Chúng tôi cung cấp 1 giải pháp phân tích dữ liệu tối ưu cho các doanh nghiệp và cá nhân mang tên HD InfoSystem. Một cách tổng quát, đây là một hệ thống quản lý, lưu trữ và phân tích dữ liệu dựa trên các mô hình AI, nhằm mang tới một môi trường làm việc với dữ liệu tiện lợi và thân thiện nhất ngay cả đối với những người dùng không chuyên nhờ việc mọi thứ sẽ được thao tác trên giao diện web, trực quan, nhanh và hiệu quả.

Đi vào chi tiết, người dùng hệ thống của chúng tôi sẽ đi qua một quá trình gồm 8 phần, theo đúng thứ tự hoặc tùy chỉnh theo nhu cầu:  
1. Select Data Information ( Thông tin dữ liệu)

2. Preview ( Xem trước)

3. Upload data (Tải lên dữ liệu)

4. Data Insight (Thông tin chi tiết về dữ liệu)

5. Data Visualization ( Trực quan hóa dữ liệu bằng biểu đồ)

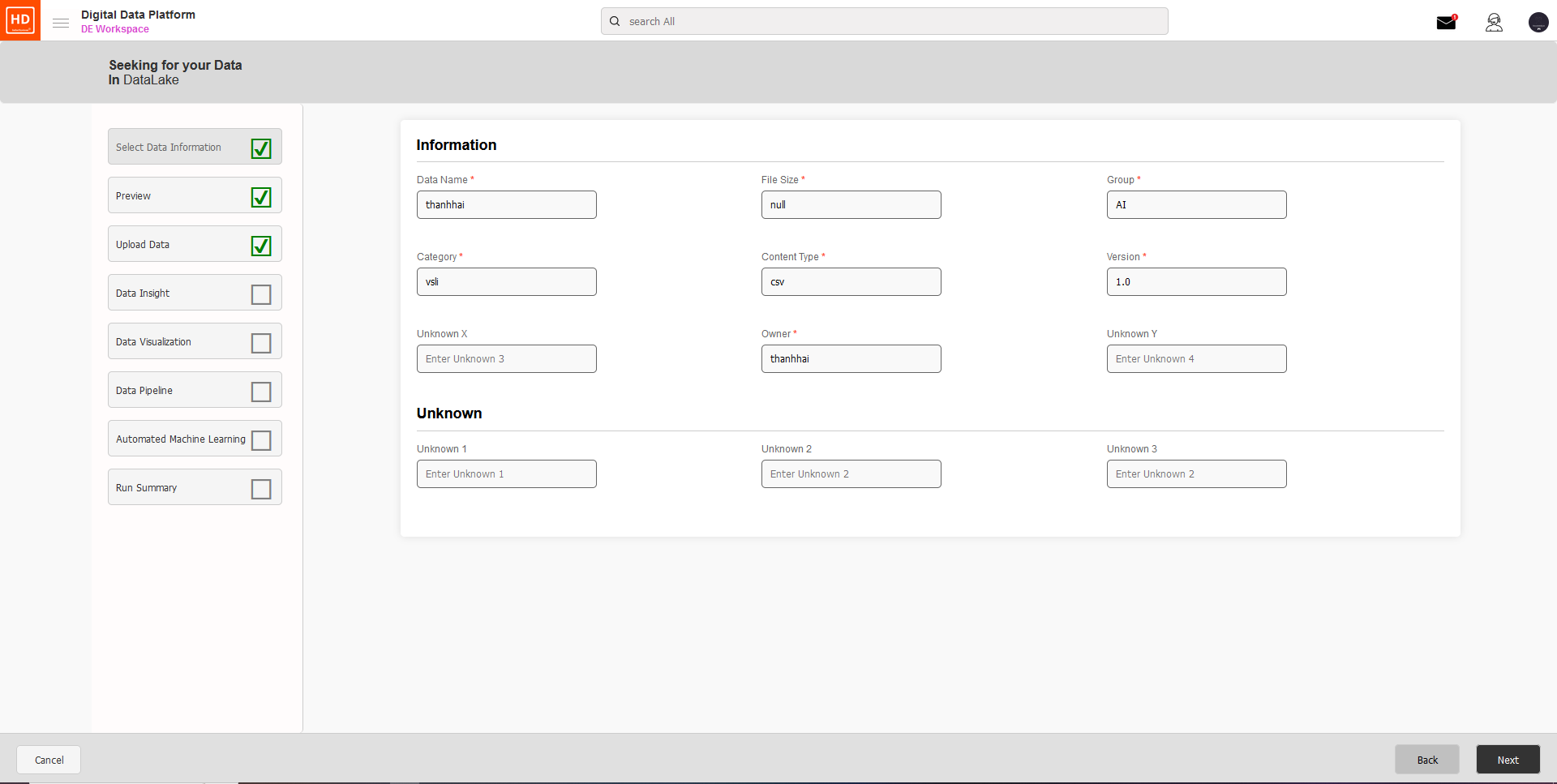
6. Data Pipeline ( Đường ống dữ liệu)

7, Automated Machine Learning (Học máy tự động - Sử dụng các mô hình AI)

8. Run Summary ( Tổng kết quá trình và hiển thị các kết quả phân tích dữ liệu)

Sau đây, chúng ta sẽ cùng đi vào chi tiết từng quá trình:

