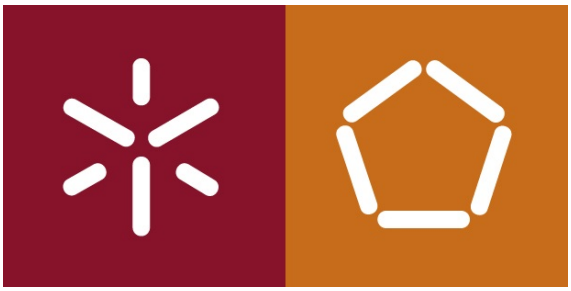

Universidade do Minho

Mestrado Integrado em Engenharia Informática

Aplicações Informáticas na Biomedicina

Definição, criação e execução de jobs no Talend

Adriana Meireles (A82582)
Bárbara Cardoso (A80453)
Carla Cruz (A80564)
Inês Alves (A81368)
Shahzod Yusupov (A82617)

19 de Novembro de 2019

Resumo

No âmbito da Unidade Curricular de Aplicações Informáticas na Biomedicina, complementar do Mestrado em Engenharia Informática, foi-nos proposta a realização de três *jobs* utilizando a ferramenta *Talend* para definição, criação e execução dos mesmos.

Conteúdo

1 Introdução

Nesta Unidade Curricular foi-nos proposto a criação de três jobs, com o auxílio da ferramenta *Talend*. É utilizado o dataset fornecido pelos docentes, *mental_health.csv*, para a criação desses jobs.

Esse dataset contêm dados reais de um estudo realizado em 2014 que foi responsável pela avaliação das atitudes dos trabalhadores em empresas de Tecnologia de Informação relativamente à sua saúde mental, assim como à frequência de transtornos mentais em ambientes de trabalho.

Após uma breve análise do dataset, observámos que possuía vários campos cada um com valores distintos:

- Timestamp
- Age
- Country
- State (Qual o estado/território dos Estados Unidos, por exemplo)
- Self_Employed (Trabalhador por conta própria)
- Family_History (Se possui alguém na família com doenças mentais)
- Treatment (Se procurou tratamento para a doença mental)
- Work_Interfere (Se a doença mental interfere ou não no trabalho)
- No_Employees (Número de Trabalhadores)
- Remote_Work (Se o trabalho é maioritariamente dentro ou fora de um escritório)
- Tech_Company (Se o empregador é ou não principalmente uma empresa de tecnologia)
- Benefits (Se o empregador oferece ou não benefícios para quem tem problemas mentais)
- Care_Options (Se o empregador fornece cuidados para saúde mental)
- Wellness_Program (Se o empregador discutiu o programa de saúde mental de um trabalhador)
- Seek_Help (Se o empregador fornece informações sobre saúde mental e como procurar ajuda)
- Anonymity (Se o anonimato está contemplado caso o trabalhador diga que tem uma doença)
- Leave (Se é fácil tirar uma baixa caso possua doença mental)
- Mental_Health_Consequence (Se o trabalhador acha que se falar com o empregador sobre saúde mental pode ter consequências negativas)
- Phys_Health_Consequence (Se o trabalhador acha que se falar com o empregador sobre saúde física pode ter consequências negativas)
- Coworkers (Se estariam dispostos a falar com colegas de trabalho sobre doenças mentais)
- Supervisor (Se estariam dispostos a falar com o seu supervisor sobre doenças mentais)
- Mental_Health_Interview (Se o trabalhador abordaria um problema mental na altura da entrevista)
- Phys_Health_Interview (Se o trabalhador abordaria um problema físico na altura da entrevista)

- Mental_vs_Physical (Se o trabalhador acha que o empregador leva doenças mentais menos a sério do que doenças físicas)
- Obs_Consequence (Se observou consequências negativas para quem trabalha e tem doenças mentais)
- Comments (Comentários adicionais)

2 Jobs

Com o intuito de explicar da melhor maneira o raciocínio adotado, iremos, de seguida, apresentar os *jobs* criados pelo grupo e quais os seus objetivos.

Atentemos que através dos *jobs* podemos usar uma determinada transformação várias vezes, não sendo necessário repetir todo o processo de construção realizado anteriormente.

