Sebastian Smoliński 26. 01.2020 r.

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Laboratorium Wspomagania Decyzji (WDEC)

Laboratorium 6

Hurtownia danych oraz moduły analizay danych

Opis ogólny

Na początku hurtownia pobiera pliki danych z odpowiedniego folderu *newData* i sprawdza ich poprawność oraz format według zadanego:

```
1  /*zmienne globalne*/
2  %let LIBRARY_EXCEL =LAB6;
3  %let LIBRARY_EXCEL_DATA_WAREHOURSE =LAB6.EXCEL_DATA_WAREHOUSE; *Hurtowania danych;
4  %let FILE_NAME_OR_WILDCARD = *.xlsx;
5  %let MY_PATH = /folders/myfolders/data_warehouse;
6
7  /*
8   * utowrzenie magazynu, oraz usuniecie zawartosci,
9   * aby na poczatku nie bylo pustych rekordow przy starcie programu
10  */
11
12 data &LIBRARY_EXCEL_DATA_WAREHOURSE;
13     delete;
14 run;
15
```

Następnie po wczytaniu danych hurtownia wykonuje moduł ładowania danych z plików do bazy odpowiednio przypisując

```
/*
 * wywołanie procedury execute, ktora rozpoczyna dzialanie programu
 */

data testData;
    call execute('%wczytajInformacjeOPlikach');
    proc SGPLOT DATA = lab6.excel_data_warehouse;
    vbar produkt_id / group=ilosc;
    RUN;
    proc SGPLOT DATA = lab6.excel_data_warehouse;
    vbar sklep_id / group=ilosc;
    RUN;
    proc SGPLOT DATA = lab6.excel_data_warehouse;
    vbar sklep_id / group=data;
    RUN;
    proc print data=lab6.excel_data_warehouse;
    var sklep_id Ilosc;
    RUN;
run;
```

Dane są wczytywane z pliku excelowego, poddawane preprocessingowi i następnie, w zależności od poprawności danych, przerzucane do odpowiedniego folderu:

```
%macro wczytajDane(nazwaPliku);

    PROC IMPORT OUT= &LIBRARY_EXCEL
                DATAFILE= "&MY_PATH/newData/&nazwaPliku"
                DBMS=xlsx REPLACE;
                GETNAMES=YES;
    RUN;

    /* makra ktore sprawdzaja poprawnosc danych */
    %czyszczeniePustychWierszy;
    %czyNumerSklepuSieZgadza("&nazwaPliku");
    %czyDataSieZgadza("&nazwaPliku");
    %czySaPustePola;
    %czyIloscJestPoprawna;
    %czyProduktIdJestPoprawne;



    /* Jezeli poprawne dane to sa przenoszone do archiwum i us' ane
       Jezeli niepoprawne dane to sa tylko usuwane */
    data &LIBRARY_EXCEL;
        set &LIBRARY_EXCEL;
        if symget('czyZwalidowane')=0 then /*m' ozmienna ustawiona w makrach sprawdzajacych poprawnosc plikow*/
            do;
                call execute('%usun('||"&nazwaPliku"||')');
                putlog "[LOG] p'ik = &nazwaPliku nie jest poprawny - usuwanie";
                delete;
                stop;
            end;
        else
            do;
                call execute('%przenies('||"&nazwaPliku"||')');
                call execute('%usun('||"&nazwaPliku"||')');
                putlog "[LOG] plik = &nazwaPliku jest poprawny - przeniesienie do archiwum";
            end;
    run;
    /*dane sa dopisywane do lacznego spisu*/
    %uploadData;
%mend wczytajDane;
```

Jeśli plik poprawnie przeszedł walidację, to jest przenoszony z folderu /newData do folderu /archive:

```
 * Kopiuje plik z katalogu /newData/ do katalogu /archive/
 */

%macro przenies(nazwaPliku);
    filename src "&MY_PATH/newData/&nazwaPliku";
    filename dst "&MY_PATH/archive/&nazwaPliku";

    data _null_;
        length msg $ 384;
        rc=fcopy('src', 'dst');
        if rc=0 then
            put ' Plik zostal powielony w newData';
        else
            do;
                msg=sysmsg(); * przechwytuje komunikaty, bledy dot. problemow z systemem plikow, nadpis, uprawnienia etc.;
                put rc= msg=;
            end;
    run;
    filename src clear;
    filename dst clear;
%mend przenies;
```

Plik jest usuwany wtedy z /newData:

```
 * Eksterminacja pliku o podanej nazwie w katalogu /newData/
 */

%macro usun(nazwaPliku);
    data _NULL_;
        putlog "usuniecie pliku &nazwaPliku";
        fname = "tempfile";
        path="&MY_PATH/newData/&nazwaPliku";
        rct=FILENAME(fname, path);
        rc=FDELETE(fname);
    run;
%mend usun;
```

Usunięcie pustych rekordów:

```
/*
 * Usuwa rekordy ktore dla wszystkich pol sa puste.
*/

%macro czyszczeniePustychWierszy;
    data &LIBRARY_EXCEL ;
        set &LIBRARY_EXCEL ;
        if missing (Data) AND missing(Godzina)  AND missing (produkt_id) AND
        missing (Ilosc) AND missing (Sklep_id) then
            delete;
    run;
%mend czyszczeniePustychWierszy;
```

Reguły sprawdzania danych:

- Przedział wartości produkt_id
- Czy są niepełne rekordy
- Czy ilość jest z zakresu 0 - > 1000
- Czy data jest taka sama jak w bazie

- Czy ostatni znak pliku odpowiada Sklep_id