ROWS on offered_ctc
3. DELETE_COLUMN on gender

Click on the back icon, and the preparation will be automatically saved and exported to different plugins, such as the Data Pipeline or AutoML. You can choose to export the transformed data to various plugins, such as the Data Pipeline or AutoML, based on your needs.

Create AutoML Workflow using DS Lab

To create an Automl Experiment using the DS Lab Plugin, follow these steps:

- Open the DS Lab Plugin from the app menu.
- Navigate to the DS Lab home page where all existing projects are displayed.

Create Project:

- Click on the "Create Project" button to proceed to the project creation page.
- On the project creation page, provide the following mandatory fields:
- Project Name: Enter a name for your project. For example, "Hiring Data".
- Project Description: Add a description for your project, such as "Creating AutoML experiment for hiring data prediction".
- Select Algorithm Types: Choose the algorithm types that are required for your project from the available options in the dropdown menu.
- Select Environment: Choose the environment for your project. In this case, select "Python TensorFlow".
- Specify Resource Allocation: Select the resource allocation based on the volume of data and the computational requirements of your project. Choose the appropriate option, such as "Medium".
- Select Idle Shutdown Time: Choose the duration after which the system should shut down if idle. For example, select "1 hr.".
- Specify External Libraries: If you require any external libraries for your project, mention them in this field.
- Mention GPU Type: If your project requires GPU acceleration, specify the GPU type.
- Once you have filled in all the necessary details, click on the "Save" button to create the project.

You should see a message confirming that the project has been successfully created.

Create AutoML Experiment

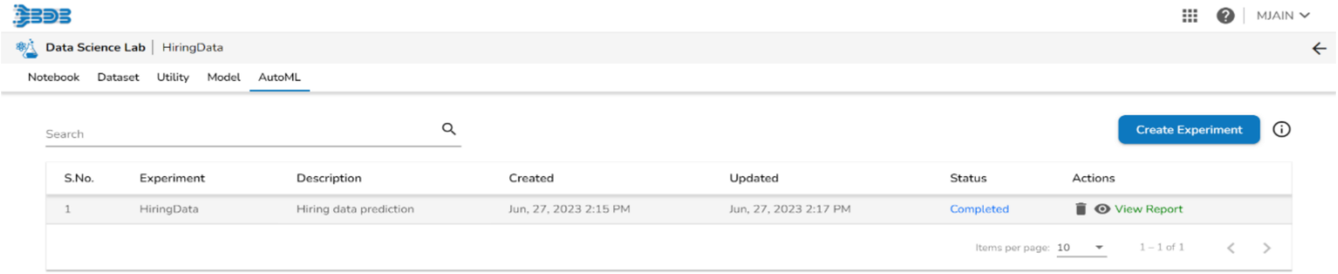
To create an AutoML experiment using the recently created "Hiring Data" project and follow these steps:

- Navigate to the "Datasets" tab in Data Science Lab.
- Click on the "Add Dataset" button.
- Select the "Data Sandbox" option from the data source dropdown menu. This will display all the files that have been uploaded into the sandbox.
- Choose the "Hiring_data_Pipeline" file.
- Click on the "Add" button to successfully add the dataset.

- Many action items will be displayed for the selected dataset, such as preview, data profile, create experiment, delete, and data preparation.
- Choose the "Create Experiment" option for the uploaded data file.
- Provide an experiment name, for example, "HiringData," and a description, such as "Hiring data prediction."
- Select the target column from the dropdown menu. You can also choose any data preparation options if required. Click on the "Next" button to proceed.
- Select the experiment type based on your requirements, such as classification or regression. In this case, select "Classification" as the gender column contains categories.
- Click on "Done" to start the experiment.
- Wait for the experiment to complete in the AutoML tab. You can track the progress as the status changes from "Started" to "Running."
- Once the experiment is completed, open and view the report. The report will display the recommended model and a run summary.

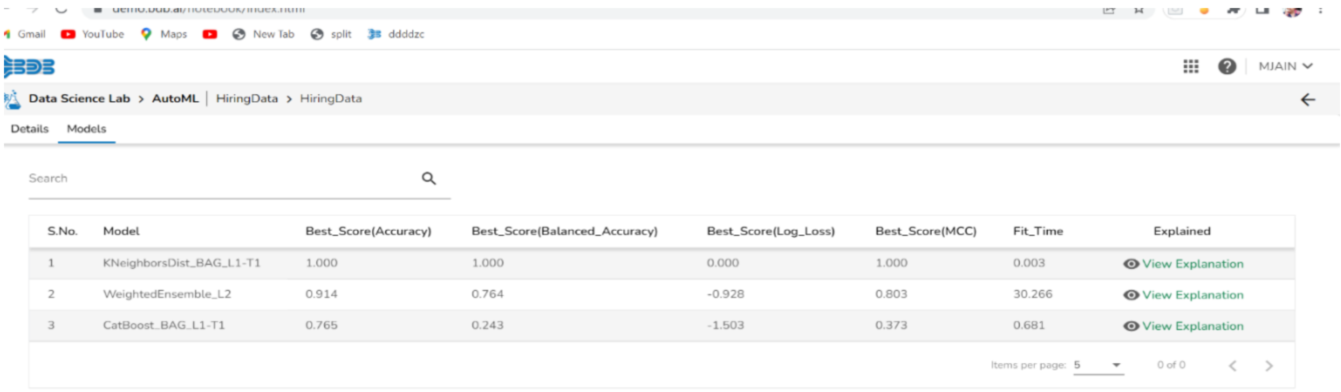
Register Model

- Switch to the "Models" tab to explore the top three models trained using AutoML.
- Click on "View Explanation" to understand the model's performance and switch to the "Dataset Explorer" tab to see the data profile.
- Navigate to the "Models" tab and expand the saved models.
- Export the models to the data pipeline using the register icon.
- Optionally, export the model into Git for version control and collaboration purposes.



S.No.	Experiment	Description	Created	Updated	Status	Actions
1	HiringData	Hiring data prediction	Jun, 27, 2023 2:15 PM	Jun, 27, 2023 2:17 PM	Completed	View Report

Items per page: 10 1 – 1 of 1


demo.bdb.ai/notebook/index.html
Gmail
YouTube
Maps
New Tab
split
ddddzc
MJAIN
Data Science Lab > AutoML | HiringData > HiringData
Details
Models
Search
Items per page: 5
0 of 0

S.No.	Model	Best_Score(Accuracy)	Best_Score(Balanced_Accuracy)	Best_Score(Log_Loss)	Best_Score(MCC)	Fit_Time	Explained
1	KNeighborsDist_BAG_L1-T1	1.000	1.000	0.000	1.000	0.003	View Explanation
2	WeightedEnsemble_L2	0.914	0.764	-0.928	0.803	30.266	View Explanation
3	CatBoost_BAG_L1-T1	0.765	0.243	-1.503	0.373	0.681	View Explanation

By following these steps, you should be able to create an AutoML experiment using the "Hiring Data" project and explore the recommended models in Data Science Lab.

Pipeline Workflow

Now Understand the process of creating a pipeline that ingests data from an SFTP server, performs data preparation, and applies automated machine learning for analysis. Let's get started!"

- First, locate and select the Data Pipeline Plugin from the app menu. This will take you to the pipeline home page. Next, look for the create icon and click on it. You'll see the option to create a new pipeline. Click on the '+' icon to proceed.
- Great! Now, let's specify the details for our end-to-end Kafka data processing pipeline. Enter a suitable name for your pipeline, such as SFTP_PipelineWorkflow2. For the description, briefly describe the workflow, such as SFTP end to end Workflow.
- Now, choose the resource allocation type based on your requirements. This feature allows you to deploy the pipeline with high, medium, or low-end configurations, depending on the velocity and volume of data that the pipeline must handle."
- Once you're done with the configuration, click on the save button to save your pipeline."

Congratulations! You've successfully created the pipeline.

Once the pipeline is saved in the pipeline list, you can add components to the canvas to create your pipeline workflow or dataflow. To add a component, simply drag the required component from the Component Palette, located on the left side of the user interface, and drop it onto the canvas. You can configure each component to define your pipeline workflow. The Pipeline Editor displays the Component Palette, which contains various components like Reader, Writer, Transformation, Consumer, Producer, Machine Learning, and more. Use these components to design your pipeline according to your specific requirements.

To follow the instructions for configuring the SFTP Monitor consumer component in the Workflow Editor, please follow these steps:

- Drag and Drop the SFTP Monitor Consumer Component:
- Locate the SFTP Monitor consumer component within the consumer section of the system component part.
- Drag and drop the SFTP Monitor consumer component into the Workflow Editor.
- Select the Invocation Type as Realtime
- Deployment Type: The deployment type is preselected based on the component.
- Container Image Version: The container image version is preselected based on the component. Failover Event: User can Select a failover event from the drop-down menu if it's created.
- Batch Size (min 1): Provide the maximum number of records to be processed in one execution cycle. The minimum limit for this field is 1.
- intelligent scaling: All components have option of Intelligent scaling which is ability of the system to dynamically adjust the scale or capacity of the reader component based on the current demand and available resources.

Move to the Meta Information Tab:

- Click on the Meta Information tab for the dragged SFTP Monitor component.
- Configure the Meta Information Settings:
- Host: Enter the IP address 10.10.28.5
- Username: If authentication is required, provide the username for accessing the SFTP server.
- Port: Specify the port number for the SFTP server 22
- Authentication: Select "Password" from the authentication dropdown to use password-based authentication.
- Directory Path: Fill in the monitor folder path using forward slashes (/). For example
- /home/ftpuser/inputData
- This is the path that the SFTP Monitor component will monitor for incoming files.
- Copy Directory Path: Specify the folder name where you want to copy the uploaded file. For example, /home/ftpuser/inputData_copy/. This is the location where the SFTP Monitor component will copy the monitored file for further processing by the SFTP Reader.
- Channel: Select the appropriate channel option from the dropdown menu. In this case, select SFTP as the supported channel for the SFTP Monitor component.
- Save Component Configuration:

- Click the 'Save Component in Storage' icon to save the configured details of the SFTP Monitor component.

To create an event and connect it with a component in the BDB platform, please follow these steps:

- Click on the Event Panel icon:
- Locate and click on the Event Panel icon in the toolbar. This will open the Event Panel.
- Add a new event:
- Within the Event Panel, click on the "+" icon to add a new event.
- Modify the event name if required.
- By default, the new event will have a generic name. Click on the event name to modify it according to your requirements.
- Add the event to the canvas.
- After modifying the event name, drag and drop the event from the Event Panel onto the canvas of the Workflow Editor.
- The event will appear as a node on the canvas.
- Connect the event to the component:
- Drag the connector line from the event node and connect it to the desired component node on the canvas.
- This establishes a connection between the event and the component.

Now, let' add transformation component.

- Drag and Drop the File Splitter component from the Transformation component palette onto the canvas. Connect it with the previous component in your workflow.
- Configure the File Splitter component by selecting the appropriate options and filling in the required fields:
- Set the Invocation Type as "Realtime" to process the files as they arrive.

Move to the "Meta Information" tab.

- Click on the "Split Type" dropdown to select the type of split option you want to use. I am selecting by File format option.
- Specify the number of outputs you want to generate from the File Splitter component. let' choose 2 from the dropdown.
- Add two events for the File Splitter output events. Each event will correspond to one of the split files generated by the File Splitter component.

Let’s Create and Add two events.

- Click on the Event Panel icon in the toolbar. This will open the Event Panel on the side of the Workflow Editor.
- click on the "+" icon to add a new event. A new event popup window will be opened
- Modify the event name by clicking on it and entering a name that suits your requirements. Repeat these steps to add one more event.
- Click on add event button.
- Drag and drop the events from the Event Panel onto the canvas of the Workflow Editor.
- Connect the events to the File Splitter component. To do this, drag the connector line from the File Splitter component and
- connect it to the desired event on the canvas. This establishes a connection between the component and the event.
- Repeat the process to connect the second event to the File Splitter component.

To configure the File Splitter component and define the output event configurations, follow these steps.

- Click on the File Splitter component to select it.
- the Meta Information tab in the component
- In the Detail section, you will find two output event configurations for the File Splitter component.
- Select the File Type for each of the output events.
- For the first output event, let's select CSV as the file type from the dropdown.
- And for the second output event, let's select Excel as the file type from the dropdown.
- Once you have selected the file types for both output events, click on the save button to save the component configuration.

Drag and drop the SFTP Stream Reader component from the Reader section onto the canvas of the Workflow Editor. Configure the component by selecting the appropriate options and filling in the required fields.

- Set the Invocation Type as "Realtime" to read the content of the file as it is updated.
- Move to the "Meta Information" tab.
- Specify the Host IP Address as "10.10.28.5".
- Enter the Username as "ftpuser".
- Set the Port Number as "22".
- Select "Password" from the Authentication dropdown.

- Provide the password for the respective user.
- Specify the Reader Path as "/home/ftpuser/inputData_copy" to indicate the location from which you want to read the file.
- Set the Channel as "SFTP".
- Click on the validate icon to initiate the connection validation process.
- If the connection is successful, you will receive a success message or indication confirming the validation.
- Once you have verified that the connection is valid, click on the save button within your workflow tool to save the component configuration.

- The configuration of the SFTP Stream Reader component will be saved, allowing you to proceed with the next steps in your pipeline workflow.

Add Event.

- Click on the Event Panel icon in the toolbar. This will open the Event Panel on the side of the Workflow Editor.
- click on the "+" icon to add a new event. A new event popup window will be opened
- Modify the event name by clicking on it and entering a name that suits your requirements.
- Click on add event button.
- Drag and drop the events from the Event Panel onto the canvas of the Workflow Editor.
- Connect the component with in and out event.

To add SFTP Reader component for reading the content of an Excel file, follow these steps.

- Drag and drop the SFTP Reader component from the Reader section onto the canvas of the Workflow Editor.
- Configure the component by selecting the appropriate options and filling in the required fields.
    - Set the Invocation Type as "Realtime" to read the content of the file as it is updated.
    - Move to the "Meta Information" tab.
    - Specify the Host IP Address as "10.10.28.5".
    - Enter the Username as "ftpuser".
    - Set the Port Number as "22".
    - Select "Password" from the Authentication dropdown.
    - Provide the password for the respective user.
    - Specify the Reader Path as "/home/ftpuser/inputData_copy" to indicate the location from which you want to read the file.

