Hurtownie Danych i Przetwarzanie Analityczne
Zadanie nr 2

Skład zespołu:
Paweł Pytel 136786
Szymon Michalak 136769
Marcin Jaskulski 136560

Pod opieką:
prof. dr hab. inż. Robert Wrembel

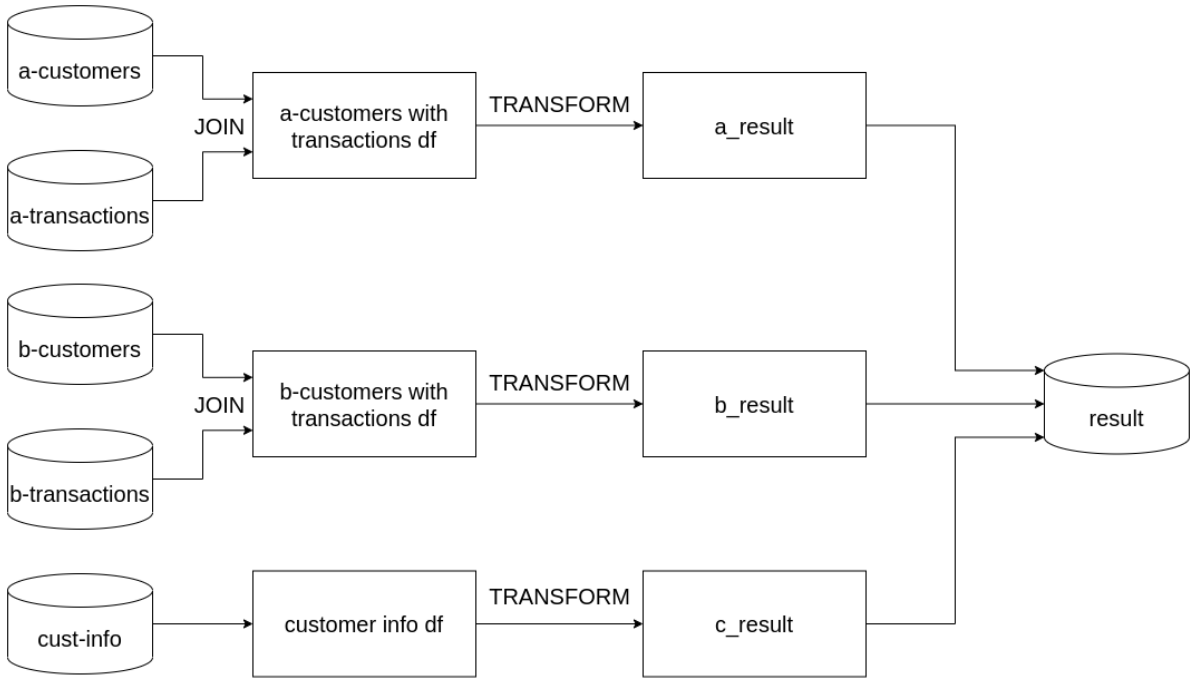
1. Opis koncepcji rozwiązania problemu

Dane z plików, które pochodzą z różnych systemów są wczytywane do pamięci operacyjnej komputera, a następnie przez przystosowane do tego funkcje są dekodowane i parsowane na klasy. Następnie klasy te konwertowane są na obiekty typu Dataframe. Późniejsze operacje, takie jak grupowanie, sortowanie czy filtrowanie są wykonywane właśnie na obiektach Dataframe.

2. Wykorzystane biblioteki i technologie

Do realizacji zadania postanowiliśmy wykorzystać język Python ze względu na jego uniwersalność (możliwość uruchomienia przy pomocy środowiska Windows czy Linux) oraz znajomość różnych bibliotek. Głównie wykorzystana została biblioteka “pandas”. Należu pamiętać, żeby ją zainstalować (np. używając komendy “sudo pip3 install pandas”). To ona pozwala na tworzenie obiektów typu Dataframe. Dodatkowo wykorzystaliśmy wbudowane w środowisko biblioteki do czytania plików tekstowych.

3. Diagram przepływu zadań



4. Pytania i odpowiedzi:

1. Jakiego typu połączenia zbiorów musimy wykonać: inner-join, outer-join?
Zastosowaliśmy połączenie typu outer-join.

2. Czy każda transakcja posiada znanego klienta? Jak to sprawdzić?
Tak, każda transakcja posiada znanego klienta. Aby to zweryfikować należy sprawdzić, czy dane po połączeniu tabel klientów i transakcji zawierają rekordy z pustym polem "custid". Oznacza to, że do danej transakcji nie przypisano żadnego klienta.

3. Czy każdy znany klient z plików A i B posiada co najmniej jedną transakcję?
Nie, niektórzy klienci nie posiadają transakcji. Żeby to zweryfikować należy analogicznie jak w poprzednim punkcie sprawdzić, czy po połączeniu tabel istnieje rekord, który ma puste pole "transid".

4. Czy każda transakcja jest związana z istniejącym klientem?
Tak, każda transakcja jest związana z istniejącym klientem. Aby to zweryfikować należy sprawdzić, czy dane po połączeniu tabel klientów i transakcji zawierają rekordy z pustym polem "custid". Oznacza to, że do danej transakcji nie przypisano żadnego klienta.

5. Czy występują problemy z jakością danych w plikach źródłowych? Jeżeli tak, to jakie? Jak to wpływa na połączenia zbiorów?
Zbiory udało się połączyć stosunkowo bezproblemowo.

6. W jaki sposób przeczytać dane z serwerów mainframe?
Należy odczytać plik używając parametru encoding="ibm037", następnie podzielić go na rzędy wysług zadanej długości linii, a na końcu podzielić wyodrębnione rzędy na pola, korzystając z informacji dotyczącej długości pól.

7. Który rekord (imię i nazwisko) ma największy dochód, a który największą wartość transakcji?
Największy dochód: Leon Bulderberg (109940)
Największa wartość transakcji: zbiór danych B, Michalina Garga (689360)

8. Ile rekordów spełnia kryteria z punktu 1?
Dla domyślnych parametrów wynikowy plik zawiera 5003 rekordy.

9. Ile rekordów odrzucimy w punkcie (1.2)?
Ponieważ punkt 1.2 nic nie mówi o odrzucaniu rekordów, dlatego odpowiedź zostałą udzielona dla punktu 1.3:
Dla domyślnych parametrów dochód o wartości VIP_INCOME przekroczyło 3909 rekordów, jednak zgodnie ze specyfikacją zadania odrzucono 200 rekordów, czyli ich maksymalną liczbę.

10. Czy zmieni się kod jeżeli wolumen danych wzrośnie 100 krotnie (dla każdego pliku)?
Nie, rozwiązanie problemu jest na tyle uniwersalne, że nie wymaga zmiany kodu przy większej ilości danych. Oczywiście im więcej danych tym dłuższy czas przetwarzania.

5. Podział pracy:
   - czytanie plików - Marcin Jaskulski - 4h
   - łączenie danych - Szymon Michalak - 4h
   - testowanie i analiza - Paweł Pytel - 4h
   - raport - wszyscy - 3h (w sumie)