Sofia University „St. Kliment Ohridski”

Faculty of Mathematics & IT

PKDD'99 Discovery Challenge

Document prepared by:

Ivo Pesev, № 62387

Maria Stancheva, № 81862

Tsvetelina Botcheva, №62394

Course:

Data Warehouse and Business Analysis,

Winter Semester 2021 / 2022

Teacher:

Todor Kichukov

February, 2022

Sofia

Table of Contents

1. Description of the task

Once upon a time, there was a bank offering services. Among them: account management, loan offering, etc. The aforementioned bank aspires to improve its services by differentiating between potentially “good” and “bad” clients. The bank managers have only vague ideas about whom they would like to offer some additional services to and whom to watch carefully, in order to try and minimise risk and subsequently - bank losses. Fortunately, the bank stores data about their clients, their accounts, completed transactions, the granted loans and the credit cards issued. So the bank managers hope that there is a way they could take advantage of the data and extract some important insights.

Data description:

- relation **account** (4500 objects in the file ACCOUNT.ASC) - each record describes static characteristics of an account,
- relation **client** (5369 objects in the file CLIENT.ASC) - each record describes characteristics of a client,
- relation **disposition** (5369 objects in the file DISP.ASC) - each record relates together a client with an account,
- relation **permanent order** (6471 objects in the file ORDER.ASC) - each record describes characteristics of a payment order,
- relation **transaction** (1056320 objects in the file TRANS.ASC) - each record describes one transaction on an account,
- relation **loan** (682 objects in the file LOAN.ASC) - each record describes a loan granted for a given account,
- relation **credit card** (892 objects in the file CARD.ASC) - each record describes a credit card issued to an account,
- relation **demographic data** (77 objects in the file DISTRICT.ASC) - each record describes demographic characteristics of a district.

Each account has both static characteristics (e.g. date of creation, address of the branch) given in relation "account" and dynamic characteristics (e.g. payments debited or credited, balances) given in relations "permanent order"

and "transaction". Relation "client" describes characteristics of people who can manipulate the accounts. One client can have more than one account and vice versa; clients and accounts are related to one another through the "disposition" relation . Relations "loan" and "credit card" describe some services which the bank offers to its clients; more than one credit card can be issued to every single account, while at most one loan can be granted per account. Relation "demographic data" provides some publicly available information about the districts (e.g. the unemployment rate); additional information about the clients can be deduced from this.

2. Justification of the selected DWH building approach

To perform the task, our team focused on Ralph Kimball's approach to data warehousing. The main reason for this choice is the simplicity that the data model offers. A star schema is easy to understand and navigate, with dimensions joined only through the fact table. These joins are more significant to the end user, because they represent the fundamental relationship between parts of the underlying business. All in all, the star schema offers simplified querying and analysis and is easy to understand by both technical and business personnel.

We believe that it is not fatal to lose some of the advantages of the other data warehouse design approaches. For this task, we are using a single data source, the incoming data is predictable and its scope is limited, which is why we did not choose Bill Inmon's model. We also think that many of the best benefits of Data Vault like traceability and real-time loading of data are not particularly important for the analysis in our case.

3. Description of the ETL process

The CBS (Core banking system) is the one and only data source. In accordance with the project's goals, all ETL processes are executed once. (in other words, the data is uploaded / updated only once and not daily / weekly / monthly as it happens to be in reality).

- From CBS (from the source) to the SA (staging area)

  All records from the CBS tables are inserted into the corresponding SA tables. The query execution date is added as “loading_date” in each SA table.

  Please refer to the MS Excel document (ETL_CBS_to_SA.xlsx) to access the exact matching table.

- From the staging area to the data warehouse

  Due to the specifics of the task and the fact that the ETL process is performed only once, some dimensions are filled in advance.

