**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 4 ( Ngày 05/09/2017): Gom nhóm và trích lọc dữ liệu**

**Bước 1**: Tạo 2 *file* dữ liệu **Address File.txt** và **Person File.txt**.

***Address File.txt:***

*Id;Street;Town;County;Postcode*

*1;19 West Close;Shefford;West Sussex;WE24 8ST*

*2;140 Great Square;Slough;Cornwall;CO43 3RN*

*3;178 North Close;Biggleswade;Hereford and Worcester;HE91 7RE*

*4;89 Windsor Street;Warrington;Gloucestershire;GL38 5OU*

*5;153 Dee Avenue;Shefford;County Down;CO63 5UN*

***Person File.txt:***

*Id;Title;FirstName;LastName;AddressId*

*1;Mr;Austin;Patel;4*

*2;Ms;Sophia;Watson;2*

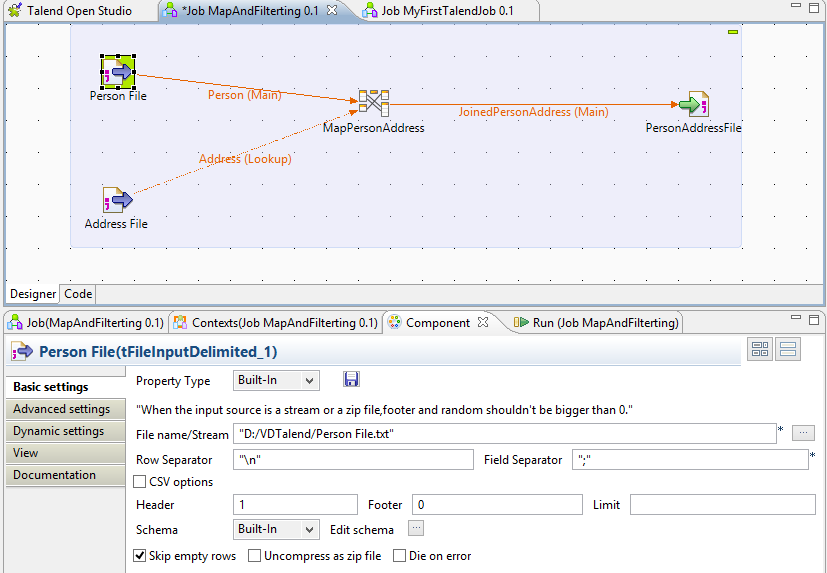
*3;Mr;Ewan;Parker;3*

*4;Ms;Evie;Cunningham;2*

*5;Mr;Dexter;Booth;*

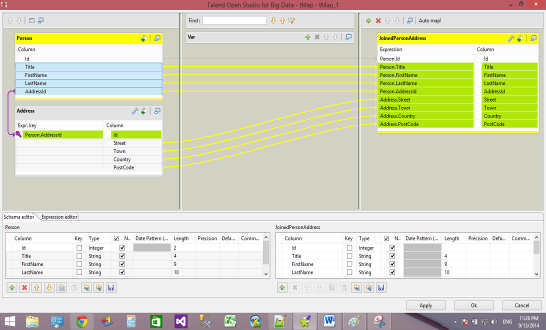
**Bước 2**: Thiết kế **Job** và cấu hình các *components*:

* **Person File** **(tFileInputDelimited\_1):** chứa đường dẫn tới *Person File.txt*.
* **Address File (tFileInputDelimited\_2):** chứa đường dẫn tới *Address File.txt*.
* **PersonAddressFile (tFileOutputDelimited\_1):** chứa đường dẫn tới *PersonAddress.csv* – file kết quả.

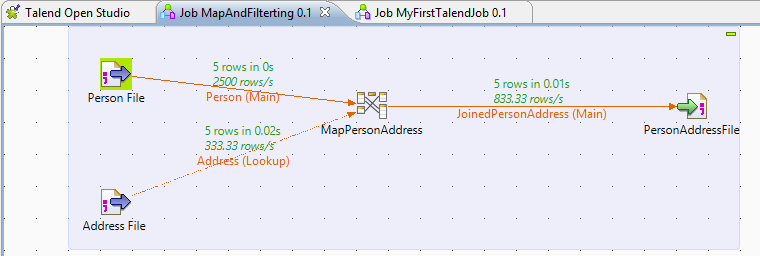


Tiếp theo, thực hiện cấu hình cho **MapPersonAddress.**

Như có thể thấy hình chụp dưới đây, 2 files đã được thêm vào *Job* và đã được **join** theo **AddressId.** có thể thực hiện điều này bằng cách kéo thả cột là khóa ngoại của bảng này vào khóa chính của bảng kia hoặc nhập bằng tay.



Sau khi cấu hình xong thì có thể **Run** xem thử kết quả. Kết quả được như hình dưới đây.



Ta có kết quả file **MapPersonAddress** như sau:

*Id;Title;FirstName;LastName;AddressId;Street;Town;County;Postcode*

*1;Mr;Austin;Patel;4;89 Windsor Street;Warrington;Gloucestershire;GL38 5OU*

*2;Ms;Sophia;Watson;2;140 Great Square;Slough;Cornwall;CO43 3RN*

*3;Mr;Ewan;Parker;3;178 North Close;Biggleswade;Hereford and Worcester;HE91 7RE*

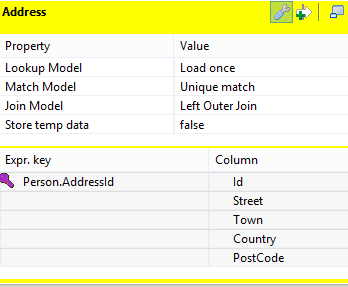
*4;Ms;Evie;Cunningham;2;140 Great Square;Slough;Cornwall;CO43 3RN*

*5;Mr;Dexter;Booth;;;;;*

Ta nhận xét kết quả như sau:

* *Addresses Id*=1, *Addresses*  *Id*=5 không xuất hiện trong *Person File*.
* *Id*=5 không có Addresses *Id* nên không thể dò tìm kết nối với *Addresses File.*
* Hai người có *Addresses Id*=2.

