ProjectProposal

Daily Covid data analysis

Project members:

Rasmus Laitila, rasmus23002

Iiro Käki (iiro23001)

Lauri Raatikainen

**Dataset Name & Source:** worldometer\_coronavirus\_daily\_data.csv, contains COVID-19 information, including cumulative total infections, daily infections, cumulative total deaths, and daily deaths for each country from February 15, 2020, to May 14, 2022. The dataset is sourced from publicly available COVID-19 tracking databases such as Our World in Data, the WHO, or national health agencies.

**Justification for Choosing This Dataset:** This dataset was chosen due to its high relevance to public health, its large volume of records, and its structured format, making it ideal for Big Data analysis. Additionally, the dataset provides an opportunity to analyze the spread and impact of the pandemic on a global scale, which remains a critical area of study.

### **Problem Statement**

**What Problem Are You Solving?** The project aims to analyze the trends and patterns in COVID-19 infections and deaths across different countries. It seeks to identify key factors influencing infection rates, mortality, and the effectiveness of interventions over time.

**Why Is This Problem Important in a Big Data Context?** The COVID-19 pandemic generated massive amounts of real-time data, making it an ideal case for Big Data analytics. By leveraging Big Data technologies such as Apache Spark, we can process, analyze, and derive meaningful insights that traditional data processing tools would struggle with.

**Project Objectives**

* **Analyze** global and regional trends in COVID-19 infections and deaths.
* **Identify** correlations between infection rates, government interventions, and public health measures.
* **Utilize** Big Data technologies such as Apache Spark for large-scale data processing.
* **Visualize** trends through interactive dashboards and reports to enhance data-driven decision-making.

### **Expected Outcomes**

* **Detailed insights** into how COVID-19 evolved across different regions and time periods.
* **Predictive models** for forecasting infection and death trends using machine learning techniques.
* **Efficient data processing** pipelines using Spark and Hadoop to handle large-scale datasets.
* **Actionable recommendations** for improving pandemic response strategies based on data-driven findings.

This proposal serves as the foundation for our final Big Data project, ensuring a structured approach to data collection, analysis, and interpretation.