  - ***dim_Date*** is a pre-filled table with dates from 01/01/1993 to 31/12/1998 and their characteristics
  - ***dim_Credit_Card_Type*** is a pre-filled table with 3 records describing the possible values for credit card type
  - ***dim_junk_Trans_Operation_Info*** is a pre-filled table with 70 records describing all possible combinations of values of ***“type”***, ***“operation”*** and ***“k_symbol”*** attributes, related to transactions
  - ***dim _K_symbol*** is a pre-filled table with 4 records describing the possible values of ***“k_symbol”*** attribute, related to permanent orders
  - ***dim_junk_Loan_Details*** is a pre-filled table with 20 records describing all possible combinations of values of “duration” and “status” attributes, related to loans
  - ***dim_Account_Ownership*** is a pre-filled table with 2 records indicating whether a particular client is owner of the account
  - ***dim_Account_Priority_Level*** is a pre-filled table with 3 records describing account status according to its balance
  - ***dim_Demographic_Data*** is a projection of all of the attributes of SA_Demographic_Data except ***“loading date”***
  - ***dim_Client***
    - ***client_id*** is ***client_id*** from table ***SA_Client***
    - ***gender*** and ***date_of_birth*** are obtained by ***birth_number***. If the middle two characters represent a number that is:

        - less than 12, then the gender is ‘m’ and the ***date_of_birth*** is ***birth_number*** converted to DATE
        - greater than 12, then the gender is ‘w’ and the ***date_of_birth*** is ***birth_number*** in which 5000 is subtracted from the four-digit number at the end, converted to DATE
    - ***dim_Account_Details*** is a projection of attributes ***“account_id”***, ***“creation_date”*** (converted to DATE) and ***“frequency”*** from ***SA_Account***.
    - ***dim_Trans_Partner_Info*** is a union of
        - projection of attributes ***“bank”*** and ***“account”*** from ***SA_Transaction***
        - projection of attributes ***“bank_to”*** and ***“account_to”*** from ***SA_Permanent_Order***
    - ***fact_Client***
        - A client’s first appearance in ***fact_Client*** contains the date of creation of the first account of which he is the owner. ***“num_of_accounts_owner”*** is the number of acounts created on that date, ***“num_of_loans”*** is the number of loans granted on these accounts.
        - For every first day of a month a row for a client is inserted, the number of accounts being the same if no new accounts are created during this period, or adjusted by adding the number of new accounts. The same thing happens for the number of loans.
    - ***fact_Account***
        - An account’s first appearance in ***fact_Account*** contains the date of the first transaction related to this account. The balance is obtained by the same transaction. As many rows are inserted (daily) for this account as the number of clients related to it.
        - For each subsequent date (from ***dim_Date***) the same number of rows are inserted, the balance being the same as the previous date if there are no transactions on that date, or adjusted by adding the sum of the amounts of all transactions for that date.

- ***fact_Order*** is a projection of attributes ***“order_id”***, ***“account_id”***, ***“amount”***, from ***SA_Permanent_Order***. The value of attribute ***“k_symbol_id”*** is obtained by selecting ***“k_symbol_id”*** from ***dim_K_Symbol*** where the value of ***“k_symbol”*** corresponds to that in ***SA_Permanent_Order***. The value of attribute “partner_id” is obtained by selecting ***“partner_id”*** from ***dim_Trans_Partner_Info*** where the values of both ***“bank”*** and ***“partner_account”*** correspond to those in ***SA_Permanent_Order***.
- ***fact_Credit_Card*** is a projection of attributes ***“card_id”***, ***“issued”*** (converted to DATE) from ***SA_Credit_Card***. The value of attribute ***“account_id”*** is obtained by selecting ***“account_id”*** from ***SA_Disposition*** where ***“disp_id”*** in ***SA_Disposition*** is the same as ***“disp_id”*** in ***SA_Credit_Card***. The value of attribute ***“type_id”*** is obtained by selecting ***“type_id”*** from ***dim_Credit_Card_Type*** where ***“type”*** in ***dim_Credit_Card_Type*** is the same as ***“type”*** in ***SA_Credit_Card***.
- ***fact_Loan*** is a projection of attributes ***“loan_id”***, ***“account_id”***, ***“date”*** (converted to DATE), ***“amount”***, ***“payments”*** from ***SA_Loan***. The value of attribute ***“loan_details”*** is obtained by selecting “id” from ***dim_Loan*** where the values of both ***“duration”*** and ***“status”*** correspond to those in ***SA_Loan***.
- ***fact_Transaction*** is a projection of attributes ***“trans_id”***, ***“account_id”***, ***“date”*** (converted to DATE) and ***“amount”*** from ***SA_Transaction***. ***“trans_operation_info_id”*** and ***“partner_id”*** contain id-s of rows from ***dim_junk_Trans_Operation_Info*** and ***dim_Trans_Partner_Info*** corresponding to the details of each transaction.

