Business intelligence and analytics
(Data warehouse)

Cristian De Marco 204794
Department of Mathematics and Computer Science, UNICAL

May 1, 2019

Contents

Chapter 1

Dataset description

Kiva loans Kiva.org is an online crowdfunding platform to extend financial services to poor and financially excluded people around the world.

1.1 Data description

kiva_loans.csv This file contains informations about the loans disbursed in several countries from 1-1-2014 to 31-10-2016:

- **id**: Unique identifier for loan.
- **funded_amount**: The amount disbursed by Kiva to the field agent (USD).
- **loan_amount**: The amount disbursed by the field agent to the borrower (USD).
- **activity**: More granular category.
- **sector**: More general category.
- **use**: Exact usage of the loan amount.
- **country_code**: ISO country code of the country in which the loan was disbursed.
- **country**: Full name of the country in which the loan was disbursed.
- **region**: Full name of the regionin which the loan was disbursed.

- **partner_id**: Identifier of partner organization.
- **posted_time**: The time at which the loan is posted on Kiva by the field agent.
- **disbursed_time**: The time at which the loan is disbursed by the field agent to the borrower.
- **funded_time**: The time at which the loan posted to Kiva gets funded by lenders completely.
- **term_in_months**: The duration for which the loan was disbursed.
- **lender_count**: The number of lenders that contributed to this loan.
- **tags**: Additional tags.
- **borrower_genders**: Comma separated M,F letters that represent the gender of an individual in the group.
- **repayment_interval**: The type of repayment.
- **date**: The date at which the loan was disbursed.

loan_themes_by_region.csv This file has been used to get informations about partners and related sectors:

- **partner_id**: Unique identifier for partner.
- **field_partner_name**: Full name of the partner organization.
- **sector**: Sector of the partner organization.

Chapter 2

Design

2.1 Design

Three-tier architecture A three-tier data architecture has been used for the realization of the data warehouse. Such kind of architecture allows to keep the extraction and transformation phases separated from the warehouse loading phase. After cleaning and transforming the data (using Pentaho as ETL tool) , these are been stored in the operational data store, that will be subsequently used to feed the data mart tier.

