

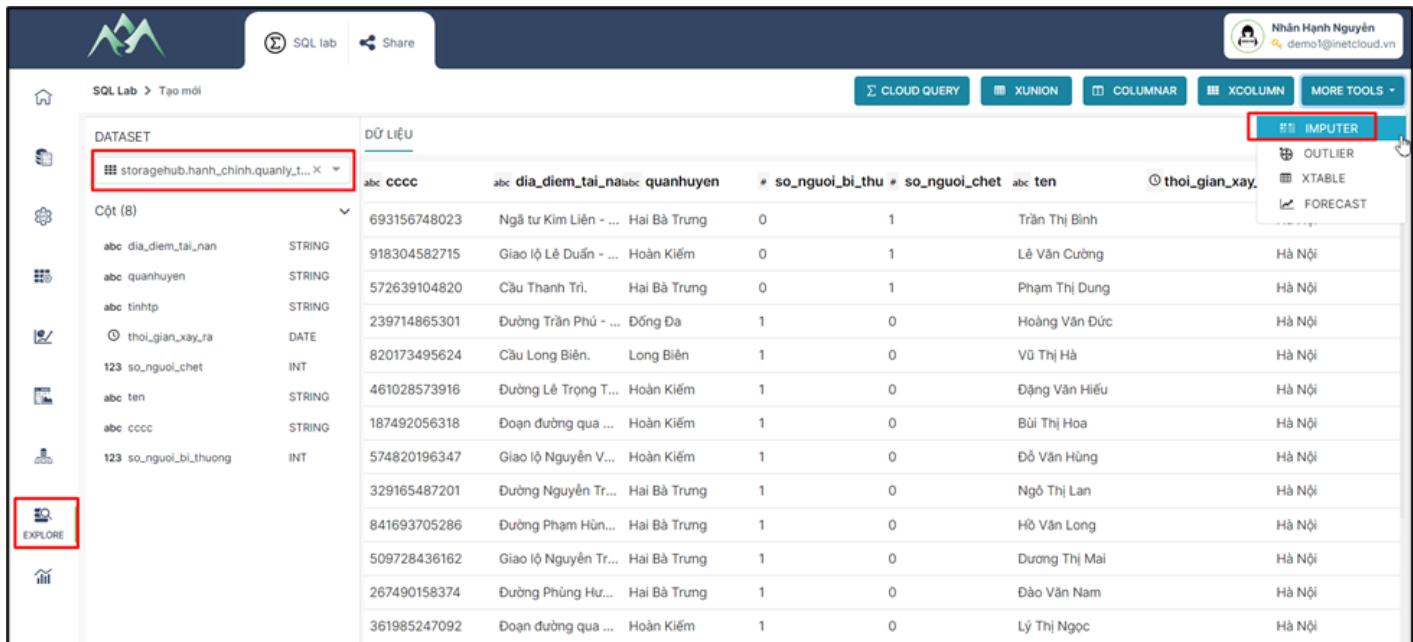
3. IMPUTER

Mục tiêu IMPUTER: được sử dụng để xử lý các giá trị thiếu trong dữ liệu, tức là các ô dữ liệu mà không có giá trị (có thể là NaN, null, hoặc các giá trị không có ý nghĩa khác). Khi dữ liệu bị thiếu, việc sử dụng Imputer giúp điền các giá trị thay thế vào những vị trí thiếu này để có thể tiếp tục phân tích dữ liệu một cách chính xác.

- [Thực hiện trên hệ thống theo các bước hướng dẫn sau:](#)

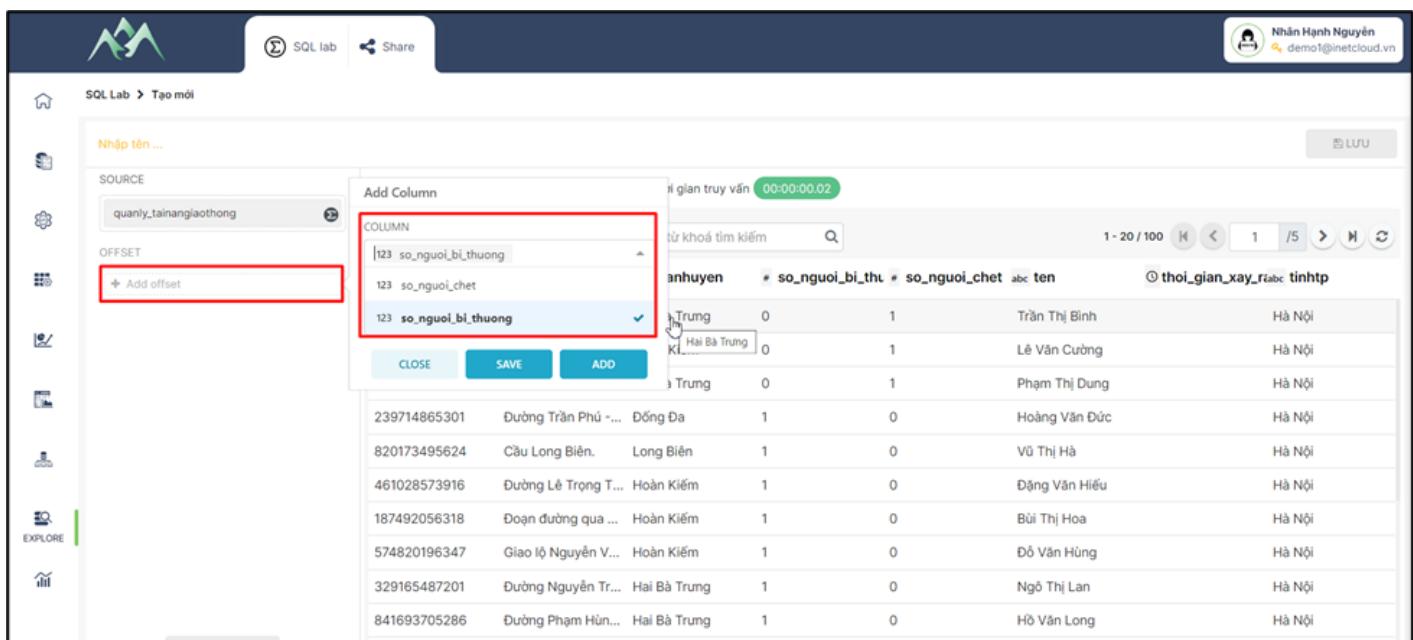
Thực hiện trên hệ thống theo các bước hướng dẫn sau:

