Project Report

Data Warehousing and Mining

ETL

Spanish Flu Pandemic

Members : Muhammad Wasiq Patel, Muhammad Hamza Mushtaque, Abdul Rehman Mehtab
Roll Numbers: CT-096, CT-092, CT-05
Section: B
Teacher: Sir Umer

Project Scope:

Create an ETL pipeline using any programming language that has the following properties
1) Extracting data from any three sources
2) Applying basic transformation operations, such as conversion, summarization, and enrichment.
3) Load data into any database, such as Oracle Apex, MongoDB, or Firebase
4) Apply different data warehouse queries (Hint: lab work).
5) Import the results into power BI.
6) Apply different visualization techniques to display your results.
7) Create the dashboard / publish reports.
Note:
Activity is based on the group. In each group, the maximum number of people is not more than four.
The business process must be different for each group.

Project Thesis:

Abstract:

The Spanish Flu pandemic of 1918 was a catastrophic global event that caused substantial morbidity and mortality worldwide. Understanding historical pandemics like the Spanish Flu aids in comprehending the trajectory of infectious diseases. This project aimed to create an ETL (Extract, Transform, Load) pipeline to extract, process, and analyze data related to the Spanish Flu pandemic, providing insights into its impact on various regions and demographics.

The Spanish Flu pandemic, caused by the H1N1 influenza virus, emerged during the final stages of World War I and led to an unprecedented loss of life. Its global impact on public health, healthcare systems, and societal structures was profound. Analyzing historical data from this pandemic offers valuable lessons and insights applicable to contemporary public health challenges.

Objectives

- Data Extraction: Gather historical datasets, including records of infections, mortality rates, geographic spread, demographic information, and public health interventions.
- Data Transformation: Clean, format, and enrich the extracted data to enable meaningful analysis. This includes data normalization, consolidation, and the creation of relevant metrics.
- Data Loading: Store processed data into a structured database for efficient retrieval and analysis.
- Analysis and Visualization: Utilize the processed data to derive insights regarding the Spanish Flu's impact on different regions, demographics, and societal responses.
- Report and Visualization: Create a comprehensive report and visualizations to present findings effectively.

Methodology

1. Data Extraction

- Sources: Kaggle
- Data Types: Infection rates, mortality statistics, geographical data, interventions, demographic breakdowns, and historical context.
- Tools: Python libraries for web scraping, data retrieval, and data import/export.

2. Data Transformation

- Cleaning and Preprocessing: Address missing values, standardize formats, and ensure data consistency.
- Enrichment: Combine datasets, calculate mortality rates, create demographic indicators, and derive additional relevant metrics.
- Normalization: Standardize data structures for seamless analysis.
- Tools: Python's Pandas library for data manipulation and transformation.

3. Data Loading

- Database Selection: Utilize a relational database (e.g MongoDB) to store structured data efficiently.
- Schema Design: Create an optimized database schema for storing processed Spanish Flu data.
- ETL Process: Implement an ETL pipeline to load cleaned and transformed data into the database.
- Tools: PostgreSQL for database management, SQL for querying, and Python for data loading.

4. Analysis and Visualization

- Queries and Analysis: Formulate SQL queries to perform analytical tasks on the stored data.
- Visualization: Utilize tools like Power BI to create visual representations of key findings and trends.
- Insights: Identify patterns, hotspots, and correlations in the data to derive meaningful insights.

5. Report and Visualization

- Documentation: Compile a detailed report outlining the project's methodology, findings, and insights.

- Visualization Tools: Generate visual representations, such as graphs, maps, and charts, to convey the data-driven conclusions effectively.
- Presentation: Summarize key findings and their implications in a clear and understandable manner.

Conclusion

This ETL project on the Spanish Flu aimed to illustrate the process of extracting, transforming, and loading historical pandemic data to derive actionable insights. By analyzing the impact of the Spanish Flu using data-driven approaches, this project provides a framework for understanding the dynamics of past pandemics, offering valuable lessons for current and future public health crises.

References

Include references to datasets, academic papers, historical records, and any other sources used in the project for transparency and academic integrity.

About the Project

Our World in Data presents the data and research to make progress against the world's largest problems. Parts of the article — mostly the mortality estimates of the various influenza pandemics — were revised in May 2023.

In the last 150 years the world has seen an unprecedented improvement in health. The visualization shows that in many countries life expectancy, which measures the average age of death, doubled from around 40 years or less to more than 80 years. This was not just an achievement across the countries shown here; life expectancy has doubled in all regions of the world.

What also stands out is how abrupt and damning negative health events can be. Most striking is the large, sudden decline of life expectancy in 1918, caused by an unusually deadly influenza pandemic that became known as the ‘Spanish flu’.

To make sense of the fact life expectancy declined so abruptly, one has to keep in mind what it measures. *Period life expectancy*, which is the precise name for this measure, captures the mortality in *one particular year*. It summarizes the mortality in a particular year by calculating the average age of death of a hypothetical cohort of people for which that year’s mortality pattern would remain constant throughout their entire lifetimes.

This influenza outbreak wasn’t restricted to Spain and it didn’t even originate there. Recent genetic research suggests that the strain emerged a few years earlier, around 1915, but did not take off until later on. The earliest recorded outbreak was in Kansas in the United States in 1918.

But it was named as such because Spain was neutral in the First World War (1914-18), which meant it was free to report on the severity of the pandemic, while countries that were fighting tried to suppress reports on how the influenza impacted their population to maintain morale and not appear weakened in the eyes of the enemies. Since it is very valuable to speak openly about the threat of an infectious disease I think Spain should be proud that it was not like other countries at that time.

The virus spread rapidly and eventually reached all parts of the world: the epidemic became a pandemic. Even in a much less-connected world the virus eventually reached extremely remote places such as the Alaskan wilderness and Samoa in the middle of the Pacific islands. In these remote places the mortality rate was often particularly high.