Design and Analysis of a Large-Scale Data Engineering Pipeline for Earthquake Data

Name : Abdallah Naser Alqwasmi
Student ID : 149548
E-mail : analqwasmi20@cit.just.edu.jo

Name : Dalaa Hussam Badarneh
Student ID : 163586
E-mail : dhbadarneh22@ciit.just.edu.jo

Name : Mohammad Hussein Radaideh
Student ID : 165532
E-mail:mhalradaideh22@cit.just.edu.jo

***Abstract*—This project presents the design and implementation of a large-scale data engineering pipeline for earthquake data analysis. The dataset consists of more than two million records collected over multiple decades. The pipeline includes data ingestion, cleaning, transformation, and exploratory data analysis using Python-based tools. Correlation analysis, statistical distributions, and temporal analysis were applied to identify meaningful patterns in earthquake magnitude, depth, and geographic distribution. The results demonstrate the effectiveness of structured data engineering workflows in handling large real-world datasets and extracting reliable insights.**

Keywords—Data Engineering, Earthquake Data, Data Cleaning, Large-Scale Data, Exploratory Data Analysis

I. INTRODUCTION

Earthquake datasets are characterized by large volumes, heterogeneous structures, and data quality challenges. Raw seismic data often contains missing values, inconsistencies, and noise, which require systematic preprocessing before analysis. Data engineering plays a crucial role in transforming such raw data into a structured and reliable format suitable for analytical tasks.

The objective of this project is to design and implement a complete data engineering pipeline for large-scale earthquake data. The pipeline processes data from ingestion to analysis, enabling the exploration of spatial, temporal, and statistical patterns within the dataset.

This project follows a complete **data engineering workflow**, starting from raw data ingestion to data cleaning, transformation, storage, and analytical visualization. The main objective is to analyze global earthquake data and extract meaningful insights related to earthquake magnitude, depth, temporal behavior, and geographic distribution.

II.Data Engineerig Pipeline

The implemented data engineering pipeline consists of four main stages:
data ingestion, data cleaning, data storage, and data analysis.
The dataset was ingested from a **raw CSV** file containing global earthquake records.
After preprocessing and cleaning, the refined dataset was stored locally **in CSV format** and processed directly using Python libraries such as Pandas and Matplotlib for analytical tasks and visualization.

III. Methodology

A. Dataset Description

The dataset used in this project integrates earthquake records collected from multiple sources. The final dataset contains over **two million earthquake events** spanning several decades. Key attributes include time, latitude, longitude, depth, magnitude, and source-related metadata.

B. Tools and Technologies

The data engineering pipeline was implemented using Python and the following libraries:

- **Pandas** for data manipulation and preprocessing
- **NumPy** for numerical computations
- **Matplotlib and Seaborn** for visualization
- **Scikit-learn** for preprocessing utilities

C. Data Cleaning and Preprocessing

Several preprocessing steps were applied to ensure data quality:

- Handling missing and null values
- Removing duplicate records
- Converting data types to appropriate formats
- Filtering invalid or inconsistent numerical values

These steps significantly improved data reliability and consistency.

During the preprocessing stage, missing and inconsistent values were identified and handled appropriately. A noticeable number of missing values existed in attributes such as depth and magnitude before cleaning. These values were either filtered out or corrected based on logical constraints to ensure data reliability. Additionally, several transformations were applied, including extracting temporal features such as year, month, and day from the time attribute. Records with non-physical or unrealistic depth values were filtered to improve the overall quality of the dataset.

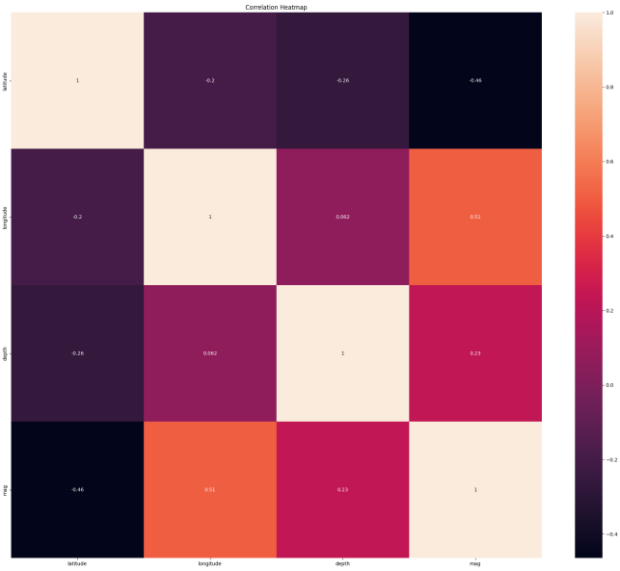
D. Data Transformation

After cleaning, the data was transformed to support efficient analysis. Temporal attributes were extracted, and numerical features were prepared for statistical analysis and visualization.