Por fim, reparemos também que nos *jobs* seguidamente apresentados encontram-se algumas colunas de tabelas que não são usadas propriamente no dito *job*, no entanto, o grupo optou por as inserir no seu estudo com o intuito de, no final, haver a possibilidade de se destacarem curiosidades interessantes no âmbito deste estudo, como é o caso do *job2* inserido neste relatório.

2.1 Job 1

Neste primeiro *job* decidimos averiguar se a saúde mental interefe com o trabalho de uma pessoa com idade superior a 35 anos. Para isso, começamos por selecionar as seguintes colunas do *dataset* fornecido:

- Age
- Gender
- work_interfere
- mental_vs_physical
- family_history

Primeiramente optamos por filtar por idade, uma vez que o nosso estudo incide apenas em pessoas com mais de 35 anos. Assim, iremos trabalhar agora apenas com pessoas com idade superior à referida. Optamos também, a título de curiosidade, por nos focar em pessoas que se encontrem já em tratamento, uma vez que estas representam, à partida, uma maior gravidade do estado da doença. Após esta seleção, decidimos guardar os dados das pessoas que tivessem menos de 35 anos de idade numa tabela na base de dados a que demos o nome de *job1*.

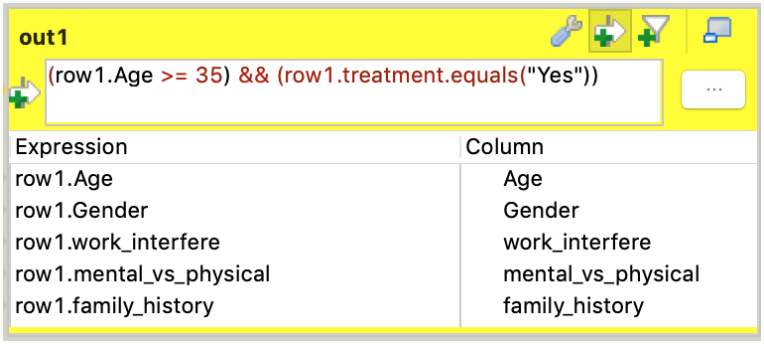

Figura 1: tMap: Job1

Por fim, de modo a facilitar a observação dos resultados, optamos por realizar uma filtragem por género e, posteriormente, uma filtragem por grau de interferência no seu desempenho no trabalho. Assim, encontramo-nos perante um conjunto de dados que irão agora ser repartidos por quatro ficheiros:

1. ficheiro correspondente aos trabalhadores do género feminino que responderam que nunca ou raramente sentem que a sua saúde mental interfere na sua perfomance profissional (out_female_Healthy);

2. ficheiro correspondente aos trabalhadores do género feminino que responderam que sentem que a sua saúde mental interfere na sua perfomance profissional muitas vezes ou, pelo menos, às vezes (out_female_Unhealthy);
3. ficheiro correspondente aos trabalhadores do género masculino que responderam que nunca ou raramente sentem que a sua saúde mental interfere na sua perfomance profissional (out_male_Healthy);
4. ficheiro correspondente aos trabalhadores do género masculino que responderam que sentem que a sua saúde mental interfere na sua perfomance profissional muitas vezes ou, pelo menos, às vezes (out_male_Unhealthy).

