**Project Proposal** (CSP 554 Big Data)

**Group Name:** Chicago Crime Analysis

**Group Members:** Manish Tandel (Voice of the team), Chirag Bhansali, Bingxin Xu

**Project Topic:** Big Data Analysis of Chicago Crime Dataset

**Project Description**: The goal of this project is creating a data pipeline that reads data from the source, transforms it, and stores it in an optimal format to derive actionable insights from the data as mentioned in the project goal below. We will use all the necessary best practices for each technology to create an efficient pipeline. We aim to provide analysis that can be used by the police department to understand the various nature of the crimes going around the city.

**Data Source:** Chicago crime datasets is sourced from City of Chicago Public Safety website, which details all the crimes that happened in the city of Chicago since 2001. The dataset has 6.99 million records and 30 attributes with a total size of 1.7 GB.

**Big Data Tool:** HDFS, Hive, Pig, Spark

We will ingest data into HDFS, perform data profiling using Spark SQL, apply necessary transformations to denormalize the data using Spark, and store it in the Hive table so that it is accessible by reporting layer to derive insights. Besides data analysis, MLlib - Machine Learning Library will be utilized for prediction modeling.

*Problems we are going to address are as follows:*

1. What are the most occurring crimes in the city?
2. How many crimes are being committed at a specific location? (e.g. Street, residence)
3. Which crimes are being committed at a specific time of the day?
4. Which crimes are being committed at a specific day of the week?
5. At what locations are the crimes being committed at a specific time of the day?
6. At what locations are the crimes being committed on a given day of the week?
7. Analysis of a particular crime type over the years

*Question to be addressed by using Spark:*

1. What kind of crime a random person is likely to encounter at a given location, date and time? (Prediction)
2. What types of crimes are more likely to happen in which part of the city? (Clustering)

**References:**

[1] Mcclendon, Lawrence, and Natarajan Meghanathan. “Using Machine Learning Algorithms to Analyze Crime Data.” *Machine Learning and Applications: An International Journal*, vol. 2, no. 1, 2015, pp. 1–12., doi:10.5121/mlaij.2015.2101.

[2] Shyam Varan Nath. 2006. Crime Pattern Detection Using Data Mining. In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology (WI-IATW '06). IEEE Computer Society, Washington, DC, USA, 41-44. DOI: <https://doi.org/10.1109/WI-IATW.2006.55>

[3] Kim, Suhong & Joshi, Param & Kalsi, Parminder & Taheri, Pooya. (2018). Crime Analysis Through Machine Learning. 415-420. 10.1109/IEMCON.2018.8614828. Available at: <http://airccse.org/journal/mlaij/papers/2115mlaij01.pdf>

**Dataset:**

[4] Crimes - 2001 to present | City of Chicago | Data Portal.

url: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>