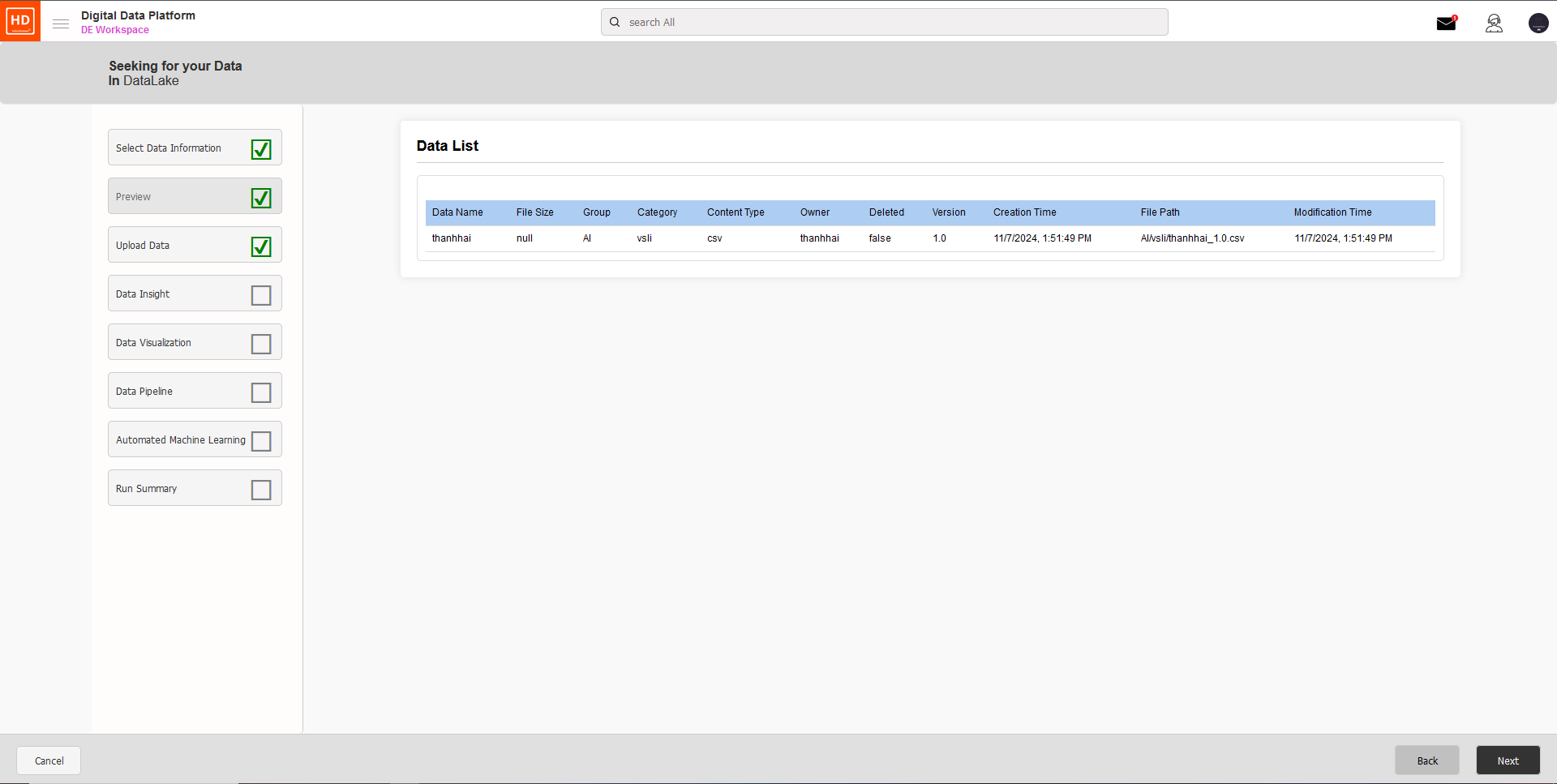
1. Select Data Information



Đầu tiên Select Data Information hay còn gọi là phần cung cấp các thông tin về dữ liệu cho hệ thống. Ở đây, người dùng có thể điền vào những thông tin sơ bộ và đặc trưng của dữ liệu, nhằm phân biệt, nhận biết và phục vụ cho mục đích lưu trữ và quản lý dữ liệu. Các thông tin bắt buộc yêu cầu là:

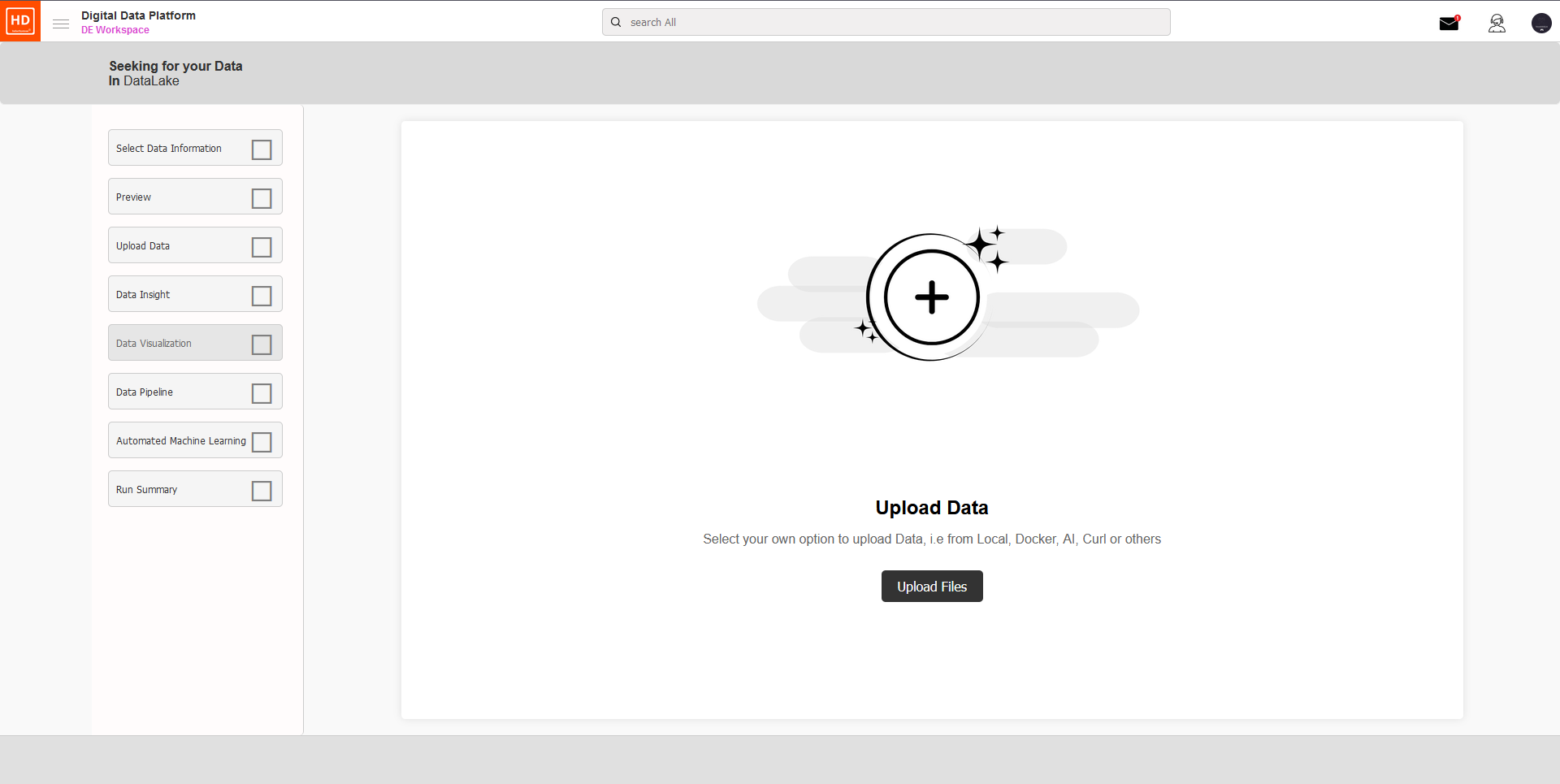
* Data Name( tên gọi duy nhất của bộ dữ liệu),
* File size ( kích thước của file dữ liệu),
* Group (nhóm phân loại của dữ liệu)
* Category (thể loại)
* Content Type(kiểu dữ liệu)
* Version( Phiên bản)
* Owner( người sở hữu, người tải lên hay là người được quyền và cấp quyền truy cập, quản lý là phân tích bộ dữ liệu).