- Set the Channel as "SFTP".
- Click on the validate icon to initiate the connection validation process.
- If the connection is successful, you will receive a success message or indication confirming the validation.
- Once you have verified that the connection is valid, click on the save button within your workflow tool to save the component configuration.
- The configuration of the SFTP Reader component will be saved, allowing you to proceed with the next steps in your pipeline workflow.

Add Event

- Click on the Event Panel icon in the toolbar. This will open the Event Panel on the side of the Workflow Editor.
- click on the "+" icon to add a new event. A new event popup window will be opened.
- Modify the event name by clicking on it and entering a name that suits your requirements.
- Click on add event button.
- Drag and drop the events from the Event Panel onto the canvas of the Workflow Editor.
- Connect the component with in and out event.

let's add Data Prep transform component..

- Drag and drop the DataPrep transform component onto the canvas.
- Configure the transform component:
- Set the Invocation Type as "batch."
- Move to the Meta Information tab.
- Select "Data Sandbox" from the Data Center Type dropdown.
- Choose the sandbox file from the Sandbox Name dropdown that you added to the data center sandbox.
- Select "Preparation" from the Preparation dropdown and choose the "transformeddatapreparation" option.
- Save the component.

Now, let's move to the Event Panel to create an event and connect it with the component:

- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "+" icon.
- You can change the displayed event name and click on add event button.
- Drag and drop the event onto the canvas.
- Connect the in-event and out event with the transform component.

You have successfully added the transform component, configured it, and connected it to an event in the canvas.

Configure the AutoML component with the following settings:

- Drag and drop the Automl Component from Machine Learning component palette onto the canvas.
- Configure the component:
- Set the invocation type as "batch." and move to meta information.
- Specify the project name by selecting it from the "Project Name" dropdown. This assumes you have already created a project within your AutoML plugin.
- Choose the model you want to use from the "Model Name" dropdown. This assumes you have previously registered or trained models using AutoML.
- Save the component configuration. click on save button.

create an event and connect it with the component:

- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "+" icon.
- You can change the displayed event name and click on add event button.
- Drag and drop the event onto the canvas.

Connect the in-event and out event with component.

Now, it's time to select the appropriate writer component for your destination.

- Locate the 'DB Writer' component in the component palette.
- Drag and drop the 'DB Writer' component onto the canvas or workspace.
- Connect the DB Writer component with the event component.
- Now, let's configure the DB Writer component to connect to your ClickHouse database:
- In the configuration window of the DB Writer component, provide the necessary connection details for your ClickHouse database.
- Select invocation type as batch and move to meta information.
- Enter the host name of your ClickHouse database.
- Specify the port number on which your ClickHouse database is running.
- Enter the database name.
- Provide the username and password for authentication.
- Specify the table or collection where you want to write the transformed data in the "Destination Table" field. This should be the table name in your ClickHouse database.
- From the "Driver" dropdown, select "ClickHouse" as the driver.

- From the "Save Mode" dropdown, select "Append Mode" to append the transformed data to the existing data in the destination table.

Once you have configured the DB Writer component with the necessary connection details and settings, click on the "Save" button to apply the configuration.

Let's add one more Writer component MongoDB WRITER LITE Component

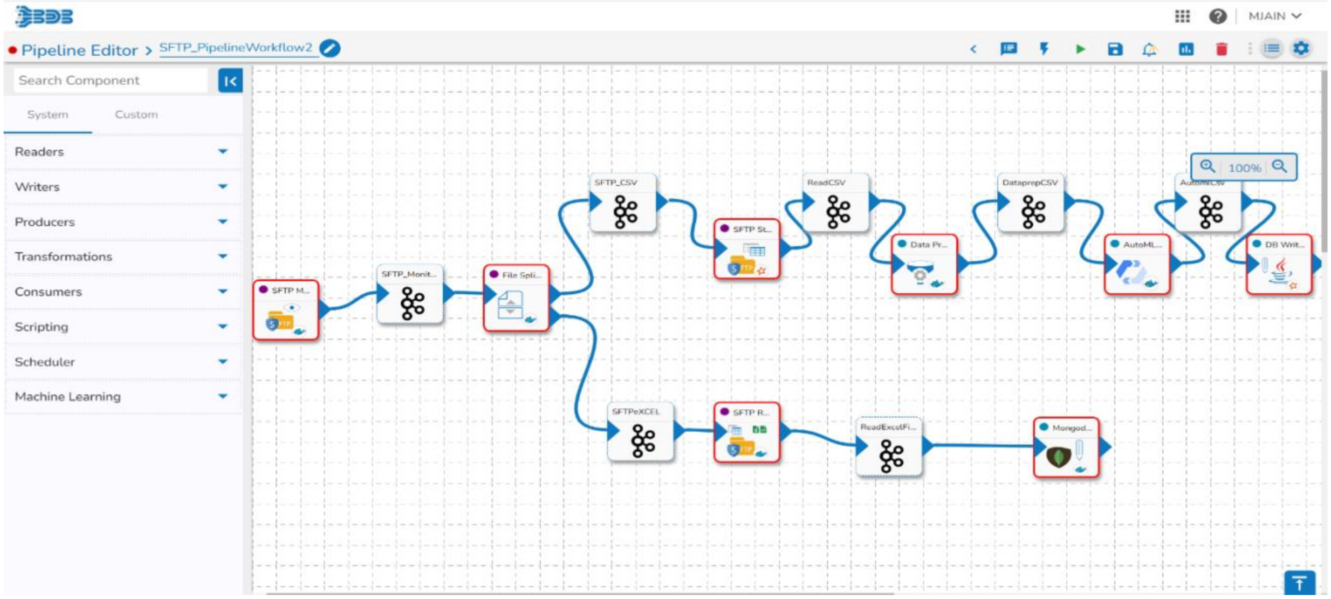
To add the MongoDB Writer Lite component to store the Excel hiring data directly from the SFTP Reader component, follow these steps:

- Locate the MongoDB WRITER LITE Component: In your pipeline development environment or component palette, find and select the MongoDB WRITER LITE component. This component is specifically designed to write data to a MongoDB database.
- Drag and drop the MongoDB WRITER LITE Component: Drag the MongoDB WRITER LITE component from the component palette onto the canvas of your pipeline workflow.
- Connect the Event to the MongoDB WRITER LITE Component: Drag a connection line from the event (previously created and added to the canvas) and connect it to the input of the MongoDB WRITER LITE component. This connection signifies that the event data will flow into the MongoDB writer component for further processing.
- Configure the MongoDB WRITER LITE Component: click on the MongoDB WRITER LITE component to open its configuration settings. Select invocation type as batch and move to the meta information and configure the necessary parameters such as the MongoDB connection details, database name, collection name, and any additional settings required for your specific use case.

To specify the connection string, database name, collection name, save mode, and composite key for the MongoDB WRITER LITE component in your Kafka data pipeline workflow, follow these steps:

- Connection Type: Select "Connection String" from the dropdown menu.
- Connection String: In the "Connection String" textbox, enter the connection string for your MongoDB database. The connection string typically includes the MongoDB server address, port number, authentication credentials (if required), and any additional parameters.
- Database Name: Specify the name of the database you want to write the data into.
- Collection Name: Enter the name for the collection in which you want to store the data. For example, you can specify "HiringDataKafka" as the collection name.
- Save Mode: Choose the "Upsert" option from the "Save Mode" dropdown menu. Upsert allows you to update existing documents or insert new documents based on a specified composite key.

Once you have configured the Writer component with the necessary connection details and settings, click on the "Save" button to apply the configuration.



Once you have configured and set up your Pipeline Workflow, it's time to update and activate the pipeline. Follow these steps to proceed.

- Locate the "Update Pipeline" icon in the toolbar of your workflow editor.
- Click on the "Update Pipeline" icon to initiate the update process.
- After updating the pipeline, click on the "Activate Pipeline" button. This will start the execution of the pipeline and initiate the data processing.
- Once the pipeline is activated, navigate to the logs section, specifically the Advanced Log section, to monitor the status of the pipeline execution.
- Check whether all the pods associated with the pipeline are up and running in the advanced log section.
- With the pipeline running and the pods up and running, your data processing tasks are now in progress.

To view the specific logs for each component, click on the corresponding pod or log entry. The logs will provide detailed information about the execution and any potential errors or issues encountered during the process.

To ingest data from WinSCP for SFTP and integrate it with your pipeline, you can follow these steps:

Launch WinSCP and connect to the SFTP server:

- Open WinSCP and enter the SFTP server's connection details, including the hostname or IP address, port, username, and password.
- Click "Login" to establish a connection to the SFTP server.
- Navigate to the directory or folder with the data files:
- Once connected, navigate to the directory or folder on the SFTP server where the data files you want to ingest are located.
- Select the files for ingestion: Select the specific files you want to ingest. In your case, you mentioned having a CSV file and an Excel file. Ensure that you have selected these files for transfer.
- Drag and drop the files into the configured folder in your pipeline: Drag and drop the selected files from WinSCP onto the specified folder that is already configured in your pipeline.
- This will initiate the file transfer process, uploading the files from your local machine to the designated folder on the SFTP server.
- Wait for the file transfer to complete: Allow some time for the file transfer to complete. The duration will depend on the file size and network speed.
- Once the transfer is finished, the files will be available in the designated folder on the SFTP server.

Now, switch to your pipeline workflow and check if the files are being picked up and processed as expected:

SFTP Monitor component: The SFTP Monitor component picks up files from the specified SFTP location.

File Splitter component: The File Splitter component splits the files based on their format. In this case, the CSV file is sent to one event, and the Excel file is sent to another event.

SFTP Reader component: The SFTP Reader component reads the contents of the received files and processes them. It will read the CSV file and send it to event1 and read the Excel file and send it to event2.

DataPrep component: DataPrep receives the data from the respective events, performs data processing based on the script or transformations applied, and sends the processed data to the output event.

AutoML component: The AutoML component reads the data from the input event, applies machine learning models or algorithms to the data, and sends the processed data to the output event.

ClickHouse DB: The processed data is stored in the ClickHouse database for further analysis or retrieval.

The processed data will be written to the specified MongoDB collection, as configured in the MongoDB Writer Lite component.

BDB 8.2 : Data Pipeline Workflow – 3

Workflow3 is designed to enable the seamless ingestion of hiring data from an API source, ensuring data quality through data preparation, leveraging AutoML techniques to extract valuable insights, and storing the processed data into a database.

It allows you to connect to the Data Center, Data Preparation and Data Pipeline and AutoML Plugins

The workflow begins by establishing a connection with the API source and retrieving real-time hiring data. The API integration allows for the efficient extraction of information such as job listings, candidate profiles, and recruitment metrics. This ensures that the hiring data is up-to-date and reflects the latest developments in the recruitment process.

To ensure data quality and reliability, the workflow incorporates a data preparation step. This involves transforming the raw hiring data by cleaning, standardizing, and structuring it for further analysis. Tasks such as handling missing values, removing inconsistencies, and performing feature engineering are carried out to enhance the accuracy and consistency of the data.

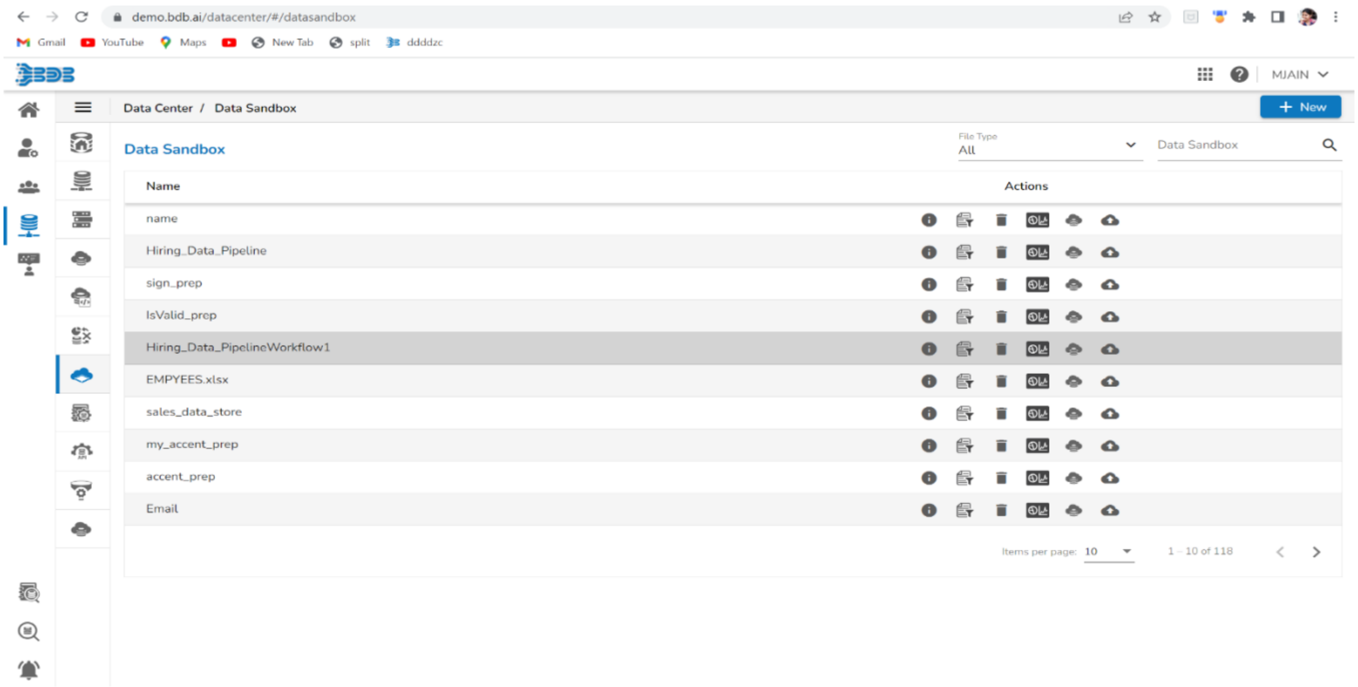
Once the data is prepared, the workflow utilizes AutoML techniques to extract valuable insights. AutoML plugin automate the process of training and evaluating machine learning models. By feeding the prepared hiring data into the AutoML system, various machine learning algorithms, model architectures, and hyperparameter configurations are explored to identify the best-performing model. Finally, the workflow includes a step to store the processed data into a database. This allows for easy retrieval, further analysis, and integration with reporting systems. By writing the hiring data, along with the derived insights from the AutoML experiment, into a database, User can access and utilize the information for decision-making, optimization, and reporting purposes.

