**Introduction to Big Data and Data Science**

**Data Analysis on Earthquake Dataset**

**Project Proposal**

**Group 18**

**Team Members:**

1. Ashoka Chakravarthy Sanjapu
2. Tejas Reddy Vepala
3. Meghana Kurapati

**Goals and Objectives**:

**Motivation**:

We choose the earthquake dataset for our project. We are aiming to analyze the earthquakes data in the history to have a better understanding of seismic patterns and possible predictions. The use of big data analytics and tools in the collaboration with machine learning algorithms is key to the success of this project. We plan on gradual development throughout the semester by working on different aspects to fulfill the requirement. Although there have been many technological advancements, forecasting information on earthquakes and their activity has been as imprecise as ever. We plan on uncovering patterns that could lead to new discoveries or forecasting information.

**Significance:**

Earthquakes cause huge loss to the environment and the human lifestyle. To prevent loss to an extent we need accurate forecasting to avoid any causalities or to do damage control. This is the reason why it is very important to access the risks and predicting methodologies. This projects insights that we gain at the end are important to mitigate the issues like economic losses, causalities and creating a preparedness in regions that are prone to have high seismic activity. This can lead to the creation of alarm systems that will help in evacuations and emergency responses.

**Objectives:**

We will delve into the dataset to find correlations and trends in the number of earthquakes in the past years. We will have to put extra attention towards the magnitude, timing, and geolocation. We are opting to use big data tools and frameworks such as Hadoop, pyspark etc. in this project. We will create a machine learning model and integrate it. We are trying to achieve an accuracy percentage that can outweigh previous algorithms. Our final objective or goal is to create a system that will help users to predict an earthquake at a certain geolocation and timing. We might as well add more models to this project to investigate the information from a different perspective for better assessment.

**Features:**

Analytical Models: using machine learning algorithms to generate pattern recognition in the earthquake or seismic activity to enhancing prediction accuracy.

Real Time Analysis: Giving outputs based on the real time data for locations and timings.

Data Integration: Integrating seismic data with geological data to provide in-detail analytics.

Deliverables - Risk Maps of Seismic activities: Generating risk maps region wise and magnitude wise.

Analysis Report: Report on the findings and irregularities in the prediction and analysis phase.

Technical Milestones

* Data Collection and preprocessing.
* Machine Learning in depth usage on the data.
* Model Training and Deployment.
* Inculcating the use of Hadoop and other tools.

**Integration of Hadoop and Data Science:**

The Integration between Hadoop and data science plays an important role in our Earthquake data analysis project. The Hadoop components like HDFS, MapReduce, Hive and Spark are the key components. Seismic data is primarily stored in HDFS, with MapReduce enabling shared processing. By using Hive, we can process the SQL-based querying and Machine Learning algorithms and Advanced analytics steps are possible using Spark.

Machine Learning: We are planning to implement Classification algorithms such as Random Forest, Decision Trees and Naive Bayes Algorithms. We will work on more algorithms as well according to the data.

To efficiently handle massive amounts of data, the seismic dataset is ingested into HDFS, utilizing its distributed storage capabilities and MapReduce is to complete data cleaning, Extraction of features and Transformation steps. And then we can integrate with Apache Spark for Machine learning models, and We use the seismic dataset to create and train predictive models using Spark's machine learning libraries like **MLlib** or **ML.** After exporting data from Hadoop then we import the data in google colab and complete all the data preprocessing and machine learning tasks in google Colab.

**Workflow:**

A diagram of a software process

Description automatically generated

**References:**

1. Earthquake Prediction Using Machine Learning by unogeeks

https://unogeeks.com/earthquake-prediction-using-machine-learning/

1. Sisodiya, N., Dube, N., Prakash, O. et al. Scalable big earth observation data mining algorithms: a review. Earth Sci Inform 16, 1993–2016 (2023).