1. Preview

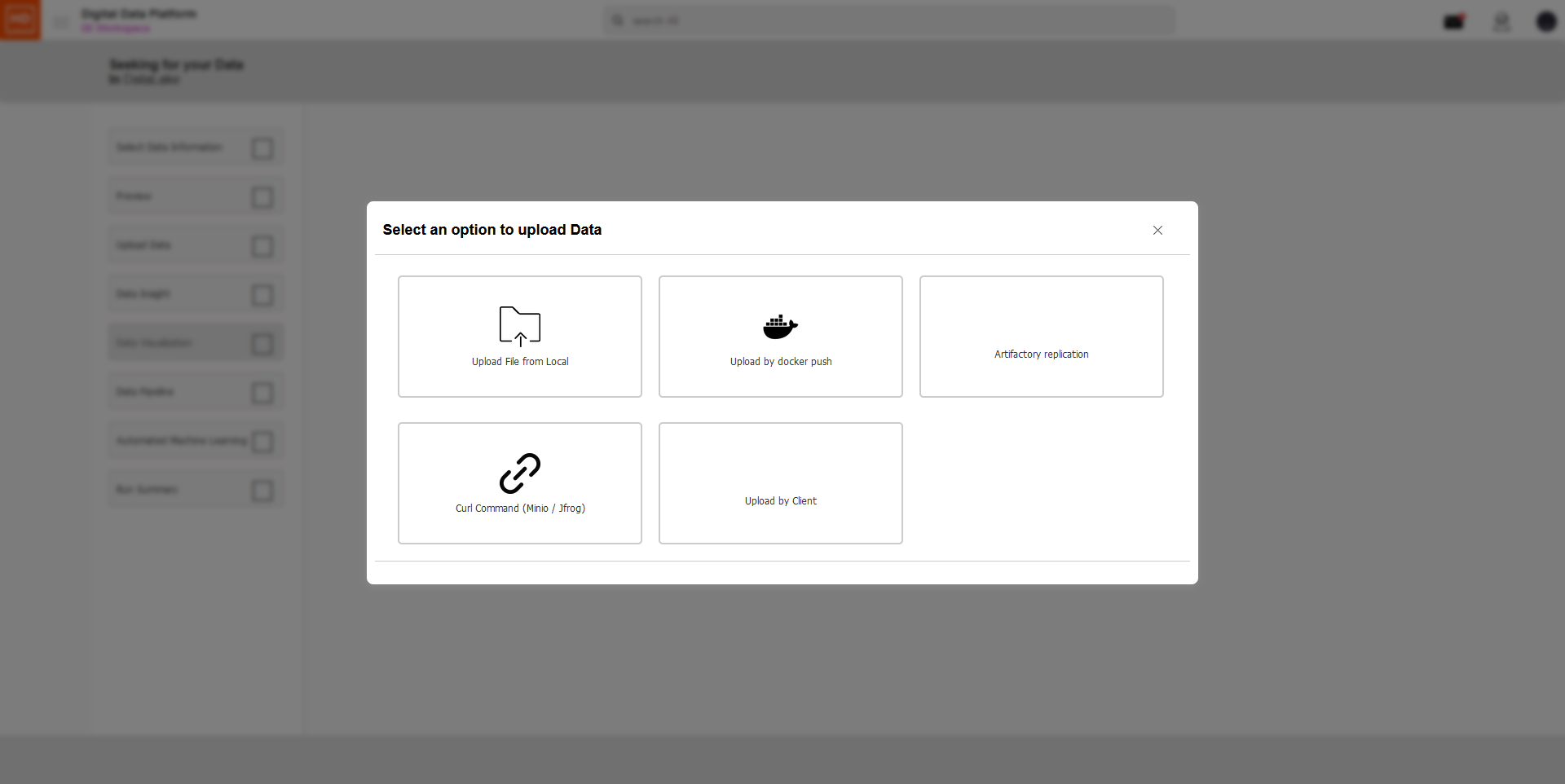


Preview hay Xem trước là phần người dùng có thể xem trước được dữ liệu của mình đã được tạo hay chưa, các thông tin người dùng khai báo đã đúng hay chưa. Bên cạnh đó, hệ thống cũng cung cấp thêm thông tin về thời gian tạo dữ liệu, thời gian thay đổi sửa chữa thông tin dữ liệu và đường dẫn tới điểm lưu trữ dữ liệu trong các nền tảng lưu trữ thông tin dữ liệu, cụ thể ở đây là Postgresql.

1. Upload data



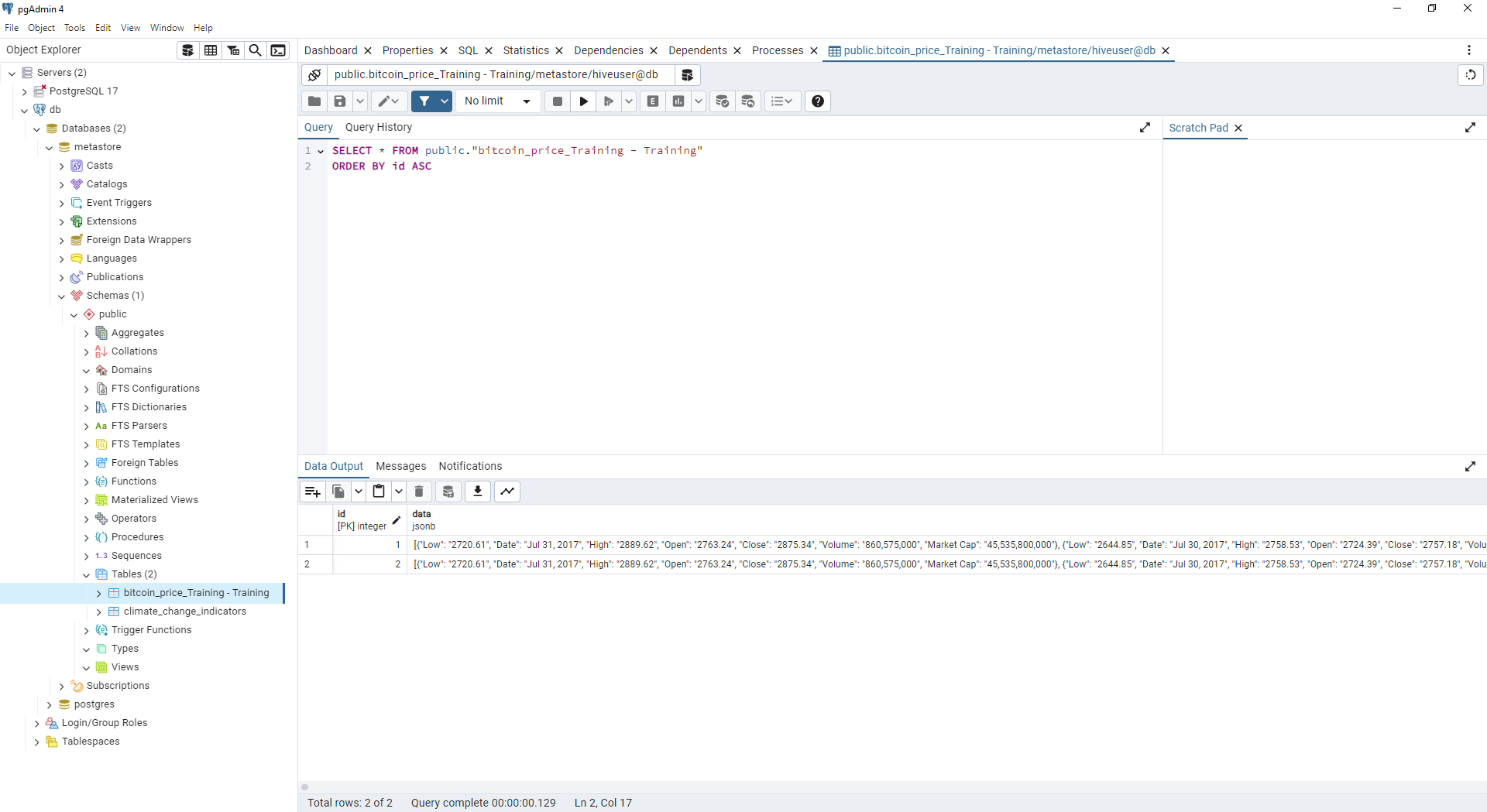
Upload data hay Tải lên bộ dữ liệu, là một trong những quá trình quan trọng nhất của hệ thống.



