**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 2( Ngày 22/08/2017):** **METADATA VÀ** **SCHEMAS.**

**các tính năng thường được sử dụng, chẳng hạn như:**

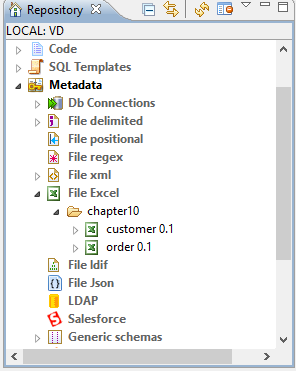
* Hand-cranking a built-in schema.
* Propagating schema changes.
* Creating a generic schema from existing metadata.
* Cutting and pasting schema information.
* Dropping schemas to empty components.

Quản lý **metadata** là một trong những khía cạnh quan trọng nhất của phát triển **Talend Jobs**, và các hình thức phổ biến nhất của *metadata* được sử dụng trong *Talend Jobs* là **schemas**.

Để phát triển thành công một **Jobs**, điều quan trọng là các *metadata* định nghĩa cho một nguồn dữ liệu mô tả chính xác định dạng cơ bản dữ liệu của nó. Nếu không xác định một cách chính xác các dữ liệu sẽ dẫn đến nhiều sai sót và lãng phí thời gian theo dõi các vấn đề định dạng dữ liệu mà không thể tránh được.

Talend cung cấp một loạt các trình thuật sĩ để giúp *metadata* có thể lấy dữ liệu từ nhiều nguồn dữ liệu khác nhau như *database tables, delimited files, and Excel worksheets ….*

**Schemas** được định nghĩa trong metadata, có thể được xây dựng trong các *components* hoặc trong *metadata repository*, như thể hiện trong hình dưới đây:



đó là cách tốt nhất để xác định nguồn và mục tiêu của *matadata* bằng cách sử dụng một *repository schema*.

Trường hợp ngoại lệ cho quy tắc này chính là khi làm việc với một lần nguồn dữ liệu được tạo ra chẳng hạn như một truy vấn cơ sở dữ liệu (*database query*). Mặc dù là một nguồn dữ liệu, nó dễ dàng hơn để lưu trữ các *schemas* cho các truy vấn hơn là làm lộn xộn các kho lưu trữ (*repository*) như các *single schemas*.

**Những lợi ích của việc sử dụng Repository schemas:**

* Talend sẽ đảm bảo cho các thay đổi được thực hiện trong *Repository schema* là mô hình thác nước cho tất cả các *Jobs* sử dụng Schema, như vậy tránh được sự cần thiết phải quét các Jobs thủ công để tìm các *schema* đã được xây dựng cần được thay đổi.
* có thể được tái sử dụng trên nhiều **Jobs**, do đó làm giảm số lượng các *key***.**
* Báo cáo phân tích ảnh hưởng có thể được tạo ra cho thấy nơi *Repository schema* đang được sử dụng trong dự án. Điều này cho thấy tác động của thay đổi nhiều đánh giá chính xác hơn khi lập kế hoạch thay đổi bất kỳ nguồn dữ liệu cơ bản nào.

**1.1. Hand-cranking a built-in schema.**

Trong phương pháp này, có một tập tin CSV mà không có một dòng tiêu đề và yêu cầu để tạo ra một *schema* cho dữ liệu. Đây là một công thức cơ bản mà hầu hết độc giả đều quen thuộc: tuy nhiên, nó cung cấp một khuôn khổ cho các cuộc thảo luận về một số nguyên tắc quan trọng của *Talend schema*.

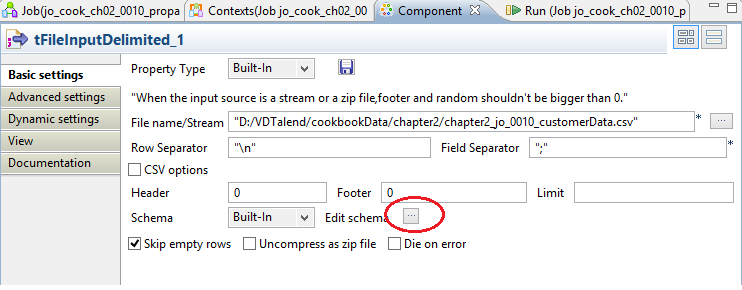
Ví dụ định nghĩa mẫu tin dưới đây:

