CS 329E Midterm Presentation

Heng Cai & Chris Cherry

Warehouse Project

The goal of this project was to **transform raw datasets into a fully functional data warehouse** that provides **insightful economic and financial analysis**.

This project was completed in **five phases**, each refining the data step-by-step:

- **Project 1:** Finding datasets and setting up cloud resources.
- **Project 2:** Extracting and validating structured data.
- **Project 3:** Cleaning and staging data by fixing anomalies.
- **Project 4:** Structuring relationships and creating universal identifiers.
- **Project 5: Building the mart layer**, which provides user-friendly dashboards for analysis.

In this presentation, I'll walk through each phase, **showing key challenges, solutions, and a live demo of the final marts**."

Project 1: Data Collection & Warehouse Setup

In **Project 1**, we started by **choosing a dataset domain** related to **state-level economic trends**.

we selected **nine major datasets**, covering:

- **Crime**
- **Demographics**
- **GDP**
- **Mortality and Natality**
- **Energy onsumption**
- **Budget**
- **Employment**
- **Cities**
- **Geo-location**

To store this data, we:

✔ **Created a Google Cloud Storage (GCS) bucket** for raw file uploads.
✔ **Set up a BigQuery project** to manage the warehouse.
✔ **Defined a structured data dictionary** to track table attributes.

A key challenge was **finding datasets that were logically related** and contained **both structured and unstructured data**.

Project 2: Extracting & Loading Structured Data

With the datasets collected, **Project 2** focused on **extracting structured data** and **ensuring validation criteria were met**.

We extracted and loaded the data following these steps:

1. **Converted unstructured data (PDF, JSON) into tabular formats** using the LLM.
2. **Ensured dataset diversity** by selecting **multiple independent sources**.
3. **Validated the schema** to ensure each table was correctly formatted.
4. **Loaded structured datasets into BigQuery**, storing them in a **raw dataset**.

One of the biggest challenges was **handling inconsistent column formats across sources**.
To solve this, we used **schema mapping** to standardize field names and ensure smooth integration.

Project 3: Cleaning & Staging Data

Once the raw data was loaded, **Project 3** focused on **cleaning and staging the data** to ensure it was **ready for analysis**.

We applied **three key transformations**:

✔ **Fixed data type mismatches** – Converting string-based numbers to proper numeric types.
✔ **Handled missing values** – Replacing empty values with `NULL`.
✔ **Split multi-value fields** – Breaking down combined attributes into separate columns.

The cleaned data was stored in a **staging dataset**, serving as an **intermediate layer before analysis**.

Project 4: Entity Decomposition & Normalization

In **Project 4**, We focused on **normalizing data and improving entity relationships** to ensure efficient queries.

Key improvements included:

✔ **Merged duplicate records** from different datasets using **universal identifiers**.
✔ **Separated multi-entity tables** into properly structured tables to reduce redundancy.
✔ **Flattened nested lists** into separate tables for better query performance.

For example, instead of storing a **single column with multiple crime categories**, we split it into **separate fields** for **crime against persons, property, and society**.

Project 5: Building the Mart Layer

Finally, **Project 5** involved creating the **mart layer**, which translates cleaned data into **business-friendly insights**.

We designed **10 marts**, each answering a key business question:

- **Which states have the highest and lowest GDP per capita?**
- **How does crime correlate with GDP and unemployment rates?**
- **What are the most energy-intensive states and sectors?**
- **How does government spending relate to economic performance?**
- **Which states have the highest mortality and natality rates?**
- **How do crime trends evolve over time in relation to economic indicators?**
- **How do state unemployment rates trend over time?**
- **Which states have the highest tax-funded expenditures per capita?**
- **Which states have the most balanced economy across energy consumption, GDP, and employment?**
- **What are the key economic drivers of population growth?**

Conclusion

In summary, this project transformed raw datasets into a **structured and insightful mart layer**.

Each phase **incrementally refined the data**: ✔ **Project 1:** Data collection & setup.
✔ **Project 2:** Extracting structured data.
✔ **Project 3:** Cleaning & staging data.
✔ **Project 4:** Structuring relationships & fixing anomalies.
✔ **Project 5: Building marts for analysis.**

Thank you for listening!