2.2 Entity-Relationship model

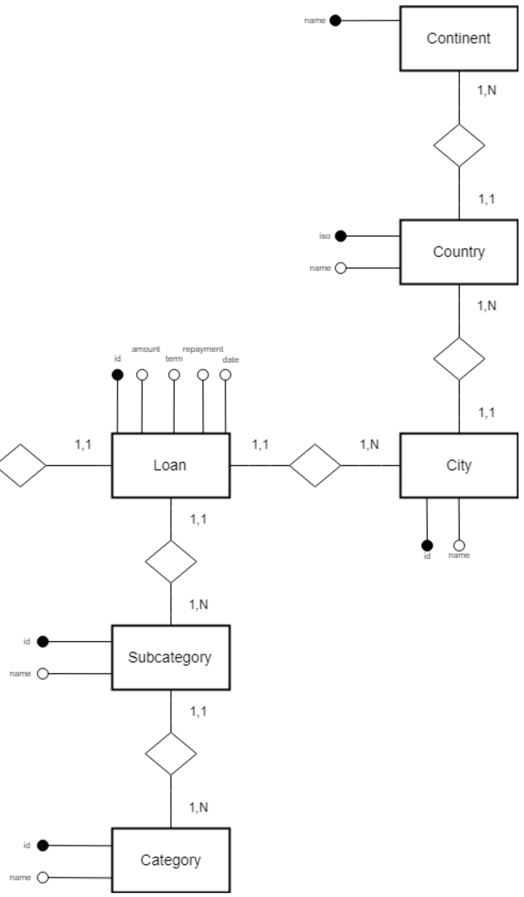


Figure 2.1: E-R model for the loans operational data store

Chapter 3

Conceptual design

3.1 Attribute tree

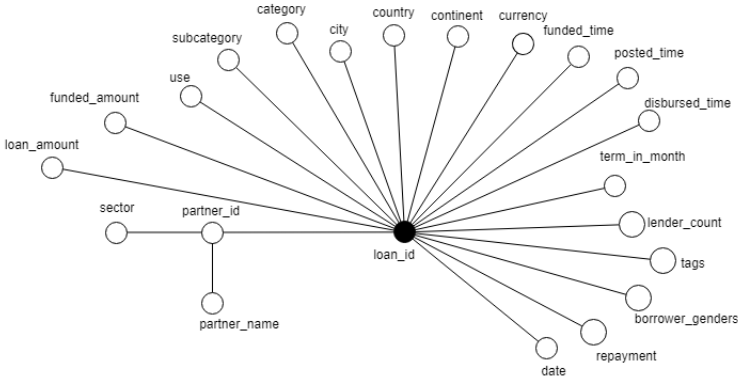


Figure 3.1: The attribute tree corresponding to the loan

3.2 Restructured attribute tree

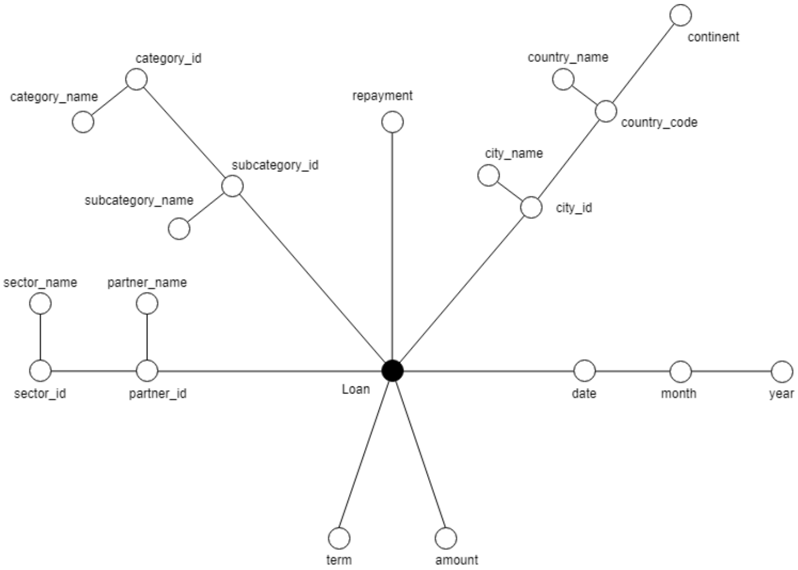


Figure 3.2: The restructured attribute tree

Pruning The following attributes has been pruned:

- funded_amount
- currency
- posted_time
- disbursed_time
- funded_time
- lender_count
- tags
- borrower_genders

Dependencies The following dependencies has been added:

- date - month - year
- city - country - continent
- subcategory - category

3.3 Dimensional fact schema

In this phase the loan has been selected as relevant fact of the source schema. Then measures and dimensions has been defined.