Upload File in Sandbox

"Let's Understand the Process of creating a new data sandbox and preparing the data for further operations. Now, let get started

- Start by clicking on the 'Apps' menu, where you'll find a list of available modules.
- From the menu, select the 'Data Center' module. This will take you to the Data Center page, where you can manage your data.

- Within the Data Center page, navigate to the 'Sandbox' section or tab. This section is specifically designed for managing sandboxes for your data.
- Once you're in the Sandbox section, look for the option to create a new sandbox. Click on it to proceed.
- Give the sandbox a descriptive name that will help you identify it later. For example, let's name it 'Hiring_Data_Pipeline'
- Provide an appropriate description for the sandbox if needed. This description can help provide more context or details about the purpose of the sandbox. In this case, let's use 'Hiring Data for Pipeline Workflow Creation' as the description.
- Now, it's time to choose your data file. Look for an option to upload or select a data file for the sandbox. Click on Upload Button



Data Center / Data Sandbox

Data Sandbox

File Type: All

Data Sandbox

Name	Actions
name	
Hiring_Data_Pipeline	
sign_prep	
IsValid_prep	
Hiring_Data_PipelineWorkflow1	
EMPYEES.xlsx	
sales_data_store	
my_accent_prep	
accent_prep	
Email	

Items per page: 10 1 – 10 of 118

Great! You've successfully uploaded the data file to the sandbox. Now, let's proceed to the next step, which is Preparation. In this step, we'll prepare the data for further analysis or processing.

Create Preparation Using DataPreparation:

Data Preparation is used to clean the data.

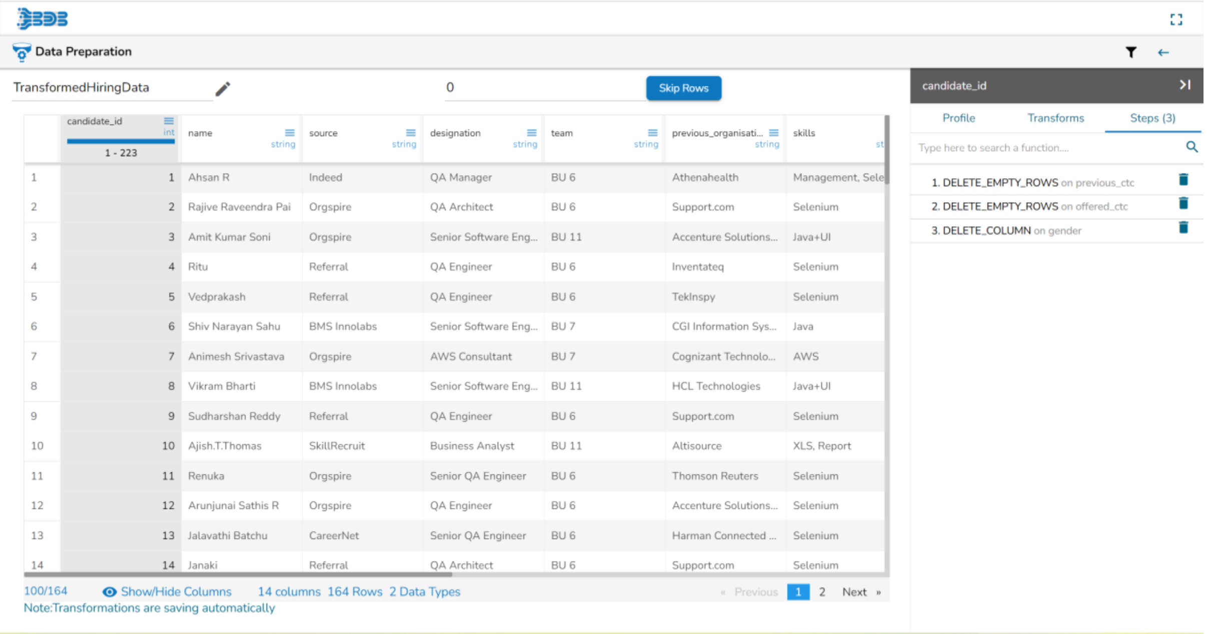
Are you ready to prepare your hiring data pipeline sandbox file? Let's get started with the data preparation process.

- Select Data Preparation Icon. will navigate to the Data preparation home page.
- Here on the Data Preparation home page, you can see your complete dataset displayed in a grid form. The Data Preparation Plugin automatically profiles the data,

- providing valuable insights into its characteristics and detecting any anomaly data. You can also view the data profiling details.
- On the right-hand side, you'll find the selected column's profile. Here, you can explore various details such as charts, information, and patterns associated with the selected column.
- All the Transformations will appear inside transform tab.

Transforms

- Now, let's start preparing and cleaning the data. To remove the empty cells in the 'Previous CTC' column, just select prevoius ctc column and
- navigate to the 'Transforms' tab and search for the 'Delete Empty Rows with Empty Cells' transform. Click on it to remove all the empty rows from the 'Previous CTC' column."
- You can see that the empty rows in the 'Previous CTC' column have been removed
- Next, let's perform the same transformation on the 'Offered CTC' column.
- Select the column and search for the 'Delete Rows with Empty Cell' transform. Click on it to remove the empty rows from the 'Offered CTC' column.
- Great! The empty rows in the 'Offered CTC' column have been successfully removed."
- Now, let's delete the 'Gender' column from the dataset. Simply select the column and search for the 'Delete Column' transform. Click on it to remove the 'Gender' column.
- Perfect! The 'Gender' column has been deleted from the dataset
- You can see that all the performed transforms are recorded in the 'Steps' section. This helps you keep track of the changes made to the dataset.
- Now, let's rename the preparation for identification purposes. Simply click on the edit icon and give it a new name, such as 'TransformedHiringData
- Great! The preparation has been renamed to 'TransformedHiringData



TransformedHiringData

0

	candidate_id int 1 - 223	name string	source string	designation string	team string	previous_organisati... string	skills st
1	1	Ahsan R	Indeed	QA Manager	BU 6	Athenahealth	Management, Sele
2	2	Rajive Raveendra Pai	Orgspire	QA Architect	BU 6	Support.com	Selenium
3	3	Amit Kumar Soni	Orgspire	Senior Software Eng...	BU 11	Accenture Solutions...	Java+UI
4	4	Ritu	Referral	QA Engineer	BU 6	Inventateq	Selenium
5	5	Vedprakash	Referral	QA Engineer	BU 6	TekInspy	Selenium
6	6	Shiv Narayan Sahu	BMS Innolabs	Senior Software Eng...	BU 7	CGI Information Sys...	Java
7	7	Animesh Srivastava	Orgspire	AWS Consultant	BU 7	Cognizant Technolo...	AWS
8	8	Vikram Bharti	BMS Innolabs	Senior Software Eng...	BU 11	HCL Technologies	Java+UI
9	9	Sudharshan Reddy	Referral	QA Engineer	BU 6	Support.com	Selenium
10	10	Ajish.T.Thomas	SkillRecruit	Business Analyst	BU 11	Altisource	XLS, Report
11	11	Renuka	Orgspire	Senior QA Engineer	BU 6	Thomson Reuters	Selenium
12	12	Arunjunai Sathis R	Orgspire	QA Engineer	BU 6	Accenture Solutions...	Selenium
13	13	Jalavathi Batchu	CareerNet	Senior QA Engineer	BU 6	Harman Connected ...	Selenium
14	14	Janaki	Referral	QA Architect	BU 6	Support.com	Selenium

100/164 Show/Hide Columns 14 columns 164 Rows 2 Data Types

Note:Transformations are saving automatically

candidate_id

1. DELETE_EMPTY_ROWS on previous_ctc
2. DELETE_EMPTY_ROWS on offered_ctc
3. DELETE_COLUMN on gender

Click on the back icon, and the preparation will be automatically saved and exported to different plugins, such as the Data Pipeline or AutoML. You can choose to export the transformed data to various plugins, such as the Data Pipeline or AutoML, based on your needs.

Create AutoML Workflow using DS Lab

To create an Automl Experiment using the DS Lab Plugin, follow these steps:

- Open the DS Lab Plugin from the app menu.
- Navigate to the DS Lab home page where all existing projects are displayed.

Create Project:

  - Click on the "Create Project" button to proceed to the project creation page.
  - On the project creation page, provide the following mandatory fields:
  - Project Name: Enter a name for your project. For example, "Hiring Data".
  - Project Description: Add a description for your project, such as "Creating AutoML experiment for hiring data prediction".
  - Select Algorithm Types: Choose the algorithm types that are required for your project from the available options in the dropdown menu.
  - Select Environment: Choose the environment for your project. In this case, select "PythonTensorFlow".
  - Specify Resource Allocation: Select the resource allocation based on the volume of data and the computational requirements of your project. Choose the appropriate option, such as "Medium".
  - Select Idle Shutdown Time: Choose the duration after which the system should shut down if idle. For example, select "1 hr.".
  - Specify External Libraries: If you require any external libraries for your project, mention them in this field.
  - Mention GPU Type: If your project requires GPU acceleration, specify the GPU type.
  - Once you have filled in all the necessary details, click on the "Save" button to create the project.

You should see a message confirming that the project has been successfully created.

Create AutoML Experiment

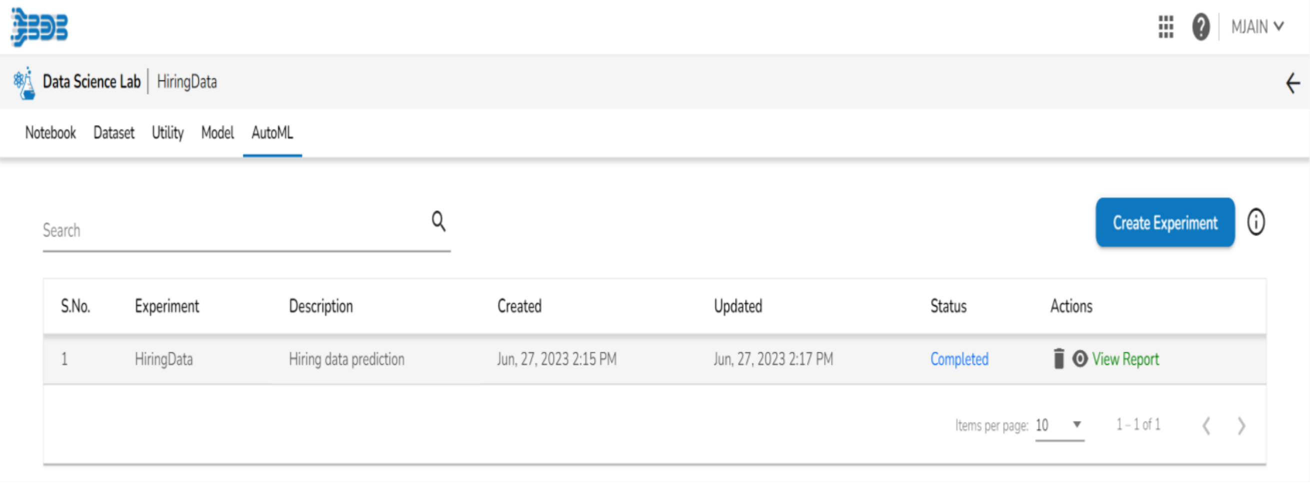
To create an AutoML experiment using the recently created "Hiring Data" project and follow these steps:

- Navigate to the "Datasets" tab in Data Science Lab.

- Click on the "Add Dataset" button.
- Select the "Data Sandbox" option from the data source dropdown menu. This will display all the files that have been uploaded into the sandbox.
- Choose the "Hiring_data_Pipeline" file.
- Click on the "Add" button to successfully add the dataset.
- Many action items will be displayed for the selected dataset, such as preview, data profile, create experiment, delete, and data preparation.
- Choose the "Create Experiment" option for the uploaded data file.
- Provide an experiment name, for example, "HiringData," and a description, such as "Hiring data prediction."
- Select the target column from the dropdown menu. You can also choose any data preparation options if required. Click on the "Next" button to proceed.
- Select the experiment type based on your requirements, such as classification or regression. In this case, select "Classification" as the gender column contains categories.
- Click on "Done" to start the experiment.
- Wait for the experiment to complete in the AutoML tab. You can track the progress as the status changes from "Started" to "Running."
- Once the experiment is completed, open and view the report. The report will display the recommended model and a run summary.

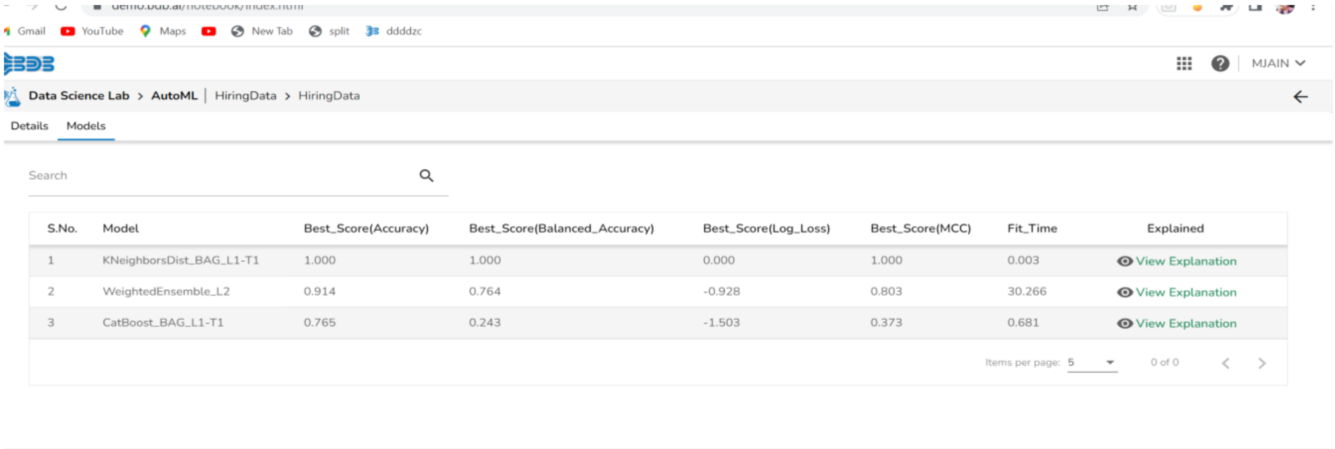
Register Model

- Switch to the "Models" tab to explore the top three models trained using AutoML.
- Click on "View Explanation" to understand the model's performance and switch to the "Dataset Explorer" tab to see the data profile.
- Navigate to the "Models" tab and expand the saved models.
- Export the models to the data pipeline using the register icon.
- Optionally, export the model into Git for version control and collaboration purposes.



