University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Fundamentals of Data Science
Project Phase 2: Design and Data Staging
Due Date: March 22, 2024, 11:59pm

Tengyang Deng 300156567
Van De Lande, Eric 300068600
Wenbo Yu 300161788

A. High-Level Data Staging Plan Schematic

Introduction to the Data Staging Plan

- In this stage, the main purpose is to integrate two original sources as one and then transform it to the expected data mart and generate desired dimensions. During the integration, the data cleaning is applied to both datasets by dropping unnecessary or duplicate attributes. After integration, the data discretization and feature engineering is applied to the generated data frame in aspects of temperature and emission. Meanwhile, the summarization and aggregation is also implemented to transform the dataset. The final step is to load the dataframe into DBMS as a data mart and use queries for creating dimensions.

Extraction:

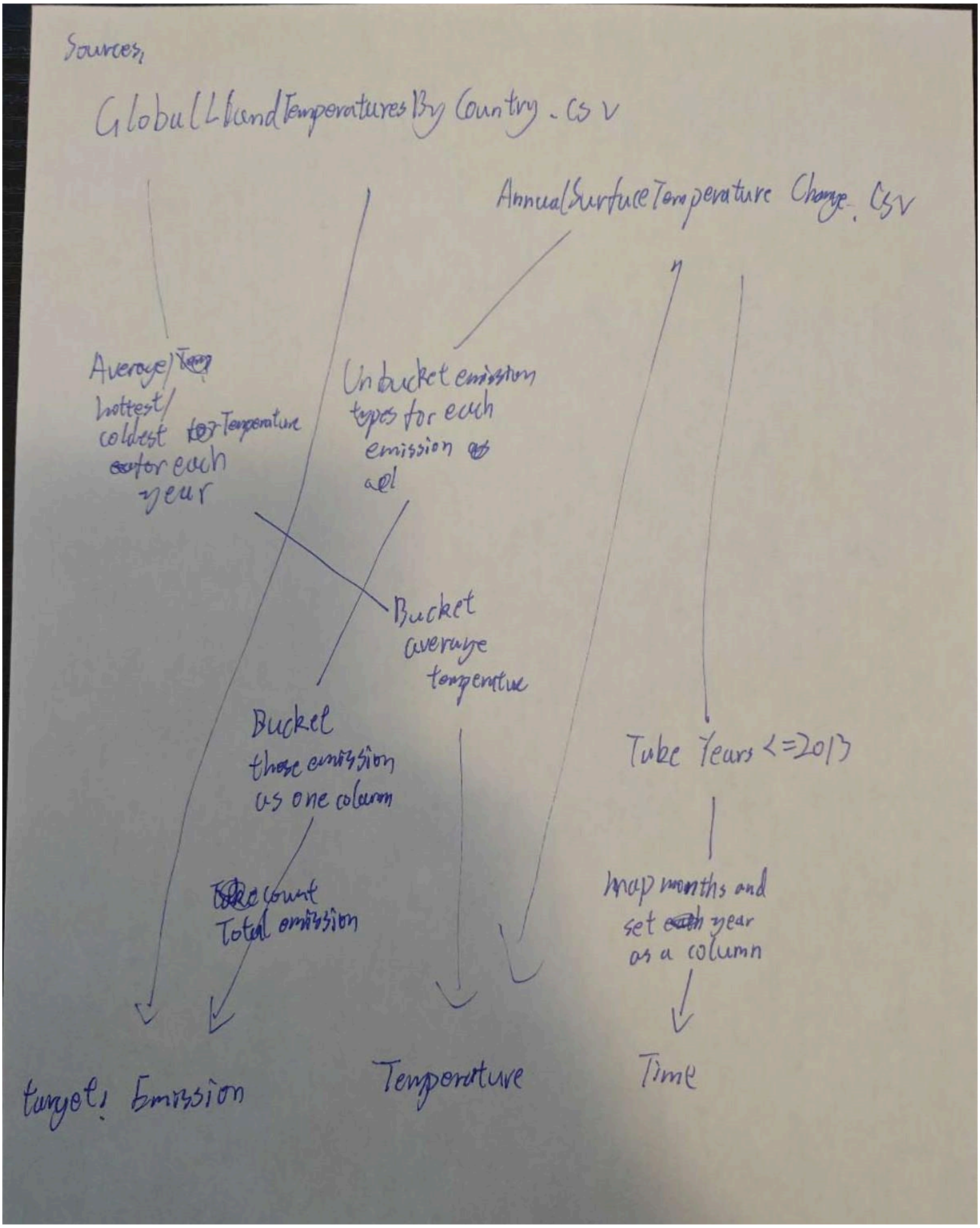
Extraction sources:

