**Dokumentasi Pengerjaan Tes Web Scraping klikbca.com**

1. Pendahuluan

Sistem harus dapat mengambil data kurs (USD, SGD, EUR, dan UAD) beli dan jual dari landing page <https://www.klikbca.com>. Data kurs tersebut diambil menggunakan 2 tools yaitu python dan pentaho, pengambilan data dilakukan otomatis 1 jam sekali berturut-turut. Data disimpan di dalam file csv dan database MySQL.



Secara garis besar proses pengambilan data dilakukan dengan cara mengambil data html dari web klikbca.com, kemudian mengekstraksi fitur yang diperlukan, dalam hal ini tabel kurs e-Rate, setelah data didapatkan, data disimpan ke dalam file csv dan database MySQL. Untuk dapat melakukan pengambilan otomatisasi perlu dilakukan scheduling, hal ini dapat dilakukan baik itu menggunakan python maupun pentaho.

1. Python

File python dibagi menjadi 2 yaitu python yang menyimpan data di file csv (**sched\_kurs\_klikbca\_1jam\_csv.py**), dan python yang menyimpan data di database (**sched\_kurs\_klikbca\_1jam\_DB.py**). Kemudian keduanya dibuat menjadi file .bat dengan path instalasi python adalahh pada direktori “**C:\Users\Acer\AppData\Local\Programs\Python\Python39\python.exe**” (perlu disesuaikan). Dalam implementasinya digunakan library diantaranya time, urllib, re, datetime, contextlib, dan mysql connector.

Adapun tahapannya adalah sebagai berikut:

1. Melakukan koneksi DB untuk kasus DB, membuat file untuk kasus csv (pembuatan file berdasarkan tanggal)
2. Melakukan request untuk mendapatkan html dari web klikbca.com
3. Mengekstrak fitur dengan menggunakan regex
4. Membersihkan data dengan memperbaiki string dan konversi ke float
5. Insert ke database dan atau csv

Untuk scheduling menggunakan scheduler, waktu looping data diganti pada baris berikut.



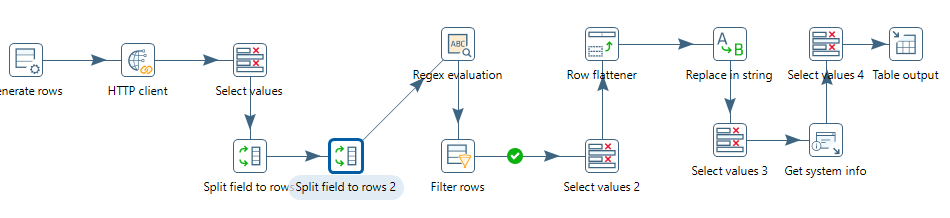
1. Pentaho

Versi pentaho yang digunakan adalah versi 9.1. Dalam pengerjaannya dibuat tiga file *Transformation* dan dua file *Job.* File **getklikbcakurs.ktr** merupakan transformation untuk mengambil data kurs dan simpan di file csv. Sedangkan **getklikbcakursdb.ktr** merupakan transformation yang sama namun dengan output database table. Berikutnya, file **createfilekursklikbca.ktr** adalah tranformation yang membuat file csv untuk menampung transformasi sebelumnya.

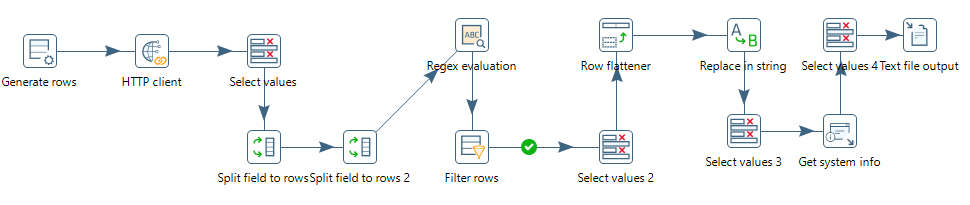
Kemudian job **scheduleklikbca.kjb** merupakan job scheduling yang menjalankan transformation **getklikbcakurs.ktr** dan **getklikbcakursdb.ktr** secara berulang setiap schedule yang ditentukan waktu schedule dapat diubah pada step start di job ini, sedangkan **finishjob.kjb** menjalankan transformastion **createfilekursklikbca.ktr** dan job **scheduleklikbca.kjb,** sehingga menjadikan **finishjob.kjb** sebagai *root process* dari sistem yang dibuat dengan Pentaho.

Adapun proses pengambilan data dilakukan dengan cara yang kurang lebih sama dengan python, yaitu:

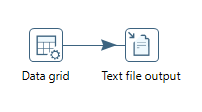
1. Mengambil data menggunakan step HTTP client
2. Row displit menggunakan delimiter ‘<table’ dan ‘<td’, hal ini bertujuan untuk membagi row dan mempermudah proses regex
3. Proses regex untuk memperoleh data yang diinginkan
4. finalisasi data, dengan membersikan simbol yang tidak diinginkan, dan konversi ke numeric, serta menambahkan kolom waktu.
5. Menyimpan data baik itu di database maupun file csv.



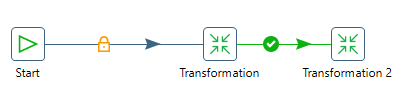
**getklikbcakursdb.ktr**



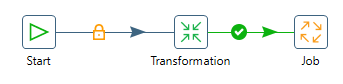
**getklikbcakurs.ktr**



**createfilekursklikbca.ktr**



**scheduleklikbca.kjb**



**finishjob.kjb**

1. Kendala

Kendala yang kerap terjadi pada sistem ini adalah gagalnya request ke website klikbca.com. hal ini terjadi pada python maupun pada pentaho. Pada python apabila hal ini terjadi maka proses akan terhenti dan muncul error, sedangkan pada pentaho akan muncul error namun proses tetap dijalakan selama proses tidak dihentikan secara manual.