*John Smith,27/11/1990,2012-01-10 10:24:54.953*

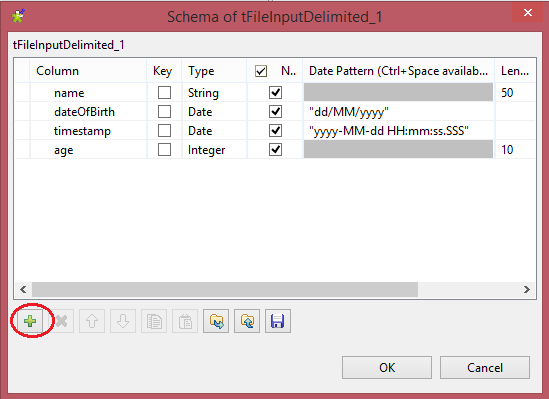
có thể thấy định nghĩa mẫu tin bao gồm: *name, Date of birth, timestamp* và *age*. Trong đó *age* có gái trị *null*.

Để tạo được nó như thế nào? Để tạo được mẫu tin đó làm như sau:

* Tạo một Job*jo\_cook\_ch02\_0000\_handCrankedSchema* và kéo *tFileInputDelimited* *component* từ *palette* và mở nó bằng cách *double-click* vào nó.
* *Click Edit Schema* **button (…)** như hình bên dưới:



* *Click* **+ button** để thêm các trường và định nghĩa các kiểu dữ liệu theo hình sau:



* *Click* **Ok** để trở lại cửa sổ *component*.

**1.2. Propagating schema changes.**

Thông thường trong quá trình phát triển, nó là cần thiết để thay đổi *schema* bằng cách thêm, loại bỏ, hoặc cột lại đặt hàng. Điều này thường là một nhiệm vụ rất nặng nề, đặc biệt là nếu một *schema* được sử dụng trong nhiều công việc.

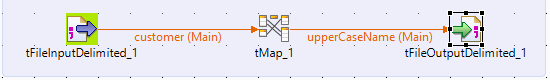
Như đã đề cập trong chương này, lưu trữ *schema* trong *metadata* cho phép các *schema* được sử dụng lại. Nếu một *schema* chia sẽ được thay đổi, sau đó Talend sẽ nhắc nhở để tìm hiểu xem những thay đổi sẽ được áp dụng cho tất cả các *Jobs*.

Nếu sự thay đổi này được thực hiện, sau lần đó sau khi công việc được mở ra, các *components* sử dụng *schema* thông thường sẽ được đánh dấu như là lỗi, vì *schema* không còn phù hợp**.**

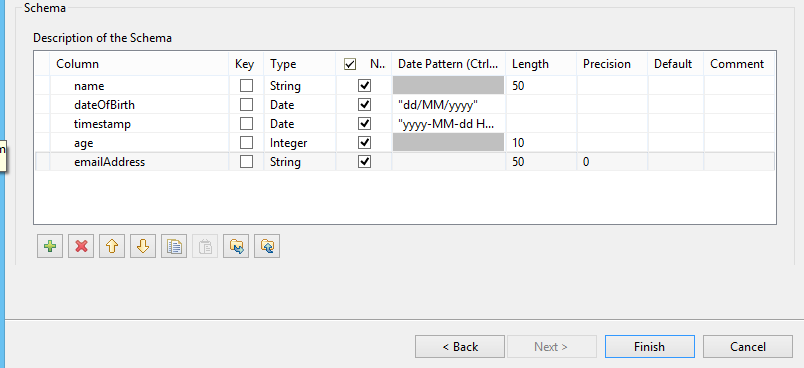
Talend cung cấp cơ chế đối thoại bên trong *schema* có một số xử lý từ việc đảm bảo rằng những thay đổi được đồng hóa vào tất cả các *Jobs*.

Để thấy được sự thay đổi ta thực hiện ví dụ sau:

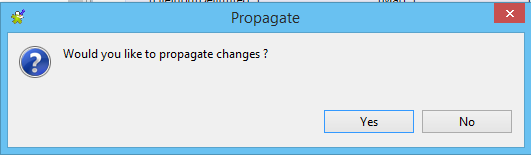
* Mở Talend Jobs *jo\_cook\_ch02\_0010\_propagateSchema* có cấu trúc như hình dưới.