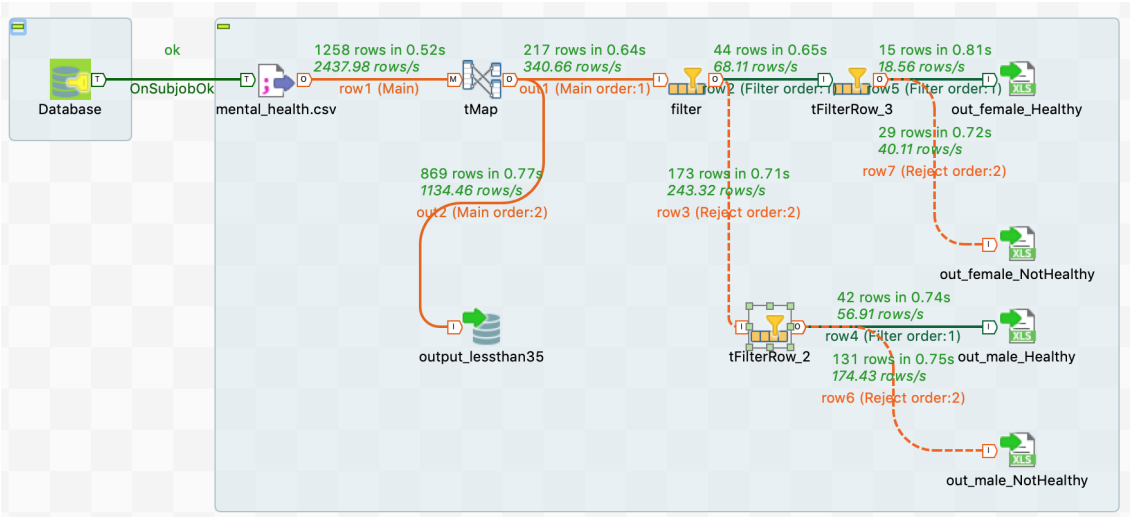

Figura 2: Job1

2.1.1 Análise de Resultados

Após uma implementação atenta e cuidada do *job* referido passamos agora para a sua análise.

É importante notar que o objetivo de todos os *jobs* implementados neste relatório era dar resposta a um problema que consideramos preocupante e que fosse útil na área da saúde, mais especificamente, neste caso, na área da saúde mental.

Posto isto, podemos agora responder à pergunta "**a saúde mental interefe ou não no trabalho de uma pessoa com mais de 35 anos?**"

A partir da imagem acima podemos verificar que apenas 217 trabalhadores têm mais de 35 anos de idade e, desse conjunto, apenas 44 são do género feminino, restanto, assim, 173 do género masculino. Passando agora à análise da questão central do problema, verificamos o seguinte:

- No género feminino, das 44 trabalhadoras, apenas 15 responderam que nunca ou raramente sentiram que a sua saúde mental interferia na sua performance no trabalho;
- No género masculino, dos 173 trabalhadores, apenas 42 responderam que nunca ou raramente sentiram que a sua saúde mental interferia na sua performance no trabalho;

Para efeitos de estudo, o grupo decidiu que só seriam considerados dados absolutamente negativos no caso das respostas serem *"Never"* ou *"Rarely"* a *"work_interfere"*, uma vez que os casos mais preocupantes serão quando a interferência no trabalho se revela contínua. Sendo assim, podemos afirmar que a saúde mental interfere, de facto, no trabalho de uma pessoa com mais de 35 anos.

2.2 Job 2

Já no segundo *job* o objetivo passou por avaliar se o histórico familiar teria alguma influencia na saúde mental dos trabalhadores.

Para isso, optamos por selecionar algumas colunas do *dataset* dado e que consideramos relevantes para o caso em questão:

- Age
- Gender
- self_employed
- family_history
- mental_health_consequence
- mental_vs_physical
- obs_consequence
- mental_health_consequence

De notar que a coluna *self_employed* foi selecionada por razões externas ao objetivo principal, isto é, com este mesmo *job* pretendemos ainda verificar se existe algum tipo de relação com o facto do trabalhor realizar o seu trabaho por conta própria ou não.

Assim, após a seleção cuidada das colunas que possuiam informação relevante para o caso de estudo, passamos para a filtragem dos dados. Esta filtragem consistia em selecionar apenas os trabalhadores que possuissem histórico familiar afirmativo.

Para finalizar, os dados foram ordenados de modo ascedente da coluna *Age* e *self_employed* a fim de facilitar a sua visualização final, sendo estes exportados para um ficheiro denominado *job2_out*.

