Part 1: Conceptual Design

CSI 4142 - Fundamentals of Data Science
Winter 2023

uOttawa

Faculté de génie
Faculty of Engineering

School of Electrical Engineering and Computer Science
University of Ottawa

Yazan Otoum, Ph.D

Group #30 : Deliverable #1

Ishanveer Gobin 300135454
Andie Samadoulougou 300209487
Kate Sin Yan Chun 300144923

Table of Content

The Grain

The grain refers to the level of details or granularity at which the model represents a concept. It may be explained by considering what data is stored in one row of the fact table.

In this conceptual model, the grain can be “A single transaction for an applicant using a specific device during a session which is either fraudulent or not considering the similarity between their name and email, and the volume of transactions over the past 6 hours, 24 hours and 4 weeks”.

Dimensions and Dimensional Attributes

In this dimensional model, we have a fact table and 4 dimensions (Transaction, Applicant, Device and Session). The fact table has 4 foreign keys from the 4 dimensions as well as a fraud_bool attribute which indicates whether the bank account is fraudulent or not.

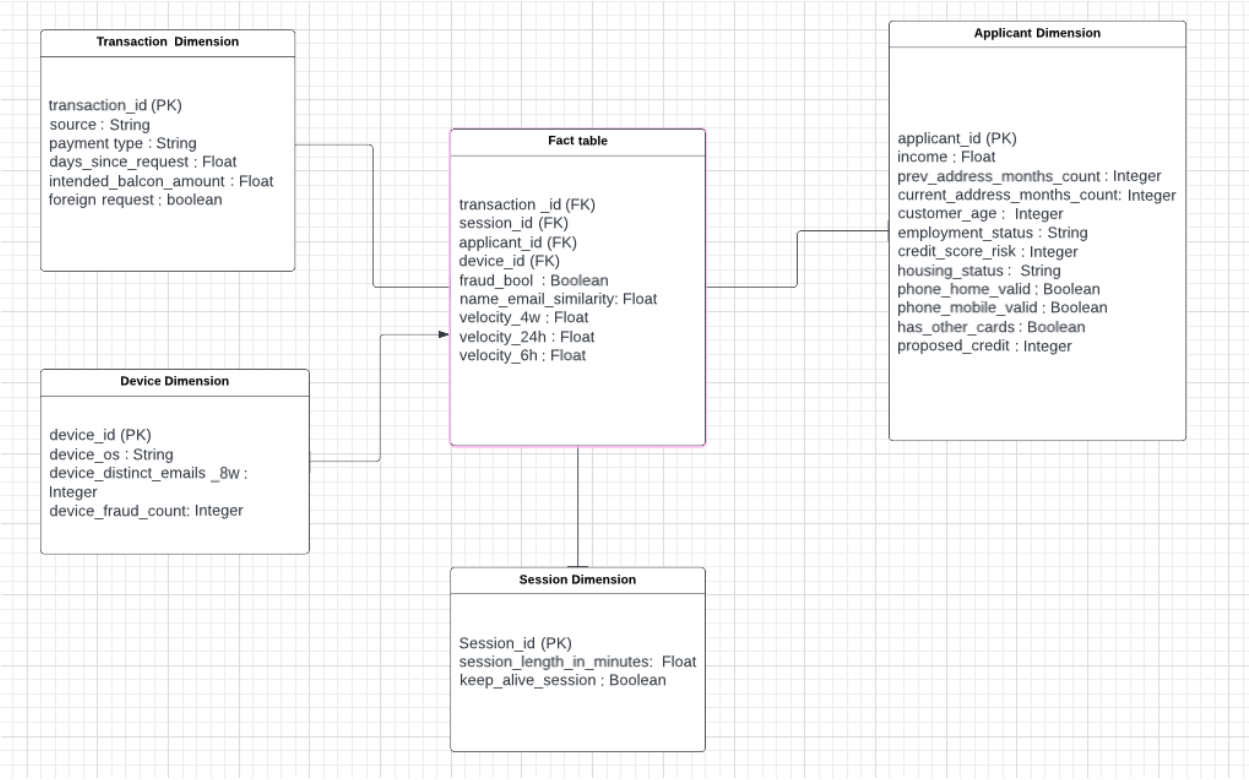
- Transaction Dimension
  - transaction_id (Primary Key)
  - Source : string, sample values = “INTERNET”
  - Payment type : string, sample value = “AB”
  - Days _since_request : float, minimum = 0, maximum 78.5, sample value = 30
  - Intended_balcon_amount : float, minimum = -1, maximum = 108, sample value = -1,205
  - Foreign_request : boolean , sample value = 0

