Project Paper Draft

**Abstract:**

In this paper, we use Big Data technologies to analyze the Chicago crime dataset and derive useful insights from it. We address 9 questions in this study. The answers to these questions and visualization of the results provide useful information to the Chicago Police department. We use machine learning techniques like linear regression and K-means clustering to achieve our aim.

1. **Introduction:**

The Chicago Police Department's Bureau of Records has been keeping records of all the crimes in Chicago since the beginning of the 20th century.[1][10] This presents us a great opportunity to get our hands on a massive real-world data. The overall crime rate in Chicago, especially the violent crime rate, is higher than the US average.[2] Even though, the nation's crime rates remained near historic lows, in 2016, Chicago was responsible for nearly half of 2016's increase in homicides in the US.[3]

We identified the answers to 7 out of 9 questions can be found using either Pig scripts or Hive queries. We also realized we need to apply machine learning techniques to solve more advanced questions.

1. **Research Questions:**
   1. Questions to be addressed by using Hive, Pig
      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. At what locations are the crimes being committed at a specific time of the day?
      5. Which crimes are being committed at a specific day of the week?
      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

For visualizations, the results of the questions 2.1.3, 2.1.4, 2.1.5, 2.1.6 will be represented by heatmaps.

* 1. Question to be addressed by using Spark MLlib
     1. What kind of a crime random person is likely to face 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)

1. **Proposed framework:**

* 1. Storing and processing data:

HDFS:

The Hadoop Distributed File system (HDFS) is a distributed file system designed to run on commodity software. It is highly fault-tolerant and designed to be deployed on low-cost hardware. HDFS provides high throughput to application data and is highly useful for applications that have large datasets. It is reliable as the data is duplicated on multiple nodes in a cluster. HDFS provides an interface for applications to move themselves closer to where the data is located. HDFS has been designed to be easily portable from one platform to another. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. Moreover, it supports a very traditional hierarchical file organization.

* 1. Analysis of Data

Hive:

Hive is an essential tool in the Hadoop ecosystem which provides an SQL dialect called Hive Query Language. The developers can focus on writing the queries and Hive translates most queries to MapReduce jobs exploiting the scalability of Hadoop. Hive provides the basic SQL operations like filtering rows, selecting certain rows, joins, aggregation using group by. Hive enforces schema on read which makes for a fast initial load since the data does not have to be read, parsed, and serialized to disk in the database’s internal format.

Pig Latin:

Pig is a simple SQL-like scripting language that parses, optimizes, and automatically executes Pig Latin scripts as a series of MapReduce jobs on a Hadoop cluster. Pig is much easier and faster to write than MapReduce, and it's easier to decipher as well. Pig Latin is a data flow language- it allows users to describe how data from one or more inputs should be read, transformed, and after that stored to one or more outputs in parallel. In other words, a Pig Latin script is a directed acyclic graph (DAG).

Spark MLlib:

Mllib is Apache Spark’s machine learning library. Its goal is to make practical machine learning easy and scalable. It provides various tools at a high level like: ML Algorithms- common learning algorithms such as regression, classification, collaborative filtering, and clustering. Featurization- feature extraction, transformation, dimensionality reduction, and selection. Pipelines- tools for constructing, evaluating, and tuning ML pipelines. Persistence- saving and loading algorithms, models, and pipelines. Utilities- linear algebra, statistics, data handling, etc.

1. **Milestones for the project**

|  |  |  |
| --- | --- | --- |
| **Tasks** | **Owners** | **Due Date** |
| Search for a dataset, decide the approach and the technologies, finalize a problem statement | Everyone | 11/04/2019 |
| Project Draft and create a Git repository | Radhika Barbole, Shivam Kulkarni | 11/11/2019 |
| Data cleaning | Shubham Jain, Vivekanandan Sakthivel | 11/13/2019 |
| Move the cleaned data in HDFS | Everyone (As everyone needs to work independently on the dataset) | 11/14/2019 |
| Find answers to questions 2.1.1, 2.1.2, 2.1.3, 2.1.4 by using Hive | Shivam Kulkarni, Shubham Jain | 11/16/2019 |
| Find answers to questions 2.1.5, 2.1.6, 2.1.7 by using Pig Latin | Radhika Barbole, Vivekanandan Sakthivel | 11/16/2019 |
| Find answer to question 2.2.1 using Spark MLlib & perform unit testing | Radhika Barbole, Vivekanandan Sakthivel | 11/20/2019 |
| Find answer to question 2.2.2 using Spark MLlib & perform unit testing | Shivam Kulkarni, Shubham Jain | 11/20/2019 |
| Integration testing | Everyone | 11/21/2019 |
| Results visualization | Everyone | 11/22/2019 |
| Final paper | Everyone | 11/23/2019 |

1. **References**

**Websites**

1. [**^**](https://en.wikipedia.org/wiki/Crime_in_Chicago#cite_ref-5) [*"Crime in Chicago, Illinois (IL): murders, rapes, robberies, assaults, burglaries, thefts, auto thefts, arson, law enforcement employees, police officers, crime map"*](http://www.city-data.com/crime/crime-Chicago-Illinois.html)*.*
2. [**^**](https://en.wikipedia.org/wiki/Crime_in_Chicago#cite_ref-7) [*"Chicago Driving Uptick in Murders; National Crime Rate Stays Near 'Historic Lows'"*](https://www.usnews.com/news/articles/2016-09-19/chicago-drives-uptick-in-murders-national-crime-rate-stays-near-historic-lows)*. U.S. News. September 19, 2016.*
3. [**^**](https://en.wikipedia.org/wiki/Crime_in_Chicago#cite_ref-6) [*"Chicago Responsible for Nearly Half of U.S. Homicide Spike"*](http://time.com/4497814/chicago-murder-rate-u-s-crime/)*. Time.*
4. [*https://hadoop.apache.org/docs/r1.2.1/hdfs\_design.html*](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)
5. [*https://spark.apache.org/docs/latest/ml-guide.html*](https://spark.apache.org/docs/latest/ml-guide.html)

**Research Papers**

[6] 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.

[7] 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>

[8] 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>

[9] Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018). Crime Analysis Through Machine Learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). doi: 10.1109/iemcon.2018.8614828

**Dataset**

[10] Crimes - 2001 to present | City of Chicago | Data Portal. url: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>