Project Proposal

Title: Exploratory Analysis of Patterns and Correlations in STD Incidence Using Large-Scale Public Health Data

---

1. Introduction

Sexually Transmitted Diseases (STDs) remain a significant public health concern globally, with persistent disparities across regions, demographics, and socioeconomic groups. As surveillance systems improve and more public health data become available, there is a growing opportunity to harness these large-scale datasets to gain deeper insights into the spread, trends, and correlates of STD incidence.

This project proposes an exploratory Big Data analysis using multi-terabyte-scale datasets collected by national and state-level health departments to identify temporal, geographic, and demographic patterns and correlations in STD incidence across the United States.

---

2. Objectives

- To analyze and visualize patterns of STD incidence over time and across locations.
- To detect statistically significant correlations between STD rates and factors such as income level, education, race, population density, and access to healthcare.
- To build a dynamic, scalable pipeline capable of ingesting and processing terabyte-scale health data.
- To identify potential clusters and anomalies using machine learning techniques such as clustering (e.g., DBSCAN) and dimensionality reduction (e.g., PCA, t-SNE).

---

3. Justification as a Big Data Project

This qualifies as a Big Data project due to the **volume, variety, velocity**, and **veracity** of the datasets:

- **Volume**: The CDC and state health departments provide comprehensive public health records including patient-level reports, geocoded data, and lab results. Aggregating

across multiple years and jurisdictions, the data volume exceeds **5 TB**.

- **Variety**: Data includes structured (CSV, JSON), semi-structured (HL7, XML), and unstructured formats (PDF reports, scanned forms).
- **Velocity**: New data is generated and published on a rolling basis (e.g., weekly surveillance reports), necessitating automated ingestion pipelines.
- **Veracity**: Data quality issues due to underreporting, anonymization, and inconsistent metadata standards require robust preprocessing and validation.

---

4. Data Sources

- **CDC STD Surveillance Dataset (2010–2025)** – Over 1.5 TB
- **State-Level Department of Health Records** – ~2.3 TB combined
- **U.S. Census and American Community Survey Data** – ~0.7 TB
- **Electronic Medical Record (EMR) Data (De-identified, if accessible)** – ~0.8 TB
- **Social Determinants of Health (SDOH) Datasets from NIH and HHS** – ~0.5 TB

Total Estimated Dataset Size: ~5.8 TB

---

5. Methodology

Data Engineering

- Design and implement ETL pipelines using **Apache Spark** and **Hadoop HDFS** for distributed storage and processing.
- Use **Apache Airflow** to orchestrate regular updates and ingestion jobs.
- Store metadata and indexes in **Elasticsearch** for fast retrieval and filtering.

Data Cleaning and Preprocessing

- De-duplication, standardization of ICD-10 and HL7 codes
- Geospatial normalization using GIS libraries (e.g., GeoPandas, PostGIS)
- Handling missingness with imputation methods and uncertainty quantification

Exploratory Data Analysis

- Time series analysis for trends (e.g., seasonality, spikes)
- Heatmaps and choropleths of incidence rates
- Multivariate correlation analysis (e.g., Pearson, Cramér's V for categorical variables)

Machine Learning & Statistical Modeling

- Clustering techniques to identify STD hotspots
- Principal Component Analysis (PCA) to reduce dimensionality
- Regression models and tree-based classifiers to infer associations
- Use **MLlib**, **scikit-learn**, and **XGBoost**

---

6. Tools and Technologies

- **Distributed Storage & Processing**: Apache Hadoop, Apache Spark
- **Visualization**: Tableau, D3.js, Plotly, Kepler.gl
- **Data Management**: PostgreSQL/PostGIS, Delta Lake
- **Programming**: Python, SQL, PySpark, R
- **Cloud Infrastructure**: AWS S3, EC2, EMR or Google BigQuery and Cloud Storage
- **Security & Privacy**: HIPAA compliance, encryption, anonymization techniques

---

7. Anticipated Challenges

- **Data Sensitivity**: Ensuring full compliance with privacy regulations (HIPAA, FERPA).
- **Data Integration**: Aligning datasets from heterogeneous sources and formats.
- **Computational Load**: Requiring a scalable architecture to process multi-TB datasets efficiently.

---

8. Deliverables

- A reproducible, cloud-deployable analysis pipeline.
- Interactive dashboards showing incidence trends and hotspot detection.
- Technical report detailing analytical findings and statistical interpretations.
- Policy brief summarizing key public health insights for non-technical stakeholders.

---

9. Impact and Significance

This project will enable public health researchers and policymakers to better understand the dynamics of STD spread, uncover hidden patterns, and design targeted interventions. The use of Big Data techniques allows the handling of datasets previously too large or complex for traditional methods, making this project a foundational effort in the intersection of epidemiology and data science.

---

10. Timeline

Phase	Duration	Tasks
Data Acquisition & Setup	1 month	Access APIs, set up storage and processing infrastructure
Data Cleaning & Integration	1.5 months	ETL design, harmonization of formats

Exploratory Analysis & Visualization	1.5 months	EDA, trend mapping, correlation analysis
Modeling & Pattern Detection	1 month	ML and statistical modeling
Reporting & Presentation	1 month	Dashboards, written reports, dissemination

---

11. Team and Roles

- **Data Engineer** – Infrastructure setup, ETL pipelines
- **Data Scientist** – Analysis, modeling, interpretation
- **Public Health Expert** – Domain expertise and policy insight
- **Visualization Specialist** – Dashboards and data storytelling

---

12. Conclusion

This proposal outlines a Big Data project that leverages multi-terabyte public health datasets to understand and interpret STD incidence patterns. Through scalable architectures, robust statistical methods, and domain expertise, this project aims to deliver actionable insights for improving health outcomes across the U.S.