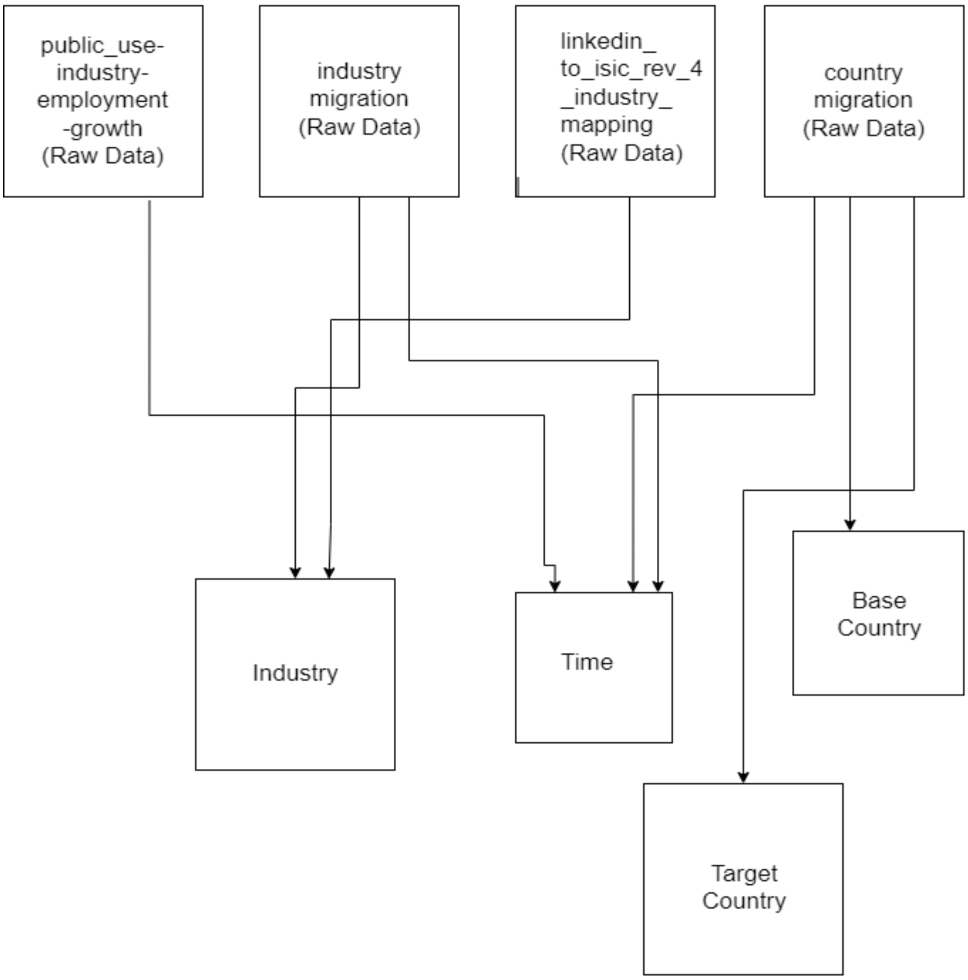
CSI4142 Project

Data Staging

Group 16

Winter 2023

A - Data Staging Plan



Data staging the dimensions

1. Base Couintry Dimensions

    Load data from country_migration.csv
    checking for null values and data types
    cleaning values and aggregate operations

2. Target Country Dimensions

    Load data from country_migration.csv
    checking for null values and data types
    cleaning values and aggregate operations

3. Industry Dimensions

    Load data from industry_migration.csv
    Load data from linkedin_to_isic_rev_4_industry_mapping.csv
    drop the unnecessary from the
    linkedin_to_isic_rev_4_industry_mapping.csv

merge the two data together
checking for null values and data types

4. Time Dimensions
   Load data from country_migration.csv
   Load data from industry_migration.csv
   Load data from public_use-industry-employment growth
   add a year column and remove the migration_year column
   merging them together