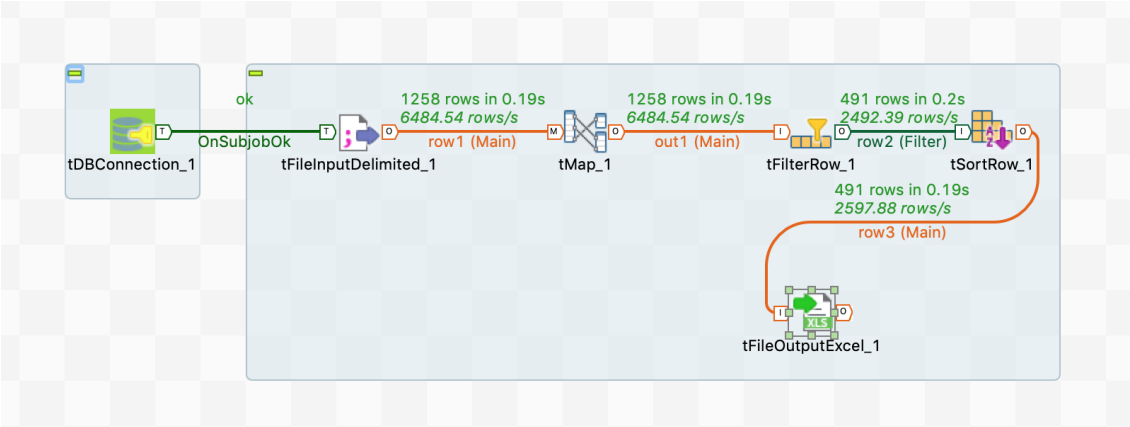

Figura 3: Job2

2.2.1 Análise de Resultados

Com o intuito de dar resposta à pergunta "**o histórico familiar de um trabalhador tem influencia na sua saúde mental?**"foi criado o *job* mencionado acima.

A resposta a esta pergunta é agora fácil de dar por observação deste mesmo *job*. Como podemos verificar, de 1258 trabalhadores representados no dataset, apenas 491 têm histórico

familiar positivo. É ainda curioso observar que nenhum destes trabalhadores trabalha por conta própria, isto é, trabalham numa empresa ou não têm emprego.

Apercebendo-nos deste facto, decidimos, então, explorar um pouco mais esta questão e fazer um outro *job* que nos ajudasse no seu entendimento.

Basicamente, este *job* seria bsatante parecido com o anterior, uma vez que advém do mesmo, mas agora pretende explorar outras hipóteses de análise dos dados.

Para isso, foram selecionadas as colunas:

- Age
- Gender
- self_employed
- family_history
- work_interfere
- Country
- no_employees
- remote_work
- tech_company

Seguidamente, decidimos focar a nossa curiosidade em trabalhadores que tivessem histórico familiar e aplicamos, sobre este conjunto, dois filtros:

1. para filtrar apenas trabalhadores que sentissem que a sua saúde mental influenciava no seu desempenho no trabalho (muitas vezes ou às vezes);
2. para selecionar os trabalhadores que trabalhassem numa empresa de IT ou que pudessem trabalhar remotamente.

Como podemos observar, 491 pessoas das 1258 têm histórico familiar afirmativo. Destas 491 pessoas, 328 afirmam que a sua saúde mental interefe no seu desempenho no trabalho e, por fim, 280 delas trabalham numa empresa de IT ou trabalham fora do escritório pelo menos 50% do tempo. No entanto, ainda numa vertente simplesmente curiosa sobre o caso, podemos verificar que apenas 12 trabalhadores dos 280 não trabalham numa empresa de IT e todos estes 12 trabalham remotamente.

