**Đề Tài: Import Data By Talend.**

|  |  |  |
| --- | --- | --- |
| STT | Thành Viên | MSSV |
| 1 | Nguyễn Công Thịnh | 15024631 |
| 2 | Võ Đông Lâm | 15014921 |

Mục đích lập nhóm: Nghiên cứu và hoàn thành đề tài nhóm đã chọn.

Nền tảng thực hiện: Talend

**Kế hoạch thực hiện đồ án**

|  |  |
| --- | --- |
| **Buổi thực hành** | **Nội Dung** |
| 1 | Tìm hiểu Talend là gì? Download, cài đặt về Talend. |
| 2 | Metadata và Schemas |
| 3 | Mapping data |
| 4 | Sử dụng java trong Talend |
| 5 | Làm việc với Database |
| 6 | Managing files |
| 7 | Làm việc với XML, Queues và Wed Services |
| 8 | Tổng kết những gì nhóm làm được |
| 9 |  |

**Buổi thực hành 3( Ngày 29/08/2017): Mapping data**

Các **tMap component** có khả năng chuyển đổi rộng lớn và do đó trở thành công cụ tích hợp dữ liệu cho nhiều nhà phát triển lựa chọn. Trong đó **tMap component** có khả năng:

* Thêm và xoá các cột.
* Áp dụng quy tắc chuyển đổi một hoặc nhiều cột.
* Lọc dữ liệu đầu vào và đầu ra.
* Kết hợp dữ liệu từ nhiều nguồn vào một hoặc nhiều kết quả đầu ra.
* Phân chia nguồn dữ liệu thành nhiều kết quả đầu ra.

Để thấy rỏ các chức năng chuyển đổi của tMap chúng ta thực hiện ví dụ sau:

**Ví dụ 1:** Trong ví dụ này ta sẽ tạo một Job để đọc dữ liệu từ một file và ghi kết quả vào một file mới. **Bước 1:** Tạo file Employee.csv có cấu trúc như sau:

Id;Name;Department;StartDate;Salary

1;John Smith;Sales;01-Mar-2012;20000

2;Jane Roberts;Sales;05-Jun-2010;22500

3;Paul Jones;Marketing;22-Jun-2010;30000

4;Mark Jacobs;Sales;05-Apr-2010;22000

5;Mary Stephenson;IT;13-Sep-2009;22500

6;Steve Roberts;IT;14-Dec-2011;18000

7;Margaret Johnson;CEO;30-Feb-2011;41000

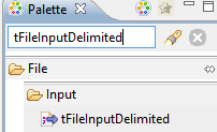
8;Paul Harrison;Sales;07-Nov-2009;19500

9;Peter Pedlar;Admin;22-Dec-2012;21000

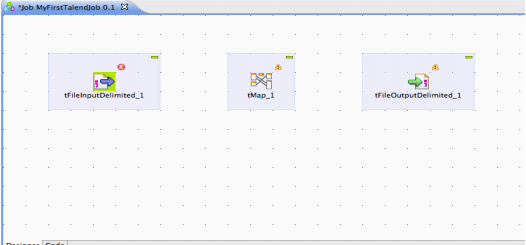
10;Douglas Peterson;Admin;28-May-2011;21500

**Bước 2:** Tiếp theo, chúng ta tạo một Talend Job mới với tên **MyFristTalendJob**. Một cửa sổ làm việc trống với Job được tạo ra.

Bước 3: Sau đó, ta thêm các **Components** vào cửa sổ thiết kế bằng cách lựa chọn chúng từ Components **Palette**. Trong cửa sổ **Palette** tìm đến 3 components **tFileInputDelimited, tMap** và **tFileOutputDelimited.**

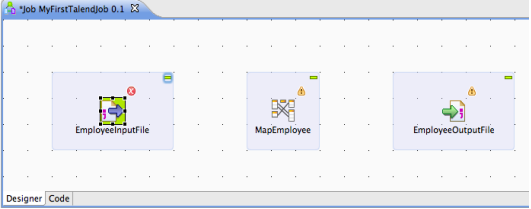


Sau khi tìm được các components, ta lần lượt kéo thả chúng vào cửa sổ thiết kế, kết quả được như hình dưới đây.