Create Fact Tables

B - Other details

1. We used Jupyter Notebook to do the data staging

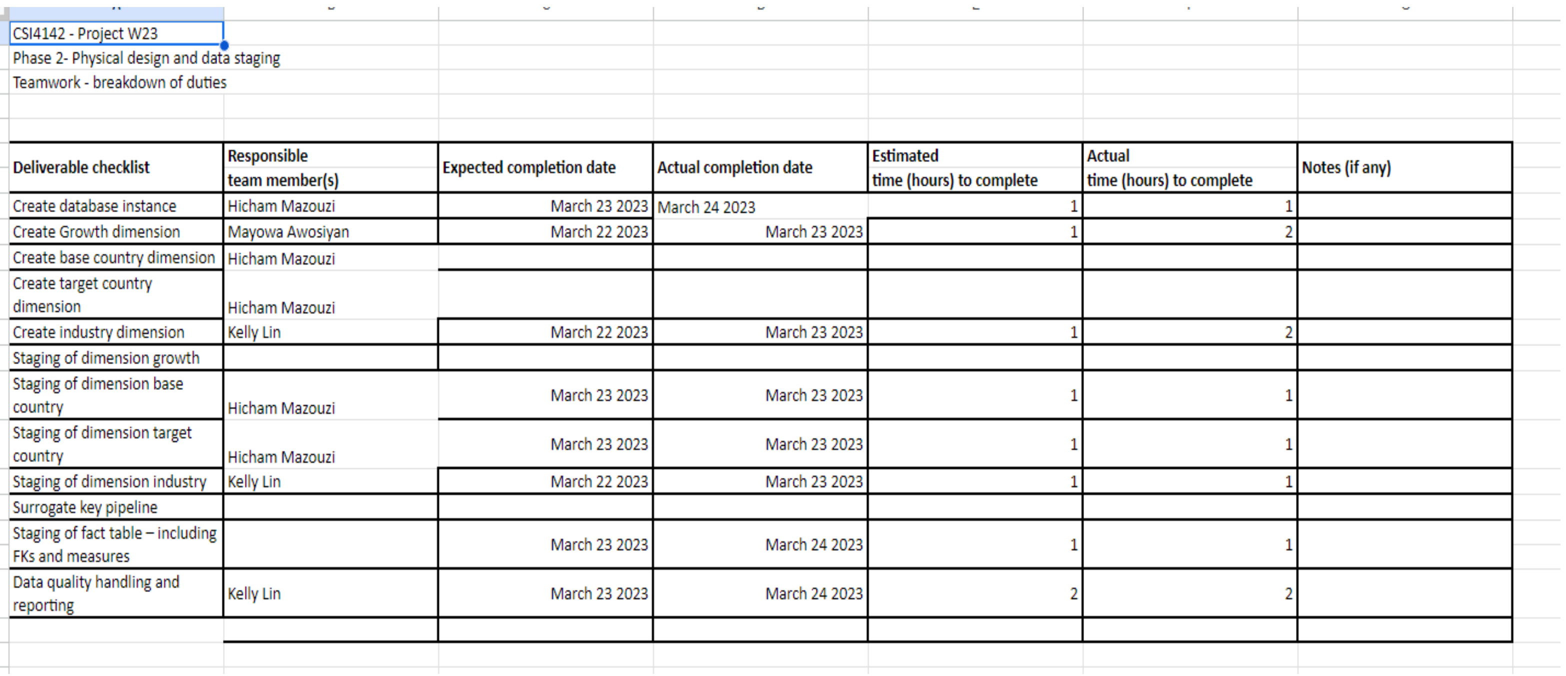
C - Quality issues

List of the quality issues encountered and how to handled

1. Missing data: We encountered missing values when merging two csv files. We found these missing values by using the data.isnull().sum() function so that we can know where the missing values are. And we use dropna() to drop the missing values
2. Noisy Data: We removed some data we deemed unnecessary like country codes in one of our dimensions because they are just another way to identify the countries but we already have a country name column in that same dimension.
3. Data integration: We need to integrate data from different sources, we use pandas.merge() to integrate two different data . We will also use the data.isnull().sum() to check if there is missing data for the combined file

4. cleaning data: we clean the data by using data.str.strip().str.lower() and data.str.replace() to replace the unnecessary word such as “and”
5. duplicate data : we found the duplicate data by using data.duplicated().sum and remove it by data.drop_duplicates()

D - Team Planning



CSI4142 - Project W23
Phase 2- Physical design and data staging
Teamwork - breakdown of duties

Deliverable checklist	Responsible team member(s)	Expected completion date	Actual completion date	Estimated time (hours) to complete	Actual time (hours) to complete	Notes (if any)
Create database instance	Hicham Mazouzi	March 23 2023	March 24 2023	1	1	
Create Growth dimension	Mayowa Awosiyan	March 22 2023	March 23 2023	1	2	
Create base country dimension	Hicham Mazouzi					
Create target country dimension	Hicham Mazouzi					
Create industry dimension	Kelly Lin	March 22 2023	March 23 2023	1	2	
Staging of dimension growth						
Staging of dimension base country	Hicham Mazouzi	March 23 2023	March 23 2023	1	1	
Staging of dimension target country	Hicham Mazouzi	March 23 2023	March 23 2023	1	1	
Staging of dimension industry	Kelly Lin	March 22 2023	March 23 2023	1	1	
Surrogate key pipeline						
Staging of fact table – including FKs and measures		March 23 2023	March 24 2023	1	1	
Data quality handling and reporting	Kelly Lin	March 23 2023	March 24 2023	2	2	