S.No.	Experiment	Description	Created	Updated	Status	Actions
1	HiringData	Hiring data prediction	Jun, 27, 2023 2:15 PM	Jun, 27, 2023 2:17 PM	Completed	View Report

Items per page: 10 1 – 1 of 1



Data Science Lab > AutoML | HiringData > HiringData

S.No.	Model	Best_Score(Accuracy)	Best_Score(Balanced_Accuracy)	Best_Score(Log_Loss)	Best_Score(MCC)	Fit_Time	Explained
1	KNeighborsDist_BAG_L1-T1	1.000	1.000	0.000	1.000	0.003	View Explanation
2	WeightedEnsemble_L2	0.914	0.764	-0.928	0.803	30.266	View Explanation
3	CatBoost_BAG_L1-T1	0.765	0.243	-1.503	0.373	0.681	View Explanation

By following these steps, you should be able to create an AutoML experiment using the "Hiring Data" project and explore the recommended models in Data Science Lab.

Pipeline Creation:

Now, Let's Understand the process of creating Data pipeline workflow.

- First, locate and select the Data Pipeline Plugin from the app menu. This will take you to the pipeline home page.
- "Next, look for the create icon and click on it. You'll see the option to create a new pipeline. Click on the '+' icon to proceed.
- Great! Now, let's specify the details for our end-to-end data processing and automl pipeline. Enter a suitable name for your pipeline, such as AutoML_Workflow3. For the description, briefly describe the workflow, such as End to End Data Processing and AutoML Workflow. '
- Now, choose the resource allocation type based on your requirements. This feature allows you to deploy the pipeline with high, medium, or low-end configurations, depending on the velocity and volume of data that the pipeline must handle."
- Once you're done with the configuration, click on the save button to save your pipeline."

Congratulations! You've successfully created the pipeline.

Once the pipeline is saved in the pipeline list, you can add components to the canvas to create your pipeline workflow or dataflow.

To add a component, simply drag the required component from the Component Palette, located on the left side of the user interface, and drop it onto the canvas. You can configure each component to define your pipeline workflow. The Pipeline Editor displays the Component Palette, which contains various components like Reader, Writer, Transformation, Consumer, Producer, Machine Learning, and more. Use these components to design your pipeline according to your specific requirements.

Next, let's add the API Ingestion component to our pipeline.

- Drag and drop the API Ingestion component from the Consumer section onto the canvas.
- Great! Now it's time to configure the API component to make it work for your specific API.
- Let's Select the Invocation type as 'Realtime and move to the Meta Information tab.
- Move to the Meta Information tab, where you'll find the ingestion ID and ingestion secrets already configured.
- component instance ID URL should be automatically generated when will update pipeline.
- Select the Ingestion type as 'API Ingestion' from the dropdown.
- Once you've configured the component, click on the Save component icon to save your changes."
- Now, let's move to the Event Panel. to create an event and connect it with the component:
- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "Create Event" button.
- Now, let's add the event component to the canvas. Drag and drop the event component from the Event Panel onto the canvas.
- Great! Now, connect the event component with the API Ingestion component by dragging and dropping a connection line between them.
- Well done! You have successfully added an event component and connected it with the API Ingestion component in your pipeline. Your workflow is taking shape. Keep up the excellent work!

Now, let's add transform component.

- Drag and drop the DataPrep transform component onto the canvas.
- Configure the transform component:
- Set the Invocation Type as "Realtime."
- Move to the Meta Information tab.
- Select "Data Sandbox" from the Data Center Type dropdown.
- Choose the sandbox file from the Sandbox Name dropdown that you added to the data center sandbox.
- Select "Preparation" from the Preparation dropdown and choose the "transformeddatapreparation" option.
- Save the component.

Now, let's move to the Event Panel to create an event and connect it with the component:

- Click on the Event Panel icon located in the toolbar.

- Add a new event by clicking on the "Create Event" button.
- Drag and drop the event onto the canvas.
- Connect the in-event and out event with the transform component.

You have successfully added the transform component, configured it, and connected it to an event in the canvas.

Configure the AutoML component with the following settings:

- Drag and drop the Automl Component from Machine Learning component palette onto the canvas.
- Configure the component:
- Set the invocation type as "batch." and move to meta information.
- Specify the project name by selecting it from the "Project Name" dropdown. This assumes you have already created a project within your AutoML plugin.
- Choose the model you want to use from the "Model Name" dropdown. This assumes you have previously registered or trained models using AutoML.
- Save the component configuration. click on save button.

create an event and connect it with the component:

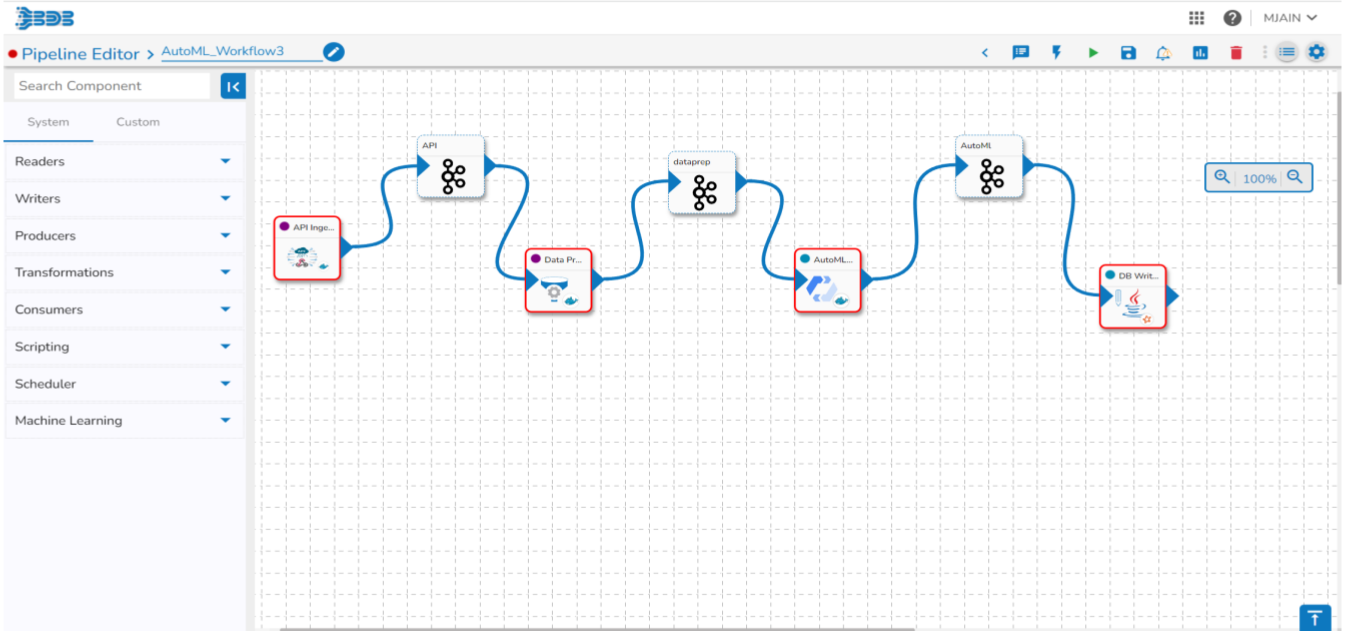
- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "Create Event" button.
- Drag and drop the event onto the canvas.
- Connect component with in and out event.

Now, it's time to select the appropriate writer component for your destination.

- Locate the 'DB Writer' component in the component palette.
- Drag and drop the 'DB Writer' component onto the canvas or workspace.
- Connect the DB Writer component with the event component.
- Now, let's configure the DB Writer component to connect to your ClickHouse database:
- In the configuration window of the DB Writer component, provide the necessary connection details for your ClickHouse database.
- Select invocation type as batch and move to meta information.
- Enter the host name of your ClickHouse database.
- Specify the port number on which your ClickHouse database is running.
- Enter the database name.
- Provide the username and password for authentication.
- Specify the table or collection where you want to write the transformed data in the "Destination Table" field. This should be the table name in your ClickHouse database.

- From the "Driver" dropdown, select "ClickHouse" as the driver.
- From the "Save Mode" dropdown, select "Append Mode" to append the transformed data to the existing data in the destination table.

Once you have configured the DB Writer component with the necessary connection details and settings, click on the "Save" button to apply the configuration.



After configuring and setting up the Pipeline Workflow, it's time to Update and activate the pipeline.

- Locate "Update Pipeline" icon in the toolbar and click on it
- Now, click on the 'Activate Pipeline button. This will Start the execution of the Pipeline and start. the data processing.
- After activating the Pipeline, navigate to the logs and advance Log section, Look for the Log Panel and click on it to access the advanced logs for detailed information.
- Within the Log Panel, you'll see the pods associated with each component. Pods are containers that hold the execution environment for the Pipeline.
- Check if the pods for each component have come up and are running. This indicates that the components are successfully deployed and ready to execute their tasks.
- To view the specific logs for each component, click on the corresponding pod or log entry. The logs will provide detailed information about the execution and any potential errors or

issues encountered during the process.

IF API component started.

To ingest data from Postman and set up the ingestion ID and ingestion secrets same as , pipeline.

follow these steps.

- Open Postman and ensure you have the necessary API requests set up to retrieve the hiring data.
- Locate the request or collection you want to use for data ingestion.
- Before sending the request, set the required headers or authentication parameters for the API
- Use Hiring data json in body of the postman.
- Look for the headers or response properties that contain the ingestion ID and ingestion secrets and set the same component instance id URL and Click on send then you will get success as true.
- Go to Pipeline workflow and click on API component event.

Congratulations on successfully setting up your pipeline for securely retrieving and processing the hiring data from the API! Now, the pipeline will perform the following tasks:

Retrieve Data: The pipeline securely retrieves the hiring data from the API source.

Data Transformation: The pipeline applies the necessary transformations and preprocessing steps to the hiring data to prepare it for further processing.

Output Event: The transformed hiring data is sent to the output event, where it can be consumed by other components or systems.

Configure AutoML Model: The pipeline configures the AutoML model, which includes selecting the appropriate algorithm, defining the model parameters, and setting up the training process.

Processed Data Storage: The processed hiring data is stored in the Clickhouse DB, ensuring secure and efficient storage for future analysis and retrieval.

With these steps in place, your pipeline is now fully functional and ready to handle the secure retrieval, transformation, AutoML model configuration, and storage of the hiring data.

Analyze the logs for each component to ensure that they are functioning as expected and that there are no errors or failures reported. If any issues arise, troubleshoot them accordingly."

Click on the data preview option inside component. This will display a sample of the records that the component will produce as its output during the execution. Analyze the sample records to verify that the component is generating the expected output. You can examine the data structure, values, and any applied transformations or Prediction.

BDB 8.2: Data Pipeline Workflow – 4

Workflow 4 Focuses on the creation of a pipeline with data preparation and two outputs (database and WebSocket)

We are thrilled to introduce you to Workflow 4, a powerful feature that enables seamless connectivity to the Data Center, Data Preparation, and Data Pipeline Plugin.

This workflow is designed to ingest hiring data from an API source. It retrieves real-time data related to hiring processes, including job listings, candidate profiles, and recruitment metrics. The API integration allows seamless extraction of hiring data, which is then transformed using data preparation plugin and stored for further analysis and reporting. The pipeline ensures the efficient and automated ingestion of hiring data, providing valuable insights for talent acquisition and recruitment efforts.

Data Sandbox and Data Preparation:

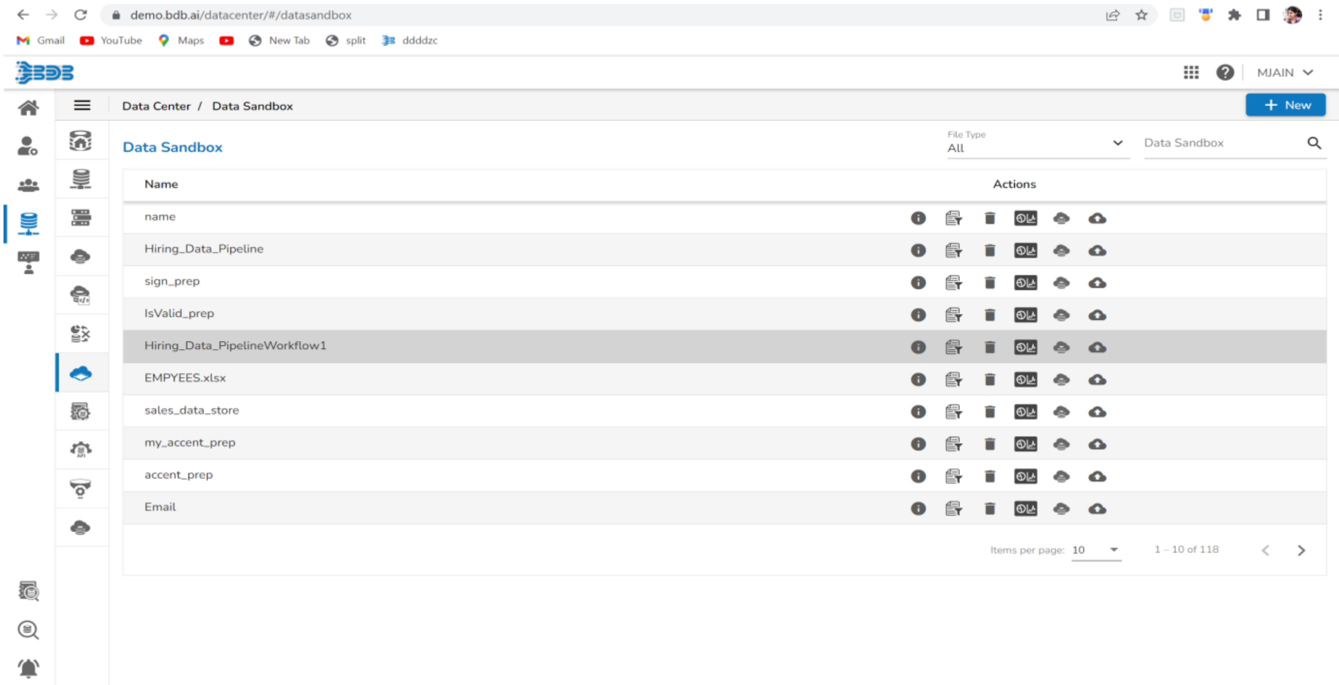
Upload File in Sandbox:

Let's Focus on the process of creating a new data sandbox and performing tasks on the data

To begin,

- click on the 'Apps' menu, which will display a list of available modules.
- From the menu, choose the 'Data Center' module. This will take you to the Data Center page, where you can manage your data.
- Now, let's navigate to the Sandbox section, which is dedicated to managing sandboxes for your data.
- Look for the 'Sandbox' section or tab within the Data Center page.
- Once you've found the Sandbox section, you should see an option to create a new sandbox. Click on that
- Now, let's give the sandbox a name. Choose a name that will help you identify it later. For example, let's name it 'Hiring_data_PipelineWorkflow1'.
- Provide an appropriate description if needed. This can help provide more context or details about the purpose of the sandbox.

