# Project Report
# Property Based Crime Analysis(2001-2010)

## Table of Contents

## I. Abstract

This Big Data Analytics project utilizes PySpark to analyze property crime data across India from 2001 to 2010. The primary objectives were to understand crime patterns, identify high-risk regions, and evaluate the effectiveness of law enforcement in recovering stolen assets. The methodology involved using PySpark for data processing, cleaning, and calculating key metrics such as property recovery rates. Initial analysis of 2,449 records revealed that Theft - Property was the most prevalent specific crime category, and the dataset demonstrated excellent quality with zero missing values or duplicate rows. The project successfully transformed raw crime statistics into meaningful, actionable intelligence for policy makers and law enforcement.

## II. Introduction

**Project Overview:** Property crime presents a significant challenge to public safety and economic stability. This project addresses the need for a scalable, data-driven analysis of crime trends over a decade-long period (2001-2010) across India. By leveraging Big Data Analytics (BDA) principles and the PySpark framework, this study provides a comprehensive view of property crime occurrences and the subsequent recovery efforts.

**Objectives:**

1. Pattern Identification: Analyze the crime types and geographical areas most affected by property crime.
2. Recovery Evaluation: Calculate and compare property recovery rates (both cases and value) to measure law enforcement efficiency.
3. Trend Analysis: Track how property crime trends evolved across different regions and crime categories over the ten-year period.

## III. Data and Methodology

### A. Dataset Description

The project utilized the 10_Property_stolen_and_recovered.csv dataset, which contains comprehensive records of property crimes reported across all Indian states and union territories from 2001 to 2010.

**Key Data Fields:**
- Area_Name: State or Union Territory
- Year: Year of occurrence (2001-2010)
- Group_Name: Major crime category (e.g., Burglary, Theft)
- Sub_Group_Name: Detailed crime category
- Cases_Property_Recovered: Number of cases where property was recovered - Cases_Property_Stolen: Number of cases where property was stolen - Value_of_Property_Recovered: Monetary value of recovered property - Value_of_Property_Stolen: Monetary value of stolen property

**B. Big Data Technology Stack**
- Framework: Apache Spark (PySpark)
- Environment: Jupyter Notebook
- Libraries: pyspark.sql, pandas, numpy

**C. Data Quality Assessment (Initial Analysis)**
- Total Records: 2,449
- Completeness: 100%
- Uniqueness: 100%
- Quality Rating: EXCELLENT
- Outliers: 13.84% to 16.86% in numeric columns

**D. Data Processing and Feature Engineering**
- Loading Data
- Cleaning (Handling Zeroes)
- Feature Creation (Recovery_Rate_Cases, Recovery_Rate_Value)

## IV. Results and Analysis

**A. Crime Type Distribution (2001–2010)**
Theft - Property accounted for over 65% of stolen property cases.

**B. Geographical and Temporal Scope**
Covers 35 regions and 10 years (2001-2010).

## V. Conclusion and Future Scope

**Conclusion:** This project demonstrated the potential of Big Data Analytics in understanding property crime trends across India from 2001 to 2010. Using PySpark, large-scale crime data was efficiently processed to reveal key insights such as the dominance of theft-related crimes, regional variations in crime incidence, and differences in recovery efficiency across states.

The analysis provides actionable insights for law enforcement, policy makers, and the public by highlighting crime patterns and recovery trends. The developed methodology not only showcases the scalability of PySpark in handling big data but also lays a strong foundation for future applications such as predictive crime modeling, geospatial analysis, and real-time crime monitoring.

**Future Scope:**

The scope of this project can be extended in multiple directions. Future work may include:

1. **Time Series Analysis –** Studying year-wise patterns in more detail to identify turning points in crime trends.
2. **Geospatial Analysis –** Mapping crime hotspots to visualize high-risk regions for better resource allocation.
3. **Predictive Modeling –** Using machine learning (PySpark MLlib) to forecast crime rates and recovery success.
4. **Real-Time Data Integration –** Adapting the pipeline to handle live crime reports for up-to-date analysis.

By implementing these extensions, the project can evolve into a **comprehensive crime analytics system**, supporting proactive law enforcement and smarter public safety policies.