4. Source data model

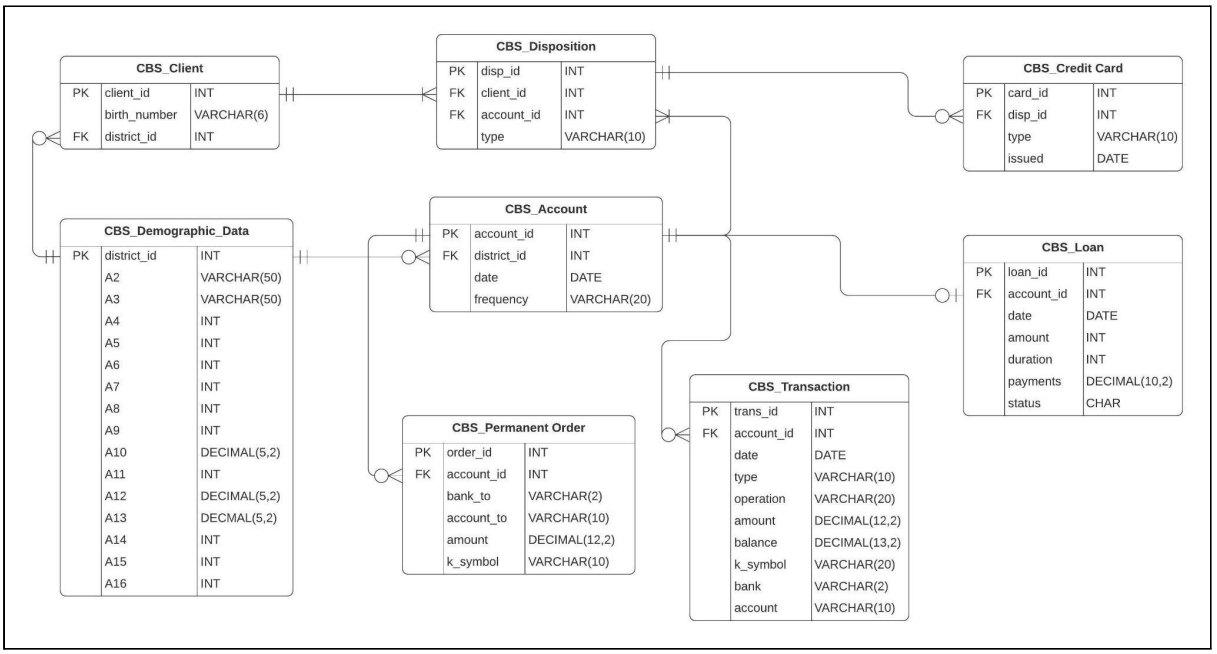

Fig.1. ERD - CBS (source)

