MANUAL DE INSTALAÇÃO DO SIPAC – SISTEMA DE INFORMAÇÃO DA PRODUÇÃO AGROPECUÁRIA CAPIXABA

Elaboração: Cristian da Silva Antério

Bolsista PIBIT/FAPES/INCAPER

2016

Table of Contents

Introdução

O Pentaho é uma suíte com várias ferramentas disponíveis para Inteligência de Negócios, incluindo:

- **Pentaho Data Integration**: Também conhecido como Kettle, é uma ferramenta de código livre para extração, transformação e carga de dados.
- **Pentaho Analysis Service**: Conhecido também como Mondrian OLAP server, é uma ferramenta de código livre para Análise Online e Processamento (OLAP).
- **Pentaho Data Mining**: Derivado do projeto Weka, um conjunto de ferramentas relacionadas com a mineração de dados.
- **Pentaho Dashboards**: Utilizado para mostrar, de uma maneira fácil, os dados trabalhados.

Neste tutorial trabalharemos com o Pentaho Data Integration(PDI), para carregar os arquivos dos arquivos csv, manipulandos-os e os inserindo em uma base de dados. O PDI tem uma interface muito amigável e é muito simples de se manipular.

> [1]. ETL é um processo para extrair dados de um sistema de arquivos, transformá-lo em um formato de banco de dados e colocá-lo no banco de dados. ↩

Instalação

A Utilização das ferramentas da suíte Pentaho é bastante simples. Para este tutorial, precisaremos apenas de duas delas: Pentaho Data Intagration (PDI) e Pentaho Business Intelligence Server (BI-Server). Os arquivos estão disponíveis em: http://sourceforge.net/projects/pentaho/files/

Após efetuar o download, precisamos fazer algumas configurações, antes de começar a utilizar a ferramenta:

- Instalação Java (Linux):
  - Precisamos instalar o Oracle Java, pois com o JDK, padrão do linux não é compatível com o Pentaho. Sendo assim, abra o terminal e adicione o repositório Java com os seguintes comandos:
  - `sudo add-apt-repository ppa:webupd8team/java`
  - `sudo apt-get update`
  - `sudo apt-get install oracle-java7-installer`
  - Feito isso, vamos configurar as variáveis de ambiente. Vá até o caminho **/usr/lib/jvm** e veja se o arquivo **java-7-oracle** se encontra lá. Caso esteja, volte até a raiz e edite o arquivo **/etc/environment** com o seguinte comando:
  - `sudo nano /etc/environment`
  - E adicione as seguinte linhas:
  - JAVA_HOME=/usr/lib/jvm/java-7-oracle
  - JRE_HOME=/usr/lib/jvm/java-7-oracle
  - Salve o arquivo, reinicie o sistema e pronto!
- Para os usuários do Windows, sugiro este tutorial de como instalar o Java:
  - https://www.youtube.com/watch?v=pZI7hdepiy0
- Vá na pasta onde você baixou e extraiu o BI-Server, abra-a no terminal e execute o arquivo **start-pentaho.sh** da seguinte forma:
  - `./start-pentaho.sh`
  - Depois disso, abra o navegador no seguinte endereço: http://localhost:8080 e seja feliz!
  - **OBS**.: **Login**: Admin, **Senha**: password.
  - **OBS²**.**: Iniciar o BI-Server era apenas para testar se estava tudo funcionando ... :p**

Transformações

Com tudo certo, vamos até a pasta do Pentaho Data Integration (PDI) e execute o arquivo `spoon.sh`.

Nesta parte iremos mostrar como carregar arquivos CSV para o banco de dados. Utilizaremos aqui o postgresql.

-- Colocar link para configuração do pgAdmin e postgres.

Com o **spoon.sh** aberto, você precisa se conectar à sua base de dados

- No Painel do lado esquerdo, vá na aba **View** (ao lado de Design). Clique com o botão direito em **Conexões** e depois **Novo**. Na tela que se abrirá, escolha:

```
Tipo de conexão:PostgreSL
Host Name: localhost
Database Name: [O nome do banco que você criou]
User Name: [Seu login]
Password: [Sua senha]
```

Clique em **Test**, se estiver tudo certo, clique em OK.

Sua base de dados estará listada em **Conexões**. Clique com o botão direito nela e entao clique em **Share** (Compartilhar), para que ela fique disponível para todos os seus futuros trabalhos no PDI.

Depois disso, podemos começar a brincar com alguns arquivos CSV's.

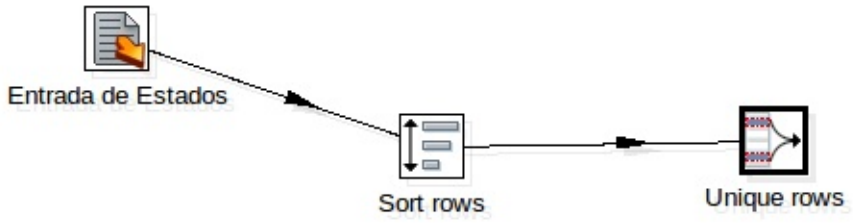
Neste primeiro exemplo, vamos utilizar o arquivo Cod_Municipio.csv. O nosso objetivo é carregar todos os estados desse arquivo (sem repetição) e inserir em nossa base de dados. Para abri-lo, basta ir em **Arquivo** > **Novo** > **Transformações** (Ou apenas Ctrl + n).

Após isso, na aba **Design** no campo **Input**, procure por CSV input file e arraste-o para tela. Dê duplo clique neste ícone e em Step name, renomeie para Entrada de Estados (ou como você desejar). No campo abaixo, filename, navegue até a pasta que se encontra o arquivo Cod_Municipio.csv.