Người dùng có thể lựa chọn đa dạng các hình thức tải lên bộ dữ liệu của mình. Hiện tại, có 3 phương thức khả dụng là:

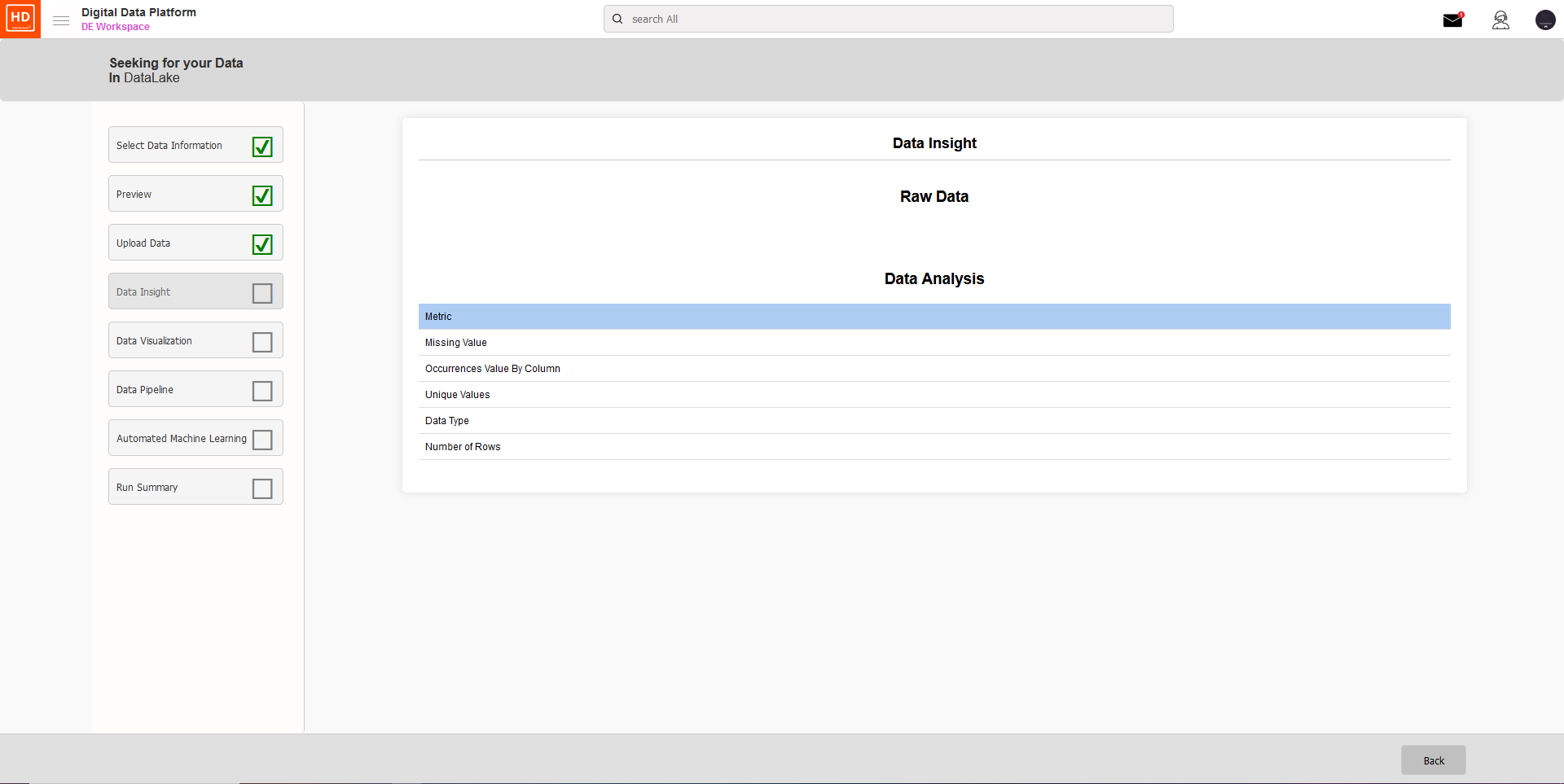
* Upload from Local Device ( Tải lên từ thiết bị)
* Upload by Docker Push ( Tải lên bằng Docker Push)
* Curl Command (Minio/Jfrog) ( Các mã lệnh Curl thông qua các hệ thống lưu trữ đối tượng như Minio và Jfrog).

Các phương thức tải lên dữ liệu khác sẽ được cập nhật và điều chỉnh trong tương lai.



Dữ liệu được tải lên sẽ được lưu trữ trong Hadoop và Postgestql, và chắc chắn, có thể được truy cập và xem trước thông qua 2 nền tảng trên.

1. Data Insight

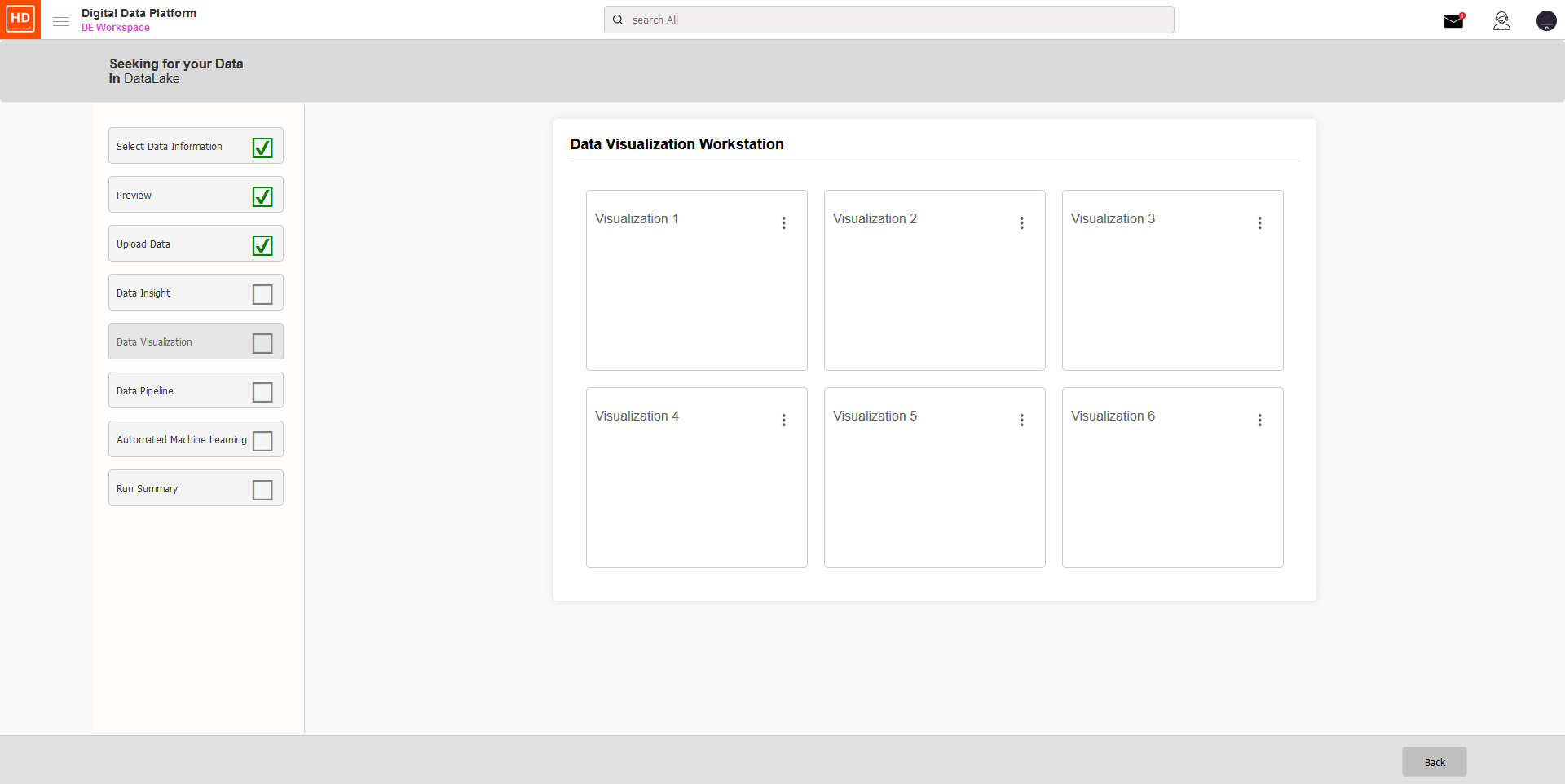


Data Insight hay thông tin chi tiết của dữ liệu, là quá trình giúp người dùng có cái nhìn cụ thể và chi tiết về các đối tượng, giá trị trong bộ dữ liệu của họ, thông qua thuật toán tiền phân tích dữ liệu.