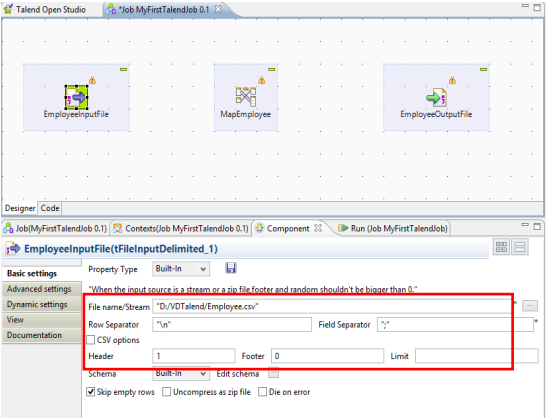


Tiếp theo, chúng ta sẽ thay đổi nhãn của các components. Thấy Talend sẽ tự động sinh ra các nhãn cho các components giống với giá trị của các components theo sau là một dấu gạch dưới và một giá trị số phía sau nhãn các components, nếu chúng xuất hiện nhiều lần trong cửa sổ thiết kế thì giá trị số này sẽ được tăng dần.

Chúng ta thay đổi nhãn của 3 components này bằng cách click vào nhãn dưới các components, sau đó click một lần nữa để cập nhật nhãn mới cho components (không phải double-click). Lần lượt thay đổi các nhãn như sau.



**Bước 4:** Cấu hình **EmployeeInputFile (tFileInputDelimited\_1)**. Lựa chọn **EmployeeInputFile** và lựa chọn tab **component** trong cửa sổ bên dưới để cấu hình lại FileInput như sau:

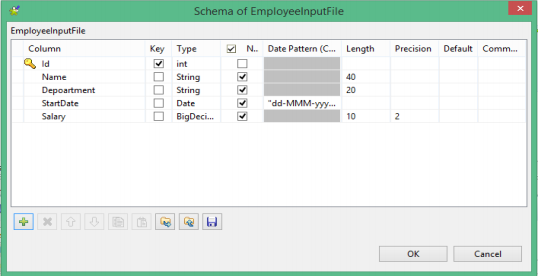


**Chú ý: File name/Stream** là đường dẫn tới file dữ liệu đầu vào cho việc xử lý dữ liệu, ở đây file Input là file Employee.csv. Thay đổi giá trị **Header** bằng 1 tức là giữ lại tiêu đề của file Input, ngược lại **Header** bằng 0 sẽ không lấy tiêu đề của file Input cho quá trình xử lý dữ liệu.

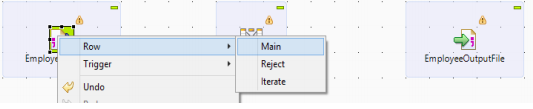
**Bước 5:** Định nghĩa **Schema** cho **EmployeeInputFile:** chọn **EmployeeInputFile** và chọn component setting trong cửa sổ bên dưới. Sau đó chọn bên phải **Edit Schema**. Talend mở ra cho hộp thoại mới.

Tại hộp thoại này ta có thể: Thêm, xóa, định nghĩa khóa chính, kiểu dữ liệu… cho các trường (các trường này là các cột dữ liệu trong file Input của chúng ta).

Ta thêm mới các trường và định nghĩa kiểu dữ liệu của các trường giống như sau:



**Bước 6:** Kết nối **EmployeeInputFile** với **MapEmployee:** Bây giờ thực hiện kết nối 2 components. Right-click lên **EmployeeInputFile** và chọn **Row -> Main** sẽ nhìn thấy một đường thẳng kéo từ component được lựa xuất hiện theo con trỏ chuột, di chuyển con trỏ đến MapEmployee và chọn nó.

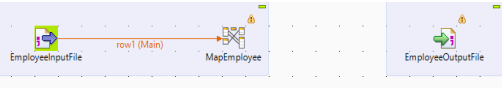


Bây giờ có thể nhìn thấy sự khác biệt trước và sau khi **Map:**

• Dấu “chấm than màu vàng” hoặc “dấu x màu đỏ” đã biến mất, điều này có nghĩa là thành phần kết nối đã có đầu ra.

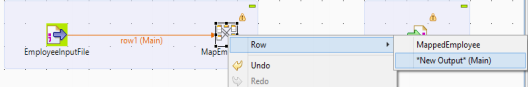
• Talend đã tạo ra một **row** kết nối với tên là **row1.**

• Kết nối component thành công thì 2 components sẽ được gom lại cùng một phần diện tích như hình bên dưới.