- AnnualSurfaceTemperatureChange.csv
- GlobalLandTemperaturesByCountry.csv

Extraction methods:

- Extracting Data from CSV Files by GitHub
- Working with Jupyter Notebooks
- data cleaning and integration
- Data Validation and Quality Checks
- Staging the Extracted Data
- Data discretization
- Feature engineering
- Aggregating or summarizing data

One-page schematic:


Sources:
GlobulLLandTemperaturesBy Country.csv
AnnualSurfaceTemperature Change.csv
Average/~~Temp~~ hottest/ coldest ~~for~~ Temperature ~~for~~ for each year
Unbucket emission types for each emission ~~as~~ col
Bucket average temperature
Bucket these emission as one column
~~Take~~ count Total emission
Take Years <=2013
map months and set ~~each~~ year as a column
target: Emission
Temperature
Time

B. Additional Details

DBMS and Data Warehousing Choices

- This project uses the relational-database management system (RDBMS) and using SQL for querying and management

- Data warehousing structure and design uses star schema for simplicity and performance.
    - Fact table: Fact_Emissions
    - Country_Dimension
    - Emissions_Dimension
    - Month_Dimension
    - Tmperature_Dimension
    - Time_Dimension

C. Data Quality Issues and Solutions

Encountered Data Quality Issues

- Missing values of temperature data (Nan or not applicable)
- Duplicates data of emission amount for each year

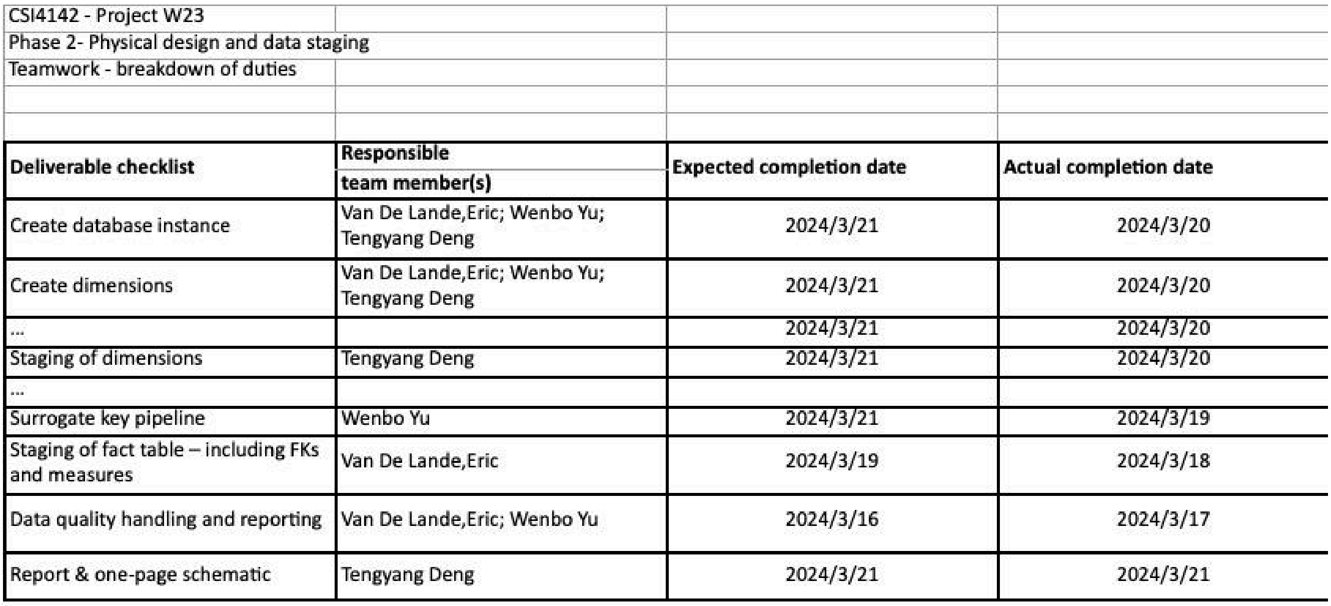
Detection and Resolution

- Find the issue by viewing data (observation) from integrated dataset
- Using drop command to drop redundant data