Bước 1: Vào hệ thống chọn module Explore, chọn tập dữ liệu cần tổng hợp và tạo mới một tập dữ liệu tổng hợp từ IMPUTER



The screenshot shows the InetCloud SQL Lab interface. On the left, there's a sidebar with icons for Home, Explore (which is highlighted with a red box), and other tools. The main area has a dark header with 'Nhân Hạnh Nguyễn' and 'demo1@inetcloud.vn'. Below the header, there are tabs for 'CLOUD QUERY', 'XUNION', 'COLUMNAR', 'XCOLUMN', and 'MORE TOOLS'. Under 'MORE TOOLS', the 'IMPUTER' button is highlighted with a red box. The central part of the screen shows a 'DATASET' section with a dropdown menu containing 'storagehub.hanh_chinh.quanly_t...'. To the right is a table titled 'DỮ LIỆU' (DATA) with columns: abc_cccc, abc_dia_diem_tai_nam, abc_quanhuyen, * so_nguo_bi_thu * so_nguo_chet, abc_ten, and @thoi_gian_xay. The table contains 15 rows of data. At the bottom left of the main area, there's a red box around the 'EXPLORE' icon in the sidebar.

Bước 2: Chọn cột dữ liệu để thuật toán Imputer giúp điền các giá trị thay thế vào những vị trí thiếu.



This screenshot shows the 'Explore' tab in the SQL Lab interface. The left sidebar has the 'EXPLORE' icon highlighted with a red box. The main area has a dark header with 'Nhân Hạnh Nguyễn' and 'demo1@inetcloud.vn'. Below the header, there are tabs for 'CLOUD QUERY', 'XUNION', 'COLUMNAR', 'XCOLUMN', and 'MORE TOOLS'. Under 'MORE TOOLS', the 'IMPUTER' button is visible. The central part of the screen shows a 'SOURCE' section with 'quanly_tainanggiaothong' selected. Below it is an 'OFFSET' section with a red box around the 'Add offset' button. To the right is a 'Add Column' dialog box with a red box around the '123 so_nguo_bi_thuong' option in the 'COLUMN' dropdown. The table below shows data with some missing values, which are being handled by the Imputer algorithm. The bottom right of the main area has a red box around the 'EXPLORE' icon in the sidebar.

Bước 3: Chọn thuật toán xử lý cho cột dữ liệu đã chọn.

The screenshot shows the 'Add Column' dialog in the SQL Lab interface. The 'COLUMN' field is set to '123 so_nguo_bi_thuong'. The 'THUẬT TOÁN' dropdown menu is open, displaying four options: 'interpolate', 'simple', 'knn', and 'bayesian'. The 'simple' option is currently selected. The main table view shows data from the 'quanly_tainangiaothong' source, with columns including 'tinh', 'so_nguo_chet', 'abc_ten', and 'thoi_gian_xay_rac'. The table has 100 rows, and the current page is 1 of 5.

Trong đó:

- o Thuật toán nội suy (interpolation) được sử dụng để ước lượng giá trị tại các điểm trung gian dựa trên các giá trị đã biết của một tập dữ liệu.
- o Thuật toán Simple được dùng để xử lý các bài toán như tìm kiếm tuyến tính, sắp xếp nổi bọt, hoặc tính giai thừa.
- o Thuật toán KNN là một thuật toán dựa trên khoảng cách để phân loại hoặc dự đoán giá trị của một điểm dữ liệu mới dựa trên các điểm dữ liệu đã biết trong không gian đa chiều.
- o Thuật toán Payesian là một trong những thuật toán phân loại phổ biến nhất trong học máy, đặc biệt là đối với các bài toán phân loại văn bản và lọc thư rác.

Bước 4: “RUN” chạy test dữ liệu được tổng hợp.

The screenshot shows the Dataiku DSS interface. On the left, there's a sidebar with icons for Home, Explore, and a chart. The main area is titled "SQL Lab > Tạo mới". A "SOURCE" section contains a dropdown set to "quanly_tainangiaothong" and an "OFFSET" section with two rows: "so_nguoibithuong(interpolate)" and "so_nguoichet(interpolate)". Below these are buttons for "TEST" (highlighted with a red box) and "Lưu" (Save). The "KẾT QUẢ" (Result) section displays a table with two columns: "# so_nguoibithuong" and "# so_nguoichet". The table has 15 rows, each containing a value from 0 or 1. A red box highlights the entire result table.

Bước 5: Nhập tên cho tập dữ liệu mới được tổng hợp, chọn Lưu thông tin tập dữ liệu được tổng hợp bằng công cụ IMPUTER

This screenshot is similar to the previous one but with some changes. The "SOURCE" dropdown now shows "imputer_tainangiaothong" (highlighted with a red box). The "Lưu" (Save) button at the top right is also highlighted with a red box. The rest of the interface, including the "TEST" button and the result table, remains the same as in the first screenshot.