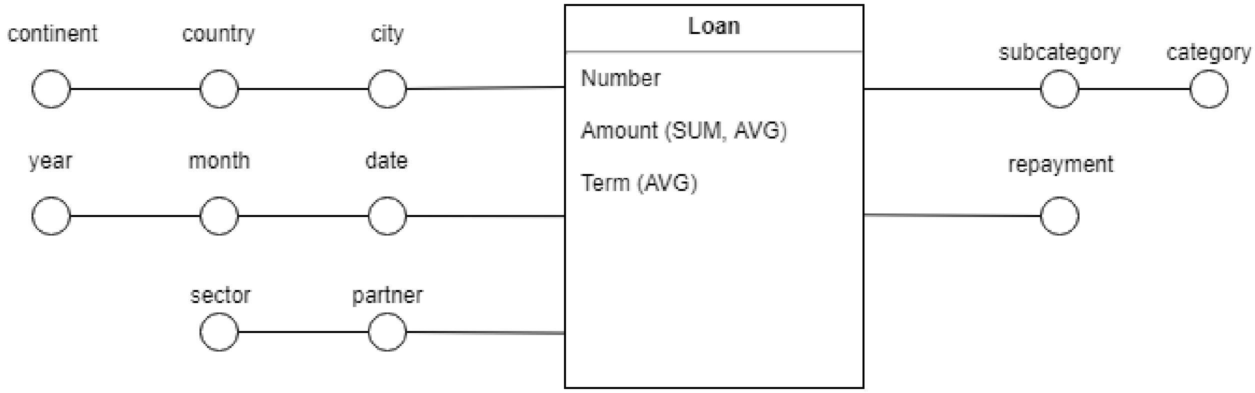


Figure 3.3: The dimensional fact schema corresponding to the loan fact

Chapter 4

Logical design

4.1 Star schema

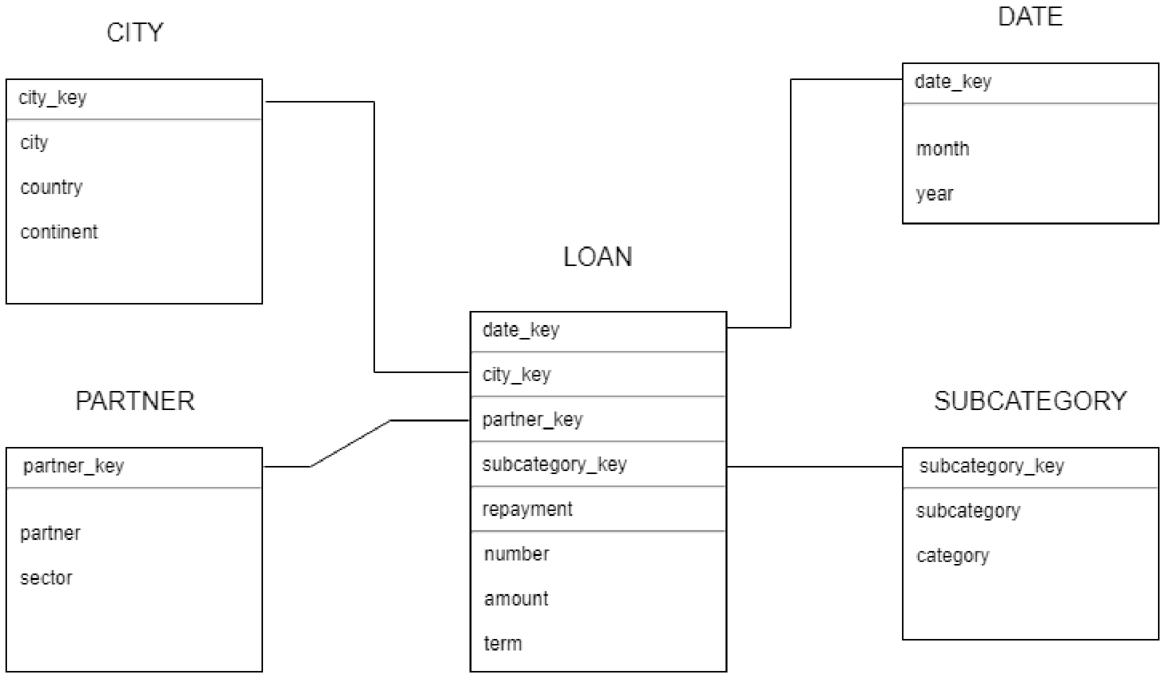


Figure 4.1: Modeling the loans with the star schema

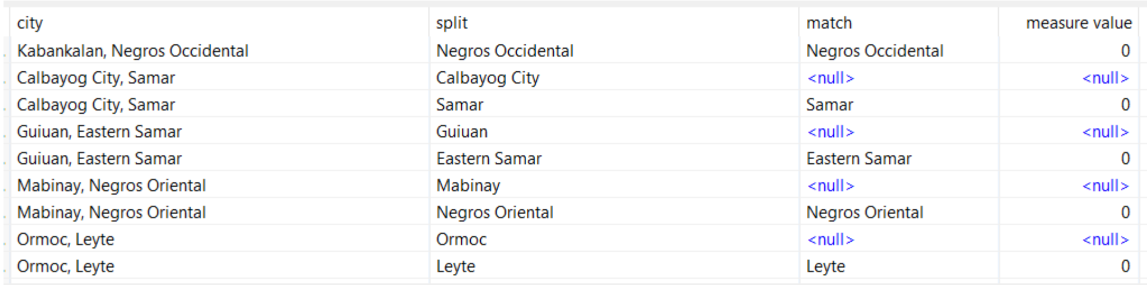
Chapter 5

Data cleaning

5.1 Kiva loans cleaning

- **Continent**: The continent field has been added by using an external file containing the countries and the relative continents. A dictionary tecnique has been used in order to set to each country in the data source the proper continent.
- **Null values** Several records had null values in the following fields:
  - **partner_id**: 13350 records (1.98%). The field has been set to Unknown.
  - **city**: 56264 records (8.37%). The field has been set to Unknown.
  - **almost all the columns**: 3983 records (0.59%). These records have been removed.
- **Bad formatted**: 17 records (0.002%). These records have been removed.
- **Duplicates**: 23559 records (3.5%). These records have been removed.

- **Cities**: This field was the dirtiest in the dataset. Cities have been cleaned by using several external files (according to the countries that occur more) and therefore by using a dictionary technique (Fuzzy match).



city	split	match	measure value
Kabankalan, Negros Occidental	Negros Occidental	Negros Occidental	0
Calbayog City, Samar	Calbayog City	<null>	<null>
Calbayog City, Samar	Samar	Samar	0
Guiuan, Eastern Samar	Guiuan	<null>	<null>
Guiuan, Eastern Samar	Eastern Samar	Eastern Samar	0
Mabinay, Negros Oriental	Mabinay	<null>	<null>
Mabinay, Negros Oriental	Negros Oriental	Negros Oriental	0
Ormoc, Leyte	Ormoc	<null>	<null>
Ormoc, Leyte	Leyte	Leyte	0

Figure 5.1: A screenshot of the fuzzy match for cities cleaning

- **Missing partners** The partners file had 303 unique partners. The loans file had 367 unique partners. The partner of the 17304 (2.57%) records with missing partners have been set to Unknown

Chapter 6

Analysis

6.1 Worksheet

- Number of loans disbursed by category / subcategory
- Number of loans disbursed by sector / partner
- Number of loans disbursed over the years / quarters
- Correlation between country poverty and amount disbursed
- Average of term by partner

6.2 Dashboard

- Number of loans and amount disbursed by country, partner and category