**Project Definition: Analyzing leading causes of deaths in Canada over 2 decades**

The project aims to analyze the leading causes of deaths in Canada from 2003 to 2022. The data will be stored in Hadoop Distributed File System (HDFS), explored and cleaned using Apache Hive, and analyzed using PySpark and DataFrame. The analysis will involve calculating descriptive statistics, aggregations, and exploring machine learning tasks such as classification, regression, or clustering based on the data characteristics.

**Project Details**

* **Data Storage**: The data on leading causes of deaths segregated by geography within Canada, age, and gender will be stored in HDFS for efficient storage and processing.
* **Data Exploration and Cleaning**: Apache Hive will be used to explore the data structures, clean the data, and prepare it for analysis using PySpark and DataFrame.
* **Data Analysis**: PySpark and DataFrame will be utilized to perform various analyses, including descriptive statistics, aggregations, and exploratory machine learning tasks.

**Analyses and Insights**

**1. Descriptive Statistics and Aggregations**

* Calculate and visualize the total number of deaths by leading causes over the two decades.
  + **Graph**: Bar chart or line graph showing the trend of leading causes of deaths over time.

**2. Age and Gender Analysis**

* Explore the distribution of leading causes of deaths by age and gender.
  + **Graph**: Stacked bar chart or grouped bar chart showing the distribution of leading causes of deaths by age and gender.

**3. Geographic Analysis**

* Analyze the leading causes of deaths by geography within Canada.
  + **Graph**: Choropleth map or bar chart showing the distribution of leading causes of deaths by region.

**4. Machine Learning Tasks**

* Experiment with classification, regression, or clustering based on the data characteristics to uncover patterns and insights.
  + **Insights**: Identify any significant trends, correlations, or clusters within the leading causes of deaths data.

**Key Findings to be Reported**

* Trends in leading causes of deaths over the two decades, highlighting any significant changes or patterns.
* Distribution of leading causes of deaths by age, gender, and geography within Canada.
* Insights from machine learning tasks, such as classification, regression, or clustering, revealing hidden patterns or correlations in the data.
* Potential implications for public health policies or interventions based on the analysis of leading causes of deaths in Canada.

By conducting the suggested analyses and exploring machine learning tasks, this project will provide comprehensive insights into the leading causes of deaths in Canada over 2 decades, enabling the identification of key trends, demographic patterns, and potential areas for targeted interventions or further research.

**Geography: Canada**

**Dataset:** <https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1310039401>

**Date: 2003 to 2022**

https://youtu.be/NDMvmnWa70M

https://www.youtube.com/watch?v=3mMTXsvPJzE&t=66s