Cụ thể, ở quá trình này, hệ thống sẽ hiển thị 2 bảng đối chiếu cho người dùng:

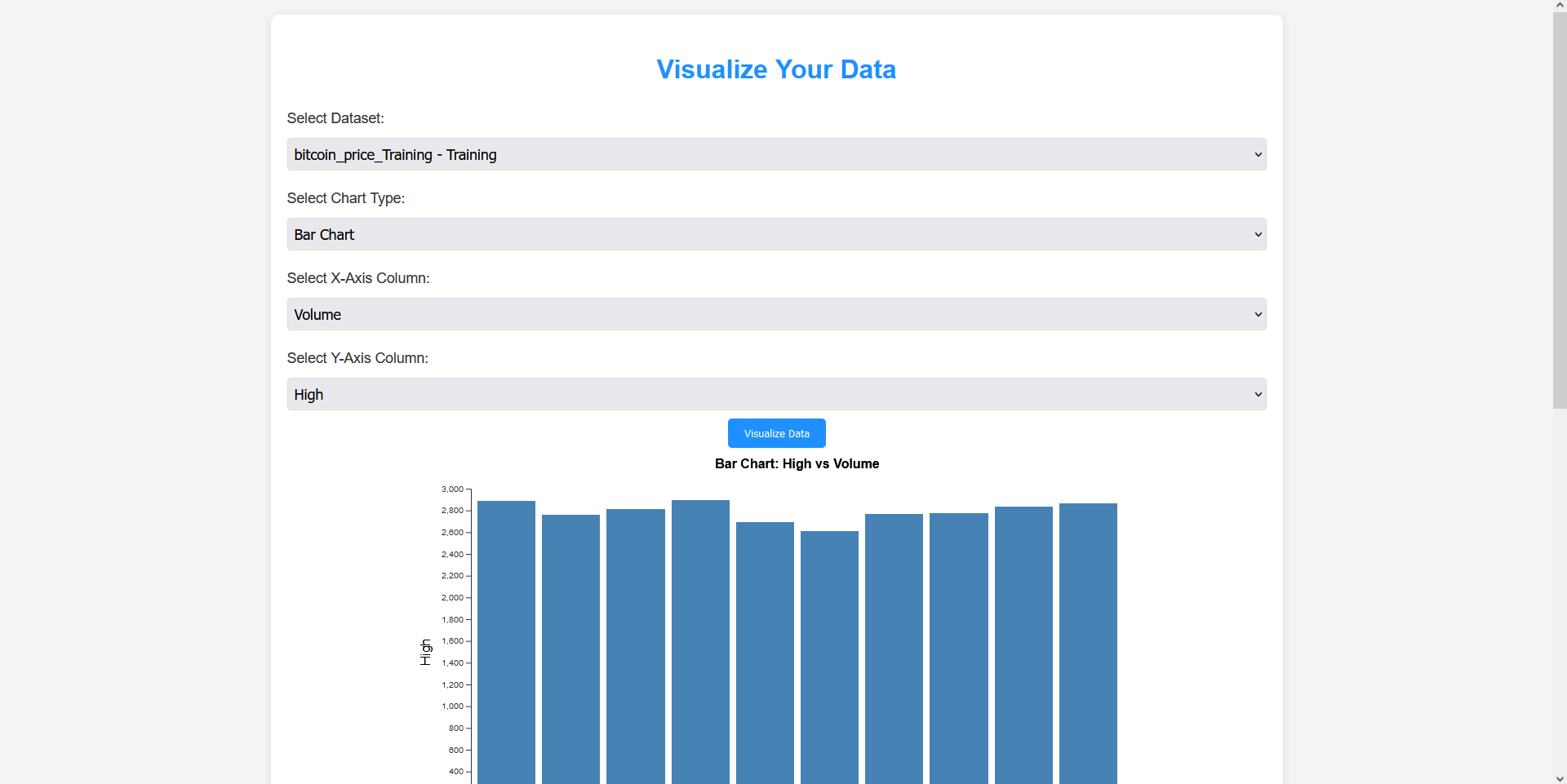
* Bảng đầu tiên, là bảng Raw Data, hay còn gọi là dữ liệu thô. Bảng này sẽ hiển thị chính xác nội dung của bộ dữ liệu mà người dùng tải lên. Từ các cột, hàng, giá trị và đối tượng
* Bảng thứ 2, là bảng Data Analysis, hay còn gọi là bảng phân tích dữ liệu. Ở đây, thuật toán của chúng tôi tự động phân tích và cung cấp 1 số thông tin, hệ mét của dữ liệu: Missing Values( Giá trị thiếu) , Occurence Values by column ( Giá trị xuất hiện theo cột) , Unique Values (Giá trị duy nhất) , Data Type (Kiểu dữ liệu) và Number of Rows (Số hàng).