- Applicant Dimension
  - Applicant_id (Primary Key)
  - Income : integer, minimum = 0.1, maximum = 0.9 , sample value = 0.2
  - Prev_address_months_count : integer, minimum -1, maximum = 383 , sample value = 92
  - current_address_months_count : integer, minimum = -1 ,maximum = 406, sample value = 88
  - Customer_age : integer , minimum = 10, maximum = 90 , sample value = 50
  - Employment_status : string , sample value = “CA”
  - Credit_score_risk : integer , minimum = -176 , maximum = 387, sample value = 185
  - Housing_status: string , sample value = “BA”
  - Phone_home_valid : boolean, sample value = 0
  - Phone_mobile_valid : boolean, sample value = 1
  - Has_other_cards : boolean, sample value = 0
  - Proposed_credit : integer, minimum = 200 , maximum = 2000, sample value = 500

- Session Dimension
  - Session_id (Primary Key)
  - Session_length_in_minutes : float, minimum = -1 , maximum = 107, sample value = 3.77
  - Keep_alive_session : boolean, sample value = 0

- Device Dimension
    - Device_id (Primary Key)
    - Device_os: String , sample value = “windows”
    - Device_distinct_emails_8w : integer, minimum = 0 , maximum = 3 , sample value = 1
    - Device_fraud_count: integer , minimum = 0, maximum = 1 , sample value = 0

Dimensional Model



Measures and Facts

Measures are the numerical values that represent the key aspects of business processes or activities. These values are used to analyze and understand the data in the data mart.

Facts are events or occurrences that are recorded in the data mart and are usually related to a measure.

In this dimensional model, the measures/facts are the following:

- fraud_bool : Boolean , sample value = 1
    - Represent if a transaction is fraudulent or not
- name_email_similarity : Float, minimum = 0, maximum = 1, sample value = 0.16682773442433269
    - Metric of similarity between email and applicant's name. Higher values represent higher similarity.
- velocity_6h : Float, minimum = -171, maximum = 16700 , sample value = 734.04
    - Average number of applications per hour in the last 6 hours
- velocity_24h : Float, minimum =1300, maximum = 9570, sample value = 2670.91
    - Average number of applications per hour in the last 6 hours
- Velocity_4w : Float, minimum= 2830, maximum = 6990, sample value = 3124.2981
    - Average number of applications per hour in the last 4 weeks

Assumptions

We do not have any assumptions at this moment.

Checklist of “10 design mistakes”

Design mistakes	Description of how we avoided/handled the mistakes
Place text attributes in the Fact table	In the fact table, we only included the keys for each dimension (as foreign keys) and added the measures/facts. Therefore, there are no text attributes in the fact table.
Limit verbose description to save space	Most of the text attributes in the dataset are descriptive and expressed in a clear and concise language.
Normalize to save space	The dimensional model has not been normalized as normalization increases the number of join operations, which can decrease query performance. To maintain high query throughput, normalization is not desired in this case.
Ignore the need to track changes	We considered the need to track changes by including slowly changing dimensions. In this case, a slowly changing dimension can be the applicant dimension where the attributes such as customer_age and current_address_months_count will change with time.
Add new hardware to solve all query performance issues	Before adding new hardware to solve query performance issues, we will first understand the root cause of the problem and weigh its benefits and costs before making a decision.
Use operational keys as the primary keys	We will be using auto generated keys that do not change as primary keys.
Neglect to declare and comply with the grain	We identified and declared the grain. Further in this project, we will ensure that the grain is consistently followed throughout the dimensional model. Additionally, we will regularly review the dimensional model and ensure that the grain is still appropriate and that compliance is maintained.

Neglect a detailed design	Our design is highly detailed and although we only have 4 dimensions, each of them contains all the relevant attributes for our project.
Expect users to query normalized data	We denormalized the data so that users are not expected to query normalized data.
Fail to conform Facts and Dimensions	We considered how to conform Facts and dimensions so that the data in our tables are consistent and the relations between them are accurate. We plan to validate the data as it is loaded, check for inconsistencies and make corrections if necessary.

References

Bank account fraud dataset suite(NeurIPS 2022). (n.d.). Retrieved 5 February 2023, from

https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

ChatGPT : https://chat.openai.com/chat

Dataset : https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?select=Base.csv

Lucid chart: https://lucid.app/lucidchart/44ce6ace-a31a-48c8-a279-eee1fd1b3e36/edit?invitationId=inv_4fe5e281-fe4a-48f0-9034-b5890caca466&page=0_0#

CSI4142: Test 1 Sample