Data Integration from Different Sources

- Applying left join to emission and temperature datasets
- Using Country and year as key and drop duplicate rows for integrated dataset. Then extract a dataframe from it and melt, organizing its emissions per year. Finally integrate the data frame with the new dataset back.

Work Distribution:

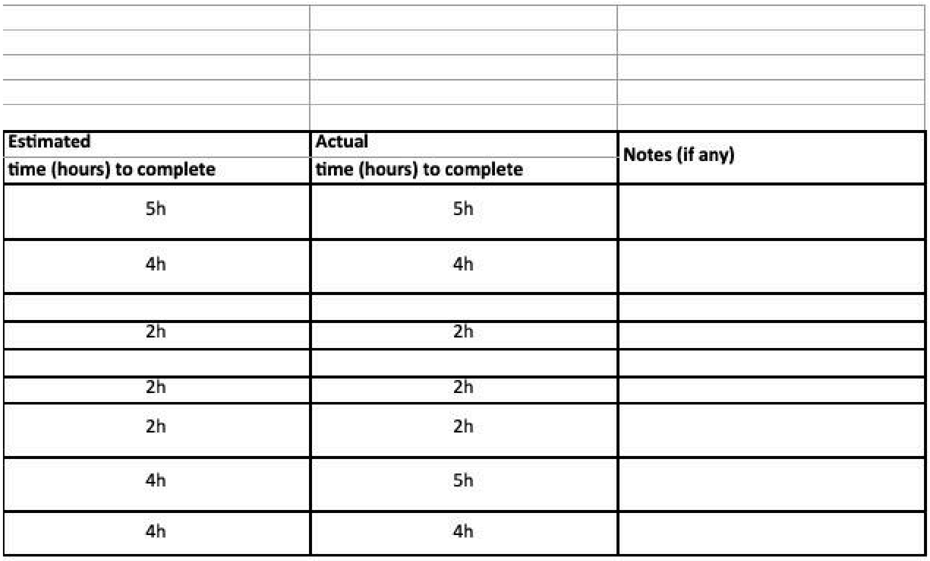


CSI4142 - Project W23

Phase 2- Physical design and data staging

Teamwork - breakdown of duties

Deliverable checklist	Responsible	Expected completion date	Actual completion date
	team member(s)		
Create database instance	Van De Lande,Eric; Wenbo Yu; Tengyang Deng	2024/3/21	2024/3/20
Create dimensions	Van De Lande,Eric; Wenbo Yu; Tengyang Deng	2024/3/21	2024/3/20
...		2024/3/21	2024/3/20
Staging of dimensions	Tengyang Deng	2024/3/21	2024/3/20
...			
Surrogate key pipeline	Wenbo Yu	2024/3/21	2024/3/19
Staging of fact table – including FKs and measures	Van De Lande,Eric	2024/3/19	2024/3/18
Data quality handling and reporting	Van De Lande,Eric; Wenbo Yu	2024/3/16	2024/3/17
Report & one-page schematic	Tengyang Deng	2024/3/21	2024/3/21



Estimated	Actual	Notes (if any)
time (hours) to complete	time (hours) to complete	
5h	5h	
4h	4h	
2h	2h	
2h	2h	
2h	2h	
4h	5h	
4h	4h	

References

- https://www.kaggle.com/datasets/rafsunahmad/global-yearly-temperature-change-in-the-surface
- https://github.com/EricVan14/ClimateChangeDataMart/blob/wenbobranch/ClimateChangeDataMart%20.ipynb
- https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data/data?select=GlobalLandTemperaturesByCountry.csv