- Next, you'll need to choose your data file. Look for an option to upload or select a data file for the sandbox.


demo.bdb.ai/datacenter/#/datasandbox
Gmail
YouTube
Maps
New Tab
split
ddddzc
MJAIN
Data Center / Data Sandbox
+ New

Data Sandbox

File Type: All

Data Sandbox

Name	Actions
name	
Hiring_Data_Pipeline	
sign_prep	
IsValid_prep	
Hiring_Data_PipelineWorkflow1	
EMPYEES.xlsx	
sales_data_store	
my_accent_prep	
accent_prep	
Email	

Items per page: 10 1 – 10 of 118

Great! You've successfully uploaded the data file to the sandbox. Now, let's proceed to the next step, which is Preparation. In this step, we'll prepare the data for further analysis or processing.

Create Preparation Using Data Preparation:

Data Preparation is used to clean the data

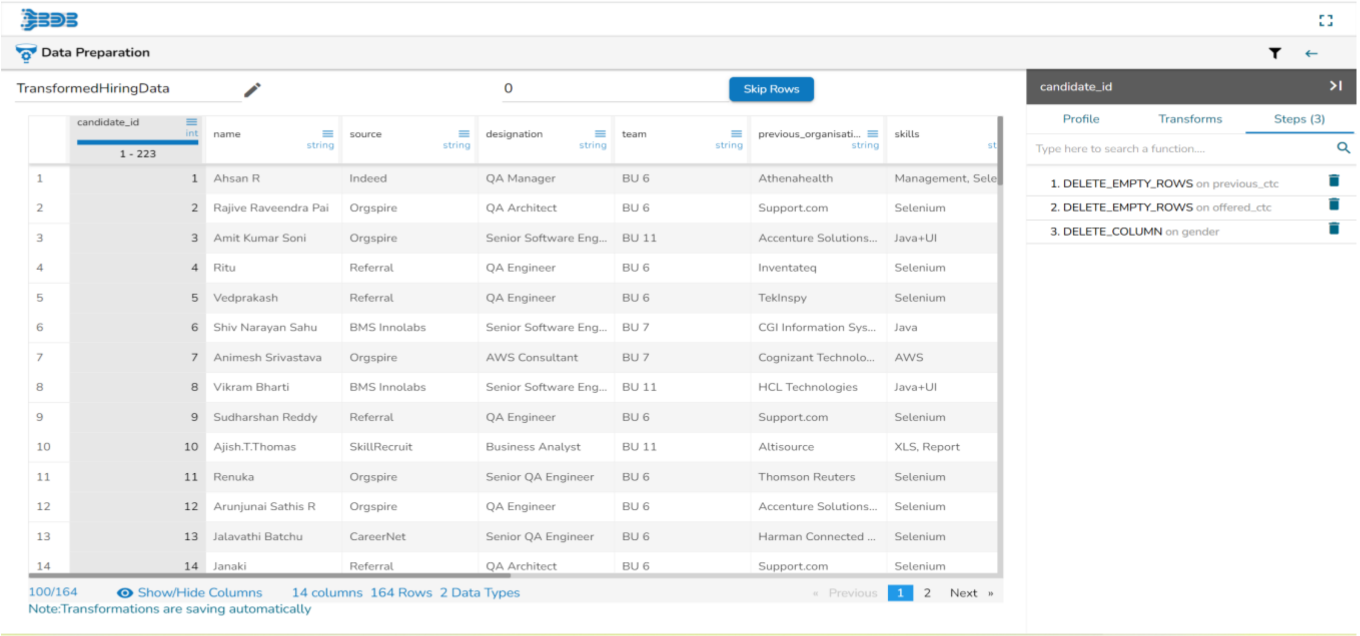
Are you ready to prepare your hiring data pipeline sandbox file? Let's get started with the data preparation process.

- Select Data Preparation Icon. will navigate to the Data preparation home page
- Here on the Data Preparation home page, you can see your complete dataset displayed in a grid form. The Data Preparation Plugin automatically profiles the data,
- providing valuable insights into its characteristics and detecting any anomaly data. You can also view the data profiling details.
- On the right-hand side, you'll find the selected column's profile. Here, you can explore various details such as charts, information, and patterns associated with the selected column.
- All the Transformations will appear inside transform tab.

Transforms

- Now, let's start preparing and cleaning the data. To remove the empty cells in the 'Previous CTC' column, just select prevoius ctc column and

- navigate to the 'Transforms' tab and search for the 'Delete Empty Rows with Empty Cells' transform. Click on it to remove all the empty rows from the 'Previous CTC' column."
- You can see that the empty rows in the 'Previous CTC' column have been removed.
- Next, let's perform the same transformation on the 'Offered CTC' column.
- Select the column and search for the 'Delete Rows with Empty Cell' transform. Click on it to remove the empty rows from the 'Offered CTC' column.
- Great! The empty rows in the 'Offered CTC' column have been successfully removed."
- Now, let's delete the 'Gender' column from the dataset. Simply select the column and search for the 'Delete Column' transform. Click on it to remove the 'Gender' column.
- Perfect! The 'Gender' column has been deleted from the dataset.
- You can see that all the performed transforms are recorded in the 'Steps' section. This helps you keep track of the changes made to the dataset.
- Now, let's rename the preparation for identification purposes. Simply click on the edit icon and give it a new name, such as 'TransformedHiringData
- Great! The preparation has been renamed to 'TransformedHiringData.



	candidate_id int 1 - 223	name string	source string	designation string	team string	previous_organisati... string	skills st
1	1	Ahsan R	Indeed	QA Manager	BU 6	Athenahealth	Management, Sele
2	2	Rajive Raveendra Pai	Orgspire	QA Architect	BU 6	Support.com	Selenium
3	3	Amit Kumar Soni	Orgspire	Senior Software Eng...	BU 11	Accenture Solutions...	Java+UI
4	4	Ritu	Referral	QA Engineer	BU 6	Inventateq	Selenium
5	5	Vedprakash	Referral	QA Engineer	BU 6	TekInspy	Selenium
6	6	Shiv Narayan Sahu	BMS Innolabs	Senior Software Eng...	BU 7	CGI Information Sys...	Java
7	7	Animesh Srivastava	Orgspire	AWS Consultant	BU 7	Cognizant Technolo...	AWS
8	8	Vikram Bharti	BMS Innolabs	Senior Software Eng...	BU 11	HCL Technologies	Java+UI
9	9	Sudharshan Reddy	Referral	QA Engineer	BU 6	Support.com	Selenium
10	10	Ajish.T.Thomas	SkillRecruit	Business Analyst	BU 11	Altisource	XLS, Report
11	11	Renuka	Orgspire	Senior QA Engineer	BU 6	Thomson Reuters	Selenium
12	12	Arunjunai Sathis R	Orgspire	QA Engineer	BU 6	Accenture Solutions...	Selenium
13	13	Jalavathi Batchu	CareerNet	Senior QA Engineer	BU 6	Harman Connected ...	Selenium
14	14	Janaki	Referral	QA Architect	BU 6	Support.com	Selenium

100/164 Show/Hide Columns 14 columns 164 Rows 2 Data Types

« Previous 1 2 Next »

Note:Transformations are saving automatically

1. DELETE_EMPTY_ROWS on previous_ctc
2. DELETE_EMPTY_ROWS on offered_ctc
3. DELETE_COLUMN on gender

Click on the back icon, and the preparation will be automatically saved and exported to different plugins, such as the Data Pipeline or AutoML. You can choose to export the transformed data to various plugins, such as the Data Pipeline or AutoML, based on your needs.

Create AutoML Workflow using DS Lab

To create an Automl Experiment using the DS Lab Plugin, follow these steps:

- Open the DS Lab Plugin from the app menu.

- Navigate to the DS Lab home page where all existing projects are displayed.

Create Project:

- Click on the "Create Project" button to proceed to the project creation page.
- On the project creation page, provide the following mandatory fields:
- Project Name: Enter a name for your project. For example, "Hiring Data".
- Project Description: Add a description for your project, such as "Creating AutoML experiment for hiring data prediction".
- Select Algorithm Types: Choose the algorithm types that are required for your project from the available options in the dropdown menu.
- Select Environment: Choose the environment for your project. In this case, select "PythonTensorFlow".
- Specify Resource Allocation: Select the resource allocation based on the volume of data and the computational requirements of your project. Choose the appropriate option, such as "Medium".
- Select Idle Shutdown Time: Choose the duration after which the system should shut down if idle. For example, select "1 hr".
- Specify External Libraries: If you require any external libraries for your project, mention them in this field.
- Mention GPU Type: If your project requires GPU acceleration, specify the GPU type.
- Once you have filled in all the necessary details, click on the "Save" button to create the project.

You should see a message confirming that the project has been successfully created.

Create AutoML Experiment

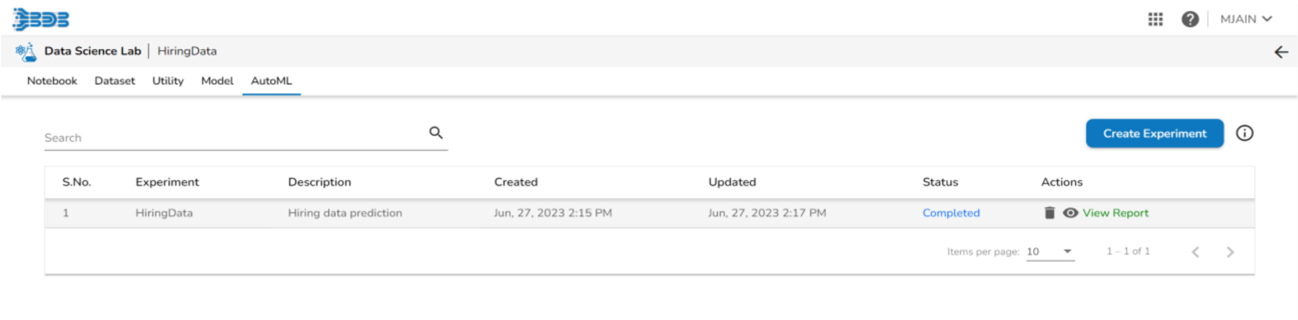
To create an AutoML experiment using the recently created "Hiring Data" project and follow these steps:

- Navigate to the "Datasets" tab in Data Science Lab.
- Click on the "Add Dataset" button.
- Select the "Data Sandbox" option from the data source dropdown menu. This will display all the files that have been uploaded into the sandbox.
- Choose the "Hiring_data_Pipeline" file.
- Click on the "Add" button to successfully add the dataset.
- Many action items will be displayed for the selected dataset, such as preview, data profile, create experiment, delete, and data preparation.
- Choose the "Create Experiment" option for the uploaded data file.

- Provide an experiment name, for example, "HiringData," and a description, such as "Hiring data prediction."
- Select the target column from the dropdown menu. You can also choose any data preparation options if required. Click on the "Next" button to proceed.
- Select the experiment type based on your requirements, such as classification or regression. In this case, select "Classification" as the gender column contains categories.
- Click on "Done" to start the experiment.
- Wait for the experiment to complete in the AutoML tab. You can track the progress as the status changes from "Started" to "Running."
- Once the experiment is completed, open and view the report. The report will display the recommended model and a run summary.

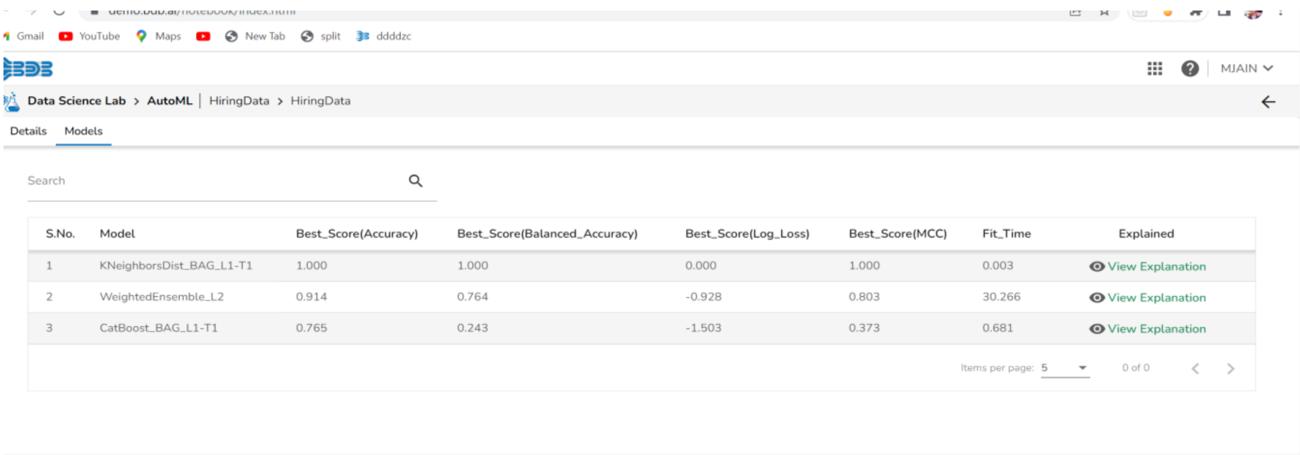
Register Model

- Switch to the "Models" tab to explore the top three models trained using AutoML.
- Click on "View Explanation" to understand the model's performance and switch to the "Dataset Explorer" tab to see the data profile.
- Navigate to the "Models" tab and expand the saved models.
- Export the models to the data pipeline using the register icon.
- Optionally, export the model into Git for version control and collaboration purposes.