IV. Results

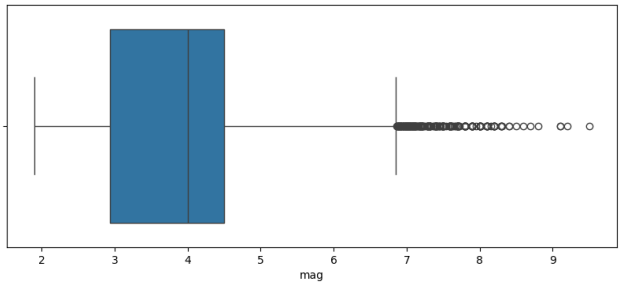
A. Correlation Analysis

Correlation analysis was performed to examine relationships between numerical features, including latitude, longitude, depth, and magnitude. The correlation heatmap shows that magnitude has a **moderate positive correlation with longitude (0.51)** and a **weak positive correlation with depth (0.23)**. A **moderate negative correlation (-0.46)** was observed between magnitude and latitude.



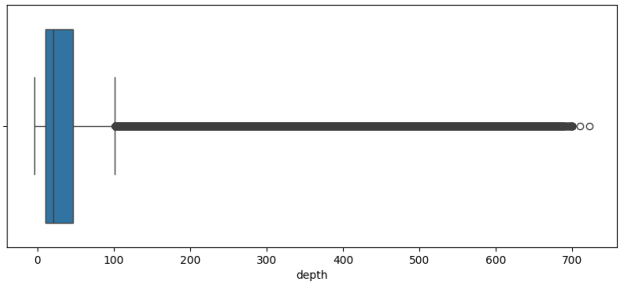
B. Magnitude Distribution

The magnitude distribution analysis shows that most earthquakes have magnitudes between **3.0 and 4.5**, with a median close to **4.0**. High-magnitude earthquakes are rare, with outliers reaching up to **9.0**.



C. Depth Distribution

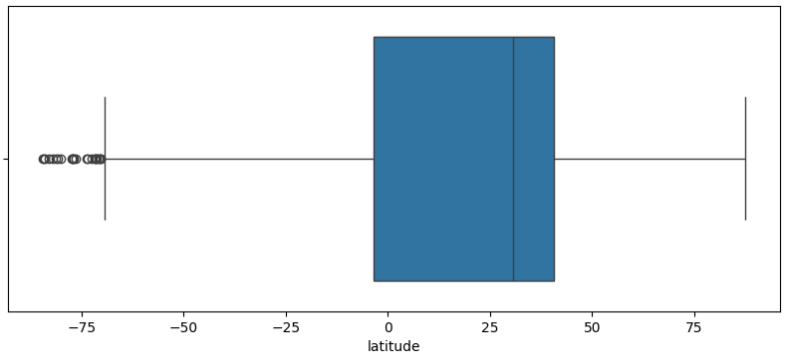
Depth analysis indicates that the majority of earthquakes occur at shallow depths below **100 km**, while deep-focus earthquakes extend beyond **700 km** as outliers.