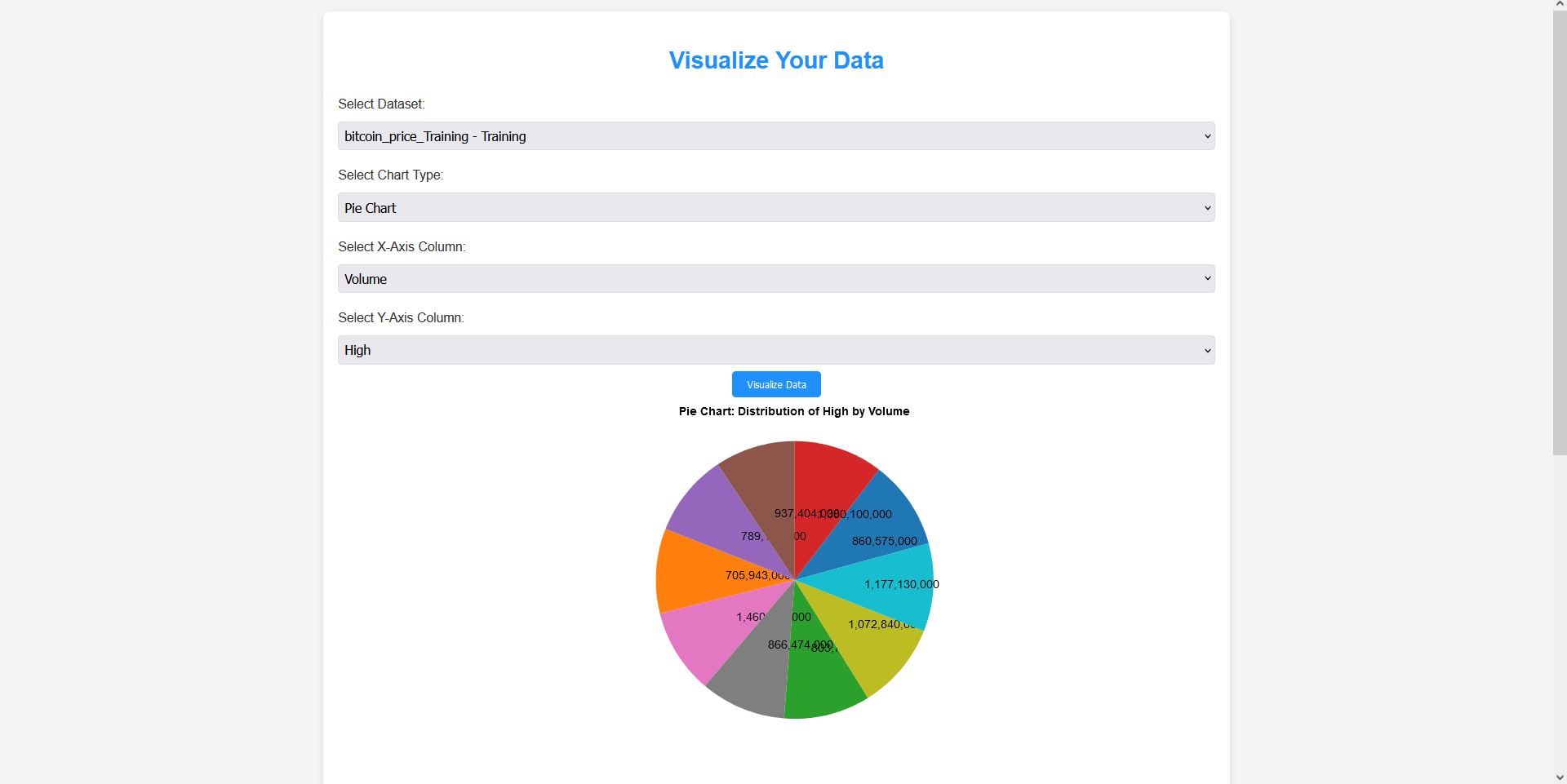
1. Data Visualization

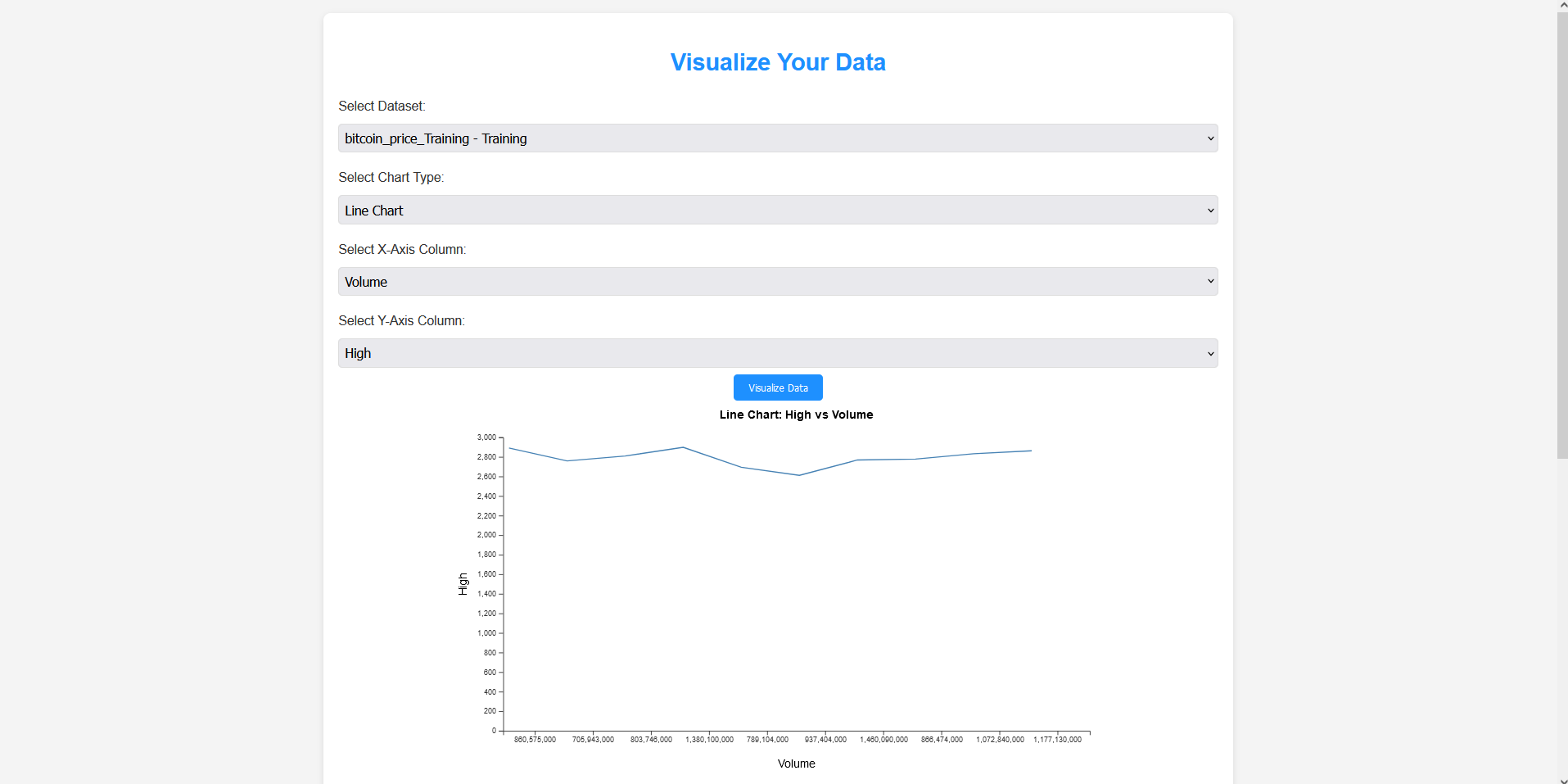


Data Visualization hay còn gọi là Trực quan hóa dữ liệu, là quá trình hệ thống sẽ vẽ biểu các biểu đồ để thể hiện các giá trị có trong bộ dữ liệu

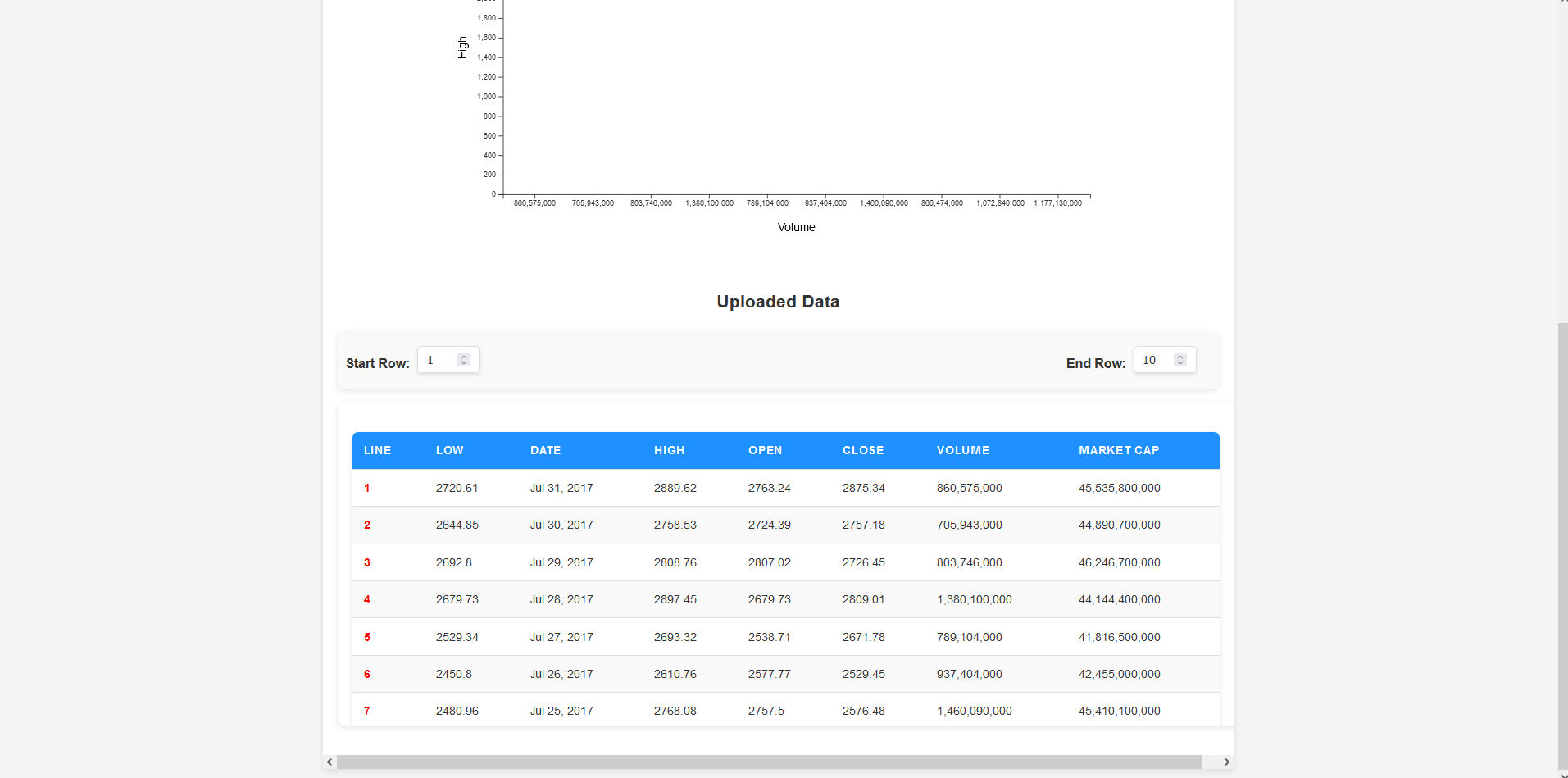
Ở phần này, hệ thống cung cấp tối đa 6 trạm làm việc cùng 1 lúc để người dùng có thể lựa chọn trực quan hóa tối đa 6 bộ dữ liệu cùng 1 lúc.