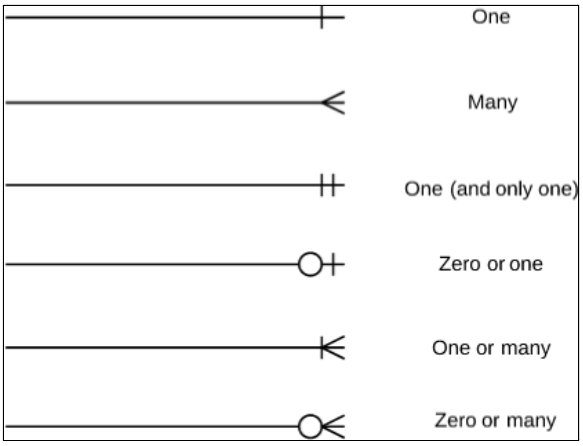

Fig. 2. ERD notations - legend

5. Staging area data model

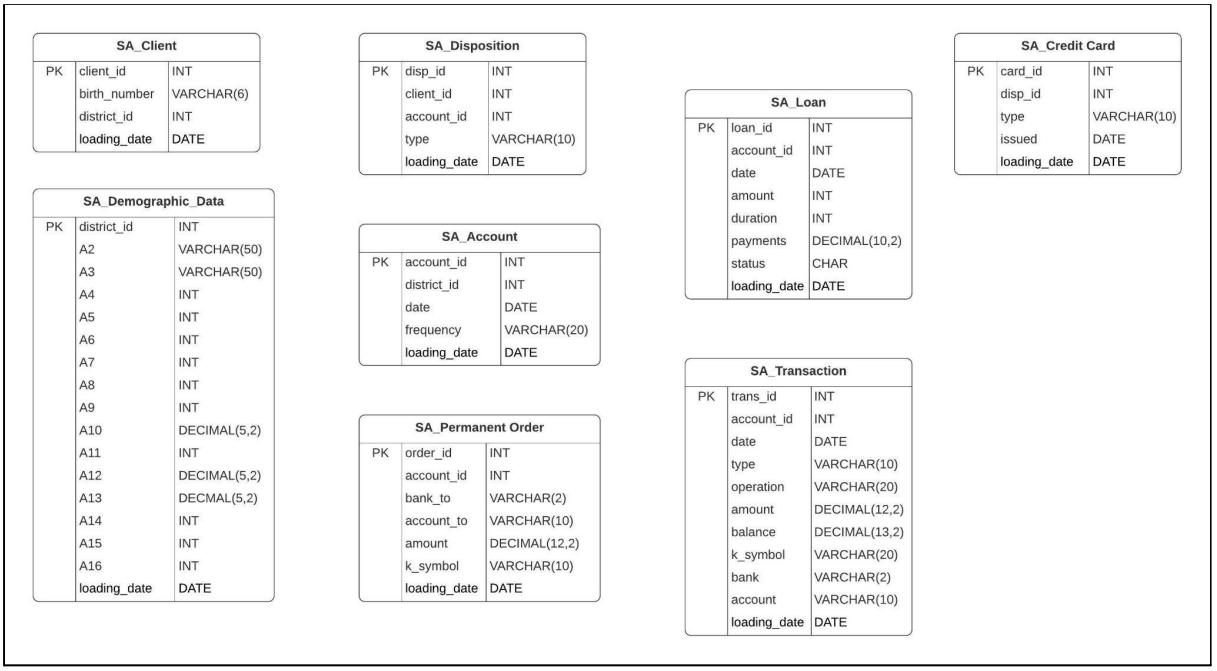

Fig.3. ERD - Staging Area

For the staging area model we have implemented the following changes when compared to the CBS model:

- The tables are not connected to one another any longer
- The are no constraints in the tables (with the exception of PKs)
- A new field - “loading_date” of type DATE - is added to each table, so that it is known exactly when the data was last loaded / updated

