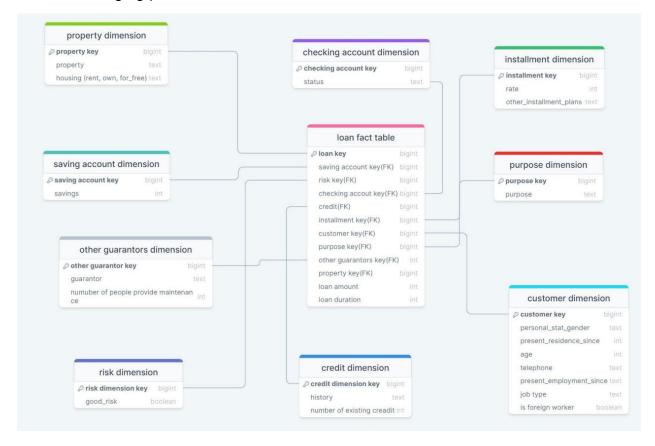Git link:

https://github.com/SuperChentao/creditDataProcessing

schematic staging plan



All the data from the same data source: German Credit Scoring Data

https://www.kaggle.com/datasets/elsnkazm/german-credit-scoring-data

(discussed with TA, single abundant data sources is acceptable)

the source data is well organized (binning, label) already, but I just carried out data cleaning procedure for the raw data. From the data source description, I grabbed the valid values/range for each column and applied the filters (made by valid values) to them, for example:

```python
# remove out range value other_installment_plans
other_installment_plans_values=['none', 'bank', 'store']
df=df[df['other_installment_plans'].isin(other_installment_plans_values)]
```

And delete missing data row (if any). I deleted them because there were enough data:

```
Check for null value

    df=df[~df.isnull().any(axis=1)]
348]   ✓  0.0s
```

I also separate some columns to better organize the data in data mart:

```
    df[['gender','marriage']]=df['personal_stat_gender'].str.split(':',expand=True)
    df=df.drop('personal_stat_gender',axis=1)
9]   ✓  0.0s
```

After creating and relating all the dimensions and fact table, I created database instance and load the data into it:

**installment**
- 🔑 id integer
- rate integer
- other_installment_plans character varying(20)

**credit**
- 🔑 id integer
- history character varying(40)
- num_credit integer

**checking_account**
- 🔑 id integer
- status character varying(20)

**saving_account**
- 🔑 id integer
- savings character varying(30)

**customer**
- 🔑 id integer
- gender character varying(10)
- marriage character varying(20)
- residence_since integer
- age integer
- has_telephone boolean
- employment_since character varying(20)
- job_tyle character varying(30)
- foreign_worker boolean

**loan**
- 🔑 id integer
- saving_account_id integer
- checking_account_id integer
- credit_id integer
- installment_id integer
- customer_id integer
- purpose_id integer
- other_guarantors_id integer
- property_id integer
- duration integer
- loan_amt integer
- good_risk boolean

**other_guarantors**
- 🔑 id integer
- guarantor character varying(20)
- num_providor integer

**property**
- 🔑 id integer
- property character varying(40)
- housing character varying(20)

**purpose**
- 🔑 id integer
- purpose character varying(20)