**Bước 3**: Chú ý đến các *button* để thiết lập dữ liệu đầu vào hoặc thiết lập các điều kiện trích lọc.



Lựa chọn *button (tMap settings)* đầu tiên. Mỗi tùy chọn đầu vào đều có các thiết lập đặc trưng, thông thường sẽ được giử thiết lập mặc định. Nếu khi bạn thay đổi các thiết lập đó thì bạn sẽ nhận thấy *button* giống như sau (đã có 2 thiết lặp đã được thay đổi).

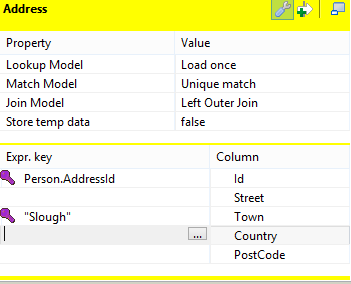
**Lookup Model**: Thông thường bạn lựa chọn tải dữ liệu dò tìm từ **Input** chỉ một lần, giá trị mặc định của **Lookup Model** là **Load Once**, nếu bạn có trường hợp cần tải lại dữ liệu từ *Input* cho mỗi *row* khóa chính thì lựa chọn **Reload at each row or Reload at each row (cache),** tùy thuộc vào kích thước của đầu vào, một trong hai tùy chọn đó sẽ ảnh hưởng nghiêm trọng đến thông lượng *Job*.

**Match Model**: Mặc định **Match Model** là **Unique match**, nếu *row* dò tìm chính phù hợp với nhiều dòng trong *file Input* thì chỉ có dòng phù hợp cuối cùng là xuất hiện trong *file Output.* Tùy chọn thứ hai là **First match, chỉ có dòng dò tìm phù hợp đầu tiên xuất hiện trong *file Output.* Tùy chọn cuối cùng là All matches tất cả các dòng phù hợp sẽ xuất hiện trong *file output*.**

**Join Model**: Mặc định **Join** **Model** là ***Left Outer Join***. Nếu không có *row* dò tìm phù hợp xuất hiện trong file *Input* thì *row* là *primary* của *Input* vẫn được giữ lại trong *file* *Output*. Nếu muốn thực hiện một **Antijoin** (dò tìm có trong A mà không có trong B) thì lựa chọn *option* và loại trừ các dòng ra, bởi *output* ra một giá trị quan trọng từ cái nhìn vào *Input* của bạn và sau đó không bao gồm hàng có giá trị này là *null*. Tùy chọn thứ hai có sẵn là **Inner Join,** trong trường hợp này thì chỉ có những *row* phù hợp sẽ được *output* ra.

**Store temp data**: Nếu đang xử lý dữ liệu lớn, có thể tìm thấy nó hữu ích khi thiết lập tùy chọn này *true*, để tiết kiệm bộ nhớ. Khi thiết lập là *true*, cần phải thiết lập **Temp data directory path**, việc này có thể được tìm thấy trên *tab* **Advanced settings** *component* của *tMap*. Lưu ý rằng điều này có thể có tác động tiêu cực đối với thời gian tổng thể *Job*.

**Bước 4**: Trở lại ví dụ ban đầu, chúng ta thiết lập điều kiện trích lọc như sau.



**Run** lại *Job* của bạn và xem kết quả. Với điều kiện trích lọc như trên thì tất cả các dòng trong *Person File* sẽ xuất hiện trong f*ile Output* tuy nhiên chỉ có **Address.Town==”*Slough*”:**

*Id;Title;FirstName;LastName;AddressId;Street;Town;County;Postcode*

*1;Mr;Austin;Patel;4;;;;*

*2;Ms;Sophia;Watson;2;140 Great Square;Slough;Cornwall;CO43 3RN*

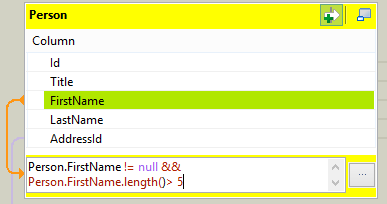
*3;Mr;Ewan;Parker;3;;;;*

*4;Ms;Evie;Cunningham;2;140 Great Square;Slough;Cornwall;CO43 3RN*

*5;Mr;Dexter;Booth;;;;;*

**Bước 5:** Mỗi *Input* đều cho phép lựa chọn **button** trong hộp thoại **tMap Expression Editor, đây là nơi thiết lập các điều kiện trích lọc phức tạp nhưng Talend lại không có biểu hiện gì khác biệt khi bạn tắt cửa sổ Expression Filter, chính vì lý do này tôi cũng tránh sử dụng các bộ lọc ẩn.**

Chúng ta cùng thiết lập các điều kiện mở rộng như hình dưới đây.



**Run** lại *Job* ta nhận được kết quả là tất cả các **FristName** có độ dài hơn 5 kí tự.

*Id;Title;FirstName;LastName;AddressId;Street;Town;County;Postcode*

*1;Mr;Austin;Patel;4;;;;*

*2;Ms;Sophia;Watson;2;140 Great Square;Slough;Cornwall;CO43 3RN*

*5;Mr;Dexter;Booth;;;;;*