6. DWH data model

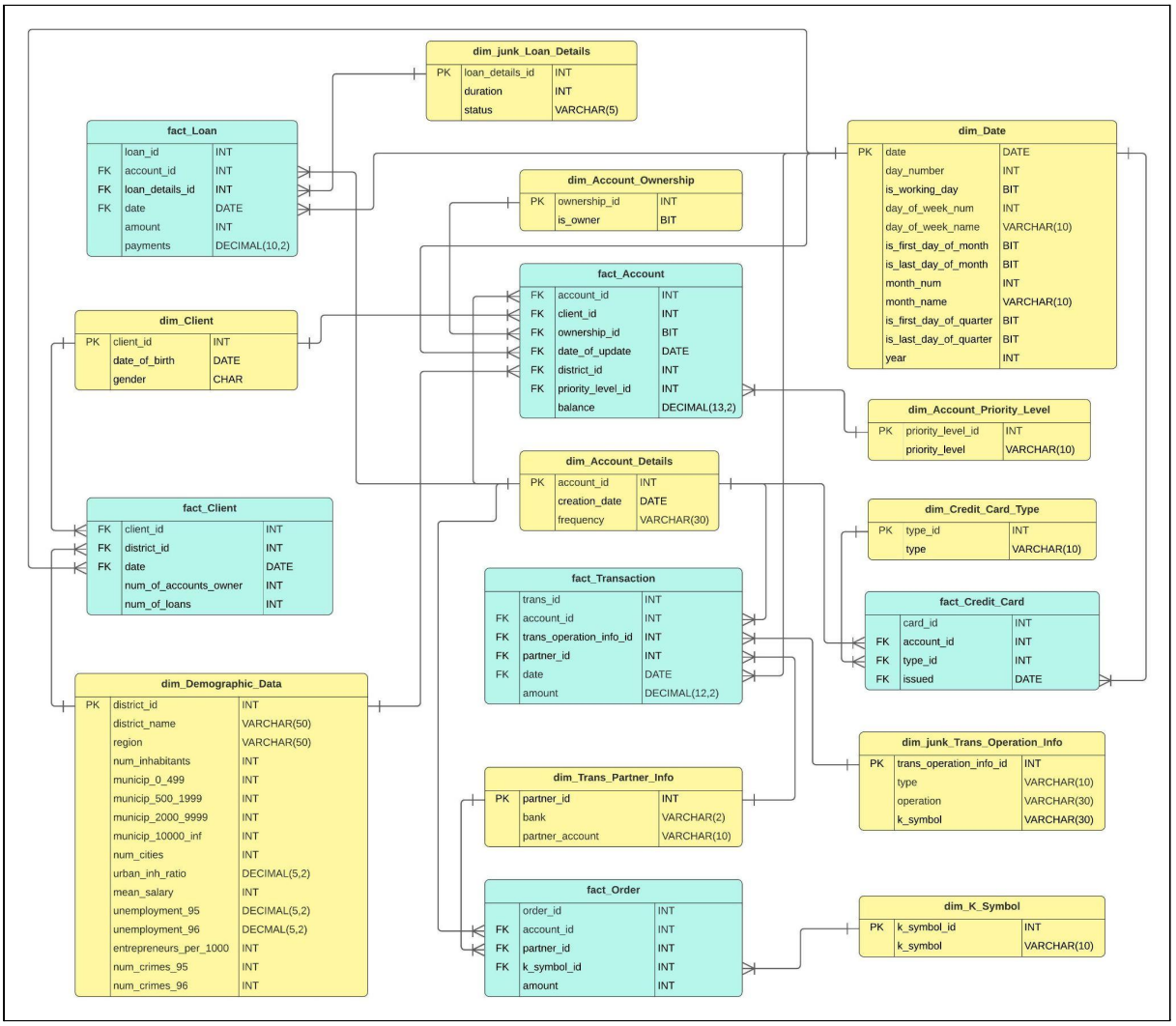

Fig. 4. ERD - Data Warehouse

Our dimensional model consists of 6 fact tables, which are complemented by 11 dimensions. Here is a brief overview of the structure. Under each fact table are listed the dimensions it is associated with:

- fact_Account
    - dim_Account_Details (on ***'account_id'***)
    - dim_Account_Ownership (on ***'ownership_id'***)
    - dim_Account_Priority_Level (on ***'priority_level_id'***)
    - dim_Client (on ***'client_id'***)
    - dim_Date (on ***'date'***)
    - dim_Demographic_Data (on ***'district_id'***)
- fact_Loan
    - dim_junk_Loan_Details (on ***'loan_details_id'***)
    - dim_Account_Details (on ***'account_id'***)
    - dim_Date (on ***'date'***)
- fact_Credit_Card
    - dim_Account_Details (on ***'account_id'***)
    - dim_Credit_Card_Type (on ***'type_id'***)
    - dim_Date (on ***'date'***)
- fact_Transaction
    - dim_Account_Details (on ***'account_id'***)
    - dim_junk_Trans_Operation_Info (on ***'trans_operation_info_id'***)
    - dim_Trans_Partner_Info (on ***'partner_id'***)
    - dim_Date (on ***'date'***)
- fact_Order
    - dim_Account_Details (on ***'account_id'***)
    - dim_Trans_Partner_Info (on ***'partner_id'***)
    - dim_K_Symbol (on ***'k_symbol_id'***)
    - dim_Date (on ***'date'***)
- fact_Client
    - dim_Client (on ***'client_id'***)
    - dim_Demographic_Data (on ***'district_id'***)
    - dim_Date (on ***'date'***)

Overview of the individual tables:

- ***dim_Date***

This is a pre-filled table with dates ranging from 01/01/1993 to 31/12/1998. Numerous fields are added, so that various types of analysis could be conducted. ***Please refer to the CSV document (dim_Date.csv).***

- ***dim_Demographic_Data***

Contains useful information about each district, including the number of inhabitants, number of cities, unemployment figures, crime statistics and more.