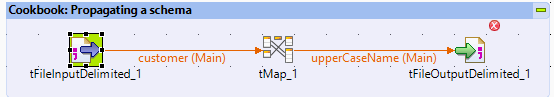
* Sau đó từ *metadata palette*, mở *Generic schema* *sc\_cook\_0010\_genericCustomer*. Thêm một trường mới *emailAddress* như hình dưới đây:



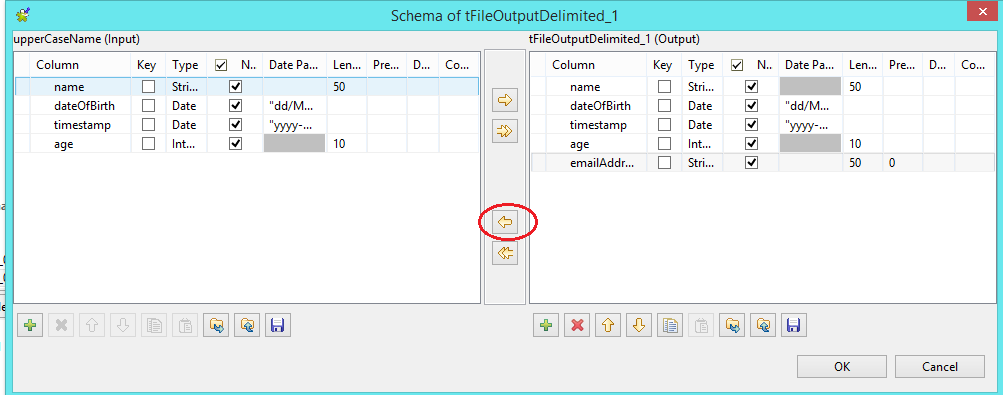
* *Click* **Finish** để lưu *schema*, sau đó *click* **Yes** để áp dụng những thay đổi.



* *Click* **Ok** để hoàn tất, sẽ nhìn thấy lỗi như hình dưới.



* Mở *tFileOutputDelimited*  và *click* chọn **Edit schema** **(…)** *button*.
* Như thấy bảng bên trái khác với bảng bên phải. chọn *emailAddress* và *click*  *button* để đưa về bảng bên trái.



* *Click* **Ok** để lưu những thay đổi.

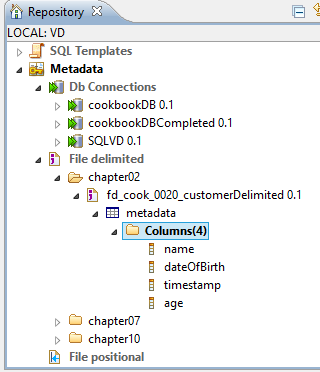
Khi Talend cập nhật *Job schema* cho một thành phần *Output*, nó không tuyên truyền thay đổi các thành phần *Input*. Sử dụng tùy chọn **<<** cho phép các nhà phát triển để sao chép tất cả các thay đổi từ *schema* kết quả trở lại vào thành phần trước, sẵn sàng cho một quy tắc được áp dụng.

**1.3. Creating a** **generic schema from the existing metadata.**

Bất kỳ *schema* nào cũng có thể dễ dàng chuyển đổi thành một **generic schema** để cho phép nó được tái sử dụng. Các công thức sau đây cho thấy hai phương pháp của việc tạo ra **generic schema**; thứ nhất là từ một *schema* tồn tại trước đó *metadata repository* và thứ hai từ một *schema* được xây dựng.

**Từ repository schema**:

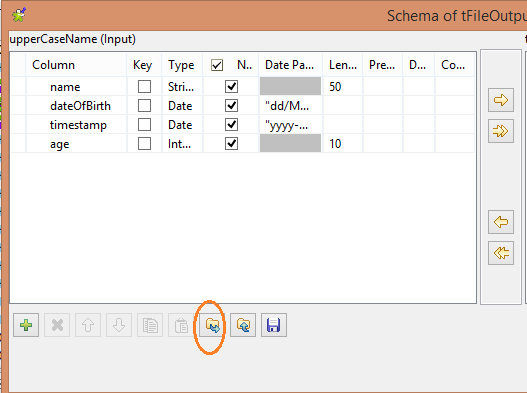
* Mở *repository schema*  *fd\_cook\_0020\_customerDelimited* như hình sau:



* *Right-click* vào *metadata* và lựa chọn **copy to Generic schema**.
* Di chuyển đến *schema* trong *Folder chapter02*, *double-click* và sửa tên thành *sc\_cook\_0020\_genericCustomer1*.