S.No.	Experiment	Description	Created	Updated	Status	Actions
1	HiringData	Hiring data prediction	Jun, 27, 2023 2:15 PM	Jun, 27, 2023 2:17 PM	Completed	View Report

Items per page: 10 1 – 1 of 1


demo.bdb.ai/notebook/index.html
Gmail
YouTube
Maps
New Tab
split
ddddzc
MJAIN
Data Science Lab > AutoML | HiringData > HiringData
Details
Models
Search

S.No.	Model	Best_Score(Accuracy)	Best_Score(Balanced_Accuracy)	Best_Score(Log_Loss)	Best_Score(MCC)	Fit_Time	Explained
1	KNeighborsDist_BAG_L1-T1	1.000	1.000	0.000	1.000	0.003	View Explanation
2	WeightedEnsemble_L2	0.914	0.764	-0.928	0.803	30.266	View Explanation
3	CatBoost_BAG_L1-T1	0.765	0.243	-1.503	0.373	0.681	View Explanation

Items per page: 5 0 of 0

By following these steps, you should be able to create an AutoML experiment using the "Hiring Data" project and explore the recommended models in Data Science Lab.

Data Pipeline:

Now, Let's understand process of creating an API-based pipeline workflow.

Let's get started!

- First, locate and select the Data Pipeline Plugin from the app menu. This will take you to the pipeline home page.
- "Next, look for the create icon and click on it. You'll see the option to create a new pipeline. Click on the '+' icon to proceed.
- Great! Now, let's specify the details for our API-based pipeline. Enter a suitable name for your pipeline, such as API_Hiring Data_Workflow4. For the description, briefly describe the workflow, such as 'Ingesting Hiring Data from API Source'.
- Now, choose the resource allocation type based on your requirements. This feature allows you to deploy the pipeline with high, medium, or low-end configurations, depending on the velocity and volume of data that the pipeline must handle."

Once you're done with the configuration, click on the save button to save your pipeline."

Congratulations! You've successfully created the pipeline.

Once the pipeline is saved in the pipeline list, you can add components to the canvas to create your pipeline workflow or dataflow. To add a component, simply drag the required component from the Component Palette, located on the left side of the user interface, and drop it onto the canvas. You can configure each component to define your pipeline workflow. The Pipeline Editor displays the Component Palette, which contains various components like Reader, Writer, Transformation, Consumer, Producer, ML, and more. Use these components to design your pipeline according to your specific requirements.

Next, let's add the **API Ingestion component** to our pipeline.

- Drag and drop the API Ingestion component from the Consumer section onto the canvas
- Let's Select the Invocation type as 'Realtime and move to the Meta Information tab.
- Move to the Meta Information tab, where you'll find the ingestion ID and ingestion secrets already configured component instance ID URL should be automatically generated when will update pipeline.
- Select the Ingestion type as 'API Ingestion' from the dropdown.
- Once you've configured the component, click on the Save component icon to save your changes."
- Now, let's move to the Event Panel. to create an event and connect it with the component:
- Click on the Event Panel icon located in the toolbar.

- Add a new event by clicking on the "Create Event" button.
- Now, let's add the event component to the canvas. Drag and drop the event component from the Event Panel onto the canvas.
- Great! Now, connect the event component with the API Ingestion component by dragging and dropping a connection line between them.

Well done! You have successfully added an event component and connected it with the API Ingestion component in your pipeline. Your workflow is taking shape. Keep up the excellent work!

Now, let's add transform component.

- Drag and drop the Data Prep transform component onto the canvas.
- Configure the transform component:
- Set the Invocation Type as "Realtime."
- Move to the Meta Information tab.
- Select "Data Sandbox" from the Data Center Type dropdown.
- Choose the sandbox file from the Sandbox Name dropdown that you added to the data center sandbox.
- Select "Preparation" from the Preparation dropdown and choose the "transformed data preparation" option.
- Save the component.

Now, let's move to the Event Panel to create an event and connect it with the component:

- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "Create Event" button.
- Drag and drop the event onto the canvas.
- Connect the event with the transform component.
- Click on event and change output filed from 1 to 2

You have successfully added the transform component, configured it, and connected it to an event in the canvas.

After applying Dataprep transformations to the data, it's time to select the appropriate writer component for your destination.

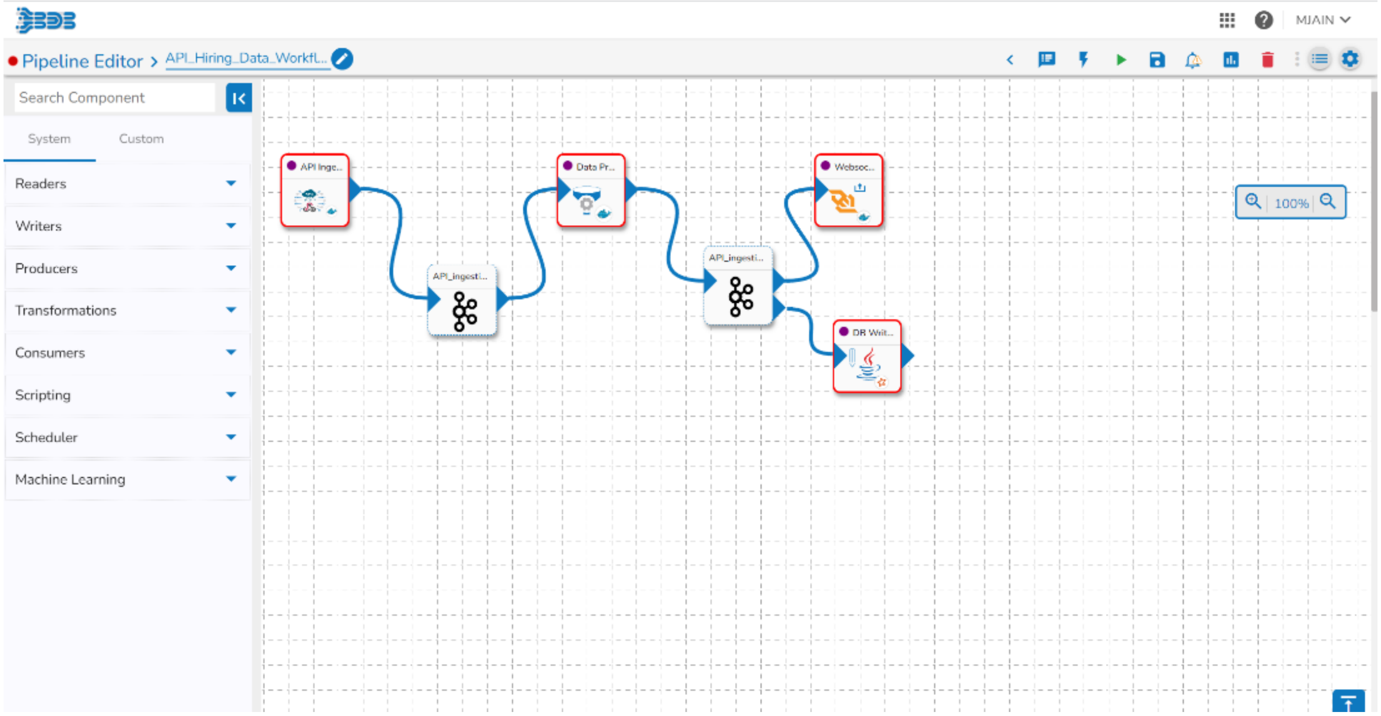
- Locate the 'DB Writer' component in the component palette.
- Drag and drop the 'DB Writer' component onto the canvas or workspace.
- Connect the DB Writer component with the event component.
- Now, let's configure the DB Writer component to connect to your ClickHouse database:

- In the configuration window of the DB Writer component, provide the necessary connection details for your ClickHouse database.
- Select invocation type as batch and move to meta information.
- Enter the host name of your ClickHouse database.
- Specify the port number on which your ClickHouse database is running.
- Enter the database name.
- Provide the username and password for authentication.
- Specify the table or collection where you want to write the transformed data in the "Destination Table" field. This should be the table name in your ClickHouse database.
- From the "Driver" dropdown, select "ClickHouse" as the driver.
- From the "Save Mode" dropdown, select "Append Mode" to append the transformed data to the existing data in the destination table.
- Once you have configured the DB Writer component with the necessary connection details and settings, click on the "Save" button to apply the configuration.

Now, let add Websocket producer component:

A WebSocket producer pipeline component can work with a dashboard to provide real-time data streaming and visualization.

- Drag and drop the Websocket Producer component from the producer section onto the canvas and connect with event.
- In the Websocket component's configuration, select "Realtime" as the invocation type from the dropdown.
- Move to the Meta Information tab, where you'll find the ingestion ID and ingestion secrets already configured.
- Note that once you update the pipeline, a new GUID (Globally Unique Identifier) will be automatically generated for the component.
- Save the component.



After configuring and setting up the Pipeline Workflow, it's time to Update and activate the pipeline.

- Locate "Update Pipeline" icon in the toolbar and click on it.
- Now, click on the 'Activate Pipeline button. This will Start the execution of the Pipeline and start the data processing.
- After activating the Pipeline, navigate to the logs section,
- Look for the Log Panel and click on it to access the advanced logs for detailed information.
- Within the Log Panel, you'll see the pods associated with each component. Pods are containers that hold the execution environment for the Pipeline.
- Check if the pods for each component have come up and are running. This indicates that the components are successfully deployed and ready to execute their tasks.
- To view the specific logs for each component, click on the corresponding pod or log entry. The logs will provide detailed information about the execution and any potential errors or issues encountered during the process.

IF API component started...

To ingest data from Postman and set up the ingestion ID and ingestion secrets same as, pipeline.

Follow these steps.

- Open Postman and ensure you have the necessary API requests set up to retrieve the hiring data.
- Locate the request or collection you want to use for data ingestion.

- Before sending the request, set the required headers or authentication parameters for the API.
- Use Hiring data json in body of the postman.
- Look for the headers or response properties that contain the ingestion ID and ingestion secrets and set the same component instance id URL and Click on send then you will get success as true.
- Go to Pipeline workflow and click on API component event.

Congratulations! You have successfully ingested the data. Now your pipeline is ready to securely retrieve and process the hiring data from the API.

Congratulations! Your pipeline is now ready to transform the hiring data, send it to the output event, and store it in ClickHouse DB. Additionally, the same ingestion ID, GUID, and secrets will be used in the dashboard for Realtime data streaming. Access the dashboard where you want to stream the data in real-time. Configure the dashboard to utilize the same ingestion ID, GUID, and secrets for data streaming. This ensures that the transformed data from the pipeline is seamlessly integrated into the dashboard for live updates.

Analyze the logs for each component to ensure that they are functioning as expected and that there are no errors or failures reported. If any issues arise, troubleshoot them accordingly.

Click on the data preview option inside component. This will display a sample of the records that the component will produce as its output during the execution. Analyze the sample records to verify that the component is generating the expected output. You can examine the data structure, values, and any applied transformations or filters.

Hope you are able to create pipeline using Pipeline components. Thankyou

BDB 8.2: Data Pipeline Workflow – 5

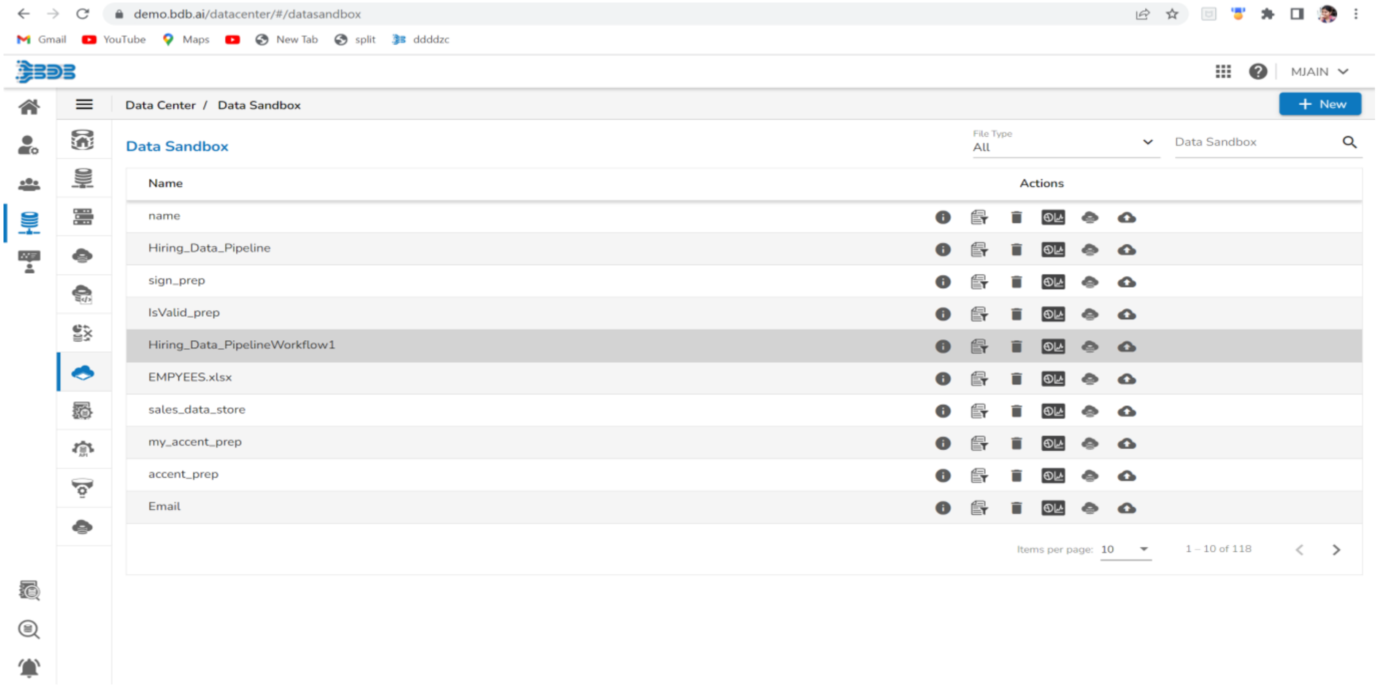
Workflow 5 is a streamlined process for handling hiring data. It starts by receiving the data from a specified Kafka topic. Next, the data undergoes data preparation to ensure its quality and consistency. AutoML techniques are then applied to extract valuable insights from the data. Finally, the processed data is stored in a database for further analysis and easy retrieval. This workflow enables organizations to efficiently handle hiring data, utilize machine learning to uncover insights, and maintain a structured and accessible data repository.