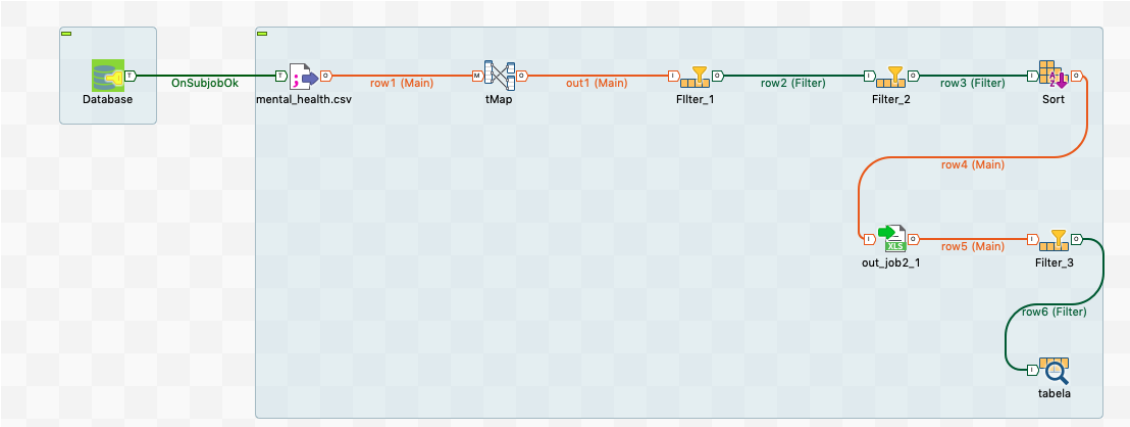

Figura 4: Job2_1

Podemos assim concluir que os trabalhadores ligados ao ramo da tecnologia têm uma grande tendência para desenvolver problemas relacionados com a saúde mental.

tabela

Age	Gender	self_employed	family_history	work_interfere	Country	no_employees	remote_work	tech_company
34	female	No	Yes	Often	United Kingdom	26-100	Yes	No
34	Male	No	Yes	Often	United Kingdom	26-100	Yes	No
23	Male	No	Yes	Sometimes	United States	26-100	Yes	No
28	female	No	Yes	Often	United States	100-500	Yes	No
34	Male	No	Yes	Sometimes	United States	26-100	Yes	No
30	cis-female/femme	Yes	Yes	Often	United States	1-5	Yes	No
36	Male	No	Yes	Sometimes	United States	More than 1000	Yes	No
29	Female	No	Yes	Sometimes	United States	26-100	Yes	No
29	Male	No	Yes	Sometimes	United States	More than 1000	Yes	No
42	male	Yes	Yes	Sometimes	United States	6-25	Yes	No
38	m	No	Yes	Sometimes	United States	100-500	Yes	No
40	Male	No	Yes	Sometimes	United States	More than 1000	Yes	No

Figura 5: Tabela de demonstração

2.3 Job 3

Por fim, no terceiro e último *job* criado pelo grupo procurámos perceber em que medida os trabalhadores sentem ou não dificuldade em comunicarem com os seus colegas de trabalho ou supervisores.

Começamos, como seria de prever, por selecionar algumas colunas que consideramos importantes para este estudo:

- Age
- Gender
- work_interfere
- wellness_program
- seek_help
- mental_health_consequence
- coworkers
- supervisor
- mental_health_interview
- mental_vs_physical
- obs_consequence
- treatment

Com isto, selecionamos os trabalhadores que nos deram respostas afirmativas às colunas *"supervisor"* ou *"coworkers"*. Ainda numa vertente de melhor perceção de resultados, filtramos também os trabalhadores que deram respostas positivas a ambas as colunas e ordenamo-los por ordem ascendente de idade e género, a fim de melhorar a organização do conjunto de dados. Estes resultados foram, posteriormente, guardados num ficheiro externo *out_job3*.

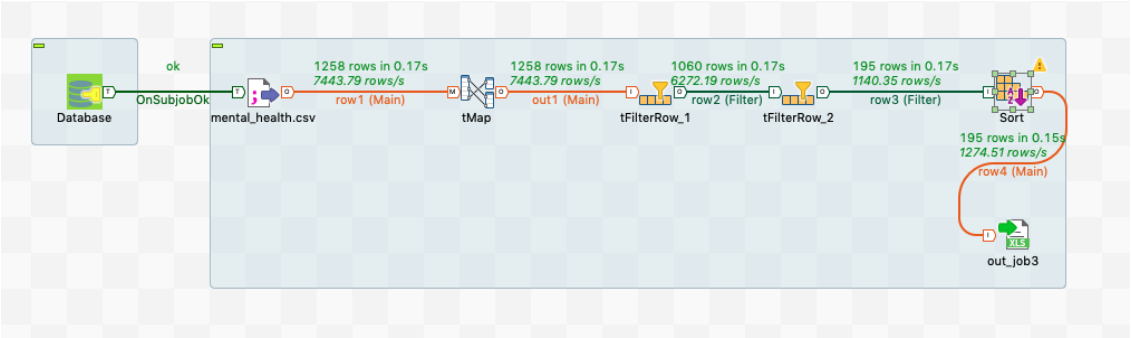

Figura 6: Job3

2.3.1 Análise de Resultados

Passando agora para a análise de resultados do *job* desenvolvido, procuramos dar resposta à questão **"os trabalhadores sentem dificuldade em falar com os seus supervisores ou colegas de trabalho acerca de saúde mental?"**. Esta pergunta torna-se extremamente importante na desmistificação da temática "a saúde mental é ou não um assunto *tabu*?", uma vez que, no quotidiano em que nos inserimos, cada vez mais pessoas sofrem de doenças do foro psíquico e sentem-se constrangidas perante a sociedade.