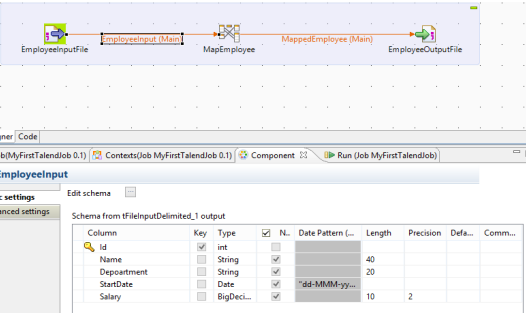


**Bước 7:** cấu hình và kết nối cho **EmployeeOutputFile**. Bây giờ chúng ta có thể cấu hình các thành phần **EmployeeOutputFile** theo một cách tương tự như **EmployeeInputFile**. Tuy nhiên Filename/Stream là nơi lưu trữ tập tin kết quả của thao tác xử lý dữ liệu, nếu file này chưa tồn tại thì Talend sẽ tạo ra file này ngược lại file này sẽ được thay thế bằng dữ liệu mới vừa được thực thi

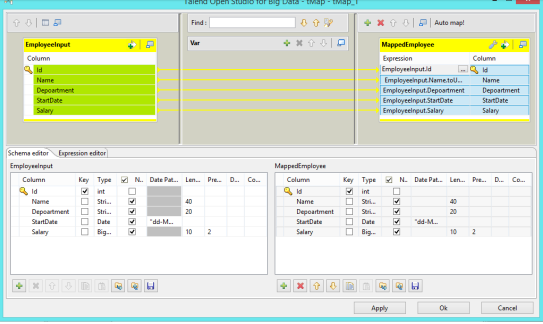
Tiếp theo, chúng ta thực hiện kết nối **MapEmployee** và **EmployeeOutputFile**. Right-click vào **MapEmployee** là chọn **Row-> \*New Output\*(Main).**



Chọn lựa chọn **New Output** Talend yêu cầu phải đặt tên cho Output vừa tạo, đặt tên là **MappedEmployee** và sửa nhãn **row1(Main)** thành **EmployeeInput**. Kết quả như hình sau:

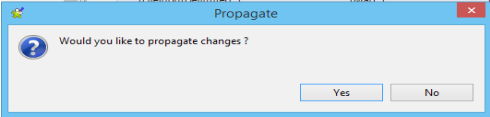


**Bước 8:** Cấu hình **MapEmployee**: để cấu hình cho MapEmployee lựa chọn vào chính nó và chọn thẻ Component properties, sau đó chọn bên phải Map Editor hoặc double-click vào Component MapEmployee trong cửa sổ thiết kế. Ngay lập tức một hộp thoại xuất hiện cho phép cấu hình MapEmployee. Theo hướng dẫn này, chúng ta sẽ tạo cấu hình cho cấu trúc của file đầu ra giống như cấu trúc của file đầu vào.

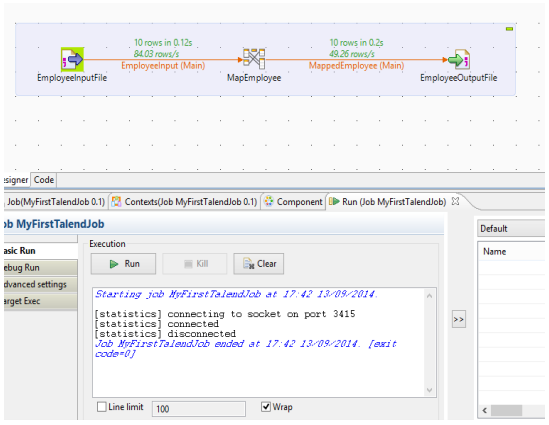


Tiến hành cấu hình MapEmployee như hình trên, đặc biệt Talend hổ trợ cho lựa chọn **Copy** và **Paste**. Sử dụng các nút lệnh để copy hoặc paste một hoặc nhiều schema từ EmployeeInput vào MappedEmployee. Khi copied các schema input vào output thành công thì Talend tự động tạo biểu đồ kết nối cho chúng ta như hình trên.

Như vậy, chúng ta đã cấu hình cho MapEmployee xong, Chọn button **Ok** để áp dụng những thay. Talend sẽ hỏi nếu muốn áp dụng nó, chọn button **Yes** để hoàn thành các thiết lập.



**Bước 9: Run Job**. Để chạy công việc chọn tab **Run** và nhấn **button Run**.



Bây giờ chúng ta có thể mở file kết quả lên và so sánh với file dữ liệu Input, kết quả là mọi thứ giống nhau hoàn toàn chỉ mất đi dòng header.

Công việc của các thành viên trong nhóm:

* Thực hành : Võ Đông Lâm + Nguyễn Công Thịnh
* Viết báo cáo và chỉnh sửa : Võ Đông Lâm

\***Tài liệu tham khảo** : 1.“*Đồ án tin học ứng dụng tìm hiểu talend*”

2.”*UG\_EN\_DI\_TOSDI*”