It allows you to connect to the Data Center, Data Preparation and Data Pipeline and AutoML Plugins

Let's Understand process of creating a new data sandbox and performing tasks on the data.

To begin,

- click on the 'Apps' menu, which will display a list of available modules.
- From the menu, choose the 'Data Center' module. This will take you to the Data Center page, where you can manage your data.
- Now, let's navigate to the Sandbox section, which is dedicated to managing sandboxes for your data.
- Look for the 'Sandbox' section or tab within the Data Center page.
- Once you've found the Sandbox section, you should see an option to create a new sandbox. Click on that
- Now, let's give the sandbox a name. Choose a name that will help you identify it later. For example, let's name it 'Hiring_data_PipelineWorkflow1'
- Provide an appropriate description if needed. This can help provide more context or details about the purpose of the sandbox.
- Next, you'll need to choose your data file. Look for an option to upload or select a data file for the sandbox.



Name	Actions
name	
Hiring_Data_Pipeline	
sign_prep	
IsValid_prep	
Hiring_Data_PipelineWorkflow1	
EMPYEES.xlsx	
sales_data_store	
my_accent_prep	
accent_prep	
Email	

Items per page: 10 1 – 10 of 118

you've successfully uploaded the data file to the sandbox. Now, let's proceed to the next step, which is Preparation. In this step, we'll prepare the data for further analysis or processing.

Create Preparation Using DataPreparation:

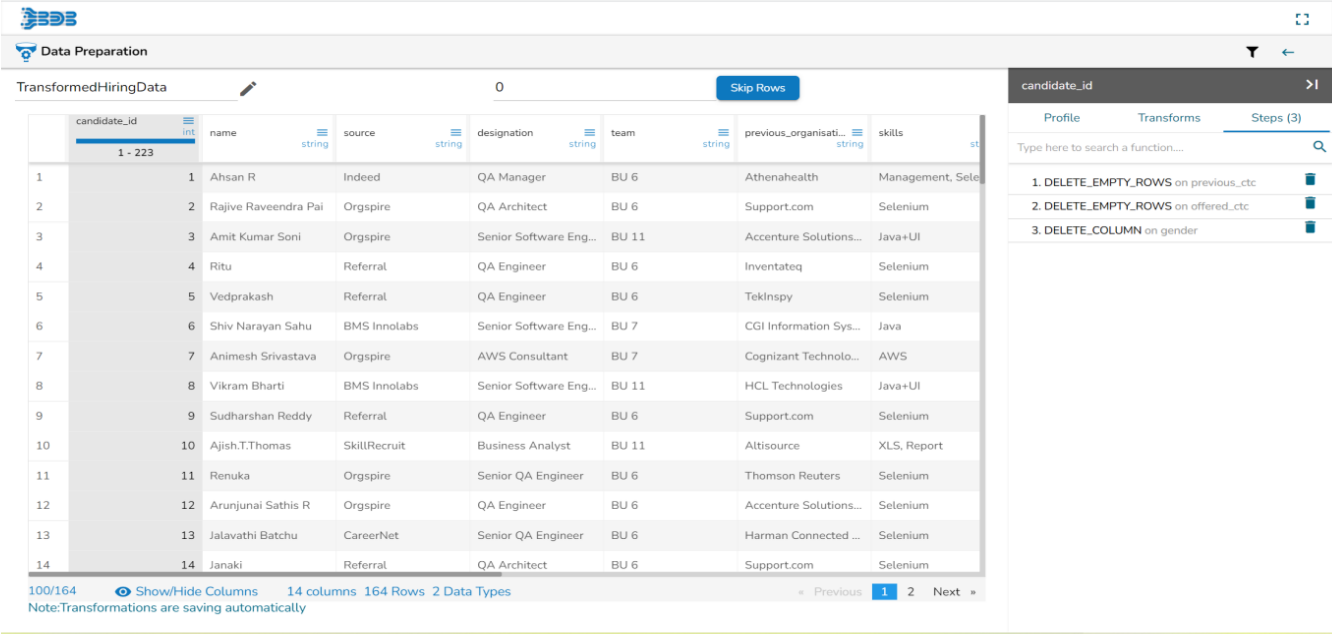
Data Preparation is used to clean the data.

Are you ready to prepare your hiring data pipeline sandbox file? Let's get started with the data preparation process.

- Select Data Preparation Icon. will navigate to the Data preparation home page.
- Here on the Data Preparation home page, you can see your complete dataset displayed in a grid form. The Data Preparation Plugin automatically profiles the data,
- providing valuable insights into its characteristics and detecting any anomaly data. You can also view the data profiling details.
- On the right-hand side, you'll find the selected column's profile. Here, you can explore various details such as charts, information, and patterns associated with the selected column.
- All the Transformations will appear inside transform tab.

Transforms

- Now, let's start preparing and cleaning the data. To remove the empty cells in the 'Previous CTC' column, just select prevoius ctc column and
- navigate to the 'Transforms' tab and search for the 'Delete Empty Rows with Empty Cells' transform. Click on it to remove all the empty rows from the 'Previous CTC' column."
- You can see that the empty rows in the 'Previous CTC' column have been removed.
- Next, let's perform the same transformation on the 'Offered CTC' column.
- Select the column and search for the 'Delete Rows with Empty Cell' transform. Click on it to remove the empty rows from the 'Offered CTC' column.
- Great! The empty rows in the 'Offered CTC' column have been successfully removed."
- Now, let's delete the 'Gender' column from the dataset. Simply select the column and search for the 'Delete Column' transform. Click on it to remove the 'Gender' column.
- Perfect! The 'Gender' column has been deleted from the dataset.
- You can see that all the performed transforms are recorded in the 'Steps' section. This helps you keep track of the changes made to the dataset.
- Now, let's rename the preparation for identification purposes. Simply click on the edit icon and give it a new name, such as 'TransformedHiringData
- Great! The preparation has been renamed to 'TransformedHiringData.



TransformedHiringData

0

	candidate_id int 1 - 223	name string	source string	designation string	team string	previous_organisati... string	skills st
1	1	Ahsan R	Indeed	QA Manager	BU 6	Athenahealth	Management, Sele
2	2	Rajive Raveendra Pai	Orgspire	QA Architect	BU 6	Support.com	Selenium
3	3	Amit Kumar Soni	Orgspire	Senior Software Eng...	BU 11	Accenture Solutions...	Java+UI
4	4	Ritu	Referral	QA Engineer	BU 6	Inventateq	Selenium
5	5	Vedprakash	Referral	QA Engineer	BU 6	TekInspy	Selenium
6	6	Shiv Narayan Sahu	BMS Innolabs	Senior Software Eng...	BU 7	CGI Information Sys...	Java
7	7	Animesh Srivastava	Orgspire	AWS Consultant	BU 7	Cognizant Technolo...	AWS
8	8	Vikram Bharti	BMS Innolabs	Senior Software Eng...	BU 11	HCL Technologies	Java+UI
9	9	Sudharshan Reddy	Referral	QA Engineer	BU 6	Support.com	Selenium
10	10	Ajish.T.Thomas	SkillRecruit	Business Analyst	BU 11	Altisource	XLS, Report
11	11	Renuka	Orgspire	Senior QA Engineer	BU 6	Thomson Reuters	Selenium
12	12	Arunjunai Sathis R	Orgspire	QA Engineer	BU 6	Accenture Solutions...	Selenium
13	13	Jalavathi Batchu	CareerNet	Senior QA Engineer	BU 6	Harman Connected ...	Selenium
14	14	Janaki	Referral	QA Architect	BU 6	Support.com	Selenium

100/164 14 columns 164 Rows 2 Data Types

Note:Transformations are saving automatically

candidate_id

1. DELETE_EMPTY_ROWS on previous_ctc
2. DELETE_EMPTY_ROWS on offered_ctc
3. DELETE_COLUMN on gender

Click on the back icon, and the preparation will be automatically saved and exported to different plugins, such as the Data Pipeline or AutoML. You can choose to export the transformed data to various plugins, such as the Data Pipeline or AutoML, based on your needs.

Create AutoML Workflow using DS Lab

To create an Automl Experiment using the DS Lab Plugin, follow these steps:

- Open the DS Lab Plugin from the app menu.
- Navigate to the DS Lab home page where all existing projects are displayed.

Create Project:

- Click on the "Create Project" button to proceed to the project creation page.
- On the project creation page, provide the following mandatory fields:
- Project Name: Enter a name for your project. For example, "Hiring Data".
- Project Description: Add a description for your project, such as "Creating AutoML experiment for hiring data prediction".
- Select Algorithm Types: Choose the algorithm types that are required for your project from the available options in the dropdown menu.
- Select Environment: Choose the environment for your project. In this case, select "PythonTensorFlow".
- Specify Resource Allocation: Select the resource allocation based on the volume of data and the computational requirements of your project. Choose the appropriate option, such as "Medium".

- Select Idle Shutdown Time: Choose the duration after which the system should shut down if idle. For example, select "1 hr".
- Specify External Libraries: If you require any external libraries for your project, mention them in this field.
- Mention GPU Type: If your project requires GPU acceleration, specify the GPU type.
- Once you have filled in all the necessary details, click on the "Save" button to create the project.

You should see a message confirming that the project has been successfully created.

Create AutoML Experiment

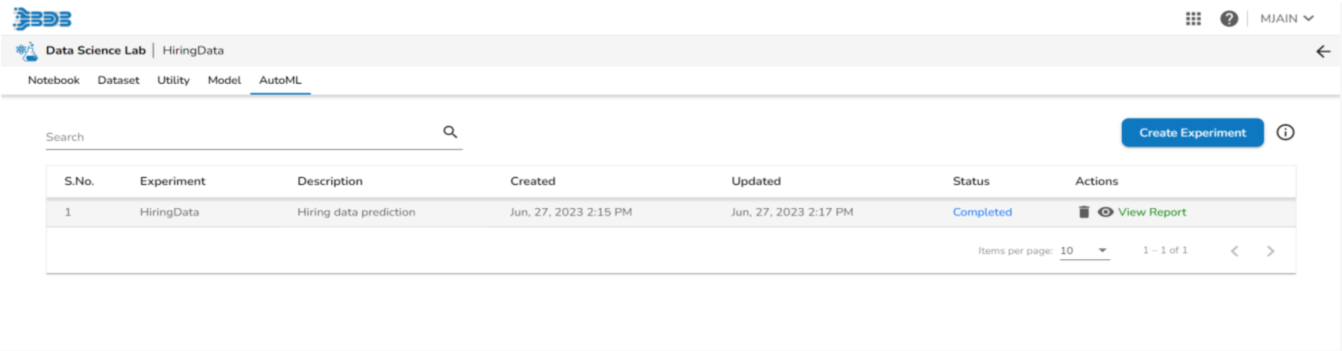
To create an AutoML experiment using the recently created "Hiring Data" project and follow these steps:

- Navigate to the "Datasets" tab in Data Science Lab.
- Click on the "Add Dataset" button.
- Select the "Data Sandbox" option from the data source dropdown menu. This will display all the files that have been uploaded into the sandbox.
- Choose the "Hiring_data_Pipeline" file.
- Click on the "Add" button to successfully add the dataset.
- Many action items will be displayed for the selected dataset, such as preview, data profile, create experiment, delete, and data preparation.
- Choose the "Create Experiment" option for the uploaded data file.
- Provide an experiment name, for example, "HiringData," and a description, such as "Hiring data prediction."
- Select the target column from the dropdown menu. You can also choose any data preparation options if required. Click on the "Next" button to proceed.
- Select the experiment type based on your requirements, such as classification or regression. In this case, select "Classification" as the gender column contains categories.
- Click on "Done" to start the experiment.
- Wait for the experiment to complete in the AutoML tab. You can track the progress as the status changes from "Started" to "Running."
- Once the experiment is completed, open and view the report. The report will display the recommended model and a run summary.

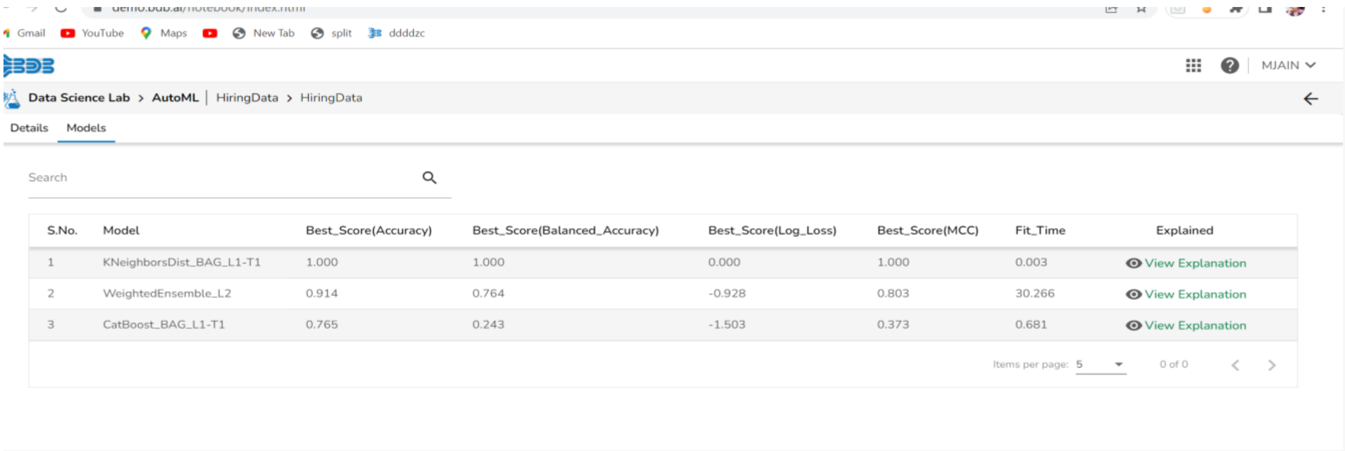
Register Model

- Switch to the "Models" tab to explore the top three models trained using AutoML.
- Click on "View Explanation" to understand the model's performance and switch to the "Dataset Explorer" tab to see the data profile.
- Navigate to the "Models" tab and expand the saved models.

- Export the models to the data pipeline using the register icon.
- Optionally, export the model into Git for version control and collaboration purposes.