Sendo assim, consideramos esta uma questão verdadeiramente útil para a área da investigação relativa à saúde mental.

Observando o *job* criado vemos que dos 1258 trabalhadores, 1060 sentir-se-iam à vontade para falar sobre problemas de saúde mental com o seu supervisor ou com colegas de trabalho ou com, pelo menos, alguns deles. No entanto, destes 1060 apenas 195 se sentiriam à vontade de falar com ambos.

No nosso entender, estes dados podem ter diversas interpretações, no entanto, consideramos os resultados bastante satisfatórios, relevando um enorme passo no que toca à aceitação das doenças mentais por parte da população.

3 Notas Extra

A fim de executar os *jobs* mencionados neste relatório é necessário relembrar que tanto no componente *tDBConnection* como no componente *tFileOutputExcel* há ajustes obrigatórios.

Repare-se que, nos exemplos descritos, a base de dados selecionada foi criada por um elemento do grupo, podendo o nome da mesma e a sua localização na máquina em uso serem diferentes.

Vejamos, a título de exemplo, o seguinte caso:

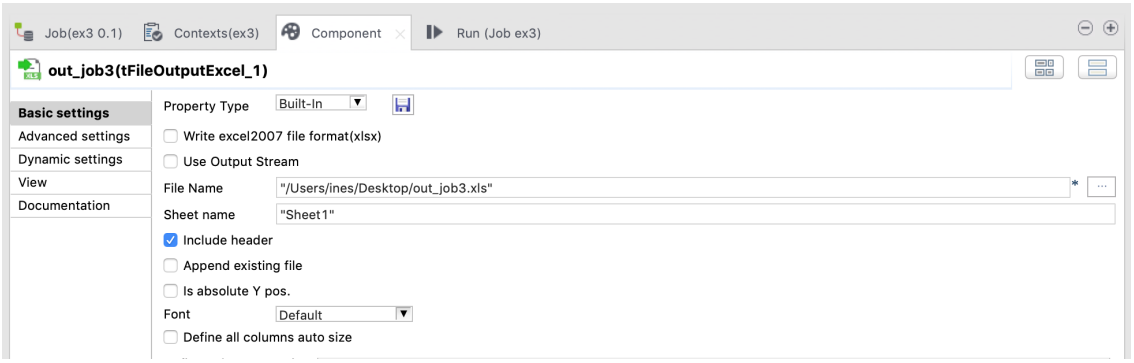

Figura 7: Exemplo de campo a ser alterado

Neste caso, no *job3* inserido neste relatório e já referido na secção anterior, o ficheiro resultante está a ser guardado em */Users/ines/Desktop/*, pelo que, para testar este *job*, o *path* deve ser alterado para um local existente na máquina que está a correr o *job*. Da mesma forma, no componente *tDBConnection*, deve ser selecionada a base de dados respetiva.

4 Conclusão

A realização deste trabalho permitiu-nos aprofundar o nosso conhecimento no uso da plataforma *Talend Open Studio*, plataforma essa que é amplamente utilizada como ferramenta ETL e que, para além de ser *Open Source*, permite combinar, converter e atualizar os dados localizados em diferentes fontes de informação.

O grupo sentiu algumas dificuldades no início, pois nenhum dos membros tinha experiência suficiente na utilização da ferramenta. Porém, após várias pesquisas e discussões entre os elementos, conseguimos realizar com sucesso as tarefas propostas no enunciado, contornando assim os obstáculos encontrados inicialmente.

Em suma, os objetivos propostos foram alcançados, estando o grupo motivado para a realização do trabalho prático, em que será necessária a utilização desta mesma ferramenta.