**Từ một schema được xây dựng.**

* Mở *Talend Job jo\_cook\_ch02\_0020\_builtInSchema* và mở *tFileOutputDelimitted*.
* *Click* chọn *button* trong hình dưới đây:



* Một hộp thoại **Save As** xuất hiện, lưu với tên sau *sc\_cook\_0020\_genericCustomer.xml.*
* Tiếp theo, tạo mới một *generic* *schema* bằng cách *right-click* vào **generic schema** và lựa chọn **Create** **generic schema from xml.**
* Sau đó lựa chọn *file xml* đã lưu, đặt tên cho **generic schema** mới là *sc\_cook\_0020\_genericCustomer2*.

Talend lưu trữ *schema* trong các tập tin định dạng XML, không phụ thuộc vào loại *schema*. Điều này có nghĩa *schema* có thể dễ dàng chuyển đổi giữa các loại, giữa xây *schema* tự xây dựng hoặc *schema repository*.

* 1. **. Cutting and pasting schema information.**

Kỹ thuật này tiết kiệm thời gian thực hiện, nhưng không phải lúc nào cũng rõ ràng và ngay lập tức, bởi vì *schema* không có một lựa chọn *right-click*.

Ví dụ: *copy Job jo\_cook\_ch02\_0020\_builtInSchema* và đổi tên thành*jo\_cook\_*

*ch02\_0030\_copySchema.*

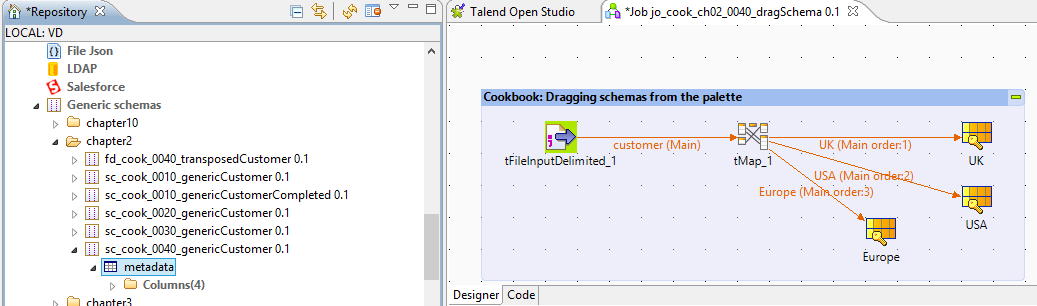
* Kéo thả *tFileOutputExcel* component từ cửa sổ *palette* bên phải vào.
* Mở *tFileOutputDelimited* và mở *Edit schema*.
* Click chọn và nhấn **Ctrl + A** để lựa chọn tất cả các cột.
* Nhấn **Ctrl+ C** để *copy* tất các các cột được chọn.
* Mở *tFileOutputExcel* *component* và chọn *Edit schema* (*tFileOutputExcelcomponent* đang rỗng).
* Nhấn *Ctrl + V* để dán các cột.

Talend cho phép sử dụng các phím tắt để copy hoặc dán thông tin của các cột trong *schema*.

* 1. **. Dropping schemas to empty components.**

Viêc này đơn giản là tiết kiệm thời gian hữu ích, đặc biệt là khi sử dụng **generic schema** và các **Hash** *component***.**

* Trong cửa sổ **Repository,** mở*generic schema sc\_cook\_0040\_genericCustomer* sẽ nhìn thấy *metadata.*



* Để **droping** schema ta chỉ kéo *metadata* bỏ vào *tHashOutput*. Khi kéo một *schema* *metadata* vào một *component*, *component* tự động phân chia với các *schema* mới.

Công việc của các thành viên trong nhóm:

* Đọc và phân tích tài liệu : Võ Đông Lâm
* Viết báo cáo và chỉnh sửa : Nguyễn Công Thịnh

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

2.”*UG\_EN\_DI\_TOSDI*”

**Tiến độ buổi thực hành 2 (22/08/2017) : chưa đảm bảo đủ nội dung thực hành của buổi thực hành 2. Tiếp tục tìm hiểu và thực hiện vào buổi thực hành 3 (29/08/2017)**