S.No.	Experiment	Description	Created	Updated	Status	Actions
1	HiringData	Hiring data prediction	Jun, 27, 2023 2:15 PM	Jun, 27, 2023 2:17 PM	Completed	View Report

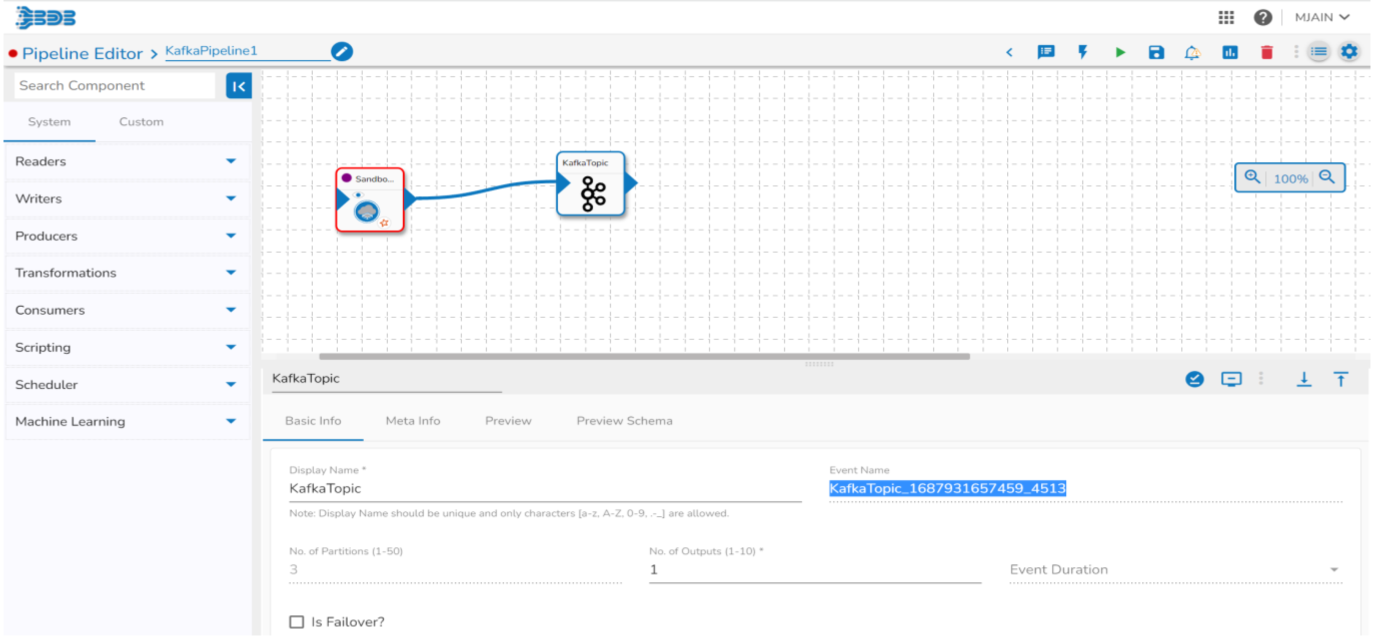

demo.bdb.ai/notebook/index.html
Gmail
YouTube
Maps
New Tab
split
ddddzc
MJAIN
Data Science Lab > AutoML | HiringData > HiringData
Details
Models
Search
Items per page: 5
0 of 0

S.No.	Model	Best_Score(Accuracy)	Best_Score(Balanced_Accuracy)	Best_Score(Log_Loss)	Best_Score(MCC)	Fit_Time	Explained
1	KNeighborsDist_BAG_L1-T1	1.000	1.000	0.000	1.000	0.003	View Explanation
2	WeightedEnsemble_L2	0.914	0.764	-0.928	0.803	30.266	View Explanation
3	CatBoost_BAG_L1-T1	0.765	0.243	-1.503	0.373	0.681	View Explanation

By following these steps, you should be able to create an AutoML experiment using the "Hiring Data" project and explore the recommended models in Data Science Lab.

Pipeline Workflow:

- Before Creating main Pipeline create kafkaPipeline1 with sandbox reader component and add event.
- Click on event and copy event name from the event name text box for ex. KafkaTopic_1687931657459_4513
- Update and Activate pipeline.



KafkaTopic

Basic Info | Meta Info | Preview | Preview Schema

Display Name *
KafkaTopic
Note: Display Name should be unique and only characters [a-z, A-Z, 0-9, .-_] are allowed.

Event Name
KafkaTopic_1687931657459_4513

No. of Partitions (1-50)
3

No. of Outputs (1-10) *
1

Event Duration

□ Is Failover?

Now, Let's Understand process of creating Kafka Data pipeline workflow.

- First, locate and select the Data Pipeline Plugin from the app menu. This will take you to the pipeline home page.
- "Next, look for the create icon and click on it. You'll see the option to create a new pipeline. Click on the '+' icon to proceed.
- Great! Now, let's specify the details for our end-to-end kafka data processing pipeline. Enter a suitable name for your pipeline, such as Kafka Pipeline_Workflow5. For the description, briefly describe the workflow, such as Kafka Data Integration Pipeline
- Now, choose the resource allocation type based on your requirements. This feature allows you to deploy the pipeline with high, medium, or low-end configurations, depending on the velocity and volume of data that the pipeline must handle."
- Once you're done with the configuration, click on the save button to save your pipeline."

Congratulations! You've successfully created the pipeline.

Once the pipeline is saved in the pipeline list, you can add components to the canvas to create your pipeline workflow or dataflow. To add a component, simply drag the required component from the Component Palette, located on the left side of the user interface, and drop it onto the canvas. You can configure each component to define your pipeline workflow. The Pipeline Editor displays the Component Palette, which contains various components like Reader, Writer, Transformation, Consumer, Producer, Machine Learning, and more. Use these components to design your pipeline according to your specific requirements.

- Locate the Event Panel: Look for the Event Panel icon in the toolbar and click on it.

- Add a New Event: Click on the "+ icon" or any designated button to add a new event. This action will open the Add Event popup or dialog box.
- Configure Event Mapping: In the Add Event popup, enable the Event Mapping option by sliding the toggle button. This option allows you to map the event to a specific Event, in this case, the Kafka topic.
- Specify Kafka Topic: specify event/kafka topic e.g., “KafkaTopic_1687931657459_4513” Use the search icon or input field in the Add Event popup to specify the Kafka topic to which the event should be mapped.
- Event Mapping Details: After specifying the Kafka topic, displays additional information about the event, such as event duration, number of partitions, or the number of output Fields.
- Map Event: Once you have configured the event details, click on the "Map Event" button in the Add Event popup to create the event and save the configuration.
- Drag and Drop Event: After the event is created and saved, you can drag and drop it onto the canvas of your pipeline workflow.

let's add transform component..

- Drag and drop the DataPrep transform component onto the canvas.
- Configure the transform component:
- Set the Invocation Type as "Realtime."
- Move to the Meta Information tab.
- Select "Data Sandbox" from the Data Center Type dropdown.
- Choose the sandbox file from the Sandbox Name dropdown that you added to the data center sandbox.
- Select "Preparation" from the Preparation dropdown and choose the "transformeddatapreparation" option.
- Save the component.
- Now, let's move to the Event Panel to create an event and connect it with the component:
- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "+" icon.
- You can change the displayed event name and click on add event button.
- Drag and drop the event onto the canvas.
- Connect the in-event and out event with the transform component.

You have successfully added the transform component, configured it, and connected it to an event in the canvas.

Configure the AutoML component with the following settings:

- Drag and drop the Automl Component from Machine Learning component palette onto the canvas.
- Configure the component:
- Set the invocation type as "batch." and move to meta information.
- Specify the project name by selecting it from the "Project Name" dropdown. This assumes you have already created a project within your AutoML plugin.
- Choose the model you want to use from the "Model Name" dropdown. This assumes you have previously registered or trained models using AutoML.
- Save the component configuration. click on save button.

create an event and connect it with the component:

- Click on the Event Panel icon located in the toolbar.
- Add a new event by clicking on the "+" icon.
- You can change the displayed event name and click on add event button.
- Drag and drop the event onto the canvas.

Connect the in-event and out event with component.

Now, it's time to select the appropriate writer component for your destination.

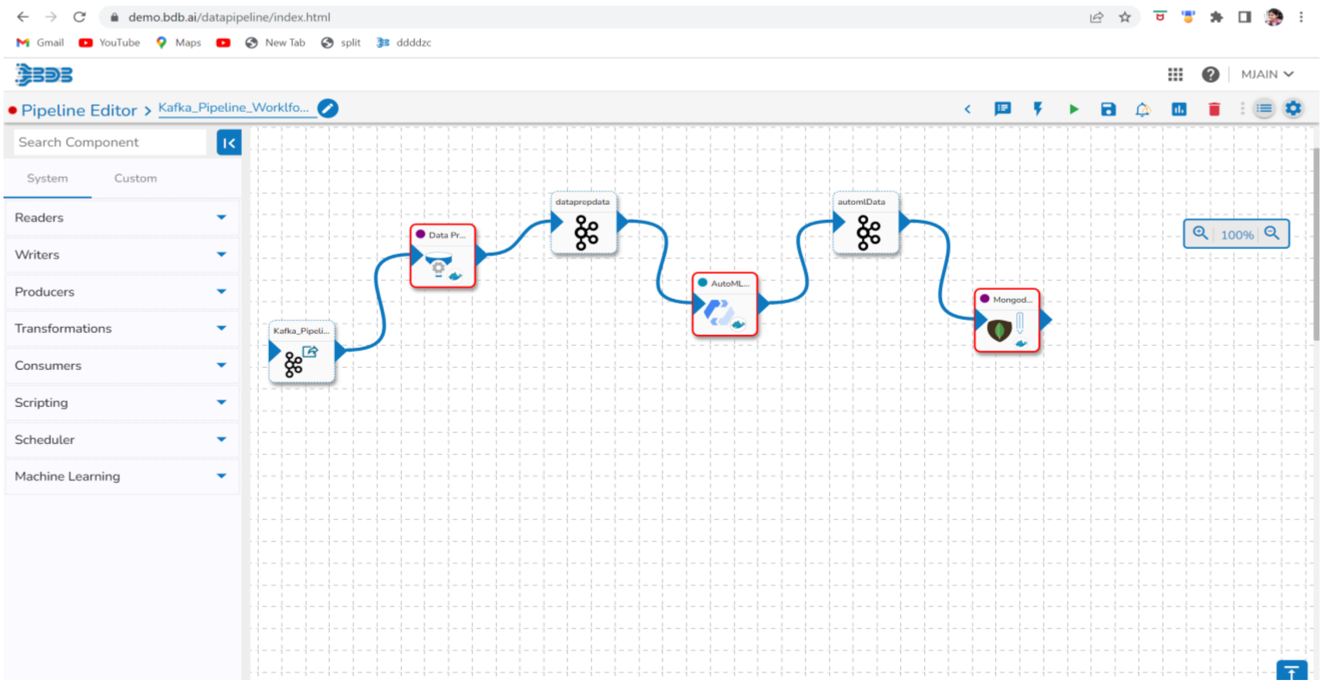
- Locate the MongoDB WRITER LITE Component: In your pipeline development environment or component palette, find and select the MongoDB WRITER LITE component. This component is specifically designed to write data to a MongoDB database.
- Drag and drop the MongoDB WRITER LITE Component: Drag the MongoDB WRITER LITE component from the component palette onto the canvas of your pipeline workflow.
- Connect the Event to the MongoDB WRITER LITE Component: Drag a connection line from the event (previously created and added to the canvas) and connect it to the input of the MongoDB WRITER LITE component. This connection signifies that the event data will flow into the MongoDB writer component for further processing.
- Configure the MongoDB WRITER LITE Component: click on the MongoDB WRITER LITE component to open its configuration settings. Select invocation type as realtime and move t o the meta information and configure the necessary parameters such as the MongoDB connection details, database name, collection name, and any additional settings required for your specific use case.

To specify the connection string, database name, collection name, save mode, and composite key for the MongoDB WRITER LITE component in your Kafka data pipeline workflow, follow these steps:

- Connection Type: Select "Connection String" from the dropdown menu.

- Connection String: In the "Connection String" textbox, enter the connection string for your MongoDB database. The connection string typically includes the MongoDB server address, port number, authentication credentials (if required), and any additional parameters.
- Database Name: Specify the name of the database you want to write the data into.
- Collection Name: Enter the name for the collection in which you want to store the data. For example, you can specify "HiringDataKafka" as the collection name.
- Save Mode: Choose the "Upsert" option from the "Save Mode" dropdown menu. Upsert allows you to update existing documents or insert new documents based on a specified composite key(candidate_id).

Once you have configured the Writer component with the necessary connection details and settings, click on the "Save" button to apply the configuration.



After configuring and setting up the Pipeline Workflow, it's time to Update and activate the pipeline.

- Locate "Update Pipeline" icon in the toolbar and click on it
- Now, click on the 'Activate Pipeline button. This will Start the execution of the Pipeline and start the data processing.
- After activating the Pipeline, navigate to the logs and advance Log section, look for the Log Panel and click on it to access the advanced logs for detailed information.
- Within the Log Panel, you'll see the pods associated with each component. Pods are containers that hold the execution environment for the Pipeline.

- Check if the pods for each component have come up and are running. This indicates that the components are successfully deployed and ready to execute their tasks.

To view the specific logs for each component, click on the corresponding pod or log entry. The logs will provide detailed information about the execution and any potential errors or issues encountered during the process. Analyze the logs for each component to ensure that they are functioning as expected and that there are no errors or failures reported. If any issues arise, troubleshoot them accordingly."

Data Preview

- Click on the Event: In the pipeline canvas or component configuration page, locate the specific event you want to analyze. Click on the event to select it.
- Navigate to the Meta Info Tab: Look for the Meta Info tab within the event configuration panel. Click on it to access the meta-information related to the event.
- Meta Info Details: In the Meta Info tab, you will find details such as the number of records processed by the component, data size, and partition details. This information provides insights into the volume and distribution of the processed data.
- Move to the Preview Tab: Switch to the Preview tab within the event configuration panel. This tab allows you to view a sample of the records that the component will produce as its output during execution.
- Analyze the Sample Records: In the Preview tab, you will see a preview of the output records generated by the component. Take a close look at the data structure, values, and any applied transformations or predictions. This will help you verify that the component is producing the expected output.

THANK YOU

For More Information,
Contact:

Email id: Sales@bdb.ai