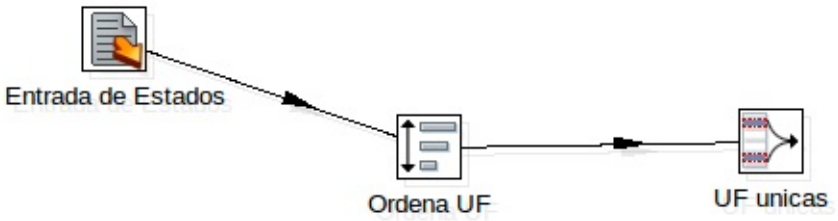
Em alguns arquivos CSV o delimitador é o **';'** (ou um outro qualquer), neste caso, vamos mudar o campo delimiter para **';'**. Logo após, clique em **Obtem Campos** e clique em OK. Você pode ver os dados carregados clicando em **Preview**. Após isso, vá em **Transform** e procure o ícone **Sort rows**. Nós iremos ordenar todos os campos do nosso arquivo, pois o passo seguinte o exige. Arraste-o para a tela, clique no passo Entrada de Estados

segurando a tecla shift e clique no passo atual (Sort Rows). Isso ligará os dois passos. Em seguida, dê duplo clique nele (redefina o nome, se quiser) e na opção **Fieldname**, seleciona o campo UF e OK.

Proximo passo é utilizar o Transform -> Unique Rows. Ligue-o com o passo anterior (Sort rows) e deverá ficar assim:

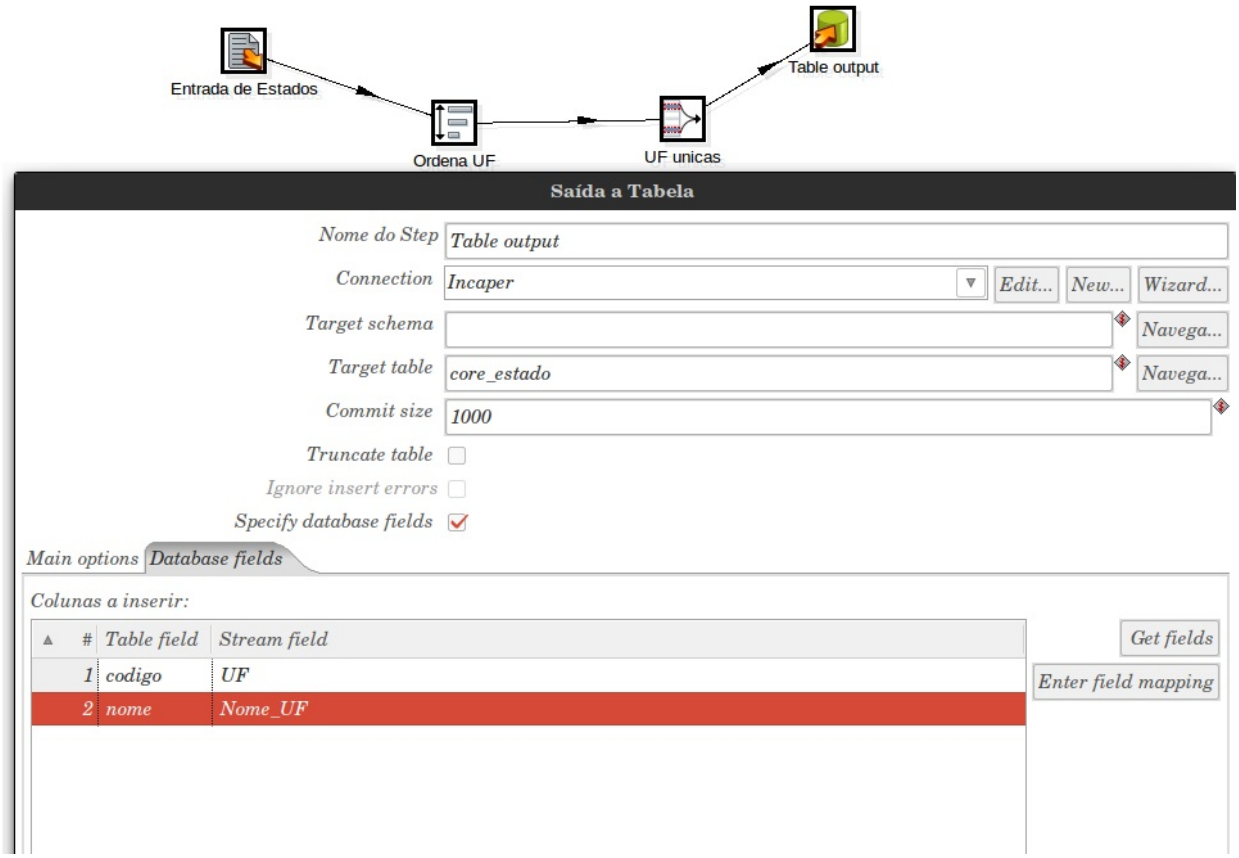


Para ficar mais organizado, vamos renomear 'Sort rows' para Ordena UF e Unique rows para UF únicas:



Duplo clique em UF únicas e no campo Fieldname escolha UF. Isso serve para que ele ignore as repetições que possivelmente encontrará.

O último passo é arrastar **Output -> Table output** para a tela. Table output será a nossa base de dados que conectamos no inicio desse tutorial. Renomeie esse passo para Tabela Estado. No campo Connection escolha a conexão que você criou e em Target table escolha a tabela **core_estado**, que é onde iremos salvar nossos dados. Depois disso marque a opção **Specify database fields** e na aba **Database fields** faça o mapeamento como está abaixo:



Clique em **OK** e tudo pronto! Agora, para executar a transformação, basta apertar a tecla **F9**. Ou **Action > Run**