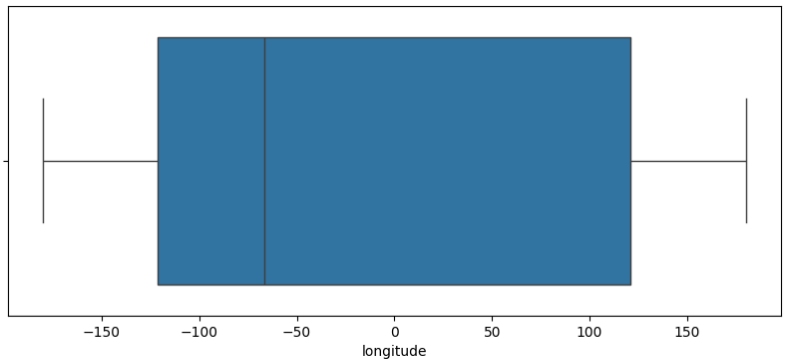


D. Geographic Distribution

Geographic analysis reveals a global distribution of earthquakes:

- Latitude values range approximately from **-60° to +85°**
- Longitude values range from **-180° to +180°**

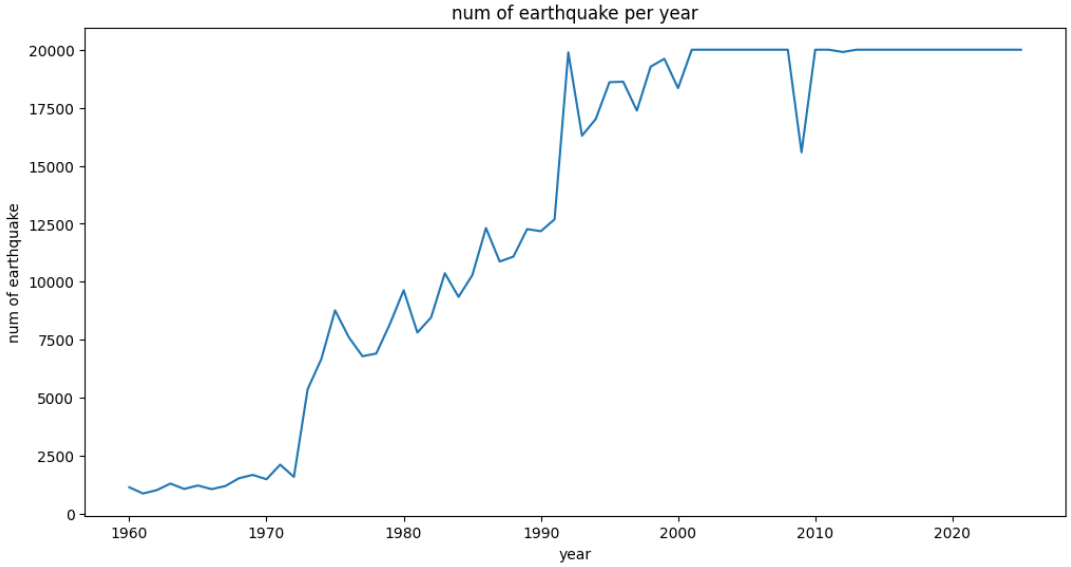




E. Temporal Analysis

Temporal analysis shows a clear increase in the number of recorded earthquakes over time:

- **1960s** : approximately **900–1,500 events/year**
- **1970s** : rapid increase beyond **5,000 events/year**
- **1990s onward** : stabilization around **18,000–20,000 events/year**



V. Discussion

The weak correlation between earthquake magnitude and geographic coordinates aligns with geophysical principles, as earthquake strength is mainly influenced by tectonic stress release rather than exact location. Shallow earthquakes occur more frequently due to crustal plate interactions, while deeper earthquakes are less common and usually associated with subduction zones.

The skewness observed in depth distribution supports the fact that most seismic activity occurs near the Earth's surface, making shallow earthquakes more impactful on human populations.

V. Conclusion

This project presented the design and implementation of a complete data engineering workflow for large-scale earthquake data analysis. Starting from raw data ingestion, the dataset was cleaned, transformed, and stored in a structured format suitable for analytical processing.

Through exploratory data analysis, meaningful insights were extracted regarding earthquake magnitude, depth, geographic distribution, and temporal behavior. The results demonstrated that most earthquakes occur at shallow depths, while extreme seismic events are rare but significant. Additionally, the analysis showed that geographic location has a limited linear relationship with earthquake magnitude, aligning with known geophysical behavior.

Overall, this project highlights the critical role of data engineering practices in managing large and complex datasets and enabling reliable data-driven analysis. The implemented workflow provides a solid foundation for future extensions involving real-time data processing, scalable storage, and predictive analytics.

Acknowledgment

The authors would like to thank **Dr. Ala'a Alhowaide**, instructor of the **Data Engineering** course, for his guidance and support throughout this project.