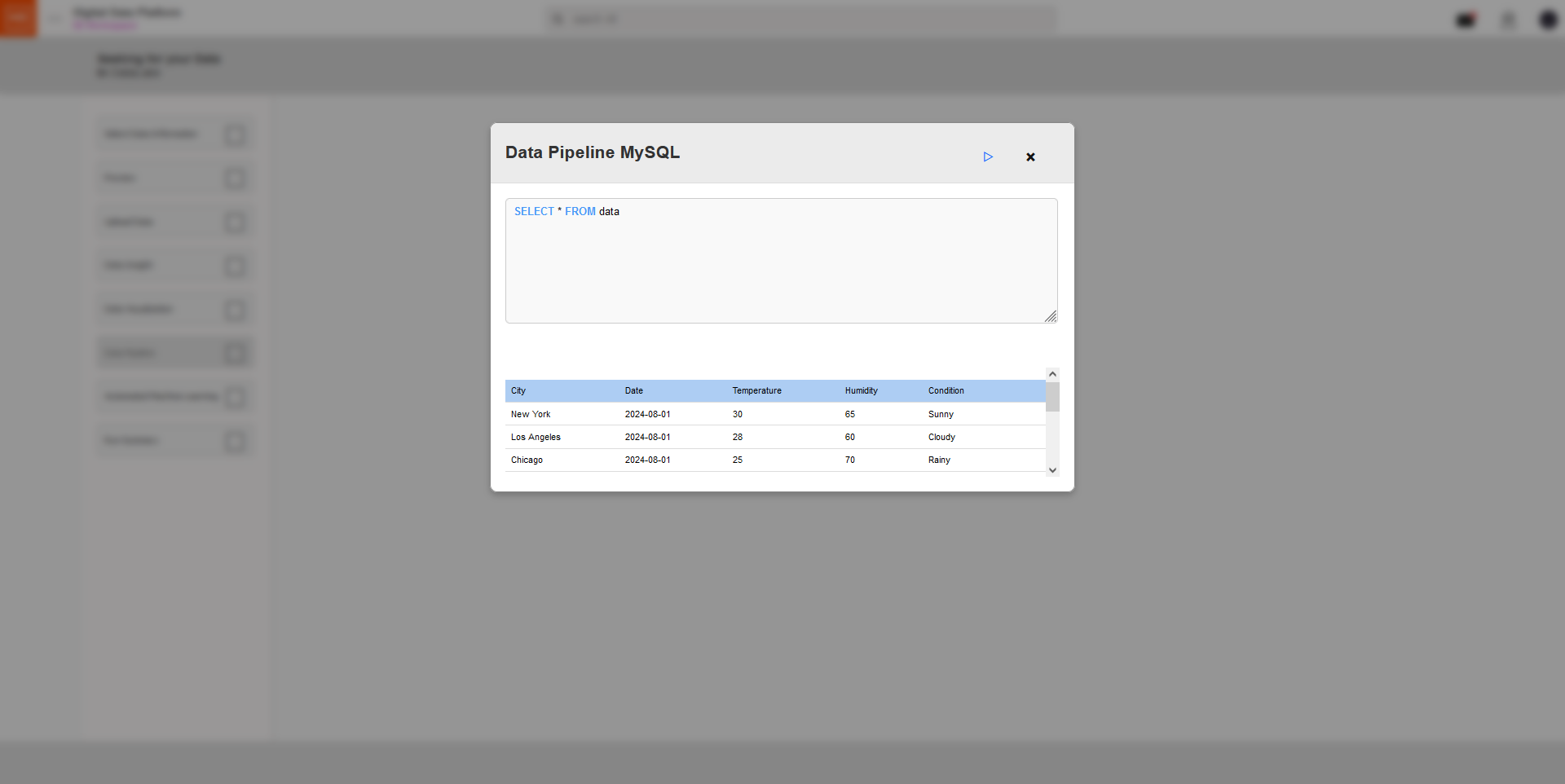




Trong mỗi trạm làm việc, người dùng đều có thể chọn để trực quan hóa bất kì 1 bộ dữ liệu nào đã được tải lên. Ngoài ra, còn có thể tùy chỉnh gán các cột làm trục X và trục Y của biểu đồ trực quan hóa, cũng như chọn loại biểu đồ để vẽ. Hiện tại chúng tôi cung cấp 3 loại biểu đồ trực quan hóa dữ liệu là Bar Chart ( Biểu đồ cột), Line Chart (Biểu đồ đường) và cuối cùng là Pie Chart ( Biểu đồ hình tròn).

