The Piggybank is an American banking start-up. So far, bank has not that many customers, however it expects rapid growth in upcoming fiscal year. Currently all customers data is stored in CSV (Comma Separated Values) files and data is processed in the MS Office – all monthly reporting such as risk monitoring DPD30+, DPD60+, DPD90+ and overall credit portfolio summary.

*DPD explanation:*

*A loan is past due if the PAST\_DUE\_AMT is greater than $100. Each loan in the data set has instalment payment date on a day before the last day of the month (e.g., 30.03.2020). Number of days past due is equal to 1 on the first reporting date, where PAST\_DUE\_AMT is greater than $100. If the amount past due continues to be greater than $100 in the following month, number of days past due grows accordingly.*

*DPD buckets are calculated based on number of days past due:*

*0 – if number of days past due is smaller than 30 days*

*1 - if number of days past due is equal or greater than 30 days and smaller than 60 days*

*2 - if number of days past due is equal or greater than 60 days and smaller than 90 days*

*3 - if number of days past due is equal or greater than 90 days*

The IT management has decided to switch from CSV files and use RDBMS (Relational Database Management System) to satisfy current and upcoming business needs.

You, as data engineer were asked to try out open-source RDBMS solution (SQLite), migrate all CSV filed to it and recreate business reports that are running now in MS Office.

Documentation regarding the data structures that are used by business is following:

Customer data is stored in four CSV files:

1. Client – this file contains information about customers, such as age or education,
2. Income – in this data set Piggybank is storing information about customer income,
3. Household – contains information regarding customer family (like marital status, number of children etc.)
4. Loan – in this file, your company stores information about loan amount, number of instalments and “default” if occurs.

(As “default” one should understand a client who has not been paying his/her debt off for more than 90 days)

Technical documentation provided to you by the data architect contains the information about attributes in each file, their business meaning and range of values (if applicable). However, there is no information about relationship between tables nor physical requirements (such as information about primary keys (PK), foreign keys (FK), or granularity).

You have been asked to figure it out and propose logical and physical model in RDBMS (such as entity names, indexation, PK, FK etc.). Mentioned documentation can be found in the attached docx file:



Work is performed with use of agile tools, such as JIRA. It is a kind of ticketing system to breakdown large tasks into smaller chunks.

On JIRA board, series of tasks have been assigned to you:

1. Install open-source SQLite ( <https://sqlitestudio.pl/> ) on your local machine.
2. Create database Piggybank and import all CSV files to table structures:
   1. Define data types,
   2. Propose indexes, PK’s, FK
   3. Provide justifications for your decisions around physical designs (indexes, PKs, FKs, other).
3. Check the data quality, create analyses reporting data quality and apply data cleansing if needed:
   1. Validate data quality for provided data (leverage on documentation where values and constraints are marked out).
   2. Measure data quality – propose metrics that can help a company to measure data quality efforts.
   3. Propose and provide visualization of above metrics.
   4. Provide explanation and description of SQL code used for data quality check and data cleansing.
4. Design and implement physical model for CUSTOMER\_CAR (Customer Analytical Record) table, which will hold all attributes from all four CSV files in one row (mind that your company is expecting large volume of data in upcoming fiscal year).
5. Create data pipeline process, that will append data to CUSTOMER\_CAR table:
   1. Discuss how often data should be loaded.
   2. How could you handle need of data correction?
6. Based on loaded data to CUSTOMER\_CAR table, recreate DPD bucket and portfolio summary reports (you can use any tool for visualization, including XLS or PPTX):
   1. Summary report – dashboard for last 3 reporting periods with information about total loan amount, past due amount and already paid loan amount.
   2. DPD bucket report – three charts (DPD30+, DPD60+, DPD90+) with past due amount per reporting date for last 12 months.
7. Prepare evidence-based analysis to IT management about open-source RDBMS and supply your recommendations.
8. Prepare the final documentation, which should have:
   1. Overview of used data, including physical model witch your justification of indexation, PK’s, FK’s (output from task no. 2).
   2. Summary of data quality metrics which can be applied, including dashboard with visualization (output from task no. 3)
   3. Entity relation diagram (ERD) for source tables
   4. Summary for data pipeline process (output from task no. 5)
   5. Your recommendation about RDBMS (output from task no. 7)