- ***fact_Account***

Each row in this table contains data about a single account instance. It should be noted that the number of records associated with an account is dependent on the number of people using it. For example, let us say that a particular account is being used by its owner and two other ‘regular’ users. In this case, in the table there would have to be three distinct records each associating the account with the concrete user or owner.

The ***balance*** field is updated daily. The ***fact_Account*** table can be classified as of type “periodic snapshot”.

- ***dim_Account_Details***

Contains qualitative fields, such as date of creation of the account and frequency of issuing statements.

- ***dim_Account_Ownership***

Contains a boolean field which indicates whether the particular account user is the owner, or not.

➢ ***dim_Account_Priority_Level***

Contains an INT field which classifies the associated account under one of the three balance brackets. The possible values are “small” (balance <= 20000), “medium” (20000 < balance < 80000) and “large” (balance > 80000).

➢ ***fact_Loan***

Contains quantitative data about the granted loans. Each account can be approved for a maximum of one loan. Naturally, the grain is limited to each row describing a single loan associated with its corresponding account. ***“loan_id”*** can be regarded as a degenerated dimension.

The ***“fact_loan”*** table can be classified as of type “transactional”.

➢ ***dim_junk_Loan_Details***

This is a junk dimension containing two qualitative fields: ***“duration”*** (of the loan) and ***“status”***.

➢ ***fact_Credit_Card***

Contains information about the issued credit cards. Each table row represents a credit card ***(“card_id”),*** the account associated with it ***(“account_id”)***, its type ***(“type_id”)*** and its issue date ***(“date”)***. For a single account one or more cards can be issued. For this reason, one and the same card id number can be encountered in several rows. The ***“card_id”*** field can be regarded as a degenerated dimension.

The ***“fact_Credit_Card”*** table can be classified as of type “transactional”.

➢ ***dim_Credit_Card_Type***

This dimension contains information about the possible credit card type

(field ***“type”***).

➢ ***fact_Client***

This table houses quantitative information about all client instances. A single row contains data on a single client of the bank (the date of update is being indicated by the ***"date"*** field), including ***"district_id"*** (address of the client), ***"num_of_accounts_owner"*** (the total number of accounts in regard to which this particular client acts as owner), ***"num_of_loans"*** (the total number of granted loans associated with the accounts of this particular client where they are the owner; ***NB*** obviously this number is always <= than ***"num_of_accounts_owner"***).

The ***fact_Client*** table can be classified as of type "periodic snapshot".

➢ ***dim_Client***

Provides additional information about each client. The fields included (besides ***"client_id"***) are "date_of_birth" and "gender".

➢ ***fact_Transaction***

Each row contains data about a single completed transaction. Namely, the fields are (besides ***"trans_id"***) ***"account_id"*** (the account which initiated the transaction), ***"trans_operation_info_id"*** (an id field which connects the fact table to the dim table containing info on the partner account), ***"partner_id"*** (an id field which connects the fact table to the dim table containing info on the partner account), ***"date"*** (an id field which connects the fact table to the "dim_Date" dimension - the calendar), ***"amount"*** (the transaction's value). ***"trans_id"*** can be regarded as a degenerated dimension.

The ***fact_Transaction*** table can be classified as of type "transactional".

➢ ***dim_junk_Trans_Operation_Info***

This is a junk dimension which contains all the possible combinations (the Cartesian product) of its fields. To be more particular - all combinations of ***"type"*** (indicates whether the transaction is of type '+' or '-'), ***"operation"***

(indicates the mode of the transaction), ***"k_symbol"*** (characterises the transaction, e.g. if it is classified as an old-age pension, loan payment, etc.).

