Earthquake Data Management and Analysis Project

Bilal Khan / 20100416 / BAK167

The study of earthquakes is vital for understanding natural disasters and preparing for their impacts. This project aims to manage and analyze earthquake data to uncover patterns, trends, and potential impacts. The project's primary focus is to collect, store, clean, transform, analyze, and visualize earthquake data using various data science techniques and tools. This report provides an overview of the projects stepss, implementation, and preliminary findings.

The project began with the collection of earthquake data. A CSV file containing earthquake data from 1995 to 2023 was used as the primary data source. The file was loaded into a Pandas DataFrame for manipulation and analysis. The first inspection of the data showed multiple attributes, like magnitude, date and time, location, depth, and additional information like alert levels and tsunami warnings. This comprehensive dataset provided a solid foundation for further more analysis.

To ensure the data was stored in a structured and accessible manner, it was saved into a SQLite database. This step facilitated easy retrieval and manipulation of data for future queries and analysis. Using a relational database helped in maintaining the data integrity and allowed for efficient data management practices.

Data cleaning was a crucial step to prepare the dataset for analysis. This involved handling missing values and detecting and handling outliers. Missing values were particularly present in the 'alert', 'location', 'continent', and 'country' columns. Rows with missing location data were dropped, while missing values in the 'alert' column were filled with 'unknown', and missing values in the 'continent' and 'country' columns were filled with 'Unknown'. Outliers in the magnitude data were identified and handled using the interquartile range method. This cleaning process ensured that the dataset was consistent and ready for further analysis.

The data transformation step involved creating new features and scaling numerical values. A new feature, 'day_of_week', was created from the 'date_time' column to analyze the distribution of earthquakes over different days of the week. The 'magnitude' values were scaled to a common range to facilitate comparison and analysis. This step enhanced the dataset by adding meaningful features and normalizing the data, making it more suitable for analysis and visualization. Exploratory Data Analysis was conducted to visualize data trends and distributions. Using Matplotlib, a histogram of earthquake magnitudes was created to visualize their distribution. The histogram revealed that most earthquakes in the dataset had magnitudes between 5.0 and 7.0, with a few outliers having higher magnitudes. This visualization helped me in understanding the frequency and distribution of earthquake magnitudes, providing valuable insights into earthquake patterns. An optional part of the project was to perform sentiment analysis on social media data related to earthquakes. Although social media data was not provided in the initial dataset, the project included a framework for performing sentiment analysis using TextBlob. This analysis aimed to understand public sentiment and reactions to earthquakes, which could be valuable for emergency response and understanding public

perception of earthquakes. To enable real-time monitoring and analysis, the project incorporated Apache Kafka for real-time data streaming. A Kafka producer was set up to stream earthquake data in real-time, allowing the system to handle live data feeds. This feature aimed to provide up-to-date information on earthquakee occurrences, enhancing the project's capability to monitor and respond to earthquakes as they happen. The final step involved creating a real-time dashboard using Plotly Dash. The dashboard provided an interactive and dynamic way to visualize earthquake data, making it easier to interpret and analyze trends over time. The dashboard was set up to update in real-time, displaying the latest data on earthquake magnitudes and other relevant features. This visualization tool aimed to provide a view of earthquake patterns, aiding researchers and decision makers in understanding and responding to earthquakes.

This project had a straightforward approach to managing and analyzing earthquake data. From data collection and storage to cleaning, transformation, analysis, and visualization, each step was implemented to ensure the data was handled effectively. The preliminary results provided valuable insights into earthquake patterns and trends, laying the groundwork for furthesr research and analysis. Future work could involve expanding data sources, performing advanced statistical and machine learning analyses, and integrating the real-time dashboard with emergency response systems. This project highlights the importance of data management in understanding natural disasters and preparing for their impacts

```
160    print(earthquake_data.head())
161
```

```
title  magnitude         date_time  ...
0  M 7.0 - 18 km SW of Malango, Solomon
Islands        7.0  22-11-2022 02:03  ...
1      M 6.9 - 204 km SW of Bengkulu,
Indonesia        6.9  18-11-2022 13:37  ...
```