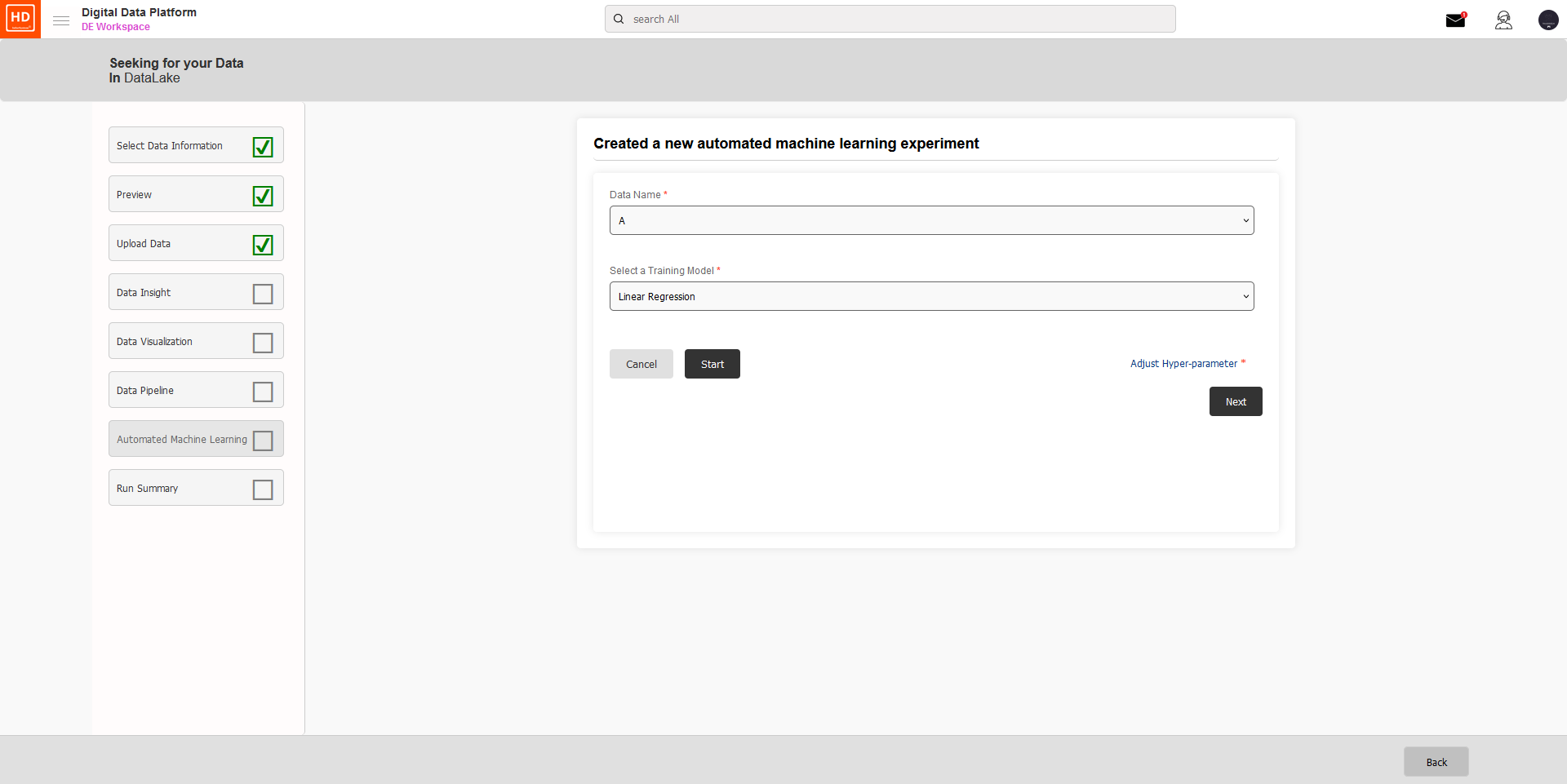


1. Data Pipeline



Ở quá trình Data Pipeline hay còn gọi là đường ống dữ liệu, hệ thống sẽ cung cấp một môi trường cho người dùng sử dụng Ngôn ngữ truy vấn có cấu trúc để quản lý cơ sở dữ liệu quan hệ. Ngôn ngữ được dùng ở đây sẽ giống như là Postgresql. Thực hiện các hàm truy vấn như trong Postgresql để truy xuất thông tin, giá trị của dữ liệu được tải lên

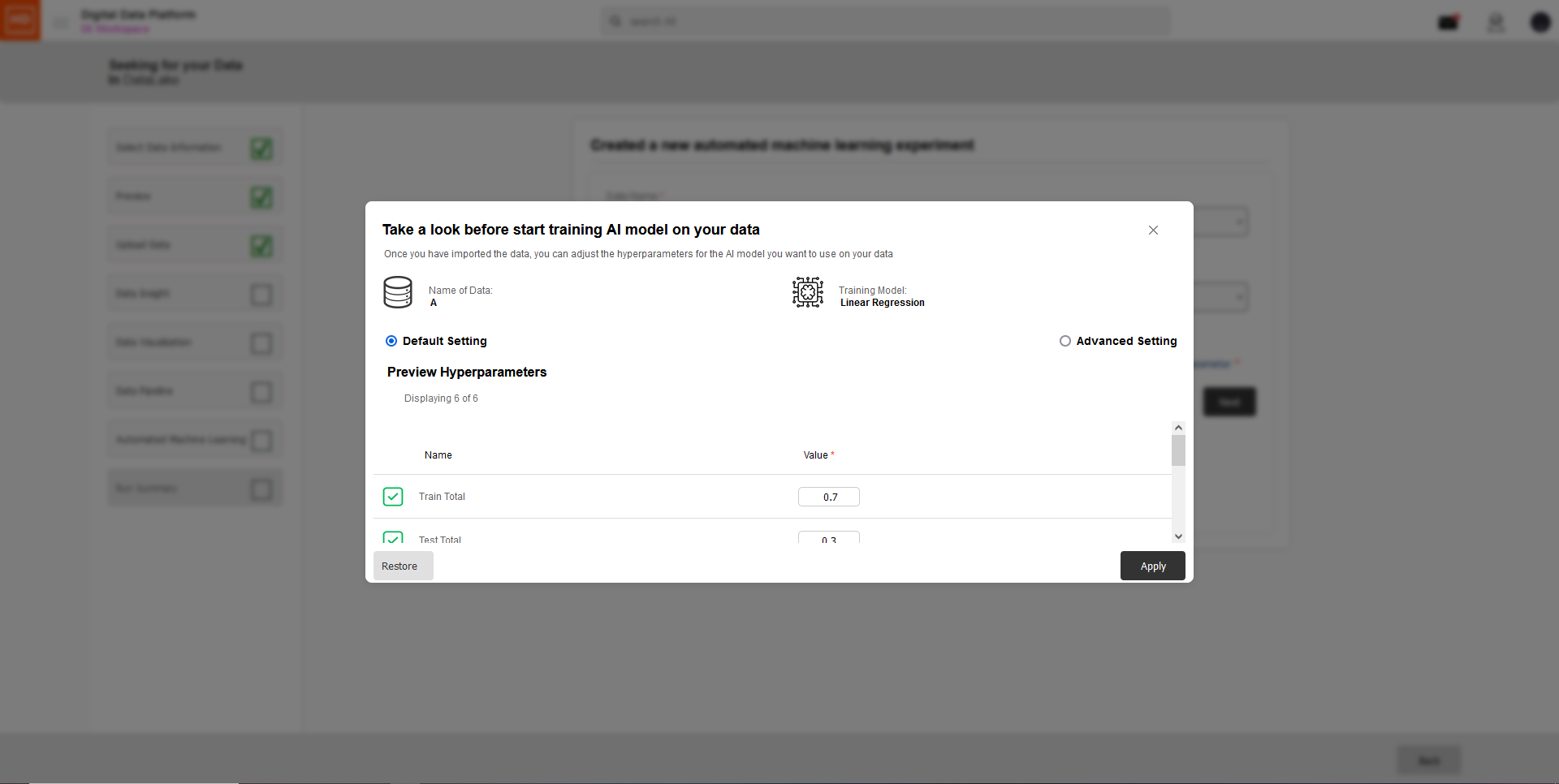
VII. Automated Machine Learning



Với bộ dữ liệu được tải lên từ các phần trước, Automated Machine Learning, hay Học máy tự động, là quá trình xử lý dữ liệu thô bằng các mô hình AI để mang tới những kết quả, dự đoán và phân tích chuyên sâu cho bất cứ đối tượng hay giá trị nào của bộ dữ liệu.

Các mô hình luyện AI chúng tôi cung cấp là:

* Linear Regression
* Logistic Regression
* …



Hệ thống cũng tích hợp sẵn chức năng mặc định thiết lập các Hyper Parameter ( siêu tham số) để người dùng sử dụng các mô hình AI trên dữ liệu của mình trong mục Default Setting ( Cài đặt mặc định) , tuy nhiên, người dùng vẫn có thể tùy biến các tham số này một cách thủ công bằng Advanced Setting (Cài đặt nâng cao) sao cho đạt được hiệu quả cao nhất cho việc phân tích dữ liệu.

VIII. Run Summary

Run Summary, hay Tổng kết quá trình, là nơi tổng hợp và lưu kết quả của tất cả các quá trình được thực hiện ở các quá trình trên, nhằm mục đích lưu trữ cho người dùng sử dụng với nhiều mục đích khác nhau trong tương lai.