➢ ***dim_Trans_Partner_Info***

Each row contains 2 distinct values (besides the ***"partner_id"*** field) - "bank" (bank of the partner / recipient) and "partner_account" (account of the partner / recipient).

➢ ***fact_Order***

Each row contains data about a single particular order. The fields include (besides ***"order_id"***) ***"account_id"*** (the account making the order), ***"partner_id"*** (an id field which connects the fact table to the dim table containing info on the receiving account), ***"k_symbol_id"*** (an id field which connects the fact table to the dim table containing info on k-symbol), ***"date"*** (a DATE field which connects to ***"dim_Date"*** and indicates the exact date when the order was issued), ***"amount"*** (the debited amount). ***"order_id"*** can be regarded as a degenerated dimension.

The ***fact_Order*** table can be classified as of type "transactional".

➢ ***dim_K_Symbol***

This dimension is used exclusively by the ***"fact_Order"*** table. It contains all 4 possible values for the k-symbol (in other words, the characterization of payment - whether it is an insurance payment, loan payment, etc.).

7. Useful data reports

- Group loans by their status ("contract finished, no problems" and "running contract, OK so far" correspond to "LOANS WITHOUT PROBLEMS"; "contract finished, loan not paid" and "running contract, client in debt" correspond to "PROBLEMATIC LOANS") and frequency of issuance of statements of the accounts related to the loans

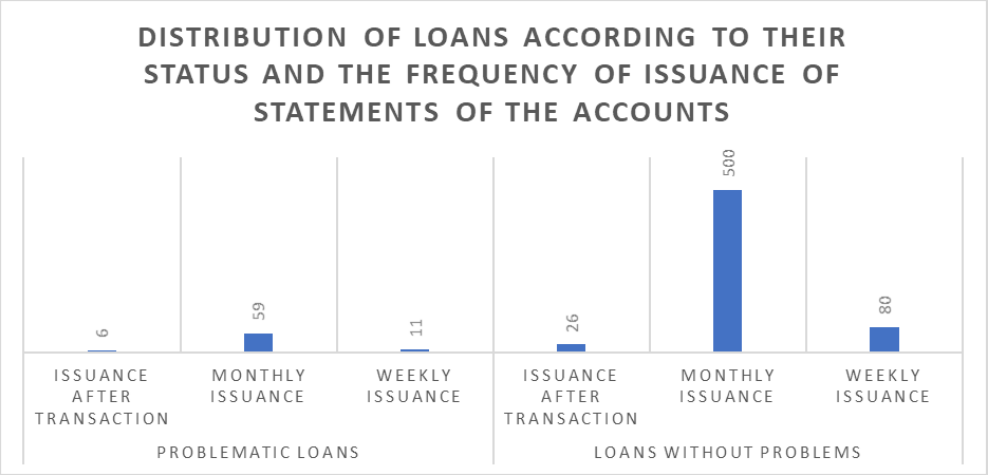

Fig.5. Report related to the loans (Visualisation)

The results show that most loans are related to accounts with monthly issuance of statements. This means that clients who are owners of accounts with lower frequency of issuance of statements are more likely to be attracted by the bank to take out loan.

The observed trend is the same for both loans without problems and problematic loans (which means that whether the client will pay off the loan is not related to the frequency).

- Find top 5 distinct accounts with biggest sanction interest. This is done by querying the ***"fact_Transaction"*** table and filtering by the attribute ***"k_symbol"*** (looking for the value "sanction interest"), part of ***"dim_junk_Trans_Operation_Info"***. The final result is the maximum sanction amount and its corresponding ***"account_id"*** number.

Table 1. Report №2 - results

account_id	***amount***
5092	332.70
3326	283.50
4356	252.50
2305	241.20
5429	229.80

8. Used software tools

- MS SQL Server
- MS Excel

9. Sources

Web:

- ***Data Warehouse Concepts: Kimball vs. Inmon Approach | Astera***
- ***Inmon or Kimball: Which approach is suitable for your data warehouse?***

Literature:

- “Data Warehouse and Business Intelligence Fundamentals” Course Lectures prepared by Todor Kichukov